Structural and Functional Annotation of the Wheat Genome

4

Frédéric Choulet, Xi Wang, Manuel Spannagl, David Swarbreck, Hélène Rimbert, Philippe Leroy, Pauline Lasserre-Zuber and Nathan Papon

Abstract

Wheat genome sequencing has passed through major steps in a decade, starting from the sequencing of large contiguous sequences obtained from chromosome-specific BAC libraries, to reach high-quality genome assemblies of a dozen of bread wheat varieties and wild relatives. While access to an assembled genome sequence is crucial for research, the resource that is mainly used by the community is not the sequence itself, but rather the annotated features, i.e., genes and transposable elements. In this chapter, we describe the work performed to predict the repertoire of 107 k high-confidence genes and 4 million TE copies in the hexaploid wheat genome (cultivar CHINESE SPRING; IWGSC RefSeq) and the procedures established to transfer the annotation through the different releases of genome assembly. Limitations and implications for building a wheat pangenome are discussed, as well as the possibilities for future improvements of structural annotation, and opportunities offered by novel approaches for functional annotation.

F. Choulet (✉) · H. Rimbert · P. Leroy ·
P. Lasserre-Zuber · N. Papon
UCA, INRAE, GDEC, Clermont-Ferrand, France
e-mail: frederic.choulet@inrae.fr

H. Rimbert
e-mail: helene.rimbert@inrae.fr

P. Leroy
e-mail: philippe.leroy.2@inrae.fr

P. Lasserre-Zuber
e-mail: pauline.lasserre-zuber@inrae.fr

N. Papon
e-mail: nathan.papon@inrae.fr

X. Wang
BASF Belgium Coordination Center CommV, Trait Research, Gent Zwijnaarde, Belgium
e-mail: xi.wang@basf.com

M. Spannagl
PGSB Plant Genome and Systems Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
e-mail: manuel.spannagl@helmholtz-muenchen.de

D. Swarbreck
Earlham Institute, Norwich Research Park, Norwich, Norfolk, UK
e-mail: david.swarbreck@earlham.ac.uk

4.1 Introduction

The International Wheat Genome Sequencing Consortium (IWGSC; http://www.wheat-genome.org) was launched in 2005 with the aim of accelerating research in wheat by delivering molecular markers and genomic resources with

the long-term goal of getting a high-quality reference genome sequence for the hexaploid wheat (Feuillet and Eversole 2007). It represents more than a decade of coordinated efforts from the completion of the first chromosome-specific BAC library construction (Paux et al. 2008) to the assembly of the 21 chromosome sequences of cultivar CHINESE SPRING (IWGSC 2018). Since the first release in 2018, the IWGSC integrated additional information coming from optical mapping and long reads in order to improve the quality of the assembly by correcting misordered scaffolds and filling gaps. This led to release RefSeq v2.0 and v2.1 in 2021 (Zhu et al. 2021).

Besides the methodological challenge of assembling this genome, the work performed to deliver an annotation is not well known and often poorly considered. Annotation consists of the identification of sequence features providing biological information, and it represents one of the most difficult tasks in genome sequencing projects. It is far from being obvious. However, annotation is the data mostly accessed by users, contrary to the genome sequence. Achieving a robust structural and functional genome sequence annotation is, thus, essential to provide the foundation for further relevant biological studies (Yandell and Ence 2012). Annotation of the RefSeq v1.0 required the coordinated effort of the IWGSC Annotation Group, bringing together researchers from three different Institutes: GDEC (France), PGSB (Germany), and Earlham Institute (UK). In addition, after the first release of the annotation, additional work has been performed in order to incorporate manual curation, and especially to update the annotation following changes to the genome assembly. This was achieved by developing fine-tuned bioinformatics approaches.

In this chapter, we present an overview of the processes that were established in order to release the first version of the annotation of RefSeq v1.0 and the updates since the first version. Besides the description of the work performed, this chapter is also a current opinion to consider the degree of approximation, the limits of the resources available and used for downstream analyses, and thus, a critical view of the quality of the data. The chapter also includes the plans for future versions not only for the structural annotation, but also for functional annotation.

4.2 Methods, Strategies, Resources for Structural Annotation of Genomes and Their Implications in Wheat Pangenomics

4.2.1 General Aspects of Structural Annotation

Depending on the sequence features targeted for study, and depending on the organism, genome annotation can be either trivial or complicated. This is why there may be a confusion for non-experts who may believe annotation is routine in genome sequencing projects. This is not the case for many species, and especially, this was not the case for wheat. For instance, in compact bacterial genomes, coding genes are intron-less and represent the very wide majority of the genome so that predicting the presence of coding open reading frames is obvious and does not even require human curation. For species already widely studied, like in human for instance, with several genomes already assembled and annotated, annotation may be routine since it is based purely on similarity with available highly conserved genomes. The difficulty of annotation increases with the size of the genome, the repeat content and active transposable element (TE) expression, the ploidy, the fragmentation of coding genes into small exons, and with the phylogenetic distance to an already well-characterized genome. The difficulty also increases with the level of conservation of the predicted features. A protein-coding gene highly conserved among distant species will be easily predicted with high confidence, while predicting poorly conserved features with a high level of accuracy is more complicated.

Annotation relies on the combination of approaches: (i) the homology-based method using alignment/mapping algorithms searching for sequence similarity either with proteins, showing that a sequence is conserved across evolution, and/or transcriptomic data, showing that a sequence is expressed; (ii) the ab initio methods, i.e., predictions using statistical models such as hidden Markov models (HMMs); (iii) structural feature-based method through the identification of intrinsic information like motifs at the borders of transposons. It thus relies on a combination of software, algorithms, and adapted reference libraries. Annotation needs to be automated, i.e., performed through a pipeline that combines all different programs and minimizes the subsequent long and laborious step of manual curation.

4.2.2   Sequence Features Usually Annotated and Common Ambiguities

In the plant genomics area, publications usually report on genes and repeats. Both terms are, however, confusing and the shortcut widely accepted by the community to distinguish genes and repeats is ambiguous. First, for convenience, the term "gene" is used as a shortcut for protein-coding gene. It will be the case in this chapter too. When a "number of genes" is given, it nearly always refers to a number of protein-coding genes. However, genomes also carry non-coding RNA (ncRNA) genes which are biologically important. In the annotation area, we distinguish two types of non-coding RNA genes: (i) highly conserved ncRNAs involved in essential cellular processes (splicing, translation) which are ribosomal RNAs, transfer RNAs, small nuclear and nucleolar RNAs, and (ii) less evolutionary conserved ncRNAs like micro-RNAs, long-non-coding RNAs, and others involved in specific regulation processes. Annotating conserved and non-conserved ncRNAs follows two completely different approaches. rRNA, tRNAs, snoRNAs,

snRNAs are easily identified by a simple similarity-search approach; however, they tend not to be annotated. The reason for that is probably that they are of interest only for research groups working specifically on them and that are able to identify them with specific tools. In contrast, annotation is much more complicated for the species-specific ncRNAs. It requires the availability of small RNASeq reads that could be mapped to identify transcribed regions as a first clue before concluding to the presence of an ncRNA gene. Second, genes are repeats. In bread wheat, the majority of the "genes" are repeated with only 17% (30,948/181,036) of single-copy genes (IWGSC 2018) so reference to genes versus repeats brings confusion particularly when some repeats carry genes. "Repeats" is a general term encompassing simple repeats as satellite DNA, telomeric repeated motifs, but also transposable elements (TEs), and their mobilizable or inactive derivatives. Usually in plant genome annotation, the term TE is used to describe all elements whatever their status, autonomous, non-autonomous, transposable, mobilizable, or inactive. TEs can carry genes and/or pseudogenes that encode proteins involved in transposition. In species like wheat, where the genome is massively comprised of TEs, it is essential to identify them to avoid calling genes that are in fact derived from TEs and, thus, are/were involved in transposition rather than a function related to a phenotype and under selection pressure.

The problems described above limit our ability to determine if a sequence is a functional protein-coding gene, a pseudogene, or part of a TE, with high confidence. In addition, the lack of evidence sometimes limits our ability to precisely determine the structure of a gene. Positions of the start codon and borders between coding exons and introns can remain doubtful in many cases. Transcriptomic data like RNASeq are extremely useful to determine exon/intron borders, the existence of alternative transcripts, and the extent of untranslated regions (UTRs of the mRNA upstream the start and downstream the stop codons). Fixing the start codon position,

however, often requires protein sequence homology. Usually in whole-genome annotation projects, for each gene, the most important is to predict the coordinates of the CDS features (i.e., the coding exons). With RNASeq, it became a routine to also annotate the positions of UTRs and all alternatively spliced mRNAs, while defining one representative mRNA/CDS per gene (usually the longest or the most conserved with other species, numbered "1" by convention). For low or non-expressed genes, UTR and mRNA coordinates may not be predicted because of a lack of information. In that case, the gene coordinates are limited to the CDS, which remains the basic essential annotation for a protein-coding gene. For wheat, our main goal was to predict CDS first and, if possible, to add the layer of UTRs and transcripts, these later ones being highly dependent on the RNASeq samples available and methods used.

Wheat gene models have been assigned a confidence category, namely high versus low confidence (HC, LC). This could be misleading since confidence may rely either on the existence of a gene or rather on its exon/intron structure. For instance, one can be highly confident that a sequence encodes a gene while weakly confident on its exact exon coordinates. Both are related. Doubt of the existence of a gene at a given locus is associated with lack of homology evidence. In RefSeq v1, the HC/LC categories classified genes based on their level of similarity (complete or partial) with proteins from other plants. The consequence is that HC genes are likely functional and conserved among *Poaceae* even if some might be predicted with a doubtful structure. LC genes share partial similarity with known proteins and can be well-defined functional genes but the qualitative judgment is of low confidence.

