Title

The art of curation at a biological database: Principles and application

Authors

Odell, Sarah G

Lazo, Gerard R

Woodhouse, Margaret R

et al.

Publication Date

2017-09-01

DOI

10.1016/j.cpb.2017.11.001

Current Plant Biology 11–12 (2017) 2–11

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

The art of curation at a biological database: Principles and application☆

Sarah G. Odell[a,b], Gerard R. Lazo[a], Margaret R. Woodhouse[a], David L. Hane[a], Taner Z. Sen[a,c,*]

[a] U.S. Department of Agriculture – Agricultural Research Service, Crop Improvement and Genetics Research Unit, Albany, CA 94710, United States
[b] University of California, Department of Plant Sciences, Davis, CA 95616, United States
[c] Iowa State University, Department of Genetics, Development, and Cell Biology, Ames, IA 50011, United States

ARTICLE INFO

Keywords:
Biological databases
Curation
Genetic markers
Genetic maps
Genomic data
Genome browsers

ABSTRACT

The variety and quantity of data being produced by biological research has grown dramatically in recent years, resulting in an expansion of our understanding of biological systems. However, this abundance of data has brought new challenges, especially in curation. The role of biocurators is in part to filter research outcomes as they are generated, not only so that information is formatted and consolidated into locations that can provide long-term data sustainability, but also to ensure that the relevant data that was captured is reliable, reusable, and accessible. In many ways, biocuration lies somewhere between an art and a science. At GrainGenes (https://wheat.pw.usda.gov;https://graingenes.org), a long-time, stably-funded centralized repository for data about wheat, barley, rye, oat, and other small grains, curators have implemented a workflow for locating, parsing, and uploading new data so that the most important, peer-reviewed, high-quality research is available to users as quickly as possible with rich links to past research outcomes. In this report, we illustrate the principles and practical considerations of curation that we follow at GrainGenes with three case studies for curating a gene, a quantitative trait locus (QTL), and genomic elements. These examples demonstrate how our work allows users, i.e., small grains geneticists and breeders, to harness high-quality small grains data at GrainGenes to help them develop plants with enhanced agronomic traits.

1. Introduction

The value of a biological database is largely defined by the breadth and accuracy of its content. If the content is becoming limited and inaccurate, a database would steadily lose its value for its users, and will eventually become obsolete. The data coverage and accuracy of a database need to be therefore continuously enhanced, and a primary way of accomplishing this goal is through a critical process called biological curation, i.e., extracting biological data from scientific literature and integrating it into a biological database. Curators, usually combining computational skills with PhD-level biological expertise, peruse peer-reviewed scientific articles, extract data sets that they judge to be the most useful for their user base, and integrate them into a back-end database, so that these data sets can be displayed through a web interface. Because curators apply a set of subjective criteria and the extracted data sets need to be integrated into specific databases with different contexts and focus, the curated content from the same journal article can sometimes be curated slightly differently at different biological repositories (even for plant databases with similar user bases, such as GrainGenes (https://wheat.pw.usda.gov; https://graingenes.org) [1], TAIR [2], MaizeGDB [3], Gramene [4], Sol Genomics Network [5], and Soybase [6]. Yet, curators follow similar routes, workflows, and principles in curating biological data. Here, we provide general curatorial principles followed at GrainGenes, along with two examples of how curation is performed in practice.

GrainGenes [1] has a long history of serving the small grains communities via curation and many other activities. The repository was established in 1992 as a central data repository focused on Triticeae and Avena species, and has been continuously supported by the U.S. Department of Agriculture, Agricultural Research Service since then as a service to geneticists and breeders of wheat, barley, rye and oat worldwide. At times, the database resource has also facilitated research progress by hosting emerging projects such as EST sequencing, mapping, genome sequencing, and tools such as scripts for generating wheat genome-specific SNPs [7]. The database contains a wide variety of data types, including genome sequences, genetic maps, genes, alleles, molecular markers, phenotypes, QTLs, experimental protocols, and publications. In addition, GrainGenes serves the small grains communities by hosting small grains community newsletters such as the Annual Wheat Newsletter and Barley Genetics Newsletter, and community

☆ This article is part of a special issue entitled "Genomic resources".

* Corresponding author at: 800 Buchanan St. Albany, CA 94710, United States.
E-mail address: taner.sen@ars.usda.gov (T.Z. Sen).

https://doi.org/10.1016/j.cpb.2017.11.001
Received 26 September 2017; Received in revised form 28 November 2017; Accepted 29 November 2017

sites, such as the Triticeae Toolbox (T3) [8] repository. In addition, job openings, news updates, and links to other sites of interest are provided. A wide range of tools can be accessed by small grains researchers who use the website, including Generic Model Organism Database (GMOD) data visualization tools such as the CMap genetic map viewer [9] and the JBrowse genome browser [10] that visualizes genetic features along a reference sequence.

1.1. The age of big data

The effectiveness of data curation depends on the initial triage of papers, choosing the ones with the most impactful research outcomes. The amount and the heterogeneity of data that are included in the papers influence triage decisions. A curator needs to consider how data sets will be entered into a back-end database and displayed through the web interface. So-called "Big Data" has made these triage decisions more important than ever [11]. What do we mean by big data? Big data is hard to define, but for most in biological fields, it means data sets from megabytes to terabytes in size with a wide variety of data types. Big data is a direct result of technological innovations. Within the last few decades, rapid advancements in high-throughput technology and high-performance computing have resulted in an explosion of biological data production, both experimental and predictive. As a result, we now have access to multiple high-quality genome assemblies, transcriptomes, proteomes, and genome-wide association studies (GWAS). The resolution and accuracy of these data sets are usually high and they have opened up new possibilities for research and scientific discovery. However, our current data infrastructure, analysis methods, and visualization capabilities are being continuously challenged. Against this data deluge, indexing and standardization are becoming more crucial for ensuring that data are available for knowledge extraction, and research communities are getting together to create guiding principles, such as FAIR, i.e., findability, accessibility, interoperability, and reusability [12]. Grassroots organizations, such as AgBioData (https://www.agbiodata.org) have formed to help standardize data representation across groups and to make recommendations for responsible data sharing and management in developing data and metadata standards in the form of templates that would facilitate scaling of curation. At GrainGenes we developed data templates for researchers to help them upload their data into GrainGenes (some GrainGenes templates with metadata fields and example data entries can be found here: https://wheat.pw.usda.gov/GG3/submit), but better standards across communities are needed. The age of big data is only starting, and big data will definitely present more opportunities and challenges for biocuration in the future.

There has been an increasing effort to use standardized ontologies for labeling of genetic data. Standardization of data labels, in an object-oriented sense, has helped to build data-connections within and between databases as resources have grown over time. The aim of the Gene Ontology (GO) Consortium is to create unity in the description of gene terminology. Out of this community, the Trait Ontology (TO) and Plant Ontology (PO) have also blossomed. The evolution of terminology into ontology terms has enabled some types of classification and curation to be automated, easing the workload of curators. However, by no means does this automation make manual curation obsolete. Rather, high-throughput automation complements manual curation by allowing curators to focus more on the tasks of curation that require a human mind – those that call for critical thinking, investigation, and creativity.

1.2. Why curate?

To researchers who would like to have access to most recent, high-quality data in their field, the importance of curation is obvious. But, unfortunately, curation is not always seen as a critical part of scientific work. Here we want to emphasize the importance of curation for the advancement of science.

If new data sets are not curated into databases for long-term sustainability and integrated with pre-existing data, they may lose their accessibility and utility over time. If new, important data sets are not used, knowledge production and discovery rates will lag behind data production rates. In other words, data must be captured, standardized, organized, and made accessible to the scientific community if it is going to have a significant and lasting impact. In addition, a database is only as good as its data. If members of the scientific community do not find the data in their popular databases up-to-date, accurate, or transferable, then the database is of little use and will be obsolete soon. Likewise, if an online database's interface is not intuitive, few researchers will utilize the database. The role of a biocurator is therefore to provide up-to-date, accurate, and accessible information, and, through this critical activity, facilitate scientific discovery.

