Pombase
Curation types: GO (including extensions), Phenotypes (FYPO) (single gene at present but multi-gene shortly), Genetic and physical interactions (BioGRID format), complementation Modifications (MOD) ,DNA and protein sequence features (SO), Domains and motifs (submitted to Pfam), Manual curation of human and S. cerevisiae orthologs
human disease associations, gene expression from small scale experiments
and a few smaller controlled vocabularies with low frequency

Reactome
placing human proteins in reactions that describe their molecular functions and subcellular locations, working with data from the primary published research literature.

DictyBase
Literature: papers in PubMed that have Dictyostelium and or discoideum in title, abstract or keywords are imported into dictyBase weekly, Literature curation: gene names, protein products, short descriptions, GO, strains, phenotypes; broad categorization of papers into literature topics, GO development, new terms requests, Development of phenotype ontology, Update of assay and environment controlled vocabularies for strain curation, Answering to user requests regarding any annotation, Gene Model curation (first part completed in summer 2011, now only occasionally): start/stop, exon/intron boundaries, Moderation of community annotations: adding summary on gene page, Updating curation status notes when done with a gene, Writing a summary paragraph for gene.

ZFin
Sequence-gene associations (not primary sequence curation), alleles, genotypes, morpholinos, antibodies, gene expression, phenotypes using EQ syntax, transgenic constructs, transgenic insertions, data exchanges (with NCBI, UniProt, GOA, GOC, MODB, GEO, mirBase, probably others!)

CGD/AspGB
literature curation of gene-specific data manual correction of gene model structural annotations user mail documentation keep static pages on web site up-to-date work with programmers on addition of new data, new species spec out improvements, interfaces, tools, testing simple queries and scripts to fulfill user requests (programmers take care of the difficult ones) community outreach preparation of posters and manuscripts, help with grant-writing periodic projects such as merging of annotated gene sets from other groups, incorporation of sequence updates In addition to manual curation of the literature, we determine orthologs (via Jaccard clustering and InParanoid) and protein domains (IprScan), and we use these data to make automated inferences of GO annotations and to supplement our free-text descriptions of otherwise uncharacterized genes. The GO-related components of this pipeline are described in detail at: - CGD Prediction of Gene Ontology (GO) annotations based on orthology: http://www.candidagenome.org/cgi-bin/reference/reference.pl?dbid=CAL0121033 - CGD Prediction of Gene Ontology (GO) annotations based on protein characteristics (e.g., domains and motifs): http://www.candidagenome.org/cgi-bin/reference/reference.pl?dbid=CAL0142013 - AspGD Prediction of Gene Ontology (GO) annotations based on protein characteristics (e.g., domains and motifs): http://www.aspergillusgenome.org/cgi-bin/reference/reference.pl?dbid=ASPL0000166200 - AspGD (2011) Prediction of Gene Ontology (GO) annotations based on orthology: http://www.aspergillusgenome.org/cgi-bin/reference/reference.pl?dbid=ASPL0000000005 Data Represented in the Databases: - Gene data including names, descriptions, GO, mutant phenotypes, notes on sequence and annotation

updates and issues, protein domains and properties, orthology (displayed on gene pages in both databases, in addition, AspGD has implemented Sybil Comparative Genomics Browser), gene model structure (including UTRs, uORFs), coordinates, gene sequence - Sequence data including chromosome/contig and gene sequence, SNPs (displayed on GBrowse), large-scale sequence datasets (new in AspGD, using GenomeView for display) - Myriad downloadable files including sequence, curated data, orthology, archived experimental datasets, etc.: see http://www.candidagenome.org/DownloadContents.shtml and http://www.aspgd.org/DownloadContents.shtml - User-contributed data including colleague profiles, images and movies, archived datasets, etc.

MGI
Separate groups curate data for alleles and phenotypes, sequence, embryonic express, tumor, and function (GO). Data is for the most part literature-based, although data-loads are the primary source of data for the sequence group.