Refinement of automated annotation pipelines to deal with the LC "challenge" is expected to engage manual curation by experts. Manual curation is required to improve the overall quality of the automated annotation. However, manual curation may be mistakenly considered as a validation. Both computer and human algorithms take a decision based on a priori knowledge on the structure of genes and on homology information. When the decision is obvious, typically for genes widely conserved, homology with known proteins and mapped transcripts, if consistent, human curation is not needed. When homology is weak or partial, with a lack of transcription evidence, manual curation does not allow to achieve high confidence neither on the existence of a gene nor on its structure. Curation has a positive impact only in particular cases: missing genes (with evidence slightly under default thresholds), chimeric tandem duplicated genes, start codon mis-assignment, and correction of gene models that are in fact pseudogenes because truncated or with frameshift mutations. These are all particular cases where the situation deviates from standard and is too complex for algorithms.

For TEs, especially in large genomes, manual curation has a much stronger impact than it has for genes. Automated TE modeling is extremely complicated in genomes like wheat where TEs cover 85% of the genome. The history of nested insertions of young elements into old ones has shaped a mosaic of TEs highly fragmented. For instance, manual curation led to identify blocks of nested TEs in which the two extremities of the older element are separated by > 200 kb (Choulet et al. 2010). Such reconstruction is a computational challenge, and manual curation still has a major impact on the quality of the TE annotation. However, with around 4 million TEs in the wheat genome, manual curation was limited to small regions for the moment.

4.2.3 TEs *Versus* Genes: The Crucial Point of Having a Manually Curated TE Library

Providing the complete (protein-coding) gene catalog of a sequenced genome is the priority of annotation. The impact of our knowledge about TEs on our ability to determine if an ORF is part of a functional gene, or if it is a TE-related ORF, is illustrated in rice, where the first releases in 2002 over-predicted around 50,000 genes (Goff et al. 2002; Yu et al. 2002;

Bennetzen et al. 2004) because of unknown TEs. In the wheat context, in the first release (RefSeq v1.1), the predicted CDSs represented 143 Mb [i.e., 107,891 HC genes; (IWGSC 2018)] which is not even 1% of the genome versus 85% for TEs. Considering the possibility that if even only 5% of the TEs are not correctly identified, the amount of "TE-related ORFs" considered as potential functional genes would exceed the total number of predicted genes. Consistent with such a high degree of uncertainty was the initial number of 908,149 candidate loci (after filtering out TE-matching loci) that matched either transcripts and/or homologous proteins in the wheat draft genome annotated in 2014 (IWGSC 2014). RNASeq analysis highlighted 976,962 potentially expressed loci in this study (generating polyA-tailed transcripts), a number considered to be well in excess compared to what was expected based on studies in model grasses. Releasing an annotation that is a good representation of the biological reality is therefore a challenge, and the availability of a curated TE library is of major importance since it could filter out thousands of mis-called genes.

In the development of a representative wheat genome sequence, the long-standing effort to build a high-quality curated TE library has provided a sound foundation. From the beginning of BAC sequencing in wheat, barley, and related *Triticeae*, which all share common TE families, several groups around the world have contributed to manually annotate TEs while defining their exact borders (by searching for terminal repeated motifs). These TEs were organized, classified, and distributed through the Triticeae Repeat (TREP) library maintained by Thomas Wicker at Zurich University, a resource extremely useful for masking TEs, a common task in genome annotation meaning that nucleotides assigned to TEs are converted to Ns (or to lowercases). In 2010, the first large contiguous wheat sequences (obtained from BAC-contigs) were published, representing 18 Mb (Choulet et al. 2010). Although it accounted only 0.1% of the genome, it doubled the amount of wheat sequences available at that time. Even though our knowledge of the wheat genome was still

extremely partial, similarity-searches against TREP already identified 75% of the sequence as TEs. This early work demonstrated that manual annotation of a small fraction of the genome allowed the identification of all the abundant TE families, highly repeated, that comprised most of the genome. It also revealed that CACTAs were underrepresented in the library, contrary to LTR-Retrotransposons (LTR-RTs) Gypsy/ Copia. The main reason being that the level of variability/diversity of LTR-RTs is low compared to CACTAs. This impacts TE annotation/masking because similarity-search (at low stringency) allows cross-matching between LTR-RT families, meaning that it is not necessary to have identified all families to mask the unknown ones. In contrast, for CACTA families, similarity between families is often limited to the extremities of the element while the internal part is much more variable. This is why a special effort was made, in 2010, to manually curate 3222 elements, especially 330 CACTAs, in order to enrich the wheat TE library (Choulet et al. 2010). This led to the proportion of predicted TEs increasing from 75 to 85% of the genome. In 2014, these ca. 3200 new elements were combined to TREP and classified de novo and a more exhaustive library called ClariTeRep was established (Daron et al. 2014). ClariTeRep is mostly enriched in CACTAs compared to the original TREP library and has a clear impact on TE annotation of *Triticeae* genomes. Several *Triticeae* sequencing projects concluded that CACTAs represent 5–6% of the genome (Jia et al. 2013; Ling et al. 2013), while their proportion is around 15% based on ClariTeRep.

4.2.4   Ab Initio, Homology-Based Predictions, and the RNASeq Revolution for Gene Calling in Complex Genomes

Pipelines for automated structural annotation usually require to combine information from ab initio predictors and evidence of similarity with known proteins in other species or transcriptome sequences (ESTs, full-length cDNAs,

RNASeq [short reads], IsoSeq [long reads] data). For large genomes like in wheat, the problem of ab initio predictors is the very high number of false positives. Indeed, since TEs are estimated to cover at least 85% of the genome, while genes would cover 1–2%, the remaining 13–14% of unannotated DNA account for approximately 2 Gb where gene finders predict gene models because of the presence of ORFs that look likely coding. The reason is that the unannotated part is shaped by low-copy TE-derived sequences, old TE relics, not identified with default TE identification approaches, that carry ORFs that are/were coding (e.g., fragment of transposase) and thus are mistakenly recognized by gene predictors.

Because of the TE-derived ambiguity, biological evidence of homology with related species has always been the criteria of choice to accurately predict genes in wheat. The bad point for wheat was that the number of related species with a sequenced genome was limited, among the *Poaceae*, to *Oryza sativa*, *Zea mays*, *Sorghum bicolor*, and *Brachypodium distachyon*. Outside the *Poaceae* (common ancestor 60 MYA), sequence similarity is too weak to ensure accurate homology-based predictions. This raised a serious problem: wheat genes conserved among the *Poaceae* were well-predicted but our ability to predict less conserved genes was very limited at the early stages of annotation before 2010, especially for species-specific genes.

Transcriptome sequencing considerably enhanced our ability to determine which regions of the genome carry genes because it showed evidence of transcription. Transcriptome sequencing started with a massive effort to sequence millions of ESTs and full-length cDNAs (Ogihara et al. 2004; Zhang et al. 2004) and was followed by the emergence of RNASeq technical capacity which provided unprecedented power to drive structural annotation. First use of an RNASeq expression atlas for wheat gene annotation at the chromosome scale was published in 2014 (Choulet et al. 2014; Pingault et al. 2015). In brief, 7264 gene models were predicted but only 5185 (71%) showed transcription evidence in an RNASeq atlas covering five plant organs at three developmental stages each. In addition, 3692 transcribed regions were detected in the unannotated sequences showing that 42% of the loci likely expressed did not correspond to predicted protein-coding genes. This indicated a high level of uncertainty in describing biological reality when annotating the wheat genome. In this chapter, we propose a critical view of automated gene annotation pipelines, namely that bioinformatics can predict but not demonstrate that a sequence is a gene and that a gene is not a pseudogene. Although RNASeq became a primary resource for structural annotation, the correspondence between RNASeq-read mapping loci and the final filtered gene set was far from perfect, with 29% of chr3B gene models showing no transcription evidence and 42% of transcribed regions not looking like protein-coding genes. Homology with related species remains an important benchmark.

4.2.5 Single-Gene Duplications Raise More Problems Than Polyploidy for Structural Annotation

Given the weight of similarity-search with transcripts and proteins in structural annotation, intrinsic features of the genome significantly impact the difficulty to identify the correct gene structure since sequence alignments underpin all the studies. A first important intrinsic feature to impact annotation is the fragmentation level, i.e., the number of exons per gene. As a CDS is fragmented into several exons, the difficulty to predict the correct intron/exon structure increases. In wheat, considering RefSeq Annotation v2.1, the average number of exons per CDS is only 4. Sixty percent of the CDSs are split into a maximum of 3 exons. Actually, only 10% of the gene set corresponds to CDSs split into ten exons or more. Thus, the fragmentation problem is limited in wheat.

Other important criteria are the lengths of exons and introns. Small exons might be missed by sequence alignments because under the

default thresholds of automated pipelines. Large introns also raise problems for spliced-alignments. In the current wheat annotation release, the average exon length is 498 bps and the average intron length is 280 bps (considering only one representative transcript per gene). Thus, exons are, on average, large enough for high-scoring alignments, and introns are small enough for the efficiency of spliced-alignments. So, although it is commented that the wheat genome is complex, some intrinsic features are rather less complex than in many other eukaryotes.

