# Next-Generation Information Systems for Genomics

Chris Mungall

PhD by Research Publication
University of Edinburgh
2010

**Abstract**

INTRODUCTION:

The advent of next-generation sequencing technologies is transforming biology by enabling individual researchers to sequence the genomes of individual organisms or cells on a massive scale. In order to realize the translational potential of this technology we will need advanced information systems to integrate and interpret this deluge of data. These systems must be capable of extracting the location and function of genes and biological features from genomic data, requiring the coordinated parallel execution of multiple bioinformatics analyses and intelligent synthesis of the results. The resulting databases must be structured to allow complex biological knowledge to be recorded in a computable way, which requires the development of logic-based knowledge structures called ontologies. To visualise and manipulate the results, new graphical interfaces and knowledge acquisition tools are required. Finally, to help understand complex disease processes, these information systems must be equipped with the capability to integrate and make inferences over multiple data sets derived from numerous sources.

RESULTS:

Here I describe research, design and implementation of some of the components of such a next-generation information system. I first describe the automated pipeline system used for the annotation of the *Drosophila* genome, and the application of this system in genomic research. This was succeeded by the development of a flexible graph-oriented database system called Chado, which relies on the use of ontologies for structuring data and knowledge. I also describe research to develop, restructure and enhance a number of biological ontologies, adding a layer of logical semantics that increases the computability of these key knowledge sources. The resulting database and ontology collection can be accessed through a suite of tools. Finally I describe how the combination of genome analysis, ontology-based database representation and powerful tools can be combined in order to make inferences about genotype-phenotype relationships within and across species.

CONCLUSION:

The large volumes of complex data generated by high-throughput genomic and systems biology technology threatens to overwhelm us, unless we can devise better computing tools to assist us with its analysis. Ontologies are key technologies, but many existing ontologies are not interoperable or lack features that make them computable. Here I have shown how concerted ontology, tool and database development can be applied to make inferences of value to translational research.

# Contents

# List of Tables

# List of Figures

# List of Examples

# List of citations of key peer-reviewed publications

# 1 Introduction

## 1.1 Introduction

The arrival of next-generation sequencing technology is transforming biology by enabling individual researchers to sequence the genomes of individual organisms or cells on a massive scale[Mardis, 2008]. However, genome sequences are not in themselves sufficient to uncover the complex relationship between genotype, environment and phenotype. If genomics is to fulfill its translational potential and accelerate healthcare outcomes, then sequence data must be processed, structured, viewed and interpreted in the context of a wide variety of other kinds of information. One valuable experiment-rich source of such information is research on model organisms, including mouse, zebrafish and fruit fly *Drosophila melanogaster*. Even distantly related species can shed light on complex human diseases – for example, expression of mutant forms of the *α-synuclein* gene in a transgenic fruit fly recapitulates some of the essential cellular phenotypes of Parkinson's Disease[Feany and Bender, 2000]. If genome sequences and knowledge derived from humans and model organisms can be systematically combined then we can use data mining to search for patterns and generate new knowledge about the relationship between genes, phenotypes and disease.

However, the volume, fragmentation and sheer complexity of this data presents significant informatics challenges[Tyers and Mann, 2003]. Traditional knowledge-sources such as scientific journals structure knowledge in a human-centric way, as a combination of semi-structured narrative text and images. This has to be re-structured and processed in order to use many computational methods, but it can be difficult to do this systematically for complex heterogeneous biological phenomena such as the manifestation of mutant phenotypes and diseases. Tolstoy observed that "Happy families are all alike; every unhappy family is unhappy in its own way". Where model organisms are concerned, wild-type phenotypes are often alike, but each mutant genotype manifests its own particular mutant phenotype. This open-ended variation is a challenge for systematic structured database representations.

A single model organism such as *Drosophila melanogaster* has a dedicated horizontally integrated database[FlyBase-Consortium, 2002][FlyBase-Consortium, 2003], populated with a multitude of datatypes such as genotype-phenotype associations, gene models, gene function and gene expression data. These databases are complex and expensive to maintain, requiring dedicated curators[Bourne and McEntyre, 2006], and represent only the tip of the data iceberg. Each such database is complemented by a wide variety of vertical domain-specific databases that cut across species boundaries, such as the ma-

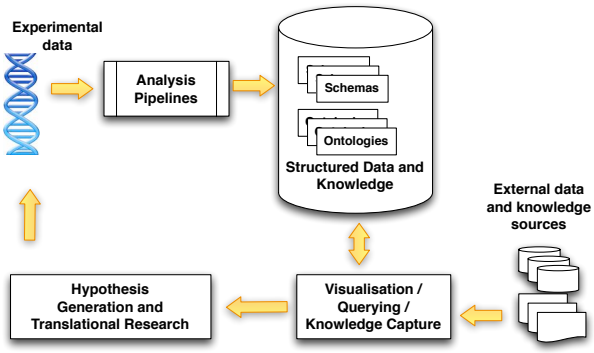


Figure 1: The key challenges of next-generation genomic information systems: genome analysis, structuring knowledge, visualisation/querying/editing and delivery of translational results

jor sequence databases[Benson et al., 1999], expression databases[Parkinson et al., 2007] and databases of protein-protein interactions[Breitkreutz et al., 2008]. In addition, there are many specialized databases intended for epidemiological data, neural network maps, neurological diseases[Martone et al., 2004]. The journal *Nucleic Acids Research* has published articles on a total of 1170 distinct databases[Galperin and Cochrane, 2008]. All of these databases are developed by multiple distributed groups employing varying technologies and data structures, resulting in fragmented data *silos*. This is a fundamental obstacle to biology as an information science and the application of data mining techniques. Currently, investigators must manually integrate data themselves, for example by visiting multiple websites[Stein, 2003].

If we are to take full advantage of the wealth of next-generation sequence data, we need *Next-Generation information systems*. These systems must be flexible enough to analyze and combine heterogeneous data and knowledge across the kingdoms of life from a variety of modalities and sources, from model organisms through to human disease data. These systems should be equipped with *intelligent inferencing capabilities* in order to answer complex questions posed by researchers, or to discover implicit correlations distributed across fragmented datasets. These systems must also be able to work together as part of a larger knowledge cloud, speaking the same language in order to allow synergistic aggregation of multiple data sources.

The key challenges in the development of next-generation genomic information systems can be broken down into four main areas (see Figure 1):

- *Genome analysis and annotation.* Given a genome sequence, how can

we harness computational resources and tools to uncover the location, structure and function of the biological features encoded in that sequence?

- *Formal structures for representing data and knowledge.* How do we structure data and knowledge from a variety of sources in a systematic form, that allows computers to answer complex queries and make inferences?

- *Visualization and knowledge acquisition.* How can we present complex data to investigators such that they can see important correlations in the data, form their own hypotheses and contribute their additional knowledge?

- *Hypothesis generation and translational research.* How can we mine multiple data sources and algorithmically generate inferences and hypotheses of relevance to human health?

The program of research and development described in this thesis addresses these challenges.

## 1.2 Thesis Structure

This thesis is structured around the four key areas, summarised above. After this introduction, the background section (section 2) covers the relevant concepts and literature in each of the four areas. The next four sections (3 - 6, pages 22 - 48) constitute the body of the thesis, and consist of the main research results organized according to the challenges outlined above. Section 3 describes an analysis pipeline devised for the analysis and annotation of the *Drosophila melanogaster* genome, and its application in a number of *Drosophila*-centric analyses. Section 4 covers the development of formal structures for representing complex biological information - an ontology-centric relational database and the construction or enhancement of a number of biological ontologies. Section 5 describes the development of tools to allow researchers to interact with and contribute to these information sources. Section 6 concludes the results by showing how the information systems described in sections 3-5 can be used to explore biological questions of relevance to human health.

Each of these four results sections are divided into subsections, with each subsection centered around a publication or group of related publications. I made contribution to all papers cited in these four sections – if a cited paper is particularly notable, it is highlighted with an asterisk and accompanied by

a footnote explaining my contribution (see page ix for a page index of all key paper citations).

The discussion section (section 7) reviews the results in the context of current research, and summarizes the impact of some of the key findings, in the context of ongoing research. Concluding statements are in section 8, and acknowledgments are in section 9.

Section 10 (page 58) lists all publications cited in the thesis. These are listed in order of citation, such that all my publications are be listed as a single contiguous block, corresponding to the four results sections (3-6). The references section is annotated with summary descriptions of each of my publications.

# 2 Background

## 2.1 Genome Analysis and Annotation

### 2.1.1 The need for accurate and comprehensive genome analysis

The sequence of *Drosophila melanogaster* was determined in 2000 in a collaboration between the public Drosophila Genome Projects and Celera genomics[Adams et al., 2000]. As the second metazoan to be sequenced, this was an event of major significance. With the increasing cost-efficiency and speed of today's next-generation sequencing technology, the delivery of a new genome sequence does not merit such fanfare. One challenge that has remained constant in the intervening decade is in understanding how these genomes instruct the cell to make the necessary components for the organism to develop and function. The DNA sequence does not yield this information readily - first it must be carefully *annotated* to locate and characterize the regions containing structures of interest such as genes, regulatory elements and transposable elements. Accurate annotation is particularly important for key model organisms such as *Drosophila melanogaster*, for which there exists a large body of knowledge derived from experimental research. This knowledge is invaluable in interpreting the genome.

### 2.1.2 Pipelines and workflows

Annotating a genome involves combining sequence alignments from programs such as BLAST[Altschul et al., 1997] and Sim4[Florea et al., 1998], as well as gene prediction programs such as Genie[Reese et al., 2000] and tRNAscan-SE[Lowe and Eddy, 1997].

For small DNA regions, researchers can run these programs themselves, and manually integrate the results. However, this approach is not scalable to large genomes whose assembled sequence is not yet stable. Automating the execution and synthesis of multiple sequence comparison and prediction tools requires coordination by a genome analysis *pipeline* executing some pre-determined *workflow*. The workflow specifies dependencies between tasks, and the pipeline ensures the tasks are executed in the correct order. Sometimes this involves dispatch of multiple tasks in parallel on multiple independent processors (Beowulf clusters)[Sterling et al., 1995].

Despite the fact that many genome analysis algorithms fall into the so-called *embarrassingly parallel* category, the construction of pipeline software is difficult. This is due to a number of reasons - (1) complex interdependencies between analyses (2) the need to be fault-tolerant in the face of system problems (3) the need to synthesize the total set of analyses into

a coherent whole and (4) dividing large datasets into manageable chunks. Unfinished genomes pose additional difficulties, as the genome is still fragmented over multiple unstable assembled regions. Additionally, a pipeline must take care of automatically triggering a new analysis on the change of an assembled sequence, and of mapping forward existing annotations (sometimes called "lifting over").

Historically, most bioinformatics workflows have been ad-hoc collections of scripts written in the programming language Perl[Stein, 1996]. These can be difficult to extend and maintain. In many cases it is desirable to allow biologists lacking programming experience to manage, edit and monitor workflows and their execution. One of the first tools to provide this kind of capability was Clone Curator (Helt et al, unpublished), which was used in the annotation of the *adh* region of the *Drosophila melanogaster* genome[Ashburner et al., 1999].

### 2.1.3 Automated and manual annotation

Workflows may be entirely automated, or they may involve expert human curation. Automated workflows are more scalable to the ever growing number of genomes (computers are less expensive than trained biologists), but also more error-prone. The Ensembl analysis pipeline[Potter et al., 2004] is an example of a scalable system that is used for automated annotation of multiple genomes. This pipeline is complemented by the Otter manual annotation system[Searle et al., 2004].

The existence of manually generated annotations based on experimental evidence is particularly important for model systems such as *Drosophila melanogaster* which acts as a source for downstream automated annotations.

### 2.1.4 The challenges of heterochromatic regions

Like most eukaryotes, *Drosophila* includes substantial fraction of heterochromatic DNA at the telomeres and centromeres. These regions pose particular problems for both contig assembly and annotation due to the presence of transposable elements and long stretches of repetitive sequence. The assembled genome sequence of heterochromatin is often in flux compared to euchromatin, necessitating a dynamic approach to annotation, in which located genome structures are frequently mapped forward to new assemblies. Tools such as RepeatMasker[Smit et al., 1996] can be executed as part of the pipeline, taking as input DNA sequences, and emitting sequences with repetitive regions masked out. This is a necessary preparatory step in the pipeline for some gene finding and alignment tools. However, this step is usually

not sufficient to accurately detect all genes and alignments in heterochromatin. One reason for this is because transposable elements have inserted themselves into the introns of the gene, resulting in uncharacteristically long introns which do not match the gene finders' trained models.

### 2.1.5 Comparative analysis

Analysis pipelines also make use of comparative data. Comparisons can be made between the genome, transcriptome and proteome of the reference species and multiple additional species. Comparisons at the DNA sequence typically only make sense for closely related species, but protein conservation is frequently observed across distantly related species. These comparisons not only inform the annotation of the genome of interest, but also shed light on molecular evolutionary processes and history - for example, the gain and loss of genome features such as introns over time [Stoltzfus, 2004]. We have still much to learn about these processes, and the growing number of sequenced genomes can help, provided that accurate experimentally-verified genome annotations are available.

## 2.2 Formal Structures for Representing Biology

### 2.2.1 The need to structure knowledge

Biological knowledge and data is frequently captured using a combination of natural language and ad-hoc semi-structured data files. This is convenient for the researchers producing data, as it requires no expertise in data modeling or formal knowledge representation. However, this kind of unstructured representation is far more difficult to compute over. Despite the advantages made in natural language processing and search engine technology, it is still difficult or impossible to pose queries to text-oriented search engines such as "find *Drosophila* genes whose function contributes to eye development", or "what zebrafish mutants exhibit similar phenotypes to Holoprosencepahly in humans?". In order for computers to answer questions such as this, we need to *structure* knowledge and data in a computable way.

There are a number of approaches one might take for this structuring, both formal and informal. Here we focus on two complementary formal paradigms for structuring data and knowledge: *Relational Schemas* and *Ontologies*.

### 2.2.2 Relational Databases and Relational Schemas

A relational database is a system for managing and querying data structured according to the relational model, the fundamental concept of which is the *n-ary relation*, a subset of the cartesian product of $n$ domains[Ullman, 1988]. These can also be thought of as tables, with each n-tuple of the relation constituting a row of the table. A database is a collection of relations, and the structure of the database is described by a relational schema, a set of constraints over the set of values for each relation. The Ensembl[Birney et al., 2004] core schema represents genomic features such as genes and regulatory regions and their component structures. For example, the exon_transcript ternary relation is constrained to take as arguments: an exon, a transcript, and an integer denoting relative order of transcription of the exon within that transcript. This is formally described as:

$$\mathsf{exon\_transcript} \subseteq E \times T \times \mathbb{N}$$

Where $E$ and $T$ represent the set of exons and transcripts respectively.

Because relational databases are based on mathematical principles it is possible to automatically optimize queries. This means that relational databases can efficiently store extremely large quantities of data, and can answer complex queries within reasonable time constraints. This makes them extremely useful in bioinformatics, especially where high data volume is concerned.

Relational systems are not ideal for all situations however. One issue is the cost of schema evolution. Another issue is the lack of *expressivity* of the relational model and relational schemas. It is difficult or unwieldy to directly model complex biological entities such as the full set of genomic feature types or possible phenotypes of an organism. This is particularly true where type hierarchies are concerned; for example, if we introduce sub-types of *transcript* such as *mRNA*, *tRNA* etc.

### 2.2.3 Ontologies and Controlled Vocabularies

An ontology is a logic-based organizational structure for knowledge[Washington and Lewis, 2008]. One type of ontology commonly encountered in biology is the anatomical ontology, a representation of the types of entities making up an organism, and constraints on the relationships that hold between these parts. If ontologies are grounded in logic they can be used for inferential reasoning[Stevens et al., 2000]; queries for genes expressed in the *brain* should return genes expressed in the *cerebellum*, based on *is_a* and *part_of* relationships in the ontology (figure 2A).

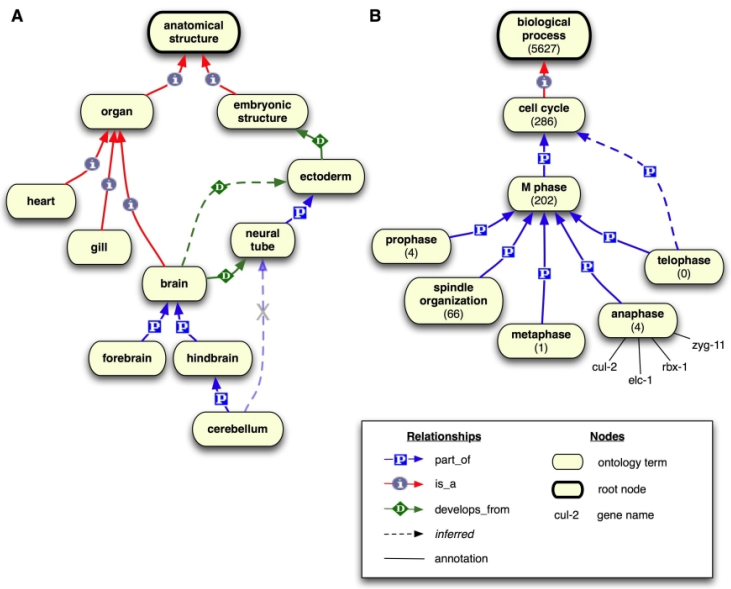


Figure 2: Example ontologies. (A) Subset of the Zebrafish Anatomical Ontology. Arrows indicate relationships between classes, relationship type denoted by single-letter icon (B) Subset of the Gene Ontology. Queries for genes with products that participate in the *cell cycle* would return *cul-2*, *elc-1* etc through logical inference. Taken from Washington and Lewis, 2008

The term "ontology" is frequently used interchangeably with *terminology*, *controlled vocabulary* or *thesaurus*. Whilst there is no widely agreed upon definition that systematically distinguish between these, typically the emphasis of an ontology is on a formal, logical structure, intended to be used by machines for automated query answering and inference. In contrast, the emphasis of a terminology or controlled vocabulary is on linguistic elements and human navigation and usage. In practice it can be difficult and unnecessary to differentiate; many so-called biological ontologies are in fact hybrid representational structures incorporating some mix of both terminological and logic-oriented ontological aspects.

### 2.2.4 The Gene Ontology

A case in point here is the Gene Ontology (GO)[Ashburner et al., 2000][GO-Consortium, 2006][Consortium, 2007], which is used to describe the function and subcellular localization of gene products. The GO consists of some 27,000 interconnected terms divided into three aspects: molecular function, biological process and cellular component (see figure 2B). The primary utility of the GO is in the large corpus of *functional annotations* - curated associations between genes and GO terms spanning thousands of species, but most heavily populated for the main model organisms and human (because of the volume of experimental data). The GO is ubiquitous in genomic research, and the large corpus of associations between genes and GO terms is frequently used in the analysis of high-throughput experiments[Khatri and Draghici, 2005].

The original incarnation of the GO took a deliberately informal approach as a design choice to avoid the "analysis paralysis"[Lewis, 2004] associated with formal ontological approaches. This strategy was successful in allowing the rapid development of a large number of terms and annotations, but the lack of logical underpinnings limited the computability of the ontology[Smith and Kumar, 2004].

For example, the GO term *negative regulation of cysteine biosynthesis* indicates to a human reader that (a) the direction of regulation is downwards; (b) the process regulated is one of *cysteine biosynthesis*, and (c) in the regulated process, the output is one of *cysteine*. However, these facts are stated in computationally opaque human-readable definitions rather than logical axioms referencing other ontologies such as `CHEBI`[Degtyarenko et al., 2007].

### 2.2.5 Formalization of Ontologies

These observations also held true for a number of other controlled-vocabulary style biological ontologies. Most of these ontologies lacked formal semantics

and computable definitions of the terms used, which limited support for automated reasoning[Soldatova and King, 2005].