2. Curation workflow at GrainGenes

Although each publication that is curated into GrainGenes might use distinct data types, the general protocol for manual curation is the same (Fig. 1). By following an established procedure, data can be formatted in a manner that is compatible with work done by past and different curators, assuring that as much meaningful information as possible is stored in the database.

What differentiates community databases like GrainGenes from primary data repositories such as NCBI and EMBL is that the content of community databases is geared toward a particular organism or a set of closely-related organisms to cater to the needs of researchers in that particular community. The manual curation required to maintain a community database is time-intensive, and making incremental updates are an ongoing challenge, but it ultimately results in an indispensable resource.

2.1. Identification of peer-reviewed journal articles for curation

Every curator has limited time for curation, and therefore the first and most important step in the curation workflow at GrainGenes is to identify the peer-reviewed, high-impact journal articles that would most enrich the database and make the database most useful to small grains researchers. The identification step is primarily done by monitoring the release of publications in scientific journals relevant to small grains research. PubMed, Google Scholar and Scopus are sites where a wide range of journal articles is regularly updated, but these listings are not all-inclusive or strictly plant-focused. Our experienced curators are familiar with the journals most likely to contain articles of interest to our users and devote special attention to each new issue of them. Citation indices also help identify research with impact. We do not however use a specific list of journals or quality-metrics to identify articles.

Although web search tools are very useful, our experience shows that personal interactions are the best way of being informed of high-impact articles (Fig. 1). There are two beneficial ways of interacting with researchers to learn of new advances and therefore new publications coming our way. First, by attending conferences and listening to presentations, our curators are apprised of cutting-edge research that has been published or is about to be published. Second, when curators establish personal relationships with small grains researchers, then the researchers are more likely to contact the curators when they have new data sets. Some journals have actually made data submission mandatory for article authors in an effort to promote open access to research data, and GrainGenes is among the biological databases that greatly benefit from this requirement.

2.2. Curation prioritization

Selected papers then go through the triage stage, where the priority

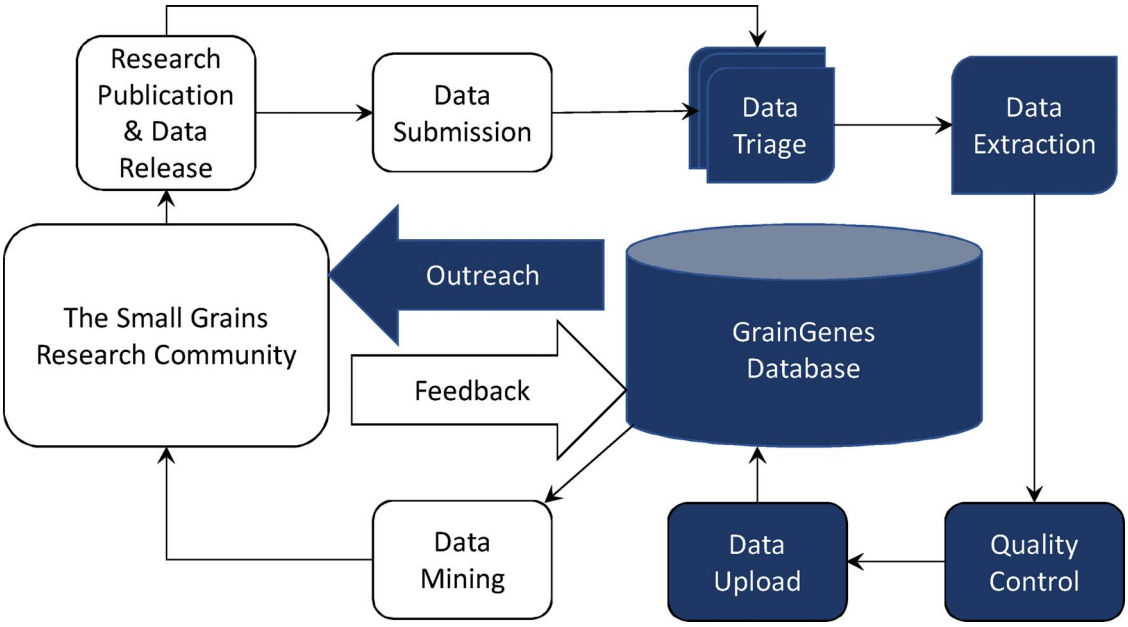

Fig. 1. A Simplified Schematic of the GrainGenes curation workflow. As part of the research life cycle, community databases such as GrainGenes and the researchers who use them exist in a mutually beneficial relationship, with the efforts of both groups enhancing the work of the other. Communication and an understanding of the needs and requirements of both groups are essential to keep the cycle running smoothly.

of the papers is assessed by the impact of the research outcomes and their importance to the small grains community. Higher priority papers such as new genome assemblies or genetic mapping studies are curated sooner. Much of this prioritization is done at the discretion of the curator. However, when we are contacted by researchers who wish to submit their data prior to their publications with tight deadlines, we do our best to accommodate them.

2.3. Formatting and standardizing the data for upload and data sharing

Once a paper has been selected, it must be parsed for relevant data, and the specific data types and data sets that will be curated must be identified. The data may come from the main journal article itself, or in supplemental tables that are provided along with the article. Extracting and formatting the data manually is perhaps the most time-consuming part of the actual curation process. After the data sets are properly extracted, further data formatting and standardizing are done with scripts developed here at GrainGenes. GrainGenes is a repository that has existed for almost a quarter century, so adherence to nomenclature and syntactical rules that were set by past GrainGenes curators and the small grains communities are both necessary to ensure that new data sets are properly connected to and expand on previously-curated data sets. Running a common workflow with established scripts ensures that once data sets have been uploaded, links and visualization are automatically created for specific data types.

3. Curatorial case studies

In order to more accurately portray the curation process at GrainGenes, the next three sections will describe case studies of curation projects. The first details a 'feature-centric' curation project, in which curators attempted to compile data on an important gene that affects yield in wheat from a collection of scientific publications. The second will follow the curator through the process of 'paper-centric' curation, in which a researcher contacted GrainGenes to submit data from their accepted paper on stem rust resistance in a collection of diverse durum wheat varieties. The last will highlight genomic curation, focusing on the curation of sequence data and gene annotations for the recently-released wild emmer wheat (*T. turgidum* ssp. *dicocoides*) genome assembly.

3.1. Curation of the TaGW2 gene for wheat yield

A Report Page for a gene (for example Fig. 2A) is far more substantial than it would first appear to be. It is the combined efforts of multiple scientists over years, potentially decades. Our knowledge of genes, their functions, sequences, and interactions, has come from researchers building on the discoveries of their peers and predecessors. That sort of cumulative knowledge paints a biological picture that is larger than the findings of each individual study alone and depends on the underlying data being accessible and reusable, often by storage in community databases and public data repositories. One example of this is the corpus of knowledge that has been developed around the gene *TaGW2* (*Triticum aestivum* grain width and weight *2*), which contributes to one aspect of wheat yield.

Yield in crop plants is known to involve very complex interactions of multiple, small-effect genes and regulations of expression. This makes breeding for higher yielding varieties quite challenging. In wheat, phenotypic measurements for yield often use the parameter thousand-grain weight (TGW) [13]. TGW is determined from a collection of different measurements, usually grain weight (GW), grain length (GL), and grain thickness (GT) [14], all three of which are positively correlated with TGW [15,16]. The wheat gene *TaGW2* is of particular interest in that it appears to have a relatively large, stable effect on thousand-grain weight and grain width across diverse germplasm, without a significant decrease in yield [17]. A great deal of research has been done to learn more about this gene and how it works. Because of its potential usefulness to wheat researchers and breeders, it is an important gene for which to have a well-curated page in the GrainGenes database.

