PRELIMINARY STUDIES

SGD

The *Saccharomyces* Genome Database (SGD) project designs, develops, and implements a database containing comprehensive biological information correlated with the genome of the budding yeast *Saccharomyces cerevisiae* and provides analysis tools that allow scientists to discover functional relationships between sequences and gene products within the context of the cell. To this end, SGD annotates the genome, assimilates new published results, and incorporates genomic information from other *S. cerevisiae* strains and other *Saccharomyces* species. SGD uses controlled vocabularies to represent biological concepts in order to facilitate computational analysis and the integration of heterogeneous data types. We also maintain and broaden relationships with the greater scientific community, and make technical improvements that allow SGD to better serve its users. The database and its associated resources are publicly available via the World Wide Web (http://www.yeastgenome.org/).

Data Type	Number
Colleague	10,744
Feature	21,796
GO annotations	64,295
Homolog alignment	3,084,247
Interactions	191,312
Literature Guide	413,608
Microarray probes (expression/tiling)	89,355
Other fungal sequences	138,807
Paragraph	1,463
PDB alignment	226,797
Phenotype	19,945
Primer	26,228
Reference	53,621
Sequence	51,029
Transcripts	53,375

Summary of SGD contents (number of rows) highlighting some of the main data categories as of January 2008.

ZFIN

The zebrafish has emerged as a premiere organism to study vertebrate biology. Powerful techniques allow rapid, efficient generation and recovery of zebrafish lines with mutations affecting genes that orchestrate developmental patterning, organogenesis, physiology and behavior. Recent advances make it easy to study gene function by generating transgenic zebrafish or by knocking down gene function with morpholino antisense oligonucleotides. The functions of many of these genes are conserved among vertebrate groups. Thus, analysis of zebrafish mutations provides insights into gene functions in other vertebrates, including humans. Ongoing genetic screens are identifying thousands of mutations. Large-scale projects have already produced sequence and expression data for over 10,000 genes, and the complete genome sequence is available.

ZFIN, the zebrafish MOD, is mandated to provide the research community with easy access to this immense and rapidly growing body of information and to integrate information obtained from studies of zebrafish genetics and genomics with information from other model organisms and humans.

The long term goals for ZFIN are a) to be *the* community database resource for the laboratory use of zebrafish, b) to develop and support integrated zebrafish genetic, genomic, developmental and physiological information, c) to maintain the definitive reference data sets of zebrafish research information, d) to link this information extensively to corresponding data in other model organism and human databases, e) to facilitate the use of zebrafish as a model for human biology and f) to help serve the broad needs of the biomedical research community.

Data Type	Number
Genes	27253
Alleles	5683
EST/cDNAs	28603
Total GO annotations	81495
Wild-type strains	21
Gene expression patterns	23375
Phenotype annotations	15736
Images (phenotypes, expression patterns)	55723
Anatomical structures	1967
Developmental stages	45
Links to other databases	305513
Publications	8876
Researchers	4371
Laboratories	529
Companies	61

Summary of ZFIN contents (number of records) as of January 2008.

RGD

The laboratory rat, *Rattus norvegicus*, is one of the primary 'physiological' model systems used in modern biomedical research. Starting from the first genetic linkage map published in 1987 (Robinson, 1987), the draft whole genome sequence was released in 2004 (Gibbs et al., 2004)and now the rat has a wide array of genetic and genomic resources available. Building on the physiology and the genomics the rat has been widely used as a model of complex genetic human diseases such as cancer, hypertension, renal failure, alcoholism, polycystic kidney disease and more. The Rat Genome Database (RGD, http://rgd.mcw.edu) was created in 1999 to provide a central repository for this increasing amount of genetic and genomic data. There are over 36,500 genes models in RGD along with their associated map locations, functional annotations (e.g. Gene Ontology), sequence links and references. At a physiological level, there are 1300 rat quantitative trait loci linking a wide variety of complex phenotypes to the genome. The focus of RGD going forward is very similar to that of ZFIN, described above – to develop and support integrated rat genetic, genomic and physiological information, particularly in ways that facilitate the use of this information in comparative genomic studies and translational research.