One proposed approach was to recast the GO in the Web Ontology Language (OWL), a Description Logic (DL) with formal semantics and built-in logical constructs for composing descriptive class expressions[Wroe et al., 2003]. This has the benefit of using mathematical set-oriented techniques to describe biological phenomena in a way that allows for the algorithmic detection of subsumption relationships and unsatisfiable classes.

For example, the cellular component *mitochondrial membrane* could be represented as the set intersection between (a) the set of all *membrane*s and (b) the set of things that stand in a *part_of* relation to some member of the *mitochondrion* set. There are various DL notations, the most human readable being OWL Manchester Syntax[M.Horridge et al., 2006], in which the above expression would be rendered as *membrane* AND *part_of* SOME *mitochondrion*.

We can formally define the GO class using an *equivalence axiom*, as in the following example:
**(Example 1)** *mitochondrial membrane* EquivalentTo: *membrane* AND *part_of* SOME *mitochondrion*

Another way to view this is as a genus-differentia definition; we are defining a class based on a genus or general class *membrane* and one or more discriminating or differentiating characteristics (being part of a mitochondrion).

Using a formalisation such as this allows us to use computational tools such as reasoners to infer for example that the set *mitochondrial membrane* is subsumed by the set *organelle membrane* (figure 3).

### 2.2.6 Formalization of Relations in Ontologies

Another factor differentiating a terminology or thesaurus from an ontology is that the former relates terms loosely in an informal fashion into a semantic network, whereas the latter uses relations with formally defined semantics. The original incarnation of the GO was more akin to a semantic network, with loosely defined relations. This meant that some relations such as *part_of* were used inconsistently, yielding incorrect answers to queries. This was also true of other emerging biological ontologies, such as anatomical ontologies. In addition, the use of relations was inconsistent across as well as within ontologies[Smith et al., 2004].

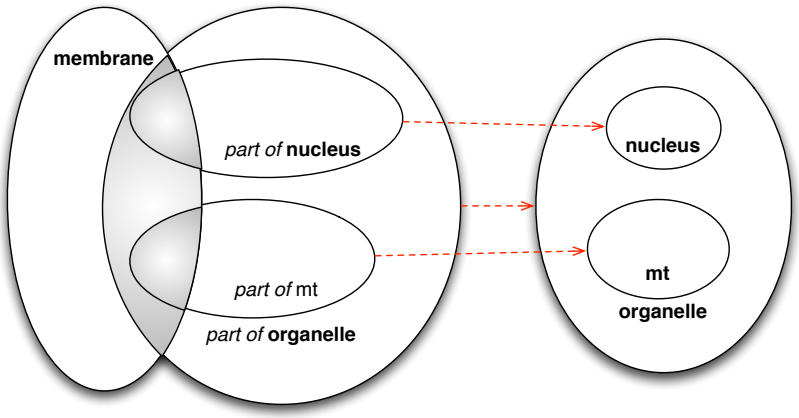


Figure 3: Venn diagram illustrating set-theoretic view of GO. Circles denote sets whose extensions are cellular component instances. Dotted arrows denote set restrictions over the part_of relation; e.g. all members of the set *part of nucleus* must be part of a member of the set *nucleus*. The diagram shows that the set intersection *membrane* ∩ *part of mt* is necessarily a subset of the set intersection *membrane* ∩ *part of organelle* (grey). These subset relationships can be inferred automatically if the ontology defines classes in terms of set intersections.

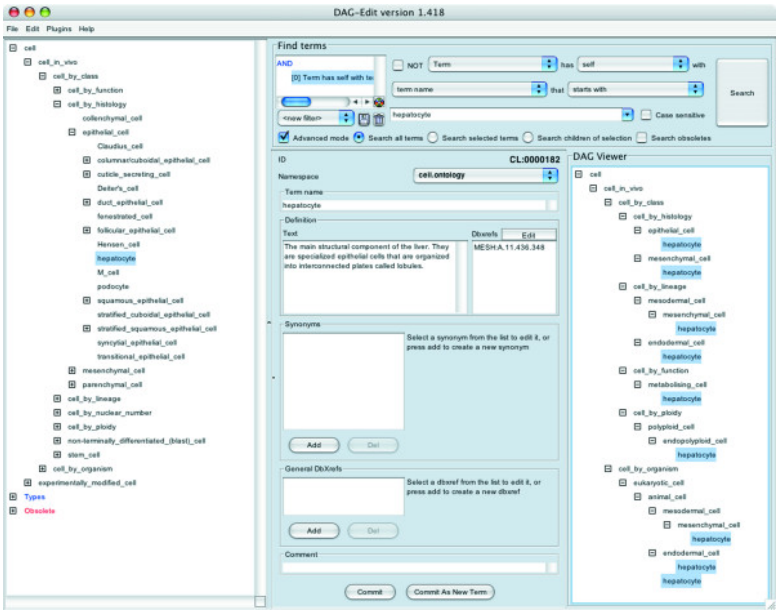


Figure 4: Screenshot of ontology editor view of Cell Ontology (CL). Taken from Bard et al, 2005.

### 2.2.7 Anatomical Ontologies

The complete set of tissues and components in an organism (its *anatome*) has to be organized in a computer-comprehensible way to support systems biology[Bard, 2005]). Anatomical ontologies are required for, amongst other things, specifying the location in an organism in which genes are expressed.

Two of the key challenges in developing anatomical ontologies are *granularity* and *variation between species*. At finer levels of granularity we see less variation which means we typically have a single ontology serving multiple species. The GO cell component ontology is applicable across all kingdoms, as is the Cell Ontology(CL)[Bard et al., 2005]. However, ontologies of gross anatomy such as the adult mouse anatomical ontology (MA)[Hayamizu et al., 2005] typically serve a single species. This creates a problem for translational research - data from one organism cannot be readily compared against data from another organism without extensive *ontology mappings*. These mappings are error-prone and expensive to maintain, especially as the number of species increases. This interoperability problem is compounded when we have multiple anatomical ontologies serving one organism, each built according to different principles, as in the case of the adult and embryonic mouse ontologies[Baldock et al., 2003].

### 2.2.8 Models and Ontologies for Phenotypes

The inherent variation in mutant phenotypes poses problems for ontologies which attempt to capture all possible combinations in advance, such as the Mammalian Phenotype (MP) ontology[Smith et al., 2005a]. The MP is an example of a *pre-composed* ontology, like the GO. This means that every phenotype is represented by a single atomic named class, composed in advance of annotation. This leads to a lot of repetition and redundancy, as can be seen when we consider the cross-product of orthogonal sets of ontology classes:

**(Example 2)**

$$\{increased, decreased\} \times \{lung, brain, ...\} \times \{size, weight, ...\}$$

In the MP we see classes like *increased lung size*, *increased lung weight*, ... composed in advance. Anatomical terms such as *lung* are referenced *implicitly* - i.e. they are visible to a human but opaque to a computer (without unreliable string-matching techniques).

The perceived problems of pre-composition led to an alternative model in which classes from different ontologies are dynamically combined (post-composition to compose descriptions, making use of a core Phenotypic And Trait Ontology (PATO) of phenotypic attribute-value qualifiers[Gkoutos et al., 2004]. This model was called the Entity-Attribute-Value (EAV) model[Gkoutos et al., 2005a][Gkoutos et al., 2005b] (figure 5).

There were a number of problems with the original specification of the EAV model. The model was specified informally, which could lead to different incompatible implementations. There was no agreed-upon exchange format or syntax. There was no mapping between EAV structures and MP pre-composed structures. Finally, the EAV employed a mix of a data model or schema and ontologies, but did not specify the interaction between the two. This meant that two different implementations could give wildly different answers to the same question.

### 2.2.9 Relationship between the relational model and ontology languages

Relational databases and ontologies are complementary paradigms for representing biological information. A conventional account is that relational schemas are for storing data and that ontologies are for representing knowledge. This is a useful guideline but in practice it can be difficult to apply

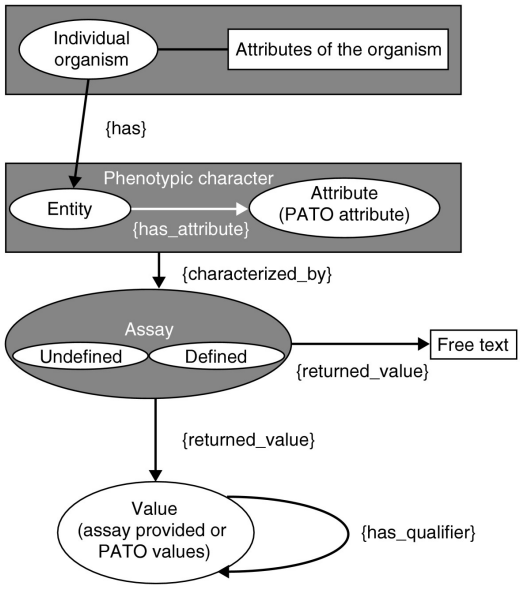


Figure 5: Proposed EAV schema from Gkoutos et al, 2005.

a hard and fast distinction between data and knowledge, and there is frequently confusion about whether to model at the relational schema level or the ontology level.

In fact relational databases and ontologies are both examples of formal structures that can be represented in first-order predicate logic (FOL). FOL can be used as a unifying framework One drawback is that FOL has certain computational properties such as undecidability that make it less attractive to use in practical next-generation information systems. However, there are subsets of FOL such as Horn logic (also known as logic programs) that have better computational properties. It is possible to map a fragment of description logics such as OWL to logic programs[Grosof et al., 2003], and logic programs largely subsume relational databases[Draxler, 1991]. Logic programs are therefore a reasonable candidate for providing a unified view over relational databases and ontologies, and this approach has been used successfully in the neurosciences[Gupta et al., 2003].

Logic Programs also subsume a class of formal grammars called *Definite Clause Grammars* (DCGs)[Clocksin and Mellish, 1981]. Formal grammars can be used to specify the generation and parsing of sequences of symbols from a collection of production rules. They have been used in fields as diverse as natural language processing and RNA secondary structure prediction[Holmes and Rubin, 2002]. In the context of the Chomsky hierarchy of expressivity[Chomsky, 1959]), DCGs are at least as expressive as Con-

text Free Grammars. Implementations include the XSB deductive database system[Sagonas et al., 1994] and SWI-Prolog[Wielemaker, 2003].

## 2.3 Visualization, Presentation, Knowledge Acquisition and Querying

### 2.3.1 Motivation

Once data has been structured according to database schemas and ontologies the next challenge is to present this data to researchers in meaningful ways, ways that allow them to hone in on data of interest and to detect correlations that would have been occluded had the data been in some other form. In some circumstances it is also desirable to give researchers the ability to create or modify annotations.

### 2.3.2 System architecture

In this day and age it is generally not acceptable for researchers to be forced to manually download all data onto their local machine before they can query it. Modern systems generally have some form of client-server architecture, with the bulk of the data residing on the server, and the researcher viewing a portion of this data on a client machine or application.

### 2.3.3 Traditional client-server

With a traditional setup, the user interacts with the data via a desktop application which they install and download on their local machine. This application takes care of communicating with a remote server using some protocol to read or write data.

This type of client-server setup has a number of disadvantages. There is an additional burden on the researcher, in that he or she must install and keep up to date their copy of the application, which discourages casual serendipitous browsing. There is also in general no way of going from a view of an entity such as a gene in one application to the corresponding view of the same gene in a different application.

An example of this setup in bioinformatics was the ACEDB system[Stein and Thierry-Mieg, 1999].

### 2.3.4 Web interfaces

The advent of the web fundamentally changed how researchers accessed data. Many biological databases could be accessed via a web-based interface that

did not require the installation of any desktop software beyond a standard web browser. With the standard web-based architecture, the servers would do the majority of the work, and generate static representations of the page that are sent to the web browser to be rendered. This resulted in a more static experience, but this was sufficient for many database interfaces.

The UCSC genome browser [Kent et al., 2002] is an example of a sophisticated visual web interface to complex data. The UCSC browser allows for scrolling and zooming along a genome sequence, with features such as exons and introns indicated via glyphs. Each user action results in the server rendering an image which is delivered to the web browser.

The development of Java applets allowed for a more interactive experience typically required for visually demanding applications such as genome viewers. Two early java applets for genome browsing were BioViews[Helt et al., 1998] and GeneScene[Lewis et al., 2000].

Applets were in some ways a reversion to pre-web style architectures, as they abandoned the hyperlink as the primary means of navigation. This and a number of other problems lead to them being used less commonly in modern applications.

### 2.3.5 Next-generation interactive web interfaces with AJAX

Historically desktop applications have been viewed as offering a richer experience with greater interactivity, whereas page-oriented web interfaces offered a more static experience. However, this view is now changing with the advent of richer web application technologies and frameworks such as *AJAX* (Asynchronous Javascript and XML), characterized by familiar web applications such as Google Maps and GMail.

### 2.3.6 Distributed architectures

In general it is not convenient for researchers to download large datasets to their local machine and browse the data there. There needs to be some kind of *distributed architecture*, involving protocols for client software to communicate with data servers. An example of such a protocol in genomics is the Distributed Annotation Server (DAS) protocol[Dowell et al., 2001], which is tailored for the retrieval of genomic features of interest within a given range. A more generic architecture is the Common Object Request Broker Architecture (CORBA)[Vinoski et al., 1997], which allows the exchange of structured data across a network.

## 2.4 Hypothesis Generation and Translational Research

### 2.4.1 Motivation

Integrating genomic data with other kinds of related experimental data in formal logical structures with advanced visualization and query interfaces provides researchers the ability to browse the data to find information they need.

We can go one step further and use data-mining techniques with next generation information systems to generate and explore hypotheses.

### 2.4.2 Identification of sequence variants involved in disease

The dbSNP[Sherry et al., 2001] database has over 12 million unique human sequence variants. Assaying every possible variant for association with diseases or adverse phenotypic affects is prohibitively expensive in terms of time, labor and reagents. A common strategy is to assay a subset of sequence variants which we have some evidence for disease association. For example, genome variants that change conserved amino acids are more likely to be disease-causing[Botstein and Risch, 2003].

The Online Mendelian Inheritance in Man (OMIM)[Hamosh et al., 2005] resource is a collection of 19,000 semi-structured records describing genes and inherited diseases. OMIM is a rich and well-curated resource, but unfortunately much of the valuable information is textual rather than structured. This applies to both the descriptions of the sequence variants and to the phenotypes themselves.

Model organisms provide a valuable source of candidate genes. Costello syndrome is a neuro-cardio-facio-cutaneous developmental syndrome resulting from mutations in the H-RAS gene. If mouse H-Ras is mutated in the orthologous position the phenotype recapitulates the disease[Schuhmacher et al., 2008]. Spontaneous models can be identified by comparing phenotypes in animals with human disease phenotypes - for example, the *fat aussie* mouse has phenotypes similar to human Alstrom syndrome which is caused by ALMS1 variants[Collin et al., 2002], and indeed further investigation showed that *fat aussie* was associated with mutations in Alms1[Arsov et al., 2006].

These examples for identifying animal models of disease relied on knowledge of the genetic basis of the human disease, but there are many human diseases for which it is not yet known. If a researcher could compare human, model organism and even ancestral phenotypes directly, they would have a mechanism to more rapidly identify candidate genes and models of disease.

### 2.4.3 Ontologies and data mining

Ontologies are key technologies for data mining. One technique is to use a subset of an ontology (sometimes called a *slim*)[Lomax, 2005]) and map all annotations to that subset. For example, on completion of functional annotation of a genome, map all genes into a set of pre-defined high-level categories.

Another technique is term enrichment[Boyle et al., 2004]. Here, a collection of annotated entities (such as genes up-regulated under certain experimental conditions) is tested for statistical enrichment across the whole ontology.

Ontologies can also be used to measure similarity between annotated entities - for example, comparing two genes based on their GO annotation profiles[Lord et al., 2003][Pesquita et al., 2007][Pesquita et al., 2009].

One limitation all these techniques have in common is that they have all been evaluated on the `GO`, and are not extensible to post-composition style annotation. This means that they can be used for comparing genotypes based on annotations to a pre-composed ontology such as `MP`, but not to annotations that use combinatorial models, such as the EAV model or OWL class expressions.

## 2.5 Background Summary

There are numerous technical challenges in the design and implementation of intelligent next-generation genomic information systems. These systems must be able to coordinate and integrate interdependent data transformation workflows across multiple connected processing units. The results must be integrated into flexible and expressive databases and knowledge bases that allow comparison across many different data types. Ontologies are key to data integration, but many ontologies are insufficiently developed or lack key features that allow them to be used for inference. Sophisticated data architectures will be required, as well as biologist-friendly knowledge acquisition and dissemination applications.

Finally, in order for the needs of translation research to be served, it will be necessary to develop and apply new analysis methods that use ontologies to perform data-mining across complex heterogeneous datasets spanning multiple species.

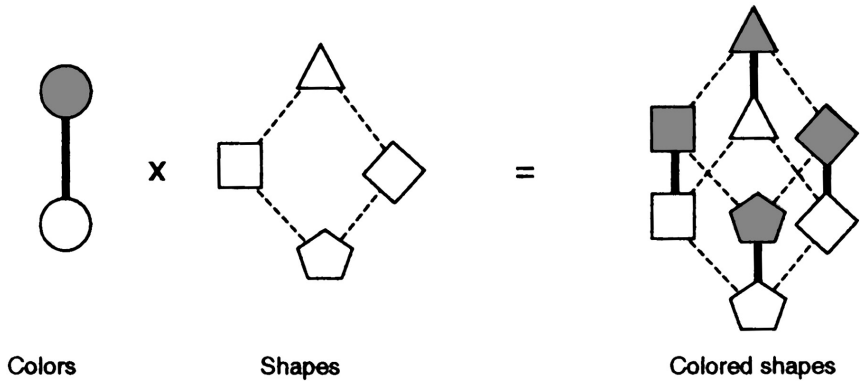


Figure 6: An example illustrating composition of nodes from graph cross-product. DAG cross-product example. In this example, a DAG whose nodes represent colors is crossed with a DAG whose nodes represent shapes. The result is a DAG whose nodes are colored shapes. Every combination is represented, so there are eight nodes in the result. From Hill et al 2002.

## 2.6 Historical Context

Many of the results described in thesis related directly or indirectly to knowledge representation and in particular, ontologies. Biologists and bioinformaticians have been rapidly developing and using ontologies over the past decade, this adoption has been ad-hoc, and has largely ignored developments made my computer scientists working in isolation biologist users. Many of these fledgling bio-ontologies made use of the same simple structure and file formats as the GO (see section 2.2.4), rather than the more advanced description logic languages and accompanying tools. One reason for this situation may be that the perceived difficulty in using these tools, whereas there was already a community of biologists making use of the GO. Coupled with the runaway success of the GO, and the perceived lack of demand for more advanced tooling, this perhaps led many biologists to ignore developments in theoretic and computational aspects of ontology engineering.

In fact, the simple data structures used for these ontologies turned out to be a technological "local minimum". For example, early on in the development of the GO, bioinformaticians realized that development of the ontology was sustainable only if algorithmic techniques were adopted to help automate construction of the ontology - the so-called "cross-product" approach[Hill et al., 2002]. The abstract idea is captured in figure 6 which shows the composition of colored shapes from orthogonal color and shape graphs, and the example below shows a subset of the cross-product obtained by combining an elemental *developmental* graph with an elemental *anatomical* graph:

**(Example 3)**

$\{development, morphogenesis, growth\} \times \{lung, brain, heart, bone, limb...\}$

In fact the approach was in part an independent re-invention of Formal Concept Analysis (FCA) and Description Logic class-intersection constructs. For example, the grey triangle at the top right corner of 6 could be represented by a DL expression:

$$GreyTriangle \equiv Triangle \sqcap Grey$$

Furthermore, description logic reasoning procedures can be used to automatically generate the full graph subsumption hierarchy. Unfortunately, the GO researchers were not immediately aware of this branch of computer science, and the computer scientists were not immediately aware that the biologists had a problem they could help with (although this was later pointed out, e.g. in [Wroe et al., 2003]).

