Title

Booly : a new data integration platform for systems biology

Author

Do, Long Hoang

Publication Date

2010

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Booly: a new data integration platform for systems biology

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biology with a Specialization in Bioinformatics

by

Long Hoang Do

Committee in Charge:

Professor Ethan Bier, Chair
Professor Steven Briggs
Professor Trey Ideker
Professor Harvey Karten
Professor William McGinnis
Professor Steven Wasserman

2010

The Dissertation of Long Hoang Do is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2010

Dedication

This dissertation is dedicated to my family. Without them I would not be in the position of completing my doctoral work. First and foremost are my parents, who in the dark of the night in 1979, fled Vietnam and risked their lives in an effort to give our family a better opportunity. Secondly, I would like to dedicate this work to the memory of my younger brother Vu, who was lost on that fateful journey at the young age of two. Next are my other siblings, three older sisters and a younger brother. They have given strength to our family, supported my parents when I was off far away in college doing whatever I was doing, and provided me with all the support and encouragement a brother could ever desire. My niece and nephews have provided me with the joys of being an uncle, and prepared me well for fatherhood. Finally, I would like to dedicate this work to my wife and son for their love and support. They inspire me to dream beyond tomorrow and fill my every day with happiness and curiosity. I am truly lucky to have them in my life.

Table of Contents

List of Figures and Tables

Acknowledgements

I would like to thank Professor Ethan Bier for his support as my advisor. His guidance and enthusiasm has proved invaluable to my work. Without his prodding and encouragement, this dissertation would surely have taken another five years to complete. I am always amazed at the breadth of his curiosity to all things in science. I have been very fortunate to have him as a mentor.

Besides my advisor, I would like to thank the rest of my thesis committee for their encouragement, insightful comments, and hard questions. They were tough on me when they needed to be and I am fortunate to have some of the most distinguished thinkers in their respective fields.

I would also like to acknowledge members of the Bier lab. They have been to many of my lab meetings, and I truly feel sorrow they had to sit through so many of my technical figures and bioinformatics related subject matter. Through it all, they have always given me support, encouragement, and a grounded voice and ear so that I may improve.

To my friends and colleagues, without them, graduate school would be unbearable. First are my poker buddies who gave me endless evenings of entertainment and an outlet to perform my math and acting skills, sprinkled with a bit of science every now and then. Next are my fellow lab mates and colleagues that joined me for lunch and coffee at the cart, an exercise *almost* worth re-doing

graduate school. Finally, I would like to thank all my friends outside the program that have kept in touch and been there for me. They are truly life-long friends.

Chapters 1, 2, 3, 4, 5, in part, has been submitted for publication of the material as it may appear in Booly: a new data integration platform, Do, Long H.; Esteves, Francisco F.; Karten, Harvey J.; Bier, Ethan, BMC Bioinformatics 2010. The dissertation author was the primary investigator and author of this paper.

Chapter 3.1, in part, is being prepared for submission for publication of the material. Esteves, Francisco E.; Do, Long H.; Karten, Harvey J; Bier, Ethan. Francisco Esteves was the primary investigator and author of this paper.

Chapter 4.2, in part, is currently being prepared for submission for publication of the material. Do, Long and Bier, Ethan. The dissertation author was the primary investigator and author of this material.

Vita

1995-1997	Research Assistant, University of California, Berkeley, School of Optometry
1997	Bachelor of Arts, University of California, Berkeley, Molecular and Cellular Biology
1998-2000	Biomedical Scientist, Lawrence Livermore National Laboratory
2000	Bioinformatics Scientist, Protogene Laboratories
2000-2003	Bioinformatics Scientist, GenoSpectra, Inc. (Affymetrix)
2003	Bioinformatics Scientist, Lawrence Livermore National Laboratory
2005-2007	Graduate Teaching Assistant (Biochemistry Lab, Bioinformatics Lab), University of California, San Diego
2010	Doctor of Philosophy, University of California, San Diego, Biology with a Specialization in Bioinformatics

Publications

1. **Do LH**, Esteves FF, Karten HJ, Bier E: Booly: a new data integration platform. *BMC Bioinformatics* 2010, **11**(1):513.

2. **Do LH**, Bier E: The Booly aliasing method (unpublished).

3. Reiter LT, **Do LH**, Fischer MS, Hong NA, Bier E: Accentuate the negative: proteome comparisons using the negative proteome database. *Fly (Austin)* 2007, 1(3):164-171.

4. Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, **Do L**, Land ML, Pelletier DA, Beatty JT, Lang AS *et al*: Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nat Biotechnol* 2004, 22(1):55-61.

Fields of Study

Major Field: Biology and Bioinformatics

Studies in Biology: Professor Ethan Bier

Studies in Bioinformatics: Professor Ethan Bier and Professor Steven Briggs

ABSTRACT OF THE DISSERTATION

Booly: a new data integration platform for systems biology

by

Long Hoang Do

Doctor of Philosophy in Biology with a Specialization in
Bioinformatics

University of California, San Diego, 2010

Professor Ethan Bier, Chair

Data integration continues to remain a difficult and escalating problem in bioinformatics. The goal of this thesis is to develop a data integration platform that addresses two recurring issues in current data integration methods: 1) the issue of naming and identity and 2) the barrier of entry for general researchers to contribute and perform analysis of data. We have developed a web tool and warehousing system, Booly, that features a simple yet flexible data model coupled with the ability to perform powerful comparative analysis, including the use of

Boolean logic to merge datasets together, and an integrated aliasing system to decipher differing names of the same gene or protein. We applied Booly across heterogeneous data sources and identified genes useful in comparing avian and mammalian brain architecture, which were validated by comprehensive *in situ* hybridization experiments. The Booly paradigm for data storage and analysis should facilitate integration between disparate biological and medical fields and result in novel discoveries that can then be validated experimentally.

Chapter 1

Introduction

In the 1960s systems theory and biology enjoyed considerable interest among eminent scientists, mathematicians and engineers as researchers took a systems approach to 'search for general biological laws governing the behavior and evolution of living matter in a way analogous to the relation of the physical laws and non-living matter' [1-4]. Recently, systems biology has re-emerged as a movement in biological research that can be described as an inter-disciplinary study field that focuses on complex interactions in biological systems and as a paradigm, is the antithesis to the reductionist paradigm [4, 5].

> Systems biology...is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different...It means changing our philosophy, in the full sense of the term (Noble 2006 [5]).

Consistent with Noble's idea of "integration" is the practice of utilizing an inter-disciplinary field such as bioinformatics to develop data integration platforms and methods for carrying out systems biology analysis. However, the difficulty in achieving lasting solutions to integration of diverse biological data continues to be a central problem in bioinformatics. A number of technologies and systems have been developed that offer a variety of potential solutions to the data integration problem. These solutions differ by the architecture they adopt and by the common "touch-points" used to integrate data (e.g., data values, names, identities, schema properties, ontology terms, Uniform Resource Identifier [URI], keywords, loci, spatial-temporal points) [6].

