Genomic and Genetic Database Resources for the Grasses[W]

Kevin L. Childs*

Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824

In the strict sense, a biological database is not a Web site. However, the interface that a researcher uses to access the data stored within a database is almost always a Web site. Nonetheless, the two terms are synonymous in the minds of most biologists. With the advent of high-throughput technologies in the last decade, the amount of data that is generated by sequencing, mapping, and expression analysis experiments can only be properly stored in databases. Furthermore, the relationships between the data in biological databases are so complex that the best method to allow wide data access by biologists is via a user-friendly, web-based interface.

Databases have become an essential tool for research in the grasses. An ideal grass database would utilize a genome sequence to provide a framework for other annotations and biological features that are derived from the genes. With the advent of high-throughput sequencing, bioinformatic, and functional genomic methods, genome sequences can be annotated with gene models depicting the exon, intron, and untranslated region structure of genes, functional descriptions of genes, numerous alignment results, promoter annotations, protein interactions, and expression data. The database should also include phenotypic data, germplasm descriptions (including mutant lines), allelic variation, and genetic maps. Therefore, the ideal database would integrate information from the organismal to the sequence level and would allow a biologist to search for any piece of information using any data type. For example, beginning with a sequence, a biologist would be able to find a gene model and integrated annotation for the gene model that would display relevant allelic variation, mutant cultivars, phenotypic and functional descriptions, and functionally related genes. Most importantly, searching the database for any of these data types would allow a biologist to traverse back to gene sequences.

Currently, this ideal database does not exist for any grass species. Most of the data types mentioned above can be found in at least one or a few grass databases, but at best, it is necessary to utilize more than one database to complete all of the searches described above. This article will review grass databases that contain genome sequence, annotation, and genetic resource data. The completeness, quality, and interconnectedness of those databases will be discussed.

WEB SITE QUALITY AFFECTS DATA USABILITY

While the content of a biological database is important, equally significant is the quality of the Web site that allows biologists to access the data. Before discussing individual grass database Web sites, a few words on Web site quality are warranted. The sole purpose for a biological database Web site is to present data in a clear and concise manner. Anything that detracts from this goal is a detriment to the database. Anything that makes a visitor's experience with a database frustrating reflects poorly on both the maintainers of the database and, by inference, the data itself. Anyone with a basic understanding of the type of biological data in a particular database should be able to quickly master how to locate their data of interest. Simplicity is key for an easily navigable database Web site. Time saved in poor Web site design results in database visitors needlessly expending their own time trying to figure out how to negotiate a confusing Web site. Thoughtful planning when designing a database Web site can go a long way to make a complex network of data a truly useful resource.

TYPES OF COMMON GRASS DATABASES

The grass databases are usually distinguished by a theme. Some only focus on data from a single species. Others contain data from a few related species or are simply general plant databases. The grass databases also typically focus on a narrow range of biological topics. The genome annotation databases are well known, but there are also databases that emphasize genetic or germplasm data. Highly specialized databases can provide information about a single biological data type such as transcription factors. The grass databases for which there are the most data are those that provide: (1) genetic information and/or (2) genomic sequence and annotation, specifically for rice (*Oryza sativa*), maize (*Zea mays*), and sorghum (*Sorghum bicolor*). There are relatively few dedicated databases for other members of the Poaceae. Because of the variable states of the genome sequence, genetic resources, and

* E-mail kchilds@plantbiology.msu.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Kevin L. Childs (kchilds@plantbiology.msu.edu).

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.108.129593

overall level of research for the grasses, the content and quality of the databases that are available for each species is quite uneven. There are too many grass databases to cover all in a short review, but a summary of many grass databases, their features, and their URLs can be found in Supplemental Table S1. Those that will be covered here each present a substantial amount of data and have added value to those data by presenting additional analyses or by relating data from multiple sources in a way that provides new biological insights. Of course, all of the grass databases are being updated regularly, and researchers should directly check these online resources for the latest data types and features that are available.

GENOME ANNOTATION DATABASES

Currently, genome sequence and annotation data exists for rice, maize, and sorghum. The three main rice genome annotation resources are the Michigan State University Rice Genome Annotation Project (formerly hosted at The Institute for Genomic Research), the Rice Annotation Project (RAP), and the Rice Information System (Rise; Zhao et al., 2004; Ouyang et al., 2007; Tanaka et al., 2008). The maize genome sequence can be viewed at the MaizeSequence database (http://www.maizesequence.org). Sorghum genome sequence and annotation is provided by the sorghum section of the Joint Genome Institute (JGI) Eukaryotic Genomics Web site (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html). Genome annotation databases are collections of sequences, loci, gene models, and descriptions of sequence features. Most genome annotation databases also provide a functional description for their gene models. Typically, functional descriptions are automatically generated, and this results in starkly utilitiarian characterizations of the possible role each gene. Only in cases where genes are manually annotated is it likely that more familiar classic gene names or symbols are appended to the functional descriptions. Graphical depictions of alignments of interspecific gene and protein sequences to gene models are often available at genome annotation Web sites to allow users to judge the quality of derived gene models and to assist in functional assignments. Raw sequence alignments are generally not available because of the large storage requirements to make such data accessible, but researchers can use the blast servers at the genome annotation sites to regenerate specific alignment results. Some of the other annotation types that are available for grass genomes are listed in Supplemental Table S1.

