Plant Biology Databases: A Needs Assessment
November 16, 2005

Authors:

Bill Beavis	<wdb@ncgr.org>	National Center for Genome Resources
Damian Gessler	<ddg@ncgr.org>	National Center for Genome Resources
Sue Rhee	<rhee@acoma.stanford.edu>	Carnegie Institution of Washington
Dan Rokhsar	<DSRokhsar@lbl.gov>	Lawrence Berkeley Laboratory
Doreen Main	<dorrie@genome.clemson.edu>	Clemson University
Lukas Mueller	<lam87@cornell.edu>	Cornell University
Eva Huala	<huala@acoma.stanford.edu>	Carnegie Institution of Washington
Lincoln Stein	<steinl@cshl.edu>	Cold Spring Harbor Laboratory
Carolyn Lawrence	<triffid@iastate.edu>	USDA ARS

ABSTRACT

We review the anticipated needs of the plant genome research community for long-lived data collections. We find that there is an increasing need for such repositories, and offer guidelines for balancing the funding of data production projects with those aimed to manage and integrate the data. In particular, we find that there is a pressing need to develop a trained cadre of skilled knowledge workers who are able to curate complex biological data, and to provide this cadre with a system of stable funding that enables data repositories to be established and maintained over extended periods of time. We note approvingly the current trend of species-specific databases to expand into comparative genomics-minded clade-oriented databases, but caution that new technologies are needed to facilitate the transparent integration of data among these databases.

Conclusions and specific recommendations begin on page 34.

INTRODUCTION

A recent draft report from the National Science Board (NSB)—an oversight board of the National Science Foundation (NSF)—recommended that the NSF establish an "agency-wide umbrella strategy" for maintaining and enhancing long-lived data collections (Pennisi 2005; NSB 2005). "Data collections" is broadly inclusive of the digital data itself as well as the brick and mortar and personnel infrastructure needed to maintain the resource in a state that is useful to the scientific community. "Long-lived" refers to resources that have life spans that exceed technological generations, so they must adapt their technological implementations while maintaining or advancing their functionality.

The rationale for establishing such long-lived data collections is straightforward. The NSF

and USDA together have invested hundreds of millions of dollars in research grants to generate large-scale data sets, most notably in the field of genomics. These data sets will have significance to the research community for an extended period of time, in some cases far longer than the duration of the research grant that originally funded their generation. In order to preserve this investment, the NSF and USDA need a strategy to identify and support groups to maintain these data sets. Another reason for investing in long-lived data collections arises from the value of data integration. It is often the case that two data sets, when integrated, are far more useful than the two data sets taken individually. An obvious example is the case of a genome sequence and a collection of clustered ESTs (expressed sequence tags). Taken individually, the genome sequence provides poor information about the location and nature of genes because of the inaccuracy of *ab initio* gene prediction, and the EST collection provides little information on gene structure and rarely provides the full-length coding sequence. Taken together, however, the genome sequence and EST collection provide a more accurate and comprehensive view of the content and structure of the genes in the genome. This latter point argues for the establishment of "living" data repositories in which the information is actively curated, as opposed to "dead" repositories of static FTP sites.

The establishment of long-lived data collections for plant biological data has been somewhat patchy to date. During the 1990's, the USDA-ARS established a series of species-specific databases for maize, rice, wheat, soy and other species of agronomic importance, but the results were not always satisfactory, possibly due to scarcity of resources. After a recent consolidation in the number of databases funded by the ARS, those databases that remained have shown robust growth, most notably MaizeGDB and the Legume Information System (LIS). The NSF DBI has been reluctant to commit long-term resources to database projects, but when it has committed substantial resources to data collections, it has had notable successes as evidenced by TAIR, Gramene and TIGR. However, it is unclear whether the current paradigm of establishing species-specific databases in response to investigator-initiated research proposals is the most efficient and forward-looking strategy.

This document looks at the nature of current and future biological data sets, and attempts to provide a framework on which administrators at NSF and USDA can manage the need for long-term data collections.

Definitions

We lead this document with a number of definitions.

Static Repository – A static data repository is an unchanging archive of information. An example of a static repository is an FTP site containing data files from a SNP discovery project. Static repositories are typically read-only so that once published, they change

rarely if at all. Compared to curated repositories, static repositories are relatively inexpensive to set up and maintain.

Curated Repository – A curated data repository is under active management. Data sets are reanalyzed on a regular basis in order to integrate them with each other and to find and correct inconsistencies within the data sets. The managers of this type of resource inject their own editorial judgment into the process in order to create an integrated data set that represents their best estimate of reality. Curated repositories are often built on top of database management systems and web-based interfaces that invite researchers to explore the connections among the component data sets.

Stock Center – A stock center is a repository of physical reagents, such as seed stocks, clones, vectors, and cell lines. It incorporates a database that describes its holdings and often offers an online catalog function that allows browsing and electronic ordering. The stock center database ideally should create a public interface for accessing its catalog, thereby allowing data repositories to create cross-references to stock center holdings.

MOD – Model organism database. This is a curated repository that focuses on a particular species. MODs are often been formed spontaneously by a research community in order to track reagents and other shared information resources needed by the community.

COD – Clade oriented database. These are a new breed of curated repositories that focus on multiple related species, for example vertebrates.

Data Set Annotation – Data set annotation is the process by which third parties add value to existing data sets using combinations of informatics tools and human judgment. Examples include predicting genes on genome sequences, identifying the genomic locations of genetic markers, establishing the correspondence between quantitative trait loci (QTLs) among two or more species based on common traits, or adding human-readable descriptions of gene function to gene records. Annotation is a service commonly made available at curated repositories.

Automated Annotation – Automated annotation is the result of running a computational pipeline on a data set. Examples of automated annotation include gene prediction, EST clustering, and ortholog set development. Automated annotation systems are expensive to set up because of the investment in software and algorithmic development required, but once established their maintenance costs are modest. A further characteristic of automated annotation is that these processes do not usually require personnel who have a detailed knowledge of the biology of the organism, because most automated annotation pipelines are species-independent. For example, an EST clustering system set up to work on poplar will also produce satisfactory results for tomato.

Manual Annotation – Manual annotation requires the judgment of a human being and is

characterized by activities that require the integration of information from multiple data sets and from the scientific literature. Examples of manual annotation activities include gene ontology annotation, the interpretation of targeted gene knockout studies, and the classification of the traits measured in a QTL study. In contrast to automated annotation, manual annotation systems may have low startup costs (they can start with one postdoc's part-time activity and grow from there), but do not decrease in cost during the lifetime of the project.

Data Providers – These are the producers of data sets, typically teams of bench biologists, computational biologists, and bioinformaticians. The managers of data repositories, whether of the static or curated types, either create interfaces that allow data providers to submit their data without assistance, or actively seek out the data providers and assist them in making their data available through the repository.

End-Users – These are consumers of the data sets, typically bench biologists. Naïve end-users require easy-to-use and intuitive interfaces that nevertheless provide them with access to the full data set. These users are often satisfied with one-object-at-a-time interfaces, such as those provided by almost all biological databases. More sophisticated users require query interfaces that allow them to integrate multiple data sets within the current repository, functionality that a few of the larger databases provide. The most sophisticated users wish to integrate multiple data sets across multiple repositories, a type of functionality that is rare in all but a few restricted cases.

Evidence and Attribution Tracking – Evidence tracking links an assertion contained within a repository to the underlying evidence that supports that assertion. For example, an assertion about the genes a transcription factor regulates may be supported by a paper that describes a knockdown of the transcription factor. Curated repositories need to scrupulously document the chain of evidence in order to prevent unsubstantiated facts from "magically" appearing in the database. Attribution tracking links a data set and annotations on the data set to the individual or group that produced it. In actively curated data sets, there is always a risk of losing attribution information. Because the data has been heavily worked over, end users lose track of where the data originated. This is not ideal, as it discourages data providers from submitting their sets, while simultaneously encouraging end users to treat the information as if it had magically truthful properties. Managers of curated repositories try to avoid this trap by propagating correct attributions and evidence tracking throughout the data.

Attribution tracking links a data set and annotations on the data set to the individual or group that produced that data set. In actively curated data sets, there is always a risk of losing attribution data. Because the data has been heavily worked over, end users lose track of where the data originated. This is not ideal, as it discourages data providers from submitting their sets, while simultaneously encouraging end users to treat the

information as if it had magically truthful properties. Managers of curated repositories try to avoid this trap by propagating correct attributions throughout the data.

Ontologies – Ontologies are sets of vocabulary terms whose meanings and relations with other terms are explicitly stated in such a way as to be comprehensible to humans and computer programs. For example, the Gene Ontology describes the function of genes. Ontology-building has emerged as a major activity of curated repositories because by annotating data sets using a shared set of ontologies, repositories can establish connections both within the data sets they curate and across data sets contained within different repositories.

The Bioinformatics Food Chain

Over time, a food chain of sorts has arisen within bioinformatics (Figure 1). An understanding of how this food chain works can assist in making decisions on how to balance competing demands on resources.

At the bottom of the food chain are LIMS (laboratory information management) systems. These are highly customized laboratory-specific systems responsible for managing the internal processes of a data provider. In the genome sequencing world, a typical LIMS system would manage the robots that set up automated sequencing runs.

