Wheat Data Integration and FAIRification: IWGSC, GrainGenes, Ensembl and Other Data Repositories

Michael Alaux, Sarah Dyer, Taner Z Sen

2 Wheat Data Integration and FAIRification: IWGSC, GrainGenes, Ensembl and Other Data Repositories

Michael Alaux, Sarah Dyer and Taner Z. Sen

Abstract

Wheat data integration and FAIRification are key to tackling the challenge of wheat improvement. The data repositories presented in this chapter play a central role in generating knowledge and allow data exchange and reuse. These repositories rely on international initiatives such as (i) the International Wheat Genome Sequencing Consortium (IWGSC), which delivers common genomics resources such as reference sequences, communal Web-based seminars and (ii) the Wheat Information System (WheatIS) of the Wheat Initiative (http://www.wheatis.org), which improves the interoperability and findability of the wheat data across the repositories.

Keywords

Wheat data · Data repositories · IWGSC · GrainGenes · Ensembl · FAIR

2.1 Introduction

According to the Food and Agriculture Organisation (FAO), wheat is the most widely cultivated crop on Earth, contributing about a fifth of the total calories consumed by humans (https://www.fao.org/faostat/en/#data). To meet the challenge of delivering safe, high-quality and health-promoting food and feed in an environmentally sensitive, economical and sustainable manner, it is generally considered that wheat improvement needs molecular breeding to complement more standard approaches. Furthermore, the efforts of breeding happen in a context of climate change but are still limited by insufficient knowledge and understanding of the molecular basis of central agronomic traits. In order to address the scientific questions related to this challenge, the wheat research community generates large and

M. Alaux (✉)
Université Paris-Saclay, INRAE, URGI, 78026 Versailles, France
e-mail: michael.alaux@inrae.fr

Université Paris-Saclay, INRAE, BioinfOmics, Plant Bioinformatics Facility, 78026 Versailles, France

S. Dyer
European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK
e-mail: sdyer@ebi.ac.uk

T. Z. Sen
Western Regional Research Center, Crop Improvement and Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, Albany, CA, USA
e-mail: taner.sen@usda.gov

University of California, Department of Bioengineering, Berkeley, CA, USA

R. Appels et al. (eds.), *The Wheat Genome*, Compendium of Plant Genomes,
https://doi.org/10.1007/978-3-031-38294-9_2

heterogeneous datasets. The greatest value of these data lies in their integration to generate new knowledge as a result of effective sharing to allow transparency and openness.

The wheat data landscape relies on repositories centred on (i) one or multiple data types (such as genomics, genetics or phenomics) that are highly curated and integrated with a common reference genome (e.g. the accession CHINESE SPRING developed by the IWGSC, 2018), (ii) projects or community of users with dedicated tools to mine the data. To improve the FAIRness (Findable, Interoperable, Accessible, and Reusable, Wilkinson et al. 2016a) of the wheat datasets and databases, the WheatIS expert working group of the Wheat Initiative recommended standards and developed a data discovery tool dedicated to improve the findability of wheat data across repositories (Dzale Yeumo et al. 2017; Sen et al. 2020).

In this chapter, we describe major wheat data repositories and tools, and how they integrate different types of wheat data following the FAIR principles.

2.2 IWGSC Data Repository

The International Wheat Genome Sequencing Consortium (IWGSC) has developed a variety of resources for bread wheat (*Triticum aestivum* L.) through its long-term efforts to achieve a high quality and functionally annotated reference wheat genome sequence (accession CHINESE SPRING). These data are available in a dedicated IWGSC data repository (https://wheat-urgi.versailles.inrae.fr/Seq-Repository, Alaux et al. 2018) categorised by data type as shown in Fig. 2.1.

2.2.1 Sequence Assemblies and Annotations

IWGSC wheat genome sequence assemblies are available for download, BLAST (Altschul et al. 1990), and display in genome browsers. The assembly dataset includes the draft and the reference sequences, along with their annotations.

The draft survey sequence assembly (IWGSC Chromosome Survey Sequence (CSS) v1, IWGSC 2014) and the chromosome 3B reference sequence (the first reference quality chromosome sequence obtained by the consortium, Choulet et al. 2014) were released in 2014, followed by two improved versions of the CSS (v2 and v3). The virtual gene order map generated for the CSS, the POPSEQ data were used to order sequence contigs on chromosomes (Mascher et al. 2013), and mapped marker sets were associated with these assemblies.

The reference sequence of the bread wheat genome released in 2018 (IWGSC RefSeq v1.0, 2018) included the whole genome, pseudomolecules of individual chromosomes or chromosome arms, scaffolds with the structural and functional annotation of genes, transposable elements (TEs) and non-coding RNAs. In addition, mapped markers as well as annotations supported with alignments of nucleic acid and protein evidence were made available. Manual annotations for specific gene families or regions of specific chromosomes (ca. 3685 genes) were included in the IWGSC RefSeq v1.1 annotations. This v1.1 annotation set was updated to v1.2 by integrating a set of 117 novel genes and 81 microRNAs manually curated by the wheat community following guidelines provided by IWGSC.

