Genome assemblies

a) Introduction to the wheat genome

Wheat is an allopolyploid of which there are two major types:

- Hexaploid common wheat (*Triticum aestivum* ssp. *aestivum;* 17 Gb genome size; AABBDD genomes), which is mainly used for bread and biscuit products.
- Tetraploid durum wheat (*Triticum turgidum* ssp. *durum*; 12 Gb genome size; AABB genomes), used mainly for pasta.

Hexaploid wheat arose from a polyploidisation event ~ 10,000 years ago, whereas tetraploid wheat arose ~ 400,000 years ago (Figure 1).

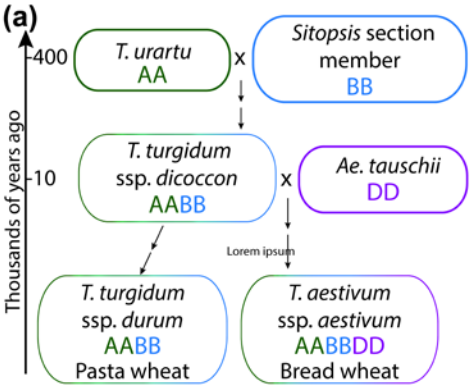

Figure 1: The evolutionary history of allopolyploid wheat. Modified from Borrill *et al.*, 2015. DOI: 10.1111/nph.13533

Hexaploid wheat contains three closely related genomes (A, B and D) which contain homoeologous genes in a conserved order. Wheat homoeologs share over 95 % sequence identity within coding regions and most wheat genes are expected to be present as three copies in the A, B and D genome. Due to the high sequence conservation between homoeologs, genes may be functionally redundant or act in a dose dependent manner. This means that often all three copies must be knocked out to cause a strong phenotype. However, in other cases the homoeologous genes have developed specialised functions or become pseudogenized over time due to reduced selection pressure on duplicated genes.

b) Multiple genome assemblies have been released

For a long time, the large and complex genome of wheat had confounded efforts to create a high-quality genome assembly. Technological advances though have prompted the release of many genome assemblies over a short period of time. While most of these assemblies are based on the landrace Chinese Spring, other cultivars and species have also been recently sequenced.
Here, we will describe all publicly available wheat genome assemblies not just to provide context but also because tools were developed for several of these assemblies.

c) Overview of assemblies (ordered by latest release)

10+ Wheat Genomes Project (Hexaploid bread wheat cultivars)

An international consortium led by C. Pozniak (Univ. of Saskatchewan) has generated chromosome-level and scaffold-level assemblies for multiple wheat cultivars representing global wheat diversity. The includes cultivars from Australia, Canada, Germany, Japan, Switzerland, the USA, the UK and CIMMYT. The majority of these cultivars have been assembled to a similar standard as the reference Chinese Spring genome (RefSeq). Two different algorithms have been used to generate these assemblies, either the NRGene DeNovoMagic or W2RAP. These assemblies will be integrated into Ensembl Plants and are available for download under Toronto Agreement.

10+ Wheat Genomes Project reference: http://www.10wheatgenomes.com/
10+ Wheat Genomes Project data access: https://wheatis.tgac.ac.uk/grassroots-portal/blast
http://webblast.ipk-gatersleben.de/wheat_ten_genomes/

Triticum 4.0 (Hexaploid bread wheat)

A new (in 2020) whole-genome assembly and comprehensive annotation of the "Chinese Spring" cultivar of bread wheat. The assembly is based on the previously published Triticum 3.1 assembly and the IWGSC v1.0 assembly, which were merged using the RaGOO software followed by a variety of custom programs. The combination of the two previous assemblies yielded a chromosome-scale assembly with a total length of 15,397,713,314 bp, of which 98% (15,070,919,678) is mapped onto chromosomes. The Triticum 4.0 assembly is both larger and more contiguous than any previous assembly.