Data Type	Number
Gene	36718
QTL	1351
Rat Strains	1404
Microsatellite Markers	12871
ESTs	593879
References	22,348
Gene Ontology Annotations	182566
Disease ontology Annotations	4443
Mammalian Phenotype ontology Annotations	2248
Pathway ontology Annotations	2380

Summary of RGD contents (number of records) as of January 2008.

InterMine and FlyMine

InterMine is an open source database project based in the Department of Genetics at the University of Cambridge. It is a generic system that was developed to enable the FlyMine project, and is now being used in a number of other projects worldwide (U.S., Canada, multiple European locations, Japan) including the NIH-funded modENCODE project that includes both *Drosophila* and *Caenorhabditis*.

FlyMine is an integrated database of genomic, expression and protein data primarily from *Drosophila* and *Anopheles*. A sophisticated web application provides flexible query access to researchers. It was developed to address the difficulty of using, in an integrated way, the large-scale datasets being generated by modern biology: data are available in a wide variety of formats in numerous different places, making their joint analysis otherwise hard. As an integrated resource, FlyMine makes it possible to run efficient data mining queries that span these domains of biological knowledge.

The key features of InterMine are outlined below and these are also introduced through the short tour available from the FlyMine homepage at http://www.flymine.org

Custom Queries: It is possible to build complex queries through the web interface without requiring knowledge of a database query language such as SQL, and without needing to refer to a database schema. The QueryBuilder web interface allows users to browse the data model, place constraints on any attribute and define the output columns. A query optimizer (see below) allows even large and complex queries to be answered rapidly. Queries can be exported and imported as XML for sharing with others.

Template Queries: The web application allows queries that have been composed through the QueryBuilder to be easily converted into reusable web-forms called "template queries". This feature allows complex queries to be encapsulated for reuse by others. Users are able to construct their own personal template queries (and export/import them as XML), while a super-user is able to publish template queries for everyone to use. It is possible to search the resulting library of published template queries that typically cover commonly-used tasks. Having found a template query of interest, the user can either adapt it using the QueryBuilder or immediately run it on one or more items. The ability of a super-user to publish new templates quickly and at any time is useful for help desk responses.

Data Export: The results of any query or template query can be viewed as a table, from which it is possible to export data in a range of formats, e.g. tab-delimited, comma-separated (for import into spreadsheet packages), for which it is easy to customize the output columns. In addition, there are specific output formats for particular types of data (e.g. FASTA format for DNA and protein sequences, GFF3 format for genome annotation).

Lists: All template queries or custom queries can be run either on single items or on lists of e.g. genes in one step. Lists can be saved by selecting columns from the output of a query or uploaded using a sophisticated tool. The upload tool resolves input that contains heterogeneous identifiers, and is able to convert between different types of identifiers e.g. gene identifiers can be derived from protein identifiers and vice versa. Lists can be saved between sessions in a personal profile (see MyMine section below). An important feature is the ability to perform set operations on lists, e.g. to be able to combine lists and find the intersection between them. A super-user can also publish lists for general use. For example, such public lists can include gene sets identified in important papers.

List Analysis & Widgets: The properties of lists can be examined through list analysis pages. These include a number of graphical and statistical 'widgets' that help elucidate the properties of the set. For instance, for a list of genes in FlyMine, graphs are generated (using JfreeChart http://www.jfree.org/jfreechart/) that show the chromosomal distribution, tissue-specific profile of RNA abundance and *in situ* localizations. Statistical enrichment is calculated for a number of attributes, e.g. Gene Ontology (GO) controlled vocabulary, protein domains, and publications. All widgets are interactive and, for instance, clicking on the bar of a bar chart returns the corresponding subset for export or creation of a new list. See Figure 1 for an illustration of some of the FlyMine widgets.

