DATA ACCESS DISCLAIMER:
These data are in open access. While scientists may freely publish using the IWGSC data, IWGSC does request that the source of the data be properly acknowledged.
For questions regarding data access and publication policy please contact IWGSC (communications@wheatgenome.org).

PLEASE NOTE:
MD5 checksums are provided to ensure complete data transfer of files, please make sure that the MD5 checksum of your copy corresponds to the one provided with the files.

QUESTIONS
In case of questions please contact Michael Alaux from INRA-URGI in the first instance at: urgi-support@inrae.fr

CREDITS:
This data set has been generated as a collaborative effort between INRA-GDEC Clermont Ferrand, France (INRA) (Frederic Choulet, Philippe Leroy, Helene Rimbert), Plant Genome and Systems Biology, Helmholtz Zentrum, Munich, Germany (PGSB) (Klaus Mayer, Manuel Spannagl, Sven Twardziok, Heidrun Gundlach) and Earlham Institute, Norwich, UK (EI) (David Swarbreck, Luca Venturini, Gemy Kaithakottil) under the coordination of IWGSC (International Wheat Genome Sequencing Consortium) (Kellye Eversole, Jane Rogers).

Gene Annotation

GENE PREDICTION VERSION:
The gene structures in this annotation are based on the bread wheat genome assembly version **IWGSC RefSeq v1.0.** This annotation is now referred to as IWGSC RefSeq Annotation v1.0. Gene structures have been called using a multi-step process, involving both INRA and PGSB running established pipelines for initial gene prediction. The two sets of gene models were evaluated by EI and combined to produce an integrated, revised gene set. The gene prediction pipelines used evidence from multiple publicly available data sets including transcriptome data (RNAseq and IsoSeq data), flCDNAs and reference proteins, applying criteria specified in each of the pipelines. Full details about the gene prediction processes, the data sets used as evidence and the gene model integration process will be included in the IWGSC publication describing the wheat reference genome assembly and initial analyses.

PROTEIN HOMOLOGY-BASED GENE MODEL CLASSIFICATION:
Predictions of gene products derived from the integrated set of gene models were used to classify the models as HC (High-confidence) or LC (Low-confidence) genes, based on their sequence homology to genes in public databases and/or sequences in TREP (transposon database).

HC genes have complete gene models with very good sequence homology to experimentally verified plant proteins from SwissProt (HC1) OR with very good sequence homology to any

annotated (also automatically) poaceae protein in SwissProt or trEMBL but NO good hits in the transposon database TREP (HC2).
“Complete” gene models are defined as containing both start and stop codons. Please note that these are not necessarily the biologically “true” start and stop sites.

LC genes have incomplete gene models with very good sequence homology to experimentally verified plant proteins from SwissProt (LC1), complete gene models with no significant (see definition below) homology to any of the three databases (LC2) OR incomplete gene models with very good sequence homology to any annotated (also automatically) poaceae protein in SwissProt or trEMBL but NO good hits in the TE database TREP (LC1). Incomplete gene models with no significant (see definition below) homology to any of the three databases were classified into LC3. Please note that gene models have not been assigned to a pseudogene category in this annotation version, but are likely represented as LC1 and LC3 genes.

TREP genes have no significant (see definition below) homology to experimentally verified plant proteins from SwissProt BUT good homology to TREP entries.

Details of the gene classification.