Gene annotation used the Liftoff software (https://github.com/agshumate/Liftoff) to map all high-confidence genes from the IWGSC v1.0 assembly onto Triticum 4.0. The vast majority of genes (100,839 out of 105,200) mapped successfully. Notably, 5,799 new genes were identified, all of them identical or near-identical copies of previously annotated genes. In addition, of the 2,691 genes that were on unplaced scaffolds in IWGSC v1.0, 2,001 were placed onto chromosomes in Triticum 4.0.

Triticum 4.0 reference: Alonge *et al.*, 2020 DOI: https://doi.org/10.1101/2020.04.06.028746
Triticum 4.0 data access: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA392179
Triticum 4.0 assembly: https://www.ncbi.nlm.nih.gov/assembly/GCA_002220415.3

Svevo v1.0 (Tetraploid durum wheat)

Using the proprietary DeNovoMagic (NRGene) algorithm, as well as optical mapping and Hi-C conformation capture data, a high-quality genome assembly of the durum wheat (*Triticum turgidum*) cultivar 'Svevo' was generated. This 10.46 Gb assembly consists of 14 pseudomolecule chromosomes (9.96 Gb) and a molecule with all unassigned scaffolds (0.499 Gb). The assembly was annotated with 66,559 high confidence and 303,404 low confidence gene models. All data has been incorporated into the Ensembl Plants database.

Svevo v1.0 reference: Maccaferri *et al.,* 2019 DOI: 10.1038/s41588-019-0381-3
Svevo v1.0 data access: http://plants.ensembl.org/Triticum_turgidum/Info/Index
http://www.interomics.eu/durum-wheat-genome

RefSeqv1.0 (Hexaploid bread wheat)

A whole genome assembly has been produced by the IWGSC; Illumina sequencing data was assembled into a 14.5 Gb draft genome using a proprietary algorithm, DeNovoMagic (NRGene). This assembly has much larger contigs than previous assemblies (N50 super-scaffold length 22.8 Mb) and equals genome assemblies of rice and other model species in terms of quality and contiguity. The sequence scaffolds have been ordered using POPSEQ data and Hi-C (chromosome conformation capture) to generate 21 pseudomolecules representing the majority of the wheat genome. Another pseudomolecule, termed chromosome U, contains all sequences that could not be assigned to one of the 21 wheat chromosomes. Gene models have been generated consisting of 110,790 high confidence genes (homology to genes in other species) and 158,793 low confidence genes (e.g. truncated genes, genes lacking transcriptional evidence or lacking homology to other species). This genome assembly represents the current gold standard and is used as the reference genome by most researchers. All data has been integrated into the Ensembl Plants database.

RefSeqv1.0 reference: IWGSC 2018, DOI: 10.1126/science.aar7191
RefSeqv1.0 data access: https://plants.ensembl.org/Triticum_aestivum/Info/Index
https://urgi.versailles.inra.fr/blast_iwgsc/blast.php

Triticum 3.1 (Hexaploid bread wheat)

A whole genome shotgun sequence assembly of Chinese Spring, which was assembled using short Illumina and long PacBio reads: the assembly was performed in several, iterative steps using the MaSuRCA and Celera Assembler software.

The combination of very long reads (average read length ~10 kb) coupled with deep sequencing of low error-rate short reads (65x coverage) produced an assembly with a total length of 15.3 Gb represented by 279,439 contigs with an N50 of 232,659 bp and average contig size of 54,912 bp. The Triticum 3.1 assembly is highly contiguous and does not contain unknown nucleotides (Ns).

It was aligned to an existing assembly of *Aegilops tauschii* (the D-genome progenitor of hexaploid wheat) to identify its D-genome portion; this data was saved into a separate assembly called TriticumD 1.0.

There is currently no annotation available for the Triticum 3.1 assembly.