Examples of FlyMine widgets that operate on lists of genes include:

General:

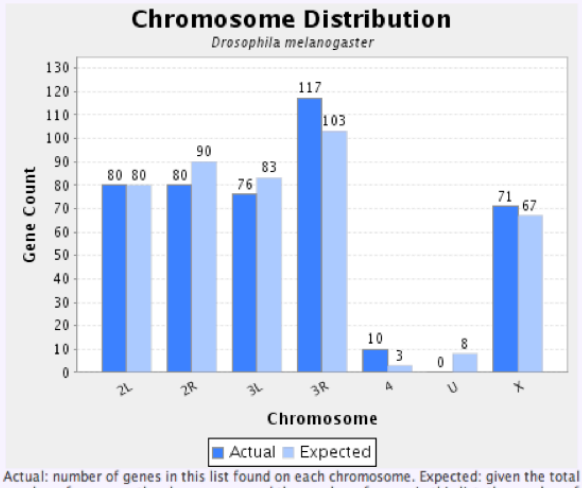
- Chromosome distribution of genome annotation features
- Statistical enrichment methods for
    - GO terms, i.e. terms that are over-represented in the set
    - protein domains, i.e. domains that are over-represented
    - publications, i.e. papers in which the list members are over-represented
    - UniProt protein features and keywords
- Pathway assignments from KEGG (Kanehisa et al., 2008)

Specific:

- Graph of tissue specific gene expression profile (FlyAtlas, Chintapalli et al., 2007)
- Graph of mRNA sub-cellular localization profile (Fly-FISH, Lécuyer et al., 2007)

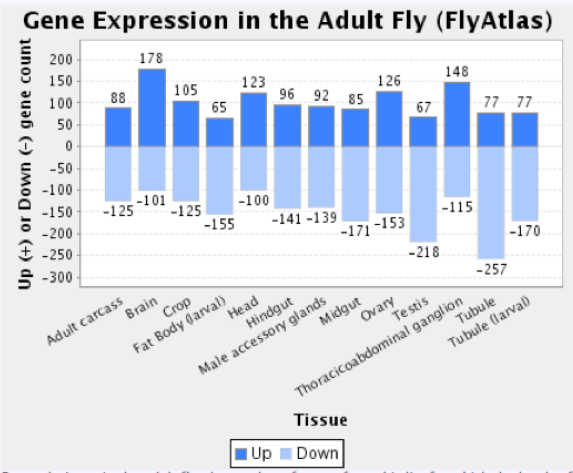
Report Pages: Every entry (e.g. gene, protein, etc) in an InterMine warehouse has a configurable report page. This displays basic attributes of the entry, and a super-user can configure template queries to be run whenever the page is accessed, thus providing the answers to more complex queries without any further action on the part of the user. This configuration is carried out through a tagging system in the web interface, which means templates can be added to report pages at any time. Report pages can also be configured to link out to other databases and to render data in integrated third party tools, e.g. GBrowse for sequence annotation and Jmol (http://jmol.sourceforge.net/) for protein structures.

a.



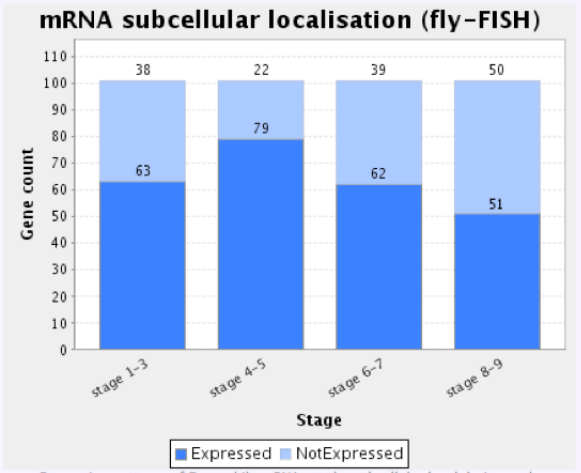
Actual: number of genes in this list found on each chromosome. Expected: given the total number of genes on the chromosome and the number of genes in this list, the number of genes expected to be found on each chromosome.

b.



