


SOFTWARE

Open Access



# DAWN: a resource for yielding insights into the diversity among wheat genomes

Nathan S. Watson-Haigh<sup>1,2\*</sup> , Radosław Suchecki<sup>1,3</sup>, Elena Kalashyan<sup>1</sup>, Melissa Garcia<sup>1</sup> and Ute Baumann<sup>1</sup>

## Abstract

**Background:** Democratising the growing body of whole genome sequencing data available for *Triticum aestivum* (bread wheat) has been impeded by the lack of a genome reference and the large computational requirements for analysing these data sets.

**Results:** DAWN (Diversity Among Wheat geNomes) integrates data from the *T. aestivum* Chinese Spring (CS) IWGSC RefSeq v1.0 genome with public WGS and exome data from 17 and 62 accessions respectively, enabling researchers and breeders alike to investigate genotypic differences between wheat accessions at the level of whole chromosomes down to individual genes.

**Conclusions:** Using DAWN we show that it is possible to visualise small and large chromosomal deletions, identify haplotypes at a glance and spot the consequences of selective breeding. DAWN allows us to detect the break points of alien introgression segments brought into an accession when transferring desired genes. Furthermore, we can find possible explanations for reduced recombination in parts of a chromosome, we can predict regions with linkage drag, and also look at diversity in centromeric regions.

**Keywords:** Wheat, Diversity, Variants, Genome, Visualisation

## Background

With advances in technology and the reducing costs of high-throughput sequencing, it has become feasible to sequence the large ( $\approx 17$  Gbp), polyploid genome of bread wheat (*Triticum aestivum*) and resequencing projects have been undertaken [1] or are currently underway [2, 3]. Ahead of these whole genome sequencing projects, data from thousands of sequenced exomes has become available, predominantly from TILLING populations [4, 5]. While exome capture provides a means to sequence and analyse many more individuals by significantly reducing the sequencing space [6], it is limited to the coding regions for which probes have been designed and is sensitive to GC content. As a result, the coverage of coding regions by exome capture has been shown to be inferior to whole genome sequencing [7]. Furthermore, whole genome sequencing is not only more powerful

in detecting exome variants [8] but is capable of capturing structural variation and non-exonic variants [9]. However, this comes at the cost of significantly more resources, not only in terms of sequencing but also the analysis.

Genetic diversity is generally estimated at the population level from SNP data and provides information on the amount of genetic diversity between individuals at the whole genome level, but not on its distribution within the genome. However, genetic diversity goes beyond SNPs and includes indels, introgressions and other structural variation such as copy-number-variation (CNV). These are all known to be important drivers of diversity. Introgressions are often the result of wide crosses with landraces, wild relatives or related species such as rye. These donor species are often more resilient and are good sources of tolerance to various diseases and abiotic stresses such as heat and drought, and have been used extensively in wheat breeding [10, 11]. The ability to access and visualise genetic diversity in detail, from whole chromosomes to individual genes, will enable a better understanding and utilisation of the

\*Correspondence: [nathan.watson-haigh@adelaide.edu.au](mailto:nathan.watson-haigh@adelaide.edu.au)

<sup>1</sup>School of Agriculture, Food and Wine, University of Adelaide, PMB 1, 5064 Glen Osmond, SA, Australia

<sup>2</sup>Bioinformatics Hub, School of Biological Sciences, University of Adelaide, SA 5005 Adelaide, Australia

Full list of author information is available at the end of the article



available diversity in a region of interest, irrespective of scale.

While whole genome sequencing resources are available to the community, their wider utility has been impeded due to several factors: 1) the lack of a high quality, contiguous reference genome assembly and gene annotations; 2) the large computational resource requirements for analysing these data sets effectively and 3) the tools for making that information available in a way that would allow breeders to access information on potentially useful genetic diversity. Addressing these issues is key to closing the gap between research and some of the challenges in plant breeding. The release of the IWGSC RefSeq v1.0 assembly [12], and associated annotations, addresses the first of these issues. Through DAWN (Diversity Among Wheat geNames) we address the other two issues.

### Implementation

DAWN provides convenient access to almost 1 terabyte of precomputed data. This is derived from a Whole Genome Shotgun (WGS) sequencing project of 17 wheat accessions, including Chinese Spring (CS), 62 exome captures, and RNA-seq data from several tissues and developmental stages from CS. For convenience, it also incorporates annotation data released with the CS IWGSC RefSeq v1.0 assembly, including gene annotations and marker location information. This resource is accessible to the wheat community through a JBrowse [13, 14] genome browser interface hosted at <http://crobiad.agwine.adelaide.edu.au/dawn/>.

### Reference genome and annotations

DAWN uses the *Triticum aestivum* CS IWGSC RefSeq v1.0 genome assembly as the reference genome [12]. To our knowledge, no currently available genome browsers, including UCSC [15], Ensembl [16], IGV [17], Tablet [18] or JBrowse [13, 14], support the CSI indexing schema. Therefore to enable the visualisation of read alignments from BAM files and variant calls from VCF files we have used a version of the reference genome where each pseudomolecule had been split into two smaller “parts”. To facilitate the conversion of coordinates between full-length pseudomolecules and these “parts” we have developed an online conversion tool (<http://crobiad.agwine.adelaide.edu.au/dawn/coord/>). IWGSC RefSeq v1.0 is accompanied by several GFF3 files which describe the physical location of gene models (v1.0 and v1.1), transposable elements and markers from various platforms (e.g. Illumina Infinium iSelect 90K and 9K SNP chip markers, DArT markers and several EST and SSR data sets). We pre-processed these files to transform coordinates to match the pseudomolecule “parts” (Additional files 1 and 2) and merged functional annotations into the v1.0 high and low confidence gene models

(Additional files 3 and 4). Resulting GFF3 files were validated using GenomeTools v1.5.9 [19].

The CS IWGSC RefSeq v1.0 derived annotation tracks are available to the user under “Annotations/RefSeq v1.0” while the various marker data sets are available as tracks under “Markers”. In addition, the v1.1 gene annotations are available to the user under “Annotations/RefSeq v1.1”. The data from these tracks have been indexed, making genes and markers searchable by name, via the location/search box.

### WGS resequencing data

Whole Genome Shotgun (WGS) resequencing data from 16 bread wheat accessions was obtained from Bioplatforms Australia (BPA) [1, 20] in addition to Chinese Spring WGS Illumina data from the ENA (accession number PRJNA392179) [21]. The data was aligned to the reference genome using minimap2 v2.10 [22]. Alignments with a MAPQ <5 were removed and then further processed to create several data tracks per accessions for visualisation. Approximately 29-50% of raw reads failed to align with a MAPQ  $\geq 5$  while 38-56% of raw reads aligned with a MAPQ  $\geq 30$  (Additional file 5) These tracks are available to the user under “Resequencing/Whole Genome Shotgun (Illumina)”. The BPA panel includes 11 Australian accessions: Baxter, Chara, Drysdale, Excalibur, Gladius, H-45, Kukri, RAC-875, Westonia, Wyalkatchem and Yitpi, and 5 Northern Hemisphere accessions: AC-Barrie (Canada), Alsen (USA), Pastor (CIMMYT), Volcani-DD-1 (Israel) and Xiaoyan-54 (China). All but Xiaoyan-54 are spring wheat.

For each accession the user can access: 1) “Coverage” tracks for visualising read coverage depth patterns at Kbp to Mbp scales. These show the mean coverage (yellow line) as well as 1 and 2 standard deviations (grey background shading). Regions with read coverage  $> 2 \times SD$  from the mean were extracted, then merged if  $\leq 500$  bp apart and reported if  $\geq 5$  kbp (above the mean) or  $\geq 50$  kbp (below the mean) in length (Additional file 6). 2) “Read Alignment” tracks for visualising individual read alignments and alignment mismatches at the 100’s of bp scale. 3) “SNP Coverage” tracks for highlighting mismatches between the read alignments and the CS reference. Vertical lines within the read coverage plot indicate the proportion of reads with mismatches to the CS reference and teardrops shown below the coverage track indicate those positions exceeding 90% alternative bases and at  $\geq 3$  reads coverage. This track is particularly useful for identifying haplotype blocks at the Kbp scale. Most tracks transition to a read coverage depth or variant density plots at the Kbp-Mbp scale, when the density of information is too high to be visually meaningful.

### Exome capture data

Public exome capture data for 62 accessions was obtained from the ENA (accession number PRJNA227449) [23] and aligned to the reference genome using the same approach as the WGS data described above. These tracks are available to the user under “Resequencing/Exome (Illumina)”. The data set comprises 6 breeding lines, 29 cultivars, 26 landraces and 1 synthetic hexaploid. As with the WGS resequencing data, “Coverage”, “Read Alignment” and “SNP Coverage” tracks are available for each accession.

### Variant calls and variant density

Variant calling was performed for each accession in the WGS resequencing and exome capture data sets, using a SAMtools v1.8 [24] and BCFTools v1.8 calling pipeline, and is accessible under the “Variant Calls” tracks. These tracks show variant positions as vertical bars, coloured according to the alternative allele. Positions that were reported as indels or triallelic are displayed in black. Variants were classified as either high quality homozygous (PASS), low quality homozygous (LowQualHom), high quality heterozygous (Het) or low quality heterozygous (LowQualHet). A summary of the number of variants for each WGS data set is presented in Table 1. By default, only the homozygous (PASS and LowQualHom) variant calls are displayed. However all classes of variants can be

toggled on/off using the “Hide sites not passing filter ...” available from the track label of “Variant Calls” tracks. These tracks are particularly useful for identifying haplotype blocks at the 10’s - 100’s Kbp scale, depending on variant density, and for marker development.

A higher-level visualisation of variant calls is provided as “Variant Call Density” tracks, calculated as the number of variant calls per 10 Kbp of non-overlapping windows. Regions with variant density  $>2*SD$  from the mean were extracted, then merged if  $\leq 40$  kbp apart and reported if  $\geq 500$  kbp in length (Additional file 7). When used in concert with the read “Coverage” tracks at the multi-Mbp, it provides a way to differentiate genomic regions which are CS-like (good read coverage and low variant density) from those which are more divergent from CS (good read coverage and high variant density).

### Gene expression data

Chinese Spring RNA-derived data was obtained from URGI [25]. Briefly, it comprised of 5 tissues (grain, leaf, root, spike and stem) at 3 different developmental stages and in 2 replicates. For each tissue and developmental stage we aligned the reads to the reference genome using STAR v2.6.0c [26] and provide access to the resulting data via tracks under “Expression/IWGSC/RNA-seq” for visualising the read alignments which transition to read coverage depth plots at the 10’s of Kbp scale. A summary of coverage profiles is also available for each tissue (under “Coverage Summary”) to help identify tissue-specific expression patterns using a smaller number of tracks. Unlike other gene expression resources (e.g. Wheat Expression Browser [27]) the information in these tracks cannot be directly compared across different samples (no normalisation performed). However, it still provides an insight into whether genes are potentially expressed and if this may be tissue or stage specific.

### Optimising data for JBrowse tracks

Due to the large size of the wheat genome and the data sets used, the size of index files can become quite large (e.g.  $\approx 40$  megabytes for each of the 16 WGS BAM files). Large index files can negatively affect the responsiveness of DAWN, especially when viewing many tracks simultaneously. Before JBrowse can render data for a region of a track to be viewed, it potentially has to download 100’s megabytes of index files. Fortunately, JBrowse offers a feature whereby it loads different index files depending on the currently loaded reference sequence (i.e. chromosome part). To take advantage of this feature, we split BAM, VCF and bigWig files into 43 chromosome parts and index these separately. As a result, the BAM indexes are on average a few hundred kilobytes in size and less than 1.3 megabytes per chromosome part. This reduces the delay until the data is rendered.

