Defining CP-CML Patient Subsets Associated with Poor Imatinib Uptake and Response

Dale B. Watkins

The Melissa White Laboratory, Cancer Theme

South Australian Health and Medical Research Institute (SAHMRI)

&

The Faculty of Health Sciences

Department of Medicine and

Centre for Personalised Cancer Medicine

The University of Adelaide

Adelaide, South Australia

A thesis submitted to The University of Adelaide

In candidature for the degree of Doctor of Philosophy

December 2013

TABLE OF CONTENTS

ABSTRACT

The introduction of tyrosine kinase inhibitor (TKI) therapy, specifically imatinib, has dramatically improved the treatment outcome for the majority of chronic phase chronic myeloid leukaemia (CP-CML) patients. Although most patients will achieve excellent clinical (haematological, cytogenetic and molecular) responses on imatinib, it is clear that a subset of patients will respond poorly, or fail imatinib therapy. Currently, up to 35% of patients treated with imatinib fit into this subset, displaying either primary or acquired resistance, leading to sub-optimal response or imatinib failure. The organic cation transport-1 (OCT-1) protein is the major active protein involved in imatinib transport. Measuring the function of OCT-1 in leukaemic mononuclear cells prior to imatinib therapy, expressed as OCT-1 activity (OA), has been demonstrated to be a strong prognostic indicator. Notably, low OA is strongly associated with patients at significant risk of poor molecular response, mutation development and leukaemic transformation during imatinib therapy. It is important to therefore determine what factors underlie the range of OA levels observed in CP-CML patients, and whether patients with very low OA and poor response to imatinib have different overall disease characteristics associated with alternative biological mechanisms.

The present study sought to 1) determine the variation in CP-CML patient immunophenotype at diagnosis, in relation to patient characteristics, including OA; 2) determine the gene expression patterns associated with OA, and identify new biomarkers for CP-CML; and 3) determine the global DNA methylation profile of CP-CML, with particular focus on very low OA, and ascertain whether aberrant epigenetic programming may underlie poor imatinib response. Specific lineage differences were identified, with patients defined as very low OA associated with a decreased T-lymphocyte signature, compared to very high OA. Furthermore, an up-regulated histone gene signature associated with very low OA was

identified. Gene expression analysis also identified *GFI1* as a novel biomarker for progression in CP-CML, as patients with low diagnostic *GFI1* expression in their white cells were at significant risk of disease transformation to blast crisis (BC), even when receiving TKI therapy. Additionally, significant differences in global DNA methylation patterns were identified between CP-CML and normal individuals; CP-CML patients with very low OA, compared to all other patients; and CP-CML versus BC. Importantly, this is the first report of global DNA methylation analysis in CML and identifies that aberrant epigenetic programming may have a significantly greater role in CML than originally first thought.

In conclusion, the findings detailed in this thesis provide further insight into the heterogeneity of CP-CML, and to a lesser extent OA. Additionally, a greater understanding of the possible factors influencing OA determination is presented, along with a new biomarker (*GFI1*) for disease progression to blast crisis in *de novo* CP-CML. Finally, the global DNA methylation results may present novel targets and pathways responsible for poor TKI response that will assist in developing new therapeutic strategies for *de novo* CP-CML patients, involving combination therapy to enhance patient outcome.

DECLARATION

I, Dale B. Watkins, certify that this thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968. The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Dale B. Watkins

1st December 2013

PUBLICATIONS

Manuscripts

Kok CH*, Watkins DB*, Leclercq T, D'Andrea RJ, Hughes TP, White DL. 2013. Low GFI1 expression in white blood cells of CP-CML patients at diagnosis is strongly associated with subsequent blastic transformation. *Leukemia*, 27(6):1427-30. Impact factor: 10.164.

* denotes equal first-authors.

Watkins DB, Hughes TP, White DL, D'Andrea RJ. 2013. NPM1 mutations occur rarely or not at all in chronic myeloid leukaemia patients in chronic phase of blast crisis. *Leukemia*, 27(2):489-90. Impact factor: 10.164.

Conference Abstracts

Watkins DB, Kok CH, D'Andrea RJ, Hughes TP, White DL. Global DNA methylation profiling in CP-CML, with emphasis on a poor-risk subset of CP-CML and during disease progression to Blast Crisis. *The Adelaide University, Faculty of Health Sciences Postgraduate Research Conference*, August 2013. Adelaide, SA. Poster presentation.

Watkins DB, Kok CH, D'Andrea RJ, Hughes TP, White DL. Global DNA methylation profiling in CP-CML, with emphasis on a poor-risk subset of CP-CML and during disease progression to Blast Crisis. Centre for Personalised Cancer Medicine (CPCM) Symposium, July 2013. Adelaide, SA. Poster presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Global DNA methylation analysis identifies key pathway differences between poor (Low OCT-1 Activity) and standard risk CP-

CML patients at diagnosis. *American Society of Hematology (ASH) Annual Meeting*, December 2012. Atlanta, GA, USA. Poster presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Global DNA methylation profiling in a poor-risk subset of CML. *HAA Annual Scientific Meeting*, October 2012. Melbourne, VIC. Oral presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Global DNA methylation profiling in a poor-risk subset of CML. *The Adelaide University, Faculty of Health Sciences Postgraduate Research Conference*, August 2012. Adelaide, SA. Poster presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Identification of differential lineage involvement and subsequent development of a predictive classifier for poor risk chronic-phase CML patients. *New Directions in Leukaemia Research (NDLR) Conference*, March 2012. Sunshine Coast, QLD. Poster presentation.

Watkins DB, Kok CH, Hughes TP, Slader C, D'Andrea RJ, White DL. Differential lineage involvement between very low and higher OCT-1 Activity chronic-phase CML patients. *American Society of Hematology (ASH) Annual Meeting*, December 2011. San Diego, CA, USA. Poster presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Development of a predictive classifier for poor risk chronic-phase CML patients at diagnosis using immunophenotyping. *HAA Annual Scientific Meeting*, October 2011. Sydney, NSW. Oral presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Development of a predictive classifier for poor risk chronic-phase CML patients at diagnosis using immunophenotyping. *The Adelaide University, Faculty of Health Sciences Postgraduate Research Conference*, August 2011. Adelaide, SA. Poster presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Immunophenotyping of chronic-phase chronic myeloid leukaemia patients at diagnosis identifies differential lineage involvement and predicts poor risk patients. *Australian Society for Medical Research (ASMR) SA Annual Meeting*, June 2011. Adelaide, SA. Oral presentation.

Watkins DB, Kok CH, Hughes TP, D'Andrea RJ, White DL. Immunophenotyping of chronic-phase chronic myeloid leukaemia patients at diagnosis identifies differential lineage involvement and predicts poor risk patients. *Royal Adelaide Hospital (RAH) Medical Staff Society Prize*, May 2011. Adelaide, SA. Oral presentation.

SCHOLARSHIP AND AWARDS

PhD Scholarship; The Leukaemia Foundation of Australia, 2010 – 2013

Support for the educational and professional development of researchers and other professionals undertaking a PhD. The award is to support research in Australia into the causes, treatment and care of people with leukaemia, lymphoma, myeloma and related blood disorders, and is awarded on the merits of the applicant and project proposal.

Dawes Top-Up Scholarship; RAH/IMVS Research Committee, 2011 – 2013

Top-up scholarships are awarded to applicants in receipt of a major external scholarship based on merit and research proposal.

SAHMRI (Poster) Prize; The University of Adelaide, Faculty of Health Sciences Postgraduate Research Conference, 2013

For the abstract entitled "Global DNA methylation profiling in CP-CML, with emphasis on a poor-risk subset of CP-CML and during disease progression to Blast Crisis", Adelaide, SA; August 2013.

Non-Member Travel Grant, HAA, 2012

Support for non-members to attend the annual HAA conference. Awarded on the basis of the submitted abstract entitled: "Global DNA methylation profiling in a poor-risk subset of CML", Melbourne, Victoria.

School of Medicine Poster Prize; The University of Adelaide, Faculty of Health Sciences Postgraduate Research Conference, 2012

For the abstract entitled “Global DNA methylation profiling in a poor-risk subset of CML”, Adelaide, SA; August 2012.

Leukaemia Foundation Poster Prize; New Directions in Leukaemia Research (NDLR) Conference, 2012

For the abstract entitled “Identification of differential lineage involvement and subsequent development of a predictive classifier for poor risk chronic-phase CML patients”, Sunshine Coast, QLD; March 2012.

Australian Postgraduate Award, Australia Government, 2010

Unable to accept due to acceptance of the LFA primary PhD scholarship.

ABBREVIATIONS

µ-bcr – Micro Breakpoint Cluster Region

µg – Microgram/s

µL – Microlitre/s

µM – Micromolar

[^{14}C] – Carbon-14 Radioactive Isotope

450K – Illumina Infinium® HumanMethylation450 BeadChip

ABC – ATP-Binding Cassette

ABL1 – Abelson Leukaemia Virus Proto-Oncogene Homolog 1

ADP – Adenosine Diphosphate

AGRF – Australian Genome Research Facility

ALL – Acute Lymphoblastic Leukaemia

AML – Acute Myeloid Leukaemia

AP – Accelerated Phase

Ara-C – Arabinofuranosyl Cytidine (Cytarabine)

ATCC – American Type Tissue Culture Collection

ATP – Adenosine Triphosphate

BC – Blast Crisis

BCR – Breakpoint Cluster Region

BCR-ABL1 – Fusion Gene

BCR-ABL1 – Fusion mRNA

Bcr-Abl1 – Fusion Protein

BH-FDR - Benjamini-Hochberg adjusted FDR

BM – Bone Marrow

BMA – Bayesian Model Averaging

bp – Base Pair

BSA – Bovine Serum Albumin

C – Celcius

CCyR – Complete Cytogenetic Response (absence of Ph-positive cells as measured by classical karyotyping or fluorescence *in-situ* hybridisation)

cDNA – Complementary DNA

CGI – CpG Island

ChIP – Chromatin Immunoprecipitation

CHR – Complete Haematological Response (sustained and significant reduction in WBCs to a normal range)

CMA – Classification for Microarrays

CML – Chronic Myeloid Leukaemia

CMR – Complete Molecular Response (BCR-ABL1 mRNA levels negative in 2 consecutive assays)

CP – Chronic Phase

CpG – Cytosine–phosphate–Guanine

Crkl – C1T10 regulator of kinase like

Ct – Cycle Threshold

CTL – Cytotoxic T-Lymphocyte

DD – Dimerization Domain

DEPC – Diethylpyrocarbonate

DMSO – Dimethyl Sulphoxide

DNA – Deoxyribonucleic Acid

dNTP - Deoxyribonucleotide

EDTA – Ethylenediaminetetraacetic Acid

EFS – Event-Free Survival

ELN – European Leukemia-Net

EMR – Early Molecular Response (3 month BCR-ABL1 mRNA levels < 10%)

ERK – Extracellular Signal-regulated Kinase

EUTOS – European Treatment and Outcome Study

FACS – Fluorescence Activated Cell Sorting

FC – Fold Change

FCS – Foetal Calf Serum

FDA – Food and Drug Administration, United States of America

FDR – False Discovery Rate

g – also known as rcf (Relative Centrifugal Force)

GDP – Guanidine Diphosphate

GSEA – Gene-Set Enrichment Analysis

GTP – Guanidine Triphosphate

h – Hour/s

HBSS – Hanks Balanced Salt Solution

HLA – Human Leukocyte Antigen

HPC – Haematopoietic Progenitor Cell

HSC/s – Haematopoietic Stem Cell/s

HTqPCR – High-Throughput Quantitative PCR

IC50 – 50% Inhibitory Concentration

IFN-α – Interferon-α

IUR – Intracellular Uptake and Retention

kD – Kilo Dalton

KD – Kinase Domain

L – Litre/s

LIMMA – Linear Models for Microarray Data

LSC/s – Leukaemic Stem Cell/s

M – Molar

M-bcr – Major Breakpoint Cluster Region

m-bcr – Minor Breakpoint Cluster Region

MAP – Mitogen-Activated Protein

MCyR – Major Cytogenetic Response (> 35% Ph+ metaphases)

MDR – Multi-Drug Resistance

MDS – Multi-Dimensional Scaling

MeV – Multi-Experiment Viewer

MFI – Mean Fluorescence Intensity

mg – milligram/s

min – Minutes/s

mL – Millilitre/s

mM – Millimolar

MMR – Major Molecular Response (BCR-ABL1 mRNA levels < 0.1%)

MNC/s – Mononuclear Cell/s

MQ – Milli-Q

mRNA – Messenger RNA

MSigDB – Molecular Signatures Database

MW – Molecular Weight

ND – Not Determined

ng – Nanogram/s

NLS – Nuclear Localisation Signal

nM – Nanomolar

OA – OCT-1 Activity

OCT-1 – Organic Cation Transporter 1

OS – Overall Survival

p- – Phosphorylated Form of Protein

P-loop – Nucleotide Binding Loop

P-value – Probability Value

PAM – Prediction Analysis for Microarrays

PB – Peripheral Blood

PBS – Phosphate Buffered Saline

PFS – Progression-Free Survival

Ph – Philadelphia Chromosome

PI-3K – Phosphoinositide 3-Kinase

PMNC/s – Polymorphonuclear Cell/s

PVDF – Polyvinylidene Difluoride

QC – Quality Control

RIPA – Radioimmunoprecipitation Buffer

RMA – Robust Multi-Array Averaging

RNA – Ribonucleic Acid

RO – Reverse Osmosis

ROC – Receiver Operating Characteristic

RPMI – Roswell Park Memorial Institute (media)

RQ-PCR – Quantitative Reverse Transcription-Polymerase Chain Reaction

SD – Standard Deviation

SDS-PAGE – Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis

sec – second/s

SEM – Standard Error of the Mean

SH – *Src*-Homology Region

SNP/s – Single Nucleotide Polymorphism/s

STAT – Signal Transducer and Activation of Transcription

SWAN – Subset-Quantile Within Array Normalisation

TBS – Tris Buffered Saline

TBST – Tris Buffered Saline +Tween®20

TF – Transcription factor

TFS – Transformation-Free Survival

TGA – Therapeutic Goods Administration, Australia

TKI/s – Tyrosine Kinase Inhibitor/s

TLDA – TaqMan® Low Density Array

TMDs – Transmembrane Domains

TSS – Transcription Start Site

TWC/s – Total White Cell/s

U/mL – Units Per Millilitre

UTR – Untranslated Region

WBC/s – White Blood Cell/s

WCC – White Cell Count

WCF – White Cell Fluid

CLINICAL TRIALS REFERRED TO IN THIS THESIS

ALLG TIDEL II (CML9)

A Phase II study in adult patients with newly-diagnosed chronic phase, chronic myeloid leukaemia of initial intensified imatinib therapy, and sequential dose-escalation followed by treatment with nilotinib in suboptimal responders to determine the rate and duration of major molecular response.

Official Title: Australasian Leukaemia and Lymphoma Group (ALLG) Therapeutic Intensification in DE-novo Leukaemia (TIDEL) II trial

Trial ID: ACTRN12607000325404 (http://www.ANZCTR.org.au)

ENESTxtnd

A Phase III study in adult patients to further investigate the safety and efficacy of nilotinib in newly diagnosed chronic myeloid leukaemia patients in the chronic phase.

Official Title: Extending Molecular Responses With Nilotinib in Newly Diagnosed Chronic Myeloid Leukemia (CML) Patients in Chronic Phase (ENESTxtnd) trial

Trial ID: NCT01254188 (http://www.clinicaltrials.gov)

IRIS

A prospective, multi-centre, open-label, randomised Phase III study in adult patients with newly diagnosed chronic-phase chronic myeloid leukaemia to compare the effectiveness of imatinib (STI571) with that of Interferon-α plus low-dose Cytarabine (Ara-C).

Official Title: International Randomised Study of Interferon and STI571 (IRIS) trial

Trial ID: NCT00006343 (http://www.clinicaltrials.gov)

ACKNOWLEDGEMENTS

Looking back, I find it hard to actually remember a time during my PhD candidature where I found myself seriously asking “why did I do this?” and for anyone who has undertaken the extremely stimulating, frustrating, rewarding, and tedious challenge of a PhD, they would know that to do so is by no means an easy feat. I believe this reflects the excellent support and guidance I have had during these past four years.

Firstly, I must thank and acknowledge my three supervisors, Prof. Deb White, Prof. Tim Hughes and Prof. Richard D’Andrea, for without whom, completing this monumental challenge would not have been possible. Deb, your support, guidance, encouragement and tireless efforts have not only helped me immensely during my PhD, but have also provided a wonderful example of the dedication required to succeed in research, I cannot thank you enough. Tim, your CML knowledge is second to none and having a clinician’s perspective when discussing my data and project direction has been invaluable, thank you. D’A, your unique perspective and ideas, while sometimes seemed perplexing, were excellent at getting me to think “outside the box” and led to my first first-author publication, which I will always remember and am extremely grateful for.

A huge thank you must also go to all the past and present members of the Melissa White Lab, who provided me with an amazing work environment, often filled with laughter as well as excellent technical assistance. Special thanks to Chung for all his ideas, suggestions and help throughout my PhD, as well as for to turning me from a pure biologist to a biologist with bioinformatics skills (I know I still have a long way to go!); to Tamara for reviewing all my thesis chapters (by no means an easy task!); to Verity, Amity, Jarrad and Jenny for performing experiments (especially all the OA assays) and processing the patient samples; to Stephanie and Bron for providing me with patient data and help when sending samples interstate to the AGRF; to Phuong for her calm PCR teaching and assistance, as well as always being ready for a chat to help with any problem I may have encountered; to Sue for her help with reviewing my discussion chapter; to Craig, who may have only been in the lab for a short period, but certainly made the days writing my thesis at my desk more enjoyable with some witty banter and who can probably, in conjunction with Jarrad, be credited with introducing me to coffee (although I still don’t see what all the fuss is about!).

A special thank you must go to two people, who are not only colleagues, but awesome friends and have certainly filled my PhD journey with fun and frivolity, Lisa and Laura. You both made sure to remind me if my ego was becoming inflated, and took a little too much delight in scaring/scheming against me, but I feel truly lucky to call you both friends. Lisa, although at times we were pitted against each other in combat (i.e. Uni poster prizes), it has been truly amazing having somebody to both celebrate with and complain to, who is experiencing the exact same thing as me. Laura, I still clearly remember the day when you were “too cool” to hang out with the young kids (i.e. me) at your parent’s house, but I guess I have to thank you for finding it within yourself to talk to me when I first came to the lab. Your sense of humour, general lack of knowing when you had “crossed the line” with a joke/comment and outlandish behaviour (especially after Friday night/conference drinks), made coming into work so much more enjoyable. I hate to admit it, but the lab/student room was a quiet place when you were away on holidays!

I must also thank the Leukaemia Foundation of Australia, for their generous scholarship which funded my PhD, and without which much of the research and conferences I attended would not have been possible. I am also indebted to my amazing group of friends, some of which I’m sure still don’t understand what it is that I’m doing, or how hard it’s been, but have been there nonetheless to celebrate the highs, and to take my mind off the lows. A special mention must go to my once colleague and best mate, Matt, who has endured a host of whinging and complaining whenever I was stressed, but still rocked up each morning or night to pump out that next gym session.

Heart-felt thanks must also go to my Mum (Carol), Dad (David), sister (Lauren) and all my extended family, who always believed in me, supported me and have shared this journey (both the good and bad) with me. The person I am today is a reflection of the love, support and drive that you have imparted on me.

Lastly, my deepest thanks go to my partner Felicity Sprod, who has been there unwavering throughout this whole journey. Your love, affection and endless support have helped me through the tough periods and always bring a smile to my face. To your loving family (Karyn and Derek, Rachel and Ryan, Belinda and Brad), who have both welcomed and supported me, I thank you all for your care and encouragement throughout this journey.

1 INTRODUCTION

1.1 Chronic Myeloid Leukaemia

1.1.1 Clinical Features

Chronic Myeloid Leukaemia (CML) is a clonal malignancy characterised by an over production of white blood cells, primarily originating from the myeloid lineage in the bone marrow of an affected individual. The first case report was published in 1845,[1] but it was not until over one hundred years later that the first major breakthrough in CML disease biology transpired. In 1960, Philadelphian researchers Peter Nowell and David Hungerford discovered an abnormally small chromosome present in blood cells from CML patients; but absent in normal individuals. This minute chromosome was designated the 'Philadelphia' chromosome,[2] and is the genetic hallmark of CML. Affecting people from all age groups, CML has a worldwide incidence of approximately 1-2 per 100 000 population, encompassing approximately 15–20% of all leukaemias.[3]

CML is characterised by a triphasic disease course, initiating in a chronic phase (CP), before progressing through an accelerated phase (AP), to the terminal blast phase or blast crisis (BC; **Figure 1.1**). Both the AP and BC can be grouped together when a patient is considered to be in advanced-phase CML.[4] Without therapy, the duration of the CP can vary between 2 and 5 years; the AP generally lasts up to a year; and the fatal BC lasts between 3 and 6 months.[4] Patients can present with CML in any of these 3 stages, but presentation in BC is quite uncommon. Significantly, in CP, the disease is fairly indolent and characterised by an expansion of the granulocyte cell population, the majority of which are of the myelocyte and neutrophil subtypes[5], however, patients will often exhibit no discernible symptoms. When symptoms are present, they are often non-specific and include signs such as weakness, weight loss, fatigue and fever.[6] Nevertheless, most patients are diagnosed in CP, often from

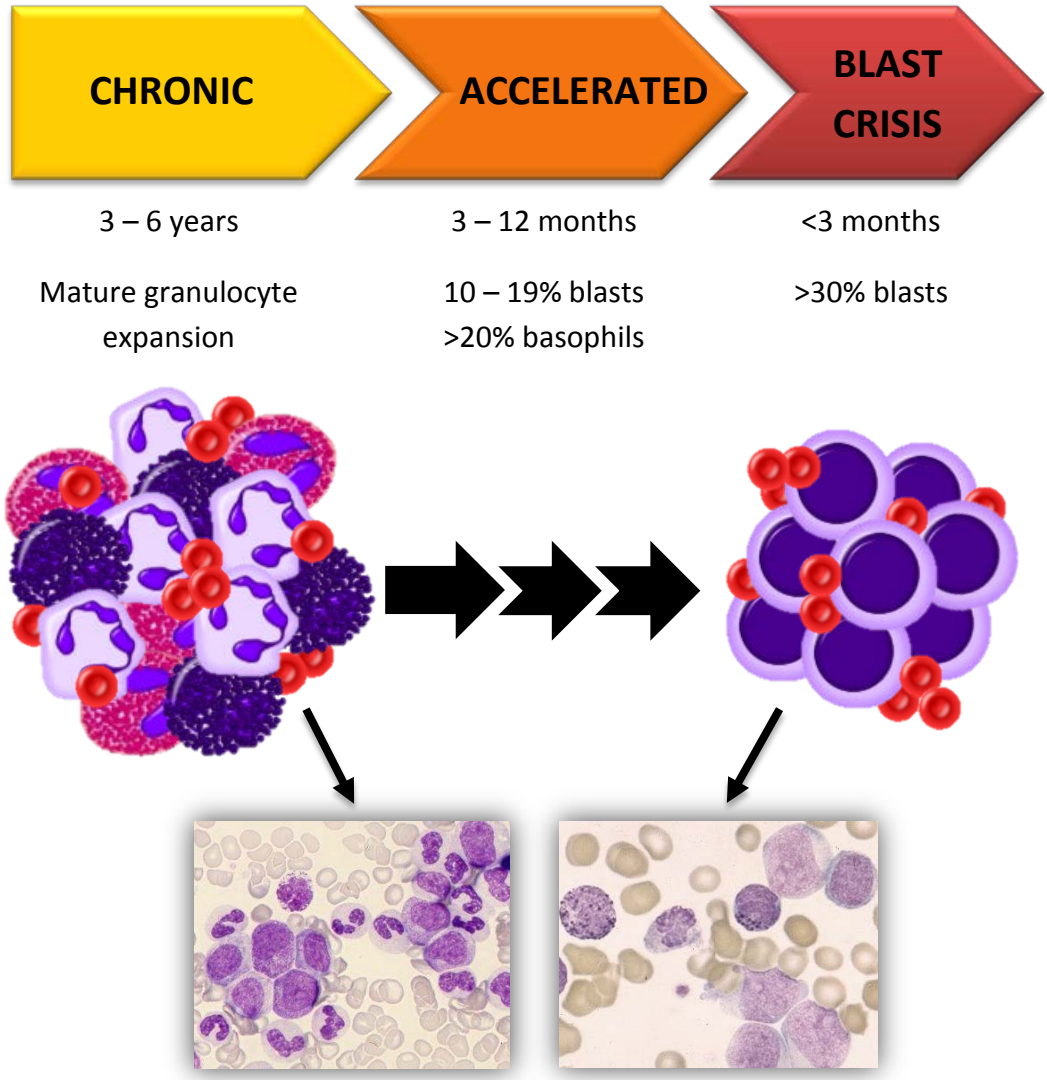


Figure 1.1 Disease progression in CML

Chronic phase CML is characterised by an expansion of the granulocytic cell lineage, consisting mainly of the myelocyte and neutrophil subtypes. Accelerated phase is distinguished by a more aggressive phenotype, including an increase in undifferentiated precursor cells. Fatal blast crisis is characterised by the rapid expansion of an immature blast cell population, resembling an acute leukaemia.

routine blood screens, which incidentally lead to the CML diagnosis. Disease progression can occur rapidly if not treated or detected early, with both the AP and BC characterised by aggressive phenotypes, including the appearance of blast cells and additional chromosomal abnormalities (clonal evolution).[7-10] The prognosis for a patient with CML is dependent upon their stage of disease at diagnosis, with patients diagnosed in CP having prolonged survival.[11] However, significant variability in survival is observed even for patients diagnosed in CP,[12] possibly indicating a subset of patients with more advanced disease that is not discernible by current methods.

1.1.2 Philadelphia Chromosome and Resulting **BCR-ABL1** *Fusion Gene*

The Philadelphia (Ph) chromosome arises from a balanced reciprocal translocation t(9;22)(q34;q11) between chromosome 9 and chromosome 22 (**Figure 1.2**).[13] This results in the formation of a unique molecular structure (a fusion gene) between the *ABL1* (Abelson leukaemia virus proto-oncogene) gene on chromosome 9[14] and the *BCR* (Breakpoint cluster region) gene on chromosome 22.[15,16] The consequence of this translocation is the production of the Bcr-Abl1 oncoprotein, a fusion protein with unregulated (constitutively active) tyrosine kinase activity.[4,17,18] While breakpoints within *ABL1* and *BCR* can occur in numerous locations specific to each gene, the majority are restricted to three regions at the 5' end of *ABL1*; either upstream of exon 1b, downstream of exon 1a, or, more frequently, between the two (**Figure 1.3**).[19] The breakpoints within *BCR* are conserved to three different breakpoint cluster regions. However, the majority of *BCR* breakpoints occur in a 5.8kb area spanning exons 12-16 (mainly at exons e13 and e14; **Figure 1.3**), known as the major breakpoint cluster region (M-*bcr*).[20]

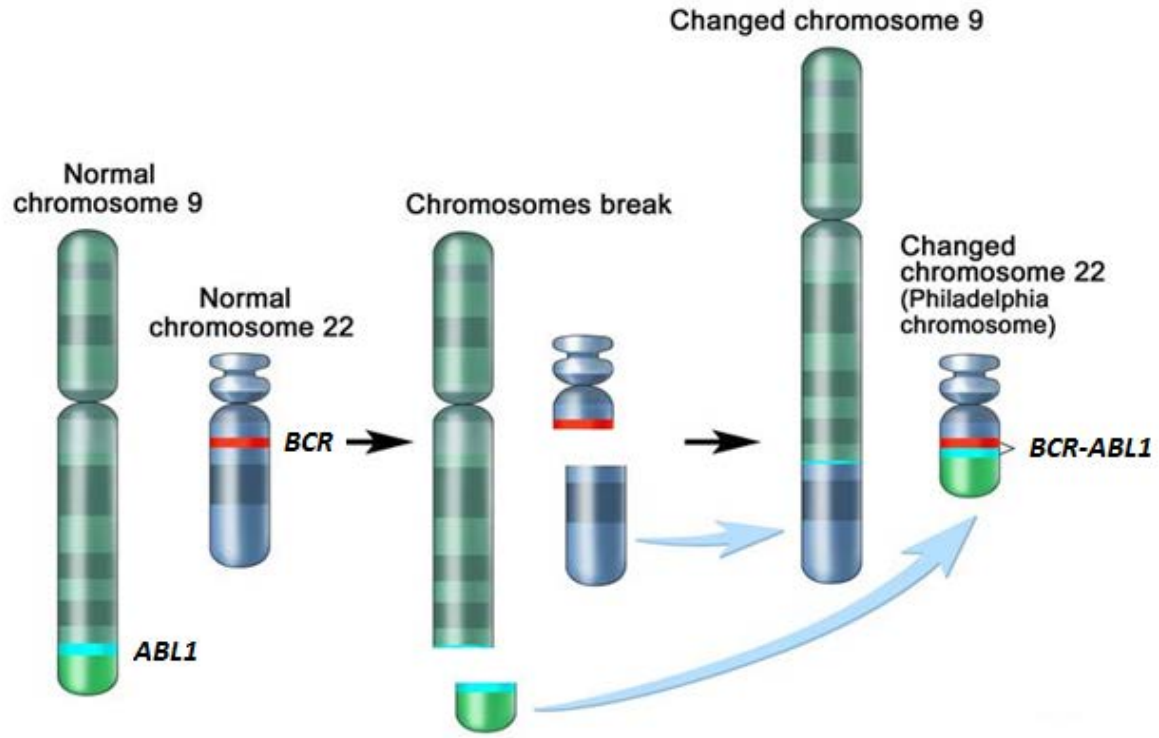


Figure 1.2 Formation of the Philadelphia Chromosome

The Philadelphia (Ph) chromosome arises from the reciprocal translocation of the long arms of chromosomes 9 and 22.[13] The *BCR-ABL1* gene is formed on chromosome 22 where the piece of chromosome 9 attaches. Figure adapted from National Cancer Institute (http://www.cancer.gov/dictionary?cdrid=44179, accessed 28th May 2013).[21]

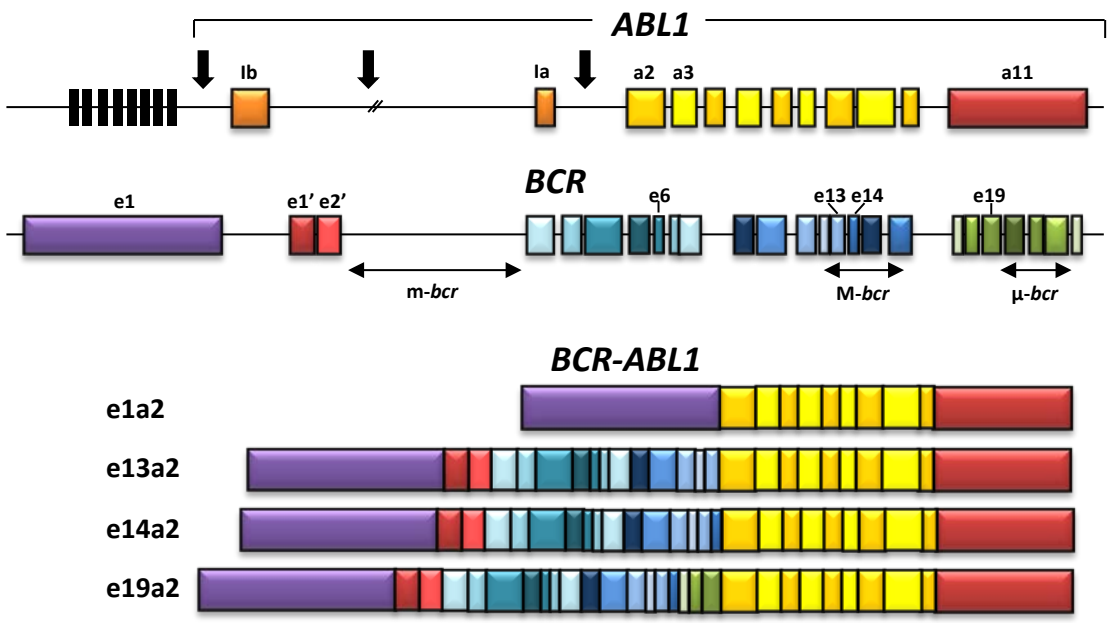


Figure 1.3 Schematic representation of the *BCR-ABL1* oncogene

Exons are represented by coloured boxes with introns depicted as horizontal lines. The three main breakpoint regions in the *ABL1* gene are shown by the solid black arrows. Although there are several points at which breakage can occur, processing of the transcript results in messenger RNA (mRNA) of the a2 portion only. The breakpoints in *BCR* occur in three breakpoint cluster regions: m-*bcr*, M-*bcr* and μ-*bcr*. The length of the Bcr-Abl1 fusion protein formed is dependent on where the breakpoints occur. Breaks within m-*bcr* result in fusion proteins with an e1a2 junction. Breaks within M-*bcr* result in either e13a2 or e14a2 (formally b2a2 and b3a2, respectively) junctions while breaks in μ-*bcr* form fusion proteins with e19a2 junctions. Figure adapted from Melo *et al*., 1996.[19]

Alternative splicing allows fusion transcripts with either e13a2 (b2a2) or e14a2 (b3a2) junctions to form, resulting in mRNA that encodes a 210-kd chimeric protein ($p210^{BCR-ABL1}$; **Figure 1.4**).[20,22] Breakpoints can also occur further upstream, in the minor breakpoint cluster region (m-*bcr*), located between *BCR* exons e2' and e2.[20] This is rare in CML, but occurs with greater frequency in patients with Ph-positive Acute Lymphoblastic Leukaemia (Ph+ ALL). This e1a2 mRNA encodes a 190-kd protein ($p190^{BCR-ABL1}$; **Figure 1.4**), which is important to distinguish when determining the disease phenotype.[19] A third breakpoint cluster region, (known as µ-*bcr*[23]), has been rarely identified downstream of M-bcr and m-bcr, between exons e19 and e20. The mRNA of which encodes a 230-kd fusion protein ($p230^{BCR-ABL1}$; **Figure 1.4**), usually associated with rare Ph-positive Chronic Neutrophilic Leukaemia.[23] It is interesting to note that the reciprocal *ABL1-BCR* gene (chromosome 9) is transcriptionally active in approximately two thirds of Ph-positive CML patients.[22]

1.1.3 Disease Biology and Pathogenesis

CML is specifically a myeloproliferative disorder, involving a clonal expansion of haematopoietic stem cells transformed with *BCR-ABL1*. These functional leukaemic cells proliferate in the bone marrow, leading to abnormally high white blood cell (WBC) counts, primarily neutrophils and myelocytes.[4] Usually, haemoglobin levels fall moderately, while the platelet count increases moderately, but can remain within normal parameters.[24] This initial expansion of the myeloid compartment is not only attributable to increased cell proliferation, but also to enhanced cell survival, additional cell divisions occurring during myeloid maturation, and to the ability of CML clones to grow in higher densities than normal.[25] The *BCR-ABL1* gene was unequivocally demonstrated to be the causative agent of CML in 1990 when Daley *et al.*[26] transfected mouse bone marrow cells with *BCR-ABL1*. These cells were then transplanted into a second group of irradiated mice who subsequently

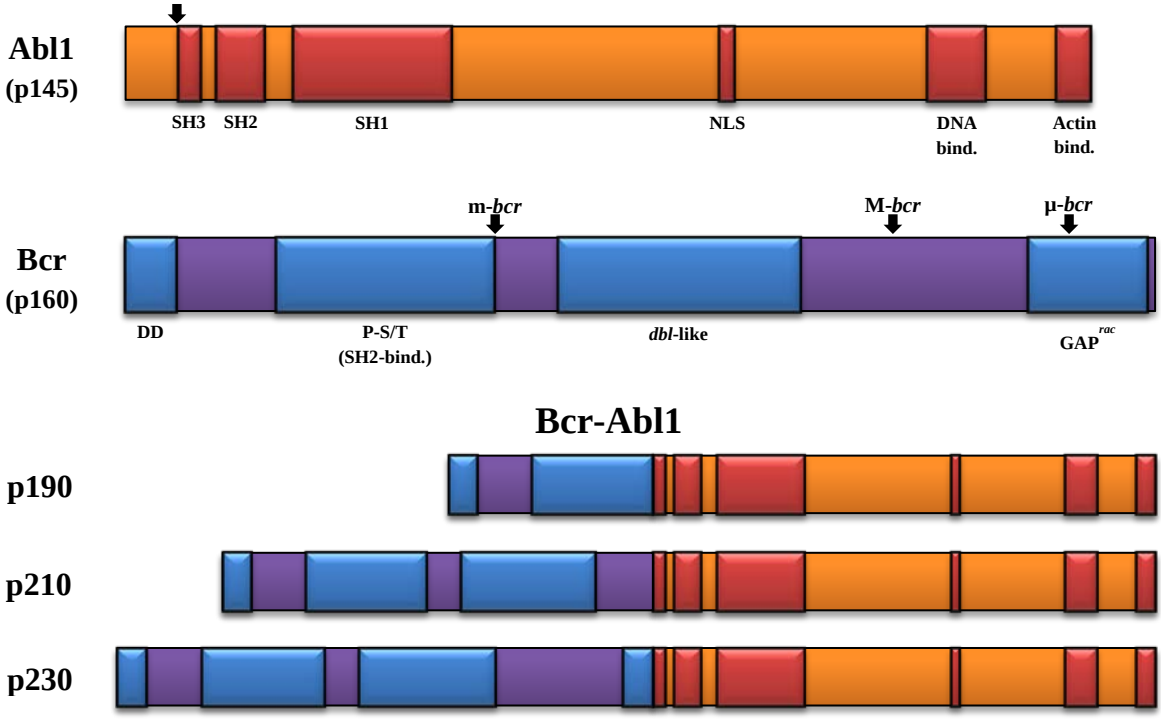


Figure 1.4 Schematic representation of the normal Abl1 and Bcr proteins, together with the three major isoforms of the Bcr-Abl1 oncoprotein

The *BCR-ABL1* oncogene results from a reciprocal translocation between the *BCR* and *ABL1* genes. The arrows indicate the three main breakpoints for Bcr, where differential splicing events give rise to fusion proteins of differing length, and intrinsic tyrosine kinase activity: m-*bcr* (p190$^{BCR-ABL1}$), M-*bcr* (p210$^{BCR-ABL1}$) and µ-*bcr* (p230$^{BCR-ABL1}$).[27] Both the Abl1 and Bcr proteins contain specific regions that have important roles when the Bcr-Abl1 fusion protein is formed. In Abl1, these include the regulatory *src*-homology (SH) regions SH3 and SH2, the SH1 kinase domain, the nuclear localization signal (NLS) domain, and the DNA- and actin-binding domains.[28,29] In Bcr, these include the dimerization domain (DD), the phospho-serine/threonine (P-S/T)-rich SH2-binding domain, and the *dbl*-like and GAP^{rac} domains.[30-32] Figure adapted from Melo, J.V., 1996.[19]

developed a disease closely resembling CML, providing definitive evidence that *BCR-ABL1* was the causative agent in CML.

Significant homology exists between the Abelson murine leukaemia virus (Ab-MuLV) oncogene, *v-abl*, and its human cellular homologue, c-Abl (encoded by *ABL1*),[33,34] which fuses with the promoter and N-terminal portion of Bcr in CML. As the *BCR-ABL1* translocation occurs, *ABL1*, which is normally present on chromosome 9, is moved to chromosome 22. The *ABL1* gene encodes a 145-kd non-receptor tyrosine kinase which is ubiquitously expressed (**Figure 1.4**).[28] The c-Abl proto-oncoprotein resides in both the nucleus and the cytoplasm of a cell.[35,36] Normally, the c-Abl protein is involved in the regulation of the cell cycle,[37] cellular response to genotoxic stress,[38] and transmission of information about the cellular environment through integrin signalling.[39] As such, the c-Abl protein has a significant and complex role, integrating extracellular and intracellular signals from various sources, thus, influencing cellular apoptosis, cell cycle decisions and ultimately cell fate.

The *BCR* gene encodes a 160-kd cytoplasmic protein that is ubiquitously expressed (**Figure 1.4**).[28] The function of this protein is not well understood, however, it does contain a number of distinct structural domains. These include a novel N-terminal serine/threonine kinase in the first exon (P-S/T domain; **Figure 1.4**),[31] a N-terminal coiled-coil domain allowing dimer formation *in vivo* (DD; **Figure 1.4**),[30] a central region that stimulates guanidine triphosphate (GTP) exchange for guanidine diphosphate (GDP) on Rho guanine-nucleotide exchange factors (*dbl*-like domain; **Figure 1.4**),[40] and C-terminal sequences that encode GTPase activity for Rac, (GAP^{rac} domain; **Figure 1.4**),[40,41] a small GTPase of the RAS

superfamily. Bcr can also be phosphorylated on multiple tyrosine residues, especially Tyr177, which binds Grb-2, an adapter molecule associated with activation of the RAS pathway.[42]

Intriguingly, despite the presence of the NLS, the Bcr-Abl1 oncoprotein may be unable to migrate between the nucleus and cytoplasm, unlike normal c-Abl, and is predominantly located in the cytoplasm,[43] where it is able to disrupt multiple signalling pathways involved in cellular homeostasis. Once the *BCR-ABL1* oncogene formation was elucidated as the causative event in CML, researchers sought to understand its molecular mechanism of action by identifying the signalling pathways associated with Bcr-Abl1 tyrosine kinase activity. It is now known that the increased tyrosine kinase activity of Bcr-Abl1 causes unrestricted phosphorylation of tyrosine residues on proteins involved in signal transduction pathways, including those controlling cell proliferation, survival, adhesion, growth factor independence and apoptosis.[3,44]

Six "hallmarks" (common traits) of cancer were originally proposed by Hanahan and Weinberg,[45] encompassing sustaining proliferative signalling; evading growth suppressors; resisting cell death; enabling replicative immortality; inducing angiogenesis; and activating invasion and metastasis. Recently, these common traits were reviewed to include the conceptual developments made in cancer research since the original review, identifying two new enabling characteristics, pivotal to the acquisition of the hallmark capabilities, and two new emerging hallmarks.[46] These enabling characteristics included genome stability and mutation, and tumour-promoting inflammation, while the emerging traits included deregulating cellular energetics and avoiding immune destruction. These hallmarks all have significant roles in CML. Specifically, the key signal transduction pathways associated with CML (**Figure 1.5**) include the Ras and mitogen-activated protein (MAP) kinase pathways. The

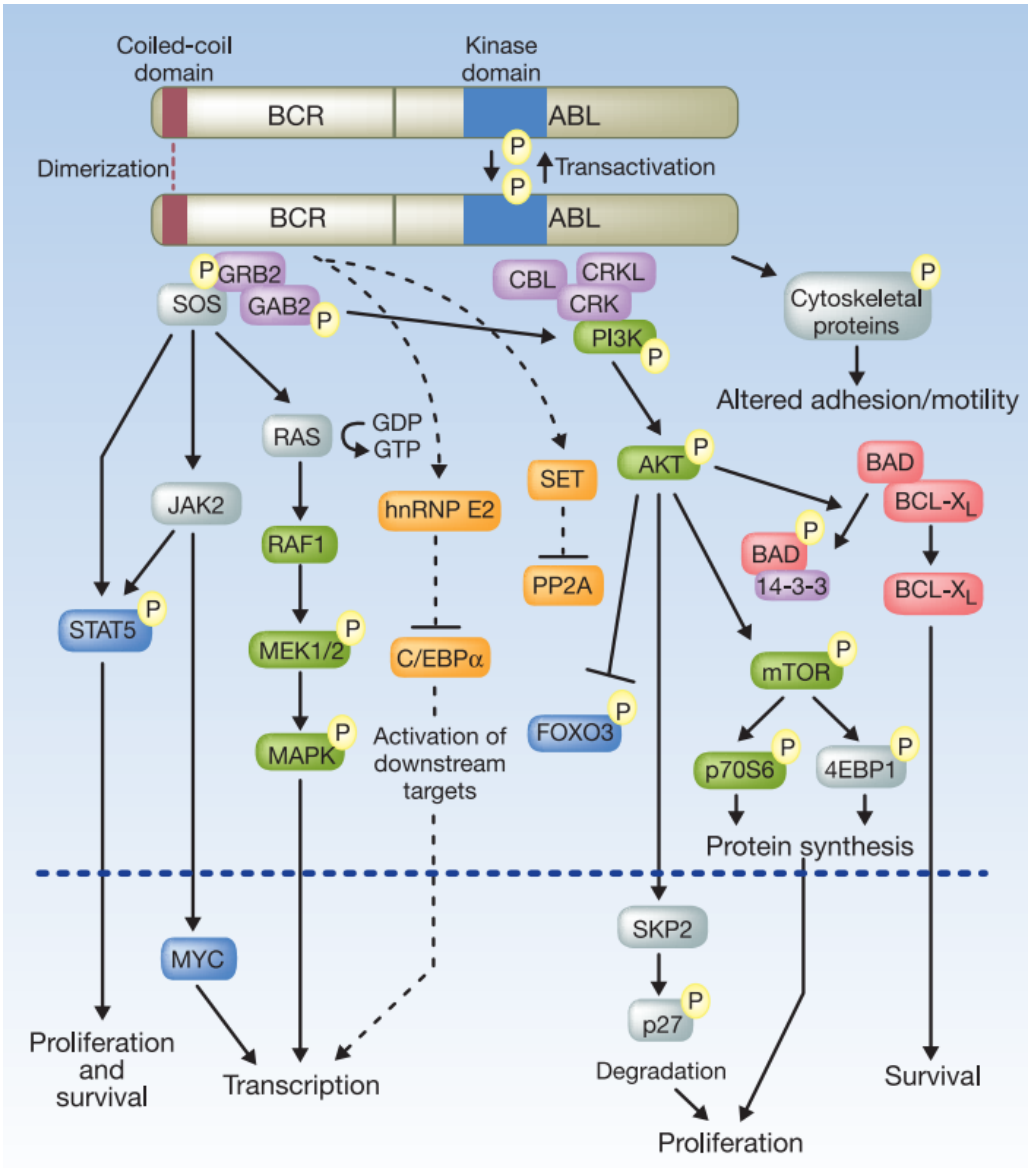


Figure 1.5 Schematic of the signalling pathways activated by Bcr-Abl1 in CML

Bcr-Abl1 dimerization triggers autophosphorylation events that activate the kinase and generate docking sites for intermediary adapter proteins (**purple**) such as GRB2. Activation of multiple downstream pathways is facilitated by Bcr-Abl1–dependent signalling including enhanced survival, inhibition of apoptosis, and perturbation of cell adhesion and migration. A subset of these pathways and their constituent transcription factors (**blue**), serine/threonine-specific kinases (**green**), and apoptosis-related proteins (**red**) are shown. Furthermore, a few pathways more recently implicated in CML stem cell maintenance and BCR-ABL1–mediated disease transformation are shown (**orange**). Figure from O'Hare *et al.*, 2011.[47]

Ras oncoprotein normally operates as an intrinsic negative-feedback mechanism, controlling active signal transmission; however, when Ras GTPase activity is compromised, enhanced proliferative signalling occurs.[46] Bcr-Abl1 is able to bind the SH2 domain of the adapter protein GRB2, which subsequently recruits Son of Sevenless (SOS), and this complex is able to stimulate the conversion of inactive Ras (GDP-bound) to the active (GTP-bound) state (**Figure 1.5**).[48] Additionally, this leads to the activation of the GRB2-associated binding protein 2 (GAB2), causing constitutive activation of downstream Ras pathways, including activation of the MAP extracellular signal-regulated kinase (ERK) 1/2 (MEK) and MAP kinase proteins (**Figure 1.5**), leading to abnormal cell proliferation,[48,49] and contributing to oncogenic transformation.[50]

Bcr-Abl1 also activates the Jak-STAT signalling pathway via constitutive phosphorylation of signal transducer and activation of transcription (STAT) transcription factors (such as STAT1 and STAT5).[51-53] The STAT5 transcription factor normally modulates transcription of multiple targets including cyclin D1, involved in cell cycle progression from G_1 to S phase.[54] In *BCR-ABL1* transformed cells, activation of STAT5 signalling is partially responsible for an anti-apoptotic effect,[55] as it up-regulates the anti-apoptotic molecule BCL-xL, in combination with the inactivation of the pro-apoptotic molecule BAD by AKT, contributing to malignant transformation.[48,56] The phosphoinositide 3-kinase (PI-3K)/AKT pathway is also activated by Bcr-Abl1 and the resultant GRB2/GAB2/SOS signalling cascade (**Figure 1.5**), leading to enhancement of cell proliferation via induction of p27 proteasomal degradation and mTOR activation.[48] Additionally, PI-3K activation leads to the activation of AKT, resulting in the phosphorylation and activation of other downstream cellular proteins (e.g. Bad; **Figure 1.5**),[48] contributing further to the anti-apoptotic affect.[52] Furthermore, PI-3K activity has been demonstrated to be essential for the proliferation of *BCR-ABL1*-positive cells (**Figure**

1.5).[57] Finally, it is important to note that Bcr-Abl1 does not only affect cell signalling pathways, but it also affects cellular adhesion properties,[58] degrades inhibitory proteins such as Abi, a family of Abl-interacting proteins that antagonize the oncogenic potential of Abl,[59] and may aid survival of leukaemic cells by influencing cell cycle and anti-apoptotic regulation.[3]

1.2 Historical Treatment Options for CML Patients

Attempts at designing therapeutic tools for CML based on the current knowledge of the molecular and cell biology of the disease have concentrated on 3 areas – the inhibition of gene expression at the translational level by "antisense" strategies, the stimulation of the immune system's capacity to recognize and destroy leukaemic cells, but mainly the modulation of protein function by specific signal transduction inhibitors.[20] However, before these molecular therapeutic strategies were employed, conventional chemotherapy was used to treat CML.

1.2.1 Conventional and Intensive Chemotherapy

The two most commonly used chemotherapy agents for CML were busulfan and hydroxyurea.[60] Essentially, both treatments have identical efficacy and median duration of response.[60,61] However, hydroxyurea became increasingly popular due to its rapid action and lower level of adverse effects, when compared to busulfan.[61] Hydroxyurea is an inhibitor of the ribonucleotide reductase,[61] and requires daily administration to adequately control the leukocyte count.[60] Significantly, both drugs are able to reduce the cellular burden associated with CML at diagnosis, but do not specifically target or eliminate the leukaemic cells. Hence, patients continue to progress through the various stages of CML, and eventually die

prematurely from the disease. New treatments that would induce remission were subsequently investigated, leading to the use of interferon-α (IFN-α).

1.2.2 Interferon-α

The interferons are a group of proteins with antiproliferative, antiviral and immunomodulatory effects.[62] Multiple clinical studies involving the use of partially pure and synthetic leukocyte interferon had demonstrated an anti-tumour effect against several cancers including breast cancer, multiple myeloma, malignant lymphoma and acute lymphatic leukaemia,[62] hence, the application of IFN-α to CML was a logical progression. IFN-α was the first drug used for the treatment of CML that was able to consistently induce cytogenetic responses in a significant number of patients,[63] with up to 20-25% of patients achieving a complete cytogenetic response (CCyR; absence of Ph-positive cells as measured by classical karyotyping or fluorescence *in-situ* hybridisation).[64] Subsequent studies with IFN-α demonstrated that not only was a cytogenetic response obtained with drug therapy, but that responses were independently associated with improved survival,[64] and importantly, could act as early surrogate markers for long-term survival.[65,66] Consequently, multiple studies comparing IFN-α and conventional chemotherapy were performed.[67-69] However, it soon emerged from these studies that IFN-α therapy had one major drawback – a high toxicity profile, which often lead to psychological changes in the patient, making drug compliance difficult. Therefore, as neither IFN-α, nor conventional chemotherapy were able to induce CCyR consistently in all patients, and IFN-α had undesirable side effects, new molecular therapeutic strategies were pursued.

1.3 Molecular Therapeutic Strategies

The most exciting of the molecularly designed therapeutic approaches was brought about by the advent of signal transduction inhibitors, which inhibit a protein from functioning in the oncogenic pathway.[20] As the Bcr-Abl1 protein induces oncogenesis through its constitutive tyrosine kinase activity, it was logical for researchers to target inhibition of this activity as a means of silencing the oncoprotein. As a result, many tyrosine kinase inhibitors were assessed for their potential application in altering CML cell phenotypes. Natural sources were initially used to identify suitable compounds, from which more potent compounds (including synthetic compounds) were developed. The aim was to design chemical structures that competed with adenosine triphosphate (ATP) or the protein substrate for the binding site in the catalytic centre of the kinase[70] (**Figure 1.6**).

1.3.1 First Generation Tyrosine Kinase Inhibitors

1.3.1.1 Imatinib Mesylate

The 2-phenylaminopyrimidine, STI571 (formerly CGP57418B; Novartis Pharmaceutics, Basel, Switzerland), a synthetically produced compound (chemical structure shown in **Figure 1.7**), specific for the Abl1 tyrosine kinase, exhibited the greatest clinical potential.[71] This compound is now known as imatinib mesylate, and has the trade names Gleevec/Glivec®. The development of imatinib started in the late 1980's when scientists at Ciba Geigy (now Novartis Pharmaceuticals) identified a derivative of 2-phenylaminopyrimidine as a lead compound with inhibitory activity against protein kinases.[72,73] The solubility and oral bioavailability of this compound was then significantly increased by the addition of a highly polar side chain, N-methylpiperazine.[74] Subsequently, imatinib emerged as the compound best suited for clinical development due to its excellent selectivity for growth inhibition of *BCR-ABL1* expressing cells.[74] Imatinib selectively inhibits all Abl1 kinases (c-Abl, Bcr-Abl1, v-

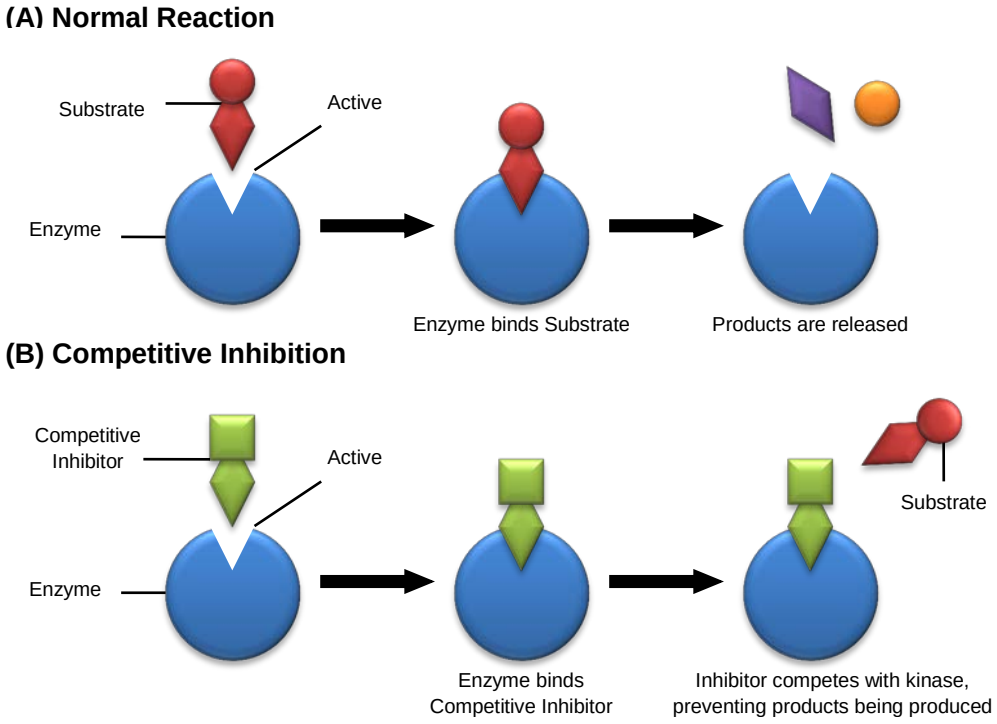


Figure 1.6 Schematic representation of normal enzyme-substrate binding, and competitive inhibitor-enzyme binding

(A) In a normal reaction, the substrate binds to the active domain of the enzyme, is catalysed and products released. **(B)** In a competitive inhibition reaction, an inhibitor binds to the active domain of the enzyme, preventing the substrate from binding. Figure adapted from Mann *et al*., 2005.[75]

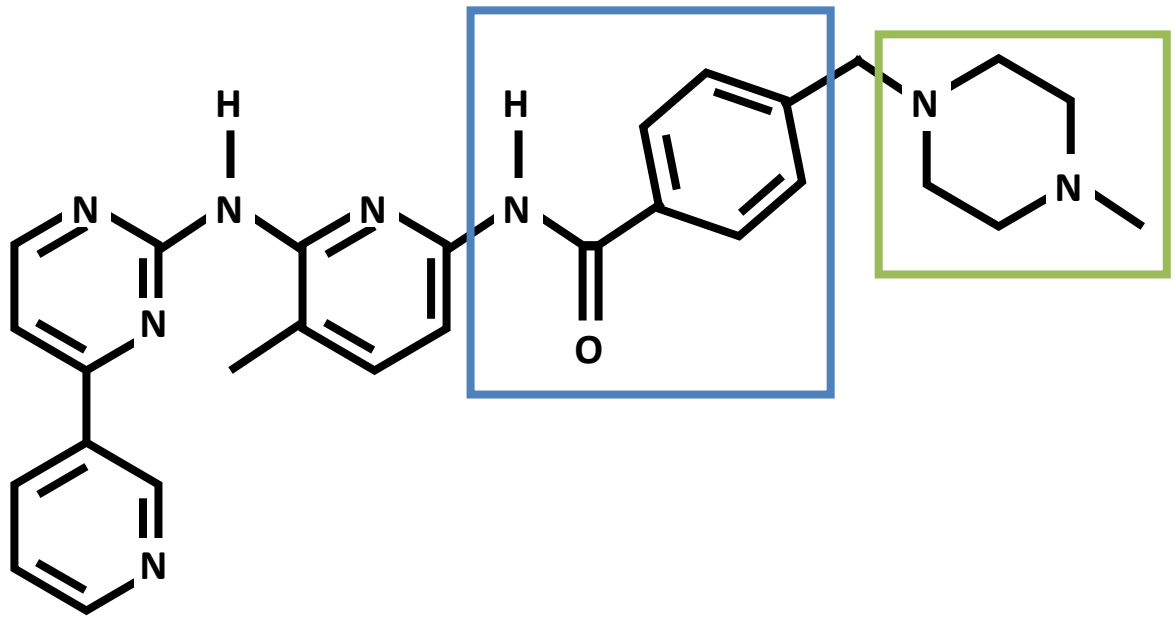


Figure 1.7 The chemical structure of the tyrosine kinase inhibitor, imatinib mesylate

Researchers at Novartis Pharmaceuticals identified this 2-phenylaminopyrimidine molecule exhibiting inhibitory activity against Abl1 tyrosine kinase. The specificity was increased via the addition of a benzamide group (BLUE), and the solubility and bioavailability was improved via the addition of the N-methylpiperazine group (GREEN).[76] The molecular formula is $C_{29}H_{31}N_70 \bullet CH_4SO_3$ with a molecular weight of 589.7g/mol. This compound was later designated imatinib mesylate, with the trade name Glivec/Gleevec®.

Abl)[74] and, in addition, also targets c-fms,[77] c-KIT,[78] and PDGF receptors,[71,74] without targeting other tyrosine kinases such as the Src family kinases.[79] Imatinib competitively binds to the ATP-binding site of the constitutively active tyrosine kinase in Bcr-Abl1 expressing cells, preventing phosphorylation of downstream proteins involved in signal transduction, leading to growth arrest or apoptosis[80] (**Figure 1.8**).

The ability of imatinib to inhibit *BCR-ABL1* positive cell lines *in vitro* was demonstrated in 1996.[74] Subsequently, multiple clinical trials were undertaken to assess the efficacy, potency and safety of imatinib treatment for patients with CML. Dose-escalating Phase I trials,[81,82] in CP-CML patients who had previously failed treatment with IFN-α, demonstrated that complete haematological responses (CHR, sustained and significant reduction in WBCs to a normal range) were obtainable with doses greater than 300mg. Phase II clinical trials[83-85] soon followed, demonstrating beneficial effects from imatinib treatment including higher rates of cytogenetic and haematologic response, lower rates of progression and increased overall survival, for patients in chronic phase,[83,86] advanced phase[85] and in blast crisis.[84] These Phase II trials confirmed the results observed in the Phase I trials, demonstrating that imatinib therapy resulted in superior treatment outcomes in all phases of the disease.

The Phase III IRIS (International Randomised Study of Interferon and STI571) trial compared imatinib with IFN-α and low dose subcutaneous cytarabine (Ara-C), for the treatment of CP-CML in *de novo* patients. Higher rates of CHR (97%) and CCyR (76%) were achieved by patients receiving imatinib than IFN-α plus Ara-C; 69% and 14% respectively (**Figure 1.9**).[86] These higher levels of CCyR in patients on imatinib therapy also correlated with a higher proportion of patients experiencing major molecular response (MMR; BCR-ABL1 mRNA transcript levels < 0.1% by 12 months).[87] The 8-year follow-up data from the IRIS trial

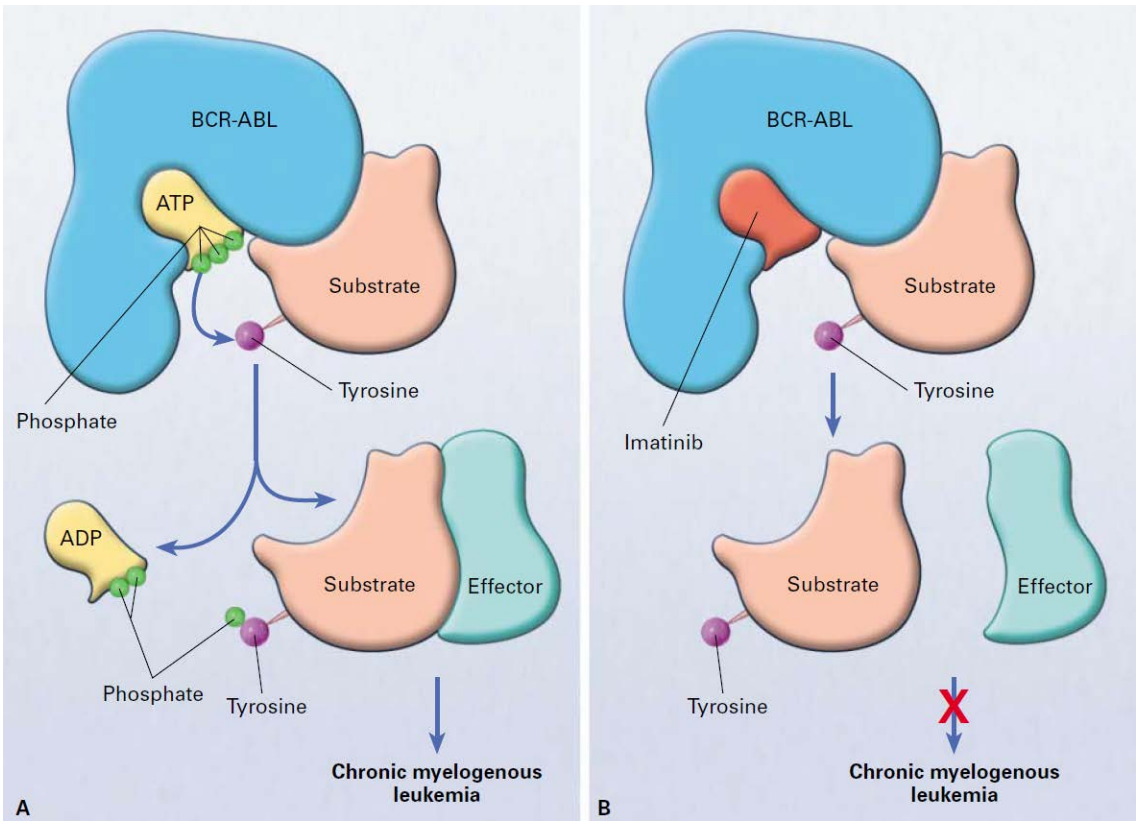


Figure 1.8 Mechanism of action of imatinib binding Bcr-Abl1

(A) In Ph+ cells, the Bcr-Abl1 oncoprotein is constitutively expressed. Adenosine triphosphate (ATP) binds to Bcr-Abl1 in the kinase pocket. The substrate is activated by the phosphorylation of a tyrosine residue, releasing adenosine diphosphate (ADP), and activating the kinase at the same time. This leads to uncontrolled production of intracellular signals, activation of other downstream effector molecules and ultimately CML. **(B)** When imatinib occupies the kinase pocket; ATP is unable to bind Bcr-Abl1, preventing the phosphorylation of the substrate and effectively inhibiting the action of Bcr-Abl1. Figure from Savage, D.G. and Antman, K.H., 2002.[88]

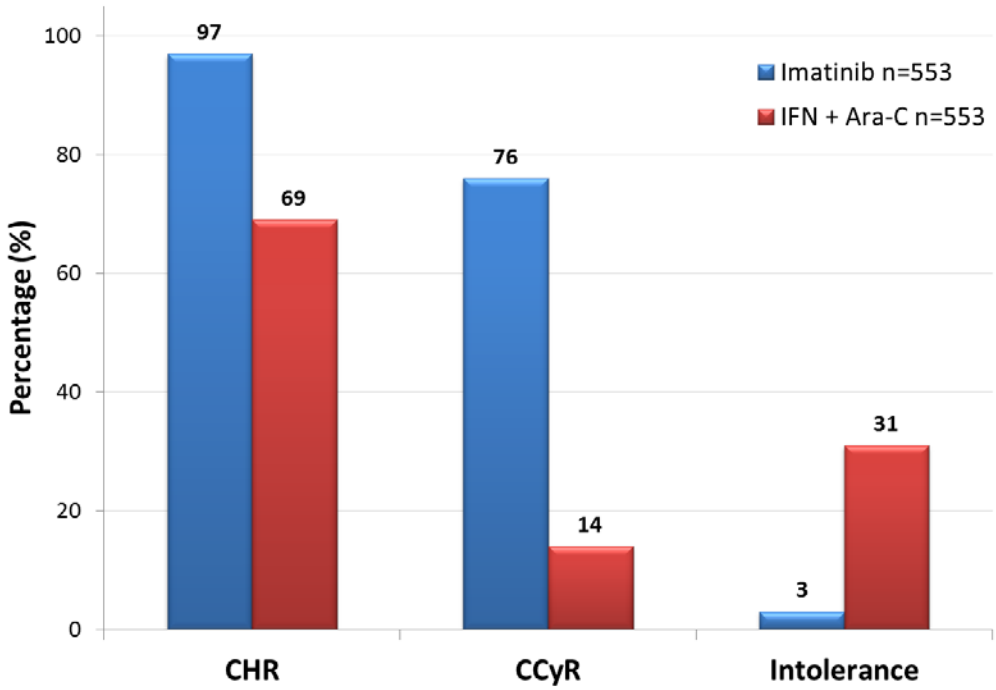


Figure 1.9 Efficacy of imatinib versus interferon-α + cytarabine for the treatment of CP-CML

Percentage of patients achieving complete haematological response (CHR), complete cytogenetic response (CCyR) and patients showing intolerance to imatinib following imatinib therapy for newly diagnosed CP-CML patients with a median follow-up of 18 months. Data demonstrates imatinib treated patients exhibit superior response and decreased intolerance, when compared to interferon-α + cytarabine (IFN + Ara-C) treated patients. Figure adapted from Druker, B.J., 2008.[89]

continued to validate the excellent efficacy and safety profile of imatinib, and demonstrated an MMR rate of 86% for patients on imatinib therapy, with no patient progressing to AP/BC if they achieved MMR by 12 months.[90] Estimated overall survival was 85%; event-free survival was 81% and 92% of patients were progression-free during 8 years of imatinib therapy, suggesting that patients responding to imatinib were likely to maintain their response long-term.

1.3.1.2 Resistance to Imatinib

The major problem associated with imatinib treatment is the risk of the patient demonstrating primary resistance, or developing secondary resistance to imatinib. Failure to achieve a desired response (e.g. CCyR) is known as primary resistance,[91] and this occurs in approximately one third of imatinib-treated patients. It is therefore imperative to understand the causes behind this resistance mechanism, and identify patients at risk prior to therapy initiation. There are two distinct processes required for the achievement of a cytogenetic response. Firstly, successful reduction of leukaemic burden via Bcr-Abl1 inhibition must occur and secondly, an adequate number of normal haematopoietic progenitors in the bone marrow is required to resume normal haematopoiesis.[91] Approximately 30% of newly diagnosed CP-CML patients on the IRIS trial did not achieve a CCyR within 1 year of imatinib therapy (defined as suboptimal response),[86] and by 5 years, an additional 10% of patients had relapsed, including 10% who had previously achieved CCyR.[92] This finding was supported by those of Marin *et al*.[93] who demonstrated that 27% of imatinib-treated patients failed to achieve CCyR by 12 months.

The causes of primary resistance are largely unknown, but are thought to result from deficiencies in several processes, all leading to reduced kinase inhibition (**Figure 1.10**).[94] Alterations to the expression and activity of drug transporters, such as the human organic

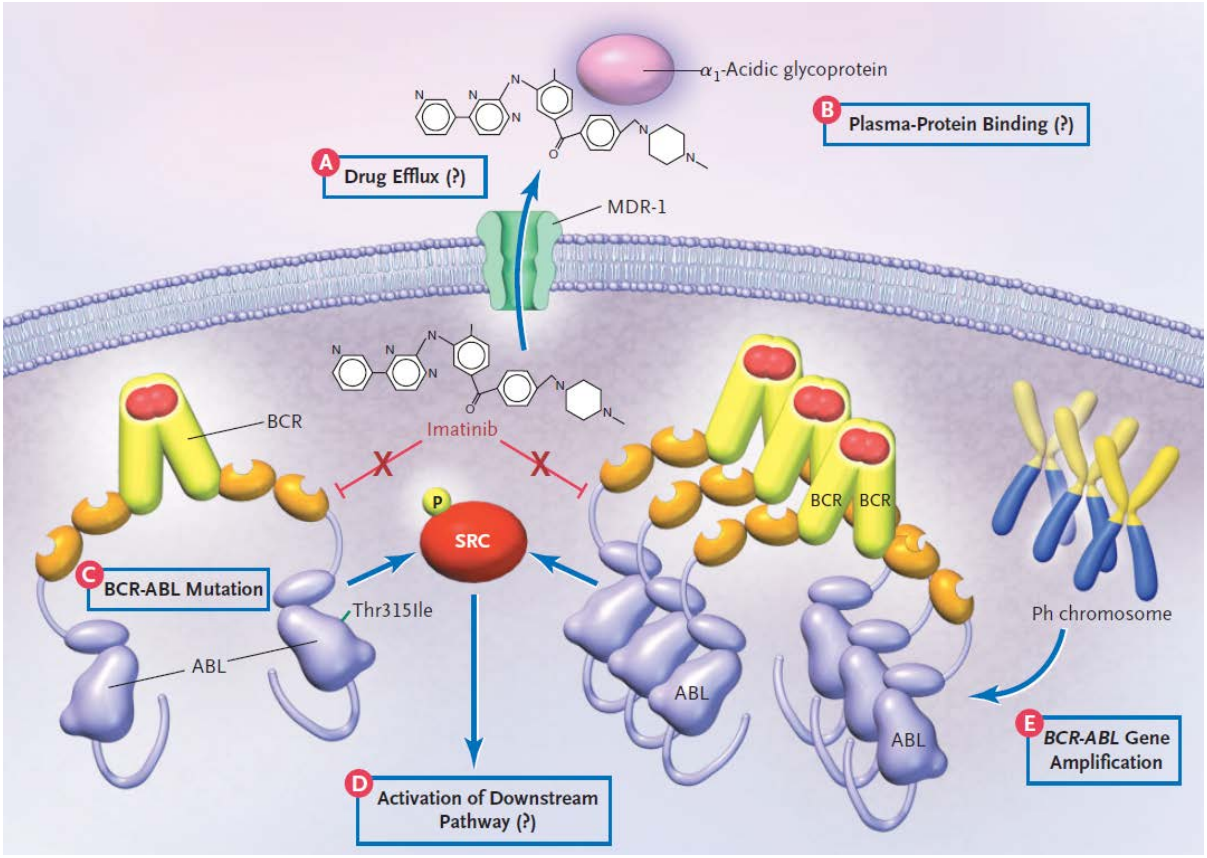


Figure 1.10 Mechanisms of resistance to TKI therapy

Increased efflux of drug from the cancer cell, mediated by membrane transporters such as multidrug-resistance gene 1 protein (MDR-1; ABCB1), can decrease intracellular concentrations of the drug **(mechanism A)**. Increased drug binding by plasma proteins such as α_1-acidic glycoprotein can decrease the drug's effective concentration **(mechanism B)**. Mutations in the Bcr-Abl1 kinase domain, including T315I, can decrease or abolish the inhibitory effect of the drug **(mechanism C)**. Constitutive activation of a signalling pathway downstream of a tyrosine kinase (TK), such as a SRC family member, can alleviate the dependence of the cancer on the original TK target **(mechanism D)**. Amplification of the *BCR-ABL1* gene leading to overproduction of the TK can confer relative resistance to an inhibitor **(mechanism E)**. Figure from Krause, D.S. and Van Etten, R.A., 2005.[95]

cation transporter-1 (OCT-1, SLC22A1), may also affect the ability of imatinib to successfully enter cells.[96,97] The adenosine triphosphate-binding cassette (ABC) transporter ABCB1 (MDR-1), of which imatinib is a substrate,[98-100] is involved in imatinib efflux, and has been linked to the development of resistance.[101] Excessive binding of imatinib to plasma proteins, such as α1-acid glycoprotein-1 (AGP1), has also been proposed to reduce the levels of active drug, limiting the therapeutic activity and leading to resistance.[102] Finally, activation of downstream signalling pathways, such as the SRC family kinases (reviewed by Quintás-Cardama *et al.*[103]), can lead to resistance as they alleviate the dependence of CML on Bcr-Abl1, allowing the cancerous cells to proliferate and survive, even in the presence of TKI.

Secondary resistance often evolves from primary resistance;[91] however, unlike primary resistance, the molecular mechanisms leading to secondary resistance are better understood (**Figure 1.10**). These mechanisms often involve escape from Bcr-Abl1 inhibition, leading to the loss of response. The major mechanism by which this occurs is clonal expansion of leukaemic cells with mutations in the Bcr-Abl1 tyrosine kinase domain.[80,104] Over 50% of all secondary resistance cases occur through this mechanism.[80] Mutations clustered within the ATP-binding region of Bcr-Abl1 prevent imatinib, but not ATP binding,[105] thus the kinase remains active and the disease persists. Of note are patients with mutations in the nucleotide-binding loop (P-loop), or the T315I (threonine to isoleucine substitution at base 315) mutation, for whom there is a particularly poor prognosis.[106]

In total, over 100 different Bcr-Abl1 mutations have been reported to date.[107] Thus, identification of patients unlikely to respond well to imatinib treatment, or further analysis of those that have already developed resistance, is quite complex. The T315I mutation[108] is also known as the gatekeeper mutation, because the substitution maps to the nucleotide-

binding site of *ABL1*.[109] The T315I mutation is located within the hinge region of the enzymatic cleft of the Abl1 kinase, where critical stabilizing hydrogen-bond interactions occur between the TKI and the active site of the kinase.[109] Therefore, because this critical hydrogen bond is unable to form, this mutation prevents the binding of imatinib (and other FDA (Food and Drug Administration, United States of America) approved TKIs, dasatinib, nilotinib and bosutinib), to Bcr-Abl1. Thus, the T315I mutation is the only mutation resistant to all frontline FDA-approved TKIs in current use. The other most commonly occurring mechanism that leads to imatinib resistance is the overproduction of *BCR-ABL1*, typically through genomic amplification or the acquisition of additional Ph chromosomes,[108,110,111] which occurs in approximately 10% of resistant patients.

1.3.1.3 Importance of CML stem cells for survival and TKI resistance

A further mechanism for TKI-resistance occurs via the CML stem cell (CD34+CD38-, CD133+;[112] also known as leukaemic stem cells [LSCs]). CML stem cells have been demonstrated to be insensitive to TKI therapy, and are thought to have a significant responsibility for minimal residual disease in TKI-treated patients.[113,114] While these LSCs have similar properties to normal haematopoietic stem cells (HSCs), as they are quiescent and have self-renewal capacity,[114] evidence is now emerging that these LSCs are not dependent on Bcr-Abl1 kinase activity for their survival.[112,113] Furthermore, functional CML LSCs are able to survive *in vitro* for a prolonged time, despite complete Bcr-Abl1 inactivation and the withdrawal of cytokines required for stem cell growth.[113] Transgenic mouse studies have demonstrated that these cells also retain long-term repopulating capacity after transplantation into recipient mice;[115] hence, these cells are thought to be responsible for relapse in patients who stop TKI therapy, even after prolonged negative molecular response, thought to be remission, with imatinib treatment.[115,116]

1.3.2 Second Generation Tyrosine Kinase Inhibitors

Crystallographic studies predicted that the ABL1 kinase domain from most imatinib-resistant mutants could be bound using inhibitors with greater flexibility in their conformational requirements than imatinib. Thus, second generation TKIs were developed (**Figure 1.11**).[117] Currently, three second generation TKIs, dasatinib and nilotinib, have shown excellent clinical potential, and both received FDA approval for front-line CML treatment in 2010[118] and subsequent TGA (Therapeutics Goods Administration, Australia) approval in 2011.

1.3.2.1 Dasatinib

Dasatinib (BMS-354825, chemical structure shown in **Figure 1.12 [A]**) is an orally bioavailable dual *Src*/Abl1 kinase inhibitor[119,120] that exhibits a two-log increased potency relative to imatinib, and demonstrates activity against 14 of 15 known imatinib-resistant Bcr-Abl1 mutants.[117,121] It is structurally unrelated to imatinib,[80] and is able to inhibit both Src family and Abl1 kinases. Dasatinib binds Bcr-Abl1 with less stringent requirements, and in both the active and inactive conformations, resulting in a 325-fold increase in potency against cells expressing wild-type Bcr-Abl1 when compared to imatinib.[120,122]

1.3.2.2 Nilotinib

Nilotinib (AMN107, chemical structure shown in **Figure 1.12 [B]**) is an aminopyrimidine that is a structural derivative of imatinib.[120,123] Like imatinib, nilotinib binds the Abl1 kinase domain in the inactive conformation,[123] but with increased affinity, resulting in a 20-fold increase in potency against wild-type *BCR-ABL1* expressing cells, when compared with imatinib.[123] This increased affinity was obtained through alteration of the side groups present on the imatinib molecule.[124] Nilotinib also exhibits anti-proliferative activity against 32 of 33 known imatinib-resistant mutations *in vitro*.[120,123] Furthermore, it should be noted that both dasatinib and nilotinib have a small, yet distinct, subset of resistance causing mutations, unique to each TKI. However, the T315I mutation is the only mutation resistant

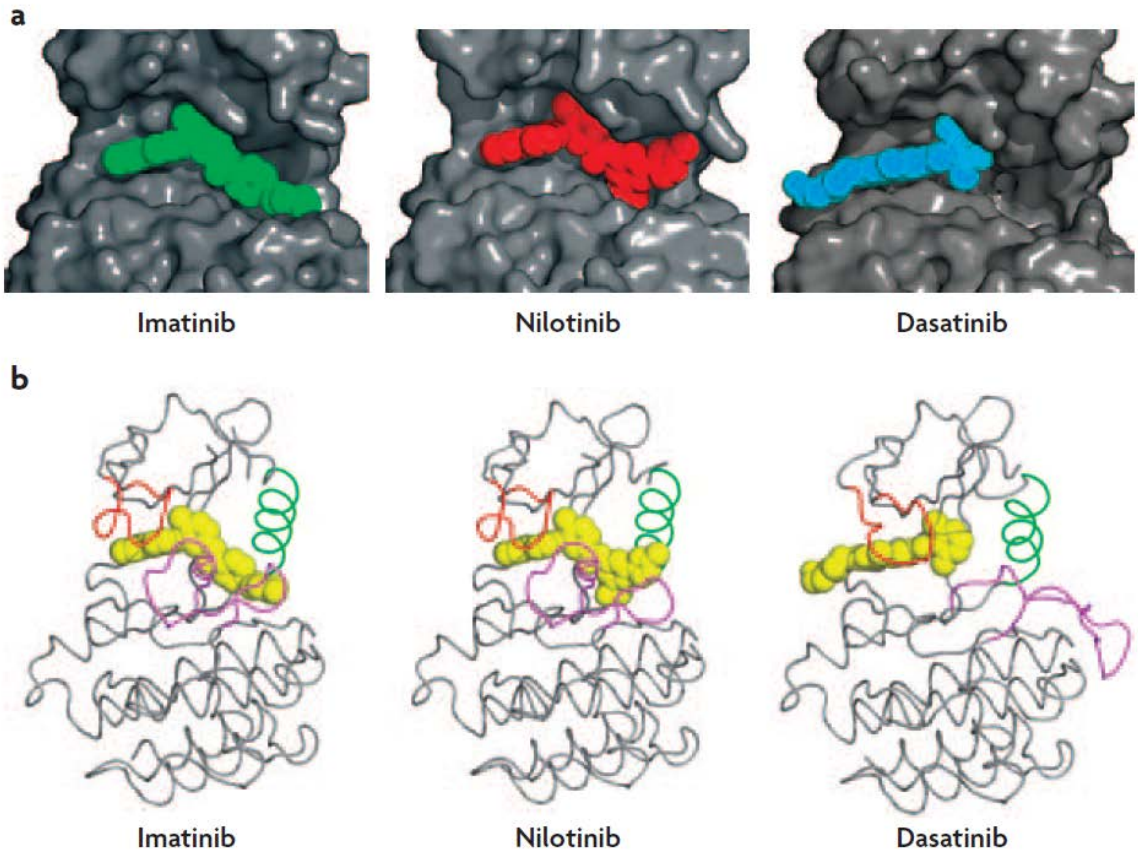


Figure 1.11 Structure of Abl1 in complex with imatinib, nilotinib and dasatinib

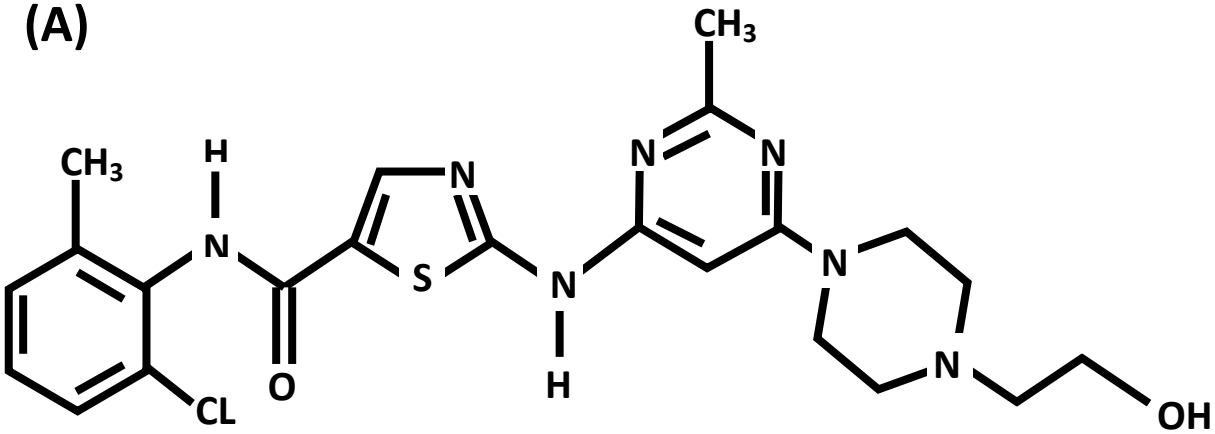
(A) Surface representations of crystals structures of Abl1 kinase domain in complex with imatinib (green), nilotinib (red) and dasatinib (blue). **(B)** Comparison of the different binding modes of the three Abl1 inhibitors: imatinib (left), nilotinib (middle) and dasatinib (right). The positions of the nucleotide-binding loop (P-loop; red) and activation loop (A-loop; magenta) vary according to whether the kinase is in an active or inactive conformation. Imatinib and nilotinib block the kinase in an active conformation only, while dasatinib can bind both the active and inactive kinase (shown here binding the inactive conformation). The green helix is helix C, which often moves between the active and inactive states of kinases. Figure from Weisberg *et al*., 2007.[125]

Figure 1.12 The chemical structures of Dasatinib (A) and Nilotinib (B)

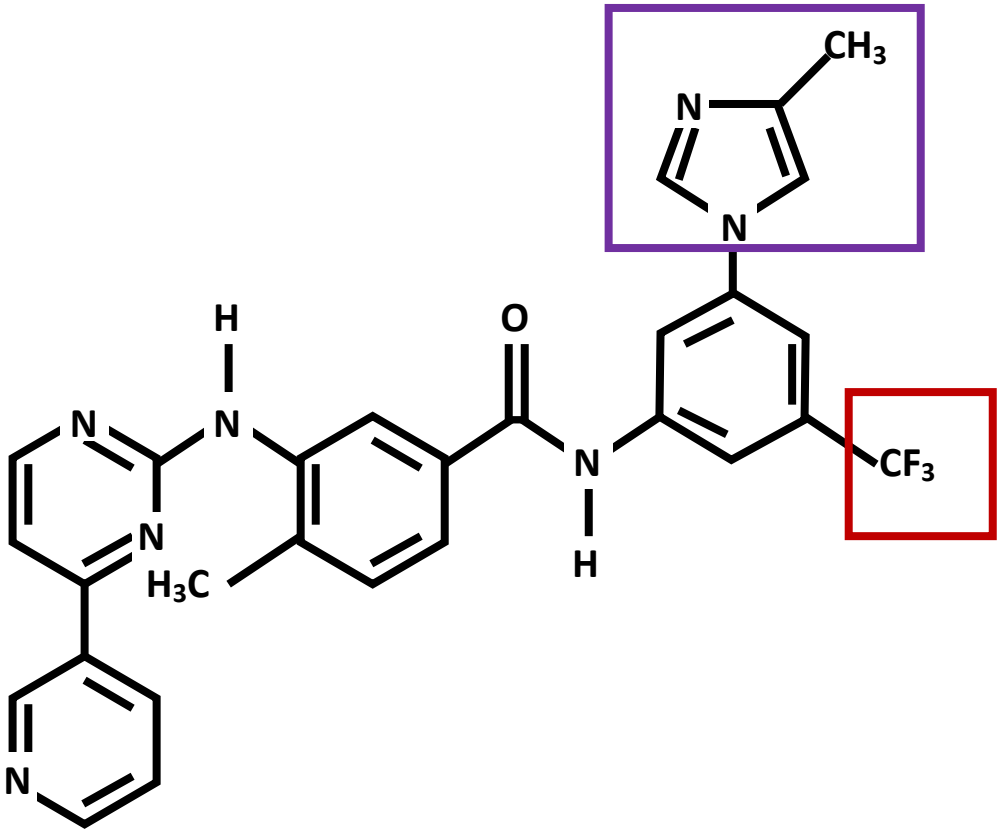
(A) Dasatinib binds both the active and inactive conformation of the Abl1 kinase domain, and of Src kinases. The structural conformation of dasatinib allows it to have potency approximately 325-fold higher than imatinib.[120] The molecular formula for Dasatinib is $C_{22}H_{26}CN_7O_2S•H_20$, with a molecular weight of 506.02g/mol (monohydrate).

(B) Nilotinib binds the inactive conformation of the Abl1 kinase domain. The only structural difference between nilotinib and imatinib is the substitution of the piperazine ring with a trifluoromethyl / imidazole phenyl moiety (RED/PURPLE) which results in increased potency and a better topological fit for nilotinib in the kinase domain.[123] The molecular formula for Nilotinib is $C_{28}H_{22}F_3N_7O•HCl•H_20$, with a molecular weight of 529.516g/mol.

(A)



(B)



to all three TKIs, as previously stated.[80,120,126] Hence, treatment of CML with these three TKIs will not be successful in all cases.

1.4 Predictors of Patient Outcome

Imatinib therapy results in excellent outcomes for the majority of CP-CML patients; however, it has been recently estimated that between 30-40% of CP-CML patients treated solely with imatinib will not benefit optimally long term due to primary or secondary imatinib resistance,[93] as discussed above. Although, approximately half of imatinib resistant patients can be salvaged with second-generation TKIs (dasatinib or nilotinib),[127] alternative treatment strategies are still imperative for this group. These strategies must be tailored towards each individual by firstly, identifying them at diagnosis, and secondly, having a greater understanding of the underlying intrinsic causes of poor response. Ultimately, the prediction of those patients who will not respond optimally to imatinib, despite subsequent salvage therapy, will allow therapeutic strategies critical to improving their clinical outcome to be developed. Currently, there are three main pre-therapy predictors of relevance to this project: the $IC50^{imatinib}$, the OCT-1 expression level and OCT-1 activity. In addition to these, a number of response predictors currently exist and these have been associated with variable success in the clinic.

1.4.1 Pre-therapy Prognostic Indicators

1.4.1.1 Patient Age

Before the advent of imatinib and the TKI era, older patients diagnosed with CML had an increased risk of poor outcome and response, which is emphasized by the fact that age, is an important determinant for both the Sokal and Hasford prognostic indices, described

below.[128,129] However, age is no longer associated with poor response in the imatinib era,[130,131] as older patients (age ≥ 60 years) were demonstrated to have similar CCyR rates, compared to younger patients (age < 60 years; 87% vs. 79% respectively; *P* = 0.28) and survival rates (12% vs. 9% respectively; *P* = 0.57).[130]

1.4.1.2 Sokal, Hasford and EUTOS Scores

The Sokal[128] and Hasford[129] prognostic scores were developed in the pre-imatinib era to predict patient outcome, while the European Treatment and Outcome Study (EUTOS)[132] score was recently developed based on imatinib-treated patient outcome. Both Sokal and Hasford scores group a patients' prognostic risk into one of three categories: low, intermediate or high, while the EUTOS score identifies low- and high-risk categories. The Sokal score was developed from a study of 813 Ph+ CML patients, and is comprised of four discriminatory factors identified by hazard ratio analysis: spleen size, blast-cell percentage, patient age and platelet count.[128] Hazard ratios were calculated for each patient, allowing each patient to be segregated into either the low-risk (Sokal score < 0.8), intermediate-risk (Sokal score 0.8 - 1.2), or high-risk group (Sokal score > 1.2). It is important to note that although the Sokal score was developed in the era where busulfan was the most common therapeutic modality for CML, it still remains relevant and useful today.

The Hasford score was developed from a large meta-analysis of 1,303 CML patients treated with IFN-α, using six covariates: patient age, spleen size, blast count, platelet count, eosinophil count, and basophil count.[129] A learning set of 908 patients was used to develop the prognostic score, with three distinct risk groups identified: low-risk (Hasford score < 780), intermediate-risk (Hasford score 780-1480), or high-risk group (Hasford score > 1480). A test set of 285 patients was then able to confirm and validate the predictive ability of the Hasford score.

As both the Sokal and Hasford scores were developed in the pre-imatinib era, the European Leukemia-Net (ELN) recently sought to develop a new scoring system using data from imatinib treated patients. The EUTOS score was generated from the data of 2,060 CML patients on imatinib-based regimes, and requires the percentage of basophils in the peripheral blood and the spleen size at diagnosis to classify patients as either low- or high-risk.[132] It was designed to predict for failure to achieve CCyR by 18 months on imatinib therapy (imatinib failure accordingly ELN guidelines),[133] and was suggested to have superior prognostic power for imatinib-treated patients when compared to the Sokal score.[132] However, as these classical prognostic scores are based on clinical and laboratory characteristics, there is concern that they may not be adequate to identify the variability in patient response observed with TKI therapy.[134]

1.4.1.3 The $IC50^{Imatinib}$

In 2005, an *in vitro* study was performed by White *et al*.[135] to determine the concentration of imatinib required for 50% inhibition (IC_{50}) of Bcr-Abl1 kinase activity (now known as $IC50^{imatinib}$), using mononuclear cells (MNCs) from 62 *de novo* CP-CML patients. However, as Bcr-Abl1 can be rapidly degraded and de-phosphorylated during cell lysis[108] it is not ideal for measuring kinase activity. An alternative measure of Bcr-Abl1 kinase activity was needed. Previously, the adaptor protein CrkL (CT10 regulator of kinase like) was demonstrated to be specifically and constitutively phosphorylated (p-CrkL) by Bcr-Abl1 in CML cells.[136-138] CrkL is able to be measured reproducibly and quantitatively in patient samples,[108] providing an excellent surrogate marker for Bcr-Abl1 kinase activity.[110] Hence, the level of *in vitro* inhibition of p-CrkL was used to determine the $IC50^{imatinib}$. The $IC50^{imatinib}$ has been demonstrated to predict molecular response in *de novo* CML by 12 months, and when combined with a Sokal score, this predictive value is increased.[135] Flow cytometry protocols for measuring pCrkL/CrkL have also shown predictive value for CCyR at 12 months[139] using

MNCs, whereas protocols involving CD34+ cells have been unable to replicate this predictive potential, for unknown reasons.[140,141]

1.4.1.4 The Human Organic Cation Transporter-1 (OCT-1)

Drug transporters are responsible for the uptake and elimination of imatinib, regulating the intracellular drug concentration. OCT-1 is now well established as the major active influx pump for imatinib.[96,97,99,142] The OCT-1 protein is a member of the solute carrier (SLC) 22 superfamily,[143,144] and specifically mediates the uptake of many drugs via a membrane associated ATP-dependent protein channel.[99] OCT-1 was cloned from humans in 1997,[145,146] after first being discovered in rats in 1994 by expression cloning[147] and is mainly expressed in the liver.[148] OCT-1 has a predicted structure comprising 12 α-helical transmembrane domains (TMDs), a large glycosylated extracellular loop between TMDs 1 and 2, a large intracellular loop between TMDs 6 and 7 with consensus sequences for phosphorylation, and an intracellular N and C-terminus (**Figure 1.13**).[148] Thomas *et al*.[99] first demonstrated imatinib was actively transported into patient cells specifically via OCT-1 using a combination of OCT inhibitors (procainamide, verapamil and amantadine) that reduced the intracellular uptake and retention (IUR) of imatinib *in vitro*.

Our laboratory has since confirmed this finding using the inhibitor prazosin in an *in vitro* IUR assay using patient PB-MNC.[96] Wang *et al.*[149] then used an OCT-1 over-expressing KCL22 cell line to demonstrate that uptake of the OCT-1 substrate, ASP (4-[4-(dimethylamino)-styryl]-*N*-methylpyridinium), was increased in cells with high OCT-1 mRNA expression, compared to control cells, demonstrating that OCT-1 mRNA correlates with OCT-1 function and imatinib uptake. Other studies[97,99] have demonstrated that the baseline level of OCT-1 mRNA expression varies within primary CML patient samples, as well as during imatinib therapy.

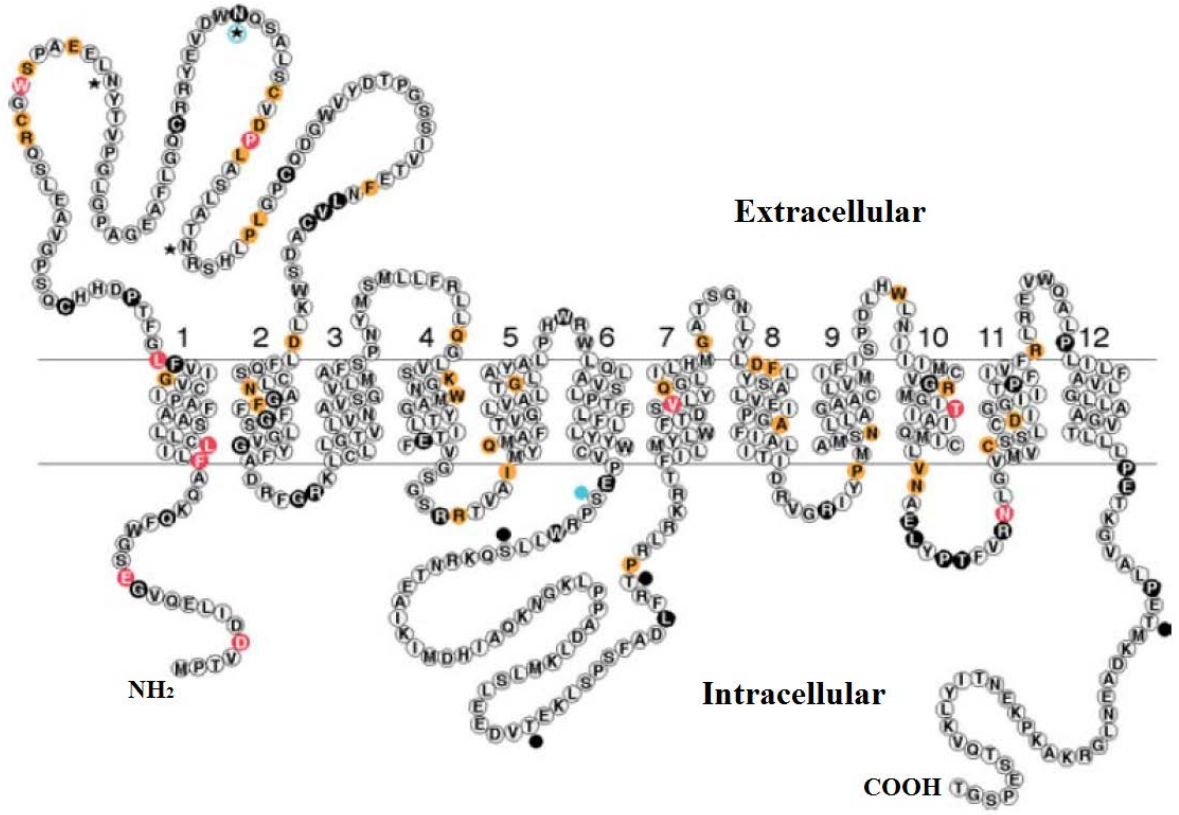


Figure 1.13 Protein structure of the organic cation transporter 1 (OCT-1)

Predicted topology and amino acid sequence of the human OCT-1 transporter. The protein has 12 transmembrane domains and a large intracellular and extracellular loop. Black dots represent protein kinase C phosphorylation sites. Stars represent N-glycosylation sites. Coloured dots represent conserved regions of the sequence with other SLC22 family members. Figure from Koepsell *et al*., 2003.[143]

Hence, it is thought that expression and activity of influx proteins may play an important role in patients' varying responses to TKI therapy. This is discussed in more detail below.

1.4.1.4.1 OCT-1 mRNA Expression

The first investigations into OCT-1 mRNA were performed by Thomas *et al.* [99] who demonstrated that imatinib was transported into cells by an active uptake mechanism (namely by OCT-1) and that OCT-1 was expressed at varying levels in CML cell lines and patient cells. They postulated that further research may reveal a correlation between the expression of these transporters and clinical prognosis. Currently, there are conflicting views on the predictive value of OCT-1 mRNA expression for patient response. Crossman *et al.*[97] demonstrated that pre-imatinib OCT-1 expression was significantly lower in patients who did not respond (at least 65% Philadelphia chromosome positive during the first 10 months of treatment), compared to patients who did respond (achieved CCyR by 12 months). Wang *et al.*[149] also confirmed that patients with higher pre-imatinib OCT-1 expression levels had a significantly increased chance of progression-free survival, overall survival and achieving CCyR after 6 months of therapy. Most recently, OCT-1 expression has also been demonstrated to predict the achievement of 6-year major molecular response (MMR).[150]

Contradicting these findings, White *et al.*[142,151] found that the OCT-1 mRNA level is not a strong predictor of molecular response, and Zhang *et al*.[152] found no correlation between OCT-1 mRNA levels and imatinib resistance. The important variable to note with the studies that found no predictive correlation between OCT-1 mRNA expression and patient response is that the patient cohorts consisted solely of newly diagnosed, untreated CP-CML patients. This is not true for the studies that found OCT-1 mRNA to be predictive, as they used a variety of cohorts, which included patients in stable CCyR;[150] patients in CP, AP and BC;[149]

patients in CP, with evidence of clonal evolution and AP;[97] and even patients who had received prior imatinib treatment; all factors which could influence the results obtained.

It is also important to note that the level of OCT-1 expression within a patient sample has been demonstrated to relate to the cell type present in that sample.[153-155] Specifically, OCT-1 expression is higher in PB-polymorphonuclear cells (PB-PMNCs), compared to PB-MNCs,[155] and CD34+ cells have lower expression than the more mature MNC compartment.[156,157] However, measuring the pre-treatment OCT-1 expression levels from isolated PB-PMNC, PB-MNC or total leukocytes did not correlate with 6 or 12 month response to imatinib therapy.[154] This highlighted that while consistent cell types should be used when measuring OCT-1 expression in any prognostic situation, it may not be the sole determining factor regarding the reported discrepancies in the prognostic ability of OCT-1 mRNA.

Recently, Giannoudis *et al.*[158] provided an important alternative explanation for why there are discrepancies observed in OCT-1 mRNA assessment in different studies. Their study investigating OCT-1 polymorphisms[158] demonstrated that if one of the PCR primers used to measure OCT-1 expression spanned a polymorphism, the apparent mRNA expression was affected. This observation was exemplified upon re-analysis of their previous work where the terminal base of the forward PCR primer hybridized to a frequent OCT-1 polymorphism. When the terminal base was omitted, the correlation between expression and imatinib outcome was lost.[149] It should be noted that the primers used by White *et al*.[142,151] were located in exon 9 of OCT-1, whereas those used by Wang *et al*.[149] and Giannoudis *et al*.[158] were located in exon 7. Thus, the analysis of OCT-1 mRNA expression has not yet been established as a consistent method for determination of patient response to imatinib therapy.

1.4.1.4.2 OCT-1 Polymorphisms

Multiple non-synonymous single nucleotide polymorphisms (SNPs) of OCT-1 have been described,[158-162] all of which dramatically affect transport function. Studies have mainly focussed on investigating the association between OCT-1 SNPs and patient response to TKI therapy. Kim *et al.* demonstrated that the L160F (rs683369) SNP was significantly associated with treatment failure, including loss of response in a cohort of 229 CML patients.[163] Bazeos *et al.* were able to show a correlation between the G401S (rs34130495) SNP and imatinib response (MMR achievement) in 132 imatinib treated CP-CML patients.[153] Additionally, Maffioli *et al.* demonstrated that the rs6935207 SNP was associated with optimal responders in a small cohort study of 65 *de novo* CP-CML patients.[164] However, the caveat of these studies is in the actual frequencies of the SNPs, with percentages of 18%, 4.5% and 6%, respectively.

Most recently, Giannoudis *et al.*[158] studied the effect of 23 OCT-1 SNPs on imatinib treatment outcome in 336 newly diagnosed CP-CML patients and were only able to demonstrate an association with outcome for one SNP, rs35191146 (M420del). Patients with the M420del were associated with an increased probability of imatinib treatment failure (defined as a change of treatment due to either unsatisfactory response or imatinib intolerance). A complex interaction was also determined to exist between the M420del and M408V SNPs using cell line and patient data, demonstrating that patients with M420del and none or one copy of M408V, had a significantly increased risk of treatment failure.[158] However, it must be noted that the frequency of this SNP (M420del) was once again low at 18.5%. Furthermore, the results of this study were recently questioned by Tzvetkov *et al*.[165] who were unable to identify the M420del-M408 haplotype in a large cohort of 371 Caucasians or in the 1000 Genomes Project dataset. Taken together, conjecture remains

regarding the clinical relevance of OCT-1 SNP analysis, as it is yet to conclusively demonstrate accurate response prediction for the majority of patients. This is supported by studies in our laboratory that were unable to demonstrate any significant association between OCT-1 SNPs and achievement of MMR by 24 months on imatinib therapy in 136 CP-CML patients.[166]

1.4.1.4.3 OCT-1 Activity

To further the work with OCT-1 expression, it was postulated that a functional assay measuring imatinib uptake in CML patient cells could have greater predictive value. Thus, the term OCT-1 activity (OA) was coined by White *et al.*[151] to describe the difference in the IUR of [^{14}C]-labelled imatinib, with and without OCT-1 inhibition, using CML patient MNCs (**Figure 1.14**). OCT-1 inhibition is achieved using the potent and selective inhibitor prazosin,[167] enabling the functional activity of the OCT-1 protein to be estimated in relation to imatinib transport. A study by White *et al.*[151] defined the median OA across 56 CP-CML patients enrolled in the Australasian Leukaemia and Lymphoma Group TIDEL (Therapeutic Intensification in *De-novo* Leukaemia) I trial, as 7.2ng/200,000 cells. Hence, patients with an OA < 7.2ng/200,000 cells are classified as having low OA, while those with OA > 7.2ng/200,000 cells have high OA. When patient response was analysed, a significant correlation was observed between the IC50^{imatinib} and OA; in particular, patients with a low IC50^{imatinib} had a significantly increased OA, compared to patients with high IC50^{imatinib} values.[151] Using patients enrolled in the TIDEL I trial, White *et al.*[151] demonstrated that over 24 months of imatinib treatment, patients with high OA achieved significantly greater molecular responses (MMR) than patients with low OA. It has now been demonstrated that this trend continues after 60 months of imatinib treatment (**Figure 1.15**).[142] Interestingly, the average daily dose of imatinib received did not affect the response of high OA patients, however low OA patients, who received an average imatinib dose of 600mg/day, achieved

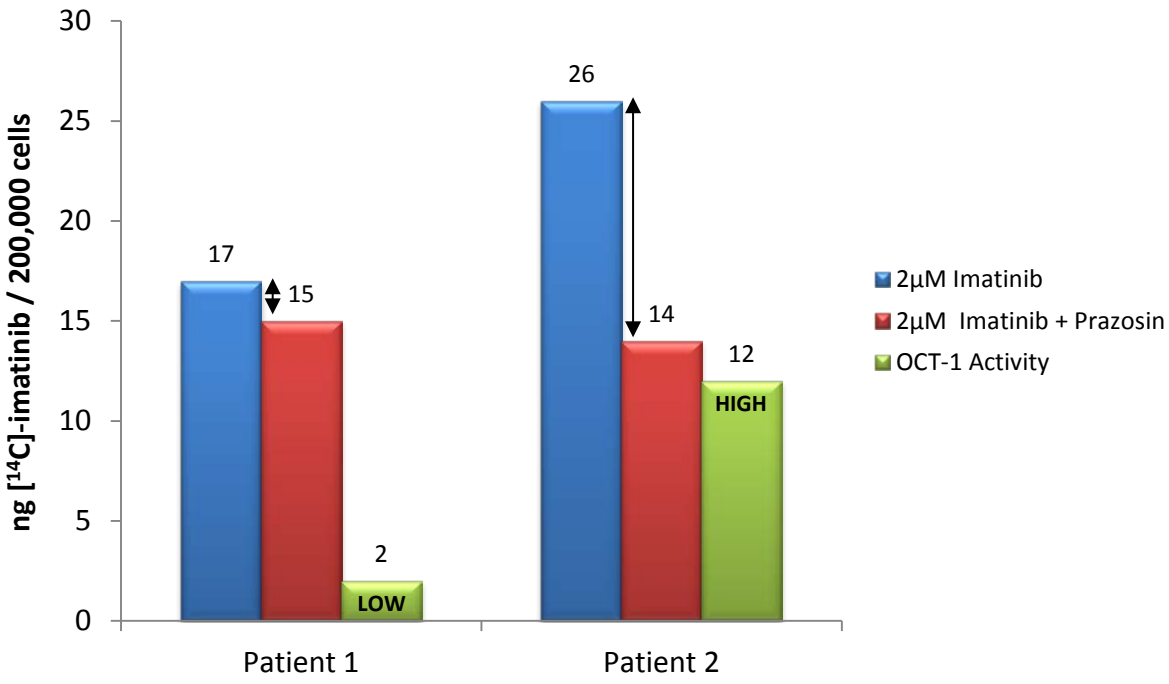


Figure 1.14 Calculation of OCT-1 activity (OA) using representative patient data

The defined median OA value is 7.2ng/200,000 cells. Patient 1 has a low OA, as the difference between the 2µM imatinib (**Blue**) and the 2µM imatinib + Prazosin (**Red**) IUR values is 2ng/200,000 cells. Patient 2 has a high OA, as the difference between the IUR values is 12ng/200,000 cells.

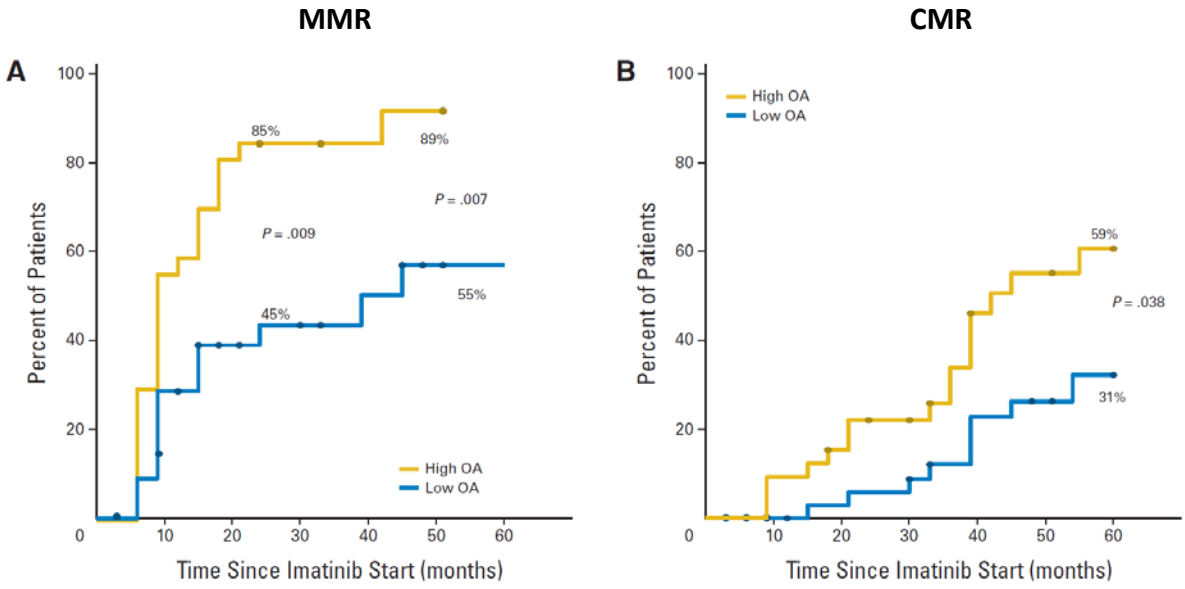


Figure 1.15 Relationship between OCT-1 activity and clinical response to imatinib therapy

The percentage of patients achieving **(A)** major molecular response (MMR) and **(B)** complete molecular response (CMR) on the basis of low and high OCT-1 activity (OA) patient groups. Kaplan-Meier curves demonstrate that a significantly greater proportion of patients who had high OA achieve MMR and CMR by 5 years compared to patients with low OA. Figure from White *et al.*, 2010.[142]

significantly better molecular responses by 24 months than those who were treated with lower doses.[151] This indicated that dose increase in patients with low OA could be an important factor for overcoming suboptimal response.

The OA also predicted for event-free survival (EFS) and overall survival (OS), as patients with high OA had significantly higher EFS (74% high OA, 48% low OA; *P* = 0.03) and OS (96% high OA, 87% low OA; *P* = 0.028), than patients with low OA.[142] The OA has also been used to further stratify the patients into quartiles, with Q1 being the lowest (< 4ng/200,000 cells) and Q4 the highest (> 10ng/200,000 cells).[142] Patients with OA levels in Q1 had a significantly increased risk of failing to achieve MMR and of disease progression (**Figure 1.16**).[142]

1.4.1.4.4 Treatment of Patients with Low OA

The optimal treatment for patients with low OA, especially those in Q1 (lowest OA quartile), is still unclear. Dose escalation can be effective,[151] if the patient is able to tolerate this; however, often Q1 patients are not able to withstand the toxicity associated with escalated imatinib dosing. The second generation TKI's, dasatinib and nilotinib, are an option for these patients, but their long-term safety profile is not well established, unlike the superior safety profile of imatinib. Also, even though is it postulated that OA should not have any relevance for dasatinib or nilotinib response, this has not yet been tested. Therefore, it is important to determine if there are factors that underlie the range of OA levels observed in CP-CML patients, and whether patients with low OA and poor response to imatinib have different overall disease characteristics associated with alternative biological mechanisms. Intriguingly, in patients who averaged at least 600mg of imatinib daily, and had high OA, approximately 20% still failed to achieve MMR by 18 months,[151] demonstrating that even with these patients who would be expected to respond favourably, there are unknown factors (e.g. drug adherence) affecting their response. Little research has currently been

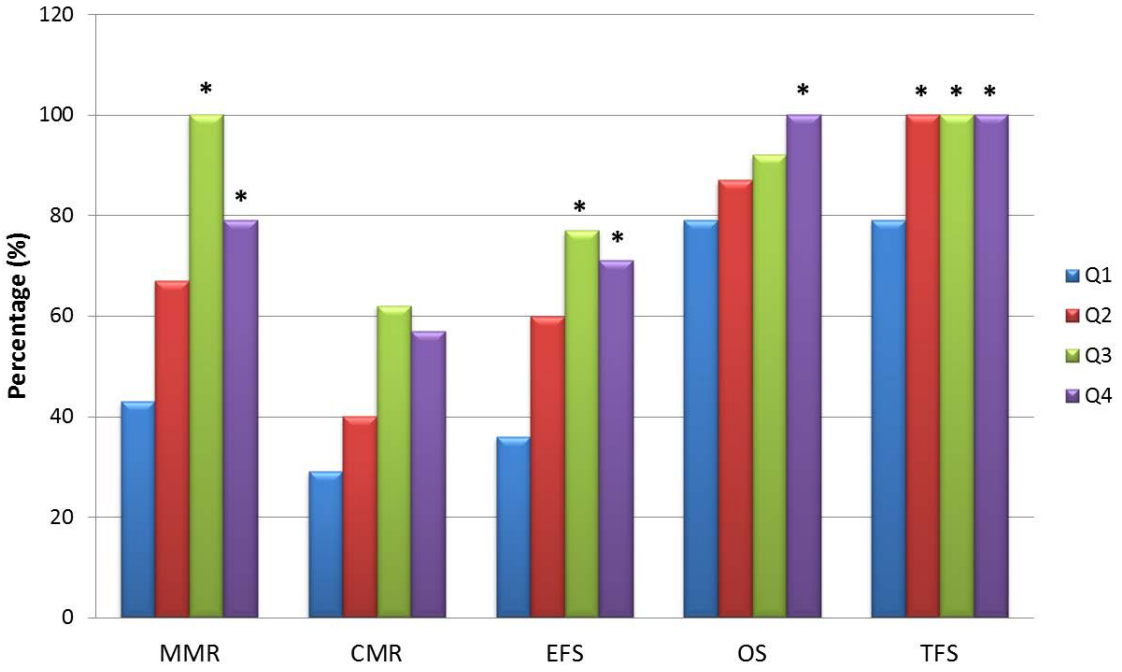


Figure 1.16 Patient response after 5 years of imatinib therapy on the basis of OCT-1 activity divided into quartiles

The percentage of patients achieving major molecular response (MMR), complete molecular response (CMR), event-free survival (EFS), overall survival (OS) and transformation-free survival (TFS) after 5 years of imatinib treatment on the TIDEL I trial, were assessed based on OCT-1 activity (OA) quartiles. Significantly, patients with OA in the lowest OA quartile (Q1; < 4ng/200,000 cells) achieved significantly lower rates of MMR, and had poorer OS, EFS and TFS. * Indicates significantly different from quartile 1 (Q1; P <0.05). Figure adapted from White *et al.,* 2010.[142]

undertaken in this area; however, it has been recently suggested that the OA assay may be providing a complex surrogate for many other transporters,[168] or that within the patient MNC fraction, the stage of differentiation or unique phenotype of the predominant cell population is influencing the OA result.[142] Consequently, the OA assay may be indicating the intrinsic biology of the disease clone, as patients with low OA (especially Q1) may have a more genotypically advanced disease stage, resulting in their increased risk of rapid disease progression,[142] or have a disease clone that has originated from a different progenitor than that of patients who respond well.

1.4.1.5 Efflux Transporters

The efflux of drugs from malignant cells is an important mechanism of resistance to various therapeutics. Variations in these transporters (e.g. SNPs) can potentially alter the intracellular concentrations of their substrates, and may explain differential response to TKI treatment. It has been well documented that ABCB1 (protein; also known as MDR1) is responsible for the efflux of imatinib from the cell, as imatinib is a substrate for ABCB1.[98-100] ABCB1 is also involved in multi-drug resistance (MDR) and the efflux of other anti-cancer agents.[98] Although more controversial,[169] imatinib has also been suggested to be a substrate for efflux by ABCG2 (breast cancer resistance protein, BCRP).[170,171] Both proteins are members of the ATP-binding cassette (ABC) transporter family.

1.4.1.5.1 ABCB1 Polymorphisms

The interaction between *ABCB1* polymorphisms and imatinib response has been well studied. Dulucq *et al.*[172] investigated the three most relevant *ABCB1* gene SNPs (rs1128503 – C1236T; rs1045642 – C3435T; and rs2032582 – G2677T/A) in 90 imatinib-treated CML patients, demonstrating patients homozygous for the 1236T allele had significantly greater MMR achievement (85% versus 47.7% for other genotypes; $P = 0.003$), and patients with the 2677G allele had significantly poorer MMR achievement (77.8%, TT/TA versus 47.1%,

GG/GA/GT; *P* = 0.018).[172] Conversely, Deenik *et al.*[173] examined 46 CP-CML patients and demonstrated that patients with a homozygous 1236C genotype had significantly greater MMR incidence after 12 months of high-dose imatinib therapy, compared to the 1236CT/TT genotype (92% versus 52% and 50%; *P* = 0.02). Hazard ratio analysis also revealed a 4-fold reduction in the probability of obtaining complete molecular response (CMR) for the CT and TT alleles (0.25 and 0.27, respectively; *P* = 0.01). This prompted Dulucq *et al.*[174] to update their earlier report using a significantly larger patient cohort (n = 557). However, they were unable to demonstrate any significant association between *ABCB1* SNPs and MMR, indicating that their earlier results were biased due to the small patient cohort (n = 90) used. Supporting this, Kim e*t al.,*[163] Marin *et al.*[150] and Seong *et al.*[175] have also previously found no association between *ABCB1* SNPs and imatinib response.

More recently, Ni *et al.*[176] reported that the rate of resistance to imatinib was higher in the C3435T CT/TT genotype, compared to the CC genotype (59.4% versus 25%, respectively; *P* = 0.023); and in the C1236T TT genotype, compared to the CT/CC genotype (75% versus 31.3%, respectively; *P* = 0.004). Supporting this, Angelini *et al.*[177] reported a significant correlation between the C3435T SNP and CMR in Caucasians with the CC genotype, compared to the CT/TT genotype (*P* = 0.005). In contrast to this, Maffioli *et al.*[164] reported a protective effect from primary failure for the 2677T allele (*P* = 0.016), while the CC genotype at the C3435T locus was associated with primary failure (*P* = 0.0004). Taken together, the impact of *ABCB1* SNPs on imatinib response is still unclear, with no predictive analysis currently recommended to determine patient response.

1.4.1.5.2 ABCG2 Polymorphisms

The interaction between *ABCG2* gene polymorphisms and imatinib are less well studied, and remain controversial. Kim *et al.*[163] investigated 16 SNPs in 5 genes from 229 CML patients

and identified that the *ABCG2* Q141K and C421A AA genotype exhibited significantly higher MMR and CMR rates, compared to the CA or CC genotypes (*P* = 0.004 and *P* = 0.006, respectively); and that the GG genotype for *ABCG2* V12M was associated with a lower CCyR rate (*P* = 0.02). This was supported recently by Seong et al.[175] who investigated 82 newly diagnosed CP-CML patients for the same *ABCG2* C421A polymorphism and found significant association between the *ABCG2* 421 AA variant and MMR achievement (CC + CA versus AA genotype; *P* = 0.02). Conversely, Angelini *et al.*[177] were unable to demonstrate an association between *ABCG2* SNPs (V12M and Q141K) and imatinib response. Further complicating these results is the fact that these SNPs were only present in a small number of patients, meaning the applicability of any predictive analysis is questionable.

1.4.1.6 Genome-wide Gene Expression Profiling

Global gene expression profiling with microarrays has been used with varying success to elucidate biomarkers of disease phase or response in CML, which is discussed in further detail in **Chapter 3**. However, it would be remiss if the studies by Oehler *et al.*[178] and McWeeney *et al*.[179] were not discussed briefly here. Oehler *et al.*[178] recently re-analysed the landmark CML disease phase (CP, AP and BC) dataset of Radich *et al*.[180] and applied a probabilistic method called Bayesian Model Averaging (BMA) to identify a small gene set which discriminated the different disease phases. A 6 gene-set (*NOB1*, *DDX47*, *IGSF2*, *LTB4R*, *SCARB1*, and *SLC25A3*) was identified that separated CP from BC, and validation of this signature with an independent 88 patient cohort demonstrated that early CP patients (untreated) were able to be discriminated from late CP (patients who had failed imatinib, but were still in CP), as well as AP and BC patients. This is one of the first microarray studies to successfully validate the resulting gene-set; however, this gene-set has not been incorporated into routine diagnostic use and independent validation by a different group is still required.

More recently, McWeeney *et al*.[179] profiled bone marrow CD34+ cells from 36 imatinib-naïve CP-CML patients to identify biomarkers predictive of imatinib failure (no CCyR by 12 months of imatinib therapy). A 75 gene-set was defined, which successfully discriminated the two groups (responders (achieve CCyR) versus non-responders (no CCyR)). When applied to a prospectively selected 23 patient validation cohort, this classifier correctly predicted 88% of responders and 83% of non-responders (86.9% overall accuracy).[179] In addition, gene ontology analysis suggested that the CP-CML patients destined to fail on imatinib therapy had a more advanced disease genotype than indicated by the morphological criteria alone. Furthermore, McWeeney *et al.* were able to demonstrate concordance between their gene-set and previously published advanced phase CML gene-sets.[12,181] While the findings of this study are important and exciting, a 75 gene-set requiring CD34+ cells is restrictive for application to routine diagnostic practice; hence, this gene-set has not been successfully utilised to stratify patients at diagnosis who are likely to respond to imatinib therapy.

1.4.2 On-therapy Prognostic Indicators

1.4.2.1 Imatinib Plasma Levels and Compliance

Imatinib plasma-trough concentrations were first measured by Picard *et al.*,[182] who used high-performance liquid chromatography-tandem mass spectrometry to measure imatinib levels in 68 CML patients. Imatinib plasma levels correlated with response (CCyR and MMR), allowing the authors to conclude that a minimum trough level of 1,002 ng/mL was necessary for optimal response. The authors also suggested that the measurement of trough levels may identify those patients who would benefit from increased imatinib dosing, and should form part of routine on-therapy disease monitoring. Larson *et al.*[183] performed a systematic study using 351 patients enrolled in the IRIS trial, and by measuring the trough imatinib levels at day 29, were able to confirm the results of Picard *et al.*[182] Patients who achieved a

CCyR and MMR had significantly higher trough imatinib levels, than those who did not, and the overall mean, steady-state trough level was 979 ng/mL for this patient cohort, confirming the findings of Picard *et al.* that a 1000 ng/mL trough concentration of imatinib is critical for molecular response.

Recently, Forrest *et al*.[184] were unable to support these data when they retrospectively analysed imatinib trough levels in 78 patients, as no significant difference was observed in the trough level of patients who did, or did not, achieve MMR. However, in contrast to this finding, Marin *et al.*[150] were able to support the results of the Picard and Larson groups, when they demonstrated in 84 patients that imatinib trough levels greater than 1,000 ng/mL were statistically associated with achievement of MMR, but not CMR. Recently, other groups have also validated these results when measuring trough levels at 6 months of imatinib treatment,[185] or by using slightly higher cut-off levels (<1,165 ng/mL).[186]

Furthermore, Marin *et al*. then sought to determine the underlying cause of low plasma levels by investigating patient adherence to TKI therapy. Using a microelectronic monitoring system (MEMS), 87 CML patients were assessed for compliance over a median 91 day period. Strikingly, compliance was strongly associated with MMR and CMR, and poor compliance was postulated to be the predominant reason for poor response in patients who had been treated with imatinib for prolonged periods (median 59.7 months).[150] This landmark study highlighted the fact that many CML patients are non-compliant to oral medications (no patient with adherence below 80% achieved MMR), indicating that all other predictive factors could be obsolete if patients do not actually take their medication as prescribed. This observation does not only apply to CML patients, but any patient with a

chronic illness who self-medicates, as approximately 50% of patients do not take their medications as prescribed.[187]

1.4.2.2 Early Molecular Response to Predict Long-Term Outcome

Measuring *BCR-ABL1* levels by RQ-PCR (quantitative reverse transcription-polymerase chain reaction) early in imatinib therapy were found to be predictive of subsequent imatinib response, when performed 3 months after imatinib initiation.[188,189] In these analyses, patients were segregated into three groups: those who achieved a 1-log reduction in *BCR-ABL1* levels (approximately equal to a *BCR-ABL1:ABL1* ratio of 10% in the international scale), those achieving between 1- to 2-log reductions or those achieving a greater than 2-log reduction following 3 months of therapy. Using data from the IRIS trial, all patients achieving greater than 2-log reduction in *BCR-ABL1* also achieved MMR by 24 months, while conversely, only 13% of patients failing to achieve a 1-log reduction achieved MMR. More recently, Quintas-Cardama *et al*.[190] reported similar findings using the cut-offs of a transcript ratio > 10% (not international scale), between 1% to 10%, or ≤ 1% at 3 months. The probability of achieving MMR was significantly lower in patients with a transcript ratio > 10%, when compared to patients with transcripts between 1% and 10%, or ≤ 1% (33% versus 53% and 80%, respectively). However, it was Marin *et al.*[191] who recently brought 3 month *BCR-ABL1* transcript levels back into the spotlight, with their study confirming the results of Quintas-Cardama *et al*. Using a receiver operating characteristic (ROC) curve, Marin *et al.* determined the optimal 3 month *BCR-ABL1* transcript level to be less than 9.84% (international scale), which not only identified those patients with the best outcome, but was independently associated with overall survival, progression-free survival, and event-free survival.[191] The predictive ability of 3 month *BCR-ABL1* transcript levels > 10% for identification of patients at risk of poor response has now been validated by multiple groups.[192-194]

Although 3 month *BCR-ABL1* transcript levels are an excellent prognostic indicator, patients are still required to receive TKI for 3 months before the transcript levels are measured and TKI therapy adjusted accordingly. Ideally however, the decision regarding optimal TKI therapy selection should be made at diagnosis to provide the greatest benefit to the patient, yet the development of a reliable assay to allow this selection to be made is still on-going.

1.5 Hypothesis and Aims

Intracellular drug transport impacts significantly on imatinib response. There are various biological factors intrinsic to CML patients that determine drug transport and they are likely related to the phenotype of the predominant leukaemic cell, as well as the dysregulation of key regulatory pathways associated with transporter expression, modification and activity.

1.5.1 Hypothesis

A combination of biological factors will delineate distinct subgroups of CP-CML patients at diagnosis. This will demonstrate that CP-CML is not a uniform disease entity, as indicated by variable patient response and provide new approaches for the sub-classification of CP-CML patients.

1.5.2 Aims

1. To determine the variation in CP-CML patient immunophenotype at diagnosis and to relate this to patient characteristics at presentation and imatinib response.

2. To investigate the gene expression patterns of newly diagnosed CP-CML patients from very low and very high OCT-1 activity groups, and then further extend these studies to include intermediate OCT-1 activity groups.

3. To perform data-mining and pathway analysis to select transporter regulatory pathways likely to impact on imatinib transport, and to examine their influence on drug uptake in these patient groups.

4. To investigate the global DNA methylation profile of CP-CML patients with a particular focus on low OCT-1 activity patients, and ascertain whether aberrant epigenetic programming may underlie poor imatinib response and disease progression.

2 MATERIALS AND METHODS

2.1 Commonly Used Reagents

Table 2.1 Suppliers of commonly used reagents

Reagent	Supplier	Catalogue Number
ABCG2 (Bcrp1) RPE antibody	R&D Systems	FAB995P
Acetic Acid	Fisher Scientific	0400
Acrylamide (40% Acrylamide/Bis solution, 37.5:1 ratio)	BioRad	161-0148
Agarose Type I, low	Sigma-Aldrich	A6013-250G
Anti-coagulant Citrate Dextrose Solution Formula A (ACD-A acid)	Baxter	P73A6
Anti-glycophorin A (GlyA) RPE antibody	BD Biosciences	340947
Anti-hCD31 (PECAM-1) APC antibody	R&D Systems	FAB3567A
Anti-hCD81 RPE antibody	R&D Systems	FAB4615P
Anti-hCD200R1 RPE antibody	R&D Systems	Fab3414P
Anti-hIntegrin αVβ5 (ITGB5) RPE antibody	R&D Systems	FAB2528P
Anti-human Frizzled-6 (FZD6) antibody	R&D Systems	AF3149
Anti-human LRP-6 antibody	R&D Systems	AF1505
Anti-hVEGF R2 (KDR) APC antibody	R&D Systems	FAB357A
Anti-goat IgG RPE Secondary antibody	R&D Systems	F0107
Anti-Mouse IgG – AP conjugate Antibody	Cell Signalling Technology	7056
β-mercaptoethanol	Sigma-Aldrich	M6250
Benchmark Protein Ladder	Invitrogen	10748-010
Bovine Serum Albumin (BSA) powder	Sigma-Aldrich	49418
Bromophenol Blue	Sigma-Aldrich	114391
CD3 (SK7) FITC antibody	BD Biosciences	349201
CD4 FITC antibody	BD Pharmingen	555346
CD7 RPE antibody	BD Biosciences	340581
CD8 RPE antibody	BD Pharmingen	555367
CD14 FITC antibody	BD Pharmingen	555397
CD15 FITC antibody	BD Pharmingen	555401
CD16 RPE antibody	BD Pharmingen	555407
CD19 FITC / PE-Cy7 antibody	BD Pharmingen	555412 / 557835
CD20 RPE antibody	BD Biosciences	347201
CD25 RPE antibody	BD Pharmingen	555432
CD33 RPE antibody	BD Biosciences	347787
CD34 (8G12) FITC antibody	BD Biosciences	348053
CD36 APC antibody	BD Pharmingen	550956
CD38 RPE antibody	BD Biosciences	347687
CD45 FITC antibody	BD Biosciences	347463
CD56 RPE antibody	BD Biosciences	347747
CD59 RPE antibody	BD Pharmingen	555764
CD63 RPE antibody	BD Pharmingen	556020
CD66b FITC antibody	BD Pharmingen	555724
CD71 FITC / RPE antibodies	BD Pharmingen	555536 / 555537
CD99 RPE antibody	Biolegend	318008
CD105 RPE antibody	BD Pharmingen	560839
CD117 RPE antibody	BD Biosciences	340529
CD133 RPE antibody	Miltenyi Biotec	130-080-801
CD243 (ABCB1) RPE antibody	Beckman Coulter	IM237OU

Chloroform	Merck	100776B
c-Myc (D84C12) Antibody (57-65 kDa) - Rabbit	Cell Signalling Technology	5605
c-Myc (phospho Ser62) Antibody (49 kDa) - Rabbit	Abcam	ab51156
c-Myc (phospho Thr58 + Ser62; E203) Antibody (57 kDa) - Rabbit	Abcam	ab32029
Complete (Protease Inhibitor Cocktail Tablets)	Roche	4693116001
D_c Protein Assay	BioRad	500-0112
Diethyl pyrocarbonate (DEPC) H_2O	MP Biomedicals Inc.	821739
Dimethyl Sulfoxide (DMSO)	Merck	K39661852
dNTP set (N = A, C, G, T)	GE Healthcare	27-2035-02
DTT (dithiothreitol) 0.1M	Invitrogen	18064-014
ECF substrate *Attaphos*	GE Healthcare	RPN 5785
E2F1 Antibody (70 kDa) - Rabbit	Cell Signalling Technology	3742
Ethanol	Merck	4.10230.2511
Ethylenediaminetetraacetic acid (EDTA)	APS	180-500G
Foetal Calf Serum (FCS)	JRH Biosciences	12003-500M
GAPDH (14C10) Antibody (37 kDa) - Rabbit	Cell Signalling Technology	2118
GFI1 TaqMan® Gene Expression Assay (Hs003822078_m1)	Applied Biosystems	4331182
GFI136N (rs34631763) TaqMan® SNP Genotyping Assay (C_25596143_10)	Applied Biosystems	4351379
Glycerol	Ajax Finechem	242
Glycine	Sigma-Aldrich	G8898
Glycogen	Roche	901393
Goat Anti-Rabbit IgG – AP conjugate Antibody	Santa Cruz	Sc-2007
GUSB TaqMan® Gene Expression Assay (Hs99999908_m1)	Applied Biosystems	4331182
Hanks Buffered Saline Solution (HBSS)	Sigma-Aldrich	H9394
HEPES 1M	Sigma-Aldrich	H0887
HLA-DR APC / RPE antibody	BD Biosciences	340549 / 347367
Hydrochloric Acid (HCl, 35%)	Fisher Scientific	1100
High Pure DNA Isolation Kit	Roche	11 796 828 001
Isopropanol	Ajax Finechem	425-2.5L PL
Isotype Control Antibody – Mouse IgG1-APC	R&D Systems	IC002A
Isotype Control Antibody – Mouse IgG1-FITC / RPE	Dako Cytomation	X0927 / X0928
Isotype Control Antibody – Mouse IgG1-FITC / RPE	BD Pharmingen	555748 / 555749
Isotype Control Antibody – Mouse IgG2a-FITC / RPE	BD Pharmingen	555573 / 555574
Isotype Control Antibody – Mouse IgM-APC	BD Pharmingen	550883
Isotype Control Antibody – Mouse IgM-FITC	BD Pharmingen	555782
Isotype Control Antibody – Mouse IgG2b-APC / RPE	BD Pharmingen	555745 / 555743
Isotype Control Antibody – Mouse IgG1-PE-Cy7	BD Pharmingen	557872
Kaleidoscope Prestained Protein Standards	BioRad	161-0375
L-Glutamine 200 mM	SAFC Biosciences	59202C-100ML
Lymphoprep	Axis Shield	1114547
Methanol	Chem Supply	MA004-P
Methyl Violet	BDH	34033
miRNeasy RNA Extraction Kit	Qiagen	217004
MLLT4 TaqMan® Gene Expression Assay (Hs00984486_m1)	Applied Biosystems	4331182
Pefabloc (Protease Inhibitor)	Roche	11429869001

Penicillin 5000U/mL Streptomycin 5000µg/mL	Sigma-Aldrich	P4458
Phosphate Buffered Saline (PBS)	SAFC Biosciences	59331C
PhosSTOP (Phosphatase Inhibitor Cocktail Tablets)	Roche	4906845001
Prazosin Hydrochloride (MW = 419.9)	Sigma-Aldrich	P7791
PVDF (Western Blot Membrane)	GE Healthcare	PRN 303F
QuantiTect™ Reverse Transcriptase Kit	Qiagen	205313
Random Hexamer Primer	Geneworks	RP-6
REPLI-g® Mini DNA Amplification Kit	Qiagen	150025
RPMI-1640 Medium w/o L-glutamine	Sigma-Aldrich	R0883
Sodium Azide (NaN_3)	Sigma-Aldrich	S2002
Sodium Chloride (NaCl)	Ajax Finechem	1128
Sodium Deoxycholate	Sigma-Aldrich	D6750
Sodium Dodecyl Sulphate (SDS)	Sigma-Aldrich	L4509
Sodium Fluoride (NaF)	Sigma-Aldrich	S7920
Sodium (ortho) Vanadate (Na_3VO_4)	Sigma-Aldrich	S6508
Superscript II Reverse Transcriptase	Invitrogen	18064-014
TaqMan® Universal Master Mix	Applied Biosystems	4318157
TaqMan® SNP Genotyping Master Mix	Applied Biosystems	4371353
TEMED	Sigma-Aldrich	T7024
Tris (hydroxymethyl) Aminomethane	Merck	1083870500
TRIzol® Reagent	Invitrogen	15596-018
Trypan Blue Solution (0.4%)	Sigma-Aldrich	T8154
Trypsin	SAFC Biosciences	59417C
Tween®20	Sigma-Aldrich	P9416
Western Blot Recycling Kit	Alpha Diagnostic	90100

2.2 Solutions, Buffers & Media

2.2.1 Adherent Cell Line Wash Buffer

HBSS	500 mL
0.53 mM EDTA (1 M)	530 µL
0.25% Trypsin (2.5%)	1 mL

2.2.2 Cell Culture Media

For all cell lines used in this study:

RPMI-1640 Medium	500 mL

To this the following was added:

2 mM L-Glutamine (200 mM)	5 mL
50,000 units/L Penicillin	5 mL
50 mg/mL Streptomycin Sulphate	5 mL
10% FCS	50 mL

Store at 4°C, and preheat to 37°C in a water bath prior to use.

2.2.3 Complete Protease Inhibitor Cocktail (25x stock)

One Complete tablet

2 mL Milli-Q H_2O

The stock was stored at -20°C.

2.2.4 dNTP Set (N = A, C, G, T)

25 mM stock = 40 µL of each dNTP

Working stock: 5 mM = 20 µL of 25 mM stock in 80 µL DEPC water

2.2.5 Flow Cytometry Fixative (FACS Fix)

1 x PBS	500 mL
40% w/v Formaldehyde	5 mL
D-glucose	10 g
0.02% NaN_3	0.1 g

The solution can be stored at 4°C for ≈ 6 months.

2.2.6 Hanks Balanced Salt Solution (HBSS)

Ca^{++} and Mg^{++} free. Add 10 mM HEPES prior to use.

2.2.7 Flow Cytometry Staining Buffer

HBSS (containing 10 mM HEPES)	500 mL

2.2.8 Freeze Mix

70% HBSS

20% FCS

10% DMSO

The solution was made up fresh for each batch of samples to be cryopreserved.

2.2.9 Imatinib Mesylate, MW = 589.72

Imatinib mesylate (imatinib; Glivec; formerly STI-571) was provided by Novartis Pharmaceuticals (Basel, Switzerland). Stock solutions of this compound were prepared at 10 mM with distilled water, sterile filtered and stored at -70°C.

2.2.10 50% [^{14}C]-Imatinib (100 µM)

[^{14}C]-imatinib (1695.72 µM)	29.5 µL
10 mM imatinib	5 µL
RPMI-1640 medium	966 µL

[^{14}C]-imatinib (specific activity 3.394 MBq) was provided by Novartis Pharmaceuticals. Stock solutions were prepared at 1 mg/mL with distilled water and stored at -70°C.

2.2.11 1x Laemmli's Buffer for cell lysis

50 mM Tris-HCl, pH 6.8 (0.2 M)	2.5 mL
10% Glycerol	1 mL
2% SDS	4 mL
5% β-Mercaptoethanol	500 µL
0.05% Bromophenol blue (1%)	250 µL
1 mM Na_3VO_4 (0.1 M)	100 µL
10 mM NaF (0.5 M)	200 µL
H_2O	1.45 mL

Stored in 1 mL aliquots at -20°C, thawed at room temperature or briefly at 100°C in heating block before use.

2.2.12 2x Laemmli's Loading Buffer

2.0 M Tris-HCl, pH 6.8	625 µL
20% Glycerol	2 mL
10% SDS	4 mL
10% β-Mercaptoethanol	1 mL
1% Bromophenol blue	500 µL

H_2O	1.875 mL

Stored in 1 mL aliquots at -20°C, thawed at room temperature or briefly at 100°C in heating block before use.

2.2.13 5x Laemmli's Loading Buffer

1.5 M Tris-HCl, pH 6.8	2 mL
50% Glycerol	5 mL
SDS	1 g
25% β-Mercaptoethanol	2.5 mL
1% Bromophenol blue	500 µL

Stored in 1 mL aliquots at -20°C, thawed at room temperature or briefly at 100°C in heating block before use.

2.2.14 Membrane Blocking Solutions

2.2.14.1 2.5% Skim Milk

Non-fat milk powder	12.5 g
1 x TBST	500 mL

The following proteins require blocking in 2.5% non-fat milk powder: p-Crkl.

2.2.14.2 5% Skim Milk

Non-fat milk powder	25 g
1 x TBST	500 mL

The following proteins require blocking in 5% non-fat milk powder: c-MYC, phospho c-MYC (S62), phospho c-MYC (T58 + S62).

2.2.14.3 5% BSA

BSA	25 g
1 x TBST	500 mL

The following proteins require blocking in 5% BSA: E2F1.

2.2.15 Modified Radioimmunoprecipitation (RIPA) Buffer

50 mM Tris-HCl, pH 7.4	790 mg
1% NP-40 (10%)	10 mL
0.25% Na-deoxycholate (10%)	2.5 mL
150 mM NaCl	900 mg

Dissolve reagents in 75 mL H_2O and then make up to 100 mL total volume. Store RIPA buffer at 2 – 8°C until ready to use. The following protease and phosphatase inhibitors should be added to the solution immediately prior to use:

- Pefabloc (50x stock, 100 mM – use at 1x)
- Complete (25x stock – use at 1x)
- PhosSTOP (10x stock – use at 1x)

2.2.16 Pefabloc Protease Inhibitor (50x stock)

Used at 2 mM from 100 mM stock

100 mM stock = 25 mg/mL Milli-Q H_2O

The stock was stored at -20°C.

2.2.17 PhosSTOP Phosphatase Inhibitor (10x stock)

One PhosSTOP tablet

1 mL Milli-Q H_2O

The stock was stored at -20°C.

2.2.18 Prazosin Hydrochloride – Inhibits OCT-1, OCT-2, OCT-3

Used at 100 µM from 10 mM stock

10 mM stock = 4.2 mg dissolved in 1 mL of methanol

The solution is made fresh on the day of use.

2.2.19 Random Hexamer Primer (100 mg stock)

Working stock: 250 ng/mL = 100 mg in 400 µL DEPC water

2.2.20 1x SDS-PAGE Running Buffer

25 mM Tris-HCl, pH 8.3	3.026 g
192 mM Glycine	14.413 g
0.1% SDS	1 g

Dissolve reagents in 500 mL H_2O and then make up to 1 L.

