THE IDENTIFICATION OF GENETIC AND EPIGENETIC CHANGES THAT CONTRIBUTE TO TYPE 1 DIABETES

YING YING WONG

Discipline of Paediatrics, School of Medicine,
University of Adelaide

June 2021

A thesis submitted to the University of Adelaide as the requirement
for the degree of Doctor of Philosophy

Table of Contents

List of Tables

List of Figures

Declaration

I, Ying Ying WONG, author of this thesis entitled, "*The Identification of Genetic and Epigenetic Changes that Contribute to Type 1 Diabetes*", certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, 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 in my name, 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 permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Adelaide Graduate Research Scholarship.

06/11/2021

Abstract

Type 1 diabetes (T1D) results from an immune cell mediated destruction of insulin-producing pancreatic β cells. Currently there is no cure for T1D. The exact cause for T1D is unknown but growing evidence points to the contribution of both genetic and environmental factors, leading to a breakdown in immunological tolerance normally maintained by Regulatory T (Treg) cells. The exact environmental contributions to T1D progression are not well characterised but emerging studies suggest that they may alter the immune system via epigenetic modification. Recent data strongly link the breakdown in tolerance in T1D and other autoimmune diseases to alterations in the transcriptional program in CD4$^+$ T cells, however, the molecular mechanisms are not well understood.

This work proposes that in T1D causal genetic risk SNPs alter the gene expression patterns in CD4$^+$ Treg and or T helper cells by either disrupting or creating new TF (transcription factor) binding sites in regulatory elements (enhancers) located in genetic susceptibility regions and this may combine with environmentally induced epigenetic change and alter chromatin accessibility. Current methods to identify the functional consequences and mechanisms of these changes are complex, time consuming and expensive as generally they can only examine one TF/binding site at a time, involve TF binding site prediction, which has a high degree of false positives/negatives and require large quantities of starting material making them challenging for application on limited clinical samples.

To overcome these limitations, and to functionally annotate genetic risk of T1D, this study employs genome wide approaches including ATAC-seq and RNA-seq to compare the DNA accessibility and transcriptomes in CD4$^+$ Treg and Th (Helper T)/Tconv (Conventional T) cells isolated from individuals with established T1D and sibling-matched healthy controls. By incorporating case-control ATAC-seq and TF footprints this study prioritises 111 and 96 T1D-associated SNPs in Treg and Tconv cells, respectively, that may play a role in mediating the

disease susceptibility and subsequently contributing to the loss of tolerance in T1D. Using a bioinformatic pipeline to integrate case-control ATAC-seq differentially accessible peaks and RNA-seq differentially expressed genes with Hi-C 3D connectivity maps this study identifies 42 and 21 dysregulated gene targets in Treg and Tconv cells, respectively. Those targets include *TIGIT*, *MAF* and *IL2* and the enhancers regulating those loci showed differential accessibility and are enriched for T1D SNPs and differential TF footprint signals. One theory to explain such observation is T1D SNPs and epigenetic alterations may alter or disrupt TF occupancy at these loci contributing to dysregulated target gene regulation.

This study identifies changes in chromatin structure in T1D samples relative to healthy controls, enabling the identification of changes driven by both genetic and epigenetic variation that correlates with an altered transcriptional program in T1D. T1D associated SNPs at these regions can then be correlated with alterations in TF binding and putative epigenetically modified T1D regions can be validated in follow-up functional assays to demonstrate causality. This study captures chromatin and transcriptional changes between T1D and healthy individuals but it does not have the capability to distinguish if the changes are the driver or the consequence of the disease because the case cohort contains only established T1D from a single time point. In order to infer causality those changes would need to be tracked and validated over a timeline of disease progression in a longitudinal cohort. Nonetheless, this work provides a novel 3D genomic approach to functionally annotating the genetic risk and epigenetic changes that directly or indirectly result in altered gene expression, and promising preliminary data warranting further investigation on the causal functional role of the dysregulated gene targets in T1D.

“Time is the most precious gift you can give to someone.”

Gloria Tesch

Publications and Awards

Publications:

BROWN, C. Y., SADLON, T., HOPE, C. M., **WONG, Y. Y.**, WONG, S., LIU, N., WITHERS, H., BROWN, K., BANDARA, V., GUNDSAMBUU, B., PEDERSON, S., BREEN, J., ROBERTSON, S. A., FORREST, A., BEYER, M. & BARRY, S. C. 2020. *Molecular Insights into Regulatory T-Cell Adaptation to Self, Environment, and Host Tissues: Plasticity or Loss of Function in Autoimmune Disease.* Frontiers in Immunology, 11.

LIU, N., SADLON, T., **WONG, Y. Y.**, PEDERSON, S., BREEN, J. & BARRY, S. C. 2020. *3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk.* bioRxiv, 2020.09.04.279554.

ZAMMIT, N. W., **WONG, Y. Y.**, WALTERS, S., WARREN, J., BARRY, S. C., GREY, S. T. 2020. *RelA Governs a Network of Islet-Specific Metabolic Genes Necessary for Beta-Cell Function.* Available at SSRN: https://ssrn.com/abstract=3733072 or http://dx.doi.org/10.2139/ssrn.3733072.

DINH, T. D., BREEN, J., NICOL, B., SMITH, K. M., NICHOLLS, M., EMERY, A., **WONG, Y. Y.**, BARRY, S. C., YAO, H. H.-C., ROBKER, R. L. & RUSSELL, D. L. 2021. *Progesterone receptor-A isoform interaction with RUNX transcription factors controls chromatin remodelling at promoters during ovulation.* bioRxiv, 2021.06.17.448908.

HOPE, C. M., HUYNH, D., **WONG, Y. Y.**, OAKEY, H., PERKINS, G. B., NGUYEN, T., BINKOWSKI, S., BUI, M., CHOO, A. Y. L., GIBSON, E., HUANG, D., KIM, K. W., NGUI, K., RAWLINSON, W. D., SADLON, T., COUPER, J. J., PENNO, M. A. S., BARRY, S. C. & ON BEHALF OF THE ENDIA STUDY, G. 2021. *Optimization of Blood Handling and Peripheral Blood Mononuclear Cell Cryopreservation of Low Cell Number Samples.* International journal of molecular sciences, 22, 9129.

Awards Received:

Adelaide Graduate Research Scholarships (AGRS), University of Adelaide, Australia (2017-2020)

The Environmental Determinants of Islet Autoimmunity (ENDIA) Study NHMRC Grant Funded Supplementary Scholarship (2017-2020)

Robinson Research Institute (RRI) Travel Grant, University of Adelaide (2019)

Hans-Jürgen & Marianne Ohff Research Grant (2019)

Australian Bioinformatics and Computational Biology Society (ABACBS) Travel Award (2018)

Adelaide Medical School Research Travel Award (2018)

Global Learning Travel Grant, University of Adelaide (2017)

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my principal supervisor, Professor Simon Barry for his continuous support, invaluable advice and dedicated involvement in every step throughout my PhD. His motivation and great enthusiasm in science have been deeply influential in shaping my academic trajectory at the University of Adelaide, and my career aspirations in research. Instead of conforming to conventional paths he has inspired me that there is always more than one path to success, and it is often non-linear. I thank him for believing in me, and all the wonderful opportunities that have been given to me throughout my PhD, such as training experiences at Germany and various research collaborations. He prioritises professional development and he looks after the well-being of his employees and students. I thank him for providing such a positive work environment that fuels sustainable growth, where it is safe to try, fail, learn and innovate, and the team members are always happy to assist each other to grow as a team.

I am also extremely grateful to my co-supervisor, Dr Timothy Sadlon, for his unwavering support, insight and belief in my abilities, which was vital in making this thesis a reality. He has extended an enormous amount of assistance throughout my academic research, in providing invaluable suggestions, reviewing and proofreading my work. His passion in science and immense knowledge in the field with focus on high quality and credibility research have helped me grow tremendously as a scientist over the years. His optimism and constant encouragement have guided me through the multitude of challenges in the conduct of research and made me see the positive side of everything, even when I am at my lowest point. It has been a great privilege and honour to work under his guidance.

I would also like to express my appreciation and gratitude to Professor Jennifer Cooper and Dr Jessica Harbison, for providing this study with valuable specimens from T1D patients and their healthy siblings, which were collected as part of their Australian Type 1 Diabetes and the Gut

(TIGs) study. Without their contributions this study would have never been accomplished. I also thank the ENDIA (Environmental Determinants of Islet Autoimmunity) study and Dr Megan Penno for providing this project with a supplementary scholarship, in addition to the Adelaide Graduate Research Scholarships awarded by the University. I also gratefully acknowledge all the participants and their families involved in the study for their contributions to the advancement in T1D research.

I am thankful to our collaborators, Dr Marc Beyer and Professor Joachim Schultze from the Life and Medical Sciences Institute (LIMES), University of Bonn, Germany for hosting my training at their laboratory to learn to perform ATAC-seq experiment as well as data analysis. I thank Maren, the former PhD student, for guiding me through each step of ATAC-seq library preparation; Kathrin and Klause for ATAC-seq data analysis workshops; Simone for Tn5 supply and insightful discussion about Tn5 manufacturing; Kristian for his expertise in genomics and epigenomics sequencing as well as his unwavering contributions through email correspondence after I returned to Australia; and all the lab members for making this trip a pleasant and rewarding experience.

I am also grateful to Dr Jimmy Breen, Dr Steve Pederson and Ning Liu for their expertise in bioinformatics and statistical analyses. I have benefited greatly from all the bioinformatics and R workshops hosted by the team, which have helped set up the groundwork for my venture in big data analyses, and their invaluable input and advice throughout my PhD have been extremely helpful. I also thank the University for providing this study information technology support and access to high performance computing (Phoenix, HPC2) and private cloud service (Cronos).

I would like to extend my sincere thanks to all the past and current members of Barry Lab: Cheryl, Vincent, Katherine, Veronika, Baggy, Silvana, Chris, Silvie, Holly, Kate, Jacqui, Leila and Emma, for being incredibly inspirational, fun and supportive, both professionally and personally. I thank Cheryl for her constructive feedback and advice throughout my PhD in improving multiple aspects of my research; Vincent who has been a great bioinformatics troubleshooting partner and contributed selflessly on some issues I encountered in my research; Katherine for being an awesome companion and providing emotional support during difficult times; Veronika for being a great mentor whose wisdom has helped me see the opportunity in every difficulty personally and professionally; Baggy for his efforts in processing and biobanking the clinical samples used in my PhD research; Chris for his expertise in flow cytometry including panel design for this project; and Silvana for her stash of endless guilty pleasure sweet treats!

No words can ever be strong enough to express my gratitude to my parents for their unconditional love, support and constant encouragement. Although I have spent a significant part of my life living abroad and away from family after high school, I never forget my upbringing and the values I have been raised on. Thanks for providing me with high quality education and the freedom to pursue what I want to do. I also thank them for always listening, letting me vent, making efforts to be healthy and providing rescue to kitchen disasters over the phone!

Last but not least, I very much appreciate my partner, Liam for his understanding, guidance and encouragement through this venture. Thanks for being a great teacher, instilling positive values in me and navigating me during rough times with your wisdom. I have learnt from him that life is not about making the right decisions but sticking to the decisions you have committed to and making them right. I look forward to continuing growing together. Also, thanks for all the sweet treats and amazing meals from the mum, Anita, who has always made sure I am well fed in my hectic schedule.

Abbreviations

°C	degrees Celsius
µg	microgram
µL	microLitre
µM	microMolar
3C	Chromosome Conformation Capture
3D	Three-dimensional
3'UTR	3 prime untranslated region
APC	Antigen Presenting Cells
ATAC-seq	Assay for Transposase-Accessible Chromatin Sequencing
bp	base pairs
BP	Biological Process
Cat	Catalogue
CC	Cellular Component
CD	Crohn's Disease
cDNA	complimentary DNA
ChIP-seq	Chromatin Immunoprecipitation sequencing
CO_2	Carbon Dioxide
CPM	Counts Per Million
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeat
DA	Differentially accesibile
DE	Differentially expressed
DGE	Differential gene expression
DNA	Deoxyribonucleic acid
FACS	Fluorescence-Activated Cell Sorting

FBS	Fetal Bovine Serum
FDR	False discovery rate
FOXP3	Forkhead box Protein3
g	gravitational force
GLM	General Linear Model
GO	Gene Ontology
GWAS	Genome-wide association studies
H3K27ac	Acetylation of histone H3 on lysine 27
Hi-C	a variant of chromosome conformation capture assay
HiChIP	Hi-C with chromatin immunoprecipitation
HREC	Human Research Ethics Committee
IBD	Inflammatory bowel disease
IFN	Interferons
Ig	Immunoglobulin
IL	Interleukin
IPEX	Immunodysregulation polyendocrinopathy enteropathy X-linked
iTreg	Induced regulatory T cell
kb	kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
M	Molar
MF	Molecular Function
mg	miligram
MHC	Major histocompatibility complex
mL	miliLitre
mRNA	messenger Ribonucleuc acid

MS	Multiple Sclerosis
MSigDB	Molecular Signature Database
NFκB	Nuclear factor kappa B
ng	nanogram
NK	Natural Killer
nM	nanoMolar
ns	not significant
PBMC	Peripheral blood mononuclear cell
PBS	Phosphate-buffered saline
PCA	Principal component analysis
pcHi-C	Promoter Capture Hi-C
PCR	Polymerase chain reaction
qRT-PCR	Quantitative Real Time PCR
RA	Rheumatoid arthritis
RIN	RNA Integrity Number
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
RT	room temperature
RUV	remove unwanted variation
SEM	standard error of measurement
SNP	Single Nucleotide Polymorphisms
T1D	Type 1 diabetes
Tconv	T conventional
TCR	T-cell receptor
TF	transcription factor

TGFβ	Transforming growth factor beta
Th	T helper
TNF	Tumor necrosis factor
TNFSF	Tumor necrosis factor superfamily
Treg	Regulatory T cell
UC	Ulcerative Colitis
Ver	Version
WCH	The Women's and Children's Hospital

CHAPTER 1: LITERATURE REVIEW

1.1 Immune system

The human immune system is made up of a group of highly specialised cells that cooperate to protect the body from harmful, pathogens, which is essential for survival. The main duties of the body's immune system include neutralising and eliminating foreign invaders such as bacteria, viruses or parasites that have come into contact with our body, as well as reacting against abnormal cells such as tumour cells while suppressing self-reactive cells[14]. There are two main categories of the immune system – innate immunity and adaptive immunity (Figure 1-1). Innate immunity is the first line of defence which confers non-specific immune responses that come into play immediately or within hours of an antigen's appearance in the body. The main players in the innate immune response include natural killer cells, dendritic cells, macrophages, eosinophils and basophils and neutrophils [46]. They do not demonstrate immunological memory and thus are unable to be tuned to respond more rapidly to repeat exposures. However, they are essential for presenting antigens to cells of the adaptive immune system. In contrast, adaptive immunity takes longer to come into effect but it elicits a more lasting and specific immune response. The main players of adaptive immunity are B and T cells. The adaptive immunity generally comes into play within 5 or 6 days after the barrier breach and first exposure, which is then followed by a gradual resolution of the infection[16]. The key feature of adaptive immunity is immunological memory, meaning ensuring a more rapid and greater magnitude of response to a repeat exposure. As this system needs to be able to respond to the vast number of pathogens the body is potentially exposed to throughout life this necessitates a highly diverse set of cells. Antigen recognition is driven by specific receptors on the B-cell and T-cell termed the B-cell (BCR) and T-cell (TCR) receptors respectively. The high diversity of antigen recognition is achieved by specific BCR and TCR domain gene rearrangements during the differentiation of B and T cells to generate a vast collection of clones with a range of affinities for different subsets of antigens[47]. T and B cells will undergo activation and proliferation upon antigen encounter and they demonstrate a much better antigen recognition, elimination and memory during subsequent exposure to the same antigen[16] (Figure 1-2).

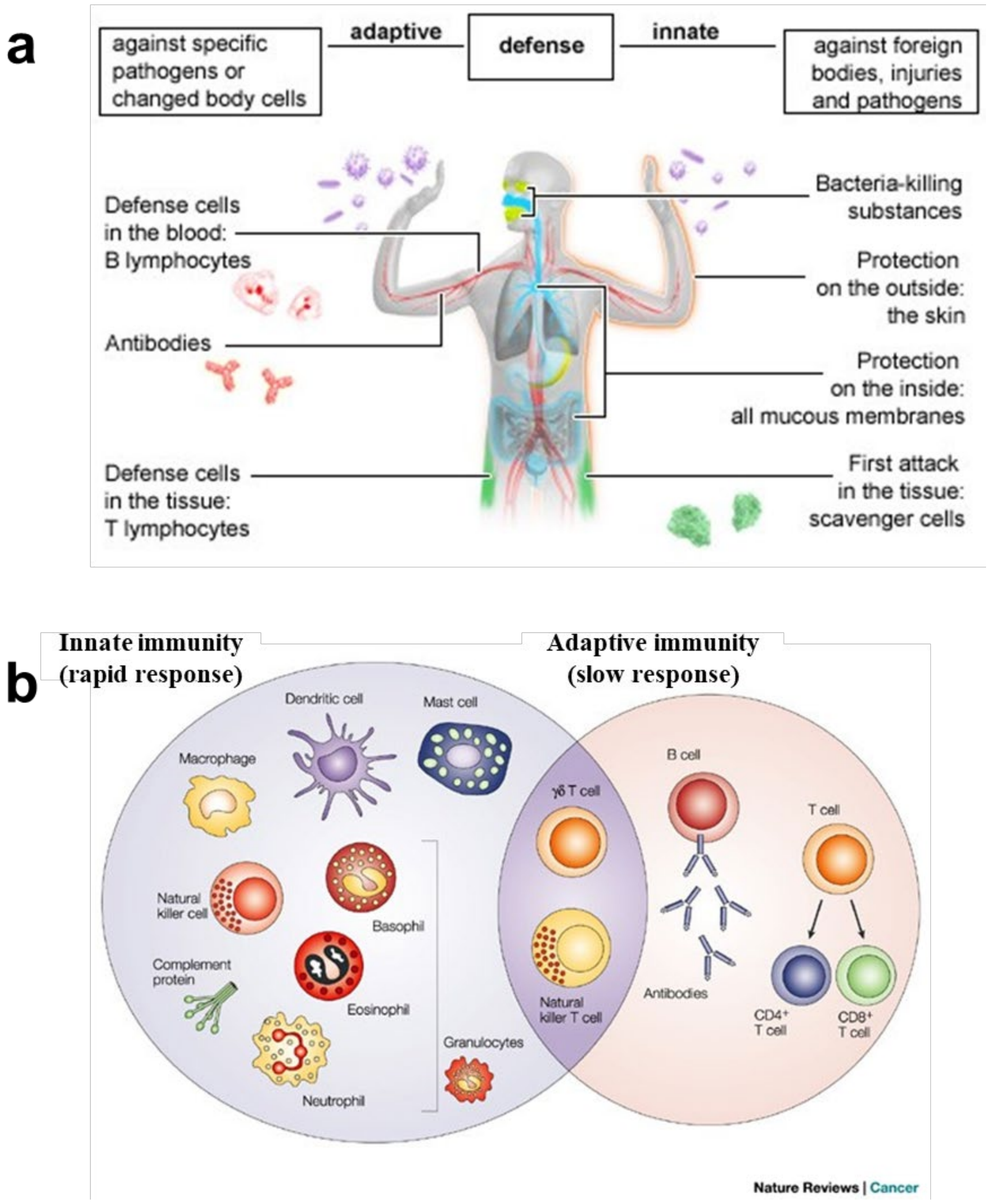

a
against specific pathogens or changed body cells
adaptive
defense
innate
against foreign bodies, injuries and pathogens
Defense cells in the blood: B lymphocytes
Antibodies
Defense cells in the tissue: T lymphocytes
Bacteria-killing substances
Protection on the outside: the skin
Protection on the inside: all mucous membranes
First attack in the tissue: scavenger cells
b
Innate immunity (rapid response)
Adaptive immunity (slow response)
Dendritic cell
Mast cell
Macrophage
Natural killer cell
Basophil
Complement protein
Eosinophil
Granulocytes
Neutrophil
γδ T cell
Natural killer T cell
B cell
T cell
Antibodies
CD4+ T cell
CD8+ T cell
Nature Reviews | Cancer

Figure 1-1 | Overview of the innate and adaptive immunity (Online [14]) (a) and the elements or cell types involved in each mechanism (Dranoff [34]) (b).

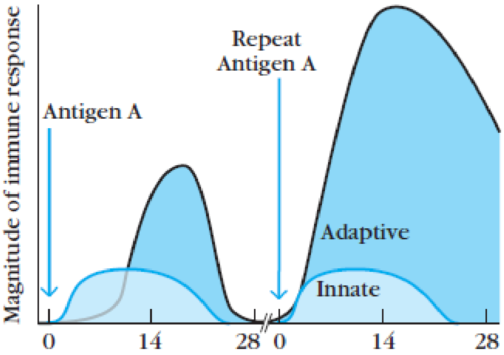


Figure 1-2 | Differences in the magnitude of primary and secondary immune response between innate and adaptive mechanism (Jones [16]).
In adaptive immunity, a second exposure to the same antigen results in a much stronger immune response than innate immunity which peaks earlier in time. On the contrary, innate immunity has no antigenic memory and it elicits similar response during subsequent exposures to the same antigen.

1.2 Development and differentiation of CD4$^+$ T cell Subsets

Fundamental roles of CD4$^+$ T cell subsets in adaptive immune response. There are two common types of T cells – CD4$^+$ helper T cells and the CD8$^+$ T cells (Figure 1-1; b). They are defined by the type of co-receptor glycoprotein expressed on the cell surface and the class of MHC (major histocompatibility complex molecules) recognised by the TCR and co-receptor during antigen presentation. CD4 glycoprotein is expressed on CD4$^+$ helper T (Th) and it recognises antigens presented by MHC class II molecules, whereas CD8 glycoprotein is expressed on CD8$^+$ cytotoxic T cells and it recognises antigens presented by MHC class I molecules (Figure 1-3).

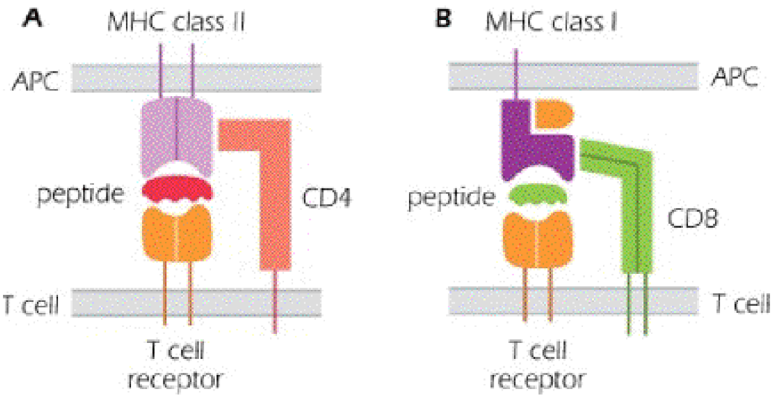


Figure 1-3 | Association of CD4$^+$ Helper T cell (left) and CD8$^+$ T cell (right) with MHC class II and Class I molecules[13].

The pivotal role of Conventional T (Tconv) cells is to coordinate the different arms of the adaptive immune system to shape an effective and highly specific secondary response against target antigens. Cytotoxic T cells are important for killing cancerous or virally infected cells mainly through exocytosis of cytolytic granule contents, thereby triggering apoptosis of the target cell [48]. CD4$^+$ T cells are generated in the thymus and released into the circulation as antigen "inexperienced" or Naïve CD4$^+$ Conventional T (Tconv) cells that can differentiate into different Tconv subtypes that shape the immune response to particular pathogens. These include Type 1 T helper (Th1) cells which are involved in host immunity against intracellular bacteria and viruses; Type 2 T helper (Th2) cells that are responsible for host immunity against extracellular parasites such as helminth, Type 17 T helper (Th17) cells which are critical for host defence against fungal and extracellular bacterial infections [49], as well as follicular helper T (Tfh) cells which provide specialized help to B cells in the context of germinal center formation, affinity maturation and production of high affinity antibodies[50]. A special subset of CD4$^+$ cells known as regulatory T (Treg) cells are involved in maintaining immune homeostasis[51] and preventing deleterious immune reactions by potently suppressing other T effector subsets[52]. Figure 1-4 illustrates the development of the T helper lineage in the thymus. In summary, developing thymocytes go through a series of defined developmental stages and are subject to distinct selection processes - positive selection and negative selection, ensuring that only the T-cell populations that are both self-MHC restricted (able to recognise antigen peptides in complex with MHC) and self-tolerant (do not respond to endogenous or self-peptides) survive and are released into the circulation [18].

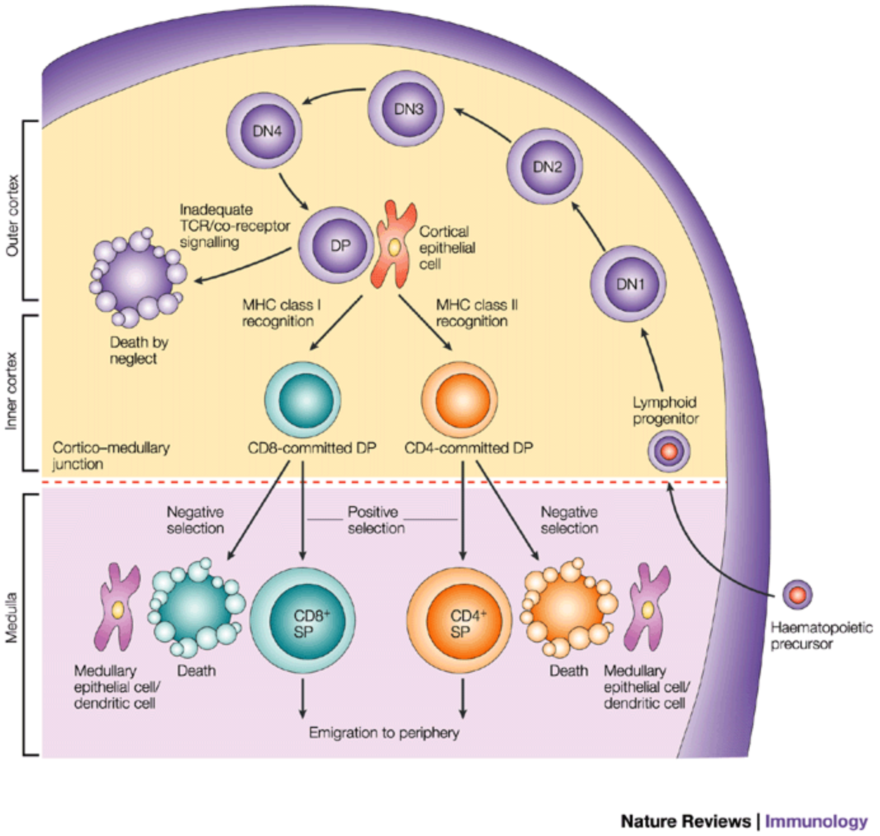


Figure 1-4 | T-cell development in the thymus (Germain [18]).

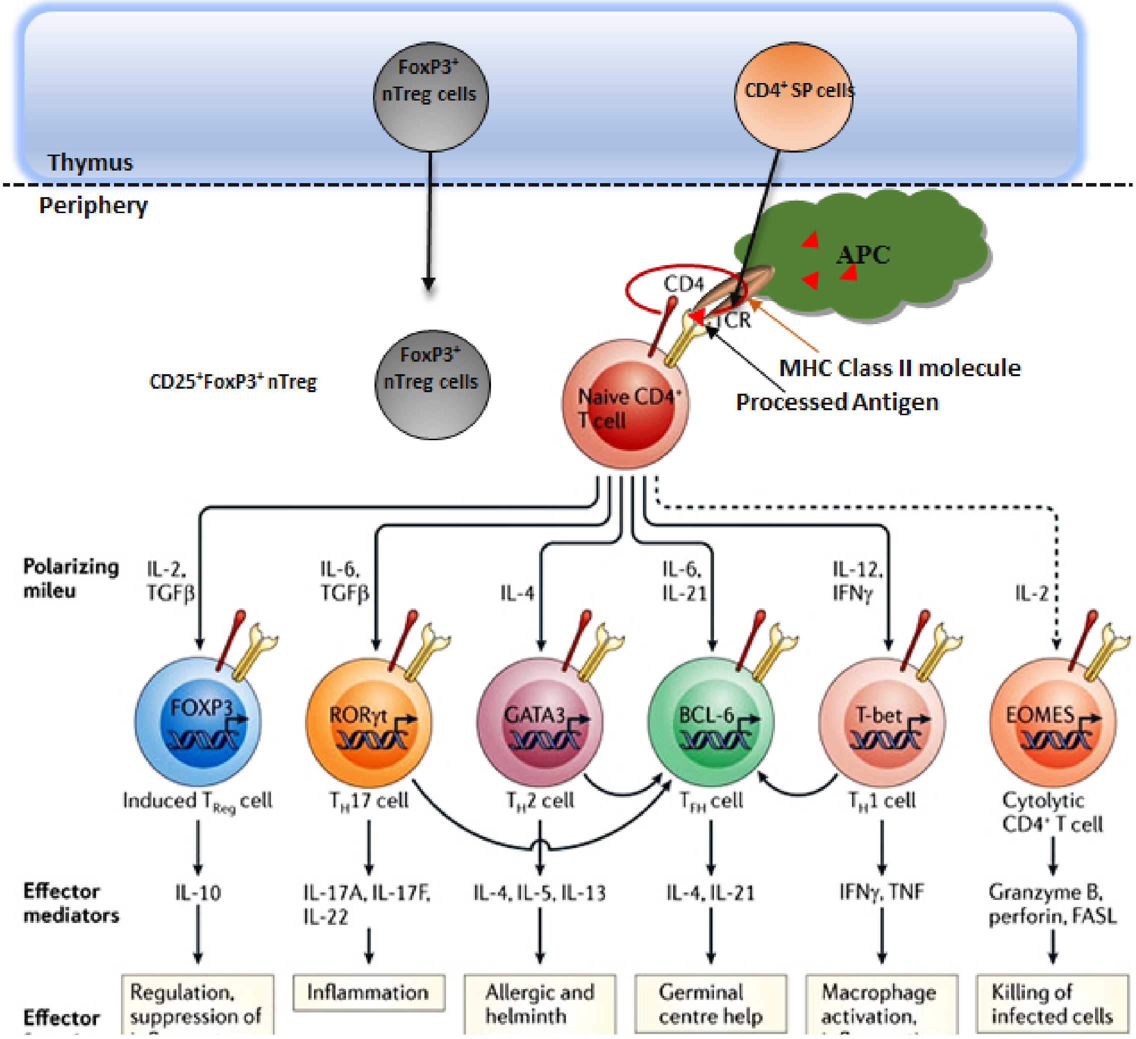


Figure 1-5 | Differentiation of naïve CD4$^+$ T cells (Figure modified from Swain, McKinstry [19]).
Following antigen stimulation, Naïve CD4+ T cells can differentiate into functionally distinct depending on the characteristic cytokines network. Regulatory T (Treg) cells can be thymic-derived (nTreg) or induced from naïve CD4+ T cells (iTreg).

CD4$^+$ Conventional T (Tconv) cell Differentiation. As shown in Figure 1-5, naïve CD4$^+$ T cells are activated upon priming by antigen-presenting cells (APC) and the presence of specific cytokines during TCR signaling directs the differentiation of naïve CD4$^+$ T cell into distinct subsets of Conventional T (Tconv) cells with different biological roles [53]. T cell subset specification is tailored by interactions with APCs such as dendritic cells and peritoneal macrophages, the dose and type of presented antigen, the affinity of TCR-peptide interaction, type of cytokines secreted, costimulatory interactions [54] and lineage-specific master transcriptional factors (Figure 1-5). The differentiation of each T cell subset relies on the induction of its master transcription factor (TF) as well as reinforcement by other lineage-restricted TFs[53] and suppression of master TFs responsible for lineage commitment in other subsets, ensuring only highly specific T cell subsets with defined roles are developed.

1.3 Development and differentiation of Regulatory T (Treg) cells

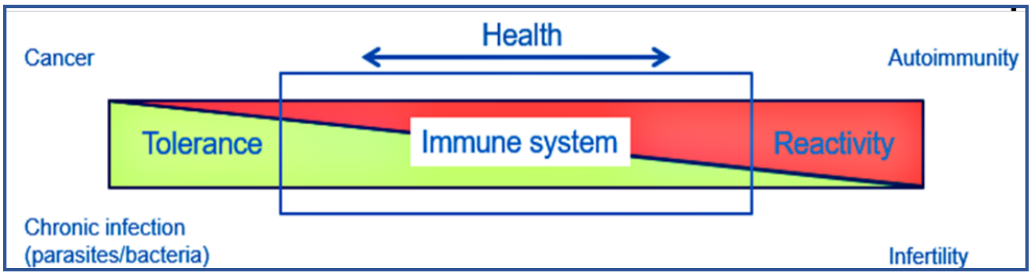


Figure 1-6 | Immunological balance for maintenance of immune homeostasis and self-tolerance.
Regulatory T (Treg) and T helper (Th) lineages play a prominent role in maintaining a fine balance in the immune response to establish a healthy responsive immune system.

Treg in immune homeostasis and tolerance. A healthy responsive immune system requires a fine balance between tolerance and reactivity. Breakdown of immune tolerance from excessive reactivity can trigger the body to develop an autoimmunity against its own tissues and cells, whereas a state of excessive nonreactivity or immune unresponsiveness may increase the susceptibility to infection and cancer (Figure 1-6). In healthy individuals, rare immune cells called Regulatory T (Treg) cells maintain tolerance and prevent inappropriate immune responses[55-58]. Due to the vast repertoire of antigen recognition potential generated during TCR gene rearrangement, a number of mechanisms exist to eliminate cells containing a TCR that can recognise self-antigens at sufficient affinity to mount an immune response. Thymic negative selection clonally deletes autoreactive thymocytes that exhibit strong binding affinity with "self" peptides by apoptosis but this process is not 100% efficient and thus, some autoreactive T cells may escape the process of thymic negative selection and eventually be released into the periphery. Tregs are a critical component of peripheral tolerance that act to suppress these self-reactive clones[59, 60]. Tregs are also involved in limiting collateral damage to the host cells as a result of normal immune responses by turning off the response once the pathogens have been cleared. In addition, Tregs are also crucial in establishing tolerance against ubiquitous commensal organisms and adaptation to harmless foreign antigens such as chemicals and drugs that have come into contact with the host cells.

Natural Treg and Induced Treg cells. It has been well established that there are two main subpopulations of regulatory T cells – natural (or thymic) and inducible (or peripheral) regulatory T cells. Thymically-derived natural regulatory T cells (tTreg) represent 5-10% of peripheral CD4$^+$ T cells and they constitutively express the transcription factor FOXP3 (Forkhead box P3) and the cell-surface marker CD25, which is the high affinity receptor for IL2 [58]. Thymic Treg cells develop through recognition of MHC/self-peptides that demonstrated medium affinity cognate interaction[61]. Induced Treg cells are induced in the periphery from naïve CD4$^+$ T cells upon antigen recognition in the presence of tolerance promoting factors such as TGF-β and All-trans Retinoic Acid (ATRA). These cells turn on the expression of FOXP3 [62] and have a fundamental role in establishing mucosal immune tolerance to non-self antigens including food and microbiota [58].

The X-linked transcription factor (TF) Forkhead box P3 protein (FOXP3) is essential for Treg's stability and phenotype[56]. Mutations in FOXP3 have been shown to cause a lethal multi-organ inflammatory disease termed scurfy in mice[56] and IPEX (Immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome) [63] in humans. In both cases loss of functional FOXP3 results in a lack of functional Tregs. Thus, Tregs are essential for suppression of complex autoimmune diseases and maintenance of immune homeostasis. The observation that deletion of regulatory T cells in genetically targeted adult mice contributes to the development of intestinal inflammation[64] and scurfy-like syndrome shows that Tregs are required throughout life to restrain inappropriate immune response[65]. Growing evidence also points to Treg defects in many autoimmune- and inflammatory diseases, including T1D.

1.4 Type 1 Diabetes (T1D)

1.4.1 Type 1 Diabetes (T1D) is a multistage disease

Type 1 diabetes (T1D) is a chronic autoimmune disease caused by the immune system attacking and eventually destroying the insulin-producing β-cells in the islets of Langerhans of the pancreas [66-70]. A lack of insulin results in loss of control of blood glucose levels and thus elevated blood glucose levels. Sustained elevated levels of blood glucose is associated with long-term damage and failure of various organs, especially the eyes, kidneys, nerves, heart, and blood vessels[71]. Currently there is no cure for T1D and patients depend on regular long-term blood glucose management through lifelong insulin administration. Worldwide incidence of T1D is increasing but it is most prevalent in Caucasian populations in developed countries including Australia. The primary drivers and mechanisms underlying rapid rates of increased incidence in T1D are unknown but appear to manifest in genetically susceptible individuals and are linked with a plethora of environmental triggers[66, 67, 69, 72-75]. Environment influences that are associated with T1D development or progression are diets, vitamin D[66, 67], viruses such as enteroviruses[66, 74, 76], rotaviruses[74], coxsackieviruses[77], hygiene hypothesis[67, 73], chemicals[73] (which exert immunomodulatory effects) and gut microflora[74]. Nonetheless, no specific agent has been identified as the major contributor in T1D and that suggests additive or synergistic contributions as the drivers of the disease pathogenesis.

Islet autoantibodies are strong predictors of T1D. T1D is characterised by the presence of β-cell autoantibodies which are the key markers of T1D and are detectable prior to any signs of overt disease. β-cell autoantibodies against the four well-characterised T1D autoantigens are insulin (IAA), glutamic acid decarboxylase 2 (GAD65), protein phosphatase-like protein (IA-2A) and zinc transporter 8 (ZnT8A)[66, 70, 72, 78-81]. The presence of islet autoantibodies presents a greater risk of developing T1D, especially in children born with increased T1D genetic risk with the lifetime risk of developing symptomatic T1D approaches 100% once two or more islet autoantibodies are detected in genetically susceptible children [70, 80, 82].

Staging of T1D. T1D is described as a continuum that progresses at variable but predictable rates through multiple developmental stages before clinical manifestations[70]. As shown in Figure 1-7, Stage 1 is the start of T1D where individuals are tested positive for two or more islet autoantibodies. Clearly the immune system has already started attacking the insulin-producing β cells before Stage 1, although there are no clinical symptoms and blood glucose levels are normal. Stage 2 of T1D includes presymptomatic individuals that have two or more islet autoantibodies and show abnormal blood glucose levels as a result of significant loss of β cell mass[17]. The 5-year risk of symptomatic disease at this stage is estimated to be 75%, while the lifetime risk approaches 100%[83]. Lastly, Stage 3 represents the clinical diagnosis phase where individuals manifest clinical symptoms of T1D such as polyuria, polydipsia, weight loss, and fatigue[83].

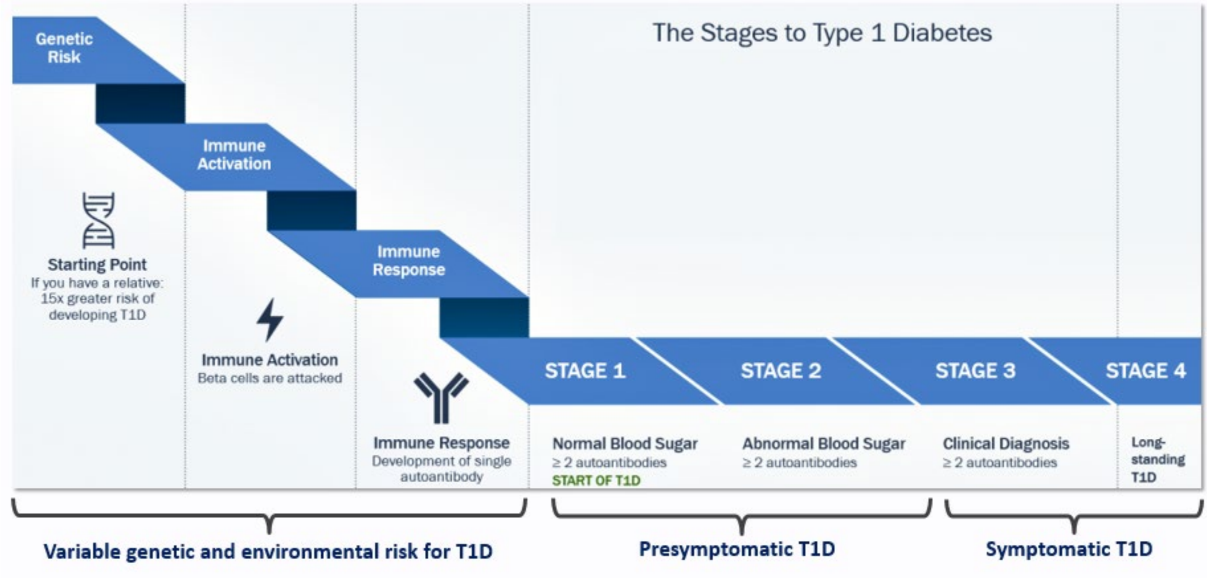
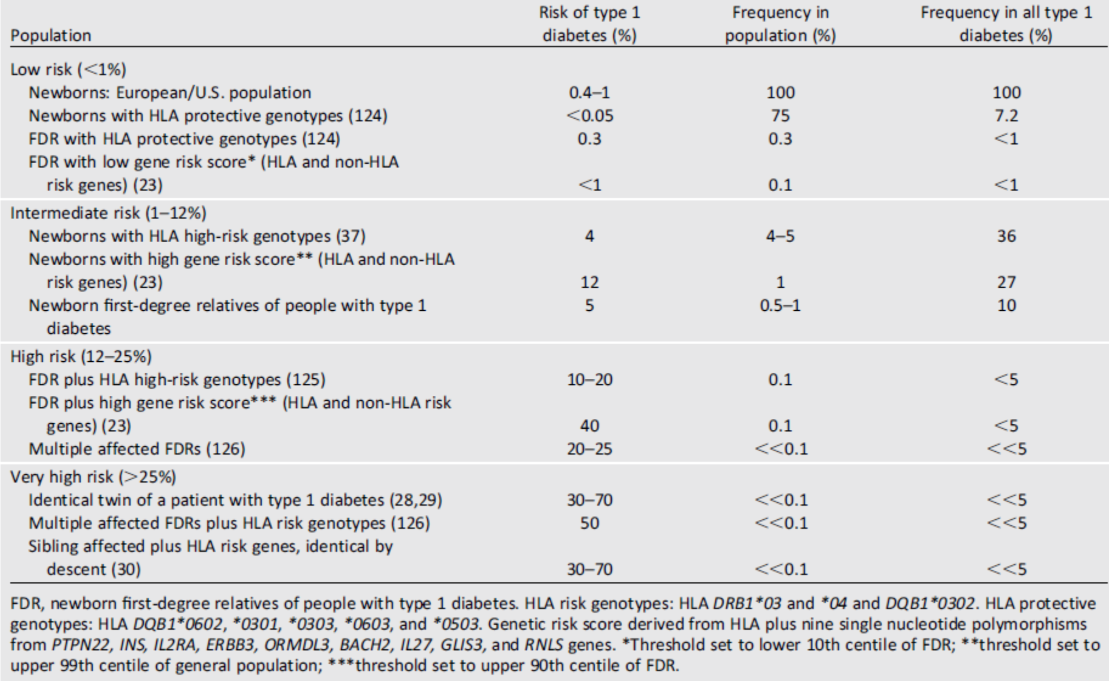


Figure 1-7 | Staging classification and disease progression of T1D (Adapted from Children's Diabetes Foundation 2016[17]).

Table 1-1. T1D risk stratification by family backgroud and genetic contribution[70].



Population	Risk of type 1 diabetes (%)	Frequency in population (%)	Frequency in all type 1 diabetes (%)
Low risk (<1%)			
Newborns: European/U.S. population	0.4–1	100	100
Newborns with HLA protective genotypes (124)	<0.05	75	7.2
FDR with HLA protective genotypes (124)	0.3	0.3	<1
FDR with low gene risk score* (HLA and non-HLA risk genes) (23)	<1	0.1	<1
Intermediate risk (1–12%)			
Newborns with HLA high-risk genotypes (37)	4	4–5	36
Newborns with high gene risk score** (HLA and non-HLA risk genes) (23)	12	1	27
Newborn first-degree relatives of people with type 1 diabetes	5	0.5–1	10
High risk (12–25%)			
FDR plus HLA high-risk genotypes (125)	10–20	0.1	<5
FDR plus high gene risk score*** (HLA and non-HLA risk genes) (23)	40	0.1	<5
Multiple affected FDRs (126)	20–25	<<0.1	<<5
Very high risk (>25%)			
Identical twin of a patient with type 1 diabetes (28,29)	30–70	<<0.1	<<5
Multiple affected FDRs plus HLA risk genotypes (126)	50	<<0.1	<<5
Sibling affected plus HLA risk genes, identical by descent (30)	30–70	<<0.1	<<5

FDR, newborn first-degree relatives of people with type 1 diabetes. HLA risk genotypes: HLA *DRB1*03* and **04* and *DQB1*0302*. HLA protective genotypes: HLA *DQB1*0602*, **0301*, **0303*, **0603*, and **0503*. Genetic risk score derived from HLA plus nine single nucleotide polymorphisms from *PTPN22, INS, IL2RA, ERBB3, ORMDL3, BACH2, IL27, GLIS3,* and *RNLS* genes. *Threshold set to lower 10th centile of FDR; **threshold set to upper 99th centile of general population; ***threshold set to upper 90th centile of FDR.

1.4.2 T1D is a complex genetic disease

Familial studies have demonstrated a strong genetic contribution in T1D. For instance, a positive association exists between disease prevalence in first-degree relatives[68], with the highest disease concordance among monozygotic (MZ) compared to dizygotic twins (DZ). However, no single gene defect causes T1D, but rather genetic variation at multiple gene regions convey disease risk to an individual. The Human Leukocyte Antigen (HLA) locus on human chromosome 6p21 demonstrates the strongest association with T1D (~50% risk). The HLA locus encodes for 3 classes of MHC proteins which are important for immune functions (Figure 1-8). HLA class II alleles are strongly associated with T1D, particularly specific alleles at the DRB1, DQA1 and DQB1 loci[68] (Figure 1-8). HLA haplotypes (combination of HLA gene variations) DRB1*0401-DQB1*0302 and DRB1*0301-DQB1*0201 confer the greatest susceptibility, whereas DRB1*1501 and DQA1*0102-DQB1*0602 show disease protection[84]. Class I HLA genes have also been shown to contribute to T1D independently of Class II[75] HLA genes. In addition, Genome-wide association studies (GWAS) and other genetic studies have identified an additional 59 non-HLA risk loci that each contributes an incremental increase in T1D risk [2, 3, 85-87]. Surprisingly, the non-HLA risk loci appear to be driving the increased incidence of T1D over the last few decades instead of HLA loci, implying the profile of genetic contributions in T1D is changing. To understand how these risk loci contribute to disease, several fine-mapping[2, 88] and eQTL[89] (expression Quantitative Trait Loci) studies have been performed to identify and refine i) the list of T1D genetic variation that reside at these risk loci, ii) the identity of the potential target genes of these risk loci. However, how many of these genetic loci contribute to disease remains unsolved[2].

1.4.3 T1D has a significant epigenetic component

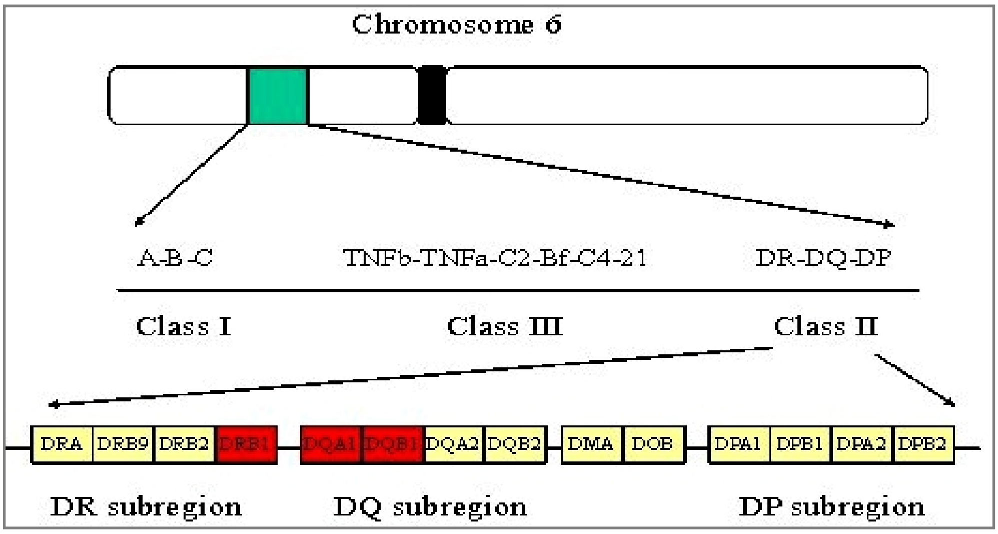


Figure 1-8 | HLA region on Chromosome 6 is associated with the highest risk of T1D[35].

Although it has been shown that certain HLA risk genotypes are associated with the highest risk of developing T1D (~50% risk), multiple lines of evidence demonstrated a strong contribution of non-genetic elements in the disease. For instance, the increasing worldwide incidence of T1D[90] over the past three decades cannot be explained by a change in genotype frequencies in the population, implying a role for non-genetic, likely an environmental influence. For instance, the increasing worldwide incidence of T1D over the past 30 years can only be explained by environmental changes and this is supported by data from migration cohorts where immigrants take on the risk of the local population[68, 91]. Furthermore, only a small percentage of children with a genetic predisposition develop T1D and the concordance rate for islet autoimmunity among monozygotic twins is never 100%, suggesting strong contribution of environmental influences acquired after birth in the pathogenesis of T1D [92]. Many potential non-genetic factors involved in triggering islet autoimmunity and/or promoting progression to T1D have been proposed such as duration of breast feeding [93], viral infections [94], standard of hygiene [95] and gut microbiome [96], but how each of these factors affects disease predisposition has remained poorly defined. A likely route for environmental factors to alter the immune system is via epigenetic mechanisms such as histone modification, DNA methylation and altered microRNA expression and alteration in all these have been reported in T1D patients[97-102]. Epigenetics refers to heritable phenotype changes in gene expression that do not involve alterations in the DNA sequence. Epigenetic modifications are not determined genetically but they can be affected by genetic makeup. In T1D, differential methylation patterns were observed within or in close proximity to T1D risk loci in T1D-discordant monozygotic twins which may suggest differential expression of the corresponding gene products between affected and unaffected siblings. It was also observed that CD4$^+$ T cells from latent autoimmune diabetes in adults (LADA) was associated with significantly higher enrichment of DNA methylation compared with healthy controls, and their corresponding *FOXP3* promoter region was hypermethylated [103]. It remains unclear how they interact with genetic effects to contribute to the loss of immune regulation and destruction of pancreatic β cells in T1D. In T1D, it is likely that the epigenetic modifications further increase the disruption (e.g. altered accessibility) at regulatory regions containing non-coding genetic risk, or alter the transcription/regulatory potential of other targets which combine with the genetic risk to amplify the possibility of developing the disease. Further characterisation of where these epigenetic modifications occur and the functional consequences of these alterations on gene expression is required to understand islet autoimmunity and progression to T1D.

1.4.4 The balance of Treg and Tconv is perturbed in T1D

The number and function of Treg and Tconv cells are tightly regulated to maintain immune homeostasis, and loss of this fine balance contributes to autoimmune disease (Figure 1-6) [104]. T1D is the consequence of loss of tolerance to self-antigens as a result of defects in central tolerance which has failed to eliminate autoreactive T cells in the thymus, as well as defects in peripheral tolerance which has failed to suppress autoreactive T cells that escaped central tolerance[105]. Hence, T1D is associated with both functional defects in the Treg cells and development of resistance to Treg-mediated immune suppression in the Tconv cells[105, 106]. The importance of Treg cells in restraining autoimmune responses is exemplified in FOXP3 mutation variants of IPEX (immune dysregulation, polyendocrinopathy, enteropathy, X-linked) syndrome in humans which manifests itself as a triad of Type 1 diabetes (the most common endocrine disorder present in IPEX individuals), diarrhea and psoriasis-like skin disorder [107]. It is interesting to note, however that there are no mutations in FOXP3 itself in other autoimmune diseases, and this suggests the Treg defect is linked to the genes directly or indirectly controlled by FOXP3 [108].

It has also been shown that adoptive transfer of Treg cells into non-obese diabetic (NOD) mice model ameliorates disease while reducing Treg number or function exacerbates disease[104]. Consistency of Treg specific contributions to T1D in humans and mice is demonstrated by the findings in both of functional defects in Treg cells, increased Treg apoptosis and Treg suppression resistance in Tconv cells [105, 109]. In T1D multiple immune cell types have also been implicated including NKT, CD8, B cells (Figure 1-9) but evidence from several studies suggests they can either alter Treg/Th cell balances or are themselves affected by a change in this balance[108].

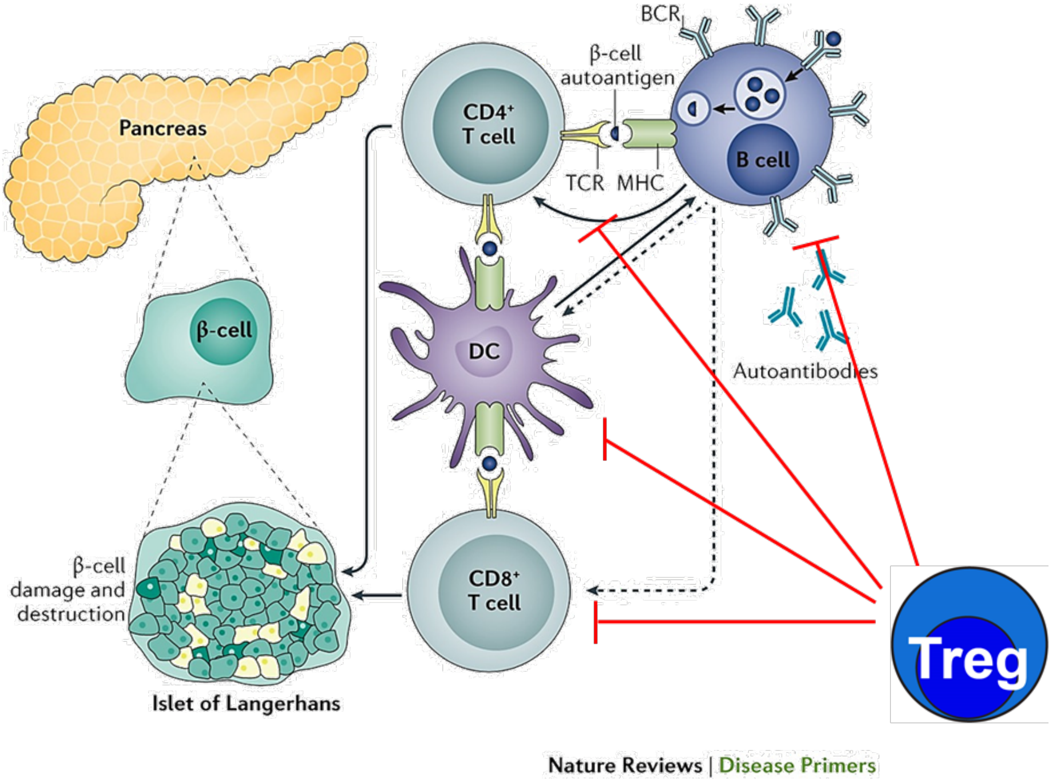


Figure 1-9 | The pathogenesis of T1D contributed by various cell types (Adapted from Katsarou, Gudbjörnsdottir [24]).
The exposure of B cells to β-cell autoantigens leads to the production of islet-specific autoantibodies, and antigen presentation by B cells and DCs drives the activation of β-cell-specific T cells. Growing evidence suggest that Tregs can suppress B cell responses and B cell-mediated antibody production but not just CD8+ and CD4+ T cell-mediated responses.

1.4.5 Many T1D risk loci are linked to transcriptional regulation in Treg and Tconv cells

Genome Wide Association Studies (GWAS) have identified more than 50 non-HLA risk loci that contribute to T1D risk and ongoing fine-mapping studies aim to refine the list of candidate SNPs at these loci[2, 88, 89]. However, a major limitation of this type of study is that it is probability-based and annotation is by linear proximity. This means that the named target of the polymorphism may not be altered. While this has led to the identification of several non-HLA genes that are associated with T1D risk, including *PTPN22*, *CTLA4*, and *IL2RA*[2, 81] most of the genetic variation of these risk regions does not alter a protein coding gene. As a result of this there has been a concerted effort to functionally annotate the genetic risk using new genomic annotations including non-coding regulatory elements such as enhancers. Integrated analysis of autoimmune disease-associated loci and epigenetic/transcriptional elements in the human genome[2, 88] now indicate that the majority of disease-associated SNPs map to lymphoid cell-specific transcriptional enhancers thereby linking disease risk with altered cell-specific transcription of target genes in these cell types[88]. Importantly, majority of the T1D associated risk loci are implicated in the immune regulation (Figure 1-10) and the risk SNPs are enriched in Tconv/Treg cell-specific enhancers[2, 88]. Each T1D genetic variant in isolation may have a subtle contribution on disease risk but each may alter a key function in the immune system and its interaction with pancreatic β cells. Their enrichment is particularly prominent at enhancers that are activated in response to antigen recognition and are enriched for activation dependant transcriptional regulation, classifying them as super enhancers[88, 110] in Tconv cells. Critically most of these enhancers that respond to stimulation in Tconv cells are subjected to repression in Treg cells [11, 111]. Together these data suggest that autoimmune disease-associated variation preferentially target enhancers that are essential for both Treg (where they are repressed), or Tconv cells (where they are activated). The impact of T1D risk loci on transcription in Treg/Tconv cells is supported by the finding of altered gene expression patterns in T1D compared with healthy controls[89, 98, 112].

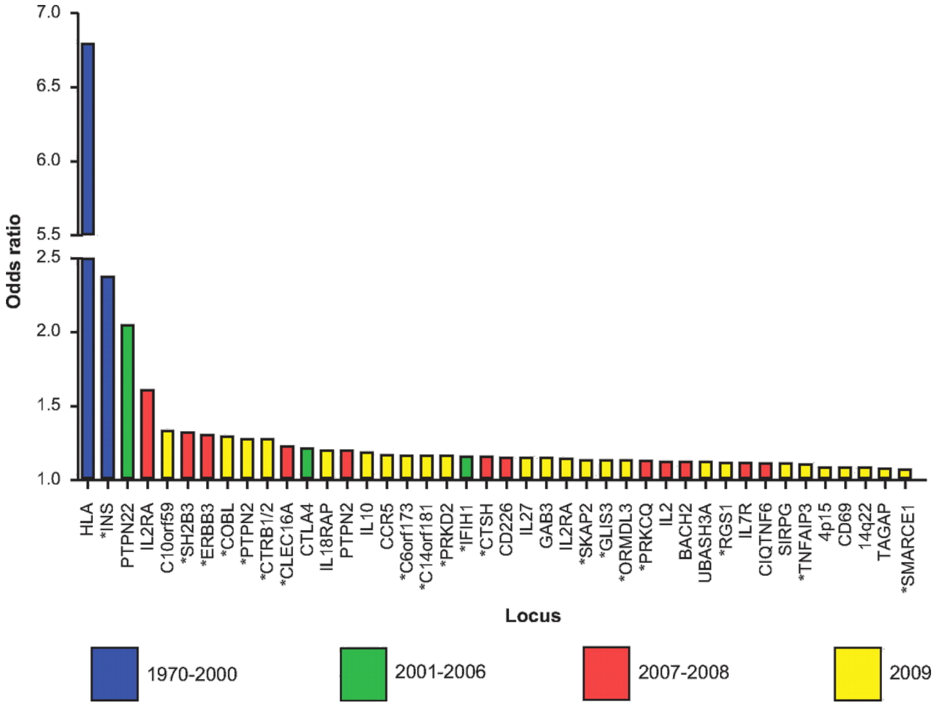


Figure 1-10 | T1D risk loci are implicated in the immune regulation (Pociot, Akolkar [10]).
Color-coding indicates the year of discovery of these candidate gene loci. The y-axis indicates the best estimate of the odds ratio for risk alleles at each of the risk loci. Asterisks indicate expression in human pancreatic islets.

1.5 Epigenetic regulation of cell type and condition specific gene expression

1.5.1 Chromatin structure and transcriptional regulation

Human genome is approximately 3 Gbp (gigabase pair) long and only about 1 percent of DNA is made up of protein-coding genes. The remaining 99% represents non-coding genome and they have to be compacted to fit into the nucleus in such a way that allows cell-type specific regulation of gene functions. Non-coding DNA does not provide instructions for making proteins, but they contain important gene regulatory elements such as enhancers that are integral for control of gene activity, determining when and where genes are turned on and off. Since every cell carries the same DNA and genome, it is therefore the levels of gene expression (epigenome) that determine the cell type of a cell. To fit the DNA into a nucleus it needs to be packaged into a highly compacted structure known as chromatin. Different states of chromatin are defined by the folding and compaction of chromatin - condensed (heterochromatin) versus relaxed (euchromatin) structure (Figure 1-11). Changes to the structure of chromatin affect gene expression - condensed chromatin is generally associated with gene repression whereas the open form of chromatin allows gene transcription. Nucleosomes are assembled as a result of interactions between DNA and histone proteins (Figure 1-11). A single nucleosome is a histone octamer formed from two copies of each of the four core histone proteins, wrapped by 147bp of DNA 1.75 times [113].

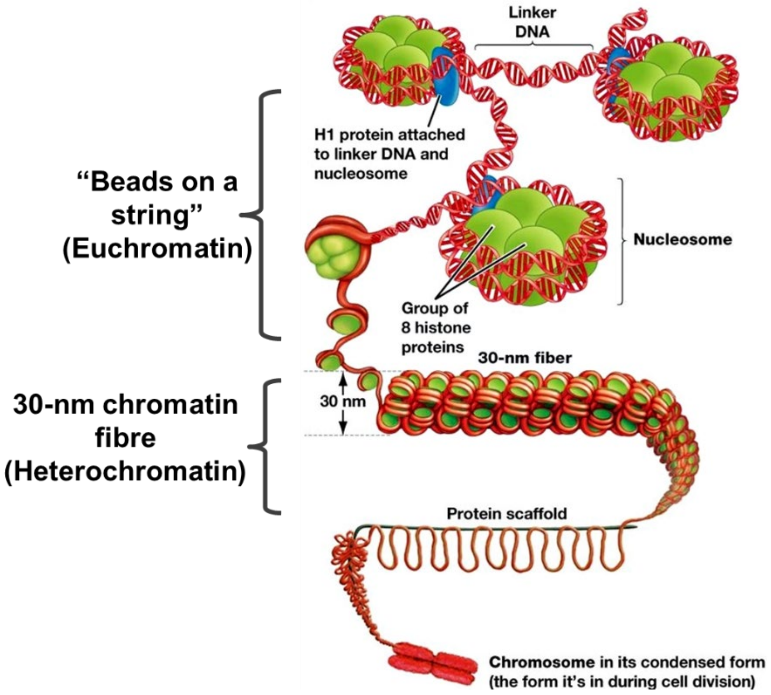

Linker DNA
H1 protein attached to linker DNA and nucleosome
"Beads on a string" (Euchromatin)
Nucleosome
Group of 8 histone proteins
30-nm fiber
30 nm
30-nm chromatin fibre (Heterochromatin)
Protein scaffold
Chromosome in its condensed form (the form it's in during cell division)

Figure 1-11 | Nucleosomes and histone proteins in DNA packaging [12].

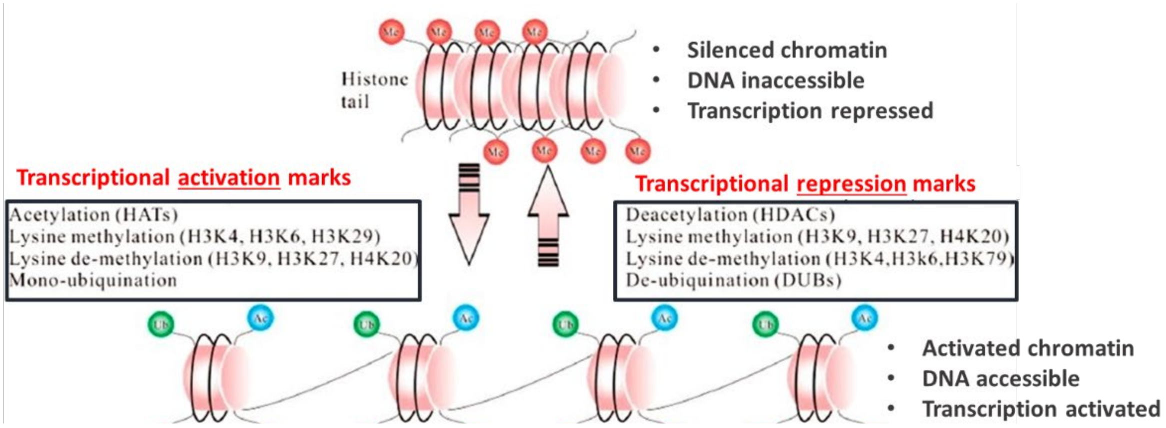

Me
Histone
tail
• Silenced chromatin
• DNA inaccessible
• Transcription repressed
Transcriptional activation marks
Acetylation (HATs)
Lysine methylation (H3K4, H3K6, H3K29)
Lysine de-methylation (H3K9, H3K27, H4K20)
Mono-ubiquination
Transcriptional repression marks
Deacetylation (HDACs)
Lysine methylation (H3K9, H3K27, H4K20)
Lysine de-methylation (H3K4,H3k6,H3K79)
De-ubiquination (DUBs)
Ub
Ac
• Activated chromatin
• DNA accessible
• Transcription activated

Figure 1-12 | Epigenetic modifications at histone N-terminal tails (adapted from Wang, Yin [23]).

The organization of nucleosomes controls the access of transcription factors (TFs) and other DNA-binding proteins to DNA. Nucleosome organization can be influenced by sequence-specific histone binding preferences, ATP-dependant chromatin modellers or other DNA binding proteins. Nucleosomes are repositioned by chromatin remodelling to control gene expression. This chromatin remodelling occurs through post-translational epigenetic modification, which refers to heritable but reversible changes in gene expression without alterations in DNA sequence. Typically every cell in our body carries the same genome but the epigenome changes with cell and tissue types[114] because it acts to program the cells to only express genes that are relevant for a particular cell type at a precise time. Some of the well-characterised epigenetic modifications that are involved in remodelling chromatin structure occur at histone N-terminal tails and include acetylation, methylation and phosphorylation (Figure 1-12; a, b). Some of these covalent histone modifications have roles in DNA replication and DNA repair[115].

Different epigenetic states, as represented by different histone codes, correlate to distinct "readouts" of the genetic information, leading to either gene activation or silencing [116]. As demonstrated in Figure 1-12 and Figure 1-13, some of the gene activation histone signatures are acetylation/methylation at histone 3 at lysine 4 (H3K4ac/H3K4me) and acetylation at histone 3 at lysine 27 (H3K27ac); whereas gene repression are associated with marks such as methylation of histone 3 at lysine 9 and 27 (H3K9 and H3K27)[23]. While histone methylation is associated with reversible transcriptional repression, DNA methylation at cytosine residues in CpG dinucleotides leads to highly stable long term gene repression[115]. Nonetheless, both modification pathways can be dependant on one another[115].

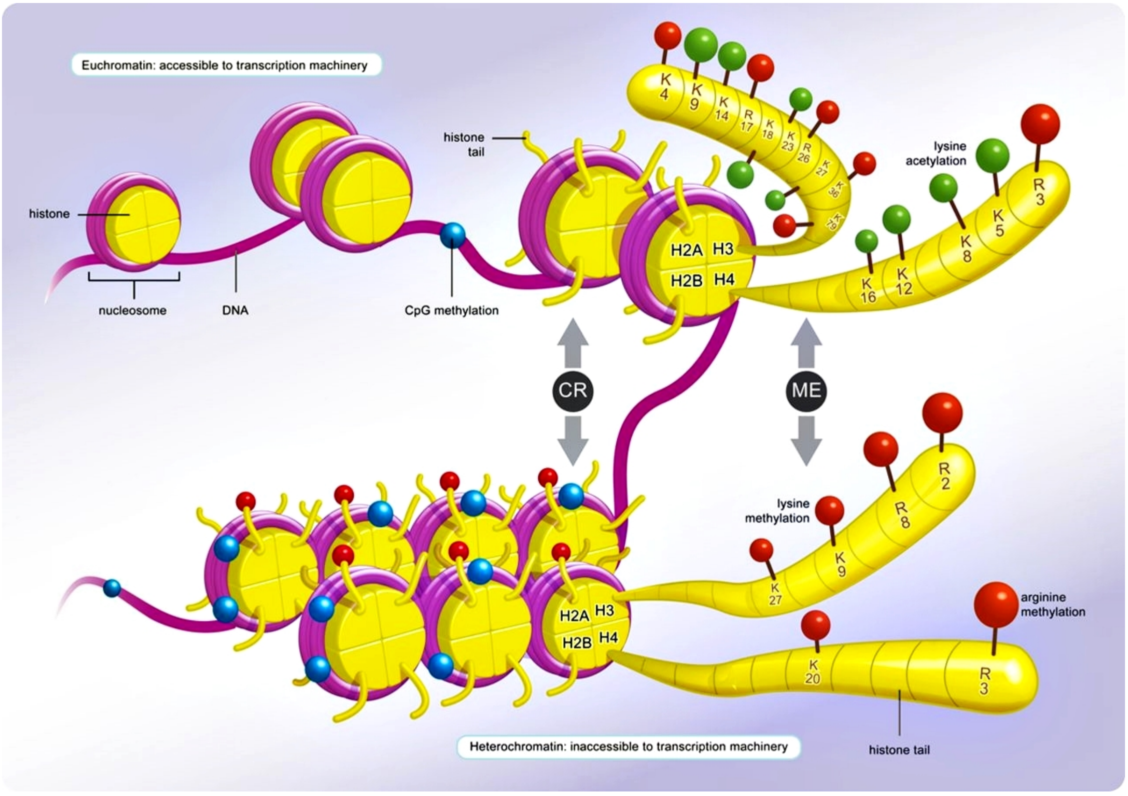


Figure 1-13 | Epigenetic modifications at histone tails and DNA turn gene on or off (Adapted from Kubicek [15] 2011).

1.5.2 Gene enhancers are critical in driving target gene expression

Transcriptional enhancers (TEs) are DNA segments that are generally a few hundred basepairs in length which contain clusters of TF binding sites. Enhancers appear to physically interact with the promoter regions of their target genes concentrating TF and co-activators through looping to activate gene transcription[117, 118]. Most of these enhancer-target promoter interactions occur within a distance of ~50 kb, although many can occur at greater distances up to several megabases[117]. Super enhancers (SEs) are a rare subset of enhancers which are comprised of large clusters of transcriptional enhancers that are associated with expression of genes that define cell identity[117, 119]. Critically super-enhancers appear to be most strongly enriched for autoimmune-disease associated sequence variation[2, 88, 110]. Localised patterns of chromatin modification, such as DNA methylation, histone modification and nucleosome remodelling correlate with enhancer activity, TF binding and initiation or silencing of gene transcription. This has led to successful efforts to catalogue the regulatory elements on a genome-wide scale[120]. Some of the consortia that catalog epigenetic features across different cell types and species are Encyclopedia of DNA Elements (ENCODE)[121], NIH Roadmap Epigenomics Mapping Consortia (REMC)[122], Functional Annotation of Mammalian Genomes 5 (FANTOM 5)[123] and International Human Epigenome Consortium (IHEC)[124].

Identification of these enhancer elements is typically based on the presence of specific histone modifications[7, 88, 110, 117, 119], coactivators[7, 88, 110, 117] (which are recruited to enhancers to deposit histone modifications) and DNaseI hypersensitive sites (DHSs)[125]. Apart from using chromatin immunoprecipitation assays with sequencing (ChIP-seq) to determine enrichment of histone modifications, local distortions in chromatin structure created by TFs binding and positioned nucleosomes have also been exploited by several methods such as Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq)[126, 127], DNAse-seq[125] and more recently Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq)[25] to identify regulatory elements due their altered accessibility to nucleases, transposases or chemicals[128].

It has been well established that linear distance does not accurately predict the interacting gene targets of enhancers as enhancer can bypass neigbouring genes to regulate genes far apart in the linear genome sequence but are proximate in 3D space within the nucleus[129-132]. Enrichment of disease-associated risk SNPs, including T1D-associated variants, at non-coding regions of the genome [2, 88] presents a strong rationale to consider spatial genome organization in identifying target genes impacted by disrupted long-range enhancer-promoter interactions. A 3D contact map is also required to understand the impact of genetic risk and epigenetic changes on gene expression in disease. This conformation-dependant DNA looping mechanisms make it difficult to bioinformatically predict the direct gene targets of enhancers from linear maps. To circumvent this, 3C (chromosome conformation capture)-based technologies such as Hi-C can be used to map distal regulatory elements like enhancers to their respective target genes based on physical connectivity evidence. Hi-C is a technique that captures the conformation of genome by measuring the frequency at which two DNA fragments physically associate in three-dimensional (3D) space. It is a powerful tool used to profile nuclear organization and chromosome architecture. In Hi–C, a biotin-labelled nucleotide is incorporated at the ligation junction, allowing selective purification of chimeric DNA ligation junctions followed by deep sequencing [133]. Hi-C has been performed in a great magnitude of cell types including Tconv and Treg cells [6, 37, 134]. This spatial context will allow us to have a better understanding on the genome regulation and the impact of genetic risk and epigenetic changes on gene expression in disease.

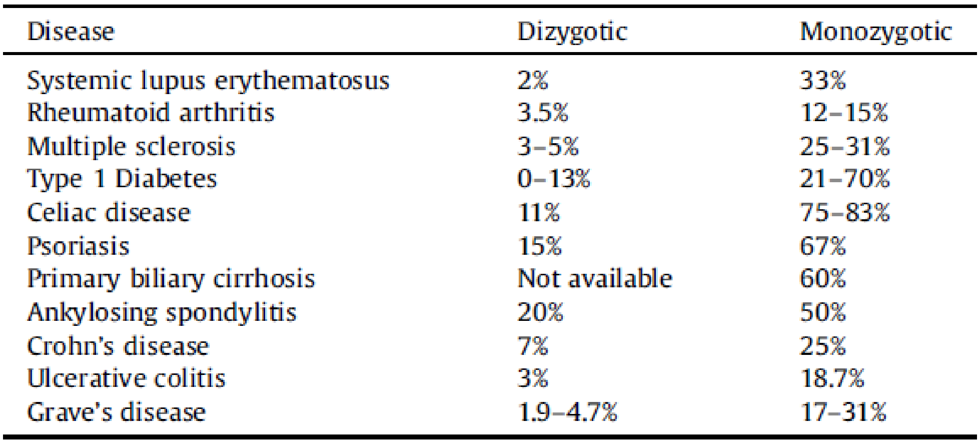
1.5.3 Stimulation induces global remodelling of accessible chromatin

In response to stimulation T cells exhibit a global, large-scale increase in chromatin accessibility [45, 135, 136]. T-cell stimulation-specific regions are enriched for GWAS-eQTLs or SNPs associated with autoimmune diseases such as IBD and RA, as well as enhancers identified from different T subsets, emphasising the importance of profiling cells under stimulation to identify disease-relevant regulatory elements[45, 136]. Probing these stimulation-responsive regions could reveal disease mechanisms previously missed from the steady-state profiling. However, most of the genomic and functional studies have been performed on unstimulated cell types and thus we lack a comprehensive interrogation of the effects of stimulation on the chromatin landscape of different immune cell types and the roles of SNPs in these response-specific regions in autoimmune disease.

1.5.4 Epigenetic variants alter gene regulation in autoimmune diseases

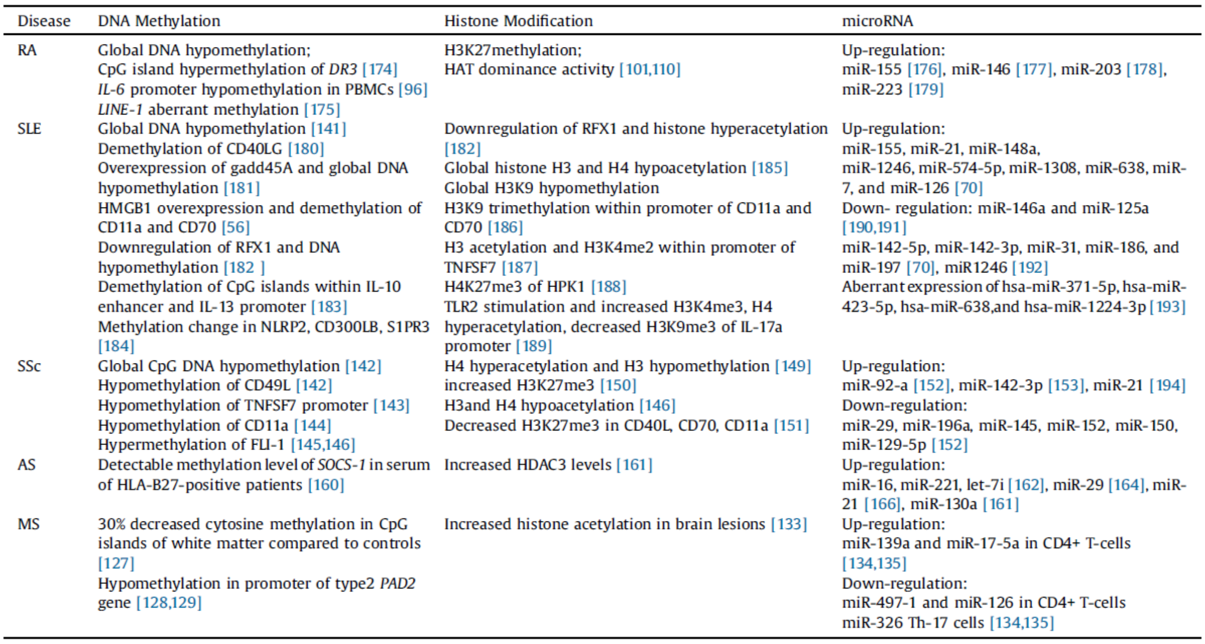
Normal cellular functions require tight control of gene regulation and maintenance of epigenomic homeostasis. There is now growing interest in exploring the roles of non-genetic component including epigenetic factors in complex disease aetiology. The involvement of environmental/epigenetic component in complex autoimmune diseases[137] is supported by the finding that the disease concordance among MZ twins is always not 100% (Table 1-2) and epigenetic alterations were found in various diseases (Table 1-3).

Table 1-2. Autoimmune disease concordance among monozygotic and dizygotic twins (Ahmadi, Gharibi [26]).



Disease	Dizygotic	Monozygotic
Systemic lupus erythematosus	2%	33%
Rheumatoid arthritis	3.5%	12–15%
Multiple sclerosis	3–5%	25–31%
Type 1 Diabetes	0–13%	21–70%
Celiac disease	11%	75–83%
Psoriasis	15%	67%
Primary biliary cirrhosis	Not available	60%
Ankylosing spondylitis	20%	50%
Crohn's disease	7%	25%
Ulcerative colitis	3%	18.7%
Grave's disease	1.9–4.7%	17–31%

Table 1-3. Epigenetic alterations in autoimmune diseases (Ahmadi, Gharibi [26]).
RA; Rheumatoid Arthritis, SLE; Systemic lupus erythematosus, SSc; Systemic sclerosis, MS; Multiple Sclerosis, AS; Ankylosing Spondylitis



Disease	DNA Methylation	Histone Modification	microRNA
RA	Global DNA hypomethylation; CpG island hypermethylation of *DR3* [174] *IL-6* promoter hypomethylation in PBMCs [96] *LINE-1* aberrant methylation [175]	H3K27methylation; HAT dominance activity [101,110]	Up-regulation: miR-155 [176], miR-146 [177], miR-203 [178], miR-223 [179]
SLE	Global DNA hypomethylation [141] Demethylation of CD40LG [180] Overexpression of gadd45A and global DNA hypomethylation [181] HMGB1 overexpression and demethylation of CD11a and CD70 [56] Downregulation of RFX1 and DNA hypomethylation [182] Demethylation of CpG islands within IL-10 enhancer and IL-13 promoter [183] Methylation change in NLRP2, CD300LB, S1PR3 [184]	Downregulation of RFX1 and histone hyperacetylation [182] Global histone H3 and H4 hypoacetylation [185] Global H3K9 hypomethylation H3K9 trimethylation within promoter of CD11a and CD70 [186] H3 acetylation and H3K4me2 within promoter of TNFSF7 [187] H4K27me3 of HPK1 [188] TLR2 stimulation and increased H3K4me3, H4 hyperacetylation, decreased H3K9me3 of IL-17a promoter [189]	Up-regulation: miR-155, miR-21, miR-148a, miR-1246, miR-574-5p, miR-1308, miR-638, miR-7, and miR-126 [70] Down- regulation: miR-146a and miR-125a [190,191] miR-142-5p, miR-142-3p, miR-31, miR-186, and miR-197 [70], miR1246 [192] Aberrant expression of hsa-miR-371-5p, hsa-miR-423-5p, hsa-miR-638,and hsa-miR-1224-3p [193]
SSc	Global CpG DNA hypomethylation [142] Hypomethylation of CD49L [142] Hypomethylation of TNFSF7 promoter [143] Hypomethylation of CD11a [144] Hypermethylation of FLI-1 [145,146]	H4 hyperacetylation and H3 hypomethylation [149] increased H3K27me3 [150] H3and H4 hypoacetylation [146] Decreased H3K27me3 in CD40L, CD70, CD11a [151]	Up-regulation: miR-92-a [152], miR-142-3p [153], miR-21 [194] Down-regulation: miR-29, miR-196a, miR-145, miR-152, miR-150, miR-129-5p [152]
AS	Detectable methylation level of *SOCS-1* in serum of HLA-B27-positive patients [160]	Increased HDAC3 levels [161]	Up-regulation: miR-16, miR-221, let-7i [162], miR-29 [164], miR-21 [166], miR-130a [161]
MS	30% decreased cytosine methylation in CpG islands of white matter compared to controls [127] Hypomethylation in promoter of type2 *PAD2* gene [128,129]	Increased histone acetylation in brain lesions [133]	Up-regulation: miR-139a and miR-17-5a in CD4+ T-cells [134,135] Down-regulation: miR-497-1 and miR-126 in CD4+ T-cells miR-326 Th-17 cells [134,135]

Among the epigenetic mechanisms DNA methylation is the most well-characterized and commonly used epigenetic marker in interrogating human diseases as it confers a more stable transcriptional signal compared with the other epigenetic mechanisms. In addition, human methylome profiles for various cell types are currently available for comparative studies, for instances, DNA methylomes from human peripheral blood mononuclear cells[138] and cardiomyocytes[139]. Epigenome changes in diseases such as cancers, neurodegerative/psychiatric disorders, metabolic disorders, as well as autoimmune and inflammatory diseases (Table 1-3) have been identified to date, supporting the notion that disease-promoting environmental factors exert their effects through changes in the epigenome in disease settings[140].

Epigenetics of T1D. Studies have identified T1D-specific epigenetic changes in discordant twins. For instance, enrichment of differentially variable CpG positions (DVPs) in T1D twins when compared with their healthy co-twins and when compared with healthy, unrelated subjects (in T cells, B cells and monocytes)[102]. Also, global hypomethylation of CpG sites within promoter regions in T1D discordant MZ twins when compared with their healthy co-twins has been reported [100], as well as enrichment of methylation variable positions (MVP) at CpG sites in monocytes between T1D MZ twins and their healthy co-twins[101]. In addition, significant methylation differences in T1D discordant MZ twin pairs were detected especially at GWAS-identified T1D associated genes such as *INS*, *IL-2RB* and *CD226*[99] that are important for immune regulation. These provide evidence of multiple types of T1D-specific epigenetic variance that is linked to disease.

1.6 Genome-wide profiling of DNA accessibility, gene expression and chromatin conformation

Assaying chromatin accessibility as a means to dissect the roles of regulatory genetic variants. Although combining GWAS and fine mapping studies with epigenetic annotation of the genome has refined the list of potential causal SNPs and target enhancers, they have not in most cases identified a causal SNP. On average eight candidate SNPs were discovered at each disease risk locus by GWAS, complicating the characterisation of molecular mechanism(s) causing the defects [2]. They also do not delineate disease causality or the role of non-genetic components such as environmental factors/epigenetic changes in the disease pathogenesis. New methods such as ATAC-seq can be used to capture epigenetics modifications such as chromatin accessibility and TF binding at these regions in an unbiased manner.

As described local distortions in chromatin structure created by TFs binding and positioned nucleosomes have been exploited by several methods such as Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq)[126, 127], DNAse-seq[125], sequencing of micrococcal nuclease sensitive sites (MNase-seq) and more recently Assay for Transposase-Accessible Chromatin with high throughput sequencing (ATAC-seq)[25] (Figure 1-14) to identify regulatory elements due their altered accessibility to nucleases, transposases or chemicals[128].

In general, though the conventional chromatin probing methods such as DNase-seq and FAIRE-seq are powerful, they require at least millions of cells as input requirement, involve complex sample preparations, and are unable to infer the interplay of nucleosome positioning, chromatin accessibility and TF binding simultaneously[25]. For these reasons ATAC-seq has become the method of choice for accessibility mapping using small numbers of input cells.

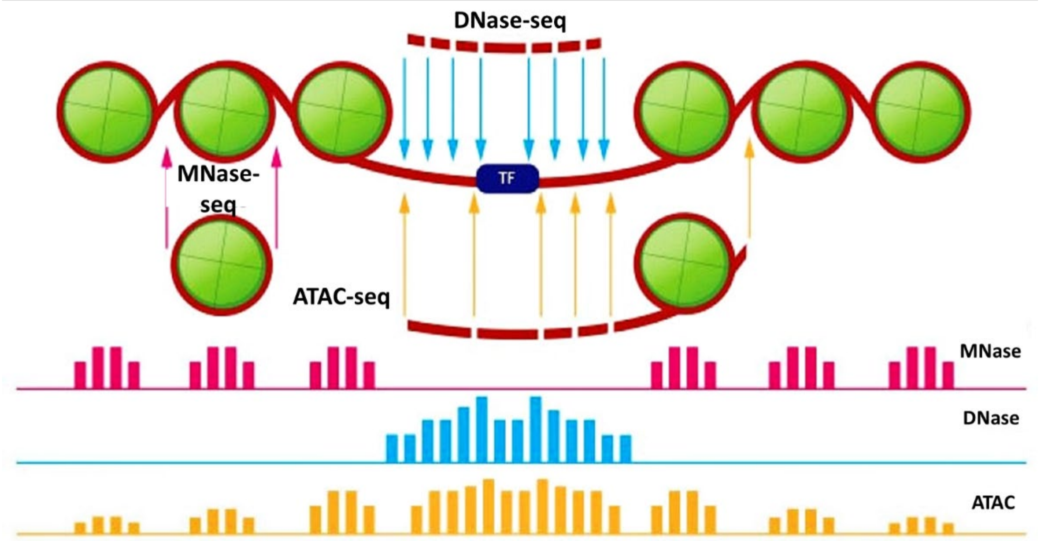


Figure 1-14 | Genome-wide methods used to probe chromatin accessibility (Adapted from Chi [31]).

ATAC-seq (Assay of Transposase Accessible Chromatin with high throughput sequencing) probes chromatin accessibility, TF occupancy and nucleosome positioning with low starting material. In ATAC-seq, Tn5 transposase, preloaded with sequencing adaptors, simultaneously cuts and ligates sequencing adaptors at accessible open chromatin regions (Figure 1-15) [25, 141]. The captured fragments between the adapters can then be PCR-amplified and subjected to high throughput next generation sequencing[25, 141]. As compared to other genome-wide chromatin probing methods, ATAC-seq is relatively simple and rapid (Figure 1-16). It is also highly sensitive, with 50,000 to as little as 5000 cells, producing comparable signals as DNase-seq data generated from 1 – 50 million cells (Figure 1-17). ATAC-seq also has a higher signal-to-noise ratio compared to FAIRE-seq, whose data was generated from 3–5 orders of magnitude higher cell numbers than ATAC-seq[25]. The requirement of low input material makes ATAC-seq amenable to use on rare population and small clinical samples such as bio-banked material[30].

The statistical power of ATAC-seq is well reflected in one study that compared the statistical sensitivity of ATAC-seq and microarray in detecting the differential regulation of X chromosome inactivation (XCI) escapee genes between males and females[22]. In comparison to microarray analysis ATAC-seq required only 11 samples of each gender to detect a 2-fold difference in known XCI and XCI escapee genes, while mRNA microarray required 81 samples of each gender to reach a power of 0.95 (Figure 1-19), suggesting that ATAC-seq is over seven times more sensitive than microarray.

At a high-sequencing depth ATAC-seq can also be used to identify TF binding sites at single basepair resolution as TF-occupied sites prevent Tn5 cleavage and adaptor insertions, leading to protected regions (known as "footprints") in the sequencing reads[21, 25, 142, 143]. Furthermore, the sequence of these footprints can be compared to reference genomes in the case of disease samples. As shown in Figure 1-18, ATAC-seq gives higher resolution on inferring TF occupancy sites as compared to DNase-seq and ChIP-seq. ATAC-seq can also be used to infer nucleosome positioning based on insert size distribution of sequenced fragments[25].

Since it was introduced ATAC-seq has been performed on various cell types including primary CD4$^+$ T cells[22], CD8$^+$ T cells[144], naïve B cells (SLE samples)[30], islet cells (T2D samples)[145], brain cells[146], retinal cells[147], induced pluripotent stem cells (iPSCs)[148], lymphoblastoid cells[25], mouse embryonic stem cells[149, 150] and fibroblast (MEF) cells[151], suggesting its general applicability, feasibility and sensitivity in inferring chromatin accessibility and TF occupancy not just in distinct cell types, but under different cell states (for instances, healthy vs. disease; activated vs. non-activated).

ATAC-seq data from CD4$^+$ T cells from healthy donors revealed the characterized degree of variability in chromatin accessibility between different donors and demonstrated that these variable regions were enriched for autoimmune-disease associated SNPs and mapped to epigenetically modified regions, suggesting that disease-associated variants at these regions may further amplify this variation until a pathogenic threshold is crossed [22]. Importantly, ATAC-seq has not been performed on immune cells from T1D cohorts to identify changes associated specifically with T1D.

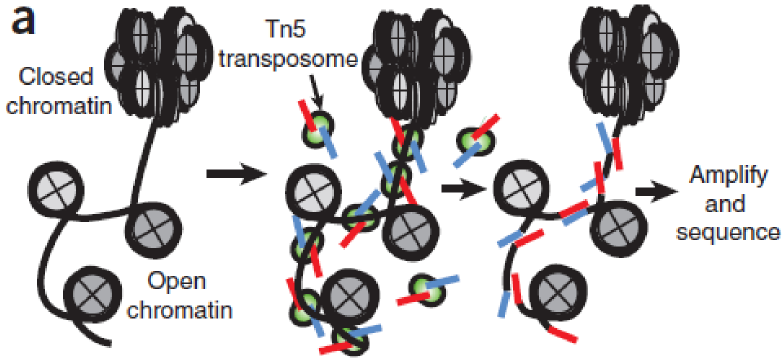


Figure 1-15 | ATAC-seq protocol (Buenrostro, Giresi [25]).
Isolated nuclei are incubated with a hyperactive mutant of Tn5 transposase preloaded with illumina sequencing adapters (red & blue lines). The transposase preferentially cleaves open chromatin and inserts the adapters at the end of each fragment. These serve as priming sites for sequencing library construction.

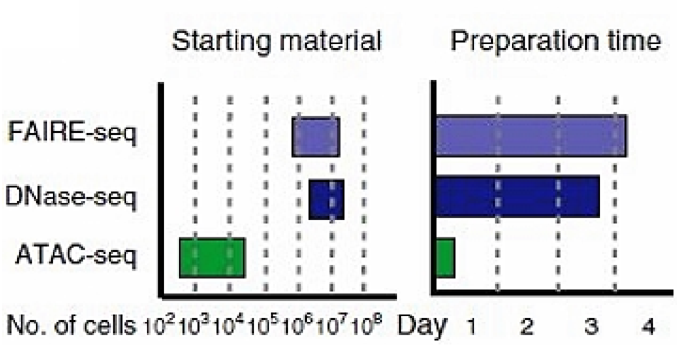


Figure 1-16 | Approximate input material and sample preparation time requirements for genome-wide methods of open-chromatin analysis (Buenrostro, Giresi [25]).

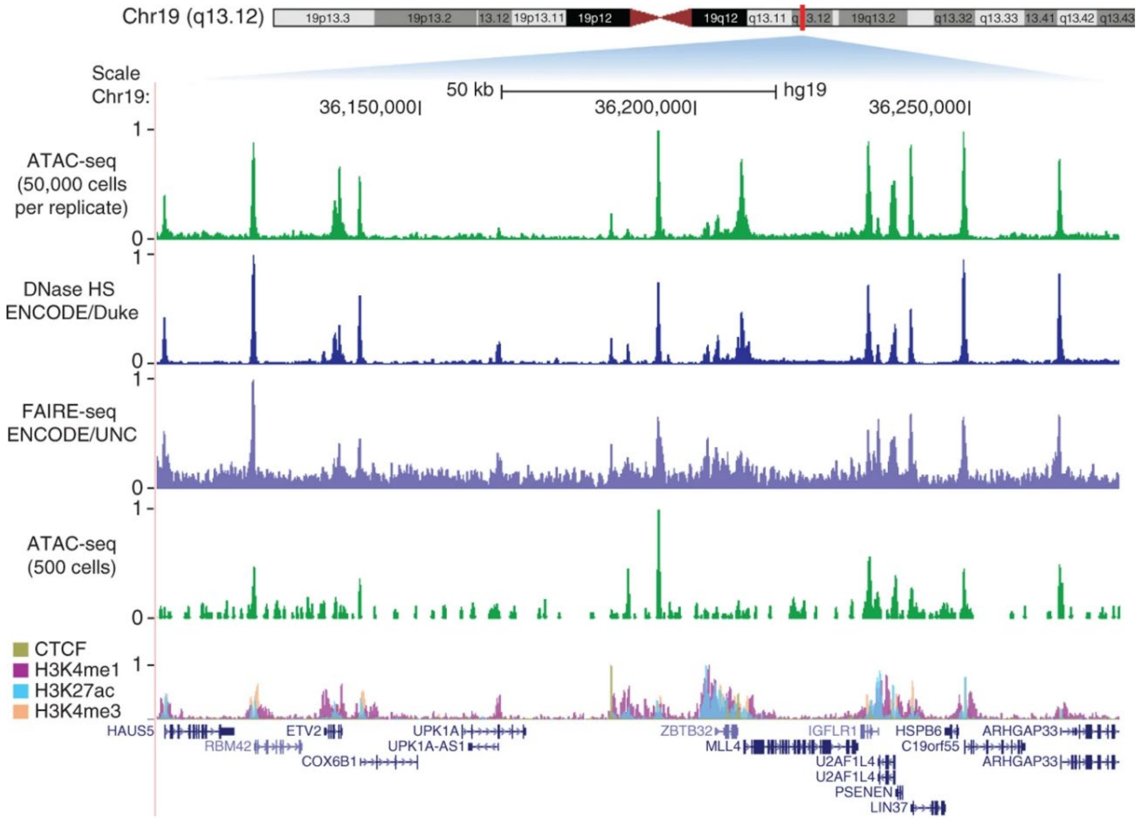


Figure 1-17 | Comparison of ATAC-seq to DNase-seq and FAIRE-seq at a locus in GM12878 lymphoblastoid cells (Adapted from Buenrostro, Giresi [25]).

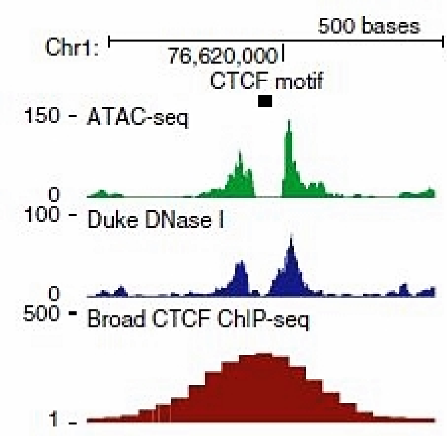


Figure 1-18 | ATAC-seq reports higher resolution in TF CTCF binding signal (Buenrostro, Giresi [25]).

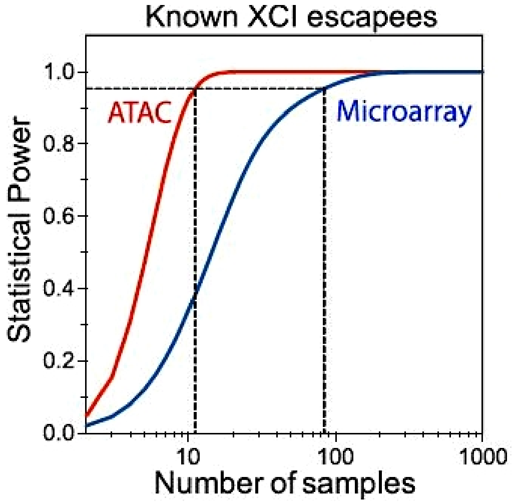


Figure 1-19 | ATAC-seq requires smaller sample size to achieve excellent statistical power compared to microarray (Qu, Zaba [22]).
Genes that escape X inactivation in females were used to compare the accuracy and statistical power of ATAC-seq data versus standard microarray data. ATAC-seq requires 11 samples of each gender, while microarray requires 81 samples of each gender to reach a power of 0.95, in detecting 2-fold difference in gene activity between males and females.

1.7 PhD Project Framework

1.7.1 Rationale for the research project

TFs are often involved in recruiting histone modification enzymes to regulate gene expression. In T1D, causal SNPs may disrupt or create TF binding sites in enhancers located in genetic susceptibility regions, resulting in alterations in chromatin structure such as DNA accessibility arising from histone modifications. This can subsequently lead to changes in gene expression patterns in CD4$^+$ T cells from T1D individuals. Current methods to identify these changes and their molecular consequences are complicated, time consuming and expensive, as generally they can only examine one TF/binding site at a time, involve TF binding site prediction which relies on TF binding site consensus motifs being correct (which currently results in a high degree of false positives/negatives due to the high redundancy of the TF binding motif information) and require large starting material (at least millions of cells) making them challenging for application on limited clinical samples. ATAC-seq (Assay of Transposase Accessible Chromatin with high throughput sequencing) is a new genome-wide method that can simultaneously probe chromatin structure (accessibility and nucleosome positioning) and importantly TF binding at high resolution **[25]**. Importantly, it is highly sensitive requiring as little as 500 cells making it amenable to limited clinical samples. ATAC-seq has not been performed on immune cells from human T1D cohorts.

1.7.2 Unanswered questions in T1D molecular mechanisms

Evidence suggests both genetic and environmental factors contribute to the islet autoimmunity and progression to T1D.

- How do the genetic and environmental factors combine to cause the immune system to attack pancreatic β cells?
- How do the genetic and environmental factors combine to cause the failure of the immune tolerance to prevent this?

A likely route for environmental factors to modify the immune system is via epigenetic mechanisms such as histone modification, DNA methylation and altered miRNA expression and alteration in all these have been reported in T1D patient.

- What epigenetic changes are involved in human T1D?
- Where do these epigenetic changes occur in the genome?
- What are the consequences of these epigenetic alterations on gene expression in Thelper/Tconv (Conventional T) and Treg cells?

Addressing these questions in an unbiased cell specific systematic approach requires the genome-wide identification and functional annotation of genetic and epigenetic changes that contribute to Type 1 Diabetes (T1D).

1. In T1D, how do genetic variants act to mediate disease susceptibility and contribute to the loss of tolerance to pancreatic β cells in susceptible individuals?

2. Do disease-associated genetic variants mediate disease susceptibility by altering the transcriptional program of Treg or Tconv cells?

1.7.3 Hypothesis

Genetic variation and epigenetic changes linked to T1D modify the activity of Treg- and Tconv-specific enhancers, leading to altered gene, resulting in a breakdown in immune homeostasis and an aberrant immune response.

1.7.4 Aims of the project

This project aims to compare the DNA accessibility, transcription factor (TF) binding profiles and gene expression in Treg and Tconv cells isolated from individuals with established T1D and age- and sex- matched healthy controls by using ATAC-seq and RNA-seq. By generating a comprehensive genome-wide protein binding profile for T1D cohorts and healthy controls, this project will identify alterations in TF occupied sites between T1D and healthy individuals in Treg and Tconv on a genome-wide scale, in addition to alterations in DNA accessibility and gene expression. In addition, it will enable annotation of DNA accessibility of T1D susceptibility regions. T1D associated SNPs at these regions can then be correlated to alterations in TF binding.

Aim 1 To establish a protocol for performing ATAC-seq on frozen T cells obtained from healthy adult human donors.

Aim 2 To establish and validate bioinformatics pipeline tools for data analysis.

Aim 2.1 To establish and validate differential chromatin accessibility (DA) analysis of ATAC-seq data.

Aim 2.2 To validate differential gene expression (DGE) analysis of RNA-seq.

Aim 2.3 To benchmark peak calling for ATAC-seq data.

Aim 3 To perform comparative genomics and map the genetic risk in a case-control human cohort.

Aim 3.1 To perform ATAC-seq to compare chromatin accessibility and TF occupancy profiles in Treg and Tconv cells from T1D and healthy cohort.

Aim 3.2 To perform RNA-seq to compare gene expression in Treg and Tconv cells from T1D and healthy cohort.

Aim 3.3 To identify T1D-associated targets by integrating epigenome, GWAS SNPs, TF ChIP data and TF binding site predictions with the generated ATAC-seq chromatin accessibility maps.

1.7.5 Significance/Contribution to the discipline

Currently there is no cure for T1D, with patients requiring careful management of blood glucose levels by lifelong insulin administration, and this is associated with enormous human and economic costs. The molecular mechanisms implicated in T1D are unknown, but evidence points to the contribution of both genetic and environmental factors. Although studies have proposed various potential environmental factors associated with T1D, how each of these factors causes disease is not well characterised. Emerging studies suggest that a likely route for environmental factors to alter the immune system is via epigenetic mechanisms.

Multiple cellular and molecular components of tolerance and effector responses are implicated in the pathogenesis of T1D. Some of the major immune subsets contributing to the onset of T1D include macrophages, dendritic cells, natural killer cells, B cells, $CD8^+$ and $CD4^+$ T cells [152]. The pathogenesis of T1D can be driven by defects in the Treg-mediated immune suppression (numbers, survival and/or function) and/or resistance to Treg control in those effector subsets [105].

This PhD project will identify differentially regulated genomic regions in Tconv and Treg cells from human T1D cohort and healthy controls in the context of chromatin accessibility, by using a genome-wide open chromatin assay known as ATAC-seq which requires much fewer starting material than other chromatin analysis methods. The identified differentially accessible gene loci in T1D can then be correlated to corresponding gene expression changes. By generating a comprehensive genome wide protein binding profile, this project will identify variable TF occupied sites between T1D and healthy individuals at T1D susceptibility regions and identify potential regions altered by epigenetic changes. This will allow us to understand how genetic risk and epigenetic changes disrupt gene regulation in immune cells and Treg function in T1D. This will generate a roadmap we can use to define how immune tolerance is disrupted in autoimmune disease and provide avenues for immunomodulatory interventions and therapies.

In addition, this will provide strong preliminary data to extend ATAC-seq analysis to longitudinal human T1D cohorts of sufficient power to correlate changes during progression from islet autoantibody seroconversion to T1D. This will allow us to understand the molecular events linked to T1D disease progression which is important for the development of better T1D prognostic tools and therapies.

CHAPTER 2: MATERIALS & METHODS

2.1 Cell Isolation and Antigens Staining

2.1.1 Isolation of peripheral blood mononuclear cells (PBMCs)

PBMCs were isolated from either fresh whole blood donated by healthy adult volunteers (with human ethics approval: REC1596/08/2019) or fresh adult human buffy coats obtained from the Australian Red Cross Blood Service by density gradient centrifugation. For PBMC isolation from whole blood samples, blood samples (~10-20mL in lithium heparin tubes) were made up to 35mL with Phosphate Buffered Saline (GE Healthcare Life Sciences, Cat. #SH30028.02) and layered on 15mL of Ficoll-Paque^{TM} PLUS (Cytiva, Cat. # 17144003) in a 50-mL Falcon tube, followed by centrifugation at 800 xg for 20 minutes in a swinging-bucket rotor without brake. The mononuclear cell layer at the interphase was transferred to a 50-mL tube and topped off to a final volume of 50mL with 2% FBS supplemented PBS. PBMCs were washed twice by centrifugation at 500 xg at RT for 10 min and rested overnight in complete X-VIVO 15 culture media (X-VIVO 15 Serum-free media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5% heat inactivated human serum) at 37°C in a humidified 5% CO_2 incubator.

For PBMCs isolation from adult human buffy coats, all buffy coats (~50mL per donor) were protected from light at room temperature under gentle agitation on a rocking platform and processed within 24 hours. Briefly, buffy coat was made up to 140mL with Phosphate Buffered Saline (GE Healthcare Life Sciences, Cat. #SH30028.02) supplemented with 2% FBS (Gibco™, Cat. #10099) and approximately 35mL was layered on top of 15mL of Ficoll-Paque^{TM} PLUS in a 50-mL Falcon tube, followed by centrifugation at 800 xg for 20 minutes in a swinging-bucket rotor without brake. The mononuclear cell layer at the interphase was transferred to a 50-mL tube and made up to a final volume of 50mL with 2% FBS supplemented PBS. PBMCs were washed three times by centrifugation at 500 xg at RT for 10 min prior to incubation overnight in complete X-VIVO 15 culture media (X-VIVO 15 Serum-free media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5% heat inactivated human serum) at 37°C in a humidified 5% CO_2 incubator.

2.1.2 Cryopreservation of PBMCs

Washed PBMCs pellets were resuspended in 1mL of freezing medium (heat-inactivated FCS containing 10% DMSO) at a concentration of 1×10^7 cells/mL, transferred to a 2-mL cryovial and placed into a Mr. Frosty™ Freezing Container. The unit was placed in the -80°C freezer overnight prior to transfer to liquid nitrogen for long-term storage.

2.1.3 Thawing of PBMCs

During the optimisation of the thawing protocol in Chapter 3 a thawing protocol was developed. Briefly, cryopreserved vials of PBMC samples were removed from the liquid nitrogen tank and thawed in a 37°C water bath for 10 minutes. Thawed PBMCs were transferred to a 10mL conical tube and diluted with pre-warmed complete X-VIVO 15 culture media (X-VIVO 15 Serum-free media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5% heat inactivated human serum, 200U/mL DNase [Worthington cat# LS002007]) at a rate of 1 mL/5 seconds. The conical tube containing the cell suspension was gently inverted to mix and spun at 500g for 10 minutes at RT. Cells were washed twice in complete X-VIVO 15 media. The supernatant was aspirated using a glass pasteur pipette and the cells were resuspended in pre-warmed complete X-VIVO 15 culture media at 3-4×10^6 cells/mL. Upon thawing the PBMCs were allowed to recover overnight in the complete X-VIVO 15 culture media in a 24-well culture plate at 37 °C in a CO2 incubator for 16-17 hours prior to T cell sorting by FACS.

2.1.4 Cell surface antigen staining and FACS sorting of Tconv and Treg cells

Upon overnight resting in culture isolated PBMCs were washed once by centrifugation at 500 xg at RT for 10 minutes in PBS supplemented with 2% FCS prior to cell staining. Cells were labelled with the following fluorochrome conjugated anti-human monoclonal antibodies: anti-

CD4, anti-CD25, anti-CD127 and viability dye for FACS analysis (Table 2-1). Cell sorting was performed using a BD FACSAria Fusion flow cytometer. Viable bulk Treg cells were identified as Horizon Fixable Viability Stain 700 negative, CD4^{+} CD25^{hi} CD127^{lo} population, whereas viable bulk Tconv cells were sorted from Horizon Fixable Viability Stain 700 negative, CD4^{+} CD25^{lo} CD127^{hi} population. A fraction of the isolated populations was analysed to assess for post-sort cell purity by expression of cell surface markers as well as intracellular FOXP3 staining by flow cytometry.

Nuclear protein such as FOXP3 requires a cell permeabilisation and fixation step to allow the fluorochrome-conjugated antibodies to enter the nucleus of the cell. Transcription Factor Staining Buffer Set from eBioscience™ (Cat# 00-5523-00, Thermo Fisher Scientific Inc.) was used for staining nuclear protein FOXP3 to confirm the purity of sorted Treg (positive) and Tconv (negative) populations. Prior to cell labelling, solutions were diluted as necessary to a 1x working concentration according to the manufacturer's instructions. Cells were fixed with 1mL of Fixation/Permeabilization working solution (1:1 Fixation:Permeabilization) and pulse-vortexed before incubating for 60 minutes at 4°C protected from light. The cells were then washed twice in 2mL of 1X Permeabilization Buffer and pelleted by centrifugation at 500xg for 5 minutes at RT. Cells were incubated with fluorochrome-conjugated FOXP3 antibody (Table 2-1) and incubated for 30 minutes at RT in the dark, then washed twice again in 2mL of 1X Permeabilization Buffer, followed by 1mL of PBS and pelleted by centrifugation at 500 xg for 5 minutes at RT. Cells were resuspended in 200µL of the PBS and analysed by flow cytometry.

Table 2-1. List of antibodies used in Flow Cytometry

Antibody	Fluorescence Channel	Clone	Isotype	Manufacturer	Location
CD4	BUV395	SK3	Mouse IgG1, κ	BD Biosciences	Cell surface
CD25	BV421	M-A251	Mouse IgG1, κ	BD Biosciences	Cell surface
CD127	PE-CF594	HIL-7R-M21	Mouse IgG1, κ	BD Biosciences	Cell surface
Viability	Alexa Fluor® 700	-	-	BD Biosciences	Cell surface
FOXP3	Alexa Fluor® 647	-	Mouse IgG1	BD Biosciences	Intracellular (nuclear protein)

2.1.5 T cell culture and stimulation

Following FACS isolation, cells were plated at 100,000 cells per well in a 96-well U-bottom plate and maintained in complete X-VIVO 15 culture media (X-VIVO 15 Serum-free media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5% heat inactivated human serum) in the presence of 500U/mL recombinant human IL-2 for 2 hours at 37ºC in a humidified 5% CO_2 incubator prior to cell preparation for ATAC-seq and RNA-seq experiments. For T cell activation, following 2-hour post-sort recovery, cells were stimulated with beads conjugated with anti-CD3 and anti-CD28 antibodies (Dynabeads Human T-Expander CD3/CD28, Gibco no. 11141D, Life Technologies) in complete X-VIVO 15 culture in the presence of 500U/mL recombinant human IL-2 at a cell/bead ratio of 1:1 for 48 hours. After 48 hours Dynabeads were removed from culture medium by magnetic separation as per the manufacturer's protocol. Non-stimulated samples were processed immediately following the 2-hour post-sort recovery for ATAC-seq and RNA-seq experiment.

2.2 Omni ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing)

2.2.1 Cell preparation and transposition

Omni ATAC-seq experiment was performed as described previously [153]. Briefly, post cell sorting, cells were pre-treated with 200U/mL DNase (Worthington cat# LS002007) for 30 minutes at 37°C prior to ATAC-seq experiment. Cell counts were performed manually using a hemocytometer and 50,000 T cells were lysed in 50µL of cold resuspension buffer (RSB: 10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM $MgCl_2$ in water) containing 0.1% NP40, 0.1% Tween-20, and 0.01% digitonin on ice for 3 minutes. Immediately after lysis, the reaction was washed with 1mL of ATAC-seq RSB containing 0.1% Tween-20 by centrifugation at 500 xg for 10 minutes at 4°C. The nuclei were resuspended in 50µL of transposition mix comprising of 25µL 2× TD buffer, 2.5µL Tn5 transposase (Illumina Inc), 16.5µL PBS, 0.5µL 1% digitonin, 0.5µL 10% Tween-20 and 5 µL water, unless otherwise stated. The transposition reaction was incubated at 37°C for 30 minutes in a thermomixer with mixing (1000 rpm), unless otherwise stated. The DNA was purified using a Zymo DNA Clean & Concentrator-5 (cat# D4014) kit as per manufacturer's instructions, eluted in 21µL of elution buffer and pre-amplified for 5 cycles using NEBNext 2x Master mix (New England Biolabs, cat# M0541S) using the reaction setup and thermal cycling conditions shown in Table 2-2 and Table 2-3, respectively. The oligo design used in PCR of ATAC-seq protocol was adopted from Buenrostro, Giresi [25] study (Appendix; Table 7-6).

Table 2-2. PCR reaction setup for ATAC-seq.

Reagent	Volume
25 uM Primer Ad1	2.5µL
25 uM Primer Ad2 (barcoded)	2.5µL
2x NEBNext Master Mix	25µL
Transposed sample	20µL

Table 2-3. PCR cycling conditions for pre-amplification of ATAC-seq reaction.

Temperature	Time
72°C	5 minutes
98°C	30 seconds
5 cycles of:	
98°C	10 seconds
63°C	30 seconds
72°C	1 minute
Hold at 4°C	

2.2.2 Determination of appropriate additional amplification steps by qPCR

To determine the number of additional cycles needed for optimal amplification qPCR was performed using 5µL of pre-amplified reaction using the reaction setup and thermal cycling conditions shown in Table 2-4 and Table 2-5, respectively. The rest of the pre-amplification reaction was placed on ice during the qPCR experiment.

Table 2-4. qPCR reaction setup for pre-amplified ATAC-seq reaction.

Reagent	Volume
Sterile water	3.76

25 uM Primer Ad1	0.5µL
25 uM Primer Ad2 (barcoded)	0.5µL
20x EvaGreen® Dye (in water) (cat# 31000)	0.24µL
2x NEBNext Master Mix	5µL
Pre-amplified sample	5µL

Table 2-5. qPCR cycling conditions for pre-amplified ATAC-seq reaction.

Temperature	Time
98°C	30 seconds
20 cycles of:	
98°C	10 seconds
63°C	30 seconds
72°C	1 minute
Hold at 4°C	

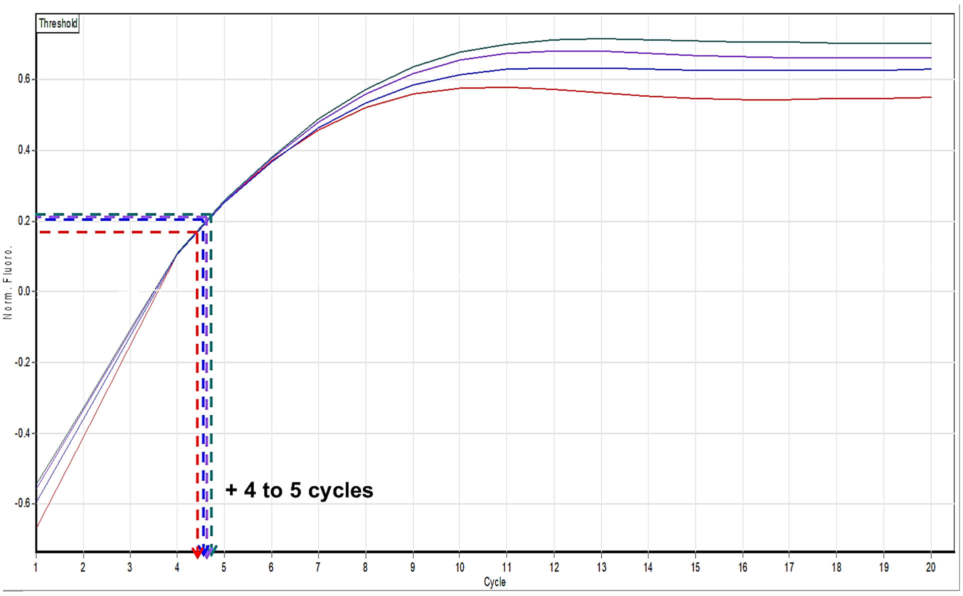


Figure 2-1 | Representative amplification plot demonstrating the correct number of additional cycles to perform for four ATAC-seq libraries.
The additional number of cycles needed is determined from the cycle number that corresponds to one-third of the maximum fluorescent intensity.

After qPCR amplification, the number of required additional PCR cycles was determined from the PCR amplification profiles (Figure 2-1) as described [141]. Briefly, one-third of the maximum fluorescence intensity was used as the threshold to calculate the number of cycles required. The remainder of the pre-amplified reaction was then amplified using the calculated additional number of cycles without addition of any further reagents. Final PCR reaction was purified using Zymo DNA Clean & Concentrator-5 (cat# D4014) kit and eluted in 20µL of elution buffer.

2.2.3 Library size selection, quality assessment and sequencing

The quality and size distribution of the constructed ATAC-seq libraries were assessed on an Experion™ Automated Electrophoresis System (Bio-Rad Laboratories, Inc., California, United States) using the Experion DNA 1K Analysis Kit (Bio-Rad Laboratories, Inc., California, United States, Cat# 7007107) according to manufacturer's protocol (Figure 2-3). Data acquisition and analysis were performed using the Experion software. Double size selection was carried out to enrich for a fragment size window of 100 to 800bp prior to sequencing as described in Buenrostro, Wu [141] (Figure 2-2; Figure 2-3) using SPRIselect (Beckman Coulter, California, United States; cat# B23317). All buffers and reagents used were provided in the kits and procedures were performed according to the manufacturer's protocols.

Amplified, barcoded ATAC-seq libraries (Table 2-4, Table 7-6) were quantified by qPCR using a KAPA Library Quantification Kit (Illumina® Platforms; Cat# KK4824) and pooled in equimolar concentrations. Pooled libraries were sequenced on an Illumina NextSeq 550 High-Output paired-end 75-cycle platform (optimisation experiments) or Illumina Hiseq X Ten paired-end 150-cycle platform (T1D case-control experiments). Demultiplexing of samples was performed the sequencing facility using Illumina bcl2fastq (ver. 2.17) according to index sequences.

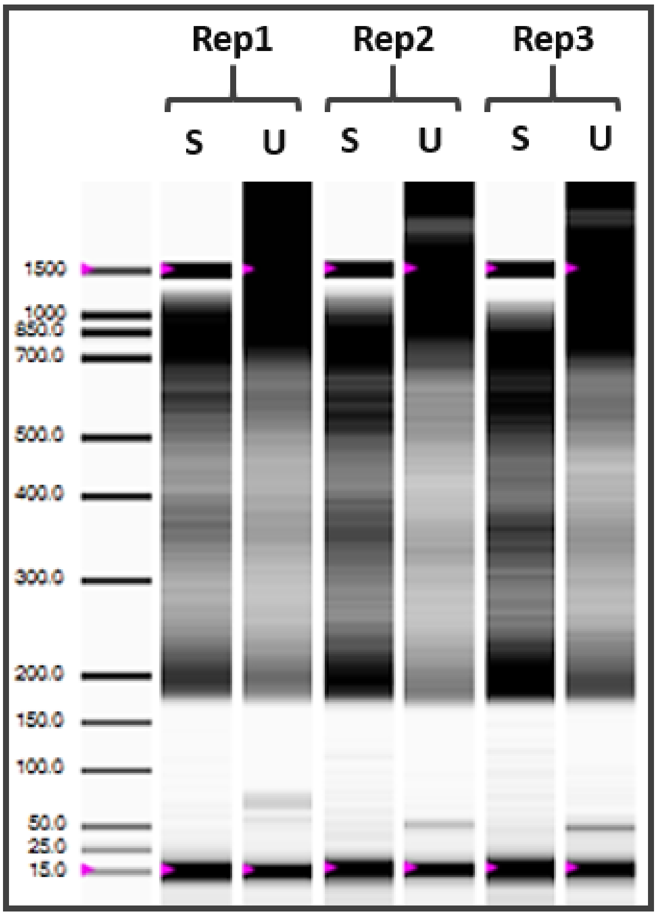


Figure 2-2 | Representative virtual gel images showing fragment size distribution of three amplified ATAC-seq libraries before and after size selection as determined by the Experion™ Automated Electrophoresis System. ***S, selected; U, unselected.***

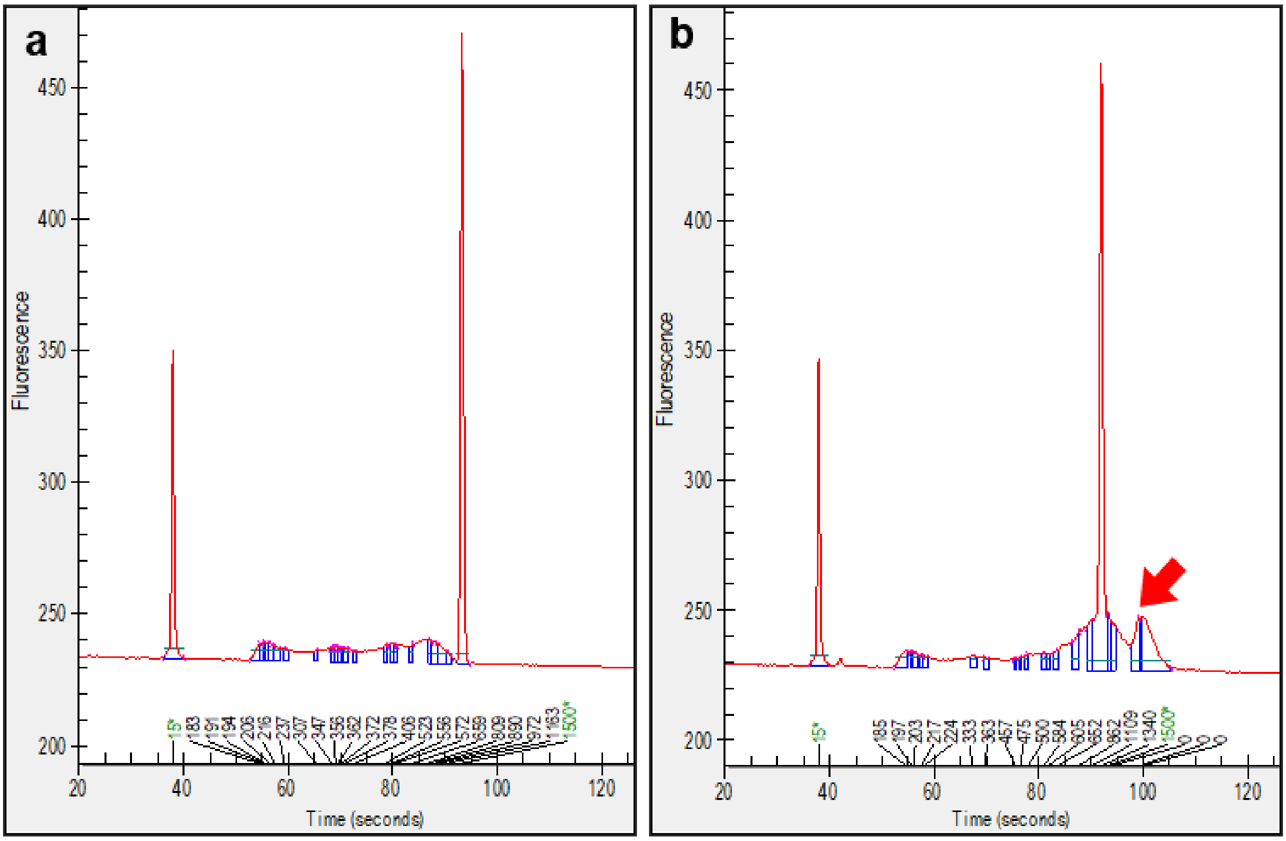


Figure 2-3 | Representative electropherograms showing the fragment size distribution of an amplified ATAC-seq before (a) and after (b) size selection.
Red arrow indicates elimination of long fragments after size selection.

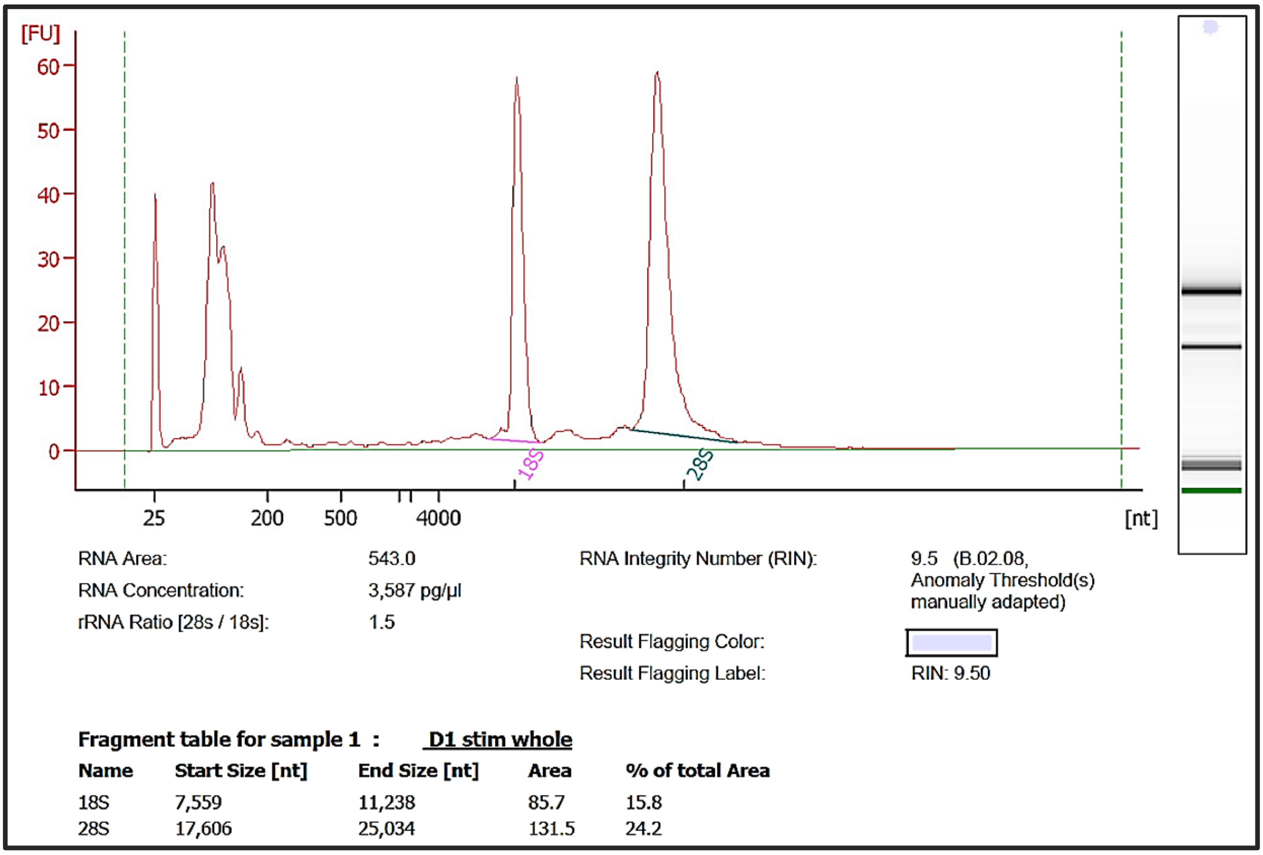
2.3 RNA-seq

2.3.1 RNA isolation and quality assessment

Depending on the nature of the experiments, total RNA was isolated from either whole cells or supernatant fractions collected from Omni ATAC-seq lysis reaction (ATAC-SN) using a miRNeasy Micro Kit (Qiagen; cat# 217084). Extraction and purification procedures were performed according to the manufacturer's protocol. Briefly, samples were homogenised in QIAzol Lysis Reagent (for whole cells experiment) (Qiagen, Cat# 79306) or TRIzol LS reagent (for ATAC-SN experiment) (Invitrogen, Cat# 10296010) was purified with a RNeasy MinElute spin column according to the manufacturer's protocol except that the RNA was eluted twice from the RNeasy MinElute spin column in a total volume 14µL of RNase-free water to ensure maximum yield recovery.

For RNA isolation from the supernatant fractions from Omni ATAC-seq lysis reaction (ATAC-SN), each ATAC-SN sample (~1mL) was divided into five 1.5-mL Eppendorf tubes (each containing ~200µL of supernatant) and 3 volumes of TRIzol™ LS Reagent (Invitrogen, 10296010) were added to each tube (for standard RNA isolation protocol, cells were pelleted and homogenised in 700µL of QIAzol Lysis Reagent). Following phase separation of ATAC-SN homogenate, the aqueous phase from each sample was mixed with 1.5 volumes of 100% ethanol and pooled, followed by RNA purification with a single RNeasy MinElute spin column per sample according to the manufacturer's protocol. The RNA was eluted twice from the RNeasy MinElute spin column in a total volume 14µL of RNase-free water.

The quantification of purified RNA (1µL) was performed with Qubit™ 3 Fluorometer (Invitrogen™, cat# Q33216) using the Qubit™ RNA HS Assay Kit (Invitrogen™, cat# Q32852) according to the manufacturer’s protocol.



RNA Area:	543.0	RNA Integrity Number (RIN):	9.5 (B.02.08, Anomaly Threshold(s) manually adapted)
RNA Concentration:	3,587 pg/µl		
rRNA Ratio [28s / 18s]:	1.5		
		Result Flagging Color:	🟦
		Result Flagging Label:	RIN: 9.50

Fragment table for sample 1 : D1 stim whole

Name	Start Size [nt]	End Size [nt]	Area	% of total Area
18S	7,559	11,238	85.7	15.8
28S	17,606	25,034	131.5	24.2

Figure 2-4 | Representative electropherogram showing the RNA concentration, the ribosomal ratio and the RNA Integrity Number (RIN) of an RNA sample determined by the Agilent Bioanalyzer.

The integrity and quality of the input RNA were assessed on an Agilent Bioanalyzer® using the Agilent RNA 6000 Pico Kit (Figure 2-4). In RNA-seq library construction highly abundant ribosomal RNAs (rRNA) are removed from total RNA either by Poly(A) mRNA enrichment or rRNA depletion (negative selection). Owing to more efficient use of sequencing depth, higher exonic coverage as well as higher accuracy in gene quantification [154], Poly(A) mRNA enrichment was the choice of rRNA removal method. The PolyA mRNA enrichment protocol requires high quality RNA with a RIN (RNA integrity number) score of above 7 (Figure 2-4). For a small number of samples with partially degraded RNA (RIN = 2 to 7) rRNA depletion protocol was used. The RNA integrity number (RIN) is an algorithm used to assign integrity values to RNA measurements [155] based on the ratio of 28S:18S ribosomal RNA. I obtained RIN score ranged from 8.7 to 9.9 for whole cell RNA and 4.7 to 9.3 for ATAC-SN samples.

2.3.2 Enrichment of Poly(A)$^+$ mRNA and RNA-seq library generation

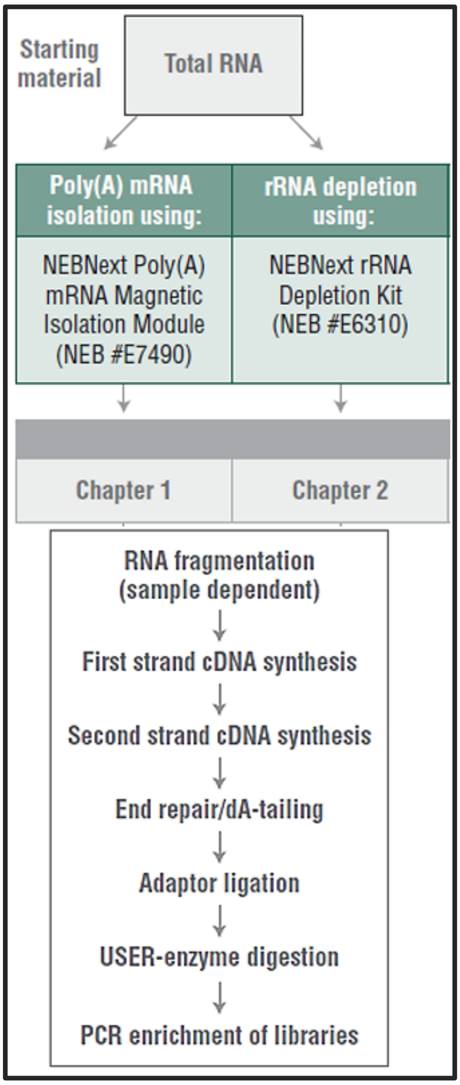

Starting material
Total RNA
Poly(A) mRNA isolation using:
rRNA depletion using:
NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB #E7490)
NEBNext rRNA Depletion Kit (NEB #E6310)
Chapter 1
Chapter 2
RNA fragmentation (sample dependent)
First strand cDNA synthesis
Second strand cDNA synthesis
End repair/dA-tailing
Adaptor ligation
USER-enzyme digestion
PCR enrichment of libraries

Figure 2-5 | Overview of RNA-seq library generation using the NEBNext Ultra II Directional RNA Protocol (Instruction manual from NEB; cat# E7760).

For RNA-seq library construction highly abundant ribosomal RNAs (rRNA) were removed from total RNA either by Poly(A) mRNA enrichment or rRNA depletion (negative selection) depending upon the RIN number of the sample. Poly(A) mRNA enrichment was selected as the method of choice for rRNA removal method for RNA samples with a RIN score of above 7 (Figure 2-5), because it results in more efficient removal of non-mRNA species resulting in higher exonic coverage and accuracy in gene quantification [154]. For a small number of samples with a lower RIN score (RIN = 4 to 7 in T1D case-control RNA-seq experiment;

Chapter 5) a rRNA depletion protocol was used using NEBNext® rRNA Depletion Kit (New England Biolabs; cat# E6310).

The RNA samples were enriched for Poly(A) RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs; cat# E7490) prior to generation of cDNA libraries from the Poly (A)- enriched RNA using a NEBNext Ultra Directional II RNA Library Preparation Kit for Illumina (New England Biolabs; cat# E7760). The Poly (A)- enriched mRNA was subjected to a fragmentation incubation time of 15 minutes to yield an RNA insert size of ~ 200bp, followed by cDNA synthesis, end repair, and adaptor ligation. Individual RNA-seq libraries (n= 8 for whole cell vs ATAC-SN RNA-seq in Chapter 3; n=47 for T1D case-control RNA-seq in Chapter 5) were barcoded using the NEBNext Multiplex Oligos for Illumina Kit (#E6609). The RNA-seq libraries were purified using the SPRIselect (Beckman Coulter, California, United States; cat# B23317) to remove primers at ~80bp and adaptor-dimers at ~128bp and to enrich for a narrow distribution of fragments with an average peak size of approximately 300 bp.

2.3.3 Library quality assessment, quantification and sequencing

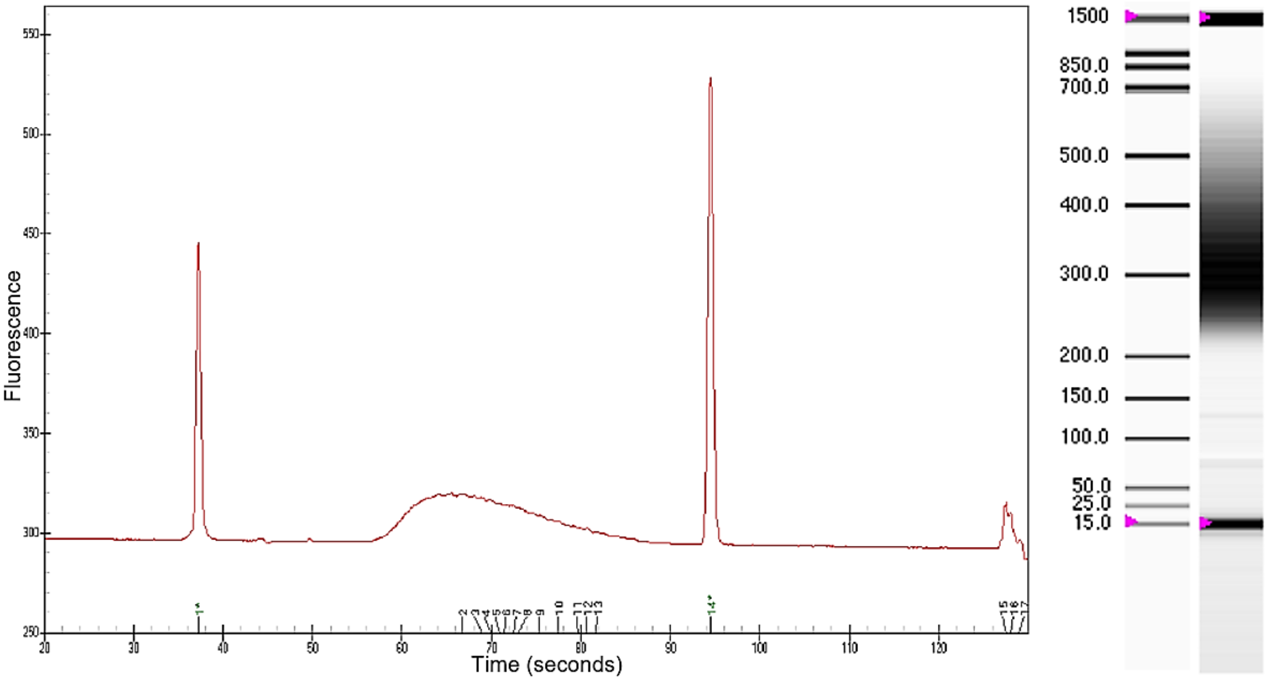


Figure 2-6 | Representative electropherograms showing the fragment size distribution of an RNA-seq library.

The quality and size distribution of the constructed RNA-seq libraries were assessed on an Experion™ Automated Electrophoresis System (Bio-Rad Laboratories, Inc., California, United States) using the Experion DNA 1K Analysis Kit (Bio-Rad Laboratories, Inc., California, United States, cat# 7007107) according to manufacturer's protocol. Data acquisition and analysis were performed using the Experion software. A representative electropherogram is shown in Figure 2-6 demonstrating a narrow distribution of library fragments with a peak size of approximately 300bp.

Barcoded RNA-seq libraries were quantified by qPCR using a KAPA Library Quantification Kit (Illumina® Platforms; cat# KK4824) and pooled in equimolar concentrations. Pooled libraries were sequenced by GENEWIZ on a paired-end 150-cyle Illumina Hiseq X Ten platform. Demultiplexing of samples was performed using Illumina bcl2fastq (ver. 2.17) according to index sequences by the sequencing facility.

2.4 Bioinformatics data analysis

2.4.1 Omni ATAC-seq data analysis pipeline

2.4.1.1 Data processing, alignment and peak calling

The analysis of ATAC-seq data used the following tools and versions: FastQC ver. 0.11.7, Samtools ver. 1.3.1, Picard ver. 2.2.4, Bowtie2 ver. 2.2.9, MACS2 ver. 2.1.2, BEDTools ver. 2.25.0, bedmap ver 2.4.36, Subread ver. 1.5.2.

The sequencing data quality was determined using FastQC ver. 0.11.7[156] followed by trimming of Nextera adapters using cutadapt (ver. 1.14). Trimmed reads were aligned to the GRCh37 genome using Bowtie2 (ver. 2.2.9) with '-X 2000' setting. For each sample quality trimming was performed with option '-q 10' with unmapped and multi-mapped reads filtered with option '-F 2828' using Samtools ver. 1.3.1. Uniquely mapped paired reads were then filtered to exclude PCR duplicates using Picard ver. 2.2.4). Mitochondrial reads, reads mapping to ENCODE hg19 blacklisted regions [157] (regions with anomalous high signal across multiple genomic assays and cell types) and mitochondrial blacklisted regions (high signal regions on the nuclear genome due to sequence homology with the mitochondrial genome) were filtered out using BEDTools (ver. 2.25.0). For peak calling and TF footprinting, the read start sites were adjusted to represent the center of a Tn5 transposase binding event as described in Adey, Morrison [158]. Briefly, as a consequence of the Tn5 transposase binding as a dimer and inserting two adapters separated by 9bp [158], all reads aligning to the forward strand were offset by +4 bp whereas all reads aligning to the reverse strand were offset by -5 bp.

2.4.1.2 Differential accessibility analysis

For differential chromatin accessibility analysis, the processed ATAC-seq reads were first concatenated from all sample replicates by groups. Peaks were called from the merged bam files using MACS2 ver. 2.1.2 [159] with parameters '*callpeak -f BAMPE -g hs --nolambda --*

min-length 100 --max-gap 50 --call-summits --bdg --keep-dup all' (unless stated otherwise). Peak calling was performed on the pooled libraries prior to the statistical analysis to control for type I error [160]. For each data set of peaks generated by MACS2 overlapping peaks were merged using BEDTools (ver. 2.25.0) and the number of reads mapping to each peak in each individual sample was calculated using csaw [160] or featureCounts [161], including only fragments with both ends successfully aligned (-B) and with reads overlapping multiple features assigned to the feature with the largest overlap *(--largestOverlap*). For the csaw pipeline, reads within a peak were quantified using csaw *regionCounts* and peaks containing low counts were removed based on average log2 counts-per-million (CPM=0.4) to correct for the different library sizes across libraries. The normalization factors across libraries were computed using csaw *normFactors* [162], which uses the trimmed mean of M-values (TMM) method to eliminate composition biases, typically in background regions of the genome, and imported into R for differential accessibility analysis using edgeR *exactTest* [38] unless otherwise stated.

For the featureCounts pipeline, counts were normalized for sample library size and count dispersions were calculated in edgeR upon removal of low read counts (minimum coverage the same as csaw workflow). Differential accessibility analysis was performed using exactTest with prior count increased to 1 to reduce overestimation of log fold-change values for low-abundance differential peaks as described in Zuberbuehler, Parker [163]. For both csaw and featureCounts differential gene expression (DGE) workflows, regions having a Benjamini–Hochberg FDR below 0.05 were considered to have significantly differential accessibility between groups. IGV ver. 2.5.2 (Broad Institute) and UCSC genome browser (University of California Santa Cruz) were used for visualization of ATAC- and RNA-seq tracks.

2.4.1.3 Transcription factor (TF) footprinting and differential footprinting analysis

HINT-ATAC[21] was used to identify TF footprints, perform motif matching and differential footprint analysis from ATAC-seq peaks. HINT-ATAC[21] was used to call footprints with parameters *‘rgt-hint footprinting --atac-seq --paired-end --organism=hg19’* and differential footprints with parameters ‘*rgt-hint differential --organism=hg19 –bc*’.

HINT-ATAC also allows the use of fragment-size decomposition of cleavage signals from ATAC-seq data as input. It has been reported that the combined use of fragments from nucleosome-free regions and regions bound by one nucleosome is best for footprinting analysis of Omni ATAC-seq data [21]. As suggested by the published literature, footprints were called from the pooled reads extracted from nucleosome-free regions (NFRs) and regions bound by one nucleosome (1N) based upon sequence length (Figure 2-7). Nucleosome-free regions (NFRs) are defined by reads with fragment size below 146bp while regions bound by 1 nucleosome (1N) are defined by reads with fragment size between 146bp and 307bp [21]. Footprint signal is computed using HINT-ATAC on the genome-wide footprints matching the known TF position frequency matrices (PFMs) obtained from JASPAR[29] database to identify potential TF occupancy.

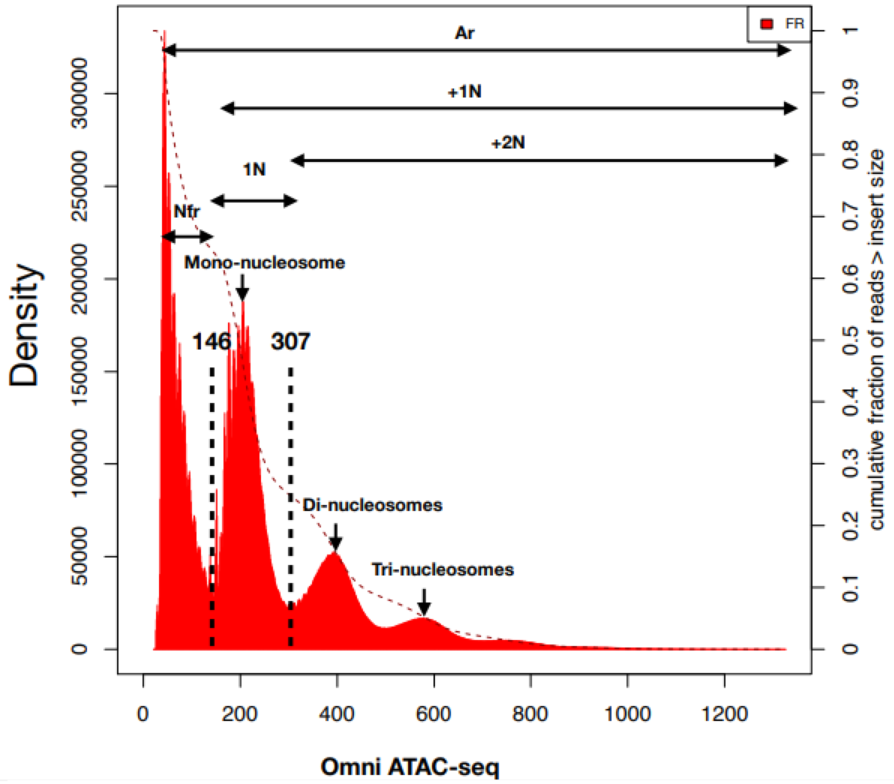


Figure 2-7 | Fragment size distribution of ATAC-seq data shows peaks corresponding to fragments with particular numbers of nucleosome [21].
The combined use of fragments from nucleosome-free regions (Nfr) and regions bound by one nucleosome (1N) is recommended for footprinting analysis of Omni ATAC-seq data. Nfr, nucleosome-free region; 1N, 1 nucleosome.

2.4.2 RNA-seq data analysis pipeline

2.4.2.1 Sequencing quality assessment, data processing and alignment

RNA-seq data quality was determined using FastQC v0.11.7 [156]. Raw reads were trimmed to remove adapters and bases with a quality score (Q) (Phred33) of < 20 using Cutadapt ver. 0.4.4_dev [164] or AdapterRemoval [165]. Adapter- and quality-trimmed reads were aligned to the GRCh38 human genome (GENCODE; hg38) using STAR ver. 2.6.0a55 [166] allowing for no novel splice junctions (--alignSJoverhangMin 500) and keeping only uniquely mapped reads (--outFilterMultimapNmax 1), unless otherwise stated. Aligned reads were assigned to genes in the GENCODE GRCh38 comprehensive gene annotation (ver. M17) using featureCounts (ver. 1.6.2), considering only fragments with minimum mapping quality score of 10.

2.4.2.2 Differential expression analysis

Differential expression analysis was performed using edgeR [38] (ver. 3.22.3) running on R. Briefly, raw counts were imported and filtered to remove genes with low or no expression (≤ 1-2 counts per million in all experimental groups). Filtered counts were then normalized for library size using edgeR *calcNormFactors()*, followed by estimation of count dispersion using *estimateDisp()*. Linear modelling in R package limma was carried out using the *lmFit* and *contrasts.fit* functions in edgeR. Significantly differentially expressed genes were defined as genes having a Benjamini-Hochberg false discovery rate (FDR) less than 0.05. For estimating transcript abundance, reads per kilobase million (RPKM) was determined from normalized read counts using the edgeR rpkm() function. Data were visualized using ggplot2 (ver. 3.0.0).

2.4.3 Gene ontology (GO) and Gene set enrichment analysis (GSEA)

In order to identify gene ontology (GO) terms or sets of genes that are significantly over-represented between two biological states in the ATAC-seq and RNA-seq data, gene ontology (GO) enrichment analysis and gene set enrichment analysis (GSEA) were performed. Whilst GO enrichment analysis [120] identifies significantly enriched terms from a list of genes (e.g. significantly differential genes), GSEA [121] takes the entire list of genes as input and ranks the genes to determine which gene sets are overrepresented at the top or bottom of a ranked list of genes.

GO enrichment analysis was performed using R package limma using the goana.default function. The GO analysis used an up-to-date summary of the Gene Ontology database [122] as represented in the R package GO.db. GO aspects molecular function, biological process and cellular component were included in the analysis.

For GSEA, the gene sets used were subsets of complete collection of the Molecular Signatures database (MSigDB release 6.2) and included the Homo sapiens hallmark (H), curated (C2), regulatory target (C3), ontology (C5) and immunologic signature (C7) gene sets. A total of

20,750 gene sets were included for analysis. GSEA was run using 1×10^6 permutations to allow for the appropriate adjustment of p-values given the large number of tests. Gene sets with more than 250 genes were excluded from analysis in order to restrict results to informative gene sets.

CHAPTER 3: ATAC-SEQ PROTOCOL OPTIMISATION

3.1 Introduction

Gene expression is predominantly regulated by the binding of TFs to the promoters and distal regulatory elements such as enhancers. Chromatin structure has a major influence on gene expression through controlling the access of TFs to binding sites in enhancers and promoters. Localised patterns of epigenetic modification of chromatin, such as histone modification, DNA methylation and nucleosome remodelling correlate with enhancer activity, TF activity and regulation of transcription. This has resulted in collaborative efforts to catalogue regulatory elements harbouring epigenetic marks and TF binding sites in hundreds of mammalian cell types and tissues on a genome-wide scale, through large, international consortia that share similar interests, such as the Encyclopedia of DNA Elements (ENCODE), The Functional Annotation of the Mammalian Genome 5 (FANTOM5), The International Human Epigenome Consortium (IHEC) and NIH Roadmap Epigenomics Consortium [8, 9, 42, 167, 168]. Briefly, these consortia have catalogued genome wide profiles of the epigenetic modifications using multiple platforms: DNA methylation, measured by bisulfite sequencing [169, 170]; histone modifications [171, 172] and TF binding by ChIP-seq [173, 174]; chromatin accessibility by DNase-seq (DNase I hypersensitive sites sequencing) [125, 175-177] and more recently by ATAC-seq [22, 25, 141, 178].

Chromatin accessibility is a measure of regulatory capacity at a gene locus and a critical determinant of chromatin organization and function. It controls the extent to which nuclear macromolecules are able to physically interact with the chromatin and is determined by nucleosome occupancy and composition, as well as physical contact of non-histone protein such as TFs that regulate access to DNA [179]. The accessible DNA makes up ~2-3% of the genome yet it captures more than 90% of the regions occupied by TFs assayed in the ENCODE project [42, 125, 179]. Therefore chromatin accessibility informs both aggregate TF activity and the regulatory capacity of a genetic locus. Chromatin accessibility is usually measured by the susceptibility of chromatin to enzymatic methylation or cleavage. However, conventional chromatin profiling methods such as DNase-seq requires high number of cells as input requirement and it is not able to probe the interplay of nucleosome organization, chromatin accessibility and TF occupancy simultaneously. The development of ATAC-seq has made it possible to quantify chromatin accessibility using a much lower number of cells, therefore amenable to limited clinical and primary tissue samples.

Biobanks provide us with a valuable resource in advancing our understanding in human health and disease. Cryopreservation in biobanking preserves structurally intact living cells and tissues and is a common practice in basic research and clinical applications. They make a significant contribution to improving our understanding, discovery, prevention, diagnosis, intervention and cure of complex diseases [180]. The osmotic, temperature and solute changes that take place during cryopreservation and thawing can have a significant impact on the viability and functionality of the recovered cells [181]. Some of the technical parameters that can influence the quality and functionality of the recovered cells during cryopreservation and thawing process include cell density, choice of cryoprotectant and thawing media, thawing duration and centrifugation speed. A robust cryopreservation and thawing protocol are imperative to ensure the quality of recovered cells closely recapitulates that of the fresh cells such that downstream applications deliver an accurate interpretation and clinical translation.

This chapter has four aims. First of all, as my work aimed to profile the chromatin accessibility of T cells recovered from the frozen, PBMC samples collected from T1D patients and sibling-matched healthy cohort biobanked between 2015 to 2017, I evaluated and optimised published thawing techniques in an attempt to improve overall thawed cell viability and recovery using cells isolated from healthy adult volunteers. Cryopreserved cells often demonstrate reduced viability and functionality upon thawing. It is critical that good methodologies are first established for isolation of sufficient viable Treg and Tconv cells from thawed PBMCs for application of ATAC-seq protocol. Secondly, I aimed to optimise the ATAC-seq protocol for use on T cells recovered from frozen PBMC samples. To address that I performed ATAC-seq on frozen T cells obtained from healthy adult volunteers and evaluated the effects of varying Tn5 amount, purification protocols and tagmentation conditions on the quality of the ATAC-seq libraries.

Thirdly, I also explored the plausibility of extracting RNA for transcriptomics from the same cell pool as the ATAC-seq lysate, derived from just 50,000 cells. This approach allows simultaneous profiling of accessibility and transcriptome information from a single ATAC-seq reaction and would greatly benefit experiments constrained by limited input material. It also improves the confidence in correlating chromatin accessibility with gene expression as the DNA and RNA material are derived from the same pool of cell populations, thereby reducing cell variability or heterogeneity in inferring the overlap between the two experiments.

Lastly, using PBMC samples obtained from healthy adult volunteers, I aimed to assess the feasibility and compatibility of ATAC-seq protocol on frozen T cells, and how closely the chromatin accessibility profile from the frozen cells recapitulates that of the fresh cells. This also serves as a reasonable proxy for assessing how good the quality of the biobanked clinical samples was going to be, as the PBMCs derived from healthy adults were frozen according to the standard biobanking protocol used for T1D and healthy control samples. To date a number of studies have evaluated the effects of cryopreservation on profiling chromatin structure using ATAC-seq [30, 182] but none looked in context of frozen primary CD4$^+$ T cells, especially at both steady state and in response to cell stimulation. To address that I measured the viability and recovery of thawed PBMCs and compared them to that of the fresh PBMCs. I also performed ATAC-seq on resting and stimulated fresh and frozen Treg cells and compared the accessibility landscape (accessible peak regions, TF footprints) between the two.

I was able to establish a thawing methodology suitable for use on the biobanked clinical samples. I was able to demonstrate that optimised thawing protocol results in significantly higher recovery (p=0.008) of live PBMCs compared with standard or unmodified thawing and sufficient number of high purity Treg and Tconv cells was isolated from the thawed PBMCs (Tconv = $1.34\text{x}10^6$ ± 103,052; Treg = $8.8\text{x}10^4$ ± 4,596; purity > 90% determined by flow cytometry). I was also able to show that ATAC-seq libraries generated from frozen PBMCs exhibit a similar accessibility fingerprint for open chromatin signatures such as transcription start sites (TSS) and distal enhancer elements as published ATAC-seq datasets. Lastly, the fresh and frozen Treg ATAC-seq libraries display comparable enrichment of accessible regions, TF occupancy footprints as well as responsiveness to stimulation.

3.2 Aims and Hypothesis

The hypothesis for this chapter is that the chromatin accessibility landscape of the frozen T cells closely recapitulates that of the fresh T cells.

The aims of this chapter are:

1. To optimise thawing protocol for improved PBMC viability and live cell recovery

2. To establish an optimised ATAC-seq protocol for use on T cells recovered from frozen PBMC samples

3. To compare the chromatin accessibility landscape of fresh and frozen Treg cells

3.3 Materials and Methods

3.3.1 Optimisation of thawing protocol

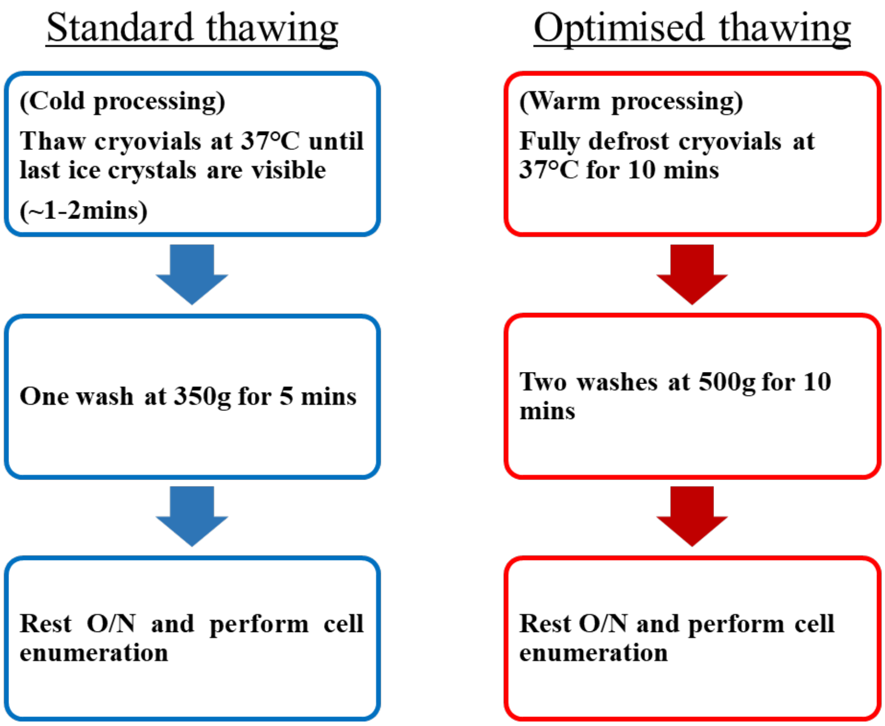

Standard thawing
(Cold processing)
Thaw cryovials at 37°C until last ice crystals are visible
(~1-2mins)
One wash at 350g for 5 mins
Rest O/N and perform cell enumeration
Optimised thawing
(Warm processing)
Fully defrost cryovials at 37°C for 10 mins
Two washes at 500g for 10 mins
Rest O/N and perform cell enumeration

Figure 3-1 | Optimisation of PBMC thawing entails some of the key modifications to the standard protocol.

The standard protocol used for thawing of cryopreserved PBMCs was adopted by the Barry Lab as a routine cell thawing procedure, and it was the baseline against which further modifications were tested in this work. The modification of parameters introduced to the standard thawing protocol was mainly based on the literature published in the Ramachandran, Laux [181] study, with adaptations. Briefly, for standard thawing, the frozen PBMC vials containing 1mL of cells (FCS with 10% DMSO) were thawed in a 37°C water bath rapidly until the last ice crystals were visible (~1-2 minutes) and 1mL of pre-warmed, 37°C complete X-VIVO 15 culture media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5% heat-inactivated human serum were added to the cells. The 2-mL cell suspension was then transferred to and mixed with 8mL of complete X-VIVO culture media (at a rate of 1 mL in < 5 seconds), followed by centrifugation at 350g for 5 minutes. The thawed cells were allowed to recover overnight in the complete X-VIVO culture media in a 24-well culture plate at 37 °C in a CO2 incubator for 16-17 hours prior to cell enumeration and T cell sorting by FACS.

As for optimised thawing, the frozen PBMC vials were thawed by incubating in a 37ºC water bath for 10 minutes to raise the temperature of the cells to 37ºC. The fully thawed cells were then quenched slowly (at a rate of 1 mL/5 seconds) with 9mL of pre-warmed, 37ºC complete X-VIVO culture media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine, 5% heat inactivated human serum and 200U/mL DNase (Worthington cat# LS002007) twice by centrifugation at 500g for 10 minutes. The thawed cells were allowed to recover overnight at a cell density of ~3.5–4.0 x 10^6/mL, in the complete X-VIVO culture media in a 24-well culture plate at 37°C in a CO2 incubator for 16-17 hours prior to cell enumeration and T cell sorting by FACS.

The assessment of cell viability and recovery was performed with a trypan blue dye exclusion test. The viability (%) of cells is defined as,

$$\%\,Viability = \left(\frac{number\ of\ viable\ cells}{total\ number\ of\ frozen\ viable\ cells}\right) \times 100$$

whereas the recovery (%) of cells after overnight culture is defined as

$$\%\,Recovery\ after\ overnight\ culture$$
$$= \left[\frac{number\ of\ viable\ cells\ after\ overnight\ culture}{(total\ number\ of\ frozen\ viable\ cells - number\ of\ cells\ removed\ for\ measurement\ directly\ after\ thawing)}\right] \times 100$$

3.3.2 Optimisation of ATAC-seq protocol

As described in Section 2.2 this work employed Omni ATAC-seq as the protocol of choice for ATAC-seq, as it was reported to be compatible with frozen samples, with substantial improvement in signal-to-background ratio and complexity in comparison with other ATAC-seq method variants [153]. At the outset of this work, part of my candidature was dedicated to establishing the ATAC-seq protocol, as well as data analysis pipeline from first principles, as it had not been performed previously within the laboratory. Part of the training undertaken was a lab placement at the Life and Medical Sciences Institute (LIMES), Bonn, Germany which trained and set me up for standard ATAC-seq methodology including competencies in bioinformatics data processing and analysis. Since returning to Adelaide I adapted and optimised the Omni ATAC-seq protocol for use on frozen primary Tconv and Treg cells. Using Tconv cells isolated from thawed PBMC obtained from whole blood donated by healthy adult volunteers, the ATAC-seq protocol was optimised in an attempt to improve overall library digestion pattern and quality through modifications to Tn5 transposase input amount (2.5 – 3.5µL), tagmentation time (30 vs. 45 minutes), choice of purification kits (MinElute PCR Purification or Zymo Research DNA Clean & Concentrator-5) and library size selection range/kits (Promega ProNex, AMPure XP or SPRIselect).

Briefly, in order to determine the optimal tagmentation duration, 50,000 lysed cells were subjected to 30 or 45 minutes of incubation at 37°C in a thermomixer with 1000 rpm mixing in 50µL of transposition mix comprising of 25µL 2× TD buffer, 2.5µL Tn5 transposase (Illumina Inc), 16.5µL PBS, 0.5µL 1% digitonin, 0.5µL 10% Tween-20 and 5 µL water. The 30- and 45-minute tagmentation reaction was purified using either a Zymo DNA Clean & Concentrator-5 (cat# D4014) or a Qiagen MinElute PCR Purification Kit (cat# 28004) kit. All libraries were amplified for a total of 9 PCR cycles (as described in Section 2.2.2) and size selection was carried out using SPRIselect beads (Beckman Coulter, cat# B23318) after PCR amplification to include a fragment size window of 100 to 800bp prior to sequencing as described [141]. Barcoded ATAC-seq libraries were quantified by qPCR using a KAPA Library Quantification Kit (Illumina® Platforms; cat# KK4824) and pooled in equimolar concentrations. Before sequencing on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 16.3 million reads (± 3.7 million) per sample.

3.3.3 Omni ATAC-seq of Fresh and Thawed Regulatory T (Treg) cells

3.3.3.1 Cryopreservation, thawing and T cell culturing

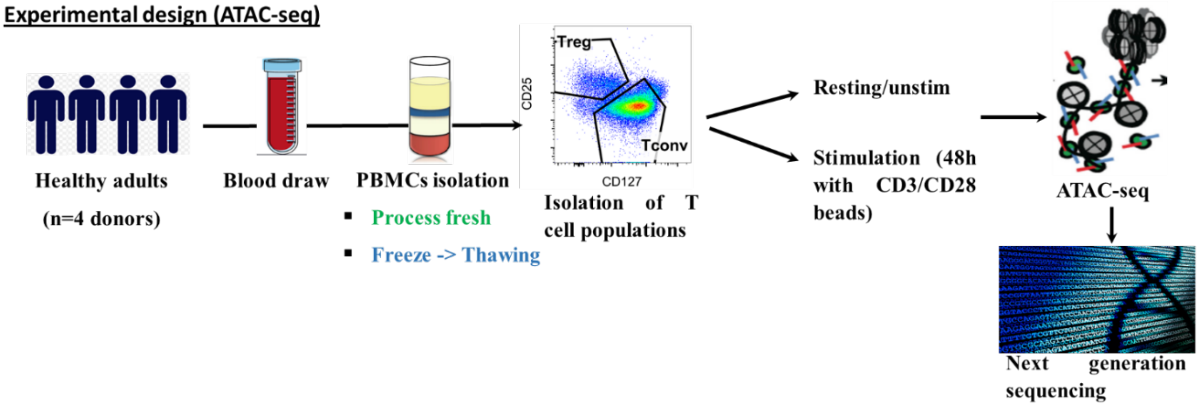

Experimental design (ATAC-seq)
Healthy adults
(n=4 donors)
Blood draw
PBMCs isolation
Process fresh
Freeze -> Thawing
Treg
CD25
Tconv
CD127
Isolation of T cell populations
Resting/unstim
Stimulation (48h with CD3/CD28 beads)
ATAC-seq
Next generation sequencing

Figure 3-2 | Experimental design for profiling chromatin accessibility in the fresh and frozen T cells from healthy adults.

PBMCs were isolated from fresh whole blood obtained from 4 healthy adult as described in Section 2.1.1 (Figure 3-2). Washed PBMCs were divided into two groups, half for fresh processing and the other half for freezing/biobanking (Figure 3-3).

As described in experimental flowchart from Figure 3-4, PBMC samples to be processed fresh were rested overnight in complete X-VIVO 15 culture media (X-VIVO 15 Serum-free media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine and 5% heat inactivated human serum) at 37ºC in a humidified 5% CO2 incubator for 16-17 hours prior to cell sorting of Tconv and Treg cells by FACS the next day.

PBMC samples to be frozen were resuspended in 1mL of freezing medium (heat-inactivated FCS containing 10% DMSO) at a concentration of 1×10^7 cells/mL, transferred to a 1.5-mL cryovial and placed into a Mr. Frosty™ Freezing Container. The unit was placed in the -80°C freezer overnight and the frozen tubes were transferred to liquid nitrogen for long-term storage. Frozen PBMC samples were left in the liquid nitrogen for at least 2-3 weeks before thawing (Figure 3-4). Thawing was performed as described in Section 2.1.3 for isolation of Tconv and Treg cells following O/N recovery.

The sorted Treg cells were either left untreated (i.e. resting/unstimulated) or stimulated with beads conjugated with anti-CD3 and anti-CD28 antibodies (Dynabeads Human T-Expander CD3/CD28, Gibco no. 11141D, Life Technologies) in complete X-VIVO 15 culture in the presence of 500U/mL recombinant human IL-2 at a cell/bead ratio of 1:1 for 48 hours. After 48 hours Dynabeads were removed from culture medium by magnetic separation.

3.3.3.2 Library construction for Omni ATAC-seq

Briefly, resting (on the sort day) and stimulated cells (48h post sorting) were pre-treated with 200U/mL of DNase (Worthington) for 30 minutes at 37°C prior to ATAC-seq experiment. Cell counts were performed manually using a hemocytometer and 50,000 Treg cells were lysed in 50µL of cold resuspension buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM $MgCl_2$ in water) containing 0.1% NP40, 0.1% Tween-20, and 0.01% digitonin on ice for 3 minutes. Immediately after lysis, the reaction was washed with 1mL of ATAC-seq cold resuspension buffer containing only 0.1% Tween-20 by centrifugation at 500g for 10 minutes at 4°C. At this stage, the supernatant (~1mL) was collected and 3 volumes of TRIzol LS reagent (Invitrogen, 10296010) was added to the supernatant (divided into five 1.5-mL Eppendorf tubes), mixed and stored at -80°C for RNA-seq experiment.

The nuclei were then resuspended in 50µL of transposition mix (30µL 2× TD buffer, 3.0µL Tn5 transposase, 16.5µL PBS, 0.5µL 1% digitonin and 0.5µL 10% Tween-20) (Illumina Inc). The transposition reaction was incubated at 37°C for 45 minutes in a thermomixer with 1000 rpm mixing. The reaction was purified using a Zymo DNA Clean & Concentrator-5 (D4014) kit. The subsequent ATAC-seq library preparation was performed as described in Section 2.2.2. All libraries were amplified for a total of 9 PCR cycles and size selection was carried out to include a fragment size window of 100 to 800bp prior to sequencing. Barcoded libraries were pooled and sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 37.2 million reads (± 6 million) per sample.

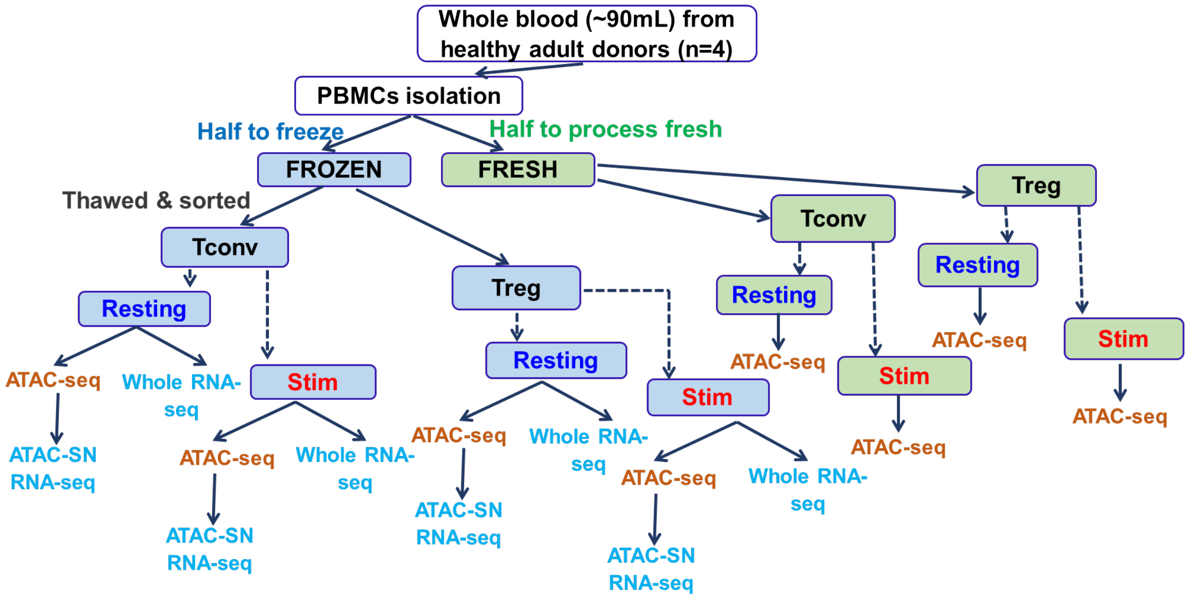


Figure 3-3 | Experimental plan showing the profiling of chromatin accessibility in the fresh and thawed T cells, alongside gene expression in the whole cells and ATAC-seq supernatant (cytoplasmic) fractions.
This approach allows simultaneous and streamlined assaying of accessibility and transcriptome from a single ATAC-seq reaction using just 50,000 cells. ATAC-SN, ATAC-seq supernatant.

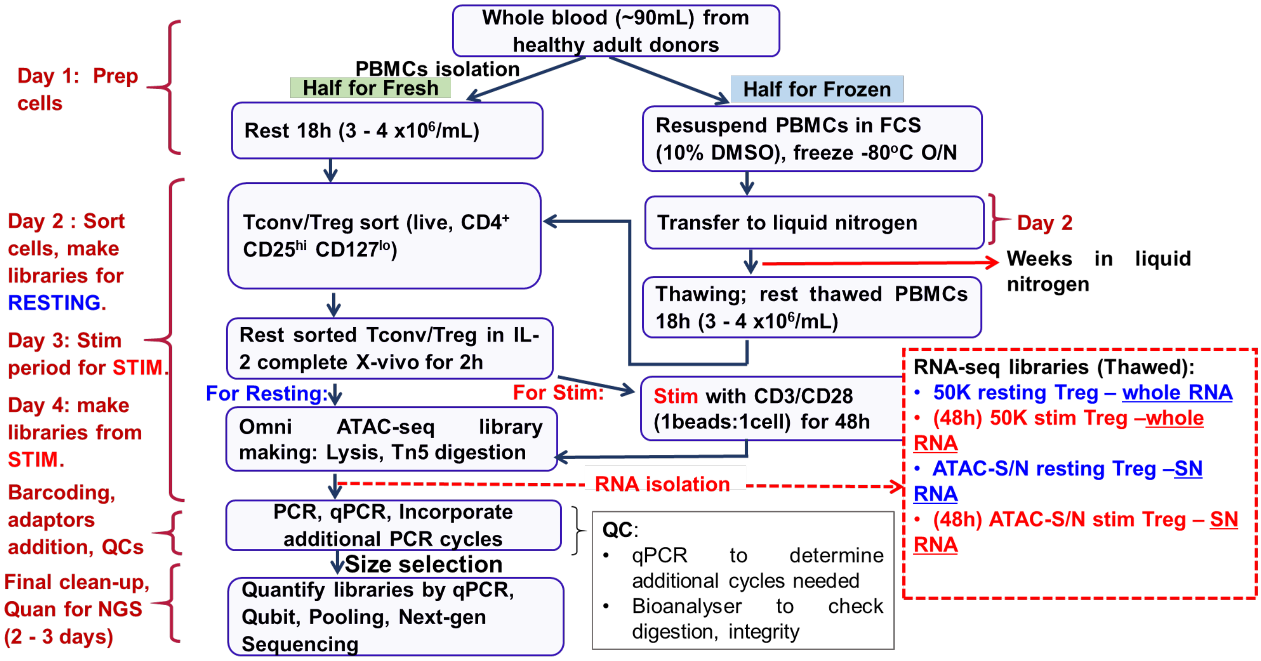

Day 1: Prep cells
Day 2 : Sort cells, make libraries for RESTING.
Day 3: Stim period for STIM.
Day 4: make libraries from STIM.
Barcoding, adaptors addition, QCs
Final clean-up, Quan for NGS (2 - 3 days)
Whole blood (~90mL) from healthy adult donors
PBMCs isolation
Half for Fresh
Half for Frozen
Rest 18h (3 - 4 x10^6/mL)
Resuspend PBMCs in FCS (10% DMSO), freeze -80°C O/N
Tconv/Treg sort (live, CD4$^+$ CD25^{hi} CD127^{lo})
Transfer to liquid nitrogen
Day 2
Weeks in liquid nitrogen
Thawing; rest thawed PBMCs 18h (3 - 4 x10^6/mL)
Rest sorted Tconv/Treg in IL-2 complete X-vivo for 2h
For Resting:
For Stim:
Stim with CD3/CD28 (1beads:1cell) for 48h
Omni ATAC-seq library making: Lysis, Tn5 digestion
RNA isolation
RNA-seq libraries (Thawed):
• 50K resting Treg – whole RNA
• (48h) 50K stim Treg –whole RNA
• ATAC-S/N resting Treg –SN RNA
• (48h) ATAC-S/N stim Treg – SN RNA
PCR, qPCR, Incorporate additional PCR cycles
Size selection
QC:
• qPCR to determine additional cycles needed
• Bioanalyser to check digestion, integrity
Quantify libraries by qPCR, Qubit, Pooling, Next-gen Sequencing

Figure 3-4 | Experimental flowchart showing the timeline for preparation and processing of fresh and frozen T cells for ATAC-seq and RNA-seq.