WormBase
http://wiki.wormbase.org/index.php/WormBase_Literature_Curation_Workflow

FlyBase
A 0.1FTE time is spent on fixing existing annotations rather than making new ones (e.g. to comply with GO QC checks and improved annotation standards, dealing with obsoletions.
To give some context, our curators are based at three sites (Cambridge University, UK, Harvard University and University of New Mexico); each site deals with different data types and set their priorities largely independently. GO curation is carried out in Cambridge along with curation of other broadly 'genetic' data types such as allele and phenotype information. Harvard curate 'molecular' information such as expression patterns, physical interaction, gene features. The University of New Mexico site is a devoted to annotating Drosophila gene models in species than Drosophila melanogaster - this does not include GO annotation).

Once a paper has been selected for curation by Cambridge, GO annotation is given equal priority with the other data types we curate - all relevant 'genetic' data from the main paper are curated (note: we do not have the resources to routinely capture all supplementary data). Papers are prioritized based on an initial triage process, primarily carried out by the authors, that flags the different data types of interest to FlyBase in a paper – broadly, papers with the highest number of Cambridge relevant flags (gene split, gene merge, gene rename, new characterization, new allele, new transgene, phenotype data) are curated first. The potential for GO annotation is not explicitly flagged in this process at present but we find the 'new characterization' and 'rename' flags find papers that generate novel GO annotation effectively.

We curate information about the following features: Aberrations, Alleles, Balancers, Cell lines, Clones, Genes, Images, Insertions, Interactions, Library collections, Natural transposons, Polypeptides, Recombinant constructs, References, Sequence features, Stocks, Transcripts, Transposons.

In addition to GO, SO and MI, we use a number of onotologies developed in house (FBbt, FBdv, FBbi, FBsv, FBcv) to describe fly anatomy, fly development, imaging methods, stocks, phenotypic class, allele class and publications.

We will curate any source of information about Drosophila (our publication types include everything from jigsaws and microscope slides to obituaries and postage stamps.) However the

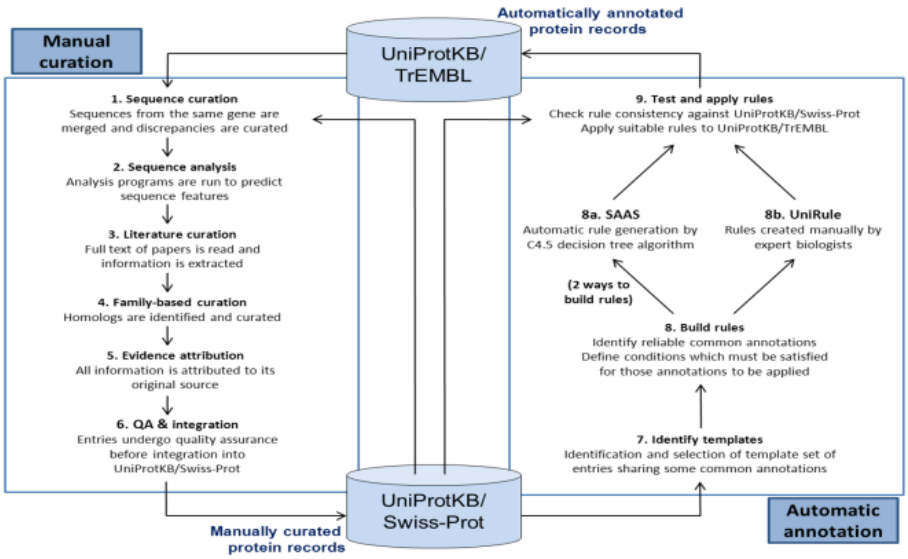
main source of information is primary research papers about Drosophila identified via PubMed (we no longer routinely curate information from review articles).

The other mains sources of information are:

directly submitted HTP data (e.g. from modENCODE)
personal communications from users (e.g as a source of gene merge)
stocks
data generated internally (e.g. GO ISS annotations, gene models)
data supplied directly by other groups (e.g. InterPro based GO annotations)

We no longer use meeting abstracts as a data source but have legacy GO annotations from this source.

