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 about the IWGSC Data Repository, please contact Michael Alaux from URGI in the first instance at: urgi-support@inrae.fr
In case of questions about the data content, you can contact the data providers.

ARTICLE
Zhu et al., Optical maps refine the bread wheat Triticum aestivum cv Chinese Spring genome assembly, Plant J. 2021 Apr 24, https://doi.org/10.1111/tpj.15289

Gene Annotation

CREDIT: Frédéric Choulet and Hélène Rimbert (INRAE) with funding from the French Government managed by the Research National Agency (ANR) under the Investment for the Future program (BreedWheat project ANR-10-BTBR-03).
A new annotation, IWGSC Annotation v2.1, to accompany RefSeq v2.1 was completed. Initially, the previous annotation was updated to IWGSC Annotation v1.2 by integrating a set of 117 novel genes and 81 microRNA, many of which had been curated manually by the wheat community and then this, in turn, was used to annotate IWGSC RefSeq v2.1. The transposable elements (TEs) in the resulting assembly IWGSC RefSeq v2.1 were reannotated and gene annotation was updated by transferring the previously known gene models (v1.1) using a fine-tuned, dedicated strategy implemented in the Marker-Assisted Gene Annotation Transfer for Triticeae (MAGATT) pipeline. The newly released IWGSC RefSeq Annotation v2.1 contains 266,753 genes comprising 106,913 HC genes and 159,840 LC genes.

Author: Helene Rimbert
Date: September 09 2020
Title: Annotation of RefSeq v2.1
Contact: <helene.rimbert@inrae.fr>; <frederic.choulet@inrae.fr>

Annotation of RefSeq v2.1

This gene annotation is based on the automatic transfer of the annotation v1.2 based on IWGSC RefSeq v1.
Annotation v1.2 is available in the IWGSC Data Repository:
https://urgi.versailles.inrae.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.2/

All gene structure are in GFF format for both High Confidence and Low Confidence genes.

* Correspondence table between Annotation v1.2 and v2.1
iwgsc_refseqv2.1_annotation_200916_IDmapping.csv

* Gene structure
iwgsc_refseqv2.1_annotation_200916_HC.gff3
iwgsc_refseqv2.1_annotation_200916_LC.gff3

GFF file of the gene structure. Contains all mRNA isoforms , UTRs and CDS features.

* Sequences
iwgsc_refseqv2.1_annotation_200916_HC.cds.fasta
iwgsc_refseqv2.1_annotation_200916_HC.proteins.fasta
iwgsc_refseqv2.1_annotation_200916_HC.proteins.valid.fasta
iwgsc_refseqv2.1_annotation_200916_HC.cds.valid.fasta
iwgsc_refseqv2.1_annotation_200916_HC.transcripts.fasta
iwgsc_refseqv2.1_annotation_200916_LC.cds.fasta
iwgsc_refseqv2.1_annotation_200916_LC.cds.valid.fasta
iwgsc_refseqv2.1_annotation_200916_LC.proteins.fasta
iwgsc_refseqv2.1_annotation_200916_LC.proteins.valid.fasta
iwgsc_refseqv2.1_annotation_200916_LC.transcripts.fasta

FASTA file of all mRNA, CDS and translated CDS (proteins).

All CDS/proteins which had no warning while running `fastavalidcds` are in the 'cds.valid.fasta' and 'proteins.valid.fasta' file.

* Validation information of CDS

All CDS sequences were analysed with `fastavalidcds` tool rom Exonerate package (version 2.4.1).
The results are saved in the following files:
iwgsc_refseqv2.1_annotation_200916_HC.cds.valid.explained.txt
iwgsc_refseqv2.1_annotation_200916_LC.cds.valid.explained.txt

* Marker-based mapping summary
iwgsc_refseqv2.1_annotation_200916_anchoringSummary.csv
iwgsc_refseqv2.1_annotation_200916_blatSummary.csv

All mapping information are compiled in these two files.
For each gene, the pair of ISBPs marker used for the targeted mapping is shown and the BLAT result on the target region is reported.