Each of the rice databases uses its own genome assembly of *japonica* (Rise, RAP, and MSU) or *indica* (Rise) rice (Zhao et al., 2004; Ouyang et al., 2007; Tanaka et al., 2008). All three projects each have their own gene model sets and provide genome browsers to graphically view their annotations (Supplemental Table S1). Although Rise is the only source for annotation of the *indica* genome sequence, the breadth of annotation at Rise is limited. The MSU and RAP resources provide more extensive annotation that includes alignment and protein domain analyses of gene models that allow users to expand on the provided functional definitions of individual genes. The RAP gene models are primarily based on rice full-length cDNAs, but they also make use of de novo gene predictions and partial cDNA sequences. All functional annotation of RAP gene models was reviewed during annotation jamborees. The MSU gene models are the product of FGENESH gene model predictions that were algorithmically refined by transcript evidence, and a subset of these models were subjected to manual curation (Salamov and Solovyev, 2000; Haas et al., 2003). The assignment of functional annotations of the MSU gene models was performed by an automated process and was supplemented by community annotation (Thibaud-Nissen et al., 2007).

Annotation for the maize genome is available from the MaizeSequence database (http://www.maizesequence.org; Supplemental Table S1). Because sequencing of maize is ongoing, pseudomolecules do not yet exist, and all maize sequence is only available as bacterial artificial chromosome (BAC) clone sequence. However, extensive physical mapping data and genetic marker data do exist and are related to each other through graphical browsers. Genome annotations are viewable on a single BAC basis using a sequence browser although track descriptions are not available at this time, thereby complicating interpretation of the annotation. The gene models are the result of automatic pipelines but have not yet been assigned functional annotation descriptions. Protein domain-level alignments do exist within individual gene description pages to allow users to make their own functional annotation assignments, but to learn the functional description of any protein domain, the user must access an external protein domain database. Because annotation is provided at the level of presumably overlapping BACs, it is unclear how much redundancy exists in the gene model set. As the maize sequencing effort progresses and a draft genome sequence is produced, presumably a nonredundant gene model set will be created and more complete functional annotations will be assigned.

A draft, whole genome shotgun version of the sorghum genome is available. Preliminary access to the initial annotation of the sorghum genome sequence is provided by the JGI Eukaryotic Genomics Web site (http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html; Supplemental Table S1). The JGI sorghum site has the official sorghum gene models, but the precise method by which these gene models were derived is unclear. The JGI sorghum database also contains gene models generated by standard prediction programs as well as transcript alignments against which the official sorghum gene models may be judged. While functional descriptions are lacking, interspecific protein alignments and protein domain analyses at the JGI Web site allow users to perform their own functional assignments.

While the databases that have been mentioned so far each focuses on a single species, Gramene is a resource for comparative grass genomics (Liang et al., 2008). It is

unique among grass databases because of the range of data that it provides and the range of species from which the data are derived. A key feature of Gramene is that genome annotations are presented for rice (both *indica* and *japonica*), the short arm of chromosome 3 from *Oryza glaberrima*, and sorghum. For maize genomic sequence, users are redirected to the MaizeSequence database, but hopefully, when the maize genome sequence is released, it will be fully integrated into this database. Gramene obtains its genome assemblies, gene models, and functional annotation from the other major grass genome annotation databases, but Gramene also constructs gene models using GeneBuilder and FGENESH so that there are gene models associated with each genome that have been built by common methods (Milanesi et al., 1999; Salamov and Solovyev, 2000). The definitions of the browser tracks at Gramene are unavailable, making full interpretation of the annotations difficult. Gramene is more than just a genome annotation resource. For rice and sorghum, genes have been mapped to biochemical pathways using the MetaCyc pathway database (Caspi et al., 2008). Additional functional annotation has been performed at Gramene by manual curation that involves literature reviews with gene sequences, and a large number of gene sequences have been assigned their classical gene symbols. Most importantly for a comparative genome database, genetic map, marker, and quantitative trait loci (QTL) data can be searched and compared for several grass species. Comparative map analysis is easily performed between species so that data from one species can be used to leverage data in a second species.