**Table 1** Number of variants (millions) called per accession from WGS data (% indel calls)

Accession	PASS	Het	LowQualHom	LowQualHet
ACBarrie	15.51 (4.0%)	3.87 (2.1%)	4.85 (6.1%)	3.06 (2.0%)
Alsen	22.90 (3.9%)	2.14 (3.9%)	5.15 (8.2%)	3.31 (3.3%)
Baxter	26.44 (4.6%)	4.99 (2.9%)	3.91 (8.8%)	5.05 (2.4%)
Chara	21.59 (4.6%)	4.31 (2.6%)	3.77 (8.5%)	7.16 (1.4%)
Drysdale	23.29 (4.3%)	2.29 (3.4%)	3.67 (10.1%)	1.80 (3.5%)
Excalibur	22.48 (3.9%)	2.75 (3.1%)	3.69 (9.6%)	5.52 (2.3%)
Gladius	23.90 (4.3%)	2.48 (3.1%)	3.41 (10.8%)	2.53 (2.4%)
H45	20.94 (4.2%)	2.88 (2.9%)	3.76 (9.4%)	4.59 (1.6%)
Kukri	17.21 (3.9%)	4.43 (4.1%)	3.13 (8.9%)	2.69 (3.6%)
Pastor	19.52 (4.2%)	5.44 (2.8%)	4.21 (8.0%)	5.56 (2.2%)
RAC875	29.25 (5.0%)	3.23 (3.9%)	2.75 (11.2%)	2.22 (4.4%)
Volcanii	20.28 (4.0%)	2.81 (3.3%)	5.61 (7.2%)	6.87 (2.5%)
Westonia	20.49 (4.5%)	3.50 (3.7%)	4.11 (8.6%)	4.67 (3.2%)
Wyalkatchem	26.67 (5.0%)	2.69 (3.8%)	3.14 (9.4%)	1.73 (3.7%)
Xiaoyan	15.43 (4.4%)	6.30 (1.7%)	3.62 (8.3%)	11.90 (1.1%)
Yitpi	19.59 (4.6%)	7.84 (1.6%)	3.45 (9.3%)	11.77 (1.1%)
Chinese spring	0.09 (26.4%)	0.43 (6.7%)	0.04 (36.9%)	0.21 (8.8%)

The filter values PASS, Het, LowQualHom and LowQualHet correspond to high quality homozygous, high quality heterozygous, low quality homozygous and low quality heterozygous calls respectively. Filter values can be used to show/hide that particular type of variant in the “Variant Calls” tracks

### Processing pipeline

The processing pipeline underpinning the DAWN data was implemented using Snakemake v5.1.4 [13]. The 2,149 jobs were executed on a compute cluster containing 2 nodes, each with 72 Intel Xeon E5-2699 v3 CPUs (2.30GHz) and 770 gigabytes RAM. The analysis of the WGS and RNA-Seq data sets took 5.7 CPU years, had a peak memory usage of 300 gigabytes and generated over 11 terabytes of data files ( $\approx 800$  gigabytes are for JBrowse tracks). The commands used in the processing of this data are available in Additional file 8 with example commands and parameters in Additional file 9.

While users have access to over 800 gigabytes of data files, only a fraction of this is downloaded to a user's computer. This is made possible by JBrowse's ability to efficiently retrieve and locally cache information for relatively small subset of data.

### Results and discussion

Below we present examples to demonstrate the utility of DAWN in the investigation of genetic diversity among wheat genomes, opportunities for discovery of new alleles or introgression segments as well as its application for marker development and breeding strategies. The ability to visualise data from several accessions at once, together with gene expression data, marker information and gene annotations provides a powerful resource for investigating genetic diversity among wheat genomes.

### Introgressions

Alien introgressions can be easily spotted with DAWN as decreases in read coverage and probably an associated increase in variant density. For example, a wheat accession that contains an introgression fragment from a distantly related species would show few, if any, sequence reads aligned over the corresponding region of the CS reference genome. Even for the more conserved genic regions we may observe few aligned reads if the sequence divergence is too great for the aligner to accurately place. For introgressed portions of closer relatives, such as durum wheat, the reduction in read alignment affects the non-coding regions more dramatically. That is, we see higher numbers of variant calls in the intergenic regions than in the coding regions. The read alignment coverage and variant density tracks allow easy identification of putative deletions and alien introgressions; their approximate physical size can be inferred from CS. This information can provide valuable insights to QTL-cloning projects since the generation and screening of mapping populations may be unsuccessful if the region harbouring the gene of interest is placed within an introgression fragment and thus unlikely to generate informative recombinants.

### Stem rust locus *Sr36*

We observed a large region of chromosome 2B in Baxter which showed a consistently reduced read depth coverage and increase variant density compared to the rest of the genome (Fig. 1a and b). The region starts at  $\approx 89.5$  Mbp on chr2B\_part1, spans the centromeric region, and ends at  $\approx 304.3$  Mbp on chr2B\_part2. This corresponds to  $\approx 668$  Mbp (83%) of the chr2B pseudomolecule and contains 4445 high-confidence gene models. Across this region, we observed increases in read coverage around genes together with increases in variant density (Fig. 1c). This suggested that while much of the intergenic space is very different in Baxter, the gene space is nevertheless similar to the CS reference genome.

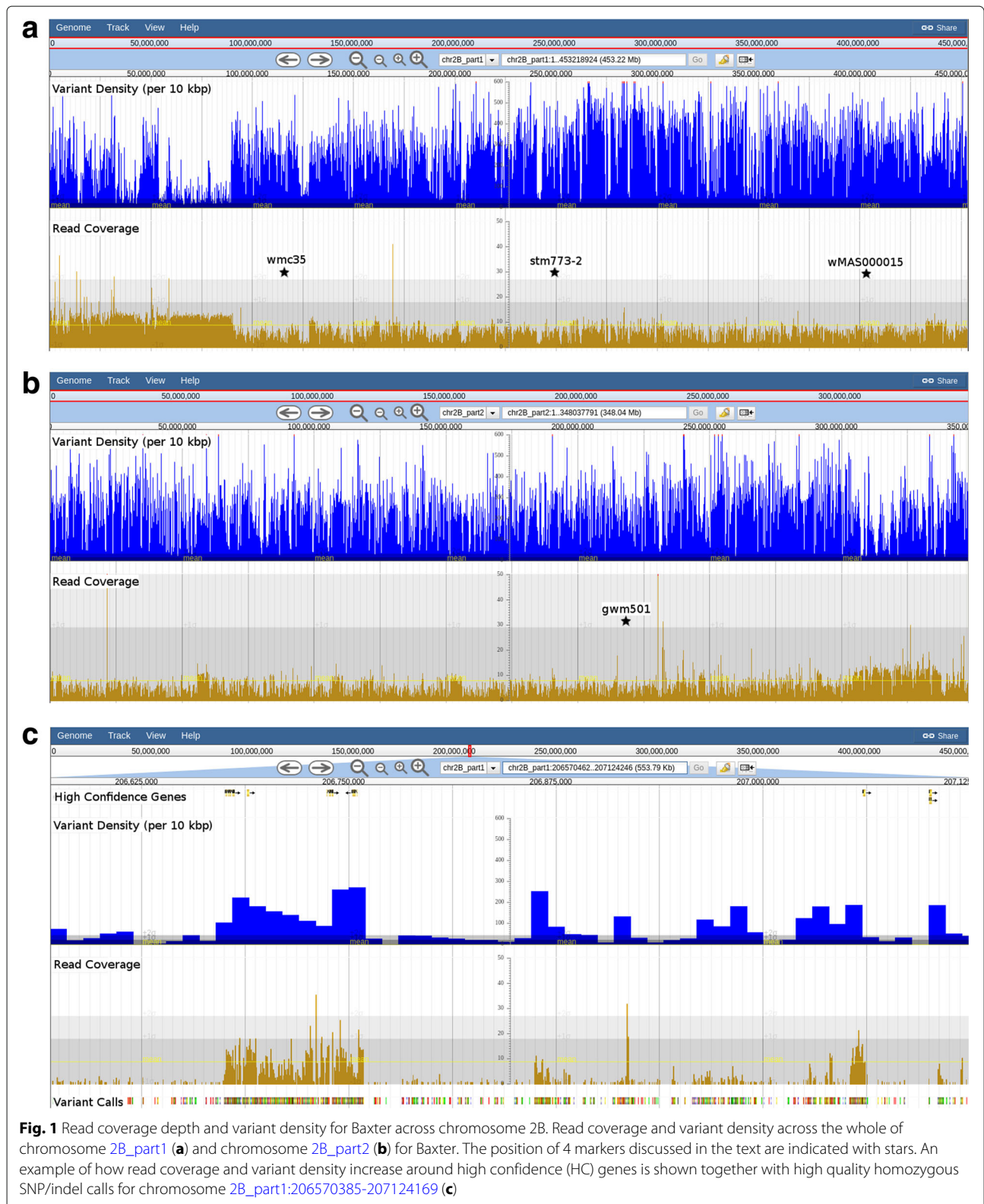
The stem rust resistance locus *Sr36*, located on chromosome 2B, is derived from *Triticum timopheevi* and confers resistance against many *Puccinia graminis* sp. *tritici* pathotypes [28]. The microsatellite marker stm773-2 has been found to be tightly linked to *Sr36* and the KASP marker, wMAS000015, is also available [29, 30]. The Australian cultivar, Cook, is derived from the hexaploid wheat CI-12633 which is one of several origins of the *T. timopheevi* *Sr36* introgression [28]. Cook has been used extensively in breeding programs and is a common source of *Sr36* in Australian wheat accessions, including Baxter. A revised genetic map of Sunco (derived from Cook and carrying *Sr36*) x Tasman (not carrying *Sr36*) suggested the *Sr36* translocation extends from marker wmc35 on the short arm, to marker gwm501 on the long arm [31]. By aligning the primer sequences for each of the four markers, to the reference genome, we were able to place wmc35, stm773-2 and wMAS000015 on chromosome 2B\_part1 at 113.3 Mbp, 249.7 Mbp and 406.3 Mbp respectively, while gwm501 was placed on chromosome 2B\_part2 at 218.9 Mbp (Fig. 1a and b). It has been shown that accessions carrying *Sr36* show no allelic diversity across most of chromosome 2B when compared to accessions which lack it [32], show segregation distortion [33] and linkage repulsion with *Sr39* [34]. This means that combining and introgressing new traits on chromosome 2B in lines possessing *Sr36* derived from CI-12633 will be difficult.

By looking at chromosome 2B in DAWN, we were not only able to find this introgressed region, but were able to delimit the area to a similar interval as previously determined by genetic mapping. However, the *Sr36*-derived gene(s) responsible for stem rust resistance remain elusive, especially given the size of this introgression.

### Root lesion nematode resistance tightly linked to yellow flour colour

Root lesion nematode (*Pratylenchus neglectus*) infections can cause significant yield losses and thus are a major problem for Australian wheat growers. Moderate





resistance has been described for two accessions in the BPA panel, Excalibur and Wyalkatchem, whereas other Australian accessions, including Kukri, Chara, Gladius

and Yitpi, are susceptible. The resistance has been attributed to the *Rlnn1* locus located in the terminal region of the long arm of chromosome 7A [35].

Jayatilake et al. [36] showed *Rlnn1* to be tightly linked to the *t* allele of *Psy-A1*, an allele highly associated with yellow pigmentation of the wheat grain. While yellow flour colour is a desirable trait for durum wheats, it is undesirable in bread wheat. Efforts towards separating the loci by recombination have so far been unsuccessful [37].

To investigate what DAWN can reveal about this region, we retrieved sequences for markers (*wri1*, *wri2*, *wri3*, *wri5* and *wPt-0790*) used by Jayatilake et al. [36] and placed these onto the genome by BLASTn. In doing so, we identified [TraesCS7A01G557300](#) as *Psy-A1*. The SNP/indel patterns and read alignment coverage clearly identified at least 4 different haplotypes, with Excalibur and Wyalkatchem being the most different from CS, as evident from few reads being aligned across this region and with higher variant density (Fig. 2). This lack of read alignment coverage extends from position  $\approx 272.1$  Mbp on chr7A\_part2 to the telomere, i.e. a  $\approx 14.6$  Mbp

long segment containing 233 high-confidence gene models (Fig. 3). Thus it appears likely that the *Rlnn1* carrying segment has been introgressed from a wild relative of wheat as a terminal substitution. Tight linkage and suppressed recombination observed between *Lr20/Sr15* (leaf rust resistance), *Pm1* (powdery mildew resistance) and *Rlnn1* are carried on this introgression and now form part of Excalibur's and Wyalkatchem's genomes [38]. Sequence differences between bread wheat and the alien introgression segment likely explains the observed suppressed recombination and the failed attempts to separate *Rlnn1* from the unfavourable *Psy-A1* allele over the last decade.