Triticum 3.1 reference: Zimin *et al.*, 2017 DOI: 10.1093/gigascience/gix097
Triticum 3.1 data access: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA392179
TriticumD 1.0 data access: ftp://ftp.ccb.jhu.edu/pub/data/Triticum_aestivum/Wheat_D_genome/

Zavitan WEWSeq v.1.0 (Wild Emmer genome assembly)

A whole genome sequencing approach was used to assemble the tetraploid wild emmer genome of accession “Zavitan”. Short Illumina reads as well as mate pair libraries were used as input for the propriety DeNovoMagic algorithm (NRGene), which resulted in a 10.5 Gb assembly. Using Hi-C (chromosome conformation capture) the assembled scaffolds could be further assembled into 14 pseudomolecule chromosomes. Another pseudomolecule, termed chromosome U, contains all sequences that could not be assigned to one of the 14 wild emmer chromosomes. Gene models were developed based on models from other grass species and transcriptome data, resulting in 65,012 high confidence (HC) genes and 45,532 low confidence (LC) genes. All data has been integrated into the Ensembl Plants database.

WEWSeq v.1.0 reference: Avni *et al.*, 2017 DOI: 10.1126/science.aan0032
WEWSeq v.1.0 access: http://plants.ensembl.org/Triticum_dicoccoides/Info/Index
http://wewseq.wixsite.com/consortium/svevo-x-zavitan-population

TGACv1 (Hexaploid bread wheat)

A whole genome shotgun sequence assembly of Chinese Spring was produced using short Illumina reads, nested long mate-pair libraries and the W2RAP pipeline. This method created an assembly of total length 13.4 Gb, with approximately 10x longer N50 than the CSS and W7984 assemblies. Gene models from IWGSC were projected onto the TGACv1 assembly, with 99 % of the total genes located on the TGACv1 assembly. *De novo* gene prediction has been carried out for the TGACv1 assembly resulting in a total of 273,739 transcripts (including non-coding and transcript variants). Of these 104,305 are high confidence protein coding genes. In general, these gene models are more complete than the CSS gene models due to the longer contig length within the TGACv1 assembly. All data from the TGACv1 assembly is available on the Ensembl Plants archive page.

TGACv1 reference: Clavijo *et al.*, 2017 DOI: 10.1101/gr.217117.116
TGACv1 data access: http://oct2017-lants.ensembl.org/index.html
W2RAP pipeline: Clavijo *et al.,* 2017 DOI: 10.1101/110999

W7984 (Synthetic hexaploid bread wheat)

A whole genome sequencing approach was undertaken in the synthetic hexaploid wheat “Synthetic W7984”. This approach used large-insert sequencing libraries and enabled separate assemblies of the three homoeologous genomes to create an assembly of 9.1 Gb. Some scaffolds from the W7984 assembly are more contiguous than their counterparts in the CSS assembly, but the opposite can also be true. Using both references gave the most complete picture of genomic regions of interest. No gene models were predicted for the W7984 assembly.

W7984 reference: Chapman *et al.*, 2015 DOI: 10.1186/s13059-015-0582-8
W7984 data access: http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/blast_WGS.php

https://wheatis.tgac.ac.uk/grassroots-portal/blast

Chinese Spring Survey Sequence (CSS) (Hexaploid bread wheat)

The International Wheat Genome Sequencing Consortium (IWGSC) used a chromosome flow-sorting approach to separate individual chromosome arms. The landrace used (Chinese Spring) had aneuploid genetic stocks available; in each line one arm of a different chromosome e.g. the short arm of chromosome 1A, is deleted. This allows for chromosome arm 1AL to be separated from the other chromosomes, before sequencing. Individual chromosome arms were sequenced to 30-240x coverage using Illumina NGS, generating a 10.2 Gb assembly of Chinese Spring (termed Chinese Spring survey sequence (CSS)).