TaGW2 was identified based on its homology to the rice gene *OsGW2*, which was found to affect yield in this cereal [17,18]. *OsGW2* was found to encode for a RING-type E3 ubiquitin ligase protein, and subsequent experiments with *TaGW2* have shown that the protein it encodes has similar activity [17,18]. Three homeologues of *TaGW2* exist on the chromosome group 6 of hexaploid wheat, with *TaGW2-6A* being the most well-characterized. *TaGW2-6A* has been shown to negatively regulate grain size [17]. Two main haplotypes identified in the promoter region of the gene have been identified, Hap-6A-G and Hap-

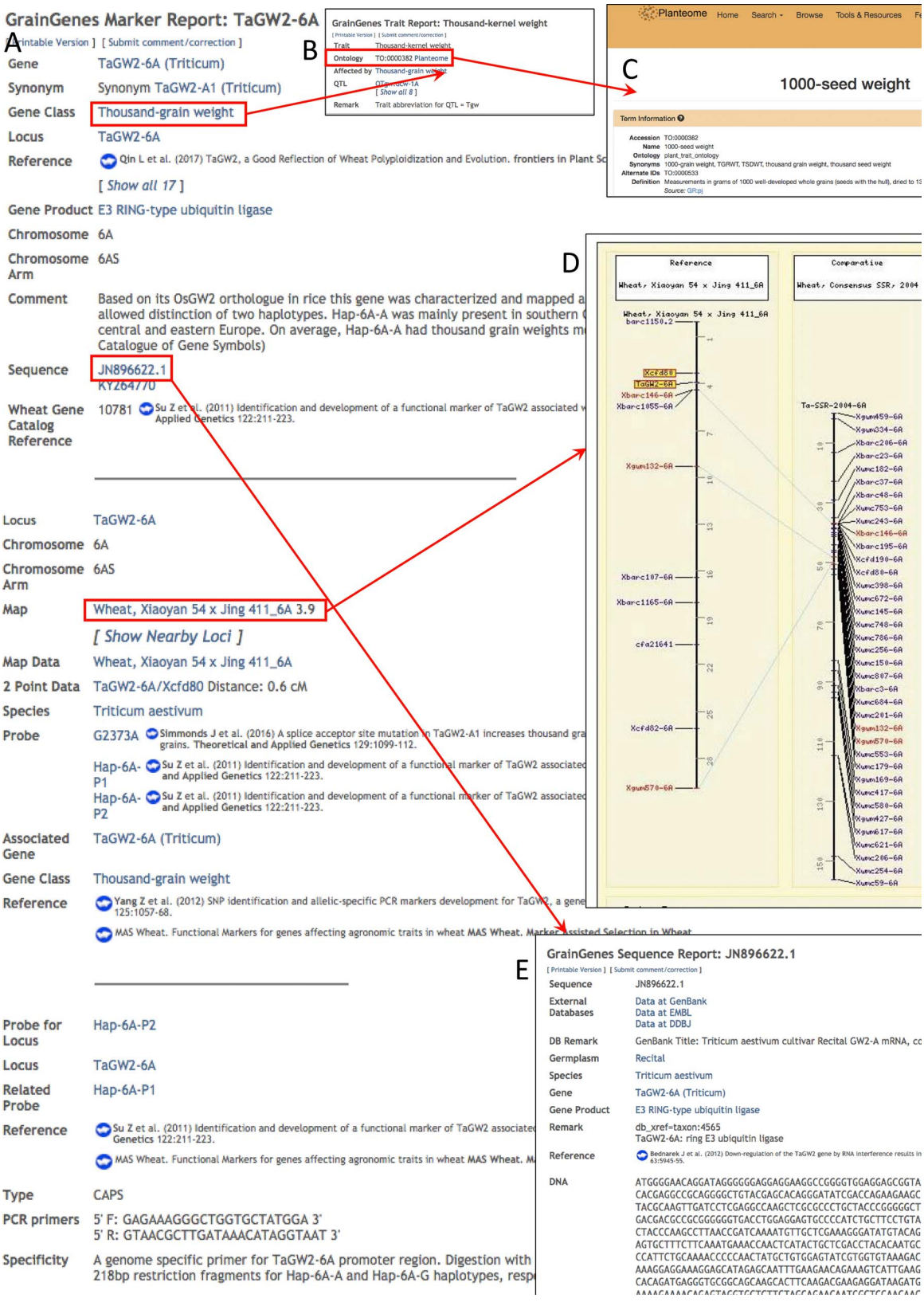

Fig. 2. (A) The TaGW2-6A Marker Report page showing gene, locus, and probe data with links to (B) Trait Report page for Thousand-kernel weight, which is affected by the Thousand-grain weight geneclass; (C) Planteome ontology page for TO:0000382: 1,000-seed weight; (D) CMap representation of the TaGW2 locus on chromosome 6A and a nearby marker, Xcfd80, highlighted in yellow (blue lines represent shared loci between the example map and the 2004 wheat SSR consensus map); (E) Sequence Report Page for the TaGW2-6A coding sequence from GenBank.

6A-A. The Hap-6A-A haplotype results in the production of a shortened peptide sequence, and genotypes possessing this haplotype show increased grain width and TGW. Single nucleotide polymorphisms (SNPs) in the *TaGW2-6A* coding sequence have also been identified that disrupt normal translation of the protein, and result in a phenotype of increased yield [19,20].

Starting with the characterization of this gene in 2011 by Su et al. [17], multiple published studies have confirmed the importance of this gene through association mapping studies: QTLs associated with yield effects are found in the same location as *TaGW2-6A*, near the centromere on the short arm of chromosome 6A [21–25]. In addition, more has been learned about the expression of the gene and its effect on yield and other traits [26], and about the variation in this gene resulting from polyploidization, domestication, and breeding for different geographic regions [27,28]. Similarly, more information is being gleaned from the *TaGW2* homeologues on 6B and 6D, and how the three genes work in concert to produce a phenotype [29].

So how does one curate this? In terms of importance, genes that have a significant effect on yield are relatively rare, so this gene would have a high priority in curation. This particular 'gene-centric' approach to curation, in which a curator chooses a gene and looks at the collective knowledge-set for it across multiple published articles before combining and filtering the information into data types in the database, can be particularly time-consuming and difficult. Each paper builds off of several others, and consolidating the data from each of them into the most distilled form requires discretion and a strong understanding of the biological methods used in each study. In addition, it often requires investigation to sift through evidence and to locate accessions in other databases and repositories that can help paint a more complete picture of the gene.

In curating this gene, we had to first look at what information was already present in the database. From there, we could determine from the literature what information was missing, and what, if anything, needed to be changed.

By searching in the Genome Browser with the term 'TaGW2*' on GrainGenes, four different records were found – two gene reports and two loci reports. It appeared as though information on both 'TaGW2-6A' and 'TaGW2-A1′ had been previously, but separately curated into the GrainGenes database, with references to the original paper [17], which identified *TaGW2*. A reference to the information page about the TaGW2 gene provided by the MASWheat project was found as well (http://maswheat.ucdavis.edu/protocols/TaGW2/), demonstrating that *TaGW2-6A* and *TaGW2-A1* were synonyms for the same gene, and curatorial steps needed to be taken to link these two records in GrainGenes.