Part of the challenge in engineering and using ontologies effectively is understanding enough of the biological problem and the computer science state-of-the art to bring the two together effectively. This is the background for many of the results presented in the forthcoming sections.

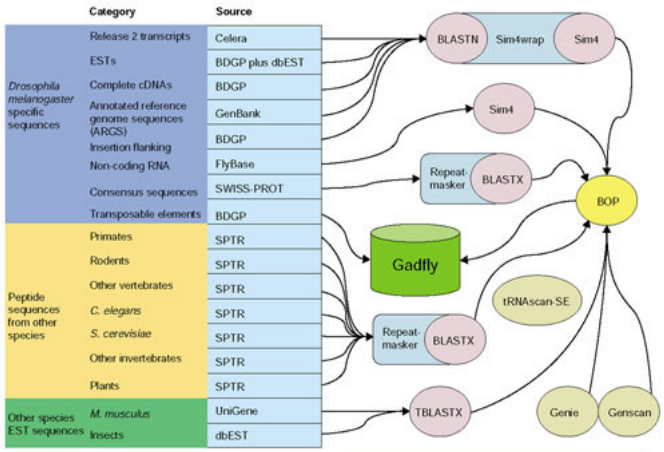


Figure 7: `GadFly` Analysis Pipeline for release 3 of *Drosophila* genome. Reproduced from Mungall et al, 2002

# 3 Systems for the Analysis of Genomic Sequences

## 3.1 An integrated computational pipeline and database to support whole-genome sequence annotation

We created the `GadFly` automated annotation system primarily for the analysis and re-annotation of the *Drosophila* genome[Mungall et al., 2002]*[1]. The input for the system is a collection of genome sequences (either whole chromosome arms or smaller assembled regions), external sequence data (genomic, transcriptomic or proteomic, in the same species or other species, both closely and distantly related) and a configuration file. The output is a collection of genome annotations collated and synthesized from the runs of a number of gene prediction and sequence similarity tools. Figure 7 shows the data sources and analyses used in the annotation pipeline.

`GadFly` incorporated a number of innovative features:

1. One of the first general-purpose configurable bioinformatics pipelines to exploit Beowulf clusters to run multiple jobs in parallel on multiple processors

2. highly configurable and allows biologists to configure workflows through a workflow specification language

[1]Peer reviewed publication in *Genome Biology*

3. the results from multiple individual analyses could be combined using an algorithm called to *Autopromote* which combines multiple analyses using voting networks into a single coherent gene models with alternately spliced transcripts.

The `GadFly` system and successor systems was used for a number of annotation projects, described in the rest of this section.

## 3.2 Annotation of the Drosophila melanogaster euchromatic genome: a systematic review

The `GadFly` system was used in the analysis and re-annotation of The *Drosophila melanogaster* genome[Misra et al., 2002]*[2]. The re-annotation yielded large benefits in improved gene models, finding additional exons, UTRs and spliceforms in the majority of genes, splits and merges of genes models, as well as entirely new protein coding genes.

My contributions included selection and configuration of analysis software used, setting of curator rules, and investigation of complex gene models. The latter includes nested, trans-spliced and dicistronic genes, as well as cases of tandemly repeated overlapping genes.

## 3.3 Heterochromatic sequences in a Drosophila whole-genome shotgun assembly

In attempting to apply the same methods to heterochromatin, we discovered that gene finders and alignment programs failed to find the outermost exons of many heterochromatin genes, even after masking out repetitive regions because of the uncharacteristically long introns.

Our solution was to go beyond simple repeat-masking. We excised and condensed the repetitive elements to fixed-length sequences, resulting in dramatically improved alignments and gene predictions, illustrated with genes such as *rolled* gene[Hoskins et al., 2002].

Heterochromatin is difficult to assemble due to the high levels of repetitive sequence. It took a number of years for us to finally finish the sequence and publish the full analysis[Smith et al., 2007a]*[3].

[2]Peer reviewed paper in *Genome Biology*. My contribution was in the automated genome analysis and in-silico experimental design

[3]Peer reviewed paper in *Science*. I devised the system used for data analysis and data management, and carried out the Gene Ontology analysis

## 3.4 Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome

We performed a comparative analysis, looking in depth at 8 genomic regions from 4 *Drosophila* species[Bergman et al., 2002]. This required an extension of the `GadFly` system to perform a TBLASTN analysis of the other genomes against the newly annotated *Drosophila melanogaster* peptide sequences.

The results revealed that the gene models were highly conserved, which helped us improve the exon-intron annotations in *Drosophila melanogaster*. The exon-intron structures themselves are highly conserved, with only one case of intron gain.

## 3.5 Large-scale trends in the evolution of gene structures within 11 animal genomes

We had previously performed a protein-centric comparative analysis of 3 model organisms and human[Rubin et al., 2000]. Using some of the tools developed for the analysis of *Drosophila* we analyzed 11 animal genes, this time focusing on the change in exon-intron structure over evolutionary time, and the correlation with protein sequence changes[Yandell et al., 2006]*[4]. This work involved extracting genome annotation data from GenBank, much of which was inconsistent or missing data. We developed a tool for inferring implicit genome annotations from GenBank records, which we contributed back to the open source BioPerl project[Stajich et al., 2002].

Our results indicated that change in exon-intron structure is gradual and largely independent of protein sequence evolution. This indicates that gene structures can be used as a way of exploring deep homology.

Figure 8 shows correlation in orthologous intron lengths between *Drosophila melanogaster* and 5 other insect species. With 5 million years until the last common ancestor, intron lengths are highly correlated. As we move back in time until 63 million years ago (the most distantly related Drosophila species in the set) we see intron lengths becoming less correlated. At 250 million years ago the correlation is lost, as can be seen in the comparison with the malaria mosquito, *Anopheles gambiae*.

My contribution to this work included the database analysis and development of software libraries. This research also made use of software developed in tandem with a new relational database system called *Chado*.

[4]Peer-reviewed paper in *PLoS Computational Biology*. I contributed to the experimental design, and I devised and implemented the software and analysis pipeline

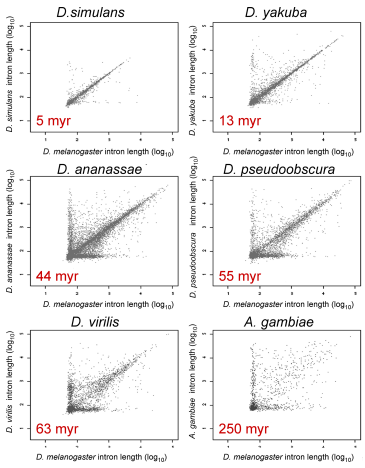


Figure 8: Intron lengths and evolution. Taken from Yandell, Mungall et al, 2006

# 4 Formal Structures for Representing Biology

## 4.1 Chado: An ontology-based modular schema for representing biological information

We developed a modular ontology-oriented relational database schema called Chado[Mungall et al., 2007a]*[5]. This schema was designed as a standard model to be used across multiple projects and application tools spanning multiple species.

This schema was innovative in a number of respects:

1. *Modular Design.* Instead of being a monolithic system, Chado used inter-related schema modules, with each modules representing a different sub-domain of biology. The core module is the *sequence* module for representing genes and genomic features.

2. *Formal graph-theoretic structures* for representing the relationships that hold between genomic entities and their derivatives (see figs 9 and 10).

3. *Hybrid relational-ontology design.* Chado has an extensible, flexible under-constrained relational structure and uses ontologies for the bulk of the modeling, making Chado a hybrid relational-ontology system.

[5]Peer-reviewed paper in *Bioinformatics*

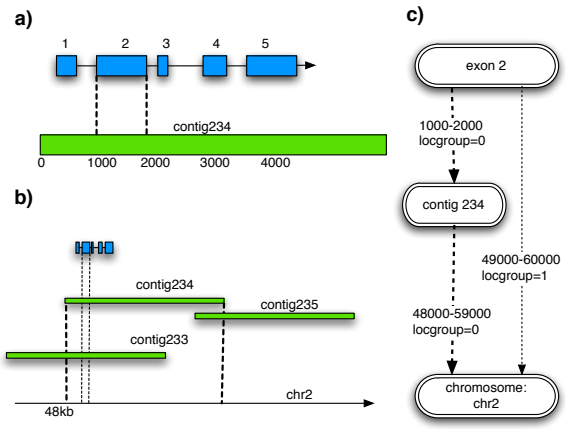


Figure 9: Example Chado Location Graph. Left panel shows visual depiction of gene location on contig. Right panel shows the LG diagrammatically. Inferred edges are indicated with dashed lines. Reproduced from Mungall et al, 2007

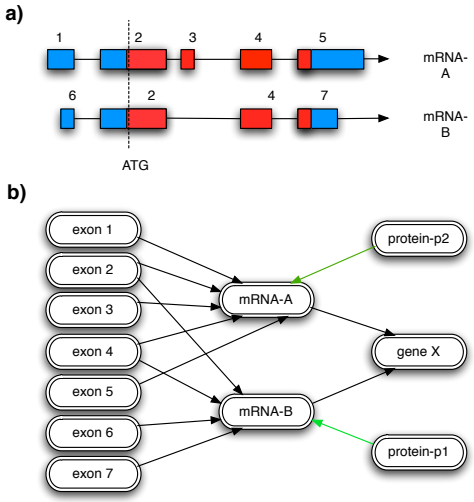


Figure 10: Example Chado Feature Graph. (a) visual depiction of gene model with alternately spliced gene. (b) Feature Graph. Reproduced from Mungall et al, 2007

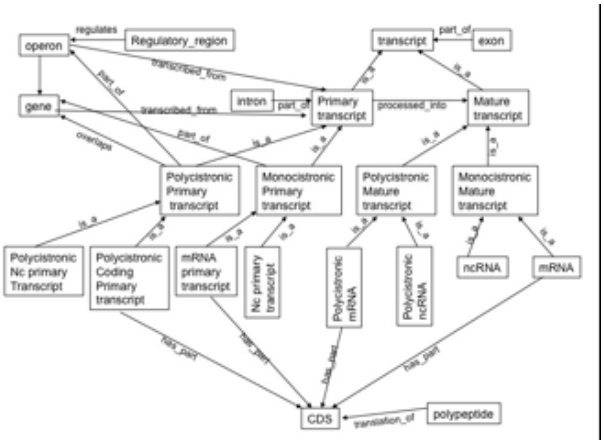


Figure 11: A section of the Sequence Ontology showing how terms and relationships are used together to describe knowledge about sequence. Reproduced from Mungall *Journal of Biomedical Informatics* [SO] 2010.

At the core of Chado is the *sequence* module, for representing genomic sequence features, as well as functional annotations made using the Gene Ontology (GO). Chado also includes modules for *expression* data, *phenotypic* data, *phylogenetic* data as well as others contributed by members of the wider community.

The Chado sequence model is deliberately minimal, and delegates the representation of feature types and relationships to an ontology.

## 4.2 The Sequence Ontology

We developed the Sequence Ontology (SO), a standard collection of feature types together with the relationships that can hold between them[Eilbeck et al., 2005]*[6].

The SO is a formal structure consisting of a class hierarchy of feature types, together with the relationships that hold between instances of these types (see figure 11). The SO is intended to be use as a type system in the Chado database, as well as in exchange formats such as GFF3, thus avoiding the rigidity associated with typing at the schema level. The SO also serves as a source of terminological information, with definitions and synonyms for each of the major feature types.

The original incarnation of SO combined both molecules and sequences, sometimes confusing the two. The ontology was structured primarily along

[6]Peer-reviewed paper in *Genome Biology*. I contributed to the design of the ontology

A

```
exon_transcript_order(X1,T,R),
exon_transcript_order(X2,T,R+1),
exon_dnaseq_pos(X1,Seq,_,X1End,Str),
exon_dnaseq_pos(X2,Seq,X2Beg,_,Str)
->
intron(I),
intron_dnaseq_pos(I,Seq,X2Beg,X1End,Str)
```

B

```
stemloop(SL1,S1,L1),
stemloop(SL2,S2,L2),
overlaps(L1,S2),
overlaps(L2,S1)
->
pseudoknot(X),
X = SL1+SL2
```

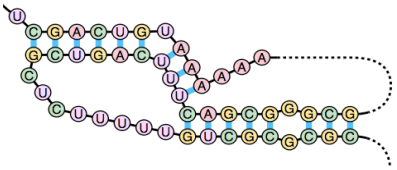

Figure 12: (A) Rule for inference of existence of intron and its genomic position based on consecutive order of exons (B) rule for the inference of pseudoknots based on connectivity relationships as defined in RNAO. Taken from Mungall 2009 (LNCS)

mereological ground, using the *part_of* relation. In order to better characterize genomic features we enhanced the ontology, adding new relations, including genomic relations based on the Allen Interval Algebra[Mungall et al., 2010a]*[7]. As part of this work we also clearly delineated one-dimensional sequences from their corresponding molecules and their properties, leading in part to the formalization of an ontology of RNA molecules[Batchelor et al., 2009][8].

Many of the axioms we would like to add go beyond what can be expressed in ontology languages such as OBO or OWL. For example, given two exons that are transcribed in succession (using a relation such as exon_transcript, as defined in the Ensembl datamodel), we should be able to infer the existence of an intron in between these two exons (see figure 12A). Logic programs are an alternative basis for expressing these kinds of rules[Mungall , 2009][9]. These rules can be extended from linear DNA sequences to RNA secondary structures, as shown in figure 12B.

[7]Peer-reviewed paper in *Journal of Biomedical Informatics*, in press

[8]Published as conference proceedings, currently being expanded into a full journal article

[9]Conference proceedings, published in *Lecture Notes in Computer Science*

| Relation | Transitive | Symmetric | Reflexive | Antisymmetric |
|---|---|---|---|---|
| *is_a* | + | - | + | + |
| *part_of* | + | - | + | + |
| *located_in* | + | - | + | - |
| *contained_in* | - | - | - | - |
| *adjacent_to* | - | - | - | - |
| *transformation_of* | + | - | - | - |
| *derives_from* | + | - | - | - |
| *preceded_by* | + | - | - | - |
| *has_participant* | - | - | - | - |
| *has_agent* | - | - | - | - |

Table 1: Some properties of type-level relations in the OBO Relation Ontology. Note that type-level adjacency is non-symmetric. Taken from Smith et al, 2005.

## 4.3 Relations in Biomedical Ontologies

One problem with the many nascent biological ontologies was informal and inconsistent use of relations such as *is_a* and *part_of*. We created a relation ontology (RO) to promote interoperability of ontologies and to support automated reasoning[Smith et al., 2005b]*[10].

Each relation can have formal properties such as transitivity and reflexivity (see table 1).

The RO makes a formal distinction between type-level and instance-level relations. Type-level relations are defined in terms of the instance-level ones, and type-level relations may have different formal properties from their instance level counterparts. For example, adjacency is symmetric on the instance level, but not on the type-level (consider: every nucleus is adjacent to some cytoplasm, but not every cytoplasm is adjacent to some nucleus).

One important feature of the RO is its treatment of time, an aspect missing in many description logic treatments of biological ontologies.

## 4.4 Obol: Integrating Language and Meaning in Bio-Ontologies

Just as a biological sequence contains latent meaning that can be elucidated by pattern matching, terms from ontologies such as the GO contain patterns that can be extracted computationally. The `Obol` grammar and inference

[10]Peer-reviewed publication in *Genome Biology*. I developed the initial version of the ontology

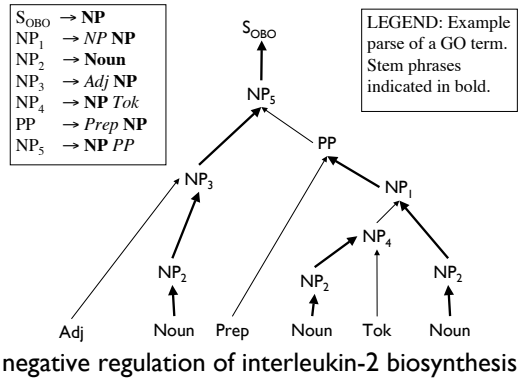


Figure 13: Obol parse of the term *negative regulation of interleukin-2 biosynthesis* using a context free grammar with 5 production rules. Taken from Mungall, 2004

system was devised to extract latent meaning hidden in terminology-oriented ontologies such as the GO[Mungall , 2004]*[11].

Obol is implemented in Prolog and uses Definite Clause Grammars to encode production rules for parsing opaque terms such as *negative regulation of interleukin-2 biosynthesis* into description logic expressions that reveal hidden semantics (see figure 13). Obol also includes rules for reasoning over these structures, using properties of relations defined in the RO. For example, automatic inference of the subsumption relationship between *negative regulation of interleukin-2 biosynthesis* and *regulation of cytokine biosynthesis*.

Obol presented an advance on previous techniques which relied on *regular grammars* to do similar tasks. Whist regular expressions have the advantage of being easily implemented in programming languages such as Perl, they lack the expressive power to fully capture the nested linguistic expressions used in GO terms. Obol DCGs are more expressive, at least as expressive as *Context Free Grammars* (CFGs).

Obol proved successful at determining the correct structure of many of the terms in GO. Combined with its built-in reasoning system, Obol was able to suggest thousands of new links that were subsequently added to the GO, as well as detecting many errors in the ontology structure or ontology definitions.

[11]Conference paper from Bio-Ontologies 2004 subsequently published in *Comparative and Functional Genomics*. Whilst this paper was peer-reviewed, it is classified as a conference paper and thus may not count as a full peer-reviewed journal article

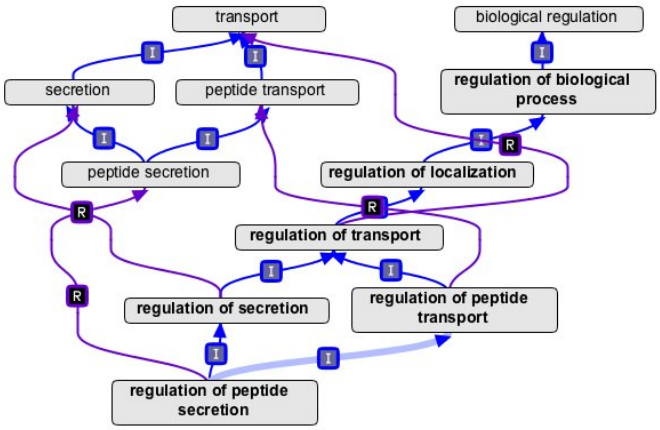


Figure 14: Use of Gene Ontology Cross-Product definitions for reasoning. Bold line between *regulation of peptide secretion* and *regulation of peptide transport* is inferred by a reasoner. Taken from Mungall *Journal of Biomedical Informatics* [GO] 2010

More recently, Obol has been extended to work with OWL ontologies and a wider range of logical class expressions[Vassiliadis et al., 2009][12].

## 4.5 Cross-product extensions of the Gene Ontology

We have used a combination of `Obol` and manual curation to generate logical definitions for 41% of the terms in the GO[Mungall et al., 2010b]*[13]. These logical definitions have been partitioned into mutually exclusive sets called cross-products (XPs). Each XP is a subset of the cross-product between one of the 3 GO ontologies and another ontology. For example, the term *oocyte differentiation* is formally defined using the parent class *cell differentiation* from the GO biological process ontology and discriminating characteristics that reference *oocyte* in the Cell Ontology (table 2). This term therefore falls into the $BP \times CL$ XP set.