Does polyploidy impact our ability to call genes? The main problem with alignment-based methods for gene calling is obviously multiple mapping, i.e., the fact that a transcript/protein matches at multiple loci along the genome. But it does not mean at all that single-copy genes are easier to predict than duplicated genes. In contrast, the fact that a gene is repeated on, e.g., chromosomes 1A, 1B, and 1D, because of polyploidy is rather in favor of accurate structural annotation. Since each copy is carried by a different chromosome, it is annotated independently and this does not generate problems due to multiple mapping. The three subgenomes A-B-D could be annotated as if they were three genomes of three different species. If a gene copy is silenced and thus does not generate an RNASeq signal, reads coming from the copies that are transcribed can be used to predict the structure of all copies. So, again, to our opinion polyploidy is an advantage here for structural annotation. To go further, we can even consider that we did not fully exploit the advantage provided by this intrinsic redundancy of the genome for structural annotation of the IWGSC RefSeq. We will present this in more detail in the paragraph below describing future plans for improvements.

Large chromosomes such as found in wheat are usually fragmented into "chunks" that are annotated independently in parallel. The problems with multiple mapping arise when repeated copies of a gene are carried by the same chunk. This is typically the case for tandemly duplicated genes. This is why automated structural annotation of tandem duplicates is the most complicated task. Single-gene duplications are much more problematic than whole-genome duplication (i.e., polyploidy). This is true for every genome to be annotated mainly via the homology-based approach. However, for wheat, this problem has strong implications because we demonstrated that single-gene duplications intensively affected the gene repertoire during its recent evolution (Glover et al. 2015). In the IWGSC RefSeq v1.1, we found that 27% of genes were present as tandem duplicates (IWGSC 2018). Multiple mapping of homologous proteins and transcripts on tandem duplicates may lead to artificially link exons from the two copies and, thus, to predict chimeric genes. This is especially the case for highly identical copies that are separated by a small intergenic region, compatible with a classical intron length. Some highly repeated gene families such as the kinase genes and disease resistance genes are well known to fall into this category. Unfortunately, these genes are often the favorite candidates to control phenotypes of interest, and in that case, manual curation is a required step to improve significantly the accuracy of automated annotation.

4.3 RefSeq V1.0 Structural Annotation

4.3.1 The Impact of Annotation Procedure on Gene Predictions Is Very Strong

Sequencing the wheat genome has a long story. Different initiatives have been launched following the advances of sequencing technologies to tackle the hexaploid genome and also the genome of the diploid and tetraploid relative species. For CHINESE SPRING itself, before completing RefSeq v1, a draft genome assembly (named CSSs for chromosome survey sequences) was released in 2014 (IWGSC 2014) together with a chromosome-scale assembly of the entire chromosome 3B using a BAC-by-BAC approach, hereafter named "3B-BAC-2014"

(Choulet et al. 2014). In addition, another version of the CHINESE SPRING genome was produced and annotated in 2017 named TGACv1 (Clavijo et al. 2017). Hence, when the annotation of RefSeq v1 started, chromosome 3B has already been annotated three times independently: 3B-BAC-2014 with the TriAnnot pipeline at GDEC Institute (Clermont-Ferrand, France), CSS-3B-v2.2 at PGSB Institute (Munich, Germany), and TGACv1 at Earlham Institute (EI, Norwich, UK) with homemade pipelines. Here, we compared these three gene catalogs to have a flavor of the impact of the methods on the results released: among the 7264 CDSs predicted on 3B-BAC-2014, only 26% (1884) and 12% (867) were strictly identical in TGACv1 and CSSv2.2 (sharing strictly identical protein sequences). These percentages appear extremely low if one considers these are three independent initiatives to sequence/annotate the same genotype. It demonstrates the impact of the annotation procedure on the released gene catalog as well as the possible impact of the sequencing strategy and assembly quality.

4.3.2 Gene Annotation Through a Federated Approach

Given the strong differences observed when comparing results obtained by different groups, the IWGSC established an Annotation Working Group in order to coordinate the efforts and establish an integrated approach to annotate RefSeq v1. Genes were predicted independently by two groups using two different pipelines and two different strategies: GDEC and PGSB. Both were then integrated at EI to end up with a single annotation. This led to v1.0 which was quickly updated into v1.1 after integrating~4000 manually curated genes (see below for details on curation).

In v1.1, 107,891 high-confidence (HC) protein-coding loci were identified, with a relatively equal distribution across the A, B, and D subgenomes (35,345, 35,643, and 34,212, respectively). In addition, 161,537 other protein-coding genes were classified as low-confidence

(LC) genes, representing partially supported gene models, gene fragments, and orphans. On ChrUn (unplaced scaffolds), 2691 HC and 675 LC gene models were identified. Evidence for transcription was found for 85% (94,114) of the HC genes versus 49% of the LC genes. In addition, 303,818 pseudogenes were also annotated. The quality of RefSeq Annotation v1.1 was estimated with BUSCO v3 (24). It revealed that 99% (1436/1440) of the BUSCO v3 genes were present in at least one complete copy and 90% (1292/1440) in three complete copies.

4.3.2.1 Gene Modeling Using TriAnnot

The TriAnnot pipeline was developed and updated over a period of more than 10 years to enable automated robust structural and functional annotation of protein-coding genes, transposable elements, and conserved non-coding RNA genes in *Triticeae* genomes (Leroy et al. 2012). It was dedicated to large-scale annotation projects and is executable through the command line on high-performance computing infrastructures for parallelization with task dependencies. TriAnnot was initially used for the annotation of BACs (Choulet et al. 2010) and then for the entire chromosome 3B (Choulet et al. 2014). Thus, it was intensively trained and customized specifically for wheat before we assembled RefSeq v1.

The specificities of the annotation strategy implemented in TriAnnot included: (i) mask TEs first in order to restrict the gene modeling to the non-TE space; (ii) use both evidence-based and ab initio approaches before selecting the best gene model at each locus. It was launched individually on each scaffold (or chunks for large ones) of RefSeq v1.0 in parallel while positions of features were subsequently calculated on pseudomolecules. The different steps and tools launched by the pipeline are described below:

- Step 1: TE annotation and sequence masking. TEs were identified by similarity-search using CLARITE and ClariTeRep (Daron et al. 2014). CLARITE used RepeatMasker with cross_match as search engine for optimized accuracy (Smit et al. 1996–2004).

Nucleotides assigned to TEs were then masked so that the following steps, i.e., ab initio predictions and similarity-searches, were all performed on the masked genome sequence.

- Step 2: Gene modeling. Ab initio gene models were predicted using two gene finders previously trained with a wheat gene dataset: FGeneSH (http://linux1.softberry.com/berry.phtml) and AUGUSTUS (Stanke et al. 2006). Evidence-driven gene predictions were also computed following three different strategies giving different weights to protein and transcript similarities. The first approach was based on homology with proteomes of related species. Similarity-search was performed using BLAST (Zhang et al. 2000) and significant hits, filtered with fine-tuned thresholds, were then used for spliced-alignment using EXONERATE (Slater and Birney 2005). The query proteins were those predicted in main *Poaceae* species for which a genome sequence was available: *O. sativa* (International Rice Genome Sequencing Project 2005), *B. distachyon* (The International Brachypodium Initiative 2010), *S. bicolor* (Paterson et al. 2009), *Z. mays* (Schnable et al. 2009), and *Hordeum vulgare* (International Barley Genome Sequencing Consortium et al. 2012). This approach is well suited to precisely determine the obvious structure of a large fraction of the protein-coding genes by taking advantage of their evolutionary conserved nature. However, the main limit here was the lack of similarity at the protein extremities which may lead to incomplete alignment that prevents from finding the start and/or stop codons. Thus, TriAnnot utilized an iterative extension in order to identify in-frame start and stop codons for gene modeling. Models with partial structure were flagged pseudogenes.

  The second evidence-driven approach (SIMsearch module) was based on transcripts first, rather than proteins. SIMsearch module is a gene modeling program based on FPGP (Amano et al. 2010) and adapted specifically for wheat to address problems generated by tandem repeated genes. SIMsearch identified the loci that are transcribed by spliced-alignment using est2genome (Mott 1997) of a series of wheat transcript libraries. The CDS coordinates were predicted afterward through similarity with *Poaceae* proteomes. SIMsearch was launched twice using two databanks of wheat transcripts: (1) predicted transcripts derived from a large RNASeq experiment that targeted five plant organs at three development stages each in two replicates (Pingault et al. 2015); (2) all available wheat full-length cDNAs available at EBI-ENA and from Ogihara et al. (2004). Thus, TriAnnot did not use RNASeq reads directly as an input. Read mapping and transcript calling were computed prior to gene annotation, and the predicted transcripts were provided as FASTA input for spliced-alignment during the process of gene modeling.

- Step 3: Selection of the best gene model at every locus. In summary, TriAnnot predicts gene models through five approaches: two ab initio and three evidence-based (one derived from spliced-alignment of homologous proteins + two derived from transcript evidence). One gene may obviously be predicted through different ways. Thus, the final step is the selection of the best gene model at each locus. Indeed, at that step, there was no combination of different overlapping models to create a new one.

A scoring process was applied in order to validate the existence of a gene and to retain its most probable structure. For scoring, TriAnnot used BLASTP to search for similarity of each model with proteomes of related *Poaceae*, including *Aegilops tauschii* and *Triticum urartu*, and calculate a score while considering metrics of the best hit alignment (percentage of identity and coverage, presence of canonical splicing sites, presence of start and stop codons).

