Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly

Tingting Zhu1,†, Le Wang1,†, Hélène Rimbert2,†, Juan C. Rodriguez1,†, Karin R. Deal1, Romain De Oliveira2, Frédéric Choulet2, Gabriel Keeble-Gagnère3, Josquin Tibbits3, Jane Rogers4, Kellye Eversole5, Rudi Appels3,4, Yong Q. Gu5, Martin Mascher6 and Ming-Cheng Luo1,∗

1Department of Plant Sciences, University of California, Davis, CA 95616, USA,
2GDEC, Université Clermont Auvergne, INRAE, Clermont-Ferrand 63000, France,
3Centre for AgriBioscience, Agriculture Victoria, AgriBio, Bundoola, VIC 3083, Australia,
4International Wheat Genome Sequencing Consortium, Eau Claire, WI 54701, USA,
5Crop Improvement and Genetics Research Unit, USDA-ARS, Albany, CA 94710, USA, and
6Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Seeland, Germany

Received 17 November 2020; revised 12 April 2021; accepted 19 April 2021.
*For correspondence (e-mail mcluo@ucdavis.edu; jdvorak@ucdavis.edu).
†These authors contributed to this work.

SUMMARY

Until recently, achieving a reference-quality genome sequence for bread wheat was long thought beyond the limits of genome sequencing and assembly technology, primarily due to the large genome size and > 80% repetitive sequence content. The release of the chromosome scale 14.5-Gb IWGSC RefSeq v1.0 genome sequence of bread wheat cv. Chinese Spring (CS) was, therefore, a milestone. Here, we used a direct label and stain (DLS) optical map of the CS genome together with a prior nick, label, repair and stain (NLRS) optical map, and sequence contigs assembled with Pacific Biosciences long reads, to refine the v1.0 assembly. Inconsistencies between the sequence and maps were reconciled and gaps were closed. Gap filling and anchoring of 279 unplaced scaffolds increased the total length of pseudomolecules by 168 Mb (excluding Ns). Positions and orientations were corrected for 233 and 354 scaffolds, respectively, representing 10% of the genome sequence. The accuracy of the remaining 90% of the assembly was validated. As a result of the increased contiguity, the numbers of transposable elements (TEs) and intact TEs have increased in IWGSC RefSeq v2.1 compared with v1.0. In total, 98% of the gene models identified in v1.0 were mapped onto this new assembly through development of a dedicated approach implemented in the MAGAAT pipeline. The numbers of high-confidence genes on pseudomolecules have increased from 105,319 to 105,534. The reconciled assembly enhances the utility of the sequence for genetic mapping, comparative genomics, gene annotation and isolation, and more general studies on the biology of wheat.

Keywords: direct label and stain, pseudomolecule, transposable element, gene collinearity, Hi-C.

INTRODUCTION

Wheat provides about one-fifth of the calories and proteins consumed by humans, and is annually planted on an area larger than any other crop. About 95% of the wheat area is planted with bread wheat (*Triticum aestivum*, 2n = 6x = 42, genome formula AABBD), which is used for bread, noodles, pastry, cookies and related products; and the remaining 5% is planted with durum wheat (*Triticum turgidum* ssp. *durum*, 2n = 4x = 28, genome formula AABB) which is used for pasta (Dubcovsky and Dvorak, 2007).

Wheat yields have not been increasing at a rate sufficient to meet anticipated demands (Ray et al., 2013). The development of genomic resources, including a reference-quality genome sequence, is critical for accelerating genetic improvement of wheat. Due to polyploidy, the large sizes of the wheat subgenomes, and the high proportion of repeated sequences (85% of the genome; Wicker et al., 2018), the assembly of a reference-quality genome sequence for a polyploid wheat species was only recently possible as a result of improvements in sequencing and sequence assembly technologies. The first reference-quality polyploid wheat genome sequence was reported for wild emmer wheat (*T. turgidum* ssp. *dicoccoides*, 2n = 4x = 28, genome formula AABB) acc. ‘Zavitan’ (Avni et al., 2017), and was followed by the publication of the
reference sequences for bread wheat cv. Chinese Spring (CS; IWGSC, 2018) and durum wheat cv. Svevo (Maccarelli et al., 2019).

The wild emmer genome sequence (WEW_v1.0) was assembled from whole-genome shotgun (WGS) reads by NRGene Inc. using their DeNovoMAGIC assembler (Avni et al., 2017). Subsequent alignment of the WEW_v1.0 pseudomolecules to wheat genome-wide optical maps revealed some inconsistencies resulting from mis-orientation or positioning of scaffolds (Dvorak et al., 2018). Using optical maps of Zavitan, about 10% of the WEW_v1.0 assembly was revised in the WEW_v2.0 assembly (Zhu et al., 2019a).

An Illumina/DeNovoMAGIC assembly was also the core resource of the CS reference genome sequence (IWGSC RefSeq v1.0). This assembly was validated, enriched and improved using numerous independent resources (e.g. physical maps and mapped sequence tags) produced between 2008 (Paux et al., 2008) and 2018 by the International Wheat Genome Sequencing Consortium (IWGSC). Genome-wide optical maps were not available for validation of IWGSC RefSeq v1.0, nor were they used in the validations of other CS genome assemblies, such as the assembly built from a combination of Illumina and Pacific Biosciences (PacBio) reads (Triticum 3.1; Zimin et al., 2017) recently updated to Triticum 4.0 (Alonge et al., 2020).

The first optical map built for a Triticeae species employed a chemistry utilizing a single-strand restriction endonuclease to nick DNA (Hastie et al., 2013; Luo et al., 2017). The nicks were enzymatically repaired and labeled, and the DNA molecules were counterstained for imaging. This chemistry was consequently called nick, label, repair and stain (NLRS). The first genome-wide optical map for the CS genome was built with the NLRS chemistry (Huo et al., 2017). Subsequent alignment of the WEW_v1.0 pseudomolecules to wheat genome-wide optical maps revealed some inconsistencies resulting from mis-orientation or positioning of scaffolds (Dvorak et al., 2018). Using optical maps of Zavitan, about 10% of the WEW_v1.0 assembly was revised in the WEW_v2.0 assembly (Zhu et al., 2019a).

Table 1 Comparison of genome-wide optical maps of the CS genome built with the DLS and NLRS chemistries

| Sample   | DLS       | NLRS     |
|----------|-----------|----------|
| Enzyme   | DLE-1     | Nt.BspQI |
| Molecule size N50 (kb) | 306 | 284 |
| Molecule min. length (kb) | 150 | 180 |
| Molecule total length (Gb) | 2096 | 1927 |
| Genome coverage (x) | 136 | 113 |
| Map contigs (no.) | 726 | 11327 |
| Map total length (Gb) | 14.41 | 14.15 |
| Map contig length N50 (Mb) | 55.01 | 1.69 |
| Max. contig length (Mb) | 331.43 | 10.52 |

DLS, direct label and stain; NLRS, nick, label, repair and stain.