The improved version IWGSC RefSeq v2.1 assembly was released in 2021 (Zhu et al. 2021), which relied on whole-genome optical maps and contigs assembled from whole-genome-shotgun Pacific Biosciences (PacBio) reads (Zimin et al. 2017). Optical maps were used to detect and resolve chimeric scaffolds, anchor unassigned scaffolds, correct ambiguities in positions and orientations of scaffolds, create super-scaffolds and estimate gap sizes more accurately. PacBio contigs were used for gap closing, and pseudomolecules of the 21 CHINESE SPRING chromosomes were re-constructed to develop this new reference sequence. The corresponding IWGSC v2.1 annotation accompanying the IWGSC RefSeq v2.1 assembly was also completed. The transposable elements (TEs) in the resulting assembly IWGSC RefSeq v2.1 were

Fig. 2.1 Homepage of the IWGSC data repository hosted by the Wheat@URGI portal [Retrieved in August 2023]

reannotated, and gene annotations were updated by transferring the known gene models from previous annotations using a fine-tuned, dedicated strategy implemented in the Marker-Assisted Gene Annotation Transfer for *Triticeae* pipeline (https://forgemia.inra.fr/umr-gdec/magatt). The released IWGSC Annotation v2.1 contains 266,753 genes comprising 106,913 high-confidence genes and 159,840 low-confidence genes (Zhu et al. 2021).

In addition to the bread wheat reference sequence, the IWGSC also sequenced the genome of the Turkish bread wheat elite cultivar SONMEZ (Nelson et al. 2005) along with seven diploid and tetraploid species: *Triticum durum* cv. CAPPELLI, *Triticum durum* cv. STRONGFIELD, *Triticum durum* cv. SVEVO, *Triticum monococcum*, *Triticum urartu*, *Aegilops speltoides* and *Aegilops sharonensis* (IWGSC 2014). Download and BLAST services are available for these datasets at https://wheat-urgi.versailles.inrae.fr/Seq-Repository/Assemblies.

More broadly, the IWGSC is responsible for organising workshops and seminars and making genomics tools available to the community (https://www.wheatgenome.org/) as shown in Fig. 2.2. For example, the Apollo portal from national Australian Research Data Commons (https://apollo-portal.genome.edu.au/) has been set up to allow the curation of the IWGSC v2.1 annotation.

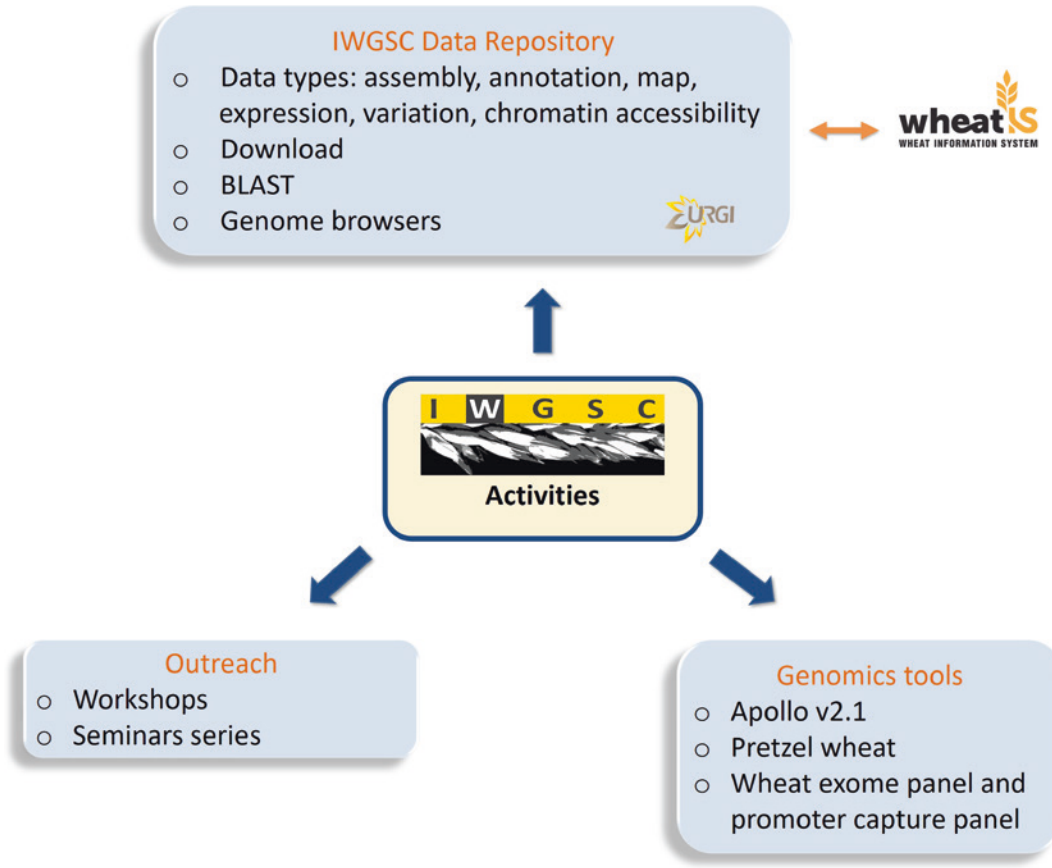

Fig. 2.2 Summary of IWGSC activities

2.2.2 Physical Maps and BAC Libraries

Physical maps of the 21 bread wheat chromosomes, based on high information content fluorescence fingerprinting (Nelson et al. 2005) or whole-genome profiling (Philippe et al. 2012) of flow-sorted chromosome or chromosome-arm specific BAC libraries, are stored and displayed in a dedicated browser. The BAC clone assemblies were produced by IWGSC members using fingerprinted contigs (Soderlund et al. 2000) or LTC (Frenkel et al. 2010) software. The positions of individual BAC clones, markers and deletion bins were mapped onto physical contigs. The wheat physical map browser also provides a link to request the BAC clones from the French plant genomic resource centre.