The combination of these logical definitions and the relationships from external ontologies allow us to use automated reasoners to find the answers to biological questions and to automatically classify the GO. For example, we can infer the *oocyte differentiation* should be an *is_a* child of *germ cell differentiation*. Using this strategy we identified and fixed over 2000 links in the GO. An example of one such fix is shown in figure 14.

[12]Published as conference proceedings. I wrote the manuscript and supervised the work
[13]Peer-reviewed paper in *Journal of Biomedical Informatics*

| XP Set | Size | Example | Def |
|---|---|---|---|
| BP × BP | 606 | S phase of mitotic cell cycle | S phase AND *part_of* mitosis |
| BP × BP (regulates) | 3529 | Regulation of neuroblast proliferation | biological regulation AND regulates neuroblast proliferation |
| BP × BP (multi-organism) | 374 | modulation of intracellular transport in other organism during symbiotic interaction | interspecies interaction between organisms AND regulates intracellular transport AND during symbiosis AND regulates_process_in external organism |
| BP × MF (regulates) | 201 | Regulation of protein kinase activity | biological regulation AND regulates protein kinase activity |
| BP × CC | 476 | Mitochondrial translation | translation AND occurs_in mitochondrion |
| CC × CC | 682 | Acrosomal membrane | membrane AND surrounds acrosome |
| CC × MF | 173 | histone deacetylase complex | protein complex AND has_function histone deacetylase activity |
| MF × MF (regulates) | 104 | Lipase activator activity | molecular function AND regulates lipase activity |
| MF × CC | 48 | Microtubule motor activity | motor activity AND results_in_movement_along microtubule |
| BP × CL | 544 | Oocyte differentiation | cell differentiation AND results_in_acquisition_of_features_of oocyte |
| BP × Uberon | 583 | Neural plate formation | anatomical structure formation AND results_in_formation_of neural plate |
| BP × PATO | 31 | Regulation of cell volume | biological regulation AND regulates (volume AND quality_of cell) |
| MF × Uberon | 9 | Structural constituent of bone | structural molecule activity AND inheres_in bone |
| CC × CL | 28 | Neuron projection | cell projection AND part_of neuron |
| BP × CHEBI | 3077 | L-cysteine catabolic process to taurine | catabolic process AND has_input L-cysteine AND has_output taurine |
| MF × CHEBI | 315 | nitrate reductase activity | oxidoreductase activity AND reduces nitrate |
| BP × *PRO* | 37 | Interleukin-1 biosynthesis | biosynthetic process AND has_output interleukin-1 |

Table 2: Gene Ontology cross-products. The set of all equivalence axioms is divided into individual XP sets denoted $A \times B$. The number of equivalence axioms in each set is show, together with an example equivalence between GO term and intersection expression. Taken from Mungall *Journal of Biomedical Informatics* [GO] 2010

These logical definitions can also be used to automatically align with pathway databases such as Reactome. For example, molecular function classes can automatically be matched to the reactions they subsume based on the reaction participants in CHEBI.

This work is part of a larger effort to apply engineering techniques on the Gene Ontology[Alterovitz et al., 2010].

In many cases we found that the referenced ontology was incorrect and the links in the GO were correct. We developed an abductive inference technique for detecting these[Bada et al., 2008][14].

This work demonstrated the importance for GO in coordinating with a set of orthogonal external ontologies.

## 4.6 The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration

The GO was constructed to serve the needs of functional annotation – assigning a gene to categories based on what the gene function is and where it is localized within the cell when it executes its function. However, the GO was not intended for representing other biological phenomena, such as where in the organism a gene is expressed, or what happens to an organism when a gene is mutated. This requires other ontologies.

The development of these other ontologies had to be coordinated such that they were mutually consistent. The Open Biological Ontologies (OBO) library[Ashburner et al., 2003][15] and later on, the OBO Foundry, were created in order to foster the development of these ontologies[Smith et al., 2007b]*[16].

The activities of the OBO Foundry include the publication of a set of best practices[17], including standardized naming conventions[Schober et al., 2009].

## 4.7 A Common Anatomy Reference Ontology

We developed a common anatomy reference ontology (`CARO`) to serve as a standard framework for all anatomical ontologies[Haendel et al., 2007][18].

[14]Published as conference proceedings

[15]Conference paper

[16]Peer-reviewed publication *Nature Biotechnology*. I played a key role in the creation of these libraries, developing the technology, infrastructure and contributing to the core principles

[17]http://obofoundry.org

[18]Book chapter

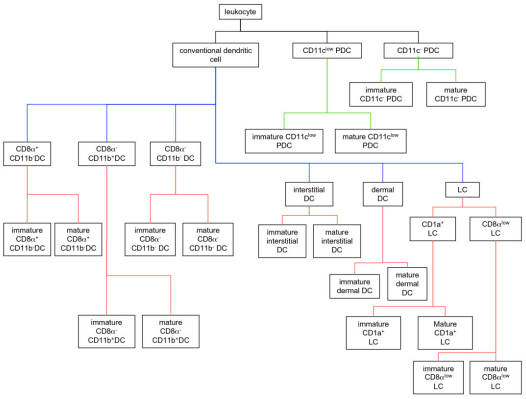


Figure 15: Subset of the Dendritic Cell Ontology. Taken from Masci et al, 2009. DC, dendritic cell; PDC, plasmacytoid dendritic cell; and LC, Langerhans Cell

CARO employs a jointly-exhaustive pairwise-disjoint hierarchy classified according to structural criteria. CARO is a small ontology consisting of only 46 very general classes such as "epithelium" and "cavitated compound organ", but lacks classes for specific cells, organs or systems such as "immune cell", "heart", "liver", "immune system" "eye" and so on. Instead, these types of entity are intended to be represented in more specific ontologies such as cell type ontologies or species-specific gross anatomical ontologies.

## 4.8 Hematopoietic Cell Types: Prototype for a Revised Cell Ontology

We developed an ontology of dendritic cells[Masci et al., 2009] (DC-CL) classified according to surface protein expression. This required the creation of new mereological relations defined using the Gene Ontology cell component ontology in order to specify the properties of cells in a compact way.

Figure 15 shows a subset of the DC-CL ontology. Table 3 shows a subset of logical definitions from the ontology.

The methods described here were broadened and applied to hematopoetic cells as a whole, leading to an overall restructuring of the Cell Ontology[Diehl et al., 2010].

| **Class** | **Genus** | **Differentia** | |
|---|---|---|---|
| conventional dendritic cell | leukocyte$_{\mathrm{CL}}$ | *has_high_plasma_membrane_amount* | CD11c$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD3$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD19$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD34$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD56$_{\mathrm{PRO}}$ |
| CD11c$^{\mathrm{low}}$ plasmacytoid dendritic cell | leukocyte$_{\mathrm{CL}}$ | *has_low_plasma_membrane_amount* | CD11c$_{\mathrm{PRO}}$ |
| | | *has_plasma_membrane_part* | CD45R$_{\mathrm{PRO}}$ |
| | | *has_plasma_membrane_part* | GR1$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD11b$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD3$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD19$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD34$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD56$_{\mathrm{PRO}}$ |
| CD11c-plasmacytoid dendritic cell | leukocyte$_{\mathrm{CL}}$ | *has_plasma_membrane_part* | CD45RA$_{\mathrm{PRO}}$ |
| | | *has_plasma_membrane_part* | CD123$_{\mathrm{PRO}}$ |
| | | *has_plasma_membrane_part* | CD303$_{\mathrm{PRO}}$ |
| | | *has_plasma_membrane_part* | ILT7$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD11c$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD3$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD19$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD34$_{\mathrm{PRO}}$ |
| | | *lacks_plasma_membrane_part* | CD56$_{\mathrm{PRO}}$ |

Table 3: Logical definitions for the three most general types in DC-CL. Adapted from Masci et al 2009

## 4.9 Representing Phenotypes in OWL

The original EAV database model for describing phenotypes lacked any formal semantics. To remedy this we proposed a way of representing phenotypes with `PATO` and OBO ontologies using a Description Logic language such as OWL-DL[Mungall et al., 2007b][19]. This retains the expressivity of the EAV model, but within a single unified framework.

Each phenotype is represented as a class expression using set-theoretic operators such as intersection and union. Most phenotype descriptions are composed using an intersection between a named `PATO` class and a relational expression formed using the formal *inheres_in* relation (table 4).

[19]Conference proceedings, non peer-reviewed

| Phenotype | DL representation |
|---|---|
| curved wing | Curved THAT *inheres_in* SOME Wing |
| photosensitivity | Sensitive THAT *towards* SOME UltravioletLight AND *inheres_in* SOME Skin |
| high permeability of mitochondrial cristae in axons of CA1 pyramidal cells | HighPermeability THAT inheresIn SOME (MitochondrialCristae THAT *part_of* SOME (Axon THAT partOf SOME PyramidalCell THAT *part_of* SOME CA1Field)) |

Table 4: Representing phenotypes as OWL-DL descriptions. OWL expressions are written using Manchester Syntax. Adapted from Mungall et al 2007.

# 5 Tools for Visualisation and Knowledge Acquisition

## 5.1 Web-based architectures for visualising genetic and genomic data

Once information has been structured according to relational schemas and ontologies it should be made accessible to researchers in a way that allows complex queries and visualisation of complex data.

One common way to present information in a database to users is through a web-based interface. The development of web interfaces can be simplified and accelerated through the use of templating systems. One such system is WebInTool[Hu et al., 1996], which is a component of the Anubis interface to the ArkDB family of genetic mapping databases[Hu et al., 2001]. Anubis is a web-based graphical interface for visualizing markers on linkage, radiation hybrid and cytogenetic maps. This interface was innovative in being one of the first interactive web-enabled graphic map browsers in biology, and also in the use of distributed architecture and HTTP-based API, similar to the DAS protocol which followed a few years later. This was later extended to use CORBA[Hu et al., 1998] with a Java front-end. CORBA had the advantage of being industry-standard, but proved far more difficult to implement.

The web-based visualization paradigm also proved useful for genomic sequence feature data, with development culminating in the Generic Genome Browser (GBrowse)[Stein et al., 2002]. GBrowse works in conjunction with Chado (or its predecessor `GadFly`) or other genome databases.

The GBrowse architecture is shown in figure 16. Figure 17 shows a screen-

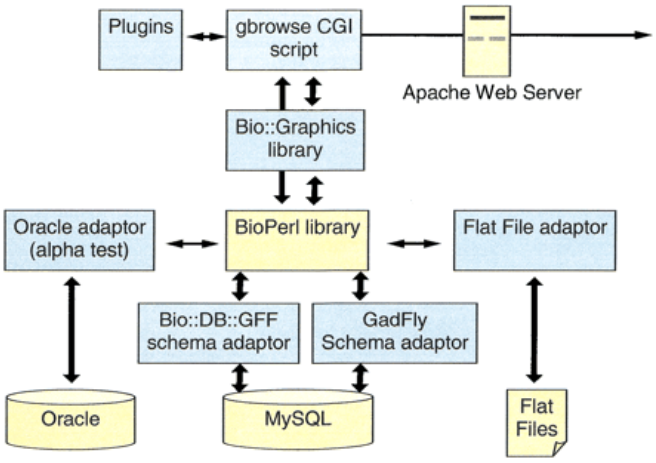


Figure 16: GBrowse Architecture, including `GadFly` database. Reproduced from Stein et al, 2002

shot of a region of the *C elegans* genome in GBrowse.

One of the limitations imposed on GBrowse by the technology of the time was the delivery of static images to the browser. We implemented a successor to GBrowse called JBrowse using Asynchronous Java and XML (AJAX) technology in order to deliver greater interactivity[Skinner et al., 2009].

## 5.2 AmiGO: online access to ontology and annotation data

We developed a web application called AmiGO that allows users to query, browse, and visualize ontologies and related gene product annotation data[Carbon et al., 2008]. As well as browsing and querying (see figure 19), AmiGO also offers term enrichment and an annotation slimmer tool.

This application is constructed on top of the `GO` Database[Harris et al., 2004] and forms a critical component of the Gene Ontology online resource[Ashburner et al., 2001].

## 5.3 Knowledge Acquisition Tools

We have developed a suite of user interfaces intended for the acquisition of complex knowledge from experts. These are all implemented as Java desktop applications.

Apollo[Lewis et al., 2002] is a sequence annotation editor which was constructed initially for and used extensively in the annotation of the *Drosophila*

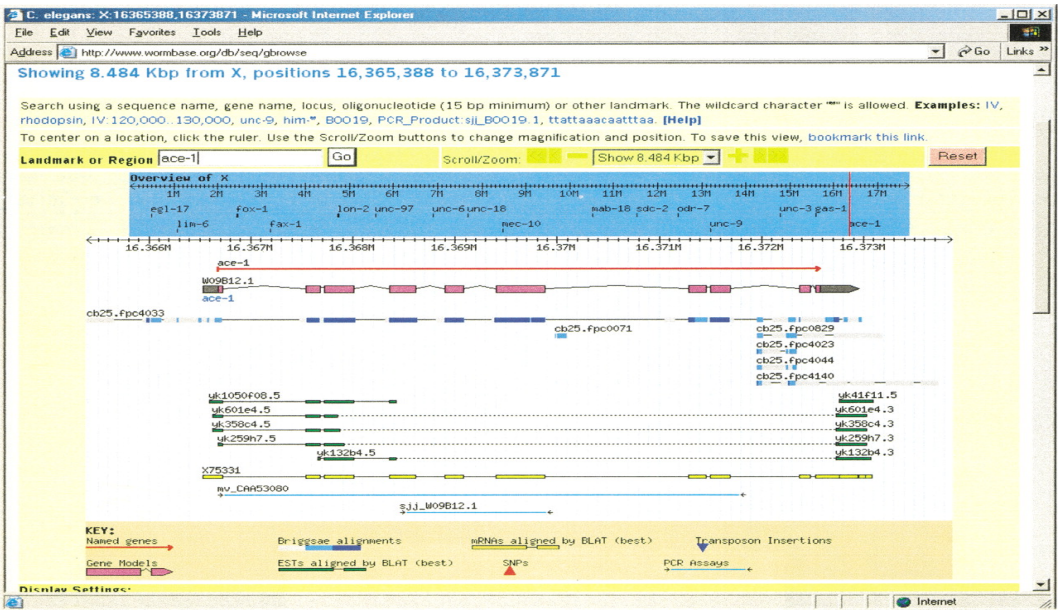


Figure 17: GBrowse Screenshot, from Stein et al, 2002

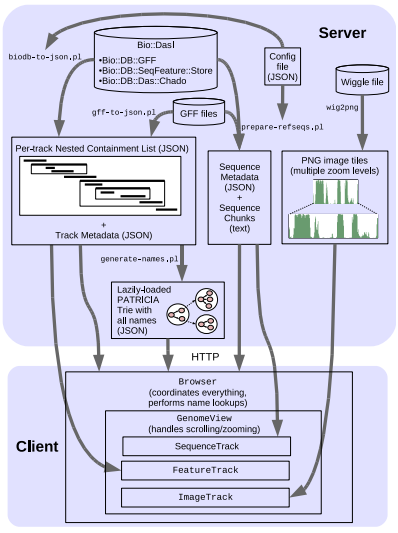


Figure 18: JBrowse architecture, from Skinner et al, 2002. Feature data from databases such as Chado are converted to JSON NCLists. These are delivered to the browser, where they are rendered client-side using a JavaScript module *GenomeView.js*

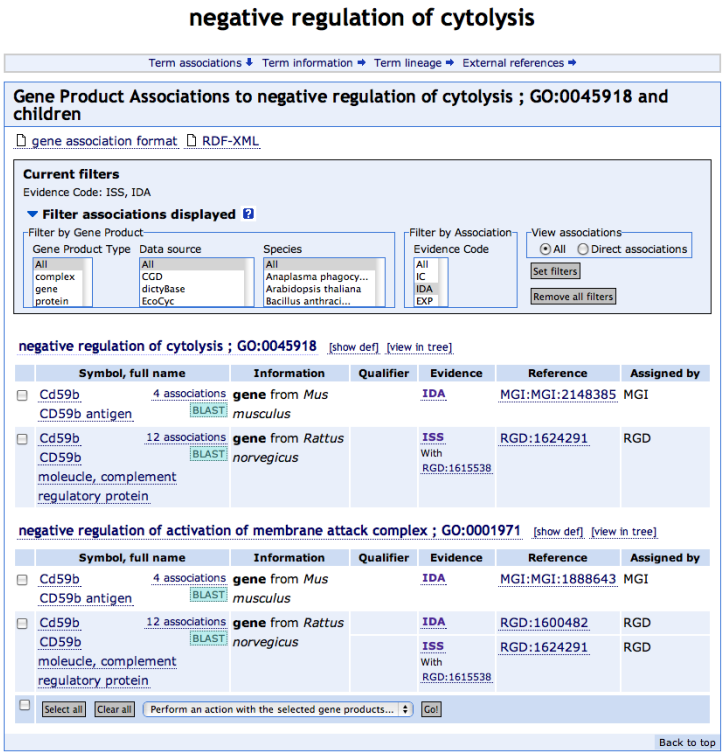


Figure 19: AmiGO screenshot showing genes annotated to *negative regulation of cytolysis*. From Carbon et al, 2008

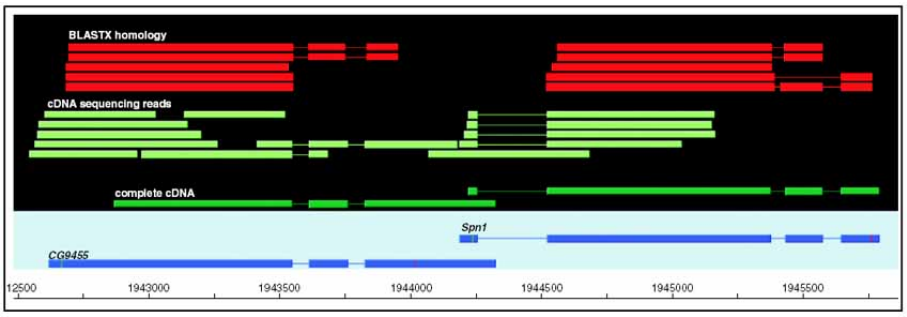


Figure 20: 3' UTR of CG9455 overlaps downstream Spn1 gene. Both genes have Gene Ontology annotations to *serine protease activity*. Figure from Misra et al 2002, the screenshot taken from Apollo (Lewis et al 2002) shows experimental evidence from cDNA alignments to the genome, stored in `GadFly`.

*melanogaster* genome (figure 20).

OBO-Edit is a graph-oriented ontology editor[Day-Richter et al., 2007], constructed initially for the Gene Ontology, but now used for the majority of ontologies in the OBO library.

# 6 Applications of Next-Generation Information Systems in Translational Research

## 6.1 GO analysis of Plasmodium

We applied tools developed for the GO Database for functional annotation analysis in the malaria parasite, *Plasmodium falciparum*[Gardner et al., 2002][20]. These tools used the graph structure of the ontology to "slim" annotations, in order to provide a high-level whole-genome overview; annotations are propagated up the graph to a predetermined level, and counted, where they can easily be plotted in a bar chart (see Figure 21).

For example, annotations to the specific term *protein import* (GO:0017038) and counted when providing a summary for the high level term *transport*. When all the GO annotations in Plasmodium were summed in this way and compared to Yeast, it showed an over-representation of genes in the cell adhesion and cell invasion categories. This correlates with what is already known about the Plasmodium lifecycle, and illustrates the utility of the GO and in particular inferences over the GO graph.

The GO analysis here is limited in that the high-level categories of interest had to be pre-determined, and there was no estimate of significance of the enrichment in these categories.