Gene models not supported by homology with *Poaceae* proteins or by transcription evidence were simply discarded (i.e., ab initio only). Models sharing similarity with known

proteins and for which splicing sites were supported by transcript evidence were classified as high confidence. Low-confidence genes also share similarity with known proteins and transcripts but lack support for some splicing sites and/or position of start/stop codons. Finally, genes sharing similarity with known proteins but over less than 70% of the length of its best BLAST hit were classified as pseudogenes. Thus, TriAnnot predicted 107,226 gene models: 65,884 HC and 41,342 LC genes, plus an additional 73,044 pseudogenes on the IWGSC RefSeq v1.

4.3.2.2 PGSB Gene Prediction Pipeline

The procedure implemented in the PGSB annotation pipeline differs in many aspects from that of TriAnnot. It is based on mapping all available evidence on unmasked genome sequence and filtering out TE-related predictions afterward. It was all evidence-driven, not using any ab initio gene finder.

- Step 1: Mapping. The PGSB annotation pipeline combined spliced-alignments of reference proteins, IsoSeq reads and full-length cDNAs (flcDNAs), and RNASeq transcript predictions. In addition to the RNASeq atlas from Pingault et al. (2015) also used in TriAnnot, additional samples were added here. There were Illumina reads produced on grain-specific samples (Pfeifer et al. 2014), whole transcriptome PacBio sequenced samples (PRJEB15048), and disease resistance gene enriched transcriptome samples (PRJEB23081). The latter were all from CHINESE SPRING but there were also transcriptomic data generated from other accessions cultivated under drought and heat stresses (SRP045409) and under infection by *Fusarium graminearum* (E-MTAB-1729). Mapping outputs were all combined, and mapped reads were assembled into transcripts with StringTie (Pertea et al. 2015).

  Protein sequences from the five species *Arabidopsis thaliana*, *B. distachyon*, *O. sativa*, *S. bicolor,* and *Setaria italica*, and complete proteins from *Triticeae* in UniProt (UniProt Consortium 2018) were aligned with GenomeThreader independently on each chromosome. flcDNAs from wheat and barley (Mochida et al. 2009), together with wheat IsoSeq reads (Clavijo et al. 2017) were mapped with Gmap (Wu and Watanabe 2005) and included in the prediction pipeline.

- Step 2: Prediction and selection of open-reading frames. Predictions originating from protein alignments, full-length transcript alignments, and RNASeq were combined while removing redundancy (using Cuffcompare and StringTie). Then, TransDecoder (https://github.com/TransDecoder/TransDecoder/) was used to predict the coding frame for each transcript while considering the most upstream start codon by default. These predictions were then aligned against a set of reference proteins from angiosperms in UniProt, and protein domains were also searched for. These data were given to TransDecoder for selecting the most probable CDS for each model.

Since TEs were not masked prior to mapping evidence, PGSB predictions were filtered out afterward based on similarity-search with TE-related proteins from the PTREP library (https://botserv2.uzh.ch/kelldata/trep-db).

4.3.2.3 Integration of TriAnnot and PGSB Gene Models with Mikado

Selection of the best representative model at each locus was applied through a rule-based approach that combined supporting evidence and intrinsic gene features. PacBio transcripts, RNASeq reads, and homologous protein alignments over the genome were used to measure the accuracy of predictions and a set of high-confidence splicing sites was established from RNASeq mapped reads. Mikado (Venturini et al. 2018) was used to cluster genes from the two pipelines into loci, to calculate an overall score to each gene model, and to select the highest-scoring gene model. The score reflected the congruence between a model and its supporting

evidence, calculated with an average F1-score (reflecting precision and recall) and metrics of gene feature, e.g., a penalty was applied to introns larger than 10 kb. After selecting the representative model, Mikado was used to identify additional high-quality alternatively spliced transcripts, only those that met a series of stringent requirements. The most important were: a CDS overlapping at least 60% of the representative CDS, without any retained intron, and with only verified exon/intron junctions. Eventually, to enrich the annotation, coordinates of UTRs were added based on comparing models and aligned transcripts with PASA (Haas et al. 2008).

4.3.2.4 Gene Confidence Assignment: HC Versus LC

Despite the sophisticated combination of both TriAnnot and PGSB predictions, the final number of models was very high: 269,428, representing approximately 90,000 protein-coding genes per (haploid) subgenome. As previously observed in wheat, regions showing traces of expression or homology with known proteins are much more abundant than expected, given that the number of protein-coding genes is a quite stable parameter in plant genomes with ~ 30,000 genes per haploid genome. It suggested that many gene models were in fact pseudogenes or doubtful non-coding transcribed regions for instance. However, both included filtering steps to discard models matching wheat transposons, before gene modeling for TriAnnot, after for PGSB. Thus, a confidence category was assigned to each gene model: high confidence versus low confidence. The idea was to provide a single filtered dataset of HC genes to people only interested in large-scale whole-genome analyses while keeping information of LC genes to people interested in the characterization of a particular region.

First classification parameter was the completeness of the model, i.e., the presence of both a start and a stop codon. HC genes were complete with significant homology with plant

(*Magnoliophyta*) proteins retrieved from Swiss-Prot and TrEMBL. LC genes were, either complete but without significant homology with plant proteins or, incomplete with or without significant homology. The 269,428 gene models were split into 107,891 HC (40%) and 161,537 LC (60%) protein-coding genes. The number of HC genes was much closer to the expected value for plants (~ 35,000 genes per haploid genome), and this became the reference dataset used by the community.

However, within all the limits explained here, we encourage users to always keep in mind the level of uncertainty behind the annotation space. To the question "how many protein-coding genes are there in wheat?" we should answer: We do not know because the proportion of doubtful predictions is just too high.

4.3.2.5 What Should Be Known About the LC Genes and Pseudogenes

The consequence of confidence assignment is that the LC category gathered genes that were non-conserved, i.e., might be species-specific, for which we did not have enough evidence to conclude it is functional, together with (highly) conserved genes that are either pseudogenes or just partially assembled or mis-predicted. One must consider that a part of the LC genes is conserved but exhibits a structure likely incomplete. This has strong implications for researchers interested in a particular gene family or a particular locus.

In addition, a specific search for pseudogenes was launched at the whole-genome level, based on finding DNA fragments sharing similarity with HC genes but only partially or with frameshifts and/or internal stop codons. In total, 288,939 pseudogenes were discovered with 10,440 corresponded to LC genes. Thus, the coding landscape is even more complicated than often believed, with 108 k HC, 162 k LC, and 279 k gene fragments and so if a gene is considered to be absent based on HC genes only, it is important to consider the pool of LC genes.

4.3.3 Comparing Genes Between A, B, and D Subgenomes

4.3.3.1 Finding Homeologous Groups Based on HC Genes Only Can Lead to False Conclusions and Highlights the Requirement of Considering LC Genes

Considering the conclusion of the latter paragraph, it implies that comparing the A-B-D gene repertoires was strongly impacted by the input gene dataset. Homeologous groups were inferred from gene trees. Initially, trees were built with the complete set of HC and LC genes which revealed that considering HC genes only led to considerably overestimate the level of variability between A-B-D subgenomes, because many LC genes were, in fact, orthologous to HC genes (i.e., homeologous in the hexaploid) even though functional annotation revealed that some LC genes represented mis-predicted TE-genes (e.g., transposase-like genes). The solution adopted was to work on a filtered gene dataset: 181,036 genes (103,757 HC and 77,279 LC genes; instead of 269 k initially) that do not correspond to either TE-related functions or to pseudogenes. This led to determine a total of 39,238 homeologous groups (i.e., clades of A-B-D orthologous deduced from gene trees) and 33% of them include LC genes. In total, 28,829 LC genes have homeologous partners and were thus valid for biological analyses.

The main conclusion of the A-B-D comparison was that the gene repertoire of the three subgenomes is much more different than previously thought. The default hypothesis is often that a gene is present in three pairs of homeologous copies in bread wheat because it is a hexaploid. The reality is that only 55% of the homeologous groups are triads, i.e., single-gene copy per subgenome (configuration 1:1:1). Thus, 45% of the groups represent cases where gene loss and/or duplications occurred after A-B-D divergence. Gene loss after A-B-D divergence represents the same proportion for A, B, and D: ~ 10% of the homeologous groups. Regarding gene duplications, they also occurred in the same proportions in A, B, and D. This analysis suggested that the three lineages leading to A-B-D genomes have independently accumulated differences (gene loss and gene duplications) at similar rates.

4.3.3.2 No Evidence of Any Biased Gene Fractionation and Importance of Gene Duplications

Regarding gene presence/absence, no evidence for biased partitioning was observed (IWGSC 2018). In contrast, comparisons support gradual loss/duplications that have occurred after A-B-D divergence in the diploid, tetraploid ancestors, and after hexaploidization event in modern bread wheat. Before gene loss, a gene may lose function because of silencing or change in expression, so that the first evidence of diploidization might be observed at the expression level. Hence, RNASeq data analyses showed that there was an equal contribution of the three homeologous genomes to the overall gene expression, demonstrating the absence of global subgenome dominance (IWGSC 2014).