2.2.3 Expression Data

RNA-Seq expression data are available as read counts and transcripts per kilobase million mapped reads for the IWGSC RefSeq v1.1 annotation. A transcriptome atlas developed from 850 RNA-Seq datasets representing 32 tissues at different growth stages and stresses were mapped to the IWGSC RefSeq annotations v1.0 and v1.1 (Ramírez-González et al. 2018).

2.2.4 Variation Data

These datasets consist of the 1000 wheat exome project (He et al. 2019), whole exome capture and genotyping-by-sequencing approaches of 62 diverse wheat lines (Jordan et al. 2015)

and varietal and intervarietal SNPs (Rimbert et al. 2018). VCF data files are downloadable, and the variant calls can be displayed in the browser (https://wheat-urgi.versailles.inrae.fr/Seq-Repository/Variations).

2.2.5 Chromatin Accessibility

Using a differential nuclease sensitivity assay, the chromatin states were investigated in the coding and TE-rich repetitive regions of the allopolyploid wheat genome. Micrococcal nuclease (MNase) scores in BigWig format for IWGSC RefSeq v1.0 assembly are available to download (Jordan et al. 2020).

2.3 Wheat@URGI

The Wheat@URGI portal, developed by INRAE (French National Research Institute for Agriculture, Food and Environment) URGI unit, hosts the IWGSC data repository and GnpIS, a dedicated information system following the Findable Accessible Interoperable Reusable (FAIR) principles: https://wheat-urgi.versailles.inrae.fr/ (Alaux et al. 2018; Pommier et al. 2019).

GnpIS encompasses a set of integrated databases to manage genomic data using well-known tools such as BLAST, JBrowse, GBrowse and InterMine. An in-house database called GnpIS-coreDB developed by URGI to manage genetic and phenomic plant data, especially wheat, has been produced from French, European and international projects since 2000. A significant amount of this data is available as open access, and some project-restricted data can be obtained through a material transfer agreement.

Data managed by GnpIS-coreDB include: genetic information (markers, quantitative trait loci (QTLs), germplasm, genome-wide association studies (GWAS), genomic information (SNP discovery experiments, genotyping and synteny) and phenomic data. The phenomic data are available as whole trials including phenotypic and environmental observations on well-identified plant material provided by reference sources such as European genebanks. Detailed descriptions of these datasets are available in Alaux et al. (2018) and Pommier et al. (2019), and Table 2.1 presents a data summary.

The genetic and phenomic data have been produced from large collaborative projects such as BreedWheat (Paux et al. 2022) and Whealbi (Pont et al. 2019).

These different types of data are linked within the GnpIS information system. This integration is organised around key data, also called "pivot data" as they are pivotal objects which allow integration between data types. The key objects used to link genomic resources to genetic data are markers and traits. Markers are mapped to the genome sequences and provide information on neighbour genes and their function. They also have links to GnpIS-coreDB genetic maps, QTLs, genotyping and GWAS data. Traits link the genetic data to the phenomic data in GnpIS-coreDB and to synteny data displayed by the PlantSyntenyViewer tool (Flores et al. 2023; Pont et al. 2013).

The FAIRness of these data (including metadata) can be summarised as follows:

- Findability: (i) the data are searchable using our data discovery tools (WheatIS data discovery and FAIDARE, see below), Web interfaces (genome browsers), analysis tool (BLAST), data mining tool (WheatMine); (ii) digital object identifiers (DOIs) were generated for each accession.
- Accessibility: phenotyping data are accessible through Breeding API (BrAPI) Web services (Selby et al. 2019) and file downloads.
- Interoperability: the data are in standard formats (gff3, VCF, MCPD, MIAPPE, Papoutsoglou et al. 2020), and phenotyping data follow an ontology developed within the BreedWheat project and merged with the international wheat crop ontology (CO_321, Shrestha et al. 2012).
- Reusability: (i) all the GnpIS tools have general terms of use and licence. Open access data including code are in CC BY 4.0; (ii) the data are sufficiently described to allow their reuse in new analysis.

Table 2.1 Genetic and phenomic wheat data summary hosted in the GnpIS-coreDB database of the Wheat@URGI portal in August 2023

Data type		Total number of data points	Open access	Restricted access
Germplasm	Taxon	56	56	0
	Accession	15,031	10,448	4583
Genetic map	Map	30	29	1
	Marker	716,745	314,390	402,355
	QTL	749	465	284
Genotyping	Experiment	23	1	22
	Sample	9556	42	9543
	Marker	680,463	0	680,463
	SNP discovery	724,132	280,321	443,811
Phenotyping	Trial	895	833	62
	Seed lot	8461	5037	3653
	Variable	405	107	301
	Observation	1,488,199	602,553	885,646
GWAS	Analysis	2013	43	1970
	Sample	3096	2361	735
	Variable	313	37	279
	Marker	160,774	4109	156,665
	Association	1,014,694	48,596	966,098