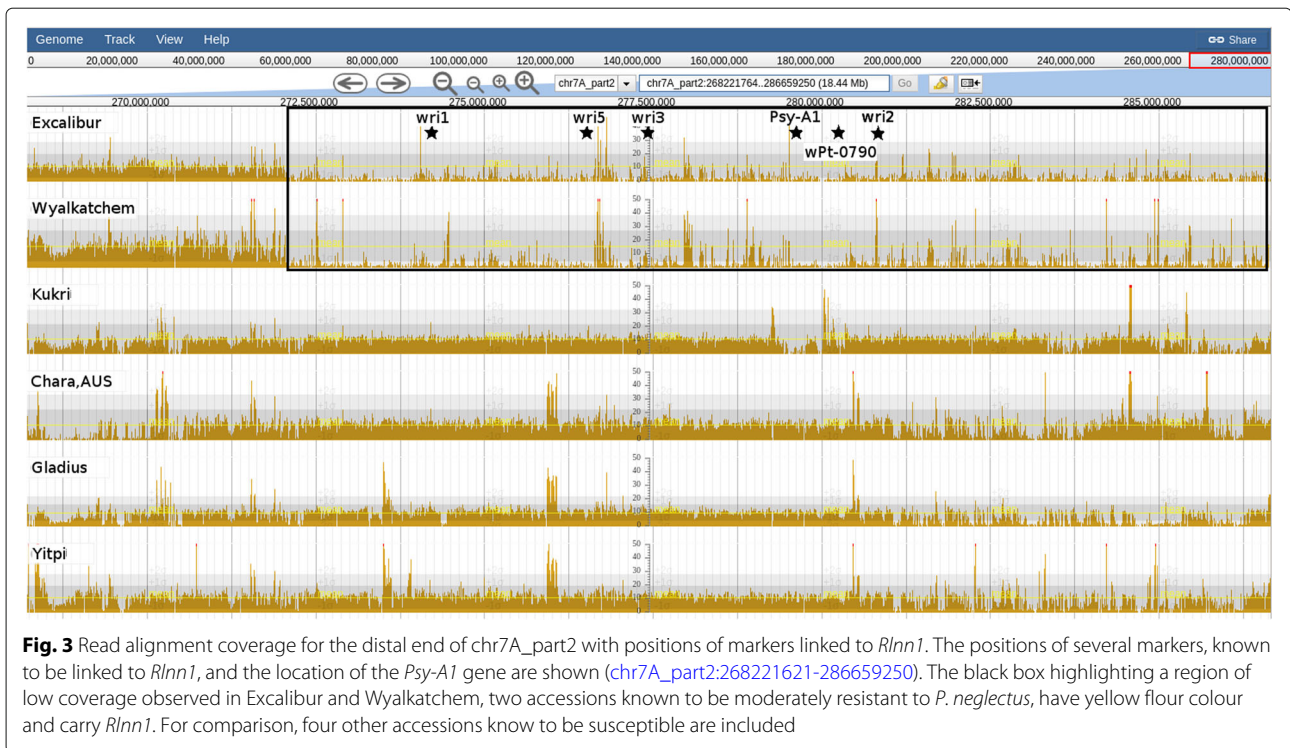
**Deletions**

***VRN-B1* deletion**

Spring and winter growth habit in hexaploid wheat is determined primarily by allelic variation in the *VRN-1* homeologues *VRN-A1* on chromosome 5A, *VRN-B1*



**Fig. 2** Genomic diversity around the *Rlnn1* marker *wri2*. SNP Coverage plots for the 16 BPA accessions have been sorted into 4 haplotypes based on the SNP/indel patterns in view ([chr7A\\_part2:281214721-281228448](#)). The position of the marker, *wri2*, is indicated by the star. Coloured drops hanging off the read coverage profile indicate the presence of putative SNPs/indels to the reference, A-green, T-red, C-blue, G-yellow, indel-grey



on chromosome 5B and *VRN-D1* on chromosome 5D [39, 40]. Briefly, hexaploid spring wheat have a deletion in the first intron of *VRN-B1* and/or *VRN-D1*. Spring types lacking these deletions are expected to have a *VRN-A1* promoter which differs from the recessive *vrn-A1* allele [39].

Using DAWN we were able to see that 12 of the 15 spring wheats in the BPA panel had a *VRN-B1* deletion, as indicated by a lack of read alignment coverage in those accessions (Fig. 4). Three of these 12 spring wheats have evidence for an  $\approx 8$  Kbp deletion (Pastor, Drysdale and Baxter) while the other 9 spring wheats seem to have an  $\approx 2.7$  Kbp deletion and the remaining  $\approx 5.3$  Kbp is significantly different to CS. Of the remaining three spring wheats (H-45, Chara and AC-Barrie) which lack a *VRN-B1* deletion, they also lack a *VRN-D1* deletion (Fig. 5). The variable read alignment coverage around *VRN-D1* and *VRN-A1* make it difficult to determine the precise combination of alleles at these loci (Fig. 6).

#### *Wx-B1* nulls

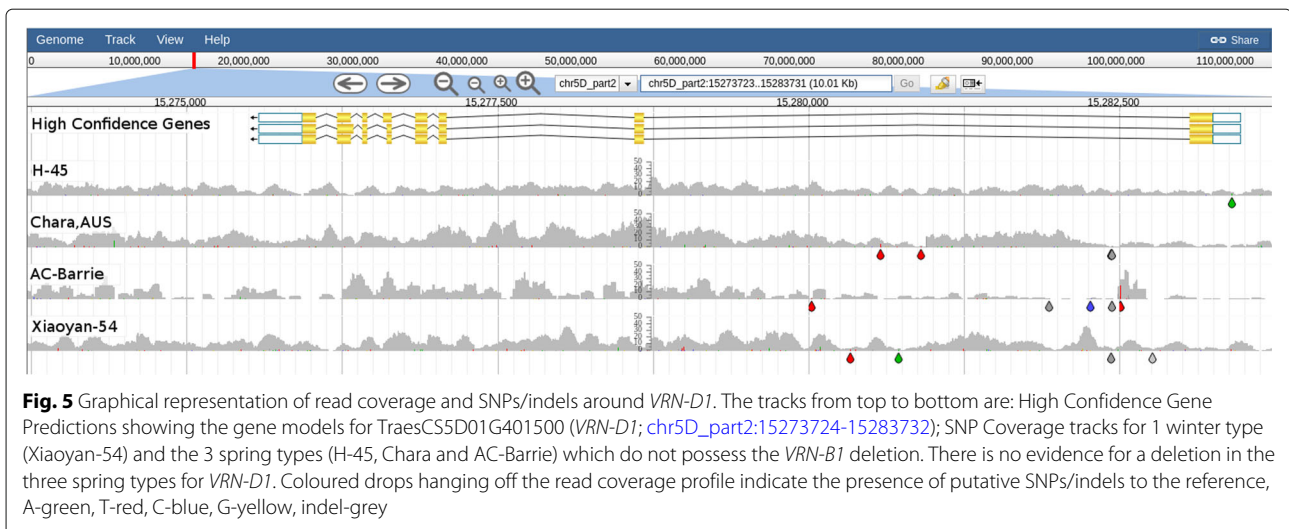
The ratio of the two main macromolecules, amylose and amylopectin, is closely related to the quality of starch in the wheat grain, with high amylose being associated with low noodle quality [41]. [TraesCS4A01G418200](#), also known as “waxy”, encodes a granular bound starch synthase. This gene is solely responsible for the synthesis

of amylose in wheat and has three homeologs: *Wx-A1* on chromosome 7A, *Wx-B1* on chromosome 4AL (7B translocation) and *Wx-D1* on chromosome 7D [42]. Null alleles of waxy genes have been described previously in a variety of studies, for instance when examining Mexican [43], Italian [44], Spanish [45] and 324 European accessions which included landraces and spelt wheats [46]. The identification of accessions with these null alleles have allowed the development of new lines with low amylose content. In all cases, *Wx-B1* appeared to be the most polymorphic locus. For a recent review on waxy proteins and their genes see Guzman & Alvarez 2016 [47]. Using DAWN we were able to see a deletion ( $\approx 8$  Kbp) in six accessions from the BPA panel (Alsen, Pastor, RAC-875, Westonia, Wyalkatchem and Yitpi). This deletion included the whole of *Wx-B1* (Fig. 7) as well as the 3' end of a neighbouring gene ([TraesCS4A01G418100](#)), annotated as coding for succinate dehydrogenase subunit 5. However, we are not aware of any phenotypic consequences resulting from the partial deletion of this gene. We also observed that *Wx-D1* was conserved across all 16 BPA accessions.

#### *TOM1* deletion

Higher plants have two strategies for the uptake of Fe(III) from the rhizosphere (Marschner et al. 1986). The grasses (including wheat, maize, rice and barley) secrete mugineic acid (MA) family phytosiderophores (PS) from their roots

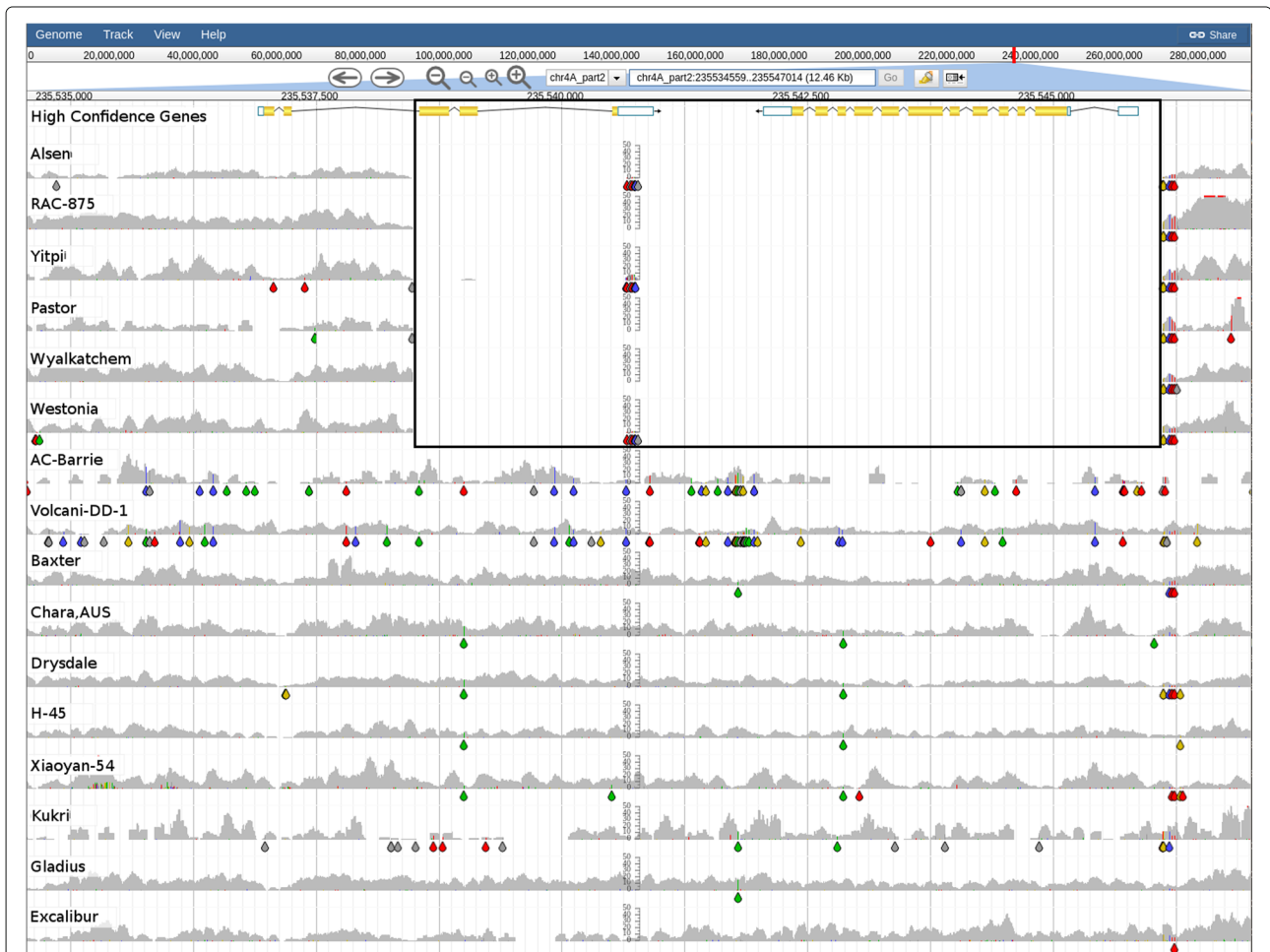








**Fig. 6** Graphical representation of read coverage and SNPs/indels around *VRN-A1*. The tracks from top to bottom are: High Confidence Gene Predictions showing the gene models for TraesCS5A01G391700 (*VRN-A1*;chr5A\_part2:134174264-134195143); SNP Coverage tracks for 1 winter type (Xiaoyan-54) and the 3 spring types (H-45, Chara and AC-Barrie) which do not possess either a *VRN-B1* or *VRN-D1* deletion. Putative SNPs/indels indicate the spring types share the same haplotype which differs from the haplotype of the winter wheat. Coloured drops hanging off the read coverage profile indicate the presence of putative SNPs/indels to the reference, A-green, T-red, C-blue, G-yellow, indel-grey