## 6.2 Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins

In mining the human genome for disease-causing sequence variants we are faced with an abundance of potential candidates (dbSNP has over 12 million unique sequence variants). We hypothesized that if a sequence variant is disease-causing, then a variant at the corresponding position in the paralogous protein is also likely to be disease-causing. This could allow for improved sequence variant candidate selection.

To explore this hypothesis we systematically examined the genome-wide distribution of sequence variants along the lengths of paralogous proteins[Yandell et al., 2008]*[21]. In order to do this, we created a database of human disease genes together with known variants and disease annotations on those variants. This database employed inferencing techniques to derive amino acid variation information and class (synonymous, non-synonymous, conservative,

[20]Peer reviewed paper in *Nature*. My contribution was the GO analysis.

[21]Peer-reviewed in *PLoS Computational Biology*

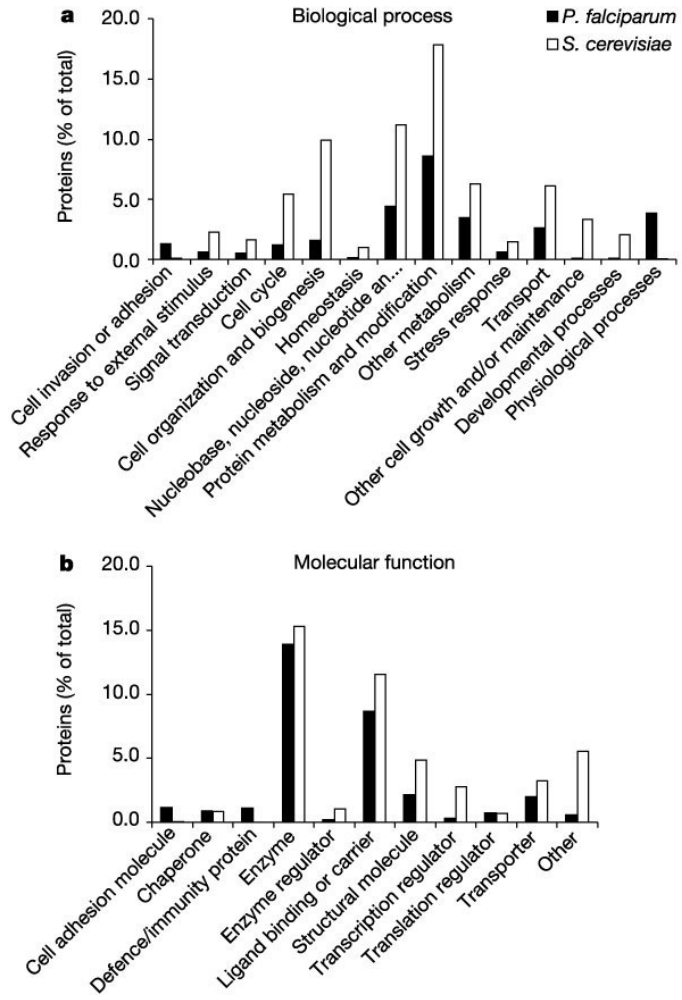


Figure 21: classification of *P. falciparum* genes using Gene Ontology. Reproduced from Gardner et al, 2002

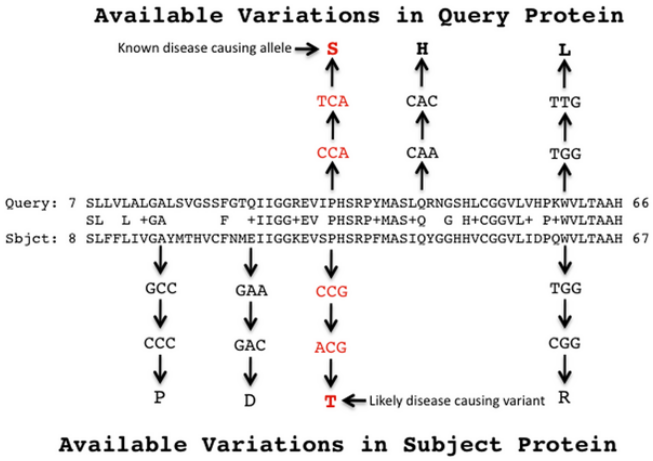


Figure 22: Using sequence similarity to identify variant pairs (taken from Yandell et al, 2008)

| Gene A | Variant A | Disease A | Gene B | Variant B | Disease B |
|---|---|---|---|---|---|
| FGFR2 | CX972741 | Pfeiffer syndrome | FGFR3 | CM950470 | Thanatophoric dysplasia |
| JAG1 | CD993777 | Alagille syndrime | FBN1 | CM972811 | Marfan syndrome |
| ATP7A | CM940140 | Menkes syndrome | ATP7B | CM970138 | Wilson disease |
| ABCA1 | CM993803 | Tangier disease | ABCA4 | CM990025 | Stargardt disease |
| CFTR | CM940275 | Cystic fibrosis | ABCC8 | CM981883 | Hyperinsulism |

Table 5: Examples of paralogous pairings of alleles in known disease genes. Cols 1 and 4 give the gene symbols for two paralogous disease-causing genes. Cols 2 and 5 give the HGMD IDs of the two variants that comprise the Class 1 pair. Columns 3 and 6 list the diseases most commonly associated with the two paralogous variants. Taken from Yandell et al, 2008

non-conservative). We also developed methods to match phenotypically characterized genotypes in OMIM with dbSNP variants. We then used BLASTP to derive candidate homologous pairs and to map variations from the known disease gene to candidate gene (see figure 22). Our results showed that disease causing genes preferentially align with each other and that on average, choosing uncharacterized variations aligned to known disease variants will enrich 6.6-fold for clinically significant variations.

Table 5 shows some examples of paralogous pairings of known disease genes. The diseases often exhibit similarity - for example, patients with Menkes and Wilson disease, for example, both suffer from abnormalities in copper metabolism and in some cases neurodegeneration.

One limitation of this study is the lack of any *quantification* of the degree of similarity between the phenotype pairs of paralogous disease alleles.

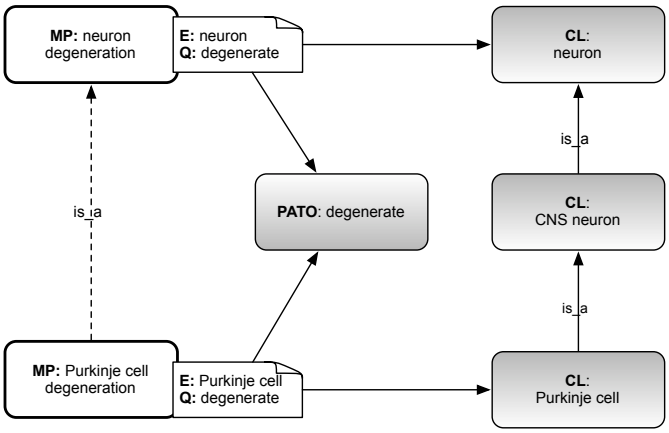


Figure 23: Logical definitions for Mammalian Phenotype Ontology classes, making use of PATO and Cell Ontology (CL). Leveraging external ontologies means we can use automated reasoners to infer relationships, such as the on between Purkinje cell degeneration and neuron degeneration. Image taken from Mungall *Genome Biology* 2010.

## 6.3 Integrating Phenotype Ontologies across Multiple Species

We took four incompatible pre-composed phenotype ontologies and generated equivalence axioms between the named classes in these ontologies and simple "Entity-Quality" (EQ) descriptions[Mungall et al., 2010]*[22]. These EQ descriptions are a syntactic variant of the OWL class expressions described previous (Mungall 2007, see previous section). We created an Obol grammar that mapped phenotype terms into class expressions and curated the results. We focused specifically on the Mammalian Phenotype (MP) ontology and mapped 72% of the 6844 classes. Figure 23 shows an example mapping for the class *Purkinje cell degeneration*.

We validated our results by attempting to recapitulate asserted relationships in MP and the Human Phenotype (HP) ontology using our mappings and automated reasoning (table 6). We discovered that in the MP over a third of the manually stated relationships could be inferred automatically. This demonstrates the benefits of using a formal approach combined with automated reasoners to partially automated construction of ontologies.

We have worked with a number of groups to help them adopt the EQ phenotype model described here, including groups involved in systematic

[22]Peer-reviewed publication in *Genome Biology*

| | HP (human) | MP (mouse) |
|---|---|---|
| Number of *is_a* relationships asserted in ontology | 10,162 | 7,950 |
| Number of *is_a* relationships that can be inferred automatically | 1,421 | 2,922 |
| Number of novel *is_a* relationships proposed (unvetted) | 407 | 478 |

Table 6: Results of using automated reasoning to recapitulate asserted relationships in pre-coordinated phenotype ontologies. Taken from Mungall 2010

mouse phenotyping pipelines[Hancock et al., 2009][23], and clinicians building ontologies of musculoskeletal system phenotypes[Gkoutos et al., 2009][24].

## 6.4 Linking Human Diseases to Animal Models using Ontology-based Phenotype Annotation

We investigated the use of PhenoBlast to identify animal models of human diseases[Washington et al., 2009]*[25]. As a first step we manually annotated 12 human disease genes using Phenote, extracting textual information from OMIM and translating these into formal phenotype expressions using `PATO`. This resulted in 1000+ genotype-phenotype associations.

We then compared these phenotypes against the full set of genotype-phenotype associations from zebrafish and mouse. The mouse phenotype annotations were from the MGI group and used the Mammalian Phenotype ontology (`MP`). We used the equivalence mappings described previously (page 44, Mungall 2010, *Integrating Phenotype Ontologies across Multiple Species*) to translate these into `PATO`-based formal phenotype expressions. The Zebrafish annotations from the ZFIN group were already in the correct form. Finally, we used the Uberon ontology in order to compare across species.

These were all loaded into an instance of the Ontology-Based Database (OBD) [Mungall, in prep] and the PhenoBlast algorithm [Mungall, in prep] was used to find similar animal genes to the 12 human disease genes based on their phenotypic profile (table 7). We also used the same algorithm to

[23]I contributed to the methods described in this paper

[24]Conference paper

[25]Peer-reviewed in *PLoS Biology*. I contributed to the analysis design, and devised and implemented all software used

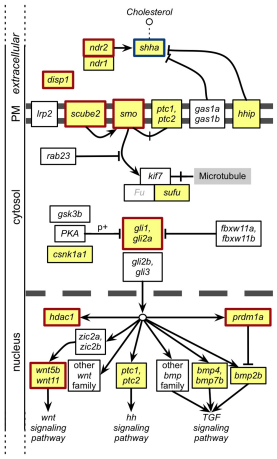


Figure 24: A phenotype similarity search for mutant phenotypes similar to zebrafish shha retrieves many known pathway members. Diagram taken from Washington 2009.

identify potential pathway members of the zebrafish Sonic hedgehog pathway (figure 24).

Since the genetic basis of human disease is often unknown, this method provides a means to identify candidate genes, pathway members, and disease models based on computationally identifying similar phenotypes within and across species.

Table 7: Comparison between human disease genes and closest matches in model organisms. Highest scoring genes when comparing a human disease gene versus either mouse or zebrafish, using four different phenotype similarity metrics. Sequence orthologs that are the top hit are in bold. From Washington et al, 2009

| | mouse | | | | zebrafish | | | |
|---|---|---|---|---|---|---|---|---|
| **Gene** | **simIC** | **simJ** | **ICCS** | **maxIC** | **simIC** | **simJ** | **ICCS** | **maxIC** |
| **ATP2A1** | Jph1 | Slc25a5 | Aldh2<br>Cisd1 | Jph1 | ryr1b | ryr1b | ryr1b | ryr1b |
| **EPB41** | Epb4.9 | Mnek1a | **Epb4.1** | **Epb4.1**<br>Epb4.2<br>Epb4.9<br>Trf | smad5 | gata1 | dtl | dtl<br>kiaa1279<br>sass6<br>stil |
| **EXT2** | Hoxd8 | Hoxd8 | Hoxc4 | Sp7 Crtap | t30212 | t30539 | t30611 + 6 unnamed | dla blo exp stb tz227c tg310a |
| **EYA1** | **Eya1** | **Eya1** | Tbx1 | Trps1<br>Gja1<br>Msx2 | rerea | fgf8a | rerea | axin1<br>chm<br>shy |
| **FECH** | Abcg2 | Abcg2 | Abcg2 | Anapc2<br>Usp8 | tal1 | abhd11 | kita | tal1 |
| **PAX2** | Rpl24 | Maf | Mitf | Mitf | lamb1 | sufu | **pax2a** | **pax2a**<br>flr<br>axin1 |
| **SHH** | Cdon | Ctnnbip1 | Alx1 | Ift57 | rerea | fgf8a | sox9a | sox9a<br>tfap2a |
| **SOX9** | Fgfr2 | Ror2 | Prrx1 | Ror2<br>Fgfr3 | fgf8a | cdc16 | fgf8a | int |
| **SOX10** | Ednrb | Ednrb | Ednrb | Ret | **sox10** | mib | **sox10** | **sox10**<br>pbx4<br>ache<br>tfap2a<br>tcf7l2<br>psoriasis |
| **TNNT2** | Hdac9 | Hdac9 | Irx4 | Hdac9 +20 tied | cx36.7 | cx36.7 | vmhc | acvr1ttna |
| **TTN** | Myl2 | Scn5a | Mybpc3 | Myl2<br>Nkx2-5 | cx36.7 | cx36.7 | **ttna** | **ttna**<br>mef2ca<br>ache<br>hey2 |

# 7 Discussion

## 7.1 Genome Analysis Pipelines

The need for genome analysis pipelines is greater than ever - as of July 2009, the Genomes Online Database (GOLD) lists nearly one thousand completely sequenced genomes and thousands more in progress[Bernal et al., 2001]. Since the initial development of the `GadFly` pipeline there have been a number of other systems developed. Taverna has a graphical interface that allows biologists to create complex interdependent workflows using web services[Oinn et al., 2004]. Pegasys also includes a similar interface[Shah et al., 2004]. MAKER is less flexible but is tuned for the rapid annotation of emerging model organism genomes[Cantarel et al., 2008]. Perhaps the most popular such tool for next-generation sequence analysis is Galaxy[Giardine et al., 2005].

The advantage of having a graphical interface to construct a workflow is a major advantage for individual researchers lacking programming skills. These researchers may have a particular analysis which they repeat regularly - for this use case a tool such as Taverna is ideal. However, for more complex pipelines a graphical language is usually not expressive enough, and a Turing-complete programming language is required. But most existing programming languages do not parallelize well. One promising approach is to use a rule-based approach such as UNIX Makefile style production rules augmented with functional and logic programming constructs[Mungall, Bioinformatics Open Source Conference 2004 presentation[26]].

Most of the systems above implement parallelization through the use of compute clusters. The growth of multi-core processors[Asanovic et al., 2006] is an opportunity for extracting more computing power from commodity hardware. However, many algorithms are designed for execution on single-core machines. Many analysis programs will need rewritten such that they can be executed in parallel, such as for example by using the MapReduce algorithm[Dean and Ghemawat, 2004].

## 7.2 Impact of the Chado Relational Schema and the Sequence Ontology

The Chado schema was originally devised for FlyBase and has since been adopted as the core schema of the Generic Model Organism Database (GMOD) project[27], which has lead to it being adopted by many nascent Model Organ-

[26] http://open-bio.org/bosc2004/presentations/biomake.pdf

[27] http://www.gmod.org

ism Database (MOD) projects (table 8). Chado is also making headway with more established model organisms and influencing the redesign of aspects of existing MOD schemas. Chado is also being used as the core database schema in projects outside the scope of the MODs such as orthology databases and the modENCODE project and *Drosophila* neurogenetics[Pfeiffer et al., 2008].

One of the reasons for the widespread adoption of Chado is the flexibility of the schema in comparison with relatively rigid schemas such as the Ensembl core schema. Chado has been criticised for "representing the biology in the applications thereby allowing flexibility in what can be stored but at the cost of not being able to force applications to be consistent."[Stoeckert, 2005]. This criticism is valid; consistency in Chado and similar models such as GFF3 has generally been encouraged through "best-practice" guidelines[Eilbeck and Lewis, 2004] and textual definitions in the SO. For example, the SO states that the 3' UTR is adjacent to the stop codon. However, this is not enforced computationally, except in ad-hoc checks in individual programming libraries. Ideally this relationship between the UTR and the stop codon would be encoded as a logical axiom in the ontology, and this axiom could be interpreted *directly* in computer programs both to validate databases in which both UTRs and codons are explicitly asserted, and to infer the existence of either UTRs or codons in databases where only one is systematically asserted.

These axioms could be encoded as a *Genome Calculus*, incorporating relations based on the Allen Interval Algebra, extended to take into account phenomena such as reverse transcription and circular genomes[Mungall et al.] [in preparation]. Preliminary research shows that encoding such a calculus goes beyond the expressivity of both the relational model and ontology languages such as OWL, bit within the expressivity of logic programming.

## 7.3 The growth of biological ontologies

Before the inception of the `GO` in 1998, very few biologists were familiar with the term "ontology". The success of the GO, the simplicity of graph-oriented model and the biologist-friendliness of associated tools such as OBO-Edit (formerly DAG-Edit) lead to the creation of a number of other biological ontologies. Much of this development was actively encouraged by the GO, because it was obvious that the success of the GO would depend on the existence of ontologies such as the Cell ontology and the CHEBI ontology for the generation of cross-products.

The proliferation of ontologies required some kind of mechanism for ensuring consistency and avoidance of redundancy. The Open Biological Ontologies (OBO) library was created to provide the infrastructure and governance

| **Project** | **Organism(s)** |
|---|---|
| FlyBase[FlyBase-Consortium, 2003] | *Drosophila* |
| modENCODE[Celniker et al., 2009] | *Drosophila melanogaster*, *C elegans* |
| Rubin Lab, Janelia Farm | *Drosophila melanogaster* neurogenetics[Pfeiffer et al., 2008] |
| XenBase[Bowes et al., 2008] | *Xenopus* |
| dictyBase[Chisholm et al., 2006] | *Dictyostelium discoideum* |
| VectorBase[Megy et al., 2008] | *Anopheles*, *Ixodes scapularis* and other vectors |
| Generation Challenge[Wanchana et al., 2008] | Crops |
| parameciumDB[Arnaiz et al., 2007] | *Paramecium tetraurelia* |
| SGD (phenotype data)[Costanzo et al., 2009] | *Saccharomyces cerevisiae* |
| SGD lite[Dolinski and Botstein, 2005] | *Saccharomyces cerevisiae* |
| Princeton Protein Orthology Database[Heinicke et al., 2007] | Multi-species |
| Sanger Institute, Pathogen sequencing[Hertz-Fowler et al., 2004] | Pathogens |
| wFleaBase[Colbourne et al., 2005] | *Daphnia* |
| BeetleBase[Wang et al., 2006] | *Tribolium* |
| AphidBase[Gauthier et al., 2007] | *Acyrthosiphon* |
| ButterflyBase[Papanicolaou et al., 2008] | *Lepidoptera* |

Table 8: Projects and databases using the Chado Relational Schema. A high proportion are new model organism databases, but some established model organism databases are using subsets of the Chado schema for particular datatypes

for the development of multiple ontologies such that they form a logical and coherent whole.

## 7.4 Formalization of ontologies

Many biological ontologies start out as informal controlled-vocabulary style resources. The lack of formal axioms limits the ability to which we can use automated means to build the ontology or to answer questions about biology using the ontology. Obol was developed as a means of transitioning controlled-vocabulary style ontologies into logical structures. Obol improves on other techniques such as perl regular expressions because Obol DCGs are at least as expressive as context free grammars (CFGs), which are more expressive than regular grammars.

Obol has been used successfully within the GO to augment the ontology with additional formal definitions connecting the GO with multiple other OBO ontologies, and has also been used to augment pre-composed phenotype ontologies in the same way. Obol is now being adopted by other communities such as the Plant Ontology Consortium [Jaiswal et al., 2005]. Obol was also featured prominently in an article in *The Scientist*[Adams, 2005].