UniProt



See flow chart in original Document.

Annotation methods applied to UniProtKB/Swiss-Prot include manual extraction and structuring of information from the literature, manual verification of results from computational analyses, mining and integration of large-scale data sets, and continuous updating as new information becomes available. UniProt has developed two prediction systems, UniRule and the Statistical Automatic Annotation System (SAAS) to automatically annotate UniProtKB/TrEMBL in an efficient and scalable manner with a high degree of accuracy

Manual GO curation fits into the literature curation activities of UniProt curators while automatic GO annotation is produced in three ways:

a) As part of the UniRule creation process
b) By the manual mapping of GO terms to corresponding concepts in the controlled vocabularies used by the UniProt Knowledgebase, including UniProt keywords and subcellular Locations, Enzyme Commission numbers and cross-references to InterPro.

An additional Ensembl electronic annotation method uses orthology data from Ensembl Compara to project experimentally evidenced GO annotations from a source species onto one or more target species. The orthology data and resulting projections are supplied by the Ensembl group, using experimentally-evidenced GO annotations supplied from UniProt

UniProt curators annotate to GO as part of providing a full set of information for a UniProt record, selected according to UniProt annotation priorities.

UniProt manual annotation programs:

- Chordata protein annotation program
- Prokaryotic protein annotation program
- Fungal protein annotation program
- Plant protein annotation program
- Drosophila protein annotation program
- Caenorhabditis protein annotation program
- Viral protein annotation program
- Animal toxin protein annotation program

For the specialized GO annotators, the annotation priorities are:
- Comprehensive annotation of human proteins
- GO Consortium targets
- User requests
- Priorities for specific grants (e.g. renal target list)

TAIR
See attached flowchart for literature curation workflow.

Within our literature curation effort, GO curation is our top priority although we do also extract gene aliases, expression patterns and phenotypes. Other important curator tasks include processing community submissions of GO and PO annotations, approving gene class symbols, reporting bugs and testing bug fixes, and answering user questions.

Domain coverage - all aspects of plant gene function
Source of data - Primary literature, collaboration with plant journals, direct community submissions, Interpro2GO pipeline
Gene structure annotation

*Gene structures from JGI, IEA annotations using InterproScan and Interpro2GO done by TAIR.

For all published articles with Arabidopsis in the title, abstract or keywords, we associate the article to any genes mentioned in the abstract (based on gene symbol or locus identifier). We choose a subset of these for curation. In addition to GO annotations we make PO (Plant

Ontology) annotations for gene expression patterns. We extract gene aliases, allele and germplasm names, and phenotype information.

AgBase
GO Curation is the number one priority for the AgBase livestock biocurators.
For the biocurator who annotates cotton, her time is equally divided between GO & PO annotation, depending on the data in the papers.
We initially provided biocuration for chicken (taxon:9031) and cow (taxon:9913) but in the new AgBase grant are expanding this to include other agricultural livestock species (pig, sheep, turkey, horse). We annotate UniProt proteins, when they are available, NCBI proteins when there is no corresponding UniProtKB record. We have been providing IEA annotation for transcripts represented on arrays (NCBI) but now our users are moving towards RNASeq, so this is becoming less critical. We will also be getting our first experimental data sets of chicken miRNAs this year and will work to provide annotations for that and expand our annotations to ncRNAs.
Chicken & horse communities are expected to have a reference gene (& gene product) set this year. This will help us considerably with annotation.
For plants we have 1 FTE dedicated to providing cotton GO & PO annotation for experimental data sets.

SGD
See flow chart in document.
GO curation is one of the high priority tasks for SGD along with data types like sequence, phenotype, pathways, HTP data.
Following types of data are curated on an ongoing basis.

- Literature
- GO
- Phenotypes
- Sequence
- Interactions, physical and genetic
- Biochemical pathways
- Gene Expression
- High-throughput datasets
- Strains

Karen Christie 2/14/12 10:33 AM
Comment [1]: What are we still curating in this area that is different that "Multiple strain representation"?

