

Additional File 1. Supplementary Methods

Wheat chromosome 3DL Assembly and Annotation

A total of 2,703 PacBio scaffolds from Chinese Spring chromosome 3DL were retrieved (Zimin et al. 2017) and overlapped and extended to 1,751 scaffolds using Fosill mate-paired reads as described previously (Lu et al. 2018). These 1,751 PacBio scaffolds were then integrated with our 3DL BAC scaffolds (International Wheat Genome Sequencing Consortium (IWGSC) et al. 2018; Lu et al. 2018) to generate a final set of 524 super-scaffolds using nucmer (default setting) and delta-filter (-i 99 -l 500) in MUMmer v3.23 (Kurtz et al. 2004) to identify joins between BAC and PacBio scaffolds, and further assembled using scaffolder v0.5.0 software (Barton and Barton 2012). BAC scaffolds were aligned to 524 templates using nucmer (default setting) and further filtered using delta-filter (-i 99 -r -q -l 500), and then were further merged to form 504 scaffolds by a custom Perl script designed to merge BACs for each template scaffolds using quickmerge v0.2 (-hco 5.0 -c 1.5 -l 1000 -ml 500) software (Chakraborty et al. 2016). A 3DL pseudomolecule was made by mapping 504 scaffolds to the IWGSC chr3D pseudomolecule (International Wheat Genome Sequencing Consortium (IWGSC) et al. 2018) using MUMmer v3.23. All scaffolds were localised and assigned to a specific order and strand and then linked using scaffolder v0.5.0. Order discrepancies were manually corrected. One hundred Ns were placed between two neighbour scaffolds to mark the sequence gap.

Repetitive sequences of the 3DL pseudomolecule were identified using RepeatMasker (v4.0.7) with the submodule Tandem Repeat Finder (v4.09) (Benson 1999). Sequence comparisons were performed using the alignment software RMBlast (v2.2.28). A database of repetitive DNA (109,726 sequences) were collected from Repbase RepBase (v20170127; 45,447 sequences) (Bao et al. 2015), PGSB-REdat(v9.3p; 61,730 sequences) (Nussbaumer et al. 2013) and 2,549 repetitive substrings (4,671,512 bp) with at least 10-time appearance that were discovered by RepeatScout (v1.0.5) (Price et al. 2005). Those 2,549 repetitive substrings were classified by PASTE Classifier in REPET package (v2.5) (Hoede et al. 2014) using the REPbase databases (REPET edition v20.05 with 38,777 nucleotide sequences and 24,192 amino acid sequences).

Genes were identified *ab initio* on the chromosome 3DL pseudomolecule using Augustus (v3.0.3) (Stanke et al. 2006) trained for hexaploid wheat. A training set was made from 7,264 CDS sequences from wheat 3B (Choulet et al. 2014) mapped to 3DL pseudomolecule using the PASA pipeline (version r20140417; parameters: --runpasa -a gmap,blat) (Haas et al. 2008). Those genes, overlapped with each other or with similarity $\geq 70\%$ or bad genes (number of bases cannot be divided by three), are removed and not used for following analysis. It generated a total of 787 gene models, of which 687 were used to train Augustus to generate wheat-specific prediction parameters, and the remaining were used for testing the precision of these trained prediction parameters. To clearly show and refine the wheat exon borders covered by Paragon RNAseq reads, a two-step RNAseq mapping strategy was employed using Gsnap2Augustus (<https://github.com/lufuhao/Gsnap2Augustus>). Firstly, RNAseq reads were mapped to the 3DL pseudomolecule using GSNAP (version: 2017-12-29; parameter: --nofails --expand-offsets=1 --novelsplicing=1 -B 3 --localsplicedist=10000 --npaths=30 --format=sam) (Wu and Nacu 2010). This generated a preliminary *ab initio* prediction of BAM intron hints. An exon-exon database was created using the script 'intron2exex.pl' included in Augustus. Secondly, RNAseq reads were mapped to the exon-exon database and the coordinates were calibrated back to the 3DL pseudomolecule using 'samMap.pl' included in Augustus. Precise BAM intron hints were generated by bam2hints included in Augustus. Finally, Augustus calculated *ab initio* predictions based on BAM intron hints.