Databases used in the classification process:
TREP: PTREP 17, a collection of *Triticeae* transposon proteins downloaded from http://botserv2.uzh.ch/kelldata/trep-db/index.html was used to identify and tag transposon-derived gene models..
UniPoa: Annotated poaceae proteins (SwissProt & trEMBL). Sequences were downloaded from Uniprot (Feb 2017) and further filtered for complete sequences with start and stop codons.
UniMag: Validated magnoliophyta proteins ***(SwissProt)***. Sequences were downloaded from Uniprot (Feb 2017) and further filtered for complete sequences with start and stop codons.
For each transcript in the initial data set the sequence homology to each of the three databases was determined (BLASTP, e-value cutoff 10-10) and classified by the query coverage value of the best hit with a 90%, or in the case of TREP a 75% overlap criterion
Step 1: A transcript was assigned to a reference database if the overlap between query and subject reached >90% for the two protein databases or >75% query coverage for the TREP database.
Step 2: classification into distinct confidence groups and sub-groups according to the following rules:

```
# Primary confidence class
unimag & complete <- "HC"
!unimag & (!trep & unipoa) & complete <- "HC"
(unimag | (!trep & unipoa)) & !complete <- "LC"
!unimag & !trep & !unipoa & complete <- "LC"
!(unimag & complete) & trep <- "TREP"
# Secondary confidence class
unimag & complete <- "HC1"
!unimag & (!trep & unipoa) & complete <- "HC2"
(unimag | (!trep & unipoa)) & !complete <- "LC1"
!unimag & !trep & !unipoa & complete <- "LC2"
!unimag & !trep & !unipoa & !complete <- "LC3"
!(unimag & complete) & trep <- "TREP"
```

The confidence group of each transcript is given in the respective GFF/GTF file. Please note that transcripts with different confidence classes can be present at a single gene locus. In these cases, the “best” confidence label of all transcripts present was assigned to the respective locus.

GENE IDs:

At each gene locus the splice variant with the longest coding sequence (CDS) has been selected as the representative transcript.
Gene IDs were assigned following the schema and rules below:

TraesCS3B01G207000.1 where

Traes = Triticum aestivum;

CS = Chinese Spring. Other cultivar names will normally be represented by three characters.

3B01 - version 01 of the 3B assembly and/or annotation - this makes allowance for the fact that the sequence assembly and gene models will be updated over time and the version number will make it clear for users which version is being referred to in analyses; “U” designates gene structures on chromosome “U” (unknown) sequences i.e. scaffolds that could not be assigned to a chromosome arm.

G indicates "gene";

207000 = increasing number in steps of “100” along the respective chromosome arm;

.1 = splice variant "1";

FILES INCLUDED IN RELEASE:
This data set contains files in different formats, providing access to the gene calls either as structural predictions (GTF/GFF files) or sequences (fasta files).
The files contained in this release are:

iwgsc_refseqv1.0_HighConf_2017Mar13.gff3.zip

Contains the HC (High Confidence) structural gene predictions on the IWGSC RefSeq v1.0 genome assembly in GFF3 format. Corresponding genome sequence files are available from https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.0/

iwgsc_refseqv1.0_HighConf_CDS_2017Mar13.fa.zip

Contains the coding (CDS) sequence of all high-confidence (HC) gene models that have been positioned on the IWGSC sequence assembly, including splice variants.

iwgsc_refseqv1.0_HighConf_PROTEIN_2017Mar13.fa.zip

Contains the protein sequence of all high-confidence (HC) gene models that have been positioned on the IWGSC sequence assembly, including splice variants.

iwgsc_refseqv1.0_HighConf_REPR_CDS_2017Apr03.fa.zip

Contains the coding (CDS) sequence of all high-confidence (HC) gene models that have been positioned on the IWGSC sequence assembly, without splice variants. Only the representative gene model per locus is given.

iwgsc_refseqv1.0_HighConf_REPR_PROTEIN_2017Apr03.fa.zip

Contains the protein sequence of all high-confidence (HC) gene models that have been positioned on the IWGSC sequence assembly, without splice variants. Only the representative gene model per locus is given.

iwgsc_refseqv1.0_HighConf_UTR_2017May05.gff3.zip

Contains the Untranslated Transcribed Region (UTR) assigned to high-confidence (HC) genes in GFF3 format.

