

Curation of the EcoCyc Database: The EcoCyc Update Project

Martha Arnaud for the EcoCyc Database, SRI International, 333 Ravenswood Avenue, Menlo Park, CA 94025, martha.arnaud@sri.com

The EcoCyc database is a model-organism database for *Escherichia coli*. EcoCyc is best known for its coverage of metabolic pathways and enzymes; however, EcoCyc is now evolving into a curated encyclopedia of *E. coli* molecular biology. Curated descriptions of transporters, transcription factors, operons, and regulatory elements have been added to EcoCyc. Recently, we have initiated a project to systematically curate every gene product in *E. coli*. Over the next few years, we will add paragraph-style text comments to describe every characterized gene product as well as comprehensive reference lists. This talk will focus on our goals, plan, priorities, and our progress to date.

EcoCyc is freely available to researchers at academic and non-profit institutes.

An Overview of Eukaryotic Annotation at TIGR

Roger Smith

The Institute for Genomic Research (TIGR) is involved in the annotation of a number of eukaryotic genomes and a systematic approach is utilized for the primary annotation of these genomes. Newly sequenced genomes are assembled into scaffolds and then directed sequencing fills in gaps in these scaffolds, a process referred to as closure. Typically, once a genome is near completion or is fully closed, the sequence data is passed through a centralized data management system known as Eukaryotic Genome Control (EGC). This fully automated yet customizable system first identifies biological features such as genes and then data is gathered from a number of sources utilizing various methods for each gene to facilitate the assignment of function and structure. Once this data is assembled in an automated fashion, it is manually reviewed and annotated by a team of scientists to produce a high quality, thorough, complete and consistent annotation of the proteome. A custom software interface, Annotation Station is used by annotators to manually inspect the evidence aligned to the genes and edit them. A web-based interface developed at TIGR, MANATEE (available at <http://manatee.sourceforge.net/>) allows annotators to easily access the computationally derived data and add functional information to gene products such as names, aliases, symbols, E.C. numbers, GO assignments, and comments. This comprehensive tool provides annotators with the best possible information to curate gene products based on functionally characterized protein matches. The overall strategy, tools, and software used at TIGR in eukaryotic annotation will be discussed in more detail with an emphasis on the manual annotation methodologies employed.

Organization and presentation of biological information in the *Saccharomyces* Genome Database

Maria C. Costanzo^a, Rama Balakrishnan^a, Karen R. Christie^a, Kara Dolinski^b, Selina S. Dwight^a, Stacia R. Engel^a, Becket Feierbach^b, Dianna G. Fisk^a, Jodi Hirschman^a, Eurie L. Hong^a, Laurie Issel-Tarver^a, Robert S. Nash^a, Anand Sethuraman^a, Barry Starr^a,

Chandra L. Theesfeld^a, Rey Andrada^a, Gail Binkley^a, Qing Dong^a, Christopher Lane^a, Mark Schroeder^b, Shuai Weng^a, David Botstein^b, J. Michael Cherry^a

^aDepartment of Genetics, School of Medicine, Stanford University, Stanford, CA 94305-5120, USA

^bLewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Washington Road, Princeton, NJ 08544, USA

The *Saccharomyces* Genome Database (SGD) collects and organizes information about genes and gene products of the model organism *S. cerevisiae*, bakers' or budding yeast. As the repository of the official *S. cerevisiae* genome sequence, SGD provides tools to access and analyze the genomic sequence and to view the overall organization of the genome. SGD's primary focus, however, is on the biology of the yeast cell and its molecular components. Reflecting this, SGD is organized around a basic unit, the Locus page, that collects information for a single gene or chromosomal feature. The layout of the Locus page is intended to be as user-friendly as possible, with simple presentation, clear organization, intuitive use of links, logical paths of navigation, and thorough help documentation. The Locus page provides links to literature, sequence, Gene Ontology (GO) annotations, expression data, functional analysis studies, and other types of information specific to the gene of the page. For instance, the "Functional Analysis" pull-down menu on a Locus page takes the user directly to expression data for that locus, in the selected dataset. Because the Locus pages are the organizational center for the database, they can be accessed and searched from a variety of other pages and tools in SGD. Recent additions to the Locus page, and projects currently in progress for its improvement, will be discussed.