3.3.4 RNA isolation from ATAC-seq lysis reaction for RNA-seq

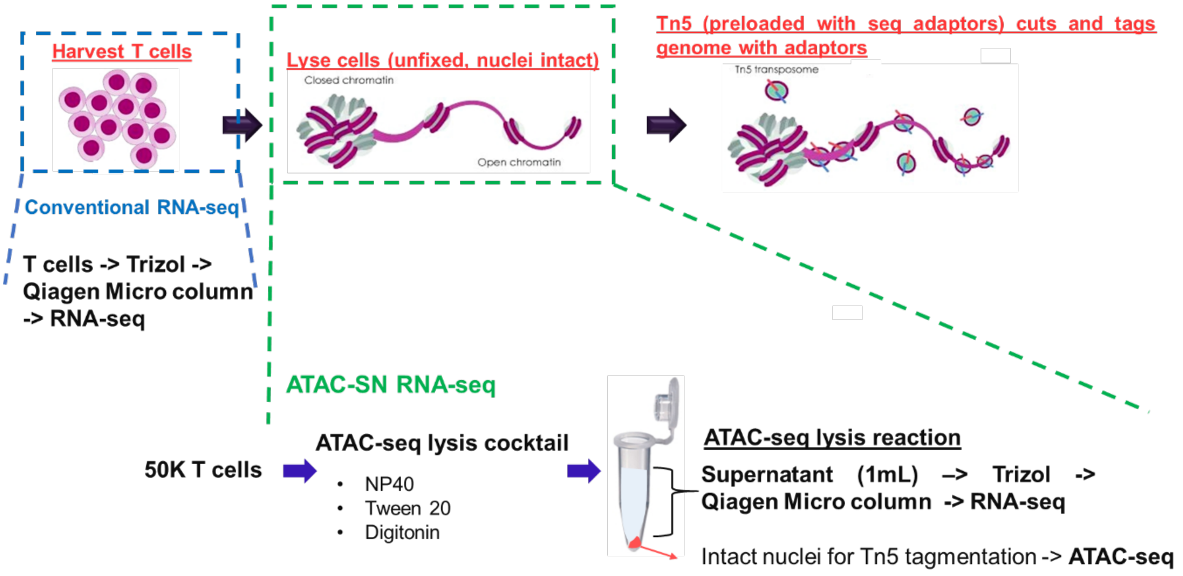

Harvest T cells
Lyse cells (unfixed, nuclei intact)
Closed chromatin
Open chromatin
Tn5 (preloaded with seq adaptors) cuts and tags genome with adaptors
Tn5 transposome
Conventional RNA-seq
T cells -> Trizol -> Qiagen Micro column -> RNA-seq
ATAC-SN RNA-seq
50K T cells
ATAC-seq lysis cocktail
• NP40
• Tween 20
• Digitonin
ATAC-seq lysis reaction
Supernatant (1mL) –> Trizol -> Qiagen Micro column -> RNA-seq
Intact nuclei for Tn5 tagmentation -> ATAC-seq

Figure 3-5 | Experimental workflow demonstrating simultaneous profiling of chromatin accessibility and gene expression from 50,000 of lysed T cells.

RNA was isolated from ATAC-seq supernatant fraction (green inset) for comparison with whole cell RNA (blue inset) in the RNA-seq experiment.

A major aim of this PhD project was to explore the plausibility of extracting transcriptome information prior to ATAC-seq lysate reactions, which allows simultaneous profiling of accessibility and transcriptome information in the same pool of 50,000 cells.

Matching RNA material was collected from the stimulated Treg cells from the same four donors that Omni ATAC-seq profiles were obtained from, as described in Section 3.3.3.2. RNA was isolated from whole Treg cells and supernatant fractions collected from Omni ATAC-seq lysis reaction (ATAC-SN), with detailed process described in Section 2.3.1 and 2.3.2 (also Figure 3-5). Briefly, samples were homogenized using TRIzol LS reagent (Invitrogen, 10296010) and total RNA was extracted using a miRNeasy Micro Kit (Qiagen; cat# 217084) and RNA integrity was determined by the Agilent RNA 6000 Pico Kit using the Agilent Bioanalyzer. Samples with RNA integrity number (RIN) > 7.5 were prepared for mRNA sequencing. All RNA samples were enriched for Poly(A) RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs #E7490) prior to generation of cDNA libraries from Poly (A)- enriched RNA using a NEBNext Ultra Directional II RNA Library Preparation Kit for Illumina (New England Biolabs #) according to the manufacturer’s protocol. Barcoded libraries were pooled and sequenced on a paired-end 150-cycle Illumina Hiseq sequencer (Illumina) to an average read depth of 37.5 million reads (± 16 million) per sample.

3.4 Results

3.4.1 Effect of different thawing conditions on viability and recovery of PBMCs and T cells

My PhD aimed to profile the chromatin accessibility and gene expression in Tconv and Treg cells from a repository of PBMC samples biobanked between 2015 to 2017 from an established T1D cohort and sibling-matched healthy controls. The biobank was established at the Women's and Children's Hospital, Adelaide, South Australia via Professor Jennifer Couper (Endocrinology and Diabetes). As Treg cells are a rare population which constitute only ~5–

15% of the peripheral T cell pool, and the resources of this study are derived from limited biobanked paediatric samples, I first established an optimised thawing protocol in an attempt to improve the viability and yield of cryopreserved cells upon thawing. Rigid and robust thawing practices are important for conservation of cells at the phenotypic and molecular level. This is crucial as cryopreserved cells often exhibit reduced viability and functionality but ATAC-seq requires a high viability population of intact cells to generate good quality datasets to avoid sequencing noise due to the tagmentation of free DNA released by the dead cells. Thus, it is imperative that good methodologies are first established for PBMC thawing and isolation of sufficient, viable Treg and Tconv cells from thawed PBMCs for downstream genetic and genomic applications.

In order to demonstrate and improve the feasibility of using biobanked PBMCs in the proposed downstream experiments I first processed fresh and frozen PBMCs obtained from healthy adult donors (buffy coats from Australian Red Cross Blood Service) according to the standard bio banking protocol. The same criteria as used for the T1D biobank cohort were followed with theses samples in order to recapitulate the conditions and the environment the T1D/sibling cells were cryopreserved in. This frozen material was used to optimise handling techniques in thawing, assess technical reproducibility and donor variation as well as to determine scalability to mimic limited biobanked samples.

Each step in the cell thawing process can be performed in a great magnitude of different ways and each variation may have a significant impact on the viability and functional properties of cells. In my optimised thawing protocol I have modified and introduced adjustments to the original standard protocol with parameters such as incubation time, temperature of washing medium, speed at which the washing medium is added, centrifugation conditions, number of washes and cell density (as specified in Section 3.3.1) in an effort to achieve maximal cell viability and recovery. Frozen PBMCs derived from buffy coats obtained from three healthy adult donors were subjected to both thawing protocols in parallel and were evaluated for cell viability and overall yield.

Using an optimised thawing protocol, which comprised incubation of cryovials at 37°C in a water bath for 10 minutes, implementation of 2 washes in warm media, slow addition of warm media to the cells and overnight culture at a cell concentration of ~3.5–4.0 x 10^6/mL, I observed significantly higher recovery (formulas as specified in Section 3.3.1) in the optimised protocol in comparison with the standard thawing procedure (Figure 3-6). The implementation of optimised thawing techniques did not affect the viability (formulas as specified in Section 3.3.1) of cells significantly.

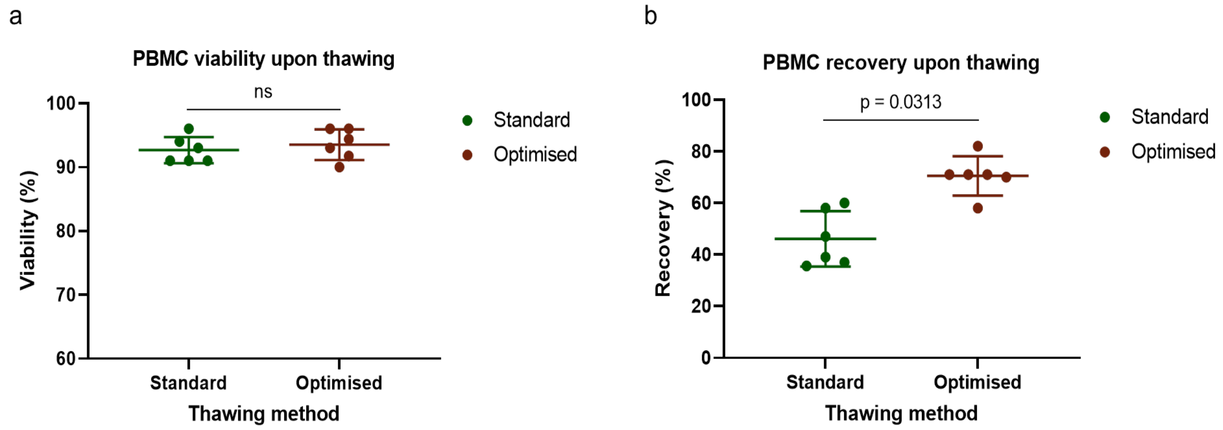


Figure 3-6 | Viability (a) and recovery (b) of thawed PBMCs following different thawing methods.
Results are presented as mean with standard deviation of 6 experiments involving 3 healthy subjects. Statistical comparisons were performed with Wilcoxon signed rank test with two-tailed p values ≤ 0.05 being considered significant.

As cell sorting has to be carried out remotely, the cell thawing and sorting workflow was streamlined and split over two days, involving an overnight resting. It has also been demonstrated that introduction of an overnight resting period in the experimental thawing procedure improves the sensitivity and functionality of T cells recovered from thawed PBMCs [183]. Thus, it was of interest to examine and track the viability and recovery of thawed PBMCs over the resting period to determine the timepoint where optimal phenotypes were observed. I evaluated potential changes in cell viability and recovery of the cryopreserved PBMCs over the 17-hour resting period upon thawing (Figure 3-7) and observed minimal fluctuations of PBMC viability and recovery within each donor.

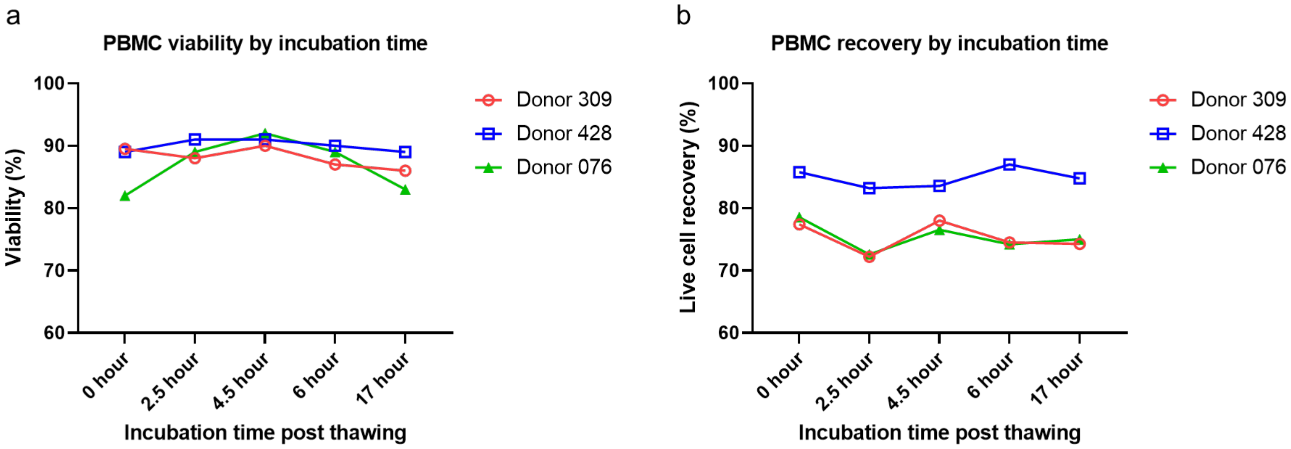


Figure 3-7 | Effect of the duration of the incubation time on the viability (a) and recovery (b) of thawed PBMCs from 3 healthy subjects.
Upon thawing the PBMCs were cultured in the complete X-VIVO 15 media at 37 °C in a 5% CO2 incubator and aliquot of cells from independent wells was enumerated at different timepoints by trypan blue exclusion assay.

Cell clumping is a common observation during PBMC thawing and it leads to cell loss and potentially compromises the phenotype and functionality of the cells. This phenomenon occurs partly owing to the nature of short-lived cells such granulocytes which release "sticky" DNA molecules and cause neighbouring cells and other debris to aggregate into large clumps. It was shown that incorporation of DNase treatment in the thawing procedure reduces cell aggregation[184, 185] and it is compatible with downstream immunophenotyping and lymphocyte functional assays, as it did not result in significant changes in cell viability and expression of leukocyte surface markers in the cryopreserved PBMCs [184]. To determine the effect of DNase on viability and recovery of thawed PBMCs, using the optimised thawing protocol (Section 3.3.1) I treated the cells with complete X-VIVO culture media supplemented with or without DNase (100U/mL) during the two washes (10 minutes each). I observed reduced or no cell aggregate formation in PBMC samples that were treated with supplementary DNase compared with cells that received no DNase treatment. The average cell viability was 87.4% ± 3.3% for DNase-treated cells and 89.3% ± 5.3% for control cells, whereas the average cell recovery was 65.5% ± 14.6% for DNase-treated cells and 62.0% ± 19.0% for control cells. The incorporation of DNase treatment in thawing procedure improved the overall recovery of cryopreserved PBMCs in some donors to some extent, although no significant difference was observed (Figure 3-8). These findings agree with observations from García-Piñeres, Hildesheim [184].

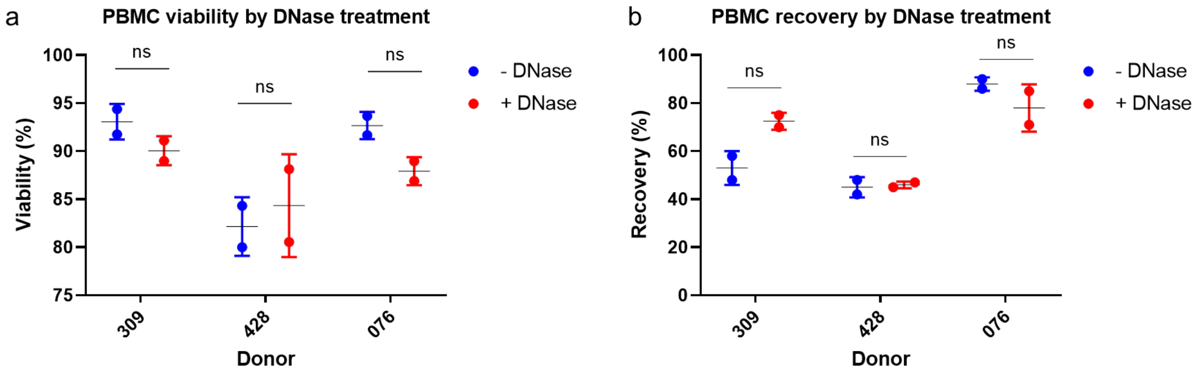


Figure 3-8 | Effect of the DNase treatment on the viability (a) and recovery (b) of thawed PBMCs from 3 healthy subjects (Optimised protocol).
PBMC thawing was performed with or without the addition of DNase in the thawing media during washing steps and cell enumeration was performed post overnight incubation at 37 °C in a 5% CO2 incubator. Two technical replicates were assayed for each donor. Statistical comparisons were performed with multiple paired T-test with p values ≤ 0.05 being considered significant.

With the PBMCs thawed in the presence or absence of supplementary DNase as described above, I also determined the proportion and number of viable Tconv (Figure 3-9) and Treg (Figure 3-10) cells recovered from the thawed PBMCs using FACS. Whilst higher proportion and number of Tconv and Treg cells were recovered from the DNase-treated PBMCs compared with control PBMCs for most donors, the difference was not statistically significant.

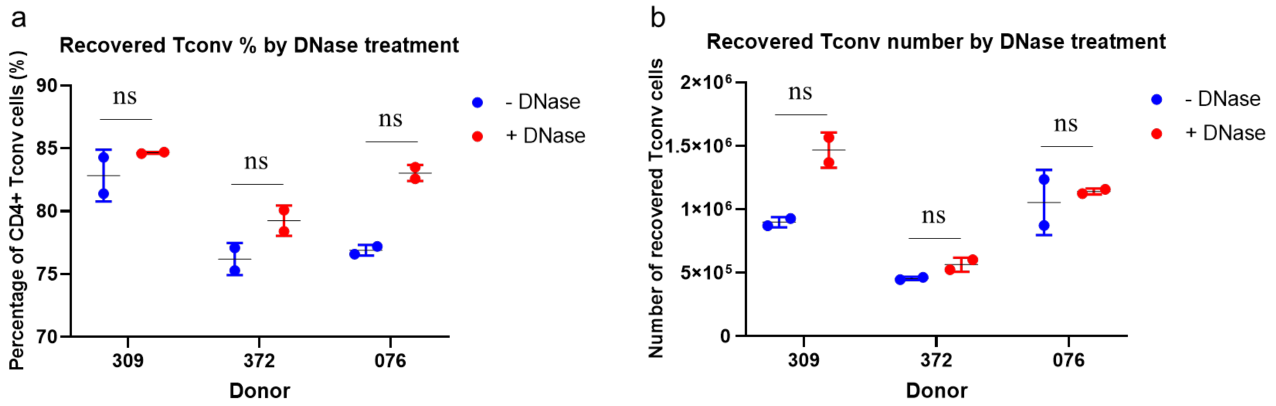


Figure 3-9 | Effect of the DNase treatment on the proportion (a) and number (b) of Tconv cells recovered from 3 healthy subjects.

Tconv cells were sorted from PBMCs thawed in the presence or absence of DNase by FACS post overnight incubation. The proportion and number of recovered Tconv cells were determined by FACS analysis. Two technical replicates were assayed for each donor. Statistical comparisons were performed with multiple paired T-test with p values ≤ 0.05 being considered significant. Tconv, Conventional T cells.

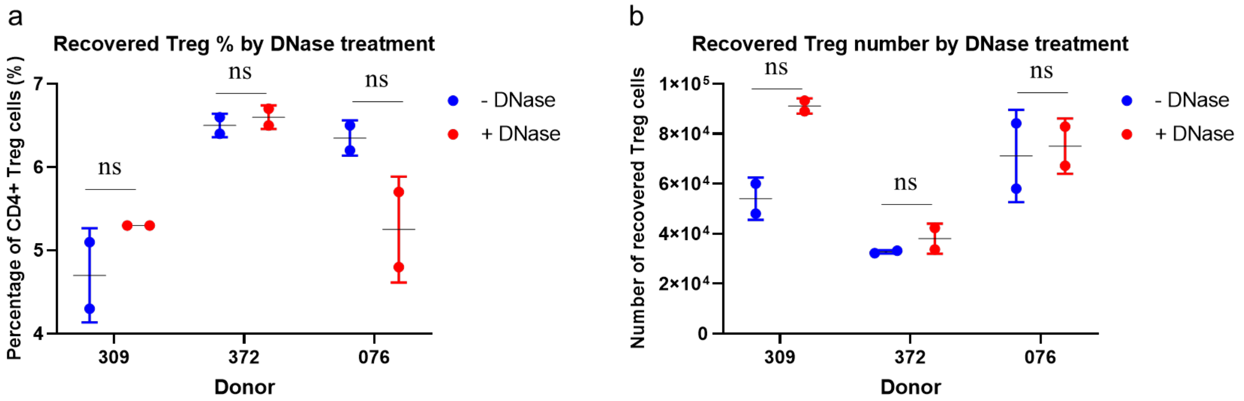


Figure 3-10 | Effect of the DNase treatment on the proportion (a) and number (b) of Treg cells recovered from 3 healthy subjects.

Treg cells were sorted from PBMCs thawed in the presence or absence of DNase by FACS post overnight incubation. The proportion and number of recovered Tconv cells were determined by FACS analysis. Two technical replicates were assayed for each donor. Statistical comparisons were performed with multiple paired T-test with p values ≤ 0.05 being considered significant. Treg, Regulatory T cells.

3.4.2 ATAC-seq protocol optimization (library preparation)

As my work proposed to use cryopreserved T1D and healthy control samples from a biobank repository for profiling chromatin accessibility using ATAC-seq, it is critical to demonstrate the feasibility and reproducibility of applying the assay on the biobanked cells as a readout for epigenetic states. I undertook a four-week training placement at the Life & Medical Sciences Institute (LIMES), University of Bonn in Germany under the guidance of Dr Marc Beyer's and Prof Joachim Schultze's Genomics and Immunoregulation group, to acquire the technical skills of performing ATAC-seq assay as well as competencies in data processing/analysis. Following this training, I established the ATAC-seq protocol in the Barry Lab and refined the methodology for use on cryopreserved T cells. We adopted Omni ATAC-seq [41], an optimised variant of the ATAC-seq protocols [25, 178], as it leads to substantial improvement in data quality and is compatible for a broad breadth of cell types and cell contexts including T cells and frozen tissues. As tagmentation efficiency of Tn5 transposase may vary in different cell types and environments, using T cells isolated from fresh or frozen PBMCs obtained from healthy adult donors (either whole blood or buffy coats from Australian Red Cross Blood Service; Section 3.3.1 and Section 3.4.1), I optimised and refined various aspects of the Omni ATAC-seq protocol, including the input amount of Tn5, tagmentation time, library size selection and purification method, in an effort to identify the optimal conditions to produce good quality ATAC-seq libraries in our system of interest.

The first batch of Omni ATAC-seq T cell libraries were constructed according to the original Omni ATAC-seq[41] protocol with no further modifications to the Tn5 input amount, tagmentation time, library size selection or purification method. Viable Tconv and Treg cells obtained from healthy adult donors were isolated using FACS from fresh or thawed PBMCs treated with DNase to further clean up the dead cell debris, followed by cell lysis, tagmentation and amplification according to the standard Omni ATAC-seq protocol (described in Section 2.2). In ATAC-seq, qPCR is performed to determine the appropriate number of additional PCR cycles required after the pre amplification step, which corresponds to one-third of the maximum fluorescent intensity, after the libraries have undergone the initial pre-amplification step (as described in Section 2.2.2). The qPCR amplification plot of the pre-amplified Omni ATAC-seq

T cell libraries showed that the libraries required additional PCR cycles ranging from 6 – 12 cycles (Figure 3-11, a). The integrity and fragment size distribution of ATAC-seq libraries were assessed on the Experion Automated Electrophoresis System. The distribution of transposed DNA fragment size from most of the libraries showed faint nucleosomal banding pattern on the electrophoresis system (Figure 3-11, b). An optimal transposition reaction should yield a DNA laddering pattern with a periodicity of about 200bp, corresponding to DNA fragments that were protected by an integer number of phased nucleosomes.

To reduce the effect of size- and GC-bias from the library construction process, qPCR cycling signal serves as a guide to stop amplification of ATAC-seq libraries prior to saturation and it is also an important QC metric as it reflects the complexity of the libraries. As described in Buenrostro, Wu [141] if more than 6 additional cycles are needed on the pre-amplified libraries (i.e. >11 total cycles) library complexity becomes a concern. It was unclear as to why the T cell ATAC-seq libraries, which were constructed in parallel under the same conditions and methodology, demonstrated such a great diversity and variation in complexity. The same library construction process was replicated in 3 independent experiments, involving four T cell samples each (n=12), and same outcome was observed.

3.4.2.1 Buffer

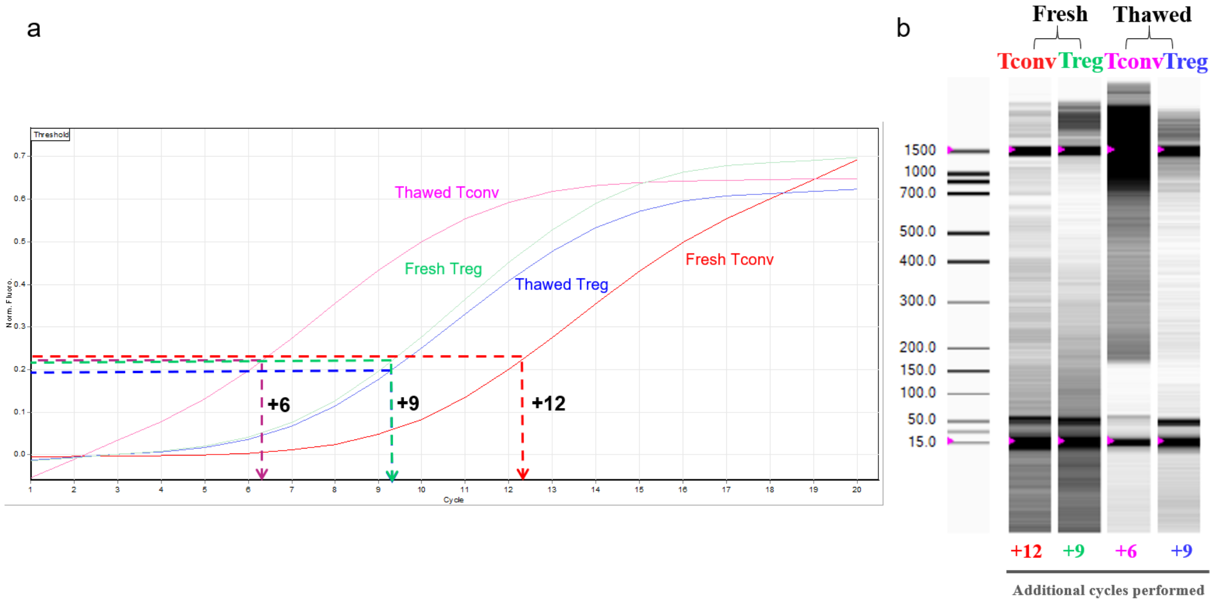


Figure 3-11 | (a) Representative amplification plot demonstrating the number of additional PCR cycles to perform for four T cell tagmentation reactions set up using Illumina Tn5 with in-house tagment buffer (Omni ATAC-seq recipe from Corces, Trevino [41]). (b) Virtual gel showing fragment sizes for amplified ATAC-seq libraries, determined by the Experion Automated Electrophoresis System.

To determine if the observed library complexity and fragmentation was a result of suboptimal transposition (over- or under-transposition), transposition was performed with a titration of Tn5 input, ranging from 2.5uL to 3.5uL, in the same reaction volume (50uL) on five T cell samples (Figure 3-12). Titration of Tn5 input amount did not improve the complexity or fragmentation of the ATAC-seq libraries (Figure 3-12).

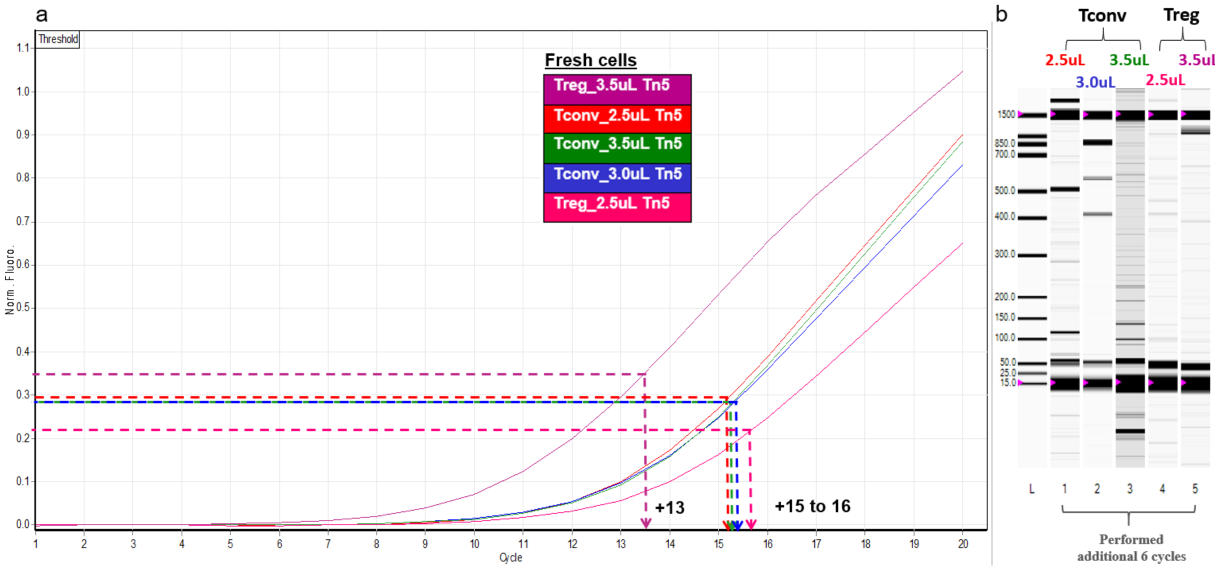


Figure 3-12 | (a) Representative amplification plot demonstrating the number of additional PCR cycles to perform for five T cell tagmentation reactions set up using varying amount of Illumina Tn5 with in-house tagment buffer. (b) Virtual gel showing fragment sizes for amplified ATAC-seq libraries, determined by the Experion Automated Electrophoresis System.

Recipes for buffers required for performing Omni ATAC-seq were included and accessible in the published protocol[41], including recipes for lysis buffer, resuspension buffer for washing as well as tagmentation DNA buffer for transposition reaction. The published Omni ATAC-seq methodology[41] used a lab-made version of Tn5 transposase complex[186] in conjunction with other homemade buffers, however, the T cell libraries (Figure 3-11 and Figure 3-12) were generated with commercial Tn5 transposase (Illumina Nextera DNA Library Prep Kit, cat# FC-121-1030). The transposition reaction was set up using the commercial Tn5 transposase (Illumina) with “homemade” tagmentation DNA buffer (Omni ATAC-seq recipe) and that raised speculation about incompatibility of buffer systems leading to suboptimal and inefficient DNA transposition in cells, as the “homemade” Tn5 transposase may have different properties and requirements on buffer system from the commercial Tn5. To determine whether the observed suboptimal library complexity and fragmentation was a result of buffer incompatibility, Omni ATAC-seq library construction was repeated on T cells (same pool of cells as in Figure 3-11) with transposition performed using the commercial Tn5 transposase and tagmentation DNA (TD) buffer available from Illumina Nextera DNA Library Prep Kit (cat# FC-121-1030). The ATAC-seq library constructed using the commercial transposition reagents (Illumina Tn5 + Illumina TD buffer) was highly complex as the number of additional PCR cycles requited was much lower (Figure 3-13; a) compared with libraries generated using the mixed transposition reagents (Illumina Tn5 + homemade TD buffer) (Figure 3-11 and Figure 3-12). The fragment size distribution of the amplified ATAC-seq library also showed a clear, characteristic DNA laddering pattern with a periodicity of 200bp, corresponding to integer multiples of nucleosomal protection as shown on the virtual gel and electropherogram trace (Figure 3-13; b and c), which is an indication of a good quality ATAC-seq library. The same transposition conditions (Illumina Tn5 + Illumina TD buffer) was adopted in library construction of another 4 T cell ATAC-seq libraries, involving a titration of Tn5 input amount ranging from 2.5uL to 3.5uL, and similar levels of library complexity (Figure 3-14, a) and nucleosomal binding pattern (Figure 3-14, b and Figure 3-15) was observed, although the titration of Tn5 input did not show appreciable change in the library complexity or fragment size distribution. All of these results suggest that incompatible buffer system was likely the cause of suboptimal transposition observed in pilot experiments.

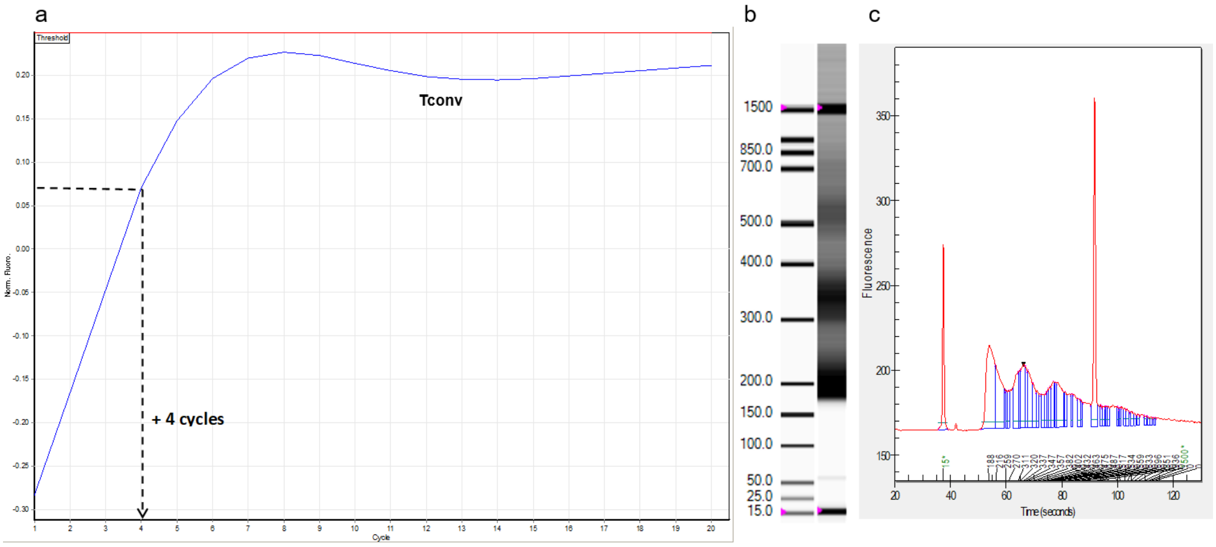


Figure 3-13 | (a) Representative amplification plot demonstrating the number of additional PCR cycles to perform for tagmentation reaction set up using Illumina Tn5 with Illumina Tagment DNA (TD) buffer. Virtual gel (b) and electropherogram (c) showing fragment size distribution of the amplified library, determined by the Experion Automated Electrophoresis System.

3.4.2.2 Tn5 amount

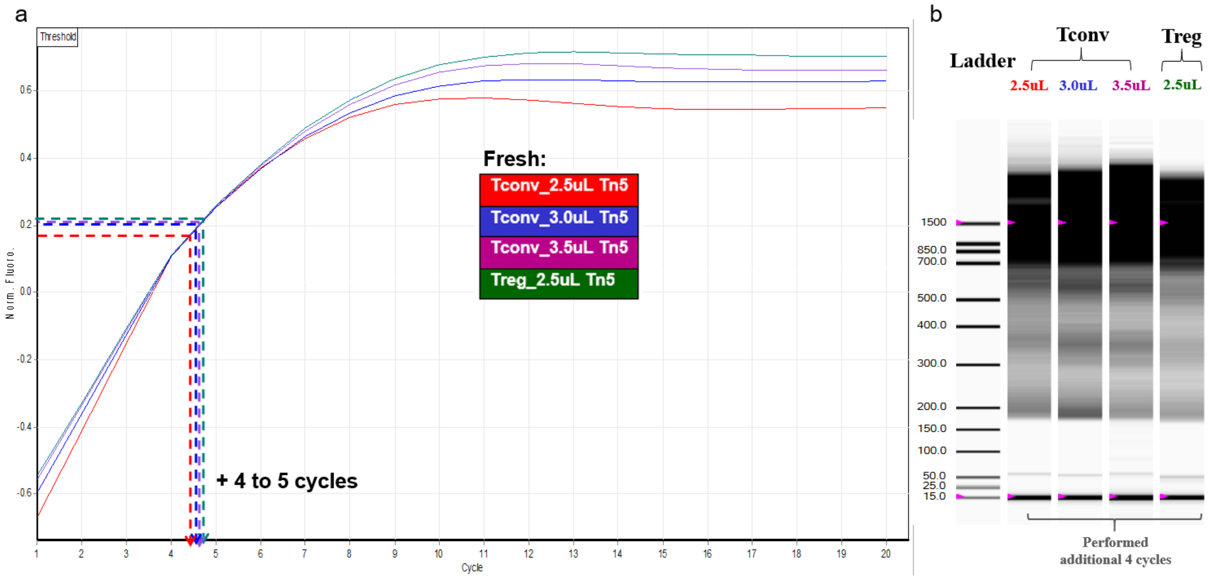


Figure 3-14 | (a) Representative amplification plot demonstrating the number of additional PCR cycles to perform for four T cell tagmentation reactions set up using varying amount of Illumina Tn5 with Illumina Tagment DNA (TD) buffer. (b) Virtual gel showing fragment sizes for amplified ATAC-seq libraries, determined by the Experion Automated Electrophoresis System.

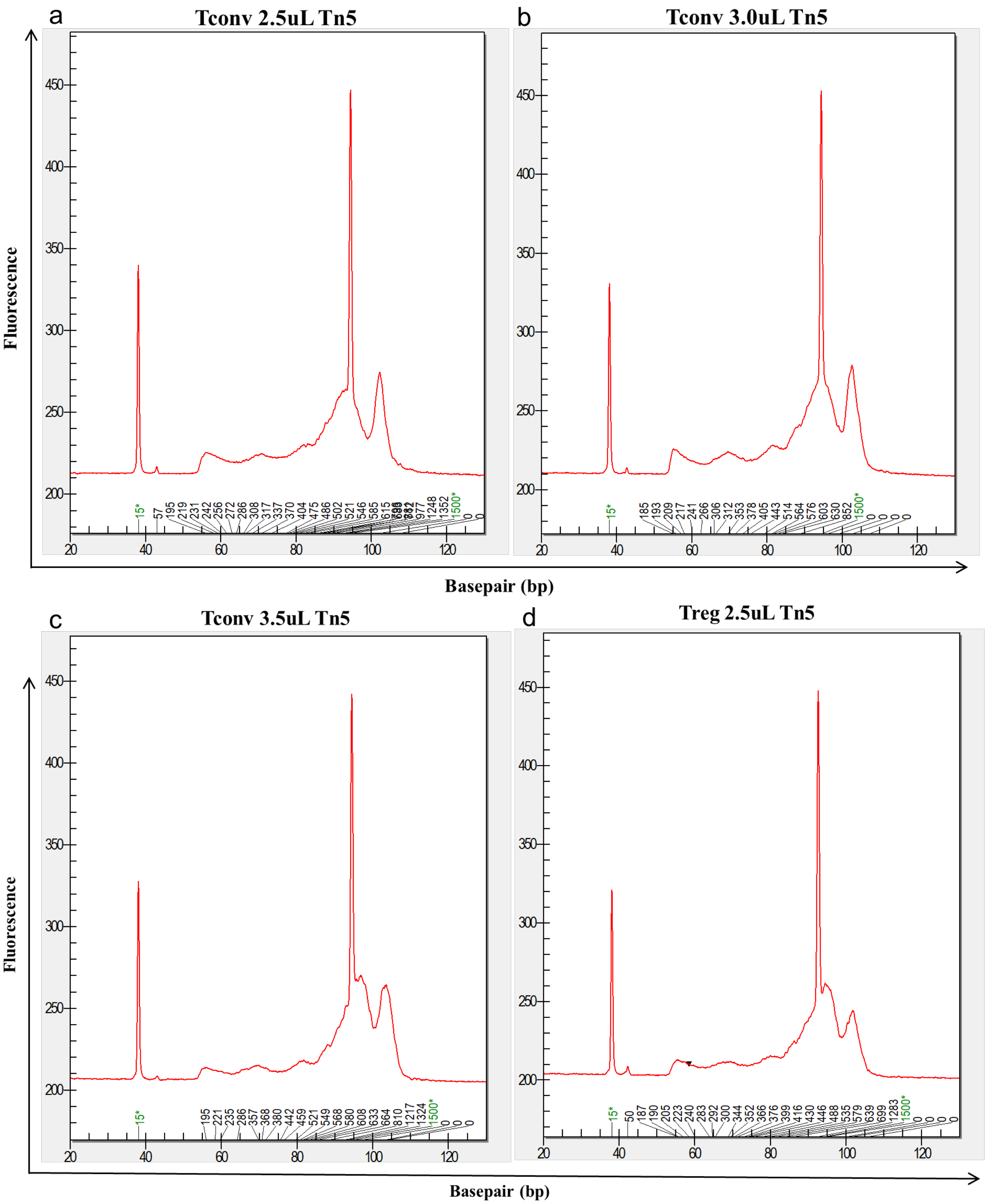


Figure 3-15 | Electropherograms (a-d) showing fragment size distribution of amplified T cell libraries (corresponding PCR amplification signal and virtual gel in Figure 3-14), determined by the Experion Automated Electrophoresis System.

3.4.2.3 Library size selection

ATAC-seq libraries contain highly diverse fragment sizes ranging from 150bp to >1500bp. Enrichment of large fragments (>1000bp) in ATAC-seq libraries can lead to inaccuracy in quantification and subsequent reduced clustering on the flow cell during sequencing. Furthermore, large fragments generally show poor hybridisation to the sequencing flow cell and thus do not get sequenced. Size selection prior to sequencing is recommended on libraries that show a preponderance of large fragments. Dual or double size selection aims to produce a population of dsDNA fragments of a desired size range, removing fragments above and below a chosen size range based on the volume/volume (v/v) ratio of size selection reagent to the DNA sample. I performed dual size selection on ATAC-seq libraries to enrich for DNA fragments between 100-700bp to eliminate contaminating adapter/primer dimers (<100bp), partial library constructs as well as confounding long fragments. I first optimised size selection using a generic dsDNA source (DNA ladder) to identify conditions that produce precise and reproducible population of DNA fragments of desired size cutoffs, before attempting to purify ATAC-seq libraries. I have mainly attempted size selection using three magnetic bead-based purification products – ProNex Size-Selective Purification System (Promega), AMPure XP and SPRIselect (Beckman Coulter Life Sciences). ProNex beads (Promega) was used initially as it was claimed to result in higher recovery of input DNA, accuracy and precision in the size distribution of selected libraries compared with AMPure XP beads. Desired size cutoffs for a population of dsDNA fragments are achieved by varying the volume to volume ratio of selection beads to the staring DNA sample. As this ratio is changed the length of fragments binding to the beads and/or eluted in the supernatant also changes. Lower ratio of bead:DNA is associated with binding of larger dsDNA fragments to the beads (shorter fragments in the supernatant/elution). For dual size selection with ProNex, first selection removes undesired, high molecular weight DNA fragments (desired DNA fragments are in the supernatant), and the second selection removes undesired, low molecular weight DNA fragments (desired DNA fragments are bound to the beads and eluted).

Whereas for double size selection with AMPure XP and SPRIselect, left side selection removes undesired, small molecular weight DNA fragments (desired DNA fragments are bound to the beads and eluted), and the right side selection removes undesired, high molecular weight DNA fragments (desired DNA fragments are in the supernatant). This occurs because DNA fragment size has an effect on the total charge per molecule where larger DNA fragments have larger charge, which promotes electrostatic interaction with the beads and displaces smaller DNA fragments. Dual size selection was initially performed on 1kb Plus DNA ladder (NEB) using ProNex (Promega) kit in an attempt to enrich for population of dsDNA fragments in the range of 100-700bp (Figure 7-1). According to manufacturer's protocol, selection at ProNex Chemistry Ratio (v/v) of 1.05/0.4x removes fragments below 150bp and fragments above 800bp. However, that ratio led to the loss of 600-bp DNA population too and retention of undesired small DNA fragments, and selection using lower ratios (<0.4x) did not successfully eliminate them (Figure 7-1). In order to recover 600-700bp DNA fragments, lower ratio (1.0x) was used in the first selection (which remove long fragments) to promote retention of 600-700bp DNA population. Though DNA fragments of ~700bp were recovered, the sample-to-sample reproducibility was poor, and second selection with lower ratios (<0.3x) did not successfully eliminate undesired short fragments (Figure 3-16; left panel). Dual size selection using ATAC-seq test libraries demonstrated suboptimal performance and poor sample-to-sample reproducibility with considerable separation artifacts on the gel electrophoresis system (which can be caused by poorly resolved fragment peaks).

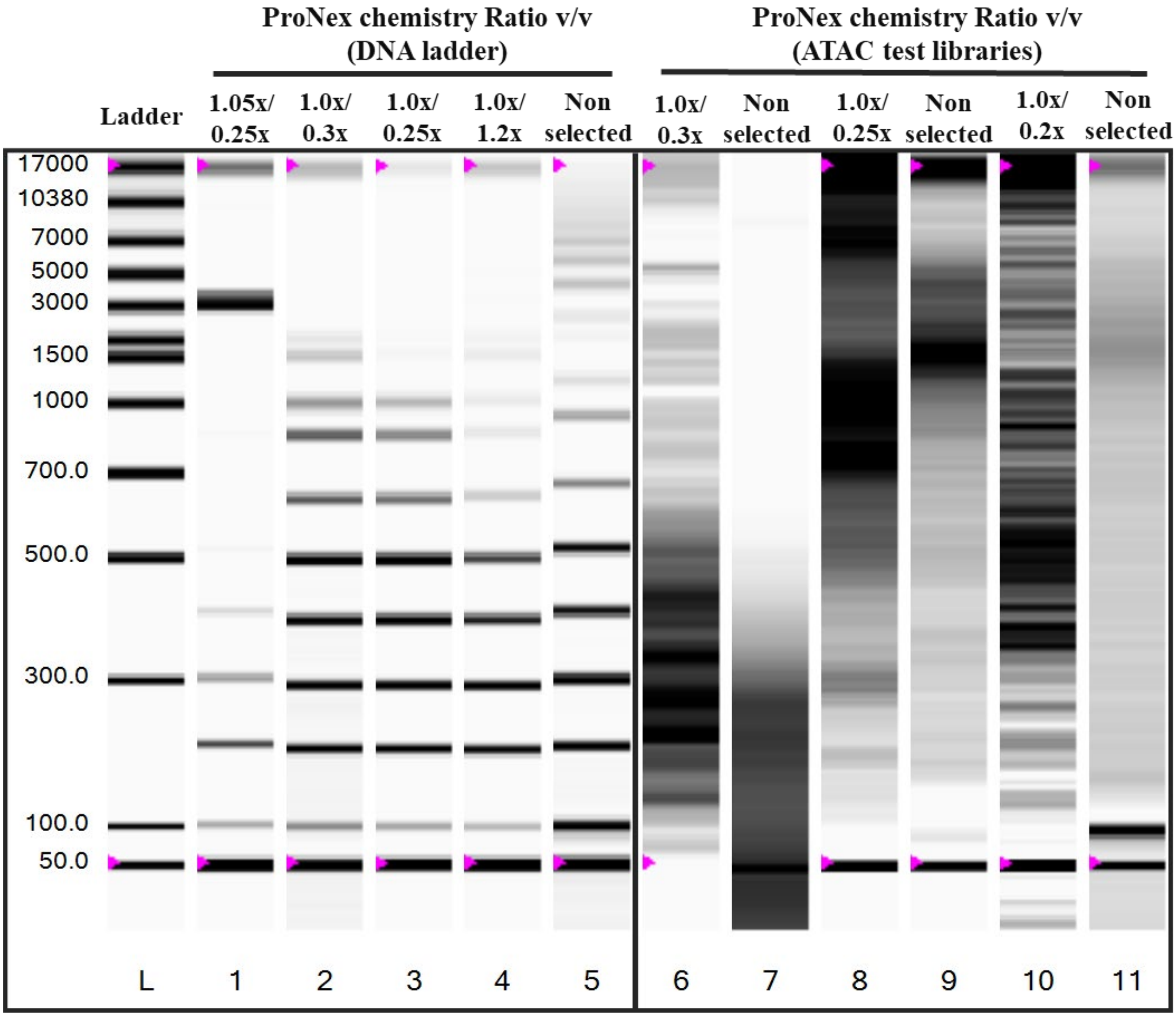


Figure 3-16 | Attempt of dual size selection using 1kb Plus DNA Ladder (left) and ATAC-seq test libraries (right) with varying Promega ProNex chemistry:sample ratios targeting first and second selection to remove undesired DNA fragments above and below desired size cutoffs.
Size-selected DNA Markers or ATAC-seq libraries were electrophoresed on an Experion Automated Electrophoresis System. Lanes labelled Non selected contains input DNA without size selection.

Subpar performance of ProNex beads (Promega) such as low specificity and accuracy of size selection with high carryover of undesired DNA (or loss of desired DNA) as well as poor sample-to-sample reproducibility have prompted me to attempt size selection using alternative products - AMPure XP and SPRIselect (Beckman Coulter Life Sciences). Using AMPure XP and SPRIselect beads I performed double size selection on 1kb Plus DNA ladder (NEB) and subsequently, ATAC-seq libraries to enrich for DNA fragments within the 200-800bp size window using manufacturer's protocol (Figure 3-17; Figure 3-18). Size selection performed using those kits resulted in highly precise, accurate and consistent DNA population of targeted size region with low carryover of undesired DNA and it also demonstrated high compatibility with ATAC-seq libraries (Figure 3-17; Figure 3-18). Double size selection of subsequent ATAC-seq libraries described in this thesis were performed with SPRIselect beads using bead:DNA ratio of 1.0x/1.3x (Beckman Coulter Life Sciences).

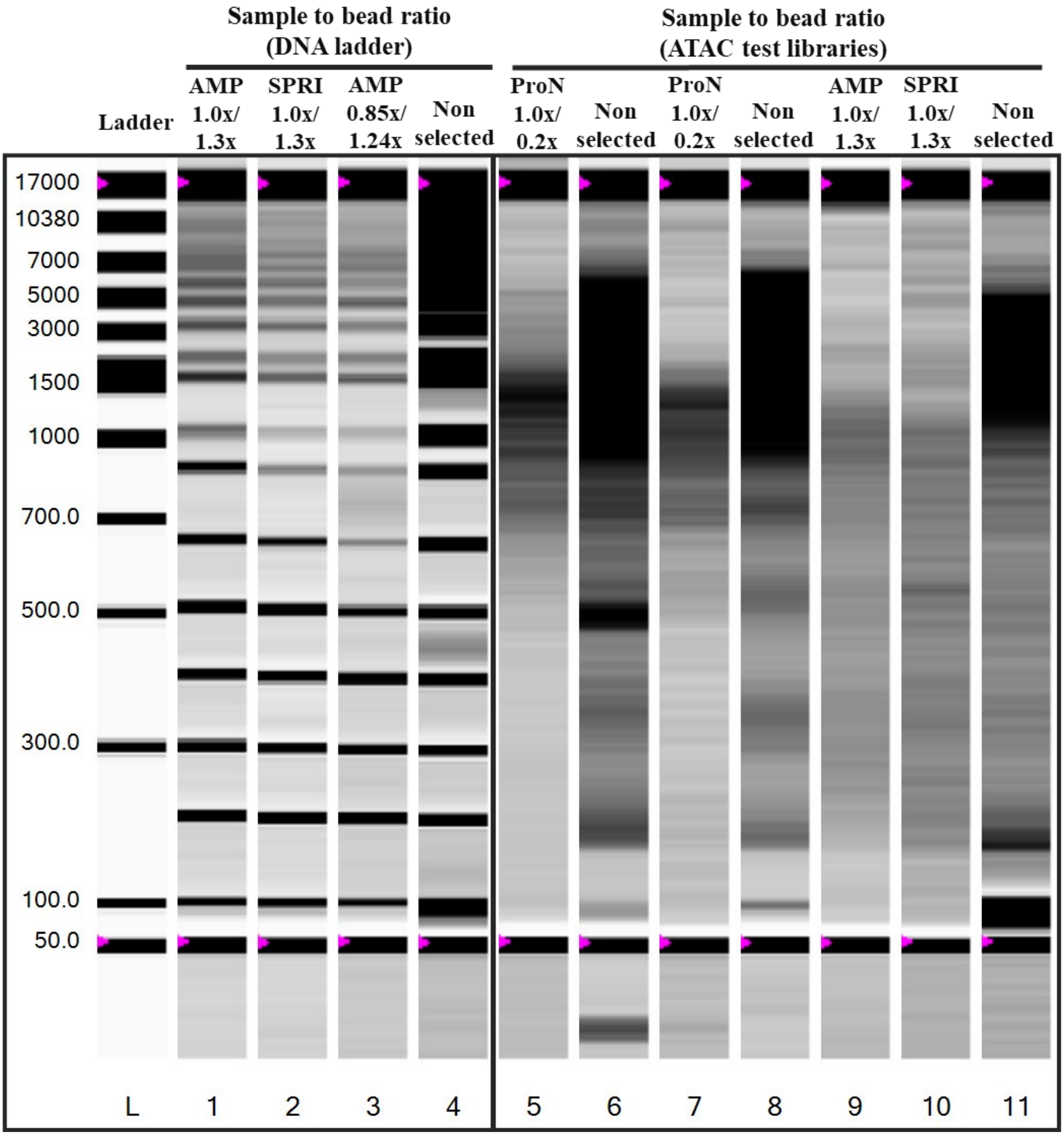


Figure 3-17 | Double size selection of NEB 1kb plus DNA ladder (left) and ATAC-seq test libraries (right) with varying reagent to sample ratios from Promega ProNex, AMPure XP or SPRIselect (Beckman Coulter Life Sciences) targeting left- and right-side selection to remove undesired DNA fragments smaller and larger than the target size.

Size-selected DNA Markers or ATAC-seq libraries were electrophoresed on an Experion Automated Electrophoresis System. The last lane of the gel for each section is labelled Non selected and contains input DNA without size selection.

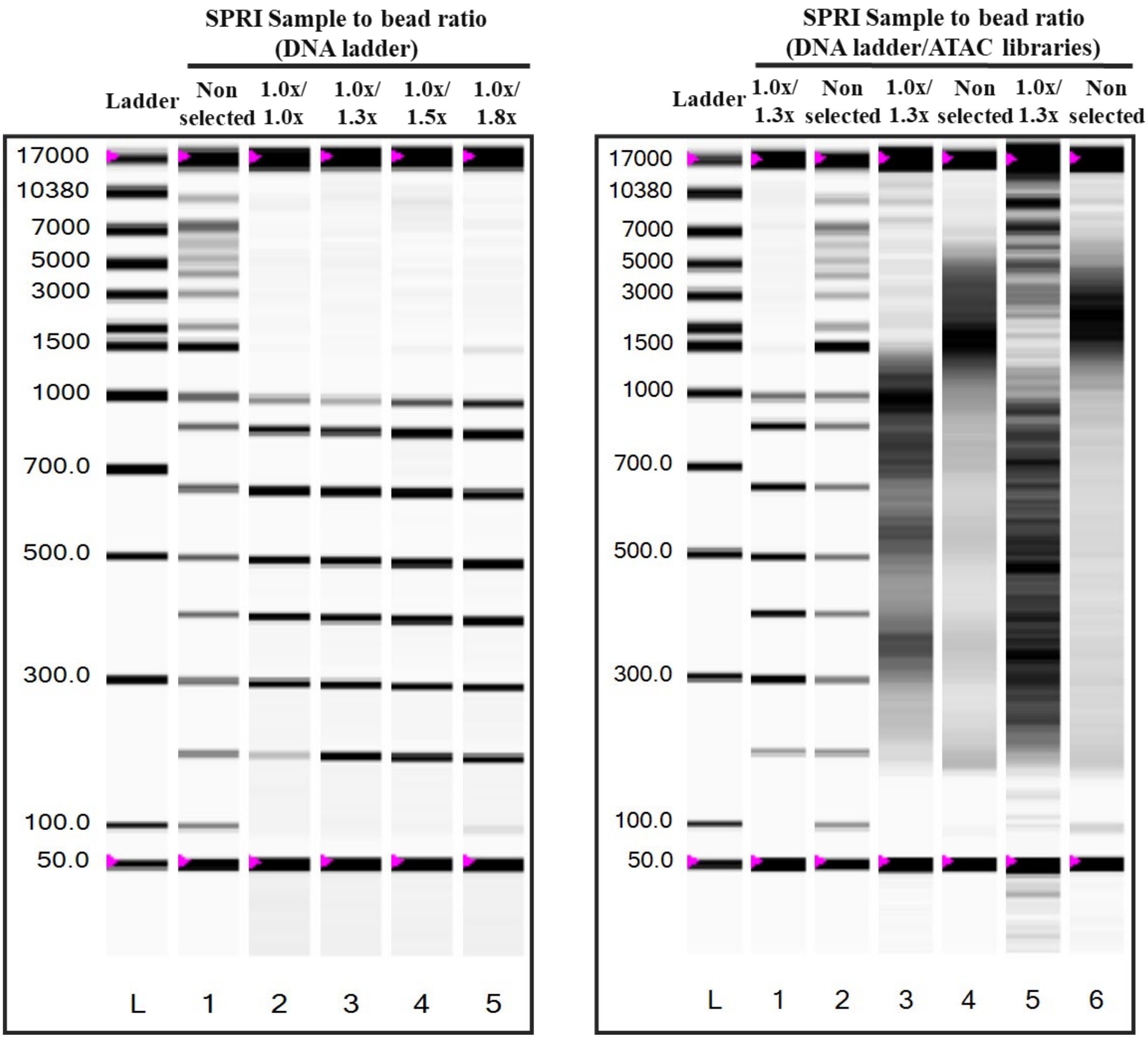


Figure 3-18 | Double size selection of NEB 1kb plus DNA ladder or ATAC-seq test libraries with varying ratios of SPRIselect (Beckman Coulter Life Sciences) volume to sample volume.
The varying ratios of SPRIselect volume to sample volume attempted here targets the right side selection to remove undesired DNA fragments larger than the target size. Size-selected DNA Markers or ATAC-seq libraries were electrophoresed on an Experion Automated Electrophoresis System.

3.4.2.4 Choice of purification kits

I also explored the impact of different commercial library purification kits - Qiagen MinElute PCR Purification and Zymo Research DNA Clean & Concentrator-5 Kit, on the complexity and fragment size distribution of ATAC-seq libraries. As original standard ATAC-seq methodology[25] recommends purification using MinElute kit, whereas Omni ATAC-seq[41] uses the latter it was of interest to determine if one led to better improvement on overall data quality than the other. To examine this, ATAC-seq libraries were purified pre- and post- PCR amplification and assessed using qPCR (Figure 7-2) and the Experion electrophoresis system (Figure 7-3), respectively. Both kits resulted in comparable, highly complex libraries (Figure 7-2) with similar distribution of fragment sizes (Figure 7-3).

3.4.2.5 Tagmentation time

In order to determine the optimal incubation time for ATAC-seq tagmentation reaction of cryopreserved CD4$^+$ T cells, 30 and 45 minutes of incubation were tested on 50,000 T cells isolated from thawed PBMCs obtained from a healthy adult volunteer. Three technical repeats were included for each tagmentation condition and libraries were size-selected to enrich for 100 - 800bp fragmentation interval prior to sequencing in order to eliminate discrepancies in quantification that may arise from large fragments and improve clustering efficiency. Size selection worked optimally as the targeted fragment size window was enriched (Figure 7-4; Figure 7-5). The fragment size distribution of the amplified ATAC-seq library showed a distinct, characteristic DNA laddering pattern with a periodicity of 200bp corresponding to integer multiples of nucleosomal protection as shown on the virtual gel (Figure 7-4) and electropherogram trace (Figure 7-5), which is an indication of good quality ATAC-seq libraries. In comparison with 30-minute tagmentation, the fragment size distribution of 45-minute tagmentation showed enrichment of fragments from lower molecular range (~200bp) to some certain on the Experion electrophoresis system (Figure 7-4).

Barcoded libraries were pooled and sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 16.3 million reads (± 3.7 million) per sample (detailed protocol in Section 3.3.2).

3.4.2.6 Fresh vs thawed Treg

Biobanked resources have contributed greatly to the advances in research and laboratory diagnostics to improve our understanding of health and disease. There are growing demands for such resources. Not only is the post-thaw viability or recovery important metric to their downstream applications but the recovered samples should closely reflect the physiological and biochemical state in their pre-freeze state.

In order to demonstrate and improve the feasibility of using biobanked PBMCs on the proposed genomic experiments I first processed and frozen PBMCs obtained from whole blood of healthy adult volunteers according to the standard biobanking protocol and criteria used for the T1D biobank cohort (Women's and Children's Hospital, South Australia) in order to recapitulate the conditions and the environment the cells were cryopreserved in (see Section 3.3.3 for detailed protocol). Treg cells were isolated from the freshly processed and thawed PBMC from the healthy adult volunteers (n=5), either untreated (resting) or stimulated, and subjected to ATAC-seq in an attempt to identify any impact freezing may have on the chromatin structure at steady state and in response to stimulation, which is unlikely to be an issue as ATAC-seq has been used previously on biobanked material [30, 41, 182].

The fresh and thawed ATAC-seq libraries demonstrated similar level of library complexity (Figure 3-19; Figure 3-20) and nucleosomal binding pattern (Figure 3-21; Figure 3-22) for most donors, as shown on the qPCR amplification plots and Experion electropherograms. Stimulation results in overall higher distinction of DNA laddering pattern at a periodicity of about 200bp compared with resting libraries. The fragment size distribution of thawed Treg libraries from Donor 4 (Figure 3-22) lacked the characteristic nucleosomal binding pattern which could be indicative of overdigestion. This could be attributed to the use of a new batch of tagmentation reagents in the library generation and the associated lot-to-lot variation could potentially contribute to discrepancies in digestion pattern. This observation has led to the adoption of 30-minute tagmentation using the new lot of Tn5 in the subsequent ATAC-seq experiments (Chapter 5) to avoid overdigestion. Barcoded resting and stimulated Treg libraries from Donor 1, 2 and 3 were sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 37.2 million reads (± 6 million) per sample.

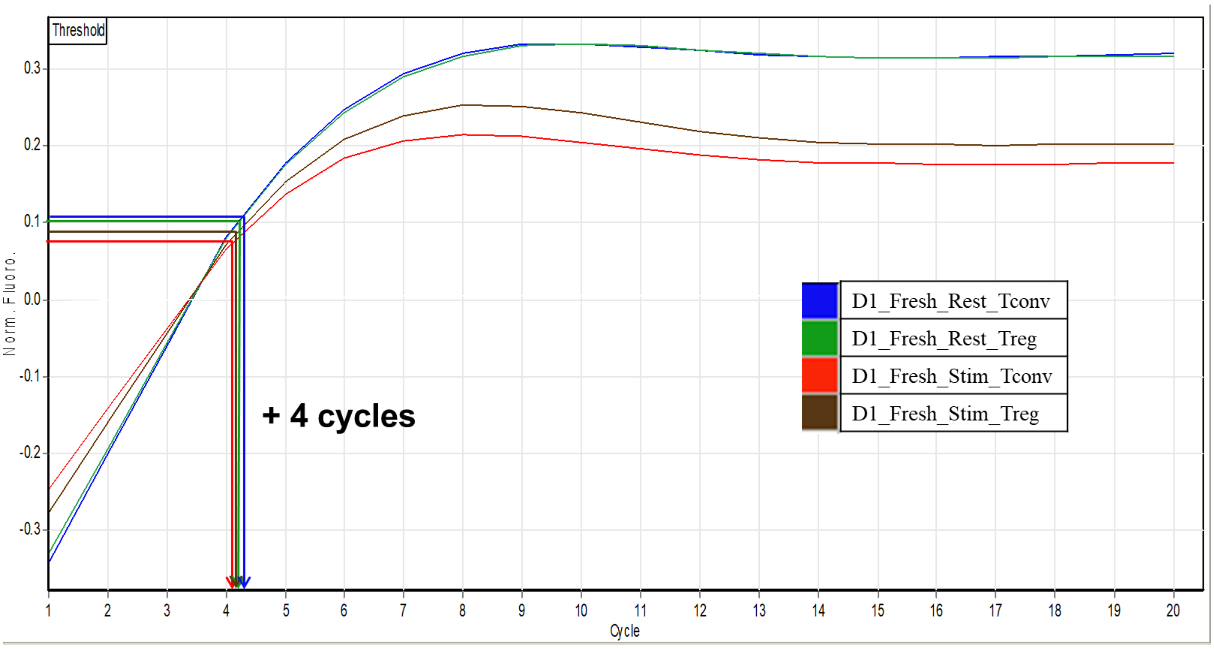


Figure 3-19 | Representative amplification plot demonstrating the number of additional PCR cycles to perform for four tagmentation reactions prepared from fresh T cells from a healthy adult donor.

Tconv or Treg cells were either left untreated (resting) or stimulated using Dynabeads Human T-Expander CD3/CD28 for 48 hours prior to transposition. Samples from a total of 4 donors were included in this part of work, however, only data from one donor is shown as representative.

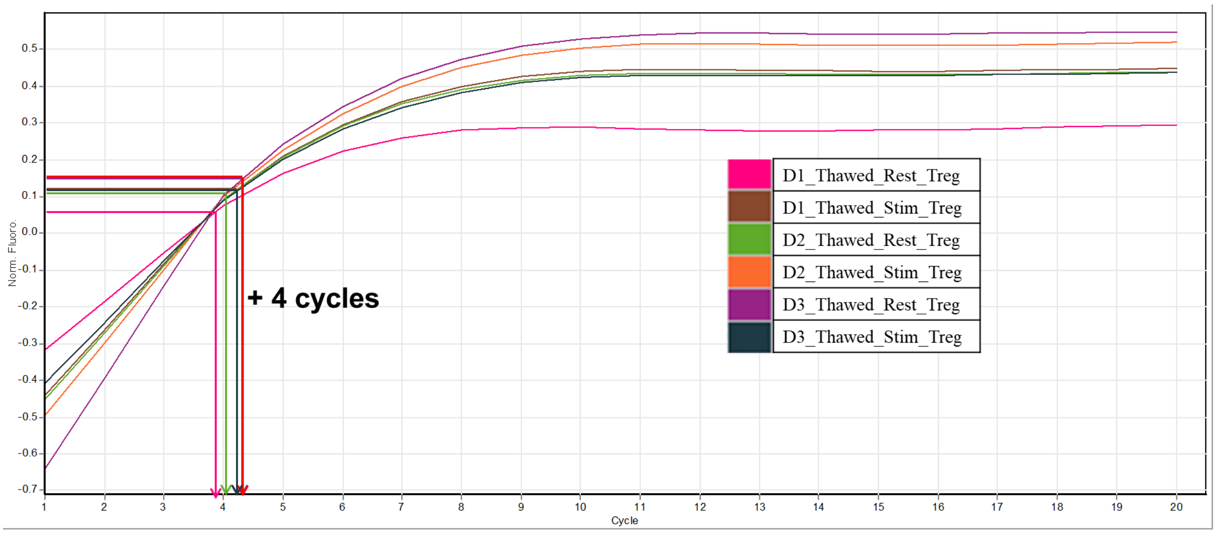


Figure 3-20 | Representative amplification plot demonstrating the number of additional PCR cycles to perform for six tagmentation reactions prepared from thawed Treg cells from three healthy adult donors.

Treg cells were either left untreated (resting) or stimulated using Dynabeads Human T-Expander CD3/CD28 for 48 hours prior to transposition.

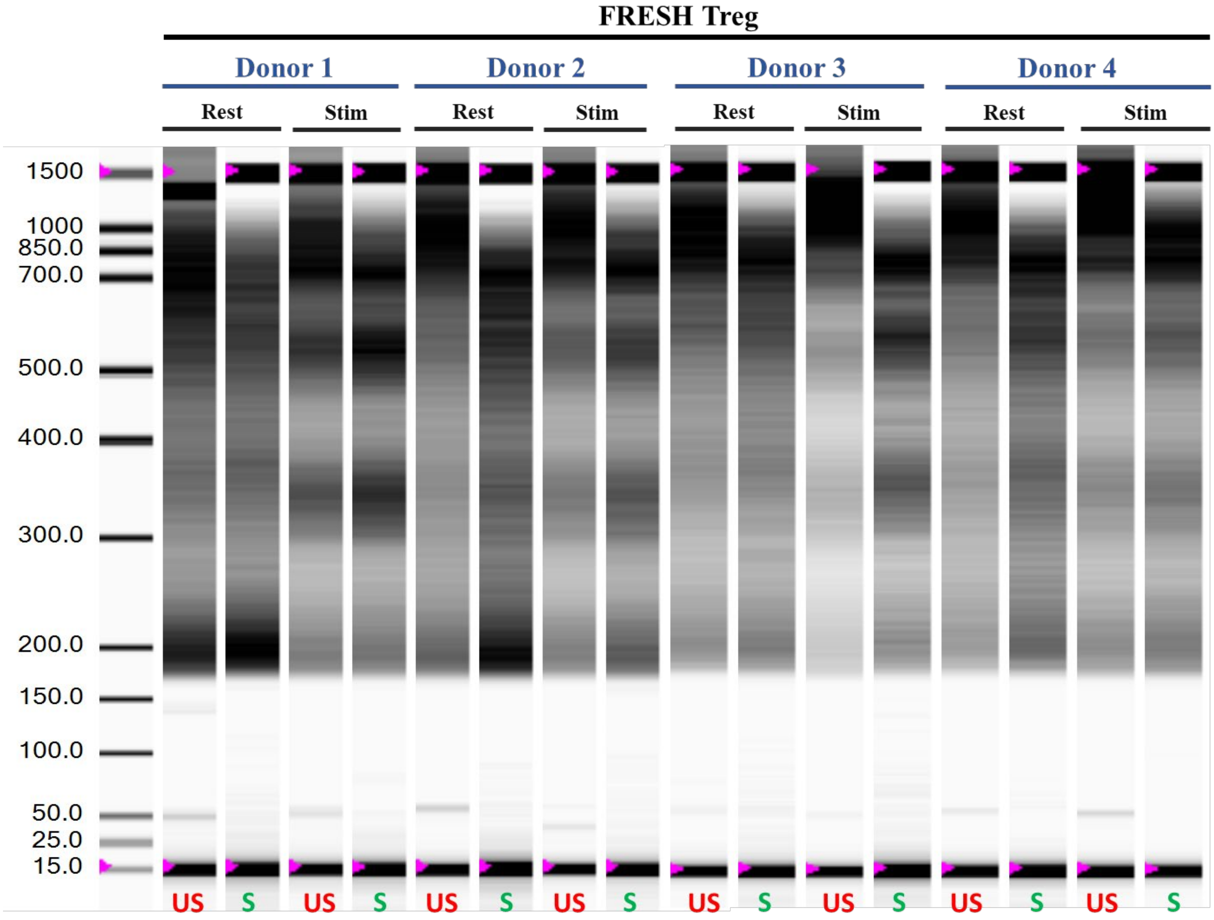


Figure 3-21 | Virtual gel demonstrating fragment size distribution of amplified fresh Treg ATAC-seq libraries from four 4 donors before and after size selection, determined by the Experion Automated Electrophoresis System.

Treg cells were either left untreated (resting) or stimulated using Dynabeads Human T-Expander CD3/CD28 for 48 hours prior to transposition. Size selection was performed using SPRIselect beads (Beckman Coulter Life Sciences). Matching PCR amplification profiles (pre-size selection) are shown in Figure 3-19. *S, size-selected; US, unselected.*

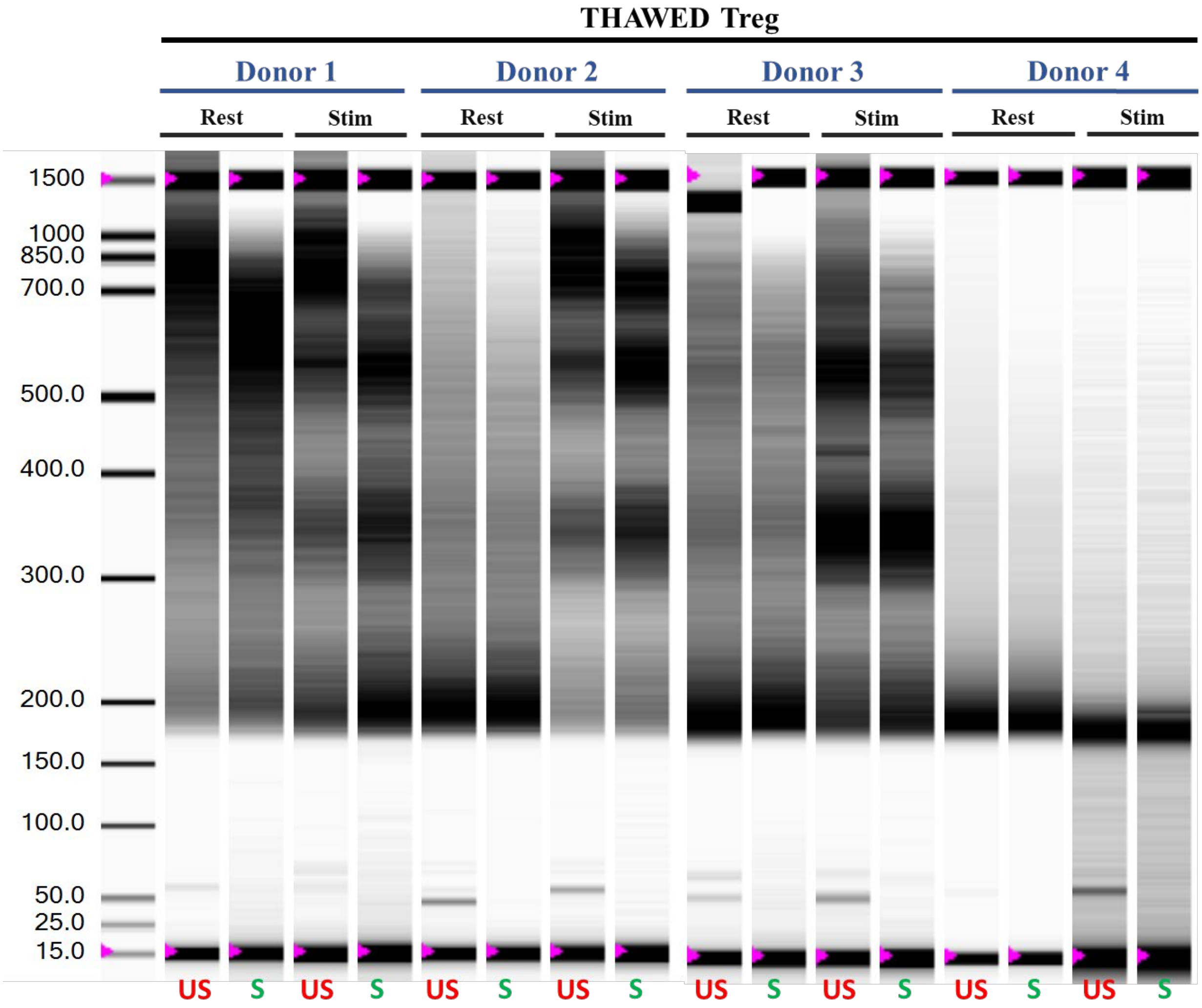


Figure 3-22 | Virtual gel demonstrating fragment size distribution of amplified thawed Treg ATAC-seq libraries from four 4 donors before and after size selection, determined by the Experion Automated Electrophoresis System.
Treg cells were either left untreated (resting) or stimulated using Dynabeads Human T-Expander CD3/CD28 for 48 hours prior to transposition. Size selection was performed using SPRIselect beads (Beckman Coulter Life Sciences). Matching PCR amplification profiles (pre- size selection) are shown in Figure 3-20. *S, size-selected; US, unselected.*

3.4.2.7 Isolation of RNA from ATAC-seq lysis reaction

I also explored the plausibility of extracting both the accessibility and transcriptome information simultaneously from a single ATAC-seq reaction using 50,000 cells. The ability to extract transcriptional signatures from ATAC-seq lysate reaction from the same population of 50,000 cells allows the capability to accurately correlate accessibility signal to transcript levels as the impact of cellular heterogeneity is minimal. It also makes the assays amenable for studies that use limited clinical samples like biobanked paediatric materials, especially on rare subset populations like Treg cells.

To examine this, I first compared different RNA isolation and purification methods in recovering intact RNA population from the ATAC-seq lysate reaction. Detailed workflow and processes for isolation of RNA from supernatant (SN) fractions from Omni ATAC-seq lysis reaction (ATAC-SN) are described in Section 2.3.1, Figure 3-5, Section 3.3.4 and Figure 3-23 (a). Briefly, each ATAC-SN sample (1mL SN per sample) was divided into five 1.5-mL Eppendorf tubes (each containing ~200µL of supernatant) and TRIzol™ LS Reagent was added to homogenise the supernatant samples. Phase separation of ATAC-SN homogenate was performed using chloroform and the upper, colourless, aqueous phase (~2ml per sample) containing RNA was harvested.

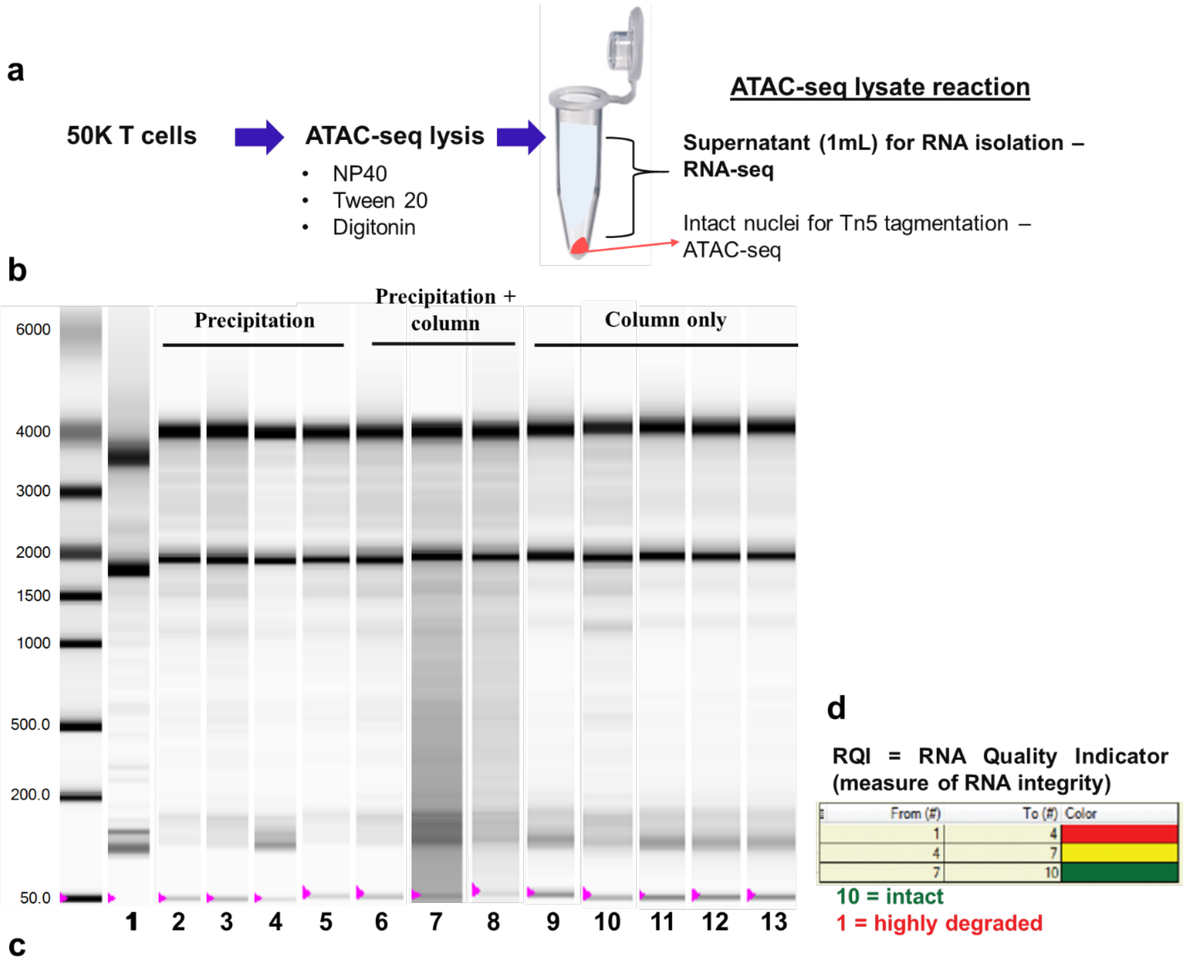
For purification by traditional precipitation, 100% isopropanol (0.5mL per 0.75mL TRIzol™ LS Reagent) was added to the harvested aqueous phase (~2mL per sample) to precipitate the total RNA by centrifugation following 10-minute incubation, which was then washed in 75% ethanol.

For purification by precipitation and column clean-up, 100% isopropanol was added to the harvested aqueous phase, the mixture was incubated for 10 minutes and loaded onto the commercial RNA clean-up columns from - Qiagen mirRNeasy (Mini, Micro) or Zymo Research Direct-zol (Mini) kits, for RNA purification according to the manufacturer's protocol.

For purification by columns clean-up only, 100% ethanol was added to the harvested aqueous phase and the mixture was loaded directly onto the commercial RNA clean-up columns from - Qiagen mirRNeasy (Mini, Micro) or Zymo Research Direct-zol (Mini) kits, for RNA purification according to the manufacturer's protocol.

The integrity and yield of the isolated RNA were determined by the Experion Automated Electrophoresis System.

Figure 3-23 (b; c) shows all three RNA isolation and purification methods recovered high integrity and comparable yield of RNA. Qiagen mirRNeasy (Micro) kit was picked as the purification method of choice for recovery of RNA from ATAC-seq supernatant fractions (ATAC-SN) as the workflow demonstrated a more streamlined process with reduced hands-on time and reproducible recovery of high RNA quality and yield (Figure 3-23, c; Lane 11-13).



From (#)	To (#)	Color
1	4	🟥
4	7	🟨
7	10	🟩

c

Lane	RNA isolation method	RNA conc. (pg/uL)	RQI
1	Experion Total RNA control	161053	8.6
2	Precipitation Rep 1	2036.92	9.1
3	Precipitation Rep 2	2261.49	8.8
4	Precipitation Rep 3	3239.92	9.1
5	Precipitation Rep 4	1963.26	9.1
6	Precipitation + Qiagen miRNeasy (Mini) column	2022.51	8.2
7	Precipitation + Zymo Direct-zol (Mini) column Rep 1	5329.00	7.7
8	Precipitation + Zymo Direct-zol (Mini) column Rep 2	5829.15	7.4
9	Qiagen miRNeasy (Mini) column Rep 1	2455.00	9.2
10	Qiagen miRNeasy (Mini) column Rep 2	2022.51	8.2
11	Qiagen miRNeasy (Micro) column Rep 1	2455.56	9.3
12	Qiagen miRNeasy (Micro) column Rep 2	2404.29	9.2
13	Qiagen miRNeasy (Micro) column Rep 3	2367.98	9.2

Figure 3-23 | (a) Experimental workflow demonstrating simultaneous profiling of chromatin accessibility and gene expression from a single ATAC-seq reaction using 50,000 lysed stimulated T cells. (b-c) Recovery of RNA from ATAC-seq supernatant fractions using different extraction and purification approaches/commercial kits. The evaluation of RNA integrity and concentration was performed on an Experion Automated Electrophoresis System (RNA HighSens Analysis). (d) Integrity of a total RNA sample is graded with a scale of 1 (lowest integrity, most degraded; red) to 10 (highest integrity, least degraded; green). ***RQI, RNA quality indicator.***

To determine if the transcript levels of RNA populations recovered from the ATAC-seq lysis reaction (ATAC-SN method) closely recapitulates that isolated from the whole cell lysate (conventional method), RNA-seq libraries (n=4 donors) were generated from RNA material purified from whole Treg cells or supernatant fractions from Omni ATAC-seq lysis reaction (ATAC-SN) (Figure 3-24) and subjected to next generation sequencing (Figure 3-25). The RNA recovered from ATAC-SN reaction exhibited comparable quality (as indicated by RNA Integrity score) to that recovered from the matching whole cell lysate for stimulated T cells across all donors (Figure 3-24). However, the RNA recovered from ATAC-SN of resting T cells from most donors was partially degraded, registering an average RIN score of 6.4 (Figure 3-24). The fragment size distribution of Treg RNA-seq libraries constructed from RNA samples prepared from ATAC-SN or whole cells showed a narrow distribution with a peak size of approximately 300bp, which is a typical size distribution for standard RNA-seq libraries constructed with NEBNext Ultra II Directional RNA Protocol (Section 3.3.4), except for two libraries prepared from ATAC-SN of resting Treg cells. The RNA-seq libraries prepared from resting ATAC-SN of Donor 2 and Donor 3 lacked a distinct fragment peak distribution and a smeared pattern was observed, potentially attributable to partially degraded starting RNA input. As a result of subpar quality observed in RNA-seq libraries prepared from ATAC-SN of resting Treg cells, they were not sequenced. The remaining RNA-seq libraries were barcoded, pooled and sequenced on a paired-end 150-cycle Illumina Hiseq sequencer (Illumina) to an average read depth of 37.5 million reads (± 16 million) per sample.

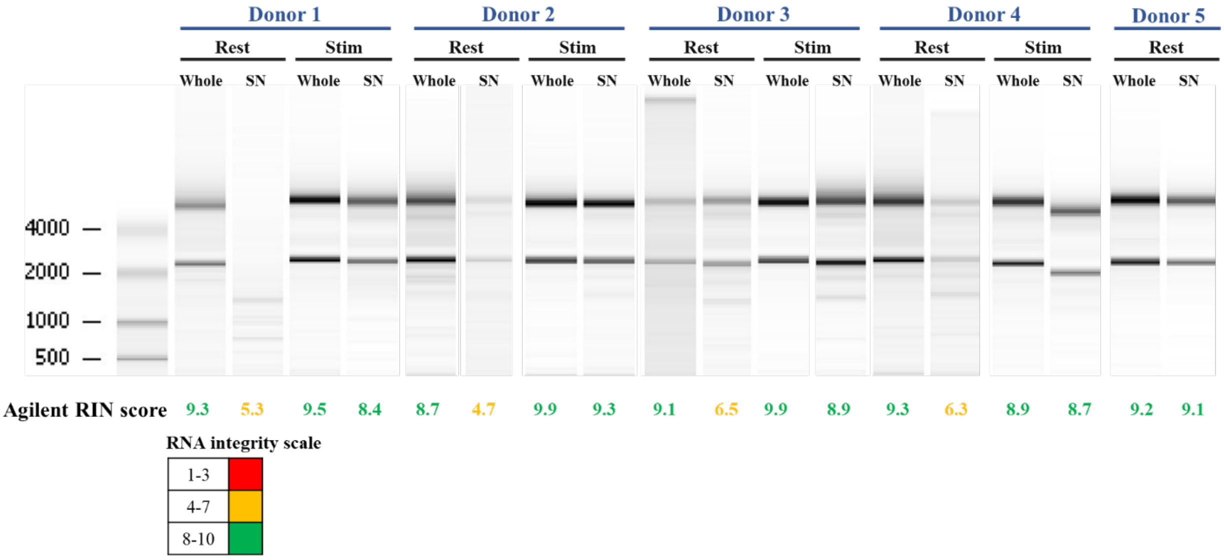


Figure 3-24 | Recovery of RNA from Treg ATAC-seq supernatant fractions using Qiagen miRNeasy Micro Kit. RNA was recovered from Treg ATAC-seq supernatant fractions collected from healthy adult donors (matching ATAC-seq profiles in Figure 3-22) for comparison with RNA isolated from whole Treg cells (conventional RNA isolation method). The assessment of RNA integrity was performed on an Agilent 2100 Bioanalyzer system (RNA HighSens Analysis).

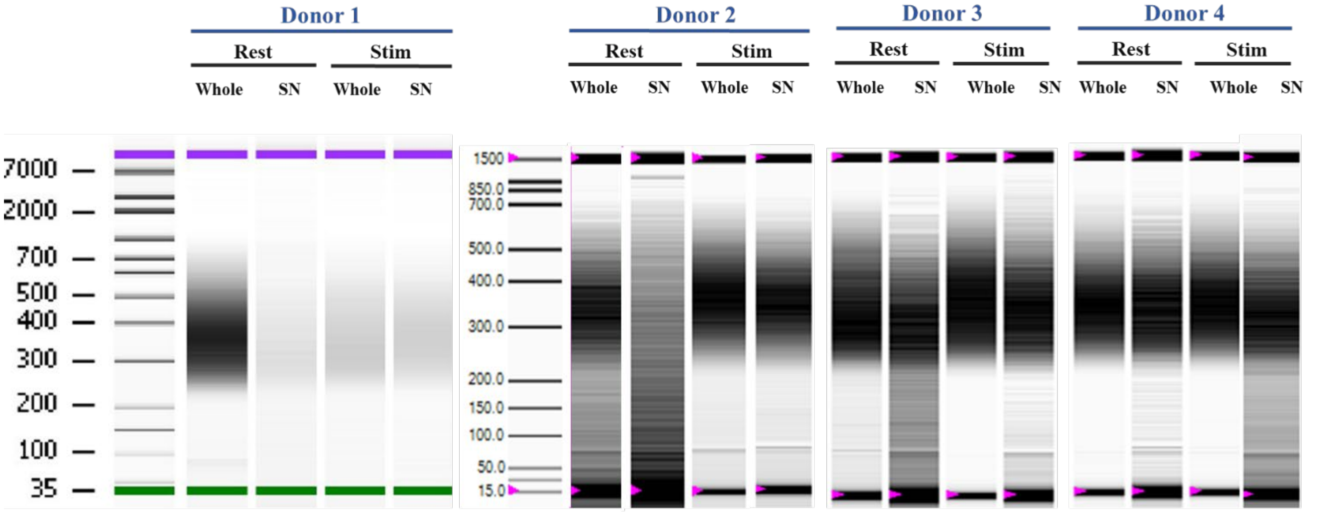


Figure 3-25 | The fragment size distribution of Treg RNA-seq libraries constructed from RNA samples prepared from ATAC-seq supernatant fractions or whole cells (matching RNA samples in Figure 3-24 and ATAC-seq samples in Figure 3-22).
The assessment of library QC was performed on either an Agilent 2100 Bioanalyzer system (High Sensitivity DNA Assay) or Experion Automated Electrophoresis System (DNA 1K Analysis).

3.4.3 Exploratory analysis of sequencing data:

This section describes the exploratory analyses performed for three sequencing experiments in relation to the optimisation of ATAC-seq protocol. First part presents the exploratory findings from ATAC-seq of 30- vs. 45-minute tagmentation, followed by ATAC-seq of fresh vs. thawed Treg and lastly, RNA-seq of whole vs. ATAC-SN lysate. Quality control assessment measures such as sequencing read yield, per base sequence quality, proportion of read duplicates, alignment metrics, library complexity and proportion of mitochondrial contamination are included in this section.

3.4.3.1 Tagmentation

ATAC-seq libraries of 30- and 45-minute tagmentation (n=3 technical repeats) (Figure 7-4) were sequenced to an average read depth of 16.3 million reads (± 3.7 million) per sample. The lowest yield library was 30-minute tagmentation from Rep 2 (11,030,634 reads) and the highest being 45-minute tagmentation from Rep 2 (20,467,119 reads). As a typical ATAC-seq library has a large window of highly diverse fragment sizes ranging from 100bp to >1000bp, the discrepancy in sequencing depth could be driven by varying proportion on the representation of short- and long-fragments between libraries, which could inaccurately skew the concentration of a library during quantification. Overall the 30- and 45-minute tagmentation libraries demonstrated comparable level of sequencing quality in metrics such as per base sequence quality (Figure 3-27; Figure 3-28), which provides the distribution of quality scores at each position in the read, sequence duplication levels (Figure 3-29) as well as alignment performance (Figure 3-30).

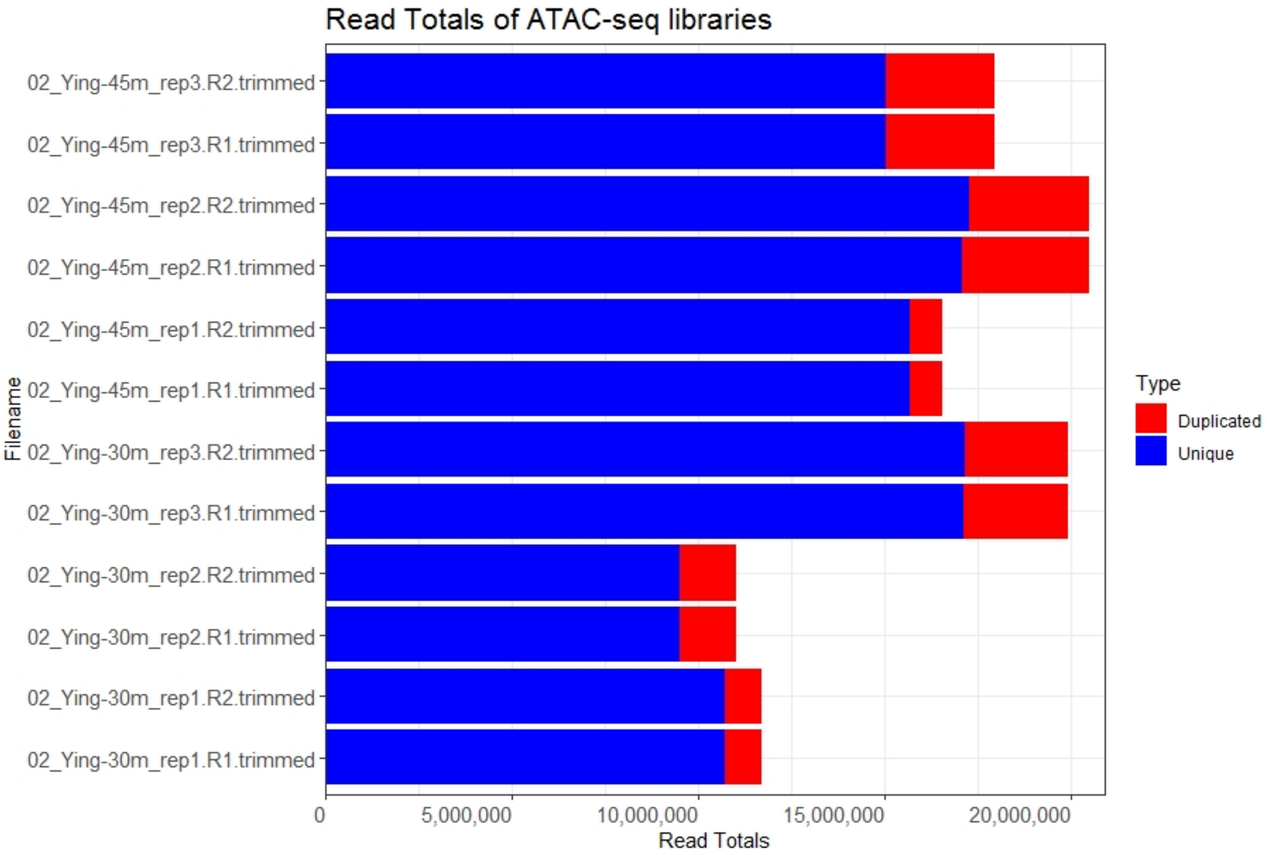


Figure 3-26 | The total number of paired-end reads of Tconv ATAC-seq libraries post adapter trimming, constructed from 30 and 45 minutes of tagmentation incubation at 37°C.
Three technical replicates were included for each tagmentation condition.

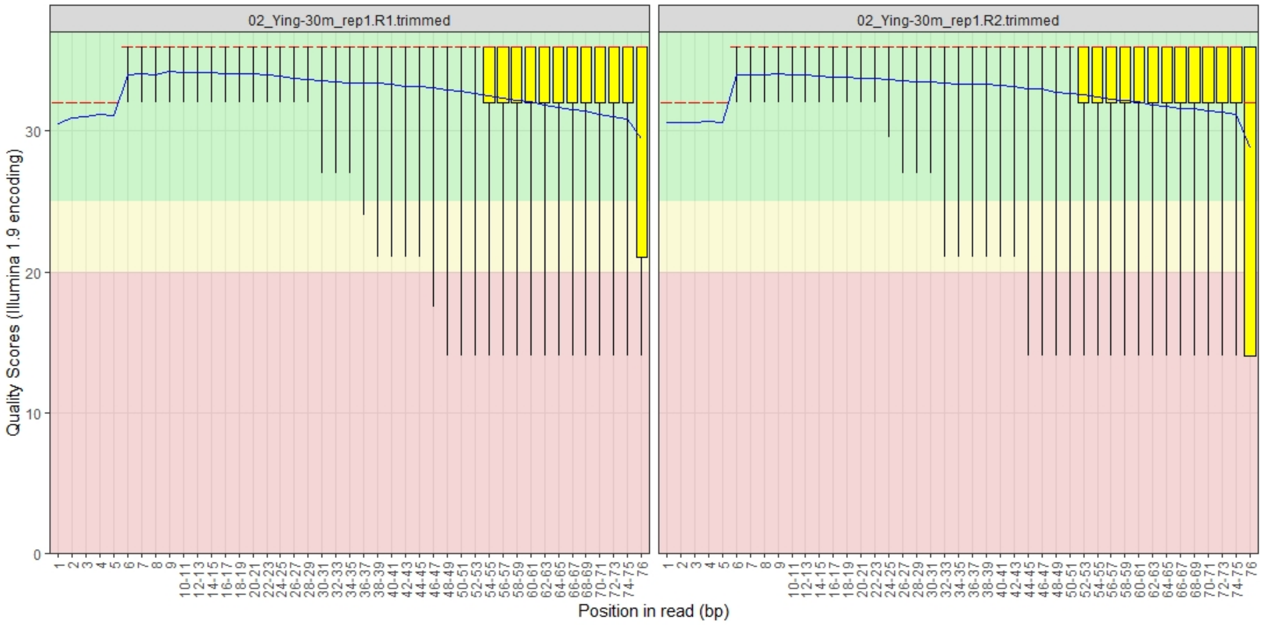


Figure 3-27 | Representative per base sequence quality plot for one paired-end Tconv ATAC-seq library (30-minute tagmentation).

This QC metric provides the distribution of quality scores at each position in the read across all reads in both orientation of paired-end sequencing.

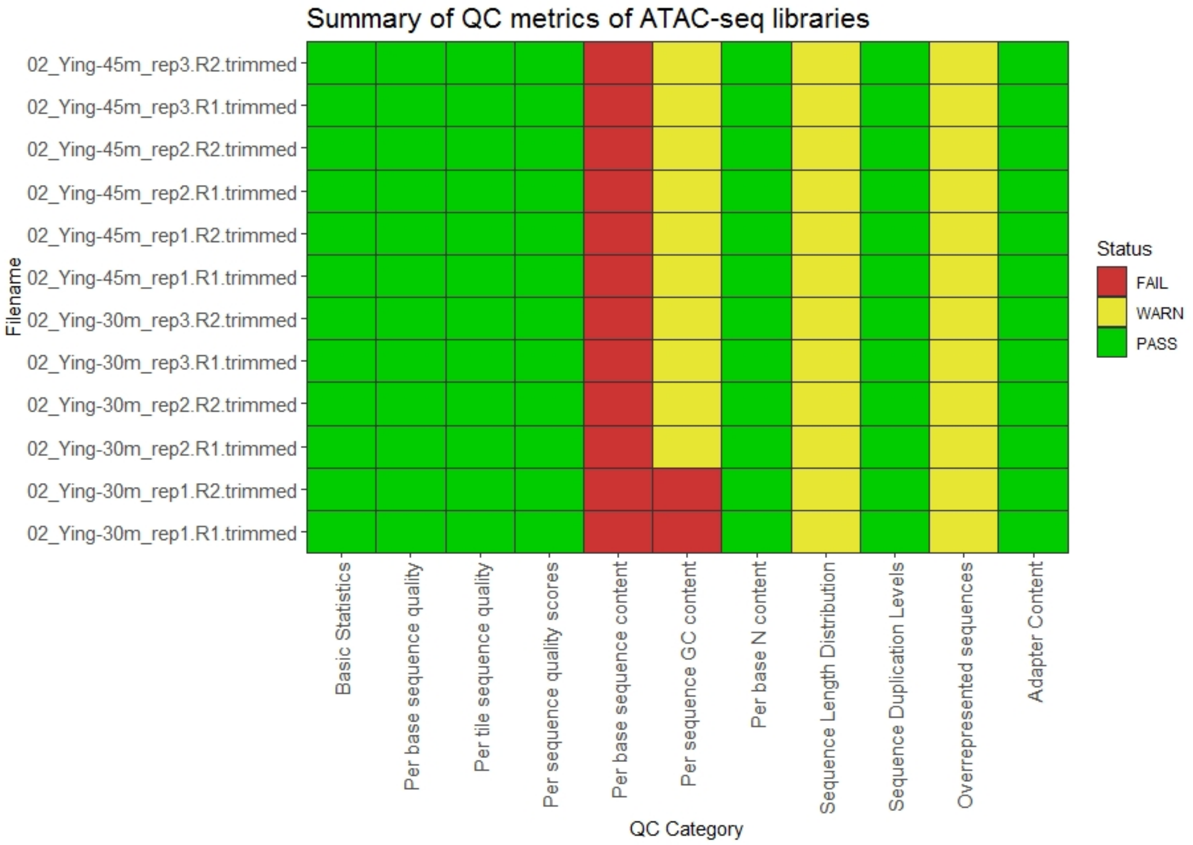


Figure 3-28 | Summary of QC metric modules for Tconv ATAC-seq libraries prepared from 30 and 45 minutes of tagmentation reactions.

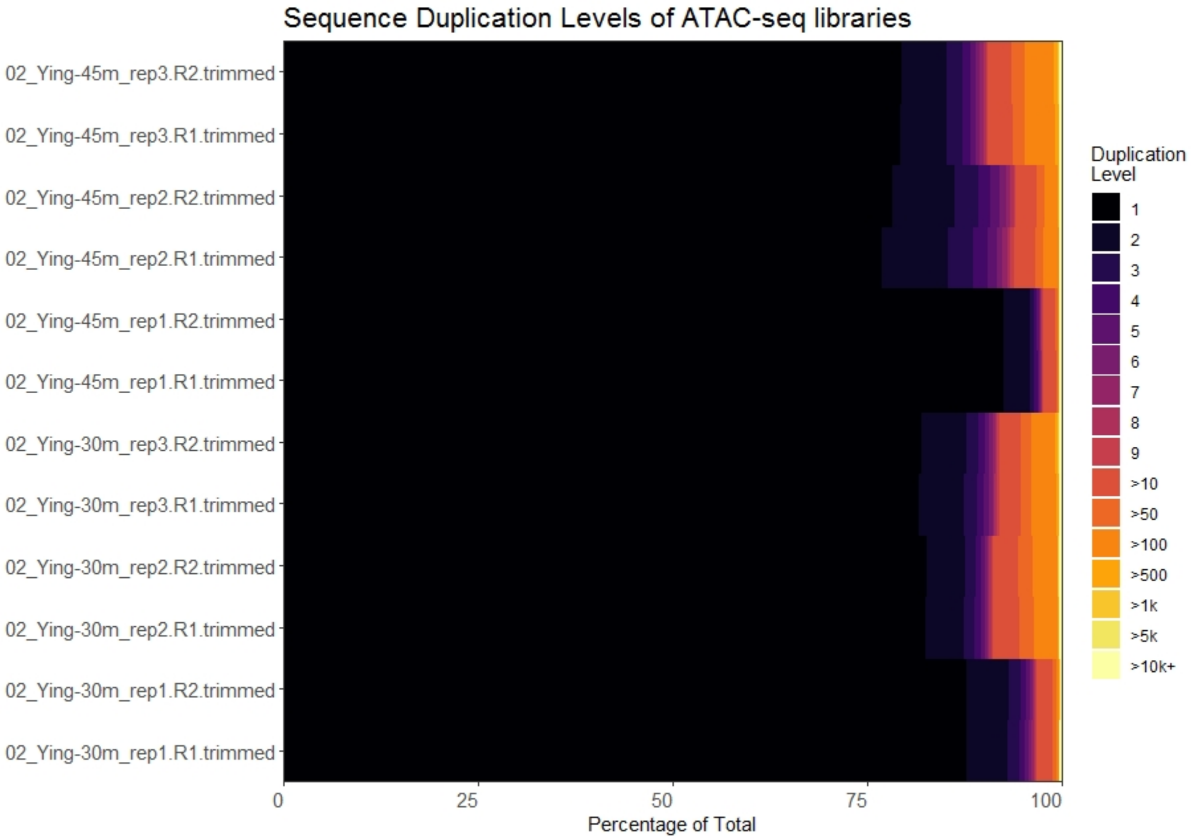


Figure 3-29 | Heatmap showing the duplication levels of 30- and 45- minute tagmentation ATAC-seq libraries.

The values are expressed as percentages of total instead of read numbers to ensure comparability across libraries.

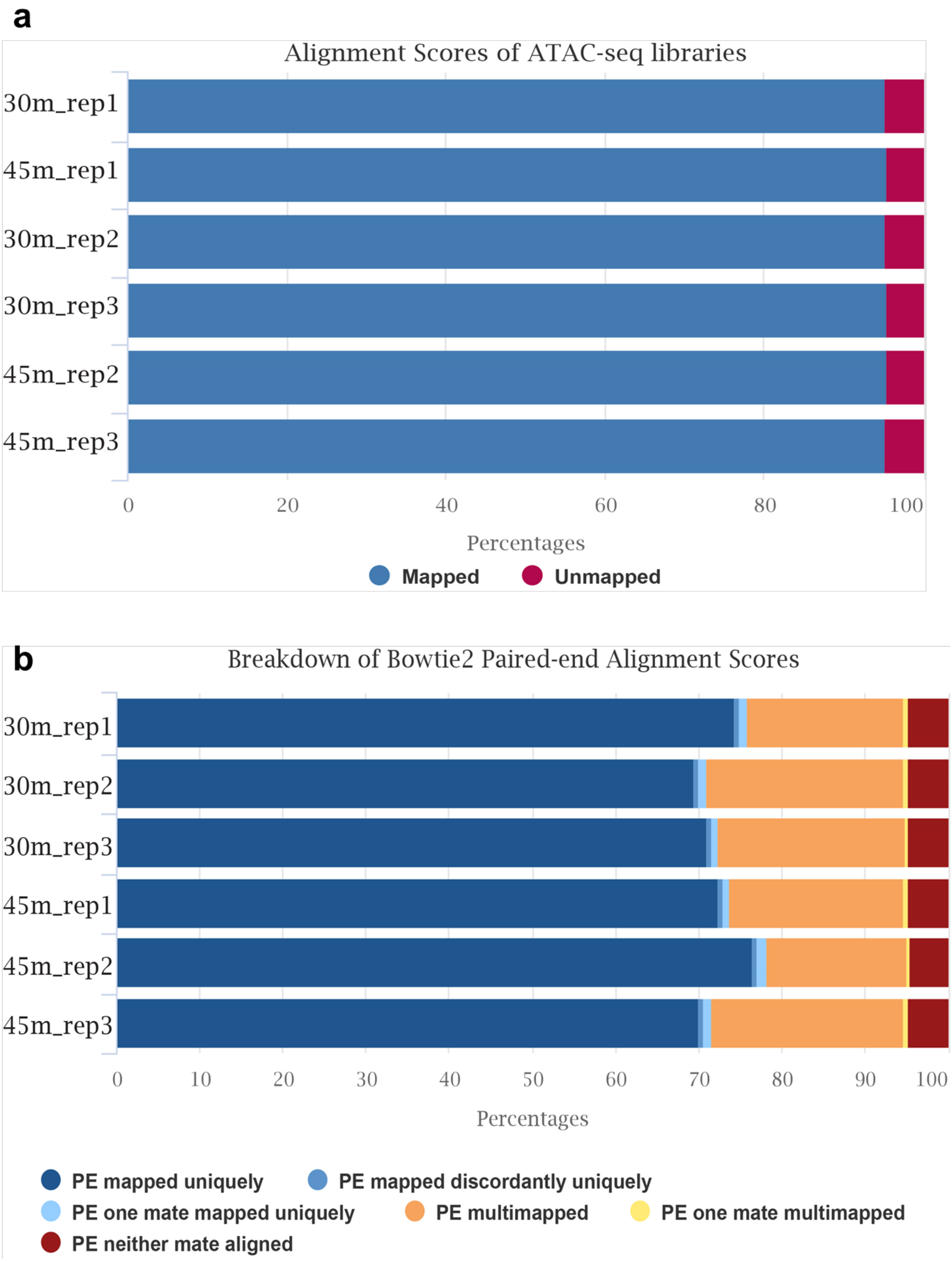


Figure 3-30 | (a) Alignment scores of 30- and 45- minute tagmentation Tconv ATAC-seq libraries. (b) Breakdown of alignment metrics for mapped ATAC-seq reads from (a), including percentage of uniquely and multi-mapped reads.

3.4.3.2 Fresh vs thawed

ATAC-seq libraries of fresh and thawed Treg cells (n=3 donors) (Figure 3-21, Figure 3-22) were sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 37.2 million reads (± 6 million) per sample (Figure 3-31). The highest yield library was Thawed_Treg_D3_Rest (44,145,618 reads), whereas the lowest yield was Thawed_Treg_D2_Stim (30,778,250 reads).

Overall the fresh and thawed Treg libraries demonstrated comparable level of sequencing quality as shown in representative plot of per base sequence quality (Figure 3-32), sequence duplication levels (Figure 3-33), adapter content of pre- and post-trimming which serves as validation before downstream analysis (Figure 3-34, Figure 7-6), alignment performance (Figure 3-35, Figure 3-36, Figure 3-37) as well as proportion of mitochondrial contamination (Figure 3-38). On the level of sequencing quality, the differences captured between the fresh and thawed Treg ATAC-seq libraries were minimal, though for some measures (Figure 3-33, Figure 3-38) more appreciable changes were observed between the rest and stimulated libraries for both the fresh and thawed samples.

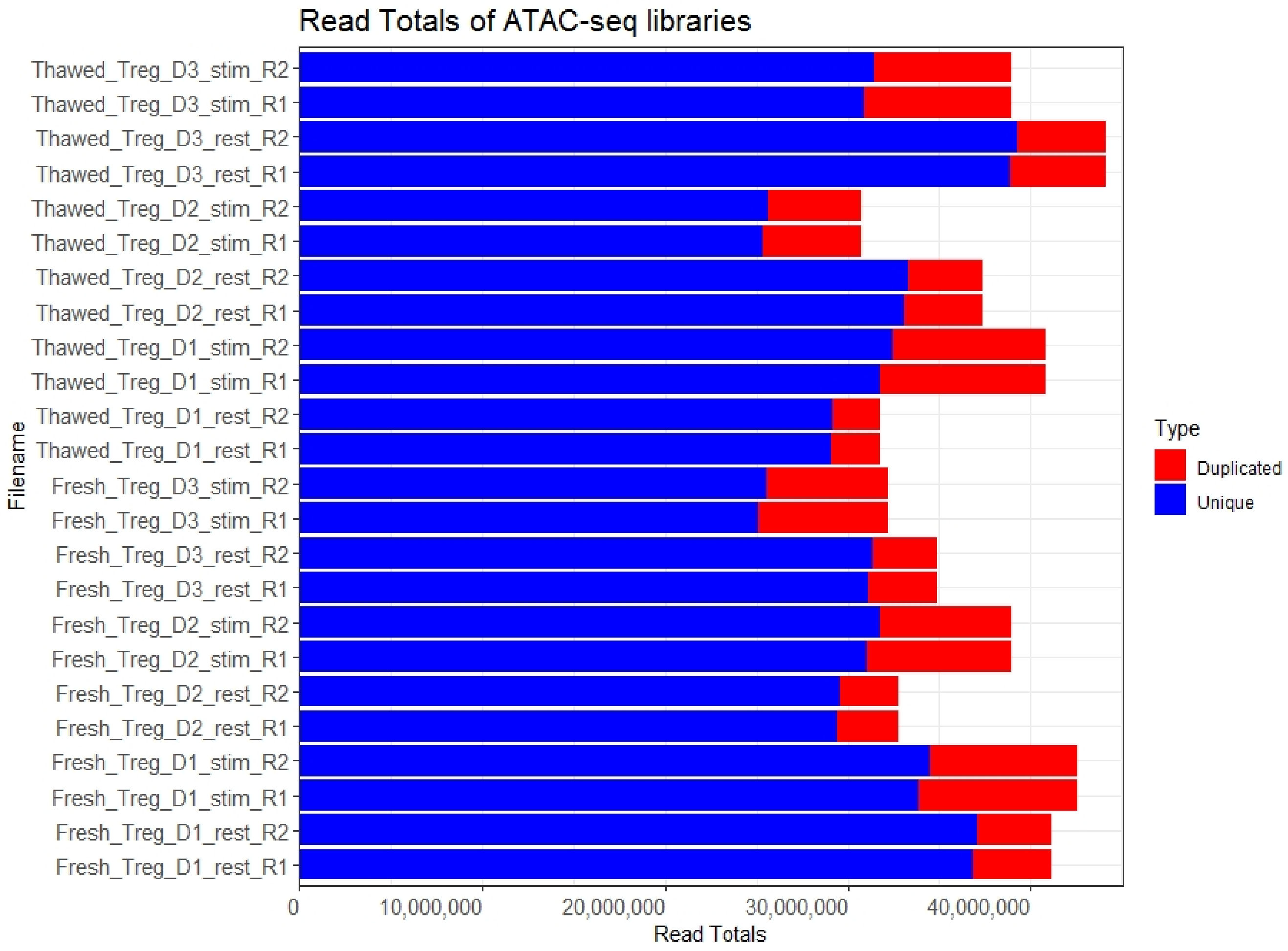


Figure 3-31 | The total number of paired-end reads of Treg ATAC-seq libraries, constructed from fresh or thawed Treg cells purified from 3 healthy adult donors.
Treg cells were either left untreated (resting) or stimulated using Dynabeads Human T-Expander CD3/CD28 for 48 hours prior to transposition for ATAC-seq library prep.

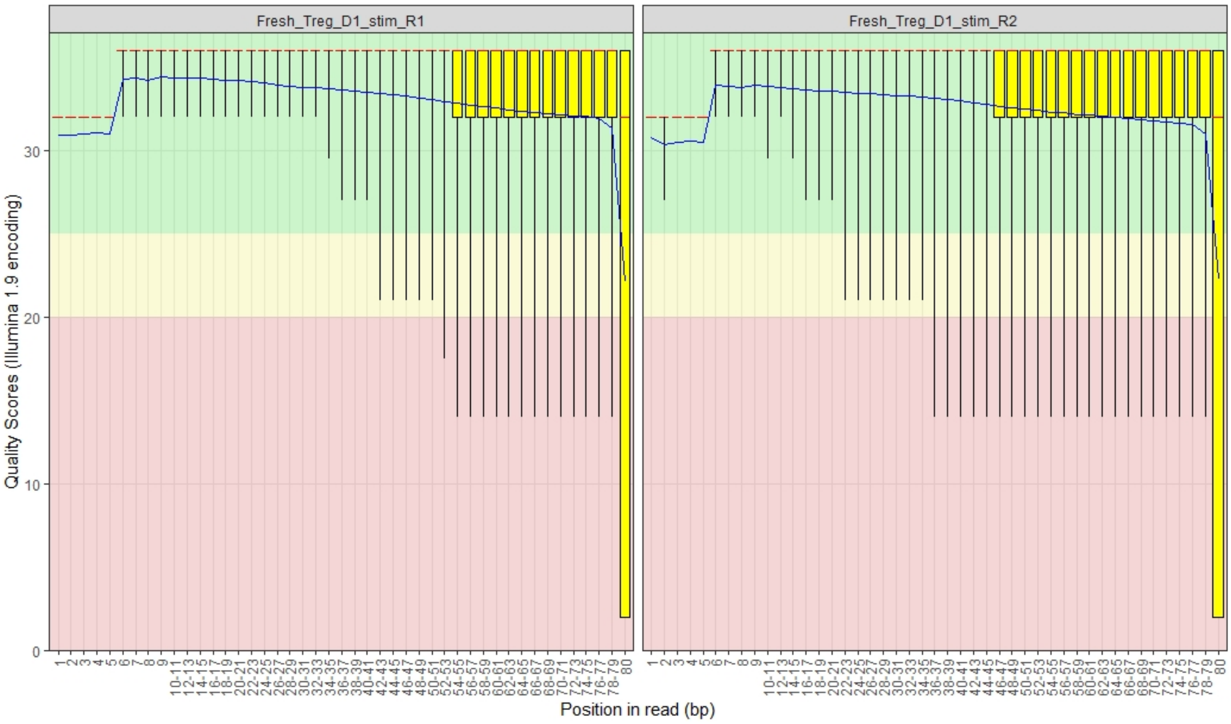


Figure 3-32 | Representative per base sequence quality plot for one paired-end Treg ATAC-seq library (Fresh).
This QC metric provides the distribution of quality scores at each position in the read across all reads in both orientation of paired-end sequencing.

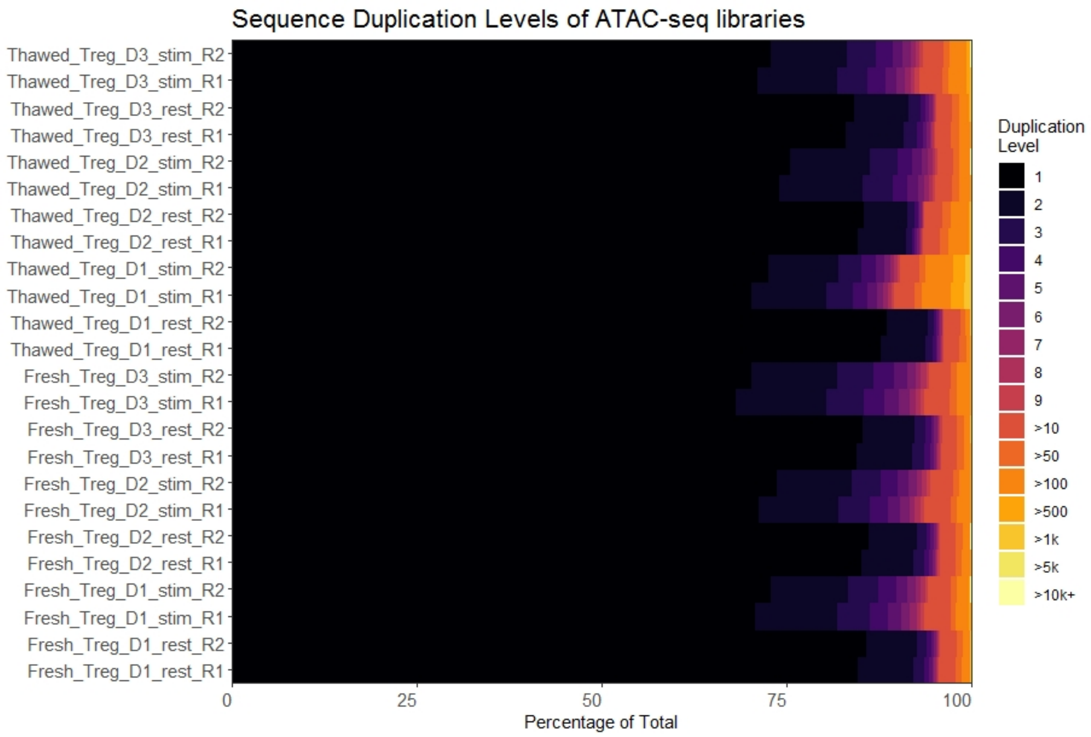


Figure 3-33 | Heatmap showing the duplication levels of fresh and thawed Treg ATAC-seq libraries.

The values are expressed as percentages of total instead of read numbers to ensure comparability across libraries.

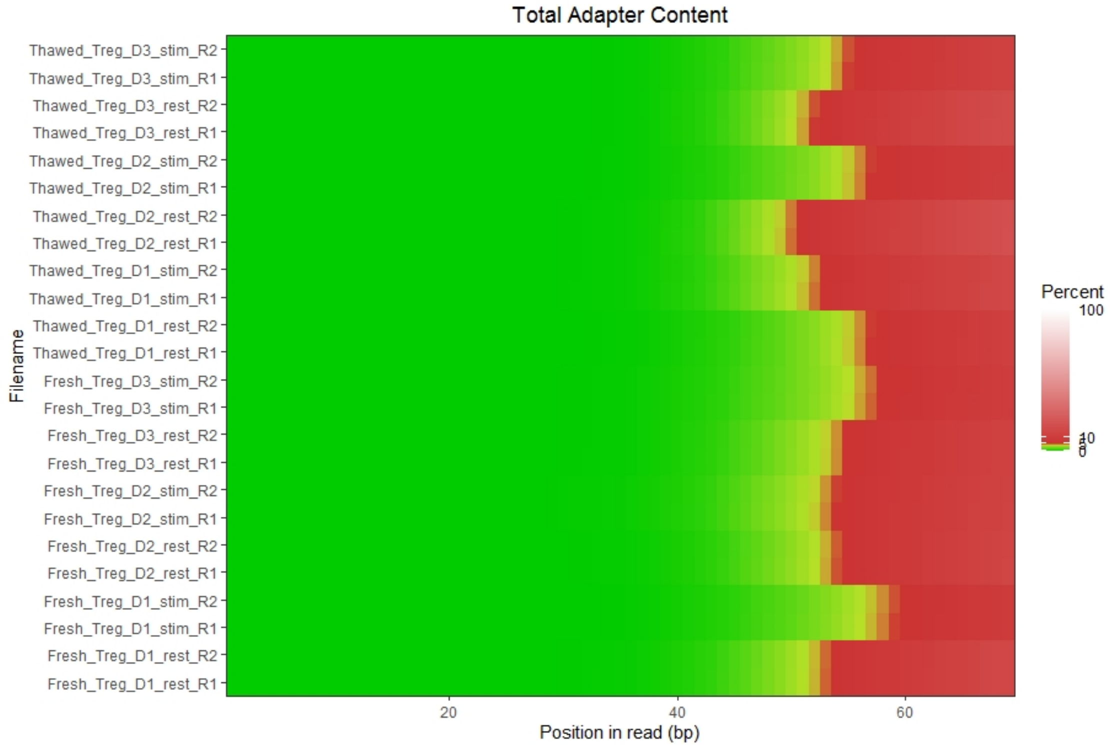


Figure 3-34 | Adapter content and position in raw sequencing reads for fresh and thawed Treg ATAC-seq libraries.

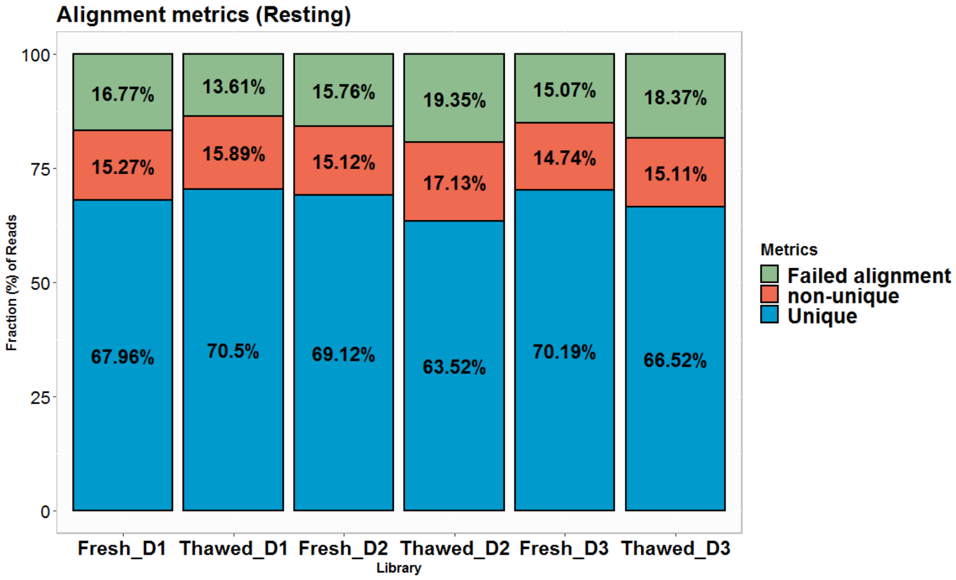


Figure 3-35 | Alignment metrics for fresh and thawed Treg ATAC-seq libraries (resting).

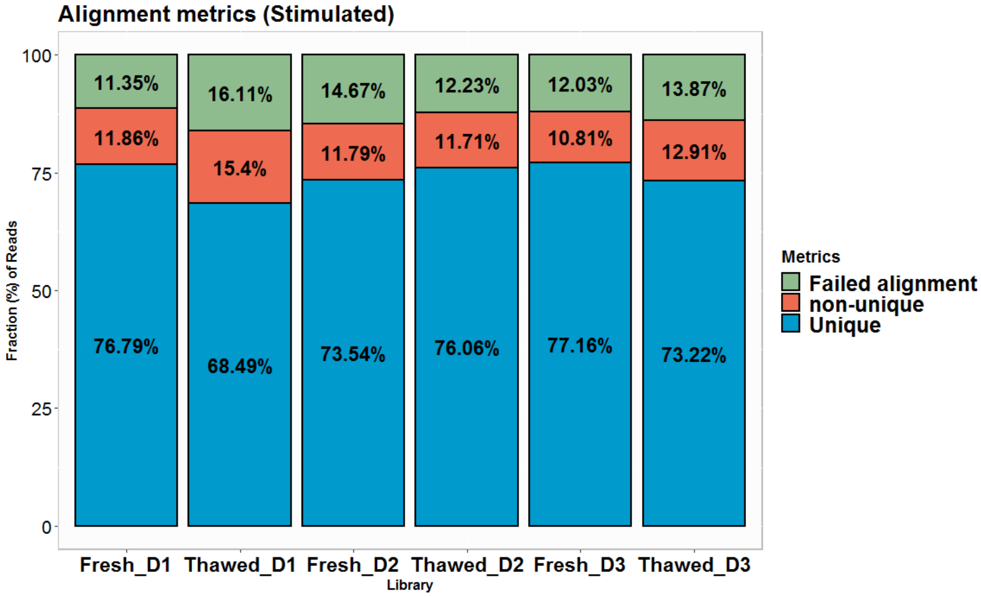


Figure 3-36 | Alignment metrics for fresh and thawed Treg ATAC-seq libraries (stimulated).

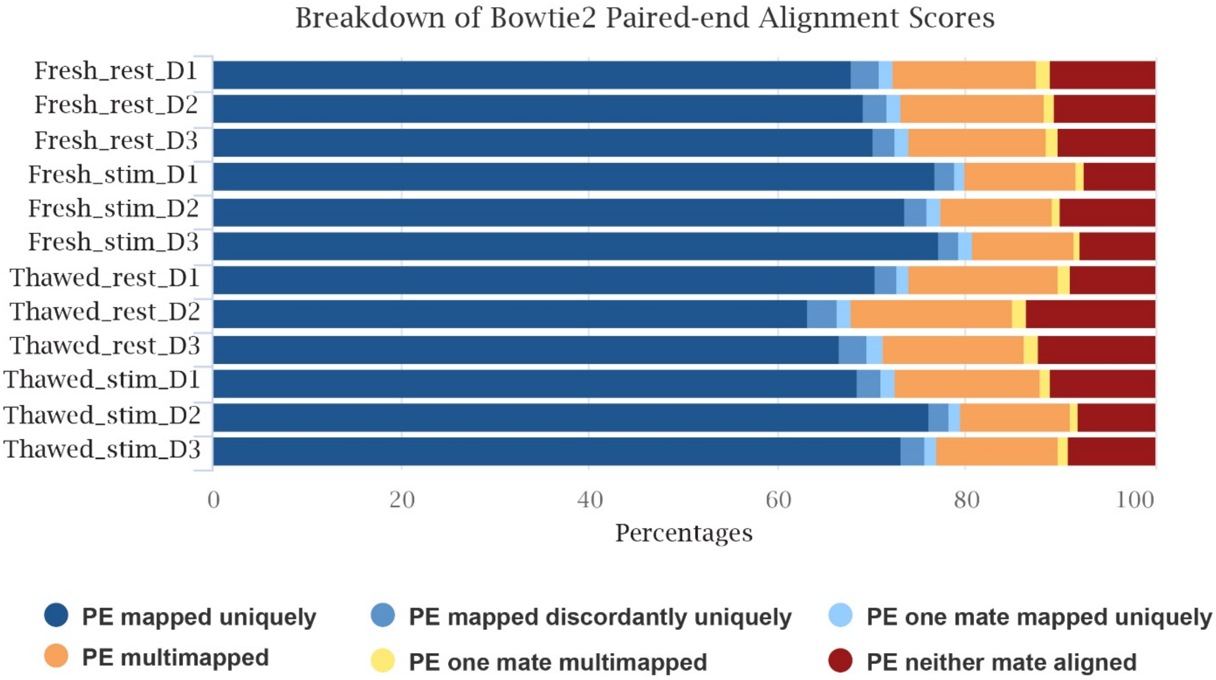


Figure 3-37 | Breakdown of alignment metrics for mapped ATAC-seq reads in fresh and thawed Treg ATAC-seq libraries, including percentage of uniquely and multi-mapped reads.

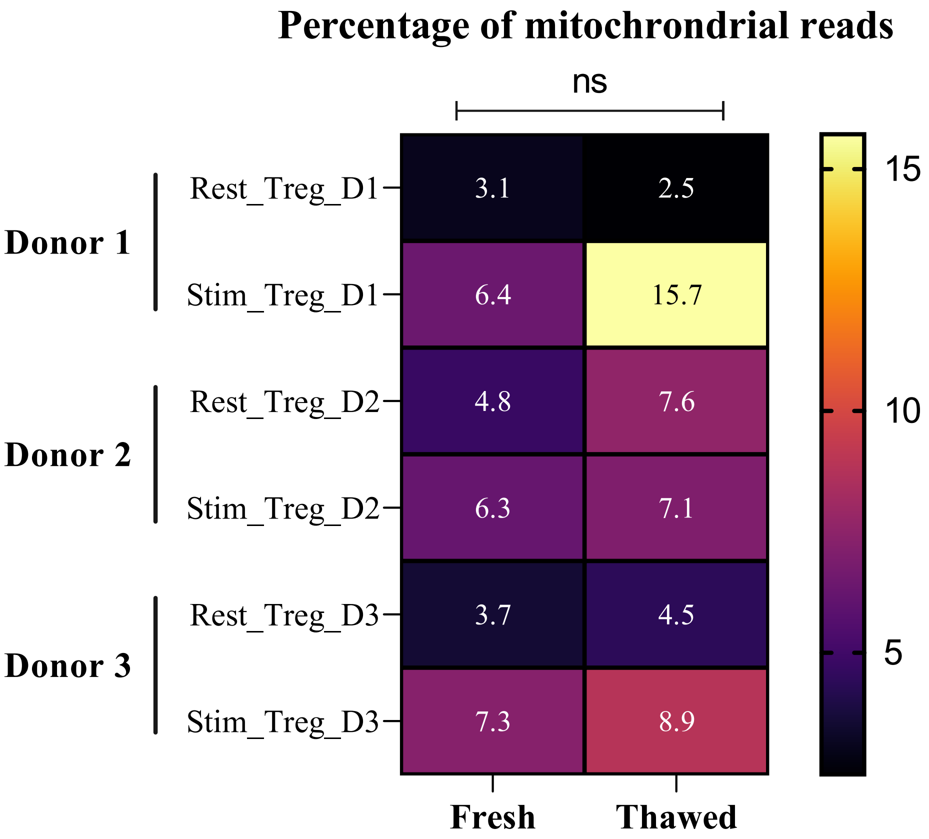


Figure 3-38 | Heatmap-based representation of ATAC-seq quality control metric – percentage of reads mapping to mitochondrial DNA, for fresh and thawed Treg ATAC-seq libraries.
Deeper color is used to depict the most desirable value of the statistic and range following a linear scale starting at 0 (black) and ending at the maximum value (yellow). All values were determined from the full depth of aligned reads.

3.4.3.3 RNA-seq of Whole RNA vs ATAC-seq lysate RNA

RNA-seq libraries constructed from RNA material purified from stimulated whole Treg cells or supernatant fractions from Omni ATAC-seq lysis reaction (ATAC-SN) (n=4 donors; Figure 3-25) were sequenced on a paired-end 150-cycle Illumina Hiseq sequencer (Illumina) to an average read depth of 33.6 million reads (± 3.6 million) per sample. Library D3-stim-whole had the highest number of sequencing reads (39,866,762) whereas D2-stim-SN had the lowest read yield (27,957,585) (Figure 3-39). The whole and ATAC-SN RNA-seq libraries demonstrated similar level of per base sequence quality (Figure 3-40) and alignment metrics (Figure 3-42). However, in comparison with whole RNA-seq libraries, RNA-seq libraries generated from

Omni ATAC-seq lysis reaction (ATAC-SN) had higher level of sequence duplication (Figure 3-41) and lower level of molecular complexity (Figure 3-43).

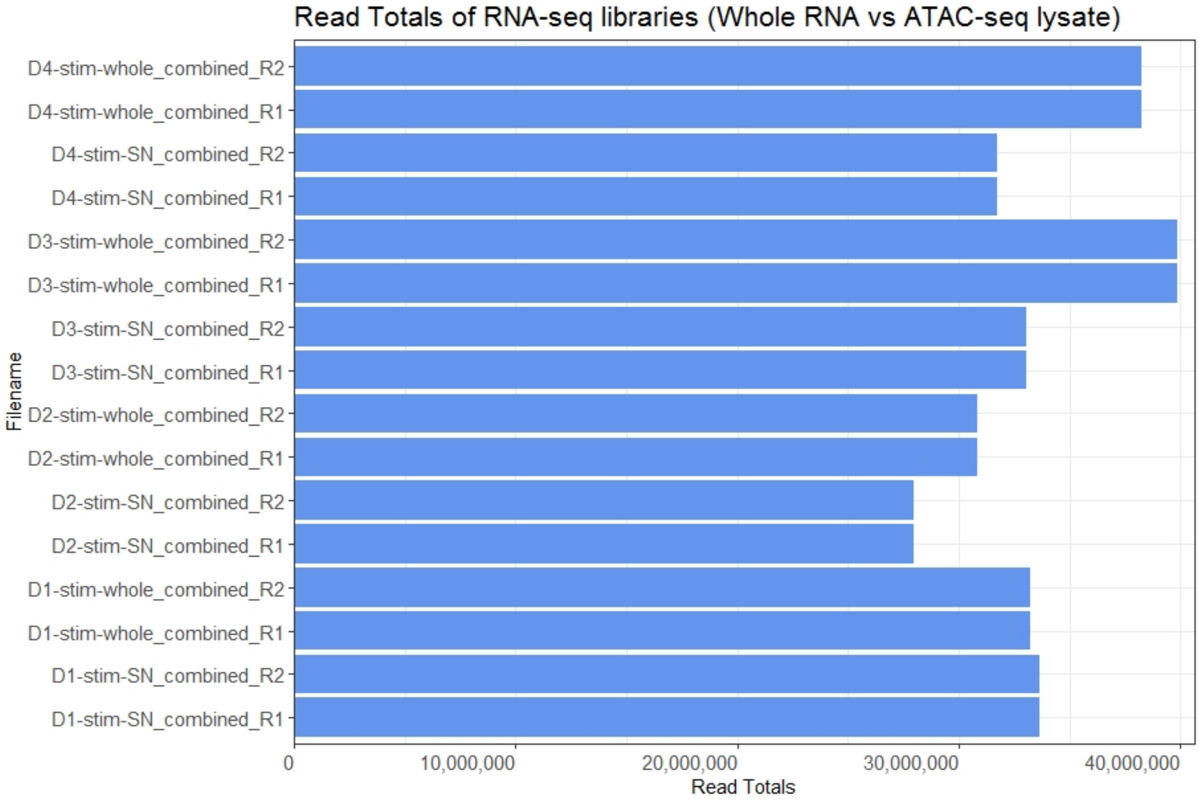


Figure 3-39 | The total number of paired-end reads of Treg RNA-seq libraries, constructed from RNA samples prepared from ATAC-seq supernatant fractions or whole Treg cells (matching ATAC-seq libraries in Figure 3-31).
Stim, stimulated; SN, supernatant.

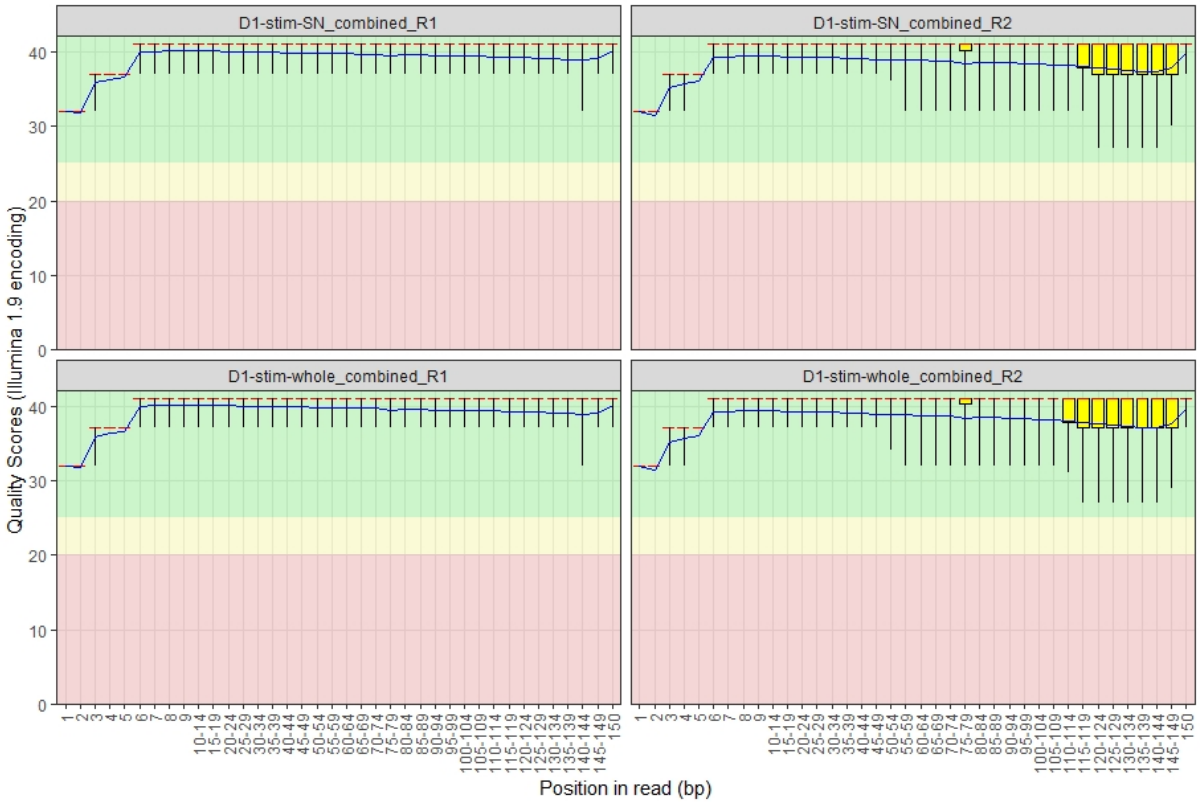


Figure 3-40 | Representative per base sequence quality plot for two paired-end Treg RNA-seq libraries (Donor 1).
This QC metric provides the distribution of quality scores at each position in the read across all reads in both orientation of paired-end sequencing.

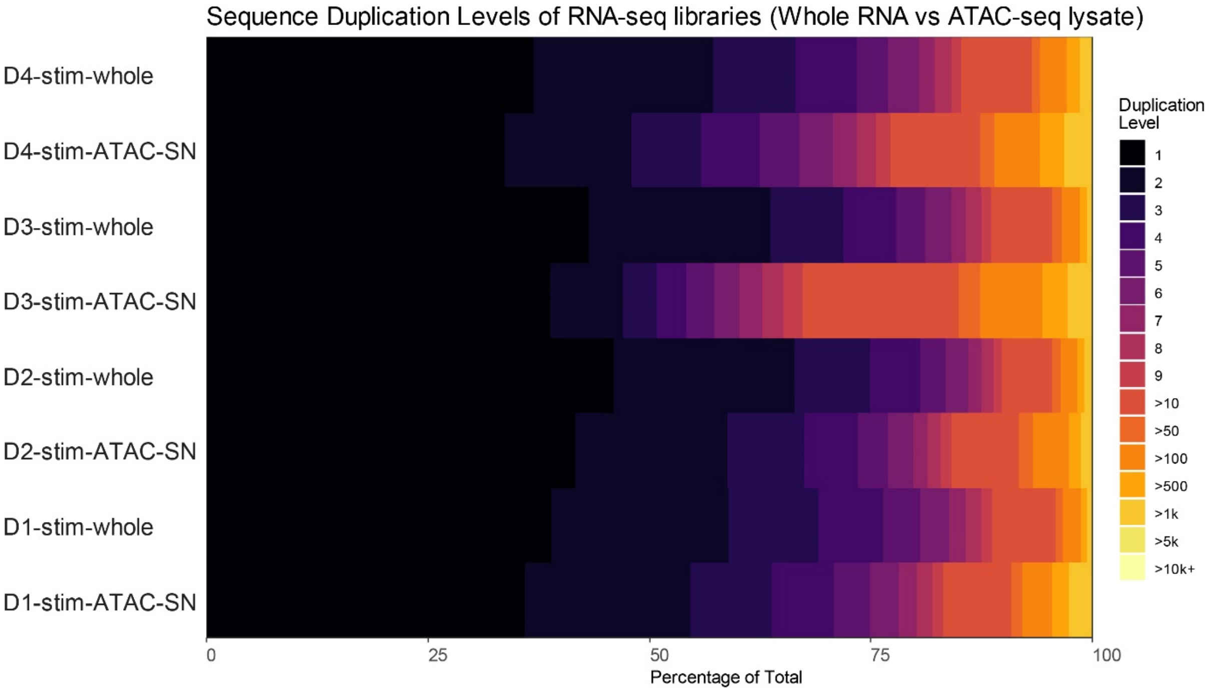


Figure 3-41 | Heatmap showing the duplication levels of whole and ATAC-SN Treg RNA-seq libraries.

The values are expressed as percentages of total instead of read numbers to ensure comparability across libraries. SN, supernatant.

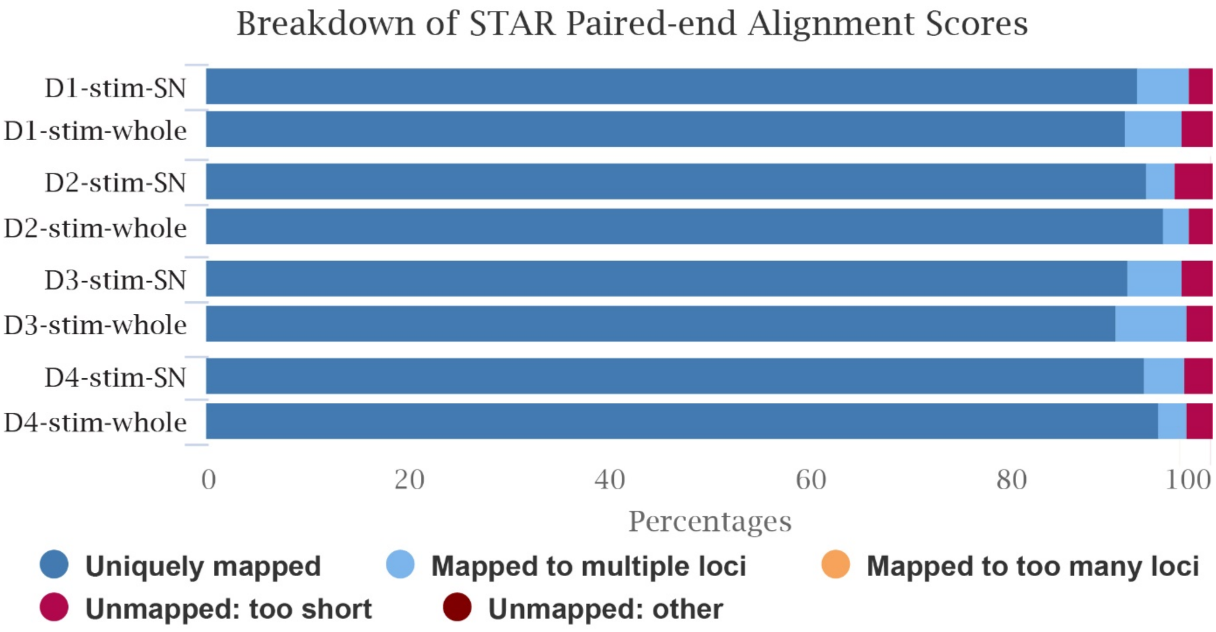


Figure 3-42 | Breakdown of alignment metrics for mapped RNA-seq reads in whole and ATAC-SN Treg RNA-seq libraries, including percentage of uniquely and multi-mapped reads.

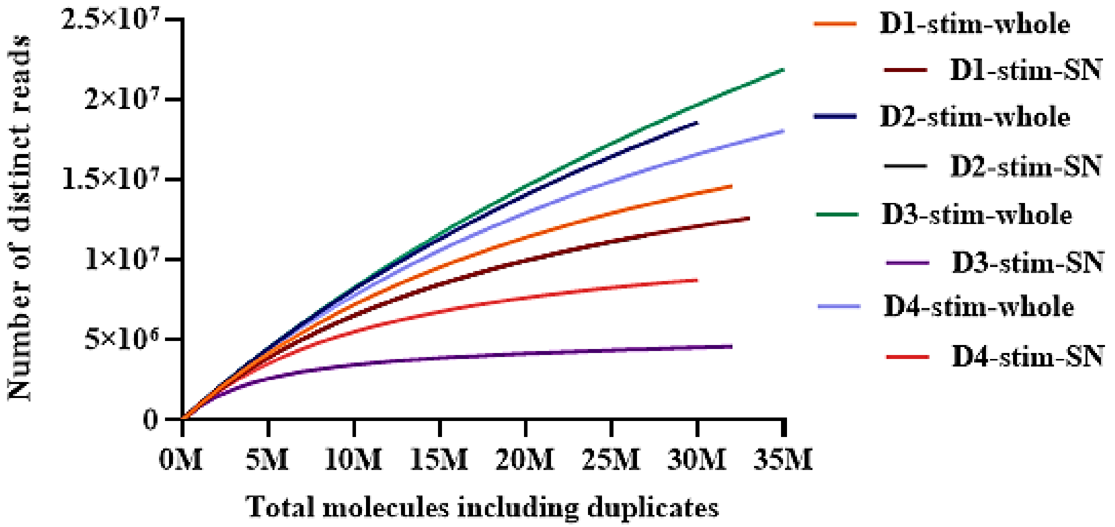


Figure 3-43 | A comparison of the full experiment library complexities of sequenced whole and ATAC-SN Treg RNA-seq libraries.
This calculation interpolates and estimates the number of redundant reads from a given sequencing depth based upon a hypergeometric formula.

3.4.4 Comparable complexity and genomic distribution observed in 30m and 45m tagmentation

The sequenced ATAC-seq libraries of 30- and 45-minute tagmentation (Figure 7-4) showed similar enrichment of read signal at open chromatin signature such as transcriptional start sites (TSS) (Figure 3-44), which is a key measure of signal to noise ratio, and multi-modal fragment size distribution with a clear periodicity of approximately 200bp (Figure 3-45) (merged reads from three technical repeats), representing DNA fragments from nucleosome-free regions (NFRs) or fragments protected by an integer number of phased nucleosomes. Both the 30- and 45-minute tagmentation libraries demonstrated comparable library complexity (Figure 3-46).

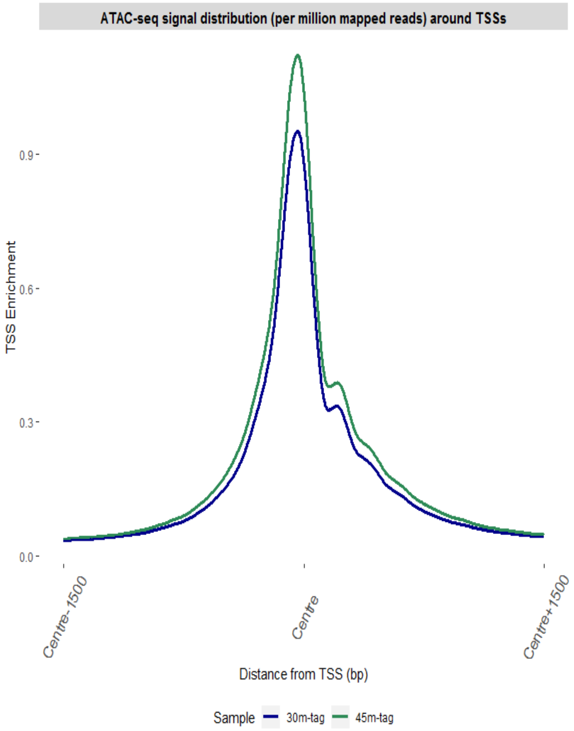


Figure 3-44 | Distribution of ATAC-seq signal at ± 1.5 kb transcriptional start sites (TSS). Signal coverage is calculated from reads per million mapped reads for each sample. The data shown are representative of reads pooled from three technical replicates.

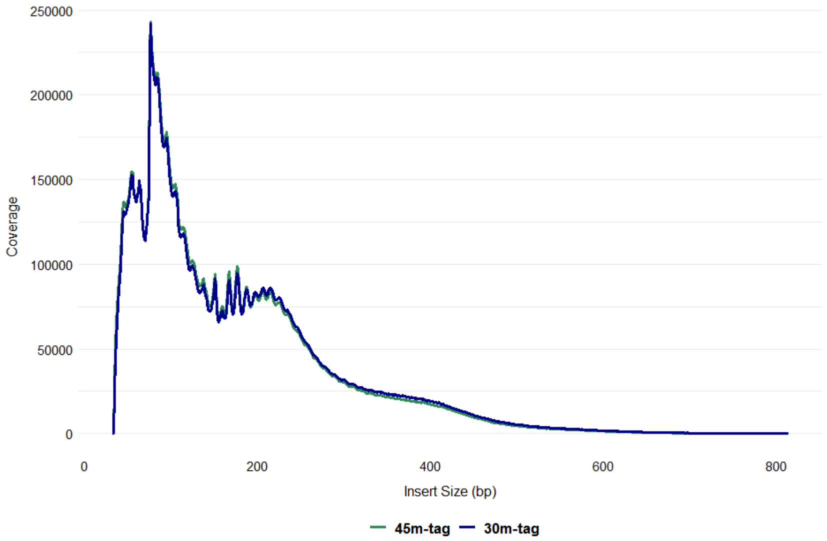


Figure 3-45 | The insert size distribution of 30- and 45-minute tagmentation ATAC-seq libraries shows clear nucleosomal banding patterns.
The data shown are representative of reads pooled from three technical replicates.

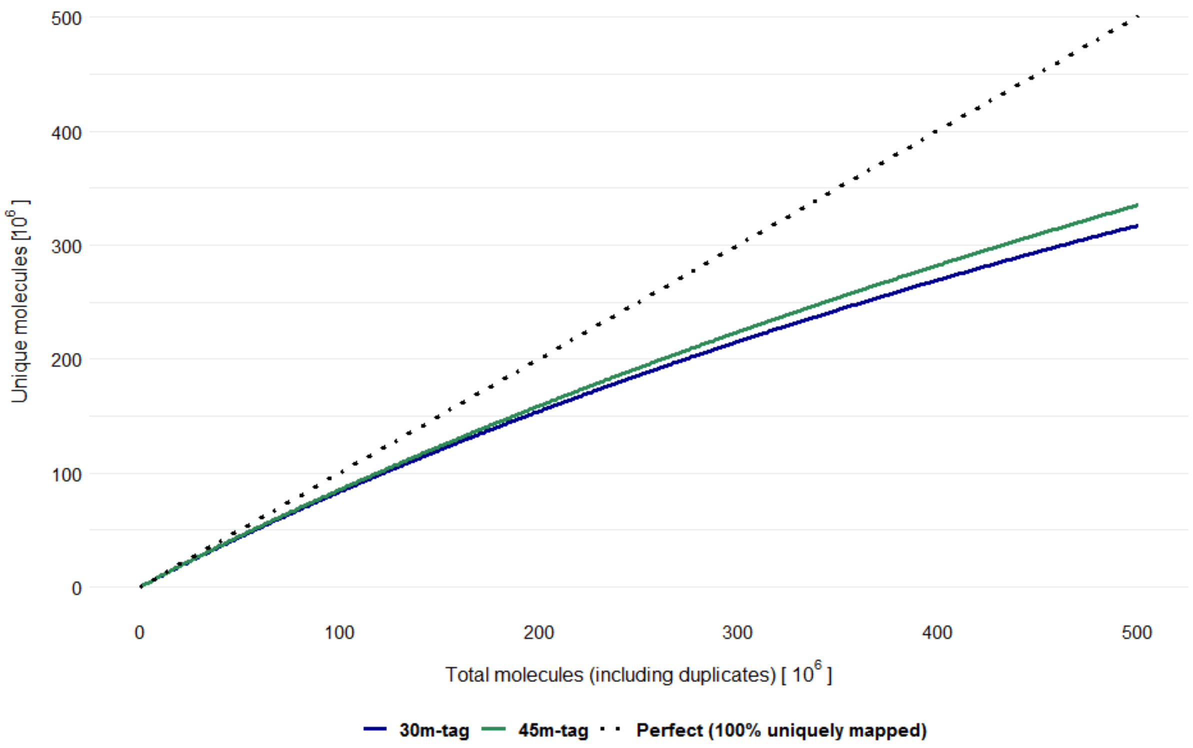


Figure 3-46 | Full library complexity and extrapolated yield curve of 30- and 45- minute tagmentation ATAC-seq libraries.
This calculation[11] uses rational function approximations of Good & Toulmin's non-parametric empirical Bayes estimator to predict the library complexity of future experiments several orders of magnitude larger than the initial run. The data shown are representative of reads pooled from three technical replicates. The black diagonal line represents an ideal library, in which every read is a distinct read.

3.4.5 Library complexity and chromatin accessibility signal is preserved through freezing process

As described in Section 3.3.3 and Section 3.4.2.6, to determine any impact freezing may have on the chromatin structure, Treg cells were isolated from the freshly processed and thawed PBMCs from the healthy adult volunteers (n=4) (Figure 3-47) and subjected to ATAC-seq (Figure 3-21, Figure 3-22 and Figure 3-31). Following thawing and cell sorting by FACS, an almost identical cell viability was observed in lymphocytes population from the freshly processed and thawed PBMC samples (Figure 3-48, Figure 3-49). The T cell population complexity was preserved as demonstrated by the representative FACS plot depicting the frequency of sorted Tconv and Treg cells through fluorescent labelling of surface markers – CD4, CD25 and CD127.

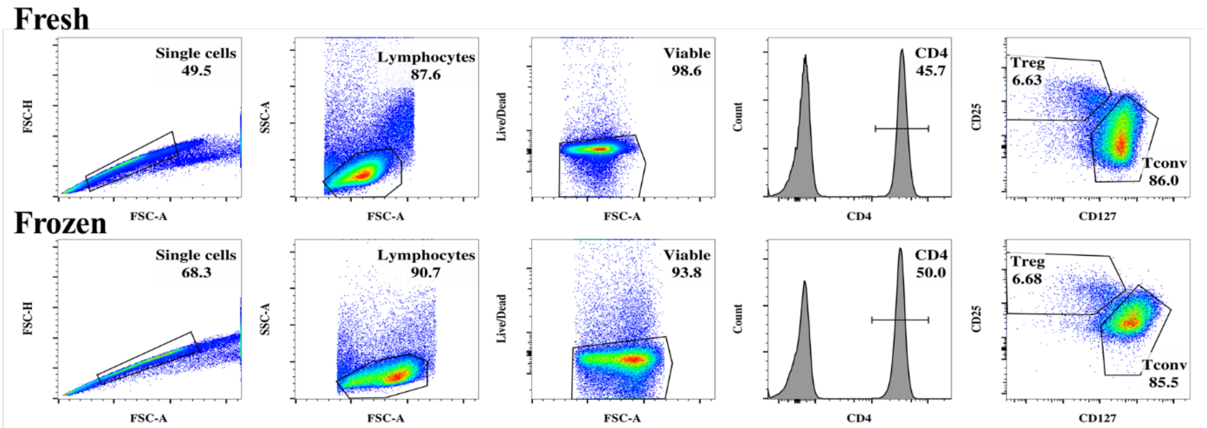


Figure 3-47 | Representative flow cytometric profile from one fresh and thawed PBMC sample demonstrating gating strategy to isolate Conventional T (Tconv) and Regulatory T (Treg) cells.

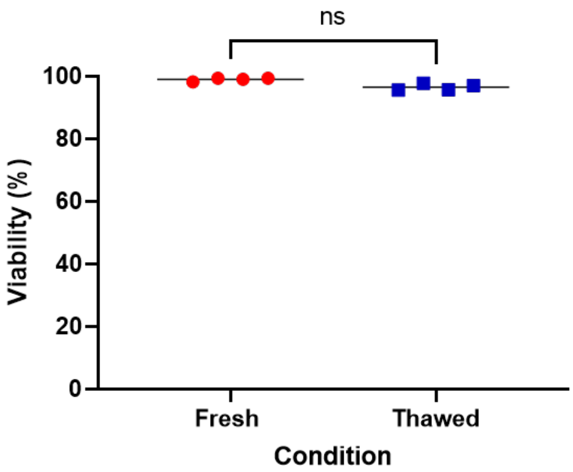


Figure 3-48 | Viability of the fresh and thawed lymphocytes for four healthy adult subjects.
Lymphocytes were identified by forward scatter (FSC) and side scatter (SSC) gating strategy and cell viability was determined by flow cytometry with the use of fixable viability dye. Statistical comparisons were performed with multiple paired T-test with p values ≤ 0.05 being considered significant.

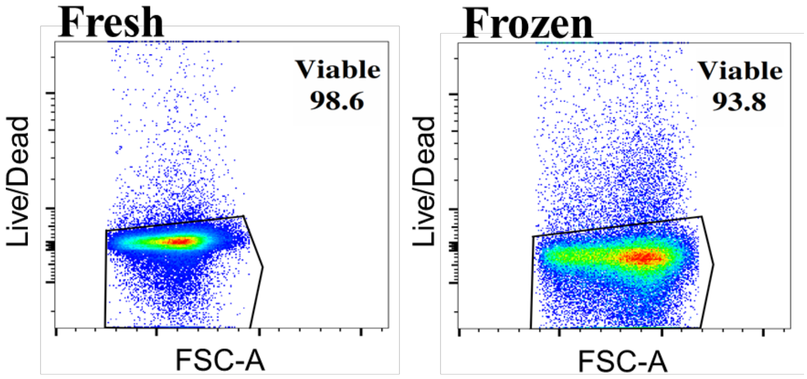


Figure 3-49 | Representative flow cytometric profile of lymphocyte viability for fresh and thawed sample.
Lymphocytes were identified by forward scatter (FSC) and side scatter (SSC) gating strategy and cell viability was determined by flow cytometry with the use of fixable viability dye.

Omni ATAC-seq was performed on CD4^{+} CD25^{hi} CD127^{lo} Treg cells recovered from freshly processed and frozen PBMCs during resting state and in response to stimulation. The fresh and thawed ATAC-seq libraries showed comparable molecular complexity (Figure 3-50, Figure 7-7), multi-modal insert size distribution indicating nucleosomal protection pattern (Figure 3-51), enrichment of open chromatin signal at transcriptional start sites (TSS) (Figure 3-52), as well as genome coverage (Figure 3-53, Figure 3-54) during resting and in response to stimulation.

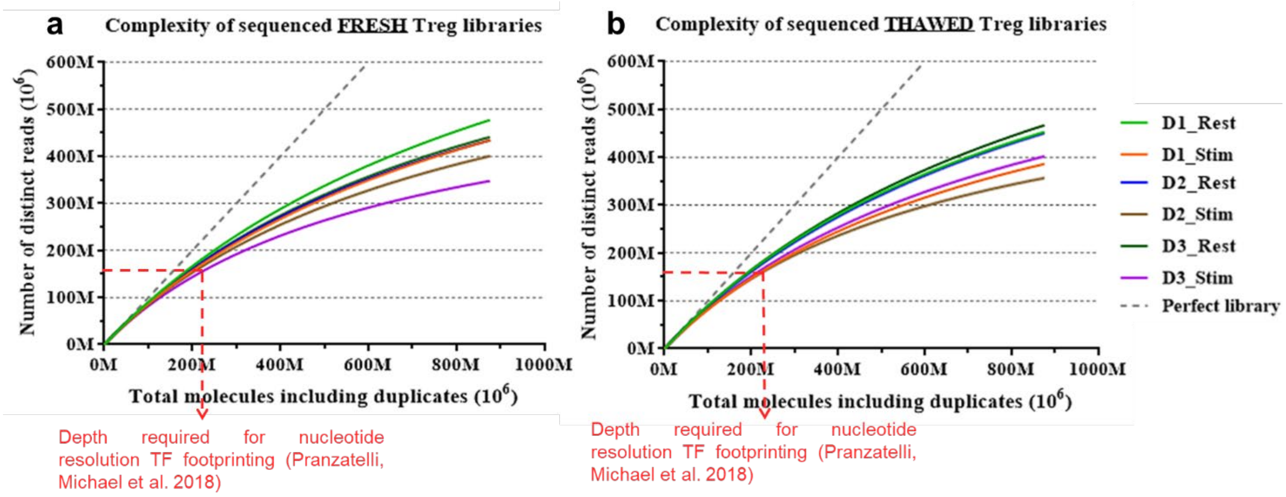


Figure 3-50 | Full library complexity and extrapolated yield curve of individual fresh and thawed Treg ATAC-seq libraries.

This calculation[20] uses rational function approximations of Good & Toulmin's non-parametric empirical Bayes estimator to predict the library complexity of future experiments several orders of magnitude larger than the initial run. Pranzatelli, Michael [43] study observed that sequencing depth of 160 million reads is an optimal read depth for recovery of TF footprints based on ChIP-seq data. The black diagonal line represents an ideal library, in which every read is a distinct read.

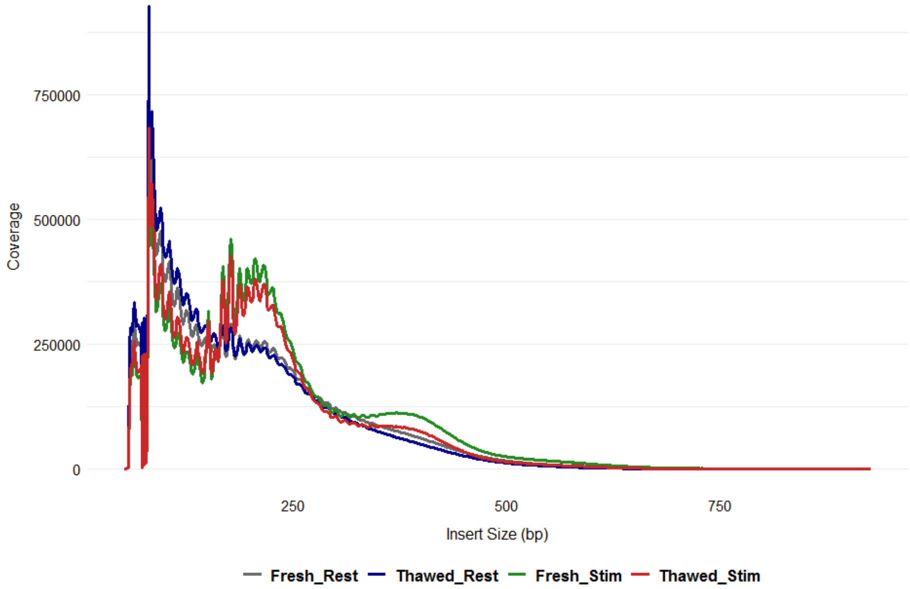


Figure 3-51 | The insert size distribution of fresh and thawed Treg ATAC-seq libraries under resting and stimulated state.

The data shown are representative of reads pooled from three healthy donors.

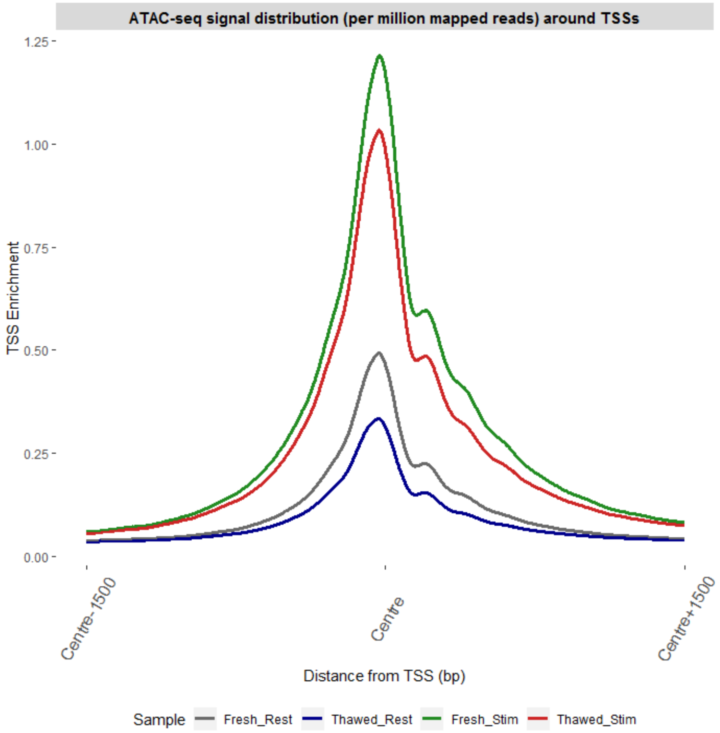


Figure 3-52 | Distribution of ATAC-seq signal at ± 1.5 kb transcriptional start sites (TSS) for fresh and thawed Treg libraries under resting and stimulated state.
Signal coverage is calculated from reads per million mapped reads for each sample. The data shown are representative of reads pooled from three healthy donors.

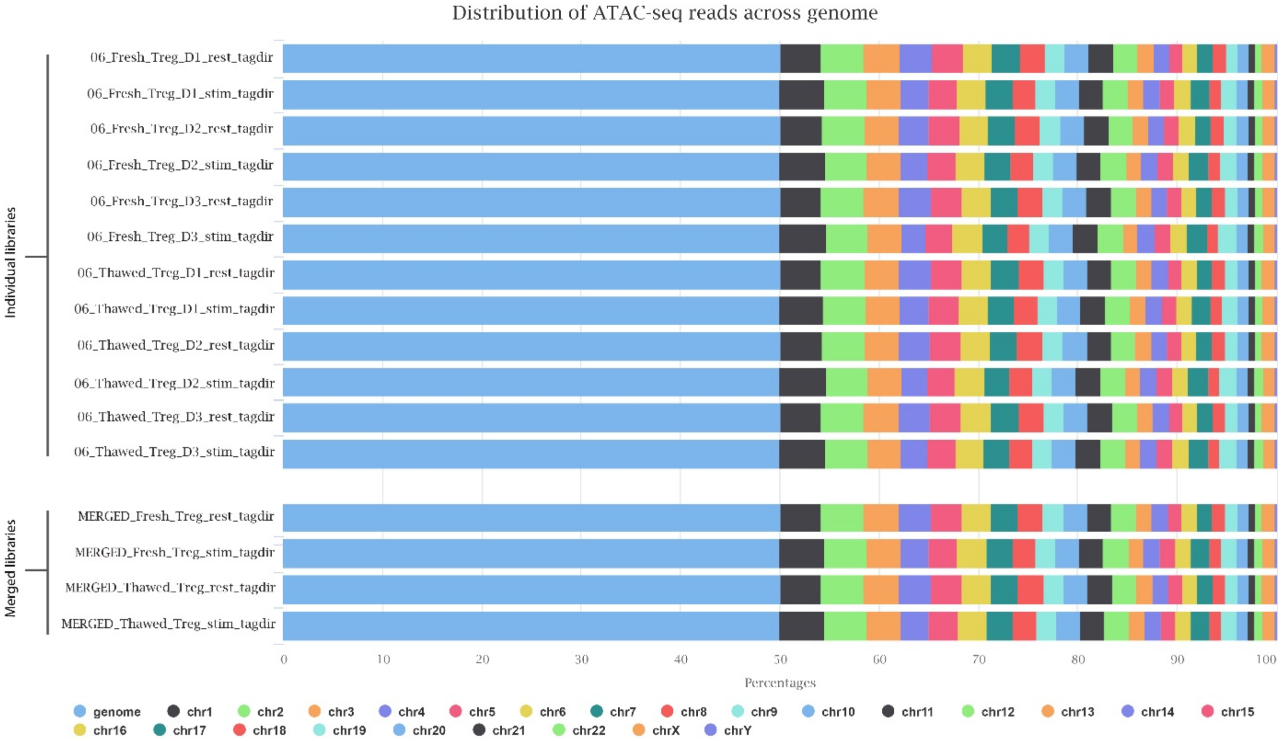


Figure 3-53 | Distribution of ATAC-seq reads mapped to the genome for individual (top panel) and merged (bottom panel) fresh and thawed Treg libraries.
Merged libraries represent reads pooled from three healthy donors.

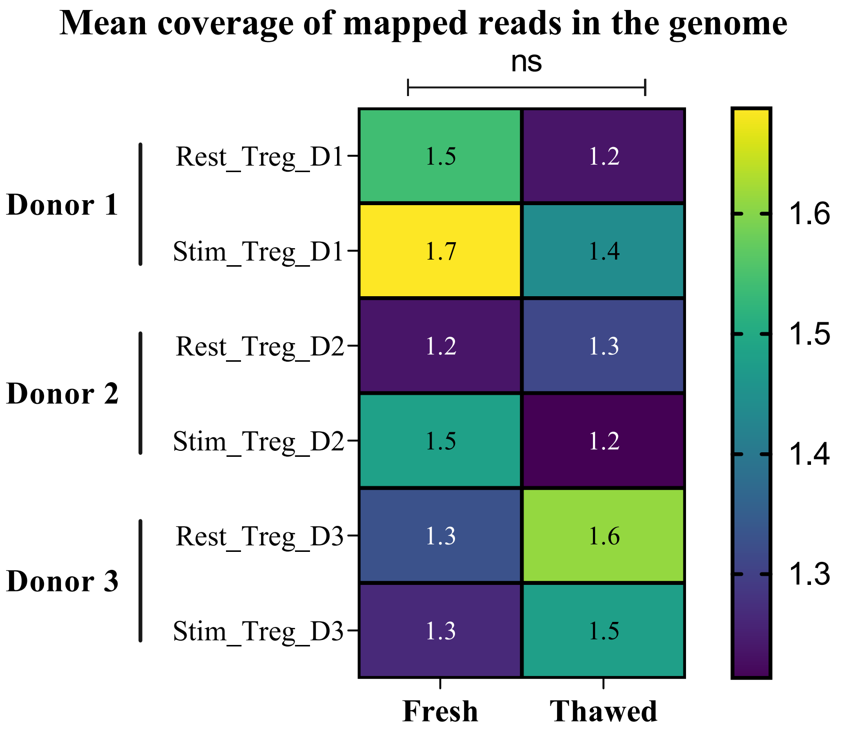


Figure 3-54 | Genome coverage of mapped ATAC-seq reads for fresh and thawed Treg libraries in resting and stimulated state.

Furthermore, a similar fraction of mapped reads in called peak regions (FRiP) from the fresh and thawed ATAC-seq samples was observed (Figure 3-55), suggesting a similar degree of global accessibility enrichment. I also observed a comparable distribution of ATAC-seq peaks between fresh and thawed samples across various annotated genomic features such as intergenic regions, exons, introns and gene promoters (Figure 3-56, Figure 3-57), suggesting the representation of the highly accessible chromatin regions from the frozen cells closely recapitulates that of the fresh cells.

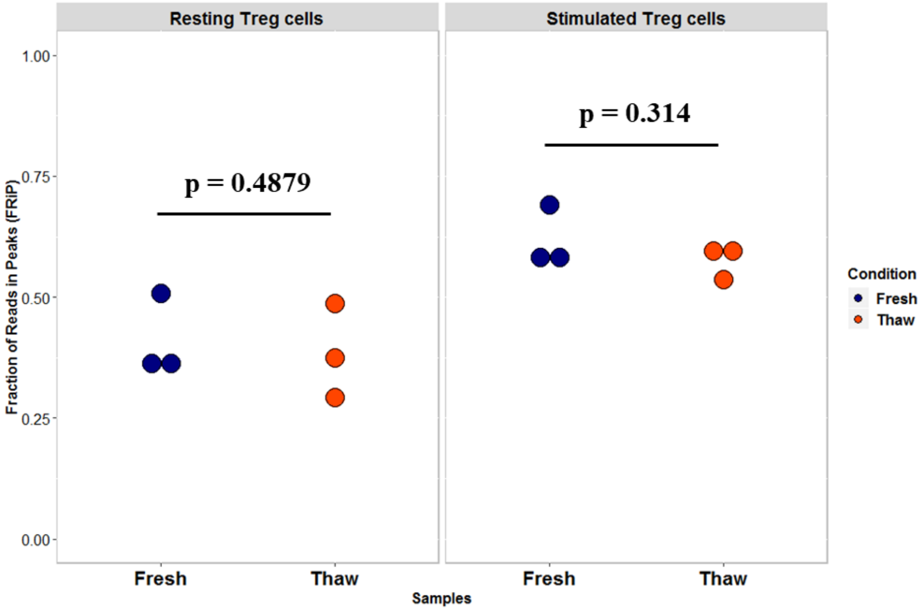


Figure 3-55 | The fraction of reads in peaks for fresh and thawed Treg cells under resting and stimulated state.
Statistical comparisons were performed with multiple paired T-test.