* Other

File Name	Description
`iwgsc_refseqv2.1_annotation_200916_gmap_differentChrom.gff3`	GMAP of isoforms which are not on the same chromosome.
`iwgsc_refseqv2.1_annotation_200916_missing.bed`	RefSeqV1.2 gene models for which we failed to transfer.

|`iwgsc_refseqv2.1_annotation_200916_overlap.bed`| RefSeqv2.1 gene models which are overlapping each other.|

Warnings

* non optimal mapping

We are aware that 133 genes should be taken carefully.
The list of genes below have a perfect match in a different location on the genome than what we have found with the targeted approach.
Here are the list of genes:

# v1	v2
TraesCS1A02G394700	TraesCS1A03G958700
TraesCS1A02G402100	TraesCS1A03G983700
TraesCS1B02G025600	TraesCS1B03G051200
TraesCS1B02G367700	TraesCS1B03G1003200
TraesCS1B02G473500	TraesCS1B03G1260300
TraesCS1B02G076800	TraesCS1B03G191800
TraesCS1B02G320500	TraesCS1B03G878900
TraesCS1D02G003200	TraesCS1D03G006900
TraesCS1D02G001700	TraesCS1D03G009400
TraesCS1D02G001600	TraesCS1D03G009500
TraesCS1D02G000200	TraesCS1D03G012400
TraesCS1D02G294400	TraesCS1D03G708200
TraesCS2A02G015400	TraesCS2A03G032000
TraesCS2A02G548500	TraesCS2A03G1267600
TraesCS2A02G548600	TraesCS2A03G1267700
TraesCS2A02G557200	TraesCS2A03G1279900
TraesCS2A02G592600	TraesCS2A03G1289600
TraesCS2A02G582900	TraesCS2A03G1310000
TraesCS2A02G576400	TraesCS2A03G1328200
TraesCS2A02G567000	TraesCS2A03G1347400
TraesCS2A02G561900	TraesCS2A03G1356500
TraesCS2A02G244700	TraesCS2A03G577700
TraesCS2A02G246700	TraesCS2A03G591800
TraesCS2A02G339500	TraesCS2A03G805300
TraesCS2A02G336500	TraesCS2A03G810600
TraesCS2A02G334200	TraesCS2A03G814200
TraesCS2A02G332000	TraesCS2A03G817600
TraesCS2A02G326300	TraesCS2A03G829300
TraesCS2A02G322600	TraesCS2A03G836900
TraesCS2A02G322400	TraesCS2A03G837300
TraesCS2B02G023900	TraesCS2B03G000100
TraesCS2B02G011500	TraesCS2B03G028400
TraesCS2B02G005700	TraesCS2B03G035500
TraesCS2B02G005300	TraesCS2B03G036700
TraesCS2B02G002600	TraesCS2B03G042700
TraesCS2B02G431600	TraesCS2B03G1099000
TraesCS2B02G431400	TraesCS2B03G1099200

TraesCS2B02G589300	TraesCS2B03G1476100
TraesCS2B02G621800	TraesCS2B03G1564900
TraesCS2B02G618100	TraesCS2B03G1573100
TraesCS2B02G617300	TraesCS2B03G1576200
TraesCS2B02G273200	TraesCS2B03G697000
TraesCS2B02G272700	TraesCS2B03G697900
TraesCS2B02G270700	TraesCS2B03G703800
TraesCS2D02G474700	TraesCS2D03G1061400
TraesCS2D02G592600	TraesCS2D03G1281700
TraesCS2D02G587800	TraesCS2D03G1293400
TraesCS2D02G573800	TraesCS2D03G1327500
TraesCS2D02G570000	TraesCS2D03G1334000
TraesCS2D02G567600	TraesCS2D03G1341200
TraesCS3A02G031500	TraesCS3A03G061800
TraesCS3A02G137100	TraesCS3A03G312000
TraesCS3A02G228800	TraesCS3A03G587100
TraesCS3A02G227900	TraesCS3A03G588300
TraesCS3A02G302000	TraesCS3A03G746000
TraesCS3A02G298900	TraesCS3A03G751500
TraesCS3A02G297400	TraesCS3A03G754500
TraesCS3A02G322200	TraesCS3A03G762000
TraesCS3B02G002000	TraesCS3B03G000300
TraesCS3B02G018400	TraesCS3B03G038200
TraesCS3B02G018200	TraesCS3B03G038600
TraesCS3B02G509100	TraesCS3B03G1261500
TraesCS3B02G510900	TraesCS3B03G1271000
TraesCS3B02G554600	TraesCS3B03G1381300
TraesCS3B02G608200	TraesCS3B03G1515800
TraesCS3B02G607600	TraesCS3B03G1517100
TraesCS3B02G605700	TraesCS3B03G1522000
TraesCS3B02G173700	TraesCS3B03G414500
TraesCS3B02G230700	TraesCS3B03G592400
TraesCS3B02G233800	TraesCS3B03G605900
TraesCS3B02G263100	TraesCS3B03G690200
TraesCS3D02G005500	TraesCS3D03G010000
TraesCS3D02G001500	TraesCS3D03G017500
TraesCS3D02G033500	TraesCS3D03G060400
TraesCS3D02G476300	TraesCS3D03G1051100
TraesCS3D02G522000	TraesCS3D03G1154000
TraesCS3D02G521600	TraesCS3D03G1154600
TraesCS3D02G521100	TraesCS3D03G1155300
TraesCS3D02G544100	TraesCS3D03G1199700
TraesCS3D02G229500	TraesCS3D03G535700
TraesCS3D02G224400	TraesCS3D03G545100
TraesCS3D02G222800	TraesCS3D03G549100
TraesCS3D02G221400	TraesCS3D03G551700
TraesCS4A02G409000	TraesCS4A03G1013100
TraesCS4A02G336800	TraesCS4A03G839600
TraesCS4A02G399500	TraesCS4A03G991900
TraesCS4B02G000900	TraesCS4B03G003200