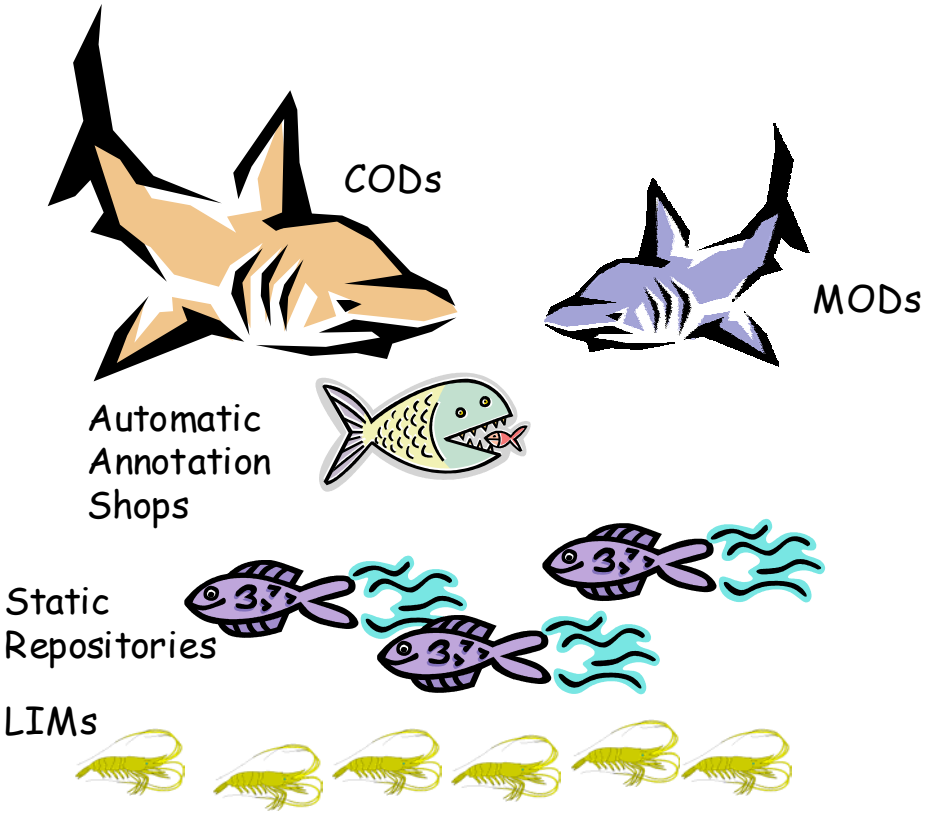

Figure 1: The Bioinformatics Food Chain

Next in the food chain are the static data repositories that are responsible for providing long-term storage for the information generated by the data providers. The primary duty of these repositories is to provide a stable, time-stamped and versioned record of the raw data. In genome sequencing, the classic example of this is GenBank (Benson *et al* 2004), which records sequence submissions. Other examples of static repositories include GEO (Barrett *et al* 2005), a repository of microarray expression data, and PDB (Westbrook *et al* 2003), a repository of x-ray crystallographic structures.

Above this level are the automatic annotation shops. These are enterprises that add value to the information contained in static repositories by performing automated annotation across the data set, producing a new set of annotations. Ensembl (Birney *et al.* 2004) is a good example of an automatic annotation shop. Its primary mission is to predict protein-coding genes on genomes using a highly automated and consistent pipeline. PlantGDB (Dong *et al.* 2005) performs consistent automatic EST assembly and annotation across multiple plant species.

The information produced by automatic annotation shops is in turn taken up by model organism databases (MODs). These are community databases focused on a single species or group of related species. MODs take the information provided by automatic annotation

shops, enhance it with manual curation, integrate it with information from the literature, and relate it to other data sets and resources. In the plant genomics world, The Arabidopsis Information Resource, TAIR (Rhee *et al.* 2003), is the oldest and best established MOD.

At the top of the food chain is a new breed of databases that we call "clade-oriented databases" or CODs, for a want of a better term. The CODs are multi-species databases, which usually have a clade-specific emphasis. They integrate information from the static data repositories, annotation shops, and MODs into a single integrated database designed expressly for making comparisons among species. The best-known database of this type is the UCSC Genome Browser (Karolchik *et al.* 2003), which contains information on all vertebrate genomes and selected model organism animals such as yeast, fly and worm. The best extant examples in the plant genomics world are Gramene (Ware *et al.* 2002) and LIS (Gonzales *et al.* 2005), which are CODs for monocots and legumes respectively.

It is important to realize that these categories are not mutually exclusive. Many databases combine these categories. For example, NCBI manages GenBank, a static repository of nucleotide sequences, a curation shop, the NCBI human gene build, and EntrezGenes (REF), which is essentially a set of mini-MODs.

PLANT BIOLOGY DATA SETS AND THEIR REQUIREMENTS

The next sections will describe the types of data sets relevant to plant biology and the long-term data gathering, integration, and analysis activities needed to maintain their value.

Genome Sequencing and Mapping

The process of genome mapping and sequencing generates a large number of reagents and information resources, including:

1. Marker collections – PCR primer pairs, oligos, clone end sequences, and other collections of markers used for identifying genomic positions.

2. Clone libraries – cDNA libraries, BAC, fosmid and other libraries that act as a valuable laboratory reagent long after the mapping and sequencing is over.

3. Physical maps –All cytological and sequence-based maps are in fact physical maps, but most often the label "physical map" is used to describe the information that describes the order and orientation of the members of clone libraries on a given genome. (*Genetic Maps and Variation* discusses Genetic Maps).

4. Raw sequence reads – Sequencing trace files, nucleotide reads, and quality score

files that are the raw evidence for the genomic sequence.

5. Genome assemblies – Long-range genomic sequence assembled from raw reads using sequence assembly algorithms.

Both static and curated repositories are needed to support these activities (Table 1). Static repositories that allow occasional correction of the information are sufficient to manage the marker collections, raw sequence reads, and the information associated with clone libraries, because these data, once generated, do not change frequently. Stock centers can manage the probes needed to detect RFLP-based markers.

However, physical maps and sequence assemblies are dynamic, changing by way of each annotation and refinement update. Physical maps typically require active curation for a period of years after their initial generation, and genome assemblies, at least for eukaryotes, appear to require active curation indefinitely (even the oldest and simplest of the eukaryotic assemblies, that of *S. cerevisiae*, is still being updated). If the clone library is intended to be a long-lasting reagent, a stock center is needed to maintain and distribute it.

The assembly and curation of physical maps requires a group that is skilled in the operation of such software as FPC. Typically physical map assembly is an iterative process that involves experimental validation at the bench, making it useful for physical map assembly and maintenance to be co-located with the laboratory that develops the clone libraries and fingerprints. After a physical map has been published, the tasks of annotating and integrating it with other data can be taken up by the curated repositories, which will increase its usefulness and value to the community.

Genome assembly is a more complex situation. There are typically three phases of the process, a rough "draft" assembly followed by a finishing phase, followed in turn by a maintenance phase. The draft assembly is both computation-intensive and dependent on sophisticated (and somewhat finicky) software, but it requires no laboratory intervention once the first set of reads has been developed. One can envision draft assemblies being performed by a specialist third party group unaffiliated with the sequencing laboratories. The finishing phase, however, involves an iterative process of human and computational inspection of sequence, laboratory experimentation, and refinement of the assembly. Finishing always takes place in the sequencing laboratories.

After the genomic assembly is "finished," it enters an important maintenance phase that has not received much attention. As the assembly is annotated (see next section) and the scientific community uses the assembled sequence in their research, discrepancies and other problems are inevitably discovered. Ideally, these problems should be resolved -- or at least formally noted -- and used to incrementally improve the assembly. This task calls for a curated repository that can act as the focal point for genome annotation, community

feedback, and the management of assembly updates and version-controlled releases.

Historically, sequencing centers have not been good fits for genome assembly maintenance and the responsibility for this activity has been taken on by MODs and more recently by CODs (crop monocots in Gramene and Medicago and Lotus in LIS). On occasion, the activity has been mired in disputes over the "ownership" of the sequence, leading to periods in which a genomic sequence has stagnated. The absence of a clearly-defined center that can receive and act on complaints about problems in the assembly leads to frustration among the end-users and loss of confidence in the assembly; this is an outcome to be avoided.

It is important to state clearly that physical mapping and genome sequencing and assembly are mutually dependent activities that are usually independent of the species or clade of the organism being sequenced. Therefore the static and curated repositories that support these activities can easily be managed by centers that operate on multiple species and do not need to bring any species-specific expertise to bear. There are also many existing facilities that can manage this type of data: for example, GenBank is the obvious choice for the static repositories for marker collections, sequence reads, traces and assembly versions.

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
Marker Collections	X	X	X	Curation involves assignment of markers to genomes which is part of genome annotation; no species-specific knowledge needed.
Clone Libraries	X		X	Use existing repositories for static data
Physical Maps	X	X	X	Does not require species-specific knowledge
Draft Genome	X	X		Static repository needed for reads – use NCBI. Curated repository need for assembly, but no species-specific knowledge (usually) required
Finished Genome		X		Manage community input; species-specific knowledge helpful but not required

Table 1: Types of Data Repositories Needed for Genome Sequencing & Mapping

Summary recommendations for genome sequencing and mapping projects:

- Molecular markers (genetic and physical) should be submitted to NCBI GenBank.

- Clone libraries should be submitted to stock centers.

- A standard file format should be developed for representing physical maps. Physical maps should be curated at the MOD or COD level.

- Genome assemblies must be curated and maintained after the original sequencing centers have moved on. Sequencing projects must develop a plan for the orderly handing over of the assembly to a repository that can manage updates of the assembly in response to community feedback and/or additional experimental data.