GeneDB: A Prokaryotic and Eukaryotic Genome Resource

Christiane Hertz-Fowler & The Pathogen Sequencing Unit
The Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SA UK

The Pathogen Sequencing Unit (PSU) at the Wellcome Trust Sanger Institute is involved in the sequencing and annotation of a diverse range of prokaryotic and eukaryotic organisms, in some instances these projects are part of collaborations between sequencing centres. GeneDB (<http://www.genedb.org>) database was developed to house genome datasets from these projects. Currently, datasets from 16 species, including from 5 bacterial, 3 fungal, 4 Apicomplexan and 3 Kinetoplastid species, are represented within GeneDB. The major emphasis in development so far has been to make the sequence and annotation of finished as well as ongoing genomes projects available via a user-friendly resource.

All data can be easily accessed using browsable catalogues, text and/or sequence searches as well as via a query interface which allows a wide range of annotation to be interrogated. Genes or feature predictions are displayed on their own pages, containing location information, neighbourhood maps, results of predictive software packages and additional curated annotation in a graphical and text based format. Queries can be extended to include datasets from multiple genomes and can be viewed, refined and downloaded in a variety of file formats. Extensive cross-referencing allows retrieval of

related information not only across species within GeneDB but also from external resources.

Datasets from six organisms are curated by biologists, aiming to integrate sequence data with the vast array of available functional, expression and phenotypic data, accessible through public databases, the literature and fed back via the research communities. With an increasing emphasis on comparative sequencing projects, curators are also involved in maintaining datasets of related organisms.

The presentation will briefly introduce GeneDB before focusing on the challenges faced by the GeneDB curators.

RegulonDB: Curation, Literature Search, Notation and Evidences about Transcriptional Regulation and Transcription Unit Organization in E.coli K-12

Gama-Castro S., Peralta-Gil M., Martínez-Antonio A., Santos-Zavaleta A., Salgado H., Jimenez V. and Collado-Vides J., Program of Computational Genomics, CIFN, UNAM, A,P, 565-A, Cuernavaca, Morelos 62100, México.

RegulonDB is a database with experimental knowledge about the elements of transcriptional regulation in Escherichia coli K-12. (Salgado, et al. Nucleic Acids Res. 2001 29:72-4). It contains information on transcription units, promoters, terminators, regulatory proteins, binding sites for regulatory proteins and conditions and the associated affected genes. All of these objects are supported by references and evidences, supporting the validity of the data. We have defined a specific notation to describe several objects in a unique and unambiguous way in the database. These include promoters, transcription units and regulatory proteins.

The curation process includes all articles that contain information about transcriptional regulation. The first step of this search is to gather abstracts from PubMed database using a set of pertinent keywords. Then the abstracts of these papers are read and selected to obtain the complete articles in order to read them. Finally, the data extracted is added through several capture forms into RegulonDB, there is a capture form for each object. The quality control of the data added is monitored automatically through reports of inconsistency in the data. We will describe a quick overview of the database, emphasizing the complete curation process.

Integration of New Data into RGD: Quality Control and Data Submission Tools

Dean Pasko, Susan Bromberg, Wenhua Wu, Chunyu Fan, Chin-Fu Chen, Gopal Gopinathrao, Rajni Nigam, Cindy Foote, Dorothy Reilly, Angela Zuniga-Meyer, Jiali Chen, Norberto de la Cruz, Mary Shimoyama, Simon Twigger, Aubrey Hughes, Jed Mathis, Nataliya Nenasheva, Victoria Petri, Weiye Wang, Lan Zhao, Peter Tonellato, Howard Jacob

Human and Molecular Genetics Center, Medical College of Wisconsin, 8701 Watertown Plank Road, Milwaukee, Wisconsin 53226, USA

One goal of RGD is to integrate data from multiple sources including the literature, research laboratories, as well as other comprehensive databases such as Swiss-Prot and LocusLink. This requires that we accurately match incoming data with our data.

Because there are many different symbols for any given gene, quality control measures have been developed to ensure that data records are associated correctly before they are loaded. This was done using a combination of gene symbol and aliases, GenBank accession numbers, and sequence. Data that does not meet the established criteria are separated according to the conflict type and resolved by manual curation. Data curated from the literature also goes through this pipeline, but in addition quality control measures have been built into the data submission forms curators use. The methods used to align the data and resolve conflicts as well as data entry forms will be described.