There are two major approaches to address data integration: the data warehousing approach, and the mediated approach [6-8]. The data warehouse technique transforms the content of multiple source databases to fit a common data model, essentially integrating all the data under a single roof. In the data warehousing approach, software is created to fetch data from remote data sources, transform it into the appropriate data model, and load it into the data warehouse. Examples of the data warehouse approach include BioMart and BioWarehouse [9, 10].

In the mediated approach, a data resource environment is built around a collection of data housed and stored autonomously in remote databases. The data in remote locations is accessed by queries made against the mediated schema, and software "drivers" or "wrappers" are used to determine where and how to fetch the remote information. View integration (Kleisli, BioMediator, BioZon, TAMBIS) and link integration (SRS, Entrez) are all variations of the mediated approach [11-16]. Another variation of the mediated approach concerns the semantic integration problem. Projects such as BioMOBY and Bio2RDF use semantic integration approaches to establish complex relationships between objects by resolving conflicts between the meaning of words and concepts [17, 18].

A number of challenges exist for current data integration efforts. At the forefront is the issue of naming and identity where the same biological object (e.g. genes, proteins) contain multiple aliases. Goble and Stevens states: "The failure to address identity will be the most likely obstacle that will stop mashups, or any other technology or strategy, becoming an effective integration mechanism [6]". For example, the BioMOBY project utilizes shared ontologies as a semantic integration approach to describe biological objects, but does not necessarily help in naming the biological objects or resolve the same objects with multiple names. One idea to address the issue of naming and identity, as Stein proposes, is to create globally unique identifiers as has been done with human gene symbols by the HUGO Gene Nomenclature Committee [7, 19]. However, Stein acknowledges the practice rarely works due to the "dynamic nature of the field" and that changes would be too difficult to keep up with for any commission.

The second major challenge in data integration is the barrier of entry for scientists and developers to contribute and perform analysis of data. The data integration systems currently in existence either do not account for the general researcher contribution or are too difficult to utilize by non-specialists. Stein opines "…the simplicity of the data models and the ease of implementation…will make or break [data integration] attempts [7]." The Bio2RDF project encourages research data be deposited in a machine-readable format through the use of the Resource Description Framework (RDF) [17]. RDF is regarded as a necessary component in the semantic web movement, allowing web content to be meaningful to machines [20]. However, RDF is overly complicated for the broad base of users who are confronted with the simple but vexing problem of integrating data from a diverse set of spreadsheets with other data sources.

The goal of this thesis is to address the two main challenges when carrying out data integration for the purpose of systems biology analysis. First, it describes a platform, Booly, with a simple yet flexible data model and tools that lower the barrier of entry for general researchers and developers to contribute, collaborate, and perform analysis of data. Secondly, it presents an on demand aliasing method integrated into Booly to address the issue of naming and identity. Finally, it presents a number of biologically relevant uses of Booly, including the identification of genes useful for comparing avian and mammalian brain architecture, and a method to integrate genetic interactions with known disease targets to discover secondary uses of marketed drugs (drug repurposing).

The novel framework Booly provides for storing and integrating biological databases, with contributions from general researchers and large data centers alike coupled with high-throughput on demand alias translations, will spark a new approach in data integration efforts. These advantages over existing data integration approaches should attract growing contributions from developers and the research community that spur important new discoveries.

Chapter 1, in part, has been submitted for publication of the material as it may appear in Booly: a new data integration platform, Do, Long H.; Esteves, Francisco F.; Karten, Harvey J.; Bier, Ethan, BMC Bioinformatics 2010. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Overview

2.1 Implementation

The Booly data integration platform consists of a data warehouse, scripts to perform alias lookups and Boolean operations, and a web interface for interaction from the user. In Booly, data from Gene Ontology [21] and PubMed are represented as individual datasets similar to a spreadsheet table consisting of rows and columns. Each dataset can be merged with others to produce an output of the requested combination of Boolean operations constrained against the identifiers and their aliases grouped by a similar fingerprint such as gene sequence or chemical formula (Figure 1, chapter 4). For example, one can merge a table of microarray data with a Gene Ontology dataset to attach annotation to previously unannotated microarray data. Furthermore, heterogeneous identifiers from the datasets are resolved by the integrated alias lookups and applied accordingly.

One can perform a combination of Boolean logic on multiple datasets by simply arranging datasets on our web interface in a manner akin to an algebra equation. We demonstrate the ability to perform powerful comparative analysis on the recently sequenced twelve *Drosophila* genomes to identify genes lost in one species of the *melanogaster* subgroup (Figure 2, Figure 3, Figure 4, Table 1, section 3.2) [22, 23].

Combining diverse datasets can be difficult when consideration must be made to map identifiers to a uniform nomenclature. A number of aliasing services exist which perform the task of alias resolution (DAVID, Synergizer, AliasServer, HGNC) [19, 24-26], however many require pre-existing knowledge of an identifier's source before translation can be performed while others lack the flexibility to allow for aliases beyond just genes and proteins (e.g. aliases for drugs or ontology terms). To resolve these shortcomings, we have implemented our own streamlined form of alias resolution and demonstrate an approximate running time performing a Booly intersection with aliasing (Figure 5, Figure 6, section 3.4, section 4.2).

We illustrate the power of Booly's alias resolution while integrating multiple sources for the purpose of comparing mammalian and avian brain architecture. Our analysis began with a homebrew dataset we curated from the Allen Institute Brain Atlas for genes that are selectively expressed at high levels in the mouse hippocampus [27]. The next step was to integrate this dataset with mouse Gene Ontology and BLAST [28] hits of the mouse genome against other species such as the fish, chicken, and fruitfly. Unfortunately, while the Allen data and Gene Ontology had identifiers mapped to official mouse gene names, our BLAST data had identifiers mapped to Ensembl [29] sequence identifiers. Using our aliasing tool, we overcame this commonly encountered problem and seamlessly integrated these datasets together, which resulted in the identification of an enriched set of genes that are expressed in a region of the avian brain believed to correspond to the mammalian hippocampus (Figures 7-10, section 3.1).

2.2 Extensions and applications

In addition to the core functionalities we have previously described, Booly can be extended further by creation of new applications. For example, we created an application that allows researchers to generate new BLAST datasets. Another application allows the user to switch "touch points" (identifiers used to map one piece of data to another) [6], which makes it possible to perform concatenated series of complex Boolean comparisons (Figure 11, section 4.3.4). An example of the utility of this tool is to integrate known *Drosophila melanogaster* genetic interaction networks with human diseases and existing uses of FDA approved drugs to develop a new approach to identify new potential uses for drugs, sometimes referred to as drug repurposing (Figure 12, Figure 13, section 3.3). Additionally, the Application Programming Interface (API) utilizing RESTful web services (http://booly.ucsd.edu/api) is will allow developers an easy way to both import and retrieve data within Booly.

2.3 Privacy and data integrity

The Booly web application allows for users to create a secure, personalized account for storage of datasets. In this manner, only the original owner of a data set will be able to view, modify, delete, and share their content. Once a data set is shared either publicly or to other individuals, permission is granted for the recipients to receive a copy of the data set, thereby preserving the original data set's integrity. The security of individual accounts is consistent with today's current web standards and will continually see improvements as the technology advances.

