

UC Berkeley

UC Berkeley Previously Published Works

Title

GrainGenes: a data-rich repository for small grains genetics and genomics

Permalink

<https://escholarship.org/uc/item/334454cg>

Authors

Yao, Eric
Blake, Victoria C
Cooper, Laurel
et al.

Publication Date

2022-05-25

DOI

10.1093/database/baac034

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

GrainGenes: a data-rich repository for small grains genetics and genomics

Eric Yao^{1,2}, Victoria C. Blake^{1,3}, Laurel Cooper⁴, Charlene P. Wight⁵, Steve Michel¹, H. Busra Cagirici¹, Gerard R. Lazo¹, Clay L. Birkett⁶, David J. Waring⁷, Jean-Luc Jannink^{6,7}, Ian Holmes², Amanda J. Waters⁸, David P. Eickholt⁸ and Taner Z. Sen^{1,*}

¹United States Department of Agriculture—Agricultural Research Service, Western Regional Research Center, Crop Improvement and Genetics Research Unit, 800 Buchanan St., Albany, CA 94710, USA

²Department of Bioengineering, University of California, Stanley Hall, Berkeley, CA 94720-1762, USA

³Department of Plant Sciences and Plant Pathology, Montana State University, 119 Plant Biosciences Building, Bozeman, MT 59717, USA

⁴Department of Botany and Plant Pathology, Oregon State University, 1500 SW Jefferson Way, Corvallis, OR 97331, USA

⁵Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, 960 Carling Ave., Ottawa, ON K1A 0C6, Canada

⁶United States Department of Agriculture—Agricultural Research Service, Robert Holley Center, 538 Tower Rd., Ithaca, NY 14853, USA

⁷Section of Plant Breeding and Genetics, Cornell University, Bradfield Hall, 306 Tower Rd, Ithaca, NY 14853, USA

⁸PepsiCo R&D, 1991 Upper Buford Circle, 210 Borlaug Hall, St. Paul, MN 55108, USA

*Corresponding author: Tel: +1 (510) 559-5982; Fax: +1 (510) 559-5963; Email: taner.sen@usda.gov

Citation details: Yao, E., Blake, V.C., Cooper, L. *et al.* GrainGenes: a data-rich repository for small grains genetics and genomics. *Database* (2022) Vol. 2022: article ID baac034; DOI: <https://doi.org/10.1093/database/baac034>

Abstract

As one of the US Department of Agriculture—Agricultural Research Service flagship databases, GrainGenes (<https://wheat.pw.usda.gov>) serves the data and community needs of globally distributed small grains researchers for the genetic improvement of the Triticeae family and *Avena* species that include wheat, barley, rye and oat. GrainGenes accomplishes its mission by continually enriching its cross-linked data content following the findable, accessible, interoperable and reusable principles, enhancing and maintaining an intuitive web interface, creating tools to enable easy data access and establishing data connections within and between GrainGenes and other biological databases to facilitate knowledge discovery. GrainGenes operates within the biological database community, collaborates with curators and genome sequencing groups and contributes to the AgBioData Consortium and the International Wheat Initiative through the Wheat Information System (WheatIS). Interactive and linked content is paramount for successful biological databases and GrainGenes now has 2917 manually curated gene records, including 289 genes and 254 alleles from the Wheat Gene Catalogue (WGC). There are >4.8 million gene models in 51 genome browser assemblies, 6273 quantitative trait loci and >1.4 million genetic loci on 4756 genetic and physical maps contained within 443 mapping sets, complete with standardized metadata. Most notably, 50 new genome browsers that include outputs from the Wheat and Barley PanGenome projects have been created. We provide an example of an expression quantitative trait loci track on the International Wheat Genome Sequencing Consortium Chinese Spring wheat browser to demonstrate how genome browser tracks can be adapted for different data types. To help users benefit more from its data, GrainGenes created four tutorials available on YouTube. GrainGenes is executing its vision of service by continuously responding to the needs of the global small grains community by creating a centralized, long-term, interconnected data repository.

Database URL: <https://wheat.pw.usda.gov>

Introduction

Large scientific datasets are being created worldwide at an unprecedented rate, volume and degree of complexity. To advance biology, these large datasets have been generated using next-generation sequencers accompanied by automated processing steps that shear large DNA chromosomes into short fragments, ligate the fragments into vectors, immobilize the fragments onto a surface, amplify the DNA by polymerase chain reaction (PCR), detect signals and interpret the data. These datasets, usually called ‘raw’ as they are not yet computationally assembled into long pseudomolecules, are required by scientific journals and funding agencies to be deposited in large archival public repositories such as GenBank (1) and the

European Nucleotide Archive (ENA) (2), depending on the data types (e.g. sequencing outputs and ‘reads’ go to ENA).

After the assembly step usually come the annotation steps, where gene model positions and other genomic features (including regulatory control elements) are predicted using sophisticated algorithms, such as machine learning, that combine available data and heuristic approaches. The data are then analyzed to extract biological knowledge in the hope of producing future publications in part. These ‘processed’ datasets are preferred, but not always required, by journals or funding agencies to be deposited in public repositories and, therefore, the fate of these datasets, including genetic maps, quantitative trait loci (QTLs) and genome-wide association

study (GWAS), usually depends on the willingness of the researchers to burden themselves with formatting the data and sharing it with the larger body of researchers.

Data availability and accessibility issues are not confined to 'Big Data'. Small datasets that are meticulously obtained in research labs, in part by taking measurements one sample at a time, are valuable, but their accessibility may be limited. Most of these datasets find their way into Supplementary Materials of individual journal articles in non-standardized formats and computationally-hard-to-process file types such as Microsoft Excel, or, worse, they stay in researchers' individual computers or data drives, until they become obsolete, are forgotten or are simply thrown away. How would other researchers access these inaccessible datasets to harness them to further their own research for the benefit of the public, and, in the case of plants and animals, agriculture? The answer in part lies with community-centered databases and their curators, usually with PhD-level biological training, who gather, arrange, format and collate them after datasets are published, and computational personnel with domain expertise who create web-based interfaces and tools and manage publicly available sites on secured servers over networks.

In the biological data ecosystem, clade-oriented and model organism databases exist alongside large public repositories. In the plant systems, some well-known databases are The Arabidopsis Information Resource/Phoenix Bioinformatics (3) (for Arabidopsis), MaizeGDB (4) (maize), Planteome (5), Gramene (6) (multiple species), Soybase (7) and Legume Information System (8) (legumes). The funding sources for these repositories vary greatly. Similar to other countries, some databases in the USA are soft-funded with federal grants (usually by National Science Foundation or Department of Energy for plants, by National Institutes of Health for species that are more relevant for human health), and some are hard-funded by US congressional funds that provide long-term funding sustainability and long-term data safekeeping.

In this ecosystem, GrainGenes (9) (<https://wheat.pw.usda.gov>; Figure 1) has been hard-funded by the US Congress for over 25 years, since 1992. Its community-centric mission is serving global small grain communities by (i) acting as a centralized repository for peer-reviewed genetic, genomic, metabolic, phenotypic, transcriptomic, proteomic and trait-related data resources for genetic improvement of plants, (ii) performing community and outreach efforts through disseminating conference announcements, open positions for technicians and researchers and available assistantships for students and (iii) managing the technical infrastructure for select online resources (such as the Oat Newsletter—<https://oatnews.org>). GrainGenes provides stewardship for digital resources so that plant and especially small grains research can be facilitated and supported.