Comparative Map Curation in Gramene Using CMap

Immanuel Yap, Ken Y. Clark, and the Gramene Team

CMap is a visualization tool for comparative mapping. It is being developed as a module for the Generic Model Organism Database Project (GMOD) and is available for download at <<http://www.gmod.org/cmap/>> CMap defines a map_set as a collection of maps, and a map as a linear array of features. Correspondences may be generated between different features on different maps. Since CMap strictly acts as a display tool, the curator is free to redefine the meaning of maps, features, and correspondences as well as how they are drawn. For instance, the curator may define a map as being genetic, physical, sequence-based, etc. A feature on a map may represent a locus, gene, molecular marker, EST, phenotype, QTL, etc. Features may be portrayed as a point or an interval of different color. Correspondences may be generated automatically, based on feature names, or specified by the curator. Data may be loaded by batch from a perl command-line script. CGI scripts allow manual addition, modification, and editing of data via a web browser. Gramene <<http://www.gramene.org/>> is using CMap to display and compare genetic, physical, and sequence-based maps of rice and other grasses. Issues faced by Gramene curators when using CMap to store and display maps and generate correspondences will be discussed.

Map Curation on GrainGenes

Victoria Carollo^a, Gerard Lazo^a, David Matthews^b, Olin Anderson^a

^a USDA-ARS-WRRC, 800 Buchanan Street, Albany, CA 94710

^b Cornell University, Dept. of Plant Breeding, Ithaca, NY 14853

The GrainGenes project is a USDA-supported compilation of molecular and phenotypic information on wheat, barley, rye and oats. The curation and delivery of web-based genetic and physical maps on GrainGenes has been a mainstay of the database since its inception in 1991. GrainGenes currently serves 90 Map_Data sets of entire genomes, 22 Linkage_Data sets of single chromosome studies, and 581 records of 2_point_data studies. Mapped loci are linked to text records containing associated data such as marker type, mapping scores, images of autoradiograms, band size, linked QTL, references, and mapping probes. Probe records associated with loci contain links to external databases, clone and source information, PCR conditions, primer sequences, etc. QTL maps on GrainGenes link the loci to supporting statistics and descriptions of trait studies. Maps selected to add to the GrainGenes map collection are usually

identified via the literature reference stream into the database, but very often curators work closely with colleagues in the Triticeae research community to publish maps on GrainGenes simultaneously with publication in scientific journals and newsletters. GrainGenes is planning a major move from an object-oriented ACEDB database to a relational database. New map viewing graphical interfaces will also be implemented. The GBrowse viewer, developed by the GMOD group, will provide a user-friendly interface for physical maps, and allow users to annotate map data and curators to include more features than currently available on the ACEDB map viewer. The CMap viewer, developed by Gramene, will facilitate map integration and comparative genomics. An overview of current map accessions, planned additions and new features will be discussed.

Sequence Curation in dictyBase

P. Fey, E. M. Just, P. Gaudet, P. A. Dyck, S. Merchant, W. A. Kibbe, R. L. Chisholm

Northwestern University, Feinberg School of Medicine, Center for Genetic Medicine,
303 E Chicago Ave, Chicago, IL 60611

dictyBase (<http://dictybase.org>) provides the scientific community with a database that aims to integrate all currently available *Dictyostelium* genome sequence, literature and as much as possible, biological knowledge associated with specific genes. dictyBase is based on the Saccharomyces Genome Database (SGD) which has worked closely with us in developing dictyBase. In dictyBase we wish to present information collected from a variety of different sources, including sequencing data from the genome sequencing centers, GenBank records submitted by research laboratories, data from cDNA sequencing projects, and other gene predictions (not from sequencing centers). Our goal is to present all of this diverse, and occasionally conflicting data to allow individual users to evaluate and make their own judgments about its validity. To accomplish this we display all the different sequence types on separate tracks in our Genome browser (Gbrowse; <http://www.gmod.org/ggb/index.shtml>): 'Verified Genes', 'Gene Predictions from Sequencing Centers', 'EST Alignments', 'HMM Gene Predictions', 'Contigs', 'Unanchored Genes'. In addition, this approach also may be valuable for other databases with incomplete genome sequences.

Since the most consistent source of sequence data comes from the genome sequencing centers, our current strategy is to use genome sequence as the primary sequence. However, to date, *Dictyostelium* genome center gene predictions (geneID) are derived in an entirely automated fashion, as are a separate set of gene predictions derived by an independent effort and distinct software package--the HMM gene predictions. Therefore we created the 'Verified Genes' track, whose coordinates come from manual entries by dictyBase curators based on all available information (GenBank records and ESTs). The coding sequences are blasted against the Chromosomal DNA from the sequencing centers to compare gene coordinates. In case of a discrepancy between the experimentally validated gene model and the sequencing center gene model, the curators assign new chromosomal coordinates according to the experimentally derived sequence. In this way dictyBase can contribute to the improvement of gene models and present users with manually annotated gene models when there is additional supporting data. We hope to stimulate discussion in the group of how other databases are handling these issues.

