Genetics: Early Online, published on August 12, 2020 as 10.1534/genetics.120.303501

Chromosome-scale assembly of the bread wheat genome reveals thousands of additional gene copies

Michael Alonge[*,1,2], Alaina Shumate[†,‡,1], Daniela Puiu[†,‡], Aleksey Zimin[†,‡], Steven L. Salzberg[*,†,‡,§,2]

[*]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218 *

[†]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218

[‡]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21211

[§]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205

[1]These authors contributed equally.

[2]Correspondence: malonge11@gmail.com, salzberg@jhu.edu

Data Reference Numbers:

NCBI bioproject: PRJNA392179

Genebank assembly accession: GCA_002220415.3

Gene annotations: https://github.com/TriticumAestivum/Annotation

RUNNING TITLE

Wheat assembly reveals many new genes

KEYWORDS

Genome assembly, gene annotation, scaffolding, wheat, Triticum, gene duplication

CORRESPONDENCE

Michael Alonge

160 Malone Hall

3400 North Charles Street

Baltimore, MD 21218

(831) 201-2788

malonge11@gmail.com

Steven L. Salzberg

Center for Computational Biology

Johns Hopkins University

Wyman Park Bldg, S261,

3100 Wyman Park Drive

Baltimore, MD 21211

(410) 516-8246

salzberg@jhu.edu

ABSTRACT

Bread wheat (*Triticum aestivum)* is a major food crop and an important plant system for agricultural genetics research. However, due to the complexity and size of its allohexaploid genome, genomic resources are limited compared to other major crops. The IWGSC recently published a reference genome and associated annotation (IWGSC CS v1.0, Chinese Spring) that has been widely adopted and utilized by the wheat community. Although this reference assembly represents all three wheat subgenomes at chromosome-scale, it was derived from short reads, and thus is missing a substantial portion of the expected 16 Gbp of genomic sequence. We earlier published an independent wheat assembly (Triticum_aestivum_3.1, Chinese Spring) that came much closer in length to the expected genome size, although it was only a contig-level assembly lacking gene annotations. Here, we describe a reference-guided effort to scaffold those contigs into chromosome-length pseudomolecules, add in any missing sequence that was unique to the IWGSC CS v1.0 assembly, and annotate the resulting pseudomolecules with genes. Our updated assembly, Triticum_aestivum_4.0, contains 15.07 Gbp of non-gap sequence anchored to chromosomes, which is 1.2 Gbps more than the previous reference assembly. It includes 108,639 genes unambiguously localized to chromosomes, including over 2,000 genes that were previously unplaced. We also discovered more than 5,700 additional gene copies, facilitating the accurate annotation of functional gene duplications including at the *Ppd-B1* photoperiod response locus.

INTRODUCTION

Bread wheat (*Triticum aestivum*) is a crop of significant worldwide nutritional, cultural and economic importance. As with most other major crops, there is a strong interest in applying advanced breeding and genomics technologies towards crop improvement. Key to these efforts are high-quality reference genome assemblies and associated gene annotations which are the foundations of genomics research. However, the bread wheat genome has some notable features that make it especially technically challenging to assemble. One such feature is allohexaploidy (2n=6x=42, AABBDD), a result of wheat's dynamic domestication history (Petersen *et al.* 2006; Dubcovsky and Dvorak 2007). This polyploidy results from the hybridization of domesticated emmer (*Triticum turgidum*, AABB) with *Aegilops tauschii* (DD). Domesticated emmer, also an ancestor of durum wheat, is itself an allotetraploid resulting from interspecific hybridization between *Triticum urartu* and a relative of *Aegilops speltoides*.

The resulting bread wheat genome is immense, with flow cytometry studies estimating the genome size to be ~16 Gbp (Arumuganathan and Earle 1991). As with most other large plant genomes, repeats, including mostly retrotransposons, make up the majority of the genome, which is estimated to be ~85% repetitive (Appels *et al.* 2018). These repeats especially make this genome difficult to assemble, even given the recent improvements in long-read sequencing and algorithmic advancements in genome assembly technology. Nonetheless, early efforts were made to establish *de novo* reference genome assemblies for wheat. In 2014, the International Wheat Genome Sequencing Consortium (IWGSC) used flow cytometry-based sorting to sequence and assemble individual chromosome arms, thus removing the repetitiveness introduced by homeologous chromosomes (Mayer *et al.* 2014). In spite of this approach, this short-read based

assembly was highly fragmented, and only reconstructed ~10.2 Gbp of the genome. Subsequent short-read assemblies using alternate strategies were also developed by the community, though each also struggled to achieve contiguity and completeness (Chapman *et al.* 2015; Clavijo *et al.* 2017).

In 2017, we released the first-ever long-read-based assembly for bread wheat (Triticum_aestivum_3.1), representing the Chinese Spring variety (Zimin *et al.* 2017). With an N50 contig size of 232.7 kbp, Triticum_aestivum_3.1 was far more contiguous than any previous assembly of bread wheat, and with a total assembly size of 15.34 Gbp, it reconstructed the highest percentage of the expected wheat genome size of any assembly. Though this assembly provided a more complete representation of the Chinese Spring genome, its contigs were not mapped onto chromosomes, and notably, it did not include gene annotation.

In 2018, the IWGSC published a chromosome-scale reference assembly and associated annotations for bread wheat (IWGSC CS v1.0, Chinese Spring), providing the best-annotated reference genome yet(Appels *et al.* 2018). Because that assembly was entirely derived from short reads, it was less complete and more fragmented than Triticum_aestivum_3.1, having a total size of 14.5 Gbp and an N50 contig size of 51.8 kpb. However, a collection of long-range scaffolding data, including physical (BACs, Hi-C), optical (Bionano), and genetic maps, enabled most of the assembled scaffolds to be mapped onto wheat's 21 chromosomes. These pseudomolecules served as a foundation for comprehensive *de novo* gene and repeat annotation, facilitating investigations into the genomic elements that drove the evolution of genome size, structure, and function in wheat.

Here, we used the IWGSC CS v1.0 assembly (Genbank accession GCA_900519105.1) to inform the scaffolding and annotation of the more complete Triticum_aestivum_3.1 assembly. The new assembly, Triticum_aestivum_4.0, contains 1.1 Gbp of additional non-gapped sequence compared to IWGSC CS v1.0, while localizing 97.9% of sequence to chromosomes. Comparative analysis revealed that Triticum_aestivum_4.0 more accurately represents the Chinese Spring repeat landscape, which is heavily collapsed in IWGSC CS v1.0. Our more-complete assembly allowed us to anchor ~2,000 genes that were previously annotated on unlocalized contigs in IWGSC CS v1.0. We also found 5,799 additional gene copies in Triticum_aestivum_4.0, showing extensive collapsing of gene duplicates in IWGSC CS v1.0 assembly. We highlighted specific examples of these extra gene copies, including at the *Ppd-B1* locus, where Triticum_aestivum_4.0 accurately reflects the expected four copies of Pseudo-Response Regulator (PRR) genes influencing photoperiod sensitivity. We additionally found three extra copies of a MADS-box transcription factor gene in T4, demonstrating the potential to find new gene copy number variants (CNVs) that influence traits. The Triticum_aestivum_4.0 assembly and annotations are available at www.ncbi.nlm.nih.gov/bioproject/PRJNA392179.