It is generally best practice to have the majority of information on a genetic feature assigned to one standard name in the database, so that users and curators do not need to do a great deal of searching to gather knowledge on that feature. Synonyms then should all link to the preferred feature name so that users can locate the information by searching either name. However, such an ideal situation rarely exists. In the literature, it appears that the two names were used interchangeably. However, the Wheat Catalogue of Gene Symbols (https://shigen.nig.ac.jp/wheat/komugi/genes/symbolClassList.jsp) listed *TaGW2-6A*, but did not mention *TaGW2-A1*, so we chose *TaGW2-6A* to be the central gene report, and created synonym links to the *TaGW2-A1* report, so that one could easily navigate between the two. (Shortly after our curation, a supplement of the Wheat Gene Catalogue was released that cites *TaGW2-A1* in association with *TaGW-A2. The supplement can be found here:* https://shigen.nig.ac.jp/wheat/komugi/genes/macgene/supplement2017.pdf*)*

Moving on to the locus data points, the same process was repeated: transferring the majority of information over to the primary report and creating synonym links for the two names. After the pre-existing information in the database had been organized, it was time to look at what new research outcomes regarding *TaGW2-6A* existed that could be curated into the database. Using publicly available curation tools such as Pubtator [30], the curation text-mining tool offered by NCBI, makes locating extractable data types within text faster and easier. Once we had identified the corpus of scientific papers that pertained to *TaGW2*, the PubTator search tool pulled up sixteen articles (as of January 2017) that mentioned 'TaGW2′, starting with the original Su et al., 2011 study [17]. The exact number of articles obtained by Pubtator changes as new articles are added to PubMed and pulled by Pubtator.

By reading through the abstracts of the papers, curators started to get an idea of the data types that would need to be created for the information. For example, sequence and probe data types needed to be created for PCR primers that were developed to differentiate the Hap-6A-A and Hap-6A-G alleles. A gene product data type needed to be created for 'RING-type E3 ubiquitin ligase', as that was not present in the database. For mapping studies, map data, loci, and QTLs needed to be created so that the overlap of the location of *TaGW2-6A* with various yield QTLs could be easily visualized in CMap.

With QTL and gene class data types, curators must do a thorough search of the data that is already in GrainGenes, not only to prevent redundancies, but also to make sure that the data that is added is as interconnected as much as possible with pre-existing data. For example, the gene class 'Thousand-grain weight', based off of the Gene Class from the Wheat Catalogue of Gene Symbols, was already in GrainGenes. By linking the *TaGW2-6A* gene report to this gene class, we also created links to the already curated 'thousand-grain weight' trait data type, and the grain weight QTLs. A plant trait ontology term: TO:0000382, for 1000 grain weight was then attached to *TaGW2-6A*, and a link to Planteome (www.planteome.org) [31], a resource for common reference ontologies related to plant biology, was also provided (Fig. 2). Several GenBank [32] accessions are available for TaGW2-6A. The full coding sequence was obtained[29], as well as the promoter region of the gene [18], and a coding sequence of the gene with a mutation that creates an alternate splice site [19].

When *TaGW2-6A* was first identified, it was also fine mapped onto chromosome 6A in hexaploid wheat (*Triticum aestivum* L.) [17]. The marker Xcfd80, which is very close to the centromere, was found to be 0.6 cM from the locus for *TaGW2-6A*. The marker and probe data for Xcfd80 were already in GrainGenes, as were most of the other markers that were used for mapping in the study, so curators simply needed to use the marker names and cM distance provided in the paper to construct a map data page (Fig. 2). The fine map created in the study from a biparental mapping population was quite small – only 12 markers – and was only for one chromosome. This made creating a CMap instance of the map and making correspondences with other maps relatively simple. After the map data type was created, it was configured for display in CMap, and corresponding links were built between maps in GrainGenes that have markers in common with the new map. A combination of automated and manual searching through the database was required to check which of the markers already existed in the database – structured datasets can facilitate this process. If there were any potentially matching markers under slightly different names, these were investigated to ascertain whether or not they were, the same marker. With curation in general, name-checking needs to be done thoroughly. Errors in naming can be extremely difficult to rectify, and may lead to lost information or repeated errors down the road.

When the process was finished, the resulting Report Page for *TaGW2* told a rather complete and interesting story (Fig. 2). By combining and distilling the research of multiple scientists since the gene's identification in 2011, curators have enabled future GrainGenes users who are interested in this gene to view a succinct and accurate snapshot of what we currently know about this gene. From there, they can formulate their own hypotheses to further our knowledge, or they can use the information from this page to incorporate varieties of wheat containing desirable alleles for this gene into their breeding schemes. The time that curators spent processing and adding value and perspective to the data from those papers allows GrainGenes clients to more efficiently use the resulting knowledge in their work.

3.2. Curation of QTL for stem rust resistance in durum wheat

The majority of curation at GrainGenes is 'paper-centric', meaning that our efforts are focused around extracting all the crucial research outcomes from one specific study published in a peer-reviewed journal. Often we are contacted by researchers involved in the study who wish to have their data put into the GrainGenes database usually because it will be publicly-available in a long-term database with sustainable funding. Increasingly, some funding agencies and open access journals have added open data requirements to researchers to create incentives for making research outcomes available through repositories such as GrainGenes.

One such example of a paper that was curated at GrainGenes was the 2017 study by Chao et al. [33] which evaluated stem rust resistance in a diverse durum wheat (*Triticum turgidum ssp. durum*) panel from the National Small Grains Collection (NSGC). A component of the National Plant Germplasm System (NPGS) of the United States Department of Agriculture – Agricultural Research Service (USDA-ARS), the NSGC is a collection of small grains germplasm from across the globe and represents a vast amount of genetic diversity (http://www.ars.usda.gov/main/docs.htm?docid=2884). The corresponding author of the study contacted us to request that the data from the study be made publically available in GrainGenes. The paper on the study had been accepted for publication in The Plant Genome after peer-review. The journal requested that the data be curated prior to publication and links to the corresponding data sets could be provided in the paper.

Stem rust (*Puccinia graminis* f. sp. *tritici*), or *Pgt*, is a fungal pathogen that has caused serious damage to both hexaploid and tetraploid wheat across the globe. Between 1955 and 1999, stem rust was relatively under control in North America and most of the world (http://maswheat.ucdavis.edu//protocols/StemRust), due to the development of resistant varieties and cultural practices that reduced stem rust infection. However, in 1999, an extremely virulent strain of *Pgt*, TTTSK, was found in wheat fields in Uganda. It was capable of overcoming the major resistance genes that were bred into wheat at the time [34]. Referred to as Ug99, the destructive strain wreaked havoc on wheat yields in East Africa and kindled a worldwide search to identify new sources of stem rust resistance. Of the 70 known stem rust resistance (*Sr*) genes that have been identified and characterized in wheat, about 34 for them are still effective against Ug99 [44,45]. Two new races, JRCQC and TRTTF, have since been found to overcome *Sr13* and *Sr9e* – two very widely used *Sr* genes [34,35]

Durum wheat germplasm is generally more resistant to Ug99 and other races of stem rust than common bread wheat [36] Therefore, the 2017 study by Chao et al. [33] aimed to survey the genetic potential of a diverse array of durum wheat germplasm held by the U.S. National Small Grains Collection to determine if new sources of stem rust resistance could be found. In total, 429 lines were evaluated, including landraces, breeding materials, and elite cultivars from 64 countries and regions across the globe.

The first step in curating this paper was to thoroughly read and understand the paper, and then to examine the provided data sets to gather as much information on the study as possible. Our deep reading of the study showed that twenty-one loci in total were found to be associated with disease responses to stem rust races through genome-wide association analysis. Of those, seventeen loci were found to affect seedling response to stem rust, and four were associated with adult plant resistance (APR) in field studies. These loci were grouped into 13 distinct quantitative trait loci due to their proximity, and whether or not they were associated or potentially associated with a known *Sr* gene.