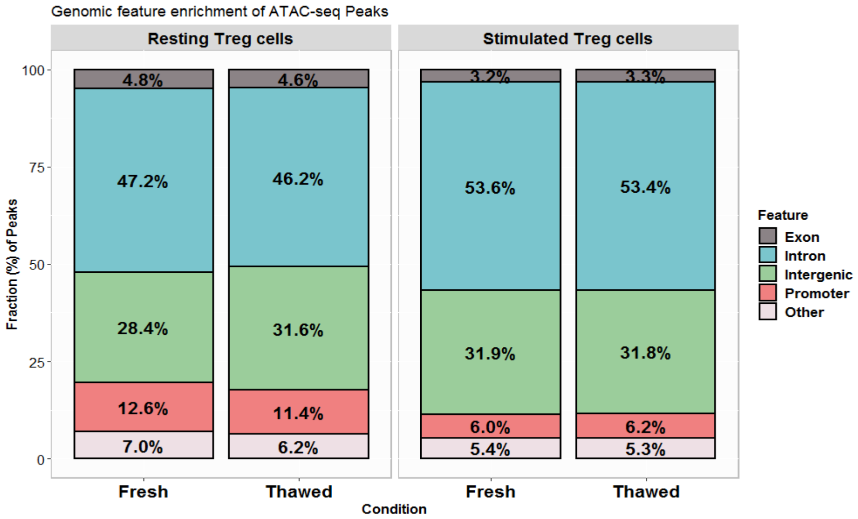


Figure 3-56 | Distribution of ATAC-seq peaks across distinct genomic features, expressed in proportion of annotated peaks.
Promoters are defined by -1kb to +100bp TSS region. Other, peaks annotated to 3' UTR, 5'UTR, miRNA, non-coding RNA and TTS (transcription termination site). The data shown are computed from pooled data representative of 3 biological replicates.

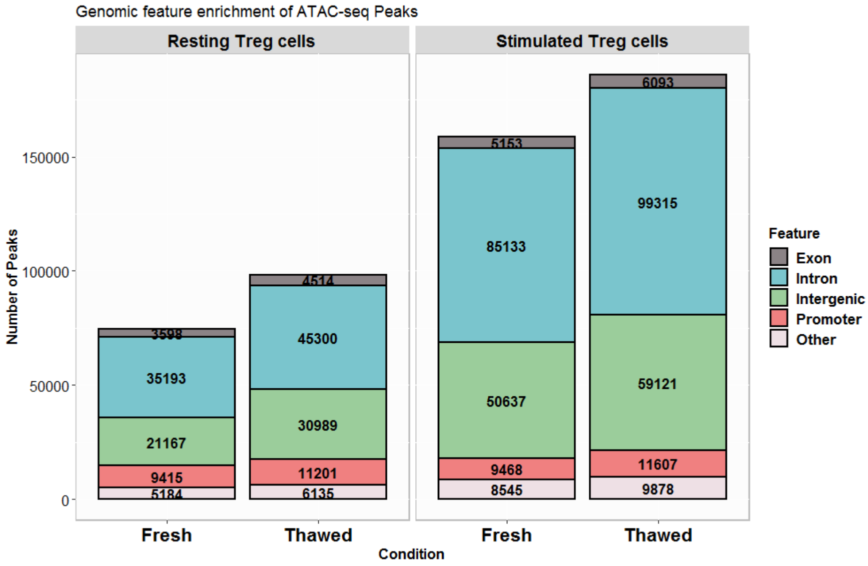


Figure 3-57 | Distribution of ATAC-seq peaks across distinct genomic features, expressed in number of annotated peaks.
Promoters are defined by -1kb to +100bp TSS region. Other, peaks annotated to 3' UTR, 5'UTR, miRNA, non-coding RNA and TTS (transcription termination site). The data shown are computed from pooled data representative of 3 biological replicates.

3.4.6 Protein-DNA interactions are conserved in frozen samples

Within accessible chromatin transcription factor (TF)-bound DNA sequences are selectively resistant to Tn5 cleavage, resulting in short regions of protected DNA or "footprints". To performed TF footprinting analysis, I either used pooled (by condition) reads from full libraries[27] or segregated the pooled reads to extract nucleosome-free regions (NFRs) and regions bound by one nucleosome (1N) (see Section 2.4.1 for detailed protocol) to identify TF occupancy sites. It has was shown that the combined use of fragments from nucleosome-free regions and regions bound by one nucleosome is best for footprinting analysis of Omni ATAC-seq data[21]. I defined nucleosome-free regions (NFRs) as reads with fragment size below 146bp whilst regions bound by 1 nucleosome (1N) as reads with fragment size between 146bp and 307bp [21].

I observed clear, discrete and deep notch of TF footprints in accessible regions identified from the fresh and thawed Treg ATAC-seq mapped reads during resting state and in response to stimulation (Figure 3-58, Figure 3-59 and Figure 3-60). Discrete, symmetrical pattern around the centers of these binding sites was observed in both the fresh and thawed Treg samples from the full library (Figure 3-60), nucleosome-free and mono-nucleosomal fragments (Figure 3-58, Figure 3-59). Similar accessibility levels were observed surrounding the positions of predicted TF binding sites for a wide range of TFs with important roles in T cells such as RUNX1, GATA3, IRF1, NFκβ, NFAT and FOXP3 for fresh and thawed Treg samples, indicating freezing or biobanking did not affect the ability to resolve the accessibility patterns of TFs and had no global effect on TF-DNA interactions.

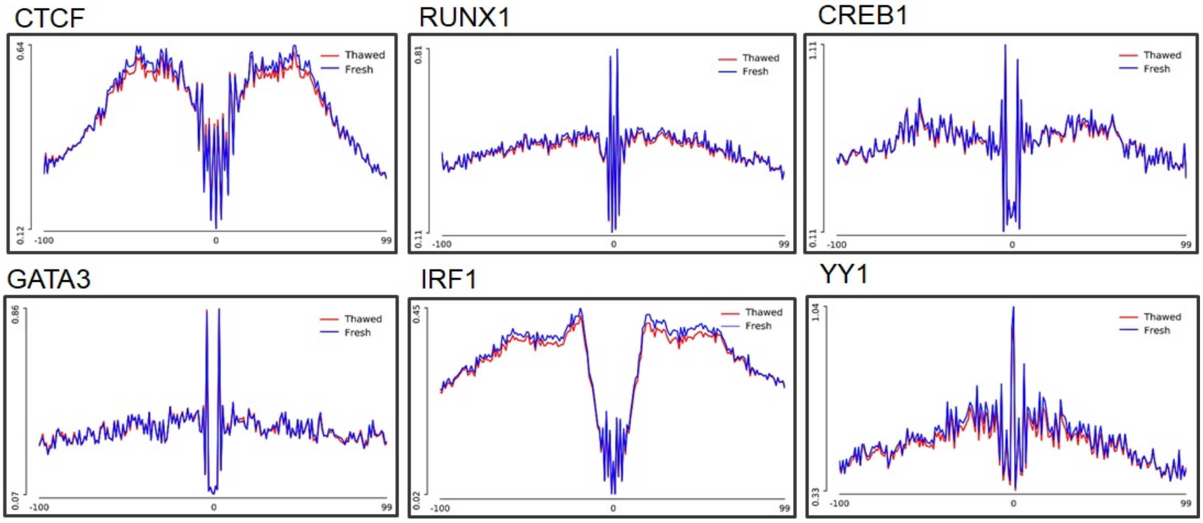


Figure 3-58 | **Chromatin accessibility patterns of TF footprints for fresh and thawed Treg cells in <u>resting</u> state.** Histogram comparing fragments corresponding to nucleosome-free and mono-nucleosomal lengths from fresh and thawed Treg samples during resting state at CTCF, RUNX1, CREB1, GATA3, IRF1 and YY1 binding motifs. Footprint signal was computed using HINT-ATAC[21] on the genome-wide footprints matching the corresponding motifs obtained from JASPAR[29] database. Higher ATAC-seq signal around the TF binding motif is associated with higher activity of that TF in a particular condition. The data shown are computed from pooled sequencing data representative of 3 adult donors.

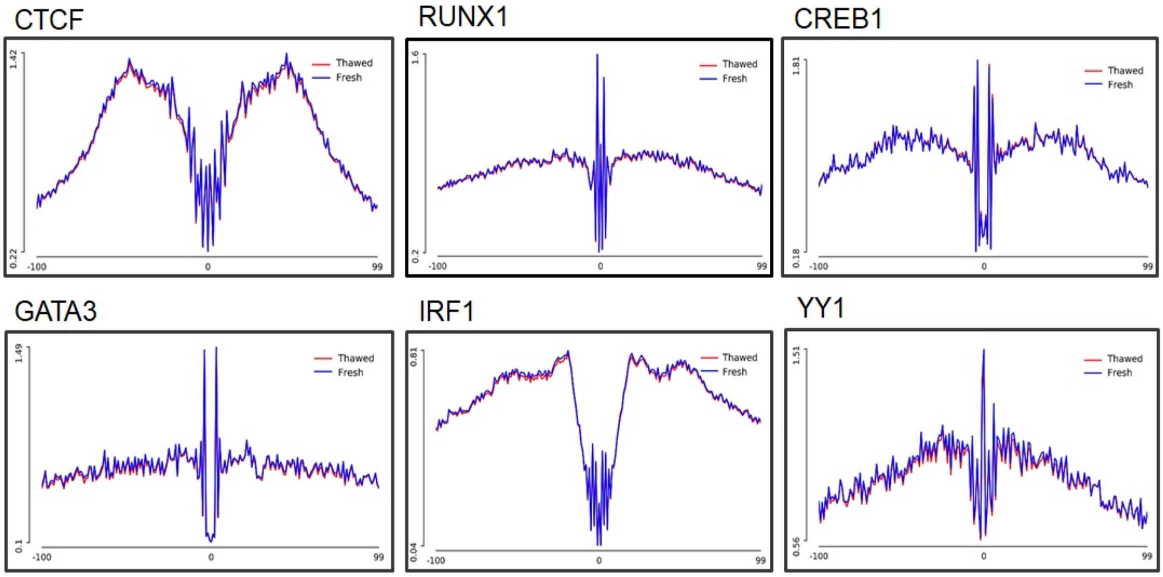


Figure 3-59 | Chromatin accessibility patterns of TF footprints for fresh and thawed Treg cells in stimulated state.

Histogram comparing fragments corresponding to nucleosome-free and mono-nucleosomal lengths from fresh and thawed Treg samples during stimulated state at CTCF, RUNX1, CREB1, GATA3, IRF1 and YY1 binding motifs. Footprint signal is computed using HINT-ATAC[21] on the genome-wide footprints matching the corresponding motifs obtained from JASPAR[29] database. Higher ATAC-seq signal around the TF binding motif is associated with higher activity of that TF in a particular condition. The data shown are computed from pooled sequencing data representative of 3 adult donors.

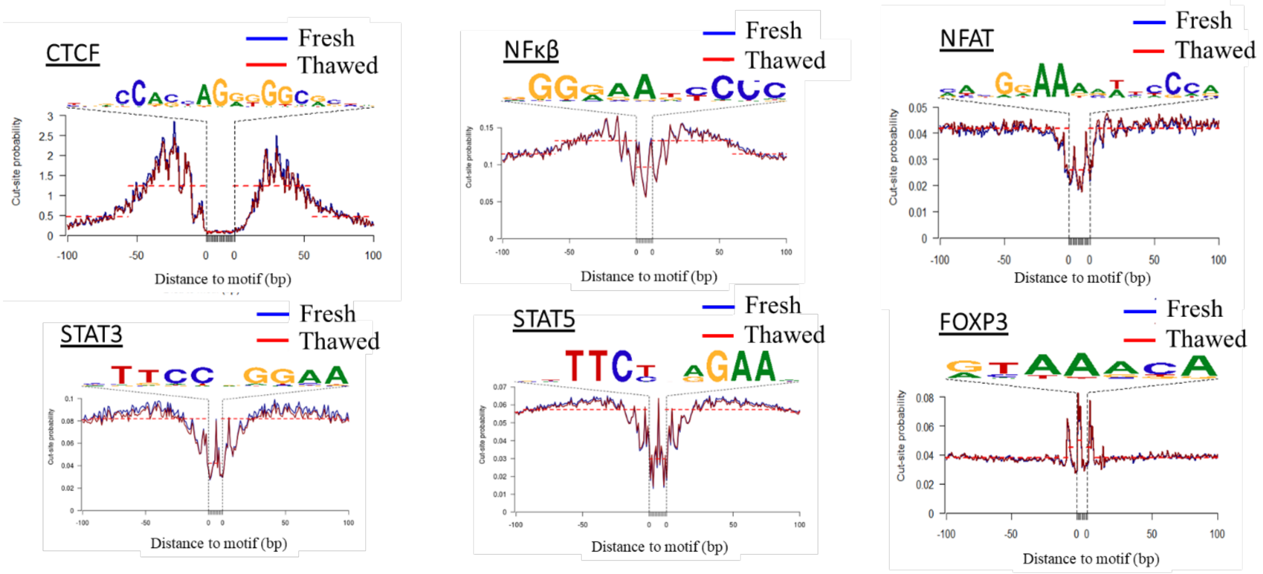


Figure 3-60 | Aggregate ATAC-seq footprints for other immune-relevant TFs generated over binding sites within the genome for fresh and thawed Treg cells in stimulated state. Histogram comparing fragments from full library between fresh and thawed Treg samples during stimulated state at CTCF, NFκβ, NFAT, STAT3, STAT5 and FOXP3 binding motifs. Footprint signal is computed using ATACseqQC [27]. The data shown are computed from pooled sequencing data representative of 3 adult donors.

3.4.7 Cell responsiveness to stimulation is preserved through freezing

It has been shown that stimulation results in a global increase in chromatin accessibility and remodelling and those stimulation-responsive functional regions are enriched in SNPs associated with autoimmune diseases [45, 136], emphasizing the importance of profiling cells under stimulation to identify disease-associated regulatory regions and their roles in autoimmunity.

ATAC-seq was performed on Treg cells isolated from the freshly processed and thawed PBMCs (Section 3.3.3). Treg cells were either left untreated (resting) or stimulated using Dynabeads Human T-Expander CD3/CD28 for 48 hours prior to transposition. Both fresh and thawed Treg libraries demonstrated highly structured ATAC-seq signal around nucleosomes as illustrated in the V-plot (Figure 3-61). A V-plot maps the density of fragment sizes versus fragment midpoint positions and it is used to show cross-correlation of ATAC-seq fragment density and nucleosome positioning along the genome [4, 187]. Both fresh and thawed Treg ATAC-seq show a V-shaped pattern in these aggregate protection profiles (Figure 3-61), where the apex represents the shortest possible fragment that is protected by a nucleosome. A more striking periodicity was observed in stimulated cells for both fresh and thawed samples compared with resting cells (Figure 3-61; right panel), suggesting that stimulation results in stronger enrichment and abundance of fragments available for global nucleosome remodelling. We observed a clear depletion of short fragments being wrapped around the nucleosome, reinforcing that short ATAC-seq fragments are enriched at nucleosome-free regions (NFR). The apex of the V shape is at approximately 117bp while the most enriched position in the V-plot corresponds to fragments of 143bp centered at the nucleosome dyad, which is the length of DNA bound by a canonical nucleosome. Our observation agrees with previous study by Schep, Buenrostro [4].

Furthermore, the fresh and thawed ATAC-seq data showed high correlation of peak signals (Figure 3-62) and the chromatin accessibility landscape is preserved genome-wide in the thawed Treg cells especially at loci indispensable for Treg function and development such as *IL2RA* (Figure 3-63). Peaks were identified at regions that show high overlap with T1D SNPs, regulatory gene enhancers and TF binding sites important for function and activation responsiveness of Treg cells in both fresh and thawed samples (Figure 3-63). Accessibility signal was preserved in thawed Treg samples (Figure 3-63; brown dotted inset) and stimulation results in global increase in chromatin accessibility in both fresh and thawed Treg cells (Figure 3-63; red inset).

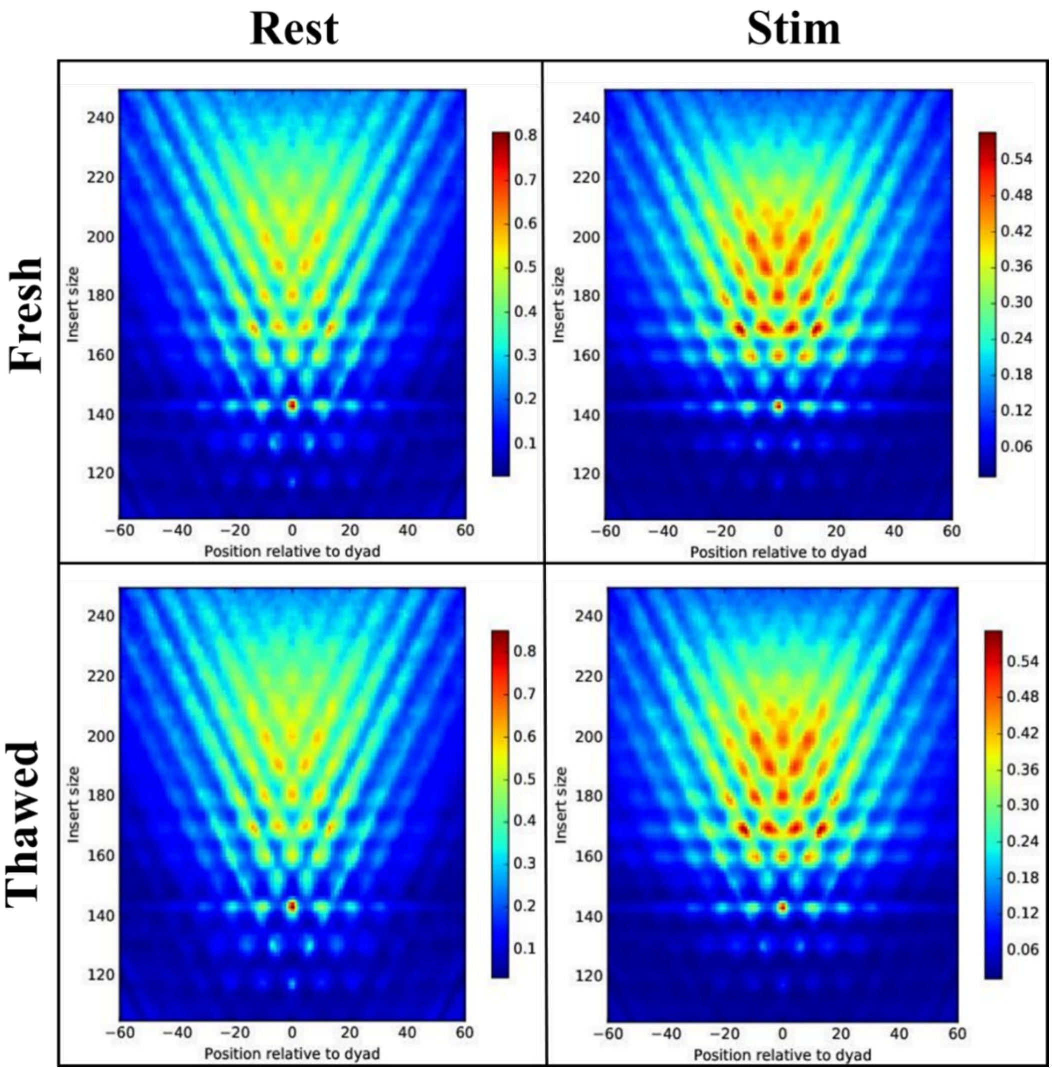


Figure 3-61 | 2D nucleosomal "fingerprint" showing ATAC-seq signal around nucleosomes for fresh and thawed Treg cells during resting and stimulated state.
V-plot, generated by nucleosome-positioning algorithm, NucleoATAC by Schep, Buenrostro [4], maps the density of fragment sizes vs fragment midpoint positions relative to nucleosome dyads called by chemical mapping. Y-axis value represents insert size of fragments (bp) and X-axis value represents distance of the fragment midpoint from nucleosomes (bp). These aggregate protection profiles depict a V-shaped structure, where the apex of the "V" represents the smallest possible fragment that spans the DNA protected by a nucleosome. The calculation was performed on pooled sequencing reads representative of 3 adult donors.

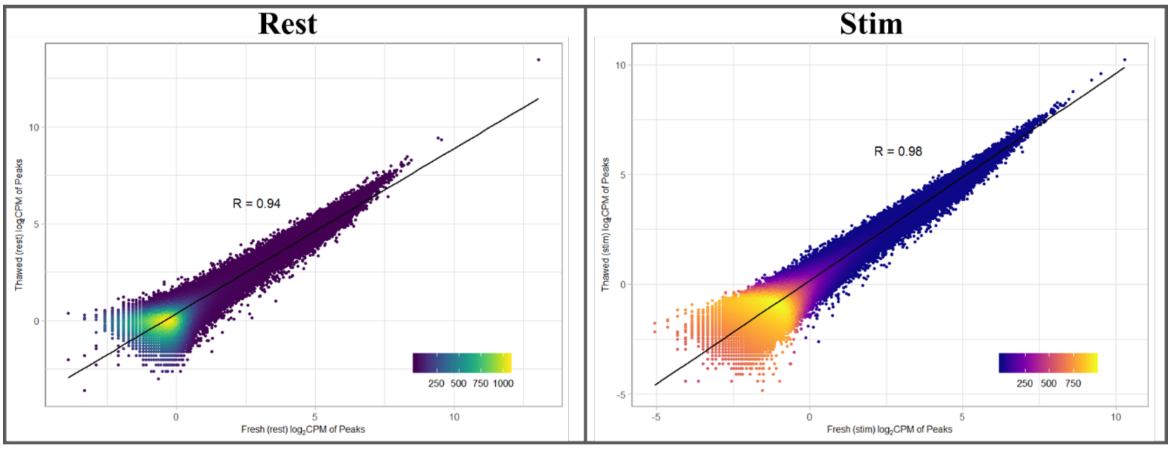


Figure 3-62 | Scatter plot of the ATAC-seq count per million (CPM) reads in peaks for fresh and thawed Treg cells during resting (left) and stimulated (right) state.
Pearson's correlation coefficient value is indicated. The calculation was performed on pooled data representative of 3 adult donors.

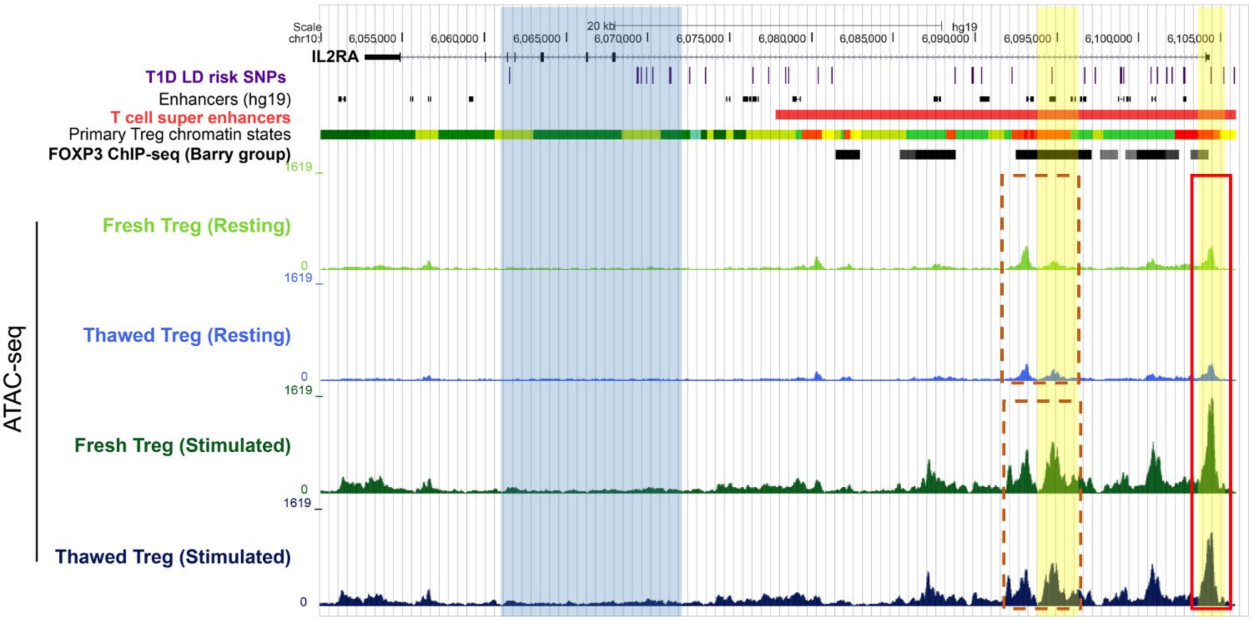


Figure 3-63 | Chromatin accessibility profile of fresh and thawed Treg cells during resting and stimulated state at the IL2RA locus.
ATAC-seq signal is intersected with T1D LD risk SNPs (GWAS database), hg19 enhancer annotation (FANTOM5 phase 2.5), T cell super enhancers [7], Treg chromatin states (Roadmap Epigenomics Project) and FOXP3 binding sites [11]. Each ATAC-seq track represents signal pooled from 3 donors. Browser view was generated using UCSC genome browser. *T1D, Type 1 Diabetes.*

3.4.8 A significant fraction of transcripts is retained in ATAC-seq lysate

This section describes an RNA-seq experiment that aimed to determine if the transcript levels of RNA material recovered from the ATAC-seq lysis reaction (ATAC-SN method) closely recapitulates that isolated from the whole cell extract (conventional method) (see Section 3.3.4 and Section 3.4.2.7). This streamlined workflow allows us to extract both the chromatin accessibility information and transcriptional signatures simultaneously from a single population of 50,000 cells.

RNA-seq libraries were constructed from RNA material (n=4 donors) recovered from ATAC-seq supernatant fractions (ATAC-SN) and compared with RNA purified from whole cell extracts. The whole cell extracts comprised RNA material from both the cell nucleus and cytoplasm whereas RNA recovered from the ATAC-seq supernatant reaction would predominantly originate from the cytosolic fraction of the cells (nucleus is extracted for use in transposition reaction in ATAC-seq workflow). The RNA-seq libraries were barcoded, pooled and sequenced on a paired-end 150-cycle Illumina Hiseq sequencer (Illumina) to an average read depth of 37.5 million reads (± 16 million) per sample.

The scatterplots of rlog-transformed counts show high correlation in gene-level counts between whole cell and ATAC-SN samples, although genes from whole cell RNA-seq appear to have higher counts than ATAC-SN (Figure 3-64). rlog-transformation is a normalisation function from Deseq2 [188] which aims to moderate the count variance across the mean. Differential gene expression (DGE) analysis revealed that the expression of 11,827 (91.1%) genes was conserved in the ATAC-SN samples, whilst some genes showed statistically significant differential transcript abundance between ATAC-SN and whole cell fractions, where significance is defined by a Benjamini-Hochberg false discovery rate (FDR) of less than 0.05 and fold change ≥ 1.5 (Figure 3-65, Figure 3-66). Overall, the ATAC-SN and whole cell fractions showed high correlation (R=0.94; p-value < $2.2\text{x}10^{-16}$) of transcript populations

(Figure 3-67). I then performed gene ontology (GO) enrichment analysis for genes that were significantly more abundant in ATAC-SN/cytosol and whole cell compartment (Figure 7-9, Figure 7-10, Figure 7-11, Figure 7-12, Figure 7-13, Figure 7-14, Figure 7-15, Figure 7-16). Transcripts that were more enriched in the ATAC-SN/cytosol fractions are involved in membrane-associated processes such as co-translational protein targeting to membrane, protein targeting to endoplasmic reticulum (ER), as well as cytoplasmic translation, catabolic and metabolic processes (Figure 7-10, Figure 7-14, Figure 7-15, Figure 7-16). Significantly overrepresented GO classes associated with ATAC-SN/cytosolic transcript also included processes associated with extracellular organelles, ribosomes and ubiquitin cycle. Whereas transcripts that were downregulated in the ATAC-SN/cytosol fraction, or upregulated in the whole cell transcriptome, were enriched for terms associated with the nucleus, including nuclear division, nuclear membrane, histone modification, chromosome organization and glycoprotein biosynthesis (Figure 7-9, Figure 7-11, Figure 7-12, Figure 7-13). These results show high overlap with literature that studied global gene expression patterns of subcellular repertoire of RNA molecules extracted from total, nuclear and cytoplasmic cellular extracts [189-191]. As it was demonstared in Zaghlool, Niazi [191] that nuclear-encoded-mitochondrial proteins (NEMPs) were significantly enriched in the cytosol compared with the nucleus, I attempted to assess the enrichment of NEMPs in my RNA-seq dataset. I retrieved a full list of genes (n=1158) encoding proteins involved in mitochondrial localization from Mitocharta2.0 [5] and conducted gene overlap analysis for differentially expressed genes up- or down-regulated in the ATAC-SN/cytosol fraction and human NEMPs (Figure 7-17; a). The statistical significance of the association between gene sets was determined using Fisher's exact test (Figure 7-17; b). The results show that a total of 71 (20.6%) genes significantly upregulated in the cytoplasmic-derived ATAC-SN overlapped with the human NEMPs dataset, whereas only 13 (1.6%) genes downregulated in the ATAC-SN were NEMPs (Figure 7-17; a). NEMPs were shown to be significantly enriched only in the cytoplasm-derived ATAC-SN fraction (p-value = 6x10^{-10}) and not in whole cellular extracts (Figure 7-17; b), corroborating the findings from Zaghlool, Niazi [191].

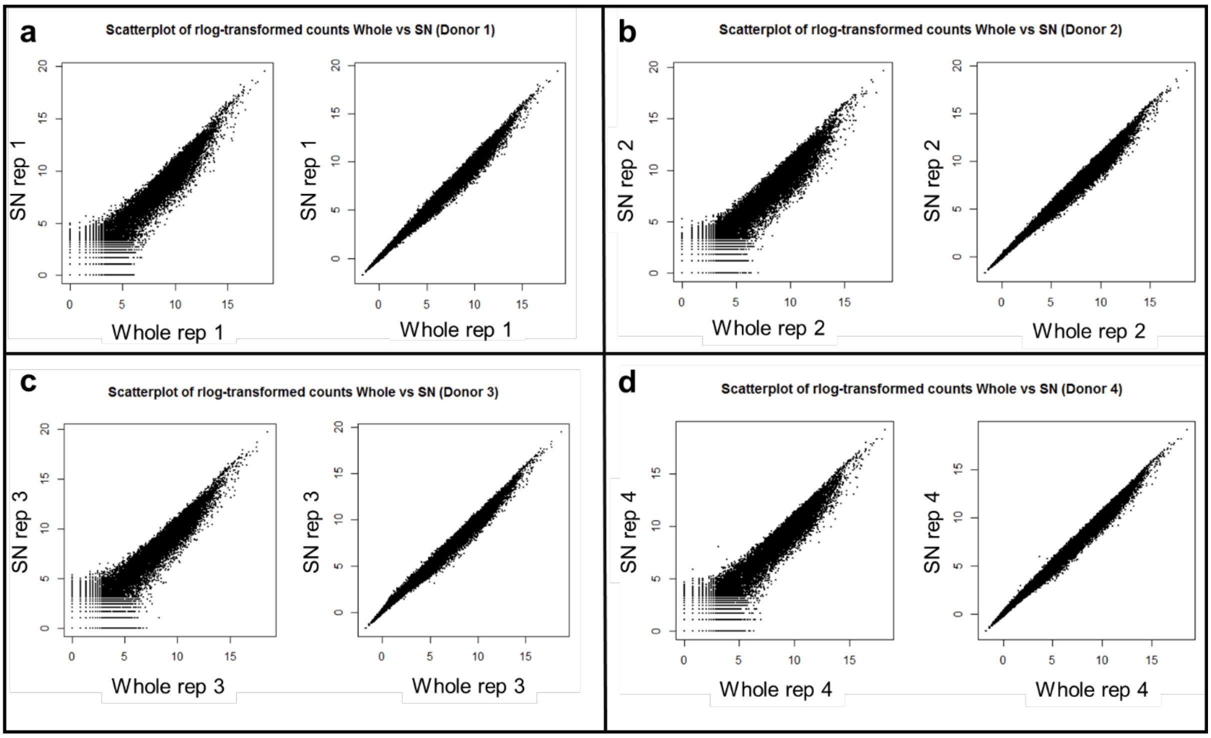


Figure 3-64 | Scatterplot of transformed counts between whole and ATAC-SN samples. This sample-level QC function in Deseq2 transforms the count data to the log2 scale in a way which minimizes variance across samples for low count genes. For each donor (i.e. rep), scatterplot on the left depicts log2 transformed of normalized counts whilst scatterplot on the right depicts "regularized log2" transformed of normalized counts for each gene.

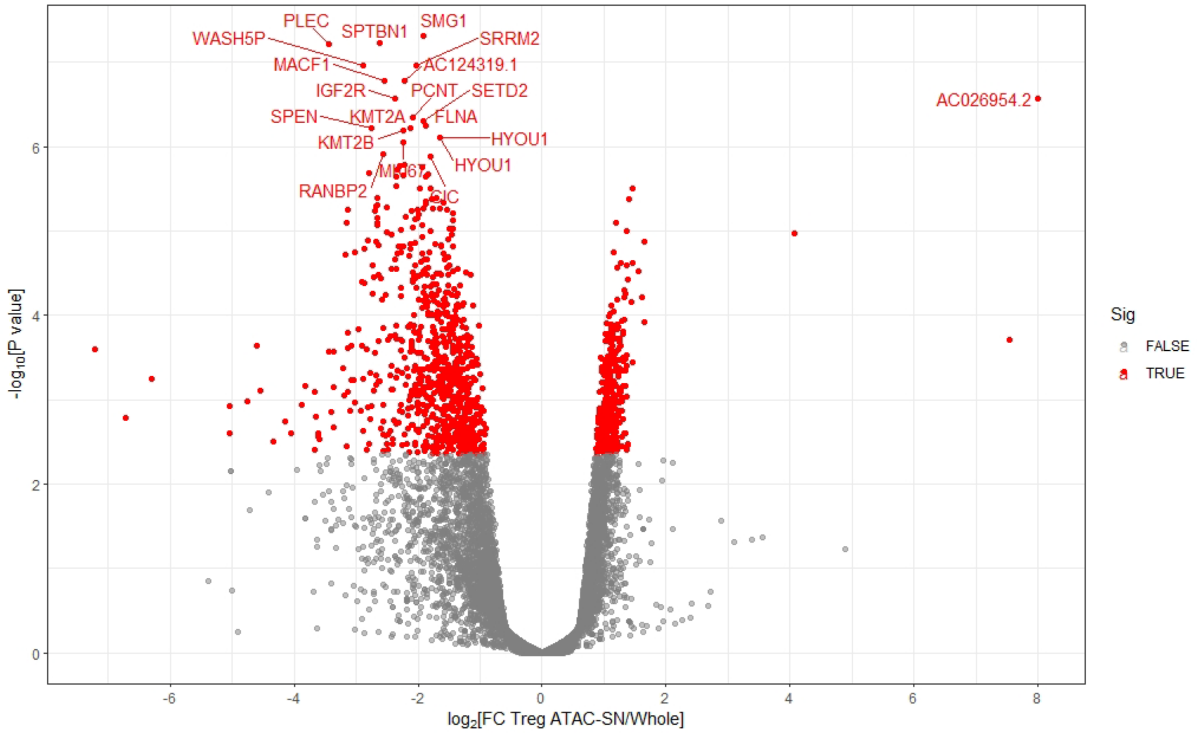


Figure 3-65 | Differential gene expression analysis between ATAC-seq and whole cell lysate purified from Treg cells.

Differential expression analysis was performed using the edgeR exactTest, with genes having a Benjamini-Hochberg false discovery rate (FDR) less than 0.05 and fold change ≥ 1.5 being considered significant (highlighted in blue and red). Top 20 differentially expressed genes are annotated. Results shown are representative of four healthy subjects.

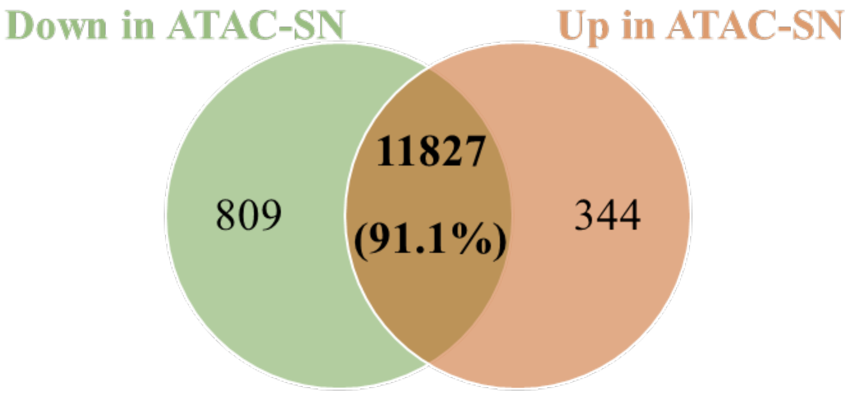


Figure 3-66 | Venn diagram displaying the differential gene expression (DGE) results (volcano plot from Figure 3-65) for RNA-seq of whole cell and ATAC-SN lysate.

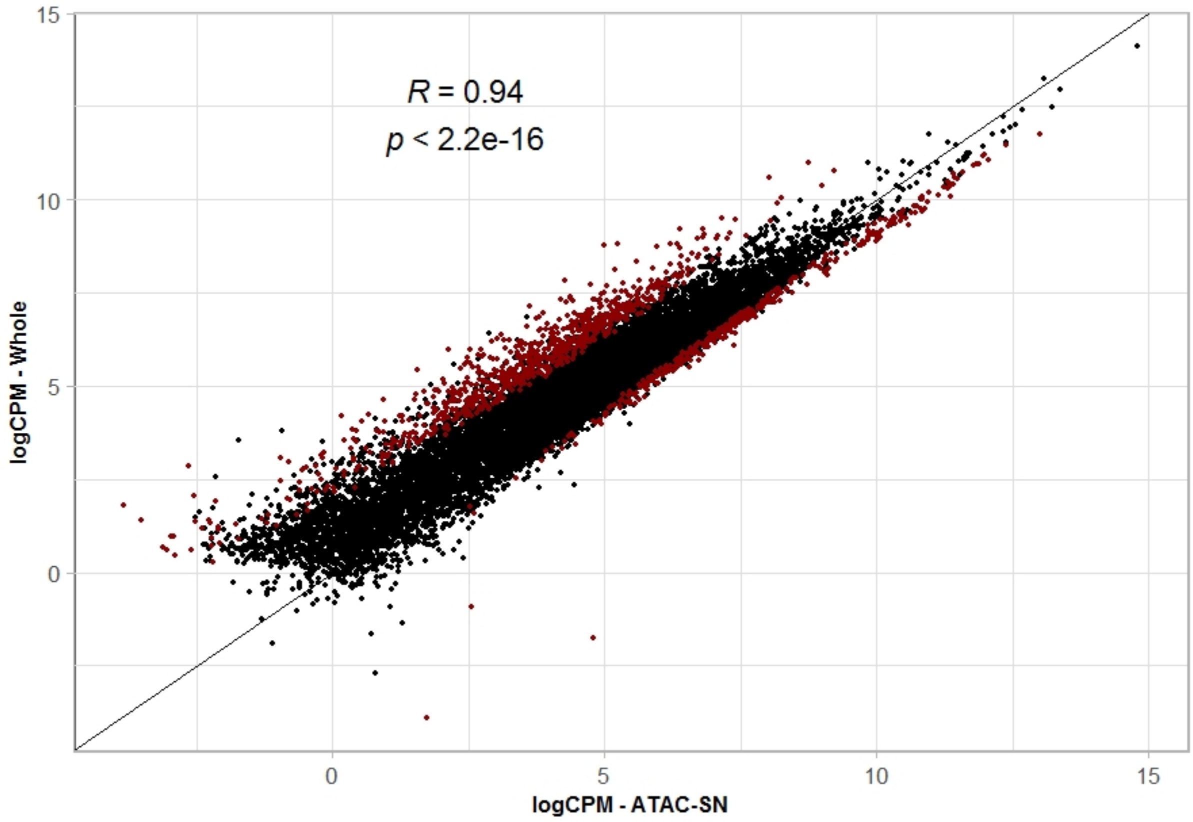


Figure 3-67 | Correlation of gene expression profiles from whole versus ATAC-SN RNA-seq.
The expression for each gene is represented as log count per million of sequencing reads. Differentially expressed genes between whole and ATAC-SN RNA-seq are annotated in red.

3.5 Discussion

My PhD research aimed to profile the chromatin accessibility and gene expression changes in T cells recovered from the frozen, biobanked PBMC (peripheral blood mononuclear cell) samples obtained from the T1D patients and sibling-matched healthy controls. Rigid and robust thawing practices are important for conservation of cells at the phenotypic and molecular level. This is crucial as cryopreserved cells often showed reduced viability and functionality but genomic assay like ATAC-seq requires a high viability population of intact cells as input to avoid sequencing noise due to the tagmentation of free DNA released by the dead cells. Thus, it is critical to first establish good methodologies for PBMC thawing and isolation of sufficient, viable Treg and Tconv cells from thawed PBMCs for downstream genetic and genomic applications.

In this chapter, I have attempted to define the optimum conditions for thawing of frozen PBMCs and for performing Omni ATAC-seq experiment on T cells recovered from thawed PBMC samples. I refined various aspects of the Omni ATAC-seq protocol, including the input amount of Tn5, tagmentation time, library size selection and purification method, in an effort to identify the optimal conditions to produce good quality ATAC-seq libraries in our system of interest.

The first question in this chapter of work sought to determine the conditions or parameters for thawing biobanked samples that lead to maximal PBMC viability and recovery. Biobanks are repositories for storing biological specimens predominantly for use in medical and clinical research, driving development in the era of genomics and personalized medicine [192]. They make a significant contribution to improving our understanding, discovery, prevention, diagnosis, intervention and cure of complex diseases [180]. PBMCs are among the large collections of biospecimens commonly archived in biobanks. Each step in the PBMC thawing process can be performed in a great magnitude of different ways and each variation may have a significant impact on the viability and functional properties of cells. Prior studies have addressed the impact of some technical parameters during PBMC cryopreservation and thawing on the quality and functionality of the recovered cells, including cell density, choice of cryoprotectant and thawing media, thawing duration, method of mixing cells and washing medium and centrifugation speed [181, 193-196]. Those studies attempted optimising cryopreservation and thawing protocols to improve the viability, yield and functionality of the cryopreserved cells in an effort to provide a reliable, convenient and more practical alternative to the use of fresh samples in downstream applications such as flow cytometry[197], immune monitoring[181] and cell transplantation[195]. Not only is the post-thaw viability or recovery important metric to their applications but the recovered samples should closely reflect the physiological and biochemical state in their pre-freeze state.

The results of my work indicate that optimised thawing protocol, which comprised incubation of cryovials at 37°C in a water bath for 10 minutes, implementation of 2 washes in warm thawing media, slow addition of warm media to the cells (at a rate of 1 mL/5 seconds) and overnight culture at a cell concentration of ~3.5–4.0 x 10^6/mL, results in significantly higher overnight recovery (Section 3.4.1) compared with the standard thawing procedure, where cryovials are thawed until last ice crystals are visible, implementation of 1 wash and addition of warm media at a rate of >1 mL/5 seconds. Although significant increase in PBMC yield was observed using the optimised thawing protocol, there was no significant difference in overall PBMC viability between the optimised and standard thawing. These results agree with the findings from Ramachandran, Laux [181] to a certain extent where they showed incubation of cryovials at 37°C in a water bath for 10 minutes with slow addition of thawing media to the cells resulted in improved viability and overall health of recovered cells, compared with thawing until the last ice crystals are visible with rapid addition of thawing media to the cells. Cryopreserved cells are usually washed as part of the thawing process to dilute out the exposure of DMSO [198] which can be toxic to the cells [199]. They also demonstrated that two washes significantly increased the viability of the recovered PBMCs when compared with a single wash, though the impact of number of washes on the PBMC recovery was not measured [181]. My findings are also consistent with that of Hønge, Petersen [197] where it was shown that a longer centrifugation duration with higher force (10 minutes at 500xg) results in significantly higher yield of recovered PBMCs upon thawing compared with a shorter centrifugation duration with lower force (5 minutes at 300xg). Although the impact of most of the parameters tested in my thawing optimisation have been addressed in the work of others, it is important to consider that the state of health and response mechanism of cells is very much dependent on the environmental context that each cell previously experienced or exposed to (e.g. cryopreservation process).

The investigators from Ramachandran, Laux [181] did not measure the impact of those thawing parameters on the overall yield/recovery of cryopreserved cells, and the discrepancies in cell viability outcomes between our studies could be attributed to the timepoint where measurements were taken. Cell viability was measured immediately after thawing in Ramachandran, Laux [181], whereas in my study the measurements of viability and recovery of thawed cells were taken post overnight culture. I introduced an overnight resting period in my thawing procedure as it has been shown that an overnight resting phase improves the elimination of dead or dying cells for a cell concentration of 2–5 x 10^6/mL, as well as the antigenic sensitivity and functionality of T cells recovered from cryopreserved PBMCs [183, 196]. Wang, Huckelhoven [196] further demonstrated that an 18-hour resting period post thawing resulted in significantly higher frequency of recovered Treg cells. This is useful for my subsequent case-control study as Treg cells represent a rare subtype of T cell pool and also the study is constrained by scarcity of paediatric resources from the biobank. Nevertheless, the data from Ramachandran, Laux [181] must be interpreted with caution because viability alone is not a good indication or metric of assessing cell health but measurement of cell recovery should also be taken into account too. This is because the population of cryopreserved cells is made up of viable, early apoptotic, late apoptotic, as well as dead cells [196] and a prolonged incubation of resting period enables apoptosis-prone cells to die and which eventually leads to clearance of dead cells, resulting in a PBMC population with a higher representation of truly viable cells. Thus, the significantly improved PBMCs viability reported in Ramachandran, Laux [181] upon, for example, implementation of two washes could be a result of better clearance of apoptotic cells (and thus, higher cell loss) from longer processing time compared with single wash protocol (10 vs 20 minutes), of which measurement of cell recovery will be able to clarify that. Significant improvement in overall PBMC yield was observed in my optimised thawing protocol, which could be attributed to a lower representation of apoptotic or dead cells resulting from the enhanced thawing parameters. High cell viability seen in optimised and standard thawing procedure is likely an indication of high clearance of apoptotic and dead cells which takes place during overnight resting in both protocols.

Nonetheless, my results build on existing evidence from Ramachandran, Laux [181] that, apart from cell viability, the optimised thawing parameters also result in higher yield of recovered cryopreserved cells. Although, it is beyond the scope of this study to evaluate the impact of various thawing conditions on the phenotype and functionality of the recovered T cells. Indeed, variations in the thawing procedure could result in alterations in phenotype and functions of recovered T cells [181, 183, 196]. This constitutes a limitation in our study. Therefore, T-cell phenotypic and functional assays such as ELISPOT and intracellular cytokine staining for co-stimulatory and inflammatory markers would add complexity to the knowledge generated by this work.

Although it has been well established by research that cryopreservation could have an impact on the viability and functionality of cells and cells purified from freshly collected blood samples are usually preferred, the use of freshly isolated samples in downstream assays poses a number of logistical challenges. The potential for introducing confounding variation in downstream experiments is heightened for fresh samples collected over a protracted period of time and at the same time sole reliance on fresh samples also limits access by future research projects for additional functional validation studies. ATAC-seq is widely applied on freshly processed samples and only a limited number of studies have been performed to benchmark the effects of cryopreservation on chromatin structure [30, 182] none to date have been performed in primary CD4$^+$ T cells.

In this chapter I asked whether freezing before nuclei isolation compromises nuclear integrity and alters chromatin structure in the Regulatory T (Treg) cells. This is an important quality control metric for demonstrating feasibility and compatibility of using ATAC-seq to infer chromatin structure in the biobanked primary T cells, which will subsequently be applied to the biobanked T1D case-control PBMC samples. I adopted an ATAC-seq variant, Omni ATAC-seq [200] that works across multiple applications including frozen tissue samples, with substantial improvement of data quality. PBMCs were obtained from whole blood isolated from healthy adult donors, processed fresh or frozen (according to WCH T1D biobanking standard) and thawed according to the optimised thawing procedure. ATAC-seq was then performed on purified fresh and frozen CD4$^+$ CD25^{lo} CD127$^+$ Treg cells under resting and stimulated state.

The results indicate that the fresh and thawed Treg ATAC-seq samples demonstrated excellent concordance of sequencing data quality including per base sequence quality, sequence duplication levels and alignment metrics (Section 3.4.3.2). Besides, the library complexity, chromatin accessibility, TF-DNA interaction and cell responsiveness to stimulation were preserved through freezing (Section 3.4.5, Section 3.4.6, Section 3.4.7). Rigorous measurements of quality control recommended by ENCODE consortium for ATAC-seq data, such as TSS signal and fraction of reads in peaks (FRiP), show that the fresh and thawed Treg samples had close to identical signal to noise ratios and that tagmentation occurred primarily at focal accessible regions. The fresh and thawed ATAC-seq data showed high correlation of peak signals indicating freeze-thaw did not have an appreciable impact on the accessible chromatin landscape and downstream identification of highly accessible regions in the genome. The overlap of peaks across genomic annotations indicated neither the biobanking nor thawing process biased the discovery of certain genomic features. The results of this work also show that freezing or biobanking did not affect the ability to resolve or discern the accessibility patterns of TFs and had no global effect on TF-DNA interactions during resting state and in response to stimulation. These TFs have a critical role in the differentiation and function of Treg cells and thus, are important in maintaining immune tolerance and homeostasis of the immune system. They include FOXP3, which is the master TF of Treg cells, as well as NFκβ, RUNX1 (Runt-related transcription factor 1) and NFAT (Nuclear factor of activated T-cells), which FOXP3 interacts with in regulating transcriptional programme for Treg differentiation [201-203]. Their association is also confirmed in Allison, Sajti [204] who found primed enhancers of CD4$^+$ T cells were enriched in motifs of Runx1, Ets1, IRF, NF-κB/Rel and AP-1, highlighting the central role of these binding partners in regulating enhancers and gene expression in the primary human Tconv and Treg cells. It was reported that the RUNX1–FOXP3 complex suppresses IL-2 and IFN-γ production and promotes the expression of Treg cell-specific markers such as CD25, CTLA4, and GITR (glucocorticoid-induced tumor necrosis factor (TNF) receptor) to exert suppressive phenotypes in Treg cells [205]. The key role of these FOXP3-interacting partners in regulating Treg transcriptional program is supported by evidence demonstrating that mutations of FOXP3 abrogate its interaction with NFAT, impairing the ability of FOPX3 to inhibit IL-2 secretion, upregulate CTLA4 and CD25 expression to exert suppressive function of Treg cells (154).

Overall, the findings from this section of work support the hypothesis that frozen T cell samples cryopreserved using the standard WCH biobanking protocol and thawed using the optimised thawing method generated high quality chromatin accessibility data, at a level closely similar to freshly processed samples. The chromatin accessibility landscape was preserved genome-wide in the thawed Treg cells especially at loci indispensable for Treg function and development such as *IL2RA*, which is a known T1D-associated risk locus [2, 3], suggesting biobanking process does not perturb epigenetic signatures and the samples can be reliably used to identify gene loci and TF networks impacted in T1D using ATAC-seq. These results strongly reflect those of Scharer, Blalock [30] who found high correlation of signal to noise ratios, accessibility levels and TF footprint patterns between fresh and biobanked CD19$^+$ naive B cell ATAC-seq samples. Study by Fujiwara, Baek [182] also reported that the chromatin landscape from cryopreserved breast cancer cells and mammary tissue closely recapitulated that of the fresh cells and tissue. While previous research has focused on the impact of freezing or biobanking on the genome-wide chromatin architecture in B cells and breast cancer cells, these results build on existing evidence and demonstrate indistinguishable chromatin accessibility and TF occupancy in biobanked T cells under resting and in response to stimulation. Previous studies have provided fundamental resources for understanding the impact of biobanking on the epigenetic regulation in resting cells, but many disease states are associated with activation of immune cells [206, 207]. Dysregulation of T-cell homeostasis and activation are known to play a role in cancer and autoimmunity including T1D and hundreds of genetic variants have been associated with transcriptional regulation and gene expression during T-cell activation and polarization [208, 209].

It is worth noting that, on the level of genome-wide chromatin accessibility, the differences captured between the fresh and thawed Treg ATAC-seq libraries were minimal, more appreciable changes were actually detected between the resting and stimulated libraries for both the fresh and thawed samples. These findings are in agreement with previous studies which show stimulation resulted in large-scale chromatin remodelling and a global increase in chromatin accessibility in primary human T cells [45, 136]. We decided to probe chromatin accessibility not just in resting but stimulation state as it has been demonstrated that genetic variants associated with autoimmune diseases can remain hidden within T-cell stimulation-responsive regions and these stimulation-response regulatory elements explained significant heritability across multiple immune cell types[45, 136], highlighting the relevant cell type and context are essential for revealing the genetic drivers of context-specific responses and associated autoimmune diseases. It is encouraging that the results of this study indicate the thawed Treg samples demonstrated responsiveness to activation at a level almost indistinguishable from the fresh samples, confirming biobanking process did not compromise the sensitivity of cells to respond to activation and epigenetic reprograming of immune responses. This supports the notion that the biobanked samples can be reliably used to identify genetic or epigenetic drivers to reveal disease mechanisms in T1D in both resting and stimulated state.

I also demonstrated the feasibility of extracting both the chromatin accessibility and transcriptome information simultaneously from a single population of 50,000 cells. This workflow would greatly benefit experiments constrained by limited input material such as using rare cell types or paediatric samples. In addition to understanding how epigenetics is altered in autoimmune disease, combining ATAC-seq and RNA-seq allows us to link altered regulatory elements to their target genes to identify molecular mechanisms leading to the loss of tolerance.

In ATAC-seq protocol, upon cell lysis, centrifugation is carried out separating the lysis reaction into two fractions - pellet which contains lysis products derived solely from the nucleus, which is used for Tn5 transposition in ATAC-seq, and supernatant fraction which contains lysates derived mainly from the cytoplasm and other cellular organelles such as ribosomes and endoplasmic reticulum (ER). As the supernatant fraction is not required in the ATAC-seq workflow I attempted to extract transcriptome from the supernatant lysate (referred to as

"ATAC-SN" thereafter) and asked if its representation was comparable to the whole cell lysate, which is commonly used as input in the conventional RNA-seq or microarray experiment. If it was not, what signatures were altered in ATAC-SN in the context of gene expression? As RNA material recovered from ATAC-SN constitutes RNA mainly from cytoplasmic compartment there is a lack of nuclear components, whereas total cellular extract is theoretically derived from both, we expect to see a subcellular localised expression pattern. The transcript profile from the ATAC-SN is likely to show enrichment for mRNA which resides exclusively in the cytoplasm and cell organelles, and depletion for nuclear RNAs. Although transcriptome analysis such as RNA-seq and microarray has predominantly been performed using RNA extracted from whole cells, growing evidence is emphasising the importance of investigating the subcellular repertoire of RNA molecules and understanding the spatial dimension that govern their distribution inside the cell [210-213]. Subcellular localization of mRNA species is thought to be a mechanism to spatially regulate protein production and control transport of target proteins to their functional domains within the cell [189, 214]. To date, a number of studies have demonstrated subcellular localization of RNA transcripts in various tissue types, and they revealed a small proportion of significant differences between the nuclear and cytosolic transcriptome [189-191, 210, 211, 213-217]. Work from Trask, Cowper-Sal-lari [190] concluded that to provide a legitimate profile of steady-state mRNA, the nuclear components must be excluded, or at least analysed separately from the cytoplasmic fractions for accurate quantification of gene expression. The investigators argued that by omitting a subset of variably expressed nuclear transcripts, more reproducible results were obtained and differentially expressed genes that would have been missed (or, statistically non-significant) using the whole cell fractions were recovered (nuclear dominance). Their findings are supported by a study from Battich, Stoeger [216] who showed nuclear compartmentalization acts as a global noise or fluctuations buffering mechanism of gene transcription in mammalian cells, ensuring that cytoplasmic transcript abundance is tightly regulated and minimally stochastic.

This study set out with the aim of assessing the global gene expression patterns of RNA recovered from cytoplasm-derived ATAC-seq supernatant extract (ATAC-SN) in primary human T cells. It allows us to gain an insight into the distribution of RNA transcripts in the ATAC-SN transcriptome and evaluate the reliability of using it for a complete view of gene expression levels in the cells. Given that a majority of the ATAC-SN fraction comprises cytoplasmic RNA and studies have reported that the global expression patterns for subcellular RNA components from total and cytoplasmic compartments were strongly correlated[215, 218], although some argued that whole cell RNA may be a misleading representation of cytoplasmic steady-state mRNA levels stemming from highly variable expression of nuclear RNA [190], this section of work hypothesised that the cytoplasm-derived ATAC-SN and whole cell fractions exhibited concordant distribution patterns of transcript abundance. Most of the current knowledge about subcellular localization of RNA transcripts is established using cell lines or tissues, such as the glioblastoma cell line[215], HepG2 human hepatoma cell line[189] and human brain tissue[191]. This study, using the cytoplasm-derived ATAC-seq lysate extracts, provides the first resource to contribute to the existing knowledge of subcellular RNA localization in the context of frozen primary human Treg cells.

In line with the hypothesis, this study revealed a high concordance in relative distribution of transcripts derived from ATAC-SN and total cellular extracts. The cytoplasm-derived ATAC-SN and whole-cell fractions exhibited high correlation (R=0.94; p-value $< 2.2\text{x}10^{-16}$) of transcript populations where 91.1% of transcriptome was significantly conserved in both compartments, confirming that there is minimal bias introduced by the method. The approximately 9% of non-conserved transcripts were predominantly nuclear RNAs as expected, which we have excluded from the overall analysis.

For a small minority of the transcripts that exhibited differential distribution between the ATAC-SN and whole-cell fractions, I performed gene ontology (GO) enrichment analysis for genes that were significantly more abundant in the ATAC-SN and whole cells independently.

Transcripts that showed higher expression in the whole cell transcriptome compared with ATAC-SN fraction, as expected, showed an overrepresentation of biological processes and molecular functions associated with the nucleus, including nuclear division, nuclear membrane, histone modification, chromosome organization and glycoprotein. These results are in accord with literature that investigated global gene expression patterns of subcellular repertoire of RNA molecules extracted from nuclear and cytoplasmic cellular extracts [189-191]. The enrichment of transcripts in the whole cell fraction with respect to ATAC-SN could be a result of high transcription rate for these genes, low rate of release to the cytosol and/or rapid degradation in the cytosol, as well as nuclear retention. In line with findings from Barthelson, Lambert [189], our nucleus-containing whole cell extracts also displayed selective enrichment of transcripts involved in membrane-associated biological processes, particularly those integral to the plasma membrane such as metal ion transmembrane transporter activity, as well as the endoplasmic reticulum (ER)/golgi apparatus such as glycoprotein biosynthetic process. These observations could be a result of physical association of the membrane proteins with the outer nuclear membrane as ER is composed of a continuous membrane system that includes the nuclear envelope (NE) [219]. In addition to that, mechanisms of nuclear retention have been widely proposed [220, 221]. Nuclear retention of incompletely spliced[222], mature or hyper-edited mRNAs[223] serves as a gene regulation mechanism, allowing the cell to rapidly respond to stress, viral infection, or changing environmental stimuli [221, 224]. As there is a lack of literature on the nuclear retention of mRNA in human primary cells, I compared my list of whole-cell enriched genes to the list of nuclear-retained genes identified in MIN6 pancreatic beta cell line by Bahar Halpern, Caspi [225] and discovered that approximately 70% of the differential genes in the whole-cell lysate are known nuclear retained genes. They include *SLC36A1* (Solute Carrier Family 36 Member 1), which has also been reported to show nuclear localization in smooth muscle cells (SMCs) from rats [226] and *SYNE3* (Spectrin repeat containing nuclear envelope family member 3) whose expression was also strongly localised in the nucleus of lung tissues from mice [226].

For ATAC-SN fraction where the RNA was primarily derived from the cytosol, I found that there was significant enrichment of transcripts involved in membrane-associated and translational machinery processes such as co-translational protein targeting to membrane, protein targeting to endoplasmic reticulum (ER), as well as cytoplasmic translation, catabolic and metabolic processes. Those differentially expressed genes also exhibited an overrepresentation of GO categories associated with extracellular organelles such as mitochondrion organisation, ribosomes and ubiquitin cycle, suggesting rapid changes taking place for active regulation of proteome turnover within the cell. These findings were expected given that a majority of ATAC-SN fraction comprises cytoplasmic RNA. The observations are consistent with the literature [189-191] that aim to delineate subcellular distribution of RNA from cytosol, nucleus and whole cell compartment. The enrichment of transcripts in the ATAC-SN with respect to the whole cell fraction could be attributed to their high stability in the cytoplasm and low transcription rates. The results also indicate that the ATAC-SN fraction showed higher enrichment of nuclear-encoded mitochondrial proteins (NEMPs) than whole cell, which corroborates findings from a recent study by Zaghlool, Niazi [191] indicating that NEMPs were significantly enriched in the cytosol of human brain tissues and displayed significantly longer half-life compared with other protein-coding transcripts, which supports their preferential and prolonged localization in our cytosolic ATAC-SN. These data suggest that NEMP mRNAs are transported from the nucleus after transcription is complete and accumulate in the cytoplasm until their translation is required by the mitochondria as demonstrated in Couvillion, Soto [227].

Conceptually, whole cell transcriptome should theoretically contain a similar amount of cytoplasmic RNA as the ATAC-SN's, with the only difference being whole cell fraction contains nucleus and ATAC-SN does not. Possible explanations for the differential abundance of cytoplasmic RNA populations observed between whole-cell and ATAC-SN fractions are that: (1) highly statistically significant differential expression of nuclear RNA diminishes the significant effects of the cytoplasmic RNA in the whole cell, resulting in some genes being identified as statistically non-significant in the whole cell compartment. As ATAC-SN fraction primarily constitutes cytoplasmic RNA whereas whole cell suffers from variable effects arising from nuclear RNA, cytoplasmic transcripts in the ATAC-SN are "enriched" in a statistical sense due to low heterogeneity of RNA populations (for ATAC-SN enriched genes). (2) library prep using RNA recovered from ATAC-SN involves additional processing steps, such as ATAC-seq lysis and concentration of RNA pooled from multiple high volumes of ATAC-seq supernatant fraction, which can be the source of material loss (for whole-cell enriched genes). In accordance with the hypothesis, Trask, Cowper-Sal-lari [190] reported that approximately 55-76% of the significantly differentially expressed genes between nuclear and whole cell compartment were dominated by the nuclear component and not found in the cytoplasm (a phenomenon known as nuclear dominance).

Overall, the transcriptome derived from ATAC-seq supernatant fraction closely recapitulates the cytosolic transcript profile of the whole cell transcriptome. It should be noted that there is a formal possibility that there is a disease-linked role for some of the nuclear transcripts that are excluded by selecting the ATAC-SN methodology but given that the majority of SNPs are not in gene bodies, this is not a major concern. As the ATAC-seq dataset includes all of these regions, it would be formally possible to analyse only these nuclear enriched regions for differential accessibility, but that is beyond the scope of this study. One limitation inherent in this study was the higher enrichment of duplicates (i.e. lower complexity) in the ATAC-SN samples in comparison with whole RNA samples, which could be attributed to the lower starting input RNA amount and possibly longer sample processing time for this method.

This finding has significant implications for studies that aim to obtain chromatin accessibility and transcriptome from low input resources such as rare cell types or biobanked paediatric samples. This work would benefit from a more streamlined RNA isolation and purification system tailored for dilute, high reaction volumes such as ATAC-seq lysate reaction. Newer single cell technologies that were not cost effective at the commencement of this study have become more economical and could be applied in a similar manner.

CHAPTER 4: BENCHMARKING AND ESTABLISHING BIOINFORMATICS DATA ANALYSIS PIPELINES

4.1 Introduction

ATAC-seq shows numerous advantages over other standard chromatin profiling assays as a readout of gene regulatory activity, owing to its low starting input material and simple protocol. The protocol generates high quality data that demonstrates high signal-to-noise ratio, statistical sensitivity[22] and allows accurate inference of transcription factor (TF) binding at high resolution[21, 25, 142, 143], making it an attractive alternative to DNase-seq. Compared with DNase-seq or ChIP-seq it is a relatively recent genome-wide technology that probes chromatin structure (accessibility and nucleosome positioning) and TF occupancy. At the outset of this project there were limited analysis software or bioinformatics tools available for processing or analyzing ATAC-seq data, although the analysis framework for ATAC-seq is similar to DNase-seq and ChIP-seq. Most available tools were designed for upstream processing of sequencing data such as adapter trimming, read alignment or removal of duplicates, whereas computational tools dedicated for ATAC-seq analyses on peak calling, data normalisation, differential accessibility analysis, functional annotation and integration with other types of genome-wide sequencing assays are lacking and underdeveloped. Although some popular computational algorithms commonly used in genome-enrichment assays such as MACS2[228, 229] (calls TF/histone modification peaks from ChIP-seq data) and edgeR [38] (identifies differentially expressed genes in RNA-seq data) offers the flexibility of cross-platform application and were claimed to work for ATAC-seq data too, they failed to address the suitable parameters or adjustments needed for the programs to perform optimally on or utilise full information contained in this type of dataset. There are also very limited comprehensive data analyses desscriptions or reports for optimal practice for analysing ATAC-seq datasets, and the conventional computational tools fall short as they fail to account for additional nucleosome positioning information contained in the Tn5 transposase-digested DNA fragments. This could potentially result in erroneous data modelling contributing to, for instances, spurious peak identification or flawed control for Type I (false positive) or Type II (false negative) errors.

As my work aimed to profile the chromatin accessibility of T cells recovered from the clinical biobanked PBMC samples collected from the T1D patients and sibling-matched healthy cohort, it was critical to master the best practices of the experimental protocol as well as the computational framework for analysing ATAC-seq dataset, to ensure high-quality epigenomic data are produced for interpretation and clinical translation. This necessitated benchmarking and establishment of a robust data analysis pipeline especially one that is compatible with our study design and cell context.

This chapter of work presents the findings from the portion of my candidature which I spent benchmarking and testing the program algorithms to establish an ATAC-seq data analysis pipeline – from raw data processing to differential accessibility analysis and inference of TF binding. I generated some Omni ATAC-seq datasets using primary T cells obtained from healthy adult volunteers as part of the methodology optimisation work (see Section 3.4.2 and Section 3.4.3) and they were used for the purpose of this work.

This chapter has four main sections. First of all, I assessed the quality of the T cell Omni ATAC-seq datasets I generated in CHAPTER 3: by comparing them with published Omni-/Standard ATAC-seq datasets generated from T or B cells. I processed my ATAC-seq and published datasets with the same bioinformatics pipeline framework and performed a series of meta-analyses to assess for QC metrics such as sequencing insert size distribution, library complexity, enrichment of transcription start sites (TSS), genomic coverage of reads, reproducibility of ATAC-seq replicates and genomic annotation of ATAC-seq peaks to demonstrate competencies of producing good quality ATAC-seq libraries.

To identify enriched accessible chromatin regions above a significance threshold in ATAC-seq libraries, MACS2 [228, 229] has been commonly employed by the majority of the studies [25, 37, 41, 178] as the program of choice for calling peaks from ATAC-seq data. However, those studies used a different set of MACS2 parameters or arguments such as p-value cutoff (`--pvalue`), peak shifting model (`--nomodel`), etc from each other for peak calling and it was unclear why these discrepancies exist. The algorithm of MACS2 was originally designed to identify TF binding and histone modification from ChIP-seq data and recent work by Gaspar [159] has highlighted some flawed MACS2 algorithms in calling peaks from ATAC-seq data and criticised a large number of studies for erroneous or poor quality peak identification from regions with low fragment coverage. His work addressed common practices of using MACS2 parameters that are inappropriate to ATAC-seq dataset and suggested modifications are needed for adaptation of MACS2 for the analysis of such data type. In particular, the peak patterns from ATAC-seq do not share the same properties as ChIP-seq's as the pileup signal generated from paired-end fragments from ATAC-seq represent both nucleosome-free and nucleosome-occupied chromatin regions, whereas ChIP-seq peaks correspond to specific TF binding regions which are narrower than ATAC-seq peaks. Most of the peak callers failed to account for such differences and thus may result in erroneous data modelling and inaccurate peak identification of ATAC-seq data. Thus in the second part of this chapter I tested and benchmarked strategies for calling ATAC-seq peaks with MACS2 to identify improvement or optimal parametrization for this dataset. The algorithms proposed in Gaspar [159] were adopted and compared with the conventional MACS2 peak calling approaches.

Identification and characterization of differentially accessible regions are a critical element in my work, as the key outcome of this study was to identify differential accessible chromatin regions between T1D and healthy controls. Computational methods for identifying differential accessible peaks from ATAC-seq data were lacking when this project was initiated and it was not until very recently edgeR [38] was proposed as the software of choice for this purpose. However, edgeR statistical package was originally designed for differential expression analysis of RNA-seq expression profiles and it encompasses a wide range of statistical methodology models such as empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests. None of the publications of date has addressed the suitability and statistical considerations of applying each of these differential tests on ATAC-seq data and thus it was unclear which edgeR model was appropriate for identifying statistically differentially accessible ATAC-seq peaks between comparison groups of interest.

A more recent work by Reske, Wilson [230] highlighted that ATAC-seq data normalisation method can have a significant impact on the outcome of differential accessibility analysis. Depending on the nature of the experiments and scale of the expected chromatin alterations, conflicting results can arise from different normalisation approaches and biases within the data should be critically assessed before further interpretation of the data output.

In the last part of this chapter, using Tconv and Treg ATAC-seq datasets, I performed differential accessibility analysis using edgeR and evaluated the impact of different normalisation methods on the outcome of the differential analysis between Tconv and Treg ATAC-seq. Based on previous work, the Barry Lab has established considerable prior biological knowledge and expected molecular changes between Tconv and Treg gene regulation and thus, this model serves as a reliable proxy for assessing and benchmarking analytical methods for calculating differential accessibility.

4.2 Aims and Hypothesis

The hypothesis for this chapter is that in-house ATAC-seq libraries exhibit comparable data quality, complexity and genomic feature distribution compared with published ATAC-seq datasets.

The aims of this chapter are:

1. To assess the quality of in-house T cell Omni ATAC-seq libraries in comparison with published Omni- and Standard- ATAC-seq datasets.
2. To compare and benchmark ATAC-seq peak calling strategies with *MACS2*.
3. To establish ATAC-seq differential accessibility (DA) analysis pipeline with *edgeR*.
4. To evaluate and compare the performance of different normalisation approaches on ATAC-seq differential accessibility (DA) analysis.

4.3 Material and Methods

The datasets (available from public data repositories or supplied by collaborating lab) used for cross-validating the in-house T cell Omni ATAC-seq libraries or benchmarking analysis methods in this chapter are listed in Table 4-1.

Table 4-1. Resources table showing datasets used in this chapter of work for benchmarking and improving ATAC-seq data analysis methods.

Source	Experiment	Repository accession	Cell type	Marker	# of replicate	Sequencing instrument
Barry Lab	Omni ATAC-seq	unpublished	Primary Human Tconv cells (Thawed)	CD4^{+} CD25lo CD127^{+}	3	NextSeq 550
Barry Lab	Omni ATAC-seq	unpublished	Primary Human Treg cells (Fresh)	CD4^{+} CD25^{+} CD127^{lo}	3	NextSeq 550

Corces et al 2017	Omni ATAC-seq	SRP103230 (SRA)	Primary Human CD4^{+} T cells	CD45, CD3, CD4	2	Illumina HiSeq 4000
Henriksson et al 2017	Standard ATAC-seq	E-MTAB-6292	Primary Human CD4^{+} T cells	CD4^{+} CD62L^{+}	1	Illumina Hiseq 2500
Scharer et al 2016	Standard ATAC-seq	GSE71338	Primary Human naïve B cells (Fresh)	IgD^{+} CD19^{+} MTG^{-}CD27^{-}CD38^{+} CD24^{+}	1	Illumina Hiseq 2500
Scharer et al 2016	Standard ATAC-seq	GSE71338	Primary Human naïve B cells (Frozen)	IgD^{+} CD19^{+} MTG^{-}CD27^{-}CD38^{+} CD24^{+}	1	Illumina Hiseq 2500
Beyer Schultze Lab	Standard ATAC-seq	unpublished	Primary Human Tconv cells	CD3^{+} CD4^{+} CD25^{lo} CD127^{+}	3	Illumina Hiseq
Beyer Schultze Lab	Standard ATAC-seq	unpublished	Primary Human Treg cells	CD3^{+} CD4^{+} CD25^{+} CD127^{lo}	3	Illumina Hiseq
Barry Lab	Standard ATAC-seq	unpublished	Primary Mouse CD4^{+} T cells	CD4^{+} T Cell Negative Selection	4	NextSeq 550
ENCODE[42]	DNase-seq	ENCFF815ZJW	Primary Human naïve T cells	CD4^{+} CD45RO^{-} CD25^{-}	1	Illumina Hiseq 2000
Vahedi, Kanno [7] 2015	ChIP-seq	GSE60482	Primary Human naïve T cells	CD4^{+} CD45RA^{+} CD45RO^{-}	4	Illumina HiSeq 2000

4.3.1 Quality assessment of in-house ATAC-seq datasets

The Tconv ATAC-seq datasets used in this section of work were generated as part of the protocol optimisation described in Section 3.3.2 and Section 3.4.2.5 in an effort to determine the optimal incubation time for ATAC-seq tagmentation reaction. 50,000 Tconv cells purified from thawed PBMCs obtained from a healthy adult volunteer (3 technical replicates each) were subjected to either 30 or 45 minutes of tagmentation and sequenced to an average read depth of 16.3 million reads (± 3.7 million) per sample.

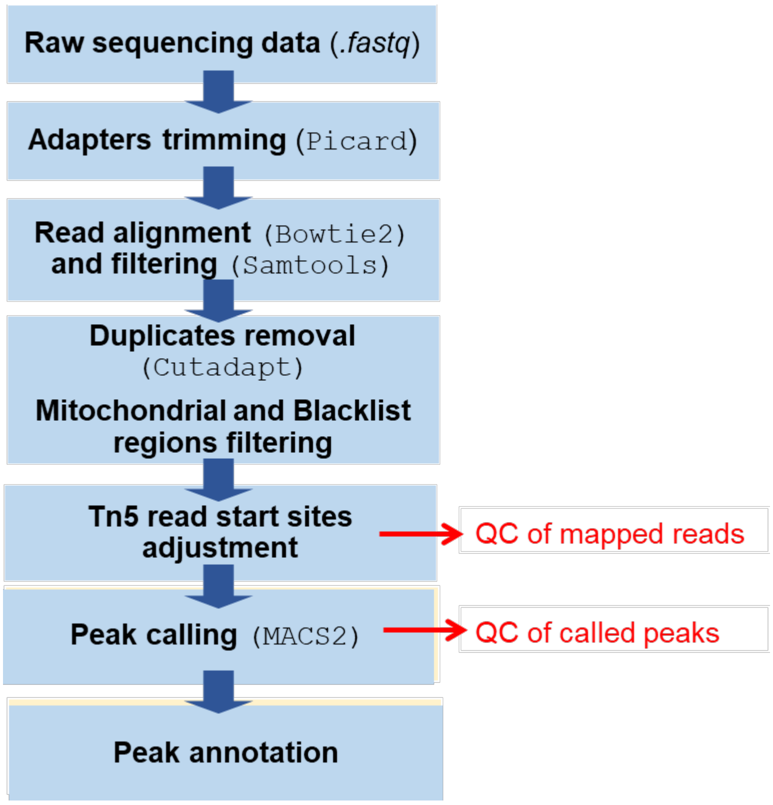


Figure 4-1 | Flowchart showing bioinformatics pipeline for ATAC-seq data processing and measurement of quality control metrics.

Raw Omni- or standard- ATAC-seq datasets were downloaded from the public data repositories (Table 4-1) and subjected to the same bioinformatics pipeline framework (Figure 4-1) as the in-house Omni ATAC-seq datasets, for evaluation of various ATAC-seq QC metrics such as fragment size distribution, library complexity, enrichment of transcription start sites (TSS), genomic coverage of reads, reproducibility of ATAC-seq replicates and genomic annotation of ATAC-seq peaks. Detailed description for the processing of ATAC-seq datasets can be found in Section 2.4.1.1.

I calculated the percentage aligning to mitochondria and the enrichment of reads at transcription start site regions (TSS enrichment) relative to 3 kb away using the RefSeq gene annotation. The insert size distribution of ATAC-seq mapped reads was computed using Picard (ver. 2.2.4), the average signal profiles and heatmaps of ATAC-seq signal relative to TSS regions were generated using Bioconductor package ChIPseeker[231], the interpolation and extrapolation of library complexity was computed using Preseq [20] and the overlap of ATAC-seq peaks to GRCh38/hg38 human reference genome, DNase-seq peaks[42], transcription start sites (TSS), T cell enhancers[7] was calculated using BEDTools (ver. 2.25.0). Irreproducibility Discovery Rate (IDR) analysis of experimental replicates was performed using the ENCODE pipeline (https://github.com/spundhir/idr). Genomic annotation of ATAC-seq peaks was performed using ChIPseeker[231].

4.3.2 Optimised ATAC-seq peak calling with MACS2

The human Treg ATAC-seq datasets used in this section of work were generated as part of the protocol optimisation described in Section 3.3.3 and 3.4.2.6, in an effort to determine the impact of freezing on chromatin accessibility in T cells. ATAC-seq libraries were generated from 50,000 Treg cells purified from fresh PBMCs obtained from three healthy adult volunteers and sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 37.2 million reads (± 6 million) per sample.

The mouse T cell ATAC-seq datasets used in this section of work were generated as part of a separate project in Barry Lab. ATAC-seq libraries were generated from 50,000 mouse CD4$^+$ T cells enriched using a MojoSort™ Mouse CD4 T Cell Isolation Kit and sequenced on a paired-end 75-cycle Illumina NextSeq 550 High-Output platform (Illumina) to an average read depth of 34.5 million reads (± 7 million) per sample.

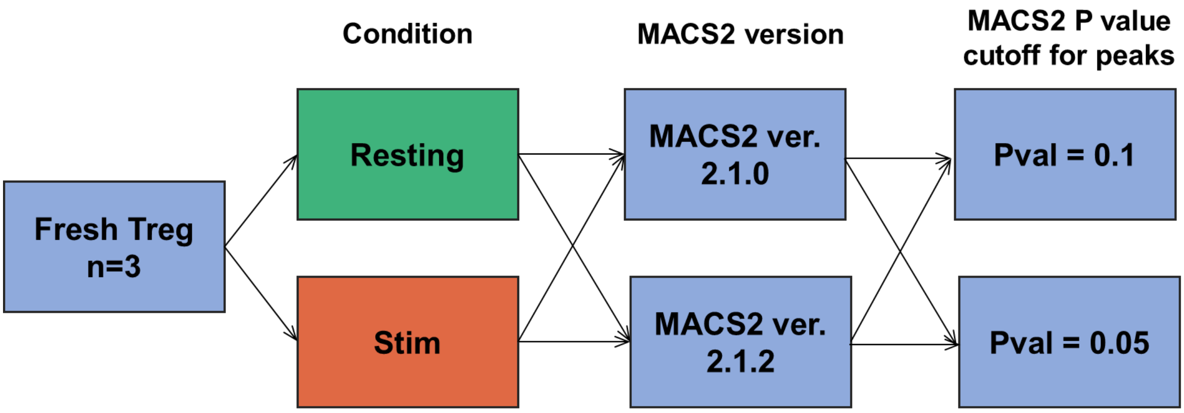


Figure 4-2 | Flowchart showing the analysis processes for benchmarking ATAC-seq peak calling strategies using MACS2.

To benchmark peak calling algorithms on ATAC-seq data, peaks were called from mapped reads pooled from three resting or stimulated Treg ATAC-seq libraries, using MACS2 ver. 2.1.0 [228] or MACS2 ver. 2.1.2 [159] (Figure 4-2). For MACS2 ver. 2.1.0, peaks were called from the merged bam files with arguments '*callpeak -f BAMPE -g hs --nolambda --nomodel --call-summits --slocal 10000 --keep-dup all*', whereas for MACS2 ver. 2.1.2, peaks were called from the merged bam files with arguments '*callpeak -f BAMPE -g hs --nolambda --min-length 100 --max-gap 50 --call-summits --bdg --keep-dup all*', at a significance cutoff of 0.1 or 0.05 *(--p 0.1 or --p 0.05*). Genomic annotation of ATAC-seq peaks was performed using ChIPseeker [231]. Fraction of reads in peaks was computed using featureCounts (v1.6.2). The overlap of peaks called from MACS2 ver. 2.1.0 and MACS2 ver. 2.1.2 was calculated using BEDTools (ver. 2.25.0) *intersect*.

4.3.3 Differential accessibility analysis of Tconv vs Treg ATAC-seq

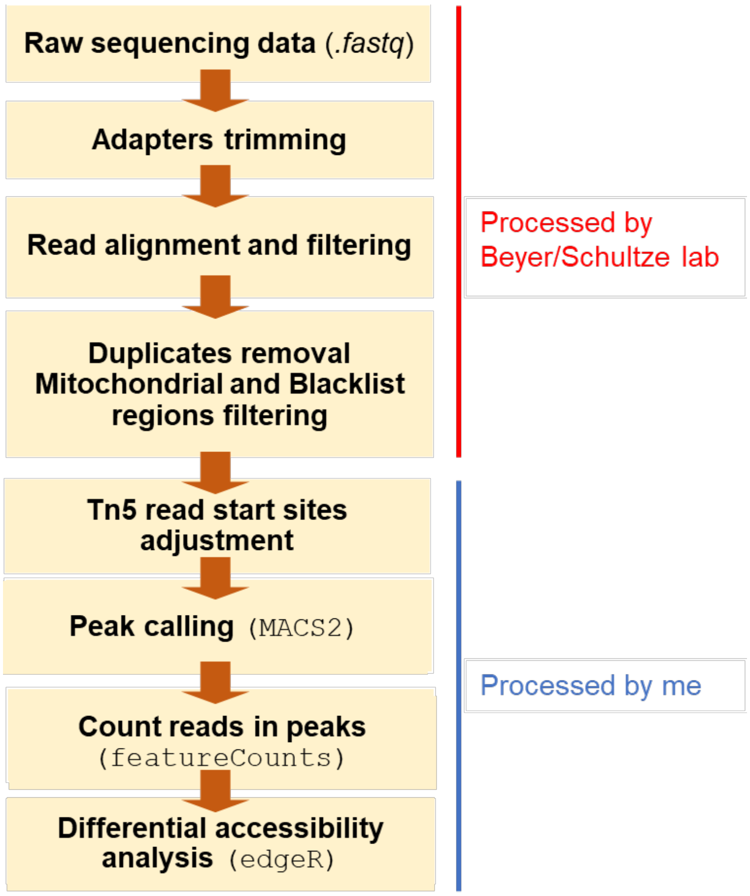

Raw sequencing data (.fastq)
Adapters trimming
Read alignment and filtering
Duplicates removal Mitochondrial and Blacklist regions filtering
Tn5 read start sites adjustment
Peak calling (MACS2)
Count reads in peaks (featureCounts)
Differential accessibility analysis (edgeR)
Processed by Beyer/Schultze lab
Processed by me

Figure 4-3 | Flowchart showing bioinformatics pipeline for ATAC-seq data processing and differential analysis.

Differential accessibility analysis in this section was performed using Tconv and Treg ATAC-seq aligned datasets (n=3 donors) kindly supplied by Professor Joachim Schultze and Dr Marc Beyer (University of Bonn, Bonn, Germany). Processing of ATAC-seq mapped reads was performed as described in Section 2.4.1.1 and Figure 4-3. Raw count processing and differential expression analysis was performed using edgeR[38] running on R. Starting with a consensus peak set (n=95,966), query peaks with fewer than 1 CPM in at least three replicates samples were eliminated from the analysis, resulting in 75,449 peaks. Normalization factors to scale the library sizes for each sample were calculated using TMM method (trimmed mean of M-values) by *calcNormFactors* function, followed by estimation of common, trended, and tagwise dispersion using *estimateDisp()*. Differential accessibility of ATAC-seq peaks was calculated using the exactTest method, with peaks having a Benjamini-Hochberg false discovery rate (FDR) of less than 0.05 being considered significant. Peaks were annotated to the nearest transcriptional start sites (TSS) using the Bioconductor package ChIPpeakAnno.

Gene Ontology (GO) enrichment analysis of significantly differentially accessible regions was performed using the default settings of *goana* function from limma [232], using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance. Results were then limited to GO terms with at least 4 steps back the root terms for each ontology (https://uofabioinformaticshub.github.io/summaries2GO/data/goSummaries.RDS).

Over-representation of KEGG pathways for significantly differentially accessible regions was performed using the default settings of *kegga* function from limma [232], using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

In order to visualise the results of the GO and KEGG pathways enrichment analyses, the R package `clusterProfiler` was used. Plots of gene-concept network were generated using *cnetplot* function from the Bioconductor package clusterProfiler [233].

Gene-Set Enrichment Analysis (GSEA)[234] was performed on the ranked list using the conventional H_0 in order to minimise issues with tied t-statistics. The gene sets used for this analysis were the complete collection defined by release 6.2 of the Molecular Signatures database (MSigDB). A master list of genesets was formed by only retaining genesets from the collections C3, C2, C5, C7 and H. GSEA was performed using the function *fgseaMultilevel* from the Bioconductor package fgsea, which allows us to exceed the results of simple sampling and calculate arbitrarily small P-values. In order to restrict results to informative gene sets, those with > 250 genes were excluded and I considered gene sets with a Bonferroni-adjusted p-value of less than 0.05 as significantly enriched in the datasets.

4.3.4 Comparison of ATAC-seq normalization methods in differential accessibility analysis

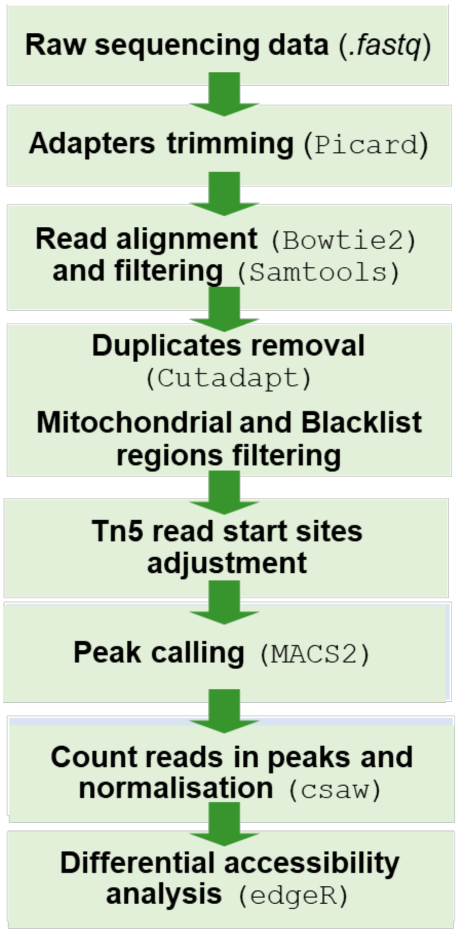

Raw sequencing data (.fastq)
Adapters trimming (Picard)
Read alignment (Bowtie2) and filtering (Samtools)
Duplicates removal (Cutadapt)
Mitochondrial and Blacklist regions filtering
Tn5 read start sites adjustment
Peak calling (MACS2)
Count reads in peaks and normalisation (csaw)
Differential accessibility analysis (edgeR)

Figure 4-4 | Flowchart showing bioinformatics pipeline for ATAC-seq data processing and analysis in benchmarking ATAC-seq normalisation methods.

This section describes the cross-comparison and evaluation of the impact of different normalisation methods (Table 4-2) on the outcome of differential analysis between Tconv and Treg ATAC-seq. Differential accessibility analysis performed in this section used Tconv and Treg ATAC-seq libraries generated from biobanked PBMCs obtained from 12 healthy subjects as part of the ATAC-seq profiling described in Chapter 5. Processing and analysis of ATAC-seq data were performed according to the workflow described in Figure 4-4 and Section 2.4.1.1. Bioconductor package csaw [162] was used to count the number of reads overlapping each ATAC-seq peak defined by MACS2 ver. 2.1.2 for each sample. Starting with a consensus peak set, average $\log_2$CPM for each query peak region was computed and query peaks with fewer than 1 CPM were eliminated from the analysis, resulting in 166,020 peaks.

Normalisation (Table 4-2) was performed using *normFactors* function from the Bioconductor package csaw using full library sizes (all mapped reads) to scale the peak library sizes for each sample. For Method 1 normalisation, count offsets were computed using the trimmed mean of M-values (TMM) method on the raw count of the ATAC-seq peaks from all the samples (Table 4-2). Method 2 normalisation implemented the trimmed mean of M values (TMM) method to generate linear scaling factors from counts in large, 10-kb genomic bins, whereas Method 3 normalisation used a non-linear loess-based (loess: locally estimated scatterplot smoothing) to compute normalisation factors as described in Reske, Wilson [230] (Table 4-2). Method 2 trims the top and bottom quantiles of bins based on fold-change and signal abundance to reduce the variability between samples at the majority of bins. TMM assumes that most peaks are not truly differentially accessible and it assesses for systematic signal differences present across the genome that are presumed to be technical [230]. Loess-based normalisation (Method 3) is a highly conservative method that adjusts the signal distribution locally based on the extent of ATAC-seq signal abundance and it assumes a symmetrical global distribution in which there are no true biological variability in ATAC-seq tagmentation efficiency and any evidence of these biases are technical and should be removed [230]. Following the indicated normalisation processes, the count matrices with the respective normalization factors or offsets were imported into R and subjected to the edgeR [38] statistical framework of estimating common, trended, and tagwise dispersions using *estimateDisp()*. Differential accessibility analysis was calculated using the exactTest method, with peaks having a Benjamini-Hochberg false discovery rate (FDR) of less than 0.05 being considered significant.

Gene Ontology (GO) enrichment analysis, over-representation of KEGG pathways of significantly differentially accessible regions as well as Gene-Set Enrichment Analysis (GSEA) [234] was performed as described in Section 2.4.3.

Table 4-2. Description of 3 normalisation strategies in calculating differential accessibility of Treg versus Tconv ATAC-seq. ***TMM, Trimmed Mean of the M-values; Loess, locally estimated scatterplot smoothing.***

Peak calling	**Quantification of reads in peaks**	**Normalisation strategy**	**DA analysis**
MACS2 **[159]**	*CSAW* [162]	TMM normalisation based on raw peak counts (Method 1)	*edgeR*
MACS2	*CSAW*	TMM normalisation based on 10-kb binned background counts (Method 2)	*edgeR*
MACS2	*CSAW*	Loess normalisation based on raw peak counts (Method 3)	*edgeR*

4.4 Results

4.4.1 Quality assessment of in-house ATAC-seq datasets

It is imperative to master the best practices of the ATAC-seq experimental protocol as well as the data analysis framework to ensure high-quality data are produced for correct interpretation and clinical translation of T1D genetic and epigenetic studies. However, computational tools for ATAC-seq-centric analyses on peak calling, data normalisation, differential accessibility analysis, functional annotation and cross-platform data integration were limited and underdeveloped. This necessitates the establishment of a robust statistical or analytical data pipeline especially one that is compatible with our study design and cell context.

ATAC-seq had not been previously performed in the Barry Lab when this project was initiated, as part of the methodology optimisation effort I generated some Omni ATAC-seq datasets using primary T cells obtained from healthy adult volunteers (see Section 3.4.2 and Section 3.4.3). To determine whether these ATAC-seq libraries meet data quality standards, I cross-validated them with other Omni- and Standard- ATAC-seq datasets available from public or collaborative genomics data repositories (see Table 4-1). I processed my ATAC-seq and published datasets with the same bioinformatics pipeline framework (Section 4.3.1 and Table 4-2) and performed a series of meta-analyses to assess for various ATAC-seq QC metrics such as fragment size distribution, library complexity, enrichment of transcription start sites (TSS), genomic coverage of reads, reproducibility of ATAC-seq replicates and genomic annotation of ATAC-seq peaks.

As described in Section 2.2 Omni ATAC-seq was chosen as the variant of choice for ATAC-seq protocol as it is compatible with frozen samples and results in substantial improvement of signal-to-background ratio and complexity in comparison with other ATAC-seq protocol variants [153]. As demonstrated in Table 4-3 and statistics summary in Table 4-4, overall the Omni ATAC-seq protocol produces accessibility data with a considerably higher proportion of uniquely mapped reads and library complexity with lower mitochondrial DNA contamination compared with the standard ATAC-seq method. Comparing the in-house human primary T cell Omni ATAC-seq libraries and the published T cell Omni ATAC-seq datasets from Corces, Trevino [41], my libraries demonstrated a higher proportion of uniquely mapped reads (72.3% ± 2.5%) and library complexity (91.1% ± 3.3%) compared with published datasets (66.4% ± 0% and 85.0% ± 1.0%) (Table 4-3 and Table 4-4). However, my libraries had higher contamination of mitochondrial DNA than the published datasets (Table 4-3 and Table 4-4).

Table 4-3. In-house and published ATAC-seq datasets used in this chapter of meta-analysis for quality assessment and benchmarking data analysis approaches.

Source of dataset	ATAC-seq variant	Experiment	Raw reads	% of uniquely mapped reads	Alignment rate (%)	Library complexity (%)	Mito-chondrial reads (%)
Barry Lab	Omni	30m_tagmentation_Rep1	11,696,222	74.37	95.26	94.8	8.0
Barry Lab	Omni	30m_tagmentation_Rep2	11,030,634	69.50	95.34	89.0	24.7
Barry Lab	Omni	30m_tagmentation_Rep3	19,924,828	71.03	95.38	89.1	19.6
Barry Lab	Omni	45m_tagmentation_Rep1	16,526,469	72.38	95.37	96.2	3.6
Barry Lab	Omni	45m_tagmentation_Rep2	20,467,119	76.55	95.52	90.3	12.0
Barry Lab	Omni	45m_tagmentation_Rep3	17,929,664	69.99	95.26	87.0	20.0
Corces et al 2017	Omni	Omni ATAC-seq CD4 T cells Rep 1	23,375,593	66.39	79.53	86.0	3.1
Corces et al 2018	Omni	Omni ATAC-seq CD4 T cells Rep 2	26,544,999	66.36	79.38	84.0	2.9
Henriksson et al 2017	Standard	Standard ATAC-seq CD4 T cells	14,425,385	58.00	98.37	70.0	68.7
Beyer, unpublished	Standard	Standard ATAC-seq Tconv cells Rep 1	45,096,412	50.50	53.74	NA	NA
Beyer, unpublished	Standard	Standard ATAC-seq Tconv cells Rep 2	51,716,189	50.93	54.30	NA	NA
Beyer, unpublished	Standard	Standard ATAC-seq Tconv cells Rep 3	27,176,779	45.08	47.53	NA	NA
Beyer, unpublished	Standard	Standard ATAC-seq Treg cells Rep 1	41,564,415	51.88	55.04	NA	NA
Beyer, unpublished	Standard	Standard ATAC-seq Treg cells Rep 2	46,345,447	41.04	43.50	NA	NA
Beyer, unpublished	Standard	Standard ATAC-seq Treg cells Rep 3	39,077,061	44.80	47.25	NA	NA
Scharer et al 2016	Standard	Standard ATAC-seq Fresh B cells	9,417,663	68.74	98.79	72.0	58.0
Scharer et al 2017	Standard	Standard ATAC-seq Frozen B cells	11,702,977	76.40	98.65	85.7	36.0
Barry Lab	Omni	Omni ATAC-seq mouse CD4 T cells Rep 1	37,503,850	66.39	85.4	91	2.490838731
Barry Lab	Omni	Omni ATAC-seq mouse CD4 T cells Rep 2	32,187,632	57.08	74.81	92	2.381384134
Barry Lab	Omni	Omni ATAC-seq mouse CD4 T cells Rep 3	24,629,585	62.91	84.18	92	3.007217706
Barry Lab	Omni	Omni ATAC-seq mouse CD4 T cells Rep 4	43,686,994	61.16	82.85	94	2.352859485

Table 4-4. Statistics of QC metrics output from Table 4-3.

	Omni ATAC-seq	**Standard ATAC-seq**
Sample size	12	9
	Mean (%) ± SD	
Mitochondrial reads (%)	8.7 (± 7.9)	54.2 (± 13.6)
Uniquely mapped reads (%)	67.8 (± 5.4)	54.2 (± 11.0)
Library complexity (%)	90.5 (± 73.5)	75.9 (± 7.0)

	Ying Omni ATAC (human CD4$^+$ T cells)	**Corces Omni ATAC (human CD4$^+$ T cells)**
Sample size	6	2
	Mean (%) ± SD	
Mitochondrial reads (%)	14.7 (± 7.4)	3.0 (± 0.1)
Uniquely mapped reads (%)	72.3 (± 2.5)	66.4 (± 0)
Library complexity (%)	91.1 (± 3.3)	85.0 (± 1.0)

The fragment size distribution of in-house Omni ATAC-seq libraries (Figure 7-18; a-b) was highly comparable with the published T cell Omni ATAC-seq datasets (Figure 7-18; c), which showed a multi-modal distribution with a clear periodicity of approximately 200bp, representing DNA fragments protected by an integer number of phased nucleosomes, however, the standard ATAC-seq libraries generated from B cells (Figure 7-18; d-e) demonstrated variability in the enrichment of nucleosomal binding pattern between the fresh and frozen conditions, as well as in comparison with T cell Omni libraries. Reassuringly, my libraries also showed similar enrichment of accessibility signal at transcriptional start sites as the published datasets (Figure 4-5) and were highly complex (Figure 4-6; Figure 7-19), even at sequencing depth several orders of magnitude higher than the original run (Figure 4-7). Although the peaks identified from the in-house ATAC-seq libraries had lower genome coverage (Figure 4-8) and lower overlap with the published T cell DNase-seq [42] peaks (Figure 4-9) in comparison with the published Omni ATAC-seq libraries, they showed comparable enrichment at known accessible chromatin signatures such as transcriptional start sites (TSS) (Figure 4-10) and enhancer elements (Figure 4-11) identified from primary human CD4$^+$ T cells [7]. DNase-seq is a “gold-standard” in profiling chromatin accessibility and was commonly used as cross-validation of ATAC-seq analysis.

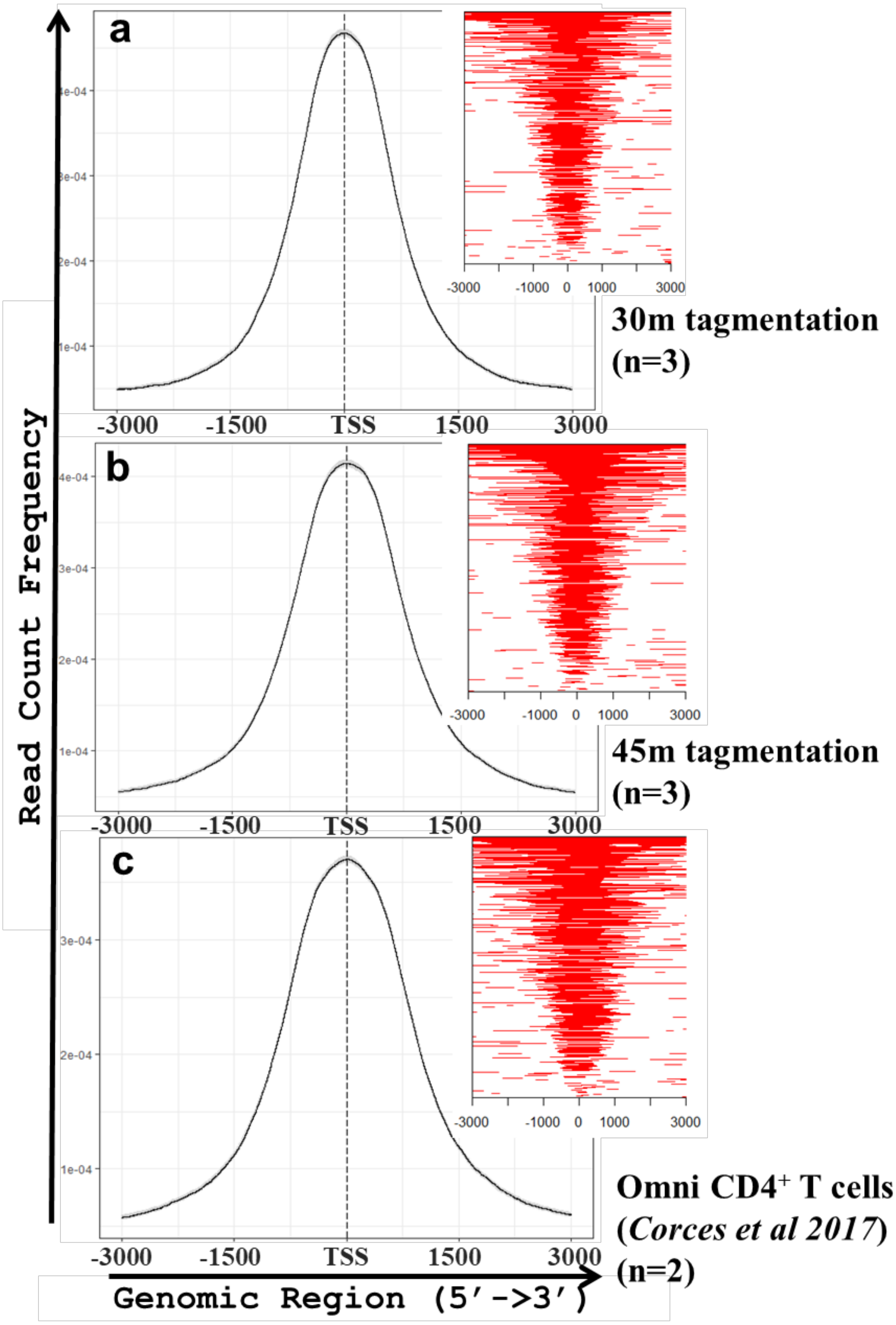


Figure 4-5 | Representative enrichment plots of Tn5 insertions at ± 3 kb transcription start sites (TSS enrichment) for Omni ATAC-seq datasets.

The enrichment of the ATAC-seq fragments at transcription start sites (TSS) over genomic background is a valuable indicator for the signal-to-noise ratio in ATAC-seq data. The data shown are representative of reads pooled from three technical replicates for (a, b) and two replicates for (c).

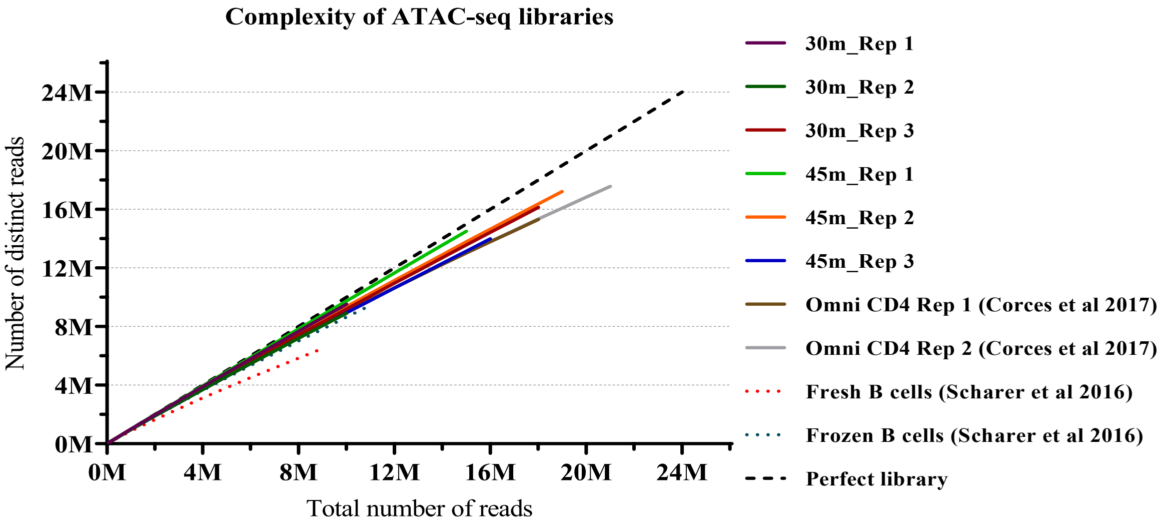


Figure 4-6 | Library complexity of in-house (30m, 45m tagmentation) and published ATAC-seq libraries.
The black diagonal line represents an ideal library, in which every read is a distinct read. The data shown are reads from single replicates.

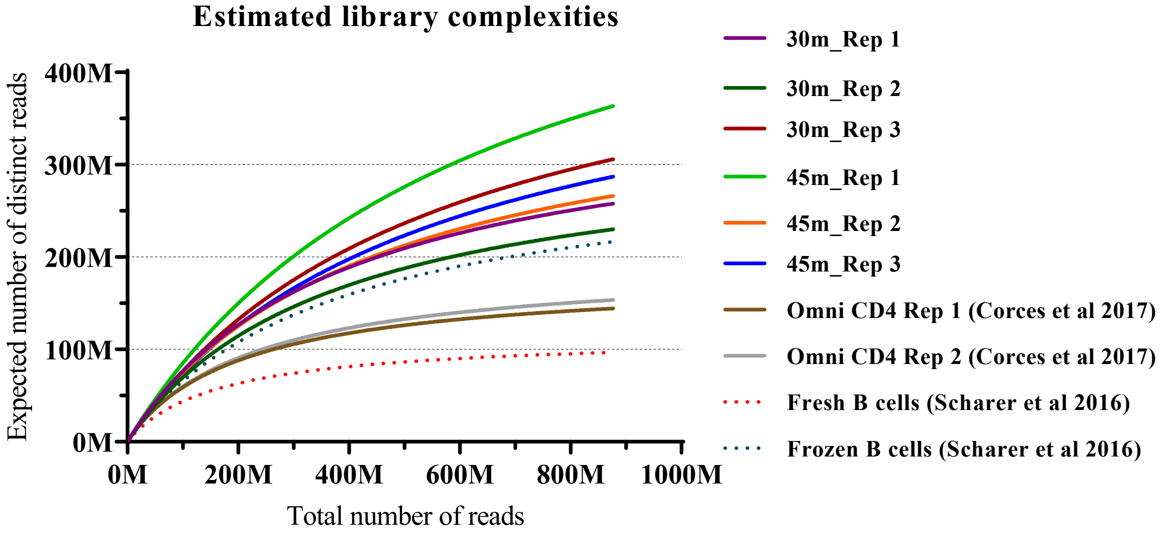


Figure 4-7 | Library complexity and extrapolated yield curve of in-house (30m, 45m tagmentation) and published T- and B- cell ATAC-seq libraries.
This calculation [20] uses rational function approximations of Good & Toulmin's non-parametric empirical Bayes estimator to predict the library complexity of future experiments several orders of magnitude larger than the initial run.

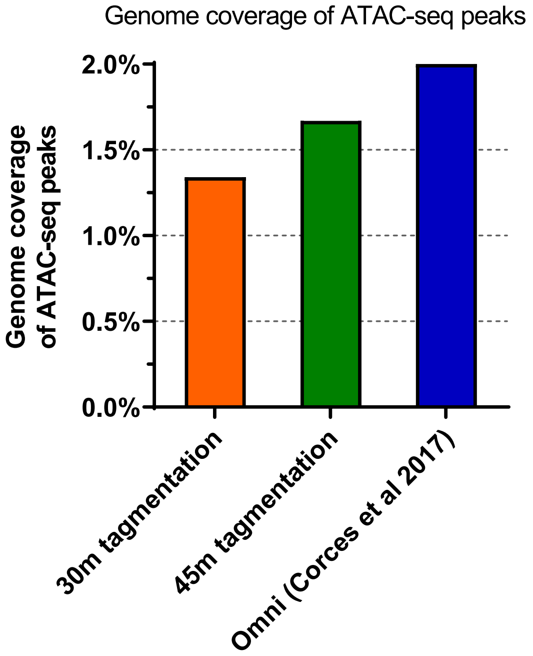


Figure 4-8 | Genomic coverage of ATAC-seq peaks for in-house and published Omni ATAC-seq libraries.
ATAC-seq peaks were called from mapped reads pooled from three technical replicates for tagmentation libraries and two replicates for published Omni ATAC-seq datasets [41].

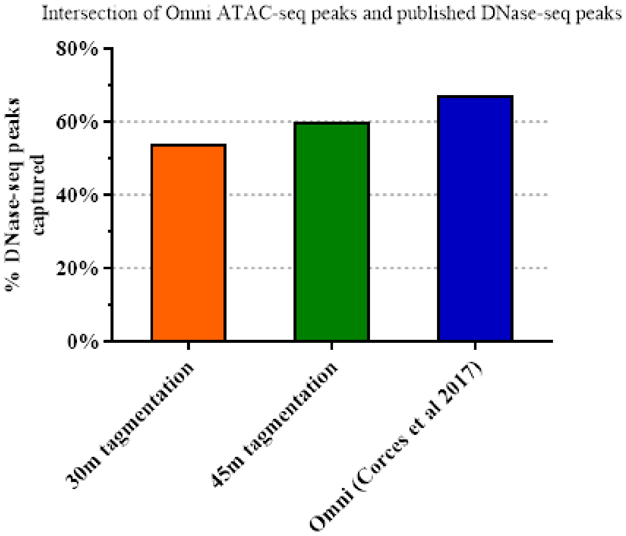


Figure 4-9 | Percentage of ATAC-seq peaks intersecting human DNase-seq peaks.
ATAC-seq peaks were called from mapped reads pooled from three technical replicates for tagmentation libraries and two replicates for published Omni ATAC-seq datasets [41]. DNase-seq dataset was generated from human adult primary CD4^{+} CD45RO^{-} CD25^{-} naïve T cells and available from ENCODE database [42].

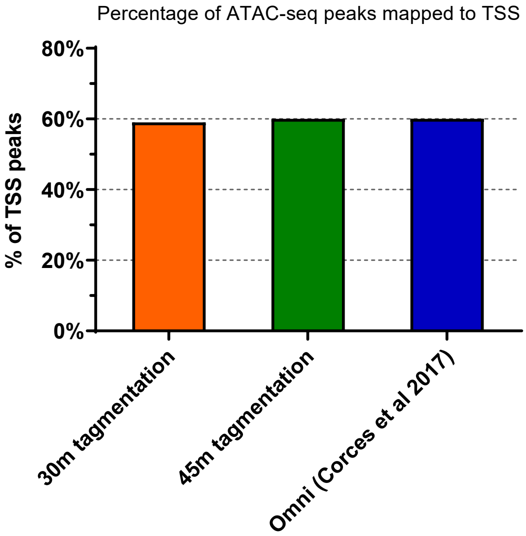


Figure 4-10 | Percentage of ATAC-seq peaks overlapping transcription start sites (TSS) of human reference genome.
ATAC-seq peaks were called from mapped reads pooled from three technical replicates for tagmentation libraries and two replicates for published Omni ATAC-seq datasets [41].

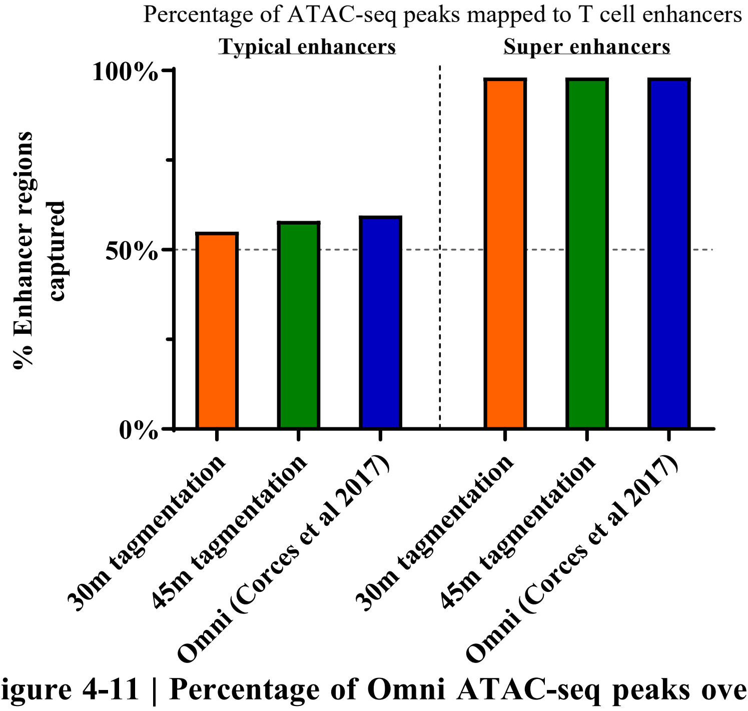
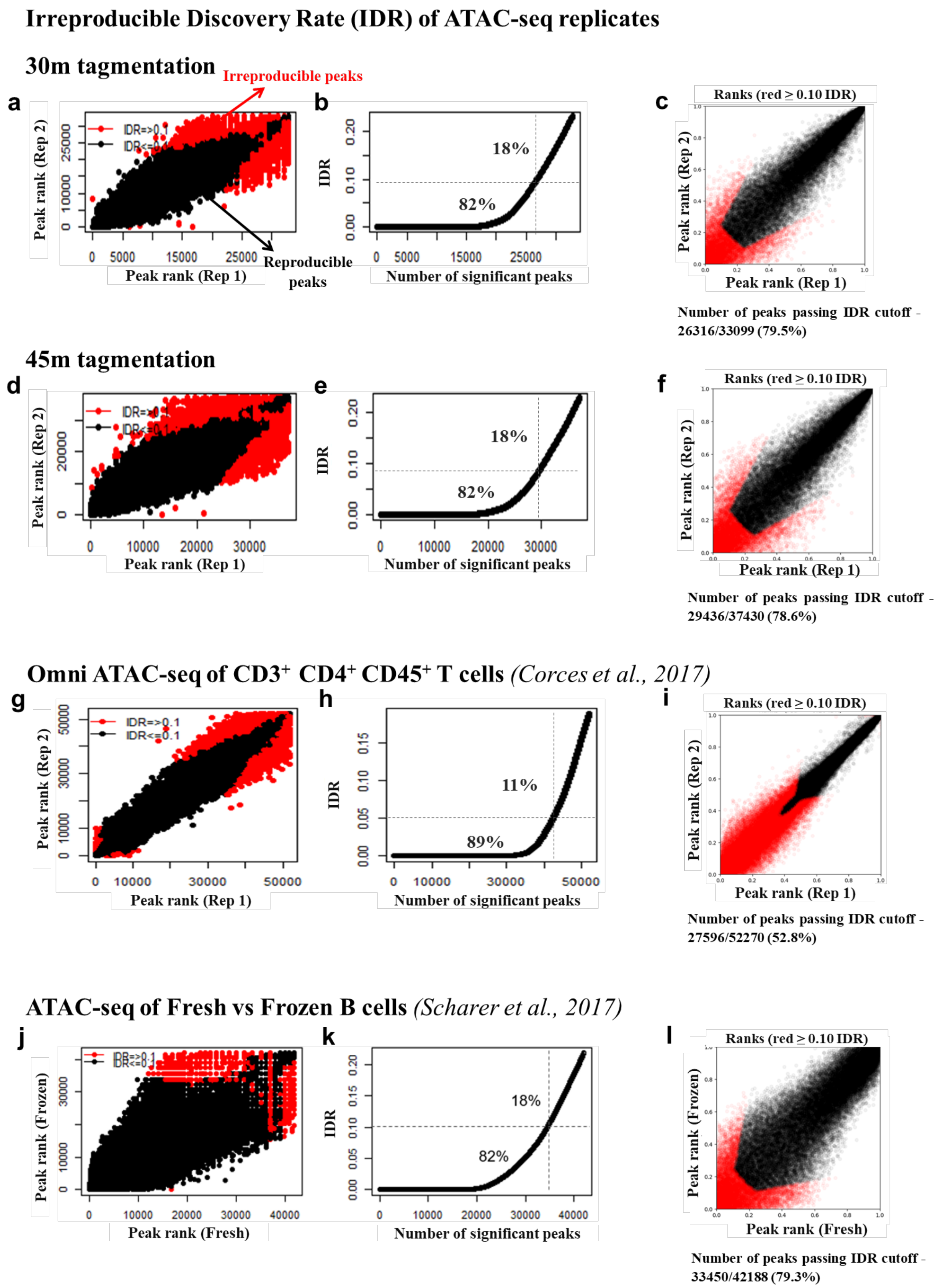


Figure 4-11 | Percentage of Omni ATAC-seq peaks overlapping human T cell typical enhancers and super enhancers.
ATAC-seq peaks were called from mapped reads pooled from three technical replicates for tagmentation libraries and two replicates for published Omni ATAC-seq datasets [41]. Human T cell enhancer datasets were generated from naïve CD4^{+} CD45RA^{+} CD45RO^{-} primary T cell population and available from Vahedi, Kanno [7] study.

IDR (Irreproducibility Discovery Rate) analysis is extensively used by the ENCODE consortium [42] as part of their criteria and standards for submission of sequencing datasets. It is used as a measure of reproducibility or concordance across experimental replicates and in ATAC-seq, it ranks the peaks based on the consistency of the peak calls between the replicates. The most significant peaks, which are likely to be genuine signals, have high consistency between replicates, whereas peaks with low significance, which are more likely to be noise, have low consistency between replicates. A high proportion of peaks were reliably identified between the experimental replicates in both the in-house 30m (Figure 4-12; a-c) and 45m (Figure 4-12; d-f) tagmentation libraries, based on measurement of statistical significance (Figure 4-12; a, d, g, j) and average enrichment signal (Figure 4-12; c, f, i, l) of peaks, at a level similar to the published ATAC-seq datasets (Figure 4-12; g-l).



Irreproducible Discovery Rate (IDR) of ATAC-seq replicates

30m tagmentation

Number of peaks passing IDR cutoff - 26316/33099 (79.5%)

45m tagmentation

Number of peaks passing IDR cutoff - 29436/37430 (78.6%)

Omni ATAC-seq of CD3$^+$ CD4$^+$ CD45$^+$ T cells *(Corces et al., 2017)*

Number of peaks passing IDR cutoff - 27596/52270 (52.8%)

ATAC-seq of Fresh vs Frozen B cells *(Scharer et al., 2017)*

Number of peaks passing IDR cutoff - 33450/42188 (79.3%)

Figure 4-12 | Irreproducibility Discovery Rate (IDR) analysis for measuring reproducibility and concordance of ATAC-seq peaks between replicates.
Scatter plots showing consistency between a pair of ATAC-seq peaks in the replicates ranked by p-value (a, d, g, j) and signal value (c, f, i, l). The estimated IDR as a function of different rank thresholds is shown in the consistency transition plots (b, e, h, k) which provides an indicator of the transition from signal to noise and suggests how many peaks have been reliably detected.

Peaks from the in-house tagmentation libraries show comparable distribution of genomic features such as promoters and intergenic regions (Figure 4-13; Figure 4-14), as well as distribution of peaks relative to TSS (Figure 4-15), with the published and collaborative datasets.

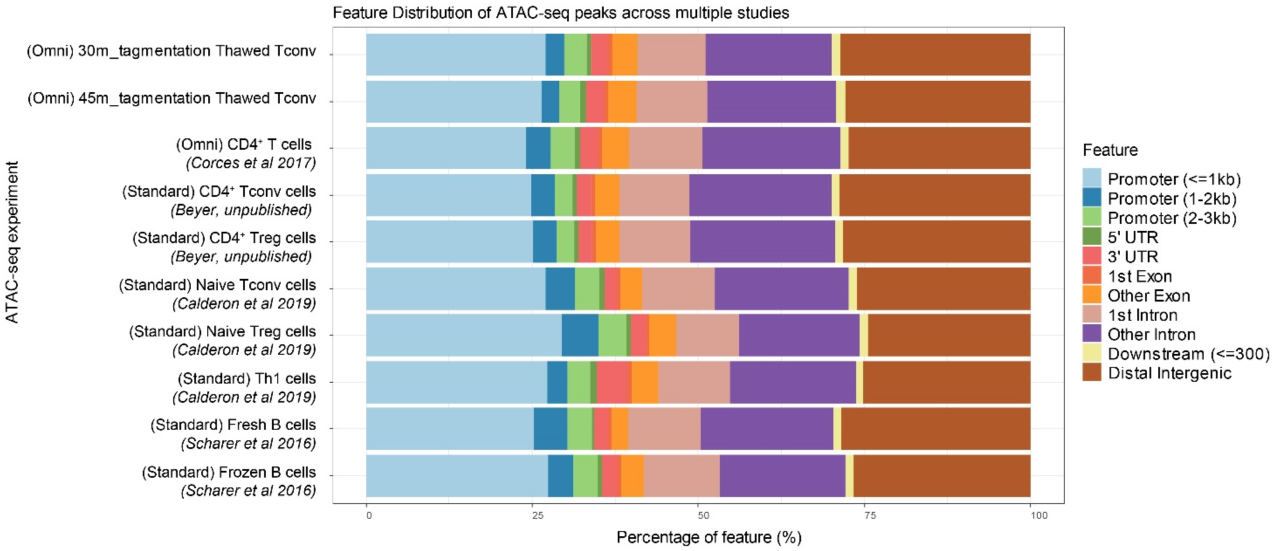


Figure 4-13 | Comparison of genomic annotation of ATAC-seq peaks for in-house tagmentation and other ATAC-seq datasets.

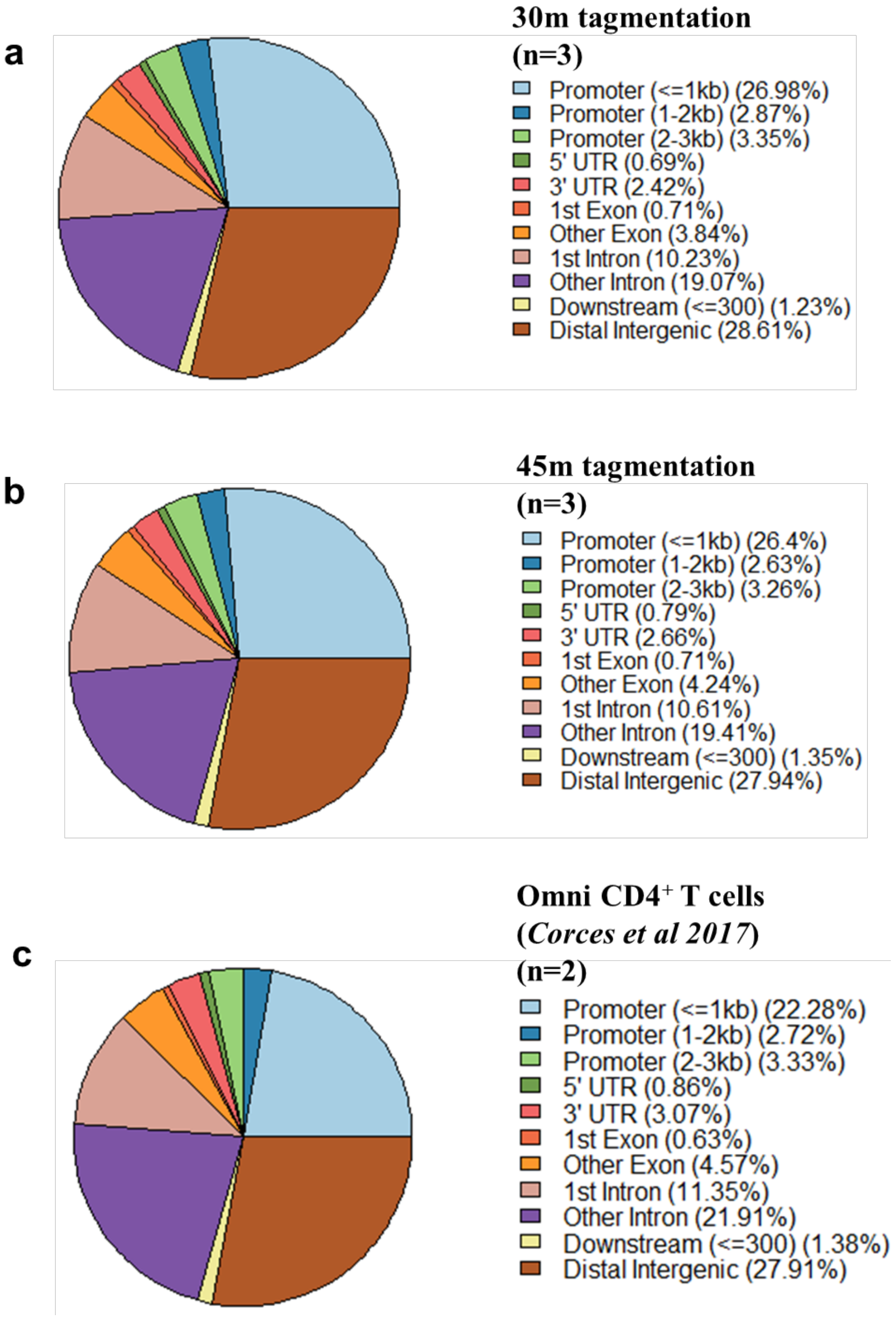


Figure 4-14 | Pie charts showing genomic annotation and percentages for peaks called from tagmentation and published Omni ATAC-seq datasets (same datasets as Figure 4-13 but only selected libraries are shown for ease of visualization).

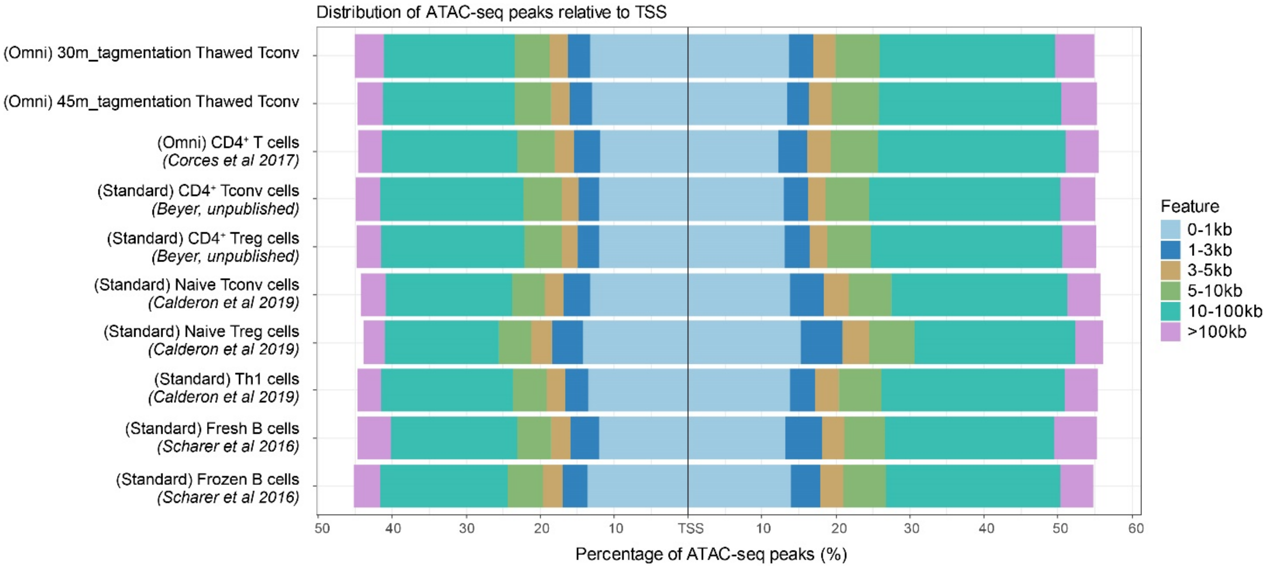


Figure 4-15 | Comparison of distribution of ATAC-seq peaks relative to transcriptional start sites (TSS) for in-house tagmentation and other ATAC-seq datasets.

4.4.2 Optimised ATAC-seq peak calling with MACS2

To identify enriched accessible chromatin regions above a statistical significance threshold, MACS2 [228], which is a popular program for calling peaks especially from ATAC-seq data has been commonly employed by the majority of the studies [25, 37, 41, 178].

Study by Gaspar [159] reviewed the source code of MACS2 and revealed that a large number of published ATAC-seq datasets including the recent Omni ATAC-seq publication by Corces, Trevino [41], had used MACS2 modelling to call peaks in mode that was inappropriate and suboptimal to the ATAC-seq datasets, resulting in erroneous or poor quality peak identification. For instance, Corces, Trevino [41] used the command-line arguments `--format BAM` and `–-nomodel` in calling peaks from their paired-end ATAC-seq datasets, `--format BAM` instructs the program to analyse the datasets in a single-end mode and ignore the second read of every read pair, whereas `–nomodel` extends every "single-end" R1 reads to 200bp by default. This results in false reporting of coverage values of the identified peaks as ATAC-seq libraries have a large window fragment size distribution containing highly diverse DNA fragment lengths ranging from 100bp to 1000bp. Furthermore, their strategy failed to utilise the full information contained in the ATAC-seq dataset as the second reads of the entire data were ignored. It was shown that those incorrect peak identification resulting from single-end analysis mode and default peak extension model occurs both in the close proximity of true peaks, as well as regions where no signal enrichment was observed.

Gaspar [159] then modified the original MACS2 (MACS2 ver. 2.1.0) source code to accept additional arguments that are optimal to ATAC-seq datasets with increased overlap between single-end and paired-end analysis modes (referred to as "MACS2 ver. 2.1.2"). This section of work presents and compares the findings from testing MACS2 ver. 2.1.0 and MACS2 ver. 2.1.2 using the human Treg (see Section 3.3.3 for preparation) and mouse $CD4^+$ T cell ATAC-seq datasets.

Using MACS2 ver. 2.1.2 developed by Gaspar [159] a higher number of peaks were identified from Treg cells both during resting and in response to stimulation compared with MACS2 ver. 2.1.0, at both p-value cutoffs of 0.1 and 0.05 (Figure 4-16, Figure 4-17). Two peak statistical thresholds (`--p` option) were tested in this analysis as several ATAC-seq publications have adopted different significance cutoffs for peak identification, and it was of interest to evaluate the differences between the two algorithms in terms of peak number or proportion on annotated genomic features. Overall MACS2 ver. 2.1.2 resulted in a higher proportion of peaks identified from the intergenic regions and a lower proportion of peaks assigned to the promoters compared with MACS2 ver. 2.1.0 (Figure 4-17). One distinct difference between peak calling from MACS2 ver. 2.1.0 and MACS2 ver. 2.1.2 was that peaks identified from MACS2 ver. 2.1.2 had smaller peak size or coverage (represented as medians and means; Figure 4-18; red inset) compared with MACS2 ver. 2.1.0. One explanation for this observation is that the peaks called by MACS2 ver. 2.1.2 were sharper and more resolved than MACS2 ver. 2.1.0 which better represented accessible chromatin regions (Figure 4-19; a). Regions with low accessibility and/or high background noise were not identified as peaks in MACS2 ver. 2.1.2 but MACS2 ver. 2.1.0 (Figure 4-19; a). Stimulation leads to increased chromatin accessibility and higher coverage of identified peaks compared with resting state in both peak-calling algorithms (Figure 4-18). As expected a higher fraction of ATAC-seq reads was captured in the stimulated compared with resting Treg cells (Table 4-5). In comparison with peaks detected by MACS2 ver. 2.1.0, a lower fraction of reads from resting cells and a higher fraction of reads from stimulated cells were captured by peaks called from MACS2 ver. 2.1.2 (Table 4-5). MACS2 ver. 2.1.2 was also able to identify high signal coverage region which was missed by MACS2 ver. 2.1.0 as peak (Figure 4-19; b). When both algorithms were used to call peaks from ATAC-seq datasets generated mouse CD4^{+} T cells, I observed a greater number of identified peaks (mean of 41.6% ± 9.5%) resulting in a mean of 56.4% (± 8.9%) overlap between peak sets generated by the two models using MACS2 ver. 2.1.2 compared with ver. 2.1.0. (Figure 4-20).

In summary peak calling using MACS 2.1.2 was selected for further data analysis because it identified a greater enrichment of enhancer/ intergenic regions, which are likely to contain more of the genetic risk that we wish to functionally annotate. Furthermore, the peaks called by MACS 2.1.2 more accurately represented the chromatin accessibility signal of the cells as lower false positive and negative were observed in comparion with peaks identified from MACS 2.1.0.

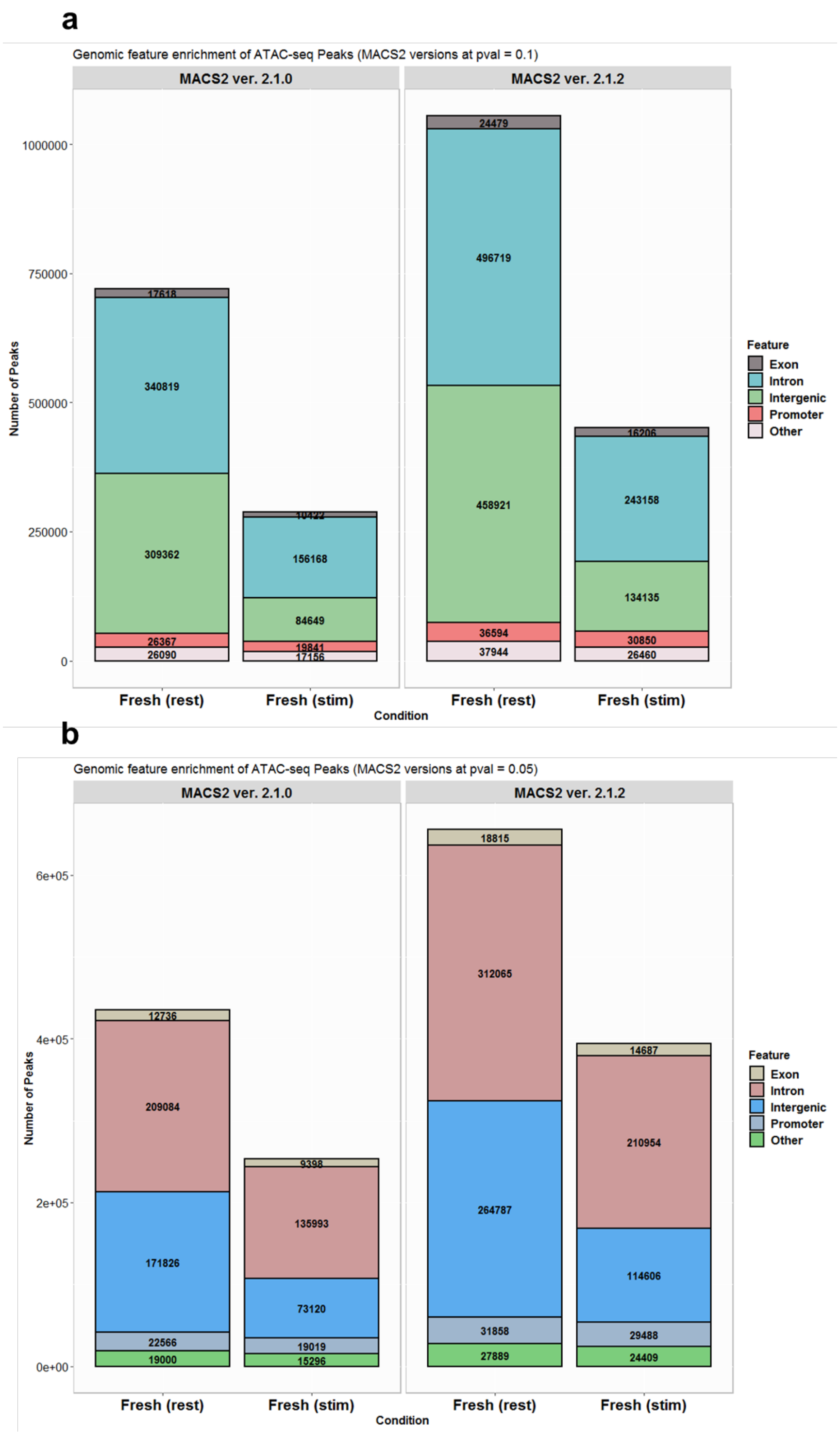


Figure 4-16 | Comparison of genomic annotation for peaks called (expressed as numbers) at significance threshold or p-value cutoff (`--pvalue`) of 0.1 (a) and 0.05 (b) using MACS2 ver. 210 (left panel) and ver. 2.1.2 (right panel).

ATAC-seq datasets were generated from resting and stimulated primary Treg cells purified from three human adult donors (see Section 3.3.3) and peaks were called from mapped reads pooled from three replicates. *Ver, version. Others, 3UTR, miRNA, ncRNA, TTS, pseudo, 5UTR, snoRNA, snRNA and rRNA.*

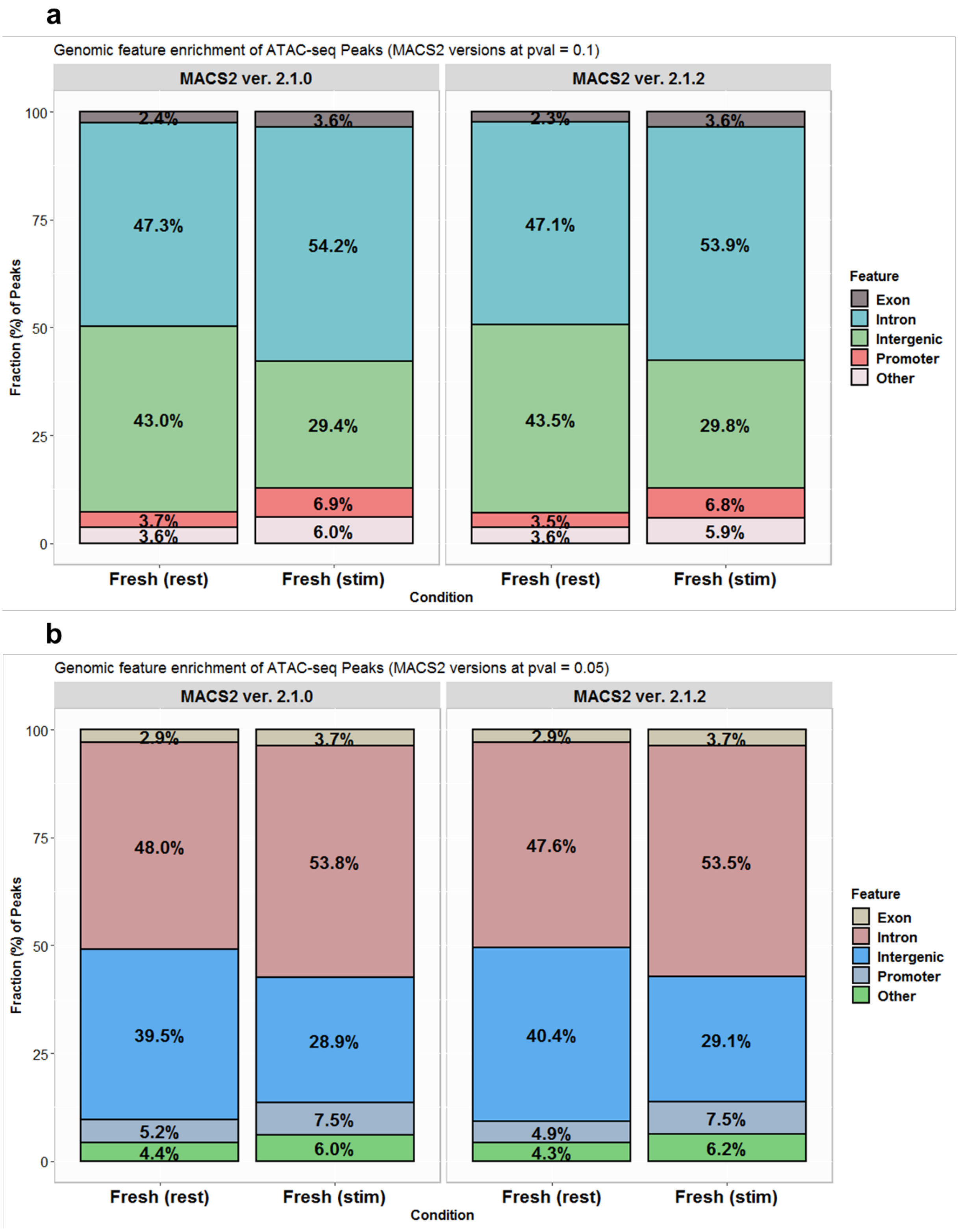
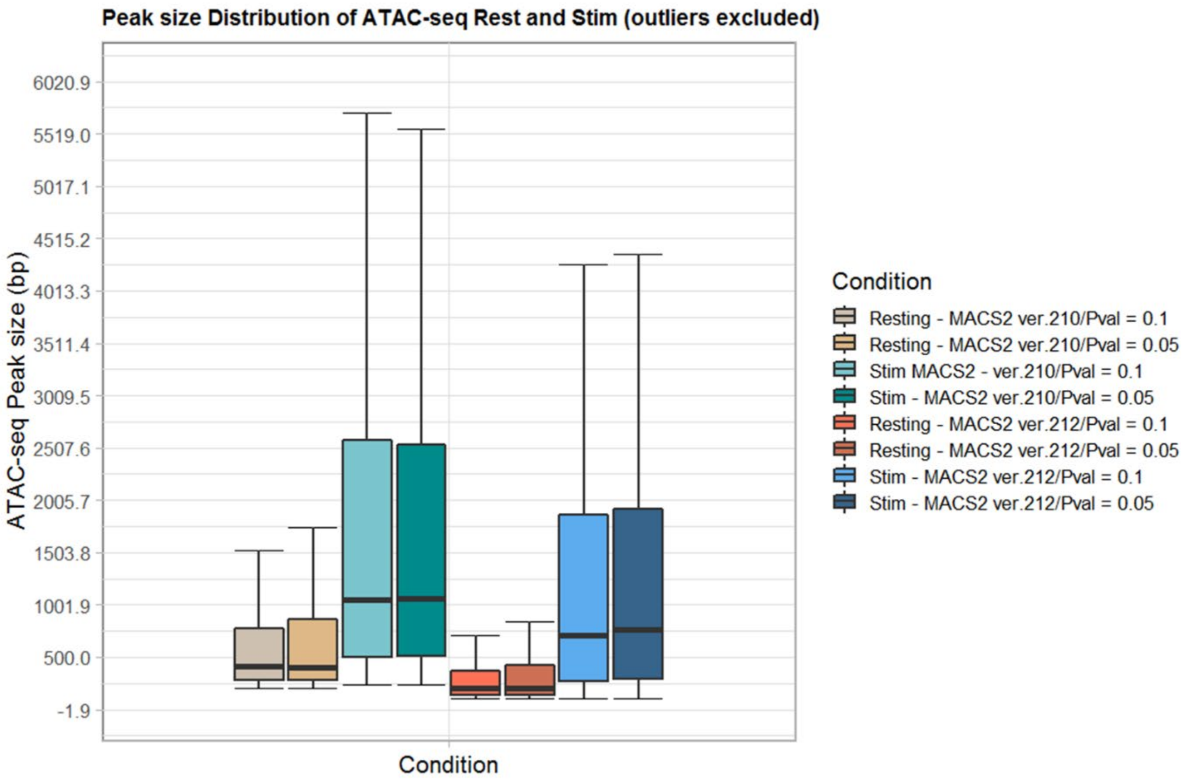
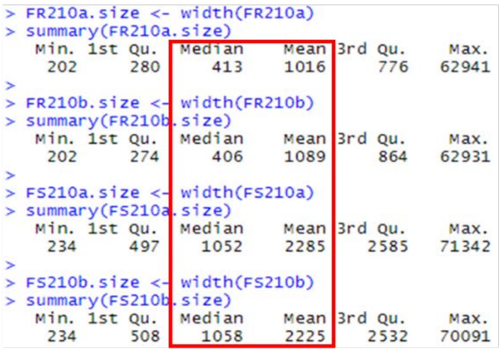


Figure 4-17 | Comparison of genomic annotation for peaks called (expressed as fractions) at significance threshold or p-value cutoff (`--pvalue`) of 0.1 (a) and 0.05 (b) using MACS2 ver. 210 (left panel) and ver. 2.1.2 (right panel).

ATAC-seq datasets were generated from resting and stimulated Treg cells purified from three adult donors (see Section 3.3.3) and peaks were called from mapped reads pooled from three replicates. *Ver, version. Others, 3UTR, miRNA, ncRNA, TTS, pseudo, 5UTR, snoRNA, snRNA and rRNA.*

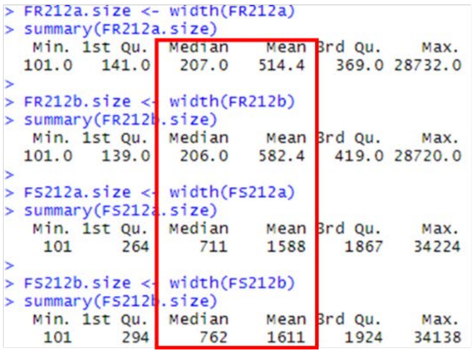


MACS2 ver. 210



```
> FR210a.size <- width(FR210a)
> summary(FR210a.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    202     280     413    1016     776   62941 
> 
> FR210b.size <- width(FR210b)
> summary(FR210b.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    202     274     406    1089     864   62931 
> 
> FS210a.size <- width(FS210a)
> summary(FS210a.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    234     497    1052    2285    2585   71342 
> 
> FS210b.size <- width(FS210b)
> summary(FS210b.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    234     508    1058    2225    2532   70091 
```

MACS2 ver. 212



```
> FR212a.size <- width(FR212a)
> summary(FR212a.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  101.0   141.0   207.0   514.4   369.0 28732.0 
> 
> FR212b.size <- width(FR212b)
> summary(FR212b.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  101.0   139.0   206.0   582.4   419.0 28720.0 
> 
> FS212a.size <- width(FS212a)
> summary(FS212a.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    101     264     711    1588    1867   34224 
> 
> FS212b.size <- width(FS212b)
> summary(FS212b.size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    101     294     762    1611    1924   34138 
```

a=pval 0.1; b=pval 0.05

Figure 4-18 | Size distribution of ATAC-seq peaks called from resting and stimulated Treg cells, at significance threshold or p-value cutoff (`--pvalue`) of 0.1 and 0.05 from MACS2 ver. 210 and ver. 2.1.2.

ATAC-seq datasets were generated from resting and stimulated Treg cells purified from three adult donors (see Section 3.3.3) and peaks were called from mapped reads pooled from three replicates. *Ver, version; F, fresh; R, rest; S, stim; a, p-value at 0.1; b, p-value at 0.05.*

Table 4-5. Fraction of reads in ATAC-seq peaks called from individual resting and stimulated Treg replicates (see Section 3.3.3) using MACS2 of indicated version.

	MACS2 ver.210	MACS2 ver.212
Fresh Treg D1 Rest	29.7	27.9
Fresh Treg D2 Rest	32.5	30.7
Fresh Treg D3 Rest	33.5	31.7
Fresh Treg D1 Stim	56.9	57.1
Fresh Treg D2 Stim	55.6	55.8
Fresh Treg D3 Stim	68.3	68.5

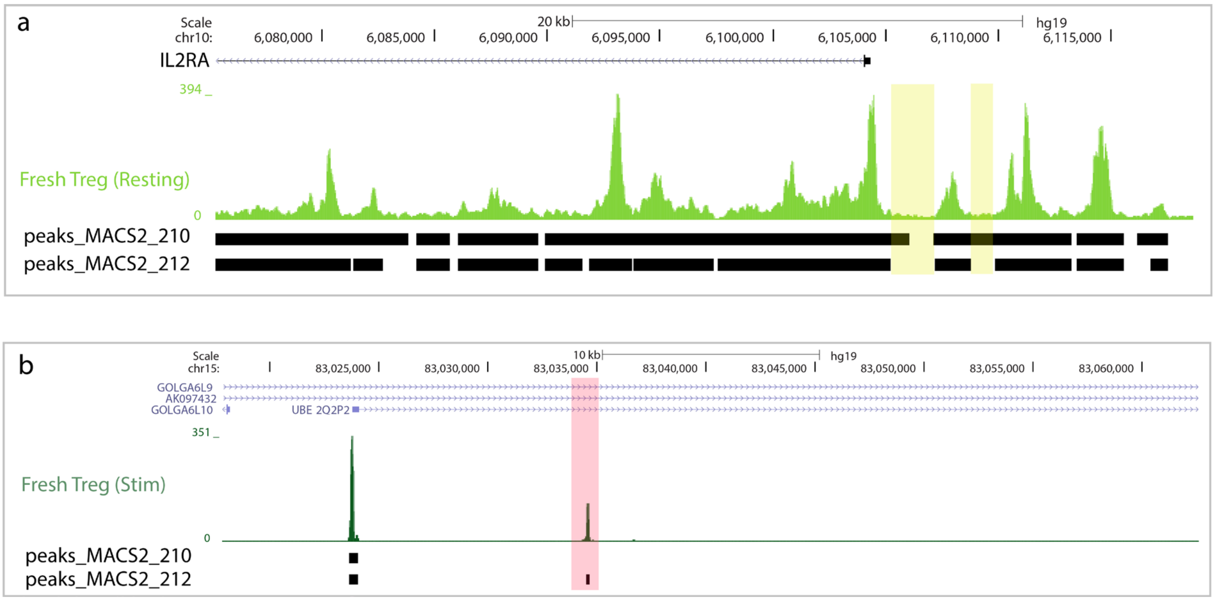


Figure 4-19 | Genome browser view showing representative ATAC-seq peaks called by different versions of MACS2.

The ATAC-seq peaks (regions defined by black block) called by MACS2 ver. 2.1.2 from the resting Treg cells at the *IL2RA* locus were more resolved and accurately represented the chromatin regions with high enrichment of sequencing reads compared with MACS2 ver. 2.1.0 (a). Low accessibility chromatin regions with low pileup signal (a; highlighted in yellow) were not identified as peaks in MACS2 ver. 2.1.2 but were in MACS2 ver. 2.1.0, which may indicate calling of false positive peaks. In addition, MACS2 ver. 2.1.0 also failed to identify ATAC-seq peaks at high coverage regions from the stimulated Treg cells (b; highlighted in red).

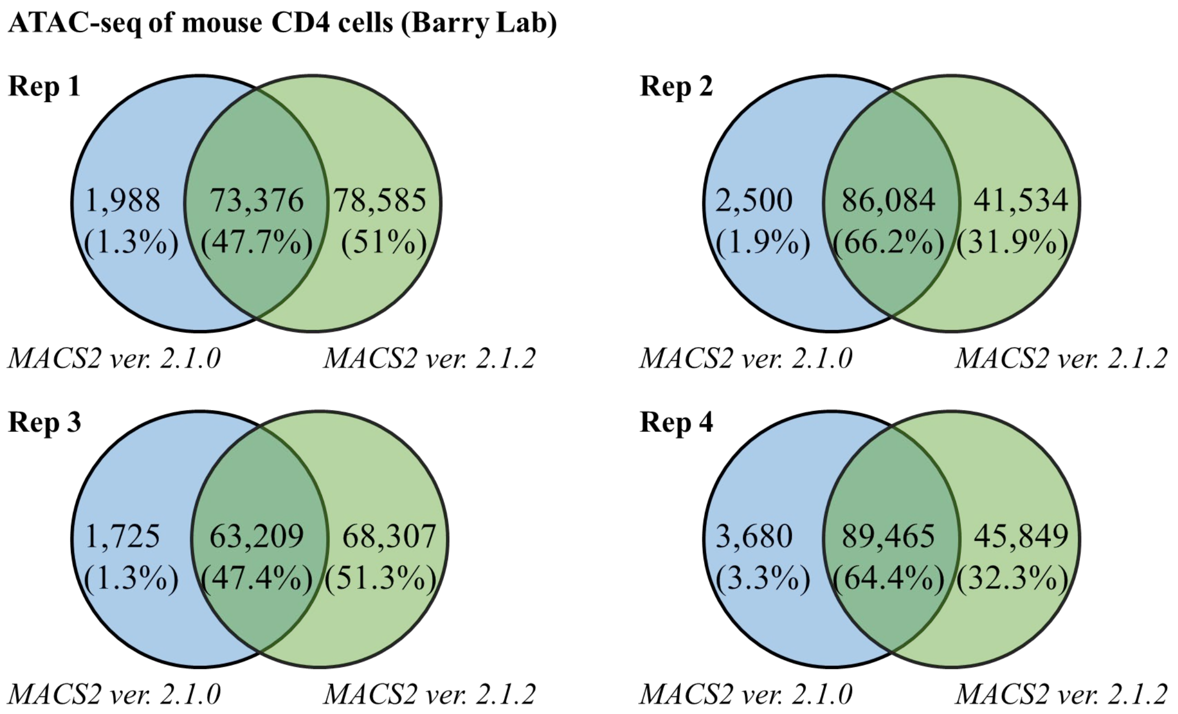


Figure 4-20 | Overlap of peaks called between from MACS2 ver. 210 and ver. 2.1.2. ATAC-seq datasets were generated from mouse CD4 T cells (n=4 mice) and peaks were called from individual replicate at a peak significance threshold of 0.05.

4.4.3 Differential accessibility analysis of Tconv vs Treg ATAC-seq

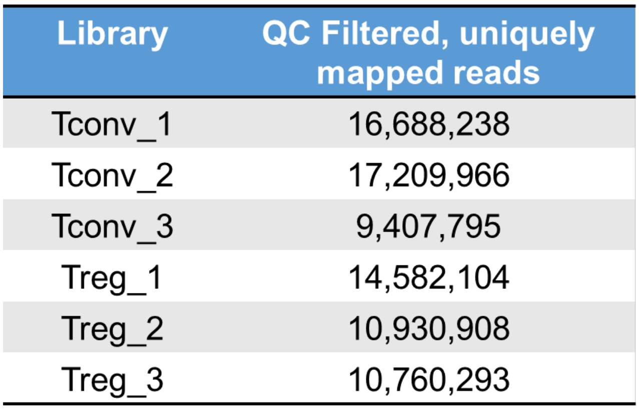
Apart from benchmarking peak calling approaches for ATAC-seq my work also aimed to establish a differential analysis pipeline for calling differentially accessible peaks between samples of interest, which is under-developed for ATAC-seq datasets. Using Tconv and Treg ATAC-seq datasets kindly supplied by Professor Joachim Schultze and Dr Marc Beyer (University of Bonn, Bonn, Germany), this section of work describes the differential accessibility analysis between Tconv and Treg cells purified from three healthy adult volunteers.

Differential accessibility analysis was performed using edgeR [38] exactTest as described[163], and peaks having a Benjamini-Hochberg (BH) false discovery rate (FDR) below 0.05 were considered to have significantly different accessibility between Treg and Tconv.

After post-alignment read processing to filter out unmapped, non-primary and low mapping quality reads I obtained 14,435,333 (± 3,561,381) reads on average per Tconv sample and 12,091,102 (± 1,762,781) reads per Treg sample (Table 4-6). The ATAC-seq signals for pooled reads (n=3) from Tconv and Treg libraries were enriched at transcriptional start sites (TSS) (Figure 4-21) and I saw similar enrichment of read signal in the ATAC-seq peaks across Tconv and Treg samples (Figure 4-22). The largest proportion of variation in the accessibility data was driven by cell type (Figure 4-23) and differential accessibility analysis shows that the top differentially regulated peaks were annotated to well-characterised Tconv/Treg gene signatures identified through RNA-seq, microarray and flow cytometry experiments [235-237] such as FOXP3, CTLA4, GATA3 and SATB1 (Figure 4-24). These signatures are well-established markers that reliably distinguish between T cell subsets and pathway and network enrichment analyses showed that those differentially accessible regions were strongly enriched for terms associated with T cell differentiation, cellular response to cytokine stimulus, T cell receptor signalling pathway, IL2-STAT5 signalling and response to chemokine (Figure 4-25, Figure 4-26, Figure 4-27, Figure 4-28, Figure 4-29, Figure 4-30, Figure 4-31, Figure 4-32). All of these indicate that the ATAC-seq differential analysis algorithm I adopted identified differentially accessible regions that correlate with well-defined Tconv/Treg transcriptional signal and corroborate observations from other studies.

Table 4-6. The total number of sequenced reads of Tconv and Treg ATAC-seq libraries after QC filtering.

ATAC-seq datasets were generated using human primary T cells purified from three adult donors by Beyer/Schultze Lab (The University of Bonn, Germany) and used in this section of analysis work (Section 4.4.3) with permission.



Library	QC Filtered, uniquely mapped reads
Tconv_1	16,688,238
Tconv_2	17,209,966
Tconv_3	9,407,795
Treg_1	14,582,104
Treg_2	10,930,908
Treg_3	10,760,293

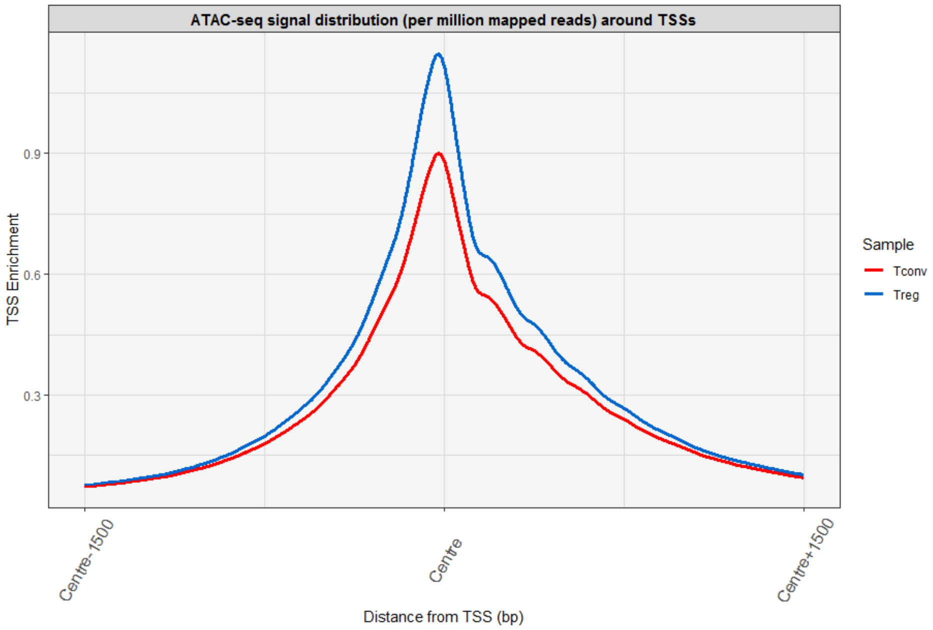


Figure 4-21 | Enrichment plot of Tn5 insertions at ± 1.5 kb transcription start sites (TSS enrichment) for T cell ATAC-seq datasets.

The enrichment of the ATAC-seq fragments at transcription start sites (TSS) is a valuable indicator for the signal-to-noise ratio in ATAC-seq data. The data shown are representative of reads pooled from three replicates.

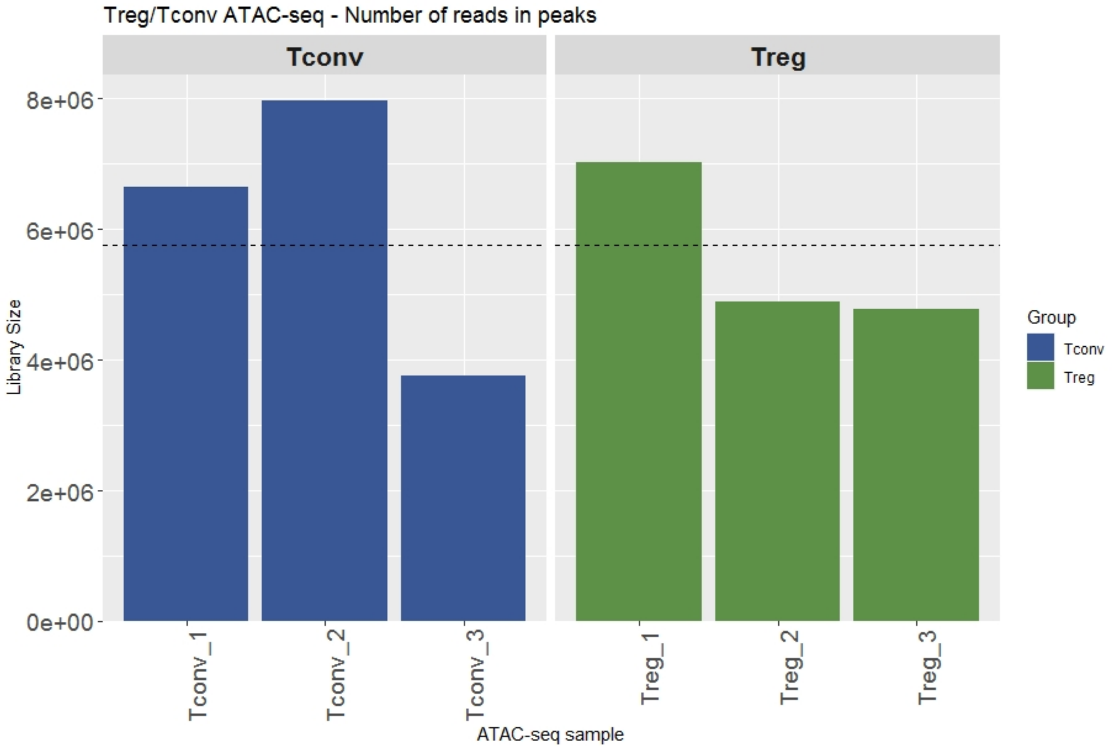


Figure 4-22 | Number of reads in peaks for T cell ATAC-seq datasets.
Count matrix was prepared by calculating the number of reads mapped to each ATAC-seq peak and used as input in differential accessibility analysis to identify differential peaks between Tconv and Treg using edgeR [38].

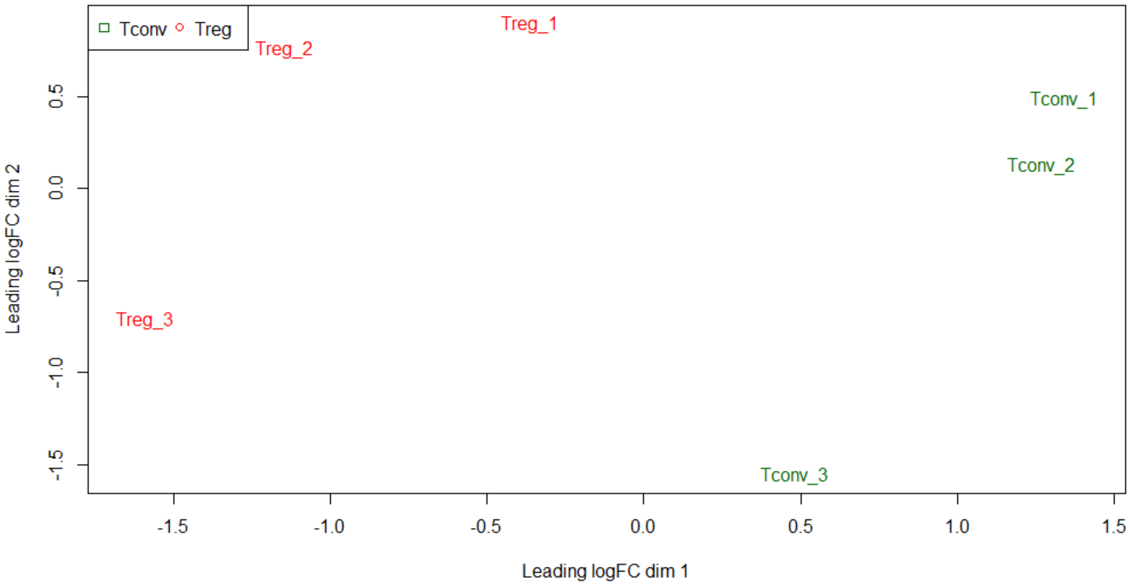


Figure 4-23 | Multidimensional scaling (MDS) plot of log-CPM values over dimensions 1 and 2 for Tconv and Treg ATAC-seq samples.

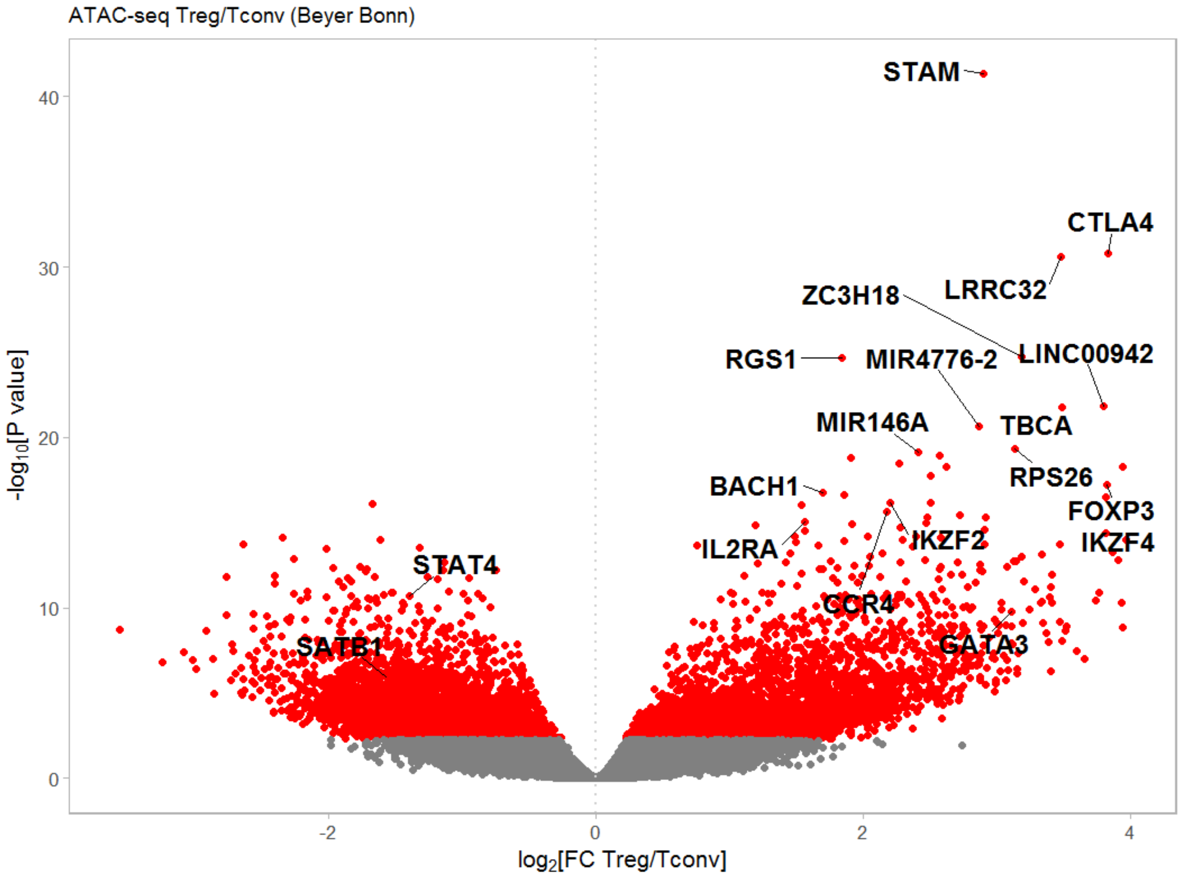


Figure 4-24 | Volcano plot showing differential ATAC-seq analysis between Treg and Tconv cells.

Regions having significantly differential accessibility (Benjamin-Hochberg FDR < 0.05) are coloured red. Differential peaks were annotated to the nearest TSS in linear genomic distance. Top 10 and selected immune-relevant differential regions are annotated with gene symbols.

Table 4-7. Enrichment analyses performed in this thesis.

Pathway analysis	Description
GO (Gene Ontology)	Annotates differential genes using ontology classification.
KEGG	Maps differential genes to molecular interaction networks.
GSEA (Gene Set Enrichment Analysis)	Identifies classes of genes that are over-represented in the whole dataset.

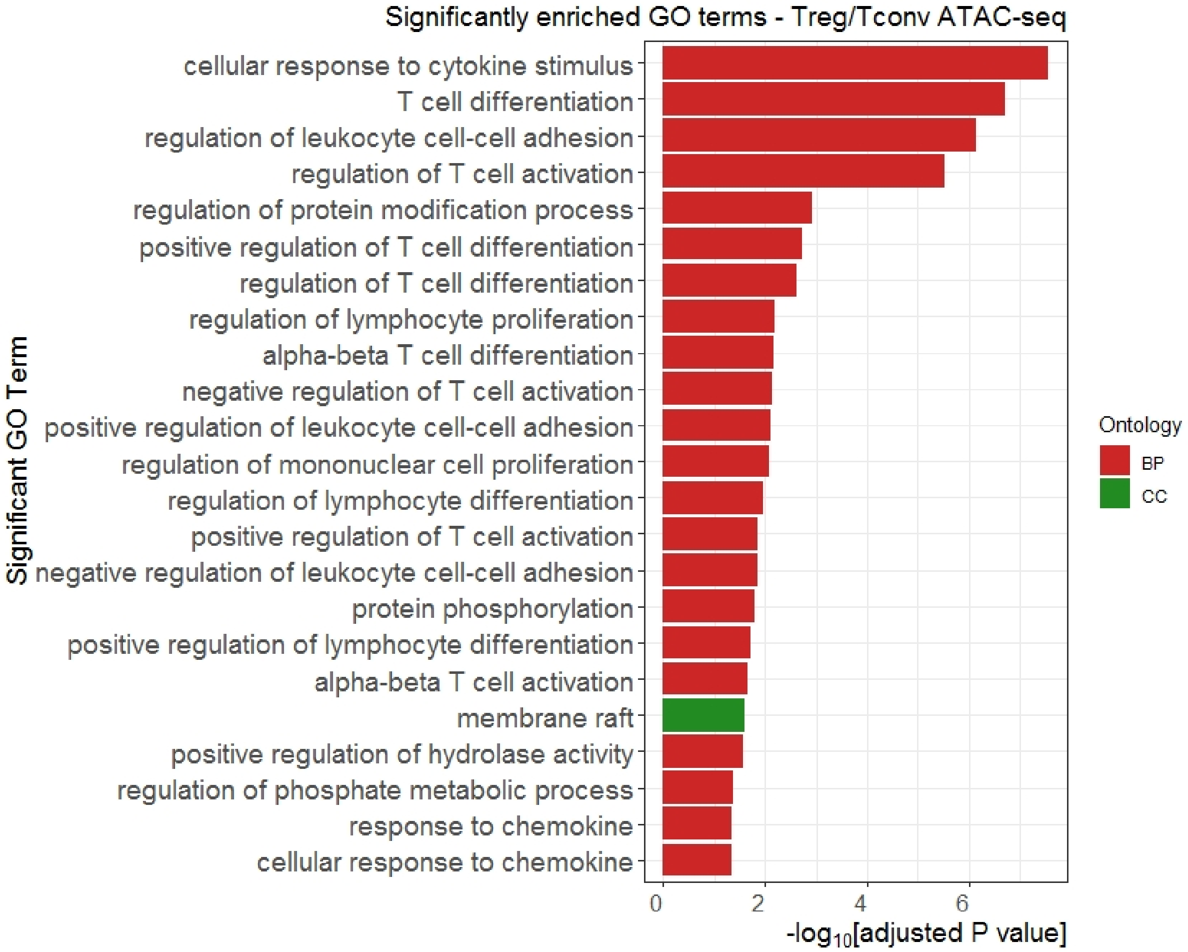


Figure 4-25 | Gene Ontology enrichment analysis showing significantly enriched GO terms associated with differential accessible peaks between Treg- and Tconv- ATAC-seq, using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

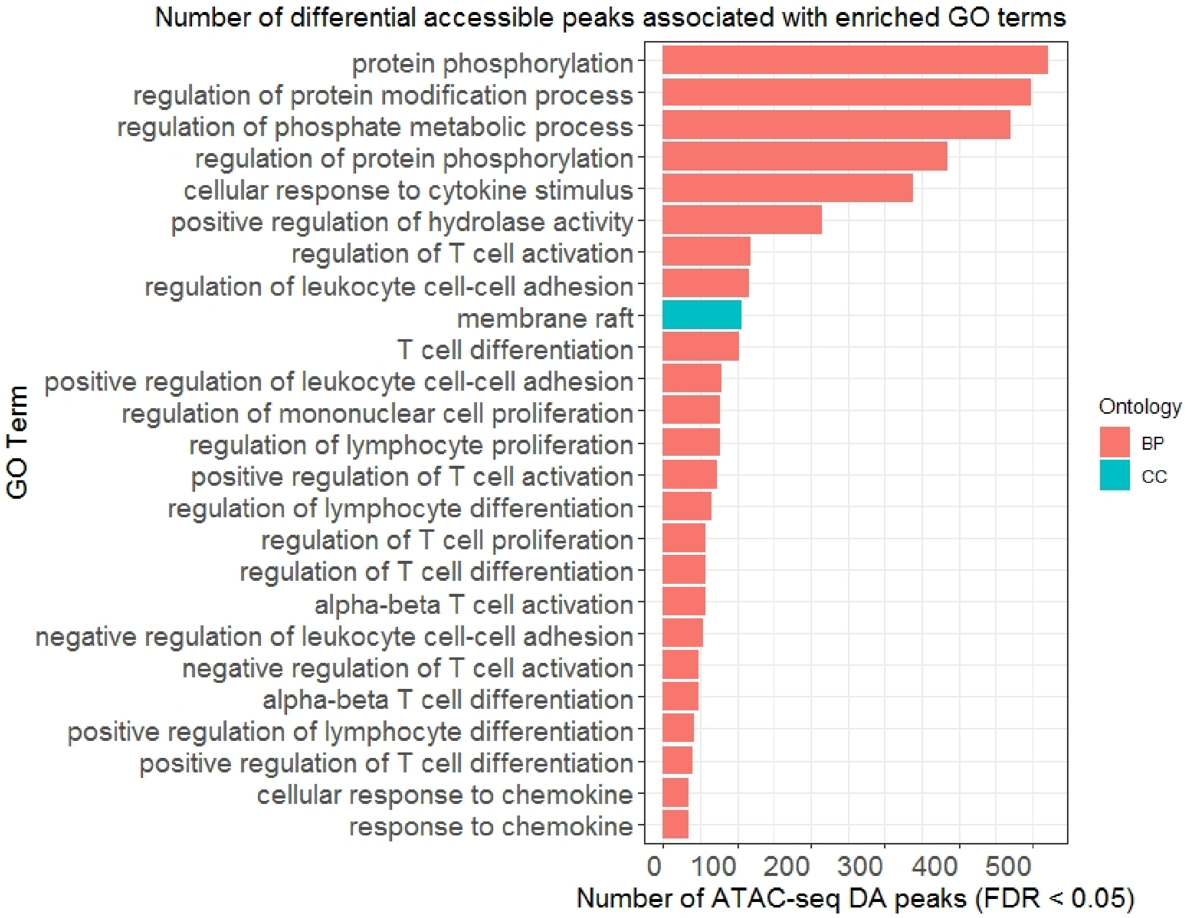


Figure 4-26 | Number of differential accessible peaks associated with each enriched GO term in Treg- versus Tconv- ATAC-seq.