EST evidence for gene predictions were generated using Exonerate (v2.4.0; parameter: --model est2genome --percent 70 --score 100 --showvulgar yes --bestn 10 --minintron 20 --softmaskquery no --softmasktarget no --showalignment no --showtargetgff yes --geneseed 250) (Slater and Birney 2005) using 3 EST datasets: *de novo* assembly of Paragon RNAseq reads using the Trinity assembler (v2.0.6; parameter: --genome_guided_max_intron 10000) (Grabherr et al. 2011); the Triticeae Full-Length CDS Database (TriFLDB) (Mochida et al. 2009); and 1,551,792 *Triticum aestivum* ESTs downloaded from NCBI. Protein evidence was produced by mapping protein sequences from 10 species (Additional File 2: Table S8) to the 3DL pseudomolecule using GenomeThreader (v1.6.2; parameter: -gff3out yes -skipalignmentout yes) (Gremme et al. 2005). EvidenceModuler (EVM; v20120625) (Haas et al. 2008) was used to combine *ab initio* gene predictions, EST and protein alignments into weighted (1:5:10) consensus gene structures. The pipeline and converters are available at <https://github.com/lufuhao/AutoEVM>.

EVM gene models were manually curated using the Integrative Genomics Viewer (IGV; v2.3.60) (Thorvaldsdóttir et al. 2013) and given a confidence score (0-5), in which evidence from RNAseq peaks, *ab initio* predictions, protein alignments, *de novo* EST alignments and NCBI EST alignments separately accounted for a single confidence value. Pseudogenes were annotated as genes with good exon-intron structures that conformed to the GT-AG intron rule, but had no consensus/translatable CDS. Finally, manually curated gene modules were transferred to the chr3DL pseudomolecule using RATT (in PAGIT v1) (Otto et al. 2011) and some remaining modules using custom scripts (<https://github.com/lufuhao/ExonerateTransferAnnotation>), which employed Exonerate to map transcripts and CDS sequences separately and then integrated them together into final gene models.

Similarity searches were carried out using blastx in the BLAST+ toolkit (v2.6.0; parameter: -evalue 1e-6 -outfmt 5 -show_gis -num_alignments 20 -max_hsps 20) against the NCBI NR (v20171024) protein database. The best 20 hits with E-value 1E-6 for each sequence were retained and used for GO-mapping. GO terms (version 07-Jan-2017) associated with these candidate sequences were assigned to each gene using BLAST2GO (v2.5; parameter: -v -annot -dat -img -ips ipsr -annex -goslim) (Götz et al. 2008). The web plotting tool WEGO (Ye et al. 2006) was used to draw GO annotations.

Gene Expression Analyses

RNA samples from Paragon and *Ae. tauschii* lines AL8/78, Clae23 and ENT336 were extracted as described by Oñate-Sánchez and Vicente-Carbajosa (Oñate-Sánchez and Vicente-Carbajosa 2008). RNA quality was assessed using a Qubit fluorometer (Invitrogen) and 2100 Bioanalyzer (Agilent Technology) in the Earlham Institute (Norwich, UK). Illumina TruSeq mRNA libraries were constructed using the Illumina TruSeq RNA Sample preparation guide v2 (Illumina Inc.) in accordance with the manufacturer's protocol. One µg of total RNA used to purified mRNA using two rounds of poly-T oligonucleotide purification attached magnetic beads. During the second elution of poly-A RNA, the RNA was fragmented and primed for cDNA synthesis. cDNA synthesis was carried out using SuperScript II Reverse Transcriptase (Invitrogen) and random primers. Second strand cDNA synthesis was carried out and the DNA was subjected to end repair, "A" tailing and ligation. cDNA templates were enriched by 15 cycles