MATERIALS AND METHODS

Establishing the initial contig set

We first sought to establish the most complete set of contigs representing the genome of *T. aestivum* Chinese Spring. We started with the Triticum_aestivum_3.1 contigs (T3) (Zimin *et al.* 2017) because they comprise 1 Gbp of additional non-gap sequence compared to the IWGSC CS v1.0 (IW) reference assembly. However, when establishing a set of contigs for downstream

scaffolding, we wanted to ensure that we incorporated any contigs unique to the reference assembly and therefore "missing" from the T3 assembly. To do this, we broke the reference assembly into "contigs" by breaking pseudomolecules at gaps (at least 20 "N" characters). We then aligned these reference contigs (query) to the T3 contigs (reference) using NUCmer (-l 250 -c 500), and filtered them using delta-filter (-1 -l 5000) to include only reciprocal best alignments at least 5 kbp long (Kurtz *et al.* 2004). Of the reference contigs that were at least 10 kbp in length, if under 25% of a contig was covered by alignments, it was deemed a putative "missing" contig.

We then checked to see if these putative missing contigs would indeed be covered by alignments produced with more sensitive parameters. The putative missing contigs (query) were aligned again to the T3 assembly with NUCmer, but with a smaller minimum seed and cluster size (-l 50 -c 200). Alignments were filtered as before, and if under 25% of a putative missing contig was covered by these more sensitive alignments, they were deemed to be validated as missing from T3. These validated missing IW contigs were combined with the T3 contigs to establish our final set of contigs for downstream scaffolding, which had an N50 length of 230,687 bp and a sum of 15,429,603,425 bp.

RaGOO scaffolding

We performed two rounds of reference-guided scaffolding with RaGOO. We first used RaGOO to look for false sequence duplications, especially those that could have arisen by incorporating "missing" IW contigs. Though RaGOO usually employs Minimap2 (Li 2018) to align query contigs to a reference genome, we used NUCmer in order to produce high specificity alignments.

We aligned our contigs (query) to the IW reference genome (reference) using a very large seed and cluster size (-l 500 -c 1000). Such specificity in alignments was necessary to unambiguously order and orient contigs with respect to the highly repetitive allohexaploid reference genome. The resulting delta file was converted to PAF format using Minimap2's paftools. Next, we ran RaGOO using these alignments rather than the default Minimap2 alignments while also specifying a minimum clustering confidence score of 0.4 (-i). We also excluded any unanchored IW sequence from consideration (-e).

To remove false duplication of missing contig sequence, we observed that such duplications would align to more than one place in these RaGOO pseudomolecules. Conversely, contigs that were truly "missing" should only align once (perfectly) to their ordered and oriented location in the RaGOO scaffolds. We aligned the RaGOO scaffolds (query) to the missing IW contigs (reference) with NUCmer (-l 50 -c 200) and filtered alignments with delta-filter (-q -l 5000) (Marçais *et al.* 2018). If a missing contig had more than one alignment with coverage at least 50% and percent identity at least 98%, it was deemed to be a false duplicate and removed from the initial contig set. With false duplicates removed, we proceeded with the second round of RaGOO scaffolding which had all of the same specifications as the first round.

We next sought to remove any unanchored contigs that had duplicated sequences amongst the anchored contigs. The same previously described process to remove false duplicates was also used here, except that the RaGOO scaffolds along with unanchored contigs (query) were aligned to the unanchored contigs (reference). Also, the minimum coverage was 75% rather than 50%. After removing these unanchored duplications, scaffolds were polished with POLCA (included

177 in MaSuRCA 3.3.5) (Zimin and Salzberg 2019). For polishing, we used the Illumina reads from
178 the NCBI SRA accession SRX2994097. POLCA introduced 595,705 bp in substitution
179 corrections and 1,033,593 bp in insertion/deletion corrections. After polishing, the final error rate
180 of the sequence was estimated at less than 0.008% or less than 1 error per 10,000 bases. Finally,
181 we removed any redundant mitochondria and chloroplast sequences from unplaced contigs, thus
182 resulting in the final Triticum_aestivum_4.0 (T4) assembly. T4/IW dotplots were made by
183 aligning the polished T4 assembly (query) to the IW reference assembly (reference) with
184 NUCmer (-l 500 -c 1000). Alignments less than 10 kbp were removed with delta-filter and were
185 plotted with mummerplot (--fat --layout).
186
187 **Shared k-mer frequency distribution**
188 101-mers were counted in T4 and IW using KMC (v3.1.0, -ci1 -cx10000 -cs10000) (Kokot *et al.*
189 2017). 101-mers shared by T4 and IW were then extracted with kmc_tools "simple" using the
190 intersection function. The 101-mer copy frequency distribution of these shared *k*-mers in both T4
191 and IW (-ocleft and -ocright) was then plotted in **Figure 3**.
192
193 **Centromere annotation**
194 We annotated centromere sequence in T4 using an approach similar to the original IW
195 publication (Appels *et al.* 2018). First, publicly available Chinese Spring CENH3 ChIP-seq data
196 (SRR1686799) was downloaded from the European Nucleotide Archive (Guo *et al.* 2016). Reads
197 were then trimmed with cutadapt (v1.18, -a AGATCGGAAGAG) and aligned to T4 with bwa
198 mem (v0.7.17-r1198-dirty) (Li and Durbin 2009; Martin 2011). Alignments with a mapq score
199 less than 20 were removed and the remaining alignments were compressed and sorted with

samtools view and samtools sort respectively (Li *et al.* 2009). Alignments were then counted in 100 kbp non-overlapping windows along the T4 genome using bedtools makewindows and bedtools coverage (v2.29.2) (Quinlan and Hall 2010). Any group of two or more consecutive windows with greater than or equal to three times the genomic average coverage was considered putative centromere sequence, and any such intervals within 500 kbp were merged together. These intervals were further merged or removed by manually comparing them with the CENH3 ChIP-seq alignments, resulting in a single inferred centromere annotation for each chromosome (**Table S1**). Some IW chromosomes have more than one centromeric position reported in the original IW publication. Accordingly, we picked the longest centromeric interval for each IW chromosome for the comparative analysis presented in this work.

Chloroplast and mitochondria genome assembly

We took the first 20 million Illumina read pairs from the SRR5815659 accession and assembled them with megahit (v1.2.8) (Li *et al.* 2015). The resulting assembly contained 145,887 contigs (74.41 Mb) with lengths ranging between 200 bp and 56,565 bp. Then we aligned these contigs to the *Triticum aestivum* reference chloroplast sequence (NC_002762.1) using NUCmer (with --maxmatch switch to align to repeats) and filtered the alignments with delta-filter, keeping the best hits to the reference NC_002762.1. The reference was covered completely by alignments of only five contigs. Then, we aligned these contigs to each other with NUCmer (--maxmatch --nosimplify) and used the alignments to manually order and orient them into a single chloroplast sequence scaffold.