© 2021 The Authors. The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd, The Plant Journal (2021), doi: 10.1111/tpj.15289
pseudomolecules, and fewer were caused by errors in the assembly of scaffolds. Discrepancies of 10 kb or less, if present, were not detected as they were below optical map resolution.

**Pseudomolecule reconstruction**

To correct ordering and orientation problems, pseudomolecule scaffolds were combined with unplaced scaffolds to produce a dataset of 138,546 scaffolds (scf_v0 in Table 2 and Figure 2) that was aligned to the DLS and NLRS optical maps to identify scaffolds with errors in their assembly. The alignments identified 81 conflicts within 72 scaffolds covering 429.408 Mb. The chimeric scaffolds were corrected, thereby increasing the number of scaffolds from 138,546 to 138,634, and decreasing their N50 from 6.871 to 6.775 Mb (scf_v1 in Table 2 and Figure 2).

Super-scaffolds were rebuilt from scf_v1 scaffolds with the aid of the DLS map. The total length of the resulting super-scaffolds was 14.712 Gb with an N50 of 64.014 Mb (scf_v2 in Table 2 and Figure 2). The scf_v2 scaffolds were then aligned to the NLRS optical map, which facilitated additional super-scaffolding of the sequences. The N50 of the resulting super-scaffolds (scf_v3 in Figure 2) reached 72.092 Mb, with the largest recording a length of 364.575 Mb (Table 2).

The scf_v3 super-scaffolds were ordered and oriented through alignment on high-density genetic maps. A total of 673 super-scaffolds (14.311 Gb) sharing two or more markers with genetic maps were anchored on the 21 wheat chromosomes to produce pseudomolecules_v1.1. A comparison with the pseudomolecules of the IWGSC RefSeq v1.0 assembly showed that 279 more scaffolds (74.960 Mb) from unallocated scaffolds (ChrUn) were now anchored on the chromosomes, 354 scaffolds (469.834 Mb) were re-oriented, and 233 scaffolds (394.868 Mb) were moved into new locations (Table S3).

**Gap closing and final refinement of the pseudomolecules**

During the scaffold correction stage, the contigs within scaffolds remained largely the same as in pseudomolecules_v1.0, retaining gaps filled with Ns. In addition, there were 4021 gaps of unknown sizes in pseudomolecules_v1.0 (Table 3). Pseudomolecules_v1.1 acquired most of the intra-scaffold gaps and had 647 gaps of unknown sizes (Table 3). To close gaps in pseudomolecules_v1.1, 97,809 PacBio contigs equaling 12.939 Gb and N50 of 264.143 bp were polished with CS DV418 Illumina reads. Polishing corrected 7,221,194 (0.056%) base errors in the PacBio contigs. The PacBio contigs were aligned on the optical maps, and 96,363 of the contigs totaling 12.291 Gb were validated. The validated PacBio contigs were used for gap closure by aligning them on pseudomolecules_v1.1 and substituting N bases with actual nucleotide sequences. Of the 527.170 gaps in pseudomolecules_v1.1, 343.566, including 12 gaps of unknown length (Table 3), were closed to generate pseudomolecules_v2.0. In addition, 91,820 unplaced scaffolds with a total length of 351.583 Mb were assigned to ChrUn. The pseudomolecules_v2.0 along with ChrUn constituted the intermediate IWGSC RefSeq v2.0 assembly.

Comparison of IWGSC RefSeq v2.0 with IWGSC RefSeq v1.0 detected SNPs and single-base indels in 12,382 coding sequences (CDSs) from 9,416 genes. Assuming that the differences were introduced into the assembly with the PacBio reads, the sequences harboring these CDSs were replaced with corresponding contigs from IWGSC RefSeq v1.0. This final refinement produced the final genome assembly IWGSC RefSeq v2.1 (Figure 3; Table 4).

**Table 2** Scaffolding steps using optical maps

| Scaffolds (no.) | scf_v0 | Chimera resolved (scf_v1) | Scaffolded using DLS map (scf_v2) | Scaffolded using NLRS map (scf_v3) |
|-----------------|--------|--------------------------|----------------------------------|-----------------------------------|
| Max. length (bp)| 45,793 | 45,793 851               | 364,574 574                      | 364,574 574                       |
| Total length (bp)| 14,533 | 14,533 405 144       | 14,711 983 205                   | 14,710 380 608                    |
| Sequence N50 (bp)| 6,870 | 6,774 542           | 64,014 232                      | 72,091 519                       |
| N/5 | 1.80 | 1.80                   | 2.99                             | 3.00                              |

DLS, direct label and stain; NLRS, nick, label, repair and stain.
Gene reannotation

Evaluation of the new sequence assembly showed that the number of differences between IWGSC RefSeq v1.0 and IWGSC RefSeq v2.1, primarily arising from gap filling, was too great to allow for a simple correlation between the two versions to generate new gene coordinates. Gene sequence alignments were, therefore, re-computed to map the gene annotation to the new assembly.

Before transferring the gene annotation onto the new genome assembly, it was updated to IWGSC Annotation v1.2 by integrating a set of 117 novel genes and 81 micro-RNAs (miRNAs), many of which had been manually curated by the wheat community. The new annotation release containing 108 010 HC and 161 535 LC gene models (Table S1) was used to annotate IWGSC RefSeq v2.1.

Mapping a query gene on the whole wheat genome often leads to spurious alignments because of the high number of homologous gene copies. In order to mitigate this problem, the interval that harbored a gene to be mapped was first identified using Insertion Site-Based Polymorphisms (ISBP) markers, which mark junctions between neighboring TEs that are present in both the IWGSC RefSeq v1.0 and IWGSC RefSeq v2.1 assemblies.

ISBPs are 150-mers representing a unique sequence shaped by the junction of a TE and its insertion site (De Oliveira et al., 2020; Paux et al., 2010; Rimbert et al., 2018). Because ISBPs can be derived at each TE extremity, they are extremely abundant and well adapted for identifying corresponding loci across different assembly versions. In wheat, almost all genes are closely flanked by a pair of unique ISBPs.