iwgsc_refseqv1.0_LowConf_2017Mar13.gff3.zip

Contains the LC (Low Confidence) structural gene predictions on the IWGSC RefSeq v1.0 genome assembly in GFF3 format. Corresponding genome sequence files are available from https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.0/

iwgsc_refseqv1.0_LowConf_CDS_2017Mar13.fa.zip

Contains the coding (CDS) sequence of all low-confidence (LC) gene models that have been positioned on the IWGSC sequence assembly, including splice variants.

iwgsc_refseqv1.0_LowConf_PROTEIN_2017Mar13.fa.zip

Contains the protein sequence of all low-confidence (LC) gene models that have been positioned on the IWGSC sequence assembly, including splice variants.

iwgsc_refseqv1.0_LowConf_REPR_CDS_2017Apr03.fa.zip

Contains the coding (CDS) sequence of all low-confidence (LC) gene models that have been positioned on the IWGSC sequence assembly, without splice variants. Only the representative gene model per locus is given

iwgsc_refseqv1.0_LowConf_REPR_PROTEIN_2017Apr03.fa.zip

Contains the protein sequence of all low-confidence (LC) gene models that have been positioned on the IWGSC sequence assembly, without splice variants. Only the representative gene model per locus is given..

iwgsc_refseqv1.0_LowConf_UTR_2017May05.gff3.zip

Contains the Untranslated Transcribed Region (UTR) assigned to low-confidence (LC) genes in GFF3 format.

iwgsc_refseqv1.0_PGSB_annotation_files.zip

Countains the syntenic gene pairs file.

Transposable elements and markers

iwgsc_refseqv1.0_TransposableElements_2017Mar13.gff3.zip

Contains the Transposable Elements (TEs) predicted by the CLARITE software on the IWGSC RefSeq v1.0 genome assembly in GFF3 format. Corresponding genome sequence files are available from
https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.0/

iwgsc_refseqv1.0_Marker_mapping_summary_2017Mar13.zip

Contains the markers mapped on the IWGSC RefSeq v1.0 genome assembly in GFF3 format. Corresponding genome sequence files are available from
https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Assemblies/v1.0/

New file added in February 2019:
280k public markers from the BreedWheat project (TaBW280k.gff3).

ISBPs

Frederic Choulet designed ISBPs on CSS contigs (filtering out 3B/1B) + 3B pseudomolecule (version 2014) + 1B scaffolds (BAC-based assembly) in order to estimate the completeness of the RefSeq_v1.0 (pseudomolecules+unplaced scaffolds).

There were 4773481 ISBPs (without Ns) in this sample.
- 4607897 mapped over their full length and with 0 mismatch => 97% of the sample
- 4512979 mapped uniquely => 98% of the mapped ISBPs
- 94918 mapped at multiple locations => 2% of the mapped ISBPs

Functional annotation of genes

Gene functions have been assigned using the AHRD tool (Automated Assignment of Human Readable Descriptions, https://github.com/groupschoof/AHRD, version 3.3.3). AHRD scores and ranks blast hits taken from searches against different databases on the basis of the knowledge of the databases and the alignment quality. The blast hit descriptions are tokenized into informative words and a lexical analysis scores these tokens according to their frequency and quality. Finally the best scoring description is assigned. Table 1 shows a summary of functional terms used in the AHRD analysis.

A postprocessing filter step (disTEG: distinction between TEs and Genes) on the AHRD and domain descriptions identified ~3,300 remaining transposons (TEs) within the high confidence (HC) gene set and tagged the genes as either *G* for canonical genes with a nonTE description , *TE* for obvious transposons, *TE?* for potential transposons (having domains that occur in true genes as well as in transposons, e.g. zinc fingers) or *U* for unknowns, without any description. Table 1 explains the header fields of the provided functional annotation files. Separate files are allocated for all HC and LC protein-coding genes:

iwgsc_refseqv1.0_FunctionalAnnotation_v1__HCgenes_v1.0.TAB and *iwgsc_refseqv1.0_FunctionalAnnotation_v1__LCgenes_v1.0.TAB*. For HC genes an additional file - *iwgsc_refseqv1.0_FunctionalAnnotation_v1__HCgenes_v1.0-repr.TEcleaned.TAB* - with only the representative splice variant per locus and without obvious TEs was extracted as the most suited input for functional studies.