To assemble the mitochondrial genome, we aligned the megahit contigs discussed above to the *Triticum aestivum* mitochondria reference sequence (MH051716) with NUCmer (--maxmatch). We then filtered the alignments with delta-filter, keeping the best matches to the MH051716 reference. This revealed 43 non-chloroplast contigs of least 500 bp in length that matched best to the mitochondria reference. We then ordered and oriented these 43 contigs using RaGOO (v1.1), setting the minimum alignment length to 500 bp. The chloroplast and mitochondria sequence are included in our data submission to NCBI.

Genome annotation

We used Liftoff to annotate the T4 genome using the IW v1.1 gene models (Shumate and Salzberg 2020). Genes were aligned to their same chromosome in T4 using BLASTN v.2.9.0 (-soft_masking False -dust no -word_size 50 -gap_open 3 -gapextend 1 -culling_limit 10). The blast hits were filtered to include only those that contained one or more exons. For each gene, the optimal exon alignments were chosen according to sequence identity and concordance with the exon/intron structure of the gene model in IW. These alignments were used to define the boundaries of each exon, transcript, and gene in T4. We excluded any transcripts that did not map with at least 50% alignment coverage. Any genes without at least one mapped isoform were then aligned against the entire T4 genome using BLASTN with the same parameters and placed given they did not overlap an already placed gene.

To place the chrUn genes, we aligned the genes to the entire T4 genome using the same parameters. We excluded any transcripts that did not meet the 50% alignment coverage threshold or overlapped an already annotated gene.

To find additional gene copies we aligned all genes (query) to the complete T4 genome (reference) using BLASTN v2.9.0 (-soft_masking False -dust no -word_size 50 -gap_open 3 -gapextend 1 -culling_limit 100, qcov_hsp_perc 100). The notable differences in these parameters are qcov_hsp_perc which requires 100% query coverage, and culling_limit which has been increased from 10 to 100 to increase the number of reported alignments for genes with a highly increased copy number. We excluded any alignments that did not have 100% exonic sequence identity or overlapped a previously placed gene. We used gffread to filter out genes with non-canonical splice sites(Pertea and Pertea 2020).

Finally, using the same methods as described for high confidence genes above, we also used Liftoff to map the IW v1.1 low confidence annotation onto T4. We successfully mapped 152,900 out of 161,537 low confidence genes. Another 1,581 genes mapped partially below the 50% alignment coverage threshold.

Ppd-B1 haplotype comparison

To find the approximate location of the *Ppd-B1* locus in the T4 and IW assemblies, we aligned a *Ppd-B1* PRR gene sequence (GenBank Accession DQ885757.1) to T4 and IW with blastn v2.6.0 (-perc_identity 95) (Beales *et al.* 2007). No matches were found on IW chr2B, though partial matches were found on chrUn. In contrast, 4 strong matches were found on T4 chr2B, corresponding to genes *T4021472*, *T4021473*, *T4021474*, and *T4021475*. We also aligned the entire Chinese Spring haplotype for this locus, which had been previously cloned and sequenced (GenBank Accession JF946485.1), to T4 using blastn v2.6.0 (-perc_identity 95) (Díaz *et al.*

2012). We used these alignments to approximately define the genomic coordinates of *Ppd-B1* in T4. In order to further validate the accuracy of this locus in T4, we aligned the GenBank JF946485.1 sequence to the T4 locus +/- 10 kbp flanking sequence in order to find pairwise maximal exact matches (MEMs) at least 50 bp in length. These alignments are depicted in **Figure 4C** and were generated with mummer v3.23 (-maxmatch -l 50 -b -c). Prior to alignment, the GenBank JF946485.1 sequence was reverse complemented in order to refer to the same strand as our T4 chr2B.

Because the PRR gene annotations used to define T4 *Ppd-B1* PRR genes were incomplete in IW, they were also initially incomplete in T4. To correctly annotate these T4 PRR genes, we used Liftoff to lift-over the GenBank JF946485.1 PRR gene annotations to T4. These genes are labeled *T4021472* , *T4021473*, *T4021474*, and *T4021475* in the final annotation.

Data Availability

The Triticum_aestivum_4.0 assembly is available at www.ncbi.nlm.nih.gov/bioproject/PRJNA392179 (GenBank accession: GCA_002220415.3). The annotation is available at https://github.com/TriticumAestivum/Annotation and ftp://ftp.ccb.jhu.edu/pub/data/Triticum_aestivum/Triticum_aestivum_4.0. All results described are in reference to annotation version v1.0. The Triticum_aestivum_4.0 inferred centromere positions are provided in Table S1. Table S2 lists the IWGSC CS v1.0 chrUn annotations that we localized, while Table S3 lists the IWGSC CS v1.0 annotations of which we found extra copies. Table S4 provides a mapping from our custom annotation IDs to IWGSC CS v1.0 annotation IDs.

RESULTS

Scaffolding the Triticum_aestivum_3.1 genome assembly

Our goal was to utilize both our previously published Triticum_aestivum_3.1 contigs (T3) and the IWGSC CS v1.0 reference assembly (IW) to establish an improved chromosome-scale genome assembly for the Chinese Spring variety of bread wheat. **Figure 1** depicts the pipeline used to derive our final Triticum_aestivum_4.0 (T4) assembly. We started with the T3 contigs because they were highly contiguous (N50 = 232.7 kbp) and contained a total of 1.1 Gbp more non-gap sequence compared to the IW assembly. However, we wanted to ensure that our final assembly did not exclude any contigs missing from T3 but present in IW. To incorporate any such "missing" IW contigs, we first derived a set of contigs from the IW assembly by breaking pseudomolecules at gaps. By aligning these IW contigs to the T3 assembly, we identified 4,702 IW contigs (89,866,936 bp) with sequence missing from the T3 assembly. These sequences along with the T3 contigs comprised our initial contig set.

We used RaGOO (Alonge *et al.* 2019), a reference-guided scaffolding tool, to order and orient these contigs into chromosome-length scaffolds. This scenario presents a near-ideal context for reference-guided scaffolding because the contigs and the reference assembly represent the same inbred genotype, and thus we expect no genomic structural differences. Although RaGOO normally utilizes Minimap2 (Li 2018) alignments between contigs and a reference assembly, we used NUCmer (Kurtz *et al.* 2004; Marçais *et al.* 2018) instead, as it offered the necessary flexibility to align these large and repetitive genomes. Specifically, NUCmer provided the specificity needed to unambiguously align contigs to a highly repetitive allohexaploid reference

genome (see **Methods**). Even with high stringency alignments, RaGOO ordered and oriented most of the assembly (97.67% of bp) into pseudomolecules.

We next sought to remove any false duplications potentially created during the process of incorporating 4,702 IW sequences. We aligned these IW contigs to the RaGOO scaffolds and removed 357 IW contigs from the initial set of 4,702 that aligned to more than one place in the assembly and therefore were no longer deemed "missing" from T3. This produced our final set of contigs, which included the T3 contigs plus 4,345 (84,909,842 bp) contigs from IW that contained sequence missing from T3. The final contigs had an N50 length of 230,687 bp (essentially the same as the T3 assembly) and a sum of 15,429,603,425 bp. We then repeated the RaGOO scaffolding step, and polished the resulting scaffolds with POLCA (Zimin and Salzberg 2019) using the original Illumina reads, yielding the final T4 chromosome-scale assembly. Finally, we removed mitochondria and chloroplast genome sequence from T4 and assembled these genomes separately with Illumina reads (see **Methods**).