2.4 GrainGenes

The GrainGenes repository (https://wheat.pw.usda.gov, Yao et al. 2022, Fig. 2.3) is a digital platform and a community service provider that has been continuously supported by U.S. congressional funds since 1992 through the U.S. Department of Agriculture. Its stakeholders are primarily global small grain research communities who work on wheat, barley, rye and oat (Blake et al. 2022). Unlike many other small grain repositories, GrainGenes has decades-worth of genetic data: GrainGenes contains rich, peer-reviewed, curated data content (Odell et al. 2017), ranging from genetic to genomic, phenotypic to traits, and people to publications, with a myriad of search and visualisation tools to enhance data findability and information discovery. GrainGenes also provides services, such as the GrainGenes email list and a Twitter feed, for small grain communities through communicating community announcements, open positions, upcoming conference information and grant opportunities.

The range of genome browsers at GrainGenes for wheat-related species attest the data growth as a result of increasingly accessible sequencing platforms, advanced assembly algorithms and annotation pipelines (https://wheat.pw.usda.gov/GG3/genome_browser). GrainGenes, in addition to IWGSC's CHINESE SPRING v1 and v2 assemblies, houses assemblies and annotations for *Aegilops longissima*, *A. speltoides*, *A. sharonensis*, five *Aegilops tauschii* accessions, wild emmer (ZAVITAN) and durum wheat SVEVO, as well as *Triticum aestivum* genomes from the 10+Wheat Genome project and the hexaploid wheat pangenome. The genome browsers at GrainGenes are shared with the Triticeae Toolbox (T3) database for the benefit of small grain researchers.

In its IWGSC CHINESE SPRING v1 genome browser, GrainGenes has many tracks overlapped with the IWGSC's data depository at Wheat@URGI and Ensembl Plants. In addition to those tracks, T3 created several tracks for variants, genome-wide association studies (GWAS), primers and quantitative trait loci (QTLs). The GrainGenes team created the guanine-quadruplex (G4) track, for this newly emergent transcription regulation element class (Cagirici and Sen 2020).

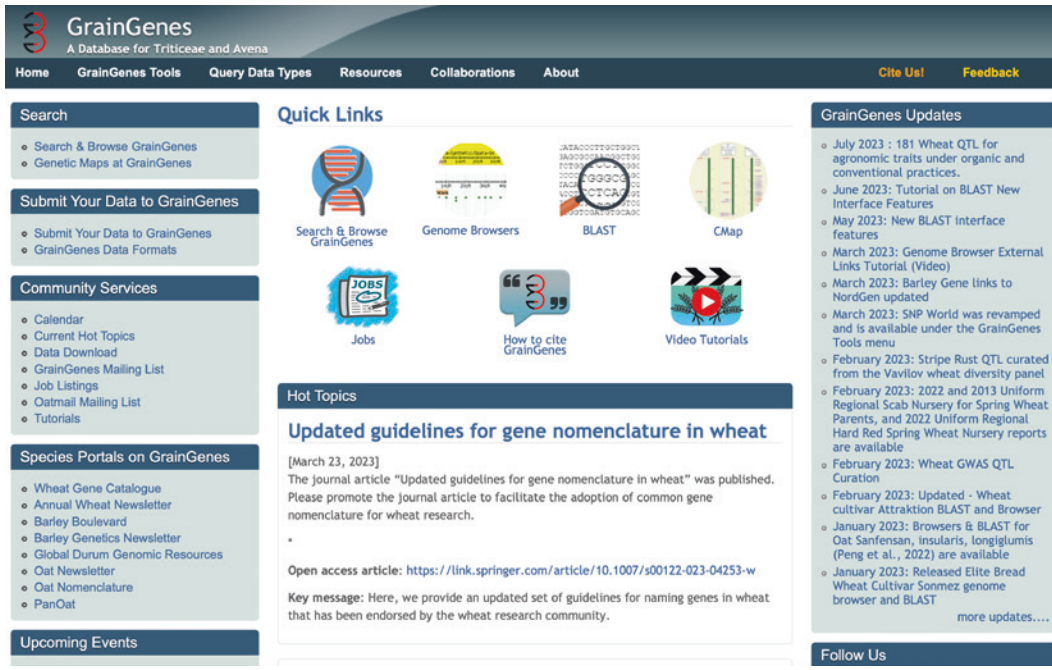

Fig. 2.3 Homepage of GrainGenes (https://wheat.pw.usda.gov) [Retrieved in July 2022]

Some of GrainGenes' genome browsers overlap with the genome browsers at other repositories such as Wheat@URGI or Ensembl Plants. This duplication of displays is not in excess, but ultimately serve the interest of small grains researchers, because having the same datasets at multiple repositories allows users to harness different tools built on top of these datasets, for example, BLAST services at GrainGenes (https://wheat.pw.usda.gov/blast/) or the Ensembl Variant Effect Predictor at Ensembl Plants.

One of the added values of using genome browsers at GrainGenes is their integration with the BLAST service at GrainGenes. When users run their nucleotide/protein sequences at GrainGenes, the results are linked to hit regions in the browsers, which allow users to go to those regions with a single mouse click. GrainGenes also uniquely allows rubber banding selection of a genome region on its JBrowse-based browsers for automatic copy pasting of underlying sequence data for subsequent BLASTing.