Chapter 2, in part, has been submitted for publication of the material as it may appear in Booly: a new data integration platform, Do, Long H.; Esteves, Francisco F.; Karten, Harvey J.; Bier, Ethan, BMC Bioinformatics 2010. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Results

3.1 Avian brain architecture

A novel and important feature of Booly is its inherent ability to combine data from diverse sources into single comparative tables. One practical example that illustrates this enabling power of Booly is in helping to define sets of candidate genes that might be expressed in corresponding functional regions of the mammalian versus avian brains.

We were interested in testing the hypothesis that even though avian and mammalian brains appear superficially different in organization, that there are clear homologous regions in these brains as indicated by shared localized patterns of gene expression [30]. As a starting point, we took advantage of the genome-wide expression data in the mouse brain generously made available by the Allen Institute. We searched for genes expressed within well-defined subdomains of the murine hippocampus such as the Dentate Gyrus and the Cornu Ammonis (CA1-CA3) subdivisions of Amon’s Horn since these discrete subdomains are defined and delimited by multiple gene expression patterns [27, 31]. As a test of the conserved brain structure hypothesis we could then ask whether the avian orthologs of these signature genes are similarly expressed in similar patterns in corresponding regions of the avian brain.

To identify well-conserved hippocampal gene markers for our studies we used the Booly's ability to resolve aliases for identifiers of different datasets in our study of the vertebrate hippocampus. We searched the Allen Institute's In Situ Brain Atlas of over 20,000 gene expression patterns for genes that are highly expressed in the mouse hippocampus and then constrained our list of candidates to those genes that exhibited high sequence conservation with genes in other vertebrate and invertebrate species including the chicken as a representative of birds and *Drosophila* as an invertebrate (Figure 7). Although the Allen Institute does an admirable job of mapping gene symbols to the identifiers of sequences, we encountered a number of examples where sequences were identified from different sources (mixtures of RefSeq identifiers and other NCBI accessions—most likely as a result of unavailable RefSeq curation). Some hippocampus specific genes were also identified in previous studies that used a different set of identifiers [27]. We therefore employed the Booly to perform the alias translations between the Allen data (mouse gene symbols) and data containing previous BLAST homologs of the mouse (Ensembl) to those of non-mammals (Ensembl). We were able to also integrate other datasets such as Gene Ontology avoiding the complication of having to conform to a singular identifier source.

Our data integration of the Allen Data with homologous BLAST datasets generated ~150 evolutionarily conserved candidate genes. We then reduced this list further to a set of 13 test genes by focusing our initial attention on genes expressed in highly localized patterns that encode transcription factors as well as those involved in restricted cell-cell signaling in the hippocampus (e.g., specifically in the CA1 region or the dentate gyrus). One clear prediction of the homology hypothesis (i.e., that there should be conserved patterns of gene expression in corresponding regions of the mammalian and avian brains) is that the chicken orthologues of mammalian hippocampal genes should at least be expressed, and presumably enriched, in regions of the bird brain that have been proposed to include the avian hippocampus based on functional studies. To test this hypothesis, we performed Real Time Quantitative Polymerase Chain Reaction (RTqPCR) analysis on RNA extracted from micro-dissected regions of the chick brain. The data provide a quantitative measure of relative gene expression levels in different regions of the brain, and are highly reliable as repeated RTqPCR runs produced highly concordant results as did analysis of RNA from brain regions dissected in independent experiments (Figure 8). We compared expression levels of these predicted avian hippocampal genes in the 1-day-old chick hippocampus to that of other regions in the chick brain. We found approximately half of the genes tested showed at least a two fold elevation in gene expression in the hippocampus relative to that of neighboring regions in the chick brain. The Allen Institute database lists 286 genes (out of approximately 20,000) as being highly expressed in the hippocampal region/hippocampal formation in the mouse brain. This constitutes only 1.5% of the total genes tested in the mouse genome. In contrast, our hit rate was approximately 30-fold higher among our predicted set of avian hippocampal genes.

We also validated our RTqPCR results in several cases by *in situ* hybridization and found that these genes were indeed expressed in a localized fashion in the proposed hippocampal region of the chick brain (F. Esteves et al., manuscript in preparation). As previously mentioned, we focused on genes expressed within well-defined subdomains of the murine hippocampus such as the dentate gyrus and the CA1-CA3 subdivisions. Although the dentate gyrus and CA1-CA3 subdivisions are readily apparent in the mammalian brain, these exact homologous formations in the avian hippocampus remain unclear (Figure 9) [32]. Our *in situ* hybridization experiments identified the homeobox gene *prox1* as a reliable marker for the dentate gyrus, localizing the gene to a cell-dense "V" shaped area of the avian hippocampus (Figure 10).

Identifying a limited set of gene candidates to test experimentally in this encouraging pilot study was greatly facilitated through use of Booly and its ability to seamlessly resolve aliases.

3.2 Species specific genes in the *melanogaster* subgroup

We performed comparative analysis on the recently sequenced twelve *Drosophila* genomes to identify genes lost in one species of the *melanogaster* subgroup [22, 23]. In this study, we performed a variety of consecutive intersections (AND) and a NOT operation to determine sets of genes that were lost at various nodes in the phylogenic tree (Figure 2, Figure 3). In one such analysis, we examined genes lost during evolution of *Drosophila ananassae* that were retained in the sister melanogaster subgroup comprised of *D. melanogaster, D. simulans, D. sechelia, D. yakuba*, and *D. errecta* and a neighboring out group, *D. pseudoobscura* (Figure 3). Interestingly, the types of genes lost specifically in the *D. ananassae* lineage (73) fall into similar functional classes as those we and others have previously identified as being common classes for organism specific genes (Figure 4, Table 1) [22, 33]. These genes include those involved in interaction with the outside world such as barrier forming chorion proteins (e.g., Cp18), and odorant binding (Or88a, Obp8a), defense (e.g., Tot family defensins), and reproductive signaling (OsC: pheromone binding).

3.3 Discovering secondary drug targets

As described by Goble et al., “touch points” are targets of various efforts in data integration [6]. Booly employs a data model where identifiers (key—short text) are attached to every row of data (values can be text or html). Booly’s initial efforts for data integration involve the use of Boolean logic against the keys of each dataset. When two keys match (either directly or through alias translation), the values are brought together and the appropriate Boolean logic is performed. Our next goal was to demonstrate how Booly could be used as an intermediate step in extracting other touch points for further data integration. For this task, we first created a utility where combined datasets integrated by our Boolean logic functions can be saved as an entirely new dataset. We then added a functionality that allows for switching the identifier columns of the primary output tables, essentially changing the "touch point" to a new identifier (Figure 11, Figure 13). Switching identifiers allows users to compare datasets that were originally not amenable to Boolean joining. This process, "switching and chaining", which can be carried out in a concatenated fashion with multiple comparisons, allows for the creation of entirely new integrated datasets extending far beyond those that could be generated by standard Boolean comparisons constrained to datasets with common identifiers.