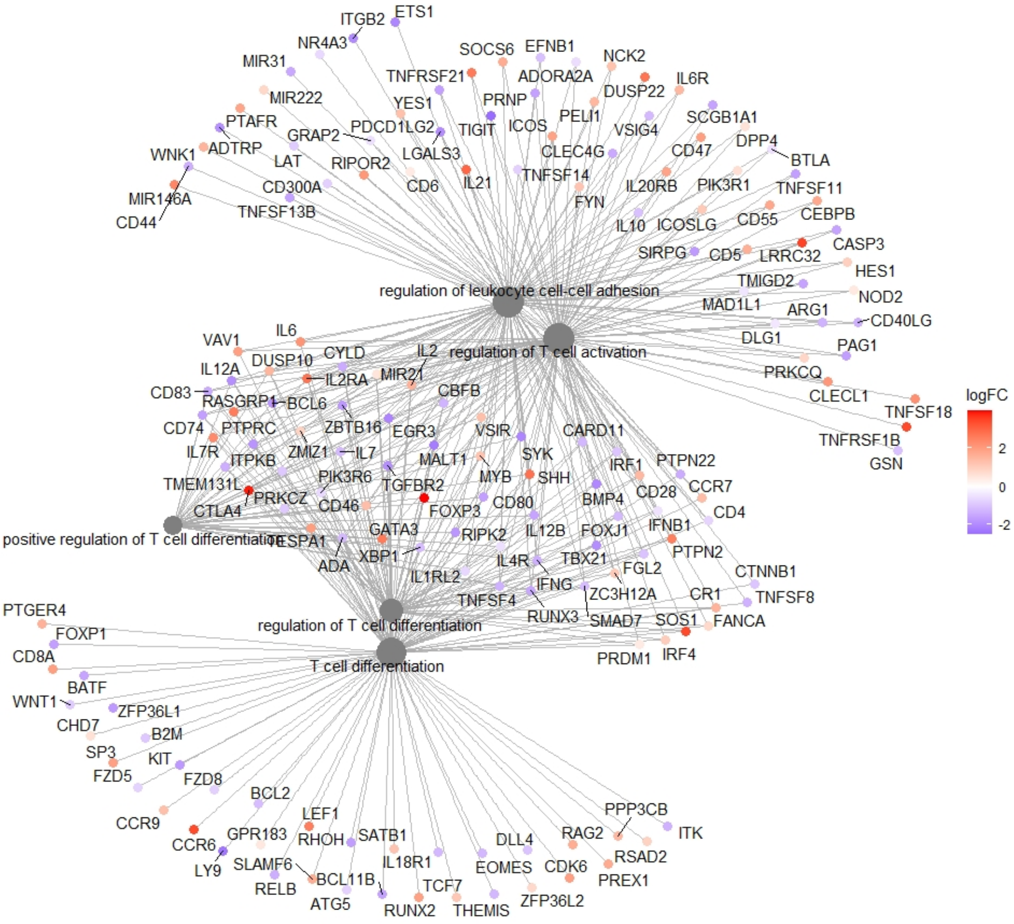


Figure 4-27 | Enrichment plot showing the most significant GO terms and related genes from the Biological Process Ontology for Treg- versus Tconv- ATAC-seq.

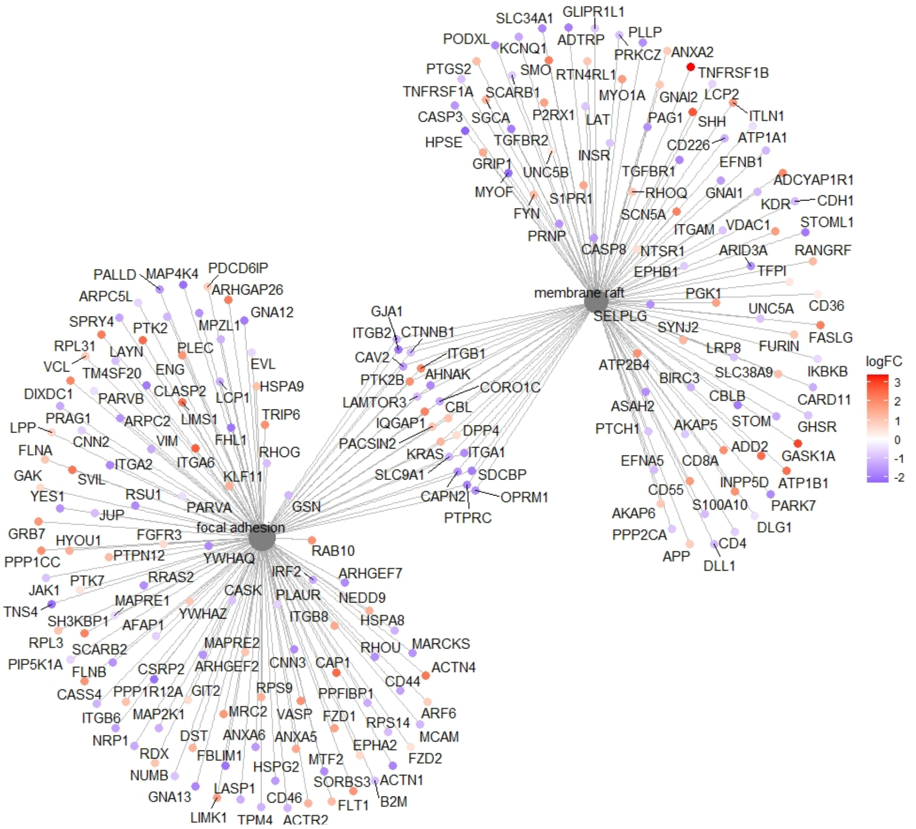


Figure 4-28 | Enrichment plot showing the most significant GO terms and related genes from the Cellular Component Ontology for Treg- versus Tconv- ATAC-seq.

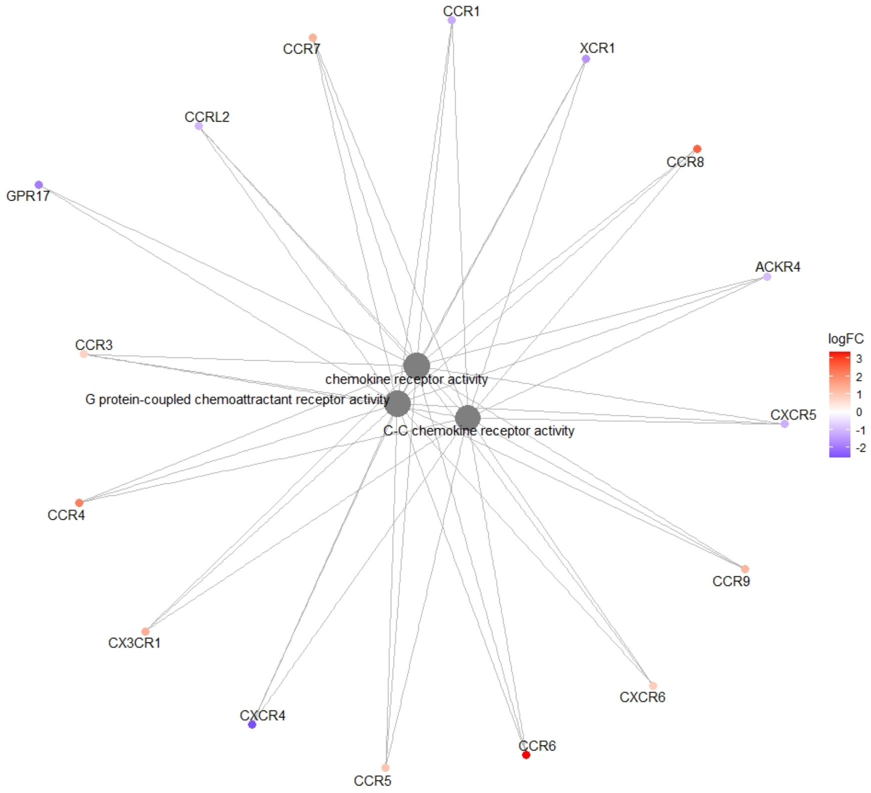

CCR1
CCR7
XCR1
CCRL2
CCR8
GPR17
ACKR4
CCR3
chemokine receptor activity
G protein-coupled chemoattractant receptor activity
C-C chemokine receptor activity
CXCR5
logFC
3
2
1
0
-1
-2
CCR4
CCR9
CX3CR1
CXCR6
CXCR4
CCR6
CCR5

Figure 4-29 | Enrichment plot showing the most significant GO terms and related genes from the Molecular Function Ontology for Treg- versus Tconv- ATAC-seq.

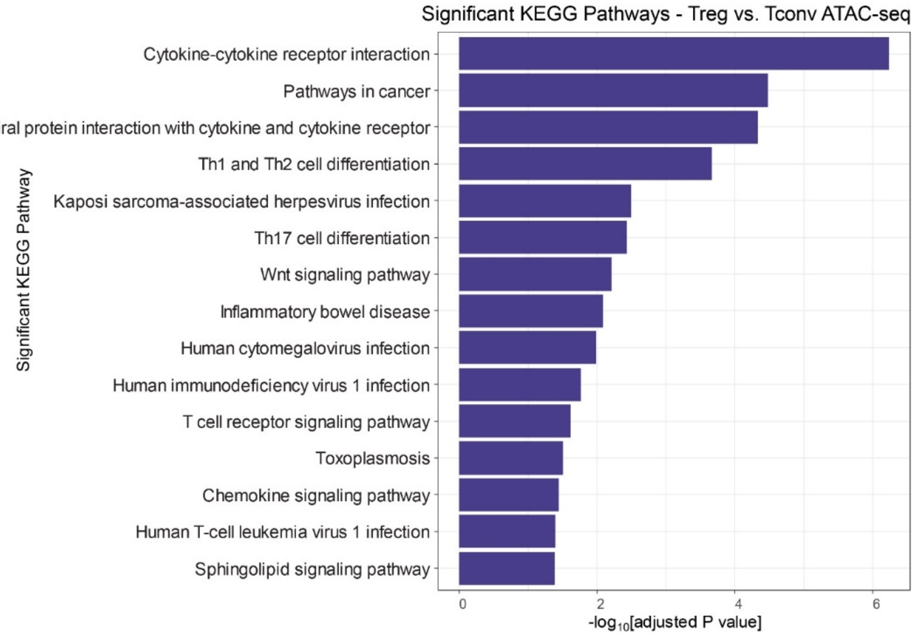


Figure 4-30 | KEGG pathway analysis showing significantly enriched pathways associated with differential accessible peaks between Treg- and Tconv- ATAC-seq.

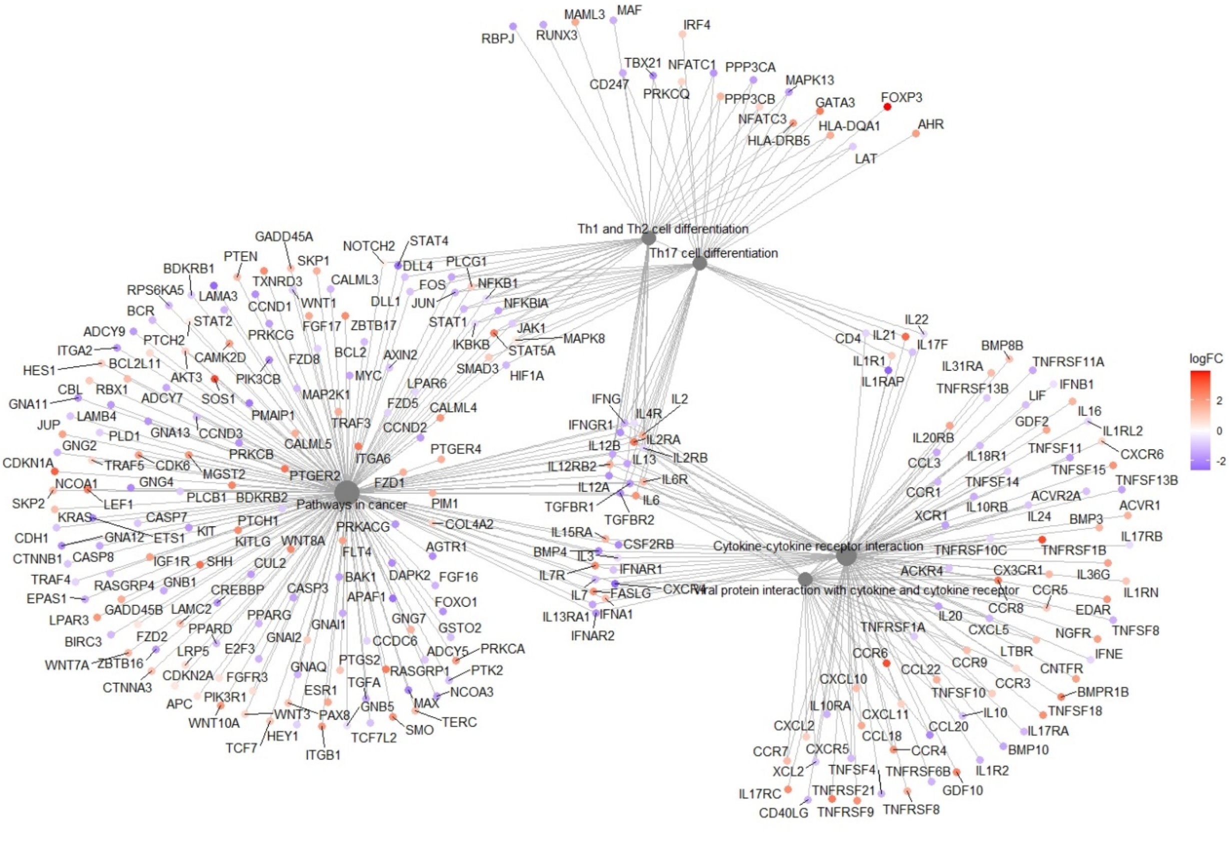


Figure 4-31 | Enrichment plot showing the most significant KEGG pathways and associated genes for Treg- versus Tconv- ATAC-seq.

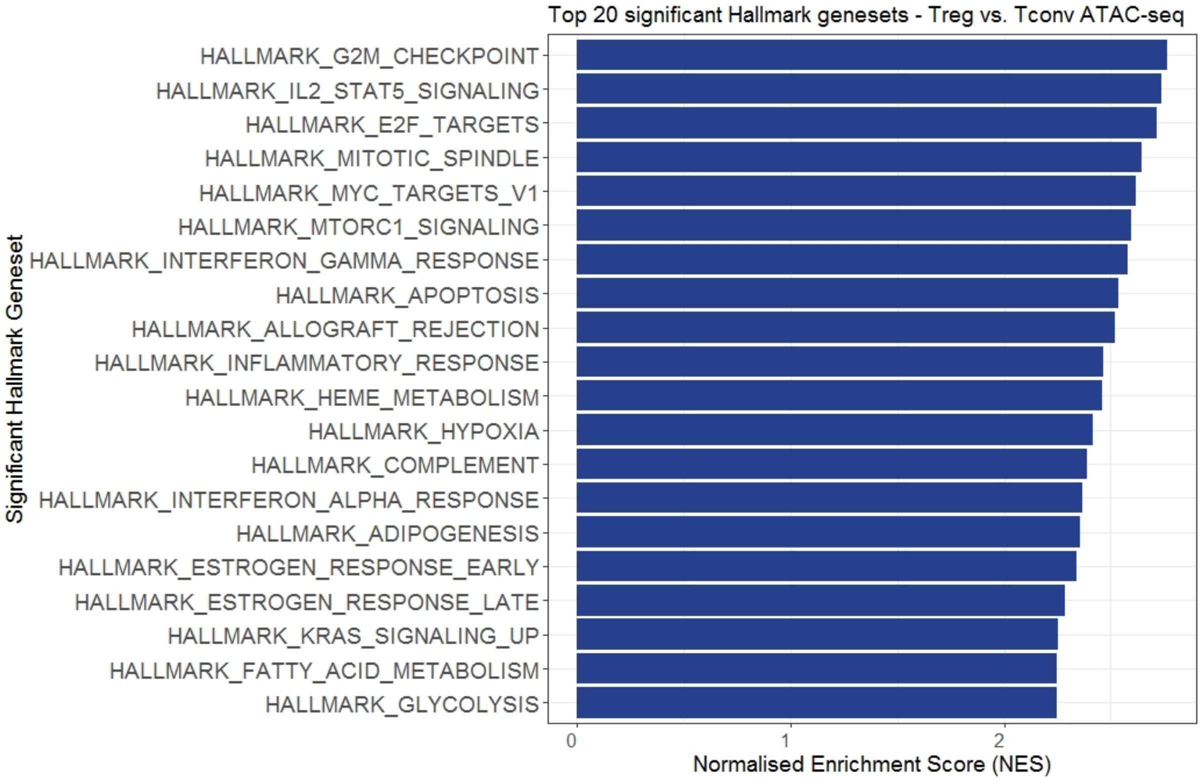


Figure 4-32 | Gene set enrichment analysis (GSEA) showing top 20 significantly enriched (FDR < 0.05) Hallmark gene sets in Treg- vs Tconv ATAC-seq.
Results are representative of accessibility profiles from three healthy subjects.

4.4.4 Comparison of ATAC-seq normalization methods in differential accessibility analysis

It was reported in Reske, Wilson [230] work that different normalisation tools and algorithms used in calculating differential accessibility in ATAC-seq data can significantly alter the biological interpretation or output evaluation. In an effort to benchmark normalisation methods for ATAC-seq differential analysis, using Tconv and Treg ATAC-seq datasets, I cross-compared and evaluated the impact of different normalisation methods (Table 4-2) on the outcome of differential analysis between Tconv and Treg ATAC-seq. Tconv and Treg ATAC-seq datasets were generated from biobanked PBMCs obtained from 12 healthy subjects as part of the ATAC-seq profiling described in CHAPTER 5:. As shown in Figure 4-33, the largest proportion of variation that was driving the separation of all sequenced ATAC-seq samples was cell type. MA plots were used to identify global accessibility patterns of differential analysis measurements between Tconv and Treg ATAC-seq following different methods of data normalisation, where each query peak region is quantified by the difference in read signal between the two groups as the *y-axis* and signal abundance as the *x-axis*. MA plots for differential peaks showed that different normalisation approaches result in varying global accessibility distributions (Figure 4-34). An upward or downward shift in the accessibility distribution could either indicate a global effect or technical bias [230]. The MA distribution of TMM normalisation approach based on raw peak counts (Method 1) (Figure 4-34; a) was similar to the Loess normalisation method (Method 3) (Figure 4-34; c), whereas TMM normalisation based on binned peak counts (Method 2) (Figure 4-34; b) which counts the number of reads overlapping a genomic window at spaced positions across the genome, resulted in more significant differential regions that demonstrated decreased accessibility than increased accessibility (in Treg). The distribution of differential analysis measurements appeared asymmetrical and was shifted downwards in the TMM normalized binned windows (Figure

4-34; b), which may or may not be technical in nature and the bias was corrected with TMM normalisation using raw counts (Figure 4-34; a) or Loess normalization (Figure 4-34; c).

The top 40 differentially accessible (DA) peaks identified between the Tconv and Treg ATAC-seq captured most of the well-characterised Tconv/Treg gene signatures such as FOXP3, TIGIT and CTLA4 following normalisation using Method 1 and Method 3, but not Method 2 (Figure 4-35, Figure 4-36, Figure 4-37, Figure 4-38, Figure 4-39 and Figure 4-40). An overlap of 82% was observed between DA regions identified from Method 1 and 3 (Figure 4-41). Genomic annotation of all the differentially accessible peaks identified following the 3 normalisation methods showed that the feature distribution was similar between Method 1 and Method 3, whereas differential peaks identified though Method 2 normalisation showed higher enrichment of promoter regions but lower enrichment of intergenic regions in comparison with Method 1 and Method 3 (Figure 4-42, Figure 4-43, Figure 4-44 and Figure 4-45). Nonetheless, the pathway and network enrichment analyses showed that the differentially accessible regions identified through all 3 normalisation methods were strongly enriched for terms associated with T cell differentiation, cellular response to cytokine stimulus, regulation of T cell activation T cell receptor signalling pathway or IL2-STAT5 signalling (Figure 4-46, Figure 4-47, Figure 4-48, Figure 4-49, Figure 4-50, Figure 4-51, Figure 4-52, Figure 4-53, Figure 4-54, Figure 4-55, Figure 4-56, Figure 4-57, Figure 4-58, Figure 4-59, Figure 4-60).

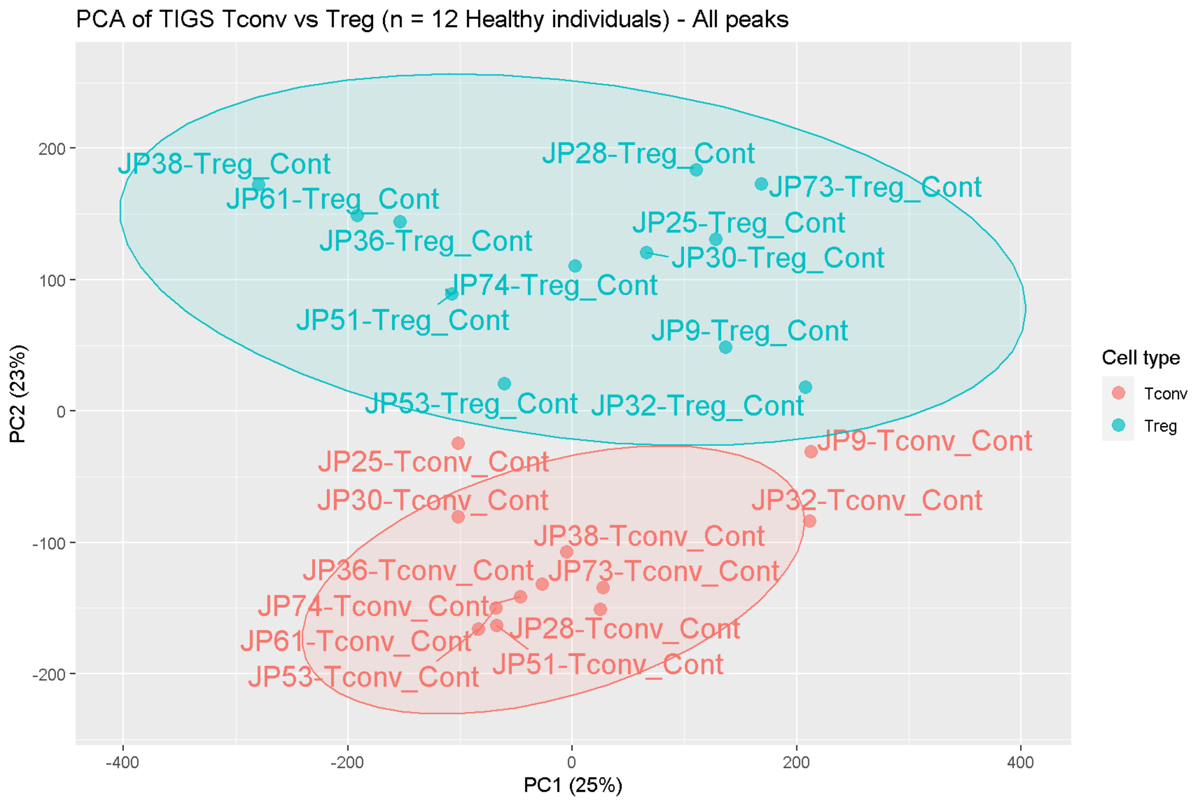
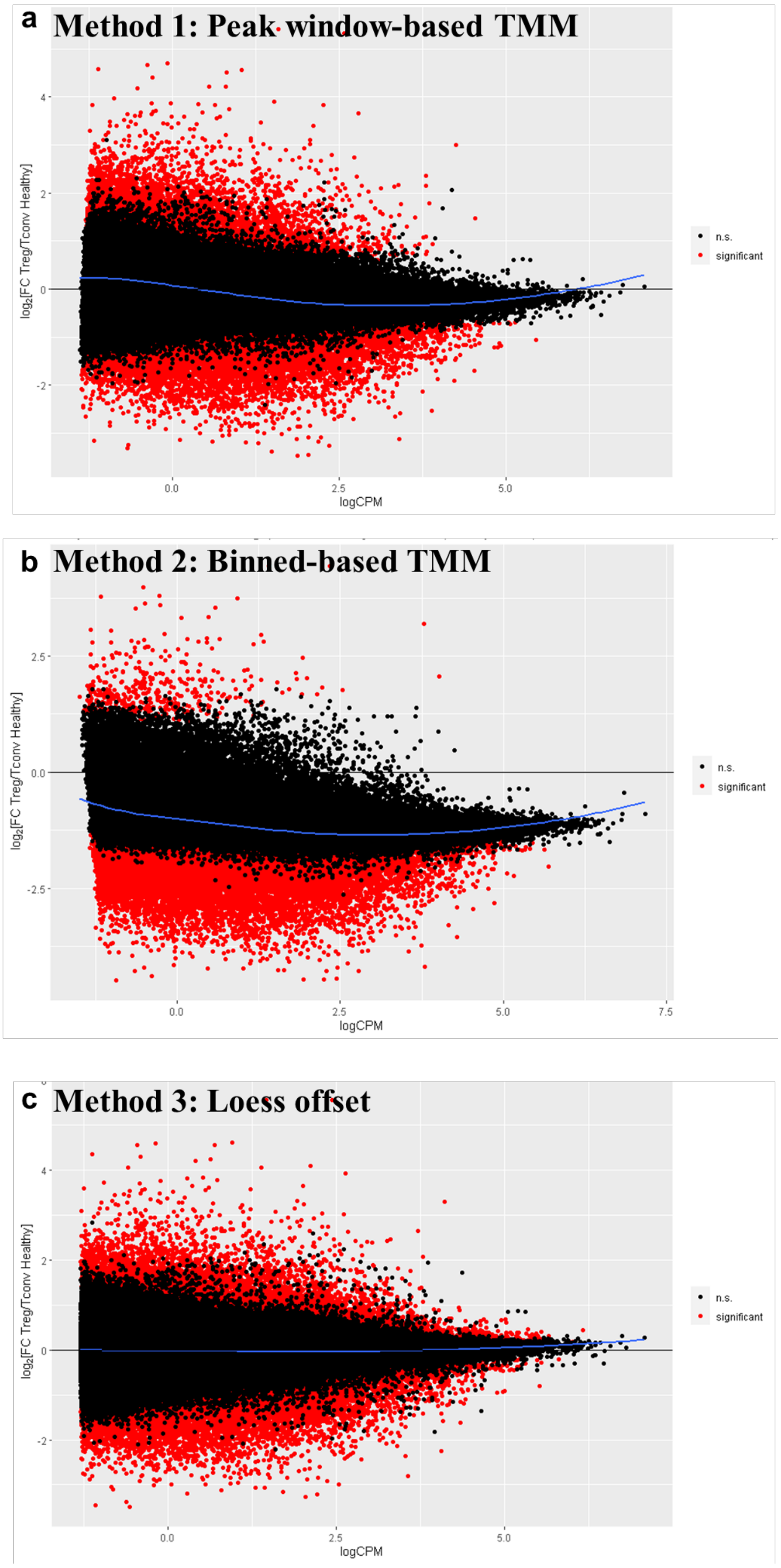


Figure 4-33 | Principal component analysis (PCA) plot of count values in peaks over first two components for Treg- vs Tconv- ATAC-seq.

ATAC-seq profiles were generated from stimulated Treg and Tconv cells obtained from 12 healthy subjects. The ATAC-seq samples are separated by cell type over the first and second principal component.



a Method 1: Peak window-based TMM

b Method 2: Binned-based TMM

c Method 3: Loess offset

Figure 4-34 | MA plots depicting global differential accessibility distributions from the same ATAC-seq dataset (Treg- vs Tconv- ATAC-seq; n=12) analysed by 3 different normalisation approaches.

X-axis of MA plot represents average peak signal abundance at that region, while Y-axis corresponds to the log2 difference in peak signal between Treg and Tconv. Black dots represent non-significant regions, and red dots represent significant (FDR < 0.05) differential regions. Blue lines are loess fits to each distribution. Refer to Table 4-2 for detailed explanation of each normalisation strategy.

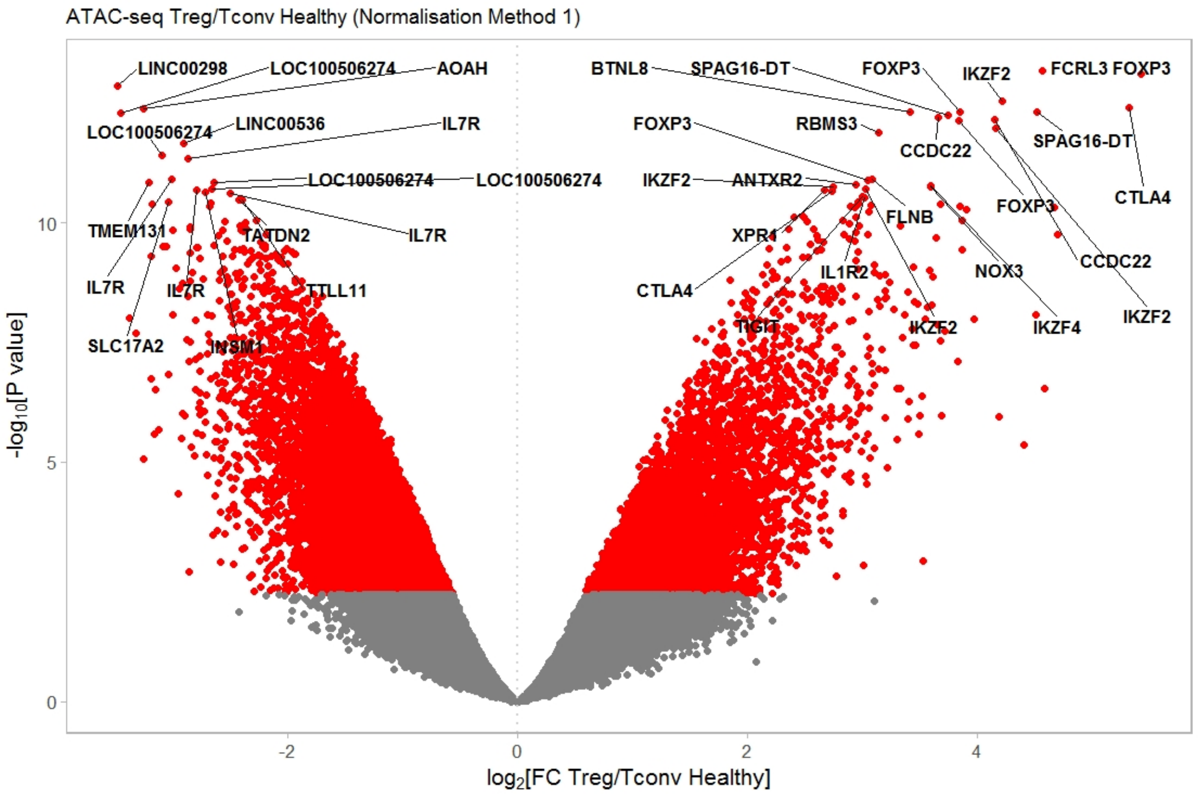


Figure 4-35 | Volcano plot showing differential ATAC-seq analysis between Treg and Tconv cells using normalisation Method 1 (see Table 4-2).

Regions having significantly differential accessibility (Benjamin-Hochberg FDR < 0.05) are coloured red. Differential peaks were annotated to the nearest TSS in linear genomic distance. Top 40 differential peaks are annotated with gene symbols. Results shown are representative of 12 healthy subjects.

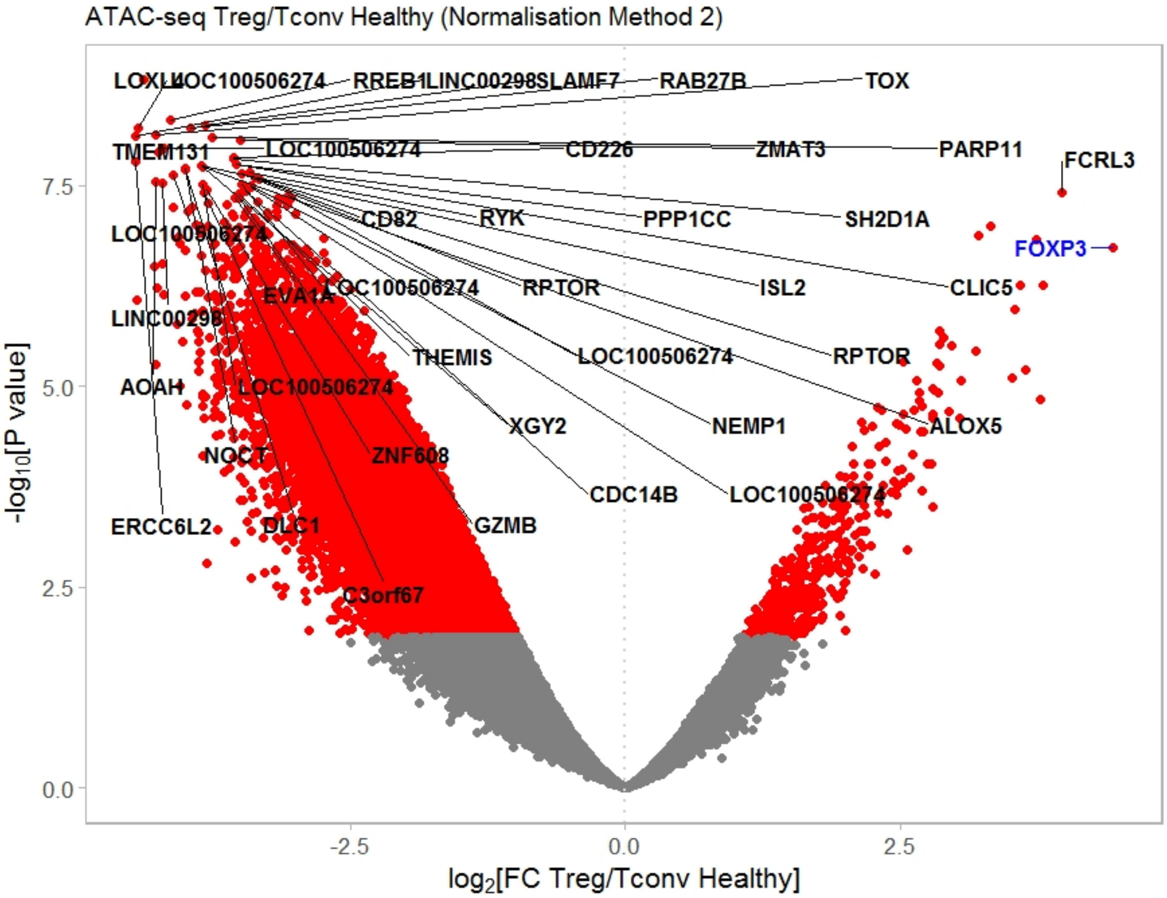


Figure 4-36 | Volcano plot showing differential ATAC-seq analysis between Treg and Tconv cells using normalisation Method 2 (see Table 4-2).

Regions having significantly differential accessibility (Benjamin-Hochberg FDR < 0.05) are coloured red. Differential peaks were annotated to the nearest TSS in linear genomic distance. Top 40 differential peaks are annotated with gene symbols (black font). Peak at the *FOXP3* locus (blue font) was not identified as top 40 differential accessible region between Treg and Tconv cells using this normalisation method. Results shown are representative of 12 healthy subjects.

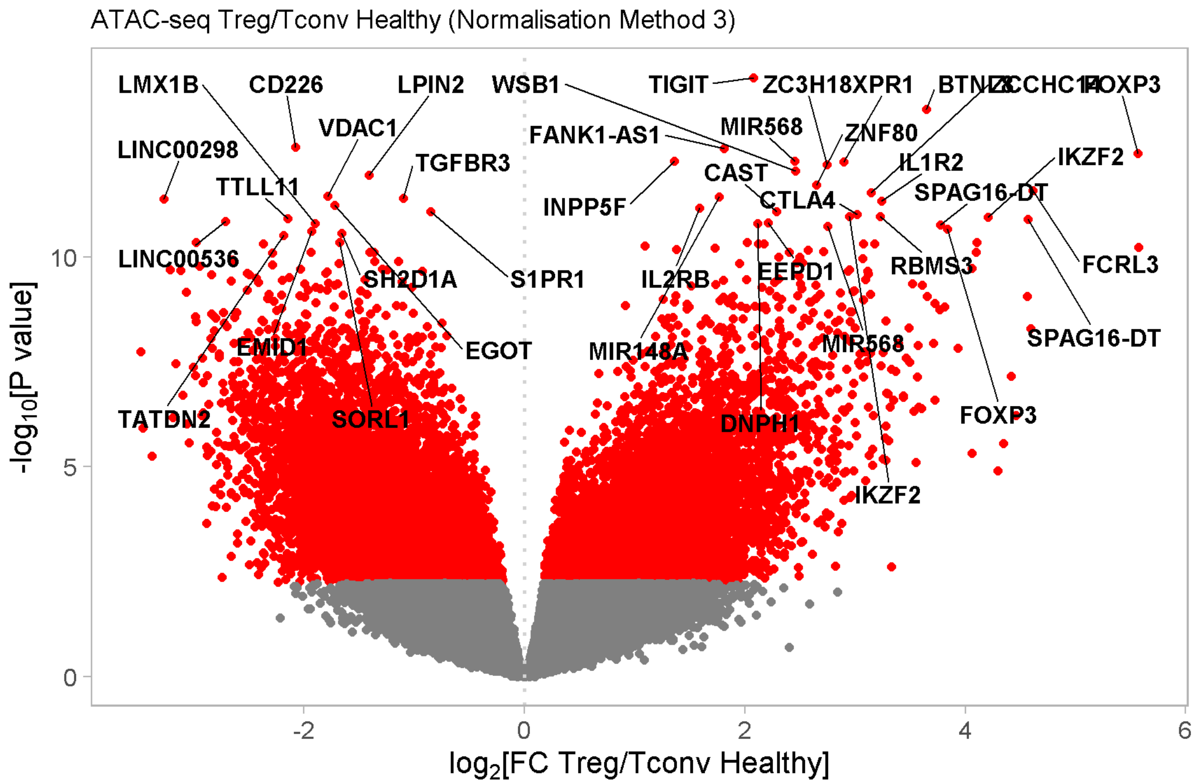


Figure 4-37 | Volcano plot showing differential ATAC-seq analysis between Treg and Tconv cells using normalisation Method 3 (see Table 4-2).
Regions having significantly differential accessibility (Benjamin-Hochberg FDR < 0.05) are coloured red. Differential peaks were annotated to the nearest TSS in linear genomic distance. Top 40 differential peaks are annotated with gene symbols. Results shown are representative of 12 healthy subjects.

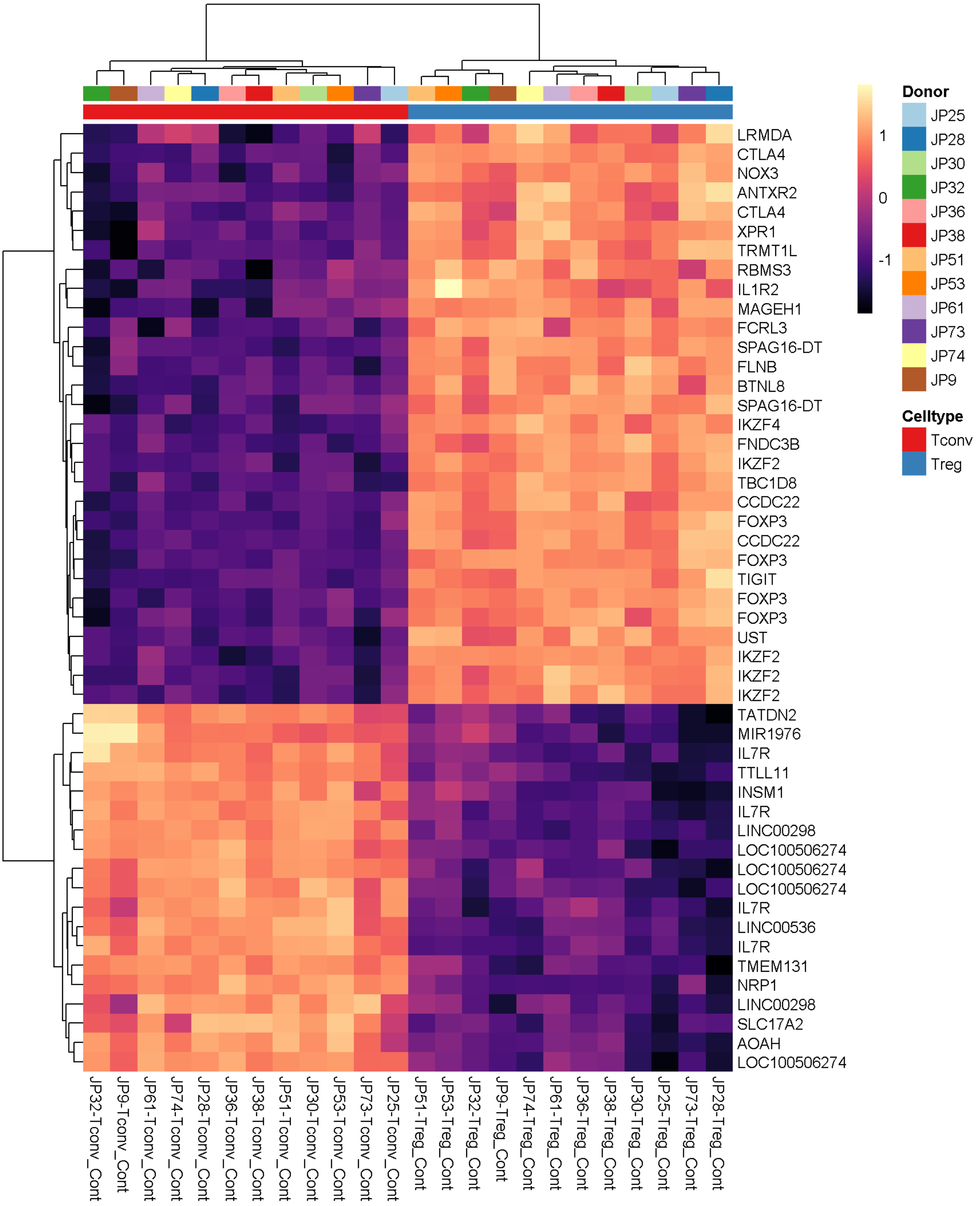


Figure 4-38 | Heatmap depicting top 50 differential ATAC-seq peaks between Treg and Tconv cells using normalisation Method 1 (see Table 4-2).
Differential peaks were annotated to the nearest TSS in linear genomic distance. Results shown are representative of 12 healthy subjects.

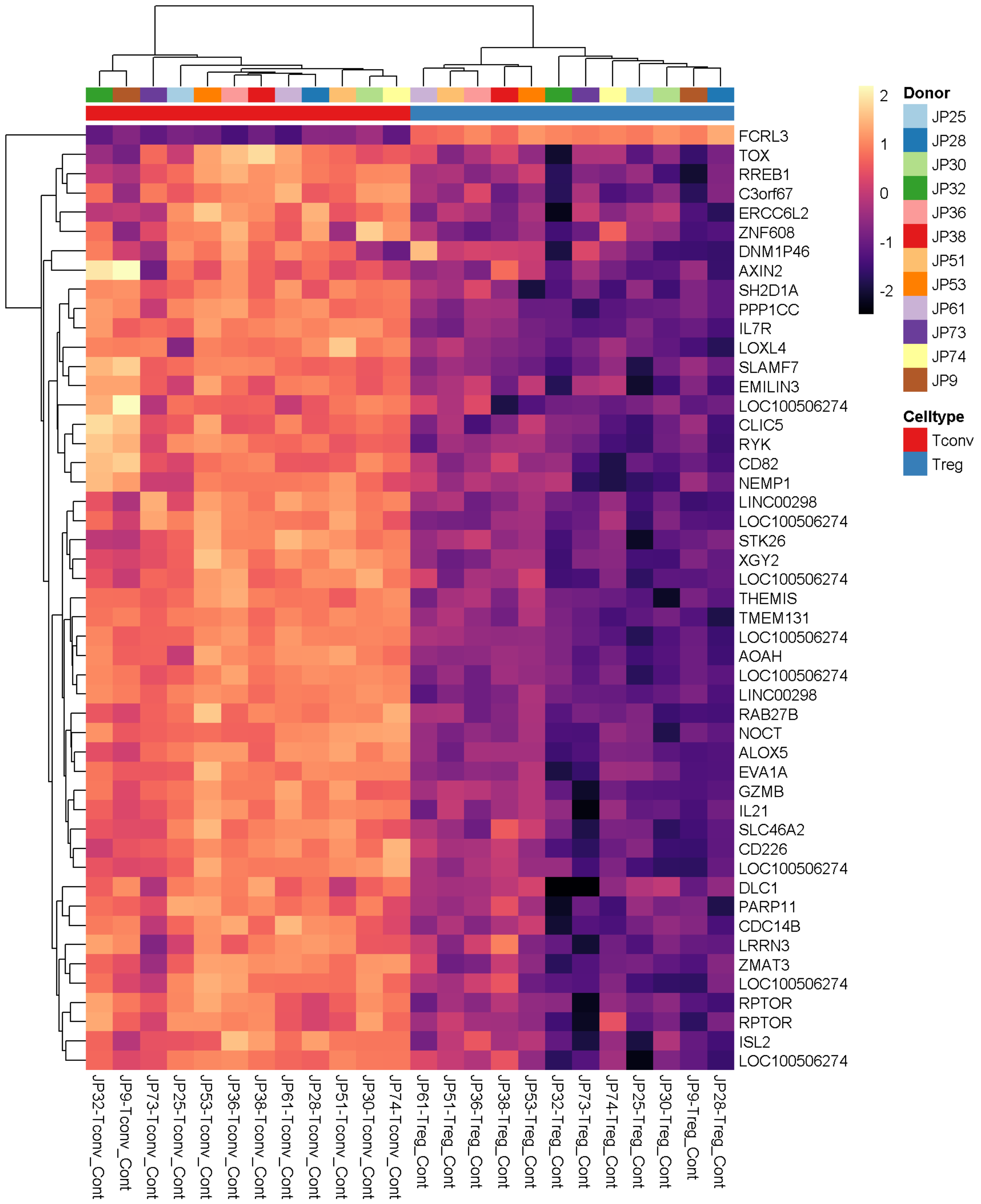


Figure 4-39 | Heatmap depicting top 50 differential ATAC-seq peaks between Treg and Tconv cells using normalisation Method 2 (see Table 4-2).
Differential peaks were annotated to the nearest TSS in linear genomic distance. Results shown are representative of 12 healthy subjects.

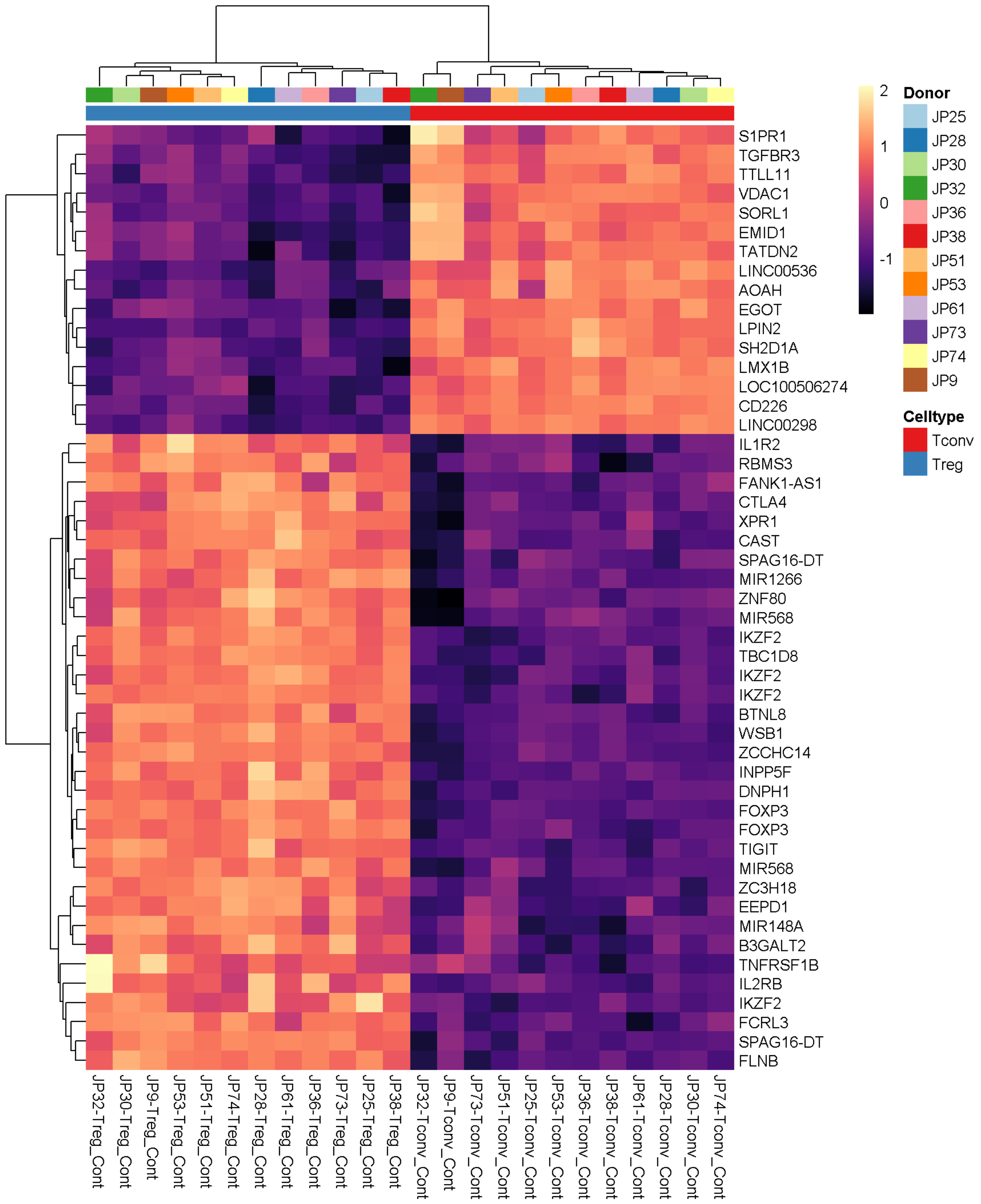


Figure 4-40 | Heatmap depicting top 50 differential ATAC-seq peaks between Treg and Tconv cells using normalisation Method 3 (see Table 4-2).

Differential peaks were annotated to the nearest TSS in linear genomic distance. Results shown are representative of 12 healthy subjects.

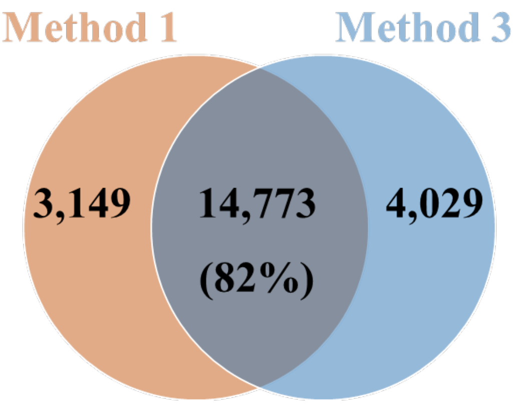


Figure 4-41 | Overlap of significantly differential ATAC-seq peaks between normalization method 1 and 3.

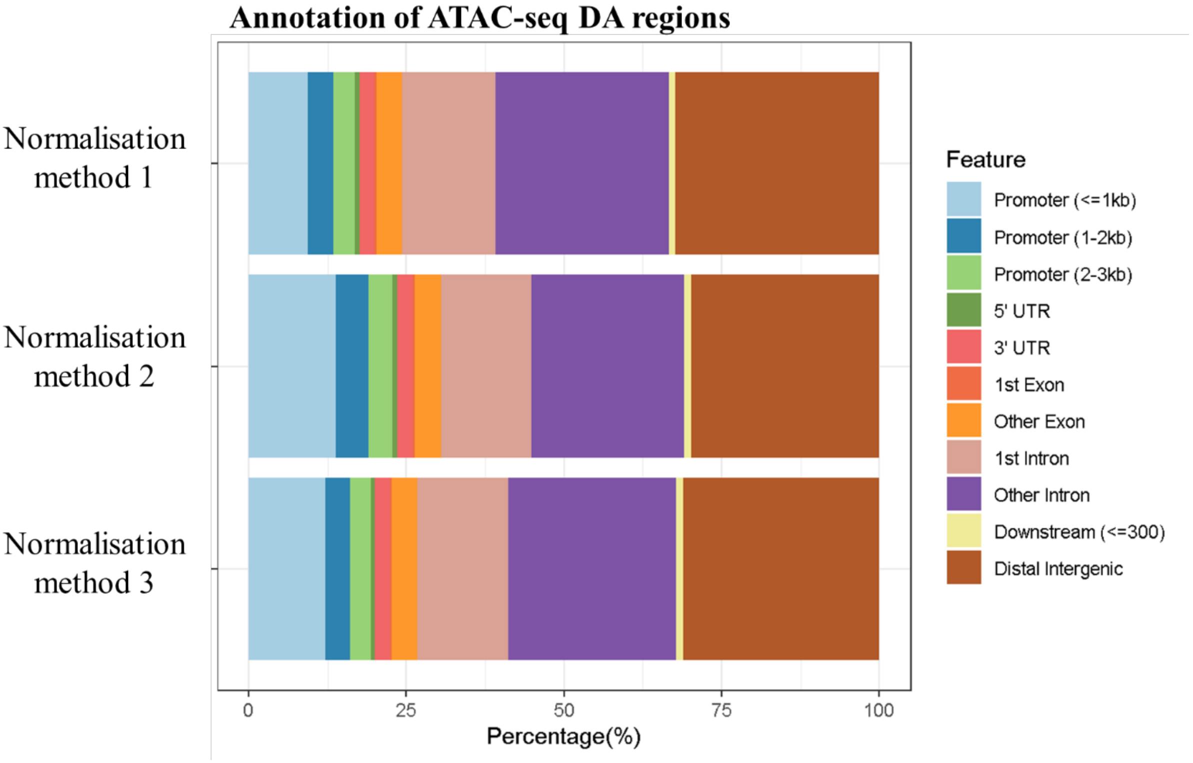


Figure 4-42 | Genomic annotation of significantly differential accessible peaks between Treg- and Tconv- ATAC-seq produced by 3 different normalisation approaches (see Table 4-2). ***DA, differential accessible.***

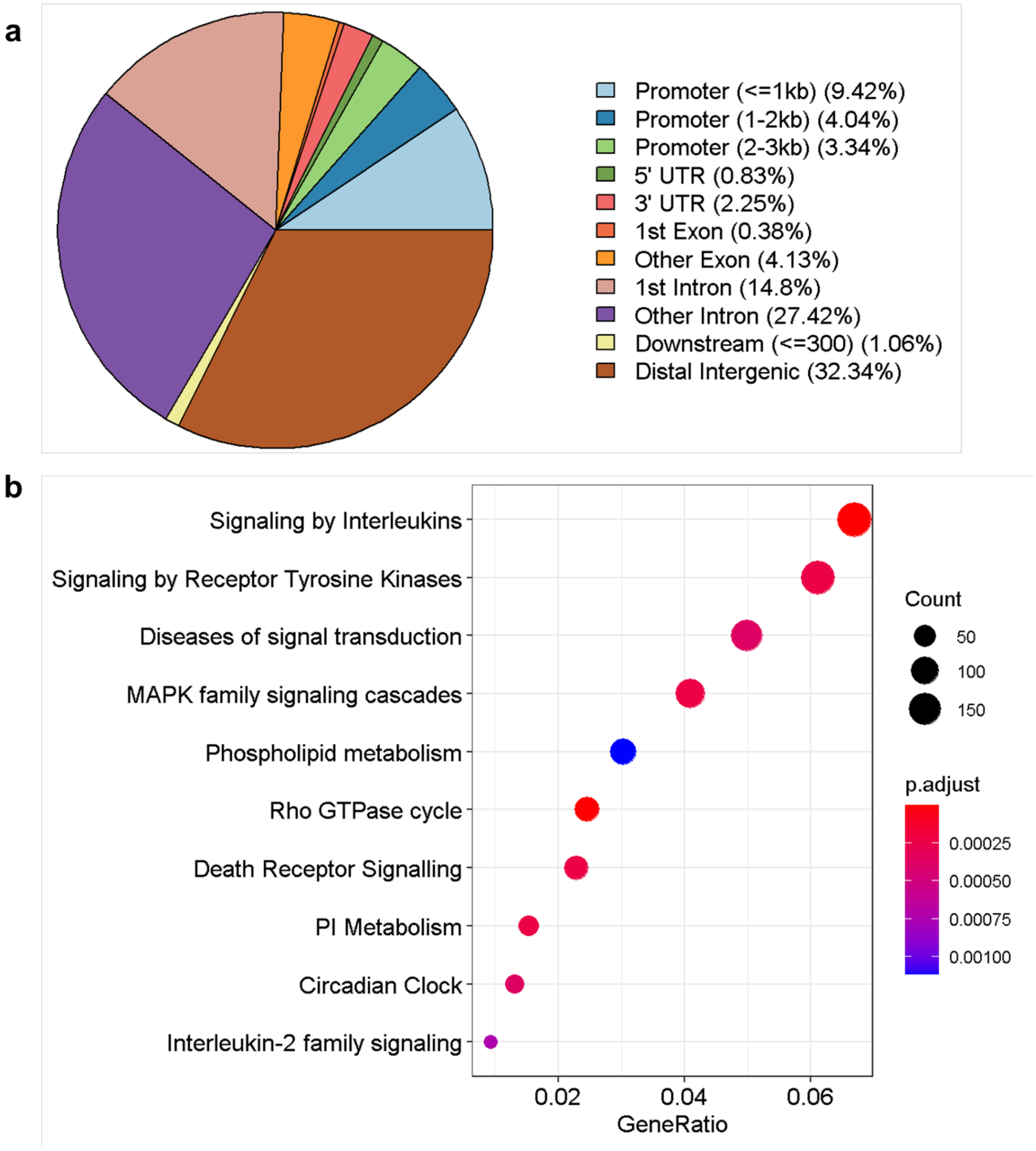


Figure 4-43 | Genomic annotation (a) and pathway enrichment analysis (b) of significantly differential accessible peaks between Treg- and Tconv- ATAC-seq calculated using normalisation Method 1 (see Table 4-2).

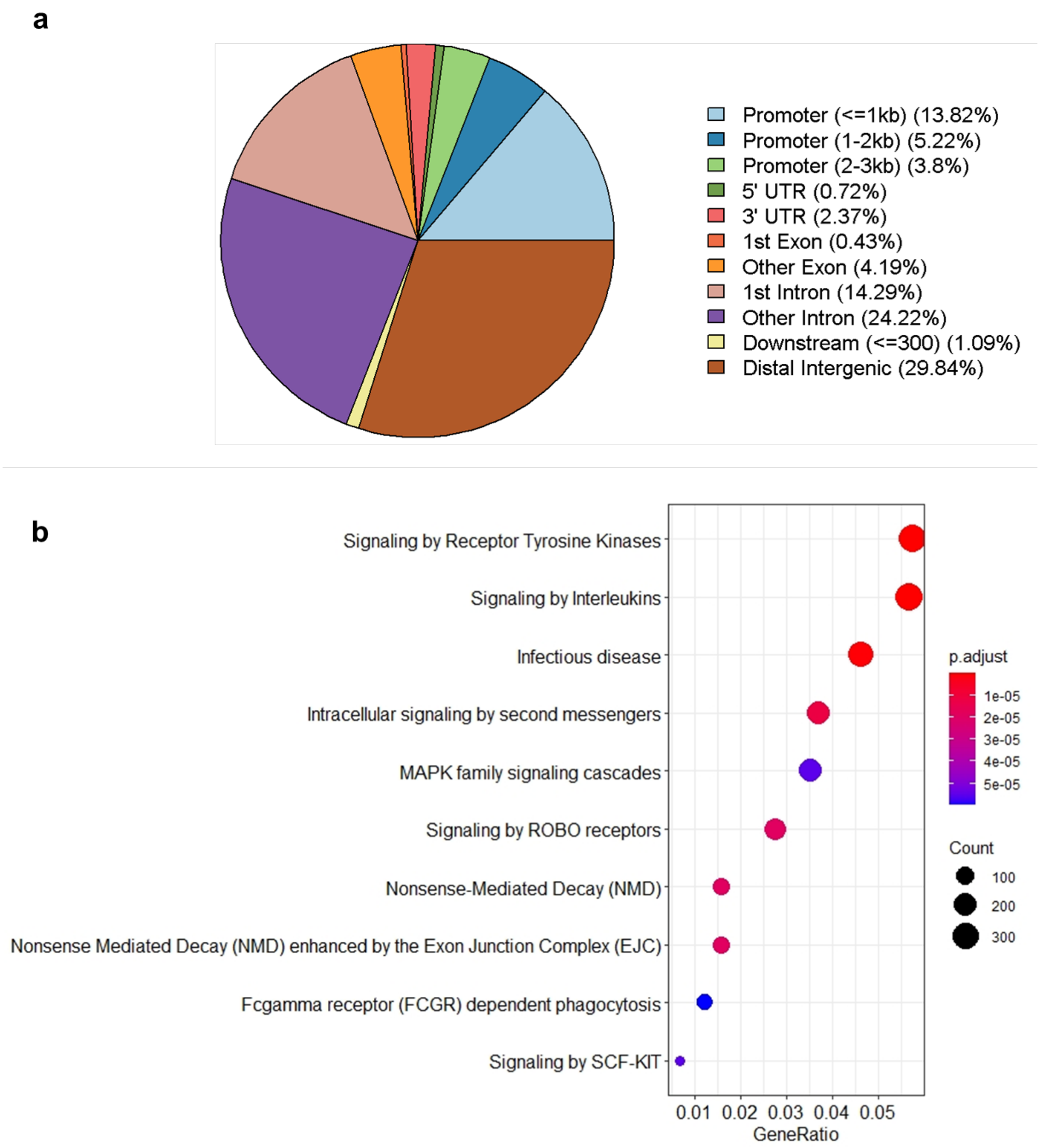


Figure 4-44 | Genomic annotation (a) and pathway enrichment analysis (b) of significantly differential accessible peaks between Treg- and Tconv- ATAC-seq calculated using normalisation Method 2 (see Table 4-2).

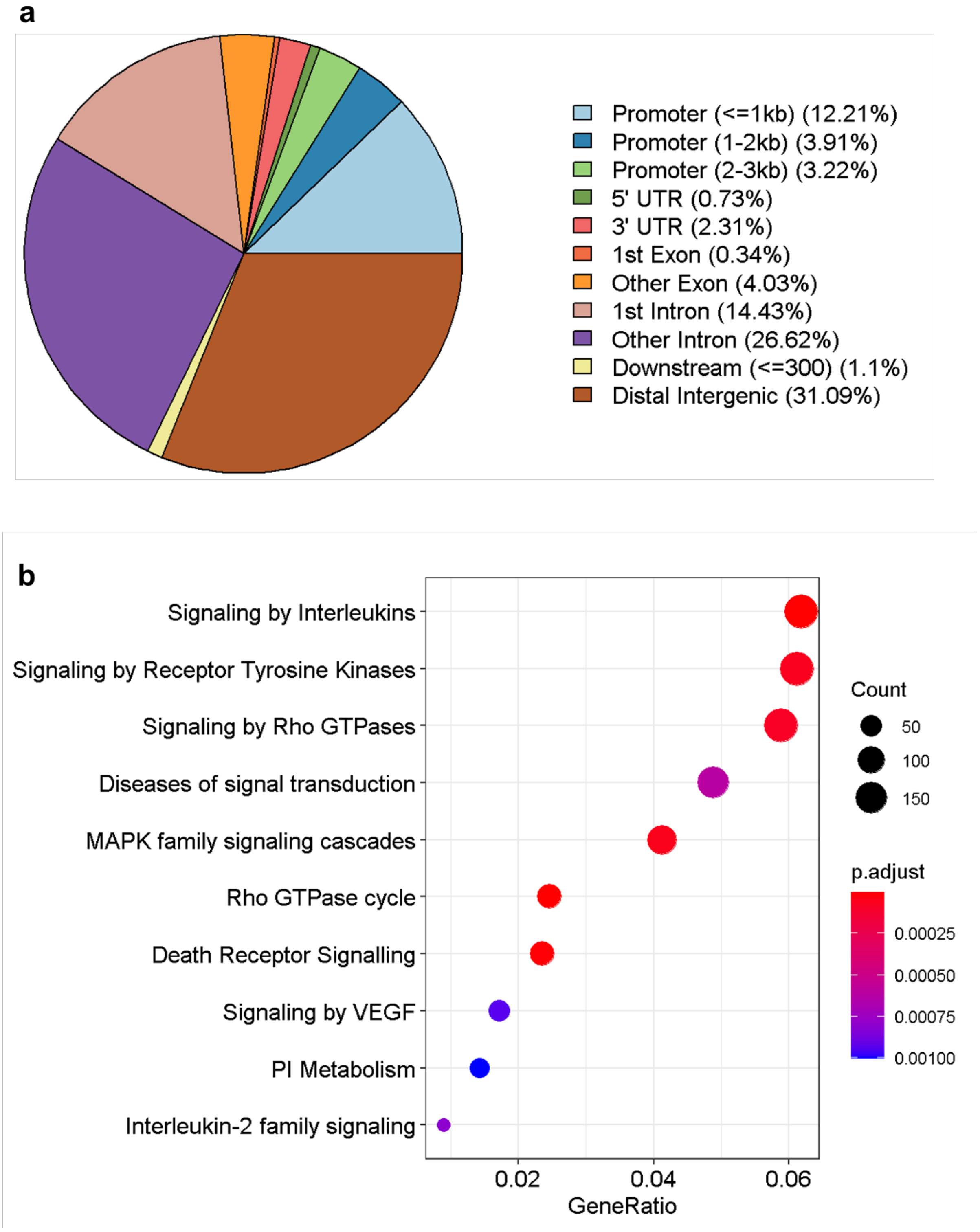


Figure 4-45 | Genomic annotation (a) and pathway enrichment analysis (b) of significantly differential accessible peaks between Treg- and Tconv- ATAC-seq calculated using normalisation Method 3 (see Table 4-2).

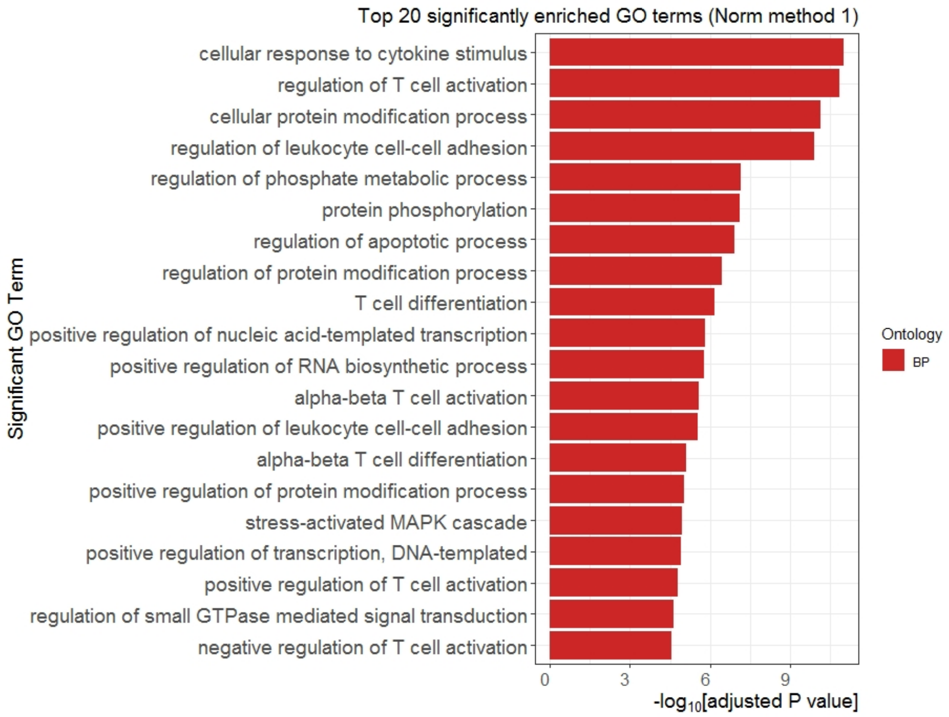


Figure 4-46 | Gene Ontology enrichment analysis showing top 20 significantly enriched GO terms associated with differential accessible peaks between Treg- and Tconv- ATAC-seq, calculated using normalisation Method 1, using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.
Results are representative of accessibility profiles from 12 healthy subjects.

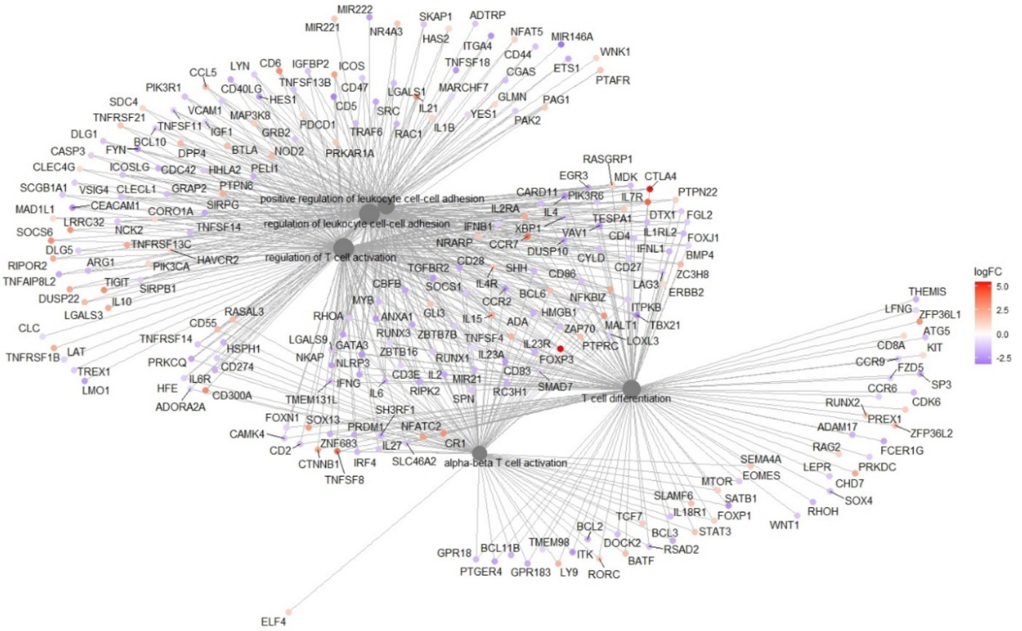


Figure 4-47 | Enrichment plot showing the most significant GO terms and related genes from the Biological Process Ontology for Treg- versus Tconv- ATAC-seq (normalisation Method 1).

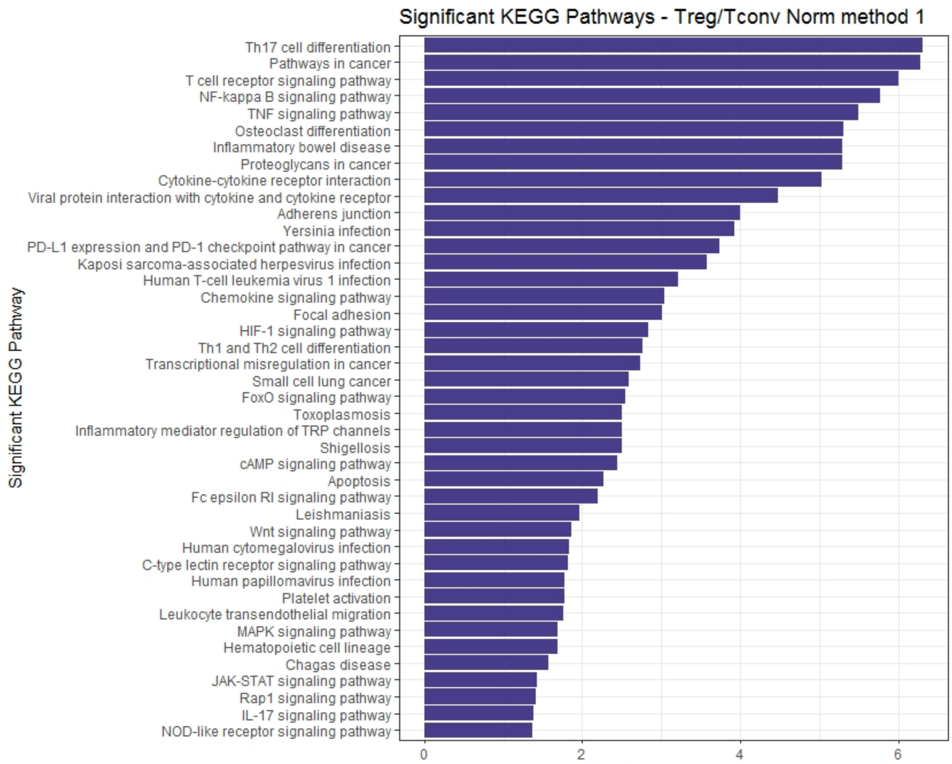


Figure 4-48 | KEGG pathway analysis showing significantly enriched pathways associated with differential accessible peaks between Treg- and Tconv- ATAC-seq (normalisation Method 1).

Results are representative of accessibility profiles from 12 healthy subjects.

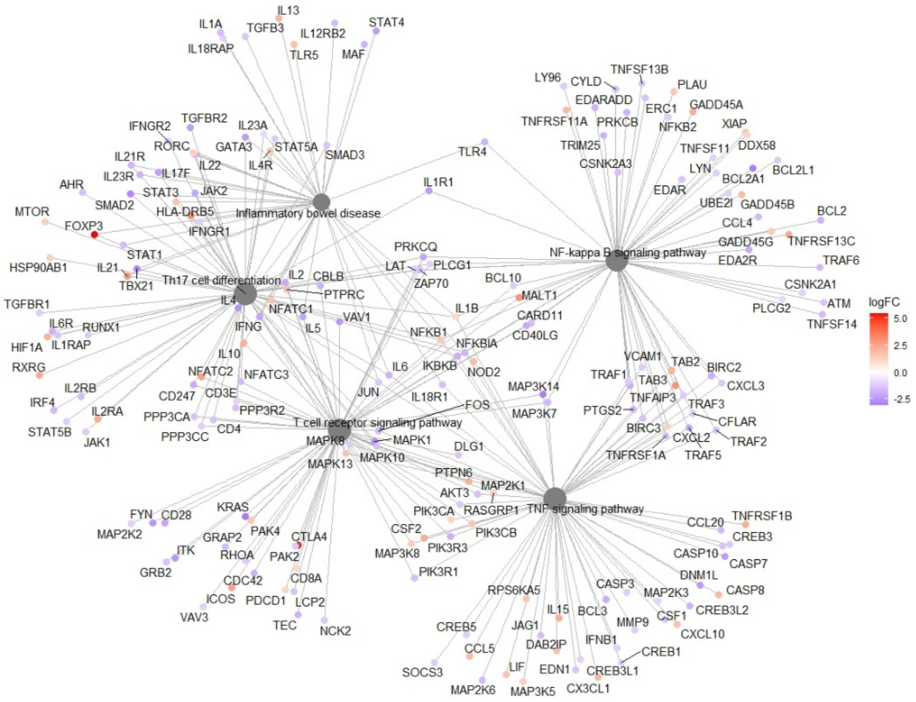


Figure 4-49 | Enrichment plot showing the most significant KEGG pathways and associated genes for Treg- versus Tconv- ATAC-seq (normalisation Method 1).

Results are representative of accessibility profiles from 12 healthy subjects.

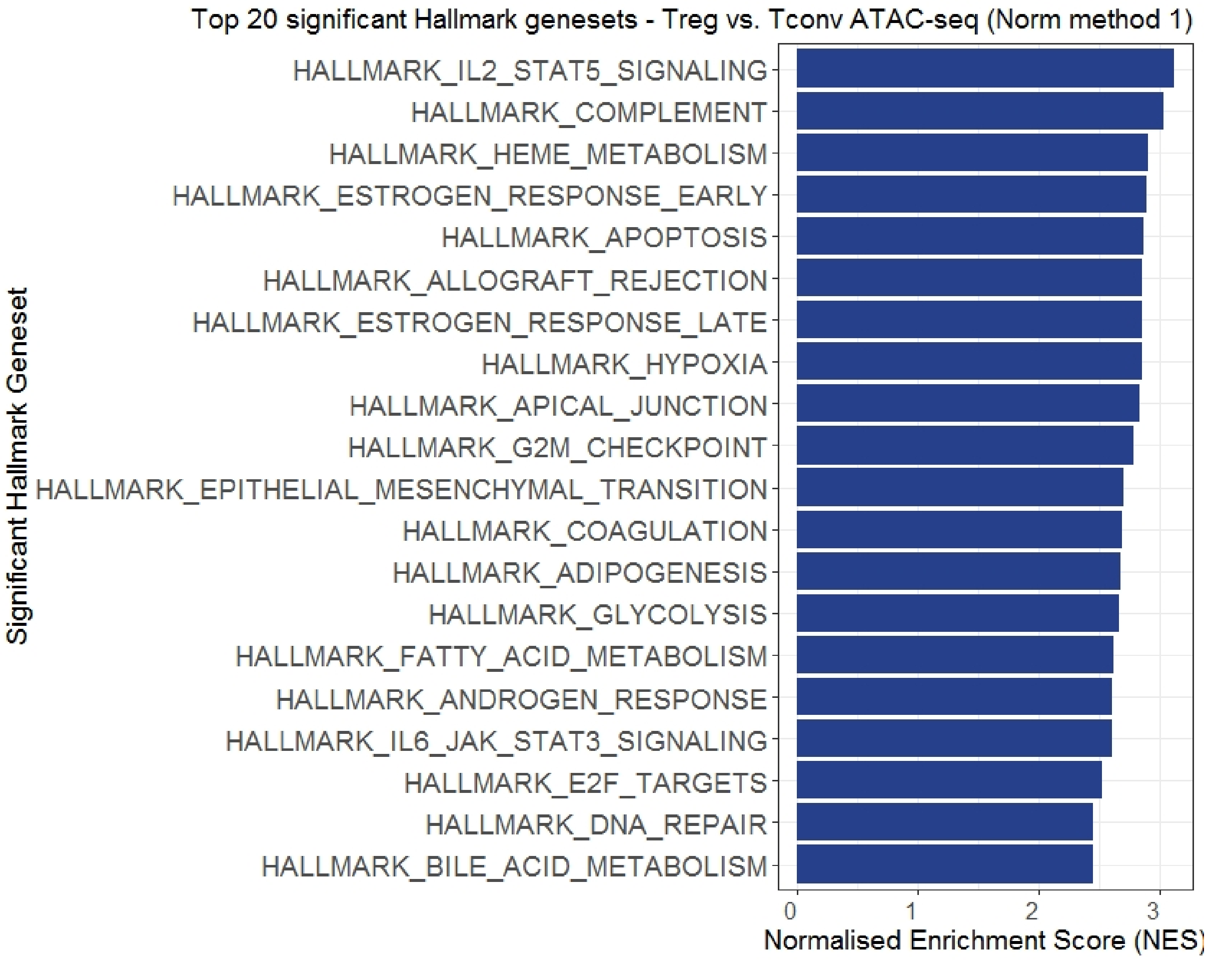


Figure 4-50 | Gene set enrichment analysis (GSEA) showing top 20 significantly enriched (FDR < 0.05) Hallmark gene sets in Treg- vs Tconv- ATAC-seq (normalisation Method 1). Results are representative of accessibility profiles from 12 healthy subjects.

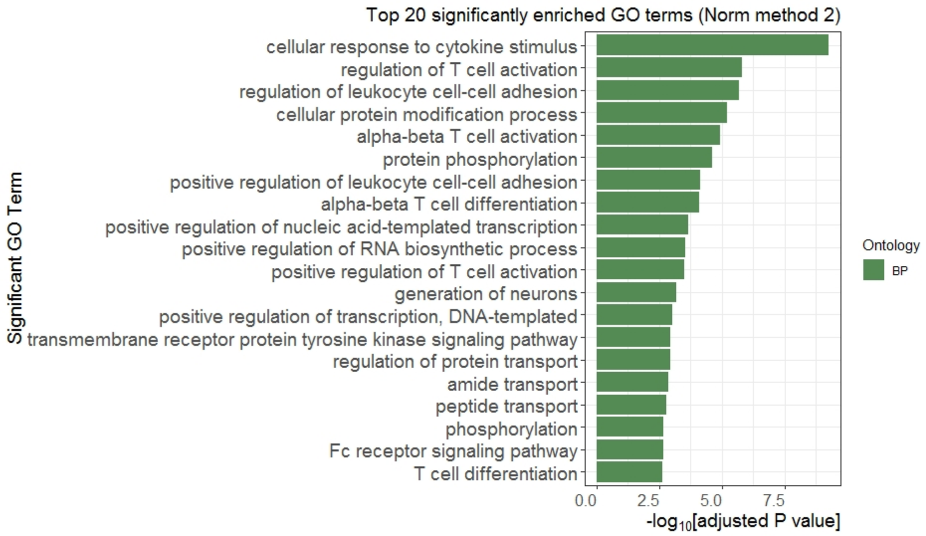


Figure 4-51 | Gene Ontology enrichment analysis showing top 20 significantly enriched GO terms associated with differential accessible peaks between Treg- and Tconv- ATAC-seq, calculated using normalisation Method 2, using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

Results are representative of accessibility profiles from 12 healthy subjects.

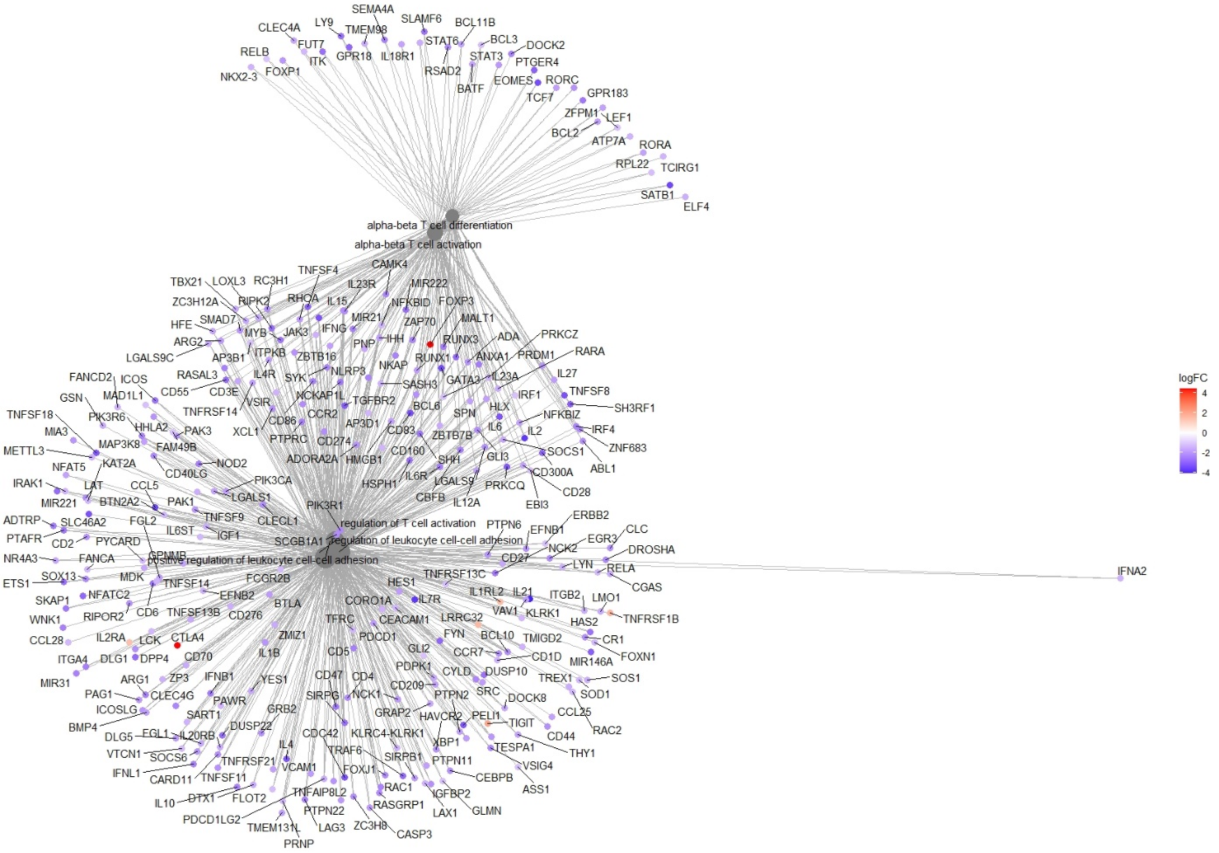


Figure 4-52 | Enrichment plot showing the most significant GO terms and related genes from the Biological Process Ontology for Treg- versus Tconv- ATAC-seq (normalisation Method 2).

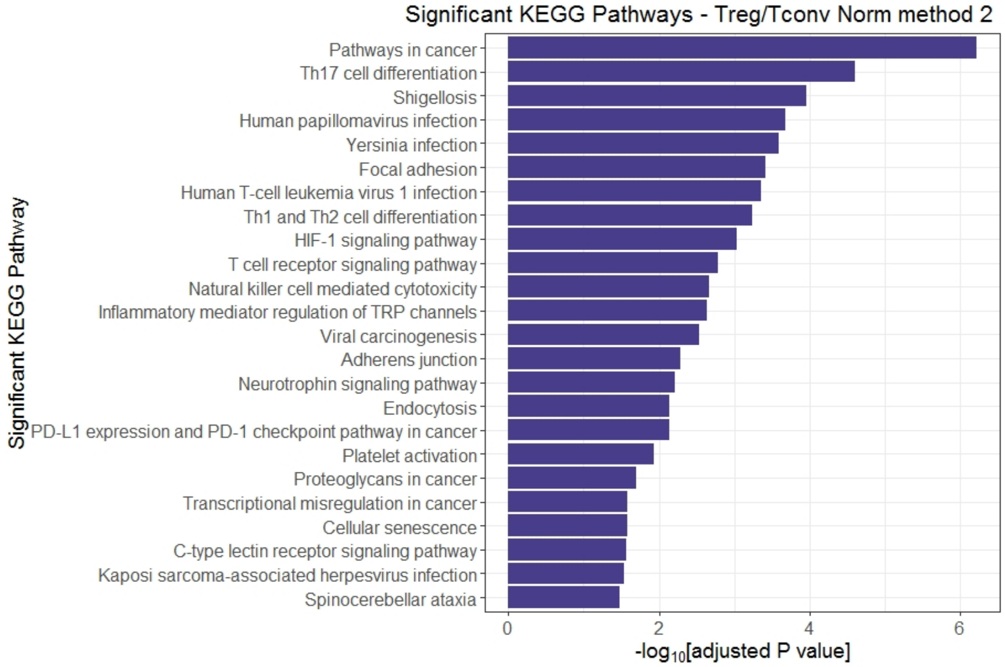


Figure 4-53 | KEGG pathway analysis showing significantly enriched pathways associated with differential accessible peaks between Treg- and Tconv- ATAC-seq (normalisation Method 2).

Results are representative of accessibility profiles from 12 healthy subjects.

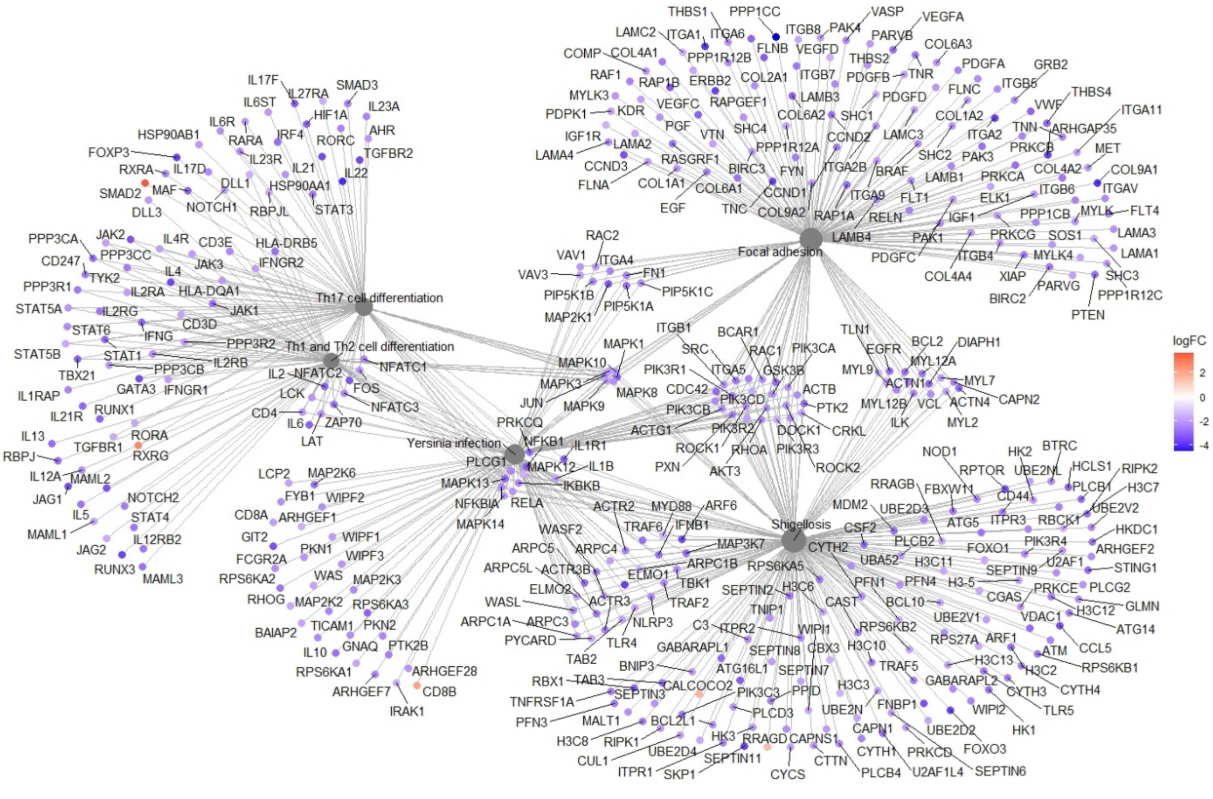


Figure 4-54 | Enrichment plot showing the most significant KEGG pathways and associated genes for Treg- versus Tconv- ATAC-seq (normalisation Method 2).

Results are representative of accessibility profiles from 12 healthy subjects.

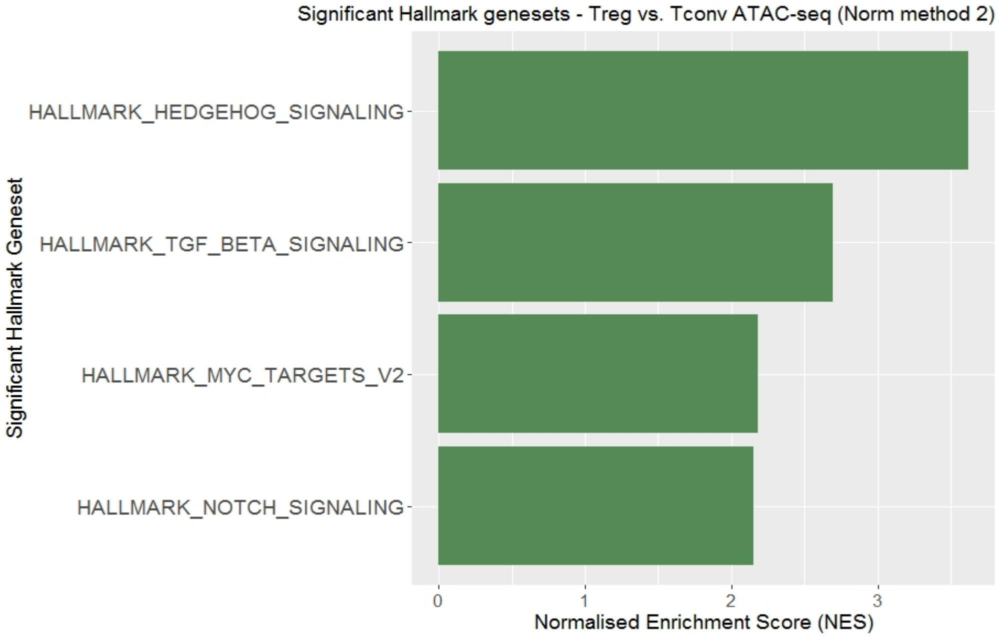


Figure 4-55 | Gene set enrichment analysis (GSEA) showing top 20 significantly enriched (FDR < 0.05) Hallmark gene sets in Treg- vs Tconv- ATAC-seq (normalisation Method 2).

Results are representative of accessibility profiles from 12 healthy subjects.

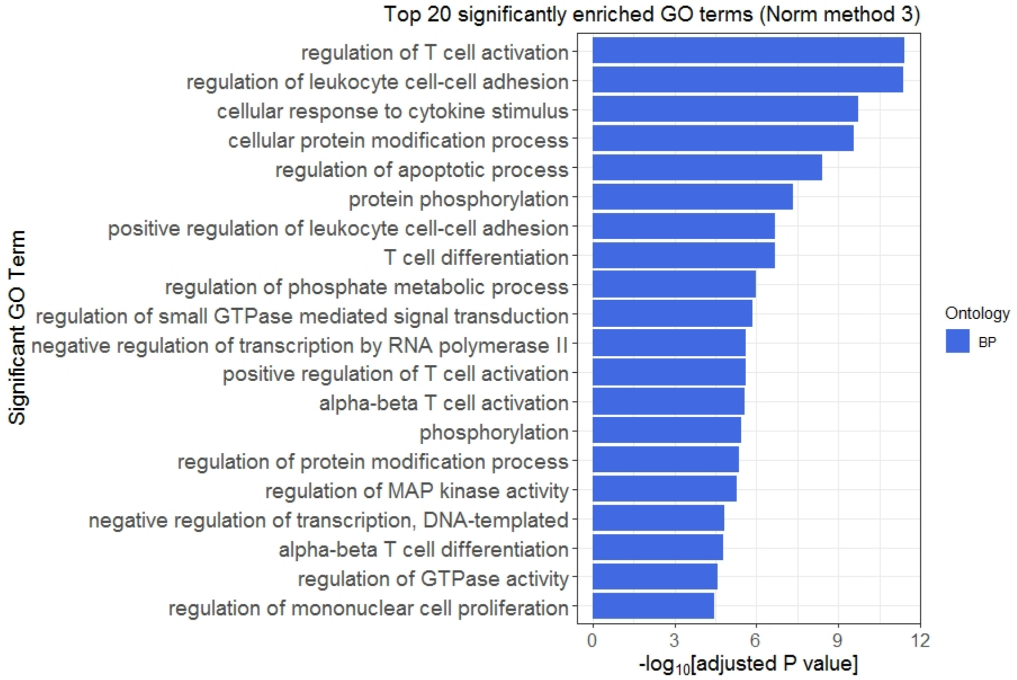


Figure 4-56 | Gene Ontology enrichment analysis showing top 20 significantly enriched GO terms associated with differential accessible peaks between Treg- and Tconv- ATAC-seq, calculated using normalisation Method 3, using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

Results are representative of accessibility profiles from 12 healthy subjects.

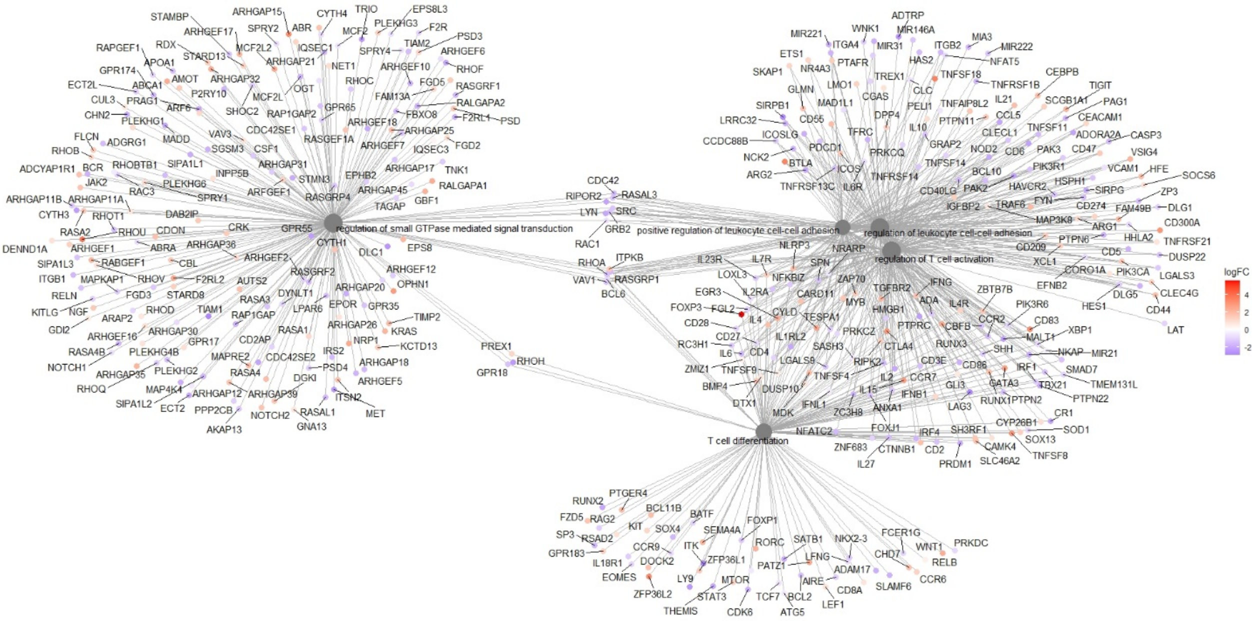

regulation of small GTPase mediated signal transduction
positive regulation of leukocyte cell-cell adhesion
regulation of leukocyte cell-cell adhesion
regulation of T cell activation
T cell differentiation
logFC
4
2
0
-2

Figure 4-57 | Enrichment plot showing the most significant GO terms and related genes from the Biological Process Ontology for Treg- versus Tconv- ATAC-seq (normalisation Method 3).

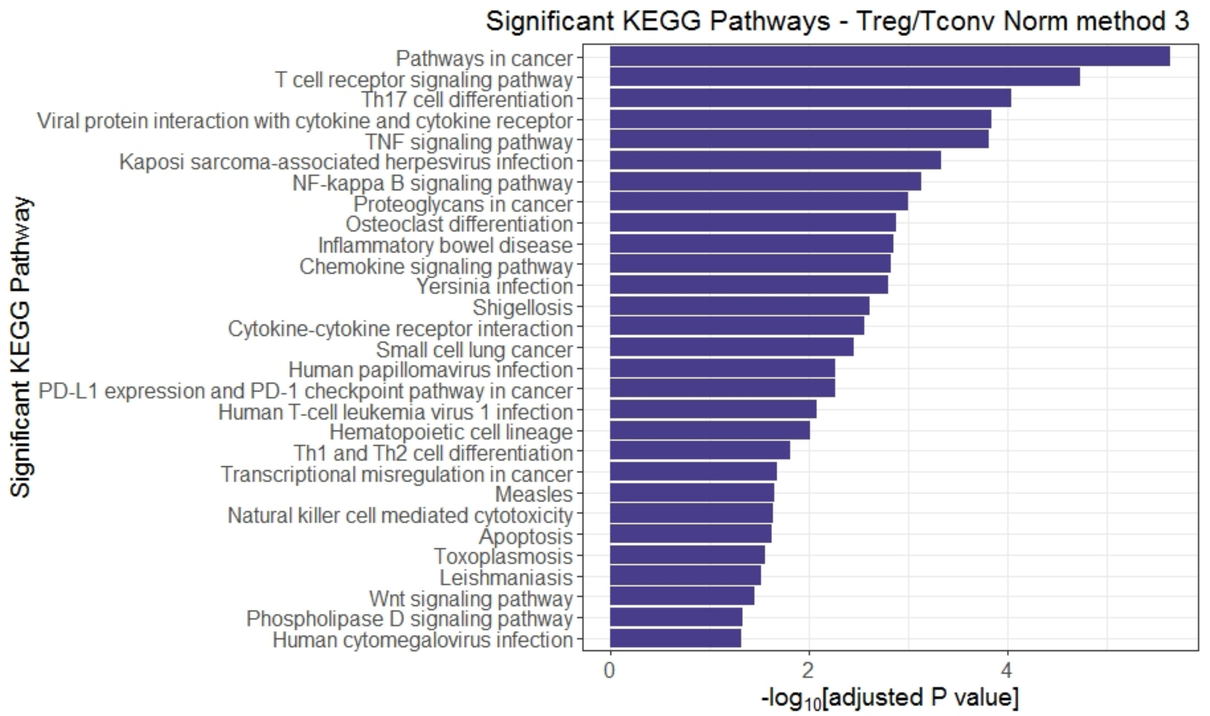


Figure 4-58 | KEGG pathway analysis showing significantly enriched pathways associated with differential accessible peaks between Treg- and Tconv- ATAC-seq (normalisation Method 3).

Results are representative of accessibility profiles from 12 healthy subjects.

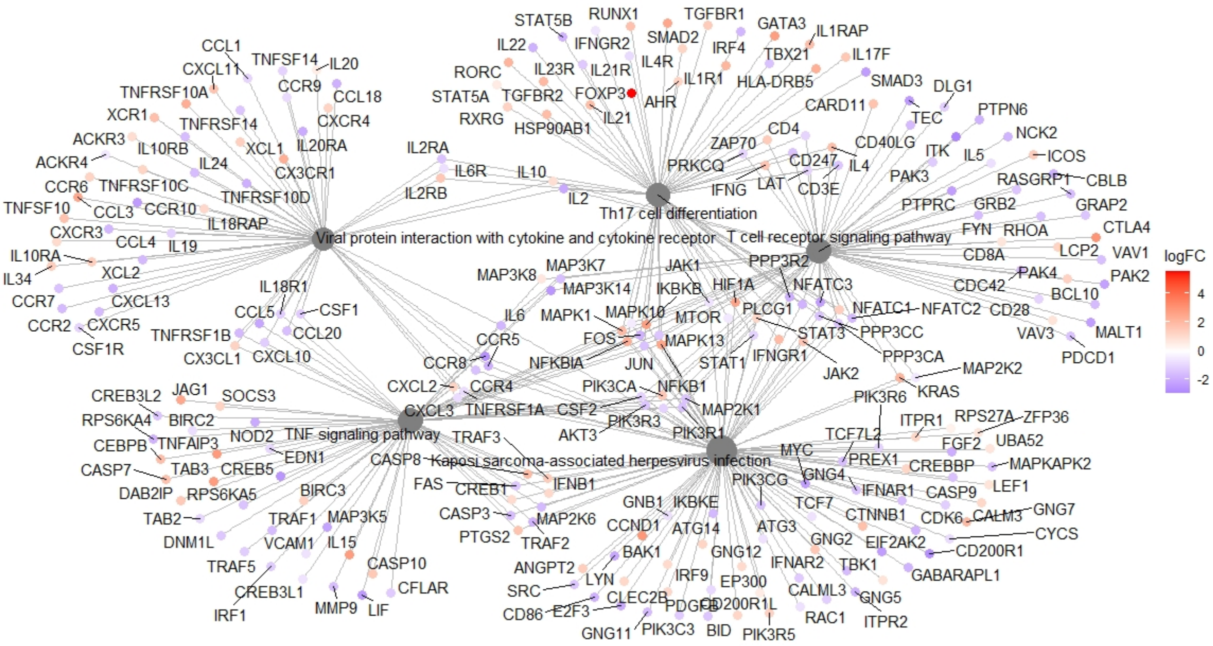


Figure 4-59 | Enrichment plot showing the most significant KEGG pathways and associated genes for Treg- versus Tconv- ATAC-seq (normalisation Method 3).

Results are representative of accessibility profiles from 12 healthy subjects.

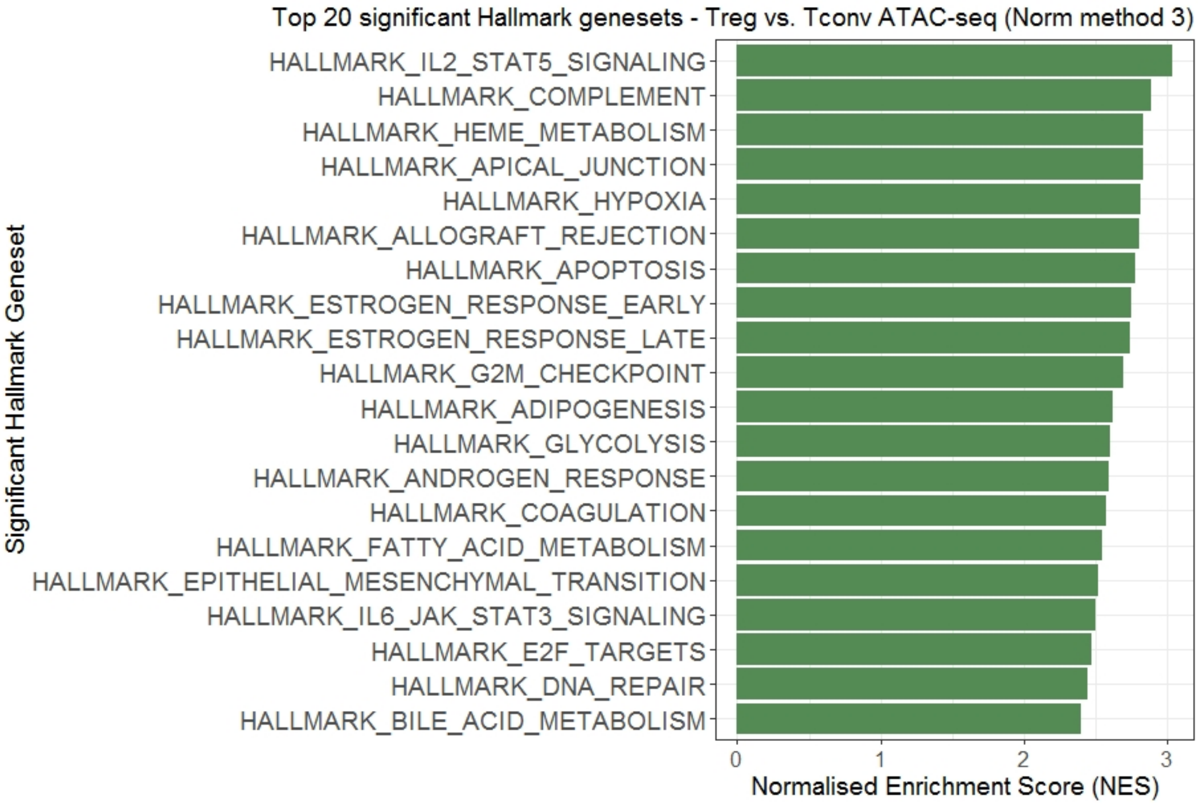


Figure 4-60 | Gene set enrichment analysis (GSEA) showing top 20 significantly enriched (FDR < 0.05) Hallmark gene sets in Treg- vs Tconv- ATAC-seq (normalisation Method 3). Results are representative of accessibility profiles from 12 healthy subjects.

4.5 Discussion

Accurate annotation of functional regulatory elements in a given cell type is critical to understand global gene regulation and dissect the molecular pathways involved in disease. It is critical to develop and apply the best practices for the experimental protocol as well as the computational framework for analysing sequencing data, to ensure high-quality epigenomic data are produced for interpretation and clinical translation. However, when this project was initiated, ATAC-seq technology, including data analysis expertise had not been developed in the Barry lab. Because ATAC-seq is a relatively new sequencing technology, the computational framework dedicated for analysing ATAC-seq data especially on peak calling, data normalisation and differential accessibility is limited and underdeveloped. Data analysis framework for ATAC-seq is similar to DNase-seq and ChIP-seq to some extent and compatible tools are available for only upstream processing of ATAC-seq data such as adapter trimming, read alignment or removal of duplicates. These conventional computational tools fall short as they usually fail to account for additional nucleosome positioning information contained in the Tn5 transposase-digested DNA fragments.

This presents a major limitation in deciphering the data as there is no “gold standard”, or even tools tailored for analysing this kind of dataset. As a result, this necessitates the need to benchmark and improve statistical or analytical methods used for properly interpreting the sequencing data generated by ATAC-seq, and it is crucial to ensure the analysis strategies are compatible with our system of interest in terms of study design and cell context.

Therefore, in this chapter, I assessed and benchmarked analysis strategies in an effort to establish a sound and robust statistical or analytical pipeline for analysing ATAC-seq data. Some of the areas addressed in this section of work include peak calling strategies, differential accessibility (DA) analysis and data normalisation approaches for ATAC-seq data. In addition, because ATAC-seq had not been performed previously within the lab before this project was initiated, this chapter also demonstrates the quality assessment of our in-house T cell Omni ATAC-seq libraries in comparison with a number of published Omni- and Standard- ATAC-seq datasets as a proxy or standard to measure the quality of generated libraries.

Quality evaluation supports the choice of biological sample handling and library generation selected based results presented in Chapter 3, and showed that the in-house Tconv Omni ATAC-seq libraries produced high quality chromatin accessibility data, at a level concordant with, if not better than the published ATAC-seq datasets generated from human primary T and B cells [30, 41, 238]. These quality control metrics include current ENCODE consortium [121] standards for ATAC-seq data, such as alignment rate, fragment size distribution, library complexity, enrichment of transcription start sites (TSS), genomic coverage of reads, reproducibility of replicates and peak annotation. Critically, the accessible chromatin regions identified from my ATAC-seq libraries also displayed high overlap with the published T cell DNase-seq[42] dataset, which is the "gold standard" for profiling accessible chromatin. In addition, I observed enrichment of T cell enhancer elements from primary human CD4$^+$ T cells from a ChIP-seq study [7] in my datasets, as well as peak calling at a level comparable with the published dataset from Corces, Trevino [41] who developed the ATAC-seq protocol.

It is encouraging that my T cell ATAC-seq libraries exhibited exceptional quality overall, as it demonstrates the reliability of methods and capacity for applying this method on the biobanked clinical samples.

Peak calling is fundamentally the most important step of analysing ATAC-seq data. Peak calling identifies enriched accessible chromatin regions above a statistical significance threshold and MACS2 [228], which is a popular program of choice for peak calling has been commonly employed by a vast majority of the ATAC-seq studies [25, 37, 41, 178]. However, MACS2 [228] was designed for ChIP-seq datasets and investigation led by Gaspar [159] revealed that a large number of ATAC-seq studies, including Corces, Trevino [41] - the group who developed Omni ATAC-seq, had used MACS2 to call peaks using parameters that were inappropriate and suboptimal for ATAC-seq datasets, resulting in erroneous or poor quality peak identification. This is due to the fact that ATAC-seq datasets are usually generated from paired-end sequencing yet surprisingly, a substantial number of studies had used a suboptimal peak modelling which resulted in disregard of at least half of the sequencing data and incorrect peak identification at regions where no signal enrichment was observed. This would have a significant negative impact on the analysis of my subsequent T1D case-control ATAC-seq data as that would potentially result in identification of false positive regions and missing of critical disease-relevant regulatory regions.

I incorporated the suggestions and modifications recommended in Gaspar [159] in my peak calling algorithms. They include peak calling in paired-end analysis mode *('-f BAMPE')* and inclusion of arguments *'--min-length'* and *'--max-gap'* to specify the minimum peak length requirement and maximum gap between significant sites for peak merging. Overall, using modified peak calling algorithms (MACS2 ver. 2.1.2) I observed higher number of peaks called, higher proportion of intergenic peak assignment and smaller peak size coverage compared with original, unmodified peak calling mode (MACS2 ver. 2.1.0) in both resting and stimulated ATAC-seq libraries. The peaks called by MACS 2.1.2 were more resolved and accurately represented the chromatin accessibility signal of the cells where lower false positive and negative were observed in comparion with peaks called from MACS 2.1.0. One important gene regulatory element captured by ATAC-seq is distal gene enhancers which are enriched in non-coding intergenic regions [7, 88] and studies have reported enrichment of ATAC-seq peaks in those genomic regions [187, 239-241]. Thus, the enrichment of open chromatin to intergenic regions using the modified algorithms demonstrates the validity and accuracy of the approach in capturing the accessible regulatory signatures. It has been reported that large ATAC-seq peaks can result in bias in differential accessibility and TF footprinting analyses and limit the ability to resolve independent peaks [242]. The lower peak width distribution observed in peaks called from the modified approach improves the resolution of peaks and confidence for subsequent count-based and motif-specific analyses. Fraction of reads in peaks (FRiP), which measures the signal-to-noise ratio, is one of the metrics used by ENCODE consortium[42] in assessing the quality of ATAC-seq datasets. It also serves as an important indicator of peak calling accuracy as sequencing reads should fall within regions of enrichment classified as peaks and reads mapped to outside of peak regions are predominantly background noise. The modified algorithms captured a lower FRiP score in resting cells and a higher FRiP score in stimulated cells compared with unmodified peak caller. This is not unexpected because stimulation results in large-scale chromatin remodelling and a global increase in chromatin accessibility in cells [45, 136], leading to increased pileup coverage and signal threshold for more reliable and confident peak calling than steady-state cell context. In resting cells where overall difference in accessibility across genome is minimal or not as pronounced as in stimulated cells, peak callers may have lower confidence in calling significant peaks as it is harder to differentiate true signals from background noise as demonstrated in MACS2 ver. 2.1.0. This explains the observation of larger peak width window in peaks identified from the unmodified approach (MACS2 ver. 2.1.0).

Overall, I saw improvements in peak identification and quality of peaks called through incorporation of modifications suggested in Gaspar [159] in my peak calling workflow. The next important question is can we reliably identify peak regions that show differential accessibility between different conditions of interest, for example, in Tconv and Treg cells where significant changes in gene regulation is expected [235-237, 243]? The ultimate goal of my PhD is to identify genomic regions that exhibit alterations or differential gene regulation between T1D patients and healthy individuals, and it is critical to ensure the algorithms I use correctly identify regions where changes are expected to occur.

Accurate identification of differential accessible (DA) chromatin regions is challenging owing to variability in transposition reaction efficiency which can be further compounded by heterogeneity in sorted cell populations and lack of comprehensive literature for optimal practices of analysing ATAC-seq data. Computational tools for differential analysis of RNA-seq and ChIP-seq data are well established and documented [38, 188, 244] however, to date there is none designed for DA analysis of ATAC-seq data. Different approaches of DA analysis often yield conflicting results and little emphasis is placed on statistical considerations of normalisation approach during DA analysis. Reske, Wilson [230] demonstrated that different ATAC-seq normalisation approaches can have a significant impact on the outcomes of DA analysis and interpretation. It is therefore crucial to compare multiple normalisation methods before DA analysis as well as to understand the impact of normalisation assumptions on the biases inherent in the data.

In this section of work, I used Tconv and Treg ATAC-seq profiles to benchmark analytical methods for calculating differential accessibility as there is substantial prior knowledge on the transcriptional regulation of Tconv and Treg[235-237, 243]. By pairing multiple data normalisation approaches with DA analysis modelling, including the approaches described in Reske, Wilson [230], I aimed to identify workflow of which the interpretations support the biological relevance of Tconv and Treg regulation. The results showed that normalisation using TMM (Trimmed mean of M values) (Method 1) and loess (locally estimated scatterplot smoothing) normalisation (Method 3) to generate linear scaling factors from raw peak counts led to a symmetric global accessibility distribution in DA analysis between Tconv and Treg, as indicated by symmetric MA plots.

MA plot is a type of Bland–Altman plot commonly applied to genomic data for visualisation of global patterns resulting from differences in measurements between two comparison groups [245]. In this context I used MA plots to identify global accessibility patterns of differential analysis measurements between Tconv and Treg ATAC-seq upon data normalisation. An upward or downward shift in the MA distribution could either indicate a global effect or substantial technical bias, and data normalisation will result either result in no changes or elimination of these features in the data [230]. In general if the MA distribution does not appear symmetrical it may be indicative of trended bias, which can be corrected using conservative normalisation approaches like loess [230].

TMM (Trimmed mean of M values) normalisation assumes that most regions are not differentially accessible and controls for technical bias by eliminating systematic errors in ATAC-seq libraries, while still allowing asymmetric differences arising from true DA regions [246]. Whereas loess (locally estimated scatterplot smoothing) normalisation is a highly conservative method which normalises the ATAC-seq signal locally and because it assumes a symmetric global distribution in which there are no true biological differences for ATAC-seq transposition efficiency or signal distribution, any observation of these biases will be thoroughly eliminated [230]. Incorporation of TMM and loess normalisation methods in my DA analysis pipeline generated promising results reflecting biologically expected changes in chromatin accessibility between Tconv and Treg cells. DA and gene ontology (GO)/gene set enrichment analyses (GSEA) for Tconv and Treg ATAC-seq identified critical loci and pathways that strongly reflect the biology and differential regulation between the two cell types, corroborating findings from a large number of studies involving RNA-seq, microarray and flow cytometry experiments [235-237, 247]. The DA analysis captured most of the gene signatures or loci known to be differentially expressed or regulated between Tconv and Treg, including TF and cell surface molecules such as *FOXP3*, *TIGIT* (T Cell Immunoreceptor With Ig And

ITIM Domains), *IKZF2* (IKAROS Family Zinc Finger 2) and *CTLA4* (cytotoxic T-lymphocyte-associated protein 4) in the top differentially accessible peaks. Whereas the pathway and network enrichment analyses showed that the differentially accessible regions were strongly enriched for categories associated with T cell differentiation, cellular response to cytokine stimulus, regulation of T cell activation, T cell receptor signalling pathway and IL2-STAT5 signalling.

Overall, these findings support the work of Reske, Wilson [230] that different normalisation approaches result in varying global accessibility distributions and DA outputs. I also observed conflicting DA outcomes from using different normalisation approaches. The results demonstrate that TMM and loess normalisation methods show high concordance in DA outputs which are biologically relevant to my T cell datasets. TMM was chosen as the normalisation method of choice with which to proceed forward for any downstream DA analysis, including T1D case-control ATAC-seq datasets, due to high computing requirements of loess normalisation and potential significant elimination of true alteration signals owing to its highly conservative nature.

CHAPTER 5: IDENTIFICATION OF GENETIC AND EPIGENETIC CHANGES THAT ASSOCIATE WITH TYPE 1 DIABETES (T1D)

5.1 Introduction

Type 1 diabetes (T1D) is a chronic autoimmune disease characterised by increased blood glucose levels (hyperglycaemia) owing to the immune system attacking and eventually destroying the insulin-producing pancreatic β-cells [66, 67, 72, 90, 248]. In 1986, work from Eisenbarth [248] led to the recognition that T1D as a chronic autoimmune disease associated with a long pre-diabetic period identified by the appearance of islet cell autoantibodies with polyendocrine deficiencies. He described that this chronic autoimmune process, triggered by unknown factors, persisted over many years and involved the destruction of insulin-producing pancreatic β-cells by autoreactive lymphocytes. Since then, extensive human and animal studies have reinforced his concept that this progressive condition is accompanied by β-cell destruction and dysfunction. In T1D, autoantibodies can appear many months or even years before symptom onset, they are not thought to be pathogenic but rather serve as biomarkers for diagnosis and prediction of T1D. The autoantibodies are mainly those targeting insulin, glutamic acid decarboxylase (GAD), a tyrosine phosphatase-like protein (islet antigen-2 [IA-2]) and zinc transporter-8 (ZnT8). Individuals with HLA-DR and HLA-DQ genotypes demonstrated increased risk of developing two or more autoantibodies and T1D and the lifetime risk of progressing to symptomatic T1D approaches 100% once two or more islet autoantibodies are detected in genetically susceptible children [70, 80, 82].

T1D is one of the most common childhood-onset chronic diseases with the incidence of T1D in Australia currently ranked as the 7th highest among the 30 OECD (Organisation for Economic Cooperation and Development) nations [249]. Currently there is no cure for T1D, with patients requiring lifelong exogenous insulin administration for management of blood glucose levels, and it is associated with significant human and economic costs. The reason a child develops T1D is not well understood but it appears to depend on a complex interaction of genetic predisposition and environmental factors that trigger or permit the autoimmune response against the β-cells. Work over the past decades has strengthened the concept that environmental factors (e.g. diet, infection) contribute strongly to the pathogenesis of T1D. This is supported by the observations that the incidence of T1D has been increasing in younger children worldwide [90] at a rate that is faster than can be accounted for genetic change, such as genotype frequencies in the population, alone. Increased T1D incidences in children with low-risk HLA haplotypes, such as HLA-DQB1*0602 strongly implied an environmental influence. This is also supported by data from migration cohorts, where immigrants take on the risk of the local population [68]. A likely route for environmental factors to alter the immune system is via epigenetic mechanisms such as histone modification, DNA methylation and altered microRNA expression and alteration in all these have been reported in T1D patients [97, 98].

The key unanswered mechanistic questions are: How do genetic and environmental factors combine to cause the loss of this immune regulation and attacking on the pancreatic β cells in T1D? What epigenetic modifications are involved in the pathogenesis of T1D and how do they regulate the genes differently to contribute to the disease?

There is compelling evidence that imbalance between regulatory T (Treg) cells and conventional T (Tconv) or helper T (Th) cells plays a key role in T1D autoimmunity. In both mouse (Scurfy) and human (IPEX), loss of FOXP3 resulting in lack of normal Treg formation, results in spontaneous autoimmune disease, including T1D from birth [250-252]. FOXP3^{+} Treg cells are a key mediator of immune tolerance and they prevent autoimmunity in healthy individuals. They modulate T-cell activation and promote immune tolerance through direct cell-cell interactions and secretion of immunosuppressive cytokines such as IL10 and TGF-β.

Although Treg cells are widely reported to contribute to the immunological defect in many autoimmune diseases including T1D, the molecular mechanisms leading to impaired Treg function in T1D remain poorly understood. Loss of Treg function can be attributed to a functional defect in the Treg themselves or the acquisition of resistance to Treg suppression by Tconv cells. Deficiency in local Treg number, reduced IL-2R signalling and instability of suppressive activities have been reported in T1D [105, 106, 253] and adoptive transfer of Treg cells into T1D mouse model, non-obese diabetic (NOD) mice, results in disease prevention and remission [254]. It is plausible that many of the genetic risk factors and environmental trigger associated with T1D disrupt this Treg-dependent immune homeostasis.

Genome-wide association studies (GWAS) [2, 3, 85-87] have revealed 119 lead SNPs spanning across 59 non-HLA genomic loci that are associated with T1D. These non-HLA risk loci appear to be implicated in the increased incidence of T1D over the last few decades instead of the HLA loci, but how these regions contribute to disease has remained elusive [68, 91, 92].

In T1D a majority of the risk variants do not alter coding sequence but rather appear to be enriched in distal regulatory elements, such as gene enhancers, that are active in lymphoid cells including CD4$^+$ T cells (Tconv and Treg) [2, 88, 89], suggesting that they are more likely to influence transcriptional regulation rather than altering protein coding sequences and function. One or more of the T1D associated risk variants in these regions, the causal SNP(s), are likely to interfere with enhancer function altering the expression of the genes normally controlled by enhancers in Treg and Tconv cells. This is supported by the observation that the majority of the identified T1D-associated risk loci are implicated in immune regulation and functions, whereas only a limited number is linked to the formation of islet autoantibodies[10, 255].

Defining and functionally characterising genetic variants is a crucial step for interpreting GWAS and understanding how these risk loci contribute to disease. This includes identifying disease causal SNPs, mechanisms by which they act and their target genes. However, it is a challenge to identify the true disease drivers because GWAS-identified lead SNPs serve only as representatives for all the SNPs in the same haplotype block and it is likely that other SNPs in high linkage disequilibrium with the lead SNPs are causal for the disease. These susceptibility regions often i) span large genomic regions, ii) contain many co-inherited bystander SNPs (SNPs in high linkage disequilibrium), and iii) may contain more than one independent association. As disease-associated genetic variants are likely to influence transcriptional regulation rather than altering protein function, they could possibly exert their function through epigenetic mechanisms such as DNA methylation and histone modification which are plausible drivers of dysregulated gene expression, and alterations in these epigenetic signatures have been reported in T1D patients [97, 98].

In genetically predisposed individuals the combination of genetic and environmental alterations is sufficient to compromise Treg-dependent immune homeostasis and trigger autoimmune response against the β-cells. Whether these epigenetic alterations occur at T1D risk loci and/or elsewhere has not been resolved. Genomic and epigenomic profiling that incorporates functional annotation of regulatory elements can help prioritize the non-coding risk variants in T1D, and experimental approaches that aim to identify their target genes would help in revealing their functional relevance in disease.

Although multiple studies have demonstrated the enrichment of T1D associated risk variants in T cell specific transcriptional enhancers in healthy individuals [2, 88], despite the indispensable roles of Tconv and Treg cells in the pathogenesis of T1D, to date, no study has examined the chromatin structure or accessibility of those enhancers in primary Tconv and Treg cells from T1D patients, or determine how modifications in these enhancers might lead to altered expression of the genes controlled by these enhancers. Here in this chapter, I described the epigenomic and transcriptomic profiles of Tconv and Treg cells from T1D patients in comparison to their healthy siblings in order to understand the links between chromatin landscape and gene expression. High resolution chromatin accessibility profiling using ATAC-seq also enabled me to identify alterations in TF occupancy in enhancer regions differentially regulated between the T1D cohort and healthy controls. By incorporating chromatin conformation capture (3C) contact maps generated from Hi-C, H3K27ac HiChIP and promoter capture Hi-C technology in the chromatin accessibility data of T1D patients, I was able to map target genes and assign molecular functions to altered enhancers in T1D, which control genes beyond the nearest gene in the linear genome sequence. This section of work revealed significant alterations in the enhancer repertoire, transcriptome and TF activity in primary Tconv and Treg cells of established T1D cohort. Intersecting my epigenomic data with a catalog of common human variants from the Genome Aggregation Database (gnomAD) [1], I identified a subset of novel candidate T1D risk variants, independently of association studies, located in enhancers altered in T1D.

5.2 Aims and Hypothesis

The hypothesis for this chapter is that genetic variation and epigenetic changes in T1D modify the activity of Treg- and Tconv-specific enhancers, leading to altered expression of the genes normally controlled by these enhancers.

The aims of this chapter are:

1. To compare chromatin accessibility in Treg and Tconv from T1D cohort and healthy controls using ATAC-seq
2. To identify T1D associated targets and TF footprints from ATAC-seq chromatin accessibility profiles by integrating Hi-C and RNA-seq data

5.3 Material and Methods

5.3.1 Cohort information and sample size

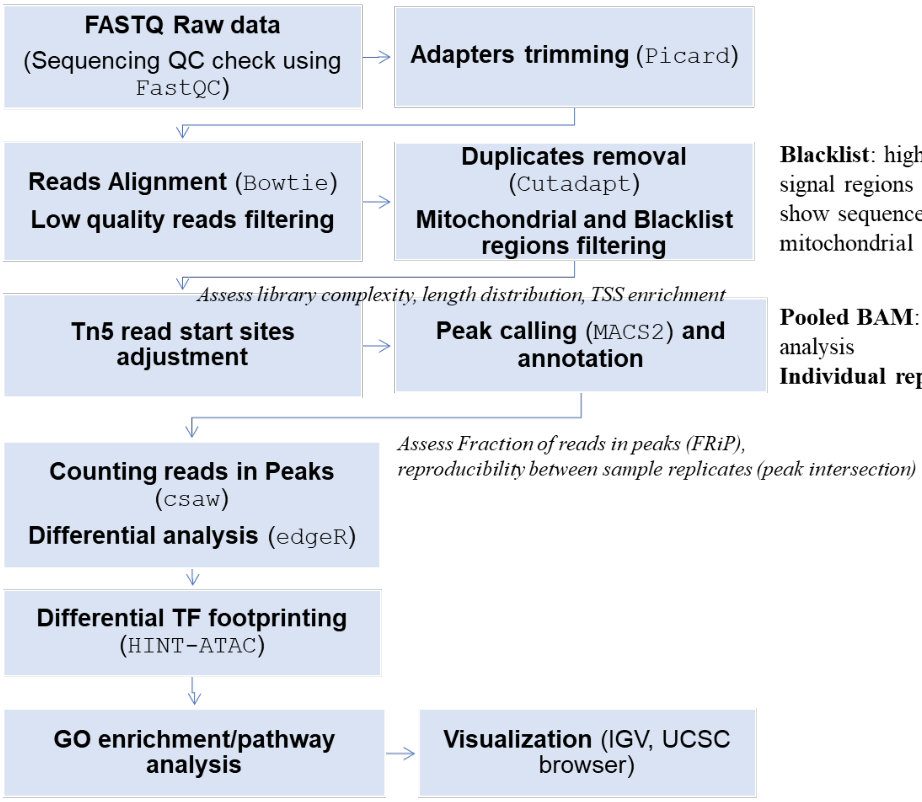
All genomics/epigenomics work was carried out in accordance with WCHN human research ethics committee approval (REC No. 1596/08/2019). The resources procured for this study came from a repository of biobanked PBMCs samples obtained from the Australian Type 1 Diabetes and the Gut (TIGs) cohort [256, 257]. The cohort used in this study included youth newly diagnosed with T1D from 4 weeks to 2 years post diagnosis and their islet autoantibody-negative siblings recruited between 2014 to 2017. The biobank was established at the Women's and Children's Hospital, Adelaide, South Australia via Professor Jennifer Couper and Dr Jessica Harbison (Endocrinology and Diabetes). Table 5-4 summarises demographic and other reported information [age = 11 ± 3.4 years; 62% male; 58% gender-matched pairing (71% males); 50% HLA-matched pairing] for PBMC samples from 12 pairs of T1D and sibling-matched healthy controls that passed cell viability (>90%) upon thawing and minimum recovery ($\geq 8\text{x}10^6$ live PBMCs) criteria.

5.3.2 Thawing of biobanked samples and cell preparation for sequencing experiments

Frozen PBMCs from the biobank were thawed in pairs (i.e. T1D samples alongside respective sibling-matched healthy controls) using the optimised thawing protocol described in Section 3.3.1. Briefly, the frozen PBMC vials were retrieved from the biobank and thawed by incubating in a 37°C water bath for 10 minutes. The fully thawed cells were then quenched slowly (at a rate of 1 mL/5 seconds) with 9mL of pre-warmed, 37°C complete X-VIVO culture media supplemented with 2 mM HEPES pH 7.8, 2 mM L-glutamine, 5% heat inactivated human serum and 200U/mL DNase (Worthington cat# LS002007) twice by centrifugation at 500g for 10 minutes. The thawed cells were allowed to recover overnight at a cell density of ~3.5–4.0 x 10^6/mL, in the complete X-VIVO culture media in a 24-well culture plate at 37°C in a CO_2 incubator for 16-17 hours prior to cell enumeration and T cell sorting by FACS as described in Section 2.1.4). Upon cell sorting by FACS, cells were plated at 100,000 cells per well in a 96-well U-bottom plate and maintained in complete X-VIVO 15 culture media in the presence of 500U/mL recombinant human IL-2 for 2 hours at 37^{o}C prior to cell preparation for ATAC-seq and RNA-seq experiment. For T cell activation, following 2-hour post-sort recovery cells stimulated with beads conjugated with anti-CD3 and anti-CD28 antibodies (Dynabeads Human T-Expander CD3/CD28, Gibco no. 11141D, Life Technologies) in complete X-VIVO 15 culture in the presence of 500U/mL recombinant human IL-2 at a cell/bead ratio of 1:1 for 48 hours. After 48 hours Dynabeads were removed from culture medium by magnetic separation for Omni ATAC-seq and RNA-seq experiments. Omni ATAC-seq was performed as described in Section 2.2 and RNA-seq libraries were prepared from whole cells and supernatant fractions collected from Omni ATAC-seq lysis reaction (ATAC-SN) as described in Section 2.3. Intact RNA samples (RIN > 7) were enriched for Poly(A) RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs; cat# E7490) whereas partially degraded RNA samples (RIN 6 – 7) were ribosomal RNA-depleted (New England Biolabs; cat# E6310) prior to generation of cDNA libraries using a NEBNext Ultra Directional II RNA Library Preparation Kit for Illumina (New England Biolabs; cat# E7760).

5.3.3 Bioinformatics data analysis

5.3.3.1 ATAC-seq peak calling and processing


FASTQ Raw data
(Sequencing QC check using FastQC)
Adapters trimming (Picard)
Reads Alignment (Bowtie)
Low quality reads filtering
Duplicates removal (Cutadapt)
Mitochondrial and Blacklist regions filtering
Assess library complexity, length distribution, TSS enrichment
Tn5 read start sites adjustment
Peak calling (MACS2) and annotation
Assess Fraction of reads in peaks (FRiP), reproducibility between sample replicates (peak intersection)
Counting reads in Peaks (csaw)
Differential analysis (edgeR)
Differential TF footprinting (HINT-ATAC)
GO enrichment/pathway analysis
Visualization (IGV, UCSC browser)

Blacklist: high signal nuclear artefact; high signal regions on the nuclear genome that show sequence homology with the mitochondrial genome.

Pooled BAM: Call "query" peaks for DA analysis
Individual replicate: QC, annotation

Figure 5-1 | Flowchart showing bioinformatics pipeline for ATAC-seq data processing, measurement of quality control metrics and differential accessibility analysis.

Processing and analysis of ATAC-seq data for Tconv and Treg samples were performed according to the workflow described in Figure 4-4, Figure 5-1 and Section 2.4.1.1. Depending on the nature of data analyses, the processed ATAC-seq reads were concatenated by cohort group (T1D/healthy) from all sample replicates (n=12 pairs) or gender-matched replicates (n=5 pairs) and peaks were called from the bam files containing pooled reads with arguments '*callpeak -f BAMPE -g hs --nolambda --min-length 100 --max-gap 50 --call-summits --bdg --keep-dup all*' using MACS2 ver. 2.1.2 [159]. The 1-bp peak summits for each peak were extended 250bp on both sides to a uniform peak size of 500bp using BEDTools (ver. 2.25.0). I chose to use fixed-width peaks for downstream analysis because it makes count based and motif specific analyses less biased to large peaks (in differential peak and TF footprinting analyses). In addition, it improves the ability to resolve independent peaks as merging overlapping peaks of varying sizes can result in many peaks being merged into one large peak. I then ranked the 500-bp peaks by summit significance value (defined by MACS2) and selected the top 100,000 peaks for analyses. As preliminary exploratory analyses revealed gender as the largest proportion of variation driving the separation of ATAC-seq samples, in order to mitigate the issue of gender bias resulting from disproportionate representation of males and females in this study design, peaks mapped to chromosome X and Y were excluded from downstream analyses. For each T1D- and healthy-specific peakset, overlapping peaks were merged using BEDTools ver. 2.25.0. The T1D- and healthy-specific peaks from Treg and Tconv datasets were merged to obtain a union peak set, resulting in 68,510 and 71,902 peaks in the Treg and Tconv datasets, respectively.

5.3.3.2 Identification of confounding variables in the sequencing data and data normalisation

To assess general patterns and identify sources of variance in the ATAC-seq data, principal component analysis (PCA) was performed on ATAC-seq count data in R Statistical Software using Bioconductor package prcomp. PCA identifies the directions of dominating variability in the dataset and is used to determine the presence of confounding factors, either biological or technical in origin, driving the variance in the data. The correlations between the principal components and various suspected confounding variables (Table 7-3) were calculated to check for any apparent biases in the data.

As preliminary exploratory analyses revealed guanine-cytosine content (GC-content) had a sample-specific effect on the chromatin accessibility and gene expression measurements, conditional quantile normalization (CQN) [40] was performed in R Statistical Software as an attempt to correct for global distortions resulting from this systematic bias in the ATAC-seq and RNA-seq data.

5.3.3.3 Differential chromatin accessibility analysis

All differential accessibility analyses between T1D and healthy controls was performed in a family-based pairwise design to control for confounding variables arising from genetic heterogeneity and environmental components between families.

Differential chromatin accessibility between T1D and healthy controls was calculated for two peak sets of genomic features of interest – all top 100,000 filtered peaks (as described in Section 5.3.3.1) and filtered peaks (top 100,000) intersecting TSS, super- and typical- enhancers of CD4$^+$ T cells [7, 8].

The number of reads mapped to each peak set was calculated using csaw [160]. Reads within the peaks were quantified using *regionCounts* function with peaks having low counts removed based on average log2 counts-per-million for each row of counts. The normalization factors across libraries were computed using *normFactors* to scale the peak library sizes for each sample and imported into R for differential accessibility analysis using edgeR exactTest [38]. Count offsets were computed using the trimmed mean of M-values (TMM) method (described in Section 4.3.4) on the raw count of the ATAC-seq peaks from all the samples. Starting with a consensus peak set, average log$_2$CPM for each query peak region was computed and query peaks with fewer than 1 CPM were eliminated from the analysis. The complete data set comprised a total of 42,095 peaks for Treg dataset, 50,229 peaks for Tconv dataset and 43,421 peaks for gender-matched (n=5 male pairs) dataset as query peak regions for differential accessibility analysis. RUVSeq [32] was used to correct for unwanted variation in the data using negative or *in-silico* empirical control regions. Prior to RUVg normalisation, *in-silico* empirical control regions, i.e., least significantly differential regions, were obtained on a first-pass differential accessibility analysis and the resulting factors believed to have captured the unwanted variation were computed and incorporated into a linear regression model with cohort and pairing information.

Estimation of common, trended, and tagwise dispersion was performed using *estimateDisp()*. Differential accessibility of ATAC-seq peaks between T1D and healthy controls was calculated using the exactTest method, with peaks having a Benjamini-Hochberg false discovery rate (FDR) of less than 0.05 being considered significant. Significantly differentially accessible peaks were assigned either to the nearest transcriptional start sites (TSS) in linear genomic distance using the Bioconductor package ChIPpeakAnno [258], or linked to their long-range gene targets in 3D space through integration with enhancer/promoter connectome data generated from conformation capture technology (described in Section 5.3.3.7).

5.3.3.4 Differential gene expression analysis

All differential expression analyses between T1D and healthy controls was performed in a family-based pairwise fashion to control for confounding variables arising from genetic heterogeneity and environmental components between families.

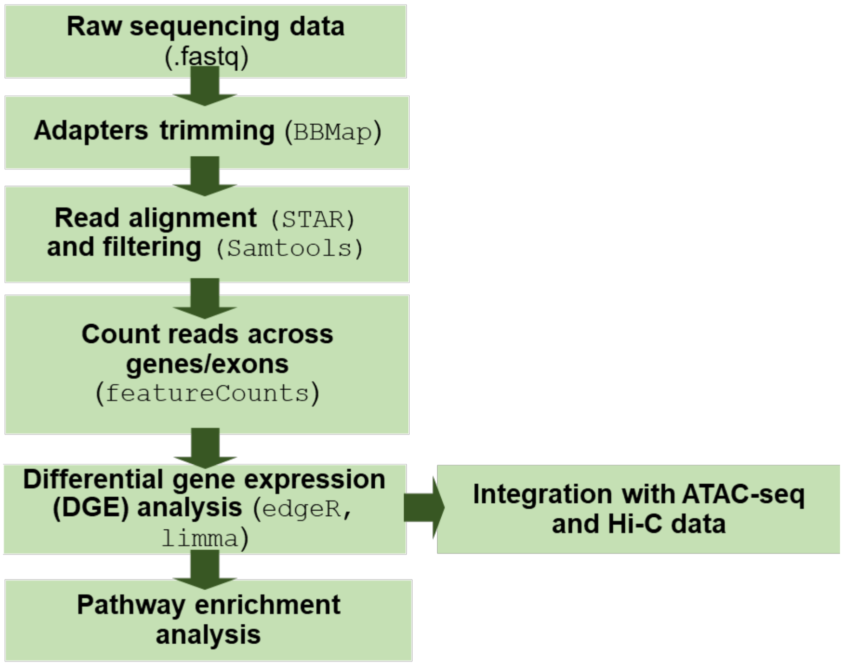

Raw sequencing data (.fastq)
Adapters trimming (BBMap)
Read alignment (STAR) and filtering (Samtools)
Count reads across genes/exons (featureCounts)
Differential gene expression (DGE) analysis (edgeR, limma)
Integration with ATAC-seq and Hi-C data
Pathway enrichment analysis

Figure 5-2 | Flowchart showing bioinformatics pipeline for RNA-seq data processing and differential analysis.

Differential gene expression analysis was performed as described in Section 2.4.2 and Figure 5-2. The sequencing data quality was determined using FastQC v0.11.7 [156]. Reads were aligned to the GRCh38 human genome (GENCODE; hg38) and quantified with GENCODE GRCh38 comprehensive gene annotation (ver. M17) using featureCounts (ver. 1.6.2).

Differential expression analysis was performed using edgeR (ver. 3.22.3). Significantly differential genes were defined as genes having a Benjamini-Hochberg false discovery rate (FDR) less than 0.05 and the estimated absolute $\log_2$(fold change) between T1D and healthy was >1.1.

5.3.3.5 Identification of enriched TF footprints and differential footprinting analysis

In ATAC-seq the binding of TFs to DNA prevents the Tn5 from accessing the otherwise nucleosome-free region, leaving footprints. I used HINT-ATAC [21] to identify TF footprints, perform motif matching, generate binding average profiles and perform differential footprint analysis in my case-control ATAC-seq data as described in Section 2.4.1.3. As I would like to identify alterations in TF binding within chromatin regions altered by epigenetic changes, i.e., regions demonstrating differential DNA accessibility between T1D cohort and healthy controls, I performed differential TF footprinting within case-control differential ATAC-seq peaks (Bonferroni adjusted p-value < 0.05). Lineplots demonstrating ATAC-seq signals around the TF binding consensus motifs were adapted from HINT-ATAC.

5.3.3.6 Enrichment of T1D LD-SNPs in ATAC-seq peaks and TF footprints

As SNPs identified from GWAS genotyping arrays (i.e. index SNPs) serve only as representatives for all the SNPs in the same haplotype block and it is possible that other SNPs in high linkage disequilibrium (LD) with the index SNPs are casual for the disease, I expanded my analyses using LD calculation together with the 1000 Genomes Project reference panels to include SNPs that are in high LD with the risk-associated index SNPs (referred to as “LD-SNPs” hereafter) in my enrichment and association studies. I first compiled a comprehensive list of variants associated with T1D from GWAS catalogue and then identified SNPs in high

LD with the GWAS SNPs using r2>0.8 as a cutoff for a proxy or LD SNP. A list of T1D associated index SNPs was downloaded from the NHGRI-EBI GWAS Catalogue on 14/12/2020 (https://www.ebi.ac.uk/gwas/), from which I selected SNPs if: (a) they demonstrate genome wide significance (p-value $\leq 9 \times 10^{-6}$) and were verified in a peer-reviewed publication and (b) the study was reported as "Type 1 diabetes", "Type 1 diabetes in high risk HLA genotype individuals", "Type 1 diabetes autoantibodies in high risk HLA genotype individuals" or "Fulminant type 1 diabetes". The database reported 342 index SNPs associated with T1D from GWAS studies. For this set of index SNPs, I identified variants in high LD (r2 > 0.8 with maximum allowable deviations for matching SNPs of 1%) using SNPsnap [259] (LD SNPs computed using PLINK ver. 1.07 [260] and 1000 Genomes Project Phase 3 reference panels from European (EUR) population (GBR, FIN, IBS, TSI and CEU). The calculation reported 2,370,237 SNPs where I then filtered out duplicate SNPs and extended 50bp on both sides to a uniform peak size of 100bp using BEDTools (ver. 2.25.0) for downstream enrichment analyses. To identify candidate causal SNPs that are more likely to be associated with T1D, I took advantage of my case-control genomic and epigenomic data and incorporated functional annotation of regulatory elements, such as ATAC-seq peaks and TF footprints differentially accessible between T1D cohort and healthy controls, in the T1D LD variants to prioritise or functionally annotate the T1D risk-associated SNPs located in regulatory regions of the genome (referred to as "refined T1D SNPs" thereafter). These refined T1D SNPs showed functional relevance to T1D and I then statistically assessed the enrichment of these variants in the case-control differential ATAC-seq peaks and TF footprints using permutation test [28], employing the number of overlaps as the evaluation function and more than 5000 random permutations in each test. P-values were calculated as the number of overlaps between the features of interest was higher compared with permuted genomic regions.

5.3.3.7 Integration of ATAC-seq data with Hi-C chromatin conformation capture data

Table 5-1. Resource table showing the chromatin conformation capture (3C) datasets used in this section (Section 5.3.3.7) for annotating the gene targets of regulatory elements differentially accessible between T1D cohort and healthy controls.
B, biological replicate; T, technical replicate.

Source	3C variant	Type of interaction captured	Dataset identifier	Cell type used in my analysis	Marker	Number of replicates
Liu et al. 2020 (Barry Lab) [134]	Hi-C	all interactions	NA	Expanded Treg	$CD4^+$ $CD25^+$	2T
Mumbach et al. 2017 [37]	H3K27ac HiChIP	enhancer connectome	GSE101498	Treg	$CD4^+ CD25^+ CD127^{lo}$	3B1T
				Th17	$CD4^+ CD25^- CD45RA^-$ $CCR6^+ CXCR5^-$	3B2T
Javierre et al. 2016 [6]	Promoter capture Hi-C	promoter interactome	EGAS00001001911	Activated total $CD4^+$ T cells	$CD4^+$	3B

It has been well established that long-range interactions between regulatory elements such as non-coding gene enhancers and promoters play a key role in transcriptional regulation and distal regulatory elements do not always link to the closest gene by linear distance. The majority of the GWAS-identified SNPs associated with complex diseases and traits are located within non-coding regions of the genome [2, 88] and therefore can alter gene expression through spatial interactions that involve distal loci. As the query peak regions of interest I used for conducting differential accessibility analysis between T1D and healthy controls comprised genomic regions corresponding to TSS, super- and typical- enhancers of $CD4^+$ T cells [7, 8], rather than assigning those T1D-associated distal regulatory elements to the nearest TSS in linear distance, I linked them to their putative target gene promoters in the 3D nuclear space by incorporating several published interactome datasets (listed in Table 5-1) generated by chromosome conformation capture (3C) technology (see schematics in Figure 5-3 and Figure 5-4). These 3C datasets include Hi-C contact map [134] generated from Treg cells which captures all chromatin interactions in the nucleus, H3K27ac HiChIP [37] generated from Treg and Th17 which enriches for enhancer interactions, as well as promoter capture Hi-C (pcHi-C) contact map [6] generated from activated CD4$^+$ T cells which enriches for promoter interactions.

In order to identify the gene targets of T1D-associated enhancers captured by my ATAC-seq data as well as to identify novel gnomAD human genetic variation [1] that might play a role in modulating the risk of T1D through long-range gene interactions, I first retrieved high-confidence loops (identified using Juicer pipeline HiCCUPS tool) called from Treg and Th17 cells in the H3K27ac HiChIP dataset from Mumbach, Satpathy [37], which comprised a total of 3,918 Treg loops and 4,241 Th17 anchor loops. Using a promoter capture Hi-C (pcHi-C) contact map generated by Javierre, Burren [6] from activated CD4$^+$ T cells, I extracted significant high-confidence promoter interactions (CHiCAGO scores ≥ 5) for my analyses and that revealed a total of 221,188 interaction loops. From an in-house Treg Hi-C contact map generated by Liu, Sadlon [134], a total of 1,397,506 interaction loops were detected in which I extracted loops that contain TSS, super- and typical- enhancer elements of CD4$^+$ T cells [7, 8] at least at one interaction anchor and extended the anchor length to a final size of 8kb to improve comparability with Treg H3K27ac HiChIP contact map [37] (which has an average contact coverage of 8kb), resulting in 249,794 TSS-/enhancer-containing loops to be used in my analyses. I used Treg Hi-C [134] and Treg H3K27ac HiChIP [37] contact maps to identify putative target genes regulated by the T1D-associated enhancer elements (FDR < 0.05) identified from my case-control Treg ATAC-seq data. Th17 H3K27ac HiChIP [37] and activated CD4$^+$ T cell [6] pcHi-C contact maps were used to annotate the target genes regulated by the T1D-associated enhancer elements (FDR < 0.05) identified from my case-control Tconv ATAC-seq data. The transcriptome changes of the identified target genes were validated by case-control Treg or Tconv RNA-seq. Intersection of gene sets was performed using VennDiagram in R Statistical Software.

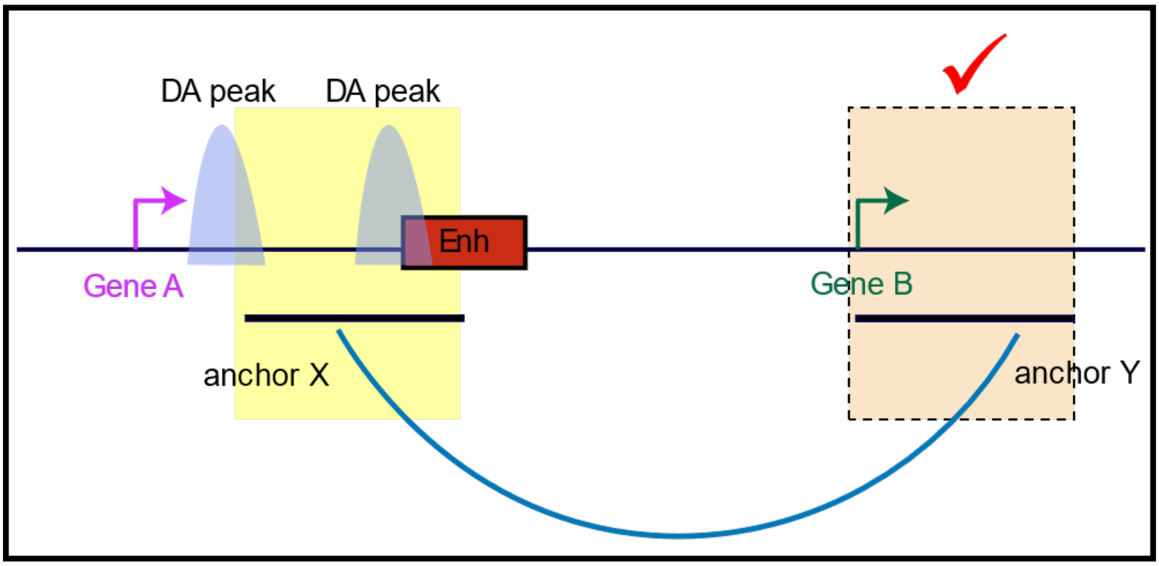


Figure 5-3 | Mapping differentially accessible ATAC-seq peaks of T1D and healthy subjects to the putative gene targets using high-resolution Hi-C contact maps, rather than to the nearest TSS in linear genome distance.
In this schematic, Gene A is the closest gene to the differentially accessible peak, which is also a gene regulatory enhancer element, however, contact map generated from chromosome conformation capture assay (e.g. Hi-C) shows that the enhancer is regulating a distant target gene, Gene B instead of Gene A. Using this approach we identify targets of long-range regulation at genomic regions far apart in the linear genome sequence but are proximate in 3D space. *DA, differential accessible.*

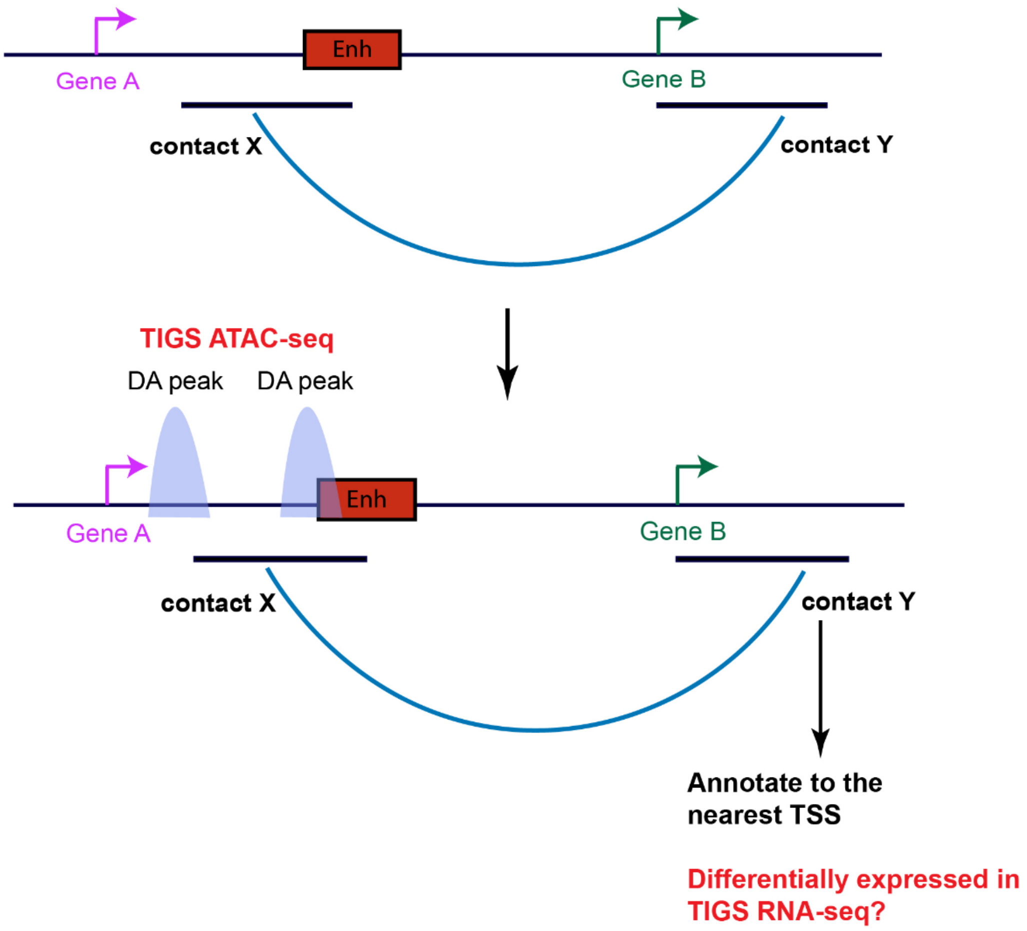


Figure 5-4 | Linking differentially accessible ATAC-seq peaks of T1D and healthy subjects to the putative gene targets using high-resolution Hi-C contact maps and RNA-seq. Hi-C contact maps were used to identify long-range gene targets of differentially accessible regulatory elements between T1D and healthy subjects and changes in transcriptomes were validated by RNA-seq. *DA, differential accessible.*

5.3.3.8 Identification of novel T1D-associated human genetic variation

In order to identify putative candidate functional SNPs that present a higher probability of having an impact on T1D independently of association studies, I sampled 4,655,805 common human variants (MAF > 0.1) from the Genome Aggregation Database (gnomAD) (version 3.0) [available from: https://gnomad.broadinstitute.org/] [1] and incorporated functional annotation of case-control regulatory elements such as differentially accessible genomic regions, TF footprints, T cell enhancer- and promoter-associated interaction loop anchors as an approach to prioritise genetic variants which may be differentially regulated between T1D cohort and healthy controls, termed "novel T1D-associated SNPs". Intersection and visualization of genomic region sets of interest were performed using Intervene [261].

5.3.3.9 Permutation analysis for T1D-associated features of interest

The enrichment of GWAS-identified and novel T1D-associated genetic variants (gnomAD) in various features of interest such as case-control differential ATAC-seq peaks, differential TF footprints, FOXP3 binding sites [11] was statistically assessed using permutation tests [26], employing the number of overlaps as the evaluation function and more than 5000 random permutations in each test. P-values were calculated as the number of overlaps between the features of interest was higher compared with permuted genomic regions.

5.4 Results

5.4.1 Thawing of biobank cohort

During my early work I established an optimised thawing methodology using healthy adult cells which resulted in significantly higher recovery (p=0.008) of live PBMCs compared with unmodified thawing protocol for use on the biobanked clinical samples (Chapter 3; Section 3.4.1). I also successfully established ATAC-seq protocol for use on T cells recovered from frozen PBMC samples. Furthermore, using the optimised thawing and ATAC-seq protocol I showed that the chromatin accessibility landscape of thawed T cells closely recapitulates that of the fresh T cells in healthy adults as described in Chapter 3 (Section 3.4.5).

It is critical to employ the best practices of bioinformatics framework for analysing high throughput sequencing data to ensure accurate interpretation and clinical translation of the data. I therefore benchmarked analysis strategies and established a robust data analysis pipeline for ATAC-seq data - from raw data processing, peak calling (Chapter 4; Section 4.4.2), data normalisation (Chapter 4; Section 4.3.4), to differential accessibility analysis (Chapter 4; Section 4.4.3) as well as inference of TF binding.

With optimal experimental and bioinformatics framework in place I proceeded to apply the methodology on the biobank cohort comprising T1D and sibling-matched healthy controls. My work aimed to profile the chromatin accessibility and transcriptome changes in Tconv and Treg cells recovered from the T1D cohort and sibling-matched healthy controls, using ATAC-seq and RNA-seq. I successfully thawed frozen PBMC samples collected from 15 pairs or families of T1D cohort and sibling-matched healthy controls (Table 5-2), corresponding to a total of 71 vials of cryopreserved PBMCs from the biobank. Multiple timepoints or vials were thawed for some donor samples owing to poor sample viability or recovery of sufficient material for sequencing experiments. Principle component analysis (PCA) of PBMCs yield reveals that the largest proportion of variability in the data was driven by the patient visit or sample processing/biobanking period, indicating batch effects as a potential confounding factor which may bias the inference of differential accessibility or expression in the subsequent sequencing experiments (Figure 5-5; a). Although overall, high correlation between viability and yield of cryopreserved PBMCs was observed in this biobank cohort (Figure 5-5; b), some samples (May – Oct 16; Nov 16 – Jan 17) with high viability demonstrated low recovered yield upon overnight incubation (Figure 5-6), implying donor or processing batch effects as the source of this discrepancy. Nonetheless, low viability PBMC samples were omitted from my study as they generally did not meet the minimum input requirement for sequencing experiment (Figure 5-6) and it has also been reported that poor quality cells such as low viability is a common culprit of poor signal-to-noise ratio in the accessibility and expression data [41, 262].

Although some cryopreserved PBMC samples from the biobank exhibited poor cell viability and/or recovery upon thawing and overnight resting, there was no significant difference between T1D cohort and healthy control samples for those QC parameters (Figure 5-7). The average cell viability was 67.3% (± 32.7%) for T1D samples and 66.8% (± 33.4%) for healthy control samples, whereas the average cell recovery upon overnight incubation was 67.8% (± 27.9%) for T1D samples and 73.0% (± 29.5%) for healthy control samples (Figure 5-7).

Table 5-2. Thawing of cryopreserved PBMCs obtained from T1D biobank cohort. Samples are identified by study ID, timepoint/visit number, biobanking date. Each cryovial contained PBMCs purified from 9mL of peripheral blood obtained from individuals with established T1D or healthy controls.

T, timepoint.

Sample	Cohort	Family/Pairing	PBMC vials thawed
JP23_T1_01/11/16	T1D	1	2
JP73_T4_01/11/16	Control	1	2
JP25_T4_10/10/16	Control	2	2
JP26_T1_04/08/18	T1D	2	1
JP27_T2_20/01/16	T1D	3	2
JP28_T1_19/08/15	Control	3	1
JP29_T2_09/03/16	T1D	4	1
JP30_T1_ 08/09/15	Control	4	1
JP32_T2_22/03/16	Control	5	2
JP33_T2_22/03/16	T1D	5	2
JP35_T2_08/03/16	T1D	6	1
JP35_T3_30/08/16	T1D	6	1
JP36_T1_28/09/15	Control	6	1
JP36_T2_08/03/16	Control	6	1
JP37 T1 - 29/09/15	T1D	7	2
JP38 T2 - 27/04/16	Control	7	2
JP38 T3 - 12/09/16	Control	7	2
JP43_T1_17/11/15	T1D	8	2
JP43_T2_23/05/16	T1D	8	2
JP60_T1_23/05/16	Control	8	2
JP60_T2_06/09/16	Control	8	2
JP50_T1_18/02/16	T1D	9	2
JP50_T2_18/08/16	T1D	9	2
JP51_T1_18/02/16	Control	9	2
JP53 T1 - 19/04/16	Control	10	2
JP53 T2 - 04/10/16	Control	10	2
JP56 T2 - 10/10/16	T1D	10	2
JP56_T1_21/04/16	T1D	10	1
JP61_T1_08/06/16	Control	11	2
JP61_T2_10/10/16	Control	11	2
JP62_T2_10/10/16	T1D	11	2
JP67_T1_26/07/16	T1D	12	2

JP67_T2_24/10/16	T1D	12	2
JP68_T1_26/07/16	Control	12	1
JP68_T2_24/10/16	Control	12	2
JP70_T1_10/08/16	T1D	13	2
JP71_T2_25/10/16	Control	13	2
JP74_T1_18/01/17	Control	14	2
JP75_T1_18/01/17	T1D	14	2
JP8_T3_21/04/15	T1D	15	2
JP9_T5_26/04/16	Control	15	1
Sum		***15 families***	***71***

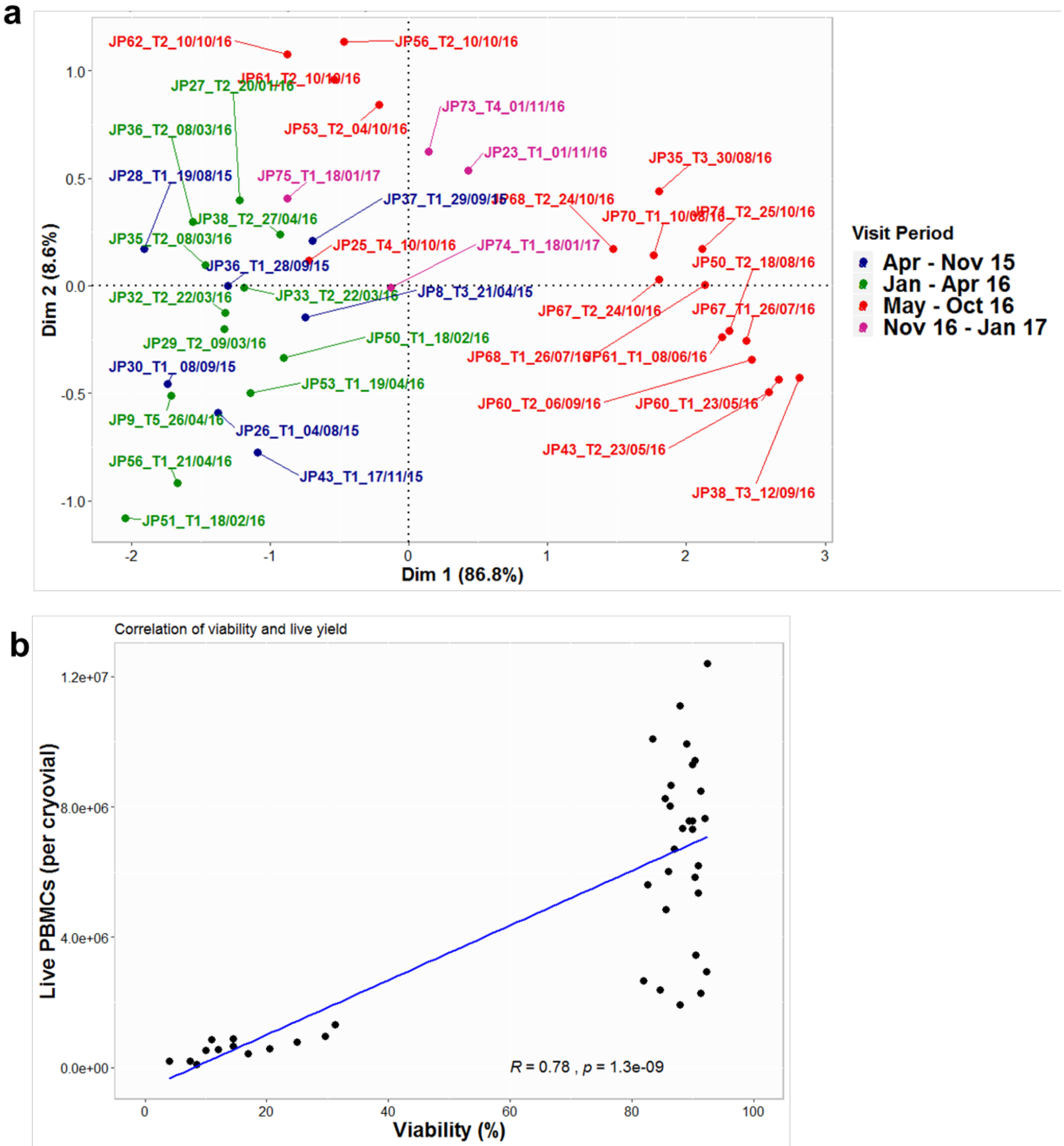


Figure 5-5 | Principal component analysis (PCA) plot of absolute yield of viable PBMCs recovered from the T1D biobank (a) and correlation between PBMCs viability and yield recovered per cryovial (b).

Samples are identified by study ID, timepoint/visit number, biobanking date. Each cryovial contained PBMCs purified from 9mL of peripheral blood obtained from individuals with established T1D or healthy controls.

T, timepoint.

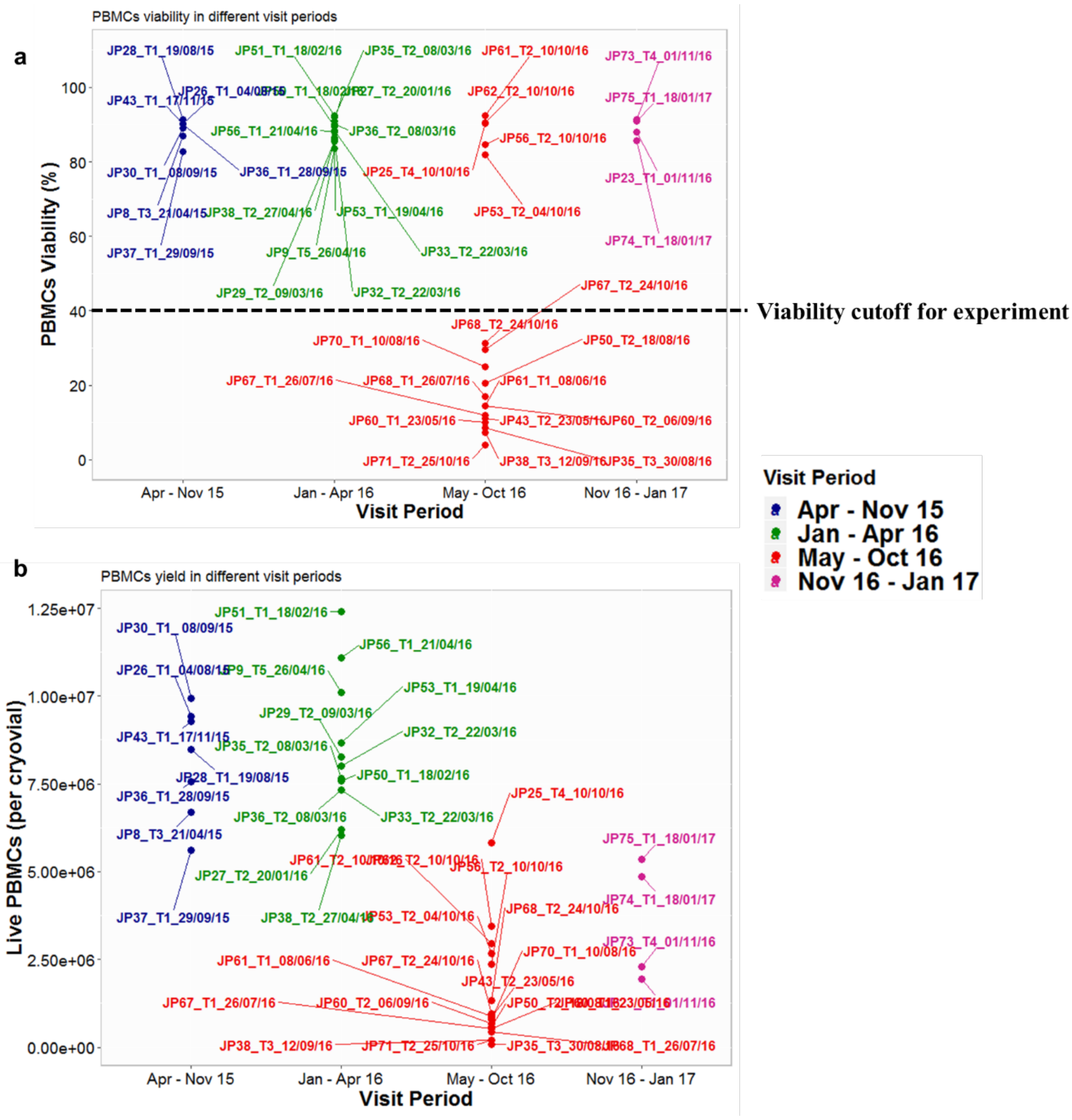
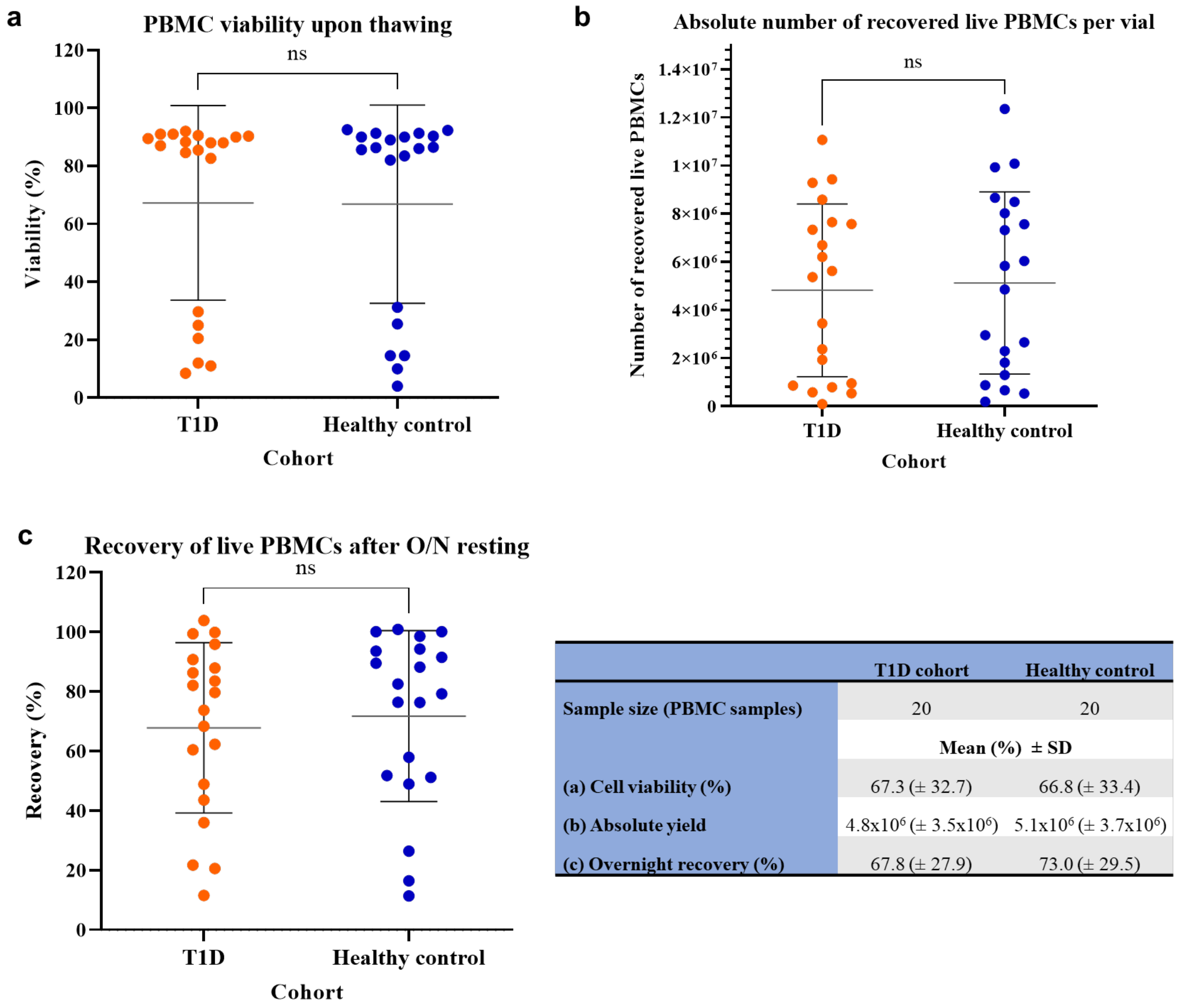


Figure 5-6 | PBMCs viability (a) and viable yield recovered (b) from the T1D biobank, arranged according to patient visit period or sample processing date.

Samples are identified by study ID, timepoint/visit number, biobanking date. Black dotted line indicates the quality control threshold on cell viability used for subsequent sequencing experiments.

T, timepoint.



	T1D cohort	Healthy control
Sample size (PBMC samples)	20	20
	Mean (%) ± SD	
(a) Cell viability (%)	67.3 (± 32.7)	66.8 (± 33.4)
(b) Absolute yield	4.8x10^6 (± 3.5x10^6)	5.1x10^6 (± 3.7x10^6)
(c) Overnight recovery (%)	67.8 (± 27.9)	73.0 (± 29.5)

Figure 5-7 | Viability and recovery of cryopreserved PBMCs from the biobank cohort upon thawing.

Cryopreserved PBMCs from the T1D biobank cohort were thawed according to optimised thawing protocol (Section 5.3.2). Upon overnight resting, the viability (a), absolute number of live (b) and recovery (c) of thawed PBMCs were measured.