Genome Annotation

After the production of a genome sequence, the next task is to add meaning to it via a process of annotation. Genome annotation spans the gamut from identifying the location of previously-identified cDNA sequences in the genome, to characterizing the interactions among different gene family members, and involves making inferences at the nucleotide, protein, and biological process levels (Stein 2001). The first steps of genome annotation are typically to identify repetitive elements, and to align ESTs, cDNAs, protein sequences and molecular markers (such as STSs) to the genome. The next step is to create a set of gene predictions, both for protein-coding and non-coding (e.g. miRNA) genes. This is followed by an involved process of annotating the genes and their products; typical steps involve identifying recognizable protein domains in the products of protein-coding genes, describing the function of gene products using the Gene Ontology and other controlled vocabularies, and integrating these annotations with information on gene product expression patterns and molecular interactions garnered from other high-throughput experimental data sets.

Much more so than mapping and sequencing, genome annotation is a dynamic ongoing process. This is so because the annotation of an organism's genome blends imperceptibly into the understanding of the organism's biology. An understanding of the genome's "parts list" leads to new discoveries at the bench. Techniques developed during the pursuit of hypothesis-driven research leads to new data sets that enhance the quality of genome annotation.

Genome annotation may be approached using fully automated methods, or a combination of automated annotation followed by manual curation. Automated genome annotation is essential both for the initial annotation of a newly sequenced genome and for keeping the annotation up to date. Following the automated steps, the annotation may be enhanced by manual curation in order to increase its reliability and coverage. Manual curation involves careful examination of the automated annotations by expert curators, who apply their biological knowledge to identifying flaws in gene predictions, Gene Ontology assignments, and other annotations. Manual curation is also necessary to link the annotated genome to the biological literature so as to provide the critical bridge between

genomics and hypothesis-driven research.

Although requiring a higher initial investment, manual curation to a high standard will result in a dataset which can more easily be maintained by automated processes requiring only limited subsequent manual intervention. Automated pipelines can incorporate newly deposited sequence information much more easily when the initial gene models are confirmed as correct.

Because manual curation is labor intensive, it will not be economically practical to apply it to all genomes. In such a case it is important to choose a "reference genome" that will act as an exemplar for a clade under study. The reference genome should be heavily hand-curated so that its annotations can later be computationally propagated to genomes of related species. The fully-automatic annotation of a genome that has not had the benefit of a hand-annotated close relative is likely to be inferior to one that does.

Regardless of whether it was produced by a fully-automated effort or a combination of automatic and manual curation, the single most important output of an annotation effort is a canonical list of genes and their genomic structure and function. The gene list serves as a reference for the entire research community and is an absolute prerequisite for subsequent studies that attempt to leverage the genome sequence. To be most useful there must be a community consensus on the nature and ownership of the gene list, and there should be a process by which updates to the gene list are tracked so that researchers can recover the name and exact structure of a gene at the time a particular experiment was performed.

Significant long term efforts and costs are required to maintain an annotated genome sequence as a useful resource (Table 2). The maintenance tasks include 1) continuous refinement of gene structures and addition of splice variants using new data (for example new cDNAs or ESTs, genome sequences of related organisms) and improved gene prediction algorithms; 2) updates to gene function annotation (including gene product information and GO function, process and cellular component annotations using both computational and manual literature-based methods); 3) annotation of other objects that can be anchored on the genome, for example cDNA clones, transposons and repeats, mutations including insertional knockouts, and SNPs and other markers that serve as research tools for the utilization of the genome sequence.

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
Canonical Gene List		X		Requires cooperation of both automatic and manual curation groups. Species-specific knowledge required for manual curation, but

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
				not for automatic gene builds.
Aligned reagents	X	X	X	Requires extensive cooperation among static repositories, stock centers and curated repositories.
Protein domains		X		Does not require species-specific knowledge
Gene function (e.g. GO)		X		Automated assignment followed by manual curation.

Table 2: Types of Data Repositories Needed for Genome Annotation

Automatic annotation shops. In plants, where comparative genomics seems likely to play an even more important role than in vertebrate genomics, uniform high-quality automatic annotation is vital both within and among plant genomes. A lack of uniformity has the potential to cripple efforts to achieve high quality genome annotation. However, automatic annotation of plant genomes has, to date, been *ad hoc.* The primary annotation of the Arabidopsis genome was performed by the sequencing group consortium, resulting in a genome in which different chromosomes were initially annotated to different standards (The Arabidopsis Genome Initiative 2000); this has since been remediated by manual curation. The *Oryza sativa* and *indica* genomes were also annotated in a piecemeal fashion, and the confusion is now being exacerbated by redundant, but perhaps unavoidable, genome annotation efforts that have given birth to multiple conflicting gene sets and assemblies. Although there are several sophisticated efforts in this direction, including the International Medicago Genome Annotation Group (IMGAG, www.medicago.org), PlantGDB, and the TIGR genome annotation group (www.tigr.org) it is fair to say that there has yet to emerge a specialist genome annotation shop on par with Ensembl that is widely trusted by the research community to produce a high quality, uniform automatic annotation. We feel there is a strong need for such a facility.

Another vital function of automatic annotation shops is the alignment of sequence-based reagents to the genome. These reagents include MPSS and SAGE tags, EST sequences, BAC end sequences, the oligonucleotides and cDNAs used in microarrays, the flanking sequences of SNPs, and genetic markers. Because of the dynamic nature of both the assembly and the gene annotations, these alignments must be performed on an ongoing basis, and changes in the alignments, such as the movement of an EST from one chromosome to another, must be versioned and tracked. The reason that active curation of alignments is so important is because biological inferences from sequence-based reagents are dependent on the correct relationship between reagent and the genome annotation. For example, if an Affymetrix array is built on top of one version of an annotated genome and then the genome assembly and its annotations are updated in light of new

knowledge, the oligonucleotides chosen for the array may no longer correspond to the genes they were chosen to represent; it is critical for researchers to know how the oligonucleotides on the array relate to the current best gene annotations.

A large number of software tools for automated genome annotation have been developed (Table 3). In principle the automated tools allow any research group with access to a compute cluster to become an annotation shop. In practice, the tools need to be extensively tweaked to accommodate the idiosyncrasies of each genome, and this, in turn, requires a good understanding of the organism's biology. In addition, a considerable amount of computer science sophistication is required to construct and manage an automatic annotation pipeline. For this reason, there are currently only a handful of groups with the capability to perform consistent automated genome-wide annotation.

Annotation Type	*Description*	*Representative Tools*
Sequence cleansing	The ability to remove superfluous sequences, i.e., vector removal, quality trimming, and poly A/T trimming	
Repeat finding	The ability to identify transposons, microsatellites, and other repetitive elements	RepeatMasker, TIGR's Plant Repeat Databases
Sequence similarity searching	The ability to compare sequences against known proteins and transcripts	BLAST, BLAT
Protein domain identification	The ability to identify protein families, domains and other functional sites	InterPro, BLOCKS, eMOTIFS
Signal peptide cleavage sites	The ability to identify signal peptide cleavage sites.	SignalP
Transcription factor identification	The ability to identify transcription factors and their binding sites	TRANSFAC
Non-coding RNA gene identification	The ability to identify non-coding RNA genes.	RFAM
Gene prediction	The ability to predict the presence and structure of a gene from the genomic nucleotide sequence.	GenScan Fgenesh
GO mapping	The ability to associate a sequence with Gene Ontology terms based on protein domain content and other characteristics.	Interpro2GO
Miscellaneous sequence to genome mapping	The ability to map other useful sequence-based features to the genome (e.g. cDNAs, ESTs, microarray elements, insertion flanks, SNPs, TILLing mutations)	
Transcript mapping	The ability to annotate new genes and update existing gene models based on transcript data	GeneWise, Exonerate, PASA
Manual curation of gene structures	The ability to manually adjust the structure of gene models (e.g. add new exons or splice variants)	Artemis, Apollo

Annotation Type	*Description*	*Representative Tools*
Literature-based annotation of gene function	The ability to assign functional annotations go genes from literature sources, using free text and/or ontologies.	PubSearch, Textpresso, Manatee
Manage community curation	The ability to accept corrections and new information from community submissions.	AtGDB, HAVANA

Table 3: Software tools for genome annotation

Some effort has also gone into developing software frameworks for automated sequence annotation (Hoon *et al.* 2003; Potter *et al.* 2004). These frameworks use a machine-readable protocol to drive pipelines of the various sequence annotation tools. Although the frameworks show promise for facilitating the setting up of an annotation shop, they have a long way to go before they are ready to be used outside their group of origin.

Manual curation. As in the animal genomics world, responsibility for maintaining and enhancing plant genome annotations by manual curation has become the domain of several plant community MODs, including TAIR and TIGR (Lee *et al.* 2005). Because manual curation is strongly tied to the biological literature, to research community needs, and to the various experimental resources for the organism (knockout collections, genetic maps), it seems likely that additional community databases will be needed to come online as new genomes are completed.

As with automatic annotation, a variety of software systems have been developed to assist with manual curation (Table 3).

Community Curation. Community curation of the canonical gene set will be needed to maintain high quality genome annotation in the long term without excessive funding requirements, but community participation is currently quite low. There are technical and social reasons for this lack of participation. The primary technical reason for this is that tools to facilitate community participation are expensive to develop because they need to be robust, easy to use, and provide mechanisms for quality control. In addition they must be sufficiently adaptable to incorporate new kinds of data. The primary social reason for this is that there is little or no reward for curated contributions to community information resources. Furthermore, the tools to incorporate community annotation are currently much more restricted than other areas of genome annotation and with doubts regarding the level of enthusiasm on the part of the communtiy, creating further resources may prove difficult. Additionally while there is a high cost in developing such tools there is also a substantial cost in regard to the manual curation which will likely be required to verify community curation.