The markers that were used in the study were from the Hexaploid Wheat Illumina 90 K iSelect SNP assay[37], and information for the markers was already present in GrainGenes. Significant trait-associated markers found in the GWAS were aligned to the consensus map (Fig. 3A) for tetraploid wheat, and those map positions were used to create a CMap instance, with the positions of the thirteen QTL included on the map (Fig. 3E).

To incorporate the QTL data types, names were created by the curators for each QTL because none had been provided in the paper. Using standard QTL naming conventions, they were labeled with an abbreviation of the trait the QTL is associated with and the chromosome it is located on, for example, QSr.locus–3B (Fig. 3C). Information such as linked markers, associated traits and/or genes, and comments on the magnitude of effect, environment, and stem rust race to which the QTL was associated were all added for each QTL data type. Plant Trait Ontology [31] terms for fungal disease resistance (TO:0000439) were also added to the report page. Germplasm data types were created with links to a germplasm accession page in the Germplasm Resource Information Network (GRIN) database (http://www.ars-grin.gov/) (Fig. 3B).

Like many genome-wide association studies, the phenotypic data sets from this study were collected from field trials at multiple locations and over the course of several years. Reading through the Methods section, and, if necessary, consulting with corresponding authors is crucial to obtaining necessary environmental data to link with phenotype scores.

Curating meaningful metadata, or data about data sets, is a crucial aspect of any database or repository. Like a table with no column labels, a dataset without proper metadata is uninformative and, therefore, quite useless. In the worst case scenario, insufficient metadata for a dataset can result in its misinterpretation or misuse. Although that is usually not the case, it is nonetheless important for curators to maintain proper metadata for phenotype and other kinds of datasets. For this study, conditions and locations for field trials were well-documented in the Methods section of the paper, and this information was formatted into trait study and environment data types in the database for each field site and year.

Some databases are better equipped than others to visualize particular data types and make them available. To prevent redundancy and to help make data as easily accessible to researchers as possible, it is important for databases to acknowledge areas in which they excel, and areas in which other databases may surpass them. Collaboration across databases enables data to be stored in a unified, findable, and easily interpretable manner. For example, the phenotypic data for stem rust resistance across the four field trials and the six seedling trials performed in the study have been stored in The Triticeae Toolbox (T3) [7] under Trial Codes:-StemRustSeedling_2013_StPaul, StemRustField_2012_StPaul, StemRustField_2013_StPaul, StemRustField_2014_StPaul, and StemRustField_2014_Ethiopia [37]. From their webpage, users can download the full dataset of phenotype scores for the 429 durum wheat lines used in the study, and identify individual lines that showed higher resistance across the various environments. By providing links from GrainGenes' mapping and QTL data to these resources at T3 (Fig. 3D), we provide the most information to the user as possible, without expending extra resources to carry the same information in a less complete form on both databases.

This study was important in helping to evaluate the germplasm of durum wheat that is available in the NSGC for resistance to stem rust. In curating the data from this paper, information on known and novel genes, QTLs, markers, and germplasm has been added to the net collection of knowledge on stem rust and existing resistance in durum wheat. There is great value added by curation in a biological database because users can look at the broader picture, and find connections that were not found by any individual study. Breeders can use this information in their efforts to breed more resistant wheat varieties. Small grains researchers can build hypotheses off of what they find in GrainGenes and other biological databases, and in this way publish more studies and create yet more data and knowledge.

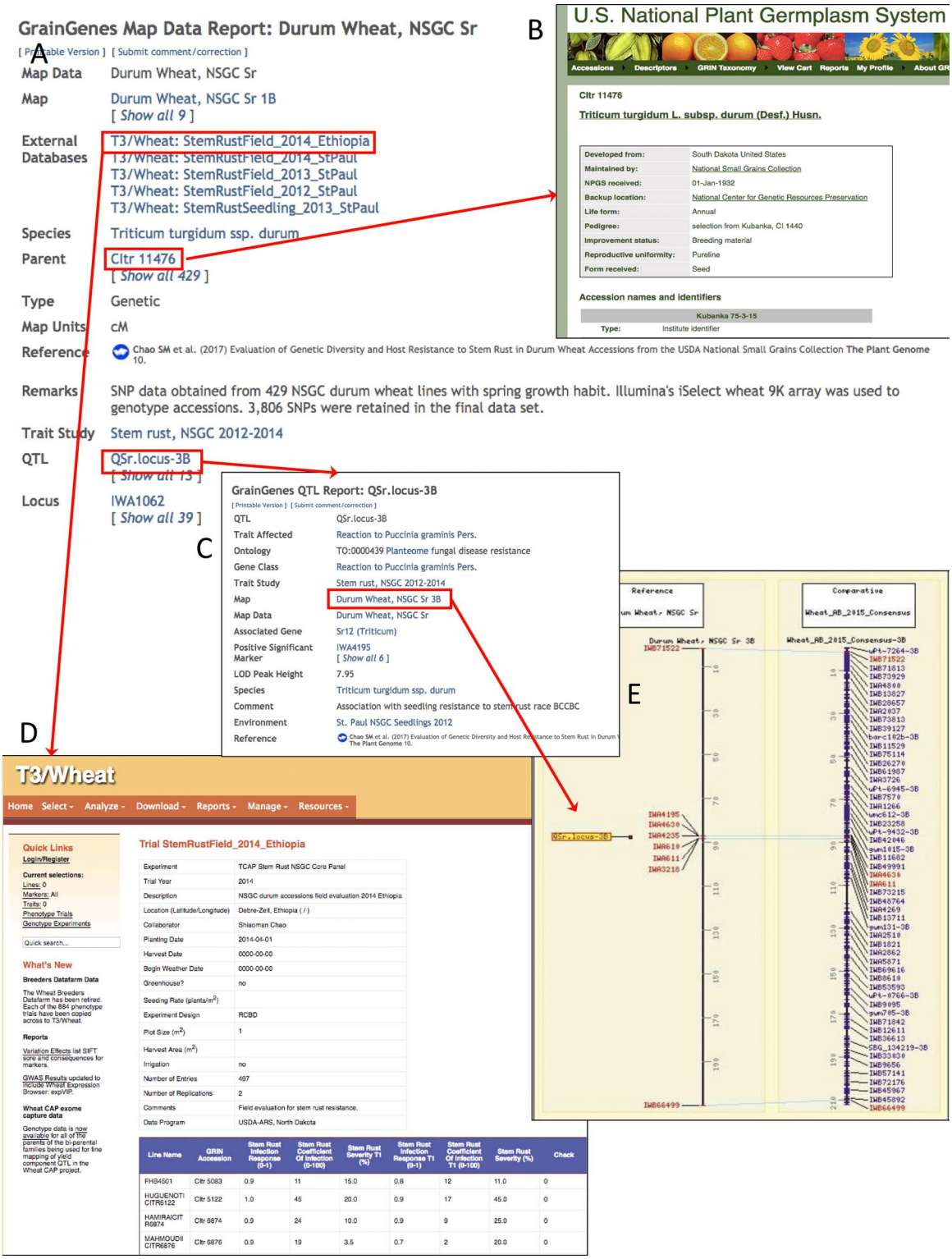

Fig. 3. (A) The Map Data Report Page for the Chao et al. 2017 GWAS study of stem rust resistance in the NSGC durum wheat collection; (B) the GRIN-Global germplasm accession page for CItr 11476, one of the lines used in the study; (C) the QTL Report page for QSr.locus-3B, detailing trait information and associated markers; (D) T3/Wheat Phenotype data for the Ethiopia 2014 field trials from this study, which are linked to from the GrainGenes report page; (E) the CMap visualization of chromosome 3B genetic map, with comparisons to the tetraploid wheat 90K SNP array map (blue lines represent shared markers between maps. The location of the QTL Qsr.locus-3B is also shown.