2.2.21 1x SDS-PAGE Transfer Buffer

25 mM Tris-HCl, pH 8.3	3.026 g
192 mM Glycine	14.413 g
20% Methanol	200 mL

Dissolve reagents in 500 mL H_2O, add 200 mL methanol and then make up to 1 L.

2.2.22 SDS-Polyacrylamide Gel

Table 2.2 Specific for Hoefer SE260 Tank

	Resolving gel (12%)	*Stacking gel (5%)*
H_20	1.9 mL	3 mL
40% Acrylamide	1.75 mL	630 µL
1.5 M Tris Buffer A	1.25 mL	-
1.5 M Tris Buffer B	-	1.26 mL
10% SDS	50 µL	50 µL
10% APS	50 µL	50 µL
TEMED	3 µL	5 µL

Table 2.3 Specific for Hoefer Tall Mighty Small Tank

	Resolving gel (10%)	*Stacking gel (5%)*
H_20	2.4 mL	3 mL
40% Acrylamide	1.25 mL	630 µL
1.5 M Tris Buffer A	1.25 mL	-
1.5 M Tris Buffer B	-	1.26 mL
10% SDS	50 µL	50 µL
10% APS	50 µL	50 µL
TEMED	5 µL	5 µL

2.2.23 10x TBS

20 mM Tris-HCl, pH 7.5	24.2 g
1.5 M NaCl	87.6 g

Dissolve reagents in 750 mL RO H_2O, adjust the pH to 7.5 with HCl and then make to 1L.

2.2.23.1 1x TBS

10x TBS stock	100 mL
RO H_2O	900 mL

2.2.23.2 1x TBST

10x TBS stock	100 mL
0.1% Tween20	1 mL
RO H_2O	900 mL

2.2.24 Thaw Solution

HBSS	500 mL
5% FCS	25 mL
5% ACD	25 mL
10 mM HEPES	5 mL

The solution (without ACD) was stored at 4°C. ACD was added and the solution was heated to 37°C in a water bath prior to use.

2.2.25 Tris Buffer A

Milli-Q® H_2O	500 mL
1.5 M Tris-HCl, pH 8.8	90.855 g

2.2.26 Tris Buffer B

Milli-Q® H_2O	500 mL
0.5 M Tris-HCl, pH 6.8	30.285 g

2.2.27 White Cell Fluid

2% Glacial Acetic Acid	2 mL
Milli-Q® H_2O	98 mL
Methyl Violet	Few crystals

Acetic acid was added slowly to Milli-Q® H_2O. The methyl violet (Gurr®, BDH) was then added, and the solution mixed well to ensure that the crystals dissolved completely. The solution was filtered using a 0.2 µM bottle top filter, and stored at room temperature.

2.3 General Techniques

2.3.1 Lymphoprep Isolation of Mononuclear Cells

Forty-sixty millilitres (40 – 60 mL) of peripheral blood (PB) from patients with CML or normal donors was collected into Lithium Heparin tubes. All samples were collected with informed consent in accordance with the Institutional Ethics approved protocols and with reference to the Declaration of Helsinki. A white cell count was performed using white cell fluid (WCF) and a maximum of 1 x 10^8 cells (maximum of 15 mL of blood) were transferred into a 50 mL polypropylene conical tube. The blood volume was brought to 35 mL using HBSS + 10 mM HEPES and underlain with 15 mL of lymphoprep. Tubes were then centrifuged at 300 *g* for 30 minutes with no brake in a Heraeus multifuge 3S-R benchtop centrifuge (Thermo Fisher Scientific, Waltham, MA, USA). The interface containing the mononuclear cells (MNCs) was then transferred to another 50 mL conical tube and washed once in HBSS. Cells were then counted using WCF.

2.3.2 Cell Counts and Viability

Cell concentration was determined by diluting the cell suspension in WCF and cell viability was assessed by diluting samples with trypan blue solution. Ten microliters (10 µL) of these suspensions were transferred to a haemocytometer counting chamber (Neubauer Improved, Assistant, Germany) and cell concentration and viability calculated accordingly.

2.3.3 Maintenance of Cell Lines

2.3.3.1 Cell Line Specifications

K562 cells originally derived from the pleural effusion of a 53 year old female with blast crisis CML and were obtained from the American Type Tissue Culture Collection (ATCC, Manassas, VA, USA). K562-ABCG2 cells were generated by transfecting K562 cells via electroporation

with a pcDNA3 vector containing full length ABCG2[195] (kindly provided by Prof. Douglas Ross, University of Maryland, Baltimore, MD, USA). The resultant K562-ABCG2 cells were cultured in 500 µg/mL G418 (Geneticin; Invitrogen, Carlsbad, CA, USA) and assessed for appropriate ABCG2 expression by quantitative PCR and flow cytometry. K562-Dox cells stably overexpress ABCB1 following continuous passage of K562 cells in doxorubicin (kindly provided by Prof. Leonie Ashman, University of Newcastle, Callaghan, Australia). KU812 cells were originally derived from the peripheral blood of a 38 year old male with blast crisis CML and were obtained from the ATCC. HeLa cells were originally derived from a 31 year old female with cervical cancer (kindly provided by Dr Bruce Lyons, School of Medicine, University of Tasmania, Hobart, Australia).

2.3.3.2 Culture of General Cell Lines

All appropriate tissue culture techniques were performed in a Class II Biosafety Cabinet (Gelman Sciences, Singapore). Suspension cell lines (K562, K562-ABCG2, K562-Dox, KU812) were maintained at a cell density between $1 \times 10^5 - 1 \times 10^6$ cells/mL in 25cm^2, 75cm^2 or 175cm^2 tissue culture flasks (Greiner). Media was pre-warmed to 37°C prior to use. Cultures were incubated in a 37°C/5% CO_2 incubator. Cell cultures were checked every second day for contamination, counted and re-cultured at the above concentrations. The adherent HeLa cell line was maintained in 75cm^2 tissue culture flasks at a cell density of 1×10^7 cells in 20 mL media. Prior to re-culture, cells were rinsed twice with pre-warmed adherent cell line wash buffer and trypsinised at 37°C for ~10 min with 0.25% trypsin. Cell aggregates were separated by pipetting and resuspended in fresh media devoid of trypsin.

2.3.4 Cryopreservation of Cells

Cells were pelleted at the desired concentration, resuspended in 1 mL of Freeze Mix per 1×10^7 cells and quickly transferred to cryo-ampoules (Nalgene, Thermo Fisher Scientific).

Patient PB-MNCs were frozen to -80°C using a Controlled Rate Freezer (Planer KRYO10 Series II) at a rate of 1°C per min. Samples were stored for up to 10 years in liquid nitrogen (-196°C).

2.3.5 Thawing of Cells

Vials of cells were removed from liquid nitrogen and thawed rapidly by immersion in a 37°C water bath with gentle agitation. In a class II biosafety cabinet, the cell suspension was quickly transferred to a 50 mL conical tube, and approximately 20 mL of thaw solution (pre-warmed to 37°C) was added drop-wise with constant mixing. The sample volume was then increased to 30 mL with thaw solution and the cells pelleted by centrifugation at 300 *g* for 10 minutes in a Heraeus multifuge 3S-R benchtop centrifuge. The supernatant was aspirated and the procedure repeated to remove all residual DMSO.

2.3.6 Antibody Staining for Flow Cytometric Analysis

2.3.6.1 Visualisation of Surface Proteins with Conjugated Antibodies

Patient cells ($2 \times 10^5 – 1 \times 10^6$ cells) were transferred to 5 mL fluorescence-activated cell sorting (FACS) tubes and following suspension in approximately 250 µL of flow cytometry staining buffer, R-phycoerythrin (PE), fluorescein isothiocyanate (FITC), or allophycocyanin (APC) conjugated antibodies were added to tubes as specified (**Appendix 1 and 2**). Cells were incubated for a period of 40 min on ice in the dark, washed twice with buffer and re-suspended in 200 µL FACS fixative. Samples were stored at 4°C in the dark until analysis was performed. Control tubes were stained with appropriate isotype control IgG (RPE, FITC, APC or PE-Cy7) antibodies, which were used to define the gates for positive and negative marker expression. The fluorescence intensity of cell suspensions was examined with a Beckman Coulter Cytomics FC500 flow cytometer (Beckman Coulter, Miami, FL, USA) using CXP

Cytometry List Mode Data Acquistion (Beckman Coulter), with FCS Express 4 software (De Novo Software, Los Angeles, CA, USA) used for post-analysis. Cell populations were analysed based on the forward and side light scattering properties and the fluorescence intensity of appropriate fluorochromes.

2.3.6.2 Visualisation of Surface Proteins with Secondary Antibodies

Patient cells ($2 \times 10^5 – 1 \times 10^6$ cells) were transferred to 5 mL FACS tubes and following suspension in approximately 250 µL of flow cytometry staining buffer, un-conjugated primary antibodies were added to tubes as specified (**Table 2.4**). Cells were incubated for a period of 40 min on ice in the dark, washed once with buffer and then re-suspended in 250 µL buffer. R-phycoerythrin (RPE) conjugated secondary antibody was added to tubes as specified (**Table 2.5**). Cells were incubated for a period of 40 min on ice in the dark, washed twice with buffer and re-suspended in 200 µL FACS fixative. Samples were stored at 4°C in the dark until analysis was performed. Control tubes were stained with isotype control IgG RPE antibodies, which were used to define the gates for positive and negative marker expression. The fluorescence intensity of cell suspensions was examined using an FC500 flow cytometer (Beckman Coulter, Miami, FL, USA), with FCS Express 4 software (De Novo Software, Los Angeles, CA, USA) used for post-analysis. Cell populations were analysed based on the forward and side light scattering properties and the fluorescence intensity of the RPE fluorochrome.

Table 2.4 Primary (unconjugated) antibody staining for surface flow cytometry

Antibody	Cell Number	Volume	Buffer
FZD6	1×10^6	20 µL	HBSS
LRP6	1×10^6	20 µL	HBSS

Table 2.5 Secondary (conjugated) antibody staining for surface flow cytometry

Antibody	Secondary Antibody	Volume	Buffer
FZD6	Anti-goat IgG RPE	20 µL	HBSS
LRP6	Anti-goat IgG RPE	20 µL	HBSS

2.3.7 Protein Concentration Determination using D_c Protein Assay

2.3.7.1 Lysate Preparation with Modified RIPA Buffer

Target cells were spun and washed once with ice-cold 1x PBS. Cell pellet was resuspended in an appropriate amount of ice-cold RIPA lysis buffer (approximately 40 µL buffer/1 x 10^6 cells, but volume can vary between 20 - 40 µL buffer/1 x 10^6 cells depending on whether cell lines or primary cells are used), and incubated on ice for a minimum of 30 min. Following incubation, centrifuge at 4°C for 20 min at 13,400 *g*. Gently remove tubes and place on ice. Carefully transfer the supernatant to a fresh tube kept on ice, discard the cell pellet and now perform the BioRad D_C Protein Assay to determine protein concentration.

2.3.7.2 D_c Protein Assay

Firstly, prepare working Reagent A' by adding 20 µL of Reagent S to 1 mL Reagent A and mix to combine. Using a blank 96-well plate (see **Appendix 3** for plate setup), pipette 5 µL H_2O in two wells (Blank; try to avoid outer wells to reduce contamination risk). Prepare 7 dilutions of the protein standard (BSA), ranging from 0 µg/µL to 2 µg/µL (**Table 2.6**). Pipette 5 µL of each standard, in duplicate, into the STD wells.

Table 2.6 Standard Curve for D_c Protein Assay

Standard Curve (µg/µL)	0.0	0.2	0.4	0.8	1.2	1.6	2.0
H2O (µL)	12.0	10.8	9.6	7.2	4.8	2.4	0.0
2 µg/µL BSA (µL)	0.0	1.2	2.4	4.8	7.2	9.6	12.0

Prepare a 1:5 dilution of each sample using H_2O, and pipette 5 µL of each, in duplicate, onto the plate. Add 25 µL of Reagent A' to each well, followed by 200 µL of Reagent B and cover the plate. Allow to incubate at room temperature for 15 min and then read the absorbance using a 595 nm wavelength (EL808 microplate reader; Bio-Tek Instruments, Inc., Winooski, VT, USA). Note – the absorbance is stable for 1 hr. Calculate the protein concentration for each sample using the standard curve generated.

2.4 Specialised Techniques

2.4.1 Imatinib Intracellular Uptake and Retention ($IUR^{imatinib}$) Assay

The IUR and OCT-1 activity assay were performed as previously described by White *et al.* 2006.[96] All assay points were performed in triplicate. Viable cells (2×10^5) were incubated in 2 mL of RPMI + 10% FCS in the presence or absence of varying concentrations of [^{14}C]-imatinib ranging from 0 – 2 µM. Inhibitors of influx (e.g. prazosin) were added as required. Cells and reagents were incubated for 2 hours at 37°C/5% CO_2. Following incubation the cellular and aqueous phases were separated by centrifugation at 4,300 *g* for 5 min, then pulse spun to 15,700 *g*, for 30 sec. A 20 µL aliquot of supernatant (S/N) from each tube was then added to 100 µL of Microscint-20 in a 96-well flat bottomed plate. The remaining supernatant was aspirated from the tubes and 50 µL of Microscint-20 was added. Tubes were vortexed thoroughly, and pulse spun for 15 sec at 15,700 *g*. Lysed cells were then transferred to a 96-well plate with wells containing 50 µL of Microscint-20. The plate was covered with an adhesive plastic seal and was then counted on a Top Count scintillation counter (Perkin Elmer, Waltham, MA, USA) as counts per minute (cpm). The incorporation of [^{14}C]-labelled imatinib in ng/200,000 cells was then calculated using the formula below:

$$\frac{(\text{cpm cells} - \text{background})}{[(\text{cpm cells} - \text{background}) + (\text{cpm S/N} - \text{background})]} \times \text{ng of } [^{14}\text{C}]\text{-imatinib added}$$

2.4.1.1 OCT-1 Activity

The OCT-1 inhibitor prazosin was used in the $IUR^{imatinib}$ assay at 100 µM. The OCT-1 activity was calculated as the difference between the imatinib IUR in the absence of prazosin and the IUR in the presence of prazosin. Where IUR values in the presence of prazosin were equal or higher than the values in the absence of prazosin, these patients were scored as having negligible (0 ng/200,000 cells) OCT-1 activity.

2.4.2 IC50 Assay, as Determined by Crkl Phosphorylation Status

BCR-ABL1 positive cell lines (2×10^5 cells) and patient MNCs (2×10^6 cells) were incubated in 50 mL tubes for 2 h at 37°C with concentrations of imatinib ranging 0 – 100 µM in 10 mL of RPMI-1640 media + 10% newborn calf serum (NCS). Following incubation cells were washed once with ice-cold 1x PBS, transferred to 1.7 mL microtubes and pelleted by centrifugation for 5 min at 3,900 *g*. All supernatant was removed and cells were lysed in 20 µL of 1x Laemmli's buffer by boiling in a 100°C heat block for 12 min. Lysates were clarified by microfugation and stored at -20°C before resolution by SDS-PAGE.

2.4.3 SDS-PAGE and Western Blot

2.4.3.1 Specific for Crkl and Phosphorylated Crkl (p-Crkl, 39 kDa)

Western blotting for Crkl and p-Crkl was performed as previously described by White *et al*. 2005.[135] Protein lysates (corresponding to 1×10^5 cell lines or 2×10^6 patient MNCs) were resolved on a SDS 12% polyacrylamide gel at 20 mA until the 39 kDa marker was 1.5 cm from the end of the gel, and the protein was electrophoretically transferred to a PVDF membrane (GE Healthcare Life Sciences, Rydalmere, NSW, Australia) at 65 mA overnight in 1x SDS-PAGE Transfer Buffer. The membrane was incubated for 1 h at room temperature with 2.5% membrane blocking solution. The membrane was then probed for 2 h at room temperature with 1:500 anti-Crkl antibody (Santa Cruz Biotechnology, Dallas, TX, USA) in 2.5% membrane blocking solution. Following this, the membrane was rinsed in 1 × TBST buffer, and then washed for 3 × 5 minutes with 1 × TBST buffer. The membrane was then incubated with 1:2000 alkaline-phosphatase (AP) conjugated anti-rabbit immunoglobulin (Santa Cruz Biotechnology) in 2.5% blocking solution for 1 hour at room temperature. The membrane was then rinsed in 1 × TBST buffer, washed for 3 × 5 minutes in 1 × TBST buffer and then washed for 3 × 5 minutes in 1 × TBS buffer. Bound antibodies were detected with ECF

attaphos substrate (Amersham Biosciences, Sunnyvale, CA, USA) by FluorImager analysis (Molecular Dynamics, Amersham Biosciences).

2.4.3.2 Specific for Low Molecular Weight Proteins (< 100 kDa)

Western blotting for the low molecular weight proteins c-MYC, phospho c-MYC (S62), phospho c-MYC (T58 + S62), E2F1 and GAPDH was performed using an adapted version of the protocol described above. Briefly, lysates from either cell lines (2×10^5 cells) or patient MNCs (2×10^6 cells) were made as described in **Section 2.3.7** and resolved on a SDS 10% polyacrylamide gel at 20 mA until the dye-front reached the end of the gel, and the protein was electrophoretically transferred to a PVDF membrane at 200 mA for 2 h. Membranes were then incubated in blocking solution for 1 h at room temperature before being probed with primary antibody as per **Table 2.7**. Following overnight incubation at 4°C, membranes were rinsed in 1 × TBST buffer, and then washed for 3 × 5 min with 1 × TBST buffer. The membrane was then incubated with secondary antibody as per **Table 2.7** for 1 h at room temperature. Following incubation, the membrane was rinsed in 1 × TBST buffer, washed for 3 × 5 min with 1 × TBST buffer and then washed in 1 × TBS buffer for 3 × 5 min. Bound antibodies were detected with ECF attaphos substrate by FluorImager analysis.

Table 2.7 Incubation conditions for lower molecular weight western blot

1° Antibody	Diluent	Conc.	2° Antibody	Diluent	Conc.
c-MYC	5% skim milk powder / TBST	1:1000	α-rabbit-AP	5% skim milk powder / TBST	1:2000
Phospho c-MYC (S62)	5% skim milk powder / TBST	1:500	α-rabbit-AP	5% skim milk powder / TBST	1:2000
Phospho c-MYC (T58 + S62)	5% skim milk powder / TBST	1:10,000	α-rabbit-AP	5% skim milk powder / TBST	1:2000
E2F1	5% BSA / TBST	1:1000	α-rabbit-AP	5% BSA / TBST	1:2000
GAPDH	5% skim milk powder / TBST	1:1000	α-rabbit-AP	5% skim milk powder / TBST	1:2000

2.4.3.3 Densitometric Analysis using ImageQuant Software

Phosphorylated and non-phophorylated Crkl bands were quantified using ImageQuant software (Molecular Dynamics) and p-Crkl, as a percentage (%) of the total Crkl protein, was calculated. The average signal for each protein band was determined, with the background signal from a blank section of membrane deducted. Bands of p-Crkl were measured and graphed as a percentage of the intensity of the total Crkl (i.e. p-Crkl + Crkl = 100%). The percentages were then normalised with 0 nM imatinib representing 100% p-Crkl and the highest concentration of imatinib (e.g. 100 µM imatinib) representing 0% p-Crkl. The IC50 value was then taken as the concentration of imatinib when normalised p-Crkl reaches 50%, using the following formula:

$$IC50 = \%\ \mathrm{p-Crkl\ at\ 0\ \mu M} - [\frac{(\%\ \mathrm{p-Crkl\ at\ 0\ \mu M} - \%\ \mathrm{p-Crkl\ at\ 100\ \mu M})}{2}]$$

2.4.3.4 Stripping of Western Blots for Re-probing

For best results, the blots required for stripping should be prepared immediately after their first probe. Dilute the 10x Antibody Stripping Solution from the Western Blot Recycling Kit (Alpha Diagnostic, San Antonio, TX, USA) with reagent quality water in a 1:10 ratio. Prepare enough solution to allow free movement of blots during incubation. Incubate blots with stripping solution for 15 min at room temperature, with gentle mixing. Wash the blot twice with 1x Blocking Buffer from the Western Blot Recycling Kit for 5 min each wash. The blot can now be re-probed with antibody.

2.4.4 Real Time Quantitative PCR (RQ-PCR)

2.4.4.1 TRIzol® RNA Extraction

Target cells ($1 \times 10^6 – 1 \times 10^7$ cells) were lysed in 1 mL of TRIzol® reagent (Invitrogen) and incubated at room temperature for 5 min. Two-hundred microliters (200 µL) of chloroform was then added, tubes were shaken vigorously for 15 sec and incubated at room

temperature for 2 – 3 min. Following this, the tubes were centrifuged at 12,000 *g* for 15 min at 4°C. The aqueous phase (top layer) was transferred to a fresh RNase/DNase free 1.5 mL tube. RNA was precipitated by the addition of 1 µL glycogen (20 µg) and 500 µL isopropanol. The samples were gently mixed and incubated at room temperature for 10 min. RNA was pelleted by centrifugation at 12,000 *g* for 10 min at 4°C. Supernatant was removed from the RNA pellet, and the pellet washed in 1 mL of 75% ethanol before being centrifuged at 7500 *g* for 5 min at 4°C. Supernatant was removed again and tubes centrifuged briefly to remove any excess ethanol. The RNA pellet was briefly dried, then dissolved in 20 µL DEPC water (depending on size of pellet) at 55°C for 10 min. RNA was quantified using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific) and samples diluted with DEPC water to achieve the desired concentration of between 250 ng – 1 µg/µL where possible and stored at -70°C.

2.4.4.2 miRNeasy™ RNA Extraction

The miRNeasy™ RNA Extraction kit (Qiagen, Chadstone Centre, VIC, Australia) was used to isolate RNA specifically for use with high-throughput assays, such as gene expression microarrays and global DNA methylation arrays. Target cells ($1 \times 10^6 – 1 \times 10^7$ cells) were lysed in 1 mL of TRIzol® reagent (Invitrogen) and homogenized by vortexing for 1 min. Tubes were then incubated at room temperature for 5 min to promote dissociation of nucleoprotein complexes. Next, 200 µL chloroform was added, tubes were shaken vigorously for 15 sec and incubated at room temperature for 2 – 3 min. Following this time the tubes were centrifuged at 12,000 *g* for 15 min at 4°C. The aqueous phase (top layer; ~ 500 µL) was transferred to a new collection tube and 1.5 volumes (~ 750 µL) of 100% ethanol was added and mixed thoroughly by pipetting several times. 700 µL of the sample was then pipetted into an RNeasy Mini spin column (Qiagen) in a 2 mL collection tube and centrifuged at 8,000 *g* for 15 sec at room temperature. The flow through was then discarded and repeated again

for the remainder of the sample. 700 µL of Buffer RWT was then added and centrifuged at 8,000 *g* for 15 sec to wash the column. Next, 500 µL Buffer RPE was pipetted onto the column and centrifuged at 8,000 *g* for 15 sec. Once complete, the flow through was discarded and the process repeated; however, this time centrifuging at 8,000 *g* for 2 min, to dry the column membrane. The RNeasy spin column was then placed into a new 2 mL collection tube and centrifuged at full speed (16,100 *g*) for 1 min to eliminate any residual Buffer RPE or flow through. The RNeasy spin column was then transferred to a new 1.5 mL collection tube, with 35 µL of RNase-free water pipetted directly onto the column membrane and then centrifuged at 8,000 *g* for 1 min. RNA was then quantified using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific) and samples stored at -70°C.

2.4.4.3 QuantiTect™ cDNA Synthesis

The QuantiTect™ Reverse Transcription kit (Qiagen) was used for fast cDNA synthesis. Briefly, 500 ng of RNA was used to generate cDNA. The genomic DNA elimination reaction was prepared on ice, according to **Table 2.8**.

Table 2.8 Genomic DNA elimination reaction components

Component	Volume (Per Reaction)
gDNA Wipeout Buffer, 7x	2 µL
Template RNA	Variable
RNase-free water	Variable
Total reaction volume	14 µL

All components were assembled and mixed by briefly vortexing the tubes. Tubes were then incubated for 2 min at 42°C, and then placed immediately on ice. The reverse-transcription master mix was then prepared on ice, according to **Table 2.9**.

Table 2.9 Reverse-transcription reaction components

Component	Volume (Per Reaction)
Quantiscript Reverse Transcriptase	1 µL
Quantiscript RT Buffer, 5x	4 µL
RT Primer mix	1 µL
Total reaction volume	6 µL

All components were mixed and kept on ice. The reverse-transcription mix was then added to the 14 µL of template RNA already prepared, mixed and stored on ice. All samples were then incubated for 15 min at 42°C, followed by 3 min incubation at 95°C to inactivate the Quantiscript Reverse Transcriptase. Once complete, all samples were placed on ice and stored at -20°C until required for PCR. The final concentration of cDNA is 25 ng/µL, if 500 ng RNA was initially used.

2.4.4.4 RQ-PCR for BCR-ABL1 mRNA Expression

Previously designed primers and probes (see **Table 2.10**)[196] for either the b3a2 or b2a2 BCR-ABL1 mRNA transcripts contained in patient cells were used. Probes were labelled with a FAM™ dye on the 5' end and a minor groove binder (MGB) non-fluorescent quencher on the 3' end. Standards for each transcript were prepared and analysed in every batch. Master mixes were made using the TaqMan® Universal PCR Master Mix (Applied Biosystems, Mulgrave, VIC, Australia) and prepared as per **Table 2.11**.

Table 2.10 Primer sequences for BCR-ABL1 PCR

Primer Name	Sequence 5' to 3'
Genes of interest	
B3A2 Forward	GGG CTC TAT GGG TTT CTG AAT G
B3A2 Reverse	CGC TGA AGG GCT TTT GAA CT
B3A2 probe	CAT CGT CCA CTC AGC CAC TGG ATT TAA GC
B2A2 Forward	ATC CGT GGA GCT GCA GAT G
B2A2 Reverse	CGC TGA AGG GCT TCT TCC TT
B2A2 probe	CCA ACT CGT GTG TGA AAC TCC AGA CTG TCC

Reference genes	
BCR Forward	CCT TCG ACG TCA ATA ACA AGG AT
BCR Reverse	CCT GCG ATG GCG TTC AC
BCR Probe	TCC ATC TCG CTC ATC ATC ACC GAC A

Table 2.11 Preparation of PCR master mix

Reagent	Volume (per sample)
TaqMan® Universal PCR Master Mix	12.5 µL
DEPC H_2O	9.55 µL
Forward Primer (50 µM)	0.1 µL
Reverse Primer (50 µM)	0.1 µL
TaqMan® Probe (10 µM)	0.25 µL

The prepared master mix (22.5 µL) was then pipetted into the appropriate wells of a 96-well PCR plate, before adding 2.5 µL of DNA standards (see Setup sheet in **Appendix 4**) to the appropriate wells. Standard wells were capped before the addition of 2.5 µL sample DNA to the corresponding wells. HeLa cells were used as a negative control and were included in every experiment along with a no template control (NTC). High and low quality control samples, representing two different levels of *BCR-ABL1* expression, were also included to ensure consistency between assays. The PCR was performed in the ABI Prism 7500 Sequence Detection System (Applied Biosystems) with the following cycler conditions:

Hold 1 @ 50°C, 2 minutes

Hold 2 @ 95°C, 10 minutes

Cycling (45 repeats) Step 1 @ 95°C, hold 15 seconds

Step 2 @ 60°C, hold 60 seconds

After completion of the PCR amplification, wells were checked for evaporation and the plate discarded. Results were analysed using the supplied ABI Prism 7500 software (Applied Biosystems).

2.4.5 TaqMan® Low Density Array (TLDA)

2.4.5.1 Preparing the Sample-Specific PCR Mix

TLDA arrays contain 384-wells pre-loaded with TaqMan® Gene Expression assays (Applied Biosystems), generally in a custom-designed layout. Relative levels of gene expression are determined from the fluorescence data generated during PCR. For full details on the TLDA formats and target genes used in each array, see **Appendices 2.5** and **2.6**. The arrays were carried out according to the manufacturer's instructions (see Applied Biosystems TaqMan® Low Density Array User Bulletin for more information, http://www3.appliedbiosystems.com/cms/groups/mcb_support/documents/generaldocuments/cms_042326.pdf). Briefly, cDNA was synthesized from 500 ng RNA as described in **Section 2.4.4.3**. Each TLDA array contains 8 fill reservoirs (ports), allowing a maximum of 8 samples to be loaded. Depending on the gene format of the array, 1 to 8 different samples can be loaded. For each sample, the components listed in **Table 2.12** were added to a labelled 1.5 mL tube, capped and gently vortexed to thoroughly mix the solution. The tubes were then briefly pulse-spun to eliminate any air bubbles from the mixture.

Table 2.12 Preparation of TLDA sample mix

Component	Volume per Port
cDNA sample (500 ng) + RNase-free water	50 µL
TaqMan® Universal PCR Master Mix (2x)	50 µL
TOTAL Volume	**100 µL**

Component (e.g. 1 patient, 4 ports)	Total Master Mix
cDNA sample (500 ng)	18 µL
RNase-free water	207 µL
TaqMan® Universal PCR Master Mix (2x)	225 µL
TOTAL Volume	**450 µL***

*Load 100 µL per port, therefore 100 ng RNA is loaded in each port.

2.4.5.2 Loading the PCR Reaction Mix into Fill Reservoirs

Allow the refrigerated TLDA card to equilibrate to room temperature for approximately 10 min. Once the original packaging (plastic tube) has reached room temperature, carefully remove the TLDA card, and place in the class II biosafety cabinet, foil side down. Then load 100 µL of the desired sample-specific PCR reaction mix into each port, ensuring that the PCR reaction mix sweeps in and around to fill the reservoir. Inspect all 8 ports for consistent sample loading. Next, the TLDA card was placed into the specific, custom-made array holder and loaded into the specific Heraeus Multifuge 3SR Plus centrifuge bucket, ensuring that each bucket is fully loaded and that the "This Side Out" label faces the front of the bucket. Program the centrifuge as follows: Bucket type – 15679, up ramp rate – 9, down ramp rate – 9 and speed – 300 *g*. Two consecutive, 1 min spins were then performed to ensure complete distribution of the reaction mix.

2.4.5.3 Sealing the TLDA Card and Performing PCR

The microfluidic card sealer isolates the wells of a TLDA card after it is loaded with cDNA samples, using a precise stylus assembly (carriage) to seal the main fluid distribution channels of the array. Position the microfluidic card sealer on a sturdy bench, with the front end (starting position) closest to you and the carriage in the starting position. Insert the TLDA card into the sealer and gently push until the card is held securely in place by the spring clips and is seated securely. Push the carriage across the base of the sealer in the direction of the arrows, using a slow, steady and deliberate motion, until the carriage reaches the mechanical stops. Remove the sealed TLDA card and inspect for proper sealing. Using scissors, trim the fill reservoirs from the TLDA card, using the edge of the card as a guide. The TLDA card is now ready to be analysed on the ABI 7900HT Real-time PCR machine (Applied Biosystems), with the following cycler conditions:

Hold 1 @ 50°C, 2 minutes

Hold 2 @ 94.5°C, 10 minutes

Cycling (40 repeats) Step 1 @ 97°C, hold 30 seconds

Step 2 @ 59.7°C, hold 60 seconds

Results were analysed using SDS 2.3, RQ Manager and DataAssist[197] 2.0 software (Applied Biosystems), with gene expression calculated using Delta Ct (ΔCt) values, normalised against the selected reference gene (either 18S, ACTB, B2M, GUSB, HMBS, HPRT1, or PES1).

2.4.6 TaqMan® RQ-PCR for Specific Genes

RQ-PCR for specific genes of interest was performed as follows. TaqMan® Gene Expression assays containing validated primers and probe, specific for the target gene, were ordered from Applied Biosystems (**Table 2.13**). Briefly, cDNA was synthesized from 500 ng RNA as described in **Section 2.4.4.3**. The PCR reaction mix was then prepared as per **Table 2.14** for each target/control gene.

Table 2.13 TaqMan® Gene Expression Assays used for PCR

TaqMan® Assay ID	Gene	Amplicon Length	Assay Design
Hs00984486_m1	MLLT4	107	Probe spans exons
Hs00382207_m1	GFI1	61	Probe spans exons
Hs99999908_m1	GUSB	81	Probe spans exons

Table 2.14 Preparation of TaqMan® RQ-PCR Reaction Mix

PCR Reaction Mix Component	Total Master Mix
20x TaqMan® Gene Expression Assay	1 µL
2x TaqMan® Gene Expression Master Mix	10 µL
cDNA template (100 ng)	4 µL
RNase-free water	5 µL
TOTAL Volume	**20 µL**

20 µL of PCR reaction mix was transferred into each well of a 96-well reaction plate (Applied Biosystems), and each well sealed with optical cap strips (MicroAmp, Applied Biosystems). The plate was then centrifuged briefly (4,000 *g*, 15 sec) and loaded into the ABI 7900HT Real-

time PCR machine (Applied Biosystems) to be amplified with the following cycler conditions:

Hold 1 @ 50°C, 2 minutes

Hold 2 @ 95°C, 10 minutes

Cycling (40 repeats) Step 1 @ 95°C, hold 15 seconds

Step 2 @ 60°C, hold 60 seconds

Results were analysed using SDS 2.3 and RQ Manager software (Applied Biosystems) and gene expression was calculated using Delta Ct (ΔCt) values, normalised against the reference gene (GUSB).

2.4.7 DNA Extraction

The High Pure PCR Template Preparation Kit (Roche Diagnostics, Castle Hill, NSW, Australia) was used to isolate genomic DNA from various cell lines and patient samples. Briefly, 2×10^6 – 1 x 10^7 cells were pelleted and all but 200 µL of supernatant aspirated. Cells were resuspended in the remaining 200 µL and transferred to a new, labelled 1.7 mL tube. Next, 200 µL of Binding Buffer and 40 µL Proteinase K were added, and the tubes vortexed. Samples were then incubated for at least 30 min at 72°C. Then 100 µL isopropanol was added, the tube vortexed and the mixture pipetted into a HighPure filter in a collection tube. Following centrifugation at 5,900 *g* for 1 min, the flow-through was discarded and 500 µL of Inhibitor Removal Buffer added before centrifugation was repeated. The filter was then transferred to a clean collection tube and washed twice with 500 µL Wash Buffer via centrifugation at 5,900 *g* for 1 min. The filter was transferred to a clean collection tube and dried by centrifugation at 15,700 *g* for 1 min. The filter was then transferred to a second clean collection tube, 100 µL pre-warmed Elution Buffer added and incubated at room temperature for 5 min. Tubes were centrifuged at 5,900 *g* for 1 min to collect DNA eluent. The DNA concentration and quality was measured using a NanoDrop ND-1000

Spectrophotometer (Thermo Fisher Scientific) and stored at 4°C (short-term storage) or -70°C (long-term storage).

2.4.8 REPLI-g® Whole Genome Amplification

The REPLI-g® Mini Kit (Qiagen) was used to amplify genomic DNA from patient samples. Briefly, prepare sufficient Buffer D1 (denaturation buffer) and Buffer N1 (neutralization buffer) for the total number of whole genome amplification reactions (**Table 2.15**). Five microlitres (5 µL) of genomic DNA was placed into a microcentrifuge tube (total amount should be > 10 ng), and 5 µL Buffer D1 was added, mixed by vortexing and pulse-spun briefly. Samples were then incubated at room temperature (15 – 25°C) for 3 min. Ten microlitres (10 µL) of Buffer N1 was added to each tube, mixed by vortexing and pulse-spun briefly.

Table 2.15 Preparation of Buffer D1 and Buffer N1 for REPLI-g® amplification

Component	Buffer D1	Buffer N1
Reconstituted Buffer DLB	9 µL	-
Stop solution	-	12 µL
Nuclease-free water	32 µL	68 µL
TOTAL Volume	**41 µL**	**80 µL**

A master mix was then prepared as per **Table 2.16**, mixed and pulse-spun briefly. Next, 30 µL of master mix was added to the 20 µL of denatured DNA for each sample prepared earlier, and incubated at 30°C for 10 – 16 hrs. Inactivation of the REPLI-g Mini DNA Polymerase was performed by heating all samples for 3 min at 65°C. A 50 µL REPLI-g reaction typically yields approximately 10 µg of DNA regardless of the amount of input genomic DNA. Samples were then stored at 4°C (short-term storage) or -20°C (long-term storage).

Table 2.16 Preparation of REPLI-g® Master Mix for genomic DNA

Component	Volume per reaction
REPLI-g Mini Reaction Buffer	29 µL
REPLI-g Mini DNA Polymerase	1 µL
TOTAL Volume	**30 µL**

2.4.9 NPM1 Mutation Fragment Analysis

DNA was extracted from the samples of interest as described in **Section 2.4.7**. One set of primers (**Table 2.17**) were used to screen for the most common *NPM1* mutations (4 nucleotide insertions in nt959 or 965 of exon 12) giving a product size of 232 bp for wild-type *NPM1* or 236 bp for mutant *NPM1*.[198]

Table 2.17 *NPM1* Primer Sequences for PCR

Primer Name	Oligonucleotide Sequence (5' to 3')	Ordered From
NPMF	VIC® - AGGACAGCCAGATATCAACTGTTAC	Applied Biosystems
NPMR	AGTTAACTCTCTGGTGGTAGAATGAAA	Geneworks

Seventy-five nanograms (75 ng) of DNA were required at 15 ng/µL concentration for the DNA PCR. PCR reaction mix was set-up as per **Table 2.18** for each sample, as well as Positive (*NPM1* mutant), Negative (*NPM1* wild-type) and H_2O controls.

Table 2.18 Preparation of *NPM1* TaqGold DNA-PCR Reaction Mix

PCR Reaction Mix Component	Final Conc.	Volume per Reaction
dNTP 10 mM	0.2 mM	1 µL
$MgCl_2$ 25 mM	1.5 mM	3 µL
10x Buffer	-	5 µL
NPMF-VIC 50 µM	0.5 µM	0.5 µL
NPMR 50 µM	0.5 µM	0.5 µL
TaqGold Polymerase	1.25 U	0.3 µL
dH_2O		34.7 µL
TOTAL Volume		**45 µL**

Forty-five microlitres (45 µL) of reaction mix was pipetted into each reaction tube. In the UV cabinet, 5 µL DNA (plus H_2O to a final concentration of 15 ng/µL) was added to each tube and mixed well. Tubes were then placed in a thermal cycler (Eppendorf Mastercycler; Eppendorf, North Ryde, NSW, Australia) and amplified according to the following cycler conditions:

Hold 1 @ 95°C, 10 minutes

Cycling (40 repeats) Step 1 @ 94°C, hold 30 seconds

Step 2 @ 59°C, hold 30 seconds

Step 3 @ 72°C, hold 30 seconds

Hold 2 @ 72°C, 10 minutes

Hold 3 @ 4°C, ∞

After completion of the amplification, PCR products were visualised on a 2% agarose gel (stained with ethidium bromide) to confirm success of reaction. Once products of the correct size were confirmed, the samples were sent to the Sequencing Centre at SA Pathology (Ground Floor, Department of Molecular Pathology, Frome Rd, Adelaide) for Fragment Analysis. Results were obtained as a printout from the GeneMapper ID v3.2 software (Applied Biosystems), with either a single peak at approx. 232 bp indicating wild-type *NPM1*, or a double peak at approx. 232 and 236 bp for mutated *NPM1*.

2.4.10 Global DNA Methylation Analysis

2.4.10.1 Illumina Infinium HumanMethylation450 array

Previously isolated genomic DNA (500 ng; **Section 2.4.7**) was subjected to sodium bisulfite conversion using the EZ DNA Methylation Kit (Zymo Research, Orange, CA, USA), according to the manufacturer's recommendations. Bisulfite DNA quality and concentration were assessed using the NanoDrop ND-1000 (Thermo Fisher Scientific). Following the Illumina

Infinium HD Methylation protocol,[199] 4 µL of bisulfite converted DNA sample was whole-genome amplified followed by enzymatic end-point fragmentation, precipitation and resuspension. The resuspended samples were then hybridized to the Illumina Infinium HumanMethylation450 (450K) BeadChips (Illumina, San Diego, CA, USA) at 48°C for 16 hrs. A wash was performed to remove any unhybridized and non-specifically hybridized DNA, and then single nucleotide extension was performed. The nucleotides incorporated were labeled with biotin (ddCTP and ddGTP) and 2,4-dinitrophenol (DNP; ddATP and ddTTP). After extension was complete, repeated rounds of staining with a combination of antibodies that differentiated DNP and biotin were performed. Finally the BeadChip was washed once more and was ready for scanning.

The 450K array is unique, as two different assay types are used to measure DNA methylation: Infinium I (type I probes) and Infinium II (type II probes). These probes are bound to beads scattered throughout the array and fluoresce when a probe successfully binds to DNA, producing a signal that is read by the Illumina scanner. Infinium I assays use two bead types for each CpG of interest: an unmethylated (U) and a methylated (M) bead, which each have a different probe design (ProbeA [U] and ProbeB [M]). This allows both type I probes to fluoresce in the same colour channel for each CpG (either red [Cy5] or green [Cy3]). However, the Infinium II assays use a single bead type for each CpG (M + U bead). Therefore, one probe is designed for each type II target CpG and the fluorescence colour produced is based on the nucleotide incorporated in the extension step. Hence, incorporation of an A or T corresponds to an unmethylated site in red (U), while the incorporation of a C or a G corresponds to a methylated site in green (M).[199]

The Illumina iScan was used to scan the BeadChips with a two-color channel to detect Cy3 labelled probes on the green channel and Cy5 labelled probes on the red channel. The images were extracted using GenomeStudio (v2011.1) Methylation module (1.9.0) software (Illumina). GenomeStudio was used to normalise data using different internal controls that are present on the HumanMethylation450 BeadChip. The raw intensity data (IDAT) files were then imported into the *R* statistical program (http://www.r-project.org/; version 3.0.1)[200] and processed using the *minfi* (v1.4.0) package[201] obtained from the Bioconductor software project (http://www.bioconductor.org/).[202] M-values were then generated using *minfi*, as this value has been demonstrated to be valid for statistical analyses.[203] Following probe type correction using subset-quantile within array normalisation (SWAN),[204] the M-values were then converted to β-values using the equation $\beta = (2M/[2M + 1])$. β-values represent a number ranging from 0 to 1, where 0 represents non-methylated and 1 represents completely methylated. Further detail on the procedure used to analyse the 450K methylation data, and the specific selection criteria used to identify differentially methylated CpGs is described in **Section 2.6.10.**

2.4.11 Bisulphite Pyrosequencing of Target CpGs

2.4.11.1 Assay Design

All pyrosequencing assays were designed and performed by the Australian Genome Research Facility (AGRF; Perth, WA, Australia) using the algorithms built into the PyroMark Assay Design Software (Version 2.0.1, Qiagen). Briefly, approximately 500bp of sequence surrounding each target CpG site were input into the software's Methylation Analysis module using default parameters adjusted to force primer design to avoid variable positions. Primers designed to target each CpG site were chosen from a list generated by the software. PCR and sequencing primers for each assay are listed in **Table 2.19**. All primers were

supplied by Integrated DNA Technologies (San Diego, CA, USA). All biotin labelled primers were HPLC purified.

Table 2.19 Primer Sequences for Pyrosequencing Assays

Sequence Name	Sequence	Tm
cg00448707-F	TTT TGA TTT TTT TGT GGG GTT TGA	53
cg00448707-RB	/5Biosg/CCA ACT TTA ACT CCA AAA AAC AAA AAC TA	54
cg00448707-S	GGA GGA TAA GTA GGG	43
cg01663953-FB	/5Biosg/AAG GGT TTA ATA AGT AAT TTA ATG AAA G	49
cg01663953-R	CCT ATA ACC CCT AAA CAA ATT CCT	52
cg01663953-S	ATA AAA CTC AAA ATT TAA CTC TTC	46
cg03343083-F	GAA ATA TTG TTA GAG TAA GGA TGG GTA	52
cg03343083-RB	/5Biosg/CAA TTA CCT ATT TTA CTT TCC TTC CA	51
cg03343083-S	TGT TAG AGT AAG GAT GGG TAA	50
cg04524088-F	GGT TTT TTG ATT ATT GTT GAG AGG AAT AGT	54
cg04524088-RB	/5Biosg/TTT CTC TTC TCC CCA AAA CTC TTT TCT TC	57
cg04524088-S	GGG TTT AGT TGG TAT TAA GAA	47
cg06173889-F	TGT GTT GGA AGT TTG TAT TTT GAT AA	52
cg06173889-RB	/5Biosg/CCC CTA AAT ATA AAA TTC TCT CCT ACC	53
cg06173889-S	GGT GAG GTT GTT GTG	47
cg08191073-F	GTT TTT TTG GAA ATT GGG GAT GG	53
cg08191073-RB	/5Biosg/CTA AAA ATA AAC ACA CCC TTC AAA TAC	51
cg08191073-S	YGT TGG AGT TTT GTA	41
cg10987071-F	AAG GGG GAT AAG AAT GTG TT	51
cg10987071-RB	/5Biosg/ACA TCC ATT CCT AAT ATT ACT ATA CTT ACC	52

cg10987071-S	GAT AAG AAT GTG TTT TAT AGG T	46
cg16334795-FB	/5Biosg/AGT TGG AGG AGG TAT AAA GAA T	51
cg16334795-R	CTA CTA TAA TCC CTA CCC TTC TCC	53
cg16334795-S	AAC TAT CTA ATT AAT ACT C	38
cg17137424-F	GGG ATG GAT GGT ATT TGA GTT	52
cg17137424-RB	/5Biosg/ATA ACC ATC TAC CAA CTA T	44
cg17137424-S	GGT TGT TGG TTT TGG	45
cg17146918-FB	/5Biosg/GGT GGA GTG GGG TTG TAT G	55
cg17146918-R	CTC CCT CCA CTC CCA ACA AAT T	57
cg17146918-S	CCA ACA AAT TCC TTT CCA A	48
cg17394237-F	GAG AGG GGT TAG ATG TTA GGT A	53
cg17394237-RB	/5Biosg/CCT TTA AAC TCT CAA TTT CTA AAT TAC C	51
cg17394237-S	AGA TGT TAG GTA TTT GTG A	45
cg21995068-F	TGT AGT ATT ATG TTT TAG TTT GGA TTA TAG	50
cg21995068-RB	/5Biosg/CAT TTT AAC CAA ACT AAT CTC AAA CTC	51
cg21995068-S	AAT TGT AGA AAT AGG TTA GG	44
cg23825092-F	AGG AAT TTT TTG AGT TTG TAA TTA GAG TA	52
cg23825092-RB	/5Biosg/TTC CCC TTC ATA CAA ACC CTC TC	56
cg23825092-S	GTT TGT AAT TAG AGT ATA AGT TGT	47
cg27053299-F	GAT TGG AGG ATT GAA AAG TAG G	51
cg27053299-RB	/5Biosg/TCC CAA ACC CCA ATA CCC	55
cg27053299-S	GAT TAG GGA GTT GTA T	41

2.4.11.2 Bisulphite Conversion of Genomic DNA and PCR

DNA samples were converted using the Epitect Bisulphite Conversion Kit (Qiagen). 500 ng of genomic DNA was converted overnight in 140 µL total volume using the standard protocol from the kit. Converted DNA was isolated on provided columns and stored at -20°C. All PCR amplifications were performed with the PyroMark PCR Kit (Qiagen). Amplification reactions were set-up in tubes as per **Table 2.20**.

Table 2.20 Pyrosequencing PCR Reaction Mix

Reaction Mix Component	Total Master Mix
PyroMark Master Mix	12.5 µL
Coral Load	2.5 µL
Forward Primer 5 µM	1 µL
Reverse Primer 5 µM	1 µL
Bisulphite-converted Template DNA	2 µL
H_2O	6 µL
TOTAL Volume	**25 µL**

Tubes were then placed in a thermal cycler and amplification performed according to the following cycler conditions:

Hold 1 @ 95°C, 15 minutes

Cycling (45 repeats) Step 1 @ 95°C, hold 30 seconds

Step 2 @ 56°C, hold 30 seconds

Step 3 @ 72°C, hold 30 seconds

Hold 2 @ 72°C, 10 minutes

Hold 3 @ 4°C, ∞

After completion of the amplification, PCR products were visualised on a 2% agarose gel (stained with ethidium bromide) to confirm success of reaction.

2.4.11.3 Pyrosequencing

Biotin-labelled amplicons were bound to Streptavidin Sepharose High Performance beads (GE Healthcare Life Sciences). The beads containing the immobilized PCR products were denatured and washed using proprietary solutions (Qiagen) on the Pyrosequencing Vacuum Prep Tool (Qiagen) to isolate a single stranded template, as per the manufacturer's recommendations. Single stranded templates were then transferred to an optically clear, 24 well sequencing plate in 0.3 µM of pyrosequencing primer. Annealing of the sequencing primer to the template was done by heating the plate to 80°C followed by cooling to room temperature. Pyrosequencing was performed on a PyroMark 24 Pyrosequencing System (Qiagen) as per the manufacturer's instructions. Data was analysed on the PyroMark Q24 software to give the percentage methylation values for each CpG site in the sample.

2.4.12 Single Nucleotide Polymorphism (SNP) Genotyping

2.4.12.1 Sequenom MassARRAY® Custom SNP Genotyping

Custom SNP genotyping was performed on the Sequenom MassARRAY® system (Sequenom, San Diego, CA, USA) and utilises a homogenous MassExtend (hME – single base extension) reaction termed iPLEX Gold, allowing analysis of 6 SNP's in a single multiplex reaction. SNP assays were designed and peformed by the AGRF as recommended by the manufacturer with reagents included in the iPLEX Gold SNP genotyping kit (Sequenom) and the software and equipment provided with the MassARRAY platform (Sequenom).

2.4.12.1.1 MassARRAY® Assay Design

All SNP assays were designed using the algorithms built into the MassARRAY Assay Design Suite 1.0 (Sequenom). Briefly, sequence information for the SNPs of interest was input into the software using default parameters and primers for both PCR and MassEXTEND® primers for multiplexed assays were designed automatically. The primers selected for each assay are available from the AGRF. All primers were supplied by Integrated DNA Technologies.

2.4.12.1.2 MassARRAY® PCR Amplification

The target sequences were simultaneously amplified from DNA extracted from the samples of interest as described in **Section 2.4.7**. DNA was resuspended to a final volume of 50 ng/µL, before an appropriate amount was added to the 5 µL final PCR volume (**Table 2.21**).

Table 2.21 MassARRAY® PCR Reaction Mix

Reaction Mix Component	Concentration in 5 µL	Total Master Mix
H_2O	NA	1.850 µL
PCR Buffer with $MgCl_2$ (10x)	1.25x	0.625 µL
$MgCl_2$ (25 mM)	1.625 mM	0.325 µL
dNTP mix (25 mM)	500 µM	0.100 µL
Primer mix (500 nM each)	100 nM	1.000 µL
Genomic DNA (5 – 10 ng/µL)	5 – 10 ng/rxn	1.000 µL
Hotstar Taq® DNA Polymerase (5 U/µL)	0.5 U/rxn	0.100 µL
TOTAL Volume		**5.000 µL**

Tubes were gently vortexed to mix samples and then placed in a thermal cycler, with amplification performed according to the following cycler conditions:

Hold 1 @ 94°C, 15 minutes

Cycling (45 repeats) Step 1 @ 94°C, hold 20 seconds

Step 2 @ 56°C, hold 30 seconds

Step 3 @ 72°C, hold 60 seconds

Hold 2 @ 72°C, 3 minutes

Hold 3 @ 4°C, ∞

After completion of the amplification, to neutralize unincorporated dNTPs, PCR products were treated with shrimp alkaline phosphatase (SAP) according to **Table 2.22**.

Table 2.22 MassARRAY® SAP Mix

Reaction Mix Component	Total Master Mix
H_2O	1.53 µL
10x SAP Buffer	0.17 µL
SAP enzyme (1 U/µL)	0.30 µL
TOTAL Volume	**2.00 µL**

The SAP mix (2 µL) was added to each 5 µL PCR reaction, gently mixed by vortexing and incubated in a thermal cycler according to the following cycler conditions:

Hold 1 @ 37°C, 20 minutes

Hold 2 @ 85°C, 5 minutes

Hold 3 @ 4°C, ∞

2.4.12.1.3 MassARRAY® iPLEX Gold Reaction

The iPLEX Gold primer extension reaction was prepared as per **Table 2.23,** vortexed to mix and briefly centrifuged.

Table 2.23 MassARRAY® iPLEX Gold Reaction Mix

Reaction Mix Component	Concentration in 9 µL total	Total Master Mix
H_2O	NA	0.755 µL
iPLEX Buffer (10x)	0.222x	0.200 µL
iPLEX termination mix	1x	0.200 µL
Primer mix (7 µM; 14µM*)	0.625 µM; 1.25 µM	0.804 µL
iPLEX enzyme	1x	0.041 µL
TOTAL Volume		**2.000 µL**

* 7 µM and 14 µM illustrate the doubled concentration of the high mass primers relative to the low mass primers. Low mass primers should be at 0.625 µM and high mass primers at 1.25 µM in the final 9 µL reaction.

Tubes were then placed in a thermal cycler, with amplification performed according to the following cycler conditions:

Hold 1 @ 94°C, 30 seconds

Cycling (40 repeats) Step 1 @ 94°C, hold 5 seconds

Cycling (5 repeats) Step 1 @ 52°C, hold 5 seconds

Step 3 @ 80°C, hold 5 seconds

Hold 2 @ 72°C, 3 minutes

Hold 3 @ 4°C, ∞

2.4.12.1.4 Dispensing to SpectroCHIP® Bioarrays and MALDI-TOF analysis

After desalting of the products using SpectroCLEAN resin (Sequenom) following the manufacturer's protocol, cleaned extension products were dispensed onto a 384-element SpectroCHIP array (Sequenom) using a nanodispenser, and finally, the array was introduced into the Sequenom MassARRAY® mass spectrometer. Spectra were acquired using SpectroAcquire software (v3.3.1; Sequenom), and data analysis, including automated allele calling, was done using MassARRAY Typer software (v3.3.0; Sequenom).

2.4.12.2 GFI1^{36N} TaqMan® SNP Genotyping Assay

The TaqMan® SNP Genotyping assay for the GFI1^{36N} SNP (rs34631763) was ordered from Applied Biosystems (Assay ID: C_25596143_10). Briefly, DNA was extracted from the samples of interest as described in **Section 2.4.7.** The SNP genotyping assay was diluted from 40x to a 20x working stock with 1x TE buffer. SNP genotyping mix was prepared as per **Table 2.24**, vortexed to mix and briefly centrifuged.

Table 2.24 Preparation of TaqMan® SNP Genotyping Reaction Mix

SNP Genotyping Mix Component	Total Master Mix
20x TaqMan® Genotyping Assay mix	1.25 µL
2x TaqMan® Genotyping Master Mix	12.5 µL
TOTAL Volume	**13.75 µL**

13.75 µL of SNP genotyping reaction mix was transferred into each well of a 96-well reaction plate (Applied Biosystems), and 10 ng DNA was added in a total volume of 11.25 µL (made up with DNase-free, RNase-free water) before each well was sealed with optical cap strips (MicroAmp, Applied Biosystems). The plate was then centrifuged briefly and loaded into the ABI 7900HT Real-time PCR machine (Applied Biosystems) to be amplified with the following cycler conditions:

Hold @ 95°C, 10 minutes	
Cycling (40 repeats)	Step 1 @ 95°C, hold 15 seconds
	Step 2 @ 60°C, hold 60 seconds

Results were analysed using the Allelic Discrimination plot in SDS 2.3 software (Applied Biosystems).

2.4.13 Affymetrix Microarray Gene Expression analysis

2.4.13.1 Sample Processing and Microarray Hybridisation

Briefly, total RNA was isolated as per **Section 2.4.4.2** using the manufacturer's protocols. RNA quality and integrity was validated using the Experion automated electrophoresis station (Bio-Rad, Hercules, CA, USA). cDNA was labeled with Cy5 or Cy3 dyes using the Cy-Scribe post-labeling kit (Amersham Biosciences) before microarray hybridisation was performed at the Adelaide Microarray Centre using the Affymetrix Human Gene ST 1.0 GeneChip array (Affymetrix, Santa Clara, CA, USA) as per manufacturer's recommended protocols. The arrays were then analysed on an Affymetrix TG3000 system.

2.4.13.2 Data analysis

The raw CEL files were subjected to Robust Multi-array Average (RMA) normalisation as implemented in *aroma.affymetrix* package. *Aroma.affymetrix* was obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/)

and was implemented in the *R* statistical program (http://www.r-project.org/; version 2.13.0).[200] Quality control was assessed based on NUSE plots. Differential gene expression was determined using linear models (LIMMA)[205] incorporating quality weighting and adjusted for multiple testing using the Benjamini-Hochberg (BH) method.[206] Differentially expressed genes were selected based on a fold change greater than 1.5 and a FDR *P*-value < 0.35.

2.4.14 Illumina Microarray Gene Expression analysis

2.4.14.1 Sample Processing and Microarray Hybridisation

Briefly, total RNA was isolated as per **Section 2.4.4.2** using the manufacturer's protocols. RNA quality and integrity was validated using the Agilent 2100 bioanalyzer (Agilent Technologies, Santa Clara, CA). cDNA was labeled with Cy5 or Cy3 dyes using the Cy-Scribe post-labeling kit (Amersham Biosciences) before microarray hybridisation was performed at the AGRF using the Illumina HumanHT-12 v4 Expression BeadChip (Illumina) as per manufacturer's recommended protocols. The raw signal intensities generated for each sample were then output using Illumina BeadStudio software by the AGRF. Explanation of the data analysis procedure used (LIMMA) is described in **Section 2.6.12.** Differentially expressed genes were selected based on a *P*-value < 0.05.

2.5 Statistical Analyses

Unless specifically stated, all figures were constructed using GraphPad Prism 6© software (GraphPad Software Inc., La Jolla, CA, USA) and Microsoft Excel 2010 software. Column graphs represent the mean plus the standard error of the mean (SEM) except where specified. Box plots display the median value, the upper 25th and lower 75th percentiles and whiskers encompass the 10th and 90th percentiles, with points outside this range depicted as individual dots. All statistical analyses were performed either using GraphPad Prism 6® or the statistical environment *R*, as specified. Normality tests were performed on each data set using a Kolmogorov-Smirnov test. The Levene Median Test was used to assess for equal variance when appropriate. The Mann-Whitney Rank Sum or the Student's *t*-test were used to determine differences between experimental groups. Where the data sets passed the normality and equal variance tests a *t*-test was applied, where the data failed either test a Mann-Whitney Rank Sum was used. Differences were considered to be statistically significant when the probability value (*P*-value) was < 0.05. Further details on the specific bioinformatics analyses utilised are described in each chapter as appropriate.

2.6 Bioinformatics Analyses

2.6.1 MeV

MultiExperiment Viewer (MeV; v4.8.1) is a Java based, versatile microarray data analysis tool, which is a part of the TM4 software package,[207] freely available online from http://www.tm4.org/mev.html. MeV incorporates sophisticated algorithms for clustering, visualization, classification, statistical analysis and biological theme discovery. MeV was used to perform unsupervised hierarchical clustering, using the HCL (Hierarchical clustering) algorithm.[208] Briefly, the dataset of interest was loaded into the MeV interface, and samples were labelled according to any subtypes of interest (i.e. OA, response etc.). Each dataset (e.g. immunophenotypic cell surface marker expression) was adjusted by the median expression value for each particular marker/gene across samples, allowing markers/genes either up- or down-regulated in relation to the median expression level to be identified. Hierarchical clustering was then performed, including optimization for both the marker/gene and sample orders, using either Pearson Correlation or Euclidean Distance as the distance metric, and either average or complete as the linkage method. A hierarchical clustering tree was then generated by MeV.

2.6.2 IterativeBMA

The iterative Bayesian Model Averaging (BMA) algorithm is a variable gene selection and classification algorithm, which is an extension of the BMA algorithm originally developed by Yeung *et al.*[209] IterativeBMA was successfully adapted for use with both immunophenotypic cell surface marker expression data and DNA methylation data. In classification, the aim is to build a classifier using a training set that will be able to predict the classes of an independent test set with high accuracy; however, this is difficult to perform when the number of genes/markers is much greater than the number of samples to be classified, which is very

common for microarray or high-throughput PCR datasets. BMA is an excellent algorithm to combat this problem, as it devises sets of predictive genes called "models". BMA is then able to take model uncertainty into consideration (i.e. there will generally be many models that fit the data well) by computing the weighted average of the posterior probabilities that a test sample belongs to a given class, based on multiple "good" models.[210] The weights produced are proportional to the goodness of fit of the model.[210] The iterativeBMA algorithm is an extension of BMA, primarily by the addition of an initial ranking step, where univariate measures are used to place the genes in order. The ratio of between-group to within-group sum of squares (BSS/WSS)[211] is used for this initial pre-processing step, and identifies genes/markers with large BSS/WSS ratios (i.e. genes/markers with relatively large variation between classes (e.g. OA groups), but small variation within classes), assigning high rankings, before the traditional BMA algorithm is applied. IterativeBMA was obtained from the Bioconductor software project (http://www.bioconductor.org/)[202] and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200]

To briefly summarize the approach when analysing the immunophenotyping data, iterativeBMA was used to rank the cell surface markers associated with OA (very low versus very high) initially using the BSS/WSS univariate measure. High rankings were assigned to the markers with large BSS/WSS ratios. The traditional BMA algorithm was then applied to the top 30 ranked markers, with the "leaps and bounds" algorithm and Occam window method, initiated. Then markers assigned low posterior probabilities (< 5%) of being in the predictive model by BMA were removed. Suppose *N* markers are removed. The next *N* markers from the BSS/WSS ordered rank are added back to the set of markers, so that a window of 30 markers is consistently maintained, and the BMA algorithm was applied again.

These steps of marker swaps and iterative applications of BMA are continued until all markers have been subsequently considered.

IterativeBMA was used to predict the probabilities that a given sample was either very low OA or very high OA, based on the cell surface marker expression (immunophenotying) of the markers populating the model. Class 0 was used to represent very high OA, and class 1 represented very low OA. In the classification algorithm, the true class labels of the test samples were unknown, allowing the true class labels of the test set to be compared to the labels predicted by the algorithm, generating the number of classification errors. For a test sample assigned to very low OA, a predicted probability close to 1 is more desirable than slightly above 0.5, whereas the opposite is true for very high OA patients. The Brier score[212] is equal to the sum of squares of the difference between the true class and the predicted probability over all samples, and when the predicted probabilities are restricted to within 0 to 1, it is equal to the number of classification errors. The Brier score was used to compare the performance of the classification models developed using iterativeBMA. This same analysis method was applied to the DNA methylation data.

2.6.3 Prediction Analysis for Microarrays (PAM)

The Prediction Analysis for Microarrays (PAM) for the R package (PAMR) algorithm contains functions for performing sample classification using the nearest shrunken centroids method. Originally developed by Tibshirani *et al.*[213,214] for use with gene expression data, but able to be adapted for multiple data types, PAM provides a simple, fast and accurate classifier, generating results that are able to be easily interpreted by a biologist. PAM was successfully adapted for use with both the immunophenotypic cell surface marker expression dataset and the DNA methylation dataset. The nearest shrunken centroids method is based on the

nearest centroid classifier,[215] but with the addition of a simple modification where the class centroids are shrunk towards the overall centroids, after being standardised by the within-class standard deviation for each marker.[213] This allows markers with stable expression in the same sample classes to be given greater weight in the classifier. PAMR was obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/) and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200]

Briefly, the *k*-nearest neighbour (KNN) impute function was first applied in PAMR, as it fills in any missing data from the dataset. For each marker with missing values, the *k*-nearest neighbour was found using a Euclidean metric, confined to the columns for which that marker was NOT missing data. Each candidate neighbour may be missing some of the coordinates used to calculate the distance metric, in which case the average distance from the non-missing coordinates was calculated. Having found the *k*-nearest neighbours for a particular marker, the missing elements were then imputed by averaging those (non-missing) elements of its neighbours. However, this can fail if ALL the neighbours are missing in a particular element, in which case the overall column mean for that block of markers was used. Once the dataset was complete with no missing values, PAMR was used to train a predictive classifier using the nearest shrunken centroid method. The classifier was then cross-validated using default settings and the error curves were plotted. A confusion matrix, providing a table of true versus predicted value for each sample using the classifier, was computed allowing the threshold, which provides the classifier model with the best accuracy to be identified. Cross-validated class probabilities were then plotted using this threshold, and the most significant markers were identified and plotted, before the false discovery rate was estimated and plotted also. Finally, to ensure that the best classifier was selected, the

threshold was adaptively scaled against 10 different iterations, allowing the best threshold to be identified and selected. Subsequently, cross-validated class probabilities were generated again using this adjusted threshold and the most accurate classifier selected.

2.6.4 Gene set enrichment analysis (GSEA) and Pathway/Gene ontology analysis

GSEA (http://broadinstitute.org/gsea/) was performed as described by Subramanian *et al*.[216] Briefly, GSEA was performed for functionally related genes across a spectrum of C2 curated gene-sets from the Molecular Signatures Database (MSigDB), including BioCarta gene sets (217 gene sets), KEGG gene sets (186 gene sets), and Reactome gene sets (430 gene sets). The complete gene list derived from the microarray results (21,014 genes) was used to construct a file for GSEA. The pre-ranked gene list, ranked according to the moderated t-statistics generated using LIMMA,[205] when compared to the two groups of interest (e.g. very low OA vs. very high OA) was used as the input file for GSEA. Gene-sets with FDR *q*-value < 0.25 after 1,000 permutation cycles were considered significantly enriched. Lists of leading edge subset genes, the cores of gene sets that account for the enrichment signal, were then generated. Protein and/or gene networks and ontology associated with a gene-set of interest were derived using either Ingenuity Pathway Analysis (IPA; Ingenuity Systems, http://www.ingenuity.com) or the GeneGo Metacore™ (Thomson Reuters, http://portal.genego.com) integrated software suite.

The MSigDB is a collection of annotated gene-sets for use with the GSEA software.[216,217] Freely available (http://www.broadinstitute.org/msigdb/), the MSigDB was used to investigate overlaps between a gene-set of interest and published gene-sets. Briefly, the list of genes of interest (e.g. histone gene list) was loaded into the MSigDB interface and was compared against the C2 curated gene-sets from online pathway databases, publications in

PubMed, and knowledge of domain experts. Gene-sets with an FDR *P*-value < 0.05 were considered significantly enriched.

2.6.5 Microarray dataset re-analysis

The Chambers *et al*.[218] dataset investigating the gene expression profiles of mouse haematopoietic stem cells (HSCs) and their differentiated progeny; including erythrocytes, granulocytes, monocytes, NK cells, activated and naïve T-cells, and B-cells, was downloaded from the NCBI Gene Expression Omnibus (GEO; accession number GSE6506). The raw data was normalised using the GC-RMA algorithm, which is available as a part of the open-source Bioconductor project (http://www.bioconductor.org/)[202] within the statistical programming language R (http://cran.r-project.org/).[200] Identification of differential gene expression was performed using empirical Bayes moderated t-statistics, and the Bioconductor package LIMMA.[205] The false-discovery rate (FDR) was controlled for multiple testing using the Benjamini-Hochberg (BH) method. Only genes with a BH-adjusted FDR *P*-value < 0.05 were identified as statistically significant.[206]

2.6.6 Selection of housekeeping genes

RefFinder (http://www.leonxie.com/referencegene.php/)[219] was used to determine the stabilities of candidate reference genes for RQ-PCR studies. RefFinder is a user-friendly web-based tool that integrates currently available major computational programs, including geNorm,[220] Normfinder,[221] BestKeeper,[222] and the comparative ΔCt method,[223] to compare and rank the stability of candidate reference genes (all four methods are reviewed in Chao *et al.*[224]). Cycle threshold (Ct) values, calculated for each candidate reference gene, were used to rank each gene in each program, assigning an appropriate weight to each individual gene. The geometric mean of these weights was then calculated, determining the final ranking and

the relative expression stability of each gene. The reference genes included in the various RQ-PCR studies were *18S*, *ACTB, B2M*, *GUSB*, *HMBS*, *HPRT* and *PES1.*

2.6.7 HTqPCR

The high-throughput quantitative PCR (HTqPCR)[225] package is designed for the analysis of Ct values from RQ-PCR data, with particular emphasis of data generated using the TLDA platform. HTqPCR was obtained from the Bioconductor software project (http://www.bioconductor.org/)[202] and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200] The general workflow implemented in HTqPCR is demonstrated in **Figure 2.1,** and the vignette available from Bioconductor was followed, with adaptations made as necessary for each dataset. Firstly, raw TLDA RQ-PCR Ct data was imported into HTqPCR and a general overview of the average Ct values was obtained. Concordance between duplicate gene measurements for each patient sample was assessed, with samples marked if the duplicate Ct values differed by more than 20% of the average for that feature. Variation between samples was also assessed in a similar manner. Reliability of the RQ-PCR data was determined, with Ct values flagged as "Undetermined" if above a selected threshold, or "Unreliable" if they were so low as to be estimated to be problematic. All flagged Ct values were removed from further analysis. Normalisation was performed using the deltaCt normalisation method, where the mean of the chosen reference gene was subtracted from all other values in the gene-set. HTqPCR has four other normalisation methods that can also be interrogated if required: quantile, norm.rankinvariant, scale.rankinvariant and geometric.mean. Next, the overall correlation between the samples was assessed, as well as the range of Ct values and coefficients of variation (CV) for each gene. Hierarchical clustering and principle component analysis (PCA) was performed to display the similarities and differences between the patient samples. Finally, significant

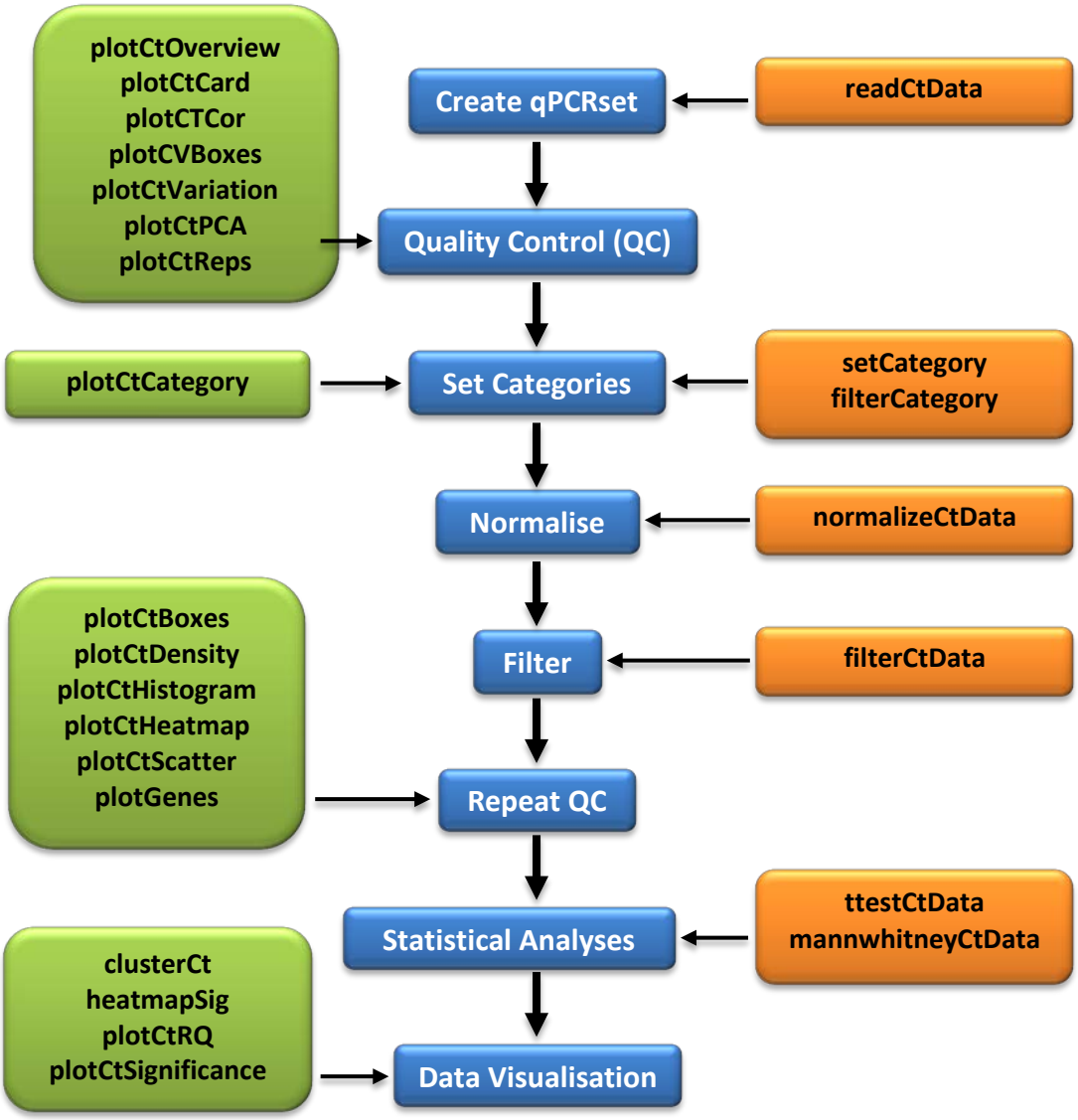


Figure 2.1 HTqPCR workflow for the analysis of RQ-PCR data

The centre column (**blue**) highlights the main procedural steps used in the HTqPCR analysis. The left column (**green**) provides examples of the visualisation functions used in HTqPCR, while the right column (**orange**) demonstrates the data analysis functions. The quality control (QC) functions are performed both before and after data normalisation, in order to examine the quality of the data, or identify trends. Figure adapted from Dvinge *et al.*[225] to demonstrate the HTqPCR analysis method utilised in this thesis.

differential expression was tested between samples using the standard *t*-test, with both t- and p-statistics generated, along with fold changes and information about whether the Ct values used in the calculations were "OK" or "Unreliable".

2.6.8 CMA package

The CMA (Classification for Microarrays) package[226,227] contains a comprehensive assortment of classification algorithms, designed for microarray-based studies, which can be used with high-throughput RQ-PCR datasets, including those generated from TLDA analysis. CMA is presented as a uniform, user-friendly interface containing a total of 21 different classification methods (**Table 2.25**). The user-friendliness of CMA allows uniform data inputs across the different methods, and outputs that are highly self-explicable and informative. Probability estimations for predicted observations are provided for the majority of classifiers, and learning sets are automatically generated. Comparison of the performance of several classifiers can be made using one or several of the following performance measures: misclassification rate, sensitivity and specificity, empirical area under the curve (AUC), Brier score or average probability of correct classification. The CMA package was obtained from the Bioconductor software project (http://www.bioconductor.org/)[202] and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200]

CMA was then used to perform a bundle of 6 different classification methods based on the principle of discriminant analysis: diagonal-, linear- and quadratic discriminant analysis (DLDA, LDA and QDA, respectively), discriminant analysis by Fisher (FDA), shrunken centroids discriminant analysis (SCDA) and Partial Least Squares followed by linear discriminant analysis (pls_LDA). Briefly, three-fold cross-validation, repeated five times in order to generate stable results, was initially performed to generate multiple training sets containing

Table 2.25 Classification algorithms available in the CMA package

Method Name	CMA Function Name	Package
Componentwise Boosting	compBoostCMA	CMA
Diagonal Discriminant Analysis	dldaCMA	CMA
Elastic Net	ElasticNetCMA	Glmpath
Fisher's Discriminant Analysis	fdaCMA	CMA
Flexible Discriminant Analysis	flexdaCMA	Mgcv
Tree-based Boosting	gbmCMA	Gbm
k-nearest neighbours	knnCMA	Class
Linear Discriminant Analysis *	ldaCMA	MASS
Lasso	LassoCMA	Glmpath
Feed-Forward Neural Networks	nnetCMA	Nnet
Probalistic nearest neighbours	pknnCMA	CMA
Penalized Logistic Regression	plrCMA	CMA
Partial Least Squares + *	pls_ldaCMA	Plsgenomics
****** logistic regression***	pls_lrCMA	Plsgenomics
****** Random Forest***	pls_rfCMA	Plsgenomics
Probabilistic Neural Networks	pnnCMA	CMA
Quadratic Discriminant Analysis *	qdaCMA	MASS
Random Forest	rfCMA	randomForest
PAM	scdaCMA	CMA
Shrinkage Discriminant Analysis	shrinkldaCMA	CMA
Support Vector Machine	svmCMA	E1071

Abbreviations: PAM – Prediction Analysis for Microarrays. * indicates methods that can be combined with the Partial Least Squares method. Table adapted from Slawski *et al.*[226]

two-thirds of the patients (n = 27), allowing 13 patients to be used as the test set. As a preliminary step to classification, variable selection was performed for the methods requiring it (linear and quadratic discriminant analysis) using the *t*-test, Kruskal-Wallis test, LIMMA and random forest test. DLDA was the first classification method performed, as it does not require variable selection or tuning. Next, the variables (genes) were ranked according to the t statistic, as a basis for variable selection necessary for LDA, QDA and FDA analyses, which were subsequently performed using default parameters. SCDA was then executed, which requires hyper-parameter tuning for the shrinkage intensity. This was performed using either the pre-specified grid by setting the "tuninglist" to an empty list, or by setting the shrinkage intensity to delta = 0.01. Finally, pls_LDA was run with default settings, and all six methods were compared and ranked by the following variables: classification, Brier score and average probability.

2.6.9 Transcription factor prediction using microarray gene-sets

To predict transcription factor regulation using gene-sets identified from the microarray and TLDA analyses, three different web-based predictive tools were used; TFactS,[228] Pscan[229] and ChEA.[230] TFactS is a web-based tool developed by the de Duve Institute MEXP Unit, to predict the transcription factors that were respectively up- or down-regulated, inhibited or activated in a biological condition based on lists of up-regulated and down-regulated genes from transcriptomics experiments.[228] Briefly for TFactS, all the significantly up- and down-regulated histone genes (FDR *P*-value < 0.05) were uploaded to the TFactS online tool (http://www.tfacts.org/) and the generated results were then exported to Microsoft Excel. Default settings (as per author's recommendations[228]) were used and transcription factors were identified as statistically significant (either regulated, activated or inhibited) if the BH-adjusted FDR *P*-value was < 0.05.

Pscan is also a web-based software tool for identifying over-represented transcription factor binding site motifs in sequences from gene-lists.[229] It is able to scan a set of sequences with motifs matching the binding specificity of known transcription factors, and identifies the motifs which are significantly positively or negatively enriched, providing insight into the transcription factors most likely to be common regulators of the initial gene-set. Briefly for Pscan, gene-sets containing the gene IDs of interest were uploaded to the Pscan web interface (http://www.beaconlab.it/pscan/), the promoter region to investigate, with respect to the transcription start sites (TSSs) of the genes, was selected (-450 to +50) and finally the database to use was chosen (either JASPAR[231] or TRANSFAC[232]). Once selected, the run is computed and the output shows the ranking of the profiles selected according to their *z*-test *P*-value as a downloadable text file.

ChIP enrichment analysis (ChEA) is another transcription factor enrichment tool, which is able to generate lists of transcription factors whose targets are over-represented in the gene-set of interest, against a manual database. ChEA utilises ChIP-seq data from published studies (database contains 189,933 interactions, manually extracted from 87 publications, describing the binding of 92 transcription factors to 31,932 target genes) to identify target enrichment.[230] Briefly for ChEA, the gene symbols matching the histone genes from the gene-set of interest were entered into the ChEA interface (http://amp.pharm.mssm.edu/lib/chea.jsp/), the experiments to analyse against were filtered for Human species and the enrichment analysis was performed. Once completed, the results were sorted by *P*-value, with only those transcriptions factors with $P < 0.05$ identified as significantly enriched. To compute statistical enrichment, the Fisher exact test with Bonferroni's correction was implemented. The proportions for the test included the number of genes in the input list, the number of genes identified in the ChIP-X experiment,

the genes shared between the two lists and the number of overall targets in the ChIP-X database (approximately 30,000).[230] Finally, the result file was exported to Microsoft Excel.

2.6.10 Minfi

The *minfi* package provides tools for the analysis of Illumina methylation arrays, with a special focus on the new Illumina Infinium HumanMethylation450 (450K) arrays. Originally developed by Hansen, K.D. and Aryee, M.,[201] *minfi* was successfully used to perform pre-processing, quality control (QC) assessment, identification of interesting methylation loci and plotting functionality. The Illumina 450K arrays have a complicated design. Each sample is measured on a single array, using two different colour channels (red and green) to provide a methylated measurement and an "un-methylated" measurement. The complexity arises by the way these values are measured, as the array contains probes with either a "Type I" or "Type II" design. CpGs measured using a Type I design are measured using a single colour, with two different probes in the same colour channel providing the methylated and unmethylated measurements. However, CpGs measured using a Type II design use a single probe and two different colours to provide the methylated and unmethylated measurements. The Type I probes were originally developed for the Illumina Infinium 27K methylations arrays, but when the 450K arrays were designed, both Type I and II probes were included, as Type II probes use less space on the BeadChip allowing the number of total probes to expand to approximately 480,000. *Minfi* was obtained from the Bioconductor software project (http://www.bioconductor.org/)[202] and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200]

To briefly summarise the approach, *minfi* was used to analyse the IDAT files from the 450K results, which contain summarized bead information. These files represent two different

colour channels prior to normalisation. The IDAT files are read into *minfi*, along with a sample sheet containing information about the samples analysed on the 450K arrays. QC was then performed, with several plots useful for identifying samples with data quality problems produced, including density plots and control probe plots, which allow individual control probes to be plotted (e.g. for bisulphite conversion, hybridisation, negative controls etc.). A QC report is generated summarising all these results. Samples that failed QC were then removed from further analysis (6/60 samples failed QC, all due to low bisulphite conversion rates; **Appendix 8 and 9**). Pre-processing (normalisation) was then performed, using background subtraction and control normalisation, as per the Illumina recommendations in Genome Studio (Illumina supplied software). Multi-dimensional scaling (MDS) plots were then generated using the top 1000 variable CpGs to obtain a quick estimate of the relationships between samples. MDS plots are a useful way to identify any outliers in the data.

The Beta-value (β-value) is a quantitative measure of DNA methylation levels of specific CpGs, and ranges from 0 for completely unmethylated to 1 for completely methylated. β-values were obtained using Illumina's formula, where M = methylated allele and U = unmethylated allele:

$$\beta = \frac{\mathrm{M}}{\mathrm{M} + \mathrm{U} + 100}$$

Next, SWAN (subset-quantile within array normalisation)[204] was performed. SWAN was used to substantially reduce the technical variability between the Type I and Type II assay designs used within the 450K array, whilst maintaining the important biological differences. These technical differences can result in aberrant β-value distributions, which are corrected by SWAN, producing a smoother overall β-value distribution. The normalised data was then

used to identify differentially methylated CpGs using "M-values", which were computed by logit(β), as recommended by Du *et al.*[203] The dmpFinder function in *minfi* was used to perform this analysis, using an F-test to identify CpGs differentially methylated between the groups of interest, with the selection criteria outlined below. A customised version of the dmpFinder function (termed "dmpScan") was also used to generate the log fold change values. Manhattan plots were generated using the *CpGassoc* package,[233] obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/), while heatmaps were generated using Z-scores and the heatmap function, available in the *stats* package from *R*. Z-score values were calculated from M-values in *R* using the following formula, where the raw value is the original M-value:

$$Z = \frac{\text{Raw value} - \text{Mean}}{\text{Standard Deviation}}$$

2.6.10.1 Selection Criteria 1A

The first selection criteria used identified CpGs as differentially methylated between the groups of interest if the Benjamini-Hochberg adjusted false discovery rate (FDR)[206] *P*-value < 0.05.

2.6.10.2 Selection Criteria 1B

The second selection criteria increased the stringency from Criteria 1A (FDR *P*-value < 0.05), as CpGs also required a log fold change (logFC) > 2 between the groups of interest, to be identified as significant.

2.6.10.3 Selection Criteria 1C

The third selection criteria increased the stringency from Criteria 1A (FDR *P*-value < 0.05) further again, but without using logFC, as any CpGs containing a single nucleotide polymorphism (SNP) within 10 bp of the CpG locus were excluded, while CpGs also required a mean difference in methylation level (Δβ) ≥ 0.17, similar to Laffaire *et al*,[234] to be identified as significantly differentially methylated.

2.6.10.4 Selection Criteria 2A

Batch effects were firstly removed using ComBat as described in **Section 2.6.11** All CpGs with a detection *P*-value < 0.01, or mapped to either chromosome X or Y were removed from the analysis, leaving 386,958 CpGs available for final analysis. CpGs were identified as differentially methylated between the groups of interest if the Benjamini-Hochberg adjusted false discovery rate (FDR)[206] *P*-value < 0.05 and did not contain a SNP within 10 bp of the CpG locus.

2.6.10.5 Selection Criteria 2B

This selection criteria increased the stringency from Criteria 2A (detection *P*-value < 0.01 and chromosome X and Y exclusion, plus FDR *P*-value < 0.05 and no SNP within 10 bp of CpG locus) as significant CpGs also required a mean difference in methylation level ($\Delta\beta$) ≥ 0.17, to be identified as differentially methylated.

2.6.11 ComBat

High-throughput methylation analyses are susceptible to various technical artefacts, similar to microarray gene expression analyses, including batch effects.[235] The ComBat (Combining Batches) algorithm adjusts for known batch effects using an Empirical Bayesian framework originally developed by Johnson *et al*.[236] and is available for use as part of the Surrogate Variable Analysis (SVA) package.[237] The principle of ComBat involves a regression analysis, which allows variation to be removed from the data, as information is pooled across genes, "shrinking" the batch effect parameter toward the overall mean of the batch estimates. Importantly, ComBat can be successfully adapted for use with 450K methylation data and Empirical Bayes correction has been previously demonstrated to effectively remove batch effects from Illumina Infinium methylation data.[235] ComBat was obtained from the *SVA* package as a part of the Bioconductor software project (http://www.bioconductor.org/)[202]

and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200]

To briefly summarise the approach, ComBat was used to remove batch effects from the background corrected, pre-processed and SWAN normalised 450K methylation data (refer **Section 2.6.10**). The batch effect to be removed was the difference in array processing dates (i.e. between the original 450K analysis and the new 450K analysis, known as the covariate), while the variable of interest (e.g. difference in OA groups) was preserved. Once the covariate and the variable of interest were identified, the ComBat function was applied to the data, using the Empirical Bayesian adjustments, producing an expression matrix with the same dimensions as the original dataset. Significance analysis was then performed with *minfi* using this batch-adjusted dataset.

2.6.12 Linear Models for Microarray Data analysis

The Linear Models for Microarray data (LIMMA) analysis package is designed for the analysis of gene expression data arising from microarray technologies. Originally developed by Smyth, G.K.,[205] LIMMA is based on the use of linear models to assess differential expression in the context of multifactor designed experiments, allowing multiple comparisons between many RNA targets to be analysed simultaneously. LIMMA was obtained from the Bioconductor software project (http://www.bioconductor.org/)[202] and was implemented in the *R* statistical program (http://www.r-project.org/; version 3.0.1).[200]

To briefly summarise the approach, LIMMA was used to analyse the expression data from the Illumina HumanHT-12 v4 Expression BeadChips. Firstly, the raw summary probe expression profiles (including regular probe profile and control probe profile), originally

obtained from the Illumina BeadStudio software, were imported into LIMMA. A target file, containing sample and probe annotation was also imported. Boxplots were then generated for both the regular and negative control probes to determine that all samples were successfully analysed. The proportion of expressed probes for each array was then determined. Pre-processing, including background correction, quantile normalisation and $\log_2$ transformation of the raw data, was then performed using the neqc pre-processing strategy developed by Shi *et al.*[238] Probes not expressed in all samples were removed from the analysis before linear models were fitted to each probe. Differential expression analysis was performed, and a list of differentially expressed probes generated. This list of significant probes was then used for further downstream analyses.

2.6.13 Oncomine analysis

Oncomine is a cancer-profiling database containing published microarray datasets that have been collected, standardised, annotated and analysed by Compendia Bioscience, aimed at facilitating discovery from gene-expression profiling.[239,240] Oncomine was accessed online (http://www.oncomine.org/) and used to search for histone gene enrichment in published microarray datasets.

3 IMMUNOPHENOTYPING OF CP-CML PATIENTS WITH LOW AND HIGH OCT-1 ACTIVITY AT DIAGNOSIS

3.1 Introduction

Imatinib is the current paradigm for targeted therapy in chronic phase chronic myeloid leukaemia (CP-CML), but has little impact on subsequent phases.[92,241] Although most patients will achieve excellent clinical (haematological, cytogenetic and molecular) responses on imatinib, it is clear that a subset of patients will respond poorly, or fail imatinib therapy. Currently, up to 35% of patients treated with imatinib fit into this subset, displaying either primary or acquired resistance, leading to sub-optimal response.[93] It is imperative that these patients be identified at diagnosis, allowing alternative treatment strategies to be implemented. To effectively enable this, a greater understanding of the underlying intrinsic biology of poor response is needed.

Our laboratory has previously demonstrated that the degree of imatinib-induced kinase inhibition achieved in CML cells *in vitro* and *in vivo* is a key predictor of patient response to imatinib.[135,242] Subsequently, we and others demonstrated that reduced kinase inhibition was intrinsically linked to the amount of drug which was actively pumped into the cell via the organic cation transporter 1 (OCT-1).[96,97,99,149,151] Hence, reduced activity of the OCT-1 protein is a likely key to primary imatinib resistance and suboptimal response in CP-CML patients. The OCT-1 activity (OA) assay, developed in our laboratory,[151] is a functional assay measuring imatinib uptake in CML patient mononuclear cells (MNCs) at diagnosis, which has excellent predictive value for molecular response in diagnostic CP-CML patient samples.[142,151] The defined median OA is 7.2ng/200,000 cells,[151] allowing patients with an OA < 7.2ng/200,000 cells to be classified as low OA, while those with OA > 7.2ng/200,000 cells have high OA.

Interestingly, we have shown that low OA can be partially overcome by dose increase in those patients who can tolerate dose escalation.[151] Analysis of the 5 year molecular data from the Australian Therapeutic Intensification in *De Novo* Leukaemia I (TIDEL I) trial (600mg up-front imatinib in newly diagnosed CP-CML patients) significantly linked low OA to the development of kinase domain mutations and disease progression.[142] Furthermore, in this same analysis low OA was demonstrated to be associated with a poorer rate of achievement of major molecular response (MMR; BCR-ABL1 mRNA levels < 0.1% (IS)) and complete molecular response (CMR; BCR-ABL1 mRNA levels negative in 2 consecutive assays) within 5 years of treatment and lower overall, transformation-free and event-free survival (TFS and EFS, respectively).[142] Hence, a low OA is a negative short- and longer term response predictor, and patients with low OA require careful management.

The OA assay was then used to further stratify patient groups, demonstrating that patients with OA in the lowest quartile (very low OA; OA < 4ng/200,000 cells) in the TIDEL I study had a significantly poorer response (lower MMR, EFS, TFS and overall survival [OS]; **Figure 3.1**) than patients in all other quartiles.[142] These results suggest that varied disease biology exists within CP-CML patients from the four OA quartiles. In particular, these results highlighted the importance of defining molecular and cellular characteristics associated with low and high OA CP-CML patients. The OA assay is likely to be providing a complex surrogate for many other transporters as previously suggested,[168] as well as for the intrinsic differences in biology across the patient cohort. As such, it is now important to determine what factors underlie the range of OA levels observed in CP-CML patients, and whether patients with low OA and poor response to imatinib have different overall disease characteristics associated with alternative biological mechanisms.

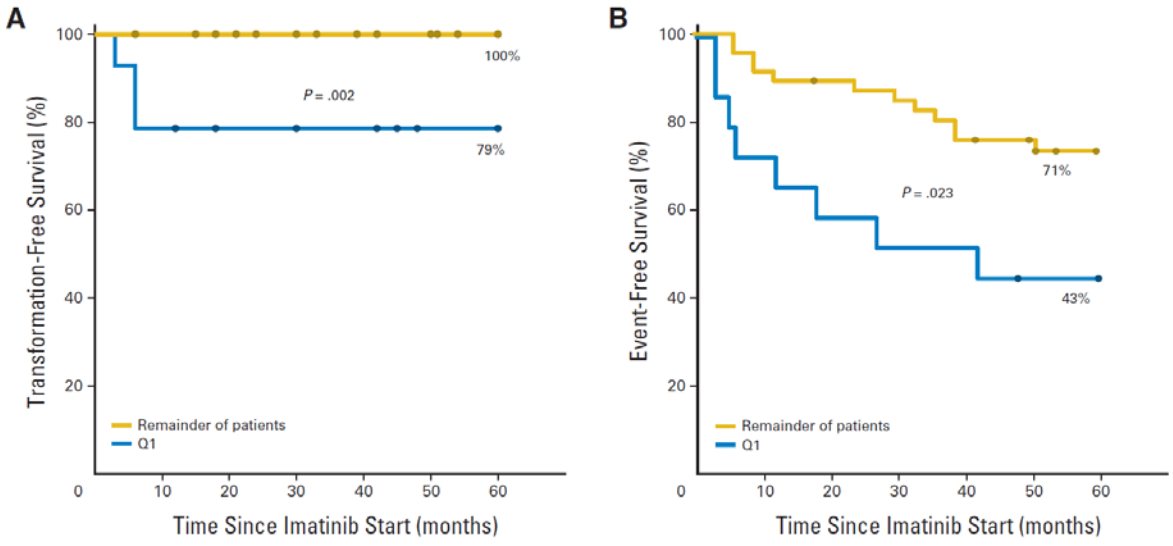


Figure 3.1 CP-CML patients with the lowest OA (Q1) are at risk of poor response to imatinib therapy

Transformation-free **(A)** and event-free survival **(B)** in quartile 1 (Q1; very low OA < 4ng/200,000 cells) patients, versus the remainder of patients (OA > 4ng/200,000 cells). Patients in Q1 have a significantly higher transformation rate (all patients who transformed to blast crisis were in Q1), and subsequently, also have a significantly higher number of events on TKI therapy. Figure from White *et al.,* 2010.[142]

3.1.1 OCT-1 activity varies with cell lineage

OCT-1 mRNA expression has been demonstrated to be influenced by the cell populations present in the peripheral blood (PB). Specifically, OCT-1 mRNA expression was significantly higher in PB polymorphonuclear cells (PMNCs), compared to MNCs, suggesting that OCT-1 mRNA expression in total PB leukocytes is influenced by the percentage of PMNCs.[153,155] Engler *et al.*[243] investigated this observation in relation to OA, and demonstrated that OA in patient MNCs was strongly related to cell lineage. In particular, OA was demonstrated to be highest in pure populations of neutrophils, followed by monocytes, and lowest in the lymphocyte cell population (**Figure 3.2**). Furthermore, MNC OA was significantly correlated with neutrophil OA, indicating that the overall OA read-out was due to the predominance of granulocytes in the MNC population.[243] These results suggested that the predictive ability of the OA assay in CP-CML patients is strongly related to the cellular composition of the PB sample, from which the MNCs are isolated. However, further immunophenotypic analysis of CML patient MNCs at diagnosis, providing highly specific cell population information in relation to OA, is lacking.

3.1.2 Immunophenotyping in CML

Flow cytometric immunophenotyping is an indispensable diagnostic tool for the diagnosis, classification, and monitoring of haematologic neoplasms. Immunophenotyping is able to evaluate individual cells in suspension for the presence and absence of specific cell surface antigens, to determine specific phenotypes.[244] However, flow cytometric immunophenotyping does not have a well-defined role in the diagnosis of CML, as genomic methods (e.g. measurement of BCR-ABL1 mRNA levels) are generally utilised for identification, classification, and disease monitoring. Yet, immunophenotyping has provided valuable insight into CML biology. Detailed three-colour immunophenotyping of peripheral

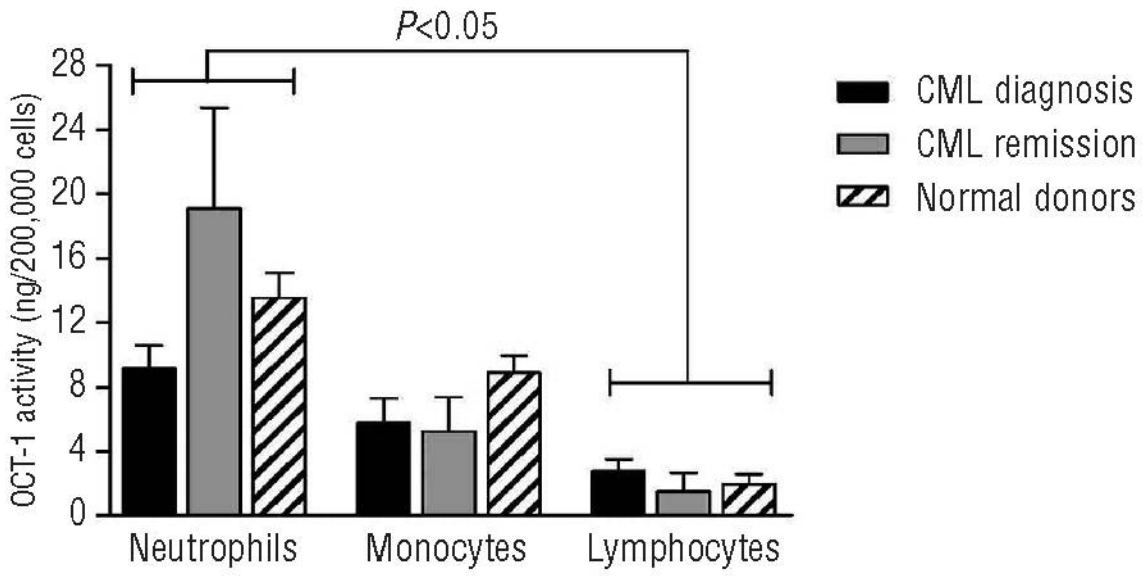


Figure 3.2 OCT-1 activity (OA) is highest in neutrophils, followed by monocytes and lymphocytes, isolated from CML patients and normal donors

Data from CML patients at diagnosis, in CCyR (remission), and normal individuals is shown. OA was significantly higher in neutrophils from all 3 patient groups, compared to lymphocytes. Columns represent the mean + standard error of the mean (SEM). Figure from Engler *et al*., 2011.[243]

lymphocytes from CP-CML patients, identified that the naïve T-lymphocyte compartment is well preserved in CML.[245] Furthermore, Normann *et al*.[246] sought to identify a unique immunophenotype of Ph+ CD34+ CML cells, and identified that all CML patients with signs of disease progression contained CD34+ cells that were also positive for CD7 (CD34+CD7+), indicating prognostic importance for CD7 expression in CML. Different immunoprofiles have also been identified in CML patients treated with imatinib, dasatinib and nilotinib,[247] and flow cytometry has been used to successfully measure imatinib uptake in CML,[248] and even to detect the presence of the Bcr-Abl1 oncoprotein.[249,250]

Recently, Rohon *et al.*[251] performed immunophenotyping on CML patient samples at diagnosis and during TKI (imatinib and dasatinib) therapy, using an extensive flow cytometry antibody panel. Leukocyte and lymphocyte subclasses were analysed, with decreased B- and dendritic cells, and increased natural killer (NK) T-like cells identified at diagnosis in the bone marrow, with all changes normalising during imatinib therapy.[251] Patients treated with dasatinib could be divided into two groups: patients with an immunoprofile that normalised with dasatinib therapy; and those demonstrating immune activation, represented by increased CD8+, NK- and NKT-like cells in the PB. These results demonstrated the differential effects of these TKIs on immune cells in CML patients. Furthermore, Janssen *et al.*[252] used a new flow cytometric approach to detect normal stem cells in newly diagnosed CML patients, identifying patients associated with lower clinical risk scores (i.e. Sokal score), lower imatinib toxicity and greater molecular response rates, than patients without detectable normal stem cells. While these studies have all generated intriguing results, flow cytometric immunophenotyping is still not routinely used in CP-CML diagnosis, nor is it used to identify patients best suited to a particular therapeutic option, due to the well-established genomic methods currently utilised.

3.1.3 Immune surveillance in CML

CML is one of the most immunogenic cancers in man. The high cure rate observed in the pre-TKI era with allogeneic haematopoietic stem cell transplant (HSCT), combined with the high success rate of donor lymphocyte infusions to eradicate CML in relapsed patients after HSCT[253] exemplifies this. Furthermore, the inherent immunogenicity of CML cells is seen in the detection of cytotoxic T-lymphocyte (CTL) responses to the leukaemia-associated antigens; proteinase 3, WT1 (Wilms Tumour 1),[254] PRAME (preferentially expressed antigen of melanoma)[255] and the polycomb group protein BMI-1 (B lymphoma Mo-MLV insertion region 1 homolog)[256] in CML patients. The presence of CTLs specific for PR1, a peptide derived from proteinase 3, have been associated with improved outcome in CML patients who received interferon-α (IFN-α), as well as those who have received HSCT.[257] In the TKI-era, CP-CML patients given a combination of imatinib and IFN-α, who subsequently discontinued imatinib and maintained treatment-free remission had elevated levels of PR1-specific CTLs.[258] Importantly, vaccination strategies using WT1 and PR1 have been demonstrated to be immunogenic and reduce leukaemia cell burden in a clinical trial of eight patients with myeloid leukaemias.[259] Furthermore, both the French and Australian imatinib cessation studies found that patients who had been treated with IFN-α prior to starting TKI therapy had a lower relapse risk.[260,261] The presence of the killer-cell immunoglobulin-like receptor (KIR) genotype KIR2DS1 was recently associated with a poor clinical response to standard dose imatinib[262] and dasatinib,[263] suggesting that natural killer effector function is relevant in modulating clinical responses to these TKIs. Conversely, the role of immune suppressor cells, regulatory T-cells and myeloid derived suppressor cells is poorly defined in the setting of TKIs, although naturally occurring regulatory T-cells have been shown to be increased in patients with an inferior response to imatinib.[264]

3.1.4 Approach

To identify patient specific immunophenotypes, flow cytometric analysis (**Section 2.3.5**) was performed prospectively on MNCs isolated from the peripheral blood (**Section 2.3.1**) of 27 CP-CML patients at diagnosis. Various antibody panels were employed, including single-, double- and triple-colour panels (refer **Appendix 1** for exact cell surface markers in Panel 1). Four different immunophenotypic datasets were obtained: total ungated percentage, total ungated mean fluorescence intensity (MFI), lymphocyte-gated percentage and lymphocyte-gated MFI (**Figure 3.3**). Data for the MFI datasets was normalised against the specific isotype control for each cell surface marker, allowing a normalised MFI value to be determined. OCT-1 activity (OA) was subsequently measured as previously published,[96] using cryopreserved MNCs (**Section 2.3.1**) isolated from the peripheral blood diagnosis samples of the same CP-CML patients, using the OCT-1 inhibitor prazosin (**Section 2.4.1.1**). OA is expressed as ng of imatinib per 200,000 cells. Using a significance cut-off ($P < 0.05$), the top cell surface markers differentially expressed between very low (OA < 4 ng/200,000 cells) and other OA (OA > 4 ng/200,000 cells) patients in each dataset were identified.

Flow cytometric immunophenotyping was then performed on a further 7 CP-CML patients (Panel 1, **Appendix 1**) to increase the overall number of patients in the cohort, enabling a 34 patient dataset to be compiled. A predictive cell surface marker classifier for OA was developed; with a 4-marker classifier (88% accuracy) using iterativeBMA (**Section 2.6.2**) which had the greatest accuracy compared to any classifier developed using PAMR (**Section 2.6.3**). Validation of this 4-marker classifier was performed using an independent cohort of 60 *de novo* CP-CML patients analysed by flow cytometric immunophenotyping (**Section 2.3.5**) using antibody Panel 2 (**Appendix 2**). Unfortunately, the classifier failed to predict very low OA patients from the rest in this independent patient cohort, indicating that perhaps an immunophenotyping classifier may not be suitable to replace OA.

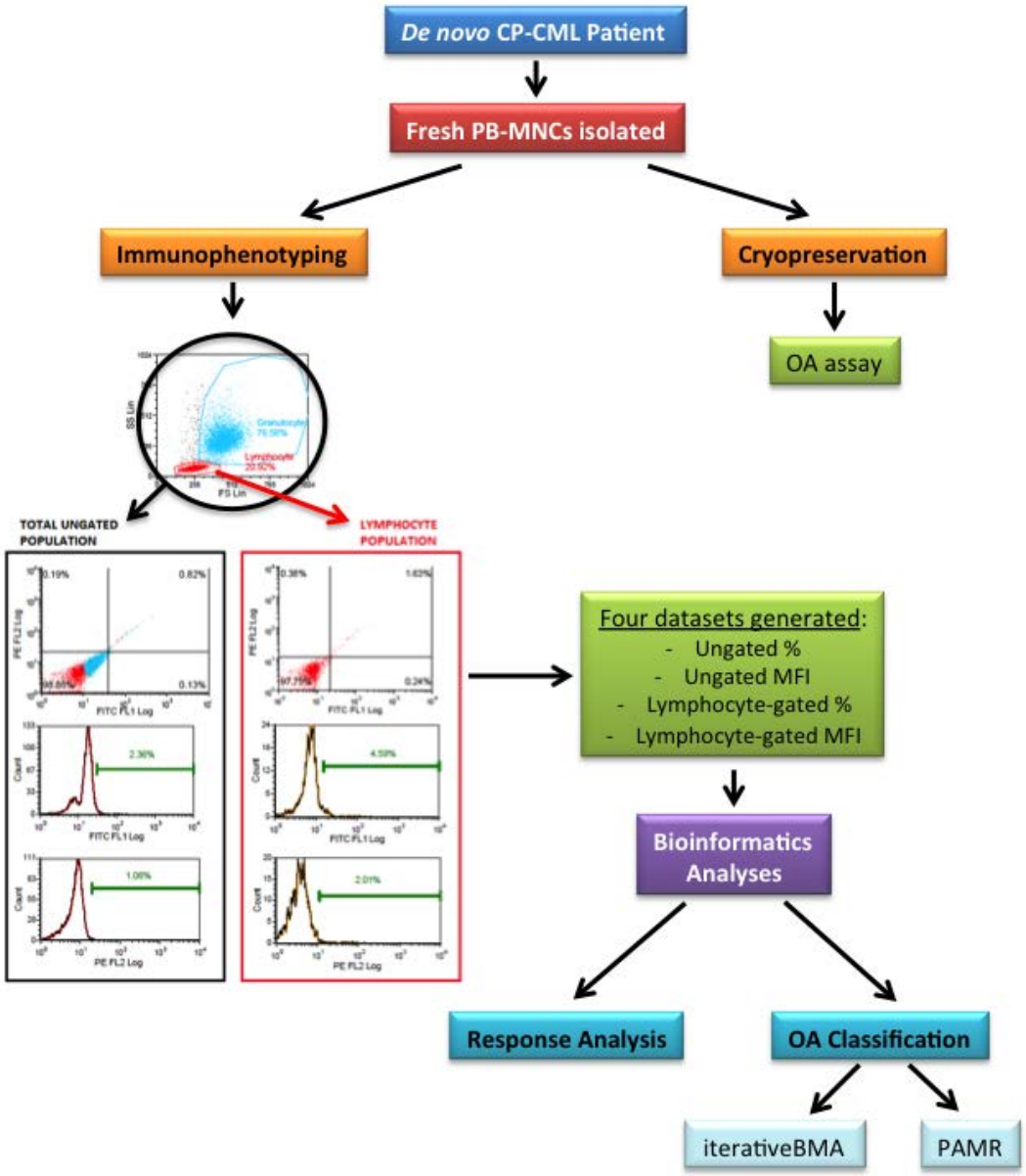


Figure 3.3 Immunophenotyping workflow for the identification of the Total and Lymphocyte-gated cell populations and subsequent downstream analyses

Peripheral blood mononuclear cells (PB-MNCs) were isolated from *de novo* CP-CML patients. Fresh MNCs were used to perform the immunophenotyping analyses, while cryopreserved MNCs were used to perform the OA assay. Specifically, the lymphocyte-gated cell population was delineated from the total cell population (including the granulocyte population [**blue**], but excluding dead cells/debris), based on forward scatter (FS) and side scatter (SS) properties (identified by the **red** gate). The four datasets generated by immunophenotyping were then used for downstream bioinformatics analyses including OA classification and response analysis.

Altogether, 95 *de novo* CP-CML patients had immunophenotyping data available for multiple cell surface markers, allowing correlation analyses with patient molecular response (early molecular response [EMR] and 12 month major molecular response [MMR]), as well as OA, to be performed. The analyses discussed in this chapter were used to determine the immunophenotype of CP-CML patients, specifically if very low OA patients had significant lineage differences from all other OA patients. This data will be used to further understand the underlying biology associated with very low OA.

3.2 Results

3.2.1 Immunophenotyping analysis of 27 CP-CML patients at diagnosis

Prospective flow cytometric immunophenotyping was performed at diagnosis for 28 CP-CML patients, of which 27 were subsequently enrolled to the Therapeutic Intensification in *De Novo* Leukaemia (TIDEL) II clinical trial run in our centre. The remaining patient failed the criteria for enrolment, and was subsequently removed from further analysis. Single- and double-colour immunophenotyping was performed for 27 different cell surface markers in 40 different combinations (Panel 1, **Appendix 1**) using peripheral blood mononuclear cells (PB-MNCs). The antibody panel was designed to cover the most common cell types observed in patient samples, including monocytes, granulocytes and lymphocytes, while cell surface markers more specific for haematopoietic progenitor cells and erythrocytes were also included. Antibodies for genes known to be associated with CML (*ABCB1* and *ABCG2*), as well as for genes previously suggested to be differentially expressed between patients with very low versus very high OA in our laboratory (*FZD6*, *LRP6* and *ITGB5*) were also included.

As immunophenotyping was performed prospectively, no bias towards patient OA values was applicable, and once the specific OA value for each patient was determined, the correlation between OA and immunophenotype was determined. Immunophenotypic analysis was performed on both the total cell population, which contains all cells including the expanded granulocyte population associated with CML; and the lymphocyte-gated cell population (**Figure 3.3**), which is more specific for the lymphocyte cell population, and includes the more immature blast cells associated with progression in CML. Both the percentage of cells positive for a marker, and the mean fluorescence intensity (MFI) were recorded for each cell population.

3.2.1.1 Analysis of the percentage immunophenotyping datasets

As OA has been demonstrated to vary with cell lineage,[243] the immunophenotyping performed here focussed on identifying further specific lineage differences between patients with very low OA (OA < 4 ng/200,000 cells), compared to the rest (OA > 4 ng/200,000 cells). Of the 40 cell surface marker combinations analysed by flow cytometry, 3 markers had significantly decreased expression (by percentage) in very low OA patients, compared to the rest, from the total ungated cell population (**Figure 3.4**). Of the markers identified, CD45 is a leukocyte marker commonly used for the identification of haematopoietic cells,[265] while both CD15 and CD66b are more specific granulocyte (neutrophil) markers.[266,267] When the same analysis was performed, this time using the lymphocyte-gated percentage data, 3 different markers were identified as differentially expressed between very low and other OA patients (**Figure 3.5**); however, two of the significant markers (CD34 and CD34+CD38-) describe a very similar cell surface marker profile. CD34 is a marker of haematopoietic progenitor cells,[268] and CD34+CD38- cells indicate a more primitive subpopulation of progenitor cells,[269] while CD38 is found on the surface of many cells.[270] CD38 is a differentiation marker and when used in combination with CD34, CD38 positivity identifies a more mature cell.

3.2.1.2 Analysis of the MFI immunophenotyping datasets

Analysis of the MFI data generated from the total ungated cell population was performed. The MFI value generated for each cell surface marker was normalised against the MFI value of the isotype control for each patient before being used for statistical analysis. Four markers were identified with significantly increased normalised MFI expression in very low OA patients, compared to the rest, from the total ungated cell population (**Figure 3.6**). The markers each identify different cell types including mature T-cells (CD3),[271] monocytes (CD14),[272] neutrophils (CD16),[273] and B-cells (CD20).[274] When the same analysis was

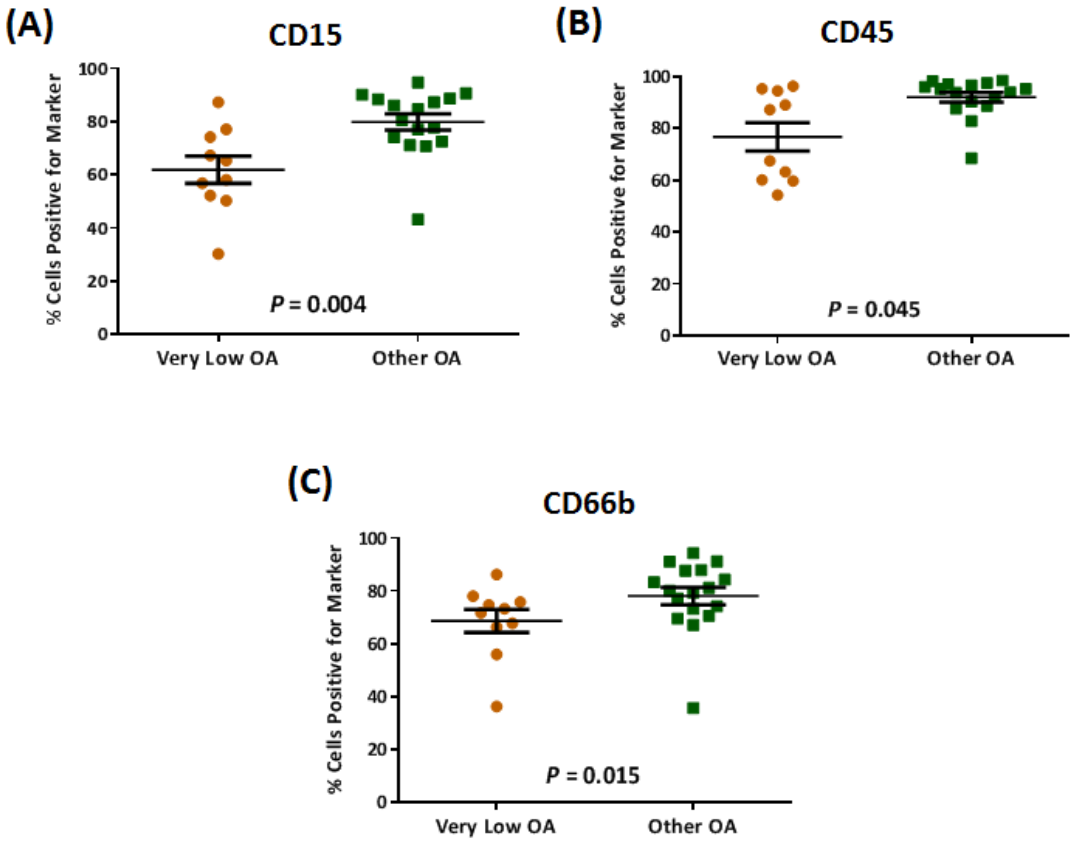


Figure 3.4 Significant cell surface marker expression between very low and other OA CP-CML patients, using total ungated percentage immunophenotyping data

The difference in the percentage of cell surface marker expression for **(A)** CD15, **(B)** CD45 and **(C)** CD66b, between very low (n = 10) and other OA (n = 17) patients, is plotted using total ungated percentage data. Scatterplots demonstrate the mean percentage (%) of cells positive ± the standard error of the mean (SEM). Mann-Whitney tests were performed in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells.

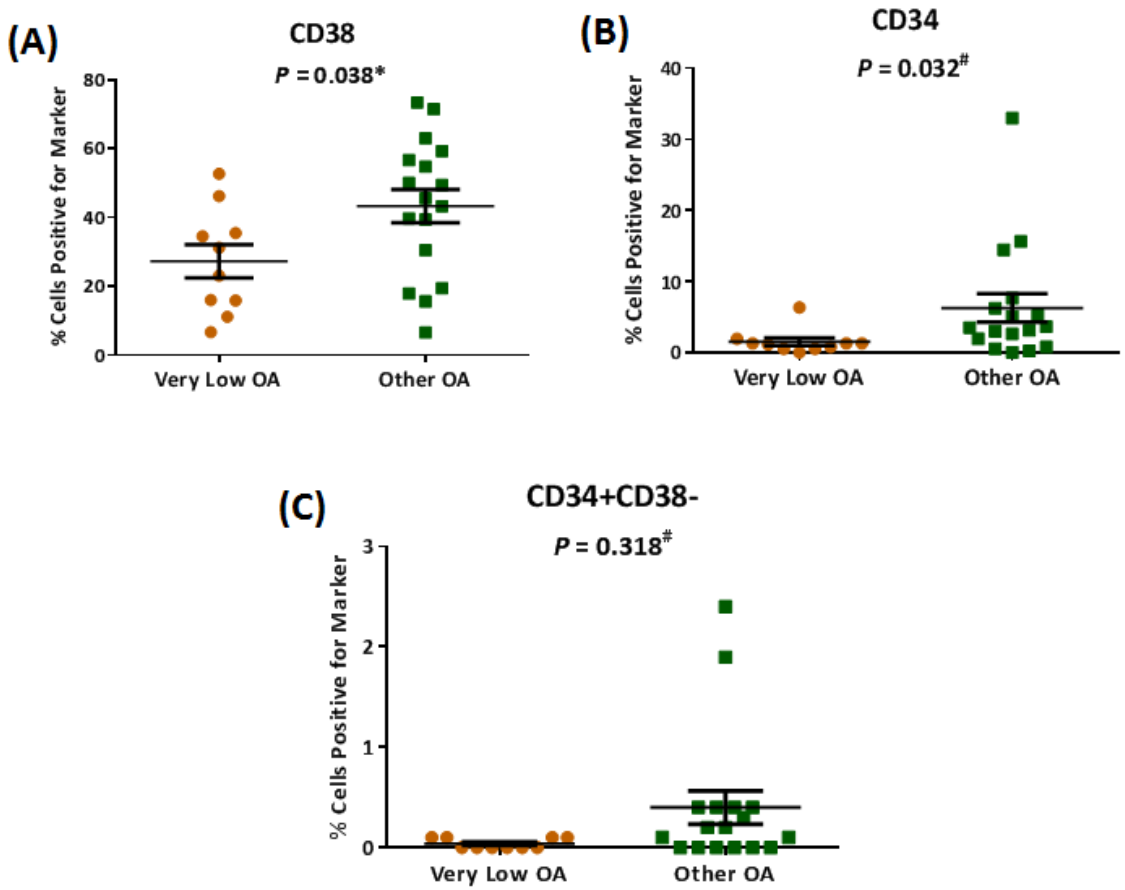


Figure 3.5 Significant cell surface marker expression between very low and other OA CP-CML patients, using lymphocyte-gated percentage immunophenotyping data

The difference in the percentage of cell surface marker expression for **(A)** CD38, **(B)** CD34 and **(C)** CD34+CD38-, between very low (n = 10) and other OA (n = 17) patients, is plotted using lymphocyte-gated percentage data. Scatterplots demonstrate the mean percentage (%) of cells positive ± SEM. Student's *t*-tests (*) or Mann-Whitney tests (#) were performed as appropriate in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells.

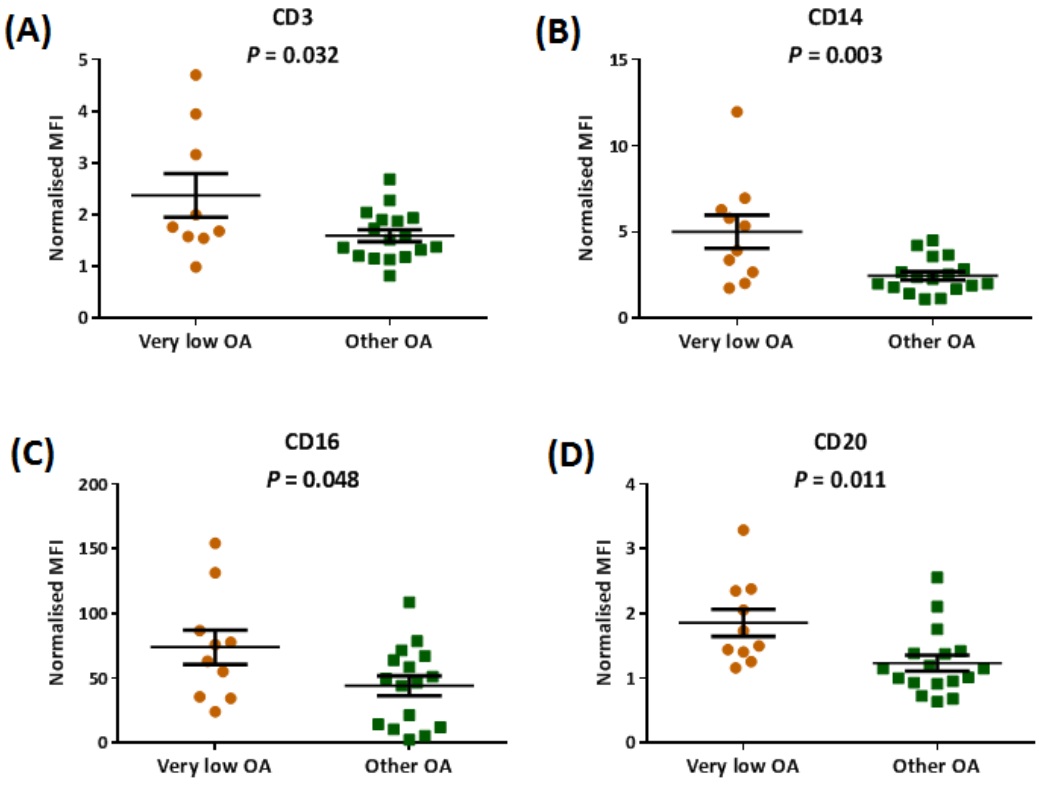


Figure 3.6 Significant cell surface marker expression between very low and other OA CP-CML patients, using total ungated MFI immunophenotyping data

The difference in the normalised mean fluorescence intensity (MFI) of the cell surface markers **(A)** CD3, **(B)** CD14, **(C)** CD16 and **(D)** CD20, between very low (n = 10) and other OA (n = 17) patients, is plotted using total ungated MFI data. Scatterplots demonstrate the mean normalised MFI of cells ± SEM. Student's *t*-tests were performed in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells.

performed, this time using the lymphocyte-gated normalised MFI data, only a single marker (ABCB1) was identified with significantly increased MFI expression in very low OA patients, compared to other OA patients (**Figure 3.7**).

Intriguingly, no cell surface marker was consistently differentially expressed across both the total ungated cell population (which includes the lymphocyte population) and lymphocyte-gated cell population alone, by either percentage or MFI analyses. This may be due to the differences between each dataset, with MFI calculated as a normalised ratio in respect to the isotype control, while the percentage dataset is not normalised, meaning smaller changes in the percentage dataset would be detected compared to the MFI dataset. Furthermore, the size of the cell populations (ungated versus lymphocyte-gated) was determined not to influence the MFI or percentage measures when the absolute cell counts (i.e. relative number of lymphocytes and granulocytes) were assessed.

However, the significant reduction of two granulocyte/neutrophil markers (CD15 and CD66b) in very low OA from the ungated percentage analysis provides insight into the reduced OA of these patients, as neutrophils were shown to possess the greatest OA by Engler *et al.*[243] Confounding this however, is the observation that CD16, another neutrophil marker, had increased MFI in very low OA patients, compared to all other patients in this same cell population (ungated total MNCs). Yet, CD14 (monocyte) MFI was also increased in very low OA, and as monocytes were demonstrated to have lower OA than neutrophils,[243] it could be the combination of these results that provides the clearest indication of factors affecting OA, and the heterogeneity observed in CP-CML.

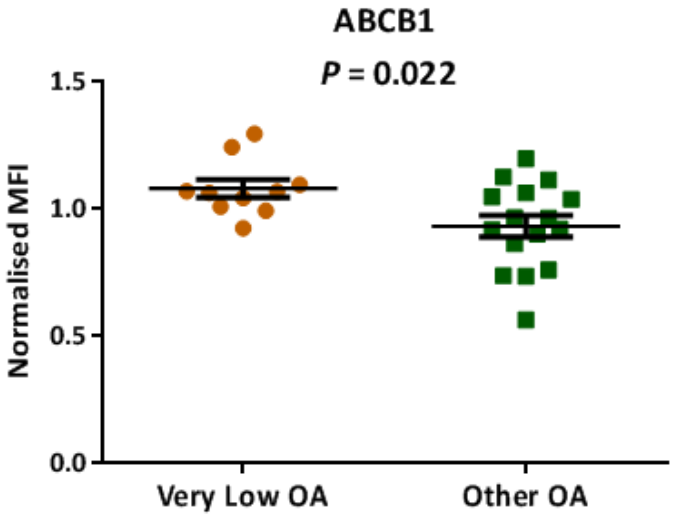


Figure 3.7 Significant ABCB1 cell surface expression between very low and other OA CP-CML patients, using lymphocyte-gated MFI immunophenotyping data

The difference between the normalised mean fluorescence intensity (MFI) of the cell surface marker ABCB1 in very low (n = 10) and other OA (n = 16) patients (data was missing for 1 patient), is plotted using lymphocyte-gated MFI data. Scatterplot demonstrates the mean normalised MFI of cells ± SEM. Student's *t*-test was performed in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells.

3.2.2 Development of an immunophenotyping classifier for OCT-1 activity

To identify if a specific combination of cell surface markers would have greater predictive power for differentiating very low OA patients from the rest, a predictive classifier using machine learning algorithms and the immunophenotyping data was developed. Classification is a supervised learning technique frequently applied to microarray (but in this case flow cytometric immunophenotyping) data to identify genes whose expression correlates with a specific sample phenotype. In this particular case, cell surface markers that are predictive for very low or other OA patients were sought. To do this, both the accuracy and number of cell surface markers required to obtain this accuracy were important (i.e. the greatest accuracy with the smallest number of markers is the goal). Two different machine learning packages were employed in the statistical environment *R*; iterative Bayesian Model Averaging (iterativeBMA)[275] and Prediction Analysis for Microarrays (PAM),[213] allowing the best classifier to be selected.

IterativeBMA is based on the Bayesian Model Averaging (BMA) method for marker selection and classification developed by Yeung *et al.*[209] BMA is a multivariate technique that takes model (small sets of predictive markers) uncertainty into consideration by computing the weighted average of the posterior probabilities that a test sample belongs to a given class, based on multiple models.[210] BMA is an excellent classification tool, containing many desirable features: computational efficiency; systematic determination of the number of predictive markers and models; posterior probabilities relevant to each prediction, marker and model are calculated; and models typically containing only a few markers are selected.[209,210] IterativeBMA was applied to the immunophenotyping data (the ungated percentage, ungated MFI, lymphocyte-gated percentage or lymphocyte-gated MFI datasets) to identify marker signatures predictive for very low OA patients. However, before iterativeBMA was applied, to improve the power of the analysis the sample size was

increased by analysing a further 10 *de novo* CP-CML patients using the immunophenotyping panel, of which 8 patients met the criteria for study enrolment. This provided a total of 35 patients with immunophenotyping data available for classification, but as an OA value was not available for one patient, a 34 patient training set was used with iterativeBMA. This same 34 patient dataset was also used to "test" the classifier, and although this method may over- or under-estimate the predictive accuracy, it does allow the smallest set of signature markers to be identified. Four different classifiers were developed using iterativeBMA, one for each of the immunophenotyping datasets (ungated percentage, ungated MFI, lymphocyte-gated percentage and lymphocyte-gated MFI).

3.2.2.1 Development of immunophenotyping classifiers using the percentage datasets and iterativeBMA

Firstly, the "KNN impute" function, available in the PAMR package,[214] was used to impute any missing values in the dataset by finding the *k*-nearest neighbours of the columns for which that marker is not missing data, within an Euclidean metric. Four markers were identified by iterativeBMA over 2 models that discriminated very low OA from the rest, using the ungated percentage immunophenotyping data for the training set (**Table 3.1**). Univariate analysis was then performed, with three of the top 4 markers corresponding with markers identified by the multivariate classification method (**Table 3.1**). The posterior probabilities were computed for every marker by averaging the posterior probabilities of the selected models in which each marker was included. Membership of the 4 signature markers in each of the 2 models selected by the iterativeBMA algorithm is demonstrated in **Figure 3.8.** Interestingly, the marker CD3+CD25+ was only ranked 20th by univariate measures, indicating that it may only be providing a very small influence to the overall classifier.

Table 3.1 The 4-marker signature discriminating very low from other OA patients as selected by a 34 patient training set using the ungated percentage immunophenotyping dataset with iterativeBMA

Cell Surface Marker	Cell Specificity	Posterior Probability (%)	Univariate (BSS/WSS) Rank
CD45+	Pan-leukocyte	100	2
CD34+HLADR-	Haematopoietic progenitor cell	100	4
CD3+CD25+	Activated T-cell	86	20
CD45-GlyA+	Erythrocyte	14	1

The table indicates chromosomal location, function, the posterior probability of each gene in multivariate modelling, and its rank in univariate modelling. OA indicates OCT-1 activity; BMA, Bayesian model averaging; GlyA, glycophorin A; and BSS/WSS, the ratio of between-group to within-group sum of squares.

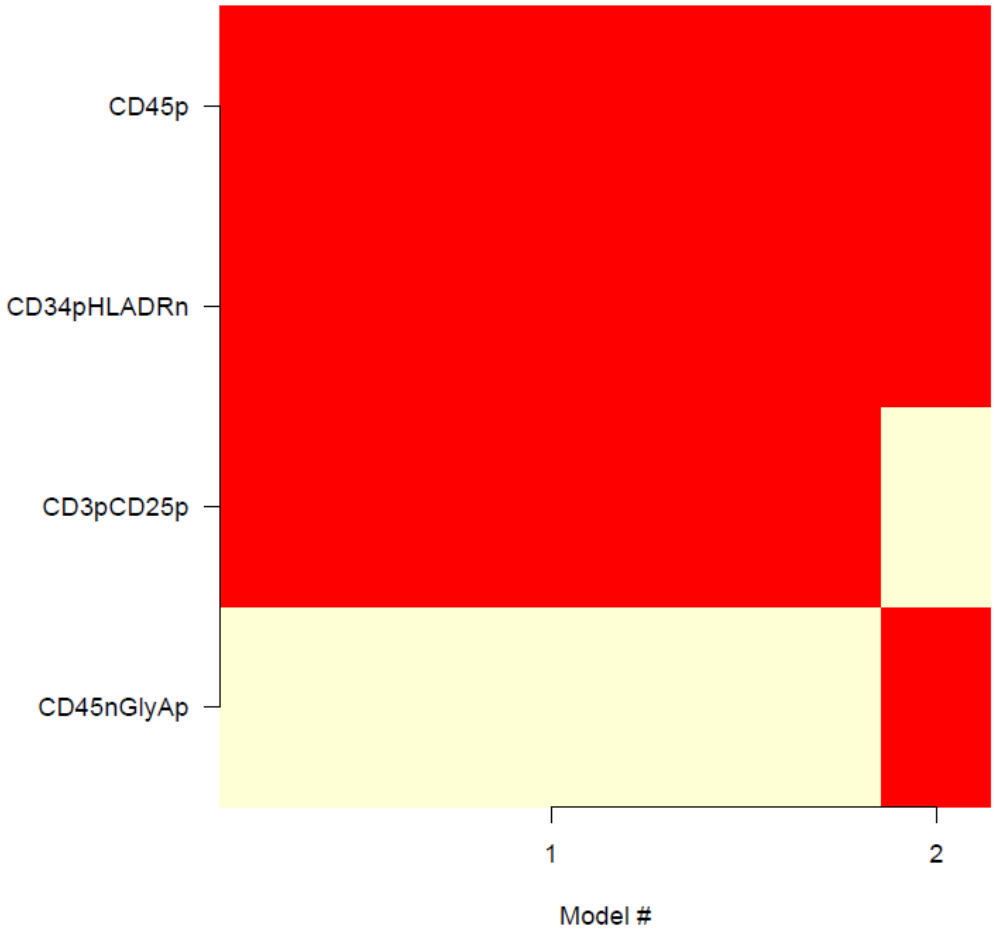


Figure 3.8 Model selection by iterativeBMA using a 4-marker signature developed from the ungated percentage immunophenotyping data

Membership of the 4 signature cell surface markers in the 2 models selected by the iterativeBMA algorithm, using a 34 patient training set derived from the total ungated percentage immunophenotyping dataset. The 4 signature markers are plotted on the vertical axis, with the 2 models on the horizontal axis. The widths of the column are proportional to the posterior probabilities of the selected models, with a **red** entry indicating when the corresponding gene is included in that particular model.

Next, to determine the predicted probability (of being very low or other OA) for each of the 34 test patient samples, the average predicted probabilities from the 2 selected models, weighted by the posterior probabilities, were calculated. The predicted probability threshold was set at 0.5; therefore if the predicted probability was less than 0.5, patients were classified as other OA, otherwise, if the predicted probability was greater than 0.5, patients were classified as very low OA. Four of the 34 patients were misclassified using the iterativeBMA gene signature (**Figure 3.9**), generating an overall accuracy of 88% (sensitivity = 77% [very low OA], specificity = 95% [other OA]), with a Brier score of 2.76. Sensitivity measures the proportion of actual positive results correctly identified (true positive rate; in this case very low OA), while specificity measures the proportion of actual negative results correctly identified (true negative rate; other OA). The Brier score is a relative probabilistic measure of the number of errors; hence, a relatively small number of errors correspond with a low Brier score, indicating higher prediction accuracy.

The lymphocyte-gated percentage dataset was used next as the training set with iterativeBMA to develop a classifier, and 6 markers were identified over 5 models that discriminated very low OA from the rest (**Table 3.2**). Interestingly, univariate analysis only ranked 2 of these markers in the top 10 (CD45+ and CD34+HLADR-), indicating that the other 4 markers may only be providing a small influence to the overall classifier. The posterior probabilities were then computed for each marker as before, with the membership of the 6 signature markers in each of the 5 models demonstrated in **Figure 3.10**. Next, the predicted probability (of being very low or other OA; threshold 0.5) for the same 34 patient samples was determined. Six of the 34 patients were misclassified using this iterativeBMA signature (3 very low and 3 other OA patients; **Figure 3.11**), generating an overall accuracy of 82% (sensitivity = 77%, specificity = 86%), with a Brier score of 3.57.

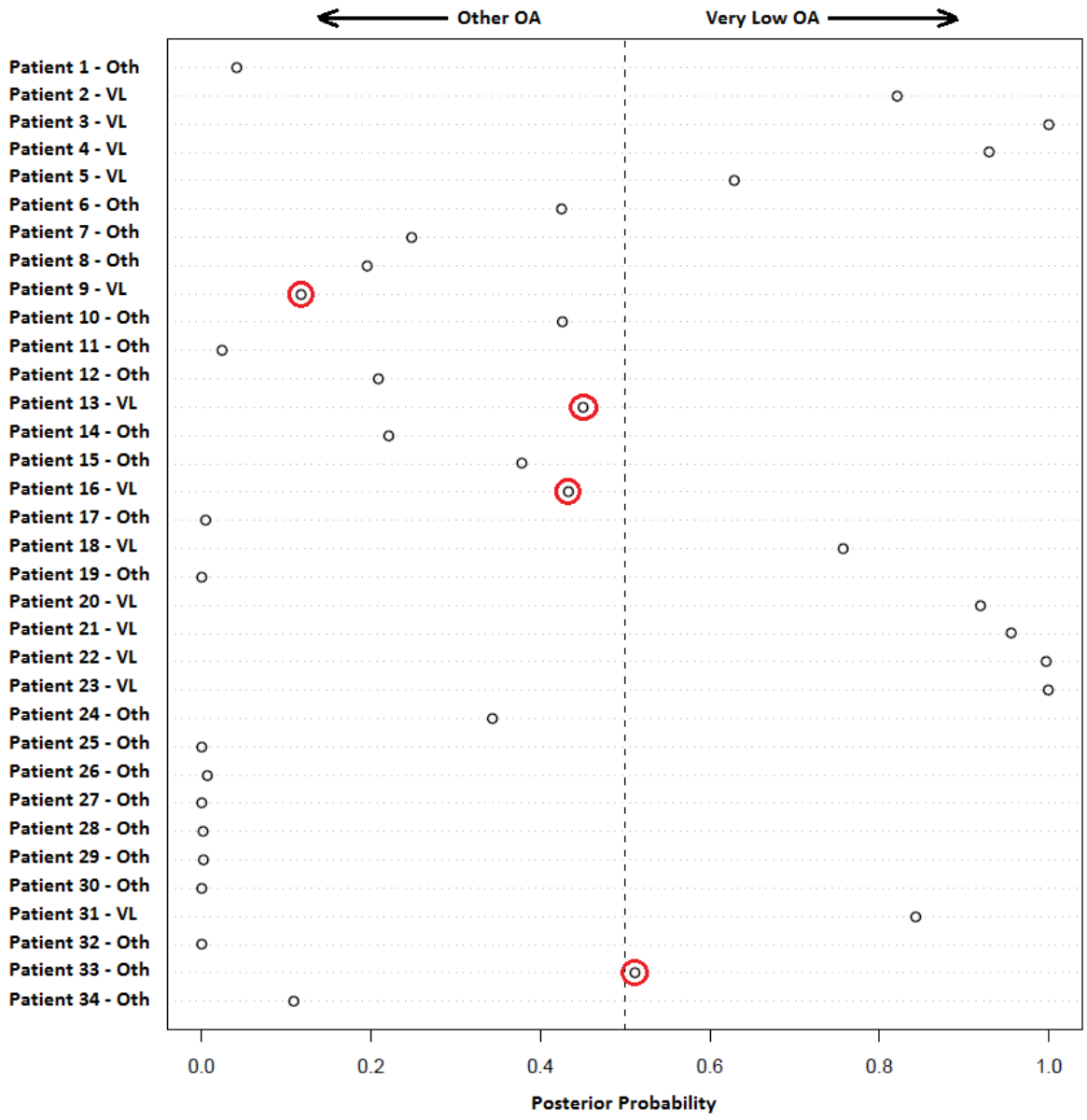


Figure 3.9 The 4-marker ungated percentage immunophenotyping classifier is 88% accurate against a test set comprising the same patients as the training set

Results of the validation of the 4-marker classifier developed by iterativeBMA (n = 34 training set), tested against the same 34 patients. A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate an "other" OA patient. Four patients were misclassified (**red** circles; 3 very low and 1 other OA patients), providing an overall accuracy of 88%, with a Brier score of 2.76. VL – very low OA (OA < 4 ng/200,000 cells); Oth – other OA (OA > 4 ng/200,000 cells).

Table 3.2 The 6-marker signature discriminating very low from other OA patients as selected by a 34 patient training set using the lymphocyte-gated percentage immunophenotyping dataset with iterativeBMA

Cell Surface Marker	Cell Specificity	Posterior Probability (%)	Univariate (BSS/WSS) Rank
CD34+HLADR-	Haematopoietic progenitor cell	100	4
CD45+	Pan-leukocyte	96.3	2
HLADR+	Antigen-presenting cell	52.1	28
CD20+	B-cell	14.8	21
CD45+GlyA+	Erythrocyte	8.9	24
CD4+	T helper cell	3.7	11

The table indicates chromosomal location, function, the posterior probability of each gene in multivariate modelling, and its rank in univariate modelling. OA indicates OCT-1 activity; BMA, Bayesian model averaging; GlyA, glycophorin A; and BSS/WSS, the ratio of between-group to within-group sum of squares.

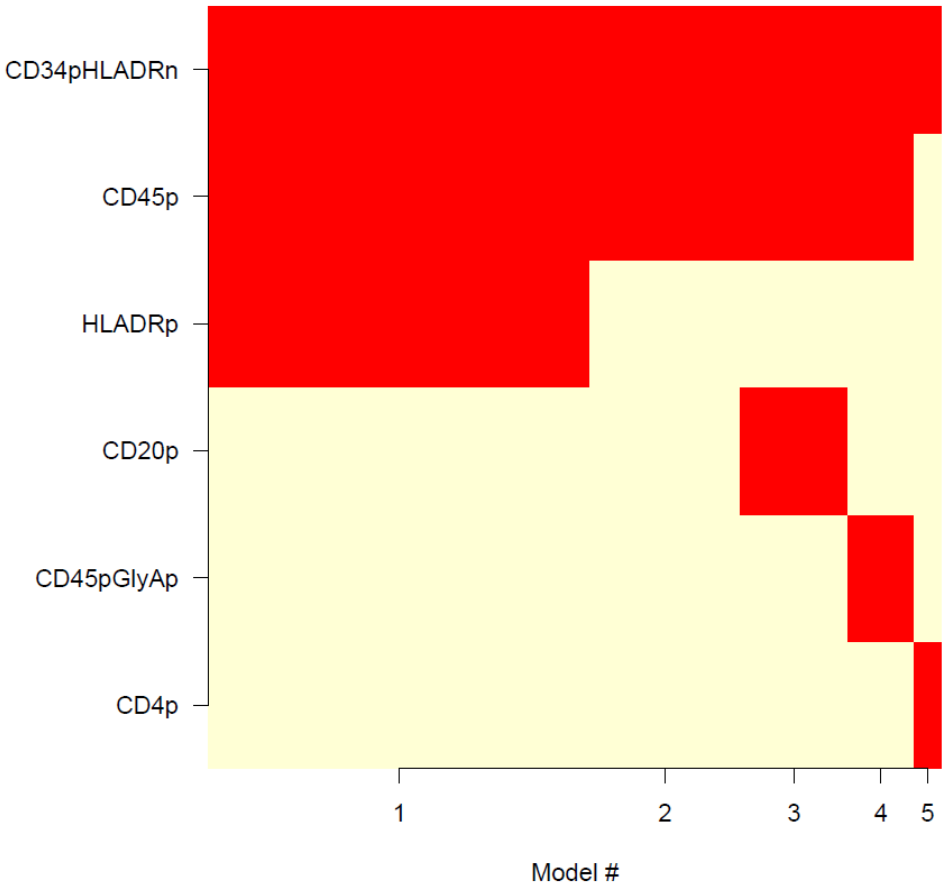


Figure 3.10 Model selection by iterativeBMA using a 6-marker signature developed from the lymphocyte-gated percentage immunophenotyping data

Membership of the 6 signature cell surface markers in the 5 models selected by the iterativeBMA algorithm, using a 34 patient training set derived from the lymphocyte-gated percentage immunophenotyping dataset. The 6 signature markers are plotted on the vertical axis, with the 5 models on the horizontal axis. The widths of the column are proportional to the posterior probabilities of the selected models, with a **red** entry indicating when the corresponding gene is included in that particular model.