GENETIC RESOURCES RELATED TO GENES, MAPS, AND FUNCTIONAL TRAITS

Although the grass genome annotation databases are invaluable to grass scientists, databases that are focused on genetic markers, maps, and mutant and natural germplasms provide an equally useful resource. While gene models and functional annotations are important for grass biologists, many research projects begin with a population segregating for an interesting trait, and relating that trait to existing maps can speed the identification of associated markers, loci, and possibly sequences. Other scientists studying a particular biological trait want to be able to identify germplasm with a relevant phenotype. Resources that can aid these types of inquiries are available through genetic databases.

The Oryzabase database contains data about available rice germplasm, mapping populations, and mutant stocks (Kurata and Yamazaki, 2006). Oryzabase contains descriptions of a variety of available germplasm and has images of relevant phenotypic traits. Genetic and comparative maps are also found at Oryzabase. Where possible, mapped markers are linked to germplasm accessions. Although this site also has genome browsers with gene models and EST/cDNA features for the *indica* and *japonica* genomes, integration between the browsers and the mapping and marker data is limited.

For the maize community, the MaizeGDB database can be used to search for genetic and QTL maps, additional markers, germplasm resources, and functional annotation (Lawrence, 2007). The number of genetic and QTL maps that are available for maize is large. MaizeGDB allows comparisons of maps with common markers. Descriptions of individual markers are extensive, and mutant germplasm is associated with many classic markers. Genetic maps are displayed as simple marker lists, but hopefully, a future update will include the use of a graphical map display tool to make intermap comparisons more intuitive. Functional annotation is available for gene loci, mutant phenotypes, and metabolic pathways, and functional term searches can return results that are linked to a molecular marker or mapped genes. Functional descriptions are associated with sequence data but not to official maize gene model sequences. For researchers who are interested in finding maize plants with particular mutations, MaizeGDB has information about stocks available from the Maize Genetics Stock Center and phenotypic variation in that germplasm. Additionally, transposon and EMS mutant populations are characterized. The germplasm data descriptions include images and are linked to genetic maps when possible. MaizeGDB does not currently have a working browser, although one is anticipated in the future. Linking sequence and marker data to a tiled maize BAC browser would greatly enhance this resource.

Another database of maize mapping, marker, phenotype, germplasm, and sequence data is Panzea (Canaran et al., 2008). Unlike MaizeGDB, Panzea has generated the majority of the marker and sequence data that it presents. In particular, the Panzea project has generated sequences from thousands of loci from maize inbred lines and teosinte (*Zea mays* ssp. *parviglumis*) to produce a survey of the nature of the polymorphisms that are present in maize. A unique feature of Panzea is the display of the geographic distribution of genetic variation using Google Maps software. By searching on a marker or gene, links to sequence, polymorphism, and geographic distributions can be made. Searches can be performed that result in displays of all polymorphisms that exist between two maize accessions, and the sequence context of these polymorphisms can be viewed. Sequence variations are not shown relative to gene models, but biologists would find it very interesting if such relations could be presented in the future. Genetic and physical maps can be displayed in both graphical and tabular format. Common linkages between maps are easily displayed and all map features can be traced back to known polymorphisms in particular cultivars. The data presented at Panzea is extensive. It does take some time to become proficient at navigating the database. However, in recognition of this fact, tutorials and a use case scenario are provided.

Genetic data for the Triticeae and oats (*Avena sativa*) can be found in the GrainGenes database (O'Sullivan,

2007). Although genomic sequence resources are scarce for these small grain grasses, mapping data for these species are richly developed. Similar to Oryzabase, MaizeGDB, and Panzea, GrainGenes allows users to query and view genetic and QTL maps, markers, and sequences. Generally, researchers can move easily between maps, marker, and sequence data. Map data is graphically viewable, and a subset of marker data is available via a genome browser. The browser data is less well developed, and from some browser annotations it is possible for a user to fall into an outdated AceDB-based section of GrainGenes. Unlike the other genetic databases described above, the germplasm section of GrainGenes is a simple list of sources for small grain germplasm, and therefore, germplasm is not yet related to the map and marker data.

While not containing the same variety of data as the genetic databases mentioned so far, the National Center for Biotechnology Information (NCBI) Map Viewer contains genetic map data for 14 grass species (Tatusova et al., 2007). Maps are easily searched and displayed, and comparative map views are convenient. However, ancillary data is lacking. Additional marker data is provided by links to GrainGenes, MaizeGDB, or entries in the GenBank, UniSTS, and Probe sections of NCBI. The exception to this is rice that has gene model and transcript data aligned to the genomic sequence, but no markers are placed within the genomic sequence. The main reason to use the NCBI Map Viewer is that your species of interest may not have its map data provided by any other more fully integrated public database, and by using the NCBI Map Viewer, it will be possible to find similarities with other species that have more extensively developed genetic and genomic resources.