At the very least, however, all annotated genomes should have a community feedback

mechanism so that those individuals who find errors and other problems in the canonical gene set can report them and be assured that their reports will be acted on.

Static and curated repositories for genome annotation. The primary repository for the static storage of genome annotation is the genome division of NCBI, also known as GenomeDB (www.ncbi.nlm.nih.gov/Genomes/). This division holds the original automatic annotations from sequencing groups. In some cases, plant MODs have reached agreement with NCBI to transmit the results of their manual curation to GenomeDB, enabling this repository to display up to date information as well. In other cases, the MODs remain the sole curated repository for genome annotation data.

GenomeDB, the MODs and CODs generally provide reliable access to the data and utilize best software engineering practices of versioning and keep information on history and evidence tracking. The main downside of having MODs be the sole repository of current genome annotation information is that this interferes with the ability of users to make comparisons among the genomes, due to historical differences in user interfaces and data representation. The Generic Model Organism Database (GMOD) project (Stein *et al. 2002*) is attempting to remediate this issue by establishing standards for representation of genomic annotation data (see for example, the Sequence Ontology (Eilbeck *et al.* 2005)), but the proposed standards have yet to be widely implemented and have yet to have a measurable impact on the research community.

User interfaces. GenomeDB provides a "one size fits all" user interface that provides basic genome visualization, browsing and querying.

The community databases provide user interfaces to genomic data based on community specifications. These include tools for graphical visualization of sequence data in relation to a genome map, query tools based on community needs, presentation of query results in the context of the biology for the species of interest, and customized bulk data access methods. As noted earlier, the use of highly customized user interfaces is a double-edged sword. While it enhances the user experience for members of a specific research community, it inhibits comparisons among species. For this reason the GMOD project has developed standardized user interface tools for viewing genome and for querying and downloading bulk data sets (Durinck *et al.* 2005). New community databases should be encouraged to adapt existing tools rather than inventing new ones.

Funding. Funding for the community databases is generally based on funding cycles of 3-5 years in length, while support for GenomeDB is tied to NCBI's more stable long-term funding. While plant biologists recognize the need for curation to keep information current, no stable long-term mechanisms for supporting such curation have been developed. In addition, effective management of the dependencies that exist between information resources is extremely difficult, given the lack of standards for versioning

and update/release notification mechanisms. Finally, each data resource is typically provided with its own custom access and interface mechanisms, forcing users to learn a special form of interaction with each provider of data.

In summary, the most pressing needs for plant genome annotation are 1) one or more dedicated annotation shops that can create a set of automatic gene predictions from a virgin genome assembly using a well-understood, reproducible annotation pipeline; 2) a policy for assigning responsibility for the canonical gene list to a group charged with the long-term maintenance and curation of the list; 3) a mechanism for involving the research community in the upkeep of the genome annotation; and 4) a well-supported "portal" for access to aggregated plant genomic data.

Summary recommendations for genome annotation:

- Sequencing projects must develop a plan for developing a public, canonical set of gene predictions over a set period of time using generally accepted best practices for gene prediction. The plan should include a mechanism for accepting and responding to community feedback on incorrect or missing gene models.

- Use of standardized genome annotation pipelines should be encouraged. This will simplify the task of cross-species comparison, and reduce redundant effort.

- Encourage partnerships between manual curation groups and genome annotation shops.

Comparative Genomics

Annotation and analysis of genomes are increasingly informed by comparisons among sequences from closely and distantly related organisms. The importance of these comparisons to plant biology will increase dramatically in the coming years as the number of available genomes grows. The identification and characterization of homologous sequences -- that is, sequences that are related by descent from a common ancestor -- is an essential step in the interpretation of genomes, since the evolutionary relatedness of these sequences across different genomes provides clues to conservation of gene and protein structure and function. Conversely, the sequence-level divergences that are overlaid upon this coarse conservation may be implicated in the diversification of gene function and the emergence of novel traits. Comparative analysis of the genomes provides the much-needed link between functional studies often pursued in model systems and the genetic mapping of traits (e.g. QTLs) that is widespread in crop species.

The dynamic nature of plant genomes makes this characterization particularly challenging, as modern genes or sequence elements may be related to each other through a series of local and/or genome-scale duplication events along one or both lineages. For

example, it is not unusual for a single gene in the common angiosperm ancestor to have given rise to multiple surviving genes in modern plants through a series of shared and/or lineage-specific gene duplications. At the largest scales, networks of tens, hundreds, or even thousands of genes may be conserved across tens of megabases of genomic territory, resulting in long "syntenic" (literally, "same strand") regions within and between genomes. At shorter scales, tandem duplication, divergence, transposition, and loss of individual genes and their associated regulatory sequences are important processes that need to be disentangled.

Ancient polyploidy and diploidization events are an essential part of plant history, including at least two in eurosid lineage leading to *Arabidopsis*, and one in the grass lineage leading to *Oryza*, with additional more recent polyploidizations known in maize, soybean, alfalfa, sugarcane, and other plant species. These superimposed duplications lead to complex hypotheses in which the function of the gene in the angiosperm ancestor may be partitioned, amplified, or otherwise distributed across multiple modern genes, in a potentially genome-specific manner. The analysis of ancient polyploidy is further complicated by the rampant loss of duplicated genes that follows these events, which makes them challenging to identify at the single gene level. Since polyploidy is not common in animals, computational advances in this area are likely to be driven by plant bioinformatics.

Commonly used "best hit" analyses are especially prone to error in the face of the dynamism of plant genomes. If used without the proper caution, such approaches to "functionally" annotate new gene sequences has the potential to contaminate plant sequence databases with faulty nomenclature that will become increasingly unreliable without a combination of new computational methods combined with machine-assisted manual curation of reference genomes distributed across plant phylogeny.

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
Genome to genome alignment		X		Primarily automatic annotation.
Gene families		X		Automatic annotation supplemented by manual curation. Extensive knowledge of gene family properties required. Probably well-suited for community curation.
Conserved functional elements		X		Active research needed. Requires collaboration among computational groups, curatorial groups and experimental groups.

Table 4: Types of Data Repositories Needed for Comparative Genomics

Active curation is needed to manage comparative genomics data (Table 4). Specific needs for comparative plant genomics are methods and tools for:

1. *Characterizing syntenic relationships among plant genomes.* This activity uses protein- and nucleotide-sequence similarity measures, supplemented by mapping data to relate the genomes of multiple species via their syntenic relationships. End-users should be able to navigate across the resulting web of synteny to understand the phylogenetic history of their segment of interest. This activity is a task for curated repositories and involves a combination of automated and manual annotation.

2. *Gene family characterization.* At the whole-gene level, characterize the pattern of duplication, divergence, and loss in each gene lineage in the context of these large-scale genomic events and local tandem events to lead to a complete understanding of the diversification of modern gene families is the long term goal. These phylogenetic efforts must be accompanied by visualization and query tools as well as easy-to-interpret confidence measures that make these, often arcane, studies accessible to the general user. This is largely an automated annotation task to be performed by curated repositories.

4. *Gene structure evolution.* At the sub-gene level, characterize the conservation of gene structure and probe the evolution of alternative splicing patterns, in order to understand possibilities for functional divergence. This annotation activity can probably be automated, but will require extensive research before it is a reality. As the experimental characterization of splice variants is unlikely to keep pace with the increase in raw genomic sequence, we will need computational methods to predict alternative splicing and to represent when and where these variants occur.

5. *Improved detection of non-coding sequences through comparative analysis of genomes*. This includes comparisons within and between genomes. We anticipate that these putative cis-regulatory sequences will be characterized systematically both empirically and computationally, through the integration of datasets from a wide array of experiments built upon genomic data, including expression microarrays, chromatin immunoprecipitation assays, proteomic studies, etc. This is also an automated annotation activity to be carried out by curated repositories, but like (4) it requires extensive research into new experimental methods. Also be aware that this is a type of genome annotation that dovetails with the requirements described in the corresponding section.

6. *User interface.* For this complex type of data to be manageable by end users, curated repositories must set up user interfaces that allow users to navigate the

web of experimentally determined functional data across multiple plant species, with easy access to the source of evidence for functional annotations. In this manner, the true power of comparative genomics can be brought to bear by linking the relatively small number of functional studies to exponentially growing number of sequence resources. The GMOD project provides some portable tools for displaying synteny data (Pan *et al.* 2005; Ware *et al.* 2002) but more development work is needed to capture the full complexity of macro and micro-synteny across phylogenetic trees.

As noted earlier, comparative genomics standards and algorithms are still very much an active research topic. For this reason it is highly appropriate for research activities to be combined with active curation.

We see the plant community as requiring the following services: 1) one or more automatic annotation shops that provide the computes necessary to generate baseline genome to genome alignments and gene family identifications; 2) curated repositories that will take the resources produced by (1) and provide hand-curated management of synteny blocks, protein families, and conserved functional elements; 3) standardized user interfaces for displaying and manipulating this type of data.

Summary recommendations for comparative genomics projects:

- Encourage the use of standardized pipelines and/or annotation shops for performing genome to genome alignments.

- Encourage the development of standardized machine-readable representations of genome to genome alignments and synteny relationships.

Genetic Mapping and Diversity

Genetic maps of plants are of importance both as a key tool for unraveling the biology of the organism and as a resource for selective breeding and improvement of agronomically important species. Natural and induced genetic variability can be detected using both phenotypic (visible) traits and a wide range of molecular technologies. Detectable genetic polymorphisms include various forms of polyploidy, chromosomal rearrangements, gene rearrangements, insertions, deletions, microsatellite repeats, RFLPs, PCR-AFLPs, SNPs, MNPs and haplotype blocks. While genetic variants are often characterized in terms of the detection technology, the utility of information from genetic variants depends on context: genomic location, population attributes and phenotypic effects.