An example that illustrates the types of sophisticated analysis that can be accomplished using the switching and chaining technique is to identify a list of prescription drugs that could potentially be used to treat additional diseases outside of its known or intended target, a goal sometimes referred to as drug repurposing. One way to create a list of such candidate alternative drug uses is to link diseases into interaction networks (such networks often represent integrated biological processes such as signaling pathways or DNA repair) under the hypothesis that one might be able to use a drug treating one particular disease in the network for a second disease belonging in the same network (Figure 12). Since many diseases can be caused by mutations in genes or are related to such diseases, we first created a link between drugs, human diseases, and human disease genes. As human disease gene interaction data is not readily available, we identified homologs of human disease genes in the model organism *Drosophila melanogaster* and then utilized high quality genetic interaction data derived from the vast published literature available for this model system. In practical terms, we first linked the human disease to a gene (gathered from human genes known to have allelic variants for the particular disease-OMIM) and its prescribed drugs (FDA drug database), then we found the homolog of the human gene to that of the fruitfly, and finally we integrated genetic interaction networks found in the fruitfly (restricted to high quality interactions, ~1400 *D. melanogaster* genes). The entire process is summarized in Figure 13, which involves 3 instances of both switching and chaining of queries.

After performing this concatenated combinatorial operation, we retrieved a list of ≈ 50 genetic interactions that suggested potential alternative drug targets (http://booly.ucsd.edu/drug-networks). In one such example, the *forkhead (fkh)* fly gene was shown to interact with *brachyenteron (byn)* via a phenotypic enhancement. It's closest human homologs were NP_036315-- linked to autoimmune diseases, and NP_005140--linked to hormone deficiencies, respectively. A particularly interesting drug, Cytomel, was retrieved for treating the general category of hormone deficiency. After closer inspection, it was found that Cytomel is used to treat cases of hypothyroidism. Hashimoto's thyroiditis, or chronic lymphocytic thyroiditis, is an autoimmune disease where the body's own T-cells attack the cells of the thyroid and is the most common form of hypothyroidism in the United States. Our study was able to reveal this particular connection between autoimmune disease and thyroid hormone deficiency. A question arising from this example is whether other drugs listed to treat various forms of hormone deficiency could be used to treat the other autoimmune disease, and vice versa, whether any of the drugs used to treat autoimmunity could be used to treat certain hormone deficiencies such as Hashimoto's thyroiditis.

Another example of a potential connection between genes that could have therapeutic implication is one between the multi-EGF domain *Crb1* involved in stabilizing the adherens junction and another cell junction molecule *DLG3* (http://booly.ucsd.edu/drug-networks-2). Mutations in *Crb1* cause Retinitis Pigmentosa (RP), a retinal degeneration disease, while disruption of *DLG3* function causes mental retardation. No treatments are currently available to treat RP, however, patients with *DLG3* mutations can be treated with anti-depression drugs such as Mirtazapine, Fluphenazine Hydrochloride, Buphenyl, or Prolixin Decanoate. Given the strong genetic interaction between the fly homologs of the human disease genes (*crb* ≈ *Crb1* and *DLG3* ≈ *sdt*) and the fact that they both play an important role in stabilizing cell-cell junctions one might wonder whether treatment of RP patients with drugs used to treat depression might have a positive effect. One could extract similar potential repurposing of drugs for the other drugs/diseases. To construct this particular list of candidate drugs for new diseases, we used only a small subset of interaction data in fruit flies and drug components. However, one could also perform additional queries of this kind based on other types of genetic interactions (e.g., in yeast, *C. elegans*, or mice) or using well validated protein-protein interaction data. This strategy offers a potentially useful alternative and complementary approach to existing attempts at drug repurposing based on categorizing disease states by virtue of shared gene expression profiles [34].

Our switching and chaining approach is only one example of how Booly can be used as an intermediate platform in data integration. The value fields can be extracted for touch points via other approaches and algorithms in a similar manner to how we extracted new identifiers. Coupled with an initial alias translation and Boolean logic functions, Booly offers core functionalities vital to future data integration efforts, which we anticipate will be further empowered as developers make use of its flexible simple format to create new functionalities that extend its utility.

3.4 Booly running time.

To simulate an approximate running, we performed Booly intersections with alias resolution of up to 10 datasets containing at most 20,000 genes apiece. The test was performed on one Xserve 2.3 GHz G5 PPC with 6GB RAM. A density of 200,000 genes resulted in a running time of approximately one minute (Figure 6), which includes the time it takes to resolve all aliases from the 200,000 genes, group them accordingly, perform necessary intersection operations among each dataset, fetch data attached to the genes, and display results to the client browser.

Chapter 3, in part, has been submitted for publication of the material as it may appear in Booly: a new data integration platform, Do, Long H.; Esteves, Francisco F.;

Karten, Harvey J.; Bier, Ethan, BMC Bioinformatics 2010. The dissertation author was the primary investigator and author of this paper.

Chapter 3.1, in part, is being prepared for submission for publication of the material. Esteves, Francisco E.; Do, Long H.; Karten, Harvey J; Bier, Ethan. Francisco Esteves was the primary investigator and author of this paper.

Chapter 4

Methodology

4.1 Booly integration algorithm

An overview of the Booly integration algorithm is shown in Figure 1 and summarized below.

1. Dataset Ordering. Boolean operations are performed based on the order of precedence:

a. Group Selection Using Parenthesis
b. NOT/Conjunction (-) Operation
c. AND/Intersection (+) Operation.
d. OR/Union (U) Operation.
e. Precedence for multiple instances of the same operator is determined by the order in which they appear in the query.

2. Alias Hash Key Conversion. When aliasing is requested, all identifiers from every dataset ($\boldsymbol{D}^{\boldsymbol{1..n}}$) are converted to a hash key from an in-house Alias lookup database. The hash key is derived by utilizing the Secure Hashing Algorithm (SHA1) 160-bit digest of a fingerprint such as a gene sequence, chemical formula, URI, etc... (Figure 5, Figure 14a, section 4.2). The hash key is unique to the fingerprint (avoiding collisions as is the problem with, e.g. numerical identifiers) and can convert any arbitrary length message into 40 hexadecimal characters. This makes the hash key ideal as a non-semantic identifier.

3. Identifiers Grouped Based on Aliases. Hash Keys as well as the original identifiers are grouped based on exact matches. Groups are then consolidated based on the criteria that identifiers are one and the same when the same hash key exists amongst all identifiers in question.

4. Consolidated Groups Undergo Boolean Operation. The first pair of datasets ($\boldsymbol{D^1}$, $\boldsymbol{D^2}$) based on Step 1 undergo the requested Boolean operation. The operation is performed iteratively until all matched aliases between the two datasets are exhausted.

5. Datasets Combined. The results of Step 4 are combined into a temporary dataset ($\boldsymbol{D^{1,2}}$). $\boldsymbol{D^{1,2}}$ is compared against $\boldsymbol{D^3}$ and steps 4-5 are repeated until a final dataset $\boldsymbol{D^{1..n}}$ emerges.