of PCR as per manufacturer's instructions. The amplified library was quantified using a Bioanalyzer DNA 100 Chip. The library was normalised to 10 nM for generation of sequence clusters on a sequencing flow-cell on the Illumina c-Bot instrument. Sequencing library cluster generation was carried out on a paired-end flow cell on the Illumina cBot according to the manufacturer's instructions. All sequencing was carried out on an Illumina HiSeq 2500, with 100 bp paired-end read metric, TruSeq SBS V3 Sequencing kit and version 1.12.4.2 RTA. FASTQ files were generated and demultiplexed according to library-specific indices by CASAVA (v. 1.8.2, Illumina). Adaptor sequences were trimmed using CutAdapt v 1.6 (Martin 2011) and low quality sequences were removed by Trimmomatic v0.30 (Bolger et al. 2014) with parameter ILLUMINACLIP:2:30:10 HEADCROP:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 (Additional File 2: Table S2).

To estimate gene expression differences between Paragon and *Ae. tauschii*, the HISAT2-StringTie pipeline (Pertea et al. 2016) was used to compute Transcripts Per Million (TPM) values. Transcripts were mapped to the Triticum3.1 wheat assembly (Zimin et al. 2017) with the 2,703 PacBio-based 3DL scaffolds replaced by the 3DL pseudomolecule, and to the *Ae. tauschii* AL8/78 assembly (Luo ref).. HISAT2 (v2.1.0) (Kim et al. 2015) was used to create genome indices and map the trimmed RNAseq reads to each genome. A two-step StringTie (v1.3.3b) (Pertea et al. 2016) strategy was used to measure transcript abundance. For each RNA-Seq sample, StringTie was used to assemble the read alignments; and then a non-redundant set of transcripts observed in all the RNA-Seq samples assembled previously was generated using the stringtie --merge mode; finally the second run was performed with the merged transcript models, for each RNA-Seq sample, using the -B/-b and -e options in order to estimate transcript abundances and generate read coverage tables.

Absolute quantitative RT-PCR was employed to validate the use of TPM values for comparing TPM values between species. Fourteen genes exhibiting balanced expression that had single copy AA and BB homoeologs in the TGACv1 wheat genome were selected for primer design for Q-RT-PCR using HomoeologPrimer (<https://github.com/lufuhao/HomoeologPrimer>). Two pairs of primers for each gene was designed, one pair to amplify all AA+BB+DD transcripts in hexaploid wheat and the DD transcript in diploid *Ae. tauschii*, and other pair to specifically amplify DD transcripts in the hexaploid and the diploid genomes. For each species, 50,000 protoplasts from young leaves were collected in triplicates, total RNA made using the Qiagen

RNeasy Plant Mini Kit (Cat No. 74904). Up to 1 µg total RNA was used to generate cDNAs using Qiagen QuantiTect Reverse Transcription Kit (Cat No. 205311), which yielded 20 µl of cDNAs (33 ng/µl for Paragon and 26 ng/µl for AL8/78). Quantitative PCR was performed on the LightCycler 480 System (Roche). A dilution curve of 8 - 80m copies of the 5,340-bp plasmid, pETnT was used to establish a standard curve of absolute molecule numbers. The 20-µl PCR profile was set as follows: 10 µl of LightCycler® 480 SYBR Green I Master, 1 µl of forward primer, 1 µl reverse primer, 0.3 1 µl cDNA, 7.7 µl of H₂O; The LightCycler 480 instrument was run as follows: 10 min of pre-incubation at 95 °C; 45 cycles amplification: 10s at 95 °C, 10s at 72 °C and 10s at 72 °C with single acquisition; Melting Curve and cooling was set as instructed in the SYBR Green I Master protocol.