A total of 5 394 172 ISBPs were designed on IWGSC RefSeq v1.0 and mapped to IWGSC RefSeq v2.1 (Table S1). Over 90% of these ISBPs (4 908 316) mapped fully and uniquely with no mismatches, giving an average density of one ISBP every 3 kb. An ISBP-flanked interval on IWGSC RefSeq v2.1 was assigned for 264 876 (98.3%) of the 269 545 gene models, corresponding to 107 877 high-confidence (HC) and 161 668 low-confidence (LC) gene models. They represented 207 575 intervals containing between 1 and 4 genes, with 83% of the intervals containing a single gene (Table S1). The gene intervals spanned 668, 683 and 512 Mb of the A-, B- and D-subgenome sequences, respectively, and 136 Mb of ChrUn scaffolds. The average size of an interval was 9.6 kb. A total of

| Chromosome | v1.0 | v1.1 | v2.0 |
|------------|------|------|------|
| Chr1A      | 16   | 13   | 145  |
| Chr2A      | 18   | 19   | 171  |
| Chr3A      | 7    | 7    | 186  |
| Chr4A      | 34   | 34   | 236  |
| Chr5A      | 13   | 13   | 151  |
| Chr6A      | 11   | 11   | 158  |
| Chr7A      | 15   | 15   | 196  |
| Chr1B      | 45   | 45   | 191  |
| Chr2B      | 38   | 38   | 273  |
| Chr3B      | 48   | 48   | 237  |
| Chr4B      | 59   | 63   | 204  |
| Chr5B      | 54   | 55   | 250  |
| Chr6B      | 61   | 62   | 276  |
| Chr7B      | 58   | 58   | 281  |
| Chr1D      | 16   | 16   | 133  |
| Chr2D      | 16   | 17   | 167  |
| Chr3D      | 19   | 20   | 159  |
| Chr4D      | 24   | 26   | 115  |
| Chr5D      | 27   | 28   | 170  |
| Chr6D      | 27   | 27   | 121  |
| Chr7D      | 32   | 32   | 201  |
| Total      | 635  | 647  | 4021 |
241,201 (90%) genes with identical sequence between the two assemblies were unambiguously mapped, and 22,578 (8.5%) genes exhibited nucleotide differences (mismatches and/or indels) that affected only introns or UTRs. Together, they represented 263,779 (98%) genes from IWGSC RefSeq v1.0. Of these, 31,520 genes (12%) would not have transferred accurately if GMAP against the whole IWGSC RefSeq v2.1 had been used, demonstrating the importance of reducing the genome complexity before mapping as implemented in our pipeline MAGATT (https://forgemia.inra.fr/umr-gdec/magatt).

Of the remaining 2% of IWGSC RefSeq v1.0 genes, 2,974 (1%) were aligned with mismatches/indels in the CDSs (or only partially aligned in rare cases), meaning that their sequence has changed in IWGSC RefSeq v2.1 compared with v1.0. The remaining 2792 IWGSC RefSeq v1.0 genes (1%) could not be identified in IWGSC RefSeq v2.1. A total of 1,258 genes previously on the ChrUn scaffolds were assigned to chromosomes in IWGSC RefSeq v2.1.

All genes were renamed to more accurately reflect the position of the gene and the annotation. For example, in a new name TraesCS1A03G000000[LC], 1A and 03 specify the location of a low-confidence gene on chromosome 1A and the third annotation version, respectively.

In total, the new release IWGSC RefSeq Annotation v2.1 contains 266,753 genes comprising 106,913 HC genes and 159,840 LC genes (Table S1). Altogether, 105,534 HC and 155,624 LC genes were located on the pseudomolecules, and 1,379 HC and 4,216 LC genes were located on scaffolds assigned to ChrUn (Table S1).

Table 4. Comparison of the total and effective lengths (excluding Ns) of IWGSC RefSeq v2.1 pseudomolecules with those of IWGSC RefSeq v1.0

| Chr ID | RefSeq v2.1 | RefSeq v1.0 |
|--------|-------------|-------------|
| Chr1A  | 598,660,471 | 590,447,024 |
| Chr2A  | 787,782,082 | 777,846,456 |
| Chr3A  | 754,129,162 | 743,748,734 |
| Chr4A  | 754,227,511 | 741,420,740 |
| Chr5A  | 713,360,525 | 704,968,565 |
| Chr6A  | 622,669,697 | 614,336,433 |
| Chr7A  | 744,491,536 | 733,447,483 |
| Chr1B  | 700,547,350 | 689,076,938 |
| Chr2B  | 812,755,788 | 798,522,160 |
| Chr3B  | 851,934,094 | 831,834,594 |
| Chr4B  | 673,810,255 | 665,179,646 |
| Chr5B  | 714,805,278 | 704,718,176 |
| Chr6B  | 731,188,232 | 717,245,234 |
| Chr7B  | 764,081,788 | 749,505,598 |
| Chr1D  | 498,638,509 | 491,298,183 |
| Chr2D  | 656,544,405 | 647,090,398 |
| Chr3D  | 619,616,552 | 611,768,429 |
| Chr4D  | 518,332,611 | 511,996,999 |
| Chr5D  | 569,951,140 | 563,314,725 |
| Chr6D  | 495,380,293 | 488,354,375 |
| Chr7D  | 642,921,167 | 633,773,979 |
| ChrUn  | 351,582,993 | 308,528,982 |
| Total  | 14,576,989,303 | 14,316,999,506 |

© 2021 The Authors. *The Plant Journal* published by Society for Experimental Biology and John Wiley & Sons Ltd, *The Plant Journal*, (2021), doi: 10.1111/tpj.15289
Gene collinearity in homoeologous pseudomolecules

The refinement of the pseudomolecules was expected to improve gene collinearity between homoeologous pseudomolecules. The numbers of collinear genes and syntenic blocks (at least five genes were required to call a synteny block) between the A-, B- and D-subgenome homoeologous pseudomolecules were counted in the IWGSC RefSeq v2.1 and IWGSC RefSeq v1.0 assemblies to assess this expectation. In the pairwise subgenome comparisons, the number of collinear genes was greater ($P = 0.019$, two-tailed paired $t$-test, $N = 3$) in IWGSC RefSeq v2.1 than in IWGSC RefSeq v1.0 (Table 5). Concomitantly, there was a reduced fragmentation of syntenic blocks ($P = 0.006$, two-tailed paired $t$-test, $N = 3$), indicating higher collinearity between homoeologous chromosomes in the IWGSC RefSeq v2.1 assembly than in IWGSC RefSeq v1.0 (Table 5).