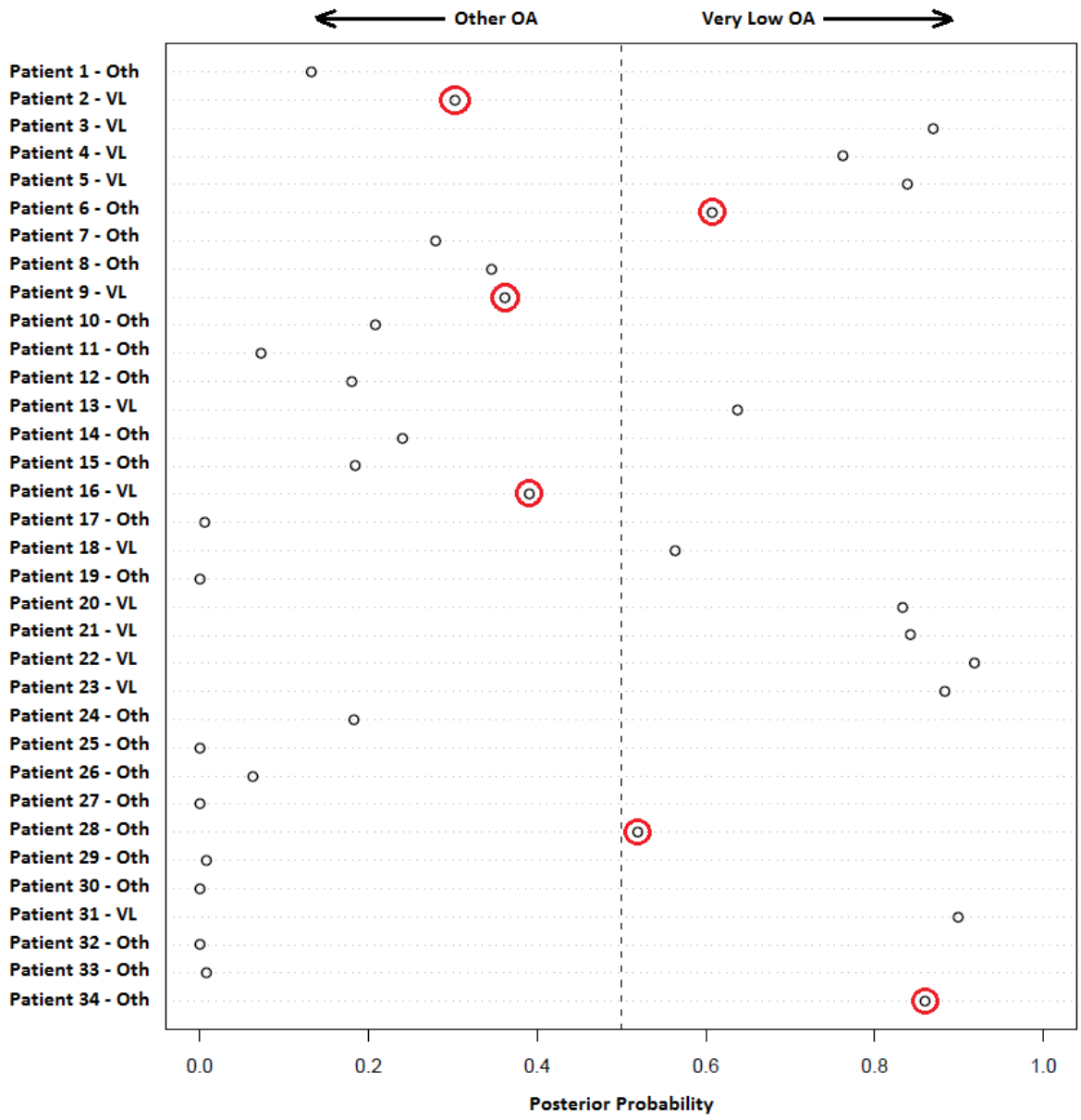


Figure 3.11 The 6-marker lymphocyte-gated percentage immunophenotyping classifier is 82% accurate against a test set comprising the same patients as the training set

Results of the validation of the 6-marker classifier developed by iterativeBMA (n = 34 training set), tested against the same 34 patients. A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate an "other" OA patient. Six patients were misclassified (**red** circles; 3 very low OA and 3 other OA patients), providing an overall accuracy of 82%, with a Brier score of 3.57. VL – very low OA (OA < 4 ng/200,000 cells); Oth – other OA (OA > 4 ng/200,000 cells).

3.2.2.1.1 Development of immunophenotyping classifiers using the MFI datasets and iterativeBMA

The ungated normalised MFI dataset was then used to develop a classifier using iterativeBMA, with 16 markers identified over 31 models. Unfortunately however, when the predicted probabilities of the 34 test samples were calculated using this classifier, 9 patients were misclassified, all of whom were very low OA patients (**Figure 3.12**), generating an overall accuracy of 73.5% (sensitivity = 69% [very low OA], specificity = 100% [other OA]), with a Brier score of 6.49. The lack of sensitivity towards very low OA makes this classifier unsuitable for further use.

Finally, the lymphocyte-gated normalised MFI dataset was used to develop a classifier using iterativeBMA, with 9 markers identified over 14 models, which discriminated very low OA from the rest (**Table 3.3**). Univariate analysis identified 5 of the 9 markers within the top 10 significant markers; however, the top marker by univariate analysis was CD15, which was not included in the classifier developed using the multivariate iterativeBMA analysis. Posterior probabilities were then calculated for each marker, with the membership of the 9 markers in the 14 models demonstrated in **Figure 3.13**. The predicted probability (of being very low or other OA; threshold 0.5) for the same 34 patient samples was then determined. Ten of the 34 patients were misclassified using this iterativeBMA signature (6 very low and 4 other OA patients; **Figure 3.14**), generating an overall accuracy of 68%, with a Brier score of 5.92.

For simplicity and ease of use, a classifier involving combinations of the 4 datasets (i.e. ungated percentage plus ungated MFI data) was not developed. Therefore, the best (greatest accuracy) classifier developed with iterativeBMA was the ungated percentage classifier, with an internal validation accuracy of 88%. However, before this classifier was

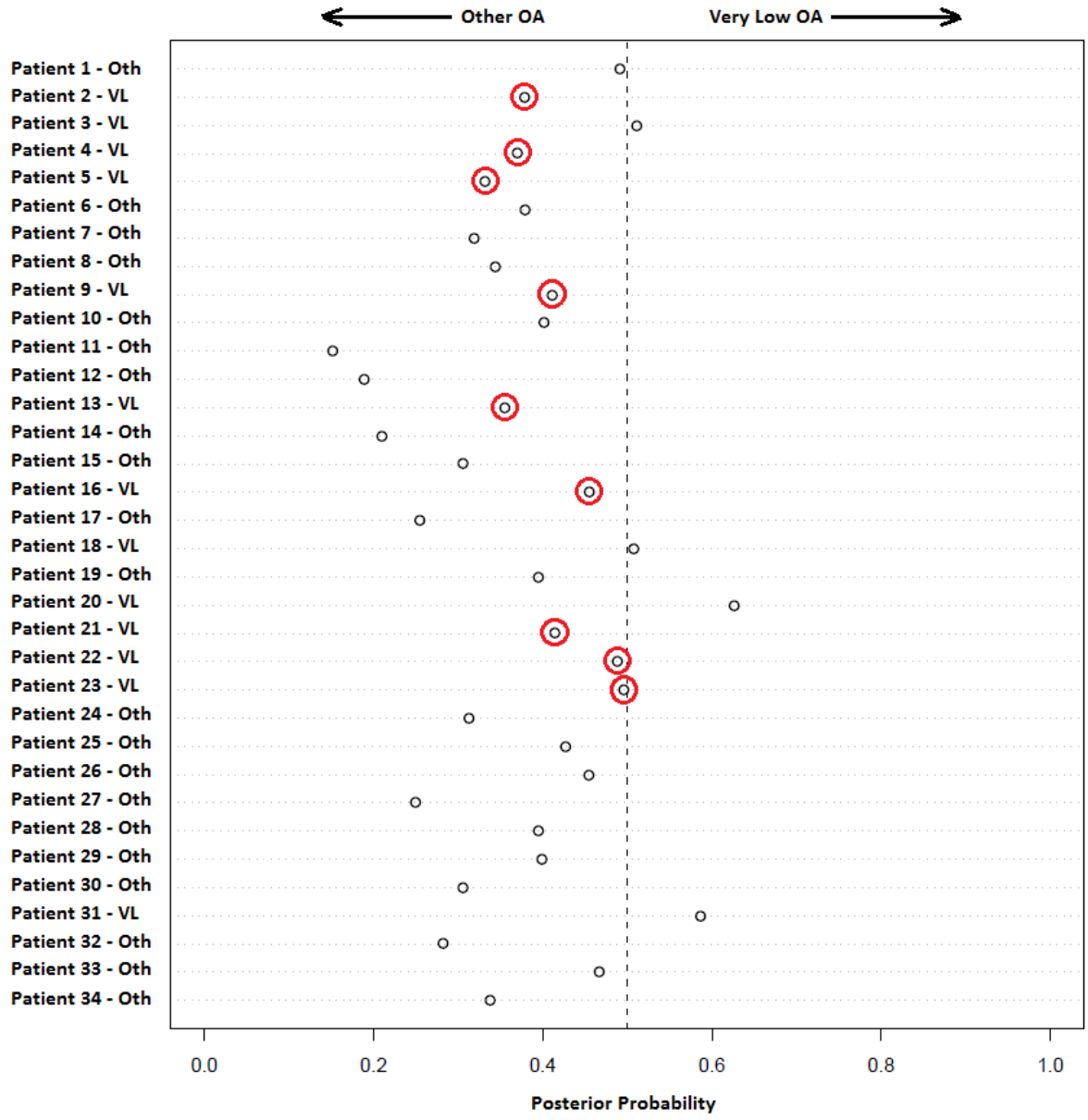


Figure 3.12 The 16-marker ungated MFI immunophenotyping classifier is 73.5% accurate against a test set comprising the same patients as the training set

Results of the validation of the 16-marker classifier developed by iterativeBMA (n = 34 training set), tested against the same 34 patients. A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate an "other" OA patient. Nine patients were misclassified (**red** circles; all very low OA patients), providing an overall accuracy of 73.5%, with a Brier score of 6.49. This classifier was not investigated further due to the low sensitivity for predicting very low OA patients (69%). VL – very low OA (OA < 4 ng/200,000 cells); Oth – other OA (OA > 4 ng/200,000 cells).

Table 3.3 The 9-marker signature discriminating very low from other OA patients as selected by a 34 patient training set using the lymphocyte-gated MFI immunophenotyping dataset with iterativeBMA

Cell Surface Marker	Cell Specificity	Posterior Probability (%)	Univariate (BSS/WSS) Rank
ABCB1	N/A	88.1	21
CD71	Erythrocyte	27.7	2
CD34	Haematopoietic progenitor cell	13.4	20
CD4	T helper cell	10.6	4
CD7	T-cell precursor	10.1	5
CD45	Pan-leukocyte	7.8	7
CD59	Erythrocyte	2.6	12
FZD6	N/A	2	3
CD38	Cell differentiation	1.9	19

The table indicates chromosomal location, function, the posterior probability of each gene in multivariate modelling, and its rank in univariate modelling. OA indicates OCT-1 activity; BMA, Bayesian model averaging; FZD6, frizzled 6; GlyA, glycophorin A; N/A, not applicable; and BSS/WSS, the ratio of between-group to within-group sum of squares.

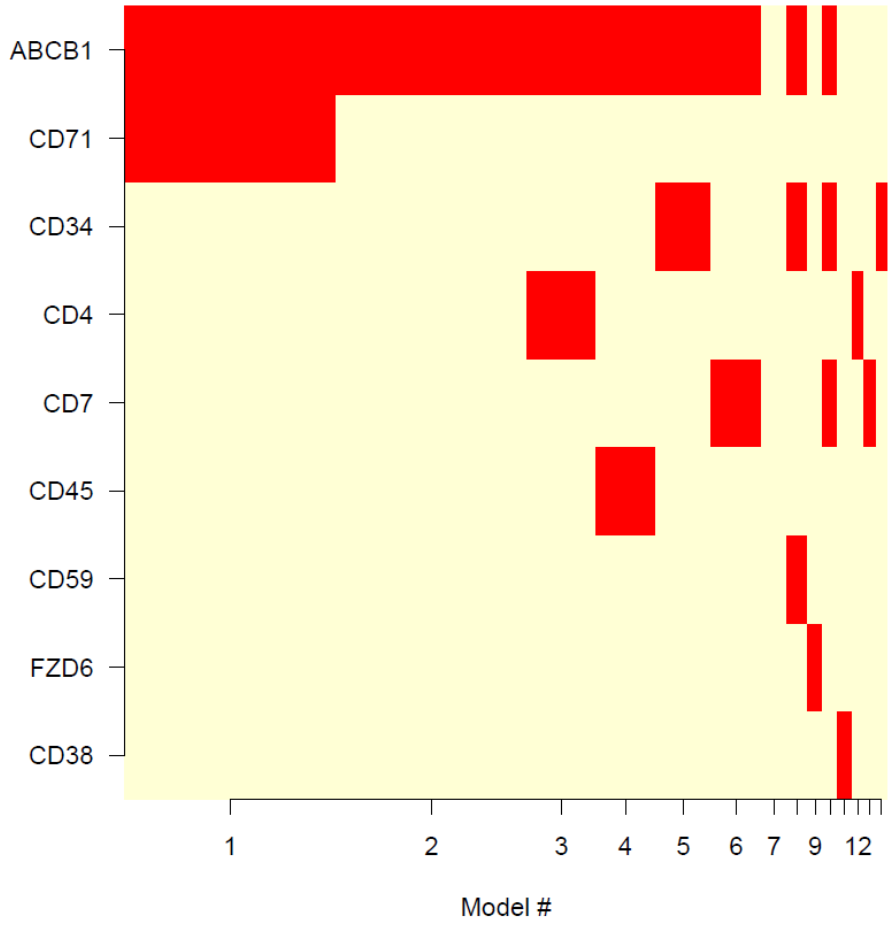


Figure 3.13 Model selection by iterativeBMA using a 9-marker signature developed from the lymphocyte-gated MFI immunophenotyping data

Membership of the 9 signature cell surface markers in the 14 models selected by the iterativeBMA algorithm, using a 34 patient training set derived from the lymphocyte-gated MFI immunophenotyping dataset. The 9 signature markers are plotted on the vertical axis, with the 14 models on the horizontal axis. The widths of the column are proportional to the posterior probabilities of the selected models, with a red entry indicating when the corresponding gene is included in that particular model.

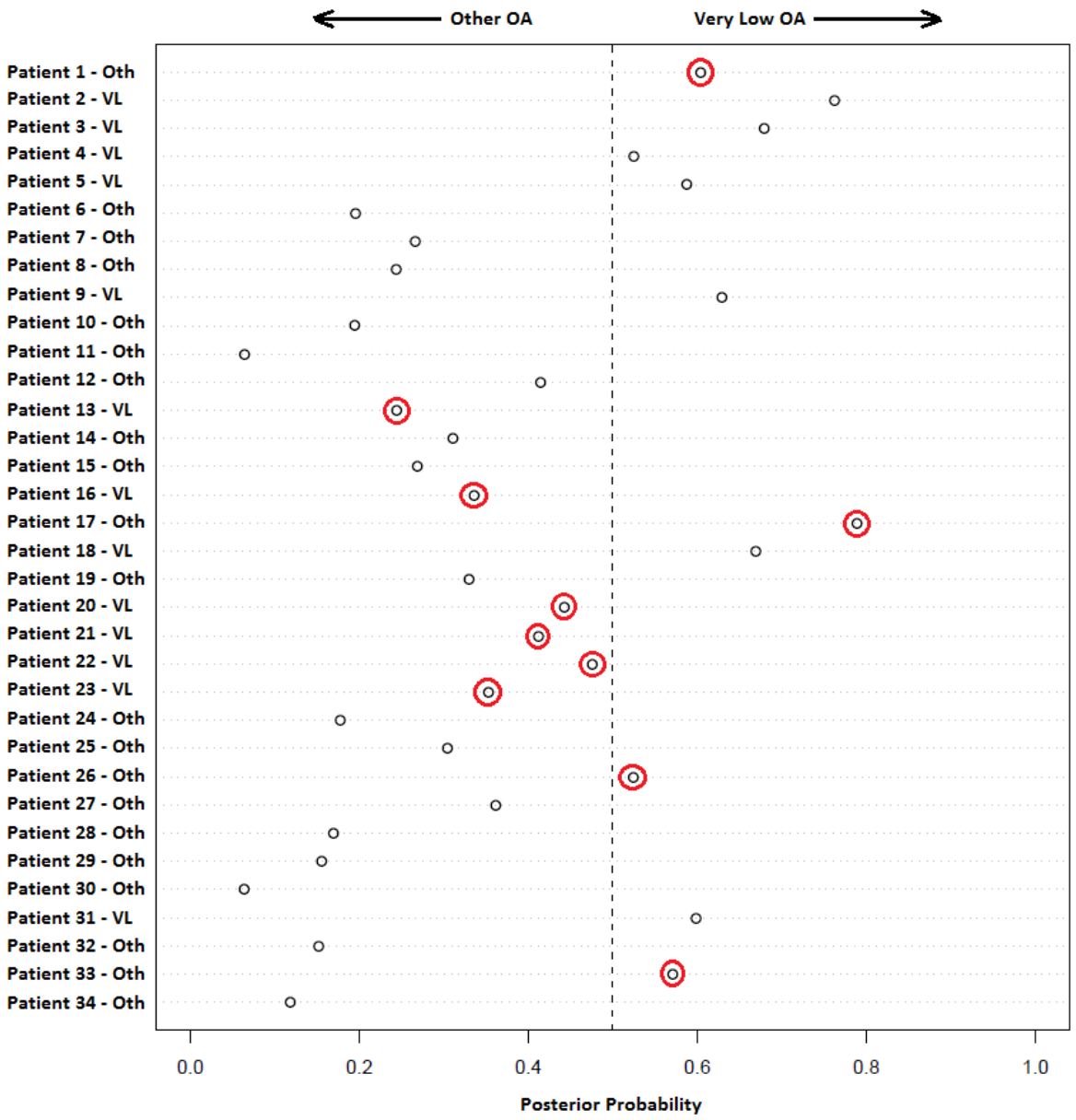


Figure 3.14 The 9-marker lymphocyte-gated MFI immunophenotyping classifier is 68% accurate against a test set comprising the same patients as the training set

Results of the validation of the 9-marker classifier developed by iterativeBMA (n = 34 training set), tested against the same 34 patients. A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate an "other" OA patient. Ten patients were misclassified (**red** circles; 6 very low OA and 4 other OA patients), providing an overall accuracy of 68%, with a Brier score of 5.92. VL – very low OA (OA < 4 ng/200,000 cells); Oth – other OA (OA > 4 ng/200,000 cells).

validated with a large, independent test set, a different classification algorithm (PAM) was analysed to determine if a superior classifier could be developed.

3.2.2.2 Development of immunophenotyping classifiers using the percentage datasets and PAMR

The Prediction Analysis for Microarrays in R (PAMR) package developed by Tibshirani *et al.*[213,214] performs sample classification using the nearest shrunken centroids method and provides a simple, fast and accurate classifier, with easily interpretable results. PAMR was applied to the total ungated percentage dataset, using the 34 patients as both the training and test sets (internal validation) and was able to generate a 9 marker classifier (threshold = 0.808) with an overall accuracy of 65% (sensitivity = 39%, specificity = 81%; **Figure 3.15**). The 9 markers selected by this classifier were CD45+, CD45-GlyA+, CD66b+, CD15+, CD4+, CD4+CD8-, CD15+CD16+, GlyA+ and CD16+. Four of 19 other OA patients and 8 of 13 very low OA patients were misclassified, indicating that this classifier was poor at predicting very low OA.

Next, the PAMR algorithm was applied to the lymphocyte-gated percentage dataset, once again using the 34 patient data for both the training and test sets. A 13 marker classifier was developed (threshold = 0.497) with an overall accuracy of 77% (sensitivity = 69%, specificity = 81%; **Figure 3.16**). The 13 markers selected by this classifier were CD45+, CD45-GlyA+, CD38+, CD7+,CD63+, CD3+CD7+, CD3+, CD66b+, CD4+, CD16+, CD34+, CD34+CD38+ and CD59+. Four of 19 Other OA patients and 4 of 13 very low OA patients were misclassified, indicating that this classifier was fairly even in its prediction of very low and other OA patients.

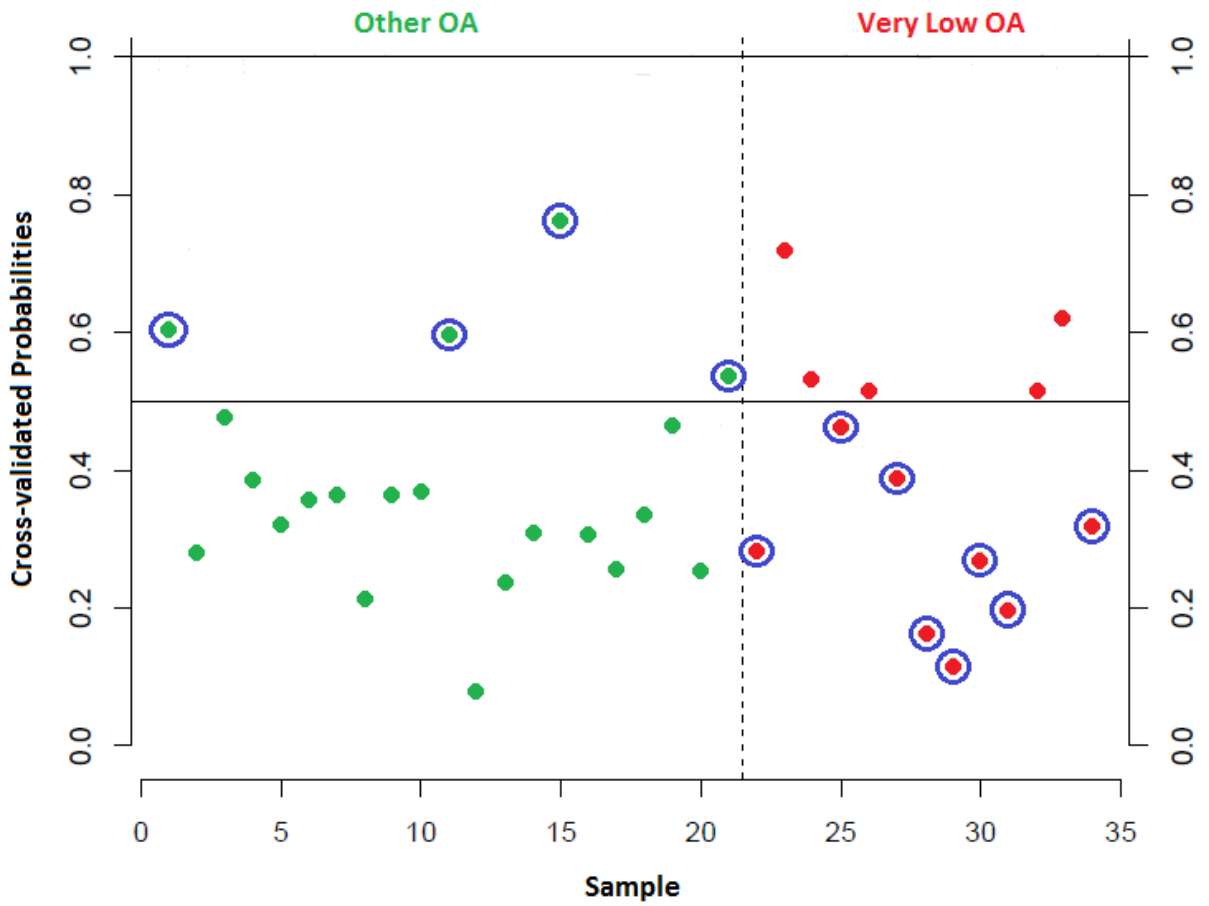


Figure 3.15 The 9-marker PAMR classifier developed using the total ungated percentage immunophenotyping dataset is 65% accurate

Results of the validation of the 9-marker classifier developed by PAMR using the total ungated percentage immunophenotyping dataset (n = 34 training set), tested against the same 34 patients. A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). Twelve patients were misclassified (**blue** circles; 8 very low OA and 4 other OA patients), providing an overall accuracy of 65%.

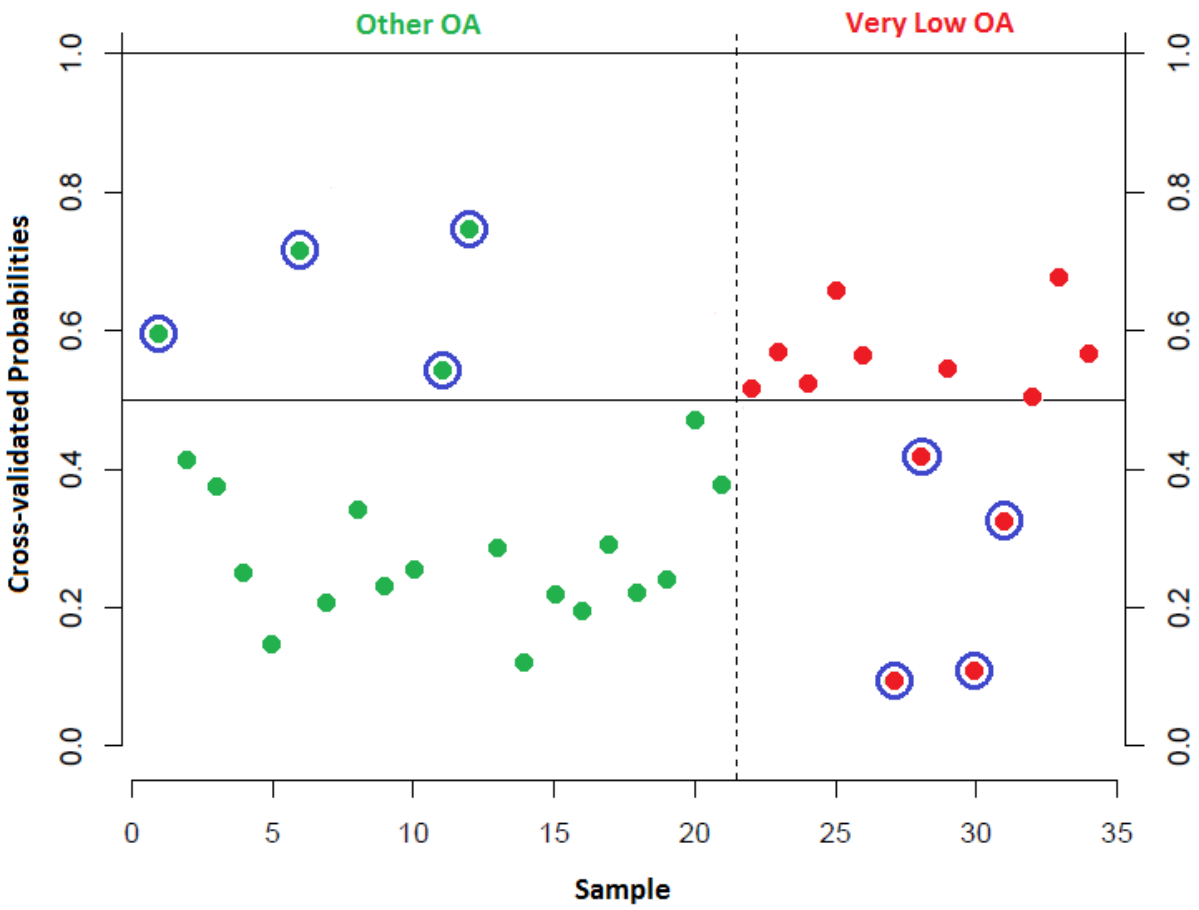


Figure 3.16 The 13-marker PAMR classifier developed using the lymphocyte-gated percentage immunophenotyping dataset is 77% accurate

Results of the validation of the 13-marker classifier developed by PAMR using the lymphocyte-gated percentage immunophenotyping dataset (n = 34 training set), tested against the same 34 patients. A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). Eight patients were misclassified (**blue** circles; 4 very low OA and 4 other OA patients), providing an overall accuracy of 77%.

3.2.2.3 Development of immunophenotyping classifiers using the MFI datasets and PAMR

Predictive classifiers using both the total ungated MFI and the lymphocyte-gated MFI datasets were also generated; however, both classifiers were not predictive for very low OA (2/13 correct, with overall accuracy of 65%; and 0/13 correct, with overall accuracy of 62%, respectively). Therefore, none of these classifiers had superior predictive power compared to the 4-marker, 88% accurate, iterativeBMA classifier developed using the total ungated percentage immunophenotyping dataset.

3.2.3 Validation of the OCT-1 activity immunophenotyping classifier in an independent patient cohort

Flow cytometric immunophenotyping using Panel 2 (**Appendix 2**) was performed using PB-MNCs from 61 CP-CML patients at diagnosis, of which 55 were subsequently enrolled to the Safety and Efficacy of Nilotinib in Newly Diagnosed Chronic Myeloid Leukaemia patients (ENESTxtnd) clinical trial run in our Centre. A further 5 patients were confirmed CP-CML, but not enrolled in this clinical trial; however, they were treated in our Centre so response data was available. Furthermore, a single patient was found to be diagnosed in BC-CML, not CP and was subsequently removed from further analysis. This meant a total of 60 CP-CML patients with immunophenotyping data and matched OA values available were used as an independent validation set. Once again, immunophenotyping analysis was performed on both the total cell population, and the lymphocyte-gated cell population, with both the percentage of cells positive for a marker, and the MFI recorded.

In order to validate the 4 marker ungated percentage classifier developed by iterativeBMA (refer **Section 3.2.2**), the 60 patients with useable immunophenotyping data were used as an independent test set against the original 34 patient training set (refer **Section 3.2.1**).

IterativeBMA was performed as before, using the total ungated percentage data with the specific 4 marker classifier originally developed. Unfortunately, the classifier was unable to be validated with this new, independent test set, as 27/60 patients were misclassified, providing a Brier score of 25.54 and overall accuracy of only 55% (**Figure 3.17**). The classifier was slightly more accurate when predicting the very low OA patients (20/30, 60%), than other OA patients (13/30, 43.3%). This result was unexpected, as immunophenotyping was performed prospectively using PB-MNCs at diagnosis for both the training (n = 34) and test (n = 60) sets, which should eliminate any bias resulting from the different clinical trials in which the patients (TIDEL II or ENESTxtnd) were enrolled.

The most likely reason that the classifier failed is due to the heterogeneity of the MNC compartment within patient samples, which is comprised primarily of a mixture of granulocytes at various stages of differentiation in CML patients.[243] However, the actual cell populations/cell ratios within the MNC compartment can vary greatly between patients, making identification of consistent markers amongst a large patient cohort extremely difficult. Furthermore, when looking back at the development of the classifier, the classification accuracy was not as high as expected (88%) after internal validation (against the same samples used to develop the classifier; expected accuracy very close to 100%), indicating that there could be a chance of failure when independent validation was performed, which is what subsequently occurred.

Additionally, the patient cohort used to develop the classifier may not have been representative of the whole CP-CML cohort, as only a small number of patients (n = 34) were analysed due to restricted primary sample availability. Hence, it is possibile that the apparent statistical differences observed in the training set are not representative of true

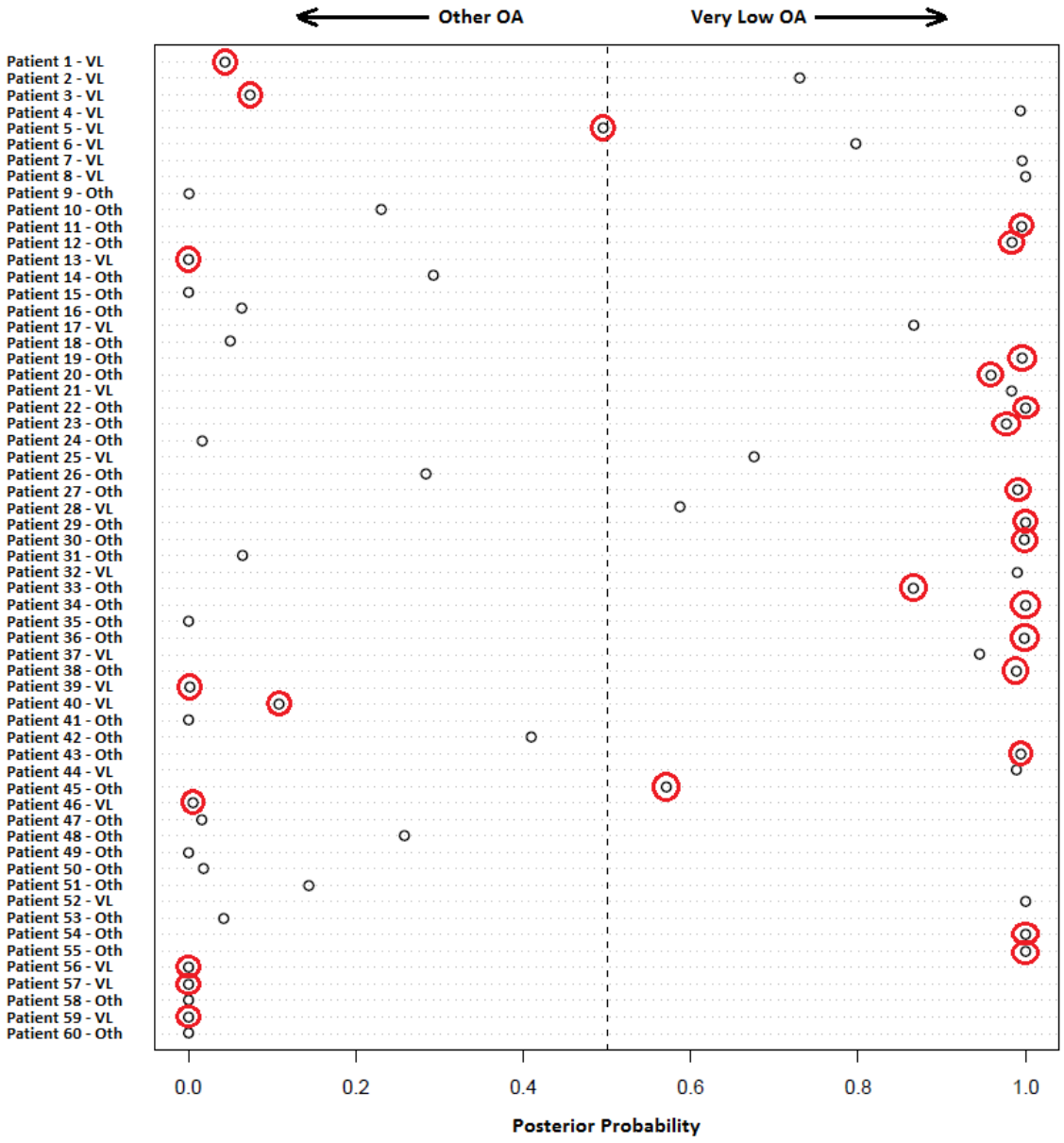


Figure 3.17 Validation of the 4-marker ungated percentage immunophenotyping classifier against an independent 60 patient test set

Results of the validation of the 4-marker classifier developed by iterativeBMA (n = 34 training set), tested against an independent 60 patient validation set. A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate an "other" OA patient. Twenty-seven patients were misclassified (**red** circles; 10 very low and 17 other OA patients), providing an overall accuracy of 55%, with a Brier score of 25.54. VL – very low OA (OA < 4 ng/200,000 cells); Oth – other OA (OA > 4 ng/200,000 cells).

differences when applied to a larger datatset (e.g. the test set). To investigate this further, unsupervised hierarchical clustering based on the expression of the 4 markers selected in the ungated percentage classifier was performed (**Figure 3.18**). The training (n = 34) and test (n = 60) patient cohorts did not appear to cluster independently of each other; however, when the number of immunophenotyping markers was increased to include all common markers analysed between both datasets, the training and test cohorts did appear to cluster separately (**Figure 3.19**). This data indicates that for an unknown reason the cellular composition of the MNC compartment appears to be different between the training and test patient cohorts, which has contributed to the failure of the classifier. Although the immunophenotyping data was unable to produce an accurate classifier for OA, it can still be used to identify single markers that may help to differentiate very low OA patients from the rest.

3.2.4 Combined 95 patient immunophenotyping analysis identifies markers associated with very low OA

The immunophenotyping data from both the 35 patient training cohort (now 35 patients as an OA value was determined for an extra patient) and the 60 patient validation cohort were combined to generate a larger, 95 patient dataset. This dataset was then investigated to determine if any cell surface markers were significantly differentially expressed for the following criteria: OA; the achievement of early molecular response (EMR; 3 month BCR-ABL1 mRNA levels < 10%); and the achievement of 12 month major molecular response (MMR; BCR-ABL1 mRNA levels < 0.1%). However, it is important to note that not all cell surface markers were analysed for all patients, due to sample and antibody availability.

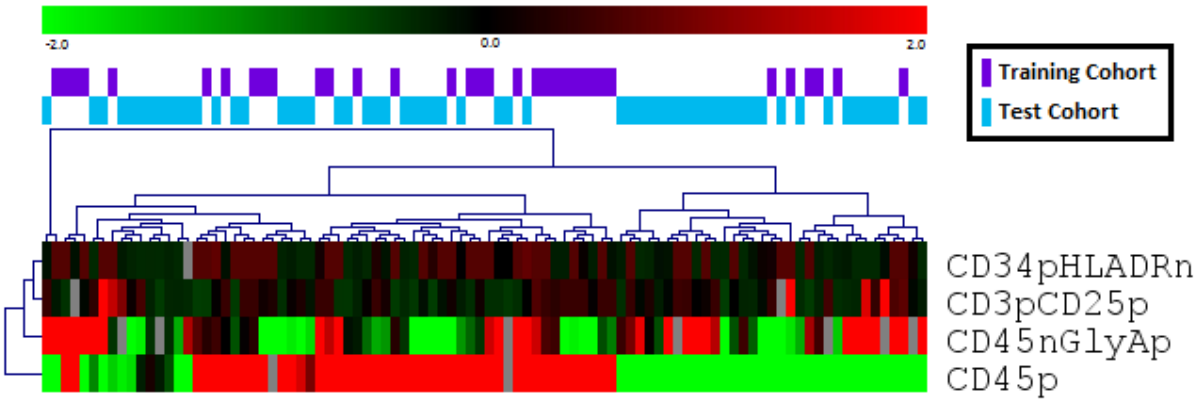


Figure 3.18 Hierarchical clustering of the 4 markers identified from the iterativeBMA ungated percentage dataset classifier

Unsupervised hierarchical clustering (Euclidean distance, complete linkage) was performed using the 4 markers identified in the predictive classifier developed by iterativeBMA using the ungated percentage immunophenotyping dataset. Each column represents one sample, and each row represents one marker. **Red** colour indicates overexpression compared to the median expression for that marker, and **green** colour represents low expression. Hierarchical clustering was performed using MultiExperiment Viewer (MeV v4.8, http://www.tm4.org/mev/).

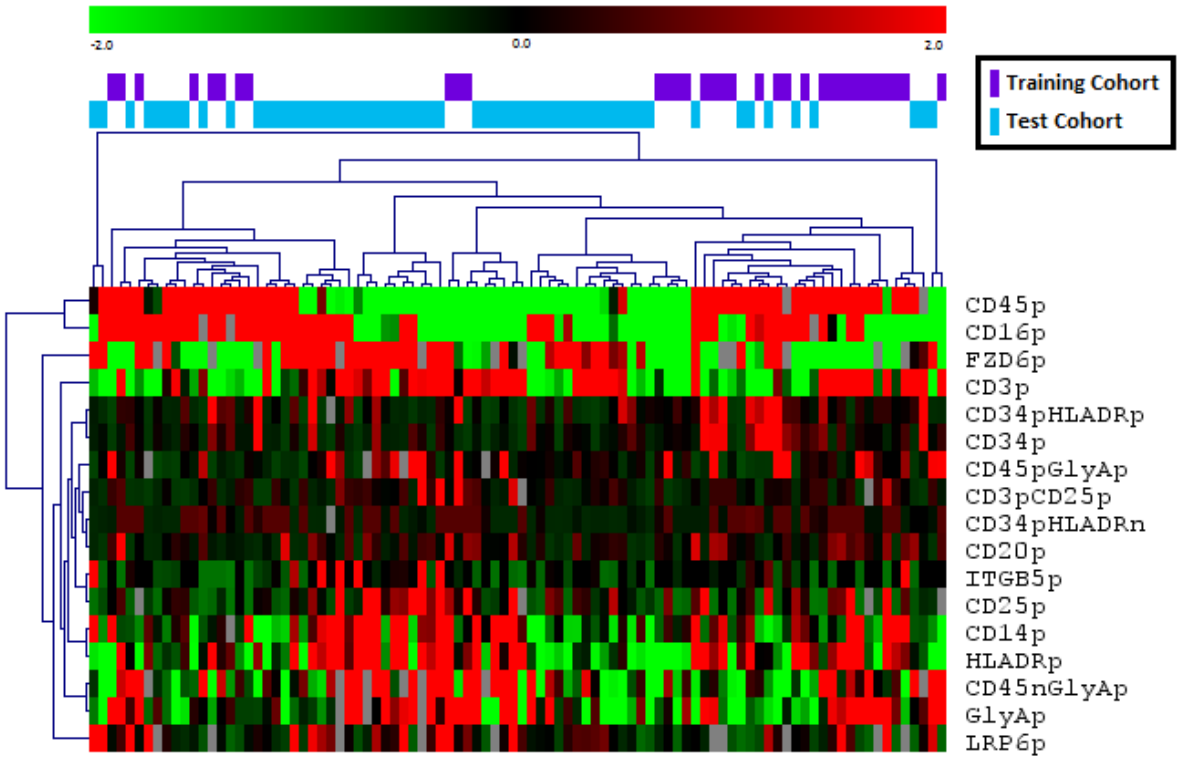


Figure 3.19 Hierarchical clustering of all the common markers between the 34 patient training immunophenotyping dataset and the 60 patient validation set

Unsupervised hierarchical clustering (Euclidean distance, average linkage) was performed using all the common markers identified from the ungated percentage immunophenotyping dataset for both the 34 patient training set and the 60 patient test/validation set. Each column represents one sample, and each row represents one marker. **Red** colour indicates overexpression compared to the median expression for that marker, and **green** colour represents low expression. Hierarchical clustering was performed using MultiExperiment Viewer (MeV v4.8, http://www.tm4.org/mev/).

3.2.4.1 Analysis of the 95 patient immunophenotypic percentage datasets

Differences in OA were assessed first in this large cohort, as this was the initial aim of the immunophenotyping analysis. Using the total ungated percentage dataset, 3 markers were differentially expressed between the very low and other OA patients (**Figure 3.20**), while 9 markers were identified as differentially expressed between very low and other OA patients using the lymphocyte-gated percentage dataset (**Figure 3.21**). One marker (CD3+CD25-; T-cell) was common between both the ungated percentage and lymphocyte-gated percentage datasets, suggesting that a T–cell-associated phenotype was apparent in very low OA, compared to other OA patients.

Patient immunophenotype was then compared with molecular response (EMR and 12 month MMR). Firstly, EMR was evaluated, with 82 of the original 95 patients analysed having EMR data available for further analysis; including 76 patients who achieved EMR, while only 6 patients failed to achieve EMR. Seventeen cell surface markers were differentially expressed with a *P*-value < 0.05 in the ungated percentage dataset; 11 of which had significantly increased expression in CP-CML patients who achieved EMR, while 6 were significantly increased in patients who failed to achieve EMR (**Table 3.4**; **Figure 3.22**). Fifteen cell surface markers were differentially expressed with a *P*-value < 0.05 when the lymphocyte-gated percentage dataset was analysed. Three of these markers had significantly increased expression in CP-CML patients who achieved EMR, while 8 were significantly increased in patients who failed to achieve EMR (**Table 3.5**; **Figure 3.23**).

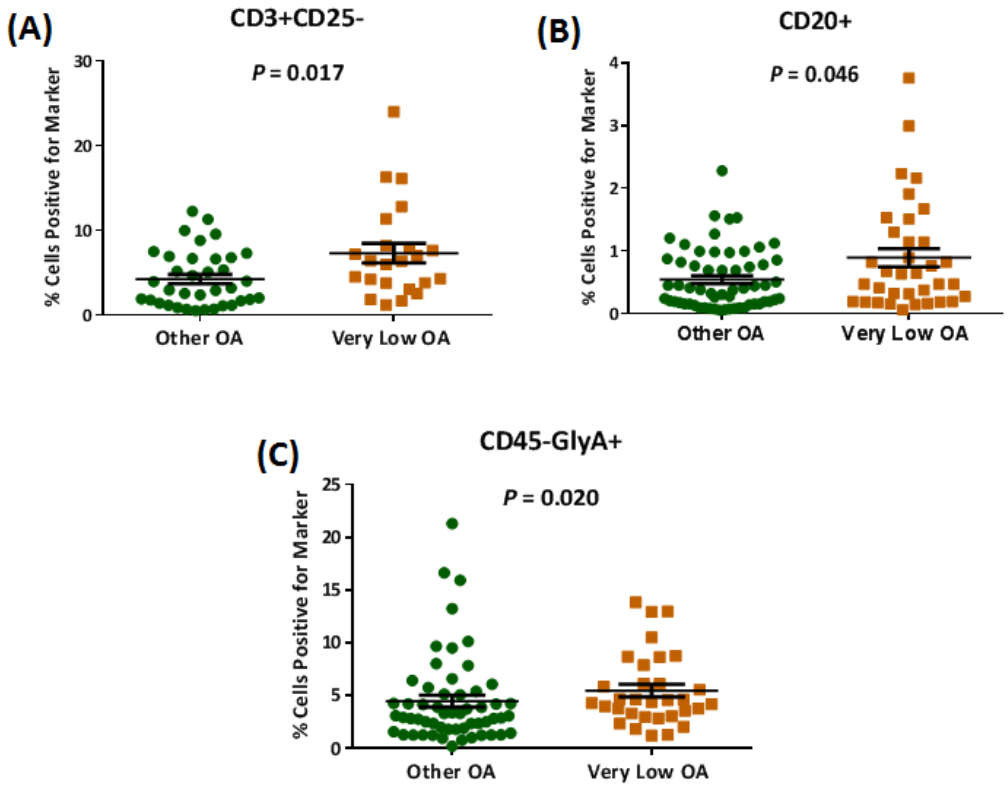


Figure 3.20 Significant cell surface marker expression between 95 very low and other OA CP-CML patients, using total ungated percentage immunophenotyping data

The difference in the percentage of cell surface marker expression for **(A)** CD3+CD25-, **(B)** CD20+ and **(C)** CD45-GlyA+ between very low and other OA patients is plotted using total ungated percentage data. Scatterplots demonstrate the mean percentage (%) of cells positive ± SEM. Mann-Whitney tests were performed as appropriate in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells. GlyA – glycophorin A.

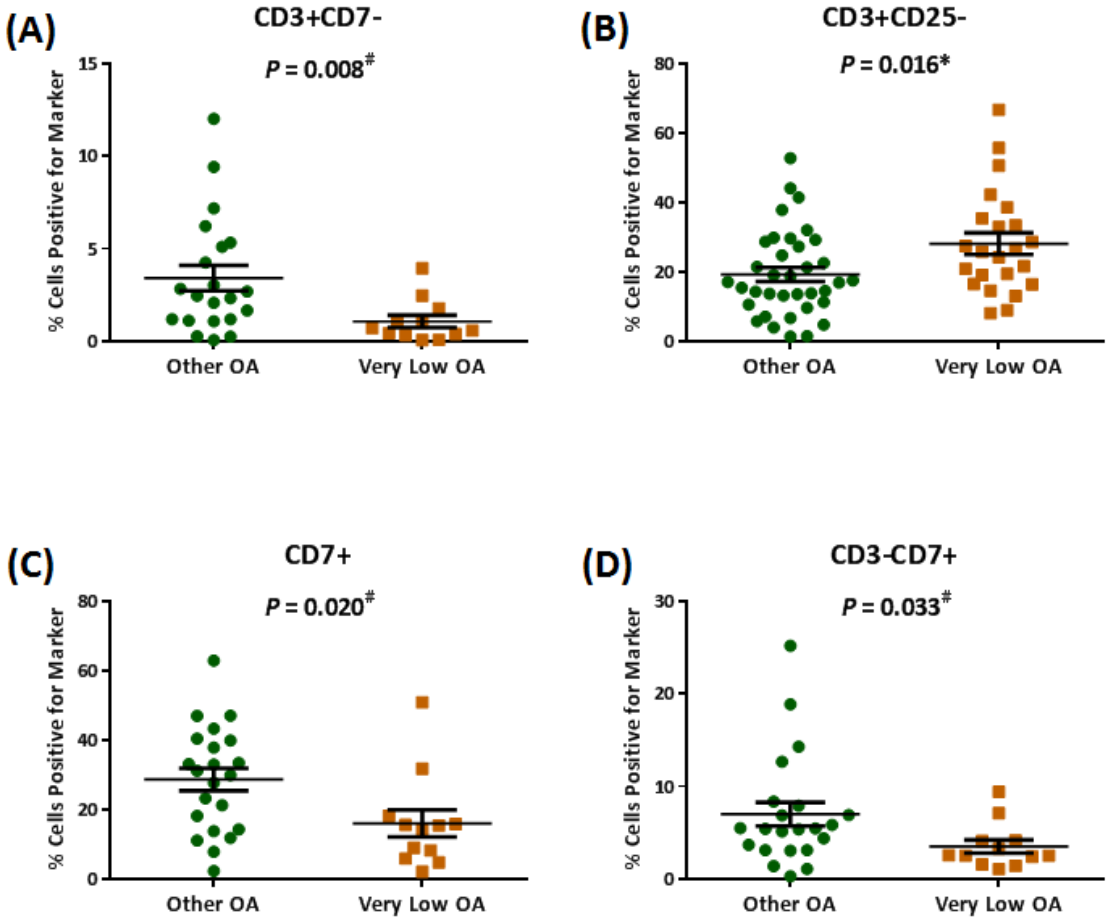


Figure 3.21 Significant cell surface marker expression between 95 very low and other OA CP-CML patients, using lymphocyte-gated percentage immunophenotyping data

The difference in the percentage of cell surface marker expression for **(A)** CD3+CD7-, **(B)** CD3+CD25-, **(C)** CD7+, **(D)** CD3-CD7+, **(E)** CD38+, **(F)** CD8+, **(G)** CD45+CD36-, **(H)** CD14-CD16+, and **(I)** CD45+ between very low and other OA patients is plotted using lymphocyte-gated percentage data. Scatterplots demonstrate the mean percentage (%) of cells positive ± SEM. Student's *t*-tests (*) or Mann-Whitney tests (#) were performed as appropriate in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells.

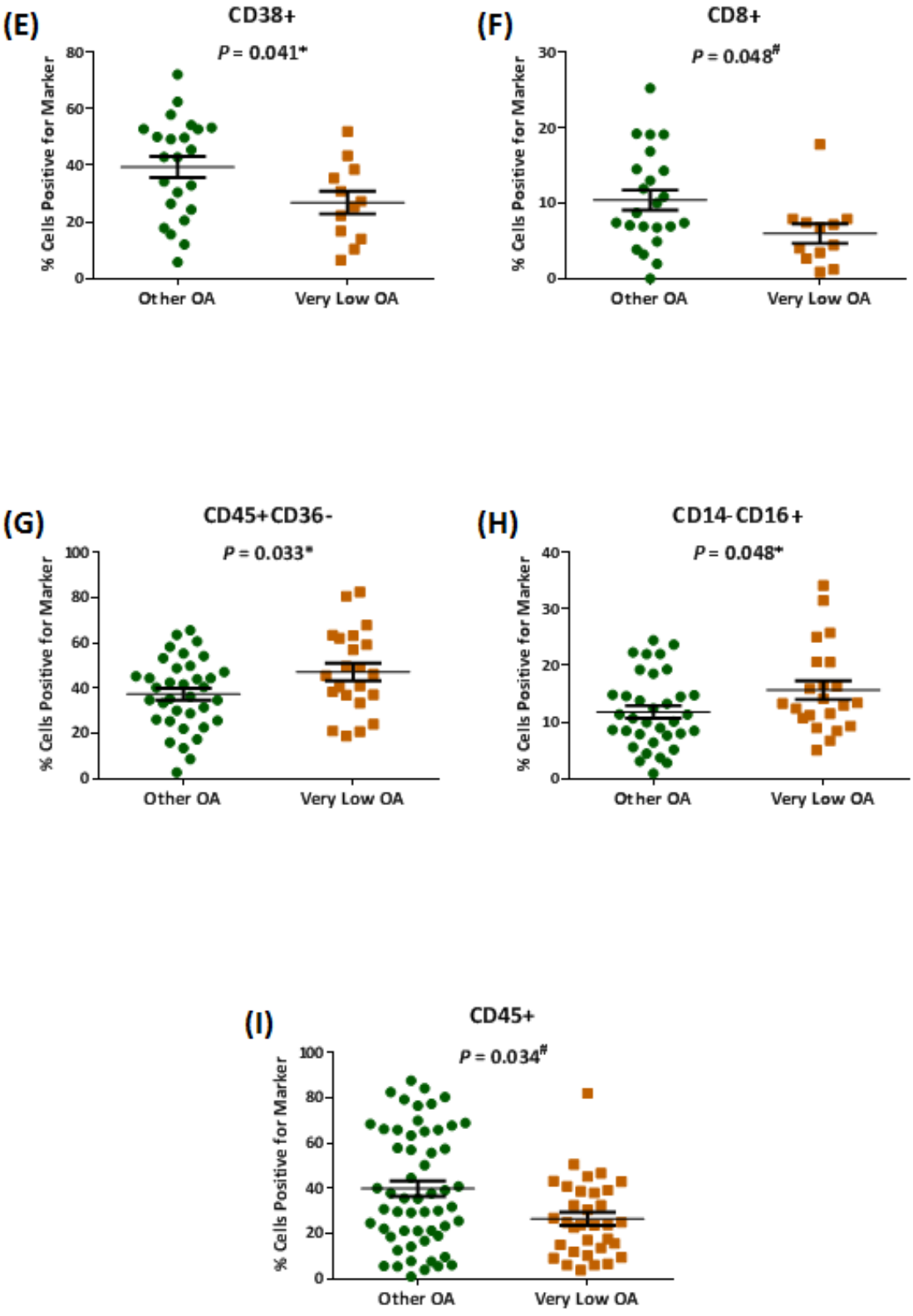

(E)
CD38+
P = 0.041*
% Cells Positive for Marker
0
20
40
60
80
Other OA
Very Low OA
(F)
CD8+
P = 0.048#
% Cells Positive for Marker
0
10
20
30
Other OA
Very Low OA
(G)
CD45+CD36-
P = 0.033*
% Cells Positive for Marker
0
20
40
60
80
100
Other OA
Very Low OA
(H)
CD14-CD16+
P = 0.048*
% Cells Positive for Marker
0
10
20
30
40
Other OA
Very Low OA
(I)
CD45+
P = 0.034#
% Cells Positive for Marker
0
20
40
60
80
100
Other OA
Very Low OA

Table 3.4 Cell surface markers with differential expression in CP-CML patients based on the achievement of EMR using the total ungated percentage immunophenotyping dataset

Cell Surface Marker	Cell Specificity	*P*-value	Up-regulated in patients who:
CD20+	B cell	0.019#	Achieve EMR
FZD6+	N/A	< 0.001#	Achieve EMR
CD34+	Haematopoietic progenitor cell	0.007#	No EMR
CD34+CD38-	Primitive haematopoietic progenitor cell	0.047#	No EMR
CD34+HLADR+	Haematopoietic progenitor cell	0.016#	No EMR
CD133+	Haematopoietic progenitor cell	0.006*	No EMR
CD34+CD133+	Haematopoietic progenitor cell	0.040#	No EMR
CD7+	T-cell precursor	0.043#	Achieve EMR
CD3+	T-cell	0.018#	Achieve EMR
CD3+CD25+	Activated T-cell	0.037#	Achieve EMR
CD3+CD7+	T-cell	0.040*	Achieve EMR
CD8+	Cytotoxic T-cell	0.030*	Achieve EMR
CD4+	T-helper cell	0.049*	Achieve EMR
CD4+CD8-	T-helper cell	0.023*	Achieve EMR
CD4-CD8+	Cytotoxic T-cell	0.036*	Achieve EMR
GlyA+	Erythrocyte	< 0.001#	Achieve EMR
CD45+	Pan-leukocyte	0.003#	No EMR

Abbreviations: FZD6, frizzled 6; GlyA, glycophorin A; N/A, not applicable; and EMR, early molecular response. *P*-values were calculated using the Student's *t*-test (*) or Mann-Whitney test (#), as appropriate in GraphPad Prism 6.

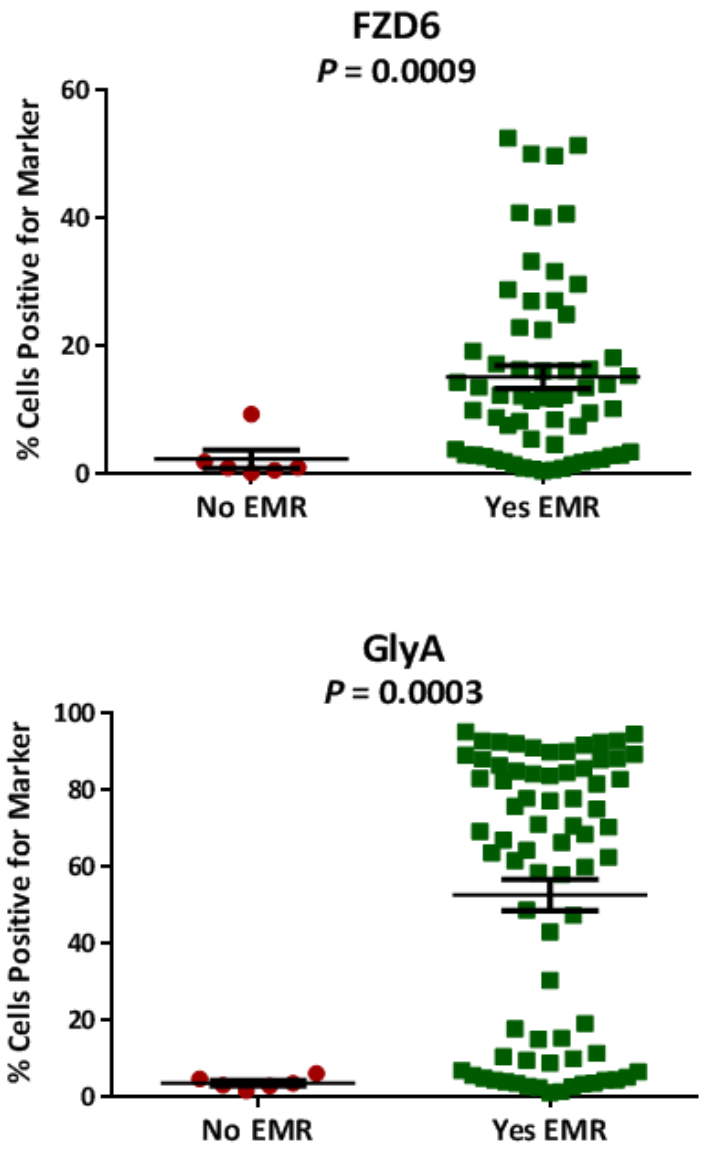


Figure 3.22 Significant cell surface marker expression between CP-CML patients who did, or did not, achieve EMR using ungated percentage immunophenotyping data

The difference in cell surface marker expression for CP-CML patients who achieved EMR, versus those who did not, using ungated percentage data is plotted for FZD6 and GlyA. Scatterplots demonstrate the mean percentage (%) of cells positive ± SEM. Mann-Whitney tests were performed using GraphPad Prism 6 as appropriate.

Table 3.5 Cell surface markers with differential expression in CP-CML patients based on the achievement of EMR using the lymphocyte-gated percentage immunophenotyping dataset

Cell Surface Marker	Cell Specificity	*P*-value	Up-regulated in patients who:
CD20+	B cell	0.039#	Achieve EMR
CD63+	Platelet	0.045*	No EMR
FZD6+	N/A	0.011#	Achieve EMR
CD34+	Haematopoietic progenitor cell	0.002#	No EMR
HLADR+	Immune cells	0.024#	No EMR
CD34+CD38+	Haematopoietic progenitor cell	0.007#	No EMR
CD34+CD38-	Primitive haematopoietic progenitor cell	0.013#	No EMR
CD34+HLADR+	Haematopoietic progenitor cell	< 0.001#	No EMR
CD133+	Haematopoietic progenitor cell	0.001#	No EMR
CD34+CD133+	Haematopoietic progenitor cell	0.002#	No EMR
CD4+CD25+	Regulatory T-cell	0.020*	Achieve EMR
CD15+	Granulocyte	0.021#	No EMR
CD15+CD16+	Neutrophil	0.016#	No EMR
CD15+CD16-	Granulocyte	0.003#	No EMR
CD14+	Monocyte	0.009#	No EMR

Abbreviations: FZD6, frizzled 6; N/A, not applicable; and EMR, early molecular response. *P*-values were calculated using the Student's *t*-test (*) or Mann-Whitney test (#), as appropriate in GraphPad Prism 6.

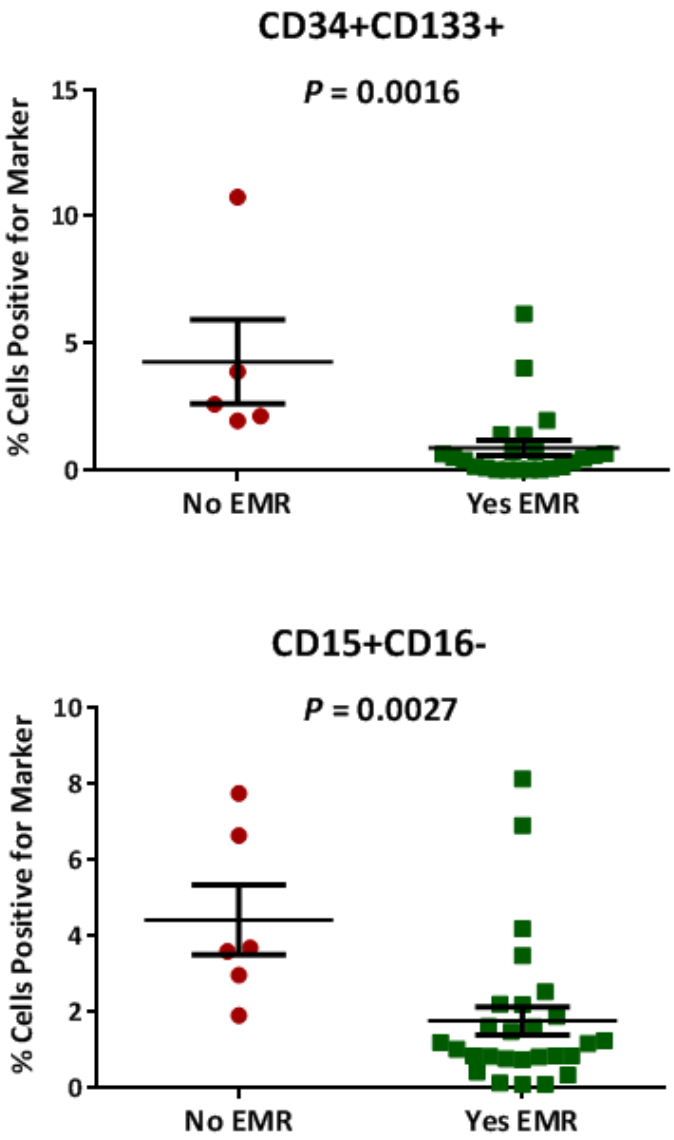


Figure 3.23 Significant cell surface marker expression between CP-CML patients who did, or did not, achieve EMR using lymphocyte-gated percentage immunophenotyping data

The difference in cell surface marker expression for CP-CML patients who achieved EMR, versus those who did not, using lymphocyte-gated percentage data is plotted for CD34+CD133+ and CD15+CD16-. Scatterplots demonstrate the mean percentage (%) of cells positive ± SEM. Mann-Whitney tests were performed using GraphPad Prism 6 as appropriate.

The achievement of 12 month MMR was then evaluated, with 76 CP-CML patients having 12 month MMR data available for further analysis; including 57 patients who achieved MMR, and 19 patients who did not. Two cell surface markers were differentially expressed with a *P*-value < 0.05 in the lymphocyte-gated percentage dataset; 1 with significantly increased expression in CP-CML patients who achieved MMR by 12 months, and the other significantly increased in CP-CML patients who failed to achieve 12 month MMR (**Figure 3.24**). Although significant, neither of these two markers would be able to confidently differentiate patients who achieved 12 month MMR, from those who did not. The degree of confidence is low because considerable overlap in cell surface marker expression occurs between both the patients who did, or did not achieve MMR by 12 months, meaning that a distinct expression cut-off value was unable to be defined to differentiate the patients. Intriguingly, no significant differential cell surface marker expression was observed in relation to 12 month MMR achievement when the ungated percentage, ungated MFI or lymphocyte-gated MFI datasets were analysed. As only two surface markers were identified as significant from the lymphocyte-gated percentage dataset, yet confidence in this result was low as explained previously, it was therefore not surprising that no significant markers were identified from the lymphocyte-gated MFI dataset. Furthermore, this result actually supports the lack of confidence expressed towards the results from the lymphocyte-gated percentage dataset. Taken together, these results indicate that patient immunophenotype (cell surface marker expression), is unlikely to be associated with the achievement of MMR by 12 months on TKI therapy.

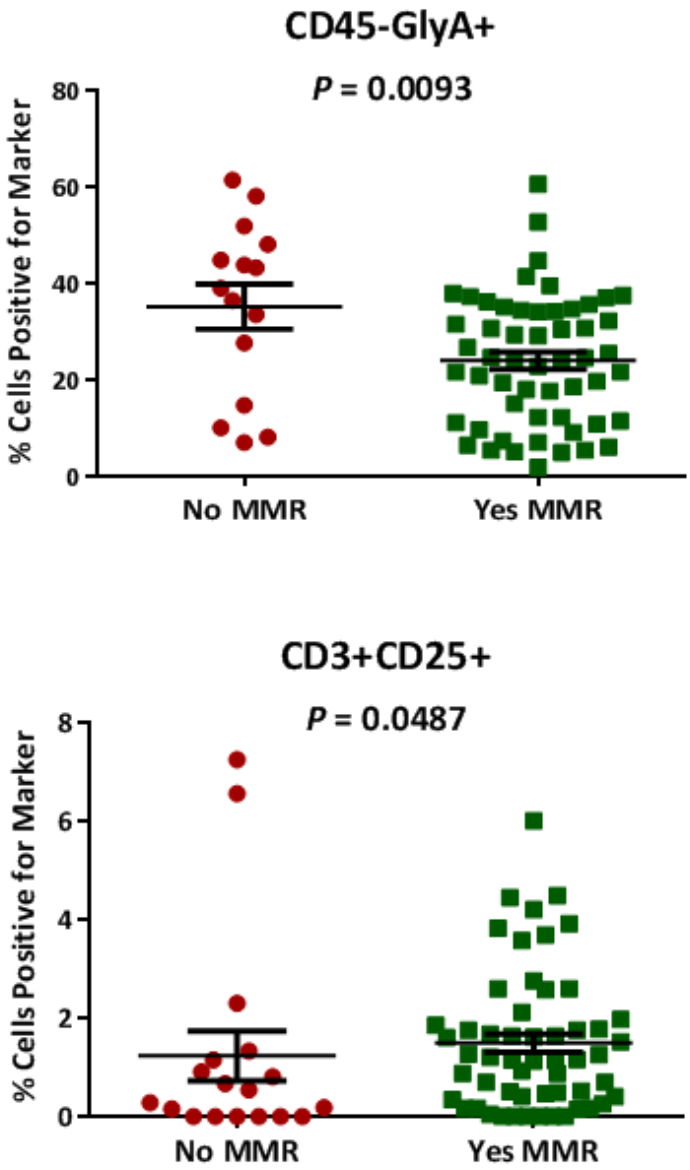


Figure 3.24 Significant cell surface marker expression between CP-CML patients who did, or did not, achieve 12 month MMR using lymphocyte-gated percentage data

The difference in cell surface marker expression for CP-CML patients who achieved MMR by 12 months, versus those who did not, using lymphocyte-gated percentage data is plotted for the CD45-GlyA+ and CD3+CD25+. Scatterplots demonstrate the mean percentage (%) of cells positive ± SEM. Student's *t*-tests or Mann-Whitney tests were performed using GraphPad Prism 6 as appropriate.

3.2.4.2 Analysis of the 95 patient immunophenotypic MFI datasets

Only 1 marker (ABCB1; **Figure 3.25**) was differentially expressed between the very low and other OA patients using the total ungated MFI dataset. There was no overlap between any of the markers identified from either of the total ungated datasets (percentage or MFI). Four markers were differentially expressed between very low and other OA patients (**Figure 3.26**) when the lymphocyte-gated MFI dataset was analysed. Interestingly, CD7 expression (another T-cell marker) was significantly increased in other OA patients, compared to very low OA patients, in both the lymphocyte-gated percentage and MFI datasets. Furthermore, expression of the T-lymphocyte marker CD3+CD7- was also increased in other OA patients, compared to very low OA patients, in the lymphocyte-gated percentage dataset. This data suggests that in the lymphocyte-gated cell population from CP-CML patients, the T-lymphocyte population is quite varied between OA sub-groups (very low versus the rest).

Patient immunophenotype was then compared to response, as before. Only a single cell surface marker, FZD6, was identified as having significantly increased expression in CP-CML patients who achieved EMR (**Figure 3.27**), compared to those patients who did not from the ungated MFI dataset. The lymphocyte-gated MFI dataset was interrogated next, with 2 cell surface markers differentially expressed with a *P*-value < 0.05 between patients who did, or did not, achieve EMR (**Figure 3.28**). Frizzled 6 (FZD6) was the only marker consistently differentially expressed across all four datasets (both percentage and both MFI datasets). Interestingly, FZD6 expression was significantly increased in patients who did achieve EMR, compared to those who did not in all four datasets, indicating that elevated FZD6 may identify patients likely to respond well to TKI therapy. However, as the cohort of patients who did not achieve EMR was very small (n = 6), these results are only preliminary.

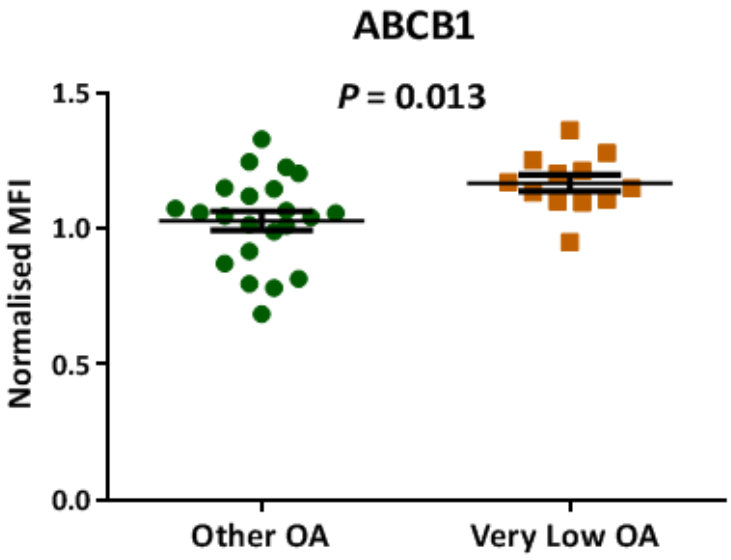


Figure 3.25 Significant ABCB1 cell surface expression between 95 very low and other OA CP-CML patients, using total ungated MFI immunophenotyping data

The difference between the MFI of the cell surface marker ABCB1 in very low and other OA patients, is plotted using total ungated MFI data. Scatterplot demonstrates the normalised MFI of cells ± SEM. Student's *t*-test was performed in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells.

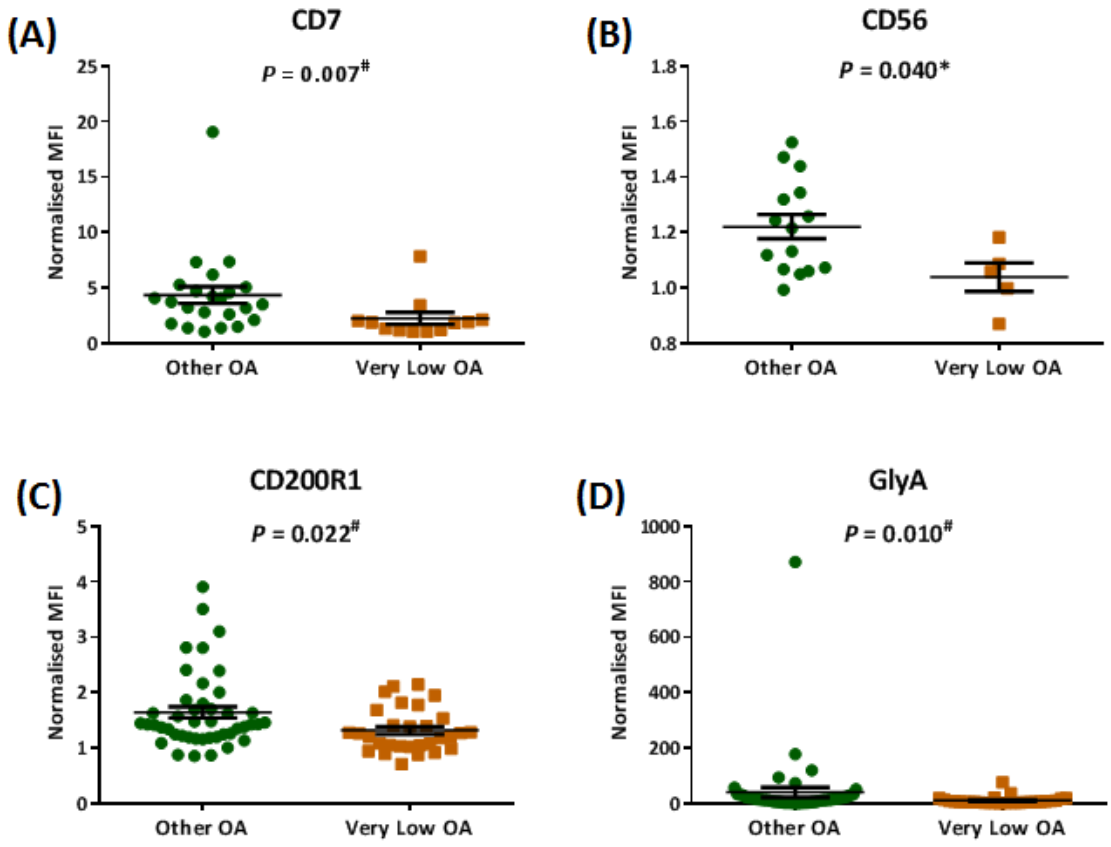


Figure 3.26 Significant cell surface marker expression between 95 very low and other OA CP-CML patients, using lymphocyte-gated MFI immunophenotyping data

The difference between the MFI of the cell surface markers **(A)** CD7, **(B)** CD56, **(C)** CD200R1, and **(D)** GlyA in very low and other OA patients, is plotted using lymphocyte-gated MFI data. Scatterplot demonstrates the normalised MFI of cells ± SEM. Student's *t*-tests (*) or Mann-Whitney tests (#) were performed as appropriate in GraphPad Prism 6. Other OA represents all patients with OA > 4 ng/200,000 cells. GlyA – glycophorin A.

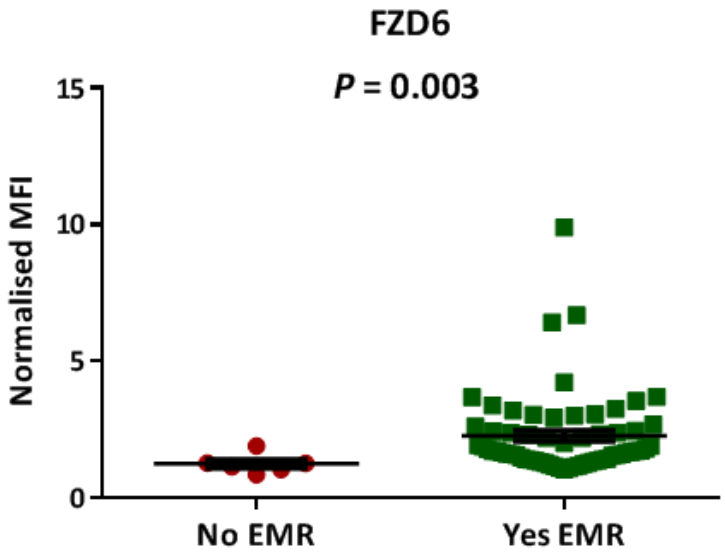


Figure 3.27 Significant FZD6 cell surface expression between CP-CML patients who did, or did not, achieve EMR using ungated percentage MFI data

The difference in FZD6 cell surface marker expression for CP-CML patients who achieved EMR (n = 62), versus those who did not (n = 6) using ungated normalised MFI data is plotted. Scatterplot demonstrates the normalised MFI of cells ± SEM. A Mann-Whitney test was performed using GraphPad Prism 6.

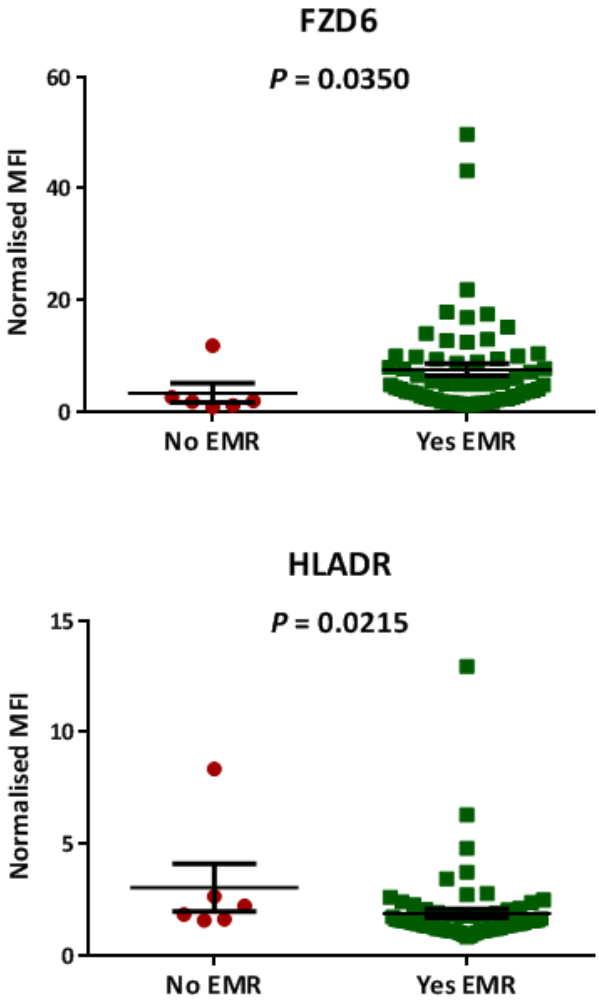


Figure 3.28 Significant cell surface marker expression between CP-CML patients who did, or did not, achieve EMR using lymphocyte-gated MFI data

The difference in cell surface marker expression for CP-CML patients who achieved EMR, versus those who did not, using lymphocyte-gated MFI data is plotted for FZD6 and HLADR. Scatterplots demonstrate the normalised MFI of cells ± SEM. Mann-Whitney tests were performed using GraphPad Prism 6 as appropriate.

Furthermore, while providing insight into some of the possible immunophenotypic differences between these patients, analysis of a significantly larger CP-CML patient cohort will need to be performed to validate these findings. As stated before, no significant differential cell surface marker expression was observed in relation to 12 month MMR achievement when the ungated MFI or lymphocyte-gated MFI datasets were analysed.

3.3 Discussion

The OCT-1 activity (OA) assay, developed in our laboratory,[151] is a functional assay measuring imatinib uptake at diagnosis in CML patient MNCs, and has excellent predictive value for molecular response at diagnosis.[142,151] The OA assay is one of the best prognostic markers for patients at risk of poor response, as patients with very low OA (OA < 4 ng/200,000 cells) are at significant risk for poor response (lower MMR, event-free survival, transformation-free survival and overall survival), compared to all other patients.[142] However, the OA assay is performed using [^{14}C]-imatinib, live cells and specialised equipment, making the applicability of the assay in diagnostic laboratories worldwide quite difficult; hence, the assay has not been adopted for routine testing. Recently, OA has been demonstrated to be strongly related to cell lineage, with the highest OA measured in pure populations of neutrophils, followed by monocytes, and lowest in the lymphocyte cell population.[243] This data strongly linked the predictive power of OA, as measured in CP-CML patient MNCs, to the cellular composition of the original PB sample from which the MNCs were derived. Therefore, the observed variation between OA and subsequent disease outcome in patients, combined with this link between OA and cell lineage, highlights the importance of defining molecular and cellular characteristics associated with low and high OA CP-CML patients at diagnosis. Furthermore, immunophenotypic analysis of CML patient MNCs at diagnosis may allow for a better understanding of the factors underlying the variation in OA levels observed in CP-CML patients.

The findings described in this chapter suggest that small, yet distinct, immunophenotypic lineage differences are present between CP-CML patients with very low OA, compared to the rest. Namely, in the initial 27 patient analyses, decreased expression of the granulocyte and neutrophil markers CD15 and CD66b, and the pan-leukocyte marker CD45 was observed

in very low OA patients compared to the rest, in the total ungated MNC percentage dataset. This dataset best represents the cell population used to perform the OA assay (thawed MNCs), and in keeping with the pattern of OA levels (neutrophils > monocytes > lymphocytes) observed by Engler *et al.*[243] demonstrates that a decreased percentage of neutrophils in the very low OA patients could be contributing to their low OA. Unfortunately, this marker combination was no longer significant when the patient cohort was increased to include 95 CP-CML patients, as discussed further below.

The 35 patient total ungated percentage immunophenotyping dataset, described in this chapter, was able to develop a 4-marker predictive classifier for OA. Unfortunately, this classifier could not be validated using an independent 60 CP-CML patient cohort. This was a somewhat unexpected result, as immunophenotyping was performed prospectively using MNCs at diagnosis for both the training (n = 34) and test (n = 60) sets, effectively eliminating any bias resulting from the different clinical trials in which the patients (TIDEL II or ENESTxtnd) were enrolled. As mentioned previously, heterogeneity of the MNC compartment in CP-CML patients, making identification of cell surface markers/lineages specific to a certain OA group extremely difficult, is the most likely reason that the classifier failed. However, other possibilities to consider include that other confounding effects are more powerful than cell surface markers (e.g. the markers will detct cell type well, but there is too great a range of OA within the cell subpopulations). Also, as the initial patient cohort used for the training set of the classifier was quite small (n = 34), this cohort may not have been a good representation of the "universal" CP-CML patient cohort, with some of the statistical associations observed in the training set representing random, rather than true, associations. Furthermore, unsupervised hierarchical clustering suggested that the training

and test cohorts clustered independently of one another when all common immunophenotyping markers were analysed.

This data suggests that for an unknown reason the immunophenotypes identified from the MNC compartment of both the training and test cohorts were different. As each patient cohort (training and test) was analysed at distinct times, the observed differences may be antibody related; as antibodies, which have slight variation themselves, were changed as necessary between both cohorts (i.e. new antibody purchased when current antibody ran out). However, controls were in place to minimise this effect, including keeping the instrument voltage and gain settings as consistent as possible between runs, while isotype control antibodies, matched to each primary antibody's host species and isotype, were performed for every run. The results from these isotype control antibodies were then used to normalise the MFI values and identify cells that were positive for a particular cell surface marker, specific to each patient sample. Furthermore, optimal data compensation was performed by using the "fluorescence minus one" (FMO) control staining procedure, where patient cells were stained with all antibodies except the one of interest, in order to determine the exact range of the negative population.[276] This was performed for all antibodies analysed in the immunophenotyping panels to ensure consistent compensation between all patients.

Patients enrolled to the ENESTxtnd trial (comprising the test cohort) were able to have up to 1 month of imatinib treatment, before being screened for trial eligibility (samples that were analysed by immunophenotyping) and commencing nilotinib therapy, under the clinical trial guidelines. Although this was unlikely to be a common occurrence, the exact number of patients who received TKI therapy prior to be screened for trial commencement is

unfortunately unknown, with this data not available for analysis (restricted as part of the locked clinical trial database). Furthermore, approximately 59% (20/34; 6 very low OA) of TIDEL II patients (training set) received hydroxyurea therapy prior to screening for trial commencement, with the effect of this to their MNC compartment also unknown. Differences relating to the enrolment site for each patient could also have impacted on the MNC composition, as patients treated here in Adelaide would have PB taken and processed on the same day, while samples taken at sites interstate and overseas (New Zealand) would be delayed by 1-2 days for travel time, prior to processing. However, only 15% (9/60: test set) of ENESTxtnd and 29% (10/34; training set) of TIDEL II patient samples were from Adelaide, indicating that this reason was unlikely to be influencing the difference in MNC composition observed between the two cohorts, as the majority of samples were from interstate/overseas. Taken together, this data provides further evidence that CP-CML is not a homogeneous disease, but rather, significant heterogeneity exists between the cellular compositions of patients diagnosed in CP-CML.

The data generated from the immunophenotyping datasets presented here did however, generate some interesting results. Analysis of the combined patient cohorts, generating a 95 patient cohort, identified an differential T-lymphocyte (CD3+CD25-), B-cell (CD20) and erythrocyte (CD45-GlyA+) signature associated with very low OA, compared to the rest, using the ungated percentage immunophenotyping data (**Figure 3.20**). However, this cell surface marker signature is quite different to the results from the 27 patient signature which included CD15, CD66b (granulocyte and neutrophil) and CD45 (leukocyte), as discussed earlier. Hence, of the significant markers identified between CP-CML patients with very low OA, compared to the rest from the initial 27 patient analysis (all 4 datasets), only 1 cell surface marker (CD38+, lymphocyte-gated percentage dataset) remained significant when

the combined 95 patient analysis was performed. Once again, this data provides further support for the suggestion that significant heterogeneity exists between CP-CML patients at diagnosis, as when the patient cohort was quite large (n = 95), the results were vastly different from the earlier sub-group (n = 27) analysis. These results also prevented independent analysis of a combination of significant markers from the initial 27 patient analysis to identify distinct immunophenotypes associated with patients with very low OA, compared to the rest, in place of a classifier from being performed.

Further analysis of the ungated percentage dataset did identify 17 markers significantly associated with the achievement of EMR. Increased expression of haematopoietic progenitor cell (HPC) markers were associated with patients not achieving EMR, while increased expression of a significant T-lymphocyte, B-lymphocyte and erythrocyte signature was associated with the achievement of EMR (**Table 3.4**). However, the most interesting result related to FZD6 expression, which was significantly increased in patients who did achieve EMR compared to those who did not, in all four datasets (ungated percentage, ungated MFI, lymphocyte-gated percentage and lymphocyte-gated MFI).

FZD6 is a member of the “frizzled” gene family, which encode 7-transmembrane domain proteins that are receptors for Wnt signalling proteins. Wnt/Fzd binding is known to regulate an array of downstream signals affecting development, tissue-specific stem cell renewal, and tumourigenesis, particularly through the canonical pathway involving stabilisation of β-catenin.[277] Recently, Wnt/β-catenin signalling has been demonstrated to be involved in protecting CML stem cells during TKI therapy, leading to leukaemic stem cell preservation.[278,279] Furthermore, Wnt/Ca(2+)/NFAT (non-canonical Wnt pathway) signalling was demonstrated to maintain survival of Bcr-Abl1+ leukaemic cells after the inhibition of

Bcr-Abl1 by TKIs.[280] However, no direct role for FZD6 has been established in CML. Although, the FZD6 protein has been suggested to be a receptor for the WNT4 ligand, involved in the non-canonical Wnt pathway,[281] and it has been demonstrated to act as a negative regulator of the canonical Wnt/β-catenin signalling cascade.[282] The data presented here indicates that elevated FZD6 expression was observed in CP-CML patients likely to respond well to TKI therapy (EMR). How these elevated FZD6 levels affect the Wnt/β-catenin signalling pathway, thought to be important to CML stem cell preservation, remain to be determined and are now the focus of ongoing investigation. However, it is important to note that as the cohort of patients who did not achieve EMR was very small (n = 6), these results are only preliminary and analysis of a significantly larger CP-CML patient cohort will need to be performed to validate these findings.

Intriguingly, no significant difference in cell surface marker expression associated with 12 month MMR was observed in the ungated percentage dataset, where the lymphocyte-gated percentage dataset was the only dataset to identify markers with significant differential expression between patients who did, or did not, achieve 12 month MMR. Increased expression of the activated T-cell marker CD3+CD25+ was observed in patients achieving MMR, while an increase in erythrocyte (CD45-GlyA+) expression was associated with patients not achieving MMR. Unfortunately, the achievement of 24 month MMR was unable to be assessed, as the majority of patients analysed were yet to reach this milestone.

Finally, the decreased T-lymphocyte signature (including the markers CD3+CD7-, CD7+, CD3-CD7+ and CD8+) associated with very low OA when compared to the rest in the 95 patient lymphocyte-gated percentage dataset, suggests a possible role for immune surveillance in OA. In particular, a decreased immune response in CP-CML patients with very low OA could

possibly have a role in the poor response to TKI therapy observed for these patients, as the immune system has an important role in the regulation of cancer development. It can both antagonize and enhance tumour development and progression through a number of mechanisms involving both innate and adaptive immune cells.[283] Innate immune cells directly contribute to cancer via free radical-induced DNA damage and paracrine regulation of intracellular pathways (via NFκβ), while adaptive immune cells modulate cancer via direct inhibition of tumour growth through cytotoxic T-cell activity and cytokine-mediated lysis of tumour cells.[283] Furthermore, Hanahan and Weinberg recently altered their seminal "Hallmarks of Cancer" to include a new emerging hallmark of cancer – the active evasion of cancer cells from attack and elimination by immune cells,[46] indicating further the significant impact the immune system can have on cancer. Furthermore, an increased T- lymphocyte signature (including the markers CD7+, CD3+, CD3+CD25+, CD3+CD7+, CD8+, CD4+, CD4+CD8- and CD4-CD8+) was observed in the 95 patient ungated percentage dataset, and was associated with patients who achieved EMR. This data suggests that an active immune response is perhaps influencing the excellent early response observed in these patients, while the decreased T-lymphocyte signature observed in CP-CML patients with very low OA may contribute to their poor response by preventing natural immune surveillance from functioning correctly.

In conclusion, the data presented here demonstrates that specific lineage differences are present between CP-CML patients at diagnosis with very low OA, compared to the rest. The findings presented in this chapter provide additional support towards the role of cell lineage in influencing the OA assay, with further evidence discussed in **Chapter 4**. Furthermore, while conjecture exists, this data also provides further understanding of the underlying biology of OA, and indicates that CP-CML heterogeneity has a significant impact on OA

determination. While a predictive classifier for OA using immunophenotyping was unable to be generated, this analysis still provided important insight into possible cell biology underlying the OA results. Additionally, the identification of an increased HPC marker signature in patients who did not achieve EMR, suggests that these patients may possess a more immature cellular phenotype; while the increased T-lymphocyte signature associated with patients who did achieve EMR could possibly indicate that active immune surveillance is contributing to the improved response observed in these patients.

4 GENE EXPRESSION PROFILING IN CP-CML PATIENTS WITH LOW AND HIGH OCT-1 ACTIVITY

4.1 Introduction

Although studies consisting solely of newly diagnosed, previously untreated CP-CML patients found no correlation between disease response and patient OCT-1 mRNA expression,[142,151,152] global expression-based studies are excellent tools to understand the different biological characteristics associated with OCT-1 activity (OA). This is mainly due to high throughput nature of microarray analysis, which allows an extremely large number of genes to be analysed simultaneously in patient samples, generating a richness and depth of data that can be used to identify pathways and specific targets for further validation (e.g. protein studies). Therefore, it is proposed that this technology can also be used to elucidate a gene signature that differentiates very low OA (OA < 4 ng/200,000 cells) patients from the remainder of the patients (OA > 4 ng/200,000 cells). An mRNA based approach has already been applied for the diagnosis of CML,[284] making the technique readily available and translatable between laboratories. There are limitations to the OA assay itself, due to the requirement of [^{14}C]-imatinib, live cells and specialised equipment. Also, the expression-based approach enables a measurement of the cellular state associated with OA, allowing direct comparisons between patients, to be made based on the global gene expression patterns. This will also provide insights into the biology of CP-CML and possible targets for further functional analyses.

DNA microarrays are a small chip on which thousands of DNA sequences from different genes are attached at specific locations. Thus, microarray gene expression studies allow the activity of thousands of genes to be simultaneously analysed in a single experiment when DNA from a sample of interest is loaded onto the microarray. A standard microarray works by reverse-transcribing a given mRNA sample to complementary DNA (cDNA), using fluorescent (generally Cy3-dNTP or Cy5-dNTP) or radio-labelled deoxynucleotides ([^{33}P]- or [^{32}P]-α-dCTP; **Figure 4.1**).[285] Purification and denaturation steps follow, where the labelled

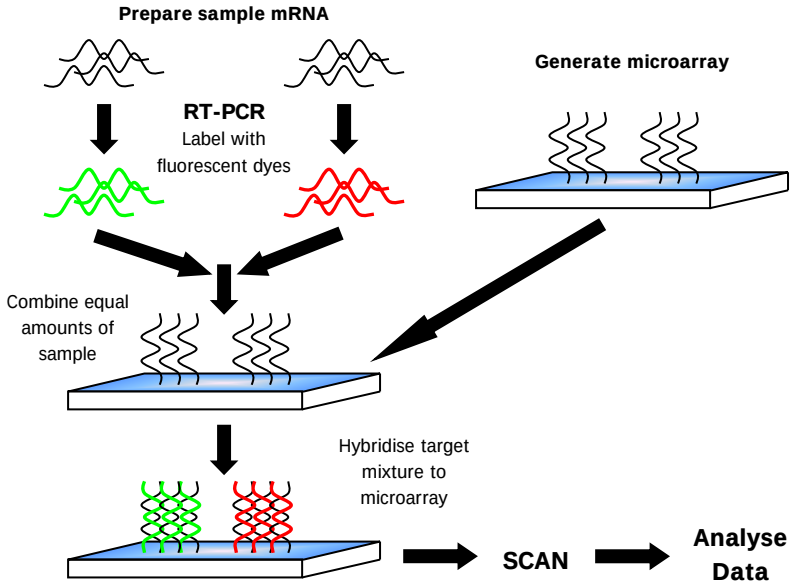


Figure 4.1 The steps involved in microarray preparation

The sample mRNA is initially reverse-transcribed to cDNA, using fluorescent dNTPs. The microarray plate is generated with the required gene set, and the sample is hybridised to the plate, in equal amounts. The microarray plate is then scanned for gene expression using computational software and the data analysed. Figure adapted from Microarray Centre.[286]

targets are hybridised to the microarray. The expression levels of the genes on the microarray are then determined using computational software that accurately measures the amount of mRNA bound to each spot on the microarray.[285] Before the data is analysed, it is normalised; a process that provides an approximation of the ratio of gene expression between the samples.[285] This step is crucial for obtaining high quality data.

4.1.1 Microarrays and CML

The understanding of CML disease biology has significantly increased with the introduction of microarray technology. Microarrays are now being used to investigate whether the differential expression of specific genes or gene expression patterns in patient samples at diagnosis is linked to the intrinsic heterogeneity observed in CML patient response. Ultimately, this approach may lead to the development of a screening assay using a selected gene-set to predict the response of patients treated with tyrosine kinase inhibitors (TKIs), such as imatinib. The ability to characterise a patient's leukaemic cells before starting treatment can provide important treatment information that can contribute to the decision regarding which TKI to administer (or to perform an allogeneic stem cell transplant), if the patient's response was predicted to be suboptimal to a particular TKI.

Gene expression analysis of primary CML cells by microarray has focused on three main areas: comparing *BCR-ABL1* positive cells with normal cells; correlating gene expression profiles with patient response; and correlating gene expression profiles with disease progression. The first CML-based microarrays compared the gene expression of *BCR-ABL1* positive cells with normal cells.[287-289] For example, Ohmine *et al.*[289] used CD133+ CML haematopoietic progenitor cells, Nowicki *et al.*[288] used mononuclear cells (MNCs), whilst Kronenwett *et al.*[287] used enriched CD34+ haematopoietic stem and progenitor cells. The use

of different cells types in these microarrays is a major confounding factor and as a result, there is not a good correlation between these studies. This highlights one of the caveats associated with comparing microarray results; the gene expression profiles generated relate directly to the cell type, and often array platform, used. Therefore, when comparing patient profiles, only microarrays using similar patient samples are able to be directly compared (e.g. peripheral blood (PB) samples cannot be compared against bone marrow (BM) samples, and MNCs cannot be directly compared with enriched CD34+ cells). Crossman *et al*.[290] demonstrated that in mixed cell populations, such as total white cells (TWCs), the gene expression profiles of CML patients who either obtained a complete cytogenetic response (CCyR) or had no response (no major cytogenetic response (MCyR); > 35% Ph+ metaphases) to imatinib therapy were similar.

However, microarrays using heterogeneous cell populations are still potentially informative, as phenotypic differences in individual patient samples may be related to leukaemia biology, therefore important information may be lost by enriching the samples for a particular sub-population.[291] Furthermore, studies involving enriched cell populations (e.g. CD34+ cells) are not easily translated to other laboratories, making the use of heterogeneous cell populations advantageous. As such, studies have been performed correlating CML gene expression profiles with cells collected from whole blood,[292,293] BM,[294] MNCs,[291] isolated CD34+ BM cells[295] and isolated CD34+ PB cells,[12] with patient response to therapy (mainly imatinib,[12,292-295] but also to IFN-α[291]). The majority of these studies compared response (generally defined as achieving a complete cytogenetic response [CCyR; 0% Ph+ cells]) with non-response (no CCyR, > 0% Ph+ cells), and were able to generate prediction models from genes that distinguished cytogenetic response from non-response.

4.1.2 Microarray Expression Profiles with CD34+ CML cells

Multiple studies have used cell fractions enriched for CD34+ haematopoietic progenitor cells to demonstrate patient differences within this more primitive cell population. Neumann *et al.*[295] compared Ph-negative CD34+ cells isolated from BM in patients with Ph-positive CML who achieved MMR, to normal CD34+ BM cells. They were however, unable to find a difference in the gene expression profiles between these cell types, which may be explained by the fact that MMR is not truly negative, but other unknown factors may have also contributed. Yong *et al.*[12] also used enriched CD34+ cells to differentiate between aggressive (developed blast crisis [BC] within 3 years of diagnosis) and indolent disease (survived for longer than 7 years before onset of BC). They were able to identify 20 genes that were differentially expressed between these two groups. Zheng *et al.*[181] examined enriched CD34+ cells from CML patients in BC and CP and identified a distinct molecular signature that distinguished the two phases. This included identifying 34 known genes with more than two-fold transcriptional changes.

Radich *et al.*[180] also compared changes in gene expression associated with progression and response in CML. Their data was consistent with CML progression occurring in a two-stage process, from chronic phase to advanced phase (including both AP and BC) rather than a three-stage process (CP-AP-BC). They postulated that CP patients, who had poor imatinib response, possess genetic features of advanced phase that are not detectable by routine examination at diagnosis.[180] These findings were extremely important as they had the potential to influence therapy, as patients identified with an advanced phase gene signature could be treated more aggressively upfront at diagnosis. This dataset was then reanalysed with the aim of developing a predictive gene signature for CML disease progression and patient response.[178] Using Bayesian Model Averaging (BMA), a probabilistic method for the identification of small gene-sets from multiple genes, a 6 gene (*NOB1*, *DDX47*, *IGSF2*, *LTB4R*,

SCARB1, and *SLC25A3*) predictive classifier that discriminated CP-CML from BC-CML, was developed. The advantage of using a method such as BMA, is that instead of segregating patients into discrete states (i.e. CP or BC), posterior probabilities between 0 and 1 are generated, allowing patients in "intermediate" (i.e. posterior probabilities close to 0.5) states to be predicted.[178] Validation of this 6 gene predictor was performed with an independent cohort of 88 CML patients, demonstrating that early CP patients were able to be discriminated from late CP, AP and BC.

More recently, McWeeney *et al.*[179] used a gene expression signature from enriched bone marrow CD34+ CML cells, to predict cytogenetic response in CP patients treated with imatinib. A gene expression classifier consisting of a 75 gene signature was elucidated, and when applied to an independent validation set, this classifier correctly predicted 88% of responders (patients achieving MCyR) and 83% of non-responders (patients not achieving MCyR). Finally, this 75 gene signature was used to perform a meta-analysis against the advanced disease datasets of Yong *et al*.[12] and Zheng *et al*.[181] Significant overlap was found with genes from both datasets, indicating that this gene expression classifier, which is the first prospectively validated classifier developed for CML, may allow for the development of intensive targeted therapy aimed at the patients most likely to benefit. Importantly, pathway and gene ontology analysis performed with this signature gene-set proposed that the patients, who failed to respond to imatinib, possessed a more advanced disease phenotype than specified by morphologic criteria alone. However, the use of CD34+ cells in the development of this classifier has not lead to wide acceptance.

4.1.3 The Limitations of Microarray Gene Expression Profiling

There are also important limitations associated with microarrays that need to be considered. Some of these have already been discussed, including the heterogeneity of the cell sample type, e.g. TWC or MNC versus enriched CD34+ cells. The caveats of the technology itself include the fact that it is quite expensive to perform, each microarray produces huge amounts of data that must be sorted to isolate the significant results and large numbers of patient samples are required for the results to be significant when comparisons are made between the studied outcome groups.[296] However, even with these limitations, microarray technology is an excellent tool to analyse gene expression in CML. Furthermore, it is helping to shape our understanding of this disease and could potentially play an important role in the understanding of how OA is associated with disease outcome.

4.1.4 Approach

In order to identify candidate genes potentially associated with OCT-1 activity (OA), preliminary microarray analysis was performed using RNA isolated from the TWCs of 14 *de novo* CP-CML patients (**Section 2.4.4.2**) from the two extreme OA groups (very low OA [OA < 4 ng/200,000 cells; n=8] and very high OA [OA > 10 ng/200,000 cells; n=6]). OA was measured as described in **Section 2.4.1.1** using thawed CML patient MNCs isolated from peripheral blood diagnosis samples (**Section 2.3.1** and **2.3.4**), with the OCT-1 inhibitor prazosin. Using a significance cut-off ($P < 0.05$), the top differentially expressed genes were selected and unsupervised hierarchical clustering performed (**Section 2.6.1**) to determine if these genes distinguished very low from very high OA. A gene-set containing 89 genes was identified. To identify genes with common biological function, gene set enrichment analysis (GSEA) was utilised (**Section 2.6.4**) and to test for lineage enrichment, the Chambers *et al.*[218] dataset was re-analysed (**Section 2.6.5**).

Validation of the microarray data was performed using custom TaqMan® Low Density Arrays (TLDAs, Applied Biosystems), containing primer and probe sets specific for the 89 identified genes, as well as 7 housekeeping controls (**Section 2.4.5**) and analysed using HTqPCR (**Section 2.6.7**). A predictive gene expression classifier for OA was developed; with a 10-gene classifier (75% accuracy) using iterativeBMA (**Section 2.6.2**) which had the greatest accuracy compared to any classifier developed using the CMA package (**Section 2.6.8**) identified. Validation of this 10-gene classifier, as well as 10 other genes of interest and 4 reference genes, was performed using custom TLDA (24-gene format; **Section 2.5.5**) with an independent cohort of 80 *de novo* CP-CML patients. Altogether, 110 *de novo* CP-CML patients had RQ-PCR data available for 17 genes (not including controls), allowing correlation analyses with patient response (cytogenetic, molecular and risk of transformation), as well as OA, to be performed.

Further interrogation of previously published CML microarrays[179,180] for differential histone gene expression was then performed due to the observed enriched histone gene signature in very low OA. Transcription factor binding enrichment analysis was performed using three different tools TFactS, Pscan and ChEA (**Section 2.6.9**) to determine the role and regulation of histones in CP-CML. To investigate whether transcription factors were responsible for the up-regulation of histone gene expression in very low OA, western blot experiments were performed (**Section 2.3.6** and **2.4.3.3**) to determine protein levels. The analyses discussed in this chapter were used to determine how the gene expression profile of very low OA patients differs initially from patients with very high OA, and subsequently, all other OA patients, to further understand the underlying biology associated with very low OA.

4.2 Results

4.2.1 Microarray analysis of very low and very high OCT-1 activity patients

Preliminary microarray analysis was performed to examine the genes associated with very low (OA < 4 ng/200,000 cells) and very high (OA > 10 ng/200,000 cells) OA CP-CML patients. Total white cells (TWCs) from 14 *de novo* CP-CML patients (very low OA, n = 8; very high OA, n = 6) were studied due to cost limitations. Statistical analyses (refer **Chapter 2, Section 2.4.13.2**) revealed approximately 3,400 genes (from a total of approximately 29,000 genes) were differentially expressed between very low and very high OA patients using a minimum statistical significance cut-off of $P < 0.05$ (Student's *t*-test). This gene-set was further consolidated using a more stringent statistical cut-off (FDR *P*-value ≤ 0.2 and $P < 0.01$), allowing the top 800 genes to be selected (**Figure 4.2**). As this number of genes was still large, and ultimately, 96-gene format high-throughput RQ-PCR was going to be used for validation, the top 89 significant genes were selected based on the greatest fold-change difference between the OA groups. Hierarchical clustering was then performed on this core set of 89 genes (40 up- and 49 down-regulated in very low, compared to very high OA; **Appendix 7**), which identified two distinct clusters; one predominantly containing very low OA patients, and one predominantly containing very high OA patients (**Figure 4.3**). The identified 89 gene-set is referred to as the "OA gene-set" from this point on.

4.2.2 Gene Set Enrichment Analysis identifies specific lineage differences between very low and very high OCT-1 activity patients

Gene Set Enrichment Analysis (GSEA; Broad Institute)[216,297] is a powerful bioinformatics tool for identifying groups of genes from a defined gene-set that are known to be functionally related to published gene-sets with well annotated biological information (e.g. metabolic

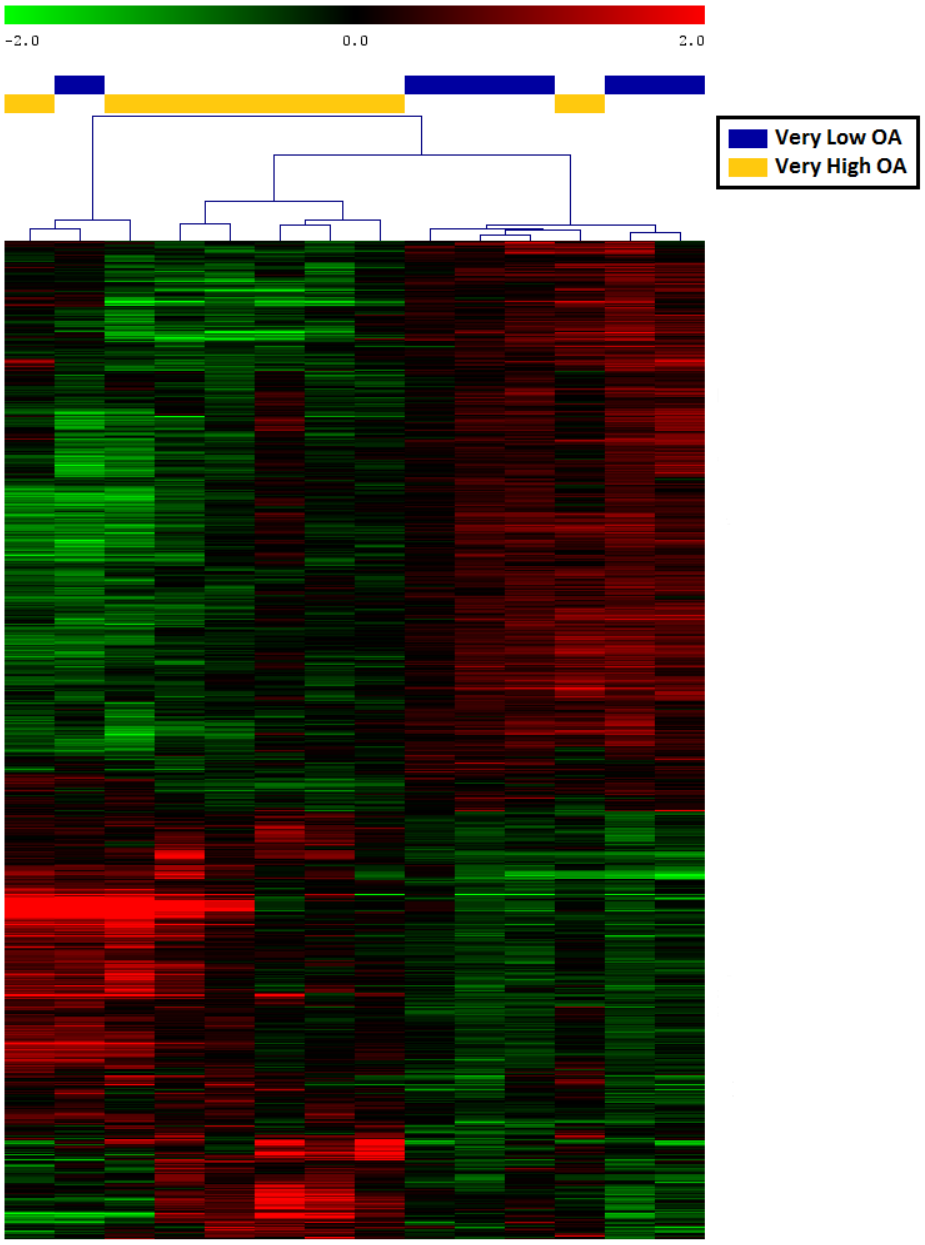


Figure 4.2 Genes differentially expressed between very low and very high OA patients

Samples from 8 patients with very low OA and 6 patients with very high OA were compared using Affymetrix microarrays. Approximately 800 genes were differentially expressed at the significance criteria FDR *P*-value ≤ 0.2 and $P < 0.01$. Each column represents one sample, and each row represents one gene. **Red** colour indicates overexpression compared to the median expression for that gene, and **green** colour represents low expression. Unsupervised hierarchical clustering (Pearson correlation, average linkage) was performed using MultiExperiment Viewer (MeV v4.8, http://www.tm4.org/mev/).

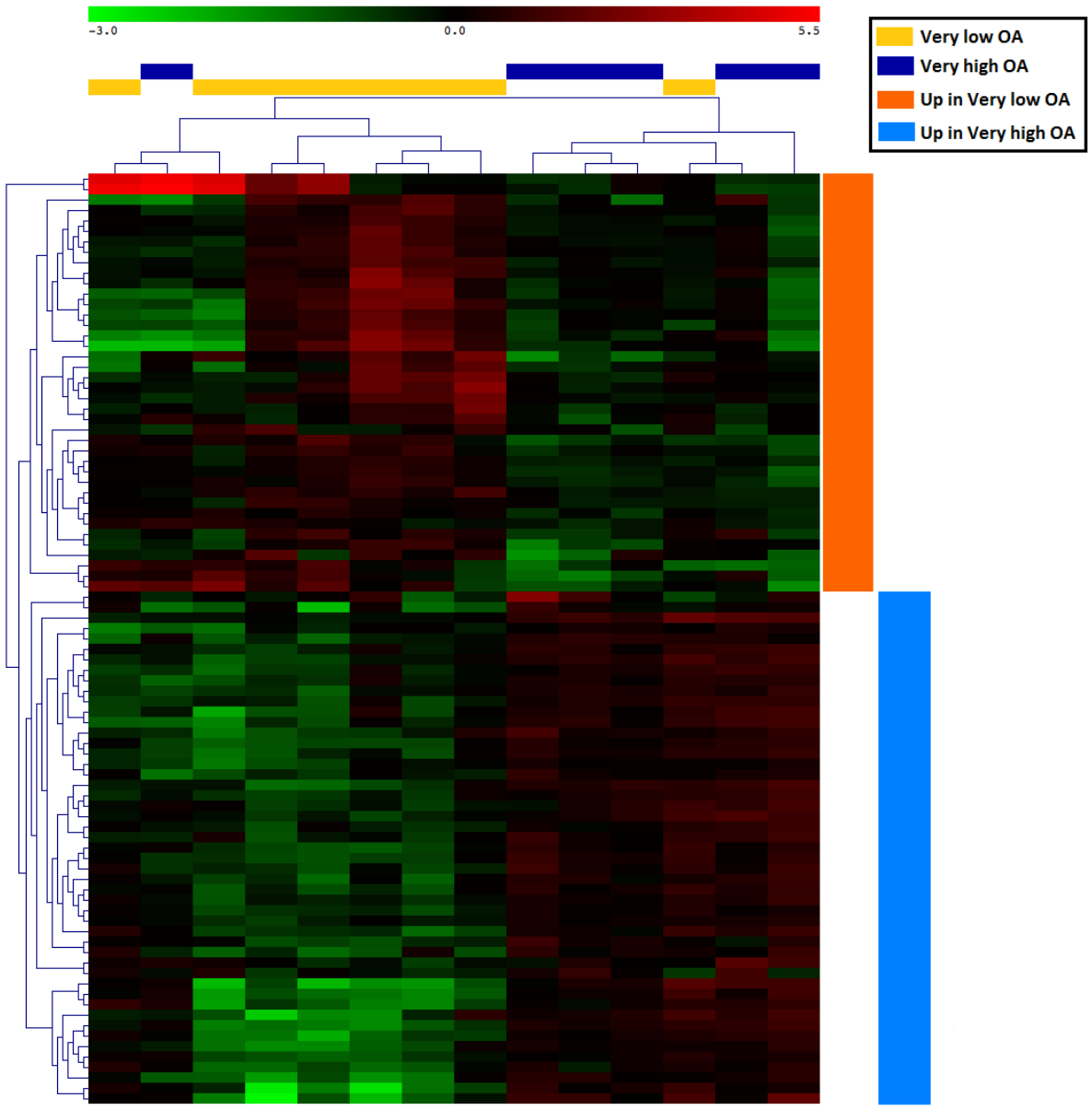


Figure 4.3 A key 89 gene-set differentiates very low and very high OA patients

The top 89 genes differentiating very low and very high OA patients were identified using the following significance criteria: FDR *P*-value ≤ 0.2, $P < 0.01$ and greatest fold-change. Each column represents one sample, and each row represents one gene. **Red** colour indicates over-expression compared to the median expression for that gene, and **green** colour represents under-expression. Unsupervised hierarchical clustering (Pearson correlation, average linkage) was performed using MultiExperiment Viewer (MeV v4.8, http://www.tm4.org/mev/).

pathways, cell signalling pathways or kinases; genes targeted by microRNAs; or common phenotypes produced from gene knockout models etc.)[298] In short, GSEA is based on the concept that gene-set differences will be more prominent in the data than individual gene expression differences. Therefore, we used GSEA to examine the proposed biological function of the genes differentially expressed between the very low and very high OA patients.

Using GSEA with the default significance cut-off (FDR *q*-value < 0.25, normalised enrichment score (NES) > 1.5 or < -1.5), 322 gene-sets were significantly up-regulated in very low OA patients, and 925 gene-sets were significantly up-regulated in very high OA patients. Furthermore, to ascertain the gene-sets with the greatest enrichment, the statistical cut-off was reduced to FDR *q*-value < 0.05, identifying 101 and 663 gene-sets enriched for very low and very high OA patients, respectively. Interestingly, a number of gene-sets involving cell lineage, specifically monocytes (up-regulated) and erythrocytes (down-regulated), were enriched in the very low OA patients (**Table 4.1**; **Figure 4.4**). In contrast, a significant enrichment of cell cycle and mitosis and DNA replication/synthesis gene sets were observed for the very high OA patients (**Table 4.2**; **Figure 4.5**). The only cell lineage-associated gene-set enriched in the top 250 gene-sets for the very high OA patients was a down-regulated monocyte gene-set (Sakai – Tumour infiltrating monocytes – Down; **Table 4.2**). Intriguingly, the direct opposite, up-regulated monocyte gene-set (Sakai – Tumour infiltrating monocytes – Up; **Table 4.1**) was enriched in the very low OA patient samples. This data suggests that there are significant lineage differences in the TWC populations between CP-CML patients with very low and very high OA, which may be contributing to the differences in OA observed and consequently, the underlying biology of the leukaemia itself.

Table 4.1 Common functional gene-sets up-regulated in very low OA patients by GSEA

	Gene-set Size*	FDR *q*-value	Nominal *P*-value	NES
Cell lineage				
***KEGG* – Haematopoietic cell lineage**	80	0.003	< 0.001	2.18
***TONKS* – Targets of RUNX1-RUNX1T1 fusion sustained in Monocytes - UP**	18	0.009	< 0.001	2.00
***BIOCARTA* – Erythroid pathway**	15	0.010	< 0.001	1.97
***HADDAD* – T lymphocyte and NK progenitor - UP**	76	0.014	< 0.001	1.94
***TONKS* – Targets of RUNX1-RUNX1T1 fusion Erythrocyte - DOWN**	15	0.014	< 0.001	1.93
***TENEDINI* – Megakaryocyte markers**	50	0.015	< 0.001	1.92
***SAKAI* – Tumour infiltrating monocytes - UP**	26	0.018	< 0.001	1.89

Abbreviations: FDR – false discovery rate, NES – normalised enrichment score. *Gene-set size refers to the number of genes overlapping between the gene-sets. The NES is the primary statistic for examining gene set enrichment results, as it accounts for differences in gene-set size and in correlations between gene-sets and the expression dataset.

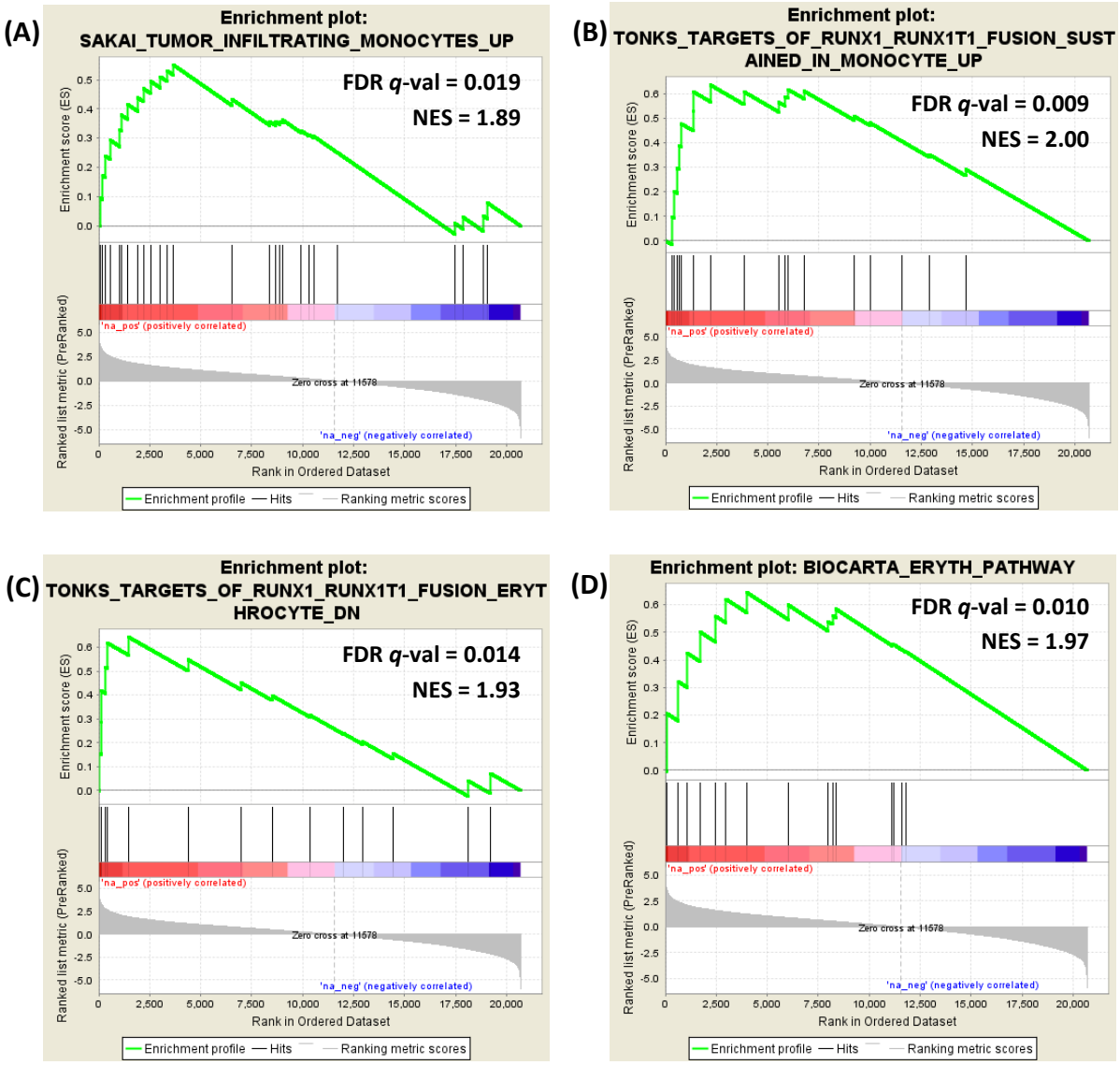


Figure 4.4 Monocytic and erythroid lineage enrichment in very low OA patients by GSEA

GSEA identified significant enrichment between the very low OA patient gene-set and multiple cell lineage-specific gene-sets, particularly monocytes [**(A)** and **(B)**] and erythrocytes [**(C)** and **(D)**]. On the GSEA enrichment plots, single genes common between the two gene-sets are shown as black vertical lines above the red – blue scale, where **Red** indicates positive association for very low OA genes and **Blue** indicates negative association for very low OA genes (positive very high OA gene association). The enrichment score is plotted at the top in **Green**, with a peak on the left indicating enrichment for very low OA-associated genes. Abbreviations: FDR – false discovery rate, NES – normalised enrichment score.

Table 4.2 Common functional gene-sets up-regulated in very high OA patients by GSEA

	Gene-set Size*	FDR *q*-value	Nominal *P*-value	NES
Cell cycle and Mitosis				
***REACTOME* – Mitotic M-M/G1 phases**	146	< 0.001	< 0.001	-3.38
***REACTOME* – S phase**	100	< 0.001	< 0.001	-3.28
***REACTOME* – Cell cycle checkpoints**	101	< 0.001	< 0.001	-3.27
***REACTOME* – Cell cycle mitotic**	282	< 0.001	< 0.001	-3.20
***REACTOME* – M-G1 transition**	60	< 0.001	< 0.001	-3.15
***REACTOME* – G1-S transition**	96	< 0.001	< 0.001	-3.14
***KEGG* – Cell cycle**	112	< 0.001	< 0.001	-2.54
DNA Replication/Synthesis				
***REACTOME* – DNA replication pre-initiation**	74	< 0.001	< 0.001	-3.25
***REACTOME* – Synthesis of DNA**	86	< 0.001	< 0.001	-3.22
***KEGG* – DNA replication**	34	< 0.001	< 0.001	-2.76
***REACTOME* – DNA strand elongation**	29	< 0.001	< 0.001	-2.75
***REACTOME* – DNA repair**	92	< 0.001	< 0.001	-2.66
Cell lineage				
SAKAI – Tumour infiltrating monocytes – DOWN	75	< 0.001	< 0.001	-2.34

Abbreviations: FDR – false discovery rate, NES – normalised enrichment score. *Gene-set size refers to the number of genes overlapping between the gene-sets. The NES is the primary statistic for examining gene set enrichment results, as it accounts for differences in gene-set size and in correlations between gene-sets and the expression dataset.

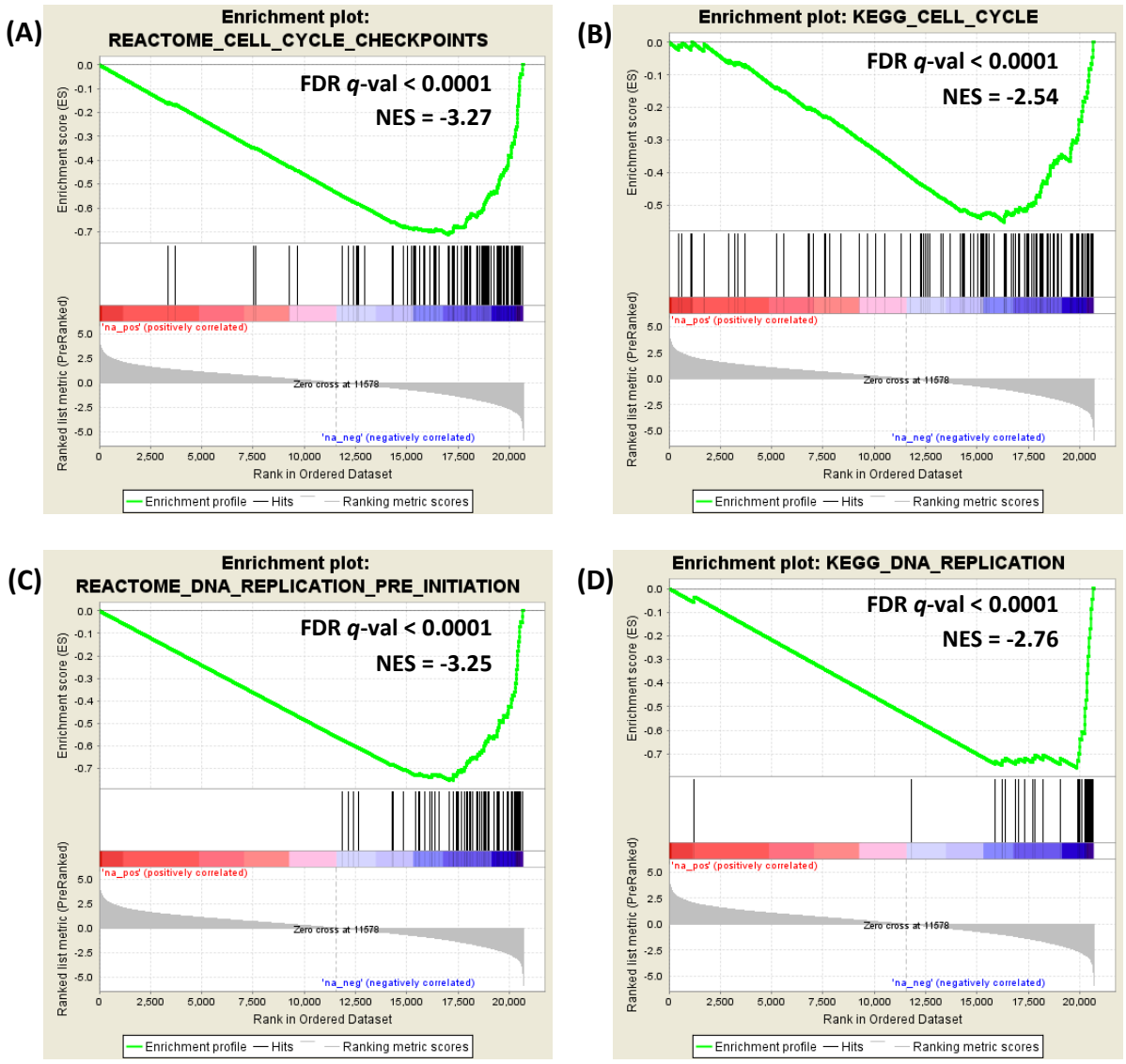


Figure 4.5 Cell cycle and DNA replication enrichment in very high OA patients by GSEA

GSEA identified significant enrichment between the very high OA patient gene-set and multiple cell cycle/mitosis [**(A)** and **(B)**] and DNA replication/synthesis gene-sets [**(C)** and **(D)**]. On the GSEA enrichment plots, single genes common between the two gene-sets are shown as black vertical lines above the red – blue scale, where **Red** indicates positive association for very low OA genes (negative very high OA gene association) and **Blue** indicates negative association for very low OA genes, hence positive very high OA gene association. The enrichment score is plotted at the top in **Green**, with a negative peak on the left indicating enrichment for very high OA-associated genes. FDR – false discovery rate, NES – normalised enrichment score.

To further investigate the cell lineage signature associated with the very low OA patients, the gene expression database of haematopoietic stem cells (HSCs) and their differentiated progeny; including erythrocytes, granulocytes, monocytes, NK cells, activated and naïve T-cells, and B-cells, generated by Chambers *et al.*[218] was interrogated. Although developed using murine HSCs, this gene expression study provides insight into the genes uniquely expressed in HSCs, as well as their differentiated progeny. Using a select number of genes identified by Chambers *et al*. as specific for particular cell lineages, the very low OA gene signature was again demonstrated to be associated with increased monocytic and erythroid gene expression, and decreased granulocytic gene expression (**Figure 4.6**). Taken together, this data demonstrates that a gene signature strongly associated with monocytic and erythroid cell lineages, exists in very low OA patients.

4.2.3 Validation of the 89 gene OCT-1 activity signature by RQ-PCR

Microarray gene expression analysis is an excellent tool to investigate the expression profiles of thousands of genes in samples at once; however, the sensitivity of this technology is low. Therefore, to validate the results from the OA microarray and in particular the key 89 gene signature, quantitative reverse transcription PCR (RQ-PCR) was performed using custom TaqMan® Low Density Arrays (TLDAs). Initially, custom 96-gene TLDAs were designed, incorporating all 89 genes identified from the microarray (**Section 4.2.1**), plus 7 reference genes. TLDA RQ-PCR was performed for 40 CP-CML patients, using *de novo* TWC RNA samples. Included in these 40 patients were the original 14 patients analysed on the microarray (see **Table 4.3** for patient details).

The reliability of a RQ-PCR experiment can be improved by including invariant endogenous controls (reference genes) in the assay design to correct for sample to sample variations in

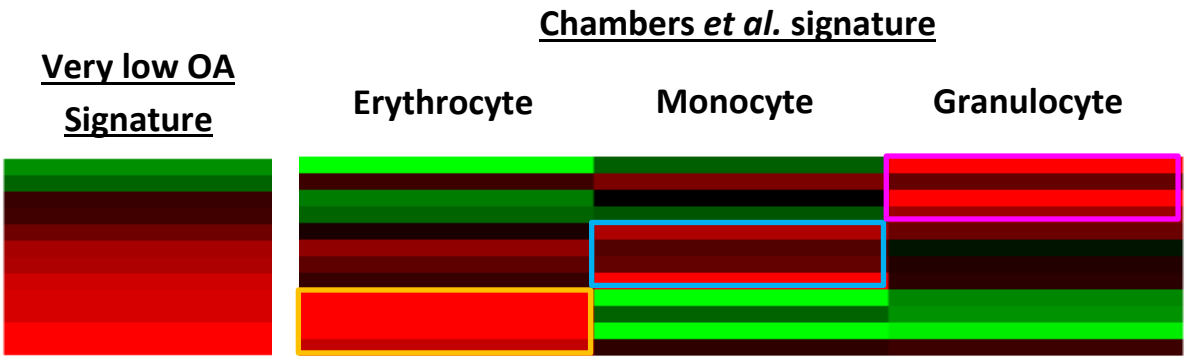


Figure 4.6 The very low OA gene expression profile is enriched with an erythroid and monocytic signature

The expression profile of the very low OA patient cohort was compared against the Chambers *et al.*[218] HSC gene expression dataset. The very low OA expression signature was associated with increased expression of erythroid (**Gold**) and monocytic (**Blue**) genes, and decreased expression of a granulocytic genes (**Pink**). The columns represent expression of the selected genes in the very low OA gene signature, and in particular cell lineages, with each row representing a single gene. **Red** colour indicates increased expression compared to the median expression for that gene, and **green** colour represents low expression.

Table 4.3 Characteristics of the 40 patients used for 96-gene TLDA RQ-PCR

		Patient Cohort	
		Very low OA (n = 19)	Very high OA (n = 21)
Median Age	**Years**	45	50
	Range	18 – 74	32 – 80
Male	**%**	84.2	61.9
Sokal Score	**High**	3	3
	Intermediate	7	5
	Low	8	12
BCR-ABL1 levels < 10%	**3 month, %**	73.7	85.7
MMR	**12 month, %**	42.1	85.7
	24 month, %	47.4	90.5
Transformation	**24 month, %**	10.5	4.8

One patient in each of the very low and very high OA cohorts did not have a Sokal score available. Abbreviations: OA – OCT-1 activity; MMR – major molecular response, BCR-ABL1 mRNA levels < 0.1%.

PCR efficiency, as well as errors in sample quantification (e.g. variation in amount of starting material or cDNA loading). These reference genes allow normalisation of the target mRNA copy numbers to be performed. However, the quality of the normalised expression data is dependent on the quality of the reference gene used for normalisation.[299] Therefore, to determine the appropriate reference gene for this RQ-PCR analysis, RefFinder[219] was used to evaluate and screen for the optimal reference gene. RefFinder integrates four major statistical algorithms (geNorm, Normfinder, BestKeeper and the comparative ΔCt method) to determine the stability of each reference gene using a ranking method. Based on the rankings from RefFinder, *GUSB* was the most stably expressed gene in this 40 patient RQ-PCR dataset (**Table 4.4**; **Figure 4.7**). Therefore, the ΔCt was determined by normalising each gene to the level of *GUSB* mRNA for each sample using the HTqPCR package[225] (**Figure 4.8**), and the ΔΔCt was calculated relative to very high OA CP-CML patients. Differential gene expression was determined using a Student's *t*-test implemented in HTqPCR and only genes with a $P < 0.05$ were considered statistically significant (refer **Figure 4.9** for data analysis workflow).

Of the 89 genes analysed by RQ-PCR, 16 genes were significantly down-regulated and one gene was up-regulated in the very high OA patient group, compared to very low OA patients (**Table 4.5**; **Figure 4.10**), validating the findings from the microarray profiling. When corrected for multiple testing (using the Benjamini-Hochberg (BH) adjusted FDR; a more stringent criteria), 9 genes remained significantly down-regulated in the very high OA patient group, compared to very low OA patients (**Figure 4.10** and **4.11**). Unsupervised hierarchical clustering based on the expression of these 9 genes was performed (**Figure 4.12**); however, very low OA patients were unable to be distinctly separated from the very high OA patients. This result indicated that although these 9 genes were statistically significant individually,

Table 4.4 Rank of reference genes based on their expression stability in 40 CP-CML patients according to RefFinder analysis of four different statistical algorithms

Method	1	2	3	4	5	6	7
Delta Ct	GUSB	HMBS	HPRT1	B2M	ACTB	18S	PES1
BestKeeper	GUSB	B2M	HPRT1	HMBS	PES1	ACTB	18S
Normfinder	GUSB	HMBS	ACTB	B2M	HPRT1	18S	PES1
geNorm	GUSB/HPRT1		B2M	HMBS	ACTB	18S	PES1
Recommended Comprehensive Ranking	**GUSB**	**HPRT1**	**HMBS**	**B2M**	**ACTB**	**18S**	**PES1**

Each reference gene is ranked in order from 1 to 7; where 1 corresponds to the best gene and 7 to the worst, for each of the four statistical algorithms analysed in RefFinder. *GUSB* was identified as the most stable reference gene in all four methods, and is the gene recommended by RefFinder for use as the reference gene within this CP-CML patient cohort.

Figure 4.7 Comprehensive reference gene stability ranking according to RefFinder analysis in 40 CP-CML patient samples

The geometric mean of the ranking values determined by all four statistical algorithms (Delta Ct, BestKeeper, Normfinder and geNorm) performed by RefFinder are plotted. *GUSB* has the lowest geometric mean, indicating that it is the most stable reference gene in this cohort of CP-CML patients. Therefore, *GUSB* was used as the reference gene for RQ-PCR data normalisation.

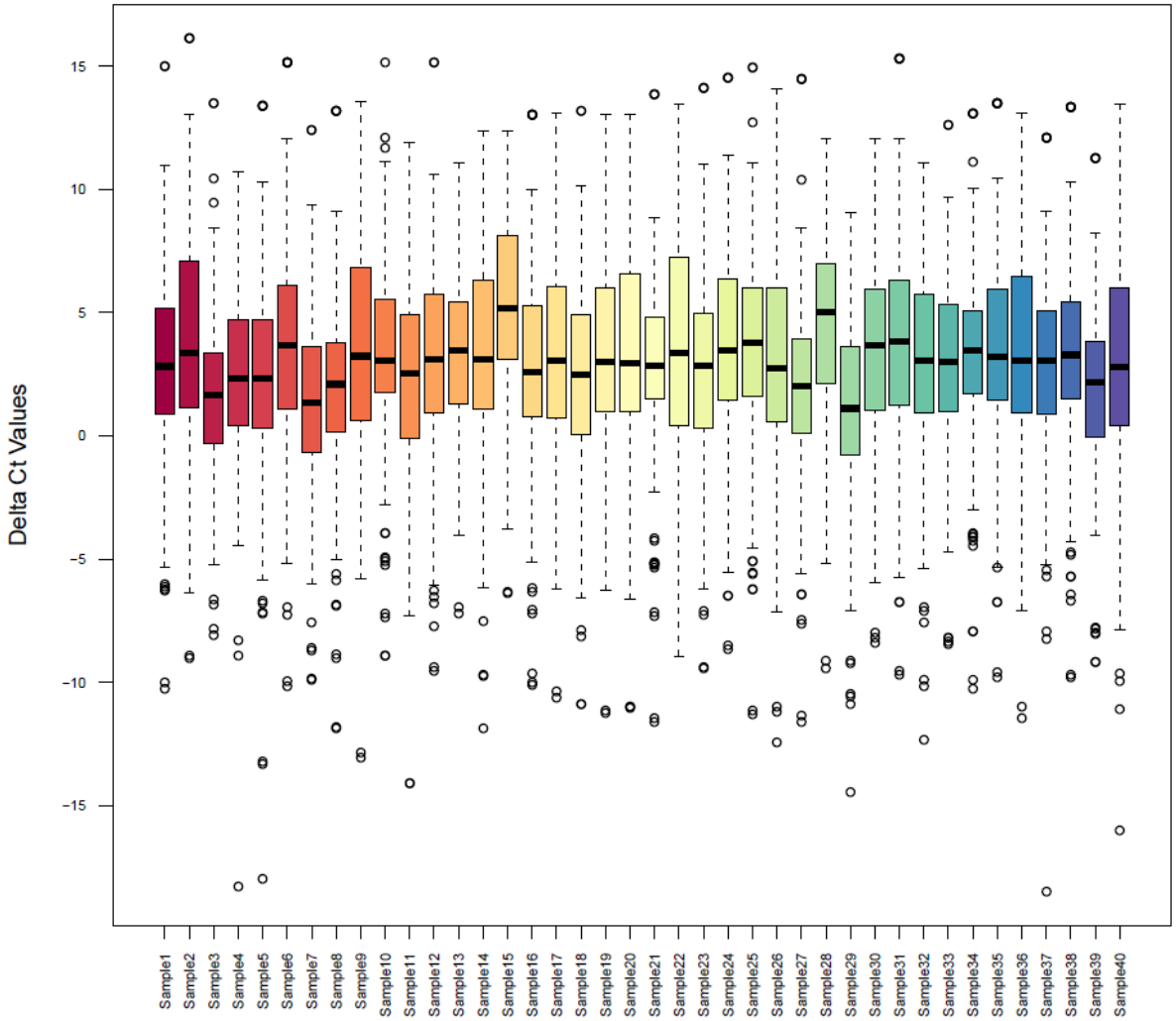


Figure 4.8 Boxplots demonstrating the range of Delta Ct values for each patient after normalisation with HTqPCR

The Delta Ct (ΔCt) values for all 89 genes normalised to *GUSB* were plotted for each of the 40 patients analysed by TLDA RQ-PCR. Boxplots represent the 25th to 75th percentile, with 50th percentile (median) represented by the centreline. Whiskers represent 1.5*IQR (interquartile range), providing roughly a 95% confidence interval, with outliers plotted as open circles.

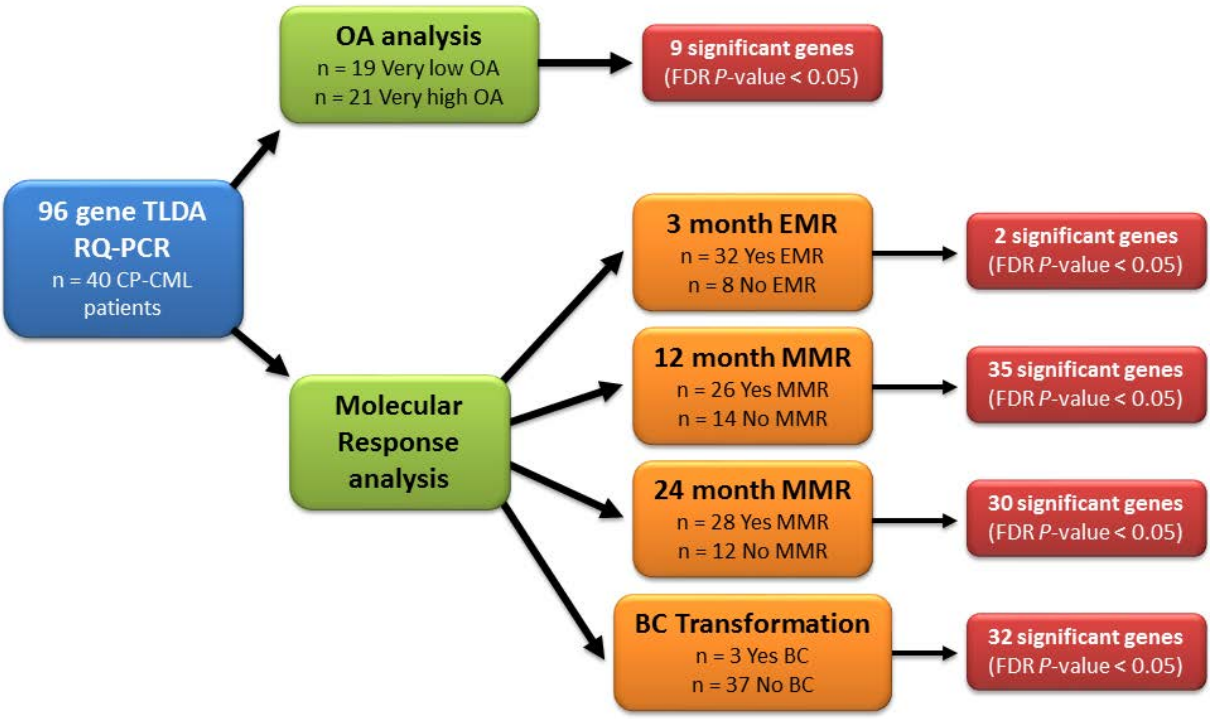


Figure 4.9 Workflow for analysis of the 96-gene RQ-PCR data

The **blue** box highlights the initial dataset used for the HTqPCR analysis. The **green** boxes indicate the two main analyses performed; firstly OA, followed by molecular response. The **orange** boxes demonstrate the particular response parameters investigated, while the **red** boxes indicated the number of statistically significant genes identified. EMR – early molecular response (3 month BCR-ABL1 mRNA levels < 10%); MMR – major molecular response (BCR-ABL1 mRNA levels < 0.1%); BC – blast crisis.