GrainGenes serves a broad range of data. Historically, and still, the repository is one of the largest centralized sites in the world for small grains genetic and genomic data and encompasses data types such as genetic markers, alleles, traits, QTLs, genetic maps and, more recently, genomic sequences and variations, structural and functional annotations, genome assemblies and pangenomes. These datasets, especially the genetic sets, have been manually curated by biologist staff members [for example (10)]. When curating and managing these datasets, GrainGenes follows the findable, accessible, interoperable and reusable principles (11) to make its datasets

more valuable to its users, which include the larger plant and agriculture community.

These datasets are visualized at GrainGenes through a wide range of web-based visualization applications that were implemented and customized for small grains. Some of these open-source applications were developed by the GrainGenes project members and disseminated worldwide for other researchers to freely download and use for their own research. GrainGenes prioritizes tool implementation at the site not based on where a given tool was developed but based on how well it will serve the small grains community broadly. As a simple example, GrainGenes recently strengthened its existing search capability by deploying the Google site search tool that not only facilitated the search at GrainGenes but also increased the visibility, discovery and accessibility of the data at GrainGenes through the popular Google search engine. For example, try googling 'graingenes qtl BYDV KxO-15-a' or 'graingenes barley yellow dwarf virus'. Even a general search of 'graingenes gene iron' could provide results (try these searches with and without quotes).

In this article, we present curated data resources, computational tools, tutorials, outreach and capabilities at GrainGenes for the benefit of small grains researchers and the broader plant community.

Genomic data integration and visualization

Similar to most biological databases [for example, MaizeGDB (12)], GrainGenes went through a genomic-centric transition that accelerated in the last couple of years. Part of this transition was the implementation of new genome browsers that display assemblies and annotations as tracks. Selection of the right genome browser (13) to serve the community of interest was critical. GrainGenes opted for the Generic Model Organism Database (GMOD) project tools, which are open-source, freely available and community-driven. Among these, GBrowse and JBrowse (14, 15) are the most popular tools for displaying genomic data. GrainGenes built all the new browsers with the JavaScript-based JBrowse version 1 because it is client-based (i.e. the computational burden is on the user), fast, uses newer technologies and relies on community development. GrainGenes has yet to make the switch to JBrowse version 2 in part because of a decision to continue supporting Basic Local Alignment Search Tool (BLAST) as a service from within JBrowse using the JBrowse Connect framework, which currently only works in JBrowse 1.

New browsers, new tracks

With the cost of sequencing decreasing and the capabilities of computational tools increasing, more and more species are being sequenced, and those sequences are assembled and annotated in many biological communities, including the small grains. Consequently, GrainGenes now hosts 51 genome browsers covering wheat, barley, rye and oat species (Supplementary Table S1). Since 2019, all the genome browsers at GrainGenes have been developed in collaboration with the Triticeae Toolbox (T3) (16), reflecting the complementary missions of both databases and the willingness to devote resources to best serve their user base. These browsers are created with the data outcomes of tremendous work and effort put in by consortiums such as the International Wheat

The screenshot shows the GrainGenes homepage with a top navigation bar containing 'Home', 'GrainGenes Tools', 'Query Data Types', 'Resources', 'Collaborations', 'About', and 'Feedback'. The main content area is divided into several sections:

- Search:** Search & Browse GrainGenes, Genetic Maps at GrainGenes.
- Submit Your Data to GrainGenes:** Submit Your Data to GrainGenes, GrainGenes Data Formats.
- Community Services:** Calendar, Current Hot Topics, Data Download, GrainGenes Mailing List, Job Listings, Oatmail Mailing List, Tutorials.
- Species Portals on GrainGenes:** Wheat Gene Catalogue, Annual Wheat Newsletter, Barley Boulevard, Barley Genetics Newsletter, Oat Newsletter, Oat Nomenclature.
- Upcoming Events:** Plant and Animal Genome Conference in San Diego, California.
- Quick Links:** Search & Browse GrainGenes, Genome Browsers, BLAST, CMap, Jobs, How to cite GrainGenes, Video Tutorials.
- Hot Topics:** Barleymap MorexV3 2021 release (October 19, 2021). Description: Barleymap (<https://floresta.eead.csic.es/barleymap>), a Web tool for mapping the position of genetic markers along the physical and genetic maps of the barley genome, has been updated and now it supports the Morex V3 genome.
- GrainGenes Updates:** December 2021: Rye Weining genome browser and BLAST are available; December 2021: Wheat Fielder genome browser and BLAST are available; October 2021: PepsiCo releases annotated gene set and associated files for OT3098 v2 genome in partnership with GrainGenes; October 2021: More MASWheat Quality Genes Curated; October 2021: Aegilops tauschii Aet v5.0 genome browser and BLAST are available; September 2021: Black awns gene class curated from Wheat Gene Catalogue; September 2021: WAPO1, a candidate 'spikelet number per spike' gene; September 2021: Polyphenol oxidase genes updated with links to MASWheat and the WGC; September 2021: Quick Links on GrainGenes' homepage are enriched; August 2021: Historical (1960-1969) hard red spring wheat performance nursery reports added; August 2021: Glume color genes curated from the WGC; August 2021: Lipoxigenase genes updated with links to MASWheat and the WGC.
- @GrainGenes Tweets:** A section for social media updates.

Figure 1. The GrainGenes' homepage (<https://wheat.pw.usda.gov>) shows a wide range of services and jump points available through an intuitive, graphic interface. The top menu buttons open up new links.

Genome Sequencing Consortium (IWGSC) (17), the Barley PanGenome Project (18) and the 10+ Wheat Project (19) among others (Figure 2).

It is important to note that a certain level of triage is already necessary to select sequenced species to create genome browsers and provide other services. Given the ease and the low cost of sequencing, assembly and annotation pipelines, triaging at biological databases will become more important with time. Triage at GrainGenes focuses on the needs of small grains communities and the level of quality of the work. In some cases, the choices are straightforward, e.g. 'first high-quality assembly of species X'. In other cases, the buy-in from a community is indicative of the importance of the work, e.g. through the establishment of a Consortium. For the in-between cases, GrainGenes relies on the guidance provided by its Advisory Board, GrainGenes Liaison Committee, that consists of scientists who are well familiar with small grains research communities.

OT3098 hexaploid oat browser

One special example among the browsers is the OT3098 hexaploid oat browser, where GrainGenes collaborated with PepsiCo and other private and public institutions to make the assembly and annotation data available through GrainGenes first for both the version 1 and version 2 genome assemblies and annotations. Although generated largely by private funds, the public can reach OT3098 data free of charge through the GrainGenes Download page (<https://wheat.pw.usda.gov/GG3/graingenes-data-downloads>) and GrainGenes genome browsers. GrainGenes also created a BLAST service,

the results of which are linked to the OT3098 hexaploid oat browsers.