**Fig. 7** Graphical representation of read coverage and SNPs/indels around the *Wx-B1* locus. The tracks from top to bottom are: High Confidence Gene Predictions showing the gene models for TraesCS4A01G418100 (left) and TraesCS4A01G418200 (*Wx-B1*, right) and SNP Coverage tracks for the 16 BPA accessions (chr4A\_part2:235534559-235547014). Six accessions (Alsen, RAC-875, Yitpi, Pastor, Wyalkatchem and Westonia) show evidence of a deletion (boxed) as indicated by the lack of read coverage, while the remaining ten do not have a deletion. Coloured drops hanging off the read coverage profile indicate the presence of putative SNPs/indels to the reference, A-green, T-red, C-blue, G-yellow, indel-grey

into the rhizosphere to chelate and solubilise iron. These iron-PS complexes are taken up into the roots through specific transporters. Transporter of MAs 1 (*TOM1*) has been identified as the likely gene encoding the efflux transporter of 2'-deoxymugineic acid (DMA) in plants [48]. Rice contains five homologues of *TOM1*, two being in tandem with *TOM1* on chromosome 11 (*TOM2* and *TOM3*) and three others in tandem on chromosome 12 [49]. The function of *TOM3* and the chromosome 12 homologues have not yet been determined, but *TOM2* is thought to be involved in the translocation of metal ions inside the plant body [49].

We identified the *TOM1* homeologues: *TOM-A1* on chromosome 4A ([TraesCS4A01G187500](#)), *TOM-B1* on chromosome 4B ([TraesCS4B01G131400](#)) and *TOM-D1* on chromosome 4D ([TraesCS4D01G125900](#)). We saw that *Gladius* and *RAC-875* have a deletion ( $\approx 2.5$  Kbp) which spans the first three exons of *TOM1* and  $\approx 1$  Kbp of promoter region (Fig. 8). While this deletion would certainly lead to a *TOM-A1* null, it is not known if these two accessions are more susceptible to iron deficiency or if the *TOM-B1* or *TOM-D1* homeologues compensate for the absence of *TOM-A1*.

### Copy number variation (CNV)

Duplications of genomic loci are known to have played an important role in the evolution of plant genomes and have been linked to disease risk in humans [50]. While it is believed that CNV predominantly affects intergenic regions, there are known CNVs which affect protein-coding genes. For example, CNV has been linked to important traits such as flowering time, plant height and

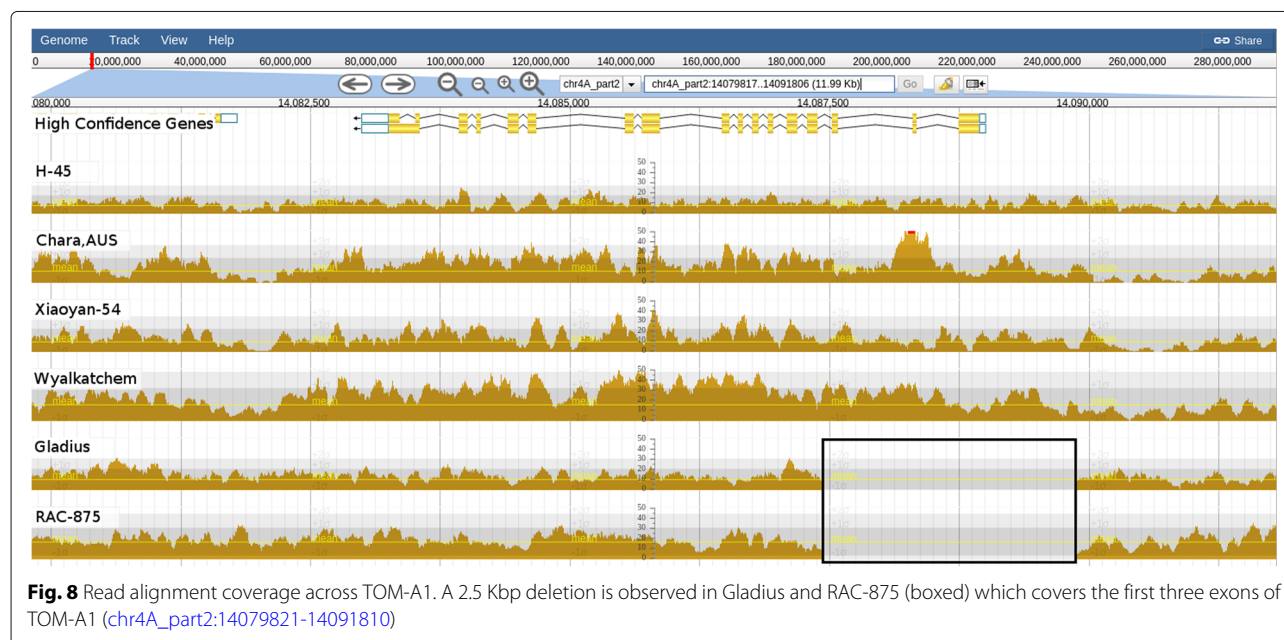
resistance to biotic and abiotic stresses, including boron tolerance in barley [51]. For a recent review of CNV in plants see Zmienko et al. 2014 [52].

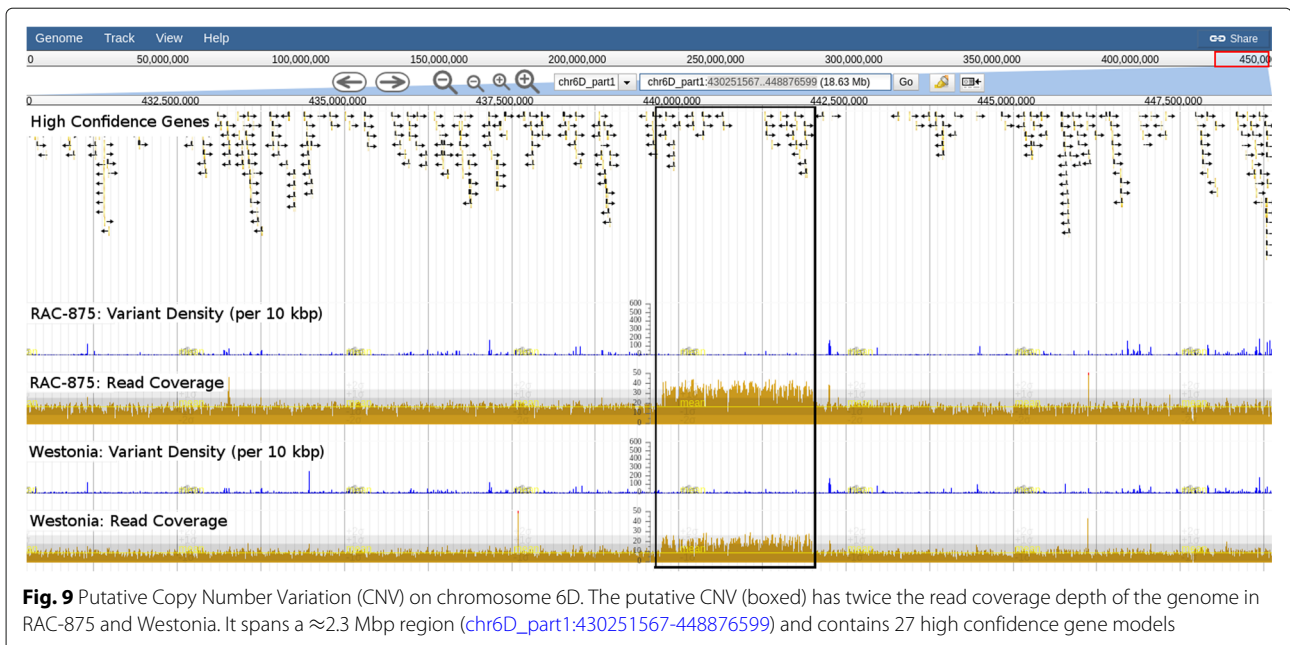
Using read coverage depth tracks it is possible to identify putative increases in CNV compared to the CS reference and to delineate the boundaries of the duplication. This is especially the case for the D genome where read coverage depth is less variable. One such example is a  $\approx 2.3$  Mbp region on chromosome 6D which shows an  $\approx 2$  fold higher coverage (and  $>2^*SD$ ) compared to the mean coverage of the rest of the genome. This putative CNV encompasses 27 high confidence gene models and is only observed for *RAC-875* and *Westonia* (Fig. 9). While Additional file 6 contains coordinates of regions with read coverage  $>2^*SD$  from the mean, we encourage those with special interest in CNV to analyse our data using the latest computational tools [53].

### Haplotypes

Haplotype blocks are usually defined using linkage disequilibrium (LD) estimated between pairs of markers. Methods to define haplotype blocks require the selection of somewhat arbitrary LD thresholds, especially for species where there is limited information on the extent of LD. DAWN allows haplotype blocks to be visualised using the distribution of SNPs/indels along the chromosomes of the 16 BPA accessions. Different haplotype alleles and recombination between blocks can also be observed.

The nucleotide polymorphisms visible in Fig. 10 allow grouping of the BPA accessions into five distinct haplotypes. Using the information carried in these SNP Coverage tracks, it is possible to see changes in haplotypes





as a result of recombination. Fig. 11 shows a region on chromosome 1A in the vicinity of [TraesCS1A01G013000](#), a gene annotated as being “disease resistance family protein”. The SNP pattern in the region immediately preceding this gene clearly shows that the seven accessions share the same haplotype. However, at about base 7,294,500 (near the 3' UTR of this gene) there is a putative recombination break point as evident from the 4 distinct haplotypes immediately following this position: 1) Drysdale; 2) AC-Barrie, Alsen, RAC-875 and Yitpi; 3) Baxter and 4) Chara. A second putative recombination break point at around base 7,300,000 results in additional haplotypes to give 6 distinct haplotypes following this position: 1) Drysdale; 2) AC-Barrie; 3) Alsen; 4) RAC-875 and Yitpi; 5) Baxter and 6) Chara.

As with the example in Fig. 10, we have observed a propensity for putative recombination break points to occur within close proximity of genes. This is consistent with observations made in yeast where the double-strand breaks, which are required for recombination, tend to occur 5' of genes near the promoters [54] and in maize where a recombination hotspot was located in the 5' transcribed region of the anthocyanin1 (*a1*) gene [55]. Similarly, a recent study of crossover events on wheat chromosome 3B showed a significant association of crossovers with genic features, particularly those which were expressed during meiosis [56].

### Centromeres

It has been known for many years that recombination events are unequally distributed along wheat

chromosomes such that their frequencies decrease from telomeres towards the centromeres [57–63]. More recently, Choulet et al. [64] partitioned the pseudomolecule for chromosome 3B with respect to centromere location, gene density and recombination rate and estimated that the centromere extended from 265 Mbp to 387 Mbp.

To explore whether identification of centromeric regions by visual inspection is possible with DAWN, and investigate the level of diversity across the centromeres among the 16 BPA accessions we first examined the variant call density and concomitant distribution of high and low confidence genes at a megabase scale. Although we expected the centromere to be contained within part1 of each pseudomolecule we analysed the complete length of the pseudomolecules (i.e. part1 and part2). While we observed a reduction of high confidence gene density for most chromosomes (for example Fig. 12), these could be subtle and did not allow us to demarcate the centromeric regions. However, using variant call density tracks we observed a lower number of changes in variant density in part1 compared to part2 of the pseudomolecules. Thus corroborating our expectation that the centromeres are located within part1 of the pseudomolecules.