De novo TE annotation

De novo annotation of TEs with CLARITE (Daron et al., 2014) annotated 4,199,592 TEs belonging to 506 families. The total length of TEs increased from 11,921,309,743 to 12,092,094,168 bp. As a result of the increase in pseudomolecule length in IWGSC RefSeq v2.1 compared with v1.0 (Table 6), the percentage of the CS genome accounted for by TEs (85.0%) in IWGSC RefSeq v2.1 was nearly identical to that reported for IWGSC RefSeq v1.0 (84.7%). Because the same approach, tools and criteria were used, the percentages of the genome represented by TEs and by individual TE super-families and families were very similar in the IWGSC RefSeq v1.0 (Wicker et al., 2018) and v2.1 assemblies (Table 6). ISBPs can be used as a measure of similarity of the TE content (De Oliveira et al., 2020; Rimbert et al., 2018). Overall, 99.9% of ISBPs was shared by the two genome sequences.

Table 5 Numbers of collinear genes and syntenic blocks between subgenomes in the pseudomolecules of IWGSC RefSeq v1.0 and IWGSC RefSeq v2.1

| Subgenome comparison | Collinear genes (no.) | Syntenic blocks (no.) |
|----------------------|------------------------|-----------------------|
|                      | v1.0 | v2.1 | Difference (v2.1-v1.0) | v1.0 | v2.1 | Difference (v2.1-v1.0) |
| A and B              | 41,999 | 42,479 | 480 | 546 | 484 | –62 |
| A and D              | 44,280 | 45,008 | 728 | 581 | 500 | –81 |
| B and D              | 43,704 | 44,484 | 780 | 510 | 442 | –68 |

Table 6 Comparison of the TE content in IWGSC RefSeq v1.0 and IWGSC RefSeq v2.1, and among subgenomes of IWGSC RefSeq v2.1

|                        | IWGSC RefSeq v1.0 | IWGSC RefSeq v2.1 | IWGSC RefSeq v2.1 subgenomes |
|------------------------|-------------------|-------------------|-----------------------------|
|                        | Pseudomolecules (bp) | 14,066,280,851 | 14,225,829,371 |
|                        | Repeats (bp) | 11,921,309,743 | 12,092,094,168 |
|                        | TEs (%) | 84.75 | 85.00 |
|                        | Class 1<sup>a</sup> | 67.6 | 66.9 |
|                        | Gypsy (RLG) | 46.7 | 46.1 |
|                        | Copia (RLC) | 16.7 | 16.5 |
|                        | Unclass. LTR RTs | 3.24 | 3.26 |
|                        | LINEs (RIX) | 0.9 | 1.05 |
|                        | SINEs (SIX) | 0.01 | 0.01 |
|                        | Class 2<sup>b</sup> | 16.5 | 17.0 |
|                        | CACTA (DTC) | 15.5 | 15.9 |
|                        | Mutator (DTM) | 0.38 | 0.44 |
|                        | Unclass. with TiRs (DTX) | 0.21 | 0.24 |
|                        | Harbinger (DTH) | 0.16 | 0.18 |
|                        | Mariner (DTT) | 0.16 | 0.17 |
|                        | Unclass. class 2 (DXX) | 0.06 | 0.06 |
|                        | hAT (DTA) | 0.006 | 0.009 |
|                        | Helitrons (DHH) | 0.004 | 0.01 |
|                        | Unclass. repeats | 0.68 | 0.95 |
| A                      | 13.2 | 15.9 | 19.5 |
| B                      | 0.35 | 0.43 | 0.55 |
| D                      | 0.24 | 0.24 | 0.25 |
| Class 2<sup>c</sup>    | 0.17 | 0.17 | 0.19 |
|                        | 0.15 | 0.17 | 0.19 |
|                        | 0.26 | 0.07 | 0.06 |
|                        | 0.009 | 0.011 | 0.008 |
|                        | 0.011 | 0.010 | 0.010 |
|                        | 0.78 | 0.93 | 0.88 |

RT, retrotransposon; TE, transposable element; TiR, terminal inverted repeats.

<sup>a</sup>3-letter TE codes in parentheses, as defined in Wicker et al. (2007).

<sup>b</sup>RNA transposons.

<sup>c</sup>DNA transposons.
super-families and families were similar among the subge-
ages of subgenome lengths represented by individual
in the D-subgenome, mirroring the total lengths of pseudo-
centages of the total lengths of the A-, B- and D-
There were more intact
ent from the numbers of intact LTR retrotransposons (RTs).
IWGSC RefSeq v2.1 compared with v1.0. This was appar-
in the A- and B-subgenomes.
family was over-represented relative to their percentages
P
The Plant Journal
© 2021 The Authors.
The Plant Journal, (2021), doi: 10.1111/tpj.15289

Figure 4. Numbers of intact LTR retrotransposons (RTs) in IWGSC RefSeq v1.0 and IWGSC RefSeq v2.1.

In IWGSC RefSeq v2.1, TEs accounted for similar per-
centages of the total lengths of the A-, B- and D-
subgenome pseudomolecules (Table 6). The largest quanti-	y of TEs was in the B-subgenome and the smallest was in the D-subgenome, mirroring the total lengths of pseudo-
molecules in the three subgenomes (Table 6). The percents-
ages of subgenome lengths represented by individual
super-families and families were similar among the subge-
Table 6), although in the D-subgenome the Gypsy
super-family was under-represented and the CACTA super-
family was over-represented relative to their percentages
in the A- and B-subgenomes.

Gap closure improved the contiguity of TE sequences in
IWGSC RefSeq v2.1 compared with v1.0. This was apparent
from the numbers of intact LTR retrotransposons (RTs).
There were more intact Copia, Gypsy and unknown LTR
RTs (P < 0.0001 in each LTR RT class, two-tailed t-test with
Bonferroni correction, N = 21) in IWGSC RefSeq v2.1 than in v1.0 (Figure 4; Table S4).

Comparisons of IWGSC RefSeq v2.1 with CS BAC sequences
Comparison of a genome sequence assembly with sequences of bacterial artificial chromosomes (BACs)
assembled from reads produced with Sanger sequencing
technology has been used in the past to assess quality of a
genome sequence assembly (Luo et al., 2017). The IWGSC
RefSeq v2.1 sequence was compared with sequences of 15
CS BACs sequenced to high quality. Fourteen of these BACs
could be mapped onto the corresponding pseudomolecules
with mean identity > 99.5% and coverage ≥ 99.99%. This
sequence identity was comparable to that of the 99.75%
identity reported for an analogous comparison made with the
 Ae. tauschii Aet v4.0 genome sequence assembly (Luo et al., 2017). Clone GU817319.1 was for an unknown reason
exceptional; it matched with only 98.85% identity and 89.92%
coverage to the IWGSC RefSeq v2.1 sequence (Table S2).