Despite the highly repetitive nature of the Chinese Spring genome, RaGOO confidence scores indicate that T4 scaffolding was consistent with the reference genome structure (**Figure S1**). This suggests that our high-specificity NUCmer parameters mitigated erroneous contig ordering and orientation resulting from repetitive alignments. Dotplots further confirm that there are no large-scale structural rearrangements between T4 and IW pseudomolecules (**Figure S2**). While borrowing its chromosomal structure from IW, T4 demonstrates superior sequence completeness. 97.9% of T4 sequence (15.09 Gbp) was placed onto 21 chromosomes yielding pseudomolecules that had 1.2 Gbp more localized non-gapped sequence than the IW reference (**Table 1**). This

extra sequence was evenly distributed across the genome, with each T4 pseudomolecule containing more sequence (average of 48.8 +/- 8.4 Mbp) than its IW counterpart while having substantially fewer gaps (**Figure 2**).

Because IW was derived from short-reads, it is conceivable that some genomic repeats were collapsed during assembly (Schatz *et al.* 2010). Therefore, we hypothesized that T4, a long-read-based assembly, more accurately represents the repeat landscape of the Chinese Spring genome. As support for this hypothesis, we observe that 101-mers shared by T4 and IW were present at higher copies in T4 (**Figure 3**). This observation holds for a wide range of 101-mer copy numbers, suggesting that T4 more accurately represents both lower-order (duplications) and higher-order (transposable elements) repeats. To investigate a specific instance of repeat collapse in IW, we compared centromere sequence content in the two assemblies. As was done in the original IW publication, we used publicly available CENH3 ChIP-seq data to infer centromere positions in T4 (see **Methods**) (**Table S1**)(Guo *et al.* 2016; Appels *et al.* 2018). This analysis indicated ChIP-seq peaks corresponding to centromeres for each of the 21 chromosomes (**Figure S3**). T4 had a total of 39.1 Mbp more centromeric sequence than IW, highlighting that the long-read-based T4 assembly localized more centromeric sequence than IW.

Annotating the Triticum_aestivum_4.0 genome assembly

We mapped the IW v1.1 high-confidence annotation onto T4 using an annotation lift-over tool we developed called Liftoff (see **Methods**) (Shumate and Salzberg 2020). Given a genome annotation, Liftoff aligns all genes, chromosome by chromosome, to a different genome of the same species using BLAST (Altschul *et al.* 1990). For all genes that fail to map to the same

chromosome, Liftoff attempts to map them across chromosomes. The best mapping for each gene is chosen according to sequence identity and concordance with the exon/intron structure of the original gene model. Out of 130,745 transcripts from 105,200 gene loci annotated on primary chromosomes in IW, we successfully mapped 124,579 transcripts from 100,831 gene loci. We define a transcript as successfully mapped if the mRNA sequence in T4 is at least 50% as long as the mRNA sequence in IW. However, the vast majority of transcripts greatly exceed this threshold, with 92% of transcripts having an alignment coverage of 98% or greater (**Figure S4A**). Sequence identity is similarly high with 92% of transcripts aligning at an identity of 95% or greater (**Figure S4B**). Of the transcripts that failed to map, 4,634 had a partial mapping with an alignment coverage < 50%, and the remaining transcripts failed to map entirely.

As expected, we observed strong gene synteny between T4 and IW (**Figure S5**). Of the 100,831 mapped IW genes, 96,148 mapped to the same chromosome in T4. The remaining 4,683 mapped to a different chromosome after failing to map to their expected chromosome. There is a clear pattern showing many of these genes mapped to a similar location on the same chromosome of a different subgenome. We also found that the sequence identity of genes mapped to different chromosomes is much lower with an average identity of 90.7% compared to 99.3% in genes mapped to the same chromosome. We therefore hypothesize that these genes are missing in the T4 assembly, and have instead mapped to paralogs in T4 that are not annotated in IW.

The IW v1.1 annotation also contains 2,691 genes annotated on unplaced contigs ("chrUn"). Using Liftoff, we were able to map 2,001 of these genes onto a primary chromosome in T4. 1,767 genes were confidently placed with a sequence identity of at least 98% while the

remaining 234 mapped with a lower identity (**Table S2**). To control for differences in annotation pipelines between IW and T4, we used Liftoff to map chrUn genes onto the primary IW chromosomes to look for additional, unannotated gene copies. Of the 2,001 chrUn genes mapped to T4 pseudomolecules, 78 of these were also mapped to primary IW chromosomes. This suggests that at least 1,923 genes were placed due to improved assembly completeness rather than differences in annotation methods.

After mapping the IW v1.1 annotation onto T4, we used Liftoff to look for additional gene copies in T4. We required 100% sequence identity in exons and splice sites to map a gene copy. We found 5,799 additional gene copies in T4 that are not annotated in IW v1.1. Of these, 4,158 genes have one extra copy and 567 genes have two or more additional copies, with a maximum of 84 additional copies (**Figure 4A**). IW collapsed most gene copies on the same chromosome rather than across homeologous chromosomes, with 4,062 of the 5,799 additional gene copies occurring on the same chromosome and 97 copies occurring on the same chromosome of a different subgenome (**Figure 4B**). 915 gene copies were placed on different chromosomes. The remaining 725 are extra copies of chrUn genes placed on chromosomes. The location and functional annotation of all additional copies is provided in **Table S3**. As was done for unplaced genes, we also looked for additional IW gene copies present elsewhere in IW. Of our 5,799 additional gene copies, 159 were also present in IW, suggesting that at least 5,640 of T4 copies are strictly the result of improved assembly completeness.

Triticum_aestivum_4.0 accurately represents gene duplications affecting traits

We searched T4 for specific examples of functionally relevant gene duplications previously collapsed or missing in IW. We focused on the *Ppd-B1* locus on chr2B because copy number variation of Pseudo-Response Regulator (PRR) genes at this locus underlies variation in photoperiod sensitivity among hexaploid wheat varieties (Beales *et al.* 2007). Others have shown that the Chinese Spring variety has four PRR genes at the *Ppd-B1* locus, with one of the copies being truncated (Díaz *et al.* 2012). Because the entire ~200 kbp Chinese Spring *Ppd-B1* locus was previously cloned and sequenced, we were able to assess if this region had been accurately assembled in both T4 and IW. IW lacks any PRR genes at the *Ppd-B1* locus, with fragments of three of the four expected paralogs (*TraesCSU02G196100*, *TraesCSU02G221500*, *TraesCSU02G199500*) residing on unplaced chrUn sequence. In contrast, T4 localizes four PRR genes (*T4021472* , *T4021473*, *T4021474*, and *T4021475*) at *Ppd-B1*, matching the expected Chinese Spring copy number state. Alignment of this T4 locus to the known Chinese Spring *Ppd-B1* sequence indicated that the entire locus had been accurately assembled, even correctly representing the 3 highly-similar intact PRR genes. (**Figure 4C**). The successful assembly of *Ppd-B1* served as a validation that T4 accurately resolves duplications with high sequence similarity.