4.3.4 TE Modeling

Given the amount of TEs shaping the wheat genome, predicting the presence of TE copies along assembled sequences has always been a prerequisite to avoid false predictions of coding genes that are in fact coding parts of TEs. Efforts to manually annotate TEs with their precise borders were made since the beginning of wheat BAC sequencing and a high-quality reference databank of wheat TE sequences was initiated in 2002 with TREP (Wicker et al. 2002) and completed in 2014 with the ClariTeRep library (Daron et al. 2014) (which includes TREP). ClariTeRep originated from manual curation of ~ 3200 TEs along the first large (Mb-sized) contiguous sequences produced on chromosome 3B (Choulet et al. 2010). This implies that the wheat TE library used for similarity-search might be biased toward elements from the B-subgenome, and depleted for A and

D subgenomes. However, it was shown that TE families that shaped the three subgenomes are the same, although subfamilies (variants) have differentially invaded the A-B-D genomes in the diploid ancestors (Wicker et al. 2018).

Thus, TE modeling in RefSeq v1.0 was performed only via a similarity-search approach against ClariTeRep. There was no de novo repeat-based discovery of new TEs. This led to the prediction of 3,968,974 copies, classified among 505 TE families, and representing 86%, 85%, and 83% of the A, B, and D genomes, respectively. Such proportions imply that TEs shape large clusters with recently inserted TEs into older ones, a mosaic of nested insertions which is a computational challenge to reconstruct. This step was dealt with CLARITE (Daron et al. 2014) for RefSeq v1.0. CLARITE uses RepeatMasker (Smit et al. 1996–2004) with the cross-match engine for the first step of similarity-search between the genome of the TE library. The main problems with using RepeatMasker in TE-rich genomes are as follows (i) the over-fragmentation: one copy is often not predicted into a single feature but rather split into adjacent fragments; (ii) the overlap of predictions, i.e., a locus could match with several reference; and (iii) scattered pieces of a TE that has been fragmented by subsequent TE insertions (nested pattern) are not joint. The CLARITE pipeline has been developed specifically for wheat, based on ClariTeRep, in order to overcome these three limitations. It uses classification information: all TEs in ClariTeRep were classified into families and subfamilies by sequence clustering. It also uses positions of LTRs in LTR-retrotransposons, which correspond to long terminal repeats (ca. hundreds of bps) that are largely involved in the fragmentation observed after RepeatMasker because both 5′ and 3′ LTRs cross-match since they are almost identical subsequences. Family classification and LTR positions are the two main points implemented in CLARITE. They allowed accurate defragmentation, while preventing chimeric merging of adjacent features, and accurate reconstruction of nested TEs.

4.4 RefSeq V1.0 Functional Annotation

Gene ontology terms, PFAM, and InterPro domains were assigned to gene models. A function was assigned to 82% (90,919) of HC genes in RefSeq Annotation v1.0. RNASeq-based transcription evidence was found for 85% and 49% of HC and LC genes, respectively. In addition, naming of gene function for each gene was performed by using the AHRD tool (Automated Assignment of Human Readable Descriptions, https://github.com/groupschoof/AHRD, version 3.3.3). This program generates informative functional annotations from BLAST outputs while avoiding retrieving too many "unknown" or "uncharacterized" functions. BLAST outputs against the following databases were parsed by AHRD: Swiss-Prot, *Arabidopsis* Araprot 11, and a subset of TrEMBL for *Viridiplantae*. A filter was then applied in order to discard genes with functions related to TEs. Genes were thus tagged as G (canonical gene), TE (obvious transposon), TE? (potential transposon), or U for unknown. Based on this, 3294 HC genes with a TE tag were moved subsequently to the LC category in RefSeq annotation v1.1.

4.5 RefSeq Annotation V1.1: Integration of Manually Curated Genes

Once Annotation v1.0 was released to the community, researchers who are experts of some specific gene families brought corrections to the automated predictions: Sometimes gene copies were missing, sometimes the predicted exon/intron structure needed to be curated. Feedback was made from the experts to the IWGSC Annotation Group in order to release an updated version 1.1. This concerns gene families CBFs, NLRs, PPRs, Prolamins, WAKs, and amino-acid transporters. A semiautomated process was developed in order to integrate manually curated gene models. It relies on a Python script using common tools like GenomeTools (Gremme et al.

2013), GFFCompare (Pertea and Pertea 2020), pyBEDTools (Dale et al. 2011). GffCompare was used to check that the curated genes did not overlap each other (different teams may have curated the same gene) and also to identify the RefSeq Annotation v1.0 models that required to be updated. Five types of correction were considered: (i) addition of a new gene model that was absent from v1.0; (ii) merging of two gene models; (iii) splitting of a gene model into two genes; (iv) correction of exon positions of a gene model; (v) complex cases which combined splitting and merging. RefSeq Annotation v1.1 includes updates of 3685 manually curated genes, of which 528 were not predicted by the automated annotation process and 354 corresponded to LC gene models. The final v1.1 HC gene set contained 107,891 genes.

4.6 RefSeq Annotation V2: The Challenge of Transferring Gene Annotation Through the Different Versions of Genome Assembly

In 2021, an update of the CHINESE SPRING IWGSC RefSeq Assembly was published (Zhu et al. 2021). Corrections were brought to the initial release by using new resources: Bionano and PacBio contigs. Inconsistencies between pseudomolecules and Bionano maps were reconciled, and 279 unplaced scaffolds were positioned into pseudomolecules. PacBio contigs publicly available (Zimin et al. 2017) were used to fill gaps. Contrary to scaffold reordering, the gap-filling step led to complete changes in the positions of gene models predicted along pseudomolecules, so that it was not possible to calculate new gene position from v1 to v2 with a simple conversion of coordinates. This raised two possibilities: compute de novo gene prediction or transferring the knowledge of the previous annotation release. Since annotation v1.1 was the outcome of an extensive effort to combine different annotation pipelines, the choice

was made to try to transfer as many models as possible while trying to optimize the traceability and to minimize the differences between Annotations v1 and v2.

However, finding the new position of a gene required sequence alignment, which raised many problems in hexaploid wheat. For example, we used GMAP to map 298,775 HC and LC genes onto Assembly v2 and observed that 32,152 (11%) could not be transferred accurately because of spurious alignments. Such high error rate was not acceptable and it was decided to develop a transfer-strategy dedicated to this task for wheat. It was implemented in the MAGATT pipeline (https://forgemia.inra.fr/umr-gdec/magatt). The strategy relies on reducing the alignment space to the shortest region predicted to carry the gene to be mapped. In wheat, genes are always flanked by TEs. Although TEs are repeats, each copy is inserted into a different site. Thus, the junction between a TE extremity and its insertion site is unique at the genome level. We derived all such tags from the TE annotation. They represent one tag every 3 kb (compared to one gene every 130 kb on average) that can be uniquely mapped from one assembly version to the other. We used these TE tags as anchors to define the smallest target interval before mapping a gene. The average size of an interval was 9.6 kb, which reduced the alignment space and avoided most problems due to multiple mapping of repeated genes. Even for clusters of tandemly repeated genes in which copies could share 100% identity, this strategy enabled the assignment of the correct interval for each copy and lead to the transfer of annotation of all copies without any cross-matching. MAGGAT succeeded to transfer 90% of HC/LC genes without any difference between v1 and v2 assemblies either in the introns or the exons, and 8% with mismatches due to nucleotide differences incorporated at the gap-filling step (in gap-flanking sequences). Indels were observed for 1% of the genes, and the remaining 1% corresponded to genes for which the sequence was discarded when assembling v2 (Zhu et al.

2021). This step gave rise to the IWGSC RefSeq Annotation v2.1.

Defining the target interval prior to mapping has a major consequence: It avoided the computation of a spliced-alignment of a query transcript/CDS. Indeed, by default MAGATT starts by mapping the entire gene feature (exons + introns + UTRs) with BLAT (Kent 2002) against the short, kb-sized, target sequence. In the majority of the cases (90%), it identified a full perfect match which enabled the repositioning of all sub-features (i.e., exons and UTRs of all alternative spliced mRNAs) from a previous to a new assembly that shared strict identity. This was of major importance because spliced-alignments could have led to errors, especially when exons are very small. When only mismatches (no Indels) were observed between the two assemblies for a given gene (3% of genes), automated repositioning was also possible. Spliced-alignments of mRNAs were computed only when BLAT returned Indels and/or partial match between a query gene and its target.

MAGATT was developed with the objective of transferring a gene annotation to a new assembly release for a given genotype. However, the strategy applies very well to the problem of annotating genes in the genome assemblies of other genotypes and is, thus, significant in the context of post-reference genome sequencing and pangenomics. Pangenomics aims at identifying conserved *versus* non-conserved genes in a series of assembled genomes. The main limit in this area is the quality of the gene predictions. It is therefore possible that presence–absence of a gene may simply be the consequence of annotation artifacts. Thus, MAGATT needs to be considered for delivering an annotation of gene models in new assemblies that mimics as much as possible the reference gene calls and avoid "polluting" the apparent dispensable gene set with differences in gene predictions.

4.7   Plans for Future Improvements

4.7.1   Improving Gene Structural Annotation

The repertoire of 107,891 genes delivered in 2018 for CHINESE SPRING is definitely a reference widely used by the community. However, the methodological limits mentioned above make us consider there are improvement levers. First of all, we must remind here that what we call genes here, by default, correspond to protein-coding genes. Non-coding RNA genes remains largely unexplored in this complex genome although we have no doubt their prediction along the genome sequence represents one of the most challenging tasks but also one of the most impacting novel information to increase our understanding of the functional sequences.

Regarding protein-coding genes, when we discuss the improvement of structural annotation, we distinguish two different things: (i) existence of the gene and (ii) structure of the gene. In other words, improvements concern, on one side, genes that are missing in the annotation and gene models that do actually not correspond to real genes. On the other side, improvements concern the exact structure of a gene and its transcripts.

