An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations.

Bernardo J. Clavijo[1*], Luca Venturini[1*], Christian Schudoma[1], Gonzalo Garcia Accinelli[1], Gemy Kaithakottil[1], Jonathan Wright[1], Philippa Borrill[2], George Kettleborough[1], Darren Heavens[1], Helen Chapman[1], James Lipscombe[1], Tom Barker[1], Fu-Hao Lu[2], Neil McKenzie[2], Dina Raats[1], Ricardo H. Ramirez-Gonzalez[1], Aurore Coince[1], Ned Peel[1], Lawrence Percival-Alwyn[1], Owen Duncan[3], Josua Trösch[3], Guotai Yu[2], Dan Bolser[4], Guy Namaati[4], Arnaud Kerhornou[4], Manuel Spannagl[5], Heidrun Gundlach[5], Georg Haberer[5], Robert P. Davey[1,6], Christine Fosker[1], Federica Di Palma[1,6], Andrew Phillips[7], A. Harvey Millar[3], Paul J. Kersey[4], Cristobal Uauy[2], Ksenia V. Krasileva[1,6,8], David Swarbreck[1,6+], Michael W. Bevan[2+] and Matthew D. Clark[1,6+].

*contributed equally to this work

+corresponding authors

[1]Earlham Institute, Norwich, UK; [2]John Innes Centre, Norwich, UK; [3]ARC Centre of Excellence in Plant Energy Biology, The University of Western Australia, Crawley WA 6009, Australia; [4]EMBL European Bioinformatics Institute, Hinxton, UK; [5]PGSB Helmholtz Center, Munich, Germany; [6]University of East Anglia, Norwich, UK; [7]Rothamsted Research, Harpenden, UK; [8]The Sainsbury Laboratory, Norwich UK;

Advances in genome sequencing and assembly technologies are generating many high quality genome sequences, but assemblies of large, repeat-rich polyploid genomes, such as that of bread wheat, remain fragmented and incomplete. We have generated a new wheat whole-genome shotgun sequence assembly using a combination of optimised data types and an assembly algorithm designed to deal with large and complex genomes. The new assembly represents more than 78% of the genome with a scaffold N50 of 88.8kbp that has a high fidelity to the input data. Our new annotation combines strand-specific Illumina RNAseq and PacBio full-length cDNAs to identify 104,091 high confidence protein-coding genes and 10,156 non-coding RNA genes. We confirmed three known and identified one novel genome rearrangements. Our approach enables the rapid and scalable assembly of wheat genomes, the identification of structural variants, and the definition of complete gene models, all powerful resources for trait analysis and breeding of this key global crop. [Supplemental material is available for this article.]

Running title: "An improved wheat genome assembly and annotation"

Keywords: wheat, whole genome shotgun assembly, annotation, transcriptome, agronomic genes

Improvements in sequencing read lengths and throughput have enabled the rapid and cost-effective assembly of many large and complex genomes (Gnerre et al., 2011; Lam et al., 2011). Comparisons between assembled genomes have revealed many classes of sequence variation of major functional significance that were not detected by direct alignment of sequence reads to a common reference (Gan et al., 2011; 1000 Genomes Project Consortium et al., 2010; Bishara et al., 2015). Therefore, accurate comparative genomics requires that genome sequences are assembled prior to alignment, but in many eukaryotic genomes assembly is complicated by the presence of large tracts of repetitive sequences (Treangen and Salzberg, 2011; Chaisson et al., 2015) and the common occurrence of genome duplications, for example in polyploids (Blanc and Wolfe, 2004; Berthelot et al., 2014).

Recent innovations in sequence library preparation, assembly algorithms, and long-range scaffolding have dramatically improved whole genome shotgun assemblies from short read sequences. These include PCR-free library preparation to reduce bias (Aird et al., 2011), longer sequence reads, and algorithms that preserve allelic diversity during assembly (Weisenfeld et al., 2014). Short-read assemblies have been linked into larger chromosome-scale scaffolds by Hi-C *in vivo* (Lieberman-Aiden et al., 2009) and *in vitro* (Putnam et al., 2016) chromatin proximity ligation, and by linked-read sequencing technologies (Mostovoy et al., 2016; Weisenfeld et al., 2016). Although it is more expensive than short read sequencing approaches, Single Molecule Real Time (SMRT) sequencing improved the contiguity and repeat representation of mammalian (Pendleton et al., 2015; Gordon et al., 2016) and diploid grass genomes (Zimin et al., 2016). SMRT technologies are also being used to generate the complete sequence of transcripts, increasing the accuracy of splicing isoform definition (Abdel-Ghany et al., 2016).

The assembly of the 17Gbp allohexaploid genome of bread wheat (*Triticum aestivum*) has posed major difficulties, as it is composed of three large, repetitive and closely related genomes (Moore et al., 1995). Despite progressive improvements, an accurate and near-complete wheat genome sequence assembly and corresponding high-quality gene annotation has not yet been generated. Initial whole genome sequencing used orthologous Poaceae protein sequences to generate highly fragmented gene assemblies (Brenchley et al., 2012). A BAC-based assembly of chromosome 3B provided major insights into wheat chromosome organisation (Choulet et al., 2014). Illumina sequencing and assembly of flow-sorted chromosome arm DNA (Chromosome Survey Sequencing, CSS) identified homoeologous relationships between genes in the three genomes, but the assemblies remained highly fragmented (IWGSC, 2014). Recently a whole genome shotgun sequence of hexaploid wheat was assembled and anchored, though not anno-

Table 1: Comparison of TGACv1 scaffolds to the IWGSC and Chapman assemblies of hexaploid wheat. Numbers are calculated using sequences greater than 500bp and including gaps (Ns) for each assembly.

	Size (Gb)	Seq. count	N20 (kb)	N50 (kb)	N80 (kb)	%Ns	% of genome
TGACv1	13.43	735,943	180.1	88.8	32.8	5.7	78.8
W7984	8.21	955,122	47.1	24.8	9.9	15.2	48.2
CSS	8.32	4,061,833	8.6	3.3	1.2	1.0	48.9

Table 2: Comparison of TGACv1 chromosome 3B scaffolds to BAC-based scaffolds (Choulet et al., 2014), and 3B scaffolds from the W7984 and CSS assemblies. Numbers are calculated using sequences greater than 500bp and including gaps (Ns) for each assembly.

	Scaffold count	N50 (kb)	Total seq. (Mb)	Gene count	% genes
3B ref.	2,808	892.4	832.8	7,703	100.0
TGACv1	29,090	116.5	790.0	6,983	90.7
W7984	26,206	30.6	479.4	5,671	73.6
CSS	272,072	3.4	557.2	5,233	67.9

tated, using an ultra-dense genetic map (Chapman et al., 2015). The assembly contained ∼48.2% of the genome with contig and scaffold N50 lengths of 8.3kbp and 25kbp, respectively.

Here we report a new sequence assembly and annotation of the allohexaploid wheat landrace Chinese Spring (CS42). Our approach is open source, rapid and scalable, and has enabled a more in-depth analysis of sequence and structural variation in this key global crop. The scaffolds cover 13.4Gbp of the genome with an N50 of 88.8kbp, and are classified into chromosome arms. The annotation, supported by extensive transcriptome sequence and 1.5 million full length SMRT cDNA sequences, contains 104,091 high confidence protein-coding genes.

Results

DNA library preparation and sequencing

We reduced bias and retained maximum sequence complexity by using unamplified libraries for contig generation (Kozarewa et al., 2009) and precisely sized mate-pair libraries for scaffolding (Heavens et al., 2015). Libraries were sequenced using Illumina paired-end (PE) 250bp reads to distinguish closely related sequences. In total, 1.1 billion PE reads were generated to provide 33× sequence coverage of the CS42 genome (Supplemental Table S4.1). For scaffolding, long mate-pair (LMP) libraries with insert sizes ranging from 2480 to 11,600bp provided 53× sequence coverage, and Tight, Amplification-free, Large insert PE Libraries (TALL) with an insert size of 690bp provided 15× sequence coverage (Supplemental Table S4.2).

Genome assembly

Nearly 3 million contigs (of length greater than 500bp) were generated using the w2rap-contigger (Clavijo, 2016) with an N50 of 16.7kbp (Supplemental Table S4.3). After scaffolding using SOAPdenovo (Luo et al., 2012), the assembly contained 1.3 million sequences with an N50 of 83.9kbp. To generate the TGACv1 assembly, scaffolds were classified to chromosome arms using raw CSS reads (IWGSC, 2014) and subsequently screened with a two-tiered filter based first on their length and their k-mer content (see Supplemental Information S4.5). The approach removed short, redundant sequences from the assembly minimising the loss of unique sequence content, leading to an increase in scaffold N50 to 88.8kbp.