Chromosome conformation capture sequencing (Hi-C) analysis
Hi-C was used to compare the structural integrity of the
IWGSC RefSeq v1.0 and 2.1 genome assemblies at the
chromosome scale level. Hi-C contact matrices (Figure S1)
showed the expected diagonal–antidiagonal matrix indica-
tive of the Rabl configuration of interphase nuclei (Mascher
et al., 2017). Contact matrices of the IWGSC RefSeq v1.0
assembly (Figure S2) indicated misassemblies in pseudo-
molecules Chr2A, Chr3A, Chr4A, Chr2B, Chr7B and Chr3D,
which were corrected in the IWGSC RefSeq v2.1 assembly.
A single discrepancy remained in the pericentromeric
region of chromosome 4A. Conflicting signals of cen-
tromeric histone H3 (CENH3) localization from chromatin
immunoprecipitation sequencing was reported for CS
chromosome 4A (IWGSC, 2018). Whether heterogeneity in
CS seed sources or some other factor is the cause of these
signals is not clear.

Comparisons with the Triticum 4.0 assembly
IWGSC RefSeq v2.1 was compared with the recent CS
assembly Triticum 4.0 (Alonge et al., 2020), which was pro-
duced from CS DV418 sequence data with the assistance
of IWGSC RefSeq v1.0. An analysis of a 10-Mb interval on
chromosome 1A (Chr1A) provides an example of differ-
ences between the assemblies (Figure 5). In that region,
the DLS optical map indicated that the IWGSC RefSeq v1.0
sequence harbored a 650-kb inversion and 602 kb of miss-
ing sequence (Figure 5a). Both the inversion and missing
sequence also appeared in the Triticum 4.0 assembly (Fig-
ure 5b), but both were rectified in RefSeq v2.1 by changing
the orientation of the 650-kb segment and inserting Ns
equivalent to 602 kb, respectively (Figure 5c). There were
only three DLE-1 restriction sites in this 602-kb interval
(Figure 5a–c), while there should have been about 90 sites
based on the genome average. The unusual structure of
this 602-kb region is most likely due to the presence of
repeats, which failed to assemble well in this region of
Chr1A. In addition to the two rearrangements incorporated
from IWGSC RefSeq v1.0, approximately 1.75 Mb of extra
sequence was present in Triticum 4.0 in this interval,
including a 250-kb sequence that our analysis assigned to
Chr7A (indicated by the orange arrow in Figure 5b).

Aligning the three CS sequences on the DLS optical
map indicated that many of the ambiguities in sequence
order, sequence orientation and chromosome location were
transferred from IWGSC RefSeq v1.0 to Triticum 4.0, while
additional conflicting arrangements were introduced by the
Triticum 4.0 assembly process itself (Table S5). In addi-
tion to conflicts in sequence arrangements, about 0.31 Gb
of sequences from other chromosomes appear mis-
assigned to pseudomolecules in Triticum 4.0, about
0.98 Gb of the sequence was missing (Table S5), and
about 1.54 Gb of the Triticum 4.0 sequence could not be
aligned with the DLS optical map. This may partially
account for the total length of the Triticum 4.0 pseudo-
molecules exceeding that of the IWGSC RefSeq v2.1 by
6%.

© 2021 The Authors.
The Plant Journal published by Society for Experimental Biology and John Wiley & Sons Ltd,
The Plant Journal, (2021), doi: 10.1111/tpj.15289
Concluding remarks

The deployment of optical maps plays two main roles in the assembly of a genome. First, an optical map represents an independent means for validating a sequence assembly. The deployment of the two CS optical maps validated about 90% of the IWGSC RefSeq v1.0 assembly (IWGSC, 2018), confirming that the assembly was indeed of high quality. Second, as the lengths of DLS optical contigs greatly exceed the lengths of sequence scaffolds, a DLS optical map is a valuable tool for scaffolding and super-scaffolding during sequence assembly. The deployment of the CS optical maps anchored 279 previously unplaced scaffolds (ChrUn) onto the pseudomolecules, reoriented 354 scaffolds and relocated 233 scaffolds, thus guiding the revision of approximately 10% of the genome sequence. Our experience illustrates the value of using multiple genome-wide optical maps to guide sequence assembly of complex genomes.

Closing gaps in the IWGSC RefSeq v1.0 sequence was another refinement of the assembly. The combined effect of anchoring 279 unplaced scaffolds (74.960 Mb) and gap filling with PacBio long-read sequences increased the length (excluding Ns) of the pseudomolecules by approximately 168 Mb (1.19%), from 13.840 Gb in IWGSC RefSeq v1.0 to 14.008 Gb in IWGSC RefSeq v2.1. Closing gaps also increased the number of intact LTR RTs by 13% as well as the overall number of identified TEs. It increased the number of HC genes on the pseudomolecules from 105 319 in Annotation v1.2 to 105 534 HC genes in Annotation v2.1, and reduced the total number of genes on unassigned scaffolds from 9449 in Annotation v1.2 to 5595 in Annotation v2.1 (Table S1).

Compared with IWGSC RefSeq v1.0, the gene space in the IWGSC RefSeq v2.1 pseudomolecules more faithfully reflects the divergence of the wheat subgenomes from a common ancestral genome in that a greater number of collinear genes and a reduced fragmentation of syntropic blocks were found in the IWGSC RefSeq v2.1 pseudomolecules. These refinements of both the TE and gene space argue for the inclusion of long reads into a reference-genome sequence assembly. Importantly, we found that it was best to use PacBio contigs validated by optical maps to close gaps. This highlights the importance of combining multiple independent sources of information to produce a consensus that is less likely to suffer from issues related to a particular technology.

Revisions present in the IWGSC RefSeq v2.1 will enhance the utility of the genome sequence in breeding and research applications. Correcting scaffold order and orientation is critical for comparative studies and genetic mapping of quantitative trait loci, as well as any other approaches relying on linkage disequilibrium such as Genomic Prediction. The assignment of previously unplaced scaffolds into correct locations of the pseudomolecule and gap closure allows the genomic regions underpinning trait associations to be more fully studied, and facilitates a more complete gene annotation. Sizing gaps of previously unknown sizes allows local scale to be assessed, and the cross-validation of sequence with optical maps enables the identification of remaining problem regions.