Results are presented as mean with standard deviation involving 20 individuals with established T1D and their respective healthy siblings from a total of 15 families (Table 5-2). Statistical comparisons were performed with multiple paired T-test with p values ≤ 0.05 being considered significant.

Absolute number of live PBMCs was collected immediately after thawing (data not shown) and after overnight incubation (b). Recovery (%) (c) was measured as the number of live PBMCs remained after overnight culture expressed as a percentage.

Using a panel of established cell surface markers, I purified viable Treg cells (CD4$^+$ CD25^{hi} CD127^{lo}) and Tconv cells (CD4$^+$ CD25^{lo} CD127^{hi}) from the thawed biobanked PBMCs derived from peripheral blood of 26 subjects, including 13 T1D patients and 13 sibling-matched healthy controls (Figure 5-8; Section 2.1.2).

There was no significant difference in the proportion and number of Treg (Figure 5-9; a, b) and Tconv (Figure 5-9; c, d) cells recovered between the T1D cohort and healthy control PBMC samples. The mean percentage of CD4$^+$ Treg population from T1D samples was 6.04% (± 1.63%) per donor sample and 6.00% (± 1.75%) for healthy control samples, with an average yield of 1.78x10^5 (± 1.03x10^5) for T1D and 1.54x10^5 (± 8.97x10^4) for healthy samples. As for representation of CD4$^+$ Tconv population, T1D samples had a mean percentage of 85.1% (± 2.41%) with an average yield of 2.45x10^6 (± 1.31x10^6), whereas healthy samples had a mean percentage of 84.3% (± 3.55%) with an average yield of 2.06x10^6 (± 9.39x10^5) (Table 5-3; Figure 5-9).

As Treg cells represent a rare population of the T cell pool, and this study was constrained by the volume of the paediatric sample and low sample yield from some biobanked samples, the number of Treg cells recovered from some biobanked samples were not sufficient for both ATAC-seq and RNA-seq experiments under both resting and activation states (Table 5-2, Table 7-2). For samples with insufficient recovery of Treg cells for all of the proposed experiments, priority was placed on the ATAC-seq in activated state (i.e. stim ATAC-seq), as study has revealed the importance of probing stimulation-responsive functional regions in immune cells including Tconv and Treg cells as they showed enrichment of autoimmune risk variants [45, 136]. RNA-seq libraries were then prepared using the RNA material recovered from the supernatant fractions of ATAC-seq lysis reaction (i.e. stim ATAC-SN) (as described in Section 3.3.4) instead of a separate pool of whole cells. Samples with sufficient Treg cells recovered were also processed for library generation of resting ATAC-seq, resting RNA-seq from ATAC-

SN (resting ATAC-SN) as well as activated whole RNA-seq (stim whole RNA-seq) (Table 5-2). Sufficient Tconv cells were recovered from all biobanked samples and thus they were processed for all 6 types of library generation - stim ATAC-seq, resting ATAC-seq, stim whole RNA-seq, stim ATAC-SN RNA-seq, resting whole RNA-seq and resting ATAC-SN RNA-seq. A total of 182 ATAC-seq and RNA-seq libraries were made in this study and 95 libraries (48 ATAC-seq and 47 RNA-seq) generated from stimulated Tconv and Treg cells from 12 pairs of T1D subjects and sibling-matched healthy controls were sequenced, selected based on library QC standards and availability of complete sibling-matched libraries for both cell types (Tconv and Treg), cell context (stimulation) and platforms (ATAC-seq and RNA-seq) (Table 7-1). Table 5-4 summarises demographic and other reported information of the biobank cohort involved in this study [age = 11 ± 3.4 years; 62% male; 58% gender-matched pairing (71% males); 50% HLA-matched pairing], collected from 12 pairs of T1D and sibling-matched healthy controls that passed cell viability (>90%) upon thawing and minimum recovery (≥ $8x10^6$ total live PBMCs) criteria.

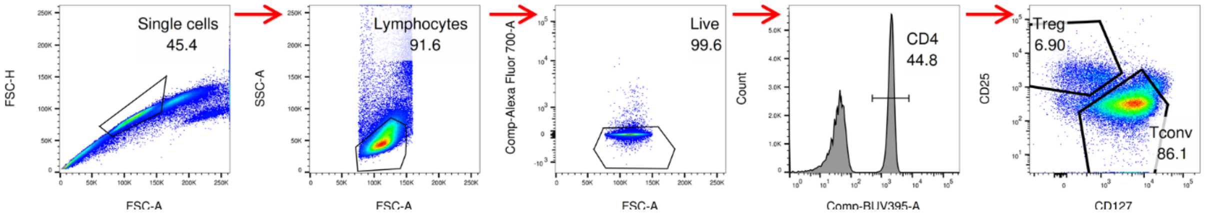


Figure 5-8 | Representative FACS sorting plots of Tconv and Treg cells from biobanked PBMC samples obtained from T1D/healthy cohort.
Gating strategy for Tconv and Treg cells from peripheral blood mononuclear cells (PBMCs), and representative FACS plots are shown.

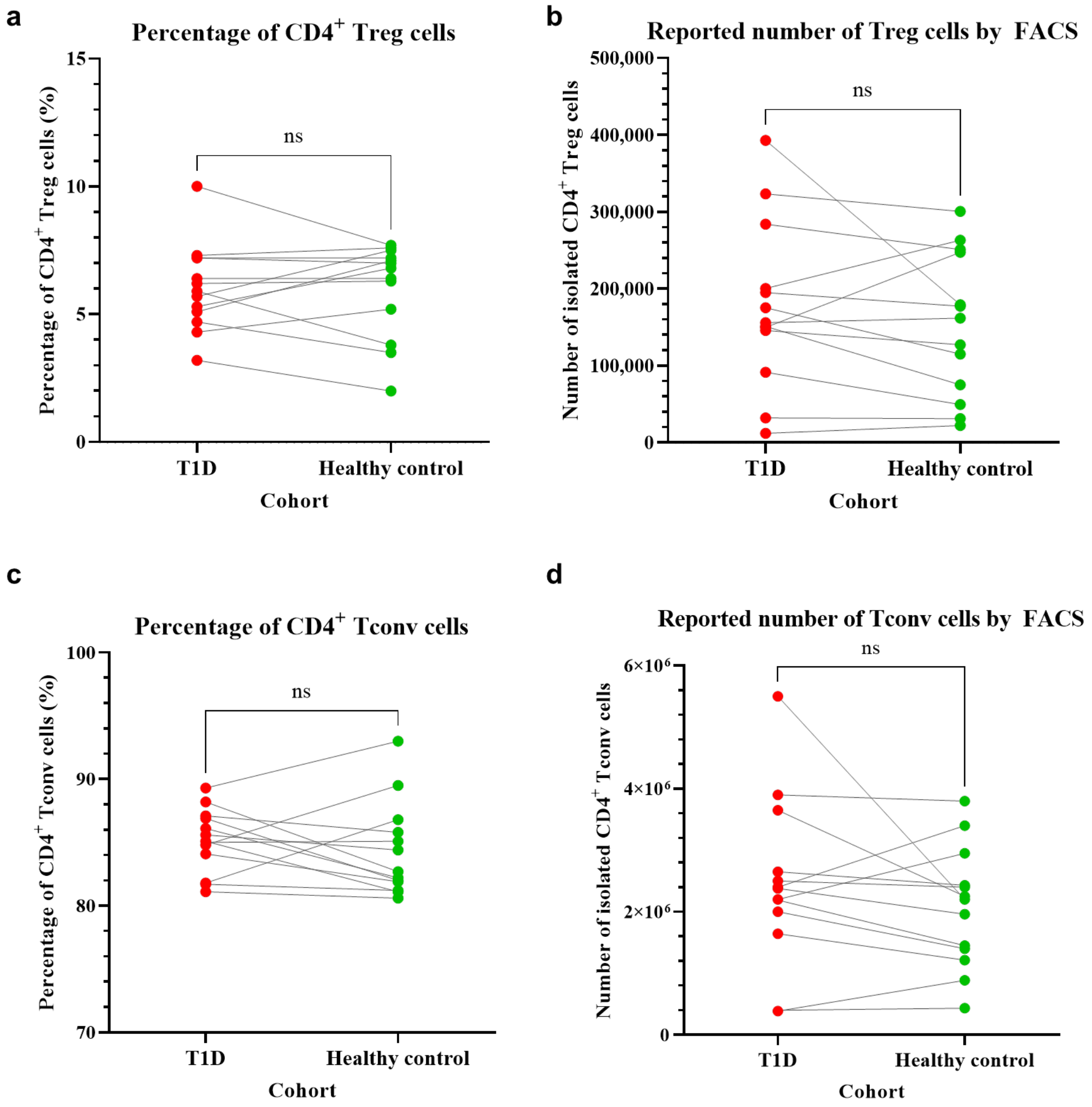


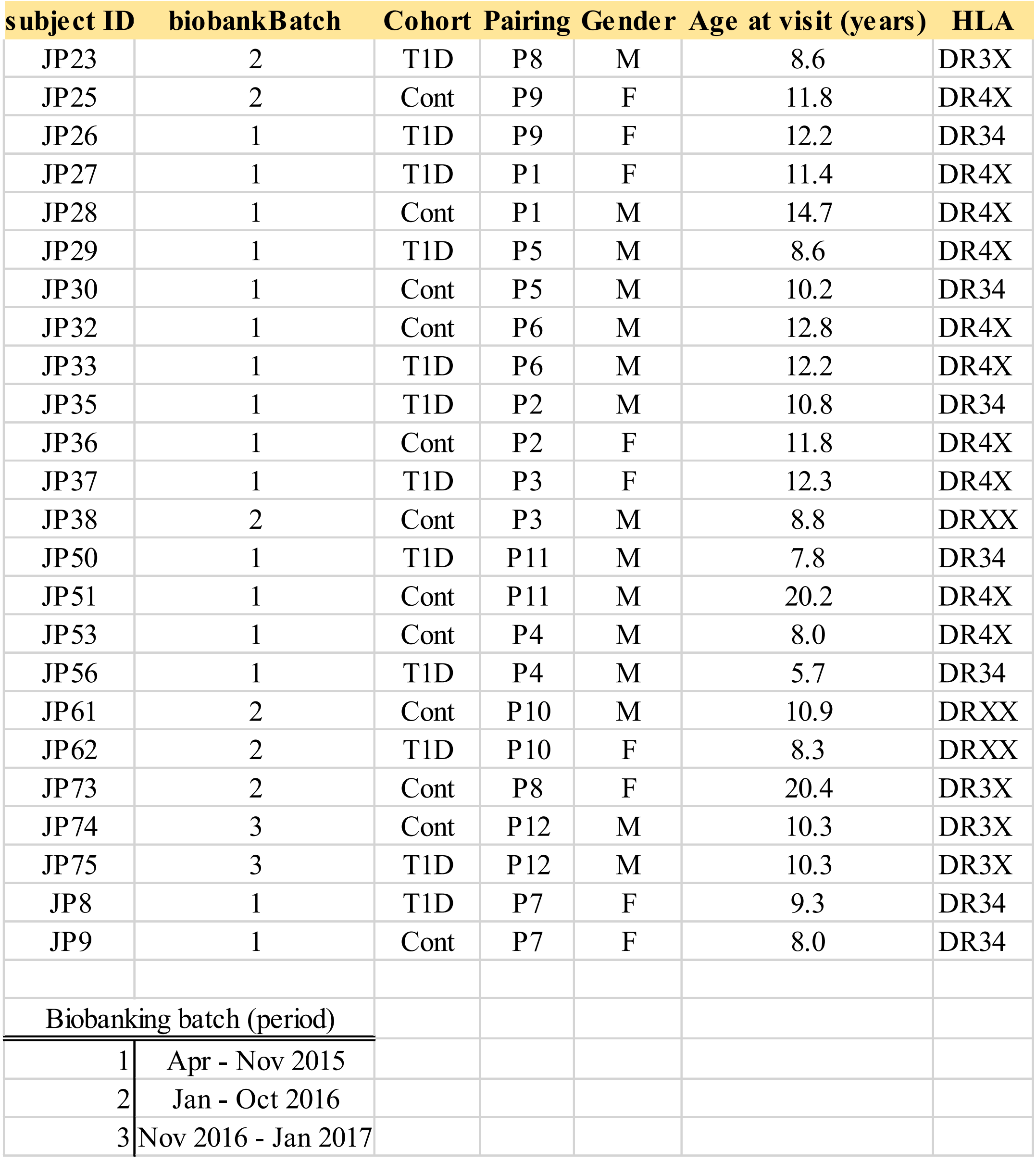
Figure 5-9 | Percentage (a, c) and number (b, d) of recovered Tconv and Treg cells per donor sample from the T1D case-control biobank cohort.
Thawed PBMCs from the case-control biobank cohort were fluorescently labelled with anti-CD4, anti-CD25, anti-CD127 and viability dye for purification of Tconv and Treg cells. The proportion and number of recovered Tconv or Treg cells were determined by FACS analysis.

Results are presented as mean with standard deviation involving 26 individuals with established T1D and their respective healthy siblings from a total of 13 families (Table 5-4). Case-control paired samples are connected by lines. Statistical comparisons were performed with multiple paired T-test with p values ≤ 0.05 being considered significant.

Table 5-3. Statistics of recovered Tconv and Treg cells per PBMC sample from the T1D case-control biobank cohort (see Figure 5-9).
Results are presented as mean with standard deviation involving 26 individuals with established T1D and their respective healthy siblings from a total of 13 families.

	T1D cohort	**Healthy control**
Sample size	13	13
	Mean (%) ± SD	
CD4$^+$ Treg percentage (%)	6.04 (± 1.63)	6.00 (± 1.75)
Recovered Treg number	1.78×10^5 (± 1.03×10^5)	1.54×10^5 (± 8.97×10^4)
CD4$^+$ Tconv percentage (%)	85.1 (± 2.41)	84.3 (± 3.55)
Recovered Tconv number	2.45×10^6 (± 1.31×10^6)	2.06×10^6 (± 9.39×10^5)

Table 5-4. Demographic and biobanking information of ATAC- and RNA-sequenced T1D and sibling-matched healthy subjects.
P, pairing.



subject ID	biobankBatch	Cohort	Pairing	Gender	Age at visit (years)	HLA
JP23	2	T1D	P8	M	8.6	DR3X
JP25	2	Cont	P9	F	11.8	DR4X
JP26	1	T1D	P9	F	12.2	DR34
JP27	1	T1D	P1	F	11.4	DR4X
JP28	1	Cont	P1	M	14.7	DR4X
JP29	1	T1D	P5	M	8.6	DR4X
JP30	1	Cont	P5	M	10.2	DR34
JP32	1	Cont	P6	M	12.8	DR4X
JP33	1	T1D	P6	M	12.2	DR4X
JP35	1	T1D	P2	M	10.8	DR34
JP36	1	Cont	P2	F	11.8	DR4X
JP37	1	T1D	P3	F	12.3	DR4X
JP38	2	Cont	P3	M	8.8	DRXX
JP50	1	T1D	P11	M	7.8	DR34
JP51	1	Cont	P11	M	20.2	DR4X
JP53	1	Cont	P4	M	8.0	DR4X
JP56	1	T1D	P4	M	5.7	DR34
JP61	2	Cont	P10	M	10.9	DRXX
JP62	2	T1D	P10	F	8.3	DRXX
JP73	2	Cont	P8	F	20.4	DR3X
JP74	3	Cont	P12	M	10.3	DR3X
JP75	3	T1D	P12	M	10.3	DR3X
JP8	1	T1D	P7	F	9.3	DR34
JP9	1	Cont	P7	F	8.0	DR34

Biobanking batch (period)	
1	Apr - Nov 2015
2	Jan - Oct 2016
3	Nov 2016 - Jan 2017

5.4.2 ATAC-seq

5.4.2.1 Libraries quality

Although Omni ATAC-seq protocol[41] recommends input requirement of 50,000 cells, it has been reported that ATAC-seq is robust to minor variations in cell number and original publication by Buenrostro, Giresi [25] demonstrated sensitivity of this technology with as few as 500 human nuclei. As Treg represent a rare cell population I was not able to recover 50,000 Treg cells from most of the samples following two-day stimulation culture for sequencing experiments (Table 7-2). As a result of that, Treg input number ranging from 11,000 to 50,000 was used in the ATAC-seq library construction, with an average input cell number of 39,125 (± 12,742) (Table 7-2).

The fragment size distribution of the amplified Tconv and Treg ATAC-seq libraries generated from the biobank cohort demonstrated a distinct, characteristic DNA laddering pattern with a periodicity of 200bp corresponding to integer multiples of phased nucleosomal protection, as revealed from the representative virtual gels (Figure 7-20; Figure 7-21), indicating a result of optimal transposition reaction. Dual size selection was performed on ATAC-seq libraries to enrich for DNA fragments between 100-700bp to eliminate contaminating adapter/primer dimers (<100bp), partial library constructs as well as confounding long fragments (Figure 7-20, Figure 7-21).

5.4.2.2 Exploratory analysis

48 ATAC-seq libraries which were comprised of activated Tconv and Treg cells generated from 12 T1D and sibling-matched healthy control samples, were sequenced on three lanes of Illumina Hiseq flowcell (GENEWIZ) with each lane covering 16 libraries. The average yield per ATAC-seq library in each sequencing lane was between 27,992,721 (± 3,819,100,12) to 29,866,766 (± 6,599,435) reads (Table 5-5; Figure 5-10, Figure 7-22, Figure 7-23). Overall, the sequenced ATAC-seq libraries showed good sequencing quality reflected in the per base sequence quality (Figure 7-24, Figure 7-25, Figure 7-26), duplication levels (Figure 7-31, Figure 7-32, Figure 7-33), post adapter trimming quality (Figure 7-27, Figure 7-28, Figure 7-29, Figure 7-30), library complexity (Figure 7-34, Figure 7-35, Figure 7-36) and insert size distribution (Figure 7-37, Figure 7-38, Figure 7-39). Paired case-control samples JP23 and JP73 demonstrated lower library complexity and higher enrichment of short fragments compared with other samples, which could be attributed to low input cell number (Table 7-2). GC content of case-control Treg ATAC-seq libraries showed that JP25 had a greater coverage spike in GC content at 60% compared with other libraries (Figure 7-46, Figure 7-47). It has been shown that GC content bias, which is commonly introduced during PCR amplification of DNA fragments in library prep [263], can confound the biological signals of interest and lead to false-positive peak calls [264]. Hence, conditional-quantile normalisation [40] was performed to correct for GC-content bias. This adds a sample-level offset for each peak count which takes into account any systemic bias. After conditional quantile normalisation, a second PCA analysis was performed in order to check for improvement in group separation. However, the normalisation appeared to contribute minimally to the sample separation at this level (Figure 7-50). The distribution of GC content for case-control Tconv ATAC-seq libraries showed uniform read coverage between samples (Figure 7-51).

Peaks corresponding to accessible genomic regions with statistically strong enrichment of sequencing reads were called from pooled replicates (n=12) for each cohort group in both cell types (Figure 5-11), using MACS2 ver. 2.1.2 [159] and optimised peak calling algorithms as evaluated in Section 4.4.2. ATAC-seq peaks called from both cohort groups for both cell types showed enrichment of intronic and intergenic genomic regions and there was no significant variation in feature distribution between T1D cohort and healthy controls (Figure 5-11).

Table 5-5. Summary of sequencing output yield for ATAC-seq experiment of T1D and healthy control samples.
48 ATAC-seq libraries, which comprised Tconv and Treg libraries generated from 12 T1D and sibling-matched healthy control samples, were sequenced on three lanes of Hiseq flowcell, each occupying 16 libraries. To avoid lane effects between comparison groups, 4 pairs of complete combination containing T1D Treg, Control Treg, T1D Tconv and Control Tconv were multiplexed and sequenced on the same lane.

Sequencing lane	Number of libraries	Total reads	Average yield per library (± SD)
L002	16	480,941,107	29,601,987 (±12,505,202)
L003	16	485,037,597	29,866,766 (±6,599,435)
L004	16	489,358,130	27,992,721 (±3,819,100)

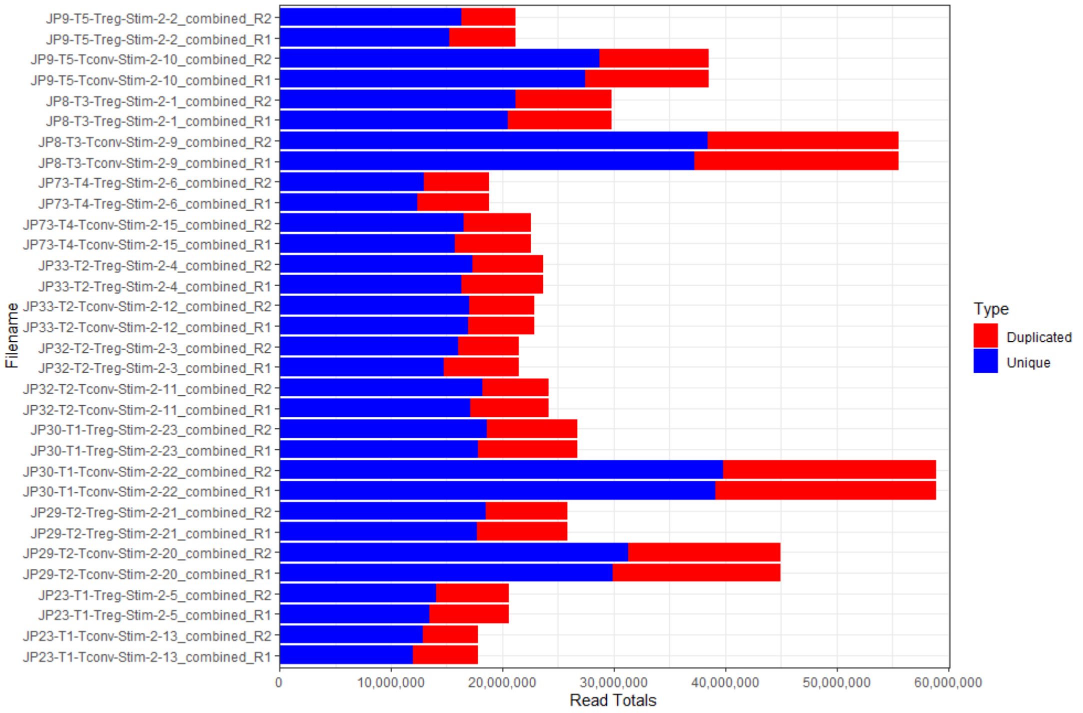


Figure 5-10 | The total number of paired-end reads for ATAC-seq libraries sequenced on Lane 002.

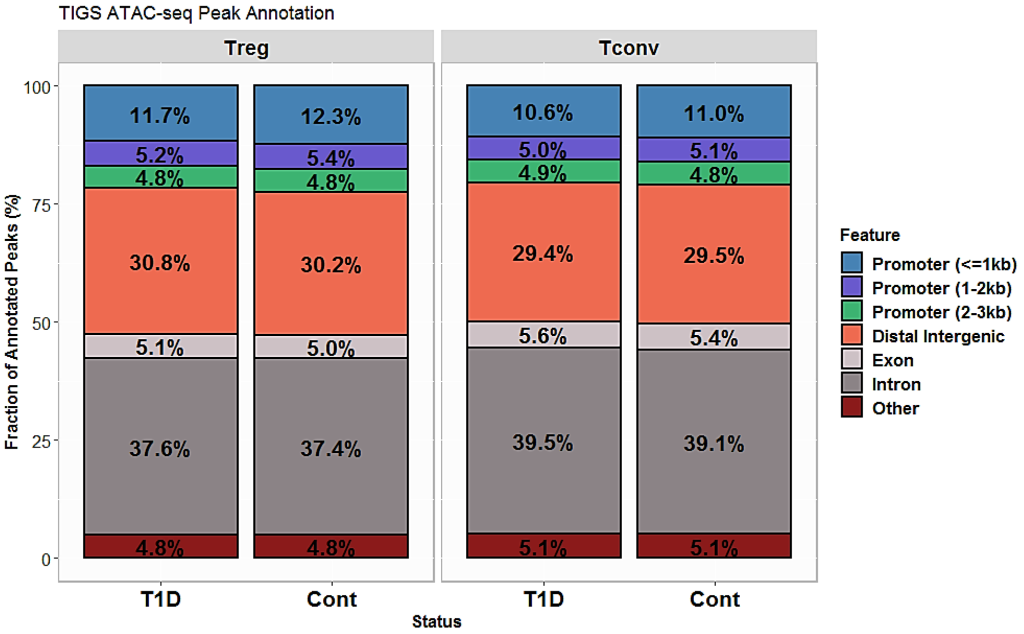


Figure 5-11 | Genomic annotation for ATAC-seq peaks called from Treg and Tconv cells from T1D or healthy control subjects (p-value > 0.9).

ATAC-seq datasets were generated from stimulated Treg and Tconv cells purified from 12 T1D and healthy control samples (see Section 5.3) and peaks were called from mapped reads pooled from all replicates within the same comparison group.

Statistical comparisons between T1D and control group for both cell types were performed with multiple paired T-test with p values ≤ 0.05 being considered significant.

Others, 3UTR , miRNA , ncRNA , TTS , pseudo , 5UTR , snoRNA , snRNA and rRNA.

In order to examine heterogeneity in the case control ATAC-seq data, I performed multidimensional scaling (MDS) and principal component analysis (PCA) on the ATAC-seq peak counts across samples. The first dimension captures the feature that best separates the samples and explains the largest proportion of variation in the data, with subsequent dimensions having a smaller effect and being orthogonal to the dimensions before them. This study involved mixed-gender case-control groups and initial data exploratory analyses showed that the paired samples did not cluster according to the primary condition of interest (cases vs controls) but gender (Figure 5-12), revealing gender as the largest proportion of variation in the data, suggesting it is a confounding variable that may obscure the biological variable of interest. It is well established that genome-wide high throughput sequencing is riddled with batch effects and other forms of unwanted variation can significantly compromise the accuracy of statistical inference in genomic experiments [265, 266]. These sources of artifacts must be normalised, corrected, modelled or removed from the data to accurately measure the biological signal of interest and statistical inference. To assess general patterns and identify additional confounders that may potentially bias the locus accessibility and gene expression measures in the data, I compiled a list of biological and technical variables (Table 7-3; methods in Section 5.3.3.2) for the cohort study and measured each of their correlation with the principal components. The experimental design of this study involved multiple factors and it is crucial to examine each variable over several dimensions in order to identify and adjust for any biases in the data before conducting differential analysis. As expected, inspection of Tconv and Treg peak-level counts revealed that cell type was the source of largest proportion of variation in the data, as indicated by highest correlation with principle component 1, followed by GC content and gender (Figure 7-52, Figure 7-53, Figure 7-54). Inspection of Treg or Tconv ATAC-seq dataset independently revealed GC variability, family pairing (genetic factors), gender, patient visit/biobanking period, fraction of reads in peaks and sample HLA genotypes as some of the confounders in the data showing correlation with the leading principal components (Figure 7-40, Figure 7-41, Figure 7-42, Figure 7-43, Figure 7-44, Figure 7-45).

The exploratory analyses showed that these case-control ATAC-seq datasets suffered from confounder effects. To mitigate the problems of unwanted variation driving either biological or technical bias in the data, either biological or technical in origin, RUV (Remove Unwanted Variation) [32], which provides offset term(s) at the sample level, was performed to estimate and adjust for those confounder effects by performing factor analysis using a set of in silico empirical control genes (i.e. least significantly differential peaks between cases and controls based on a first-pass differential analysis) for Treg and Tconv ATAC-seq samples. Owing to gender bias in the data resulting from disproportionate representation of males and females in this study design, peaks mapped to chromosome X and Y were also excluded from downstream analyses. Factor analysis was performed on those negative control regions using RUVg ([32] and the resulting factors were incorporated into a linear regression model for differential accessibility analysis using edgeR [38]. After data normalisation improvement in group separation between cases and controls was observed (Figure 5-13, Figure 5-14).

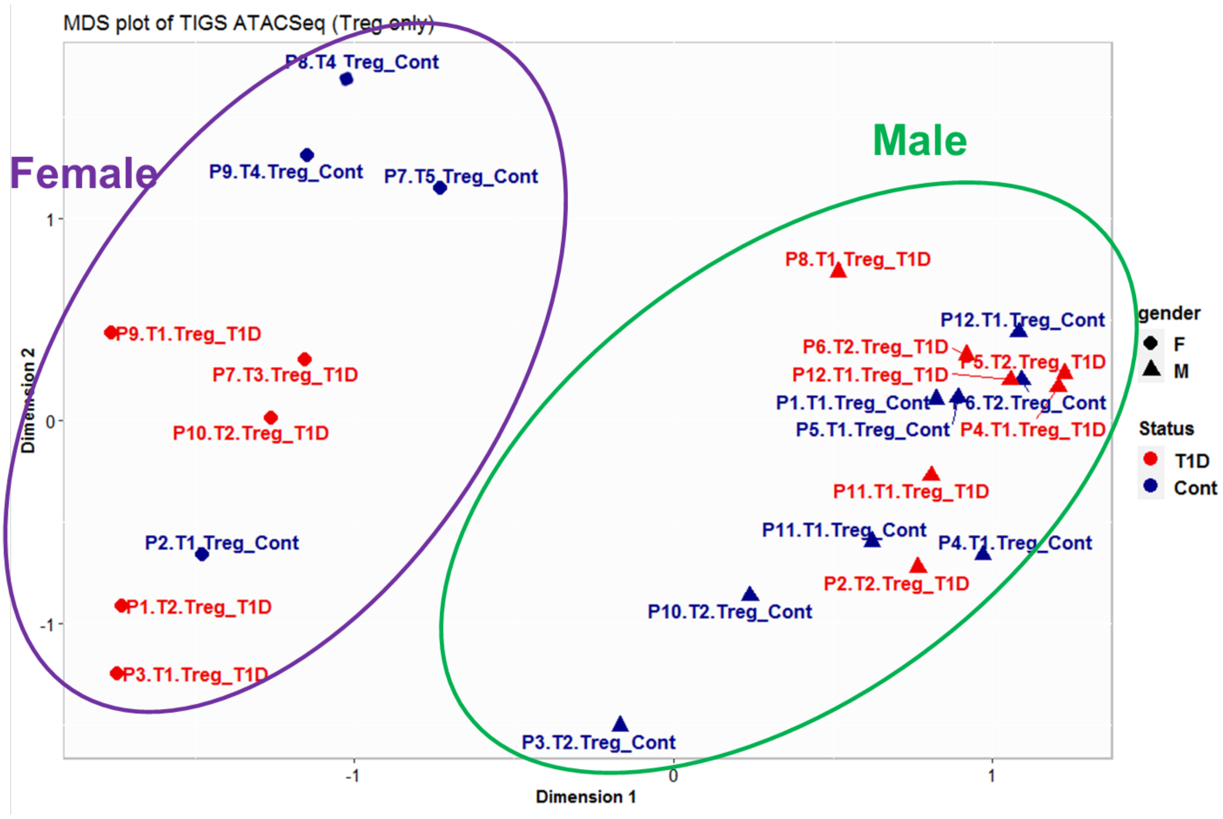


Figure 5-12 | | Multidimensional scaling (MDS) plot of counts in peaks over dimensions 1 and 2 for 24 Treg ATAC-seq libraries generated from 12 T1D and healthy control samples. *P, pairing; T, timepoint of blood draw; Cont, control.*

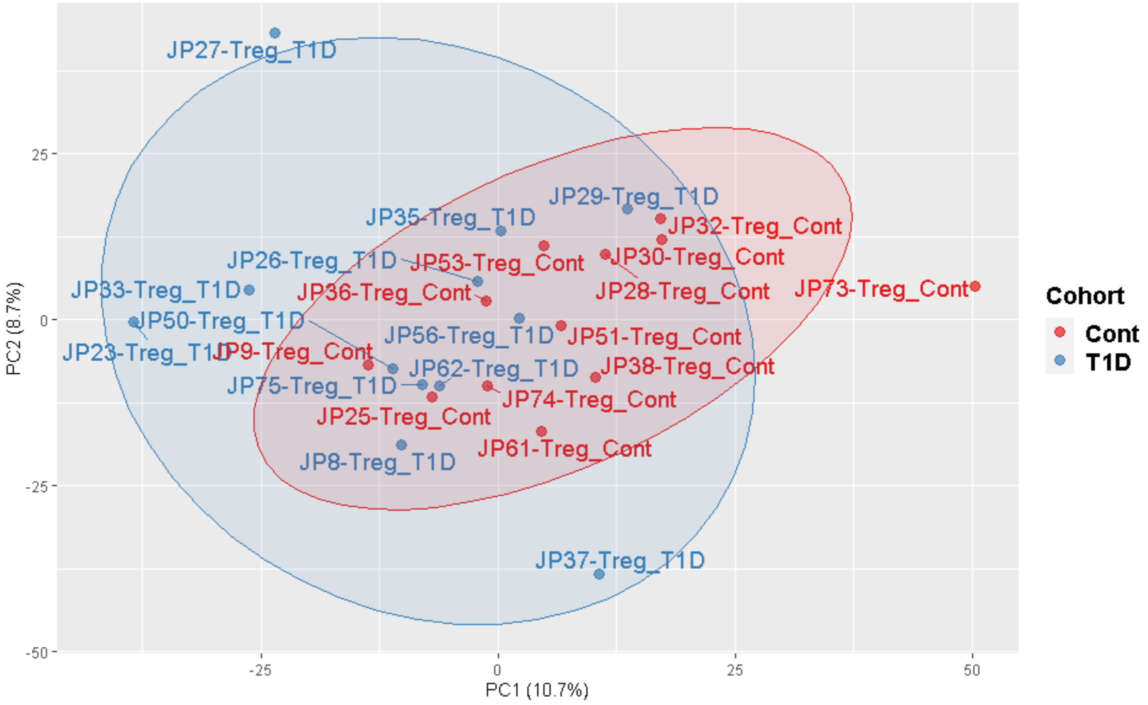


Figure 5-13 | Principal component analysis (PCA) plot of normalised counts in peaks over first two components for Treg ATAC-seq libraries generated from 12 T1D and sibling-matched healthy control samples.

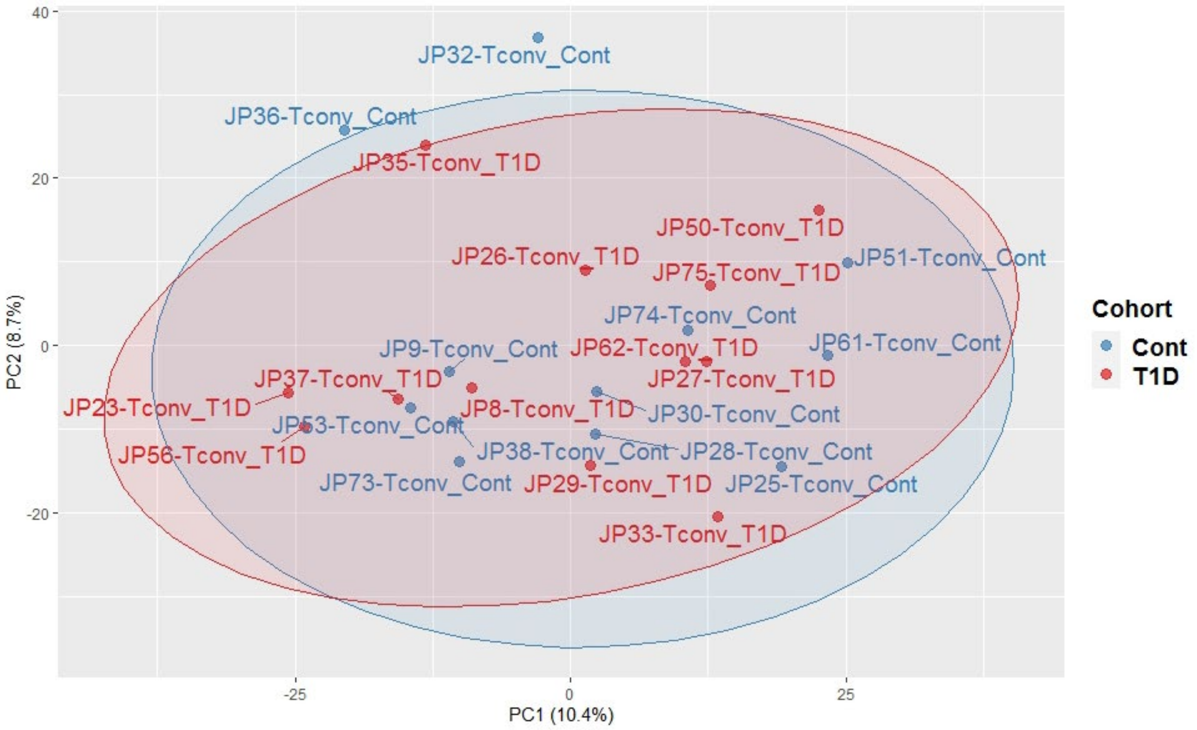


Figure 5-14 | Principal component analysis (PCA) plot of normalised counts in peaks over first two components for Tconv ATAC-seq libraries generated from 12 T1D and sibling-matched healthy control samples.

5.4.2.3 Case-control differential accessibility analysis

MA plots were used to assess the data normalisation (RUVg) performance and identify global accessibility patterns of differential analysis measurements between T1D and healthy controls in Tconv and Treg ATAC-seq, where each query peak region is quantified by the difference in read signal between the two groups as the *y-axis* and signal abundance as the *x-axis*. An upward or downward shift in the accessibility distribution could either indicate a global effect or technical bias [230]. MA plots for case-control differential peaks showed a symmetrical accessibility distribution along the horizontal axis after data adjustment using RUVg estimates and trimmed mean of M values (TMM) (optimised normalisation approach in Section 4.3.4), indicating the normalisation algorithms did not increase bias or variance of the data (Treg: Figure 5-15, Figure 5-17, Figure 7-58; Tconv: Figure 5-21). Differential DNA accessibility between Tconv and Treg cells from healthy or T1D cohort identified loci annotated to some of the well-defined Tconv/Treg gene signatures such as FOXP3, CTLA4, GATA3 and SATB1 (Figure 4-35, Figure 4-38, Figure 7-55, Figure 7-56), demonstrating data quality assurance and robustness of algorithms in identifying differentially accessible genomic regions.

Using Treg case-control ATAC-seq data, I then performed differential accessibility (DA) analyses using different sets of input query peak regions, such as all top 100,000 filtered peaks, filtered peaks (top 100,000) intersecting TSS, super- and typical- enhancers of CD4$^+$ T/Treg cells [7, 8], as well as gender-matched peaks (for 5 pairs of male case-control samples) (as described in Section 5.3.3.3). DA analyses using different peak sets of genomic features of interest as query input in Treg dataset resulted in varying results in accessibility measures and statistical inference (Figure 5-16, Figure 5-18, Figure 7-59). For instance, DA analysis using top 100,000 peaks intersecting TSS, super- and typical- enhancers of Treg cells [7, 8] identified 675 loci differentially regulated between T1D and healthy controls (Benjamin Hochberg FDR < 0.05) (Figure 5-16), whereas DA analysis using all top 100,000 peaks identified 1,822 loci

(Figure 5-18). A small subset of those differential loci were previously known to be associated with T1D [2, 10] (annotation based on linear genomic distance), such as CTLA4, ICOS and BACH2 (Figure 5-16, Figure 5-18; highlighted in blue). 15-23% of the identified DA peaks were bound by FOXP3, the master TF of Treg cells. Differential loci identified using top 100,000 peaks intersecting TSS, super- and typical- enhancer regions (n=675 peaks) showed higher enrichment of promoter features but slightly lower enrichment of distal intergenic and intronic regions compared with DA output conducted using all top 100,000 peaks (n=1,822 peaks), though there was no significant difference in feature distribution between the two (Figure 5-19). Broadcasting of representative ATAC-seq profiles generated from a case-control matched pair at one of the T1D associated locus, *IL2RA*, showed that the accessible chromatin regions were enriched for T1D associated genetic risk variants, located within T-cell specific regulatory enhancers and stimulation-responsive elements (Figure 5-20). The accessibility signal from the T1D sample at this locus appeared lower than the healthy counterpart, however, owing to high heterogeneity of the data this locus was not identified as a statistically significant differential region (Figure 5-20).

As initial exploratory analyses revealed gender as a confounding factor in this dataset, I attempted to probe differential chromatin accessibility only within the gender-matched case-control pairs (n=5 male pairs) (Figure 7-57), having the understanding that this analysis design may be compromised by sample size and statistical power limitations. Nonetheless, the gender-matched DA analysis identified 2,290 ATAC-seq peaks differentially regulated between 5 pairs of T1D patients and their healthy siblings (FDR < 0.05), including FOXP3, IL2RA and IL10RA which were important regulators of Treg development and suppressive phenotypes (Figure 7-58, Figure 7-59). 24% of the differential ATAC-seq peaks were annotated to loci previously known to be associated with T1D [2, 10]. 350 (51.9%) of differential ATAC-peaks identified from 12 case-control pairs were captured in the gender-matched DA analysis (Figure 7-60).

DA analysis of Tconv ATAC-seq (n=12 matched pairs) identified 168 upregulated and 886 downregulated peaks (Figure 5-21, Figure 5-22). As observed in Treg dataset, a small subset of the differentially regulated genomic regions in the Tconv cells were linked to known T1D-associated risk loci, including BACH2, CTLA4 and CD226.

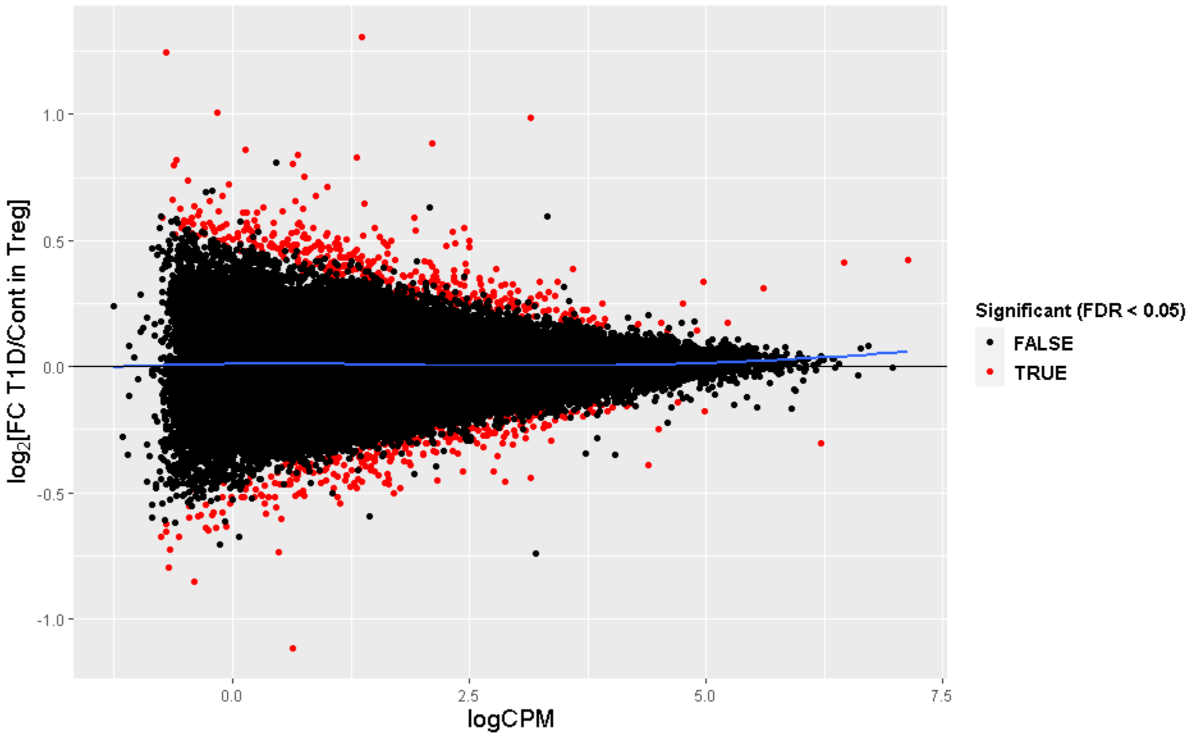


Figure 5-15 | MA plots depicting global differential accessibility distributions from the Treg ATAC-seq libraries.

The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Treg cells [8], super- and typical- enhancers of CD4$^+$ T cells [7].

The data shown are representative of counts in peaks from 12 T1D and sibling-matched healthy control samples. X-axis of MA plot represents average peak signal abundance at that region, while Y-axis corresponds to the log2 difference in peak signal between Treg and Tconv. Black dots represent non-significant regions, and red dots represent significant (FDR < 0.05) differential regions. Blue lines are loess fits to each distribution.

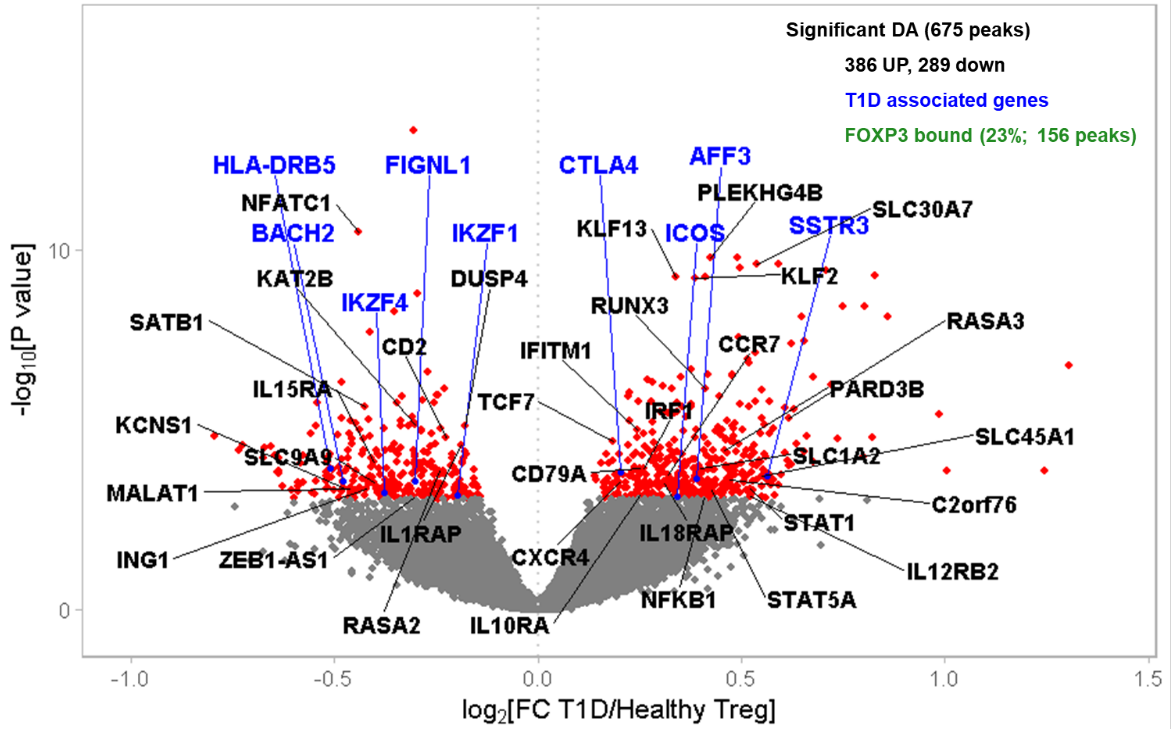


Figure 5-16 | Volcano plot showing differential accessibility analysis between Treg from 12 T1D and sibling-matched healthy subjects.

The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Treg cells [8], super- and typical- enhancers of CD4^{+} T cells [7].

Regions having significantly differential accessibility (Benjamin Hochberg FDR < 0.05) are coloured red or blue. Differentially accessible regions in blue were previously known to be associated with T1D [2, 10]. Differential peaks were annotated to the nearest TSS in linear genomic distance. Selected differentially accessible immune-relevant loci were annotated with gene symbols.

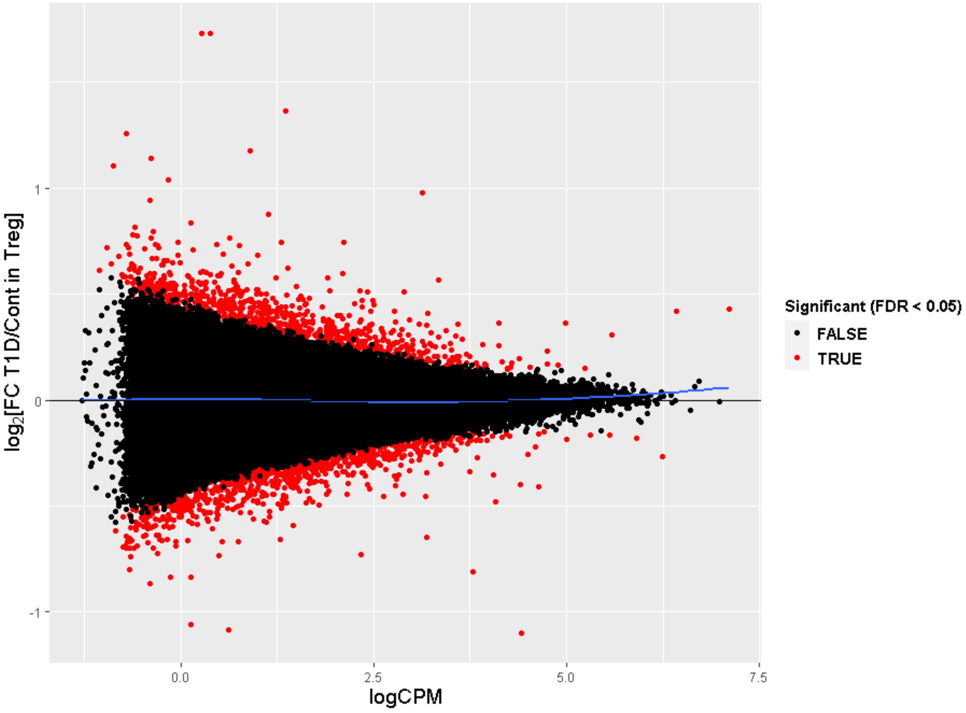


Figure 5-17 | MA plots depicting global differential accessibility distributions from the Treg ATAC-seq libraries.

The input for this differential accessibility analysis was count values derived from all top 100,000 peaks (by peak significance).

The data shown are representative of counts in peaks from 12 T1D and sibling-matched healthy control samples. X-axis of MA plot represents average peak signal abundance at that region, while Y-axis corresponds to the log2 difference in peak signal between Treg and Tconv. Black dots represent non-significant regions, and red dots represent significant (FDR < 0.05) differential regions. Blue lines are loess fits to each distribution.

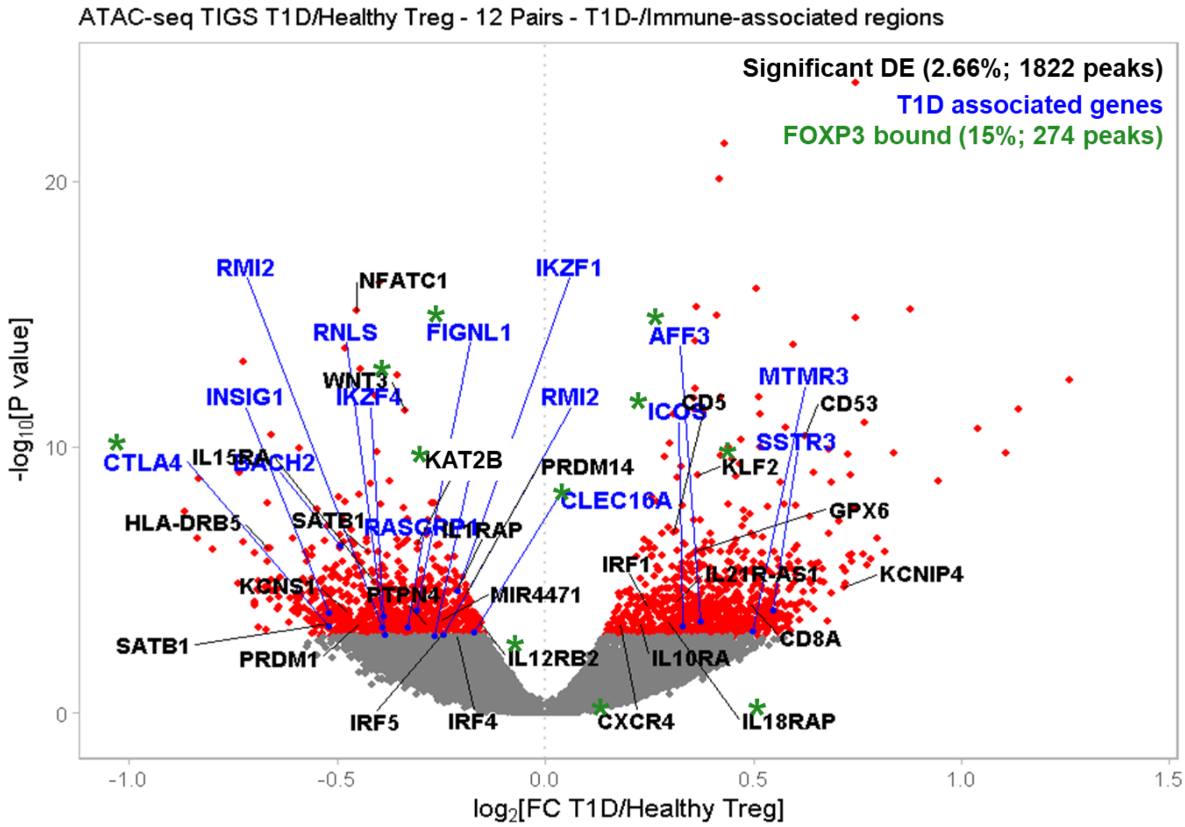


Figure 5-18 | Volcano plot showing differential accessibility analysis between Treg from 12 T1D and sibling-matched healthy subjects.
The input for this differential accessibility analysis was count values derived from all top 100,000 peaks (by peak significance).

Regions having significantly differential accessibility (Benjamin Hochberg FDR < 0.05) are coloured red or blue. Differentially accessible regions in blue were previously known to be associated with T1D [2, 10]. Differential peaks were annotated to the nearest TSS in linear genomic distance. Selected differentially accessible immune-relevant loci were annotated with gene symbols.

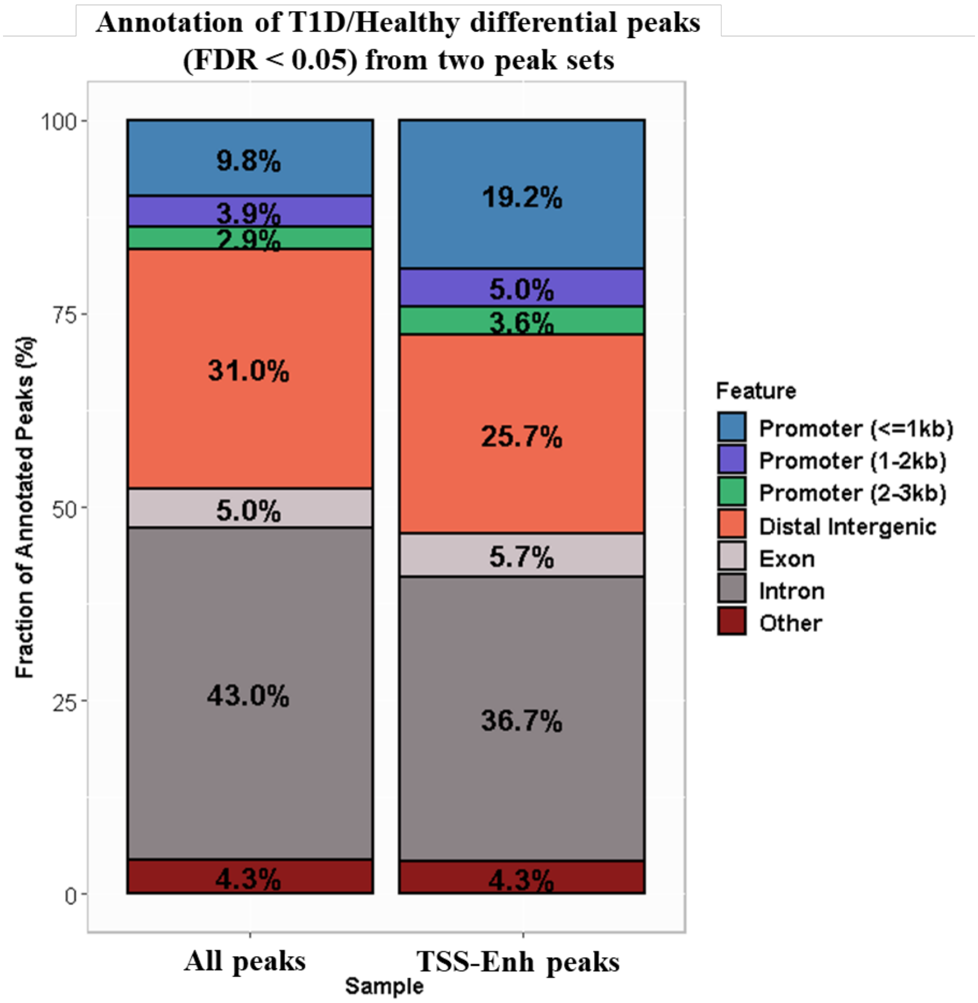


Figure 5-19 | Comparison of genomic annotation for Treg differentially accessible peaks between T1D and healthy controls *(paired T test P value > 0.9)*.

Input peakset for differential accessibility analysis was generated from all top 100,000 peaks (left column) (Figure 5-18) or top 100,000 peaks intersecting TSS, typical enhancers of Treg cells[8], super- and typical- enhancers of CD4$^+$ T cells [7] (Figure 5-16) (right column).

ATAC-seq datasets were generated from stimulated Treg cells purified from 12 T1D and healthy control samples (see Section 5.3 for methods) and peaks were called from mapped reads pooled from all replicates within the same comparison group. *Others, 3UTR , miRNA , ncRNA , TTS , pseudo , 5UTR , snoRNA , snRNA and rRNA; TSS, transcription start sites; Enh, Enhancers.*

Statistical comparisons between T1D and control group were performed with multiple paired T-test with p values ≤ 0.05 being considered significant.

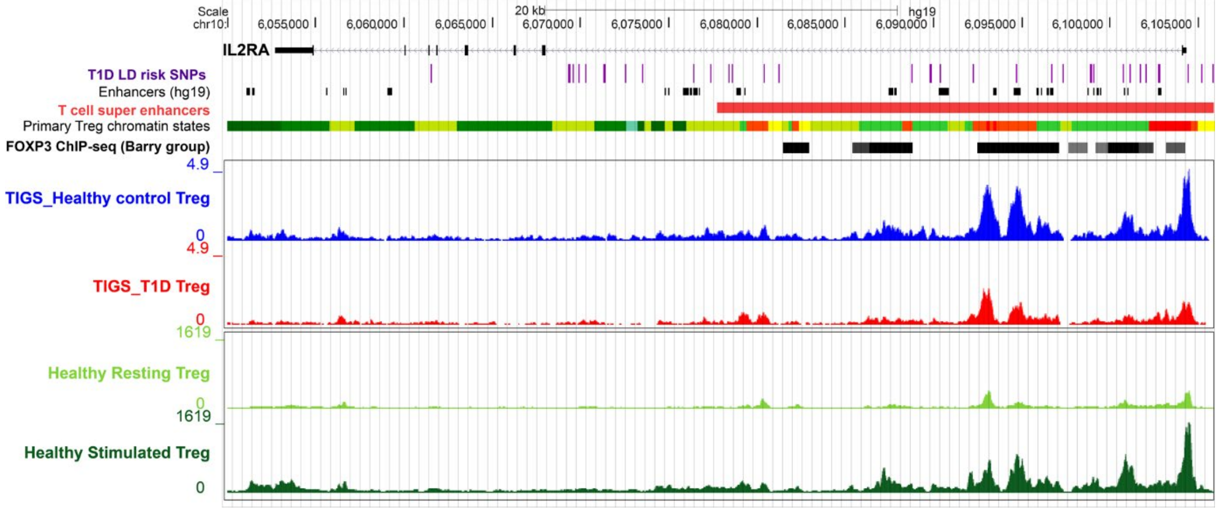


Figure 5-20 | Chromatin accessibility profiles of a representative paired T1D and healthy control samples at the IL2RA locus.

ATAC-seq signal was intersected with T1D risk SNPs in high linkage disequilibrium (R^2=0.8) with the GWAS lead SNPs [3], hg19 enhancer annotation (FANTOM5 phase2.5 [9]), T cell super enhancers [7], Treg chromatin states (Roadmap Epigenomics Project [8]) and FOXP3 binding sites [11]. ATAC-seq profiles generated from resting and stimulated fresh Treg cells from healthy adult donors (n=3) (Section 3.3.3) were broadcast for comparison. Browser view was generated using UCSC genome browser.

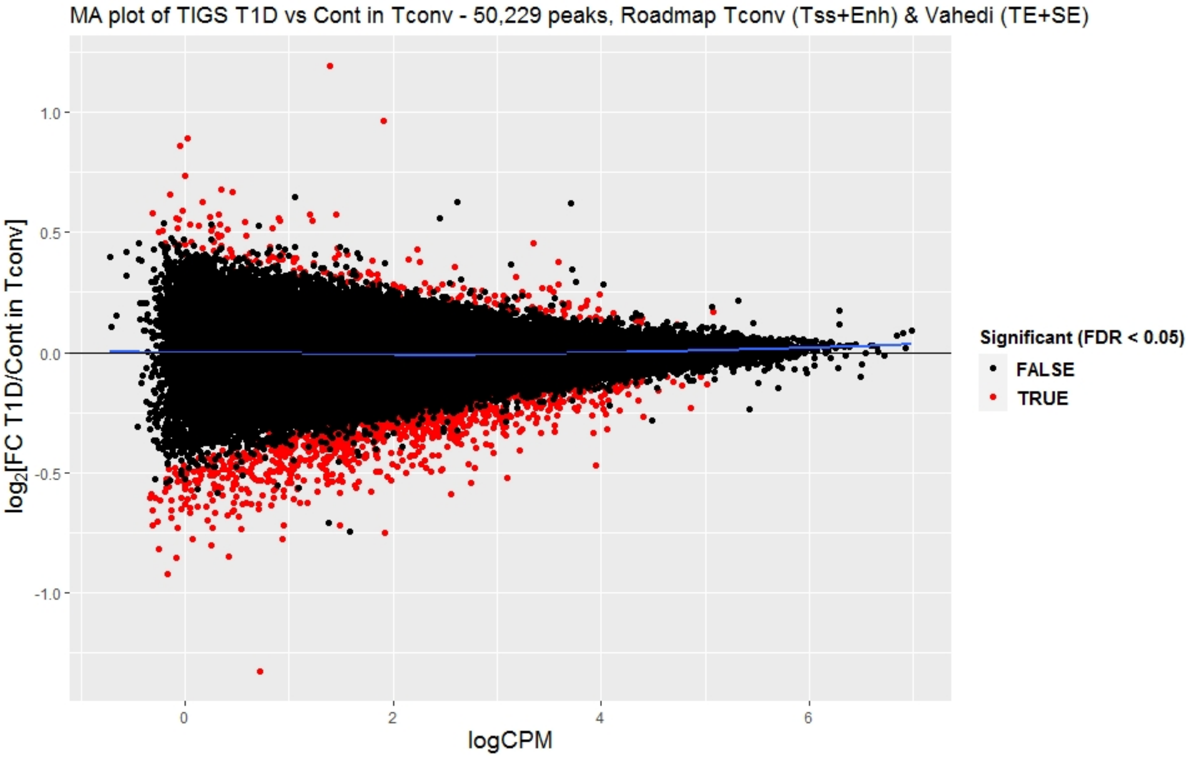


Figure 5-21 | MA plots depicting global differential accessibility distributions from the Tconv ATAC-seq libraries.

The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Helper T cells [8], super- and typical- enhancers of CD4$^+$ T cells [7].

The data shown are representative of counts in peaks from 12 T1D and sibling-matched healthy control samples. X-axis of MA plot represents average peak signal abundance at that region, while Y-axis corresponds to the log2 difference in peak signal between Treg and Tconv. Black dots represent non-significant regions, and red dots represent significant (FDR < 0.05) differential regions. Blue lines are loess fits to each distribution.

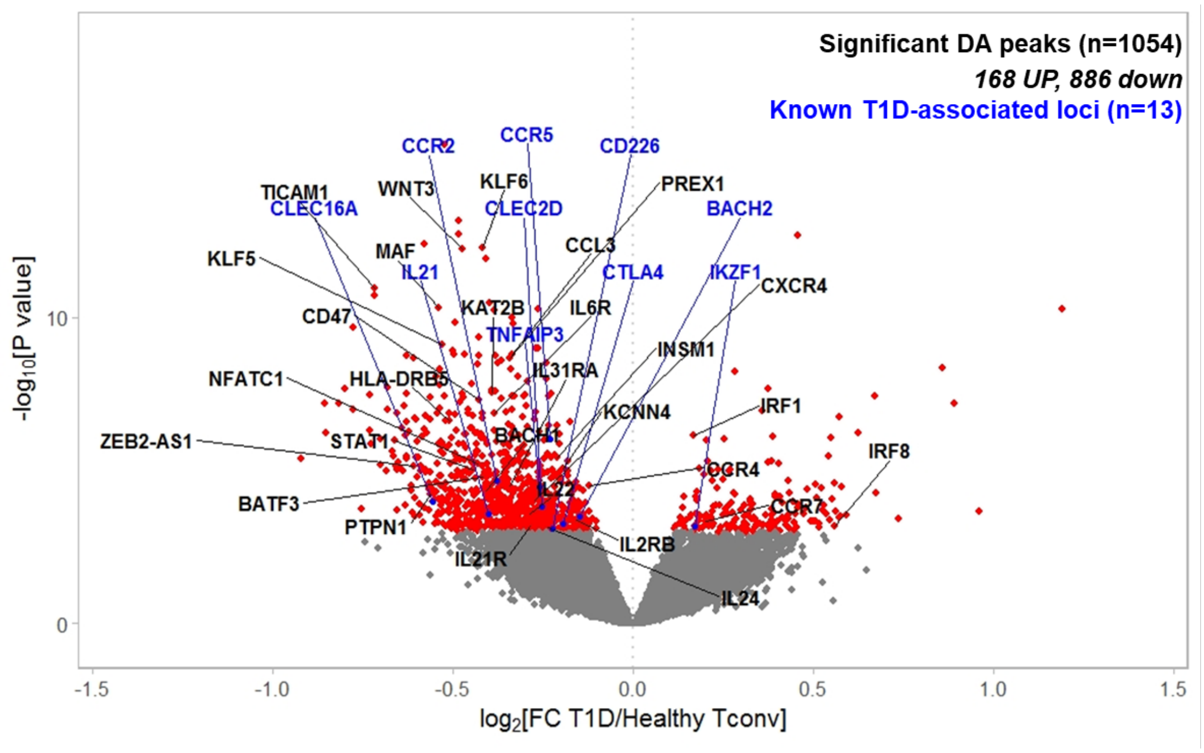


Figure 5-22 | Volcano plot showing differential accessibility analysis between Tconv from 12 T1D and sibling-matched healthy subjects.

The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Helper T cells [8], super- and typical- enhancers of CD4$^+$ T cells [7].

Regions having significantly differential accessibility (Benjamin Hochberg FDR < 0.05) are coloured red or blue. Differentially accessible regions in blue were previously known to be associated with T1D [2, 10]. Differential peaks were annotated to the nearest TSS in linear genomic distance. Selected differentially accessible immune-relevant loci were annotated with gene symbols.

5.4.2.4 Pathway analyses

Gene Set Enrichment Analysis (GSEA) of case-control Treg ATAC-seq identified several statistically significant overrepresented Hallmark gene sets such as those associated with IL2-STAT5 signalling and pro-inflammatory TNF signalling via NFκB (Figure 5-23). Some of the Immunogic signature gene sets (C7) identified in the case-control Treg ATAC-seq were linked to genes previously known to be associated with T1D (Figure 5-24). As for Tconv ATAC-seq pathway enrichment analyses showed that the differentially accessible regions were strongly enriched for terms associated with T cell activation, differentiation and cytokine-cytokine receptor interaction (Figure 5-25, Figure 5-26, Figure 5-27). A subset of the statistically significant overrepresented gene sets identified in case-control Tconv ATAC-seq showed overlap with Treg ATAC-seq, including IL2-STAT5 signalling, IFNγ/a response and pro-inflammatory TNF signalling via NFκB (Figure 5-28). As observed in Treg ATAC-seq, GSEA of case-control Tconv ATAC-seq captured immunologic signatures previously known to be associated with T1D as well (Figure 5-29).

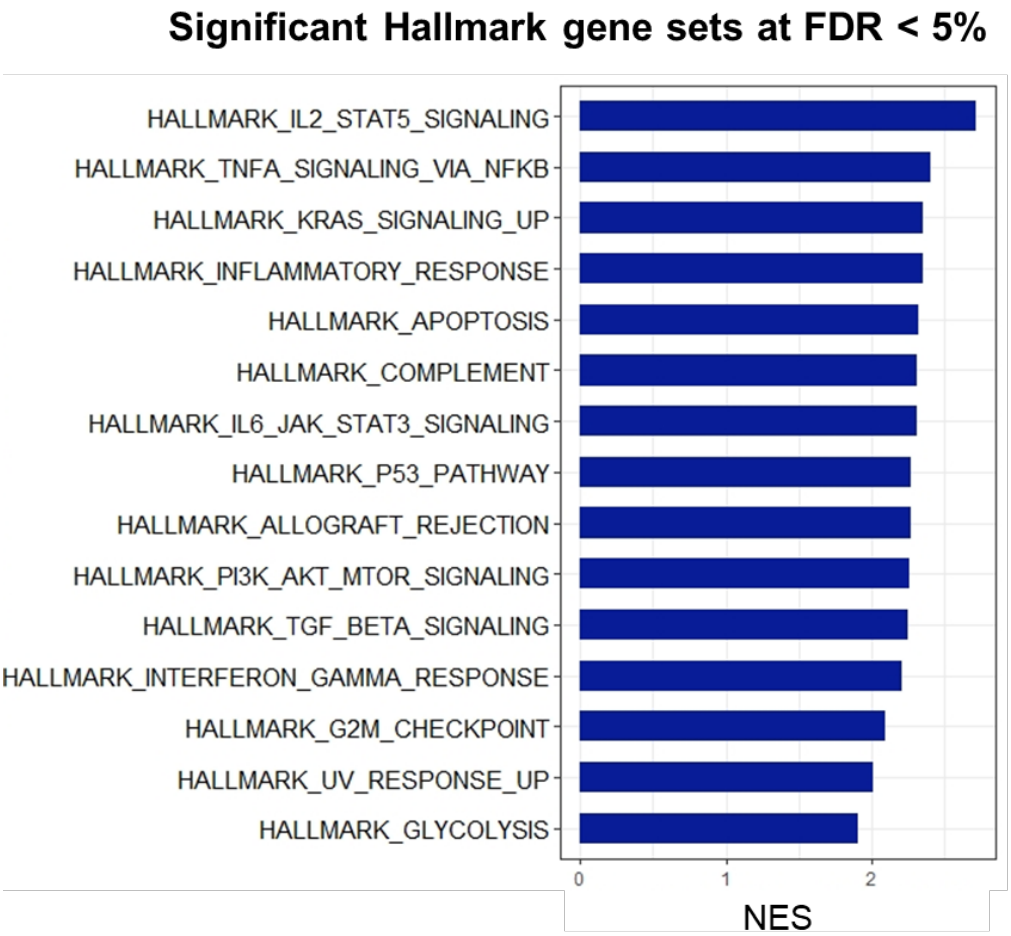


Figure 5-23 | Gene set enrichment analysis (GSEA) showing significantly enriched (FDR < 0.05) hallmark gene sets (H) between T1D and healthy Treg ATAC-seq.
Results are representative of Treg accessibility profiles from 12 T1D and sibling-matched healthy subjects.

The input of differential accessibility analysis for this GSEA analysis (and the subsequent pathway enrichment analyses) was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Treg cells [8], super- and typical- enhancers of CD4^{+} T cells [7].

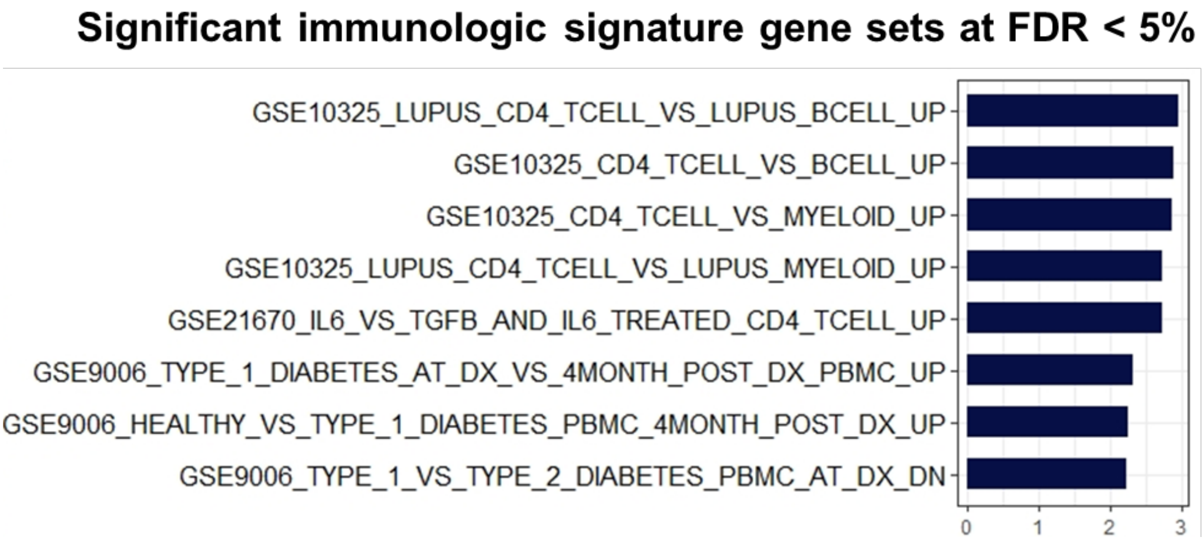


Figure 5-24 | Gene set enrichment analysis (GSEA) showing significantly enriched (FDR < 0.05) immunologic signature gene sets (C7) between T1D and healthy Treg ATAC-seq. Results are representative of Treg accessibility profiles from 12 T1D and sibling-matched healthy subjects.

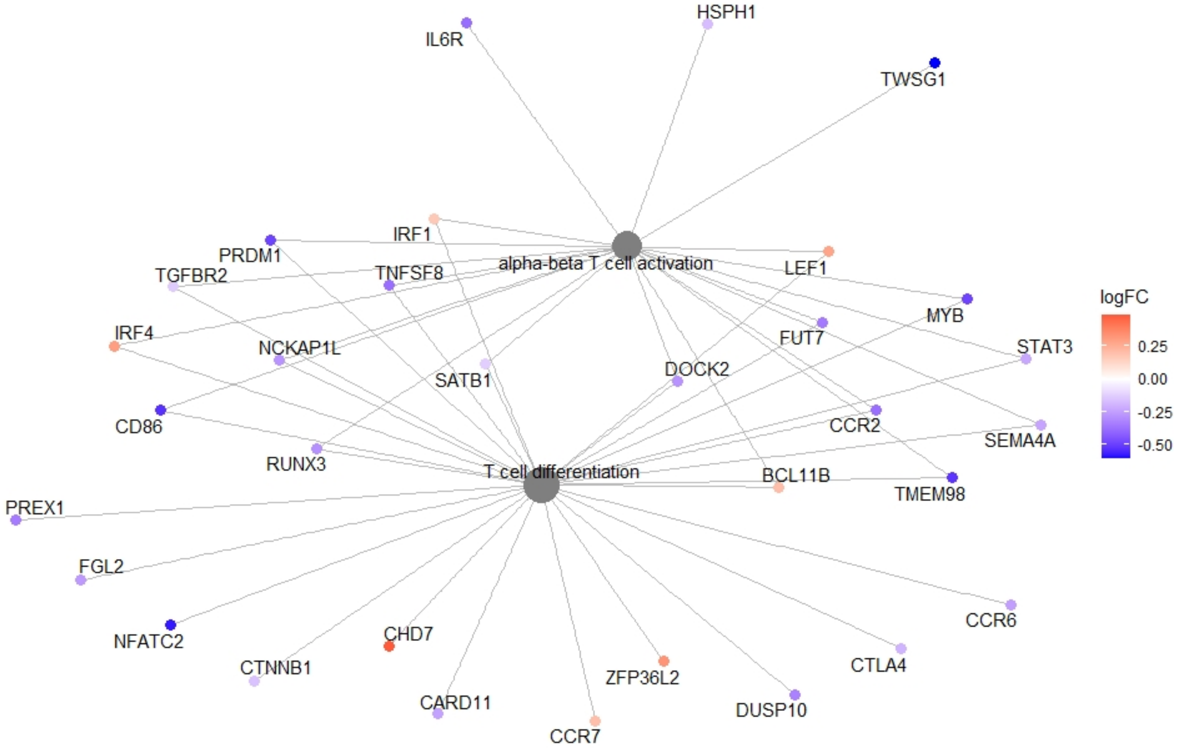


Figure 5-25 | Enrichment plot showing the most significant GO terms and related differentially accessible loci from the Biological Process Ontology for T1D vs healthy Tconv ATAC-seq, using a Bonferroni adjusted p-value < 0.05 as the criteria for significance.
Results are representative of accessibility profiles from 12 T1D and sibling-matched healthy subjects.

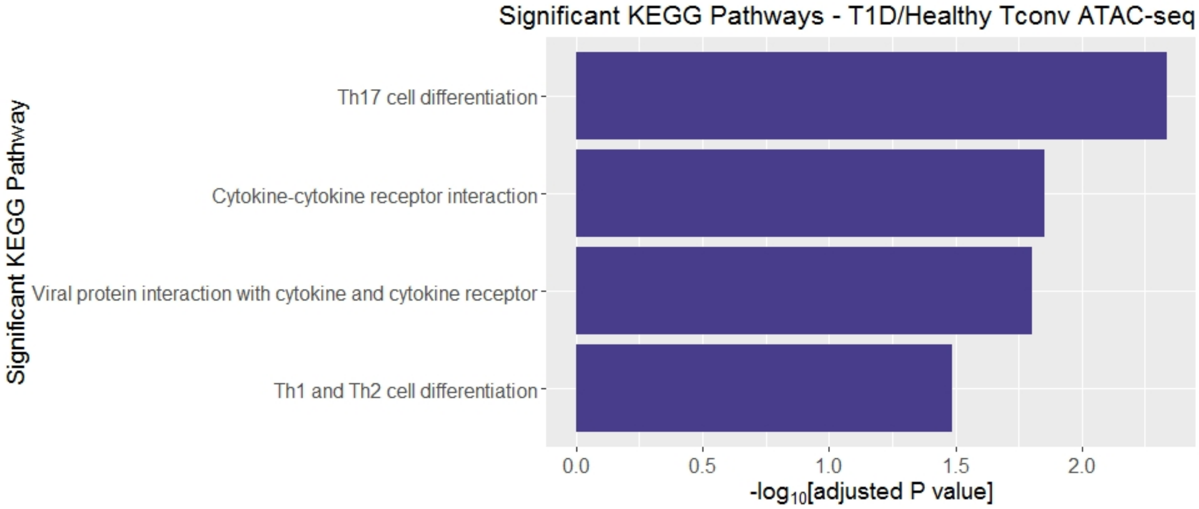


Figure 5-26 | KEGG pathway analysis showing significantly enriched pathways associated with differentially accessible peaks between T1D and healthy Tconv ATAC-seq, using a Bonferroni adjusted p-value < 0.05 as the criteria for significance.

Results are representative of accessibility profiles from 12 T1D and sibling-matched healthy subjects.

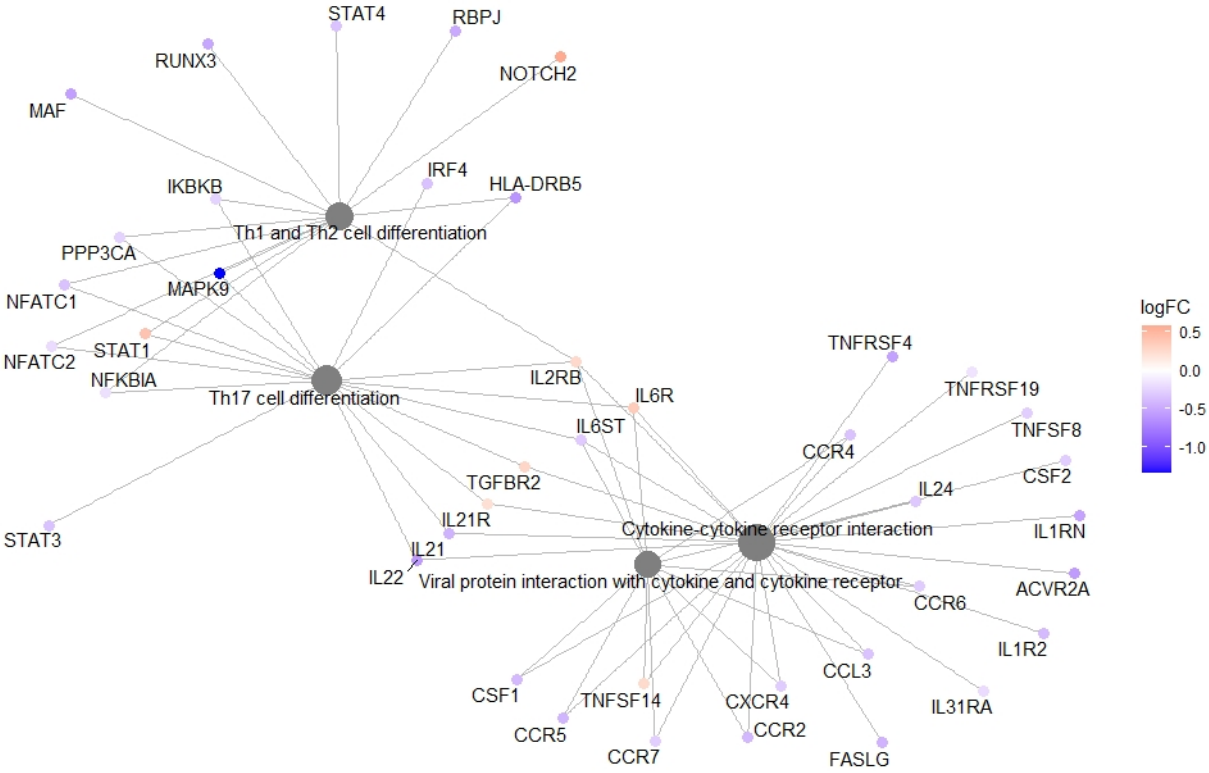


Figure 5-27 | Enrichment plot showing the most significant KEGG pathways and associated genes for T1D versus healthy Tconv ATAC-seq.

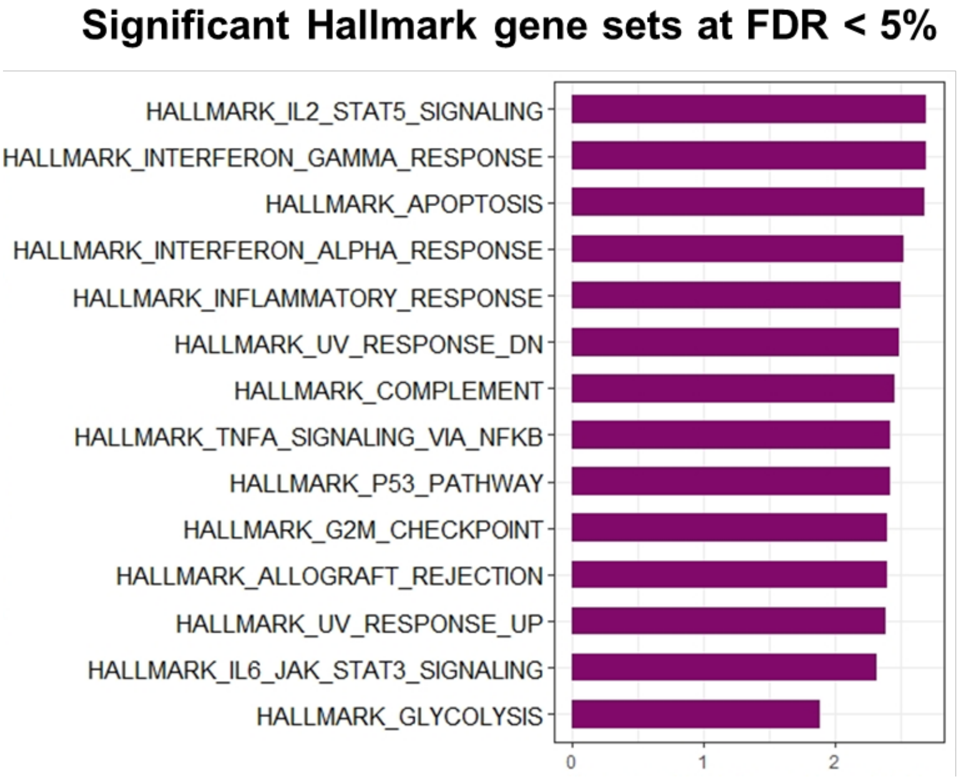


Figure 5-28 | Gene set enrichment analysis (GSEA) showing significantly enriched (FDR < 0.05) hallmark gene sets (H) between T1D and healthy Tconv ATAC-seq.
Results are representative of Tconv accessibility profiles from 12 T1D and sibling-matched healthy subjects.

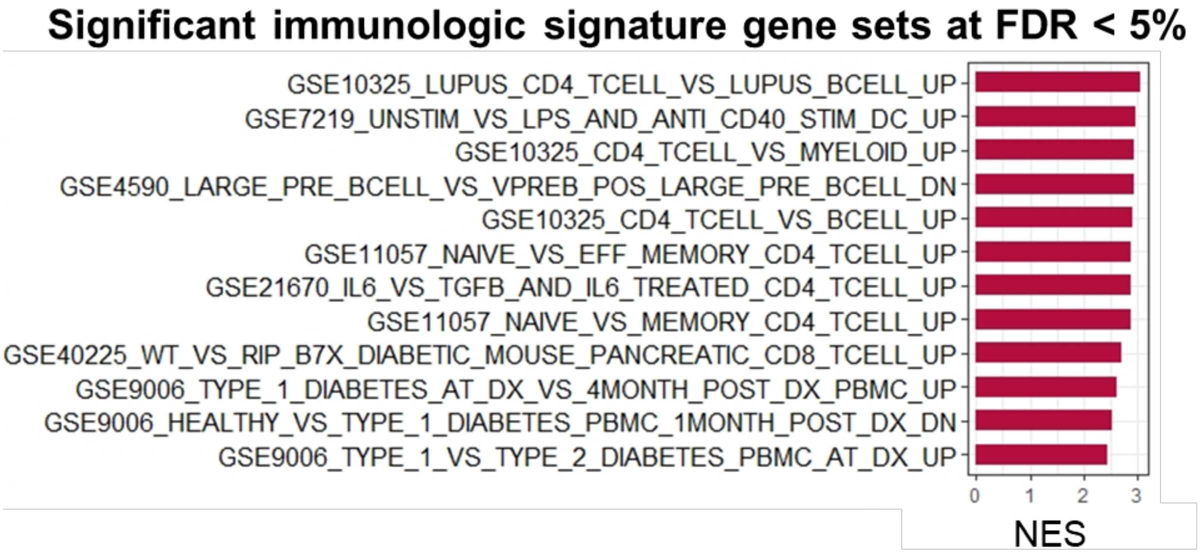


Figure 5-29 | Gene set enrichment analysis (GSEA) showing significantly enriched (FDR < 0.05) immunologic signature gene sets (C7) between T1D and healthy Tconv ATAC-seq.
Results are representative of Tconv accessibility profiles from 12 T1D and sibling-matched healthy subjects.

5.4.2.5 TF footprinting

Many transcription factors (TFs) regulate gene expression by altering chromatin accessibility so I next attempted to identify which TFs regulate chromatin accessibility in T1D.

ATAC-seq technology relies on the cleaving of open chromatin regions by Tn5, followed by sequencing and mapping of reads from the DNA fragments to identify their genomic locations and quantify the accessibility signal. However, the occupancy of TFs prevents the DNA cleavage by Tn5 at accessible chromatin regions leaving footprints. They are characterised by small DNA regions bound by TFs with a sudden drop in read coverage within peak regions of high coverage. Computational/digital footprinting is commonly used to infer TF occupancy at the nucleotide level in DNase-seq and ATAC-seq data [21, 128, 176, 182, 267]. It is an attractive alternative to the traditional approach of ChIP-seq in probing protein−DNA interactions as it circumvents antibody issues, poor resolution and high cell input requirement.

The Tconv and Treg ATAC-seq libraries generated from the T1D cohort and healthy controls were highly complex (Figure 7-34, Figure 7-35, Figure 7-36) and achieved sufficient read depth required for probing TF residence at single-nucleotide resolution. I used HINT-ATAC [21] to identify TF footprints and perform motif matching using TF position weight matrix (PWM) as well as differential footprinting from my case-control ATAC-seq data (methodology described in Section 5.3.3.5). Some of the TFs that showed statistically significant differential activity in Treg cells between T1D and healthy, as measured by difference in the ATAC-seq read coverage signal at the predicted binding sites, included RUNX2, Fos- and Jun-like TFs, FOSL1 and FOSL2 (Figure 5-30, Figure 5-31). Those TFs may interact collaboratively to alter chromatin accessibility, especially the interaction between RUNX2 and activator protein 1 (AP-1) family like Fos-like TFs have been reported [268, 269]. Non-differential TF footprints such as CTCF-L and STAT3 showed high concordance of read coverage signal between T1D and healthy controls at the predicted binding sites (Figure 5-31). A subset of the differential TFs also demonstrated altered transcript levels as revealed by our case-control RNA-seq data (Figure 5-30). To gain an insight into how alterations in TF binding might impact the transcriptional regulatory networks in T1D, I performed pathway enrichment analysis for the TF binding sites that showed differential regulation between T1D and healthy Treg cells. The genomic regions linked with alterations in TF binding in T1D were strongly associated with immunoregulatory pathways such as T cell activation and differentiation, inflammatory response to antigenic stimulus and IL2-STAT5 signalling (Figure 5-32).

Some of the TFs that showed significant alterations in binding activity in Tconv cells of T1D included EGR (Early growth response) protein family, which also showed altered transcript levels in the case-control RNA-seq, as well as Retinoic acid receptor α (RARA), which is crucial for the differentiation and functions of Th1 cell lineage (Figure 5-33, Figure 5-34). It has been shown that a T1D risk variant located within the enhancer region of T cells disrupts binding of RARA resulting in reduced expression of target genes, including *CD69*, *CELE2B*, *KLRB1*, and *EIF2S3L* [33]. As seen in Treg dataset, the altered TF footprints in Tconv of T1D were strongly associated with immune function and regulation, including leukocyte chemotaxis, T cell activation and differentiation as well as regulation of B cell proliferation (Figure 5-35).

I also explored the TF binding dynamics in the gender-matched case-control pairs, which constitute 5 male pairs of T1D and their healthy sibling counterparts. A fraction of the differential TFs discovered in the gender-matched pairs were captured in the full dataset (n=12 pairs) too, including RUNX2, Fos- and Jun-like TFs, FOSL1 and FOSL2 for Treg (Figure 7-61), and EGR proteins, Fos- and Jun-like TFs, FOSL2 for Tconv (Figure 7-62). In addition, gender-matched profiles also discovered altered TF footprints in T1D which were not captured in the full dataset (n=12), such as BACH1:MafK in Treg (Figure 7-61), which functions primarily as a transcriptional suppressor through binding to Maf recognition element (MARE) [270, 271], as well as BATF-JUN in Tconv (Figure 7-62), which is critical for IRF4 (Interferon regulatory factor 4)-mediated transcription in activated CD4$^+$ T cells and Th17 differentiated cells [272].

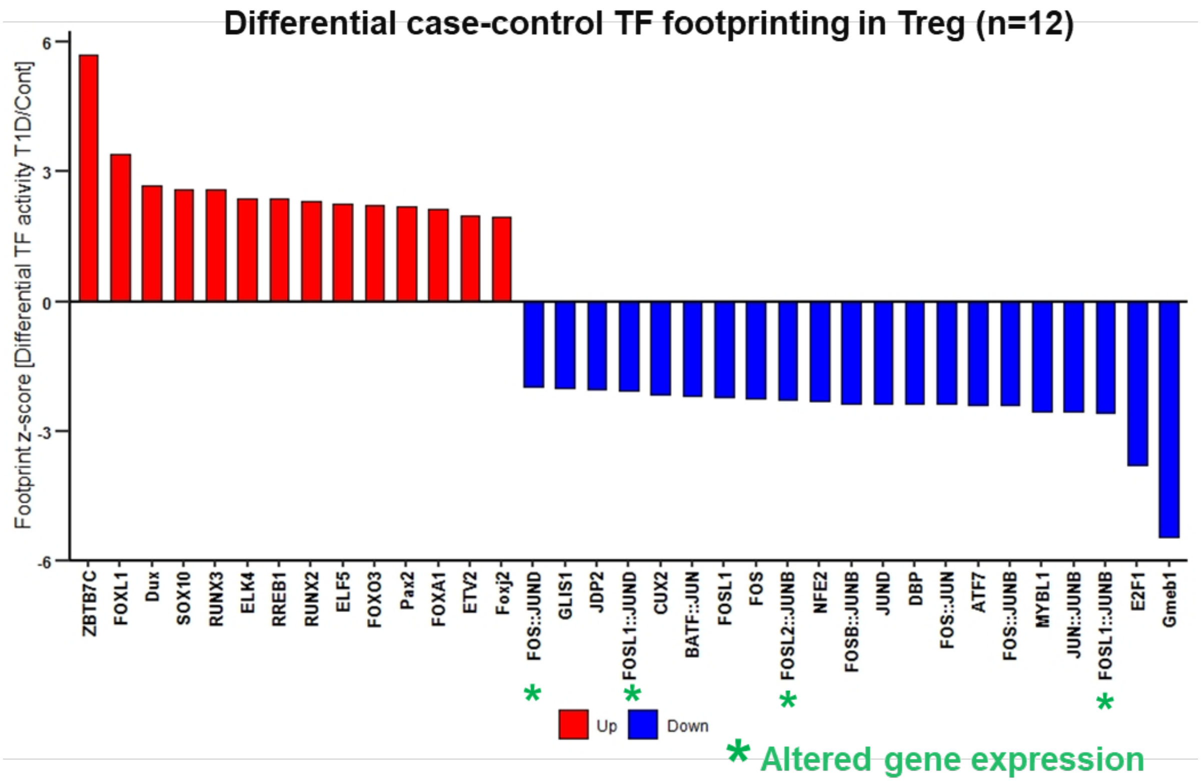


Figure 5-30 | Significantly differential transcription factor (TF) footprinting between T1D and healthy Treg cells.

Alterations in TF activity dynamics in T1D were identified from differentially accessible loci calculated from 12 T1D and sibling-matched healthy subjects. Footprints were identified from mapped sequencing reads pooled from 12 T1D and sibling-matched healthy ATAC-seq libraries and considered for differential footprinting analysis. TFs with a significant change in activity score (p-value < 0.05) were plotted with enrichment z-scores such that the activity levels are comparable across different TFs. TFs with a green asterisk showed significantly differential gene expression levels in T1D vs. healthy Treg RNA-seq (Section 5.4.3). *Up, upregulated in T1D group; down, downregulated in T1D group.*

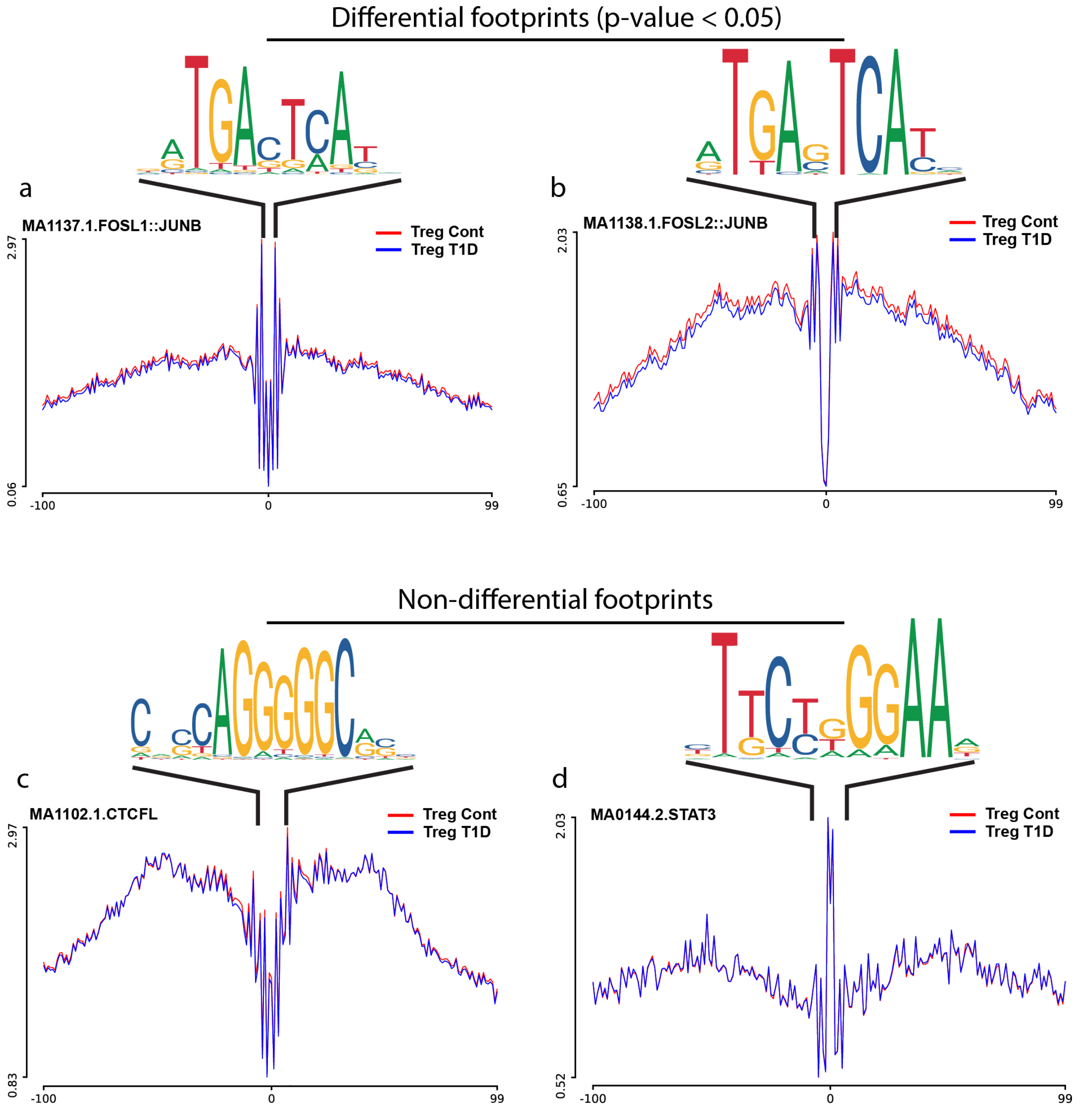


Figure 5-31 | Representative average ATAC-seq profiles around binding motif of TF footprints that showed significant alterations in activity between T1D and healthy Treg cells.

Histograms are generated from Treg chromatin accessibility signals computed from pooled sequencing data representative of 12 T1D and sibling-matched healthy subjects. Footprint signal was computed using HINT-ATAC [21] on the genome-wide footprints matching the corresponding motifs obtained from JASPAR [29] database. Higher ATAC-seq signal around the TF binding motif is associated with higher activity of that TF in a particular condition. Non-differential TF footprints with no significant change in activity between T1D and healthy controls are shown as comparison (c, d).

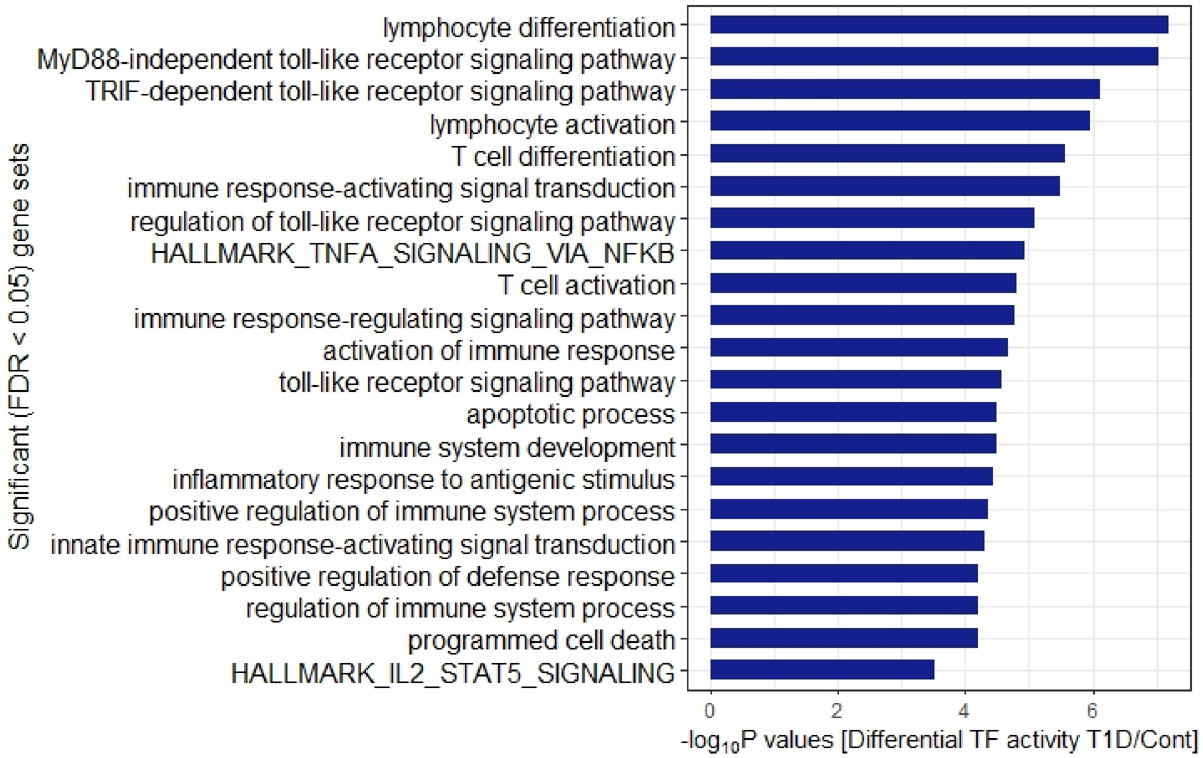


Figure 5-32 | Pathway enrichment analysis showing significantly enriched (FDR < 0.05) terms or gene sets (H) for genomic regions exhibiting differential TF footprints identified between T1D and healthy Treg ATAC-seq.
Differential TF footprints were identified using HINT-ATAC [21]. The binding sites (footprints) for all the differential TFs were extracted and subjected to ChIP-Enrich [44] for gene set enrichment testing. Results are representative of Treg accessibility profiles from 12 T1D and sibling-matched healthy subjects.

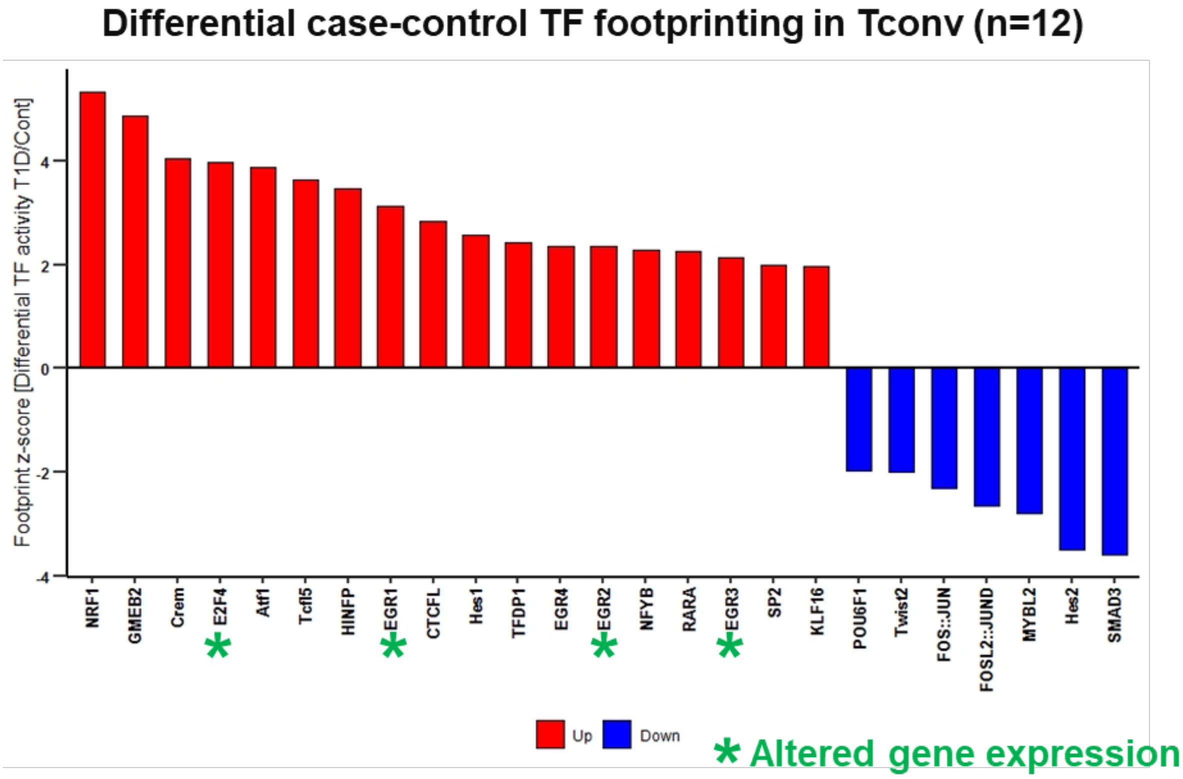


Figure 5-33 | Significantly differential transcription factor (TF) footprinting between T1D and healthy Tconv cells.

Alterations in TF activity dynamics in T1D were identified from differentially accessible loci calculated from 12 T1D and sibling-matched healthy subjects. Footprints were identified from mapped sequencing reads pooled from 12 T1D and sibling-matched healthy ATAC-seq libraries and considered for differential footprinting analysis. TFs with a significant change in activity score (p-value < 0.05) were plotted with enrichment z-scores such that the activity levels are comparable across different TFs. TFs with a green asterisk showed significantly differential gene expression levels in T1D vs. healthy Tconv RNA-seq (Section 5.4.3). *Up, upregulated in T1D group; down, downregulated in T1D group.*

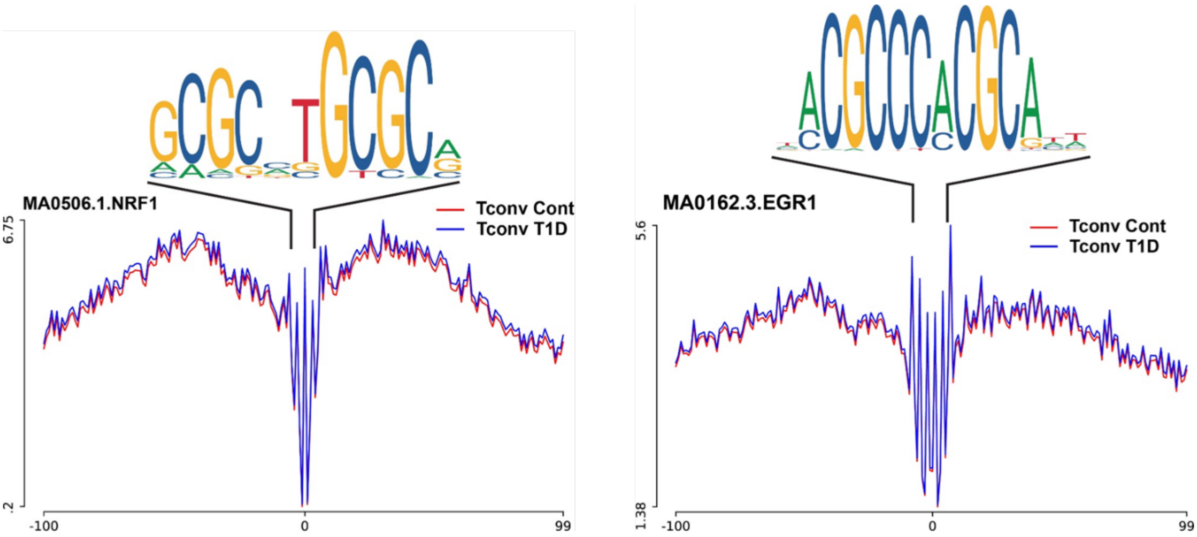


Figure 5-34 | Representative average ATAC-seq profiles around binding motif of TF footprints that showed significant alterations in activity between T1D and healthy Tconv cells.

Histograns are generated from Tconv chromatin accessibility signals computed from pooled sequencing data representative of 12 T1D and sibling-matched healthy subjects. Footprint signal was computed using HINT-ATAC [21] on the genome-wide footprints matching the corresponding motifs obtained from JASPAR [29] database. Higher ATAC-seq signal around the TF binding motif is associated with higher activity of that TF in a particular condition.

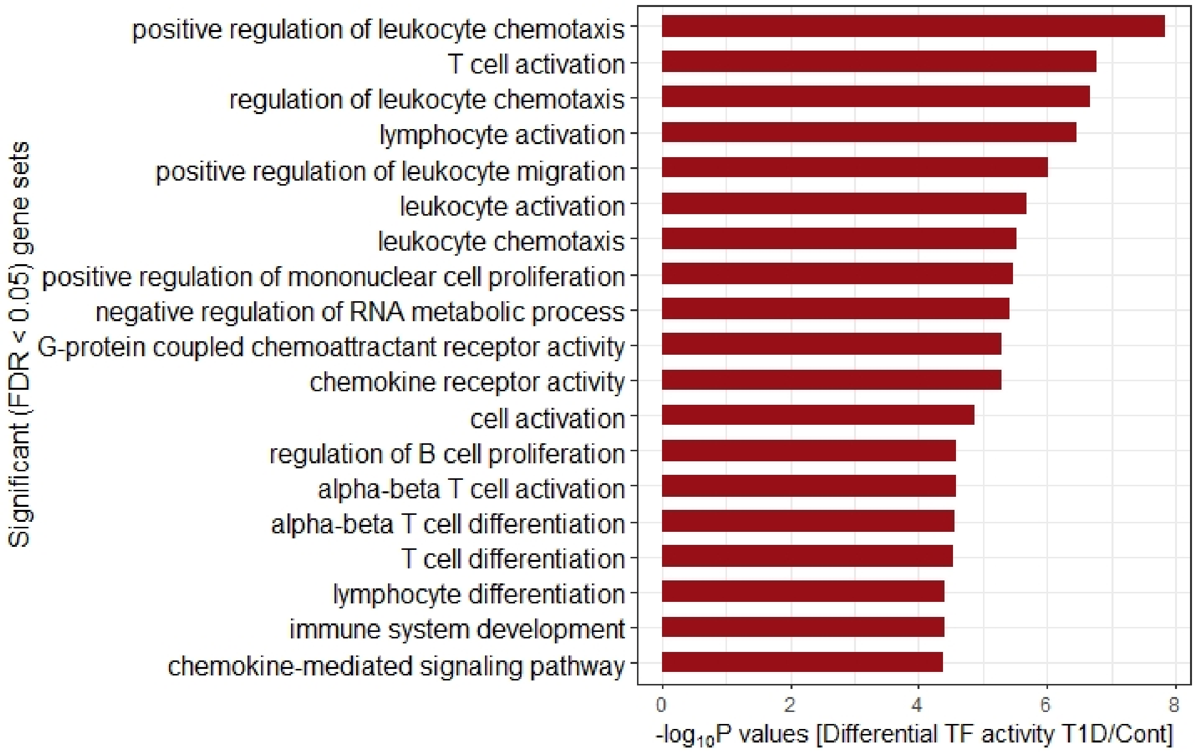


Figure 5-35 | Pathway enrichment analysis showing significantly enriched (FDR < 0.05) terms or gene sets (H) for genomic regions exhibiting differential TF footprints identified between T1D and healthy Tconv ATAC-seq.

Differential TF footprints were identified using HINT-ATAC [21]. The binding sites (footprints) for all the differential TFs were extracted and subjected to ChIP-Enrich [44] for gene set enrichment testing. Results are representative of Tconv accessibility profiles from 12 T1D and sibling-matched healthy subjects.

5.4.3 RNA-seq

5.4.3.1 Libraries quality

I then profiled the transcriptome of T1D patients and their healthy sibling counterparts using RNA-seq. Insufficient input material was recovered from Treg cells for use on ATAC-seq and RNA-seq independently. To overcome the issue of resource scarcity I extracted transcriptome from (activated) Tconv and Treg ATAC-seq lysate reaction for RNA-seq experiments (referred to as "ATAC-SN"). This approach allows simultaneous profiling of accessibility and transcriptome information from a single ATAC-seq reaction and my early optimisation work has confirmed high overlap of transcripts between ATAC-seq lysate and whole cells (Section 3.4.8). Furthermore, this approach also allows better correlation of chromatin accessibility and gene expression as the DNA and RNA material are derived from the same pool of cell population, thereby eliminating the issue of cell heterogeneity.

Intact, high quality RNA material was recovered from most of the Tconv and Treg ATAC-SN reactions of T1D and healthy control samples, as indicated by registration of high RNA quality indicator (RQI) scores (Figure 7-63). However, RNA samples recovered from some of the low complexity Treg ATAC-SN reactions – JP23, JP73, JP25, JP26, JP8 () was partially degraded (RIN 6 – 7). Intact RNA samples (RIN > 7) were enriched for Poly(A) RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs; cat# E7490) whereas partially degraded RNA samples (RIN 6 – 7) (3 case-control pairs) were ribosomal RNA-depleted (New England Biolabs; cat# E6310) prior to generation of cDNA libraries using a NEBNext Ultra Directional II RNA Library Preparation Kit for Illumina (New England Biolabs; cat# E7760) (Section 5.3.2). The fragment size distribution of RNA-seq libraries constructed from the RNA samples prepared from ATAC-SN showed a narrow distribution with a peak size of approximately 300bp (Figure 7-64), which is a typical size distribution for standard RNA-seq libraries constructed with NEBNext Ultra II Directional RNA Protocol (Section 3.3.4).

5.4.3.2 Exploratory analysis

RNA-seq libraries were generated from Tconv and Treg cells obtained from 12 T1D patients and 12 sibling-matched healthy subjects, however, one Tconv RNA sample (JP29 from T1D group) was partially degraded and thus was omitted from RNA-seq experiment (see Table 7-1 for cohort information and RNA-sequenced libraries). The remaining RNA-seq samples, which constituted of 23 Tconv and 24 Treg libraries, were barcoded, pooled and sequenced across 4 lanes of paired-end 150-cycle Illumina Hiseq sequencer to an average read depth of 25.8 million reads (± 4.6 million) per sample (Table 5-6). The average yield was 26,057,786 (± 3,754,917) per Tconv RNA-seq library and 25,619,943 (± 5,353,334) per Treg RNA-seq library (Figure 5-36).

Principal component analysis (PCA) showed that the largest proportion of variation driving the sample separation (n=47 libraries) in the case-control RNA-seq data was cell type (Figure 5-37), supporting the observation from case-control ATAC-seq accessibility profiles (Figure 7-54). The RNA-seq samples involved in this study underwent either Poly(A) mRNA enrichment or ribosomal RNA depletion depending on the integrity of the recovered RNA. For Poly(A) mRNA enrichment high integrity RNA samples (RIN > 7) is required and ribosomal RNA depletion is recommended for partially degraded RNA samples (RIN = 2 to 7). Of the 47 RNA-seq samples sequenced, most were enriched for Poly(A) mRNA prior to library generation except for Treg RNA samples recovered from 3 case-control matched pairs - JP23, JP73, JP25, JP26, JP8 and JP9, in which subpar RNA quality was observed.

Data analysis revealed RNA-seq libraries prepared using ribosomal RNA depletion protocol demonstrated low rates of uniquely mapped reads (Table 5-7), which is an important quality control parameter for RNA-seq libraries. Low unique mapping or high multimapping rate seen in some case-control RNA-seq samples may be a result of rRNA contamination from incomplete or inefficient rRNA depletion or poor-quality input RNA. Ribosomal RNAs account for 80% of the total RNA present in a cell and are present in multiple copies across the genome and thus, rRNA contamination would result in reads being mapped to multiple genomic loci and eventually discarded by the aligner. To overcome the issue of low unique mapping in rRNA-depleted RNA-seq samples, I filtered reads that mapped to the human ribosomal DNA repeat sequences before alignment which increased the fraction of uniquely mapped reads in those rRNA-depleted samples substantially (Table 5-7). Nonetheless, the reads of those rRNA-depleted RNA-seq samples showed a spike in GC content at approximately 60%, which was likely due to incomplete rRNA depletion (Figure 7-65; a). PCA showed that the largest proportion of variation in the RNA-seq data was driven by the selection approach during library prep as Poly(A) mRNA-enriched and rRNA-depleted samples appeared to cluster within their group (Figure 7-65; b). As GC content was noted as being of concern, I performed conditional-quantile normalisation [40] as an attempt to correct for global distortions resulting from this systematic bias. After conditional quantile normalisation, a second PCA analysis was performed in order to check for improvement in group separation. However, the normalisation appeared to contribute minimally to the sample separation at this level as the RNA-seq samples still appeared to segregate by selection protocol (Figure 7-65; b). As the GC content of the rRNA-depleted RNA-seq samples appeared remarkably affected and upon exclusion of rDNA reads the resulting coverage depth was significantly reduced which can impair the transcript distribution and compromise the differential analysis, these rRNA-depleted libraries were omitted from downstream analysis.

Exploratory analyses also revealed GC content was not the only source of confounding variable in the case-control RNA-seq data as it was clearly shown from the PCA that the effects from family pairing (genetic factors) was stronger than the cohort group, i.e., the samples cluster within their family group rather than the cohort (Figure 7-65; b). As it was done with the case-control ATAC-seq data, I decided to adopt RUV (Remove Unwanted Variation) [32] to correct for these confounder effects that obscure any true association between the biological factor of interest and gene expression levels. After data normalisation a significant improvement in group separation between cases and controls was observed (Figure 7-66).

Table 5-6. Summary of sequencing output yield for RNA-seq experiment of T1D and healthy control samples.
47 RNA-seq libraries, which comprised of Tconv and Treg libraries generated from 12 T1D and sibling-matched healthy control samples, were sequenced across four lanes of Hiseq flowcell. One Tconv RNA-seq library - JP29 from T1D group, had low RNA integrity number (RIN) and was excluded from sequencing.

Sequencing lane	Number of libraries	Total reads	Average yield per library (± SD)
Combined reads from 4 lanes	47 (23 Tconv + 24 Treg libraries)	1,214,207,711	25,834,207 (±4,645,597)

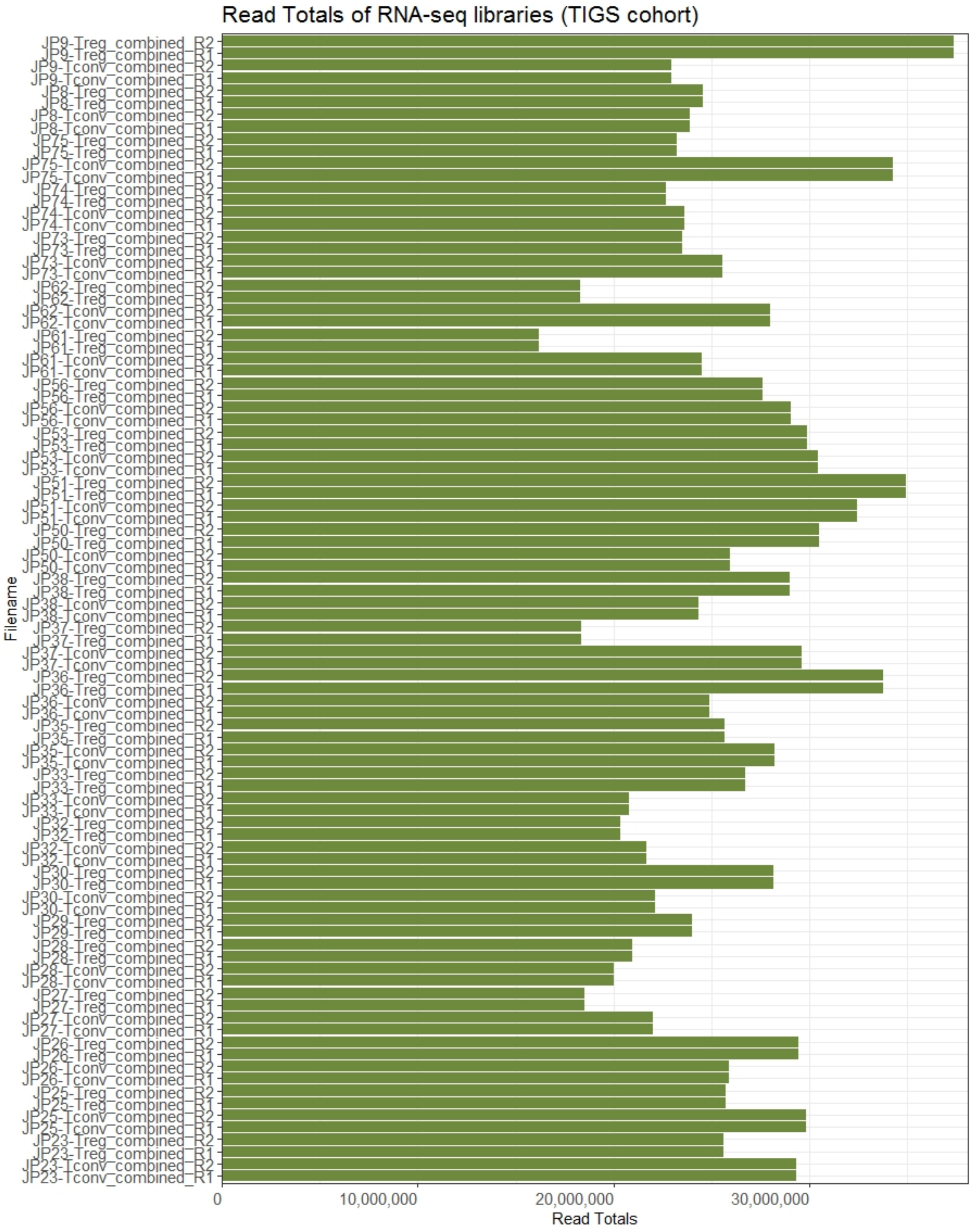


Figure 5-36 | The total number of paired-end reads for all case-control RNA-seq libraries.

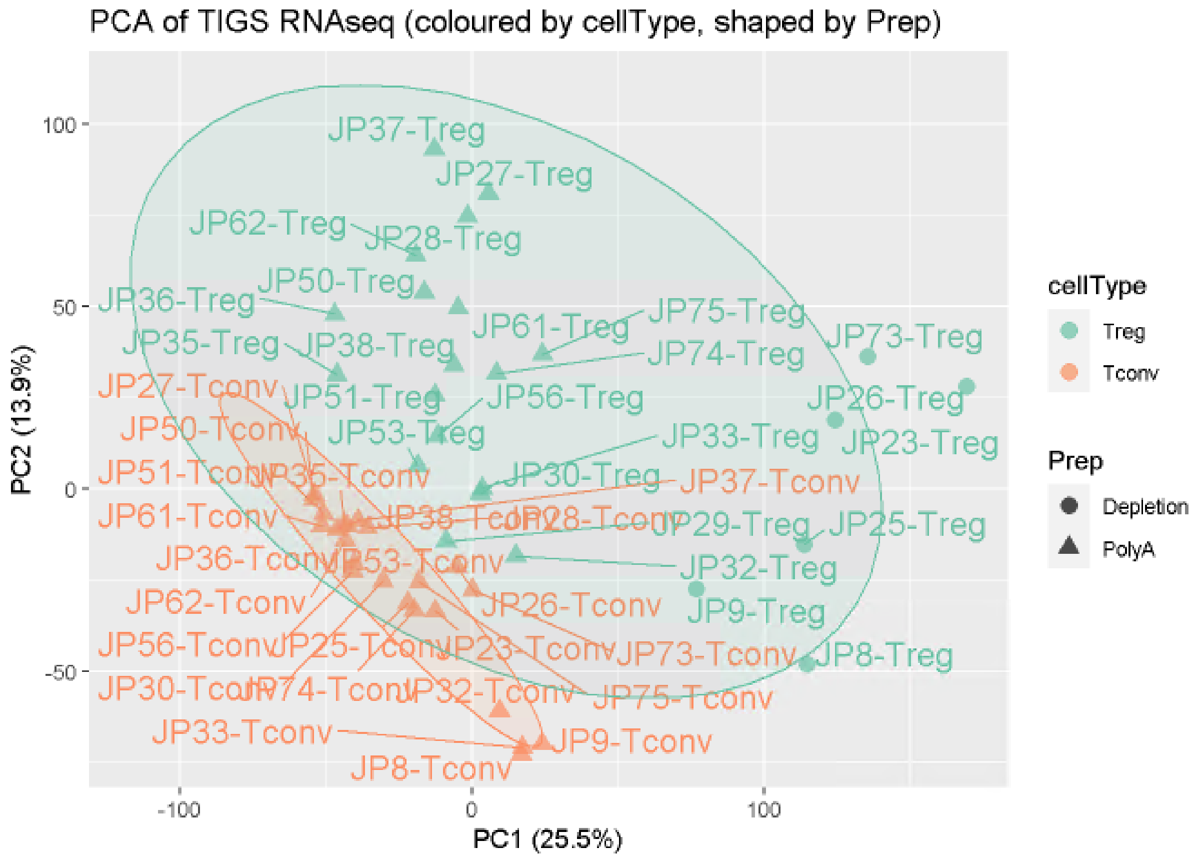


Figure 5-37 | Principal component analysis (PCA) plot of counts in peaks over first two components for Tconv and Treg RNA-seq libraries generated from 12 T1D and healthy control samples.

The RNA-seq samples are separated by cell type and selection method (enrichment for poly-A or depletion for ribosomal RNA transcripts) over the first and second principal component.

Table 5-7. Low unique mapping rates observed in ribosomal RNA (rRNA)-depleted RNA samples, indicating of incomplete or inefficient depletion during sample preparation. The contaminating rRNA reads result in high proportion of multimappers and upon removal of rRNA reads the percentage of uniquely mapped reads for those RNA-seq samples increased.

Sample Name	Prep	% Uniquely Aligned (Pre-rRNA-clean)	% Uniquely Aligned (Post-rRNA-clean)
JP23-Tconv_combined	PolyA	93.40%	94.70%
JP23-Treg_combined	rRNA depletion	43.00%	74.40%
JP25-Tconv_combined	PolyA	92.70%	93.10%
JP25-Treg_combined	rRNA depletion	48.10%	85.10%
JP26-Tconv_combined	PolyA	92.60%	92.90%
JP26-Treg_combined	rRNA depletion	74.20%	77.60%
JP27-Tconv_combined	PolyA	94.00%	94.10%
JP27-Treg_combined	PolyA	91.70%	92.10%
JP28-Tconv_combined	PolyA	92.90%	94.30%
JP28-Treg_combined	PolyA	90.10%	92.20%
JP29-Treg_combined	PolyA	91.20%	92.40%
JP30-Tconv_combined	PolyA	93.80%	94.70%
JP30-Treg_combined	PolyA	86.30%	89.80%
JP32-Tconv_combined	PolyA	92.70%	93.40%
JP32-Treg_combined	PolyA	90.20%	91.10%
JP33-Tconv_combined	PolyA	92.50%	93.90%
JP33-Treg_combined	PolyA	91.20%	93.70%
JP35-Tconv_combined	PolyA	94.30%	94.60%
JP35-Treg_combined	PolyA	93.40%	94.00%
JP36-Tconv_combined	PolyA	93.10%	93.50%
JP36-Treg_combined	PolyA	92.10%	92.60%
JP37-Tconv_combined	PolyA	91.00%	94.20%
JP37-Treg_combined	PolyA	81.20%	81.90%
JP38-Tconv_combined	PolyA	92.50%	93.90%
JP38-Treg_combined	PolyA	89.90%	90.50%
JP50-Tconv_combined	PolyA	92.60%	93.50%
JP50-Treg_combined	PolyA	93.70%	93.80%
JP51-Tconv_combined	PolyA	93.70%	94.40%
JP51-Treg_combined	PolyA	93.10%	94.50%
JP53-Tconv_combined	PolyA	94.20%	94.30%
JP53-Treg_combined	PolyA	93.60%	93.80%
JP56-Tconv_combined	PolyA	93.80%	94.30%
JP56-Treg_combined	PolyA	93.90%	94.20%
JP61-Tconv_combined	PolyA	92.10%	93.30%
JP61-Treg_combined	PolyA	93.00%	93.20%
JP62-Tconv_combined	PolyA	93.80%	94.10%
JP62-Treg_combined	PolyA	93.10%	93.20%
JP73-Tconv_combined	PolyA	94.00%	94.40%
JP73-Treg_combined	rRNA depletion	71.40%	77.90%
JP74-Tconv_combined	PolyA	94.20%	94.40%
JP74-Treg_combined	PolyA	91.80%	91.90%
JP75-Tconv_combined	PolyA	94.40%	94.60%
JP75-Treg_combined	PolyA	91.20%	92.40%
JP8-Tconv_combined	PolyA	92.90%	93.70%
JP8-Treg_combined	rRNA depletion	61.60%	80.80%
JP9-Tconv_combined	PolyA	93.10%	93.70%
JP9-Treg_combined	rRNA depletion	7.60%	84.00%

5.4.3.3 Case-control differential expression analysis

I profiled the transcriptome of T1D patients and healthy donor subjects using RNA-seq and identified 352 and 337 differentially expressed (DE) genes between T1D and healthy control cohorts for Treg and Tconv cells, respectively (Figure 5-38, Figure 5-39, Table 7-7, Table 7-8). Several T1D-associated genes (compiled from NHGRI-EBI GWAS Catalog [3] and Onengut-Gumuscu, Chen [2]) were differentially expressed between T1D and control cohorts, including *GSDMA* (Gasdermin A) and *SPEF2* (Sperm Flagellar 2) in Treg cells (Figure 5-38) and *IL2* (Interleukin-2), *ICOSLG* (Inducible T Cell Costimulator Ligand), *DDC* (Dopa Decarboxylase) and *C1QTNF6* (Complement C1q tumor necrosis factor-related protein 6) in Tconv cells (Figure 5-39). Besides those known T1D-associated genes, a subset of genes linked with other autoimmune diseases are also differentially expressed, including *IL5* (Interleukin-5), *FOSL2* (Fos-related antigen 2) and *TIGIT* (T Cell Immunoreceptor With Ig And ITIM Domains) in Treg cells, and *MAF* (MAF BZIP Transcription Factor), *RORC* (RAR-related orphan receptor C) and *IL17F* (Interleukin 17F) in Tconv cells. Consistent with Gao, Uzun [33] study that performed RNA-seq experiment on fresh whole Treg cells obtained from 6 recent onset T1D patients and 5 healthy controls, *IL5* demonstrated lower expression in the T1D cohort compared with healthy controls. *IL5* was shown to promote induction of antigen-specific Treg cells upon activation by *IL-4* (of which expression was downregulated in our T1D group too) to suppress autoimmunity and reduced expression of *IL5* is likely to compromise Treg-dependent immune homeostasis [273]. *FOSL2* was identified as a T1D associated risk locus in a large-scale meta-analysis of genome-wide association studies conducted by Bradfield, Qu [86] in 2011 and it has also been shown to affect the expression of FoxP3 and other Treg development genes in mouse model [274]. Loss of *Maf*, which was downregulated in the Tconv cells of T1D group, has been shown to promote islet inflammation [275]. Maf has also been shown to have a critical role in enforcing Th17 effector identity by promoting chromatin accessibility and expression of key Th17 program genes, notably *Rorc* and *Il17* [163], which were both downregulated in our T1D group, and Maf deficiency in T cells was linked with dysregulation of Treg – Th17 balance leading to spontaneous colitis [276].

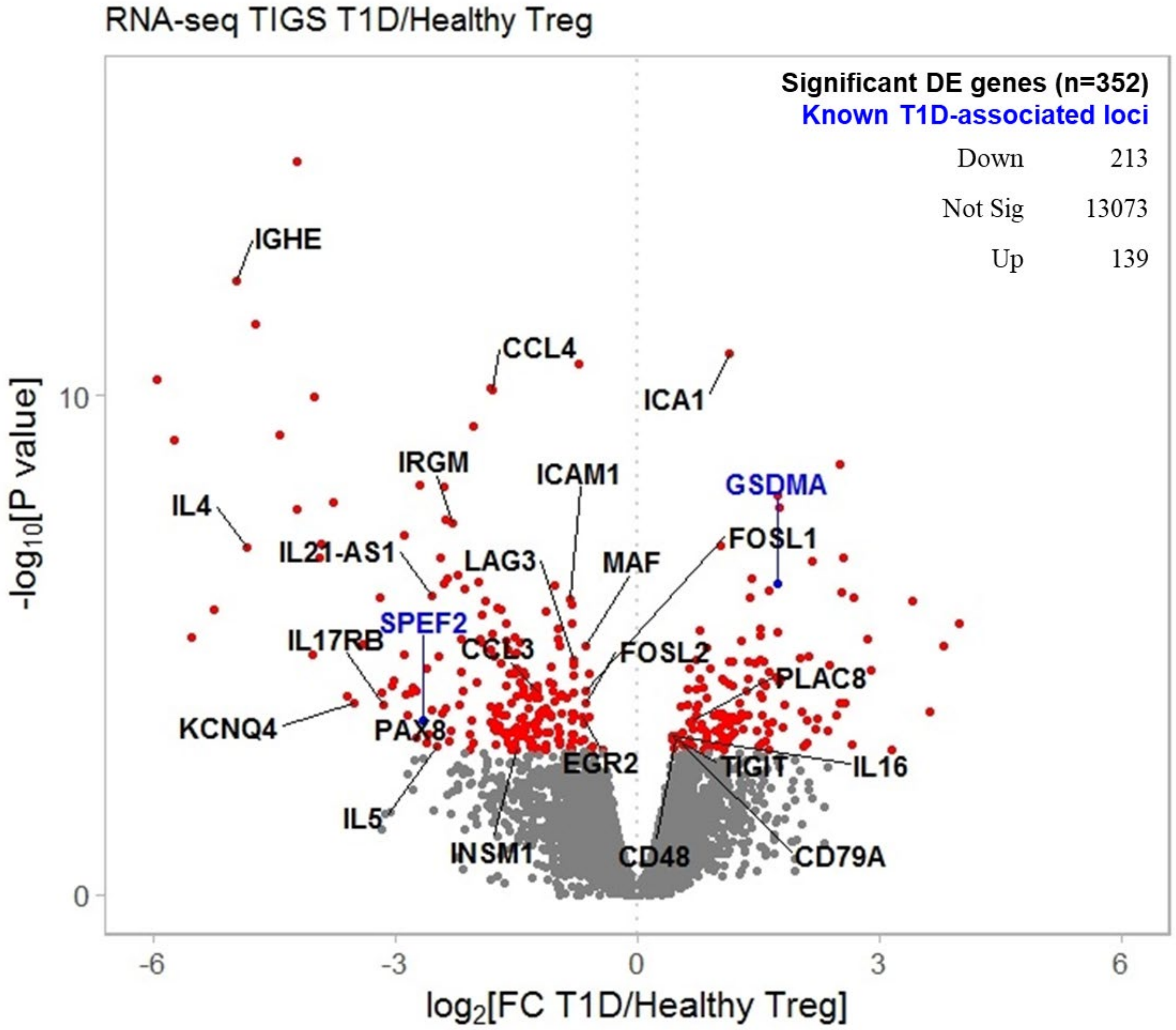


Figure 5-38 | Volcano plot showing differential gene expression (DGE) analysis between Treg from 9 T1D and 9 sibling-matched healthy subjects.
Genes having significantly differential gene expression (Benjamin Hochberg FDR < 0.05) are coloured red or blue. Differentially expressed genes in blue were previously known to be associated with T1D [2, 10]. Selected differentially expressed immune-relevant genes were annotated with gene symbols.

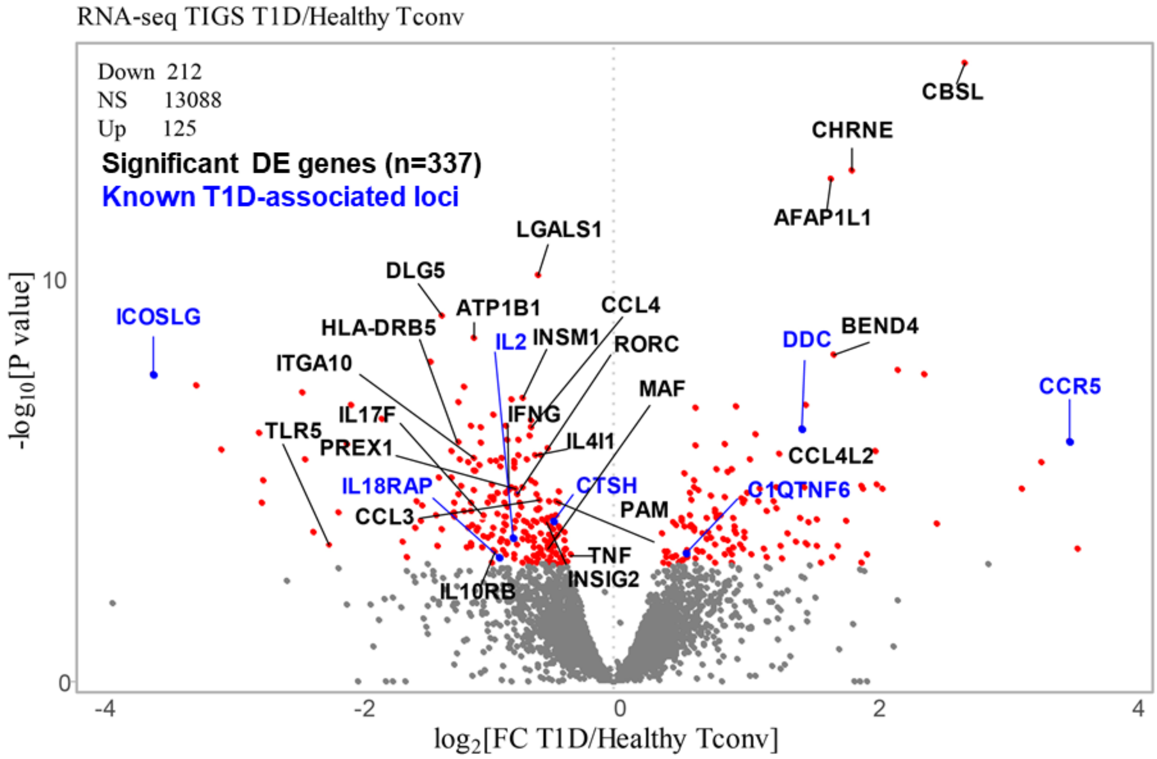


Figure 5-39 | Volcano plot showing differential gene expression (DGE) analysis between Tconv from 11 T1D and 11 sibling-matched healthy subjects.

Genes having significantly differential gene expression (Benjamin Hochberg FDR < 0.05) are coloured red or blue. Differentially expressed genes in blue were previously known to be associated with T1D [2, 10]. Selected differentially expressed immune-relevant genes were annotated with gene symbols.

5.4.3.4 Pathway analyses

Gene Set Enrichment Analysis (GSEA) identified several statistically significant overrepresented Hallmark gene sets such as those associated with NF-kB regulation in response to TNF, IL-2, IL-4 and IL-13 signalling in the case-control Treg RNA-seq dataset (Figure 5-40), whereas gene sets associated with hypoxia, glycolysis and effector memory T cells phenotype were enriched in the Tconv RNA-seq (Figure 5-41). Consistent with the pathogenic role of Tconv cells in T1D, ontology analyses for the differentially expressed (DE) genes (FDR < 0.05) between T1D and healthy Tconv cells also identified pathways associated with chemokine receptor-mediated chemotaxis (Figure 5-42), diabetes-linked voltage-gated potassium channels (Figure 5-43), autoimmune disorder such as inflammatory bowel disease (IBD) and cytokine-cytokine receptor interaction (Figure 5-44, Figure 5-45). The interaction between the interconnected networks on IBD and cytokine receptor interaction in the Tconv RNA-seq involved various TFs and cytokines such MAF, RORC, IFNG and IL10 receptor (Figure 5-45).

A subset of the dysregulated DE genes identified in Tconv cells were implicated in multiple autoimmune disorders (Figure 5-46, Figure 5-47). Although expression variability was observed in the RNA-seq samples, the expression patterns for genes such as *C1QTNF6* and *IL2* demonstrated high agreement and consistency across most of the case-control matched pairs (Figure 5-46). The implication of T1D dysregulated genes in multiple autoimmune disorders suggests they regulate a complex shared network of interconnected molecular and functional pathways in T cells leading to the loss of tolerance and induction of autoimmunity (Figure 5-47).

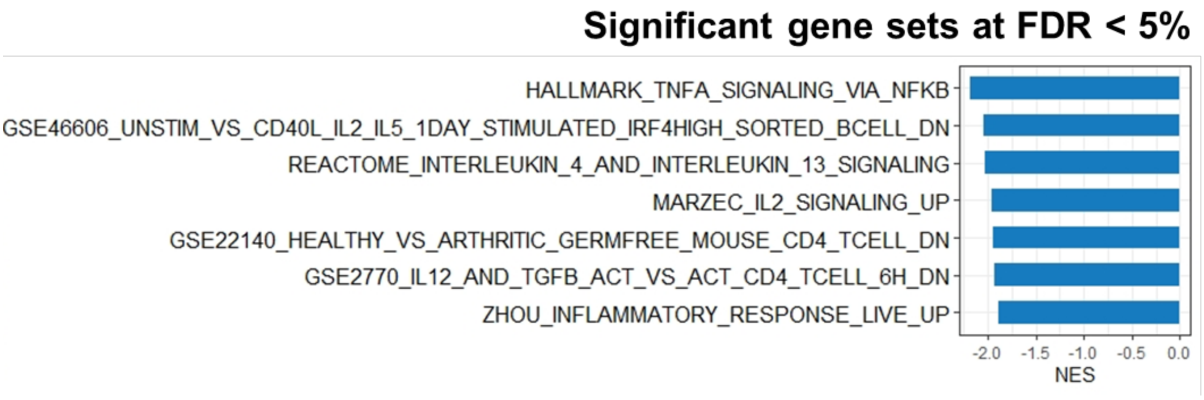


Figure 5-40 | Gene set enrichment analysis (GSEA) showing significantly enriched (FDR < 0.05) gene sets between T1D and healthy Treg RNA-seq.
Results are representative of Treg expression profiles from 9 T1D and sibling-matched healthy subjects.

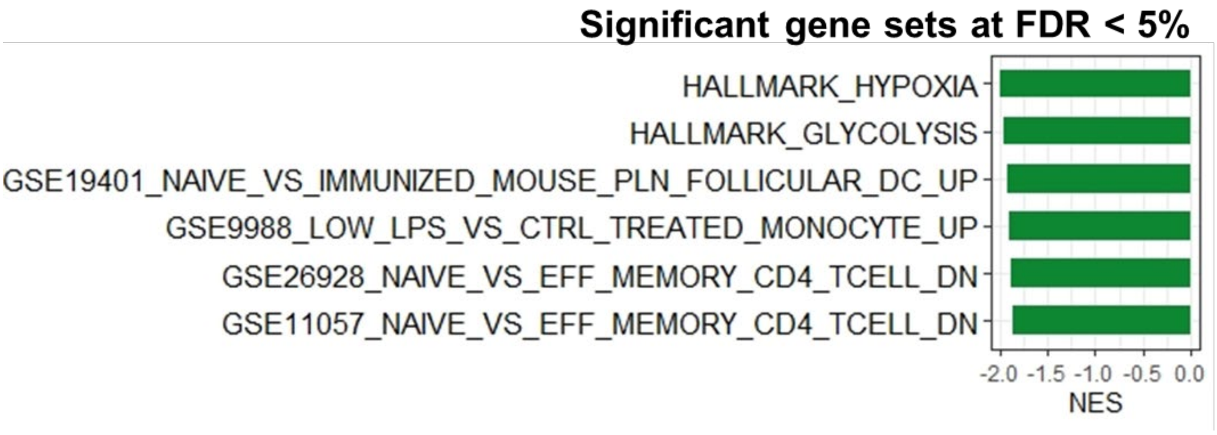


Figure 5-41 | Gene set enrichment analysis (GSEA) showing significantly enriched (FDR < 0.05) gene sets between T1D and healthy Tconv RNA-seq.
Results are representative of Tconv expression profiles from 9 T1D and sibling-matched healthy subjects.

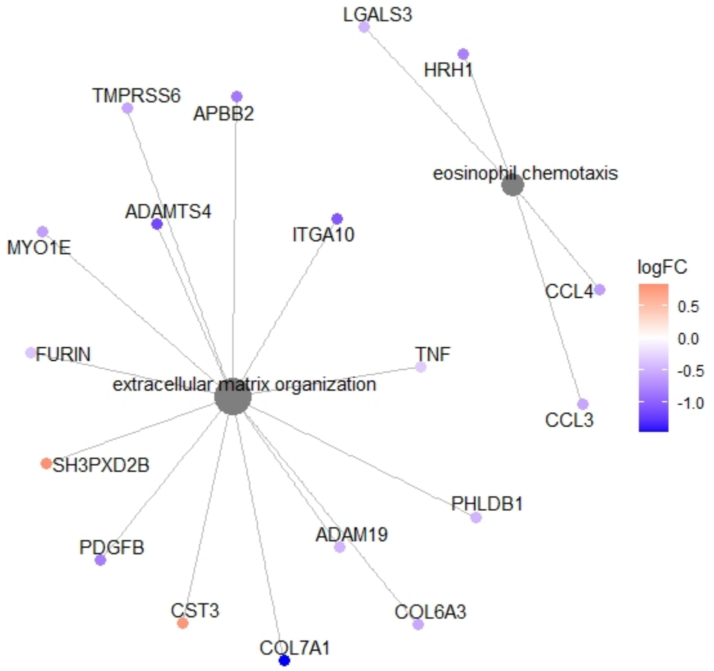


Figure 5-42 | Enrichment plot showing the most significant GO terms and related differentially expressed genes from the Biological Process Ontology for T1D vs healthy Tconv RNA-seq, using a Bonferroni adjusted p-value < 0.05 as the criteria for significance.
Results are representative of accessibility profiles from 11 T1D and sibling-matched healthy subjects.

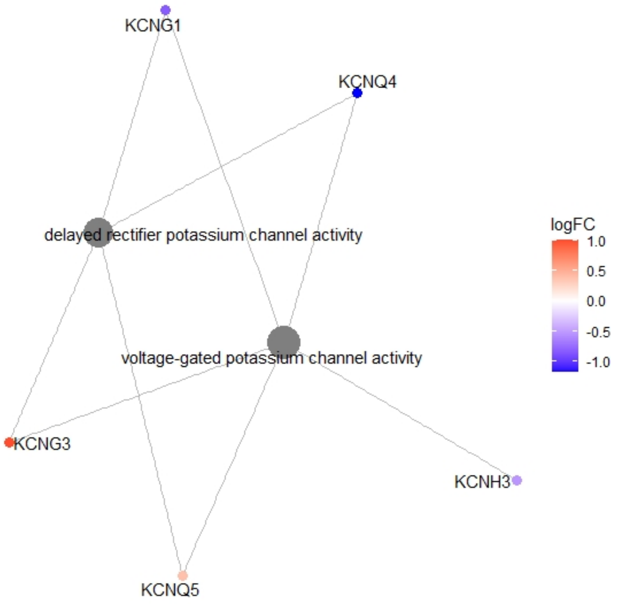


Figure 5-43 | Enrichment plot showing the most significant GO terms and related differentially expressed genes from the Molecular Function Ontology for T1D vs healthy Tconv RNA-seq, using a Bonferroni adjusted p-value < 0.05 as the criteria for significance.
Results are representative of accessibility profiles from 11 T1D and sibling-matched healthy subjects.

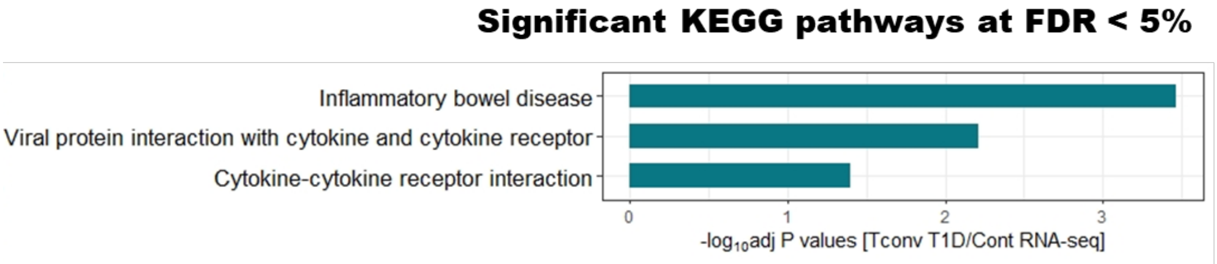


Figure 5-44 | KEGG pathway analysis showing significantly enriched pathways associated with differentially expressed genes between T1D and healthy Tconv RNA-seq, using a Bonferroni adjusted p-value < 0.05 as the criteria for significance.
Results are representative of accessibility profiles from 11 T1D and sibling-matched healthy subjects.

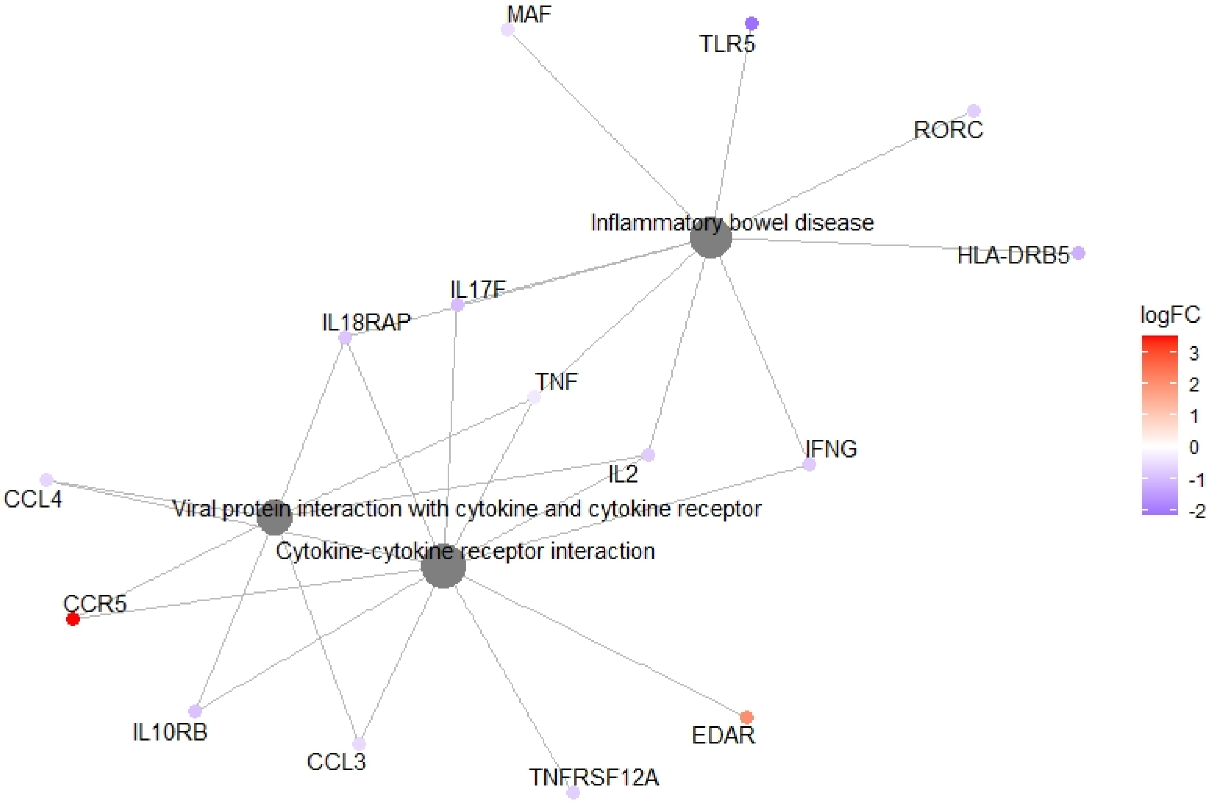

MAF
TLR5
RORC
Inflammatory bowel disease
HLA-DRB5
IL17F
IL18RAP
TNF
IFNG
IL2
CCL4
Viral protein interaction with cytokine and cytokine receptor
Cytokine-cytokine receptor interaction
CCR5
IL10RB
CCL3
EDAR
TNFRSF12A
logFC
3
2
1
0
-1
-2

Figure 5-45 | Enrichment plot showing the most significant KEGG pathways and associated genes for T1D versus healthy Tconv RNA-seq.

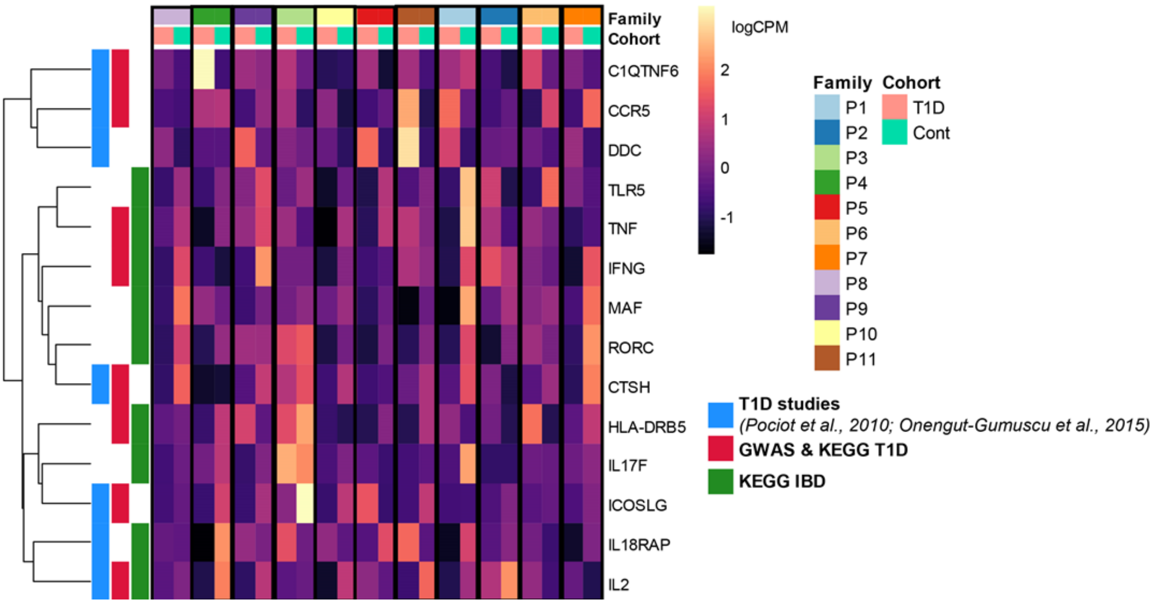


Figure 5-46 | Heatmap showing the expression patterns of differentially expressed genes associated with T1D and Inflammatory Bowel Disease (IBD) from different databases in the Tconv cells across 11 pairs of T1D and sibling-matched healthy subjects.

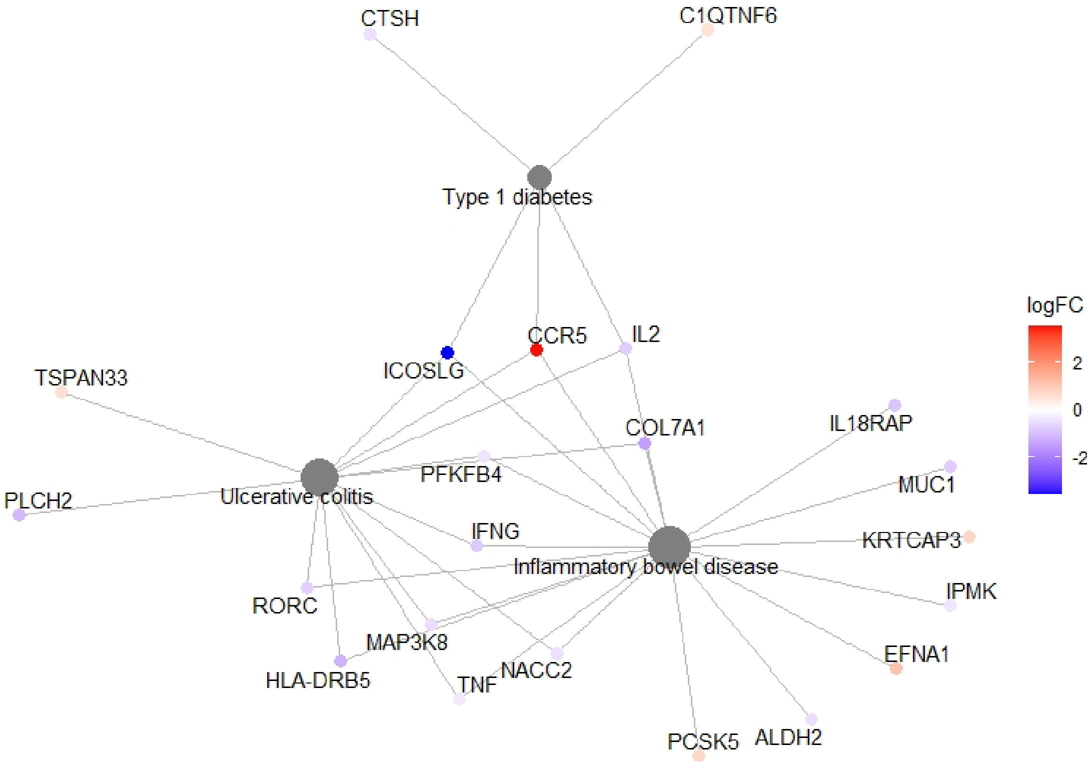


Figure 5-47 | Enrichment plot showing the interaction of differentially expressed genes associated with Inflammatory bowel disease, T1D and Ulcerative colitis.

Several differentially expressed genes are implicated in multiple autoimmune disorders, suggesting regulation of common or shared network of interconnected molecular and functional pathways leading to the loss of immune tolerance.

Genes associated with specific autoimmune diseases were obtained from GWAS database.

5.4.4 Integration of ATAC-seq data with Hi-C (chromatin conformation capture) contact maps to identify altered genes in T1D

To gain an insight into epigenetic regulation of the transcriptome, I correlated chromatin accessibility with gene expression by annotating ATAC-seq peaks differentially accessible (DA) between T1D and healthy controls to the nearest transcription start sites (TSS) in the linear genome sequence (Figure 5-48, Figure 5-49). Using this annotation approach, a small subset of significant ATAC-seq peak signal (3%) corroborated gene expression observations, for instance, *FOSL2* (Fos-related antigen 2), which was added to the growing repertoire of gene networks predisposing to T1D in a meta-analysis of genome-wide association studies [86], demonstrated reduced levels of local chromatin accessibility and transcripts in Treg cells from the T1D cohort. For Tconv cells, enrichment was observed in genes that showed both reduced local accessibility and transcript levels in the T1D cohort, for instances, *IL4I1* (Interleukin 4 Induced 1), which was recently shown to exert immunosuppressive function by suppressing the proliferation and differentiation of T cells and the proliferation of B cells [277], and *MAF*, which was reported to have a protective role against islet inflammation [275] and a key enforcer of Th17 effector identity [163].

It is well established in the research literature that long-range interactions between distal regulatory elements such as non-coding gene enhancers and promoters play a key role in transcriptional regulation and enhancers do not always link to the closest gene by linear distance. Linear distance does not accurately infer regulatory contacts as enhancer can bypass nearby genes to regulate genes far apart in the linear genome sequence but are proximate in 3D space within the nucleus. This complexity has always been overlooked when assigning target genes to distal non-coding regulatory elements like gene enhancers. Furthermore, the majority of the GWAS-identified disease-associated risk SNPs are located within non-coding regions of the genome [2, 88] and thus, they can alter gene expression through spatial interactions involving distal loci. This orientation-independent and 3D DNA looping enhancer-promoter interaction mechanisms makes it difficult to predict what gene is directly regulated by an enhancer.

As the query peak regions I used for conducting differential accessibility analysis between T1D and healthy controls comprised genomic regions corresponding to TSS, super- and typical-enhancers of CD4^{+} T cells [7, 8], here in this section of analysis, instead of assigning the T1D-associated enhancers to their nearest TSS in close physical proximity, I linked them to their putative target gene promoters in the 3D nuclear space by incorporating several published T cell interactome datasets (listed in Table 5-1) generated by chromosome conformation capture (3C) technology in healthy individuals (see schematics in Figure 5-3 and Figure 5-4). This was necessary because ATAC-seq does not provide evidence that formally links distal regulatory elements to their target genes. These high resolution 3C datasets include: Hi-C contact map [134] generated from Treg cells, which captures essentially all chromatin interactions in the nucleus; H3K27ac HiChIP contact maps [37] generated from Treg and Th17 which captures enhancer interactions; and promoter capture Hi-C (pcHi-C) contact map [6] generated from activated CD4^{+} T cells, which enriches for promoter interactions (Table 5-1). I included only significant, high-confidence enhancer or promoter loops (e.g., CHiCAGO scores ≥ 5) from these published 3D contact maps to annotate the target genes regulated by the T1D-associated enhancer elements (Benjamini-Hochberg FDR < 0.05) identified from my case-control ATAC-seq data (see Section 5.3.3.7 for methodology). I then confirmed transcriptome changes of the identified target gene products using my case-control Treg or Tconv RNA-seq.

High correlation in enhancer connectome was observed between H3K27ac HiChIP [37] and Hi-C [134] experiment in the Treg cells, demonstrating strong data quality, reproducibility and credibility (Figure 5-50). 12.8% (n=87) of the significant differentially accessible ATAC-seq peaks identified in Treg cells of T1D patients were located within the enhancer-promoter anchor loops identified in H3K27ac HiChIP [37] and Hi-C [134] experiments (Figure 5-51). Whereas for Tconv cells, 5.2% (n=55) of the differentially accessible ATAC-seq peaks identified in T1D patients were captured by the H3K27ac HiChIP (Th17 cells) [37] and pcHi-C (activated CD4$^+$ T cell) [6] contact maps (Figure 5-53).

By incorporating published chromatin conformation capture (3C) contact maps generated from Hi-C experiment in Treg cells, H3K27ac HiChIP in Treg and Th17 cells, as well as promoter capture Hi-C in activated CD4$^+$ T cells into my case-control chromatin accessibility profiles using a stringent set of criteria, I was able to map target genes and assign molecular functions to altered enhancers in Treg and Tconv cells of T1D patients, which control genes beyond the neighbouring genes in the linear genome sequence (Figure 5-52, Figure 5-54, Table 5-8). These gene targets also exhibited alterations in transcript levels in T1D patients, suggesting epigenetic modifications in T1D interfere with the activity of those differentially accessible Treg- and Tconv-specific enhancers, resulting in altered expression of the genes normally controlled by these enhancers. Through integration of case-control ATAC-seq, RNA-seq and Hi-C based methods I identified 42 and 21 altered gene targets in Treg and Tconv cells of T1D cohort, respectively (Figure 5-52, Figure 5-54 and Table 5-8). Overall, 70% of the gene targets identified from this approach were not picked up from just linear annotation alone using ATAC-seq and RNA-seq. Some of the altered gene targets I identified included *FOSL2*, *TIGIT* and *ICAM1* in Treg cells (Figure 5-52) and *MAF* and *IL2* in Tconv cells (Figure 5-54). *ICAM1* was revealed to confer susceptibility to the development of T1D in a meta-analysis of genome-wide association study conducted by Ma, Möllsten [278]. Pathway enrichment analyses revealed that gene targets identified in the Treg dataset are strongly associated with IL2-STAT5 signalling, whereas gene targets from Tconv are enriched in pathways associated with known T1D signatures and Th2 differentiation (Figure 5-59).

Integration of 3D genome interactome data with the case-control Tconv and Treg ATAC-seq and RNA-seq allows accurate identification of altered enhancer targets in T1D. For instance, as shown in Figure 5-55 (a), an ATAC-seq peak located within a T cell enhancer [7] was differentially accessible (DA) in Treg cells between T1D cohort and healthy controls, and DRD3 was the closest gene in physical proximity to the DA peak in linear genome distance (default annotation). However, the 3D connectivity map from Hi-C showed that this DA peak region was physically contacting a distant promoter, *TIGIT*, located more than 70kb away (Figure 5-55; a). The observation was confirmed by case-control RNA-seq in which alterations to the transcript level of *TIGIT* was detected (FDR < 0.05) in the T1D cohort (Figure 5-55; b). As the DA peak region appeared to overlap with a group of TF occupancy footprints (Table 7-9) identified from ATAC-seq, including YY1 whose binding was shown to be disrupted in T1D [33], as well as a T1D risk-associated risk variant (Table 5-8). One likely mode of mechanism is that the genetic variant disrupts the TF binding site(s) and/or interferes with enhancer function by altering chromatin accessibility, resulting in altered expression of *TIGIT* in Treg cells of T1D patients.

It was recently shown in Gao, Uzun [33] that a SNP (rs883868) in high linkage disequilibrium with the T1D-associated GWAS lead SNP disrupts the binding of Yin and Yang 1 (YY1) TF to a Treg enhancer in T1D, resulting in the loss of long-range enhancer–promoter interaction. Integration of case-control ATAC-seq and published H3K27ac HiChIP [37] revealed that an altered peak located within the T-cell super enhancer region physically contacted the promoter of *YY1* rather than its nearest gene in linear space, *WARS* (Tryptophanyl-TRNA Synthetase), and that long-range enhancer-promoter interaction was cell type-restricted and conserved only in Treg cells (Figure 5-56).

The same observation was captured in the Tconv cells at *MAF* (Figure 5-57) and *IL2* (Figure 5-58) gene loci. For instance, two ATAC-seq peaks located within the gene body of *WWOX*

(WW Domain Containing Oxidoreductase) and T cell enhancer regions were differentially accessible in Tconv cells between T1D cohort and healthy controls. Th17 enhancer HiChIP [37] and CD4$^+$ T cell pcHi-C [6] showed that the DA peak regions were contacting a distant promoter, *MAF*, which was located more than 700kb away. The observation was confirmed by case-control RNA-seq in which alterations to the transcript level of *MAF* was detected (FDR < 0.05) in the T1D cohort. The observation that the DA peak regions also harboured T1D risk-associated SNPs suggests differential regulation of the interacting enhancer is likely to be driven by local genetic variants through altering chromatin accessibility, resulting in altered expression of *MAF* in Tconv cells of T1D patients.

Overall, this approach also uncovered the distal regulatory control regions which may, upon functional validation, prove to be critical for shaping the gene expression and thus function of Treg and Tconv cells in T1D. Over the past two decades GWAS [3] have provided valuable insights into the genetic architecture of complex diseases, however, they focus exclusively on statistical evidence and disregard the biological plausibility. Using a combinatorial approach, I identified a small subset of altered gene targets that have already been associated with T1D risk by genome-wide association studies, supporting the claim that this proof-of-concept combinatorial, integrative evidence-based approach that leveraged multiple functional annotated features in the disease setting is a powerful approach for linkage probabilities between non-coding regulatory elements and biological phenotypes in informing disease biology.

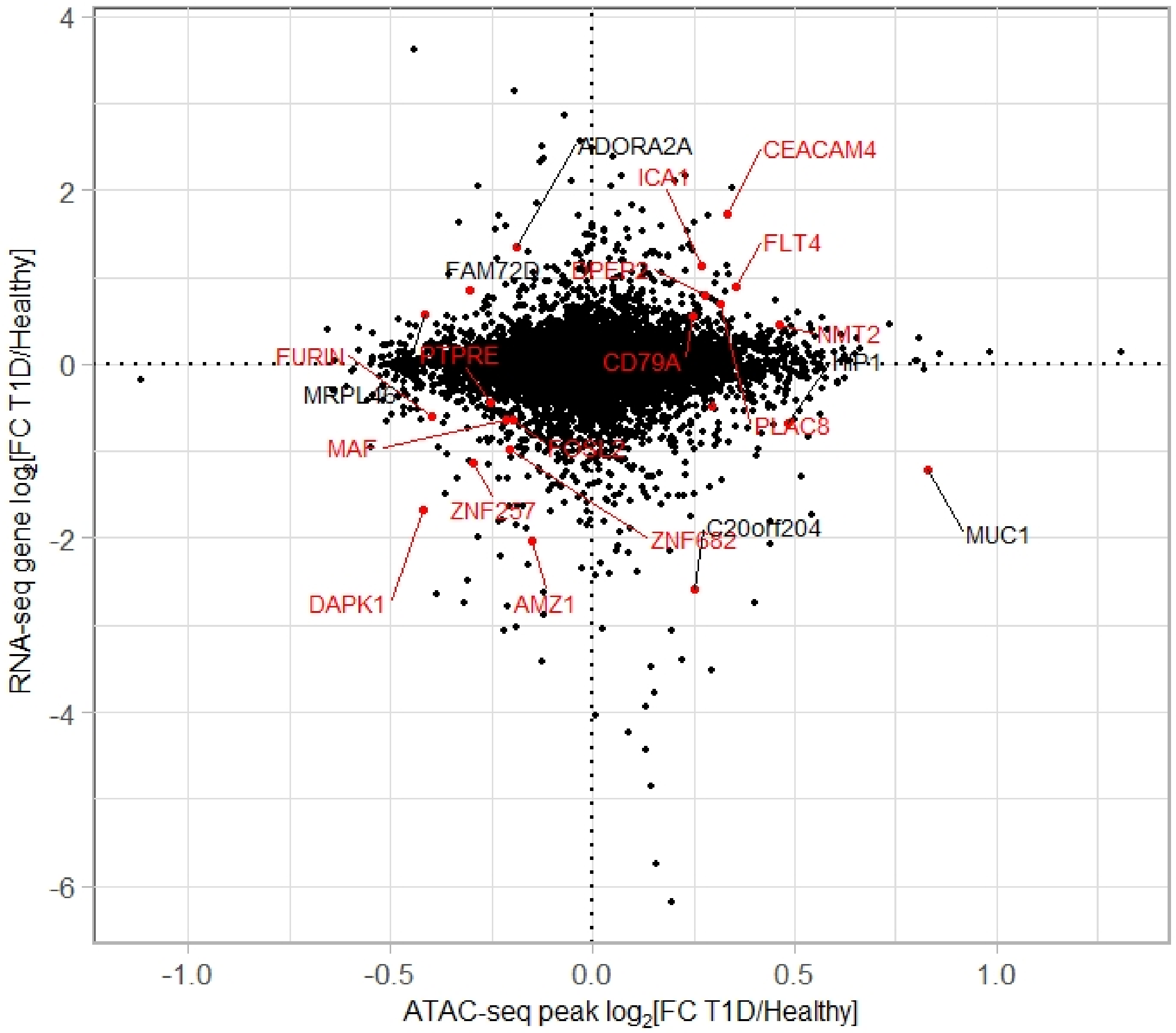


Figure 5-48 | Scatterplot showing the intersection between differentially accessible chromatin regions (ATAC-seq) and differentially expressed genes (RNA-seq) between T1D and healthy Treg cells.

Differentially accessible ATAC-seq peaks were annotated to the nearest TSS in linear genomic distance. Results are representative of 12 pairs of accessibility- and 9 pairs of expression-profiles from T1D and sibling-matched healthy subjects. Genes having Benjamin Hochberg FDR < 0.1 in both datasets are annotated, with representative genes demonstrating same directionality in both ATAC-seq and RNA-seq (i.e up or down in both ATAC-seq and RNA-seq) labelled in red font.