To determine whether the visual observations indeed coincided with the centromeric regions of the pseudomolecules we analysed previously described centromeric sequences and their distributions along the chromosomes (Additional file 9). Once approximate borders of the putative centromeric regions were established we investigated conservation of haplotypes from the left border to the right border, employing the “Variant Calls” tracks



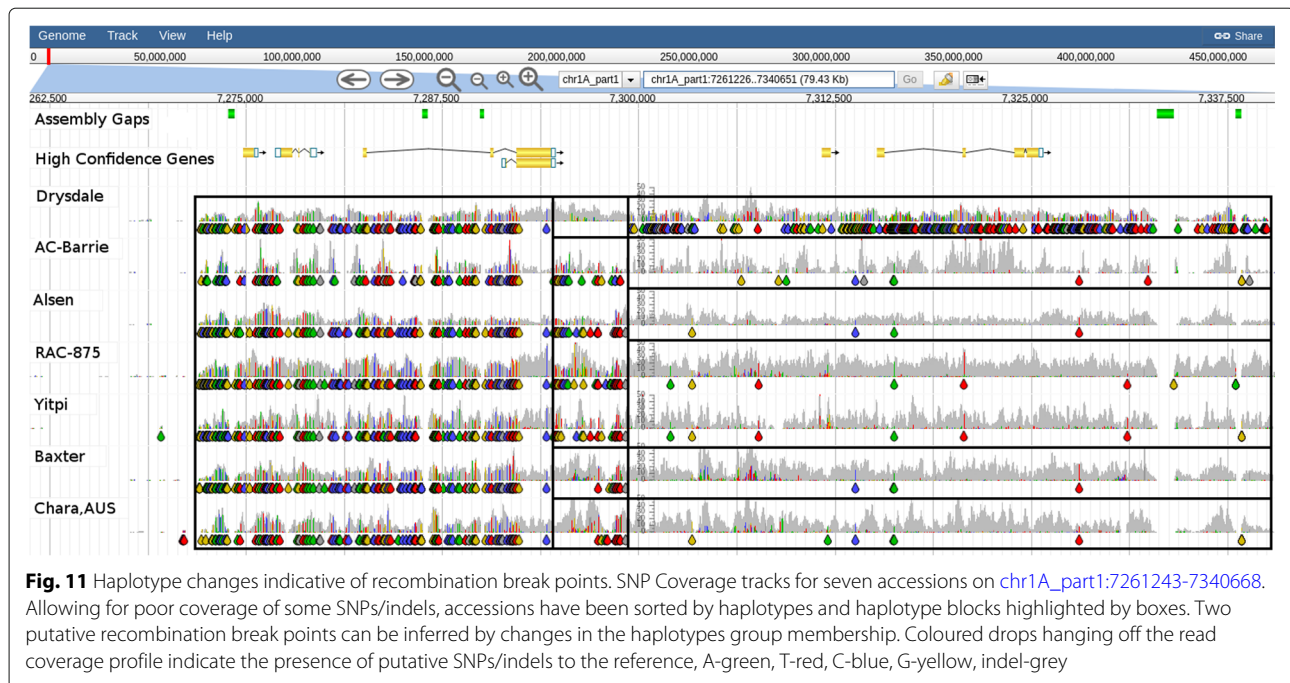
which allows visualization of several 100 Kbp and the “SNP Coverage” tracks for visualizing variants together with read coverage at scales up to 30 Kbp.

Table 2 shows the results of distribution of centromeric sequences and the investigation of the haplotypes across the delimited regions. For a BPA accession to be assigned to a haplotype group it had to be clearly identical with other genotypes. When an accession was similar to another but showed occasional additional variants we did not assign it to a group. In cases where a cultivar was found to be distinct from all other haplotypes, it would form its own single member group, for instance for chromosome 2B, Baxter, AC-Barrie and Xiaoyan-54 all have unique haplotypes. We found varying numbers of discrete haplotypes and groupings for most chromosomes, with the conspicuous exception of the centromeric region of chromosome 3B; all 16 BPA cultivars displayed the

same haplotype which was clearly distinct from Chinese Spring. As shown in Table 2 our approximate positioning of the centromere borders occasionally extended beyond observed changes in haplotypes. While a more accurate determination of the centromere regions awaits experimental verification, the fact that the observed haplotype changes are close to our predictions are encouraging.

The discovery of a single shared haplotype for the centromeric region of 3B present in the BPA accessions is peculiar, since these accessions originate from different regions of the world. Horvath et al. [65] calculated that chromosome 3B had a lower diversity than average for the entire B-genome, but their finding was based on markers located along the whole chromosome, and was not observed in other diversity studies [66, 67]. Cubizolles et al. [68] corroborate our results, they also found only two haplotypes, with the minor haplotype being present





in mostly Asian derived lines (possibly the Chinese Spring haplotype). Wheat breeding can impose strong selection pressures in favour of loci encoding disease resistance, contributing to yield or quality [69]. Whether this could be the cause for the low haplotype diversity observed at the 3B centromere warrants further investigation.

### Marker development

QTL positional cloning projects rely on the development of new polymorphic markers that are then used to screen the population under investigation for informative recombinants. Information on SNP/indel positions is the starting point for the design of new high-throughput markers such as KASP™ [70]. DAWN facilitates marker design by providing SNP/indel positions among the 16 BPA accessions as well as allowing the visualisation of previously developed markers such as the 90K SNP array and the 820K Axiom arrays.

The sequences of the markers flanking a QTL can be aligned to the CS IWGSC RefSeq v1.0 using BLASTn to find their position in the reference genome. Alternatively, if the flanking markers are among the data sets already included in DAWN (e.g. Illumina Infinium iSelect 90K and 9K SNP chip, 820K Axiom array), they can be easily located through a search in the DAWN interface. By visualising the QTL interval in DAWN, one can obtain information on the number of predicted genes, the number of haplotypes among the 16 BPA accessions and the SNPs/indels present in the region. For large QTL intervals, markers can be designed and spaced based on the

knowledge of the haplotype blocks present in the area, reducing the number of markers to be developed. In the case of small QTL intervals, markers can be designed to target genes in the region or even specific regions of target genes. SNP/indel positions more likely to be polymorphic can also be selected based on the frequency of a particular allele among the 16 BPA accessions.

### Limitations

#### Read alignments

Deletion and regions which are highly divergent from Chinese Spring may be almost devoid of read alignments. As such, users are encouraged to investigate further which of these situations is likely to be true. One option is to look for the existence of genes, which fall within such regions, in the wheat pangenome [71, 72].

Short read Illumina data is the predominant sequencing data available today. While Illumina data can be produced at sufficient volumes for sequencing wheat genomes and transcriptomes, it also contains inherent biases and limitations mainly due to the nature of short reads and GC biases [73]. Therefore, regions with high GC bias tend to be under-represented in terms of read coverage depth for both WGS and RNA-Seq data sets, reducing the power to detect variation between accessions.

#### Interface

Low and uneven read coverage depth in data sets used for calling variants, leads to missing data in the “Variant Calls” tracks and can mislead the user. However, used together



with the “SNP Coverage” or read coverage depth tracks, one can mitigate this risk by also considering the read coverage depth at a variant site. Care should be taken in overinterpreting the “SNPs” shown in a “SNP Coverage” track as they are based on read alignment mismatches rather than robust variant calling. In addition, they are rendered on-the-fly and so do not scale well, particularly when visualising many accessions over large physical distances. In order to improve the visualisation of coverage and variant information we would look to develop a new JBrowse track capable of rendering read alignment coverage from a BigWig file while superimposing variant information from a VCF file.

### Reference genome to pan-genome

In addition to the IWGSC RefSeq v1.0, two further CS assemblies have been released [21, 74], both of which are a significant improvement over the 2014 published

Chromosomal Survey Sequences [75]. We chose IWGSC RefSeq v1.0 as the reference sequence for DAWN because of the availability of pseudomolecules which facilitate identification of diversity at the chromosomal level. We envision that a future consolidation of all CS assemblies will resolve discrepancies and fill gaps, leading to a single CS reference for the community. Until then, the existence of multiple genome assemblies presents a challenge for existing wheat resources [76–79] as it demands a decision on which assembly to use as the reference, or alternatively, to consider all assemblies as a reference. Furthermore, WGS, RNA-Seq and exome capture data will require reprocessing to leverage these improvements and highlights the importance of data sets conforming to FAIR Data Principles [80, 81].

As the number of resequenced genomes increases, the benefits of using a “pan-genome” to represent the genomic repertoire of all the sequenced accessions becomes more

**Table 2** Location of centromeres and number of haplotype groups present in BPA accessions

Chr	Start (Mbp)	End (Mbp)	Size (Mbp)	# Haplotype groups	Haplotype block break points (Mbp)	Groups
1A	160	260	100	2	Xia: 256.967 Exc: 258.867	Group1: Xia, Yit, Exc Group2: Kuk, Gla, Wyal, Wes, Bax, H45, Cha, Dry, Pas, RAC, ACB, Vol
1B	180	280	100	5		Group1: Xia Group2: Vol Group3: ACB, Exc, Als, RAC Group4: Kuk, Gla, Wyal, Yit, Wes Group5: Bax, Pas, Cha, Dry
1D	146	196	50	2		Group1: Cha, Dry, Exc, Wes, Vol Group2: H45, Pas, Wyal, RAC, Xia
2A	270	370	100	3		Group1: ACB, Als, Bax, Dry, Exc, H45, Pas, Kuk, Wyal Group2: Cha, RAC, Yit, Gla, Wes, Vol Group3: Xia
2B	270	385	115	7	Yit: 379.837	Group1: Bax Group2: ACB Group3: Xia Group4: Wes, Cha, Als Group5: RAC, Yit, Gla, Exc Group6: Pas, H45, Dry Group7: Vol, Wyal, Kuk
2D	220	300	80	3		Group1: Yit, Xia, Als, H45 Group2: Kuk, Cha, Dry, Pas, RAC, Gla, Exc Group3: Bax, Wes, Wyal
3A	240	370	130	4		Group1: Yit, Kuk, Xia, Dry Group2: Cha, H45, Bax, Wes, Wyal Group3: Gla, RAC Group4: Vol, Exc
3B	280	390	110	1		All
3D	200	280	80	2	Bax: 276.588	Group1: Yit, Wes, Bax, Als, Cha, Wyal, Exc Group2: Dry, Gla, RAC
4A	245	360	115	3		Group1: Pas, Wes, Yit, H45, Bax Group2: Als, Cha, Dry, Wyal, Exc Group3: RAC, Gla
4B	270	375	105	6		Group1: ACB Group2: RAC, Dry, Kuk, Exc, Wyal Group3: Bax, H45, Xia Group4: Wes, Gla, Cha Group5: Als Group6: Yit
4D	194	226	32	4		Group1: RAC, Dry Group2: Xia, H45 Group3: Wyal, Bax Group4: Cha, Vol
5A	185	310	125	4		Group1: Als Group2: Vol Group3: Wyal, Gla, Bax, Dry Group4: Yit, Cha, H45, Xia
5B	145	240	95	3		Group1: RAC, Cha, Exc, H45, Dry, Als Group2: Wes, Bax, Wyal Group3: Yit, Gla
5D	145	245	100	2	RAC: 244.646 Gla: 244.646	Group1: RAC, Gla, H45, Dry, Wes Group2: Exc, Wyal
6A	240	355	115	3		Group1: Bax, Cha, H45, Wyal, Wes, Xia Group2: Kuk, Als, Dry, Pas, ACB, Yit Group3: RAC, Exc, Gla, Vol

**Table 2** Location of centromeres and number of haplotype groups present in BPA accessions (Continued)

Chr	Start (Mbp)	End (Mbp)	Size (Mbp)	# Haplotype groups	Haplotype block break points (Mbp)	Groups
6B	270	360	90	2	Exc: 345.417	Group1: Kuk, Xia, Exc, Dry, Wes, Yit Group2: Bax, Vol, Als, H45, Pas, Cha, Wyal, RAC, Gla
6D	184	230	46	3		Group1: Vol, Als, Dry, H45 Group2: Cha, Kuk, Wes, Wyal Group3: RAC, Yit, Gla
7A	300	410	110	3		Group1: Vol, ACB, Pas, Wes Group2: Exc, Kuk, Bax, Cha, Dry, H45, Xia Group3: Wyal, RAC, Yit, Gla
7B	250	315	65	5		Group1: Vol Group2: Kuk, Als Group3: Yit, Wyal, Exc, Gla, RAC, Wes Group4: Dry, H45, Pas, Xia Group5: Bax, Cha
7D	300	385	85	4		Group1: Kuk, Xia, Wyal Group2: RAC, Exc, Gla Group3: Yit, H45, Als, Wes Group4: Dry, Bax, Cha (Pas)

apparent [82, 83]. Initial attempts to create a wheat pan-genome have focused on supplementing an existing genome assembly with contigs assembled from reads which had failed to align to the reference genome [71]. Similarly, the rice pan-genome (RPAN) developed for the 3,000 Rice Genome Project (3K RGP) performed a *de novo* assembly of each individual and then removed redundancy to derive a set of sequences that constitute the rice pan-genome [84]. While these approaches provide a convenient linear representation of a pan-genome, it cannot easily be extended to iteratively include newly sequenced genomes without significant effort. While graph representations of the pan-genome are an attractive way to represent the genetic variation which exists between individuals, their utility is hampered by the lack of tools which can utilise these data structures for read alignments, variant calling and visualisation [85]. We expect the future will see a paradigm shift away from linear representations of a (pan-)genome to more sustainable graph based representations [86].