TRANSCRIPT ASSEMBLIES AS PROXIES FOR GENE SETS

Besides the major sequence databanks, sequence resources for other grass species are limited. However, the Plant Genome Database (PlantGDB), the Gene Index Project, the Plant Transcript Assemblies database, and the NCBI UniGenes project assemble cDNA and EST transcript sequences from individual species into contiguous sequences that represent putative mRNA transcripts (Lee et al., 2005; Childs et al., 2007; Duvick et al., 2008; Wheeler et al., 2008). The Gene Index Project uses GenBank gene entries in addition to cDNA and EST sequences to make transcript assemblies (Lee et al., 2005). The collection of transcript assemblies from a single species represents a subset of the possible coding potential of that species. The transcript assemblies provided by these databases are very popular because they are the best-existing representation of gene sets from grass species that are not yet targeted for genome sequencing. All of these databases have sets of transcript assemblies from numerous grass species, and all of the projects provide functional annotation for their transcript assemblies. Unfortunately, due to sequencing errors and natural cultivar sequence variation, the assembly process can result in many more putative transcripts than are actually produced by any given gene. These databases are self contained and do not offer linkages with other plant genetic or genomic resources, but genome annotation projects often display alignments of these transcripts assemblies within their genome browsers.

CONTINUING CHALLENGE FOR POACEAE DATABASES

The databases that have been discussed here are some of the more popular and useful grass databases, and the goal of each of these databases is to allow plant scientists to access complex data in a convenient manner so that they can advance their research. Successful grass databases accomplish this goal by making easy to use Web sites, by providing results from complex analyses, and by supplying researchers with useful computational tools. However, despite the success of grass databases, there are challenges for the developers of these resources. Given the reality that not all data for a given species will exist at a single database, there need to be convenient mechanisms for biologists to move between databases. Often gene models act as a link from one database to another, but this is not always a reciprocal relationship. A technology called the Semantic Web has been promoted to allow biologists to make connections between unrelated data sources, but the Semantic Web has not advanced to the point that it has been widely adopted by database developers (Good and Wilkinson, 2006; Pasquier, 2008). Until this technology matures or another technology is developed, the only mechanism to ensure that biologists have convenient access to all available data is for database developers to cooperate and share sufficient information so that links can be easily established to allow biologists to move between databases. The continued success of grass database projects will be measured by how easily grass researchers are able to discover data that allows them to gain insight into the biology at the center of their research.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Table S1. Data types available at grass database Web sites.

Received September 8, 2008; accepted November 6, 2008; published January 7, 2009.

LITERATURE CITED

Canaran P, Buckler ES, Glaubitz JC, Stein L, Sun Q, Zhao W, Ware D (2008) Panzea: an update on new content and features. Nucleic Acids Res **36:** D1041–D1043

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al (2008) The MetaCyc

Database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res **36:** D623–D631

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP (2007) The TIGR Plant Transcript Assemblies database. Nucleic Acids Res **35:** D846–D851

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2008) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res **36:** D959–D965

Good BM, Wilkinson MD (2006) The Life Sciences Semantic Web is full of creeps! Brief Bioinform **7:** 275–286

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654–5666

Kurata N, Yamazaki Y (2006) Oryzabase: an integrated biological and genome information database for rice. Plant Physiol **140:** 12–17

Lawrence CJ (2007) MaizeGDB: the Maize Genetics and Genomics Database. Methods Mol Biol **406:** 331–346

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res **33:** D71–D74

Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, et al (2008) Gramene: a growing plant comparative genomics resource. Nucleic Acids Res **36:** D947–D953

Milanesi L, D'Angelo D, Rogozin IB (1999) GeneBuilder: interactive in silico prediction of gene structure. Bioinformatics **15:** 612–621

O'Sullivan H (2007) GrainGenes: a genomic database for Triticeae and Avena. Methods Mol Biol **406:** 301–314

Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res **35:** D883–D887

Pasquier C (2008) Biological data integration using Semantic Web technologies. Biochimie **90:** 584–594

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res **10:** 516–522

Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, et al (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res **36:** D1028–D1033

Tatusova T, Smith-White B, Ostell J (2007) A collection of plant-specific genomic data and resources at NCBI. Methods Mol Biol **406:** 61–88

Thibaud-Nissen F, Campbell M, Hamilton JP, Zhu W, Buell CR (2007) EuCAP, a Eukaryotic Community Annotation Package, and its application to the rice genome. BMC Genomics **8:** 388

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36:** D13–D21

Zhao W, Wang J, He X, Huang X, Jiao Y, Dai M, Wei S, Fu J, Chen Y, Ren X, et al (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. Nucleic Acids Res **32:** D377–D382