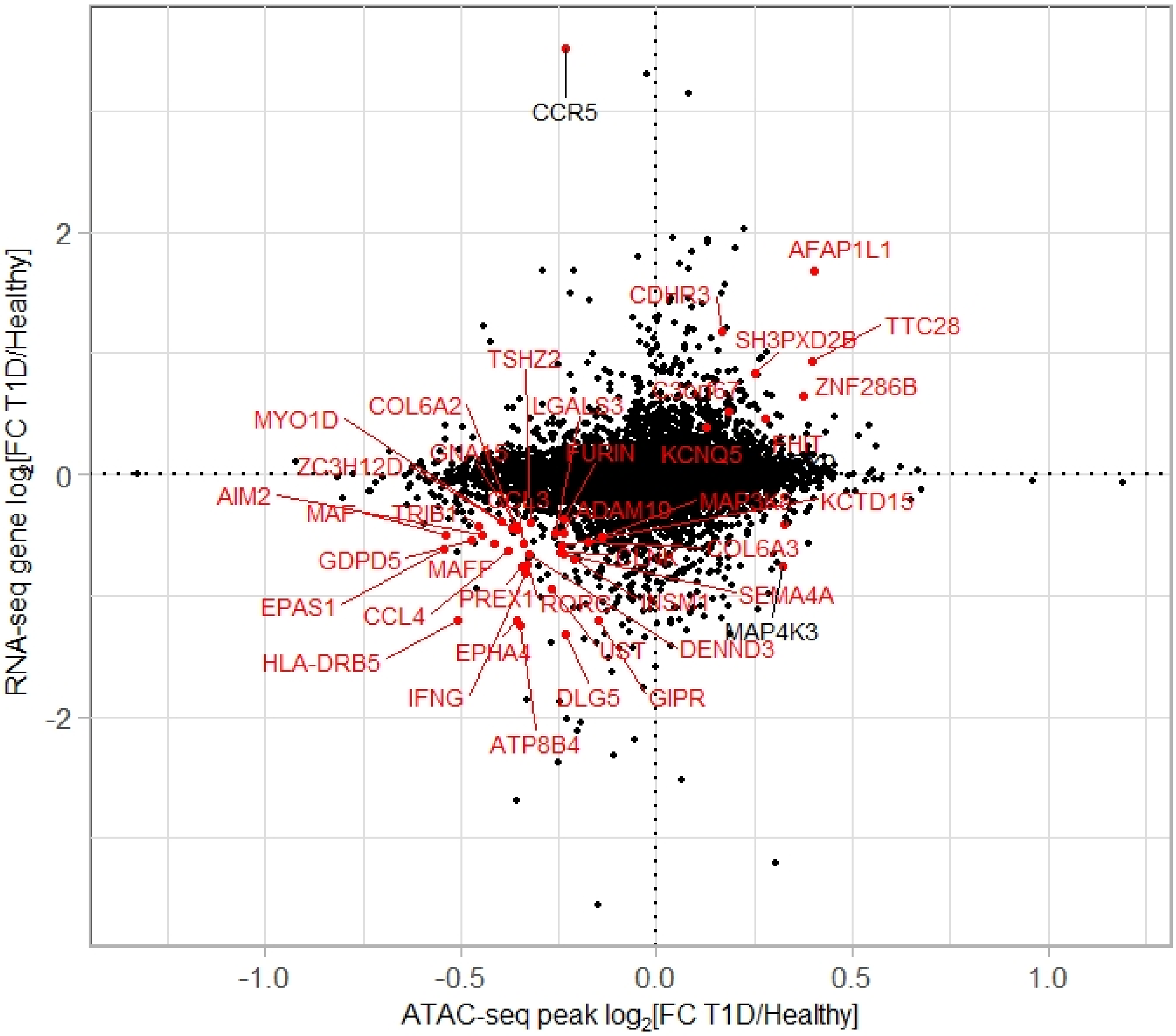


Figure 5-49 | Scatterplot showing the intersection between differentially accessible chromatin regions (ATAC-seq) and differentially expressed genes (RNA-seq) between T1D and healthy Tconv cells.

Differentially accessible ATAC-seq peaks were annotated to the nearest TSS in linear genomic distance. Results are representative of 12 pairs of accessibility- and 11 pairs of expression-profiles from T1D and sibling-matched healthy subjects. Genes having Benjamin Hochberg FDR < 0.1 in both datastes are annotated, with representative genes demonstrating same directionality in both ATAC-seq and RNA-seq (i.e up or down in both ATAC-seq and RNA-seq) labelled in red font.

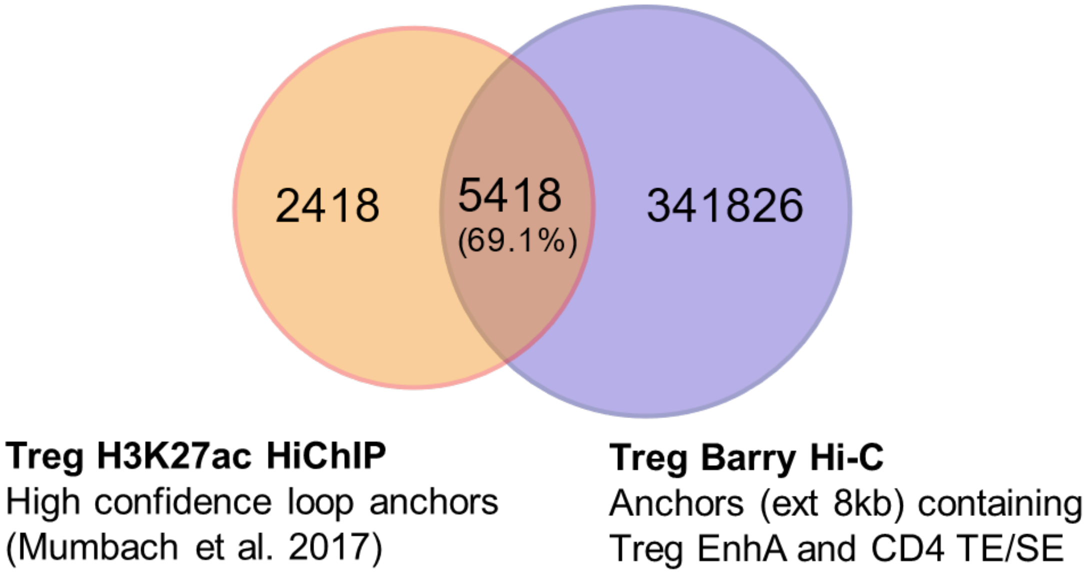


Figure 5-50 | The overlap of Treg enhancer connectome between HiChIP from Mumbach, Satpathy [37] study and Hi-C contact map from Barry Lab.
70% of enhancer- and promoter-associated loop anchors (histone modification H3K27ac HiChIP) identified in Mumbach, Satpathy [37] were captured in Hi-C contact map from Barry Lab. Both datasets were incorporated in my Treg case-control ATAC-seq and RNA-seq data to map the landscape of enhancer-promoter regulatory interactions in the T1D setting.

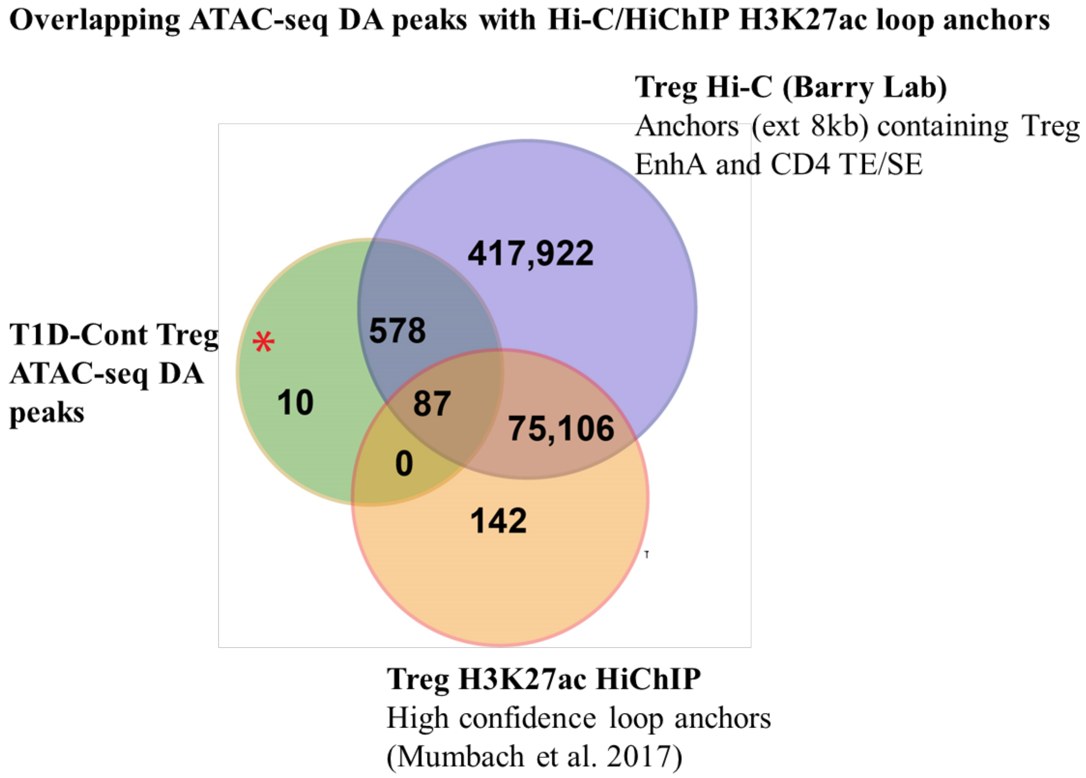
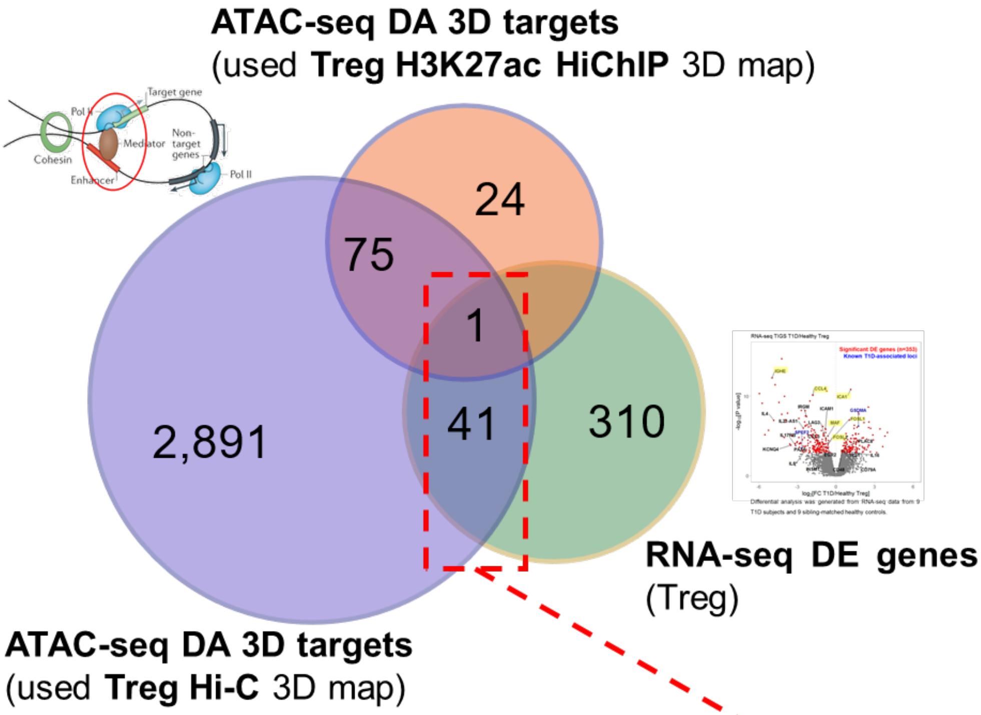


Figure 5-51 | Overlap of ATAC-seq differentially accessible peaks between T1D and healthy controls with enhancer- and promoter-associated interactome in Treg cells. 12.8% (87) of differentially accessible chromatin regions identified between T1D and healthy Treg cells from ATAC-seq (n=12) was within the Treg enhancer-promoter interaction anchors captured in HiChiP [37] and Hi-C contact maps.

Red asterisk indicates the “driver” dataset for intersection of genomic regions, for e.g., 87 case-control differentially accessible peaks were found to overlap the enhancer HiChIP loop anchors from Mumbach, Satpathy [37].

Cont, control; DA, differentially accessible.



RAP1GAP
FAM72D
LY9
SEMA4A
NMT2
CHRNA10
UCP3
TMEM45B

TNFSF11
GRTP1
LINC00221
SMG1P3
MAF
DPEP2
HIC1
LRG1

ICAM1
ZNF682
CD79A
CCDC155
FOSL2
MYO1B
ZBP1
C20orf204

NOD1
STRIP2
HIP1
SAMD9
LRRN3
DAPK1
EGFL7
CDC14B
PLAC8

CYP2D7
MAFF
GPR15
TIGIT
ADAM19
RNF32
ICA1
AMZ1
PTCH1

Known T1D loci

TF, underlined

Figure 5-52 | Identification of altered gene targets in T1D Treg cells through integration of ATAC-seq, RNA-seq and HiChIP/Hi-C contact maps.

Enhancer connectome generated from Hi-C provides 3D spatial mapping of active enhancers to their target genes. Using Hi-C chromatin structure maps, the differentially accessible enhancers identified between T1D and healthy controls (n=12) were mapped to their target genes in 3D nuclear space (termed “3D targets”), instead of to the nearest TSS in the linear genome sequence. The transcript levels of those gene targets were validated by case-control RNA-seq.

DA, differentially accessible; DE, differentially expressed.

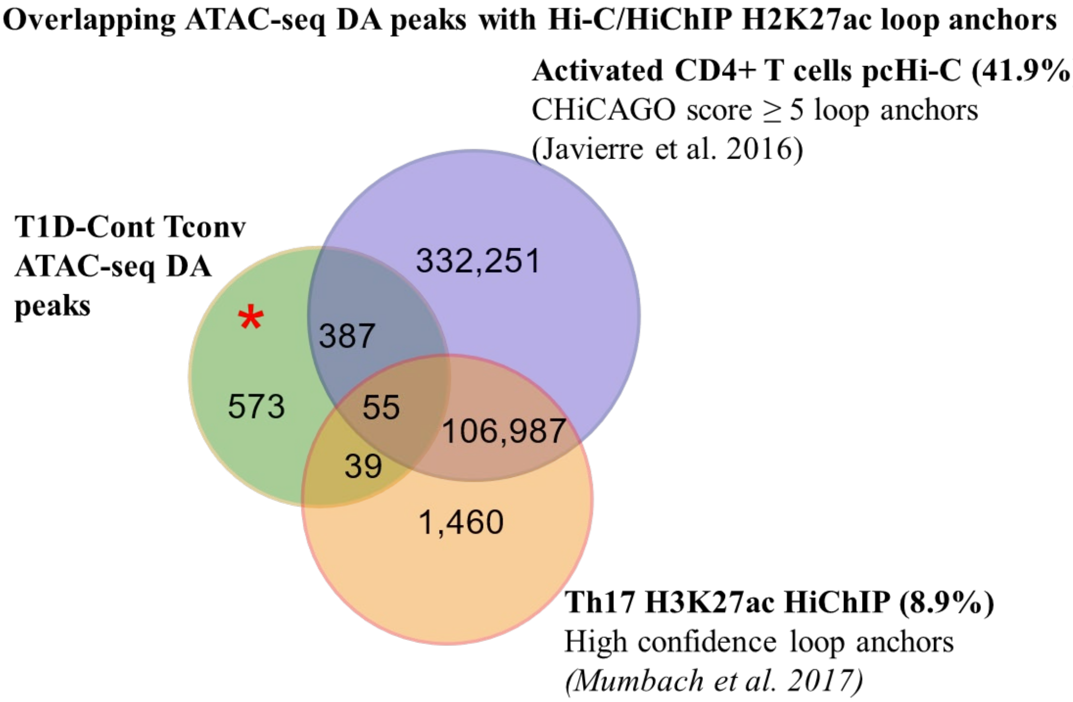
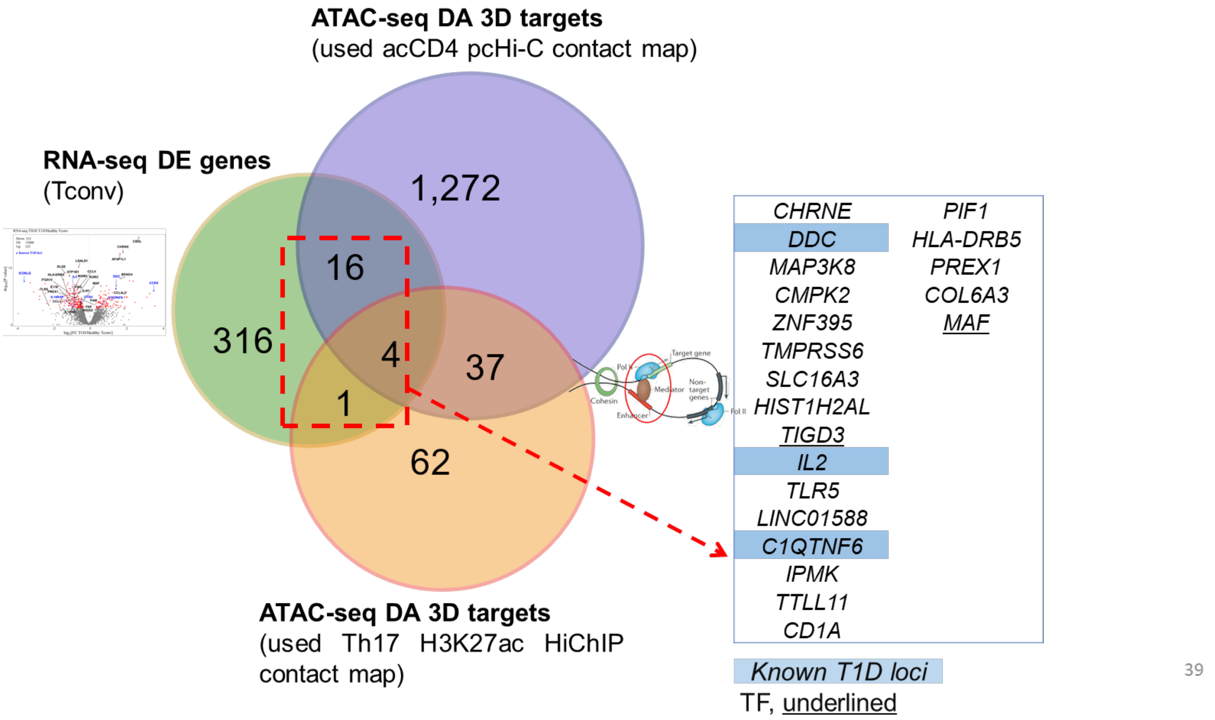


Figure 5-53 | Overlap of ATAC-seq differentially accessible peaks between T1D and healthy controls with enhancer- and promoter-associated interactome in Tconv cells. 5.2% (55) of differentially accessible chromatin regions identified between T1D and healthy Tconv cells from ATAC-seq (n=12) was captured by the Th17 enhancer interaction anchors (HiChiP) [37] and promoter interactome from activated CD4^{+} T cell (pcHi-C) [6].

Red asterisk indicates the "driver" dataset for intersection of genomic regions, for e.g., 94 case-control differentially accessible peaks were found to overlap the Th17 enhancer HiChIP loop anchors from Mumbach, Satpathy [37].

Cont, control; DA, differentially accessible; pc, promoter capture.



CHRNE
DDC
MAP3K8
CMPK2
ZNF395
TMPRSS6
SLC16A3
HIST1H2AL
TIGD3
IL2
TLR5
LINC01588
C1QTNF6
IPMK
TTLL11
CD1A

PIF1
HLA-DRB5
PREX1
COL6A3
MAF

Known T1D loci
TF, underlined

Figure 5-54 | Identification of altered gene targets in T1D Tconv cells through integration of ATAC-seq, RNA-seq and pcHi-C/HiChIP contact maps.
Promoter- and enhancer-associated connectome generated from promoter capture Hi-C and H3K27ac HiChIP provides 3D spatial mapping of active enhancers to their target genes. Using Hi-C chromatin contact maps, the differentially accessible enhancers identified between T1D and healthy controls (n=12) were mapped to their target genes in 3D nuclear space (termed “3D targets”), instead of to the nearest TSS in the linear genome sequence. The transcript levels of those gene targets were validated by case-control RNA-seq.

DA, differentially accessible; DE, differentially expressed, pc, promoter capture.

Table 5-8 | Tconv- and Treg- T1D targets identified through integration of ATAC-seq, RNA-seq and Hi-C datasets.
Listed gene targets were identified from connecting the case-control differentially accessible T cell enhancers to their target promoters using published Hi-C datasets. The expression pattern of the target genes was identified through case-control RNA-seq data.

Red and blue indicate upregulation and downregulation in T1D group, respectively. ATAC-seq peaks containing T1D risk variants (linkage disequilibrium of r2= 0.8 of GWAS T1D SNPs) were marked with an asterisk. Plus (+) signs indicate associations with other autoimmune disease as determined by GAAD (Gene and Autoimmiune Disease Association Database) [39].

Treg enhancer targets (ATAC-seq, Hi-C and RNA-seq)		
	ATAC direction	RNA direction
ADAM19	🟦	🟦
AMZ1	🟥	🟦
C20orf204	🟥 *	🟦
CCDC155	🟥	🟦
CD79A	🟥	🟥
CDC14B	🟦	🟦
CHRNA10	🟥 *	🟦
DPEP2	🟥	🟥
FOSL2	🟦	🟦
GPR15	🟥	🟥
GRTP1	🟥	🟥
HIC1	🟥	🟥
HIP1	🟥	🟦
ICA1	🟥	🟥
ICAM1	🟦	🟦
LRG1	🟥	🟦
LY9	🟥	🟥
MAF	🟥	🟦
MAFF	🟥	🟦
MYO1B	🟥	🟦
NOD1	🟥	🟥
PLAC8	🟥	🟥
PTCH1	🟦 *	🟦
RNF32	🟦	🟥
SAMD9	🟥	🟦
SEMA4A	🟦	🟥
SMG1P3	🟦	🟦
STRIP2	🟥	🟥
TIGIT	🟥 *	🟥
UCP3	🟥	🟦
ZBP1	🟥	🟥
ZNF682	🟦	🟦

Tconv enhancer targets (ATAC-seq, Hi-C and RNA-seq)		
	ATAC direction	RNA direction
C1QTNF6	🟦 *+	🟥
CD1A	🟦	🟥
CHRNE	🟦	🟥
CMPK2	🟥	🟥
COL6A3	🟥	🟦
HIST1H2AL	🟦	🟥
IL2	🟦	🟦
IPMK	🟥	🟦
LINC01588	🟦	🟦
MAF	🟦 *+	🟦
PIF1	🟦	🟦
PREX1	🟦 *+	🟦
SLC16A3	🟦	🟦
TIGD3	🟦	🟦
TLR5	🟦	🟦
TMPRSS6	🟦 *+	🟦
TTLL11	🟦	🟥

🟦	Down in T1D
🟥	Up in T1D

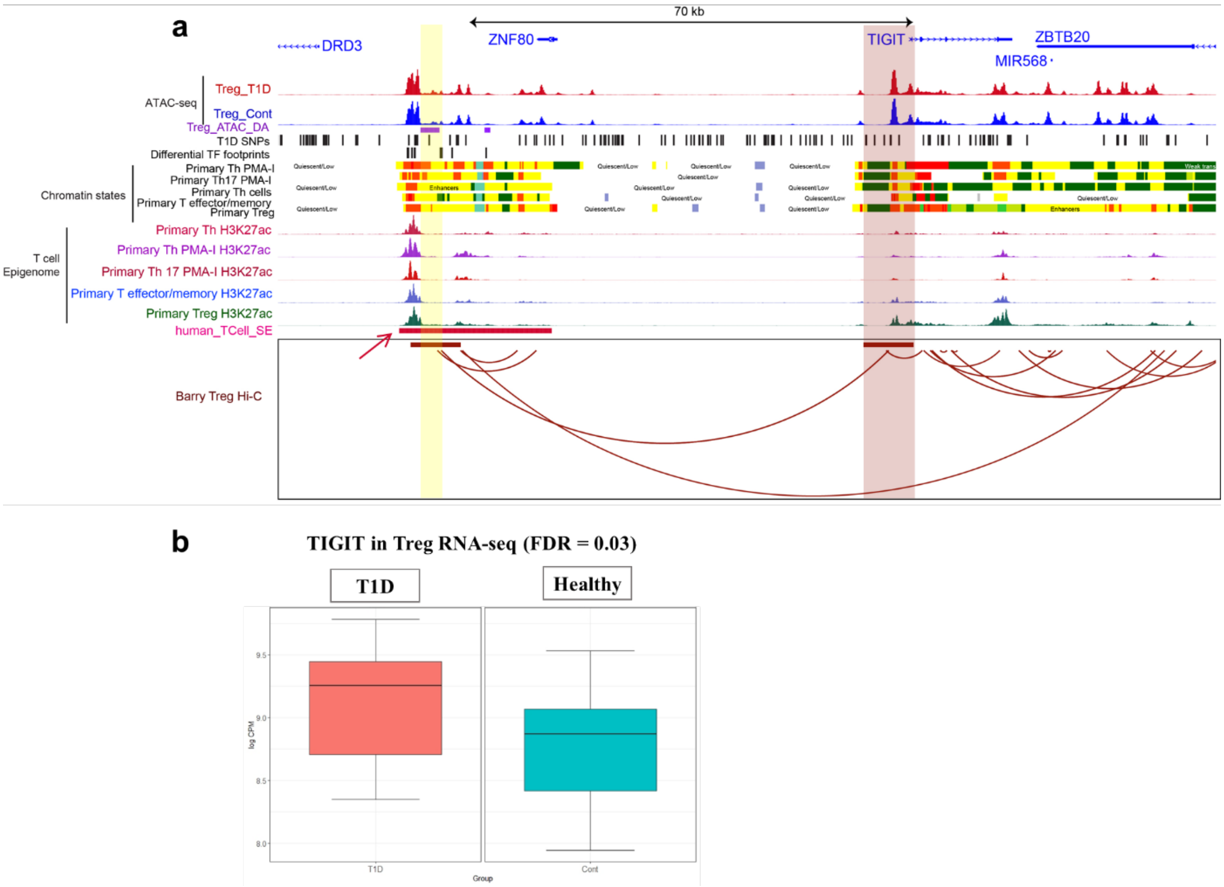


Figure 5-55 | Integration of 3D enhancer connectome from Hi-C contact maps links altered enhancer in T1D to *TIGIT* in Treg.

An ATAC-seq peak located within the T cell enhancer region (red arrow) was differentially accessible between T1D and healthy Treg cells (n=12). Although *DRD3* was the nearest gene to the differential peak in linear genome distance (default annotation), 3D connectivity map from Hi-C showed that the differential region is contacting a distant promoter, *TIGIT,* located more than 70kb away (a; highlighted in red). The observation was confirmed by case-control RNA-seq where *TIGIT* was shown to be differentially expressed (FDR = 0.03) between T1D and healthy controls (n=9).

Case-control ATAC-seq signal was intersected with T1D risk SNPs in high linkage disequilibrium (R^2=0.8) with the GWAS lead SNPs [3], differential TF footprints, T cell super enhancers [7], T cell epigenome and chromatin states (Roadmap Epigenomics Project [8]). Browser view was generated using WashU Epigenome Browser.

DA, differential accessibility.

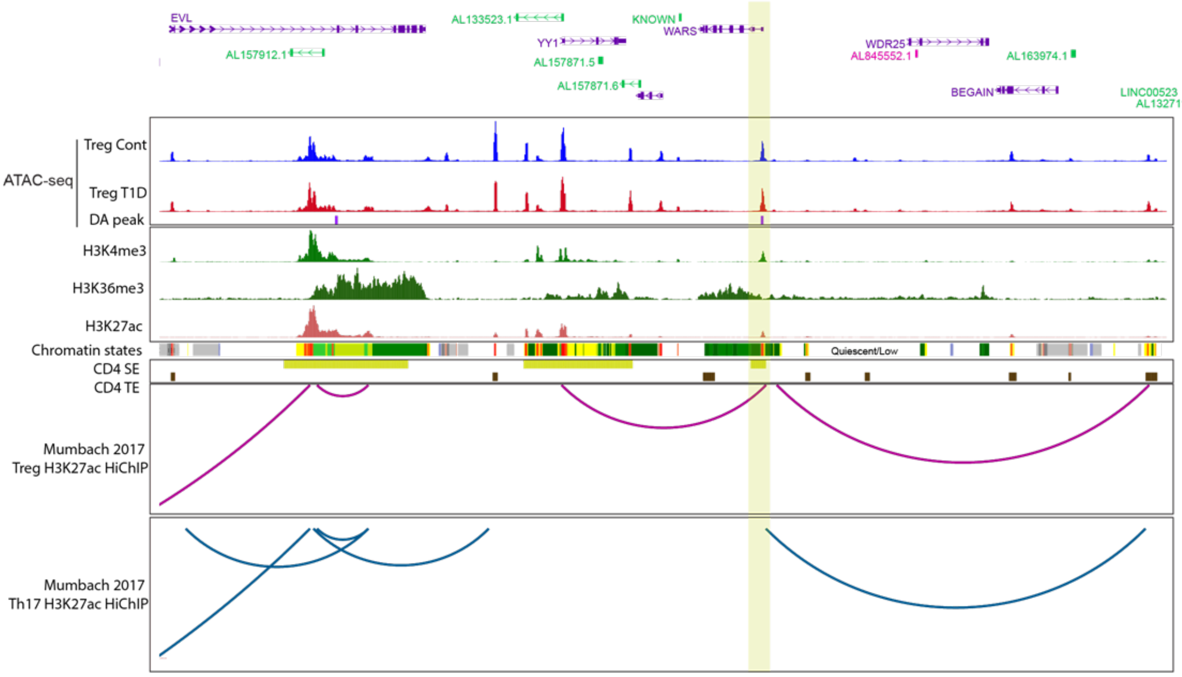


Figure 5-56 | 3D enhancer connectome captures putative cell-type-specific dyregulated chromatin contact in T1D.

An ATAC-seq peak located within the T cell enhancer region (highlighted in yellow) was differentially accessible between T1D and healthy Treg cells (n=12). Although *WARS* was the nearest gene to the differential peak in linear genome distance (default annotation), HiChIP enhancer interactome from Treg showed that the differential region is contacting a distant promoter, *YY1*. YY1 is a TF and it has been shown that T1D risk SNP disrupts its binding at Treg enhancer and results in loss of long-range enhancer–promoter interaction [33]. The same enhancer-promoter interaction was not conserved in the Th17 cells, suggesting cell-type-specific long-range gene regulation.

Cont, control; SE, super enhancer; TE, typical enhancer.

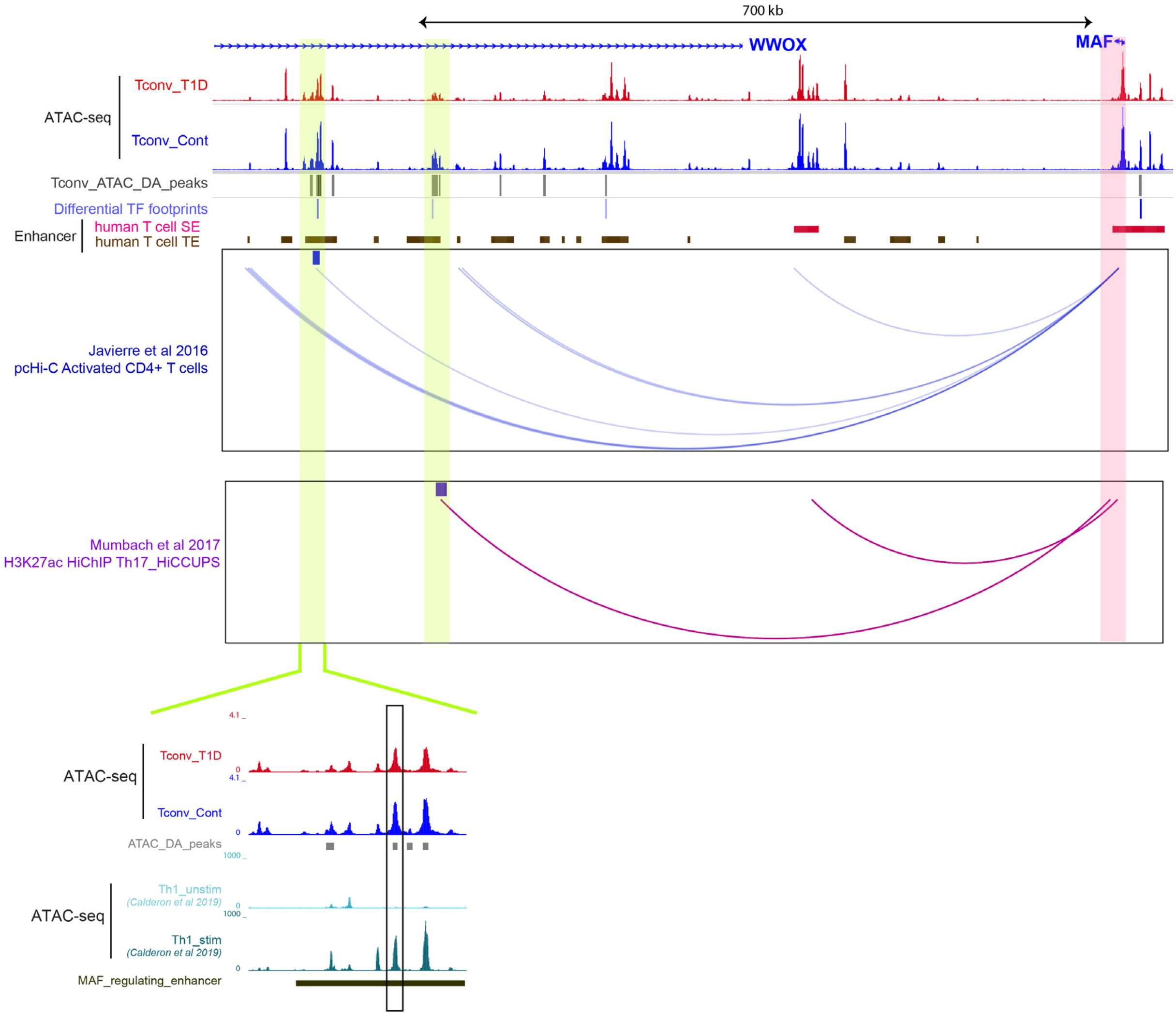


Figure 5-57 | Integration of 3D enhancer and promoter interactome connects altered enhancers in T1D to *MAF* in Tconv.

Two ATAC-seq peaks located within the gene body of *WWOX* and T cell enhancer regions (highlighted in yellow) were differentially accessible between T1D and healthy Tconv cells (n=12). 3D connectivity maps from Th17 enhancer HiChIP [37] and CD4^{+} T cell pcHi-C [6] showed that the differential regions are contacting a distant promoter, *MAF,* located more than 700kb away (highlighted in red). The observation was confirmed by case-control RNA-seq where *MAF* was shown to be differentially expressed (FDR = 0.027) between T1D and healthy controls (n=9) (Figure 5-39).

Case-control ATAC-seq signal was intersected with T1D risk SNPs in high linkage disequilibrium (R^2=0.8) with the GWAS lead SNPs [3], differential TF footprints, T cell super enhancers [7], T cell epigenome and chromatin states (Roadmap Epigenomics Project [8]). Browser view was generated using WashU Epigenome Browser.

Bottom panel; zoom-in of the enhancer region. Differentially accessible (DA) peaks identified in the T1D vs healthy control cohort and stimulation-dependent changes in the chromatin accessibility of this region in Th1 cells [45] are shown.

DA, differential accessibility.

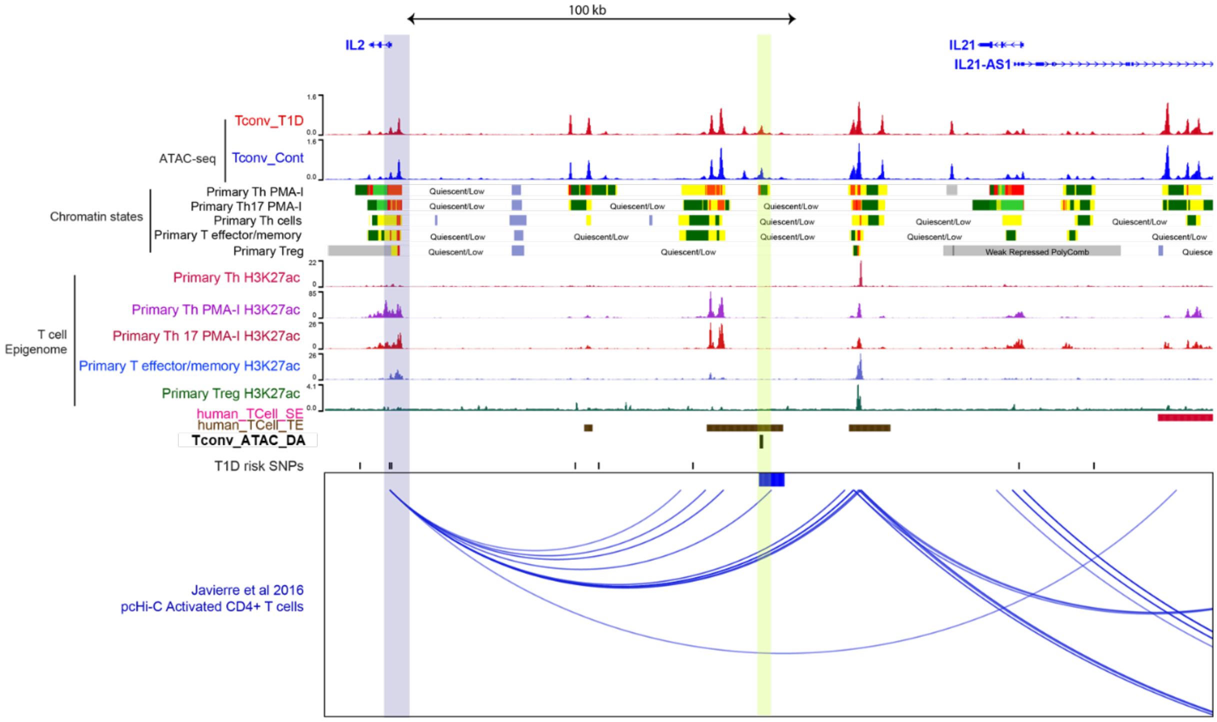


Figure 5-58 | 3D promoter connectome maps an altered enhancer in T1D to *IL2* in Tconv. An ATAC-seq peak located approximately 50kb downstream of *IL21* and T cell enhancer region (highlighted in yellow) was differentially accessible between T1D and healthy Tconv cells (n=12). 3D promoter interactome from CD4$^+$ T cell pcHi-C [6] showed that the differential region is contacting a distant promoter, *IL2,* located approximately 100kb away (highlighted in blue). The observation was confirmed by case-control RNA-seq where *IL2* was shown to be differentially expressed (FDR = 0.017) between T1D and healthy controls (n=9) (Figure 5-39).

Case-control ATAC-seq signal was intersected with T1D risk SNPs in high linkage disequilibrium (R^2=0.8) with the GWAS lead SNPs [3], T cell super enhancers [7], T cell epigenome and chromatin states (Roadmap Epigenomics Project [8]). Browser view was generated using WashU Epigenome Browser.

DA, differential accessibility.

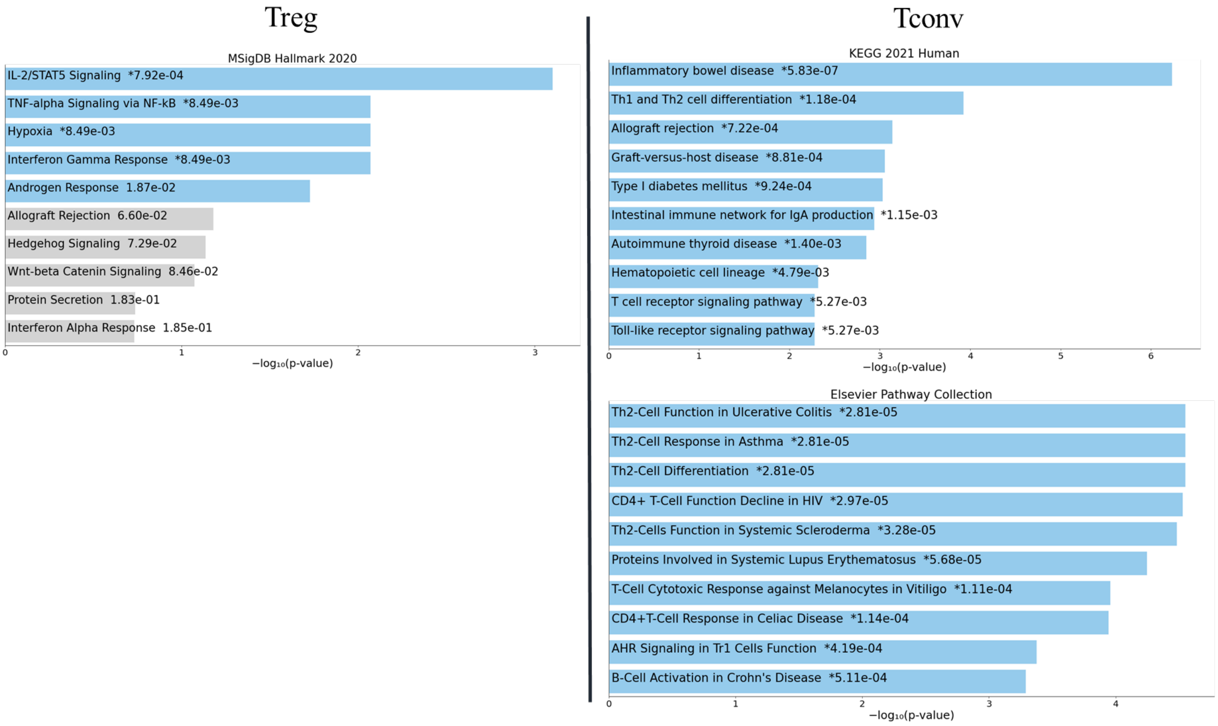


Figure 5-59 | Pathway enrichment analyses showing enriched pathways for Treg (left panel) and Tconv (right panel) gene targets identified from integration of ATAC-seq, Hi-C and RNA-seq.

5.4.5 Integrative analysis - enrichment of SNPs, cis-regulatory elements and TF occupancy in ATAC-seq peaks altered in T1D

GWAS contribute a wealth of data on the genetic architecture of complex diseases. However, most common (>5% minor allele frequency) variants identified confer relatively small increments in risk individually or in combination and represent only a small fraction of heritability predicted from familial clustering [279]. A number of hypotheses to explain for the remaining, "missing heritability" have been proposed, including much larger numbers of highly polygenic variants with small effect sizes that have yet to be captured [280, 281]; rare or low frequency (MAF < 5%) variants [282, 283], possibly with large effect sizes, were overlooked in current GWAS genotyping arrays which focus on common variants present in 5% or more of the population; low power to capture gene-gene interaction (epistasis) and insufficient accounting for shared familial environments or interactions among genotypes and environments [284-287]. As a result, many low frequency variants (MAF < 5%) with small effect sizes are unlikely to reach genome wide significance in current GWAS whereas rare variants are often under-represented on GWA genotyping arrays [288] owing to insufficiently large sample size or high-density comprehensive arrays. Furthermore, the majority of GWAS and genetic studies have been restricted to populations of European ancestry which can result in over-representation of variants with a higher minor allele frequency in Europeans compared with other populations, potentially constraining the relevance of these variants to the associated traits in non-Europeans [289].

As an alternative approach to identify novel putative T1D-associated human genetic variation independently of GWAS, I sampled 4,655,805 common variants (MAF > 10%) from a large population-scale variant resource, Genome Aggregation Database (gnomAD) (version 3.0) [1] as input to my filtering workflow, comprising of differentially accessible genomic regions between T1D cohort and healthy controls, enhancer- and promoter-associated long range regulatory regions (Figure 5-60, Figure 5-61, Figure 5-64). The same set of gnomAD common variants were studied in Liu, Sadlon [134] work and the authors demonstrated 55.7% of the gnomAD variants were not included in the largest meta-analysis of T1D genome-wide genotyped datasets to date [86]. gnomAD is the world's largest public catalog of human genetic variation and it harbours data from a number of large-scale human sequencing projects, including population and disease-specific genetic studies. Functionally annotating by incorporating regulatory elements produced from multiple experimental approaches, including high resolution conformation-dependent features and T1D-associated traits, enabled me to prioritise candidate functional, T1D risk-associated common genetic variants for follow-up studies (Figure 5-61, Figure 5-64). A total of 92 and 50 candidate gnomAD SNPs were located within accessible enhancers differentially regulated between T1D cohort and healthy controls identified from my ATAC-seq dataset (n=12 case-control pairs) and involved in long-range enhancer-promoter interaction in Treg and Tconv, respectively, suggesting they can potentially alter T cell-specific gene expression in T1D through spatial interactions (Figure 5-61, Figure 5-64). The identified candidate variants were enriched in TF binding motifs that have a key role in T cell development and differentiation, including YY1, FOXP3 and STAT6 for Treg cells, and MAF and GATA proteins for Tconv cells (Figure 5-62, Figure 5-65 (a)). SNPs in those regulatory elements can alter, by either disrupting or creating new TF binding sites in T1D-associated enhancers or alter chromatin accessibility, resulting in loss of long-range, physical enhancer-target interaction. To support this, Gao, Uzun [33] recently showed that the binding of YY1 was disrupted by a SNP (rs883868) in high linkage disequilibrium with the T1D-associated GWAS lead SNP at Treg enhancer, resulting in loss of long-range enhancer–promoter interaction. Consistent with the pathogenic role of Treg and Tconv cells in T1D, our pathway enrichment analyses revealed that the candidate T1D-associated gnomAD variant sites were associated with key Treg pathways such as IL-2 signalling pathway, IL-4 regulation of apoptosis in Treg (Figure 5-63) and inflammatory IL-1-induced NF-kappaB activation in Tconv cells (Figure 5-65; b). As initial TF motif enrichment analyse revealed that candidate T1D-associated gnomAD variants were enriched in TF binding motifs (Figure 5-62, Figure 5-65), I then incorporated the TF footprinting data I generated from the T1D ATAC-seq datasets (n=12 case-control pairs) into the filtering workflow (Figure 5-66, Figure 5-68). I identified 10 and 54 TF footprints, in Treg and Tconv cells, respectively, from T1D patients that coincided with accessible chromatin differentially regulated between T1D and healthy controls, gnomAD common SNPs (MAF > 10%) and involved in long-range enhancer-promoter gene regulation (Figure 5-66, Figure 5-68). These TF footprint regions were strongly predicted to be bound by TFs with evidence of local protection from Tn5 cleavage, and SNPs in these regions can disrupt TF binding sites and interfere with enhancer function in T1D in Tconv and Treg cells, potentially resulting in loss of physical long-range target gene regulation.

The filtered TF footprints in Treg and Tconv cells were strongly associated with immune signalling pathways involving multiple cytokines such as IFN-β, IL6, IL-12/STAT4 and IFN-γ (Figure 5-67, Figure 5-69). As STAT4 (Signal transducer and activator of transcription 4) is required for the development of Th1 cells from naive CD4$^+$ T cells and IFN-γ production in response to IL-12 [290], this observation supports the implication of Th1- and Th2-mediated pathology in T1D [291].

As genetic variants/SNPs can be distributed at random across the genome, I assessed if the credible T1D risk SNPs, which were defined in Onengut-Gumuscu, Chen [2] using dense genotyping array with Bayesian approach, were statistically significantly associated (more than expected by chance) with the differentially accessible ATAC-seq peaks I identified in the Tconv and Treg cells from T1D cohort and healthy controls (n=12) (Figure 5-70, Figure 5-71) using permutation test [28]. The results indicated that the credible T1D risk-associated SNPs were statistically significantly (p-value < 0.05) enriched in the case-control differential ATAC-seq peaks identified in both Tconv and Treg cells, when using all ATAC-seq peaks (i.e. universe region set; peaks below statistical significance included) as randomised region sets (Figure 5-70, Figure 5-71). This data strongly suggest that the credible T1D risk SNPs are more likely to exert their effect through accessible enhancers significantly altered in T1D in T cells, compared with common or ubiquitous accessible genomic regions not expected to be altered in T1D.

To gain insight into the functional relevance of non-coding T1D risk-associated variants, I then incorporated functional annotation of regulatory elements identified from my case-control data to prioritize the list of T1D-associated SNPs and delineate biological plausibility. It is critical to consider that SNPs identified from GWAS genotyping arrays (i.e. index/lead SNPs) serve only as representatives for all the SNPs in the same haplotype block and it is possible that other SNPs in high linkage disequilibrium (LD) with the index SNPs are casual for the disease. However, many current studies that leverage GWAS data and functional annotation to inform disease causality include only genome-wide significant SNPs and do not account for linkage disequilibrium (LD). To improve detection power in identifying variants underlying association signals from GWAS arrays, I expanded my analyses using LD calculation together with the 1000 Genomes Project Phase 3 reference panels to include SNPs that are in high LD with the T1D index SNPs (referred to as “LD-SNPs”) in my enrichment analyses. Using this approach I identified 111 and 96 LD-SNPs, in Treg and Tconv cells, respectively, from T1D patients that coincided with accessible chromatin regions and TF footprints differentially regulated between T1D and healthy controls (Figure 5-72, Figure 5-73). These filtered T1D associated SNPs are enriched in TF binding sites that are crucial for Treg cell activation and suppressive functions such as FOXP3 and ETS2 [292], as well as differentiation of Tconv cells such as MAF (Figure 5-78). Those filtered risk regions are associated with multiple pathways involved in T cell regulation such as IL-2/STAT5 signalling in the Treg cells and IFN response in the Tconv cells

(Figure 5-78). Attempts were made to extract genotype variants at these regulatory locations from ATAC-seq data and mismatches were identified between T1D individuals and their sibling-matched healthy controls (Table 7-10).

To assess if the filtered LD-SNPs randomly overlapped with or enriched in the differential case-control ATAC-seq peaks and TF footprints, I performed permutation tests to evaluate their associations. It was revealed that the filtered T1D LD-SNPs were statistically significantly (p-value < 0.05) enriched in the ATAC-seq peaks as well as TF footprints differentially regulated (FDR < 0.05) between T1D cohort and healthy controls in both Tconv and Treg cells, when using all ATAC-seq peaks and TF footprints (i.e. universe region set; regions below statistical significance included) as randomised region sets (Figure 5-74, Figure 5-75, Figure 5-76, Figure 5-77). This observation supports direct involvement of Tconv and Treg cells in the pathogenesis of T1D. It also suggests that T1D SNPs are more likely to exert their effect through accessible enhancers and TF binding sites significantly altered in T1D in T cells, compared with common or ubiquitous enhancers or TF binding sites not expected to be altered in T1D.

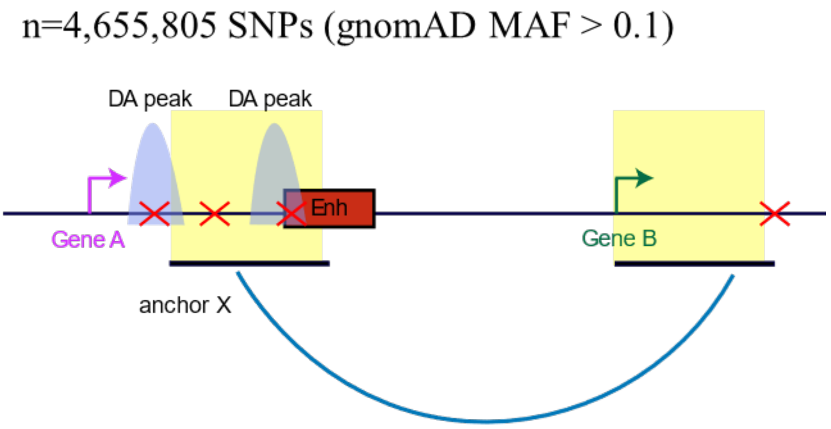


Figure 5-60 | Schematic showing the identification of novel putative T1D-associated human genetic variation (>10% allele frequency) from the Genome Aggregation Database (gnomAD)[1].
gnomAD is the world's largest public catalog of human genetic variation and it harbours data from a number of large-scale human sequencing projects, including population and disease-specific genetic studies. I sampled 4,655,805 common variants (MAF > 0.1) from the Genome Aggregation Database (gnomAD) (version 3.0) as input to my filtering workflow comprising of differentially accessible genomic regions between T1D and healthy controls, enhancer- and promoter-associated interaction anchors.

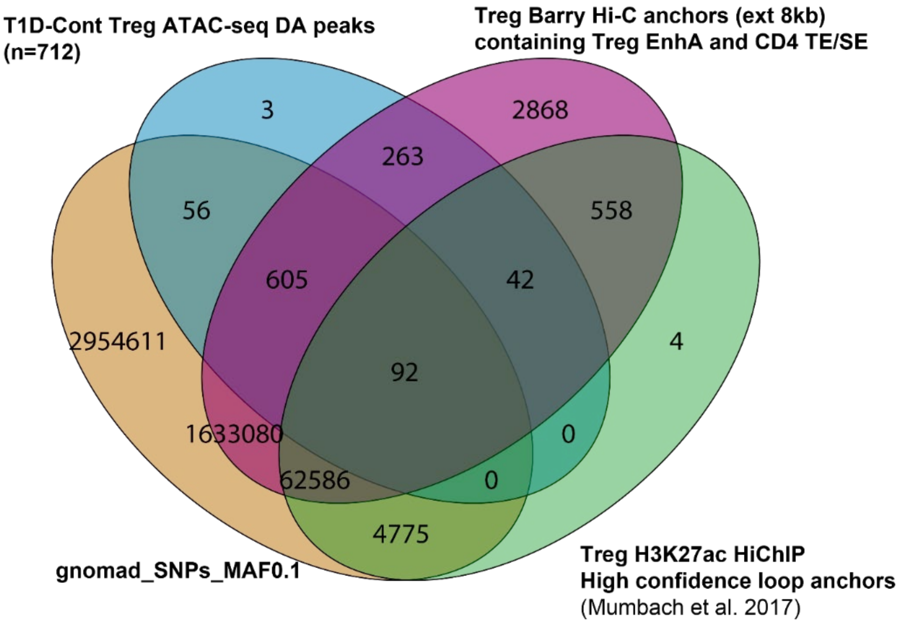


Figure 5-61 | Identification of novel putative T1D-related gnomAD human variants (MAF > 0.1) [1] in Treg cells.
The 92 gnomAD SNPs were located within ATAC-seq peaks differentially regulated between T1D and healthy controls (n=12), enhancer- and promoter- interaction anchors from Treg Hi-C (Barry Lab) and H3K27ac HiChIP [37], suggesting long-range gene regulation involving enhancer-target looping.

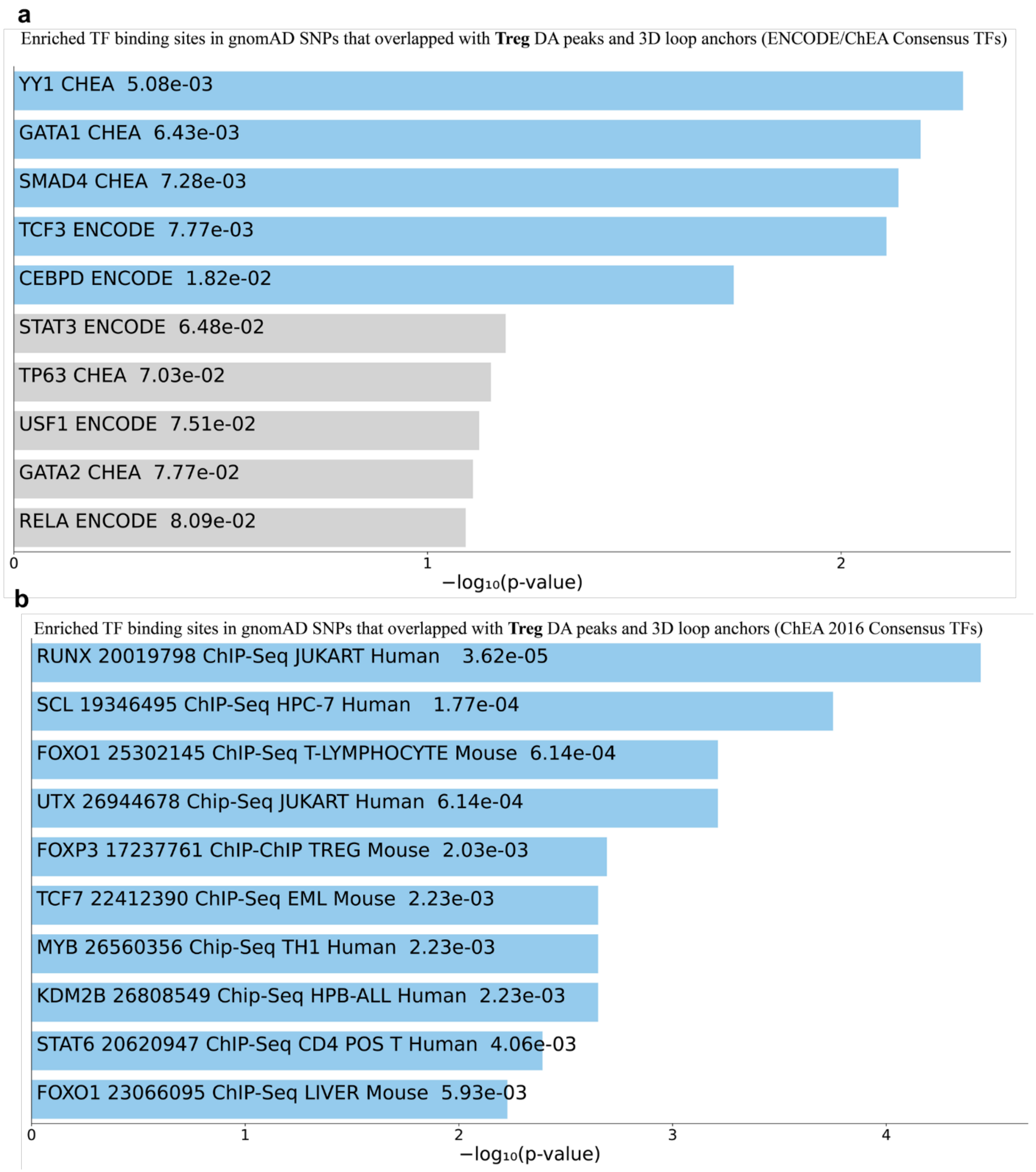


Figure 5-62 | Enrichment of TF binding sites for 92 Treg-specific gnomAD human genetic variants (MAF > 0.1) potentially implicated in T1D, using ENCODE (a) and ChEA [36] (b) databases.

The 92 Treg-specific gnomAD SNPs (venn diagram from Figure 5-61) used in this enrichment analysis were located within the case-control Treg ATAC-seq peaks and enhancer- and promoter- 3D loop anchors.

The bar chart visualizes the enriched terms and their p-values.

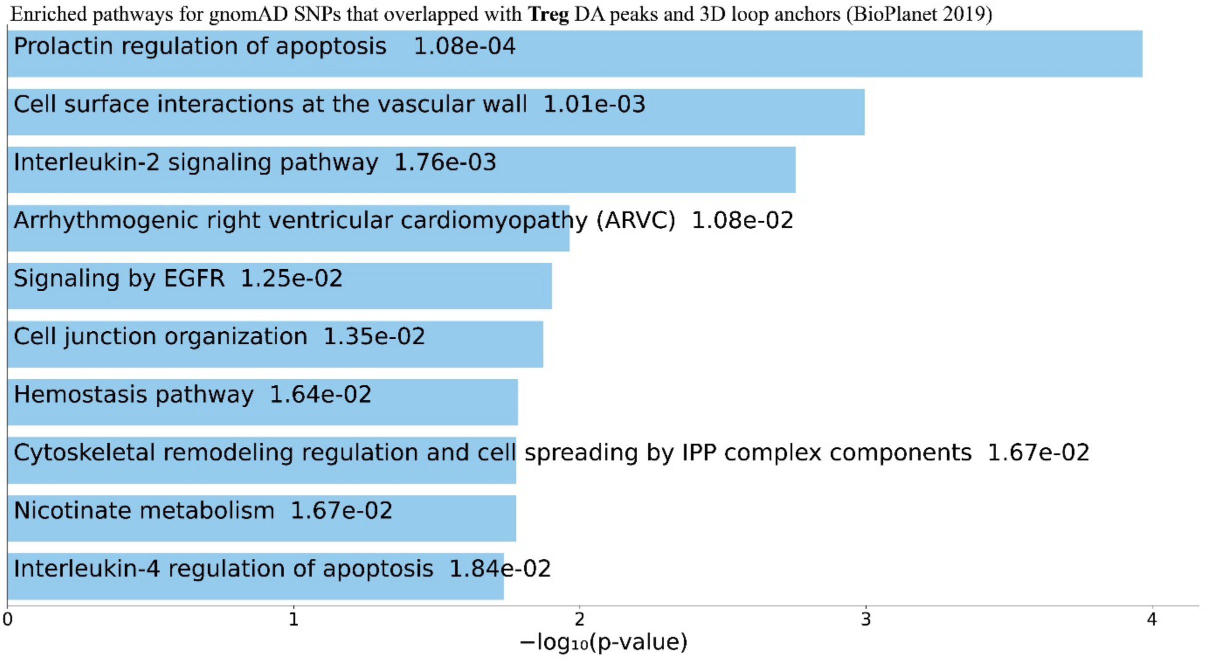


Figure 5-63 | Pathway enrichment analysis for 92 Treg-specific gnomAD human genetic variants (MAF > 0.1) potentially implicated in T1D.
The 92 Treg-specific gnomAD SNPs (venn diagram from Figure 5-61) used in this enrichment analysis were located within the case-control Treg ATAC-seq peaks and enhancer- and promoter- 3D loop anchors.

The bar chart visualizes the enriched terms and their p-values.

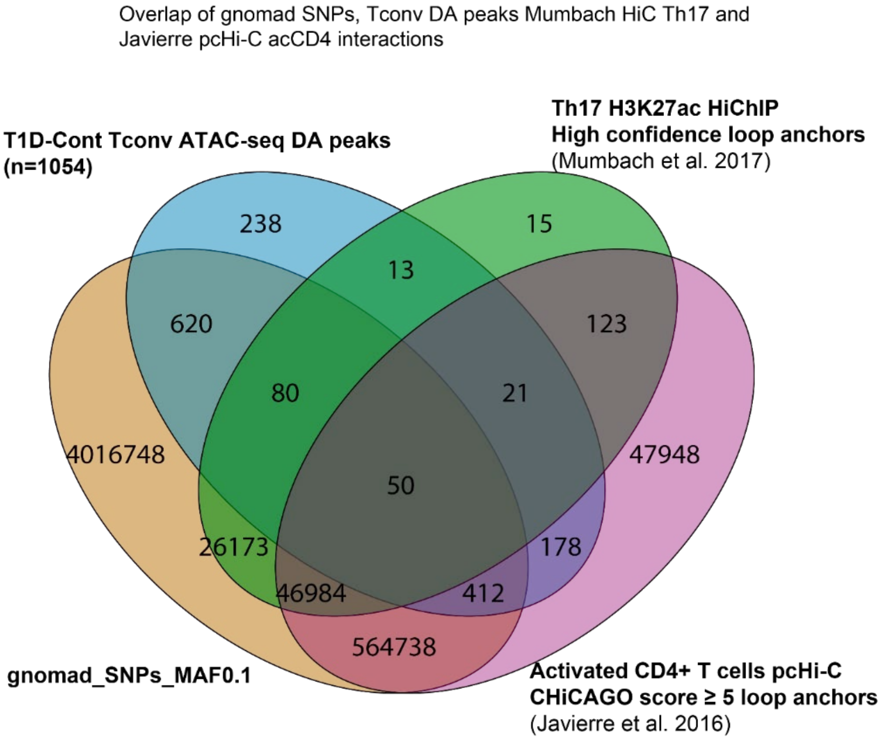


Figure 5-64 | Identification of novel putative T1D-associated gnomAD human variants (MAF > 0.1) [1] in Tconv cells.

The 50 gnomAD SNPs were located within ATAC-seq peaks differentially regulated between T1D and healthy controls (n=12), enhancer- and promoter- interaction anchors from Th17 H3K27ac HiChIP [37] and CD4^{+} T cell promoter capture Hi-C [6] , suggesting long-range gene regulation involving enhancer-target looping.

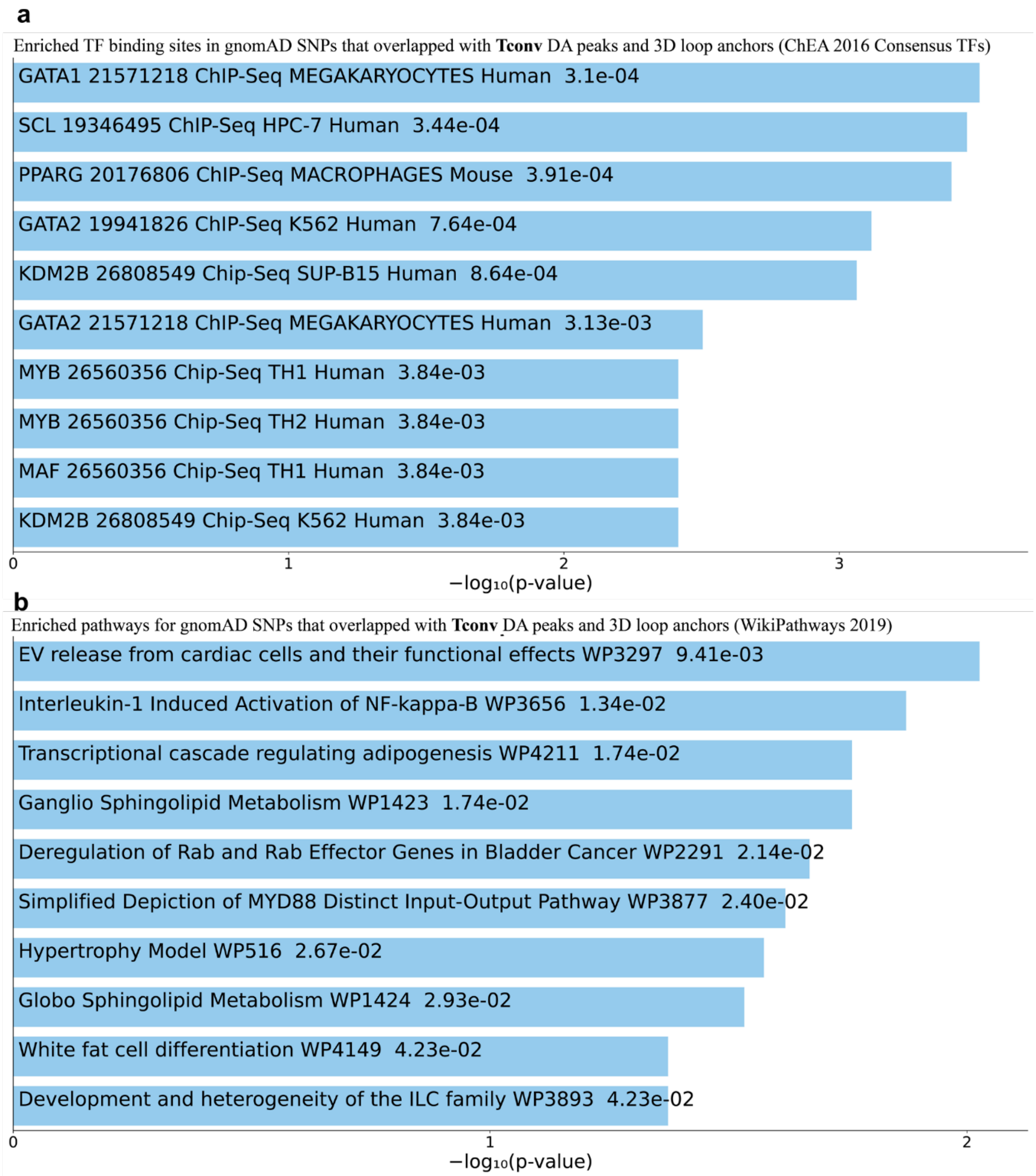


Figure 5-65 | Enrichment of TF binding sites for 50 Tconv-specific gnomAD human genetic variants (MAF > 0.1) potentially implicated in T1D, using ChEA[36] (a) and WikiPathways (b) databases.

The 50 Treg-specific gnomAD SNPs (venn diagram from Figure 5-64) used in this enrichment analysis were located within the case-control Tconv ATAC-seq peaks and enhancer- and promoter- 3D loop anchors.

The bar chart visualizes the enriched terms and their p-values.

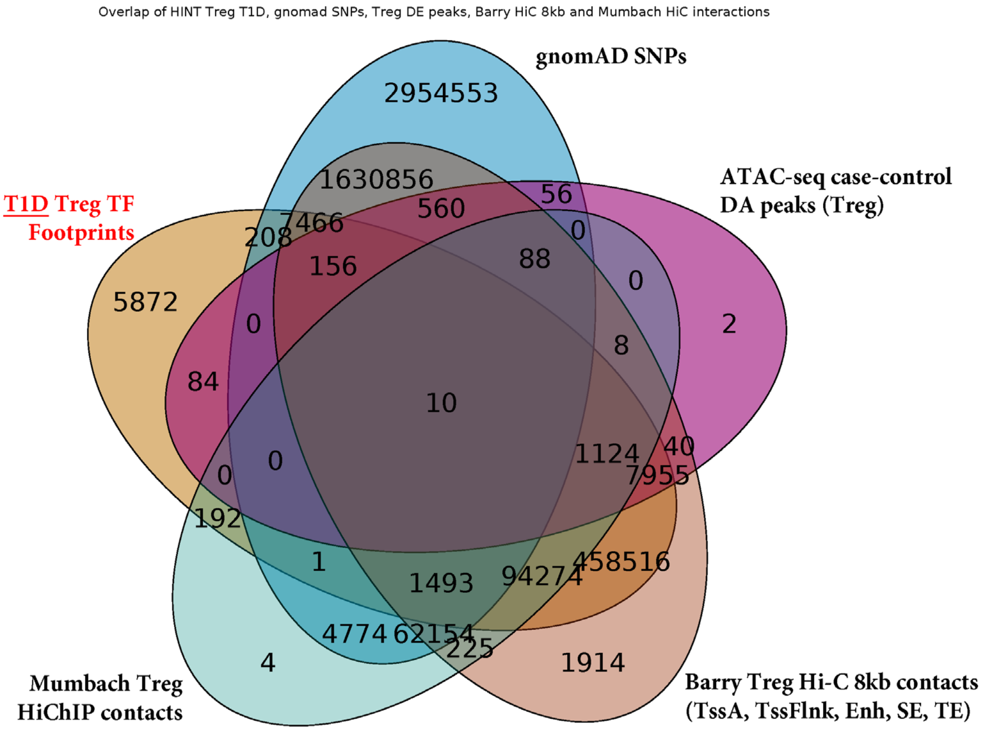

Overlap of HINT Treg T1D, gnomad SNPs, Treg DE peaks, Barry HiC 8kb and Mumbach HiC interactions
gnomAD SNPs
ATAC-seq case-control DA peaks (Treg)
T1D Treg TF Footprints
Mumbach Treg HiChIP contacts
Barry Treg Hi-C 8kb contacts (TssA, TssFlnk, Enh, SE, TE)
2954553
1630856
560
56
0
7466
208
156
88
0
5872
0
2
8
84
10
0
1124
40
0
7955
192
1
1493
94274
458516
4774
62154
225
4
1914

Figure 5-66 | Identification of TF footprints from T1D Treg cells (n=12) that coincide with differential ATAC-seq peaks, gnomAD human variants (MAF > 0.1) [1], and enhancer- and promoter- 3D interaction anchors [37].

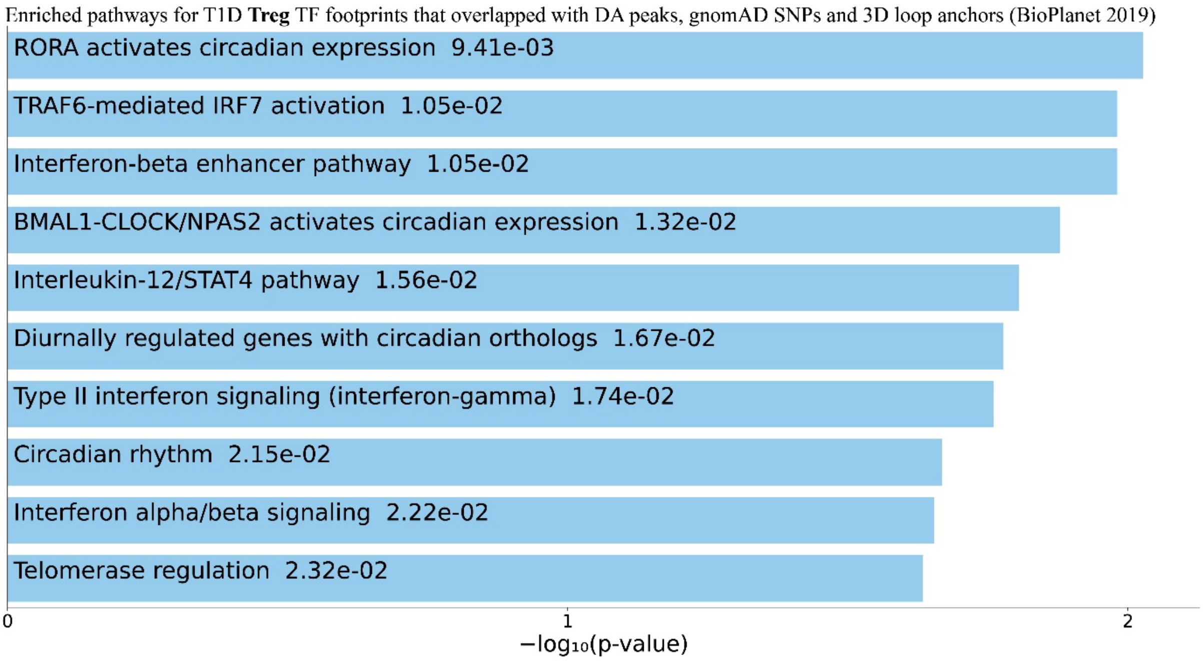


Figure 5-67 | Pathway enrichment analysis for 10 T1D Treg-specific TF footprints that coincide with gnomAD SNPs, case-control Treg ATAC-seq peaks and enhancer- and promoter- 3D loop anchors (venn diagram from Figure 5-66).

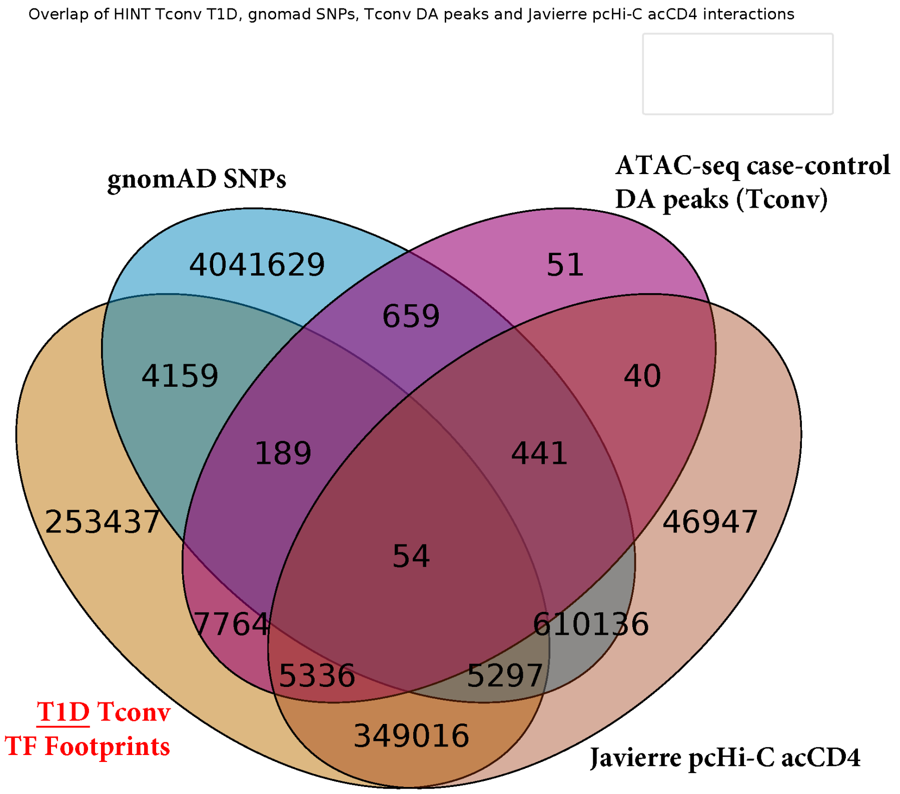

Overlap of HINT Tconv T1D, gnomad SNPs, Tconv DA peaks and Javierre pcHi-C acCD4 interactions
gnomAD SNPs
ATAC-seq case-control DA peaks (Tconv)
T1D Tconv TF Footprints
Javierre pcHi-C acCD4
4041629
51
659
4159
40
189
441
253437
46947
54
7764
610136
5336
5297
349016

Figure 5-68 | Identification of TF footprints from T1D Tconv cells (n=12) that coincide with differential ATAC-seq peaks, gnomAD human variants (MAF > 0.1) [1], and enhancer- and promoter- 3D interaction anchors [6].

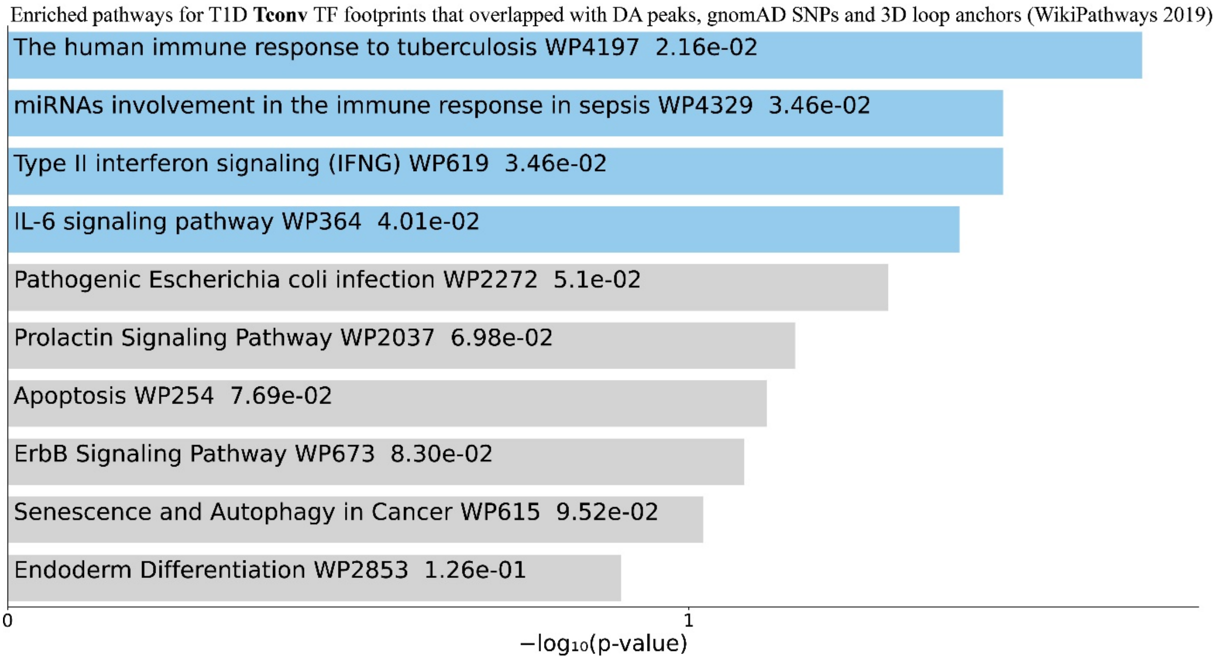


Figure 5-69 | Pathway enrichment analysis for 54 T1D Tconv-specific TF footprints that coincide with gnomAD SNPs, case-control Tconv ATAC-seq peaks and enhancer- and promoter- 3D loop anchors (venn diagram from Figure 5-68).

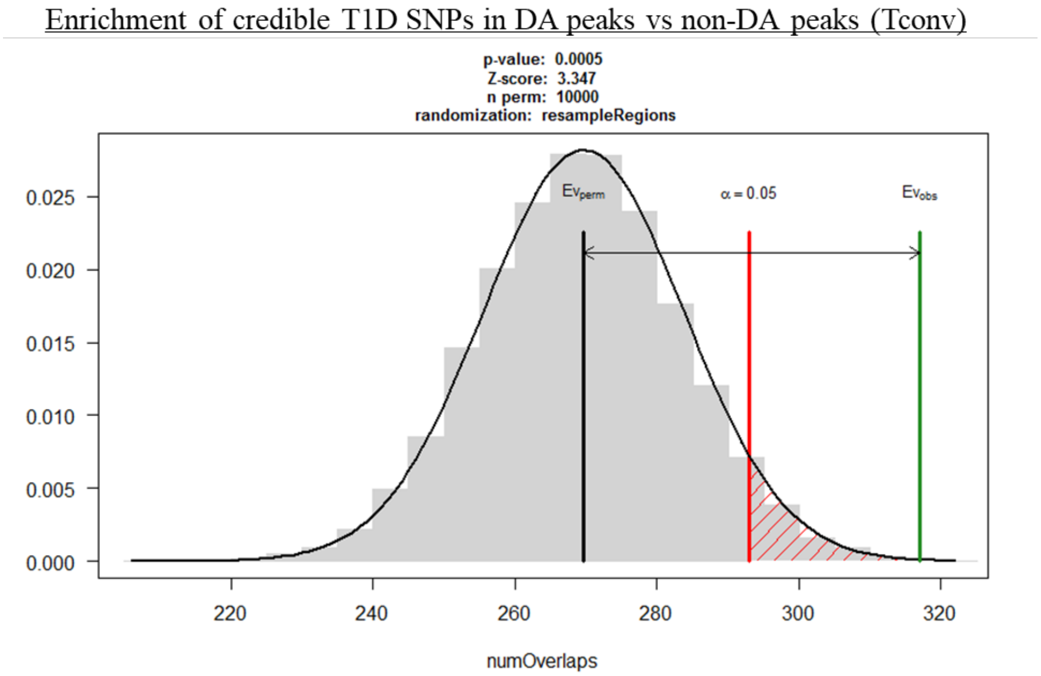

Enrichment of credible T1D SNPs in DA peaks vs non-DA peaks (Tconv)

Figure 5-70 | Plot of the results of a permutation test assessing the association between credible T1D risk SNPs [2] and differentially accessible ATAC-seq peaks between T1D and healthy controls (n=12) in the <u>Tconv</u>, using the number of overlaps as the evaluation function and 10,000 permutations.

Association test was performed using regioneR [28] which implements a permutation test framework. In this test we asked if T1D risk SNPs overlapped with differential ATAC-seq peaks more than expected by chance based on 10,000 permutations. The randomised region sets consisted of all the ATAC-seq peaks, differential or non-differential (universe region set).

The results of permutation test are presented in a histogram representing the evaluation of the randomized peak set with a fitted normal, black bar representing the mean of the randomized evaluations and a green bar representing the evaluation of the original region set (SNPs in differential peaks). A red bar (and red shading) represents the significance limit (0.05). In this case as the green bar is in the red shaded region (p-value = 0.0005) it indicates the T1D risk SNPs were statistically significantly associated with the case-control differential ATAC-seq peaks identified in our case-control Tconv cells.

nperm = number of randomizations

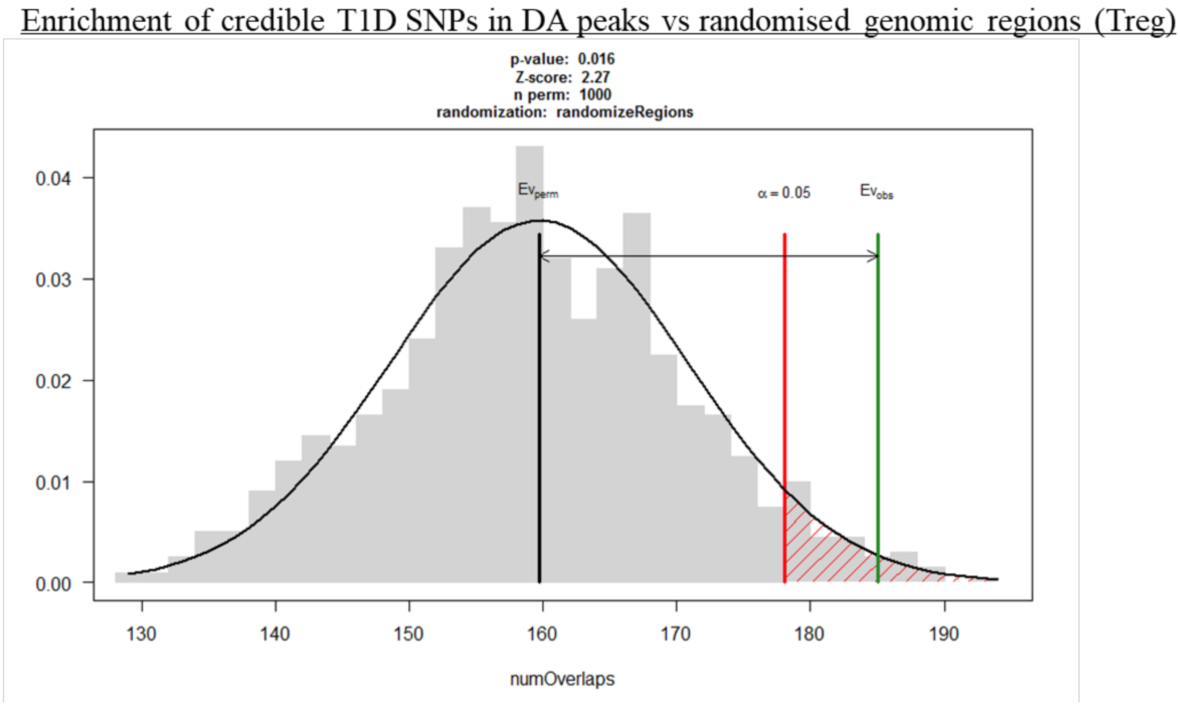


Figure 5-71 | Plot of the results of a permutation test assessing the association between credible T1D risk SNPs [2] and differentially accessible ATAC-seq peaks between T1D and healthy controls (n=12) in the Treg.

In this test we asked if T1D risk SNPs overlapped with differential ATAC-seq peaks in the Treg more than expected by chance based on 1000 permutations. The randomised region sets consisted of non-masked parts of the genome (same number and width as the differential peaks).

The results of permutation test are presented in a histogram representing the evaluation of the randomized peak set with a fitted normal, black bar representing the mean of the randomized evaluations and a green bar representing the evaluation of the original region set (SNPs in differential peaks). A red bar (and red shading) represents the significance limit (0.05). In this case as the green bar is in the red shaded region (p-value = 0.016) it indicates the T1D risk SNPs were statistically significantly associated with the differential ATAC-seq peaks identified in our case-control Treg cells.

nperm = number of randomizations

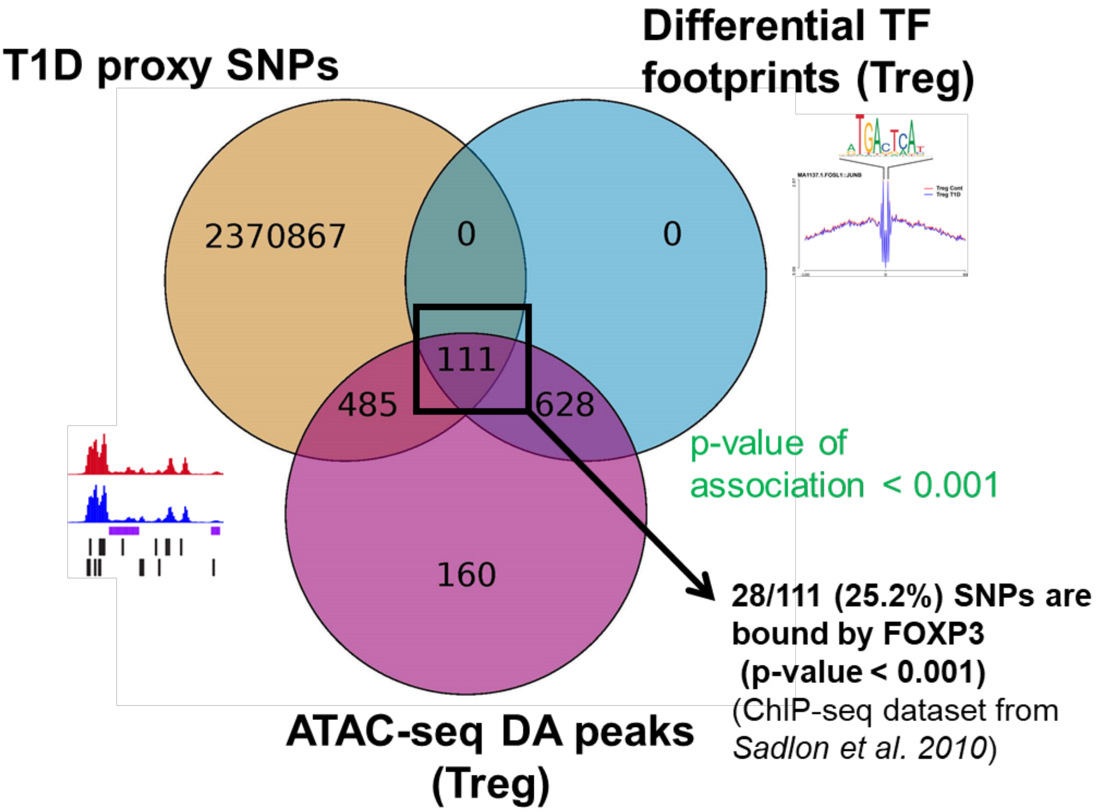

T1D proxy SNPs
Differential TF footprints (Treg)
ATAC-seq DA peaks (Treg)
2370867
0
0
111
485
628
160
MA1137.1.FOSL1::JUNB
Treg Cont
Treg T1D
p-value of association < 0.001
28/111 (25.2%) SNPs are bound by FOXP3
(p-value < 0.001)
(ChIP-seq dataset from *Sadlon et al. 2010*)

Figure 5-72 | Prioritisation of T1D associated risk SNPs through mapping to case-control regulatory regions in the Treg.

Proxy SNPs in high linkage disequilibrium with the GWAS T1D index SNPs[3] were identified using r^2 threshold of 0.8 and ±100bp window of proxy SNPs was considered in this analysis. Differential case-control ATAC-seq peaks and TF footprints were identified from Treg cells from 12 T1D and healthy controls. The association between filtered T1D SNPs (n=111) and FOXP3 binding sites [11] was assessed using permutation test [28], using the number of overlaps as the evaluation function and 5000 permutations.

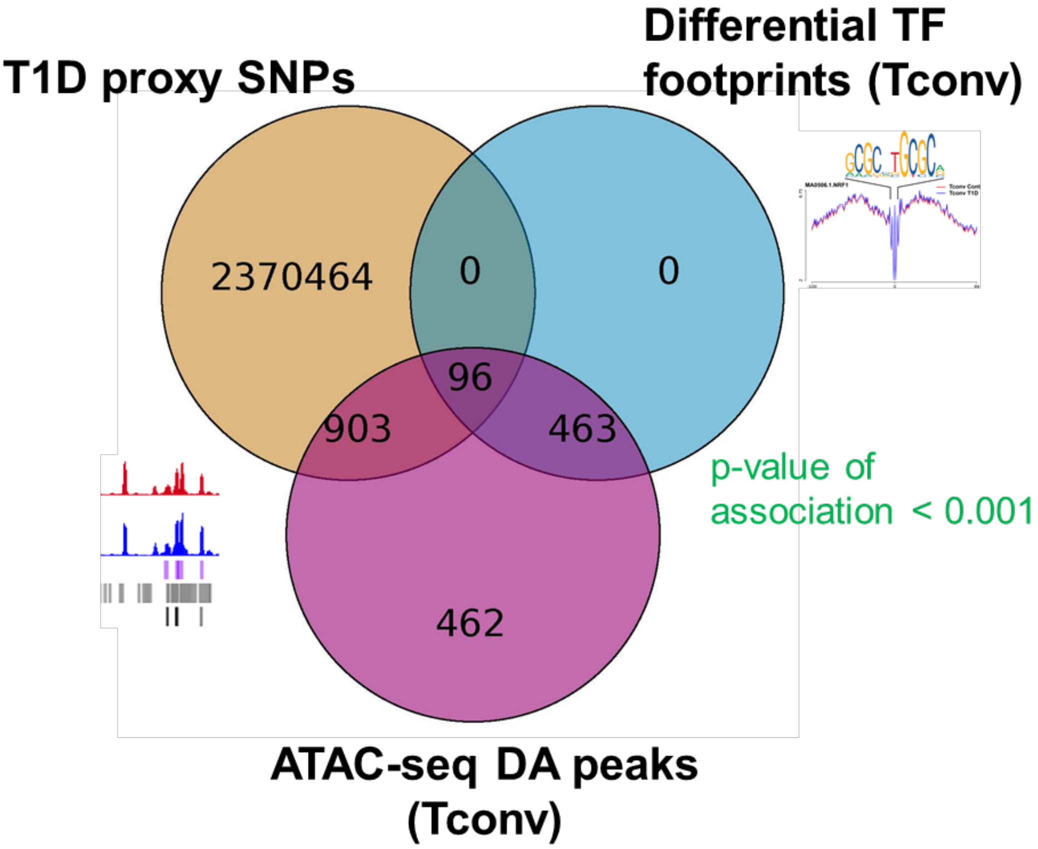

T1D proxy SNPs
Differential TF footprints (Tconv)
2370464
0
0
96
903
463
462
p-value of association < 0.001
ATAC-seq DA peaks (Tconv)

Figure 5-73 | Prioritisation of T1D associated risk SNPs through mapping to case-control regulatory regions in the Tconv.

Proxy SNPs in high linkage disequilibrium with the GWAS T1D index SNPs[3] were identified using r^2 threshold of 0.8 and ±100bp window of proxy SNPs was considered in this analysis. Differential case-control ATAC-seq peaks and TF footprints were identified from Tconv cells from 12 T1D and healthy controls. The association of filtered T1D SNPs (n=96) with differential case-control ATAC-seq peaks and TF footprints was assessed using permutation test [28].

Treg - **Intervened** T1D SNPs in DA peaks vs non-DA peaks

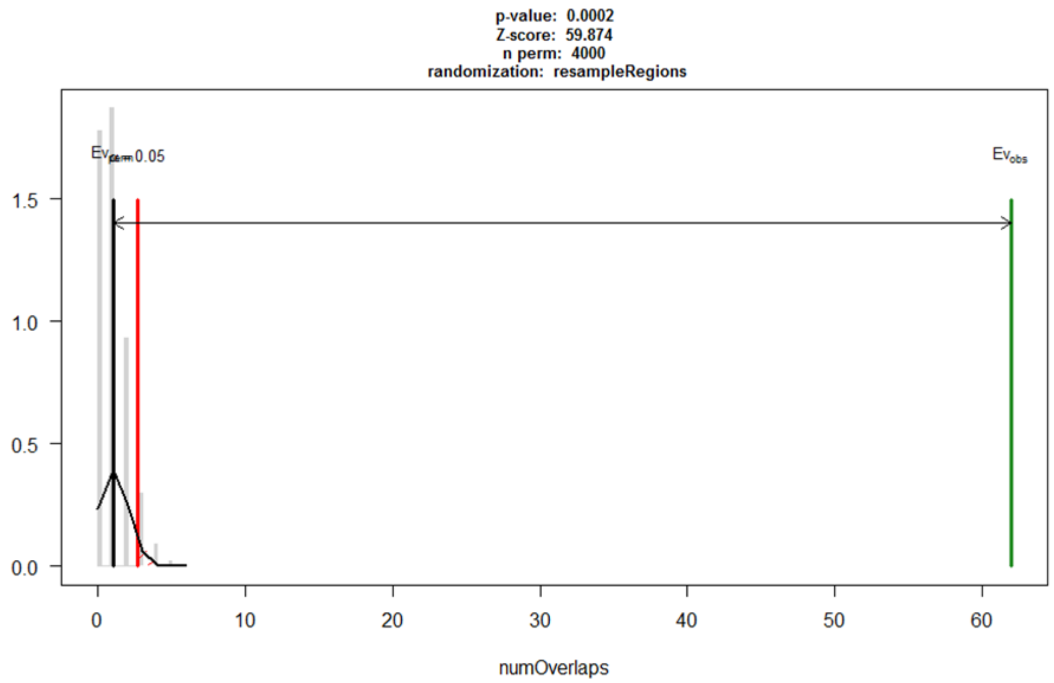


Figure 5-74 | Association analysis of filtered T1D SNPs (n=111) and differentially accessible ATAC-seq peaks between T1D and healthy controls (n=12) in the Treg.
The 111 T1D SNPs were identified based on their location in the regulatory elements - ATAC-seq peaks and TF footprints differentially regulated between T1D and healthy controls (venn diagram from Figure 5-72).

4000 permutations were performed to statistically evaluate the association between filtered T1D SNPs and case-control ATAC-seq peaks. The randomised region sets consisted of a universe of ATAC-seq peaks (differential and non-differential).

Tconv - **Intervened** T1D SNPs in DA peaks vs non-DA peaks

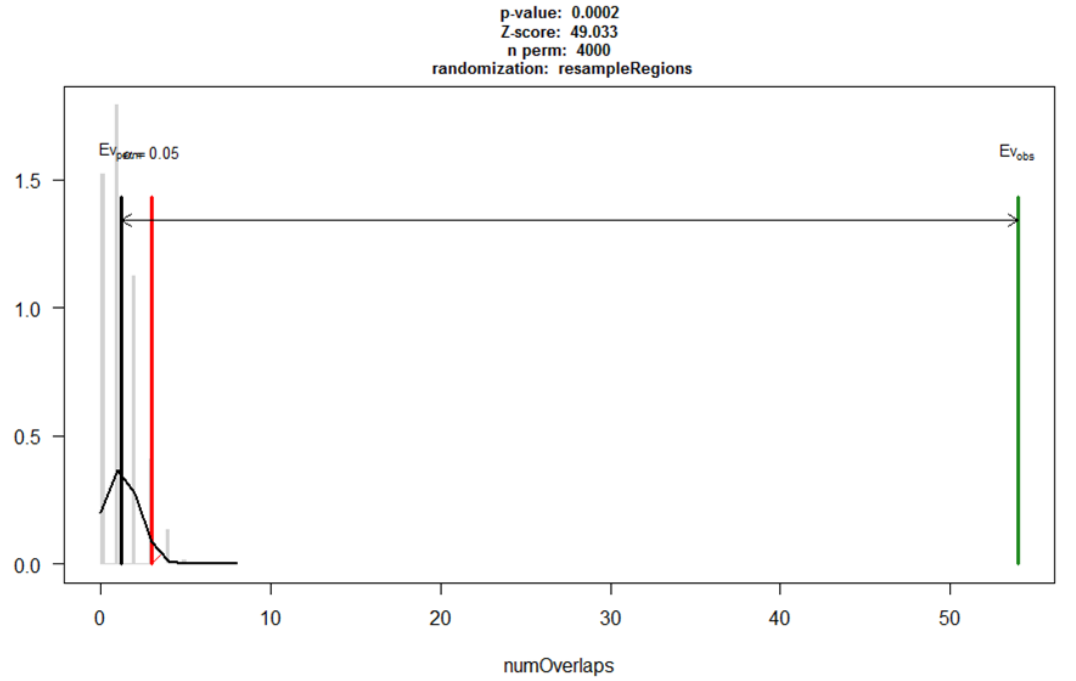


Figure 5-75 | Association analysis of filtered T1D SNPs (n=96) and differentially accessible ATAC-seq peaks between T1D and healthy controls (n=12) in the Tconv.
The 96 T1D SNPs were identified based on their location in the regulatory elements - ATAC-seq peaks and TF footprints differentially regulated between T1D and healthy controls (venn diagram from Figure 5-73).

4000 permutations were performed to statistically evaluate the association between filtered T1D SNPs and case-control ATAC-seq peaks. The randomised region sets consisted of a universe of ATAC-seq peaks (differential and non-differential).

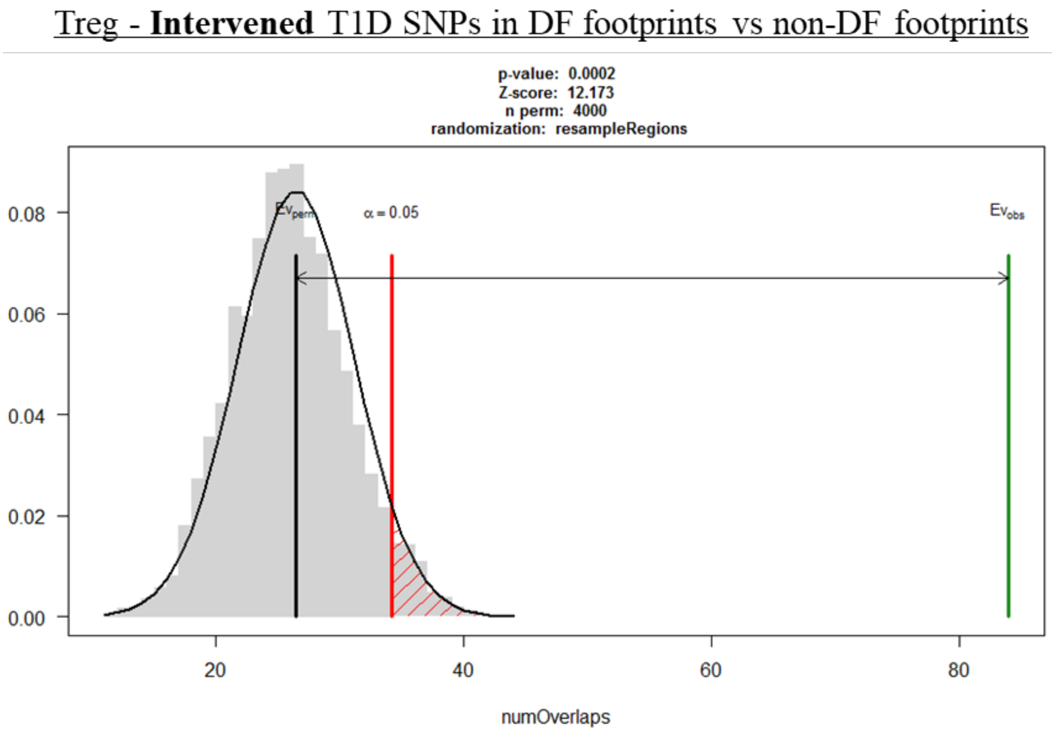


Figure 5-76 | Association analysis of filtered T1D SNPs (n=111) and differential TF footprints from T1D and healthy controls (n=12) in the Treg.
The 111 T1D SNPs were identified based on their location in the regulatory elements - ATAC-seq peaks and TF footprints differentially regulated between T1D and healthy controls (venn diagram from Figure 5-72).

4000 permutations were performed to statistically evaluate the association between filtered T1D SNPs and case-control differential TFs. The randomised region sets consisted of a universe of TF footprints (differential and non-differential).

Tconv - **Intervened** T1D SNPs in DF footprints vs non-DF footprints

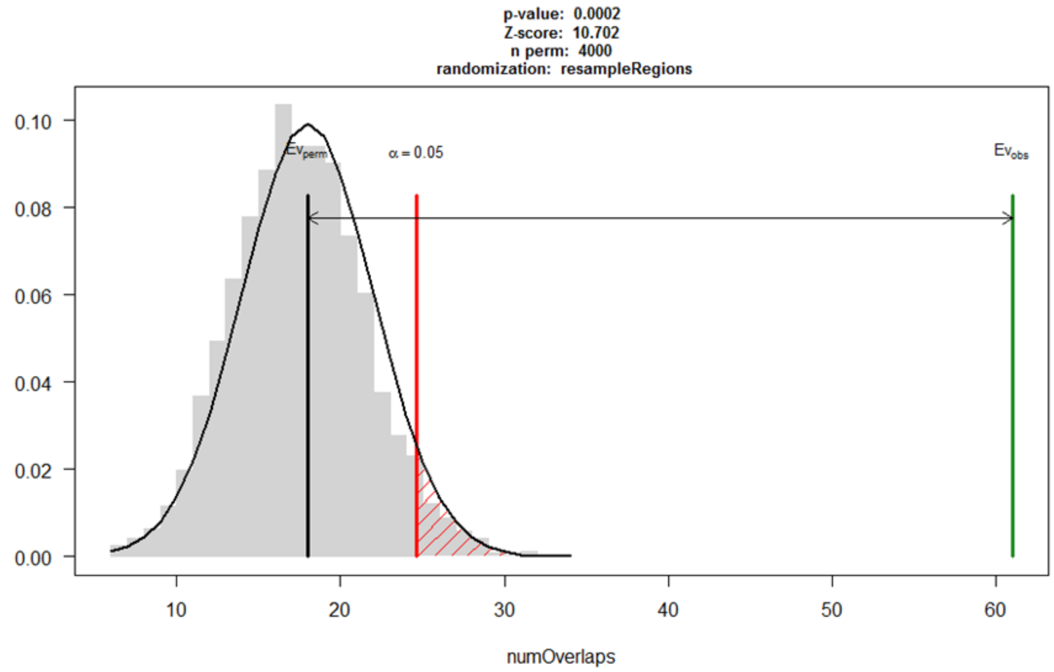


Figure 5-77 | Association analysis of filtered T1D SNPs (n=96) and differential TF footprints from T1D and healthy controls (n=12) in the Tconv.
The 111 T1D SNPs were identified based on their location in the regulatory elements - ATAC-seq peaks and TF footprints differentially regulated between T1D and healthy controls (venn diagram from Figure 5-73).

4000 permutations were performed to statistically evaluate the association between filtered T1D SNPs and case-control differential TFs. The randomised region sets consisted of a universe of TF footprints (differential and non-differential).

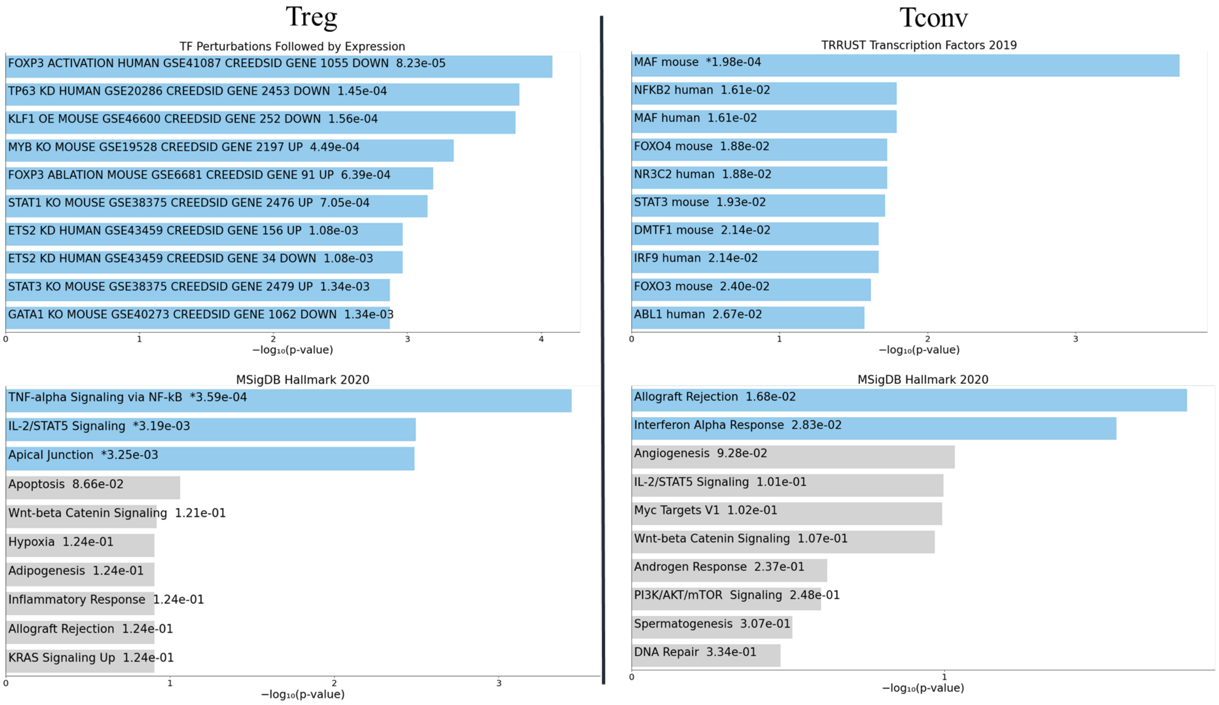


Figure 5-78 | Pathway enrichment analysis for 111 and 95 T1D LD SNPs that coincide with differential case-control ATAC-seq peaks and TF footprints in the Treg and Tconv, respectively.

5.5 Discussion

Type 1 diabetes (T1D) is an autoimmune disease caused by the immune system attacking and destroying the insulin-producing pancreatic β-cells. The exact cause for T1D is unknown and the reason a child develops T1D depends on a combination of their genetic makeup and environmental influence [76, 91, 293], leading to a breakdown in tolerance normally maintained by Treg cells. Environmental contributions associated with T1D are not well characterised but emerging studies suggest that they may alter the immune system via epigenetic modification such as DNA methylation, histone modification and microRNA [97, 102]. Recent data strongly link the breakdown in tolerance in multiple autoimmune diseases to alterations in the transcriptional program in CD4$^+$ T cells [88, 294], however, the molecular mechanisms are not well understood.

Genome-wide association studies (GWAS) and other genetic studies [2, 3, 85-87, 295] have identified 59 non-HLA genomic loci associated with the risk of developing T1D. They show higher likelihood of having an impact on T1D susceptibility but not much is known about how they contribute to the disease pathogenesis or how environmental factors combine with them to cause the loss of tolerance in T1D. Most of the T1D risk loci do not mediate disease susceptibility through changes to protein-coding regions but in fact impact non-coding regulatory elements like gene enhancers. Importantly, those T1D risk variants are highly associated with Tconv/Treg-specific enhancers [2, 88] and this enrichment is particularly pronounced at enhancers that are activated in response to antigen recognition and form super enhancers in T cells [88, 110]. Because they do not change protein sequence, they are likely to alter gene expression rather than protein function.

It was hypothesised that genetic variation and epigenetic changes in T1D modify the activity of Treg- and Tconv-specific enhancers, resulting in altered expression of the genes normally controlled by these enhancers. Toward this hypothesis, this chapter describes the alterations in chromatin accessibility and transcriptomes in Treg and Tconv cells from T1D cohort and healthy controls, determined using ATAC-seq and RNA-seq experiments. It also demonstrates genome-wide alterations in TF dynamics as well as dysregulated gene targets in T1D through integration of ATAC-seq, RNA-seq and published Hi-C contact maps.

High resolution linear epigenomics map generated by ATAC-seq revealed a subset of T cell enhancer-associated alterations in Tconv and Treg cells of 12 T1D patients and 12 sibling-matched healthy controls. A fraction of the differentially accessible (DA) regions were annotated to gene loci previously known to be associated with T1D [2, 3, 85-87, 295], corroborating the findings from other studies and validating the capability of this approach in capturing known disease signatures. They include *CTLA4* and *FIGNL1* (Fidgetin-like protein 1) which showed decreased accessibility in Treg cells of this T1D cohort, and *CCR2* (C-C chemokine receptor type 2) and *CCR5* (C-C chemokine receptor type 5) which showed decreased accessibility in Tconv cells of this T1D cohort. CTLA-4 was shown to attenuate T-cell activation through competing with CD28 binding to B7 ligands and inhibit TCR signalling by direct interaction with the TCR complex [296-299]. Downregulation of CTLA4 in Treg cells of T1D patients may indicate impaired inhibition of T cell activity leading to autoimmune destruction of pancreatic beta cells. CCR2 and CCR5, which are expressed on activated CD4$^+$ T cells, are involved in leukocyte trafficking which promotes migration of immune cells from the peripheral circulation to the sites of inflammation [300]. Although downregulation of chemokine receptors was expected to be accompanied by reduced β-cell destruction as less leukocytes and activated lymphocytes migrate to the site of inflammation, redundancy and promiscuous ligand/receptor usage of chemokines [301, 302] could compensate for the lack of CCR2/CCR5-mediated chemotactic ability and contribute to the pathogenesis of T1D. The enrichment of T1D risk-associated SNPs in the T1D-associated enhancers identified from our study reinforces observations from previous studies [2, 88] and suggests differential regulation of the interacting enhancers is likely to be driven by local genetic variants through altering chromatin accessibility, resulting in altered expression of the target gene products. Overall, the chromatin accessibility data supports the idea that, like any other autoimmune disorders, in T1D, development and differentiation of Tconv/Treg cells involving cytokine-cytokine receptor interaction are impaired [303, 304].

The next question is what underlies alterations in chromatin accessibility dynamics in T1D? As TF are key regulatory drivers of gene transcription [305] it was hypothesized that alterations in TF binding at these regions are involved in T1D. Using high resolution chromatin accessibility profiles, this study demonstrates the capability of identifying alterations in TF dynamics in enhancer regions differentially regulated between T1D cohort and healthy controls. Footprinting results revealed a subset of TFs demonstrating alterations in binding dynamics in T1D patients. Interestingly, a fraction of the differential TF binding signal identified in this case-control study was also reported to be disrupted in NOD mice (in comparison to WT mice), such as TF family of RUNX (Runt-related Transcription Factor), EGR (Early growth response), SP (Specificity protein) and SOX (SRY-related HMG-box) [306]. A key finding of this study is we observed alterations in binding of Activator Protein-1 (AP-1) TF family such as FOS, FOSL1 and FOSL2 which all exhibited lower binding dynamics in both Tconv and Treg cells of T1D patients compared with healthy controls. AP-1 proteins are expressed in CD4^{+} Tconv and Treg cells [307] and their transcript levels were found to be lower in the Treg cells from T1D patients compared with healthy controls in our case-control RNA-seq. FOS proteins are known to regulate cell proliferation, differentiation and transformation[308], and they demonstrate functional redundancy which enables them to compensate for the loss of each other in a non-synergistic manner [309, 310]. Genome-wide association studies (GWAS) has also identified *FOSL2* as a risk locus associated with several autoimmune disorders including T1D[86], Crohn’s disease and ulcerative colitis [311, 312]. The roles of FOSL2 have been extensively studied in murine models in the context of Th17 cell specification and systemic autoimmunity [274, 313, 314], as well as human T2D [315], but their significance in human T1D are largely unexplored. Although it was shown that T cell-specific Fosl2 deletion reduces disease severity in the experimental autoimmune encephalomyelitis (EAE) mouse model [274] and thus, it is reasonable to assume its protective phenotype in T1D, studies have revealed heterogeneity and opposing roles of FOS proteins in the context of T cell differentiation between human and mouse [313, 314, 316]. These discrepancies highlight the need for great caution while extrapolating findings from murine studies in regard to roles of FOS proteins and the necessity for validating their significance in human cells. It was demonstrated in Shetty, Tripathi [317] that FOSL1 and FOSL2 inhibit the expression of key regulators of Th17 (*IL-17, IL23R and IL23R*) in human naive CD4$^+$ T cells, and Lund, Löytömäki [318] reported that FOSL2 is preferentially expressed during early stage of Th2 cell differentiation. It has also been shown that FOSL1, FOSL2 and BATF TF colocalized in the proximity of key Th17 genes [317]. Hence, the reduced chromatin accessibility surrounding FOS binding sites and FOS transcript levels in Treg cells of T1D patients may suggest pathogenic reprogramming of suppressive Treg cells into a Th17-like, but not Th2-like lineage. The reduced expression of IL-4, which promotes differentiation of naïve CD4$^+$ T cells into Th2 cells, in the T1D Treg cells in comparison to healthy controls supports this proposition. Previous research has confirmed the pathogenic roles of FOXP3$^+$ Treg cells which acquired Th-like phenotype characterised by secretion of proinflammatory cytokines in autoimmune context [319-322] rendering them defective in suppressive capacity and as a result, they may fail to elicit an effective regulatory T (Treg) cell response in T1D. Growing evidence has established a connection between Th17 cells and pathogenesis of T1D in both animal and human studies [323-328]. Nonetheless this work has been unable to demonstrate evidence of significant upregulation of Th17 proinflammatory cytokine genes or signatures that define the Th17 phenotype in the T1D Treg cells through the epigenomic and transcriptomic experiments, which is likely to be related to the lack of adequate sample size as well as the presence of strong covariates in the data such as batch effects, gender, family pairing and HLA genotype status. The use of a mixed circulating population of CD4$^+$ T cells containing both effector and memory pools may further constrain the discovery strength of Th17 phenotype in our case-control studies as it has been demonstrated that IL-17-secreting cells was found in the memory population of CD4$^+$ T cells from adult T1D patients [329], and monocytes from children with T1D induced IL-17 production in CD4$^+$CD45RO$^+$ memory T cells [330]. Ferraro, Socci [106] further demonstrated, through the cytokine production and chemokine receptor profiles, that the upregulation in Th17 phenotype and functional defects in Treg cells were found in the pancreatic lymph nodes of T1D subjects but not in their peripheral blood. These limitations may render this study underpowered for capturing T1D-associated accessibility or expression signals especially those with small effect sizes [331]. Notwithstanding the shortcomings, the alterations in TF binding profiles, epigenome and transcriptome observed in our T1D cohort were strongly pointing to biological pathways crucial for T cell activation and differentiation. This supports evidence from many studies that reported involvement of dysregulated T cell differentiation programs in T1D setting [291, 326, 332, 333]. It is also reassuring that a subset of the T1D alterations identified in our study were linked to previously known T1D signatures, validating the capability of the present study in capturing true disease phenotype.

Although conflicting results were observed in differential TF footprinting from gender-matched (n=5 male pairs) and the full dataset (n=12 pairs including 5 mixed-gender pairs), this study is inclined to observations from full dataset (n=12 pairs) as the general consensus agrees that higher sequencing depth leads to improvements in detection power, robustness and reliability in the inference of TF footprints [141, 267, 334, 335], especially given the high variability nature of this study. Nonetheless, this study identified alterations in a network of TFs that may interact collaboratively or co-regulate targets implicated in T1D to drive changes in chromatin accessibility.

In recent years growing evidence began to overturn the long-held perception that the regulation of a gene is steered by its closest regulatory elements in linear space. Linear distance does not accurately infer regulatory contacts as enhancers can bypass nearby genes to regulate genes far apart in the linear genome sequence but which are proximate in 3D space within the nucleus [129-132]. This complexity has mostly been overlooked when assigning target genes to distal non-coding regulatory elements like gene enhancers. Enrichment of disease-associated risk SNPs, including T1D-associated variants, at non-coding regions of the genome [2, 88] presents a strong rationale to consider spatial genome organization in identifying target genes impacted by disrupted long-range enhancer-promoter interactions. As ATAC-seq does not provide evidence that formally links individual distal regulatory elements to their respective target genes and target assignment is performed on the basis of linear proximity, it may not be the best model for establishing disease candidacy and understanding molecular circuits due to the lack of physical connectivity evidence. Incorporation of published healthy T cell genome interactome data [6, 37, 134] generated by chromatin conformation capture (3C)-based technology (Hi-C experiment in Treg cells; H3K27ac HiChIP in Treg and Th17 cells; promoter capture Hi-C in activated CD4$^+$ T cells) with my case-control ATAC-seq profiles allows me to map altered enhancers in T1D to their putative target gene promoters in the 3D nuclear space, and matching case-control RNA-seq revealed the consequential transcriptome changes which are likely to result from dysregulation of gene enhancers. Using this combinatorial, integrative evidence-based approach that leveraged multiple functional annotated features from ATAC-seq, RNA-seq and Hi-C I identified 42 and 21 dysregulated enhancer-associated targets in Treg and Tconv cells of T1D cohort, respectively. The majority of these long-range gene targets would have been missed or overlooked using target assignment based on linear proximity.

Although accurate correlation between epigenome and transcriptome can be achieved by taking into consideration of spatial genome organisation in target mapping, it is also important to note that gene regulatory regions can be accessible in multiple immune populations, but as a result of occupancy of lineage determining pioneer TFs, they may only control or translate to gene expression changes in a few specific conditions [336]. By incorporating spatial genome organization information from Hi-C contact maps in my case-control ATAC-seq and RNA-seq data, I observe that many of the altered enhancers in T1D engage in long-distance interactions and often skip their nearest neighbouring genes in linear distance to interact with more distal promoters, reinforcing evidence from previous observations [6, 37, 129, 131]. In T1D these gene targets exhibited alterations in transcript levels and accessibility changes at their respective enhancers, suggesting epigenetic modifications in T1D interfere with the activity Treg- and Tconv-specific enhancers, resulting in altered expression of the genes normally controlled by these enhancers. Some of the altered gene targets in T1D included *FOSL2* and *TIGIT* in Treg cells, and *MAF* and *IL2* in Tconv cells.

TIGIT (T cell Ig and ITIM domain) is highly expressed on Treg cells and it acts as a coinhibitory receptor which inhibits T cell responses through engagement with the ligand CD155 expressed on the antigen-presenting dendritic cells (DCs), thereby inhibiting IL-12 and promoting IL-10 production [337-339]. Although the perception of the field is that TIGIT is strongly associated with enhanced immunosuppressive functions in Treg cells, intriguingly, our case-control data indicate TIGIT was upregulated in the Treg cells of T1D cohort. This observation could be a result of a compensatory mechanism of Treg cells to suppress pro-inflammatory responses in T1D. Nonetheless, over the past decade accumulating evidence has demonstrated that TIGIT expression is linked with spared or enhanced Th2 immunity [340-344]. Joller, Lozano [344] observed that TIGIT-expressing FOXP3^{+} Treg cells selectively suppress pro-inflammatory Th1 and Th17, but not Th2 cell responses, by inducing the secretion of the soluble effector molecule fibrinogen-like protein 2 (Fgl2) which promotes Th2 polarization. Kourepini, Paschalidis [340] reported that *Tigit* expression in Th2 cells is crucial for Th2 development and blockade of TIGIT restrained Th2 responses but had no effect on the Th1 and Th17 polarization in the allergic model. demonstrated that the TIGIT locus is hypomethylated and TIGIT was overexpressed in CD4^{+} T cells of children presenting with allergic asthma. Lastly, Kamran, Takai [341] showed that splenocytes and CD4^{+} T cells of immunized *Cd155*$^{-/-}$ (ligand of TIGIT) mice produced significantly less IL-4 and the Th2 promoting transcription factor GATA-3. I postulate that in T1D the upregulation of TIGIT in Treg cells could potentially direct the differentiation of Tconv cells skewing toward a Th2 phenotype or shift the cytokine profile towards a proinflammatory Th2 response as it selectively suppresses Th1 and Th17 responses while promoting Th2 polarization. Although earlier observations reported that Th1 cytokines promote whereas Th2 cytokines protect from the development of T1D [345, 346], other studies have demonstrated Th2-skewed immune response in T1D, reinforcing this concept of selective Treg-mediated suppression in T1D [332, 347, 348]. It has also been suggested that β-cell destruction is mediated by the dual role of Th1 and Th2 cytokines, depending on the cytokine and time after disease onset [348, 349]. Our transcriptome analysis has been unable to provide a strong link between Th2 differentiation and T1D in the Tconv compartment, which could be attributed to the presence of significant covariate effects within the data, and/or under-representation Th2 signal owing to the use of bulk populations of CD4^{+} Tconv cells in our study. As the Th2 population represents only 0.27–3.57% of the CD4^{+} T cell population [350] this study would benefit from larger sample size and thus statistical power in capturing such small significant differences. However, this study has been able to establish a connection between T1D and downregulation of Th1 and Th17 lineages in the Tconv cells of T1D patients. IFN-γ, which is predominantly expressed by Th1 cells [351, 352] and IL2, which was necessary and sufficient to drive Th1 differentiation in the CD4^{+} T cells of NOD mice [353], showed reduced transcript levels in the T1D Tconv cells. IL-17F, one of the six members of IL-17 family of cytokines expressed by Th17 cells [354, 355], also showed decreased transcript levels in the T1D Tconv cells. Yang, Chang [356] further support the hypothesis of Th2-mediated immune regulation by demonstrating that IL-17F-deficient mice exhibited enhanced Th2 cytokine production such as IL-4, IL-5, and IL-13. *MAF*, which was downregulated in chromatin accessibility and transcript levels in the Tconv cells of T1D patients, has been reported to have a protective role against islet inflammation [275]. Prior studies have also shown that *Maf* has a critical role in enforcing Th17 effector identity by promoting chromatin accessibility and expression of key Th17 program genes such as *Rorc* and *Il17* [163, 357, 358], which were both downregulated in our T1D group, and Maf deficiency in T cells has been associated with dysregulation of Treg – Th17 balance leading to spontaneous colitis in mice [276]. Furthermore, Xu, Yang [359] and Aschenbrenner, Foglierini [358] groups have demonstrated that MAF regulates the production of anti-inflammatory cytokine IL-10 production during in both human and mouse Th17 cells through suppression of IFN-γ and Th1 signals. The lack of IL-10 production by pro-inflammatory Th17 cells was linked with the inability of Th17 cells to upregulate c-MAF expression after being stimulated [358]. It has been reported in multiple studies that Th17 secrete immunosuppressive factors such as IL-10 to negatively regulate immune responses (termed "non-pathogenic Th17 cells") [360-363]. Finding from Bevington, Ng [364] further reinforced the immunoregulatory function of MAF as they observed induction of tolerance in T cells involves the opening of chromatin at known c-MAF binding sites. Consistent with the literature the expression of *IL-10* was downregulated in the T1D Tconv cells. IL-10 plays a crucial role in suppressing autoimmunity [365] and it was identified by GWAS as one of the T1D-associated risk loci [2]. It has been shown that NOD mice have reduced levels of serum IL-10 prior to diabetic onset [366] and intriguingly, administration of IL-10 protected NOD mice from developing diabetes [367]. Therefore, downregulation of IL-10 in the T1D patients in our case-control studies suggest impaired immunosuppressive capacity leading up to loss of tolerance against pancreatic β-cells. Overall, the dysregulated epigenome in the context of chromatin accessibility and TF dynamics as well as transcriptome as a whole strongly suggest that in T1D, cytokine balance, T cell development and differentiation are disrupted. To functionally validate the novel T1D gene candidates identified in this study such as TIGIT and MAF, the activity of their interacting enhancers can be measured in luciferase reporter assays and TF binding *in vivo* can be confirmed by ChIP-qPCR.

Study by Gao, Uzun [33] revealed transcriptome and enhancer-associated changes in Th1 and Treg cells isolated from 6 recent-onset T1D patients and 5 healthy controls using RNA-seq and H3K27ac ChIP-seq (active enhancer marks). Comparing my case-control results with theirs revealed minimal overlap (<1% overlap of DE genes; <5% of dysregulated enhancers) which could be attributed to a number of reasons, such as sample size of study cohort, cell type (bulk Treg vs memory Treg; bulk Tconv vs Th1), cell context (stimulated vs resting), source of RNA (ATAC-supernatant fraction vs whole cells) and study design (sibling-matched vs unpaired). Nonetheless, consistent with finding from this work, the dysregulated gene networks in T1D discovered in Gao, Uzun [33] were strongly associated with lymphocyte activation, T cell differentiation, cell proliferation and immune response.

Enrichment of genetic variants in cell type-specific *cis*-regulatory elements is observed in other complex human traits and autoimmune disorders as well [368] and it has been reported that they modulate gene expression in human immune cell populations including CD4$^+$ T cells (termed expression quantitative locus; eQTL) [88, 208, 294]. Because they are located in non-coding regions of the genome, uniformly proportioned across intergenic and intronic elements [369, 370], it is challenging to delineate the mechanisms by which they act to control target gene expression or mediate disease susceptibility. Functionally characterizing them is crucial to the interpretation of GWAS study and advancement of therapies in human diseases, that includes identifying the causal SNPs, mechanisms by which they act and lastly their target genes. However, it remains a great challenge to identify the true disease drivers because the genetic variants often span large genomic region due to coinheritance of many bystander or passenger SNPs, especially in the case of European-ancestry populations (upon whom most fine-mapping studies have been conducted) in which the LD blocks are larger than in other populations[371, 372], and they may involve more than one independent association. Since epigenetic changes such as chromatin accessibility, DNA methylation and histone modifications are required for normal gene regulation, genetic variants are likely to exert their function through epigenetic mechanisms.

As SNPs identified from GWAS genotyping arrays serve only as representatives for all the SNPs in the same haplotype block and it is possible that other SNPs in high linkage disequilibrium (LD) with the index SNPs are casual for the disease, I expanded my analyses using LD calculation together with the 1000 Genomes Project Phase 3 reference panels to include SNPs that are in high LD with the T1D index SNPs (i.e. "LD-SNPs") in my enrichment analyses. This has substantially expanded the list of putative causal SNPs from less than 200 index SNPs for T1D to more than 2 million. In an attempt to prioritise SNPs with high regulatory potential and functional consequences in T1D for functional follow-up studies, I employed a combinatorial, integrative evidence-based approach that take into account of T1D-associated ATAC-seq peaks and differential TF binding. Using this approach, I narrowed down the list of T1D LD-SNPs to 111 and 96 for Treg and Tconv cells, respectively, that coincided with accessible chromatin regions and TF footprints differentially regulated between T1D and healthy controls. Identification of genotype at these positions for individual patients will allow association between the genotype and TF binding as well as gene expression in this. Attempt was made to identify genotype variants at these regulatory locations from ATAC-seq profiles from T1D individuals (Table 7-10), however, no methodology for doing had yet been published during my PhD and it was not until after this thesis was submitted for examination, a study led by Massarat, Sen [373] systematically reviewed the performance and feasibility of genotyping and variant calling on ATAC-seq data. Some of the concerns about identifying genotypes from ATAC-seq data were variable insert sizes arising from nucleosomal-free and nucleosomal-occupied regions which may cause errors for indel calling, and due to the relative rarity of the alternative allele the majority of ATAC-seq reads are derived from the reference allele, variants callers may fail to identify heterozygous variants. Furthermore, variants within the regulatory genomic loci may not be detectable if they substantially disrupt those regions resulting in the loss of accessibility. Massarat, Sen [373] has systematically reviewed the performance of seven indel and variant callers and developed an ensemble workflow, VarCA (variants in chromatin accessible regions), which combines features from multiple callers and achieves optimal performance for use on ATAC-seq data. This novel workflow provides us with confidence and a tool to identify variant genotypes from ATAC-seq dataset, which we will test and apply on our case-control profiles thereafter.

Each variant in isolation may have a subtle contribution on disease risk but each may alter a key function in the immune system and its interaction with β cells. The candidate causal LD-SNPs were significantly associated with the T1D-associated regulatory elements identified from this study, supporting strong functional relevance to disease causality. This observation supports the involvement of Tconv and Treg cells in the pathogenesis of T1D. It also suggests that T1D SNPs are more likely to exert their effect through accessible enhancers and TF binding sites significantly altered in T1D in T cells. To reveal potential disease mechanisms attempt was made to identify the target genes of these candidate LD-SNPs through incorporation of spatial genome connectome data from Hi-C contact maps generated from healthy adults [6, 37, 134], however, they do not reflect corresponding transcriptome changes in the case-control

RNA-seq, suggesting the conformation is most likely context- or genotype- specific. To experimentally identify the target genes of the candidate LD-SNPs, one can manipulate the enhancers containing the LD-SNPs in question and determine if expression of the putative target genes is in fact altered. Physical interactions between the SNP-containing enhancers and putative target genes can be validated through looping assay such as 3C-qPCR. CRISPR genome editing can be used to epigenetically modify the enhancers to introduce SNPs in their endogenous context in primary T cells and functional readout can be accomplished by qRT-PCR.

Most disease-associated variants identified by GWAS confer relatively small increments in risk individually or in combination and represent only a small fraction of heritability predicted from familial clustering [279]. It has been proposed that a large number of highly polygenic causal variants with small effect sizes have yet to be captured which can account for the “missing heritability” [280, 281]. To detect these genetic effects, in this study I have also explored whether common genetic variation found within populations have the potential of modulating T1D risk by altering enhancer activity and target gene expression. As an alternative approach to identify novel putative T1D-associated human genetic variation independently of GWAS, I surveyed 4,655,805 common variants (MAF > 10%) from a large population-scale variant resource, Genome Aggregation Database (gnomAD) (version 3.0) [1] as input to my filtering workflow, comprising of differentially accessible genomic regions between T1D cohort and healthy controls, enhancer- and promoter-associated long range regulatory regions. The same set of gnomAD common variants were studied in Liu, Sadlon [134] and the authors demonstrated 55.7% of the gnomAD variants were not included in the largest meta-analysis of T1D genome-wide genotyped datasets to date [86] and their potential in mediating T1D risk was not assessed in sampled case-control populations. A total of 92 and 50 candidate gnomAD SNPs were located within accessible enhancers differentially regulated in T1D patients and involved in long-range enhancer-promoter interaction in Treg and Tconv, respectively, suggesting they can potentially alter T cell-specific gene expression in T1D through spatial interactions. The filtered variants were enriched in TF binding motifs that have a key role in T cell development and differentiation, including YY1 and FOXP3 for Treg cells, and MAF for Tconv cells. SNPs in those regulatory elements can alter, by either disrupting or creating new TF binding sites in T1D-associated enhancers or alter chromatin accessibility, resulting in loss of long-range, physical enhancer-target interaction. In accordance with this hypothesis, Gao, Uzun [33] demonstrated that the binding of YY1 was disrupted by a SNP in high linkage disequilibrium with the T1D-associated GWAS lead SNP (rs883868) at Treg enhancer, resulting in loss of long-range enhancer–promoter interaction. Polymorphisms in *Foxp3* lead to systemic autoimmune pathology in human and mice, including T1D [108, 250]. Finding from Kivling, Nilsson [374] broadly supports the work of other studies in this area demonstrating PBMCs from children with T1D exhibited a lower FOXP3 expression [375] than children with celiac disease (CD), or with T1D in combination with CD, establishing a link between impaired regulation of FOXP3 and T1D. Supporting the pathogenic role of Treg and Tconv cells in T1D, the candidate gnomAD variants were significantly enriched in genomic regions differentially regulated in T1D from T cells and they have a key role in IL-2 signalling pathway, regulation of apoptosis in Treg and IFN signalling in Tconv cells. Functionally annotating by incorporating evidence of T1D-associated traits and regulatory elements enabled us to identify a subset of novel putative T1D variants for follow-up studies, as well as revealing potential molecular circuits by which non-coding variants regulate causal genes in disease-relevant cells. It is hoped that associations identified by GWAS or other genomic approaches will eventually achieve disease prevention or treatment, either through functional characterization of existing variants, or identification of new variants in which true functionality lies.

Not only has this work allowed insight into the molecular mechanism of action by which the genetic risk loci may contribute to the loss of tolerance in T1D, but it has also revealed a list of novel biomarkers and gene candidates potentially for use in future immunomodulatory intervention and therapies in T1D. These novel targets, such as TIGIT and MAF, and their signalling pathways can be incorporated in combination therapy regimens together with existing T1D targets in trials that involve autoantigen-specific immune modulation and induction of Treg to achieve long-term tolerance in T1D. At the stage at which T1D can be predicted accurately or is clinically manifested the loss of β cells is significant, rendering therapies less effective in reversing or halting the autoimmune attack. If we were able diagnose or identify the risk for T1D at early stages, we could potentially intervene its progression with treatments to preserve functional β cells or prevent long-term complications of T1D.

Although this preliminary work has successfully identified changes in epigenome and transcriptome between T1D patients and healthy individuals, it does not reveal if they are the driver or the consequence of the disease pathogenesis because the T1D cohort involved in this study was derived from a single time point i.e. post clinical diagnosis. It is important to distinguish driver from passenger roles for epigenetic alterations to facilitate diagnosis, prognosis and development of therapies in T1D. In order to tell causality, we would need to track those signatures over disease progression in longitudinal studies, preferably from pregnancy into early childhood, by applying the same molecular technologies and bioinformatics pipelines used in this study. In addition, the credibility of the findings presented in this work could also be improved with experimental perturbations models as described previously.

Apart from single timepoint measurements, this study is also constrained by a number of other limitations such as modest sample size and presence of strong confounding variables arising from covariate effects such as biobanking batch, gender, HLA genotypes and high variability between study subjects which limits the statistical power of the present study. Although data normalisation has been performed to reduce the effects of dominating confounding variables as part of the data analysis workflow, associated biological effect of interest might be removed along with the unwanted variation during the process. Future studies should also take into account of the identity and source of cell populations as it has been reported that disease signatures were observed at the site of infection, i.e. pancreatic lymph nodes rather than circulating peripheral blood in T1D subjects. The use of a mixed population of CD4$^+$ T cells further renders this study underpowered for capturing disease signature arising from individual subtypes as they are underrepresented in the bulk population and thus, this field would greatly benefit from a comprehensive characterisation of cellular diversity in this context using low input genomic assay such as single cell technology.

CHAPTER 6: General Discussion and Future Directions

6.1 General Discussion

The reason a child develops T1D is not well understood but it appears to depend on a complex interaction of genetic predisposition and environmental factors that trigger or permit the autoimmune response against the pancreatic β-cells. Work over the past decades has strengthened the concept that environmental factors contribute strongly to the pathogenesis of T1D and they are likely to exert their function through epigenetic mechanisms, leading to a breakdown in tolerance normally maintained by Treg cells. But the reason why Treg cells stop working in T1D is not yet defined. Genome-wide association studies (GWAS) [2, 3, 85-87] have identified 59 non-HLA genomic loci that are mostly associated with the risk of developing T1D. Importantly, those T1D risk variants are located within T cell enhancers [2, 88] and thus, are likely to alter the transcriptional regulation of genes. Enhancers ensure the tight control of their target genes and loss of this regulation arising from genetic polymorphisms can lead to defective suppression by Treg cells in T1D. However, how they, in concert with environmental factors, mediate disease susceptibility and contribute to the loss of tolerance in T1D has remained elusive.

The general hypothesis for this thesis is that genetic variation and epigenetic changes in T1D modify the activity of Treg- and Tconv-specific enhancers, leading to altered expression of the genes normally controlled by these enhancers. In testing the hypothesis this study has successfully created T cell epigenome and transcriptome maps of human T1D. This work has revealed significant alterations in the enhancer repertoire, transcriptome and TF dynamics in primary Tconv and Treg cells of an established T1D cohort. To establish a connection between T1D-relevant enhancers and transcriptome changes, this work has successfully identified enhancer-promoter interactions by integrating features derived from epigenome, transcriptome data and spatial interactome maps. Intersecting the T1D epigenomic data with a catalogue of SNPs located in previously reported T1D risk loci identified by GWAS and other genetic studies [2, 3, 85-87, 295], this work has identified a subset of associations with high regulatory potential and functional consequences in T1D, for follow-up studies. In addition to prioritisation of T1D associations, incorporation of a catalogue of common human variants from the Genome Aggregation Database (gnomAD) [1, 134] in my epigenomic data has identified a subset of novel putative T1D-associated genetic variants, independently of association studies, located within T cell enhancers that I have now identified and demonstrated to be altered in clinical samples from a T1D cohort. The study presents the first comprehensive resource of the chromatin accessibility and matching transcriptomic profiles in effector and suppressive compartments of $CD4^+$ T cells in T1D patients, supported by 3D connectivity evidence, in establishing strong candidacy of dysregulated gene targets in T1D.

The findings in this thesis presents the first resource, using cutting edge ATAC- (Assay for Transposase-Accessible Chromatin) and RNA- sequencing technologies, that integrates the epigenome and transcriptome of human T1D in the T cell compartment to better understand the genetic and epigenetic control of tolerance and effector function in T1D. This allows identification of changes driven by both genetic and epigenetic variation that correlates with an altered cell type-specific transcriptional program in T1D.

This project had the advantage of having access to a biobank repository of PBMC samples obtained from an established T1D cohort with sibling-matched healthy controls. I first established an optimised thawing protocol to improve the viability and recovery of cryopreserved cells for use on downstream ATAC-seq and RNA-seq experiments. This effort was important because cryopreserved cells often suffer from poor viability and functionality but ATAC-seq requires a high viability population of intact cells [153] to generate good quality datasets, as tagmentation of free DNA released by the dead cells leads to high background noise. Using PBMCs obtained from healthy adult donors, I identified thawing conditions that led to optimal PBMC viability and recovery for isolation of sufficient number of viable T cells. It has been reported that an overnight resting phase improves the elimination of dead or dying cells from cryopreserved PBMC samples, aside from increased functionality of recovered T cells [183, 196]. This is because the population of cryopreserved cells consist of viable, early apoptotic, late apoptotic, as well as dead cells [196] and a resting period enables apoptosis-prone cells to die and clearance of apoptotic cells results in a PBMC population with a higher representation of truly viable cells. Hence, cell viability alone is not sufficient to assess cell health upon thawing as high cell viability can be a result of high rate of cell death, apoptosis and dead cell clearance. Nonetheless, my results accord with observations from Hønge, Petersen [197] where it was reported that a longer centrifugation duration with higher force (10 minutes at 500xg) results in significantly higher yield of recovered PBMCs upon thawing in comparison with a shorter centrifugation duration with lower force (5 minutes at 300xg). Because the optimisation was performed on adult PBMCs following the protocol used for the cryopreservation of the case-control PBMC samples, I anticipate the results generated here serve as an appropriate proxy for performance of biobanked cohorts in general. To further confirm the qualitative and quantitative attributes of the thawed material, future work can incorporate T-cell phenotypic and functional assays such as ELISPOT and intracellular cytokine staining to evaluate the expression of co-stimulatory and inflammatory markers in cryopreserved cells.

Having identified the optimal conditions for thawing cryopreserved PBMCs I then sought to determine the feasibility and reproducibility of using ATAC-seq to infer chromatin structure in the biobanked T1D case-control samples. Using PBMC samples derived from healthy adult donors, I performed Omni ATAC-seq [153] on purified fresh and frozen Treg cells under resting and stimulated states. The results were very promising as it indicates that the fresh and thawed Treg ATAC-seq samples had excellent concordance in terms of library complexity, chromatin accessibility, and signal to noise ratios, TF-DNA interaction and responsiveness to stimulation. The chromatin accessibility landscape was preserved genome-wide in the thawed Treg cells especially at loci critical for Treg function such as *IL2RA*, which has been previously associated with T1D risk [2, 3], suggesting that the biobanking process does not perturb epigenetic signatures and the samples can be reliably used to identify gene loci and TF networks impacted in T1D using ATAC-seq. It is encouraging that cryopreservation did not impair the ability to resolve TF footprints from the accessibility profiles and had no global effect on the TF-DNA interactions during resting and stimulated states. These TFs include FOXP3, NFκβ and RUNX1 and they play a critical role in the maintenance of immune tolerance and homeostasis of the immune system [201-203]. This finding broadly supports the work of other studies in this area including Scharer, Blalock [30] who observed high correlation of signal to noise ratios, chromatin accessibility and TF footprint patterns between fresh and biobanked CD19^{+} naive B cell ATAC-seq samples. In addition, Fujiwara, Baek [182] also observed that the chromatin landscape from cryopreserved breast cancer cells and mammary tissue closely recapitulated that of the fresh cells and tissue. My findings build on existing evidence and demonstrate indistinguishable chromatin accessibility and TF occupancy in biobanked primary T cells under resting conditions and in response to stimulation. This gave me confidence that the methodological approach was applicable to the subsequent case-control study as many autoimmune disease states are linked with activation of immune cells [206, 207]. Aberrant T-cell homeostasis and activation have been reported in autoimmunity, including T1D, and a fraction of genetic variants have been linked to transcriptional regulation and gene expression taking place during T-cell activation and polarization [208, 209]. Overall, the results indicate that the biobanked samples can be reliably used to identify alterations in epigenetic regulation in T1D using ATAC-seq during both resting and stimulated state.

As this project proposed to profile both the global chromatin accessibility and gene expression changes in T1D, and was constrained by limited resources of biobanked samples, I explored the feasibility of extracting both the chromatin accessibility and transcriptome information simultaneously from a single population of 50,000 cells. The rationale was that, in ATAC-seq protocol, centrifugation is carried out after cell lysis separating the lysis reaction into two fractions - pellet which contains lysis products derived from the nucleus (for ATAC-seq) and supernatant fraction which contains lysates derived mainly from the cytoplasm and other cellular organelles. As the supernatant fraction (ATAC-SN) is not required in the ATAC-seq workflow, I attempted to recover transcriptome from it and asked whether its representation was comparable with the donor matched whole cell lysate. As RNA material derived from ATAC-SN constitutes RNA mainly from cytoplasmic compartment and lacks nuclear components, whereas total cellular extract theoretically has both, we expect to see a subcellular localised expression pattern in the analysis. A number of studies have observed subcellular localization of RNA transcripts and they reported a small proportion of significant differences between the nuclear and cytosolic transcriptome [189-191, 210, 211, 213-217]. Given that a majority of the ATAC-SN fraction consists of cytoplasmic RNA and studies have reported that the expression patterns for subcellular RNA components from total and cytoplasmic compartments were strongly correlated [215, 218], while some even argued that whole cell RNA may be a misleading representation of cytoplasmic steady-state mRNA levels stemming from highly variable expression of nuclear transcripts [190], this section of work hypothesised that the cytoplasm-derived ATAC-SN and whole cell fractions demonstrated concordant distribution of transcript abundance.

In line with the hypothesis, this study observed a high concordance in relative distribution of transcripts derived from ATAC-SN and total cellular extracts. The cytoplasm-derived ATAC-SN and whole-cell fractions displayed a high overlap of transcript populations where 91.1% of transcriptome was conserved in both compartments. As expected, transcripts that showed increased expression in the whole cell transcriptome are linked with biological processes and molecular functions associated with the nucleus, including nuclear division, nuclear membrane, histone modification, chromosome organization and glycoprotein. Whereas for transcripts that showed increased expression in the cytosol-derived ATAC-SN fraction, they are involved in membrane-associated and translational machinery processes such as co-translational protein targeting to membrane, protein targeting to endoplasmic reticulum (ER), ribosomes, as well as cytoplasmic translation, catabolic and metabolic processes.

An alternate explanation for the differential expression between whole cell lysate and ATAC-SN RNA seq could be due to the enrichment of transcripts in the whole cell fraction with respect to ATAC-SN could also indicate high transcription rate for these genes, low rate of release to the cytosol and/or rapid degradation in the cytosol. Whereas the enrichment of transcripts in the ATAC-SN with respect to the whole cell fraction is likely to be related to their high stability in the cytoplasm and low transcription rates. Importantly, the overall percentage overlap and gene annotation suggest that the explanation is simply that by excluding nuclear RNA in the ATAC-SN, that gene set is missing compared with the whole cell lysate. As all samples prepared from ATAC-SN in the case control are treated identically, this nuclear transcript signal does not contribute to the final experimental differential expression analysis. Overall, the data suggest that transcriptome derived from ATAC-seq supernatant fraction closely recapitulates that of the whole cell transcriptome and thus this approach can be reliably applied to the biobanked paediatric case-control samples to infer transcriptome changes alongside chromatin accessibility alterations.

When the research presented in this PhD thesis was initiated there was very little published on the computational framework dedicated to analysing ATAC-seq data. The optimal approaches for ATAC-seq peak calling, data normalisation and differential accessibility remained largely unexplored. Although some popular computational algorithms commonly used for peak calling and differential analysis such as MACS2[228, 229] and edgeR [38] offer the flexibility of cross-platform application on ATAC-seq data, they failed to address the suitable parameters or adjustments needed for the programs to perform optimally on or utilise full information contained in ATAC-seq data. These conventional computational tools fall short as they usually fail to account for additional nucleosome positioning information contained in the Tn5 transposase-digested DNA fragments and therefore could potentially result in erroneous data modelling contributing to spurious peak identification or flawed control for type I/II errors. As my work aimed to identify the changes in chromatin accessibility in T cells recovered from the biobanked T1D cohort, it is critical to master the best practices of the computational framework for analysing ATAC-seq dataset to ensure accurate interpretation of epigenomic data for clinical translation. To achieve that I assessed and benchmarked analysis strategies in the areas of peak calling, differential accessibility (DA) analysis and data normalisation for ATAC-seq data.

Interestingly, investigation led by Gaspar [159] revealed that a large number of ATAC-seq studies had used MACS2 to call peaks using parameters that were inappropriate and suboptimal for ATAC-seq datasets, resulting in erroneous or poor quality peak identification. I incorporated the modifications recommended in Gaspar [159] in my peak calling algorithms and compared the outcomes with the original algorithms. Overall, using modified peak calling algorithms I observed higher number of peaks called, higher proportion of intergenic peak assignment where enhancers are enriched in [7, 88], excellent signal-to-noise ratio and smaller peak size coverage compared with original peak calling mode in both resting and stimulated T cell ATAC-seq libraries, which leads to improvements in the resolution of peaks and confidence for subsequent count-based (differential analysis) and motif-specific (TF footprinting) analyses. These modifications were adopted for incorporation in peak calling of case-control ATAC-seq data.

With optimal peak calling algorithms established, I then set out to determine whether the differential accessibility analysis can reliably identify genomic regions that show differential regulation between Tconv and Treg cells. As the Barry Lab has established considerable prior biological knowledge on gene regulation in Tconv and Treg cells based on previously published FOXP3 ChIP-seq study [11] this model can be reliably used for assessing and benchmarking analytical methods for calculating differential accessibility (DA).

Different methods of ATAC-seq DA analysis can lead to conflicting results, and statistical considerations of normalisation are still being developed. Using Tconv and Treg ATAC-seq data, I tested different data normalisation approaches, including the approaches described in Reske, Wilson [230], to determine the workflow that supports biologically expected changes between Tconv and Treg gene regulation. The DA results showed that normalisation using TMM (Trimmed mean of M values) and loess (locally estimated scatterplot smoothing) for generation of linear scaling factors from raw peak counts are robust strategies as a symmetric global accessibility distribution was observed, indicating absence of trended bias within the data. The DA results between Tconv and Treg were very encouraging as they reflect biologically expected changes. It successfully captured critical loci and signalling pathways that strongly reflect the biology and differential regulation between the two cell types, in accordance with findings from a large number of RNA-seq, microarray and flow cytometry studies [235-237, 247]. Some of DA gene loci identified between Tconv and Treg include TF and cell surface molecules such as *FOXP3*, *TIGIT* (T Cell Immunoreceptor with Ig and ITIM Domains), *IKZF2* (IKAROS Family Zinc Finger 2) and *CTLA4* (cytotoxic T-lymphocyte-associated protein 4) in the top differentially accessible peaks. Whereas the pathway enrichment analyses showed that the DA regions were strongly involved in T cell differentiation, cellular response to cytokine stimulus, T cell receptor signalling pathway and IL2-STAT5 signalling. Overall, the results show that TMM and loess normalisation methods show high concordance in DA outputs which are biologically relevant to my T cell datasets. TMM was chosen as the normalisation method of choice with which to proceed forward for use on the T1D case-control ATAC-seq.

In summary, initial work from my PhD has successfully established a thawing protocol which results in significant recovery of cryopreserved cells. In addition to that, my work also confirmed that neither the cryopreservation nor thawing process compromised the chromatin accessibility architecture and TF activity of Treg cells, and it is feasible to infer transcriptome from cytosol-derived ATAC-seq lysate fraction. Importantly, I have also successfully identified the optimal peak calling, data normalisation and differential analysis strategies for accurate annotation of genomic regions differentially regulated between primary conditions of interest. Through these prior work I have established and mastered the technical methodologies and bioinformatics competence necessary to proceed and apply the techniques and pipeline on the biobanked T1D cohort and healthy controls.

I then proceeded to conduct epigenomic and transcriptomic profiling of Treg cells isolated from a biobank cohort of 12 individuals with established T1D and 12 sibling-matched healthy controls. Although thawing revealed that some PBMC samples suffered from poor cell viability which was a biobanking batch effect, there was no significant difference in cell viability and recovery between T1D cohort and healthy control samples. This was expected as blood specimens obtained from the cohorts were processed in pairs, i.e. T1D and accompanied matched healthy control, and thus batch effects within pairings would presumably be lower than those observed between pairs. No significant differences were observed in the proportion and frequency of Treg or Tconv cells recovered between the T1D cohort and healthy subjects. This finding is contrary to study from Viisanen, Gazali [376] which has reported altered frequency of total ($CD4^+CD25^+CD127^{lo}$) and naïve ($CD4^+CD25^+CD127^{lo}CD45^{RA+}$) Treg cells in children with recent-onset T1D. A small, but statistically significant (6.3 ± 1.7% vs. 5.3 ± 1.7%, $P < 0.000$) increase in Treg frequency was observed in the T1D cohort compared with healthy controls [376]. This rather contradictory result may be driven by large sample size (74

T1D and 180 healthy controls) and HLA-matched pairings in their study design, which increases the statistical power and strength for detection of small differences.

Delving into the epigenome and transcriptome of T1D cohort and sibling-matched healthy controls reveals significant alterations in chromatin accessibility, transcript levels and TF dynamics in Treg and Tconv cells.

High resolution linear epigenomics map I generated by ATAC-seq revealed a subset of T cell enhancer-associated alterations in Tconv and Treg cells of 12 T1D patients and 12 sibling-matched healthy controls. A fraction of the differentially accessible (DA) regions or differentially expressed genes were annotated to risk loci previously known to be associated with T1D [2, 3, 85-87, 295], corroborating the gene expression-based findings from other studies. However, none of these studies considered the chromatin accessibility or 3D connectivity. The epigenomic and transcriptomic profiles of our case-control cohorts functionally annotate some of the previously identified T1D risk loci (on average ~1.2% overlap between case-control genomics experiments and GWAS associations) showing that genetic risk can be linked by chromatin conformation to regions not previously identified, but not all of the 59 loci are captured in this cohort. There are several possible explanations for this observation. First, genetic variants identified by GWAS have small effect sizes that together only represent a fraction of the heritability [279] and thus there may be other gene signatures involved in controlling the T1D susceptibility that are not yet captured or sampled in published GWAS. Secondly, this case-control study involves a small sample size, and it is unlikely that this will cover all of the known genetic risk. In addition, the modifications of chromatin accessibility that intersect with a specific SNP could be temporally active, and as this cohort is captured at a single time point (i.e. established T1D cohort) some changes may have normalised by this stage of disease. It is also possible that there may be differences between epigenetic signatures in the cells isolated from circulating peripheral blood compared with pancreatic lymph nodes draining from the site of pathology, which are not feasible to collect from children [106].

By combining transcriptomics and chromatin accessibility datasets, it has been possible to infer mechanisms by which transcription factor expression and footprint accessibility in their target genes could cause an additive gain or loss of function. For example, my results show that the binding dynamics of Activator Protein-1 (AP-1) TF family such as FOS, FOSL1 and FOSL2 was weaker in both the Tconv and Treg cells from the T1D cohort in comparison with healthy controls and in addition the expression levels of FOSL2 are reduced in Treg cells in T1D (Figure 5-38). This has implications for the altered expression of a large number of genes that are themselves not targets of T1D genetic risk.

AP-1 proteins are expressed in CD4$^+$ Tconv and Treg cells [307] and *FOSL2* is a risk locus associated with multiple autoimmune disorders including T1D [86], Crohn's disease and ulcerative colitis [311, 312]. The reduced chromatin accessibility surrounding FOS binding sites and FOS transcript levels in Treg cells of T1D patients may suggest pathogenic reprogramming of suppressive Treg cells into a Th17-like lineage, as supported by enrichment of pathways that promotes differentiation of Th17 cells, such as TNF-α and TGF-β signaling, in the case-control transcriptome data (Figure 5-40). Previous research has confirmed the pathogenic roles of FOXP3$^+$ Treg cells which acquired Th-like phenotype characterised by secretion of proinflammatory cytokines in an autoimmune context [319-322] rendering them defective in suppressive capacity and as a result, they fail to elicit an effective regulatory T (Treg) cell response in T1D. Increasing evidence has established a connection between Th17 cells and pathogenesis of T1D in both animal and human studies [323-328].

However it was not possible in this study to demonstrate direct evidence of upregulation of Th17 signatures in the T1D Treg cells, which could be attributable to the lack of adequate sample size as well as the presence of multiple covariates in the data, such as batch effects, gender, family pairing and HLA genotype status. The fact that this study performed profiling using a bulk CD4$^+$ Tconv/Treg population may further compound the inability to discover a Th17 gene set enrichment may simply reflect the fact that Th17 are a rare subset so are underrepresented. It has been reported that IL-17-secreting cells were found in the memory population of CD4$^+$ T cells from adult T1D patients [329], and monocytes from children with T1D induced IL-17 production in CD4$^+$CD45RO$^+$ memory T cells [330]. Ferraro, Socci [106] observed that the upregulation in Th17 phenotype and functional defects in Treg cells were found in the pancreatic lymph nodes of T1D subjects, but not in their peripheral blood. These limitations are likely to render this study underpowered for capturing T1D-associated accessibility or expression signals with small effect sizes in the PBMCs [331]. Nonetheless, the alterations captured in the TF binding, epigenome and transcriptome in the T1D cohort were associated with biological pathways crucial for T cell activation and differentiation. This supports evidence from a great deal of work which reported involvement of dysregulated T cell differentiation programs in T1D setting [291, 326, 332, 333]. This type of analysis can connect changes in expression to accessibility and transcriptional regulation, shedding new light on the mechanism of action in T1D Treg and T conv, but there is still a significant gap in knowledge because genetic risk of T1D in non-coding regions has to be accurately linked to the target genes it affects to model the defect.

To understand the impact of differentially accessible enhancers in T1D on the transcriptome, we need to know their target genes. Increasing evidence over the past decade begin to overturn the commonly held assumption that genes are regulated by their closest regulatory features in linear genome sequence [129-132]. Furthermore, enrichment of disease-associated risk SNPs, including T1D-associated variants, at non-coding regions of the genome [2, 88] presents a strong rationale to take into account the 3D genome conformation in mapping target genes impacted by disrupted long-range enhancer-promoter interactions. As ATAC-seq does not provide evidence that formally links individual distal regulatory elements to their respective target genes, it is not the best model for establishing disease candidate connectivity by DNA looping, due to the lack of physical connectivity evidence. To circumvent this, I integrated published T cell Hi-C-based contact maps [6, 37, 134] in my case-control ATAC-seq data to connect altered enhancers in T1D to their putative target genes in 3D space. I then correlated the target genes with matching case-control RNA-seq profiles to infer transcriptome changes. This combinatorial strategy leveraged functional annotated features from ATAC-seq, RNA-seq and Hi-C and identified 42 and 21 dysregulated enhancer-associated targets in Treg and Tconv cells of T1D cohort, respectively (Figure 5-52, Figure 5-54 and Table 5-8). Gene ontology analyses revealed that gene targets identified from the Treg dataset are strongly associated with IL2-STAT5 signalling, whereas gene targets identified from the Tconv cells are significantly enriched in pathways associated with known T1D signatures and Th2 differentiation (Figure 5-59). In T1D these gene targets exhibited alterations in transcript levels and accessibility changes at their respective enhancers, suggesting epigenetic modifications in T1D interfere with the activity of both Treg- and Tconv-specific enhancers, resulting in altered expression of the genes normally controlled by these enhancers.

I have made novel findings by adding 3D connectivity to the case-control ATAC-seq data to reveal long-range promoter-enhancer associations, and asked if their expression patterns were also significantly altered in our case-control transcriptomic data (Figure 5-52, Figure 5-54 and Table 5-8). *TIGIT* (T cell Ig and ITIM domain), which was upregulated in the Treg cells of T1D cohort, is highly expressed on Treg cells and it inhibits T cell responses through engagement with the ligand CD155 expressed on the antigen-presenting dendritic cells (DCs) [337-339]. Upregulation of TIGIT in Treg cells of T1D subjects from our case-control data could be a result of a compensatory mechanism of Treg cells to suppress pro-inflammatory responses in

T1D. New evidence has demonstrated that TIGIT expression is linked with positive regulation of Th2 immunity [340-344]. Joller, Lozano [344] observed that TIGIT-expressing FOXP3$^+$ Treg cells selectively suppress pro-inflammatory Th1 and Th17, but not Th2 cell responses, by inducing the secretion of fibrinogen-like protein 2 (Fgl2) to promote Th2 polarization. Kourepini, Paschalidis [340] reported that *Tigit* expression in Th2 cells is crucial for Th2 development and blockade of TIGIT restrained Th2 responses but had no effect on the Th1 and Th17 polarization in the allergic model. Interestingly, TIGIT locus is also hypomethylated [342] (a marker of activation of gene transcription) and overexpression was reported in CD4$^+$ T cells of children presenting with allergic asthma. Kamran, Takai [341] further reinforced its role in Th2 immunity by demonstrating that splenocytes and CD4$^+$ T cells of immunized *Cd155*$^{-/-}$ (ligand of TIGIT) mice produced significantly less IL-4 and the Th2 promoting transcription factor GATA-3. As a result, in T1D the upregulation of TIGIT in Treg cells could potentially direct the differentiation of Tconv cells skewing toward a Th2 phenotype or shifts the cytokine profile towards a proinflammatory Th2 response, as the TIGIT$^+$ Treg selectively suppresses Th1 and Th17 responses. Existing evidence has suggested a pathogenic role of Th2-skewed immune response in T1D [347, 348]. One limitation of this T1D cohort is the sample size as we were unable to establish a strong connection between Th2 differentiation and T1D in the corresponding Tconv compartment, which could further be compounded by under-representation of Th2 signal owing to the use of bulk population of CD4$^+$ Tconv cells in our study. In addition, linking changes to the timing of loss of tolerance is not possible, as there is only a single time point samples, and it is post insulin dependence, by which time the gene expression may have normalised. However, this study has been able to establish a plausible connection between T1D and downregulation of Th1 and Th17 lineages in the Tconv cells of T1D patients. Th1 signatures such as IFN-γ [351, 352] and IL2 [353] showed decreased expression in the T1D Tconv cells (Figure 5-39). Th17 signatures such as IL-17F [354, 355] and MAF [163, 357, 358] also showed reduced expressed in the T1D Tconv cells (Figure 5-39). Finding from Yang, Chang [356] support our hypothesis of Th2-mediated immune regulation in T1D as they reported enhanced Th2 cytokine production in IL-17F-deficient mice.

Incorporation of Hi-C in our case-control ATAC-seq and RNA-seq data has also connected altered enhancers in Tconv and Treg cells to *MAF* locus. *MAF* expression was downregulated in both the Tconv and Treg cells of T1D patients. Our analysis revealed that the differentially accessible enhancer (down in T1D compared with healthy) which regulates *MAF* in the Tconv cells contains a T1D-associated LD SNP and the locus has also been previously linked to other autoimmune disorders (Table 5-8). MAF has been reported to have a protective role against islet inflammation [275] and spontaneous autoimmunity [377]. Maf deficiency in CD4$^+$ and CD8$^+$ T cells has been associated with dysregulation of Treg – Th17 balance, leading to spontaneous colitis in mice [276]. Singh, Colberg [275] showed that loss of Maf impairs both β- and T cell function, promoting islet inflammation and augmenting the risk of autoimmune islet cell destruction. Furthermore, Xu, Yang [359] and Aschenbrenner, Foglierini [358] groups demonstrated that MAF regulates the production of anti-inflammatory cytokine IL-10 production in Th17 cells through suppression of IFN-γ and Th1 signals. *IL-10* has been previously associated with T1D risk [2], and consistent with the literature, *IL-10* was downregulated in our T1D Tconv cells. It has been reported in multiple studies that secretion of immunosuppressive IL-10 in Th17 cells is a phenotype of non-pathogenic Th17 population crucial for negative regulation of immune response [360-363]. Finding from Bevington, Ng [364] further strengthened the immunoregulatory function of MAF as they observed induction of tolerance in T cells involves the opening of chromatin at known c-MAF binding sites. We speculate that MAF deficiency in T1D Treg cells impairs the suppressive functions of Treg in preventing immune destruction of Tconv cells against pancreatic β cells. Work from Neumann, Blume [378] has demonstrated that Maf enforces the identity and function of intestinal Treg cells, in line with this published study [378], common signatures resulting from mouse Maf deficiency were replicated in our human T1D Treg cells, including decreased expression of TNFS11, FOS, FOSL2, EGR2 and LAG3. This suggests that in T1D, loss of MAF may render Treg cells defective in establishing immune homeostasis, leading to autoimmune destruction against islet cells. Furthermore, our differential footprinting data has revealed significant downregulation of TF activity of AP-1 proteins such as FOS and FOSL2 in the T1D Tconv cells (Figure 5-33), and it has been shown that the consensus binding sites of AP-1 and MAF are identical [379]. Overall, the dysregulated epigenome in the context of chromatin accessibility and TF dynamics as well as transcriptome as a whole strongly suggest that in T1D, cytokine balance, T cell development and differentiation are disrupted.

Having demonstrated alterations in epigenome, transcriptome and enhancer-promoter interactions, I next sought to determine and prioritise T1D SNPs that have potential to drive a clinically relevant regulatory alteration, leading to functional consequences in T1D, for future functional follow-up studies. Because GWAS arrays do not contain all the mapped SNPs but representative SNPs in the same linkage disequilibrium (LD) block, and it is equally likely that other SNPs in high LD with the array SNPs are causal for the disease, I expanded my analyses with LD calculation resulting in ~2 million of T1D associated LD-SNPs. I successfully narrowed down the list of T1D LD-SNPs to 111 and 96 for Treg and Tconv cells (Figure 5-72, Figure 5-73), respectively, that intersected with accessible chromatin regions and TF footprints differentially regulated between T1D and healthy controls. SNPs in these regions can alter, by either disrupting or creating new TF binding sites in T1D-associated enhancers or alter chromatin accessibility, resulting in altered expression of the target genes. The candidate causal LD-SNPs were significantly enriched in the T1D-associated regulatory elements, suggesting that T1D SNPs are more likely to exert their effect through accessible enhancers and TF binding sites significantly altered in T1D in T cells. Identification of genotype at these positions for individual patients will allow correlation between the genotype and TF binding as well as gene expression in this cohort. Some of the conundrums with identifying genotypes or variants from ATAC-seq data was variable insert sizes deriving from nucleosomal-free and nucleosomal-occupied regions which may cause erroneous indel identification, as well as the inability to identify heterozygous variants as ATAC-seq reads primarily originate from one allele. It was not until after this thesis was submitted for examination, a study led by Massarat, Sen [373] systematically reviewed the performance of multiple callers to genotype and call variants from ATAC-seq data, and they have also developed a novel workflow, VarCA, which was shown to achieve optimal performance for use on ATAC-seq data. Massarat, Sen [373] provides this study with the confidence and tool to identify the genotypes of T1D risk variants from our case-control ATAC-seq profiles thereafter.

Genetic variants identified by GWAS confer relatively small increments in risk individually or in combination and account for only a small fraction of heritability [279]. Some explanations for the “missing heritability” include a large number of highly polygenic variants with small effect sizes have yet to be captured, and rare variants which are excluded from current genotyping array technologies that focus on variants present in 5% or more of the population [280, 281]. To detect these genetic effects, this study also explored whether common genetic variation found within populations have the potential of modulating T1D risk by altering enhancer activity and target gene expression. To identify novel putative T1D-associated human genetic variation independently of GWAS, I surveyed 4,655,805 common variants (MAF > 10%) from a large population-scale variant resource, Genome Aggregation Database (gnomAD) (version 3.0) [1] and identified SNPs that fall within case-control differentially accessible genomic regions, enhancer- and promoter-associated long range regulatory regions. Liu, Sadlon [134] demonstrated that 55.7% of the gnomAD variants were not included in the largest meta-analysis of T1D genome-wide genotyped datasets to date [86] and therefore their associations with T1D susceptibility was not assessed. I identified a total of 92 and 50 candidate gnomAD SNPs located within accessible enhancers differentially regulated in T1D patients and involved in long-range enhancer-promoter interaction in Treg and Tconv cells, respectively, suggesting they can potentially alter T cell-specific gene expression in T1D through spatial interactions. Connecting the filtered candidate gnomAD SNPs to their long-range target genes using Hi-C datasets has revealed targets that were also captured in earlier analysis (integration of Hi-C with case-control ATAC-seq and RNA-seq). They include DPEP2 and CD79A for Treg cells, and COL6A3 and TLR5 for Tconv cells, reinforcing the implication of those gene targets in the pathogenesis of T1D. The filtered candidate gnomAD variants were also enriched in TF binding sites that have a key role in T cell development and differentiation, including YY1 and FOXP3 for Treg cells, and MAF for Tconv cells. This suggests that SNPs in those regulatory elements can alter, by either disrupting or creating new TF binding sites in T1D-associated enhancers, resulting in loss of long-range, enhancer-promoter interaction. In accordance with the present finding, Gao, Uzun [33] reported that the binding of YY1 was disrupted by a SNP in high linkage disequilibrium with the T1D-associated GWAS lead SNP (rs883868) at Treg enhancer, resulting in loss of long-range enhancer–promoter interaction, and our SNP prioritisation work has also been able to discover this independently. Polymorphisms in *Fopx3* lead to systemic autoimmune pathology in human and mice, including T1D [108, 250] and significant association between *FOXP3* and T1D was reported [380]. Kivling, Nilsson [374] reinforces findings from others [375] in this area demonstrating PBMCs from children with T1D exhibited a lower FOXP3 expression than children with celiac disease (CD), establishing a link between impaired regulation of FOXP3 and T1D. It is important to note that owing to gender bias in the data resulting from disproportionate representation of males and females in our study cohort, data points mapped to chromosome X (where FOXP3 resides) and Y were excluded from case-control differential analyses. Consistent with the pathogenic role of Treg and Tconv cells in T1D, overall, the candidate gnomAD variants were significantly enriched in genomic regions differentially regulated in T1D from T cells and they are involved in IL-2 signalling pathway, regulation of apoptosis in Treg and IFN signalling in Tconv cells. Functionally annotating by incorporating evidence of T1D-associated traits and regulatory elements enabled us to identify a subset of novel putative T1D variants potentially mediating the risk of developing T1D and revealing their potential biological phenotypes to aid in T1D intervention or treatment.

This work has strengthened the concept that T1D is a polygenic, complex disease that is influenced by genetic and epigenetic factors. Genetic factors have low penetrance [381] and are not sufficient to cause disease. Combination of many risk loci combines with effects of environmental factors to establish the T1D risk profile, resulting in loss of immune tolerance and assault of pancreatic β-cells. Each genetic variant in isolation may have a subtle effect on T1D risk, but each may alter a key function in the immune system and its interaction with pancreatic β-cells. How genetic variants mediate disease susceptibility and contribute to the loss of tolerance, and how the combination of genetic and environmental factors regulate the genes differently to contribute to T1D remain elusive. The knowledge generated from this PhD project carries important implications in the field of T1D. Using a small cohort of likely varying genetic risk profiles, not only has this work allowed insight into the molecular circuits by which the non-coding genetic variants may contribute to the loss of tolerance in T1D in the Tconv and Treg cells, but it has also revealed a list of novel biomarkers and gene candidates potentially for use in future immunomodulatory intervention and therapies in T1D, shedding light on the multiple, sophisticated pathways involved in T1D. T1D is a progressive disease characterised by gradual loss of β cells. The onset of increased blood glucose levels does not indicate loss of insulin reserve but rather inadequacy of β cell mass or insulin production [382]. The ability to diagnose or identify the risk of T1D during early disease development bears greater success as it allows intervention of disease progression before symptom onset or critical organ damage, preserving the survival and self-regenerative capacity of β cells. The novel immunologic markers identified in this study and their signalling pathways can be incorporated in combination with other interventions to enhance the induction of Treg cells or modulate the pro-inflammatory effector function.

6.2 Future Directions

Future research should validate the functional roles of the altered enhancers/gene targets and risk variants identified in this study using a combination of luciferase reporter assays (enhancer activity), CRISPR genome editing (modification of enhancers in primary cells), 3C-qPCR (enhancer-promoter interaction) and ChIP-qPCR (TF binding *in vivo*). Identification of genotypes at candidate SNP positions using ATAC-seq profiles for individual patients will allow correlation between the genotype and gene expression in T1D, although larger sample size might be required. These experimental models will be beneficial in providing more evidence for the clinical implications of the gene candidates.

As described this case-control study is constrained by a number of limitations such as modest sample size, which limits the statistical power of the study, resource scarcity which prevents the delineation of cell type diversity in T cell population and compromises the complexity of sequencing libraries, as well as strong confounding variables arising from covariate effects such as biobanking batch, gender, HLA genotypes and high variability between study subjects. Although data normalisation has been performed to reduce the effects of dominating confounding variables in the data, associated biological effect of interest might be removed along with the unwanted variation during the process. These factors may constrain the present study in capturing all true disease signatures, especially those with small effect sizes. However, it is important to consider that the absence of statistically significant difference is not a reflection of the absence of clinically significant difference.

Future research will therefore greatly benefit from a larger, longitudinal case-control cohort study which also allows the differentiation of disease drivers and the consequences of the T1D pathogenesis. Age-, gender- and HLA- matched study design would limit the confounding effects and improve the statistical strength for capturing true disease signatures. The case cohort involved in this study was derived from a single time point, i.e. established T1D, and therefore it lacks the capability to infer disease causality. The epigenetic alterations or risk factors identified in this study can be tracked over disease progression in longitudinal studies, preferably from pregnancy into early childhood using the same molecular technologies and bioinformatics approaches.

Future case-control investigations should also address the identity and source of cell populations as it has been reported that disease signatures were detected at the site of infection rather than circulating peripheral blood in T1D subjects. In addition, this study can be replicated with a comprehensive characterisation of segregated T cell populations to enrich for cell type-specific contribution, using low input genomic assay such as single cell sequencing as their pathogenic role in T1D has been reported [33, 291, 383].

6.3 Summary of Major Findings

This PhD project aimed to identify the epigenetic and transcriptomic alterations in Tconv and Treg cells from T1D cohort and healthy controls. Based on the results from this study, while preliminary, it is clear that in T1D, biological pathways involved in T cell cytokine balance, T cell development and differentiation are disrupted. An integrative, combinatorial approach that leveraged multiple functional annotated features from ATAC-seq, RNA-seq and Hi-C identified 42 and 21 dysregulated enhancer-associated targets in Treg and Tconv cells of T1D cohort, respectively, contributing to the repertoire of gene networks predisposing to T1D. Differential regulation was observed at non-coding enhancers regulating genes important for immune tolerance and effector differentiation such as *TIGIT*, *FOSL* and *MAF*. Previously known T1D-associated associations show significant enrichment with the altered enhancers identified from the present case-control study.