Tracks

The richness of genome browser content provides users a genomic context for regions, loci or genes of interest. In the GrainGenes genome browsers, especially in more recently sequenced and assembled species, users have access to genomic sequences of pseudomolecules and structural and functional annotation of genes. Some of our genome browsers are more richly populated than the others because they usually involve large, international consortiums. One example is the wheat reference genome Chinese Spring, created by the IWGSC (17), which has functional annotation with links to Pfam (20), InterPro (21) and AmiGO (22); links to external sites such as expVIP (23) and KnetMiner (24); and several variance tracks representing the outcomes from the Wheat Target Induced Local Lesions In Genome project (25), WHEAt and barley Legacy for Breeding Improvement (26) and MNase chromatin states (27). Another example is the Morex barley reference sequence generated by the International Barley Genome Sequencing Consortium (28). The following list includes some of the tracks available on the Morex barley genome browser: Barley Infinium 9 K single nucleotide polymorphism (SNP) markers (29), Barley 50k SNP markers (30) and Hutton 50k SNP markers (30). As part of the IWGSC Chinese Spring wheat gene assembly, we have created a unique track for expression quantitative trait loci (eQTL) data (31) that has many-to-many relationships between the expression of genes and gene loci (discussed below). To be



Figure 2. Genome Browsers page at GrainGenes. Note that the 10+ Wheat Genome Project and Barley PanGenome Project buttons open up multiple single-assembly browsers.

able to represent these relationships, we created links from gene information pages to other genes (Figure 3).

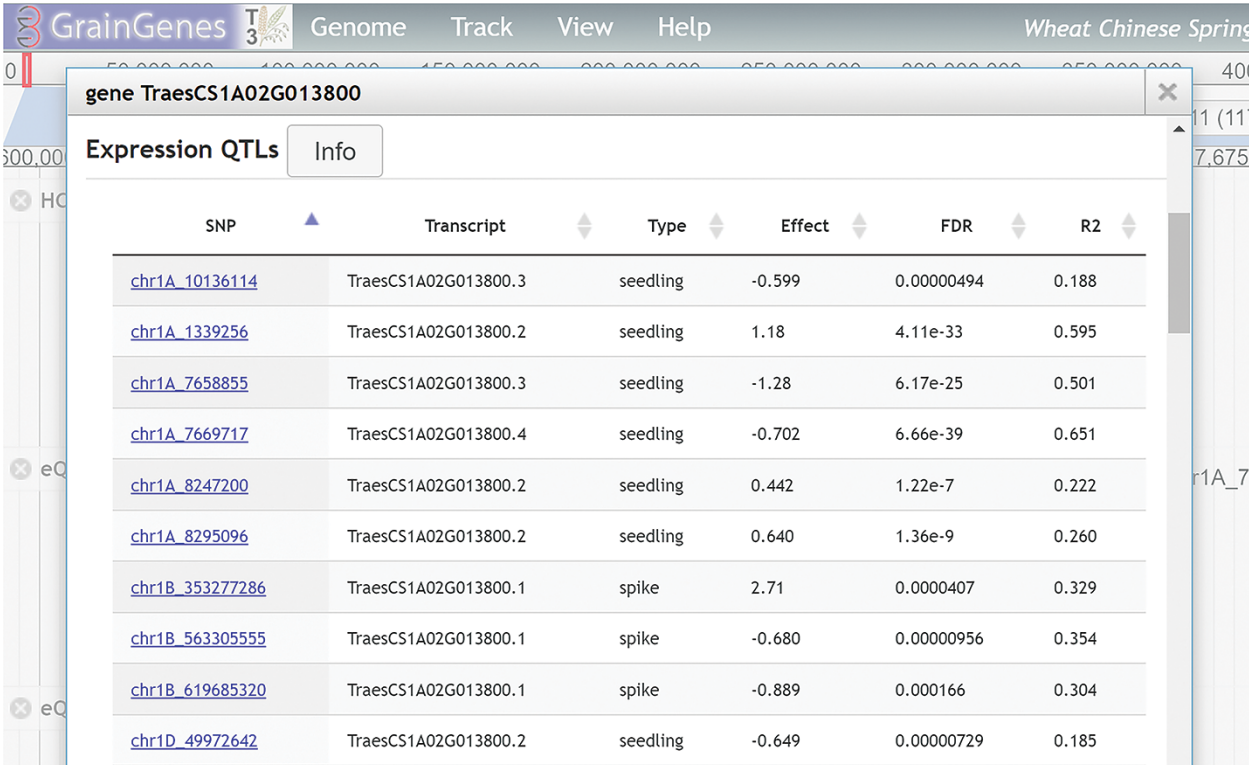
G-quadruplex tracks

G-quadruplexes (G4) are formed in guanine-rich nucleic acid sequences folded into four-stranded secondary structures. G4 structures have been strongly associated with genomic and epigenetic instability (32), DNA replication (33) and gene expression (34). In *Arabidopsis* and rice, G4 structure folding was associated with the regulation of plant development and growth (35). Based on the biophysical studies of known G4 structures (36), canonical G4 motifs were defined as at least three consecutive guanine bases separated by flexible loop regions. In GrainGenes, G4 motif tracks are now available for wheat (IWGSC Chinese Spring RefSeq v1) (37),

barley (MorexV2 and MorexV3) (38), hexaploid oat (PepsiCo OT3098 v2) and rye (Lo7 v1) (39).

Dataset project management

Browser dataset projects are managed through GitHub (<https://github.com/graingenes>) in private repositories, providing issue tracking and modification histories for the datasets. The repositories are private due to the repositories containing some system sensitive information. Each repository contains scripts for acquiring data, building datasets, building BLAST databases as well as JBrowse instances and genome browser configuration metadata. These files include scripts to (1) download raw data from their respective source locations (2), process and refine track data from FASTA, GFF3, VCF, BED files, etc. (3), build track metadata and



gene TraesCS1A02G013800

Expression QTLs Info

SNP	Transcript	Type	Effect	FDR	R2
chr1A_10136114	TraesCS1A02G013800.3	seedling	-0.599	0.00000494	0.188
chr1A_1339256	TraesCS1A02G013800.2	seedling	1.18	4.11e-33	0.595
chr1A_7658855	TraesCS1A02G013800.3	seedling	-1.28	6.17e-25	0.501
chr1A_7669717	TraesCS1A02G013800.4	seedling	-0.702	6.66e-39	0.651
chr1A_8247200	TraesCS1A02G013800.2	seedling	0.442	1.22e-7	0.222
chr1A_8295096	TraesCS1A02G013800.2	seedling	0.640	1.36e-9	0.260
chr1B_353277286	TraesCS1A02G013800.1	spike	2.71	0.0000407	0.329
chr1B_563305555	TraesCS1A02G013800.1	spike	-0.680	0.00000956	0.354
chr1B_619685320	TraesCS1A02G013800.1	spike	-0.889	0.000166	0.304
chr1D_49972642	TraesCS1A02G013800.2	seedling	-0.649	0.00000729	0.185

Figure 3. eQTL search result table in gene feature details.

configurations, build search indexes, configuration link-outs to associated external databases, banner and image graphics as such (4), process and build BLAST nucleotide and protein databases and (5) insert such configurations and process track data into our well-defined discoverable locations.

The actual raw and processed data are stored on our build/staging server, graingenes.org, not in the repositories. The raw and processed data are essentially organized on our servers in directories respective to their given repository names combined with repository scripts. We are continuously transferring and evolving new and existing scripts on our site under the GitHub project management.

As any problems, bugs, change requests and new track requests arise, they are entered as GitHub issues in their respective GitHub projects. As issues are addressed, resulting changes in scripts and files are committed and issues are closed. By utilizing this process, we keep track of how datasets were processed and what tools were used and therefore maintain a historical context.