The genome of a synthetic wheat line W7984 was previously assembled with an improved version of meraculous (Chapman et al., 2011) using 150bp PE libraries with varying insert sizes, for a combined genome coverage of 34.3×, together with 1.5kbp and 4kbp LMP libraries for scaffolding (Chapman et al., 2015). This contig assembly, with an N50 of 8.3kbp, covered 8Gbp of the genome while the scaffold assembly covered 8.21Gbp with an N50 of 24.8kbp. In comparison, the TGACv1 assembly represents almost 80% of the 17Gbp genome, a 60% improvement in genome coverage. The contiguity of the TGACv1 assembly is nearly four times that of the W7984 assembly and thirty times that of the CSS assembly (Table 1; IWGSC (2014)).

The KAT spectra-cn plot generated from TGACv1 (Figure 1**A**) showed that k-mers found at low frequency (<12), representing sequencing errors, were not found in the assembly (shown by the black distribution at k-mer multiplicity <12). Most sequence content was represented in the assembly once (shown by the main red distribution), with k-mers originating from the repetitive and the homoeologous regions of the genome represented at higher frequencies (>50). This indicated a low level of chimeric assemblies and established the accuracy of the assembly and scaffolding methods. K-mer spectra analysis of the CSS assembly (Figure 1**B**), revealed a larger fraction of absent content, corresponding to the black distribution between k-mer multiplicity 15 and 45, and greater amounts of duplication in the single copy regions of the assembly, corresponding to the purple and green areas of the main red distribution. A large amount of content in the CSS assembly did not appear as sequenced content of the reads in our PCR-free paired end data, as shown in the red bar at k-mer multiplicity equal to 0, indicating a high level of chimeric sequences or consensus problems.

90.7% of the genes previously identified on the 3B BAC-based assembly (Choulet et al., 2014) aligned to TGACv1 scaffolds (Table 2), compared with 68–74% aligned to W7984 and 67.9% aligned to CSS chromosome 3B scaffolds. This demonstrated both the improved representation of the TGACv1 assembly and the precision of chromosome classifications.

Alignment of TGACv1 3B scaffolds to the 3B BAC-based pseudomolecule (Figure 2**A**,**C**) showed that they were largely in agreement. Two examples of apparent disagreement are shown in Figure 2**B**,**D**. Scaffold_221671_3B spanned a gap of 700kbp in the 3B BAC assembly, and re-oriented and removed a duplication, by identifying both ends of a CACTA element (Figure 2**B**). Scaffold_220592_3B spanned 582kbp and diverged in one location (Figure 2**D**), and contained a Sabrina solo-LTR with a characteristic ATCAG target site duplication (TSD). In scaffold_220592_3B the TSD was present on either side of the Sabrina_3231 element, while in the BAC-based scaffold Sabrina homology ended in Ns. In the BAC-based assembly only one side of the disjunction showed alignment similarity to CACTA_3026, which was found to be complete in scaffold_220592_3B and spanned the disjunction (Figure 2**D**). These two examples illustrate how the TGACv1 assembly generated accurate scaffolds spanning typical complex and long tracts of repetitive DNA characterising the wheat genome, which were misassembled in the BAC-based approach.

Repetitive DNA composition

Supplemental Table S7.1 shows the class I and class II mobile element composition of wheat based on the TGACv1 assembly. More than 80% of the 13.4Gbp assembly was composed of approximately 9.7 million annotated transposable element entities, of which nearly 70% were retroelements (class I) and 13% DNA transposons (class II). Among the class I elements, Gypsy and Copia LTR retroelements comprised the major component of the repeats, while CACTA DNA elements were highly predominant among class II DNA repeat types. No major differences in the repeat composition of the three genomes were apparent. Compared with *Brachypodium distachyon*, which has a related but much smaller genome (Vogel et al., 2010), there has been a greater than 100× increase in repeat content, driven by both class I and class II expansion. The preponderance of CACTA DNA elements in the wheat genome emerged during this massive expansion.

Gene prediction and annotation

A total of 217,907 loci and 273,739 transcripts were identified from a combination of cross-species protein alignments, 1.5 million high-quality long PacBio cDNA reads, and over 3.2 billion RNA-Seq read pairs covering a range of tissues and developmental stages (Table 3, Supplemental Information S8).

Loci were identified as coding, long ncRNA, or repeat associated, and were classified as high or low confidence based on similarity to known plant protein sequences and supporting evidence from wheat transcripts (Supplemental Information S8.5.5). We assigned 104,091 coding genes (154,798 transcripts) as high confidence (HC), of which 95,827 spanned at least 80% of the length of the best identified homolog (termed protein rank 1, P1, in the annotation; Supplemental Figure S8.1 and Supplemental Information S8.5.1). The HC protein-coding set contained 51,851 genes confirmed by a PacBio transcript (Transcript rank 1, T1) and an additional 29,996 genes fully supported by assembled RNA-Seq data (T2), providing full transcriptome support for 81,847 (78.63%) HC genes. Gene predictions were assessed by identifying 2707 single copy genes common to *B. distachyon*, *Oryza sativa*, *Sorghum bicolor*, *Setaria italica* and *Zea mays*. A single orthologous wheat gene was identified for 2686 (99.22%) of these, with 2665 (98.45%) classified as HC and 21 (0.78%) in the low confidence (LC) set. A high coherence in gene length (r=0.969) was found between wheat and *B. distachyon* proteins (Supplemental Figure S8.2). These findings show that the HC gene set is robust and establishes a lower bound estimate for the total number of protein-coding genes in wheat. An additional 103,660 loci were defined as LC (i.e. gene models with all their transcripts either having less than 60% protein coverage or lacking wheat transcript support). These include bona fide genes that were fragmented due to breaks in the current assembly, wheat specific genes, and genes without transcriptome support (Supplementary Table S8.8).

We also identified 10,156 HC non-coding genes with little similarity in protein databases and low protein coding potential. The majority of these genes are located in intergenic regions (8854,

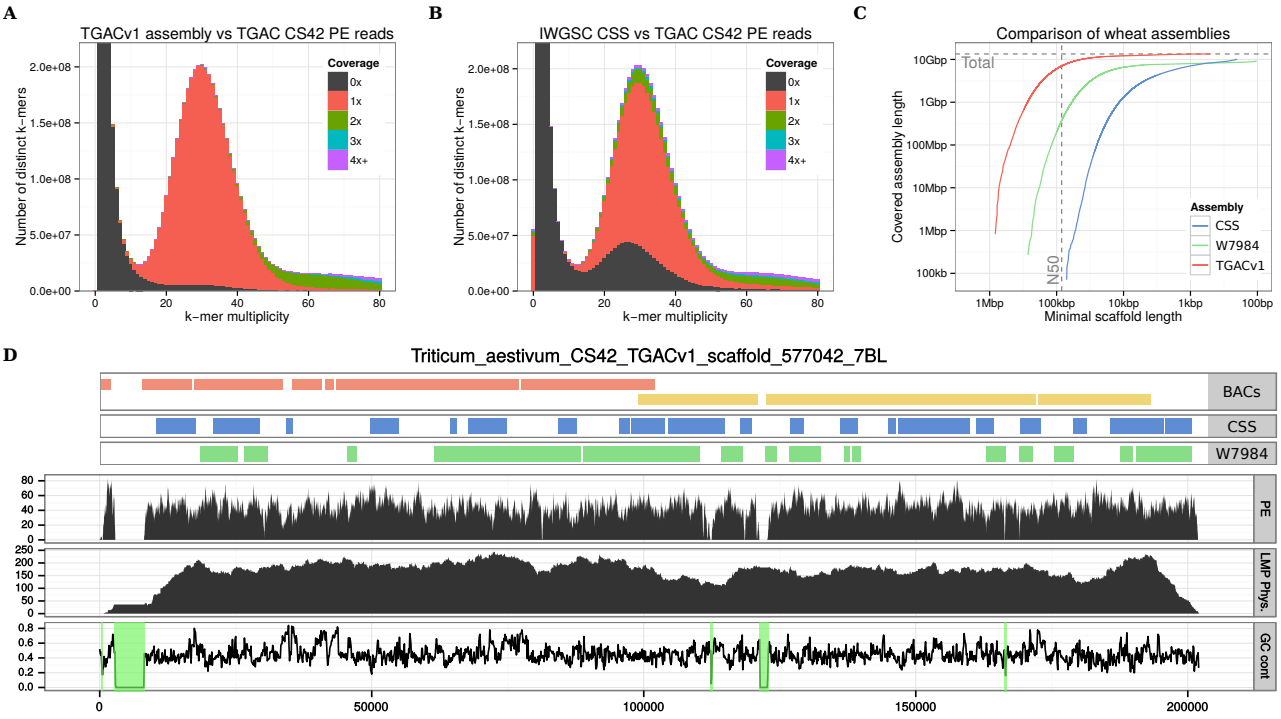

Figure 1: Summary of the TGACv1 wheat genome sequence assembly. A,**B**: KAT spectra-cn plots comparing the PE reads to the TGACv1 scaffolds (**A**) and CSS scaffolds (**B**). Plots are coloured to show how many times fixed length words (k-mers) from the reads appear in the assembly; frequency of occurrence (multiplicity, x-axis) and number of distinct k-mers (y-axis). Black represents k-mers missing from the assembly, red represents k-mers that appear once in the assembly, green twice, etc. Plots were generated using k = 31. The black distribution between k-mer multiplicity 15 and 45 in (B) represents k-mers that do not appear in the CSS assembly.
C: Comparison of scaffold lengths and total assembly sizes of the TGACv1, W7984, and CSS assemblies. **D**: Scaffold 577042 of the TGACv1 assembly. Tracks from top to bottom: aligned BAC contigs, CSS contigs, W7984 contigs, coverage of PE reads, coverage of LMP fragments, and GC content with scaffolded gaps (N stretches) with 0% GC highlighted in green. There are two BACs (comprised of 7 and 4 contigs each), 22 CSS contigs, and 15 W7984 contigs across the single TGACv1 scaffold.