4.2 Aliasing

A common problem confronted by bioinformaticians is the need to resolve whether two or more identifiers are identical, i.e., are aliases of each other. A number of aliasing services have attempted to resolve the differing naming conventions created by both computational and manual labeling methods (AliasServer, DAVID, HGNC, SEGUID, MagicMatch, NCBI, ENSEMBL) [19, 24, 26, 29, 35-37]. These services differ by their technology and solutions with the general strategy of 1) using either in-house generated unique identifiers (NCBI, DAVID, ENSEMBL), or 2) the generation of unique fingerprints (AliasServer, MagicMatch, SEGUID) by way of cryptographic hashing algorithms which digest large arbitrary blocks of data (e.g., sequence) and returns a fixed-size bit string [38, 39]. As each of these systems is designed with a specific goal in mind, none of them are optimized for specifically answering the single root question: are two identifiers the same (Fig. 1a)?

4.2.1 Aliasing motivation

In the course of designing our comprehensive data warehousing and comparison application, Booly [40], we recognized a need for a dedicated aliasing tool designed to efficiently and flexibly resolve alias identities. One of the main tasks of Booly is to mix and match datasets together using combinations of the Boolean operations. A common usage of such a tool is data aggregation between multiple sources (e.g. the aggregation of Gene Ontology data to that of a home brew spreadsheet table for annotation). When identifiers from both datasets are in the same format (e.g. gene symbol), the process of integrating the data can be performed trivially. However, the process of integrating the data becomes more challenging when converting formats is needed, thus becoming an unwieldy aliasing problem. This aliasing problem is compounded when comparing multiple datasets with differing identifier formats. Furthermore, Booly was created to compare content that extends well beyond integrating sequence entities (e.g. pharmaceutical drugs, human diseases, etc.). With these requirements in mind, we designed an aliasing system (Booly-hashing) that can quickly resolve heterogeneous identifiers from multiple sources while maintaining flexibility to handle aliases from multiple entities.

4.2.2 Booly-hashing

Booly-hashing is an aliasing database resource that utilizes a 160-bit SHA-1 hash key to generate unique fingerprints of sequences and their identifiers represented as a 40 character hexadecimal number (Fig 14a) [41]. SHA-1 like other cryptographic hash functions convert large, variable sized data into a fixed-sized bit string (hash value) such that any change in the original data will result in a completely changed hash value [37]. The property of cryptographic hash functions allows for fast discrimination of sequences millions of bases long that differ even by just a single nucleotide, making them ideal for use as fingerprints for genetic sequences.

Our streamlined approach requires the storage of only the hash key and its associated identifier. Current aliasing methods utilizing the hashing technology require the source of the identifiers to be known (AliasServer, SEGUID) [26, 35]. This limits the ability to find aliases of identifiers from heterogeneous sources. Our simplified technique is more broadly applicable as it allows for conversion to known hash keys for any identifier regardless of originating source.

Aliasing technologies that utilize in-house generated unique identifiers have also been developed (DAVID) [24], however, such systems actually add to the growing number of alias identifiers. Furthermore, these unique identifiers are often assigned to identifiers belonging within computed gene or alias clusters (single-linkage clustering--DAVID), which can restrict the type of clustering that can be performed. In contrast, Booly-hashing does not perform any clustering of identifiers, but rather leaves the process of clustering aliases together to the end user.

Finally, unlike other sequence-related aliasing technologies, we have developed our Booly-hashing infrastructure to accommodate aliases from other sources such as pharmaceutical drugs or keyword aliases. As an example, in the case of pharmaceutical drugs, the unique fingerprint is the chemical formula that remains intact despite multiple branding names. A comparison table of the differences in features among our approach and other aliasing tools can be found in Figure 14b.

In aggregate, our aliasing method allows one to efficiently and accurately ascertain whether two or more identifiers are aliases of each other. Furthermore, our streamlined approach is flexible and easy to modify and update. We have incorporated this aliasing model as part of a necessary component in Booly, our data integration platform designed to aid researchers in making new connections leading to novel discoveries in the laboratory. This generalized aliasing system should be of similar utility for development of other comparative tools that also have the simple requirement of rapidly deciding whether two identifiers are the same.

4.2.3 Booly aliasing complexity

The Booly-hashing alias resolution method utilizes the cryptographic SHA-1 hashing algorithm to create unique fingerprints attached to identifiers. Unlike other alias solutions that require knowledge of the identifier source (AliasServer, SEGUID) or a valid output source (DAVID), our approach accepts any identifier and converts it to all known fingerprint hash keys (Figure 5, Figure 14a). Not having to iterate through all known sources or all possible output sources reduces our run-time complexity by a factor of n as shown below.

Run-time Complexity

Run-time complexity is a useful generalized measure of speed that can be used to compare various algorithms in a hardware independent fashion. Although under certain simplified conditions, run-time complexities of different algorithms can be similar, a discriminating test is to ask how these various algorithms perform in worst case scenarios.

Instance 1. Identity of Source Required.

For the first scenario, the alias application requires knowledge of the identifier's source (SEGUID, AliasServer). In a worse case scenario, we must iterate through each source database (*s*) when our list contains a heterogeneous mix of identifiers (*m*) (e.g. mix of REFSEQ, Gene Name, Entrez ID, etc…). The inner loop executes a total of $m*s$ times resulting in a total complexity of $O(n^2)$.

Pseudocode:

```
for $i \leftarrow 1$ to $m$
    for $j \leftarrow 1$ to $s$
        return fingerprint($i,j$)
```

```
4       done
```

Instance 2. Identity of Output Source Required.

For the second scenario, the alias application requires knowledge of the output source for alias conversion. A common approach to determine whether two aliases refer to the same gene is to convert the aliases into a reference identifier shared between the two. For example, choosing conversion to REFSEQ would convert two identifiers into a common REFSEQ alias. However, there are many instances where aliases may exist in one source but not another, that is, a lack of redundancy across different databases (Table 2). Therefore, one must iterate through all known alias source databases to ensure completeness.

The DAVID gene conversion tool, which has over 1600 citations annually, is one of the most popular resources for gene alias lookups and conversions [24]. The DAVID tool utilizes a single linkage method to cluster similar aliases together into gene groups, assigning a unique identifier for each cluster. However, the DAVID identifier is an incremented value reset after each new update and is therefore not ideal for lasting comparison usage, unlike unique fingerprints such as a sequence hash key. In fact, the DAVID identifier is not available via the public query web interface but rather only through a bulk download request. Therefore, if we bypass the unique identifiers supplied by the DAVID conversion tool due to its limitations as described above, and utilize DAVID to convert heterogeneous identifiers (*p*) into all known output source identifiers (*t*), then this tool can be classified within instance 2, which also has a time complexity of $O(n^2)$.

Pseudocode:

```
for $i \leftarrow 1$ to $p$
    for $j \leftarrow 1$ to $t$
        return identifier($i,j$)
done
```

Instance 3. Booly Method.