A key question that impacts on both aspects is the presence of pseudogenes. Pseudogenes are sequences derived from functional genes but that have accumulated mutations (frameshift, in-frame stop codon, truncation) which switched its function off. Pseudogenes are hard to model automatically because gene modeling usually uses structural features (coding frame, start and stop codons) to call a gene while in case of pseudogenes, these features are disturbed. Manual curation of genes remains the best way to classify a sequence as a pseudogene. Although community annotation (jamboree) event was not organized in the framework of the

IWGSC, the IWGSC did establish a procedure in order to integrate curation made by different expert groups at the international level. This led to several updates: annotation releases v1.1, v1.2, and v2.1. Manual curation by experts represents 2–3% of the gene content in v2.1.

The current status with respect to wheat gene models is: 108 k HC genes, 162 k LC genes plus an additional 279 k gene fragments found by scanning for fragments of coding DNA in the unannotated part of the genome. It is clear that, with such a complicated landscape, manual curation is an endless task. However, lots could be done through bioinformatic approaches combined with manual curation in order to increase annotation quality. But even curators need information for taking decision on the most probable gene structure to consider and an open question is "which information/resources are lacking and which strategies could be useful for helping with increasing the quality of gene model predictions?".

4.7.1.1 Transcription Evidence, Gene Finders, and Homology with Related Species: Comparing A-B-D is the Most Highly Valuable Option to Improve the Quality of Structural Annotation

Finding a gene is based on three pieces of evidence: (i) a sequence is transcribed (RNASeq); (ii) a sequence shares similarity with proteins already predicted in divergent genomes; (iii) a sequence has a high probability to be protein-coding (based on hidden Markov models).

Do we miss transcript data? As early as in 2014, up to one million loci matching RNASeq data (short reads) were highlighted but even then, there were still 15% of the HC genes for which no transcription evidence was found (IWGSC 2018).

What about gene finders? The wheat genome is made of ca. 12 Gb of transposon-derived sequences while gene models represent 0.13–0.23 Gb (depending on whether or not LC genes are considered). The wheat genome is full of coding-like DNA but the very wide majority is related to TEs (transposase, reverse

transcriptase, integrase, etc.). The consequence is that the unannotated part of the genome, representing ca. 10–15% (1.5–2.0 Gb), i.e., 10 times more than the gene space, often corresponds to unidentified degenerated TEs. This means that ORFs derived from degenerated TEs are an extremely abundant source of false positive predictions for gene finders.

Detecting sequence homology with related genomes appears to us an underestimated lever of improvement. This evidence relies the evolutionary definition of a gene: an entity submitted to selection pressure. If a sequence is conserved across millions of years of evolution, we can be confident it is a gene. Predicted proteomes of *Poaceae* have been used in wheat gene modeling. However, improvements seem here obvious since there were not that many genomes available. Among the *Poaceae*, knowledge from the sequenced and annotated genomes of *O. sativa*, *Z. mays*, *S. bicolor*, *B. distachyon*, and *S. italica* were used for wheat gene modeling. They share a common ancestor with wheat between 30 and 60 MYA. Outside the *Poaceae*, fewer genes are conserved and sequence identity, even at the protein level, is low (around 55% with *Arabidopsis* for instance) which would not be of great interest to improve the annotation. Indeed, widely conserved genes are the easiest to annotate. In contrast, the challenge of annotation relies on finding genes that are specific to the *Triticeae* tribe, the *Triticum/Aegilops* genera, or even to the *T. aestivum* species. So, the most helpful resource to ensure efficient gene modeling in wheat is the *Triticeae* species, where genomes diverged 3–13 MYA, and which share high level of synteny and high level of gene sequence conservation. For instance, 88% of the predicted wheat genes (IWGSC v2.1) share on average 84% protein identity with barley predicted proteins (based on first BLAST hit alignment with thresholds 50% query overlap, 35% identity) (Mascher et al. 2017). But even TEs share sequence similarity between *Triticeae* genomes, meaning that conservation is not synonymous of selection pressure when aligning barley and wheat genomes. However, we could take advantage of the near-complete TE turnover (Wicker
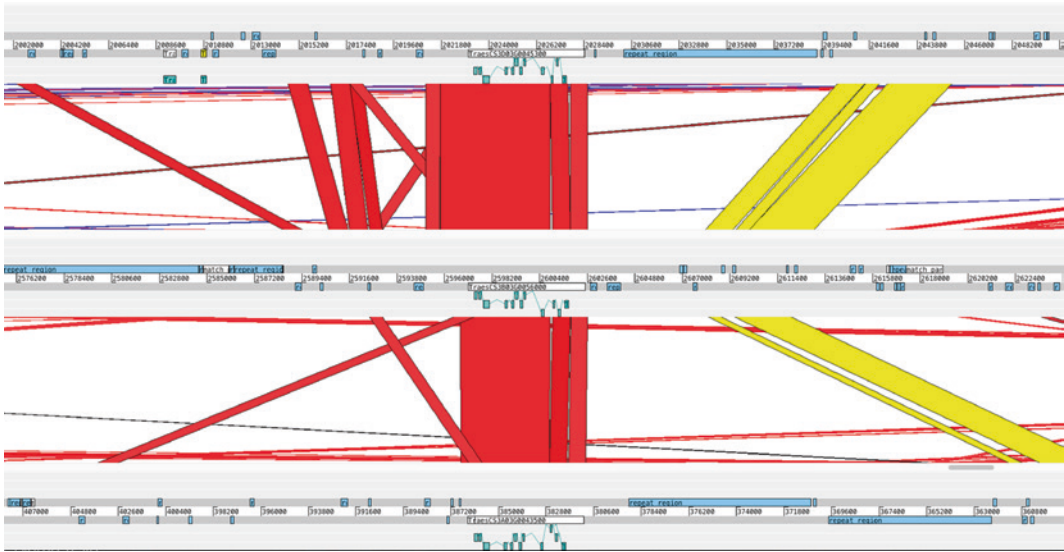
Fig. 4.1 Sequence alignments visualized with ACT (Carver et al. 2005) of three wheat homeologous regions of chromosomes 3A, 3B, and 3D. CDSs are represented in light blue, genes in white, and TEs in blue, across the six coding frames. Red blocks represent sequence conservation (> 85% identity) between A-B-D regions carrying homeologous genes and surrounding regions while TEs are not conserved between homeologous loci. Yellow blocks indicate the presence of a highly conserved unannotated sequence (neither gene nor TE) between A-B-D which strongly suggests the presence of a functional sequence subject to selection pressure that may correspond to a yet uncharacterized gene

et al. 2018) that led to erase ancestral TEs so that there are (almost) no syntenic/orthologous TEs between A, B, D (*Triticum* and *Aegilops*), H (barley; *Hordeum*), and R (rye; *Secale*) genomes. All these genomes diverged between 3 and 13 MYA, a timeframe consistent with (1) a complete TE turnover (2) within a conserved gene backbone. This is the ideal situation to identify new genes based on aligning syntenic regions. Each segment of conserved sequence between A-B-D-H-R genomes (and others) at a micro-syntenic location is evidence for selection pressure and, thus, for the presence of a gene (protein-coding or not) or a sequence involved in regulation processes called conserved non-coding sequence (CNS) as shown in Fig. 4.1.

4.7.1.2 To What Extent Sequencing More Wheat Genomes Help Improving the Reference Wheat Gene Catalog?

As explained above, the divergence window 3–13 MYA of *Triticum*, *Aegilops*, *Hordeum*, *Secale*, and others combines the advantages of a high level of gene conservation with the (almost) absence of orthologous TEs. Sequencing more *T. aestivum* genomes will not be useful in that regard. Indeed, divergence is too low so that sequence conversation is not evidence for selection pressure. Most TEs are conserved (orthologous) even between divergent accessions from the Asian and European pools, as highlighted by the Renan *versus* Chinese Spring comparison (Aury et al. 2022). However, sequencing more wheat genomes will be exploited for building the wheat pangenome.

4.7.2 De Novo Annotation *Versus* Annotation Transfer

With the advances made in sequencing technologies, assembling reference-quality wheat genome sequences is not a limit anymore (Guo et al. 2020; Walkowiak et al. 2020; Sato et al. 2021; Athiyannan et al. 2022; Aury et al. 2022). Building a wheat pangenome is thus a crucial objective in order to distinguish core *versus*

dispensable genes, especially since dispensable genes are the best candidates for adaptation to the environment, like response to specific pathogens. In contrast, core genes are enriched in essential genes, somehow not the privileged targets to search for genetic diversity controlling contrasted phenotypes.