## 7.5 Querying data: data warehouses, marts and mediators

One solution to the problem of data integration is to build a data warehouse - typically a denormalized relational database tuned towards answering questions rather than managing data. Two such systems are BioMART (formerly EnsMart)[Kasprzyk et al., 2004] and InterMine (formerly FlyMine)[Lyne et al., 2007]. The InterMine system can be used in conjunction with Chado. These systems provide a powerful means for researchers to specify complex queries and get back data in a custom report.

One problem with the warehouse approach is data latency - the warehouse must be constructed from external sources, a task that is mostly automatable but still time consuming. This means that querying a warehouse means the data may not be up to date or in sync with other data in the system. Another issue is that the databases that are combined are determined in advance, limiting the types of questions that can be answered.

Ideally different datasources could be combined *dynamically* at query time, avoiding data latency issues and allowing a wider range of datasets to be combined. This is known as the mediator approach or federated database approach, and groups such as the Neurosciences Information Framework (NIF) are applying this approach[Gupta et al., 2008].

## 7.6 Phenotype ontology based data integration

The original motivation for our phenotype ontology work was biomedical and translational applications, for example, semantically integrating model organism and clinical datasets to be able to find animal models for human diseases[Rubin et al., 2006]. However, the systematic representation of phenotypes is also vital for the study of evolution. We have been working with scientists from the National Evolutionary Synthesis Center (NESCent)[Senkowsky, 2007] to encode phylogenetic character state matrices using ontologies in order to systematically compare across different evolutionary datasets[Mabee, Ashburner, Cronk, Gkoutos, Haendel, Segerdell, Mungall , and Westerfield, 2007][28]. One of the most exciting aspects of this preliminary research is the promise of semantically integrating clinical, model organism and evolutionary datasets, allowing us to search for potential new model organisms based on the similarity between evolutionary phenotypes and human diseases.

## 7.7 Future Directions

The advancement of the life sciences could be accelerated by the creation of advanced information systems that are capable of intelligently analysing and synthesising data and knowledge from multiple sources to generate new hypotheses. This is a grand challenge requiring coordinated progress in a variety of areas. Some of the opportunities and potential future directions for research continuing the work described in this thesis include:

1. **Embracing multi-core computing.** If we are to maximize use of forthcoming multicore architecture computing, we need to develop more intelligent, expressive and flexible analysis pipeline systems. These systems should also be better integrated into upstream databases, curation and analyses workflows. Ultimately, these systems should be imbued with greater autonomy and intelligent inferencing capabilities.

2. **A unified formalism for data and knowledge.** Databases and ontologies are currently poorly integrated. We need systems that subsume both the relational model and description logic languages that combine the best of both worlds. Preliminary work (Mungall LNCS 2009) shows logic programming as one possible paradigm deserving further study. This could be used as a basis for a **Genome Calculus**, a formalization of the relations and dynamic transformations that hold between genomic entities.

[28] I contributed to the writing of the manuscript for this position paper

3. **Applications of enhanced biological ontologies.** Many applications such as term enrichment tools treat ontologies as simple directed acyclic graphs, ignoring the semantics of edge labels. By equipping these applications with greater inferencing capabilities, we create opportunities for more sensitive and specific analyses. We also need better tools to allow domain-experts to contribute to and use ontologies containing complex logical axioms.

4. **Integration of probabilistic and logical inference.** Explore the gap between statistical data mining and probabilistic modeling and logic-based ontology formalisms. Initial work on data mining using ontologies described in this thesis is promising, but this needs to be placed on a firmer probabilistic modeling basis.

5. **Knowledge-driven phenotype studies.** Genome Wide Association Studies (GWAS) have yielded a plethora of new findings, and these will soon be augmented by data coming from the 1000 genomes project and other personal genomics and next-generation sequencing sources. Yet these studies are often inconclusive or difficult to biologically interpret due to the large hypothesis space. By combining these analyses with systems biology data and detailed phenome resources we can employ a more knowledge-driven approach, yielding greater insight into underlying biological mechanisms.

## 7.8 Historical Context, Revisited

The background material in this thesis outlined the need to bring computer science research and biological problems together, particularly in the realm of knowledge representation and ontologies. One of the hindrances has been the perceived difficulty of adopting more complex representational structures and tools. The work described in this thesis has attempted to bridge that gap. For example, the Obol tool (section 4.4) has been used to assist in the migration of simple terminology-style ontologies to include richer description-logic style constructs. This was complemented by extensive manual work in translating biological knowledge into computable form (see for example, subsections 6.3, 4.5 and 4.2).

There are still a number of challenges remaining. The application of description logic technology has been instrumental in the development, consistency checking and integration of biological ontologies, but has not yet found a great deal of success in applying these ontologies to answer key biological questions. One reason is the lack of scalability over large bioinformatics

datasets - in part alleviated by using alternative reasoning strategies (see for example the methods section of the papers in subsections 6.3 and 6.4).

Another reason is the rigidity of purely deductive approaches to question answering. These only return answers that are guaranteed to be true; to formulate biological hypotheses we sometimes have to explore possibilities that cannot be proved to be true. An alternative approach is inductive reasoning[Muggleton, 1991], which has already been successfully applied in biological applications[King et al., 2004]. Another approach is meld deductive reasoning with semantic similarity measures - in subsection 6.4 in showed how this can be applied to explore relationship between genotypes and phenotype in a novel way. These hybrid approaches represent a rich mine of potential future research.

# 8 Conclusions

## 8.1 Summary

The key developments in this thesis are as follows:

1. The design and implementation of an automated genomic annotation pipeline to execute multiple interdependent analysis programs in parallel across a compute farm, and to synthesize the results into gene predictions. We used this system in concert with the Apollo tool to comprehensively annotate the *Drosophila melanogaster* genome, and to explore questions about the evolutionary relationship between *Drosophila* and other organisms on a molecular level. This research is described by a series of papers in *Genome Biology* (see in particular Misra *Genome Biology* 2002). My primary contribution was the development of workflows and databases [Mungall *Genome Biology* 2002].

2. The specification of a modular database schema for representing genomic features in the context of the biology of the organism[Mungall *Bioinformatics* 2007]. This schema makes use of formal structures called ontologies such as the Sequence Ontology [Eilbeck *Genome Biology* 2005, Mungall *Journal of Biomedical Informatics* (SO) 2010] to represent complex biological phenomena in a flexible and expressive fashion . The modularity and flexibility of this schema led to its widespread adoption for a number of model organism projects, as well as a number of genomic analyses [Yandell *PLoS Computational Biology* 2006]

3. The initiation and development of a number of biological ontologies forming the framework of the OBO Foundry[Smith *Nature Biotechnology* 2007], and the development of tools and methods for enhancing the content and expressivity of these ontologies such that logic programming and automated reasoning engines can be used [Mungall *Comparative and Functional Genomics* 2004], [Mungall *Genome Biology* 2010, Mungall *Journal of Biomedical Informatics* (GO) 2010].

4. The development of a number of tools that present complex genomic data to researchers in a visually intuitive way, and allow knowledge to be acquired from domain experts.

5. The exploration of the relationship between genotype and phenotype, at the allele level[Yandell *PLoS Computational Biology* 2008], and at

the phenome level[Washington *PLoS Biology* 2009] using multiple integrated ontologies spanning multiple species[Mungall *Genome Biology* 2010].

## 8.2 Next Generation Information Systems and Translational Research

High-throughput biology and next-generation sequencing technology promises to deliver unprecedented quantities of heterogeneous data. Translating this into meaningful biomedical results will require sophisticated information systems for data analysis, integration and knowledge-based interpretation.

The components of these *next-generation information systems* include: a flexible **analysis pipeline and workflow** system for analyzing genome-scale data across multiple processors; a powerful **database system** and multiple expressive **ontologies** for structuring data and knowledge; a suite of knowledge acquisition and visualization **tools** embedded in a distributed architecture.

These next generation systems will be invaluable tools for **translational research**. This thesis describes work applying these systems to combine and reason over data and knowledge from multiple genotype-phenotype sources, demonstrating a method for identifying candidate genes, pathway members and models for human diseases. By extending and refining these systems it should be possible to explore further the pathological processes, genetic mechanisms and shared evolutionary mechanisms underlying diseases, taking advantage of the ever-increasing volume of functional genomics data.

# 9 Acknowledgments

This thesis summarizes work related to a number of different projects funded by a variety of institutions, including the Howard Hughes Medical Institute, the National Institutes of Health NIH Grant HG00739 to FlyBase (W.M. Gelbart), roadmap initiative grant U54 HG004028 from the NIH National Human Genome Research Institute (P41 grant 5P41HG002273-08 to Gene Ontology Consortium). Early work was funded by the UK BBSRC, 1996-2003 PAGA and GAIT Bioinformatics grants.

I would like to thank Mike Tyers for his advice and sponsorship. I am grateful to the support of Gerald Rubin and to the members of the GO Consortium, Michael Ashburner and Judith Blake for their support throughout my career and during the writing of this thesis. This thesis would in all likelyhood not have been without the constant encouragement and support of Suzanna Lewis, Jonathan Bard and Ian Holmes - thank you (you can stop with the post-it notes now!). Finally, a giant thank you to my wife Rachel for the support and the lost weekends and evenings I spent writing this.

# 10 References

The following references are listed in order of citation. My name is underlined on papers on which I made a contribution. I also include annotations in highlighting the significance of the work and my specific contribution.

# References


E.R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, 2008.

M.B. Feany and W.W. Bender. A Drosophila model of Parkinson's disease. *Nature*, 404(6776):394–398, 2000.

M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422:193–197, 2003.

FlyBase-Consortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*, 30(1):106–108, Jan 2002.

FlyBase-Consortium. The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res*, 31(1):172–175, Jan 2003.

PE Bourne and J. McEntyre. Biocurators: contributors to the world of science. *PLoS Comput Biol*, 2(10):e142, 2006.

DA Benson, MS Boguski, DJ Lipman, J. Ostell, BF Ouellette, BA Rapp, and DL Wheeler. GenBank. *Nucleic acids research*, 27(1):12, 1999.

H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, et al. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic acids research*, 35(Database issue):D747, 2007.

Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H. Lackner, Jrg Bhler, Valerie Wood, Kara Dolinski, and Mike Tyers. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Research*, 36:D637D640, 2008. doi: 10.1093/nar/gkm1001. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238873`.

M.E. Martone, A. Gupta, and M.H. Ellisman. E-neuroscience: challenges and triumphs in integrating distributed data from molecules to brains. *Nature neuroscience*, 7(5):467–472, 2004.

M.Y. Galperin and G.R. Cochrane. Nucleic acids research annual database issue and the NAR online molecular biology database collection in 2009. *Nucleic Acids Research*, 2008.

L. D. Stein. Integrating biological databases. *Nature Reviews Genetics*, 4(5): 337–345, 2003.

Mark D. Adams, Susan E. Celniker, Robert A. Holt, Cheryl A. Evans, Jeannine D. Gocayne, Peter G. Amanatides, Steven E. Scherer, Peter W. Li, Roger A. Hoskins, Richard F. Galle, Reed A. George, Suzanna E. Lewis, Stephen Richards, Michael Ashburner, Scott N. Henderson, Granger G. Sutton, Jennifer R. Wortman, Mark D. Yandell, Qing Zhang, Lin X. Chen, Rhonda C. Brandon, Yu-Hui C. Rogers, Robert G. Blazej, Mark Champe, Barret D. Pfeiffer, Kenneth H. Wan, Clare Doyle, Evan G. Baxter, Gregg Helt, Catherine R. Nelson, George L. Gabor Miklos, Josep F. Abril, Anna Agbayani, Hui-Jin An, Cynthia Andrews-Pfannkoch, Danita Baldwin, Richard M. Ballew, Anand Basu, James Baxendale, Leyla Bayraktaroglu, Ellen M. Beasley, Karen Y. Beeson, P. V. Benos, Benjamin P. Berman, Deepali Bhandari, Slava Bolshakov, Dana Borkova, Michael R. Botchan, John Bouck, Peter Brokstein, Phillipe Brottier, Kenneth C. Burtis, Dana A. Busam, Heather Butler, Edouard Cadieu, Angela Center, Ishwar Chandra, J. Michael Cherry, Simon Cawley, Carl Dahlke, Lionel B. Davenport, Peter Davies, Beatriz de Pablos, Arthur Delcher, Zuoming Deng, Anne Deslattes Mays, Ian Dew, Suzanne M. Dietz, Kristina Dodson, Lisa E. Doup, Michael Downes, Shannon Dugan-Rocha, Boris C. Dunkov, Patrick Dunn, Kenneth J. Durbin, Carlos C. Evangelista, Concepcion Ferraz, Steven Ferriera, Wolfgang Fleischmann, Carl Fosler, Andrei E. Gabrielian, Neha S. Garg, William M. Gelbart, Ken Glasser, Anna Glodek, Fangcheng Gong, J. Harley Gorrell, Zhiping Gu, Ping Guan, Michael Harris, Nomi L. Harris, Damon Harvey, Thomas J. Heiman, Judith R. Hernandez, Jarrett Houck, Damon Hostin, Kathryn A. Houston, Timothy J. Howland, Ming-Hui Wei, Chinyere Ibegwam, Mena Jalali, Francis Kalush, Gary H. Karpen, Zhaoxi Ke, James A. Kennison, Karen A. Ketchum, Bruce E. Kimmel, Chinnappa D. Kodira, Cheryl Kraft, Saul Kravitz, David Kulp, Zhongwu Lai, Paul Lasko, Yiding Lei, Alexander A. Levitsky, Jiayin Li, Zhenya Li, Yong Liang, Xiaoying Lin, Xiangjun Liu, Bettina Mattei, Tina C. McIntosh, Michael P. McLeod, Duncan McPherson, Gennady Merkulov, Natalia V.

Milshina, Clark Mobarry, Joe Morris, Ali Moshrefi, Stephen M. Mount, Mee Moy, Brian Murphy, Lee Murphy, Donna M. Muzny, David L. Nelson, David R. Nelson, Keith A. Nelson, Katherine Nixon, Deborah R. Nusskern, Joanne M. Pacleb, Michael Palazzolo, Gjange S. Pittman, Sue Pan, John Pollard, Vinita Puri, Martin G. Reese, Knut Reinert, Karin Remington, Robert D.C. Saunders, Frederick Scheeler, Hua Shen, Bixiang Christopher Shue, Inga Siden-Kiamos, Michael Simpson, Marian P. Skupski, Tom Smith, Eugene Spier, Allan C. Spradling, Mark Stapleton, Renee Strong, Eric Sun, Robert Svirskas, Cyndee Tector, Russell Turner, Eli Venter, Aihui H. Wang, Xin Wang, Zhen-Yuan Wang, David A. Wassarman, George M. Weinstock, Jean Weissenbach, Sherita M. Williams, Trevor Woodage, Kim C. Worley, David Wu, Song Yang, Q. Alison Yao, Jane Ye, Ru-Fang Yeh, Jayshree S. Zaveri, Ming Zhan, Guangren Zhang, Qi Zhao, Liansheng Zheng, Xiangqun H. Zheng, Fei N. Zhong, Wenyan Zhong, Xiaojun Zhou, Shiaoping Zhu, Xiaohong Zhu, Hamilton O. Smith, Richard A. Gibbs, Eugene W. Myers, Gerald M. Rubin, and J. Craig Venter. The Genome Sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000. doi: 10.1126/science.287.5461.2185. URL `http://www.sciencemag.org/cgi/content/abstract/287/5461/2185`.

SF Altschul, TL Madden, AA Schaffer, J. Zhang, Z. Zhang, W. Miller, and DJ Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389, 1997.

L. Florea, G. Hartzell, Z. Zhang, G. M. Rubin, and W. Miller. *A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence*, volume 8. Cold Spring Harbor Lab, 1998.

M.G. Reese, D. Kulp, H. Tammana, and D. Haussler. Genie-gene finding in Drosophila melanogaster, 2000.

TM Lowe and SR Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25 (5):955, 1997.

T. Sterling, D.J. Becker, D. Savarese, J.E. Dorband, U.A. Ranawake, and C.V. Packer. BEOWULF: A parallel workstation for scientific computation. In *In Proceedings of the 24th International Conference on Parallel Processing*, 1995.

L. Stein. How Perl saved the human genome project. *The Perl Journal*, 1 (0001), 1996.

M. Ashburner, S. Misra, J. Roote, SE Lewis, R. Blazej, T. Davis, C. Doyle, R. Galle, R. George, N. Harris, et al. An exploration of the sequence of a 2.9-Mb region of the genome of Drosophila melanogaster the Adh region. *Genetics*, 153(1):179–219, 1999.

S. C. Potter, L. Clarke, V. Curwen, S. Keenan, E. Mongin, S. M. Searle, A. Stabenau, R. Storey, and M. Clamp. The Ensembl analysis pipeline. *Genome Res*, 14(5):934–41, 2004. 1088-9051 Journal Article.

S. M. Searle, J. Gilbert, V. Iyer, and M. Clamp. The otter annotation system. *Genome Res*, 14(5):963–70, 2004. 1088-9051 Journal Article.

AFA Smit, R. Hubley, and P. Green. RepeatMasker Open-3.0, 1996.

A. Stoltzfus. Molecular evolution: introns fall into place. *Current Biology*, 14(9):351–352, 2004.

J. D. Ullman. *Principles of database and knowledge-base systems, Vol. I.* Computer Science Press, Inc., 1988.

E. Birney, T. D. Andrews, P. Bevan, M. Caccamo, Y. Chen, L. Clarke, G. Coates, J. Cuff, V. Curwen, T. Cutts, T. Down, E. Eyras, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. R. Hotz, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, K. C. Woodwark, G. Cameron, R. Durbin, A. Cox, T. Hubbard, and M. Clamp. An overview of Ensembl. *Genome Res*, 14(5):925–8, 2004. 1088-9051 Journal Article Review Review, Tutorial.

N. Washington and S. Lewis. Ontologies: Scientific Data Sharing Made Easy. *Nature Education*, 1(3), 2008.

R. Stevens, C. A. Goble, and S. Bechhofer. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, 1(4):398–414, 2000. 1467-5463 Journal Article Review Review, Tutorial.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene

ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000. doi: 10.1038/75556. URL `http://dx.doi.org/10.1038/75556`.

GO-Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, 34(Database issue):D322–D326, Jan 2006.

The Gene Ontology Consortium. The Gene Ontology project in 2008. *Nucleic Acids Res*, Nov 2007.

Purvesh Khatri and Sorin Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, Sep 2005. doi: 10.1093/bioinformatics/bti565. URL `http://dx.doi.org/10.1093/bioinformatics/bti565`.

S.E. Lewis. Gene Ontology: looking backwards and forwards. *Genome biology*, 6(1):103, 2004.

Barry Smith and Anand Kumar. Controlled vocabularies in bioinformatics: a case study in the gene ontology. *Drug Discovery Today: BIOSILICO*, 2 (6):246–252, 2004. TY - JOUR.

K. Degtyarenko, P. Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcantara, M. Darsow, M. Guedj, and M. Ashburner. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 2007.

L.N. Soldatova and R.D. King. Are the current ontologies in biology good ontologies? *Nature Biotechnology*, 23(9):1095–1098, 2005.

C. J. Wroe, R. Stevens, C. A. Goble, and M. Ashburner. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pac Symp Biocomput*, pages 624–35, 2003. Journal Article.

M.Horridge, N.Drummond, J.Goodwin, A.Rector, R.Stevens, and H.Wan. The Manchester OWL Syntax. *OWL: Experience and Directions 2006*, 2006.