Table 4.5 List of genes with significant differential expression between very low and very high OA CP-CML patients, as determined by HTqPCR analysis

Gene	*P*-value	FDR *P*-value	Log FC
HIST1H2AG	< 0.001	0.001	-1.48
HIST1H2BB	< 0.001	0.022	-1.19
JUNB	< 0.001	0.022	-1.27
MT-ND2	0.001	0.022	-0.73
ANP32E	0.001	0.022	-0.81
SPRY2	0.001	0.022	-0.78
MLLT4	0.002	0.034	-1.15
GBE1	0.003	0.038	-0.71
HIST2H2BE	0.004	0.038	-1.33
HIST1H3H	0.011	0.100	-0.82
MAD2L1	0.011	0.100	-0.41
LDHA	0.018	0.146	-0.38
HIST1H2BJ	0.023	0.171	-0.59
OLFM4	0.026	0.181	0.82
HIST1H4A	0.029	0.183	-0.73
NIPSNAP3A	0.033	0.197	-0.54
LRP6	0.042	0.236	-0.60

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to very low OA patients.

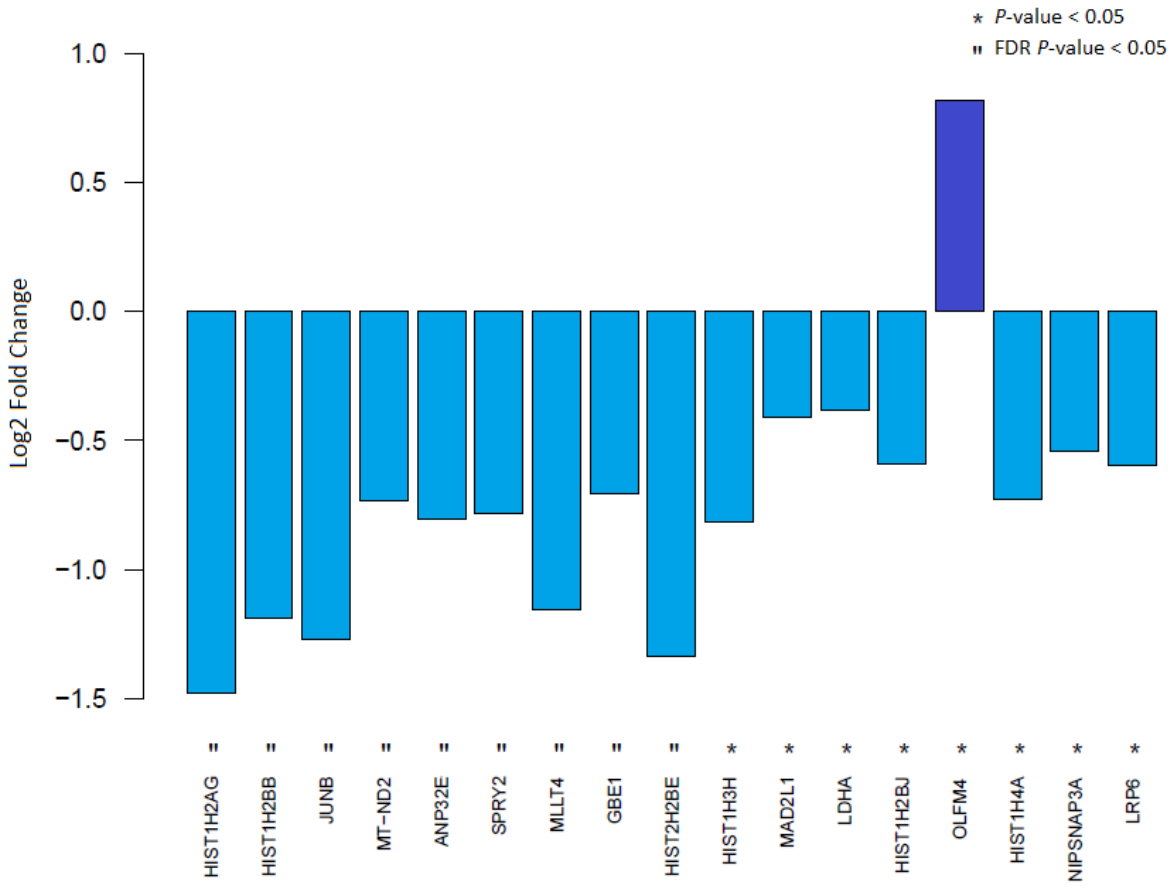
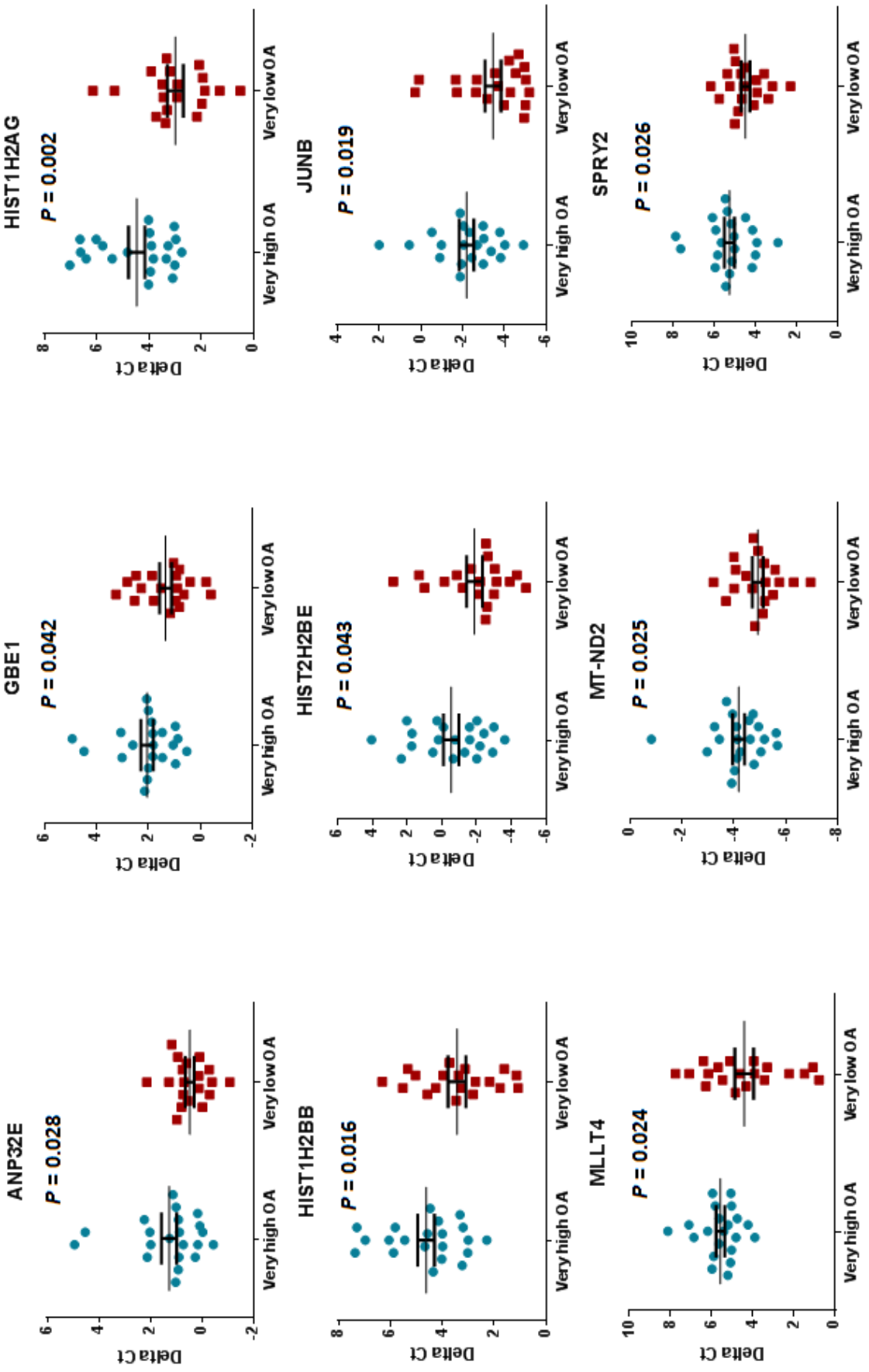


Figure 4.10 Significant gene expression between very low and very high OA CP-CML patients by TLDA analysis using TWC mRNA

The log2-transformed fold change (in relation to very high OA) values corresponding to the top 17 genes differentially expressed between very low and very high OA CP-CML patients, with a *P*-value < 0.05, were plotted. Sixteen genes demonstrated increased expression in very low OA patients, while only 1 gene had increased expression in very high OA patients. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

Figure 4.11 Nine genes had significant differential expression between very low and very high OA CP-CML patient samples

The difference between gene expression for very low (n = 19) and very high OA (n = 21) patients is plotted. Scatterplots demonstrate the mean Delta Ct ± standard error of the mean (SEM). Delta Ct values were generated against the reference gene, *GUSB*. Student's *t*-tests were performed in HTqPCR.


HIST1H2AG
P = 0.002
JUNB
P = 0.019
SPRY2
P = 0.026
GBE1
P = 0.042
HIST2H2BE
P = 0.043
MT-ND2
P = 0.025
ANP32E
P = 0.028
HIST1H2BB
P = 0.016
MLLT4
P = 0.024
Delta Ct
Very high OA
Very low OA

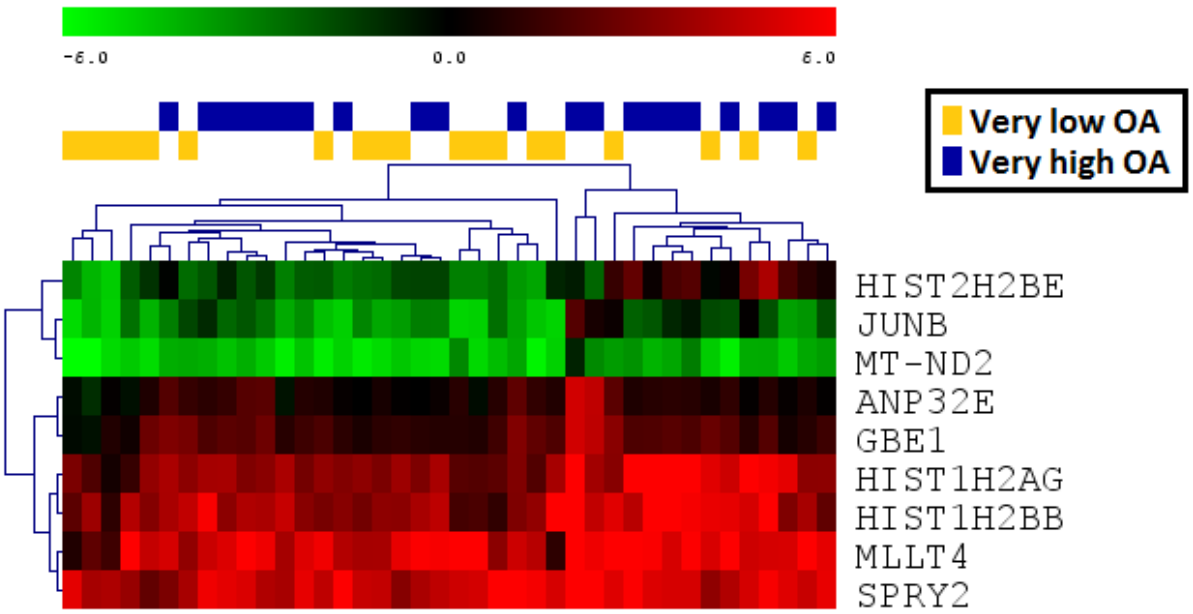


Figure 4.12 Hierarchical clustering of significant genes with differential expression between very low and very high OA CP-CML patients

Unsupervised hierarchical clustering (Euclidean distance, complete linkage) was performed using top 9 significant genes (FDR *P*-value < 0.05) differentiating very low and very high OA patients, from the 40 patient TLDA RQ-PCR analysis. Each column represents one sample, and each row represents one gene. **Red** colour indicates overexpression compared to the median expression for that gene, and **green** colour represents under-expression. Hierarchical clustering was performed using MultiExperiement Viewer (MeV v4.8, http://www.tm4.org/mev/).

when used in combination, the expression values were unable to predict very low OA patients from very high OA patients in this larger patient group than the initial microarray. The results suggest significant heterogeneity amongst the CP-CML patient cohort; however, analysis of a larger patient cohort is required to validate this inference.

Interestingly, interrogation of these 9 significantly regulated (FDR *P*-value) genes revealed that 3 were histone genes (*HIST2H2BE*, *HIST1H2AG* and *HIST1H2BB*), suggesting a possible role for histones in influencing OA. The proto-oncogene *JUNB* (decreased expression in very low OA patients), is a transcription factor with an important role in promoting myeloid differentiation,[300] providing further evidence for the differential lineage involvement between very low and very high OA patients. Previous studies in CML have also demonstrated that *JUNB* expression is down-regulated in advanced phase disease, compared to CP.[301,302] Sprouty homolog 2 (*SPRY2;* decreased expression in very low OA patients), is a negative feedback regulator of multiple receptor tyrosine kinases; including fibroblast growth factor (FGF),[303] epidermal growth factor (EGF),[304] and hepatocyte growth factor (HGF).[305] Although no direct relationship between *SPRY2* and CML has been previously revealed, *SPRY2* is associated with protein phosphatase 2A (PP2A) inactivation in tumorigenesis[306] and PP2A inactivation has been observed in BC and imatinib-resistant CML,[307,308] providing a possible mechanism for *SPRY2* involvement in CML. Intriguingly, *ANP32E* (decreased expression in very low OA patients), a member of the acidic nuclear phosphoprotein 32 family, has also been demonstrated to regulate PP2A activity.[309] Finally, *MLLT4* (myeloid/lymphoid or mixed-lineage leukaemia, translocated to, 4; decreased expression in very low OA patients), is the fusion partner of the acute lymphoblastic leukaemia (MLL) gene, involved in acute myeloid leukaemia (AML) patients demonstrated to have the t(6;11)(q27;q23) translocation.[310] While no specific role in CML has been described,

MLLT4 encodes the Ras-interacting AF-6 (afadin) protein, which when phosphorylated by the Bcr kinase, allows Bcr to bind to its PDZ (Post synaptic density protein [PSD95], Drosophilia disc large tumour suppressor [Dlg1], and Zonula occludens-1 protein [zo-1]) domain, interfering with the regulation of Ras-mediated signalling;[311] a pathway closely linked with Bcr-Abl1 activation.[312]

4.2.4 Interrogation of the 96 gene RQ-PCR dataset identifies genes predictive of response

To determine if, like OA, any of the key 89 genes analysed by TLDA RQ-PCR were predictive for various CML patient response criteria (early molecular response [EMR; 3 month BCR-ABL1 mRNA levels < 10% by the international standard]; 12 month MMR; 24 month MMR; and BC transformation), the relationship between gene expression and response was assessed. Firstly, EMR was evaluated, with patients divided by those who did (n = 32), or did not (n = 8), achieve 3 month BCR-ABL1 mRNA levels < 10% on TKI therapy. Nineteen genes were differentially expressed with a *P*-value < 0.05; 16 of which had significantly increased expression in CP-CML patients who achieved EMR, and 3 were significantly increased in patients who failed to achieve EMR (**Table 4.6**; **Figure 4.13**). When corrected for multiple testing (BH-FDR), 2 genes were still significantly up-regulated in CP-CML patients who achieved EMR, compared to those who did not (**Table 4.6**; **Figure 4.13**).

Next, the achievement of 12 month MMR was evaluated, with patients divided by those who did (n = 26), or did not (n = 14), achieve BCR-ABL1 mRNA levels < 0.1% by 12 months on TKI therapy. Using the same significance criteria as before (*P*-value < 0.05), 38 differentially expressed genes were identified; 37 of which had significantly increased expression in CP-CML patients who achieved MMR by 12 months, while 1 gene was significantly increased in patients who failed to achieve MMR by 12 months (**Table 4.7**; **Figure 4.14**). When corrected

Table 4.6 Genes with differential expression in CP-CML patients based on the achievement of EMR by HTqPCR analysis

Gene	*P*-value	FDR *P*-value	LogFC
SMOX	< 0.001	0.006	-1.51
CLU	< 0.001	0.038	-1.31
CD63	0.003	0.098	-0.82
NRGN	0.004	0.102	-1.19
GRB14	0.007	0.127	-1.74
ENKUR	0.010	0.144	-1.46
FAXDC2	0.010	0.144	-0.90
TSPAN33	0.015	0.183	-1.00
VAMP8	0.024	0.215	0.32
GNAZ	0.024	0.215	-0.99
HIST1H3H	0.025	0.215	-0.95
ENDOD1	0.028	0.222	-0.89
CYYR1	0.034	0.239	0.72
MYL9	0.035	0.239	-1.06
MLLT4	0.037	0.239	1.11
NFIB	0.040	0.239	-1.13
DDIT4	0.046	0.246	-0.97
C2orf88	0.047	0.246	-1.14
RPIA	0.049	0.246	-0.36

Abbreviations: EMR – early molecular response; FDR – false discovery rate; FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to achieving EMR by 3 months.

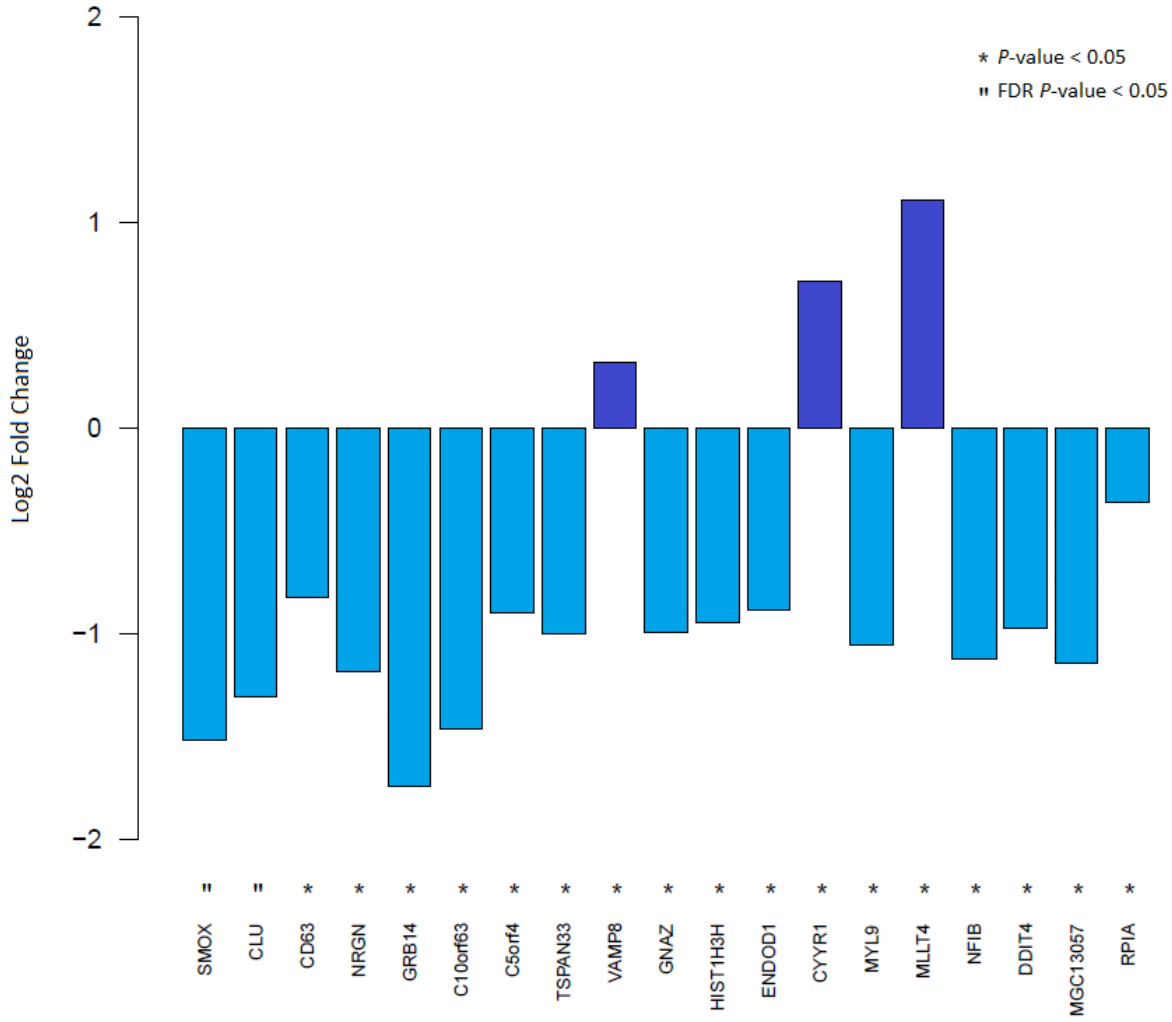


Figure 4.13 Significant gene expression between CP-CML patients who do, or do not, achieve EMR using TWC mRNA

The log2-transformed fold change (in relation to not achieving EMR) values corresponding to the top 19 differentially expressed genes associated with EMR (3 month BCR-ABL1 mRNA levels < 10%) in CP-CML patients, with a *P*-value < 0.05 were plotted. Sixteen genes demonstrated increased expression in CP-CML patients who achieved EMR, while 3 genes had increased expression in CP-CML patients who did not achieve EMR. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

Table 4.7 Genes with differential expression in CP-CML patients based on the achievement of MMR by 12 months with HTqPCR analysis

Gene	*P*-value	FDR *P*-value	LogFC	Gene	*P*-value	FDR *P*-value	LogFC
CLU	< 0.001	< 0.001	-2.51	***PTK2***	< 0.001	< 0.001	-1.97
NRGN	< 0.001	< 0.001	-2.41	***JAM3***	< 0.001	0.001	-1.91
SMOX	< 0.001	< 0.001	-1.97	***UGT2B11***	0.001	0.004	-1.39
TSPAN33	< 0.001	< 0.001	-1.89	***C2orf88***	0.001	0.006	-1.82
BEND2	< 0.001	< 0.001	-3.43	***ALDH2***	0.002	0.010	-1.09
ENDOD1	< 0.001	< 0.001	-2.45	***SLC15A4***	0.003	0.011	-0.96
TSPAN18	< 0.001	< 0.001	-3.38	***HIST1H3H***	0.003	0.012	-0.92
MYL9	< 0.001	< 0.001	-2.40	***UIMC1***	0.004	0.013	-0.59
FAXDC2	< 0.001	< 0.001	-1.86	***ZNF410***	0.006	0.020	-0.43
GNAZ	< 0.001	< 0.001	-2.59	***EGR1***	0.007	0.021	-1.06
RAB27B	< 0.001	< 0.001	-2.07	***LPAR1***	0.007	0.022	-1.16
ENKUR	< 0.001	< 0.001	-2.67	***ZC2HC1A***	0.008	0.025	-0.96
TREML1	< 0.001	< 0.001	-2.32	***GRB14***	0.009	0.026	-1.75
NFIB	< 0.001	< 0.001	-2.21	***ALOX12***	0.017	0.048	-1.55
GNG11	< 0.001	< 0.001	-2.06	***CD63***	0.017	0.048	-0.57
DAB2	< 0.001	< 0.001	-2.14	***AIG1***	0.018	0.048	-0.37
JUP	< 0.001	< 0.001	-1.41	***HIST1H4A***	0.025	0.064	0.82
GFI1B	< 0.001	< 0.001	-1.31	***FZD6***	0.032	0.080	-0.99
SLC45A3	< 0.001	< 0.001	-1.85	***PES1***	0.042	0.102	-0.47

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to achieving 12 month MMR.

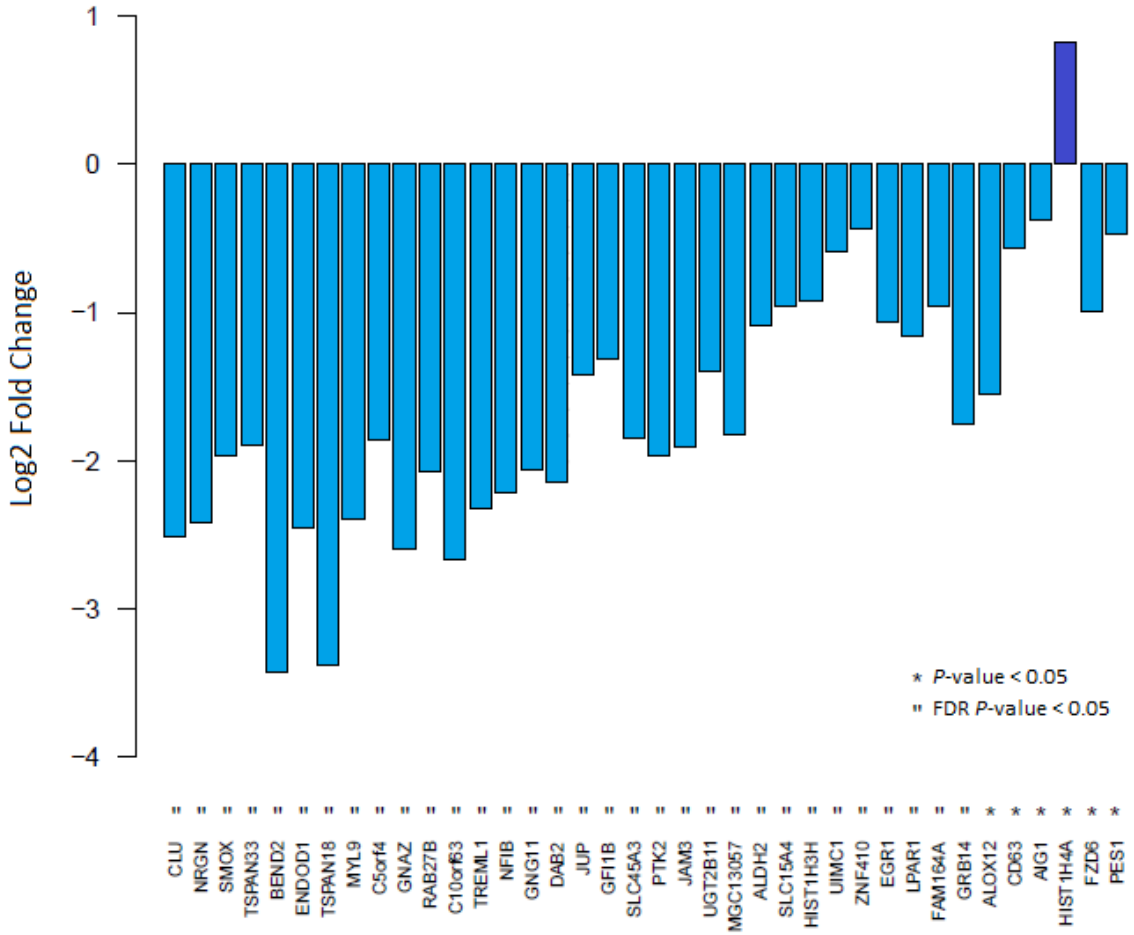


Figure 4.14 Significant gene expression between CP-CML patients who do, or do not, achieve MMR by 12 months using TWC mRNA

The log2-transformed fold change (in relation to not achieving MMR) values corresponding to the top 38 differentially expressed genes associated with 12 month MMR (BCR-ABL1 mRNA level < 0.1%) in CP-CML patients, with a *P*-value < 0.05 were plotted. Thirty-seven genes demonstrated increased expression in CP-CML patients who achieved 12 month MMR, while 1 gene had increased expression in CP-CML patients who did not achieve 12 month MMR. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

for multiple testing (BH-FDR), 35 genes remained significantly up-regulated in CP-CML patients who achieved MMR by 12 months, compared to those who did not (**Table 4.7**; **Figure 4.14**), indicating that the vast majority of these genes were highly significant.

The achievement of 24 month MMR was then evaluated, with patients divided by those who did (n = 28), or did not (n = 12), achieve MMR by 24 months on TKI therapy. Thirty-six genes were differentially expressed with a *P*-value < 0.05; 35 of which had significantly increased expression in CP-CML patients who achieved MMR by 24 months, while 1 gene was significantly increased in patients who failed to achieve MMR by 24 months (**Table 4.8**; **Figure 4.15**). When corrected for multiple testing (BH-FDR), 30 genes were still significantly up-regulated and 1 gene significantly down-regulated in CP-CML patients who achieved MMR by 24 months, compared to those who did not (**Table 4.8**; **Figure 4.15**), indicating that the vast majority of these genes were highly significant. Interestingly, when the significant genes from the 12 month and 24 month MMR gene-sets were compared, 33 genes overlapped (**Figure 4.16**), demonstrating that this core set of genes were predictive for MMR, without time-point bias. Additionally, when the EMR significant gene list was also compared, 13 genes were common between all 3 gene-sets (**Figure 4.16**), further indicating a core set of genes predictive of molecular response.

Finally, the correlation between gene expression and transformation to BC while on TKI therapy was evaluated, with patients divided by those who did (n = 3), or did not (n = 37), transform to BC within 24 months of starting TKI therapy. Notably, of the 3 patients who transformed, two did so quite early (within the first 5 months of TKI therapy), while the third transformed after 13 months of TKI therapy. Forty-one genes were differentially expressed (*P*-value < 0.05); 40 of which had significantly increased expression in CP-CML patients who

Table 4.8 Genes with differential expression in CP-CML patients based on the achievement of MMR by 24 months with HTqPCR analysis

Gene	*P*-value	FDR *P*-value	LogFC	Gene	*P*-value	FDR *P*-value	LogFC
SMOX	< 0.001	< 0.001	-1.88	***ENKUR***	0.001	0.007	-1.81
CLU	< 0.001	< 0.001	-2.15	***DAB2***	0.002	0.009	-1.70
FAXDC2	< 0.001	< 0.001	-1.89	***GNG11***	0.002	0.010	-1.47
NRGN	< 0.001	< 0.001	-2.12	***JAM3***	0.003	0.015	-1.59
BEND2	< 0.001	< 0.001	-3.27	***GFI1B***	0.004	0.017	-0.92
TSPAN18	< 0.001	< 0.001	-3.12	***NFIB***	0.005	0.018	-1.32
ENDOD1	< 0.001	< 0.001	-2.17	***SLC15A4***	0.005	0.020	-1.00
MYL9	< 0.001	< 0.001	-2.01	***CD63***	0.006	0.022	-0.69
TSPAN33	< 0.001	< 0.001	-1.44	***UGT2B11***	0.008	0.029	-1.16
HIST1H4A	< 0.001	0.001	1.34	***AIG1***	0.009	0.032	-0.42
GNAZ	< 0.001	0.001	-2.26	***ALOX12***	0.010	0.032	-1.78
JUP	< 0.001	0.001	-1.33	***CD59***	0.013	0.041	-0.60
GRB14	< 0.001	0.002	-2.19	***C2orf88***	0.015	0.046	-1.49
SLC45A3	< 0.001	0.002	-1.69	***ZNF410***	0.018	0.054	-0.39
TREML1	< 0.001	0.003	-1.78	***RPIA***	0.026	0.076	-0.37
RAB27B	< 0.001	0.004	-1.49	***FZD6***	0.032	0.090	-1.11
ALDH2	0.001	0.006	-1.32	***ZC2HC1A***	0.041	0.113	-0.78
PTK2	0.001	0.006	-1.89	***P4HB***	0.043	0.116	-0.59

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to achieving 24 month MMR.

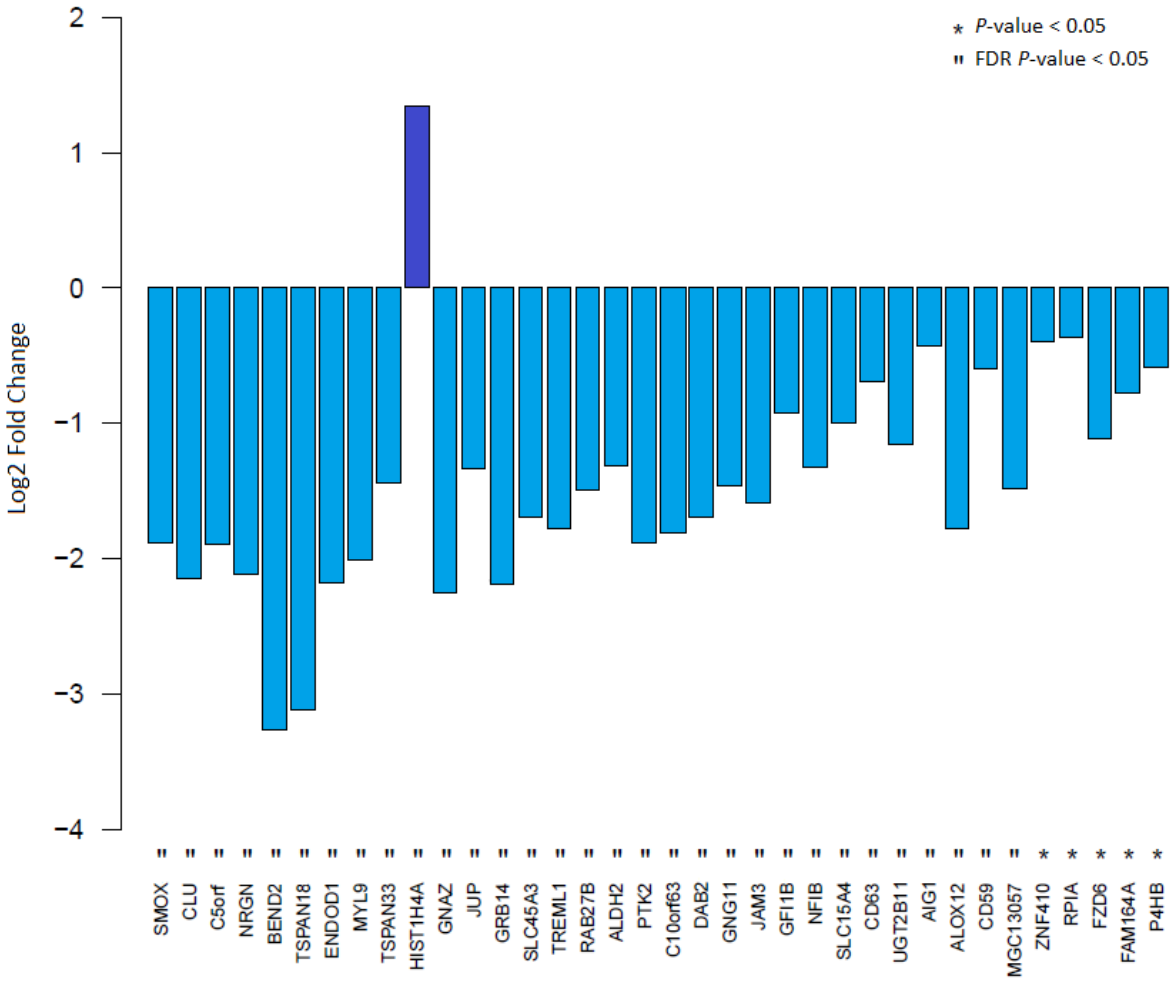


Figure 4.15 Significant gene expression between CP-CML patients who do, or do not, achieve MMR by 24 months using TWC mRNA

The log2-transformed fold change (in relation to not achieving MMR) values corresponding to the top 36 differentially expressed genes associated with 24 month MMR in CP-CML patients, with a *P*-value < 0.05 were plotted. Thirty-five genes demonstrated increased expression in CP-CML patients who achieved 24 month MMR, while 1 gene had increased expression in CP-CML patients who did not achieve 24 month MMR. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

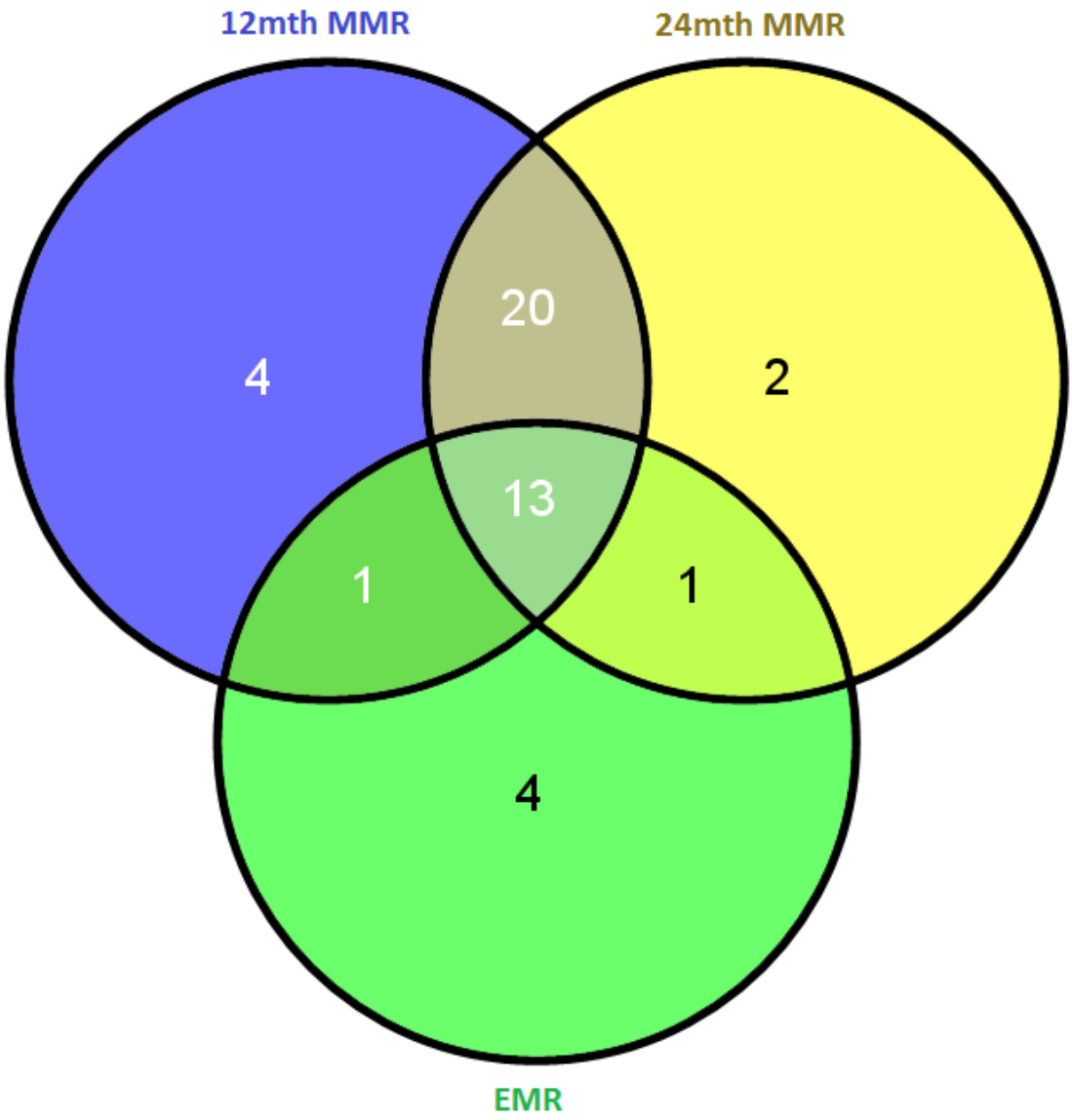


Figure 4.16 Significant overlap was observed between the significant gene-sets identified from the 12 month MMR, the 24 month MMR and the EMR analyses

Venn diagram demonstrating the overlap between the significant genes identified from the 12 month MMR analysis (**Blue**), compared to 24 month MMR analysis (**Yellow**), and the EMR analysis (**Green**). The Venn diagram was developed using Venny (http://bioinfogp.cnb.csic.es/tools/venny/).[313]

did not transform to BC, while 1 gene was significantly increased in patients who did transform (**Table 4.9**; **Figure 4.17**). Thirty-two genes were still significantly up-regulated in CP-CML patients who did not transform to BC on TKI therapy, compared to those who did not when corrected for multiple testing (BH-FDR; **Table 4.9** and **Figure 4.17**). However, as there were only 3 patients who transformed to BC in this cohort, these results must be interpreted with caution.

4.2.5 Genes significantly associated with multiple patient response parameters were identified from the 96 gene RQ-PCR dataset

The MMR gene-set (common genes from both the 12 and 24 month analysis) and the BC transformation gene-set were compared to determine if there were any common genes enriched in both datasets. Nine genes were identified; 8 of which were consistently up-regulated in patients who achieved MMR and patients who did not progress to BC, while 1 gene was consistently down-regulated in these patient groups (**Table 4.10**). Although no common pathways were enriched when pathway analysis was performed, it is interesting to note that HIST1H4A, a member of the histone H4 family, was consistently the only gene up-regulated in the poor response group (patients who failed to achieve MMR by 12 or 24 months; or who transformed to BC on TKI therapy). Furthermore, HIST1H4A was not differentially expressed between patients who did, or did not, achieve EMR, indicating that it may be better suited as a longer-term response predictor. However, when OA was investigated, HIST1H4A up-regulation was once again significantly associated with poor response (very low OA), as was the expression of a number of other genes of the histone H2A, H2B and H3 families. This demonstrates that histone gene enrichment was most significant for the OA analysis, suggesting that histones may influence OA through a yet to be determined mechanism.

Table 4.9 Genes with differential expression in CP-CML patients who transformed to BC on TKI therapy by HTqPCR analysis

Gene	*P*-value	FDR *P*-value	LogFC	Gene	*P*-value	FDR *P*-value	LogFC
CD63	< 0.001	< 0.001	-2.21	***GBE1***	0.003	0.012	-2.11
CPNE3	< 0.001	< 0.001	-1.93	***ENPP4***	0.003	0.012	-2.48
SLC2A5	< 0.001	< 0.001	-3.76	***JUNB***	0.004	0.015	-3.24
GRB14	< 0.001	< 0.001	-2.77	***TSPAN18***	0.005	0.018	-3.59
CDCA7L	< 0.001	< 0.001	-3.06	***HMGN1***	0.006	0.022	-2.60
SLC15A4	< 0.001	< 0.001	-2.83	***DEFA4***	0.007	0.024	-1.06
PIWIL4	< 0.001	< 0.001	-2.62	***HSPA4L***	0.008	0.025	-4.44
CAMP	< 0.001	< 0.001	-1.84	***SMOX***	0.013	0.039	-1.64
RPIA	< 0.001	< 0.001	-1.06	***CHIT1***	0.014	0.042	-3.72
P4HB	< 0.001	< 0.001	-2.52	***FZD6***	0.016	0.047	-3.59
SCD	< 0.001	0.001	-4.76	***HIST1H2BJ***	0.017	0.049	-1.45
ALDH2	< 0.001	0.001	-3.71	***DDIT4***	0.020	0.055	-1.58
OLR1	< 0.001	0.002	-2.42	***HIST1H2BB***	0.021	0.055	-1.83
GPI	< 0.001	0.002	-1.44	***ANP32E***	0.032	0.084	-2.05
CD59	< 0.001	0.003	-2.08	***JAM3***	0.033	0.084	-3.14
AZU1	< 0.001	0.004	-2.69	***OR2L8***	0.034	0.084	-2.09
GFI1	< 0.001	0.004	-1.51	***MT-ND2***	0.035	0.085	-1.83
OLFM4	< 0.001	0.004	-2.21	***ENDOD1***	0.041	0.092	-2.49
RETN	< 0.001	0.004	-3.83	***HIST1H4A***	0.041	0.092	1.20
CWH43	0.001	0.005	-4.17	***NIPSNAP3A***	0.043	0.094	-2.30
SUCNR1	0.002	0.007	-1.75				

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to patients who did not transform to BC.

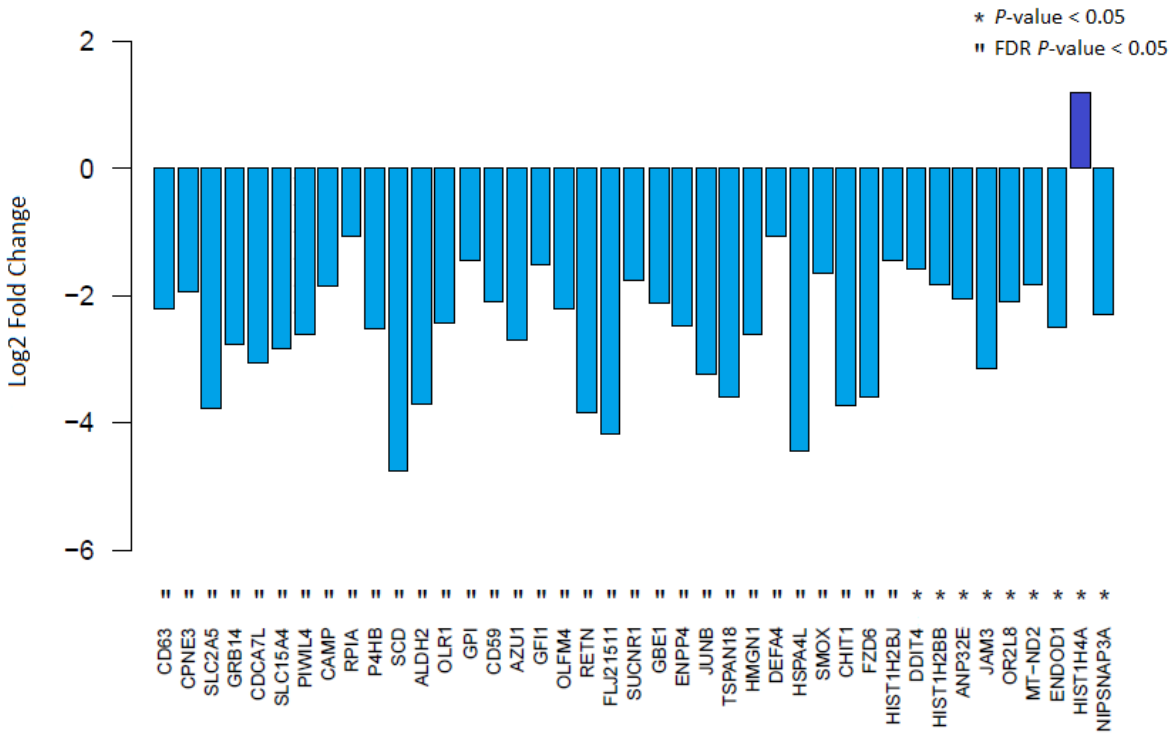


Figure 4.17 Significant gene expression between CP-CML patients who did, or did not, transform to BC on TKI therapy using TWC mRNA

The log2-transformed fold change (in relation to transforming to BC) values corresponding to the top 41 differentially expressed genes associated with transformation to BC on TKI therapy in CP-CML patients, with a *P*-value < 0.05 were plotted. Forty genes demonstrated increased expression in CP-CML patients who did not transform to BC, while 1 gene had increased expression in CP-CML patients who did transform. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

Table 4.10 Common genes between the MMR gene-set and the BC transformation gene-set

Gene	12 month MMR		24 month MMR		BC Transformation	
	P-value	LogFC	*P*-value	LogFC	*P*-value	LogFC
CD63	0.017	-0.57	0.006	-0.69	< 0.001	-2.21
ENDOD1	< 0.001	-2.45	< 0.001	-2.17	0.041	-2.49
FZD6	0.032	-0.99	0.032	-1.11	0.016	-3.59
GRB14	0.009	-1.75	< 0.001	-2.19	< 0.001	-2.77
HIST1H4A	0.025	0.82	< 0.001	1.34	0.041	1.2
JAM3	< 0.001	-1.91	0.003	-1.59	0.033	-3.14
SLC15A4	0.003	-0.96	0.005	-1.00	< 0.001	-2.83
SMOX	< 0.001	-1.97	< 0.001	-1.88	0.013	-1.64
TSPAN18	< 0.001	-3.38	< 0.001	-3.12	0.005	-3.59

Abbreviations: logFC – log2 fold change. *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to patients who did achieve MMR (by either 12 or 24 months), or who did not transform to BC.

4.2.6 Development of a gene expression classifier for OCT-1 activity

To further determine the genes with the greatest predictive power for differentiating very low and very high OA patients, a predictive classifier was developed using machine learning algorithms and the TLDA RQ-PCR data. Two different machine learning packages were employed in the statistical environment *R*; iterative Bayesian Model Averaging (iterativeBMA)[275] and CMA,[226] allowing the best classifier to be selected.

4.2.6.1 Development of an OA classifier using a 35 patient training set and 5 patient test set with iterativeBMA

IterativeBMA was applied to the TLDA RQ-PCR dataset (reference genes were excluded) as previously described in **Chapter 3**, to identify signature genes predictive for very low OA patients (**Figure 4.18**). The 40 patients analysed by TLDA RQ-PCR were randomly assigned into either a training set (n = 35) or test set (n = 5). Although small with only 5 patients, the purpose of this initial test set was to gain some insight into the accuracy of the classifier before the arduous and expensive task of performing independent validation with a larger cohort of patients as the test set was undertaken. Eleven genes were identified by iterativeBMA over 19 models that discriminated very low and very high OA patients in our training set (**Table 4.11**). Using univariate methods, the top 11 ranked genes were selected, all of which corresponded to the genes selected by the multivariate iterativeBMA method (**Table 4.11**). The posterior probabilities were computed for every gene by averaging the posterior probabilities of selected models in which each gene was included. Membership of the 11 signature genes in each of the 19 models selected by the iterativeBMA algorithm is demonstrated in **Figure 4.19**.

Next, to determine the predicted probability of patients being classified as having very low or very high OA for each of the 5 test patient samples, the average predicted probabilities from the 19 selected models, weighted by the posterior probabilities, were calculated. The

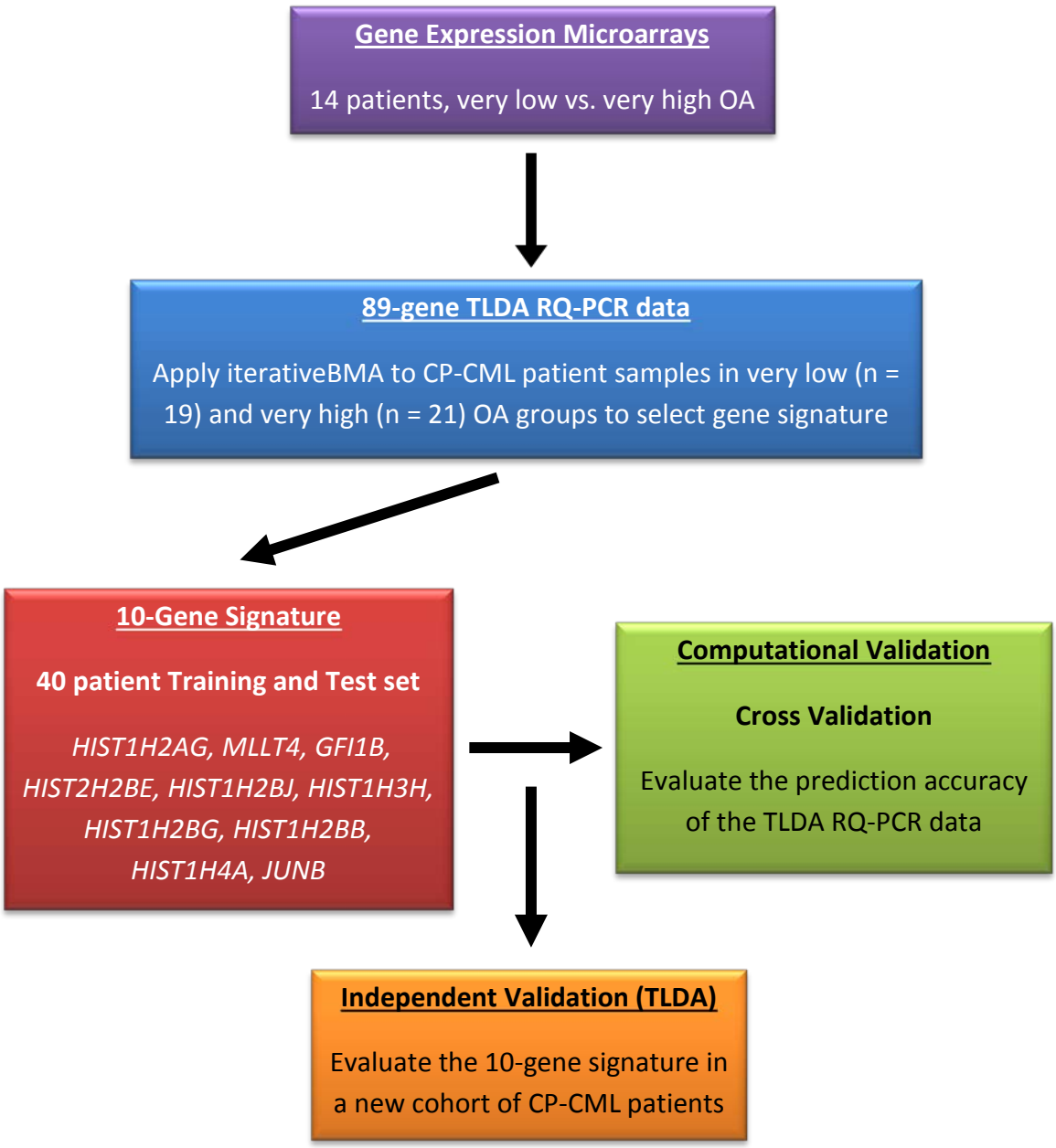


Figure 4.18 IterativeBMA 89-gene TLDA RQ-PCR classification overview

A summary of the overall approach used for OA classification with the iterativeBMA algorithm is shown. IterativeBMA is able to produce a high predictive accuracy with relatively few genes, using multivariate modelling and cross validation. The identified 10-gene signature (from the 40 patient training and test [internal validation] dataset; **Section 4.2.6.2**) will be validated by TLDA RQ-PCR using an independent CP-CML patient cohort.

Table 4.11 The 11-gene signature discriminating very low from very high OA patients as selected by a 35 patient training set with iterativeBMA

ID	Gene	Chr Location	Cellular Function	Posterior Probability (%)	Univariate (BSS/WSS) Rank
NM_021064	*HIST1H2AG*	6p22.1	Nucleosome assembly	98.6	1
NM_003528	*HIST2H2BE*	1q21.2	Nucleosome assembly	9.7	2
NM_003536	*HIST1H3H*	6p22.1	Nucleosome assembly	8.9	3
NM_003538	*HIST1H4A*	6p22.1	Nucleosome assembly	2.2	4
NM_003518	*HIST1H2BG*	6p21.3	Nucleosome assembly	4.5	5
NM_004188	*GFI1B*	9q34.13	Transcription regulation	34.9	6
NM_016108	*AIG1*	6q24.2	Unknown	100	7
NM_001040000	*MLLT4*	6q27	Signal transduction	44.6	8
NM_006418	*OLFM4*	13q14.3	Cell adhesion	2	9
NM_021062	*HIST1H2BB*	6p21.3	Nucleosome assembly	4.3	10
NM_021058	*HIST1H2BJ*	6p22.1	Nucleosome assembly	9.7	11

The table indicates chromosomal location, function, the posterior probability of each gene in multivariate modelling, and its rank in univariate modelling. OA indicates OCT-1 activity; BMA, Bayesian model averaging; Chr, chromosome; and BSS/WSS, the ratio of between-group to within-group sum of squares.

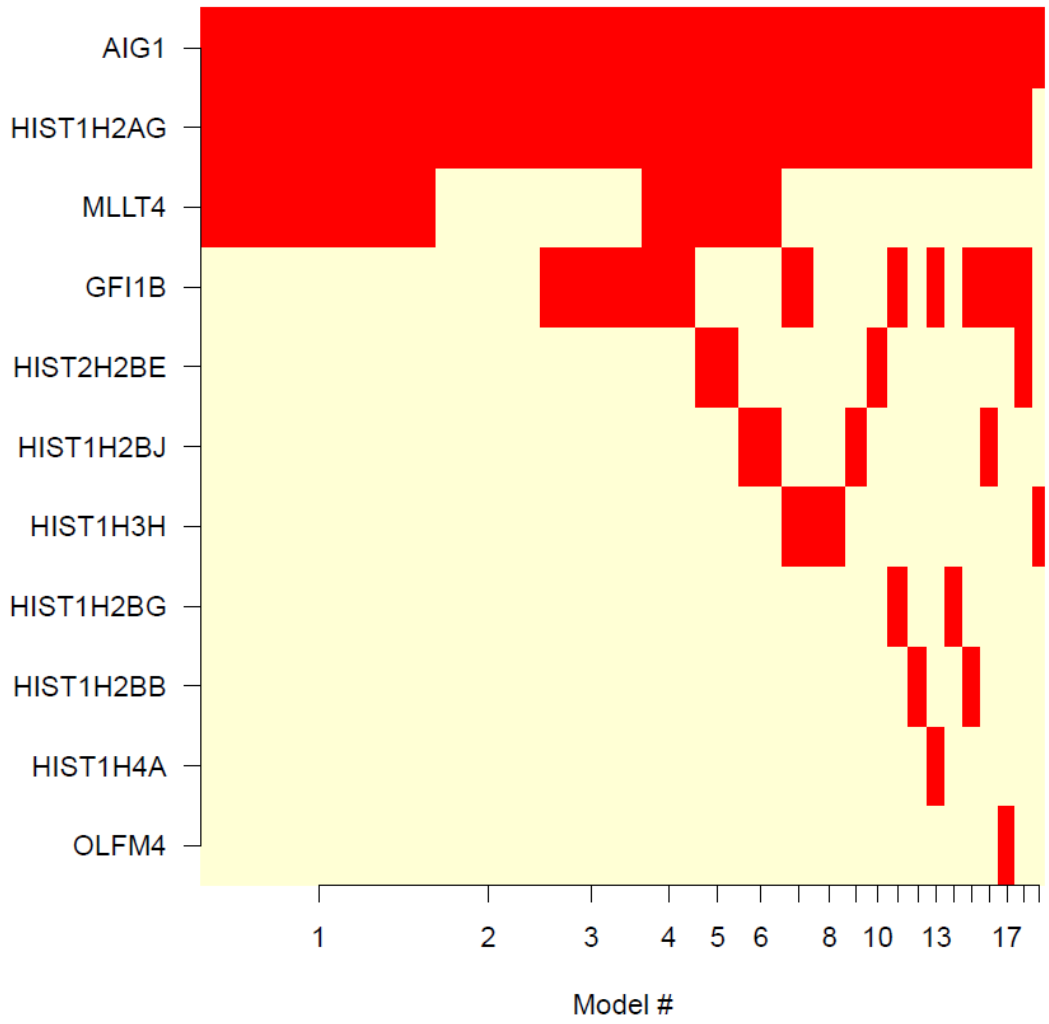


Figure 4.19 Model selection by iterativeBMA using an 11-gene signature

Membership of the 11 signature genes in the 19 models selected by the iterativeBMA algorithm, using a 35 patient training set derived from the 96-gene TLDA RQ-PCR dataset. The 11 signature genes are plotted on the vertical axis, with the 19 models on the horizontal axis. The widths of the column are proportional to the posterior probabilities of the selected models, with a red entry indicating when the corresponding gene is included in that particular model.

predicted probability threshold was set at 0.5; therefore if the predicted probability was less than 0.5, patients were classified as very high OA, otherwise, if the predicted probability was greater than 0.5, patients were classified as very low OA. Two of the 5 patients were misclassified using the iterativeBMA gene signature (**Figure 4.20**), generating an overall accuracy of 60%, with a Brier score of 1.46. The Brier score is a relative probabilistic measure of the number of errors; hence, a relatively small number of errors correspond with a small Brier score, indicating higher prediction accuracy. However, as the number of patients in the test set was small, only an approximation of the overall predictive accuracy of the 11 signature genes can truly be estimated.

4.2.6.2 Development of an OA classifier using a 40 patient training set and internal validation with iterativeBMA

Therefore, the iterativeBMA algorithm was re-applied, with all 40 patients being used as both the training and test sets (internal validation), to obtain the best overall signature gene-set. While this method may over- or under-estimate the predictive accuracy, it allows the smallest set of signature genes to be identified. Ten genes were identified over 38 models that discriminated very low from very high OA patients using the 40 patient training set (**Table 4.12**). Univariate analysis also ranked the same 10 genes as the top ranked genes, and the posterior probabilities were computed for each gene as before. Membership of the 10 signature genes in each of the 38 models selected by the iterativeBMA algorithm is demonstrated in **Figure 4.21**. Next, the predicted probability (of being very low or very high OA) for the same 40 patient samples was determined, with the predicted probability threshold once again set at 0.5. Ten of the 40 patients were misclassified using this iterativeBMA gene signature (3 very high and 7 very low OA patients; **Figure 4.22**), generating an overall accuracy of 75%, with a Brier score of 6.02. Furthermore, 9 genes were common between both signature gene-sets (35 patient training set and 40 patient training

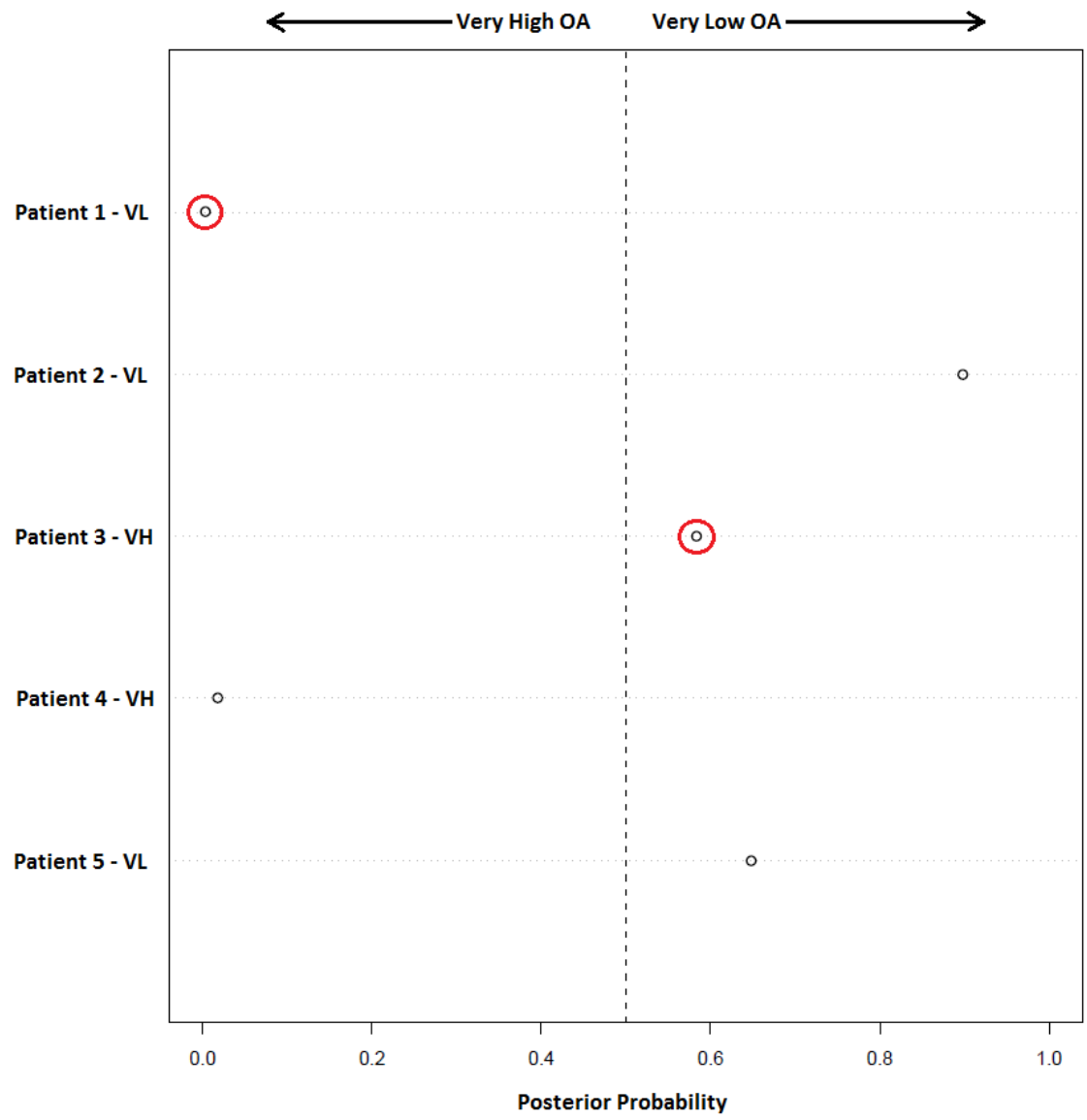


Figure 4.20 The 11-gene classifier is 60% accurate using an independent test set

Results of the validation of the 11-gene classifier developed by iterativeBMA (n = 35 training set) using an independent test set (n = 5). A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate a very high OA patient. Two patients were misclassified (**red** circles; 1 very low and 1 very high OA patient), providing an overall accuracy of 60%. VL – very low OA; VH – very high OA.

Table 4.12 The 10-gene signature discriminating very low from very high OA patients as selected by a 40 patient training set with iterativeBMA

ID	Gene	Chr Location	Cellular Function	Posterior Probability (%)	Univariate (BSS/WSS) Rank
NM_021064	*HIST1H2AG*	6p22.1	Nucleosome assembly	89	1
NM_003536	*HIST1H3H*	6p22.1	Nucleosome assembly	45.6	2
NM_003528	*HIST2H2BE*	1q21.2	Nucleosome assembly	9.4	3
NM_003538	*HIST1H4A*	6p22.1	Nucleosome assembly	16.8	4
NM_003518	*HIST1H2BG*	6p21.3	Nucleosome assembly	7.1	5
NM_021062	*HIST1H2BB*	6p21.3	Nucleosome assembly	4.7	6
NM_021058	*HIST1H2BJ*	6p22.1	Nucleosome assembly	4.8	7
NM_001040000	*MLLT4*	6q27	Signal transduction	48.1	8
NM_004188	*GFI1B*	9q34.13	Transcription regulation	70	9
NM_002229	*JUNB*	19p13.2	Transcription regulation	7.4	10

The table indicates chromosomal location, function, the posterior probability of each gene in multivariate modelling, and its rank in univariate modelling. OA indicates OCT-1 activity; BMA, Bayesian model averaging; Chr, chromosome; and BSS/WSS, the ratio of between-group to within-group sum of squares.

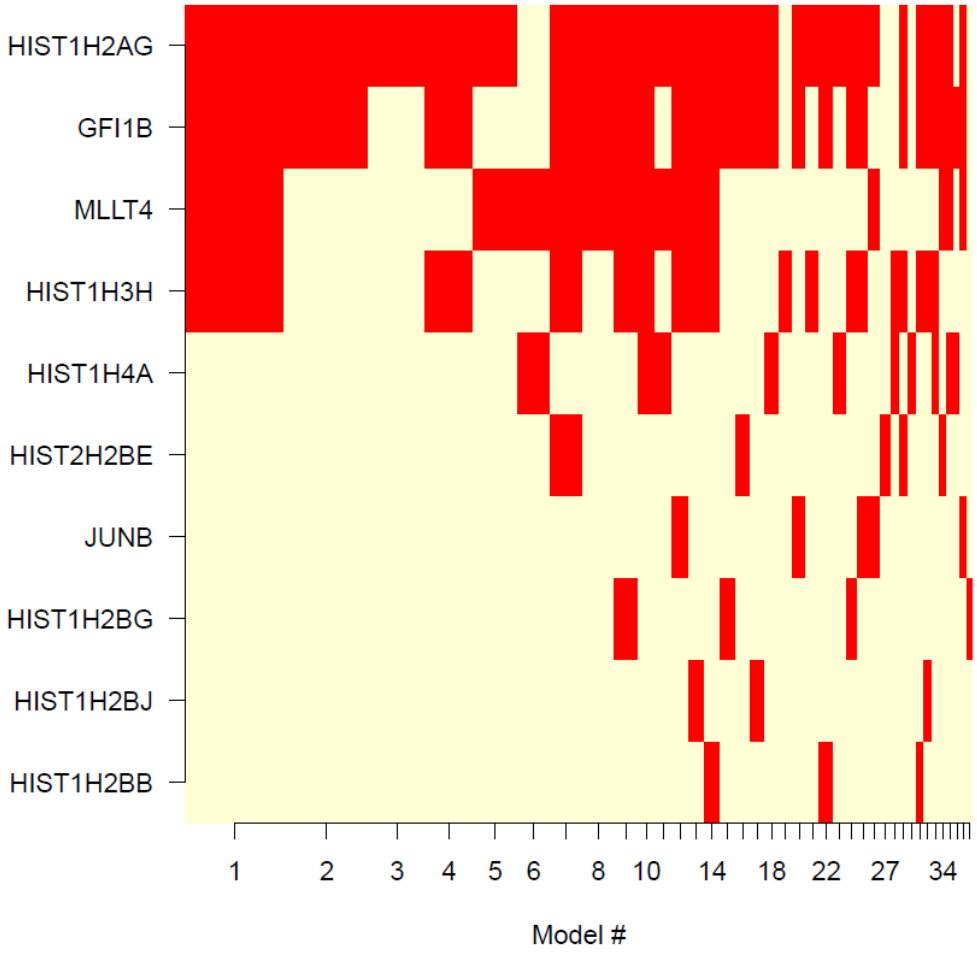


Figure 4.21 Model selection by iterativeBMA using a 10-gene signature

Membership of the 10 signature genes in the 38 models selected by the iterativeBMA algorithm, using a 40 patient training set (all patients) derived from the TLDA RQ-PCR dataset. The 10 signature genes are plotted on the vertical axis, with the 38 models on the horizontal axis. The widths of the column are proportional to the posterior probabilities of the selected models, with a red entry indicating when the corresponding gene is included in that particular model.

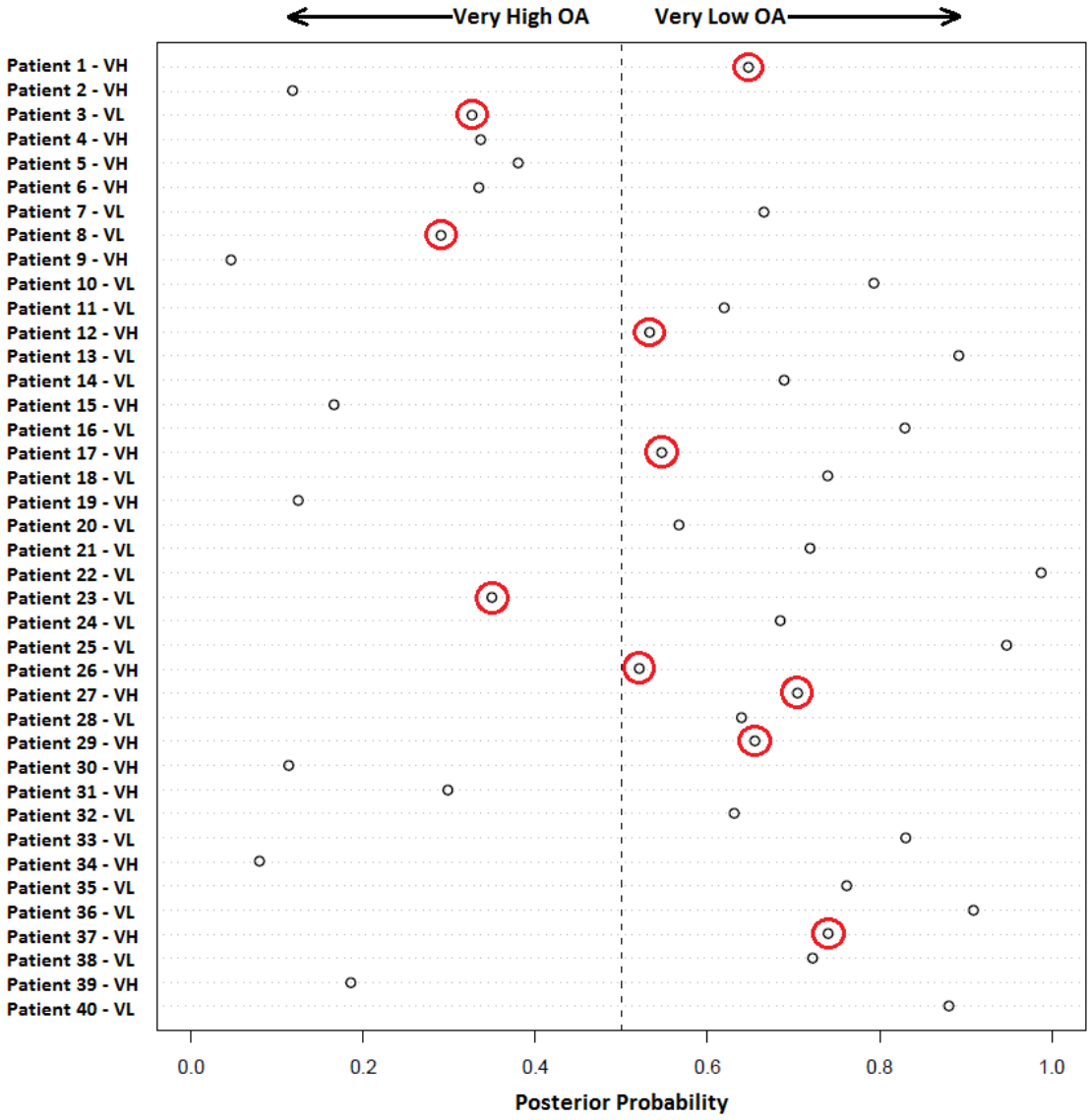


Figure 4.22 The 10-gene classifier is 75% accurate against a test set comprising the same patients

Results of the validation of the 10-gene classifier developed by iterativeBMA (n = 40 training set), tested against the same 40 patients. A posterior probability close to 1 should indicate a very low OA patient, while a posterior probability close to 0 should indicate a very high OA patient. Ten patients were misclassified (**red** circles; 3 very low and 7 very high OA patients), providing an overall accuracy of 75%, with a Brier score of 6.022. VL – very low OA; VH – very high OA.

set). However, as the overall accuracy of this classifier was lower than expected, a different package of classification algorithms, known as CMA, were analysed to determine if a superior classifier could be developed.

4.2.6.3 CMA was unable to develop an OA classifier with greater accuracy than iterativeBMA

The CMA (Classification for Microarrays) package[226,227] provides a comprehensive collection of various microarray-based classification algorithms that can be applied to high-throughput RQ-PCR data, such as that generated by TLDA analysis. Included are 21 different classification methods, comprising classical approaches (e.g. discriminant analysis), as well as more sophisticated methods (e.g. Support Vector Machines [SVM]). Three-fold cross-validation, repeated 5 times in order to generate stable results was performed to generate a training set containing two-thirds of the patients (n = 27). CMA was then used to perform a bundle of 6 different classification methods based on the principle of discriminant analysis: diagonal-, linear- and quadratic-discriminant analysis; discriminant analysis by Fisher; shrunken centroids discriminant analysis; and Partial Least Squares followed by linear discriminant analysis, against the remaining 13 "test" samples with mixed success (**Table 4.13**). Partial Least Squares plus linear discriminant analysis (pls_LDA) was the most successful classifier (lowest misclassification and highest average probability); however, the average misclassification was still 35.5% (average accuracy 64.5%). Therefore, none of these classifiers had superior predictive power for OA, compared to the 10-gene iterativeBMA classifier.

4.2.7 Selection of genes to be validated from the RQ-PCR analyses and OA classifier development

The analyses performed so far compared very low OA patients against very high OA patients,

Table 4.13 Summary of the 6 different discriminant analyses performed in CMA

	Misclassification	Brier Score	Average Probability
DLDA	0.430	0.779	0.570
LDA	0.438	0.750	0.538
FDA	0.453	0.491	0.505
QDA	0.415	0.830	0.576
SCDA	0.445	0.541	0.544
pls_LDA	0.355	0.511	0.602

Abbreviations: DLDA – diagonal discriminant analysis; LDA – linear discriminant analysis; FDA – Fisher's discriminant analysis; QDA – quadratic discriminant analysis; SCDA – shrunken centroid discriminant analysis (PAM); pls_LDA – partial least squares + linear discriminant analysis.

to identify genes with significant differential expression to generate gene classifiers that could specifically predict these patients. However, it remained unknown if these genes and classifiers could still be predictive for very low OA when patients from all four OA quartiles were included in the dataset. To do this, a new custom 24-gene TLDA card was designed, which would allow a new cohort of 80 CP-CML patients, covering all OA quartiles, to be investigated. To select the genes to be included in this 24-gene TLDA, all the gene-lists developed from the statistical tests and classification algorithms were combined, and the top 15 genes common between these methods were identified (**Table 4.14**). Four other interesting genes (*ENKUR*, *JAM3*, *C2orf88* and *NFIB*), which demonstrated differential expression between very low and very high OA patients, but were not included in the classification algorithms due to missing/incomplete data, were also selected. Furthermore, the *SLC22A1* (OCT-1) gene was included to verify if OCT-1 mRNA expression varied between these patients. Finally, 4 reference genes (*18S*, *B2M*, *GUSB* and *HMBS*) were also included for normalisation options (see **Appendix 6** for final TLDA configuration).

4.2.8 Validation of the OA classifier in an independent CP-CML patient cohort

RQ-PCR was performed on 80 CP-CML patients, using *de novo* TWC RNA samples and 24-gene TLDAs (see **Table 4.15** for patient details), to validate the 10-gene OA classifier developed by iterativeBMA. Using *GUSB* as the reference gene, the ΔCt values were determined using HTqPCR (**Figure 4.23**) as before, and the ΔΔCt values were calculated relative to "Other" (OA quartiles 2, 3 and 4) CP-CML patients. Nine samples failed quality control in HTqPCR and were excluded from further analysis (**Figure 4.24**; 1 sample failed to amplify during PCR and 8 samples [1 TLDA card] failed due to the PCR run being interrupted).

Table 4.14 Ranking of TLDA genes, in order of importance, by various statistical tests and classification algorithms

The top 15 genes were selected based on frequency and average rank, and are identified by the yellow and gold shading. Genes are ranked in descending order from highest to lowest rank for each statistical test or classification algorithm, with 1 indicating the top ranked gene. Abbreviations: LIMMA – linear models for microarray data, statistical analysis; KW – Kruskal-Wallis test, statistical test; RF – random forest, classification; BMA – iterative Bayesian Model Averaging, classification; PAM – prediction analysis of microarrays (shrunken centroid discriminant analysis), classification.

Gene	*Statistical Test or Classification Algorithm*						Frequency	Average Rank
	LIMMA	T-test	KW	RF	BMA	PAM		
HIST1H2AG	1	1	1	2	1	1	**6**	1.17
HIST2H2BE	2	2	2	4	6	2	**6**	3.00
HIST1H3H	3	3	3	3	4	4	**6**	3.33
HIST1H4A	5	4	4	5	5	5	**6**	4.67
HIST1H2BG	4	5	5	11	8	3	**6**	6.00
HIST1H2BB	6	6	11	8	10	6	**6**	7.83
HIST1H2BJ	9	8	8	7	9	10	**6**	8.50
MLLT4	7	9	14	12	3	8	**6**	8.83
JUNB	11	13	10	6	7	9	**6**	9.33
GFI1B	10	11	13	15	2	11	**6**	10.33
AIG1	12	7	6	9	---	19	**5**	10.60
CD59	13	12	7	10	---	18	**5**	12.00
OLFM4	8	10	12	---	---	7	**4**	9.25
SMOX	14	14	16	---	---	16	**4**	15.00
EGR1	17	15	17	---	---	12	**4**	15.25
CAMP	15	16	20	---	---	15	**4**	16.50
SLC2A5	16	17	---	---	---	13	**3**	15.33
CHIT1	19	---	18	---	---	14	**3**	17.00
RETN	---	19	15	18	---	---	**3**	17.33
LPAR1	18	18	---	---	---	17	**3**	17.67
ALDH2	---	---	9	1	---	---	**2**	5.00
PTK2	---	20	---	16	---	---	**2**	18.00
FZD6	---	---	19	20	---	---	**2**	19.50
AZU1	---	---	---	13	---	---	**1**	13.00
UIMC1	---	---	---	14	---	---	**1**	14.00
RAB27B	---	---	---	17	---	---	**1**	17.00
UGT2B11	---	---	---	19	---	---	**1**	19.00

Table 4.15 Characteristics of the 80 patients used for the 24-gene TLDA RQ-PCR

		Patient Cohort	
		Very low OA (n = 37)	Other OA (n = 42)
Median Age	**Years**	49	54
	Range	19 – 77	17 – 81
Male	**%**	51.3	52.4
Sokal Score	**High**	13	3
	Intermediate	7	19
	Low	17	18
BCR-ABL1 levels < 10%	**3 month, %**	86.5	97.6
MMR	**12 month, %**	48.6	73.8
	24 month, %	73.0	80.9
Transformation	**24 month, %**	2.7	0

One patient analysed by 24-gene TLDA did not have an OA value, hence this patient was excluded from further analysis. Two patients in the Other OA cohort did not have Sokal scores available. Other OA refers to all OA patients not included in the very low OA cohort. Abbreviations: OA – OCT-1 activity; MMR – major molecular response, BCR-ABL1 mRNA levels < 0.1%.

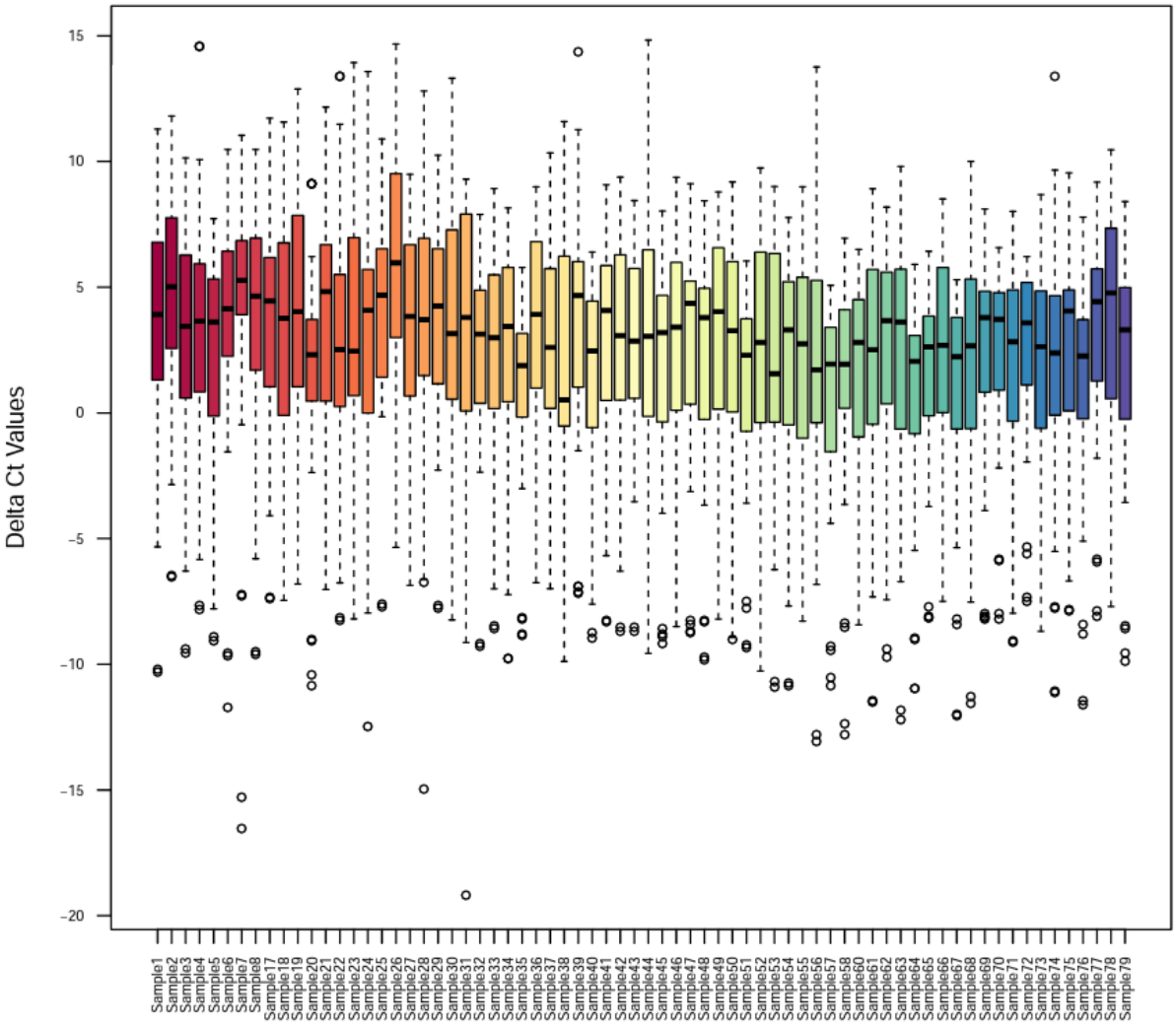


Figure 4.23 Boxplots demonstrating the range of Delta Ct values for 80 patients after normalisation with HTqPCR from 24-gene TLDA

The Delta Ct (ΔCt) values for all 20 included test genes, normalised to *GUSB,* were plotted for each of the 80 patients analysed by TLDA RQ-PCR. Boxplots represent the 25th to 75th percentile, with 50th percentile (median) represented by the centreline. Whiskers represent 1.5*IQR (interquartile range), providing roughly a 95% confidence interval, with outliers plotted as the open circles.

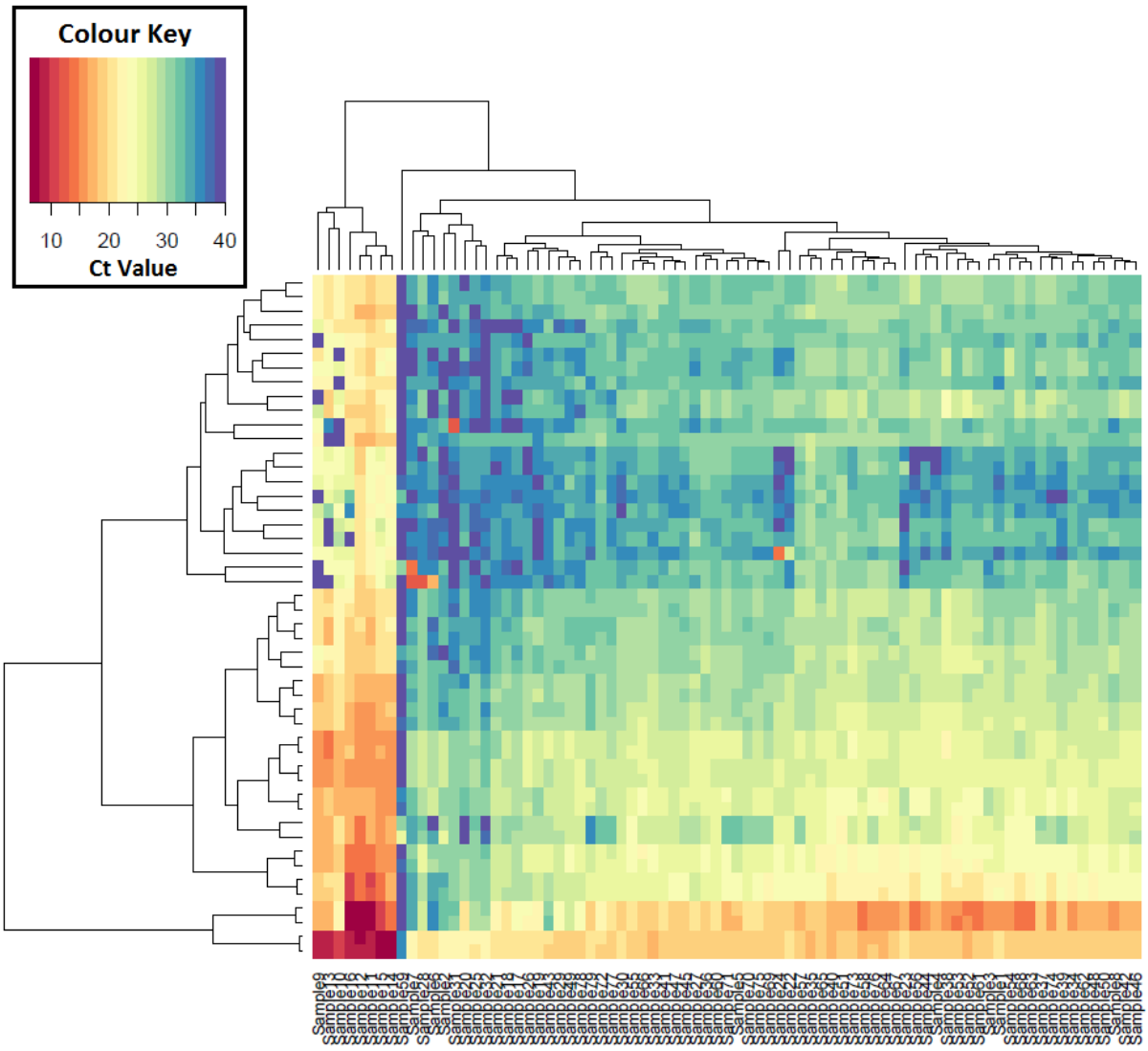


Figure 4.24 Unsupervised clustering of the 24-gene TLDA Ct values identifies 9 patients that failed QC

The raw Ct values from each gene (48; 24 genes in duplicate) are plotted. Each column represents one sample, and each row represents one gene. Unsupervised hierarchical clustering (Euclidean distance) was performed using HTqPCR. Nine samples clustered on the left, due to divergent gene expression patterns. These samples subsequently failed QC (high variation in Ct values due to failed PCR amplification) and were excluded from further analysis.

One further sample was excluded as no OA value was available, leaving 70 samples for the final analysis (n = 35 very low OA, and n = 35 Other OA; refer **Figure 4.25** for data analysis workflow).

In order to validate the 10 gene classifier developed by iterativeBMA (refer **Section 4.2.6**), the 70 patients with TLDA RQ-PCR data that passed QC were used as an independent test set against the original 40 patient training set (refer **Section 4.2.3**). IterativeBMA was performed as before, using the specific 10 gene classifier originally developed and the new 70 patient test set. Sixty-four of the 70 patients were able to be used in the final analysis, as 6 patients had missing values for at least one of the genes in the classifier. Unfortunately, the classifier was unable to be validated with this new, independent test set, as 35/64 patients were misclassified, providing a Brier score of 20.31 and overall accuracy of only 45.31%. The classifier was slightly more accurate when predicting the "Other" OA patients (20/32, 62.5%), when compared to the very low OA patients (9/32, 28.13%).

The classifier may have failed to validate for a number of reasons. However, the main reason was most likely related to the fact that the classifier was originally developed using a patient cohort consisting of CML patients from the extreme OA quartiles (very low versus very high OA); while the validation cohort comprised patients from all four OA quartiles. As the classifier was not optimised to discriminate very low OA patients from the other three quartiles (only patients from the outermost quartiles [very low and very high OA] were included in the original training set; refer **Table 4.3**), this may explain why the accuracy was so poor and it failed to validate successfully. Although this data was unable to produce an accurate classifier for OA, it can be used to identify possible biomarkers for other criteria, including various response markers, as discussed below.

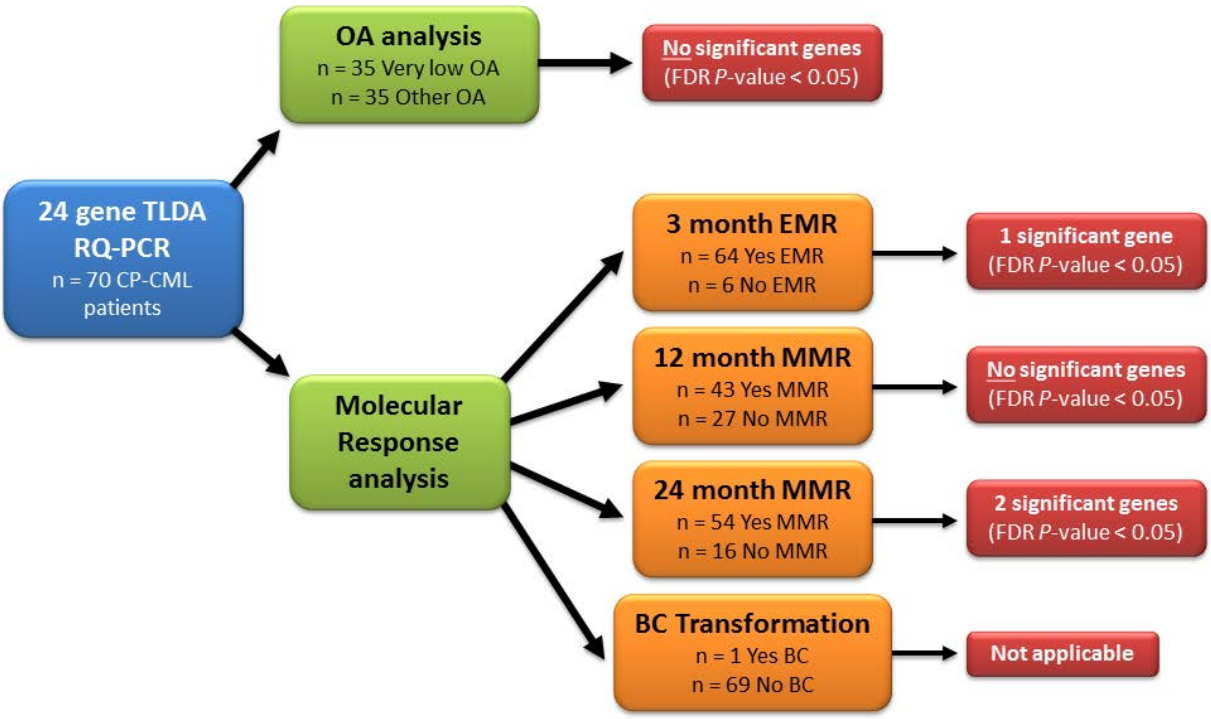


Figure 4.25 Workflow for analysis of the 24-gene RQ-PCR data

The **blue** box highlights the initial dataset used for the HTqPCR analysis. The **green** boxes indicate the two main analyses performed; firstly OA, followed by molecular response. The **orange** boxes demonstrate the particular response parameters investigated, while the **red** boxes indicated the number of statistically significant genes identified. EMR – early molecular response (3 month BCR-ABL1 mRNA levels < 10%); MMR – major molecular response (BCR-ABL1 mRNA levels < 0.1%); BC – blast crisis.

4.2.9 Interrogation of the 24 gene RQ-PCR dataset identifies genes predictive of response

Differential gene expression was determined using a 2-tailed *t*-test implemented in HTqPCR and only genes with a $P < 0.05$ were considered statistically significant. Intriguingly, of the 20 genes analysed by TLDA RQ-PCR, only 1 gene was significantly up-regulated in the Other OA patient group, compared to very low OA patients (*SLC22A1*, OCT-1; $P = 0.046$). No other gene demonstrated significant differential expression, and when corrected for multiple testing (using BH-adjusted FDR), *SLC22A1* lost significance ($P = 0.61$), indicating that differential *SLC22A1* expression could not be confirmed in this cohort. Importantly, it should be noted that the genes included in this 24-gene TLDA analysis were not necessarily selected based on exhibiting significant differential expression between patients with very low and very high OA. Rather, they were selected as they either contributed to the OA classifier or were ranked highly by statistical analyses, but notably, for both of these criteria the genes did not necessarily exhibit differential expression. Therefore, the observed lack of genes with significantly differential expression between CP-CML patients with very low OA, compared to the rest, was not completely unexpected.

Next, correlation with EMR was analysed. Patients were divided by those who did (n = 64), or did not (n = 6), achieve 3 month BCR-ABL1 mRNA levels < 10% on TKI therapy. Two genes were differentially expressed with a *P*-value < 0.05; both of which had significantly increased expression in CP-CML patients who achieved EMR by 3 months (**Table 4.16**; **Figure 4.26**). However, when corrected for multiple testing (BH-FDR), only 1 gene was still significantly up-regulated in CP-CML patients who achieved EMR by 3 months (*HIST1H3H*; **Table 4.16**).

The achievement of 12 month MMR was subsequently evaluated, and again, patients were divided by those who did (n = 43), or did not (n = 27), achieve MMR by 12 months on TKI

Table 4.16 Genes with significant differential expression in CP-CML patients based on achievement of EMR by HTqPCR analysis using 24-gene TLDA data

Gene	*P*-value	FDR *P*-value	Log FC
HIST1H3H	0.001	0.028	-0.62
AIG1	0.008	0.093	-0.37

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to achieving EMR.

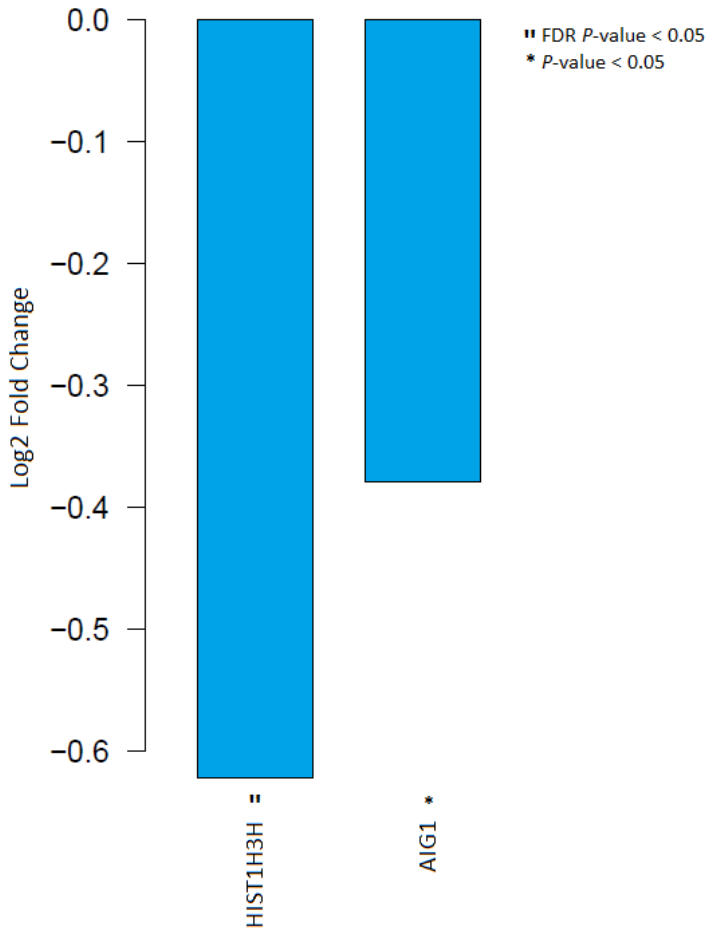


Figure 4.26 Significant gene expression between CP-CML patients who did, or did not, achieve EMR using 24-gene TLDA

Two differentially expressed genes associated with EMR (3 month BCR-ABL1 mRNA levels < 10%) in CP-CML patients, with a *P*-value < 0.05 were plotted. Both genes demonstrated increased expression in CP-CML patients who achieved EMR. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

therapy. Employing the same significance criteria as before (*P*-value < 0.05) identified 3 genes that were differentially expressed; 2 of which had significantly increased expression in CP-CML patients who achieved MMR by 12 months, while 1 gene was significantly increased in patients who failed to achieve MMR by 12 months (**Table 4.17**). Yet, when corrected for multiple testing (BH-FDR), all 3 genes lost significance; indicating that these results were unlikely to be statistically valid.

The achievement of 24 month MMR was evaluated, with patients divided by those who did (n = 54), or did not (n = 16), achieve MMR by 24 months on TKI therapy. Eight genes were differentially expressed with a *P*-value < 0.05, all of which had significantly increased expression in CP-CML patients who failed to achieve MMR by 24 months (**Table 4.18**; **Figure 4.27**). When corrected for multiple testing (BH-FDR), 2 genes remained significantly down-regulated in CP-CML patients who achieved MMR by 24 months, compared to those who did not (**Table 4.18**; **Figure 4.27**); indicating that these genes had high statistical confidence. Notably, 3 patients had achieved MMR during TKI treatment, but failed to remain in MMR at the 24 month time-point. Therefore, the achievement of MMR at the exact 24 month time-point was evaluated; however, no significant difference in gene expression between the patients who were (n = 55), or were not (n = 15), in MMR at 24 months on TKI therapy was observed. Finally, as only 1 patient in this cohort transformed to BC while on imatinib therapy, the correlation between gene expression and transformation to BC was unable to be evaluated.

Table 4.17 Genes with significant differential expression in CP-CML patients based on achievement of 12 month MMR by HTqPCR analysis using 24-gene TLDA data

Gene	*P*-value	FDR *P*-value	Log FC
HIST1H2BB	0.008	0.106	0.93
AIG1	0.009	0.106	-0.45
GFI1B	0.022	0.178	-0.72

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to achieving 12 month MMR.

Table 4.18 Genes with significant differential expression in CP-CML patients based on the achievement of 24 month MMR by HTqPCR analysis using 24-gene TLDA data

Gene	*P*-value	FDR *P*-value	Log FC
HIST1H2BB	< 0.001	0.003	1.38
OLFM4	< 0.001	0.003	1.63
JUNB	0.008	0.064	1.42
HIST2H2BB	0.012	0.076	0.73
EGR1	0.019	0.076	0.70
MLLT4	0.033	0.101	0.97

Abbreviations: FDR – false discovery rate, FC – fold change. *P*-values, FDR *P*-values and the log FC were calculated using the Bioconductor package HTqPCR, in *R*. Log FC was calculated using log2 calculations, in relation to achieving 24 month MMR.

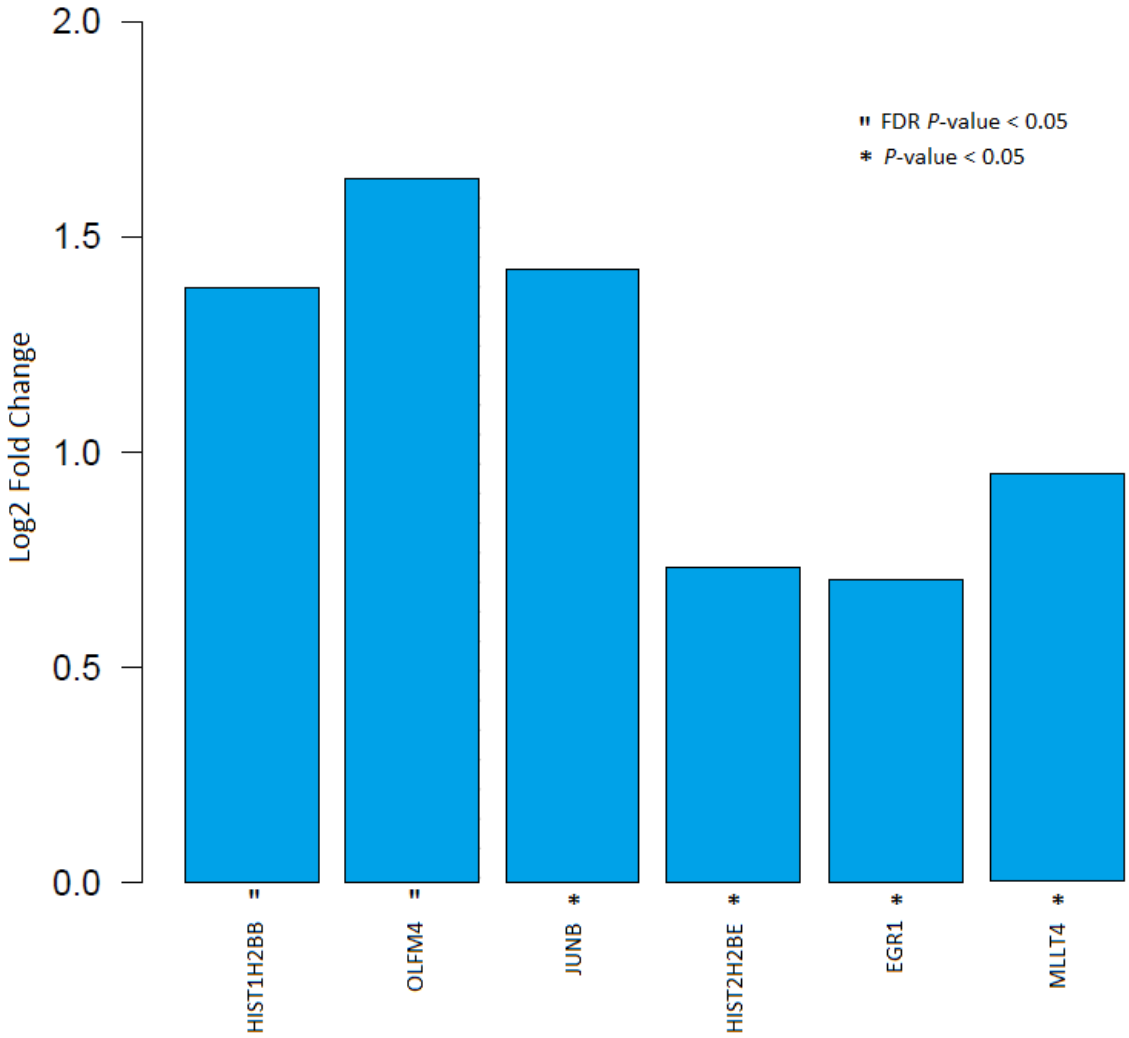


Figure 4.27 Significant gene expression between CP-CML patients who did, or did not, achieve MMR by 24 months using 24-gene TLDA

The top 6 differentially expressed genes associated with 24 month MMR in CP-CML patients, with a *P*-value < 0.05 were plotted. All genes demonstrated increased expression in CP-CML patients who did not achieve 24 month MMR. *P*-values were calculated using the Student's *t*-test in HTqPCR. All genes with a FDR *P*-value < 0.05 also had a *P*-value < 0.05.

4.2.10 No commonality was observed between the genes significantly associated with patient response from both the 96- and 24- gene RQ-PCR datasets

A comparison between the significant gene-lists (BH-FDR < 0.05) identified from both the 24- and 96-gene TLDA analysis was then performed. No comparison was able to be made for the OA or MMR by 12 month endpoints, as the 24-gene analysis did not identify any significant genes discriminating these two clinically relevant observations. Two genes (*SMOX* and *CLU*) were identified by the 96-gene analysis as significantly associated with EMR; however, neither of these genes was identified by the 24-gene analysis (only 1 significant gene; *HIST1H3H*). Furthermore, 31 genes were significantly associated with MMR by 24 months from the 96-gene analysis (**Table 4.8**); yet surprisingly, there was no overlap with the 2 significant genes (*HIST1H2BB* and *OLFM4*) from the 24-gene analysis. The lack of common genes between these two analyses was unexpected, but may relate to each patient cohort being selected based on OA criteria only. Hence, patient cohorts were not optimised for molecular response, leading to cohorts containing more responders than non-responders being analysed.

4.2.11 Combined 110 patient RQ-PCR analysis identifies genes associated with poor response to TKI therapy

In order to interrogate the data in its entirety, in an effort to identify genes associated with poor response to TKI therapy, both the 96- and 24-gene TLDA datasets were combined, and the resulting 17 common genes (110 samples, all OA quartiles) were analysed for their association with the various response endpoints, as previously analysed and OA. Four genes were differentially expressed within the OA group (very low [n = 54] versus Other OA [n = 56]; **Figure 4.28**), of which all 4 were histone genes. Two genes were differentially expressed between patients who did (n = 96), or did not (n = 14), achieve EMR by 3 months on TKI therapy (**Figure 4.29**), including 1 histone gene (*HIST1H3H*). The achievement of MMR by

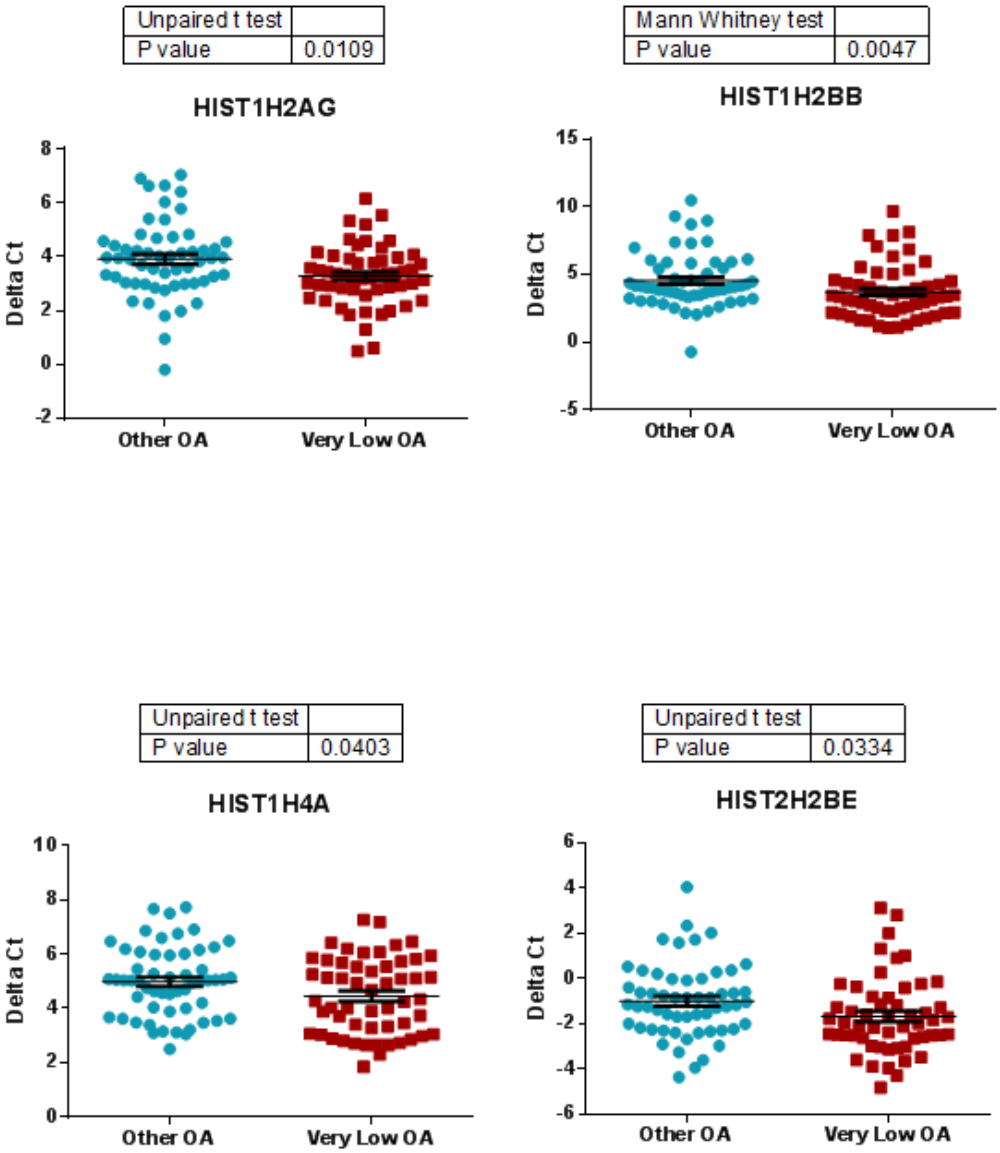


Figure 4.28 Four genes had significant differential expression between Very low and Other OA CP-CML patient samples

The difference between gene expression for very low (n = 54) and Other OA (n = 56) patients is plotted. Scatterplots demonstrate the mean Delta Ct ± standard error of the mean (SEM). Delta Ct values were generated against the reference gene, *GUSB*. A student's *t*-test or Mann-Whitney test were performed using GraphPad Prism 6 as appropriate (if the data was normally distributed, or not, respectively).

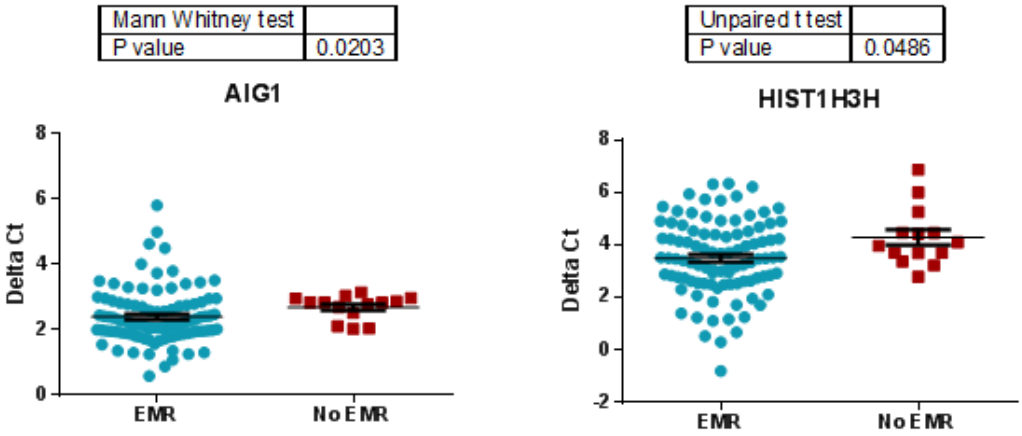


Figure 4.29 Two genes had significant differential expression associated with the achievement of EMR by 3 months on TKI therapy

The difference in gene expression for CP-CML patients who achieved EMR by 3 months (n = 96), versus those who did not (n = 14) is plotted. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. A student's *t*-test or Mann-Whitney test were performed using GraphPad Prism 6 as appropriate (if the data was normally distributed, or not, respectively).

12 months was investigated next, and 3 genes were identified with significant differential expression between patients who did (n = 69), or did not (n = 41), achieve MMR by 12 months on TKI therapy (**Figure 4.30**). One gene, *AIG1*, was significantly up-regulated in both the patients who achieved EMR by 3 months and the patients who achieved MMR by 12 months, compared to those who did not. Four genes were statistically significant when the analysis was extended to the achievement of MMR by 24 month time-point (**Figure 4.31**); however, only HIST1H4A (*P* = 0.042; Student's *t*-test) remained significant when the achievement of MMR at exactly 24 months was analysed.

Interestingly, only 1 gene (*GFI1B*) was consistent between both the MMR by 12 and 24 months gene-lists, indicating that this gene was highly associated with MMR, independent of response time-point. Two genes (*HIST1H2BB* and *HIST1H4A*) were also common between the 24 month MMR and OA gene-lists. Finally, transformation to BC while on TKI therapy was assessed and 2 genes were differentially expressed between those patients who did (n = 4), or did not (n = 106), transform to BC (**Figure 4.32**). One of these genes (*JUNB*) was common with the 24 month MMR gene-list. However, it is important to note that these results need to be validated in an independent patient cohort before any of these genes can be used as potential biomarkers for CML response. Unfortunately, due to limited primary sample availability and funding requirements, these analyses have yet to be performed, but are the focus of future studies.

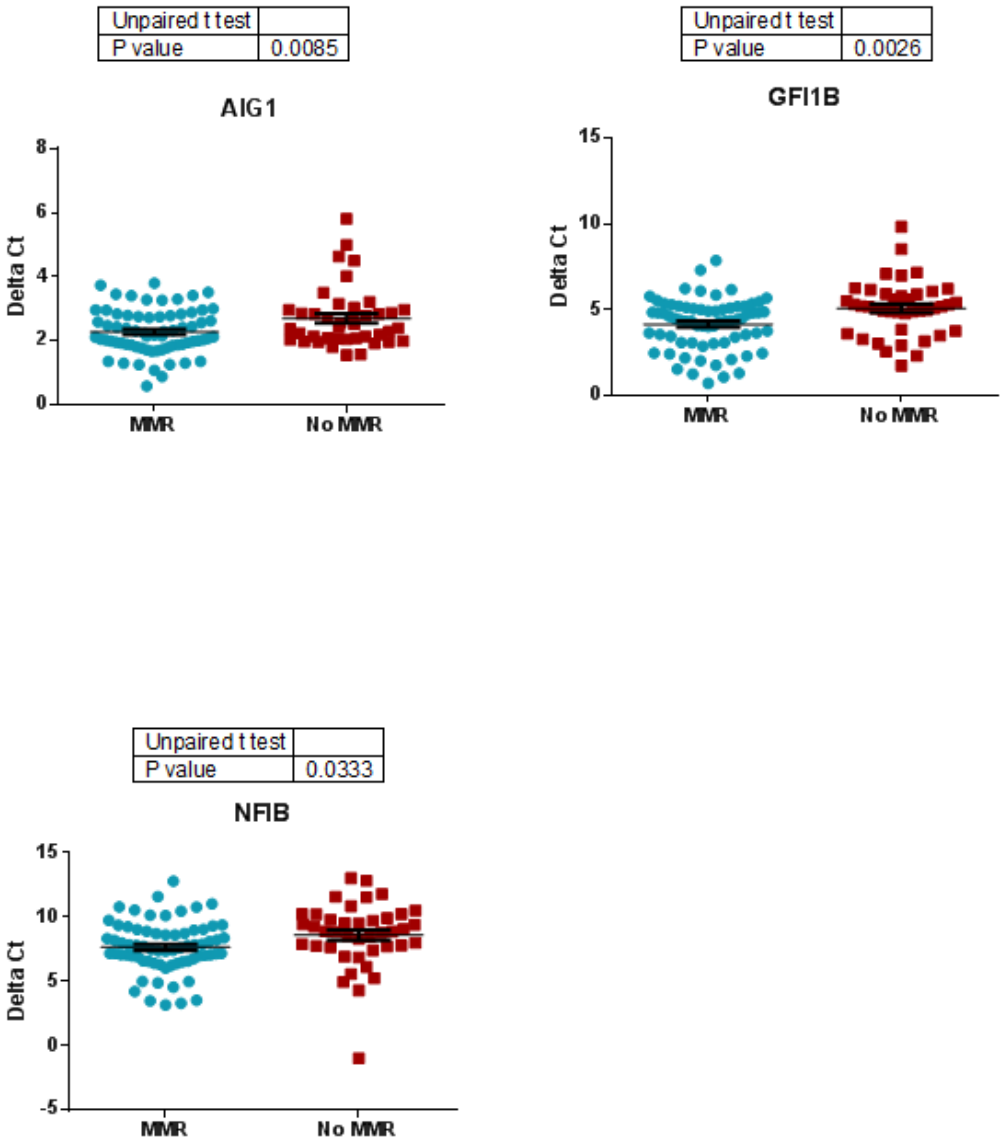


Figure 4.30 Three genes had significant differential expression associated with the achievement of MMR by 12 months on TKI therapy

The difference in gene expression for CP-CML patients who achieved MMR by 12 months (n = 69), versus those who did not (n = 41) is plotted. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. A student's *t*-test or Mann-Whitney test were performed using GraphPad Prism 6 as appropriate (if the data was normally distributed, or not, respectively).

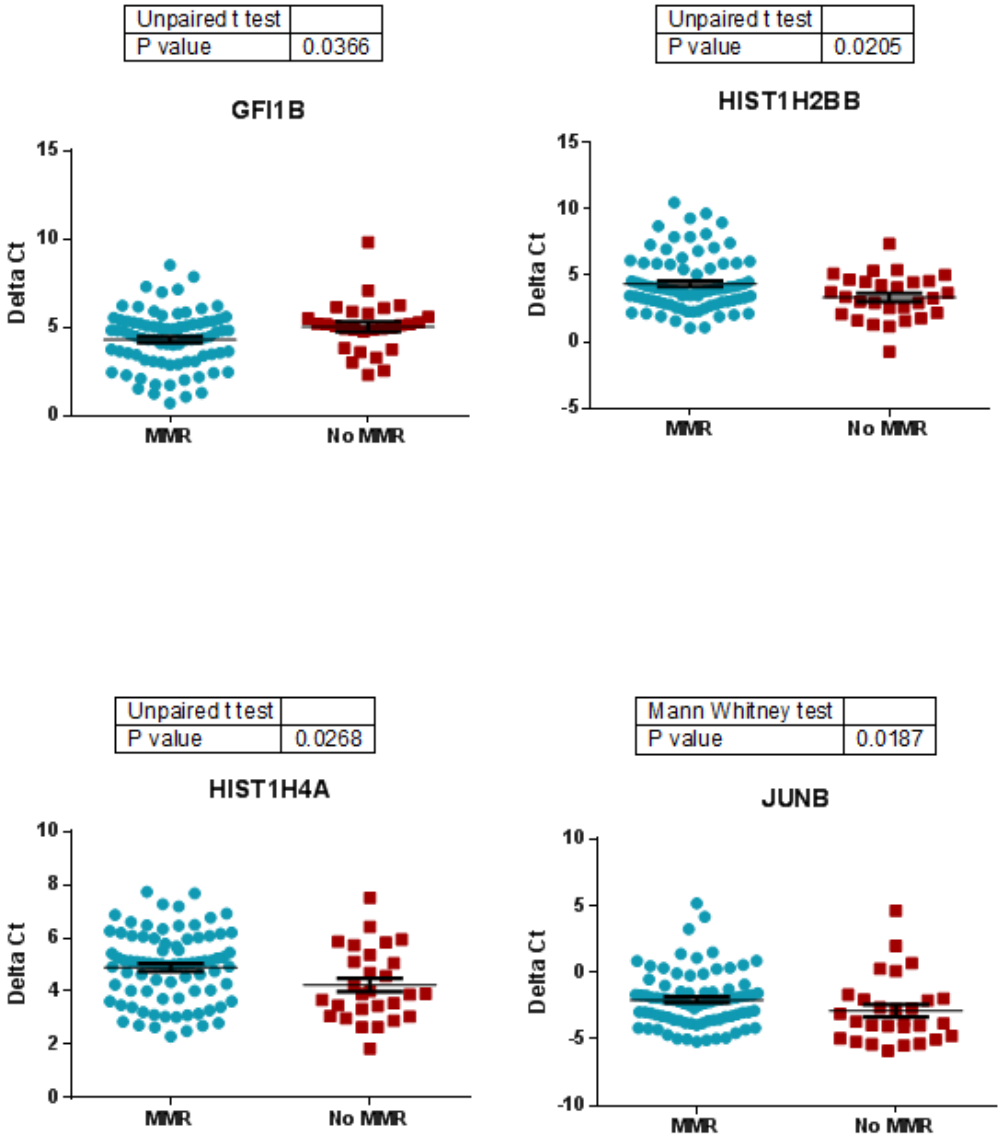


Figure 4.31 Four genes had significant differential expression associated with the achievement of MMR by 24 months on TKI therapy

The difference in gene expression for CP-CML patients who achieved MMR by 24 months (n = 82), versus those who did not (n = 28) is plotted. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. A student's *t*-test or Mann-Whitney test were performed using GraphPad Prism 6 as appropriate (if the data was normally distributed, or not, respectively).

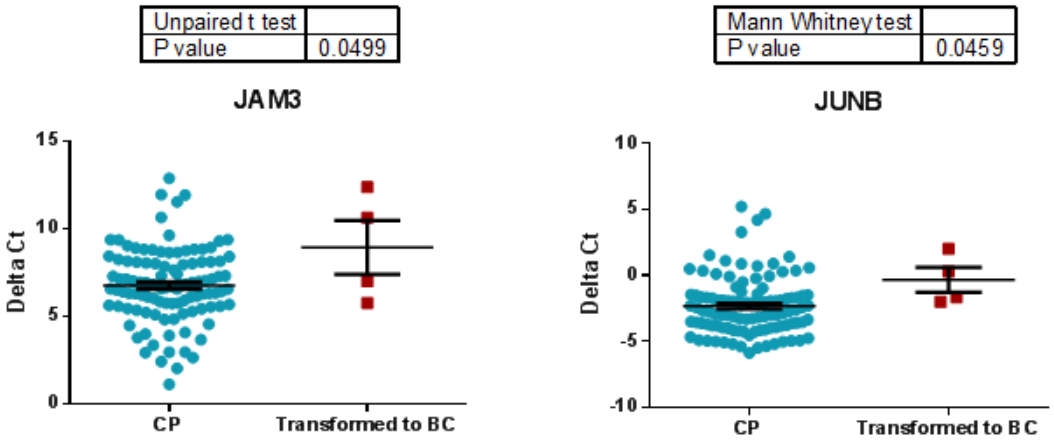


Figure 4.32 Two genes had significant differential expression associated with transformation to BC on TKI therapy

The difference in gene expression for CP-CML patients who transformed to BC while on TKI therapy (n = 4), versus those who did not (CP; n = 106) is plotted. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. A student's *t*-test or Mann-Whitney test were performed using GraphPad Prism 6 as appropriate (if the data was normally distributed, or not, respectively).

4.2.12 Histone genes are associated with very low OA, and with CML poor responders from microarray data mining

First emerging from the very low versus very high OA microarray analysis, and then consistently evident in both the 96- and 24-gene TLDA RQ-PCR analyses, was a histone gene signature, significantly up-regulated in the very low OA patients. Differential expression of some of these histone genes was also significantly associated with various response time-points, including EMR (3 months) and MMR (at both 12 and 24 months). Furthermore, the original 10-gene OA classifier developed using iterativeBMA, contained significant enrichment of histone genes (7/10 genes were histones). To determine if this histone signature was identified by pure chance, or if in fact, histone gene enrichment had been previously observed in CML, a meta-analysis of previously published CML microarray studies was performed.

The CML response datasets of McWeeney *et al.*[179] (signature associated with major cytogenetic response in imatinib-treated CP-CML patients), Yong *et al.*[12] (signature associated with survival in CML patients), Zheng *et al.*[181] (signature associated with BC-CML), Radich *et al.*[180] (signature associated with progression and response in CML), Crossman *et al.*[290] (signature associated with cytogenetic response to imatinib in CML) and Tipping *et al.*[314] (signature associated with innate imatinib resistance in CML), were all investigated for significant histone gene expression between the various responder/non-responder patient groups. Significant differential histone gene expression was detected in the McWeeney *et al*., Yong *et al.*, Radich *et al.* and Crossman *et al.* datasets. Twenty-five histone genes were identified across all 7 datasets (including the very low versus very high OA dataset; **Table 4.19**), including 3 histones (*HIST1H2BG*, *HIST1H3H* and *HIST2H2BE*) that were common

Table 4.19 Summary of significantly expressed histone genes in various CML microarray datasets

Accession No.	Gene Symbol	Chr	logFC	Direction	Very low vs. Very high OA	McWeeney	Yong	Zheng	Radich	Crossman	Tipping
NM_021064	***HIST1H2AG***	6p22.1	-0.93	Up in low OA/NR	+	+	-	-	-	-	-
NM_003518	***HIST1H2BG***	6p21.3	-1.13	Up in low OA/NR	+	+	+	-	-	-	-
NM_003536	***HIST1H3H***	6p22-p21.3	-1.09	Up in low OA/NR	+	+	-	-	-	+	-
NM_003528	***HIST2H2BE***	1q21-q23	-1.53	Up in low OA/NR	+	+	+	-	-	-	-
NM_021062	***HIST1H2BB***	6p21.3	-0.94	Up in low OA	+	-	-	-	-	-	-
NM_021058	***HIST1H2BJ***	6p22.1	-1.3	Up in low OA	+	-	-	-	-	-	-
NM_003527	***HIST1H2BO***	6p22.1	-0.61	Up in low OA	+	-	-	-	-	-	-
NM_003538	***HIST1H4A***	6p21.3	-0.82	Up in low OA	+	-	-	-	-	-	-
NM_001024599	***HIST2H2BF***	1q21.2	-1.74	Up in low OA	+	-	-	-	-	-	-
NM_001025303	***HIST2H3PS2***	1q21.2	-0.8	Up in low OA	+	-	-	-	-	-	-
NM_005319	***HIST1H1C***	6p21.3	1.64	Up NR	-	+	-	-	-	-	-
NM_003512	***HIST1H2AC***	6p21.3	2.27	Up NR	-	+	-	-	-	-	-
NM_021052	***HIST1H2AE***	6p22.1	1.87	Up NR	-	+	-	-	-	-	-
NM_003510	***HIST1H2AK***	6p22.1	1.59	Up NR	-	+	-	-	-	-	-
NM_003526	***HIST1H2BC***	6p21.3	2.71	Up NR	-	+	-	-	-	-	-

NM_021063	***HIST1H2BD***	6p21.3	1.97	Up NR	-	+	-	-	-	-	-
NM_003522	***HIST1H2BF***	6p21.3	1.53	Up NR	-	+	-	-	-	-	-
NM_080593	***HIST1H2BK***	6p21.33	1.5	Up NR	-	+	-	-	-	-	-
NM_003530	***HIST1H3D***	6p21.3	2.51	Up NR	-	+	-	-	-	-	-
NM_003516	***HIST2H2AA3/4***	1q21.2	1.78	Up NR	-	+	-	-	-	-	-
NM_003532	***HIST1H3E***	6p21.3	-	Difference in phase	-	-	-	-	+	-	-
NM_021018	***HIST1H3F***	6p22.2	-	Difference in phase	-	-	-	-	+	-	-
NM_003534	***HIST1H3G***	6p21.3	-	Difference in phase	-	-	-	-	+	-	-
NM_003535	***HIST1H3J***	6p22.1	-	Difference in phase	-	-	-	-	+	-	-
NM_003493	***HIST3H3***	1q42	-	Difference in phase	-	-	-	-	+	-	-

Histone genes that were identified in a given gene expression dataset are identified by the "+" sign and gold highlighting. Abbreviations: chr – chromosome, logFC – log2 fold change, NR – non-responder.

between three different datasets and a further one histone (*HIST1H2AG*) that was common between two different datasets. A prevalence of histones belonging to the H2A, H2B and H3 histone families was apparent, with all histone genes consistently up-regulated in the poor response patient groups (i.e. very low OA, imatinib resistant, progression to BC or no cytogenetic response).

4.2.13 Transcription factor binding analysis identifies MYC and E2F1 as significantly enriched with the histone gene signature

To determine how these histone genes may be regulated and how they may be impacting CML patient response, transcription factors involved in their regulation were identified by analysing global ChIP and transcription factor databases using Pscan,[229] TFactS[228] and ChEA.[230] Pscan analysis was performed, with all 25 histone genes analysed using the JASPAR profile and a 1000 bp length promoter region (-950 to +50). Forty-one different transcription factors were significantly enriched for the histone gene-set (**Table 4.20**). TFactS analysis was then performed using default settings, with 6 transcription factors identified as significantly enriched for the histone gene-set (**Table 4.21**). Finally, ChEA was performed using default settings, and 45 transcription factors were found to be significantly enriched for the histone gene-set (**Table 4.22**). However, as some transcription factors were identified multiple times, each time from a different ChIP dataset, 34 unique transcription factors were identified by ChEA.

Overall, 6 transcription factors were common between at least two of the three different database tools, with *E2F1* the only transcription factor common to all 3 methods (**Table 4.23**). Four of the 6 transcriptions factors (*E2F1*, *MYC*, *IRF1* and *POU5F1*) have a published association with CML; however, *MYC* (c-Myc) and *E2F1* have the most well characterised role

Table 4.20 List of transcription factors significantly enriched for the 25 histone gene-set using Pscan

TF	*Z*-score	*P*-value	TF	*Z*-score	*P*-value
TBP	7.45	4.72E-14	***MIZF***	2.42	0.01
NFYA	7.11	5.62E-13	***Esrrb***	2.39	0.01
E2F1	6.01	9.00E-10	***YY1***	2.25	0.01
MEF2A	4.72	1.17E-06	***FOXD1***	2.22	0.01
Nobox	4.48	3.63E-06	***NFE2L1::MafG***	2.20	0.01
PBX1	3.61	1.52E-04	***Nr2e3***	2.20	0.01
Nkx2-5	3.28	5.21E-04	***REST***	2.13	0.02
FOXL1	3.17	7.55E-04	***FOXI1***	2.12	0.02
Prrx2	3.01	1.29E-03	***GATA3***	2.11	0.02
En1	2.91	1.77E-03	***IRF1***	2.09	0.02
Foxa2	2.88	1.99E-03	***NR1H2::RXRA***	2.08	0.02
ARID3A	2.84	2.23E-03	***IRF2***	2.02	0.02
NKX3-1	2.79	2.62E-03	***Foxd3***	1.93	0.03
NFIL3	2.68	3.70E-03	***HNF1A***	1.91	0.03
Lhx3	2.62	4.39E-03	***FOXO3***	1.88	0.03
FOXA1	2.61	4.51E-03	***Pou5f1***	1.85	0.03
SRF	2.52	0.01	***FOXF2***	1.72	0.04
Gfi	2.49	0.01	***Evi1***	1.70	0.04
CEBPA	2.48	0.01	***NR3C1***	1.66	0.05
Pdx1	2.44	0.01	***Hltf***	1.64	0.05
HOXA5	2.43	0.01			

Abbreviations: TF – transcription factor. *P*-values and *Z*-scores were calculated in Pscan.

Table 4.21 List of transcription factors significantly enriched for the 25 histone gene-set using TFactS

TF	*P*-value	FDR *P*-value
GATA1	0.004	0.008
E2F4	0.010	0.017
E2F1	0.012	0.025
TBP	0.025	0.033
POU2F1	0.033	0.042
MYC	0.242	0.050

Abbreviations: TF – transcription factor, FDR – Benjamini-Hocherg adjusted false discovery rate. *P*-values and FDR *P*-values were calculated using TFactS.

Table 4.22 List of transcription factors significantly enriched for the 25 histone gene-set using ChEA

TF	Dataset	*P*-value	TF	Dataset	*P*-value
RNF2	16625203	8.75E-19	***CHD1***	19587682	1.53E-04
SOX2	16153702	1.29E-13	***SOX17***	20123909	1.03E-03
SOX2	18692474	2.22E-13	***TCF3***	18692474	1.30E-03
CNOT3	19339689	1.42E-12	***EED***	16625203	0.002
TAL1	20566737	1.55E-11	***POU5F1***	18692474	0.003
E2F4	17652178	2.23E-10	***SOX2***	20726797	0.004
HOXB4	20404135	4.20E-09	***POU5F1***	18700969	0.005
ZFP42	18358816	9.17E-09	***IRF1***	19129219	0.007
KLF4	19030024	1.05E-08	***FLI1***	20887958	0.007
NANOG	16153702	3.13E-08	***MYC***	19915707	0.009
STAT4	19710469	1.15E-07	***CCND1***	20090754	0.009
SMAD4	19686287	1.64E-07	***TCF3***	18347094	0.011
KLF4	18358816	4.81E-07	***E2F1***	18555785	0.011
ERG	20887958	1.63E-06	***THAP11***	20581084	0.016
SMAD1	18555785	1.73E-06	***ELK1***	19687146	0.019
POU5F1	16153702	1.93E-06	***POU3F2***	20337985	0.020
DACH1	20351289	5.78E-06	***LMO2***	20887958	0.021
MYC	19030024	8.06E-06	***ZIC3***	20872845	0.022
SOX2	19030024	1.24E-05	***ELF1***	20517297	0.025
CUX1	19635798	5.56E-05	***CLOCK***	20551151	0.026
E2F1	21310950	5.97E-05	***SOX2***	18555785	0.038
SMAD2	18955504	1.34E-04	***POU5F1***	18347094	0.040
SMAD3	18955504	1.34E-04			

Abbreviations: TF – transcription factor. *P*-values were calculated in ChEA.

Table 4.23 Common transcription factors significantly enriched for the 25 histone gene-set using Pscan, TFactS and ChEA

TF	Pscan	TFactS	ChEA	Process
E2F1	+	+	+	Cell cycle regulation, DNA replication, apoptosis
E2F4		+	+	Cell cycle regulation, DNA replication
MYC		+	+	Transcription
TBP	+	+		Transcription
IRF1	+		+	Transcription, Tumour suppressor
POU5F1	+		+	Transcription

Abbreviations: TF – transcription factor.

in CML pathogenesis. *MYC* is critical for *BCR-ABL1* transformation;[315,316] while *BCR-ABL1* has also been demonstrated to induce *MYC* transcription through *E2F1* in *BCR-ABL1* transformed myeloid cells.[317] *MYC* expression was demonstrated to be decreased by imatinib therapy in K562 cells[318] and MYC protein, but not mRNA levels, are increased in CD34+ bone marrow cells from advanced phase (AP and BC) CML patients, compared to normal individuals.[319] CML peripheral blood MNCs also display significantly higher *E2F1* mRNA levels, compared to healthy controls.[320]

4.2.14 Ingenuity Pathway Analysis identifies interactions between the transcription factors E2F1, E2F4 and MYC and the histone gene-set

To further understand the pathways involved with, and the interactions between, these transcription factors and the histone gene-set, Ingenuity Pathway Analysis (IPA) was subsequently performed. IPA identified three transcription factors, *E2F1*, *MYC* and *E2F4*; that appeared to play a major role in the regulation of the histone genes and other histone-associated genes (**Figure 4.33**). Investigation of the Molecular Signatures Database (MSigDB; Broad Institute) further confirmed the enrichment of the *E2F1*, *E2F4* and *MYC* transcription factors for the histone gene-set (**Table 4.24**). Taken together, these data were suggestive of a possible role in histone regulation by either *MYC* or the E2F family of transcription factors (*E2F1* or *E2F4*); yet no significant change in the mRNA expression levels of these transcription factors between very low and very high OA patients was observed in the original OA microarray. Thus, protein expression levels of *MYC* and *E2F1* were determined to understand whether they were functionally active in CML cell lines and patient samples from the very low and very high OA patient groups.

Figure 4.33 Interactions between histone genes and transcription factors by Ingenuity Pathway Analysis

Ingenuity Pathway Analysis (IPA) was performed using the 25 histone gene-set and the 6 transcription factors identified from published datasets to interact with these histone genes. IPA identified *E2F1*, *MYC* and *E2F4* as having the greatest interaction with the histone gene-set. The six transcription factors are coloured **PINK**. The histone genes are coloured **RED**, if they were identified from data mining of published microarray datasets, or **GREEN** if they were significant in the very low versus very high OA microarray, with the shading representing the fold change (FC) value associated with each histone, increasing from dull (small FC) to bright (large FC). **GREEN** arrows indicate *E2F1* pathway interactions. **BLUE** arrows indicate *E2F4* pathway interactions. **YELLOW** arrows indicate *MYC* pathway interactions. **GREY** arrows represent interactions not involving *MYC*, *E2F1* or *E2F4*. Finally, **GREY** colouring indicates additional genes/transcription factors that interact with either the histone gene-set, or the 6 transcription factors initially identified by IPA.

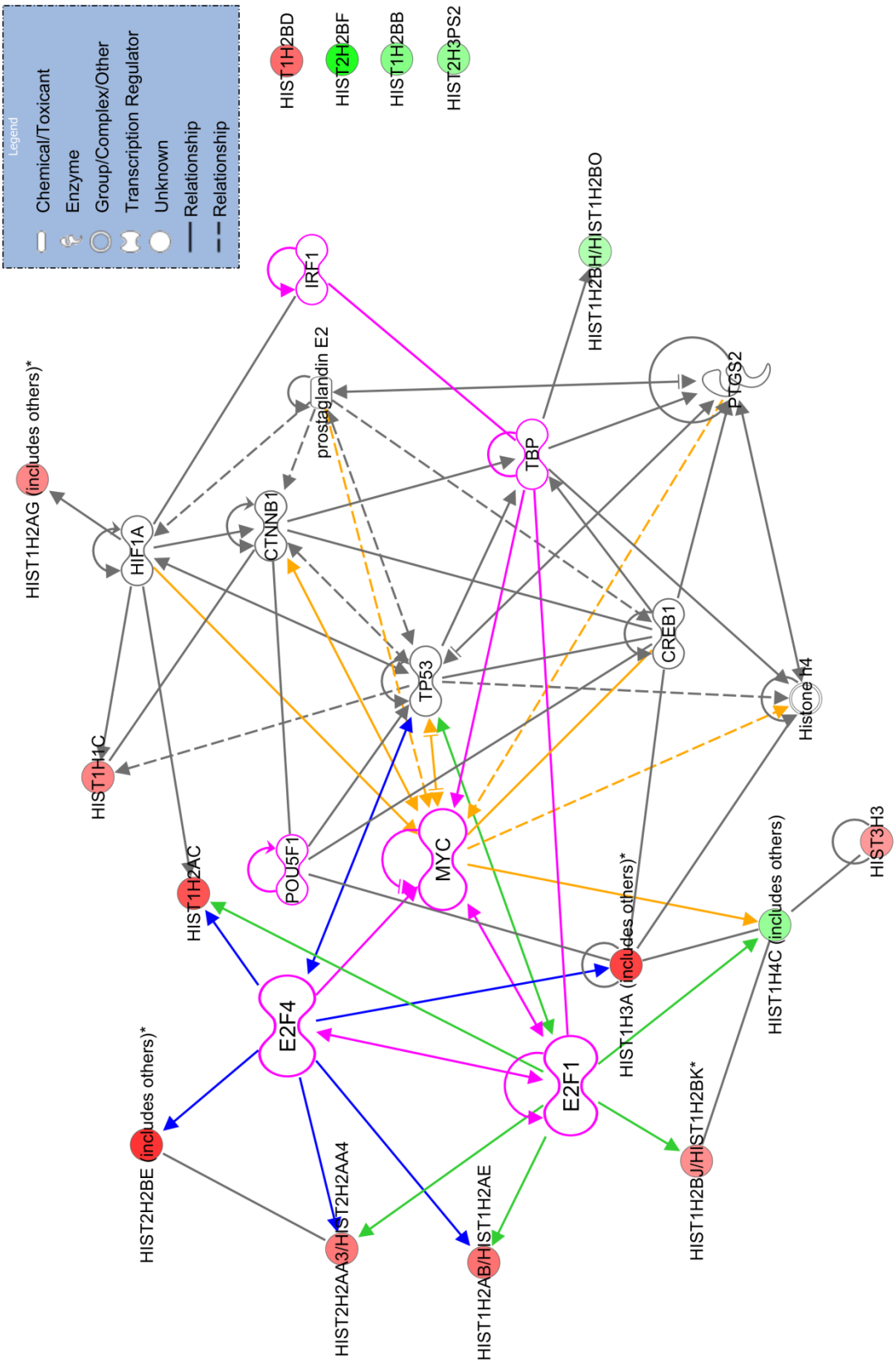

Legend
Chemical/Toxicant
Enzyme
Group/Complex/Other
Transcription Regulator
Unknown
Relationship
Relationship
HIST1H2BD
HIST2H2BF
HIST1H2BB
HIST2H3PS2
IRF1
HIST1H2BH/HIST1H2BO
HIST1H2AG (includes others)*
prostaglandin E2
PTGS2
TBP
HIF1A
CTNNB1
CREB1
TP53
Histone h4
HIST1H1C
HIST1H2AC
POU5F1
MYC
HIST3H3
HIST1H4C (includes others)
HIST1H3A (includes others)*
E2F4
E2F1
HIST2H2BE (includes others)*
HIST2H2AA3/HIST2H2AA4
HIST1H2AB/HIST1H2AE
HIST1H2BJ/HIST1H2BK*

Table 4.24 Common functional gene-sets enriched for the histone gene-set by MSigDB analysis

	Gene-set Size*	FDR *q*-value	*P*-value
Transcription Factors			
***MARSON* – Bound by E2F4**	16	< 0.001	< 0.001
***DANG* – Bound by MYC**	10	< 0.001	< 0.001
***BENPORATH* – MYC/MAX targets**	9	< 0.001	< 0.001
***REN* – Bound by E2F**	4	< 0.001	< 0.001

Abbreviations: FDR – false discovery rate. *Gene-set size refers to the number of genes overlapping between the gene-sets.