The successful resolution of the *Ppd-B1* locus suggested that new functionally relevant CNVs may be discovered among the large number of localized or duplicated genes in T4. One notable example was a MADS-box transcription factor gene, *TraesCS6A02G022700*, which had three additional tandem copies (T4 genes *T4081597*, *T4081598*, *T4081599*, and *T4081600*) on T4 chr6A (**Figure 4D**). MADS-box transcription factors are known to influence traits such as flowering time and floral organ development (Coen and Meyerowitz 1991; Ng and Yanofsky

2001). Furthermore, MADS-box gene duplications can quantitatively impact gene expression and domestication phenotypes in a dosage dependent manner (Soyk *et al.* 2019). To provide further evidence that this gene is part of a collapsed repeat in IW, we aligned Chinese Spring Illumina reads to IW and calculated the coverage across the gene +/- 50 kbp of flanking sequence. We observed a spike in coverage indicating a collapsed repeat in IW containing *TraesCS6A02G022700* (**Figure 4E**). We further note that this region contains 10,205 bp of gap sequence suggesting that this locus had been misassembled in IW. This duplication of a MADS-box transcription factor gene as well as our analysis of the *Ppd-B1* locus highlights how T4, with its superior genome completeness, resolves functionally relevant genic sequence previously misassembled, missing, or unlocalized in IW.

DISCUSSION

In one critical aspect, the bread wheat genome exemplifies the challenge of eukaryotic genome assembly. Repeats, which remain difficult to assemble, are pervasive in this transposon-rich allohexaploid plant genome. Therefore, the accurate and complete resolution of the bread wheat genome and the subsequent study of genomic structure especially depends on high-quality data and advanced genome assembly techniques. In 2017, we published the first near-complete and highly contiguous representation of the bread wheat genome (Triticum_aestivum_3.1), demonstrating the value of long reads for wheat genome assembly (Zimin *et al.* 2017). In our efforts described here, we used Triticum_aestivum_3.1 as our foundation while leveraging the strengths of the IWGSC CS v1.0 reference genome to establish the most complete chromosome-scale and gene-annotated reference assembly yet created for bread wheat. By scaffolding and annotating our contigs, we created the genomic context needed to quantify and qualify the completeness of the Triticum_aestivum_4.0 assembly, especially relative to its predecessors.

Compared to the IWGSC CS v1.0 assembly, Triticum_aestivum_4.0 resolves more repeat sequence, exemplified by the improved centromere localization and by the many additional gene copies. The discovery of these extra gene copies, as well as the localization of 2,001 previously unplaced genes, also demonstrates how Triticum_aestivum_4.0 provides an enhanced representation of Chinese Spring genic sequence.

Gene copy number variants (CNVs) are pervasive in hexaploid wheat and are associated with traits such as frost tolerance (*Fr-A2*), vernalization requirement (*Vrn-A1*), and photoperiod sensitivity (*Ppd-B1*) (Díaz *et al.* 2012; Würschum *et al.* 2015, 2017, 2018). These and other CNVs contributed to the adaptive success of domesticated wheat, which now thrives in diverse conditions and geographies. This is exemplified by the *Ppd-B1* locus, where variation of Pseudo-Response Regulator (PRR) gene copy number influences photoperiod sensitivity. Our successful assembly of the *Ppd-B1* locus, which was unanchored and incomplete in IWGSC CS v1.0, highlights a specific example where our improved assembly accurately reflected a known CNV genotype in Chinese Spring. This validation suggests that other functional gene duplications may also be directly encoded in the Triticum_aestivum_4.0 assembly and identifiable by our annotation of extra gene copies. We indicated one such potential candidate, the MADS-box transcription factor gene, which appears with three extra copies in Triticum_aestivum_4.0. We expect that further investigation of the extensive gene duplications presented in this work will provide additional insights into the role of CNVs in wheat phenotypes.

Structural variants (SVs), including CNVs, comprise a vast source of natural genetic variation influencing traits. As sequencing technologies continue to advance, plant scientists are

increasingly using pan-genome analyses to study genome structure among diverse varieties and ecotypes (Alonge *et al.* 2020; Song *et al.* 2020; Liu *et al.* 2020). These studies especially rely on structurally accurate reference genomes to discover SVs. Our work introduces Triticum_aestivum_4.0 as an improved reference genome resource ideal for future structural variant analyses in wheat. Furthermore, our comparative genomics analysis showed that a substantial portion of the Chinese Spring genome was collapsed, missing, or misrepresented when assembled with short reads. This emphasizes the utility of long reads in future wheat pan-genome analyses, where structural accuracy is key. Generally, our work provides a preview of the computational genomics analyses that are possible with an accurate wheat reference genome.

ACKNOWLEDGMENTS

This work was supported in part by NIH under grants R01-HG006677 and R35-GM130151, and by the USDA National Institute of Food and Agriculture under grant 2018-67015-28199.

AUTHOR CONTRIBUTIONS

S.S. designed the project. M.A., A.S., D.P., A.Z., and S.S. designed analysis. M.A., A.S., D.P., A.Z., and S.S. analyzed data. M.A., A.S., and S.S. wrote the manuscript. All authors read and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interest.

REFERENCES

Alonge M., S. Soyk, S. Ramakrishnan, X. Wang, S. Goodwin, *et al.*, 2019 RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. Genome Biol. 20. https://doi.org/10.1186/s13059-019-1829-6

Alonge M., X. Wang, M. Benoit, E. Van Der Knaap, M. C. Schatz, *et al.*, 2020 Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. Cell 182: 1–17. https://doi.org/10.1016/j.cell.2020.05.021

Altschul S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. https://doi.org/10.1016/S0022-2836(05)80360-2

Appels R., K. Eversole, C. Feuillet, B. Keller, J. Rogers, *et al.*, 2018 Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science (80-.). 361. https://doi.org/10.1126/science.aar7191

Arumuganathan K., and E. D. Earle, 1991 Nuclear DNA content of some important plant species. Plant Mol. Biol. Report. 9: 208–218. https://doi.org/10.1007/BF02672069

Beales J., A. Turner, S. Griyths, J. W. Snape, and D. A. Laurie, 2007 A Pseudo-Response Regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.). Theor Appl Genet 115: 721–733. https://doi.org/10.1007/s00122-007-0603-4

Chapman J. A., M. Mascher, A. Buluç, K. Barry, E. Georganas, *et al.*, 2015 A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 16. https://doi.org/10.1186/s13059-015-0582-8

Clavijo B. J., L. Venturini, C. Schudoma, G. G. Accinelli, G. Kaithakottil, *et al.*, 2017 An improved assembly and annotation of the allohexaploid wheat genome identifies complete

families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res. 27: 885–896. https://doi.org/10.1101/gr.217117.116

Coen E. S., and E. M. Meyerowitz, 1991 The war of the whorls: genetic interactions controlling flower development. Nature 353: 31–37. https://doi.org/https://doi.org/10.1038/353031a0

Díaz A., M. Zikhali, A. S. Turner, P. Isaac, and D. A. Laurie, 2012 Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum aestivum), (S. P. Hazen, Ed.). PLoS One 7: e33234. https://doi.org/10.1371/journal.pone.0033234

Dubcovsky J., and J. Dvorak, 2007 Genome plasticity a key factor in the success of polyploid wheat under domestication. Science (80-.). 316: 1862–1866.