Those who are not familiar with genome browser operations and their relationship to other pages at GrainGenes can benefit from the several YouTube tutorial videos that were created by the GrainGenes team. This is especially useful for those who would like to learn how to jump from genomic to reach genetic data, and vice versa in GrainGenes. The videos are linked at https://wheat.pw.usda.gov/GG3/tutorials.

2.5 Ensembl Plants

The Ensembl Plants platform (https://plants.ensembl.org) provides a Web browser, databases, tools and programmatic access to integrated public genomic data for a breadth of plant species (Cunningham et al. 2022, Fig. 2.4). Ensembl Plants imports genomes and community gene annotations into the platform, annotates genomic repeat regions, imports variation data and identifies homologues via Ensembl's comparative genomics analysis pipeline. Users can access bioinformatics tools such as BLAST (Altschul et al. 1990) for sequence similarity searching or the Ensembl Variant Effect

Predictor (VEP, McLaren et al. 2016) to predict the functional consequences of variants.

The first version of the IWGSC Chromosome Survey Sequence (CSS) and gene annotation for the cultivar CHINESE SPRING was made available in Ensembl Plants in 2014. At that time there were three other triticeae genomes also included: *A. tauschii*, *Hordeum vulgare* and *T. urartu*. The TGACv1 whole-genome assembly (Clavijo et al. 2017) which used the CSS reads to assign scaffolds to chromosome arms became available via Ensembl Plants in 2015 and was subsequently replaced by the release of IWGSC RefSeq v1.0 in 2018, although all assemblies can still be accessed via Ensembl's archive sites. As of April 2023, Ensembl Plants contains an additional 17 bread wheat cultivar genomes from the 10+ project (Walkowiak et al. 2020, https://www.wheatinitiative.org/10-wheat-genome-project), making 26 triticeae genomes in total. Each of the bread wheat cultivars displays the annotation from IWGSC RefSeq v1.1 projected onto the cultivar assembly. In addition, de novo genes predicted by the Plant Genome and Systems Biology Group (PGSB) at Helmholtz, Munich and the Earlham Institute (EI) for the nine chromosome-level assemblies are also displayed.

In addition to genome annotations, Ensembl Plants also displays variation data, primarily from the 35 K and 820 K Axiom SNP breeders array, as provided by CerealsDB (Wilkinson et al. 2016b) and also EMS mutations mapped from the EMS TILLing populations (Krasileva et al. 2017) maintained by JIC's SeedStor (https://www.seedstor.ac.uk) for CADENZA (hexaploid bread wheat) and KRONOS (tetraploid durum wheat). This allows users to visualise where variants are located with respect to the IWGSC genome, and where those variants occur in the proximity of gene models the Ensembl Variant Effect Predictor will provide estimates of the likely impact of those variants on predicted gene and protein sequences. This helps users to identify those variants most likely to cause disruption to genes, and Ensembl Plants also provides a route to connect to SeedStor to order materials from the EMS populations which have those variants.

Ensembl's comparative genomics pipelines (Cunningham et al. 2019) provide gene/protein trees based on sequence homology and whole-genome alignments (WGA) between the majority of species within the platform. The IWGSC v1.0 assembly has gene trees and WGA available which allow users to explore gene family loss and expansions, identifying orthologues and paralogues and regions of synteny between genomes in Ensembl Plants. The 10+ wheat cultivars have wheat-specific gene trees available which provide a mechanism for users to explore gene conservation within the current bread wheat pan-genome (Fig. 2.5).

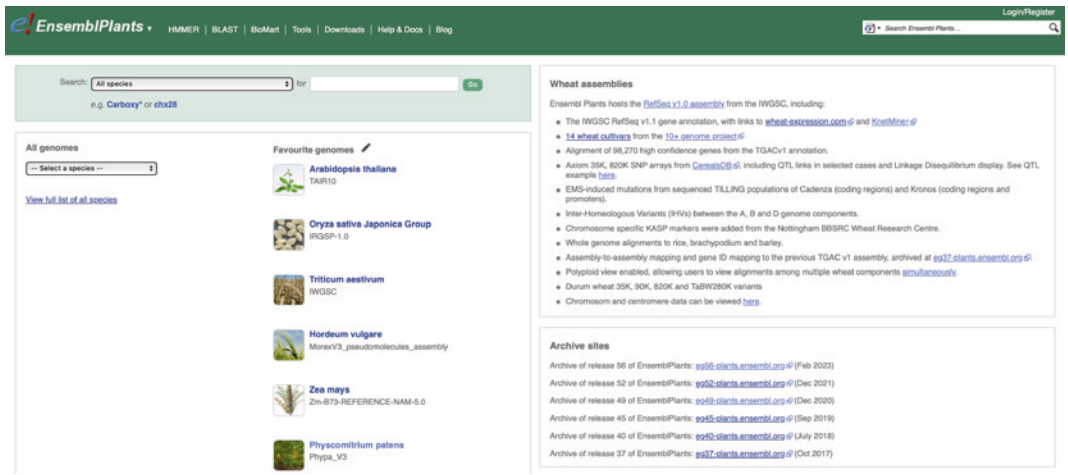

Fig. 2.4 Homepage of Ensembl Plants (https://plants.ensembl.org) [Retrieved in August 2023]