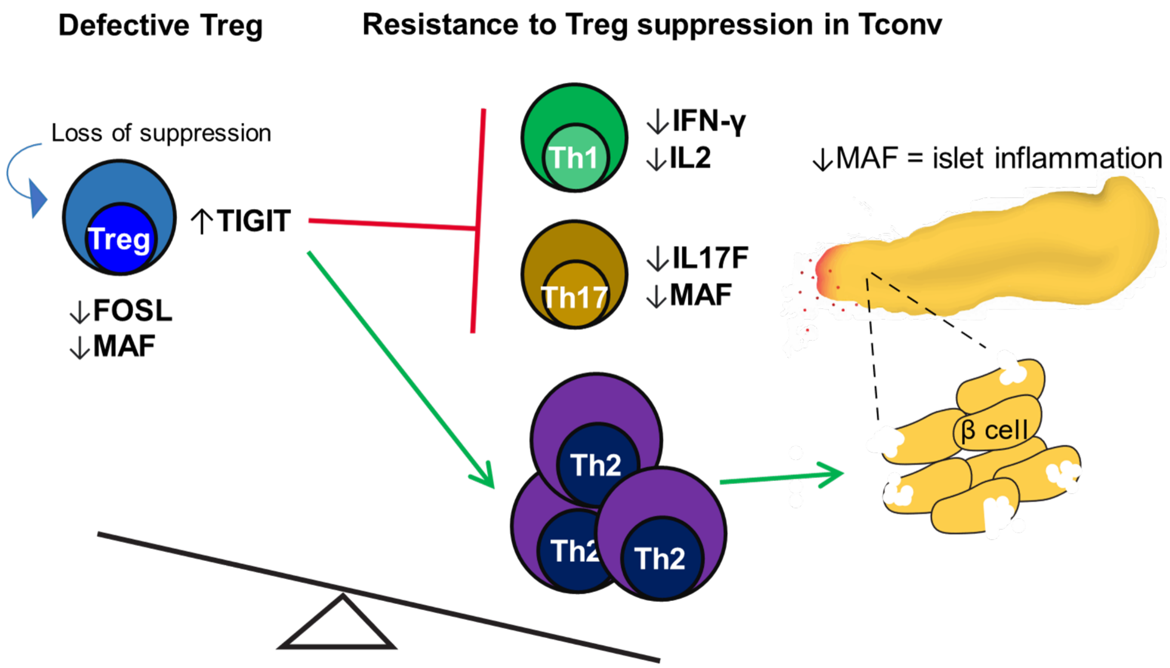


Figure 6-1 | Proposed mechanisms underlying T1D in CD4$^+$ Tconv and Treg cells. In T1D the balance of tolerance and effector reactivity is disrupted, resulting from the functional defects in Treg and resistance to immune suppression in Tconv. The work of this thesis speculates that perturbed immune homeostasis in human T1D is driven by MAF deficiency in both Treg and Tconv compartments, and failure to restrain Th2 cells, leading to autoimmune destruction of pancreatic islets in T1D.

This work proposes that MAF deficiency in both Treg and Tconv cells of T1D patients contributes to the underlying pathogenesis of T1D. MAF deficiency may render Treg cells defective in suppressing Tconv cells (loss of tolerance), and/or Tconv cells resistant to suppression by Treg cells (pro-inflammatory), contributing to the autoimmune destruction of pancreatic β cells in T1D (Figure 6-1). This hypothesis supports the findings in mouse from Singh, Colberg [275] which showed that loss of Maf impairs both T- and pancreatic β- cell function, promoting islet inflammation and augmenting the risk of autoimmune islet cell destruction. Loss of tolerance in T1D may further be compounded by upregulation of TIGIT in Treg cells, which could potentially direct T helper cell differentiation towards Th2 lineages or shift the cytokine profile towards a proinflammatory Th2 response, as it selectively suppresses

Th1- and Th17- specific responses in T1D Tconv cells. This corroborates existing evidence that suggested a pathogenic role of Th2-skewed immune response in T1D [347, 348].

This work also captured novel putative T1D-associated human genetic variation independently of GWAS which may potentially account for missing heritability in this disease. The findings from this PhD project add to a growing repertoire of evidence that dysregulation of epigenetics is implicated in T1D, impacting the expression of genes that are critical to the maintenance of immune tolerance and effector function in T cells. This resource can therefore assist efforts to understand the molecular mechanism of action by which the risk factors may influence disease susceptibility and contribute to the loss of tolerance in T1D. Functional validation studies as mentioned above will provide stronger evidence for the clinical implications of the gene candidates in T1D and a larger, longitudinal case-control cohort will allow inference of disease causality, enabling intervention to reverse autoimmune assault on pancreatic β cells.

In conclusion, the multi omics approach applied herein enabled discovery of enhancer-gene pairs that are altered in T1D, including novel targets of enhancers that are T1D perturbed. It also enabled connection of risk SNPs that are either known or novel to altered enhancer activity in T1D, and it identified regulatory motifs that were altered in T1D. Together these allow molecular diagnostic insight, and reveal disease linked transcriptional network effects that conspire to facilitate loss of tolerance resulting in the β cell destruction driving disease.

CHAPTER 7: Appendix

7.1 ATAC-seq protocol optimization

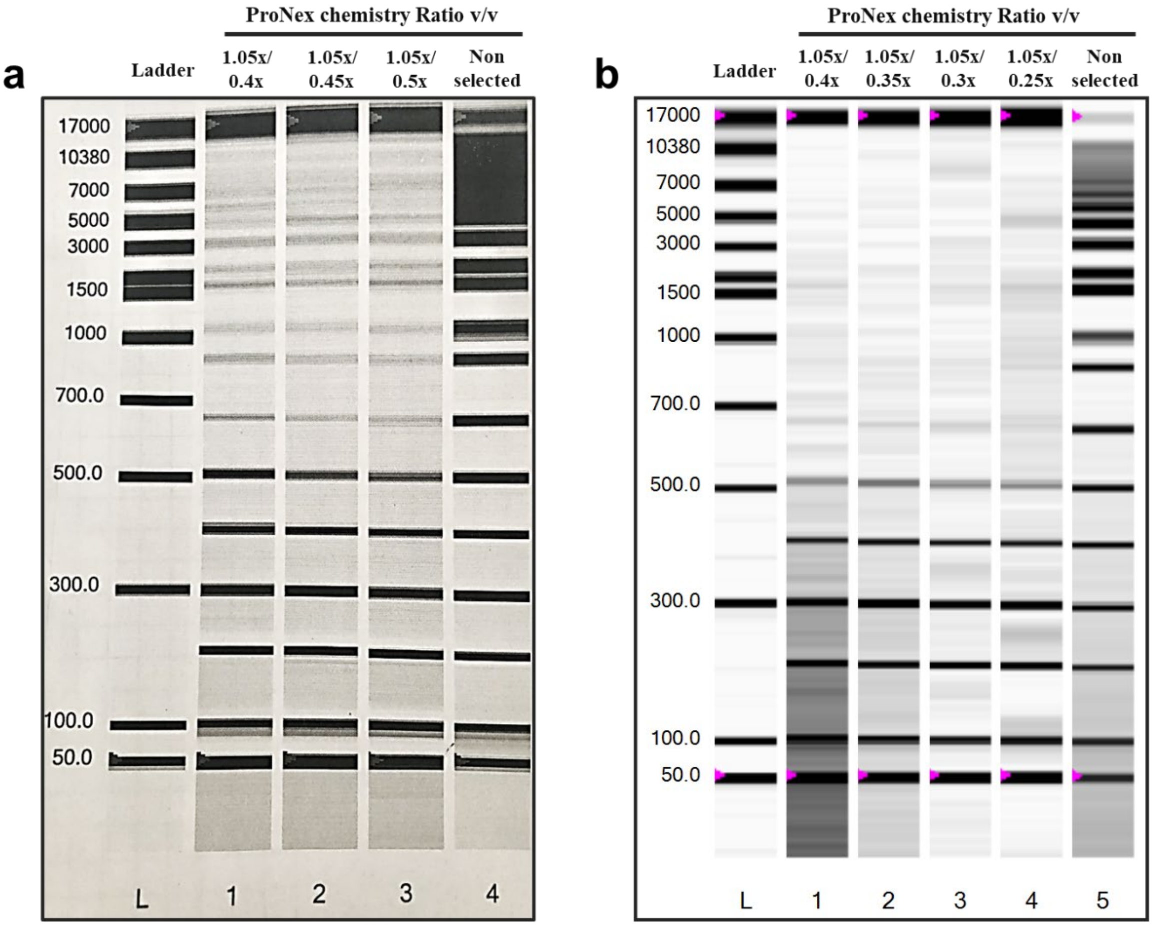


Figure 7-1 | Dual size selection of DNA ladder (1 kb Plus DNA Ladder; NEB) with varying Promega ProNex chemistry:sample ratios.

Dual size selection aims to produce a population of DNA fragments of a desired size range, removing fragments above and below a chosen size range based on the volume/volume (v/v) ratio of size selection reagent to the DNA sample. The conditions of size selection for producing desired size cutoffs were optimised with a generic dsDNA source (DNA ladder in this case). The varying chemistry v/v ratios attempted here targets the second ProNex Chemistry Selection to remove the lower molecular weight DNA fragments (such as dsDNA adapters, ssDNA oligonucleotides and nucleotides in the sequencing libraries) below a desired size. Size-selected DNA Markers were electrophoresed on an Experion Automated Electrophoresis System. The last lane of the gel is labelled Non selected and contains input DNA without size selection.

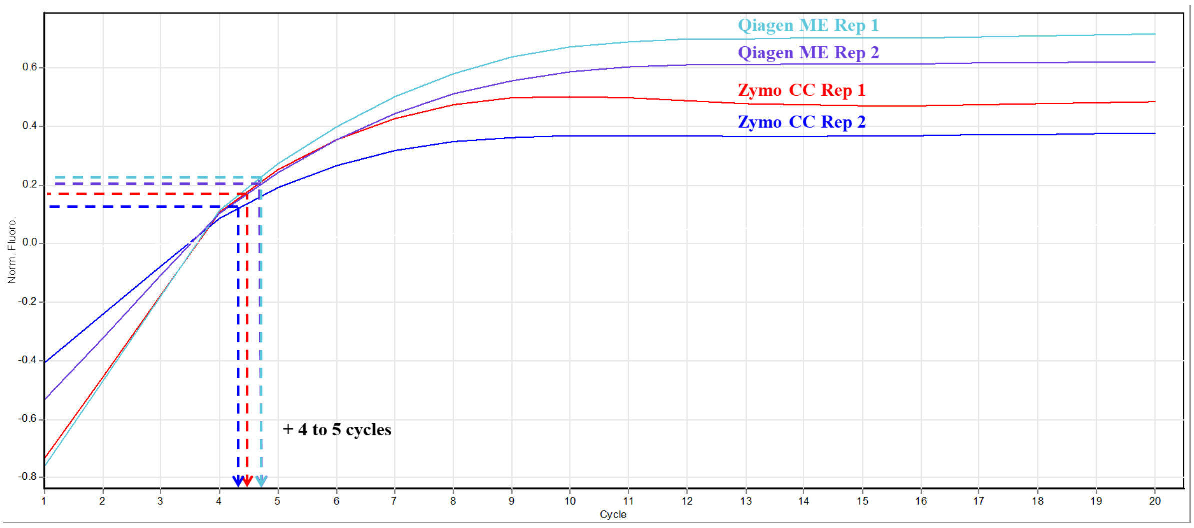


Figure 7-2 | Amplification plot demonstrating the number of additional PCR cycles to perform for four pre-amplified T cell ATAC-seq libraries purified using different commercial DNA purification kits - MinElute PCR Purification or Zymo Research DNA Clean & Concentrator-5 Kit.

Cleanup was performed on ATAC-seq tagmentation reactions followed by 5 cycles of PCR pre-amplification before assessment of complexity using qPCR amplification profiles to determine the required number of additional cycles.

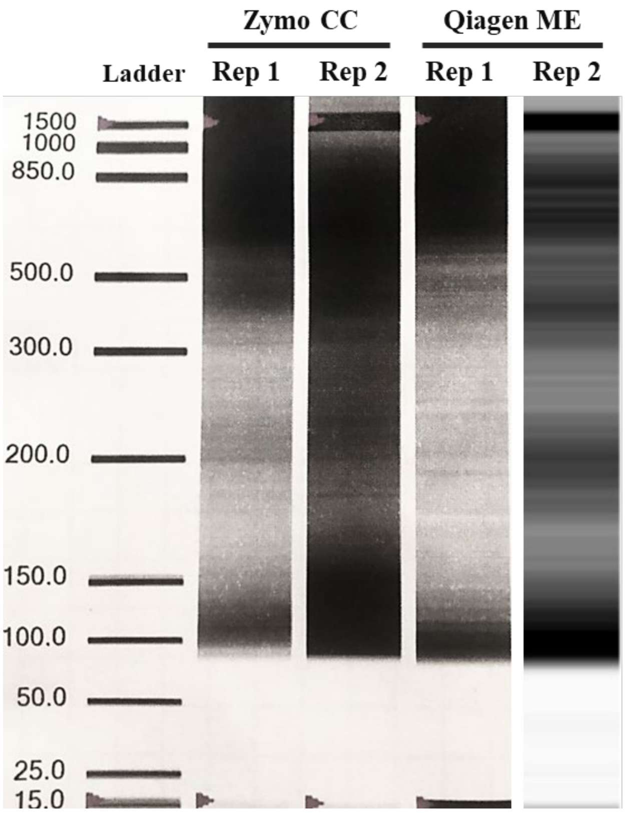


Figure 7-3 | Virtual gel demonstrating fragment sizes for amplified ATAC-seq libraries purified using different commercial DNA purification kits (matching amplification plot in Figure 7-2), determined by the Experion Automated Electrophoresis System.
All ATAC-seq libraries have undergone a total number of 11 PCR cycles and were purified using either the MinElute PCR Purification or Zymo Research DNA Clean & Concentrator-5 Kit as indicated. The ATAC-seq libraries shown this figure were not size-selected.

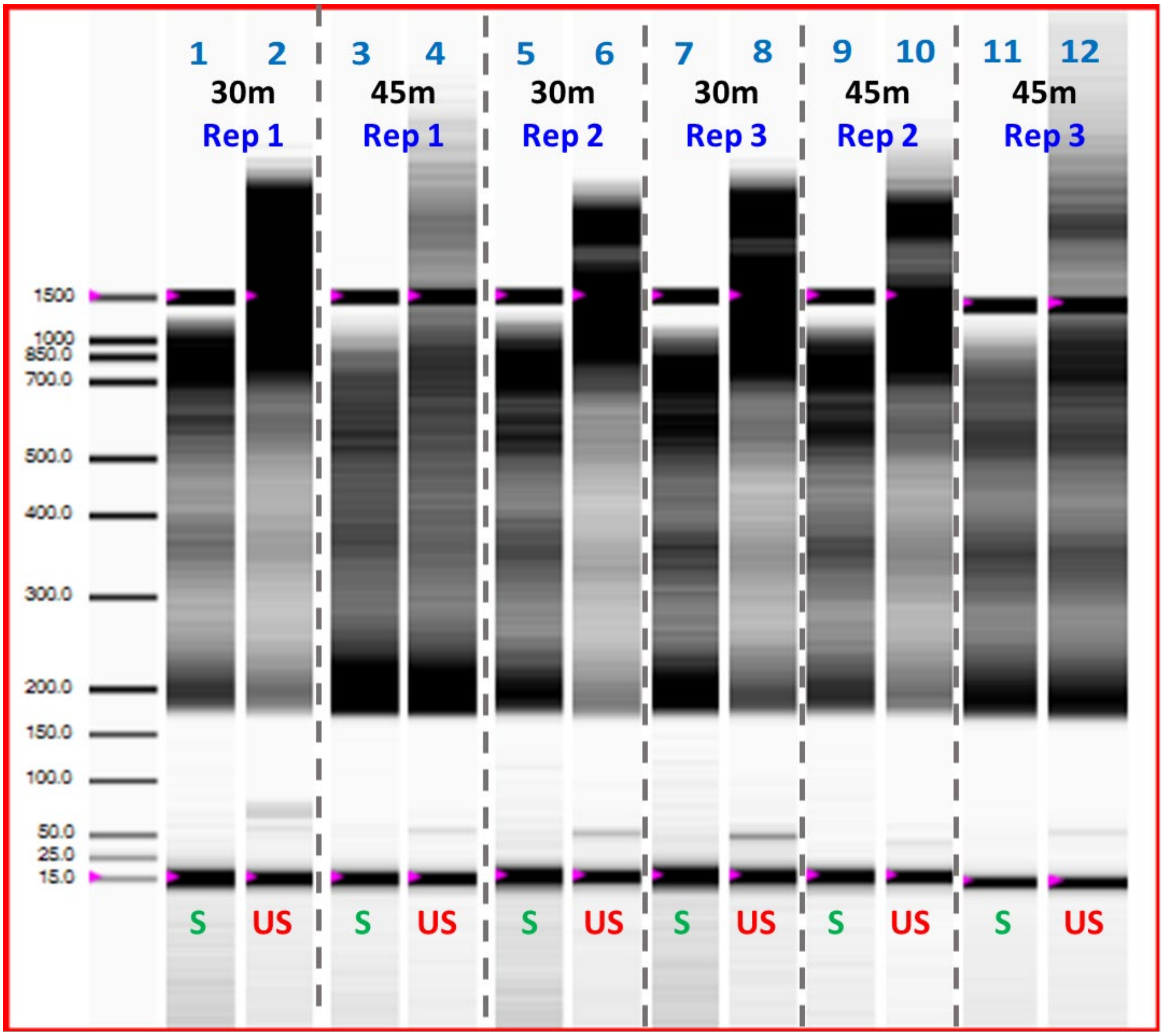


Figure 7-4 | Virtual gel demonstrating fragment size distribution of amplified Tconv libraries following varying tagmentation incubation time, before and after size selection, determined by the Experion Automated Electrophoresis System.

Three technical replicates were assayed for each tagmentation condition. S, size-selected; US, unselected.

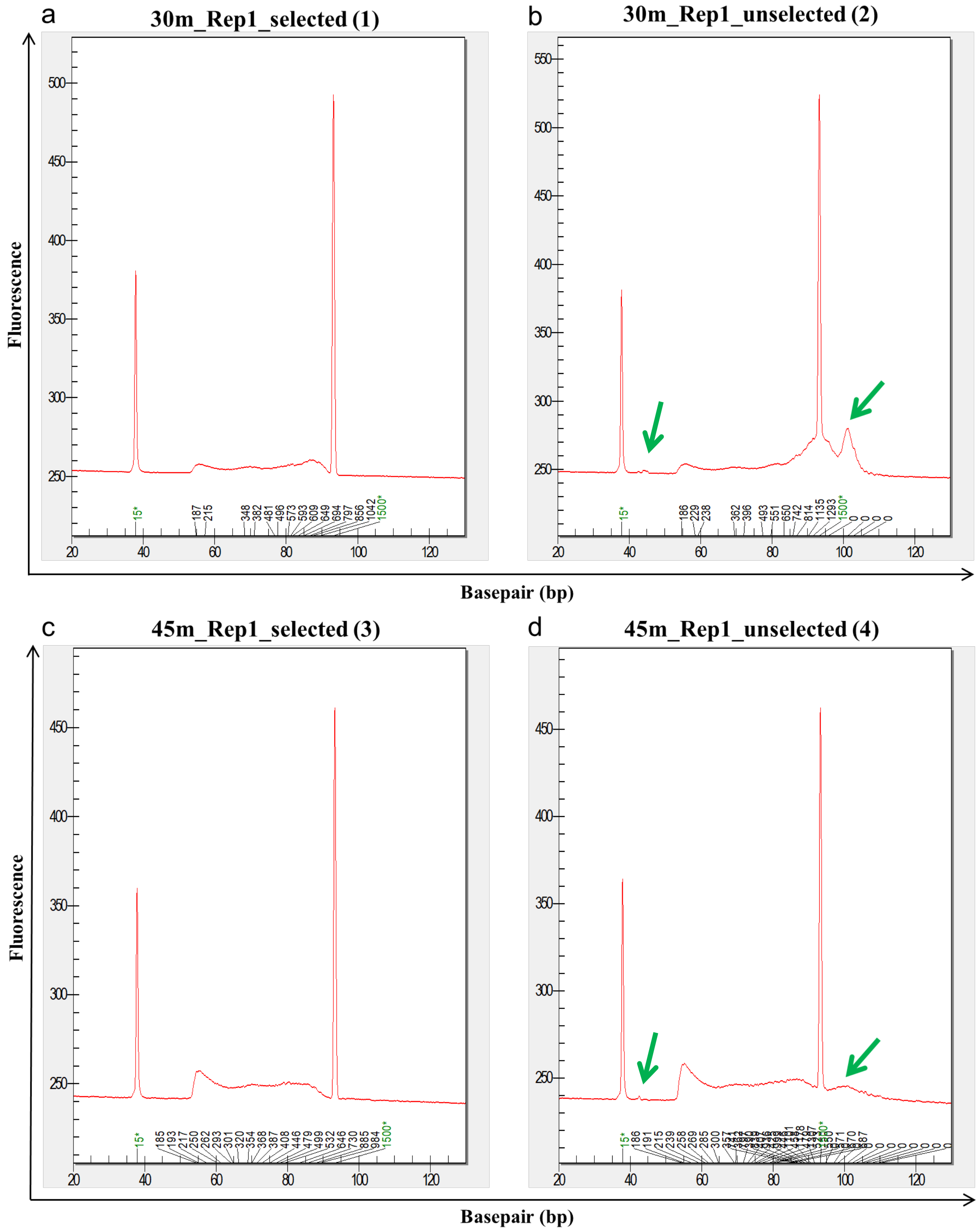


Figure 7-5 | Representative electropherograms (a-d; Rep 1 from Figure 7-4) showing fragment size distribution of the amplified T cell libraries from various tagmentation conditions, before and after size selection.
Green arrow indicates targeted region of size selection. Bracket indicates corresponding virtual gel lane in (Figure 7-4).

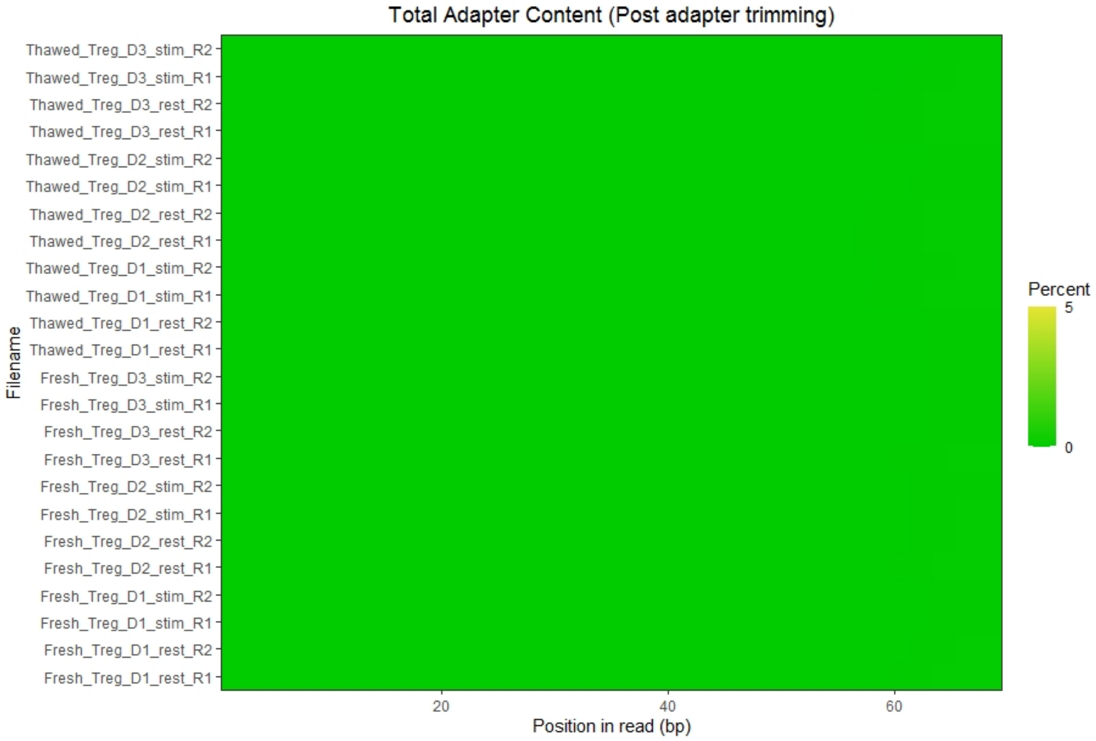


Figure 7-6 | Screening for adapter contamination in fresh and thawed Treg ATAC-seq libraries post trimming.

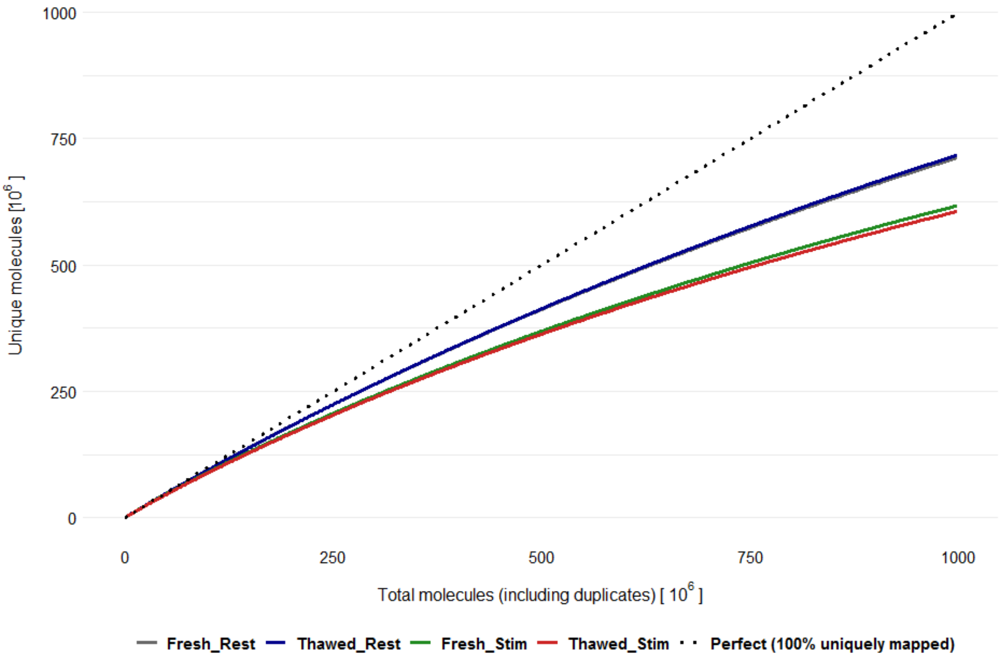


Figure 7-7 | Full library complexity and extrapolated yield curve of <u>pooled</u> fresh and thawed Treg ATAC-seq libraries under resting and stimulated state.
The data shown are representative of reads pooled from three healthy donors. The black diagonal line represents an ideal library, in which every read is a distinct read.

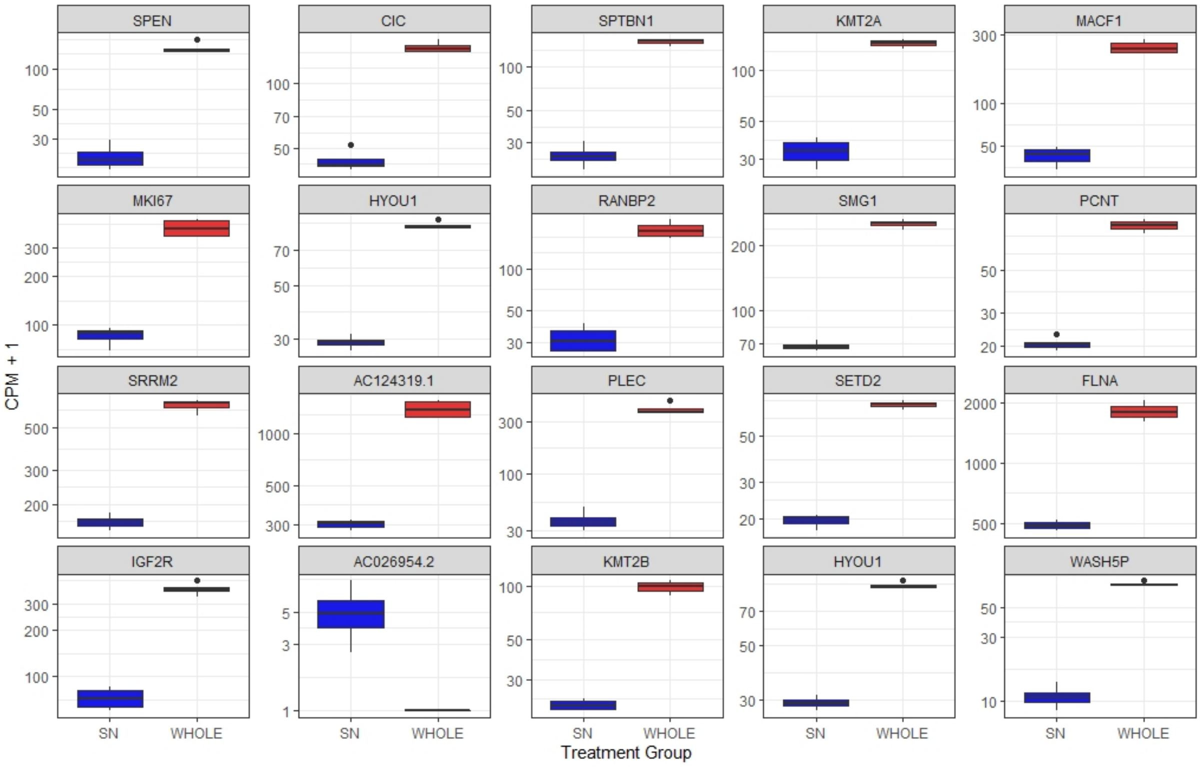


Figure 7-8 | Expression profile top 20 most differentially expressed genes between ATAC-SN and whole cell lysate purified from Treg cells.

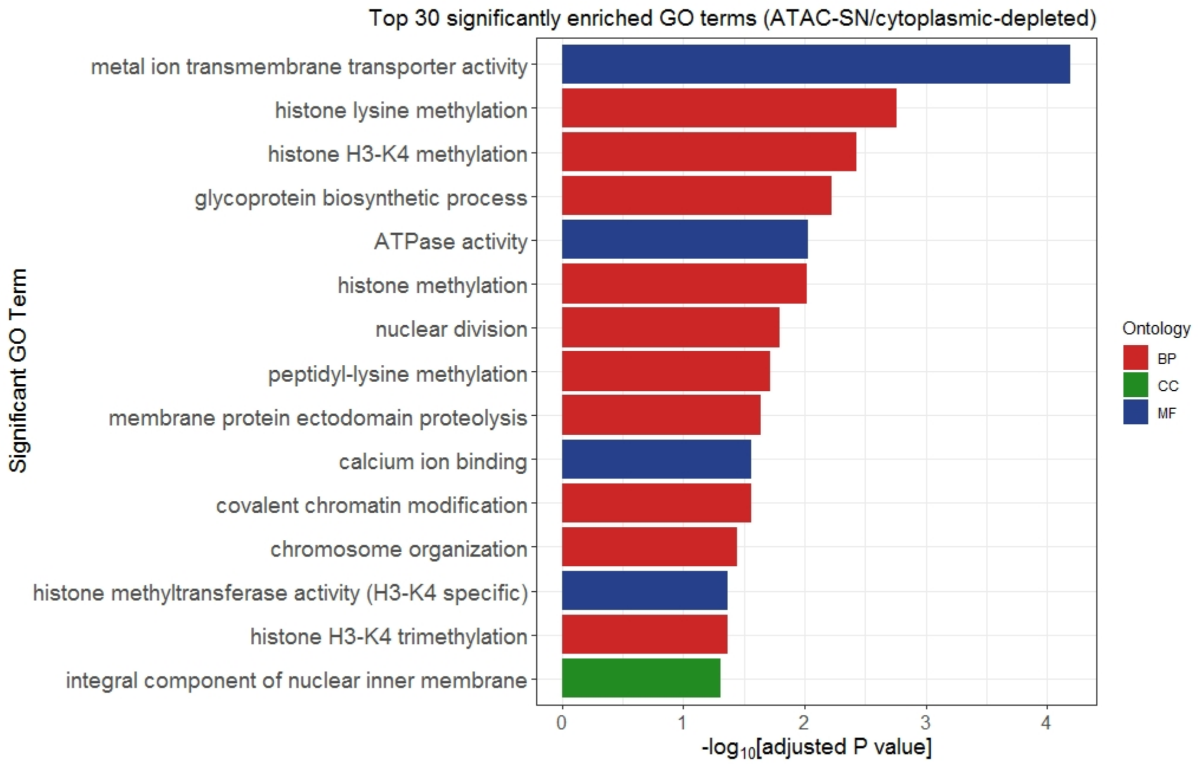


Figure 7-9 | Gene Ontology enrichment analysis showing significantly enriched GO terms associated with genes under-represented in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

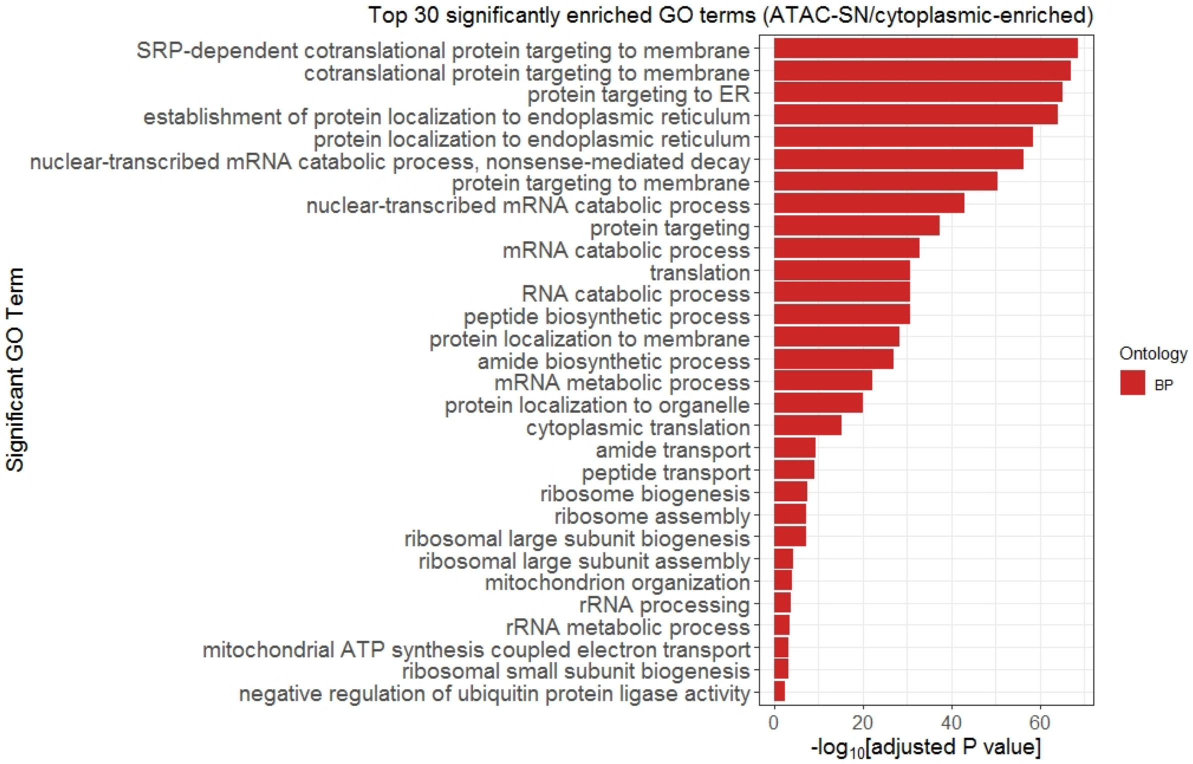


Figure 7-10 | Gene Ontology enrichment analysis showing significantly enriched GO terms associated with genes enriched/upregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

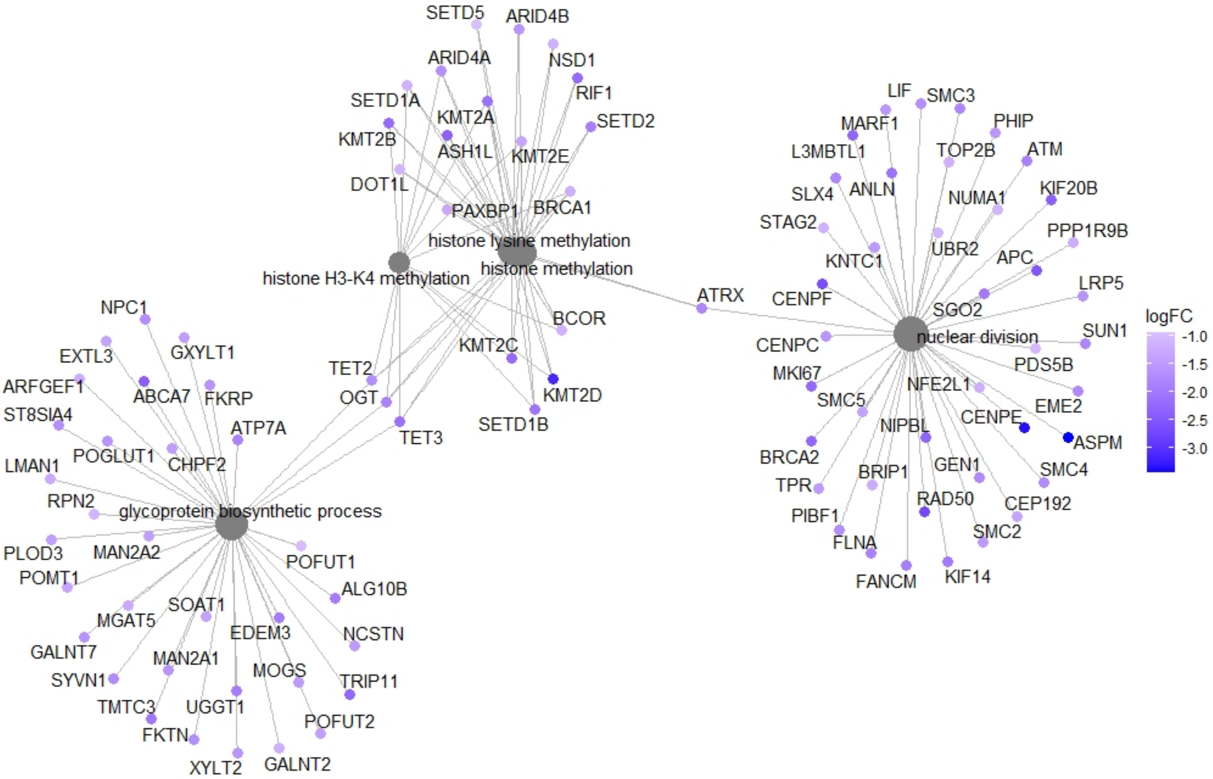


Figure 7-11 | Enrichment plot showing the most significant GO terms and related genes from the Biological Process Ontology for genes depleted/downregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

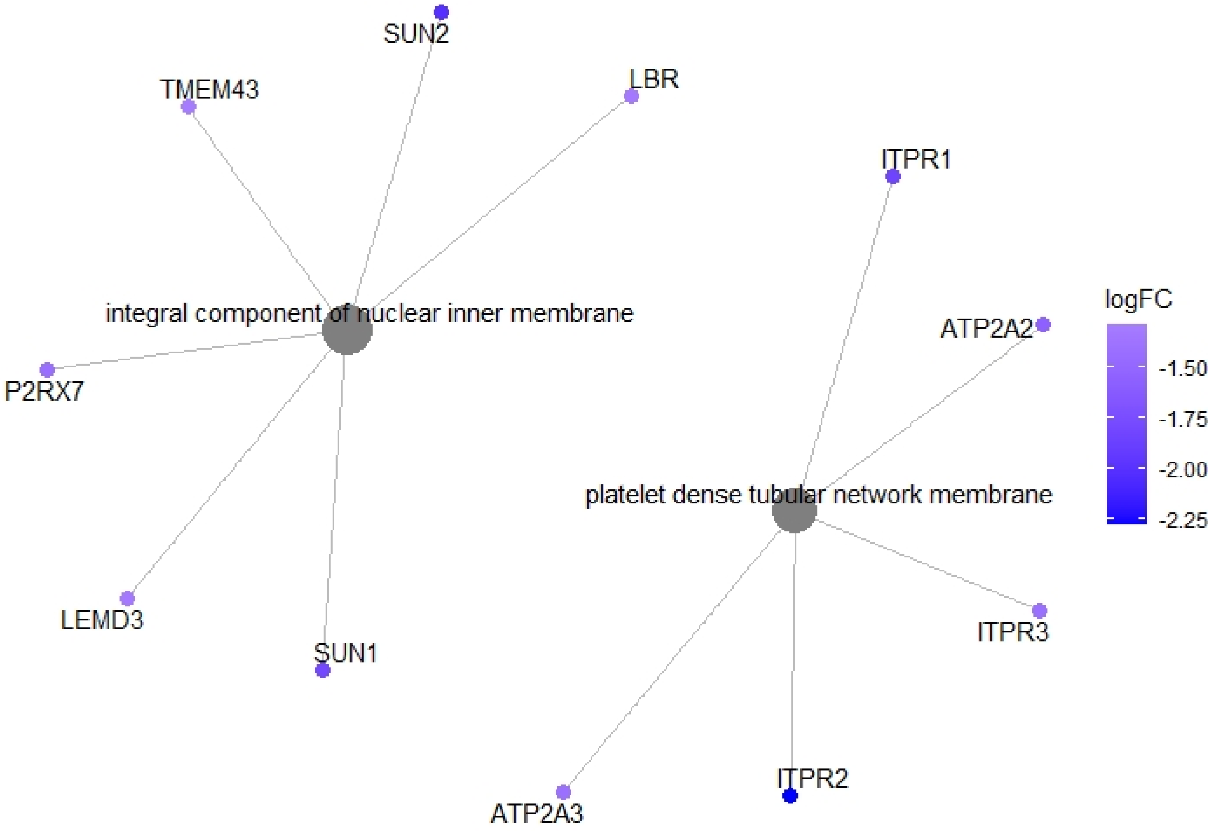

SUN2
TMEM43
LBR
ITPR1
integral component of nuclear inner membrane
ATP2A2
logFC
-1.50
-1.75
-2.00
-2.25
P2RX7
platelet dense tubular network membrane
LEMD3
SUN1
ITPR3
ITPR2
ATP2A3

Figure 7-12 | Enrichment plot showing the most significant GO terms and related genes from the Cellular Component Ontology for genes depleted/downregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

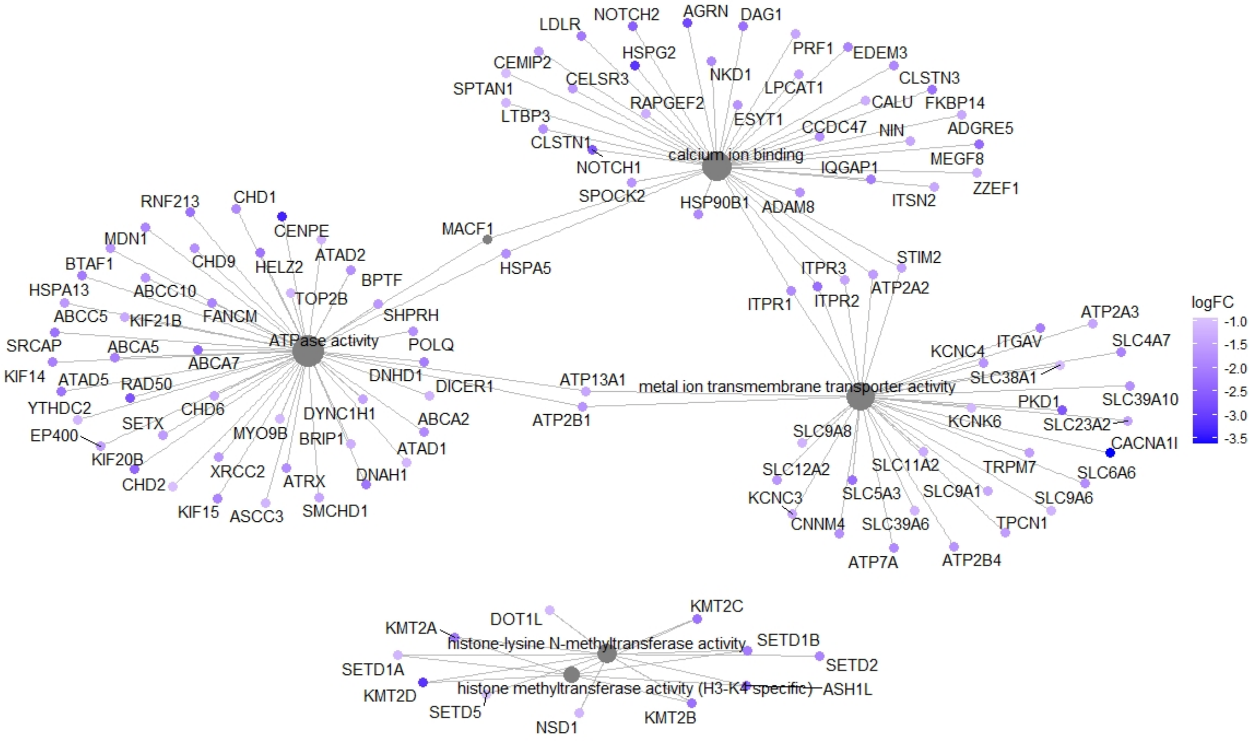


Figure 7-13 | Enrichment plot showing the most significant GO terms and related genes from the Molecular Function Ontology for genes depleted/downregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

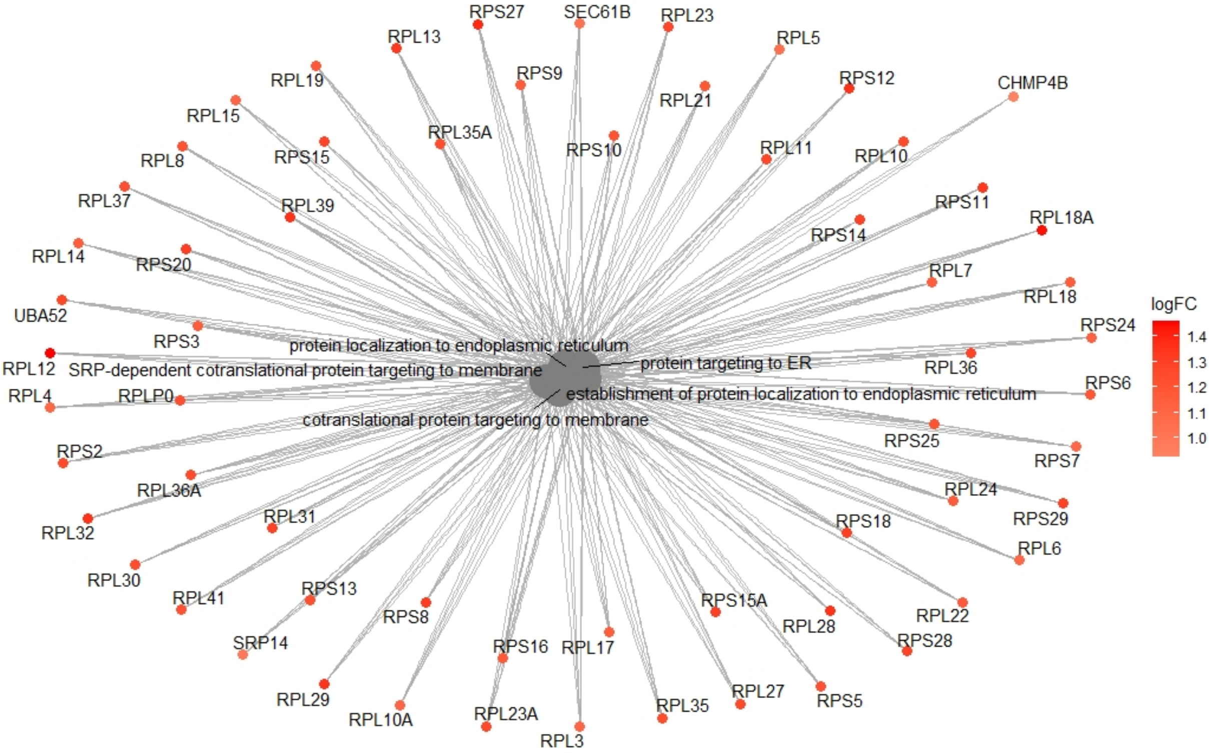


Figure 7-14 | Enrichment plot showing the most significant GO terms and related genes from the Biological Process Ontology for genes enriched/upregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

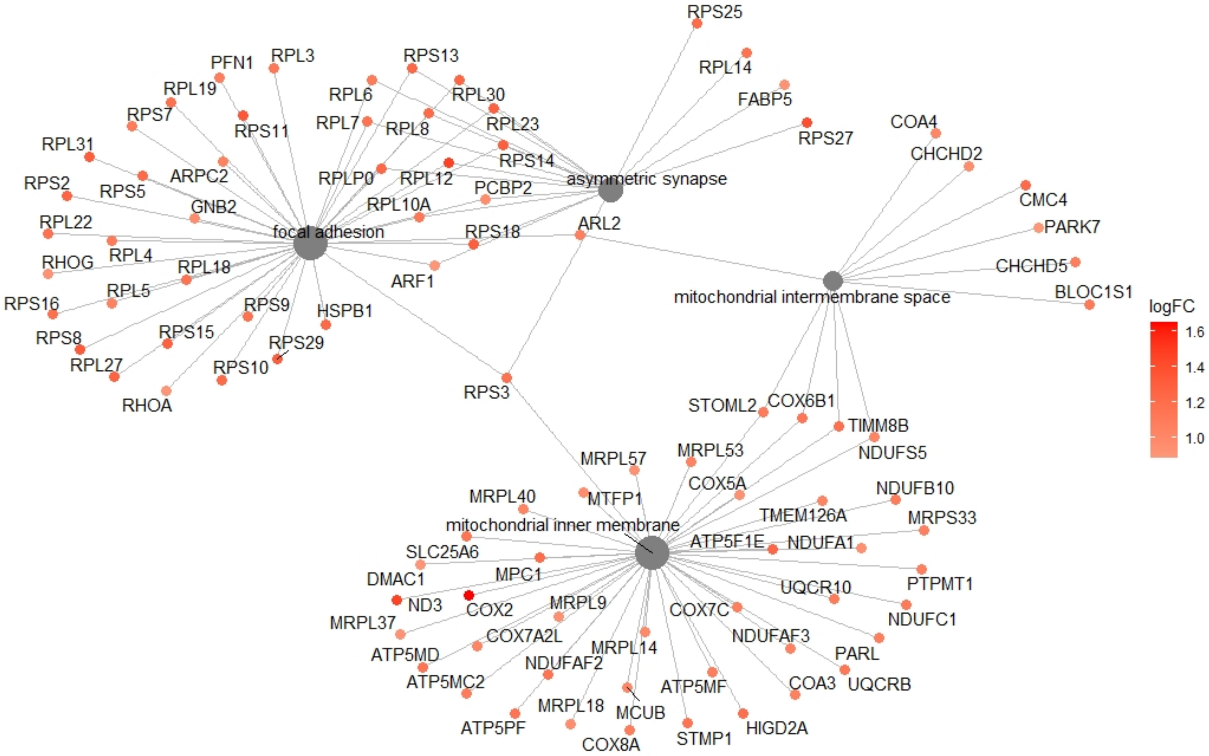


Figure 7-15 | Enrichment plot showing the most significant GO terms and related genes from the Cellular Component Ontology for genes enriched/upregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

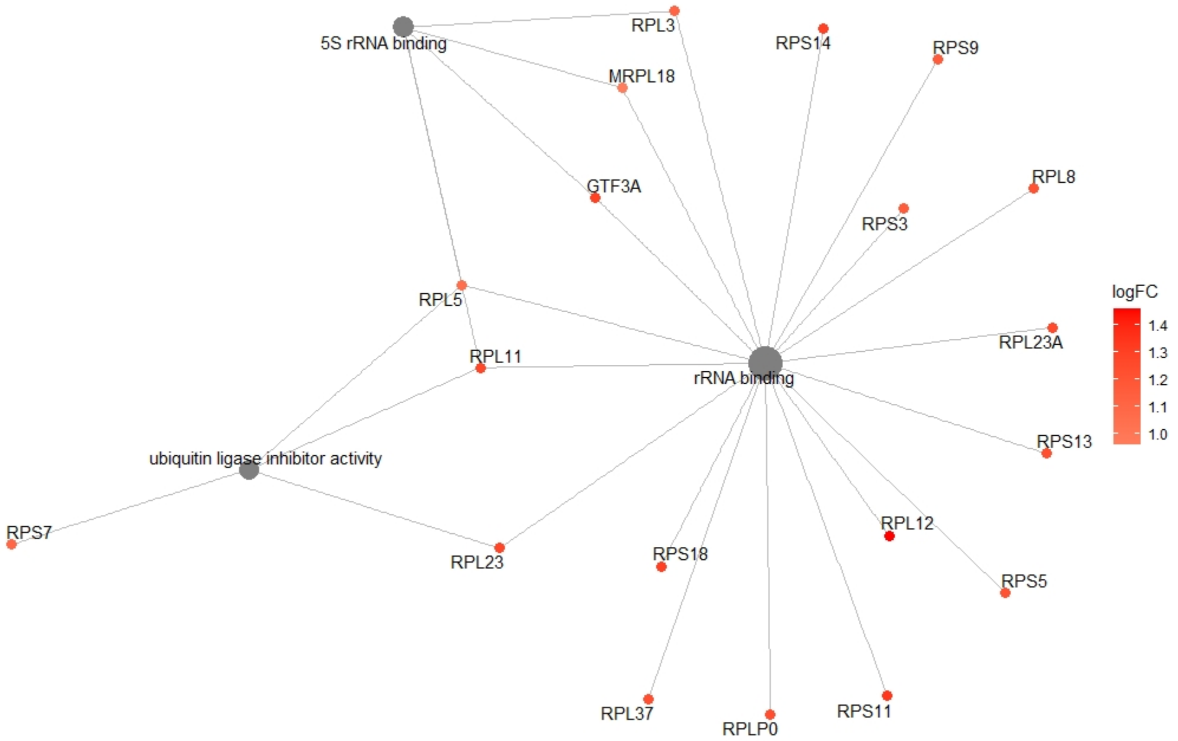


Figure 7-16 | Enrichment plot showing the most significant GO terms and related genes from the Molecular Function Ontology for genes enriched/upregulated in ATAC-SN/cytoplasmic lysate (relative to whole cell lysate), using a Bonferroni-adjusted p-value < 0.05 as the criteria for significance.

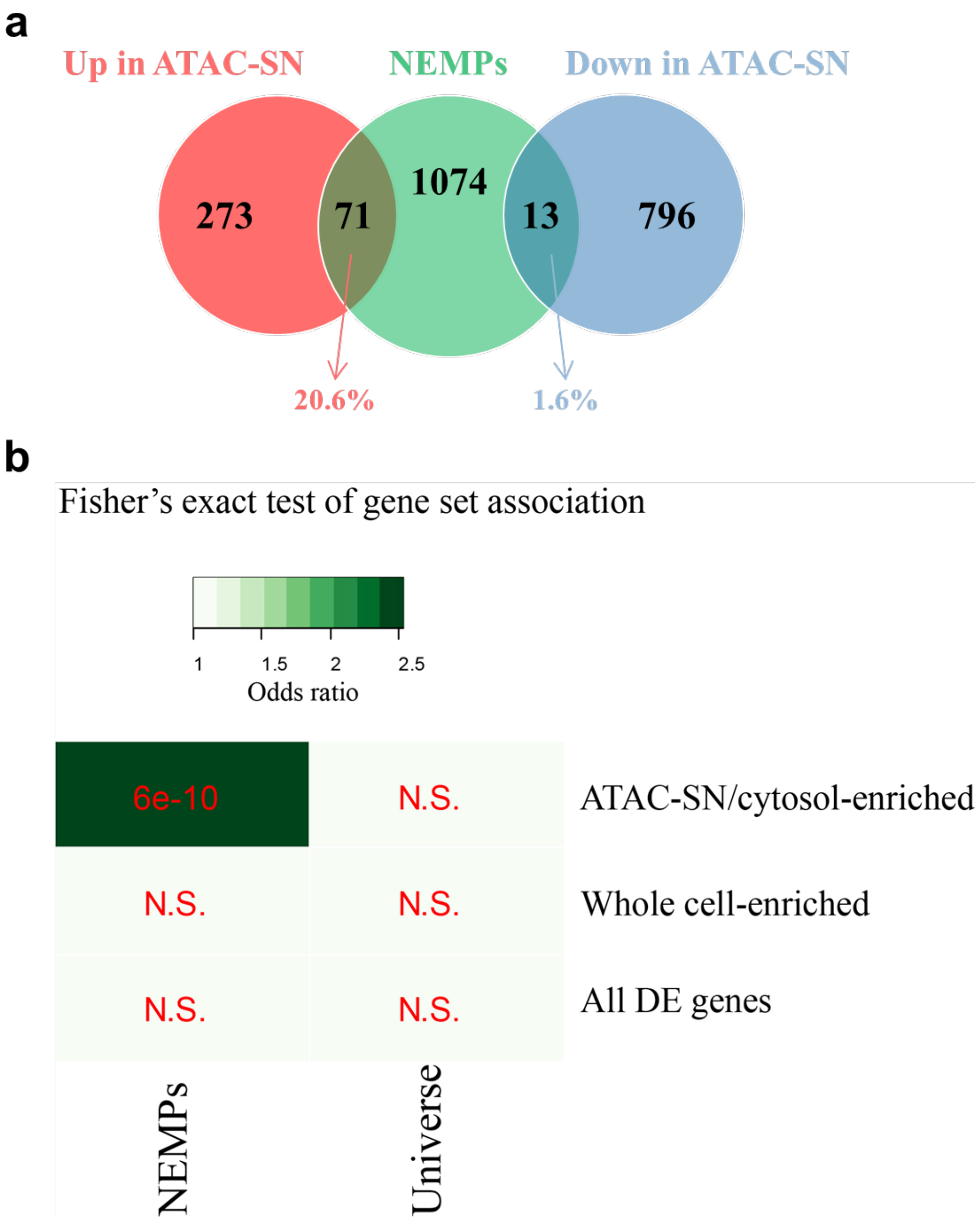


Figure 7-17 | Gene overlap (a) and Fisher's exact test (b) for assessing the strength of enrichment between genes up- or down-regulated in the ATAC-SN/cytosol fractions and human NEMPs (Nuclear-encoded-mitochondrial proteins)[5].
NEMPs database was obtained from the Broad Institute's human MitoCarta2.0[5]. These genes (n=1158) encode proteins with evidence for mitochondrial localization.

The heatmap color scale represents the odds ratios and the significant p-values, computed from Fisher's exact test, are superimposed on the grids.
All DE gene set represents all significantly differentially expressed genes between ATAC-SN and whole cell RNA-seq. Universe gene set contains all genes (DE and non-DE genes) detected in the RNA-seq dataset.

N.S., not significant; DE, differentially expressed.

7.2 Benchmarking and establishing bioinformatics data analysis pipelines

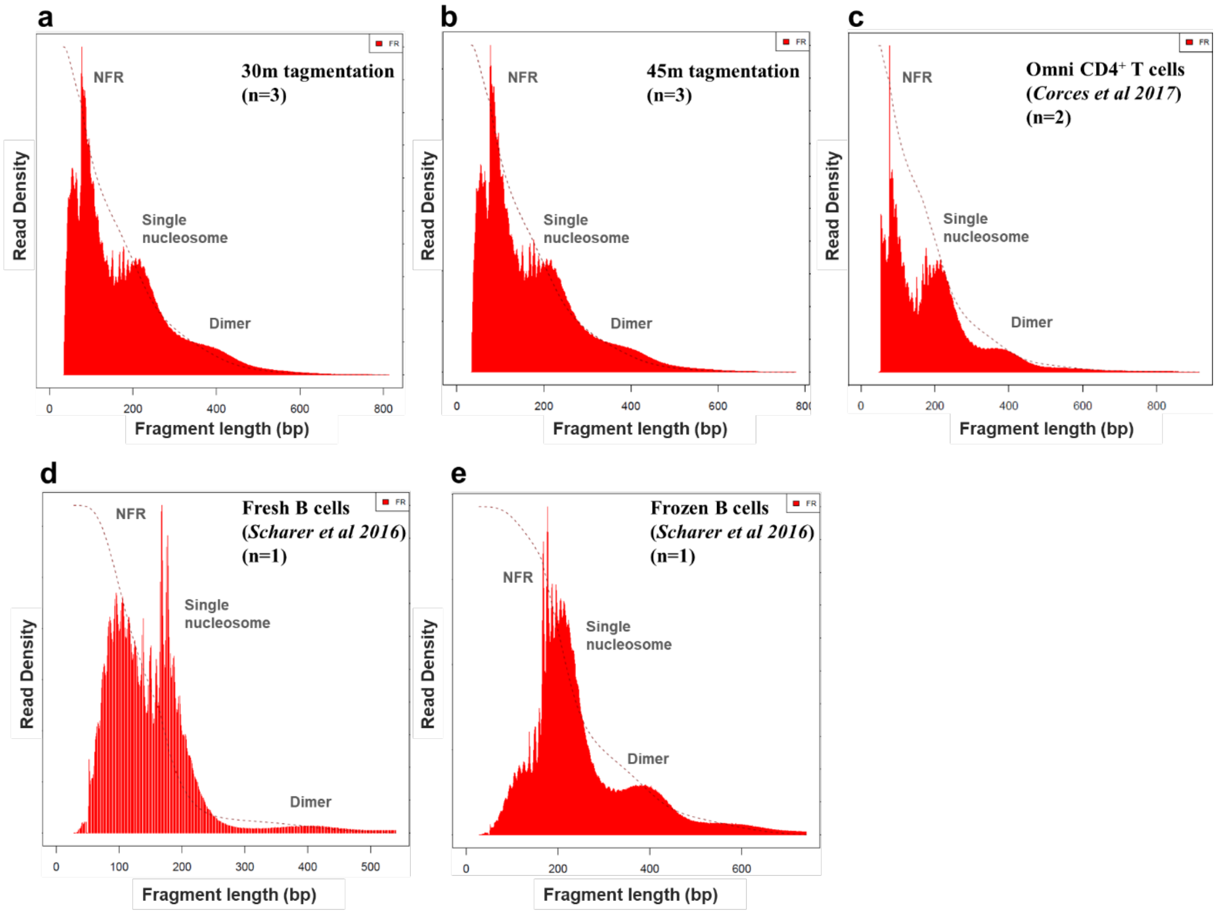


Figure 7-18 | The insert size distribution of my 30- (a) and 45-minute (b) tagmentation ATAC-seq libraries, alongside published Omni- (c) and Standard- (d, e) ATAC-seq datasets generated from primary T and B cells.

T cells for Omni ATAC-seq (c) were sorted based on the expression of CD3, CD4 and CD45, whereas naïve B cells used in Standard ATAC-seq (d, e) [30] were sorted as IgD$^+$ CD19$^+$ MTG$^-$ CD27$^-$ CD38$^+$ CD24$^+$ population. The data shown are representative of reads pooled from three technical replicates for (a) and (b), two replicates for (c) and single replicate for (d) and (e).

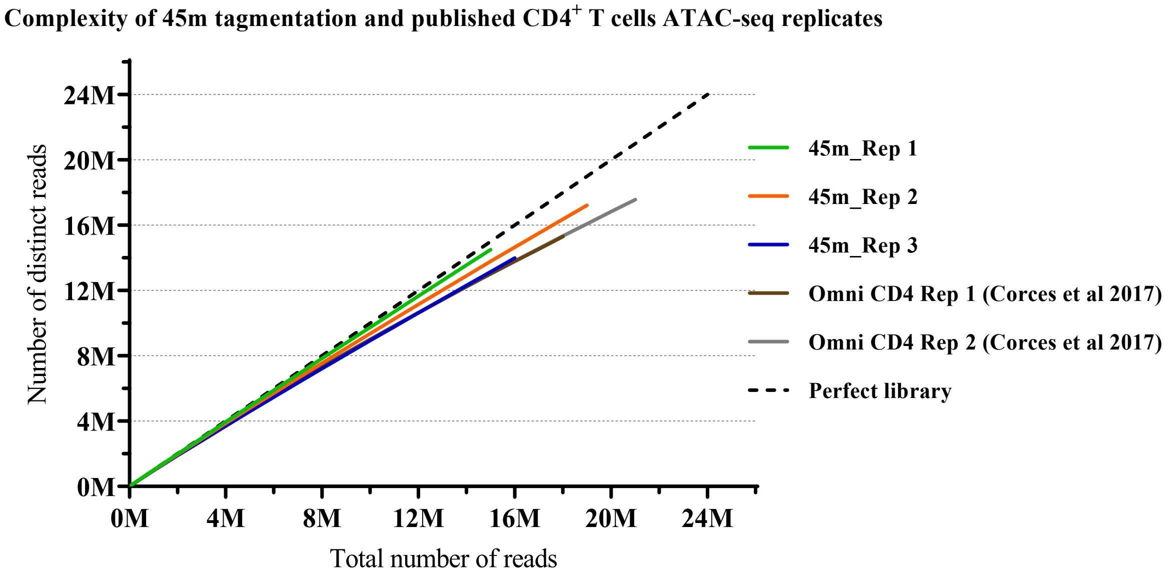


Figure 7-19 | Library complexity of in-house (30m, 45m tagmentation) and published Omni ATAC-seq libraries (same datasets as Figure 4-6 but only Omni ATAC-seq libraries are plotted for ease of visualisation).

The black diagonal line represents an ideal library, in which every read is a distinct read. The data shown are reads from single replicates.

7.3 Identification of genetic and epigenetic changes that contribute to Type 1 diabetes (T1D)

Table 7-1. Biobank cohort information and generation of sequencing libraries.
This table summarises the cell type (Tconv/Treg) and sequencing library type (ATAC-seq/RNA-seq whole or SN) successfully generated for each donor sample. Cells shaded in blue indicate sequenced ATAC-seq and RNA-seq libraries.

C, control; T, T1D; M, male; F, female; ATAC-SN, ATAC-seq supernatant fraction.

Sample	Cohort - Gender	Resting ATAC-seq	Stim ATAC-seq	Resting ATAC-SN RNA	Stim ATAC-SN RNA	Resting Whole RNA	Stim Whole RNA
JP29 - Pair 1	T - M	Tconv	Tconv, Treg	Tconv	Treg	Tconv	Tconv
JP30 - Pair 1	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP8 - Pair 2	T - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP9 - Pair 2	C - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP32 - Pair 3	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP33 - Pair 3	T - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP23 - Pair 4	T - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP73 - Pair 4	C - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP74 - Pair 5	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP75 - Pair 5	T - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP25 - Pair 6	C - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP26 - Pair 6	T - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP50 - Pair 7	T - M	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv	Tconv
JP51 - Pair 7	C - M	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv	Tconv
JP61 - Pair 8	C -M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP62 - Pair 8	T - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP27 - Pair 9	T - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv, Treg
JP28 - Pair 9	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP37 - Pair 10	T - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv, Treg
JP38 - Pair 10	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv, Treg
JP53 - Pair 11	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP56 - Pair 11	T - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP53 - Pair 11	C - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP56 - Pair 11	T - M	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
JP67 - Pair 12	T - M	Tconv	Tconv	Tconv	Tconv	Tconv	Tconv
JP68 - Pair 12	C - F	Tconv	Tconv	Tconv	Tconv	Tconv	Tconv
JP35 - Pair 13	T - M	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv	Tconv, Treg
JP36 - Pair 13	C - F	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv, Treg	Tconv	Tconv, Treg
JP43 (odd pair)	T - F	Tconv	Tconv, Treg	Tconv	Tconv, Treg	Tconv	Tconv
Total ATAC-seq/RNA-seq libraries made		33	58	NA	57	NA	34
Overall total libraries made		182					
Sequenced							
NA = Cell lysate collected but libraries not made.							

Table 7-2. Input Treg cell number for generation of activated ATAC-seq libraries. FACS-sorted Treg cells were stimulated with beads conjugated with anti-CD3 and anti-CD28 antibodies for 48 hours, followed by cell enumeration by Trypan Blue dye exclusion assay for generation of ATAC-seq libraries.

Case-control paired samples are indicated by black outside border.

ATAC-seq library	Input Treg number
JP23-Treg_T1D	11,000
JP73-Treg_Cont	11,000
JP8-Treg_T1D	36,000
JP9-Treg_Cont	50,000
JP26-Treg_T1D	19,000
JP25-Treg_Cont	22,000
JP27-Treg_T1D	50,000
JP28-Treg_Cont	50,000
JP29-Treg_T1D	40,000
JP30-Treg_Cont	40,000
JP33-Treg_T1D	43,000
JP32-Treg_Cont	50,000
JP35-Treg_T1D	50,000
JP36-Treg_Cont	50,000
JP37-Treg_T1D	50,000
JP38-Treg_Cont	50,000
JP50-Treg_T1D	50,000
JP51-Treg_Cont	41,000
JP56-Treg_T1D	39,000
JP53-Treg_Cont	42,000
JP62-Treg_T1D	34,000
JP61-Treg_Cont	20,000
JP75-Treg_T1D	50,000
JP74-Treg_Cont	41,000
Mean (± SD)	**39,125 (± 12,742)**

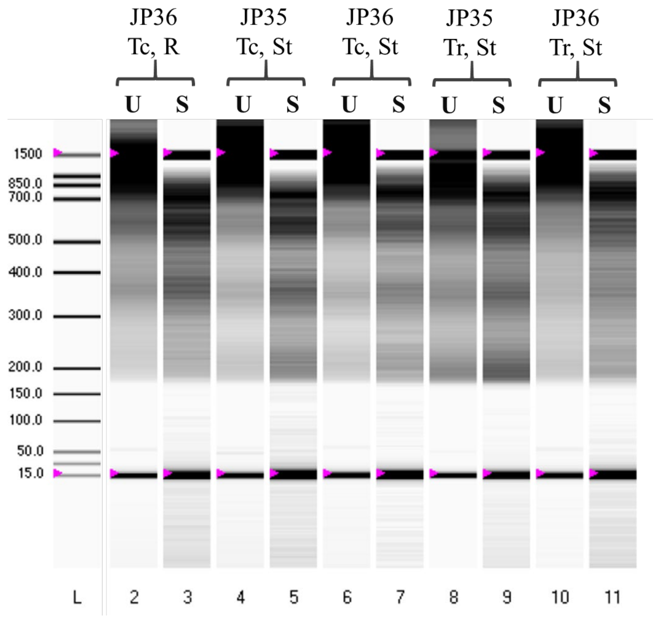


Figure 7-20 | Representative virtual gel demonstrating fragment size distribution of pre- and post-size selected T cell ATAC-seq libraries generated from biobanked T1D and healthy control samples.
The assessment of library QC was performed on an Experion Automated Electrophoresis System (DNA 1K Analysis).

Tr, Treg; Tc, Tconv; R, resting; St, stimulated; U, unselected; S, selected.

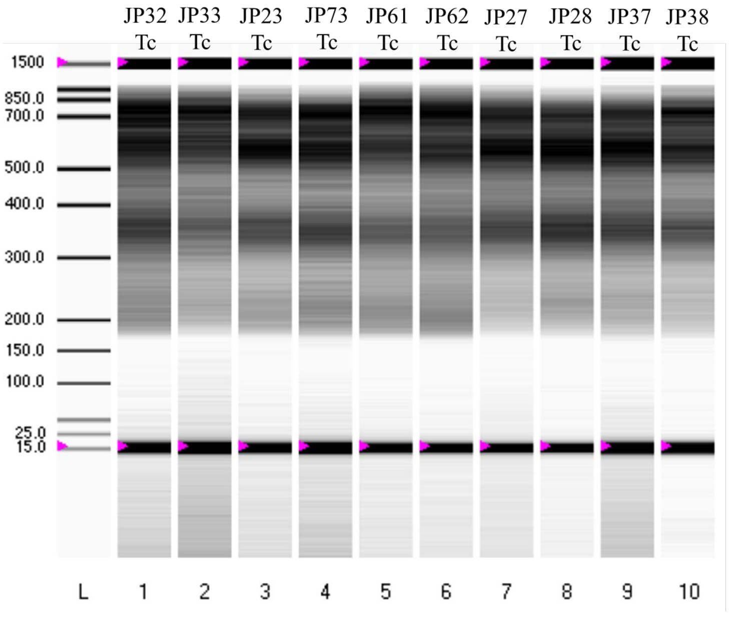

JP32 JP33 JP23 JP73 JP61 JP62 JP27 JP28 JP37 JP38
Tc Tc Tc Tc Tc Tc Tc Tc Tc Tc
1500
850.0
700.0
500.0
400.0
300.0
200.0
150.0
100.0
25.0
15.0
L 1 2 3 4 5 6 7 8 9 10

Figure 7-21 | Representative virtual gel demonstrating fragment size distribution of amplified T cell ATAC-seq libraries generated from biobanked T1D and healthy control samples.
The assessment of library QC was performed on an Experion Automated Electrophoresis System (DNA 1K Analysis).

Tc, Tconv.

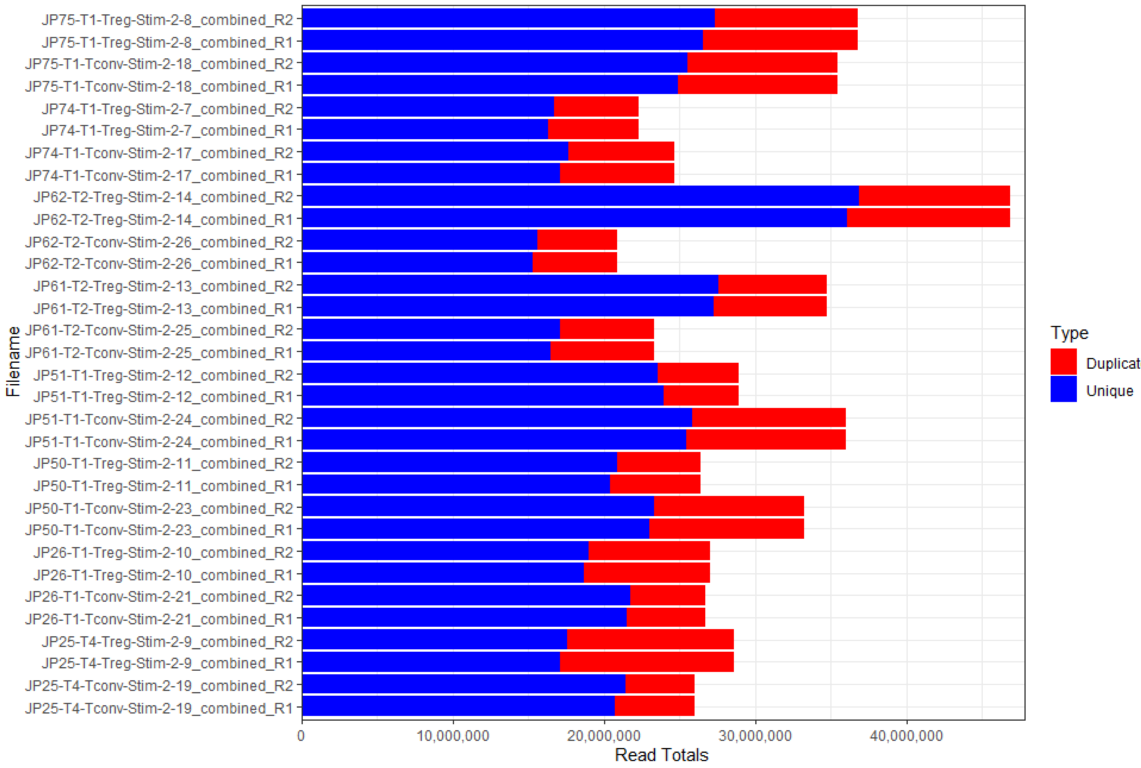


Figure 7-22 | The total number of paired-end reads for ATAC-seq libraries sequenced on Lane 003.

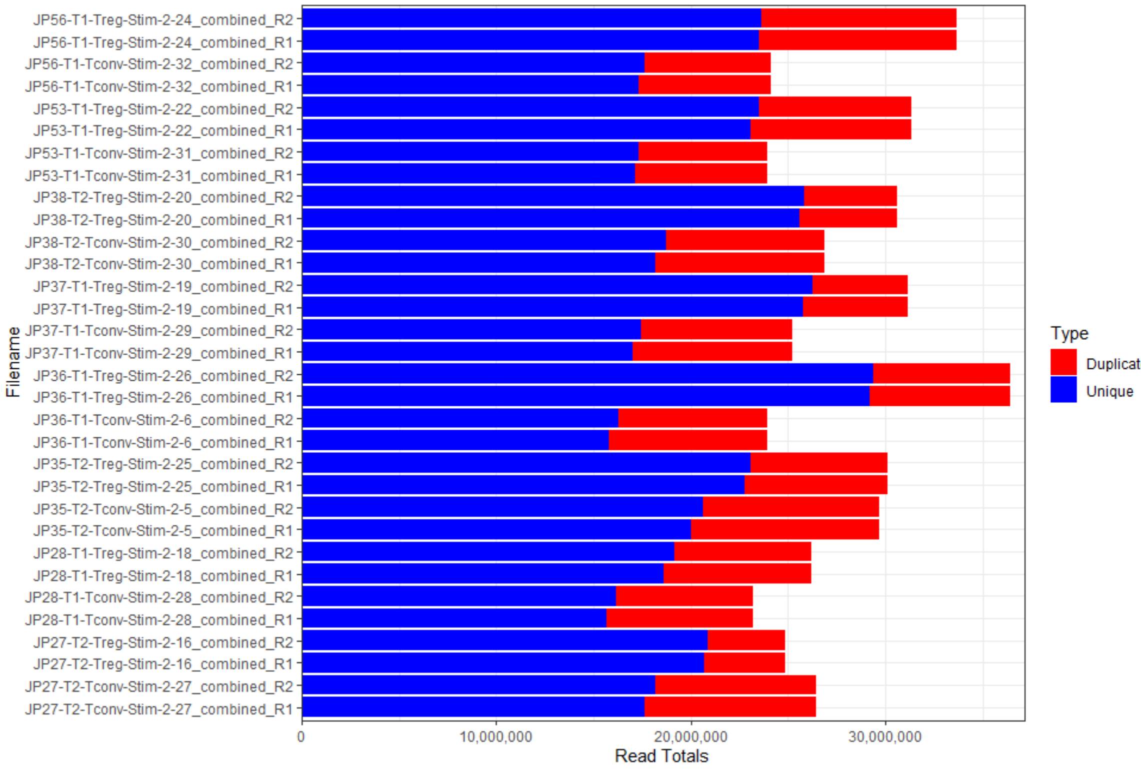


Figure 7-23 | The total number of paired-end reads for ATAC-seq libraries sequenced on Lane 004.

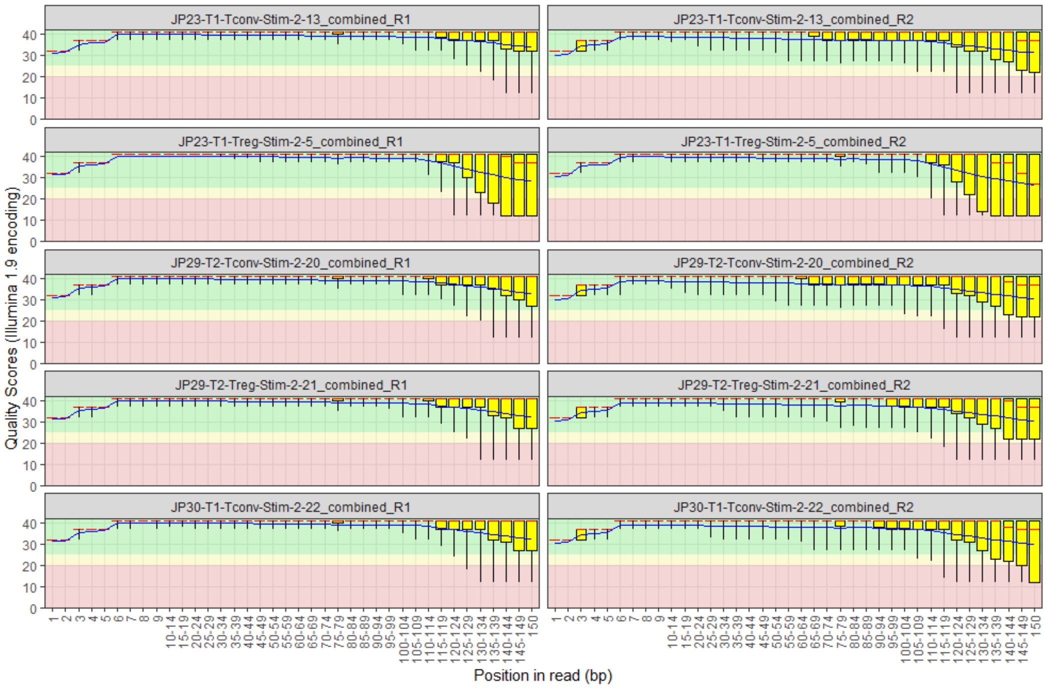


Figure 7-24 | Representative per base sequence quality plots for ATAC-seq libraries generated from biobanked T1D and healthy control samples from one sequencing lane, L002.

This QC metric provides the distribution of quality scores at each position in the read across allreads in both orientation of paired-end sequencing.

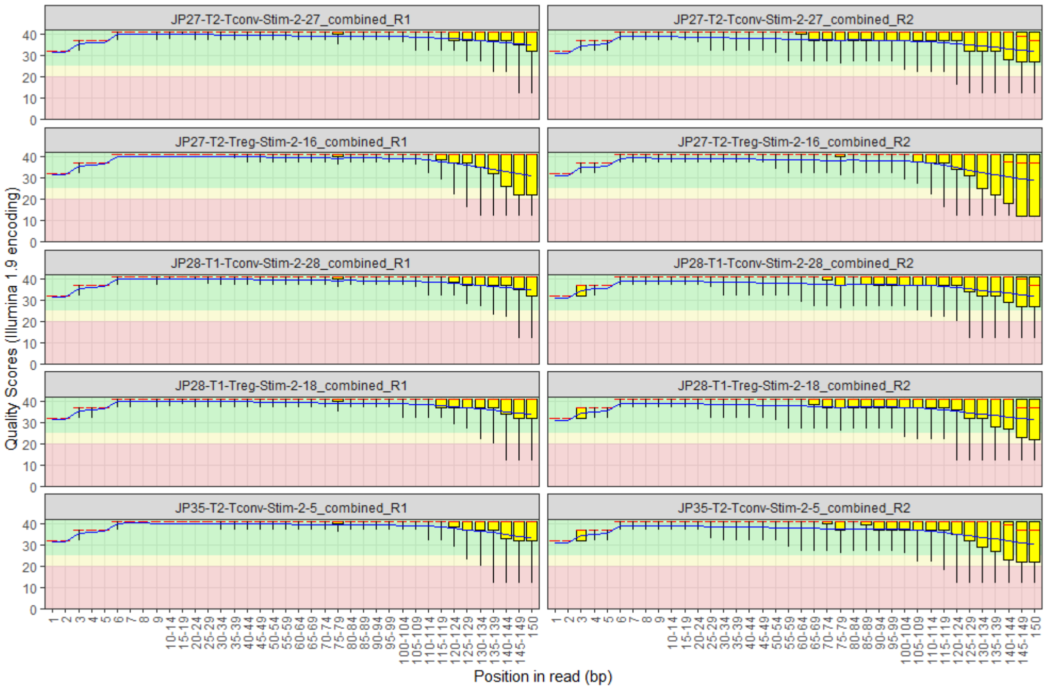


Figure 7-25 | Representative per base sequence quality plots for ATAC-seq libraries generated from biobanked T1D and healthy control samples from one sequencing lane, L003.

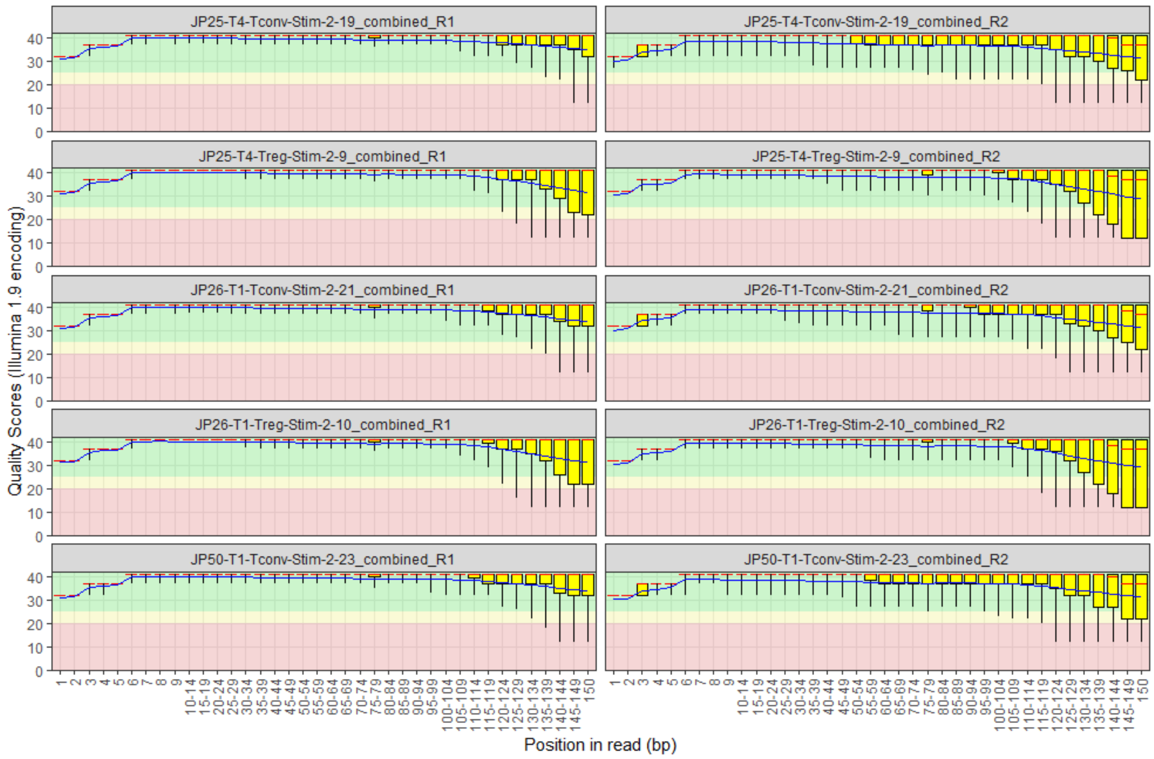


Figure 7-26 | Representative per base sequence quality plots for ATAC-seq libraries generated from biobanked T1D and healthy control samples from one sequencing lane, L004.

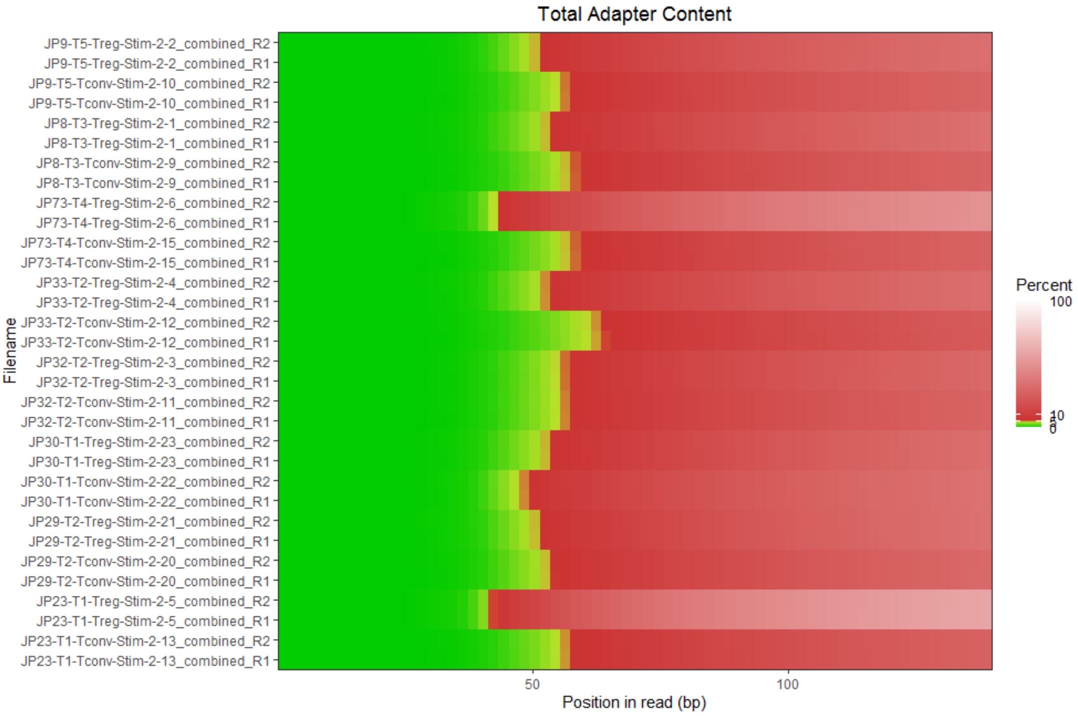


Figure 7-27 | Adapter content and position in raw sequencing reads for ATAC-seq libraries generated from biobanked T1D and healthy control samples from one sequencing lane, L002.

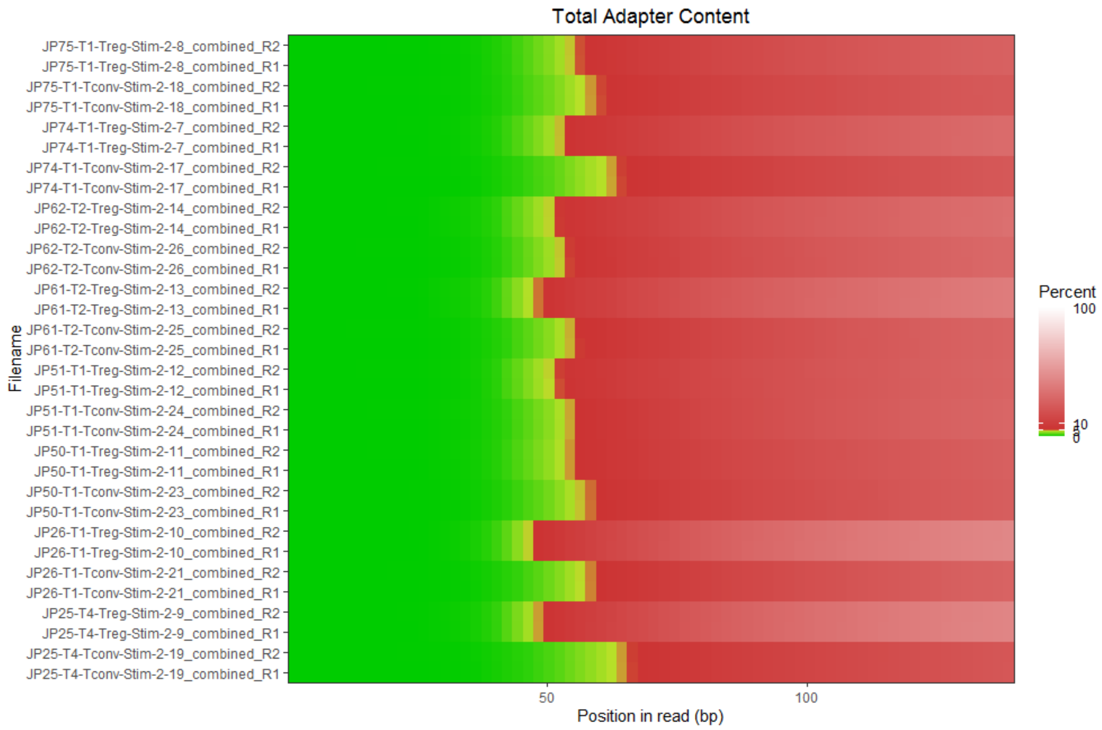


Figure 7-28 | Adapter content and position in raw sequencing reads for ATAC-seq libraries generated from biobanked T1D and healthy control samples from one sequencing lane, L003.

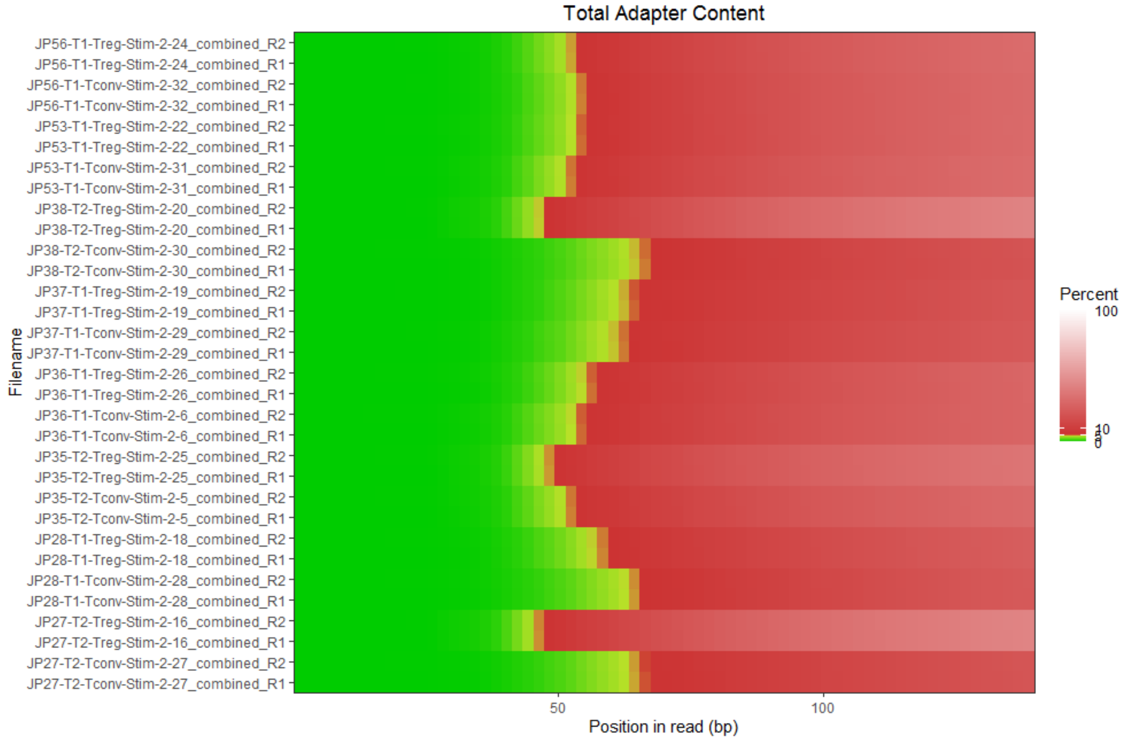


Figure 7-29 | Adapter content and position in raw sequencing reads for ATAC-seq libraries generated from biobanked T1D and healthy control samples from one sequencing lane, L004.

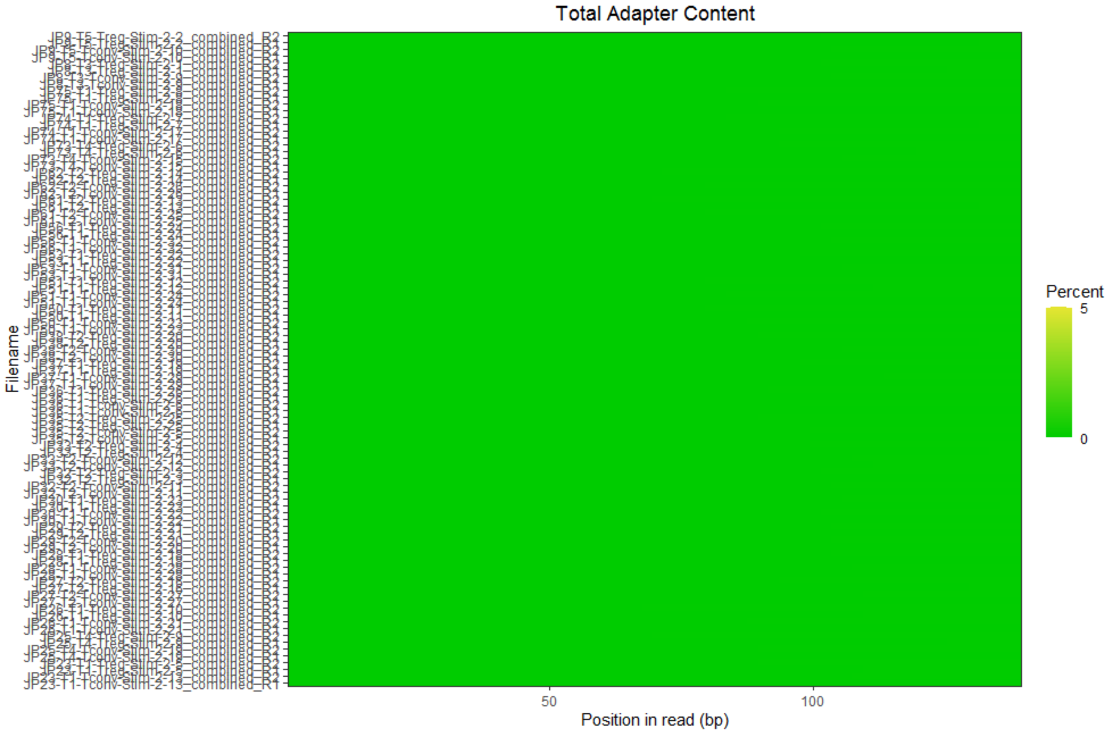


Figure 7-30 | Screening for adapter contamination in case-control ATAC seq libraries post trimming (samples from all three lanes).

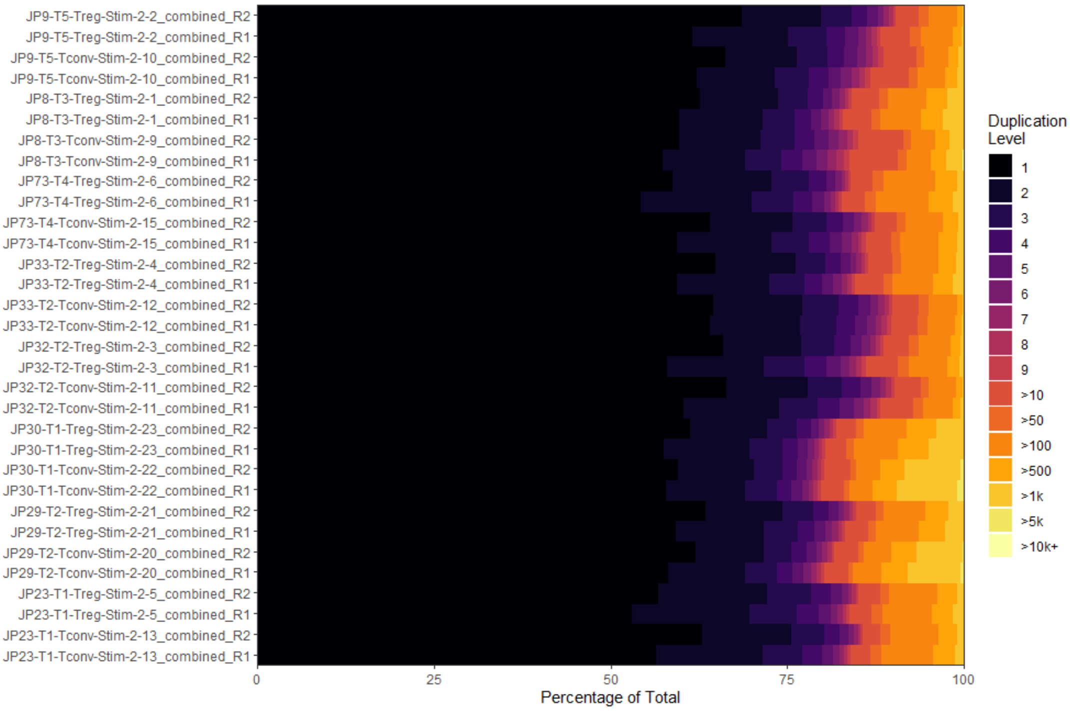


Figure 7-31 | Heatmap showing the duplication levels of case-control ATAC-seq libraries from one sequencing lane, L002.

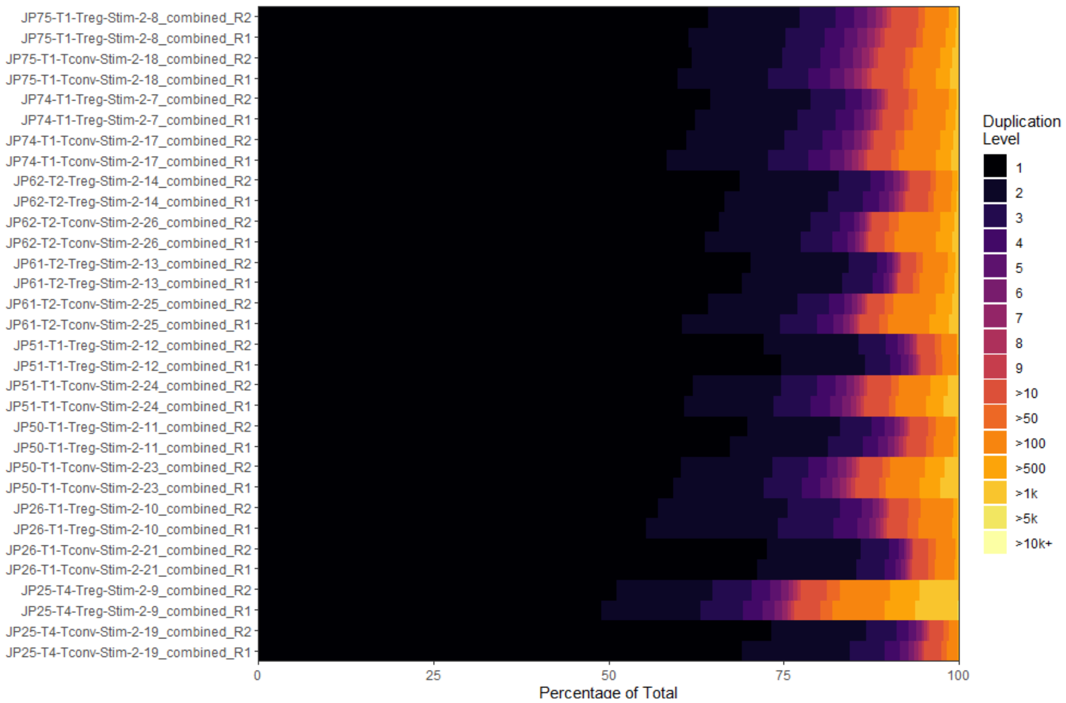


Figure 7-32 | Heatmap showing the duplication levels of case-control ATAC-seq libraries from one sequencing lane, L003.

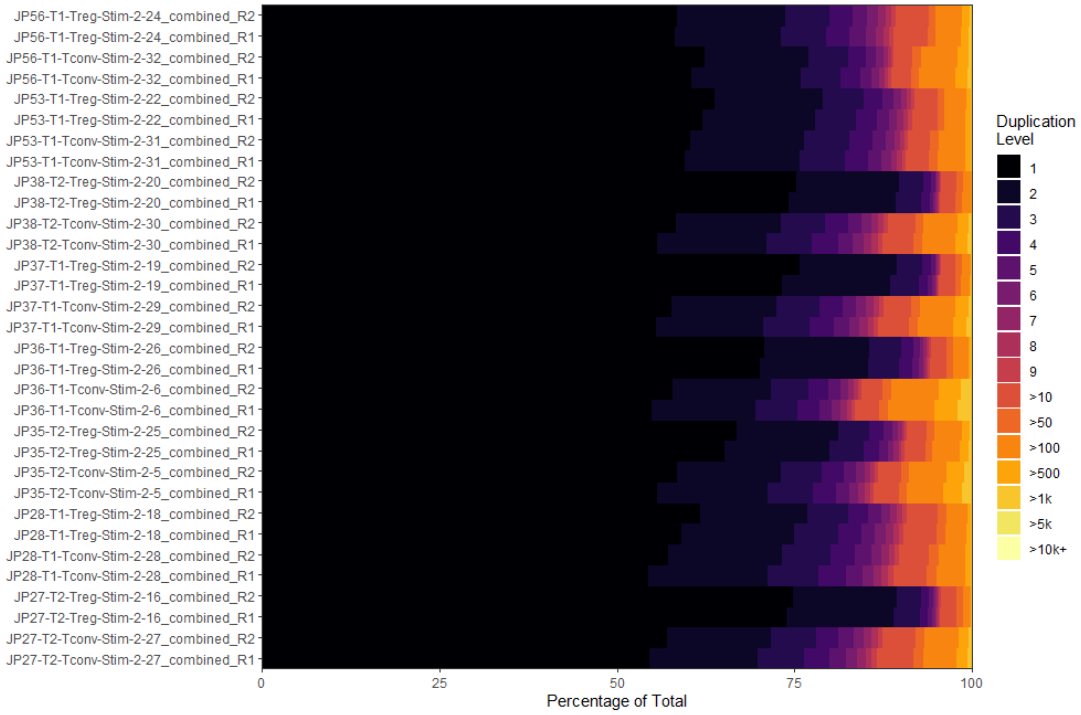


Figure 7-33 | Heatmap showing the duplication levels of case-control ATAC-seq libraries from one sequencing lane, L004.

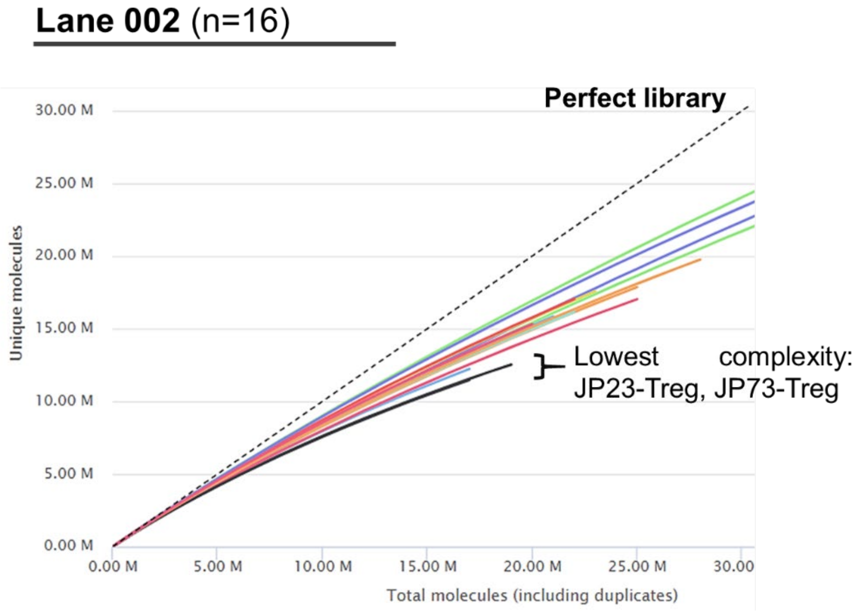


Figure 7-34 | Library complexity of sequenced case-control ATAC-seq libraries from one sequencing lane, L002.
Libraries with the lowest complexity for this sequencing lane are as indicated.

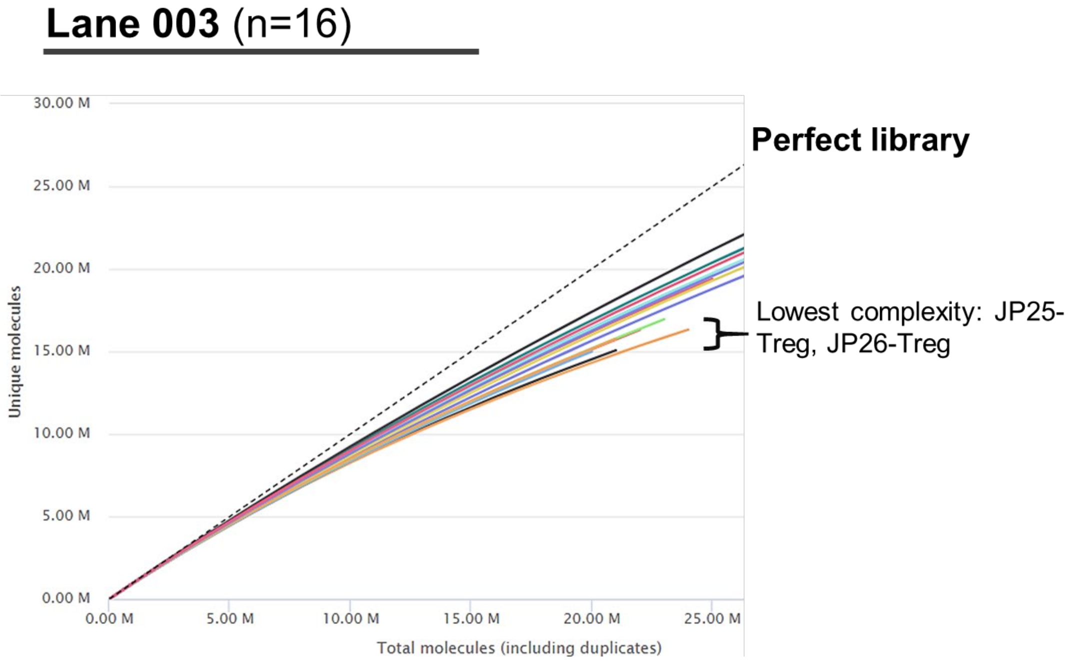


Figure 7-35 | Library complexity of sequenced case-control ATAC-seq libraries from one sequencing lane, L003.
Libraries with the lowest complexity for this sequencing lane are as indicated.

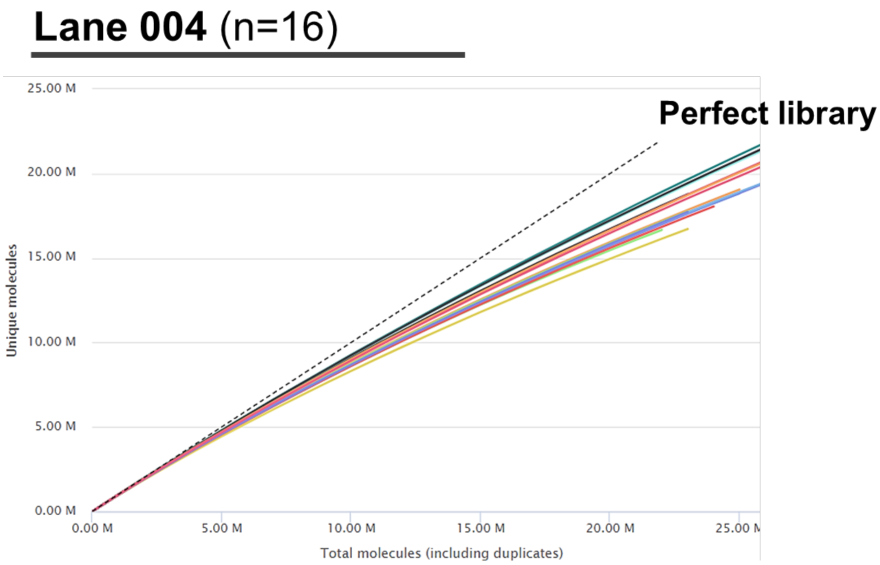


Figure 7-36 | Library complexity of sequenced case-control ATAC-seq libraries from one sequencing lane, L004.

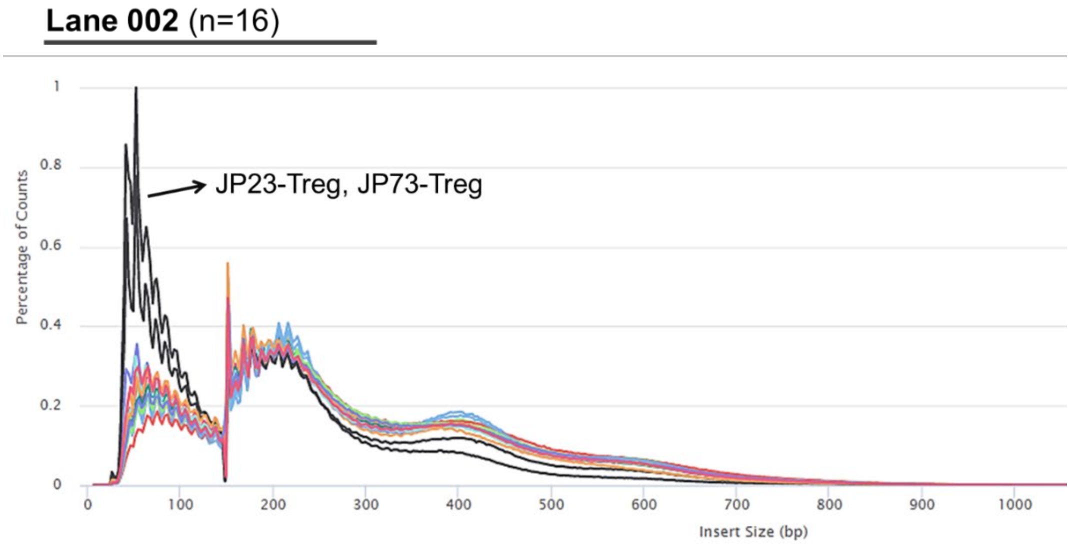


Figure 7-37 | The insert size distribution of case-control ATAC-seq libraries from one sequencing lane, L002.
Treg ATAC-seq libraries generated from JP23 and JP73 demonstrated higher enrichment of smaller DNA fragments compared with other libraries in this sequencing lane.

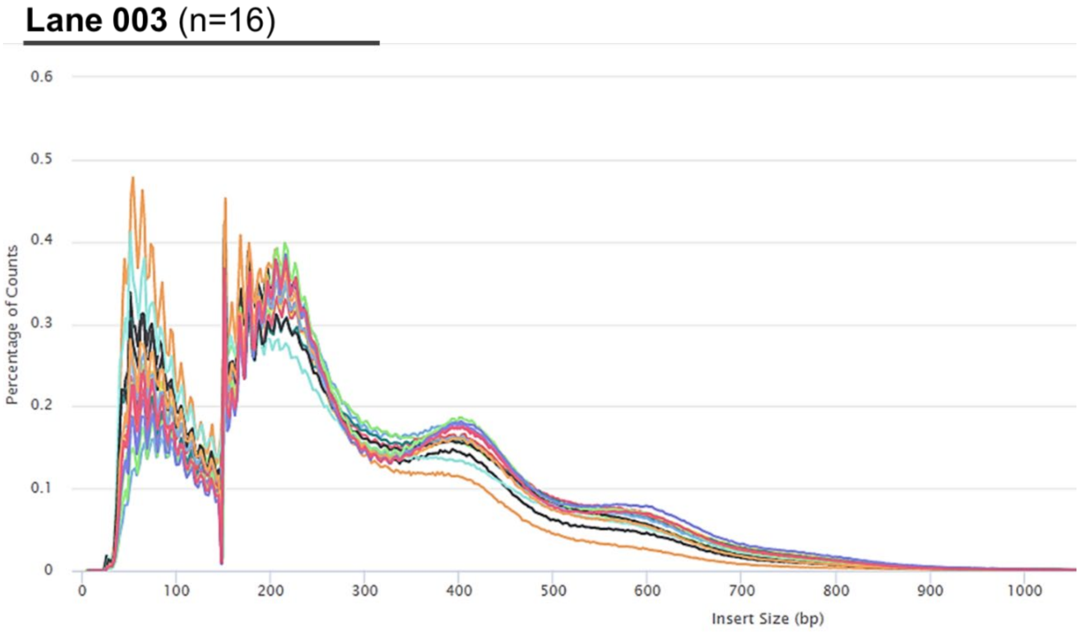


Figure 7-38 | The insert size distribution of case-control ATAC-seq libraries from one sequencing lane, L004.

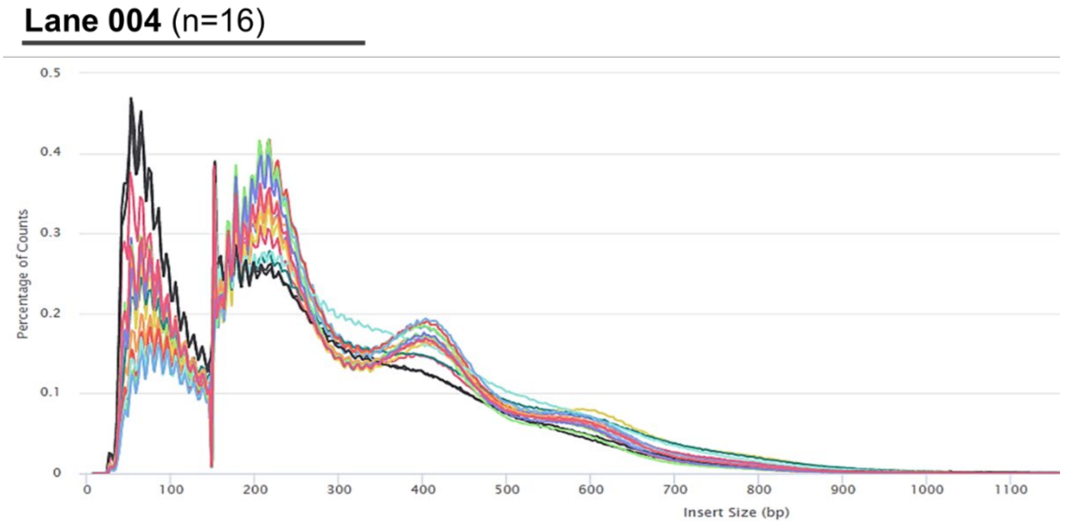
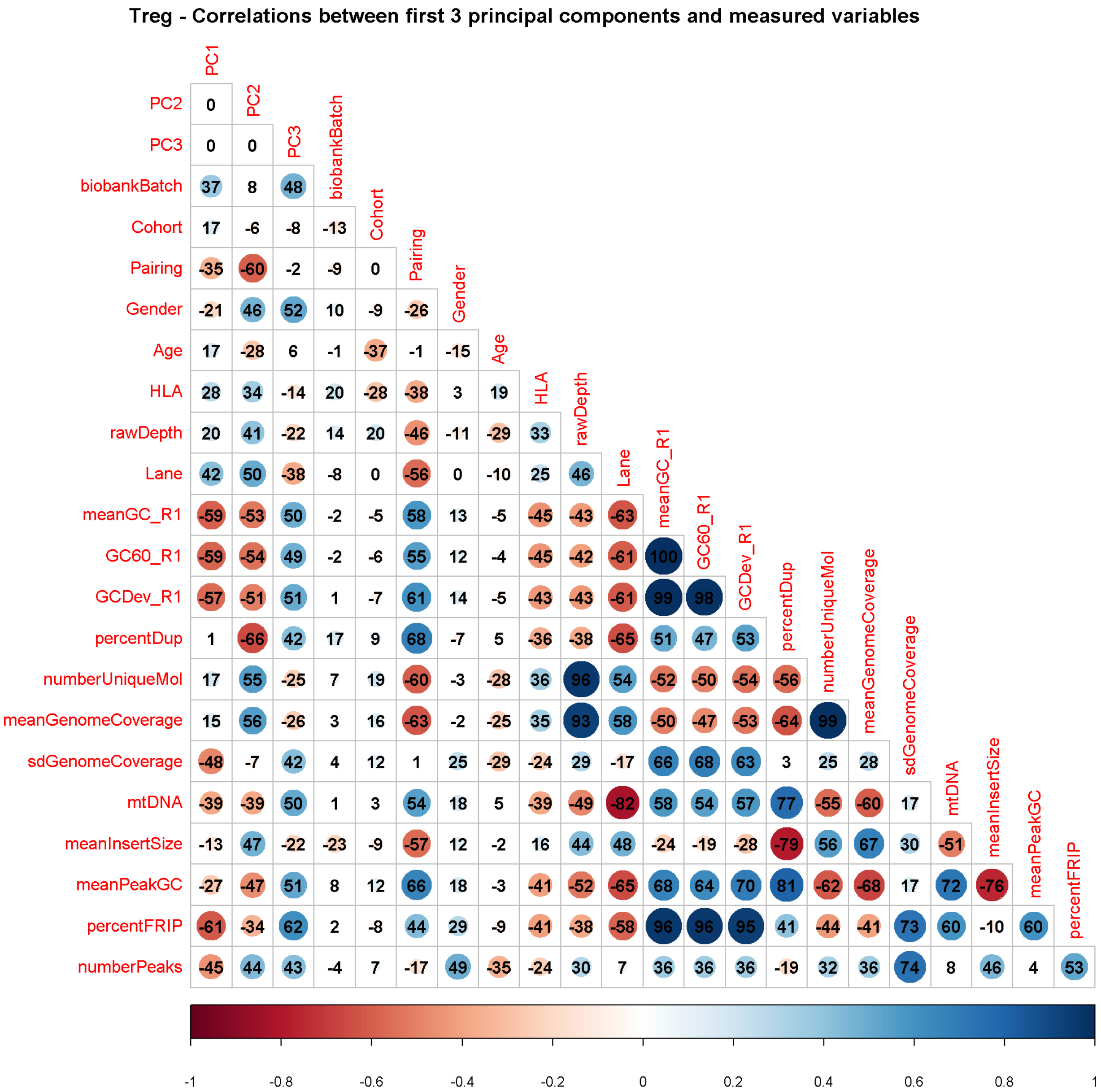


Figure 7-39 | The insert size distribution of case-control ATAC-seq libraries from one sequencing lane, L004.



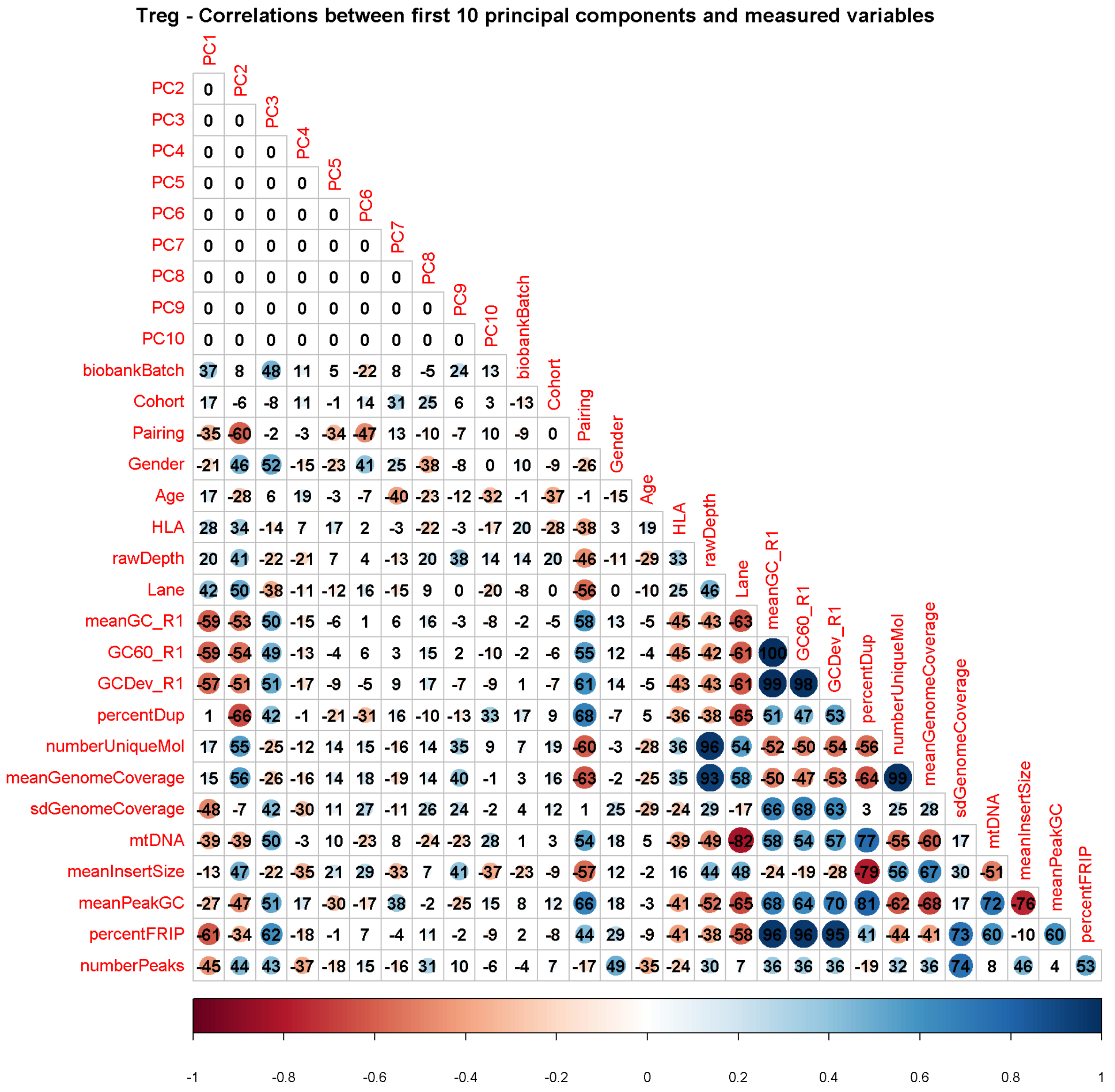
	PC1	PC2	PC3	biobankBatch	Cohort	Pairing	Gender	Age	HLA	rawDepth	Lane	meanGC_R1	GC60_R1	GCDev_R1	percentDup	numberUniqueMol	meanGenomeCoverage	sdGenomeCoverage	mtDNA	meanInsertSize	meanPeakGC	percentFRIP
PC2	0																					
PC3	0	0																				
biobankBatch	37	8	48																			
Cohort	17	-6	-8	-13																		
Pairing	-35	-60	-2	-9	0																	
Gender	-21	46	52	10	-9	-26																
Age	17	-28	6	-1	-37	-1	-15															
HLA	28	34	-14	20	-28	-38	3	19														
rawDepth	20	41	-22	14	20	-46	-11	-29	33													
Lane	42	50	-38	-8	0	-56	0	-10	25	46												
meanGC_R1	-59	-53	50	-2	-5	58	13	-5	-45	-43	-63											
GC60_R1	-59	-54	49	-2	-6	55	12	-4	-45	-42	-61	100										
GCDev_R1	-57	-51	51	1	-7	61	14	-5	-43	-43	-61	99	98									
percentDup	1	-66	42	17	9	68	-7	5	-36	-38	-65	51	47	53								
numberUniqueMol	17	55	-25	7	19	-60	-3	-28	36	96	54	-52	-50	-54	-56							
meanGenomeCoverage	15	56	-26	3	16	-63	-2	-25	35	93	58	-50	-47	-53	-64	99						
sdGenomeCoverage	-48	-7	42	4	12	1	25	-29	-24	29	-17	66	68	63	3	25	28					
mtDNA	-39	-39	50	1	3	54	18	5	-39	-49	-82	58	54	57	77	-55	-60	17				
meanInsertSize	-13	47	-22	-23	-9	-57	12	-2	16	44	48	-24	-19	-28	-79	56	67	30	-51			
meanPeakGC	-27	-47	51	8	12	66	18	-3	-41	-52	-65	68	64	70	81	-62	-68	17	72	-76		
percentFRIP	-61	-34	62	2	-8	44	29	-9	-41	-38	-58	96	96	95	41	-44	-41	73	60	-10	60	
numberPeaks	-45	44	43	-4	7	-17	49	-35	-24	30	7	36	36	36	-19	32	36	74	8	46	4	53

Figure 7-40 | Correlations between the first three principal components and measured variables for Treg ATAC-seq libraries generated from 12 T1D and healthy control samples.

PCA was performed using peak-level counts to assess general patterns in the data. Cohort, pairing, gender and HLA information were converted to an ordered categorical variable for the purposes of visualisation. See Table 7-3 for detailed description of measured variables.

Table 7-3. Description of measured variables for calculating correlations with the Principal Components (Figure 7-40, Figure 7-41, Figure 7-43 and Figure 7-44).

biobankBatch	Period where the PBMCS samples were biobanked, between 2015 to 2017.
Cohort	T1D or Control.
Pairing	Sibling-matched T1D-Control pairs, 12 families were involved in this study.
Age	Age at visit.
HLA	HLA genotypes of samples.
rawDepth	Number of raw reads obtained from sequencing.
Lane	Libraries were sequenced on three lanes of Hiseq platform.
meanGC_R1	Mean GC content of the libraries post adapter trimming (R1 reads only).
GC60_R1	Proportion of each library containing >60% GC content post adapter trimming (R1 reads only).
GCDev_R1	Standard deviation of observed GC frequencies from the expected GC frequencies (R1 reads only).
percentDup	Proportion of duplication in the mapped reads.
numberUniqueMol	Estimated number of unique molecules based on PE duplication.
meanGenomeCoverage	Mean coverage depth of mapped reads computed by *GenomicAlignments*.
sdGenomeCoverage	Standard deviation of coverage depth computed by *GenomicAlignments*.
mtDNA	Proportion of mitochondrial mapped reads.
meanInsertSize	Mean insert size of the libraries.
meanPeakGC	Mean GC content after summarisation to peak- level counts (computed from concatenated sequence in the master peaks).
percentFRIP	Proportion of reads in the peaks.
numberPeaks	Number of peaks called.



Treg - Correlations between first 10 principal components and measured variables

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	biobankBatch	Cohort	Pairing	Gender	Age	HLA	rawDepth	Lane	meanGC_R1	GC60_R1	GCDev_R1	percentDup	numberUniqueMol	meanGenomeCoverage	sdGenomeCoverage	mtDNA	meanInsertSize	meanPeakGC	percentFRIP
PC2	0																												
PC3	0	0																											
PC4	0	0	0																										
PC5	0	0	0	0																									
PC6	0	0	0	0	0																								
PC7	0	0	0	0	0	0																							
PC8	0	0	0	0	0	0	0																						
PC9	0	0	0	0	0	0	0	0																					
PC10	0	0	0	0	0	0	0	0	0																				
biobankBatch	37	8	48	11	5	-22	8	-5	24	13																			
Cohort	17	-6	-8	11	-1	14	31	25	6	3	-13																		
Pairing	-35	-60	-2	-3	-34	-47	13	-10	-7	10	-9	0																	
Gender	-21	46	52	-15	-23	41	25	-38	-8	0	10	-9	-26																
Age	17	-28	6	19	-3	-7	-40	-23	-12	-32	-1	-37	-1	-15															
HLA	28	34	-14	7	17	2	-3	-22	-3	-17	20	-28	-38	3	19														
rawDepth	20	41	-22	-21	7	4	-13	20	38	14	14	20	-46	-11	-29	33													
Lane	42	50	-38	-11	-12	16	-15	9	0	-20	-8	0	-56	0	-10	25	46												
meanGC_R1	-59	-53	50	-15	-6	1	6	16	-3	-8	-2	-5	58	13	-5	-45	-43	-63											
GC60_R1	-59	-54	49	-13	-4	6	3	15	2	-10	-2	-6	55	12	-4	-45	-42	-61	100										
GCDev_R1	-57	-51	51	-17	-9	-5	9	17	-7	-9	1	-7	61	14	-5	-43	-43	-61	99	98									
percentDup	1	-66	42	-1	-21	-31	16	-10	-13	33	17	9	68	-7	5	-36	-38	-65	51	47	53								
numberUniqueMol	17	55	-25	-12	14	15	-16	14	35	9	7	19	-60	-3	-28	36	96	54	-52	-50	-54	-56							
meanGenomeCoverage	15	56	-26	-16	14	18	-19	14	40	-1	3	16	-63	-2	-25	35	93	58	-50	-47	-53	-64	99						
sdGenomeCoverage	-48	-7	42	-30	11	27	-11	26	24	-2	4	12	1	25	-29	-24	29	-17	66	68	63	3	25	28					
mtDNA	-39	-39	50	-3	10	-23	8	-24	-23	28	1	3	54	18	5	-39	-49	-82	58	54	57	77	-55	-60	17				
meanInsertSize	-13	47	-22	-35	21	29	-33	7	41	-37	-23	-9	-57	12	-2	16	44	48	-24	-19	-28	-79	56	67	30	-51			
meanPeakGC	-27	-47	51	17	-30	-17	38	-2	-25	15	8	12	66	18	-3	-41	-52	-65	68	64	70	81	-62	-68	17	72	-76		
percentFRIP	-61	-34	62	-18	-1	7	-4	11	-2	-9	2	-8	44	29	-9	-41	-38	-58	96	96	95	41	-44	-41	73	60	-10	60	
numberPeaks	-45	44	43	-37	-18	15	-16	31	10	-6	-4	7	-17	49	-35	-24	30	7	36	36	36	-19	32	36	74	8	46	4	53

Figure 7-41 | Correlations between the first ten principal components and measured variables for Treg ATAC-seq libraries generated from 12 T1D and healthy control samples.

PCA was performed using peak-level counts to assess general patterns in the data. Cohort, pairing, gender and HLA information were converted to an ordered categorical variable for the purposes of visualisation. See Table 7-3 for detailed description of measured variables.

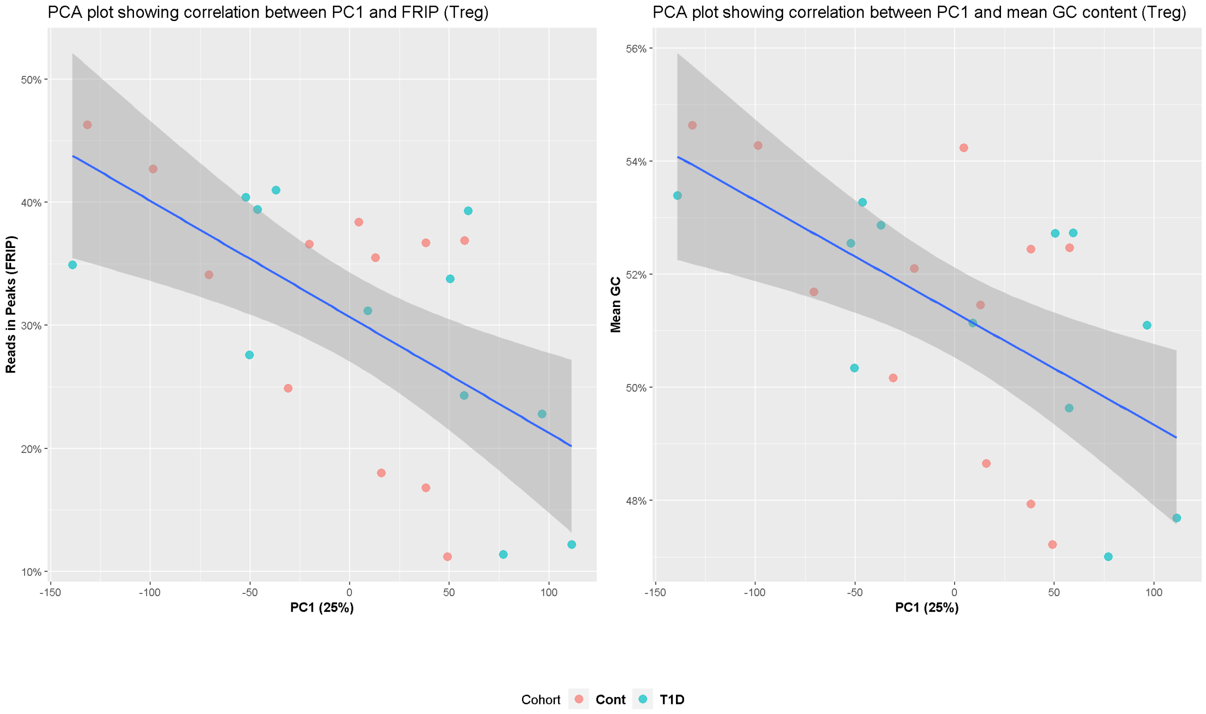
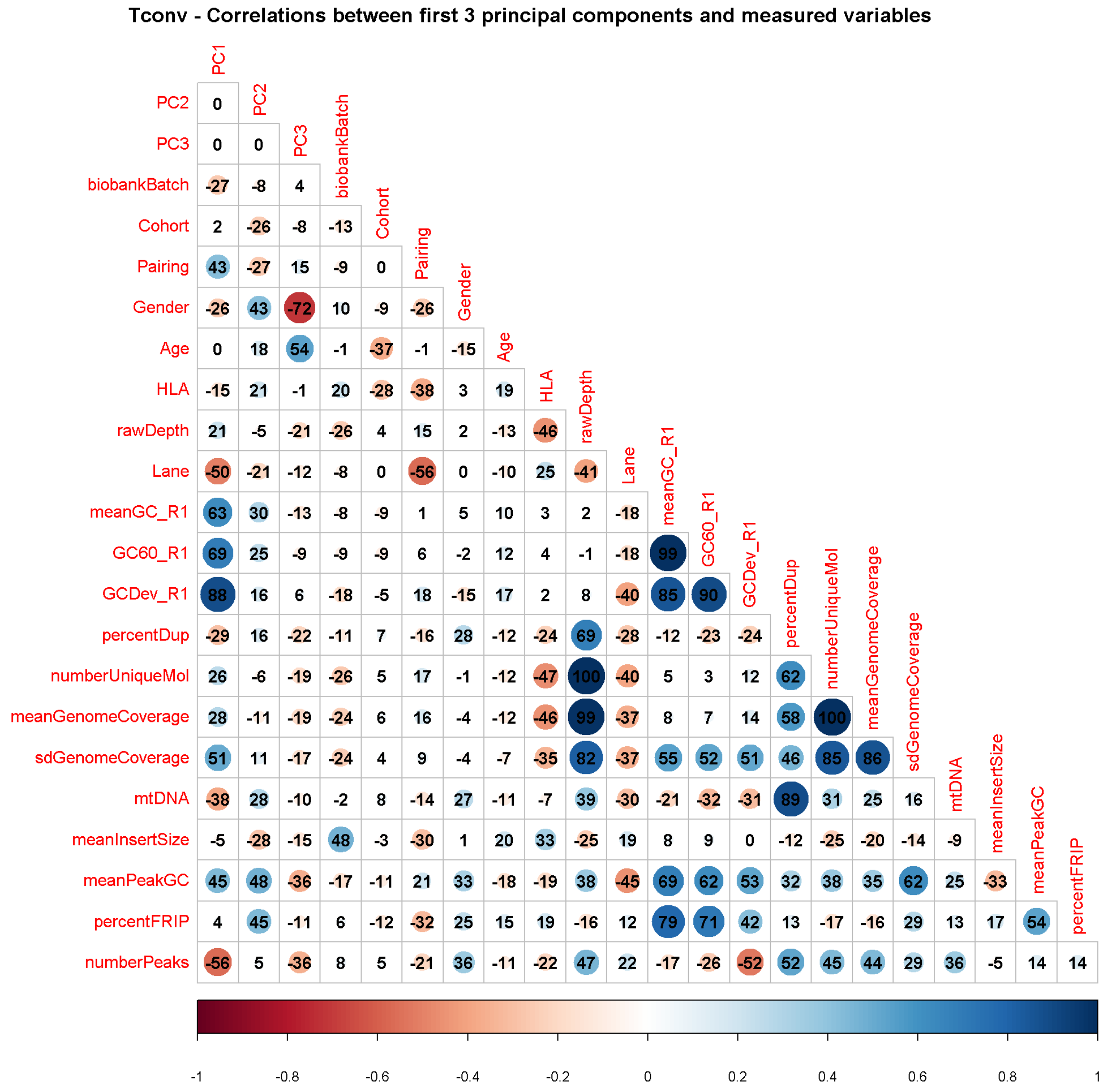


Figure 7-42 | Comparison of Principal component 1 and measured variables – fraction of reads in peaks (left) and mean QC (right) for Treg ATAC-seq libraries.
Regression lines are shown along with standard error bands for each comparison.

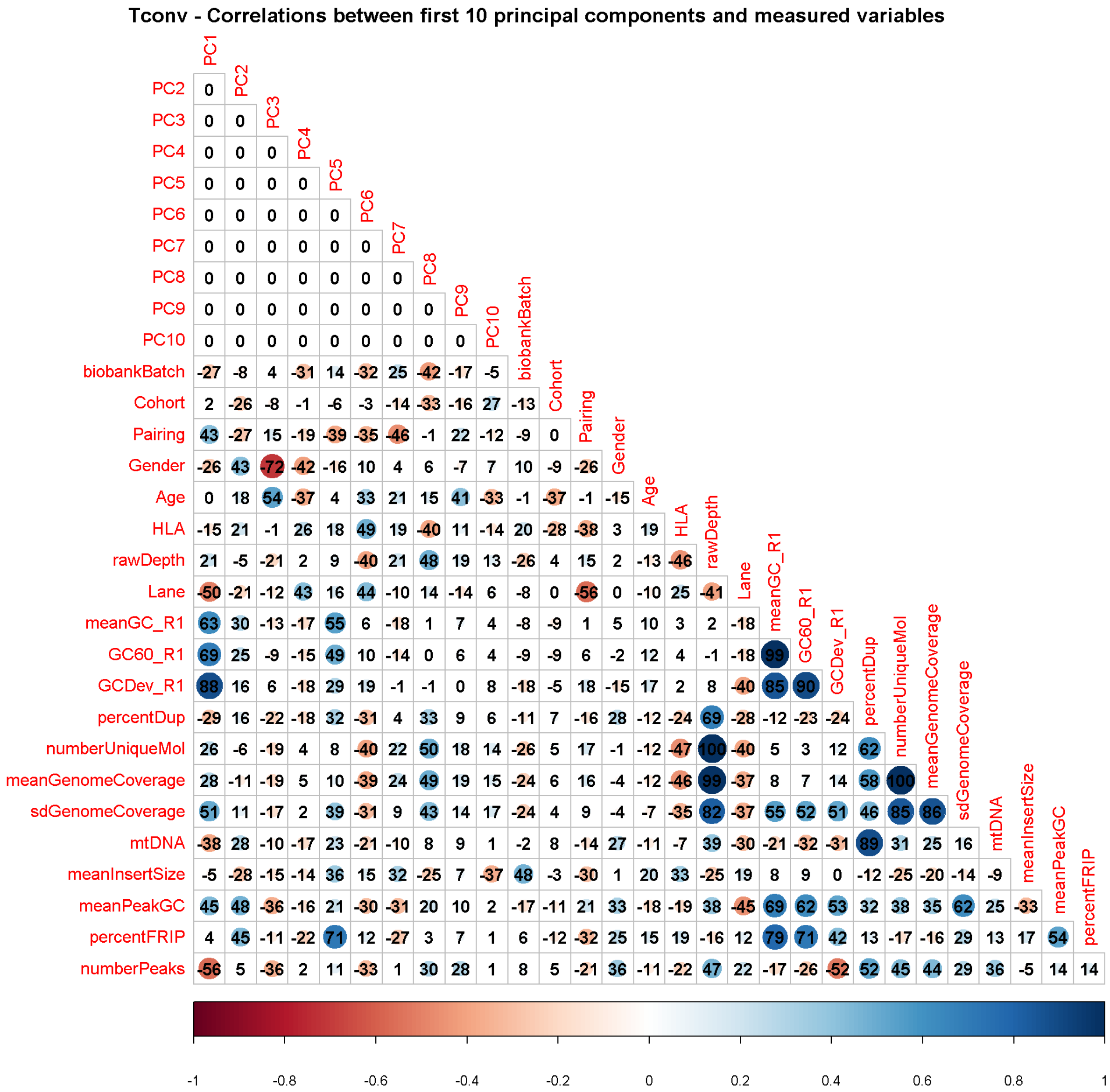


Tconv - Correlations between first 3 principal components and measured variables

	PC1	PC2	PC3	biobankBatch	Cohort	Pairing	Gender	Age	HLA	rawDepth	Lane	meanGC_R1	GC60_R1	GCDev_R1	percentDup	numberUniqueMol	meanGenomeCoverage	sdGenomeCoverage	mtDNA	meanInsertSize	meanPeakGC	percentFRIP
PC2	0																					
PC3	0	0																				
biobankBatch	-27	-8	4																			
Cohort	2	-26	-8	-13																		
Pairing	43	-27	15	-9	0																	
Gender	-26	43	-72	10	-9	-26																
Age	0	18	54	-1	-37	-1	-15															
HLA	-15	21	-1	20	-28	-38	3	19														
rawDepth	21	-5	-21	-26	4	15	2	-13	-46													
Lane	-50	-21	-12	-8	0	-56	0	-10	25	-41												
meanGC_R1	63	30	-13	-8	-9	1	5	10	3	2	-18											
GC60_R1	69	25	-9	-9	-9	6	-2	12	4	-1	-18	99										
GCDev_R1	88	16	6	-18	-5	18	-15	17	2	8	-40	85	90									
percentDup	-29	16	-22	-11	7	-16	28	-12	-24	69	-28	-12	-23	-24								
numberUniqueMol	26	-6	-19	-26	5	17	-1	-12	-47	100	-40	5	3	12	62							
meanGenomeCoverage	28	-11	-19	-24	6	16	-4	-12	-46	99	-37	8	7	14	58	100						
sdGenomeCoverage	51	11	-17	-24	4	9	-4	-7	-35	82	-37	55	52	51	46	85	86					
mtDNA	-38	28	-10	-2	8	-14	27	-11	-7	39	-30	-21	-32	-31	89	31	25	16				
meanInsertSize	-5	-28	-15	48	-3	-30	1	20	33	-25	19	8	9	0	-12	-25	-20	-14	-9			
meanPeakGC	45	48	-36	-17	-11	21	33	-18	-19	38	-45	69	62	53	32	38	35	62	25	-33		
percentFRIP	4	45	-11	6	-12	-32	25	15	19	-16	12	79	71	42	13	-17	-16	29	13	17	54	
numberPeaks	-56	5	-36	8	5	-21	36	-11	-22	47	22	-17	-26	-52	52	45	44	29	36	-5	14	14

Figure 7-43 | Correlations between the first three principal components and measured variables for Tconv ATAC-seq libraries generated from 12 T1D and healthy control samples.

PCA was performed using peak-level counts to assess general patterns in the data. Cohort, pairing, gender and HLA information were converted to an ordered categorical variable for the purposes of visualisation. See Table 7-3 for detailed description of measured variables.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	biobankBatch	Cohort	Pairing	Gender	Age	HLA	rawDepth	Lane	meanGC_R1	GC60_R1	GCDev_R1	percentDup	numberUniqueMol	meanGenomeCoverage	sdGenomeCoverage	mtDNA	meanInsertSize	meanPeakGC	percentFRIP
PC2	0																												
PC3	0	0																											
PC4	0	0	0																										
PC5	0	0	0	0																									
PC6	0	0	0	0	0																								
PC7	0	0	0	0	0	0																							
PC8	0	0	0	0	0	0	0																						
PC9	0	0	0	0	0	0	0	0																					
PC10	0	0	0	0	0	0	0	0	0																				
biobankBatch	-27	-8	4	-31	14	-32	25	-42	-17	-5																			
Cohort	2	-26	-8	-1	-6	-3	-14	-33	-16	27	-13																		
Pairing	43	-27	15	-19	-39	-35	-46	-1	22	-12	-9	0																	
Gender	-26	43	-72	-42	-16	10	4	6	-7	7	10	-9	-26																
Age	0	18	54	-37	4	33	21	15	41	-33	-1	-37	-1	-15															
HLA	-15	21	-1	26	18	49	19	-40	11	-14	20	-28	-38	3	19														
rawDepth	21	-5	-21	2	9	-40	21	48	19	13	-26	4	15	2	-13	-46													
Lane	-50	-21	-12	43	16	44	-10	14	-14	6	-8	0	-56	0	-10	25	-41												
meanGC_R1	63	30	-13	-17	55	6	-18	1	7	4	-8	-9	1	5	10	3	2	-18											
GC60_R1	69	25	-9	-15	49	10	-14	0	6	4	-9	-9	6	-2	12	4	-1	-18	99										
GCDev_R1	88	16	6	-18	29	19	-1	-1	0	8	-18	-5	18	-15	17	2	8	-40	85	90									
percentDup	-29	16	-22	-18	32	-31	4	33	9	6	-11	7	-16	28	-12	-24	69	-28	-12	-23	-24								
numberUniqueMol	26	-6	-19	4	8	-40	22	50	18	14	-26	5	17	-1	-12	-47	100	-40	5	3	12	62							
meanGenomeCoverage	28	-11	-19	5	10	-39	24	49	19	15	-24	6	16	-4	-12	-46	99	-37	8	7	14	58	100						
sdGenomeCoverage	51	11	-17	2	39	-31	9	43	14	17	-24	4	9	-4	-7	-35	82	-37	55	52	51	46	85	86					
mtDNA	-38	28	-10	-17	23	-21	-10	8	9	1	-2	8	-14	27	-11	-7	39	-30	-21	-32	-31	89	31	25	16				
meanInsertSize	-5	-28	-15	-14	36	15	32	-25	7	-37	48	-3	-30	1	20	33	-25	19	8	9	0	-12	-25	-20	-14	-9			
meanPeakGC	45	48	-36	-16	21	-30	-31	20	10	2	-17	-11	21	33	-18	-19	38	-45	69	62	53	32	38	35	62	25	-33		
percentFRIP	4	45	-11	-22	71	12	-27	3	7	1	6	-12	-32	25	15	19	-16	12	79	71	42	13	-17	-16	29	13	17	54	
numberPeaks	-56	5	-36	2	11	-33	1	30	28	1	8	5	-21	36	-11	-22	47	22	-17	-26	-52	52	45	44	29	36	-5	14	14

Figure 7-44 | Correlations between the first ten principal components and measured variables for Tconv ATAC-seq libraries generated from 12 T1D and healthy control samples.

PCA was performed using peak-level counts to assess general patterns in the data. Cohort, pairing, gender and HLA information were converted to an ordered categorical variable for the purposes of visualisation. See Table 7-3 for detailed description of measured variables.

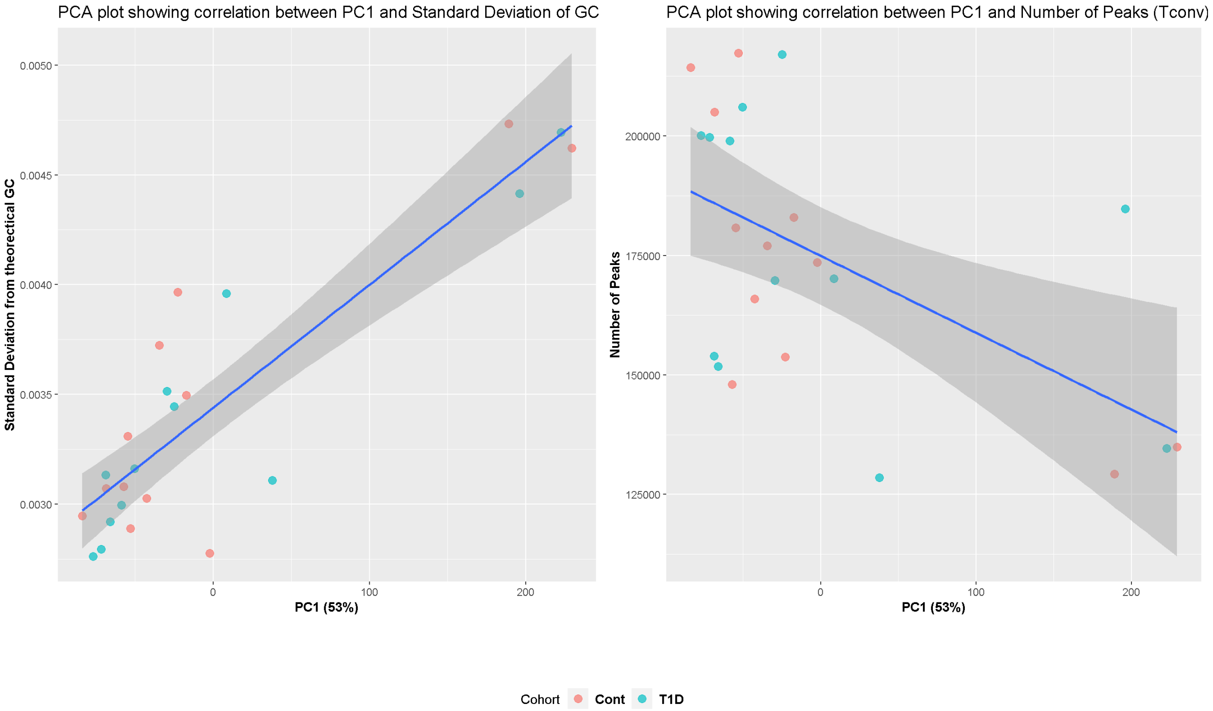


Figure 7-45 | Comparison of Principal component 1 and measured variables – standard deviation of observed GC frequencies from the expected GC frequencies (left) and number of peaks (right) for Tconv ATAC-seq libraries.

Regression lines are shown along with standard error bands for each comparison.

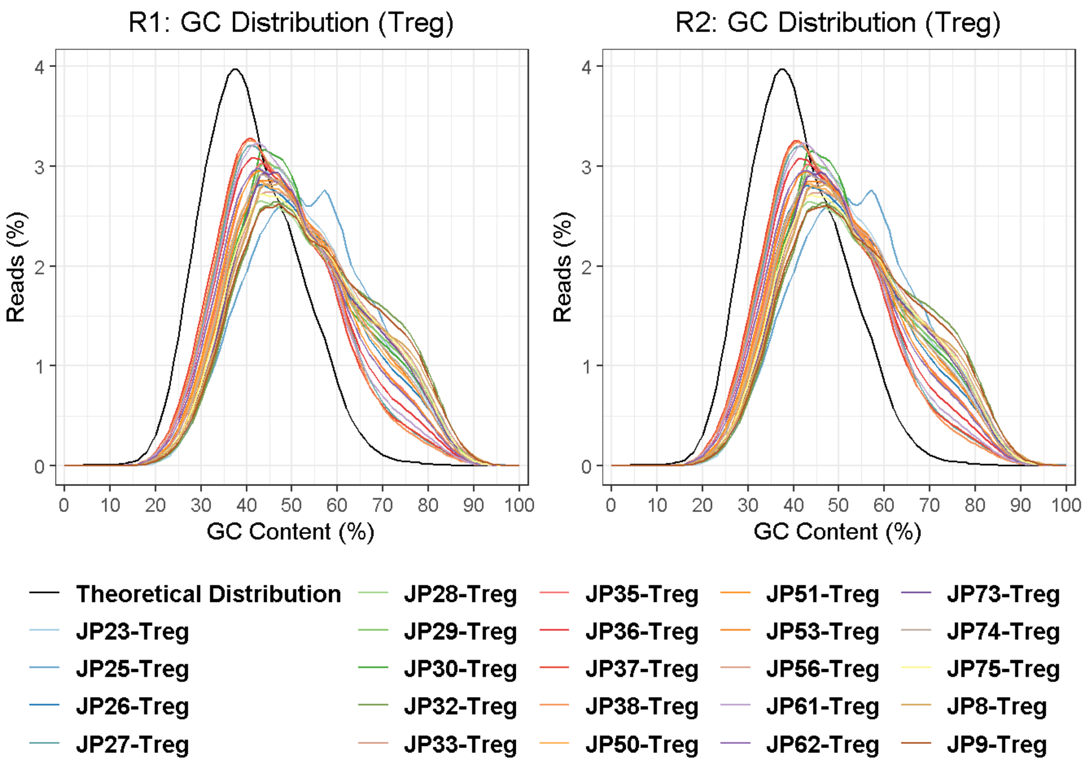


Figure 7-46 | GC content of case-control Treg ATAC-seq libraries from 12 T1D and sibling-matched healthy controls.

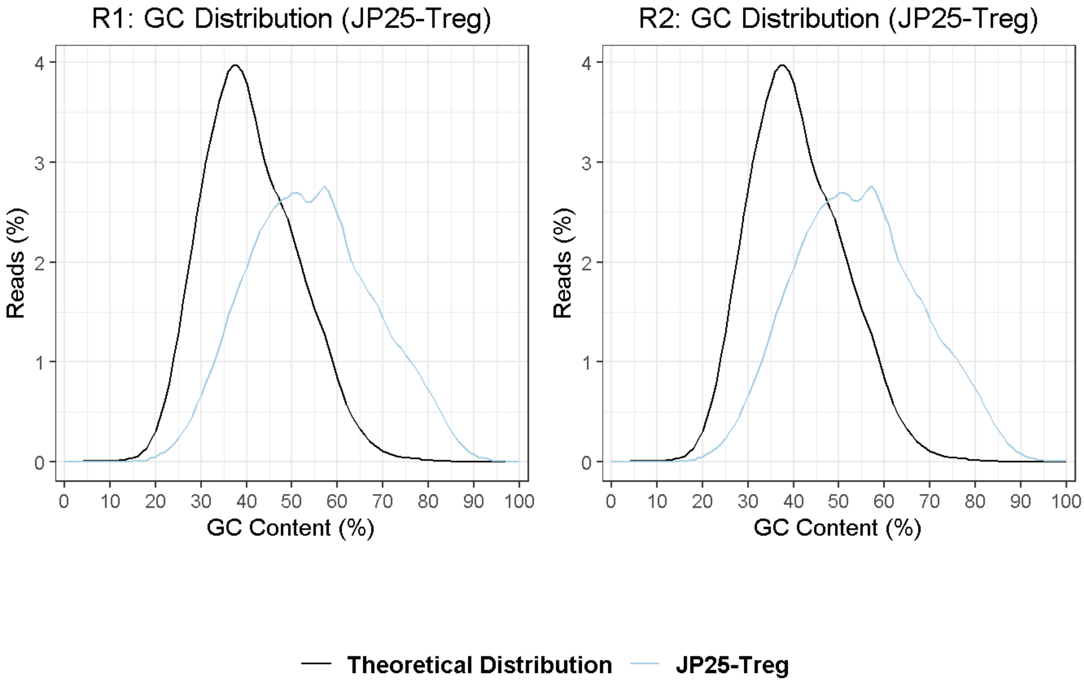


Figure 7-47 | Treg ATAC-seq library generated from JP25 showed a greater spike in GC content at 60% compared with other libraries (see Figure 7-46).

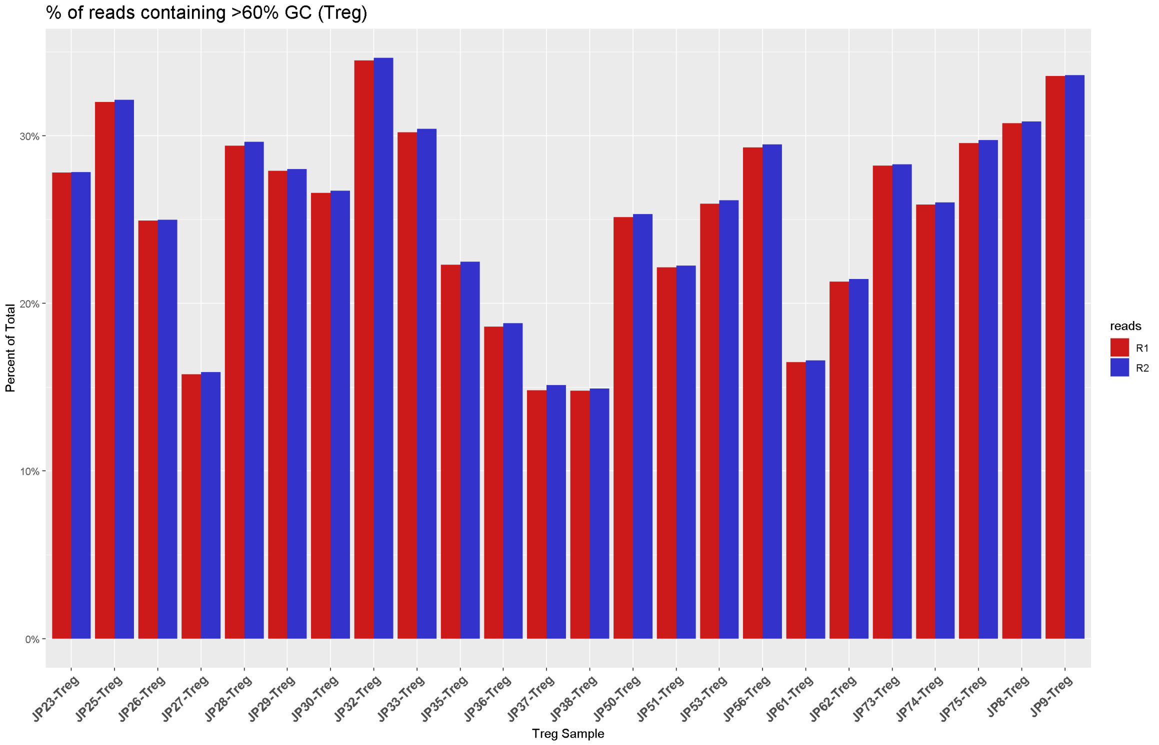


Figure 7-48 | Percentage of reads from each case-control Treg library which contain more than 60% GC content.

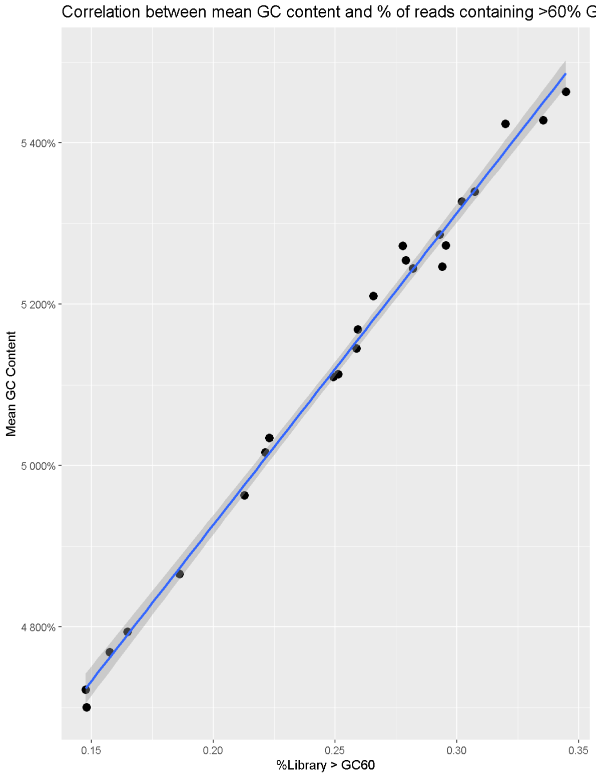


Figure 7-49 | Correlation between mean GC content and percentage of reads containing more than 60% GC content for case-control ATAC-seq libraries.

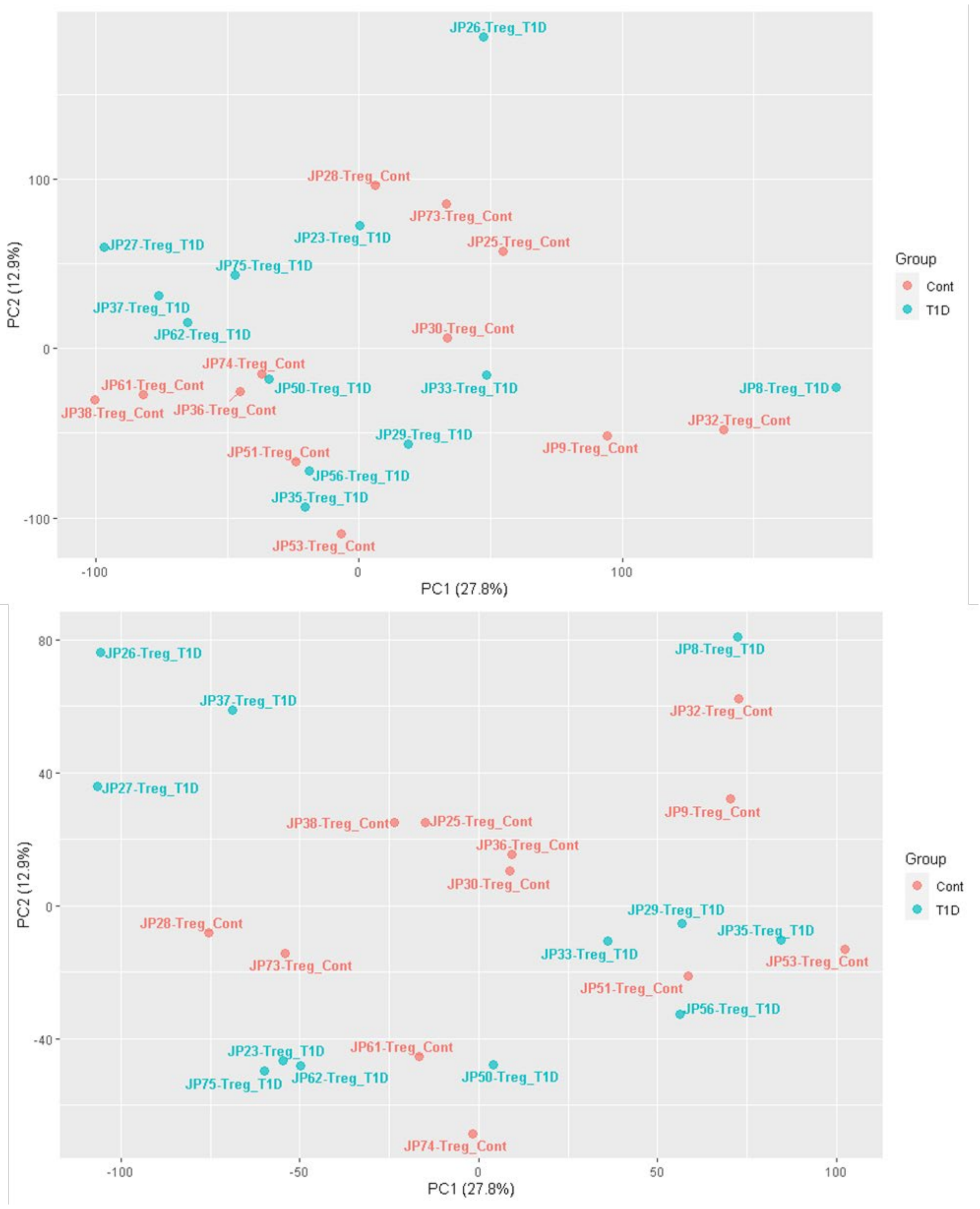


Figure 7-50 | Principal component analysis (PCA) plot of counts in peaks over first two components for case-control Treg ATAC-seq libraries before (top) and after (bottom) Conditional Quantile Normalisation.

As GC content was noted as being of concern for some of the samples in this dataset, conditional-quantile normalisation was performed using the cqn[40] package. This adds a peak and sample-level offset for each count which takes into account any systemic bias.

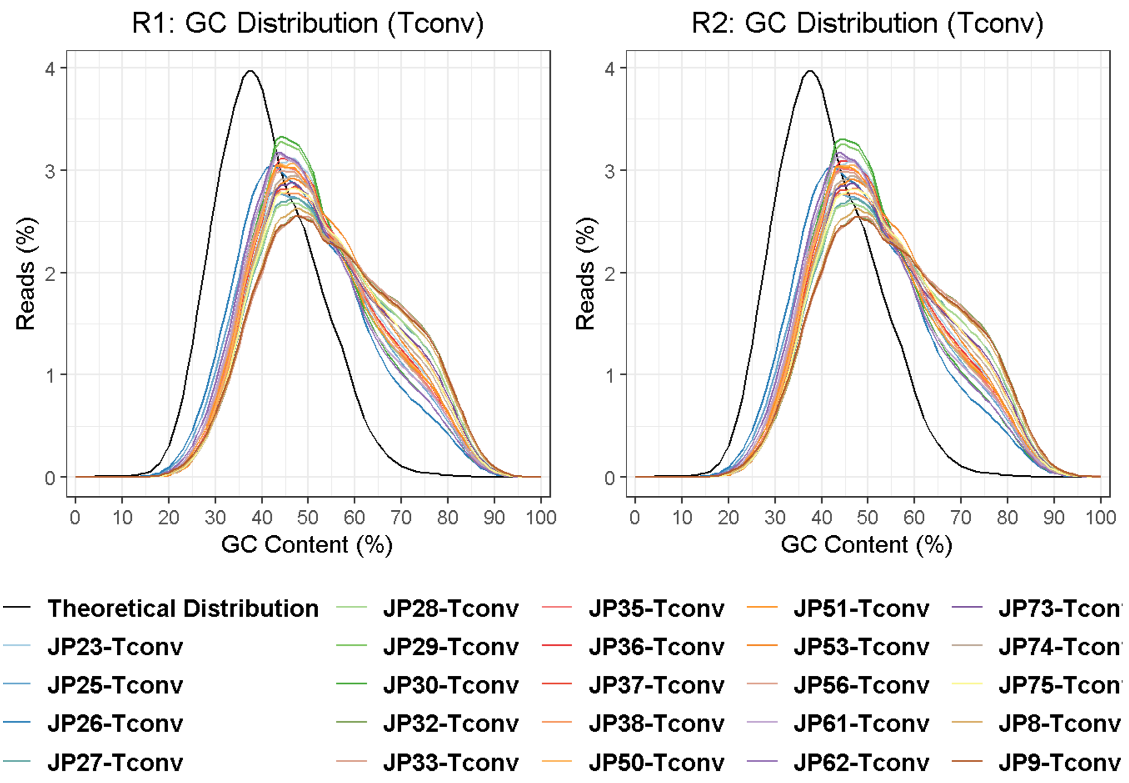
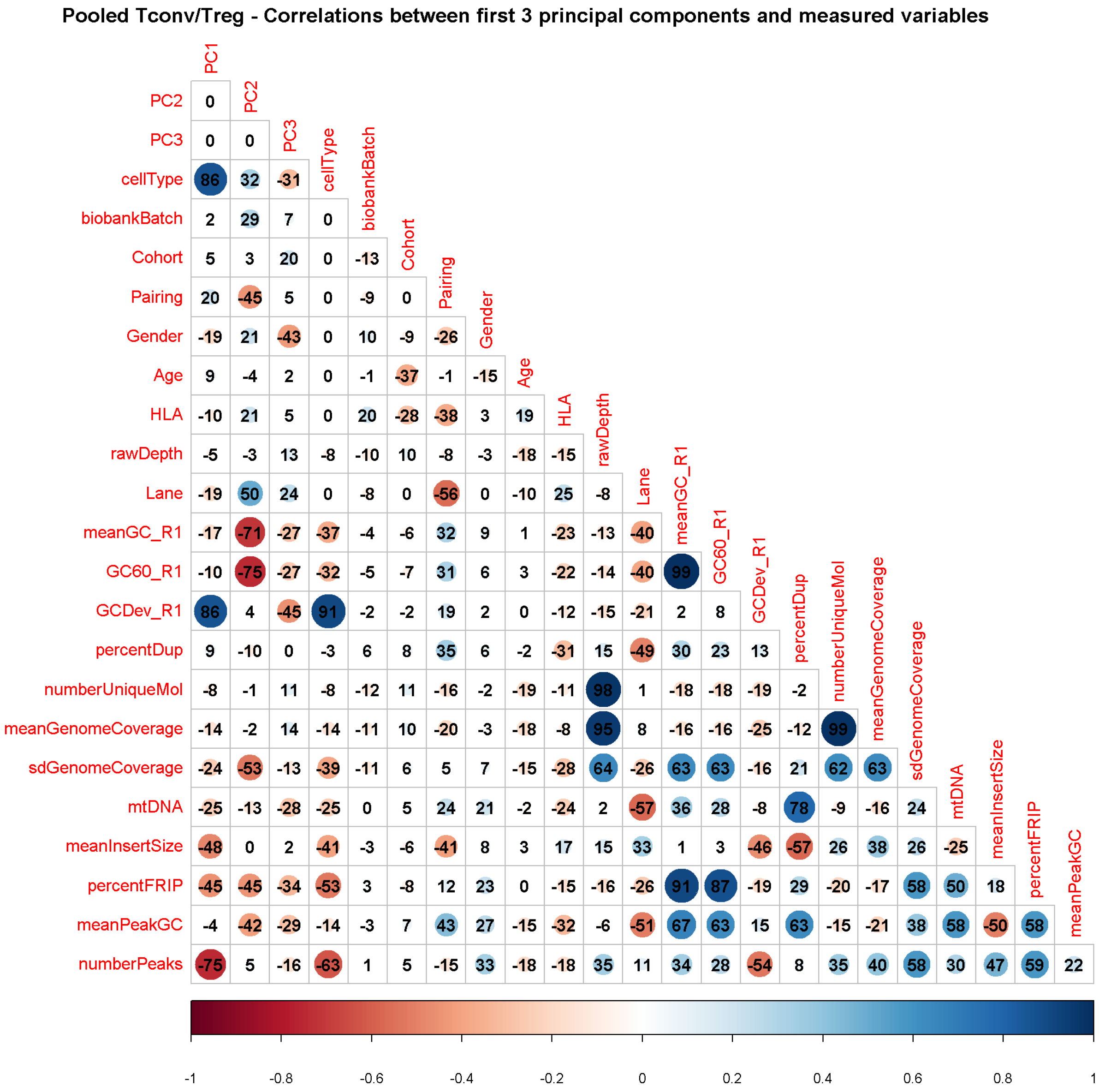


Figure 7-51 | GC content of case-control Tconv ATAC-seq libraries from 12 T1D and sibling-matched healthy controls.