Bisulphite Sequencing

Leaf material was harvested from triplicated samples DNA was extracted using the Qiagen DNeasy Plant Mini Kit. Samples from *Ae. tauschii* were processed at the CGR (University of Liverpool, UK) using the Agilent SureSelect capture probe sets for targeted gene enrichment followed by bisulfite treatment using the Zymo Research EZ DNA Methylation-Gold Kit, standard illumina library preparation and sequencing using the Hiseq 4000 (2 x 150 bp reads). Triplicated samples from hexaploid Paragon wheat were pooled and processed to generate a whole genome bisulfite treated sequencing library using the Zymo Research EZ DNA Methylation-Gold Kit and sequencing was carried out on a Hiseq 2500 at Cold Spring Harbor Laboratories, USA (2 x 250 bp reads). Bisulfite-converted Paragon and *Ae. tauschii* paired end sequences were aligned to the full Paragon genome assembly or *Ae. tauschii* 3L assemblies using Bismark (version 0.18.1) (Krueger and Andrews 2011). Duplicate sequencing reads were then filtered using Picard tools. The Bismark methylation extractor tool was then used to identify the methylation status at each cytosine residue across the sequencing reads. A custom Perl script was then used to calculate the % of reads methylated per cytosine residue across the reference sequence. To map Paragon methylation data to the 3DL pseudomolecule, methylation data was aligned to the Paragon genome assembly (https://opendata.earlham.ac.uk/opendata/data/Triticum_aestivum/) using Nucmer (Kurtz et al. 2004) and then the methylation coordinates were transferred to the Chinese Spring 3DL pseudomolecule. The longest contiguous alignments between the sequences with identities >98% and lengths ≥500 bp were identified and the methylation status of 3DL genic DNA was identified in the pseudomolecule. This yielded a space of 13,046,879 bp across 3,541

sequences that were identified as part of the 3DL pseudomolecule. For *Ae. tauschii*, genic sequences in chromosome 3L spanned 19,519,314 bp across 4,130 sequences (18,975,440 bp of unique sequence).

Gene sequences for probe design for targeted gene enrichment included 80,562,496 bp from *Ae. tauschii* chromosome 3L. This included 2000 bp upstream of each gene (from the start codon), the gene body including predicted introns, and 500 bp downstream of the termination codon. Agilent SureSelect Target Enrichment probes contained 120 bp probes tiled across the design space at 40 bp intervals. A total of 227,969 probes covering a potential 27,356,280 bp were submitted for synthesis using the Agilent SureDesign capture design website.

ATAC Sequencing

Sterilised seeds were grown on moist filter paper in petri dishes at 25°C for 11-12 days. Leaf tissue (1 - 2 g) was cut into 2 - 3 cm lengths in a sterile petri dish containing 10 ml 0.6 M Mannitol, 3% Cellulase RS (Duchefa), 1% Macerozyme R10 (Duchefa), 10mM MES, pH5.7, 1mM CaCl₂, 5mM β-Mercaptoethanol, 0.1% BSA, 50 ppm ampicillin. Leaf material was finely chopped to 1 - 2 mm segments and vacuum infiltrated at 20 mm Hg for 20 mins. Protoplasts were released by shaking (50 rpm) at 25 °C for 4-4.5 hour in the dark. The lysate was filtered through a 100 µm Cell Strainer (Falcon) into a 50 ml Falcon tube, and the digested leaf material was washed three times with 6 ml suspension buffer (0.6 M Mannitol, 20mM KCl, 4mM MES, pH5.7 with KOH). The filtrate was re-filtered through a 70 µm Cell Strainer (Falcon) and centrifuged at 70g for 15 min at 12 °C. The protoplast pellet was gently resuspended in 2 ml of suspension buffer and layered onto a Percoll gradient (2.25 ml Percoll (Sigma) and 5.25 ml Suspension buffer) in a 15ml Falcon tube. The gradient was centrifuged at 3000 rpm for 15 min, and the upper phase was discarded and the protoplast layer (cloudy light green phase) was collected into new 15 ml Falcon tube. This was washed in 9 ml suspension buffer, centrifuged at 70 g for 10 min, and the pellet resuspended in 2 ml suspension buffer. Protoplast yields and integrity were tested using Evans Blue (400 mg/l in 0.5M mannitol) and a haemocytometer. This yielded more than 200,000 viable high quality protoplasts.