Presence/absence (and copy number) variations of genes between two genotypes are limited to a few percent (De Oliveira et al. 2020). Using resequencing data of chromosome 3B from 20 *T. aestivum* accessions, it was shown that variable genes represent between 2 and 6% of pairwise comparisons with CHINESE SPRING. This weak percentage implies that approximations due to incomplete genome assembly and differences in gene predictions will strongly impact our capabilities to determine if a gene a really absent. Thus, an underestimated limit that prevents from accurate pangenome construction is the annotation step. Automated gene modeling is strongly dependent on the methods, tools, thresholds, used so that two annotations of the same genome are systematically different. Additionally, these differences are not only background noise. For instance, when the IWGSC RefSeq Annotation v1.0 was produced by combining independent predictions from two pipelines (TriAnnot and PGSB), 20% of each gene set did not overlap any prediction from the other one. Moreover, only 67 and 48% of TriAnnot and PGSB gene models were predicted with highly similar structures. These differences exceed largely the real presence/absence variations. The consequence is that pangenomic analyses are dependent on accurate mapping of a reference gene annotation to another assembly. This is why we believe annotation transfer tools like MAGATT (see paragraph RefSeq Annotation v2) are highly valuable in the pangenomic area as well as for maintaining improvements performed through manual curation. Eventually, in future wheat genome assemblies, genes will be transferred/ projected from a reference pangenome and de novo annotation should be restricted to specific (non-conserved) regions. Indeed, gene projection was already applied for the annotation of

chromosome pseudomolecule assemblies of 15 wheat accessions with the objective of building a wheat pangenome (Walkowiak et al. 2020). Besides the methodological challenge, issues of multiple identifiers (IDs) for a gene will become more and more problematic, as exemplified in the review of Adamski et al. (2020). Authors have highlighted the fact that one gene is already represented by many IDs, sometimes following different nomenclatures, due to the existence of multiple assemblies of the CHINESE SPRING genome sequence itself plus the release of gene models from wild wheat relatives and other cultivated genotypes. There is, thus, a strong need for integrating these data.

4.7.3 Functional Annotation: Opportunities

Automated functional annotation workflow based on sequence similarity and domain search has been established by IWGSC to assign gene ontology (GO) and function descriptions to the wheat reference gene set (IWGSC 2018). Although approaches based on local alignment search such as BLAST are straightforward and work well for certain species and gene families, the drawbacks are clear. It suffers from low sensitivity or specificity, depending on threshold choice and evolutionary distance of query gene set to species in the annotation source (Sasson et al. 2006). In addition, error or lack of robust annotation evidence in the source databases hinder or bias the large-scale functional annotation analysis, especially in non-model crop species.

To overcome these limitations, integrating various omics datasets from high-throughput experiments in combination with novel computational approaches has been considered for complementation to local sequence alignment methods, facilitating annotation of unknown genes or transferring functional knowledge from one gene to another. For example, generation and analysis of large-scale biomolecule interaction networks is a useful approach that utilizes omics data beyond gene/protein sequences. The basic idea is "guilt by association," where

a gene can be assigned a particular function if it is co-expressed with one or several genes of same known function, as the chance that they are co-regulated and needed for the same process or pathway is high (Tohge and Fernie 2012; Aoki et al. 2016). In addition to co-expression, gene–gene relationships such as protein-DNA binding and protein–protein interactions can be used to assign and transfer function from one gene to the other (Cho et al. 2016). Such interactome data can now be generated with advanced high-throughput experimental techniques such as single/bulk RNAseq, Yeast 2-Hybrid, and DNA affinity purification sequencing (DAPseq). Each type of interactome networks can be analyzed separately or in a combined manner to build multi-omics integrated network, followed by computational interpretation, from naive method of evidence aggregation to probabilistic modeling (Yu et al. 2015). The ranking or scoring reflecting proximity or connectivity of genes in the network is then used to link and transfer function from one gene to the other. Beyond the classic "single-gene" approach, integrated network-based approaches provide a more holistic view of gene function and gene–gene relationships, enabling functional annotation of unknown genes that are not related on sequence level but functionally interacted with studied genes (see also Chap. 11).

Choice of cutoff for sequence similarity-search and network mining is crucial but highly arbitrary, which can create bias or error in functional annotation process. In addition, link between various protein features (structure, text description, and interaction) and annotation label that can be utilized for functional annotation are sometimes beyond human knowledge and difficult to be revealed. In contrast, machine learning tools are suitable to identify these hidden features and assess their contribution to functions by analyzing a training set where a group of genes with these features are functionally characterized (Mahood et al. 2020). Quantitative contribution of different features learnt by computer is then exploited to predict the most possible function of unknown

genes possessing same feature types. Several tools have been developed to learn relationship between GO and heterogeneous data (text and sequence information, protein structure) and propose a predictor for annotating unknown genes (Törönen et al. 2018; You et al. 2018, 2019).

Although highly advantageous compared to classical approaches, conventional machine learning is achieved using handcrafted features. Deep learning using neural networks, on the other hand, can extract abstracted and high-level features from raw data directly and build a predictor, without human inference. The availability of omics data and computational resources allows to develop sophisticated deep learning algorithms for large-scale functional annotation. Various deep learning architectures have been built using, e.g., deep, convolutional and recurrent neural network, which have specific strength in learning different features (Cao et al. 2017; Sureyya Rifaioglu et al. 2019; Du et al. 2020). Tools built on these architectures predict GO terms either by learning protein sequence (Kulmanov and Hoehndorf 2020; Cao and Shen 2021), protein structure (Tavanaei et al. 2016; Jumper et al. 2021) or heterogenous data and networks (Cai et al. 2020; Peng et al. 2021). Several factors limit the application of deep machine learning approach for functional annotation in large-scale and unbiased manner. Firstly, although various omics and structure data are useful, only primary sequence is available for majority of unknown genes. Secondly, imbalance and incompleteness of GO database with respect to species and function categories can bias the learning step, and GO prediction task itself is a complex multi-label problem. Lastly, the quality of transferring gene model information between species that are evolutionarily distant needs to be assessed carefully. Nevertheless, despite these challenges, deep machine learning-based functional annotation and GO assignment have been successfully applied and will continue in many studies, with the support of the continuing expansion of high-quality omics and experimental datasets.

References

Adamski NM, Borrill P, Brinton J, Harrington SA, Marchal C, Bentley AR, Bovill WD, Cattivelli L, Cockram J, Contreras-Moreira B, Ford B, Ghosh S, Harwood W, Hassani-Pak K, Hayta S, Hickey LT, Kanyuka K, King J, Maccaferrri M, Naamati G, Pozniak CJ, Ramirez-Gonzalez RH, Sansaloni C, Trevaskis B, Wingen LU, Wulff BB, Uauy C (2020) A roadmap for gene functional characterisation in crops with large genomes: lessons from polyploid wheat. Elife 9

Amano N, Tanaka T, Numa H, Sakai H, Itoh T (2010) Efficient plant gene identification based on inter-species mapping of full-length cDNAs. DNA Res 17:271–279

Aoki Y, Okamura Y, Tadaka S, Kinoshita K, Obayashi T (2016) ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. Plant Cell Physiol 57:e5

Athiyannan N, Abrouk M, Boshoff WHP, Cauet S, Rodde N, Kudrna D, Mohammed N, Bettgenhaeuser J, Botha KS, Derman SS, Wing RA, Prins R, Krattinger SG (2022) Long-read genome sequencing of bread wheat facilitates disease resistance gene cloning. Nat Genet 54:227–231

Aury JM, Engelen S, Istace B, Monat C, Lasserre-Zuber P, Belser C, Cruaud C, Rimbert H, Leroy P, Arribat S, Dufau I, Bellec A, Grimbichler D, Papon N, Paux E, Ranoux M, Alberti A, Wincker P, Choulet F (2022) Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. GigaScience 11

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W (2004) Consistent over-estimation of gene number in complex plant genomes. Curr Opin Plant Biol 7:732–736

Cai Y, Wang J, Deng L (2020) SDN2GO: an integrated deep learning model for protein function prediction. Front Bioeng Biotechnol 8:391

Cao Y, Shen Y (2021) TALE: transformer-based protein function annotation with joint sequence-label embedding. Bioinformatics 37:2825–2833. https://doi.org/10.1093/bioinformatics/btab198

Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z (2017) ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. Molecules 22

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the Artemis comparison tool. Bioinformatics 21:3422–3423

Cho H, Berger B, Peng J (2016) Compact integration of multi-network topology for functional analysis of genes. Cell Syst 3:540-548.e545

Choulet F, Wicker T, Rustenholz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, Couloux A, Budak H, Breen J, Pumphrey M, Liu S, Kong X, Jia J, Gut M, Brunel D, Anderson JA, Gill BS, Appels R, Keller B, Feuillet C (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. Plant Cell 22:1686–1701

Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, Leroy P, Mangenot S, Guilhot N, Le Gouis J, Balfourier F, Alaux M, Jamilloux V, Poulain J, Durand C, Bellec A, Gaspin C, Safar J, Dolezel J, Rogers J, Vandepoele K, Aury JM, Mayer K, Berges H, Quesneville H, Wincker P, Feuillet C (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721

Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, Lipscombe J, Barker T, Lu FH, McKenzie N, Raats D, Ramirez-Gonzalez RH, Coince A, Peel N, Percival-Alwyn L, Duncan O, Trösch J, Yu G, Bolser DM, Namaati G, Kerhornou A, Spannagl M, Gundlach H, Haberer G, Davey RP, Fosker C, Palma FD, Phillips AL, Millar AH, Kersey PJ, Uauy C, Krasileva KV, Swarbreck D, Bevan MW, Clark MD (2017) An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res 27:885–896

Dale RK, Pedersen BS, Quinlan AR (2011) Pybedtools: a flexible python library for manipulating genomic datasets and annotations. Bioinformatics 27:3423–3424

Daron J, Glover N, Pingault L, Theil S, Jamilloux V, Paux E, Barbe V, Mangenot S, Alberti A, Wincker P, Quesneville H, Feuillet C, Choulet F (2014) Organization and evolution of transposable elements along the bread wheat chromosome 3B. Genome Biol 15:546

De Oliveira R, Rimbert H, Balfourier F, Kitt J, Dynomant E, Vrána J, Doležel J, Cattonaro F, Paux E, Choulet F (2020) Structural variations affecting genes and transposable elements of chromosome 3B in wheats. Front Genet 11:891