Upon deployment, the processed data and configuration metadata are copied to our live server, wheat.pw.usda.gov, also, respectively, in their repo-named directories. Recently, we have started to use Docker (<https://www.docker.com>) command line and prebuilt dockerized bioinformatics tools to gain system independence and alleviate the issue of use of multi-versioned bioinformatics tools. Such tools include Sam Tools, Genome Tools, VCF tools, FASTX and BLAST from such Docker container collections as [biocontainers](https://biocontainers.pro), [bschiffthaler](https://bschiffthaler.com), [clinicalgenomics](https://clinicalgenomics.com) and [cjh4zavolab](https://cjh4zavolab.com). The tools are incorporated on an as-needed basis.

Genome browser framework

The genome browser framework provides a consistent look and feel for the various genome assemblies. This includes a

standardized, collapsible, informational banner for displaying details about a dataset. [Supplementary Table S1](#) shows a list of each JBrowse dataset. A dropdown genome menu within JBrowse provides access to all GrainGenes genome browsers and the framework also drives the content of the Drupal Genome Browsers page ([Figure 2](#)). The backgrounds used in the informational banners are also applied to the buttons in the Drupal Genome Browser page for cross-over consistency.

This framework also provides collapsible sub-groups for multi-genome collections, such as the Pan-barley and Wheat 10+ genomes. This serves to preserve on-screen real estate for our growing datasets while maintaining navigational consistency.

A password security system was created for browser datasets, allowing for limited access temporarily to datasets and views before the official release. Researchers are free to share their passwords with their collaborators and even with the reviewers of their manuscripts. The password protection functionality is only offered temporarily to data generators with the intent that the content, views and functionalities have been correctly built before their manuscript is published. Once a manuscript is published, GrainGenes expects the associated datasets to be shared publicly and freely.

Improved browser Linkout framework

For cross-linking with external and internal databases, we have improved the code in our datasets to link out to GrainGenes resource pages and various external databases, such as T3 (16), Pfam (20), AmiGO (22), expVIP (23), KnetMiner (24), PhyloGenes (40) and Ensembl (41). The new code is more extensible and we are moving to expand adoption throughout existing legacy datasets.

eQTL framework—special IWGSC tracks

The eQTL framework leverages GrainGenes Application Programming Interface (GGAPI) (described below) and provides a cross-reference between High Confidence v1.1 matches for Seedling and Spike eQTL tracks in wheat. A MongoDB database was created for eQTL SNP data, accessed through a RESTful interface. JBrowse plugin code results produce tables of within-feature details of the eQTLs collection of tracks, which are sortable by column. Roll-over summaries are also available on eQTL track features (Figure 3).

Pangenomes

Separate genome browsers were created for each of the genomes that were released through the 10+ Wheat and Barley PanGenome projects. For the 10+ Wheat project, only those that have full pseudomolecules were selected (10 browsers; Supplementary Table S1).

BLAST

We replaced the deprecated The National Center for Biotechnology Information (NCBI) `wwwblast` with the more advanced and customized Sequence Server (<http://sequence-server.com>) (42), which takes advantage of the latest NCBI BLAST version and also provides new visualization of results. We designed our BLAST (43) page in such a way that BLAST results can link directly to our JBrowse datasets, where available. GrainGenes introduced a new BLAST service for our wheat, barley, oat and rye collections at <https://wheat.pw.usda.gov/blast/> (see Figure 4). The new BLAST service harnesses the latest NCBI BLAST+ 2.10.0 with all our databases provided in the new version 5 database format, and processing is load-balanced by multi-threading. The new interface provides a drag-and-drop interface, a feature of multiple database selections with multiple query sequence support for multiple species. The BLAST result page shows the results of BLAST queries, identifying hit distribution and various hit details. If a dataset exists in our genome browser collection, a link will be provided to display the hit results in our JBrowse instance.

We have a collection of 120 BLAST databases, with 51 JBrowse-linked databases (where hit results have links back into our JBrowse instances). See Supplementary Table S1 for a complete list of BLAST databases.

BLAST security framework

Occasionally, we have had to secure a particular BLAST database when in stages of pre-release. A security framework was built for Sequence Server so that a database can be password protected, where the database entry is listed but is not accessible without a password.

Other computational tools

Two applications that were previously on the GrainGenes infrastructure are now accessible from the GrainGenes homepage through the Tools menu. These applications are RJPrimers (44) and BatchPrimer3 (45), which are used for transposable element junction-based PCR and other primer design strategies. Although the assemblies underlying these applications rely on the rice genome and genomic sequences of *Aegilops tauschii* from 2009, we were alerted to the fact

that some small grains researchers still find these applications useful in their research. Hence, we created links to these applications from the GrainGenes homepage.

Improved search capabilities

GrainGenes Application Programming Interface

GGAPI, the new MongoDB-based genome browser search engine, has been implemented for genome browser feature searches and is intended to provide a more robust general search function across GrainGenes report pages and GrainGenes browsers. For the browser search function, it provides a significantly smaller footprint than the legacy browser search and has a much faster indexing capability. It also captures a richer set of search terms in descriptions, accessions, GO terms and arbitrary column-9 GFF3 attributes.

GGAPI is a RESTful service that integrates with JBrowse, while also providing a mechanism to enable other applications to use the service to find features linked to a GrainGenes Genome Browser. Currently, newer genome datasets have been implemented with the new search engine, including 20 Barley PanGenomes, Rye-lo7 v3, Morex v2 and v3 and PepsiCo QT3098 v2, with more datasets adopting it moving forward.

GGAPI currently consists of four components:

1. *ggapi-jb*: The role of this component API is to serve feature search results in a package format specified in the JBrowse REST Names API. The search results are provided in JSON format indicating the location of the query feature(s), which then informs JBrowse to navigate to a desired track and location. The service, thus, replaces JBrowse's Names indexing feature and provides search results from our Mongo database. In preparation of each new JBrowse instance for a new grain species or cultivar, we use a tool in the module to harvest keys from the dataset into the Mongo database. The keys may be feature names, ids and any attribute.
2. *ggapi-eQTL*: This is a specialized component and was described in the 'eQTL Framework' section.
3. *ggapi-probes*: This API component serves to identify probes and loci that exist in GrainGenes' MySQL database. This is generally used by our JBrowse instances, in Feature Details display, to identify whether a feature in JBrowse dataset has a report page that can be linked. Otherwise, a link is not created.
4. *ggapi-proxy*: This component is the head or front-end of GGAPI, serving as a proxy to the three other API components. It provides extensibility in that a component can be dynamically loaded or unloaded without affecting other components, providing extensibility and isolation. This enables sections of the API to be in development while maintaining the stability of other sections of the API, whereas, from the presentation perspective, they are extensions of the same API.

Google search was added to search GrainGenes pages

Along with GrainGenes Class search and browsers, our users can now search the whole website with Google at <https://wheat.pw.usda.gov/cgi-bin/GG3/browse.cgi>.