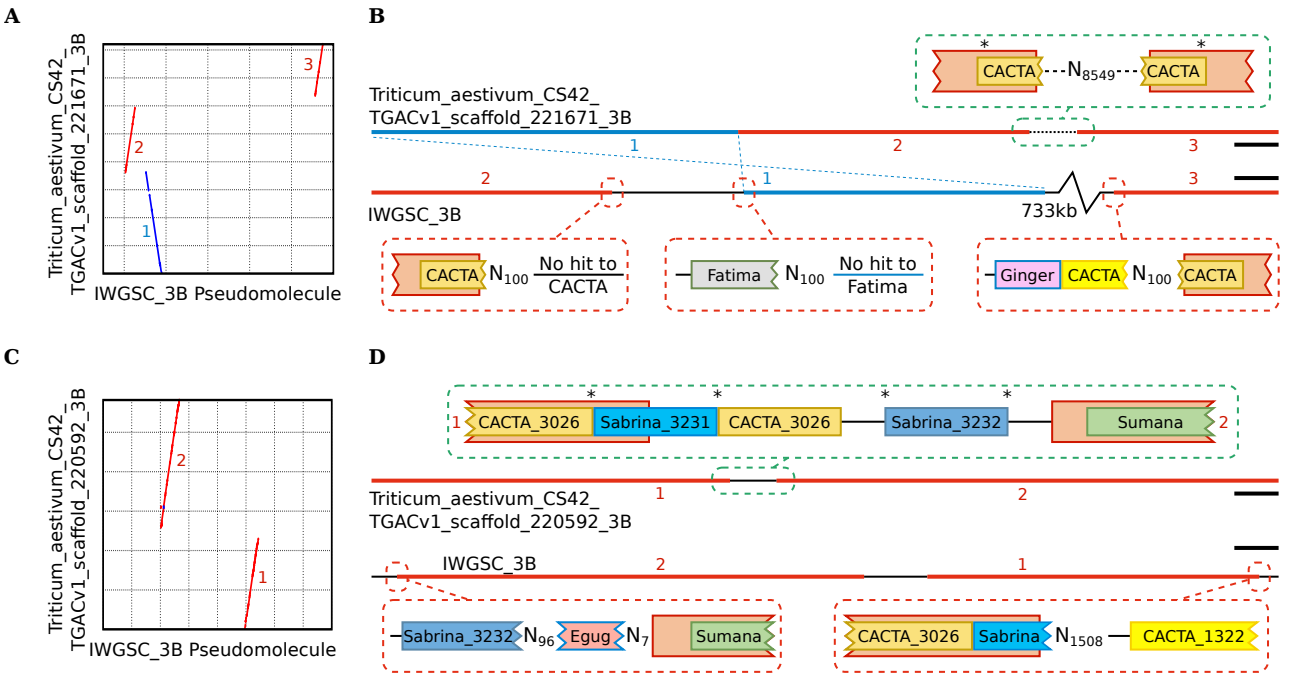

Figure 2: Comparative alignment of TGACv1 scaffolds with the 3B BAC-based pseudomolecule. A,**C**: Dot plots between TGACv1 scaffolds and 3B show disruptions in sequence alignment including rearrangements (red) and inversions (blue). **B**,**D**: Graphical representation of sequence annotations in disrupted regions. Junctions in the TGACv1 scaffolds are consistent with a complete retroelement spanning the junction that includes identical TSD on either side of the retroelement (asterisks). Corresponding regions in the 3B BAC-based pseudomolecule are characterised by Ns that produce inconsistent alignment of retroelements across putative junctions. Retroelements of the same family (CACTA, Sabrina) but matching distinct members in the TREP database are indicated by different colours. Numbers adjacent to sequences correspond to regions shown in panel **A** and **C**, respectively. Scale bars correspond to 10kbp (**B**) and 30kbp (**D**).

or 87.18%), while most of the remaining 1302 are antisense to coding genes (1082, or 10.65%; see Supplementary Information S8.5.8). 5413 of wheat non-coding genes (53.30%) were detected in at least one, of the two sequenced wheat diploid progenitor species *T. urartu* and *Ae. tauschii* (at least 90% coverage and 90% identity; see Supplementary Information S8.5.8).

To obtain additional support for gene predictions, a proteome map was constructed from 27 wheat tissues (Supplemental Information S9). This identified 2,106,323 significant peptide spectrum matches corresponding to 102,379 distinct peptides. Of these, 96.20% matched HC genes, while 13.29% were assigned to LC genes. For 56,391 genes (43,431 HC, 12,960 LC) we were able to identify at least one peptide confirming the predicted coding sequence. Due to the hexaploid nature of wheat, only 22.1% of the peptides could be assigned to a single gene. Applying progressively stricter filters, by requiring at least 2 or 5 peptides, confirmed the protein sequence of 30,607 and 17,316 HC genes, respectively. 10,819 genes met the criteria of having support from multiple peptides with at least one uniquely identifying peptide, and were considered as unambiguously corroborated by proteomic data. Among the LC genes, only 368 were identified by two or more peptides that did not match any HC gene, further supporting confidence assignments. Among these, 343 were classified as LC due to having less than 60% the length of the identified homolog, while the remaining 25 genes were classified as LC due to either repeat association or lack of wheat transcript support.

We compared the TGACv1, CSS (IWGSC, 2014) and chromosome 3B (Choulet et al., 2014) gene models. Of the 100,344 HC genes in the IWGSC annotation (PGSB/MIPS version 2.2 and INRA version 1.0 from Ensembl release 29) we were able to transfer 97,072 (97%) to the TGACv1 assembly with stringent alignment parameters (at least 90% coverage and 95% identity). Fewer (72%) of IWGSC (IWGSC, 2014) low confidence, unsupported, repeat associated and non-coding loci could be aligned (at least 90% coverage and 95% identity) likely reflecting differences between the assemblies of repeat rich and difficult to assemble regions. Of the TGACv1 HC genes, 61% overlapped with an aligned IWGSC HC gene and 78% to the full IWGSC gene set (Supplemental Information S8.5.7). Less agreement was found between TGACv1 LC and ncRNA genes and the IWGSC annotation, with only 8% overlapping IWGSC HC loci and 40% overlapping the full IWGSC gene set (Figure 3**A**). Of the 22,904 (22%) high confidence TGACv1 genes not overlapping a transferred IWGSC gene, 19,810 (86%) had cross species protein similarity support with 6665 (29%) fully supported by a PacBio transcript (Figure 3**A**). We identified 13,609 TGACv1 genes that were overlapped by transcripts originating from two or more IWGSC genes in our annotation, indicating that they were likely fragmented in the CSS assembly. In 8175 of these cases (60%) we were able to find a PacBio read fully supporting our gene model. These differences reflect improvements in contiguity, a more comprehensive representation of the wheat gene space in our assembly, and improved transcriptome support for annotation.