3.3. Curation of genomic data

The curation of genomic data (including functional annotations of genomic elements in a comparative genomic context, usually without any associated genetic coordinates) has become more important as GrainGenes moves toward a genome-centric platform. Currently, many wheat, barley, and species-specific pan-genomes are in the process of being sequenced and assembled, or are already available for genome curation [38–40]. With so much genomic data, we must decide which genomes to include in GrainGenes, and which should receive top priority. How do we prioritize these genomic datasets?.

As an example of genomic curation, we will use the wild emmer

wheat variety Zavitan [41]. Common bread wheat, *Triticum aestivum,* has a hexaploid genome made up of three subgenomes: The A genome (originating from *T. urartu)*, the B genome (originating from *Ae. speltoides*), and the D genome (from *Ae. tauschii)* [42–44]. Wild emmer wheat, *Triticum turgidum ssp. dicoccoides*, resulted from an initial polyploidy event merging the A and B genomes [45]. A second polyploidization occurred later, which introduced the D genome, creating the common bread wheat we use today [46]. Tetraploid durum wheat (*Triticum turgidum ssp. durum*) is also of the genome formula AABB, and is a close relative of wild emmer. The emmer wheat genome is therefore of interest to both wheat breeders and plant evolutionary biologists, and GrainGenes prioritized the curation of the Zavitan genome as a result.

Our curation process for genomic data is quite different from literature curation in both protocol and scale. As with genetic data, we receive notice of genome data through the research groups directly, though we also triage the literature to look for genomes that might be of interest to our stakeholders. The researchers that were involved with the sequencing and assembly of the Zavitan genome (http://wewseq.wixsite.com/consortium) contacted GrainGenes because they wished to make the assembly available also through our database. Under the Toronto agreement [47], GrainGenes provided pre-publication access to the WEWseq v1.0 Zavitan assembly via a BLAST database (https://wheat.pw.usda.gov/GG3/wildemmer_blast) [48] and a JBrowse[10] instance. For Zavitan, our workflow included creating and indexing BLAST databases for chromosomes and gene models, and creating a reference sequence track and gene model annotation tracks for the JBrowse instance by converting the genomic data into a gff format. Fig. 4 shows the GrainGenes Zavitan Genome Browser, showing all gene model annotations as well as high-confidence gene model annotations.

In the future, we intend to prioritize hosting the genomes of other small grains genomes as they become available. By hosting all these genomes in one centralized database and linking them to several already existing types of data at GrainGenes, we can make it easier for breeders and biologists to perform comparative analyses and make connections between genomic, genetic and QTL data via JBrowse and BLAST.

We anticipate a rapid increase in both the quantity and quality of Triticeae and Avena genomes that will be released in the near future. In order to prepare for these data sets, GrainGenes and other databases and repositories must work to create pipelines for efficient upload of genomic data and to implement tools for meaningful visualization, analysis, and comparison of genomes. These efforts will provide a knowledge-base to better navigate the genomes to make informed decisions for crop improvement.

3.4. Data linking to external sites

When users query data at GrainGenes using a keyword, they are provided with search results that are pointing to pages associated with that keyword. If we use markers as an example, marker pages may contain associated genetic map information, species, and remarks. Some of these fields are clickable and lead to other GrainGenes internal pages where users can find more information, such as other markers on the same genetic map. In some cases, these fields lead to external sites that provide more information, such as GenBank [32], T3 database [8], Gramene database [4], and the Germplasm Resource Information Network (GRIN) database (http://www.ars-grin.gov/).

When providing external links, these links need to be 1) accurate, 2) complete, and 3) up-to-date. So-called “dead” links (i.e., links that do not lead to any functional webpage) reduce the value of a database for users, and make users question the accuracy of the rest of the database. Different synchronization approaches are used to make sure that these links are accurate and up-to-date. In some cases, staff at different

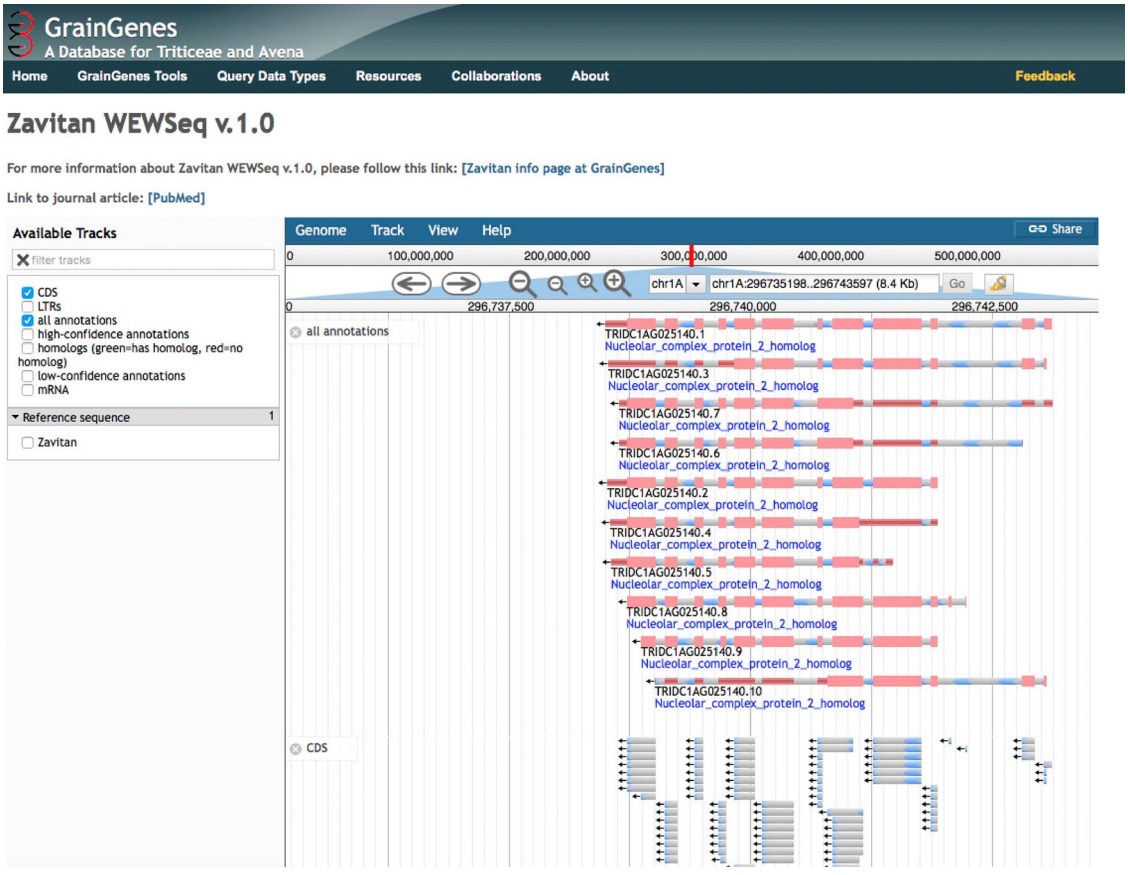

Fig. 4. The GrainGenes Genome Browser with tracks of all gene models for Zavitan WEWseq v1.0 and of the high-confidence gene subset are shown.

databases communicates with each other on a regular basis to exchange indexes over emails based on common keywords, such as GenBank ids or locus names. Although this approach is easy to implement between long-term collaborators for small number of datasets; in recent years more automated options started facilitating data sharing for larger and more heterogeneous datasets. Among them are exchanging indexes through automated servers; such as Apache Solr (for example Wheat Information System (http://wheatis.org) uses this approach). Another approach; which requires more labor in the beginning; but allows more automation; is to use Application Programming Interfaces (APIs) that automatically exchange indices. Projects such as BrAPI (https://brapi.org/) use this approach. GrainGenes use a combination of these approaches when providing links to external databases.