AHRD term	Explanation
Gene-ID	Gene identifier
is_repr	'1' if the transcript is the representative splice variant (longest cds), else '0'
AHRD-Quality-Code	a three character string, where each character is either '*' if the respective criteria is met or '-' otherwise, position 1; Bit score of the blast result is >50 and e-value is <e-10 , position 2: Overlap of the blast result is >60%, position 3: Top token score of assigned HRD is >0.5, ~ 70% are high confident *** assignments, --*,--- should be treated as low confidence.
Blast-Hit-Accession	accession of the protein the assigned description was taken from
Human-Readable-Description	from AHRD pipeline
Pfam-IDs-(Description)	Pfam accessions from InterProScan, with description for manual inspection
Interpro-IDs-(Description)	InterPro accessions from InterProScan, with description for manual inspection
GO-IDs-(Description)-via-Interpro	gene ontology IDs from InterProScan, with description for manual inspection
Gene-or-TE-TE?-U (via-function)	one of 4 tags derived from the AHRD and domain descriptions G: canonical gene, TE: transposon, TE?: potential transposons, include both canonical genes and TEs, e.g. zinc finger domains, RNAseH, helicase, needs further curation to disentangle U: unknown
Pfam-IDs	from InterProScan, only accessions for parsing
Interpro-IDs	from InterProScan, only accessions for parsing
GO-IDs-via-Interpro	from InterProScan, only accessions for parsing

Table 1. Headers used in functional annotation files

nonconding RNAs

CREDIT: Hikmet Budak.

FILES INCLUDED:

iwgsc_refseqv1.0_lncRNA_2017Dec13.gff3.zip

iwgsc_refseqv1.0_miRNA_2017Oct27.gff3.zip (new version using whole miRNAs ~9000 in miRBASE)
iwgsc_refseqv1.0_miRNA_mature-miRNA_sequences_2017May26.fsa.zip
iwgsc_refseqv1.0_miRNA_pre-miRNA_sequences_2017May26.fsa.zip

New file added in February 2019:
iwgsc_refseqv1.0_tRNA.zip
including the gff3 and the output of the tRNAscan software.

Alignment with other assemblies

CSS assembly

CREDITS: Hélène Rimbert, Frédéric Choulet

New file added in February 2023:
iwgsc_refseqv1.0_vs_CSS2014.zip
including the mapping, log and stats

TGACv1 assembly

CREDITS: Luca Venturini, Gemy Kaithakottil, David Swarbreck.

In order to compare the TGACv1 annotation (Clavijo and Venturini et al., 2017) to the new IWGSC RefSeq v1.0, the TGACv1 models were aligned against the IWGSC RefSeq v1.0 genome using GMAP v2016-09-23. The alignment was performed twice, once with stringent filtering on the terminal introns:

--min-intronlength=20 --max-intronlength-middle=50000 --max-intronlength-ends=5000 --nthreads=8 --format=gff3_gene --npaths=1 --ordered --min-identity=0.95

and once without such stringent filtering, using the default parameters:

--min-intronlength=20 --max-intronlength-middle=200000 --max-intronlength-ends=10000 --nthreads=8 --format=gff3_gene --npaths=1 --ordered --min-identity=0.95

Alignments were then filtered to retain only those whose trimmed coverage was at least 80% of the model length. Alignments were preferentially retained from the alignment with stringent filtering; alignments from the lenient run were retained only for models for which it was impossible to find a satisfactory mapping in the stringent run. Only the best match was retained for each model.