Guo X., H. Su, Q. Shi, S. Fu, J. Wang, *et al.*, 2016 De Novo Centromere Formation and Centromeric Sequence Expansion in Wheat and its Wide Hybrids. PLoS Genet. https://doi.org/10.1371/journal.pgen.1005997

Kokot M., M. Dlugosz, and S. Deorowicz, 2017 KMC 3: counting and manipulating k-mer statistics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx304

Kurtz S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, *et al.*, 2004 Versatile and open software for comparing large genomes. Genome Biol. 5: R12. https://doi.org/10.1186/gb-2004-5-2-r12

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li D., C. M. Liu, R. Luo, K. Sadakane, and T. W. Lam, 2015 MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv033

Li H., 2018 Minimap2: pairwise alignment for nucleotide sequences, (I. Birol, Ed.). Bioinformatics 34: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Liu Y., H. Du, P. Li, Y. Shen, H. Peng, *et al.*, 2020 Pan-Genome of Wild and Cultivated Soybeans. Cell 0. https://doi.org/10.1016/j.cell.2020.05.023

Marçais G., A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, *et al.*, 2018 MUMmer4: A fast and versatile genome alignment system. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.1005944

Martin M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. https://doi.org/10.14806/ej.17.1.200

Mayer K. F. X., J. Rogers, J. Dole el, C. Pozniak, K. Eversole, *et al.*, 2014 A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science (80-.). 345: 1251788–1251788. https://doi.org/10.1126/science.1251788

Ng M., and M. F. Yanofsky, 2001 Function and evolution of the plant MADS-box gene family. Nat. Rev. Genet. 2: 186–195.

Pertea M., and G. Pertea, 2020 GFF Utilities: GffRead and GffCompare. F1000Research 9: 304. https://doi.org/10.12688/f1000research.23297.1

Petersen G., O. Seberg, M. Yde, and K. Berthelsen, 2006 Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum). Mol. Phylogenet. Evol. 39: 70–82. https://doi.org/10.1016/j.ympev.2006.01.023

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

Schatz M. C., A. L. Delcher, and S. L. Salzberg, 2010 Assembly of large genomes using second-generation sequencing. Genome Res.

Shumate A., and S. Salzberg, 2020 Liftoff: an accurate gene annotation mapping tool. bioRxiv. https://doi.org/10.1101/2020.06.24.169680

Song J.-M., Z. Guan, J. Hu, C. Guo, Z. Yang, *et al.*, 2020 Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of Brassica napus. Nat. Plants 6: 34–45. https://doi.org/10.1038/s41477-019-0577-7

Soyk S., Z. H. Lemmon, F. J. Sedlazeck, J. M. Jiménez-Gómez, M. Alonge, *et al.*, 2019 Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. Nat. Plants 5: 471–479.

Würschum T., P. H. G. Boeven, S. M. Langer, C. F. H. Longin, and W. L. Leiser, 2015 Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. BMC Genet. 16: 1–8. https://doi.org/10.1186/s12863-015-0258-0

Würschum T., C. F. H. Longin, V. Hahn, M. R. Tucker, and W. L. Leiser, 2017 Copy number variations of *CBF* genes at the *Fr-A2* locus are essential components of winter hardiness in wheat. Plant J. 89: 764–773. https://doi.org/10.1111/tpj.13424

Würschum T., S. M. Langer, C. F. H. Longin, M. R. Tucker, and W. L. Leiser, 2018 A three-component system incorporating Ppd-D1, copy number variation at Ppd-B1, and numerous small-effect quantitative trait loci facilitates adaptation of heading time in winter wheat cultivars of worldwide origin. Plant Cell Environ. 41. https://doi.org/10.1111/pce.13167

589 Zimin A. V., D. Puiu, R. Hall, S. Kingan, B. J. Clavijo, *et al.*, 2017 The first near-complete
590 assembly of the hexaploid bread wheat genome, Triticum aestivum. Gigascience.
591 https://doi.org/10.1093/gigascience/gix097
592 Zimin A. V., and S. L. Salzberg, 2019 The genome polishing tool POLCA makes fast and
593 accurate corrections in genome assemblies. bioRxiv 2019.12.17.864991.
594 https://doi.org/10.1101/2019.12.17.864991
595

FIGURE LEGENDS

Figure 1. The Triticum_aestivum_4.0 assembly scaffolding pipeline. A diagram depicting the Triticum_aestivum_4.0 (T4) assembly scaffolding pipeline, which takes the Triticum_aestivum_3.0 (T3) and IWGSC CS v1.0 (IW) assemblies as input. Grey cylinders represent input or output genome assemblies, while orange boxes show the steps of the scaffolding process.

Figure 2. A comparison of Triticum_aestivum_4.0 and IWGSC CS v1.0 assembly completeness. An ideogram showing the distribution of gap sequence in the Triticum_aestivum_4.0 (T4) and IWGSC CS v1.0 (IW) assemblies. The heatmap color intensity corresponds to the percentage of gap sequence in non-overlapping 1 Mbp windows along each chromosome. Chromosomes are sorted by T4 length (left to right, top to bottom), highlighting that each T4 chromosome across all three subgenomes has more sequence and fewer gaps than its IW counterpart.

Figure 3. Shared assembly k-mer count distribution. Histogram of 101-mer copy number in the Triticum_aestivum_4.0 (T4) and IWGSC CS v1.0 (IW) assemblies. Only 101-mers shared by both assemblies are considered. While IW has more single-copy 101-mers, T4 represents more 101-mers at higher copy numbers.

Figure 4. Triticum_aestivum_4.0 resolves previously collapsed genic repeats. (A) Histogram depicting the distribution of the number of additional gene copies found in Triticum_aestivum_4.0. **(B)** Circos plot showing the locations of all additional gene copies

(http://omgenomics.com/circa/). Lines are drawn from the location of the gene in IWGSC CS v1.0 (IW) on the right half of the diagram to the location of each copy in Triticum_aestivum_4.0 (T4) on the left half. **(C)** Dotplot depicting maximal exact matches (MEMs) between T4 *Ppd-B1* (x-axis) and a publicly available Chinese Spring *Ppd-B1* sequence (GenBank accession JF946485.1) (y-axis). Dashed lines indicate the co-linear positions of four PRR genes (red labels). **(D)** Diagram of the MADS-box transcription factor gene, *TraesCS6A02G022700,* present in three additional tandem copies in T4 as relative to IW. Ideograms are not drawn to scale. **(E)** Plot of the short-read coverage in IW starting 5kb upstream of *TraesCS6A02G02270* and extending to the first gap downstream of the gene. The pink dashed lines show the location of the gene.