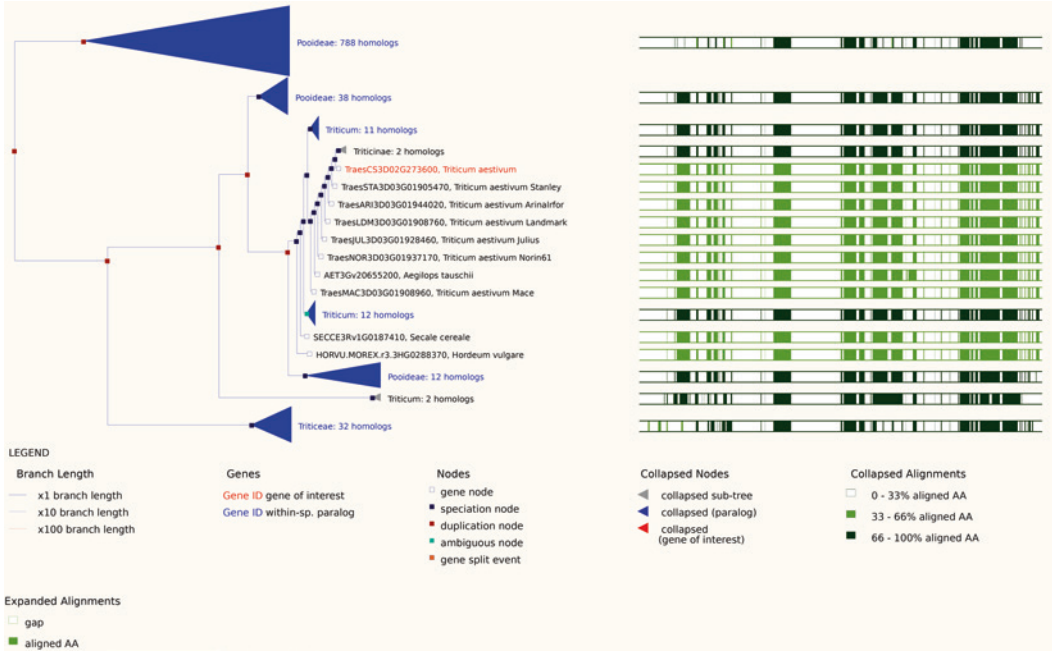

Fig. 2.5 Cultivar comparative gene tree for gene *TraesCS3D02G273600*, a heat shock protein located on chromosome 3D in IWGSC CHINESE SPRING v1.0, shown in red [Retrieved in September 2022]

Ensembl (Cunningham et al. 2022) provides user access via Web-based searches through the Ensembl browser or BioMart (which allows structured user querying to select subsets of data), FTP download access to complete sets of sequence data, annotations, gene trees and databases and programmatic access via Ensembl's APIs. All Ensembl data and tools are open access and freely available, and extensive documentation, training materials (https://training.ensembl.org) and a helpdesk are available to support user access. Ensembl Plants can also be accessed through the Gramene resource (https://www.gramene.org, Tello-Ruiz et al. 2022).

2.6 Some Other Repositories

It is beyond the purview of this chapter to provide all available wheat repositories worldwide, but the following are extremely valuable sites that we will mention briefly. Reading the publications for these repositories will be useful to learn more about their data content and features.

The Triticeae Toolbox (T3) (https://wheat.triticeaetoolbox.org, Blake et al. 2016). T3's mission is to create tools for researchers that work on genotypic–phenotypic relationships. As such, T3 played a centralised role in past projects with a strong breeding focus, such as Triticeae Coordinated Agricultural Project (TCAP) in the past, and, currently, in the Wheat Coordinated Agricultural Project (WheatCAP), both funded by the U.S. Department of Agriculture, National Institute of Food and Agriculture.

T3 houses many Web-based tools for breeders. It has capabilities that allow users to (1) upload their raw genome-wide association (GWAS) and genotype-by-sequencing datasets onto the Website, (2) perform computations such as principal component analyses and (3) visualise histograms for phenotypic observations, screen-plots of principal component eigenvalues, Q–Q plots displaying observed and expected -$\log_{10}$ *p*-values and Manhattan plots. In addition, T3 provides Web-based tools to generate selection indexes for multiple

traits simultaneously, which is a useful method for breeding programs to select and advance germplasms. As mentioned in the previous section, T3 has a very close collaboration with GrainGenes. Both databases maintain and share the same genome browsers, which enable users to go back and forth between two databases seamlessly.

Gramene (https://www.gramene.org, Tello-Ruiz et al. 2021). Gramene offers a rich data content and a wide range of tools for comparative functional genomics for 118 reference genomes and 124,010 gene family trees (Release #65, May 2022). These genomes encompass a wide range of species, including various accessions of wheat (similar to other databases discussed in this chapter). Gramene is also the home of the Plant Reactome portal (Gupta et al. 2022), which contains pathways information and gene expression displays for 106 species. Gramene has a close partnership with Ensembl Plants and displays genomes, gene models, variations and annotations collaboratively. In addition to multiple visualisation and analysis tools, such as Ensembl genome browsers, BLAST and FTP download, it also houses the Ensembl-Compara-based GeneTrees visualiser tool for sequence-based protein family classification (Vilella et al. 2009).