Because the reproductive biology of most plant species supports inbreeding, it is possible to maximize linkage disequilibrium across the genome for the inbred. Thus it is straightforward to generate large segregating families from a bi-parental cross of two

inbred lines, thereby generating genetic linkage maps for most plant species of interest to plant biologists. These same genetic variants in segregating families are also the basis for identification of large genomic regions that are likely to be in linkage disequilibrium with genes that influence complex and quantitative traits.

The utility of genetic variants from a population genetics perspective, i.e., estimating allelic frequencies, finding regions under selection, constructing haplotypes and associating allelic effects with phenotypes, is determined in the context of the breeding population. In the extreme case of a population consisting of progeny from an inbred line all genetic markers are in complete linkage disequilibrium. This is the basis for associating specific lines or accessions with genomic haplotypes or fingerprints. Actual breeding populations consist of many individual accessions and determining how to sample the breadth of breeding populations and evaluate sub-structure within a species is an active area of research. Thus, estimating allelic frequencies, haplotype blocks and genetic effects of an allele all depend upon the definition of the breeding population.

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
Marker Collections	X	X	X	Static repositories would be sufficient if standards for reporting polymorphic markers existed and were enforced.
Linkage Maps		X	X	Extensive manual curation currently required. Stock centers needed to capture germplasm of breeding populations and/or parental lines.
Quantitative Trait Loci		X	X	Extensive manual curation required. Controlled vocabularies to describe traits highly recommended. Stock centers needed to capture germplasm of breeding populations and/or parental lines.
Diversity Data		X	X	Extensive manual curation required. Stock centers needed to capture strains used in study.

Table 5: Types of Data Repositories Needed for Genetic Mapping & Diversity

The long-term storage of genetic mapping and variation data is the domain of actively curated repositories such as MODs, CODs and stock centers (Table 5). This is largely due to the complexity of the data types that need to be described, such as breeding populations and phenotypes. However, another important impediment to automated storage is the absence of standards for describing genetic maps and their components.

Even the identifiers used for polymorphic molecular markers are not standardized, and researchers routinely change marker names when using markers described by other groups in their own studies. As a result, in order to make a genetic study comparable to other studies, curators must expend great effort in order to understand the idiosyncrasies of a genetic mapping study, to normalize marker names and pedigree information, and to describe the phenotype under study. Even with extensive curation, it is often impossible to associate a genetic mapping study with identifiable germplasm accessions, due to the lack of standards for identifying the breeding population upon which the study was based.

The types of data produced by genetic mapping and population genetics studies are as follows:

1. *Polymorphic markers.* These are naturally or induced polymorphisms that can be assayed by PCR or other detection techniques. Polymorphisms are the basic components of genetic linkage maps, phenotypic association studies and population-based surveys for natural selection. NCBI dbSNP (Wheeler *et al.* 2005) is a long term storage repository for polymorphic markers, but because it relies on voluntary submissions, and is not actively curated, its contents are heavily skewed towards large-scale dbSNP discovery efforts in human and model vertebrates. It has not been heavily utilized by the plant genetics community, and as a result it contains only a handful of Arabidopsis genetic variants and no genetic variants from other plant species. Instead, plant polymorphic marker information can be found in one or more of the MODs and CODs. In Arabidopsis, TAIR has curated a large number of SNPs as well as descriptions of most types of naturally occurring polymorphisms. Similarly, maize, soybean, rice, and wheat polymorphisms can be found in the curated databases MaizeGDB, Soybase (REF), LIS, Gramene and GrainGenes (REF).

2. *Linkage Maps.* These are ordered sets of polymorphic markers whose relative position and distance are determined by examining crossover frequencies during meitotic recombination in breeding populations. All of the active community databases provide access to this type of map through a process of active curation.

3. *Quantitative Trait Loci (QTLs).* These are maps of the association between a complex trait, such as plant height, against the alleles of a set of polymorphisms that have previously been assigned to a linkage map. QTL maps are the basis for scientific breeding programs as well as a key ingredient in positional gene cloning. Because of the difficulties inherent in describing phenotypes systematically, QTLs require heavy active curation and are handled by several of the extant plant MODs and CODs.

4. *Diversity data.* These are data sets gathered from plants "in the wild" and are key to reconstructing the historical processes of natural variation and selection on plant populations. For example, by comparing the frequencies of alleles in modern maize races to the frequencies in the wild ancestor of maize, teosinte, researchers have identified genetic variants that were selected for during domestication and improvement (Wright *et al.* 2005). Population diversity data is currently captured only by a very few extant plant databases, such as the Panzea database of maize diversity (www.panzea.org).

5. *Genetic mapping reagents.* In addition to generating information, genetic mapping and variation studies generate such physical reagents as PCR primers for detecting SNPs, genotyping arrays, hybridization probes for RFLPs and AFLPs, and recombinant inbred breeding populations. In order to be preserved for future use, these reagents need to be maintained and distributed by stock centers.

Limitations of existing resources. To date, genetic mapping and diversity data, as well as the physical reagents associated with them, have been gathered in a haphazard way. The maize and Arabidopsis genomics community databases do provide access to information on genetic mutants and stocks, but other plant genetics research communities have much more dispersed resources. The unpredictable nature of funding for curated plant databases has been to some extent responsible for this state of affairs.

As noted earlier, a critical issue is the lack of a reliable connection between molecular polymorphisms, genetic mapping studies, and germplasm resources. While all plant MODs and CODs provide information on molecular polymorphisms, they often lack links to the germplasm accessions on which the polymorphisms were characterized. Similarly, the germplasm collections at international stock centers typically provide little if any information on the molecular characterization of their stocks.

Another issue has been the lack of a standardized format for representing even simple genetic data types such as genetic linkage maps. The Polymorphism Markup Language (PML) has been proposed as a standard reporting format for this purpose (Sugawara, Mizushima et al. 2005).

In order to improve the capture and maintenance of this important type of data, we recommend 1) that researchers who develop molecular polymorphic markers be required to submit the information on these assays to dbSNP or another long term repository; 2) that the plant research community move quickly to adopt PML and other emerging standards for representing genetic mapping and variation data; and 3) that stock centers and MODs receive the support necessary to coordinate capture and curation of breeding population germplasm information.

Summary recommendations for genetic mapping & diversity projects:

- Genetic markers and maps should be submitted to long-term (static or curated) repositories using publicly-recognizable names. Genetic markers based on molecular sequences should use recognizable sequence IDs. Genetic maps are probably best handled by curated repositories (MODs or CODs).

- Encourage the development and use of standardized machine-readable representations for genetic maps, diversity data, association and QTL studies.

- When feasible, important germplasm (such as seed stock for parental lines used in mapping crosses) should be submitted to stock centers prior to publication.

Pathways

Biological pathways connect the genes, proteins and chemical compounds of an organism into network of knowledge that represents a first step in understanding biology on a systems level. This knowledge can be used as a basis to model a system and to drive hypothesis driven research. Although almost any biological process can be thought of in the form of a pathway, biological pathways are usually considered to represent biochemical pathways or regulatory pathways. In the case of biochemical pathways, the proteins have enzymatic properties and usually operate on low molecular weight substrates and sometimes also bio-polymers derived from them. Regulatory pathways often involve protein-protein interactions, or covalent modifications of protein substrates, such as phosphorylation, methylation, acylation, etc., that change the activities of enzymes in regulatory or signal cascades. Obviously, biochemical and regulatory networks represent an important aspect of cell function, and their elucidation, description, and understanding provides insights into the nature of diseases and nutrition, and provides opportunities for the improvement of agriculture, biotechnology, and human well-being. In addition, the pathway data intersects naturally with large-scale genome analyses, such as genomics, proteomics, and metabolomics. Indeed, the community is turning increasingly towards network analysis tools to understand these heavily-funded data sets.

Pathway data are complex: pathways are networks of different data-types, can span different subcellular compartments which often involve transport reactions, enzymes consist of protein complexes, and reactions can require multiple co-factors, depend on substrate and enzyme concentrations, have complex enzymatic properties, and be affected by feedback and other types of inhibition. An adequate description of pathways is therefore a daunting task. Representing such knowledge is one of the primary functions of biological databases, and the curation of the metabolism of a species is best done at the MOD or COD level. However, because pathways are frequently conserved across wide evolutionary distances, several large projects take advantage of this conservation to create databases of biological pathways across multiple species. In Japan, the KEGG project at

the University of Kyoto (Kaneshisa *et al.* 2000), provides a comprehensive website with overview diagrams of about 200 biochemical pathways, along with a number of analysis tools. In Russia, the EMP Project (www.empproject.com) has created a large curated database of pathways based on the comprising several thousand journal articles. In the US, reactome.org (www.reactome.org/), focuses on human and animal pathways and currently contains 659 pathways. Another large US effort, the Metacyc project (www.metacyc.org/), collects pathway information from the scientific literature. Currently, MetaCyc contains pathways from more than 240 species (including many bacterial species, but with a particular focus also on plants), comprising more than 500 pathways with 8000 metabolites. MetaCyc uses a model that should be particularly appealing to MODs: Species specific databases can be generated quickly using the MetaCyc collection of pathways and Pathologic, a program that pulls the appropriate pathways out of the MetaCyc databases. New pathways can then be added to the species specific database, which can be fed back to MetaCyc, where they are available for future predictions.