InterPro
We integrate ~ 2,500 new signatures from member databases per year into InterPro, adding extensive annotation, including GO terms. We also maintain and refine the existing database (currently 31,979 member database entries integrated into 22,361 InterPro entries) and provide data for each release of the UniProt database. InterPro currently provides matches to ~80% of the sequences in UniProt, providing 69,115,915 GO annotations to 11,965,074 distinct proteins across a wide range of species.

BHF-UCL
We focus on the annotation of 4000 human genes identified as relevant to cardiovascular processes. However, we do annotate whole papers, and therefore will annotate other genomes, additionally we annotate other mammalian genomes if there is limited human data describing a cardiovascular priority gene.

Occasionally we write brief gene summaries (maximum 1 page) for individual genes identified as risk factors for cardiovascular disease. These are requested by individual scientists working in the

cardiovascular genetics research group.

MaizeDB
We curate all data related to maize genomics and genetics. This includes literature, genetic maps, loci, gene/gene models, qtls, stocks, cytogentics, variations, metabolic pathways, sequences, gene products, images, people, and organizations.

RGD
Gene curation - disease portal curation - disease annotations, GO annotations, pathway annotations, phenotype annotations
Gene curation - pathway curation - pathway annotations, GO annotations
Gene curation - reference genome curation, QTL candidate gene curation, and other - GO annotations
Curation (source):
GO curation: for rat genes only (literature)
Disease curation: rat, mouse and human genes, rat and human quantitative trait loci (QTL), Rat strains (literature)
Mammalian Phenotype curation: rat and human genes, rat and human QTL, rat strains (literature)
Pathway ontology curation: rat, mouse and human genes (literature)
Quantitative phenotype data (clinical measurements, measurement methods and experimental conditions): rat strains only (literature and from web- or file-based high throughput data)
QTL curation: rat and human (literature)
Data representation from automated pipelines (source; frequency):
GO annotations: mouse and human (GOC; weekly)
Gene and QTL phenotype annotations: mouse (MGI; weekly)
Disease annotations: human (OMIM, GAD; historic only—not ongoing)
Pathway annotations (KEGG; historic only—not ongoing)

SRI-coli
Domain coverage: all gene products encoded in the E. coli genome (i.e. not plasmids), regulation (transcriptional, post-transcriptional), protein complex formation, enzyme function, enzymatic reactions, metabolic and regulatory pathways
Source of data: peer-reviewed literature; on very rare occasions, personal communications (noted as such), patent applications (but neither of these are used for GO term evidence)

SoyBase
Glycine max (soybase), manual literature review
Legumes (LIS)

GeneDB
We predominantly curate genomes of parasitic organisms, and classify our efforts as ‘proactive’ (that is, we actively read the literature and update the database ourselves) and ‘reactive’ (that is, we rely on user comments and transfer of annotations across closely

related organisms). Our curation efforts are broken down into: GO annotation, phenotype curation, and structural annotation. There is often overlap between the phenotype and GO annotation activity.

SGN
The locus detail page has the following sections, all of which can be curated by community curators:

o Notes and figures
o Accessions (associate accessions with mutant phenotype - associations can also be described using ontology terms)
o Alleles (list of alleles and allele metadata)
o associated loci - regulation, interaction etc by other genes
o cyc links (for biochemical genes, link to Cyc database reaction
o sequence annotations: Genbank ids, SGN unigene ids, genome positions of sequences
o literature annotations (list of articles describing this locus)
o Ontology annotations (full ontology annotations with evidence, reference, etc.)
o User comments. Comments by users on the locus.

Source of data: literature, Genbank, other databases, community curators

MTB
GO curation creates data for later inclusion in reactome trees.
Reversely, reactome work sporadically finds missing papers for GOA.
Creation of reactome data of those biological processes of M.tb.
where most participating gene products are functionally characterized.