B. Smith, J. K?hler, and A. Kumar. On the Application of Formal Principles to Life Science Data: a Case Study in the Gene Ontology. In *Data Integration in the Life Sciences 2004*, pages 79–94, 2004.

JB Bard. Anatomics: the intersection of anatomy and bioinformatics. *Journal of anatomy*, 206(1):1, 2005.

Jonathan Bard, Seung Y Rhee, and Michael Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21, 2005. doi: -2005-6-2-r21. URL `http://dx.doi.org/-2005-6-2-r21`.

T.F. Hayamizu, M. Mangan, J.P. Corradi, J.A. Kadin, and M. Ringwald. The Adult Mouse Anatomical Dictionary: a tool for annotating and integrating data. *Genome Biology*, 6(3):R29, 2005.

R. A. Baldock, J. B. Bard, A. Burger, N. Burton, J. Christiansen, G. Feng, B. Hill, D. Houghton, M. Kaufman, J. Rao, J. Sharpe, A. Ross, P. Stevenson, S. Venkataraman, A. Waterhouse, Y. Yang, and D. R. Davidson. EMAP and EMAGE: a framework for understanding spatially organized data. *Neuroinformatics*, 1(4):309–25, 2003. 1539-2791 Journal Article.

Cynthia L Smith, Carroll-Ann W Goldsmith, and Janan T Eppig. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6(1):R7, 2005a. doi: 10.1186/gb-2004-6-1-r7. URL `http://dx.doi.org/10.1186/gb-2004-6-1-r7`.

G. V. Gkoutos, E. C. Green, A. M. Mallon, J. M. Hancock, and D. Davidson. Building mouse phenotype ontologies. *Pac Symp Biocomput*, pages 178–89, 2004. Journal Article.

G. V. Gkoutos, E. C. Green, A. M. Mallon, J. M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome Biol*, 6(1):R8, 2005a. 1465-6914 Journal Article.

G. V. Gkoutos, E. C. Green, S. Greenaway, A. Blake, A. M. Mallon, and J. M. Hancock. CRAVE: a database, middleware and visualization system for phenotype ontologies. *Bioinformatics*, 21(7):1257–62, 2005b. 1367-4803 Evaluation Studies Journal Article.

B.N. Grosof, I. Horrocks, R. Volz, and S. Decker. Description logic programs: Combining logic programs with description logic. In *Proc. 12th Intl. Conf. on the World Wide Web (WWW-2003)*, 2003.

C. Draxler. *Accessing Relational and Higher Databases Through Database Set Predicates*. PhD thesis, PhD thesis, Zurich University, 1991, 1991.

A. Gupta, B. Ludascher, and M.E. Martone. BIRN-M: a semantic mediator for solving real-world neuroscience problems. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 678–678. ACM New York, NY, USA, 2003.

W.F. Clocksin and C.S Mellish. *Programming in Prolog*. Springer-Verlag, New York, 1981.

I. Holmes and G.M. Rubin. Pairwise RNA structure comparison with stochastic context-free grammars. In *Pac Symp Biocomput*, volume 2002, pages 163–174, 2002.

Noam Chomsky. On certain formal properties of grammars. *Information and Control*, (2):137–167, 1959.

K. Sagonas, T. Swift, and D.S. Warren. XSB as an efficient deductive database engine. *ACM SIGMOD Record*, 23(2):442–453, 1994.

J. Wielemaker. An overview of the SWI-Prolog programming environment. In *13th International Workshop on Logic Programming Environments*, pages 1–16, 2003.

LD Stein and J. Thierry-Mieg. AceDB: A genome database management system. *Computing in Science & Engineering*, 1(3):44–52, 1999.

W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. The human genome browser at UCSC, 2002.

G.A. Helt, S. Lewis, A.E. Loraine, and G.M. Rubin. BioViews: Java-based tools for genomic data visualization, 1998.

S. Lewis, M. Ashburner, and M.G. Reese. Annotating eukaryote genomes. *Current Opinion in Structural Biology*, 10(3):349–354, 2000.

R. D. Dowell, R. M. Jokerst, A. Day, S. R. Eddy, and L. Stein. The distributed annotation system. *BMC Bioinformatics*, 2(7):1471–2105, 2001.

S. Vinoski et al. CORBA: Integrating diverse applications within distributed heterogeneous environments. *IEEE Communications Magazine*, 35(2):46–55, 1997.

ST Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, EM Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308, 2001.

D. Botstein and N. Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *nature genetics*, 33:228–237, 2003.

A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, and V.A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(Database Issue): D514, 2005.

A.J. Schuhmacher, C. Guerra, V. Sauzeau, M. Cañamero, X.R. Bustelo, and M. Barbacid. A mouse model for Costello syndrome reveals an Ang II–mediated hypertensive condition. *The Journal of Clinical Investigation*, 118(6):2169, 2008.

G.B. Collin, J.D. Marshall, A. Ikeda, W.V. So, I. Russell-Eggitt, P. Maffei, S. Beck, C.F. Boerkoel, N. Sicolo, M. Martin, et al. Mutations in ALMS1 cause obesity, type 2 diabetes and neurosensory degeneration in Alstr "om syndrome. *Nature genetics*, 31(1):74–78, 2002.

T. Arsov, D.G. Silva, M.K. O'Bryan, A. Sainsbury, N.J. Lee, C. Kennedy, S.S.M. Manji, K. Nelms, C. Liu, C.G. Vinuesa, et al. Fat Aussie–A New Alstrom Syndrome Mouse Showing a Critical Role for ALMS1 in Obesity, Diabetes, and Spermatogenesis. *Molecular Endocrinology*, 20(7):1610, 2006.

J. Lomax. Get ready to GO! A biologist's guide to the Gene Ontology. *Briefings in bioinformatics*, 6(3):298–304, 2005.

E. I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. GO::TermFinder open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20:37103715, 2004.

P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–83, 2003. 1367-4803 Evaluation Studies Journal Article Validation Studies.

C. Pesquita, D. Faria, H. Bastos, A. O. Falco, and F. M. Couto. Evaluating GO-based semantic similarity measures. In *Proceedings of the 10th Annual Bio-Ontologies Meeting (Bio-Ontologies 2007). Stevens R, Lord P, McEntire R, Sansone SA, eds*, page 3740, 2007.

Catia Pesquita, Daniel Faria, Andr O. Falco, Phillip Lord, and Francisco M. Couto. Semantic Similarity in Biomedical Ontologies. *PLoS Comput Biol*, 5(7):e1000443, 07 2009. doi: 10.1371/journal.pcbi.1000443.

D. P. Hill, J. A. Blake, J. E. Richardson, and M. Ringwald. Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res*, 12(12):1982–91, 2002. 1088-9051 Journal Article.

C. J. Mungall , S. Misra, B. P. Berman, J. Carlson, E. Frise, N. Harris, B. Marshall, S. Shu, J. S. Kaminker, S. E. Prochnik, C. D. Smith, E. Smith, J. L. Tupy, C. Wiel, G. M. Rubin, and S. E. Lewis. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol*, 3(12):RESEARCH0081, 2002.

**This paper describes the design and implementation of an automated genome annotation system. This system underpinned many of the other concurrent publications in this issue of Genome Biology.**

Sima Misra, Madeline A Crosby, Christopher J Mungall , Beverley B Matthews, Kathryn S Campbell, Pavel Hradecky, Yanmei Huang, Joshua S Kaminker, Gillian H Millburn, Simon E Prochnik, Christopher D Smith, Jonathan L Tupy, Eleanor J Whitfied, Leyla Bayraktaroglu, Benjamin P Berman, Brian R Bettencourt, Susan E Celniker, Aubrey D N J de Grey, Rachel A Drysdale, Nomi L Harris, John Richter, Susan Russo, Andrew J Schroeder, Sheng Qiang Shu, Mark Stapleton, Chihiro Yamada, Michael Ashburner, William M Gelbart, Gerald M Rubin, and Suzanna E Lewis. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome biology*, 3(12), 2002.

**This paper describes the results of the analysis of the Drosophila genome. My contribution was in the automated genome analysis and in-silico experimental design and analysis.**

Roger A Hoskins, Christopher D Smith, Joseph W Carlson, A. Bernardo Carvalho, Aaron Halpern, Joshua S Kaminker, Cameron Kennedy, Chris J Mungall , Beth A Sullivan, Granger G Sutton, Jiro C Yasuhara, Barbara T Wakimoto, Eugene W Myers, Susan E Celniker, Gerald M Rubin, and Gary H Karpen. Heterochromatic sequences in a Drosophila whole-genome shotgun assembly. *Genome Biol*, 3(12):RESEARCH0085, 2002.

**I implemented a modified gene-finding strategy for detecting hard-to-find genes in highly repetitive heterochromatic sequence.**

Christopher D Smith, Shengqiang Shu, Christopher J Mungall , and Gary H Karpen. The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science*, 316:1586–1591, June 2007a. doi: 10.1126/science.1139815. URL `http://dx.doi.org/10.1126/science.1139815`.

**This paper concludes the analysis of Drosophila heterochromatin. I devised the system used for data analysis and data management, and carried out the Gene Ontology analysis.**

Casey M Bergman, Barret D Pfeiffer, Diego E Rinc?n-Limas, Roger A Hoskins, Andreas Gnirke, Chris J Mungall , Adrienne M Wang, Brent Kronmiller, Joanne Pacleb, Soo Park, Mark Stapleton, Kenneth Wan, Reed A George, Pieter J de Jong, Juan Botas, Gerald M Rubin, and Susan E Celniker. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. *Genome Biol*, 3(12): RESEARCH0086, 2002.

**I performed the genome analysis for multiple related *Drosophila* partially completed genomes.**

G. M. Rubin, M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall , P. H. O'Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng, and S. Lewis. Comparative genomics of the eukaryotes. *Science*, 287(5461):2204–2215, Mar 2000.

**This paper compared the first pass Drosophila genome with other eukaryotes. I contributed to part of the genome analysis effort.**

Mark Yandell, Chris J Mungall , Chris Smith, Simon Prochnik, Joshua Kaminker, George Hartzell, Suzanna Lewis, and Gerald M Rubin. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Computational Biology*, 2 (3):e15, Mar 2006. doi: 10.1371/journal.pcbi.0020015. URL `http://dx.doi.org/10.1371/journal.pcbi.0020015`.

**An analysis of changes in gene structure compared to protein sequence. I contributed to the experimental design, and I devised and implemented the software and analysis pipeline.**

J. E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall , B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, 2002.

**BioPerl is a fundamental piece of bioinformatics infrastructure. I actively contribute code to bioperl, including heavily used software for generating structured genome relationships from GenBank records. I also contributed to the BioSQL design and implementation.**

Christopher J. Mungall , David B. Emmert, and The FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13):i337–346, 2007a. doi: 10.1093/bioinformatics/btm189. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/13/i337`.

**The Chado schema is in use for multiple model organism databases and next generation genome projects. Chado is tightly integrated with ontologies.**

K. Eilbeck, S. E. Lewis, C. J. Mungall , M. D. Yandell, L. D. Stein, R. Durbin, and M. Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5), 2005.

**The Sequence Ontology (SO) represents types of genomic entities and the relations that hold between them. The SO is widely used in genomics, and in particular in the Chado schema. I contributed to the design of the ontology and writing of the paper.**

Christopher Mungall , Colin Batchelor, and Karen Eilbeck. Evolution of the Sequence Ontology terms and relationships. *Journal of Biomedical Informatics*, 44(1):87 – 93, 2011a. doi: 10.1016/j.jbi.2010.03.002. URL `http://www.sequenceontology.org/resources/pre-prints/jbi-preprint.pdf`. Ontologies for Clinical and Translational Research.

Colin Batchelor, Thomas Bittner, Karen Eilbeck, Chris Mungall , Jane Richardson, Rob Knight, Jesse Stombaugh, Craig Zirbel, Eric Westhof, and Neocles Leontis. The RNA ontology (RNAO): An ontology for integrating RNA sequence and structure data. In *Proceeeings of the First International Conference on Biomedical Ontology*, 2009. URL `http://precedings.nature.com/documents/3561/version/1`.

**This paper is published as conference proceedings pending final acceptance for publication as a peer-reviewed journal article. I contributed to the ontology design, particularly the logical axioms for classification of RNA secondary structures.**

C. Mungall . Experiences Using Logic Programming in Bioinformatics. volume Volume 5649/2009, pages 1–21. Springer, 2009. ISBN 978-3-642-02845-8. doi: 10.1007/978-3-642-02846-5. URL `http://www.blipkit.org/blip-iclp09.pdf`.

**Conference proceedings from *International Conference on Logic Programming 2009*. This paper describes multiple uses of logic-based programming techniques in biology, in particular for reasoning about sequences and sequence molecules such as RNAs.**

B. Smith, W. Ceusters, J. Kohler, A. Kumar, J. Lomax, C.J. Mungall , F. Neuhaus, A. Rector, and C. Rosse. Relations in Biomedical Ontologies. *Genome Biology*, 6(5), 2005b. URL `http://genomebiology.com/2005/6/5/R46`.

**This paper describes the formal properties of certain relations used in biological ontologies. My contribution was the development of the ontology, and in the development of the axioms described in the paper.**

Christopher J. Mungall . Obol: Integrating Language and Meaning in Bio-Ontologies. *Comparative and Functional Genomics*, 5(7):509–520, 2004. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2447432`.

**This conference paper describes a novel technique for extracting latent meaning from biological ontologies. The tool described here was central to the ontology restructuring efforts described in this thesis.**

Vangelis Vassiliadis, Jan Wielemaker, and Chris Mungall . Processing OWL2 ontologies using Thea: An application of logic programming. In *6th OWL Experiences and Directions Workshop (OWLED 2009)*, 2009. URL http://www.webont.org/owled/2009/papers/owled2009_submission_43.pdf.

**This conference paper describes a logic programming library and related techniques for OWL ontologies. It also describes an updated version of Obol (Mungall 2004). I drafted the manuscript and supervised and advised on the development of the tools described.**

Christopher J. Mungall , Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-Product Extensions of the Gene Ontology. *Journal of Biomedical Informatics*, 44(1):80 – 86, 2011b. ISSN 1532-0464. doi: 10.1016/j.jbi.2010.02.002. URL http://www.berkeleybop.org/people/cjm/Mungall -GO-JBI-2010.pdf. Ontologies for Clinical and Translational Research.

**This paper describes enhancements to the Gene Ontology to allow automated reasoning.**

Gil Alterovitz, Michael Xiang, David P. Hill, Jane Lomax, Jonathan Liu, Michael Cherkassky, Jonathan Dreyfuss, Chris Mungall , Midori A. Harris, Mary E. Dolan, Judith A. Blake, and Marco F. Ramoni. Ontology engineering. *Nature Biotechnology*, 28(2):128–130, February 2010. ISSN 1087-0156. doi: 10.1038/nbt0210-128. URL http://dx.doi.org/10.1038/nbt0210-128.

Michael Bada, Chris Mungall , and Lawrence Hunter. A Call for an Abductive Reasoning Feature in OWL-Reasoning Tools toward Ontology Quality Control. In *5th OWL Experiences and Directions Workshop (OWLED 2008)*, 2008. URL http://www.webont.org/owled/2008/papers/owled2008eu_submission_44.pdf.

**This conference paper describes inverse entailment techniques central to the ontology restructuring efforts described in this thesis.**

M. Ashburner, CJ Mungall , and SE Lewis. Ontologies for biologists: a community model for the annotation of genomic data. In *Cold Spring Harbor symposia on quantitative biology*, volume 68, pages 227–235, 2003.

**This conference paper lays out the requirements for what would later become the OBO Foundry (see Smith 2007), as well as future efforts to restructure the GO (see Mungall [GO] 2010).**

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall , The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, Nov 2007b. doi: 10.1038/nbt1346. URL `http://dx.doi.org/10.1038/nbt1346`.

**I am one of the founders of the OBO Foundry, and have made multiple contributions to its principles and development, as described in this paper.**

Daniel Schober, Barry Smith, Suzanna Lewis, Waclaw Kusnierczyk, Jane Lomax, Chris Mungall , Chris Taylor, Philippe Rocca-Serra, and Susanna-Assunta Sansone. Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics*, 10(1): 125, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-125. URL `http://www.biomedcentral.com/1471-2105/10/125`.

**I contributed to the standards described in this paper.**

Melissa A. Haendel, Fabian Neuhaus, David Osumi-Sutherland, Paula M. Mabee, José L.V. Jr. Mejino, Chris J. Mungall , and Barry Smith. CARO - The Common Anatomy Reference Ontology. In *Anatomy Ontologies for Bioinformatics, Principles and Practice*, volume Albert Burger, Duncan Davidson and Richard Baldock (Eds.). Springer, 2007.

**Book chapter describing an upper-ontology unifying multiple species-specific anatomy ontologies. I contributed to the ontology design and writing of the manuscript.**

Anna Maria Masci, Cecilia Arighi, Alexander Diehl, Anne Lieberman, Chris Mungall , Richard Scheuermann, Barry Smith, and Lindsay Cowell. An improved ontological representation of dendritic cells as a paradigm for all cell types. *BMC Bioinformatics*, 10(1):

70, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-70. URL http://www.biomedcentral.com/1471-2105/10/70.

**Describes an ontology of dendritic cells. I contributed to ontology design, relation axiomatization and writing the manuscript.**

Alexander D. Diehl, Alison Deckhut Augustine, Judith A. Blake, Lindsay G. Cowell, Elizabeth S. Gold, Timothy A. Gondr-Lewis, Anna Maria Masci, Terrence F. Meehan, Penelope A. Morel, NIAID Cell Ontology Working Group, Anastasia Nijnik, Bjoern Peters, Bali Pulendran, Richard H. Scheuerman, Q. Alison Yao, Martin S. Zand, and Christopher J. Mungall . Hematopoietic Cell Types: Prototype for a Revised Cell Ontology. *Journal of Biomedical Informatics*, Epub ahead of print, 2010. ISSN 1532-0464. doi: 10.1016/j.jbi.2010.01.006.

**Describes methods for restructuring the Cell Ontology, and an implementation for cells of the hematopoietic system. I assisted with and coordinated ontology design, and supervised the development of the ontology. This paper has been accepted for publication and is in press.**

Christopher J. Mungall , Georgios Gkoutos, Nicole Washington, and Suzanna Lewis. Representing Phenotypes in OWL. In Christine Golbreich, Aditya Kalyanpur, and Bijan Parsia, editors, *Proceedings of the OWLED 2007 Workshop on OWL: Experience and Directions*, Innsbruck, Austria, 2007b. URL http://www.webont.org/owled/2007/PapersPDF/paper_40.pdf.

**This conference paper (non peer-reviewed) describes a means of modeling phenotypes using a formal ontology language. The methods described here are used in some the following papers.**

Jian Hu, D. Nicholson, C. Mungall , A.L. Hillyard, and A.L. Archibald. WebinTool: a generic Web to database interface building tool. In *Database and Expert Systems Applications, 1996. Proceedings., Seventh International Workshop on*, pages 285–290, 9-10 Sept. 1996. doi: 10.1109/DEXA.1996.558323.

J. Hu, C. Mungall , A. Law, R. Papworth, J. P. Nelson, A. Brown, I. Simpson, S. Leckie, D. W. Burt, A. L. Hillyard, and A. L. Archibald. The ARKdb: genome databases for farmed and other animals. *Nucleic Acids Res*, 29(1): 106–110, Jan 2001.

**I co-designed the database schema and wrote the web-based graphical map viewing interface.**

J. Hu, C. Mungall , D. Nicholson, and A. L. Archibald. Design and implementation of a CORBA-based genome mapping system prototype. *Bioinformatics*, 14(2):112–120, 1998.

**I contributed to the object modeling and implementation.**

L. D. Stein, C. Mungall , S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis. The generic genome browser: a building block for a model organism system database. *Genome Res*, 12(10):1599–610, 2002.