### Conclusions

Through DAWN we have removed the burden of analysing the whole genome sequencing data of 16 accessions and made these valuable data sets easily accessible to the community through a JBrowse interface in the context of the newly released IWGSC RefSeq v1.0 genome assembly and annotations. By providing examples, we have shown how DAWN can be utilised by researchers to: a) discover diversity of different types among genomes; b) find explanations for reduced recombination; c) identify markers for tracking important traits and d) identify candidate genes under QTL. The example where we explored possible explanations for the tight linkage between *Rlnn1* and yellow flour colour shows the

power of visualisation of genomic data at different resolutions; it took little time to place the wri-markers onto chromosome 7A and find a possible explanation for experimental observations. Moreover, DAWN could also be used in a predictive way, accelerating research direction and discovery. As more wheat genomes are resequenced and other genomic resources are made available, we can make these available through DAWN. Our processed data sets, which include BAM and VCF files, are made freely available for others to use and explore.

### Availability and requirements

We provide convenient access to the computed data files via a JBrowse interface available via <http://crobiad.agwine.adelaide.edu.au/dawn/>.

- **Project name:** Diversity Among Wheat geNomes (DAWN)
- **Project home page:** <http://crobiad.agwine.adelaide.edu.au/dawn/>
- **Operating system(s):** Platform independent
- **Other requirements:** HTML5 compatible web browser
- **License:** GNU GPL

### Additional files

**Additional file 1:** IWGSC RefSeq v1.0 annotation preprocessing shell script. Shell script to download, validate and modify IWGSC RefSeq v1.0 annotation files, including splitting GFF3 files based on boundaries defined by the IWGSC. (SH 9 kb)

**Additional file 2:** Perl helper script for splitting GFF3 files. Perl script for splitting a GFF3 file at chromosomal coordinates defined in a BED file. This converts GFF3 coordinates from full length pseudomolecules to their "parts" and is called by the preprocessing shell script (Additional file 1). (PL 3 kb)



**Additional file 3:** Script for merging functional annotations into gene models. Shell script to merge functional annotation information into the High and Low confidence gene model GFF3 files. (SH 2 kb)

**Additional file 4:** Awk helper script for creating correctly formatted GFF3 attribute information. Awk script for generating a file of correctly formatted GFF3 attributes of functional annotation information. It is called by the shell script for merging functional annotations (Additional file 3). (AWK 1 kb)

**Additional file 5:** MAPQ distribution table. Table showing MAPQ distribution for each WGS accession. Number of aligned reads with a given MAPQ are shown together with the cumulative sum of reads  $\geq$  to a given MAPQ (expressed both as a percentage of the aligned read as well as the raw reads). (TAB 28 kb)

**Additional file 6:** ZIP file containing BED5 format files of high and low read coverage regions. BED5 format file for each WGS accession. The score column (5th column) indicates if the feature (row) represents a region which has  $>2*SD$  above (High) or below (Low) the mean read coverage. (ZIP 1455 kb)

**Additional file 7:** ZIP file containing BED5 format files of high and low variant density regions. BED5 format file for each WGS accession. The score column (5th column) indicates if the feature (row) represents a region which has  $>2*SD$  above (High) or below (Low) the mean variant density. (ZIP 526 kb)

**Additional file 8:** Script to generate data files required for the DAWN data tracks. A script containing all the commands needed to generate all the data files used in DAWN. (SH 9667 kb)

**Additional file 9:** Materials and Methods. A description of materials and methods used. (PDF 229 kb)

## Abbreviations

BAM: A binary representation of the sequence alignment map (SAM) file format; CS: Chinese Spring; DAWN: Diversity among wheat geNomes; EST: Expressed sequence tag; IWGSC: International wheat genome sequencing consortium; SNP: Single nucleotide polymorphism; SSR: Simple sequence repeat (a.k.a. microsatellite, STR: short tandem repeats); VCF: Variant call format; WGS: Whole genome shotgun

## Acknowledgements

The authors thank the International Wheat Genome Sequencing Consortium (IWGSC) for pre-publication access to IWGSC RefSeq v1.0. This initiative was supported by data generated using funding from Bioplatforms Australia through the Australian Government National Collaborative Research Infrastructure Strategy (NCRIS). This research was supported by use of the Nectar Research Cloud, a collaborative Australian research platform supported by the National Collaborative Research Infrastructure Strategy (NCRIS). The authors acknowledge that this research used high-performance computing, cloud and storage services provided by eResearch SA Ltd (eRSA). The authors would like to thank Margaret Pallotta and Elise Tucker for their contribution in selecting examples on which to report and Margaret's help reviewing the manuscript. We also thank the following early adopters/users of DAWN: Pauline Thomelin, Vanessa Melino, Alberto Casartelli, Muhammad Ahsan Asif and Million Erena.

## Funding

This work was funded by the Australian Research Council, the South Australian government, the Grains Research and Development Corporation and the University of Adelaide.

## Availability of data and materials

- DAWN figshare data collection (BAM, VCF, bigWig etc) - <https://doi.org/10.4225/55/5a76e03723567>
- CS IWGSC RefSeq v1.0 genome assembly - [https://urgi.versailles.inra.fr/download/iwgc/IWGSC\\_RefSeq\\_Assemblies/v1.0/](https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Assemblies/v1.0/)
- CS IWGSC RefSeq v1.0 annotations - [https://urgi.versailles.inra.fr/download/iwgc/IWGSC\\_RefSeq\\_Annotations/v1.0/](https://urgi.versailles.inra.fr/download/iwgc/IWGSC_RefSeq_Annotations/v1.0/)
- WGS data for the 16 accessions - <https://data.bioplatforms.com/organization/about/bpa-wheat-cultivars>

## Authors' contributions

NSW-H developed the concepts, led the code development, performed data analyses, interpreted examples and wrote the manuscript. RS contributed

code and performed data analyses. EK contributed code. MG interpreted examples and wrote the manuscript. UB developed the concepts, performed data analyses, interpreted examples and wrote the manuscript. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>School of Agriculture, Food and Wine, University of Adelaide, PMB 1, 5064 Glen Osmond, SA, Australia. <sup>2</sup>Bioinformatics Hub, School of Biological Sciences, University of Adelaide, SA 5005 Adelaide, Australia. <sup>3</sup>CSIRO Agriculture and Food, Glen Osmond, Locked Bag 2, SA 5064 Adelaide, Australia.

Received: 30 June 2018 Accepted: 6 November 2018

Published online: 17 December 2018

## References

1. Edwards D, Wilcox S, Barrero RA, Fleury D, Cavanagh CR, Forrest KL, Hayden MJ, Moolhuijzen P, Keeble-Gagnère G, Bellgard MI, Lorenc MT, Shang CA, Baumann U, Taylor JM, Morell MK, Langridge P, Appels R, Fitzgerald A. Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnol J*. 2012;10(6):703–8. <https://doi.org/10.1111/j.1467-7652.2012.00717.x>.
2. 10 Wheat Genomes Project. <http://www.wheatinitiative.org/activities/associated-programmes/10-wheat-genomes-project>. Accessed 21 Nov 2018.
3. Homepage: 10 Wheat Genomes. <http://www.10wheatgenomes.com/>. Accessed 21 Nov 2018.
4. King R, Bird N, Ramirez-Gonzalez R, Coghill JA, Patil A, Hassani-Pak K, Uauy C, Phillips AL. Mutation Scanning in Wheat by Exon Capture and Next-Generation Sequencing. *PLOS ONE*. 2015;10(9):1–18. <https://doi.org/10.1371/journal.pone.0137549>.
5. Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, Simmonds J, Ramirez-Gonzalez RH, Wang X, Borrill P, Fosker C, Ayling S, Phillips A, Uauy C, Dubcovsky J. Uncovering hidden variation in polyploid wheat. *Proc Natl Acad Sci*. 2017;114(6):E913–E921. <https://doi.org/10.1073/pnas.1619268114>. <http://www.pnas.org/content/114/6/E913.full.pdf>.
6. Warr A, Robert C, Hume D, Archibald A, Deeb N, Watson M. Exome sequencing: Current and future perspectives. *G3 (Bethesda)*. 2015;5(8):1543–50. <https://doi.org/10.1534/g3.115.018564>. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4528311/>.
7. Meienberg J, Bruggmann R, Oexle K, Matyas G. Clinical sequencing: is WGS the better WES? *Hum Genet*. 2016;135:359–62. <https://doi.org/10.1007/s00439-015-1631-9>. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4757617/>.
8. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, Shang L, Boisson B, Casanova J-L, Abel L. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci*. 2015;112(17):5473–8. <https://doi.org/10.1073/pnas.1418631112>. <http://www.pnas.org/content/112/17/5473.full.pdf>.
9. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, Thiruvahindrapuram B, Merico D, Jobling R, Nalpathamkalam T, Pellecchia G, Sung WWL, Wang Z, Bikangaga P, Boelman C, Carter MT, Cordeiro D, Cytynbaum C, Dell SD, Dhir P, Dowling JJ, Heon E, Hewson S, Hiraki L, Inbar-Feigenberg M, Klatt R, Kronick J, Laxer RM, Licht C, MacDonald H, Mercimek-Andrews S, Mendoza-Londono R, Piscione T, Schneider R, Schulze A, Silverman E, Siriwardena K, Snead OC, Sondheimer N, Sutherland J, Vincent A, Wasserman JD, Weksberg R, Shuman C, Carew C, Szego MJ, Hayeems RZ, Basran R, Stavropoulos DJ, Ray PN, Bowdin S, Meyn MS, Cohn RD, Scherer SW, Marshall CR.

- Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med*. 2017;20:435–443. Original Research Article. <https://doi.org/10.1038/gim.2017.119>.
10. Sorrells ME, Gustafson JP, Somers D, Chao S, Benschler D, Guedira-Brown G, Huttner E, Kilian A, McGuire PE, Ross K, Tanaka J, Wenzl P, Williams K, Qualset CO. Reconstruction of the Synthetic W7984 x Opata M85 wheat reference population. *Genome*. 2011;54(11):875–82. <https://doi.org/10.1139/g11-054>.
  11. Schlegel I R, Korzun V. About the origin of 1RS.1BL wheat-rye chromosome translocations from Germany. *Plant Breed*. 1997;116(6): 537–40. <https://doi.org/10.1111/j.1439-0523.1997.tb02186.x>.
  12. Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Choulet F, Distelfeld A, Poland J, Ronen Gil, Sharpe AG, Pozniak C, Barad O, Baruch K, Keeble-Gagnère G, Mascher M, Sharpe AG, Ben-Zvi G, Josselin A-A, Himmelbach A, Balfourier F, Gutierrez-Gonzalez J, Hayden M, Koh C, Muehlbauer G, Pasam RK, Paux E, Rigault P, Tibbits J, Tiwari V, Spannagl M, Lang D, Gundlach H, Haberer G, Mayer KFX, Ormanbekova D, Prade V, Šimková H, Wicker T, Swarbreck D, Rimbart H, Felder M, Guilhot N, Kaithakottil G, Keilwagen J, Leroy P, Lux T, Twardziok S, Venturini L, Juhász A, Abrouk M, Fischer I, Uauy C, Borrill P, Ramirez-Gonzalez RH, Arnaud D, Chalabi S, Chalhoub B, Cory A, Datla R, Davey MW, Jacobs J, Robinson SJ, Steuernagel B, van Ex F, Wulff BBH, Benhamed M, Bendahmane A, Concia L, Latrasse D, Alaux M, Bartoš J, Bellec A, Berges H, Doležel J, Frenkel Z, Gill B, Korol A, Letellier T, Olsen O-A, Singh K, Valárik M, van der Vossen E, Vautrin S, Weining S, Fahima T, Glikson V, Raats D, Čížková J, Toegelová H, Vrána J, Sourdille P, Darrier B, Barabaschi D, Cattivelli L, Hernandez P, Galvez S, Budak H, Jones JG, Witek K, Yu G, Small I, Melonek J, Zhou R, Belova T, Kanyuka K, King R, Nilsen K, Walkowiak S, Cuthbert R, Knox R, Wiebe K, Xiang D, Rohde A, Golds T, Čížková J, Akpinar BA, Biyiklioglu S, Gao L, N'Daiye A, Kubaláková M, Šafář J, Alfama F, Adam-Blondon A-F, Flores R, Guerche C, Loaec M, Quesneville H, Condie J, Ens J, Maclachlan R, Tan Y, Alberti A, Aury J-M, Barbe V, Couloux A, Cruaud C, Labadie K, Mangenot S, Wincker P, Kaur G, Luo M, Sehgal S, Chhuneja P, Gupta OP, Jindal S, Kaur P, Malik P, Sharma P, Yadav B, Singh NK, Khurana JP, Chaudhary C, Khurana P, Kumar V, Mahato A, Mathur S, Sevanthi A, Sharma N, Tomar RS, Holušová K, Plíhal O, Clark MD, Heavens D, Kettleborough G, Wright J, Balcáková B, Hu Y, Salina E, Ravin N, Skryabin K, Beletsky A, Kadnikov V, Mardanov A, Nesterov M, Rakitin A, Sergeeva E, Handa H, Kanamori H, Katagiri S, Kobayashi F, Nasuda S, Tanaka T, Wu J, Cattonaro F, Jiumeng M, Kugler K, Pfeifer M, Sandve S, Xun X, Zhan B, Batley J, Bayer PE, Edwards D, Hayashi S, Tulpová Z, Visendi P, Cui L, Du X, Feng K, Nie X, Tong W, Wang L. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*. 2018;361(6403). <https://doi.org/10.1126/science.aar7191>. <http://science.sciencemag.org/content/361/6403/eaar7191>.
  13. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. 2012;28(19):2520–2. <https://doi.org/10.1093/bioinformatics/bts480>.
  14. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis S, Stein L, Holmes IH. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol*. 2016;17(1):66. <https://doi.org/10.1186/s13059-016-0924-1>.
  15. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The Human Genome Browser at UCSC. *Genome Res*. 2002;12(6):996–1006. <https://doi.org/10.1101/gr.229102>. <http://arxiv.org/abs/http://genome.cshlp.org/content/12/6/996.full.pdf+html>. <http://genome.cshlp.org/content/12/6/996.abstract>.
  16. Zerbito DR, Achuthan P, Akanni W, Amodè MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, Gil L, Gordon L, Haggerty L, Haskell E, Hourlier T, Izuogo OG, Janacek SH, Juettemann T, To JK, Laird MR, Lavidas I, Liu Z, Loveland JE, Maurel T, McLaren W, Moore B, Mudge J, Murphy DN, Newham V, Nuhn M, Ogeh D, Ong CK, Parker A, Patricio M, Riat HS, Schuilenburg H, Sheppard D, Sparrow H, Taylor K, Thormann A, Vullo A, Walts B, Zadissa A, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Cunningham F, Yates A, Flicek P. Ensembl 2018. *Nucleic Acids Res*. 2018;46(D1):D754–D761. <https://doi.org/10.1093/nar/gkx1098>.
  17. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178–92. <https://doi.org/10.1093/bib/bbs017>.
  18. Milne I, Stephen G, Bayer M, Cock PJA, Pritchard L, Cardle L, Shaw PD, Marshall D. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 2013;14(2):193–202. <https://doi.org/10.1093/bib/bbs012>.
  19. Gremme G, Steinbiss S, Kurtz S. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Trans Comput Biol Bioinform*. 2013;10(3):645–56. <https://doi.org/10.1109/TCBB.2013.68>.
  20. Wheat Sequencing Framework Data Initiatives. <http://www.bioplatforms.com/wheat-sequencing/>. Accessed 21 Nov 2018.
  21. Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. The first near-complete assembly of the hexaploid bread wheat genome, triticum aestivum. *GigaScience*. 2017;6(11):1–7. <https://doi.org/10.1093/gigascience/gix097>.
  22. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
  23. Jordan KW, Wang S, Lun Y, Gardiner L-J, MacLachlan R, Hucl P, Wiebe K, Wong D, Forrest KL, Sharpe A, Sidebottom CH, Hall N, Toomajian C, Close T, Dubcovsky J, Akhunova A, Talbert L, Bansal UK, Bariana H, Hayden MJ, Pozniak C, Jeddeloh JA, Hall A, Akhunov E. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. *Genome Biol*. 2015;16(1):48. <https://doi.org/10.1186/s13059-015-0606-4>.
  24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and samtools. *Bioinformatics*. 2009;25(16):2078–9. <https://doi.org/10.1093/bioinformatics/btp352>.
  25. Wheat RNA-Seq. <https://urgi.versailles.inra.fr/files/RNASeqWheat/>. Accessed 21 Nov 2018.
  26. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. Star: ultrafast universal rna-seq aligner. *Bioinformatics*. 2013;29(1):15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
  27. Borrill P, Ramirez-Gonzalez R, Uauy C. expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant Physiol*. 2016;170(4): 2172–86. <https://doi.org/10.1104/pp.15.01667>. <http://www.plantphysiol.org/content/170/4/2172.full.pdf>.
  28. Nyquist WE. Differential fertilization in the inheritance of stem rust resistance in hybrids involving a common wheat strain derived from triticum timopheevi. *Genetics*. 1962;47(8):1109–24. <http://www.genetics.org/content/47/8/1109.full.pdf>.
  29. Bariana H, Hayden MJ, Ahmed NU, Bell JA, Sharp PJ, McIntosh RA. Mapping of durable adult plant and seedling resistances to stripe rust and stem rust diseases in wheat. *Aust J Agric Res*. 2001;52(12):1247–55. <https://doi.org/10.1071/AR01040>.
  30. CerealsDB MAS data. [http://www.cerealsdb.uk/net/cerealgenomics/CerealsDB/kasp\\_download.php](http://www.cerealsdb.uk/net/cerealgenomics/CerealsDB/kasp_download.php). Accessed 21 Nov 2018.
  31. Lehmensiek A, Eckermann P, Verbyla A, Appels R, Sutherland M, Daggard G. Curation of wheat maps to improve map accuracy and qtl detection. *Crop Pasture Sci*. 2005;56(12):1347–54.
  32. Flemmig E. Molecular Markers to Deploy and Characterize Stem Rust Resistance in Wheat: North Carolina State University; 2012. <http://www.lib.ncsu.edu/resolver/1840.16/8947>.
  33. Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK, Cavanagh CR. A multiparent advanced generation inter-cross population for genetic analysis in wheat. *Plant Biotechnol J*. 2012;10(7):826–39. <https://doi.org/10.1111/j.1467-7652.2012.00702.x>.
  34. Chemayek B, Bansal UK, Qureshi N, Zhang P, Wagoire WW, Bariana HS. Tight repulsion linkage between Sr36 and Sr39 was revealed by genetic, cytogenetic and molecular analyses. *Theor Appl Genet*. 2017;130(3): 587–95. <https://doi.org/10.1007/s00122-016-2837-5>.
  35. Williams K, Taylor S, Bogacki P, Pallotta M, Bariana H, Wallwork H. Mapping of the root lesion nematode (*Pratylenchus neglectus*) resistance gene Rln1 in wheat. *Theor Appl Genet*. 2002;104(5):874–9. <https://doi.org/10.1007/s00122-001-0839-3>.
  36. Jayatilake DV, Tucker EJ, Bariana H, Kuchel H, Edwards J, McKay AC, Chalmers K, Mather DE. Genetic mapping and marker development for resistance of wheat against the root lesion nematode *Pratylenchus neglectus*. *BMC Plant Biol*. 2013;13(1):230. <https://doi.org/10.1186/1471-2229-13-230>.