Alternative splicing

Alternative splicing is an important mRNA processing step that increases transcriptome plasticity and proteome diversity (Staiger and Brown, 2013). The TGACv1 annotation includes high-quality

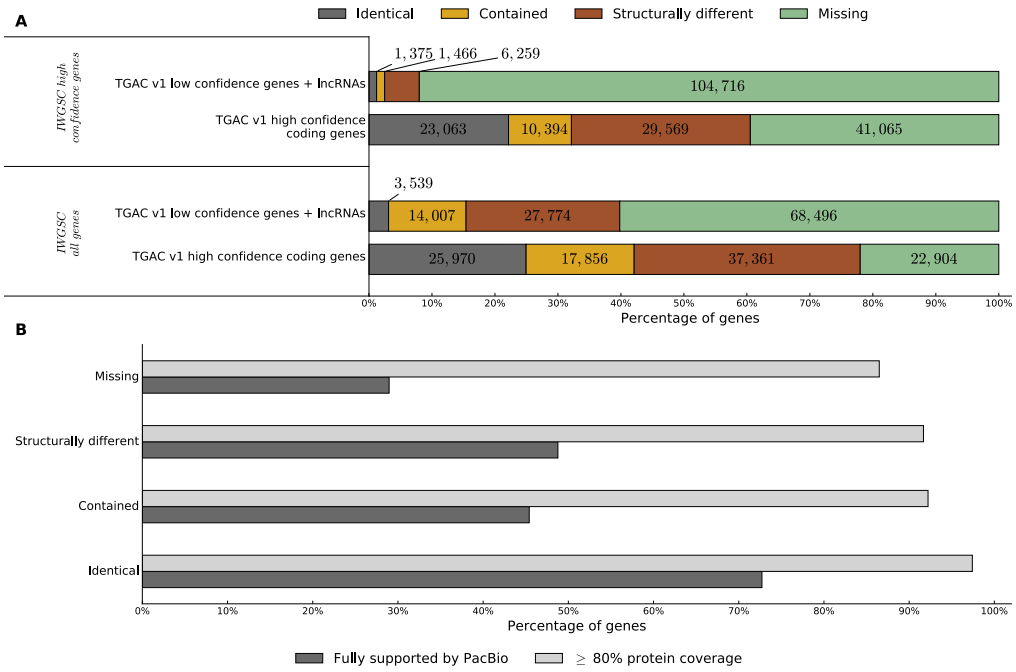

Figure 3: Comparison between IWGSC annotation and TGACv1 high and low confidence genes.IWGSC genes were aligned to the TGACv1 assembly (gmap, $\geq$ 90% coverage, $\geq$ 95% identity) and classified based on overlap with TGACv1 genes. **A**: Identical - shared exon-intron structure; Contained - exactly contained within the TGACv1 gene; Structurally different - alternative exon-intron structure; Missing - no overlap with IWGSC. **B**: Bar plot showing proportion of high confidence TGACv1 protein coding genes supported by protein similarity or PacBio data, genes are classified based on overlap with the full set of IWGSC genes.

alternative splicing variants identified from PacBio transcriptome reads. To provide a more comprehensive representation of alternative splicing we subsequently integrated transcript assemblies generated from six strand-specific Illumina libraries (Supplementary Information S8.6 and Table S8.1). This added a further 121,997 transcripts, increasing the number of genes with splice variants from 15% in the TGACv1 annotation to 31% in the supplemented set of transcripts (i.e. incorporating Illumina RNA-Seq assemblies), and increasing the average number of transcripts per gene from 1.26 to 1.88. When considering only HC genes, the number of alternatively spliced genes was increased from 27.48% to 48.80% (2.36 transcripts per gene), similar to that observed in a wide range of plant species (Zhang et al., 2015).

Intron retention (IR) was the prevalent alternative splicing event in wheat (34%) followed by alternative 3' splice sites (A3SS; 27%), exon skipping (ES; 20%), alternative 5' splice sites (A5SS; 19%) and mutually exclusive exons (MXE; 0.04%). This was similar to previous analyses of chromosome 3B (Pingault et al., 2015), and IR is also predominant in barley (Panahi et al., 2015). Alternative splicing coupled to nonsense mediated decay (NMD) regulates gene expression (Lykke-Andersen and Jensen, 2015). We found 22% of all transcripts (17% of all genes) and 29% of multi-exonic HC protein coding transcripts (33% genes) may be potential targets for NMD. Intron retention was the most common splicing event leading to NMD sensitivity, with 40% of IR transcripts identified as potential NMD targets (34% ES, 38% A5SS, 34% A3SS, 26% MXE). This suggests a potentially substantial role for alternative splicing / NMD in regulating gene expression in wheat.

Gene families

HC and LC gene families were analysed separately using OrthoMCL version 2.0 (Li (2003); Supplemental Figures S10.1 and S10.2). Splice variants were removed from the HC gene data set, keeping the representative transcript for each gene model (see Supplemental Information S8.5.6 and S10.1), and datasets were fil-

Table 3: Characteristics of predicted high (HC) and low (LC) confidence wheat genes including coding (mRNA) and long non-coding (ncRNA) RNA.

	All TGAC Models	mRNA HC	mRNA LC	ncRNA HC	ncRNA LC	Repeat-associated
Genes	217,907	104,091	83,217	10,156	9,933	10,510
Transcripts	273,739	154,798	85,778	11,591	10,438	11,134
Transcripts per gene	1.26	1.49	1.03	1.14	1.05	1.06
Transcript mean cDNA size (bp)	1,766.12	2,119.52	1,304.53	1,368.24	1,083.98	1,462.71
Exons per transcript	4.48	5.83	2.8	2.58	2.76	2.27
Exon mean size (bp)	394.15	363.73	465.27	530.25	392.24	644.09
Transcript mean CDS size (bp)	1,165.52	1,361.82	839.97	-	-	891.05
Mono-exonic transcripts	60,322	19,034	30,479	3,061	3,044	4,704
	22.04%	12.30%	35.53%	26.41%	29.16%	42.25%
Genes with alternative splicing	32,616	28,608	2,033	1,037	460	478
	14.97%	27.48%	2.44%	10.21%	4.63%	4.55%

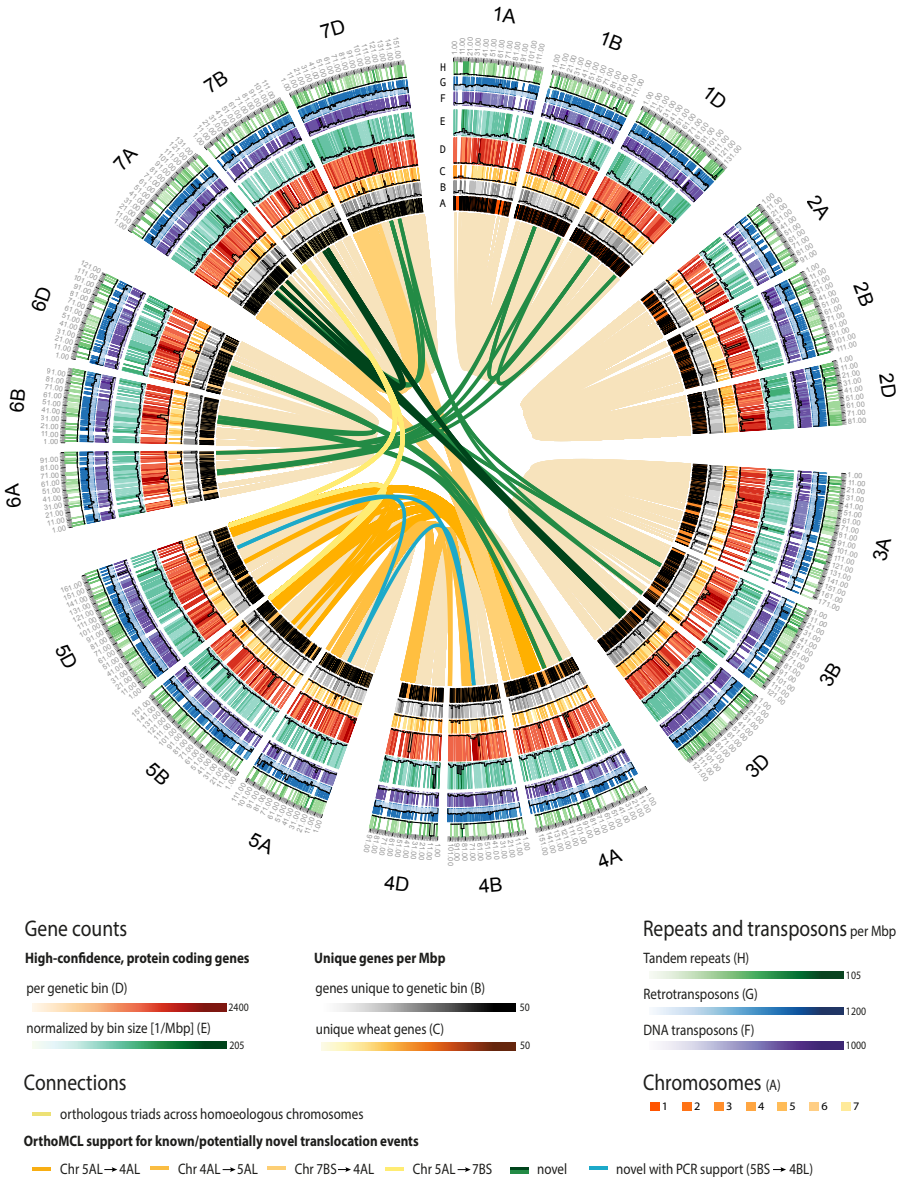

Figure 4: Circular representation of the TGACv1 CS42 assembly. Chromosomes, genetic bins, and genomic features are visualised on the outer rings (**A**-**H**) and interchromosomal links identify known and potentially novel translocation events. (**A**) The seven chromosome groups of the A, B and D genomes, scaled by number of genetic bins (black bands). (**B**-**H**) Combined heatmap/histogram representations of genomic features per genetic bin. With the exception of (**D**) all counts are normalised by the size of the genetic bin in Mbp, calculated as the total size of scaffolds assigned to the bin. (**B**) Distribution of unique genes, i.e. genes that did not have orthologs in a genome-wide OrthoMCL screen. (**C**) Distribution of wheat-specific genes (**D**-**E**) Number of HC protein-coding genes. (**F**) Distribution of DTC, DTM, and DTH DNA transposons (Supplemental Table S7.1). (**G**) Distribution of RLX, RLC, RLG, RXX, and RIX retrotransposons. (**H**) Distribution of tandem duplications. Light yellow links connect homoeologous OrthoMCL triads. Dark yellow-coloured links connect genetic bins harboring OrthoMCL outlier triads (Supplemental Information S6) that identify known translocation events. Dark green links connect genetic bins harbouring at least 3 OrthoMCL outlier triads that may support novel translocation events. The cyan link shows a novel PCR-validated translocation event between chromosomes 5BS-4BL.