2.7 WheatIS Data Discovery

An expert working group of the international Wheat Initiative has built an international wheat information system, called WheatIS, with the aim of providing Web-based one-stop access to all available wheat data resources, bioinformatics tools and recommended standards (http://wheatis.org/, Dzale Yeumo et al. 2017; Sen et al. 2020). The data repositories described in this chapter are major data providers of the WheatIS federation that facilitate the availability of genomic, genetic and phenomic data to the community using a data discovery tool. This tool developed by INRAE-URGI is a search engine that indexes the metadata of each database of the federation and provides links back to the source repositories. Long-term sustainability has been achieved through a close collaboration with the ELIXIR European infrastructure for Life Science to develop a common data discovery tool usable both for WheatIS and for ELIXIR (FAIDARE, FAIR Data-finder for Agricultural REsearch, https://urgi.versailles.inrae.fr/faidare/) extended to all plants data.

Figure 2.6 and Table 2.2 present the wheat resources queried by the WheatIS data discovery tool in August 2023: https://urgi.versailles.inrae.fr/wheatis/.

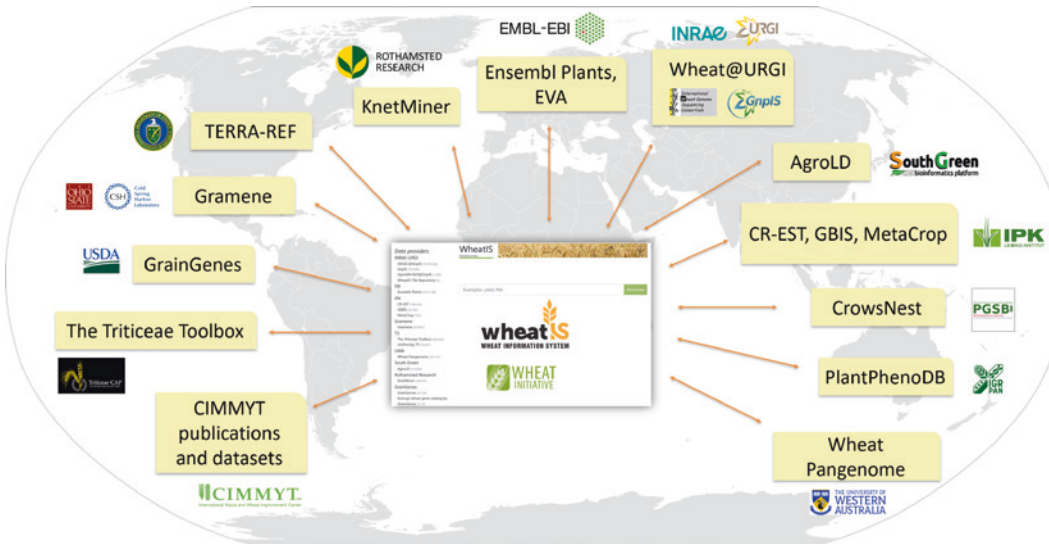

Fig. 2.6 Wheat resources queried by the WheatIS data discovery tool

Table 2.2 Number of data per wheat resource indexed by the WheatIS data discovery

Resource	Institution	Number of indexed data
TERRA-REF	U.S. Department of Energy	284
Wheat@URGI (including IWGSC data repository and GnpIS)	INRAE-URGI	19,844,409
GrainGenes (including Wheat Gene Catalogue at Komugi)	USDA-ARS	23,309
Ensembl Plants (including EVA)	EMBL-EBI	3,071,899
The Triticeae Toolbox (including UniProt)	Triticeae CAP	223,013
Gramene	CSH, OSU	229,851
AgroLD	SouthGreen	137,060
CIMMYT publications and datasets	CIMMYT	1,788
CR-EST, GBIS and MetaCrop	IPK	250,877
CrowsNest	PGSB	13,324
KnetMiner	Rothamsted Research	108,474
PlantPhenoDB	IPG PAS	6
Wheat pangenome	UWA	167,167

ARS Agricultural Research Service; *EMBL* European Molecular Biology Laboratory; *EBI* European Bioinformatics Institute; *INRAE* French National Research Institute for Agriculture, Food and Environment; *URGI* Research Unit in Genomics and Bioinformatics; *USDA* U.S. Department of Agriculture, *Triticeae CAP* Triticeae Coordinated Agricultural Product; *CSH* Cold Spring Harbor Laboratory; *OSU* Ohio State University; *CIMMYT* International Maize and Wheat Improvement Center; *IPK* Leibniz Institute of Plant Genetics and Crop Plant Research; *PGSB* Plant Genome and Systems Biology; *IPG PAS* Institute of Plant Genetics of the Polish Academy of Sciences; *UWA* University of Western Australia

2.8 Conclusion

In a context of increasingly dispersed and numerous wheat data production, the data integration and FAIRification are fundamental. The resources detailed in this chapter contribute to facilitating data discovery by helping researchers and breeders to use genetic and genomic information to improve wheat varieties. The involvement of the wheat bioinformatics community in global initiatives, such as AgBioData, ELIXIR or Research Data Alliance for an open science through standardisation, requires a long-term commitment in order to continue to contribute to research and plant breeding worldwide.

Acknowledgements The authors would like to thank the following people from Ensembl Plants: Guy Naamati, Shradha Saraf and former members Bruno Contreras-Moreira, Dan Bolser, Arnaud Kerhornou and Paul Kersey; from INRAE-URGI: Anne-Françoise Adam-Blondon, Cyril Pommier, Célia Michotey, Raphaël Flores, Nicolas Francillonne, Erik Kimmel; the GrainGenes team members.