GrainGenes
BLAST Service **BETA** Legacy BLAST service

Examples: Wheat Barley Oat Rye Multi Sequence

ATGGCAATGGCGTCATCGTGTGAGGTGCCGCCCTGCTCCTCACTCATCGCTTCGGCCTCGCCGTGCGCGGAGCAGCCGCGAGCAAGGCAAGCCCTGACCCGACGCCAGAACCCGCTGATCTCGTGTGAGTGTGACCG
TGCTGATCTAACTCAGCGAAATGGGGCCGCGCAGGCCACGGTGACGCGCGGATCTGGCAAGGAGTACGACTACTTCTACGACTCCGACGTGCGCCACCGGCTACGGCGTGGCCGCTGCTCCTCTGTCGCCGCGACGGCCGTG
TGCTGCGCAGCCGGTGCTCTGCTGCGGGCCGGGCTCAAGCGGGGGGCTCGCGCCCTGCGC

- databases with corresponding genome browsers

Wheat ABD Collections [Select all]

- Wheat Fielder pseudomolecules, Sato (Jun 2021)
- Wheat Chinese Spring IWGSC RefSeq v2.1 genome assembly (2021)
- 10+ Genome - ArinaLrFor v3.0 pseudomolecules (2020)
- 10+ Genome - Jagger v1.0 pseudomolecules (2020)
- 10+ Genome - Julius v1.0 pseudomolecules (2020)
- 10+ Genome - LongReach Lancer v1.0 (2020)
- 10+ Genome - CDC Landmark v1.0 pseudomolecules (2020)
- 10+ Genome - Mace v1.0 pseudomolecules (2020)
- 10+ Genome - SY Mattis v1.0 pseudomolecules (2020)
- 10+ Genome - Norin61 v1.1 pseudomolecules (2020)
- 10+ Genome - Triticum spelta P1190962 v1.0 pseudomolecules (2020)
- 10+ Genome - CDC Stanley v1.2 stanley (2020)

Wheat AB Collections [Select all]

- Wild Emmer Wheat Zavitan WEWSeq v2.0 pseudomolecules (2019)
- Triticum turgidum Durum Wheat Svevo Rel. 1.0 pseudomolecules (2019)
- Wild Emmer Wheat Zavitan WEWSeq v1.0 pseudomolecules (2017)
- Zavitan RefSeq v1 mapped gene set (Apr 2017)
- Zavitan RefSeq v1 unmapped gene set (Apr 2017)
- Triticum turgidum ssp. durum cv. Svevo pseudomolecules (Feb 2019)
- Triticum turgidum subsp. durum cv. Kronos Earham Inst. v1 scaffolds (Jan 2017)

Wheat A Collections [Select all]

- Triticum dicoccoides cv. Zavitan v1, A-genome (May 2017)
- Triticum dicoccoides cv. Zavitan RefSeq v.1.0, B-genome (May 2017)
- Triticum monococcum cv. DV92 RNA-Seq transcriptome - OSU - Jaiswal (Aug 2012)

Advanced parameters: -task blastn -evaluate 1e-5 ? **BLASTN**

Figure 4. The BLAST page allows users to enter a DNA sequence and select from our database collection.

WheatIS search

We have added a capability to search WheatIS directly from the GrainGenes Search and Browse page at <https://wheat.pw.usda.gov/cgi-bin/GG3/browse.cgi>. After entering keywords, users are then directed to the WheatIS (46) results page at <https://urgi.versailles.inrae.fr/wheatis/search>.

Curated data content

Curators at GrainGenes are experts in plant molecular biology and genetics and are dedicated to providing accurate and informative content. To that end, curators rely on peer-reviewed publications, genetics newsletters and catalogues, and direct contributions from colleagues to populate the database. Within the GrainGenes database, the most valuable resources are mapped loci, genes and QTL that can be viewed and expanded in CMap (47), the comparative mapping tool [another GMOD (<http://gmod.org>) resource]. QTL maps that are decades old can now be enriched with dense consensus maps, assisting researchers to identify genes responsible for traits of interest.

QTL mapping is generally the first step to identify breeding targets and GrainGenes currently has 203 QTL maps in 40 mapdata sets. As of October 2021, there are 6621 mapped QTL records for 253 traits. These are summarized by general category, species and specific traits in [Supplementary Table S2](#).

Topics below highlight recent curation efforts at GrainGenes:

MASWheat

As part of a mandate to link all wheat genes to the Catalogue of Gene Symbols for Wheat (<https://wheat.pw.usda.gov/GG3/wgc>) and (when available) protocols at the

Marker-Assisted Selection in Wheat (MASWheat) project (<https://maswheat.ucdavis.edu/>), gene classes were selected by the curation team and all existing data were reviewed and updated with current references and links to external projects. New gene and allele records were also created and, when complete, the work was announced on the GrainGenes Updates section of the homepage.

Stripe rust curation

Wheat rusts are pathogen-caused diseases with high economic cost. Most prevalent among them are leaf, stem and stripe rust diseases. A major curation effort to bring all stripe rust genetic content in wheat is summarized at <https://wheat.pw.usda.gov/GG3/content/october-2020-stripe-rust-update-graingenes>.

Consensus GWAS map for stripe rust resistance

GWAS results for stripe rust resistance published since 2015 from five studies on diverse wheat collections were placed on the integrated map from Maccaferri *et al.* (48). These are now available on the mapdata set Wheat, Yr Update, 2021. This map set uses the Maccaferri 2015 integrated base maps, but with the entire set of QTL reported on [Supplementary Table S2](#) for the numbered or provisionally named Yr genes. Many of the QTL were reported as simple loci with the QTL start position in 'Wheat, Yr genes and QTL'. Markers, mapped at more than one locus, were edited to include the chromosome extension. Map units for the 2021 base maps are centiMorgans. The positions in 'Wheat, Yr genes and QTL' are relative to the chromosome and reported as per cent. Both positions are reported by Maccaferri *et al.* (48). See also: 'Wheat, Yr genes and QTL' for the original

curation with 169 QTLs: <https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=mapdata&name=&id=337>.

Locus/probe pages for genome assemblies

GrainGenes database records were created for five assemblies. Records in GrainGenes contain links to the locus positions on the JBrowse genome browser as well as links to external databases [Pfam (20), AmiGO (22)] that are included in the functional annotation of the gene. Likewise, gene records on the browsers contain links to GrainGenes probe records. Assemblies with curated GrainGenes database probe records are Barley, IBSC Morex v1, 2017 (49), Barley, Morex v2 Tritex, 2019 (28), IBSC Morex v3 (50), Oat, *Avena atlantica* and *Avena eriantha* assemblies (51) and Wheat IWGSC Chinese Spring RefSeq v1.0 (17).

Axiom maps

Six map sets were from Allen *et al.* (52). All data are linked to the reference report PBJ-15-390. These maps, originally available at CerealsDB (<https://www.cerealsdb.uk.net/cerealsgenomics/CerealsDB/>), can now be viewed in the comparative map viewer CMap.

KASP markers for yield QTL on GrainGenes

Data including 15 yield-related QTLs, four maps, 26 KASP probes with primer information and germplasm were added to GrainGenes for Yang *et al.* (35). QTL mapping for grain yield-related traits in bread wheat was performed via SNP-based selective genotyping. All data are available via the mapdata page: https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=mapdata;name=Wheat-2019-KASP_YldQTL.