4.2.15 Western blot analysis was unable to detect the protein levels of MYC and E2F1 in CP-CML patient samples

Both total and phosphorylated MYC, at serine 62 (S62; phosphorylated by *ERK* and indicative of protein stability) and S62 plus threonine 58 (T58; phosphorylated by *GSK-3β* and indicative of protein degradation); and total E2F1 protein levels were assessed to elucidate the role of these transcription factors in CML. Initial experiments demonstrated that total E2F1 and MYC (total and p-MYC) protein levels were detectable in CML cell lines (KU812 and K562) and the HL60 cell line, which was derived from an acute myeloblastic leukaemia with maturation (FAB-M2;[321] **Figure 4.34**). However, total protein levels of E2F1, MYC and phosphorylated MYC (S62 or T58 + S62), were unable to be detected in CML patient cells with very low or very high OA (**Figure 4.35**).

From the results at hand, the negligible protein levels detected in CML patient samples for these transcription factors, combined with the lack of significant gene expression variance, indicates that the *MYC* and *E2F1* transcription factors are unlikely to be involved in the up-regulation of histone gene expression observed in very low OA CML patients. However, with further western blot optimisation, the protein detection may be improved, allowing both E2F1 and MYC to be detected in CML patient samples, which is now the focus of ongoing studies. Alternatively, the employment of other assays for detecting MYC and E2F1 activity, such as an ELISA assay (enzyme-linked immunosorbent assay), are being investigated.

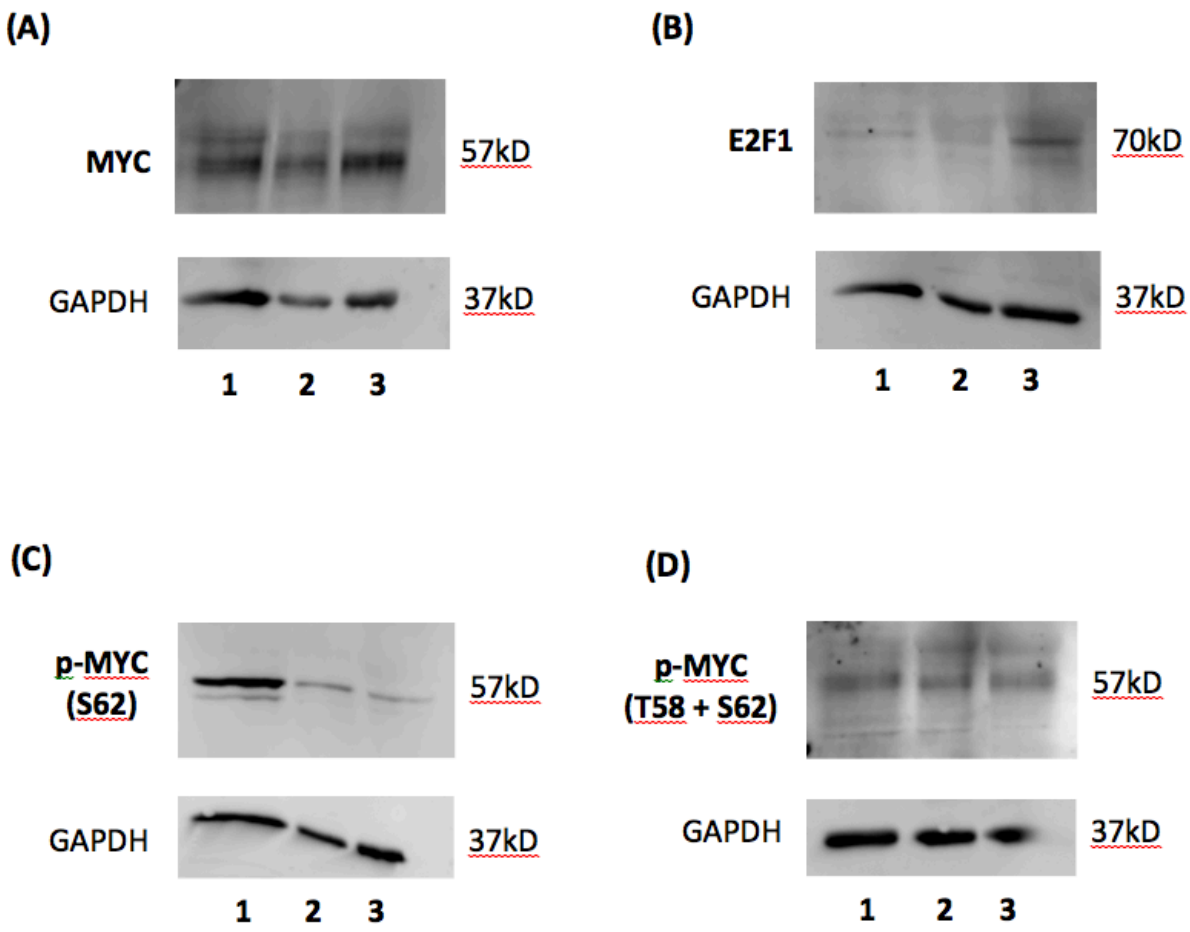
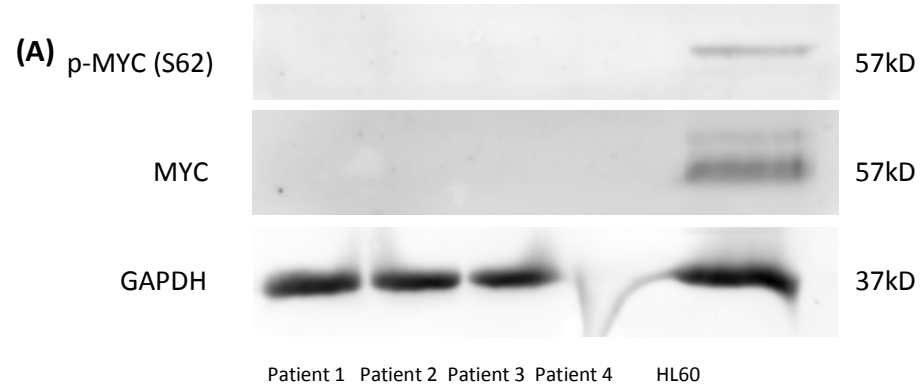


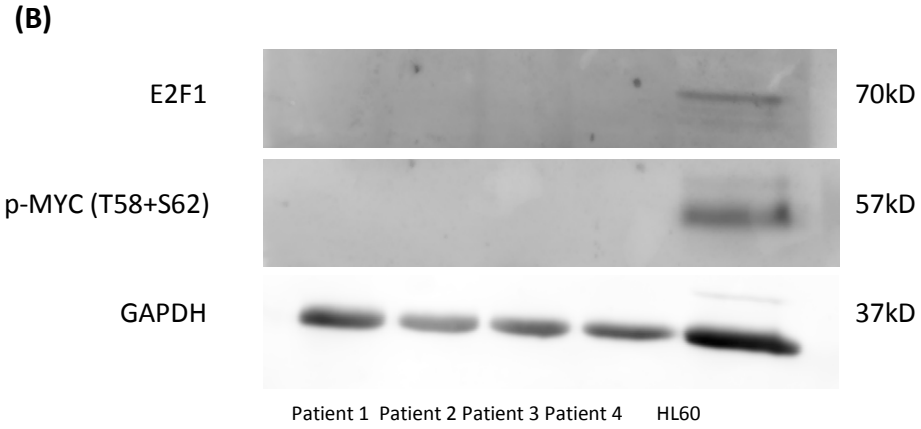
Figure 4.34 CML cell lines exhibit MYC and E2F1 protein expression

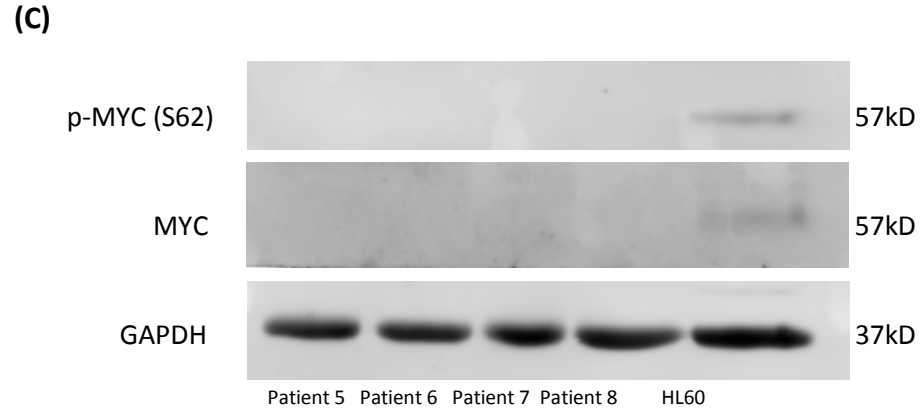
Protein expression levels of **(A)** MYC; **(B)** E2F1; **(C)** phosphorylated MYC (S62); and **(D)** phosphorylated MYC (T58 + S62) were assessed in KU812 (lane 1), K562 (lane 2) and HL60 (lane 3) cells. Loading control protein (GAPDH) was detected for all blots. Data demonstrate that all three cell lines express MYC, phospho-MYC and E2F1; therefore, lysates from these cell lines can be used as positive controls for further investigation into the detection of MYC and E2F in primary patient cells. Western blot analyses are representative of at least two separate experiments.

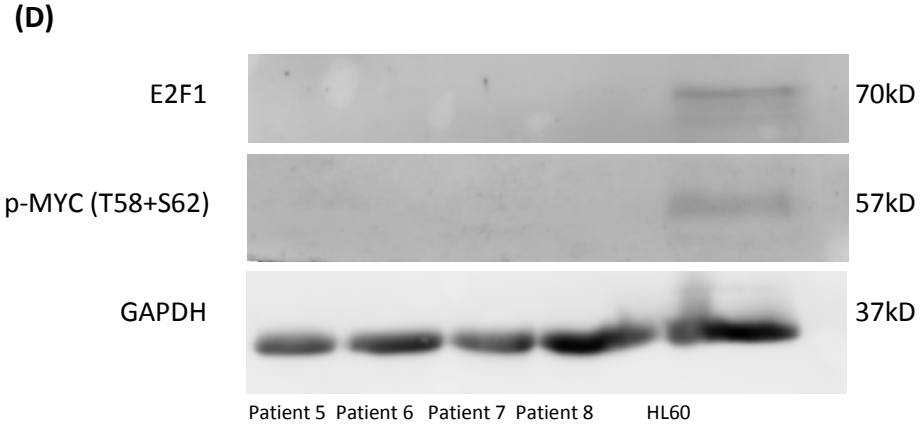
Figure 4.35 Very low and very high OA patient cells do not exhibit detectable MYC or E2F1 protein levels

Protein expression levels of p-MYC (S62) and MYC **(A & C)**; and E2F1 and p-MYC (T58 + S62) **(B & D)** were assessed in four CML patients with very low OA **(A & B)**, four CML patients with very high OA **(C & D)** and HL60 control cells. Data demonstrates no detectable protein expression in either the very low or very high OA CML patients; however, protein was detected in the control cells and loading control protein (GAPDH) was detected in all lysates. In western blots **(A & C)**, samples were initially probed for p-MYC (S62), before the blot was stripped, and re-probed for MYC. In western blots **(B & D)**, samples were initially probed for E2F1, before the blot was stripped and re-probed for p-MYC (T58 + S62); however, the same lysates was used. Western blot analyses are representative of two separate experiments, due to limited sample availability. Fifty micrograms (50 µg) total protein was loaded in each lysate.


(A)
p-MYC (S62)
57kD
MYC
57kD
GAPDH
37kD
Patient 1 Patient 2 Patient 3 Patient 4 HL60


(B)
E2F1
70kD
p-MYC (T58+S62)
57kD
GAPDH
37kD
Patient 1 Patient 2 Patient 3 Patient 4 HL60


(C)
p-MYC (S62)
57kD
MYC
57kD
GAPDH
37kD
Patient 5 Patient 6 Patient 7 Patient 8 HL60


(D)
E2F1
70kD
p-MYC (T58+S62)
57kD
GAPDH
37kD
Patient 5 Patient 6 Patient 7 Patient 8 HL60

4.3 Discussion

As previously discussed, the OCT-1 activity (OA) assay is an excellent prognostic indicator of molecular response at diagnosis,[142,151] and identifies patients with very low OA (OA < 4 ng/200,000 cells) who are at significant risk for poor response (lower MMR, event-free survival, transformation-free survival and overall survival), compared to all other patients.[142] However, the difficult applicability of the OA assay has prevented it from being adopted world-wide for routine testing. Yet, the variation between OA at diagnosis and subsequent disease outcome in patients identifies the importance of further investigating the factors underlying OA in CP-CML patients. The findings described in this chapter suggest that significant lineage gene expression differences are present between patients with very low and very high OA, with a strong monocytic and erythroid signature being associated with very low OA; however, a gene expression signature was unable to successfully distinguish low and high OA patient groups. This data supports previous work from Engler *et al.*[243] who demonstrated that OA, as measured in CML patient MNCs, was strongly related to cell lineage, with the highest OA measured from neutrophils, followed by monocytes; and the lowest measured in lymphocytes.

The RQ-PCR gene expression data, generated in order to validate the initial microarray, described in this chapter was unfortunately unable to successfully develop, and validate a predictive classifier for OA. This was a somewhat unexpected result, but not one without possible explanation. Firstly, the initial OA microarray gene expression profiling was performed using only a very small patient cohort (n = 14), representing the two extreme OA groups, very low and very high OA. Hence, if this small cohort did not provide an accurate representation of these groups and the whole CP-CML patient cohort, which is possible, failure to validate the results in downstream analyses could be expected. Furthermore, upon

subsequent analysis, one of the 14 patients originally analysed in this cohort was incorrect classified to the very low OA group, instead of the high OA group. Secondly, the development of the initial classifier was based on RQ-PCR data generated using only patients from the two extreme OA groups (i.e. very low versus very high OA); hence, the genes selected by the classification algorithm were specific for these patient groups. Yet, when the classifier was validated, the cohort of patients used included all OA groups, as this resembled a more realistic clinical situation. This alone would have added significant "noise" to the classification algorithm, thus preventing accurate classification of these patients.

Thirdly, it is becoming increasing clear that CP-CML is not a homogeneous disease, but rather, significant heterogeneity exists between patients diagnosed in CP-CML. This is not only highlighted by the OA assay itself, which already suggests variation or disease within the CP-CML cohort, coinciding with their predicted response to imatinib therapy; but also historically, by the fact that no published predictive gene signature or classifier developed to predict response in CML has been adopted as a diagnostic tool. Furthermore, the OA assay is a functional assay, measuring imatinib uptake in CML patient MNCs at diagnosis, yet no difference in OCT-1 (SLC22A1) mRNA was observed between the very low OA and other OA patient groups from the RQ-PCR analysis. Additionally, the changes observed in the RQ-PCR analysis may be mediated by post-transcriptional changes, specific to the patient and independent of OA. Finally, Hu *et al.*[168] recently suggested that OCT-1 itself might not be causative for the observed association between OA and outcome, but rather, that the OA assay is providing a complex surrogate readout for expression of various transporters relevant to the IUR of imatinib. They demonstrated that imatinib was in fact a weak substrate for OCT-1, with the transporter itself unlikely to contribute substantially to the

observed association between OA and response. If this is the case, it is highly unlikely to be determined by mRNA analysis based on OA patient groups.

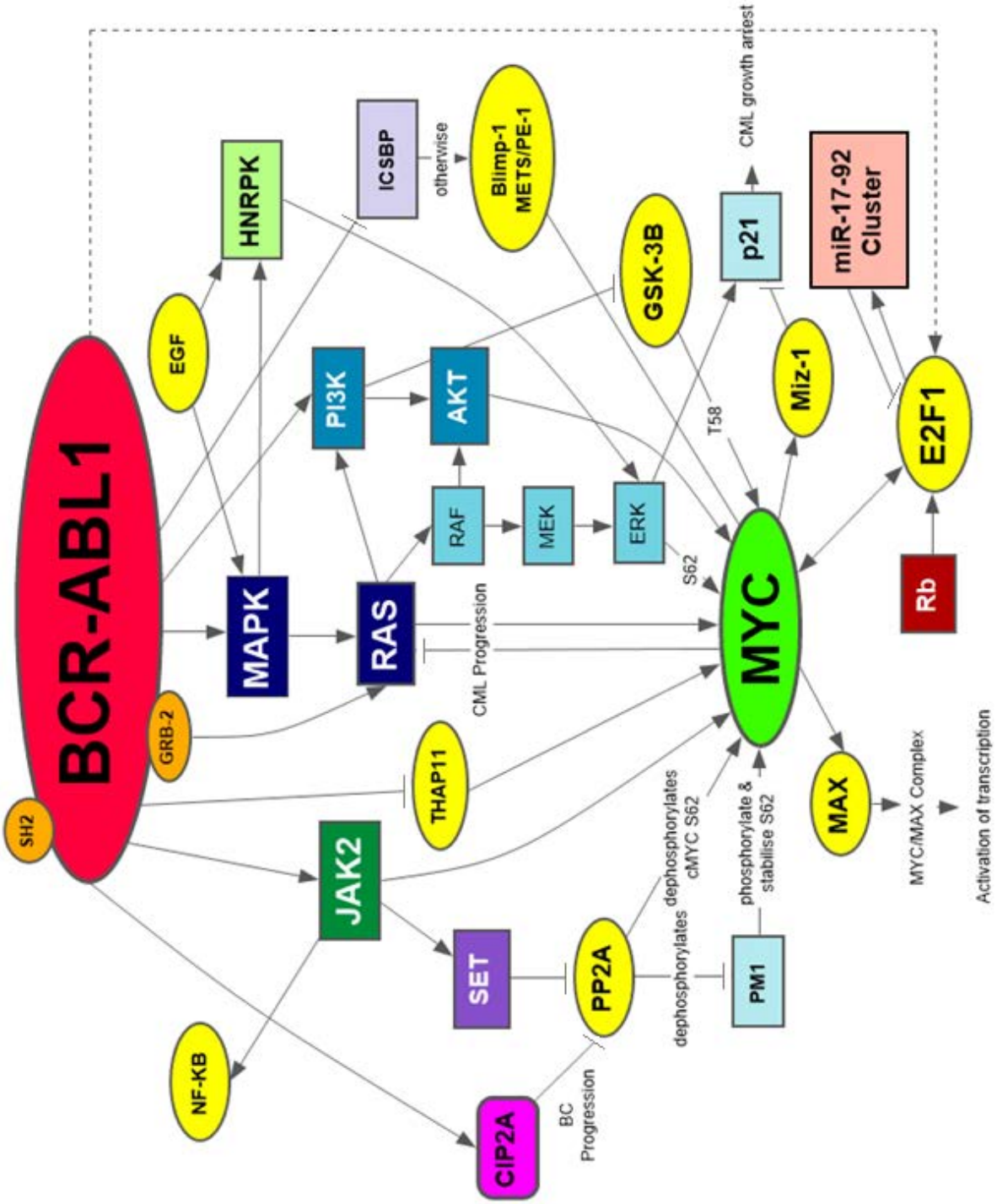
The data generated from the RQ-PCR studies presented here did however, generate some surprising results. Namely, a significant up-regulation of histone genes associated with very low OA patient cells was identified. Four histone genes (*HIST1H2AG*, *HIST1H2BB*, *HIST1H4A* and *HIST2H2BE*) were found to be differentially expressed between very low OA patients, compared to the rest, when 110 CP-CML patients were analysed. Further investigation revealed significant differential histone gene expression present in CML microarray datasets investigating response (McWeeney *et al.*,[179] Yong *et al.*[12] and Crossman *et al.*[290]) or disease progression (Radich *et al.*[180]), with up-regulation of histone genes associated with the poor-responder patient groups. Histone proteins are essential for the packaging of DNA into chromosomes, with approximately 150 bp of DNA being associated with a protein octamer containing two molecules of the core histone proteins H2A, H2B, H3 and H4; known as the nucleosome core particle.[322] The linker histone (H1) protein associates with nucleosomes to allow the formation of higher-order structures, which condense to become chromosomes.[322] HIST1 genes are found in a large cluster on chromosome 6 (6p21-p22; 55 genes), while two smaller clusters are located on chromosome 1, containing HIST2 (1q21; 6 genes) and HIST3 (1q42; 3 genes).[323] Importantly, all CML patients analysed were found to be cytogenetically normal, apart from the expected t(9;22) *BCR-ABL1* translocation, indicating that no copy number aberrations or mutations were present in these histone (HIST1, HIST2 or HIST3) clusters. Histone gene expression is cell-cycle-regulated, and is controlled at both the transcriptional and post-transcriptional levels by nuclear RNA processing, producing mRNAs lacking poly(A) tail, translation and mRNA stability control.[324] As such, histone modification levels have been demonstrated to be predictive for gene expression.[325]

In order to determine if transcriptional regulation was influencing the differential expression of histones in very low OA patients, transcription factor binding analysis was performed. The data generated identified that *MYC* and E2F family (specifically, *E2F1* and *E2F4*) transcription factors were significantly enriched for the histone gene-set. Both *MYC* and *E2F1* have established pathway interactions with *BCR-ABL1* in CML (**Figure 4.36**). The involvement of *MYC* in CML has not been fully addressed; however, *MYC* expression is up-regulated by *BCR-ABL1*[318] and *MYC* is critical for *BCR-ABL1* transformation.[315,316,326] The role of *MYC* in genomic instability and differentiation arrest has also led to it being associated with CML progression.[327] Recently, *MYC* has been demonstrated to promote aberrant DNA replication in the presence of imatinib, and a positive correlation was observed between *MYC* expression at diagnosis and poor response to imatinib therapy, unrelated to *BCR-ABL1* expression.[328] Unfortunately, the data presented in this chapter was unable to demonstrate a predictive association between *MYC* or *E2F1* and very low OA patients, at both the gene and protein level, inferring that these transcription factors are unlikely to be responsible for the up-regulation of histone genes associated with very low OA. Yet, as mentioned previously, histone genes can also be regulated at a post-transcriptional level, where DNA methylation provides an attractive target for modification of this gene expression, and is discussed further in **Chapter 5**.

In conclusion, data presented here demonstrates that the specific lineage characteristics of very low OA patients differ from very high OA patients, further confirming the role of cell lineage in OA determination. The findings presented in this chapter also identify genes predictive for response to TKI therapy, some of which will be discussed further in **Chapter 6**; and identify a significantly up-regulated histone gene signature associated with very low OA patients. However, the preliminary investigations presented here demonstrated that the

Figure 4.36 Schematic summary of signalling pathway interactions between *MYC* and *BCR-ABL1*

The transcription factor *MYC* has been implicated in CML disease progression, via multiple signalling pathways including *RAS* and *CIP2A*. *MYC* has also been demonstrated to interact with *E2F1*, another transcription factor postulated to interact with *BCR-ABL1*. Phosphorylation of MYC at serine 62 (S62) is associated with stabilisation of the MYC protein, whereas phosphorylation of MYC at threonine 58 (T58), which only occurs after S62 phosphorylation, is associated with MYC protein degradation.


BCR-ABL1
SH2
GRB-2
EGF
HNRPK
ICSBP
otherwise
Blimp-1
METS/PE-1
GSK-3B
p21
CML growth arrest
miR-17-92
Cluster
PI3K
AKT
Miz-1
E2F1
RAF
MEK
ERK
T58
S62
MAPK
RAS
CML Progression
MYC
Rb
THAP11
JAK2
dephosphorylates
cMYC S62
phosphorylate &
stabilise S62
MAX
MYC/MAX Complex
Activation of transcription
SET
PP2A
PM1
dephosphorylates
NF-KB
CIP2A
BC
Progression

transcription factors *MYC* and *E2F1* are unlikely to be "solely" responsible for the up-regulation of these identified histone genes. While a predictive gene expression classifier for OA was unable to be generated, this analysis did provide further insight into the complexity of OA and CP-CML, indicating that CP-CML patient heterogeneity is most likely having a significant impact on the OA variability, and subsequent therapeutic outcomes, observed in the CP-CML patient cohort.

5 GLOBAL DNA METHYLATION ANALYSIS OF CP-CML PATIENTS WITH LOW AND HIGH OCT-1 ACTIVITY

5.1 Introduction

5.1.1 Epigenetics and DNA methylation

Genetic differences alone cannot fully account for all the phenotypic differences observed between individuals. Epigenetics is responsible for a substantial part of this additional phenotypic control and is highlighted by the modern definition of epigenetics – information heritable during cell division other than the DNA sequence itself.[329] Furthermore, this is demonstrated in genetically identical individuals, including monozygotic twins,[330,331] and cloned animals;[332] where significant DNA methylation and phenotypic differences have been observed.

In mammals, the only know epigenetic modification of DNA is cytosine (C) methylation at position 5 (C5), when this nucleotide occurs next to a guanine (G) in CpG dinucleotides,[333] with approximately 28 million CpGs existing in the haploid human genome in methylated, hydroxymethylated or unmethylated states.[334] Approximately 70% of these CpG sites are thought to exist in the methylated state;[335] however, short regions of high CpG density are located in approximately 60% of gene promoters, known as CpG islands (CGIs), and generally exist in the unmethylated state.[336,337] The DNA methylation machinery is composed of two parts: the DNA methyltransferases (DNMTs), responsible for the establishment and maintenance of DNA methylation patterns; and the methyl-CpG binding proteins (MBDs), responsible for "reading" methylation marks (**Table 5.1**).[338] Therefore, differential DNA methylation is a crucial epigenetic modification of the genome involved in regulating many cellular processes, while also controlling the expression of particular genes during development. This occurs due to the ability of DNA methylation to function as a switch, controlling when to "turn" the expression of certain genes on and off (**Figure 5.1**).[338,339]

Table 5.1 Summary of the mammalian DNA methylation machinery

A)

DNA methyltransferase	Function	Localisation	Activity
DNMT1	Maintenance of methylation, repression of transcription	Replication foci in S-phase	Reference for hemi-methylated DNA
DNMT2	Unknown	Nuclear, not well characterised	Unknown
DNMT3A	Methylation remodelling during embryogenesis, imprint establishment, repression	Heterochromatin and nucleoplasm	*De novo* methylation
DNMT3B	Methylation remodelling during embryogenesis, repeat methylation, repression	Heterochromatin and nucleoplasm	*De novo* and maintenance methylation
DNMT3L	Methylation remodelling during embryogenesis, imprint establishment, spermatogenesis, repression	Heterochromatin and nucleoplasm	Unknown

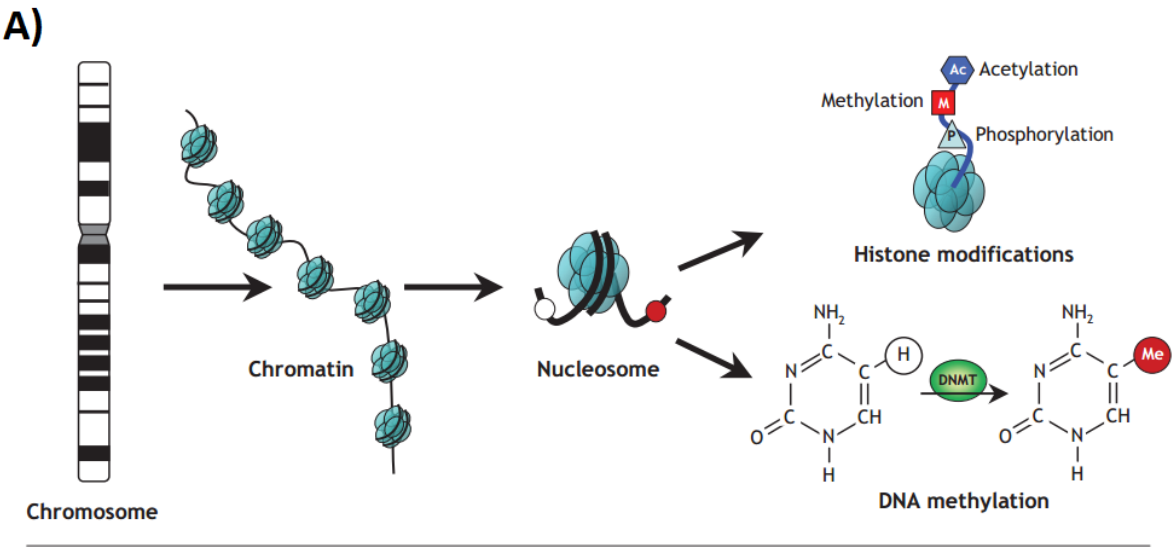
B)

Methyl-CpG binding protein	Function	Repression mediated by:	Binding Specificity
MeCP2	Transcriptional repression	SIN3A, HDAC complex	Single symmetrically methylated CpGs
MBD1	Transcriptional repression	Partially HDAC-dependent repression	Methylated and unmethylated DNA
MBD2	Transcriptional repression	MeCP1 chromatin remodelling complex	Methylated DNA
MBD3	Transcriptional repression	NuRD chromatin remodelling complex	Unmethylated DNA
MBD4	DNA repair, glycosylase domain, repair of deaminated 5-methyl C	NA	$^{5\text{-methyl}}$CpG/TpG mismatches

The DNA methylation machinery in mammals comprises a two-component system. Firstly, DNA methyltransferases **(A)** establish and maintain methylation marks, before the methyl-CpG binding proteins **(B)** then interpret these marks. Table adapted from Robertson.[338]

Figure 5.1 Schematic of epigenetic modifications and how they influence gene expression

A) Nucleosomes are formed by strands of DNA wrapping around histone octamers. These nucleosomes are then organised into chromatin, the building blocks of a chromosome. Reversible and site-specific histone modifications can occur at multiple sites through mechanisms including acetylation, methylation and phosphorylation. DNA methyltransferases (DNMTs) are then able to catalyse the methylation of cytosine residues at position 5 (C5), resulting in DNA methylation. Taken together, these modifications provide a unique epigenetic signature which regulates chromatin organisation and gene expression. **B)** Schematic of the reversible changes in chromatin organisation that influence gene expression. Genes are expressed (switched on) when the chromatin is open (active; unmethylated), and they are inactivated (switched off) when the chromatin is condensed (silent; methylated). White circles represent unmethylated cytosines; red circles represent methylated cytosines. Figure from Rodenhiser and Mann, 2006.[340]


A)
Chromosome
Chromatin
Nucleosome
Acetylation
Methylation
Phosphorylation
Histone modifications
DNMT
DNA methylation
B)
Gene "switched on"
• Active (open) chromatin
• Unmethylated cytosines (white circles)
• Acetylated histones
Transcription possible
Gene "switched off"
• Silent (condensed) chromatin
• Methylated cytosines (red circles)
• Deacetylated histones
Transcription impeded

Epigenetics has a very important role in many biological processes, as the methylation status of the cytosines within each CpG influences chromatin structure and stability, and protein-DNA interactions. Therefore, epigenetics plays a vital role in transcription, X chromosome inactivation, genomic imprinting, host defence against endogenous parasitic sequences, and embryonic development.[341-347] DNA methylation has also been demonstrated to globally decrease with age, both *in vitro*[348] and *in vivo*,[349,350] leading to the development of a model which measures the rate at which an individual's methylome ages,[351] providing evidence that the differences in aging rates are reflected in the transcriptome. Furthermore, DNA methylation plays an important role in cell lineage commitment and differentiation, as every cell lineage has CpGs that are specifically methylated or unmethylated related to that cell type.[352,353] As such, DNA methylation analysis has been proposed as a tool for cell typing,[354] and can be used as a surrogate measure of cell composition in leukocyte samples from diseased or normal states.[355]

DNA methylation is often described as a "silencing" epigenetic mark, due to its role in transcriptional regulation, where dense gene promoter hypermethylation near the transcription start site (TSS) is often associated with gene repression.[329,347] However, the process is far more complex, as the function of DNA methylation varies with context, and the relationship between DNA methylation and transcription is more subtle than first thought.[347] As the majority of DNA methylation work has focussed on CGIs at TSSs, it is important to note that the commonly accepted adage that promoter hypermethylation leads to gene silencing, while promoter hypomethylation leads to gene activation (**Figure 5.1 [B]**), is not always true, as this generally applies to CGIs only. To demonstrate this further, substantial differences in methylation levels have been observed in promoters that are CpG-poor at the TSS.[347] Furthermore, most gene bodies are CpG-poor, yet extensive gene body methylation

has been demonstrated to be a feature of transcribed genes.[356] Recently, extensive positive correlations between active transcription and gene body methylation were demonstrated on the active X chromosome,[357] indicating that gene bodies can become extensively methylated without blocking transcription elongation.[353] This data indicates that methylated sites beyond TSSs/promoters have an important role at the transcriptional level, but a clear understanding of the detailed mechanisms associated with this methylation is still lacking.

5.1.2 Role of epigenetics in cancer

Cancer, as well as many other human diseases, has been associated with aberrant epigenetic regulation.[358] In cancer in particular, the epigenome is often characterised by global DNA methylation changes in combination with altered histone modifications.[359] Furthermore, a key feature of cancer pathogenesis is global DNA hypomethylation and site-specific CGI promoter hypermethylation (**Figure 5.2**), which is commonly observed in both benign neoplasms and early-stage tumours.[359,360] Aberrant DNA hypomethylation has been demonstrated to account for the activation of some proto-oncogenes and lead to the loss of genomic imprinting (the silencing of a single copy of a gene, leaving one working copy to be inherited), of which the *IGF2* gene (insulin-like growth factor-2) is an example.[360-363] However, CGI promoter hypermethylation of tumour-suppressor genes, associated with their transcriptional silencing, is the most documented epigenetic mechanism in human cancers, with genes including *CDKN2A* (cyclin-dependent kinase inhibitor 2A), *BRCA1* (breast cancer-associated-1) and *VHL* (von Hippel-Lindau tumour suppressor), known to be affected by this phenomenon.[359,360] Additionally, while epigenetic alterations are generally thought of as surrogates for genetic change in cancer, it is probable that they also have a critical role in the initiation of cancer progression through the disruption of normal stem/progenitor cell function.[329]

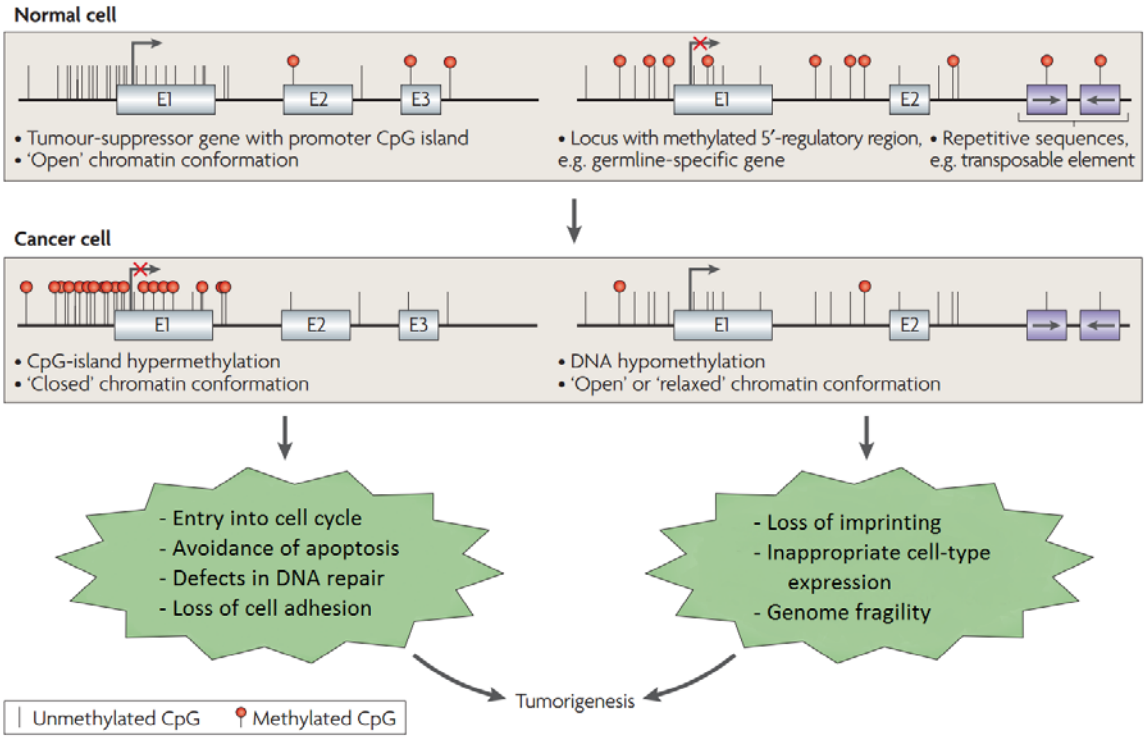


Figure 5.2 Altered DNA methylation patterns in cancer

A common alteration in cancer is the hypermethylation of CpG islands of tumour-suppressor genes. This leads to the transcriptional inactivation of these genes, corresponding to a loss of their normal cellular functions, contributing to many of the hallmarks of cancer including unregulated cell cycle entry, evasion of apoptosis, defects in DNA repair and loss of cell adhesion. Concurrently, in cancer the genome often undergoes global hypomethylation. In some cases, this hypomethylation contributes to the cancer phenotype by affecting mechanisms such as the loss of imprinting and causing inappropriate cell-type expression. It is postulated that this mechanism may contribute to the genomic instability observed in tumours. Abbreviations: E, exon. Figure from Esteller M., 2007.[358]

5.1.3 Role of DNA methylation in haematological cancers

Aberrant DNA methylation patterns have also been demonstrated to play a role in leukaemogenesis, with the majority of research performed in myeloid malignancies such as acute myeloid leukaemia (AML). Recently, Figueroa *et al.*[364] examined a cohort of 344 AML patients for both their gene expression and DNA methylation profiles, using the "HELP" assay which measures over 50,000 CpGs contained within approximately 14,000 unique gene loci. Using the DNA methylation data alone, patients could be segregated into 16 epigenetically unique subtypes. Importantly, 5 of these groups were new AML subtypes, defined only by DNA methylation signatures, with no other common feature. Therefore, AML patients are characterised by distinct epigenetic alterations, with aberrant methylation occurring in both directions; i.e. with either hypo- or hypermethylation predominating; however, gene hypermethylation was the most common.

Aberrant DNA methylation profiles are not only restricted to AML, but are also observed in other leukaemias. Schafer *et al*.[365] recently demonstrated that infants with MLL-rearranged acute lymphoblastic lymphoma (ALL) have a specific hypermethylation profile, while other genetic lesions in ALL have also been associated with aberrant DNA methylation profiles.[366] Furthermore, myelodysplastic syndromes have demonstrated increased aberrant DNA methylation profiles, compared to *de novo* AML, which may explain the responsiveness to DNA methyltransferase inhibitors observed in myelodysplastic syndromes.[367]

5.1.4 Epigenetics and DNA methylation in CML

As CML characterised by the presence of the Bcr-Abl1 oncoprotein, encoded by the Philadelphia chromosome [t(9;22)(q34;q11)], a genetic alteration; research into epigenetic mechanisms involved in the pathogenesis of CML is not well explored. However, aberrant

epigenetic regulation of CML-associated genes may contribute to overall disease pathogenesis, stem cell survival, and therapeutic resistance to tyrosine kinase inhibitor (TKI) therapy,[368] as explained further below.

Methylation of the *ABL1* gene at its Pa promoter, which remains intact during the *BCR-ABL1* translocation,[369] was one of the first epigenetic marks investigated in CML. The Pa promoter has been demonstrated to undergo progressive *de novo* methylation during the course of the disease,[370] with hypermethylation associated with advanced phase (including blast crisis [BC]) CML.[371,372] However, as methylation at the Pa promoter of *ABL1* is observed to vary significantly between studies, ranging from 26%[373] to 77%,[374] 78%[371] and 81%,[375] the significance of *ABL1* hypermethylation in CML is highly disputed. Methylation of the major breakpoint cluster region (M-*bcr*) in CML has also been studied, with increased methylation also observed during the transition to BC; however, differing patterns of methylation were observed depending on the location of the breakpoint in BCR.[376] Although a subsequent study was unable to locate an exact M-*bcr* nucleotide position responsible for the aberrant methylation observed.[377] Importantly, no studies investigating global DNA methylation patterns have been performed in CML, with the majority of studies performed either investigating only a single or a select number of genes of interest, often based on observations from other malignancies using a candidate gene approach.

The cell cycle regulating gene *p15* (*CDKN2B*) is frequently studied in leukaemias, due to its role in inhibition of cell cycle progression; however, the clinical relevance of *p15* methylation in CML is not fully understood.[368] This is because multiple studies have identified both p15 promoter hypermethylation[374,378] and hypomethylation[379,380] in CML patient samples. These contrasting results are likely due to the small sample sizes analysed in each study and the

different methodologies used. More studies have focussed on identifying the methylation of genes associated with progression, response or TKI-resistance in CML. Dunwell *et al*.[381] identified two genes, *TFAP2A* (transcription factor AP-2 alpha) and *EBF2* (early B-cell factor 2) with increased methylation in 8 BC-CML patients, compared to 55 CP-CML patients. Methylation of the *ATG16L2* (autophagy related 16-like 2) gene, found in 69% of CML patients, was also demonstrated to be associated with a lower achievement of major molecular response (MMR) by both 12 and 18 months.[381] Qian *et al.*[382] investigated the methylation patterns of the candidate tumour suppressor gene, *DAPK1* (death-associated protein kinase 1), in a cohort of 49 CML patients covering all disease stages (29 CP, 3 accelerated phase [AP], and 17 BC). A significant association between *DAPK1* promoter hypermethylation and CML disease progression was observed, indicating that *DAPK1* promoter methylation may have a significant role in CML progression.[382]

More recently, Jelinek *et al.*[375] investigated the methylation profile of 10 genes in 120 CML patients from various disease phases (65 CP, 40 AP and 15 BC). Five genes (*DPYS*, *NPM2*, *OSCP1*, *PDLIM4* and *TFAP2E*) were identified from methylation analysis of the K562 cell line, and 3 genes (*CDH13*, *PGR-A* and *PGR-B*) from previous analysis of myeloproliferative neoplasms, while both *ABL1* and *CDKN2B* were included for their previously identified roles in CML.[375] Increased methylation was observed in advanced disease, compared to CP, and in patients resistant or intolerant to imatinib. Furthermore, abnormal methylation of *PDLIM4* was associated with shortened survival, independent of CML stage and imatinib response.[375]

Therefore, due to the lack of global DNA methylation studies performed in CML and the strong evidence of a role for DNA methylation, in the initiation and progression of various cancers, we sought to determine whether epigenetic regulation played a significant role in

OA and whether the global DNA methylation profiles of CP-CML patients with very low OA differed significantly from all other patients.

5.1.5 Approach

The Illumina Infinium HumanMethylation450 (450K) BeadChip was used to perform global DNA methylation profiling (**Section 2.4.10**) using DNA isolated from the total white cells (TWCs; **Section 2.4.7**) of 55 CP-CML patients (n =29, very low OA [OA < 4 ng/200,000 cells]; n = 26, other OA [OA > 4 ng/200,000 cells]). The 450K arrays were performed by the Australian Genome Research Facility (AGRF; Melbourne, Australia). Methylation analysis was performed using the *minfi* package in *R* (**Section 2.6.10**), with various selection criteria used to identify differentially methylated CpGs between patients with very low OA, compared to the rest (**Sections 2.6.10.1** to **2.6.10.3**). Spatial CpG mapping, and regional and gene structural analysis was also performed to identify the locations of the differentially methylated CpGs. Gene enrichment analysis was then performed to identify genes (mapped to the significant CpGs) with common biological function using the GeneGo Metacore™ software suite (**Section 2.6.4**). A predictive DNA methylation classifier for OA was developed with PAMR (**Section 2.6.3**), firstly using the significant CpG list from selection criteria 1A, and secondly using the CpG list from selection criteria 1B. A 10-CpG classifier (94% accuracy) developed using CpGs from selection criteria 1B had the greatest accuracy.

However, before validation could be performed, the presence of single nucleotide polymorphisms at the CpG locus (termed “SNP-CpGs”) were identified in a number of the significant CpGs. SNP genotyping was performed for 6 SNPs identified from the differentially methylated CpG list using the Sequenom MassArray® (**Section 2.4.12.1**) for 100 CP-CML patients (55 matched to the original 450K analysis and 45 independent patient samples).

This assay was performed by the AGRF. SNP variation was observed for 4 of the 5 sites investigated. Therefore, the PAMR predictive OA classifier was generated again using the re-analysed 450K dataset with all SNP-CpGs removed (selection criteria 1C). A 14-CpG classifier (98% accuracy) was subsequently developed. Validation of the 450K methylation results for these 14 CpGs was then performed using bisulphite pyrosequencing (**Section 2.4.11**). Eleven of the 14 CpGs had pyrosequencing assays successfully designed and run by the AGRF using TWC DNA samples from 92 CP-CML patients (**Section 2.4.7**), run in two cohorts (n = 46 each). Cohort 1 contained 44 CP-CML patients matched to the original 450K analysis. The 14-CpG PAMR classifier was then re-analysed using the pyrosequencing data, with a 10-CpG classifier generated (93% accuracy) using data from patients in cohort 1. This classifier was then validated using the 46 independent CP-CML patient samples from cohort 2 of the pyrosequencing analysis.

A further 8 CP-CML patients were then analysed using the 450K methylation arrays (**Section 2.4.10**; AGRF). To ensure no batch effects were present between the methylation results from the original 55 patient analysis and this new 8 patient dataset, complete re-analysis of the combined 450K dataset was performed using the *ComBat* function from the *SVA* package in *R* (**Section 2.6.11**). Methylation analysis was once again performed using the *minfi* package in *R* (**Section 2.6.10**), with a more stringent selection criteria used to identify differentially methylated CpGs between patients with very low OA, compared to the rest (**Sections 2.6.10.4** and **2.6.10.5**). Spatial CpG mapping, and regional and gene structural analysis was then performed for the new CpG list to identify the locations of the CpGs. Gene-set enrichment analysis was performed by analysing the Molecular Signatures Database (MSigDB; **Section 2.6.4**). Finally, gene expression microarray profiling was performed using TWC RNA isolated from 45 CP-CML patients (**Section 2.4.4.2**), matched to the patients used

for the 450K methylation analysis, using the Illumina HumanHT-12 v4 Expression BeadChip (**Section 2.4.14;** AGRF) and analysed using LIMMA (**Section 2.6.11**). The analyses discussed in this chapter were used to determine how the global DNA methylation profiles of CP-CML patients with very low OA differ from all other patients ("other OA"), to further understand the underlying biology associated with very low OA.

5.2 Results

5.2.1 Global DNA methylation profiling identified significant epigenetic differences between CP-CML patients with very low OCT-1 activity compared to the rest

Global DNA methylation profiling was performed using the Illumina Infinium HumanMethylation450 BeadChip (450K), which contains more than 480,000 CpG probes, covering 99% of RefSeq genes and 96% of CpG island (CGI) regions.[199,383] The genome-wide DNA methylation profiles obtained from the TWCs of 55 CP-CML patients with very low OA (n = 29) and other OA (n = 26) were compared (**Table 5.2**) to determine if epigenetics was contributing to the poor response observed in these patients. Analysis was performed using the *minfi* package developed by Hansen, K.D. and Aryee, M.,[201] in the statistical program *R* (**Section 2.6.10**). Six patients failed to pass quality control measures and were excluded from further analysis, leaving 49 patient samples available for the final analysis (n = 27 very low OA and n = 22 other OA). Using a significance cut-off of FDR *P*-value < 0.05 (**Selection criteria 1A**, as described in **Section 2.6.10.1**); 9,861 differentially methylated CpGs were identified, of which 9,403 were hypermethylated and 458 were hypomethylated in patients with very low OA, compared to the rest. A spatial view of these sites mapped across the genome (chromosomes 1 – 22) revealed no bias in the distribution of these CpGs across all chromosomes (**Figure 5.3**).

Furthermore, regional analysis (refer **Figure 5.4**) of these CpG sites in relation to CGIs, including CpG shore (a 2kb region located directly either side of CGIs) and CpG shelf regions (a 2kb region located directly either side of the CpG shore), revealed that the largest number of CpGs were located in CGIs. Approximately 50% were located in CGIs (4,823/9,861 sites), 24% in CpG shores (2,373/9,861 sites), 5% in CpG shelves (493/9,861 sites) and 22% (2,172/9,861 sites) in non-annotated regions in relation to the CGI (**Figure 5.5**). Interestingly,

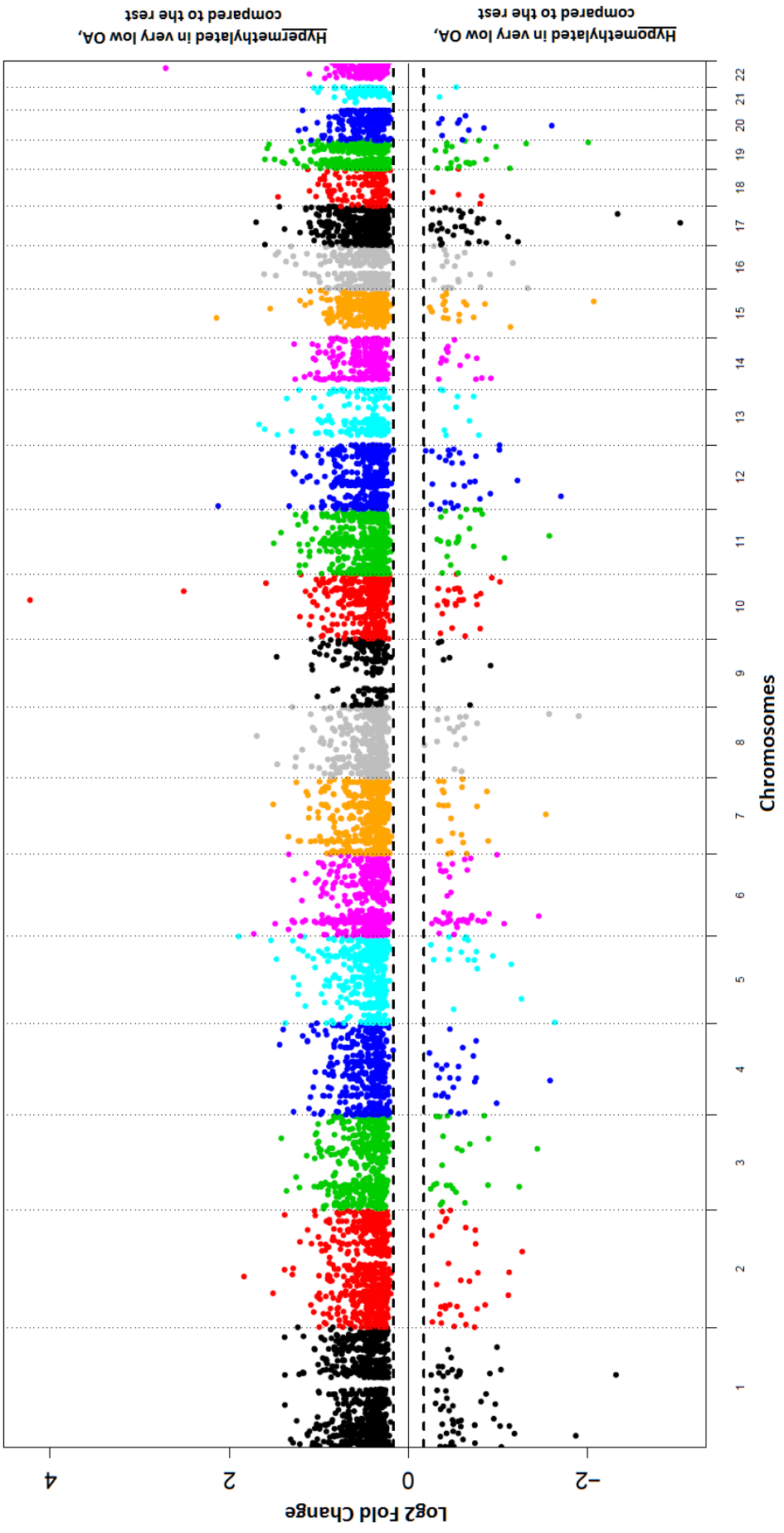
Table 5.2 Characteristics of the 55 CP-CML patients investigated by 450K methylation analysis

		Patient Cohort	
		Very low OA (n = 29)	Other OA (n = 26)
Median Age	**Years**	49	57
	Range	21 – 74	22 – 83
Male	**%**	58.6	53.8
Sokal Score	**High**	6	3
	Intermediate	9	13
	Low	13	9
BCR-ABL1 levels < 10%	**3 month, %**	89.6	96.1
MMR	**12 month, %**	57.1	84.6
	24 month, %	60.7	83.3

One patient in each of the very low and other OA cohorts did not have a Sokal score available. Abbreviations: OA – OCT-1 activity; MMR – major molecular response, *BCR-ABL1* mRNA levels < 0.1%.

Figure 5.3 The spatial view of all significant CpGs differentially methylated between the very low OA patient cohort, compared to the rest revealed an even distribution across all chromosomes

A Manhattan plot was generated using all 9,861 differentially methylated CpGs with a FDR *P*-value < 0.05 and demonstrates an even distribution of the CpGs across all chromosomes. All 22 chromosomes are plotted on the X-axis, with the log2 fold change for each individual CpG plotted on the Y-axis. A negative fold change value indicates a CpG hypomethylated in CP-CML patients with very low OA, compared to the rest, while a positive fold change value indicates CpG hypermethylated in CP-CML patients with very low OA, compared to the rest. Each CpG is mapped to the exact position within its respective chromosome. CpGs mapped to chromosomes X and Y were not included in this plot. Manhattan plots were generated using the *CpGassoc* package,[233] obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/).


Hypomethylated in very low OA,
compared to the rest
Hypermethylated in very low OA,
compared to the rest
Chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
−2
0
2
4
Log2 Fold Change

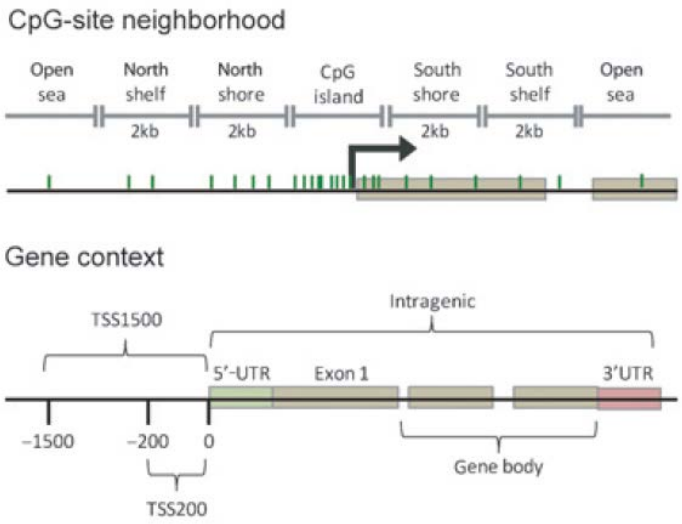


Figure 5.4 CpG probe annotation for both the regional and gene structure analyses

Each CpG probe was assigned to its CpG site neighbourhood, in relation to the CpG island (CGI), for the regional analysis. The CpG shore is a 2kb region located directly either side of the CGI, while the CpG shelf is a 2kb region located either side of the CpG shore. The furthermost region from the CGI is known as the open sea. Each CpG probe was also assigned toward the closest transcription start site (TSS) of RefSeq-annotated genes to determine its location in the gene context. The gene promoter region was defined as containing the TSS1500, TSS200 and the 5'UTR. Abbreviations: TSS200, within 200 bp of the TSS; TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region. Figure from Busche *et al*.[384]

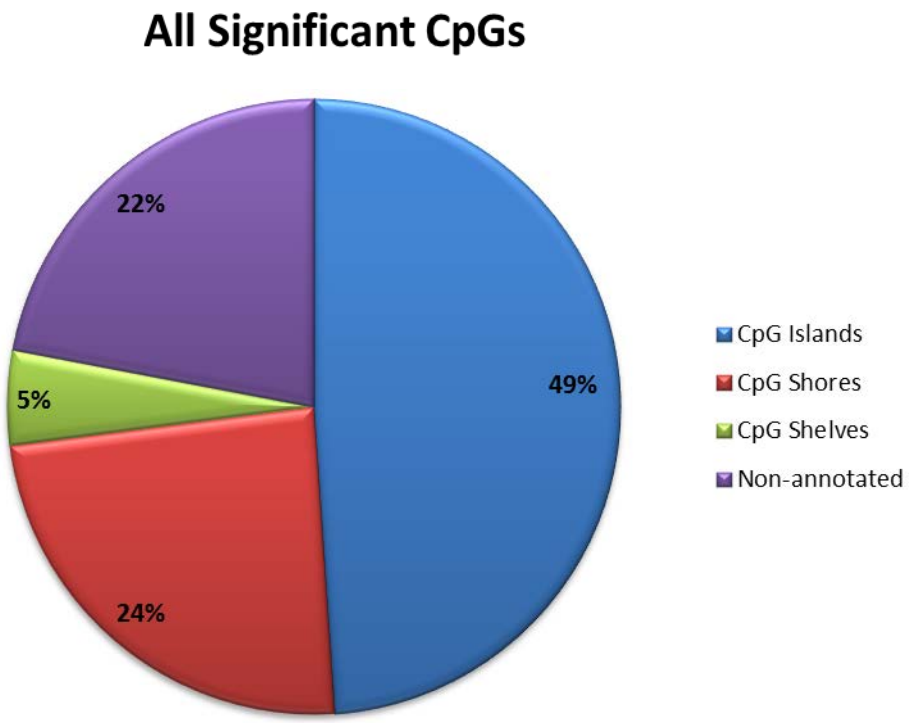


Figure 5.5 The largest proportion of the significant CpG sites from the OA methylation analysis were located in CpG islands

All differentially methylated CpGs (9,861 sites) from CP-CML patients with very low OA, compared the rest, with a known (CpG island, shore or shelf) or unknown (non-annotated) position were plotted. The largest proportion of CpGs were located CpG islands. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays.

when only the significant CpGs hypomethylated in patients with very low OA were investigated, the regional analysis was markedly different, as the majority (60%) of CpGs were located in non-annotated regions (**Figure 5.6 [A]**). However, the CpGs significantly hypermethylated in patients with very low OA (**Figure 5.6 [B]**) had a regional distribution almost identical to the overall distribution, which was expected as these CpGs were responsible for 95% of all significant CpGs. Investigation into the position of these differentially methylated CpG sites in relation to the gene structure (refer **Figure 5.4**) was performed next. The majority of significant CpGs were situated in either the gene body, within 200 bp or 1,500 bp of the transcription start site (TSS; TSS200 and TSS1500, respectively) or within non-annotated regions (**Figure 5.7**). However, as the promoter region of a gene contains the TSS200, TSS1500 and the 5'UTR (5' untranslated region), the largest proportion of CpGs (48%) were located in the promoter when these regions were combined. Once again, this distribution pattern was consistent for the CpGs hypermethylated in patients with very low OA (**Figure 5.8 [A]**); however, the majority of CpG sites hypomethylated in patients with very low OA were situated within the gene body (**Figure 5.8 [B]**).

Using a more stringent selection criterion (**Selection criteria 1B**, as described in **Section 2.6.10.2**), a total of 10 CpGs were identified as differentially methylated in patients with very low OA, compared to the rest (**Table 5.3; Figure 5.9**). Five CpGs were hyper- and 5 hypomethylated in very low OA, compared to the rest, with the largest proportion situated in the gene body (4/10). Only 3 of the CpGs had annotated genomic distribution information available, with 2/3 located in CGIs.

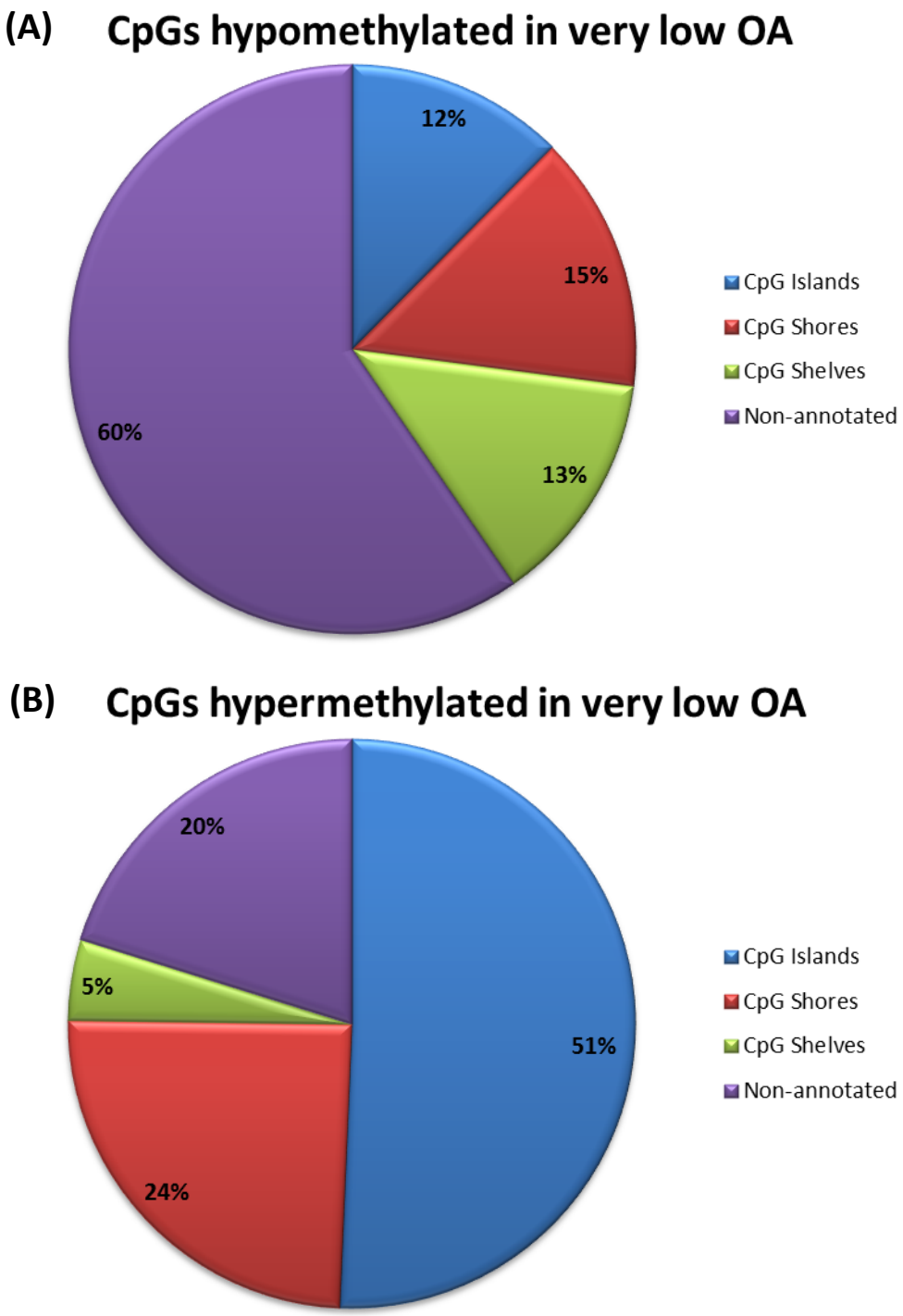


Figure 5.6 Regional location of significant hypo- and hypermethylated CpG sites in patients with very low OA, compared to the remainder of the patients

The regional location of all **(A)** 458 CpGs significantly hypomethylated and **(B)** 9,403 CpGs significantly hypermethylated in CP-CML patients with very low OA, compared to the rest, are demonstrated. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays.

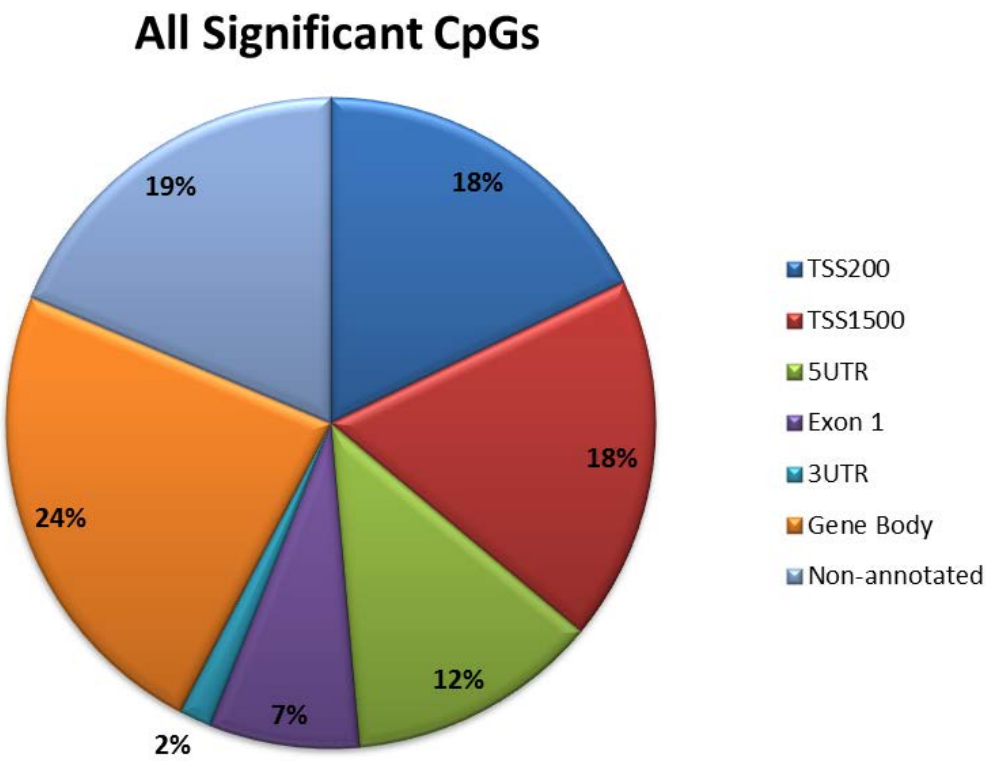


Figure 5.7 The genomic distribution of all significant CpG sites from the OA methylation analysis revealed the largest proportion were located in the promoter region

The positon of all differentially methylated CpGs (9,861 sites) from CP-CML patients with very low OA, compared the rest, in relation to the gene structure are demonstrated. The largest proportion of CpGs were located in the promoter, which consists of the TSS200, TSS1500 and 5'UTR. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

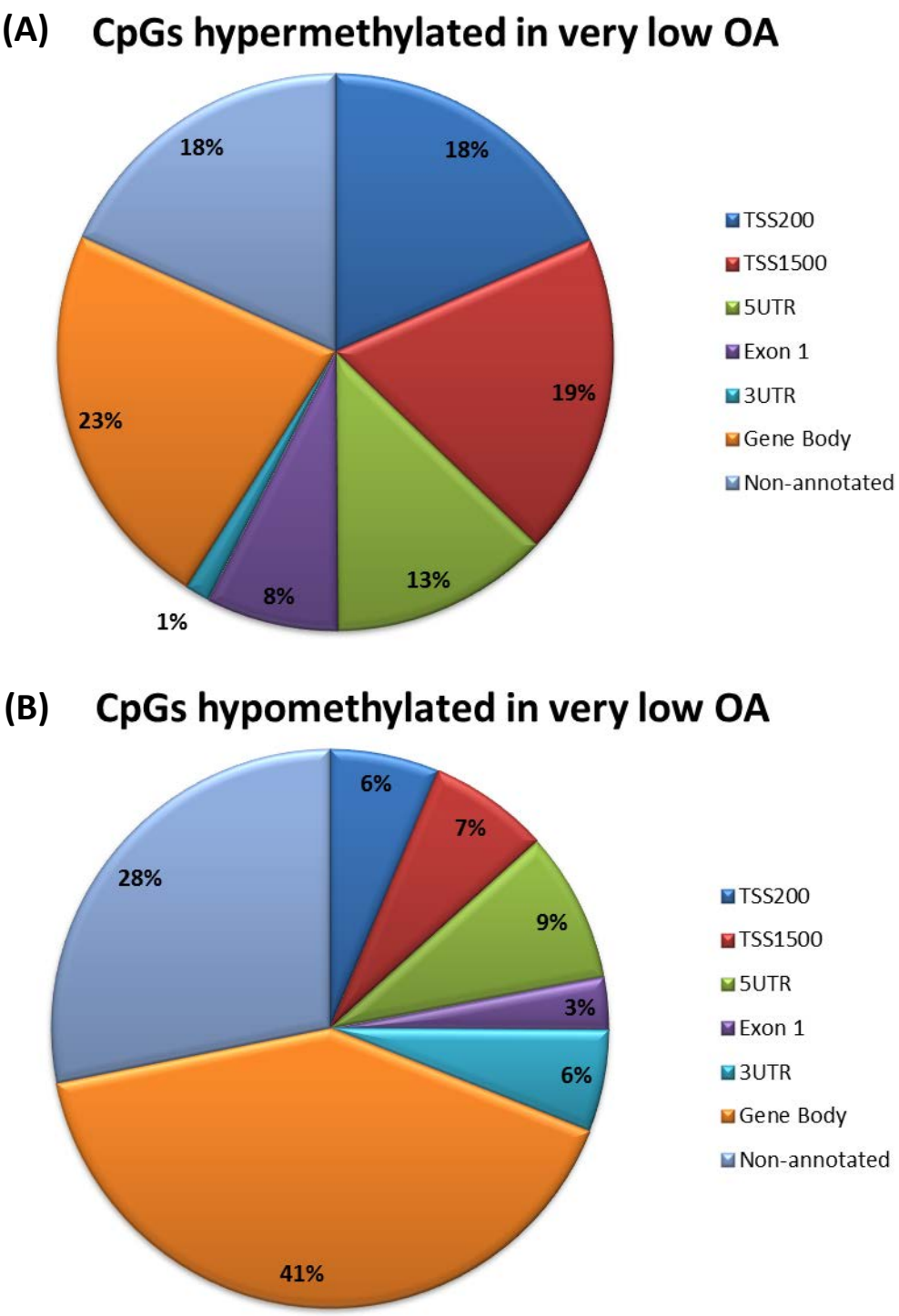


Figure 5.8 Genomic distribution of significant hyper- and hypomethylated CpG sites in patients with very low OA, compared to the remainder of the patients

The position of all **(A)** 9,403 CpGs significantly hypermethylated and **(B)** 458 CpGs significantly hypomethylated in CP-CML patients with very low OA, compared to the rest, in relation to the gene structure are demonstrated. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

Table 5.3 Summary of the 10 CpGs identified from the OA DNA methylation analysis

CpG	Gene	LogFC	FDR *P*-value	Chr	Regional location	Gene location
cg23825092	*TDRD1*	-3.03	0.008	10	TSS200	Island
cg17146918	*TENM4*	2.12	0.009	11	5'UTR	NA
cg14351440	*MCF2*	2.51	0.022	X	Body	NA
ch.4.3383604F	*TRAPPC11*	2.71	0.026	4	Body	NA
cg14179288	*TLE4*	4.22	0.028	9	Body	NA
cg21017569	*KCNAB3*	-2.07	0.032	17	TSS200	Island
cg20510474	NA	2.14	0.041	2	NA	NA
cg24100293	*CASC15*	-2.33	0.042	6	Body	NA
cg26767960	*SMARCAL1*	-2.01	0.043	2	TSS1500	Shore
cg01591343	*WDR27*	-2.31	0.045	6	3'UTR	NA

A negative logFC correlates with hypomethylation in patients with very low OA, compared to the rest; while a positive logFC correlates with hypermethylation. Abbreviations: LogFC, log2 fold change; FDR, false discovery rate; Chr, chromosome; NA, non-annotated, TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5’ untranslated region; and 3UTR, 3’ untranslated region.

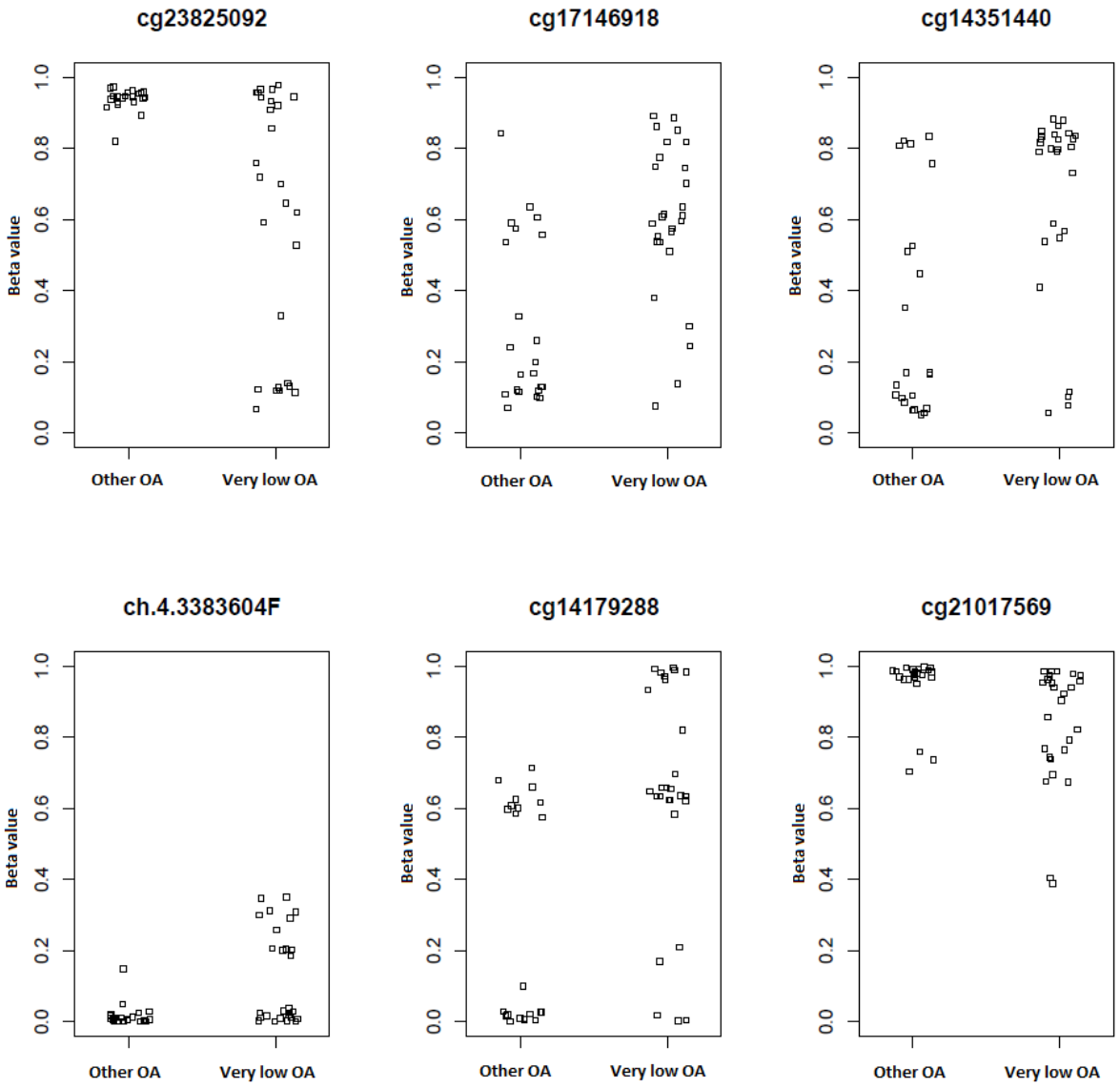
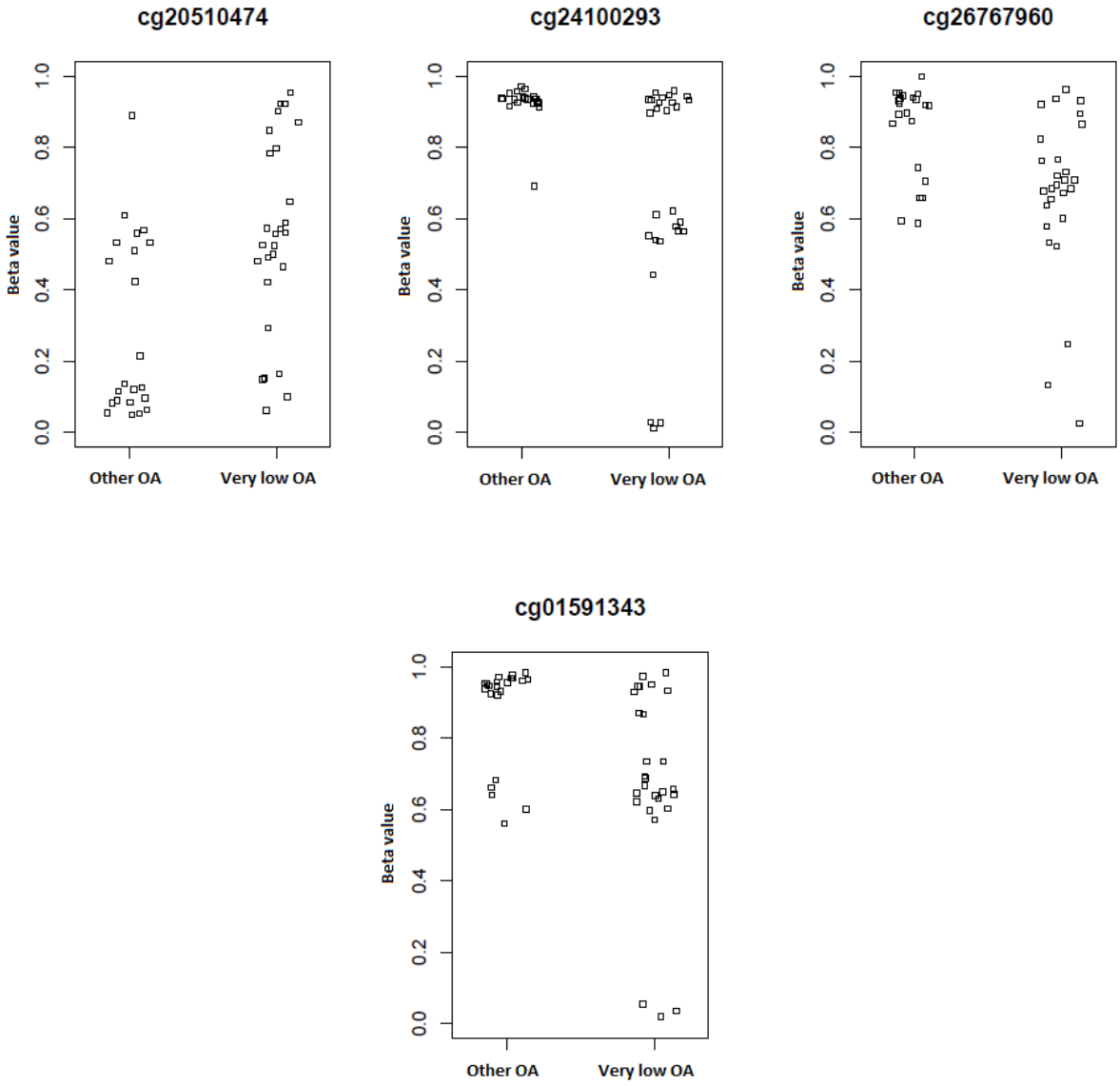


Figure 5.9 Beta-value plots for the 10 significant CpGs identified from the very low versus other OA 450K methylation analysis

The beta-values for each of the 49 CP-CML patients analysed by the 450K methylation array are plotted for each of the 10 CpGs identified as differentially methylated between patients with very low OA, compared to the rest. A beta-value close to 0 represents unmethylated (0%), while a beta-value close to 1 represents methylated (100%). Beta-values were determined using the *minfi* package in R.


cg20510474
cg24100293
cg26767960
cg01591343
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA

5.2.2 Pathway analysis identified significant enrichment of multiple cellular processes associated with very low OA patients, compared to the rest

GeneGo pathway and gene ontology enrichment analysis was performed using the 5,745 genes corresponding to the differentially methylated CpGs identified by the 450K array between CP-CML patients with very low OA, compared the rest. A number of highly significant cellular processes and pathways were enriched for this gene list (**Table 5.4**). Included in this list were pathways involved with signal transduction (NOTCH and WNT signalling), development (Hedgehog and TGF-β signalling), cell adhesion and immune response. The p53 signalling pathway, involved in transcription and cell cycle regulation was also significantly enriched between CP-CML patients with very low OA, compared to the rest (**Figure 5.10**). Notably, these pathways all have established roles in CML, including disease progression. Enrichment analysis was not performed for the 10 CpGs identified from the more stringent analysis, which correlated with 9 known RefSeq genes, due to the small number of genes in the gene-set.

5.2.3 Development of a predictive classifier for OA using the 450K DNA methylation data

DNA methylation classifiers have the potential to yield rapid and accurate prognostic and predictive markers for use in the clinical setting. This is exemplified by multiple studies which have confirmed that the DNA methylation of specific genes may be associated with clinical outcome in both acute myeloid leukaemia (AML)[364,385] and myelodysplastic syndrome.[386] Therefore, the Prediction Analysis for Microarrays in R (PAMR) package, which uses machine learning algorithms developed by Tibshirani *et al.*[213,214] to perform sample classification using the nearest shrunken centroids method, was utilised to develop a predictive classifier from the 450K DNA methylation data. Furthermore, this classifier will help to determine the CpGs with the greatest predictive power for differentiating patients with very low OA from the rest.

Table 5.4 Common cellular processes and pathways enriched in very low OA patients compared to the rest by GeneGo analysis

	Total gene-set size	Genes in dataset	*P*-value
Process Networks			
Signal transduction – NOTCH signalling	236	117	$7.94e^{-09}$
Cell cycle – G1-S Growth factor regulation	195	92	$4.93e^{-06}$
Development – Hedgehog signalling	254	112	$2.42e^{-05}$
Signal transduction – WNT signalling	177	82	$3.67e^{-05}$
Cell adhesion – Amyloid proteins	195	87	$1.17e^{-04}$
Cellular Pathways			
Cytoskeleton remodelling – TGF, WNT and cytoskeleton remodelling	111	53	$6.27e^{-12}$
Development – TGF-β receptor signalling	50	29	$1.22e^{-09}$
Transcription – p53 signalling pathway	39	24	$6.40e^{-09}$
Cell adhesion – Chemokines and adhesion	100	44	$9.58e^{-09}$
Immune respone – CD40 signalling	65	32	$1.08e^{-07}$

The total gene-set size column represents the total number of genes in the GeneGo gene-set, while the genes in dataset column represents the number of genes overlapping between the gene list of interest (from the very low OA versus the rest analysis) and the GeneGo gene-set.

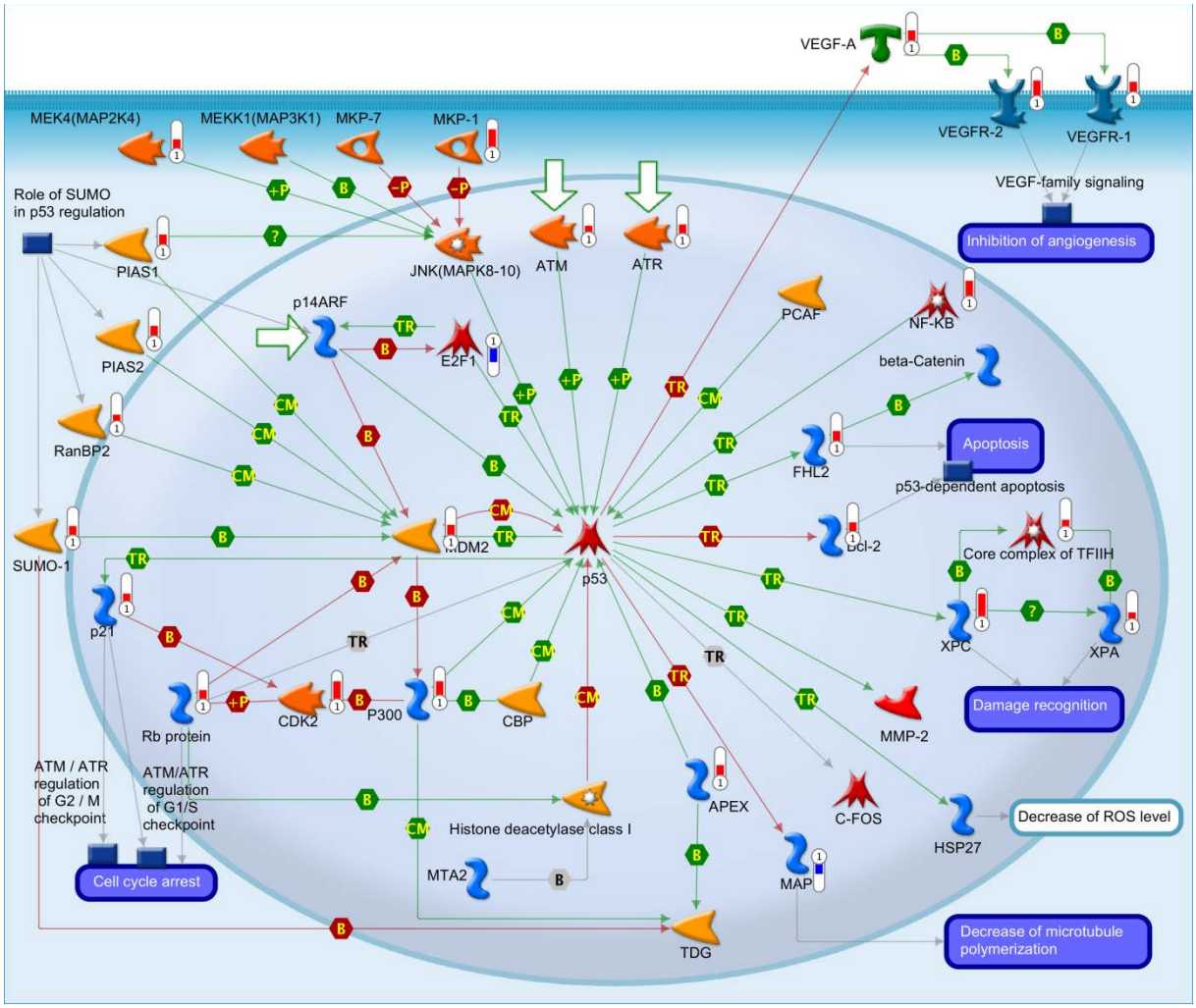


Figure 5.10 p53 signalling pathway enriched in CP-CML patients with very low OA compared to the rest

p53 is a tumour suppressor protein involved in regulation of the cell cycle and changes in p53 are commonly associated with blast crisis CML.[387] Therefore, the possible involvement of this pathway in the poor response observed in very low OA is of significant interest. GeneGo™ enrichment analysis identified significant enrichment (P = 6.40e^{-09}) of the p53 signalling pathway in CP-CML patients with very low OA, compared to the rest. A **red** column next to a gene/complex indicates increased methylation in CP-CML patients with very low OA, while a **blue** column indicates decreased methylation in CP-CML patients with very low OA, compared to the rest. This figure was generated using GeneGo™ software.

PAMR was applied to the significant 450K DNA methylation dataset (9,816 CpGs) to identify signature CpG sites predictive for patients with very low OA. The 49 patient dataset was used as both the training and test sets (internal validation) and was able to generate a 24 CpG classifier (threshold = 4.327) with an overall accuracy of 84% (sensitivity = 89%, specificity = 77%; **Figure 5.11**). The 24 CpG sites selected by this classifier are listed in **Table 5.5**. Three of the 27 very low OA patients were misclassified and 5 of the 22 other OA patients were misclassified, indicating that this classifier was very good at predicting CP-CML patients with both very low and other OA.

To determine if a predictive classifier with a smaller number of CpGs could be developed, PAMR was then applied to the 10 significant CpG list (**Table 5.3**) identified from the 450K DNA methylation analysis of patients with very low OA, compared to the rest (**Selection criteria 1B**). Once again, the 49 patient dataset was used as both the training and test sets (internal validation). A classifier using all 10 CpGs was generated (threshold = 0.9), with an overall accuracy of 94% (sensitivity = 93%, specificity = 95%; **Figure 5.12**). As only 2 of the 27 very low OA patients, and 1 of the 22 other OA patients, were misclassified, this classifier was excellent at predicting patients with both very low and other OA. Therefore, this classifier had superior predictive power compared to 24-CpG classifier.

5.2.4 Single nucleotide polymorphisms at a CpG locus influences the validity of the DNA methylation results produced

Further investigation into the 10 CpGs differentially methylated between patients with very low OA, compared to the rest, identified that 6 of the identified CpGs had single nucleotide polymorphisms (SNPs) within the CpG site (SNP-CpGs; **Table 5.6**). This is a highly significant observation, as the differential methylation thought to be observed between the OA

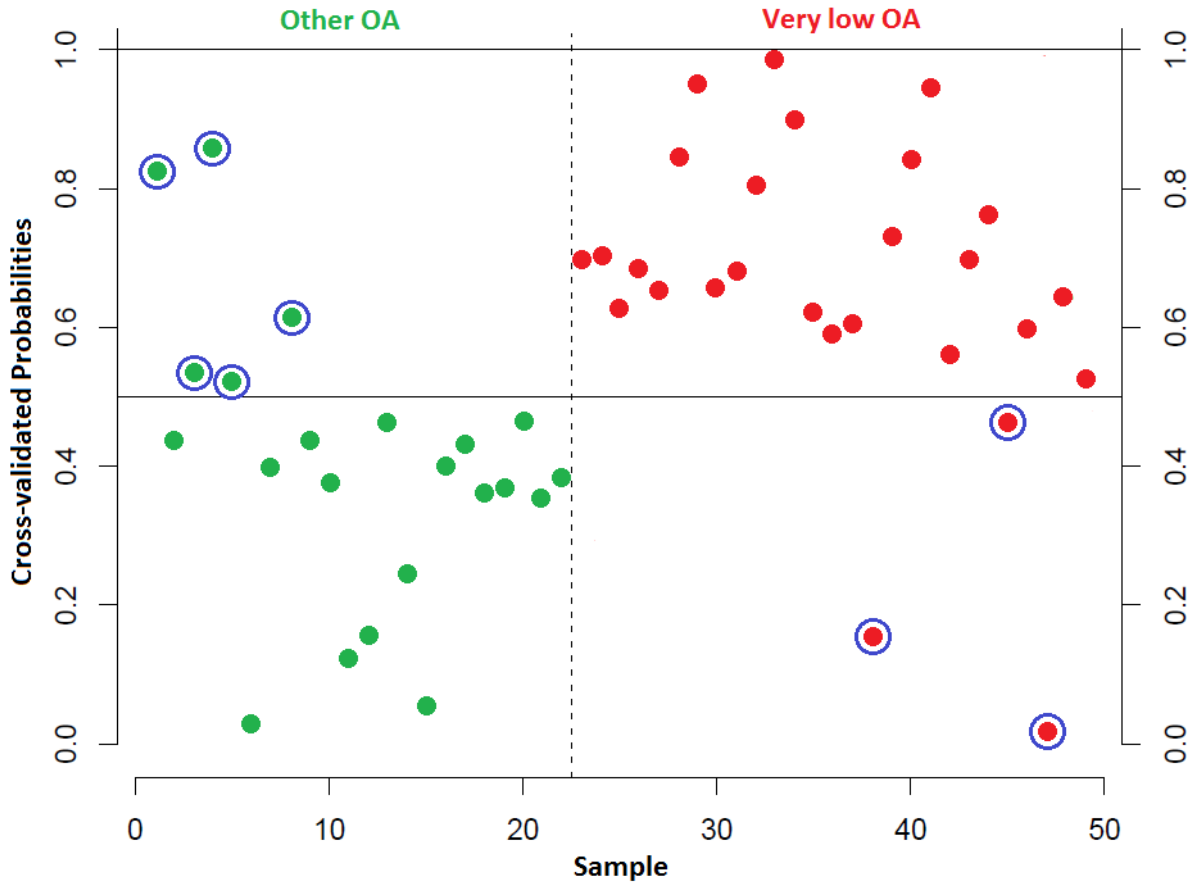


Figure 5.11 The 24-CpG PAMR OA classifier developed using the significant 450K DNA methylation dataset is 84% accurate

Results of the validation of the 24-CpG classifier developed by PAMR using the significant (9,816 CpGs) 450K DNA methylation dataset (n = 49 training set), validated against the same 49 patients. A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). Eight patients were misclassified (**blue** circles; 3 very low OA and 5 other OA patients), providing an overall accuracy of 84%.

Table 5.5 Summary of the 24 CpGs identified from the PAMR classifier for OA developed using the 450K DNA methylation dataset

CpG	Gene	Chr	Regional location	Gene location
cg00290627	NA	7	NA	Island
cg00297767	*HMGN1*	21	TSS1500	Island
cg01663953	*CCDC81*	11	3'UTR	NA
cg01972394	*ALDH4A1*	1	TSS1500	Island
cg05203949	NA	19	NA	S_Shelf
cg05451912	*HIST1H4H*	6	TSS200	Island
cg06513149	*N4BP2L1*	13	TSS200	S_Shore
cg06826278	*ONECUT1*	15	Body	N_Shore
cg08446143	*UBA1*	X	TSS1500	N_Shore
cg08813325	*SORCS2*	4	Body	NA
cg09411922	NA	10	NA	S_Shelf
cg10888523	NA	10	NA	NA
cg11706469	*FGF8*	10	Body	N_Shore
cg12415687	*PTPRN2*	7	Body	NA
cg12830327	*TMEM126B*	11	TSS200	Island
cg14609407	*SLPI*	20	1stExon	NA
cg16715129	*AMPD3*	11	1stExon	NA
cg17932802	*SLC16A3*	17	TSS200	Island
cg19212550	*DNMBP*	10	5'UTR	NA
cg22136351	*RFESD*	5	TSS200	Island
cg22961716	*DPY19L2P2*	7	TSS1500	Island
cg23889010	*SLPI*	20	Body	NA
cg24877792	*ZFP62*	5	TSS200	N_Shore
cg25417551	*TADA2A*	17	1stExon	S_Shore

Abbreviations: Chr, chromosome; NA, non-annotated, TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5’ untranslated region; and 3UTR, 3’ untranslated region..

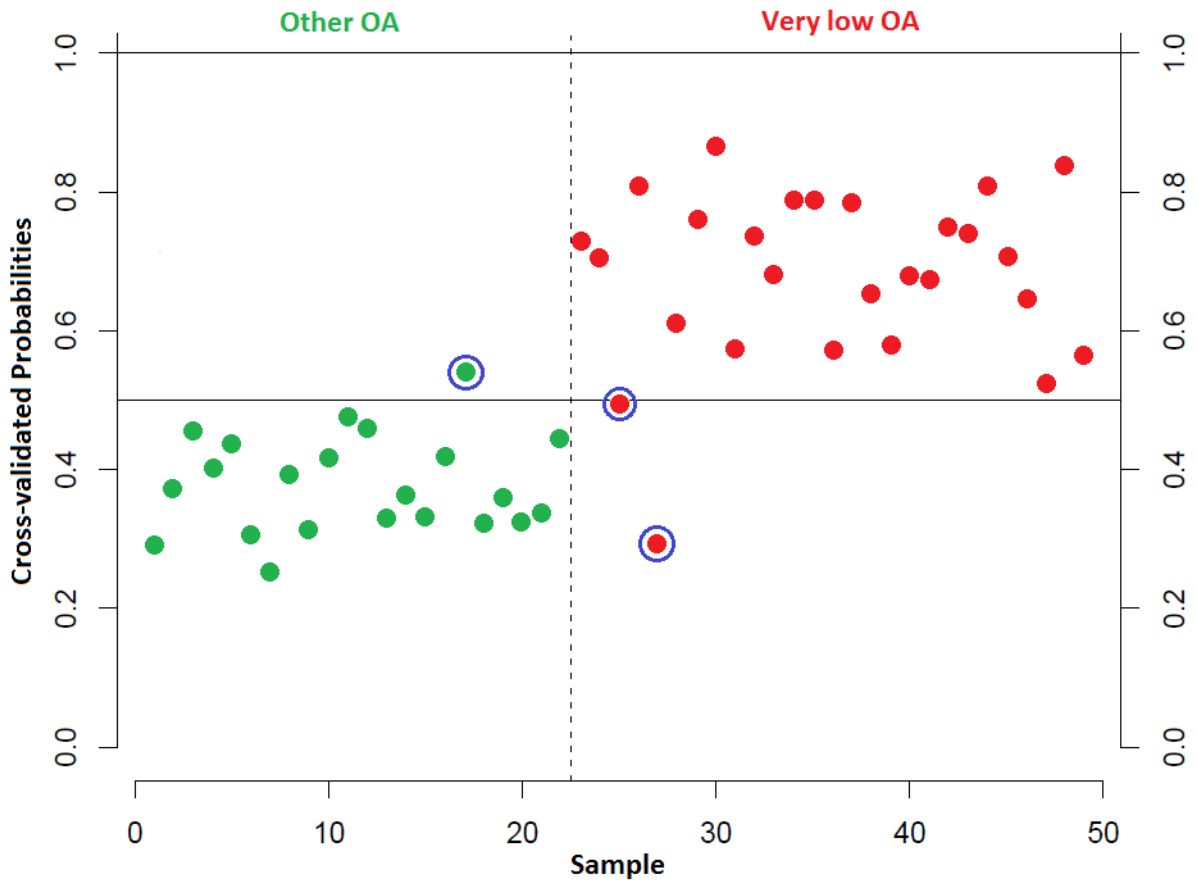


Figure 5.12 The 10-CpG PAMR OA classifier developed using the 10 significant CpG dataset from the 450K DNA methylation analysis is 94% accurate

Results of the validation of the 10-CpG classifier developed by PAMR using the 10 CpG significant 450K DNA methylation dataset (n = 49 training set), valiated against the same 49 patients. A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). Three patients were misclassified (**blue** circles; 2 very low OA and 1 other OA patient), providing an overall accuracy of 94%.

Table 5.6 The 6 CpGs with known SNPs, at or near the CpG locus, identified within the 10 CpG list

CpG	SNP	Distance from CpG locus (bp)	MAF	Gene
cg01591343	rs1047910	1	0.394	WDR27
cg14179288	rs2297498	1	0.484	TLE4
cg14351440	rs5907612	1	0.008	MCF2
cg21017569	rs9898955	5	0.349	KCNAB3
cg23825092	rs78932533	47	0.024	TDRD1
cg24100293	rs9368372	0	0.500	*CASC15*

The minor allele frequency (MAF) refers to the frequency at which the least common allele occurs in a given population.

subgroups, may in fact be a surrogate read-out of the SNP variation within the patient samples and not differential methylation at all. Therefore, SNP genotyping was performed for the 6 SNPs of interest using with 100 CP-CML patient samples, including the 55 patients originally used for the 450K DNA methylation analysis and an independent 45 patients to be used as an independent validation cohort in downstream analyses.

One SNP (rs9898955; cg21017569, KCNAB3) failed to produce any results, most likely due to unforseen primer interactions in the multiplex assay, and was subsequently omitted from all further analysis. Therefore, high quality genotyping data was obtained for 5 SNPs, which demonstrated that for 4 targets, SNP variation was present at the SNP-CpG locus (**Table 5.7**). Interestingly, no SNP variation was observed for rs78932533 (cg23825092; TDRD1) as all patients analysed were determined to contain the common homozygote (A) allele. Further supporting these results was the observation that for the patients matched to the 450K methylation analysis, the "methylation" levels determined from the array could be grouped into high, medium and low methylation groups, which correlated with the allele frequency (e.g. common homozygote, heterozygote, or rare homozygote; **Appendix 10**). Taken together, this data demonstrates that for 4 out of the 5 SNPs investigated, SNP variation at the SNP-CpG locus was responsible for the "differential methylation" read-out produced from the 450K array.

5.2.5 Development of a predictive classifier for OA using re-analysed 450K DNA methylation data to exclude SNP-CpGs

As 4 of the original 10 CpGs used to develop the predictive classifier for OA were found to contain SNPs within the CpG locus, responsible for a false differential methylation read-out, and the effect of a SNP on an additional single CpG could not be excluded, only 5 CpGs with confident DNA methylation data remained. Therefore, PAMR was applied to the 5 significant

Table 5.7 Genotyping identified SNPs, at or near the CpG locus, were responsible for the differential methylation read-out obtained from 4 out of 5 CpGs

Assay	CpG	Common Homozygote	Hetero-zygote	Rare Homozygote	Fails	Assay Pass Rate
rs1047910	cg01591343	54	41	4	1	99%
rs2297498	cg14179288	38	47	14	1	99%
rs5907612	cg14351440	40	21	38	1	99%
rs78932533	cg23825092	98	0	0	2	98%
rs9368372	cg24100293	75	21	3	1	99%

The successful pyrosequencing results for 5 CpG-SNPs are listed. Only one SNP (rs78932533, cg23825092) had no variation present in the 100 patient cohort analysed, with all successful assays indicating that the common homozygote was present. The common homozygote column lists the number of patients with two wild-type alleles, while heterozygote indicates the patients with both a wild-type and variant allele and the rare homozygote column represents the patients with two variant alleles. The fails column indicates the number of patient samples that failed the assay, which is further represented by the assay pass rate.

CpG list (cg17146918, cg20510474, cg26767960, ch.4.3383604F and cg23825092) with no SNP-CpGs, identified from the 450K DNA methylation analysis of patients with very low OA, compared to the rest. As before, the 49 patient dataset was used as both the training and test sets (internal validation). A classifier using all 5 CpGs was generated (threshold = 0.0), with a surprisingly high overall accuracy of 90% (sensitivity = 89%, specificity = 91%). As only 3 of the 27 very low OA patients, and 2 of the 22 other OA patients, were misclassified, this classifier was excellent at predicting CP-CML patients with both very low and other OA. However, although this classifier had a higher accuracy than expected due to the low number of CpGs used in its development, the overall predictive power (90%) was lower than the best classifier (albeit including SNP-CpGs) developed previously (94%).

Therefore, to identify the best possible classifier for OA from the 450K DNA methylation, the entire dataset was analysed again this time excluding any CpGs that contained a known SNP within 10 bp of the CpG probe. Also, to identify the CpGs with most significant differences in methylation between the subgroups (very low versus other OA), the CpG probes with a difference in mean methylation levels, as determined by the β-value ≥ 0.17 ($\Delta\beta \geq 0.17$) were selected (**Selection criteria 1C**, **Section 2.6.10.3**). Using this stringent selection criterion, 14 CpGs were identified as differentially methylated in patients with very low OA, compared to the rest (**Table 5.8**; **Figure 5.13**). Nine CpGs were hyper- and 5 hypomethylated in very low OA, compared to the rest. Finally, PAMR was applied to this new 14 significant CpG list, with the 49 patient dataset used once again as both the training and test sets (internal validation). A classifier using all 14 CpGs was developed (threshold = 0.0), with an excellent overall accuracy of 98% (sensitivity = 100%, specificity = 96%, as expected; **Figure 5.14**), as only 1 of the 22 other OA patients, and none of the patients with very low OA, were misclassified.

Table 5.8 Summary of the 14 CpGs identified from the OA DNA methylation analysis, excluding SNP-CpGs and SNPs within 10 bp of the target CpG

CpG	Gene	Δβ	FDR *P*-value	Chr	Regional location	Gene location
cg16334795	*BACE2*	0.192	0.027	21	TSS1500	N_Shore
cg01663953	*CCDC81*	0.187	0.011	11	3'UTR	NA
cg27053299	*CLYBL*	0.182	0.031	13	3'UTR	Island
cg00448707	*DIRAS3*	-0.172	0.020	1	Body	Island
cg03343083	*LEPREL1*	0.179	0.012	3	5'UTR	NA
cg21995068	*LOC100134868*	0.192	0.031	20	TSS1500	N_Shore
cg04524088	*MIR129-1*	0.196	0.038	7	TSS200	NA
cg17146918	*TENM4*	0.282	0.010	11	5'UTR	NA
cg10987071	*SMARCAL1*	-0.200	0.031	2	TSS1500	N_Shore
cg17394237	*SORCS2*	-0.235	0.047	4	Body	S_Shore
cg06173889	*SOX11*	-0.219	0.048	2	3'UTR	N_Shore
cg23825092	*TDRD1*	-0.327	0.009	10	TSS200	Island
cg17137424	*TRIM67*	0.190	0.029	1	1stExon	Island
cg08191073	*TSHZ2*	0.211	0.004	20	Body	NA

A negative Δβ value correlates with hypomethylation in patients with very low OA, compared to the rest; while a positive Δβ value correlates with hypermethylation in patients with very low OA, compared to the rest. Abbreviations: Δβ, difference in mean β-value; FDR, false discovery rate; Chr, chromosome; NA, non-annotated, TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

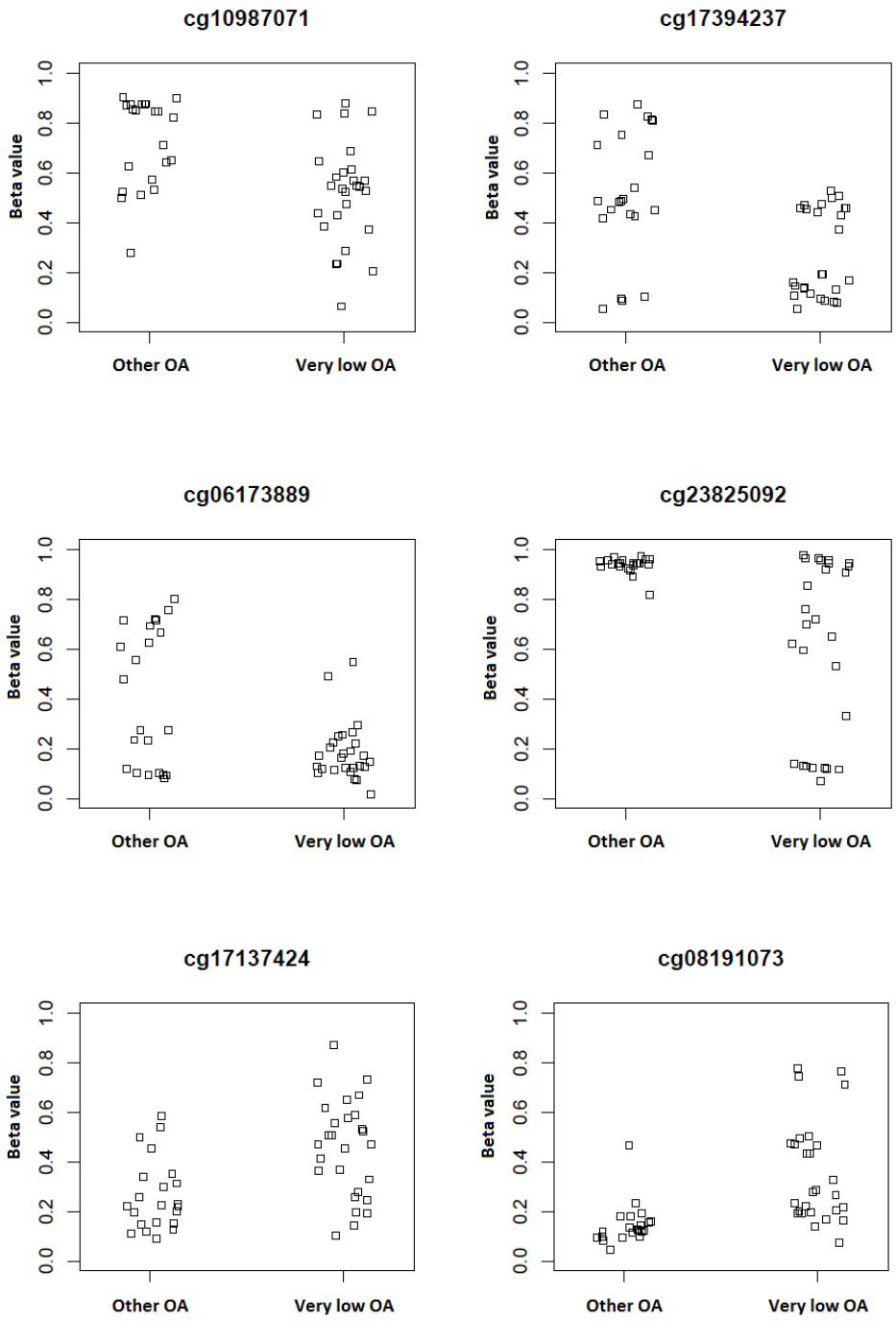
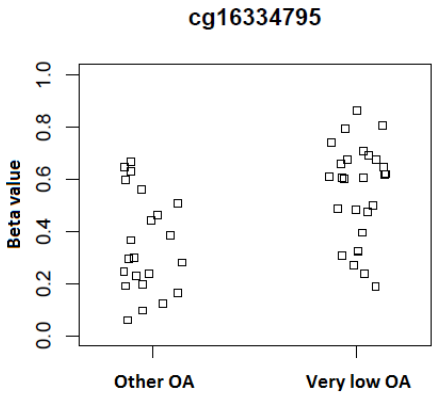
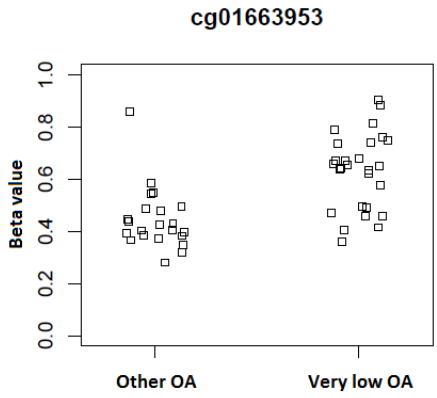
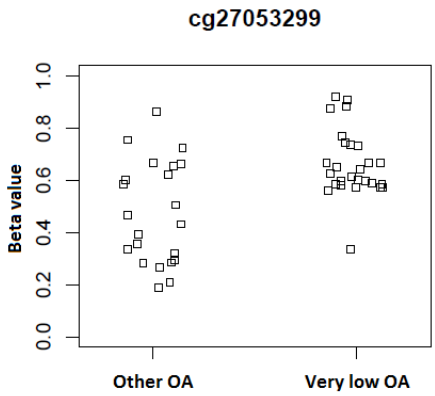


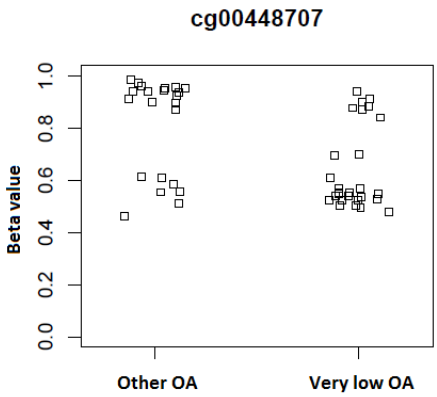
Figure 5.13 Beta-value plots for the 14 significant CpGs identified from the very low versus other OA 450K methylation analysis, excluding CpGs confounded by SNPs

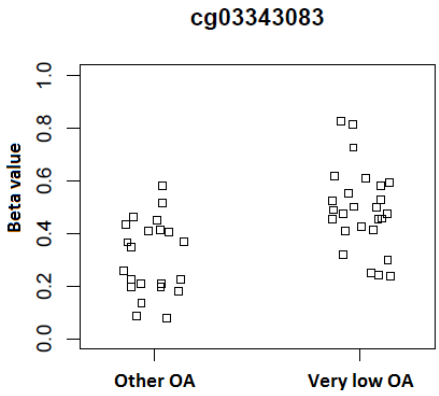
The beta-values for each of the 49 CP-CML patients analysed by the 450K methylation array are plotted for each of the 14 CpGs identified as differentially methylated between patients with very low OA, compared to the rest. A beta-value close to 0 represents unmethylated (0%), while a beta-value close to 1 represents methylated (100%).

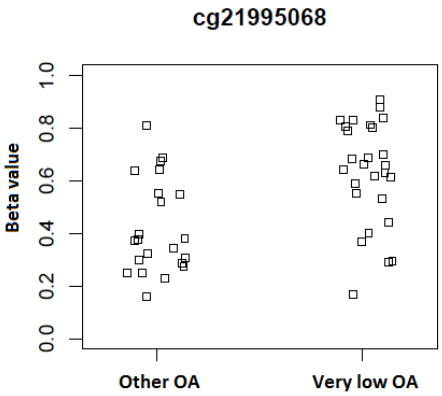

cg16334795
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA

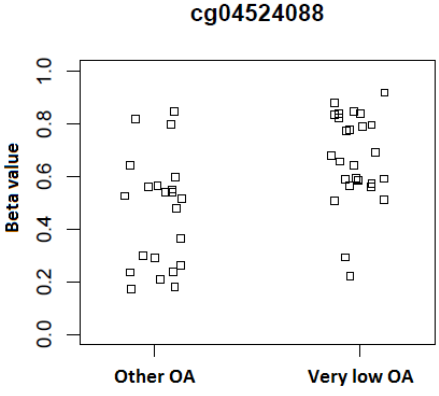

cg01663953
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA

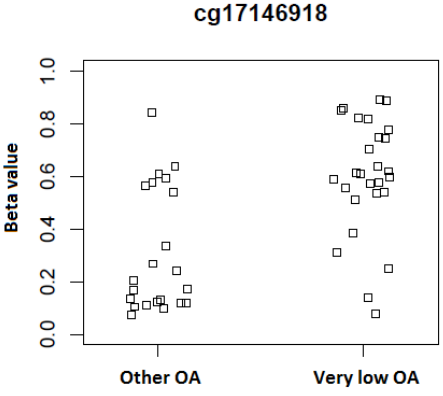

cg27053299
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA


cg00448707
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA


cg03343083
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA


cg21995068
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA


cg04524088
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA


cg17146918
Beta value
0.0
0.2
0.4
0.6
0.8
1.0
Other OA
Very low OA

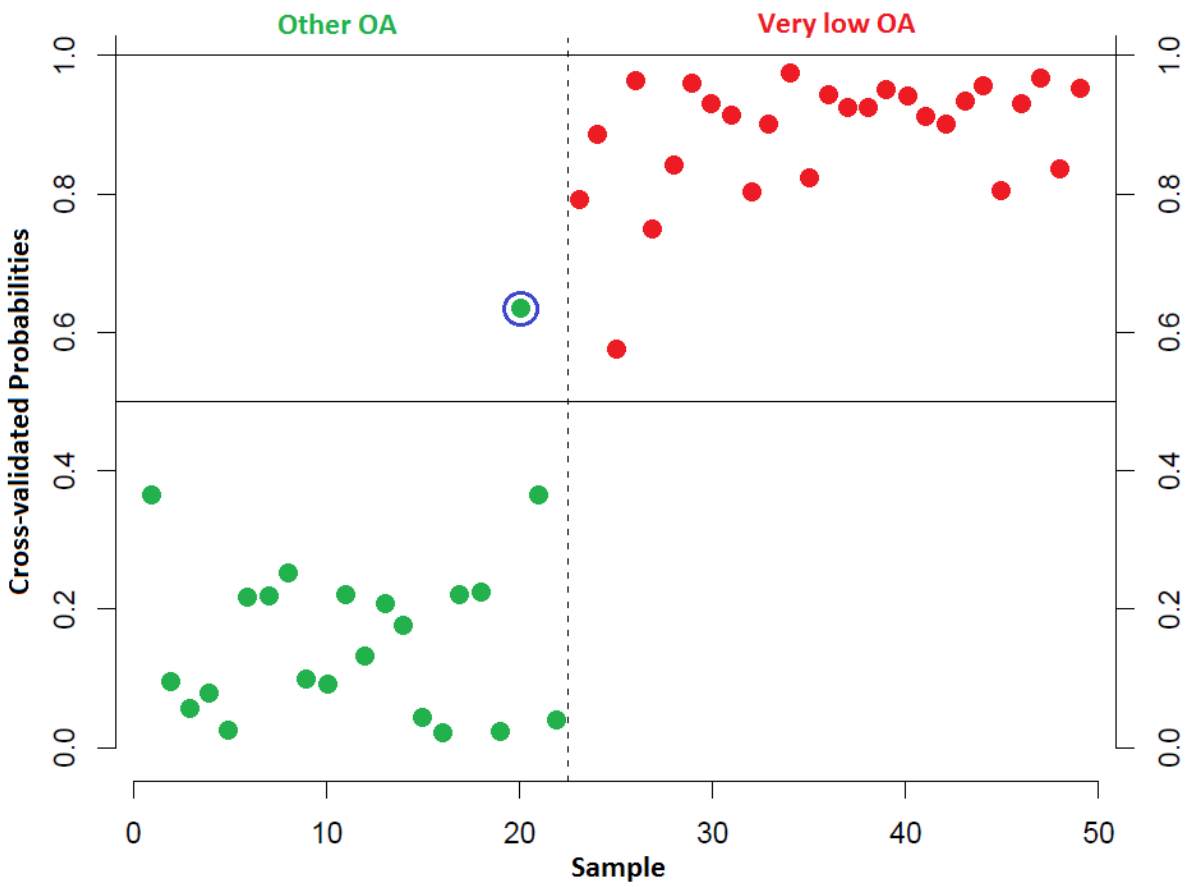


Figure 5.14 The 14-CpG PAMR OA classifier developed using the 14 significant CpG dataset from the 450K DNA methylation analysis is 98% accurate

Results of the validation of the 14-CpG classifier developed by PAMR using the 14 CpG significant 450K DNA methylation dataset (n = 49 training set), validated against the same 49 patients. A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). One patient was misclassified (**blue** circle; 1 other OA patient), providing an overall accuracy of 98%.

Therefore, this classifier had superior predictive power compared to all previously developed classifiers and was used for further analyses.

5.2.6 Pyrosequencing successfully validates target CpGs from the 450K dataset

To validate the 14 significant CpG sites identified from the classifier and 450K analysis between CP-CML patients with very low OA, compared to the rest, bisulphite pyrosequencing was performed (**Chapter 2**, **Section 2.4.11**). Pyrosequencing allows DNA methylation quantitation of the exact CpG site targeted by the 450K array to be performed. Of the 100 CP-CML patients for which SNP genotyping was performed (**Section 5.2.4**), 92 were analysed by pyrosequencing in 2 independent cohorts. Cohort 1 contained 46 patients (**Table 5.9 [A]**), including 44 patients matched to those analysed by the 450K DNA methylation array. Only these 44 matched patient samples were used for validation. Unfortunately, due to restricted sample availability, 5 patients from the original 450K array were unable to be analysed by pyrosequencing. Cohort 2 contained an independent 46 patients (**Table 5.9 [B]**), not previously analysed on the 450K array. Of the 14 CpGs initially validated by pyrosequencing, primer design and assay validation was successful for 11 CpGs. One CpG mapped to a micro RNA (miRNA; cg04524088) where pyrosequencing was unavailable, while the other 2 CpGs (cg16334795 and cg01663953) were located in densely packed CpG islands, where the presence of homopolymers (single base repeat regions) downstream of the target sequence prevented optimal primer design. For 7 of the CpGs analysed by pyrosequencing assays, methylation levels for a number of CpGs flanking the initial CpG of interest were also provided (**Table 5.10**).

Pyrosequencing analysis demonstrated excellent correlation with the methylation levels from the 450K analysis for 10 of the 11 CpGs investigated (**Figure 5.15**). However, for 1 CpG

Table 5.9 Characteristics of the 92 CP-CML patients used for pyrosequencing methylation analysis

(A)

		Patient Cohort 1	
		Very low OA (n = 26)	Other OA (n = 20)
Median Age	**Years**	45	52
	Range	19 – 72	18 – 81
Male	**%**	57.7	45.0
BCR-ABL1 levels < 10%	**3 month, %**	92.0	94.4
MMR	**12 month, %**	50.0	80.0
	24 month, %	61.5	80.0

(B)

		Patient Cohort 2	
		Very low OA (n = 17)	Other OA (n = 29)
Median Age	**Years**	40	52
	Range	18 – 74	17 – 71
Male	**%**	47.1	68.9
BCR-ABL1 levels < 10%	**3 month, %**	70.6	82.8
MMR	**12 month, %**	58.8	72.4
	24 month, %	70.6	82.7

Patient cohort 1 **(A)** and cohort 2 **(B)** as used for the pyrosequencing analysis. Abbreviations: OA – OCT-1 activity; MMR – major molecular response, *BCR-ABL1* mRNA levels < 0.1%.

Table 5.10 Summary of the 11 CpGs analysed by bisulphite pyrosequencing

CpG from 450K	Gene	Number of CpGs in assay (including target CpG)
cg27053299	*CLYBL*	3
cg00448707	*DIRAS3*	3
cg03343083	*LEPREL1*	1
cg21995068	*LOC100134868*	2
cg17146918	*TENM4*	1
cg10987071	*SMARCAL1*	1
cg17394237	*SORCS2*	1
cg06173889	*SOX11*	2
cg23825092	*TDRD1*	2
cg17137424	*TRIM67*	3
cg08191073	*TSHZ2*	4

The design of the pyrosequencing assays meant that for a number of the CpGs identified from the 450K analysis, methylation at multiple CpGs, flanking the CpGs of interest, was also measured. The number of CpGs actually measured by each pyrosequencing assay is demonstrated.

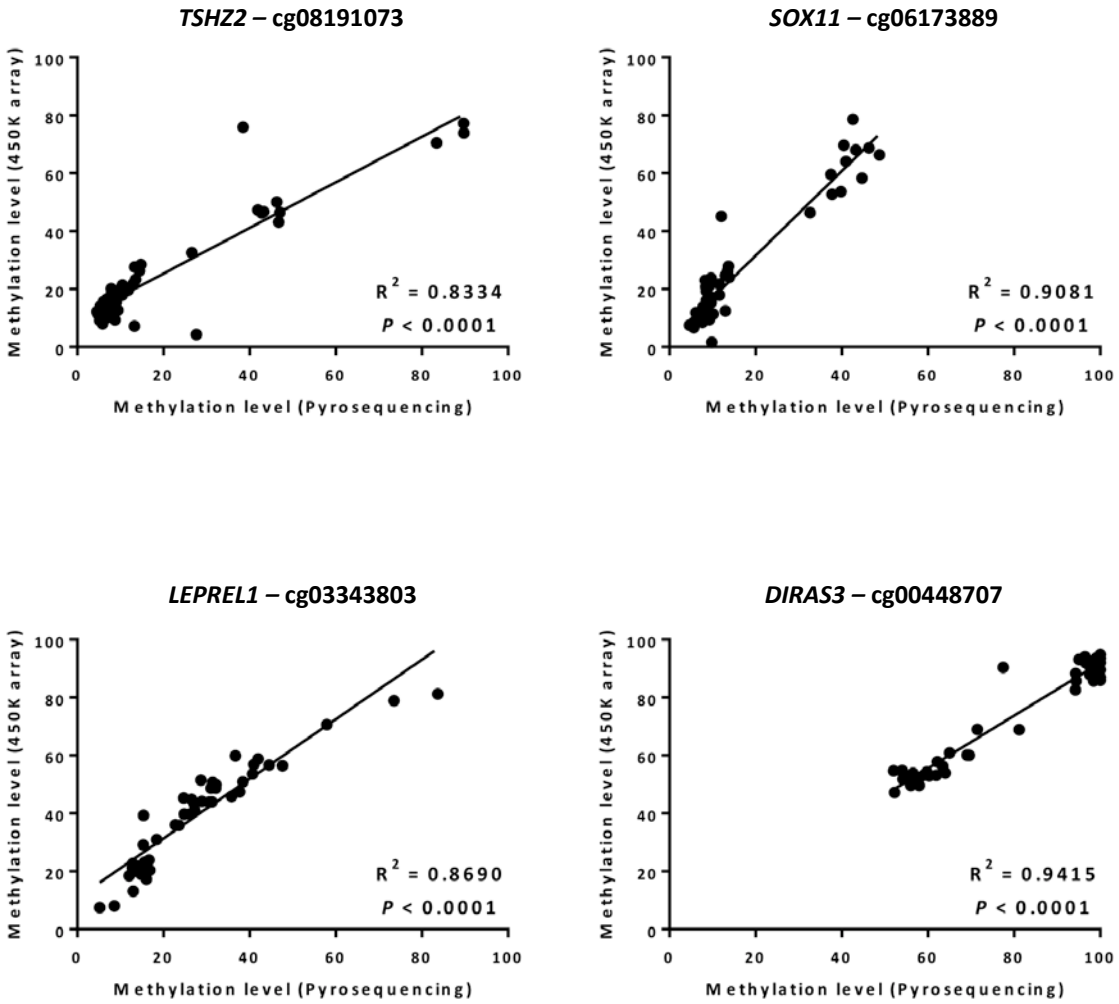
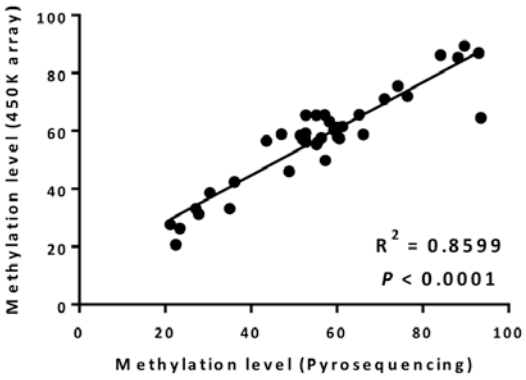


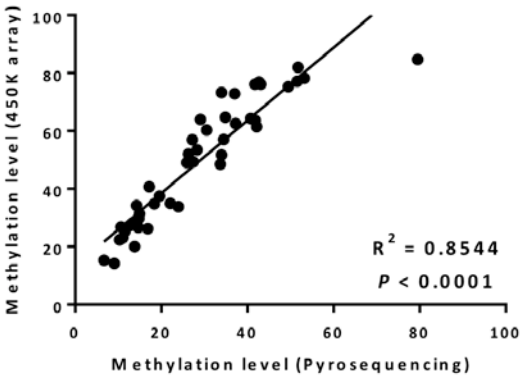
Figure 5.15 Pyrosequencing successfully validated 450K DNA methylation levels for 10 CpGs

The 44 CP-CML patients with matched 450K array and pyrosequencing (Cohort 1) methylation data are plotted. The methylation level from the 450K array data was determined by converting the β-value into a percentage (i.e. a β-value of 0.2 = 20% methylation). Correlation analysis was performed using Pearson's *r* in GraphPad® Prism 6, with a line of best-fit plotted. Significant correlation was observed between the 450K and pyrosequencing data if the *P*-value < 0.05 and the R^2 value approached 1, indicating a linear relationship was present.

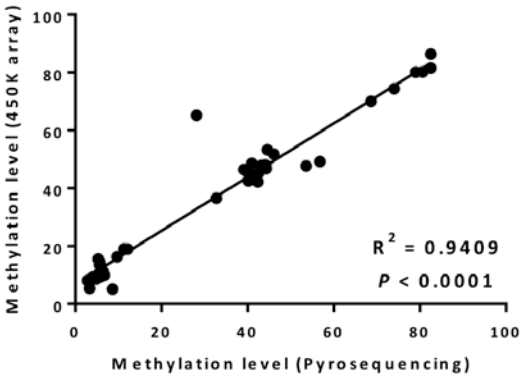
***CLYBL* – cg27053299**



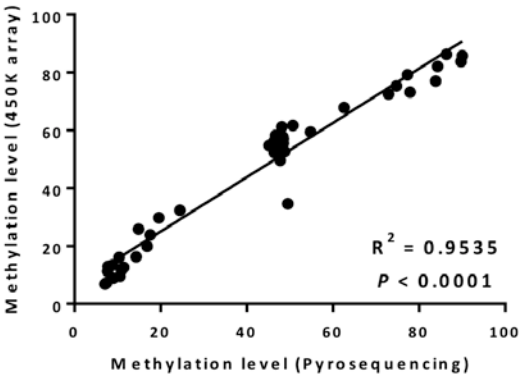
***LOC10034868* – cg21995068**



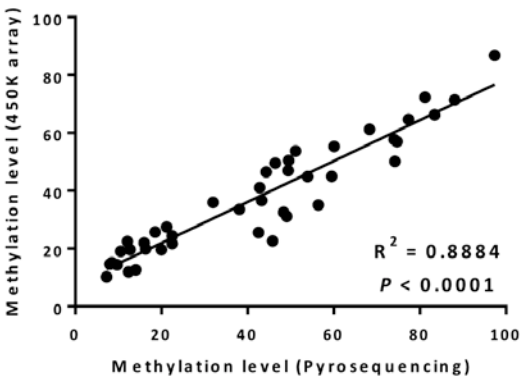
***SORCS2* – cg17394237**



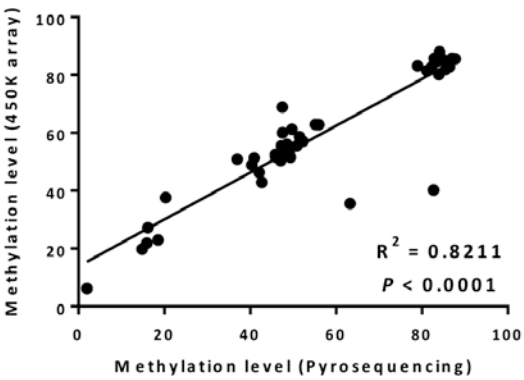
***TENM4* – cg17146918**



***TRIM67* – cg17137424**



***SMARCAL1* – cg10987071**



(cg23825092; TDRD1) no correlation was observed with the pyrosequencing results (**Figure 5.16 [A]**). This result was unexpected and difficult to explain, with no correlation also observed for the flanking CpG, located next to the original CpG of interest analysed by the pyrosequencing assay (**Figure 5.16 [B]**). The samples were confirmed to have not been mislabelled, leaving primer design as the most likely cause for the lack of correlation. However, as the pyrosequencing assay worked correctly for all samples analysed, this suggests that the pyrosequencing primers were unlikely to be an issue, indicating that perhaps the results from the 450K array were misleading for this individual CpG. Due to this unexplainable lack of correlation, this CpG was excluded from further analysis. Taken together, this data demonstrates that pyrosequencing was able to validate the differential methylation observed from the 450K analysis for 10 out of 11 CpGs.

5.2.7 Replication of the OA methylation classifier using pyrosequencing data

Using a methylation classifier based on data from the 450K array may not be easily reproducible in other laboratories or in the diagnostic setting due to the cost of performing the 450K array and the time involved in data analysis. Therefore, the OA classifier was replicated using the 10 CpG data from the pyrosequencing analysis. PAMR was applied to this 10 CpG list (no cg23825092), with the 46 patient dataset from the pyrosequencing analysis (Cohort 1) initially used as both the training and test sets (internal validation). A classifier using all 10 CpGs was developed (threshold = 0.0), with an excellent overall accuracy of 93% (sensitivity = 100%, specificity = 84%; **Figure 5.17**), as 3 of the 18 other OA patients, and none of the patients with very low OA, were misclassified. Therefore, this classifier had approximately equivalent predictive power (93%) compared to the previously developed classifiers using the 450K data (94%) and was used for further analyses.

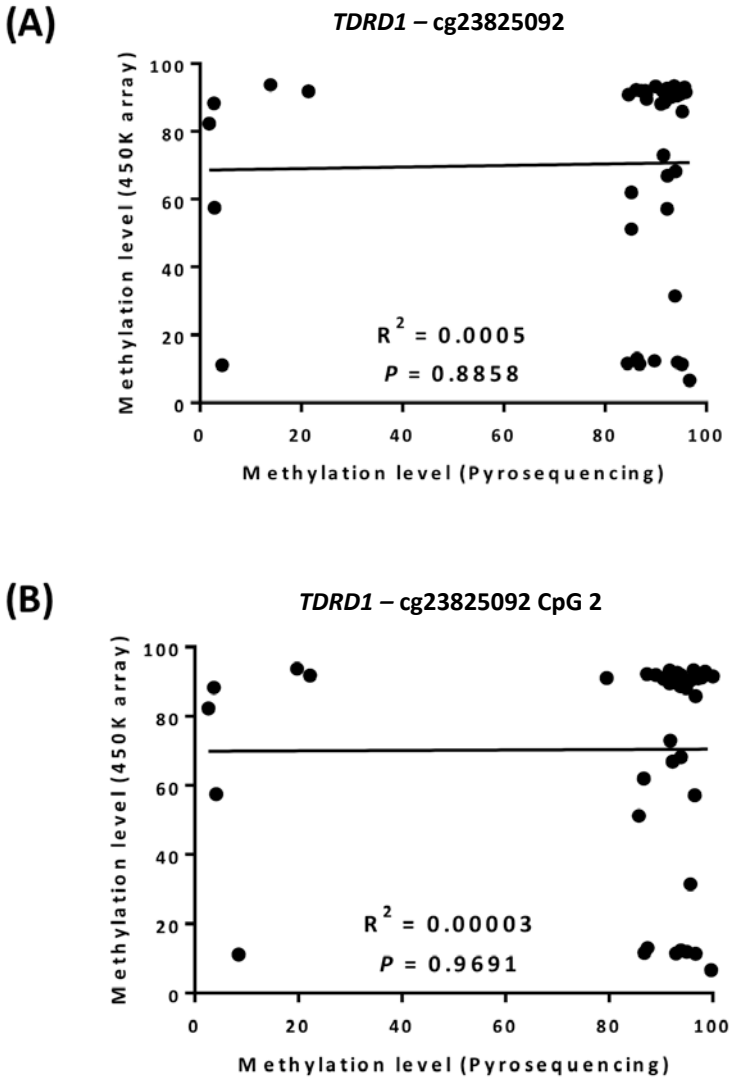


Figure 5.16 Pyrosequencing was unable to validate the 450K methylation results for cg2385092

The 44 CP-CML patients with matched 450K array and pyrosequencing (Cohort 1) methylation data are plotted. No correlation was observed for the CpG (cg23825092) of interest **(A)** or the second flanking CpG analysed by the pyrosequencing assay **(B)**. The methylation level from the 450K array data was determined by converting the β-value into a percentage (i.e. a β-value of 0.2 = 20% methylation). Correlation analysis was performed using Pearson's *r* in GraphPad® Prism 6, with a line of best-fit plotted. Significant correlation was observed between the 450K and pyrosequencing data if the *P*-value < 0.05 and the R^2 value approached 1, indicating a linear relationship was present.

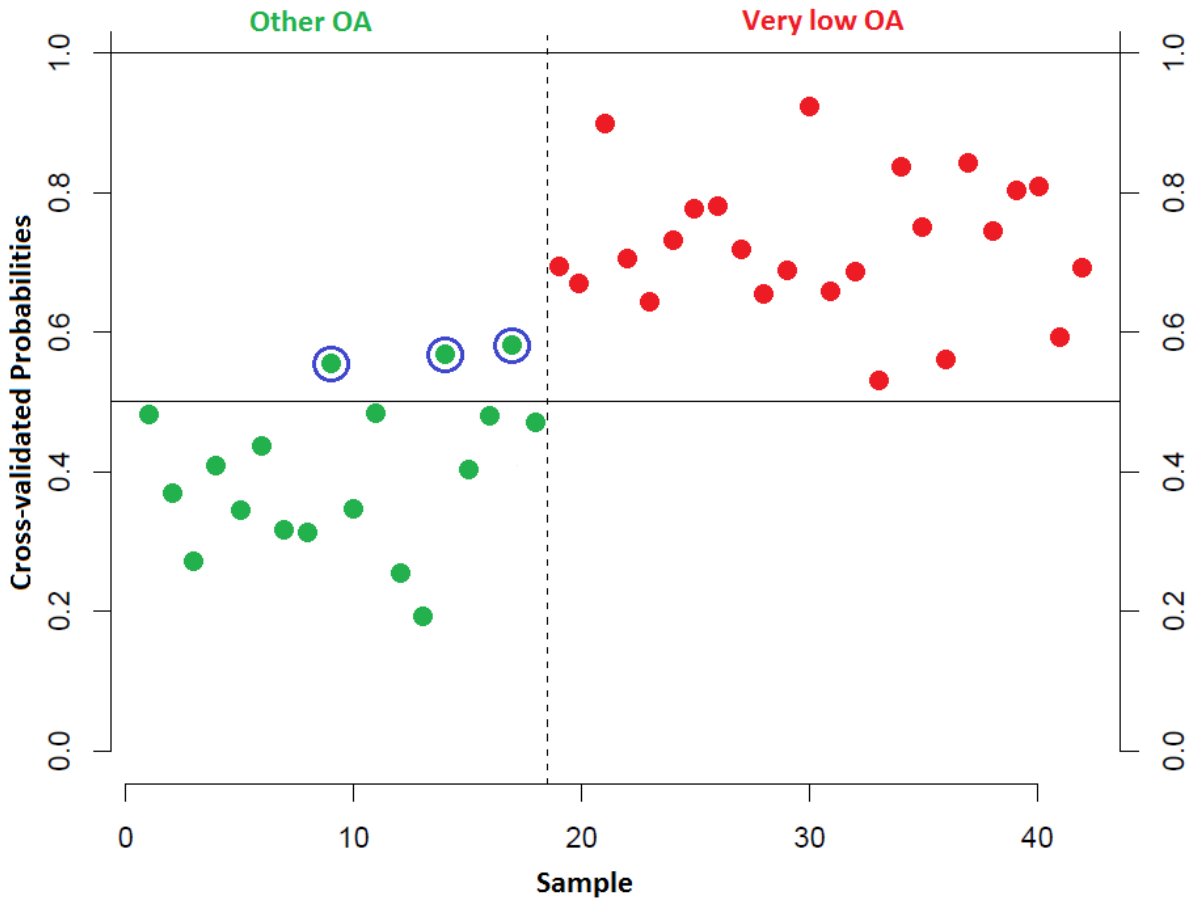


Figure 5.17 The 10-CpG PAMR OA classifier developed using the 10 CpG dataset from the pyrosequencing analysis is 93% accurate

Results of the 10-CpG classifier developed by PAMR using the 10 CpG pyrosequencing methylation dataset (n = 46 training set), tested against the same 46 patients. A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). Three patients were misclassified (**blue** circle; 3 other OA patient), providing an overall accuracy of 93%.

5.2.8 Validation of the OA pyrosequencing classifier using an independent patient cohort

In order to validate the 10-CpG classifier developed by PAMR using the pyrosequencing data, the 46 independent (not analysed by 450K array) CP-CML patients analysed in Cohort 2 of the pyrosequencing analysis were used as an independent test set against the original 46 patient training set. PAMR was performed as before using the specific 10-CpG classifier originally developed and the new 46 patient test set. Unfortunately, the classifier was unable to be validated with this new, independent test set, as 14/46 patients were misclassified, providing an overall accuracy of only 69.6% (**Figure 5.18**). In particular, this classifier was very poor at predicting the "other" OA patients (5/18, 27.8%), when compared to the very low OA patients (23/24, 95.8%). The classifier may have failed to validate for a number of reasons. Firstly, as the pyrosequencing analysis was performed in two separate cohorts (cohort 1 – training, cohort 2 – test), there is a risk of batch effects being present between the results from each cohort. However, batch effect was assessed using multi-dimensional scaling plots and no significant differences were demonstrated between cohort 1 and cohort 2 (**Figure 5.19**). As batch effect was unlikely to be the cause of the classification failure, a closer investigation of the methylation profiles of the patients in the validation set (cohort 2) was then performed, as discussed below.