The discrepancy between the number of compounds and pathways found in nature and the number found in databases is considerable. This is particulary a concern for plants, for which hundreds of thousands of compounds have been described in the literature, mostly in secondary metabolism, yet the databases contain at most a few thousand. The need for manual curation of these data into databases cannot be overemphasized. An important consideration is that a large fraction of pathway annotation work has focused on prokaryotes and animal systems. However, many of the secondary metabolite pathways in plants do not occur in animals or bacteria. Therefore, curated plant repositories will need targeted funding to annotate the plant-specific pathways. Ideally, all the annotated pathways would flow into a central database that could be used to derive the pathway complement of a new genome to be annotated. The closest current example of such a database is the previously mentioned MetaCyc database.

In addition to these heavily manually curated, dynamic databases focusing on the pathway themselves, static repositories are needed for other data types, such as storing chemical, chromatographic, mass spec, and other information on small molecules (Table 6). This is particularly important to large-scale methodologies such as metabolomics which generate data for hundreds of compounds. ChEBI (www.ebi.ac.uk/chebi) is a good start at this, but currently has fewer than 6000 curated compounds. Other static collections for enzymatic reactions, such as the Enzyme Commission database, BRENDA and ENZYME, are also important resources.

In contrast to some of the other biological data types discussed in this document, standardized file formats for describing pathways exist and are now widely accepted. The two most important ones are the BioPAX format (www.biopax.org/) and the Systems Biology Markup Language format (www.sbml.org). The first is more suitable for

describing regulatory networks, while the second is more suitable for describing biochemical reactions.

The availability of pathway data in an electronically accessible and computational format will greatly enhance the efficiency of biological and medical research and represent a first step towards a hypothesis-driven systems biology approach. Although some day it may be possible to predict pathways automatically from high-throughput data sets, pathway annotation is currently a painstaking process of human judgment and curation, and is a vital part of genome annotation.

Data Type	*Static Repository*	*Curated Repository*	*Comments, Examples*
Small molecules	X	X	No comprehensive database available
Enzyme nomenclature		X	Enzyme Commission, BRENDA, MetaCyc
Reference Pathway Set		X	MetaCyc (automatic pathways based on curated data sets in reference species)
Species-specific pathways		X	AraCyc

Table 6: Types of Data Repositories Needed for Pathway Annotation

Summary recommendations for pathway data:

- Plant pathway databases should be encouraged.
- Whenever feasible, such databases should make use of existing pathway resources, such as MetaCyc.

Ontologies & Controlled Vocabularies

An ontology is a set of vocabulary terms whose meanings and relations with other terms are explicitly stated in such a way as to be comprehensible to humans and computer programs. Ontologies provide a way to unambiguously describe data and, in effect, are vehicles for standardizing data description.

A growing number of shared ontologies are being built and used in biology. Examples include ontologies for describing gene and protein function, cell types, anatomies and developmental stages of organisms, microarray experiments, and metabolic pathways. A list of open source ontologies used in biology can be found on the Open Biological Ontologies website (obo.sourceforge.net/). The Gene Ontology (www.geneontology.org)

is a biological ontology that has garnered extensive community acceptance, and is a set of over 16,000 controlled vocabulary terms for the biological domains of 'molecular function', 'subcellular compartment', and 'biological process'. Like other biological ontologies, GO is organized as a directed acyclic graph, a type of hierarchical tree that allows a term to exist as a specific concept belonging to more than one general term. Other examples of ontologies currently in development are the Sequence Ontology (SO) project, a collection of all the terms needed to describe genome sequence annotation, and the Plant Ontology (PO) project (www.plantontology.org), a set of terms describing structure and growth stages in flowering plants.

Ontologies are used mainly to annotate data such as sequences, gene expression clusters, experiments, and strains. Data sets that have been described in this systematic way can be efficiently compared, merged, and searched. Most importantly, ontology annotations can be used as the basis for interpreting noisy functional genomics experiments, thereby inferring knowledge. For example, when interpreting a gene expression array, one can ask whether any functions and processes, as represented by ontology terms, are statistically significantly over-represented at one measured time point versus another.

There are two linked tasks in the creation and use of biological ontologies (Table 7). The first task is to create the ontology framework. This is typically performed by a small team of domain experts who meet, develop the basic topology of the ontology (the root terms and the major branches), and then flesh out the term list and definitions with increasingly specific concepts. In latter phases of ontology development, community members are invited to contribute their expertise to specific portions of the ontology. This phase of ontology development may take months to years, after which the ontology enters a slower maintenance phase.

The second task is to put the ontology to work by associating its terms with biological data. This is an ongoing task that is usually performed by curators at MODs and CODs. The exact nature of the work depends on the ontology domain. For example, a phenotype ontology could be used to describe morphological traits of plant mutants and/or naturally occurring variants. The experience of the GO and Plant Ontology groups suggests that it is best to begin the association work while the ontology is still in development, so as to stress-test the ontology while it is still plastic.

A mature suite of software tools for using ontologies is available (www.geneontology.org/GO.tools.shtml), and these are sufficient for the basic tasks of creating ontologies, refining them, performing associations, and searching databases of ontologies and their associations. However, additional tools are needed to perform data integrity checks and to explore complex ontologies. For example, term definitions are currently given in natural language form, which is fine for human comprehension but does not easily allow computers and software to be developed that can help check for

ontology integrity and provide more semantically powerful search functions. We also see an opportunity for the creation of an international repository of ontology standards that could oversee the development and maintenance of the ontologies.

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
Ontologies		X		Curation involves development and updating ontologies
Annotations		X		Curation involves annotations of data objects using the ontologies

Table 7: Types of Data Repositories Needed for Ontology Development

Summary recommendations for ontologies:

- Ontology development should be encouraged. Whenever possible, ontologies should leverage existing database schemas and software tools.

Phenotypic (Functional) Data

High-throughput methods for collecting, storing, and analyzing phenotypic data, also known as "functional genomics," ties the structural information of the genome to the biology of the organism. It comprises a broad and expanding number of techniques that generate data collections that require all the types of data repositories that we have discussed previously:

1. Tissue- and stage-specific EST library analysis.

2. Oligo- and cDNA-based microarray expression studies.

3. SAGE (Series Analysis of Gene Expression) and MPSS (Massively Parallel Signature Sequencing) data.

4. Reporter-gene tissue-specific expression data. A typical example is a gene's promoter coupled to a GFP reporter.

5. High-throughput deletion analysis, including targeted and non-targeted gene knockouts and genome-wide RNAi.

6. Traditional mutation, recombinant inbred and QTL analysis based on morphological and/or quantitative traits.

7. Chromatin immunoprecipitation data, protein interactions information, and even crystallographic structures, which can be used to speculate about the function of a given gene product.

As described in the previous section, ontologies are vital for the interpretation of phenotypic data. For example, RNA-based techniques such as microarrays, SAGE and MPSS, require standard ontologies that describe the tissue, organ, and growth stage from which the RNA was extracted, as well as ontologies that describe the precise environmental and growth conditions of the source organism. For interventions that result in a measurable phenotype, ontologies are needed to describe the portion of the organism affected and the nature of the change. Without shared ontologies, it is impossible to compare the results of functional genomics experiments across multiple experiments or species.

Given the availability of suitable ontologies, much of the curation of EST, SAGE, and MPSS datasets can be generalized to the following tasks:

1. For EST-based data sets, cluster the ESTs. This is a computational task that, though not perfect, is well understood. It is a task that is independent of a given species and which can be performed by a number of extant groups including TIGR and PlantGDB.

2. Integrate the sequence information that underlies the data set (EST sequence, SAGE, or MPSS tag) with the genomic data, when available. This involves identifying the genomic location of the EST read, EST cluster, or sequence tag. This usually a highly-automatable task and can be done by groups that do not have any special species- or clade-specific expertise.

3. Associate the RNA source with the appropriate set of ontology terms. This is a task that requires detailed understanding of the developmental biology of the organism and is best suited to database groups that focus on species- or clade-specific biology. A logical alternative is to have the data providers document the association between an RNA source used in an experiment and a set of ontology terms, but there is so far no precedent for this type of activity.

Microarray data sets require a static repository for the raw microarray results as well as a curated repository for associating the target RNAs with ontology terms that describe the tissue, stage, and environment of the plant from which the RNA was derived. Whereas the static repository can be managed by a species-independent center, such as the NCBI GEO database, the association and annotation of the data set needs to be performed by a group that has extensive knowledge of the specific organism's biology.

Reporter gene data sets require a stock center to maintain and distribute the derived lines,

and a curated repository to associate the stage- and tissue-specific expression patterns with ontology terms and to establish the connection between the reporter gene construct and the genome annotation. There may also be images associated with the data set which must be annotated. These tasks require a group with extensive knowledge of the specific organism's biology.

The requirements for the knockout and RNAi-based knockdown resources are similar to those for reporter gene sets. A stock center is needed to manage the knockout strain or the small hairpin library, and a curated repository is needed to associate the resulting phenotypes with the appropriate ontology terms as well as to establish connections to the genome annotation. Like reporter gene sets, this activity requires a group that has extensive knowledge of the specific organism's normal and abnormal biology.

Finally, the management of the traditional types of phenotypic analysis which studies spontaneously-arising variants, mutants derived from a mutagenesis screen, or agronomically important quantitative traits that differ among two strains usually requires the involvement of a stock center to curate the germplasm stocks that arise from the study and a curated repository to manage the information on the experimental design and the results. Good shared ontologies are key to managing this type of data so as to facilitate comparisons among multiple experimental studies. Manual annotation by biologists who have a detailed understanding of the organism's biology is required for anything but the most superficial curation of this type of data.