Our method does not require the identifier's source database or the output source database. The method simply iterates through every identifier (*b*) and performs a lookup of its fingerprint hash-key resulting in a run-time complexity of $O(n)$.

Pseudocode:

```
for $i \leftarrow 1$ to $b$
    return fingerprint($i$)
done
```

4.2.4 Alias clusters

Unlike other aliasing methods (DAVID, HGNC), the Booly Aliasing approach is generalizable to diverse types of data since it can create clusters of aliases beyond just gene clusters. One particular usage is to be able to distinguish between different protein variants yet still cluster aliases of the same protein. For example, our approach can distinguish between protein variants *spir*-PA (NP_724254, FBpp0080884) and *spir*-PB (NP_524854, FBpp0080885) of the *D. melanogaster* spire gene since each variant has a different unique sequence hash key while aliases (e.g. *spir*-PA/NP_724254/ FBpp0080884) have the same sequence hash key. Our approach is sufficiently flexible to create gene clusters by labeling genes with their protein sequence hash keys in addition to the gene sequence hash keys. It can also be used to resolve identifiers of data such as aliases of chemical structure names, which can then be used in combination with gene identifiers to create associations between drugs and genetic disease phenotypes [40].

4.2.5 Booly-hashing creation and updates

The Booly Alias database table is a simple three column table consisting of 1) a variable sized column for the identifier, 2) a fixed sized (40) column for the hash key, and 3) a timestamp column. Additionally, a uniqueness constraint is placed on column 1 and 2 (identifier and sequence are the same) to avoid duplicate entries.

During the creation and update steps of our Alias table, FASTA formatted files are processed for sequences and associated identifiers. Sequences free of spaces and gaps are converted to upper case letters, submitted to the database SHA-1 function, and deposited row by row attached to associated identifiers. The entire process is fast, efficient, and easily replicated.

4.3 Booly infrastructure

Specialized databases such as PUBMED or GenBank have been optimized in such a way that one can, for example, rapidly retrieve published references or sequence information. This specialization requires unique organization and entry methods to accommodate the differences in each database. Our goal has been to create an infrastructure such that any web based content can be stored inside our system without the overhead of more specialized databases; i.e., we want to be able to store PUBMED entries, GenBank sequence entries, images--all types of web-based data, using a singular database schema (Figure 15).

4.3.1 Keys and values

At the heart of our singular database schema is the key to value relationship. The “key”, or identifier, is simply a label for each row of data while the “value” is the actual row of data. Keys can be gene names, protein names, or any symbolic string (50 maximum). Values on the other hand, are data that do not have length constraints. Values can be any length of text, including HTML. The use of HTML allows one to create nested tables, multiple columns, hyperlinks, and numerous formatting options. The minimal requirements of a key to value relationship in any piece of data added makes Booly amenable to both diverse current and future data storage needs.

4.3.2 User data management

The most dynamic component of Booly is the ability of users to create their own content and store this information on the server. To keep track of each user's data, we integrated an open source forum based login system (phpBB) into Booly[42]. This feature offers not only a reliable account management system, but also a forum for communication with developers and other Booly members. Users can share datasets while also having access to any datasets made publicly available by other members. For example, users that have access to a list of *C. elegans* RNAi feeding libraries can easily submit their dataset to the public repository for others to use and perform Boolean queries against their list of *C. elegans* genes. If privacy is a concern, one can restrict sharing of datasets to colleagues through private email rather than by submitting the data to the open access public repository. These lists can then be added for comparison to the publicly available data from the users own local terminal.

A foremost concern in the construction of Booly is its potential to grow exponentially in size as more and more users store data. Although space concerns can be remedied by hardware upgrades, the speed at which user data is accessed can degrade as the database grows. To address this issue, a horizontal partitioning scheme is used to separate users into different groups. In this manner, the load can be spread across different locations (i.e. across different tables as well as different hardware).

4.3.3 Query interface

The user constructs a Boolean operation by ordering datasets in an appropriate sequence and placing Boolean commands between each dataset (Figure 2). For example, the "OR" disjunction operation is constructed by adding the "or" command between two datasets, while the "NOT" negation operation requires addition of the "not" command between them. By default, the "AND" conjunction command is inferred if no Boolean operators are specified between two datasets. Therefore, adding two datasets to a list without any operators would result in output containing entries that exist in both datasets.

4.3.4 Exporting results and external applications

Another important function that we have incorporated into Booly is the ability to export results once a Boolean query has been performed. Two options exist when exporting: 1) a local save (as an xml file or html file) on a user's personal computer, or 2) a remote save back onto Booly as an entirely new data set. The latter ability to create a new data set within Booly opens up new avenues of data integration. We briefly describe two applications, switching keys and keyword filtering, which both take advantage of exporting results as a new Booly dataset.

Switching Keys

The “key” in a key-value relationship is significant in Booly since it is the key that is used when comparing datasets against one another using Boolean logic. However, in some instances, the values may actually contain identifiers within them that could be used as keys for further Boolean operations (Figure 11). By allowing users to switch keys during the export of Boolean results, one can chain together datasets and rename keys with identifiers derived from text within the values. This creates the possibility of integrating data within values and not just the keys. The powerful combination of switching keys and chaining together concatenated series of Boolean queries allows users to make sophisticated links between otherwise unconnected datasets.

Keyword Filter

Another feature to increase the value of exporting results of a Booly merge is the ability to filter out those results based on a keyword. An example of this usage is the ability to search for results that have gene ontology involved in immune response after a dataset of gene ontology has been merged with gene expression data. Users can search for key words and then export these results as a new Booly dataset.

Further Integration Using External Applications

An important aspect of Booly is the ability for users to store content such as output from a computer program. For example, we have developed a tool that allows users to perform BLAST comparisons and store its results directly inside Booly. The output of the job is tailored specifically for Booly and stored in the user’s account, allowing the user to retrieve and compare other datasets to the newly created BLAST dataset. We plan to provide a variety of additional plug-in modules in the form of a Web API for developers in the near future. By allowing applications to directly submit output into Booly for data storage, users of the application would receive integration of their generated data with other applications as well as the ability to perform Boolean operations and alias resolution.

Chapter 4, in part, has been submitted for publication of the material as it may appear in Booly: a new data integration platform, Do, Long H.; Esteves, Francisco F.; Karten, Harvey J.; Bier, Ethan, BMC Bioinformatics 2010. The dissertation author was the primary investigator and author of this paper

Chapter 4.2, in part, is currently being prepared for submission for publication of the material. Do, Long H. and Bier, Ethan. The dissertation author was the primary investigator and author of this material.

Chapter 5

Summary and Conclusions

5.1 Summary

The growing volume of biological and medical information deposited within disparate databases has created an organization and data integration dilemma within the research community. Furthermore, new data not configuring to pre-existing specialized databases must await creation of new dedicated inclusive databases. We have created a novel tool, Booly, as a web application that solves key problems impeding current data integration efforts. An important feature of this system is a real time alias translation system, which we used to successfully integrate datasets with heterogeneous identifiers between Ensembl, gene symbols and gene ontology. Secondly, we addressed the issue of the entry barrier by creating an easy to use contribution model for both developers and researchers. Users are able to easily add datasets by copying and pasting their spreadsheet tables or by utilizing applications designed to create new Booly datasets. Lastly, we showed how Booly could be used as an intermediate step in data mining and data integration through our implementation of the switching and chaining technique to change "touch points".