Oat genetic and physical maps

With collaborating partners at Agriculture and Agri-Food Canada (AAFC), a major effort to curate all important genetic maps for *Avena* spp. resulted in 117 new map sets containing 762 maps, 21 376 markers and 785 new QTLs being added since 2018. The full list of maps for diploid and hexaploid oat can be found via the ‘Search: Genetic Maps at GrainGenes’ link on the homepage or at <https://wheat.pw.usda.gov/GG3/node/876#oats6x>. In addition, the curation team added or enhanced ~300 oat germplasm records that link to T3 (<https://triticeaetoolbox.org>) and GRIN (<https://www.ars-grin.gov>) records. From these links, users can go to the T3 website and use a pedigree viewing tool for germplasm records.

Notable new map sets are discussed below:

Oat Dal x Exeter fatty acid QTL maps

Fifty QTLs for fatty acid content and agronomic traits were mapped with diversity arrays technology (DArT) on a Dal x Exeter *Avena sativa* population by (53).

See Oat-2012-Dal_x_Exeter mapping data for links to the maps and QTL: https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=mapdata;name=Oat-2012-Dal_x_Exeter;show=locus.

Oat crown rust and powdery mildew GWAS

Nine QTLs from (54) were mapped to DArT and simple-sequence repeat markers in a GWAS analysis of 177 Spanish oat cultivars and landraces. These include five for adult resistance to crown rust, three for adult resistance to powdery mildew and one for seedling powdery mildew resistance. See GrainGenes mapdata record Oat-2015-Spanish_Pc_Pm for links to qtl and single marker/qtl maps that are available for comparison with the CMap tools: https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=mapdata;name=Oat-2015-Spanish_Pc_Pm;id=480.

Oat NSGC collection GWAS maps

GWAS on 759 cultivars and landraces from the USDA National Small Grains Collection (NSGC) for *Avena* by (55) mapped 18 QTLs for agronomic traits and disease resistance. See Oat-2016-NSGC mapping data for links to the maps and QTL: <https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=mapdata;name=Oat-2016-NSGC>.

Oat Pc53 maps

Pc53 for resistance to *Puccinia coronata* (crown rust) was mapped in two oat populations by Admassu-Yimer *et al.* (56). See Oat-2018-Pc53 mapping data for links to the maps: <https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=mapdata;name=Oat-2018-Pc53>.

Curation of the WGC

The Catalogue of Gene Symbols for Wheat, also known as the WGC, is the product of more than 50 years of curation of wheat genetic information by a group of expert wheat researchers, led by Robert A. McIntosh (<https://wheat.pw.usda.gov/GG3/WGC>).

The WGC is organized based on the following categories: Morphological and Physiological Traits, Proteins and Pathogenic Disease/Pest Reactions and contains information about wheat genes and gene classes, gene nomenclature, loci, alleles, QTLs, mapping and germplasm information and laboratory designations for markers, along with the associated references. The goal of the WGC is to provide helpful information about wheat genetics to a wide range of users, from researchers at the lab bench to farmers in the field.

Full editions of the WGC are compiled periodically (the last full version published in 2013), with yearly supplements. KOMUGI, the Wheat Genetics Resources Database of Japan (<https://shigen.nig.ac.jp/wheat/komugi/>), currently hosts an electronic version of the Catalogue and supplements up through 2017 (<https://shigen.nig.ac.jp/wheat/komugi/genes/symbolClassList.jsp>). GrainGenes has a dedicated page for the WGC at <https://wheat.pw.usda.gov/GG3/WGC> with links to the WGC documents from 1999 to 2020. In addition, the GrainGenes database contains reference records for each of the WGC editions and recent supplements (<https://wheat.pw.usda.gov/cgi-bin/GG3/report.cgi?class=journal;name=Catalogue%20of%20Gene%20Symbols%20for%20Wheat;id=281>).

Starting in August 2019, GrainGenes curators began curating the data in the WGC into the interactive MySQL

database at GrainGenes. This work included updating existing information at GrainGenes and adding new gene classes, genes, loci, maps, references and germplasm records along with links to the KOMUGI database for gene class, gene and allele records. In addition, new information from recently published peer-reviewed publications is included. The curation team at GrainGenes submits a report of the curation work to the WGC committee on a quarterly basis to report any discrepancies observed in the information and also any new information added. Updates for this ongoing curation effort are available at <https://wheat.pw.usda.gov/GG3/node/824>.

A summary table of WGC curation can be found in [Supplementary Table S3](#).

Contribution to Wheat Information System

Since 2017, GrainGenes has collaborated with the WheatIS (46) (wheatis.org) and the personnel at Unité de Recherche Génomique Info in France. Operating under the Wheat Initiative, WheatIS is a platform that provides a single hub of access through a common API to the wheat data that are distributed among the small grains databases worldwide. A shortcut to all GrainGenes data at WheatIS can be found at <https://urgi.versailles.inra.fr/wheatis/#result/term=graingenes>. Through its local node, GrainGenes indexed:

- *Germplasm*. 16 106 germplasm records including lines from the Global Tetraploid Wheat Collection, and all other diploid, tetraploid and hexaploid accessions in GrainGenes.
- *QTLs*. 3925 QTLs, of which 3408 are new, and the previous 548 have enhanced descriptions.
- *Genetic Maps*. 147 genetic maps, of which 56 are new.
- *Physical Maps*. 12 physical maps, of which two are new.
- *WGC*. As a service to the small grains community, 3119 genes from the WGC point to specific gene pages at the KOMUGI database in Japan.

Nursery reports

Uniform regional nursery data

The following reports were uploaded to GrainGenes and can be found at <https://wheat.pw.usda.gov/GG3/germplasm>:

2019 Uniform Regional Scab Nursery for Spring Wheat Parents.

2020 Uniform Regional Hard Red Spring Wheat Nursery reports.

2020 Uniform Regional Scab Nursery for Spring Wheat Parents.

Mississippi valley barley nursery reports

We have added several Mississippi Valley Barley Nursery Reports covering the years between 1999 and 2014 for agronomic traits and malting quality. These reports can be reached on our Germplasm page following this link: <https://wheat.pw.usda.gov/GG3/germplasm#barley>.

Service to the small grains community

GrainGenes is not only a centralized repository for small grains data but also a digital platform for other repositories and community newsletters:

- *Annual Wheat Newsletter*. GrainGenes hosts the Annual Wheat Newsletter issues published since the 37th issue in 1991 (<https://wheat.pw.usda.gov/ggpages/awn/>). Wheat Annual Newsletter Volume 67 (2021) was recently made available.
- *Barley Genetics Newsletter*. GrainGenes has hosted the Barley Genetics Newsletter issues since the first issue in 1971 (<https://wheat.pw.usda.gov/ggpages/bgn/>). The most recent Barley Genetics Newsletters, v48 (2018) and v49 (2019–2020), created by Phil Bregitzer and Udda Lundqvist, are now available at GrainGenes.
- *Oat Newsletter*. We host a site for the Oat Newsletter that has been published since 1950. The Newsletter was converted into a website where it not only hosts research reports and old issues but also blog entries by oat researchers (<https://oatnews.org>).
- *USDA-ARS Small Grains Genotyping Labs*. This is the website describing the four ARS genotyping labs in the USA. They use the site for general information, links and contact information (<https://wheat.pw.usda.gov/GenotypingLabs/>).
- *Grains email list*. The grains mailing list consists of 1057 members working on wheat, barley, oat and related species. It has operated since 1992. Messages are curated by GrainGenes personnel.
- *OatMail email list*. OatMail is a new email list that consists of 90 members working on oat. Messages to the OatMail list are curated by Charlene Wight and GrainGenes personnel. OatMail is jointly administered with the Oat Newsletter and AAFC Canada to facilitate communication between members of the oat research community.