For each tissue in the adult fly, the number of genes from this list for which the levels of expression are significantly high (Up) or low (Down) according to FlyAtlas AffyCall.

c.



Expression patterns of Drosophila mRNAs at the subcellular level during early embryogenesis - data from fly-FISH. Note that not all genes have yet been assayed by fly-FISH

d.

Gene Ontology Enrichment

Go terms enriched for genes in this list. Smaller p-values show greater enrichment. Method: Hypergeometric test with Bonferroni error correction (using a significance value of 0.05).

Reference population: All genes from: [Drosophila melanogaster].

ontology biological_process Update results

	p-value	
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process [GO:0019219]	0.0000000E0	[284 Genes]
regulation of metabolic process [GO:0019222]	0.0000000E0	[285 Genes]
regulation of cellular metabolic process [GO:0031323]	0.0000000E0	[284 Genes]
regulation of transcription [GO:0045449]	0.0000000E0	[282 Genes]
regulation of cellular process [GO:0050794]	0.0000000E0	[296 Genes]

e.

Protein Domain Enrichment

Protein Domains enriched for genes in this list. Smaller p-values show greater enrichment. Method: Hypergeometric test with Bonferroni error correction (using a significance value of 0.05).

Reference population: All genes from: [Drosophila melanogaster].

	p-value	
Homeobox [IPR001356]	3.3290025E-149	[87 Genes]
Homeodomain-related [IPR012287]	5.4036038E-144	[92 Genes]
Homeodomain-like [IPR009057]	5.0345805E-141	[91 Genes]
Basic helix-loop-helix dimerisation region bHLH [IPR001092]	4.9187248E-98	[54 Genes]
Helix-loop-helix DNA-binding [IPR011598]	4.2622180E-93	[54 Genes]
Zinc finger, C2H2-type [IPR007087]	2.0892021E-67	[71 Genes]
Helix-turn-helix motif, lambda-like repressor [IPR000047]	1.7167652E-52	[31 Genes]
Winged helix repressor DNA-binding [IPR011991]	4.5778598E-51	[42 Genes]

f.

Publication Enrichment

Publications enriched for genes in this list. Smaller p-values show greater enrichment. Method: Hypergeometric test with Bonferroni error correction (using a significance value of 0.05).

Reference population: All genes from: [Drosophila melanogaster].

	p-value	
Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. [12537568]	1.8037833E-135	[443 Genes]
The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. [12537573]	1.8073981E-135	[441 Genes]
Combined evidence annotation of transposable elements in genome sequences. [16110336]	1.8073981E-135	[441 Genes]
The genome sequence of Drosophila melanogaster. [10731132]	2.4740644E-135	[441 Genes]
Annotation of the Drosophila		

Figure 1: Some example widgets displayed on a FlyMine list analysis page for a list of 457 high confidence *D. melanogaster* transcription factors from FlyTF (Adryan et al, 2006). The bars on each graph and the gene counts can be clicked to get lists of genes in that category. a) counts of genes on each chromosome, b) numbers of genes found up/down regulated in tissues from FlyAtlas (Chintapalli et al, 2007), c) mRNA expression in early embryogenesis from fly-FISH (Lecuyer et al, 2007) d,e,f) statistical enrichment of Gene Ontology terms, protein domains and related publications, using a hypergeometric test with Bonferroni error correction (Benjamini Hochberg error correction now also available).

MyMine: By providing an email address and choosing a password, users enable a 'MyMine' account that allows them to store their own queries, template queries and lists between sessions as well as mark templates as favorites.