TraesCS4B02G004500	TraesCS4B03G009200
TraesCS4B02G009200	TraesCS4B03G013900
TraesCS4B02G006600	TraesCS4B03G019900
TraesCS4B02G341700	TraesCS4B03G885800
TraesCS5A02G195100	TraesCS5A03G499200
TraesCS5A02G193700	TraesCS5A03G501400
TraesCS5A02G193300	TraesCS5A03G502200
TraesCS5A02G192600	TraesCS5A03G503600
TraesCS5A02G211100	TraesCS5A03G553100
TraesCS5B02G454000	TraesCS5B03G1115300
TraesCS5B02G547200	TraesCS5B03G1319000
TraesCS5B02G544200	TraesCS5B03G1327400
TraesCS5B02G542300	TraesCS5B03G1331600
TraesCS5B02G538100	TraesCS5B03G1340100
TraesCS5D02G005000	TraesCS5D03G028200
TraesCS5D02G005100	TraesCS5D03G028300
TraesCS5D02G565400	TraesCS5D03G1211400
TraesCS5D02G134300	TraesCS5D03G339600
TraesCS5D02G329000	TraesCS5D03G741000
TraesCS5D02G374200	TraesCS5D03G837900
TraesCS6A02G022400	TraesCS6A03G047400
TraesCS6A02G419900	TraesCS6A03G1048300
TraesCS6A02G419100	TraesCS6A03G1050000
TraesCS6B02G005400	TraesCS6B03G016500
TraesCS6B02G215300	TraesCS6B03G578500
TraesCS6B02G227700	TraesCS6B03G629000
TraesCS6D02G061800	TraesCS6D03G133300
TraesCS6D02G087300	TraesCS6D03G177400
TraesCS6D02G085600	TraesCS6D03G180600
TraesCS6D02G083600	TraesCS6D03G184400
TraesCS6D02G082600	TraesCS6D03G187300
TraesCS6D02G248300	TraesCS6D03G598400
TraesCS6D02G245000	TraesCS6D03G605200
TraesCS6D02G351600	TraesCS6D03G813400
TraesCS7A02G513500	TraesCS7A03G1247700
TraesCS7A02G513800	TraesCS7A03G1249300
TraesCS7B02G046600	TraesCS7B03G123600
TraesCS7B02G461800	TraesCS7B03G1239000
TraesCS7B02G460100	TraesCS7B03G1242800
TraesCS7B02G486600	TraesCS7B03G1306900
TraesCS7B02G493900	TraesCS7B03G1333100
TraesCS7B02G099400	TraesCS7B03G261000
TraesCS7D02G159100	TraesCS7D03G1263800
TraesCS7D02G548800	TraesCS7D03G1294400
TraesCSU02G066000	TraesCSU03G063100
TraesCSU02G212600	TraesCSU03G337800

* Different strand for isoforms of the same gene

We also found 7 genes for which the mapping with GMAP leaded to multiple isoforms on different strands. In such cases, we split the gene in two models, and increased the number in the ID by '50'. One on each strand.