Data Type	*Static Repository*	*Curated Repository*	*Stock Center*	*Comments*
ESTs	X	X	X	EST clustering can be performed in a species-independent way
Microarray Expression Studies	X	X		Use existing repositories for static data
SAGE, MPSS		X		Genome mapping does not require species-specific knowledge, but ontology association does
Knockouts, knockdowns		X	X	Species-specific knowledge required
Reporter Constructs		X	X	Species-specific knowledge required
Mutant & QTL Analysis		X	X	Species-specific knowledge required

Table 8: Types of Data Repositories Needed for Phenotypic Data:

Summary recommendations for phenotypic data:

- Data sets that require species-independent computation or services, such as EST clustering and microarray storage and analysis, should leverage existing resources whenever feasible.

- Phenotypic data repositories should be encouraged to develop shared ontologies to describe assay and phenotype data.

Reagents and Stock Centers

This section deals with the specific need for stock centers to manage and distribute the physical reagents that are created by genome-scale projects. The central task of a stock center is to (1) enable individual researchers who are not directly connected to the projects to locate the reagents generated by large-scale projects; and (2) to acquire those physical entities for use in their own experimental analyses. Although it is simple to state the need, further examination of the topic reveals several thorny issues.

Stock centers must deal with the logistics of receiving reagents, storing them, and distributing them in a timely and cost-effective manner. Given that reagents are often living organisms (seed stock or even growing plants that must be propagated vegetatively) the logistical issues are substantial. Stock centers have the additional challenge of maintaining the integrity of their stock. There needs to be a verifiable link between the reagent that was used in a published experiment and the reagent that the stock center ships out upon request. No system being perfect, there is always the chance of sample mixup or contamination (either within the stock center or before it even receives the reagent), and it is desirable that stock centers have mechanisms in place to identify each sample unambiguously, for example by using molecular polymorphism fingerprints.

Finally, stock centers must establish reciprocal connections with static and curated data repositories so that the experimental data described in the repository has an unambiguous connection to a physical reagent in the stock center. In practice, this means that stock centers must implement a system of stable public IDs that can be shared with the data repositories and updated at regular intervals. A good example of a working relationship between data generators, stock centers and data repositories are the SALK SIGnAL service, the ABRC (www.biosci.ohio-state.edu/ ~plantbio/Facilities/abrc/abrchome.htm), and TAIR, which together give researchers access to a valuable collection of Arabidopsis insertional mutant stocks. Some stock centers have been effective at providing integrated search and query facilities (for example the Nottingham Arabidopsis Stock Center, NASC), but many have not had the resources to develop more than a very simple online catalog of their stocks.

We now consider existing resources for several common types of plant biology reagents.

Seed Stocks. Many MODs and project databases offer resources for locating and ordering seed stocks for plants that are genetically modified and/or for natural germplasm accessions. An example can be found at TAIR (www.arabidopsis.org/), which collaborates with ABRC at Ohio State (www.biosci.ohio-state.edu/~plantbio/Facilities/abrc/abrchome.htm) to integrate genome annotation data with biological reagents. MODs often include varying levels of pedigree data, depending on the database and the thoroughness and availability of such information. Other resources providing data on seed stocks include MaizeGDB (www.maizegdb.org/), NASC (arabidopsis.info/), the CerealsDB SNP repository (www.cerealsdb.uk.net/discover.htm), Panzea (a project database for maize diversity data; www.panzea.org/), Gramene (www.gramene.org), IRRI (the International Rice Research Institute; www.irri.org), and GRIN (the Germplasm Resource Information Network; www.ars-grin.gov/npgs/).

GRIN is of particular interest because it houses information describing natural plant genetic resources for over 11,000 species of plants (nearly a half-million accessions) and allows available stocks to be ordered online. However, unlike the other repositories listed here, GRIN does not currently store molecular information with stock data, or provide connections to data repositories that do store such information. This makes it difficult for researchers to identify and acquire useful germplasm via sequence information, and makes it impossible to verify the correct identity of a seed stock.

Transgenes. A special case of a seed stock is a transgene, a piece of DNA (generally coding DNA) that has been introduced into cells or organisms to modify the genome. Transgenic plants are created using various methods including promoter-enhancer traps, T-DNA insertional mutagenesis, and EMS mutagenesis. In the case of a transgene, there is always some form of molecular characterization of the line, typically performed by the lab that generated it, and there is often some characterization of the phenotypic consequences of the transgenesis. The key to maintaining the usefulness of the transgene is for the data repository to index the transgene by its molecular signature (e.g. the insertion site of the engineered DNA), its phenotypic effect (typically using a searchable ontology), its stock center ID, and, when appropriate, a reference to the paper in which the transgene was published. The stock center, for its part, should be able to verify that the seed stock it receives carries the correct molecular change and should provide researchers to whom the stock is distributed the information needed to verify the identity of the stock.

Vectors & Cloned Sequences. Stock centers can be called upon to store vectors, clones, and whole clone libraries. For example, the Maize Gene Discovery Project (MGDP; PI Virginia Walbot) deposited its clones at stock centers located at Texas A&M and the Arizona Genomics Institute and transmitted the clones' molecular data to the maize data repository at MaizeGDB. Even though the MGDP project is now finished, researchers can

still identify clones of interest to them and obtain the reagents. Without this foresightedness, the funding agencies' investment in the project might have been lost.

Locating Resources. A recurrent complaint from plant researchers is the difficulty of locating available data and reagents. In part this is because data providers have often established *ad hoc* solutions for archiving reagents and information about them, leading inevitably to a proliferation of distribution sites and online databases. One way to reduce the confusion would be to strongly encourage groups that are developing reagent resources to establish relationships with existing stock centers and data repositories. Some funding mechanism – perhaps subcontracts from the resource generator to the stock center and repository – would need to be found to allow this type of arrangement to scale.

Another way to make the existence of resources more transparent is the establishment of a plant molecular reagent data "portal" in which all resources are organized by species and resource type. This was the model preferred by an NSF discussion in 2000 (panel members included Howard Rines, Jennifer Normanly, David Frisch, Hongbin Zhang, Robin Buell, Jan Dvorak, and Virginia Walbot), in which it was concluded that stock centers and individual labs alone should suffice for making reagents available to researchers. We feel that the best location for such a portal is a MOD or COD, where data integration and organization happens routinely.

Summary recommendations for stock centers:

- Stock centers should be encouraged, and provided with sufficient resources, to collect, utilize and publish molecular characterization data on germplasm and other reagents.

Data Integration

Currently, many information resources are encyclopedic. They excel at collecting, curating, indexing, and presenting a broad array of data types, both within and across species. With only a few mouse clicks, scientists can see visual layouts of gene sequences annotated with functional information, 3D protein structures, and a suite of alternative displays and analysis tools. As electronic encyclopedic reference portals, these data collections have helped set standards in electronic data organization and presentation.

Yet the very technologies that allow these collections to excel, such as their heavy reliance on keyword searches, pull-down menus, and the traditional web interface of HTML over HTTP, present substantial obstacles to empowering them as high-throughput research resources. Biology is increasingly becoming a high throughput, information science, and as such, this places demands on the necessity for machines to translate our simple requests into complex queries, execute those queries over distributed resources, filter and collate the returned information, and present the results in an organized manner. This

demand is not well served by either the traditional point-and-click web browser interface or an *ad hoc* FTP download of bulk data. To see the severity of this problem, bring up a web browser and try to answer any of the following questions:

1. *Which drought tolerance genes in maize have homologs in Arabidopsis that are significantly up- or down-regulated when experimental plants are exposed to desiccating conditions?*

2. *TAIR returns 21 loci associated with the Gene Ontology term "meiosis." Arabidopsis is likely to have hundreds of genes involved in meiosis. Which ones share motifs suitable for determining ancient gene duplication events that could elucidate the process' evolution?*

3. *What information does PDB have on these genes that would support or refute common ancestry?*

The difference between asking these questions today versus ten years ago is that today much—if not all—of the information needed to make a reasonable advance is already available over the web; it is just not available in a readily-accessible, high throughput manner. In fact, the amount of point-and-click, cut-and-paste effort needed to answer them is so high that it can take a full-time postdoctoral fellow weeks to confidently discover and execute the manual workflow. The challenge for today's data collections is to allow scientists to access and extract the information the resources already have in a high throughput, efficient manner. This requirement is placed upon them because biology is increasingly becoming a high-throughput, information science.

Information must be integrated in order to answer the above questions, and a prerequisite for integration is interoperability. That is, we cannot expect machines to integrate before they can interoperate. Currently, there are neither broadly accepted nor implemented interoperability standards. Both interoperability and integration are hampered by the fact that HTML encoding tends to confound the raw data content with its structure and presentation. Disentangling the data from how it is organized and presented is an important benefit that is likely to arise from well-constructed interoperability standards, and one that will be key to achieving integration.

For us to move data collections from low-throughput, electronic encyclopedias, to high throughput, research resources, we will need to develop interoperability standards in a manner that allows machines to assess suitability-for-purpose on a request-by-request basis. This will require semantically tagging information and making it available for logical discrimination, either via document-based models, such as RDF (Resource Description Framework), OWL (Ontology Web Language), or SWRL (Semantic Web Rule Language); or via procedural access from traditional computer languages. We note approvingly that NSF has recently funded a research project to utilize these technologies in the creation of a Virtual Plant Information Network (BioMOBY 2005).