5.2.9 Pyrosequencing on an independent 46 patient cohort failed to replicate the differential methylation patterns observed from the 450K analysis

Analysis of the differential methylation patterns observed from the pyrosequencing data of patients from cohort 2 identified an interesting and possibly significant result; no differential methylation was observed between CP-CML patients with very low OA, compared to the rest (**Figure 5.20**). This was in stark contrast to the original observations for these 11 CpGs, from both the 450K array analysis (**Section 5.2.5**) and the matched pyrosequencing analysis

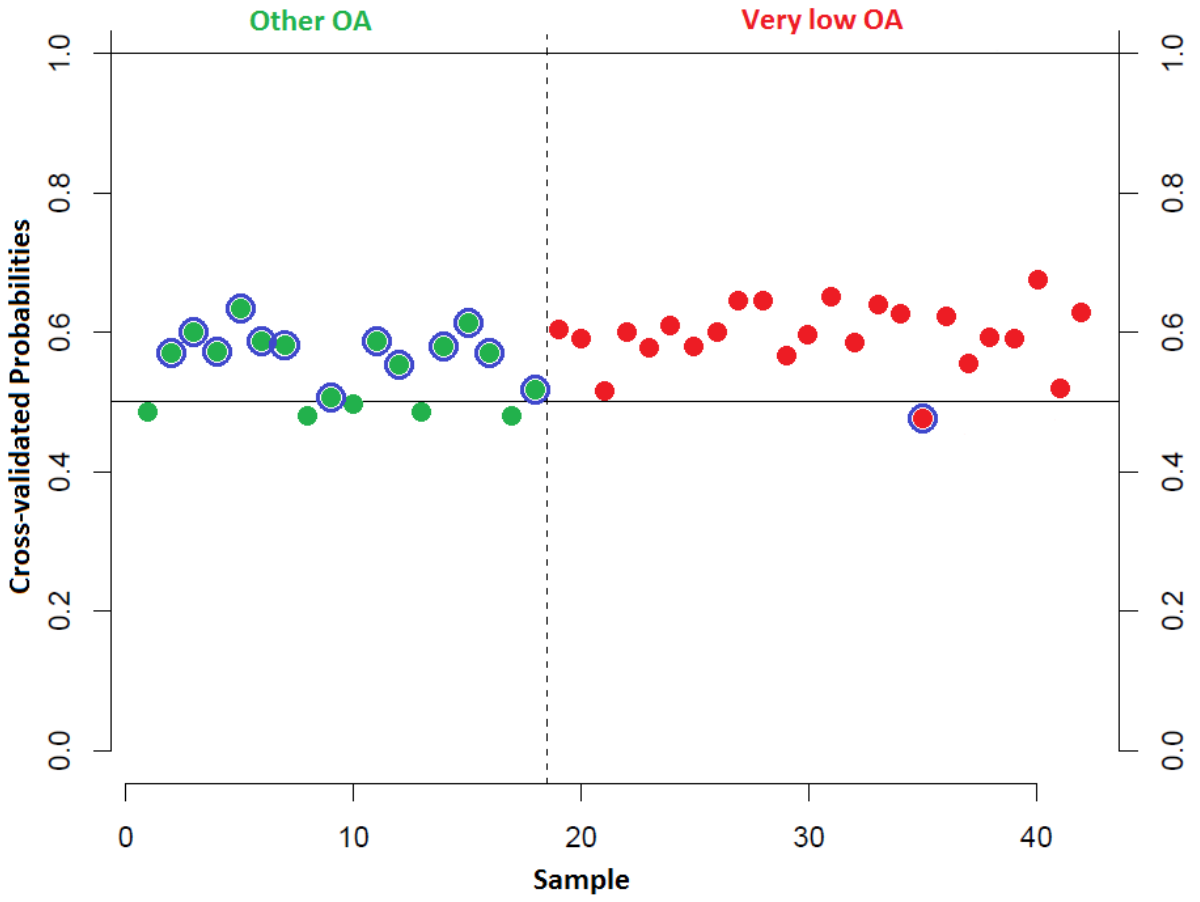


Figure 5.18 Validation of the 10-CpG PAMR OA classifier developed using the pyrosequencing data with an independent 46 patient test set

Results of the validation of the 10-CpG OA classifier developed by PAMR using the pyrosequencing methylation dataset (n = 46 training set), tested against an independent 46 patient test set (pyrosequencing Cohort 2). A 10-fold cross-validated probability close to 1 should indicate a very low OA patient (**red**; clustered on the right of the graph), while a cross-validated probability close to 0 should indicate an "other" OA patient (**green**; clustered on the left of the graph). Fourteen patients were misclassified (**blue** circle; 13 other OA patients and 1 very low OA patient).

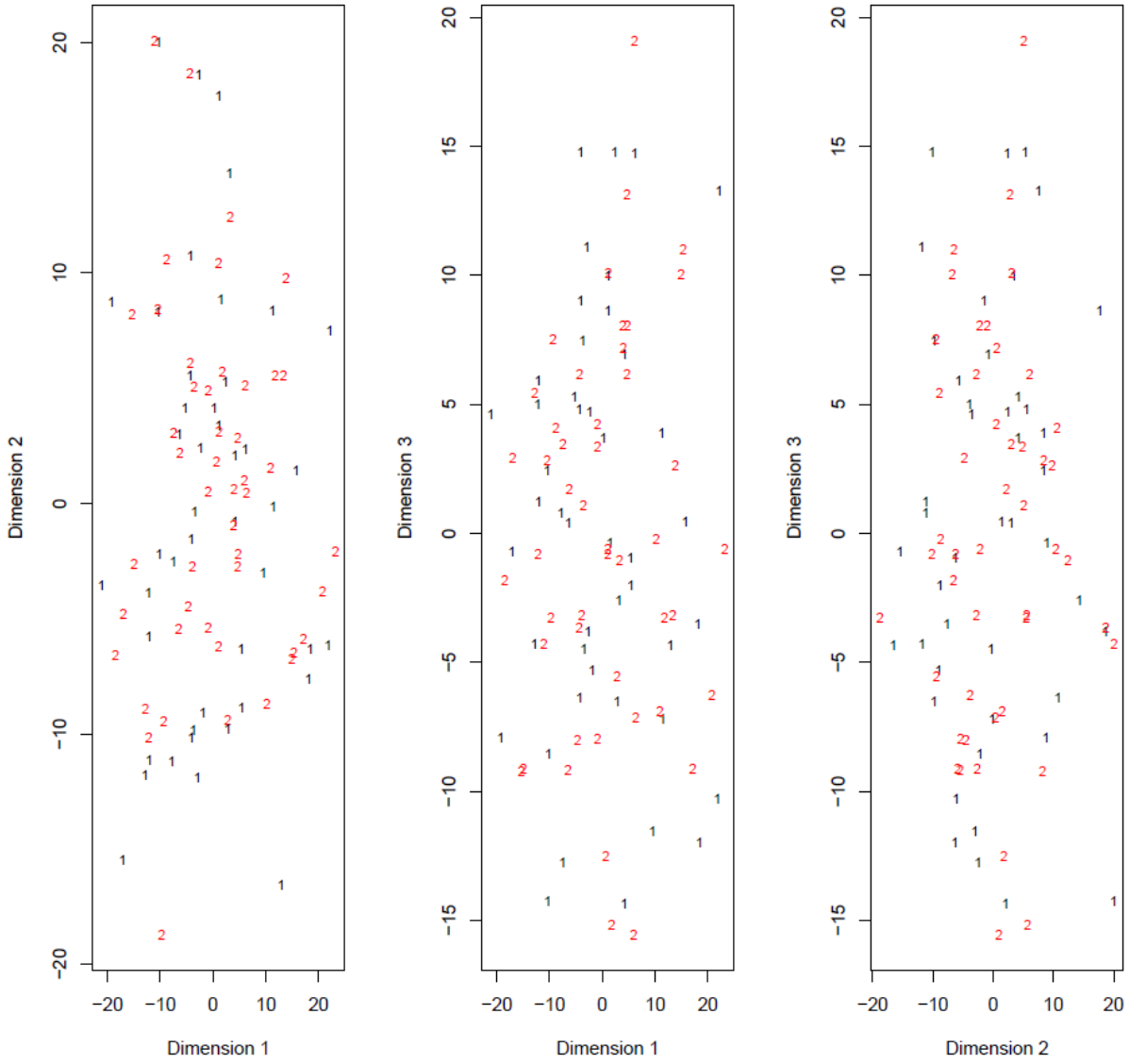


Figure 5.19 Multi-dimensional scaling plots demonstrate that no batch effects were present in the pyrosequencing data

Multi-dimensional scaling (MDS) plots allow the level of similarity between individual cases in a dataset to be compared. The pyrosequencing data from cohort 1 (n = 46 patients, **1**) was compared to the data from cohort 2 (n = 46 patients, **2**). No differences were observed between both cohorts over all 3 arbitrary dimensions. MDS plots were performed in the statistical environment *R*, v3.0.1, using the *LIMMA* package.

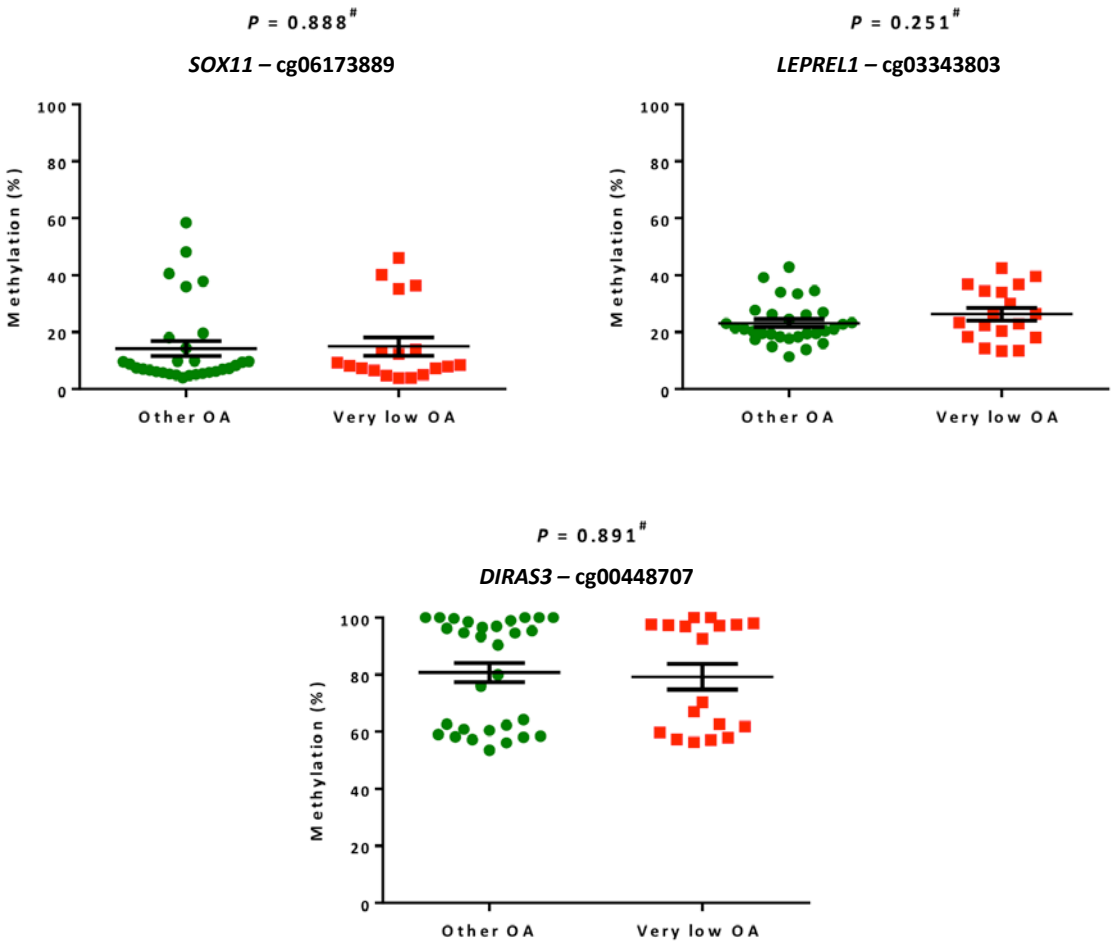
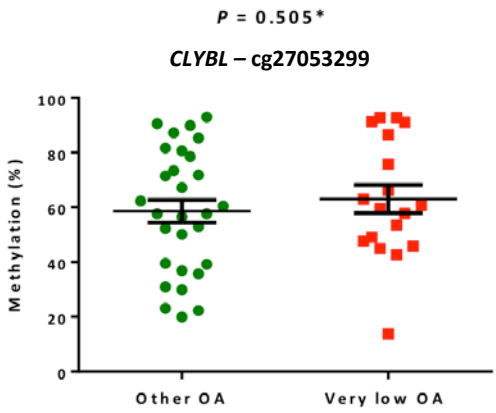
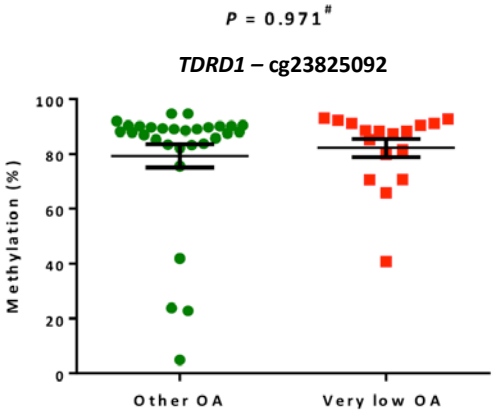
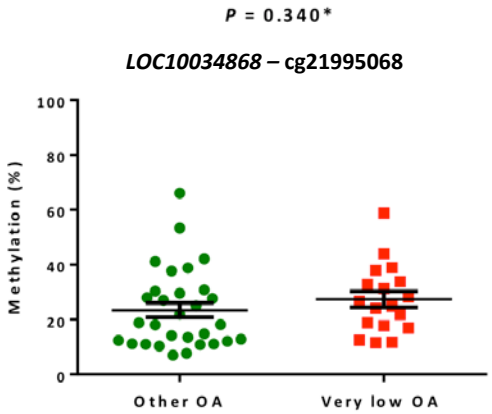


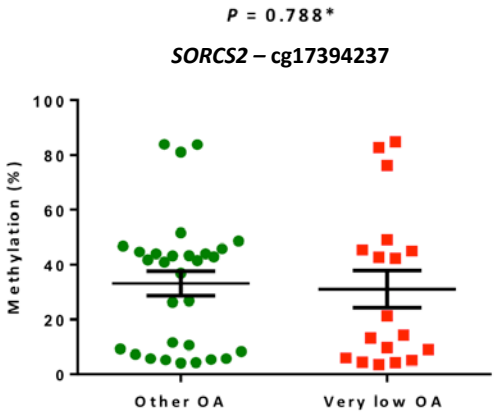
Figure 5.20 Pyrosequencing analysis of an independent patient cohort failed to demonstrate differential methylation

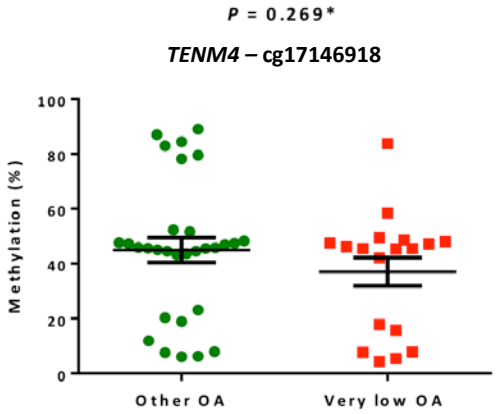
The 46 CP-CML patients analysed by pyrosequencing (Cohort 2) for the 11 CpGs of interest from the 450K array analysis data are plotted. Scatterplots demonstrate the mean methylation level (%) ± SEM. Student's *t*-tests (*) or Mann-Whitney tests (#) were performed as appropriate in GraphPad Prism 6.

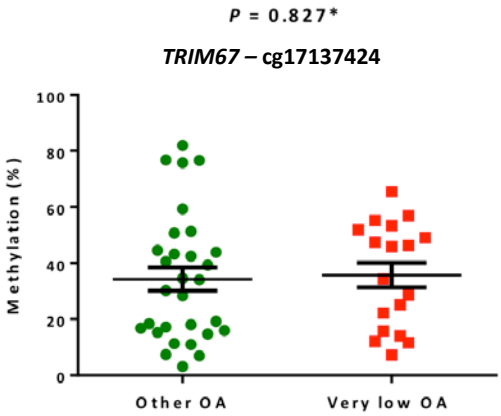

P = 0.505*
CLYBL – cg27053299
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA

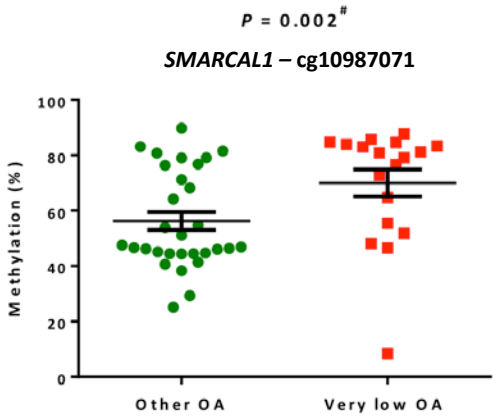

P = 0.971#
TDRD1 – cg23825092
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA

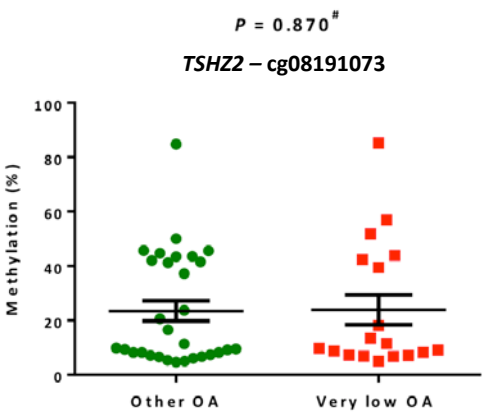

P = 0.340*
LOC10034868 – cg21995068
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA


P = 0.788*
SORCS2 – cg17394237
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA


P = 0.269*
TENM4 – cg17146918
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA


P = 0.827*
TRIM67 – cg17137424
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA


P = 0.002#
SMARCAL1 – cg10987071
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA


P = 0.870#
TSHZ2 – cg08191073
Methylation (%)
100
80
60
40
20
0
Other OA
Very low OA

(**Section 5.2.6**), which demonstrated significant differential methylation between CP-CML patients with very low OA, compared to the rest, for all CpGs excluding cg23825092 (TDRD1; not validated by the pyrosequencing analysis). Further complicating these results, significant differential methylation was observed for cg10987071 (SMARCAL1; *P* = 0.002, **Figure 5.20**), with increased (hyper) methylation observed in patients with very low OA, compared to the rest. However, the results from the original 450K array analysis, which were validated by pyrosequencing in patients from cohort 1, demonstrated the opposite, with decreased (hypo) methylation observed in patients with very low OA, compared to the rest for this CpG. This result is difficult to explain and may possibly relate to biological differences within each patient cohort.

Taken together, this data provides evidence why the pyrosequencing OA classifier failed to be validated when the independent patient test set (cohort 2) was applied to the classification algorithm (**Section 5.2.8**). However, the exact reason why the results from the randomly selected patients in cohort 2 are so vastly different from patient cohort 1, are not easily explained. As previously stated, the patients in cohort 1 were matched to the patients analysed on the 450K arrays; hence, it was expected that the results obtained by pyrosequencing would correlate well with the 450K results, which was the case. Yet, as cohort 2 contained independent patients, this assumption could not be made and the results indicate that perhaps the patients in either cohort 1 or cohort 2 alone are not an accurate representation of the general CP-CML population. Conversely, batch effect analysis demonstrated that there were no obvious differences between the pyrosequencing methylation results from cohort 1 and cohort 2, adding further complexity to the interpretation of these results.

5.2.10 Re-analysis of the global 450K DNA methylation profiling from CP-CML patients with very low OCT-1 activity compared to the rest identified significant differential methylation

Due to the lack of correlation observed when pyrosequencing was performed on the independent patient cohort (cohort 2), as described above, the original 450K DNA methylation array data was reviewed to determine if the optimal analysis pipeline was used. MDS plots demonstrated that patients grouped into two distinct clusters based on the 450K methylation patterns from each patient (**Figure 5.21**). Further investigation identified that these clusters were directly related to the gender of each patient, with female patients clustering distinctly from male patients (**Figure 5.22**). This data indicated that although there was no extreme bias between the number of male and female patients in both the very low OA and other OA groups (**Table 5.2**), the distinct methylation profiles observed between males and females may be influencing the overall results. Therefore, it was determined to re-analyse the 450K data, this time excluding all CpGs mapped to chromosomes X and Y to remove any gender bias. However, before this was performed, a further 8 CP-CML patients were analysed by the 450K array to increase the overall patient number to 63 (not including patients who failed QC; n = 36 very low OA, n = 27 other OA).

In general, datasets cannot be combined without adjusting for batch effects. To ensure no batch effects confounded the methylation results obtained between the first batch of 49 patients and the second batch of 8 patients analysed on the 450K arrays the *ComBat* (Combining Batches) algorithm, developed by Johnson *et al*.[236] was applied. *ComBat* is available for use in the Surrogate Variable Analysis (SVA) package,[237] and was used to directly remove known batch effects from the datasets (**Section 2.6.11**). Once performed, differential methylation analysis was performed using *minfi* as previously described (**Section 2.6.10**), allowing the genome-wide DNA methylation profiles obtained from the TWCs of 63

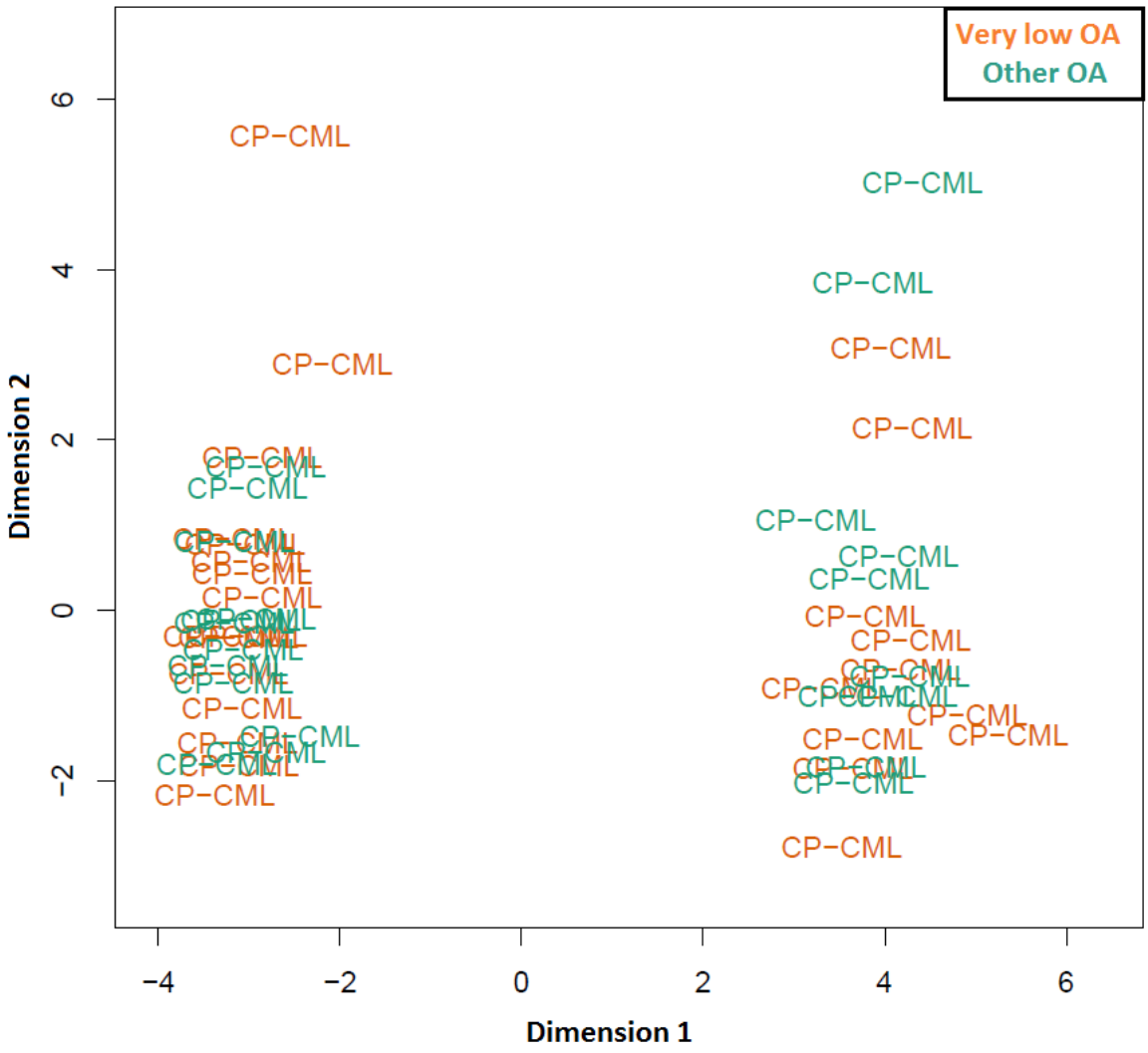


Figure 5.21 MDS plot demonstrating methylation differences between patients with very low OA compared to the rest

The top 1000 variable CpGs between CP-CML patients with very low OA, compared to the rest, were used to generate an MDS plot. Two distinct clusters are visible on the plot; however, OA grouping (either **very low OA** or **other OA**) was unable to distinguish these clusters. The MDS plot was generated using the *minfi* package in *R* v3.0.1.

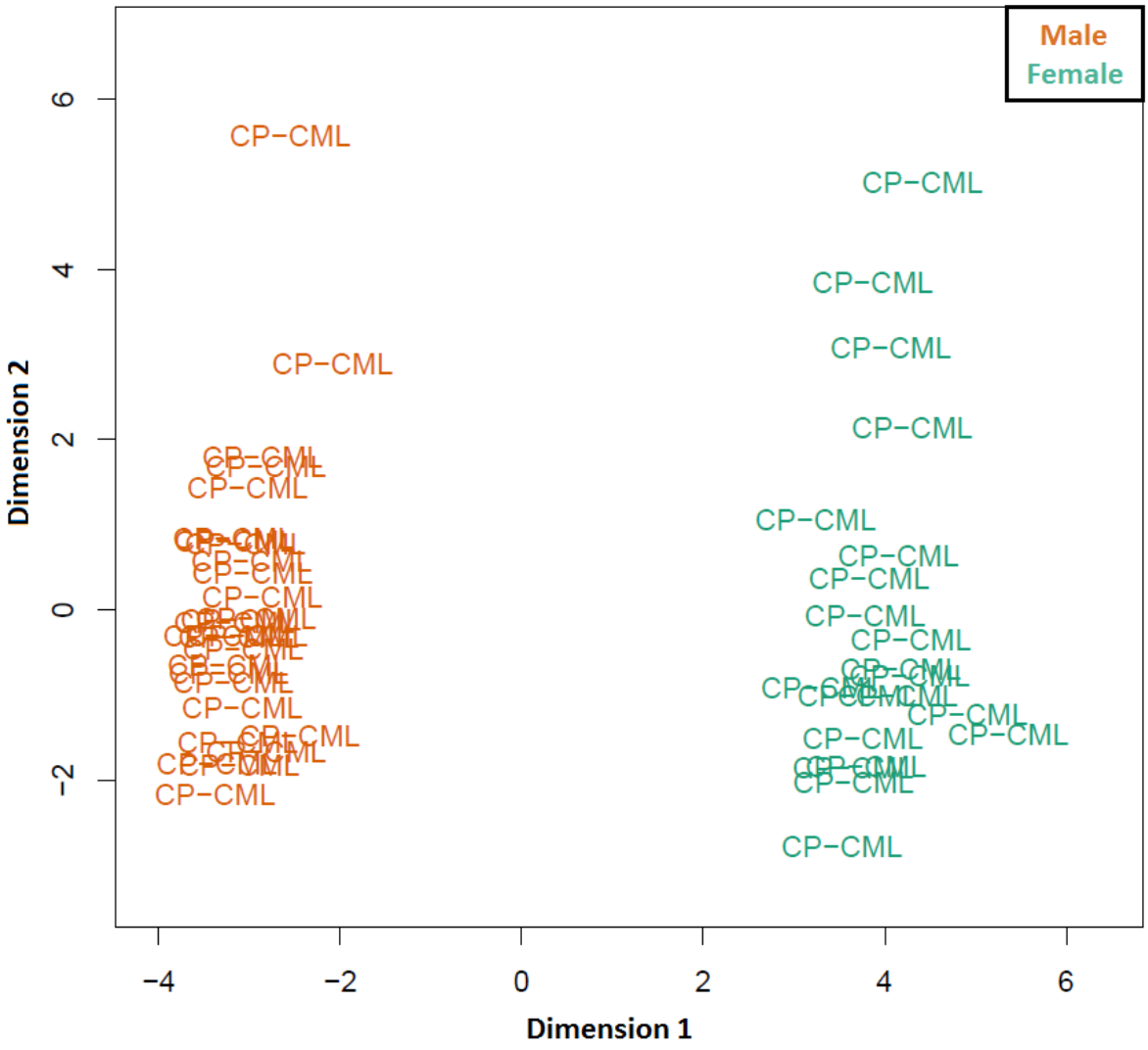


Figure 5.22 MDS plot demonstrating methylation differences between patients with very low OA compared to the rest are influenced by gender

The top 1000 variable CpGs between CP-CML patients with very low OA, compared to the rest, were used to generate an MDS plot. Two distinct clusters are visible on the plot, one for **male** patients (left) and the second for **female** patients (right). The MDS plot was generated using the *minfi* package in *R* v3.0.1.

CP-CML patients with very low OA (n = 36) and other OA (n = 27) to be compared. **Selection criteria 2A** (as described in **Section 2.6.10.4**) was initially used to identify differentially methylated CpGs, which included removing all CpGs located on chromosomes X and Y, with a detection *P*-value < 0.01, and excluding those containing a SNP within 10 bp of the CpG locus. An MDS plot was generated using the top 1000 most variable CpGs and demonstrated that the gender bias was now removed (**Figure 5.23**). Using a significance cut-off of FDR *P*-value < 0.05, 99,199 differentially methylated CpGs were identified, of which 51,771 were hypermethylated and 47,428 were hypomethylated in patients with very low OA, compared to the rest. A spatial view of these sites mapped across the genome (chromosomes 1 – 22) revealed no bias in the distribution of these CpGs across all chromosomes (**Figure 5.24**).

Using a more stringent selection criterion (**Selection criteria 2B**, as described in **Section 2.6.10.5**), a total of 1,825 CpGs were identified as differentially methylated in patients with very low OA, compared to the rest. Of these CpGs, 685 were hypermethylated and 1,140 hypomethylated in very low OA, compared to the rest, with these CpGs corresponding to 912 different annotated RefSeq genes. Regional analysis of these CpG sites in relation to CGIs revealed that the largest proportion of CpGs were located in CpG shore regions (**Figure 5.25**). Specifically, 36% were located in CpG shores (657/1,825 sites), 23% in CGI (426/1,825 sites), 6% in CpG shelves (102/1,825 sites) and 35% (640/1,825 sites) in non-annotated regions in relation to the CGI (**Figure 5.25**). Interestingly, the same regional distribution was also observed when only the significant CpGs, either hypomethylated or hypermethylated in patients with very low OA were analysed. Investigation into the position of these differentially methylated CpG sites in relation to the gene structure was performed next. The greatest proportions of significant CpGs were situated in either the gene body or within non-annotated regions (**Figure 5.26**). Once again, this distribution pattern was consistent

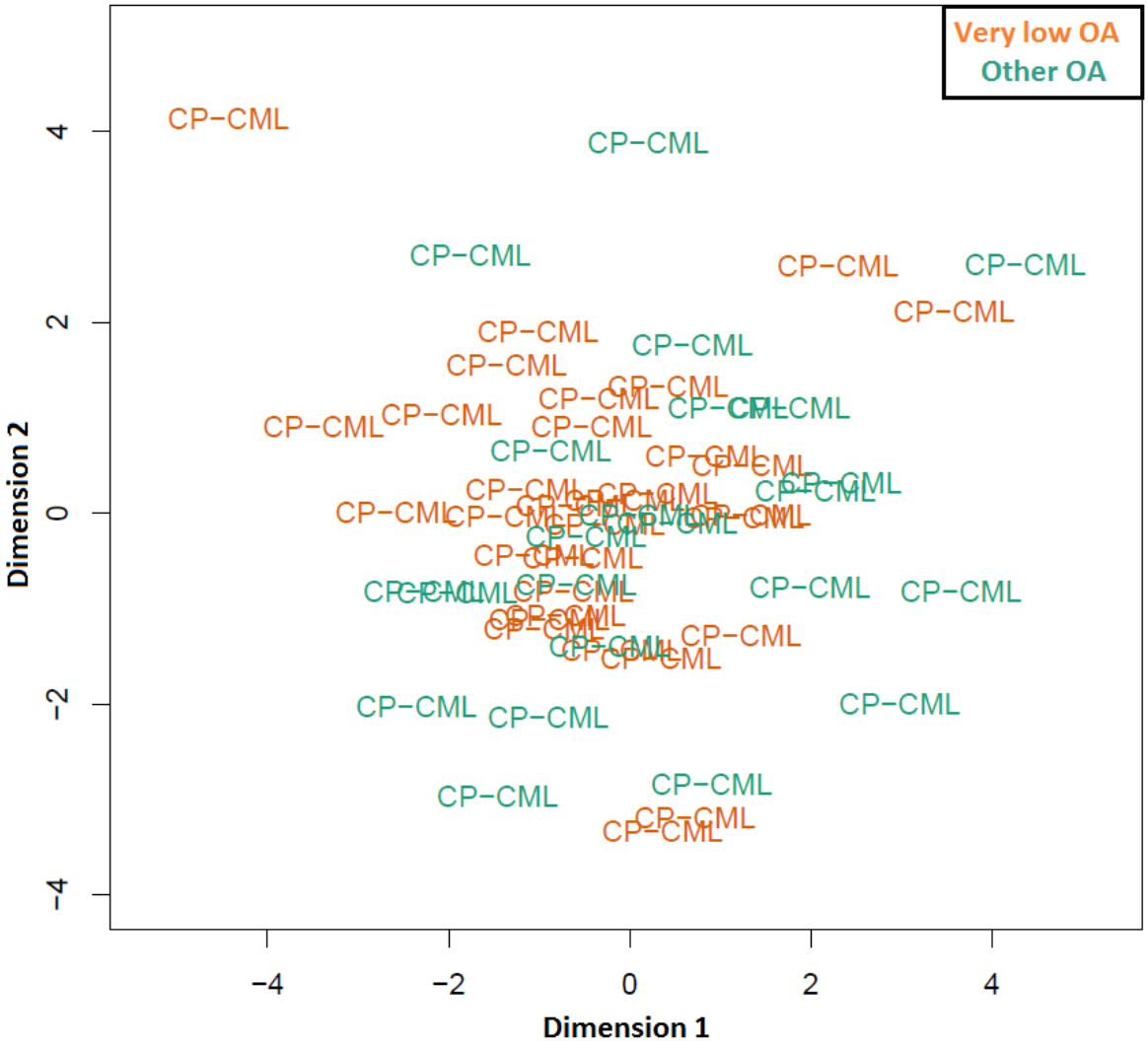
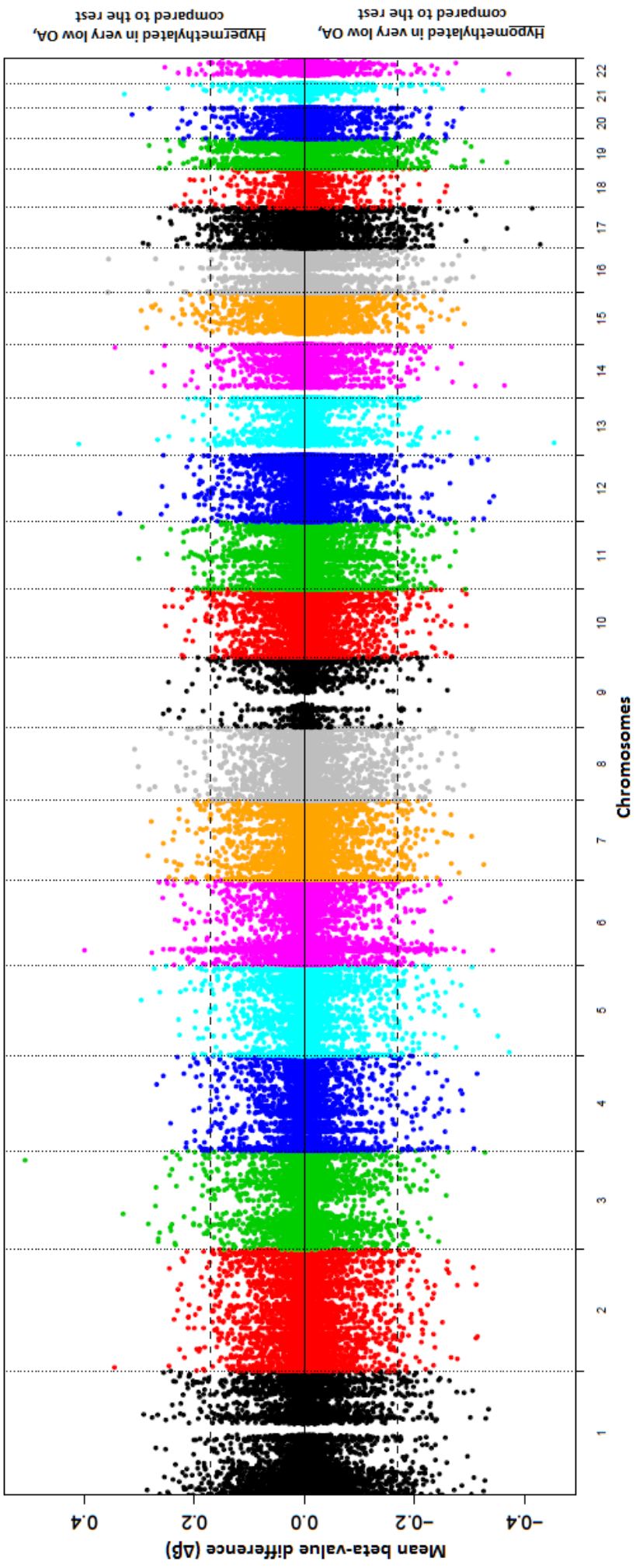


Figure 5.23 MDS plot demonstrating methylation differences between patients with very low OA compared to the rest after *ComBat*

The top 1000 variable CpGs between CP-CML patients with very low OA, compared to the rest, after *ComBat* was performed, were used to generate an MDS plot. The two distinct clusters due to gender are no longer visible on the plot. CP-CML patients are identified as either **very low OA** or **other OA.** The MDS plot was generated using the *minfi* package in *R* v3.0.1.

Figure 5.24 The spatial view of all significant CpGs differentially methylated from the re-analysed very low OA versus the rest DNA methylation data revealed an even distribution across all chromosomes

A manhattan plot was generated using all 99,199 differentially methylated CpGs with a FDR *P*-value < 0.05 and demonstrates an even distribution of the CpGs across all chromosomes. All 22 chromosomes are plotted on the X-axis, with the mean beta-value difference (Δβ) for each individual CpG plotted on the Y-axis. A negative Δβ value indicates a CpG hypomethylated in patients with very low OA, compared to the rest, while a positive Δβ value indicates a CpG hypermethylated in patients with very low OA, compared to the rest. Each CpG is mapped to the exact position within its respective chromosome. Manhattan plots were generated using the *CpGassoc* package,[233] obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/).


Hypomethylated in very low OA,
compared to the rest
Hypermethylated in very low OA,
compared to the rest
Mean beta-value difference (Δβ)
−0.4
−0.2
0.0
0.2
0.4
Chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

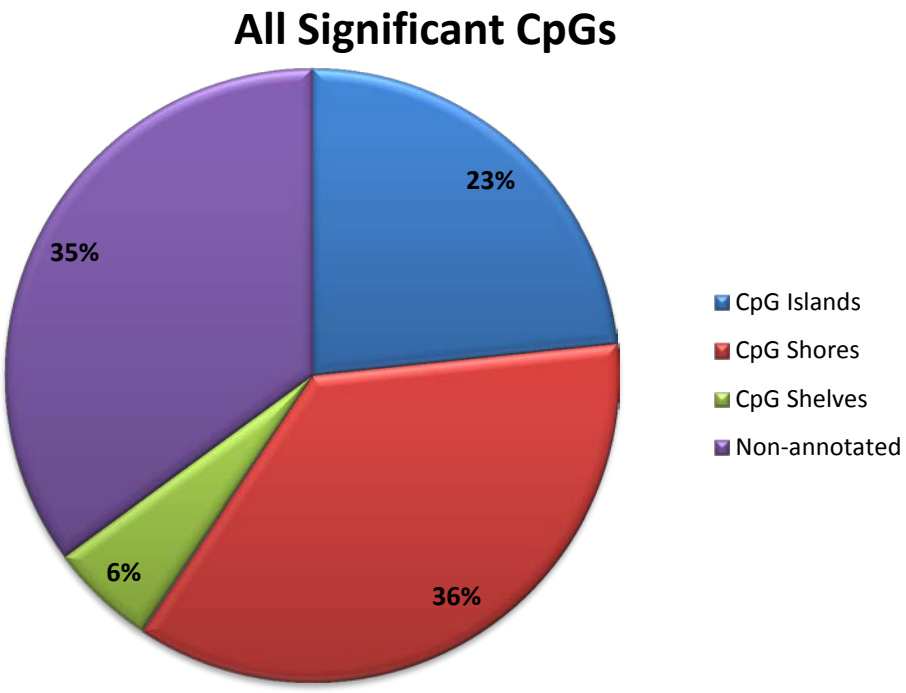


Figure 5.25 The largest proportion of the significant CpG sites from the re-analysed 450K OA methylation analysis were located in CpG shore regions

All differentially methylated CpGs (**Selection criteria 2B**; 1,825 sites) from CP-CML patients with very low OA, compared to the rest, with a known (CpG island, shore or shelf) or unknown (non-annotated) position are demonstrated. The largest proportion of CpGs were located CpG shore regions. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays.

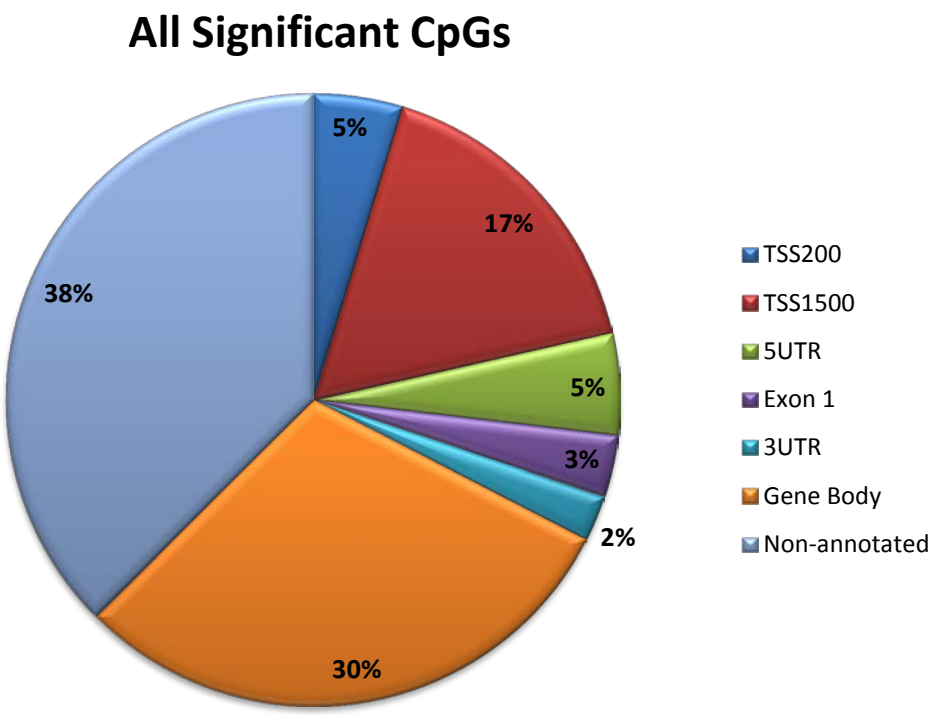


Figure 5.26 The genomic distribution of all highly significant CpG sites from the re-analysed 450K OA methylation analysis revealed the largest proportion were located in non-annotated regions

The positon of all differentially methylated CpGs (**Selection criteria 2B**; 1,825 sites) from CP-CML patients with very low OA, compared to the rest, in relation to the gene structure are demonstrated. The largest proportion of CpGs were located in non-annotated regions. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

for both the CpGs either hypomethylated or hypermethylated in patients with very low OA. Furthermore, the majority (12/14) of the CpGs originally identified as differentially methylated between patients with very low OA compared to the rest (selection criteria 1C), were located in this re-analysed significant CpG list (**Figure 5.27**). Hierarchical clustering using the top 100 differentially methylated CpGs was able to successfully distinguish patients with very low OA from the rest (**Figure 5.28**).

5.2.11 Gene-set enrichment analysis identifies significant enrichment of Polycomb group proteins associated with very low OA patients, compared to the rest

Gene-set enrichment analysis was performed by investigating the Molecular Signatures Database (MSigDB, Broad Institute) using the 912 gene list that corresponded to the 1,825 differentially methylated CpGs identified by the re-analysed ComBat 450K array between CP-CML patients with very low OA, compared the rest. A number of highly significant gene-sets (from the MSigDB) were enriched for this gene list of interest (**Table 5.11**). Included in this list were pathways involved with Polycomb group proteins (EED, SUZ12 and PRC2), and G protein-coupled receptor (GPCR) signalling. Notably, Polycomb group proteins have a known role in stem cells and cancer,[388] with the BMI1 gene (a member of the Polycomb group) demonstrated to be a prognostic marker in CML.[320]

5.2.12 Integrated genome-wide DNA methylation and gene expression analysis reveals significant interplay between genes significantly associated with very low OA, compared to other OA

As DNA methylation can function as a switch, controlling when the expression of certain genes is activated or silenced, the extent to which DNA methylation affects gene expression in CP-CML patients was examined. The gene expression profiles of 45 CP-CML patients with very low (n = 24) and other OA (n = 21) matched to the 450K analysis, were compared by

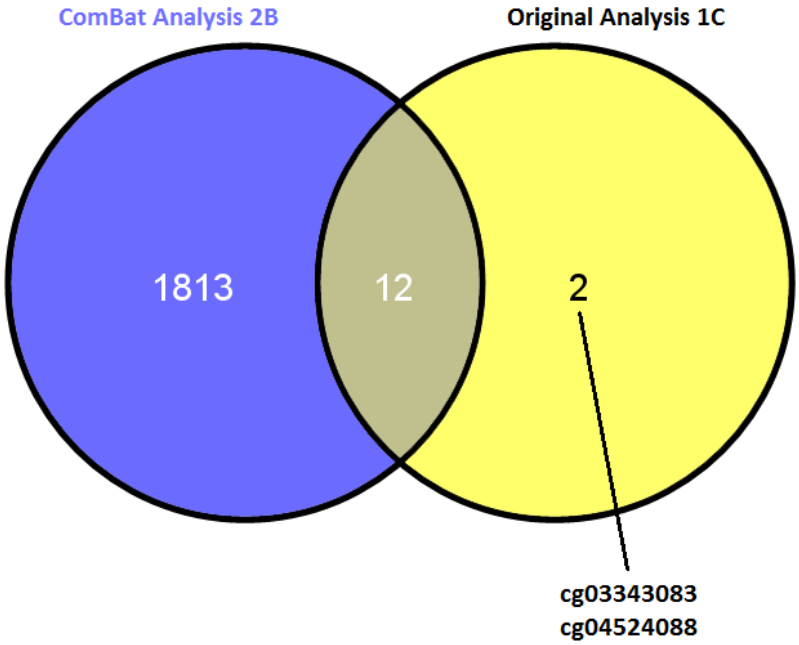


Figure 5.27 Significant overlap was observed between the significant CpGs identified from the original and ComBat analyses for patients with very low versus other OA

Venn diagram demonstrating the overlap between the significant CpGs identified from the original very low versus other OA analysis with selection criteria 1C (termed "Original Analysis 1C"), compared to the re-analysed results with selection criteria 2B (termed "ComBat Analysis 2B"). Only 2 CpGs (cg03343083, LEPREL1; and cg04524088, MIR129-1) did not overlap between the original analysis and the results from the re-analysed dataset. The Venn diagram was developed using Venny (http://bioinfogp.cnb.csic.es/tools/venny/).[313]

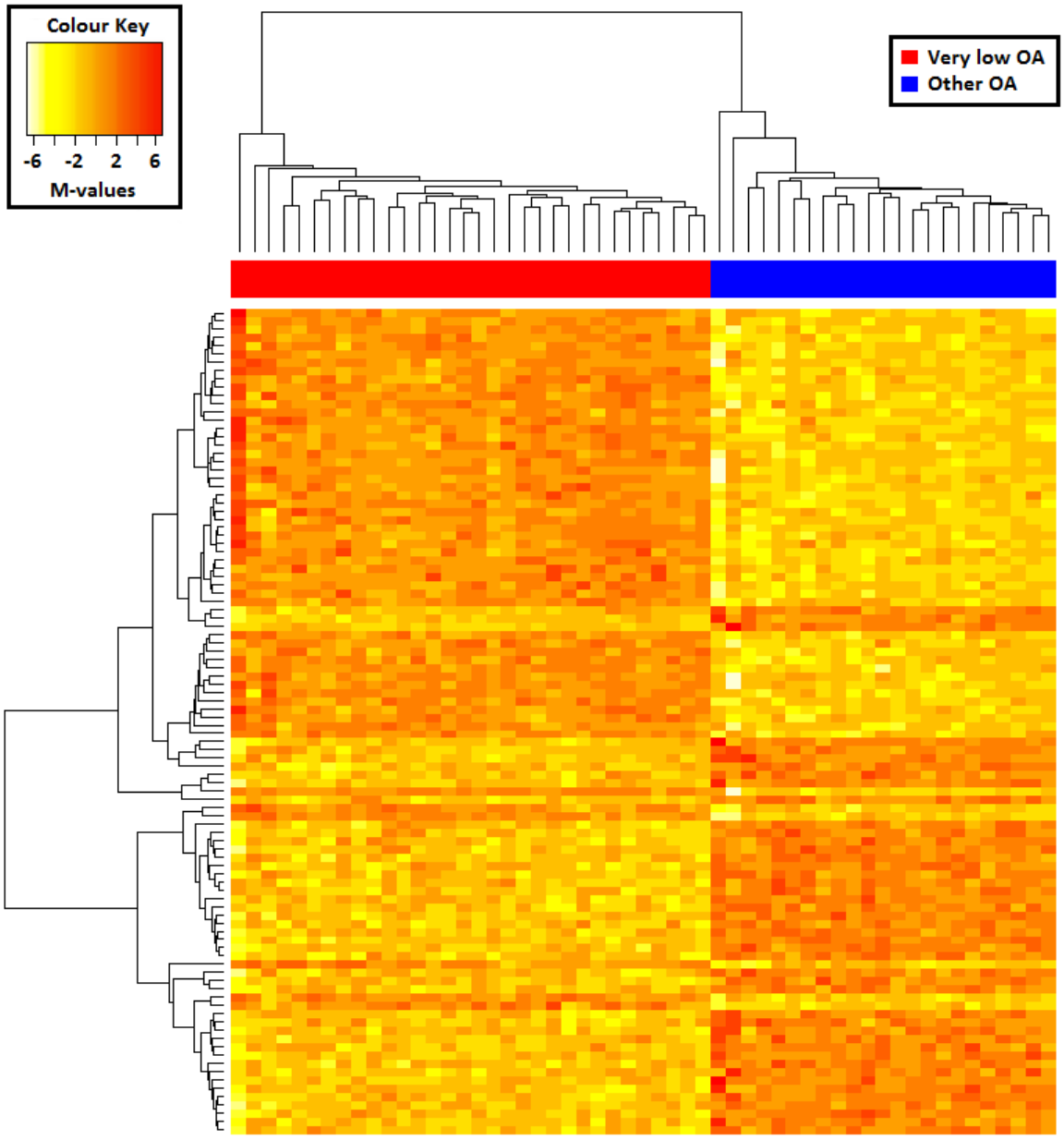


Figure 5.28 Hierarchical clustering using the top 100 differentially methylated CpGs distinguishes patients with very low OA from other OA

The top 100 CpGs differentiating patients with very low OA from those with other OA were used to perform unsupervised hierarchical clustering. Each column represents one sample, and each row represents one CpG. M-values were used to perform the clustering.[203] Unsupervised hierarchical clustering (Euclidean, complete linkage) was performed using the heatmap function from the *Stats* package in *R*, v 3.0.1.

Table 5.11 Gene-sets enriched in very low OA patients compared to the rest by MSigDB enrichment analysis

	Gene-set Size*	*P*-value	FDR *q*-value
Polycomb group genes			
Benporath – EED targets	97	< 0.0001	< 0.0001
Benporath – SUZ12 targets	100	< 0.0001	< 0.0001
Benporath – PRC2 targets	72	< 0.0001	< 0.0001
GPCR signalling			
Reactome – GPCR ligand binding	28	< 0.0001	< 0.0001
Reactome – Signalling by GPCR	44	< 0.0001	< 0.0001

Abbreviations: FDR – false discovery rate. *Gene-set size refers to the number of genes overlapping between the gene-sets from the MSigDB with the gene list corresponding to the significant CpGs from the 450K DNA methylation analysis between very low versus other OA.

microarray technology using the Illumina HumanHT-12 v4 Expression BeadChip. Analysis was performed using the LIMMA package in the statistical program *R* (**Section 2.6.12**). Using a significance cut-off of *P*-value < 0.05, 1,112 differentially expressed genes were identified between CP-CML patients with very low OA, compared to the rest. Interestingly, hierarchical clustering using the top 100 differentially expressed genes was unable to clearly separate the patients classified as very low OA from the rest (**Figure 5.29**), suggesting that in this cohort of patients, substantial biological gene expression differences were not very exclusive for patients with very low OA compared to the rest.

Next, the 450K DNA methylation and gene expression data was compared. To successfully integrate the differentially methylated gene list with the gene expression data, the methylation probe level data was summarized to gene level based on Entrez gene IDs, and merged with gene expression data based on common Entrez gene IDs. Of the total 912 differentially methylated genes (corresponding to the 1,825 differentially methylated CpGs [**Section 5.2.10**]), 598 genes were present on the gene expression arrays and when multiple CpGs or probes for particular genes were considered, a total of 997 matched data points were obtained. The majority of these 997 data points (961/997, 96%) did not show any differential gene expression. However, a set of 5 genes (1 in duplicate), which were hypermethylated in patients with very low OA, were also associated with downregulated gene expression (**Red** box, **Figure 5.30**; **Table 5.12**), synonymous with gene silencing. Interestingly, all 5 genes were hypermethylated at gene body CpG locations (**Table 5.12**). Furthermore, an 11 gene-set hypomethylated in patients with very low OA were associated with up-regulated expression of their corresponding genes (**Orange** box, **Figure 5.30**), with 5 of these genes hypomethylated at promoter CpG sites (**Table 5.13**), synonymous with gene activation. A 6 gene-set hypermethylated in patients with very low OA, but associated with

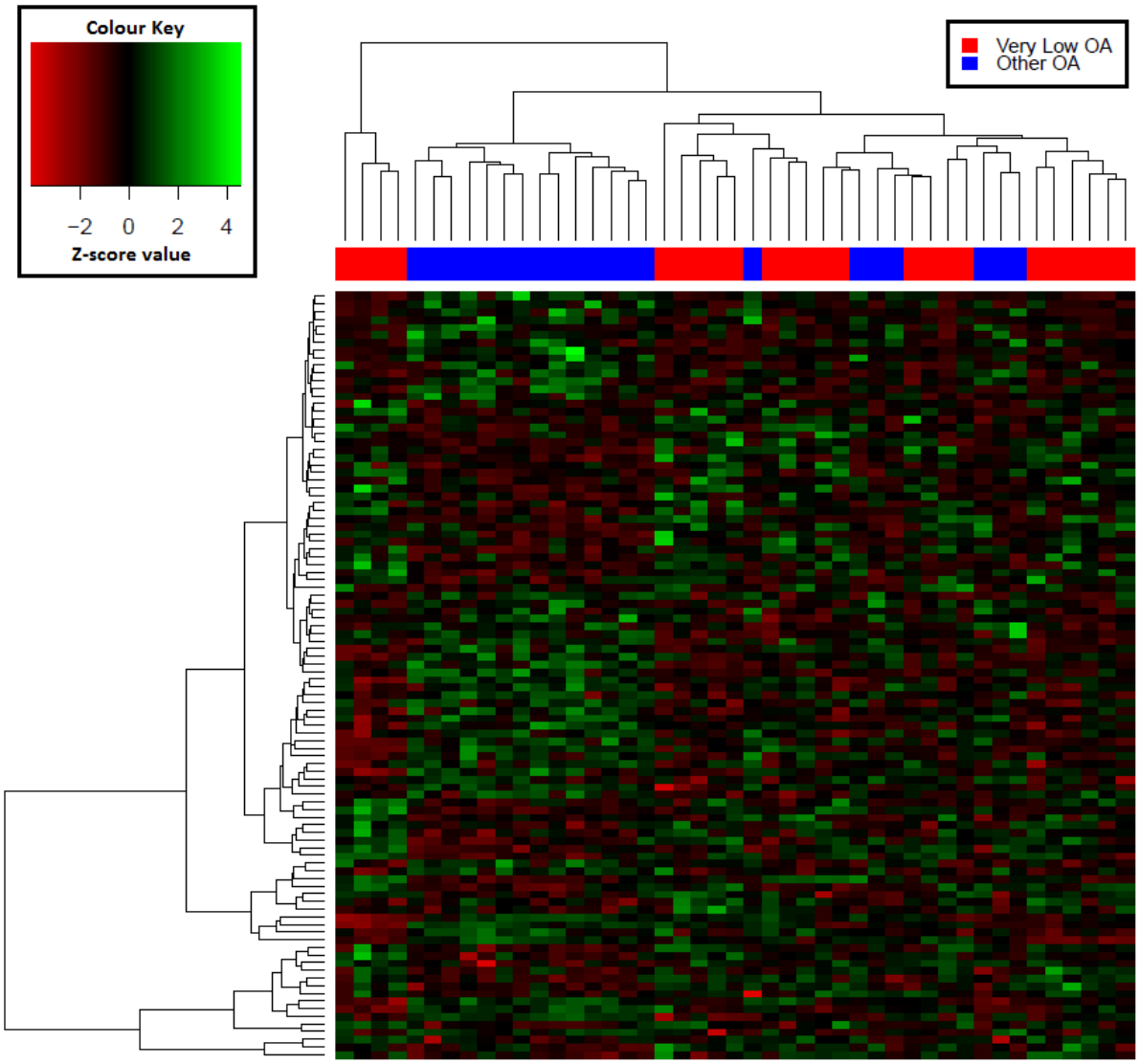


Figure 5.29 Hierarchical clustering using the top 100 differentially expressed genes was unable to distinguish patients with very low OA from other OA

The top 100 genes differentiating patients with very low and other OA were used to perform unsupervised hierarchical clustering. Each column represents one sample, and each row represents one gene. Z-score values (transformation of the expression values in relation to the mean) were used to perform the clustering. Unsupervised hierarchical clustering (Euclidean, complete linkage) was performed using the heatmap function from the *Stats* package in *R*, v 3.0.1.

Figure 5.30 Correlation between genes differentially methylated and differently expressed in patients with very low OA compared to other OA

The relationship between significant DNA methylation and gene expression in CP-CML patients with very low OA compared to other OA, at diagnosis. The scatterplot demonstrates the change in DNA methylation versus the change in gene expression. Both the DNA methylation and gene expression values are adjusted in relation to the fold change difference between very low and other OA, where a positive methylation value indicates hypermethylation in very low OA, while a negative methylation value indicates hypomethylation in very low OA, compared to other OA. The same is also true for the gene expression values, where a positive value indicates upgregulation in very low OA, while a negative value indicates downregulation in very low OA, compared to other OA. Gates are set at the significance level for DNA methylation (FDR *P*-value < 0.05) and gene expression (*P*-value < 0.05). The **Red** box indicates genes hypermethylated and downregulated in very low OA; the **Orange** box indicates genes hypomethylated and upregulated in very low OA; the **Blue** box indicates genes hypomethylated and downregulated in very low OA; and the **Green** box indicates genes hypermethylated and upregulated in very low OA, compared to other OA.

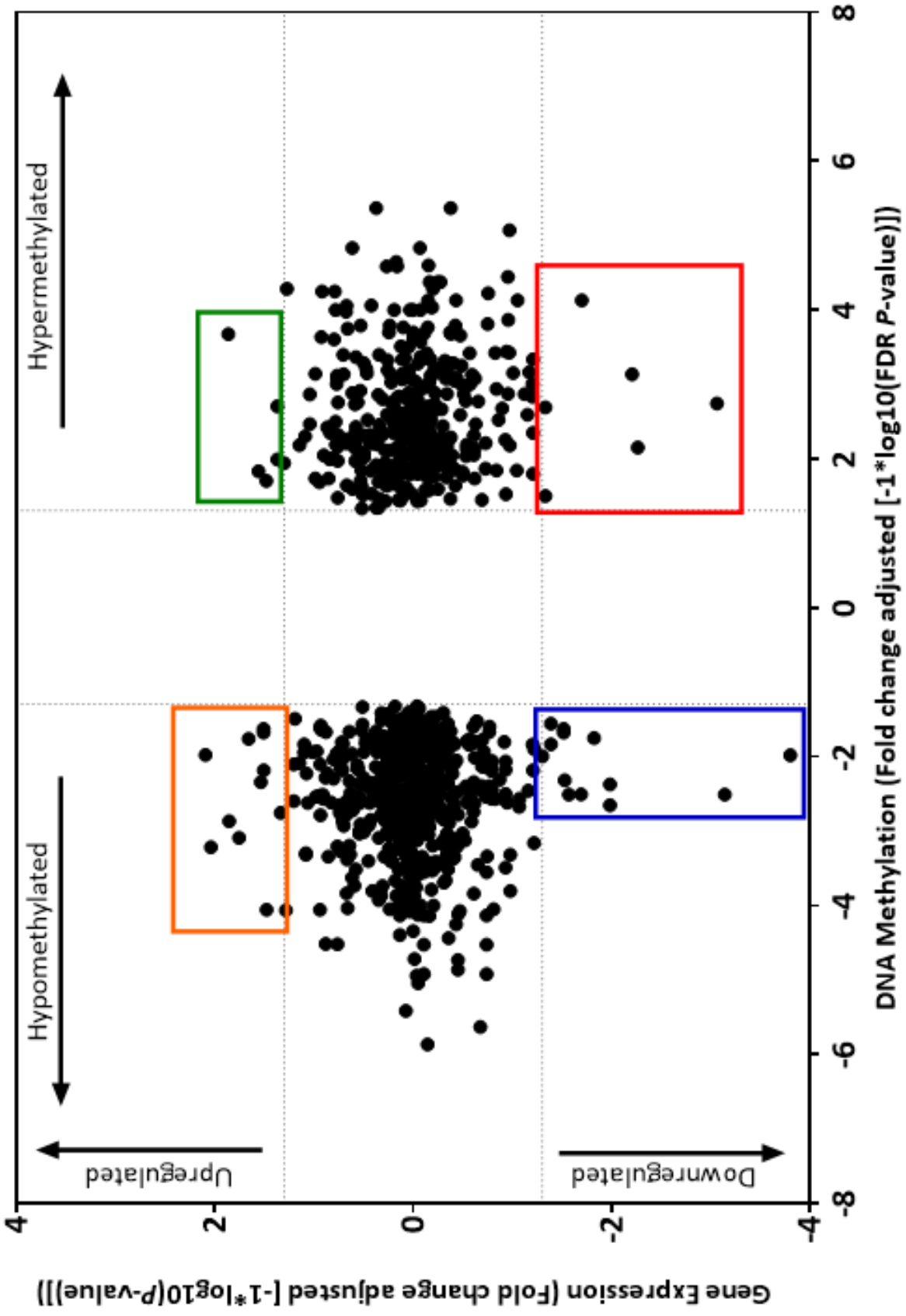

Hypermethylated
Hypomethylated
Upregulated
Downregulated
DNA Methylation (Fold change adjusted [-1*log10(FDR P-value)])
Gene Expression (Fold change adjusted [-1*log10(P-value)])
8
6
4
2
0
-2
-4
-6
-8
4
2
0
-2
-4

Table 5.12 Summary of the 5 genes hypermethylated and downregulated in the OA DNA methylation and gene expression analysis

		DNA Methylation		Gene Expression				
CpG	**Gene**	**Δβ**	**FDR *P*-value**	**LogFC**	***P*-value**	**Chr**	**Regional location**	**Gene location**
cg03025830	*FGF17*	0.183	0.032	-0.086	0.046	8	Body	Island
cg15246619	*FGF17*	0.198	0.002	-0.086	0.046	8	Body	Island
cg13469851	*PLAA*	0.193	< 0.001	-0.281	0.006	9	Body	NA
cg23084506	*PRAMEF5*	0.172	0.007	-0.147	0.005	1	Body	NA
cg00754604	*RBMS1*	0.204	< 0.001	-0.389	0.020	2	Body	NA

A positive Δβ value correlates with hypermethylation in patients with very low OA, compared to the rest. Abbreviations: Δβ, difference in mean β-value; FDR, false discovery rate; LogFC, $\log_2$ fold change; Chr, chromosome; NA, non-annotated.

Table 5.13 Summary of the 11 genes hypomethylated and upregulated in the OA DNA methylation and gene expression analysis

		DNA Methylation		Gene Expression				
CpG	Gene	Δβ	FDR *P*-value	LogFC	*P*-value	Chr	Regional location	Gene location
cg05977462	*CUX2*	-0.198	0.004	0.095	0.029	12	Body	Island
cg07343027	*DUS1L*	-0.215	0.010	0.702	0.008	17	TSS1500	S_Shore
cg16806041	*FHIT*	-0.258	0.002	0.323	0.046	3	Body	NA
cg10596483	*JRK*	-0.266	0.023	0.213	0.031	8	TSS1500	S_Shore
cg24634471	*JRK*	-0.258	0.021	0.213	0.031	8	TSS1500	S_Shore
cg00932063	*LHPP*	-0.219	0.001	0.406	0.014	10	Body	NA
cg11704876	*PPIEL*	-0.190	0.017	0.276	0.022	1	Body	NA
cg00815325	*LONP1*	-0.192	< 0.001	0.310	0.009	19	3'UTR	Island
cg12741184	*MYO1C*	-0.193	< 0.001	0.119	0.017	17	TSS1500	S_Shore
cg05420738	*RRAD*	-0.235	< 0.001	0.379	0.033	16	3'UTR	N_Shore
cg23825092	*TDRD1*	-0.188	0.006	0.251	0.031	10	TSS200	Island

A positive Δβ value correlates with hypermethylation in patients with very low OA, compared to the rest. Abbreviations: Δβ, difference in mean β-value; FDR, false discovery rate; LogFC, log2 fold change; Chr, chromosome; TSS1500, within 1500 bp of the transcription start site (TSS); TSS200, within 200 bp of the TSS; S_Shore, south shore; NA, non-annotated; N_Shore, north shore.

up-regulated gene expression (**Green** box, **Figure 5.30**) was also identified, with the majority of these genes (4/6) hypermethylated at gene body CpG locations (**Table 5.14**). Finally, a set of 13 genes, which were hypomethylated in patients with very low OA, but unexpectedly associated with down-regulated gene expression (**Blue** box, **Figure 5.30**) were identified. These genes were located between promoter and gene body CpG locations (**Table 5.15**). Interestingly, gene body hypomethylation has previously been negatively associated with gene expression in cancer, including chronic lymphocytic leukaemia[389] and breast cancer.[390]

The 5 gene hypermethylated/downregulated and 11 gene hypomethylated/upregulated gene-sets were of particular interest due to the consistent pattern expected for gene silencing and activation which they displayed. Among the hypermethylated and downregulated genes, *FGF17* (fibroblast growth factor 17; two significant CpGs), has been demonstrated to be upregulated in other cancers including hepatocellular carcinoma[391] and prostate cancer.[392,393] *FGF17* belongs to the FGF family, where it shares high sequence homology and expression patterns with *FGF8*. This has led to the suggestion that *FGF17* is a potent mediator of *FGF8* function, enhancing the survival of tumour cells under stress conditions.[391,393] Furthermore, among the hypomethylated and upregulated genes *JRK*, *RRAD* and *FHIT* were of interest. The Jerky homolog (mouse; *JRK*) gene encodes a conserved protein similar to DNA-binding proteins, such as the major centromere autoantigen B (*CENPB*). Recently, Jerky was demonstrated to facilitate Wnt signalling in colon carcinoma cells by promoting the association of β-catenin with T-cell factor, and recruiting β-catenin to chromatin.[394] Two CpGs mapped to *JRK* were significantly hypomethylated in patients with very low OA; however, confounding this was the identification of 2 significant probes from the gene expression analysis, with one probe upregulated and the other downregulated in very low OA, compared to the rest. This suggests that *JRK* may possibly have a role in Wnt/β-

Table 5.14 Summary of the 6 genes hypermethylated and upregulated in the OA DNA methylation and gene expression analysis

		DNA Methylation		Gene Expression				
CpG	**Gene**	**Δβ**	**FDR *P*-value**	**LogFC**	***P*-value**	**Chr**	**Regional location**	**Gene location**
cg04343101	*CACNA1H*	0.265	0.002	0.181	0.042	16	Body	S_Shore
cg08101977	*CACNA1H*	0.208	0.010	0.181	0.042	16	Body	S_Shore
cg07869907	*CDH22*	0.195	0.012	0.114	0.049	20	Body	Island
cg21361856	*CHN1*	0.357	< 0.001	0.261	0.013	2	TSS1500	S_Shore
cg21995068	*LOC100134868*	0.189	0.015	0.252	0.027	20	TSS1500	N_Shore
cg01499522	*MLXIPL*	0.193	0.020	0.173	0.033	7	Body	N_Shore

A positive Δβ value correlates with hypermethylation in patients with very low OA, compared to the rest. Abbreviations: Δβ, difference in mean β-value; FDR, false discovery rate; LogFC, $\log_2$ fold change; Chr, chromosome; TSS1500, within 1500 bp of the transcription start site (TSS); TSS200, within 200 bp of the TSS; S_Shore, south shore; N_Shore, north shore.

Table 5.15 Summary of the 13 genes hypomethylated and downregulated in the OA DNA methylation and gene expression analysis

A positive Δβ value correlates with hypermethylation in patients with very low OA, compared to the rest. Abbreviations: Δβ, difference in mean β-value; FDR, false discovery rate; LogFC, $\log_2$ fold change; Chr, chromosome; TSS1500, within 1500 bp of the transcription start site (TSS); TSS200, within 200 bp of the TSS; S_Shore, south shore; N_Shore, north shore.

		DNA Methylation		Gene Expression				
CpG	**Gene**	**Δβ**	**FDR *P*-value**	**LogFC**	***P*-value**	**Chr**	**Regional location**	**Gene location**
cg23034373	*CLDN14*	-0.170	0.018	-0.254	0.015	21	5'UTR	NA
cg06386478	*CSGALNACT1*	-0.198	0.004	-0.438	0.010	8	TSS200	Island
cg26286961	*CSGALNACT1*	-0.195	0.002	-0.438	0.010	8	TSS200	S_Shore
cg02519263	*HAR1B*	-0.209	0.014	-0.137	0.041	20	Body	N_Shore
cg01735357	*HAR1B*	-0.182	0.027	-0.137	0.041	20	Body	N_Shore
cg24266485	*HLA-DPB2*	-0.199	0.010	-0.116	0.049	6	Body	Island
cg10596483	*JRK*	-0.266	0.023	-0.112	0.030	8	TSS1500	S_Shore
cg24634471	*JRK*	-0.258	0.021	-0.112	0.030	8	TSS1500	S_Shore
cg07571531	*TCF21*	-0.284	0.003	-0.160	0.020	6	3'UTR	N_Shore
cg10698920	*TEX2*	-0.185	0.005	-0.707	0.029	17	5'UTR	N_Shore
cg20234855	*TMEM185B*	-0.186	0.003	-0.329	< 0.001	2	Body	Island
cg20234855	*TMEM185B*	-0.186	0.003	-0.230	0.027	2	Body	Island
cg09737019	*ZNF354A*	-0.179	0.010	-0.431	< 0.001	5	TSS1500	S_Shore

catenin signalling in very low OA. Additionally, the Ras-related associated with diabetes (*RRAD*) gene was recently suggested to act as a functional tumour suppressor gene in nasopharyngeal carcinoma.[395] Finally, expression and methylation status of the tumour suppressor, fragile histidine triad (*FHIT*) gene, has been well studied in many malignancies.[396] FHIT methylation has been studied previously in 16 BC-CML patient samples, where methylation was only observed in a single patient (1/16, 6%);[378] however, FHIT methylation was demonstrated to accumulate through disease progression in acute myeloid leukaemia (AML) and myelodysplastic syndrome (MDS).[397] Further investigation into these and other genes of interest, including validation of the gene expression array results, is now the focus of ongoing studies.

5.3 Discussion

Aberrant DNA methylation is a potential mechanism for activating or repressing the expression of genes that has been implicated in multiple haematological malignancies. In CML, DNA methylation analysis is poorly understood, with very few studies performed, and often only investigating genes identified from studies in other malignancies. Additionally, global DNA methylation analysis has not been documented at all in the CML literature, highlighting an area requiring significant interrogation. To address this issue, while also gaining valuable insight into the underlying biology associated with very low OA, global DNA methylation analysis was performed on a total of 63 CP-CML patients (n = 36 very low OA; n = 27 other OA) using the Illumina Infinium HumanMethylation450 BeadChip (450K). The Illumina Infinium 450K global DNA methylation array is a new technology that allows for the investigation of over 480,000 human CpG probes covering 99% of RefSeq genes and 96% of CpG island (CGI) regions.[199,383] As this technology is novel, the ideal analysis and data interpretation pipeline is still being sought; however, it is generally agreed that the default GenomeStudio analysis, as recommended by Illumina, is not optimal. Therefore, multiple publications in the literature have tried to address this issue, with each recommending the "best" analysis pipeline for 450K arrays.[204,398-404] At the time of writing, Dedeurwaerder *et al.*[405] reviewed these different analysis methods, developing an excellent guide for future 450K data analysis and interpretation. For the data presented in this chapter, the *minfi* package in combination with SWAN pre-processing, was used for all DNA methylation analyses, due to the ease of use and customisation afforded by this analysis pipeline. The *minfi*-SWAN pipeline was rated highly by Dedeurwaerder *et al.*[405]

The findings described in this chapter suggest that differential global DNA methylation is present between patients with very low OA, compared to all other patients. Although the

presence of SNP-CpGs and a possible gender bias did initially affect the results, once SNP-CpGs and sex-linked CpGs were excluded, in combination with batch effect removal, the "true" list of significant CpGs was identified. The bisulphite pyrosequencing data, generated in order to validate the 450K DNA methylation array and to develop a predictive classifier for OA, was quite confounding. The vast majority of the first cohort (cohort 1) of patients analysed by pyrosequencing, matched the patients used for 450K analysis, with excellent correlation observed between the methylation results from pyrosequencing and the 450K array. A predictive classifier for OA was also successfully developed using the pyrosequencing data from cohort 1. However, this classifier was unable to be validated using the independent patient data from cohort 2. Furthermore, the differential methylation observed from cohort 1 was not replicated when cohort 2 was analysed. No easy explanation for these conflicting results is forthcoming as 10 of the 11 CpGs analysed by pyrosequencing were still identified as differentially methylated between patients with very low OA compared to the rest, after the 450K data was re-analysed (**Section 5.2.10**). Therefore, this data suggests that by chance, the patients used for the 450K global DNA methylation analysis and in cohort 1 of the pyrosequencing analysis, may not be representative of the whole CP-CML population. The opposite could also be true, with the cohort 2 patients not being representative of the whole patient cohort. However, the patient characteristics between the two cohorts did not appear to differ significantly, apart from a higher percentage of males in cohort 2 (refer **Table 5.9**), yet, this confounding factor should have been removed as all sex-linked CpGs were removed from the analysis.

In order to determine if the observed differential methylation between CP-CML patients with very low OA, compared to the rest, was affecting the expression of the associated genes, gene expression microarray analysis was performed. The data generated from the

microarray analysis demonstrated that significant differential expression was observed between CP-CML patients with very low OA, compared to the rest. However, the highly significant differences expected to be observed were not, as none of the identified genes were differentially methylated with a FDR P-value < 0.05. Therefore, the lower stringency cut-off of P-value < 0.05 was used to identify significant genes. Correlation analysis between the gene list matched to the significant CpG analysis with the gene-set from the gene expression analysis, identified a small, but significant set of genes (n = 36) significant by both analyses. Importantly, 16 genes were observed to exhibit the common methylation and expression profiles associated with gene silencing (hypermethylated and downregulated, n = 5) or activation (hypomethylated and upregulated, n = 11). Included in this list were genes with documented roles as tumour suppressors in other malignancies (*RRAD* and *FHIT*) and Wnt/β-catenin signalling (*JRK*). GeneGo enrichment analysis had previously identified Hedgehog, TGF-β, NOTCH and WNT signalling as being associated with the very low OA significant CpG list, all of which have established roles in CML progression.[406,407] However, this data must be interpreted cautiously, as independent validation of the results from the gene expression array, and of particular target genes/CpGs is still required.

In conclusion, the findings detailed in this chapter strongly demonstrate that the specific global DNA methylation profile of patients with very low OA differs significantly from all other patients, providing evidence for an epigenetic influence in OA. The data presented in this chapter also identifies genes exhibiting differential methylation, which correlates with differential gene expression, providing possible targets for further study. Although these findings are preliminary, and require further validation, they do suggest a previously unknown epigenetic influence that may affect the underlying biology of CP-CML patients with very low OA.

6 GLOBAL DNA METHYLATION ANALYSIS OF CP-CML COMPARED TO NORMAL INDIVIDUALS AND DURING DISEASE PROGRESSION

6.1 Introduction

As described in **Chapter 5**, aberrant epigenetic regulation and in particular global DNA methylation changes are associated with many human diseases, including cancer.[358] Aberrant DNA hypomethylation is often associated with the activation of some proto-oncogenes, while CpG Island (CGI) promoter hypermethylation is associated with transcriptional silencing of tumour-suppressor genes.[360-363] The epigenetic impact of DNA methylation in CML is not well explored, especially by global DNA methylation analysis. This is particularly important for disease progression from chronic phase (CP) to blast crisis (BC) in CML, where aberrant DNA methylation of single genes has previously been demonstrated; however, global DNA methylation analysis is lacking. When 10 genes were investigated for promoter-associated CGI methylation, the average number of methylated genes significantly increased from 4.5 in the CP, to 6.4 during progression to BC.[375] Furthermore, global DNA methylation analysis comparing CP-CML patients to normal individuals is also not well explored, and may lead to the identification of targets unique to CML for alternative combination therapies to more effectively target the *BCR-ABL1* oncogene. This chapter aims to assess the differences in the global DNA methylation profile of CP-CML, compared to normal individuals and following progression to BC.

6.1.1 Approach

The Illumina Infinium HumanMethylation450 (450K) BeadChip was used as described in **Chapter 5** to perform global DNA methylation profiling for 63 CP-CML patients (n =36, very low OA [OA < 4 ng/200,000 cells]; n = 27, other OA [OA > 4 ng/200,000 cells]), 4 BC-CML patients and 5 normal individuals. The 450K arrays were performed by the Australian Genome Research Facility (AGRF; Melbourne, Australia). Batch effect removal was performed using the *ComBat* function from the *SVA* package in *R* (**Section 2.6.11**).

Methylation analysis was performed using the *minfi* package in *R* (**Section 2.6.10**), with various selection criteria used to identify differentially methylated CpGs (**Sections 2.6.10.4** and **2.6.10.5**). Spatial CpG mapping, and regional and gene structural analysis was also performed to identify the locations of the significant CpGs. Gene enrichment analysis was then performed to identify genes (mapped to the significant CpGs) with common biological function using the GeneGo Metacore™ software suite and the Molecular Signatures Database (MSigDB; **Section 2.6.4**). The analyses discussed in this chapter were used to determine how the global DNA methylation profiles of CP-CML patients differ from normal individuals and from patients in BC, to further understand epigenetic influence in CML.

6.2 Results

6.2.1 Global DNA methylation profiling identified significant epigenetic differences between CP-CML patients and normal individuals

Global DNA methylation profiling was performed using the Illumina Infinium HumanMethylation450 BeadChip (450K), which contains more than 480,000 CpG probes, covering 99% of RefSeq genes and 96% of CpG island (CGI) regions.[199,383] Firstly, global DNA methylation was compared between diagnostic CP-CML patient samples and normal individuals. To do this, the genome-wide DNA methylation profiles obtained from the TWCs of 63 CP-CML patients (refer **Chapter 5**) and 5 normal individuals were compared. To ensure batch effects were not confounding the results, the *ComBat* algorithm was applied. Once completed, differential methylation analysis was performed using the *minfi* package. Six CP-CML patients failed to pass QC measures and were excluded from further analysis, leaving 57 CP-CML patient samples available for the final analysis against the 5 normal samples.

6.2.1.1 Global hypomethylation is observed in CP-CML patients, compared to normal individuals

Identification of differentially methylated CpGs was initially performed using **Selection criteria 2A** (as described in **Section 2.6.10.4**), which included removing all CpGs located on chromosomes X and Y, with a detection *P*-value < 0.01, and containing a SNP within 10 bp of the CpG locus. Therefore, using a significance cut-off of FDR *P*-value < 0.05, 45,924 differentially methylated CpGs were identified, of which 19,248 CpGs were hypermethylated and 26,676 CpGs were hypomethylated in CP-CML patients, compared to normal individuals. As this number was quite high, a more stringent selection criteria was used (**Selection criteria 2B, Section 2.6.10.5**) which included only selecting the CpG probes with a difference in mean methylation levels, as determined by the β-value, ≥ 0.17 ($\Delta\beta \geq 0.17$). Using this criterion 11,597 differentially methylated CpGs were identified, of which 2,826 CpGs were

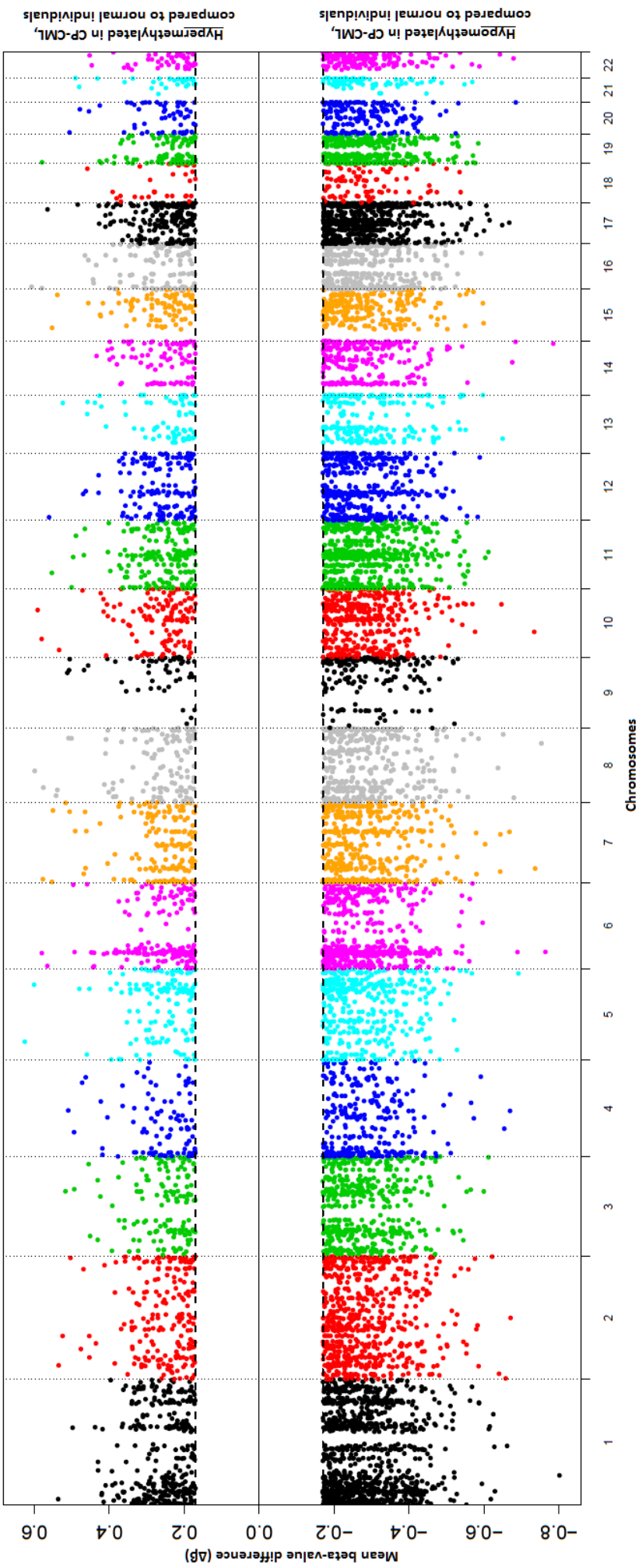
hypermethylated and 8,771 CpGs were hypomethylated in CP-CML patients, compared to normal individuals. Hence, this data indicated that global hypomethylation was more prevalent in CP-CML, compared to normal individuals. A spatial view of these sites mapped across the genome (chromosomes 1 – 22) revealed no bias in the distribution of these CpGs across all chromosomes (**Figure 6.1**).

6.2.1.2 The majority of differentially methylated CpGs between CP-CML patients and normal individuals are located in the gene body

Furthermore, regional analysis of these CpG sites in relation to CGIs revealed that the majority of CpGs were located in non-annotated regions. Specifically, 51% were located in non-annotated regions in relation to the CGI (5,908/11,597 sites), 9% in CGIs (1,036/11,597 sites), 28% in CpG shores (3,211/11,597 sites) and 12% in CpG shelves (1,442/11,597 sites; **Figure 6.2**). Interestingly, when the significant CpGs, hypermethylated or hypomethylated in CP-CML compared to normal individuals, were investigated separately, no difference in the distribution of the CpGs throughout the regional analyses was observed (**Figure 6.3**). Investigation into the position of these differentially methylated CpG sites in relation to the gene structure was then performed. The majority of significantly methylated CpGs were situated in the gene body, followed by non-annotated regions (**Figure 6.4**). Once again, this genomic distribution pattern was consistent for both the CpGs hyper- and hypomethylated in CP-CML patients, compared to normal individuals (**Figure 6.5**). Hierarchical clustering using the top 500 differentially methylated CpGs was able to successfully distinguish CP-CML patients from normal individuals (**Figure 6.6**).

Figure 6.1 The spatial view of all significant CpGs differentially methylated between CP-CML and normal individuals revealed an even distribution across all chromosomes

A manhattan plot was generated using all 11,597 differentially methylated CpGs with a FDR *P*-value < 0.05 and $\Delta\beta \geq 0.17$ and demonstrated an even distribution of the CpGs across all chromosomes. All 22 chromosomes are plotted on the X-axis, with the mean beta-value difference ($\Delta\beta$) for each individual CpG plotted on the Y-axis. A negative $\Delta\beta$ value indicates a CpG hypomethylated in CP-CML, compared to normal individuals, while a positive $\Delta\beta$ value indicates a CpG hypermethylated in CP-CML, compared to normal individuals. Each CpG is mapped to the exact position within its respective chromosome. Manhattan plots were generated using the *CpGassoc* package,[233] obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/).


Hypomethylated in CP-CML, compared to normal individuals
Hypermethylated in CP-CML, compared to normal individuals
Mean beta-value difference (Δβ)
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
Chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

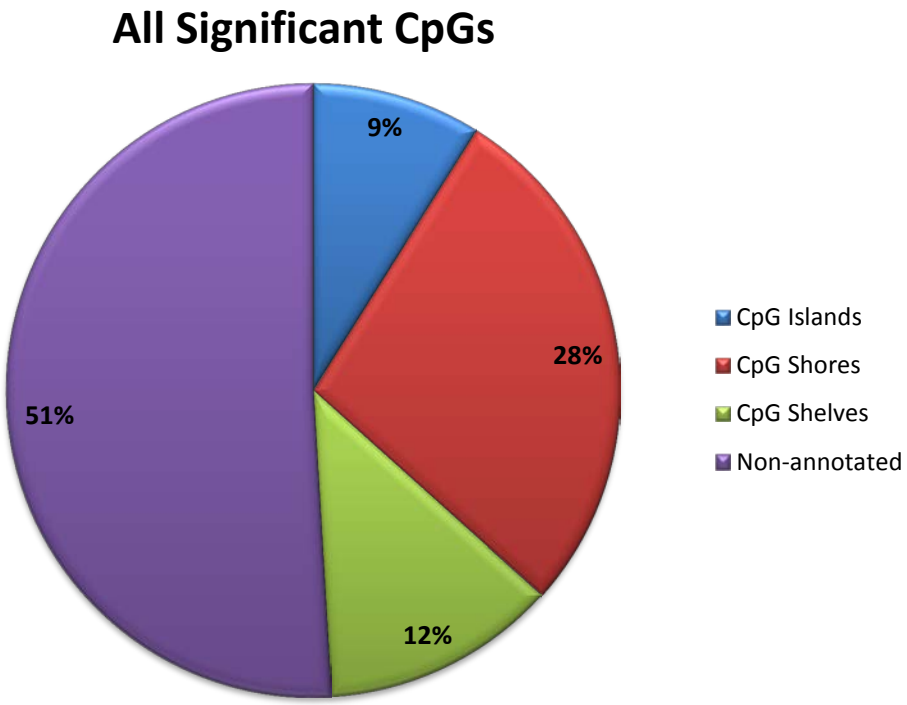


Figure 6.2 The majority of significant CpG sites from the CP-CML versus normal methylation analysis were located in non-annotated regions

All differentially methylated CpGs (11,597 sites) from CP-CML patients compared to normal individuals, with a known (CpG island, shore or shelf) or unknown (non-annotated) position were plotted. The majority (51%) of CpGs were located in non-annotated regions. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays.

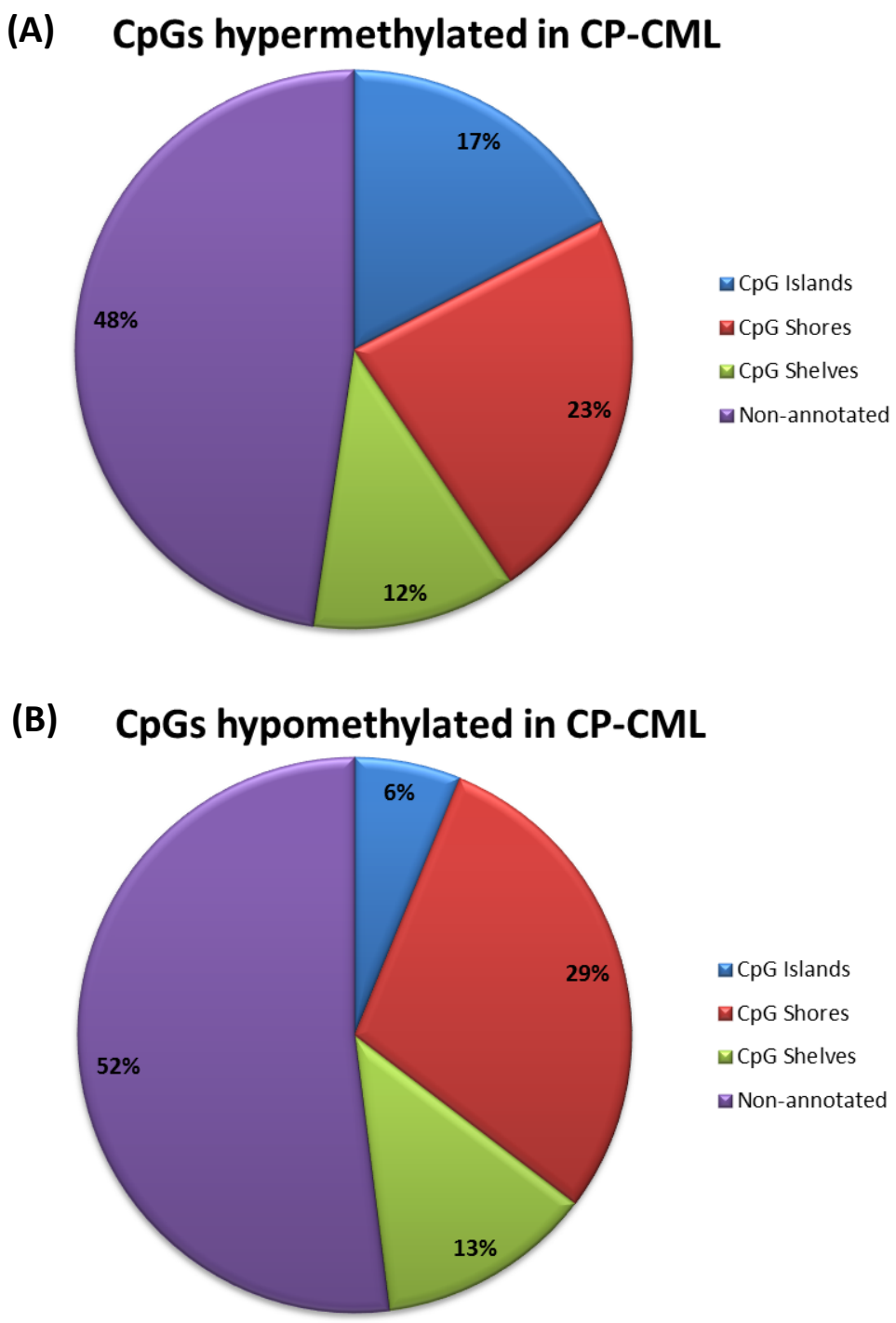


Figure 6.3 Regional location of significant hyper- and hypomethylated CpG sites in CP-CML patients, compared to normal individuals

The regional location of all **(A)** 2,826 CpGs significantly hypermethylated and **(B)** 8,771 CpGs significantly hypomethylated in CP-CML patients, compared to normal individuals, are demonstrated. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays.

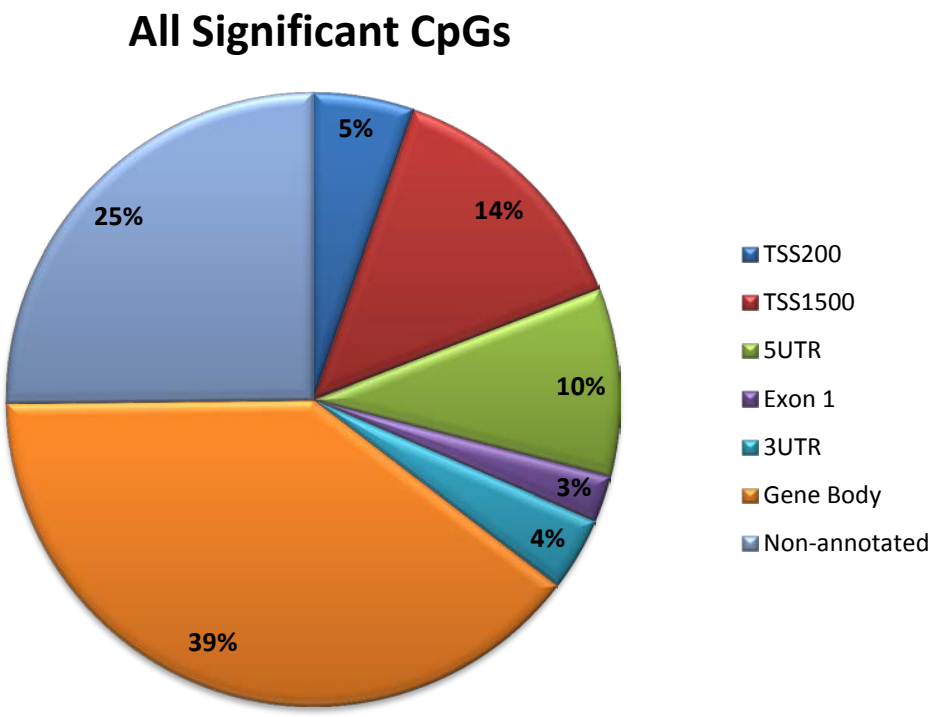


Figure 6.4 The genomic distribution of significant CpG sites from the methylation analysis comparing CP-CML to normal individuals revealed that the greatest proportion were located in gene bodies

The position of all differentially methylated CpGs (11,597 sites) from CP-CML patients compared to normal individuals, in relation to the gene structure are demonstrated. The greatest proportion of CpGs were located in the gene body. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

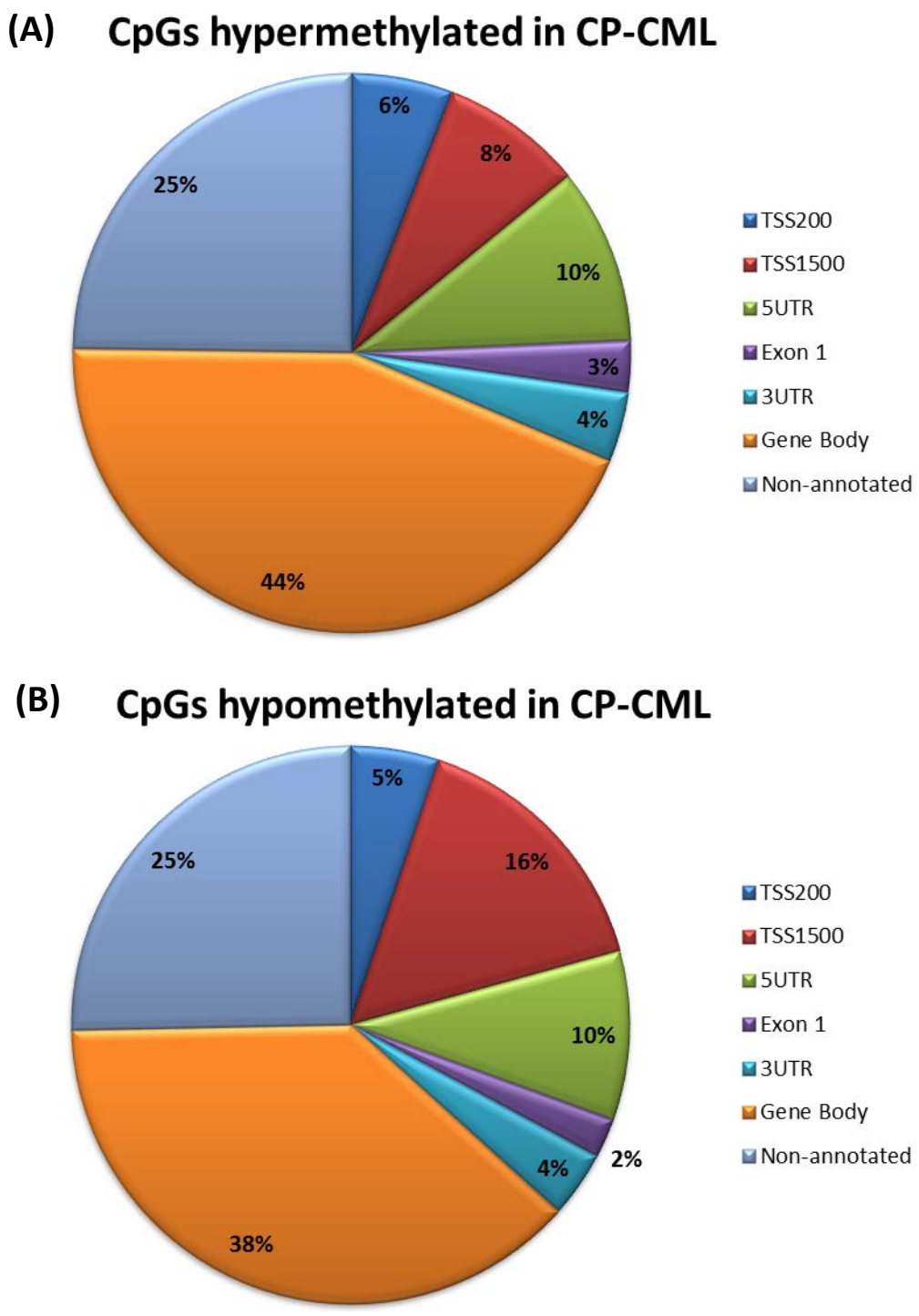


Figure 6.5 Genomic distribution of significant hyper- and hypomethylated CpG sites in CP-CML patients, compared to normal individuals

The position of all **(A)** 2,826 CpGs significantly hypermethylated and **(B)** 8,771 CpGs significantly hypomethylated in CP-CML patients, compared to normal individuals, in relation to the gene structure are demonstrated. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

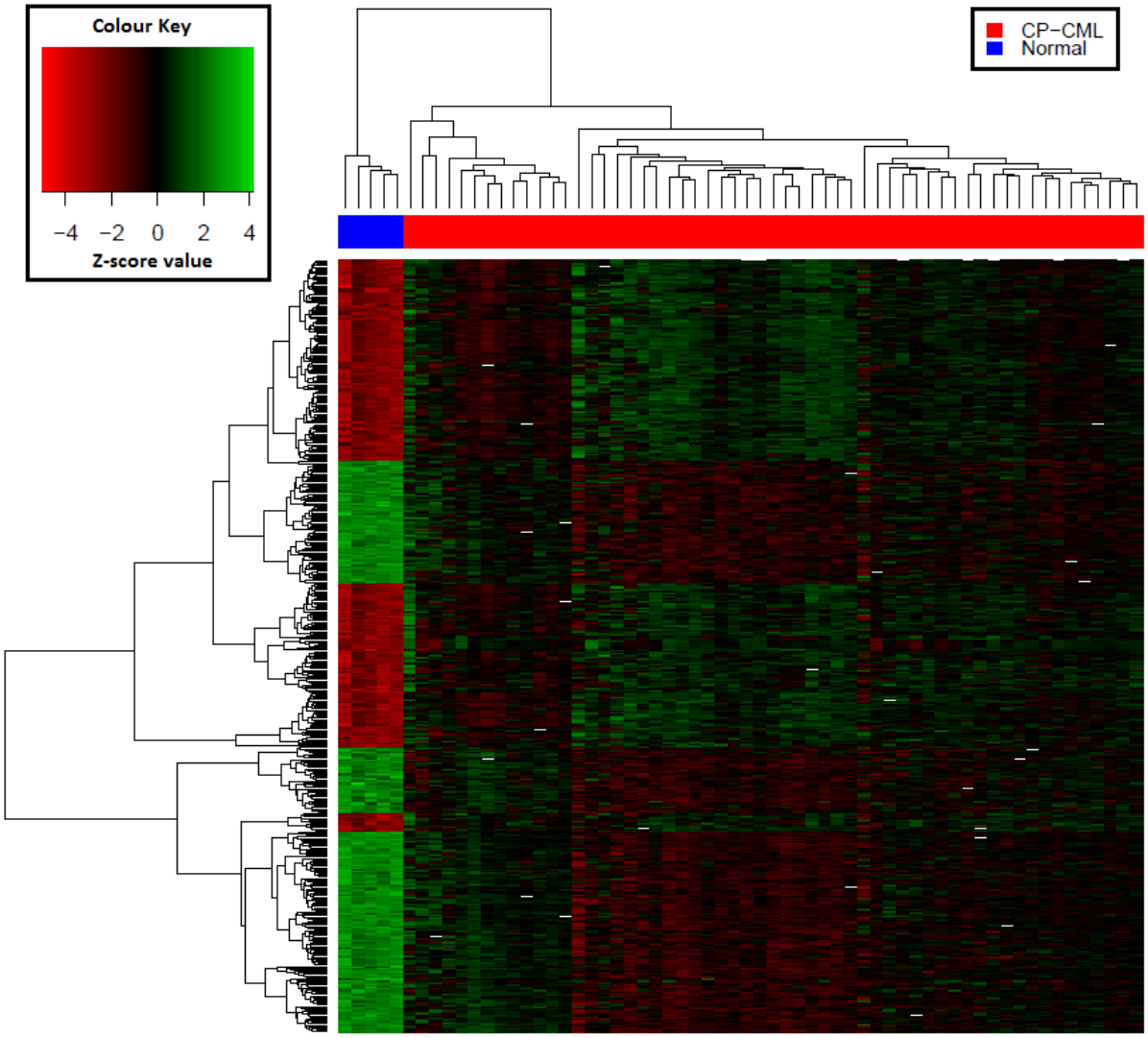


Figure 6.6 Hierarchical clustering using the top 500 differentially methylated CpGs distinguishes CP-CML patients from normal individuals

The top 500 CpGs differentiating CP-CML patients from normal individuals were used to perform unsupervised hierarchical clustering. Each column represents one sample, and each row represents one CpG. Z-score values (transformation of the methylation M-values in relation to the mean) were used to perform the clustering. A positive Z-score value (bright **green**) indicates a methylated CpG, while a negative Z-score value (bright **red**) indicates an unmethylated CpG. The intensity of the colour corresponds to the degree of methylation or non-methylation. Unsupervised hierarchical clustering (Euclidean, complete linkage) was performed using the heatmap function from the *Stats* package in *R*, v 3.0.1.

6.2.2 Pathway analysis identified significant enrichment of genes encoding multiple cellular processes associated with CP-CML patients, compared to normal individuals

GeneGo pathway and gene ontology enrichment analysis was performed using the 4,822 genes corresponding to the differentially methylated CpGs identified by the 450K array between CP-CML patients and normal individuals. Genes associated with a number of highly significant cellular processes and pathways were enriched from this gene-set (**Table 6.1**). Included in this list were pathways involved with cell adhesion and immune response. Furthermore, cellular pathways involved in development (granulocyte macrophage-colony stimulating factor [GM-CSF] signalling), apoptosis via the tumour necrosis factor (TNF)-family and cytoskeleton remodelling via the transforming growth factor (TGF) and WNT pathways, known to be involved in the protection of CML leukaemic stem cells from TKI therapy,[278,408-410] were also enriched. Additionally, signal transduction encoded by the phosphatase and tensin homolog (PTEN) gene, which is a known tumour suppressor gene in CML,[411] was also enriched in this gene-set.

6.2.3 Gene-set enrichment analysis identifies significant enrichment of Polycomb group proteins and cancer pathways associated with CP-CML patients, compared to normal individuals

Gene-set enrichment analysis was performed by investigating the Molecular Signatures Database (MSigDB, Broad Institute) using the 4,822 gene list that corresponded to the 11,597 differentially methylated CpGs identified by the 450K array between CP-CML patients and normal individuals. A number of highly significant gene-sets were enriched from the gene list corresponding to the differentially methylated CpGs (**Table 6.2**). Included in this list were pathways involving Polycomb group proteins (EZH2, PRC2, EED and SUZ12), as well as various oncogenic pathways (Liver and Colon). Notably, Polycomb group proteins have a known role in stem cells and cancer,[388] with the Polycomb group *BMI1* gene previously

Table 6.1 Common cellular processes and pathways enriched from the CP-CML versus normal gene-set by GeneGo analysis

	Total gene-set size	Genes in dataset	*P*-value
Cellular Pathways			
Cytoskeleton remodelling – TGF, WNT and cytoskeleton remodelling	111	54	$3.05e^{-14}$
Development – GM-CSF signalling	50	29	$1.24e^{-10}$
Apoptosis and survival – Apoptotic TNF-family pathways	42	26	$1.50e^{-10}$
Signal transduction – PTEN pathway	46	24	$8.21e^{-08}$

The total gene-set size column represents the total number of genes in the GeneGo gene-set, while the genes in dataset column represents the number of genes overlapping between the gene list of interest (from the CP-CML versus normal analysis) and the GeneGo gene-set.

Table 6.2 Gene-sets enriched in CP-CML patients compared to normal individuals by MSigDB enrichment analysis

	Gene-set Size*	*P*-value	FDR *q*-value
Polycomb group proteins			
Nuytten – EZH2 targets Down	211	< 0.0001	< 0.0001
Benporath – PRC2 targets	132	< 0.0001	< 0.0001
Benporath – EED targets	222	< 0.0001	< 0.0001
Benporath – SUZ12 targets	233	< 0.0001	< 0.0001
Lu – EZH2 targets Down	97	< 0.0001	< 0.0001
Cancer pathways			
Acevedo – Liver cancer Up	189	< 0.0001	< 0.0001
Grade – Colon cancer Up	173	< 0.0001	< 0.0001
Acevedo – Methylated in liver cancer Down	198	< 0.0001	< 0.0001

Abbreviations: FDR – false discovery rate. *Gene-set size refers to the number of genes overlapping between the gene-sets from the MSigDB with the gene list corresponding to the significant CpGs from the 450K DNA methylation analysis between CP-CML versus normal.

demonstrated to be a prognostic marker in CML.[320] Furthermore, Polycomb group proteins were also significantly enriched in the very low OA, compared to the rest, DNA methylation analysis (refer **Chapter 5, Section 5.2.11**).

6.2.4 Global DNA methylation profiling identified significant epigenetic differences between CP-CML and BC-CML patients

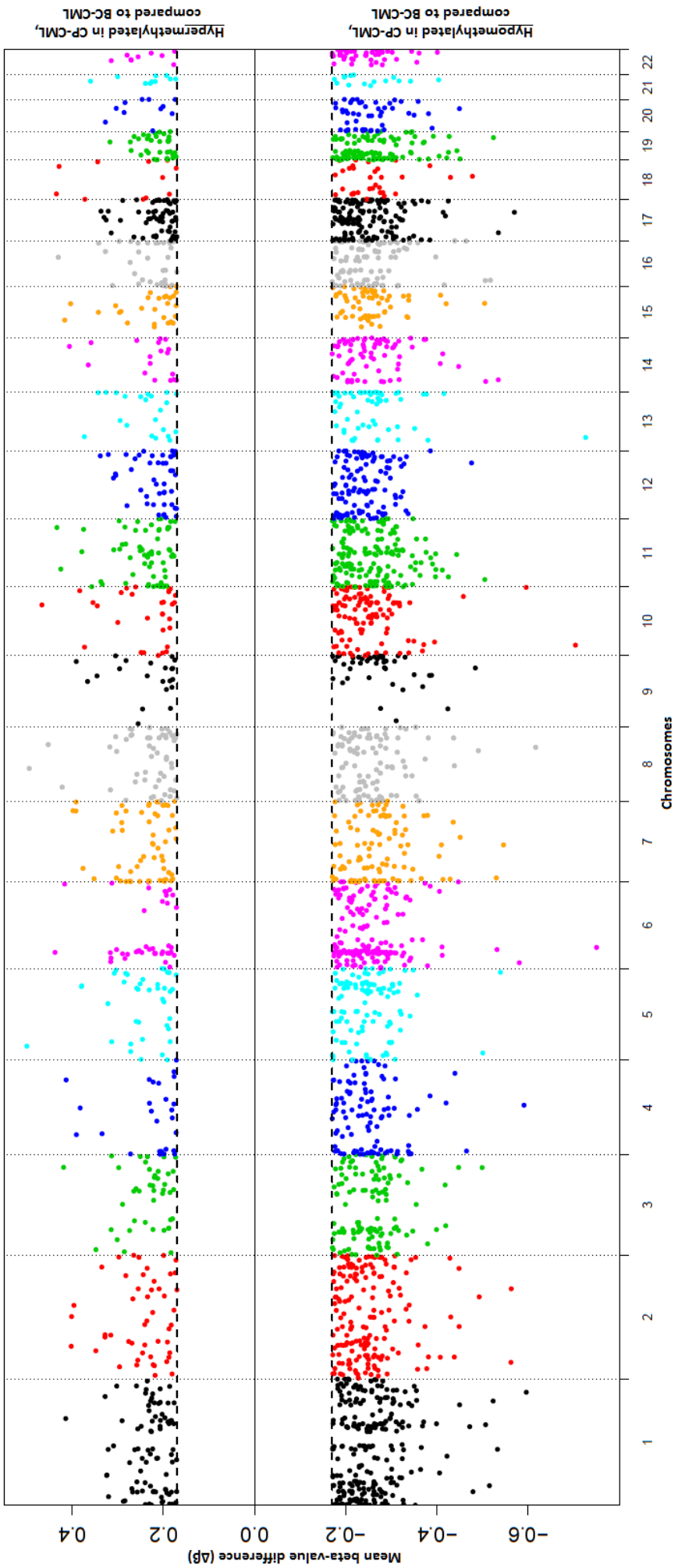
The global DNA methylation differences between patients with CML in CP versus those with disease progression to BC (BC-CML) were then compared. Once again, the Illumina 450K arrays were used to perform global DNA methylation profiling using TWCs samples (at diagnosis for CP-CML and at the time of progression for the BC-CML patient samples) as previously discussed in **Section 6.2.1**. Final analysis compared 57 CP-CML patient samples against 4 BC-CML patient samples.

6.2.4.1 Global hypermethylation is observed in BC-CML patients, compared to CP-CML patients

Identification of differentially methylated CpGs was performed using the stringent **Selection criteria 2B** (as described in **Section 2.6.10.5**), which included removing all CpGs located on chromosomes X and Y, with a detection *P*-value < 0.01, containing a SNP within 10 bp of the CpG locus and a difference in mean methylation levels, as determined by the β-value, ≥ 0.17 ($\Delta\beta \geq 0.17$). Therefore, using a significance cut-off of FDR *P*-value < 0.05, 3,014 differentially methylated CpGs were identified, of which 2,227 CpGs were hypermethylated and 787 CpGs hypomethylated in BC-CML, compared to CP-CML. A spatial view of these sites mapped across the genome (chromosomes 1 – 22) revealed no bias in the distribution of these CpGs across all chromosomes (**Figure 6.7**).

Figure 6.7 The spatial view of all significant CpGs differentially methylated between CP-CML and BC-CML patients revealed an even distribution across all chromosomes

A manhattan plot was generated using all 3,014 differentially methylated CpGs with a FDR *P*-value < 0.05 and $\Delta\beta \geq 0.17$ and demonstrated an even distribution of the CpGs across all chromosomes. All 22 chromosomes are plotted on the X-axis, with the mean beta-value difference ($\Delta\beta$) for each individual CpG plotted on the Y-axis. A negative $\Delta\beta$ value indicates a CpG hypomethylated in CP-CML, compared to BC-CML, while a positive $\Delta\beta$ value indicates a CpG hypermethylated in CP-CML, compared to BC-CML. Each CpG is mapped to the exact position within its respective chromosome. Manhattan plots were generated using the *CpGassoc* package,[233] obtained from the Comprehensive R Archive Network (CRAN) software project (http://www.cran.r-project.org/).


Hypomethylated in CP-CML, compared to BC-CML
Hypermethylated in CP-CML, compared to BC-CML
Mean beta-value difference (Δβ)
−0.6
−0.4
−0.2
0.0
0.2
0.4
Chromosomes
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

6.2.4.2 The majority of differentially methylated CpGs between CP-CML and BC-CML patients are located in the gene body

Regional analysis of the significant CpG sites in relation to CGIs revealed that the majority of CpGs were located in non-annotated regions. Specifically, approximately 55% were located in non-annotated regions in relation to the CGI (1,660/3,014 sites), 9% in CGIs (270/3,014 sites), 24% in CpG shores (722/3,014 sites) and 12% in CpG shelves (360/3,014 sites; **Figure 6.8**). When the significant CpGs hypermethylated in BC-CML, compared to CP-CML, were investigated separately, no difference in the distribution of the CpGs was observed (**Figure 6.9 [A]**). However, an increase in the percentage of non-annotated CpGs was observed for the CpGs hypomethylated in BC-CML, compared to CP-CML (**Figure 6.9 [B]**). Investigation into the structural genomic position of these differentially methylated CpG sites revealed the majority of significant CpGs were situated in the gene body, followed by non-annotated regions (**Figure 6.10**). As before, the genomic distribution pattern was different between the CpGs hyper- and hypomethylated in BC-CML patients, compared to CP-CML (**Figure 6.11**). Specifically, no difference in the genomic distribution pattern was observed for CpGs significantly hypermethylated in BC-CML, compared to CP-CML (**Figure 6.11 [A]**) from the overall distribution. However, the genomic distribution pattern of the CpGs hypomethylated in BC-CML, compared to CP-CML had a vastly different distribution, as the majority of CpGs were now located in non-annotated regions, followed by the 5'UTR (5' untranslated region; **Figure 6.11 [B]**). Unfortunately, the effect of this different genomic distribution is hard to determine with such a large proportion of non-annotated CpGs, which provide no information towards the CpG location. Hierarchical clustering using the top 100 differentially methylated CpGs was able to successfully distinguish CP-CML patients from BC-CML patients (**Figure 6.12**).

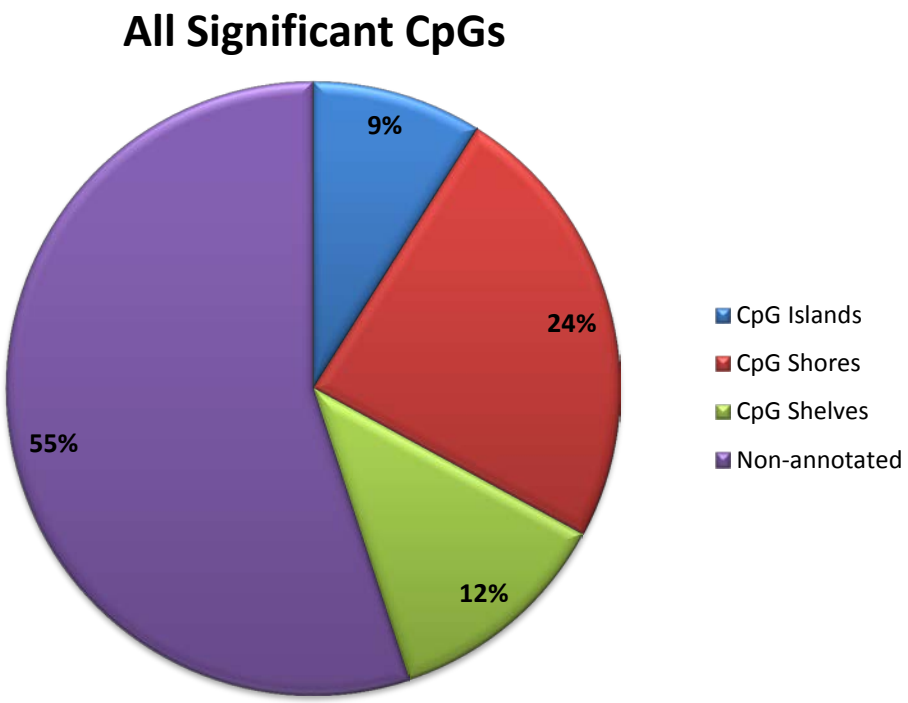


Figure 6.8 The majority of significant CpG sites from the CP-CML versus BC-CML methylation analysis were located in non-annotated regions

All differentially methylated CpGs (3,014 sites) from CP-CML compared to BC-CML patients, with a known (CpG island, shore or shelf) or unknown (non-annotated) position were plotted. The majority of CpGs were located in non-annotated regions. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file, which contains complete annotation for the Illumina 450K arrays.

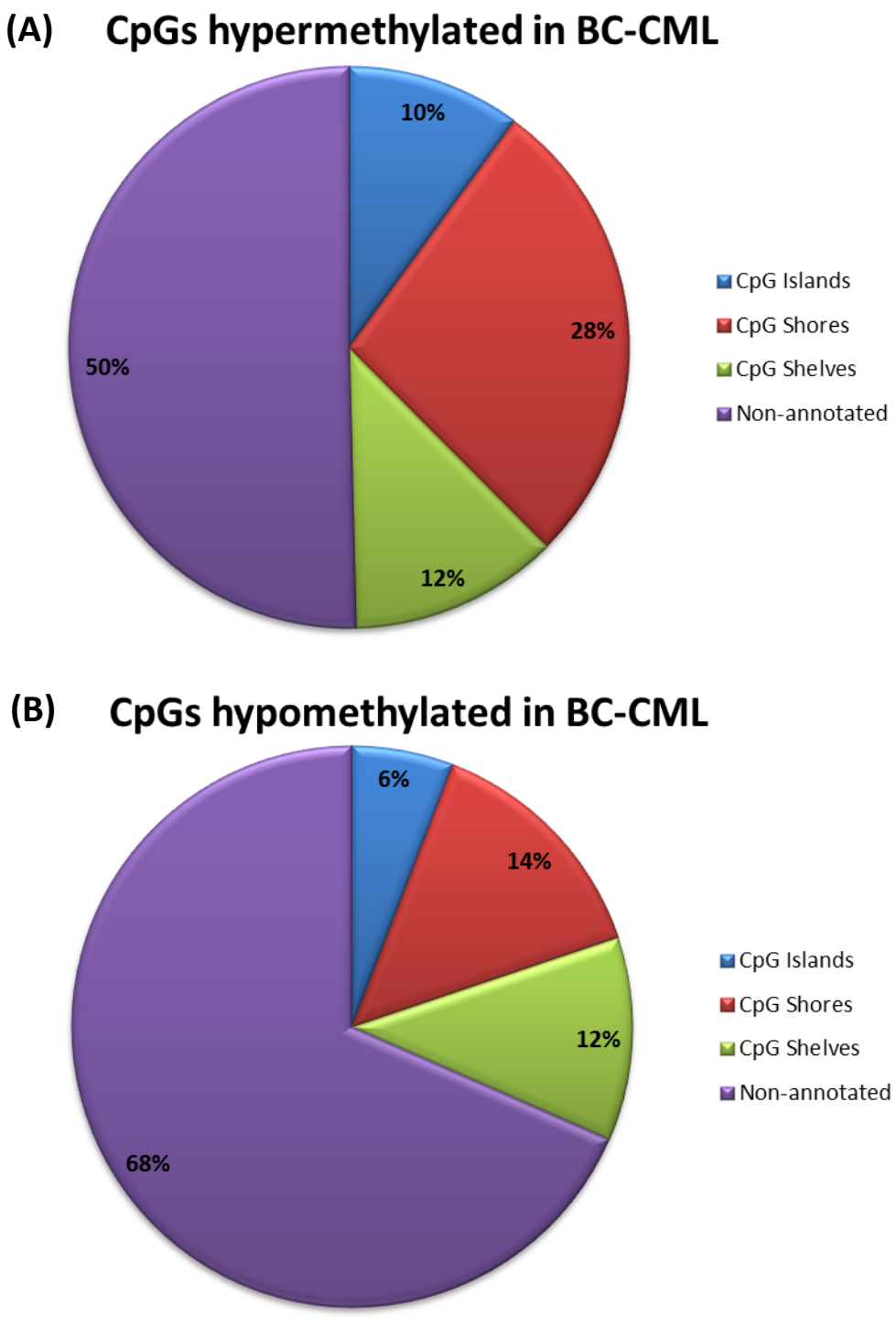


Figure 6.9 Regional location of significant hyper- and hypomethylated CpG sites in BC-CML patients, compared to CP-CML

The regional location of all **(A)** 2,227 CpGs significantly hypermethylated and **(B)** 787 CpGs significantly hypomethylated in BC-CML patients, compared to CP-CML patients, are demonstrated. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays.

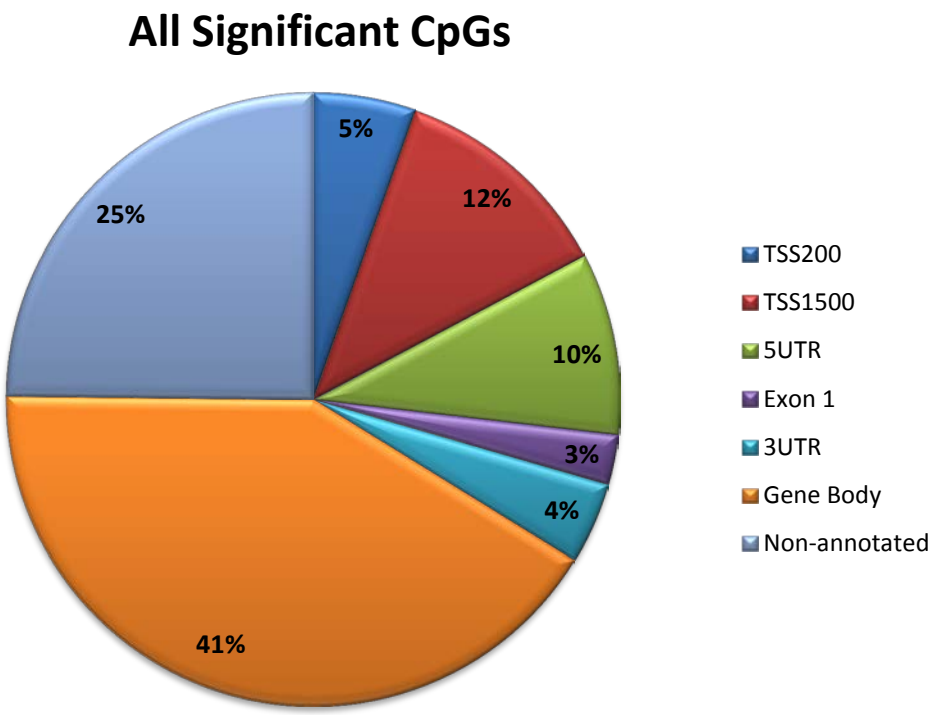


Figure 6.10 The genomic distribution of significant CpG sites from the methylation analysis comparing CP-CML to BC-CML patients revealed that the greatest proportion were located in gene bodies

The position of all differentially methylated CpGs (3,014 sites) from CP-CML patients compared to BC-CML patients, in relation to the gene structure are demonstrated. The greatest proportion of CpGs were located in the gene body. Regional locations for each CpG were obtained using the IlluminaHumanMethylation450manifest file which contains complete annotation for the Illumina 450K arrays. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

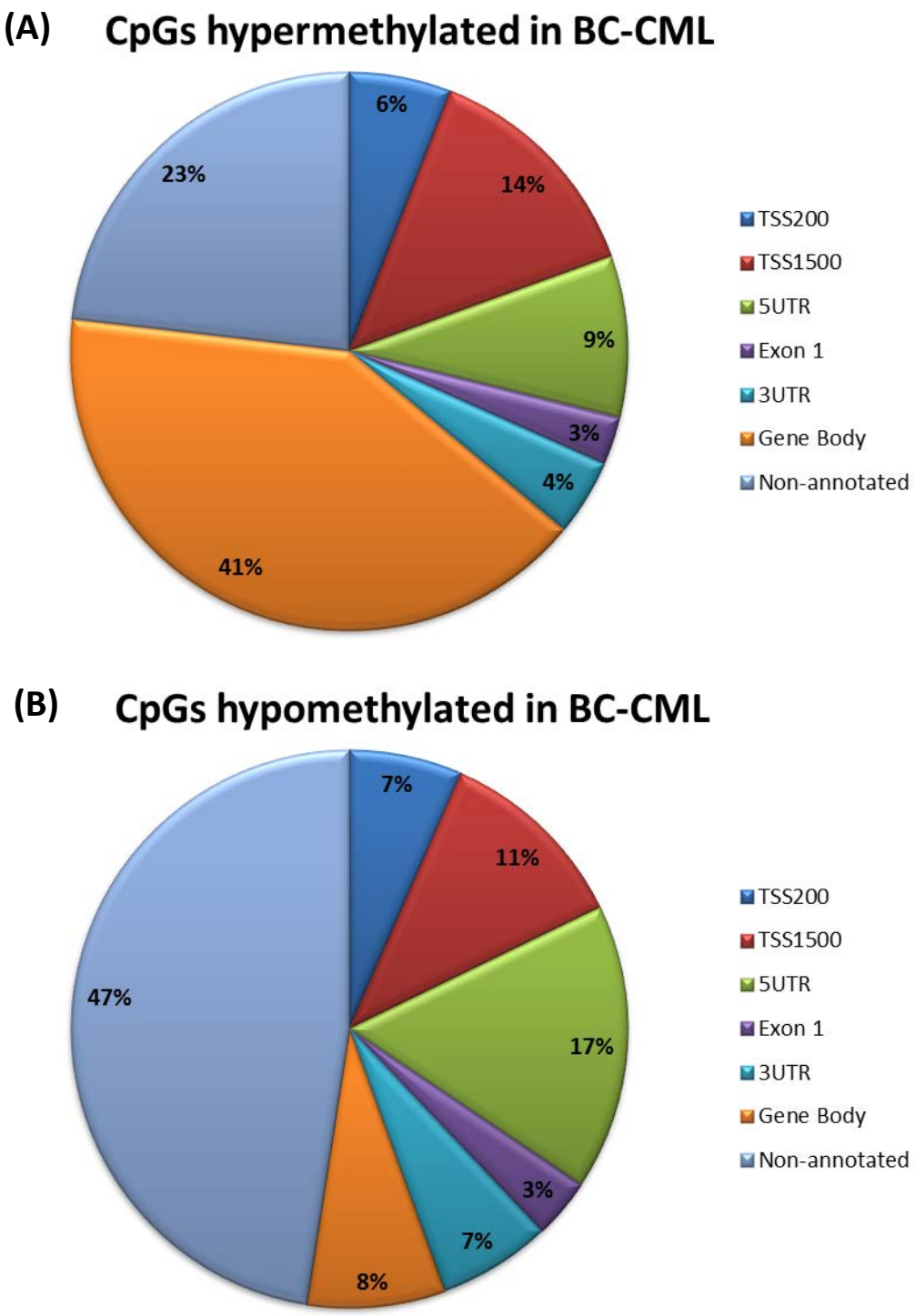


Figure 6.11 Genomic distribution of significant hyper- and hypomethylated CpG sites in BC-CML patients, compared to CP-CML patients

The position of all **(A)** 2,227 CpGs significantly hypermethylated and **(B)** 787 CpGs significantly hypomethylated in BC-CML patients, compared to CP-CML patients, in relation to the gene structure are demonstrated. Abbreviations: TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

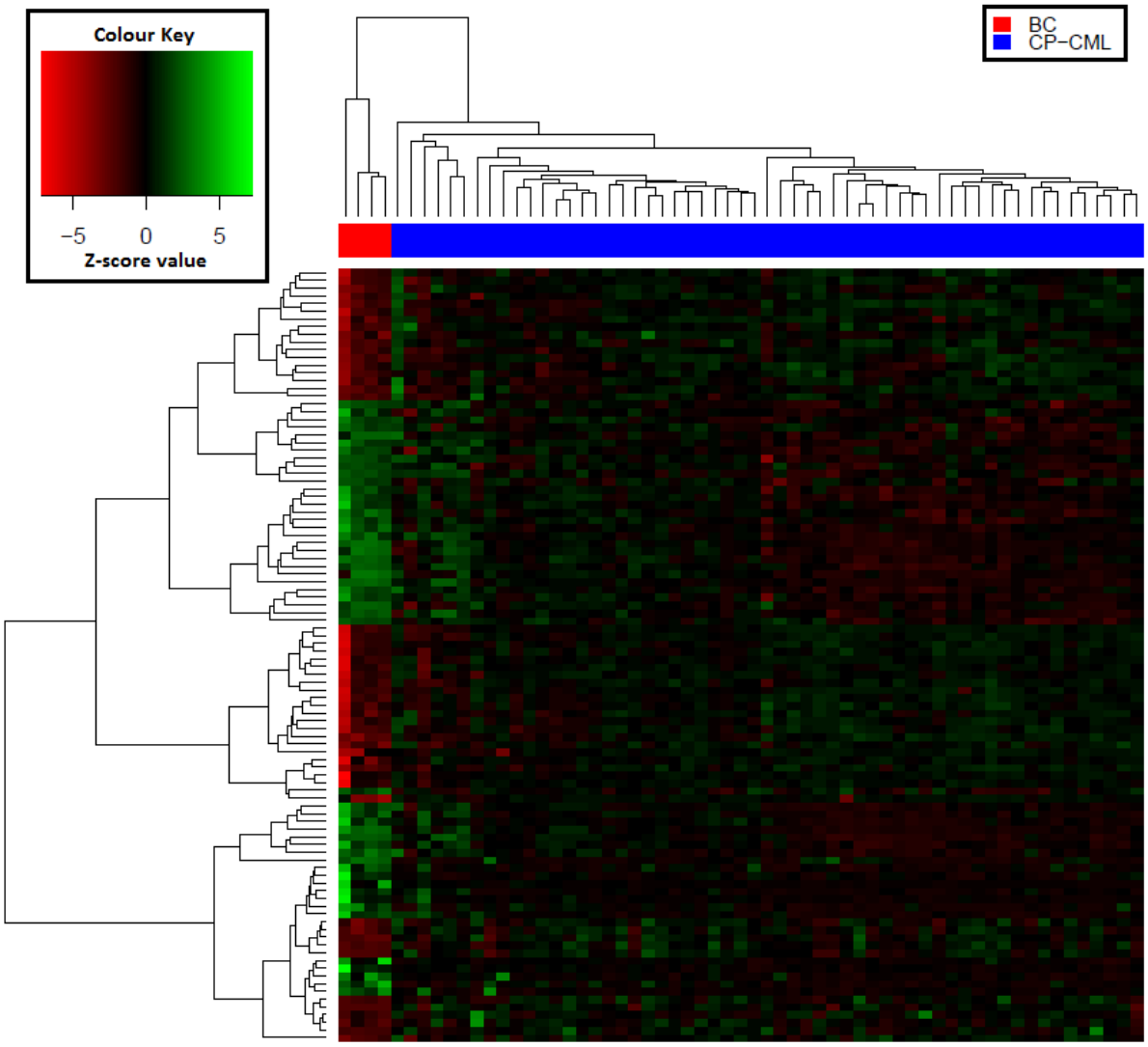


Figure 6.12 Hierarchical clustering using the top 100 differentially methylated CpGs distinguishes CP-CML patients from BC-CML patients

The top 100 CpGs differentiating CP-CML from BC-CML patients were used to perform unsupervised hierarchical clustering. Each column represents one sample, and each row represents one CpG. Z-score values (transformation of the methylation M-values in relation to the mean) were used to perform the clustering. A positive Z-score value (bright **green**) indicates a methylated CpG, while a negative Z-score value (bright **red**) indicates an unmethylated CpG. The intensity of the colour corresponds to the degree of methylation or non-methylation. Unsupervised hierarchical clustering (Euclidean, complete linkage) was performed using the heatmap function from the *Stats* package in *R*, v 3.0.1.

6.2.5 Pathway analysis identified significant enrichment of genes encoding multiple cellular processes associated with BC-CML, compared to CP-CML

GeneGo pathway and gene ontology enrichment analysis was performed using the 1,732 genes corresponding to the differentially methylated CpGs identified by the 450K array comparing CP-CML and BC-CML patients. Genes associated with a number of highly significant cellular processes and pathways were enriched in this gene-set (**Table 6.3**). Included in this list were pathways involved with cell adhesion, cell cycle and apoptosis (specifically anti-apoptosis mediated by external signals via NF-κB). Furthermore, cellular pathways involved in immune response (IL-2 activation and signalling) and cytoskeleton remodelling via the TGF and WNT pathways, known to be involved in CML and identified in the enrichment analysis from **Section 6.2.2**, were also enriched.

6.2.6 Gene-set enrichment analysis identifies significant enrichment of Polycomb group proteins and p53 and RB1 gene-sets associated with BC-CML, compared to CP-CML

Gene-set enrichment analysis was performed by investigating the Molecular Signatures Database (MSigDB, Broad Institute) using the 1,732 gene list that corresponded to the 3,014 differentially methylated CpGs identified by the 450K array between CP-CML and BC-CML patients. A number of highly significant gene-sets were enriched from the gene list of interest (**Table 6.4**). Included in this list were pathways involved with Polycomb group genes (SUZ12 and EZH2), and the tumour-suppressor genes *p53* and *RB1* (retinoblastoma 1). Interestingly, Polycomb group genes were also enriched in the CP-CML versus normal individuals significant gene-set (**Section 6.2.3**), suggesting that not only do they have a role in CML, but also possibly in the progression of CML from CP to BC. Furthermore, genetic or functional inactivation of the tumour-suppressor genes *p53* and *RB1*, is known to be one of the most common abnormalities observed in BC-CML.[412]

Table 6.3 Common cellular processes and pathways enriched from the CP-CML versus BC-CML gene-set by GeneGo analysis

	Total gene-set size	Genes in dataset	*P*-value
Process Networks			
Cell adhesion	184	42	$1.90e^{-06}$
Cell cycle – G1-S Growth factor regulation	195	36	$9.63e^{-04}$
Apoptosis – Anti-apoptosis mediated by external signals via NF-κB	111	23	$1.66e^{-03}$
Cellular Pathways			
Immune response – IL-2 activation and signalling pathway	49	12	$1.53e^{-05}$
Cytoskeleton remodelling – TGF, WNT and cytoskeleton remodelling	111	18	$3.17e^{-04}$
Gene Ontology Processes			
Negative regulation of granulocyte differentiation	21	15	$4.90e^{-12}$
Positive regulation of cell proliferation	1012	109	$5.56e^{-11}$

The total gene-set size column represents the total number of genes in the GeneGo gene-set, while the genes in dataset column represents the number of genes overlapping between the gene list of interest (from the CP-CML versus BC-CML analysis) and the GeneGo gene-set.

Table 6.4 Gene-sets enriched in CP-CML patients compared to BC-CML patients by MSigDB enrichment analysis

	Gene-set Size*	*P*-value	FDR *q*-value
Polycomb group proteins			
Benporath – SUZ12 targets	96	< 0.0001	< 0.0001
Nuytten – EZH2 targets Up	112	< 0.0001	< 0.0001
p53 and RB1 pathways			
Perez – TP53 targets	124	< 0.0001	< 0.0001
Martinez – RB1 targets Up	74	< 0.0001	< 0.0001
Martinez – RB1 and TP53 targets	67	< 0.0001	< 0.0001
Martinez – TP53 targets Down	70	< 0.0001	< 0.0001

Abbreviations: FDR – false discovery rate; RB1 – retinoblastoma 1. *Gene-set size refers to the number of genes overlapping between the gene-sets from the MSigDB with the gene list corresponding to the significant CpGs from the 450K DNA methylation analysis between CP-CML versus BC-CML.

6.2.7 Global DNA methylation profiling identified very slight epigenetic differences between CP-CML patients and CP-CML patients who later transformed to BC while receiving TKI therapy

Included in the 63 CP-CML patient dataset, originally analysed by the Illumina 450K DNA methylation arrays, was a subset of 8 CP-CML patients (termed CP-BC) who subsequently transformed to BC, despite receiving tyrosine kinase inhibitor (TKI) therapy. These patients represent a unique subgroup of patients with particularly poor prognosis. Importantly, if these CP-BC patients can be identified at diagnosis, personalised therapeutic strategies can be developed aimed towards improving their overall response. Therefore, the global DNA methylation profiles of these 8 CP-BC patients, using TWCs at diagnosis (when the patient was in CP), were compared against the remaining 55 CP-CML patients who did not transform to BC on TKI therapy, as discussed in **Section 6.2.1**. Once again, *ComBat* was used to remove batch effects and differential methylation analysis was performed using the *minfi* package. Six CP-CML and 1 CP-BC patients failed to pass quality control measures and were excluded from further analysis, leaving 49 CP-CML patient samples available for the final analysis against 7 CP-BC patient samples.

Identification of differentially methylated CpGs was performed using the **Selection criteria 2A** (as described in **Section 2.6.10.4**), with slight modification. As no CpGs were identified with a FDR *P*-value < 0.05, significant CpGs were selected based on a *P*-value < 0.05. Therefore, using this significance cut-off, 1,752 differentially methylated CpGs were identified, of which 632 CpGs were hypermethylated and 1,119 CpGs were hypomethylated in CP-BC patients, compared to CP-CML patients. Of these differentially methylated CpGs, only 3 had a $\Delta\beta \geq 0.17$ (**Selection criteria 2B**), with 1 CpG hypermethylated and 2 CpGs hypomethylated in CP-BC patients, compared to CP-CML patients (**Table 6.5**).

Table 6.5 Summary of the 3 CpGs identified from the CP-CML compared to CP-BC DNA methylation analysis

CpG	Gene	*P*-value	Δβ	Chr	Regional location	Gene location
cg19015611	*VAX2*	0.0009	-0.247	2	TSS1500	N_Shore
cg01775802	*RGS6*	0.0082	0.179	14	Body	NA
cg20862097	*DAAM2*	0.0201	0.187	6	TSS200	S_Shore

A negative Δβ correlates with hypomethylation in CP-CML patients, compared to CP-BC patients; while a positive Δβ correlates with hypermethylation in CP-CML patients, compared to CP-BC patients. Abbreviations: Δβ, difference in mean methylation (β-value) levels; Chr, chromosome; TSS1500, within 1,500 bp of the transcription start site (TSS); TSS200, within 200 bp of the TSS; NA, non-annotated.

Two CpGs were located in the gene promoter region (TSS1500 and TSS200), and 1 in the gene body. Surprisingly, 5 out of the 7 CP-BC patients transformed to BC within the first 9 months of TKI therapy, indicating that if a BC-epigenetic signature was present, it should have been detected in these patients, as if the patients transformed beyond 12 months of TKI therapy, it was less likely to be revealed. Taken together, this data demonstrates that at an epigenetic level, the diagnostic global DNA methylation profiles of CP-CML patients and CP-CML patients who later progress to BC while receiving TKI therapy (CP-BC) do not differ significantly. Therefore, suggesting that DNA methylation profiling is not informative for identifying these patients at diagnosis; however, these results should be interpreted with caution due to the small CP-BC patient cohort analysed.

6.2.8 A high correlation was observed between the significant CpG lists identified from the CML versus normal, the CP-CML versus BC-CML, and the very low versus other OA DNA methylation analyses

To determine if there was any overlap between the significant CpGs identified from any of the DNA methylation analyses performed in this chapter (i.e. CP-CML versus normal individuals and CP-CML versus BC-CML), correlation analysis was performed. The CP-CML versus CP-BC analysis was excluded as no CpGs were significant with a FDR *P*-value < 0.05. Interestingly, two-thirds (1,998/3,014, 66.3%) of the significant CpGs from the CP-CML versus BC-CML analysis were also identified as differentially methylated between CP-CML and normal individuals (**Figure 6.13**). This also identified a significant CpG list, specific for each analysis, as 9,599 CpGs were differentially methylated only when CP-CML patients were compared to normal individuals, and 1,016 CpGs were specific for the CP-CML versus BC-CML analysis. Furthermore, the analysis from **Chapter 5**, **Section 5.2.10**, where CP-CML patients classified as very low OA were compared to other OA, was also compared against the results generated in this chapter.