tered for premature termination codons and incompatible reading frames. For the HC gene set, a total of 87,519 coding sequences were clustered into 25,132 gene families. The vast majority of HC gene families contained members from the A, B and D genomes, consistent with the relatively recent common ancestry of the A and B genomes and the proposed hybrid origin of the D genome from ancestral A and B genomes (Marcussen et al., 2014). Subsets of gene families and singleton genes (those not clustered into any family) were classified to identify a) genes and families that are A, B or D genome-specific, b) gene families with expanded numbers in one genome, and c) wheat gene families that are expanded relative to other species. These gene sets were analysed for over-represented Gene Ontology (GO) terms, shown in Supplementary File **F2**. Gene families that were significantly expanded in wheat compared to *Arabidopsis*, rice, sorghum and *Brachypodium* include those encoding proteins involved in chromosome maintenance and reproductive processes, and protein and macromolecule modification and protein metabolism processes. The D genome has expanded gene families encoding phosphorylation, phosphate metabolism and macromolecule modification activities, while the B genome has expanded gene families encoding components of chromosome organisation, DNA integration and conformation/unwinding, and telomere maintenance. The B genome is derived from the *Sitopsis* section of the Triticeae, which has contributed genomes to many polyploid Triticeae species (Riley et al., 1961), suggesting B genomes may have contributed gene functions for establishing and maintaining polyploidy in the Triticeae. This is supported by the location of the major chromosome pairing *Ph1* locus on chromosome 5B (Griffiths et al., 2006).

Genome organisation

A corrected version of the POPSEQ genetic map (Chapman et al., 2015) was used to order TGACv1 scaffolds along chromosomes (Supplemental information S5). This uniquely assigned 128,906 (17.5%) of the 735,943 TGACv1 scaffolds to 1051 of 1187 genetic bins (class 1, Supplemental Information S5) to form the final TGACv1 map. The total length of these scaffolds is 8,551,191,083bp, representing 63.68% of the TGACv1 assembly and 50.52% of the 17Gbp wheat genome. The TGACv1 map includes 3927 (3.05%) scaffolds that were not previously assigned to a chromosome arm (Supplemental Information S5) and 380 (0.295%) scaffolds whose CSS-based chromosome assignment disagrees with its position according to the TGACv1 map. A further 13,019 (1.77%) scaffolds were ambiguously assigned to different cM positions on the same chromosome (class 2), 489 (0.07%) scaffolds were assigned to homoeologous chromosomes (class 3), and 3320 (0.45%) scaffolds had matching markers with conflicting bin assignment (class 4). The TGACv1 map encompasses 38,958 of the 53,792 scaffolds containing at least one annotated HC protein-coding gene (72.42%), comprising gene sequences of 307,085 968bp (73.28% of total predicted gene sequence space). In total, we were able to assign genetic bins to 75,623 (72.65%) of the HC genes.

Chromosomal locations of related genes were identified by anchoring to the TGACv1 map and are displayed in Figure 4. Analysis of OrthoMCL outlier triads (Supplemental Information S6 and S10) provided genomic support for known ancestral reciprocal translocations between chromosome arms 4AL and 5AL, a combination of pericentromeric inversions between chromosome arms 4AL and 5AL, and a reciprocal exchange between chromosome arms 4AL and 7BS (Devos et al., 1995). Several putative novel chromosomal translocations were also identified (Figure 4 and Supplemental File **F3**). As these may have originated in the parental lines used in the POPSEQ map rather than in CS42, nine genes in the predicted translocations (6 previously known and 3 novel) were tested using PCR assays on Chinese Spring chromosomal deletion stocks (Sears, 1966). Three known translocation events, 4AL-5AL and 4AL-7BS (Devos et al., 1995) and 5AL-7BS (Ma et al., 2013), and one previously unidentified translocation 5BS-4BL were validated by PCR assays.

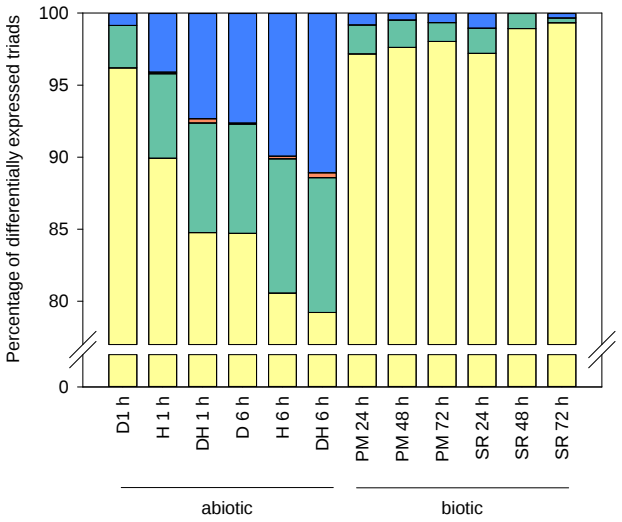

Figure 5: Response of differentially expressed (DE) triads to stress treatments according to the number and pattern of DE homoeologs. Triads were classified as having one homoeolog DE (yellow), two homoeologs DE with same direction of change (green), three homoeologs DE with same direction of change (orange), or opposite direction of change between DE homoeologs (blue). The stresses applied were drought (D), heat (H), drought and heat combined (DH), powdery mildew (PM) and stripe rust (SR), with the duration of stress application indicated in hours (h).

Gene expression

To explore global gene expression patterns we mapped multiple wheat RNA-Seq datasets to the TGACv1 transcriptome (Supplemental Table S11.1). Seventy-five percent of RNA-Seq reads mapped to the TGACv1 transcriptome (Supplemental Table S11.1), and 78% of the HC protein coding transcripts were expressed above the background level of 2 tpm (Wagner et al., 2013). Interestingly, 23% of the LC genes were also expressed above 2 tpm. Expression levels of genes across chromosomes were similar, with the exception of 19 genetic bins that had increased expression (defined as "hotspots" with a median expression level greater than 20 tpm, containing on average 5 genes) across the six tissues examined (Supplemental Figure S11.1). Hotspots tended to be enriched for genes encoding components of the cytoskeleton, ribosome biogenesis, and nucleosome assembly that were expressed at high levels in all tissues. Other notable hotspots were enriched in genes of photosystem I formation in leaf tissues, and nutrient reservoir activity in seed tissues.

The more complete and accurate annotation provided an opportunity to analyse patterns of transcript levels in homoeologous triads. Transcript levels of 9642 triads were analysed in response to biotic and abiotic stress using publicly available RNA-Seq (Supplemental Table S11.2), selected as they all used 7-day old seedlings, were replicated, and assessed dynamic transcriptional responses to standardised treatments. Across treatments, 26% (2424 of 9159) of

expressed triads showed higher expression in one or two genomes in at least one stress condition (rather than balanced expression of three genomes; see Supplemental Information S11.5). Abiotic stress led to more differentially-regulated transcripts, compared to biotic stress responses, across all three genomes. To assess the conservation of this stress response between homoeologs, we classified each homoeolog as either up-regulated (greater than 2-fold change, UP), down-regulated (less than 0.5-fold change, DOWN) or flat (between 0.5 - 2-fold change). We then assessed whether the individual homoeolog response to stress compared to control conditions was consistent (Supplemental Table S11.3). 80% ($\pm$ 5.1% SE) of triads were not differentially expressed in response to the stress treatments and were excluded from further analysis. The most frequent pattern of differential triad expression was a single homoeolog UP or DOWN, with the other two remaining flat (Figure 5; 79–99% across conditions). Triads in which either all homoeologs were expressed in the same pattern ("3 UP" or "3 DOWN") were rare, as were triads in which homoeologs were expressed in opposite directions. This is consistent with Liu et al. (2015), who identified between 13% and 41% of homoeolog triads in which homoeologs did not respond to the same degree in response to stress conditions.