There are a myriad of other enabling applications for Booly. For example, as personalized genomes become available to the general population, Booly is poised to offer individuals space to house their biological and medical information such that it can also be used to compare with publicly available content in a safe and secure fashion. Booly is also a resource for developers to add content without the obstacle of creating an online storage facility or the troublesome nature of alias resolution. Booly thus offers a fundamentally new paradigm for storing, sharing, and integrating current and future health and biological content.

5.2 Future Directions

Boolean modeling is a formal description of a broad array of biological phenomena, one notable example being gene regulation [43]. To this extent, many biological processes can be modeled by using Boolean Networks. Booly offers an important functionality for system level studies as it greatly facilitates integration of diverse datasets from multiple experimental sources, providing the first step in gathering data into a Boolean model. Further development of algorithms that apply networking or clustering of touch points within the groupings created by Booly could similarly lead to novel systems based hypotheses.

Booly offers a Uniform Resource Locator (URL) based web API, allowing developers to easily integrate their applications and datasets into Booly. In this manner, developers will be able to create their tool or database and use Booly as a repository for the tool’s output. For example, an external database may allow users to directly download all of the results from a search and place them directly into the user’s Booly account. The output generated from these tools, once placed inside Booly, will inherit all its functionality, including the ability to easily share the data, to perform Boolean logic comparisons with other data sets, and to resolve aliases.

An obvious concern for comprehensive databases, and thus for Booly, is the issue of scalability. That is, how will Booly deal with the exponential growth of data deposited into its systems? For example, as personalized genomes become a reality, as is currently being implemented in the 1000 Genomes Initiative [44], a means for an individual to store and explore this information will be highly desirable. We have created Booly in such a way that as the data grows, additional machines can be introduced in parallel into the system for load balancing and data partitioning without adversely affecting the Booly’s efficiency (speed) and reliability (uptime).

A large component of Booly is the user contribution model as similarly applied to such online applications as Wikipedia and more relevantly, WikiGene [45]. However, a major concern is quality control of user-contributed data. Our plan to address this dilemma is to implement a community based review system for each dataset (similar to Amazon product ratings). In this manner, users will be able to search and add datasets based on “collective intelligence”, a key element of Web 2.0 [46].

With most if not all data integration platforms, there is a concern that data can quickly grow out of date and require updating. The goal for Booly is to become a publically driven data repository reviewed and updated by its community. To aid the community in their efforts, we hope to implement a notification system such that when a new dataset is available, subscribers to the old dataset will be notified and allowed to upgrade or add the new dataset. In the meantime, we have created a forum message board so that contributors can disseminate update information to the community.

Finally, there is a growing movement in the life sciences to develop tools for semantic integration by way of the RDF model [17, 47]. Semantic integration approaches involve establishing complex relationships and meanings between objects, which can then be used to classify them or extract novel information regarding their behaviors. The goal of Booly is more modest, to establish identity between objects and to use this information to integrate data in which distinct names refer to the same objects. We felt our initial challenge was to help researchers and developers get their data quickly onto the web and to address the identity problem directly. However, to aid in interoperability with other data integration efforts that utilize RDF and other semantic integration approaches, we plan to provide export of data into a structured model such as RDF. It is our hope that the streamlined but efficient and user friendly comparative tools offered by Booly attract a broad base of users who are confronted by the simple but vexing problem of integrating data from a diverse set of spread sheets. Such users once adept at using Booly would presumably be primed to expand their sphere of comparison by trying out new tools such as those offered by semantic integration approaches.

5.3 Conclusions

Booly offers a new platform for the creation, storage, and integration of both personalized and public biological databases. As more applications are developed around the Booly platform, we anticipate these additions will further enhance the user experience. Booly presents a great opportunity to engage the research community in sharing data and adding combinatorial depth to potential queries. Such advances as offered by Booly should greatly aid researchers in formulating new questions that lead to novel discoveries in the laboratory.

Chapter 5, in part, has been submitted for publication of the material as it may appear in Booly: a new data integration platform, Do, Long H.; Esteves, Francisco F.; Karten, Harvey J.; Bier, Ethan, BMC Bioinformatics 2010. The dissertation author was the primary investigator and author of this paper.