Assembly	**T4**	**IW**	**T3**
All sequence (bp)	**15,397,713,314**	14,271,578,887	15,344,693,583
Anchored sequence (bp)	**15,070,919,678**	13,840,498,961	N/A

Table 1. Non-gapped sequence length of the Triticum_aestivum_4.0 (T4), IWGSC CS v1.0 (IW), and Triticum_aestivum_3.1 (T3) assemblies.

Figure 1

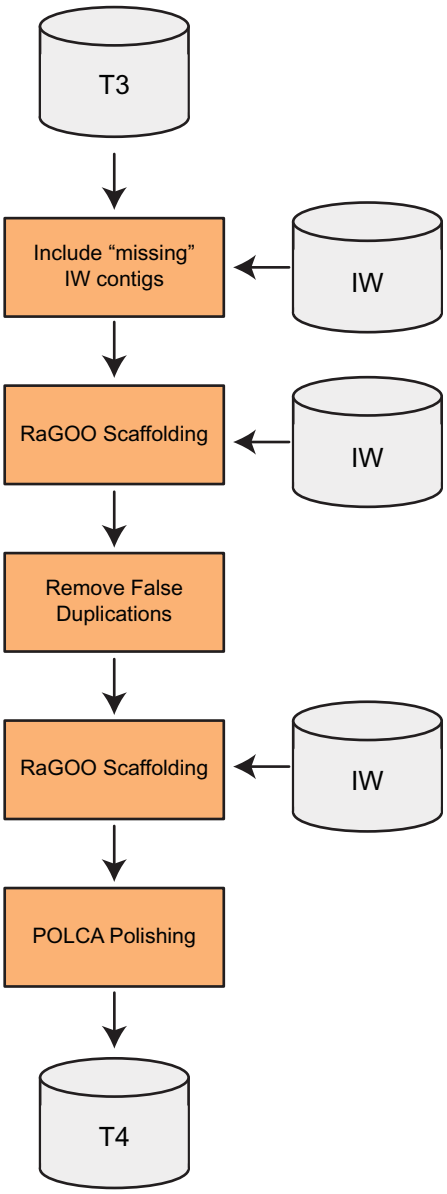

T3
Include “missing” IW contigs
IW
RaGOO Scaffolding
IW
Remove False Duplications
RaGOO Scaffolding
IW
POLCA Polishing
T4

Figure 2

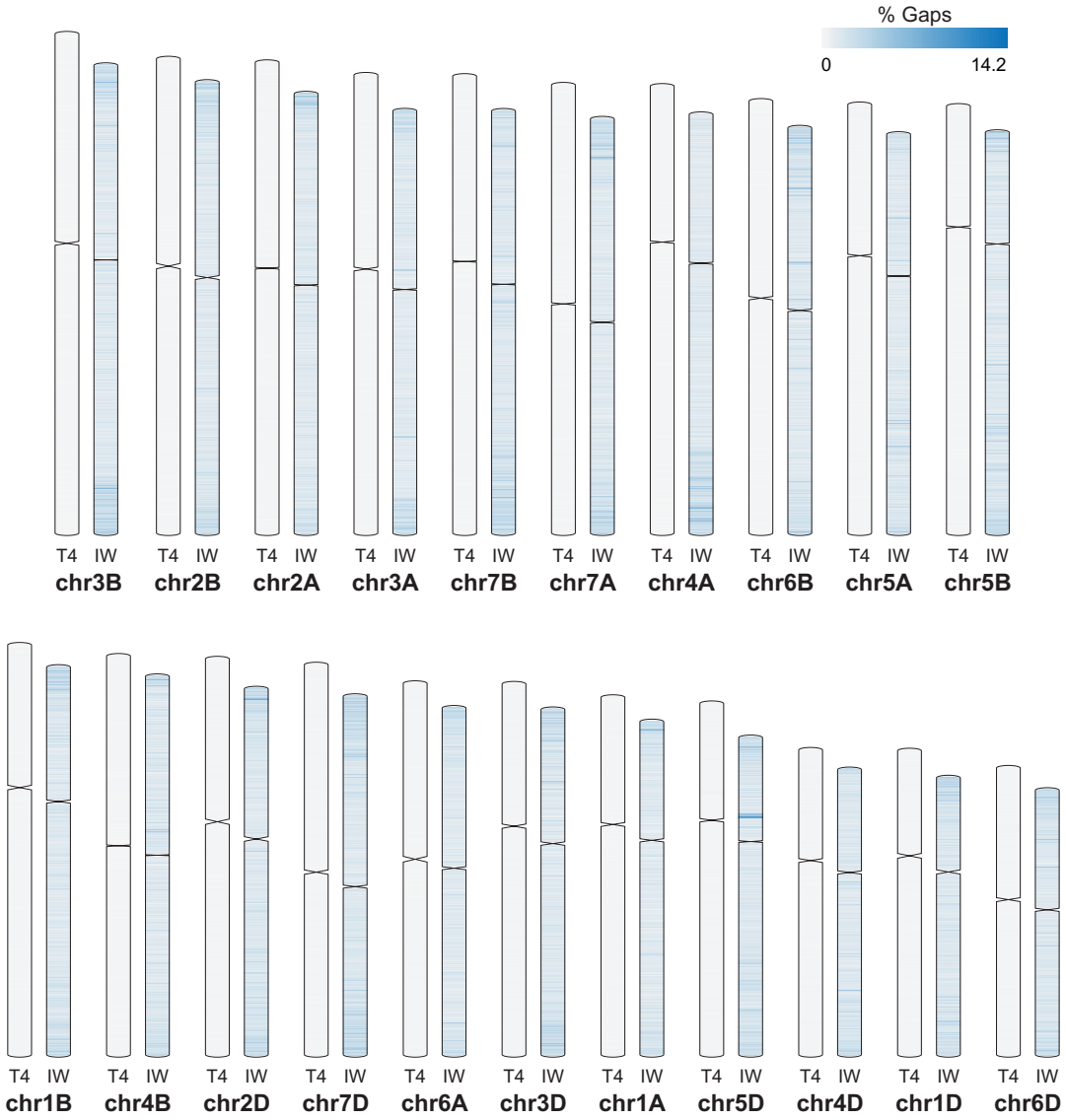

% Gaps
0
14.2
T4 IW
chr3B
chr2B
chr2A
chr3A
chr7B
chr7A
chr4A
chr6B
chr5A
chr5B
chr1B
chr4B
chr2D
chr7D
chr6A
chr3D
chr1A
chr5D
chr4D
chr1D
chr6D