**I implemented the architecture for the fruitfly database version of GBrowse.**

M.E. Skinner, A.V. Uzilov, L.D. Stein, C.J. Mungall , and I.H. Holmes. JBrowse: A next-generation genome browser. *Genome Research*, 2009. URL `http://genome.cshlp.org/content/19/9/1630.long`.

**I contributed to the design and implementation of the first iteration of JBrowse.**

Seth Carbon, Amelia Ireland, Christopher J Mungall , Shengqiang Shu, Brad Marshall, Suzanna Lewis, the AmiGO Hub, and the Web Presence Working Group. AmiGO: online access to ontology and annotation data. *Bioinformatics*, Nov 2008. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/25/2/288`.

**I implemented the architecture, and supervised the development of the interface.**

M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall , J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe,

R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32 Database issue:D258–61, 2004. 1362-4962 Journal Article.

**I designed the database schema and contributed to ontology development.**

M. Ashburner, C. A. Ball, JA Blake, H Butler, JM Cherry, J Corradi, K Dolinski, JT Eppig, M Harris, DP Hill, S Lewis, B Marshall, C Mungall , L Reiser, S Rhee, JE Richardson, J Richter, M Ringwald, GM Rubin, G Sherlock, and J Yoon. Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–1433, Aug 2001.

**I contributed to the design of the GO, and to the tools described in this paper, including the GO database and browser.**

S. E. Lewis, S. M. Searle, N. Harris, M. Gibson, V. Lyer, J. Richter, C. Wiel, L. Bayraktaroglir, E. Birney, M. A. Crosby, J. S. Kaminker, B. B. Matthews, S. E. Prochnik, C. D. Smithy, J. L. Tupy, G. M. Rubin, S. Misra, C. J. Mungall , and M. E. Clamp. Apollo: a sequence annotation editor. *Genome Biol*, 3(12):0082–1, 2002.

**I contributed to the development of this genome structure editing tool.**

John Day-Richter, Midori A Harris, Melissa Haendel, Gene Ontology OBO-Edit Working Group, and Suzanna Lewis. OBO-Edit–an ontology editor for biologists. *Bioinformatics*, 23(16):2198–2200, Aug 2007. doi: 10.1093/bioinformatics/btm112. URL `http://dx.doi.org/10.1093/bioinformatics/btm112`.

**My contribution to OBO-Edit was the design and delelopment of the built-in logical reasoner and development of the OWL Adapter (Note that my name is listed as an author on this paper under "OBO-Edit working group").**

M. J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R. W. Hyman, J. M. Carlton, A. Pain, K. E. Nelson, S. Bowman, I. T. Paulsen, K. James, J. A. Eisen, K. Rutherford, S. L. Salzberg, A. Craig, S. Kyes, M. S. Chan, V. Nene, S. J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea,

J. Allen, J. Selengut, D. Haft, M. W. Mather, A. B. Vaidya, D. M. Martin, A. H. Fairlamb, M. J. Fraunholz, D. S. Roos, S. A. Ralph, G. I. McFadden, L. M. Cummings, G. M. Subramanian, C. Mungall , J. C. Venter, D. J. Carucci, S. L. Hoffman, C. Newbold, R. W. Davis, C. M. Fraser, and B. Barrell. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906):498–511, 2002.

**My contribution was the Gene Ontology analysis.**

Mark Yandell, Barry Moore, Fidel Salas, Chris Mungall , Andrew MacBride, Charles White, and Martin G Reese. Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS Computational Biology*, 4:e1000218, November 2008. ISSN 1553-7358. URL `http://www.ncbi.nlm.nih.gov/pubmed/18989397`. PMID: 18989397.

**I developed the software and database in addition to devising and performing some of the genomic analysis.**

Christopher Mungall , Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1): R2, 2010. ISSN 1465-6906. doi: 10.1186/gb-2010-11-1-r2. URL `http://genomebiology.com/2010/11/1/R2`.

**This paper describes a method and results in unifying phenotype data from multiple databases, and was a key component in the linking of animal models to human diseases (see Washington 2009).**

John Hancock, Ann-Marie Mallon, Tim Beck, Georgios Gkoutos, Chris Mungall , and Paul Schofield. Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mammalian Genome*, 2009. doi: 10.1007/s00335-009-9208-3. URL `http://dx.doi.org/10.1007/s00335-009-9208-3`.

**This paper described some of the applications of the methods described in Mungall *Genome Biology* 2010 to integrating multiple mouse and human phenomics resources.**

GV Gkoutos, C Mungall , S Doelken, M Ashburner, S Lewis, J Hancock, P Schofield, S Khler, and PN Robinson. Entity/Quality-Based

Logical Definitions for the Human Skeletal Phenome using PATO. In *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2009)*, 2009. URL https://embs.papercept.net/conferences/scripts/abstract.pl?ConfID=9&Number=997

**This conference paper describes preliminary work applying the methods described in Mungall *Genome Biology* 2010 to musculoskeletal diseases.**

Nicole L Washington, Melissa A Haendel, Christopher J Mungall, Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking Human Diseases to Animal Models using Ontology-based Phenotype Annotation. *PLoS Biology*, 7(11), 2009. URL http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000247

**This paper describes a novel method for querying model organism genomes based on phenotypic profiles of human diseases. My contributions included devising and implementing the reasoning and similarity algorithms.**

A. Bernal, U. Ear, and N. Kyrpides. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Research*, 29(1): 126, 2001.

T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. R. Pocock, A. Wipat, and P. Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–54, 2004. 1367-4803 Journal Article.

S. P. Shah, D. Y. He, J. N. Sawkins, J. C. Druce, G. Quon, D. Lett, G. X. Zheng, T. Xu, and B. F. Ouellette. Pegasys: software for executing and integrating analyses of biological sequences. *BMC Bioinformatics*, 5(1):40, 2004. 1471-2105 Journal Article.

B.L. Cantarel, I. Korf, S. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, and M. Yandell. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1):188, 2008.

B. Giardine, C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome research*, 15(10):1451, 2005.

K. Asanovic, R. Bodik, B.C. Catanzaro, J.J. Gebis, P. Husbands, K. Keutzer, D.A. Patterson, W.L. Plishker, J. Shalf, S.W. Williams, et al. The landscape of parallel computing research: A view from berkeley. *Electrical Engineering and Computer Sciences, University of California at Berkeley, Technical Report No. UCB/EECS-2006-183, December*, 18(2006-183):19, 2006.

Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI '04)*, pages 137–150, San Francisco, USA, December 2004. URL `http://labs.google.com/papers/mapreduce-osdi04.pdf`.

Barret D Pfeiffer, Arnim Jenett, Ann S Hammonds, Teri-T B Ngo, Sima Misra, Christine Murphy, Audra Scully, Joseph W Carlson, Kenneth H Wan, Todd R Laverty, Chris Mungall , Rob Svirskas, James T Kadonaga, Chris Q Doe, Michael B Eisen, Susan E Celniker, and Gerald M Rubin. Tools for neuroanatomy and neurogenetics in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 105:9715–20, July 2008. ISSN 1091-6490. URL `http://www.ncbi.nlm.nih.gov/pubmed/18621688`. PMID: 18621688.

S.E. Celniker, L.A.L. Dillon, M.B. Gerstein, K.C. Gunsalus, S. Henikoff, G.H. Karpen, M. Kellis, E.C. Lai, J.D. Lieb, D.M. MacAlpine, et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.

J. B. Bowes, K. A. Snyder, E. Segerdell, R. Gibb, C. Jarabek, E. Noumen, N. Pollet, and P. D. Vize. Xenbase: a Xenopus biology and genomics resource. *Nucleic Acids Research*, 36:D761, 2008.

R.L. Chisholm, P. Gaudet, E.M. Just, K.E. Pilcher, P. Fey, S.N. Merchant, and W.A. Kibbe. dictyBase, the model organism database for Dictyostelium discoideum. *Nucleic acids research*, 34(Database Issue):D423, 2006.

K. Megy, M. Hammond, D. Lawson, R. V. Bruggner, E. Birney, and F. H. Collins. Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. *Infection, Genetics and Evolution*, 2008.

Samart Wanchana, Supat Thongjuea, Victor Jun Ulat, Mylah Anacleto, Ramil Mauleon, Matthieu Conte, Mathieu Rouard, Manuel Ruiz, Nandini Krishnamurthy, Kimmen Sjolander, Theo van Hintum, and Richard M. Bruskiewich. The Generation Challenge Programme comparative plant stress-responsive gene catalogue. *Nucl.*

*Acids Res.*, 36:D943–946, 2008. doi: 10.1093/nar/gkm798. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D943.

O. Arnaiz, S. Cain, J. Cohen, and L. Sperling. ParameciumDB: a community resource that integrates the Paramecium tetraurelia genome sequence with genetic data. *Nucleic Acids Research*, 35(Database issue):D439, 2007.

M.C. Costanzo, M.S. Skrzypek, R. Nash, E. Wong, G. Binkley, S.R. Engel, B. Hitz, E.L. Hong, and J.M. Cherry. New mutant phenotype data curation system in the Saccharomyces Genome Database. *Database*, 2009(0), 2009.

K. Dolinski and D. Botstein. Changing perspectives in yeast research nearly a decade after the genome sequence, 2005.

S. Heinicke, MS Livstone, C. Lu, R. Oughtred, F. Kang, et al. The Princeton Protein Orthology Database (P-POD): A Comparative Genomics. *PLoS ONE*, 2(8):766, 2007.

C. Hertz-Fowler, C.S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, et al. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic acids research*, 32(Database Issue):D339, 2004.

J.K. Colbourne, V.R. Singan, and D.G. Gilbert. wFleaBase: the Daphnia genome database. *BMC bioinformatics*, 6(1):45, 2005.

L. Wang, S. Wang, Y. Li, M.S.R. Paradesi, and S.J. Brown. BeetleBase: the model organism database for Tribolium castaneum. *Nucleic Acids Research*, 2006.

J.P. Gauthier, F. Legeai, A. Zasadzinski, C. Rispe, and D. Tagu. AphidBase: a database for aphid genomic resources. *Bioinformatics*, 23(6):783, 2007.

A. Papanicolaou, S. Gebauer-Jung, M. L. Blaxter, W. Owen McMillan, and C. D. Jiggins. ButterflyBase: a platform for lepidopteran genomics. *Nucleic Acids Research*, 36:D582, 2008.

C.J. Stoeckert. Functional genomics databases on the web. *Cellular Microbiology*, 7(8):1053–1059, 2005.

Karen Eilbeck and Suzanna E. Lewis. Sequence Ontology Annotation Guide. *Comparative and Functional Genomics*, 5:642–647, 2004. URL http://www.hindawi.com/GetArticle.aspx?doi=10.1002/cfg.446&e=cta.

Christopher J Mungall , Karen EIlbeck, and Suzanna Lewis. The Genome Interval Calculus: Extensions to the Sequence Ontology. *In preparation.*

P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, et al. Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6(7):388–397, 2005.

Amy Adams. Your Database is Talking: is anyone listening? *The Scientist*, 19(17):26, Sept 2005. URL `http://www.the-scientist.com/article/display/15711/`.

**This article highlights the uses of the Obol grammars (Mungall 2004) in bio-ontologies.**

A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*, 14(1): 160–9, 2004. 1088-9051 Journal Article.

R. Lyne, R. Smith, K. Rutherford, M. Wakeling, A. Varley, F. Guillier, H. Janssens, W. Ji, P. Mclaren, P. North, et al. FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome biology*, 8(7): R129, 2007.

A. Gupta, W. Bug, L. Marenco, X. Qian, C. Condit, A. Rangarajan, H.M. Mueller, P.L. Miller, B. Sanders, J.S. Grethe, et al. Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics*, 6(3):205–217, 2008.

Daniel L Rubin, Suzanna E Lewis, Chris J Mungall , Sima Misra, Monte Westerfield, Michael Ashburner, Ida Sim, Christopher G Chute, Harold Solbrig, Margaret-Anne Storey, Barry Smith, John Day-Richter, Natalya F Noy, and Mark A Musen. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS: A Journal of Integrative Biology*, 10(2):185–198, 2006. doi: 10.1089/omi.2006.10.185. URL `http://dx.doi.org/10.1089/omi.2006.10.185`.

S. Senkowsky. The Ascent of NESCent. *BioScience*, 57(2):106–111, 2007.

Paula M Mabee, Michael Ashburner, Quentin Cronk, Georgios V Gkoutos, Melissa Haendel, Erik Segerdell, Chris Mungall , and Monte Westerfield. Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol*, Apr 2007. doi: 10.1016/j.tree.2007.03.013. URL `http://dx.doi.org/10.1016/j.tree.2007.03.013`.

**This paper describes preliminary work unifying evolutionary biology resources and model organism databases.**

S. Muggleton. Inductive logic programming. *New generation computing*, 8 (4):295–318, 1991.

R.D. King, K.E. Whelan, F.M. Jones, P.G.K. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, and S.G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971): 247–252, 2004.

# 11 Index

# Index

# Appendices

# A Certification Statement

This section is provides to satisfy the regulations of the University of Edinburgh PhD by Research Publications.

I hereby certify that all works cited in the results sections of this thesis (pages 22 - 48) are either my own, or the products of collaborative projects on which I made a significant contribution. I have written or contributed to a total of 35 peer-reviewed publications cited in this thesis (3 of these are currently in press, or accepted pending reviewer comments on revised manuscripts). In this statement I specifically select 14 significant flagship papers for consideration, and certify my precise contributions. For the specific nature of my contribution to the other 21 publications, please see either the footnotes in the results sections, or the annotated bibliography.

## A.1 First author publications

The following six peer-reviewed papers are ones in which either I am sole author, or my contibution was the most significant or jointly most significant:

C. J. Mungall , S. Misra, B. P. Berman, J. Carlson, E. Frise, N. Harris, B. Marshall, S. Shu, J. S. Kaminker, S. E. Prochnik, C. D. Smith, E. Smith, J. L. Tupy, C. Wiel, G. M. Rubin, and S. E. Lewis. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol*, 3(12):RESEARCH0081, 2002.

**I designed the system from the ground-up. My colleagues assisted with the testing and use of the system, system administration, and with writing the manuscript.**

Christopher J. Mungall . Obol: Integrating Language and Meaning in Bio-Ontologies. *Comparative and Functional Genomics*, 5(7):509–520, 2004.

**I certify the work described here is entirely my own.**

Christopher Mungall , Georgios Gkoutos, Cynthia Smith, Melissa Haendel, Suzanna Lewis, and Michael Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biology*, 11(1):R2, 2010. ISSN 1465-6906.

**The following is the author contributions section quoted verbatim from the manuscript: *CJM conceived of and coordinated the study, drafted the manuscript, created the initial***

***mappings and performed the reasoner analysis. GG maintains mappings and coordinates changes with PATO. CS evaluated MP-XP for biological validity, evaluated reasoners results and coordinated changes with the MP. MAH and CJM conceived of and created Uberon. SEL and MA supervised the work and assisted with the manuscript.***

Christopher J. Mungall , David B. Emmert, and The FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, 23(13): i337–346, 2007a.

**This work is formally considered a joint-first author publication. DE and I jointly designed the system and jointly drafted the manuscript. DE supervised the implementation of the system, I contributed the majority of the design. Our colleagues within the FlyBase Consortium provided valuable feedback.**

Christopher Mungall , Colin Batchelor, and Karen Eilbeck. Evolution of the Sequence Ontology terms and relationships. *Journal of Biomedical Informatics*, (accepted), 2010a.

**This work was supervised by KE, who is also the chief editor of the Sequence Ontology. KE and I jointly wrote the manuscript, and the work described is the result of a collaborative effort between myself, KE and CB. Please note that whilst the final version of this manuscript has been formally accepted for publication, it is still in press**

Christopher J. Mungall , Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-Product Extensions of the Gene Ontology. *Journal of Biomedical Informatics*, In Press 2010b. ISSN 1532-0464

**I devised, coordinated and implemented the work described in this paper. My colleagues from the Gene Ontology consortium (TB, JD, AI, MH, DH, JL) assisted with the biological validation. MB assisted with the formalization.**

**Please note that whilst the final version of this manuscript has been formally accepted for publication, it is still in press**

## A.2 Additional publications

The following peer-reviewed paper describe the work of research collaborations, in which I played a significant role.

Sima Misra, Madeline A Crosby, Christopher J Mungall , Beverley B Matthews, Kathryn S Campbell, Pavel Hradecky, Yanmei Huang, Joshua S Kaminker, Gillian H Millburn, Simon E Prochnik, Christopher D Smith, Jonathan L Tupy, Eleanor J Whitfied, Leyla Bayraktaroglu, Benjamin P Berman, Brian R Bettencourt, Susan E Celniker, Aubrey D N J de Grey, Rachel A Drysdale, Nomi L Harris, John Richter, Susan Russo, Andrew J Schroeder, Sheng Qiang Shu, Mark Stapleton, Chihiro Yamada, Michael Ashburner, William M Gelbart, Gerald M Rubin, and Suzanna E Lewis. Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome Biol*, 3(12):RESEARCH0083, 2002.

**My contribution was in the automated genome analysis and in-silico experimental design. I designed the workflow system and database, devised the in-silico experimental parameters, conducted trial experiments, wrote data mining software to analyze complex genome events, assisted in the analysis and annotation.**

Christopher D Smith, Shengqiang Shu, Christopher J Mungall , and Gary H Karpen. The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science*, 316:1586–1591, June 2007a.

**I devised the system used for data analysis and data management, and carried out the Gene Ontology analysis.**

Mark Yandell, Chris J Mungall , Chris Smith, Simon Prochnik, Joshua Kaminker, George Hartzell, Suzanna Lewis, and Gerald M Rubin. Large-scale trends in the evolution of gene structures within 11 animal genomes. *PLoS Computational Biology*, 2(3):e15, Mar 2006.

**An analysis of changes in gene structure compared to protein sequence. I contributed to the experimental design, and I devised and implemented the software and analysis pipeline.**

K. Eilbeck, S. E. Lewis, C. J. Mungall , M. D. Yandell, L. D. Stein, R. Durbin, and M. Ashburner. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5), 2005.

**I contributed to the design of the ontology and specified its implementation in Chado.**

B. Smith, W. Ceusters, J. Kohler, A. Kumar, J. Lomax, C.J. Mungall , F. Neuhaus, A. Rector, and C. Rosse. Relations in Biomedical Ontologies. *Genome Biology*, 6(5), 2005b.

**My contribution was the development of the ontology, and in the development and documentation of the axioms described in the paper, as well as the creation of the OBO Relations web resource.**

Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall , The OBI Consortium, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H Scheuermann, Nigam Shah, Patricia L Whetzel, and Suzanna Lewis. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, Nov 2007b.

**I am one of the founders of the OBO Foundry, and have made multiple contributions to its principles and development, as described in this paper.**

Mark Yandell, Barry Moore, Fidel Salas, Chris Mungall , Andrew MacBride, Charles White, and Martin G Reese. Genome-wide analysis of human disease alleles reveals that their locations are correlated in paralogous proteins. *PLoS Computational Biology*, 4:e1000218, November 2008. ISSN 1553-7358.

**I developed the software and database architecture. I also integrated data from multiple sources and carried out a portion of the analysis.**

Nicole L Washington, Melissa A Haendel, Christopher J Mungall , Michael Ashburner, Monte Westerfield, and Suzanna E. Lewis. Linking Human Diseases to Animal Models using Ontology-based Phenotype Annotation. *PLoS Biology*, 7(11), 2009.

**I designed and implemented the system used to perform the analysis described in this paper. I designed and implemented the reasoning algorithm and phenotype similarity engine, as well as the database and user interface. I developed one of the ontologies central to the inter-species analysis (Uberon), and contributed to another (PATO). I also contributed to the analysis itself.**