3.5. Data download options

Curation of data into GrainGenes allows users to query and view the data in relation to other types of data through GrainGenes web-based interface in a facilitated manner. In some cases, however, researchers may like to access and process these datasets by downloading them to their computers. GrainGenes allows several different ways to access and download data. Genomic data residing as JBrowse tracks can be easily downloaded from the track info buttons using the menu option “Save track data”. For non-genomic data, GrainGenes personnel created several “bulk download” options that are available for different data types under the “Resources” menu on the GrainGenes front page.

But the most powerful query and data download option GrainGenes offers is the SQL-based search interface under “GrainGenes/GrainGenes Tools/Advanced Queries” (https://wheat.pw.usda.gov/GG3/advanced-queries). Through this interface, users can utilize their SQL expertise to query ANY data residing on the GrainGenes back-end database. Although this option is geared more towards technically-oriented researchers, database schema, table diagrams, and some beginner tutorials are provided to help users download their datasets of interest (https://wheat.pw.usda.gov/GG3/tutorials).

4. Conclusion

The principles and practices for manual curation are based around the preservation and sharing of biologically relevant and significant data sets. At GrainGenes, we work to maintain a high standard by inputting peer-reviewed, high impact research outcomes into the database. In this way, GrainGenes strives to be a reliable and accurate resource for the small grains community that it serves.

The flow of data from biological research is not abating any time soon. To the contrary, data production is expected to accelerate rapidly as technologies for high-throughput sequencing, genotyping and phenotyping become cheaper, faster, and more accurate. As biological research advances, the work that will be involved in biocuration will naturally evolve as well. The release of genome sequences provides excellent opportunities for further research in small grains. The collection of historical and current marker, QTL, and mapping data housed in GrainGenes provides the database with a unique opportunity to connect genomic and genetic data to facilitate further discoveries. In addition, the development and implementation of ontologies will aid in making cross-species comparisons and in organizing gene models and plant traits.

Dealing with this overwhelming influx of data in a responsible way means ensuring that (1) it is of high quality, (2) it has meaningful metadata, (3) it is stored in such as way that it will persist over time, and (4) it is viewed in the context of similar data, so that comparisons and new insights can be made. Biocurators are fully or partially responsible for all of these tasks. The work that curators of biological databases do should, therefore, be seen as a valuable part of the research lifecycle.

Declaration of interest

Conflicts of interest: none.

Acknowledgements

Research was supported by USDA Agricultural Research Service CRIS project 2030-21000-021-00D. 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 US Department of Agriculture. USDA is an equal opportunity provider and employer. The authors would like to thank Ann Blechl for critically reading the manuscript and providing useful suggestions, Shiaoman Chao and Assaf Distelfeldt for their help to deposit and visualize their data sets at GrainGenes.

Conflict of interest statement

The authors have no conflict of interest.

References

[1] V. Carollo, D.E. Matthews, G.R. Lazo, T.K. Blake, D.D. Hummel, N. Lui, D.L. Hane, O.D. Anderson, GrainGenes 2. 0. an improved resource for the small-grains community, Plant Physiol. 139 (2005) 643–651.
[2] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D.L. Alexander, M. Garcia-Hernandez, et al., The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 40 (2012) D1202–1210.
[3] C.M. Andorf, E.K. Cannon, J.L. Portwood 2nd, J.M. Gardiner, L.C. Harper, M.L. Schaeffer, B.L. Braun, D.A. Campbell, A.G. Vinnakota, V.V. Sribalusu, et al., MaizeGDB update: new tools, data and interface for the maize model organism database, Nucleic Acids Res. 44 (2016) D1195–D1201.
[4] M.K. Tello-Ruiz, J. Stein, S. Wei, J. Preece, A. Olson, S. Naithani, V. Amarasinghe, P. Dharmawardhana, Y. Jiao, J. Mulvaney, et al., Gramene 2016: comparative plant genomics and pathway resources, Nucleic Acids Res. 44 (2016) D1133–1140.
[5] N. Fernandez-Pozo, N. Menda, J.D. Edwards, S. Saha, I.Y. Tecle, S.R. Strickler, A. Bombarely, T. Fisher-York, A. Pujar, H. Foerster, et al., The sol genomics network (SGN)?from genotype to phenotype to breeding, Nucleic Acids Res. 43 (2015) D1036–D1041.
[6] D. Grant, R.T. Nelson, S.B. Cannon, R.C. Shoemaker, SoyBase, the USDA-ARS soybean genetics and genomics database, Nucleic Acids Res. 38 (2010) D843–846.
[7] G.R. Lazo, S. Chao, D.D. Hummel, H. Edwards, C.C. Crossman, N. Lui, D.E. Matthews, V.L. Carollo, D.L. Hane, F.M. You, et al., Development of an expressed sequence tag (EST) resource for wheat (Triticum aestivum L.): EST generation, unigene analysis, probe selection and bioinformatics for a 16, 000-locus bin-delineated map, Genetics 168 (2004) 585–593.
[8] V.C. Blake, C. Birkett, D.E. Matthews, D.L. Hane, P. Bradbury, J.L. Jannink, The triticeae toolbox: combining phenotype and genotype data to advance small-grains breeding, Plant Genome (2016) 9.
[9] K. Youens-Clark, B. Faga, I.V. Yap, L. Stein, D. Ware, CMap 1.01: a comparative mapping application for the Internet, Bioinformatics 25 (2009) 3040–3042.
[10] R. Buels, E. Yao, C.M. Diesh, R.D. Hayes, M. Munoz-Torres, G. Helt, D.M. Goodstein, C.G. Elsik, S.E. Lewis, L. Stein, et al., JBrowse: a dynamic web platform for genome visualization and analysis, Genome Biol. 17 (2016) 66.
[11] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D.P. Hill, R. Kania, M. Schaeffer, S. St Pierre, et al., Big data: the future of biocuration, Nature 455 (2008) 47–50.
[12] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B. da Silva Santos, P.E. Bourne, et al., The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3 (2016) 160018.
[13] K.G. Campbell, C.J. Bergman, D.G. Gualberto, J.A. Anderson, M.J. Giroux, G. Hareland, R.G. Fulcher, M.E. Sorrells, P.L. Finney, Quantitative trait loci associated with kernel traits in a soft × hard wheat cross, Crop Sci. 39 (1999) 1184–1195.
[14] B.B. Dholakia, J.S.S. Ammiraju, H. Singh, M.D. Lagu, M.S. Roder, V.S. Rao, H.S. Dhaliwal, P.K. Ranjekar, V.S. Gupta, W.E. Weber, Molecular marker analysis of kernel size and shape in bread wheat, Plant Breed. 122 (2003) 392–395.
[15] F. Breseghello, M.E. Sorrells, Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars, Genetics 172 (2006) 1165–1177.
[16] X.Y. Sun, K. Wu, Y. Zhao, F.M. Kong, G.Z. Han, H.M. Jiang, X.J. Huang, R.J. Li, H.G. Wang, S.S. Li, QTL analysis of kernel shape and weight using recombinant inbred lines in wheat, Euphytica (2008) 165.
[17] Z. Su, C. Hao, L. Wang, Y. Dong, X. Zhang, Identification and development of a functional marker of TaGW2 associated with grain weight in bread wheat (Triticum aestivum L.), Theor. Appl. Genet. 122 (2011) 211–223.
[18] X.J. Song, W. Huang, M. Shi, M.Z. Zhu, H.X. Lin, A QTL for rice grain width and weight encodes a previously unknown RING-type E3 ubiquitin ligase, Nat. Genet.