Nuclei were prepared by pelleting approximately 200,000 protoplasts at 70g for 10 min at 12 °C and resuspending in 6 ml MEB buffer containing 0.1% Triton X-100 (1.0M 2-methyl-2,4-pentanediol (Aldrich), 10 mM PIPES-KOH, 10 mM MgCl₂, 0.1% Triton X-100, 2%

polyvinylpyrrolidone (PVP-10 Sigma), 10mM sodium metabisulfite, 5 mM mercaptoethanol, pH 6.0) and mixed by gentle rotation at 4°C for 5 mins. The crude nuclear prep was centrifuged at 650g for 5 mins at 4°C and resuspended in 3 ml MPDB buffer containing 0.1% Triton X-100 buffer (0.375M 2-methyl-2,4-pentanediol (Aldrich), 7.5 mM PIPES-KOH, 7.5 mM MgCl₂, 0.1% Triton X-100, 7.5 mM sodium metabisulfite, 5 mM mercaptoethanol, pH 7.0). The suspended nuclei were layered onto a 37.5% Percoll gradient (3.75ml Percoll,(Sigma) and 6.25ml MPDB 0.1% Triton X-100 buffer) in a 15ml Falcon tube and centrifuged at 1000g for 10 mins at 4°C. The purified nuclear pellet was resuspended in 7.5 ml MPDB 0.1% Triton X-100 buffer, centrifuged at 650g for 5 mins at 4°C, and resuspended in 1-2 ml MPDB 0.1% Triton X-100 buffer. The yield and integrity of purified nuclei was assessed using a haemocytometer and Methylene blue staining. Typical yields were 25-30% of starting protoplast numbers. Nuclear preparation should be completed within 30 mins of protoplast lysis.

For ATAC reactions on nuclei, reagents were from the Illumina Nextera DNA Library Prep Kit FC-121-1030/15028212 (24 samples). For a single reaction 50,000 intact isolated nuclei were pelleted in a microfuge at 650g 5 mins 4 °C. The nuclear pellet was resuspended in 25 µl 2x Tagment DNA Buffer. 22.5 µl or 20 µl nuclease free water and 2.5 µl or 5.0 µl Tagment DNA Enzyme 1 was added and the reaction transferred to a 0.2ml PCR tube and incubated in thermal cycler block pre-warmed to 37°C for 30 mins, with gentle mixing by hand every 5 mins. Immediately following the transposition reaction, DNA was purified using a Qiagen PCR Purification MinElute kit. Elute DNA with 12 µl Elution Buffer and resuspended in 10 µl. Tagmented DNA can be stored at -20°C at this stage.

Amplification reactions were set up in a PCR tube:

Tagmented DNA	10 µl
Nuclease Free Water	10 µl
25uM Customized Universal P7 Primer	2.5 µl
25uM Customized Barcoded P5 Primer	2.5 µl
NEB High Fidelity 2x PCR Master Mix	25 µl

PCR cycles:

1 Cycle	5 Mins	72°C
	30 Secs	98°C
11 Cycles	10 Secs	98°C

	30 Secs	63°C
	1 Min	72°C
Hold		4°C

Amplified DNA was purified using a Qiagen PCR Purification MinElute kit. DNA was eluted with 22 µl Elution Buffer. An additional clean up step to removed excess primers used 20 µl Ampure XP Beads (Beckman Coulter A63880) and purified DNA was recovered in 25 µl 0.1 x TE. DNA was quantified using a Qubit Fluorometer and a HS DNA kit. The size distributions of tagmented amplified DNA was assessed using MultiNA MCE-202 Bioanalyser (Shimadzu) with a DNA 1000 Kit, or TapeStation Screen tape Kit D1000, or the Agilent Bioanalyser HSDNA chip.