Despite the refinements of the IWGSC RefSeq v2.1 assembly reported here, it is evident that more remains to be done. This study corrects most of the macro-scale assembly issues that were present in v1.0. Issues that are more difficult to resolve remain in localized regions and will require integration of new data types, especially in areas where the resolution of optical maps was inadequate to detect or resolve discrepancies. Also, some small scaffolds could not be aligned on the optical maps and were discarded. Some optical map contigs and aligned scaffolds may have been incorporated into the pseudomolecule either arbitrarily or were kept as in IWGSC RefSeq v1.0,
because of a lack of markers or ambiguity in genetic maps. Moreover, 351.583 Mb of assembled scaffolds remain unplaced and assigned to ChrUn. The sequences of the centromeric regions, rRNA loci, telomeric and subtelomeric regions, regions containing satellite DNA, and other hard-to-assemble regions remain largely unanchored and/or collapsed. The successful assembly of centromeric and telomeric regions of a walnut hybrid (~1.3 Gb genome size) with a combination of short- and long-read sequencing technologies and optical maps (Zhu et al., 2019b) suggests that with these and other technologies the IWGSC sequence of the CS genome can be further refined. The IWGSC RefSeq v2.1 assembly we are now providing is a major advance for applied and basic applications.

EXPERIMENTAL PROCEDURES

Plants

_Triticum aestivum_ cv. Chinese Spring DV418 was used to construct the DLS optical map and to produce the WGS PacBio SMRT long reads. DV418 is a doubled haploid of a CS accession provided by D.R. Knott, University of Saskatchewan, Saskatoon, Canada.

Comparison of the parental CS stocks

Illumina reads of CS stock DV418 (Zimin et al., 2017) were downloaded from NCBI. These reads were mapped to the IWGSC RefSeq v1.0 assembly (IWGSC, 2018) using BWA-MEM (Li and Durbin, 2010) with default settings. After removing polymerase chain reaction duplicates, variant calling was performed and filtered with QC \( \geq \) 20 and depth \( \geq 3 \) using SAMAtools (Li et al., 2009) and BCFtools (Narasimhan et al., 2016). Only homozygous sites were kept and analyzed.

Construction of the CS DLS optical map

Seeds of CS were germinated and grown in the dark, and young leaves were collected from these seedlings. Ultra-high-molecular-weight (uHMW) DNA was isolated using the Plant DNA Isolation Kit (Bionano Genomics, San Diego, CA, USA). The uHMW DNA molecules were labeled with the DLE-1 enzyme (Bionano Genomics), and were then stained with the Bionano PrepTM DLS Kit (Bionano Genomics). A consensus optical map was de novo assembled with the Assembler tool of the Bionano Solve v3.3 package using significance cutoffs of \( P < 1 \times 10^{-11} \) to generate draft consensus contigs, \( P < 1 \times 10^{-12} \) for draft consensus contig extension, and \( P < 1 \times 10^{-16} \) for final merging of the draft consensus contigs. A recipe of ‘non-haplotype’, ‘noES’ and ‘noCut’ was chosen. The initial optical map was then checked for potential chimeric contigs. Problematic regions were manually analyzed by altering contig assembly stringency, and by interrogating the NLRS CS optical map and DNA sequences.

Reconstruction of pseudomolecules

The stepwise process of pseudomolecule reconstruction is summarized in Figure 2.

_Chimera resolution._ The 21 pseudomolecules of the IWGSC RefSeq v1.0 genome assembly (IWGSC, 2018) were split into individual scaffolds at positions comprised of 100 Ns, which represented gaps of unknown lengths between scaffolds. The scaffolds were combined with unplaced scaffolds (ChrUn) of the IWGSC RefSeq v1.0 assembly. All scaffolds were validated by aligning them on the CS DLS optical map using RefAligner tool of the Bionano Solve v3.3 package (Bionano Genomics), with an initial alignment cutoff of \( P < 1 \times 10^{-10} \). Alignments in which a scaffold disagreed with the DLS optical map were aligned to the CS NLRS optical map (Huo et al., 2018) to determine if the sequence or the DLS contig was incorrectly assembled. Incorrectly assembled scaffolds were disjoined.

Super-scaffolding. Conflict-free scaffolds were super-scaffolded using the Hybrid Scaffold pipeline of Bionano Solve v3.3 package (Bionano Genomics), with an alignment cutoff of \( P < 1 \times 10^{-10} \). Gaps were filled with the number of Ns corresponding to the estimated length in bp between restriction sites flanking a gap on the DLS optical map. The NLRS optical map was used to further super-scaffold the sequence, which produced the final super-scaffolds.

Anchoring of super-scaffolds onto the chromosomes. High-density genetic maps of hexaploid wheat (Wang et al., 2014), wild emmer (Avni et al., 2014; Jorgensen et al., 2017) and Ae. tauschii acc. AL8/78 (Luo et al., 2017) were used to determine the order and orientation of super-scaffolds on each chromosome. The flow-sorted chromosome arm DNA (Chromosome Survey Sequencing) sequences (IWGSC, 2014) were used to confirm chromosome arm and subgenome allocation of super-scaffolds. Super-scaffolds were then linked with 100 Ns and anchored onto the 21 CS chromosomes.

Additional gap closure with PacBio contigs. Approximately 545 Gb of raw reads (~36 × coverage) of CS DV418 (NCBI SRX2994550) generated with PacBio sequencing technology were assembled with Falcon (Zimin et al., 2017). To polish the PacBio contigs further, raw CS DV418 Illumina reads (NCBI SRX2994097) were downloaded from NCBI and filtered using Sickle v1.33 (Joshi and Fass, 2011) with default parameters. The clean illumina reads were mapped on to the PacBio contigs using BWA v0.7.16 (Li and Durbin, 2010) and unique alignments were kept. ‘SNP’ or ‘INDEL’ calls were corrected with Pilon v1.18 (Walker et al., 2014). These PacBio contigs were validated with the CS DLS and NLRS optical maps. Only those contigs that agreed with both maps were used for gap closing. The validated contigs were then aligned to pseudomolecules_v1.1 using BLAST. When a region in pseudomolecules_v1.1 aligned unambiguously with a PacBio contig and contained gap(s), it was replaced by the corresponding PacBio contig; thus, N bases in the pseudomolecules were filled with the actual nucleotide sequence, which resulted in Pseudomolecules_v2.0. The ChrUn super-scaffolds were created by linking adjacent unanchored scaffolds in a random order with 100 Ns. The pseudomolecules_v2.0 along with ChrUn constituted the IWGSC RefSeq v2.0 assembly.