Web Services: A feature that will shortly be released (Q1 2008) is the ability to run the XML-representation of any query, including template queries, remotely via a RESTful web service. This has great potential for broadening access by other databases and programs as well as increasing interoperability. Other features of InterMine, e.g. lists and widgets, will also be accessible via web services soon (Q2 2008).

The generic nature of the FlyMine project and underlying InterMine infrastructure means that it is possible to construct similar integrated databases rapidly based on data from different organisms and datasets. The FlyMine project has already generated a substantial number of data parsers for commonly used data formats and data sources, and this forms a rich environment for the proposed project.

Figure 2 shows an overview of the architecture of InterMine. Web users interact with a web application that communicates with an object-relational mapping layer that executes queries in a PostgreSQL database (http://www.postgresql.org/). The object-relational mapping layer automatically maps between the real-world objects that biologists work with (genes, proteins, etc.) and their representations in a relational database. A key feature underlying the high performance of InterMine is the query optimizer that intercepts incoming SQL queries and greatly enhances performance.

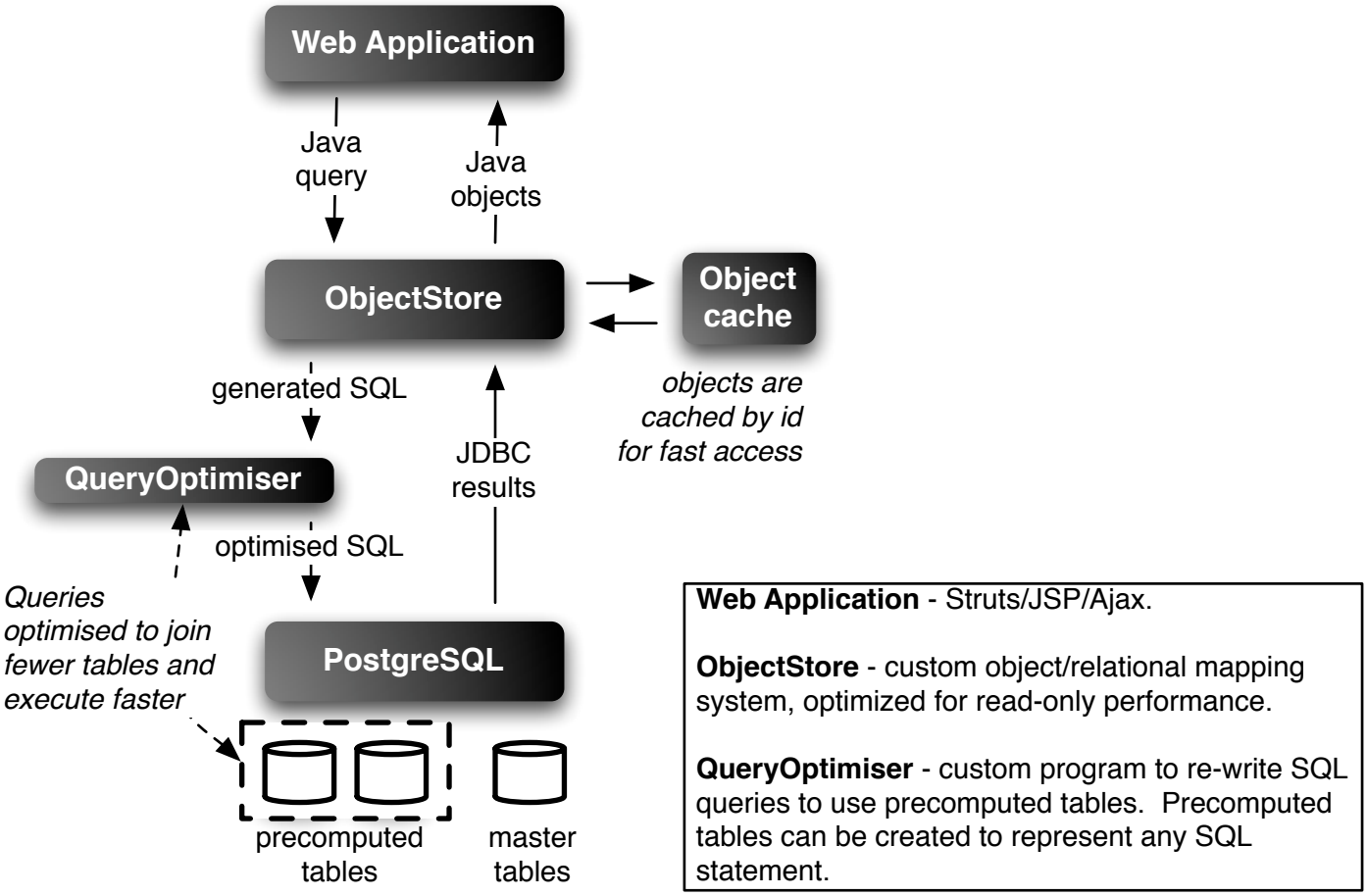


Figure 2: An overview of InterMine architecture.

Key technical features of InterMine are:

1. **A query optimizer:** A limitation of data warehouses is that they generally need to be designed and built with specific types of query in mind (schema denormalization). InterMine takes a different approach which involves transparent re-writing of queries to make use of precomputed tables (materialized views) that join any number of the master tables. The aim is to reduce the number of table joins needed and thus improve performance. As pre-computed tables can be added at any time the performance optimization of the warehouse is effectively uncoupled from the data model. All template and queries are pre-computed so that common tasks can be executed very fast.
2. **Automatic code generation:** Given a starting data model, the code to define the system is automatically written. Thus, little effort is needed to generate the user interface web application, the database itself, the Java classes of the related object model, or the programming interface. This greatly reduces the effort required to build and maintain new data warehouses.