Du Z, He Y, Li J, Uversky VN (2020) DeepAdd: protein function prediction from k-mer embedding and additional features. Comput Biol Chem 89:107379

Feuillet C, Eversole K (2007) Physical mapping of the wheat genome: a coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. Israel J Plant Sci 55:307–313

Glover NM, Daron J, Pingault L, Vandepoele K, Paux E, Feuillet C, Choulet F (2015) Small-scale gene duplications played a major role in the recent evolution of wheat chromosome 3B. Genome Biol 16:188

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A,

Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gremme G, Steinbiss S, Kurtz S (2013) GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform 10:645–656

Guo W, Xin M, Wang Z, Yao Y, Hu Z, Song W, Yu K, Chen Y, Wang X, Guan P, Appels R, Peng H, Ni Z, Sun Q (2020) Origin and adaptation to high altitude of Tibetan semi-wild wheat. Nat Commun 11:5085

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol 9:R7

International Barley Genome Sequencing Consortium, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

IWGSC (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788

IWGSC (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361

Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Yang H, Liu X, He Z, Mao L, Consortium IWGS (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589

Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664

Kulmanov M, Hoehndorf R (2020) DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 36:422–429

Leroy P, Guilhot N, Sakai H, Bernard A, Choulet F, Theil S, Reboux S, Amano N, Flutre T, Pelegrin C, Ohyanagi H, Seidel M, Giacomoni F, Reichstadt M, Alaux M, Gicquello E, Legeai F, Cerutti L, Numa H, Tanaka T, Mayer K, Itoh T, Quesneville H, Feuillet C (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. Front Plant Sci 3:5

Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C, Wu H, Li Y, Cui Y, Guo X, Zheng S, Wang B, Yu K, Liang Q, Yang W, Lou X, Chen J, Feng M, Jian J, Zhang X, Luo G, Jiang Y, Liu J, Wang Z, Sha Y, Zhang B, Tang D, Shen Q, Xue P, Zou S, Wang X, Liu X, Wang F, Yang Y, An X, Dong Z, Zhang K, Luo MC, Dvorak J, Tong Y, Yang H, Li Z, Wang D, Zhang A (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90

Mahood EH, Kruse LH, Moghe GD (2020) Machine learning: a powerful tool for gene function prediction in plants. Appl Plant Sci 8:e11376

Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Staňková H, Vrána J, Chan S, Muñoz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doležel J, Ayling S, Lonardi S, Kersey P, Langridge P, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427–433

Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K (2009) TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiol 150:1135–1146

Mott R (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. Comput Appl Biosci 13:477–478

Ogihara Y, Mochida K, Kawaura K, Murai K, Seki M, Kamiya A, Shinozaki K, Carninci P, Hayashizaki Y, Shin IT, Kohara Y, Yamazaki Y (2004) Construction of a full-length cDNA library from young spikelets of hexaploid wheat and its characterization by large-scale sequencing of expressed sequence tags. Genes Genet Syst 79:227–232

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus

FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboobur R, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Paux E, Sourdille P, Salse J, Saintenac C, Choulet F, Leroy P, Korol A, Michalak M, Kianian S, Spielmeyer W, Lagudah E, Somers D, Kilian A, Alaux M, Vautrin S, Berges H, Eversole K, Appels R, Safar J, Simkova H, Dolezel J, Bernard M, Feuillet C (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. Science 322:101–104

Peng J, Xue H, Wei Z, Tuncali I, Hao J, Shang X (2021) Integrating multi-network topology for gene function prediction using deep neural networks. Brief Bioinform 22:2096–2105

Pertea G, Pertea M (2020) GFF utilities: GffRead and GffCompare. F1000Res 9

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33:290–295

Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, International Wheat Genome Sequencing C, Mayer KF, Olsen OA (2014) Genome interplay in the grain transcriptome of hexaploid bread wheat. Science 345:1250091

Pingault L, Choulet F, Alberti A, Glover N, Wincker P, Feuillet C, Paux E (2015) Deep transcriptome sequencing provides new insights into the structural and functional organization of the wheat genome. Genome Biol 16:29

Sasson O, Kaplan N, Linial M (2006) Functional annotation prediction: all for one and one for all. Protein Sci 15:1557–1562

Sato K, Abe F, Mascher M, Haberer G, Gundlach H, Spannagl M, Shirasawa K, Isobe S (2021) Chromosome-scale genome assembly of the transformation-amenable common wheat cultivar 'Fielder'. DNA Res 28

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser

W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinform 6:31

Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0. http://www.repeatmasker.org

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435-439

Sureyya Rifaioglu A, Doğan T, Jesus Martin M, Cetin-Atalay R, Atalay V (2019) DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. Sci Rep 9:7344

Tavanaei A, Maia AS, Kaniymattam A, Loganantharaj (2016) Towards recognition of protein function based on its structure using deep convolutional networks. In: IEEE international conference on bioinformatics and biomedicine (BIBM). https://doi.org/10.1109/BIBM.2016.7822509

The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

Tohge T, Fernie AR (2012) Annotation of plant gene function via combined genomics, metabolomics and informatics. J Vis Exp e3487

Törönen P, Medlar A, Holm L (2018) PANNZER2: a rapid functional annotation web server. Nucleic Acids Res 46:W84–W88

UniProt Consortium (2018) UniProt: the universal protein knowledgebase. Nucleic Acids Res 46:2699

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D (2018) Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience 7. https://doi.org/10.1093/gigascience/giy093

Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, Klymiuk V, Byrns B, Gundlach H, Bandi V, Siri JN, Nilsen K, Aquino C,

Himmelbach A, Copetti D, Ban T, Venturini L, Bevan M, Clavijo B, Koo DH, Ens J, Wiebe K, N'Diaye A, Fritz AK, Gutwin C, Fiebig A, Fosker C, Fu BX, Accinelli GG, Gardner KA, Fradgley N, Gutierrez-Gonzalez J, Halstead-Nussloch G, Hatakeyama M, Koh CS, Deek J, Costamagna AC, Fobert P, Heavens D, Kanamori H, Kawaura K, Kobayashi F, Krasileva K, Kuo T, McKenzie N, Murata K, Nabeka Y, Paape T, Padmarasu S, Percival-Alwyn L, Kagale S, Scholz U, Sese J, Juliana P, Singh R, Shimizu-Inatsugi R, Swarbreck D, Cockram J, Budak H, Tameshige T, Tanaka T, Tsuji H, Wright J, Wu J, Steuernagel B, Small I, Cloutier S, Keeble-Gagnère G, Muehlbauer G, Tibbets J, Nasuda S, Melonek J, Hucl PJ, Sharpe AG, Clark M, Legg E, Bharti A, Langridge P, Hall A, Uauy C, Mascher M, Krattinger SG, Handa H, Shimizu KK, Distelfeld A, Chalmers K, Keller B, Mayer KFX, Poland J, Stein N, McCartney CA, Spannagl M, Wicker T, Pozniak CJ (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. Trends Plant Sci 7:561–562

Wicker T, Gundlach H, Spannagl M, Uauy C, Borrill P, Ramirez-Gonzalez RH, De Oliveira R, International Wheat Genome Sequencing C, Mayer KFX, Paux E, Choulet F (2018) Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol 19:103

Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875

Yandell M, Ence D (2012) A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:329–342

You R, Huang X, Zhu S (2018) DeepText2GO: improving large-scale protein function prediction with deep semantic text representation. Methods 145:82–90

You R, Yao S, Xiong Y, Huang X, Sun F, Mamitsuka H, Zhu S (2019) NetGO: improving large-scale protein function prediction with massive network information. Nucleic Acids Res 47:W379–W387. https://doi.org/10.1093/nar/gkz388

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Li J, Liu Z, Qi Q, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Zhao W, Li P, Chen W, Zhang Y, Hu J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Tao M, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Yu G, Rangwala H, Domeniconi C, Zhang G, Zhang Z (2015) Predicting protein function using multiple kernels. IEEE/ACM Trans Comput Biol Bioinform 12:219–233

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7:203–214

Zhang D, Choi DW, Wanamaker S, Fenton RD, Chin A, Malatrasi M, Turuspekov Y, Walia H, Akhunov ED, Kianian P, Otto C, Simons K, Deal KR, Echenique V, Stamova B, Ross K, Butler GE, Strader L, Verhey SD, Johnson R, Altenbach S, Kothari K, Tanaka C, Shah MM, Laudencia-Chingcuanco D, Han P, Miller RE, Crossman CC, Chao S, Lazo GR, Klueva N, Gustafson JP, Kianian SF, Dubcovsky J, Walker-Simmons MK, Gill KS, Dvorák J, Anderson OD, Sorrells ME, McGuire PE, Qualset CO, Nguyen HT, Close TJ (2004) Construction and evaluation of cDNA libraries for large-scale expressed sequence tag sequencing in wheat (*Triticum aestivum* L.). Genetics 168:595–608

Zhu T, Wang L, Rimbert H, Rodriguez JC, Deal KR, De Oliveira R, Choulet F, Keeble-Gagnère G, Tibbits J, Rogers J, Eversole K, Appels R, Gu YQ, Mascher M, Dvorak J, Luo MC (2021) Optical maps refine the bread wheat *Triticum aestivum* cv Chinese spring genome assembly. Plant J 107:303–314

Zimin AV, Puiu D, Hall R, Kingan S, Clavijo BJ, Salzberg SL (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience 6:1–7