Apollo: a genome annotation tool

Lynn Crosby, FlyBase, Department of Molecular and Cellular Biology,
Harvard University, Cambridge, MA 02138-2020, USA

Apollo, developed jointly by members of the Berkeley Drosophila Genome Project and the Sanger Institute (UK), is a powerful and versatile tool for the annotation of genomic sequence. It is a Java application, and can be downloaded and run on Windows, Mac OS X, or any Unix-type system (including Linux). Apollo is freely available; see <http://www.fruitfly.org/annot/apollo/> and <http://www.fruitfly.org/annot/apollo/install.html>. User documentation is remarkably detailed and complete. The BDGP version, which has been used for Drosophila genome annotation, is slightly different than the Sanger version, used for viewing the human genome. Documentation and software for developers are also freely available, from SourceForge (which can be accessed from the page above).

Apollo is both a viewer and an interactive annotating tool. The view is divided into an annotation zone and an evidence zone; one or both strands may be viewed. The tool allows the user to view very large amounts of aligned data quickly and effectively. Innumerable aspects of the view can be customized, and moving around within the region being viewed is fast and intuitive. One may zoom in and out; at the highest resolution DNA and protein sequence data appear, superimposed and aligned on the genomic sequence base line, on the annotated objects, and on the evidence objects. Annotation may be done at varying levels of detail, using alternative views and interactive manipulations between the evidence and the annotations.

Apollo is a cooperative work in progress. Currently under development is the ability to add evidence in a dynamic fashion (rather than being pre-loaded), and a version of Apollo that will simultaneously present two aligned genomes. In addition, a number of user groups have developed applications and extensions of Apollo. Developers at Berkeley and Sanger provide extensive technical support for such efforts; these communications may also be accessed via SourceForge.

Clustering MeSH Representations of Medical Literature

Craig Struble, Department of Mathematics and Computer Science, Marquette University,
Milwaukee WI

Clustering documents is an important problem with applications in concept formation, knowledge extraction, and classification. Clustering papers in Medline, a collection of abstracts from medical related publications, has been previously investigated, with many successful applications. Many approaches for document representation and clustering are based on a full text analysis of the abstracts and bodies of each paper in a document collection.

We have been investigating an alternative approach for representing documents based on Medical Subject Headings (MeSH), an ontology for indexing papers in Medline and PubMed, an online database of publication abstracts that encompasses Medline. Using

MeSH based representations, we have been able to visualize and identify structure not readily seen with full text based representations. We present our results of clustering documents contained in the Rat Genome Database (RGD), comparing full text and MeSH based representations.

Textpresso: An Information Retrieval and Extraction System for *C. elegans* Literature

Eimear Kenny, Hans-Michael Mueller and Paul Sternberg

A major challenge facing researchers in biomedical sciences is extracting the vast amount of information available only in biological literature, most of it contained in individual papers. Manual extraction of information from scientific papers is tedious and slow. We have therefore designed Textpresso, a web-based system that aids the *C. elegans* researcher and professional curator in retrieving and efficiently extracting information from papers and abstracts. Textpresso recognizes words and phrases in text as belonging to word categories in the Textpresso Ontology. The Textpresso search corpus is automatically preprocessed so that these words and phrases are annotated with their corresponding ontology category. For example, the ontology category "Regulation" would be annotated to words such as "enhance", "repress", "regulate" etc. The ontology includes all terms from the Gene Ontology (GO) Consortium. The semantically marked-up text is presented in XML format, making it available to XML-processing software tools. A web-based user interface (<http://www.textpresso.org/>) offers two ways to search this corpus using keywords and/or categories; the easy-to-use "Simple Retrieval" and the more complex and powerful "Advanced Retrieval". The Textpresso corpus comprises ~18,000 abstracts and ~2,700 full text papers information rich in *C. elegans* biology. The researcher is able to view sentences that match their query, as well as paragraphs, whole articles and citation information. The project currently focuses on *C. elegans* literature, however, an expansion to the literature of *S. cerevisiae* is planned and other model organisms should be straightforward. The project is part of WormBase (<http://www.wormbase.org/>) and GMOD (<http://www.gmod.org/>).