Outreach and training

Since GrainGenes' mission is to serve small grains researchers, interacting with them through scientific conferences to learn their priorities and concerns and keep up with the ever-changing research landscape is a priority. Consequently, GrainGenes personnel attend many conferences in person (although due to the coronavirus disease 2019 pandemic, there has been a hiatus) to give talks on GrainGenes. In addition, GrainGenes personnel give guest lectures at workshops, as well as in undergraduate and graduate classes. We have presented at the University of California, Davis, the University of California, Berkeley, and Montana State University. To reach a geographically broader audience, GrainGenes personnel also create training videos and upload them to the 'GrainGenes Official' channel on YouTube. Since 2019, the following four tutorials were created and made available on YouTube as well as at GrainGenes (<https://wheat.pw.usda.gov/GG3/tutorials/>):

- 'Navigating between GrainGenes Database Records and Genome Browsers'.
- 'Saving Information from Genome Browsers in GrainGenes'.
- 'Using CMap in GrainGenes to Improve Marker Density around a Gene of Interest'.
- 'How to submit data to GrainGenes'.

Collaborations

GrainGenes is actively involved in both the Steering Committee and Working Groups of the AgBioData Consortium

(57) (<https://www.agbiodata.org>) that was formed by agricultural databases to develop solutions to common issues for biological repositories such as developing data/metadata standards, gene naming nomenclatures and data sharing practices. Another collaboration to which GrainGenes is actively contributing is the WheatIS (46), operating under the International Wheat Initiative (wheatis.org). GrainGenes also works with Ag Data Commons (<https://data.nal.usda.gov>), which is under the National Agricultural Library of the US Department of Agriculture, to house, organize and make data accessible to the public.

Conclusion

GrainGenes is the flagship repository of the US Department of Agriculture for genetics and genomics data for small grains. Although it is primarily funded by hard funds apportioned by the US Congress, its mission is to serve the global small grains research communities for the improvement of agriculture, crop development and contribution to food production.

GrainGenes houses small grains data for Triticeae and *Avena* species, such as wheat, barley, rye and oat. It has a wide range of data types, including genetic (genetic markers, maps and QTLs), genomic (assemblies, functional and structural annotations and genomic elements) and ontology datasets. Community outreach is part of GrainGenes' mission, so users also have access to tutorials, job boards, meeting announcements and information about community initiatives.

In the future, we expect not only more assemblies and annotations for more small grains species, but also pangenomic and phenotypic, metabolic pathway, protein-protein network and, with the recent watershed publication of DeepMind AlphaFold (58), protein structural data. The access to standardized data and the ability to search, query and extract data/metadata information will be an ongoing challenge, along with the bigger challenge of reaching information across separate databases. Initiatives such as AgBioData and WheatIS are steps in the right direction to find solutions to problems common among biological databases and apply those solutions for the benefit of a global user base.

In this article, we presented an update of the genomic and genetic data resources, computational tools, improvements to infrastructure and web interface, and outreach efforts at GrainGenes that small grains researchers and the larger plant community can access and use for crop improvement to respond to the food needs of the rising population worldwide within the context of the environmental uncertainties associated with climate change.

Supplementary data

Supplementary data are available at *Database* Online.

Acknowledgements

GrainGenes is supported by the United States Department of Agriculture—Agricultural Research Service (USDA-ARS) under the CRIS project 2030-21000-024-00D. Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. USDA is an equal

opportunity provider and employer. GrainGenes benefits significantly from the oversight provided by the members of the GrainGenes Liaison Committee: Mark Sorrells (Chair), Jorge Dubcovsky, Catherine Howarth, Kevin Smith and Roger Wise. The opinions in this study are those of the authors and do not necessarily represent the opinions or policies of PepsiCo, Inc. All URLs were accessed in November 2021.

Funding

United States Department of Agriculture—Agricultural Research Service (2030-21000-024-00D).

Conflict of interest

None declared.

References

1. Sayers, E.W., Cavanaugh, M., Clark, K. *et al.* (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
2. Harrison, P.W., Ahamed, A., Aslam, R. *et al.* (2021) The European Nucleotide Archive in 2020. *Nucleic Acids Res.*, **49**, D82–D85.
3. Reiser, L., Subramaniam, S., Li, D. *et al.* (2017) Using The Arabidopsis Information Resource (TAIR) to find information about Arabidopsis genes. *Curr. Protoc Bioinform.*, **60**, 1–11.
4. Portwood, J.L., 2nd, Woodhouse, M.R., Cannon, E.K. *et al.* (2019) MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.*, **47**, D1146–D1154.
5. Cooper, L., Meier, A., Laporte, M.A. *et al.* (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Res.*, **46**, D1168–D1180.
6. Tello-Ruiz, M.K., Naithani, S., Gupta, P. *et al.* (2021) Gramene 2021: harnessing the power of comparative genomics and pathways for plant research. *Nucleic Acids Res.*, **49**, D1452–D1463.
7. Brown, A.V., Connors, S.I., Huang, W. *et al.* (2021) A new decade and new data at SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **49**, D1496–D1501.
8. Dash, S., Campbell, J.D., Cannon, E.K. *et al.* (2016) Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family. *Nucleic Acids Res.*, **44**, D1181–D1188.
9. Blake, V.C., Woodhouse, M.R., Lazo, G.R. *et al.* (2019) GrainGenes: centralized small grain resources and digital platform for geneticists and breeders. *Database (Oxford)*, **2019**, 1–7.
10. Odell, S.G., Lazo, G.R., Woodhouse, M.R. *et al.* (2017) The art of curation at a biological database: principles and application. *Curr. Plant Biol.*, **11–12**, 2–11.
11. Gundersen, S., Boddu, S., Capella-Gutierrez, S. *et al.* (2021) Recommendations for the FAIRification of genomic track metadata. *F1000Res*, **10**, 1–22.
12. Woodhouse, M.R., Cannon, E.K., Portwood, J.L., 2nd *et al.* (2021) A pan-genomic approach to genome databases using maize as a model system. *BMC Plant Biol.*, **21**, 385.
13. Sen, T.Z., Harper, L.C., Schaeffer, M.L. *et al.* (2010) Choosing a genome browser for a Model Organism Database: surveying the maize community. *Database (Oxford)*, **2010**, baa007.
14. Buels, R., Yao, E., Diesh, C.M. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
15. Skinner, M.E., Uzilov, A.V., Stein, L.D. *et al.* (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
16. Blake, V.C., Birkett, C., Matthews, D.E. *et al.* (2016) The Triticeae Toolbox: combining phenotype and genotype data to advance small-grains breeding. *Plant Genome*, **9**, 1–10.