This procedure led to obtaining an alignment for 268,583 models out of 273,739 (98.12%).

It is important to note that these alignments should not be considered the best possible lift for the annotation onto the new reference genome.
No check has been performed on the congruence of the structure between the original annotation and the one present in the new alignment; moreover, as the coverage threshold is set at 80%, some

models will have been mapped only partially. As each model has been aligned independently, it is conceivable that alternative splicing events from the same gene might have been assigned to different genomic locations; or conversely, that sequences from different genomic loci in TGACv1 might have collapsed onto the same genomic locus on the IWGSC RefSeq v1.0 reference genome. Finally, CDS coordinates have not been rederived for the aligned models, so at the moment it is not yet possible to confirm the fidelity of the protein sequence of lifted models. It is therefore not possible, as of yet, to perform a comparison only on the coding fraction of sequences.

The coordinates of the lifted models have been compared against the IWGSC RefSeq v1.0 official release using Mikado compare (https://mikado.readthedocs.io/en/latest/Usage/Compare.html). We direct to the detailed documentation online for an explanation of the file formats. The comparison produced the following three files:

- iwgsc_refseqv1.0_vs_TGACv1.tmap: this file reports the best match in the reference IWGSC RefSeq v1.0 annotation for all TGACv1 alignments. Models which are found to span more than one IWGSC1 model are classified as "fusions" and therefore will be reported over multiple lines, one for each of the matched reference IWGSC1.0 models.
- iwgsc_refseqv1.0_vs_TGACv1.refmap: this file reports, for each reference IWGSC RefSeq v1.0 annotation, the best match among the TGACv1 annotations.
- iwgsc_refseqv1.0_vs_TGACv1.stats: this file reports a summary of the analysis. For this comparison, it reports that ~1.2% of TGACv1 transcripts and IWGSC RefSeq v1.0 transcripts perfectly agree with each other, and that the congruence increases if we consider as perfect matches transcripts whose structure is conserved but whose nucleotide accuracy is over 80% - including therefore models for which the terminal ends are different.

GBS maps

CREDITS: Poland and Muehlbauer labs, postdocs: Liangliang Gao (Poland lab) and Juan Gutierrez-Gonzalez (Muehlbauer lab).

Funding: NSF “GPF-PG: Genome structure and diversity of wheat and its wild relatives” Award Number: 1339389

Three mapping populations were genotyped with high density, sequence-based markers and genetic maps were created and used for anchoring, ordering and correcting scaffold positions in the v0.4, v0.5 and v1.0 pseudomolecules. All three populations were developed from crosses between the parental lines Synthetic W7984 (M6) and Opata M85, and are described in detail in Sorrells et al., Genome 2011. These populations are: SynOpDH88, a population of double haploids of 88 individuals, and 2 populations of RILs: SynOpRIL173 and SynOpRIL993, of 173 and 993 individuals, respectively. The three linkage maps constructed were used to validate the order and to assist in correcting inconsistencies. All three maps were constructed using GBS-SNP-markers, using the reference CS sequence: 160509_Chinese_Spring.

New files included in February 2019:
2 files with the SNP calls used for the RIL (993 individuals) and DH (88 individuals) populations for the quality control of the wheat genome sequence:

- SynOpDH88_ChineseSpring_v0.5.genotype.calls
- SynOpRIL993_ChineseSpring_v0.5.genotype.calls

Optical maps

CREDITS:
The optical maps were generated from wheat chromosome arms flow-sorted at the Institute of Experimental Botany (IEB), Olomouc, Czech Republic. Data were collected and optical maps were assembled at Bionano Genomics, San Diego, USA (Saki Chan, Alex Hastie – chromosome arm 7DS) and IEB (chromosome arms 7AS/L, 7BS/L and 7DL). The IEB team included Helena Toegelová, Hana Šimková, Jan Vrána, Jarmila Číhalíková, Marie Kubaláková and Jaroslav Doležel. Construction of optical maps from 7A and 7B chromosomes was co-funded by Rudi Appels (La Trobe University, Bundoora, Australia) and Odd-Arne Olsen (Norwegian University of Life Sciences, Ås, Norway), respectively.