3. **Easy configurability:** Many aspects of the web application are configurable and easy to extend. For example, the contents of report pages, data category 'home pages', external links and the inclusion of widgets are all governed by configuration files. Little or no code is required to create widgets for new types of data. A modular approach to code design means that, for example, it is relatively easy to write and include Java code to support a new data export format.
4. **A large range of supported data types**: InterMine includes several 'sources' that are able to parse data from one of a number of supported biological formats (see Table 1) and define how those data should be integrated. The data model has a modular structure, each source is able to add to the model as required. A central configuration system allows construction of a new Mine simply by selecting sources and specifying the data sets and organisms required. A framework is provided for simple addition of new sources.

Table 1: Currently supported data formats

InterMine can currently import data from many standard formats. Data for many organisms is available in these formats. Most need minimal extra configuration. GFF3 and chado may need additional handler code to deal with custom attributes. Data parsers are developed to convert data sources into a standard simple InterMine XML format that is directly loaded. This makes it easy to write data parsers for custom formats in any language. There are Java and Perl APIs to support this process.

Import formats:

Chado database	standard GMOD database format
Ensembl database	data from Ensembl database schemas
FASTA	DNA and protein sequences
GFF3	genome features
GO gene association	GO term assignments to genes/proteins
InParanoid tables	calculated orthologs
KEGG	assignments of genes/proteins to pathways
NCBI FASTA	FASTA with standard NCBI header
NCBI organism XML	organism details, automatically fetched by taxon identifier from Entrez web service
NCBI PubMed XML	publication details, automatically fetched by PubMed identifier from Entrez web service
OBO files	terms and structure of OBO ontologies
PDB XML	protein structure data
PSI XML	protein-protein interaction
UniProt XML	protein details from UniProt

Export formats:

TAB delimited	
CSV	for import into spreadsheets such as Excel and OpenOffice
GFF3	genome features
FASTA	sequences
SIF	protein-protein interactions to view in CytoScape
Integrated tools	
GBrowse	genome annotation
Jmol	protein structures