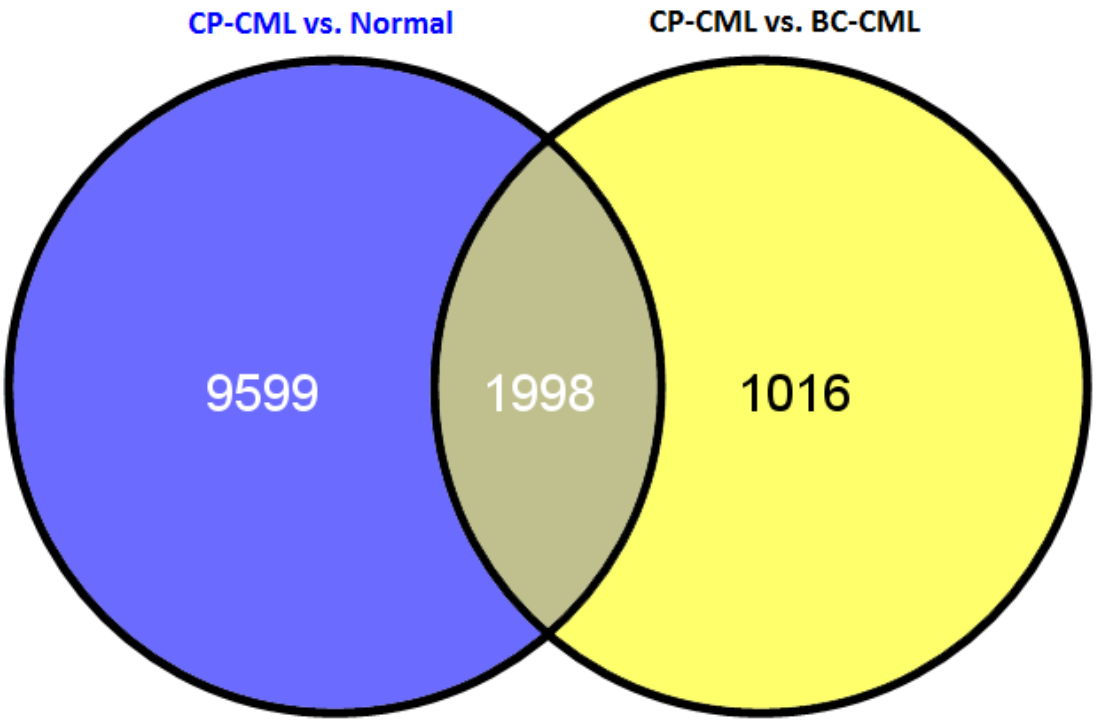


Figure 6.13 Significant overlap was observed between the significant CpGs identified from the CP-CML versus normal individuals and the CP-CML versus BC-CML analyses

Venn diagram demonstrating the overlap between the significant CpGs identified from the CP-CML versus normal individuals analysis (**Blue**), compared to the CP-CML versus BC-CML patient analysis (**Yellow**). The Venn diagram was developed using Venny (http://bioinfogp.cnb.csic.es/tools/venny/).[313]

Twenty CpGs were identified as differentially methylated between all 3 analyses (**Figure 6.14**; **Table 6.6**). The majority (11/20) of these CpGs are located in the gene body region. Interestingly, a common methylation pattern was observed. CpGs that were significantly hypermethylated in CP-CML patients compared to normal individuals, were also found to be significantly hypermethylated in CP-CML patients compared to BC-CML patients, and were hypermethylated in patients with other OA compared to very low OA. Furthermore, the exact opposite was also true, with CpGs significantly hypomethylated in CP-CML patients compared to normal individuals, also hypomethylated in CP-CML patients compared to BC-CML patients, and hypomethylated in patients with other OA compared to very low OA. Interestingly, the observed common methylation relationship between very low OA and BC (**Table 6.6**; i.e. CpGs hypermethylated in BC, were also hypermethylated in very low OA) for this subgroup of 20 CpGs suggests that CP-CML patients with very low OA may have a similar epigenetic profile to BC. This relationship was also observed with the sole other CpG (cg01761968) common to both the CP-CML vs. BC-CML and very low vs. other OA analyses (**Figure 6.14**). However, it is unlikely to extend to all CpGs, as the overall methylation patterns observed were opposite, with the majority of CpGs hypomethylated in very low OA, while global hypermethylation was associated with BC.

Taken together, these data suggest that for these common, differentially methylated CpGs, patients with poor prognosis (very low OA) had methylation profiles similar to BC-CML patients. Confounding this however, was the observation that these patients also had methylation profiles more similar to normal individuals than CP-CML patients. These findings indicate the complexity of such analyses and may allude to the fact that this subset of CpGs is distinguishing some other common feature of the samples, apart from low OA and BC. As such, further investigation into these results is now a part of ongoing studies.

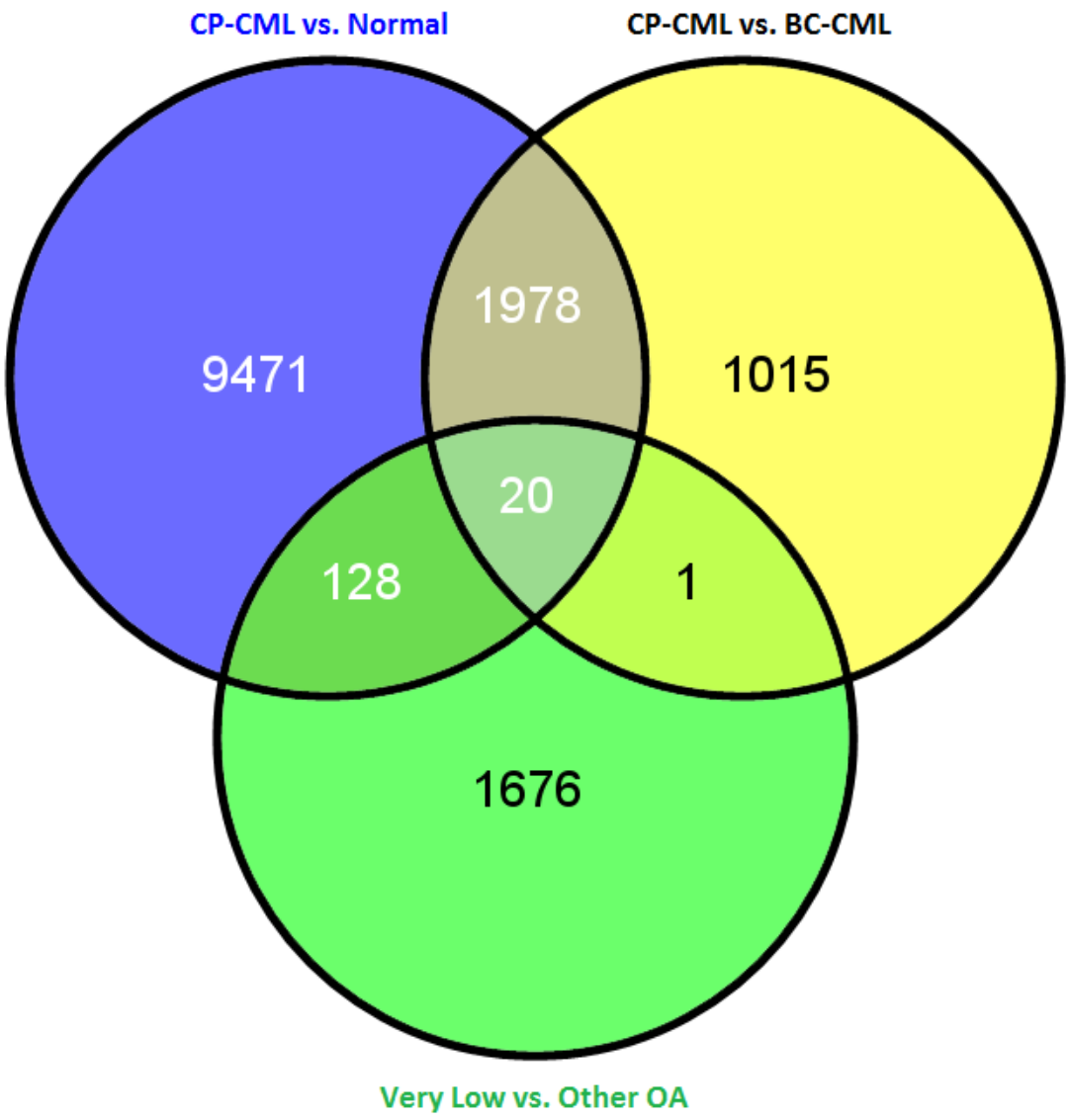


Figure 6.14 Significant overlap was observed between the significant CpGs identified from the CP-CML versus normal individuals, the CP-CML versus BC-CML and the very low versus other OA analyses

Venn diagram demonstrating the overlap between the significant CpGs identified from the CP-CML versus normal individuals analysis (**Blue**), compared to the CP-CML versus BC-CML patient analysis (**Yellow**), and the very low versus other OA analysis (**Green**). The Venn diagram was developed using Venny (http://bioinfogp.cnb.csic.es/tools/venny/).[313]

Table 6.6 Summary of the 20 CpGs common between the 3 different DNA methylation analyses

CpG	Gene	Chr	Gene location	Regional location	CML vs. Normal	CP vs. BC	Very low vs. other OA
cg00344445	*FLI1*	11	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg00511167	*CEP63*	3	S_Shore	5'UTR	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg00754604	*RBMS1*	2	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg01040749	*INPP5A*	10	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg01660999	*ANGPT1*	8	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg02222791	NA	14	NA	NA	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg02722672	*GRK4*	4	S_Shore	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg04431002	*USP10*	16	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg06152215	*CCDC92*	12	S_Shore	Body	Hyper in CML	Hyper in CP	Hyper in Other OA
cg08165361	*SPIDR*	8	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg08708961	*PSEN2*	1	NA	Body	Hyper in CML	Hyper in CP	Hyper in Other OA
cg12127991	*RHOU*	1	N_Shore	TSS1500	Hyper in Normal	Hyper in BC	Hyper in VL OA

cg12209946	NA	3	N_Shore	NA	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg15289427	*FAM65B*	6	Island	5'UTR	Hyper in CML	Hyper in CP	Hyper in Other OA
cg15520845	*NUCKS1*	1	NA	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg18325315	*GTF2IRD1*	7	S_Shelf	5'UTR	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg19037107	*ACAD8*	11	S_Shelf	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg20384132	*RANBP3L*	5	NA	TSS1500	Hyper in CML	Hyper in CP	Hyper in Other OA
cg24235882	*CHIC2*	4	N_Shore	Body	Hyper in Normal	Hyper in BC	Hyper in VL OA
cg26861460	*PARVG*	22	NA	TSS1500	Hyper in Normal	Hyper in BC	Hyper in VL OA

"Hyper in…" indicates that the CpG is hypermethylated in that patient group, i.e. hyper in normal indicates that a CpG is hypermethylated in normal individuals compared to CP-CML patients. Abbreviations: Chr, chromosome; NA, non-annotated; VL OA, very low OCT-1 activity; TSS200, within 200 bp of the transcription start site (TSS); TSS1500; within 1,500 bp of the TSS; 5UTR, 5' untranslated region; and 3UTR, 3' untranslated region.

.

6.3 Discussion

While much is known about the genetic mechanisms involved in CML, to date, very little has been published on any epigenetic mechanisms of CML, particularly global DNA methylation. A better understanding of how DNA methylation affects not only CML as a disease entity, but also in the disease progression from CP to BC, may lead to new advances in the treatment of CML. Specifically, combination therapies involving TKIs plus other drugs targeting specific epigenetic pathways may aid in optimizing overall patient response. As discussed in **Chapter 5**, global DNA methylation analysis is poorly understood in CML. The findings described in this chapter sought to address this issue and strongly suggest that distinct, yet very complex, DNA methylation patterns are present in CP-CML, compared to normal individuals and also during disease progression from CP to BC.

These results provide further support for the selected use of DNA methylation inhibitors (DNMTis) in conjunction with TKIs, as a combination therapy for some patients with CML. Demethylating agents targeting DNA methyltransferase 1 (DNMT1), such as 5-aza-2'-deoxycitidine (decitabine), have been previously used to treat imatinib-resistant CML patients, mainly in the advanced phases (either accelerated phase or BC).[380,413,414] These studies demonstrated combination therapy with decitabine and imatinib to be well tolerated and active in advanced phase CML.[380] However, further studies evaluating low-dose, long-exposure decitabine treatment schedules, aimed towards improving treatment efficacy, are required.[413] Taken together, these results indicate that further investigation into DNMTis in CML is warranted.

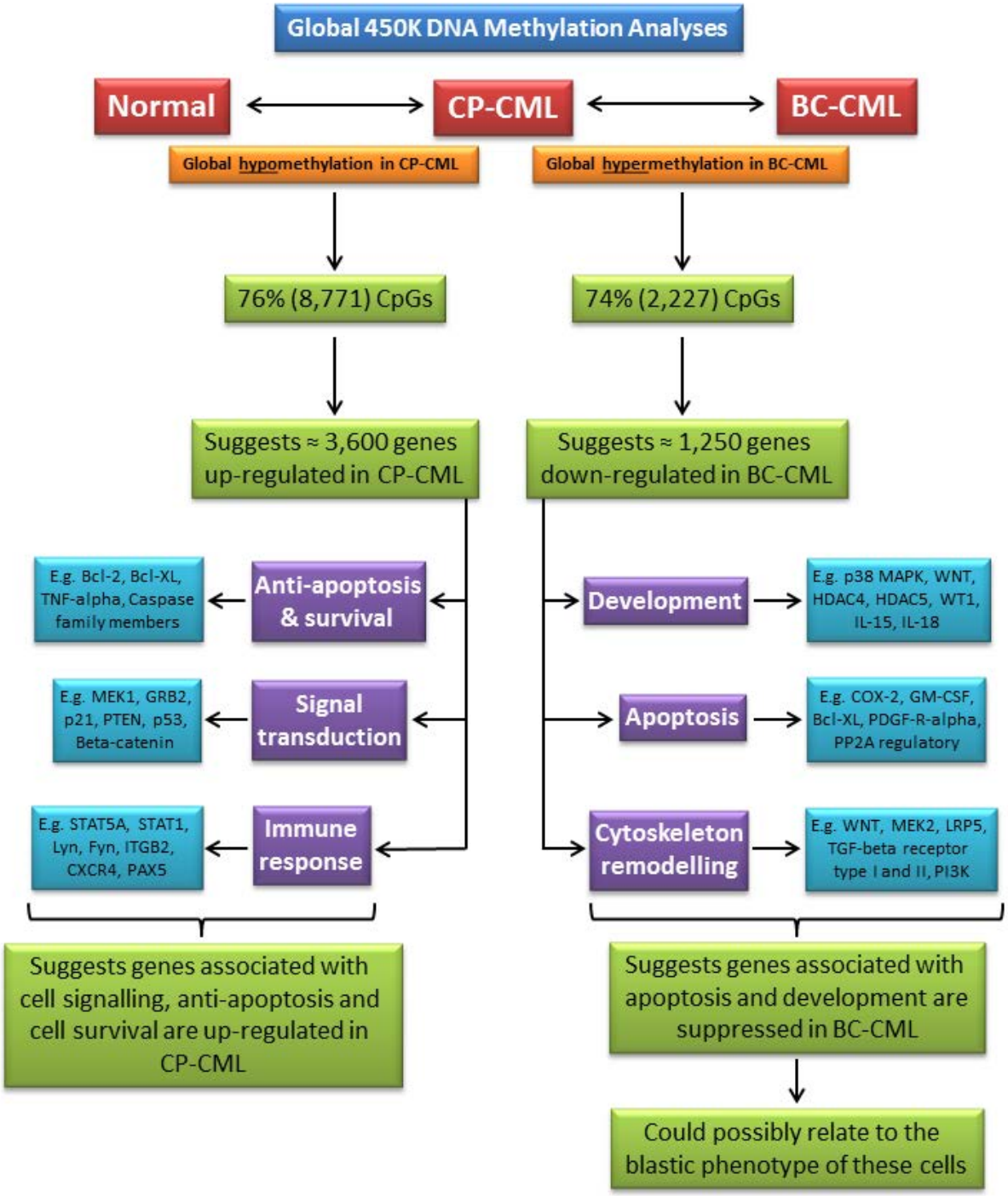
Pathway enrichment analysis identified common CML-associated pathways, including cell adhesion, apoptosis and TGF and WNT signal transduction were enriched with the gene-set

corresponding to the differentially methylated CpGs from the CP-CML versus normal analysis. Similar pathways (cell adhesion, apoptosis, immune response and cell proliferation), were also enriched from the CP-CML versus BC-CML dataset. Furthermore, significant enrichment of the Polycomb group genes (*EZH2*, *PRC2*, *EED* and *SUZ12*) was observed in the CML versus normal dataset, the CP-CML versus BC-CML dataset, and the OA analysis (very low versus other OA) performed in **Chapter 5**. Polycomb group proteins form two complexes, known as the polycomb-repressive complex 1 and 2 (PRC1 and PRC2), which repress target genes by chromatin remodelling. There is also increasing evidence that these complexes have a role in cancer progression and development by blocking cell differentiation and promoting stem cell self-renewal.[388,415-417] Additionally, PRC2 contains the EZH2, EED and SUZ12 polycomb proteins.[388] In CML, the Polycomb group gene *BMI1* has been demonstrated to be a prognostic marker for poor repsonse,[320] while SUZ12 has been suggested to be a target of the non-canonical WNT pathway and involved in CML disease progression.[418] Taken together, this data indicates that Polycomb group genes may not only have a significant role in CML disease overall, but also in progression to BC, possibly through an epigenetic mechanism. However, further investigation and validation of these results is ongoing.

The data presented in this chapter also demonstrates that global DNA hypomethylation occurs in CP-CML, compared to normal individuals; while global DNA hypermethylation occurs in BC-CML, compared to CP-CML. This data, in combination with the results from the pathway enrichment analysis, allows for speculation as to the effects of the global hypo- or hypermethylation of the corresponding genes from the different disease states (**Figure 6.15**). Global hypomethylation in CP-CML was associated with genes involved in anti-apoptosis and

Figure 6.15 Summary of the 450K methylation results from normal, CP-CML and BC-CML patient samples, indicating a prevalence of different gene pathways influenced by DNA methylation

A summary of the complex results from the CP-CML versus normal and the CP-CML versus BC-CML global DNA methylation analyses is presented; combining the significant CpG results with the pathway enrichment analyses performed using GeneGo and the corresponding gene lists. Global hypomethylation was observed in CP-CML, compared to normal individuals; while global hypermethylation was observed in BC-CML, compared to CP-CML.


Global 450K DNA Methylation Analyses
Normal
CP-CML
BC-CML
Global hypomethylation in CP-CML
Global hypermethylation in BC-CML
76% (8,771) CpGs
74% (2,227) CpGs
Suggests ≈ 3,600 genes up-regulated in CP-CML
Suggests ≈ 1,250 genes down-regulated in BC-CML
E.g. Bcl-2, Bcl-XL, TNF-alpha, Caspase family members
Anti-apoptosis & survival
E.g. MEK1, GRB2, p21, PTEN, p53, Beta-catenin
Signal transduction
E.g. STAT5A, STAT1, Lyn, Fyn, ITGB2, CXCR4, PAX5
Immune response
Development
E.g. p38 MAPK, WNT, HDAC4, HDAC5, WT1, IL-15, IL-18
Apoptosis
E.g. COX-2, GM-CSF, Bcl-XL, PDGF-R-alpha, PP2A regulatory
Cytoskeleton remodelling
E.g. WNT, MEK2, LRP5, TGF-beta receptor type I and II, PI3K
Suggests genes associated with cell signalling, anti-apoptosis and cell survival are up-regulated in CP-CML
Suggests genes associated with apoptosis and development are suppressed in BC-CML
Could possibly relate to the blastic phenotype of these cells

pro-survival, signal transduction and immune response, suggesting that these genes may be up-regulated in CP-CML, compared to normal individuals (**Figure 6.15**). Furthermore, global hypermethylation in BC-CML was associated with genes involved in development, apoptosis and cytoskeleton remodelling, indicating that these genes may be suppressed in BC-CML, compared to CP-CML (**Figure 6.15**). However, it is also important to note the different cellular phenotypes observed between these disease stages, with mature granulocytes observed in CP-CML while more immature blast cells are present in BC-CML. These phenotypic differences could be significantly influencing the different global methylation profiles observed; however, the only way to accurately determine this would be to use a homogenous cell population, such as isolated CD34+ cells, which requires a significant amount of patient sample. Unfortunately, limited patient sample availability prevented this analysis from being performed; however, these studies are planned as a part of future investigation.

Finally, the lack of differential DNA methylation observed between CP-CML and CP-BC (CP-CML patients who progress to BC during TKI therapy) patients was not expected, and indicates that epigenetic mechanisms are unlikely to be involved at diagnosis in this subset of patients. Importantly, only a small subset of CP-BC patients were analysed (n = 7). Therefore, if the differences in DNA methylation at diagnosis between CP-BC and CP-CML patients are only small, as this data suggests, a much larger patient cohort needs to be examined before the involvement of underlying epigenetic mechanisms are excluded from this patient cohort. It is of vital importance to identify these patients at diagnosis, allowing tailored therapeutic strategies to be administered, as these CP-BC patients are unlikely to respond well to normal therapeutic strategies. Hence, further gene expression analysis has been performed on a subset of these patients and is described in **Chapter 7**.

In conclusion, the findings detailed in this chapter strongly demonstrate that the specific global DNA methylation profile of CP-CML patients differs significantly from normal individuals, with global hypomethylation observed in CML. Furthermore, the global DNA methylation profile of CP-CML was also significantly different to BC-CML, with global hypermethylation observed in BC-CML. Therefore, these results, in combination with the results from **Chapter 5**, indicate that aberrant epigenetic regulation is likely to play a critical role in CML, during disease progression and in poor response to TKI therapy (very low OA).

7 IDENTIFICATION OF INDEPENDENT BIOMARKERS FOR OCT-1 ACTIVITY AND CML DISEASE PROGRESSION

7.1 Introduction

The term "biomarker" refers to a broad subcategory of medical indicators, which are able to be accurately and reproducibly measured. While many more precise definitions are available in the literature, the definition specified by the National Institutes of Health Biomarkers Definitions Working Group of "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention,"[419] provides an excellent summary. The use of biomarkers in clinical research is not new, with examples of biomarkers including a range of tests from blood pressure, to complex laboratory tests performed on blood or other tissues.[420] The aim of biomarker development is not only to identify ways to predict the efficacy of targeted therapies including response rate, progression-free survival (PFS) and overall survival (OS; **Figure 7.1**); but to provide valuable information for developing new drugs, specifically targeting the pathways associated with the biomarker.[421]

7.1.1 Biomarkers in Chronic Myeloid Leukaemia

In chronic myeloid leukaemia (CML), biomarkers are generally used to identify patients who are likely to respond to a certain treatment, such as tyrosine kinase inhibitors (TKIs), or patients who are at risk of disease progression. Effective biomarkers have become increasingly important due to the introduction of TKI therapy, as the outcome of chronic phase CML (CP-CML) has now altered from a short life expectancy, to long-term survival.[421] Additionally, as three TKIs (imatinib, nilotinib and dasatinib) are now available for front line therapy in CP-CML, and two more TKIs (ponatinib and bosutinib) are available for second- and third-line therapy, it is important to identify which agent should be used, and when, to optimize long-term patient outcomes. A number of prognostic biomarkers have been identified in CML patients, including certain clinical characteristics at diagnosis (e.g. Sokal

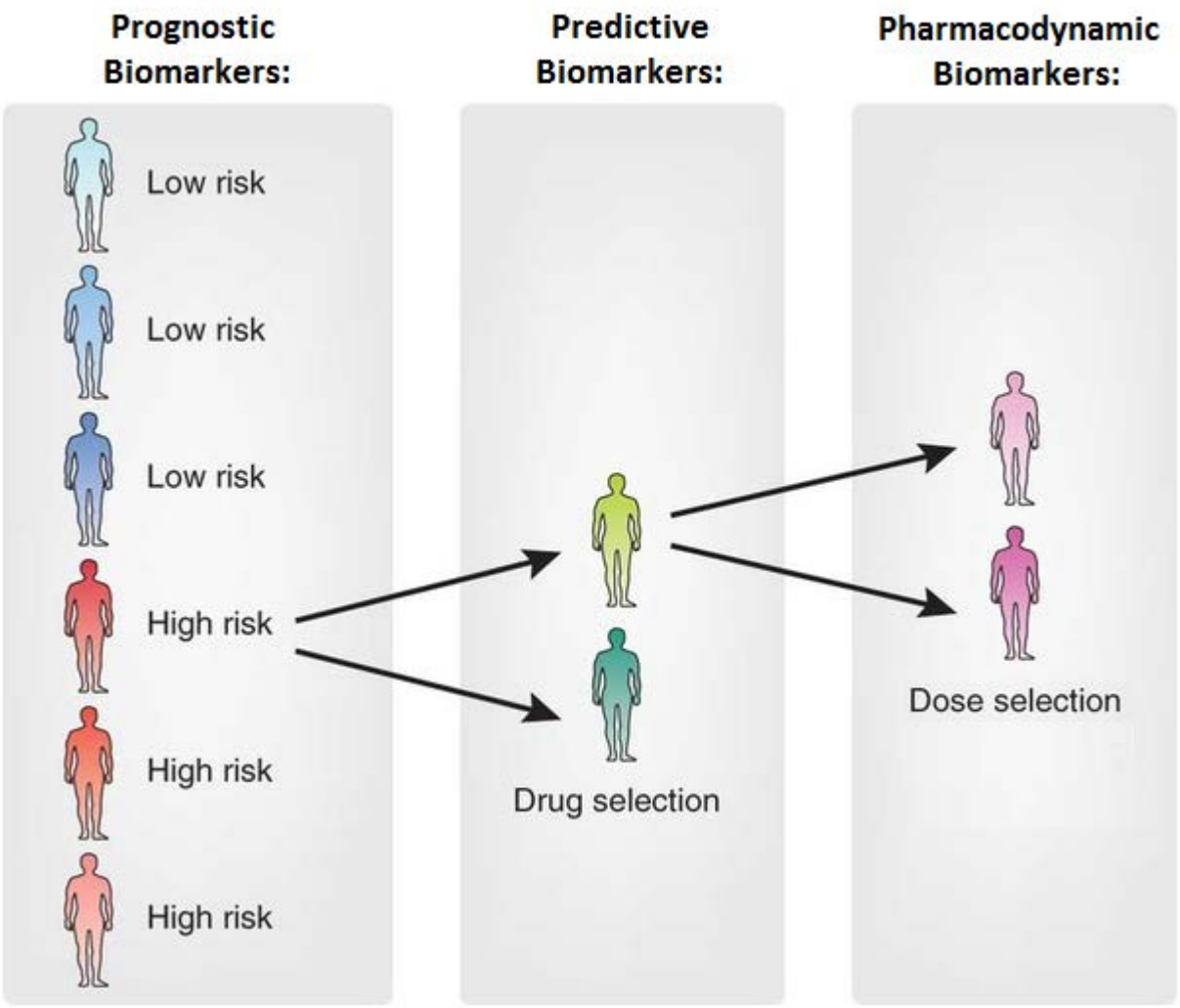


Figure 7.1 Example of the implementation of different biomarkers in the treatment of cancer

Prognostic biomarkers are able to identify patients who are at risk of disease recurrence or progression on standard therapy, and should receive tailored therapy. Predictive biomarkers help to identify the drug which a patient is most likely to be responsive (or unresponsive) to. Pharmacodynamic biomarkers are useful to identify the best drug dose for an individual patient. For example, a prognostic biomarker identifies high from low risk patients, then a predictive biomarker identifies the most suitable therapy for the high risk patients, and a pharmacodynamic biomarker determines the optimal dosage. The utilisation of biomarkers for drug and dose selection occurs in unison for the low risk patients. Figure adapted from Majewski *et al.* 2011.[422]

score); *BCR-ABL1* kinase activity; human Organic Cation Transporter 1 (OCT-1) mRNA expression, polymorphisms and activity (known as OCT-1 activity [OA]); and additional molecular changes, including early molecular response (EMR); as previously discussed in **Chapter 1**. The prognostic ability of the OA assay is of significant importance, as low OA (OA < 7.2 ng/200,000 cells) can identify at diagnosis CP-CML patients at risk of poor molecular response and significantly lower event-free and overall survival.[142,151] Very low OA (OA < 4 ng/200,000 cells) also predicts for a subset of patients with a significantly increased risk of disease progression.[142] However, as the OA assay is not easily performed in diagnostic laboratories, there is a need to identify new biomarkers to either replace the assay itself, or to predict the same clinical endpoints with similar "power".

Many groups have used microarray gene expression profiling to identify specific genes at diagnosis that can either predict response to TKI therapy, or progression to blast crisis (BC); as previously discussed in **Chapter 4**. Radich *et al.*[180] investigated gene expression changes associated with disease progression to accelerated phase (AP) and BC, and identified components of the WNT/β-catenin pathway, the transcription factors *JUN-B* and *FOS* and the marker *PRAME*, as being associated with advanced phase CML. Oehler *et al.*[178] further interrogated this data, identifying a six-gene signature (*NOB1*, *DDX47*, *IGSF2*, *LTB4R*, *SCARB1*, and *SLC25A3*) that discriminated early CP from late CP, CP from AP, and CP from BC; however, this signature is yet to be independently validated.

Furthermore, specific approaches, targeting known genes important in cell survival, proliferation and differentiation, have also identified biomarkers in CML. Wilms tumour 1 (*WT1*) is a tumour suppressor gene normally expressed at very low levels, but is significantly elevated in several chronic and acute leukaemias.[423] Cilloni *et al.*[424] investigated a cohort of

36 CML patients for *WT1* expression and found that higher WT1 mRNA levels were observed at diagnosis in imatinib non-responders, compared to responders. Furthermore, a short *in vitro* incubation of patient cells with imatinib reduced elevated *WT1* in patients who responded to therapy, but not in patients with suboptimal response. The polycomb group gene *BMI1* is able to regulate proliferation and self-renewal in both normal and leukaemic cells.[425] *BMI1* expression at diagnosis was significantly increased in AP-CML, compared to CP-CML, and was demonstrated to correlate with time to transformation to BC.[320] Low *BMI1* expression, in combination with high proteinase-3 (*PR-3*) expression was also associated with improved overall survival. BMI1 protein expression, as measured by flow cytometry, was also demonstrated to increase during disease progression,[426] further demonstrating the potential of BMI1 as a biomarker in CML. Additionally, recent work has identified that BMI1 cooperates with *BCR-ABL1* during leukaemic transformation,[427,428] validating the biological role of this protein as a biomarker for CML progression. The cancerous inhibitor of protein phosphatase 2A (*CIP2A*) has recently been described as another biomarker of disease progression in CML. Lucas *et al.*[429] identified that CIP2A protein levels were significantly higher at diagnosis, in patients who later progressed to BC on TKI therapy, than in patients who remained in CP. While these studies have all identified various biomarkers associated with either CML disease progression and/or poor response, validation in large, independent multicentre patient cohorts is currently lacking. Here we further investigate biomarkers predictive of CML disease progression, and attempt to identify specific biomarkers for OA, to potentially replace the OA assay itself.

7.1.2 Approach

The histone gene signature identified from the original OA microarray data (**Chapter 4**) was compared against the Molecular Signatures Database (**Section 2.6.4**), with further analysis

performed using the Oncomine database (**Section 2.6.13**). *NPM1* mutation analysis (**Section 2.4.9**) was performed on DNA extracted from a cohort of 14 BC-CML and 33 CP-CML patients (**Section 2.4.7**). Re-analysis of the TLDA RQ-PCR data discussed in **Chapter 4** using HTqPCR (**Section 2.6.7**) was performed, identifying significant differential *GFI1* expression. *GFI1* expression was then assessed in MNCs using microarray analysis (**Section 2.5.13**) and in CD34+ cells by re-analysing the McWeeney *et al.*[179] dataset (**Section 2.6.5**). Furthermore, the CML disease progression dataset of Radich *et al.*[180] was also interrogated (**Section 2.6.5**). Total white cell (TWC) RNA (**Sections 2.4.4**) from CP-CML patients treated with imatinib for 7 days was then analysed for *GFI1* expression. Next, the presence of the *GFI1*^{36N} SNP in 32 CP-CML patients was assessed by SNP genotyping (**Section 2.4.12.2**). Finally, imatinib- (K562 2μM IM1 and IM2) and nilotinib-resistant (K562 2μM NIL and K562 DOX 2[1] and 2[2] NIL) cell lines were assessed for *MLLT4* expression (**Section 2.4.6**). The analyses discussed in this chapter were used to identify possible biomarkers (genes) in CP-CML, potentially enabling personalised therapeutic strategies to be developed for CP-CML patients at diagnosis, thus, optimizing patient outcome. However, this is reliant on independent validation to confirm the accuracy of these potential biomarkers.

7.2 Results

7.2.1 Pathway analysis identified RNA polymerase I pathways as significantly enriched for the histone gene signature identified from RQ-PCR analysis

As described in **Chapter 4**, microarray analysis was performed to identify if differential gene expression profiles were present between CP-CML patients with very low and very high OA. Significant differences were observed and subsequently validated by RQ-PCR analysis. Interestingly, this analysis also identified a significant up-regulation of histone gene expression associated with very low OA. Furthermore, a meta-analysis of published CML microarray datasets investigating CML response to TKI therapy, or disease progression, was performed which identified additional histone genes were up-regulated in association with the poor responder groups. This data was unexpected, and although transcription factor binding analysis identified enrichment for the transcription factors *MYC*, *E2F1* and *E2F4*, subsequent protein analysis was unable to confirm this link (refer **Section 4.2.15**). Therefore, to further understand the pathways involved with and the regulation of this histone gene-set, investigation of the Molecular Signatures Database (MSigDB) was performed.

MSigDB analysis was performed using the complete 25 histone gene-set, developed from the meta-analysis (**Table 7.1 [A]**), and the 10 histone gene-set, containing only the histones identified from the very low versus very high OA analysis (**Table 7.1 [B]**), so that the OA association was not confounded by the results of the other studies. However, both analyses identified very similar results, with significant enrichment for gene-sets associated with transcription, in particular RNA polymerase (Pol) I, identified. In retrospect, this result was somewhat expected, as histones play a crucial role in transcription, through regulation of the chromatin state. Interestingly, MSigDB analysis of the 25 histone gene-set also identified significant enrichment of 4 genes associated with the Verhaak *et al.*[430] AML with mutated

Table 7.1 Common functional gene-sets enriched for the histone gene-set by MSigDB analysis

Table **(A)** lists pathway hits from the Molecular Signatures Database (MSigDB) using all 25 histone genes identified from a meta-analysis of published CML microarray data. Table **(B)** lists hits from the MSigDB for only the 10 histone genes identified from the very low versus very high OA microarray analysis. The same pathways were identified from both analyses. REACTOME is an open-source, open access and peer-reviewed pathway database incorporated in the MSigDB,[431] from which a number of enriched gene-sets originated. Abbreviations: FDR – false discovery rate, Pol – polymerase, AML – Acute Myeloid Leukaemia, NPM1 – nucleophosmin 1. *Gene-set size refers to the number of genes overlapping between the gene-sets. The FDR *q*-value is adjusted for multiple testing and provides a far superior indication of significance, as a *q*-value of 0.05 implies that 5% of significant tests will result in false positives; unlike a normal *P*-value, which applies to all tests.

(A)

	Gene-set Size*	FDR *q*-value	*P*-value
Transcription			
***REACTOME* – RNA Pol I, RNA Pol III and mitochondrial transcription**	19	< 0.001	< 0.001
***REACTOME* – Transcription**	19	< 0.001	< 0.001
***REACTOME* – RNA Pol I transcription**	19	< 0.001	< 0.001
***REACTOME* – RNA Pol I promoter opening**	19	< 0.001	< 0.001
AML			
***VERHAAK* – AML with NPM1 mutated Up**	4	< 0.001	< 0.001

(B)

	Gene-set Size*	FDR *q*-value	*P*-value
Transcription			
***REACTOME* – RNA Pol I, RNA Pol III and mitochondrial transcription**	7	< 0.001	< 0.001
***REACTOME* – Transcription**	7	< 0.001	< 0.001
***REACTOME* – RNA Pol I transcription**	7	< 0.001	< 0.001
***REACTOME* – RNA Pol I promoter opening**	7	< 0.001	< 0.001

NPM1 (nucleophosmin 1) gene-set (**Table 7.1 [A]**), indicating a possible association between histone expression and *NPM1* mutation.

7.2.2 Nucleophosmin 1 mutations are associated with increased histone expression and RNA Pol I pathways

Acute myeloid leukaemia (AML) is a heterogeneous disease with diverse genetic abnormalities, emphasized by multiple chromosomal translocations and inversions. NPM1 is a nucleolar phosphoprotein, known to shuttle between the nucleus and cytoplasm during different stages of the cell cycle and has a diverse range of cellular functions.[432] Mutations in *NPM1* (localised to 5q32) are the second most frequent mutations, following *FLT3* mutations, observed in AML,[433] and are associated with distinctive transcription signatures and a reduced ability of NPM1 to shuttle between the nucleolus, nucleus and cytoplasm.[434] *NPM1* mutations are associated with significantly better overall survival (OS) and event-free survival (EFS) in intermediate cytogenetic risk AML without *FLT3-ITD* mutations (normal karyotype), compared to patients without *NPM1* mutations, while multivariate analysis also identified mutant *NPM1* as an independent prognostic marker in AML.[430] Intriguingly, NPM1 is active in several stages of ribosome biogenesis, including facilitating the interaction between ribosomal proteins and rRNA,[435] which is transcribed by RNA Pol I. As such, RNA Pol I transcription has been demonstrated to be stimulated by NPM1.[436] Taken together, these data suggest a supporting role for *NPM1* in RNA Pol I transcription.

Analysis of the Oncomine cancer microarray database and integrated data-mining platform[239,240] was then performed, and identified that up-regulated histone gene expression correlated with *NPM1* mutations in AML samples from the Wouters *et al.*[437] dataset (**Figure 7.2**). Hence, providing further evidence of a possible role for *NPM1* in the up-regulation of histone gene expression observed in very low OA CML patients.

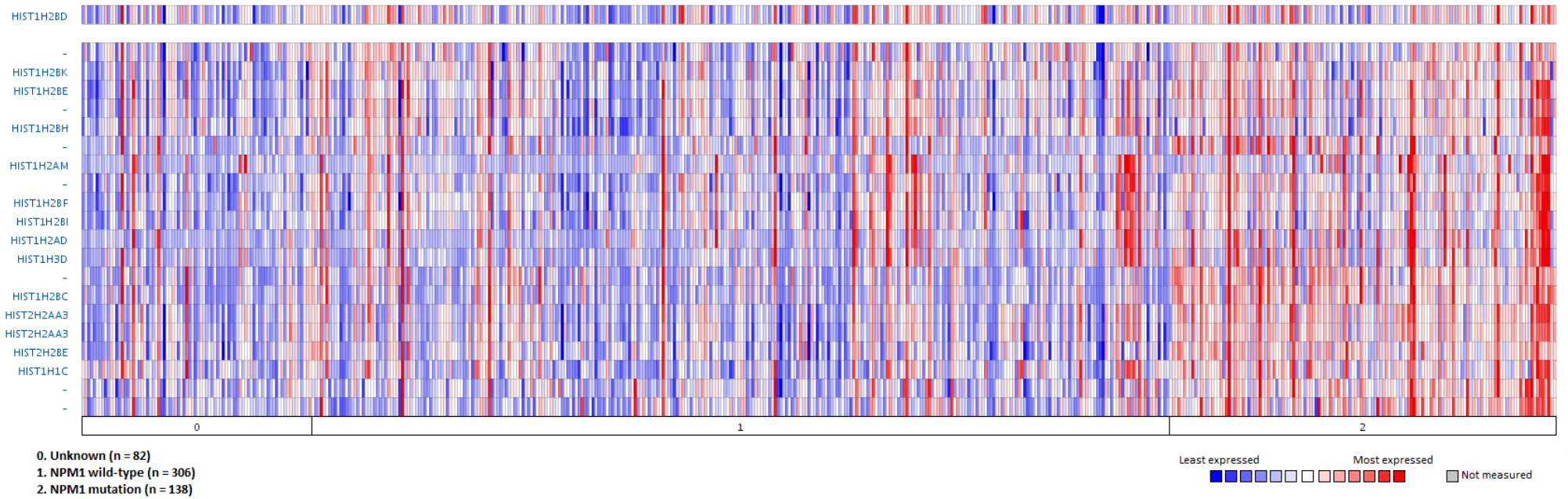


Figure 7.2 Histone gene expression is up-regulated in AML patients with *NPM1* mutations

Using Oncomine, the co-expression of histone family genes with *NPM1* mutation status in AML patients was assessed. Enrichment of histone family genes was identified as associated with *NPM1* mutations, compared to *NPM1* wild-type AML patients. Log2 median-centred expression intensity is shown, with **red** bars representing genes with the greatest expression, while **blue** bars represent the lowest expression.

7.2.3 NPM1 mutations are not associated with very low OA, and occur rarely or not at all in CML patients in chronic phase or blast crisis

Recently, Georgiou *et al.*[438] reported the case of a CML patient who achieved complete cytogenetic and molecular response on 400 mg imatinib daily, and subsequently developed AML with normal karyotype and *NPM1*^{mut}/*FLT3-ITD*^{neg}. Interestingly, subsequent testing demonstrated that the *NPM1* mutation was present at low levels 12 months prior to AML diagnosis, suggesting a possible role in the transformation of CML. *NPM1* mutations have also been reported in Ph+ AML,[439] indicating an ability to cooperate with *BCR-ABL1* in leukaemic transformation. One case of *NPM1*^{mut} myeloid BC-CML has also previously been reported in the literature.[440] In this particular case, the *NPM1* mutation did appear to function in the transformation of BC-CML, as the mutation was found in conjunction with the *BCR-ABL1* rearrangement in almost all of the blastic leukaemic cells. However, other studies have failed to definitively show *NPM1* mutations in CML (both CP-[434,441] and BC-CML[439,442]) using various techniques, but CP-CML patients at risk of disease progression have not been fully interrogated. Therefore, to investigate whether *NPM1* mutations have a role in BC-CML or early in the progression of CP-CML, and in particular whether such mutations may be associated with poor-risk CP-CML, we determined the *NPM1* mutation status of a cohort of BC-CML patients and CP-CML patients, including a subgroup identified by very low OCT-1 activity.

Initially, 14 BC-CML patients were analysed for the presence of *NPM1* mutations using a fluorescence-based PCR assay, capillary electrophoresis (ABI Prism 3100 Genetic Analyzer, Applied Biosystems) and GeneScan™ analysis (Applied Biosystems), as previously described.[198] Of these patients, 5/14 were diagnosed in blast crisis, 5 progressed from CP-CML and 4 were referred from other Centres in blast crisis. A single set of primers was used to screen for the most common *NPM1* mutations (a four-nucleotide insertion located in

nt959 or 965 of exon 12, covering all known NPM1 mutations),[434] producing a 232 bp product for wild-type *NPM1* or 236 bp product for mutant *NPM1 (NPM1^{mut}*; **Appendix 11**). No *NPM1* mutations were observed in the 14 BC-CML patients analysed using this assay, consistent with the results of previous studies.[439,442]

Next, 33 CP-CML patients were screened to determine their *NPM1* mutation status, including 17 poor-risk patients defined by very low OA. Once again we detected no *NPM1* mutations in any of the CP-CML patients, consistent with previous small-cohort CP-CML studies.[434,441] Included in this CP-CML cohort was a single patient who progressed from CP- to BC-CML four months after starting imatinib therapy. Matched samples at diagnosis and BC progression were available, and tested, with this patient remaining *NPM1^{mut}* negative at both time points. Therefore, this data lends further support to the conclusion that *NPM1* mutations are unlikely to play a role in CML disease initiation or progression to BC.

7.2.4 Low GFI1 expression in TWCs at diagnosis is strongly associated with subsequent blastic transformation

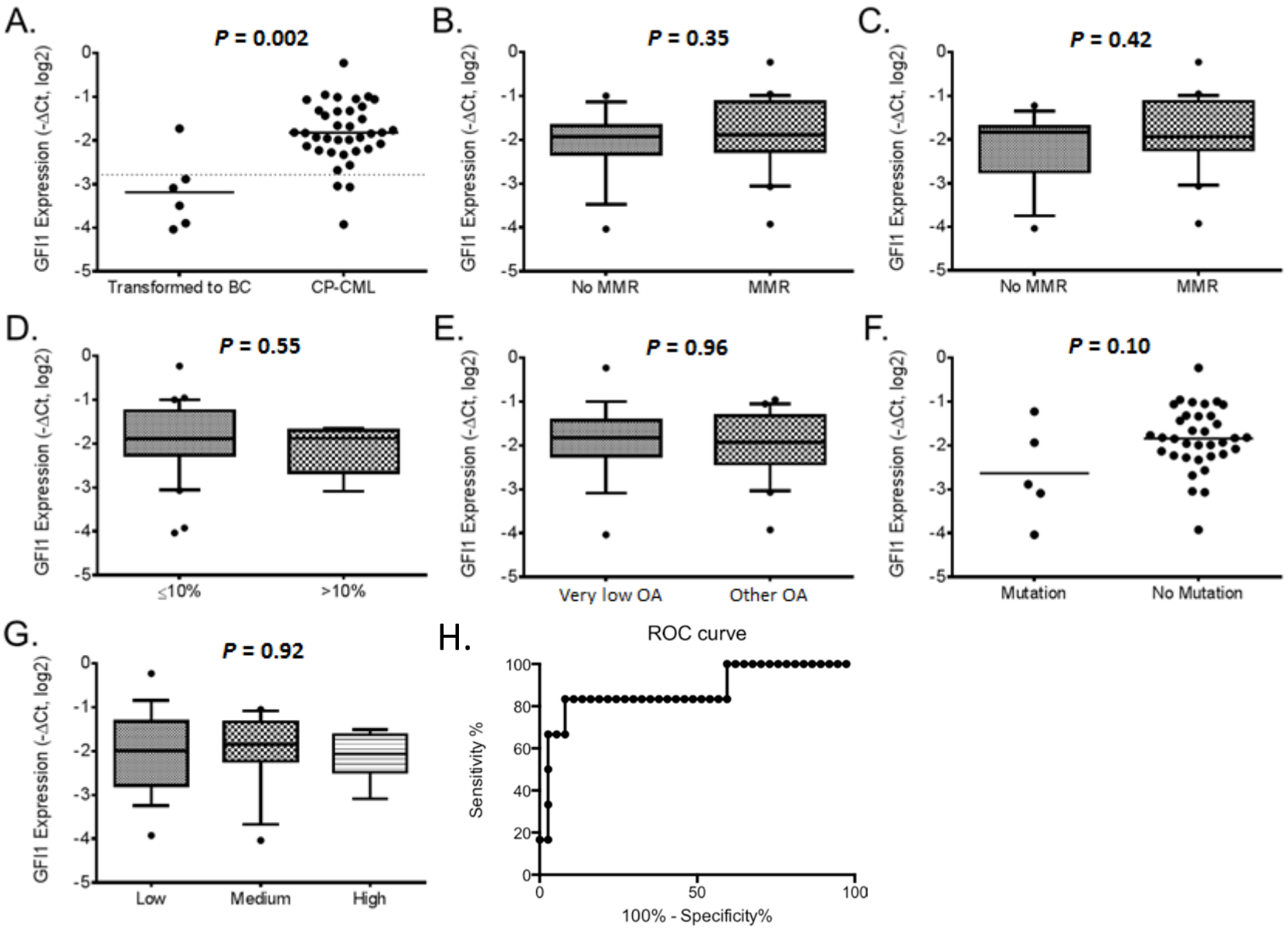
Growth factor independence 1 (*GFI1*) is a transcription factor with a crucial role in haematopoiesis, including preserving haematopoietic stem cell (HSC) quiescence and enhancing granulocytic differentiation, but is not required for inducing myeloid differentiation in p210BCR/ABL1-transformed cells.[443] Recently, Soliera *et al*.[444] demonstrated that ectopic *GFI1* expression inhibited proliferation and colony formation both in p210BCR/ABL expressing cell lines and in primary CD34+ CML cells through the repression of *STAT5B* and/or *Mcl-1*. This study, along with their previous work[443] demonstrated the biological importance of the *GFI1*/*STAT5B*/*Mcl-1* regulatory pathway on proliferation and survival of CML cells. However, the association between *GFI1* expression and good response to imatinib therapy, be it major molecular response (MMR; BCR-ABL1

mRNA levels < 0.1%) or early molecular response (EMR; 3 month BCR-ABL1 mRNA levels ≤ 10%) in *de novo* CP-CML patients was not examined. Intriguingly, when the data previously generated from the 14 patient TWC OA microarray analysis (refer **Section 4.2.1**) was re-analysed, significant differential *GFI1* expression between patients from the very low and very high OA groups (P = 0.002, FDR q-value = 0.17, logFC = 0.81), was observed. Therefore, based on the findings of Soliera *et al.,*[444] it was hypothesized that increased *GFI1* expression would be associated with a favourable outcome in CP-CML patients treated with imatinib.

The TLDA RQ-PCR data, previously generated in **Section 4.2.3**, using RNA isolated from the TWCs of 40 *de novo* CP-CML patients, plus three additional *de novo* CP-CML patients, was analysed using HTqPCR. As *GFI1* is primarily expressed in granulocytes,[445] we rationalized that the use of TWCs would enable accurate representation of total *GFI1* expression from each patient. As before, *GFI1* expression was normalised to the reference gene *GUSB*. The differential expression of *GFI1* at diagnosis, in those patients who transformed to BC (time range: 3.7-13 months, n = 6) while receiving imatinib as front-line therapy, was examined first (**Figure 7.3 [A]**). Significantly lower *GFI1* expression was observed at diagnosis in patients who transformed to BC on imatinib therapy, compared to those who did not (n = 37, P = 0.002), suggesting that lower *GFI1* expression is associated with transformation in CML. Therefore, *GFI1* was identified as a probable biomarker of CML patients at risk of disease progression. Furthermore, receiver-operator characteristic (ROC) analysis revealed that *de novo* CP-CML patients with *GFI1* expression < -2.782 (-ΔCt, log2; 92% specificity and 83% sensitivity; 87% power, P = 0.0036, likelihood ratio = 10.28) are approximately 10 times more likely to transform to BC, compared to CP-CML patients with *GFI1* expression > -2.782 (-ΔCt, log2) at diagnosis (**Figure 7.3 [H]**).

Figure 7.3 Examination of *GFI1* expression using CP-CML TWC mRNA

TLDA RQ-PCR was performed for *GFI1* expression on TWCs from 43 CP-CML patients at diagnosis. The assay was performed in duplicate and normalized to the housekeeping gene *GUSB*. The average of the duplicate assays was used for data presentation and statistical analysis. The differential expression of *GFI1* was examined for: **A.** Transformation to blast crisis (BC), **B.** MMR outcome by 12 months (including 12 months), **C.** MMR outcome by 24 months (including 24 months), **D.** 3-month BCR-ABL1 mRNA level (EMR), **E.** OCT-1 activity (OA), and **F.** Kinase domain (KD) mutation development during imatinib therapy. **G.** Differential *GFI1* expression between low, medium and high Sokal score[128] groups (one-way ANOVA). **H.** ROC curve for predicting patients likely to transform to BC, compared to CP-CML patients. All statistical analysis were performed using the Mann-Whitney U-test, unless otherwise stated. Box plots indicate the interquartile range (25 – 75 %) around the median. Whiskers represent the 10th and 90th percentiles. Very low OA – OA < 4 ng/200,000 cells; and Other OA – OA > 4 ng/200,000 cells. Figure from Kok, C.H. and Watkins, D.B. *et al.* 2013.[446]


A.
P = 0.002
GFI1 Expression (-ΔCt, log2)
Transformed to BC
CP-CML
B.
P = 0.35
No MMR
MMR
C.
P = 0.42
No MMR
MMR
D.
P = 0.55
≤10%
>10%
E.
P = 0.96
Very low OA
Other OA
F.
P = 0.10
Mutation
No Mutation
G.
P = 0.92
Low
Medium
High
H.
ROC curve
Sensitivity %
100% - Specificity%

The expression of *GFI1* in CP-CML patients who did, or did not, achieve MMR by 12 months was examined next. Intriguingly, there was no statistically significant difference observed in *GFI1* expression (P = 0.35; **Figure 7.3 [B]**) between these patients, nor when the response endpoint was extended to 24 months (**Figure 7.3 [C]**). *GFI1* expression was then assessed to determine if it was a biomarker for the achievement of EMR, a time-point recently demonstrated to be predictive of longer term response in imatinib-treated patients.[191] However, again no significant difference was observed in *GFI1* expression between patients who did, or did not achieve EMR (**Figure 7.3 [D]**). Furthermore, the initial differential *GFI1* expression between very low and very high OA patients observed from the OA microarray was unable to be verified with the TLDA RQ-PCR analysis (P = 0.96; **Figure 7.3 [E]**), suggesting *GFI1* expression is not associated with OCT-1 activity.

Point mutations in the kinase domain (KD) of *BCR-ABL1* are the most frequently described mechanism of acquired resistance to kinase inhibitor therapy in CML patients.[447] Therefore, the relationship between mutation development and *GFI1* expression in the cohort of patients with mutations (n = 4, within 24 months) was assessed; however, no significant association was observed (**Figure 7.3 [F]**). Finally, patients were grouped according to their Sokal score (low, medium or high), as previously defined,[128] but again no significant difference with respect to *GFI1* expression was observed between these groups (**Figure 7.3 [G]**).

7.2.5 GFI1 expression in MNCs or CD34+ cells is not associated with CP-CML patient response

To determine if the lack of correlation between *GFI1* expression and CP-CML patient response, apart from the significant association with blastic transformation, was due to the cell type analysed by RQ-PCR, we examined *GFI1* expression in MNCs from 93 *de-novo* CP-

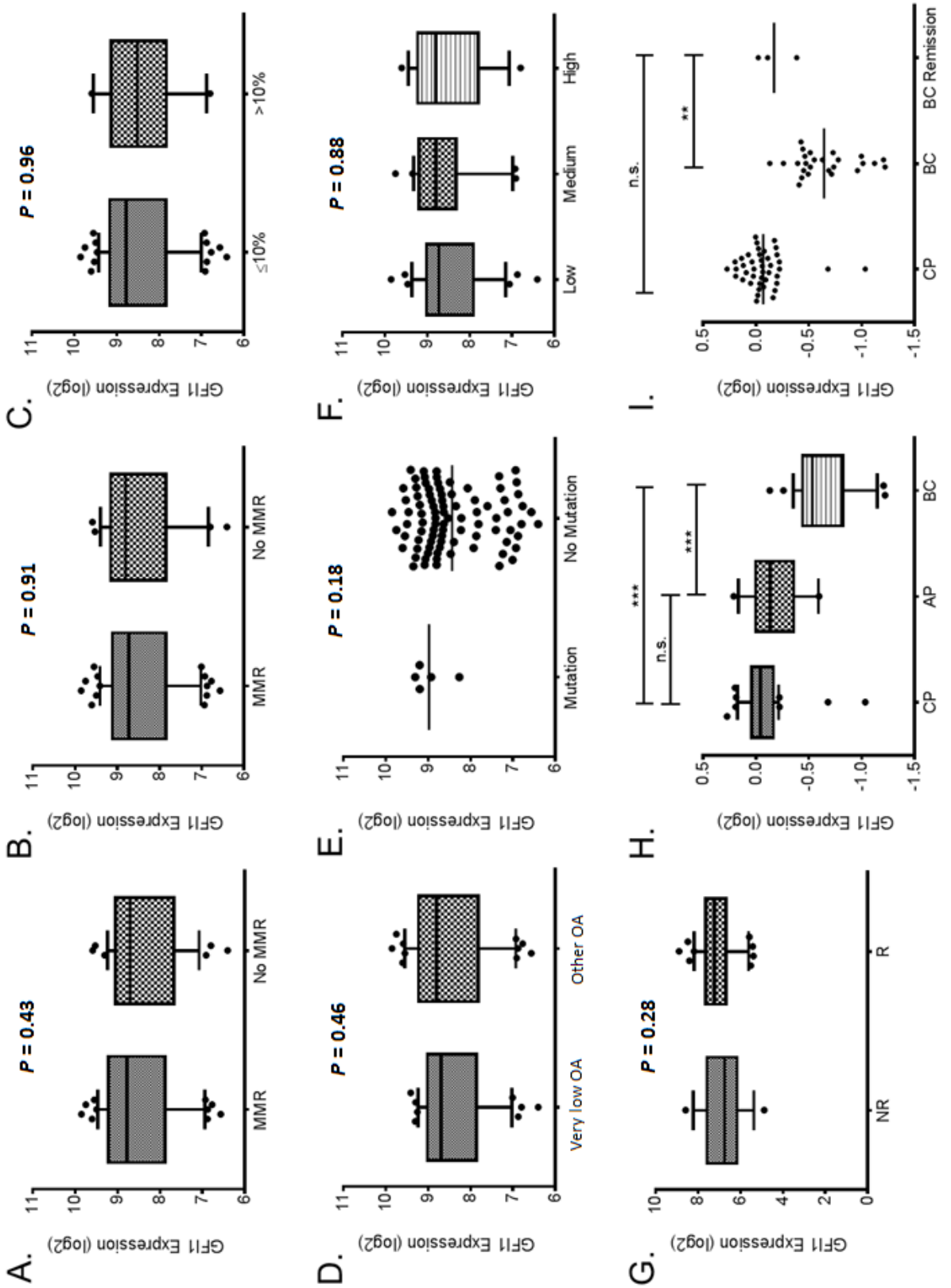
CML patients. There was no significant difference in *GFI1* expression between patients who achieved MMR by 12 months (P = 0.43; **Figure 7.4 [A]**), or by 24 months (P = 0.91; **Figure 7.4 [B]**), compared to those who did not. Again, as seen in the analysis of TWCs, no significant association between *GFI1* expression and the achievement of EMR (P = 0.96; **Figure 7.4 [C]**), OCT-1 activity (P = 0.46; **Figure 7.4 [D]**), *BCR-ABL1* kinase domain mutation development (P = 0.18; **Figure 7.4 [E]**), or Sokal score (P = 0.88; **Figure 7.4 [F]**) was observed. Furthermore, as only 1 patient transformed to BC on TKI therapy in this 93 patient cohort, the association between *GFI1* expression at diagnosis and BC transformation could not be determined in MNC samples.

As *GFI1* has a known role in haematopoietic stem cell maintenance, and the biological importance of *GFI1* was originally determined in CD34+ cells by Soliera *et al*.,[444] *GFI1* expression in CD34+ cells from CP-CML patients was examined. To do this, the McWeeney *et al*.[179] (GSE14671) microarray dataset involving CD34+ cells from 59 primary CP-CML patients originally generated to identify genes associated with imatinib response, was re-analysed for *GFI1* expression. No significant difference in *GFI1* expression between imatinib responders (R; patients who achieved at least partial cytogenetic response within 12 months of therapy) and non-responders (NR; all other patients), was observed in CP-CML CD34+ cells (P = 0.28; **Figure 7.4 [G]**). Taken together, this data suggests that *GFI1* expression levels are not significantly altered in cases of imatinib resistance regardless of the cell type examined.

As the preliminary data identified that lower *GFI1* expression at diagnosis was associated with CML transformation, we interrogated the Radich *et al.*[180] microarray dataset (GSE4170) to investigate the expression of *GFI1* in CP (n = 42; defined by < 10 % blasts), accelerated phase (AP, n = 17; defined by 10 – 30 % blasts or < 10 % blasts with clonal evolution) and BC-

Figure 7.4 Examination of *GFI1* expression using CP-CML MNC and CD34+ mRNA, and at different stages of disease course

Microarray data from 93 CP-CML mononuclear (MNC) samples was normalized using Robust Multi-array Average (RMA) and used to examine the relationship between *GFI1* expression with: **A.** MMR outcome by 12 months (including 12 months), **B.** MMR outcome by 24 months (including 24 months), **C.** 3-month BCR-ABL1 mRNA level (EMR), **D.** OCT-1 activity, **E.** Kinase domain (KD) mutation development during imatinib therapy, **F.** Sokal score (one-way ANOVA), **G.** Imatinib non-responders (NR) and responders (R) from the McWeeney *et al*.[179] CD34+ CP-CML microarray dataset, **H.** 3 phases of CML disease progression from the Radich *et al*.[180] microarray dataset (*** $P < 0.001$), and **I**. CML patients in CP, BC or BC patients who achieved remission after chemotherapy (n = 3; ** $P < 0.01$, Mann-Whitney). n.s indicates not significant ($P > 0.05$). All statistical analyses were performed using a 2-tailed t-test, unless otherwise stated. Box plots indicate the interquartile range (25 – 75 %) around the median. Whiskers represent the 10th and 90th percentiles. Very low OA – OA < 4 ng/200,000 cells; and Other OA – OA > 4 ng/200,000 cells. Figure from Kok, C.H. and Watkins, D.B. *et al.* 2013.[446]


A.
GFI1 Expression (log2)
P = 0.43
MMR
No MMR
B.
GFI1 Expression (log2)
P = 0.91
MMR
No MMR
C.
GFI1 Expression (log2)
P = 0.96
≤10%
>10%
D.
GFI1 Expression (log2)
P = 0.46
Very low OA
Other OA
E.
GFI1 Expression (log2)
P = 0.18
Mutation
No Mutation
F.
GFI1 Expression (log2)
P = 0.88
Low
Medium
High
G.
GFI1 Expression (log2)
P = 0.28
NR
R
H.
GFI1 Expression (log2)
n.s.
***
***
CP
AP
BC
I.
GFI1 Expression (log2)
n.s.
**
CP
BC
BC Remission

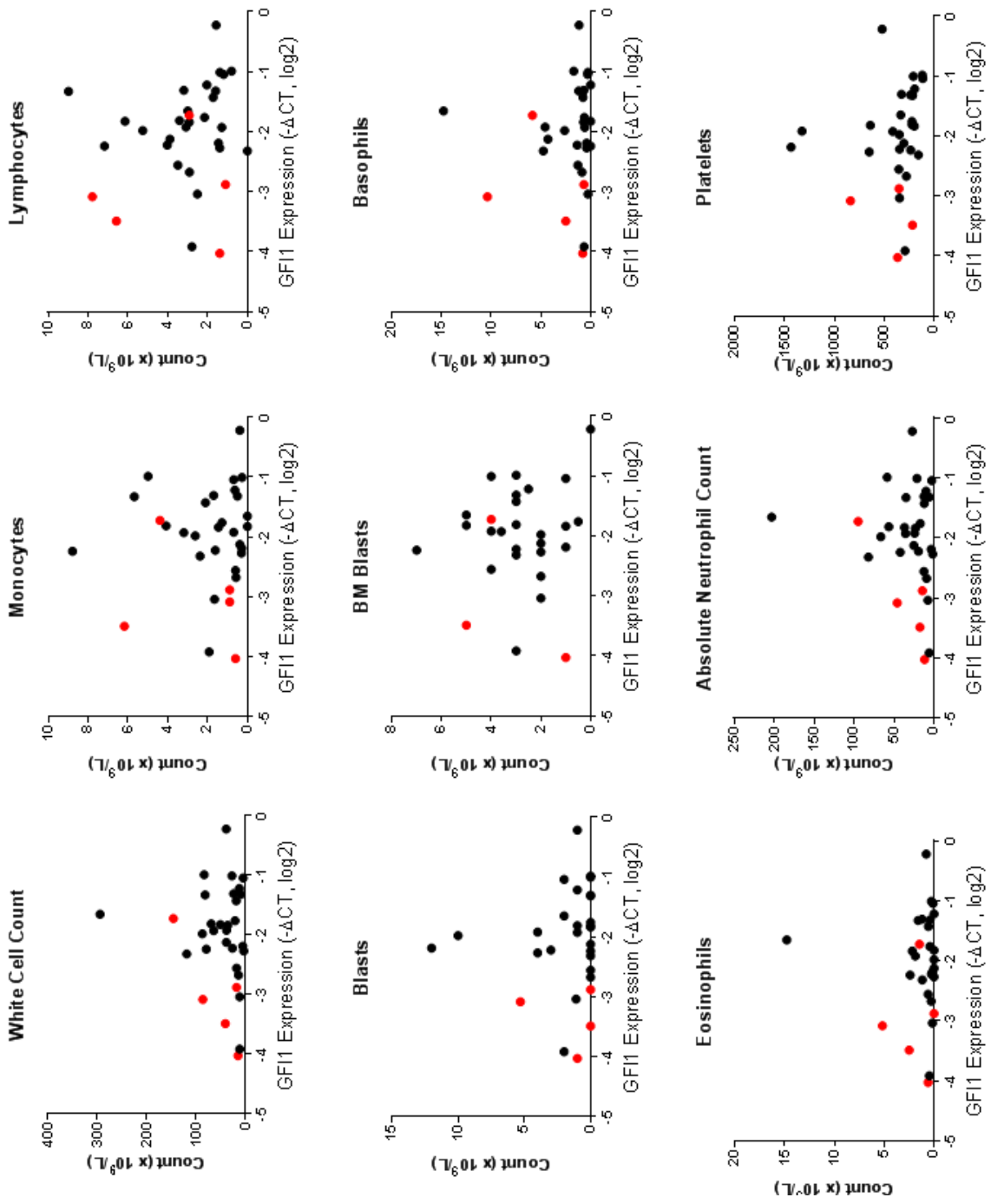
CML (n = 26; defined by > 30 % blasts). Interestingly, material from CML patients in BC had significantly lower *GFI1* expression compared to CP-CML ($P < 0.001$; **Figure 7.4 [H]**) and AP-CML ($P < 0.001$; **Figure 7.4 [H]**); supporting the observation that decreased *GFI1* expression is associated with disease progression in CML. The *GFI1* expression of patients in BC versus BC patients who were now in remission after chemotherapy (non-matched samples, BC-rem; n = 3), was then compared, with a significant increase in *GFI1* expression in the BC-rem patients ($P = 0.009$; **Figure 7.4 [I]**) to levels comparable with CP-CML ($P = 0.33$; **Figure 7.4 [I]**) being observed. Furthermore, this data is in keeping with the lower *GFI1* expression observed in the cohort of *de-novo* CP-CML patients who later transformed to BC. However, it must be noted that an independent study of larger sample size is required to confirm the observation that low *GFI1* expression at diagnosis may be associated with transformation and progression to blast crisis in CML.

7.2.6 GFI1 expression does not correlate with specific cell types within the TWC or MNC populations

To determine if the observed variation in *GFI1* expression between CP-CML patients could be attributable to differences in cell type composition within the TWC and MNC populations, the correlation between *GFI1* expression and differential cell count information, available in 32 patients, was examined. No significant correlation ($P > 0.05$) was observed between the total white cell count, or absolute blast, granulocyte, lymphocyte or monocyte numbers, and *GFI1* expression (**Figure 7.5**), suggesting that the composition of cell types in the total cell population is not a critical determinant of *GFI1* expression. Additionally, as the number of blast cells is increased in advanced phase CML, any correlation between *GFI1* expression and the percentage of blast cells was assessed in these samples. However, no significant correlation was observed (**Figure 7.6**), indicating that the initial number of blast cells was not influencing *GFI1* expression.

Figure 7.5 Correlation between *GFI1* expression and different cell populations in 32 CP-CML patients

Coloured (**RED**) points indicate CP-CML patients who transformed to BC during imatinib therapy (n = 5); however, not all patients had differential cell count data available for each analysis. No significance ($P > 0.05$) was observed with any correlation analysis performed in GraphPad® Prism 6.


Lymphocytes
Basophils
Platelets
Monocytes
BM Blasts
Absolute Neutrophil Count
White Cell Count
Blasts
Eosinophils
GFI1 Expression (-ΔCT, log2)
Count (x 10⁹/L)

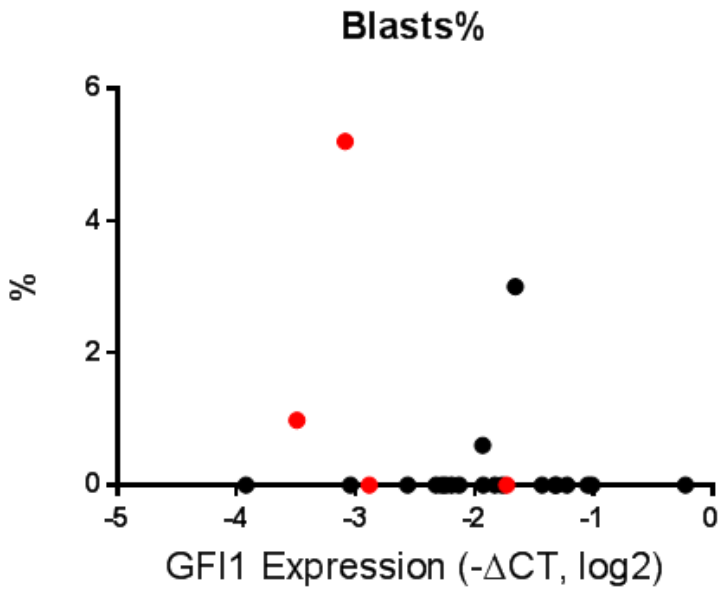


Figure 7.6 Correlation between *GFI1* expression and the percentage of blasts at diagnosis in 32 CP-CML patients

Coloured (**RED**) points indicate CP-CML patients who transformed to BC during imatinib therapy; however, blast percentage data was only available for 4/5 patients for analysis. No significance ($P > 0.05$) was observed with the correlation analysis performed with GraphPad® Prism 6.

7.2.7 GFI1 expression is consistently decreased in CP-CML patients after TKI therapy

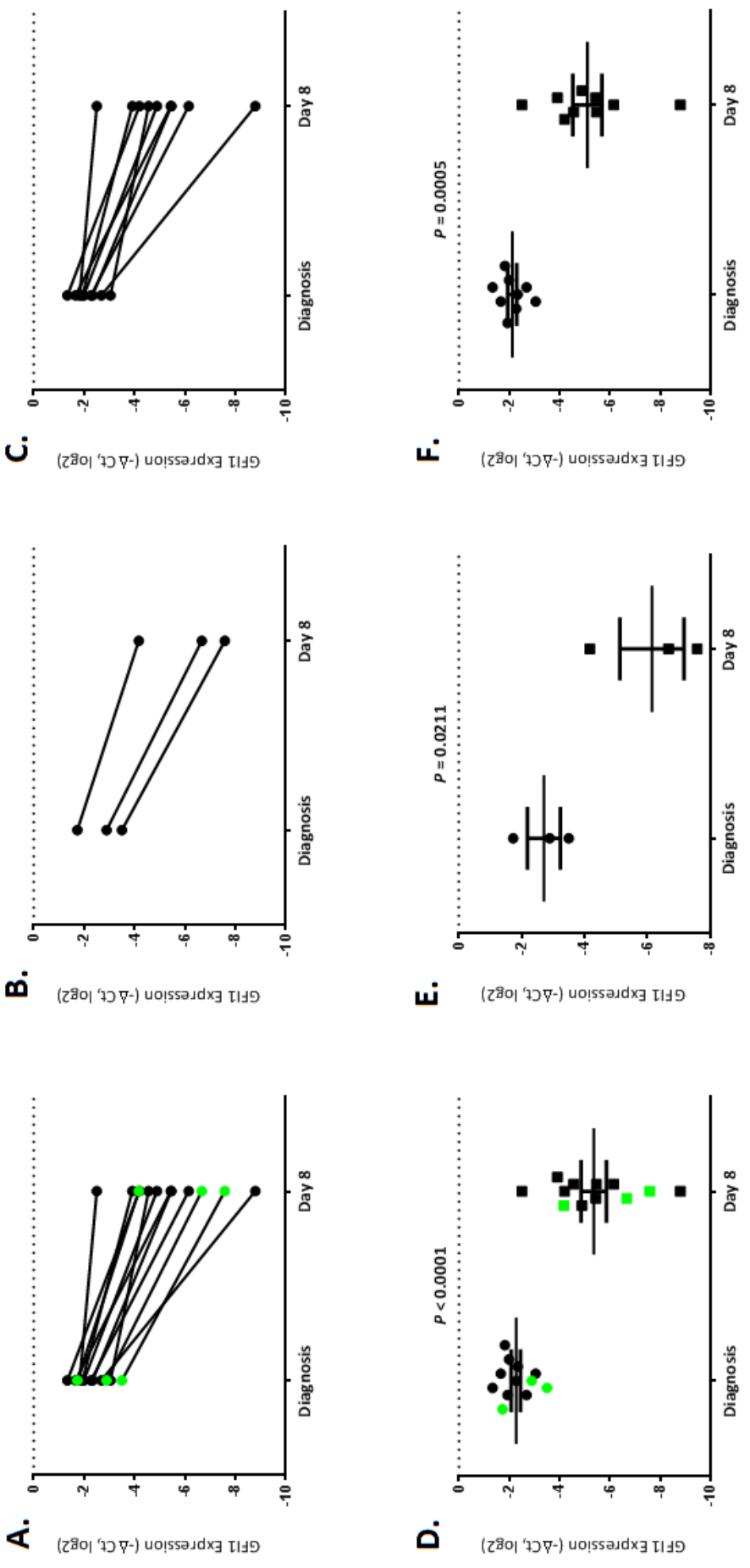
The impact of TKI treatment on *GFI1* expression *in vivo*, and in particular whether patients with low OA and risk of blastic transformation maintained low *GFI1* expression after treatment was determined next. To do this, *GFI1* expression in blood collected after the first 8 days of imatinib therapy (n = 12; including 3 patients who later transformed to BC), was compared with that collected from the same patients at diagnosis, pre-TKI therapy. Overall, a significant decrease in *GFI1* expression was observed in all patients after 8 days of imatinib therapy ($P < 0.0001$; **Figure 7.7 [A] & [D]**). Importantly, this decrease was observed in both the patients who transformed to BC (n = 3, $P = 0.02$; **Figure 7.7 [B] & [E]**) and those who did not (n = 9, $P = 0.0005$; **Figure 7.7 [C] & [F]**), indicating a consistent impact of imatinib treatment in all patients. Intriguingly, GFI1 expression at diagnosis, pre-TKI therapy, was demonstrated to be lower in both normal MNCs[448] and CD34+ cells,[180] compared to CML cells from all disease phases. This data is consistent with the previously published results of Huang *et al.*[448] who demonstrated a 10-fold reduction in CML patient *GFI1* expression back to levels similar to normal individuals after 9 – 12 months of imatinib therapy, compared to before imatinib initiation.

7.2.8 Low GFI1 expression does not correlate with the GFI136N single nucleotide polymorphism

Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation observed in human genes. They can act as biological markers, linking certain genes with disease, and when located within a gene, or gene regulatory region, they can directly influence gene expression, mRNA stability and/or subcellular localization.[449] In congenital neutropenia, mutations in the *GFI1* gene have been associated with a dominant-negative loss of *GFI1* function.[450] Furthermore, Khandanpour *et al.*[451] identified a non-synonymous

Figure 7.7 In vivo effect of TKI treatment on *GFI1* expression in CP-CML patients

A. & D. *GFI1* gene expression measured in matched CP-CML patient samples at diagnosis and at Day 8 (after 7 days of imatinib therapy; n = 12). Coloured (**GREEN**) points represent patients that transformed to BC during imatinib therapy (n = 3). **B. & E.** *GFI1* gene expression measured only in CP-CML patients who transformed to BC on imatinib therapy, at diagnosis and at Day 8. **C. & F.** *GFI1* gene expression measured only in CP-CML patients who did not transform, at diagnosis and at Day 8. Statistical analysis was performed using paired *t*-tests in GraphPad® Prism 6. Error bars indicate the mean ± the standard error of the mean (SEM).


A.
B.
C.
D.
E.
F.
GFI1 Expression (-ΔCt, log2)
0
-2
-4
-6
-8
-10
Diagnosis
Day 8
P < 0.0001
P = 0.0211
P = 0.0005

SNP, resulting in the replacement of serine with asparagine in the N-terminal domain of the coding region of *GFI1*, known as *GFI1*^{36N}, as being significantly associated with AML. The GFI1^{36N} allele, although present in only 3 – 7 % of the white population, increases the risk for AML by 60%.[451] Therefore, these observations prompted the postulation that patients who progressed to BC, identified by their low *GFI1* expression at diagnosis, may possess the *GFI1*^{36N} SNP.

To determine the presence of the *GFI1*^{36N} SNP in the cohort of CP-CML patients analysed for *GFI1* expression, genotyping was performed using a TaqMan® SNP genotyping assay (Applied Biosystems) with allele-specific primers for the *GFI1*^{36N} SNP and wild-type *GFI1*^{36S}. Available TWC DNA was investigated in 32/43 CP-CML patients originally analysed for *GFI1* expression, including 5 CP-CML patients who transformed to BC on TKI therapy. *GFI1*^{36N} SNPs were identified in 3 CP-CML patients (**Figure 7.8**); however, these were not the patients who transformed to BC on TKI therapy, who all contained the *GFI1*^{36S} allele, along with the rest of the patient cohort. No significant difference in *GFI1* expression was observed between the patients with the *GFI1*^{36N} allele or the *GFI1*^{36S} allele (P = 0.75; **Figure 7.9**). Therefore, the *GFI1*^{36N} SNP was not associated with patients at risk of BC progression, or low *GFI1* expression in the tested CML patient cohort.

7.2.9 Increased MLLT4 gene expression is associated with advanced phase CML, but not OCT-1 activity

The myeloid/lymphoid or mixed-lineage leukaemia (trithorax homolog, Drosphila); translocated to 4 (*MLLT4*; located on 6q27) gene is normally involved with a chromosomal abnormality associated with AML, involving the acute lymphoblastic leukaemia (*MLL-1*) gene [t(6;11)(q27;q23)], and encoding the MLL-AF6 fusion protein.[310] *MLLT4* encodes the afadin/AF6 (ALL-fused gene from chromosome 6) protein. MLL-AF6 constitutes the largest

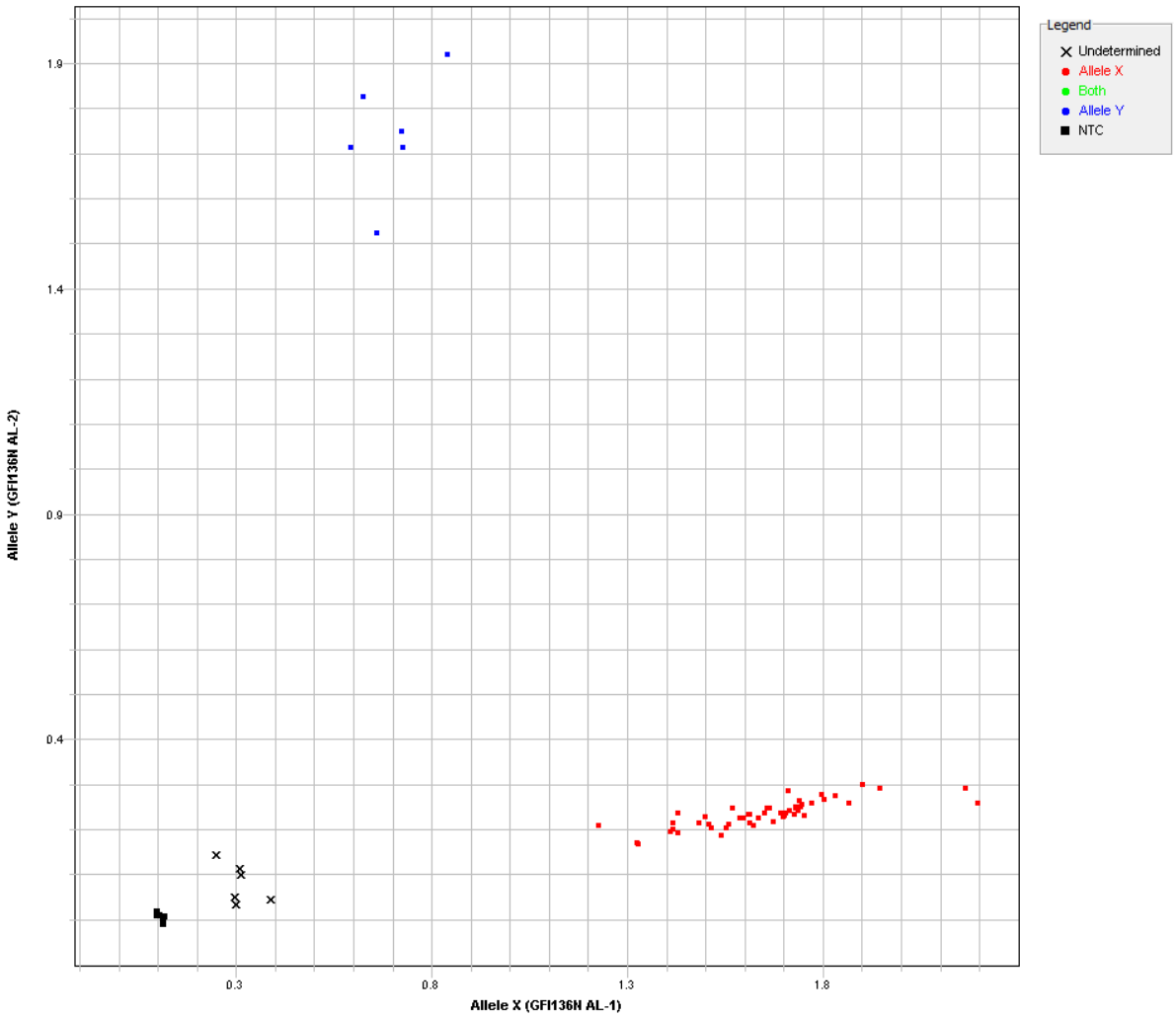


Figure 7.8 Genotyping of CP-CML patients for the *GFI1*^{36N} SNP

SNP genotyping was performed on the ABI PRISM 7900 (Applied Biosystems) using a TaqMan® SNP genotyping assay specific for *GFI1*^{36N} (assay ID: C_25596143_10; Applied Biosystems). All CP-CML samples were run in duplicate. Three samples (in duplicate) failed to amplify during the PCR step (represented as Undetermined). Three samples were positive for *GFI1*^{36N} (**BLUE**), while the rest were negative for *GFI1*^{36N} (positive for *GFI1*^{36S}; **RED**). Allele X (GFI136N AL-1) represents the *GFI1*^{36S} allele, while Allele Y (GFI136N AL-2) represents the *GFI1*^{36N} allele.

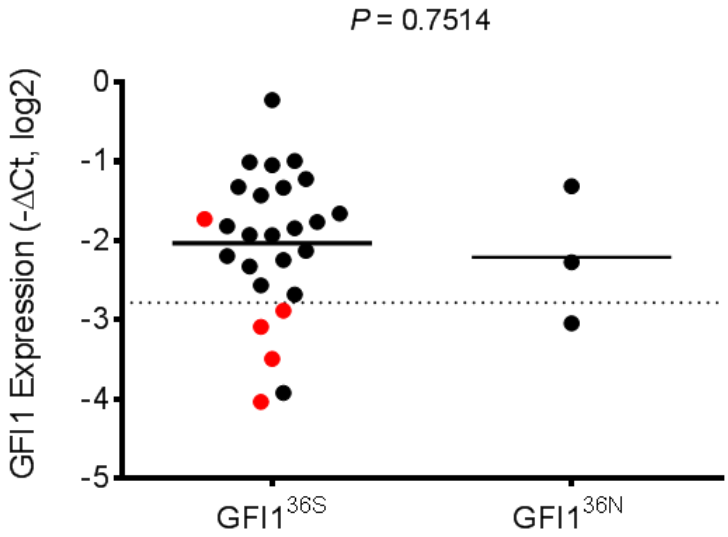


Figure 7.9 *GFI1*^{36N} SNP is not associated with low *GFI1* expression in CP-CML

The 29 CP-CML patients with available *GFI1*^{36N} SNP genotyping are plotted. Coloured (**RED**) points indicate CP-CML patients who transformed to BC during imatinib therapy (n = 5), with the mean value indicated. Statistical analysis was performed using Student's *t*-test in GraphPad® Prism 6.

subgroup of *MLL*-rearranged leukaemias fused with a predominately cytoplasmic protein (AF6).[310] Expression of *MLLT4* was identified from the OA TWC microarray as being significantly increased in very low OA patients, compared to very high OA patients (*P* = 0.0034; FDR *P* = 0.182; logFC = -0.903). Validation of the microarray results by 96-gene TLDA RQ-PCR further confirmed these results, demonstrating that increased *MLLT4* expression was associated with very low OA, compared to very high OA (*P* = 0.002, FDR *P* = 0.034, logFC = -1.15; **Figure 7.10**). However, when this analysis was extended to include patients from all OA groups, significance was lost (*P* = 0.086; **Figure 7.11**), although a trend towards increased *MLLT4* expression in very low OA, compared to other patients was observed. Significantly, *MLLT4* expression was also not associated with any response end-point; including EMR, 12 month MMR, 24 month MMR or BC transformation, when the 110 CP-CML patient cohort was analysed.

To determine if *MLLT4* expression correlated with CML disease phase, the Radich *et al.*[180] CML progression microarray dataset was interrogated. *MLLT4* expression was found to be significantly increased in both AP and BC, compared to CP (*P* < 0.0001; **Figure 7.12**), indicating that *MLLT4* is associated with disease progression in CML. Intriguingly, very low OA is associated with an increased risk of disease progression,[142] which may provide reasoning as to why *MLLT4* expression was increased in very low OA. However, when *MLLT4* expression was analysed by OA quartile, no significant difference was observed between patients with very low OA (Q1) and patients in any other OA quartile (ANOVA, *P* = 0.348; **Figure 7.13**).

Furthermore, to determine if *MLLT4* was associated with resistance to TKI therapy, *MLLT4* expression was measured in five different TKI resistant cell lines, previously developed in the

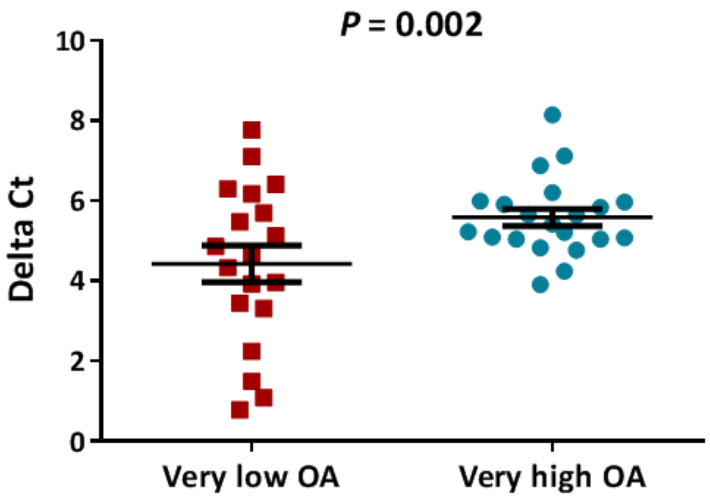


Figure 7.10 *MLLT4* expression is increased in very low OA patients, compared to very high OA patients

The difference between *MLLT4* gene expression for very low (n = 19) and very high OA (n = 21) CP-CML patients is plotted. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. Statistical analysis was performed using Student's *t*-test in HTqPCR.

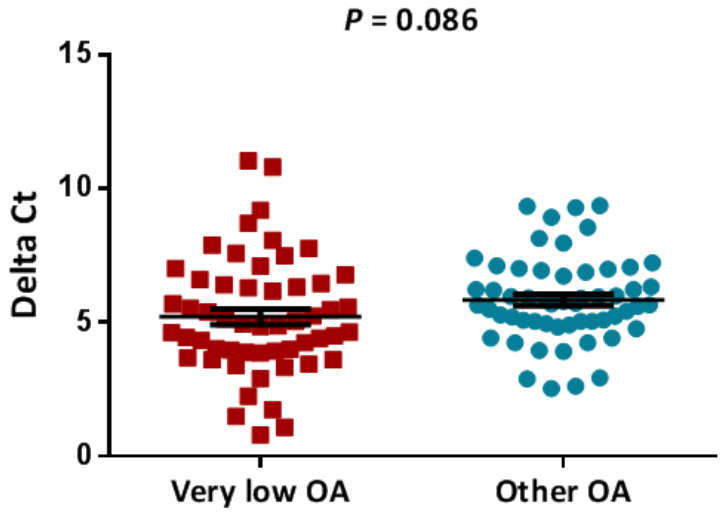


Figure 7.11 *MLLT4* expression is not significantly different between very low and Other OA patients

The difference between *MLLT4* gene expression, as measured by TLDA RQ-PCR, for very low (n = 54) and Other OA (all other patients; n = 56) CP-CML patients is plotted. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. Statistical analysis was performed using Student's *t*-test in GraphPad® Prism 6.

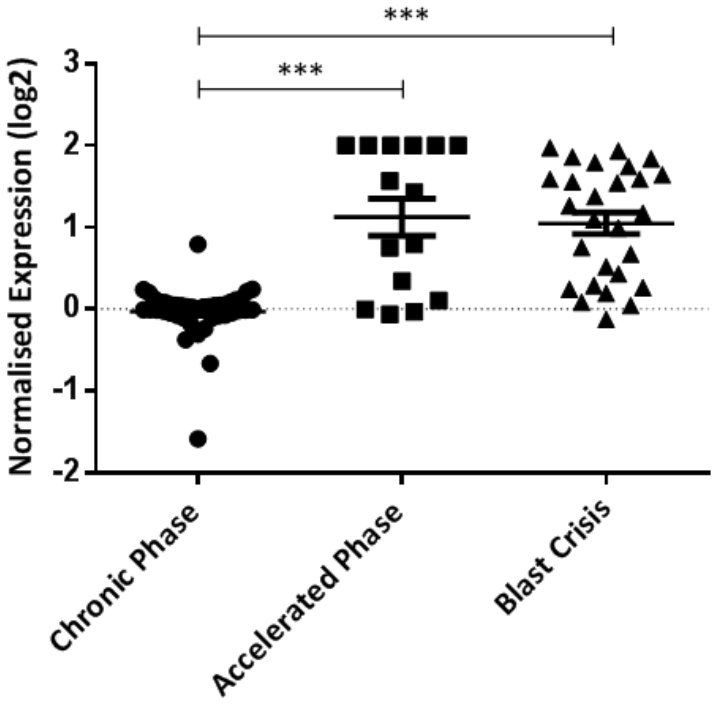


Figure 7.12 *MLLT4* expression is increased in AP- and BC-, compared to CP-CML

Interrogation of the Radich *et al.*[180] CML disease phase dataset identified *MLLT4* expression was significantly increased in both accelerated phase (AP) and blast crisis (BC), compared to chronic phase (CP). No significant difference in expression was observed between AP and BC. Statistical analysis was performed using Student's *t*-test in GraphPad® Prism 6. Error bars indicate the mean ± SEM. *** $P < 0.0001$.

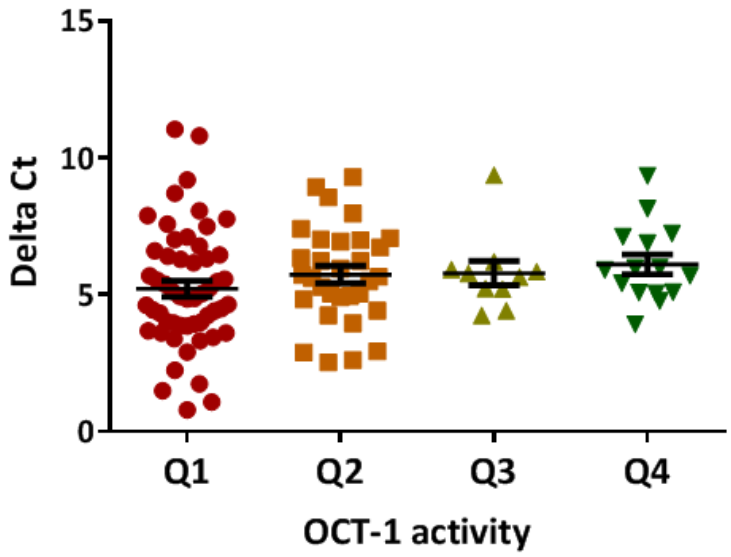


Figure 7.13 *MLLT4* expression does not correlate with specific OCT-1 activity quartiles

The difference between *MLLT4* gene expression, as measured by TLDA RQ-PCR, is plotted for CP-CML patients according to their OCT-1 activity quartile (Q1: OA < 4 ng/200,000 cells; Q2: OA > 4 – 7.2 ng/200,000 cells; Q3: OA > 7.2 – 10 ng/200,000 cells; Q4: OA > 10 ng/200,000 cells). No significant difference was observed. Scatterplots demonstrate the mean Delta Ct ± SEM. Delta Ct values were generated against the reference gene, *GUSB*. ANOVA test was performed using GraphPad® Prism 6.

laboratory.[452,453] Two cell lines (K562 2μM IM1 and IM2) were resistant to imatinib,[452] while three cell lines (K562 2μM NIL, and K562 DOX 2[1] and 2[2] NIL) were resistant to nilotinib.[453]

The K562 cell line was originally derived from a 53 year old female with BC-CML, and the K562 DOX cell line has been induced to stably overexpress *ABCB1* following continuous passage in doxorubicin. Intriguingly, a trend towards decreased *MLLT4* expression in the K562 2μM IM1 cell line, compared to naïve K562 cells was observed; however, this was not significant (P = 0.074; **Figure 7.14 [A]**). This trend was not present in the K562 2μM IM2 cell line, compared to naïve K562 cells (P = 0.576); however, the K562 2μM IM1 cell line had significantly lower *MLLT4* expression, compared to the K562 2μM IM2 cell line (P = 0.008; **Figure 7.14 [A]**). Notably, these resistant cell lines were determined to possess similar imatinib resistance profiles, namely significant *BCR-ABL1* overexpression (approximately 1200% and 400%,[452] respectively; **Table 7.2**). However, the K562 2μM IM2 cell line did demonstrate increased phosphorylated Lyn (p-Lyn) protein levels, which were not observed in the K562 2μM IM1 cell line (**Table 7.2**).

MLLT4 expression was also significantly decreased in the K562 2μM NIL cell line, compared to the K562 125nM NIL intermediate cell line, and the K562 DMSO control (P = 0.024 and P = 0.007, respectively; **Figure 7.14 [B]**). This cell line demonstrated multiple resistance mechanisms including increased p-Lyn, p-Axl and Syk protein levels. No significant differences in *MLLT4* expression were observed in the K562 DOX cell lines (**Figure 7.14 [B]**). Importantly, different nilotinib resistance mechanisms are present between these cell lines (K562 versus K562 DOX), with the K562 2μM NIL cell line exhibiting a more *BCR-ABL1* dependent resistance profile, while both K562 DOX NIL cell lines exhibit *BCR-ABL1* independent resistance profiles (**Table 7.2**).[453]

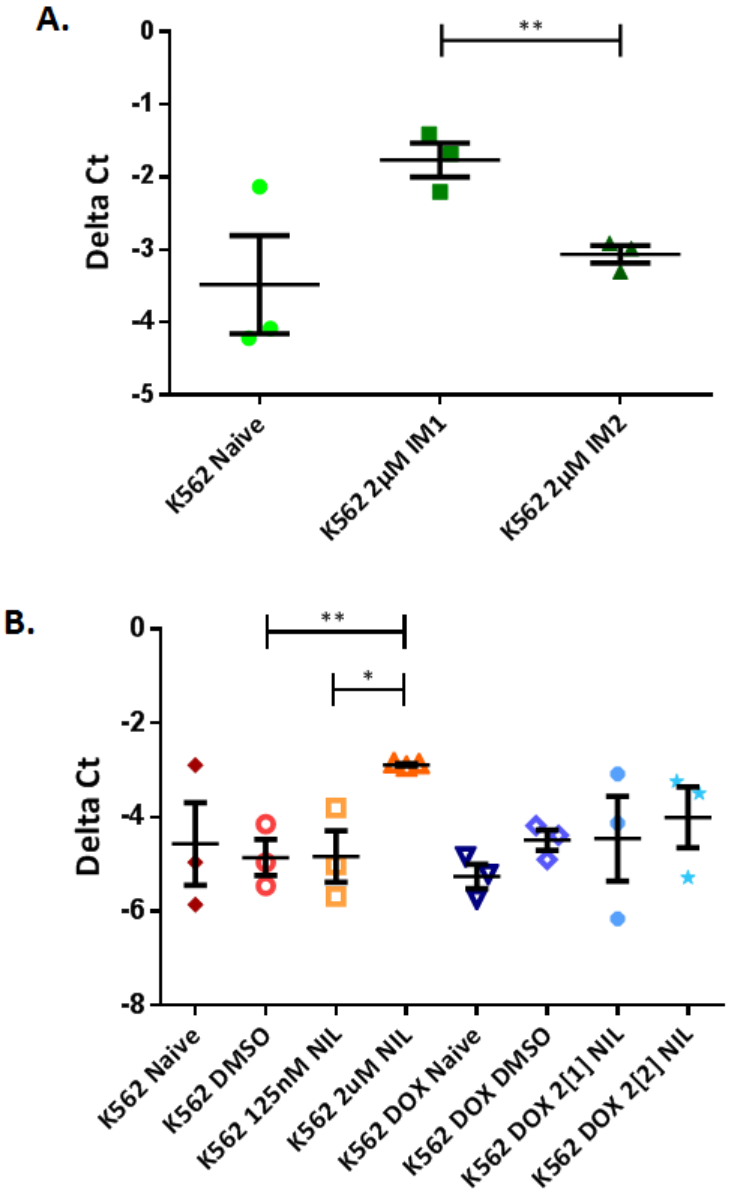


Figure 7.14 *MLLT4* expression in cell lines resistant to either imatinib or nilotinib

MLLT4 expression was determined for imatinib-resistant K562 cell lines **(A)** and nilotinib-resistant K562 and K562 DOX cell lines **(B).** Delta Ct values were generated against the reference gene, *GUSB*, and represent the results of three independent experiments, performed in duplicate. Statistical analysis was performed using Student's *t*-test in GraphPad® Prism 6. Error bars indicate the mean ± SEM. IM = imatinib, NIL = nilotinib, * $P < 0.05$, ** $P < 0.01$.

Table 7.2 Summary of resistance mechanisms in imatinib-resistant K562 cells and nilotinib-resistant K562 and K562 DOX cells

	K562 2µM IM1	K562 2µM IM2	K562 125nM NIL	K562 2µM NIL	K562 DOX 2[1] NIL	K562 DOX 2[2] NIL
Resistance to all front-line TKIs	✓	✓	✓	✓	✓	✓
IC50^{NILOTINIB}	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↓↓↓	↓↓↓
IC50^{IMATINIB}	↑↑↑	↑↑↑	↑↑↑	↑↑↑	↓↓↓	↓↓↓
IC50^{DASATINIB}	↑↑↑	↑	↑↑↑	↑	↓↓↓	↓↓↓
ABCB1 protein overexpression	×	×	✓	×	×	×
BCR-ABL1 mRNA overexpression	✓	✓	×	×	ND	ND
Increased Bcr-Abl1 protein	ND	ND	×	×	×	×
Increased Lyn protein	×	×	×	×	×	×
Increased p-Lyn	×	✓	✓	✓	×	×
Increased Axl protein	✓	✓	×	×	×	×
Increased p-Axl	ND	ND	ND	✓	×	×
Increased Syk protein	ND	ND	ND	✓	No expression	No expression
Increased p-Syk	ND	ND	ND	No phosphorylation	No phosphorylation	No phosphorylation

Abbreviations: ✓ = yes; × = no; ↑ = increase; ↓ = decrease; ND, not determined. Table adapted from Tang *et al.*[452] and Eadie, L.[453]

7.3 Discussion

Biomarkers have an important role in CML, not only to identify patients destined for optimal response, but to identify patients at risk of disease progression or treatment failure on a particular TKI therapy. Effective biomarkers are even more vital now that three different TKIs (imatinib, nilotinib and dasatinib) are available for front line therapy in CP-CML, making the decision regarding which TKI to prescribe to a patient, so that long-term outcome is optimized, increasingly complex. Although many biomarkers have been proposed in CML,[421] a lack of independent validation has stalled the progress of most, hence, limited biomarkers have actually been adopted for day-to-day clinical use. The OCT-1 activity (OA) assay has excellent prognostic capabilities,[142,151] but is restricted by its requirement for [^{14}C]-imatinib, live cells and specialised equipment. For this reason, detailed analysis of the gene expression profiling performed in **Chapter 4** was performed so that new biomarkers, to either replace the assay itself, or to predict the same clinical endpoints, could be identified.

7.3.1 NPM1 mutations are not a biomarker for either CP- or BC-CML

Significant enrichment of a histone gene signature, identified from both the microarray gene expression profiling and RQ-PCR data, was established as an interesting feature of very low OA (**Chapter 4**). Further investigation of this signature with the MSigDB identified significant enrichment for RNA Pol I associated gene-sets, and an AML with mutated *NPM1* (*NPM1^{mut}*) gene-set (Verhaak *et al.*[430]). RNA Pol I has a direct role in the production of ribosomal RNA (rRNA), driving cell growth and determining the fate of the cell.[454] Furthermore, given the established role of *NPM1* in ribosome biogenesis, particularly RNA Pol I transcription,[435,436] a possible role for *NPM1* in histone expression and CML was indicated. The Wouters *et al.*[437] AML dataset, where *NPM1^{mut}* correlated with an up-regulation in histone gene expression; and the recent publication by Georgiou *et al.*[438] describing a CML patient who achieved

complete molecular response, but subsequently developed AML with *NPM1*^{mut}; also support this role. Taken together, this data lead to the postulation that *NPM1*^{mut} would be associated with very low OA CP-CML. However, no *NPM1* mutations were detected in a cohort of 14 BC-CML and 33 CP-CML patients (of which 17 had very low OA), including a single patient who progressed to BC from CP while on TKI therapy. Therefore, the data described here demonstrates that mutations in the *NPM1* gene are rarely, if at all, associated with CP- or BC-CML patients, including those at risk of disease progression (very low OCT-1 activity), and does not support a role for *NPM1*^{mut} in the up-regulation of histone gene expression observed in very low OA CML patient cells.

7.3.2 Diagnostic GFI1 expression is a biomarker for disease progression in CML

The growth factor independence 1 *(GFI1)* gene was originally identified as differentially expressed from the OA microarray gene expression profiling (up in very high OA). However, in combination with the data from Soliera *et al.*[444] indicating ectopic *GFI1* expression inhibited proliferation and colony formation in *BCR-ABL1* positive cell lines and primary CD34+ cells, it was postulated that increased *GFI1* would be associated with favourable outcome in imatinib treated CP-CML. However, *GFI1* expression, as measured by RQ-PCR at diagnosis in TWCs, was unable to be correlated with multiple response endpoints, including MMR (12- and 24- month achievement), EMR, mutation status or Sokal score. Furthermore, the original differences in OA observed from the microarray were unable to be replicated by RQ-PCR.

Interestingly, *GFI1* expression was significantly decreased in the diagnostic sample from CP-CML patients who subsequently transformed to BC while receiving imatinib as front-line therapy, compared to those patients who remained in CP. *De novo* CP-CML patients with

GFI1 ΔCt < -2.782 (-ΔCt, log2; likelihood ratio = 10.28) were approximately 10-fold more likely to transform to BC, identifying low *GFI1* expression as a promising biomarker of transformation in CML. Additionally, advanced phase (both AP and BC) CML patients were found to have lower *GFI1* expression than CP-CML patients, further supporting the observed association between decreased *GFI1* expression and progression in CML. This work supports the previously described role of *GFI1* in inhibition of proliferation and colony formation of p210BCR/ABL-transformed cells and primary CD34+ CML cells.[444] Furthermore, the blockade of granulopoiesis in *GFI1*-/- myeloid progenitors via elevated levels of HoxA9, Pbx1, and Meis1,[455] the development of myelo-proliferative disease,[456] and defects in T-cell development ,[457] have also been exhibited by *GFI1*-/- mice. Taken together, these results suggest that unlike other negative prognostic markers (such as low OA, 3 month BCR-ABL1 mRNA levels ≥ 10%, and high Sokal score), low *GFI1* expression correlates selectively with a high risk of transformation to BC while receiving TKI therapy. However, no correlation was observed for mutation development, primary resistance or secondary resistance in CP-CML patients. Furthermore, of the 6 CP-CML patients who transformed to BC, 3 achieved EMR, while 3 did not; indicating that *GFI1* expression at diagnosis was independent of this prognostic marker. However, it must be noted that an independent study of larger sample size is required to confirm this observation.

Deletion of *GFI1* (*GFI1*$^{-/-}$), in combination with induction of Ras activation, has been demonstrated to induce a transplantable AML in mice in 17 days,[455] providing the first definitive evidence that *GFI1*$^{-/-}$ predisposes to myeloid leukaemia. Khandanpour *et al.*[456] recently demonstrated that GFI1$^{-/-}$xBcl-2 transgenic mice contain a systemic expansion of Mac-1^{+}Gr-1^{-} myeloid cells in bone marrow and peripheral lymphoid organs, which appear as monocytoid-blastoid cells in the peripheral blood, and result in the mice eventually

succumbing to a myeloproliferative-like disease resembling a pre-leukaemic state. Furthermore, low *GFI1* levels are associated with an increased risk of AML development in myelodysplastic patients.[458] Taken together, this data, in combination with the observed correlation between low *GFI1* expression and increased risk of BC transformation (to either myeloid or lymphoid BC) in CP-CML patients demonstrated here, provides further evidence that *GFI1* functions to suppress myeloid progenitor transformation.

The mechanism behind low *GFI1* expression is currently unknown; however, the *GFI1*^{36N} variant is associated with a 60% increased risk of AML,[451] and *GFI1* mutations have been associated with an increased risk of AML in patients with neutropenia.[450,458] Yet, the data presented here demonstrates that *GFI1*^{36N} was not associated with CP-CML patients who progressed to BC on imatinib therapy, nor was it associated with low *GFI1* expression. Thus, while low *GFI1* expression at diagnosis was identified as a potential biomarker for CP-CML patients at risk of disease transformation on TKI therapy, the mechanism regulating *GFI1* expression remains unknown, but is not associated with the *GFI1*^{36N} SNP.

7.3.3 MLLT4 expression is not a biomarker for OA, but is increased in advanced phase CML

MLLT4 expression was significantly increased in very low OA patients from the OA TWC microarray; however, this was not validated by RQ-PCR in a larger patient cohort, including patients from all OA groups. Although, a trend towards increased expression in very low OA compared to the rest was apparent. Interestingly, *MLLT4* expression was significantly increased in advanced phase CML (AP and BC), compared to CP. Yet, when measured at diagnosis in CP-CML patients, it was not predictive of transformation, indicating that *MLLT4* is not an initiating factor in CML progression, but more likely a secondary partner. *MLLT4* expression was not significantly associated with imatinib resistance in two K562 cell lines,

resistant to 2 μM imatinib, compared to naïve K562 cells; although a trend towards decreased expression in the resistant cell lines was apparent. *MLLT4* expression was however, significantly associated with nilotinib resistance in the K562 2μM NIL cell line, compared to both the K562 DMSO control and 125nM NIL intermediate resistant cell line. Confounding this observation though is the fact that this cell line was identified to contain *BCR-ABL1*-dependent resistance mechanisms (**Table 7.2**), including *ABCB1* overexpression as the initiator of resistance, which decreased after prolonged exposure to nilotinib, and increased *SRC* family kinase expression.[453]

Interestingly, *MLLT4* encodes the cytoplasmic AF6 protein, a Ras-binding protein[459] that has been demonstrated to interact with the Bcr kinase, which down-regulates Ras signalling by phosphorylating AF6 and binding to its PDZ domain.[311] Importantly, the PDZ domain-binding motif (STEV) of Bcr is at the extreme C-terminus, and is not present in the Bcr-Abl1 fusion oncoprotein.[311] AF6 also has a role in the regulation of c-Src, as c-Src binding to the PDZ domain of AF6 results in moderate activation of the c-Src kinase.[460] This role of AF6 with c-Src is quite interesting when the K562 nilotinib resistant cell line results are considered, allowing the speculation that *MLLT4*/AF6 could contribute to a possible mechanism for TKI resistance in CML. However, this is outside the scope of this work, as the aim was to develop biomarkers for CML, and unfortunately, no significance with any critical response end-point was observed from the *MLLT4* gene expression studies. Hence, AF6 protein analysis was not pursued, although future studies of AF6 would be beneficial.

In conclusion, the findings detailed in this chapter strongly identify low *GFI1* expression at diagnosis as a promising biomarker for CP-CML patients likely to progress to BC, during TKI therapy, while excluding both *NPM1* and *MLLT4* as possible biomarkers in CML. Thus,

patients in whom *GFI1* expression is low at diagnosis may be predisposed to BC transformation. *NPM1*^{mut} were confirmed as being not present in both CP- and BC-CML, while *MLLT4* was determined to not correlate with CML response, although expression is increased in advanced phase disease. Overall, the findings described identify a valuable new biomarker for progression in CML, and when independently validated, may allow tailored therapeutic strategies to be developed for the patients identified at risk with low *GFI1* expression.

8 DISCUSSION

8.1 Introduction

The organic cation transporter 1 (OCT-1) protein specifically mediates the uptake of many drugs via its membrane associated ATP-dependent protein channel.[99] OCT-1 is now well established as the major active influx pump for imatinib,[96,97,99,142] and is involved in the regulation of the overall intracellular imatinib concentration. Furthermore, a key predictor of patient response, decreased imatinib-induced Bcr-Abl kinase inhibition,[135,242] has been demonstrated to be intrinsically linked to the amount of drug actively pumped into the cell via OCT-1.[96,97,99,149,151] Hence, reduced activity of the OCT-1 protein is a likely key to primary imatinib resistance and suboptimal response in CP-CML patients.

The OCT-1 activity (OA) assay (developed in our laboratory[151]) is a functional assay measuring imatinib uptake in CML patient mononuclear cells (MNCs) at diagnosis. The OA assay is one of the best predictors of response to imatinib and provides excellent predictive value in diagnostic CP-CML patient samples for subsequent molecular response.[142,151] Patients with very low OA (OA < 4 ng/200,000 cells) are at significant risk for poor response (lower probability of MMR, EFS, TFS and OS), compared to all other patients.[142] However, the assay has not been widely adopted for routine testing. Importantly, not only are varying OA levels observed between CP-CML patients, but varying molecular outcomes are also observed, suggesting that divergent disease biology may exist within the CP-CML patient cohort at diagnosis, leading to varied activity of the OCT-1 transporter. Furthermore, the OA assay is performed using [^{14}C]-imatinib, requires live cells and specialised equipment, making the worldwide applicability of the assay in diagnostic laboratories unfeasible. It was therefore important to elucidate the underlying biology of OA and understand the key biological drivers which may differentiate the range of OA levels observed in CP-CML patients, potentially identifying whether patients with very low OA and poor response to

imatinib have different overall disease characteristics associated with alternative biological mechanisms. Secondly, identification of a more practical, alternate assay with greater diagnostic applicability to measure OA, or at least identify low OA patients, was sought.

8.2 Major Findings

8.2.1 Specific cell lineage differences are present between patients with very low OA compared to all other patients

Previous research from our laboratory demonstrated that OA, as measured in CP-CML patient MNCs, was strongly related to cell lineage.[243] While these results suggested that the predictive nature of the OA assay in CP-CML patients was potentially dependent on the cellular composition of the original peripheral blood sample used to isolate the MNCs, further immunophenotypic analysis of CML patient MNCs at diagnosis was lacking. Thus, the variation in immunophenotype at diagnosis was determined for a cohort of CP-CML patients, and related to patient characteristics at presentation, including OA and compared with subsequent imatinib response.

Small, yet distinct, immunophenotypic differences were identified between CP-CML patients with very low OA, compared to all other patients (**Chapter 3**). The initial analysis (n = 27) identified a decreased neutrophil population in very low OA. However, this data was confounded when the analysis was extended to include a further 68 patients (total n = 95), as different cell lineage populations were subsequently associated with very low OA (T-lymphocyte, B-cell and erythrocyte). The observed variation between these patient cohorts supports the established notion that significant heterogeneity exists between CP-CML patients at diagnosis. Importantly, previous immunophenotyping studies have demonstrated

lineage heterogeneity in BC-CML,[461,462] which can be postulated to extend from lineage heterogeneity in CP-CML. Furthermore, no individual variable was identified to be significantly different between the patient cohorts, including prior treatment before TKI, such as hydroxyurea, which reduces the WBC count[463] and may influence the immunophenotyping results. However, the only consistent difference between the two patient cohorts was the clinical trial for which each cohort was recruited (TIDEL II [first cohort, n = 27 plus an extra 8 patients, total n = 35] and ENESTxtnd [second cohort, n = 60]). Although unlikely to fully explain the observed lack of correlation, it is postulated that the lineage differences observed between these two patient cohorts is a result of CP-CML patient heterogeneity, which would have a greater impact on the small number of patients analysed in the first cohort (n = 27), compared to the overall, combined 95 patient data. Differences in the clinical trial parameters, including the type and duration of pre-trial therapy, may have also contributed.

Additionally, significant lineage differences were observed between very low and very high (OA > 10 ng/200,000 cells) OA patients when gene-set enrichment analysis was performed using microarray data, as detailed in **Chapter 4**. Significant enrichment of the very low OA gene signature with up-regulated monocyte and down-regulated erythrocyte gene-sets was identified. The erythrocyte signature supports the results from the immunophenotypic studies in **Chapter 3**; however, these studies were unable to fully support the monocyte signature. Taken together, the results presented in this thesis, combined with the previous results from Engler *et al.*[243] suggest that significant lineage differences exist between CP-CML patients with very low OA, compared to all other patients, and these are likely to be responsible for the variation in OA values observed between these patients.

8.2.2 Decreased T-lymphocyte population in very low OA suggests impaired immune surveillance

A decreased T-lymphocyte signature (including the markers CD3+CD7-, CD7+, CD3-CD7+ and CD8+, although CD3+CD25- was increased) was found to be associated with CML patients with very low OA, compared to all others, in the lymphocyte-gated percentage dataset (**Chapter 3**). In CML, it is believed that a leukaemia-specific immune response contributes to disease control.[464] Furthermore, the importance of immune surveillance has previously been recognized, with leukaemic stem cells evading immune surveillance through protection by a combination of human leukocyte antigen molecules and targets of adaptive immunity.[12,465]

This data agrees with a previous study which demonstrated an increase in the number of lymphocytes in the bone marrow of CP-CML patients who had a good response to imatinib, reflecting possible immune activation in these optimal responders.[466] Furthermore, imatinib has been demonstrated to enhance antigen-presenting cell function via tyrosine phosphorylation of intracellular targets, indicating an important role in T-cell activation.[466-468] Therefore, one can speculate that the decreased T-lymphocyte signature observed in CP-CML patients with very low OA results in decreased T-cell activation when patients are treated with imatinib. This in turn prevents the natural immune surveillance mechanism from functioning optimally; hence, contributing to the poor response observed in these patients. This hypothesis is further supported by the observation of an increased T-lymphocyte signature associated with patients who did achieve early molecular response (EMR; 3 month BCR-ABL1 mRNA levels < 10%), in the percentage immunophenotyping datasets (**Chapter 3**). The exact initial cause for these lineage differences remains elusive; however, their impact on immune surveillance may have an important role warranting further immunophenotypic examination. Unfortunately, this was unable to be performed as a part of this thesis due to time constraints and a lack of *de novo* CP-CML sample availability.

Consequently, this data raise the hypothesis that T-lymphocyte mediated immunity may contribute to an improved response in CP-CML patients with high OA, as suggested by the increased T-lymphocyte signature observed in these patients. Further investigation is warranted to investigate this, and to exclude any confounding associated with this T-lymphocyte signature, such as bias within the sample cell populations.

8.2.3 FZD6 expression is associated with early molecular response in CP-CML patients

The frizzled (Fzd) gene family member, FZD6, was the only marker different in all four (ungated and lymphocyte-gated percentage and MFI) immunophenotyping datasets, with FZD6 expression significantly increased in patients who did achieve early molecular response (EMR), compared to those who did not (**Chapter 3**). Although no direct role for FZD6 has been established in CML, the Fzd gene family are receptors for Wnt signalling proteins. Wnt/β-catenin signalling has been demonstrated to be involved in protecting CML stem cells from TKI-induced cell death, leading to leukaemic stem cell preservation.[278,279] Importantly, FZD6 has been demonstrated to act as a negative regulator of the canonical Wnt/β-catenin signalling cascade, as FZD6 repressed Wnt3a-induced canonical signalling when co-expressed with FZD1.[282] The repressive effect elicited by FZD6 was not associated with a decrease in the overall β-catenin level, suggesting that FZD6 does not affect β-catenin stabilisation, but rather repressed the canonical Wnt pathway by transmitting an antagonistic signalling cascade.[282] Therefore, elevated FZD6, observed in CP-CML patients likely to respond well to TKI therapy (achieve EMR), is postulated to repress the canonical Wnt/β-catenin signalling pathway, preventing the Wnt/β-catenin-mediated protection of CML progenitor cells from TKI therapy.

8.2.4 Up-regulation of histone gene expression is enriched in CP-CML patients with very low OA

Gene expression profiling identified an enrichment of up-regulated histone genes in patients with very low OA, compared to all other patients (**Chapter 4**). Histones are essential for the packaging of DNA into chromosomes,[322] with histone gene expression being cell-cycle regulated and controlled at both the transcriptional and post-transcriptional level.[324] The transcription factors (TFs) MYC and E2F1 were subsequently identified to be likely involved in the regulation of histones in CML by TF binding and pathway analysis was performed (**Chapter 4**). However, western plot analysis was unable to confirm this link.

Importantly, there are limitations associated with using TF databases to predict the binding of TFs to genes of interest, including the assumption of independence between regulators and targets. It is known that both the kinome and transcriptional regulatory networks consist of tightly coupled protein kinases and TFs, which are often involved in the regulation of other kinases and TFs.[469-471] Therefore, the results from the TF binding analyses may be misleading if the networks identified are closely involved in the regulation of other TFs, which may not always be identified. Although unlikely, it is possible that the TFs identified from the TF prediction analyses may have been false-positives. Otherwise, it is speculated that the lack of association observed between MYC and E2F1 with OA may have related to experimental design and assay performance. To address these issues, further optimisation of the MYC and E2F1 western blot assays is currently being performed using a more sensitive protocol which may allow both MYC and E2F1 to be detected in *de novo* CP-CML patient samples.

Nevertheless, this data suggests that the up-regulated histone expression was unlikely to be transcriptionally regulated, but rather, may be controlled by a post-transcriptional

mechanism. DNA methylation is an attractive target for post-transcriptional modification of histone expression. Furthermore, DNA methylation and histone modification are often dependent on one another and can influence normal development and tumourigenesis.[472] Therefore, investigation of global DNA methylation in CP-CML patients from the different OA groups may reveal an epigenetic mechanism influencing very low OA.

8.2.5 Global DNA methylation patterns significantly vary between CP-CML patients with very low OA compared to all other patients

Global DNA methylation analysis was performed as described in **Chapter 5**, to ascertain whether aberrant epigenetic programming may underlie the poor response associated with very low OA. Significantly divergent global DNA methylation patterns were observed between CP-CML patients with very low OA, compared to all other patients, with 62.5% of the significant CpGs identified hypomethylated in patients with very low OA. This data indicates that global hypomethylation is likely to be contributing to the poor molecular response observed in patients with very low OA and may be relevant to the gene expression differences observed.

This data is highly novel, with no CML-specific global DNA methylation studies having been published to our knowledge. Aberrant epigenetic regulation occurs in many cancers, with global DNA hypomethylation often observed.[359,360] This aberrant hypomethylation is often associated with the activation of some proto-oncogenes, such as *IGF2*, and can lead to the loss of genomic imprinting and contribute to tumourigenesis and disease progression.[360-363] To obtain a more complete understanding of the impact this aberrant methylation had on very low OA, matched gene expression data was generated (**Chapter 5**). A number of targets were identified, where both significant DNA methylation and gene expression differences

were observed. Validation of these targets and the novel global DNA methylation results are now part of important ongoing studies, which includes further pyrosequencing to validate new CpG targets. While preliminary, this data allows one to speculate that the poor response associated with CP-CML patients with very low OA is influenced in part by aberrant epigenetic mechanisms, including both DNA hyper- and hypomethylation of genes, which contributes to disease persistence and affects the underlying biology of very low OA.

8.2.6 Global hypomethylation occurs in CP-CML compared to normal individuals

Global DNA methylation analysis was also performed to understand whether epigenetic mechanisms contributed to overall CP-CML disease pathogenesis. Global DNA methylation patterns were found to vary significantly between CP-CML patients and normal individuals (**Chapter 6**), with hypomethylation observed in CP-CML. Global DNA hypomethylation has been reported to occur in many human cancers, and has an independent role in cancer formation.[363,473] Pathway enrichment analysis of the genes differentially methylated identified enrichment of a number of pathways including anti-apoptosis, signal transduction and immune response. Therefore, genes associated with these pathways are demonstrated to be up-regulated in CP-CML compared to normal individuals, potentially contributing to the increased cell proliferation, evasion of apoptosis and survival, which are hallmarks of CML. Furthermore, the results described in this thesis indicate that there are previously unknown, yet significant, global epigenetic involvement in CML, which may warrant the use of combination therapy, where TKIs are used in conjunction with epigenetic agents such as decitabine and azacitidine to inhibit DNA replication and DNA methyltransferase activity.

8.2.7 Global hypermethylation occurs in BC-CML compared to CP-CML

CML disease progression from CP to BC has been studied extensively; however, the global DNA methylation analysis performed in **Chapter 6** suggested that there may be a previously unknown epigenetic mechanism contributing to disease progression. Specifically, significant variation was observed between the DNA methylation patterns from CP-CML and BC-CML patients, with hypermethylation observed in BC-CML. Site specific promoter hypermethylation is often associated with the transcriptional silencing of tumour-suppressor genes, including *CDKN2A*, *BRCA1* and *VHL*,[359,360] and hence, is implicated in disease progression.[473] Pathway enrichment analysis suggested that genes associated with apoptosis and development were suppressed in BC-CML, which is likely related to the blastic phenotype of the cells present in BC-CML patients. Further gene expression studies are now being performed to identify target CpGs corresponding to genes with both aberrant methylation and gene expression signatures, which may be possible therapeutic targets. This data identified a novel global DNA hypermethylation signature associated with BC-CML, which is likely to contribute to disease progression in these patients. Furthermore, new treatment strategies which include epigenetic therapies may now be developed based on these observations to aid in the treatment of patients who have progressed to BC, leading to an improved survival outcome.

8.2.8 Low diagnostic GFI1 expression is a biomarker for disease progression in TKI-treated CP-CML patients

Gene expression data also led to the identification of the most promising biomarker identified in this research, low *GFI1* (growth factor independence 1) expression at diagnosis (**Chapter 7**). *GFI1* is a major regulator of haematopoietic stem cells,[474] and was identified as being significantly up-regulated in patients with very high OA from the original OA microarray analysis performed in **Chapter 4**. Although validation was unable to confirm this

association with OA, further investigation did identify that *GFI1* expression was significantly decreased at diagnosis in CP-CML patients who subsequently transformed to BC while receiving imatinib as front-line therapy, compared to those patients who remained in CP. Thus, low *GFI1* expression at diagnosis was identified as a promising biomarker of transformation in CML.

The exact mechanism as to how decreased *GFI1* expression contributes to transformation is still under investigation; however, Soliera *et al.*[444] previously described a role for *GFI1* in inhibition of proliferation and colony formation of p210BCR/ABL-transformed cells and primary CD34+ CML cells. Furthermore, the deletion of *GFI1* (*GFI1*$^{-/-}$), in combination with induction of Ras activity has been demonstrated to induce a transplantable AML in mice in approximately 17 days,[455] while alone, it is associated with the development of a myeloproliferative-like disease.[456] Interestingly, *GFI1*$^{-/-}$ has also been associated with defects in T-cell development,[457] while low *GFI1* levels have been associated with an increased risk of AML development in myelodysplastic patients.[458] Taken together, it is speculated that *GFI1* functions to suppress myeloid progenitor transformation; hence, low *GFI1* expression may be associated with CML disease progression due to a decrease in this suppression, leading to unregulated leukaemic stem cell production. Overall, the results presented in this thesis identified a valuable new biomarker for progression in CML – low diagnostic *GFI1* expression, contributing to an important unmet need in CML. Once independently validated, tailored therapeutic strategies involving more aggressive upfront therapy, such as second- or third-generation TKIs (e.g. nilotinib, dasatinib or ponatinib), may be developed for the patients identified at risk of poor response, in order to improve their response and prevent disease progression.

8.2.9 A predictive OA classifier could not be developed to replace the OA assay

The ability to classify *de novo* CP-CML patients into those likely to respond well to TKI (e.g. imatinib) therapy versus those likely to respond poorly before therapy is commenced, is an enticing prospect for clinicians. Predictive classifiers, based on OA, were developed using the immunophenotyping data (**Chapter 3**), gene expression data (**Chapter 4**) and the DNA methylation data (450K and pyrosequencing; **Chapter 5**) presented in this thesis. Unfortunately, although these classifiers all showed promise when first developed using the training set, they all lost their high predictive accuracy when validated with an independent test set.

Immunophenotypic algorithms have been used previously to divide subtypes of diffuse large B-cell lymphomas (DLBCLs) with generally good results;[475-479] however, not all classifiers were able to demonstrate an association with patient outcome. The immunophenotypic OA classifier originally developed had an overall predictive accuracy of 88% using a 34 patient training set; however, this decreased to only 55% when an independent 60 patient test set was used for validation (**Chapter 3**). The immunophenotypic classifier may have failed due to technical challenges that limit its utility, including antibody specificity and sensitivity, and the inherent subjectivity of data interpretation. These factors may have contributed to the failure to validate between the different patient cohorts, even though as much care as possible was taken to avoid any discrepancies and differences between the cohorts.

DNA methylation classification algorithms have previously been generated for subtype classification and outcome prediction in paediatric acute lymphoblastic leukaemia,[366] acute myeloid leukaemia,[364,385] myelodysplastic syndrome,[386] and non-small cell lung cancer.[480] The original DNA methylation classifier, generated using the 450K DNA methylation data in **Chapter 5**, had an overall predictive accuracy of 98% using the 49 patient training set, and

was able to be replicated using the 46 patient pyrosequencing methylation data with 93% predictive accuracy. Unfortunately, when an independent 46 patient cohort was used as the test set, the pyrosequencing classifier was unable to be validated. Walter *et al.*[480] state that *"a perceived limitation of using DNA methylation as biomarker is that it is by definition, indirect",* as it is thought that DNA methylation itself does not necessarily cause gene silencing (or activation), but is more likely to be a marker of the gene's transcriptional state. Studies have demonstrated that epigenetic patterns can be prognostic for outcome in cancers such as glioblastoma[481] and breast cancer,[482] as well as the cancers mentioned previously. The data from these studies, in combination with that from Walter *et al.,*[480] suggest that DNA methylation profiling is in fact, at least as informative as gene expression profiling for some cancers, as it can accurately define both biological and clinical outcomes. The exact reason why this classifier failed to validate is difficult to ascertain; however, it is speculated that small patient numbers, as discussed further below, contributed significantly.

When performing classification using machine-learning algorithms, it is important to be aware of and understand the limitations and general caveats involved. Classification using any dataset (i.e. gene expression, immunophenotyping or methylation) can provide numerous challenges to typical machine learning methods. The first major challenge is known as the *"small N, large P*" problem of statistical learning,[483] where the number of variables (*P*; e.g. genes) is typically much greater than the number of patient samples (*N*), with only a small subset of these variable likely to be relevant in distinguishing the different patient classes.[178,483] This challenge can be overcome by using classification algorithms that have been designed to combat this issue, such as iterativeBMA,[178] which was used in this thesis (**Chapters 3, 4** and **5**). A second challenge to machine learning relates to the reliability of the training datasets used to build the classifier. As when using clinical samples, the ability

to accurately detect false-positives, i.e. samples incorrectly assigned to a certain class even though they do not belong to that class in the training dataset used to develop the classifier, is of upmost importance as this could lead to misclassification.[483] This leads to the third challenge relating to the classifier's performance, which can significantly decline when the sample to be classified does not belong to any of the available classes.[483] In these situations, a high rate of misclassification (or false-positives) can occur, which, when the classifier is being used in a diagnostic situation, is unacceptable. This is the most likely cause for the failure of the gene expression classifier developed in **Chapter 4**, which importantly, was developed using a training set comprising only CP-CML with OA values in either the very low (lowest quartile) or very high (highest quartile) ranges. When validation of this classifier was attempted using an independent test set, it failed. However, it should be noted that this test set contained CP-CML patients with OA values in all four quartiles (representative of the typical clinical situation), not only the two extreme quartiles as in the training set, which may have contributed significantly towards the failure to validate.

Furthermore, some of the datasets used to develop the classifiers were comprised of variables (e.g. genes) selected based on statistical testing (**Chapters 4** and **5**), i.e. they contained only statistically significant genes. It is important to note that the *P*-value reported by statistical tests relates to probabilistic significance and does not necessarily indicate a biological one.[484] Therefore, there is the possibility that any number of the selected variables identified by statistical testing is in fact, a false-positive. When using a statistical cut-off of $P < 0.05$, by design, 5% of the results may be false-positive, impacting the gene-list used to develop the classifier and leading to misclassification when an independent test set is analysed.

These caveats and limitations can generally be avoided by good experimental design and by using large patient cohorts for the training sets. Unfortunately, for the experiments performed in this thesis, patient sample availability was limited in some instances, which may have contributed to the inability to successfully validate any of the classifiers developed. While unexpected, the inability to develop a predictive classifier using the data generated was not necessarily unexplainable, as described above. Importantly, this data was still able to be used to identify biomarkers for patient response and OA, while also providing greater insight into the underlying biology of OA and CP-CML.

8.3 Summary and Conclusion

The introduction of TKI therapy has dramatically improved the treatment outcomes observed for the majority of CML patients. However, not all patients respond optimally and now the challenge exists to identify these patients at diagnosis and develop personalised therapies aimed towards obtaining the best possible therapeutic outcome for each patient. It is now known that intracellular drug transport impacts significantly on imatinib response, with the OA assay demonstrated to be an excellent prognostic marker. Therefore, it was postulated that various biological factors intrinsic to CML patients that determine drug transport, are likely related to the phenotype of the predominant leukaemic cell, as well as the deregulation of key signalling pathways associated with transporter expression, modification and activity.

This study investigated the hypothesis that a combination of biological factors would delineate distinct subgroups of CP-CML patients at diagnosis and would impact on patient response, demonstrating that CP-CML is not a uniform disease entity, and providing new

approaches for the sub-classification of CP-CML patients. Three different approaches were used to investigate the underlying biology of OA in CP-CML patients; immunophenotyping, gene expression and global DNA methylation analyses. While a predictive classifier, or single biomarker, for OA was unable to be identified, the data generated was able to provide a detailed insight into the biology of very low OA. Based on the findings presented in this thesis, **Figure 8.1** presents the current hypothesis for the underlying biology associated with very low OA. Differential cell lineage involvement, particularly a decreased T-lymphocyte and an increased erythroid signature; an increased histone gene expression signature; and global hypomethylation, were all observed to be associated with CP-CML patients with very low OA, compared to the other patient groups. Additionally, the identification of a strong association between low *GFI1* expression at diagnosis in CP-CML patients and subsequent blastic transformation even on TKI therapy, may be used as a biomarker of progression in CML. Importantly, global DNA hypomethylation was observed in CP-CML, compared to normal individuals, while global DNA hypermethylation was observed in BC-CML, compared to CP-CML patients, representing the first time global DNA methylation studies have been performed in CML.

In conclusion, the underlying biology of CP-CML patients with very low OA is quite complex, with a combination of differential cell lineage involvement, gene expression and global DNA methylation all contributing. Thus, epigenetic mechanisms, particularly DNA methylation, are likely to play an important role in the underlying biology of CP-CML, contributing to the heterogeneity observed and providing an important indication for combinational therapies involving epigenetic agents to be utilised in CP-CML, particularly in patients likely to respond poorly to TKI alone.

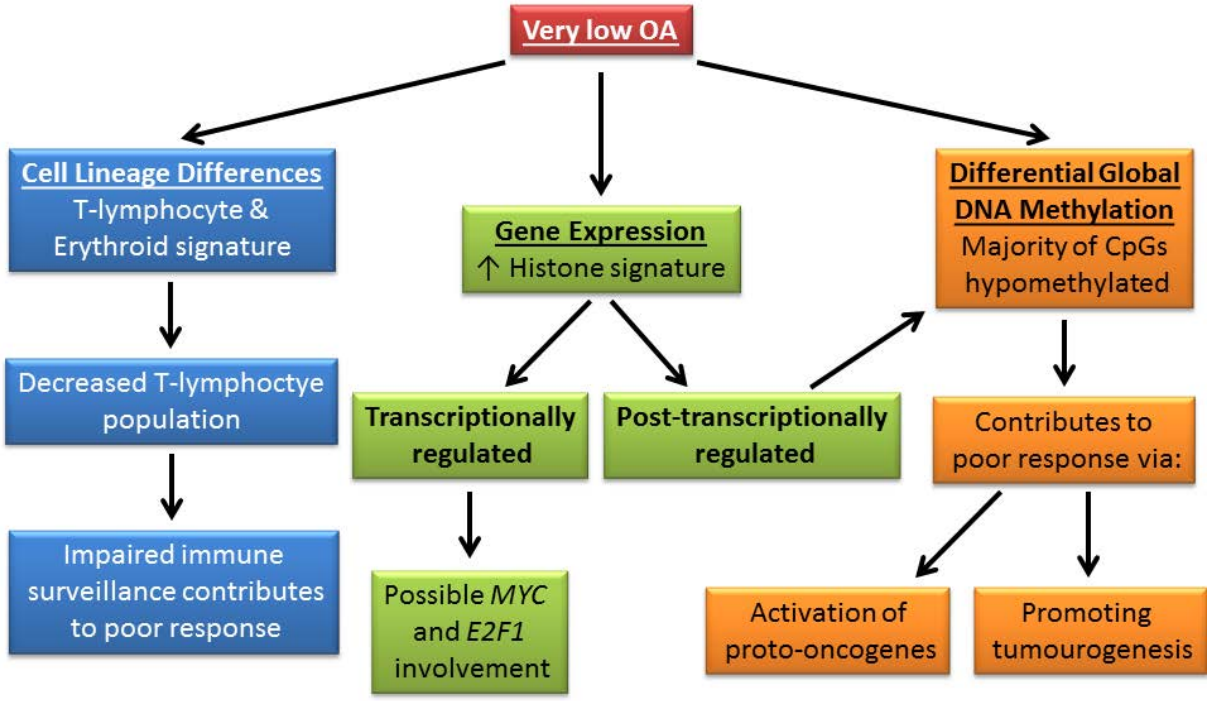


Figure 8.1 Schematic summary of the possible underlying biology associated with CP-CML patients characterised by very low OA

The OCT-1 activity (OA) assay is an excellent prognostic marker for molecular response, with CP-CML patients defined as having very low OA (OA < 4 ng/200,000 cells) associated with poor molecular response, mutation development and disease progression. When very low OA patients were compared to other OA patients (OA > 4 ng/200,000 cells), differential cell lineage, gene expression and DNA methylation patterns were observed. The differential cell lineage involvement identified a possible role for immune surveillance in very low OA. Although the up-regulated histone gene signature observed in very low OA correlated with an enrichment of the *MYC* and *E2F1* transcription factors (TFs), no difference in the regulation of these TFs was observed between OA groups. Finally, global DNA hypomethylation was observed in very low OA, and matched gene expression analysis identified a select number of genes also displaying differential gene expression, indicating that these genes may contribute to the poor response observed in patients with very low OA.