The genomic context of differences in homoeolog expression was explored in genomic regions containing at least five high confidence genes in syntenic order on all three genomes, of which at least one homoeolog was expressed over background levels in root, shoot and endosperm tissue at 10 and 20 days post anthesis (DPA) (Supplemental Table S11.1, DRP000768 and ERP004505; Pfeifer et al. (2014)). Of the four blocks meeting these criteria, one showed equal expression of all fifteen homoeologs in at least one of the tissues, while the other three blocks showed unbalanced expression of at least one homoeolog (Supplementary Figure S11.3). All blocks exhibited major structural and promoter sequence differences, and variant transcription start sites (Supplementary Figure S11.3). These multiple types of genomic differences all have the potential to contribute to unbalanced expression. To facilitate further expression studies the expression atlas at `http://www.wheat-expression.com` has been updated with the TGACv1 annotation and expression data from 424 RNA samples (Borrill et al., 2016).

Gene families of agronomic interest

Wheat disease resistance genes

Plant disease resistance (*R*-) genes termed Nucleotide Binding Site- Leucine Rich Receptors (NBS-LRRs; Dodds and Rathjen (2010)) are challenging to assemble as they are often organised in multi-genic clusters with many tandem duplications and rapid pseudogenisation. The TGACv1 assembly contains 2595 NBS-containing genes (Table 4) of which 1185 are NBS-LRR genes. Among these, 98% have complete transcripts compared to only 2% in the CSS assembly. We also used NLR-parser (Steuernagel et al., 2015) to predict the coiled-coil (CC-) NBS-LRR subclass of *R*-genes. We identified 859 complete CC-NBS-LRR genes supported by specific MEME motifs (Jupe et al., 2012), compared to 225 in the CSS assembly (Table 4). The total of 1185 wheat NBS-NLRs was consistent with that found in diploid wheat progenitors (402 NLRs in *T. urartu*) and diploid relatives (438 in *O. sativa*; Sarris et al. (2016)). Nearly 90% of CS42 *R*-genes were unambiguously assigned to chromosome arms and 57% (674/1185) were anchored to the TGACv1 map. The number of *R*-genes per scaffold ranged from 1 to 31, compared to only 2 to 3 *R*-gene per scaffold in the CSS wheat assembly (IWGSC, 2014). This finding is corroborated by BAC sequence assemblies (Supplementary Figure S12.1).

Gluten genes

Glutens form the major group of grain storage proteins accounting for 10–15% of grain dry weight and confer visco-elastic properties essential for bread-making (Shewry et al., 1995). Gluten genes encode proteins rich in glutamines and prolines that form low complexity sequences composed of PxQ motifs, and occur in tandem repeats in highly complex loci that have posed significant challenges for their assembly and annotation. We characterised the gluten genes in the TGACv1 assembly and showed that most of the known genes were fully assembled. Gluten loci, while still fragmented, exhibit much greater contiguity than in the CSS assembly (IWGSC, 2014) with up to 6 genes per scaffold (Supplementary Figure S12.2). We identified all assembly regions with nucleotide similarity to publicly available gluten sequences, adding an additional 33 gluten genes to the annotation and manually correcting 21 gene models. In total, we identified 105 full length or partial gluten genes and 13 pseudogenes in the TGACv1 assembly (Table 4, Supplementary infomation S12.2).

The gibberellin biosynthetic and signalling pathway

Mutations in the gibberellin (GA) biosynthetic and signal transduction pathways have been exploited in wheat, where gain-of-function mutations in the GA signalling protein Rht-1 confer GA-insensitivity and a range of dwarfing effects. Most modern wheat cultivars carry semi-dominant *Rht-1* alleles (Phillips, 2016), but these alleles also confer negative pleiotropic effects, including reduced male fertility and grain size. Hence, there is considerable interest in developing alternative dwarfing alleles based on GA-biosynthetic genes such as *GA20ox2*. A prerequisite for this is access to a complete set of genes encoding the biosynthetic pathway. Figure 6 shows that the TGACv1 assembly contains full-length sequences for 67 of the expected 72 GA pathway genes, in contrast to only 23 genes in the CSS assembly (IWGSC, 2014). Two paralogues of *GA20ox3* on chromosome 3D are separated by 460kbp, and *GA1ox-B1* and *GA3ox-B3* are separated by 3.2kbp, suggesting common ancestry of these two enzymes with different catalytic activities (Pearce et al., 2015).

Discussion

Access to a complete and robust wheat genome assembly is essential for the continued improvement of wheat, a staple crop of global significance with 728m tonnes produced in 2014 (`http://fenix.fao.org/faostat/beta/en/#home`). The capacity to assemble and annotate wheat genomes accurately, rapidly and cost-effectively addresses key social, economic and academic priorities by facilitating trait analyses, by exploiting diverse germplasm resources, and by accelerating plant breeding. However, polyploidy and the extensive repeat structure present in wheat have limited the completeness of previous assembly efforts (Brenchley et al., 2012; IWGSC, 2014; Chapman et al., 2015), reducing their utility.

Here we report the most complete wheat genome assembly to date, representing almost 80% of the 17Gbp genome in large scaffolds. We combined high-quality PCR-free libraries and precisely size-selected LMP libraries (Heavens et al., 2015) with the w2rap assembly software (Clavijo, 2016) to generate contiguous and complete assemblies from relatively low ($\sim$33$\times$) Illumina paired-end

Table 4: Disease Resistance and Gluten gene repertoires in the TGACv1 assembly. Resistance genes were identified by their characteristic domain architecture (Sarris et al., 2016). Gluten genes were identified by sequence similarity to either a gliadin, glutenin, or generic prolamin class, representing prolamin-like glutens discovered in oat (avenin), wheat (farinin), or barley (hordein). See Supplementary Information, Section 12.

R-genes	CSS	TGACv1
NBS-containing (Pfam)	1,224	2,595
fragmented	1,188	65
complete transcript	36	2,530
number of scaffolds	1,195	1,853
max genes per scaffold	3	31
NBS-LRR (Pfam)	627	1,185
partial genes	611	11
full length genes	16	1,174
number of scaffolds	613	979
max genes per scaffold	2	13
CC-NBS-LRR (NLR-parser)	225	859

Gluten genes	CDS	Pseudogenes
Gliadins		
alpha	29	9
gamma	18	0
unknown	14	1
omega	10	0
Glutenins		
HMW	6	1
LMW	16	1
Prolamins		
Avenin	4	0
Farinin	4	0
Globulin	2	1
Hordein	1	0
unknown	1	0
Total	105	13

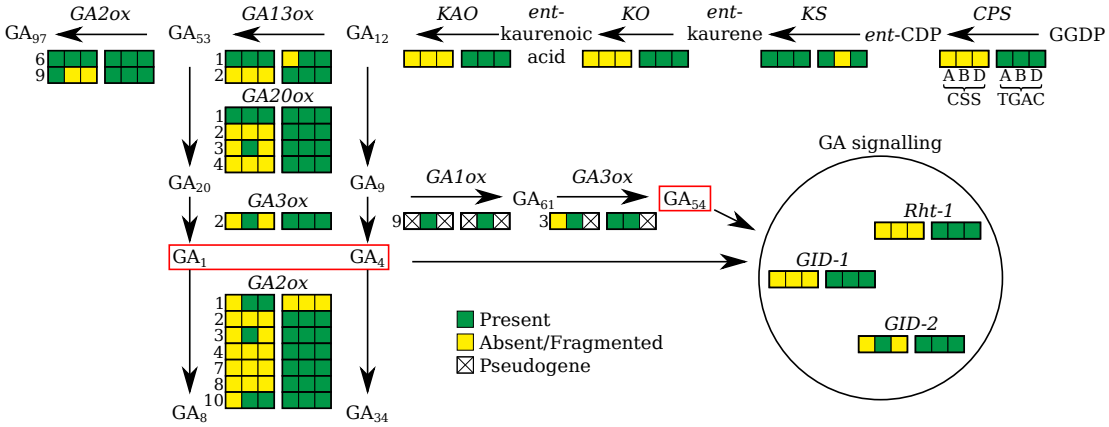

Figure 6: Genes encoding the Gibberellin Biosynthetic and Signalling pathway in bread wheat. The GA biosynthesis, inactivation and signal transduction pathway, illustrating the representation of the gene sequences in CSS and TGACv1 assemblies. If more than one paralog is known for a gene, its number according to the classification by Pearce et al. (2015) is indicated on the left of the box. Bioactive GAs are boxed in red.

read coverage and long-mate pair libraries. The contiguity of the TGACv1 assembly allowed us to create a greatly improved gene annotation supported by extensive transcriptome data. Over 78% of the 104,091 high confidence protein coding genes are fully supported by RNA-Seq data. These improvements idenitfied 22,904 genes that were absent from previous wheat gene sets (IWGSC, 2014; Choulet et al., 2014), almost all of which have a homolog in other species (Figure 3**B**). The robustness of the annotation is further supported by the use of high-quality PacBio data and agreement with proteomic data, with 42% of the HC gene models supported by sequenced peptides. This new wheat gene set provides an improved foundation for wheat research. Finally, incorporation of strand-specific Illumina RNA-Seq libraries into the annotation showed that nearly half of the high confidence genes were alternatively spliced, in line with observations in many other plants (Zhang et al., 2015).