17. International Wheat Genome Sequencing, C., investigators, I.R.P., Appels, R. *et al.* (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, 361, 1–13.
18. Jayakodi, M., Padmarasu, S., Haberer, G. *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, 588, 284–289.
19. Walkowiak, S., Gao, L., Monat, C. *et al.* (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature*, 588, 277–283.
20. Mistry, J., Chuguransky, S., Williams, L. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, 49, D412–D419.
21. Blum, M., Chang, H.Y., Chuguransky, S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, 49, D344–D354.
22. Munoz-Torres, M. and Carbon, S. (2017) Get GO! retrieving GO data using AmiGO, QuickGO, API, files, and tools. *Methods Mol. Biol.*, 1446, 149–160.
23. Adams, T.M., Olsson, T.S.G., Ramirez-Gonzalez, R.H. *et al.* (2021) Rust expression browser: an open source database for simultaneous analysis of host and pathogen gene expression profiles with expVIP. *BMC Genomics*, 22, 166.
24. Hassani-Pak, K., Singh, A., Brandizi, M. *et al.* (2021) KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. *Plant Biotechnol. J.*, 19, 1670–1678.
25. Krasileva, K.V., Vasquez-Gross, H.A., Howell, T. *et al.* (2017) Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.*, 114, E913–E921.
26. Pont, C., Leroy, T., Seidel, M. *et al.* (2019) Tracing the ancestry of modern bread wheats. *Nat. Genet.*, 51, 905–911.
27. Jordan, K.W., He, F., de Soto, M.F. *et al.* (2020) Differential chromatin accessibility landscape reveals structural and functional features of the allopolyploid wheat chromosomes. *Genome Biol.*, 21, 176.
28. Monat, C., Padmarasu, S., Lux, T. *et al.* (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biol.*, 20, 284.
29. Comadran, J., Kilian, B., Russell, J. *et al.* (2012) Natural variation in a homolog of *Antirrhinum* CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat. Genet.*, 44, 1388–1392.
30. Bayer, M.M., Rapazote-Flores, P., Ganai, M. *et al.* (2017) Development and evaluation of a barley 50k iSelect SNP array. *Front Plant Sci.*, 8, 1792.
31. Fauteux, F., Wang, Y., Rocheleau, H. *et al.* (2019) Characterization of QTL and eQTL controlling early *Fusarium graminearum* infection and deoxynivalenol levels in a Wuhan 1 x Nyubai doubled haploid wheat population. *BMC Plant Biol.*, 19, 536.
32. Guilbaud, G., Murat, P., Recolin, B. *et al.* (2017) Local epigenetic reprogramming induced by G-quadruplex ligands. *Nat. Chem.*, 9, 1110–1117.
33. Lopes, J., Piazza, A., Bermejo, R. *et al.* (2011) G-quadruplex-induced instability during leading-strand replication. *EMBO J.*, 30, 4033–4046.
34. Reina, C. and Cavalieri, V. (2020) Epigenetic modulation of chromatin states and gene expression by G-quadruplex structures. *Int. J. Mol. Sci.*, 21, 1–22.
35. Yang, X., Cheema, J., Zhang, Y. *et al.* (2020) RNA G-quadruplex structures exist and function in vivo in plants. *Genome Biol.*, 21, 226.
36. Huppert, J.L. and Balasubramanian, S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, 33, 2908–2916.
37. Cagirici, H.B. and Sen, T.Z. (2020) Genome-wide discovery of G-quadruplexes in wheat: distribution and putative functional roles. *G3 (Bethesda)*, 10, 2021–2032.
38. Cagirici, H.B., Budak, H. and Sen, T.Z. (2021) Genome-wide discovery of G-quadruplexes in barley. *Sci. Rep.*, 11, 7876.
39. Rabanus-Wallace, M.T., Hackauf, B., Mascher, M. *et al.* (2021) Chromosome-scale genome assembly provides insights into rye biology, evolution and agronomic potential. *Nat. Genet.*, 53, 564–573.
40. Zhang, P., Berardini, T.Z., Ebert, D. *et al.* (2020) PhyloGenes: An online phylogenetics and functional genomics resource for plant gene function inference. *Plant Direct*, 4, e00293.
41. Howe, K.L., Contreras-Moreira, B., De Silva, N. *et al.* (2020) Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.*, 48, D689–D695.
42. Priyam, A., Woodcroft, B.J., Rai, V. *et al.* (2019) Sequenceserver: a modern graphical user interface for custom BLAST databases. *Mol. Biol. Evol.*, 36, 2922–2924.
43. Altschul, S.F., Gish, W., Miller, W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, 403–410.
44. You, F.M., Wanjugi, H., Huo, N. *et al.* (2010) RJPrimers: unique transposable element insertion junction discovery and PCR primer design for marker development. *Nucleic Acids Res.*, 38, W313–320.
45. You, F.M., Huo, N., Gu, Y.Q. *et al.* (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinform.*, 9, 253.
46. Sen, T.Z., Caccamo, M., Edwards, D. *et al.* (2020) Building a successful international research community through data sharing: the case of the Wheat Information System (WheatIS). *F1000Res*, 9, 536.
47. Youens-Clark, K., Faga, B., Yap, I.V. *et al.* (2009) CMap 1.01: a comparative mapping application for the Internet. *Bioinformatics*, 25, 3040–3042.
48. Maccaferri, M., Zhang, J., Bulli, P. *et al.* (2015) A genome-wide association study of resistance to stripe rust (*Puccinia striiformis* f. sp. tritici) in a worldwide collection of hexaploid spring wheat (*Triticum aestivum* L.). *G3 (Bethesda)*, 5, 449–465.
49. Beier, S., Himmelbach, A., Colmsee, C. *et al.* (2017) Construction of a map-based reference genome sequence for barley, *Hordeum vulgare* L. *Sci. Data*, 4, 170044.
50. Mascher, M., Wicker, T., Jenkins, J. *et al.* (2021) Long-read sequence assembly: a technical evaluation in barley. *Plant Cell*, 33, 1888–1906.
51. Maughan, P.J., Lee, R., Walstead, R. *et al.* (2019) Genomic insights from the first chromosome-scale assemblies of oat (*Avena* spp.) diploid species. *BMC Biol.*, 17, 92.
52. Allen, A.M., Winfield, M.O., Burrige, A.J. *et al.* (2017) Characterization of a Wheat Breeders' Array suitable for high-throughput SNP genotyping of global accessions of hexaploid bread wheat (*Triticum aestivum*). *Plant Biotechnol. J.*, 15, 390–401.
53. Hizbai, B.T., Gardner, K.M., Wight, C.P. *et al.* (2012) Quantitative trait loci affecting oil content, oil composition, and other agronomically important traits in oat. *Plant Genome*, 5, 164–175.
54. Montilla-Bascón, G., Rispaill, N., Sánchez-Martín, J. *et al.* (2015) Genome-wide association study for crown rust (*Puccinia coronata* f. sp. avenae) and powdery mildew (*Blumeria graminis* f. sp. avenae) resistance in an oat (*Avena sativa*) collection of commercial varieties and landraces. *Front Plant Sci.*, 6, 103.
55. Winkler, L.R., Michael Bonman, J., Chao, S. *et al.* (2016) Population structure and genotype-phenotype associations in a collection of oat landraces and historic cultivars. *Front Plant Sci.*, 7, 1077.
56. Admassu-Yimer, B., Bonman, J.M. and Esvelt Klos, K. (2018) Mapping of crown rust resistance gene Pc53 in oat (*Avena sativa*). *PLoS One*, 13, e0209105.
57. Harper, L., Campbell, J., Cannon, E.K.S. *et al.* (2018) AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database (Oxford)*, 2018, 1–32.
58. Varadi, M., Anyango, S., Deshpande, M. *et al.* (2021) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, 50, D439–D444.