APPENDICES

Appendix 1 Panel 1 - Cell surface antibodies used for immunophenotyping

Tube	Antibody 1	Antibody 2	Volume	Cell Specificity
1	**IgG1 – PE**	-	5 µL	*Isotype control*
2	**CD20 – PE**	-	20 µL	B-cell
3	**CD63 – PE**	-	20 µL	Platelet activation
4	**ITGB5 – PE**	-	20 µL	Integrin beta 5 (cell adhesion)
5	**FZD6 – PE***	-	20 µL	Frizzled 6
6	**LRP6 – PE***	-	20 µL	Low-density lipoprotein receptor-related protein 6
7	**IgG2b – PE**	-	5 µL	*Isotype control*
8	**CD14 – PE**	-	20 µL	Monocyte
9	**ABCG2 – PE**	-	20 µL	MDR protein; transporter involved in CML
10	**IgG2a – PE**	-	5 µL	*Isotype control*
11	**CD71 – PE**	-	20 µL	Erythrocyte
12	**CD59 – PE**	-	20 µL	Erythrocyte
13	**ABCB1 – PE**	-	20 µL	MDR protein; transporter involved in CML
14	**IgG1 – FITC**	-	5 µL	*Isotype control*
15	**CD19 – FITC**	-	20 µL	B-lineage cell
16	**IgM – FITC**	-	5 µL	*Isotype control*
17	**CD66b – FITC**	-	20 µL	Granulocytes
18	**IgG1 – FITC**	**IgG1 – PE**	5 µL/5 µL	*Isotype control*
19	**IgG1 – FITC**	**CD38 – PE**	5 µL/20 µL	Differentiation marker
20	**CD34 – FITC**	**IgG1 – PE**	20 µL/5 µL	Haematopoietic progenitor cell
21	**CD34 – FITC**	**CD38 – PE**	20 µL/20 µL	Haematopoietic progenitor cell
22	**IgG1 – FITC**	**CD7 – PE**	5 µL/20 µL	T-cell precursor
23	**IgG1 – FITC**	**CD25 – PE**	5 µL/20 µL	T-cell

24	**CD3 – FITC**	**IgG1 – PE**	20 µL/5 µL	T-cell
25	**CD3 – FITC**	**CD25 – PE**	20 µL/20 µL	Activated T-cell
26	**CD3 – FITC**	**CD7 – PE**	20 µL/20 µL	T-cell
27	**IgG1 – FITC**	**CD8 – PE**	5 µL/20 µL	Cytotoxic T-cell
28	**CD4 – FITC**	**IgG1 – PE**	20 µL/5 µL	Helper T-cell
29	**CD4 – FITC**	**CD8 – PE**	20 µL/20 µL	Differentiated T-cell
30	**CD4 – FITC**	**CD25 – PE**	20 µL/20 µL	Regulatory T-cell
31	**IgG1 – FITC**	**CD56 – PE**	5 µL/20 µL	NK cell
32	**CD34 – FITC**	**CD56 – PE**	20 µL/20 µL	NK cell
33	**IgG1 – FITC**	**CD33 – PE**	5 µL/20 µL	Myeloid cell
34	**CD34 – FITC**	**CD33 – PE**	20 µL/20 µL	Haematopoietic progenitor cell
35	**IgG1 – FITC**	**CD133 – PE**	5 µL/20 µL	Haematopoietic progenitor cell
36	**CD34 – FITC**	**CD133 – PE**	20 µL/20 µL	Haematopoietic progenitor cell
37	**IgG1 – FITC**	**IgG2a – PE**	5 µL/5 µL	*Isotype control*
38	**IgG1 – FITC**	**HLADR – PE**	5 µL/20 µL	Antigen-presenting cell
39	**CD34 – FITC**	**IgG2a – PE**	20 µL/5 µL	Haematopoietic progenitor cell
40	**CD34 – FITC**	**HLADR – PE**	20 µL/20 µL	Haematopoietic progenitor cell
41	**IgM – FITC**	**IgG1 – PE**	5 µL/5 µL	*Isotype control*
42	**IgM – FITC**	**CD16 – PE**	5 µL/20 µL	Neutrophil
43	**CD15 – FITC**	**IgG1 – PE**	20 µL/5 µL	Granulocyte
44	**CD15 – FITC**	**CD16 – PE**	20 µL/20 µL	Neutrophil
45	**IgG1 – FITC**	**IgG2b – PE**	5 µL/5 µL	*Isotype control*
46	**IgG1 – FITC**	**GlyA – PE**	5 µL/20 µL	Erythrocyte
47	**CD45 – FITC**	**IgG2b – PE**	20 µL/5 µL	Leukocyte
48	**CD45 – FITC**	**GlyA – PE**	20 µL/20 µL	Erythrocyte

* Secondary antibody conjugated with fluorochrome is required

Appendix 2 Panel 2 - Cell surface antibodies used for immunophenotyping

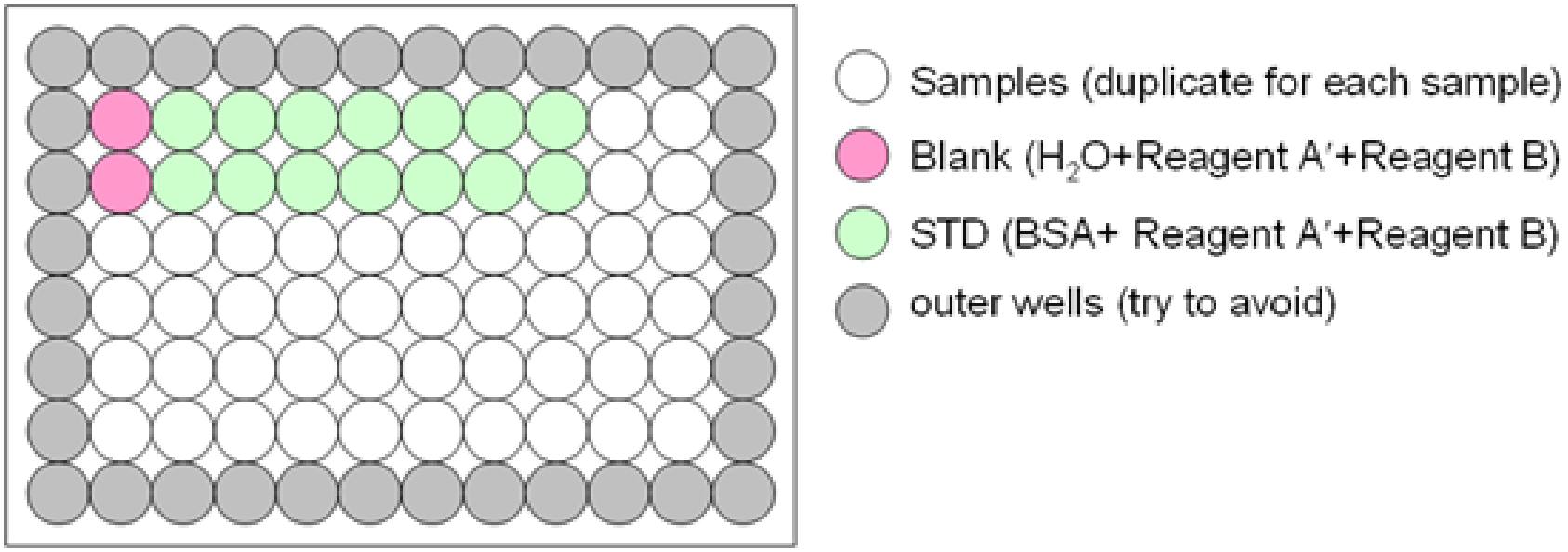
Tube	Antibody 1	Antibody 2	Antibody 3	Volume	Cell Specificity
1	**IgG1 – PE**	**-**	-	5 µL	*Isotype control*
2	**CD20 – PE**	**-**	-	20 µL	B-cell
3	**ITGB5 – PE**	**-**	-	20 µL	Integrin beta 5 (cell adhesion)
4	**CD200R1 – PE**	**-**	-	20 µL	Myeloid cell
5	**FZD6 – PE***	**-**	-	20 µL	Frizzled 6
6	**LRP6 – PE***	**-**	-	20 µL	Low-density lipoprotein receptor-related protein 6
7	**IgG2b – PE**	**-**	-	5 µL	*Isotype control*
8	**CD81 – PE**	**-**	-	20 µL	Signal transduction and adhesion
9	**IgG2a – PE**	**-**	-	5 µL	*Isotype control*
10	**CD99 – PE**	**-**	-	20 µL	Migration and adhesion
11	**IgG1 – FITC**	**IgG1 – PE**	-	5 µL/5 µL	*Isotype control*
12	**IgG1 – FITC**	**CD25 – PE**	-	5 µL/20 µL	T-cell
13	**CD3 – FITC**	**IgG1 – PE**	-	20 µL/5 µL	T-cell
14	**CD3 – FITC**	**CD25 – PE**	-	20 µL/20 µL	Activated T-cell
15	**IgG2a – FITC**	**IgG1 – PE**	-	5 µL/5 µL	*Isotype control*

16	**IgG2a – FITC**	**CD16 – PE**	-	5 µL/20 µL	Neutrophil
17	**CD14 – FITC**	**IgG1 – PE**	-	20 µL/5 µL	Monocyte
18	**CD14 – FITC**	**CD16 – PE**	-	20 µL/20 µL	Monocyte
19	**IgG1 – FITC**	**IgG2b – PE**	**IgM – APC**	5 µL/5 µL/5 µL	*Isotype control*
20	**CD45 – FITC**	**IgG2b – PE**	**IgM – APC**	20 µL/5 µL/5 µL	Leukocyte
21	**IgG1 – FITC**	**GlyA – PE**	**IgM – APC**	5 µL/20 µL/5 µL	Erythrocyte
22	**IgG1 – FITC**	**IgG2b – PE**	**CD36 – APC**	5 µL/5 µL/20 µL	Erythrocyte
23	**CD45 – FITC**	**GlyA – PE**	**IgM – APC**	20 µL/20 µL/5 µL	Erythrocyte
24	**CD45 – FITC**	**IgG2b – PE**	**CD36 – APC**	20 µL/5 µL/20 µL	Erythrocyte
25	**IgG1 – FITC**	**GlyA – PE**	**CD36 – APC**	5 µL/20 µL/20 µL	Erythrocyte
26	**CD45 – FITC**	**GlyA – PE**	**CD36 – APC**	20 µL/20 µL/20 µL	Erythrocyte
27	**IgG1 – FITC**	**IgG1 – PE**	**IgG2b – APC**	5 µL/5 µL/5 µL	*Isotype control*
28	**CD34 – FITC**	**IgG1 – PE**	**IgG2b – APC**	20 µL/5 µL/5 µL	Haematopoietic progenitor cell
29	**IgG1 – FITC**	**CD117 – PE**	**IgG2b – APC**	5 µL/20 µL/5 µL	c-Kit
30	**IgG1 – FITC**	**IgG1 – PE**	**HLA-DR – APC**	5 µL/5 µL/20 µL	Antigen-presenting cell
31	**CD34 – FITC**	**CD117 – PE**	**IgG2b – APC**	20 µL/20 µL/5 µL	Haematopoietic progenitor cell

32	**CD34 – FITC**	**IgG1 – PE**	**HLA-DR – APC**	20 µL/5 µL/20 µL	Haematopoietic progenitor cell
33	**IgG1 – FITC**	**CD117 – PE**	**HLA-DR – APC**	5 µL/20 µL/20 µL	Haematopoietic progenitor cell
34	**CD34 – FITC**	**CD117 – PE**	**HLA-DR – APC**	20 µL/20 µL/20 µL	Haematopoietic progenitor cell

* Secondary antibody conjugated with fluorochrome is required

Appendix 3 96-well plate layout for D_c Protein Assay



Appendix 4 BCR, B3A2 and B2A2 RQ-PCR set-up template

	1	2	3	4	5	6	7	8	9	10	11	12
A	BCR std $6x10^3$	BCR std $6x10^4$	BCR std $6x10^5$	BCR std $6x10^6$	HeLa	Low	High	NTC				
B												
C												
D	B3A2 std $3.2x10^1$	B3A2 std $3.2x10^1$	B3A2 std $3.2x10^2$	B3A2 std $3.2x10^2$	B3A2 std $3.2x10^3$	B3A2 std $3.2x10^4$	B3A2 std $3.2x10^5$	B3A2 std $3.2x10^6$	HeLa	Low	High	NTC
E												
F												
G	B2A2 std $2.7x10^1$	B2A2 std $2.7x10^1$	B2A2 std $2.7x10^2$	B2A2 std $2.7x10^2$	B2A2 std $2.7x10^3$	B2A2 std $2.7x10^4$	B2A2 std $2.7x10^5$	B2A2 std $2.7x10^6$	HeLa	Low	High	NTC
H												

Appendix 5 96 gene TLDA plate format for validation of the OA microarray data

Assay Map:

Hs00152928_m1	Hs00156049_m1	Hs00156390_m1	Hs00156548_m1	Hs00157252_m1	Hs00157731_m1	Hs00158408_m1	Hs00161600_m1	Hs00161720_m1	Hs00167524_m1	Hs99999901_s1	Hs00168586_m1
Hs00152928_m1	Hs00156049_m1	Hs00156390_m1	Hs00156548_m1	Hs00157252_m1	Hs00157731_m1	Hs00158408_m1	Hs00161600_m1	Hs00161720_m1	Hs00167524_m1	Hs99999901_s1	Hs00168586_m1
Hs00197437_m1	Hs00198320_m1	Hs00203102_m1	Hs00204666_m1	Hs00208830_m1	Hs00211234_m1	Hs00211518_m1	Hs00212459_m1	Hs00220767_m1	Hs00221828_m1	Hs00228221_m1	Hs00228938_m1
Hs00197437_m1	Hs00198320_m1	Hs00203102_m1	Hs00204666_m1	Hs00208830_m1	Hs00211234_m1	Hs00211518_m1	Hs00212459_m1	Hs00220767_m1	Hs00221828_m1	Hs00228221_m1	Hs00228938_m1
Hs00357891_s1	Hs00364793_m1	Hs00368313_s1	Hs00368330_s1	Hs00369999_m1	Hs00374317_s1	Hs00375898_m1	Hs00376633_m1	Hs00377326_m1	Hs00381509_m1	Hs00382207_m1	Hs00382922_m1
Hs00357891_s1	Hs00364793_m1	Hs00368313_s1	Hs00368330_s1	Hs00369999_m1	Hs00374317_s1	Hs00375898_m1	Hs00376633_m1	Hs00377326_m1	Hs00381509_m1	Hs00382207_m1	Hs00382922_m1
Hs00855332_g1	Hs00914578_m1	Hs00976711_m1	Hs00984486_m1	Hs00984874_m1	Hs01072206_m1	Hs01111686_g1	Hs01120081_m1	Hs01554515_g1	Hs01633572_g1	Hs01635324_s1	Hs01650625_s1
Hs00855332_g1	Hs00914578_m1	Hs00976711_m1	Hs00984486_m1	Hs00984874_m1	Hs01072206_m1	Hs01111686_g1	Hs01120081_m1	Hs01554515_g1	Hs01633572_g1	Hs01635324_s1	Hs01650625_s1
Hs00152928_m1	Hs00156049_m1	Hs00156390_m1	Hs00156548_m1	Hs00157252_m1	Hs00157731_m1	Hs00158408_m1	Hs00161600_m1	Hs00161720_m1	Hs00167524_m1	Hs99999901_s1	Hs00168586_m1
Hs00152928_m1	Hs00156049_m1	Hs00156390_m1	Hs00156548_m1	Hs00157252_m1	Hs00157731_m1	Hs00158408_m1	Hs00161600_m1	Hs00161720_m1	Hs00167524_m1	Hs99999901_s1	Hs00168586_m1
Hs00197437_m1	Hs00198320_m1	Hs00203102_m1	Hs00204666_m1	Hs00208830_m1	Hs00211234_m1	Hs00211518_m1	Hs00212459_m1	Hs00220767_m1	Hs00221828_m1	Hs00228221_m1	Hs00228938_m1
Hs00197437_m1	Hs00198320_m1	Hs00203102_m1	Hs00204666_m1	Hs00208830_m1	Hs00211234_m1	Hs00211518_m1	Hs00212459_m1	Hs00220767_m1	Hs00221828_m1	Hs00228221_m1	Hs00228938_m1
Hs00357891_s1	Hs00364793_m1	Hs00368313_s1	Hs00368330_s1	Hs00369999_m1	Hs00374317_s1	Hs00375898_m1	Hs00376633_m1	Hs00377326_m1	Hs00381509_m1	Hs00382207_m1	Hs00382922_m1
Hs00357891_s1	Hs00364793_m1	Hs00368313_s1	Hs00368330_s1	Hs00369999_m1	Hs00374317_s1	Hs00375898_m1	Hs00376633_m1	Hs00377326_m1	Hs00381509_m1	Hs00382207_m1	Hs00382922_m1
Hs00855332_g1	Hs00914578_m1	Hs00976711_m1	Hs00984486_m1	Hs00984874_m1	Hs01072206_m1	Hs01111686_g1	Hs01120081_m1	Hs01554515_g1	Hs01633572_g1	Hs01635324_s1	Hs01650625_s1
Hs00855332_g1	Hs00914578_m1	Hs00976711_m1	Hs00984486_m1	Hs00984874_m1	Hs01072206_m1	Hs01111686_g1	Hs01120081_m1	Hs01554515_g1	Hs01633572_g1	Hs01635324_s1	Hs01650625_s1

Gene Symbols:

EGR1	*AZU1*	*CD63*	*CLU*	*DEFA4*	*GNAZ*	*JUP*	*SCP2*	*SLC2A5*	*ALOX12*	*18S*	*P4HB*
EGR1	*AZU1*	*CD63*	*CLU*	*DEFA4*	*GNAZ*	*JUP*	*SCP2*	*SLC2A5*	*ALOX12*	*18S*	*P4HB*
OLFM4	*PLK2*	*ZNF215*	*HSPA4L*	*ENPP4*	*ZC2HC1A*	*AIG1*	*UIMC1*	*RETN*	*ZNF410*	*CWH43*	*LY6G6C*
OLFM4	*PLK2*	*ZNF215*	*HSPA4L*	*ENPP4*	*ZC2HC1A*	*AIG1*	*UIMC1*	*RETN*	*ZNF410*	*CWH43*	*LY6G6C*
JUNB	*CYYR1*	*HIST1H2BJ*	*HIST1H2AG*	*TSPAN18*	*HIST1H2BG*	*RPIA*	*ENKUR*	*SLC15A4*	*PIWIL4*	*GFI1*	*NRGN*
JUNB	*CYYR1*	*HIST1H2BJ*	*HIST1H2AG*	*TSPAN18*	*HIST1H2BG*	*RPIA*	*ENKUR*	*SLC15A4*	*PIWIL4*	*GFI1*	*NRGN*
LDHA	*GNG11*	*GPI*	*MLLT4*	*ANXA4*	*RAB27B*	*DDIT4*	*DAB2*	*MAD2L1*	*HMGN1*	*HMGN2*	*KBTBD6*
LDHA	*GNG11*	*GPI*	*MLLT4*	*ANXA4*	*RAB27B*	*DDIT4*	*DAB2*	*MAD2L1*	*HMGN1*	*HMGN2*	*KBTBD6*
EGR1	*AZU1*	*CD63*	*CLU*	*DEFA4*	*GNAZ*	*JUP*	*SCP2*	*SLC2A5*	*ALOX12*	*18S*	*P4HB*
EGR1	*AZU1*	*CD63*	*CLU*	*DEFA4*	*GNAZ*	*JUP*	*SCP2*	*SLC2A5*	*ALOX12*	*18S*	*P4HB*
OLFM4	*PLK2*	*ZNF215*	*HSPA4L*	*ENPP4*	*ZC2HC1A*	*AIG1*	*UIMC1*	*RETN*	*ZNF410*	*CWH43*	*LY6G6C*
OLFM4	*PLK2*	*ZNF215*	*HSPA4L*	*ENPP4*	*ZC2HC1A*	*AIG1*	*UIMC1*	*RETN*	*ZNF410*	*CWH43*	*LY6G6C*
JUNB	*CYYR1*	*HIST1H2BJ*	*HIST1H2AG*	*TSPAN18*	*HIST1H2BG*	*RPIA*	*ENKUR*	*SLC15A4*	*PIWIL4*	*GFI1*	*NRGN*
JUNB	*CYYR1*	*HIST1H2BJ*	*HIST1H2AG*	*TSPAN18*	*HIST1H2BG*	*RPIA*	*ENKUR*	*SLC15A4*	*PIWIL4*	*GFI1*	*NRGN*
LDHA	*GNG11*	*GPI*	*MLLT4*	*ANXA4*	*RAB27B*	*DDIT4*	*DAB2*	*MAD2L1*	*HMGN1*	*HMGN2*	*KBTBD6*
LDHA	*GNG11*	*GPI*	*MLLT4*	*ANXA4*	*RAB27B*	*DDIT4*	*DAB2*	*MAD2L1*	*HMGN1*	*HMGN2*	*KBTBD6*

Hs00170369_m1	Hs00171574_m1	Hs00173500_m1	Hs00174141_m1	Hs00175044_m1	Hs00178587_m1	Hs00180261_m1	Hs00182949_m1	Hs00185753_m1	Hs00186320_m1	Hs00186809_m1	Hs00189038_m1
Hs00170369_m1	Hs00171574_m1	Hs00173500_m1	Hs00174141_m1	Hs00175044_m1	Hs00178587_m1	Hs00180261_m1	Hs00182949_m1	Hs00185753_m1	Hs00186320_m1	Hs00186809_m1	Hs00189038_m1
Hs00229474_m1	Hs00230289_m1	Hs00232149_m1	Hs00233935_m1	Hs00234028_m1	Hs00251542_m1	Hs00259677_s1	Hs00260478_m1	Hs00263701_m1	Hs00263832_m1	Hs00269023_s1	Hs00355914_m1
Hs00229474_m1	Hs00230289_m1	Hs00232149_m1	Hs00233935_m1	Hs00234028_m1	Hs00251542_m1	Hs00259677_s1	Hs00260478_m1	Hs00263701_m1	Hs00263832_m1	Hs00269023_s1	Hs00355914_m1
Hs00396772_m1	Hs00535083_s1	Hs00542102_m1	Hs00602494_m1	Hs00606684_s1	Hs00609186_m1	Hs00697086_m1	Hs00698316_m1	Hs00740127_m1	Hs00747492_s1	Hs00818527_s1	Hs00826684_m1
Hs00396772_m1	Hs00535083_s1	Hs00542102_m1	Hs00602494_m1	Hs00606684_s1	Hs00609186_m1	Hs00697086_m1	Hs00698316_m1	Hs00740127_m1	Hs00747492_s1	Hs00818527_s1	Hs00826684_m1
Hs01682761_m1	Hs01894900_gH	Hs01921749_s1	Hs02338632_g1	Hs02520942_g1	Hs02596874_g1	Hs00609297_m1	Hs00362795_g1	Hs99999903_m1	Hs99999907_m1	Hs99999908_m1	Hs99999909_m1
Hs01682761_m1	Hs01894900_gH	Hs01921749_s1	Hs02338632_g1	Hs02520942_g1	Hs02596874_g1	Hs00609297_m1	Hs00362795_g1	Hs99999903_m1	Hs99999907_m1	Hs99999908_m1	Hs99999909_m1
Hs00170369_m1	Hs00171574_m1	Hs00173500_m1	Hs00174141_m1	Hs00175044_m1	Hs00178587_m1	Hs00180261_m1	Hs00182949_m1	Hs00185753_m1	Hs00186320_m1	Hs00186809_m1	Hs00189038_m1
Hs00170369_m1	Hs00171574_m1	Hs00173500_m1	Hs00174141_m1	Hs00175044_m1	Hs00178587_m1	Hs00180261_m1	Hs00182949_m1	Hs00185753_m1	Hs00186320_m1	Hs00186809_m1	Hs00189038_m1
Hs00229474_m1	Hs00230289_m1	Hs00232149_m1	Hs00233935_m1	Hs00234028_m1	Hs00251542_m1	Hs00259677_s1	Hs00260478_m1	Hs00263701_m1	Hs00263832_m1	Hs00269023_s1	Hs00355914_m1
Hs00229474_m1	Hs00230289_m1	Hs00232149_m1	Hs00233935_m1	Hs00234028_m1	Hs00251542_m1	Hs00259677_s1	Hs00260478_m1	Hs00263701_m1	Hs00263832_m1	Hs00269023_s1	Hs00355914_m1
Hs00396772_m1	Hs00535083_s1	Hs00542102_m1	Hs00602494_m1	Hs00606684_s1	Hs00609186_m1	Hs00697086_m1	Hs00698316_m1	Hs00740127_m1	Hs00747492_s1	Hs00818527_s1	Hs00826684_m1
Hs00396772_m1	Hs00535083_s1	Hs00542102_m1	Hs00602494_m1	Hs00606684_s1	Hs00609186_m1	Hs00697086_m1	Hs00698316_m1	Hs00740127_m1	Hs00747492_s1	Hs00818527_s1	Hs00826684_m1
Hs01682761_m1	Hs01894900_gH	Hs01921749_s1	Hs02338632_g1	Hs02520942_g1	Hs02596874_g1	Hs00609297_m1	Hs00362795_g1	Hs99999903_m1	Hs99999907_m1	Hs99999908_m1	Hs99999909_m1
Hs01682761_m1	Hs01894900_gH	Hs01921749_s1	Hs02338632_g1	Hs02520942_g1	Hs02596874_g1	Hs00609297_m1	Hs00362795_g1	Hs99999903_m1	Hs99999907_m1	Hs99999908_m1	Hs99999909_m1

CPNE3	*FZD6*	*LPAR1*	*CD59*	*PTH2R*	*PTK2*	*GFI1B*	*GRB14*	*CHIT1*	*HAT1*	*VAMP8*	*CAMP*
CPNE3	*FZD6*	*LPAR1*	*CD59*	*PTH2R*	*PTK2*	*GFI1B*	*GRB14*	*CHIT1*	*HAT1*	*VAMP8*	*CAMP*
ANP32E	*JAM3*	*NFIB*	*LRP6*	*OLR1*	*CDCA7L*	*KBTBD7;KBTBD6*	*C2orf88*	*SUCNR1*	*SLC45A3*	*HIST2H2BE*	*ALDH2*
ANP32E	*JAM3*	*NFIB*	*LRP6*	*OLR1*	*CDCA7L*	*KBTBD7;KBTBD6*	*C2orf88*	*SUCNR1*	*SLC45A3*	*HIST2H2BE*	*ALDH2*
TSPAN33	*FAXDC2*	*BEND2*	*SMOX*	*HIST1H2BB*	*GBE1*	*MYL9*	*TREML1*	*NIPSNAP3A*	*HIST1H4A*	*HIST1H3H*	*ENDOD1*
TSPAN33	*FAXDC2*	*BEND2*	*SMOX*	*HIST1H2BB*	*GBE1*	*MYL9*	*TREML1*	*NIPSNAP3A*	*HIST1H4A*	*HIST1H3H*	*ENDOD1*
SCD	*UGT2B11*	*SPRY2*	*OR2L8*	*REXO1L1P*	*MT-ND2*	*HMBS*	*PES1*	*ACTB*	*B2M*	*GUSB*	*HPRT1*
SCD	*UGT2B11*	*SPRY2*	*OR2L8*	*REXO1L1P*	*MT-ND2*	*HMBS*	*PES1*	*ACTB*	*B2M*	*GUSB*	*HPRT1*
CPNE3	*FZD6*	*LPAR1*	*CD59*	*PTH2R*	*PTK2*	*GFI1B*	*GRB14*	*CHIT1*	*HAT1*	*VAMP8*	*CAMP*
CPNE3	*FZD6*	*LPAR1*	*CD59*	*PTH2R*	*PTK2*	*GFI1B*	*GRB14*	*CHIT1*	*HAT1*	*VAMP8*	*CAMP*
ANP32E	*JAM3*	*NFIB*	*LRP6*	*OLR1*	*CDCA7L*	*KBTBD7;KBTBD6*	*C2orf88*	*SUCNR1*	*SLC45A3*	*HIST2H2BE*	*ALDH2*
ANP32E	*JAM3*	*NFIB*	*LRP6*	*OLR1*	*CDCA7L*	*KBTBD7;KBTBD6*	*C2orf88*	*SUCNR1*	*SLC45A3*	*HIST2H2BE*	*ALDH2*
TSPAN33	*FAXDC2*	*BEND2*	*SMOX*	*HIST1H2BB*	*GBE1*	*MYL9*	*TREML1*	*NIPSNAP3A*	*HIST1H4A*	*HIST1H3H*	*ENDOD1*
TSPAN33	*FAXDC2*	*BEND2*	*SMOX*	*HIST1H2BB*	*GBE1*	*MYL9*	*TREML1*	*NIPSNAP3A*	*HIST1H4A*	*HIST1H3H*	*ENDOD1*
SCD	*UGT2B11*	*SPRY2*	*OR2L8*	*REXO1L1P*	*MT-ND2*	*HMBS*	*PES1*	*ACTB*	*B2M*	*GUSB*	*HPRT1*
SCD	*UGT2B11*	*SPRY2*	*OR2L8*	*REXO1L1P*	*MT-ND2*	*HMBS*	*PES1*	*ACTB*	*B2M*	*GUSB*	*HPRT1*

Appendix 6 24 gene TLDA plate format for validation of the Very Low versus Very High OA RQ-PCR signature

Assay Map:

Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1
Hs00368330_s1	Hs00269023_s1	Hs00747492_s1	Hs00606684_s1	Hs00984486_m1	Hs00374317_s1	Hs00197437_m1	Hs00211518_m1	Hs00818527_s1	Hs00180261_m1	Hs99999901_s1	Hs00152928_m1

Gene Symbols:

HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*
HIST1H2AG	*HIST2H2BE*	*HIST1H4A*	*HIST1H2BB*	*MLLT4*	*HIST1H2BG*	*OLFM4*	*AIG1*	*HIST1H3H*	*GFI1B*	*18S*	*EGR1*

Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1
Hs00368313_s1	Hs00357891_s1	Hs00174141_m1	Hs00602494_m1	Hs99999907_m1	Hs99999908_m1	Hs00609297_m1	Hs00427552_m1	Hs00376633_m1	Hs00230289_m1	Hs00260478_m1	Hs00232149_m1

HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*
HIST1H2BJ	*JUNB*	*CD59*	*SMOX*	*B2M*	*GUSB*	*HMBS*	*SLC22A1*	*ENKUR*	*JAM3*	*C2orf88*	*NFIB*

Appendix 7 Gene expression levels and annotation for the top 89 genes identified from the very low versus very high OA microarray

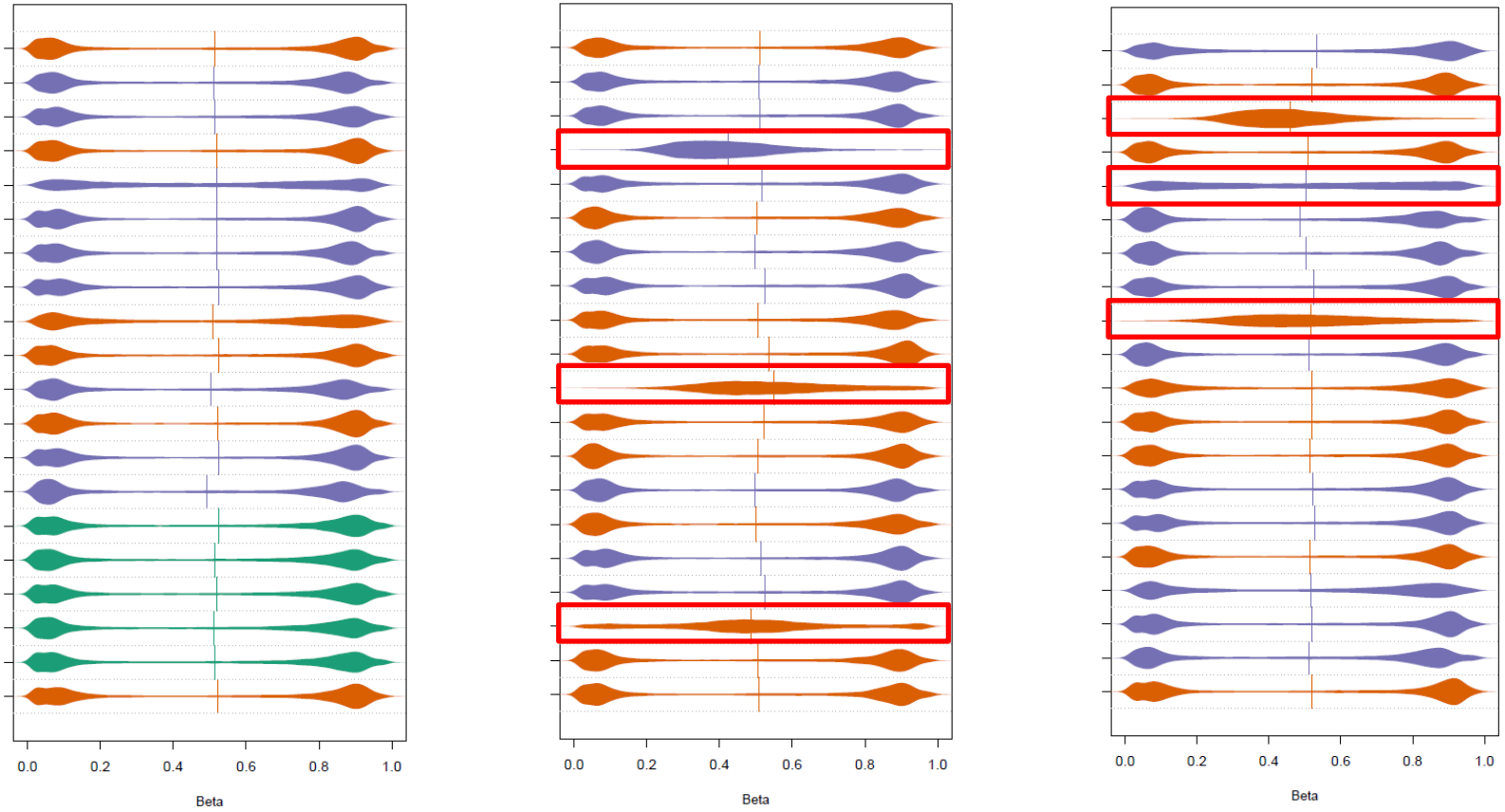
Gene Symbol	Probeset ID	Description	Chr Location	*P*-value	FDR *q*-value	Log Fold Change (VH-VL)	Transcript ID
ZNF410	7975645	Zinc finger protein 410	14q24.3	2.52E-05	0.085	0.709	NM_021188
HMGN2	7982204	High-mobility group nucleosomal binding domain 2	1p36.1	3.03E-05	0.085	1.216	CR542122
SLC2A5	7912224	Solute carrier family 2 (facilitated glucose fructose transporter), member 5	1p36.2	3.80E-05	0.085	1.652	NM_003039
JUNB	8026047	Jun B proto-oncogene	19p13.2	4.86E-05	0.089	-1.018	NM_002229
KBTBD6	7971208	Kelch repeat and BTB (POZ) domain containing 6	13q14.11	8.32E-05	0.125	1.042	NM_152903
KBTBD7	7971218	Kelch repeat and BTB (POZ) domain containing 7	13q14.11	0.00011	0.128	1.013	NM_032138
ND2	8165658	NADH dehydrogenase, subunit 2 (complex I)	Mitochondrial DNA	0.00016	0.140	-1.166	ENST00000361453
RPIA	8043413	Ribose 5-phosphate isomerase A	2p11.2	0.00016	0.140	1.030	NM_144563
C2orf88	8047062	Hypothetical protein MGC13057	2q32.2	0.00017	0.142	-0.964	NM_001042519
HMGN1	8070389	High-mobility group nucleosome binding domain 1	21q22.3|21q22.2	0.00021	0.142	0.624	NM_004965
SCP2	7901513	Sterol carrier protein 2	1p32	0.00021	0.142	0.891	NM_002979
HIST2H2BE	7919637	Histone cluster 2, H2be	1q21-q23	0.00022	0.142	-1.531	NM_003528
SPRY2	7972217	Sprouty homolog 2 (Drosophila)	13q31.1	0.00025	0.142	1.076	NM_005842
SCD	7929816	Stearoyl-CoA desaturase (delta-9-desaturase)	10q24.31	0.00030	0.142	1.404	NM_005063
CDCA7L	8138489	Cell division cycle associated 7-like	7p15.3	0.00030	0.142	1.451	NM_018719
EGR1	8108370	Early growth response 1	5q31.1	0.00037	0.142	-1.507	NM_001964
NFIB	8160138	Nuclear factor IB	9p24.1	0.00040	0.142	-0.897	NM_005596
P4HB	8019762	Prolyl 4-hydroxylase, beta polypeptide	17q25	0.00042	0.142	0.987	NM_000918
LDHA	7938777	Lactate dehydrogenase A	11p15.4	0.00047	0.142	0.998	NM_005566
PTH2R	8047910	Parathyroid hormone 2 receptor	2q33	0.00048	0.142	1.044	NM_005048
LY6G6C	8125042	Lymphocyte antigen 6 complex, locus G6C	6p21.33	0.00049	0.142	-0.517	NM_025261
GRB14	8056327	Growth factor receptor-bound protein 14	2q22-q24	0.00050	0.142	-0.914	NM_004490
RETN	8025278	Resistin	19p13.2	0.00052	0.142	0.923	NM_020415
BEND2	8171545	BEN domain containing 2	Xp22.13	0.00055	0.142	-1.012	NM_153346
VAMP8	8043197	Vesicle-associated membrane protein 8 (endobrevin)	2p12-p11.2	0.00060	0.144	1.124	NM_003761
CAMP	8079590	Cathelicidin antimicrobial peptide	3p21.3	0.00065	0.144	1.798	NM_004345
CHIT1	7923562	Chitinase 1 (chitotriosidase)	1q31-q32	0.00067	0.144	1.609	NM_003465
PIWIL4	7943240	Piwi-like 4 (Drosophila)	11q21	0.00069	0.145	1.415	NM_152431
JUP	8015412	Junction plakoglobin	17q21	0.00072	0.146	-0.813	NM_002230
TSPAN18	7939559	tetraspanin 18	11p11.2	0.00073	0.147	-0.895	NM_130783
FAXDC2	8115397	Fatty acid hydroxylase domain ontaining 2	5q31-q32	0.00081	0.151	-0.933	NM_032385
NIPSNAP3A	8157021	Nipsnap homolog 3A (C. elegans)	9q31.1	0.00083	0.151	1.167	NM_015469
AIG1	8122396	Androgen-induced 1	6q24.2	0.00085	0.151	0.920	NM_016108

ENPP4	8120061	Ectonucleotide pyrophosphatase phosphodiesterase 4 (putative function)	6p12.3	0.00087	0.151	1.041	NM_014936
ANP32E	7919715	Acidic (leucine-rich) nuclear phosphoprotein 32 family, member E	1q21.2	0.00097	0.154	1.013	NM_030920
MYL9	8062312	Myosin, light chain 9, regulatory	20q11.23	0.00107	0.154	-0.908	NM_006097
UGT2B11	8100768	UDP glucuronosyltransfer-ase 2 family, polypeptide B11	4q13.2	0.00111	0.154	-1.226	NM_001073
HIST1H2BJ	8124484	Histone cluster 1, H2bj	6p22.1	0.00135	0.157	-1.302	NM_021058
CD63	7963911	CD63 molecule	12q12-q13	0.00146	0.157	0.635	NM_001780
CPNE3	8147172	Copine III	8q21.3	0.00146	0.157	1.257	NM_003909
OR2L8	7911241	Olfactory receptor, family 2, subfamily L, member 8	1q44	0.00148	0.157	-1.342	NM_001001963
PTK2	8153223	PTK2 protein tyrosine kinase 2	8q24-qter	0.00148	0.157	-0.994	NM_153831
GFI1B	8158939	Growth factor independent 1B transcription repressor	9q34.13	0.00153	0.157	-0.677	NM_004188
HSPA4L	8097335	Heat shock 70kDa protein 4-like	4q28	0.00155	0.157	1.069	NM_014278
HIST1H2BB	8124394	Histone cluster 1, H2bb	6p21.3	0.00190	0.159	-0.942	NM_021062
JAM3	7945262	Junctional adhesion molecule 3	11q25	0.00191	0.159	-1.141	NM_032801
GPI	8027621	Glucose phosphate isomerase	19q13.1	0.00198	0.161	0.914	NM_000175
CYYR1	8069668	Cysteine tyrosine-rich 1	21q21.2	0.00199	0.161	1.060	NM_052954
TREML1	8126269	Triggering receptor expressed on myeloid cells-like 1	6p21.1	0.00221	0.165	-1.024	NM_178174
OLFM4	7969288	Olfactomedin 4	13q21.1	0.00237	0.170	2.262	NM_006418
ENKUR	7932598	Enkurin	10p12.1	0.00251	0.170	-1.369	NM_145010
GNAZ	8071671	Guanine nucleotide binding protein (G protein), alpha z polypeptide	22q11.22	0.00252	0.170	-0.891	NM_002073
GFI1	7917697	Growth factor independent 1 transcription repressor	1p22	0.00255	0.170	0.809	NM_005263
OLR1	7961142	Oxidized low density lipoprotein (lectin-like) receptor 1	12p13.2-p12.3	0.00263	0.170	1.555	NM_002543
GBE1	8088958	Glucan (1,4-alpha-), branching enzyme 1	3p12.3	0.00263	0.170	1.026	NM_000158
ANXA4	8042468	Annexin A4	2p13	0.00264	0.170	1.078	NM_001153
GNG11	8134257	Guanine nucleotide binding protein (G protein), gamma 11	7q21	0.00277	0.174	-1.475	NM_004126
SLC15A4	7967624	Solute carrier family 15, member 4	12q24.32	0.00285	0.175	0.940	NM_145648
TSPAN33	8136067	Tetraspanin 33	7q32.1	0.00287	0.175	-0.868	NM_178562
DDIT4	7928308	DNA-damage-inducible transcript 4	10pter-q26.12	0.00289	0.175	1.139	NM_019058
ALDH2	7958784	Aldehyde dehydrogenase 2 family (mitochondrial)	12q24.2	0.00307	0.179	0.878	NM_000690
REXO1L1P	8151629	REX1, RNA exonuclease 1 homolog (S. cerevisiae)-like 1, pseudogene	8q21.2	0.00325	0.182	-1.619	AF495523
DAB2	8111772	Disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)	5p13	0.00329	0.182	-1.329	NM_001343
HIST1H4A	8117334	Histone cluster 1, H4a	6p21.3	0.00330	0.182	-0.824	NM_003538
NRGN	7944876	Neurogranin (protein kinase C substrate, RC3)	11q24	0.00333	0.182	-1.144	NM_006176
AZU1	8024038	Azurocidin 1	19p13.3	0.00341	0.182	1.134	NM_001700

MLLT4	8123407	Myeloid lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 4	6q27	0.00341	0.182	-0.903	NM_001040001
PLK2	8112202	Polo-like kinase 2 (Drosophila)	5q12.1-q13.2	0.00347	0.182	-1.115	NM_006622
ZC2HC1A	8147019	Zinc finger, C2HC-type containing 1A	8q21.12	0.00365	0.182	-1.225	AF151820
RAB27B	8021301	RAB27B, member RAS oncogene family	18q21.2	0.00369	0.182	-1.775	NM_004163
ZNF215	7938183	Zinc finger protein 215	11p15.4	0.00378	0.183	0.904	NM_013250
HIST1H2BG	8124423	Histone cluster 1, H2bg	6p21.3	0.00379	0.183	-1.125	NM_003518
HAT1	8046346	Histone acetyltransferase 1	2q31.2-q33.1	0.00387	0.184	0.666	NM_003642
CLU	8149927	Clusterin	8p21-p12	0.00388	0.184	-1.377	NM_001831
HIST1H3H	8117589	Histone cluster 1, H3h	6p22-p21.3	0.00394	0.184	-1.086	NM_003536
ALOX12	8004221	Arachidonate 12-lipoxygenase	17p13.1	0.00394	0.184	-1.029	NM_000697
SLC45A3	7923792	Solute carrier family 45, member 3	1q32.1	0.00394	0.184	-1.011	NM_033102
ENDOD1	7943293	Endonuclease domain containing 1	11q21	0.00402	0.186	-0.968	NM_015036
DEFA4	8149109	Defensin, alpha 4, corticostatin	8p23	0.00402	0.186	1.805	NM_001925
UIMC1	8165703	Ubiquitin interaction motif containing 1	5q35.2	0.00414	0.186	-0.987	AF284753
LPAR1	8163257	Lysophosphatidic acid receptor 1	9q31.3	0.00445	0.189	-0.919	NM_057159
CD59	7947425	CD59 molecule, complement regulatory protein	11p13	0.00449	0.190	0.791	NM_203330
SMOX	8060745	Spermine oxidase	20p13	0.00476	0.192	-0.970	NM_175839
MAD2L1	8102560	MAD2 mitotic arrest deficient-like 1 (yeast)	4q27	0.00482	0.192	0.945	NM_002358
CWH43	8095005	Cell wall biogenesis 43 C-terminal homolog	4p12-p11	0.00485	0.192	-1.495	ENST00000226432
HIST1H2AG	8117535	Histone cluster 1, H2ag	6p22.1	0.00533	0.198	-0.934	NM_021064
LRP6	7961339	Low density lipoprotein receptor-related protein 6	12p11-p13	0.00583	0.200	-0.505	NM_002336
FZD6	8147766	Frizzled homolog 6 (Drosophila)	8q22.3-q23.1	0.00599	0.200	0.827	NM_003506
SUCNR1	8083422	Succinate receptor 1	3q24-q25.1	0.00602	0.200	1.227	NM_033050

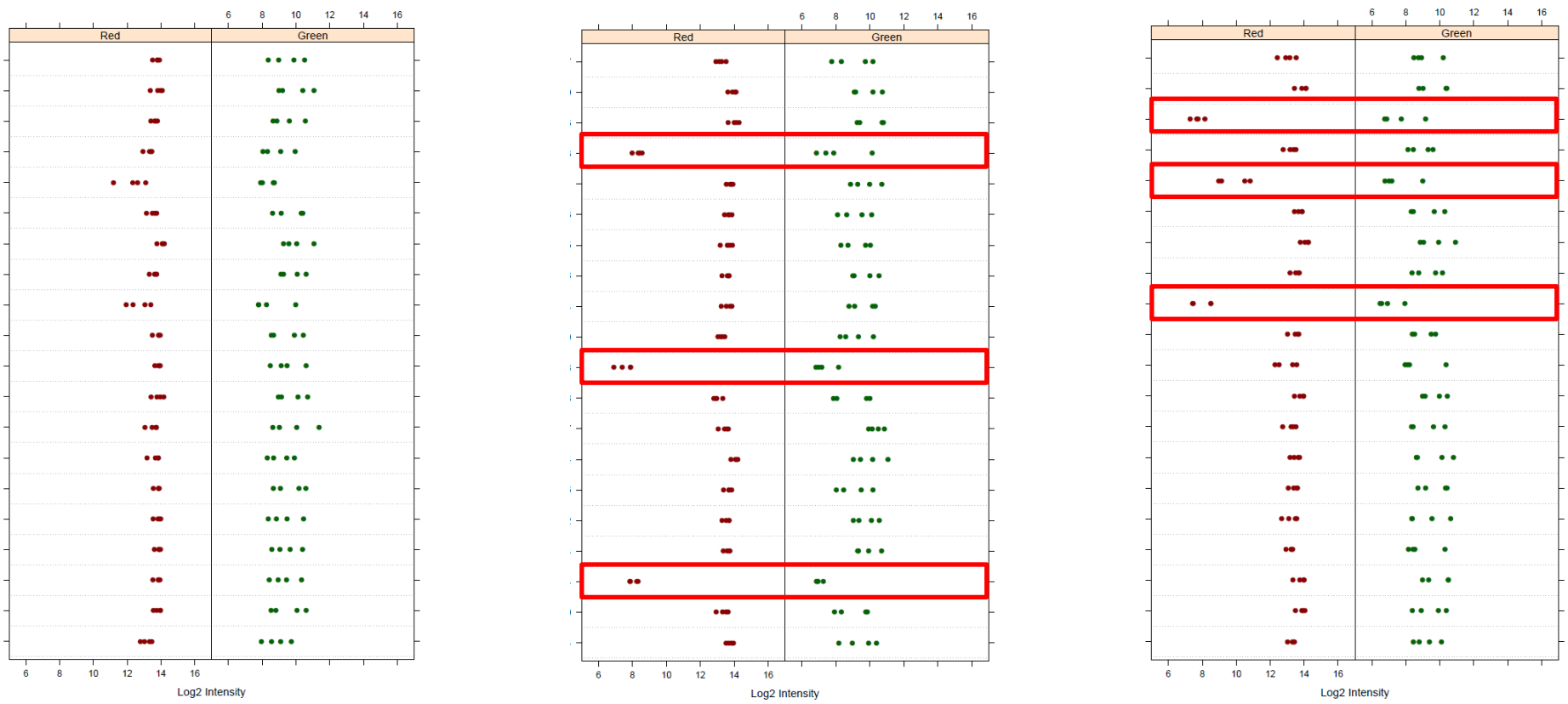
Abbreviations: Chr – chromosome, FDR – false discovery rate, VH – very high OA, VL – very low OA.

Appendix 8 Quality control analysis in *minfi* identifies samples with abnormal methylation profiles



Beanplots demonstrating the range of β-values for each of the 60 patients analysed on the 450K DNA methylation arrays. **Purple** represents patients with very low OA, while **orange** represents patients with other OA. **Green** represents normal individuals. **Red** squares indicate the 6 patients that failed QC.

Appendix 9 The 6 samples identified as abnormal by QC all failed at the bisulphite conversion step



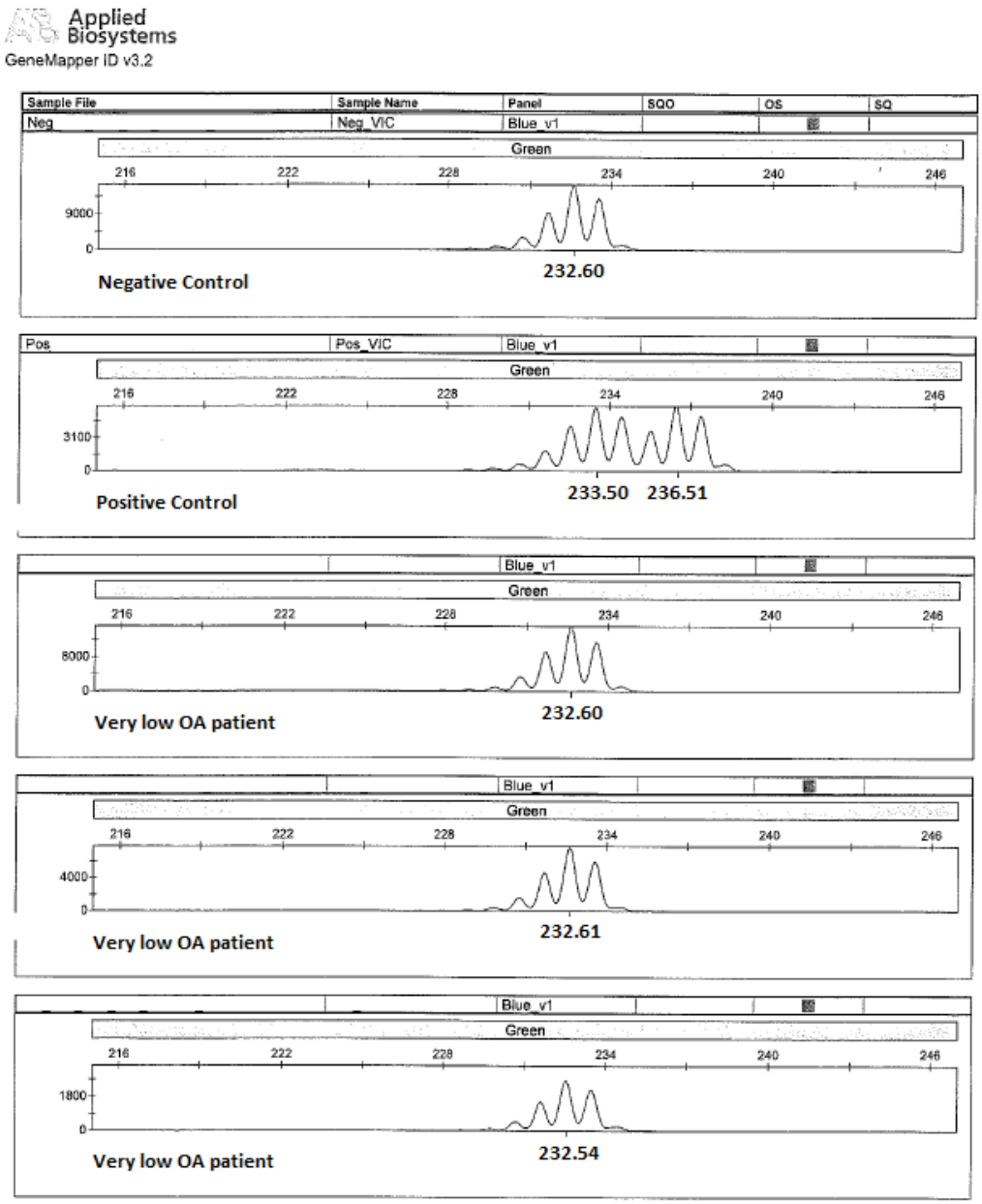
Control strip plots for the bisulphite conversion II stage demonstrate that the 6 patients identified with abnormal methylation profiles during QC (refer **Section 2.6.10**) all failed at the bisulphite conversion II step. Bisulphite conversion II control probes are plotted for each patient. **Red** squares indicate the 6 patients that failed bisulphite conversion. Red and Green refers to the fluorescence colour of the probe.

Appendix 10 Correlation between SNP allele frequency and DNA methylation levels for 5 SNP-CpGs

	cg01591343 - *WDR27*			cg14179288 - *TLE4*			cg14351440 - *MCF2*			cg23825092 - *TDRD1*			cg24100293 – *CASC15*		
	450K β-value	**Methylation group**	**rs1047910 SNP genotype**	**450K β-value**	**Methylation group**	**rs2297498 SNP genotype**	**450K β-value**	**Methylation group**	**rs5907612 SNP genotype**	**450K β-value**	**Methylation group**	**rs78932533 SNP genotype**	**450K β-value**	**Methylation group**	**rs9368372 SNP genotype**
Patient1	0.856	High	C	0.668	Med	GA	0.161	Low	A	0.882	High	A	0.888	High	C
Patient2	0.888	High	C	0.938	High	G	0.815	High	G	0.123	Low	A	0.898	High	C
Patient3	0.814	High	C	0.610	Med	GA	0.756	High	G	0.667	Low	A	0.915	High	C
Patient4	0.912	High	C	0.017	Low	A	0.093	Low	A	0.926	High	A	0.574	Low	CA
Patient5	0.913	High	C	0.566	Med	GA	0.055	Low	A	0.919	High	A	0.930	High	C
Patient6	0.914	High	C	0.636	Med	GA	0.079	Low	A	0.932	High	A	0.910	High	C
Patient7	0.930	High	C	0.588	Med	GA	0.516	Med	GA	0.567	Low	A	0.560	Low	CA
Patient8	0.877	High	C	0.026	Low	A	0.098	Low	A	0.890	High	A	0.899	High	C
Patient9	0.883	High	C	0.004	Low	A	0.787	High	G	0.115	Low	A	0.906	High	C
Patient10	0.865	High	C	0.008	Low	A	0.123	Low	A	0.914	High	A	0.911	High	C
Patient11	0.881	High	C	0.007	Low	A	0.062	Low	A	0.907	High	A	0.891	High	C
Patient12	0.890	High	C	0.004	Low	A	0.756	High	G	0.929	High	A	0.909	High	C
Patient13	0.871	High	C	0.016	Low	A	0.777	High	G	0.909	High	A	0.914	High	C
Patient14	0.889	High	C	0.521	Med	GA	0.484	Med	GA	0.918	High	A	0.899	High	C
Patient15	0.916	High	C	0.019	Low	A	0.766	High	GA	0.889	High	A	0.942	High	C
Patient16	0.816	High	C	0.939	High	G	0.073	Low	A	0.919	High	A	0.599	Low	CA
Patient17	0.894	High	C	0.932	High	G	0.797	High	G	0.933	High	A	0.931	High	C
Patient18	0.907	High	C	0.622	Med	GA	0.064	Low	A	0.919	High	A	0.928	High	C
Patient19	0.883	High	C	0.890	High	G	0.527	Med	GA	0.893	High	A	0.924	High	C
Patient20	0.909	High	C	0.566	Med	GA	0.426	Med	GA	0.906	High	A	0.909	High	C
Patient21	0.912	High	C	0.596	Med	GA	0.720	High	G	0.921	High	A	0.902	High	C
Patient22	0.886	High	C	0.009	Low	A	0.092	Low	A	0.900	High	A	0.913	High	C

Patient23	0.897	High	C	0.025	Low	A	0.153	Low	A	0.936	High	A	0.909	High	C
Patient24	0.898	High	C	0.559	Med	GA	0.099	Low	A	0.879	High	A	0.901	High	C
Patient25	0.860	High	C	0.001	Low	A	0.340	Med	GA	0.781	Low	A	0.660	Low	CA
Patient26	0.898	High	C	0.095	Low	A	0.061	Low	A	0.912	High	A	0.906	High	C
Patient27	0.018	Low	N/A	0.156	Low	N/A	0.776	High	N/A	0.862	High	N/A	0.904	High	N/A
Patient28	0.048	Low	T	0.760	Med	GA	0.557	Med	G	0.823	Low	A	0.867	High	C
Patient29	0.032	Low	T	0.627	Med	GA	0.795	High	G	0.505	Low	A	0.548	Low	CA
Patient30	0.512	Med	TC	0.021	Low	A	0.156	Low	A	0.892	High	A	0.898	High	C
Patient31	0.611	Med	TC	0.626	Med	GA	0.745	High	G	0.615	Low	A	0.525	Low	CA
Patient32	0.590	Med	TC	0.605	Med	GA	0.791	High	G	0.312	Low	A	0.521	Low	CA
Patient33	0.577	Med	TC	0.933	High	G	0.740	High	G	0.915	High	A	0.591	Low	CA
Patient34	0.593	Med	TC	0.917	High	G	0.800	High	G	0.065	Low	A	0.911	High	C
Patient35	0.606	Med	TC	0.940	High	G	0.756	High	G	0.681	Low	A	0.027	Low	A
Patient36	0.560	Med	TC	0.657	Med	GA	0.542	Med	GA	0.908	High	A	0.028	Low	A
Patient37	0.638	Med	TC	0.933	High	G	0.771	High	G	0.112	Low	A	0.537	Low	CA
Patient38	0.598	Med	TC	0.591	Med	GA	0.833	High	G	0.119	Low	A	0.011	Low	A
Patient39	0.573	Med	TC	0.544	Med	GA	0.686	High	G	0.130	Low	A	0.881	High	C
Patient40	0.558	Med	TC	0.573	Med	GA	0.049	Low	A	0.903	High	A	0.914	High	C
Patient41	0.671	Med	TC	0.603	Med	GA	0.828	High	G	0.112	Low	A	0.878	High	C
Patient42	0.593	Med	TC	0.189	Low	A	0.393	Med	GA	0.109	Low	A	0.894	High	C
Patient43	0.446	Med	TC	0.576	Med	GA	0.112	Low	A	0.571	Low	A	0.428	Low	CA
Patient44	0.598	Med	TC	0.586	Med	GA	0.805	High	G	0.885	High	A	0.550	Low	CA
Patient45	0.626	Med	TC	0.004	Low	A	0.805	High	G	0.910	High	A	0.899	High	C
Patient46	0.588	Med	TC	0.548	Med	GA	0.495	Med	GA	0.917	High	A	0.910	High	C
Patient47	0.687	Med	TC	0.600	Med	GA	0.811	High	G	0.727	Low	A	0.898	High	C
Patient48	0.552	Med	TC	0.611	Med	GA	0.053	Low	A	0.907	High	A	0.917	High	C
Patient49	0.604	Med	TC	0.025	Low	A	0.757	High	G	0.857	Low	A	0.935	High	C

Appendix 11 GeneScan analysis of CP-CML patients to identify *NPM1* mutations



This printout from the GeneScan™ analysis is representative of the results obtained for all BC- and CP-CML patients analysed for *NPM1* mutations. A negative (no *NPM1* mutation) and positive control (*NPM1* mutation present; identified by the double peak at ≈232 and 236 bp) were included in each run. Three very low OA patients are presented, with only peaks at 232 bp present, indicating no *NPM1* mutation is present.

PUBLICATIONS ARISING FROM THIS THESIS

Statement of Authorship

Title of Paper	NPM1 mutations occur rarely or not at all in chronic myeloid leukaemia patients in chronic phase of blast crisis
Publication Status	◉ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	Watkins DB, Hughes TP, White DL, D'Andrea RJ. 2013. NPM1 mutations occur rarely or not at all in chronic myeloid leukaemia patients in chronic phase of blast crisis. Leukemia, 27(2):489-90. Impact factor: 10.164.

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Dale B. Watkins		
Contribution to the Paper	Performed analysis on all samples, interpreted data and wrote manuscript.		
Signature		Date	25/11/13

Name of Co-Author	Timothy P. Hughes		
Contribution to the Paper	Supervised development of work, helped in data interpretation and evaluated and edited the manuscript.		
Signature		Date	29/11/13

Name of Co-Author	Deborah L. White		
Contribution to the Paper	Supervised development of work, helped in data interpretation and evaluated and edited the manuscript.		
Signature		Date	28/11/2013

Name of Co-Author	Richard J. D'Andrea		
Contribution to the Paper	Supervised development of work, helped in data interpretation, evaluated and edited the manuscript and acted as corresponding author.		
Signature		Date	28/11/13

Watkins, D.B., Hughes, T.P., White, D.L. & D'Andrea, R.J. (2013) NPM1 mutations occur rarely or not at all in chronic myeloid leukaemia patients in chronic phase of blast crisis.
Leukemia, v. 27(2), pp. 489-490

Statement of Authorship

Title of Paper	Low *GFI1* expression in white blood cells of CP-CML patients at diagnosis is strongly associated with subsequent blastic transformation
Publication Status	◉ Published, ○ Accepted for Publication, ○ Submitted for Publication, ○ Publication style
Publication Details	Kok CH*, Watkins DB*, Leclercq T, D'Andrea RJ, Hughes TP, White DL. 2013. Low *GFI1* expression in white blood cells of CP-CML patients at diagnosis is strongly associated with subsequent blastic transformation. Leukemia, 27(6):1427-30. Impact factor: 10.164. * denotes equal first-authors.

Author Contributions

By signing the Statement of Authorship, each author certifies that their stated contribution to the publication is accurate and that permission is granted for the publication to be included in the candidate's thesis.

Name of Principal Author (Candidate)	Dale B. Watkins		
Contribution to the Paper	Devised experiment, performed analysis on all samples, interpreted data and wrote manuscript.		
Signature		Date	25/11/13

Name of Co-Author	Chung H. Kok		
Contribution to the Paper	Devised experiment, performed analysis on all samples, interpreted data and wrote manuscript.		
Signature		Date	28/11/13

Name of Co-Author	Tamara Leclercq		
Contribution to the Paper	Performed some data analysis, helped in data interpretation and evaluated the manuscript.		
Signature		Date	28/11/13

Name of Co-Author	Richard J. D'Andrea		
Contribution to the Paper	Supervised development of research, helped in data interpretation, evaluated and edited the manuscript.		
Signature		Date	28/11/13

Continued on the next page...

Name of Co-Author	Timothy P. Hughes		
Contribution to the Paper	Supervised development of research, helped in data interpretation, evaluated and edited the manuscript.		
Signature		Date	29/11/13

Name of Co-Author	Deborah L. White		
Contribution to the Paper	Supervised development of research, helped in data interpretation, evaluated and edited the manuscript, and acted as the corresponding author.		
Signature		Date	28/11/2013

Kok, C.H., Watkins, D.B., Leclercq, T., D'Andrea, R.J., Hughes, T.P. & White, D.L. (2013) Low GFI1 expression in white blood cells of CP-CML patients at diagnosis is strongly associated with subsequent blastic transformation.
Leukemia, v. 27(6), pp. 1427-1430

REFERENCES

1. Bennett, J. Case of hypertrophy of the spleen and liver in which death took place from suppuration of the blood. *Edinburgh Medical and Surgical Journal*, 413-423 (1845).
2. Nowell, P.C. & Hungerford, D.A. A minute chromosome in human chronic granulocytic leukemia. *Science* **132**, 1497-1501 (1960).
3. Clarkson, B., Strife, A., Wisniewski, D., Lambek, C.L. & Liu, C. Chronic myelogenous leukemia as a paradigm of early cancer and possible curative strategies. *Leukemia* **17**, 1211-1262 (2003).
4. Sessions, J. Chronic myeloid leukemia in 2007. *J Manag Care Pharm* **13**, 4-7 (2007).
5. Spiers, A.S. The clinical features of chronic granulocytic leukaemia. *Clin Haematol* **6**, 77-95 (1977).
6. *Chronic myelogenous leukemia (CML)*, (Antigenics Inc., New York City, U.S.A, Antigenics. 2008).
7. Fleischman, E.W*., et al.* Correlations between the clinical course, characteristics of blast cells, and karyotype patterns in chronic myeloid leukemia. *Hum Genet* **58**, 285-293 (1981).
8. Avanzi, G.C*., et al.* Duplication of Ph and of 9q+ chromosomes during the blastic transformation of a CML case. *Cancer Genet Cytogenet* **29**, 57-63 (1987).
9. Hernandez-Boluda, J.C., Cervantes, F., Costa, D., Carrio, A. & Montserrat, E. Blast crisis of Ph-positive chronic myeloid leukemia with isochromosome 17q: report of 12 cases and review of the literature. *Leuk Lymphoma* **38**, 83-90 (2000).
10. Wang, Y*., et al.* Determination of secondary chromosomal aberrations of chronic myelocytic leukemia. *Cancer Genet Cytogenet* **153**, 53-56 (2004).
11. Sokal, J.E., Baccarani, M., Russo, D. & Tura, S. Staging and prognosis in chronic myelogenous leukemia. *Semin Hematol* **25**, 49-61 (1988).
12. Yong, A.S., Szydlo, R.M., Goldman, J.M., Apperley, J.F. & Melo, J.V. Molecular profiling of CD34+ cells identifies low expression of CD7, along with high expression of proteinase 3 or elastase, as predictors of longer survival in patients with CML. *Blood* **107**, 205-212 (2006).
13. Rowley, J.D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290-293 (1973).
14. Bartram, C.R*., et al.* Translocation of c-ab1 oncogene correlates with the presence of a Philadelphia chromosome in chronic myelocytic leukaemia. *Nature* **306**, 277-280 (1983).
15. Groffen, J*., et al.* Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. *Cell* **36**, 93-99 (1984).
16. Heisterkamp, N*., et al.* Localization of the c-ab1 oncogene adjacent to a translocation break point in chronic myelocytic leukaemia. *Nature* **306**, 239-242 (1983).
17. Lugo, T.G., Pendergast, A.M., Muller, A.J. & Witte, O.N. Tyrosine kinase activity and transformation potency of bcr-abl oncogene products. *Science* **247**, 1079-1082 (1990).
18. McLaughlin, J., Chianese, E. & Witte, O.N. In vitro transformation of immature hematopoietic cells by the P210 BCR/ABL oncogene product of the Philadelphia chromosome. *Proc Natl Acad Sci U S A* **84**, 6558-6562 (1987).
19. Melo, J.V. The diversity of BCR-ABL fusion proteins and their relationship to leukemia phenotype. *Blood* **88**, 2375-2384 (1996).
20. Deininger, M.W., Goldman, J.M. & Melo, J.V. The molecular biology of chronic myeloid leukemia. *Blood* **96**, 3343-3356 (2000).
21. Institute, N.C. NCI Dictionary of Cancer Terms: Philadelphia Chromosome. (2013).

22. Melo, J.V., Gordon, D.E., Cross, N.C. & Goldman, J.M. The ABL-BCR fusion gene is expressed in chronic myeloid leukemia. *Blood* **81**, 158-165 (1993).
23. Pane, F*., et al.* Neutrophilic-chronic myeloid leukemia: a distinct disease with a specific molecular marker (BCR/ABL with C3/A2 junction). *Blood* **88**, 2410-2414 (1996).
24. Hillman, R.S., Ault, K.A. & Rinder, H.M. *Hematology in clinical practice: A guide to diagnosis and management*, (The McGraw-Hill Companies, Inc., United States of America, 2005).
25. Clarkson, B. & Strife, A. Linkage of proliferative and maturational abnormalities in chronic myelogenous leukemia and relevance to treatment. *Leukemia* **7**, 1683-1721 (1993).
26. Daley, G.Q., Van Etten, R.A. & Baltimore, D. Induction of chronic myelogenous leukemia in mice by the P210bcr/abl gene of the Philadelphia chromosome. *Science* **247**, 824-830 (1990).
27. Li, S., Ilaria, R.L., Jr., Million, R.P., Daley, G.Q. & Van Etten, R.A. The P190, P210, and P230 forms of the BCR/ABL oncogene induce a similar chronic myeloid leukemia-like syndrome in mice but have different lymphoid leukemogenic activity. *J Exp Med* **189**, 1399-1412 (1999).
28. Laneuville, P. Abl tyrosine protein kinase. *Semin Immunol* **7**, 255-266 (1995).
29. Jackson, P. & Baltimore, D. N-terminal mutations activate the leukemogenic potential of the myristoylated form of c-abl. *EMBO J* **8**, 449-456 (1989).
30. McWhirter, J.R., Galasso, D.L. & Wang, J.Y. A coiled-coil oligomerization domain of Bcr is essential for the transforming function of Bcr-Abl oncoproteins. *Mol Cell Biol* **13**, 7587-7595 (1993).
31. Maru, Y. & Witte, O.N. The BCR gene encodes a novel serine/threonine kinase activity within a single exon. *Cell* **67**, 459-468 (1991).
32. Pendergast, A.M., Muller, A.J., Havlik, M.H., Maru, Y. & Witte, O.N. BCR sequences essential for transformation by the BCR-ABL oncogene bind to the ABL SH2 regulatory domain in a non-phosphotyrosine-dependent manner. *Cell* **66**, 161-171 (1991).
33. Groffen, J., Heisterkamp, N., Reynolds, F.H., Jr. & Stephenson, J.R. Homology between phosphotyrosine acceptor site of human c-abl and viral oncogene products. *Nature* **304**, 167-169 (1983).
34. Abelson, H.T. & Rabstein, L.S. Lymphosarcoma: virus-induced thymic-independent disease in mice. *Cancer Res* **30**, 2213-2222 (1970).
35. McWhirter, J.R. & Wang, J.Y. An actin-binding function contributes to transformation by the Bcr-Abl oncoprotein of Philadelphia chromosome-positive human leukemias. *EMBO J* **12**, 1533-1546 (1993).
36. Van Etten, R.A., Jackson, P. & Baltimore, D. The mouse type IV c-abl gene product is a nuclear protein, and activation of transforming ability is associated with cytoplasmic localization. *Cell* **58**, 669-678 (1989).
37. Sawyers, C.L., McLaughlin, J., Goga, A., Havlik, M. & Witte, O. The nuclear tyrosine kinase c-Abl negatively regulates cell growth. *Cell* **77**, 121-131 (1994).
38. Yuan, Z.M*., et al.* p73 is regulated by tyrosine kinase c-Abl in the apoptotic response to DNA damage. *Nature* **399**, 814-817 (1999).
39. Lewis, J.M. & Schwartz, M.A. Integrins regulate the association and phosphorylation of paxillin by c-Abl. *J Biol Chem* **273**, 14225-14230 (1998).
40. Boguski, M.S. & McCormick, F. Proteins regulating Ras and its relatives. *Nature* **366**, 643-654 (1993).

41. Diekmann, D*., et al.* Bcr encodes a GTPase-activating protein for p21rac. *Nature* **351**, 400-402 (1991).
42. Ma, G., Lu, D., Wu, Y., Liu, J. & Arlinghaus, R.B. Bcr phosphorylated on tyrosine 177 binds Grb2. *Oncogene* **14**, 2367-2372 (1997).
43. Daley, G.Q. & Ben-Neriah, Y. Implicating the bcr/abl gene in the pathogenesis of Philadelphia chromosome-positive human leukemia. *Adv Cancer Res* **57**, 151-184 (1991).
44. Bedi, A., Zehnbauer, B.A., Barber, J.P., Sharkis, S.J. & Jones, R.J. Inhibition of apoptosis by BCR-ABL in chronic myeloid leukemia. *Blood* **83**, 2038-2044 (1994).
45. Hanahan, D. & Weinberg, R.A. The hallmarks of cancer. *Cell* **100**, 57-70 (2000).
46. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
47. O'Hare, T., Deininger, M.W., Eide, C.A., Clackson, T. & Druker, B.J. Targeting the BCR-ABL signaling pathway in therapy-resistant Philadelphia chromosome-positive leukemia. *Clin Cancer Res* **17**, 212-221 (2011).
48. Cilloni, D. & Saglio, G. Molecular pathways: BCR-ABL. *Clin Cancer Res* **18**, 930-937 (2012).
49. Sawyers, C.L., McLaughlin, J. & Witte, O.N. Genetic requirement for Ras in the transformation of fibroblasts and hematopoietic cells by the Bcr-Abl oncogene. *J Exp Med* **181**, 307-313 (1995).
50. Melo, J.V. & Chuah, C. Novel agents in CML therapy: tyrosine kinase inhibitors and beyond. *Hematology Am Soc Hematol Educ Program*, 427-435 (2008).
51. Chai, S.K., Nichols, G.L. & Rothman, P. Constitutive activation of JAKs and STATs in BCR-Abl-expressing cell lines and peripheral blood cells derived from leukemic patients. *J Immunol* **159**, 4720-4728 (1997).
52. Frank, D.A. & Varticovski, L. BCR/abl leads to the constitutive activation of Stat proteins, and shares an epitope with tyrosine phosphorylated Stats. *Leukemia* **10**, 1724-1730 (1996).
53. Carlesso, N., Frank, D.A. & Griffin, J.D. Tyrosyl phosphorylation and DNA binding activity of signal transducers and activators of transcription (STAT) proteins in hematopoietic cell lines transformed by Bcr/Abl. *J Exp Med* **183**, 811-820 (1996).
54. Nosaka, T*., et al.* STAT5 as a molecular regulator of proliferation, differentiation and apoptosis in hematopoietic cells. *EMBO J* **18**, 4754-4765 (1999).
55. Sillaber, C., Gesbert, F., Frank, D.A., Sattler, M. & Griffin, J.D. STAT5 activation contributes to growth and viability in Bcr/Abl-transformed cells. *Blood* **95**, 2118-2125 (2000).
56. de Groot, R.P., Raaijmakers, J.A., Lammers, J.W., Jove, R. & Koenderman, L. STAT5 activation by BCR-Abl contributes to transformation of K562 leukemia cells. *Blood* **94**, 1108-1112 (1999).
57. Skorski, T*., et al.* Phosphatidylinositol-3 kinase activity is regulated by BCR/ABL and is required for the growth of Philadelphia chromosome-positive cells. *Blood* **86**, 726-736 (1995).
58. Gordon, M.Y., Dowding, C.R., Riley, G.P., Goldman, J.M. & Greaves, M.F. Altered adhesive interactions with marrow stroma of haematopoietic progenitor cells in chronic myeloid leukaemia. *Nature* **328**, 342-344 (1987).
59. Dai, Z*., et al.* Oncogenic Abl and Src tyrosine kinases elicit the ubiquitin-dependent degradation of target proteins through a Ras-independent pathway. *Genes Dev* **12**, 1415-1424 (1998).
60. Sureda, A. & Hernandez-Bronchud, M. The management of chronic myeloid leukaemia--a case history. *Ann Oncol* **8**, 91-96 (1997).

61. Hehlmann, R.*, et al.* Randomized comparison of busulfan and hydroxyurea in chronic myelogenous leukemia: prolongation of survival by hydroxyurea. The German CML Study Group. *Blood* **82**, 398-407 (1993).
62. Talpaz, M., McCredie, K.B., Mavligit, G.M. & Gutterman, J.U. Leukocyte interferon-induced myeloid cytoreduction in chronic myelogenous leukemia. *Blood* **62**, 689-692 (1983).
63. Talpaz, M.*, et al.* Hematologic remission and cytogenetic improvement induced by recombinant human interferon alpha A in chronic myelogenous leukemia. *N Engl J Med* **314**, 1065-1069 (1986).
64. Kantarjian, H.M.*, et al.* Prolonged survival in chronic myelogenous leukemia after cytogenetic response to interferon-alpha therapy. The Leukemia Service. *Ann Intern Med* **122**, 254-261 (1995).
65. Hehlmann, R.*, et al.* Randomized comparison of interferon-alpha with busulfan and hydroxyurea in chronic myelogenous leukemia. The German CML Study Group. *Blood* **84**, 4064-4077 (1994).
66. Kantarjian, H.M.*, et al.* Complete cytogenetic and molecular responses to interferon-alpha-based therapy for chronic myelogenous leukemia are associated with excellent long-term prognosis. *Cancer* **97**, 1033-1041 (2003).
67. Interferon alfa-2a as compared with conventional chemotherapy for the treatment of chronic myeloid leukemia. The Italian Cooperative Study Group on Chronic Myeloid Leukemia. *N Engl J Med* **330**, 820-825 (1994).
68. Allan, N.C., Richards, S.M. & Shepherd, P.C. UK Medical Research Council randomised, multicentre trial of interferon-alpha n1 for chronic myeloid leukaemia: improved survival irrespective of cytogenetic response. The UK Medical Research Council's Working Parties for Therapeutic Trials in Adult Leukaemia. *Lancet* **345**, 1392-1397 (1995).
69. Hehlmann, R.*, et al.* The German CML study, comparison of busulfan vs. hydroxyurea vs. interferon alpha and establishment of prognostic score 1. *Leuk Lymphoma* **11 Suppl 1**, 159-168 (1993).
70. Levitzki, A. & Gazit, A. Tyrosine kinase inhibition: an approach to drug development. *Science* **267**, 1782-1788 (1995).
71. Buchdunger, E.*, et al.* Selective inhibition of the platelet-derived growth factor signal transduction pathway by a protein-tyrosine kinase inhibitor of the 2-phenylaminopyrimidine class. *Proc Natl Acad Sci U S A* **92**, 2558-2562 (1995).
72. Zimmermann, J.*, et al.* Phenylamino-pyrimidine (PAP) -- derivatives: a new class of potent and highly selective PDGF-receptor autophosphorylation inhibitors. *Bioorganic & Medicinal Chemistry Letters* **6**, 1221-1226 (1996).
73. Zimmermann, J., Buchdunger, E., Mett, H., Meyer, T. & Lydon, N.B. Potent and selective inhibitors of the Abl-kinase: phenylamino-pyrimidine (PAP) derivatives. *Bioorganic & Medicinal Chemistry Letters* **7**, 187-192 (1997).
74. Druker, B.J.*, et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* **2**, 561-566 (1996).
75. Mann, J.C. Enzymes - Competitive Inhibition Diagram. *New World Encyclopedia* (2005).
76. Deininger, M., Buchdunger, E. & Druker, B.J. The development of imatinib as a therapeutic agent for chronic myeloid leukemia. *Blood* **105**, 2640-2653 (2005).
77. Dewar, A.L.*, et al.* Macrophage colony-stimulating factor receptor c-fms is a novel target of imatinib. *Blood* **105**, 3127-3132 (2005).
78. Heinrich, M.C.*, et al.* Inhibition of c-kit receptor tyrosine kinase activity by STI 571, a selective tyrosine kinase inhibitor. *Blood* **96**, 925-932 (2000).

79. Jabbour, E., Cortes, J.E., Giles, F.J., O'Brien, S. & Kantarjian, H.M. Current and emerging treatment options in chronic myeloid leukemia. *Cancer* **109**, 2171-2181 (2007).
80. Ohno, R. Treatment of chronic myeloid leukemia with imatinib mesylate. *Int J Clin Oncol* **11**, 176-183 (2006).
81. Druker, B.J.*, et al.* Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N Engl J Med* **344**, 1031-1037 (2001).
82. Peng, B.*, et al.* Pharmacokinetics and pharmacodynamics of imatinib in a phase I trial with chronic myeloid leukemia patients. *J Clin Oncol* **22**, 935-942 (2004).
83. Kantarjian, H.*, et al.* Hematologic and cytogenetic responses to imatinib mesylate in chronic myelogenous leukemia. *N Engl J Med* **346**, 645-652 (2002).
84. Sawyers, C.L.*, et al.* Imatinib induces hematologic and cytogenetic responses in patients with chronic myelogenous leukemia in myeloid blast crisis: results of a phase II study. *Blood* **99**, 3530-3539 (2002).
85. Talpaz, M.*, et al.* Imatinib induces durable hematologic and cytogenetic responses in patients with accelerated phase chronic myeloid leukemia: results of a phase 2 study. *Blood* **99**, 1928-1937 (2002).
86. O'Brien, S.G.*, et al.* Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *N Engl J Med* **348**, 994-1004 (2003).
87. Hughes, T.P.*, et al.* Frequency of major molecular responses to imatinib or interferon alfa plus cytarabine in newly diagnosed chronic myeloid leukemia. *N Engl J Med* **349**, 1423-1432 (2003).
88. Savage, D.G. & Antman, K.H. Imatinib mesylate--a new oral targeted therapy. *N Engl J Med* **346**, 683-693 (2002).
89. Druker, B.J. Translation of the Philadelphia chromosome into therapy for CML. *Blood* **112**, 4808-4817 (2008).
90. Deininger, M.*, et al.* International Randomized Study of Interferon Vs STI571 (IRIS) 8-Year Follow up: Sustained Survival and Low Risk for Progression or Events in Patients with Newly Diagnosed Chronic Myeloid Leukemia in Chronic Phase (CML-CP) Treated with Imatinib. *ASH Annual Meeting Abstracts* **114**, 1126- (2009).
91. Shah, N.P. Medical management of CML. *Hematology Am Soc Hematol Educ Program*, 371-375 (2007).
92. Druker, B.J.*, et al.* Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. *N Engl J Med* **355**, 2408-2417 (2006).
93. Marin, D.*, et al.* European LeukemiaNet criteria for failure or suboptimal response reliably identify patients with CML in early chronic phase treated with imatinib whose eventual outcome is poor. *Blood* **112**, 4437-4444 (2008).
94. Kantarjian, H.M.*, et al.* Dose escalation of imatinib mesylate can overcome resistance to standard-dose therapy in patients with chronic myelogenous leukemia. *Blood* **101**, 473-475 (2003).
95. Krause, D.S. & Van Etten, R.A. Tyrosine kinases as targets for cancer therapy. *N Engl J Med* **353**, 172-187 (2005).
96. White, D.L.*, et al.* OCT-1-mediated influx is a key determinant of the intracellular uptake of imatinib but not nilotinib (AMN107): reduced OCT-1 activity is the cause of low in vitro sensitivity to imatinib. *Blood* **108**, 697-704 (2006).
97. Crossman, L.C.*, et al.* hOCT 1 and resistance to imatinib. *Blood* **106**, 1133-1134; author reply 1134 (2005).
98. Mahon, F.X.*, et al.* MDR1 gene overexpression confers resistance to imatinib mesylate in leukemia cell line models. *Blood* **101**, 2368-2373 (2003).

99. Thomas, J., Wang, L., Clark, R.E. & Pirmohamed, M. Active transport of imatinib into and out of cells: implications for drug resistance. *Blood* **104**, 3739-3745 (2004).
100. Illmer, T.*, et al.* P-glycoprotein-mediated drug efflux is a resistance mechanism of chronic myelogenous leukemia cells to treatment with imatinib mesylate. *Leukemia* **18**, 401-408 (2004).
101. Mahon, F.X.*, et al.* Selection and characterization of BCR-ABL positive cell lines with differential sensitivity to the tyrosine kinase inhibitor STI571: diverse mechanisms of resistance. *Blood* **96**, 1070-1079 (2000).
102. Gambacorti-Passerini, C.*, et al.* Role of alpha1 acid glycoprotein in the in vivo resistance of human BCR-ABL(+) leukemic cells to the abl inhibitor STI571. *J Natl Cancer Inst* **92**, 1641-1650 (2000).
103. Quintas-Cardama, A., Kantarjian, H.M. & Cortes, J.E. Mechanisms of primary and secondary resistance to imatinib in chronic myeloid leukemia. *Cancer control : journal of the Moffitt Cancer Center* **16**, 122-131 (2009).
104. Martinelli, G.*, et al.* Monitoring minimal residual disease and controlling drug resistance in chronic myeloid leukaemia patients in treatment with imatinib as a guide to clinical management. *Hematol Oncol* **24**, 196-204 (2006).
105. Branford, S.*, et al.* High frequency of point mutations clustered within the adenosine triphosphate-binding region of BCR/ABL in patients with chronic myeloid leukemia or Ph-positive acute lymphoblastic leukemia who develop imatinib (STI571) resistance. *Blood* **99**, 3472-3475 (2002).
106. Branford, S.*, et al.* Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. *Blood* **102**, 276-283 (2003).
107. Ohno, R. & Nakamura, Y. Prediction of response to imatinib by cDNA microarray analysis. *Semin Hematol* **40**, 42-49 (2003).
108. Gorre, M.E.*, et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876-880 (2001).
109. Nagar, B.*, et al.* Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res* **62**, 4236-4243 (2002).
110. Hochhaus, A.*, et al.* Molecular and chromosomal mechanisms of resistance to imatinib (STI571) therapy. *Leukemia* **16**, 2190-2196 (2002).
111. Campbell, L.J.*, et al.* BCR/ABL amplification in chronic myelocytic leukemia blast crisis following imatinib mesylate administration. *Cancer Genet Cytogenet* **139**, 30-33 (2002).
112. Corbin, A.S.*, et al.* Human chronic myeloid leukemia stem cells are insensitive to imatinib despite inhibition of BCR-ABL activity. *The Journal of clinical investigation* **121**, 396-409 (2011).
113. Hamilton, A.*, et al.* Chronic myeloid leukemia stem cells are not dependent on Bcr-Abl kinase activity for their survival. *Blood* **119**, 1501-1510 (2012).
114. Kinstrie, R. & Copland, M. Targeting chronic myeloid leukemia stem cells. *Current hematologic malignancy reports* **8**, 14-21 (2013).
115. Chu, S.*, et al.* Persistence of leukemia stem cells in chronic myelogenous leukemia patients in prolonged remission with imatinib treatment. *Blood* **118**, 5565-5572 (2011).
116. Chomel, J.C.*, et al.* Leukemic stem cell persistence in chronic myeloid leukemia patients with sustained undetectable molecular residual disease. *Blood* **118**, 3657-3660 (2011).

117. Shah, N.P.*, et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **305**, 399-401 (2004).
118. Okimoto, R.A. & Van Etten, R.A. Navigating the road toward optimal initial therapy for chronic myeloid leukemia. *Current opinion in hematology* **18**, 89-97 (2011).
119. Lombardo, L.J.*, et al.* Discovery of N-(2-chloro-6-methyl- phenyl)-2-(6-(4-(2-hydroxyethyl)- piperazin-1-yl)-2-methylpyrimidin-4- ylamino)thiazole-5-carboxamide (BMS-354825), a dual Src/Abl kinase inhibitor with potent antitumor activity in preclinical assays. *J Med Chem* **47**, 6658-6661 (2004).
120. O'Hare, T.*, et al.* In vitro activity of Bcr-Abl inhibitors AMN107 and BMS-354825 against clinically relevant imatinib-resistant Abl kinase domain mutants. *Cancer Res* **65**, 4500-4505 (2005).
121. Doggrell, S.A. BMS-354825: a novel drug with potential for the treatment of imatinib-resistant chronic myeloid leukaemia. *Expert Opin Investig Drugs* **14**, 89-91 (2005).
122. Olivieri, A. & Manzione, L. Dasatinib: a new step in molecular target therapy. *Ann Oncol* **18 Suppl 6**, vi42-46 (2007).
123. Weisberg, E.*, et al.* Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer Cell* **7**, 129-141 (2005).
124. Manley, P.W.*, et al.* Urea derivatives of STI571 as inhibitors of Bcr-Abl and PDGFR kinases. *Bioorg Med Chem Lett* **14**, 5793-5797 (2004).
125. Weisberg, E., Manley, P.W., Cowan-Jacob, S.W., Hochhaus, A. & Griffin, J.D. Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. *Nature reviews. Cancer* **7**, 345-356 (2007).
126. Bradeen, H.A.*, et al.* Comparison of imatinib mesylate, dasatinib (BMS-354825), and nilotinib (AMN107) in an N-ethyl-N-nitrosourea (ENU)-based mutagenesis screen: high efficacy of drug combinations. *Blood* **108**, 2332-2338 (2006).
127. Cortes, J., Hochhaus, A., Hughes, T. & Kantarjian, H. Front-line and salvage therapies with tyrosine kinase inhibitors and other treatments in chronic myeloid leukemia. *J Clin Oncol* **29**, 524-531 (2011).
128. Sokal, J.E.*, et al.* Prognostic discrimination in "good-risk" chronic granulocytic leukemia. *Blood* **63**, 789-799 (1984).
129. Hasford, J.*, et al.* A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. Writing Committee for the Collaborative CML Prognostic Factors Project Group. *J Natl Cancer Inst* **90**, 850-858 (1998).
130. Cortes, J.*, et al.* Effects of age on prognosis with imatinib mesylate therapy for patients with Philadelphia chromosome-positive chronic myelogenous leukemia. *Cancer* **98**, 1105-1113 (2003).
131. Gugliotta, G.*, et al.* Frontline imatinib treatment of chronic myeloid leukemia: no impact of age on outcome, a survey by the GIMEMA CML Working Party. *Blood* **117**, 5591-5599 (2011).
132. Hasford, J.*, et al.* Predicting complete cytogenetic response and subsequent progression-free survival in 2060 patients with CML on imatinib treatment: the EUTOS score. *Blood* **118**, 686-692 (2011).
133. Baccarani, M.*, et al.* Chronic myeloid leukemia: an update of concepts and management recommendations of European LeukemiaNet. *J Clin Oncol* **27**, 6041-6051 (2009).
134. White, D.L. & Hughes, T.P. Predicting the response of CML patients to tyrosine kinase inhibitor therapy. *Current hematologic malignancy reports* **4**, 59-65 (2009).
135. White, D.*, et al.* In vitro sensitivity to imatinib-induced inhibition of ABL kinase activity is predictive of molecular response in patients with de novo CML. *Blood* **106**, 2520-2526 (2005).

136. Nichols, G.L.*, et al.* Identification of CRKL as the constitutively phosphorylated 39-kD tyrosine phosphoprotein in chronic myelogenous leukemia cells. *Blood* **84**, 2912-2918 (1994).
137. ten Hoeve, J., Arlinghaus, R.B., Guo, J.Q., Heisterkamp, N. & Groffen, J. Tyrosine phosphorylation of CRKL in Philadelphia+ leukemia. *Blood* **84**, 1731-1736 (1994).
138. Oda, T.*, et al.* Crkl is the major tyrosine-phosphorylated protein in neutrophils from patients with chronic myelogenous leukemia. *J Biol Chem* **269**, 22925-22928 (1994).
139. Lucas, C.M.*, et al.* BCR-ABL1 tyrosine kinase activity at diagnosis, as determined via the pCrkL/CrkL ratio, is predictive of clinical outcome in chronic myeloid leukaemia. *Br J Haematol* **149**, 458-460 (2010).
140. Khorashad, J.S.*, et al.* The level of BCR-ABL1 kinase activity before treatment does not identify chronic myeloid leukemia patients who fail to achieve a complete cytogenetic response on imatinib. *Haematologica* **94**, 861-864 (2009).
141. Simara, P.*, et al.* BCR-ABL activity measured by 50% inhibitory concentration for imatinib, p-CrkL/CrkL ratio or p-CrkL ratio in CD34+ cells of patients with chronic myeloid leukemia does not predict treatment response. *Leuk Lymphoma* **53**, 1627-1629 (2012).
142. White, D.L.*, et al.* Functional activity of the OCT-1 protein is predictive of long-term outcome in patients with chronic-phase chronic myeloid leukemia treated with imatinib. *J Clin Oncol* **28**, 2761-2767 (2010).
143. Koepsell, H., Schmitt, B.M. & Gorboulev, V. Organic cation transporters. *Reviews of physiology, biochemistry and pharmacology* **150**, 36-90 (2003).
144. Jonker, J.W. & Schinkel, A.H. Pharmacological and physiological functions of the polyspecific organic cation transporters: OCT1, 2, and 3 (SLC22A1-3). *J Pharmacol Exp Ther* **308**, 2-9 (2004).
145. Gorboulev, V.*, et al.* Cloning and characterization of two human polyspecific organic cation transporters. *DNA and cell biology* **16**, 871-881 (1997).
146. Zhang, L.*, et al.* Cloning and functional expression of a human liver organic cation transporter. *Molecular pharmacology* **51**, 913-921 (1997).
147. Grundemann, D., Gorboulev, V., Gambaryan, S., Veyhl, M. & Koepsell, H. Drug excretion mediated by a new prototype of polyspecific transporter. *Nature* **372**, 549-552 (1994).
148. Koepsell, H. & Endou, H. The SLC22 drug transporter family. *Pflugers Archiv : European journal of physiology* **447**, 666-676 (2004).
149. Wang, L.*, et al.* Expression of the uptake drug transporter hOCT1 is an important clinical determinant of the response to imatinib in chronic myeloid leukemia. *Clin Pharmacol Ther* **83**, 258-264 (2008).
150. Marin, D.*, et al.* Adherence is the critical factor for achieving molecular responses in patients with chronic myeloid leukemia who achieve complete cytogenetic responses on imatinib. *J Clin Oncol* **28**, 2381-2388 (2010).
151. White, D.L.*, et al.* Most CML patients who have a suboptimal response to imatinib have low OCT-1 activity: higher doses of imatinib may overcome the negative impact of low OCT-1 activity. *Blood* **110**, 4064-4072 (2007).
152. Zhang, W.W.*, et al.* Predictors of primary imatinib resistance in chronic myelogenous leukemia are distinct from those in secondary imatinib resistance. *J Clin Oncol* **27**, 3642-3649 (2009).
153. Bazeos, A.*, et al.* hOCT1 transcript levels and single nucleotide polymorphisms as predictive factors for response to imatinib in chronic myeloid leukemia. *Leukemia* **24**, 1243-1245 (2010).

154. Razga, F.*, et al.* The predictive value of human organic cation transporter 1 and ABCB1 expression levels in different cell populations of patients with de novo chronic myelogenous leukemia. *Int J Hematol* **94**, 303-306 (2011).
155. Racil, Z.*, et al.* The assessment of human organic cation transporter 1 (hOCT1) mRNA expression in patients with chronic myelogenous leukemia is affected by the proportion of different cells types in the analyzed cell population. *Am J Hematol* **85**, 525-528 (2010).
156. Jiang, X.*, et al.* Chronic myeloid leukemia stem cells possess multiple unique features of resistance to BCR-ABL targeted therapies. *Leukemia* **21**, 926-935 (2007).
157. Clark, R.E., Davies, A., Pirmohamed, M. & Giannoudis, A. Pharmacologic markers and predictors of responses to imatinib therapy in patients with chronic myeloid leukemia. *Leuk Lymphoma* **49**, 639-642 (2008).
158. Giannoudis, A.*, et al.* The hOCT1 SNPs M420del and M408V alter imatinib uptake and M420del modifies clinical outcome in imatinib-treated chronic myeloid leukemia. *Blood* **121**, 628-637 (2013).
159. Kerb, R.*, et al.* Identification of genetic variations of the human organic cation transporter hOCT1 and their functional consequences. *Pharmacogenetics* **12**, 591-595 (2002).
160. Shu, Y.*, et al.* Evolutionary conservation predicts function of variants of the human organic cation transporter, OCT1. *Proc Natl Acad Sci U S A* **100**, 5902-5907 (2003).
161. Sakata, T.*, et al.* Novel single nucleotide polymorphisms of organic cation transporter 1 (SLC22A1) affecting transport functions. *Biochemical and biophysical research communications* **313**, 789-793 (2004).
162. Itoda, M.*, et al.* Seven novel single nucleotide polymorphisms in the human SLC22A1 gene encoding organic cation transporter 1 (OCT1). *Drug metabolism and pharmacokinetics* **19**, 308-312 (2004).
163. Kim, D.H.*, et al.* Clinical relevance of a pharmacogenetic approach using multiple candidate genes to predict response and resistance to imatinib therapy in chronic myeloid leukemia. *Clin Cancer Res* **15**, 4750-4758 (2009).
164. Maffioli, M.*, et al.* Correlation between genetic polymorphisms of the hOCT1 and MDR1 genes and the response to imatinib in patients newly diagnosed with chronic-phase chronic myeloid leukemia. *Leuk Res* **35**, 1014-1019 (2011).
165. Tzvetkov, M.V.*, et al.* Does the haplotype Met408-Del420, which was apparently predictive for imatinib efficacy, really exist and how strongly may it affect OCT1 activity? *Blood* **123**, 1427-1429 (2014).
166. White, D.L., Saunders, V.A., Dang, P., Engler, J. & Hughes, T.P. OCT-1 activity measurement provides a superior imatinib response predictor than screening for single-nucleotide polymorphisms of OCT-1. *Leukemia* **24**, 1962-1965 (2010).
167. Hayer-Zillgen, M., Bruss, M. & Bonisch, H. Expression and pharmacological profile of the human organic cation transporters hOCT1, hOCT2 and hOCT3. *Br J Pharmacol* **136**, 829-836 (2002).
168. Hu, S.*, et al.* Interaction of imatinib with human organic ion carriers. *Clin Cancer Res* **14**, 3141-3148 (2008).
169. Mitchell, B. & Deininger, M. Techniques for risk stratification of newly diagnosed patients with chronic myeloid leukemia. *Leuk Lymphoma* **52 Suppl 1**, 4-11 (2011).
170. Burger, H.*, et al.* Chronic imatinib mesylate exposure leads to reduced intracellular drug accumulation by induction of the ABCG2 (BCRP) and ABCB1 (MDR1) drug transport pumps. *Cancer biology & therapy* **4**, 747-752 (2005).
171. Ozvegy-Laczka, C.*, et al.* High-affinity interaction of tyrosine kinase inhibitors with the ABCG2 multidrug transporter. *Molecular pharmacology* **65**, 1485-1495 (2004).

172. Dulucq, S.*, et al.* Multidrug resistance gene (MDR1) polymorphisms are associated with major molecular responses to standard-dose imatinib in chronic myeloid leukemia. *Blood* **112**, 2024-2027 (2008).
173. Deenik, W.*, et al.* Polymorphisms in the multidrug resistance gene MDR1 (ABCB1) predict for molecular resistance in patients with newly diagnosed chronic myeloid leukemia receiving high-dose imatinib. *Blood* **116**, 6144-6145; author reply 6145-6146 (2010).
174. Dulucq, S., Preudhomme, C., Guilhot, F. & Mahon, F.X. Is there really a relationship between Multidrug Resistance Gene (MDR1) polymorphisms and major molecular response to imatinib in chronic myeloid leukemia? *Blood*, 6145-6146 (2010).
175. Seong, S.J.*, et al.* Influence of enzyme and transporter polymorphisms on trough imatinib concentration and clinical response in chronic myeloid leukemia patients. *Ann Oncol* **24**, 756-760 (2013).
176. Ni, L.N.*, et al.* Multidrug resistance gene (MDR1) polymorphisms correlate with imatinib response in chronic myeloid leukemia. *Medical oncology* **28**, 265-269 (2011).
177. Angelini, S.*, et al.* Association between imatinib transporters and metabolizing enzymes genotype and response in newly diagnosed chronic myeloid leukemia patients receiving imatinib therapy. *Haematologica* **98**, 193-200 (2013).
178. Oehler, V.G.*, et al.* The derivation of diagnostic markers of chronic myeloid leukemia progression from microarray data. *Blood* **114**, 3292-3298 (2009).
179. McWeeney, S.K.*, et al.* A gene expression signature of CD34+ cells to predict major cytogenetic response in chronic-phase chronic myeloid leukemia patients treated with imatinib. *Blood* **115**, 315-325 (2010).
180. Radich, J.P.*, et al.* Gene expression changes associated with progression and response in chronic myeloid leukemia. *Proc Natl Acad Sci U S A* **103**, 2794-2799 (2006).
181. Zheng, C.*, et al.* Gene expression profiling of CD34+ cells identifies a molecular signature of chronic myeloid leukemia blast crisis. *Leukemia* **20**, 1028-1034 (2006).
182. Picard, S.*, et al.* Trough imatinib plasma levels are associated with both cytogenetic and molecular responses to standard-dose imatinib in chronic myeloid leukemia. *Blood* **109**, 3496-3499 (2007).
183. Larson, R.A.*, et al.* Imatinib pharmacokinetics and its correlation with response and safety in chronic-phase chronic myeloid leukemia: a subanalysis of the IRIS study. *Blood* **111**, 4022-4028 (2008).
184. Forrest, D.L.*, et al.* Cytogenetic and molecular responses to standard-dose imatinib in chronic myeloid leukemia are correlated with Sokal risk scores and duration of therapy but not trough imatinib plasma levels. *Leuk Res* **33**, 271-275 (2009).
185. Sohn, S.K.*, et al.* Trough plasma imatinib levels are correlated with optimal cytogenetic responses at 6 months after treatment with standard dose of imatinib in newly diagnosed chronic myeloid leukemia. *Leuk Lymphoma* **52**, 1024-1029 (2011).
186. Guilhot, F.*, et al.* Plasma exposure of imatinib and its correlation with clinical response in the Tyrosine Kinase Inhibitor Optimization and Selectivity Trial. *Haematologica* **97**, 731-738 (2012).
187. Sabate, E. (ed.) *Adherence to Long-Term Therapies: Evidence for Action*, (World Heath Organization, Geneva, Swizerland, 2003).
188. Branford, S.*, et al.* Imatinib produces significantly superior molecular responses compared to interferon alfa plus cytarabine in patients with newly diagnosed chronic myeloid leukemia in chronic phase. *Leukemia* **17**, 2401-2409 (2003).
189. Hughes, T. & Branford, S. Molecular monitoring of chronic myeloid leukemia. *Semin Hematol* **40**, 62-68 (2003).

190. Quintas-Cardama, A*., et al.* Delayed achievement of cytogenetic and molecular response is associated with increased risk of progression among patients with chronic myeloid leukemia in early chronic phase receiving high-dose or standard-dose imatinib therapy. *Blood* **113**, 6315-6321 (2009).
191. Marin, D*., et al.* Assessment of BCR-ABL1 transcript levels at 3 months is the only requirement for predicting outcome for patients with chronic myeloid leukemia treated with tyrosine kinase inhibitors. *J Clin Oncol* **30**, 232-238 (2012).
192. Hanfstein, B*., et al.* Early molecular and cytogenetic response is predictive for long-term progression-free and overall survival in chronic myeloid leukemia (CML). *Leukemia* **26**, 2096-2102 (2012).
193. Kim, D.D., Lee, H., Kamel-Reid, S. & Lipton, J.H. BCR-ABL1 transcript at 3 months predicts long-term outcomes following second generation tyrosine kinase inhibitor therapy in the patients with chronic myeloid leukaemia in chronic phase who failed Imatinib. *Br J Haematol* **160**, 630-639 (2013).
194. Neelakantan, P*., et al.* Combining BCR-ABL1 transcript levels at 3 and 6 months in chronic myeloid leukemia: implications for early intervention strategies. *Blood* **121**, 2739-2742 (2013).
195. Hiwase, D.K*., et al.* Dasatinib cellular uptake and efflux in chronic myeloid leukemia cells: therapeutic implications. *Clin Cancer Res* **14**, 3881-3888 (2008).
196. Branford, S., Hughes, T.P. & Rudzki, Z. Monitoring chronic myeloid leukaemia therapy by real-time quantitative PCR in blood is a reliable alternative to bone marrow cytogenetics. *Br J Haematol* **107**, 587-599 (1999).
197. Xia, M., Sherlock, J, Hegerich, P, You, X, Lee, K, Walworth, C and Spier, E. DataAssist - Data Analysis Software for TaqMan Real-Time PCR Data. *Proceedings of the International MultiConference of Engineers and Computer Scientists* **Vol I**(2010).
198. Lin, L.I*., et al.* A novel fluorescence-based multiplex PCR assay for rapid simultaneous detection of CEBPA mutations and NPM mutations in patients with acute myeloid leukemias. *Leukemia* **20**, 1899-1903 (2006).
199. Sandoval, J*., et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics : official journal of the DNA Methylation Society* **6**, 692-702 (2011).
200. Team, R.C. R: A Language and Environment for Statistical Computing. (R Foundation for Statistical Computing, Vienna, Austria, 2013).
201. Hansen, K.D. & Aryee, M. minfi: Analyze Illumina's 450k methylation arrays. (R package version 1.6.0, 2013).
202. Gentleman, R.C*., et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* **5**, R80 (2004).
203. Du, P*., et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics* **11**, 587 (2010).
204. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome biology* **13**, R44 (2012).
205. Smyth, G.K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3**, Article3 (2004).
206. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B Method* **57**, 289-300 (1995).
207. Saeed, A.I*., et al.* TM4: a free, open-source system for microarray data management and analysis. *BioTechniques* **34**, 374-378 (2003).
208. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).

209. Yeung, K.Y., Bumgarner, R.E. & Raftery, A.E. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* **21**, 2394-2402 (2005).
210. Raftery, A.E. Bayesian model selection in social research (with discussion). *Sociological Methodology* **25**, 111-196 (1995).
211. Dudoit, S. & Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology* **3**, RESEARCH0036 (2002).
212. Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon Weather Rev* **78**, 1-3 (1950).
213. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A* **99**, 6567-6572 (2002).
214. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. PAMR: PAM: Prediction Analyis for Microarrays. (R package version 1.54.1., 2013).
215. Hastie, T., Tibshirani, R. & Friedman, J. *Th Elements of Statistical Learing: Data Mining, Inference and Prediction*, (Springer, New York, 2001).
216. Subramanian, A*., et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
217. Liberzon, A*., et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739-1740 (2011).
218. Chambers, S.M*., et al.* Hematopoietic fingerprints: an expression database of stem cells and their progeny. *Cell Stem Cell* **1**, 578-591 (2007).
219. Xie, F., Sun, G., Stiller, J.W. & Zhang, B. Genome-wide functional analysis of the cotton transcriptome by creating an integrated EST database. *PloS one* **6**, e26980 (2011).
220. Vandesompele, J*., et al.* Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome biology* **3**, RESEARCH0034 (2002).
221. Andersen, C.L., Jensen, J.L. & Orntoft, T.F. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res* **64**, 5245-5250 (2004).
222. Pfaffl, M.W., Tichopad, A., Prgomet, C. & Neuvians, T.P. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper--Excel-based tool using pair-wise correlations. *Biotechnology letters* **26**, 509-515 (2004).
223. Silver, N., Best, S., Jiang, J. & Thein, S.L. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC molecular biology* **7**, 33 (2006).
224. Chao, W.S., Dogramaci, M., Foley, M.E., Horvath, D.P. & Anderson, J.V. Selection and validation of endogenous reference genes for qRT-PCR analysis in leafy spurge (Euphorbia esula). *PloS one* **7**, e42839 (2012).
225. Dvinge, H. & Bertone, P. HTqPCR: high-throughput analysis and visualization of quantitative real-time PCR data in R. *Bioinformatics* **25**, 3325-3326 (2009).
226. Slawski, M., Boulesteix, A.-L. & Bernau, C. CMA: Synthesis of microarray-based classification. (R package version 1.18.0, 2009).
227. Slawski, M., Daumer, M. & Boulesteix, A.L. CMA: a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC bioinformatics* **9**, 439 (2008).

228. Essaghir, A.*, et al.* Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Res* **38**, e120 (2010).
229. Zambelli, F., Pesole, G. & Pavesi, G. Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes. *Nucleic Acids Res* **37**, W247-252 (2009).
230. Lachmann, A.*, et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438-2444 (2010).
231. Bryne, J.C.*, et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **36**, D102-106 (2008).
232. Matys, V.*, et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108-110 (2006).
233. Barfield, R.T., Kilaru, V., Smith, A.K. & Conneely, K.N. CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics* **28**, 1280-1281 (2012).
234. Laffaire, J.*, et al.* Methylation profiling identifies 2 groups of gliomas according to their tumorigenesis. *Neuro-oncology* **13**, 84-98 (2011).
235. Sun, Z.*, et al.* Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC medical genomics* **4**, 84 (2011).
236. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127 (2007).
237. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012).
238. Shi, W., Oshlack, A. & Smyth, G.K. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res* **38**, e204 (2010).
239. Rhodes, D.R.*, et al.* Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* **9**, 166-180 (2007).
240. Rhodes, D.R.*, et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1-6 (2004).
241. Cortes, J. & Kantarjian, H. Advanced-phase chronic myeloid leukemia. *Semin Hematol* **40**, 79-86 (2003).
242. White, D.*, et al.* Measurement of in vivo BCR-ABL kinase inhibition to monitor imatinib-induced target blockade and predict response in chronic myeloid leukemia. *J Clin Oncol* **25**, 4445-4451 (2007).
243. Engler, J.R.*, et al.* OCT-1 function varies with cell lineage but is not influenced by BCR-ABL. *Haematologica* **96**, 213-220 (2011).
244. Craig, F.E. & Foon, K.A. Flow cytometric immunophenotyping for hematologic neoplasms. *Blood* **111**, 3941-3967 (2008).
245. Isoda, A.*, et al.* The naive T-lymphocyte compartment is well preserved in patients with chronic myelogenous leukaemia in chronic phase. *Br J Haematol* **119**, 949-955 (2002).
246. Normann, A.P., Egeland, T., Madshus, I.H., Heim, S. & Tjonnfjord, G.E. CD7 expression by CD34+ cells in CML patients, of prognostic significance? *Eur J Haematol* **71**, 266-275 (2003).
247. Hayashi, Y.*, et al.* Different immunoprofiles in patients with chronic myeloid leukemia treated with imatinib, nilotinib or dasatinib. *Leuk Lymphoma* **53**, 1084-1089 (2012).
248. Bourgne, C.*, et al.* Measurement of imatinib uptake by flow cytometry. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **81**, 996-1004 (2012).

249. Hevessy, Z*., et al.* Laboratory evaluation of a flow cytometric BCR-ABL immunobead assay. *Clinical chemistry and laboratory medicine : CCLM / FESCC* **50**, 689-692 (2012).
250. Weerkamp, F*., et al.* Flow cytometric immunobead assay for the detection of BCR-ABL fusion proteins in leukemia patients. *Leukemia* **23**, 1106-1117 (2009).
251. Rohon, P., Porkka, K. & Mustjoki, S. Immunoprofiling of patients with chronic myeloid leukemia at diagnosis and during tyrosine kinase inhibitor therapy. *Eur J Haematol* **85**, 387-398 (2010).
252. Janssen, J.J*., et al.* Residual normal stem cells can be detected in newly diagnosed chronic myeloid leukemia patients by a new flow cytometric approach and predict for optimal response to imatinib. *Leukemia* **26**, 977-984 (2012).
253. Gratwohl, A. & Heim, D. Current role of stem cell transplantation in chronic myeloid leukaemia. *Best Pract Res Clin Haematol* **22**, 431-443 (2009).
254. Gao, L*., et al.* Selective elimination of leukemic CD34(+) progenitor cells by cytotoxic T lymphocytes specific for WT1. *Blood* **95**, 2198-2203 (2000).
255. Quintarelli, C*., et al.* Cytotoxic T lymphocytes directed to the preferentially expressed antigen of melanoma (PRAME) target chronic myeloid leukemia. *Blood* **112**, 1876-1885 (2008).
256. Yong, A.S*., et al.* Improved outcome following allogeneic stem cell transplantation in chronic myeloid leukemia is associated with higher expression of BMI-1 and immune responses to BMI-1 protein. *Leukemia* **25**, 629-637 (2011).
257. Yong, A.S*., et al.* High PR3 or ELA2 expression by CD34+ cells in advanced-phase chronic myeloid leukemia is associated with improved outcome following allogeneic stem cell transplantation and may improve PR1 peptide-driven graft-versus-leukemia effects. *Blood* **110**, 770-775 (2007).
258. Burchert, A*., et al.* Sustained molecular response with interferon alfa maintenance after induction therapy with imatinib plus interferon alfa in patients with chronic myeloid leukemia. *J Clin Oncol* **28**, 1429-1435 (2010).
259. Rezvani, K*., et al.* Leukemia-associated antigen-specific T-cell responses following combined PR1 and WT1 peptide vaccination in patients with myeloid malignancies. *Blood* **111**, 236-242 (2008).
260. Mahon, F.X*., et al.* Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre Stop Imatinib (STIM) trial. *The lancet oncology* **11**, 1029-1035 (2010).
261. Ross, D.M*., et al.* Patients with chronic myeloid leukemia who maintain a complete molecular response after stopping imatinib treatment have evidence of persistent leukemia by DNA PCR. *Leukemia* **24**, 1719-1724 (2010).
262. Marin, D*., et al.* KIR2DS1 genotype predicts for complete cytogenetic response and survival in newly diagnosed chronic myeloid leukemia patients treated with imatinib. *Leukemia* **26**, 296-302 (2012).
263. Kreutzman, A*., et al.* Killer-cell immunoglobulin-like receptor gene profile predicts good molecular response to dasatinib therapy in chronic myeloid leukemia. *Exp Hematol* **40**, 906-913 e901 (2012).
264. Rojas, J.M*., et al.* Naturally occurring CD4+ CD25+ FOXP3+ T-regulatory cells are increased in chronic myeloid leukemia patients not in complete cytogenetic remission and can be immunosuppressive. *Exp Hematol* **38**, 1209-1218 (2010).
265. Trowbridge, I.S., Ostergaard, H.L. & Johnson, P. CD45: a leukocyte-specific member of the protein tyrosine phosphatase family. *Biochim Biophys Acta* **1095**, 46-56 (1991).
266. Kerr, M.A. & Stocks, S.C. The role of CD15-(Le(X))-related carbohydrates in neutrophil adhesion. *The Histochemical journal* **24**, 811-826 (1992).

267. Skubitz, K.M., Campbell, K.D. & Skubitz, A.P. CD66a, CD66b, CD66c, and CD66d each independently stimulate neutrophils. *Journal of leukocyte biology* **60**, 106-117 (1996).
268. Satterthwaite, A.B., Burn, T.C., Le Beau, M.M. & Tenen, D.G. Structure of the gene encoding CD34, a human hematopoietic stem cell antigen. *Genomics* **12**, 788-794 (1992).
269. Terstappen, L.W., Huang, S., Safford, M., Lansdorp, P.M. & Loken, M.R. Sequential generations of hematopoietic colonies derived from single nonlineage-committed CD34+CD38- progenitor cells. *Blood* **77**, 1218-1227 (1991).
270. Malavasi, F.*, et al.* Human CD38: a glycoprotein in search of a function. *Immunology today* **15**, 95-97 (1994).
271. Kronenberg, M., Siu, G., Hood, L.E. & Shastri, N. The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition. *Annual review of immunology* **4**, 529-591 (1986).
272. Griffin, J.D., Ritz, J., Nadler, L.M. & Schlossman, S.F. Expression of myeloid differentiation antigens on normal and malignant myeloid cells. *The Journal of clinical investigation* **68**, 932-941 (1981).
273. Ravetch, J.V. & Perussia, B. Alternative membrane forms of Fc gamma RIII(CD16) on human natural killer cells and neutrophils. Cell type-specific expression of two genes that differ in single nucleotide substitutions. *J Exp Med* **170**, 481-497 (1989).
274. Cragg, M.S., Walshe, C.A., Ivanov, A.O. & Glennie, M.J. The biology of CD20 and its potential as a target for mAb therapy. *Current directions in autoimmunity* **8**, 140-174 (2005).
275. Ka Yee Yeung, U.o.W., Seattle, WA, with contributions from Adrian Raftery and Ian Painter. iterativeBMA: The Iterative Bayesian Model Averaging (BMA) algorithm. (R package version 1.18.0, 2009).
276. Roederer, M. Spectral compensation for flow cytometry: visualization artifacts, limitations, and caveats. *Cytometry* **45**, 194-205 (2001).
277. Wu, Q.L., Zierold, C. & Ranheim, E.A. Dysregulation of Frizzled 6 is a critical component of B-cell leukemogenesis in a mouse model of chronic lymphocytic leukemia. *Blood* **113**, 3031-3039 (2009).
278. Zhang, B.*, et al.* Microenvironmental protection of CML stem and progenitor cells from tyrosine kinase inhibitors through N-cadherin and Wnt-beta-catenin signaling. *Blood* **121**, 1824-1838 (2013).
279. Heidel, F.H.*, et al.* Genetic and pharmacologic inhibition of beta-catenin targets imatinib-resistant leukemia stem cells in CML. *Cell Stem Cell* **10**, 412-424 (2012).
280. Gregory, M.A.*, et al.* Wnt/Ca2+/NFAT signaling maintains survival of Ph+ leukemia cells upon inhibition of Bcr-Abl. *Cancer Cell* **18**, 74-87 (2010).
281. Lyons, J.P.*, et al.* Wnt-4 activates the canonical beta-catenin-mediated Wnt pathway and binds Frizzled-6 CRD: functional implications of Wnt/beta-catenin activity in kidney epithelial cells. *Exp Cell Res* **298**, 369-387 (2004).
282. Golan, T., Yaniv, A., Bafico, A., Liu, G. & Gazit, A. The human Frizzled 6 (HFz6) acts as a negative regulator of the canonical Wnt. beta-catenin signaling cascade. *J Biol Chem* **279**, 14879-14888 (2004).
283. de Visser, K.E., Eichten, A. & Coussens, L.M. Paradoxical roles of the immune system during cancer development. *Nature reviews. Cancer* **6**, 24-37 (2006).
284. Baccarani, M.*, et al.* European LeukemiaNet recommendations for the management of chronic myeloid leukemia: 2013. *Blood* **122**, 872-884 (2013).
285. Macgregor, P.F. & Squire, J.A. Application of microarrays to the analysis of gene expression in cancer. *Clin Chem* **48**, 1170-1177 (2002).

286. Microarray Centre - Microarray analysis overview. Vol. 2010 (Microarray Centre of the Centre de Recherche Public de la Sante, 2008).
287. Kronenwett, R.*, et al.* Distinct molecular phenotype of malignant CD34(+) hematopoietic stem and progenitor cells in chronic myelogenous leukemia. *Oncogene* **24**, 5313-5324 (2005).
288. Nowicki, M.O.*, et al.* Chronic myelogenous leukemia molecular signature. *Oncogene* **22**, 3952-3963 (2003).
289. Ohmine, K.*, et al.* Characterization of stage progression in chronic myeloid leukemia by DNA microarray with purified hematopoietic stem cells. *Oncogene* **20**, 8249-8257 (2001).
290. Crossman, L.C.*, et al.* In chronic myeloid leukemia white cells from cytogenetic responders and non-responders to imatinib have very similar gene expression signatures. *Haematologica* **90**, 459-464 (2005).
291. Hagberg, A.*, et al.* Gene expression analysis identifies a genetic signature potentially associated with response to alpha-IFN in chronic phase CML patients. *Leuk Res* **31**, 931-938 (2007).
292. Frank, O.*, et al.* Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients. *Leukemia* **20**, 1400-1407 (2006).
293. McLean, L.A., Gathmann, I., Capdeville, R., Polymeropoulos, M.H. & Dressman, M. Pharmacogenomic analysis of cytogenetic response in chronic myeloid leukemia patients treated with imatinib. *Clin Cancer Res* **10**, 155-165 (2004).
294. Villuendas, R.*, et al.* Identification of genes involved in imatinib resistance in CML: a gene-expression profiling approach. *Leukemia* **20**, 1047-1054 (2006).
295. Neumann, F.*, et al.* Gene expression profiling of Philadelphia chromosome (Ph)-negative CD34+ hematopoietic stem and progenitor cells of patients with Ph-positive CML in major molecular remission during therapy with imatinib. *Leukemia* **19**, 458-460 (2005).
296. Yong, A.S. & Melo, J.V. The impact of gene profiling in chronic myeloid leukaemia. *Best Pract Res Clin Haematol* **22**, 181-190 (2009).
297. Mootha, V.K.*, et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* **34**, 267-273 (2003).
298. Clark, N.R. & Ma'ayan, A. Introduction to statistical methods for analyzing large data sets: gene-set enrichment analysis. *Science signaling* **4**, tr4 (2011).
299. Bustin, S.A. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *Journal of molecular endocrinology* **29**, 23-39 (2002).
300. Lord, K.A., Abdollahi, A., Hoffman-Liebermann, B. & Liebermann, D.A. Proto-oncogenes of the fos/jun family of transcription factors are positive regulators of myeloid differentiation. *Mol Cell Biol* **13**, 841-851 (1993).
301. Hoshino, K.*, et al.* Downregulation of JUNB mRNA expression in advanced phase chronic myelogenous leukemia. *Leuk Res* **33**, 1361-1366 (2009).
302. Strathdee, G., Ferguson, S., Sim, A. & Brown, R. DNA methylation does not regulate JUNB expression in CML: comment on "Downregulation of JUNB mRNA expression in advanced phase chronic myelogenous leukemia" by Hoshino et al. [Leuk. Res. 33 (2009) 1361-1366]. *Leuk Res* **34**, 685-686 (2010).
303. Hacohen, N., Kramer, S., Sutherland, D., Hiromi, Y. & Krasnow, M.A. sprouty encodes a novel antagonist of FGF signaling that patterns apical branching of the Drosophila airways. *Cell* **92**, 253-263 (1998).

304. Frank, M.J*., et al.* Expression of sprouty2 inhibits B-cell proliferation and is epigenetically silenced in mouse and human B-cell lymphomas. *Blood* **113**, 2478-2487 (2009).
305. Lee, C.C*., et al.* Overexpression of sprouty 2 inhibits HGF/SF-mediated cell growth, invasion, migration, and cytokinesis. *Oncogene* **23**, 5193-5202 (2004).
306. Patel, R*., et al.* Sprouty2, PTEN, and PP2A interact to regulate prostate cancer progression. *The Journal of clinical investigation* **123**, 1157-1175 (2013).
307. Neviani, P*., et al.* The tumor suppressor PP2A is functionally inactivated in blast crisis CML through the inhibitory activity of the BCR/ABL-regulated SET protein. *Cancer Cell* **8**, 355-368 (2005).
308. Perrotti, D. & Neviani, P. ReSETting PP2A tumour suppressor activity in blast crisis and imatinib-resistant chronic myelogenous leukaemia. *British journal of cancer* **95**, 775-781 (2006).
309. Costanzo, R.V*., et al.* Anp32e/Cpd1 regulates protein phosphatase 2A activity at synapses during synaptogenesis. *The European journal of neuroscience* **23**, 309-324 (2006).
310. Krivtsov, A.V. & Armstrong, S.A. MLL translocations, histone modifications and leukaemia stem-cell development. *Nature reviews. Cancer* **7**, 823-833 (2007).
311. Radziwill, G., Erdmann, R.A., Margelisch, U. & Moelling, K. The Bcr kinase downregulates Ras signaling by phosphorylating AF-6 and binding to its PDZ domain. *Mol Cell Biol* **23**, 4663-4672 (2003).
312. Puil, L*., et al.* Bcr-Abl oncoproteins bind directly to activators of the Ras signalling pathway. *EMBO J* **13**, 764-773 (1994).
313. Oliveros, J.C. VENNY. An interactive tool for comparing lists with Venn Diagrams. (http://bioinfogp.cnb.csic.es/tools/venny/index.html, 2007).
314. Tipping, A.J., Deininger, M.W., Goldman, J.M. & Melo, J.V. Comparative gene expression profile of chronic myeloid leukemia cells innately resistant to imatinib mesylate. *Exp Hematol* **31**, 1073-1080 (2003).
315. Sawyers, C.L. The bcr-abl gene in chronic myelogenous leukaemia. *Cancer surveys* **15**, 37-51 (1992).
316. Lugo, T.G. & Witte, O.N. The BCR-ABL oncogene transforms Rat-1 cells and cooperates with v-myc. *Mol Cell Biol* **9**, 1263-1270 (1989).
317. Stewart, M.J*., et al.* Role for E2F1 in p210 BCR-ABL downstream regulation of c-myc transcription initiation. Studies in murine myeloid cells. *Leukemia* **9**, 1499-1507 (1995).
318. Gomez-Casares, M.T*., et al.* C-myc expression in cell lines derived from chronic myeloid leukemia. *Haematologica* **89**, 241-243 (2004).
319. Notari, M*., et al.* A MAPK/HNRPK pathway controls BCR/ABL oncogenic potential by regulating MYC mRNA translation. *Blood* **107**, 2507-2516 (2006).
320. Mohty, M., Yong, A.S., Szydlo, R.M., Apperley, J.F. & Melo, J.V. The polycomb group BMI1 gene is a molecular marker for predicting prognosis of chronic myeloid leukemia. *Blood* **110**, 380-383 (2007).
321. Dalton, W.T., Jr*., et al.* HL-60 cell line was derived from a patient with FAB-M2 and not FAB-M3. *Blood* **71**, 242-247 (1988).
322. Rattray, A.M. & Muller, B. The control of histone gene expression. *Biochemical Society transactions* **40**, 880-885 (2012).
323. Marzluff, W.F., Gongidi, P., Woods, K.R., Jin, J. & Maltais, L.J. The human and mouse replication-dependent histone genes. *Genomics* **80**, 487-498 (2002).
324. Nicholson, P. & Muller, B. Post-transcriptional control of animal histone gene expression--not so different after all. *Molecular bioSystems* **4**, 721-725 (2008).

325. Karlic, R., Chung, H.R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**, 2926-2931 (2010).
326. Reavie, L.*, et al.* Regulation of c-Myc ubiquitination controls chronic myelogenous leukemia initiation and progression. *Cancer Cell* **23**, 362-375 (2013).
327. Quintas-Cardama, A. & Cortes, J. Molecular biology of bcr-abl1-positive chronic myeloid leukemia. *Blood* **113**, 1619-1630 (2009).
328. Albajar, M.*, et al.* MYC in chronic myeloid leukemia: induction of aberrant DNA synthesis and association with poor response to imatinib. *Mol Cancer Res* **9**, 564-576 (2011).
329. Feinberg, A.P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433-440 (2007).
330. Fraga, M.F.*, et al.* Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A* **102**, 10604-10609 (2005).
331. Kaminsky, Z.A.*, et al.* DNA methylation profiles in monozygotic and dizygotic twins. *Nature genetics* **41**, 240-245 (2009).
332. Humpherys, D.*, et al.* Epigenetic instability in ES cells and cloned mice. *Science* **293**, 95-97 (2001).
333. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).
334. Zhang, B.*, et al.* Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome research* (2013).
335. Lister, R.*, et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).
336. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* **103**, 1412-1417 (2006).
337. Wang, Y. & Leung, F.C. An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics* **20**, 1170-1177 (2004).
338. Robertson, K.D. DNA methylation and human disease. *Nature reviews. Genetics* **6**, 597-610 (2005).
339. Deaton, A.M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022 (2011).
340. Rodenhiser, D. & Mann, M. Epigenetics and human disease: translating basic biology into clinical applications. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* **174**, 341-348 (2006).
341. Watt, F. & Molloy, P.L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* **2**, 1136-1143 (1988).
342. Boyes, J. & Bird, A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64**, 1123-1134 (1991).
343. Khulan, B.*, et al.* Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome research* **16**, 1046-1055 (2006).
344. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nature reviews. Genetics* **9**, 465-476 (2008).
345. Laird, P.W. Principles and challenges of genomewide DNA methylation analysis. *Nature reviews. Genetics* **11**, 191-203 (2010).
346. Day, J.J. & Sweatt, J.D. Epigenetic mechanisms in cognition. *Neuron* **70**, 813-829 (2011).

347. Jones, P.A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* **13**, 484-492 (2012).
348. Wilson, V.L. & Jones, P.A. DNA methylation decreases in aging but not in immortal cells. *Science* **220**, 1055-1057 (1983).
349. Bjornsson, H.T*., et al.* Intra-individual change over time in DNA methylation with familial clustering. *JAMA : the journal of the American Medical Association* **299**, 2877-2883 (2008).
350. Boks, M.P*., et al.* The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. *PloS one* **4**, e6767 (2009).
351. Hannum, G*., et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell* **49**, 359-367 (2013).
352. Bocker, M.T*., et al.* Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. *Blood* **117**, e182-189 (2011).
353. Ji, H*., et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467**, 338-342 (2010).
354. Baron, U*., et al.* DNA methylation analysis as a tool for cell typing. *Epigenetics : official journal of the DNA Methylation Society* **1**, 55-60 (2006).
355. Houseman, E.A*., et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 86 (2012).
356. Wolf, S.F., Jolly, D.J., Lunnen, K.D., Friedmann, T. & Migeon, B.R. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proc Natl Acad Sci U S A* **81**, 2806-2810 (1984).
357. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141-1143 (2007).
358. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature reviews. Genetics* **8**, 286-298 (2007).
359. Rodriguez-Paredes, M. & Esteller, M. Cancer epigenetics reaches mainstream oncology. *Nat Med* **17**, 330-339 (2011).
360. Esteller, M. Epigenetics in cancer. *N Engl J Med* **358**, 1148-1159 (2008).
361. Ogawa, O*., et al.* Relaxation of insulin-like growth factor II gene imprinting implicated in Wilms' tumour. *Nature* **362**, 749-751 (1993).
362. Wilson, A.S., Power, B.E. & Molloy, P.L. DNA hypomethylation and human diseases. *Biochim Biophys Acta* **1775**, 138-162 (2007).
363. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1**, 239-259 (2009).
364. Figueroa, M.E*., et al.* DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell* **17**, 13-27 (2010).
365. Schafer, E*., et al.* Promoter hypermethylation in MLL-r infant acute lymphoblastic leukemia: biology and therapeutic targeting. *Blood* **115**, 4798-4809 (2010).
366. Milani, L*., et al.* DNA methylation for subtype classification and prediction of treatment outcome in patients with childhood acute lymphoblastic leukemia. *Blood* **115**, 1214-1225 (2010).
367. Figueroa, M.E*., et al.* MDS and secondary AML display unique patterns and abundance of aberrant DNA methylation. *Blood* **114**, 3448-3458 (2009).
368. Machova Polakova, K., Koblihova, J. & Stopka, T. Role of epigenetics in chronic myeloid leukemia. *Current hematologic malignancy reports* **8**, 28-36 (2013).
369. Chissoe, S.L*., et al.* Sequence and analysis of the human ABL gene, the BCR gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27**, 67-82 (1995).

370. Zion, M.*, et al.* Progressive de novo DNA methylation at the bcr-abl locus in the course of chronic myelogenous leukemia. *Proc Natl Acad Sci U S A* **91**, 10722-10726 (1994).
371. Issa, J.P.*, et al.* Methylation of the ABL1 promoter in chronic myelogenous leukemia: lack of prognostic significance. *Blood* **93**, 2075-2080 (1999).
372. Asimakopoulos, F.A.*, et al.* Prognostic significance of c-ABL methylation in chronic myelogenous leukemia: still an open question. *Blood* **94**, 1141 (1999).
373. Ben-Yehuda, D.*, et al.* Molecular follow-up of disease progression and interferon therapy in chronic myelocytic leukemia. *Blood* **90**, 4918-4923 (1997).
374. Nguyen, T.T.*, et al.* Quantitative measure of c-abl and p15 methylation in chronic myelogenous leukemia: biological implications. *Blood* **95**, 2990-2992 (2000).
375. Jelinek, J.*, et al.* Aberrant DNA methylation is associated with disease progression, resistance to imatinib and shortened survival in chronic myelogenous leukemia. *PloS one* **6**, e22110 (2011).
376. Mills, K.I., Sproul, A.M. & Burnett, A.K. Methylation of the major breakpoint cluster region (M-bcr) in Philadelphia-positive CML. *Leukemia* **7**, 707-711 (1993).
377. Litz, C.E., Vos, J.A. & Copenhaver, C.M. Aberrant methylation of the major breakpoint cluster region in chronic myeloid leukemia. *Blood* **88**, 2241-2249 (1996).
378. Uehara, E.*, et al.* Aberrant methylation in promoter-associated CpG islands of multiple genes in chronic myelogenous leukemia blast crisis. *Oncology letters* **3**, 190-192 (2012).
379. Herman, J.G.*, et al.* Distinct patterns of inactivation of p15INK4B and p16INK4A characterize the major types of hematological malignancies. *Cancer Res* **57**, 837-841 (1997).
380. Oki, Y.*, et al.* Phase II study of low-dose decitabine in combination with imatinib mesylate in patients with accelerated or myeloid blastic phase of chronic myelogenous leukemia. *Cancer* **109**, 899-906 (2007).
381. Dunwell, T.*, et al.* A genome-wide screen identifies frequently methylated genes in haematological and epithelial cancers. *Molecular cancer* **9**, 44 (2010).
382. Qian, J.*, et al.* Aberrant methylation of the death-associated protein kinase 1 (DAPK1) CpG island in chronic myeloid leukemia. *Eur J Haematol* **82**, 119-123 (2009).
383. Bibikova, M.*, et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295 (2011).
384. Busche, S.*, et al.* Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. *Cancer Res* **73**, 4323-4336 (2013).
385. Bullinger, L.*, et al.* Quantitative DNA methylation predicts survival in adult acute myeloid leukemia. *Blood* **115**, 636-642 (2010).
386. Shen, L.*, et al.* DNA methylation predicts survival and response to therapy in patients with myelodysplastic syndromes. *J Clin Oncol* **28**, 605-613 (2010).
387. Lanza, F. & Bi, S. Role of p53 in leukemogenesis of chronic myeloid leukemia. *Stem Cells* **13**, 445-452 (1995).
388. Richly, H., Aloia, L. & Di Croce, L. Roles of the Polycomb group proteins in stem cells and cancer. *Cell death & disease* **2**, e204 (2011).
389. Kulis, M.*, et al.* Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature genetics* **44**, 1236-1242 (2012).
390. Hon, G.C.*, et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome research* **22**, 246-258 (2012).

391. Gauglhofer, C*., et al.* Up-regulation of the fibroblast growth factor 8 subfamily in human hepatocellular carcinoma for cell survival and neoangiogenesis. *Hepatology* **53**, 854-864 (2011).
392. Polnaszek, N., Kwabi-Addo, B., Wang, J. & Ittmann, M. FGF17 is an autocrine prostatic epithelial growth factor and is upregulated in benign prostatic hyperplasia. *The Prostate* **60**, 18-24 (2004).
393. Heer, R., Douglas, D., Mathers, M.E., Robson, C.N. & Leung, H.Y. Fibroblast growth factor 17 is over-expressed in human prostate cancer. *J Pathol* **204**, 578-586 (2004).
394. Benchabane, H*., et al.* Jerky/Earthbound facilitates cell-specific Wnt/Wingless signalling by modulating beta-catenin-TCF activity. *EMBO J* **30**, 1444-1458 (2011).
395. Mo, Y*., et al.* Promoter hypermethylation of Ras-related GTPase gene RRAD inactivates a tumor suppressor function in nasopharyngeal carcinoma. *Cancer Lett* **323**, 147-154 (2012).
396. Wali, A. FHIT: doubts are clear now. *TheScientificWorldJournal* **10**, 1142-1151 (2010).
397. Iwai, M*., et al.* Expression and methylation status of the FHIT gene in acute myeloid leukemia and myelodysplastic syndrome. *Leukemia* **19**, 1367-1375 (2005).
398. Touleimat, N. & Tost, J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325-341 (2012).
399. Zhang, X., Mu, W. & Zhang, W. On the analysis of the illumina 450k array data: probes ambiguously mapped to the human genome. *Frontiers in genetics* **3**, 73 (2012).
400. Roessler, J*., et al.* Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. *BMC research notes* **5**, 210 (2012).
401. Wang, D*., et al.* IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* **28**, 729-730 (2012).
402. Pan, H*., et al.* Measuring the methylome in clinical samples: improved processing of the Infinium Human Methylation450 BeadChip Array. *Epigenetics : official journal of the DNA Methylation Society* **7**, 1173-1187 (2012).
403. Wu, D., Gu, J. & Zhang, M.Q. FastDMA: An Infinium HumanMethylation450 Beadchip Analyzer. *PloS one* **8**, e74275 (2013).
404. Marabita, F*., et al.* An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. *Epigenetics : official journal of the DNA Methylation Society* **8**, 333-346 (2013).
405. Dedeurwaerder, S*., et al.* A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics* (2013).
406. Sengupta, A*., et al.* Deregulation and cross talk among Sonic hedgehog, Wnt, Hox and Notch signaling in chronic myeloid leukemia progression. *Leukemia* **21**, 949-955 (2007).
407. Miyazono, K. Tumour promoting functions of TGF-beta in CML-initiating cells. *Journal of biochemistry* **152**, 383-385 (2012).
408. Naka, K*., et al.* TGF-beta-FOXO signalling maintains leukaemia-initiating cells in chronic myeloid leukaemia. *Nature* **463**, 676-680 (2010).
409. Hamad, A., Sahli, Z., El Sabban, M., Mouteirik, M. & Nasr, R. Emerging therapeutic strategies for targeting chronic myeloid leukemia stem cells. *Stem cells international* **2013**, 724360 (2013).
410. Gallipoli, P*., et al.* Autocrine TNF-alpha production supports CML stem and progenitor cell survival and enhances their proliferation. *Blood* (2013).
411. Peng, C*., et al.* PTEN is a tumor suppressor in CML stem cells and BCR-ABL-induced leukemias in mice. *Blood* **115**, 626-635 (2010).

412. Calabretta, B. & Perrotti, D. The biology of CML blast crisis. *Blood* **103**, 4010-4022 (2004).
413. Kantarjian, H.M.*, et al.* Results of decitabine (5-aza-2'deoxycytidine) therapy in 130 patients with chronic myelogenous leukemia. *Cancer* **98**, 522-528 (2003).
414. Issa, J.P.*, et al.* Phase 1 study of low-dose prolonged exposure schedules of the hypomethylating agent 5-aza-2'-deoxycytidine (decitabine) in hematopoietic malignancies. *Blood* **103**, 1635-1640 (2004).
415. Richter, G.H.*, et al.* EZH2 is a mediator of EWS/FLI1 driven tumor growth and metastasis blocking endothelial and neuro-ectodermal differentiation. *Proc Natl Acad Sci U S A* **106**, 5324-5329 (2009).
416. Dovey, J.S., Zacharek, S.J., Kim, C.F. & Lees, J.A. Bmi1 is critical for lung tumorigenesis and bronchioalveolar stem cell expansion. *Proc Natl Acad Sci U S A* **105**, 11857-11862 (2008).
417. Suva, M.L.*, et al.* EZH2 is essential for glioblastoma cancer stem cell maintenance. *Cancer Res* **69**, 9211-9218 (2009).
418. Pizzatti, L.*, et al.* SUZ12 is a candidate target of the non-canonical WNT pathway in the progression of chronic myeloid leukemia. *Genes, chromosomes & cancer* **49**, 107-118 (2010).
419. Biomarkers Definitions Working, G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* **69**, 89-95 (2001).
420. Strimbu, K. & Tavel, J.A. What are biomarkers? *Current opinion in HIV and AIDS* **5**, 463-466 (2010).
421. Sweet, K., Zhang, L. & Pinilla-Ibarz, J. Biomarkers for determining the prognosis in chronic myelogenous leukemia. *Journal of hematology & oncology* **6**, 54 (2013).
422. Majewski, I.J. & Bernards, R. Taming the dragon: genomic biomarkers to individualize the treatment of cancer. *Nat Med* **17**, 304-312 (2011).
423. Cilloni, D.*, et al.* Quantitative assessment of WT1 expression by real time quantitative PCR may be a useful tool for monitoring minimal residual disease in acute leukemia patients. *Leukemia* **16**, 2115-2121 (2002).
424. Cilloni, D.*, et al.* Sensitivity to imatinib therapy may be predicted by testing Wilms tumor gene expression and colony growth after a short in vitro incubation. *Cancer* **101**, 979-988 (2004).
425. Lessard, J. & Sauvageau, G. Bmi-1 determines the proliferative capacity of normal and leukaemic stem cells. *Nature* **423**, 255-260 (2003).
426. Bhattacharyya, J.*, et al.* BMI-1 expression is enhanced through transcriptional and posttranscriptional regulation during the progression of chronic myeloid leukemia. *Ann Hematol* **88**, 333-340 (2009).
427. Rizo, A.*, et al.* BMI1 collaborates with BCR-ABL in leukemic transformation of human CD34+ cells. *Blood* **116**, 4621-4630 (2010).
428. Sengupta, A., Ficker, A.M., Dunn, S.K., Madhu, M. & Cancelas, J.A. Bmi1 reprograms CML B-lymphoid progenitors to become B-ALL-initiating cells. *Blood* **119**, 494-502 (2012).
429. Lucas, C.M.*, et al.* Cancerous inhibitor of PP2A (CIP2A) at diagnosis of chronic myeloid leukemia is a critical determinant of disease progression. *Blood* **117**, 6660-6668 (2011).
430. Verhaak, R.G.*, et al.* Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* **106**, 3747-3754 (2005).

431. Croft, D.*, et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* **39**, D691-697 (2011).
432. Lim, M.J. & Wang, X.W. Nucleophosmin and human cancer. *Cancer detection and prevention* **30**, 481-490 (2006).
433. Patel, J.P.*, et al.* Prognostic relevance of integrated genetic profiling in acute myeloid leukemia. *N Engl J Med* **366**, 1079-1089 (2012).
434. Falini, B.*, et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med* **352**, 254-266 (2005).
435. Lindstrom, M.S. NPM1/B23: A Multifunctional Chaperone in Ribosome Biogenesis and Chromatin Remodeling. *Biochemistry research international* **2011**, 195209 (2011).
436. Murano, K., Okuwaki, M., Hisaoka, M. & Nagata, K. Transcription regulation of the rRNA gene by a multifunctional nucleolar protein, B23/nucleophosmin, through its histone chaperone activity. *Mol Cell Biol* **28**, 3114-3126 (2008).
437. Wouters, B.J.*, et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088-3091 (2009).
438. Georgiou, G.*, et al.* Development of acute myeloid leukemia with NPM1 mutation, in Ph-negative clone, during treatment of CML with imatinib. *Leukemia* **26**, 824-826 (2012).
439. Konoplev, S.*, et al.* Molecular characterization of de novo Philadelphia chromosome-positive acute myeloid leukemia. *Leuk Lymphoma* **54**, 138-144 (2013).
440. Piccaluga, P.P.*, et al.* Cytoplasmic mutated nucleophosmin (NPM1) in blast crisis of chronic myeloid leukaemia. *Leukemia* **23**, 1370-1371 (2009).
441. Oki, Y.*, et al.* Mutations and promoter methylation status of NPM1 in myeloproliferative disorders. *Haematologica* **91**, 1147-1148 (2006).
442. Grossmann, V.*, et al.* A deep-sequencing study of chronic myeloid leukemia patients in blast crisis (BC-CML) detects mutations in 76.9% of cases. *Leukemia* **25**, 557-560 (2011).
443. Lidonnici, M.R.*, et al.* Expression of the transcriptional repressor Gfi-1 is regulated by C/EBP{alpha} and is involved in its proliferation and colony formation-inhibitory effects in p210BCR/ABL-expressing cells. *Cancer Res* **70**, 7949-7959 (2010).
444. Soliera, A.R.*, et al.* Gfi-1 inhibits proliferation and colony formation of p210BCR/ABL-expressing cells via transcriptional repression of STAT 5 and Mcl-1. *Leukemia* **26**, 1555-1563 (2012).
445. van der Meer, L.T., Jansen, J.H. & van der Reijden, B.A. Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia* **24**, 1834-1843 (2010).
446. Kok, C.H.*, et al.* Low GFI1 expression in white blood cells of CP-CML patients at diagnosis is strongly associated with subsequent blastic transformation. *Leukemia* **27**, 1427-1430 (2013).
447. O'Hare, T., Eide, C.A. & Deininger, M.W. Bcr-Abl kinase domain mutations, drug resistance, and the road to a cure for chronic myeloid leukemia. *Blood* **110**, 2242-2249 (2007).
448. Huang, M.*, et al.* The growth factor independence-1 (Gfi1) is overexpressed in chronic myelogenous leukemia. *Acta haematologica* **123**, 1-5 (2010).
449. Shastry, B.S. SNPs: impact on gene function and phenotype. *Methods in molecular biology* **578**, 3-22 (2009).
450. Person, R.E.*, et al.* Mutations in proto-oncogene GFI1 cause human neutropenia and target ELA2. *Nature genetics* **34**, 308-312 (2003).

451. Khandanpour, C*., et al.* A variant allele of Growth Factor Independence 1 (GFI1) is associated with acute myeloid leukemia. *Blood* **115**, 2462-2472 (2010).
452. Tang, C*., et al.* Tyrosine kinase inhibitor resistance in chronic myeloid leukemia cell lines: investigating resistance pathways. *Leuk Lymphoma* **52**, 2139-2147 (2011).
453. Eadie, L. The University of Adelaide (2013).
454. Russell, J. & Zomerdijk, J.C. RNA-polymerase-I-directed rDNA transcription, life and works. *Trends in biochemical sciences* **30**, 87-96 (2005).
455. Horman, S.R*., et al.* Gfi1 integrates progenitor versus granulocytic transcriptional programming. *Blood* **113**, 5466-5475 (2009).
456. Khandanpour, C*., et al.* Growth factor independence 1 protects hematopoietic stem cells against apoptosis but also prevents the development of a myeloproliferative-like disease. *Stem Cells* **29**, 376-385 (2011).
457. Yucel, R., Karsunky, H., Klein-Hitpass, L. & Moroy, T. The transcriptional repressor Gfi1 affects development of early, uncommitted c-Kit+ T cell progenitors and CD4/CD8 lineage decision in the thymus. *J Exp Med* **197**, 831-844 (2003).
458. Huh, H.J*., et al.* CD34, RAB20, PU.1 and GFI1 mRNA expression in myelodysplastic syndrome. *International journal of laboratory hematology* **31**, 344-351 (2009).
459. Kuriyama, M*., et al.* Identification of AF-6 and canoe as putative targets for Ras. *J Biol Chem* **271**, 607-610 (1996).
460. Radziwill, G*., et al.* Regulation of c-Src by binding to the PDZ domain of AF-6. *EMBO J* **26**, 2633-2644 (2007).
461. Lopez-Karpovitch, X., Cardenas, R. & Piedras, J. Immunophenotypic characteristics of the blast crisis in chronic myeloid leukemia. *Revista de investigacion clinica; organo del Hospital de Enfermedades de la Nutricion* **49**, 31-36 (1997).
462. Khalidi, H.S*., et al.* The immunophenotype of blast transformation of chronic myelogenous leukemia: a high frequency of mixed lineage phenotype in "lymphoid" blasts and A comparison of morphologic, immunophenotypic, and molecular findings. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* **11**, 1211-1221 (1998).
463. Kantarjian, H.M., Deisseroth, A., Kurzrock, R., Estrov, Z. & Talpaz, M. Chronic myelogenous leukemia: a concise update. *Blood* **82**, 691-703 (1993).
464. Schurch, C., Riether, C., Matter, M.S., Tzankov, A. & Ochsenbein, A.F. CD27 signaling on chronic myelogenous leukemia stem cells activates Wnt target genes and promotes disease progression. *The Journal of clinical investigation* **122**, 624-638 (2012).
465. Tang, M*., et al.* Dynamics of chronic myeloid leukemia response to long-term targeted therapy reveal treatment effects on leukemic stem cells. *Blood* **118**, 1622-1631 (2011).
466. Mustjoki, S., Lundan, T., Knuutila, S. & Porkka, K. Appearance of bone marrow lymphocytosis predicts an optimal response to imatinib therapy in patients with chronic myeloid leukemia. *Leukemia* **21**, 2363-2368 (2007).
467. Wang, H*., et al.* Imatinib mesylate (STI-571) enhances antigen-presenting cell function and overcomes tumor-induced CD4+ T-cell tolerance. *Blood* **105**, 1135-1143 (2005).
468. Appel, S*., et al.* Effects of imatinib on monocyte-derived dendritic cells are mediated by inhibition of nuclear factor-kappaB and Akt signaling pathways. *Clin Cancer Res* **11**, 1928-1940 (2005).
469. Chen, E.Y*., et al.* Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics* **28**, 105-111 (2012).

470. Asif, H.M. & Sanguinetti, G. Large-scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics* **27**, 1277-1283 (2011).
471. Novershtern, N., Regev, A. & Friedman, N. Physical Module Networks: an integrative approach for reconstructing transcription regulation. *Bioinformatics* **27**, i177-185 (2011).
472. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification: patterns and paradigms. *Nature reviews. Genetics* **10**, 295-304 (2009).
473. Ehrlich, M. DNA methylation in cancer: too much, but also too little. *Oncogene* **21**, 5400-5413 (2002).
474. Duan, Z. & Horwitz, M. Gfi-1 takes center stage in hematopoietic stem cells. *Trends in molecular medicine* **11**, 49-52 (2005).
475. Colomo, L.*, et al.* Clinical impact of the differentiation profile assessed by immunophenotyping in patients with diffuse large B-cell lymphoma. *Blood* **101**, 78-84 (2003).
476. Hans, C.P.*, et al.* Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood* **103**, 275-282 (2004).
477. Muris, J.J.*, et al.* Immunohistochemical profiling based on Bcl-2, CD10 and MUM1 expression improves risk stratification in patients with primary nodal diffuse large B cell lymphoma. *J Pathol* **208**, 714-723 (2006).
478. Choi, W.W.*, et al.* A new immunostain algorithm classifies diffuse large B-cell lymphoma into molecular subtypes with high accuracy. *Clin Cancer Res* **15**, 5494-5502 (2009).
479. Meyer, P.N.*, et al.* Immunohistochemical methods for predicting cell of origin and survival in patients with diffuse large B-cell lymphoma treated with rituximab. *J Clin Oncol* **29**, 200-207 (2011).
480. Walter, K.*, et al.* DNA methylation profiling defines clinically relevant biological subsets of non-small cell lung cancer. *Clin Cancer Res* **18**, 2360-2373 (2012).
481. Noushmehr, H.*, et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510-522 (2010).
482. Fang, F.*, et al.* Breast cancer methylomes establish an epigenomic foundation for metastasis. *Science translational medicine* **3**, 75ra25 (2011).
483. Benso, A., Di Carlo, S. & Politano, G. A cDNA microarray gene expression data classifier for clinical diagnostics based on graph theory. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM* **8**, 577-591 (2011).
484. Krzywinski, M. & Altman, N. Points of Significance: Significance, P values and t-tests. *Nature methods* **10**, 1041-1042 (2013).