Thanks to the International Wheat Genome Sequencing Consortium and its sponsors, the Wheat Initiative and especially the WheatIS expert working group, the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12), the following projects: BreedWheat (ANR-10-BTBR-03, France Agrimer, FSOV), Whealbi (EU FP7–613556), AGENT (European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 862613), Wheat Genomics for Sustainable Agriculture (BB/J00328X/1), Designing Future Wheat (BB/P016855/1), Elixir: the research infrastructure for life-science data and the European Molecular Biology Laboratory.

References

Alaux M et al (2018) Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. Genome Biol 19:111

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Blake VC et al (2016) The Triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding. Plant Genome 9

Blake VC, Wight CP, Yao E, Sen TZ (2022) GrainGenes: tools and content to assist breeders improving oat quality. Foods 11:914

Cagirici HB, Sen TZ (2020) Genome-wide discovery of G-quadruplexes in wheat: distribution and putative functional roles. G3(Bethesda) 10:2021–2032

Choulet F et al (2014) Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721

Clavijo BJ et al (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

Cunningham F et al (2019) Ensembl 2019. Nucleic Acids Res 47:D745–D751

Cunningham F et al (2022) Ensembl 2022. Nucleic Acids Res 50:D988–D995

Dzale Yeumo E et al (2017) Developing data interoperability using standards: a wheat community use case. F1000Res 6:1843

Flores et al (2023) SyntenyViewer: a comparative genomics-driven translational research tool. Database 2023:baad027

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A (2010) LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. BMC Bioinformatics 11:584

Gupta P et al (2022) Plant reactome and PubChem: the plant pathway and (Bio)chemical entity knowledgebases. Methods Mol Biol 2443:511–525

He F et al (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51:896–904

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

International Wheat Genome Sequencing Consortium (IWGSC) (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 361:eaar7191

Jordan KW et al (2015) A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16:48

Jordan KW, He F, de Soto MF, Akhunova A, Akhunov E (2020) Differential chromatin accessibility landscape reveals structural and functional features of the allopolyploid wheat chromosomes. Genome Biol 21:176

Krasileva KV et al (2017) Uncovering hidden variation in polyploid wheat. Proc Natl Acad Sci USA 114:E913–E921

Mascher M et al (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). Plant J 76:718–727

McLaren W et al (2016) The ensembl variant effect predictor. Genome Biol 17:122

Nelson WM et al (2005) Whole-genome validation of high-information-content fingerprinting. Plant Physiol 139:27–38

Odell SG, Lazo GR, Woodhouse MR, Hane DL, Sen TZ (2017) The art of curation at a biological database: principles and application. Curr Plant Biol 11–12:2–11

Papoutsoglou EA et al (2020) Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol 227:260–273

Paux E et al (2022) Breeding for economically and environmentally sustainable wheat varieties: an integrated approach from genomics to selection. Biology (Basel) 11:149

Philippe R et al (2012) Whole genome profiling provides a robust framework for physical mapping and sequencing in the highly complex and repetitive wheat genome. BMC Genomics 13:47

Pommier C et al (2019) Applying FAIR principles to plant phenotypic data management in GnpIS. Plant Phenomics 2019:1671403

Pont C et al (2013) Wheat syntenome unveils new evidences of contrasted evolutionary plasticity between paleo- and neoduplicated subgenomes. Plant J 76:1030–1044

Pont C et al (2019) Tracing the ancestry of modern bread wheats. Nat Genet 51:905–911

Ramírez-González RH et al (2018) The transcriptional landscape of polyploid wheat. Science 361:eaar6089

Rimbert H et al (2018) High throughput SNP discovery and genotyping in hexaploid wheat. PLoS ONE 13:e0186329

Selby P et al (2019) BrAPI-an application programming interface for plant breeding applications. Bioinformatics 35:4147–4155

Sen TZ, Caccamo M, Edwards D, Quesneville H (2020) Building a successful international research community through data sharing: the case of the wheat information system (WheatIS). F1000Res 9:536

Shrestha R et al (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the crop Ontology developed by the crop communities of practice. Front Physiol 3:326

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res 10:1772–1787

Tello-Ruiz MK et al (2021) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. Nucleic Acids Res 49:D1452–D1463

Tello-Ruiz MK, Jaiswal P, Ware D (2022) Gramene: a resource for comparative analysis of plants genomes and pathways. Methods Mol Biol 2443:101–131

Vilella AJ et al (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res 19:327–335

Walkowiak S et al (2020) Multiple wheat genomes reveal global variation in modern breeding. Nature 588:277–283

Wilkinson MD et al (2016a) The FAIR guiding principles for scientific data management and stewardship. Sci Data 3:160018

Wilkinson PA et al (2016b) CerealsDB 3.0: expansion of resources and data integration. BMC Bioinform 17:256

Yao E et al (2022) GrainGenes: a data-rich repository for small grains genetics and genomics. Database (Oxford) 2022:baac034

Zhu T et al (2021) Optical maps refine the bread wheat *Triticum aestivum* cv. Chinese Spring genome assembly. Plant J 107:303–314

Zimin AV et al (2017) The first near-complete assembly of the hexaploid bread wheat genome, *Triticum aestivum*. Gigascience 6:1–7