PROJECT DESCRIPTION:
Bionano optical maps of chromosome arms 7AS, 7AL, 7BS, 7BL, 7DS and 7DL were constructed following the protocol of Staňková et al., Plant Biotech. J. 14:1523, 2016. A total of 2.8 million copies of each telosome, corresponding to 2.0 – 3.1 µg DNA, were purified by flow cytometric sorting from respective ditelosomic lines cv. Chinese Spring with purities ranging from 80 to 87%. DNA from each chromosome arm was nicked with Nt.BspQI (GCTCTTC recognition site), the nick sites were labelled and the DNA was stained with IrysPrep® Reagent Kit and IrysPrep® DNA Stain, respectively (Bionano Genomics, San Diego, USA). The labelled molecules were analyzed on the Irys platform (Bionano Genomics). A total of 78 to 248 Gb data > 150 kb was collected per chromosome arm, corresponding to 192 - 689 arm equivalents, respectively, and was used to assemble optical maps de novo using pairwise comparison of all single molecules and graph building in IrysSolve software (Bionano Genomics). A p-value threshold of 10e-10 was used during the pairwise assembly, 10e-11 for extension and refinement steps, and 10e-15 for a final refinement.

OPTICAL MAP FILES:

Chromosome arm	File name
7AS	7AS_optical_map.cmap
7AL	7AL_optical_map.cmap
7BS	7BS_optical_map.cmap
7BL	7BL_optical_map.cmap
7DS	7DS_optical_map.cmap
7DL	7DL_optical_map.cmap

RH map

CREDIT: Vijay Tiwari, Jesse Poland, Etienne Paux.

FILES INCLUDED:
RH_MAPS_FINAL_2017.xlsx
SEQUENCES_Marker.xlsx

Varietal SNP

CREDIT: Jorge Dubcovsky, Hans Vasquez-Gross, Eduard Akhunov.

FILES INCLUDED:
iwgsc_refseqv1.0_varietal_SNP_vcf.zip

Varietal SNP tracks generated by the Akhunov lab and aligned to the IWGSC RefSeq v1.0 by the Dubcovsky lab are available.

Recombination rate analysis

CREDIT: Etienne Paux, Frédéric Choulet.

FILES INCLUDED:
iwgsc_refseqv1.0_mapping_data.txt
Physical and genetic mapping data for the 47K markers that have been used in this study (we selected a set of highly reliable SNPs)

iwgsc_refseqv1.0_recombination_rate.txt
The recombination rate in sliding windows of 10 Mb (step 1Mb) for each of the 21 chromosomes.

1000 wheat exomes

CREDIT: Eduard Akhunov.

PROJECT:
About 1000 wheat exomes project
Genome-level DNA sequence variation map is required to establish links between causal variants and phenotypes as well as to understand the role of environmental, demographic and human-driven factors in shaping the genomic diversity of modern wheat. Here, we used a reference wheat genome IWGSC RefSeq v1.0 to generate a haplotype map based on the targeted re-sequencing of more than 1,000 diverse wheat landraces and cultivars, and tetraploid wild and domesticated relatives.

1000 wheat exomes project data is described in ***He et al., Nature Genetics, 2019*** **pdf**

Web page: http://wheatgenomics.plantpath.ksu.edu/1000EC/

FILES INCLUDED:
PassportData_160809.xlsx
Wheat diversity panel metadata

1kEC_genotype01222019.vcf.gz
VCF file (before imputation): # accessions = 1026, # of SNP sites ~8.87 million

all.GP08_mm75_het3_publication01142019.vcf.gz
Processed VCF file (after imputation and filtering). # of accession = 811, # of SNP sites ~3 million