Gene models were created by mapping RNA-seq data to the assembly and using gene models from related species as guides and quality controls. A total of 99,386 protein-coding genes were predicted, with 193,667 transcripts and splice variants. The accuracy of these gene models relies on the assembly quality; in some cases, gene models are incomplete because the underlying genomic scaffolds are too short. For example, a gene may be split between two different scaffolds, which would result in two gene models (with two separate gene model names) being predicted for each half of the gene (shown below). Alternatively, only one incomplete gene model could be predicted on one of the scaffolds.

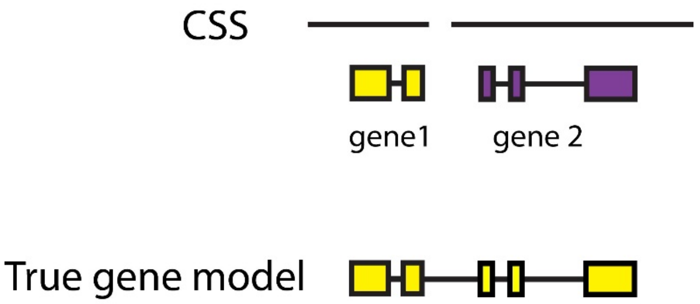

Figure 2: CSS gene models are affected by truncated scaffolds.

Chromosome 3B of Chinese Spring was assembled using a BAC by BAC approach and was considered the “gold standard” for a reference genome assembly. 3B gene models were created independently of the CSS assembly using RNA-seq data and follow their own nomenclature.

CSS reference: IWGSC 2014, DOI: 10.1126/science.1251788
CSS data access: http://archive.plants.ensembl.org/Triticum_aestivum/Info/Index

d)Comparison between assemblies

The following table gives an overview of the various wheat genome assemblies released in recent years. It highlights the rapid changes in technology that allowed ever better assemblies to be built, but also the difficulty for researchers to keep up with these changes.

Table 1. Comparison between different genome assemblies (ordered by release date).

Common name	Variety	Assembly size	Gene models	Related resources	Description	References
CSS	Chinese Spring (6×)	10.2 Gb total N50$_{contig}$ = 4.2 kb	✓	TILLING mutants Expression browsers SNP markers	Flow-sorted chromosome arms sequenced and assembled individually, allowing efficient separation of homoeologous sequences	IWGSC (2014)
W7984	Synthetic W7984 (6×)	9.1 Gb total N50$_{contig}$ = 6.7 kb			WGS assembly combining paired-end and mate-pair libraries.	Chapman *et al*. (2015)
TGAC	Chinese Spring (6x)	13.4 Gb total N50$_{contig}$ = 88 kb	✓	TILLING mutants Expression browser SNP markers	WGS assembly combining paired-end and mate-pair libraries with the W2RAP contigger.	Clavijo *et al*. (2017)
Triticum 3.1	Chinese Spring (6x)	15.3 Gb total N50$_{contig}$ = 232 kb			WGS assembly combining short Illumina and long PacBio reads.	Zimin *et al*. (2017)
Triticum 4.0	Chinese Spring (6x)	15.07 Gb total N50$_{contig}$ = 230 kb	✓		WGS assembly combining short Illumina and long PacBio reads. Missing sequences from RefSeqv1.0 were added using RaGOO.	Alonge *et al*. (2020)

RefSeqv1.0	Chinese Spring (6x)	14.5 Gb total pseudomolecule chromosomes	✓	TILLING mutants Expression browsers SNP markers	WGS assembly incorporating Hi-C and optical sequencing information. The propriety DeNovoMagic assembler was used.	IWGSC (2018)
Svevo v1.0	Svevo (4x)	10.46 Gb total pseudomolecule chromosomes	✓	SNP markers	WGS assembly incorporating Hi-C and optical sequencing information. The propriety DeNovoMagic assembler was used.	Maccaferri *et al.* (2019)
10+ Wheat Genomes Project	Mace Lancer CDC Landmark Julius Norin61 Arina Jagger Cadenza Paragon Kronos Robigus Claire				DeNovoMagic or W2RAP	