PubFetch: Collecting literature from multiple data sources

Vijay Narayanasamy & Simon Twigger
Rat Genome Database
Medical College of Wisconsin, Milwaukee WI 53226

Scientific literature curation to extract information forms an essential component of any model organism database (MOD). The literature data is available from multiple sources. Some of the publicly available electronic sources include PubMed and Agricola. There are several subscription based literature data sources as well. PubFetch provides a generic way of searching and retrieving literature data from online literature data sources so that the downstream applications don't have to deal with the idiosyncrasies of the individual literature databases. Apart from fetching the documents, PubFetch has functionalities to filter duplicate documents and to present the documents in the desired format. The current version of PubFetch retrieves documents from PubMed and Agricola and formats them into MEDLINE Display format. PubFetch is available as a stand-alone

command line Java application, web application, and also as a service in the BioMOBY webservices framework.

BioCreAtlvE: Critical Assessment of Information Extraction

Marc Colosimo, Alexander Yeh, Alexander Morgan, Lynette Hirschman
MITRE

This talk will describe MITRE's work in evaluation for text data mining applied to biology. We first summarize Task 1 for the KDD (Knowledge Discovery and Data Mining) Challenge Cup 2002, run by Alex Yeh at MITRE. The challenge was to develop an automated system for a task early in the FlyBase Harvard curation pipeline: identifying which articles to curate, based on whether they contained experimental evidence for *Drosophila* gene products. The rest of the talk will focus on the ongoing BioCreAtlvE evaluation which MITRE is running in conjunction with Christian Blaschke and Alfonso Valencia from CNB-Madrid. BioCreAtlvE includes a task on listing genes mentioned in abstracts (using data provided by the fly, mouse and yeast databases, as well as data provided by NCBI). The second subtask focuses on functional annotation of full text articles: specifically, automatic creation of Gene Ontology terms for proteins (with data provided by SWISS-PROT).

Curatorial procedures at Mouse Genome Informatics, with an emphasis on expression data

Constance M. Smith for the Gene Expression Database at Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor ME 04609

Mouse Genome Informatics (MGI) provides integrated access to data on the genetics, genomics, and biology of the laboratory mouse. Information in MGI is obtained via manual curation of the published literature, from electronic downloads, and from electronic submission. The Gene Expression Database (GXD), one of the databases comprising MGI, collects and integrates information about gene expression in the developing mouse. GXD is designed to integrate data obtained from many different kinds of assay types, including RNA in situ hybridization, immunohistochemistry, in situ reporter (knock in), Northern and Western blots, and RT-PCR. In GXD, as well as the rest of MGI, controlled vocabularies are extensively used to ensure uniform data coding and to enable complex queries of the data. For instance, expression patterns are described using an extensive dictionary of standardized anatomical terms, enabling the recording of expression results from assays with different spatial resolution in a consistent manner. Whenever possible text annotations are complemented by digitized images of original expression data to further interpretation of the primary data. GXD is available at the Mouse Genome Informatics site at www.informatics.jax.org.

Gene Expression Curation in WormBase

Wen J. Chen, Igor Antoshechkin, WormBase Consortium

WormBase gene expression curation focus on three major parts: descriptive analysis of individual genes, microarray and gene regulation.

We have screened all of the *C. elegans* publications (~7000) accumulated to date and manually extracted spacial and temporal gene expression data. Our first-pass curation pipeline continue to flag new papers for extraction. New expression data are released fort-nightly. WS112 contains experimental results from 2.354 experiments that studied ~1,700 genes. Thus, we consider ourselves complete and up-to-date on the curation of this type of data.

We have developed database models for microarray data that are based on MIAME (Minimum Information About a Microarray Experiment) recommendations and established a pipeline for paper curation and data entry into the database. The majority of data processing is carried out via Perl scripts and is automated, although some script modification is required due to differences in primary data file formats, which are usually obtained directly from authors. There are currently twenty one papers containing microarray data in *C. elegans* literature describing developmental expression profiles as well as gene expression under certain conditions such as genetic mutations or drug treatment. Seven of them have been curated and the data have been entered into WormBase. These data contain 595,451 individual expression level data points and 175 clusters. We estimate that we will process all data available in the literature by the end of this year. We are also working on developing database models for SAGE (Serial Analysis of Gene Expression) data. SAGE is another high-throughput method of gene expression analysis, which is gaining popularity. There are currently two SAGE papers in *C. elegans* literature.