TraesCS7B03G750400LC
TraesCS7B03G509700LC
TraesCS6A03G509000LC
TraesCS5D03G1015500LC
TraesCS5D03G1015200LC
TraesCS2A03G565000LC
TraesCS1B03G500600

Transposable elements annotation

CREDIT: Romain De Oliveira, Frédéric Choulet

* TE

TE annotation was performed similarly to the Chinese Spring refSeq V1.0 using CLARITE (IWGSC, 2018; Wicker et al., 2018) and the *Triticeae* TE database ClariTeRep (Daron et al., 2014).

Related files:

Tae.Chinese_Spring.refSeqv2.1.TE_annotation.gff3.gz

Updated version uploaded 17 December 2020.

* ISBPs

ISBPs markers were designed using the pipeline described in (De Oliveira et al., 2020) that first extract the 150pbs which include the junction between TE and its insertion site (perl script getISBPfromBed.pl) and then filter out redundant markers and ISBPs sequence with Ns.

Related file:

Tae.Chinese_Spring.refSeqv2.1.ISBPs.bed

RefSeqV2.1 TE annotation reported 4,199,592 TEs belongings to 506 families, accounting for 85.0% (AA: 86.1%, BB: 84.9%, DD: 83.8%) of the refSeqv2.1 pseudomolecules.

Markers mapping

CREDIT: Helene Rimbert, Frédéric Choulet

Author: Helene Rimbert
Date: September 2020
Contact: <helene.rimbert@inrae.fr>; <frederic.choulet@inrae.fr>

Anchoring markers from RefSeq-v1 to RefSeq-v2.1

Markers already anchored on RefSeq-V1

Based on dataset on the IWGSC repository:
https://urgi.versailles.inra.fr/download/iwgsc/IWGSC_RefSeq_Annotations/v1.0/iwgsc_refseqv1.0_Marker_mapping_summary_2017Mar13.zip

$ ll iwgsc_refseqv1.0_Marker_mapping_summary_2017Mar13/
axiom_820k.summary.gff
barc-accessions.summary.gff
bristol_SNP.summary.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.barc.ssr.trimmed.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.cfa.ssr.trimmed.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.cfd.ssr.trimmed.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.gdm.ssr.trimmed.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.mas.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.wmc.ssr.trimmed.gff
Chinese_spring_pseudomolecules_v1.0.repeatmasked.wms.ssr.trimmed.gff
DArT_gbs.summary.gff
DArT_public.summary.gff
DArT_ver3.summary.gff
infinium90K.summary.gff
infinium9K.summary.gff
iSelect.summary.gff
NSF-EST.summary.gff
Sourdille-EST.summary.gff
TaBW280k.gff3
wEST_mapped.summary.gff

All markers were mapped based on the RefSeq-v1 data listed above (FASTA sequence extracted using the GFF) except for:

* Markers from the Axiom 820k and 35k were mapped using the data on https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/axiom_820K_search.php

* Markers from the iSelect 83k were mapped using the data on https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/iselect_mapped_snps.php

* Markers from the KASP core marker set were mapped using the https://www.cerealsdb.uk.net/cerealgenomics/CerealsDB/kasp_mapped_snps.php

Mapping on RefSeq-v2.1

All markers as FASTA sequences were mapped with GMAP as follow:
$ gmapl --min-identity=0.95 --min-trimmed-coverage=0.8 --ordered --npaths 3 --nosplicing -f 4

Results are in GFF format.

marker set	num markers in fasta	mapped	no hit	% mapped
affy_35K_array_data	35143	34194	949	s97,30 %
affy_820K_array_data	819571	788915	30656	96,26 %
barc	495	481	14	97,17 %
bristol_SNP	3998	3998	0	100,00 %
DArT_gbs	24029	24001	28	99,88 %
DArT_public	1498	1495	3	99,80 %
DArT_ver3	4207	4193	14	99,67 %
infinium90K	81590	81540	50	99,94 %
infinium9K	6192	6185	7	99,89 %
iSelect_80k	83022	72669	10353	87,53 %
KASP_core_marker_set_July_2013	960	763	197	79,48 %
NSF-EST	11208	11014	194	98,27 %
Sourdille-EST	2846	2803	43	98,49 %
tabw280k	253419	252910	509	99,80 %
TaBW35K_SNP_array	34746	34692	54	99,84 %
wEST_mapped	6155	6035	120	98,05 %
mas	56	56	0	100,00 %
barc.ssr	143	141	2	98,60 %
cfa.ssr	22	19	3	86,36 %
cfd.ssr	58	57	1	98,28 %
gdm.ssr	20	18	2	90,00 %
wmc.ssr	241	206	35	85,48 %
wms.ssr	111	107	4	96,40 %