Final refinement of the RefSeq v2.0 assembly. Primary transcripts of 133 744 HC genes (IWGSC, 2018) were mapped to RefSeq v2.0 to obtain new coordinates. Alignments of CDSs showing mis-matches and/or indels were labeled as ‘problematic’ and as needed to be corrected because of containing nucleotide errors due to the integration of PacBio contigs into the new assembly. Original contigs from the IWGSC RefSeq v1.0 harboring the ‘problematic’ CDSs were aligned against the RefSeq v2.0 using minimap2 (Li, 2018) with the default parameters and these regions in IWGSC RefSeq v2.0 were reverted to their original version (v1.0).
leading to a final genome assembly IWGSC RefSeq v2.1 (Figure 2).

TE annotation

Intact LTR RTs were identified with the LTR_finder (Xu and Wang, 2007). A non-redundant LTR library was generated. LTRs were annotated by searches with RepeatMasker 4.0.9 (http://www.repeatmasker.org) against the non-redundant LTR library, Repbase (Bao et al., 2015) and TREP (Wicker et al., 2002) databases.

To annotate all TEs, de novo TE models were generated using CLARITE and the curated wheat TE library ClariTeRep (Daron et al., 2014) as described previously (IWGSC, 2018; Wicker et al., 2018). CLARITE uses RepeatMasker for raw similarity search, to merge adjacent TE fragments into complete TE models, and eventually identifies nested insertions. The TE content similarity between v1.0 and v2.1 was compared by searching for overlaps with BEDTools intersect (Quinlan and Hall, 2010) between v1.0 ISBPs hits on v2.1 (from BAM file) and annotated v2.1 TEs (.gff3 file). Statistics and metrics for TE families were produced using bash command lines.

Gene annotation

Integration of manually curated genes (Annotation v1.2). Before performing a gene transfer onto the new genome assembly, a set of 198 genes manually curated by experts and communicated in 2020 to the IWGSC RefSeq Annotation group was integrated. A method described in IWGSC (2018) was applied to update gene annotation, which led to include an additional 117 novel genes and to correct 81 miRNAs. We generated a new release: Annotation v1.2, containing 108 010 HC and 161 535 LC gene models, which was used for gene transfer on the new assembly.

Transfer of curated genes on RefSeq v2.1. A Snakemake pipeline called MAGATT (Marker-Assisted Gene Annotation Transfer for Triticeae) was developed to transfer gene annotation from the IWGSC Annotation v1.2 onto the IWGSC RefSeq v2.1 assembly.

The pipeline performs this transfer automatically. First, 150-bp tags corresponding to ISBP markers in IWGSC RefSeq v1.0 were designed as described in De Oliveira et al. (2020), and mapped on IWGSC RefSeq v2.1 using BWA (Li and Durbin, 2010). Only ISBPs that were fully mapped with no mismatch and with a mapping quality of at least 30 (i.e. uniquely mapped) were selected. For each gene, the pipeline identified the closest 5’ and 3’ ISBPs in order to predict the coordinates of the smallest target interval harboring a query gene of IWGSC RefSeq v2.1. If there was no 5’ or 3’ ISBPs, the start/end coordinates of the chromosome were used as borders. Only intervals between 500 bp and 10 Mb were considered. When such an interval was defined, BLAT (Kent, 2002) was used to align the genomic sequence of each query gene (from transcription start site to termination site) on the extracted sequence of the target interval. If a full-length perfect match was observed (100% identity over 100% of the gene length, including Ns with BLAT option -”extendThroughN” between the two assembly versions, or if a full-length match with only mismatches was observed (but no indels), the pipeline calculated the positions of all gene-related features (mRNA, exons, CDSs and UTRs) on IWGSC RefSeq v2.1. If indels were observed (that may occur because of the gap-filling step), the pipeline ran GMAP to perform a spliced alignment of the query mRNA on the target interval and to retrieve new feature positions. Finally, if no ISBP-based interval was found to encompass a gene, MAGATT ran GMAP on the whole genome assembly. Details can be found in Table S1.

Assessing the quality of IWGSC RefSeq v2.1

The sequences of 15 CS BACs (Table S2) assembled from Sanger reads were downloaded from the NCBI. The BACs originated from chromosomes 3A, 3B, 5A and 5D. They were mapped on the Chr3A, Chr3B, Chr5A and Chr5D pseudomolecules of IWGSC RefSeq v2.1 using BLAST, and identity and coverage were recorded.

Orthologous genes and syntenic blocks

Primary transcripts of 133 744 HC genes (IWGSC, 2018) were mapped to the pseudomolecules of IWGSC RefSeq v2.1 to obtain new coordinates. The all versus all bidirectional BLAST (Altschul et al., 1990), with default parameters, was performed among protein sequences of the corresponding primary transcripts to identify putative orthologous genes. The top five hits based on identity and coverage of each HC gene were used to identify syntonic blocks among subgenomes using MCScanX (Wang et al., 2012) with default settings.

Chromatin architecture analysis (Hi-C)

Chromosome conformation capture (Hi-C) sequencing data (IWGSC, 2018) were retrieved from ENA accession PRJEB25248. Hi-C reads were mapped with the TRITEX pipeline (Monat et al., 2019). The IWGSC RefSeq v1.0 and v2.1 assemblies were digested in silico with HindIII using EMBOSS restrict (Rice et al., 2000). Reads were trimmed at junction sites (AAAGCTAGCTT) with cutadapt (Martin, 2011) and aligned to the assemblies with Minimap2 (Li, 2018). Alignment records were converted to Binary Sequence Alignment/Map (BAM) format with SAMtools (Li 2009) and sorted with Novosort (http://www.novocraft.com/products/novosort/). A list of Hi-C links was extracted from Hi-C alignments using BEDTools (Quinlan and Hall, 2010) and TRITEX scripts. A table of Hi-C links was imported to the R statistical environment (R Core Team, 2018) and contact matrices were plotted with TRITEX scripts.

ACKNOWLEDGEMENTS

The authors are grateful to the Mésocentre Clermont Auvergne University (France) and AuBi platform for providing computing and storage resources. The authors thank Dr Patrick E. McGuire for valuable suggestions and assistance with drafting the manuscript. This work was supported by the US National Science Foundation grant IOS-1929053 (JD and M-CL) and USDA Agricultural Research Service CRIS project 2030-21430-014-00-D (YOG). RDO was funded by a grant from the French Government managed by the Research National Agency (ANR) under the Investment for the Future program (BreedWheat project ANR-10-BTBR-03).

AUTHOR CONTRIBUTIONS

M-CL, JD, TZ and LW conceived the project. TZ, JCR, KRD and M-CL generated optical maps. LW performed comparative analysis of the two Cs stocks. TZ and M-CL resolved chimera, and performed super-scaffolding and pseudomolecule construction. LW and TZ closed gaps and corrected base errors. HR and FC re-annotated genes, and RDO annotated TEs. LW, GK-G, JT, RA, YOG, HR, FC, JR, KE and MM contributed to evaluation and validation of the assembly. TZ, LW, FC, RDO, RA, M-CL and JD wrote the
CONFLICT OF INTEREST
The authors declare that they have no competing interest.