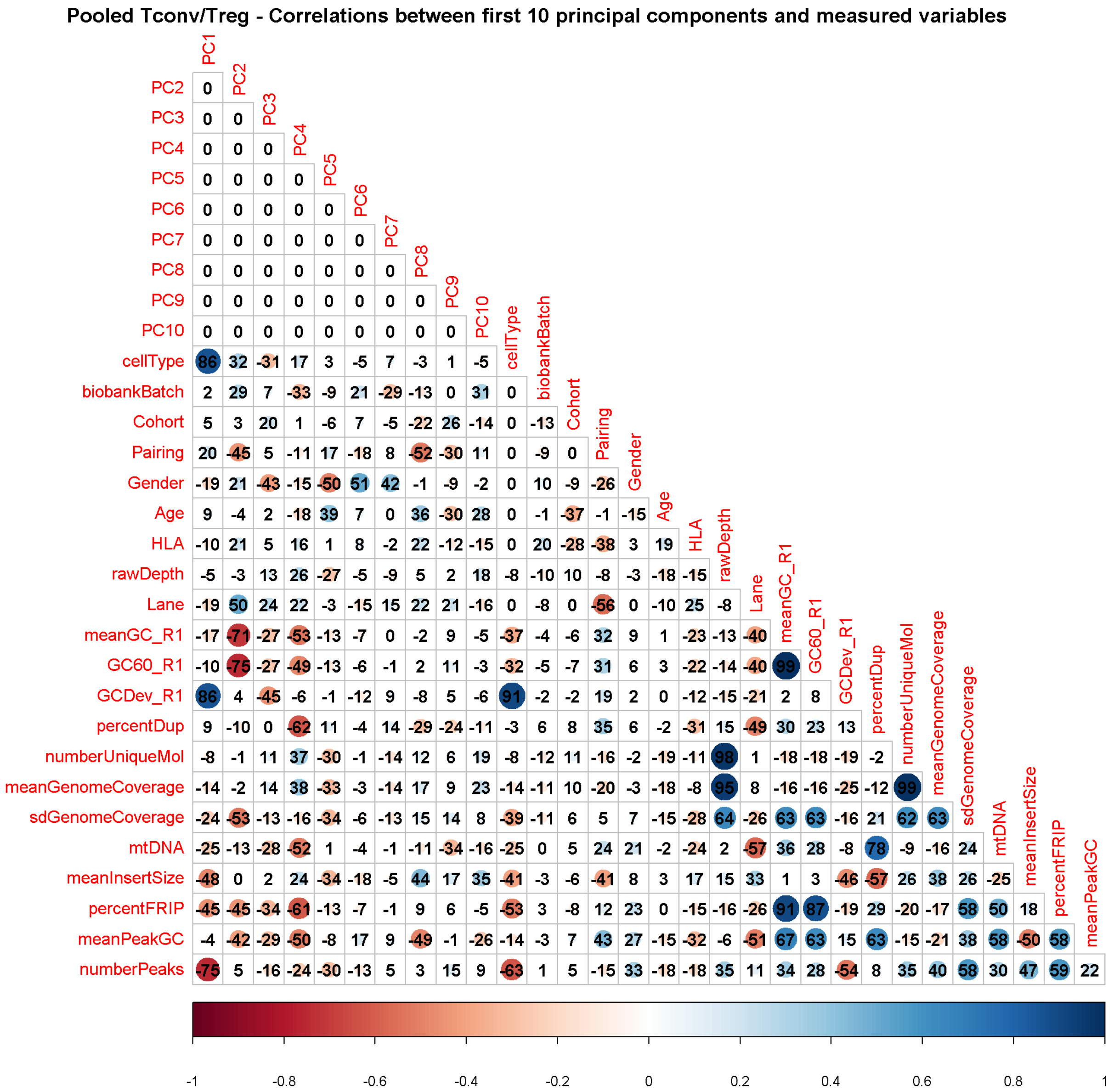


Pooled Tconv/Treg - Correlations between first 3 principal components and measured variables

	PC1	PC2	PC3	cellType	biobankBatch	Cohort	Pairing	Gender	Age	HLA	rawDepth	Lane	meanGC_R1	GC60_R1	GCDev_R1	percentDup	numberUniqueMol	meanGenomeCoverage	sdGenomeCoverage	mtDNA	meanInsertSize	percentFRIP	meanPeakGC
PC2	0																						
PC3	0	0																					
cellType	86	32	-31																				
biobankBatch	2	29	7	0																			
Cohort	5	3	20	0	-13																		
Pairing	20	-45	5	0	-9	0																	
Gender	-19	21	-43	0	10	-9	-26																
Age	9	-4	2	0	-1	-37	-1	-15															
HLA	-10	21	5	0	20	-28	-38	3	19														
rawDepth	-5	-3	13	-8	-10	10	-8	-3	-18	-15													
Lane	-19	50	24	0	-8	0	-56	0	-10	25	-8												
meanGC_R1	-17	-71	-27	-37	-4	-6	32	9	1	-23	-13	-40											
GC60_R1	-10	-75	-27	-32	-5	-7	31	6	3	-22	-14	-40	99										
GCDev_R1	86	4	-45	91	-2	-2	19	2	0	-12	-15	-21	2	8									
percentDup	9	-10	0	-3	6	8	35	6	-2	-31	15	-49	30	23	13								
numberUniqueMol	-8	-1	11	-8	-12	11	-16	-2	-19	-11	98	1	-18	-18	-19	-2							
meanGenomeCoverage	-14	-2	14	-14	-11	10	-20	-3	-18	-8	95	8	-16	-16	-25	-12	99						
sdGenomeCoverage	-24	-53	-13	-39	-11	6	5	7	-15	-28	64	-26	63	63	-16	21	62	63					
mtDNA	-25	-13	-28	-25	0	5	24	21	-2	-24	2	-57	36	28	-8	78	-9	-16	24				
meanInsertSize	-48	0	2	-41	-3	-6	-41	8	3	17	15	33	1	3	-46	-57	26	38	26	-25			
percentFRIP	-45	-45	-34	-53	3	-8	12	23	0	-15	-16	-26	91	87	-19	29	-20	-17	58	50	18		
meanPeakGC	-4	-42	-29	-14	-3	7	43	27	-15	-32	-6	-51	67	63	15	63	-15	-21	38	58	-50	58	
numberPeaks	-75	5	-16	-63	1	5	-15	33	-18	-18	35	11	34	28	-54	8	35	40	58	30	47	59	22

Figure 7-52 | Correlations between the first <u>three</u> principal components and measured variables for Tconv and Treg ATAC-seq libraries generated from 12 T1D and healthy control samples.

PCA was performed using peak-level counts to assess general patterns in the data. Cohort, pairing, gender and HLA information were converted to an ordered categorical variable for the purposes of visualisation. See Table 7-3 for detailed description of measured variables.



Pooled Tconv/Treg - Correlations between first 10 principal components and measured variables

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	cellType	biobankBatch	Cohort	Pairing	Gender	Age	HLA	rawDepth	Lane	meanGC_R1	GC60_R1	GCDev_R1	percentDup	numberUniqueMol	meanGenomeCoverage	sdGenomeCoverage	mtDNA	meanInsertSize	percentFRIP	meanPeakGC
PC2	0																													
PC3	0	0																												
PC4	0	0	0																											
PC5	0	0	0	0																										
PC6	0	0	0	0	0																									
PC7	0	0	0	0	0	0																								
PC8	0	0	0	0	0	0	0																							
PC9	0	0	0	0	0	0	0	0																						
PC10	0	0	0	0	0	0	0	0	0																					
cellType	86	32	-31	17	3	-5	7	-3	1	-5																				
biobankBatch	2	29	7	-33	-9	21	-29	-13	0	31	0																			
Cohort	5	3	20	1	-6	7	-5	-22	26	-14	0	-13																		
Pairing	20	-45	5	-11	17	-18	8	-52	-30	11	0	-9	0																	
Gender	-19	21	-43	-15	-50	51	42	-1	-9	-2	0	10	-9	-26																
Age	9	-4	2	-18	39	7	0	36	-30	28	0	-1	-37	-1	-15															
HLA	-10	21	5	16	1	8	-2	22	-12	-15	0	20	-28	-38	3	19														
rawDepth	-5	-3	13	26	-27	-5	-9	5	2	18	-8	-10	10	-8	-3	-18	-15													
Lane	-19	50	24	22	-3	-15	15	22	21	-16	0	-8	0	-56	0	-10	25	-8												
meanGC_R1	-17	-71	-27	-53	-13	-7	0	-2	9	-5	-37	-4	-6	32	9	1	-23	-13	-40											
GC60_R1	-10	-75	-27	-49	-13	-6	-1	2	11	-3	-32	-5	-7	31	6	3	-22	-14	-40	99										
GCDev_R1	86	4	-45	-6	-1	-12	9	-8	5	-6	91	-2	-2	19	2	0	-12	-15	-21	2	8									
percentDup	9	-10	0	-62	11	-4	14	-29	-24	-11	-3	6	8	35	6	-2	-31	15	-49	30	23	13								
numberUniqueMol	-8	-1	11	37	-30	-1	-14	12	6	19	-8	-12	11	-16	-2	-19	-11	98	1	-18	-18	-19	-2							
meanGenomeCoverage	-14	-2	14	38	-33	-3	-14	17	9	23	-14	-11	10	-20	-3	-18	-8	95	8	-16	-16	-25	-12	99						
sdGenomeCoverage	-24	-53	-13	-16	-34	-6	-13	15	14	8	-39	-11	6	5	7	-15	-28	64	-26	63	63	-16	21	62	63					
mtDNA	-25	-13	-28	-52	1	-4	-1	-11	-34	-16	-25	0	5	24	21	-2	-24	2	-57	36	28	-8	78	-9	-16	24				
meanInsertSize	-48	0	2	24	-34	-18	-5	44	17	35	-41	-3	-6	-41	8	3	17	15	33	1	3	-46	-57	26	38	26	-25			
percentFRIP	-45	-45	-34	-61	-13	-7	-1	9	6	-5	-53	3	-8	12	23	0	-15	-16	-26	91	87	-19	29	-20	-17	58	50	18		
meanPeakGC	-4	-42	-29	-50	-8	17	9	-49	-1	-26	-14	-3	7	43	27	-15	-32	-6	-51	67	63	15	63	-15	-21	38	58	-50	58	
numberPeaks	-75	5	-16	-24	-30	-13	5	3	15	9	-63	1	5	-15	33	-18	-18	35	11	34	28	-54	8	35	40	58	30	47	59	22

Figure 7-53 | Correlations between the first <u>ten</u> principal components and measured variables for Tconv and Treg ATAC-seq libraries generated from 12 T1D and healthy control samples.

PCA was performed using peak-level counts to assess general patterns in the data. Cohort, pairing, gender and HLA information were converted to an ordered categorical variable for the purposes of visualisation. See Table 7-3 for detailed description of measured variables.

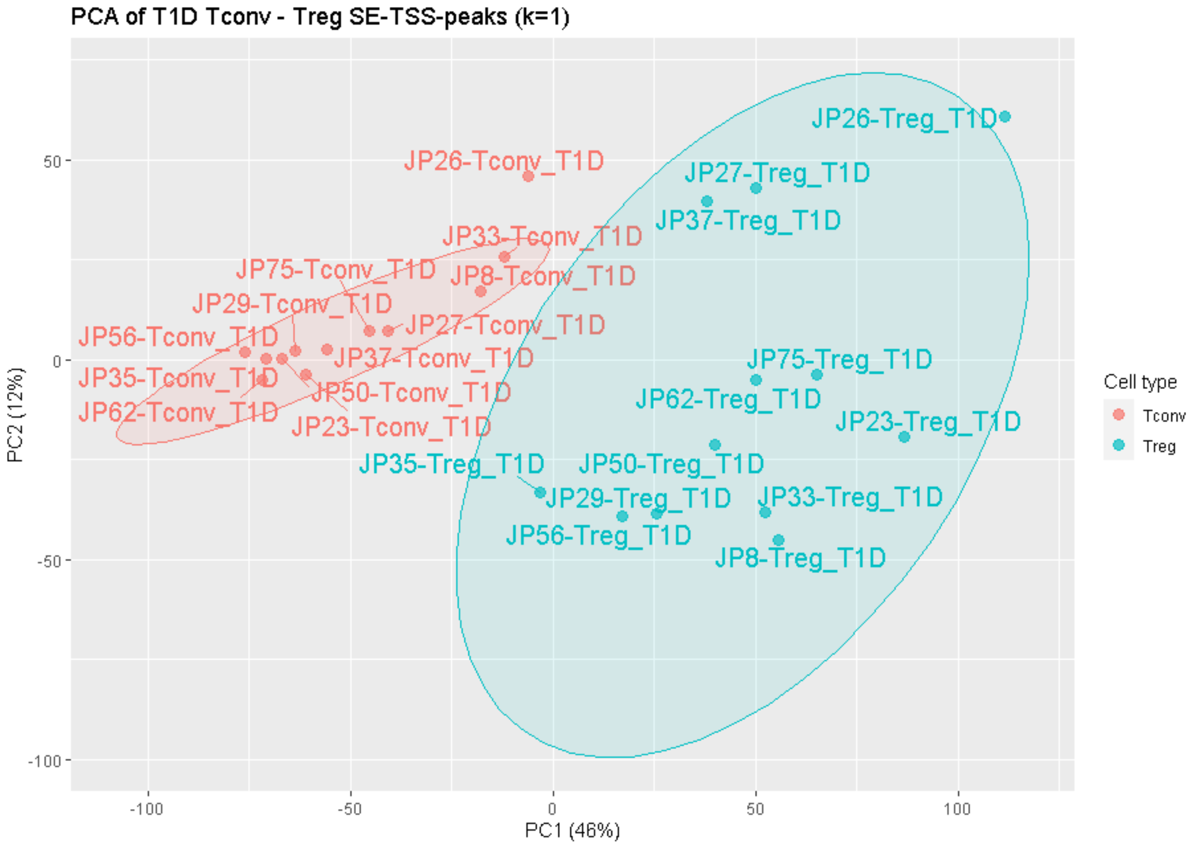


Figure 7-54 | Principal component analysis (PCA) plot of normalised counts in peaks over first two components for Tconv and Treg ATAC-seq libraries generated from 12 T1D and sibling-matched healthy control samples.

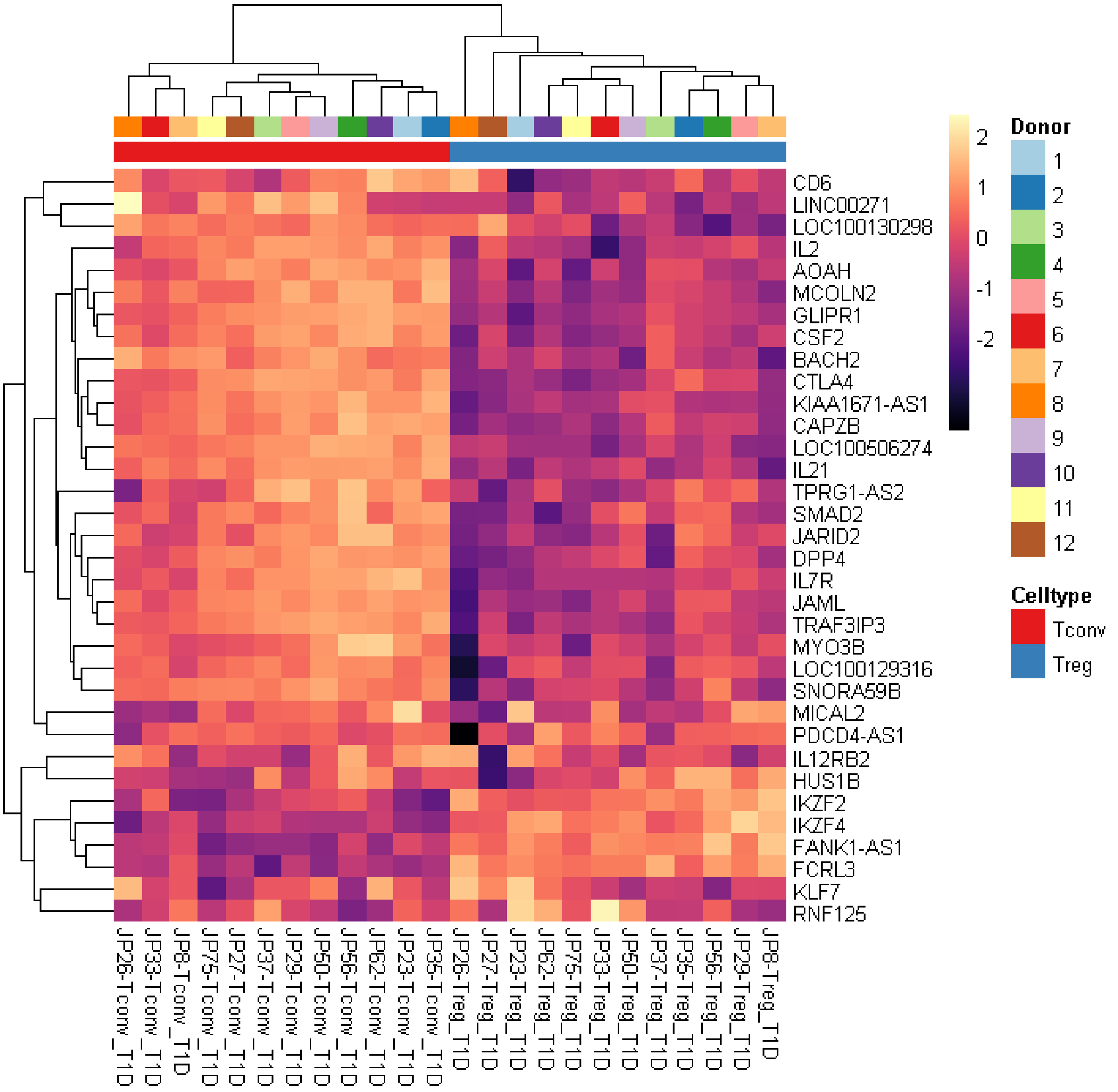


Figure 7-55 | Heatmap showing top 30 differential ATAC-seq peaks between Treg and Tconv cells from 12 T1D patients.

Differential peaks were annotated to the nearest TSS in linear genomic distance. Results shown are representative of 12 T1D subjects.

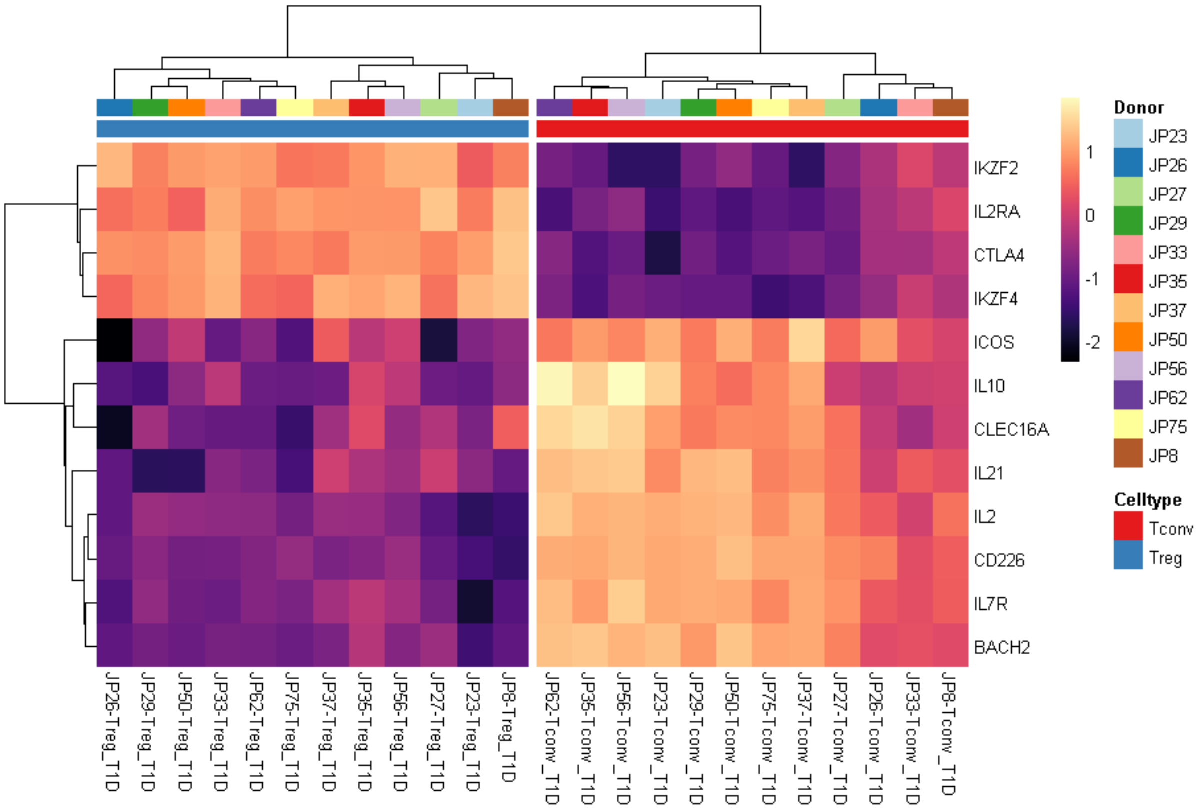


Figure 7-56 | Heatmap showing accessibility patterns of differential ATAC-seq peaks at known T1D loci between Treg and Tconv cells from 12 T1D patients.
Differential peaks were annotated to the nearest TSS in linear genomic distance. Results shown are representative of 12 T1D subjects.

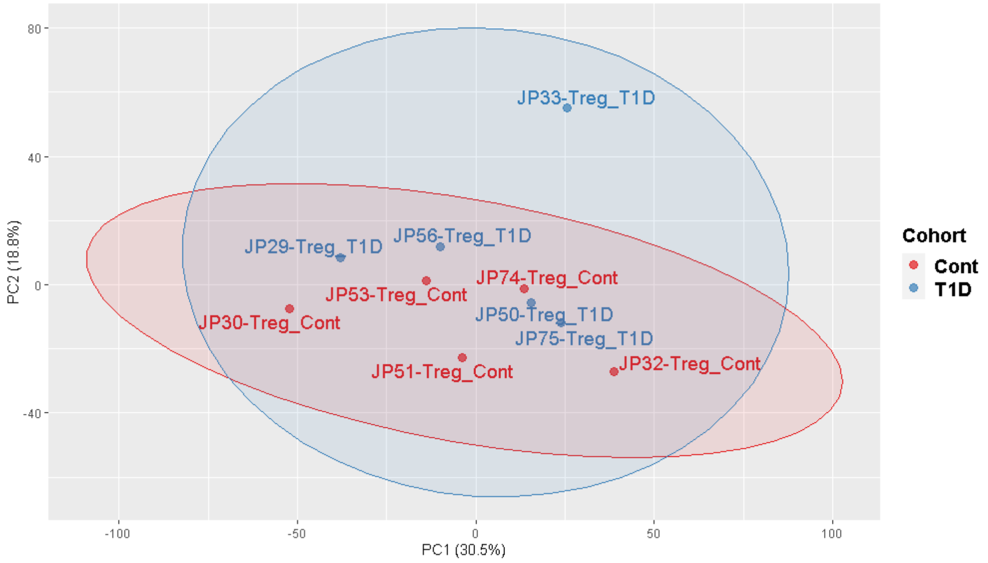


Figure 7-57 | Principal component analysis (PCA) plot of normalised counts in peaks over first two components for Treg ATAC-seq libraries generated from 5 T1D and gender- and sibling- matched healthy control samples.

Input peakset was generated from top 100,000 peaks called from pooled reads of 5 T1D or gender- and sibling- matched healthy ATAC-seq libraries intersecting TSS, typical enhancer regions of Treg cells[8], super- and typical- enhancer regions of CD4^{+} T cells [7](Figure 5-18) (right column).

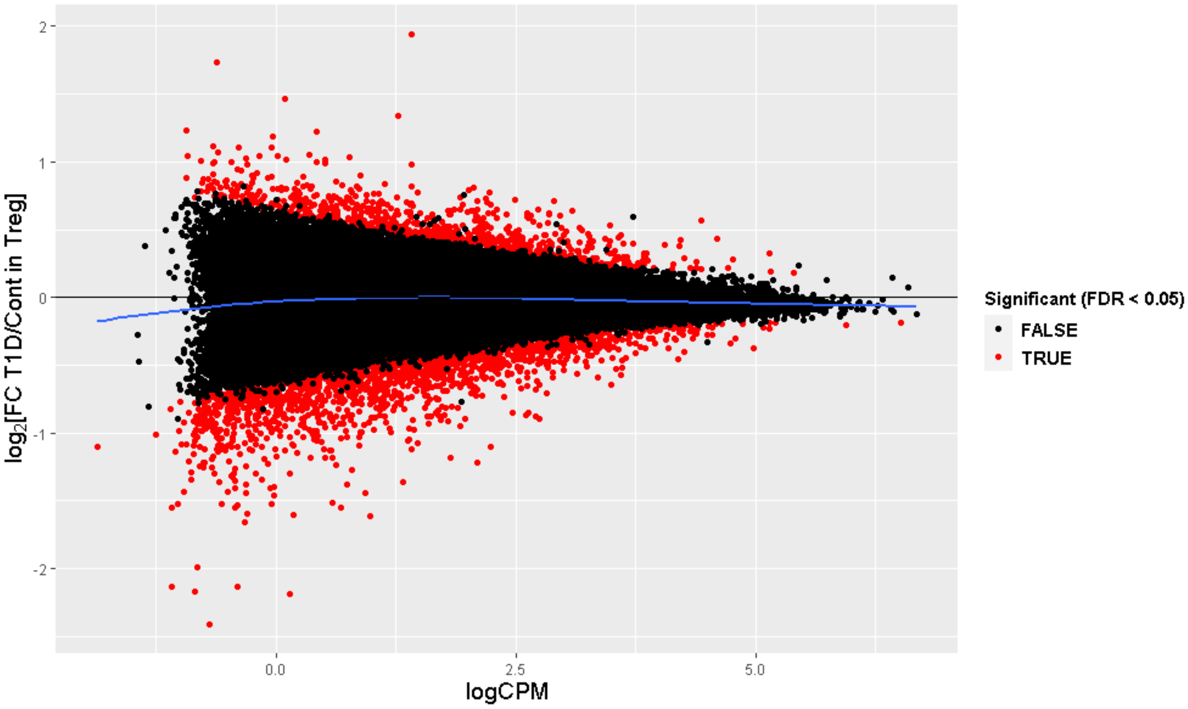


Figure 7-58 | MA plots depicting global differential accessibility distributions from the Treg gender-matched (5 male T1D-healthy pairs) ATAC-seq libraries.
The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Treg cells [8], super- and typical- enhancers of CD4^{+} T cells [7].

The data shown are representative of counts in peaks from 5 T1D and gender- and sibling-matched healthy control samples. X-axis of MA plot represents average peak signal abundance at that region, while Y-axis corresponds to the log2 difference in peak signal between Treg and Tconv. Black dots represent non-significant regions, and red dots represent significant (FDR < 0.05) differential regions. Blue lines are loess fits to each distribution.

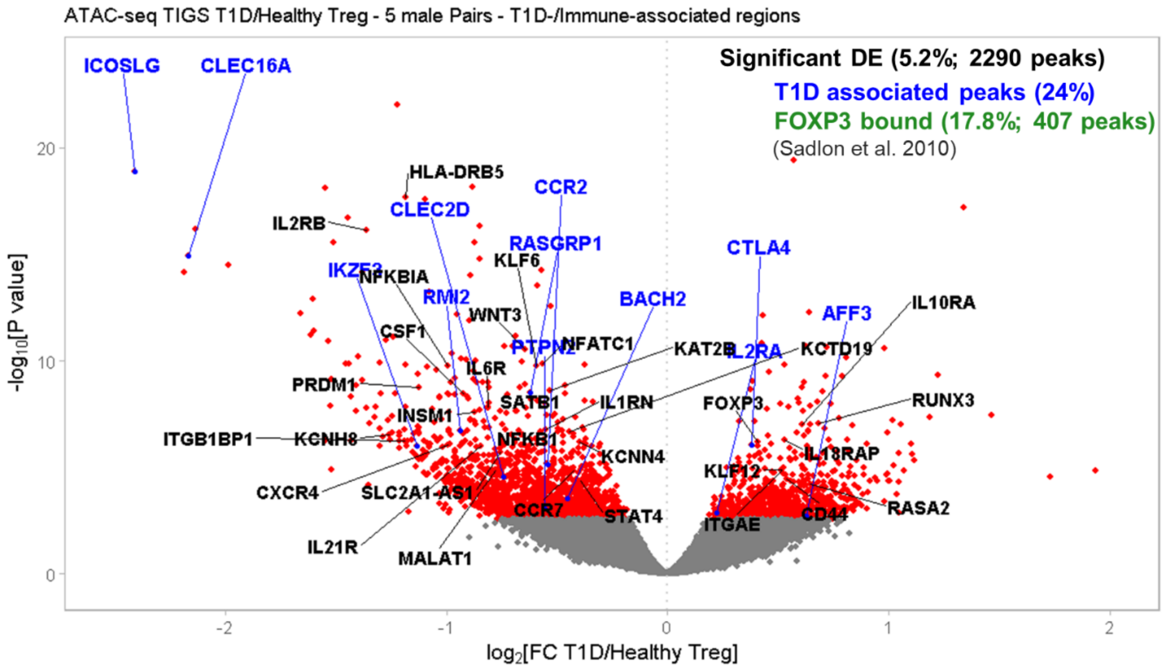


Figure 7-59 | Volcano plot showing differential accessibility analysis between Treg from 5 T1D and gender- and sibling- matched healthy subjects.

The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Treg cells [8], super- and typical- enhancers of CD4$^+$ T cells [7].

Regions having significantly differential accessibility (Benjamin Hochberg FDR < 0.05) are coloured red or blue. Differentially accessible regions in blue were previously known to be associated with T1D [2, 10]. Differential peaks were annotated to the nearest TSS in linear genomic distance. Selected differentially accessible immune-relevant loci were annotated with gene symbols.

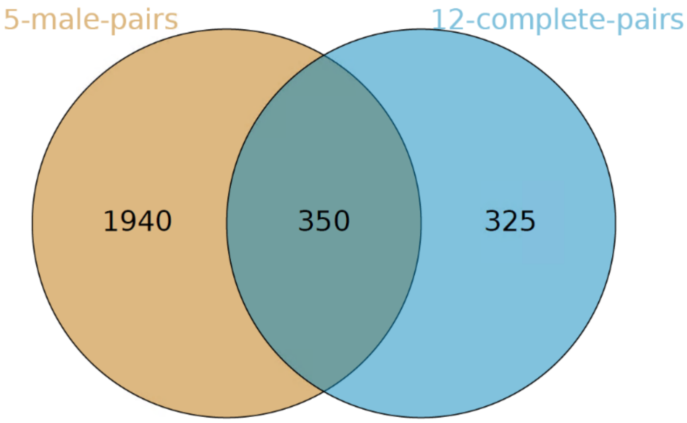


Figure 7-60 | Intersection of differential accessibility analyses performed between Treg from 5 T1D and gender- and sibling- matched healthy subjects (Figure 7-59), and Treg from all 12 T1D and sibling-matched healthy subjects (Figure 5-16) (see Appendix supplementary Table for annotated common regions).
The input for this differential accessibility analysis was count values derived from top 100,000 peaks (by peak significance) intersecting TSS, typical enhancers of Treg cells [8], super- and typical- enhancers of $CD4^+$ T cells [7].

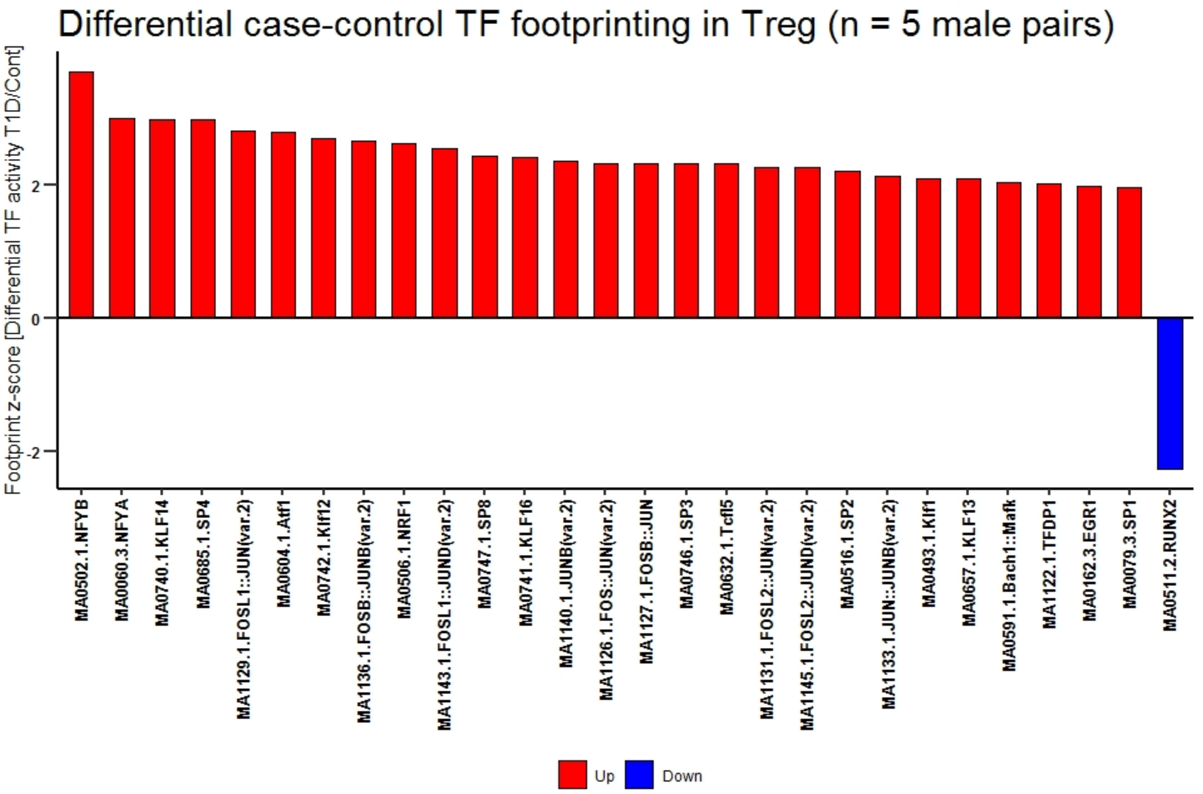


Figure 7-61 | Significantly differential transcription factor (TF) footprinting between <u>gender-matched</u> T1D and healthy <u>Treg</u> cells.

Alterations in TF activity dynamics in T1D were identified from differentially accessible loci calculated from 5 T1D and gender- and sibling-matched healthy subjects. Footprints were identified from mapped sequencing reads pooled from 5 T1D and gender- and sibling-matched healthy ATAC-seq libraries and considered for differential footprinting analysis. TFs with a significant change in activity score (p-value < 0.05) were plotted with enrichment z-scores such that the activity levels are comparable across different TFs. *Up, upregulated in T1D group; down, downregulated in T1D group.*

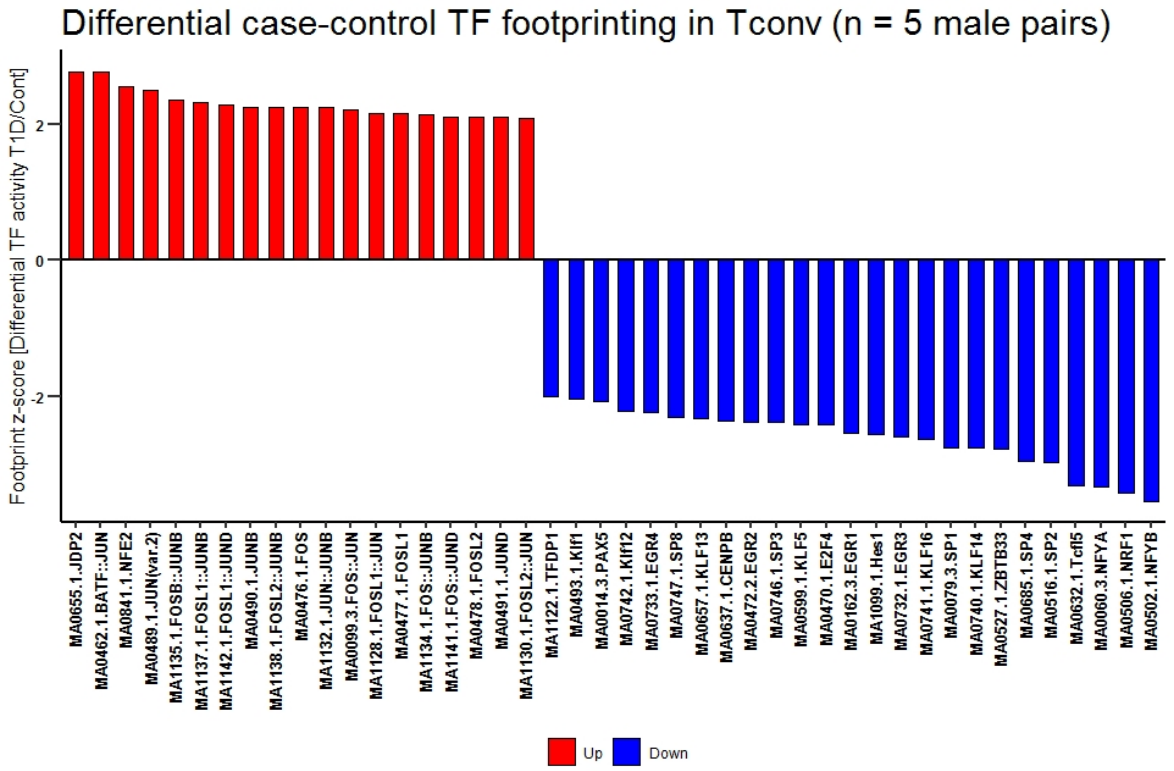


Figure 7-62 | Significantly differential transcription factor (TF) footprinting between <u>gender-matched</u> T1D and healthy <u>Tconv</u> cells.

Alterations in TF activity dynamics in T1D were identified from differentially accessible loci calculated from 5 T1D and gender- and sibling-matched healthy subjects. Footprints were identified from mapped sequencing reads pooled from 5 T1D and gender- and sibling-matched healthy ATAC-seq libraries and considered for differential footprinting analysis. TFs with a significant change in activity score (p-value < 0.05) were plotted with enrichment z-scores such that the activity levels are comparable across different TFs. *Up, upregulated in T1D group; down, downregulated in T1D group.*

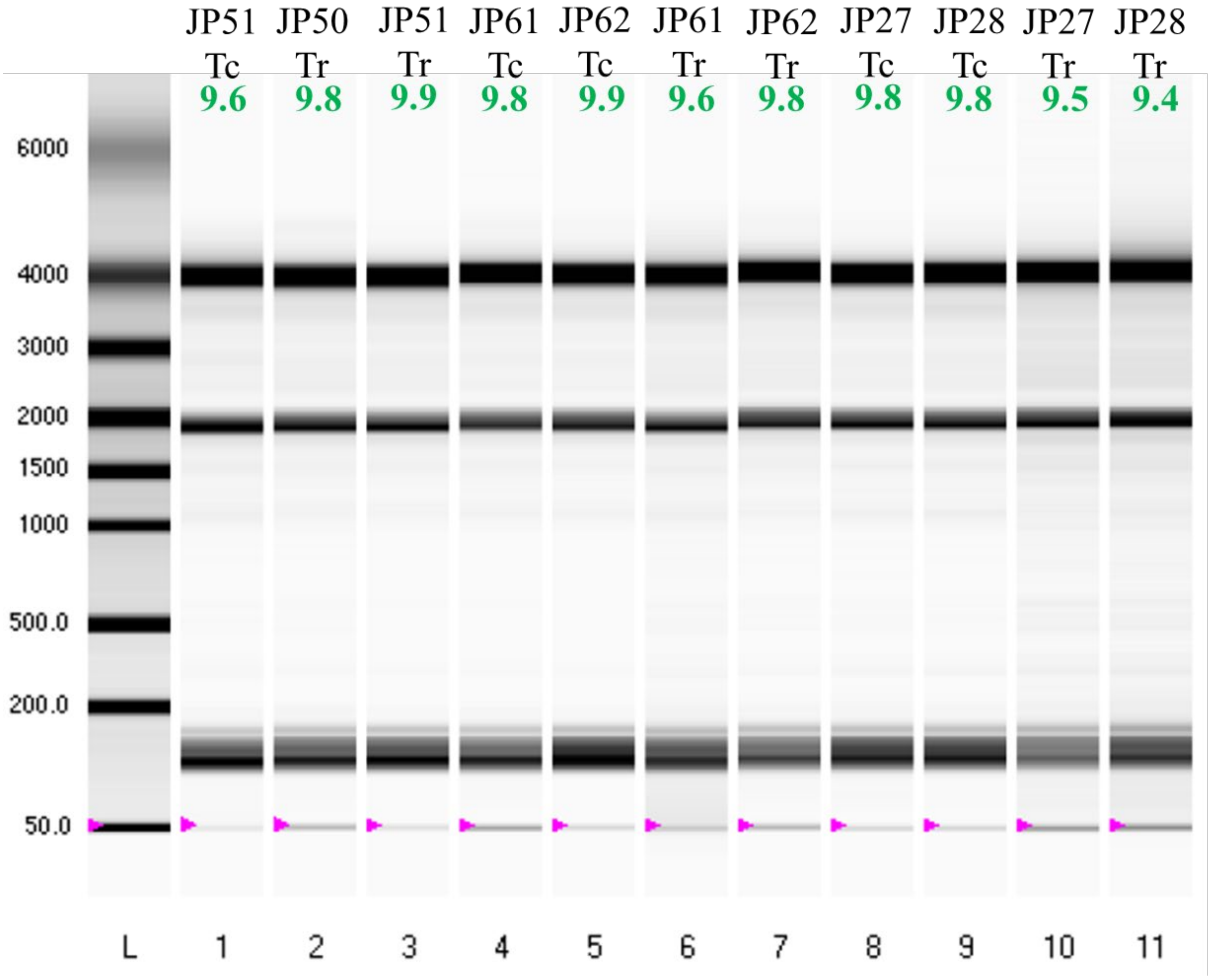


Figure 7-63 | Representative virtual gel showing RNA integrity of RNA recovered from supernatant fractions of ATAC-seq reactions prepared from biobanked T1D and healthy control samples.

The evaluation of RNA integrity was performed on an Experion Automated Electrophoresis System (RNA HighSens Analysis) and the RQI (RNA quality indicator) scores are indicated in green.

Tr, Treg; Tc, Tconv.

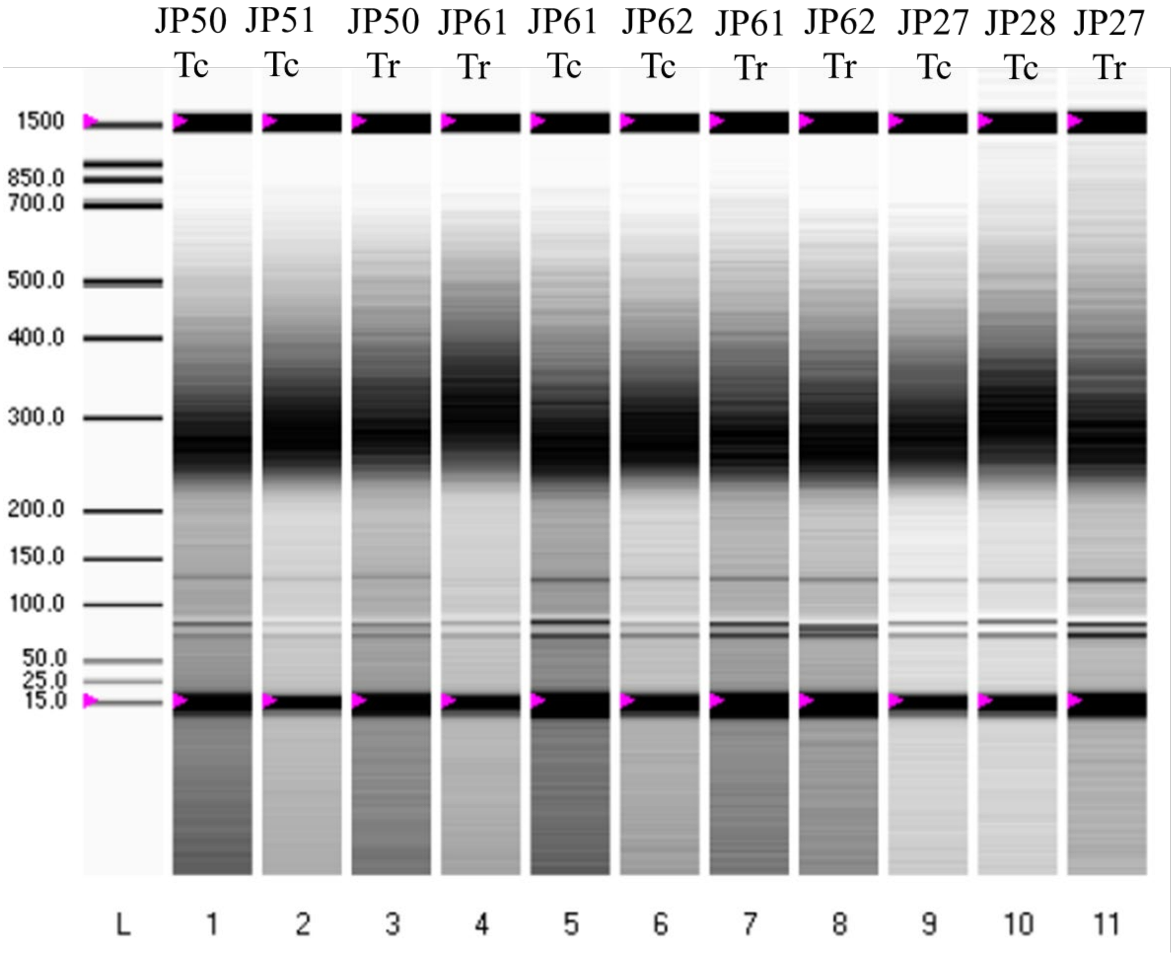


Figure 7-64 | Representative virtual gel showing the fragment size distribution of RNA-seq libraries prepared from biobanked T1D and healthy controls.
RNA seq libraries were constructed from RNA samples recovered and purified from ATAC seq supernatant fractions. The assessment of library QC was performed on an Experion Automated Electrophoresis System (DNA 1K Analysis).

Tr, Treg; Tc, Tconv.

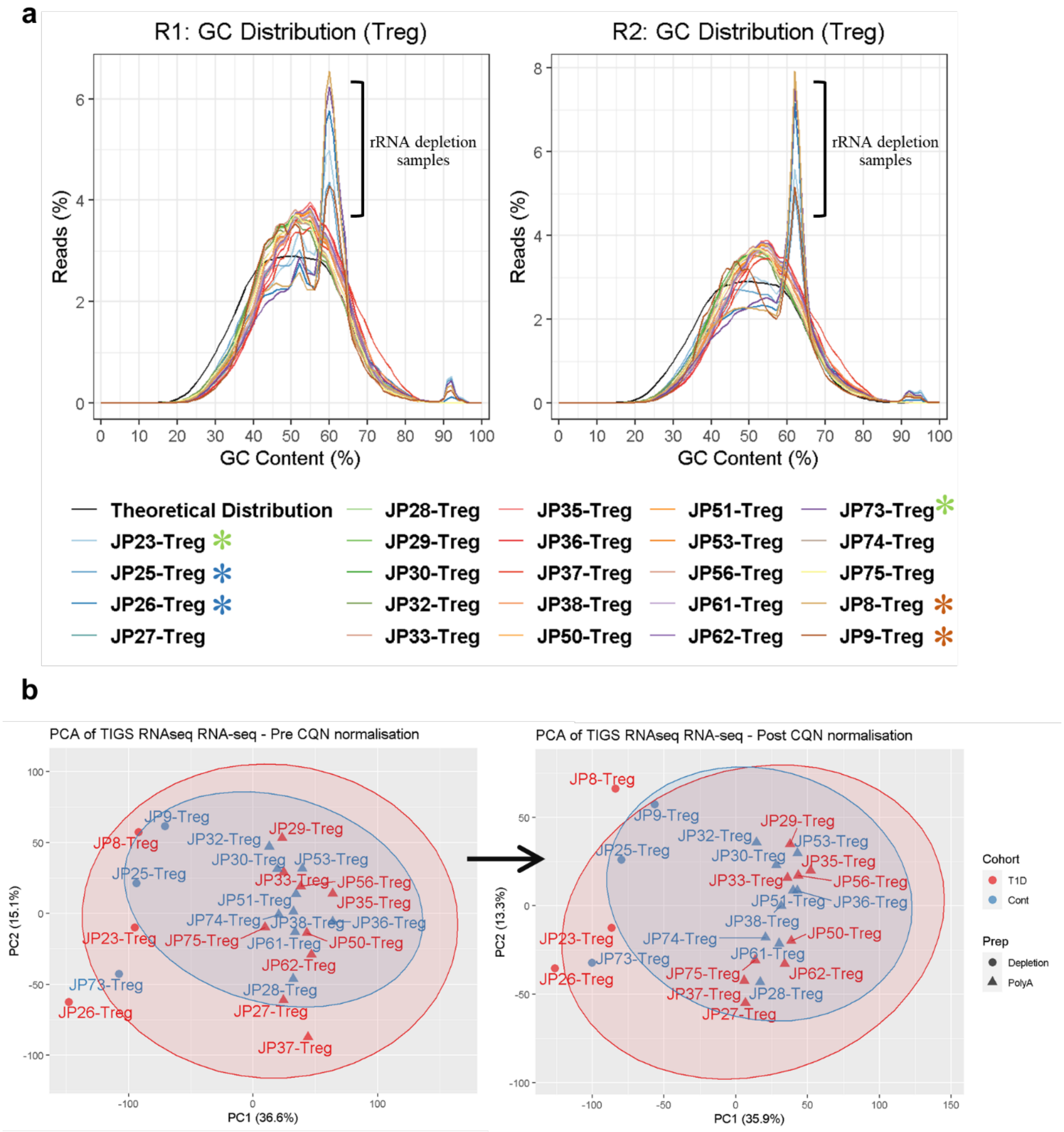


Figure 7-65 | (a) GC content of rRNA-depleted RNA samples shows a high spike in GC content at approximately 60-65%. (b) PCA plots showing distribution of RNA-seq samples before and after correcting for GC content bias using CQN (Conditional Quantile Normalization)[40].

Paired samples are indicated in coloured asterisks (for a).

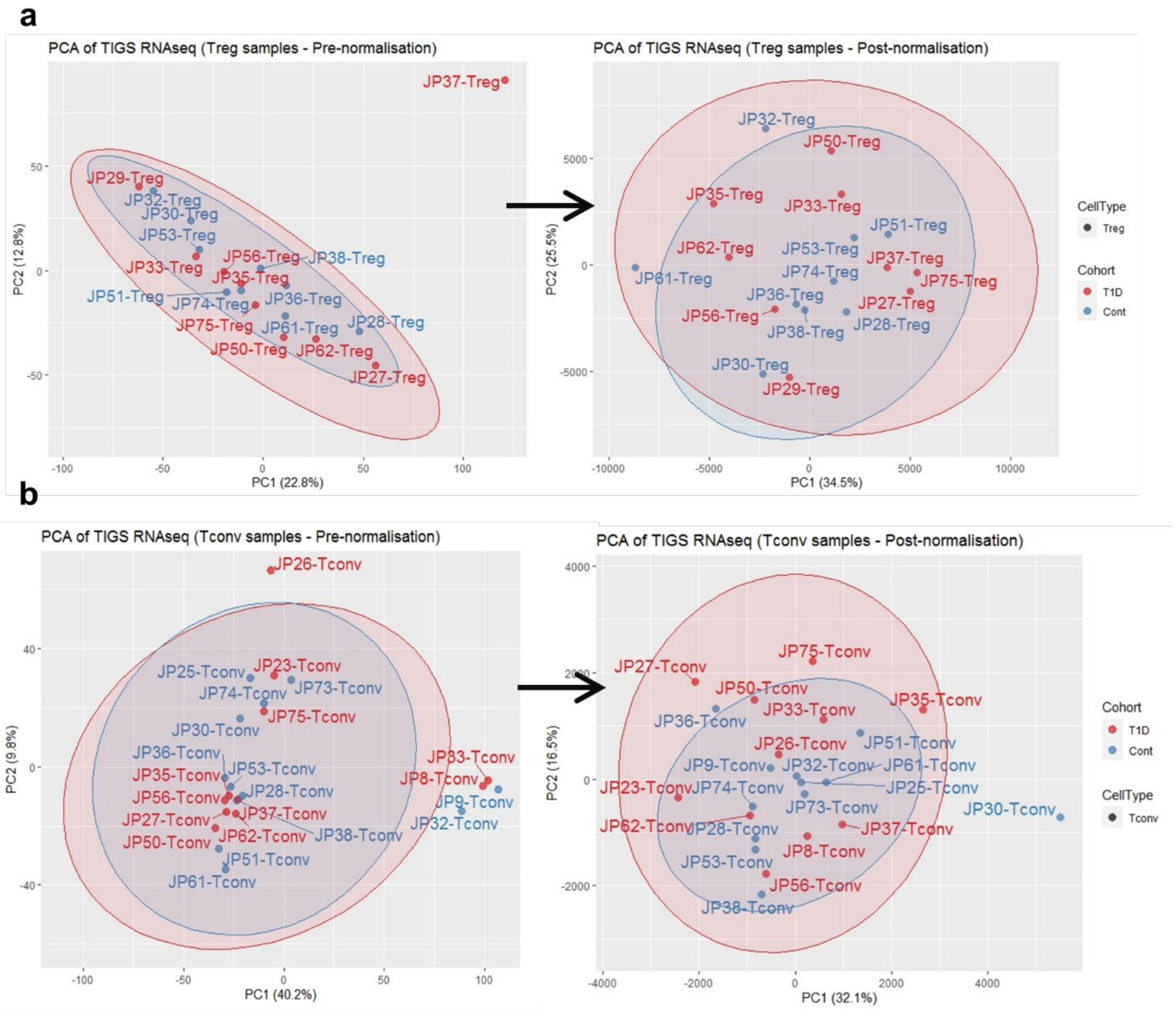


Figure 7-66 | To mitigate the problems of unwanted variation such as batch effects and library preparation, RUV (Remove Unwanted Variation) [32] was performed to adjust for nuisance technical effects by performing factor analysis using a set of *in silico* empirical control genes for Treg (a) and Tconv (b) RNA-seq samples.
Left panel shows sample distribution before normalisation whereas right panel shows sample distribution after normalisation.

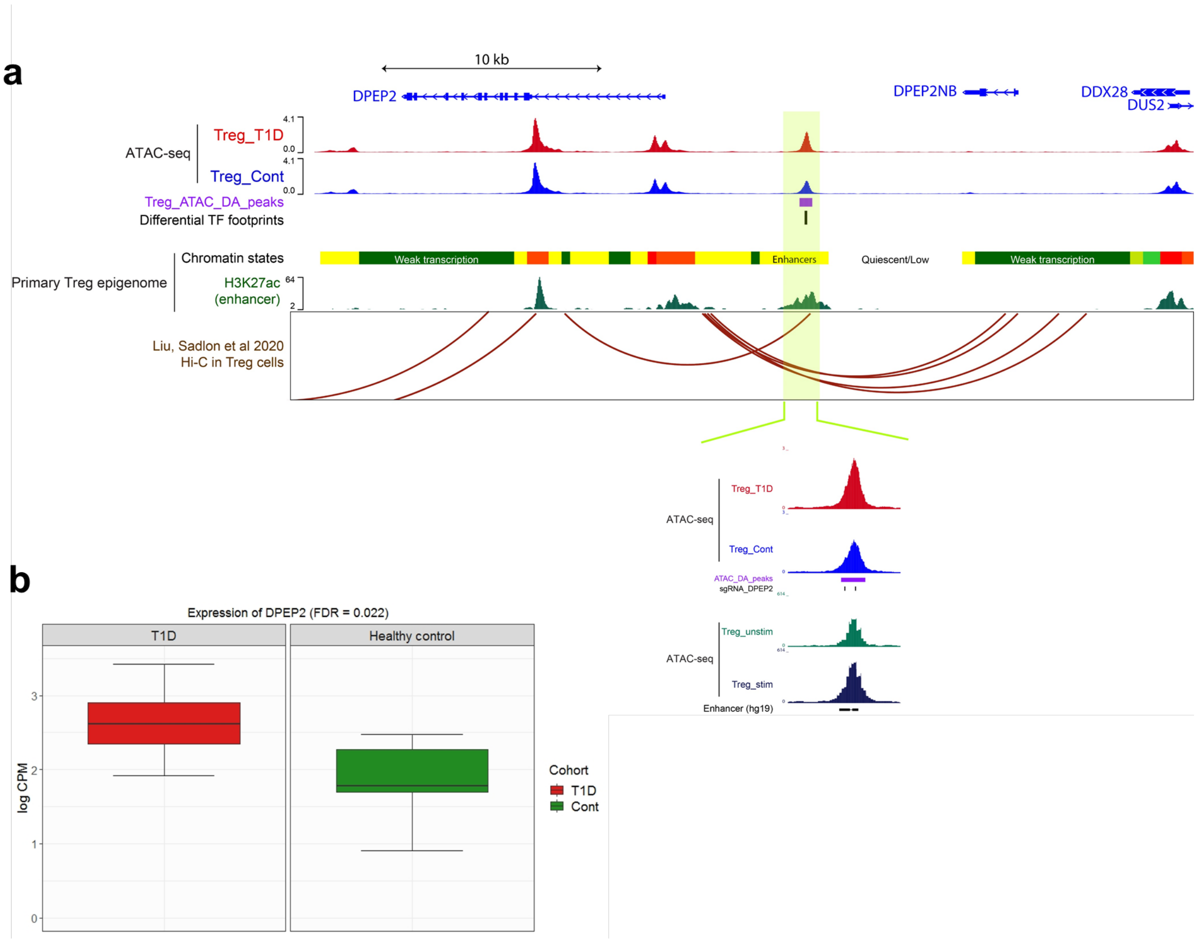


Table 7-4 | Integration of 3D enhancer connectome from Hi-C contact maps links altered enhancer in T1D to *DPEP2* in Treg.
An ATAC-seq peak (highlighted in yellow) located within the T cell enhancer region was differentially accessible between T1D and healthy Treg cells (n=12). 3D connectivity map from Hi-C showed that the differential region is contacting *DPEP2* locus. The observation was confirmed by case-control RNA-seq (b) where *DPEP2* was shown to be differentially expressed (FDR = 0.02) between T1D and healthy controls (n=9).

Case-control ATAC-seq signal was intersected with differential TF footprints, T cell epigenome and chromatin states (Roadmap Epigenomics Project [8]). Bottom panel; zoom-in of the enhancer region. Differentially accessible (DA) peaks identified in the T1D vs healthy control cohort and stimulation-dependent changes in the chromatin accessibility of this region in Treg cells from healthy adult donors.

Browser view was generated using WashU Epigenome Browser.

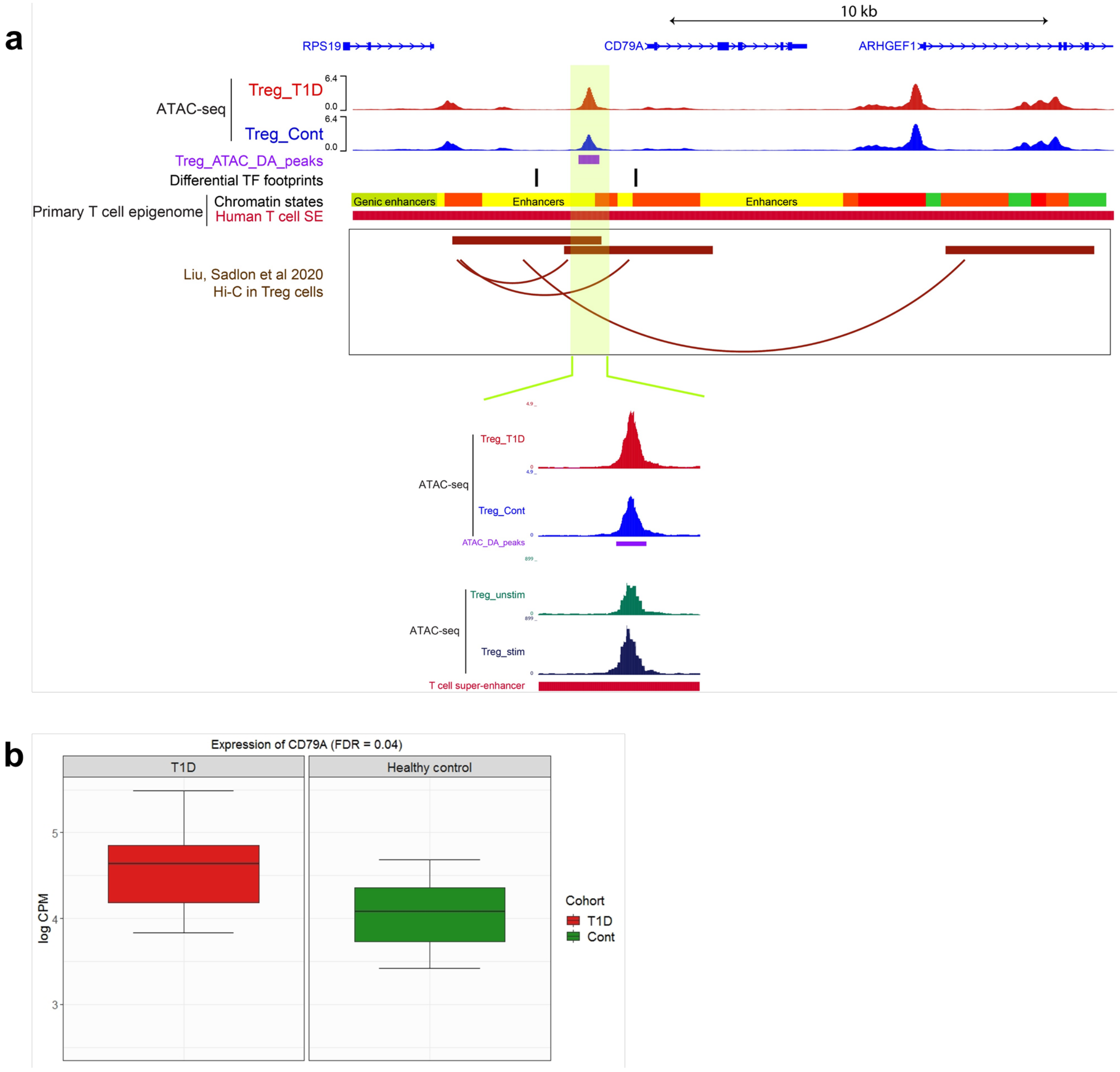


Table 7-5 | Integration of 3D enhancer connectome from Hi-C contact maps links altered enhancer in T1D to *CD79A* in Treg.

An ATAC-seq peak (highlighted in yellow) located within the T cell enhancer was differentially accessible between T1D and healthy Treg cells (n=12). 3D connectivity map from Hi-C showed that the differential region is contacting *CD79A* locus. The observation was confirmed by case-control RNA-seq (b) where *CD79A* was shown to be differentially expressed (FDR = 0.04) between T1D and healthy controls (n=9).

Case-control ATAC-seq signal was intersected with differential TF footprints, T cell super enhancers [7], T cell epigenome and chromatin states (Roadmap Epigenomics Project [8]). Bottom panel; zoom-in of the enhancer region. Differentially accessible (DA) peaks identified in the T1D vs healthy control cohort and stimulation-dependent changes in the chromatin accessibility of this region in Treg cells from healthy adult donors.

Browser view was generated using WashU Epigenome Browser.

DA, differential accessibility.

Table 7-6 | Oligo designs for generation of Omni ATAC-seq libraries.
A list of oligos used for PCR in the ATAC-seq experiment, as adapted from Buenrostro, Giresi [25].

Ad1_noMX:	AATGATACGGCGACCACCGAGATCTACACTCGTCGGCAGCGTCAGATGTG
Ad2.1_TAAGGCGA	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGT
Ad2.2_CGTACTAG	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGGAGATGT
Ad2.3_AGGCAGAA	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGGAGATGT
Ad2.4_TCCTGAGC	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGGAGATGT
Ad2.5_GGACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGGAGATGT
Ad2.6_TAGGCATG	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGGAGATGT
Ad2.7_CTCTCTAC	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGGAGATGT
Ad2.8_CAGAGAGG	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGGAGATGT
Ad2.9_GCTACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGGAGATGT
Ad2.10_CGAGGCTG	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGGAGATGT
Ad2.11_AAGAGGCA	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGGAGATGT
Ad2.12_GTAGAGGA	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGGAGATGT
Ad2.13_GTCGTGAT	CAAGCAGAAGACGGCATACGAGATATCACGACGTCTCGTGGGCTCGGAGATGT
Ad2.14_ACCACTGT	CAAGCAGAAGACGGCATACGAGATACAGTGGTGTCTCGTGGGCTCGGAGATGT
Ad2.15_TGGATCTG	CAAGCAGAAGACGGCATACGAGATCAGATCCAGTCTCGTGGGCTCGGAGATGT
Ad2.16_CCGTTTGT	CAAGCAGAAGACGGCATACGAGATACAAACGGGTCTCGTGGGCTCGGAGATGT
Ad2.17_TGCTGGGT	CAAGCAGAAGACGGCATACGAGATACCCAGCAGTCTCGTGGGCTCGGAGATGT
Ad2.18_GAGGGGTT	CAAGCAGAAGACGGCATACGAGATAACCCCTCGTCTCGTGGGCTCGGAGATGT
Ad2.19_AGGTTGGG	CAAGCAGAAGACGGCATACGAGATCCCAACCTGTCTCGTGGGCTCGGAGATGT
Ad2.20_GTGTGGTG	CAAGCAGAAGACGGCATACGAGATCACCACACGTCTCGTGGGCTCGGAGATGT
Ad2.21_TGGGTTTC	CAAGCAGAAGACGGCATACGAGATGAAACCCAGTCTCGTGGGCTCGGAGATGT
Ad2.22_TGGTCACA	CAAGCAGAAGACGGCATACGAGATTGTGACCAGTCTCGTGGGCTCGGAGATGT
Ad2.23_TTGACCCT	CAAGCAGAAGACGGCATACGAGATAGGGTCAAGTCTCGTGGGCTCGGAGATGT
Ad2.24_CCACTCCT	CAAGCAGAAGACGGCATACGAGATAGGAGTGGGTCTCGTGGGCTCGGAGATGT

Table 7-7 | Differentially expressed genes from Treg cells between T1D and healthy control samples.
Data shown are representative of 9 pairs of T1D and sibling-matched healthy control samples.

gene_id	gene_name	logFC	FDR
ENSG00000182511	FES	-4.22776	2.82E-11
ENSG00000211891	IGHE	-4.97419	3.44E-09
ENSG00000205702	CYP2D7	-4.73002	1.64E-08
ENSG00000003147	ICA1	1.137032	5.01E-08
ENSG00000248746	ACTN3	-0.73534	6.13E-08
ENSG00000222724	RNU2-63P	-5.95049	1.10E-07
ENSG00000163888	CAMK2N2	-1.83395	1.33E-07
ENSG00000275302	CCL4	-1.79822	1.33E-07
ENSG00000088367	EPB41L1	-4.01129	1.56E-07
ENSG00000112812	PRSS16	-2.03115	5.76E-07
ENSG00000074416	MGLL	-4.43087	7.54E-07
ENSG00000119121	TRPM6	-5.74241	8.70E-07
ENSG00000154165	GPR15	2.508431	2.52E-06
ENSG00000275454	AC105020.6	-2.69304	5.84E-06
ENSG00000189134	NKAPL	-2.39317	6.26E-06
ENSG00000105352	CEACAM4	1.730004	8.85E-06
ENSG00000165959	CLMN	-3.77953	1.08E-05
ENSG00000269220	LINC00528	1.763716	1.31E-05
ENSG00000275185	AC130324.3	-4.22271	1.32E-05
ENSG00000175793	SFN	-2.38358	2.07E-05
ENSG00000237693	IRGM	-2.28643	2.25E-05
ENSG00000177340	AC024940.1	-2.88563	3.95E-05
ENSG00000223459	TCAF1P1	-3.93094	5.67E-05
ENSG00000181381	DDX60L	1.034324	5.78E-05
ENSG00000113520	IL4	-4.85094	5.78E-05
ENSG00000103723	AP3B2	-3.94637	8.75E-05
ENSG00000142798	HSPG2	-2.43812	8.75E-05
ENSG00000187187	ZNF546	2.553018	8.75E-05
ENSG00000169242	EFNA1	2.164781	0.0001
ENSG00000171817	ZNF540	-2.23748	0.000174
ENSG00000196639	HRH1	-6.1825	0.000197
ENSG00000185272	RBM11	-2.36914	0.000198
ENSG00000273669	AC015819.1	1.411934	0.000198
ENSG00000154874	CCDC144B	-1.97474	0.000216
ENSG00000167914	GSDMA	1.729772	0.000223
ENSG00000234814	SVIL2P	-2.3946	0.000225
ENSG00000208028	MIR616	-1.02277	0.000225
ENSG00000197557	TTC30A	-2.14992	0.000267
ENSG00000166341	DCHS1	1.620864	0.000289
ENSG00000275719	AC008622.2	2.534644	0.0003
ENSG00000227145	IL21-AS1	-2.56164	0.00034
ENSG00000211857	TRAJ32	2.688661	0.00034

ENSG00000024422	EHD2	-3.19983	0.00034
ENSG00000105270	CLIP3	1.384753	0.00034
ENSG00000090339	ICAM1	-0.83899	0.000373
ENSG00000158089	GALNT14	3.402392	0.000373
ENSG00000254750	CASP1P2	-1.88067	0.000386
ENSG00000170345	FOS	-0.80542	0.000439
ENSG00000260920	AL031985.3	-1.73574	0.000515
ENSG00000123870	ZNF137P	-1.6933	0.000515
ENSG00000167772	ANGPTL4	-5.24883	0.000538
ENSG00000128266	GNAZ	-1.14127	0.00056
ENSG00000116675	DNAJC6	-1.92998	0.000633
ENSG00000273004	AL078644.2	3.981359	0.000897
ENSG00000204959	ARHGEF34P	-1.62809	0.000897
ENSG00000120738	EGR1	-0.82357	0.0009
ENSG00000221946	FXYD7	1.526816	0.001102
ENSG00000117016	RIMS3	-0.99086	0.001118
ENSG00000177374	HIC1	0.770163	0.001155
ENSG00000263482	ANTXRLP1	1.741056	0.001275
ENSG00000129749	CHRNA10	-1.81072	0.001328
ENSG00000090932	DLL3	1.519371	0.00136
ENSG00000114670	NEK11	-1.51294	0.001468
ENSG00000275265	AC127002.1	-5.5293	0.001469
ENSG00000102934	PLLP	2.850907	0.001505
ENSG00000172380	GNG12	-2.18072	0.001505
ENSG00000188215	DCUN1D3	-0.97846	0.001548
ENSG00000170689	HOXB9	-1.96052	0.00155
ENSG00000211710	TRBV4-1	1.277177	0.001616
ENSG00000140939	AC074143.1	-1.62739	0.001616
ENSG00000261644	AC007728.2	-1.45227	0.00166
ENSG00000110811	P3H3	-1.93414	0.001661
ENSG00000140678	ITGAX	-3.40793	0.00179
ENSG00000135074	ADAM19	-0.96313	0.001853
ENSG00000178573	MAF	-0.64595	0.001921
ENSG00000119715	ESRRB	3.794355	0.001921
ENSG00000276557	TRBV18	0.851443	0.001921
ENSG00000175564	UCP3	-1.61445	0.002162
ENSG00000099974	DDTL	-1.82607	0.002167
ENSG00000042832	TG	-1.49924	0.002472
ENSG00000170365	SMAD1	-1.52011	0.002549
ENSG00000184500	PROS1	-4.03503	0.002549
ENSG00000231672	DIRC3	-2.88709	0.002549
ENSG00000282164	PEG13	-1.78163	0.002699
ENSG00000271204	AC016831.5	-2.46392	0.002701
ENSG00000258526	AL049828.1	2.119645	0.002726
ENSG00000137441	FGFBP2	2.017333	0.002902
ENSG00000089692	LAG3	-0.78652	0.00294

ENSG00000183426	NPIPA1	1.514929	0.003028
ENSG00000273445	AC133644.2	0.717648	0.003059
ENSG00000137731	FXYD2	1.363702	0.003143
ENSG00000211720	TRBV11-1	1.134342	0.003143
ENSG00000170049	KCNAB3	1.554546	0.003143
ENSG00000155016	CYP2U1	1.182994	0.003143
ENSG00000225975	LINC01534	1.450697	0.003376
ENSG00000277258	PCGF2	-1.68894	0.003376
ENSG00000172824	CES4A	-0.80237	0.003602
ENSG00000197191	CYSRT1	2.375981	0.003602
ENSG00000146205	ANO7	-1.08317	0.003842
ENSG00000118308	LRMP	0.643627	0.003874
ENSG00000162522	KIAA1522	-1.49649	0.003915
ENSG00000051128	HOMER3	-2.62233	0.003915
ENSG00000122224	LY9	0.901493	0.003942
ENSG00000284602	AL031432.4	2.887961	0.003988
ENSG00000123243	ITIH5	1.618191	0.004305
ENSG00000260778	AC009065.4	-1.41243	0.004366
ENSG00000155980	KIF5A	-2.18496	0.004366
ENSG00000185920	PTCH1	-1.45446	0.004366
ENSG00000140564	FURIN	-0.59598	0.004596
ENSG00000178502	KLHL11	-1.58572	0.00468
ENSG00000280187	AC022107.1	-1.51803	0.00468
ENSG00000185022	MAFF	-0.79399	0.004856
ENSG00000196739	COL27A1	-1.40378	0.004902
ENSG00000286162	AL354744.1	1.712746	0.005054
ENSG00000105750	ZNF85	1.373909	0.005376
ENSG00000135898	GPR55	0.774237	0.005414
ENSG00000277602	AC005363.2	1.774878	0.005581
ENSG00000205413	SAMD9	1.04664	0.005654
ENSG00000274752	TRBV12-3	1.069072	0.005917
ENSG00000182621	PLCB1	-3.02934	0.005945
ENSG00000172889	EGFL7	-1.10319	0.006026
ENSG00000128040	SPINK2	1.755818	0.006026
ENSG00000235576	LINC01871	-1.4633	0.006026
ENSG00000284713	SMIM38	-2.0012	0.006026
ENSG00000263655	AC090125.1	1.53889	0.006097
ENSG00000279161	AC093503.3	1.138123	0.006279
ENSG00000151640	DPYSL4	-1.05815	0.006505
ENSG00000237989	LINC01679	-1.23784	0.006613
ENSG00000179397	CATSPERE	-3.04581	0.006613
ENSG00000258181	AC008083.2	1.186493	0.006906
ENSG00000108176	DNAJC12	1.578888	0.006926
ENSG00000166866	MYO1A	-1.62557	0.006926
ENSG00000273702	AC091271.1	-1.28351	0.006926
ENSG00000229692	SOS1-IT1	-2.79704	0.00706

ENSG00000211777	TRAV3	0.806184	0.007096
ENSG00000160460	SPTBN4	-1.41456	0.007145
ENSG00000205336	ADGRG1	-1.28892	0.007257
ENSG00000074706	IPCEF1	0.615668	0.007366
ENSG00000231789	PIK3CD-AS2	0.722729	0.007467
ENSG00000176945	MUC20	-2.15765	0.0078
ENSG00000128271	ADORA2A	1.341199	0.0078
ENSG00000149591	TAGLN	-1.42385	0.007825
ENSG00000188818	ZDHHC11	-2.75478	0.00783
ENSG00000175592	FOSL1	-0.63424	0.007969
ENSG00000277632	CCL3	-1.24792	0.007969
ENSG00000268575	AL031282.2	-1.05157	0.007969
ENSG00000085733	CTTN	-0.79422	0.008186
ENSG00000102878	HSF4	-3.16988	0.008227
ENSG00000272084	AL137127.1	-2.80537	0.008376
ENSG00000249741	AC093890.1	-2.86442	0.008483
ENSG00000169258	GPRIN1	-0.95378	0.008483
ENSG00000285399	AC104162.2	-1.43748	0.008483
ENSG00000157680	DGKI	1.528995	0.00851
ENSG00000140057	AK7	-1.50065	0.009158
ENSG00000164877	MICALL2	-1.27999	0.009199
ENSG00000081377	CDC14B	-0.96585	0.009199
ENSG00000253842	AP003472.2	-3.59747	0.009387
ENSG00000120659	TNFSF11	-1.30225	0.009387
ENSG00000180747	SMG1P3	1.532876	0.009442
ENSG00000215784	FAM72D	0.85262	0.009457
ENSG00000211818	TRAV39	-0.8731	0.009587
ENSG00000169213	RAB3B	-1.3562	0.009587
ENSG00000258366	RTEL1	-1.2171	0.009587
ENSG00000168917	SLC35G2	-1.2614	0.009631
ENSG00000164008	C1orf50	-1.03153	0.010881
ENSG00000228544	CCDC183-AS1	2.515273	0.011565
ENSG00000100311	PDGFB	-1.56581	0.011625
ENSG00000075426	FOSL2	-0.64222	0.011625
ENSG00000178440	LINC00843	2.562385	0.011625
ENSG00000124256	ZBP1	1.256477	0.01181
ENSG00000117013	KCNQ4	-3.51433	0.01181
ENSG00000056736	IL17RB	-3.14915	0.011985
ENSG00000173114	LRRN3	0.754431	0.011985
ENSG00000125618	PAX8	-2.21083	0.011985
ENSG00000163564	PYHIN1	0.618264	0.012112
ENSG00000176681	LRRC37A	1.817078	0.012149
ENSG00000285825	AP003501.3	-1.82417	0.012472
ENSG00000233016	SNHG7	0.528552	0.012472
ENSG00000142530	FAM71E1	-1.71765	0.013094
ENSG00000134278	SPIRE1	-1.34207	0.013481

ENSG00000214725	CDIPTOSP	-2.37635	0.014021
ENSG00000256576	LINC02361	0.723936	0.014021
ENSG00000113790	EHHADH	-1.14668	0.014021
ENSG00000237296	SMG1P1	-1.18581	0.014222
ENSG00000207005	RNU1-2	-2.55179	0.014317
ENSG00000139835	GRTP1	2.040783	0.01586
ENSG00000151715	TMEM45B	3.618608	0.016237
ENSG00000122970	IFT81	-2.41362	0.01635
ENSG00000110777	POU2AF1	-0.81795	0.01635
ENSG00000147889	CDKN2A	-0.7992	0.01635
ENSG00000128578	STRIP2	-1.13432	0.01635
ENSG00000229162	AL445471.1	-1.22373	0.016395
ENSG00000080854	IGSF9B	-0.93946	0.016395
ENSG00000222009	BTBD19	-1.09174	0.016395
ENSG00000228878	SEPT7-AS1	1.058973	0.016395
ENSG00000262580	AC087741.1	1.292936	0.016499
ENSG00000226806	AC011893.1	0.988029	0.016535
ENSG00000262712	AC012676.1	-1.74939	0.016535
ENSG00000182310	SPACA6	1.079855	0.016571
ENSG00000100027	YPEL1	1.549227	0.016625
ENSG00000108641	B9D1	0.943464	0.016625
ENSG00000283537	AC073264.3	-2.84214	0.016625
ENSG00000106484	MEST	0.73132	0.016625
ENSG00000267481	AC011477.2	-1.68177	0.016649
ENSG00000163141	BNIPL	-1.63343	0.016734
ENSG00000008277	ADAM22	2.475416	0.016755
ENSG00000267254	AC020928.1	1.381902	0.016764
ENSG00000128641	MYO1B	-1.07658	0.016764
ENSG00000119508	NR4A3	-0.59044	0.017181
ENSG00000196189	SEMA4A	-1.07197	0.017183
ENSG00000279082	LINC01727	1.160404	0.017506
ENSG00000130304	SLC27A1	-0.78348	0.017506
ENSG00000159761	C16orf86	1.184214	0.017506
ENSG00000122877	EGR2	-0.68761	0.017506
ENSG00000175505	CLCF1	-1.42104	0.017506
ENSG00000153982	GDPD1	1.030772	0.017506
ENSG00000231150	AL034345.2	1.289038	0.017506
ENSG00000157429	ZNF19	-1.46601	0.017812
ENSG00000254996	ANKHD1-EIF4EBP3	-1.43681	0.018177
ENSG00000273373	AL355488.1	2.2392	0.018559
ENSG00000260852	FBXL19-AS1	-1.23012	0.018658
ENSG00000258732	AC025884.1	2.105644	0.019057
ENSG00000189223	PAX8-AS1	-2.05347	0.019057
ENSG00000141294	LRRC46	1.241669	0.019175
ENSG00000145287	PLAC8	0.6896	0.019175
ENSG00000152582	SPEF2	-2.65489	0.019175

ENSG00000072195	SPEG	1.198633	0.019557
ENSG00000137101	CD72	-0.80508	0.019837
ENSG00000105137	SYDE1	-1.78758	0.01988
ENSG00000260404	AC110079.1	-1.23872	0.020061
ENSG00000275632	AL035461.2	-1.25815	0.020093
ENSG00000163235	TGFA	-1.81779	0.020355
ENSG00000106100	NOD1	0.652695	0.020394
ENSG00000279673	AC092919.2	1.677876	0.020394
ENSG00000177721	ANXA2R	0.652079	0.020394
ENSG00000188763	FZD9	1.600768	0.020817
ENSG00000285761	AL645939.5	-1.28121	0.020817
ENSG00000125744	RTN2	-0.6624	0.020817
ENSG00000112164	GLP1R	1.102523	0.020879
ENSG00000167261	DPEP2	0.795177	0.021923
ENSG00000237840	FAM21FP	-1.71215	0.022939
ENSG00000144115	THNSL2	1.128204	0.022939
ENSG00000146072	TNFRSF21	-1.36479	0.023094
ENSG00000249592	AC139887.2	1.032299	0.023638
ENSG00000076864	RAP1GAP	-1.50755	0.023638
ENSG00000062282	DGAT2	-1.41936	0.023741
ENSG00000279696	AP001273.1	1.797638	0.025076
ENSG00000104497	SNX16	-0.98798	0.02509
ENSG00000143167	GPA33	0.794619	0.02509
ENSG00000279088	AC022400.8	0.800765	0.02509
ENSG00000279602	AC109326.1	-1.78361	0.02509
ENSG00000276476	AL136962.1	-1.26702	0.025208
ENSG00000136378	ADAMTS7	2.217182	0.02521
ENSG00000107159	CA9	1.298979	0.025241
ENSG00000165140	FBP1	0.680453	0.025241
ENSG00000206561	COLQ	-1.4233	0.025847
ENSG00000170074	FAM153A	-1.53553	0.026213
ENSG00000114270	COL7A1	2.355631	0.026213
ENSG00000270574	AC010680.2	-2.30998	0.026874
ENSG00000259952	AC009133.2	-1.39916	0.027033
ENSG00000230615	AL139220.2	-1.50849	0.027033
ENSG00000184441	AP001062.1	-1.75658	0.027128
ENSG00000270175	AC023509.3	1.468424	0.027165
ENSG00000232859	LYRM9	0.956182	0.027388
ENSG00000143502	SUSD4	0.826599	0.027388
ENSG00000280417	AC096887.2	1.619571	0.027592
ENSG00000075643	MOCOS	-1.6052	0.027675
ENSG00000163251	FZD5	-1.57445	0.027699
ENSG00000226660	TRBV2	0.556297	0.027752
ENSG00000134369	NAV1	-1.14328	0.027752
ENSG00000196730	DAPK1	-1.67736	0.027996
ENSG00000197124	ZNF682	-0.9703	0.027996

ENSG00000160117	ANKLE1	-1.09602	0.028032
ENSG00000213025	COX20P1	-1.4281	0.028184
ENSG00000138613	APH1B	0.85638	0.028451
ENSG00000066279	ASPM	1.207522	0.028451
ENSG00000255046	AC069185.1	-1.2171	0.028465
ENSG00000196421	C20orf204	-2.58762	0.02922
ENSG00000226200	SGMS1-AS1	-1.24056	0.029299
ENSG00000283554	LINC02341	-0.79789	0.029331
ENSG00000169992	NLGN2	-0.81038	0.029641
ENSG00000268362	AC092279.1	1.027483	0.029752
ENSG00000137809	ITGA11	-2.07541	0.029908
ENSG00000154153	RETREG1	0.50482	0.030498
ENSG00000164542	KIAA0895	-1.44502	0.030781
ENSG00000172349	IL16	0.43304	0.031024
ENSG00000189042	ZNF567	1.151573	0.031109
ENSG00000215417	MIR17HG	-1.48439	0.031646
ENSG00000281357	ARRDC3-AS1	-1.63236	0.031836
ENSG00000131018	SYNE1	0.485134	0.031836
ENSG00000184731	FAM110C	-2.73853	0.032218
ENSG00000197497	ZNF665	0.866444	0.033189
ENSG00000181847	TIGIT	0.449713	0.033226
ENSG00000266066	POLRMTP1	1.108384	0.033965
ENSG00000278993	AC002350.1	-1.33817	0.034226
ENSG00000231584	FAHD2CP	0.647177	0.034404
ENSG00000197885	NKIRAS1	-0.99608	0.034587
ENSG00000070371	CLTCL1	-1.35853	0.035139
ENSG00000127946	HIP1	-0.67172	0.035646
ENSG00000117091	CD48	0.478711	0.03581
ENSG00000160654	CD3G	0.435707	0.036032
ENSG00000257906	LINC02156	-0.97307	0.036184
ENSG00000267519	AC020916.1	-0.70142	0.036884
ENSG00000255569	TRAV1-1	0.839783	0.036949
ENSG00000272667	AC012306.2	1.585347	0.037139
ENSG00000160867	FGFR4	-2.34452	0.037681
ENSG00000175866	BAIAP2	-0.81528	0.038016
ENSG00000112799	LY86	2.106853	0.038824
ENSG00000261766	AC133550.2	1.032506	0.03964
ENSG00000161609	CCDC155	-2.61543	0.03964
ENSG00000198133	TMEM229B	-1.30292	0.03964
ENSG00000008256	CYTH3	0.641265	0.039857
ENSG00000255435	AP001267.3	1.111866	0.039857
ENSG00000112655	PTK7	1.53435	0.040163
ENSG00000174945	AMZ1	-2.03508	0.040163
ENSG00000165238	WNK2	-1.60822	0.040163
ENSG00000227591	HSD11B1-AS1	-1.51663	0.040318
ENSG00000277117	FP565260.3	2.647886	0.040532

ENSG00000171236	LRG1	-1.1332	0.040698
ENSG00000105369	CD79A	0.553308	0.040698
ENSG00000196689	TRPV1	0.988668	0.041257
ENSG00000203711	C6orf99	-1.7675	0.041769
ENSG00000204161	TMEM273	-0.96921	0.04279
ENSG00000185189	NRBP2	-0.84344	0.04279
ENSG00000186235	LINC02610	2.049323	0.043167
ENSG00000229474	PATL2	0.571899	0.043458
ENSG00000216895	AC009403.1	0.984666	0.043477
ENSG00000113525	IL5	-2.49322	0.043585
ENSG00000128590	DNAJB9	-0.54893	0.043678
ENSG00000266947	AC022916.1	0.68914	0.044157
ENSG00000130489	SCO2	-1.29375	0.044288
ENSG00000129473	BCL2L2	1.504463	0.044288
ENSG00000213707	HMGB1P10	0.905144	0.044542
ENSG00000270816	LINC00221	-1.33999	0.044542
ENSG00000136643	RPS6KC1	-0.55844	0.044562
ENSG00000152465	NMT2	0.446176	0.044622
ENSG00000138061	CYP1B1	-1.25279	0.044622
ENSG00000205309	NT5M	0.676413	0.044816
ENSG00000131400	NAPSA	-0.56904	0.047204
ENSG00000185482	STAC3	0.861994	0.047371
ENSG00000196466	ZNF799	-0.80534	0.047384
ENSG00000275791	TRBV10-3	1.078683	0.048481
ENSG00000204389	HSPA1A	-0.57933	0.048481
ENSG00000137135	ARHGEF39	0.812709	0.048481
ENSG00000197134	ZNF257	-1.14057	0.048481
ENSG00000261071	AL441883.1	1.634207	0.04934
ENSG00000273604	EPOP	-0.43942	0.04934
ENSG00000173404	INSM1	-1.50423	0.04934
ENSG00000274471	AC242376.2	-1.56539	0.04934
ENSG00000166073	GPR176	-2.09069	0.049512
ENSG00000106701	FSD1L	-1.26882	0.049719
ENSG00000105982	RNF32	3.140771	0.049806

Table 7-8 | Differentially expressed genes from Tconv cells between T1D and healthy control samples.
Data shown are representative of 11 pairs of T1D and sibling-matched healthy control samples.

gene_id	gene_name	logFC	FDR
ENSG00000226278	PSPHP1	-7.45026	5.25E-35
ENSG00000213058	AL365357.1	4.872129	6.28E-34
ENSG00000274276	CBSL	2.710144	2.19E-12
ENSG00000066248	NGEF	-5.4981	4.62E-10
ENSG00000108556	CHRNE	1.844008	6.04E-10
ENSG00000157510	AFAP1L1	1.679174	8.07E-10
ENSG00000100097	LGALS1	-0.58102	1.78E-07
ENSG00000151208	DLG5	-1.32018	1.47E-06
ENSG00000143153	ATP1B1	-1.07963	4.61E-06
ENSG00000188848	BEND4	1.699438	1.15E-05
ENSG00000157985	AGAP1	-1.4156	1.53E-05
ENSG00000285077	AC091057.6	2.198222	2.22E-05
ENSG00000276832	AL354718.3	2.395747	2.44E-05
ENSG00000160223	ICOSLG	-3.55234	2.44E-05
ENSG00000176945	MUC20	-3.21447	4.36E-05
ENSG00000261324	AC010168.2	-1.14759	4.36E-05
ENSG00000184347	SLIT3	-2.39747	5.57E-05
ENSG00000173404	INSM1	-0.69496	7.24E-05
ENSG00000146021	KLHL3	-0.78767	7.69E-05
ENSG00000174672	BRSK2	-1.19994	8.70E-05
ENSG00000231793	DOC2GP	1.483776	8.73E-05
ENSG00000185736	ADARB2	-2.0197	8.73E-05
ENSG00000144115	THNSL2	0.944482	9.26E-05
ENSG00000179862	CITED4	0.638708	9.26E-05
ENSG00000136842	TMOD1	-0.923	0.0001422
ENSG00000142549	IGLON5	-1.78664	0.00017274
ENSG00000281741	AC241377.3	-0.63476	0.00018159
ENSG00000115523	GNLY	-1.1034	0.00022637
ENSG00000129521	EGLN3	-0.82417	0.00022637
ENSG00000087250	MT3	-1.02492	0.00022637
ENSG00000275302	CCL4	-0.6317	0.00022637
ENSG00000132437	DDC	1.453621	0.00024469
ENSG00000222032	AC112721.2	-2.73838	0.000296
ENSG00000160345	C9orf116	1.097709	0.00029901
ENSG00000105339	DENND3	-0.65145	0.00033367
ENSG00000006327	TNFRSF12A	-0.72994	0.00038801
ENSG00000067141	NEO1	-0.81323	0.00039583
ENSG00000160791	CCR5	3.519749	0.00041266
ENSG00000154309	DISP1	-1.02035	0.00042224
ENSG00000198502	HLA-DRB5	-1.20152	0.00042224
ENSG00000166796	LDHC	-2.0585	0.00042875
ENSG00000071282	LMCD1	0.862521	0.00043049

ENSG00000115641	FHL2	0.630719	0.00045218
ENSG00000107968	MAP3K8	-0.50859	0.00051626
ENSG00000163827	LRRC2	-3.02207	0.00055217
ENSG00000226435	ANKRD18DP	-1.24736	0.00057749
ENSG00000285668	AC126544.2	2.024229	0.00057749
ENSG00000276070	CCL4L2	1.280927	0.00065598
ENSG00000104951	IL4I1	-0.59831	0.00068229
ENSG00000106853	PTGR1	-0.72936	0.00068347
ENSG00000114405	C3orf14	-0.56426	0.00070436
ENSG00000146205	ANO7	-0.86305	0.00076324
ENSG00000143127	ITGA10	-1.0798	0.00076886
ENSG00000271737	AC008608.2	1.058345	0.0007898
ENSG00000168994	PXDC1	-2.38133	0.0007898
ENSG00000007866	TEAD3	-0.66809	0.0007898
ENSG00000117013	KCNQ4	-1.18084	0.00079712
ENSG00000204172	AGAP9	-0.76127	0.0008068
ENSG00000235034	C19orf81	-0.89971	0.0008068
ENSG00000167772	ANGPTL4	-0.94548	0.0008068
ENSG00000095932	SMIM24	-1.11701	0.00085084
ENSG00000179397	CATSPERE	3.303855	0.00085084
ENSG00000185499	MUC1	-0.76635	0.00087821
ENSG00000136205	TNS3	-1.0309	0.00095818
ENSG00000273080	AC009309.1	-0.87056	0.00102932
ENSG00000011566	MAP4K3	-0.76477	0.00106357
ENSG00000104723	TUSC3	0.913172	0.00116066
ENSG00000277149	TYW1B	0.793732	0.00116066
ENSG00000170074	FAM153A	-1.06523	0.0011698
ENSG00000102878	HSF4	-1.07729	0.0012313
ENSG00000262050	AC005696.1	0.879006	0.00129884
ENSG00000134326	CMPK2	0.544987	0.00138653
ENSG00000228863	AL121985.1	-1.3424	0.00164824
ENSG00000104043	ATP8B4	-1.25015	0.00171297
ENSG00000164292	RHOBTB3	0.620273	0.00193337
ENSG00000184731	FAM110C	-2.69679	0.00193337
ENSG00000211935	IGHV1-3	-0.63619	0.00193337
ENSG00000116016	EPAS1	-0.60844	0.00225055
ENSG00000135960	EDAR	2.029881	0.00225055
ENSG00000175183	CSRP2	-0.93178	0.0023818
ENSG00000158859	ADAMTS4	-1.16816	0.00246377
ENSG00000237254	TRBV30	1.912392	0.00247099
ENSG00000188368	PRR19	0.620239	0.00247099
ENSG00000164904	ALDH7A1	-0.69852	0.00247099
ENSG00000125510	OPRL1	1.260105	0.00251825
ENSG00000228162	AC097713.1	1.471741	0.00252626
ENSG00000107736	CDH23	-0.78794	0.00254539
ENSG00000058091	CDK14	3.149398	0.0026237

ENSG00000092098	RNF31	1.12647	0.0026237
ENSG00000124126	PREX1	-0.75315	0.0026237
ENSG00000215068	AC025171.2	1.924137	0.0026237
ENSG00000185065	AC000068.1	2.073638	0.0026237
ENSG00000204934	ATP6V0E2-AS1	0.657863	0.00270558
ENSG00000284882	AL359762.1	-1.07714	0.00272332
ENSG00000111537	IFNG	-0.80889	0.00272651
ENSG00000186918	ZNF395	-0.42405	0.00275791
ENSG00000107104	KANK1	0.590033	0.00306305
ENSG00000173156	RHOD	-0.89073	0.00306305
ENSG00000178053	MLF1	-0.8263	0.00306305
ENSG00000141294	LRRC46	1.037271	0.00306305
ENSG00000143365	RORC	-0.74807	0.00306305
ENSG00000010310	GIPR	-1.19784	0.00310254
ENSG00000177181	RIMKLA	-0.59065	0.00310254
ENSG00000169131	ZNF354A	0.752202	0.00345105
ENSG00000255605	AP000820.1	-0.92857	0.00348013
ENSG00000272599	AC016394.2	-0.92454	0.00349365
ENSG00000270175	AC023509.3	0.863306	0.00351078
ENSG00000111962	UST	-0.94854	0.00352662
ENSG00000180881	CAPS2	0.987658	0.00366817
ENSG00000261071	AL441883.1	1.006397	0.00383838
ENSG00000157653	C9orf43	-0.92708	0.00383838
ENSG00000141665	FBXO15	0.600458	0.00383838
ENSG00000277632	CCL3	-0.56982	0.00391002
ENSG00000227388	AL133410.1	1.235152	0.00405141
ENSG00000275832	ARHGAP23	-1.51316	0.00409218
ENSG00000145730	PAM	-0.43952	0.00409218
ENSG00000169242	EFNA1	1.121219	0.00409218
ENSG00000163082	SGPP2	-0.49481	0.00411588
ENSG00000114268	PFKFB4	-0.41805	0.00411588
ENSG00000071909	MYO3B	1.00295	0.00411588
ENSG00000026559	KCNG1	-0.83675	0.00411588
ENSG00000105523	FAM83E	-1.10347	0.00411588
ENSG00000274922	AL139384.1	-0.96656	0.00413176
ENSG00000168765	GSTM4	-0.71829	0.00413176
ENSG00000171236	LRG1	-2.71144	0.004205
ENSG00000271361	HTATSF1P2	-1.22709	0.00422951
ENSG00000214425	LRRC37A4P	-1.47068	0.00496164
ENSG00000238121	LINC00426	-0.81754	0.00496164
ENSG00000213988	ZNF90	-0.70691	0.00527662
ENSG00000209042	SNORD12C	1.261944	0.00532488
ENSG00000185022	MAFF	-0.57116	0.00545431
ENSG00000143847	PPFIA4	-0.58527	0.00577673
ENSG00000187045	TMPRSS6	-0.58768	0.00577673
ENSG00000255153	TOLLIP-AS1	0.783191	0.00585987

ENSG00000064201	TSPAN32	-1.14992	0.00601983
ENSG00000080200	CRYBG3	-1.09809	0.00610797
ENSG00000128536	CDHR3	1.180642	0.00622664
ENSG00000282164	PEG13	-2.12565	0.00645371
ENSG00000114450	GNB4	0.787088	0.00666147
ENSG00000245017	LINC02453	-1.2252	0.00669212
ENSG00000152076	CCDC74B	-1.23561	0.00689584
ENSG00000274070	CASTOR2	-0.88955	0.00692158
ENSG00000158457	TSPAN33	0.59438	0.00714899
ENSG00000112116	IL17F	-1.01303	0.00714899
ENSG00000174705	SH3PXD2B	0.832137	0.00714899
ENSG00000074660	SCARF1	-0.73725	0.00714899
ENSG00000116014	KISS1R	-0.44259	0.00718195
ENSG00000162069	BICDL2	-1.36656	0.00739081
ENSG00000104324	CPQ	-0.82525	0.00742695
ENSG00000215252	GOLGA8B	-0.4746	0.00742695
ENSG00000100311	PDGFB	-0.83752	0.00742695
ENSG00000270441	AC135506.1	-1.22841	0.00769298
ENSG00000101439	CST3	0.797517	0.00769923
ENSG00000049192	ADAMTS6	1.44053	0.0078399
ENSG00000234420	ZNF37BP	0.717005	0.0078399
ENSG00000162600	OMA1	0.525122	0.00804866
ENSG00000141526	SLC16A3	-0.50843	0.0081065
ENSG00000213433	RPLP1P6	0.928284	0.0081065
ENSG00000186088	GSAP	0.794796	0.0081065
ENSG00000148411	NACC2	-0.46038	0.0081065
ENSG00000263482	ANTXRLP1	1.574937	0.008125
ENSG00000113739	STC2	-0.54395	0.00818199
ENSG00000054793	ATP9A	-1.00939	0.00844238
ENSG00000114270	COL7A1	-1.48142	0.00877129
ENSG00000100299	ARSA	0.531621	0.00877129
ENSG00000103811	CTSH	-0.46371	0.00877129
ENSG00000026652	AGPAT4	-0.61831	0.00882654
ENSG00000162415	ZSWIM5	-0.69142	0.00887199
ENSG00000276903	HIST1H2AL	1.791515	0.00887199
ENSG00000267317	AC027307.2	-0.75558	0.00887199
ENSG00000125629	INSIG2	-0.52003	0.00897659
ENSG00000272221	AL645933.2	1.503005	0.00900274
ENSG00000105205	CLC	-0.66901	0.009034
ENSG00000237550	RPL9P9	0.464728	0.009034
ENSG00000145736	GTF2H2	-0.58806	0.00933711
ENSG00000278390	AL354696.2	2.491986	0.00994135
ENSG00000272977	AL008721.2	-0.84345	0.00999703
ENSG00000226510	UPK1A-AS1	-0.82505	0.01021352
ENSG00000205309	NT5M	0.488397	0.01021352
ENSG00000261377	PDCD6IPP2	0.607328	0.01037003

ENSG00000149527	PLCH2	-1.08498	0.01057083
ENSG00000258704	SRP54-AS1	-0.63732	0.01064131
ENSG00000275457	AL117332.1	0.595551	0.01074588
ENSG00000135074	ADAM19	-0.48056	0.01084952
ENSG00000221886	ZBED8	0.934837	0.0108569
ENSG00000151474	FRMD4A	0.545675	0.01104172
ENSG00000142089	IFITM3	-0.41976	0.01112262
ENSG00000230002	ALMS1-IT1	0.814106	0.01112262
ENSG00000163141	BNIPL	-1.53246	0.01130165
ENSG00000228696	ARL17B	-1.04274	0.01130165
ENSG00000196576	PLXNB2	-0.9291	0.01130165
ENSG00000116690	PRG4	-1.11499	0.01130165
ENSG00000197852	INKA2	0.619652	0.01141228
ENSG00000273802	HIST1H2BG	-0.76931	0.01162336
ENSG00000166866	MYO1A	-1.32391	0.01178104
ENSG00000128039	SRD5A3	-0.45958	0.0117945
ENSG00000128641	MYO1B	-0.40429	0.0121464
ENSG00000160207	HSF2BP	0.939995	0.01260125
ENSG00000236397	DDX11L2	-0.85598	0.01260125
ENSG00000158023	WDR66	-1.05087	0.01307834
ENSG00000279170	TSTD3	0.61938	0.01310178
ENSG00000118985	ELL2	-0.37361	0.01315068
ENSG00000134955	SLC37A2	-2.31454	0.01320355
ENSG00000173825	TIGD3	-0.67477	0.01335282
ENSG00000250251	PKD1P6	-0.69077	0.01339951
ENSG00000264885	AC026271.3	1.008492	0.01366765
ENSG00000160201	U2AF1	0.732866	0.01374698
ENSG00000163359	COL6A3	-0.5552	0.01374698
ENSG00000179431	FJX1	0.947091	0.01374698
ENSG00000284602	AL031432.4	1.341166	0.01374698
ENSG00000140961	OSGIN1	0.683667	0.01378108
ENSG00000168961	LGALS9	0.377129	0.014596
ENSG00000161653	NAGS	-0.93318	0.01502828
ENSG00000156804	FBXO32	-0.49455	0.01508737
ENSG00000060558	GNA15	-0.42726	0.01558808
ENSG00000271964	AC090948.1	-0.99953	0.01597334
ENSG00000130294	KIF1A	-0.65182	0.01673783
ENSG00000099139	PCSK5	0.749191	0.01708213
ENSG00000140939	AC074143.1	-0.45694	0.01708213
ENSG00000109471	IL2	-0.77856	0.01732036
ENSG00000184378	ACTRT3	-0.61353	0.01732036
ENSG00000113645	WWC1	1.015071	0.01732425
ENSG00000136830	FAM129B	-0.46644	0.0173821
ENSG00000198142	SOWAHC	-0.96864	0.01755256
ENSG00000099377	HSD3B7	-0.70037	0.01815343
ENSG00000163697	APBB2	-0.86548	0.01837608

ENSG00000257086	AP001453.4	0.721071	0.01837608
ENSG00000106336	FBXO24	0.854241	0.01837961
ENSG00000183856	IQGAP3	-0.47086	0.01837961
ENSG00000162433	AK4	-0.37729	0.01888325
ENSG00000196189	SEMA4A	-0.6561	0.01898866
ENSG00000135362	PRR5L	-0.5226	0.01926106
ENSG00000186642	PDE2A	0.883439	0.01932617
ENSG00000111275	ALDH2	-0.49116	0.01991014
ENSG00000272150	NBPF25P	-1.62026	0.01991014
ENSG00000157483	MYO1E	-0.61322	0.01995827
ENSG00000172159	FRMD3	1.411591	0.02050686
ENSG00000213214	ARHGEF35	-0.82081	0.02061532
ENSG00000254531	AP001816.1	-0.87427	0.02080812
ENSG00000196639	HRH1	-0.81392	0.02132528
ENSG00000171695	LKAAEAR1	-1.0986	0.02138933
ENSG00000250539	KRT8P33	1.480856	0.02229478
ENSG00000157992	KRTCAP3	0.783899	0.0223612
ENSG00000144214	LYG1	1.492945	0.0223612
ENSG00000157680	DGKI	-0.74206	0.0223612
ENSG00000187554	TLR5	-2.19806	0.02281735
ENSG00000206140	TMEM191C	-1.10055	0.0234184
ENSG00000225151	GOLGA2P7	0.979723	0.02342014
ENSG00000277117	FP565260.3	1.736022	0.0237122
ENSG00000225828	FAM229A	-1.10127	0.02378099
ENSG00000228168	HNRNPA1P21	-1.35166	0.02380041
ENSG00000131981	LGALS3	-0.48185	0.0238276
ENSG00000136866	ZFP37	-1.36183	0.02388405
ENSG00000285331	AC090517.5	1.295056	0.0240207
ENSG00000109943	CRTAM	1.572551	0.0240207
ENSG00000166780	C16orf45	-0.45178	0.0240207
ENSG00000214900	LINC01588	-0.5062	0.02413279
ENSG00000196422	PPP1R26	-1.13294	0.02474781
ENSG00000255569	TRAV1-1	0.922168	0.02479652
ENSG00000100504	PYGL	-1.01683	0.02479652
ENSG00000205336	ADGRG1	-0.60987	0.02552416
ENSG00000197905	TEAD4	-0.57333	0.02552416
ENSG00000123689	G0S2	-0.47045	0.02552416
ENSG00000140564	FURIN	-0.36888	0.026776
ENSG00000182704	TSKU	-0.56004	0.02684996
ENSG00000171126	KCNG3	1.002577	0.02710576
ENSG00000178573	MAF	-0.50543	0.02743317
ENSG00000064989	CALCRL	1.158209	0.02750556
ENSG00000235532	LINC00402	3.586579	0.02755696
ENSG00000120903	CHRNA2	0.897536	0.02755696
ENSG00000276141	WHAMMP3	-0.92239	0.02755696
ENSG00000172548	NIPAL4	-0.84638	0.02845767

ENSG00000225978	HAR1A		0.565013	0.02881439
ENSG00000114480	GBE1		-0.38113	0.02882183
ENSG00000163449	TMEM169		0.649649	0.02915296
ENSG00000019144	PHLDB1		-0.47578	0.02966147
ENSG00000249459	ZNF286B		0.650884	0.03019463
ENSG00000105499	PLA2G4C		-0.55746	0.03087785
ENSG00000100578	KIAA0586		0.420613	0.03087785
ENSG00000253570	RNF5P1		0.68697	0.03087785
ENSG00000243646	IL10RB		-0.91008	0.03087785
ENSG00000277072	STAG3L2		0.40276	0.03087785
ENSG00000133466	C1QTNF6		0.561645	0.03148842
ENSG00000064687	ABCA7		-0.44479	0.03232795
ENSG00000186522		Sep-10	-0.7631	0.03375119
ENSG00000178199	ZC3H12D		-0.44353	0.03385014
ENSG00000077935	SMC1B		1.065569	0.03385014
ENSG00000180834	MAP6D1		0.864359	0.03405316
ENSG00000128656	CHN1		-0.57725	0.03405316
ENSG00000224114	AL591846.1		1.95322	0.03405316
ENSG00000158806	NPM2		-0.83301	0.03456949
ENSG00000178665	ZNF713		0.910109	0.03467329
ENSG00000153885	KCTD15		-0.58475	0.03478854
ENSG00000176092	CRYBG2		-0.61205	0.03520324
ENSG00000135519	KCNH3		-0.53486	0.03583361
ENSG00000008283	CYB561		-0.59622	0.03594837
ENSG00000151151	IPMK		-0.38804	0.03672819
ENSG00000232810	TNF		-0.3314	0.03692746
ENSG00000175764	TTLL11		0.483371	0.03701999
ENSG00000166352	C11orf74		0.57358	0.03721997
ENSG00000196782	MAML3		1.686801	0.03763015
ENSG00000115607	IL18RAP		-0.88136	0.03783681
ENSG00000135315	CEP162		-1.59216	0.03826111
ENSG00000205089	CCNI2		0.450543	0.03832401
ENSG00000158477	CD1A		0.746406	0.03841084
ENSG00000247400	DNAJC3-DT		1.187338	0.03861374
ENSG00000109586	GALNT7		0.405756	0.03861678
ENSG00000106484	MEST		0.462581	0.03871845
ENSG00000213889	PPM1N		-0.48359	0.03882523
ENSG00000171055	FEZ2		-0.40816	0.0388466
ENSG00000147852	VLDLR		-0.66772	0.03903893
ENSG00000005379	TSPOAP1		-0.58377	0.03903893
ENSG00000285219	AL591485.1		0.636624	0.03903893
ENSG00000226210	WASH8P		-0.76444	0.04001659
ENSG00000252481	SCARNA13		1.302824	0.04100272
ENSG00000174137	FAM53A		0.537446	0.04187378
ENSG00000267169	AC022098.1		-0.6374	0.04636546
ENSG00000166845	C18orf54		0.517878	0.04683039

ENSG00000131941	RHPN2	-0.50497	0.04711322
ENSG00000117139	KDM5B	-0.35807	0.04711322
ENSG00000272463	AL357054.4	-0.70208	0.04747207
ENSG00000150457	LATS2	-0.65051	0.04747207
ENSG00000150687	PRSS23	-0.48121	0.04776336
ENSG00000186235	LINC02610	0.786276	0.04776336
ENSG00000124006	OBSL1	-0.59584	0.04814248
ENSG00000254681	PKD1P5	-0.93191	0.04814248
ENSG00000256628	ZBTB11-AS1	0.501347	0.04814248
ENSG00000140451	PIF1	-0.50149	0.04823697
ENSG00000276368	HIST1H2AJ	1.915651	0.04823697
ENSG00000165617	DACT1	-0.40115	0.04832631
ENSG00000185760	KCNQ5	0.385419	0.0483422
ENSG00000278002	AL627171.1	-0.63638	0.0483422
ENSG00000272654	AL358472.2	0.916733	0.0487273
ENSG00000269825	AC022150.4	1.602542	0.049162
ENSG00000162616	DNAJB4	-0.4546	0.049162
ENSG00000256073	URB1-AS1	-0.42645	0.04964828
ENSG00000178904	DPY19L3	0.432599	0.04964828

Table 7-9. Transcription factor (TF) footprints identified in case-control differentially accessible (DA) enhancer peak regulating TIGIT in Treg cells.
The data shown are representative of ATAC-seq profiles from 12 T1D and sibling-matched healthy control samples.

TF footprints overlapping TIGIT enhancer
MA0642.1.EN2
MA0644.1.ESX1
MA0078.1.Sox17
MA0621.1.mix-a
MA0707.1.MNX1
MA0902.1.HOXB2
MA0903.1.HOXB3
MA0132.2.PDX1
MA0618.1.LBX1
MA0722.1.VAX1
MA0723.1.VAX2
MA0725.1.VSX1
MA0726.1.VSX2
MA0087.1.Sox5
MA0095.2.YY1
MA0866.1.SOX21
MA1152.1.SOX15
MA0087.1.Sox5
MA0084.1.SRY

Table 7-10. Number of identified genotype variants in Treg cells from T1D subjects. 111 T1D GWAS LD-SNPs (87 after removal of overlapping SNPs) were shown to coincide with ATAC-seq accessible chromatin regions and TF footprints differentially regulated between T1D and healthy controls. Genotypes were identified and the number of mismatches between T1D and matched healthy controls at these locations was computed.

Pairing	Control	T1D	Number of genotype variants identified in Treg cells (T1D to matched control)
1	JP73	JP23	8/87
2	JP9	JP8	8/87
3	JP30	JP29	9/87
4	JP32	JP33	9/87
5	JP25	JP26	10/87
6	JP53	JP56	6/87
7	JP36	JP35	9/87
8	JP28	JP27	11/87
9	JP51	JP50	7/87
10	JP38	JP37	12/87
11	JP74	JP75	10/87
12	JP61	JP62	7/87

CHAPTER 8: REFERENCES

References

1. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.

2. Onengut-Gumuscu, S., et al., *Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers.* Nat Genet, 2015. **47**(4): p. 381-6.

3. Buniello, A., et al., *The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.* Nucleic Acids Res, 2019. **47**(D1): p. D1005-d1012.

4. Schep, A.N., et al., *Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions.* Genome research, 2015. **25**(11): p. 1757-1770.

5. Calvo, S.E., K.R. Clauser, and V.K. Mootha, *MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins.* Nucleic Acids Res, 2016. **44**(D1): p. D1251-7.

6. Javierre, B.M., et al., *Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters.* Cell, 2016. **167**(5): p. 1369-1384.e19.

7. Vahedi, G., et al., *Super-enhancers delineate disease-associated regulatory nodes in T cells.* Nature, 2015. **520**(7548): p. 558-62.

8. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-330.

9. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-461.

10. Pociot, F., et al., *Genetics of Type 1 Diabetes: What's Next?* Diabetes, 2010. **59**(7): p. 1561.

11. Sadlon, T.J., et al., *Genome-wide identification of human FOXP3 target genes in natural regulatory T cells.* J Immunol, 2010. **185**(2): p. 1071-81.

12. Inc, P.E., Pearson Addison-Wesley, 2011.

13. Hawai‘i, U.o., *Chapter 8 - Biology of the T Lymphocyte. Microbiology 161-Immunology and Protein Chemistry.* . Honolulu, Hawaii, U.S.

14. Online, I.H., *How does the immune system work?* 2016, Institute for Quality and Efficiency in Health Care (IQWiG): Cologne, Germany.

15. Kubicek, S., *Epigenetics—A Primer*, in *TheScientist*. 2011, LabX Media Group: Ontario, Canada.

16. Jones, J.A.O.J.P.S.A.S.P.P., *KUBY Immunology*. 7th ed. 2013, New York, United States of America: W. H. Freeman and Company.

17. Foundation, C.s.D., *New Type 1 Diabetes Staging Classification Opens Door for Intervention*. 2017 Children's Diabetes Foundation: Colorado, USA.

18. Germain, R.N., *T-cell development and the CD4–CD8 lineage decision.* Nature Reviews Immunology, 2002. **2**(5): p. 309-322.

19. Swain, S.L., K.K. McKinstry, and T.M. Strutt, *Expanding roles for CD4+ T cells in immunity to viruses.* Nat Rev Immunol, 2012. **12**(2): p. 136-148.

20. Daley, T. and A.D. Smith, *Predicting the molecular complexity of sequencing libraries.* Nature methods, 2013. **10**(4): p. 325-327.

21. Li, Z., et al., *Identification of transcription factor binding sites using ATAC-seq.* Genome Biology, 2019. **20**(1): p. 45.

22. Qu, K., et al., *Individuality and Variation of Personal Regulomes in Primary Human T Cells.* Cell Systems. **1**(1): p. 51-61.

23. Wang, Z., et al., *Histone Posttranslational Modifications of CD4(+) T Cell in Autoimmune Diseases.* International Journal of Molecular Sciences, 2016. **17**(10): p. 1547.

24. Katsarou, A., et al., *Type 1 diabetes mellitus.* Nature Reviews Disease Primers, 2017. **3**: p. 17016.

25. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Meth, 2013. **10**(12): p. 1213-1218.

26. Ahmadi, M., et al., *Epigenetic modifications and epigenetic based medication implementations of autoimmune diseases.* Biomedicine & Pharmacotherapy, 2017. **87**: p. 596-608.

27. Ou, J., et al., *ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data.* BMC Genomics, 2018. **19**(1): p. 169.

28. Gel, B., et al., *regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests.* Bioinformatics, 2016. **32**(2): p. 289-291.

29. Khan, A., et al., *JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.* Nucleic Acids Res, 2018. **46**(D1): p. D1284.

30. Scharer, C.D., et al., *ATAC-seq on biobanked specimens defines a unique chromatin accessibility structure in naïve SLE B cells.* Scientific Reports, 2016. **6**: p. 27030.

31. Chi, K.R., *Reveling in the Revealed*, in *The Scientist*. 2016, LabX Media Group: Ontario, Canada.

32. Risso, D., et al., *Normalization of RNA-seq data using factor analysis of control genes or samples.* Nature biotechnology, 2014. **32**(9): p. 896-902.

33. Gao, P., et al., *Risk variants disrupting enhancers of TH1 and TREG cells in type 1 diabetes.* Proceedings of the National Academy of Sciences, 2019. **116**(15): p. 7581.

34. Dranoff, G., *Cytokines in cancer pathogenesis and cancer therapy*, in *Nat Rev Cancer*. 2004, Nature Publishing Group. p. 11-22.

35. Mehers, K.L. and K.M. Gillespie, *The genetic basis for type 1 diabetes.* Br Med Bull, 2008. **88**(1): p. 115-29.

36. Lachmann, A., et al., *ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments.* Bioinformatics, 2010. **26**(19): p. 2438-44.

37. Mumbach, M.R., et al., *Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements.* Nature Genetics, 2017. **49**(11): p. 1602-1612.

38. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics (Oxford, England), 2010. **26**(1): p. 139-140.

39. Lu, G., et al., *GAAD: A Gene and Autoimmiune Disease Association Database.* Genomics, Proteomics & Bioinformatics, 2018. **16**(4): p. 252-261.

40. Hansen, K.D., R.A. Irizarry, and Z. Wu, *Removing technical variability in RNA-seq data using conditional quantile normalization.* Biostatistics, 2012. **13**(2): p. 204-16.

41. Corces, M.R., et al., *An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues.* Nat Meth, 2017. **14**(10): p. 959-962.

42. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

43. Pranzatelli, T.J.F., D.G. Michael, and J.A. Chiorini, *ATAC2GRN: optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference.* BMC Genomics, 2018. **19**(1): p. 563.

44. Welch, R.P., et al., *ChIP-Enrich: gene set enrichment testing for ChIP-seq data.* Nucleic Acids Research, 2014. **42**(13): p. e105-e105.

45. Calderon, D., et al., *Landscape of stimulation-responsive chromatin across diverse human immune cells.* Nature Genetics, 2019. **51**(10): p. 1494-1505.

46. Charles A Janeway, J., Paul Travers, Mark Walport, Mark J Shlomchik, *Immunobiology: The Immune System in Health and Disease*. 5th edition ed. 2001, New York: Garland Science.

47. Nemazee, D., *Receptor Selection in B and T Lymphocytes.* Annual review of immunology, 2000. **18**: p. 10.1146/annurev.immunol.18.1.19.

48. Edwards, K.M., et al., *Anti-viral strategies of cytotoxic T lymphocytes are manifested through a variety of granule-bound pathways of apoptosis induction.* Immunol Cell Biol, 1999. **77**(1): p. 76-89.

49. Milner, J.D., et al., *Impaired T(H)17 cell differentiation in subjects with autosomal dominant hyper-IgE syndrome.* Nature, 2008. **452**(7188): p. 773-6.

50. Crotty, S., *T Follicular Helper Cell Differentiation, Function, and Roles in Disease.* Immunity, 2014. **41**(4): p. 529-542.

51. Adeegbe, D.O. and H. Nishikawa, *Natural and induced T regulatory cells in cancer.* Front Immunol, 2013. **4**: p. 190.

52. Soghoian, D.Z. and H. Streeck, *Cytolytic CD4(+) T cells in viral immunity.* Expert Rev Vaccines, 2010. **9**(12): p. 1453-63.

53. Li, P., et al., *Complex interactions of transcription factors in mediating cytokine biology in T cells.* Immunol Rev, 2014. **261**(1): p. 141-56.

54. Caza, T. and S. Landas, *Functional and Phenotypic Plasticity of CD4+ T Cell Subsets.* BioMed Research International, 2015. **2015**: p. 13.

55. Cretney, E., A. Kallies, and S.L. Nutt, *Differentiation and function of Foxp3 effector regulatory T cells*, in *Trends in Immunology*. 2012, Elsevier. p. 74-80.

56. Grzanka, J., et al., *FoxP3, Helios, and SATB1: roles and relationships in regulatory T cells.* Int Immunopharmacol, 2013. **16**(3): p. 343-7.

57. Fu, W., et al., *A multiply redundant genetic switch 'locks in' the transcriptional signature of regulatory T cells.* Nat Immunol, 2012. **13**(10): p. 972-80.

58. Mills, K.H.G., *Regulatory T cells: friend or foe in immunity to infection?*, in *Nat Rev Immunol*. 2004. p. 841-855.

59. Wang, P. and S.G. Zheng, *Regulatory T cells and B cells: implication on autoimmune diseases.* International Journal of Clinical and Experimental Pathology, 2013. **6**(12): p. 2668-2674.

60. Vignali, D.A.A., L.W. Collison, and C.J. Workman, *How regulatory T cells work.* Nature reviews. Immunology, 2008. **8**(7): p. 523-532.

61. Lan, Q., et al., *Induced Foxp3+ regulatory T cells: a potential new weapon to treat autoimmune and inflammatory diseases?* Journal of Molecular Cell Biology, 2012. **4**(1): p. 22-28.

62. Peterson, R.A., *Regulatory T-cells: diverse phenotypes integral to immune homeostasis and suppression.* Toxicol Pathol, 2012. **40**(2): p. 186-204.

63. Campbell, D.J. and S.F. Ziegler, *FOXP3 modifies the phenotypic and functional properties of regulatory T cells.* Nat Rev Immunol, 2007. **7**(4): p. 305-310.

64. Boehm, F., et al., *Deletion of Foxp3+ regulatory T cells in genetically targeted mice supports development of intestinal inflammation.* BMC Gastroenterol, 2012. **12**: p. 97.

65. Kim, J.M., J.P. Rasmussen, and A.Y. Rudensky, *Regulatory T cells prevent catastrophic autoimmunity throughout the lifespan of mice.* Nat Immunol, 2007. **8**(2): p. 191-7.

66. van Belle, T.L., K.T. Coppieters, and M.G. von Herrath, *Type 1 diabetes: etiology, immunology, and therapeutic strategies.* Physiol Rev, 2011. **91**(1): p. 79-118.

67. Atkinson, M.A., G.S. Eisenbarth, and A.W. Michels, *Type 1 diabetes.* Lancet, 2014. **383**(9911): p. 69-82.

68. Borchers, A.T., R. Uibo, and M.E. Gershwin, *The geoepidemiology of type 1 diabetes.* Autoimmun Rev, 2010. **9**(5): p. A355-65.

69. Pociot, F. and A. Lernmark, *Genetic risk factors for type 1 diabetes.* Lancet, 2016. **387**(10035): p. 2331-9.

70. Insel, R.A., et al., *Staging presymptomatic type 1 diabetes: a scientific statement of JDRF, the Endocrine Society, and the American Diabetes Association.* Diabetes Care, 2015. **38**(10): p. 1964-74.

71. American Diabetes, A., *Diagnosis and Classification of Diabetes Mellitus.* Diabetes Care, 2009. **32**(Suppl 1): p. S62-S67.

72. Katsarou, A., et al., *Type 1 diabetes mellitus.* Nat Rev Dis Primers, 2017. **3**: p. 17016.

73. Rønningen, K.S., J.M. Norris, and M. Knip, *Environmental Trigger(s) of Type 1 Diabetes: Why Is It So Difficult to Identify?* BioMed Research International, 2015. **2015**: p. 847906.

74. Knip, M. and O. Simell, *Environmental Triggers of Type 1 Diabetes.* Cold Spring Harbor Perspectives in Medicine, 2012. **2**(7): p. a007690.

75. Noble, J.A., et al., *HLA class I and genetic susceptibility to type 1 diabetes: results from the Type 1 Diabetes Genetics Consortium.* Diabetes, 2010. **59**(11): p. 2972-9.

76. Stene, L.C., et al., *Enterovirus infection and progression from islet autoimmunity to type 1 diabetes: the Diabetes and Autoimmunity Study in the Young (DAISY).* Diabetes, 2010. **59**(12): p. 3174-80.

77. Tracy, S., et al., *Toward testing the hypothesis that group B coxsackieviruses (CVB) trigger insulin-dependent diabetes: inoculating nonobese diabetic mice with CVB markedly lowers diabetes incidence.* J Virol, 2002. **76**(23): p. 12097-111.

78. Frohnert, B.I., et al., *Late-onset islet autoimmunity in childhood: the Diabetes Autoimmunity Study in the Young (DAISY).* Diabetologia, 2017.

79. Knip, M., et al., *Prediction of type 1 diabetes in the general population.* Diabetes Care, 2010. **33**(6): p. 1206-12.

80. Ziegler, A.G., et al., *Seroconversion to multiple islet autoantibodies and risk of progression to diabetes in children.* Jama, 2013. **309**(23): p. 2473-9.

81. Tandon, N., *Understanding type 1 diabetes through genetics: Advances and prospects.* Indian Journal of Endocrinology and Metabolism, 2015. **19**(Suppl 1): p. S39-S43.

82. Yu, L., Z. Zhao, and A.K. Steck, *T1D Autoantibodies: room for improvement?* Current opinion in endocrinology, diabetes, and obesity, 2017. **24**(4): p. 285-291.

83. Krischer, J.P. and G. the Type 1 Diabetes TrialNet Study, *The use of intermediate endpoints in the design of type 1 diabetes prevention trials.* Diabetologia, 2013. **56**(9): p. 1919-1924.

84. Erlich, H., et al., *HLA DR-DQ Haplotypes and Genotypes and Type 1 Diabetes Risk: Analysis of the Type 1 Diabetes Genetics Consortium Families.* Diabetes, 2008. **57**(4): p. 1084-1092.

85. Barrett, J.C., et al., *Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes.* Nat Genet, 2009. **41**(6): p. 703-7.

86. Bradfield, J.P., et al., *A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci.* PLoS genetics, 2011. **7**(9): p. e1002293-e1002293.

87. Burton, P.R., et al., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-678.

88. Farh, K.K.-H., et al., *Genetic and epigenetic fine mapping of causal autoimmune disease variants.* Nature, 2015. **518**(7539): p. 337-343.

89. Ram, R., et al., *Systematic Evaluation of Genes and Genetic Variants Associated with Type 1 Diabetes Susceptibility.* J Immunol, 2016. **196**(7): p. 3043-53.

90. Patterson, C., et al., *Diabetes in the young - a global view and worldwide estimates of numbers of children with type 1 diabetes.* Diabetes Res Clin Pract, 2014. **103**(2): p. 161-75.

91. Rewers, M. and J. Ludvigsson, *Environmental risk factors for type 1 diabetes.* Lancet (London, England), 2016. **387**(10035): p. 2340-2348.

92. Redondo, M.J., et al., *Heterogeneity of type I diabetes: analysis of monozygotic twins in Great Britain and the United States.* Diabetologia, 2001. **44**(3): p. 354-62.

93. Holmberg, H., et al., *Short duration of breast-feeding as a risk-factor for beta-cell autoantibodies in 5-year-old children from the general population.* Br J Nutr, 2007. **97**(1): p. 111-6.

94. Stene, L.C. and M. Rewers, *Immunology in the clinic review series; focus on type 1 diabetes and viruses: the enterovirus link to type 1 diabetes: critical review of human studies.* Clinical and experimental immunology, 2012. **168**(1): p. 12-23.

95. Viskari, H., et al., *Relationship between the incidence of type 1 diabetes and maternal enterovirus antibodies: time trends and geographical variation.* Diabetologia, 2005. **48**(7): p. 1280-1287.

96. Brown, C.T., et al., *Gut microbiome metagenomics analysis suggests a functional model for the development of autoimmunity for type 1 diabetes.* PLoS One, 2011. **6**(10): p. e25792.

97. Dang, M.N., R. Buzzetti, and P. Pozzilli, *Epigenetics in autoimmune diseases with focus on type 1 diabetes.* Diabetes Metab Res Rev, 2013. **29**(1): p. 8-18.

98. Zhang, Y., et al., *MicroRNAs in CD4(+) T cell subsets are markers of disease risk and T cell dysfunction in individuals at risk for type 1 diabetes.* J Autoimmun, 2016. **68**: p. 52-61.

99. Stefan, M., et al., *DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology.* J Autoimmun, 2014. **50**: p. 33-7.

100. Elboudwarej, E., et al., *Hypomethylation within gene promoter regions and type 1 diabetes in discordant monozygotic twins.* J Autoimmun, 2016. **68**: p. 23-9.

101. Rakyan, V.K., et al., *Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis.* PLoS Genet, 2011. **7**(9): p. e1002300.

102. Paul, D.S., et al., *Increased DNA methylation variability in type 1 diabetes across three immune effector cell types.* Nat Commun, 2016. **7**: p. 13555.

103. Li, Y., et al., *Abnormal DNA methylation in CD4+ T cells from people with latent autoimmune diabetes in adults.* Diabetes Res Clin Pract, 2011. **94**(2): p. 242-8.

104. Josefowicz, S.Z., L.F. Lu, and A.Y. Rudensky, *Regulatory T cells: mechanisms of differentiation and function.* Annu Rev Immunol, 2012. **30**: p. 531-64.

105. Jeker, L.T., H. Bour-Jordan, and J.A. Bluestone, *Breakdown in peripheral tolerance in type 1 diabetes in mice and humans.* Cold Spring Harb Perspect Med, 2012. **2**(3): p. a007807.

106. Ferraro, A., et al., *Expansion of Th17 cells and functional defects in T regulatory cells are key features of the pancreatic lymph nodes in patients with type 1 diabetes.* Diabetes, 2011. **60**(11): p. 2903-13.

107. van der Vliet, H.J.J. and E.E. Nieuwenhuis, *IPEX as a result of mutations in FOXP3.* Clinical & developmental immunology, 2007. **2007**: p. 89017-89017.

108. Feuerer, M., et al., *How punctual ablation of regulatory T cells unleashes an autoimmune lesion within the pancreatic islets.* Immunity, 2009. **31**(4): p. 654-64.

109. Mercadante, E.R. and U.M. Lorenz, *Breaking Free of Control: How Conventional T Cells Overcome Regulatory T Cell Suppression.* Front Immunol, 2016. **7**: p. 193.

110. Hnisz, D., et al., *Super-Enhancers in the Control of Cell Identity and Disease.* Cell, 2013. **155**(4): p. 934-947.

111. Schmidl, C., et al., *The enhancer and promoter landscape of human regulatory and conventional T-cell subpopulations.* Blood, 2014. **123**(17): p. e68-78.

112. Ferraro, A., et al., *Interindividual variation in human T regulatory cells.* Proceedings of the National Academy of Sciences, 2014. **111**(12): p. E1111-E1120.

113. Jiang, C. and B.F. Pugh, *Nucleosome positioning and gene regulation: advances through genomics.* Nat Rev Genet, 2009. **10**(3): p. 161-72.

114. *Beyond the genome.* Nature, 2015. **518**(7539): p. 273.

115. Cedar, H. and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms.* Nat Rev Genet, 2009. **10**(5): p. 295-304.

116. Jenuwein, T. and C.D. Allis, *Translating the Histone Code.* Science, 2001. **293**(5532): p. 1074.

117. Whyte, Warren A., et al., *Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes.* Cell, 2013. **153**(2): p. 307-319.

118. Kulaeva, O.I., et al., *Distant activation of transcription: mechanisms of enhancer action.* Mol Cell Biol, 2012. **32**(24): p. 4892-7.

119. Heinz, S., et al., *The selection and function of cell type-specific enhancers.* Nat Rev Mol Cell Biol, 2015. **16**(3): p. 144-54.

120. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. **4**(3): p. 317-24.

121. The, E.P.C., *An Integrated Encyclopedia of DNA Elements in the Human Genome.* Nature, 2012. **489**(7414): p. 57-74.

122. Bernstein, B.E., et al., *The NIH Roadmap Epigenomics Mapping Consortium.* Nat Biotechnol, 2010. **28**(10): p. 1045-8.

123. Lizio, M., et al., *Gateways to the FANTOM5 promoter level mammalian expression atlas.* Genome Biology, 2015. **16**(1): p. 22.

124. Stunnenberg, H.G. and M. Hirst, *The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery.* Cell, 2016. **167**(5): p. 1145-1149.

125. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.

126. Simon, J.M., et al., *Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA.* Nat. Protocols, 2012. **7**(2): p. 256-267.

127. Giresi, P.G., et al., *FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin.* Genome Research, 2007. **17**(6): p. 877-885.

128. Baek, S. and M.H. Sung, *Genome-Scale Analysis of Cell-Specific Regulatory Codes Using Nuclear Enzymes.* Methods Mol Biol, 2016. **1418**: p. 225-40.

129. Novo, C.L., et al., *Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition.* Cell Rep, 2018. **22**(10): p. 2615-2627.

130. Miguel-Escalada, I., et al., *Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes.* Nature Genetics, 2019. **51**(7): p. 1137-1148.

131. Jeng, M.Y., et al., *Enhancer Connectome Nominates Target Genes of Inherited Risk Variants from Inflammatory Skin Disorders.* J Invest Dermatol, 2019. **139**(3): p. 605-614.

132. Spurrell, C.H., D.E. Dickel, and A. Visel, *The Ties That Bind: Mapping the Dynamic Enhancer-Promoter Interactome.* Cell, 2016. **167**(5): p. 1163-1166.

133. Belton, J.M., et al., *Hi-C: a comprehensive technique to capture the conformation of genomes.* Methods, 2012. **58**(3): p. 268-76.

134. Liu, N., et al., *3DFAACTS-SNP: Using regulatory T cell-specific epigenomics data to uncover candidate mechanisms of Type-1 Diabetes (T1D) risk.* bioRxiv, 2020: p. 2020.09.04.279554.

135. Alasoo, K., et al., *Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response.* Nature Genetics, 2018. **50**(3): p. 424-431.

136. Gate, R.E., et al., *Genetic determinants of co-accessible chromatin regions in activated T cells across humans.* Nature Genetics, 2018. **50**(8): p. 1140-1150.

137. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-541.

138. Li, Y., et al., *The DNA methylome of human peripheral blood mononuclear cells.* PLoS Biol, 2010. **8**(11): p. e1000533.

139. Johnson, M.D., et al., *Genetic Analysis of the Cardiac Methylome at Single Nucleotide Resolution in a Model of Human Cardiovascular Disease.* PLoS Genetics, 2014. **10**(12): p. e1004813.

140. Tost, J., S. Gay, and G. Firestein, *Epigenetics of the immune system and alterations in inflammation and autoimmunity.* Epigenomics, 2017. **9**(4): p. 371-373.

141. Buenrostro, J.D., et al., *ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.* Curr Protoc Mol Biol, 2015. **109**: p. 21.29.1-9.

142. Tripodi, I.J., M.A. Allen, and R.D. Dowell, *Detecting Differential Transcription Factor Activity from ATAC-Seq Data.* Molecules, 2018. **23**(5).

143. Xu, S., et al., *regSNPs-ASB: A Computational Framework for Identifying Allele-Specific Transcription Factor Binding From ATAC-seq Data.* Frontiers in Bioengineering and Biotechnology, 2020. **8**: p. 886.

144. Mognol, G.P., et al., *Exhaustion-associated regulatory regions in CD8+ tumor-infiltrating T cells.* Proc Natl Acad Sci U S A, 2017. **114**(13): p. E2776-e2785.

145. Varshney, A., et al., *Genetic regulatory signatures underlying islet gene expression and type 2 diabetes.* Proc Natl Acad Sci U S A, 2017. **114**(9): p. 2301-2306.

146. Fullard, J.F., et al., *Open chromatin profiling of human postmortem brain infers functional roles for non-coding schizophrenia loci.* Hum Mol Genet, 2017.

147. Hughes, A.E., et al., *Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors.* Sci Rep, 2017. **7**: p. 43184.

148. Ueki, J., et al., *Myotonic dystrophy type 1 patient-derived iPSCs for the investigation of CTG repeat instability.* Sci Rep, 2017. **7**: p. 42522.

149. Simon, C.S., et al., *Functional characterisation of cis-regulatory elements governing dynamic Eomes expression in the early mouse embryo.* Development, 2017. **144**(7): p. 1249-1260.

150. Xu, J., et al., *Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells.* Nat Genet, 2017. **49**(3): p. 377-386.

151. Timashev, L.A., et al., *The DDR at telomeres lacking intact shelterin does not require substantial chromatin decompaction.* Genes & Development, 2017. **31**(6): p. 578-589.

152. Tan, T., et al., *Alteration of Regulatory T Cells in Type 1 Diabetes Mellitus: A Comprehensive Review.* Clinical Reviews in Allergy & Immunology, 2014. **47**(2): p. 234-243.

153. Corces, M.R., et al., *Omni-ATAC-seq: Improved ATAC-seq protocol.* 2017.

154. Zhao, S., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion.* Scientific Reports, 2018. **8**(1): p. 4781.

155. Schroeder, A., et al., *The RIN: an RNA integrity number for assigning integrity values to RNA measurements.* BMC Molecular Biology, 2006. **7**(1): p. 3.

156. Andrews, S. *FastQC: a quality control tool for high throughput sequence data*. 2010; Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

157. Amemiya, H.M., A. Kundaje, and A.P. Boyle, *The ENCODE Blacklist: Identification of Problematic Regions of the Genome.* Scientific Reports, 2019. **9**(1): p. 9354.

158. Adey, A., et al., *Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.* Genome biology, 2010. **11**(12): p. R119-R119.

159. Gaspar, J.M., *Improved peak-calling with MACS2.* bioRxiv, 2018: p. 496521.

160. Lun, A.T.L. and G.K. Smyth, *De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly.* Nucleic acids research, 2014. **42**(11): p. e95-e95.

161. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.

162. Lun, A.T. and G.K. Smyth, *csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows.* Nucleic Acids Res, 2016. **44**(5): p. e45.

163. Zuberbuehler, M.K., et al., *The transcription factor c-Maf is essential for the commitment of IL-17-producing γδ T cells.* Nature immunology, 2019. **20**(1): p. 73-85.

164. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* 2011, 2011. **17**(1): p. 3.

165. Schubert, M., S. Lindgreen, and L. Orlando, *AdapterRemoval v2: rapid adapter trimming, identification, and read merging.* BMC Research Notes, 2016. **9**(1): p. 88.

166. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics (Oxford, England), 2013. **29**(1): p. 15-21.

167. Gasperini, M., J.M. Tome, and J. Shendure, *Towards a comprehensive catalogue of validated and target-linked human enhancers.* Nature Reviews Genetics, 2020. **21**(5): p. 292-310.

168. Stunnenberg, H.G. and M. Hirst, *The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery.* Cell, 2016. **167**(7): p. 1897.

169. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation.* Genome Res, 2010. **20**(3): p. 320-31.

170. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.

171. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.

172. Shaw, E.G., et al., *Postoperative radiotherapy of supratentorial low-grade gliomas.* Int J Radiat Oncol Biol Phys, 1989. **16**(3): p. 663-8.

173. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.

174. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.

175. Sabo, P.J., et al., *Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays.* Nat Methods, 2006. **3**(7): p. 511-8.

176. Hesselberth, J.R., et al., *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nat Methods, 2009. **6**(4): p. 283-9.

177. Boyle, A.P., et al., *High-resolution mapping and characterization of open chromatin across the genome.* Cell, 2008. **132**(2): p. 311-22.

178. Corces, M.R., et al., *Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution.* Nat Genet, 2016. **48**(10): p. 1193-203.

179. Klemm, S.L., Z. Shipony, and W.J. Greenleaf, *Chromatin accessibility and the regulatory epigenome.* Nature Reviews Genetics, 2019. **20**(4): p. 207-220.

180. Health, G.o.W.A.D.o. *Guidelines for Human Biobanks, Genetic Research Databases and Associated Data.* 2010; Available from: http://www.health.wa.gov.au/circularsnew/circular.cfm?Circ_ID=12748.

181. Ramachandran, H., et al., *Optimal thawing of cryopreserved peripheral blood mononuclear cells for use in high-throughput human immune monitoring studies.* Cells, 2012. **1**(3): p. 313-24.

182. Fujiwara, S., et al., *High Quality ATAC-Seq Data Recovered from Cryopreserved Breast Cell Lines and Tissue.* Scientific reports, 2019. **9**(1): p. 516-516.

183. Kutscher, S., et al., *Overnight Resting of PBMC Changes Functional Signatures of Antigen Specific T- Cell Responses: Impact for Immune Monitoring within Clinical Trials.* PLOS ONE, 2013. **8**(10): p. e76215.

184. García-Piñeres, A.J., et al., *DNAse treatment following thawing of cryopreserved PBMC is a procedure suitable for lymphocyte functional studies.* J Immunol Methods, 2006. **313**(1-2): p. 209-13.

185. Smith, J.G., et al., *Development and validation of a gamma interferon ELISPOT assay for quantitation of cellular immune responses to varicella-zoster virus.* Clinical and diagnostic laboratory immunology, 2001. **8**(5): p. 871-879.

186. Picelli, S., et al., *Tn5 transposase and tagmentation procedures for massively scaled sequencing projects.* Genome Res, 2014. **24**(12): p. 2033-40.

187. Yan, F., et al., *From reads to insight: a hitchhiker's guide to ATAC-seq data analysis.* Genome Biology, 2020. **21**(1): p. 22.

188. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.

189. Barthelson, R.A., et al., *Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells.* BMC Genomics, 2007. **8**(1): p. 340.

190. Trask, H.W., et al., *Microarray analysis of cytoplasmic versus whole cell RNA reveals a considerable number of missed and false positive mRNAs.* Rna, 2009. **15**(10): p. 1917-28.

191. Zaghlool, A., et al., *Characterization of the nuclear and cytosolic transcriptomes in human brain tissue reveals new insights into the subcellular distribution of RNA transcripts.* Scientific Reports, 2021. **11**(1): p. 4076.

192. Olson, J.E., et al., *Biobanks and personalized medicine.* Clin Genet, 2014. **86**(1): p. 50-5.

193. Baboo, J., et al., *The Impact of Varying Cooling and Thawing Rates on the Quality of Cryopreserved Human Peripheral Blood T Cells.* Scientific Reports, 2019. **9**(1): p. 3417.

194. Bull, M., et al., *Defining blood processing parameters for optimal detection of cryopreserved antigen-specific responses for HIV vaccine trials.* J Immunol Methods, 2007. **322**(1-2): p. 57-69.

195. Terry, C., et al., *Optimization of the cryopreservation and thawing protocol for human hepatocytes for use in cell transplantation.* Liver Transplantation, 2010. **16**(2): p. 229-237.

196. Wang, L., et al., *Standardization of cryopreserved peripheral blood mononuclear cells through a resting process for clinical immunomonitoring--Development of an algorithm.* Cytometry A, 2016. **89**(3): p. 246-58.

197. Hønge, B.L., et al., *Optimizing recovery of frozen human peripheral blood mononuclear cells for flow cytometry.* PLOS ONE, 2017. **12**(11): p. e0187440.

198. Riedhammer, C., D. Halbritter, and R. Weissert, *Peripheral Blood Mononuclear Cells: Isolation, Freezing, Thawing, and Culture.* Methods Mol Biol, 2016. **1304**: p. 53-61.

199. Kloverpris, H., et al., *Dimethyl sulfoxide (DMSO) exposure to human peripheral blood mononuclear cells (PBMCs) abolish T cell responses only in high concentrations and following coincubation for more than two hours.* J Immunol Methods, 2010. **356**(1-2): p. 70-8.

200. Corces, M.R., et al., *An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues.* Nature Methods, 2017. **14**: p. 959.

201. Colamatteo, A., et al., *Molecular Mechanisms Controlling Foxp3 Expression in Health and Autoimmunity: From Epigenetic to Post-translational Regulation.* Frontiers in Immunology, 2020. **10**(3136).

202. Zheng, Y., et al., *Role of conserved non-coding DNA elements in the Foxp3 gene in regulatory T-cell fate.* Nature, 2010. **463**(7282): p. 808-812.

203. Ono, M., *Control of regulatory T-cell differentiation and function by T-cell receptor signalling and Foxp3 transcription factor complexes.* Immunology, 2020. **160**(1): p. 24-37.

204. Allison, K.A., et al., *Affinity and dose of TCR engagement yield proportional enhancer and gene activity in CD4+ T cells.* Elife, 2016. **5**.

205. Ono, M., et al., *Foxp3 controls regulatory T-cell function by interacting with AML1/Runx1.* Nature, 2007. **446**(7136): p. 685-689.

206. Elinav, E., et al., *Inflammation-induced cancer: crosstalk between tumours, immune cells and microorganisms.* Nature Reviews Cancer, 2013. **13**(11): p. 759-771.

207. Donath, M.Y. and S.E. Shoelson, *Type 2 diabetes as an inflammatory disease.* Nature Reviews Immunology, 2011. **11**(2): p. 98-107.

208. Schmiedel, B.J., et al., *Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression.* Cell, 2018. **175**(6): p. 1701-1715.e16.

209. Ye, C.J., et al., *Intersection of population variation and autoimmunity genetics in human T cell activation.* Science, 2014. **345**(6202): p. 1254665.

210. Taliaferro, J.M., E.T. Wang, and C.B. Burge, *Genomic analysis of RNA localization.* RNA Biol, 2014. **11**(8): p. 1040-50.

211. Martin, K.C. and A. Ephrussi, *mRNA localization: gene expression in the spatial dimension.* Cell, 2009. **136**(4): p. 719-30.

212. Kuersten, S. and E.B. Goodwin, *The power of the 3' UTR: translational control and development.* Nat Rev Genet, 2003. **4**(8): p. 626-37.

213. Zhang, T., et al., *RNALocate: a resource for RNA subcellular localizations.* Nucleic Acids Research, 2017. **45**(D1): p. D135-D138.

214. Marrison, J.L., et al., *Subcellular Visualization of Gene Transcripts Encoding Key Proteins of the Chlorophyll Accumulation Process in Developing Chloroplasts.* Plant Physiology, 1996. **110**(4): p. 1089.

215. Solnestam, B.W., et al., *Comparison of total and cytoplasmic mRNA reveals global regulation by nuclear retention and miRNAs.* BMC genomics, 2012. **13**: p. 574-574.

216. Battich, N., T. Stoeger, and L. Pelkmans, *Control of Transcript Variability in Single Mammalian Cells.* Cell, 2015. **163**(7): p. 1596-610.

217. Chen, L., *A global comparison between nuclear and cytosolic transcriptomes reveals differential compartmentalization of alternative transcript isoforms.* Nucleic Acids Res, 2010. **38**(4): p. 1086-97.

218. Pastro, L., et al., *Nuclear Compartmentalization Contributes to Stage-Specific Gene Expression Control in Trypanosoma cruzi.* Front Cell Dev Biol, 2017. **5**: p. 8.

219. Schwarz, D.S. and M.D. Blower, *The endoplasmic reticulum: structure, function and response to cellular signaling.* Cellular and molecular life sciences : CMLS, 2016. **73**(1): p. 79-94.

220. Kumar, M. and G.G. Carmichael, *Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3542-7.

221. Prasanth, K.V., et al., *Regulating gene expression through RNA nuclear retention.* Cell, 2005. **123**(2): p. 249-63.

222. Boutz, P.L., A. Bhutkar, and P.A. Sharp, *Detained introns are a novel, widespread class of post-transcriptionally spliced introns.* Genes & development, 2015. **29**(1): p. 63-80.

223. Chen, L.L. and G.G. Carmichael, *Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA.* Mol Cell, 2009. **35**(4): p. 467-78.

224. Miyamoto, Y., et al., *Cellular stresses induce the nuclear accumulation of importin alpha and cause a conventional nuclear import block.* The Journal of cell biology, 2004. **165**(5): p. 617-623.

225. Bahar Halpern, K., et al., *Nuclear Retention of mRNA in Mammalian Tissues.* Cell Rep, 2015. **13**(12): p. 2653-62.

226. Jensen, A., et al., *PAT1 (SLC36A1) shows nuclear localization and affects growth of smooth muscle cells from rats.* Am J Physiol Endocrinol Metab, 2014. **306**(1): p. E65-74.

227. Couvillion, M.T., et al., *Synchronized mitochondrial and cytosolic translation programs.* Nature, 2016. **533**(7604): p. 499-503.

228. Feng, J., et al., *Identifying ChIP-seq enrichment using MACS.* Nature protocols, 2012. **7**(9): p. 1728-1740.

229. Zhang, Y., et al., *Model-based Analysis of ChIP-Seq (MACS).* Genome Biology, 2008. **9**(9): p. R137.

230. Reske, J.J., M.R. Wilson, and R.L. Chandler, *ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation.* Epigenetics & Chromatin, 2020. **13**(1): p. 22.

231. Yu, G., L.-G. Wang, and Q.-Y. He, *ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization.* Bioinformatics, 2015. **31**(14): p. 2382-2383.

232. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.

233. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* Omics : a journal of integrative biology, 2012. **16**(5): p. 284-287.

234. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545.

235. Sakaguchi, S., et al., *FOXP3+ regulatory T cells in the human immune system.* Nature Reviews Immunology, 2010. **10**(7): p. 490-500.

236. Miyara, M., et al., *Functional Delineation and Differentiation Dynamics of Human CD4+ T Cells Expressing the FoxP3 Transcription Factor.* Immunity, 2009. **30**(6): p. 899-911.

237. Chtanova, T., et al., *Identification of T Cell-Restricted Genes, and Signatures for Different T Cell Responses, Using a Comprehensive Collection of Microarray Datasets.* The Journal of Immunology, 2005. **175**(12): p. 7837.

238. Henriksson, J., et al., *Genome-wide CRISPR screens in T helper cells reveal pervasive cross-talk between activation and differentiation.* bioRxiv, 2017.

239. Bysani, M., et al., *ATAC-seq reveals alterations in open chromatin in pancreatic islets from subjects with type 2 diabetes.* Scientific Reports, 2019. **9**(1): p. 7785.

240. Zhao, X., et al., *Genome-wide identification of accessible chromatin regions in bumblebee by ATAC-seq.* Scientific Data, 2020. **7**(1): p. 367.

241. Zhao, Y., D. Zheng, and A. Cvekl, *Profiling of chromatin accessibility and identification of general cis-regulatory mechanisms that control two ocular lens differentiation pathways.* Epigenetics & chromatin, 2019. **12**(1): p. 27-27.

242. Corces, M.R., et al., *The chromatin accessibility landscape of primary human cancers.* Science, 2018. **362**(6413).

243. Radens, C.M., et al., *Meta-analysis of transcriptomic variation in T-cell populations reveals both variable and consistent signatures of gene expression and splicing.* Rna, 2020. **26**(10): p. 1320-1333.

244. Ross-Innes, C.S., et al., *Differential oestrogen receptor binding is associated with clinical outcome in breast cancer.* Nature, 2012. **481**(7381): p. 389-393.

245. Sandrine Dudoit , Y.H.Y., Matthew J. Callow , Terence P. Speed, *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* STATISTICA SINICA, 2002. **12**: p. 111-139.

246. Robinson, M.D. and A. Oshlack, *A scaling normalization method for differential expression analysis of RNA-seq data.* Genome Biology, 2010. **11**(3): p. R25.

247. Bhairavabhotla, R., et al., *Transcriptome profiling of human FoxP3+ regulatory T cells.* Human immunology, 2016. **77**(2): p. 201-213.

248. Eisenbarth, G.S., *Type I diabetes mellitus. A chronic autoimmune disease.* N Engl J Med, 1986. **314**(21): p. 1360-8.

249. Welfare, A.I.o.H.a., *Prevalence of Type 1 diabetes in Australian children, 2008.* Diabetes series. Vol. 15. 2011. 30.

250. Wildin, R.S., et al., *X-linked neonatal diabetes mellitus, enteropathy and endocrinopathy syndrome is the human equivalent of mouse scurfy.* Nat Genet, 2001. **27**(1): p. 18-20.

251. Bennett, C.L., et al., *The immune dysregulation, polyendocrinopathy, enteropathy, X-linked syndrome (IPEX) is caused by mutations of FOXP3.* Nat Genet, 2001. **27**(1): p. 20-1.

252. Zemmour, D., et al., *Single-cell analysis of FOXP3 deficiencies in humans and mice unmasks intrinsic and extrinsic CD4+ T cell perturbations.* Nature Immunology, 2021. **22**(5): p. 607-619.

253. Brusko, T.M., et al., *Functional defects and the influence of age on the frequency of CD4+ CD25+ T-cells in type 1 diabetes.* Diabetes, 2005. **54**(5): p. 1407-14.

254. Tang, Q., et al., *In vitro-expanded antigen-specific regulatory T cells suppress autoimmune diabetes.* J Exp Med, 2004. **199**(11): p. 1455-65.

255. Törn, C., et al., *Role of Type 1 Diabetes-Associated SNPs on Risk of Autoantibody Positivity in the TEDDY Study.* Diabetes, 2015. **64**(5): p. 1818-1829.

256. Harbison, J.E., et al., *Associations between diet, the gut microbiome and short chain fatty acids in youth with islet autoimmunity and type 1 diabetes.* Pediatr Diabetes, 2021. **22**(3): p. 425-433.

257. Harbison, J.E., et al., *Gut microbiome dysbiosis and increased intestinal permeability in children with islet autoimmunity and type 1 diabetes: A prospective cohort study.* Pediatr Diabetes, 2019. **20**(5): p. 574-583.

258. Zhu, L.J., et al., *ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data.* BMC Bioinformatics, 2010. **11**(1): p. 237.

259. Pers, T.H., P. Timshel, and J.N. Hirschhorn, *SNPsnap: a Web-based tool for identification and annotation of matched SNPs.* Bioinformatics, 2015. **31**(3): p. 418-20.

260. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

261. Khan, A. and A. Mathelier, *Intervene: a tool for intersection and visualization of multiple gene or genomic region sets.* BMC Bioinformatics, 2017. **18**(1): p. 287.

262. Haque, A., et al., *A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.* Genome medicine, 2017. **9**(1): p. 75-75.

263. Teng, M. and R.A. Irizarry, *Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data.* Genome research, 2017. **27**(11): p. 1930-1938.

264. Benjamini, Y. and T.P. Speed, *Summarizing and correcting the GC content bias in high-throughput sequencing.* Nucleic Acids Research, 2012. **40**(10): p. e72-e72.

265. Leek, J.T., *svaseq: removing batch effects and other unwanted noise from sequencing data.* Nucleic acids research, 2014. **42**(21): p. e161-e161.

266. Gagnon-Bartsch, J.A. and T.P. Speed, *Using control genes to correct for unwanted variation in microarray data.* Biostatistics, 2012. **13**(3): p. 539-552.

267. Sung, M.-H., S. Baek, and G.L. Hager, *Genome-wide footprinting: ready for prime time?* Nat Meth, 2016. **13**(3): p. 222-228.

268. Hess, J., et al., *AP-1 and Cbfa/runt physically interact and regulate parathyroid hormone-dependent MMP13 expression in osteoblasts through a new osteoblast-specific element 2/AP-1 composite element.* J Biol Chem, 2001. **276**(23): p. 20029-38.

269. D'Alonzo, R.C., et al., *Physical interaction of the activator protein-1 factors c-Fos and c-Jun with Cbfa1 for collagenase-3 promoter activation.* J Biol Chem, 2002. **277**(1): p. 816-22.

270. Sun, J., et al., *Heme regulates the dynamic exchange of Bach1 and NF-E2-related factors in the Maf transcription factor network.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(6): p. 1461.

271. Okita, Y., et al., *Transforming growth factor-β induces transcription factors MafK and Bach1 to suppress expression of the heme oxygenase-1 gene.* J Biol Chem, 2013. **288**(28): p. 20658-67.

272. Li, P., et al., *BATF-JUN is critical for IRF4-mediated transcription in T cells.* Nature, 2012. **490**(7421): p. 543-6.

273. Tran, G.T., et al., *IL-5 promotes induction of antigen-specific CD4+CD25+ T regulatory cells that suppress autoimmunity.* Blood, 2012. **119**(19): p. 4441-50.

274. Renoux, F., et al., *The AP1 Transcription Factor Fosl2 Promotes Systemic Autoimmunity and Inflammation by Repressing Treg Development.* Cell Rep, 2020. **31**(13): p. 107826.

275. Singh, T., et al., *Loss of MafA and MafB expression promotes islet inflammation.* Scientific reports, 2019. **9**(1): p. 9074-9074.

276. Imbratta, C., et al., *Maf deficiency in T cells dysregulates Treg - TH17 balance leading to spontaneous colitis.* Scientific Reports, 2019. **9**(1): p. 6135.

277. Liu, W., C. Dong, and X. Liu, *[The role of IL4I1 in immunoregulation: An update].* Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi, 2021. **37**(1): p. 79-83.

278. Ma, J., et al., *Genetic influences of the intercellular adhesion molecule 1 (ICAM-1) gene polymorphisms in development of Type 1 diabetes and diabetic nephropathy.* Diabetic medicine : a journal of the British Diabetic Association, 2006. **23**(10): p. 1093-1099.

279. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-753.

280. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nature Genetics, 2010. **42**(7): p. 565-569.

281. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.

282. Bodmer, W. and C. Bonilla, *Common and rare variants in multifactorial susceptibility to common diseases.* Nature Genetics, 2008. **40**(6): p. 695-701.

283. Saint Pierre, A. and E. Génin, *How important are rare variants in common disease?* Brief Funct Genomics, 2014. **13**(5): p. 353-61.

284. Sampson, J.N., et al., *Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types.* Journal of the National Cancer Institute, 2015. **107**(12): p. djv279-djv279.

285. Muñoz, M., et al., *Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank.* Nature genetics, 2016. **48**(9): p. 980-983.

286. Wei, W.-H., G. Hemani, and C.S. Haley, *Detecting epistasis in human complex traits.* Nature Reviews Genetics, 2014. **15**(11): p. 722-733.

287. Dempfle, A., et al., *Gene-environment interactions for complex traits: definitions, methodological requirements and challenges.* Eur J Hum Genet, 2008. **16**(10): p. 1164-72.

288. Auer, P.L. and G. Lettre, *Rare variant association studies: considerations, challenges and opportunities.* Genome Med, 2015. **7**(1): p. 16.

289. Martin, A.R., et al., *Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.* Am J Hum Genet, 2017. **100**(4): p. 635-649.

290. Bacon, C.M., et al., *Interleukin 12 induces tyrosine phosphorylation and activation of STAT4 in human lymphocytes.* Proceedings of the National Academy of Sciences of the United States of America, 1995. **92**(16): p. 7307-7311.

291. Walker, L.S.K. and M. von Herrath, *CD4 T cell differentiation in type 1 diabetes.* Clinical and experimental immunology, 2016. **183**(1): p. 16-29.

292. Bhat, N.K., et al., *Reciprocal expression of human ETS1 and ETS2 genes during T-cell activation: regulatory role for the protooncogene ETS1.* Proceedings of the National Academy of Sciences of the United States of America, 1990. **87**(10): p. 3723-3727.

293. Steck, A.K., et al., *Association of non-HLA genes with type 1 diabetes autoimmunity.* Diabetes, 2005. **54**(8): p. 2482-6.

294. Bossini-Castillo, L., et al., *Immune disease variants modulate gene expression in regulatory CD4+ T cells and inform drug targets*. 2019, bioRxiv.

295. Morahan, G., et al., *Tests for genetic interactions in type 1 diabetes: linkage and stratification analyses of 4,422 affected sib-pairs.* Diabetes, 2011. **60**(3): p. 1030-40.

296. van der Merwe, P.A., et al., *CD80 (B7-1) binds both CD28 and CTLA-4 with a low affinity and very fast kinetics.* J Exp Med, 1997. **185**(3): p. 393-403.

297. Kavvoura, F.K. and J.P.A. Ioannidis, *CTLA-4 Gene Polymorphisms and Susceptibility to Type 1 Diabetes Mellitus: A HuGE Review and Meta-Analysis.* American Journal of Epidemiology, 2005. **162**(1): p. 3-16.

298. Kristiansen, O.P., Z.M. Larsen, and F. Pociot, *CTLA-4 in autoimmune diseases--a general susceptibility gene to autoimmunity?* Genes Immun, 2000. **1**(3): p. 170-84.

299. Ostrov, D.A., et al., *Structure of murine CTLA-4 and its role in modulating T cell responsiveness.* Science, 2000. **290**(5492): p. 816-9.

300. Luther, S.A. and J.G. Cyster, *Chemokines as regulators of T cell differentiation.* Nat Immunol, 2001. **2**(2): p. 102-7.

301. Gale, L.M. and S.R. McColl, *Chemokines: extracellular messengers for all occasions?* Bioessays, 1999. **21**(1): p. 17-28.

302. Mantovani, A., *The chemokine system: redundancy for robust outputs.* Immunol Today, 1999. **20**(6): p. 254-7.

303. Sasaki, K., et al., *Modulation of autoimmune pathogenesis by T cell-triggered inflammatory cell death.* Nature Communications, 2019. **10**(1): p. 3878.

304. Skapenko, A., et al., *The role of the T cell in autoimmune inflammation.* Arthritis Research & Therapy, 2005. **7**(2): p. S4.

305. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Res, 2012. **22**(9): p. 1658-67.

306. Fasolino, M., et al., *Genetic Variation in Type 1 Diabetes Reconfigures the 3D Chromatin Organization of T Cells and Alters Gene Expression.* Immunity, 2020. **52**(2): p. 257-274.e11.

307. Uhlén, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.

308. Shaulian, E. and M. Karin, *AP-1 in cell proliferation and survival.* Oncogene, 2001. **20**(19): p. 2390-2400.

309. Shaulian, E. and M. Karin, *AP-1 as a regulator of cell life and death.* Nat Cell Biol, 2002. **4**(5): p. E131-6.

310. Hess, J., P. Angel, and M. Schorpp-Kistner, *AP-1 subunits: quarrel and harmony among siblings.* J Cell Sci, 2004. **117**(Pt 25): p. 5965-73.

311. Liu, J.Z., et al., *Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.* Nature Genetics, 2015. **47**(9): p. 979-986.

312. Jostins, L., et al., *Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease.* Nature, 2012. **491**(7422): p. 119-124.

313. Moon, Y.M., et al., *The Fos-Related Antigen 1-JUNB/Activator Protein 1 Transcription Complex, a Downstream Target of Signal Transducer and Activator of Transcription 3, Induces T Helper 17 Differentiation and Promotes Experimental Autoimmune Arthritis.* Front Immunol, 2017. **8**: p. 1793.

314. Ciofani, M., et al., *A validated regulatory network for Th17 cell specification.* Cell, 2012. **151**(2): p. 289-303.

315. Li, J., et al., *The Expression Level of mRNA, Protein, and DNA Methylation Status of<i> FOSL2</i> of Uyghur in XinJiang in Type 2 Diabetes.* Journal of Diabetes Research, 2016. **2016**: p. 5957404.

316. Tripathi, S.K., et al., *Quantitative Proteomics Reveals the Dynamic Protein Landscape during Initiation of Human Th17 Cell Polarization.* iScience, 2019. **11**: p. 334-355.

317. Shetty, A., et al., *The AP-1 factors FOSL1 and FOSL2 co-regulate human Th17 responses.* bioRxiv, 2021: p. 2021.04.26.441472.

318. Lund, R.J., et al., *Genome-wide identification of novel genes involved in early Th1 and Th2 cell differentiation.* J Immunol, 2007. **178**(6): p. 3648-60.

319. Komatsu, N., et al., *Pathogenic conversion of Foxp3+ T cells into TH17 cells in autoimmune arthritis.* Nat Med, 2014. **20**(1): p. 62-8.

320. Noval Rivas, M., et al., *Regulatory T cell reprogramming toward a Th2-cell-like lineage impairs oral tolerance and promotes food allergy.* Immunity, 2015. **42**(3): p. 512-23.

321. Kitz, A. and M. Dominguez-Villar, *Molecular mechanisms underlying Th1-like Treg generation and function.* Cellular and molecular life sciences : CMLS, 2017. **74**(22): p. 4059-4075.

322. Kannan, A.K., et al., *IL-23 induces regulatory T cell plasticity with implications for inflammatory skin diseases.* Scientific Reports, 2019. **9**(1): p. 17675.

323. Emamaullee, J.A., et al., *Inhibition of Th17 Cells Regulates Autoimmune Diabetes in NOD Mice.* Diabetes, 2009. **58**(6): p. 1302.

324. Abdel-Moneim, A., H.H. Bakery, and G. Allam, *The potential pathogenic role of IL-17/Th17 cells in both type 1 and type 2 diabetes mellitus.* Biomedicine & Pharmacotherapy, 2018. **101**: p. 287-292.

325. Li, Y., Y. Liu, and C.-Q. Chu, *Th17 Cells in Type 1 Diabetes: Role in the Pathogenesis and Regulation by Gut Microbiome.* Mediators of inflammation, 2015. **2015**: p. 638470-638470.

326. Honkanen, J., et al., *IL-17 immunity in human type 1 diabetes.* J Immunol, 2010. **185**(3): p. 1959-67.

327. Shao, S., et al., *Th17 cells in type 1 diabetes.* Cellular Immunology, 2012. **280**(1): p. 16-21.

328. Solt, L.A. and T.P. Burris, *Th17 cells in Type 1 diabetes: a future perspective.* Diabetes management (London, England), 2015. **5**(4): p. 247-250.

329. Ferreira, R.C., et al., *IL-21 production by CD4+ effector T cells and frequency of circulating follicular helper T cells are increased in type 1 diabetes patients.* Diabetologia, 2015. **58**(4): p. 781-90.

330. Bradshaw, E.M., et al., *Monocytes from Patients with Type 1 Diabetes Spontaneously Secrete Proinflammatory Cytokines Inducing Th17 Cells.* The Journal of Immunology, 2009. **183**(7): p. 4432.

331. Nayak, B.K., *Understanding the relevance of sample size calculation.* Indian journal of ophthalmology, 2010. **58**(6): p. 469-470.

332. Anderson, J.T., et al., *Insulin-dependent diabetes in the NOD mouse model. II. Beta cell destruction in autoimmune diabetes is a TH2 and not a TH1 mediated event.* Autoimmunity, 1993. **15**(2): p. 113-22.

333. Kenefeck, R., et al., *Follicular helper T cell signature in type 1 diabetes.* J Clin Invest, 2015. **125**(1): p. 292-303.

334. Karabacak Calviello, A., et al., *Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling.* Genome Biology, 2019. **20**(1): p. 42.

335. Vierstra, J. and J.A. Stamatoyannopoulos, *Genomic footprinting.* Nature Methods, 2016. **13**(3): p. 213-221.

336. Chun, S., et al., *Shared effect modeling reveals that a fraction of autoimmune disease associations are consistent with eQTLs in three immune cell types.* bioRxiv, 2016: p. 053165.

337. Joller, N., et al., *Cutting edge: TIGIT has T cell-intrinsic inhibitory functions.* J Immunol, 2011. **186**(3): p. 1338-42.

338. Levin, S.D., et al., *Vstm3 is a member of the CD28 family and an important modulator of T-cell function.* Eur J Immunol, 2011. **41**(4): p. 902-15.

339. Yu, X., et al., *The surface protein TIGIT suppresses T cell activation by promoting the generation of mature immunoregulatory dendritic cells.* Nat Immunol, 2009. **10**(1): p. 48-57.

340. Kourepini, E., et al., *TIGIT Enhances Antigen-Specific Th2 Recall Responses and Allergic Disease.* J Immunol, 2016. **196**(9): p. 3570-80.

341. Kamran, N., et al., *Toll-like receptor ligands induce expression of the costimulatory molecule CD155 on antigen-presenting cells.* PLoS One, 2013. **8**(1): p. e54406.

342. Yang, I.V., et al., *DNA methylation and childhood asthma in the inner city.* J Allergy Clin Immunol, 2015. **136**(1): p. 69-80.

343. Lucca, L.E., et al., *TIGIT signaling restores suppressor function of Th1 Tregs.* JCI Insight, 2019. **4**(3).

344. Joller, N., et al., *Treg cells expressing the coinhibitory molecule TIGIT selectively inhibit proinflammatory Th1 and Th17 cell responses.* Immunity, 2014. **40**(4): p. 569-581.

345. Rapoport, M.J., et al., *Interleukin 4 reverses T cell proliferative unresponsiveness and prevents the onset of diabetes in nonobese diabetic mice.* J Exp Med, 1993. **178**(1): p. 87-99.

346. Bradley, L.M., et al., *Islet-Specific Th1, But Not Th2, Cells Secrete Multiple Chemokines and Promote Rapid Induction of Autoimmune Diabetes.* The Journal of Immunology, 1999. **162**(5): p. 2511.

347. Vaseghi, H. and Z. Jadali, *Th1/Th2 cytokines in Type 1 diabetes: Relation to duration of disease and gender.* Indian journal of endocrinology and metabolism, 2016. **20**(3): p. 312-316.

348. Almawi, W.Y., H. Tamim, and S.T. Azar, *T Helper Type 1 and 2 Cytokines Mediate the Onset and Progression of Type I (Insulin-Dependent) Diabetes.* The Journal of Clinical Endocrinology & Metabolism, 1999. **84**(5): p. 1497-1502.

349. Faulkner-Jones, B.E., et al., *Both TH1 and TH2 cytokine mRNAs are expressed in the NOD mouse pancreas in vivo.* Autoimmunity, 1996. **23**(2): p. 99-110.

350. Niu, H.-Q., et al., *Characteristics and reference ranges of CD4+T cell subpopulations among healthy adult Han Chinese in Shanxi Province, North China.* BMC Immunology, 2020. **21**(1): p. 44.

351. Cope, A., et al., *The Th1 life cycle: molecular control of IFN-γ to IL-10 switching.* Trends in Immunology, 2011. **32**(6): p. 278-286.

352. Smeltz, R.B., et al., *Role of IFN-gamma in Th1 differentiation: IFN-gamma regulates IL-18R alpha expression by preventing the negative effects of IL-4 and by inducing/maintaining IL-12 receptor beta 2 expression.* J Immunol, 2002. **168**(12): p. 6165-72.

353. Zhou, W., F. Zhang, and T.M. Aune, *Either IL-2 or IL-12 Is Sufficient to Direct Th1 Differentiation by Nonobese Diabetic T Cells.* The Journal of Immunology, 2003. **170**(2): p. 735.

354. Dong, C., *TH17 cells in development: an updated view of their molecular identity and genetic programming.* Nature Reviews Immunology, 2008. **8**(5): p. 337-348.

355. Chang, S.H. and C. Dong, *IL-17F: regulation, signaling and function in inflammation.* Cytokine, 2009. **46**(1): p. 7-11.

356. Yang, X.O., et al., *Regulation of inflammatory responses by IL-17F.* J Exp Med, 2008. **205**(5): p. 1063-75.

357. Bauquet, A.T., et al., *The costimulatory molecule ICOS regulates the expression of c-Maf and IL-21 in the development of follicular T helper cells and TH-17 cells.* Nature Immunology, 2009. **10**(2): p. 167-175.

358. Aschenbrenner, D., et al., *An immunoregulatory and tissue-residency program modulated by c-MAF in human TH17 cells.* Nature Immunology, 2018. **19**(10): p. 1126-1136.

359. Xu, J., et al., *c-Maf regulates IL-10 expression during Th17 polarization.* J Immunol, 2009. **182**(10): p. 6226-36.

360. Wu, X., J. Tian, and S. Wang, *Insight Into Non-Pathogenic Th17 Cells in Autoimmune Diseases.* Frontiers in immunology, 2018. **9**: p. 1112-1112.

361. Zielinski, C.E., et al., *Pathogen-induced human TH17 cells produce IFN-γ or IL-10 and are regulated by IL-1β.* Nature, 2012. **484**(7395): p. 514-8.

362. Ono, Y., et al., *T-helper 17 and Interleukin-17–Producing Lymphoid Tissue Inducer-Like Cells Make Different Contributions to Colitis in Mice.* Gastroenterology, 2012. **143**(5): p. 1288-1297.

363. Esplugues, E., et al., *Control of TH17 cells occurs in the small intestine.* Nature, 2011. **475**(7357): p. 514-518.

364. Bevington, S.L., et al., *Chromatin Priming Renders T Cell Tolerance-Associated Genes Sensitive to Activation below the Signaling Threshold for Immune Response Genes.* Cell Reports, 2020. **31**(10): p. 107748.

365. Groux, H. and F. Cottrez, *The complex role of interleukin-10 in autoimmunity.* J Autoimmun, 2003. **20**(4): p. 281-5.

366. Schloot, N.C., et al., *Serum IFN-gamma and IL-10 levels are associated with disease progression in non-obese diabetic mice.* Diabetes Metab Res Rev, 2002. **18**(1): p. 64-70.

367. Pennline, K.J., E. Roque-Gaffney, and M. Monahan, *Recombinant human IL-10 prevents the onset of diabetes in the nonobese diabetic mouse.* Clin Immunol Immunopathol, 1994. **71**(2): p. 169-75.

368. Kundaje, A., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-330.

369. Blattler, A., et al., *Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes.* Genome Biol, 2014. **15**(9): p. 469.

370. Freedman, M.L., et al., *Principles for the post-GWAS functional characterization of cancer risk loci.* Nat Genet, 2011. **43**(6): p. 513-8.

371. Edwards, S.L., et al., *Beyond GWASs: illuminating the dark road from association to function.* Am J Hum Genet, 2013. **93**(5): p. 779-97.

372. Amin Al Olama, A., et al., *Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans.* Hum Mol Genet, 2015. **24**(19): p. 5589-602.

373. Massarat, A.R., et al., *Discovering single nucleotide variants and indels from bulk and single-cell ATAC-seq.* Nucleic Acids Research, 2021. **49**(14): p. 7986-7994.

374. Kivling, A., et al., *Diverse foxp3 expression in children with type 1 diabetes and celiac disease.* Ann N Y Acad Sci, 2008. **1150**: p. 273-7.

375. Hope, C.M., et al., *Peptidase inhibitor 16 identifies a human regulatory T-cell subset with reduced FOXP3 expression over the first year of recent onset type 1 diabetes.* Eur J Immunol, 2019. **49**(8): p. 1235-1250.

376. Viisanen, T., et al., *FOXP3+ Regulatory T Cell Compartment Is Altered in Children With Newly Diagnosed Type 1 Diabetes but Not in Autoantibody-Positive at-Risk Children.* Frontiers in Immunology, 2019. **10**: p. 19.

377. Pauza, M.E., et al., *Variable Effects of Transgenic c-Maf on Autoimmune Diabetes.* Diabetes, 2001. **50**(1): p. 39.

378. Neumann, C., et al., *c-Maf-dependent Treg cell control of intestinal TH17 cells and IgA establishes host–microbiota homeostasis.* Nature Immunology, 2019. **20**(4): p. 471-481.

379. Hai, T., D. Lu, and C.C. Wolford, *TRANSCRIPTION FACTORS | ATF*, in *Encyclopedia of Respiratory Medicine*, G.J. Laurent and S.D. Shapiro, Editors. 2006, Academic Press: Oxford. p. 257-260.

380. Bjørnvold, M., et al., *FOXP3 polymorphisms in type 1 diabetes and coeliac disease.* J Autoimmun, 2006. **27**(2): p. 140-4.

381. Steck, A.K., et al., *Stepwise or linear decrease in penetrance of type 1 diabetes with lower-risk HLA genotypes over the past 40 years.* Diabetes, 2011. **60**(3): p. 1045-9.

382. Staeva-Vieira, T., M. Peakman, and M. von Herrath, *Translational mini-review series on type 1 diabetes: Immune-based therapeutic approaches for type 1 diabetes.* Clinical and experimental immunology, 2007. **148**(1): p. 17-31.

383. Ozgur, B.A., et al., *255-LB: The Role of Th1, Th2, Th17, and Treg Cells in Various Clinical Phases of Type 1 Diabetes.* Diabetes, 2019. **68**(Supplement 1): p. 255-LB.