To exclude ATAC-seq reads from mitochondrial and chloroplast genomes, which arise from contamination of plant nuclei preparations, we included the *Triticum aestivum* chloroplast genome (GenBank accession No. NC_002762), *Triticum aestivum* mitochondrial genome (GenBank accession No. AP008982) and *Ae. tauschii* chloroplast genome (GenBank accession No. NC_022133) in the reference genome. As the *Ae. tauschii* mitochondrial genome has not yet been sequenced that from wheat was used. Trim_Galore (version 0.5.0; <https://github.com/FelixKrueger/TrimGalore>) was used to remove Nextera adaptor sequences and Trimmomatic (Bolger et al. 2014) were used to filter out those short (<70 bp) reads. The resulting clean reads were aligned to the reference sequences using Bowtie (v1.2.2; Options: -X 2000 --fr -m 1) (Langmead et al. 2009). After filtering reads matching mitochondrial and chloroplast genomes (which also removed reads matching chloroplast and mitochondrial insertions in the nuclear genome), we included paired reads with high mapping quality (MAPQ score >10, qualified reads) through SAMtools (Li et al. 2009) for further analysis. Duplicate reads was removed using Picard tools MarkDuplicates (<http://broadinstitute.github.io/picard/>). All reads aligning to the forward strand were offset by +4bp , and all reads aligning to the reverse complement strand were offset -5 bp (Adey et al. 2010). ATAC-Seq peak regions of each sample were called using MACS2 (v2.1.2_dev) (Zhang et al. 2008) with parameters --nomodel --shift -37 --extsize 73. To generate a consensus set of unique peaks, we next merged ATAC-Seq peaks for which the distance between proximal ends was less than 10 base pairs.

References

- Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11**: R119.
- Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**: 11.
- Barton MD, Barton HA. 2012. Scaffolder - software for manual genome scaffolding. *Source Code Biol Med* **7**: 4.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.
- Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res* **44**: e147.
- 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**: 1249721.
- Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**: 3420–3435.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.
- Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**: 965–978.
- Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVIDENCEModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**: R7.
- Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. 2014. PASTEC: an automatic transposable element classification tool. *PLoS One* **9**: e91929.
- International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq principal investigators:, Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, IWGSC whole-genome assembly principal investigators:, Pozniak CJ, et al. 2018. Shifting the limits

- in wheat research and breeding using a fully annotated reference genome. *Science* **361**.
<http://dx.doi.org/10.1126/science.aar7191>.
- Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360.
- Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571–1572.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. 10.1186/gb-2004-5-2-r12. *Genome Biol* **5**: R12.
<http://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-2-r12>.
- Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Lu F-H, McKenzie N, Kettleborough G, Heavens D, Clark MD, Bevan MW. 2018. Independent assessment and improvement of wheat genome sequence assemblies using Fosill jumping libraries. *Gigascience* **7**. <http://dx.doi.org/10.1093/gigascience/giy053>.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j* **17**: 10.
- Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K. 2009. TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiol* **150**: 1135–1146.
- Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M. 2013. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Res* **41**: D1144–51.
- Oñate-Sánchez L, Vicente-Carbajosa J. 2008. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. *BMC Res Notes* **1**: 93.
- Otto TD, Dillon GP, Degraeve WS, Berriman M. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res* **39**: e57.
- Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**: 1650–1667.
- Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**: i351–8.
- Slater G, Birney E. 2005. 10.1186/1471-2105-6-31. *BMC Bioinformatics* **6**: 31.
<http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-31>.
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio

prediction of alternative transcripts. *Nucleic Acids Res* **34**: W435–9.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–192.

Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L, et al. 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res* **34**: W293–7.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL. 2017. The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *Gigascience* **6**: 1–7.