37. Jayatilake DV. Fine Mapping of Nematode Resistance Genes Rlnn1 and Cre8 in Wheat (*Triticum aestivum*). <http://hdl.handle.net/2440/97789>. Accessed 21 Nov 2018.
38. Neu C, Stein N, Keller B. Genetic mapping of the Lr20-Pm1 resistance locus reveals suppressed recombination on chromosome arm 7AL in hexaploid wheat. *Genome*. 2002;45(4):737–44.
39. Yan L, Helguera M, Kato K, Fukuyama S, Sherman J, Dubcovsky J. Allelic variation at the *vrn-1* promoter region in polyploid wheat. *Theor Appl Genet*. 2004;109(8):1677–86. <https://doi.org/10.1007/s00122-004-1796-4>.
40. Fu D, Szűcs P, Yan L, Helguera M, Skinner JS, von Zitzewitz J, Hayes PM, Dubcovsky J. Large deletions within the first intron in *vrn-1* are associated with spring growth habit in barley and wheat. *Mol Gen Genomics*. 2005;273(1):54–65. <https://doi.org/10.1007/s00438-004-1095-4>.
41. Guerrieri N, Cavaletto M. Cereals proteins. In: Yada RY, editor. *Proteins in Food Processing*. Woodhead Publishing Series in Food Science, Technology and Nutrition. 2nd. Cambridge: Woodhead Publishing; 2018. p. 223–244. <https://doi.org/10.1016/B978-0-08-100722-8.00009-7>. <https://www.sciencedirect.com/science/article/pii/B9780081007228000097>.
42. Yamamori M, Yamamoto K. Effects of two novel *wx-a1* alleles of common wheat (*triticum aestivum* L.) on amylose and starch properties. *J Cereal Sci*. 2011;54(2):229–35. <https://doi.org/10.1016/j.jcs.2011.06.005>. <http://www.sciencedirect.com/science/article/pii/S0733521011001032>.
43. Guzmán C, Ortega R, Yamamori M, Peña RJ, Alvarez JB. Molecular characterization of two novel null waxy alleles in mexican bread wheat landraces. *J Cereal Sci*. 2015;62(Supplement C):8–14. <https://doi.org/10.1016/j.jcs.2014.11.003>. <http://www.sciencedirect.com/science/article/pii/S0733521014002124>.
44. Boggini G, Cattaneo M, Paganoni C, Vaccino P. Genetic variation for waxy proteins and starch properties in italian wheat germplasm. *Euphytica*. 2001;119(1-2):111–4. <https://doi.org/10.1023/A:1017527430353>. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034931853&doi=10.1023%2FA%3a1017527430353&partnerID=40&md5=970c1d99ccc1028a2ee6b63188a0a4c9>, cited By 21.
45. Rodriguez-Quijano M, Nieto-Taladriz MT, Carrillo JM. Polymorphism of waxy proteins in iberian hexaploid wheats. *Plant Breed*. 1998;117(4):341–4. cited By 34. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0031715361&partnerID=40&md5=887b272922ee83d8b423ab9b01e12c98>.
46. Marcoz-Ragot C, Gateau I, Koenig J, Delaire V, Branlard G. Allelic variants of granule-bound starch synthase proteins in european bread wheat varieties. *Plant Breed*. 2000;119(4):305–9. <https://doi.org/10.1046/j.1439-0523.2000.00510.x>. cited By 13. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-0033858340&doi=10.1046%2Fj.1439-0523.2000.00510.x&partnerID=40&md5=36f23f756b46c3a94f0a7a0013ec7df6>.
47. Guzmán C, Alvarez JB. Wheat waxy proteins: polymorphism, molecular characterization and effects on starch properties. *Theor Appl Genet*. 2016;129(1):1–16. <https://doi.org/10.1007/s00122-015-2595-9>.
48. Nozoye T, Nagasaka S, Kobayashi T, Takahashi M, Sato Y, Sato Y, Uozumi N, Nakanishi H, Nishizawa NK. Phytosiderophore efflux transporters are crucial for iron acquisition in graminaceous plants. *J Biol Chem*. 2011;286(7):5446–54. <https://doi.org/10.1074/jbc.M110.180026>. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037657/>.
49. Nozoye T, Nagasaka S, Kobayashi T, Sato Y, Uozumi N, Nakanishi H, Nishizawa NK. The phytosiderophore efflux transporter tom2 is involved in metal transport in rice. *J Biol Chem*. 2015;290(46):27688–99. <https://doi.org/10.1074/jbc.M114.635193>. <http://www.jbc.org/content/290/46/27688.full.pdf+html>.
50. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305(5683):525–8. <https://doi.org/10.1126/science.1098918>. <http://science.sciencemag.org/content/305/5683/525.full.pdf>.
51. Sutton T, Baumann U, Hayes J, Collins NC, Shi B-J, Schnurbusch T, Hay A, Mayo G, Pallotta M, Tester M, Langridge P. Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science*. 2007;318(5855):1446–9. <https://doi.org/10.1126/science.1146853>. <http://science.sciencemag.org/content/318/5855/1446.full.pdf>.
52. Żmierko A, Samelak A, Kozłowski P, Figlerowicz M. Copy number polymorphism in plant genomes. *Theor Appl Genet*. 2014;127(1):1–18. <https://doi.org/10.1007/s00122-013-2177-7>.
53. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z. Computational tools for copy number variation (cnv) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*. 2013;14(1):S1. <https://doi.org/10.1186/1471-2105-14-S11-S1>.
54. Wu T, Lichten M. Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science*. 1994;263(5146):515–8. <https://doi.org/10.1126/science.8290959>. <http://arxiv.org/abs/http://science.sciencemag.org/content/263/5146/515.full.pdf>.
55. Xu X, Hsia AP, Zhang L, Nikolau BJ, Schnable PS. Meiotic recombination break points resolve at high rates at the 5' end of a maize coding sequence. *The Plant Cell*. 1995;7(12):2151–61. <https://doi.org/10.1105/tpc.7.12.2151>. <http://www.plantcell.org/content/7/12/2151.full.pdf>.
56. Darrier B, Rimbart H, Balfourier F, Pingault L, Josselin A-A, Servin B, Navarro J, Choulet F, Paux E, Sourdille P. High-resolution mapping of crossover events in the hexaploid wheat genome suggests a universal recombination mechanism. *Genetics*. 2017;206(3):1373–88. <https://doi.org/10.1534/genetics.116.196014>. <http://www.genetics.org/content/206/3/1373.full.pdf>.
57. Delaney DE, Nasuda S, Endo TR, Gill B, Hulbert SH. Cytologically based physical maps of the group-2 chromosomes of wheat. *Theor Appl Genet*. 1995;91(5):568–73. <https://doi.org/10.1007/BF00223281>.
58. Delaney DE, Nasuda S, Endo TR, Gill B, Hulbert SH. Cytologically based physical maps of the group 3 chromosomes of wheat. *Theor Appl Genet*. 1995;91(5):780–2. <https://doi.org/10.1007/BF00220959>.
59. Dvorák J, Chen K-C. Distribution of nonstructural variation between wheat cultivars along chromosome arm 6bp: Evidence from the linkage map and physical map of the arm. *Genetics*. 1984;106(2):325–33. <https://pubmed.ncbi.nlm.nih.gov/17246194/>. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1202259/>.
60. Gill KS, Gill B, Endo TR, Taylor T. Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics*. 1996;144(4):1883–91. <http://www.genetics.org/content/144/4/1883.full.pdf>.
61. Gill KS, Gill B, Endo TR, Boyko EV. Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics*. 1996;143(2):1001–12. <http://www.genetics.org/content/143/2/1001.full.pdf>.
62. Hohmann U, Endo TR, Gill KS, Gill B. Comparison of genetic and physical maps of group 7 chromosomes from *triticum aestivum* L. *Mol Gen Genet*. 1994;245(5):644–53. <https://doi.org/10.1007/BF00282228>.
63. Mickelson-Young L, Endo TR, Gill B. A cytogenetic ladder-map of the wheat homoeologous group-4 chromosomes. *Theor Appl Genet*. 1995;90(7):1007–11. <https://doi.org/10.1007/BF00222914>.
64. Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoel K, Aury J-M, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C. Structural and functional partitioning of bread wheat chromosome 3b. *Science*. 2014;345(6194). <https://doi.org/10.1126/science.1249721>. <http://science.sciencemag.org/content/345/6194/1249721.full.pdf>.
65. Horvath A, Didier A, Koenig J, Exbrayat F, Charmet G, Balfourier F. Analysis of diversity and linkage disequilibrium along chromosome 3b of bread wheat (*triticum aestivum* L.) *Theor Appl Genet*. 2009;119(8):1523. <https://doi.org/10.1007/s00122-009-1153-8>.
66. Hao C, Wang L, Ge H, Dong Y, Zhang X. Genetic diversity and linkage disequilibrium in chinese bread wheat (*triticum aestivum* L.) revealed by *ssr* markers. *PLOS ONE*. 2011;6(2):1–13. <https://doi.org/10.1371/journal.pone.0017279>.
67. Rimbart H, Darrier B, Navarro J, Kitt J, Choulet F, Leveugle M, Duarte J, Rivière N, Eversole K, on behalf of The International Wheat Genome Sequencing Consortium, Le Gouis J, on behalf of The International Wheat Genome Sequencing Consortium, Davassi A, Balfourier F, Le Paslier M-C, Berard A, Brunel D, Feuillet C, Poncet C, Sourdille P, Paux E. High throughput snp discovery and genotyping in hexaploid wheat. *PLOS ONE*. 2018;13(1):1–19. <https://doi.org/10.1371/journal.pone.0186329>.
68. Cubizolles N, Rey E, Choulet F, Rimbart H, Laugier C, Balfourier F, Bordes J, Poncet C, Jack P, James C, Gielen J, Argillier O, Jaubertie J-P, Auzanneau J, Rohde A, Ouwerkerk PBF, Korzun V, Kollers S, Guerreiro L, Hourcade D, Robert O, Devaux P, Mastrangelo A-M, Feuillet C, Sourdille P, Paux E. Exploiting the repetitive fraction of the wheat genome for high-throughput single-nucleotide polymorphism discovery and

- genotyping. Madison: Crop Science Society of America. 2016;9. <https://doi.org/10.3835/plantgenome2015.09.0078>. 1.
69. Able JA, Langridge P, Milligan AS. Capturing diversity in the cereals: many options but little promiscuity. *Trends Plant Sci.* 2007;12(2):71–79. <https://doi.org/10.1016/j.tplants.2006.12.002>.
  70. Semagn K, Babu R, Hearne S, Olsen M. Single nucleotide polymorphism genotyping using kompetitive allele specific pcr (kasp): overview of the technology and its application in crop improvement. *Mol Breeding.* 2014;33(1):1–14. <https://doi.org/10.1007/s11032-013-9917-x>.
  71. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel J, Batley J, Edwards D. The pangenome of hexaploid bread wheat. *The Plant Journal.* 2017;90(5):1007–13. <https://doi.org/10.1111/tpj.13515>.
  72. Wheat pangenome BLAST. <http://www.appliedbioinformatics.com.au/gbrowseblast/cgi-bin/>. Accessed 21 Nov 2018.
  73. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S. Sequence-specific error profile of illumina sequencers. *Nucleic Acids Res.* 2011;39(13):e90. <https://doi.org/10.1093/nar/gkr344>.
  74. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu F-H, McKenzie N, Raats D, Ramirez-Gonzalez R, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Di Palma F, Phillips A, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. 2017. <https://doi.org/10.1101/gr.217117.116>. <http://dx.doi.org/10.1101/gr.217117.116>.
  75. Mayer KFX, Rogers J, Doležel J, Pozniak C, Eversole K, Feuillet C, Gill B, Friebe B, Lukaszewski AJ, Sourdil P, Endo TR, Kubaláková M, Čiháliková J, Dvorská Z, Vrána J, Šperková R, Šimková H, Febre M, Clissold L, McLay K, Singh K, Chhuneja P, Singh NK, Khurana J, Akhunov E, Choulet F, Alberti A, Barbe V, Wincker P, Kanamori H, Kobayashi F, Itoh T, Matsumoto T, Sakai H, Tanaka T, Wu J, Ogihara Y, Handa H, Maclachlan PR, Sharpe A, Klassen D, Edwards D, Batley J, Olsen O-A, Sandve SR, Lien S, Steuernagel B, Wulff B, Caccamo M, Ayling S, Ramirez-Gonzalez R, Clavijo BJ, Wright J, Pfeifer M, Spannagl M, Martis MM, Mascher M, Chapman J, Poland JA, Scholz U, Barry K, Waugh R, Rokhsar DS, Muehlbauer GJ, Stein N, Gundlach H, Zytynicki M, Jamilloux V, Quesneville H, Wicker T, Faccioli P, Colaiaacovo M, Stanca AM, Budak H, Cattivelli L, Glover N, Pingault L, Paux E, Sharma S, Appels R, Bellgard M, Chapman B, Nussbaumer T, Bader KC, Rimbart H, Wang S, Knox R, Kilian A, Alaux M, Alfama F, Couderc L, Guilhot N, Viseux C, Loaec M, Keller B, Praud S. A chromosome-based draft sequence of the hexaploid bread wheat (*triticum aestivum*) genome. *Science.* 2014;345(6194). <https://doi.org/10.1126/science.1251788>. <http://arxiv.org/abs/http://science.sciencemag.org/content/345/6194/1251788.full.pdf>.
  76. Blake VC, Birkett C, Matthews DE, Hane DL, Bradbury P, Jannink J-L. The triticeae toolbox: Combining phenotype and genotype data to advance small-grains breeding. 2016;9. <https://doi.org/10.3835/plantgenome2014.12.0099>. 2.
  77. Ware D, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR. Gramene, a tool for grass genomics. *Plant Physiol.* 2002;130(4):1606–13. <https://doi.org/10.1104/pp.015248>. <http://www.plantphysiol.org/content/130/4/1606.full.pdf>.
  78. Carollo V, Matthews DE, Lazo GR, Blake TK, Hummel DD, Lui N, Hane DL, Anderson OD. Graingenes 2.0. an improved resource for the small-grains community. *Plant Physiol.* 2005;139(2):643–51. <https://doi.org/10.1104/pp.105.064485>. <http://www.plantphysiol.org/content/139/2/643.full.pdf>.
  79. Kersey PJ, Allen JE, Allot A, Barba M, Boddus S, Bolt BJ, Carvalho-Silva D, Christensen M, Davis P, Grabmueller C, Kumar N, Liu Z, Maurel T, Moore B, McDowall MD, Maheswari U, Naamati G, Newman V, Ong CK, Paulini M, Pedro H, Perry E, Russell M, Sparrow H, Tapanari E, Taylor K, Vullio A, Williams G, Zadissia A, Olson A, Stein J, Wei S, Tello-Ruiz M, Ware D, Luciani A, Potter S, Finn RD, Urban M, Hammond-Kosack KE, Bolser DM, De Silva N, Howe KL, Langridge N, Maslen G, Staines DM, Yates A. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* 2018;46(D1):802–8. <https://doi.org/10.1093/nar/gkx1011>.
  80. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hoofstede R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The fair guiding principles for scientific data management and stewardship. *Scientific Data.* 2016;3:160018. Comment.
  81. Griffin P, Khadake J, LeMay K, Lewis S, Orchard S, Pask A, Pope B, Roessner U, Russell K, Seemann T, Treloar A, Tyagi S, Christiansen J, Dayalan S, Gladman S, Hangartner S, Hayden H, Ho W, Keeble-Gagnère G, Korhonen P, Neish P, Prestes P, Richardson M, Watson-Haigh N, Wyres K, Young N, Schneider M. Best practice data life cycle approaches for the life sciences [version 1; referees: 2 approved with reservations]. *F1000Research.* 2017;6(1618):. <https://doi.org/10.12688/f1000research.12344.1>.
  82. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial “pan-genome”. *Proc Natl Acad Sci.* 2005;102(39):13950–5. <https://doi.org/10.1073/pnas.0506758102>. <http://www.pnas.org/content/102/39/13950.full.pdf>.
  83. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol.* 2015;23:148–54. <https://doi.org/10.1016/j.mib.2014.11.016>. Host–microbe interactions: bacteria • Genomics.
  84. Sun C, Hu Z, Zheng T, Lu K, Zhao Y, Wang W, Shi J, Wang C, Lu J, Zhang D, Li Z, Wei C. Rpan: rice pan-genome browser for ≈3000 rice genomes. *Nucleic Acids Res.* 2017;45(2):597–605. <https://doi.org/10.1093/nar/gkw958>.
  85. Consortium TCP-G. Computational pan-genomics: status, promises and challenges. *Brief Bioinform.* 2018;19(1):118–35. <https://doi.org/10.1093/bib/bbw089>.
  86. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. *Genome Res.* 2017;27(5):665–76. <https://doi.org/10.1101/gr.214155.116>. 2836023[pmid].

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