A well-defined gene set in large sequence scaffolds is an essential foundation for trait analyses in wheat. We identified the complement of disease resistance genes, gluten protein genes that confer nutritional and bread-making quality of wheat grains, and the set of gibberellin biosynthetic and signal transduction genes that are important determinants of crop height and yield. An accurate gene set is also essential for understanding expression of gene families in complex allopolyploid genomes. We observed that 20% of homoeologous triads showed differential expression in seedling leaves subject to biotic and abiotic stress conditions. This is consistent with co-expression analyses in developing grains (Pfeifer et al., 2014), where most differentially expressed genes were single homoeologs being up/down-regulated. Taken together, these results identify widespread sub-functionalisation of homoeologous genes due to differential regulation. The new assembly and annotation will enable the identification of multiple sequence differences in promoters, transcription start sites, gene splicing and other features among strict homoeologs, providing a foundation for systematic analyses of the causes of these differences.

Our rapid, accurate and cost-effective assembly approach will enable multiple wheat and Triticeae genomes to be assembled in robust and comparable ways, using relatively inexpensive sequencing technologies and open-source software. These assemblies will reveal a wide spectrum of genetic variation, including large-scale structural changes such as translocations and chromosome additions that are known to play a major role in the adaptation of the

wheat crop to different growing environments. By adopting this pan-genomics approach, we will enrich our understanding of complex genome evolution and the plasticity of genome regulation, and empower new approaches to wheat improvement.

Methods

DNA library preparation and sequencing

A full description of the DNA preparation and sequencing methods is in Supplemental Information. PCR-free Paired-End (PE) libraries were sequenced using 2×250bp reads on HiSeq2500 platforms for contig generation. Tight, Amplification-free, Large insert Libraries (TALL) libraries and Nextera LMP libraries (Heavens et al., 2015) were used for scaffolding. Insert size distributions (Supplemental Figures S4.1, S4.3 and S4.2) were checked by mapping to the CS42 chromosome 3B pseudo-molecule (Choulet et al., 2014) using the DRAGEN co-processor (`http://www.edicogenome.com/dragen/`).

Assembly

Assembly was performed using the Wheat/Whole Genome Robust Assembly Pipeline, w2rap (Clavijo, 2016). It combines the w2rap-contigger, based on DISCOVAR *de novo* (Weisenfeld et al., 2014), an LMP preparation approach based on FLASH (Magoc and Salzberg, 2011) and Nextclip (Leggett et al., 2014), and scaffolding with SOAPdenovo2 (Luo et al., 2012). The w2rap-contigger takes advantage of DISCOVAR (Weisenfeld et al., 2014; Love et al., 2016) algorithms to preserve sequence variation during assembly, but has been further developed to enable processing of much larger data volumes and complex genomic repeats. The paired end read dataset was assembled into contigs on a SGI UV200 machine with 64 cores and 7TB of shared RAM, using the default settings of the w2rap-contigger from `https://github.com/bioinfologics/w2rap-contigger/releases/tag/CS42_TGACv1`. Contigs were scaffolded using the PE, LMP and TALL reads and the SOAPdenovo2 (Luo et al., 2012) prepare->map->scaffold pipeline, run at k=71 for both the prepare and map steps on the same machine using 128 cores. Contigs and scaffolds were QC'ed using KAT spectra-cn plots (Mapleson et al., 2016a) to assess motif representation.

Gene annotation

A high quality gene set for wheat was generated using a custom pipeline integrating wheat-specific transcriptomic data, protein similarity, and evidence-guided gene predictions generated with AUGUSTUS (Stanke and Morgenstern, 2005). Full methods are in Supplementary Information 8. RNA-Seq reads (ERP004714, ERP004505, and 250bp paired-end strand-specific reads from six different tissues) were assembled using four alternative assembly methods (Trapnell et al., 2010; Pertea et al., 2015; Song et al., 2016; Haas et al., 2013) and integrated with PacBio transcripts into a coherent and non-redundant set of models using Mikado (Venturini et al., 2016). PacBio reads were then classified according to protein similarity and a subset of high quality (e.g. full length, canonical splicing, non-redundant) transcripts used to train an AUGUSTUS wheat-specific gene prediction model. AUGUSTUS was then used to generate a first draft of the genome annotation, using as input Mikado-filtered transcript models, reliable junctions identified with Portcullis (Mapleson et al., 2016b), and peptide alignments of proteins from five close wheat relatives (*B. distachyon*, maize, rice, *S. bicolor* and *S. italica*). This draft annotation was refined by correcting probable gene fusions, missing loci and alternative splice variants. The annotation was functionally annotated and all loci were assigned a confidence rank based on their similarity to known proteins and their agreement with transcriptome data.

Data access

All data generated in this study has been submitted to the EMBL-EBI European Nucleotide Archive. PE and LMP reads used for genome assembly and scaffolding are available in study accession PRJEB15378, and assembled scaffolds are available in study accession PRJEB11773. The Illumina and PacBio reads used for genome annotation are available in study accession PRJEB15048. The assembly and annotation is available in Ensembl Plants (release 32) at `http://plants.ensembl.org/Triticum_aestivum` and from the Earlham Institute server at `http://opendata.earlham.ac.uk/Triticum_aestivum/TGAC/v1`. BLAST services for these datasets are available at `https://wheatis.tgac.ac.uk/grassroots-portal/blast`.

Competing interest statement

The authors declare no competing interests.

Acknowledgements

We thank Burkhard Steuernagel for assistance with NLR-parser. This work was funded by a Biotechnology and Biological Sciences Research Council (BBSRC) strategic LOLA Award to MWB and CU (BB/J003557/1), MDC (BB/J003743/1), PK (BB/J00328X/1) and AP (BB/J003913/1), a BBSRC Anniversary Future Leader Fellowship (BB/M014045/1) to PB, BBSRC Institute Strategic Programme Grants GRO (BB/J004588/1) to MWB and CU, "2020 Wheat" (BBS/E/C/00005202) to AP, and Bioinformatics BB/J004669/1 to FDP. The German Ministry of Education and Research (BMBF) grant 031A536 "de.NBI" supported MS, and the Australian Research Council (LP120200102, CE140100008) and Agilent Technologies Australia supported AHM. Next-generation sequencing and library construction was delivered via the BBSRC National Capability in Genomics (BB/J010375/1) at the Earlham Institute (EI, formerly The Genome Analysis Centre, Norwich), by members of the Platforms and Pipelines Group. Open data access, and BLAST databases and service, are provided by the EI Data Infrastructure group.

References

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., and McVean, G. A., 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–73.

Abdel-Ghany, S. E., Hamilton, M., Jacobi, J. L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., and Reddy, A. S. N., 2016. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, **7**:11706.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A., 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**(2):R18.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Bento, P., Da Silva, C., Labadie, K., Alberti, A., *et al.*, 2014. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, **5**.

Bishara, A., Liu, Y., Weng, Z., Kashef-Haghighi, D., Newburger, D. E., West, R., Sidow, A., and Batzoglou, S., 2015. Read clouds uncover variation in complex regions of the human genome. *Genome Research*, **25**(10):1570–1580.

Blanc, G. and Wolfe, K. H., 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Society*, **16**(July):1667–1678.

Borrill, P., Ramirez-Gonzalez, R., and Uauy, C., 2016. expVIP: a Customizable RNA-seq Data Analysis and Visualization Platform. *Plant Physiology*, **170**(4):2172–2186.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G. L. A., D'Amore, R., Allen, A. M., McKenzie, N., Kramer, M., Kerhornou, A., Bolser, D., *et al.*, 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, **491**(7426):705–710.

Chaisson, M. J. P., Wilson, R. K., and Eichler, E. E., 2015. Genetic variation and the de novo assembly of human genomes. *Nature Reviews Genetics*, **16**(11):627–640.

Chapman, J. A., Ho, I., Sunkara, S., Luo, S., Schroth, G. P., and Rokhsar, D. S., 2011. Meraculous: De Novo Genome Assembly with Short Paired-End Reads. *PLoS ONE*, **6**(8):e23501.

Chapman, J. A., Mascher, M., Buluç, A., Barry, K., Georganas, E., Session, A., Strnadova, V., Jenkins, J., Sehgal, S., Oliker, L., *et al.*, 2015. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. *Genome Biology*, **16**(1):26.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., *et al.*, 2014. Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**(6194):1249721–1249721.

Clavijo, B., 2016. w2rap. `https://github.com/bioinfologics/w2rap-contigger`.

Devos, K. M., Dubcovsky, J., Dvořák, J., Chinoy, C. N., and Gale, M. D., 1995. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theoretical and Applied Genetics*, **91**(2):282–288.