DATA AVAILABILITY
The IWGSC RefSeq v2.1 assembly and annotations are available at the IWGSC data repository hosted by URGI-INRAE (https://wheat-urgi.versailles.inra.fr/Seq-Repository/Assemblies) and NCBI under project PRJNA669381. The DLS optical map is available at https://doi.org/10.25338/B8K917.

SUPPORTING INFORMATION
Additional Supporting Information may be found in the online version of this article.

REFERENCES
Alonge, M., Shumate, A., Puiu, D., Zimin, A.V. & Salzberg, S.L. (2020) Rapid genome mapping in nanochannel arrays for highly complex genomes. Nature, 579, 563–566. 
Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.
Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S.O., Gundlach, H. et al. (2018) Ultra-dense genetic map of durum wheat × wild emmer wheat developed using the 90K iSelect SNP genotyping assay. *Molecular Breeding*, 34, 1549–1562.
Bao, W., Kojima, K.K. & Kohany, O. (2015) Rebase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, 6, 11.
Brown, J., Daron, J., Glover, N., Pingault, D., Theil, S., Jamilloux, V., Faure, S., Choulet, F., Roger, D., Gauthier, V., Martinant, J.-P. et al. (2014) Intra-chromosomal Hi-C contact matrices of the IWGSC RefSeq v2.1 and IWGSC RefSeq v1.0.
Choulet, F., Leroy, P., Paux, E., Sourdille, P., Salse, J., Saintenac, C., Choulet, F., Roger, D., Gauthier, V., Martinant, J.-P. et al. (2013) Ultra-dense genetic map of durum wheat × wild emmer wheat developed using the 90K iSelect SNP genotyping assay. *Molecular Breeding*, 34, 1549–1562.
Huang, J., Dong, L., Smith, A., Finklestein, J., Lam, E.T., Huo, N. et al. (2013) Rapid genome mapping in nanochannel arrays for highly complex and accurate de novo sequence assembly of the complex Aegeilo ps tauschii genome. *PloS One*, 8, e55864.
Iken, C., Luo, M.-C., Ramasamy, R., Dawson, M., Gill, B.S., Koral, A.B. et al. (2017) A high-density genetic map of wild emmer wheat from the Karaca Dag region provides new evidence on the structure and evolution of wheat chromosomes. *Frontiers in Plant Science*, 8, 1798.
Joshi, N. & Fass, J. (2011) Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). Available from https://github.com/najoshi/sickle
Kent, W.J. (2002) BLAT—The BLAST-like alignment tool. *Genome Research*, 12, 656–664.
Li, H. (2018) Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, 34, 3994–3997.
Li, H. & Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26, 589–595.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N. et al. (2009) The sequence alignment/map format and SAMTools. *Bioinformatics*, 25, 2078–2079.
Liu, D., Zhang, L., Hao, M., Ning, S., Yuan, Z., Dai, S. et al. (2018) Wheat breeding in the hometown of Chinese Spring. *The Crop Journal*, 6, 82–90.
Luo, M.-C., Gu, Y.Q., Puiu, D., Wang, H., Twardziok, S.O., Deal, K.R. et al. (2017) Genome sequence of the progenitor of the wheat D genome Aegilops tauschii. *Nature*, 551, 498–502.
Maccarrelli, M., Harris, N.S., Twardziok, S.O., Pasam, R.K., Gundlach, H., Spannagl, M. et al. (2019) Durum wheat genome highlights past domestication signatures and future improvement targets. *Nature Genetics*, 51, 885–895.
Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17, 10–12.
Mascher, M., Gundlach, H., Himmelbach, A., Seier, S., Twardziok, S.O., Wicker, T. et al. (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, 544, 427–433.
Monat, C., Padmarasu, S., Lux, T., Wicker, T., Gundlach, H., Himmelbach, A. et al. (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biology*, 20, 284.
Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C. & Durbin, R. (2016) BCFToolsR0H: a hidden Markov model approach for detecting autoryosity from next-generation sequencing data. *Bioinformatics*, 32, 1749–1751.
Paux, E., Faure, S., Choulet, F., Roger, D., Gauthier, V., Martinan, J.-P. et al. (2010) Insertion site-based polymorphism markers open new perspectives for genome saturation and marker-assisted selection in wheat. *Plant Biotechnology Journal*, 8, 196–210.
Paux, E., Sourdille, P., Salje, S., Sainthac, C., Choulet, F., Leroy, P. et al. (2008) A physical map of the 1-gigabase wheat chromosome 3B. *Science*, 322, 101–104.
Quinlan, A.R. & Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.
R Core Team. (2018) R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.
Ray, D.K., Mueller, N.D., West, P.C. & Foley, J.A. (2013) Yield trends are insufficient to double global crop production by 2050. *PLoS One*, 8, e66429.
Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16, 276–277.
Rimbeth, H., Barlier, B., Navarro, J., Kott, J., Choulet, F., Leveugle, M. et al. (2018) High throughput SNP discovery and genotyping in hexaploid wheat. *PloS One*, 13, e0186329.
Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelkelie, A., Sakhitkumar, S. et al. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9, e112963.
Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X. et al. (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research*, 40, e49.
Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E. et al. (2014) Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, 12, 787–796.

Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-González, R.H. et al. (2018) Impact of transposable elements on genome structure and evolution in bread wheat. *Genome biology*, 19, 103.

Wicker, T., Matthews, D.E. & Keller, B. (2002) TREP: a database for Triticeae repetitive elements. *Trends in Plant Science*, 7, 561–562.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B. et al. (2007) A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8, 973–982.

Xu, Z. & Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35, W265–W268.

Zhu, T., Wang, L., Rodriguez, J.C., Deal, K.R., Avni, R., Distelfeld, A. et al. (2019a) Improved genome sequence of wild emmer wheat Zavitan with the aid of optical maps. *G3: Genes, Genomes, Genetics*, 9, 619–624.

Zhu, T., Wang, L., You, F.M., Rodriguez, J.C., Deal, K.R., Chen, L. et al. (2019b) Sequencing a *Juglans regia × J. microcarpa* hybrid yields high-quality genome assemblies of parental species. *Horticulture Research*, 6, 55.

Zimin, A.V., Puiu, D., Hall, R., Kingan, S., Clavijo, B.J. & Salzberg, S.L. (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. *GigaScience*, 6, 1–7.