CONCLUSIONS AND SPECIFIC RECOMMENDATIONS

Our plant biology database needs assessment has come back time and again to a single overriding conclusion: the research community's need for a system of curated data repositories where information is actively acquired, organized, maintained and distributed. This in turn requires a trained cadre of skilled knowledge workers who are able to curate complex biological data, as well as a system of stable funding that enables such repositories to be established and maintained for extended periods of time. We will discuss global recommendations first and then summarize recommendations reached earlier that are specific for particular types of biological data.

1) Develop a funding mechanism that would give curated repositories a longer cycle time than currently feasible.

Most curated databases are now funded as research projects under a process of competitive grant review for cycles of 3-5 years. This is insufficient to establish a stable resource and to create an environment that will be attractive to those biologists who wish to make a professional career of data curation. We recommend that funding agencies develop a mechanism to fund static and curated repositories for renewable periods of 7-10 years. During this time the repositories would be subject to annual review by an advisory board, and would be held to a defined set of milestones and objective measurements of performance. This would allow successful repositories to provide the community with long-term stable maintenance of data, while allowing funding agencies to weed out unsuccessful repositories.

2) Foster curation as a career path.

The funding agencies as well as educational institutions should put renewed emphasis on data curation as a respected career path. This will involve addressing issues of curriculum development, mentoring, specialty conferences, and the development of peer reviewed journals that specialize in curation research and methodology. One promising recent development is an embryonic movement to establish a Society of Biocurators (see biocurator.org), which we feel should be encouraged. A possible mechanism for supporting students who wish to explore curation as a career would be to establish a career development award for individuals seeking to enter the discipline.

3) Balance data generation and information management.

Because the storage of the data and/or reagents generated by high-throughput studies is so vital to the community, we feel that funding agencies should insist that potential data providers include in their proposals a plan for the long term storage and maintenance of the data set and any reagents, if any, associated with it. A minimum set of standards for the publication of data sets includes using publicly recognizable identifiers for biological

data objects, using accepted nomenclature to describe the data set, using standard formats for data files, and linking the IDs of reagents submitted to stock centers to the IDs given in data files. Whenever possible, data providers should make arrangements with existing repositories and stock centers rather than planning to implement an entirely new information resource. If managing a data set will strain the existing resources of data repository and/or stock center, then the data provider should establish the appropriate subcontractual arrangements to close the gap.

4) Separate the technical infrastructure from the human infrastructure.

As noted earlier, there are many automated computational tasks that do not require specialized species- or clade-specific knowledge. These tasks include such things as gene prediction, EST assembly, genome alignment and protein family identification. In order to avoid redundant and inconsistent efforts, funding agencies should encourage partnerships between groups that can provide technical infrastructure for automated annotation tasks and groups that are skilled at manual curation. In the animal world, a successful example of this type of partnership is the relationship between Ensembl and MGD (www.informatics.jax.org); the former provides an automated gene prediction set on the mouse genome, while the latter integrates this information with allelic information, phenotypic data, genetic maps, and other heavily curated biological resources.

5) Standardize data formats and user interfaces

The lack of standard file formats for genetic maps and several other key biological data types provides friction that increases the cost and decreases the pace of active curation. The lack of standardization of data repository user interfaces leads to frustration on the part of researchers who cannot easily move from one repository to another.

Data providers should be encouraged to use standard file formats whenever available. Data repositories should provide standard user interfaces in addition to any custom ones they wish to develop. When suitable standards do not exist, there should be a push to develop them. We feel that it would be appropriate to establish a working group to develop a "Best Practices" document to describe recommended data formats and user interfaces for common biological data types. This could then be used as one guideline for evaluating data generation and management proposals.

6) Encourage CODs

Existing MODs should increasingly exchange data with and create reciprocal linkages to CODs currently in operation. In order to avoid an unsustainable proliferation of species-specific databases, and to encourage the emerging discipline of comparative genomics, we also recommend that existing MODs should be encouraged to take on new species and gradually evolve into CODs.

7) Encourage the development and deployment of new technologies explicitly aimed at integrating across MODs and CODs.

Just as MODs and CODs deliver value greater than the sum of their particulate data sets, the integration of data and services across MODs and CODs has enormous potential for achieving high value in comparative bioinformatics. Realizing this value will require work explicitly aimed at solving the distributed data integration equation. We recommend recognizing this is a distinct area of research and effort necessary to achieving a national network of integrated plant biology databases.

8) Specific recommendations for genome sequencing and mapping projects

- Molecular markers (genetic and physical) should be submitted to NCBI GenBank.
- Clone libraries should be submitted to stock centers.
- A standard file format should be developed for representing physical maps. Physical maps should be curated at the MOD or COD level.
- Genome assemblies must be curated and maintained after the original sequencing centers have moved on. Sequencing projects must develop a plan for the orderly handing over of the assembly to a repository that can manage updates of the assembly in response to community feedback and/or additional experimental data.

9) Specific recommendations for genome annotation

- Sequencing projects must develop a plan for developing a public, canonical set of gene predictions over a set period of time using generally accepted best practices for gene prediction. The plan should include a mechanism for accepting and responding to community feedback on incorrect or missing gene models.
- Use of standardized genome annotation pipelines should be encouraged. This will simplify the task of cross-species comparison, and reduce redundant effort.
- Encourage partnerships between manual curation groups and genome annotation shops.

10) Specific recommendations for comparative genomics

- Encourage the use of standardized pipelines and/or annotation shops for performing genome to genome alignments.
- Encourage the development of standardized machine-readable representations of genome to genome alignments and synteny relationships.

11) Specific recommendations for genetic mapping

- Genetic markers and maps should be submitted to long-term (static or curated) repositories using publicly-recognizable names. Genetic markers based on molecular sequences should use recognizable sequence IDs. Genetic maps are probably best handled by curated repositories (MODs or CODs).

- Encourage the development and use of standardized machine-readable representations for genetic maps, diversity data, association and QTL studies.

- When feasible, important germplasm (such as seed stock for parental lines used in mapping crosses) should be submitted to stock centers prior to publication.

12) Specific recommendations for pathway data

- Plant pathway databases should be encouraged.

- Whenever feasible, such databases should make use of existing pathway resources, such as MetaCyc.

13) Specific recommendations for ontologies

- Ontology development should be encouraged. Whenever possible, ontologies should leverage existing database schemas and software tools.

14) Specific recommendations for phenotypic data

- Data sets that require species-independent computation or services, such as EST clustering and microarray storage and analysis, should leverage existing resources whenever feasible.

- Phenotypic data repositories should be encouraged to develop shared ontologies to describe assay and phenotype data.

15) Specific recommendations for stock centers

- Stock centers should be encouraged, and provided with sufficient resources, to collect, utilize and publish molecular characterization data on germplasm and other reagents.

LITERATURE CITED

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R. (2005). NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res*. 33:D562-566.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL (2004). *Nucleic Acids Res.* 32:D23-D26.

BioMOBY (2005). A Virtual Plant Information Network (VPIN). http://biomoby.org/VPIN.doc

Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, Curwen V, Cutts T, Down T, Durbin R, Eyras E, Fernandez-Suarez XM, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Kahari A, Jekosch K, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark C, Clamp M, Hubbard T. (2004). Ensembl 2004. *Nucleic Acids Res*. 32:D468-470.

Dong Q, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V. (2005). Comparative Plant Genomics Resources at PlantGDB. *Plant Physiol.* (2005) 139:610-618.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* 21:3439-3440.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* 6:R44.

Gonzales MD, Archuleta E, Farmer A, Gajendran K, Grant D, Shoemaker R, Beavis WD, Waugh ME. (2005). The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.* 33:D660-D665.

Hoon S, Ratnapu KK, Chia JM, Kumarasamy B, Juguang X, Clamp M, Stabenau A, Potter S, Clarke L, Stupka E. (2003). Biopipe: a flexible framework for protocol-based bioinformatics analysis. *Genome Res.*13:1904-1915.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28:27-30.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31:51-54.

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:D71-74.

NSB (2005) Long-lived digital data collections: Enabling research and education in the 21st century. Draft, March 30 (2005). Available on-line at http://www.nsf.gov/nsb/meetings/2005/LLDDC_Comments.pdf.

Pan X, Stein L, Brendel V. (2005). SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics*. 21:3461-3468.

Pennisi, E. (2005) National Science Foundation: Boom in Digital Collections Makes a Muddle of Management. *Science*, 308(5719): 187-189.

Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. (2004). The Ensembl analysis pipeline. *Genome Res.* 14:934-941.

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, Miller N, Mueller LA, Mundodi S, Reiser L, Tacklind J, Weems DC, Wu Y, Xu I, Yoo D, Yoon J, Zhang P. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community.*Nucleic Acids Res.* 31:224-8.

Stein, L. (2001). Genome Annotation: From Sequence to Biology. *Nature Reviews Genetics* 2:493-503.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S. (2002). The generic genome browser: a building block for a model organism system database. *Genome Res.* 12:1599-1610.

Sugawara, H., H. Mizushima, et al. "Polymorphism Markup Language." from http://www.jsbi.org/journal/GIW04/GIW04P170.pdf.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* 408:796-815. (Also see the multiple chromosome-sequencing papers in the same *Nature* edition).

Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S, McCouch S, Stein L. (2002). Gramene: a resource for comparative grass

genomics. *Nucleic Acids Res*. 30:103-105.

Westbrook J, Feng Z, Chen L, Yang H, Berman HM. (2003). The Protein Data Bank and structural genomics. *Nucleic Acids Res*. 31:489-491.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, DiCuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L, Yaschenko E. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33:D39-D45.

Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS. (2005). The effects of artificial selection on the maize genome. *Science.* 308:1310-1314.