We just started the curation on gene regulation, which includes all the experimental studies on how a gene or an environmental condition regulates the expression of other genes. This kind of data is also curated manually. Current curation on gene regulations follows our first-pass curation pipeline. Newly published articles are curated first so that there will soon be a complete collection of gene regulation data that are published after 2003. Curation on earlier articles will be done later.

To ensure consistency and to facilitate data analysis, we have been using developmental(life stage) ontology to code temporal data. In the near future, we will begin to apply a cell and anatomy ontology to describe spatial data.

Biological Interaction Curation In FlyBase

Chihiro Yamada
FlyBase Cambridge, Department of Genetics
University of Cambridge, CB1 3NY UK.

With a large number of genome sequencing projects having been completed, and more on the way, there has been a growth in interest in not only what genes are present in the genome, but how those genes interact. There are various ways to study this and computational studies of interactions can be facilitated by some of the types of data that Model Organism Databases curate. In FlyBase there are a number of classes of curated data that could be used in these studies and I shall be discussing them in my talk.

Interactions between mutant alleles of different genes have been curated in FlyBase for three years and now constitute a large body of data that Bioinformaticists are starting to examine. I will briefly discuss mutant allele curation in FlyBase, and then go on to explain how it is extended to cover interactions between mutant alleles of different genes.

I'll finish by talking about other ways our curation captures studies on interactions, including describing interactions with GO terms, and some initial work being carried out by Harvard with the aim to describe molecular interactions.

Mutant Manifests: toward a zebrafish phenotype ontology

David Fashena 1, Erik Segerdell 1, Melissa Haendel 1, Judy Sprague 1, Monte Westerfield 1,2

1 Zebrafish Information Network, 5291 University of Oregon, Eugene, OR USA 97403-5291 <http://zfin.org>

2 Institute of Neuroscience, 1254 University of Oregon, Eugene, OR USA 97403-1254

Abstract:

The function of genes during embryonic development is illustrated by the phenotypes resulting from mutated genes. Phenotype data from zebrafish mutants and gene knock-downs are being generated at an increasing rate. To accommodate the flood of new data in a way that will facilitate the phenotypic analysis of gene function, we are investigating the use of a zebrafish phenotype ontology. The phenotype ontology would complement our existing ontology of anatomical structures and behavioral and physiological processes. The structure of any phenotype would be:

Phenotype = observable + attribute + value + qualifier

The “observables” are species-specific and come from the zebrafish anatomical ontology. To facilitate cross-species comparisons, we ensure that as many anatomical terms as possible are shared between zebrafish and mouse. Homologous structures like fins and limbs are cross-listed. The “attributes, values and qualifiers” come from the cross-species phenotype ontology being developed in collaboration with the Phenotype Ontology Consortium. The zebrafish phenotype ontology will allow the annotation of zebrafish phenotypes in a format that enables ready comparison to mutant phenotypes in other organisms. The system is sufficiently flexible to accommodate mutants, morpholinos and environmental factors. This is a significant expansion in the way ZFIN curates mutant phenotypes and will entail new challenges for our curators.

Community Curation at MaizeGDB

Carolyn J. Lawrence, Mary L. Polacco, Trent Seigfried, and Volker Brendel

The Maize Genetics and Genomics Database (MaizeGDB) is a central repository for maize sequence, stock, phenotype, genotypic and karyotypic variation, and chromosomal mapping data. The MaizeGDB team endeavors to make use of the maize

community's expertise and willingness to provide expert annotation. To this end, community curation tools have been created and are available for public use. How community curators can use the curation tools to contribute data directly to the database will be presented, and some of the protocols implemented to ensure that new records added to the database by community curators are of the highest quality will be discussed.

Community Interactions: Feedback, Support and Curation

Eva Huala

TAIR serves the needs of a large community of plant biology researchers ranging from professors to undergraduates along with teachers, students, and others. Our mandate to serve the needs of the plant biology community requires that we be responsive to user input of all kinds, ranging from simple questions about where to find information or how to use tools, to suggestions for improvement of datasets and tools or requests for specialized datasets. To make TAIR more accessible to all users we provide several avenues for gathering data, hearing community feedback and keeping in touch with the needs and desires of our community. We accomplish this by including allowing users to post comments directly on object detail pages, using open source tracking software (Jitterbug) to assign, respond to and archive user questions to TAIR, making custom datasets available in our User Requests ftp directory (ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/User_Requests), and giving workshops on how to use TAIR.