Dodds, P. N. and Rathjen, J. P., 2010. Plant immunity: towards an integrated view of plant–pathogen interactions. *Nature Reviews Genetics*, **11**(8):539–548.

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., *et al.*, 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. *Nature*, **477**:419–423.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Burton, J. N., Walker, B. J., Sharpe, T., Hall, G., Shea, T. P., Sykes, S., *et al.*, 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, **108**(4):1513–1518.

Gordon, D., Huddleston, J., Chaisson, M. J. P., Hill, C. M., Kronenberg, Z. N., Munson, K. M., Malig, M., Raja, A., Fiddes, I., Hillier, L. W., *et al.*, 2016. Long-read sequence assembly of the gorilla genome. *Science*, **352**(6281):aae0344–aae0344.

Griffiths, S., Sharp, R., Foote, T. N., Bertin, I., Wanous, M., Reader, S., Colas, I., and Moore, G., 2006. Molecular characterization of Ph1 as a major chromosome pairing locus in polyploid wheat. *Nature*, **439**(7077):749–752.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., *et al.*, 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, **8**(8):1494–1512.

Heavens, D., Accinelli, G. G., Clavijo, B., and Clark, M. D., 2015. A method to simultaneously construct up to 12 differently sized Illumina Nextera long mate pair libraries with reduced DNA input, time, and cost. *BioTechniques*, **59**(1):42–45.

IWGSC, 2014. A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**(6194):1251788–1251788.

Jupe, F., Pritchard, L., Etherington, G. J., MacKenzie, K., Cock, P. J., Wright, F., Sharma, S. K., Bolser, D., Bryan, G. J., Jones, J. D., *et al.*, 2012. Identification and localisation of the NB-LRR gene family within the potato genome. *BMC Genomics*, **13**(1):75.

Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J., 2009. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, **6**(4):291–295.

Lam, H. Y. K., Clark, M. J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F. E., Habegger, L., Ashley, E. A., Gerstein, M. B., *et al.*, 2011. Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology*, **30**(1):78–82.

Leggett, R. M., Clavijo, B. J., Clissold, L., Clark, M. D., and Caccamo, M., 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. *Bioinformatics*, **30**(4):566–568.

Li, L., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, **13**(9):2178–2189.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al.*, 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, **326**(5950):289–93.

Liu, Z., Xin, M., Qin, J., Peng, H., Ni, Z., Yao, Y., and Sun, Q., 2015. Temporal transcriptome profiling reveals expression partitioning of homeologous genes contributing to heat and drought acclimation in wheat (Triticum aestivum L.). *BMC Plant Biology*, **15**(1):152.

Love, R. R., Weisenfeld, N. I., Jaffe, D. B., Besansky, N. J., and Neafsey, D. E., 2016. Evaluation of DISCOVAR de novo using a mosquito sample for cost-effective short-read genome assembly. *BMC Genomics*, **17**(1):187.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., *et al.*, 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**(1):18.

Lykke-Andersen, S. and Jensen, T. H., 2015. Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology*, **16**(11):665–677.

Ma, J., Stiller, J., Berkman, P. J., Wei, Y., Rogers, J., Feuillet, C., Dolezel, J., Mayer, K. F., Eversole, K., Zheng, Y.-L., *et al.*, 2013. Sequence-Based Analysis of Translocations and Inversions in Bread Wheat (Triticum aestivum L.). *PLoS ONE*, **8**(11):e79329.

Magoc, T. and Salzberg, S. L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**(21):2957–2963.

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B., 2016a. KAT: A K-mer Analysis Toolkit to quality control NGS datasets and genome assemblies. *bioRxiv*, :64733.

Mapleson, D. L., Venturini, L., and Swarbreck, D., 2016b. Portcullis. `https://github.com/maplesond/portcullis`.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., Olsen, O.-A., *et al.*, 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**(6194):1250092–1250092.

Moore, G., Devos, K., Wang, Z., and Gale, M., 1995. Cereal Genome Evolution: Grasses, line up and form a circle. *Current Biology*, **5**(7):737–739.

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., Lee, J., Chu, C., Lin, C., Džakula, Ž., *et al.*, 2016. A hybrid approach for de novo human genome sequence assembly and phasing. *Nature Methods*, **13**(7):587–590.

Panahi, B., Mohammadi, S. A., Khaksefidi, R. E., Fallah Mehrabadi, J., and Ebrahimie, E., 2015. Genome-wide analysis of alternative splicing events in Hordeum vulgare : Highlighting retention of intron-based splicing and its possible function through network analysis. *FEBS Letters*, **589**(23):3564–3575.

Pearce, S., Huttly, A. K., Prosser, I. M., Li, Y.-d., Vaughan, S. P., Gallova, B., Patil, A., Coghill, J. A., Dubcovsky, J., Hedden, P., *et al.*, 2015. Heterologous expression and transcript analysis of gibberellin biosynthetic genes of grasses reveals novel functionality in the GA3ox family. *BMC Plant Biology*, **15**(1):130.

Pendleton, M., Sebra, R., Pang, A. W. C., Ummat, A., Franzen, O., Rausch, T., Stütz, A. M., Stedman, W., Anantharaman, T., Hastie, A., *et al.*, 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*, **12**(8):780–786.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, **33**(3):290–295.

Pfeifer, M., Kugler, K. G., Sandve, S. R., Zhan, B., Rudi, H., Hvidsten, T. R., International Wheat Genome Sequencing Consortium, Mayer, K. F. X., and Olsen, O.-A. O.-A., 2014. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, **345**(6194):1250091.

Phillips, A. L., 2016. Genetic control of gibberellin metabolism and signalling in crop improvement. In *Annual Plant Reviews, Volume 49*, pages 405–430. John Wiley & Sons, Ltd, Chichester, UK.

Pingault, L., Choulet, F., Alberti, A., Glover, N., Wincker, P., Feuillet, C., and Paux, E., 2015. Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. *Genome Biology*, **16**(1):29.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., *et al.*, 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*, **26**(3):342–350.

Riley, R., Kimber, G., and Chapman, V., 1961. Origin of genetic control of diploid-like behavior of polyploid wheat. *Journal of Heredity*, **52**(1):22–25.

Sarris, P. F., Cevik, V., Dagdas, G., Jones, J. D. G., and Krasileva, K. V., 2016. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. *BMC Biology*, **14**(1):8.

Sears, E. R., 1966. Nullisomic-Tetrasomic Combinations in Hexaploid Wheat. In *Chromosome Manipulations and Plant Genetics*, pages 29–45. Springer US, Boston, MA.

Shewry, P. R., Tatham, A. S., Barro, F., Barcelo, P., and Lazzeri, P., 1995. Biotechnology of breadmaking: unraveling and manipulating the multi-protein gluten complex. *Bio/technology*, **13**(11):1185–1190.

Song, L., Sabunciyan, S., and Florea, L., 2016. CLASS2: accurate and efficient splice variant annotation from RNA-seq reads. *Nucleic Acids Research*, **44**(10):e98–e98.

Staiger, D. and Brown, J. W. S., 2013. Alternative Splicing at the Intersection of Biological Timing, Development, and Stress Responses. *The Plant Cell*, **25**(10):3640–3656.

Stanke, M. and Morgenstern, B., 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*, **33**(Web Server):W465–W467.

Steuernagel, B., Jupe, F., Witek, K., Jones, J. D. G., and Wulff, B. B. H., 2015. NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics*, **31**(10):1665–1667.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, **28**(5):511–515.

Treangen, T. J. and Salzberg, S. L., 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, **13**(1):36–46.

Venturini, L., Caim, S., Mapleson, D. L., Kaithakottil, G. G., and Swarbreck, D., 2016. Mikado. `https://github.com/lucventurini/mikado`.

Vogel, J. P., Garvin, D. F., Mockler, T. C., Schmutz, J., Rokhsar, D., Bevan, M. W., Barry, K., Lucas, S., Harmon-Smith, M., Lail, K., *et al.*, 2010. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**(7282):763–768.

Wagner, G. P., Kin, K., and Lynch, V. J., 2013. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences*, **132**(3):159–164.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D., and Jaffe, D. B., 2016. Direct determination of diploid genome sequences. *bioRxiv*, :070425.

Weisenfeld, N. I., Yin, S., Sharpe, T., Lau, B., Hegarty, R., Holmes, L., Sogoloff, B., Tabbaa, D., Williams, L., Russ, C., *et al.*, 2014. Comprehensive variation discovery in single human genomes. *Nature Genetics*, **46**(12):1350–1355.

Zhang, C., Yang, H., and Yang, H., 2015. Evolutionary Character of Alternative Splicing in Plants. *Bioinformatics and biology insights*, **9**(Suppl 1):47–52.

Zimin, A. V., Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Yorke, J. A., Dvorak, J., and Salzberg, S., 2016. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the mega-reads algorithm. *bioRxiv*, :66100.