39 (2007) 623–630.
[19] J. Simmonds, P. Scott, J. Brinton, T.C. Mestre, M. Bush, A. Del Blanco, J. Dubcovsky, C. Uauy, A splice acceptor site mutation in TaGW2-A1 increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains, Theor. Appl. Genet. 129 (2016) 1099–1112.
[20] V. Jaiswal, V. Gahlaut, S. Mathur, P. Agarwal, M.K. Khandelwal, J.P. Khurana, A.K. Tyagi, H.S. Balyan, P.K. Gupta, Identification of novel SNP in promoter sequence of taGW2-6A associated with grain weight and other agronomic traits in wheat (Triticum aestivum L.), PLoS One 10 (2015) e0129400.
[21] K. Zhang, J. Wang, L. Zhang, C. Rong, F. Zhao, T. Peng, H. Li, D. Cheng, X. Liu, H. Qin, et al., Association analysis of genomic loci important for grain weight control in elite common wheat varieties cultivated with variable water and fertiliser supply, PLoS One 8 (2013) e57853.
[22] Y. Xu, R. Wang, Y. Tong, H. Zhao, Q. Xie, D. Liu, A. Zhang, B. Li, H. Xu, D. An, Mapping QTLs for yield and nitrogen-related traits in wheat: influence of nitrogen and phosphorus fertilization on QTL expression, Theor. Appl. Genet. 127 (2014) 59–72.
[23] J. Simmonds, P. Scott, M. Leverington-Waite, A.S. Turner, J. Brinton, V. Korzun, J. Snape, C. Uauy, Identification and independent validation of a stable yield and thousand grain weight QTL on chromosome 6A of hexaploid wheat (Triticum aestivum L.), BMC Plant Biol. 14 (2014) 191.
[24] X.Q. Huang, H. Kempf, M.W. Ganal, M.S. Roder, Advanced backcross QTL analysis in progenies derived from a cross between a German elite winter wheat variety and a synthetic wheat (Triticum aestivum L.), Theor. Appl. Genet. 109 (2004) 933–943.
[25] X.Q. Huang, S. Cloutier, L. Lycar, N. Radovanovic, D.G. Humphreys, J.S. Noll, D.J. Somers, P.D. Brown, Molecular detection of QTLs for agronomic and quality traits in a doubled haploid population derived from two Canadian wheats (Triticum aestivum L.), Theor. Appl. Genet. 113 (2006) 753–766.
[26] Y. Hong, L. Chen, L.P. Du, Z. Su, J. Wang, X. Ye, L. Qi, Z. Zhang, Transcript suppression of TaGW2 increased grain width and weight in bread wheat, Funct. Integr. Genomics 14 (2014) 341–349.
[27] L. Qin, J. Zhao, T. Li, J. Hou, X. Zhang, C. Hao, TaGW2, a good reflection of wheat polyploidization and evolution, Front. Plant Sci. 8 (2017) 318.
[28] L. Qin, C. Hao, J. Hou, Y. Wang, T. Li, L. Wang, Z. Ma, X. Zhang, Homologous haplotypes, expression, genetic effects and geographic distribution of the wheat yield gene TaGW2, BMC Plant Biol. 14 (2014) 107.
[29] J. Bednarek, A. Boulaflous, C. Girousse, C. Ravel, C. Tassy, P. Barret, M.F. Bouzidi, S. Mouzeyar, Down-regulation of the TaGW2 gene by RNA interference results in decreased grain size and weight in wheat, J. Exp. Bot. 63 (2012) 5945–5955.
[30] C.H. Wei, H.Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucleic Acids Res. 41 (2013) W518–522.
[31] L. Cooper, P. Jaiswal, The plant ontology: a tool for plant genomics, Methods Mol. Biol. 1374 (2016) 89–114.
[32] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K.D. Pruitt, E.W. Sayers, GenBank, Nucleic Acids Res. (2017), http://dx.doi.org/10.1093/nar/gkx1094/4621329.
[33] S. Chao, M.N. Rouse, M. Acevedo, A. Szabo-Hever, H. Bockelman, J.M. Bonman, E. Elias, D. Klindworth, S. Xu, Evaluation of genetic diversity and host resistance to stem rust in USDA NSGC durum wheat accessions, Plant Genome (2017) 10.
[34] R.P. Singh, D.P. Hodson, Y. Jin, E.S. Lagudah, M.A. Ayliffe, S. Bhavani, M.N. Rouse, Z.A. Pretorius, L.J. Szabo, J. Huerta-Espino, et al., Emergence and spread of new races of wheat stem rust fungus: continued threat to food security and prospects of genetic control, Phytopathology 105 (2015) 872–884.
[35] M. Rahmatov, M.N. Rouse, J. Nirmala, T. Danilova, B. Friebe, B.J. Steffenson, E. Johansson, A new 2DS.2RL Robertsonian translocation transfers stem rust resistance gene Sr59 into wheat, Theor. Appl. Genet. 129 (2016) 1383–1392.
[36] W. Zhang, S. Chen, Z. Abate, J. Nirmala, M.N. Rouse, J. Dubcovsky, Identification and characterization of Sr13, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) E9483–E9492.
[37] S. Wang, D. Wong, K. Forrest, A. Allen, S. Chao, B.E. Huang, M. Maccaferri, S. Salvi, S.G. Milner, L. Cattivelli, et al., Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array, Plant Biotechnol. J. 12 (2014) 787–796.
[38] J.D. Montenegro, A.A. Golicz, P.E. Bayer, B. Hurgobin, H. Lee, C.K. Chan, P. Visendi, K. Lai, J. Dolezel, J. Batley, et al., The pangenome of hexaploid bread wheat, Plant J. 90 (2017) 1007–1013.
[39] M. Mascher, H. Gundlach, A. Himmelbach, S. Beier, S.O. Twardziok, T. Wicker, V. Radchuk, C. Dockter, P.E. Hedley, J. Russell, et al., A chromosome conformation capture ordered sequence of the barley genome, Nature 544 (2017) 427–433.
[40] E. Bauer, T. Schmutzer, I. Barilar, M. Mascher, H. Gundlach, M.M. Martis, S.O. Twardziok, B. Hackauf, A. Gordillo, P. Wilde, et al., Towards a whole-genome sequence for rye (Secale cereale L.), Plant J. 89 (2017) 853–869.
[41] R. Avni, M. Nave, O. Barad, K. Baruch, S.O. Twardziok, H. Gundlach, I. Hale, M. Mascher, M. Spannagl, K. Wiebe, et al., Wild emmer genome architecture and diversity elucidate wheat evolution and domestication, Science 357 (2017) 93–97.
[42] D. Chalupska, H.Y. Lee, J.D. Faris, A. Evrard, B. Chalhoub, R. Haselkorn, P. Gornicki, Acc homoeoloci and the evolution of wheat genomes, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 9691–9696.
[43] J. Dvorak, P. Terlizzi, H.B. Zhang, P. Resta, The evolution of polyploid wheats: identification of the A genome donor species, Genome 36 (1993) 21–31.
[44] J. Dvorak, E.D. Akhunov, Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the Aegilops-Triticum alliance, Genetics 171 (2005) 323–332.
[45] J. Dvorak, H.B. Zhang, Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes, Proc. Natl. Acad. Sci. U. S. A. 87 (1990) 9640–9644.
[46] G. Willcox, Anthropology. The roots of cultivation in southwestern Asia, Science 341 (2013) 39–40.
[47] A. Toronto International Data Release Workshop, E. Birney, T.J. Hudson, E.D. Green, C. Gunter, S. Eddy, J. Rogers, J.R. Harris, S.D. Ehrlich, R. Apweiler, et al., Prepublication data sharing, Nature 461 (2009) 168–170.
[48] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.