Figure 3

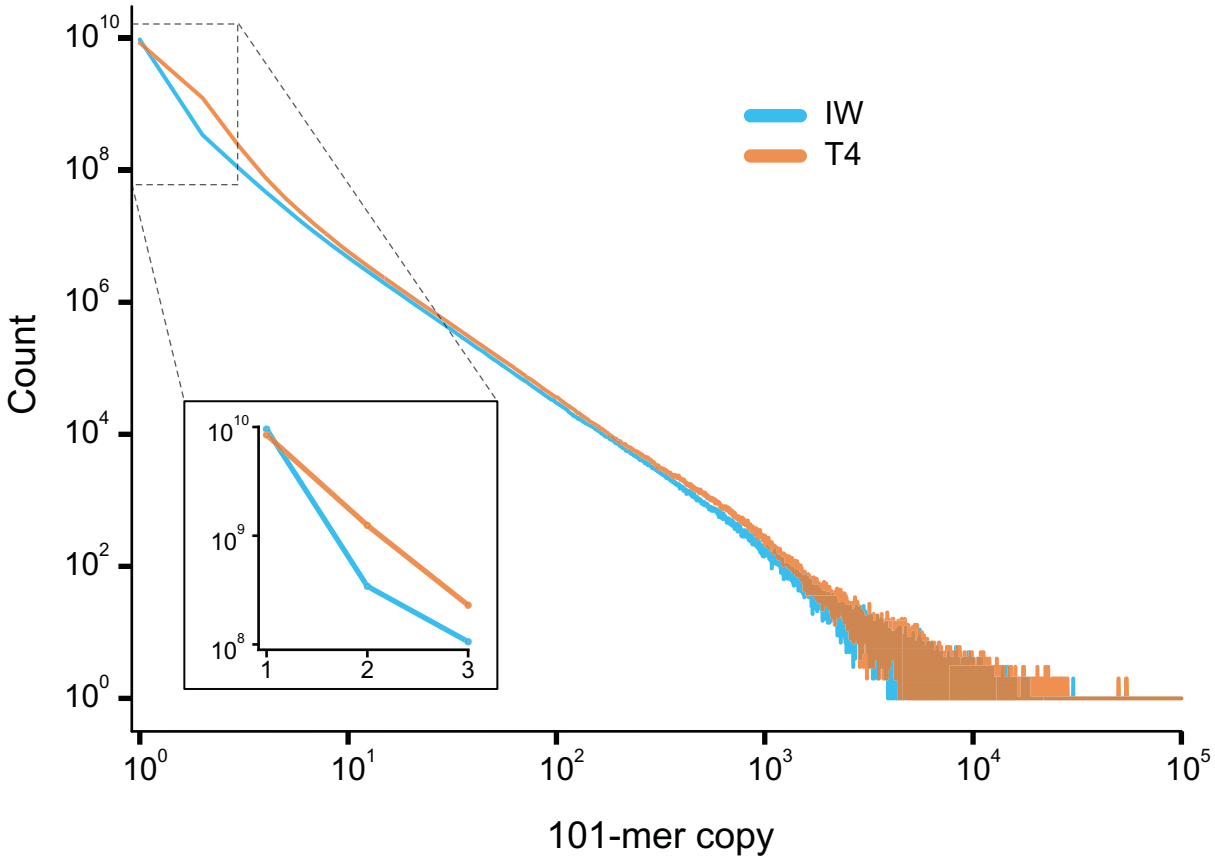
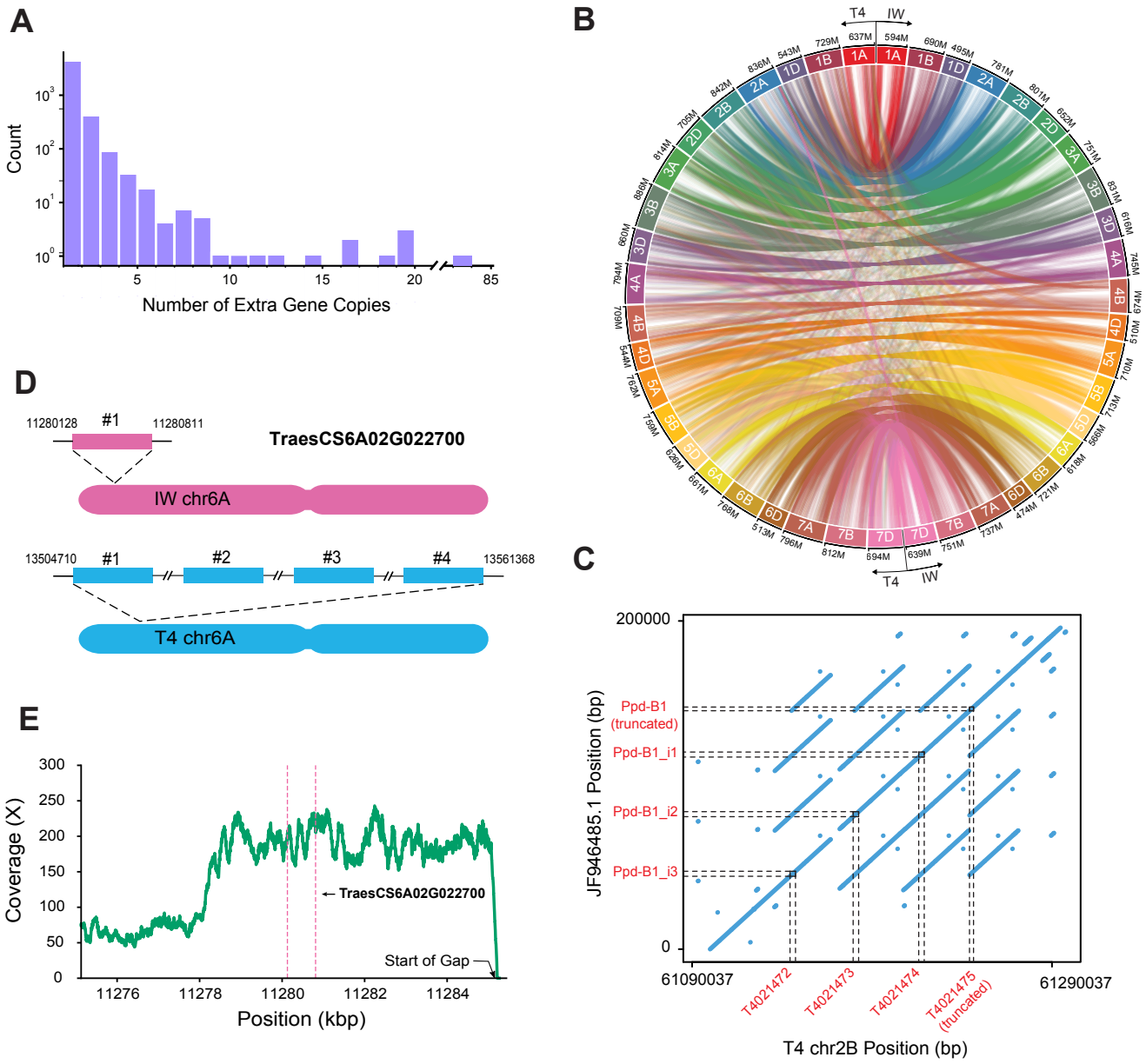

IW
T4
Count
101-mer copy
10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
10^{0}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
10^{9}
1
2
3

Figure 4


A
Count
Number of Extra Gene Copies
B
T4
IW
C
JF946485.1 Position (bp)
Ppd-B1 (truncated)
Ppd-B1_i1
Ppd-B1_i2
Ppd-B1_i3
T4021472
T4021473
T4021474
T4021475 (truncated)
T4 chr2B Position (bp)
D
11280128
#1
11280811
TraesCS6A02G022700
IW chr6A
13504710
#2
#3
#4
13561368
T4 chr6A
E
Coverage (X)
TraesCS6A02G022700
Start of Gap
Position (kbp)