Appendix

Figures and Tables

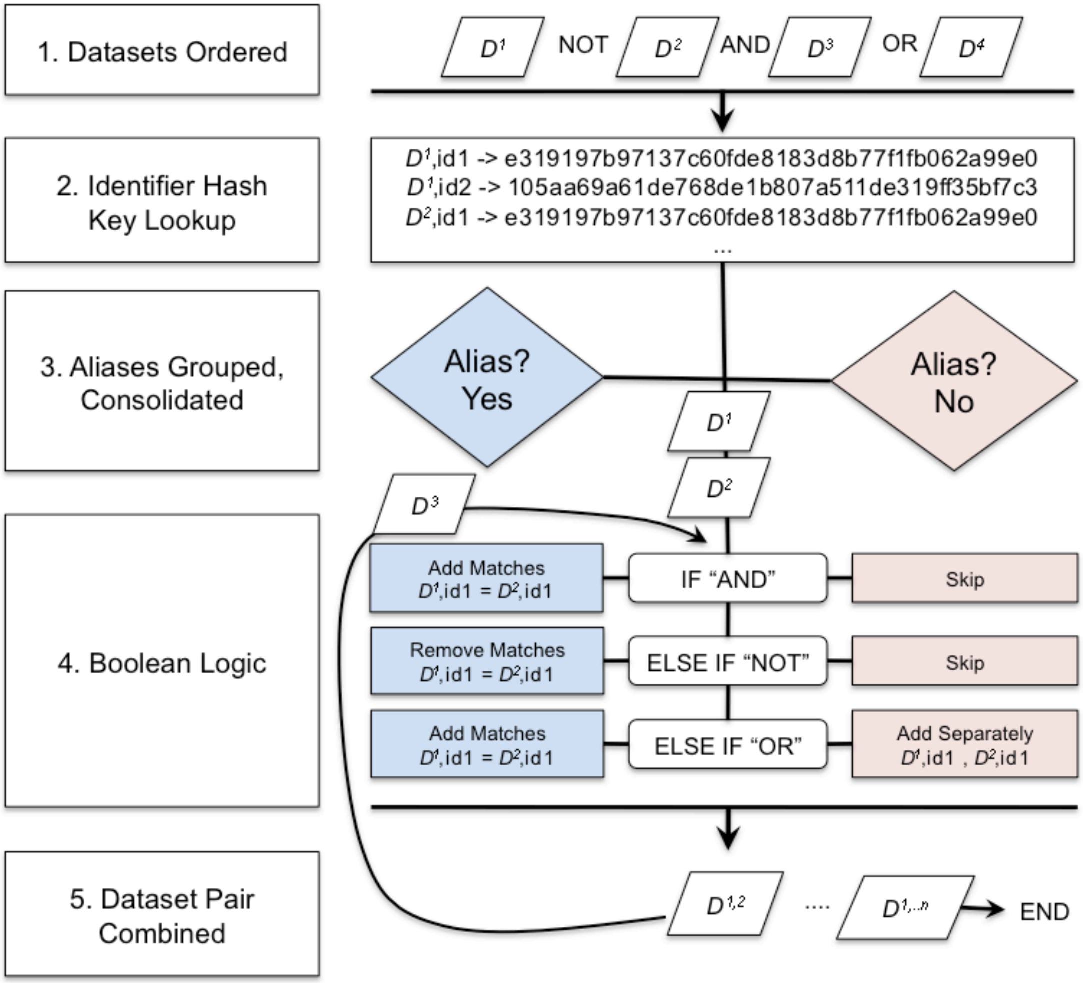
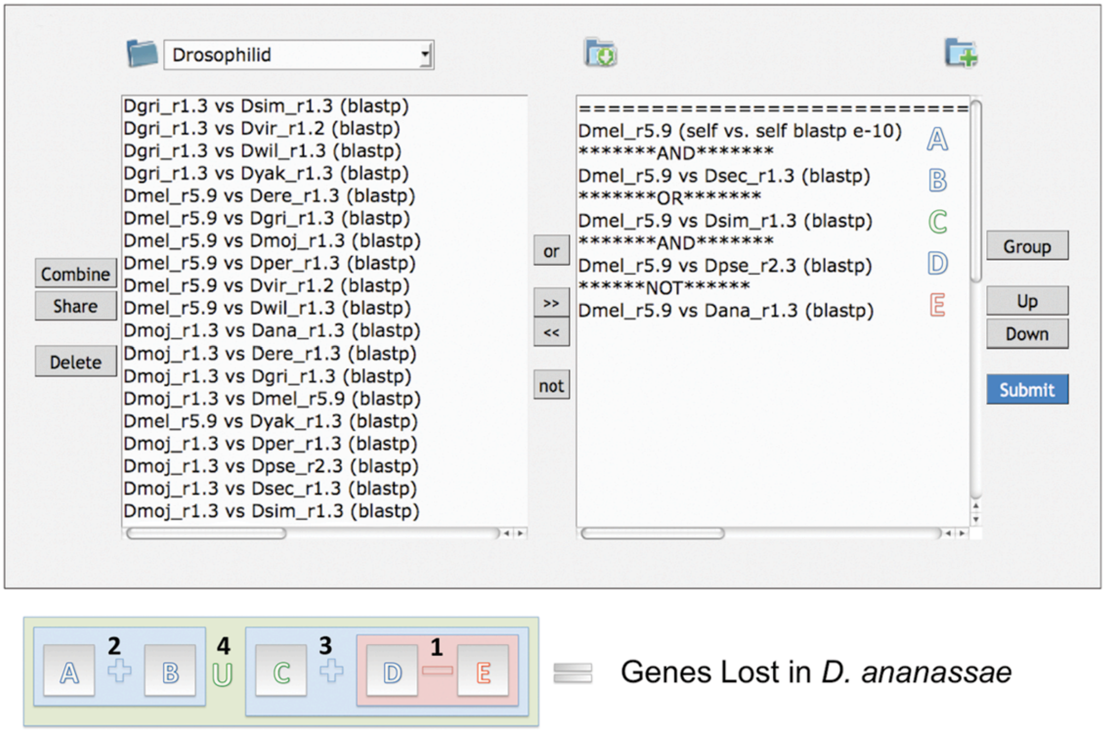

1. Datasets Ordered
2. Identifier Hash Key Lookup
3. Aliases Grouped, Consolidated
4. Boolean Logic
5. Dataset Pair Combined
D1 NOT D2 AND D3 OR D4
D1,id1 -> e319197b97137c60fde8183d8b77f1fb062a99e0
D1,id2 -> 105aa69a61de768de1b807a511de319ff35bf7c3
D2,id1 -> e319197b97137c60fde8183d8b77f1fb062a99e0
...
Alias? Yes
Alias? No
D1
D2
D3
Add Matches D1,id1 = D2,id1
IF "AND"
Skip
Remove Matches D1,id1 = D2,id1
ELSE IF "NOT"
Skip
Add Matches D1,id1 = D2,id1
ELSE IF "OR"
Add Separately D1,id1 , D2,id1
D1,2
....
D1,...n
END

Figure 1. The Booly data integration algorithm. Overview of the steps involved in performing a Booly query with aliasing. Access Booly at http://booly.ucsd.edu.



```
Dgri_r1.3 vs Dsim_r1.3 (blastp)
Dgri_r1.3 vs Dvir_r1.2 (blastp)
Dgri_r1.3 vs Dwil_r1.3 (blastp)
Dgri_r1.3 vs Dyak_r1.3 (blastp)
Dmel_r5.9 vs Dere_r1.3 (blastp)
Dmel_r5.9 vs Dgri_r1.3 (blastp)
Dmel_r5.9 vs Dmoj_r1.3 (blastp)
Dmel_r5.9 vs Dper_r1.3 (blastp)
Dmel_r5.9 vs Dvir_r1.2 (blastp)
Dmel_r5.9 vs Dwil_r1.3 (blastp)
Dmoj_r1.3 vs Dana_r1.3 (blastp)
Dmoj_r1.3 vs Dere_r1.3 (blastp)
Dmoj_r1.3 vs Dgri_r1.3 (blastp)
Dmoj_r1.3 vs Dmel_r5.9 (blastp)
Dmel_r5.9 vs Dyak_r1.3 (blastp)
Dmoj_r1.3 vs Dper_r1.3 (blastp)
Dmoj_r1.3 vs Dpse_r2.3 (blastp)
Dmoj_r1.3 vs Dsec_r1.3 (blastp)
Dmoj_r1.3 vs Dsim_r1.3 (blastp)
```

```
============================
Dmel_r5.9 (self vs. self blastp e-10)    A
*******AND*******
Dmel_r5.9 vs Dsec_r1.3 (blastp)          B
*******OR*******
Dmel_r5.9 vs Dsim_r1.3 (blastp)          C
*******AND*******
Dmel_r5.9 vs Dpse_r2.3 (blastp)          D
******NOT******
Dmel_r5.9 vs Dana_r1.3 (blastp)          E
```

Figure 2. Illustration of Booly list form and Boolean logic precedence. The right textbox depicts a list of datasets ready for Boolean operations. Numbered cartoon demonstrates the order of operations performed for this query. This query identifies genes lost in *D. ananassae* but are retained in the *melanogaster* subgroup and in the outgroup *D. pseudoobscura* (see Suppl. Figs. S4a, S4b). Booly Precedence: 1) Group Selection Using Parenthesis. 2) NOT/Conjunction (-) Operation. 3) AND/Intersection (+) Operation. 4) OR/Union (U) Operation. Precedence for multiple instances of the same operator is determined by the order in which they appear in the query.

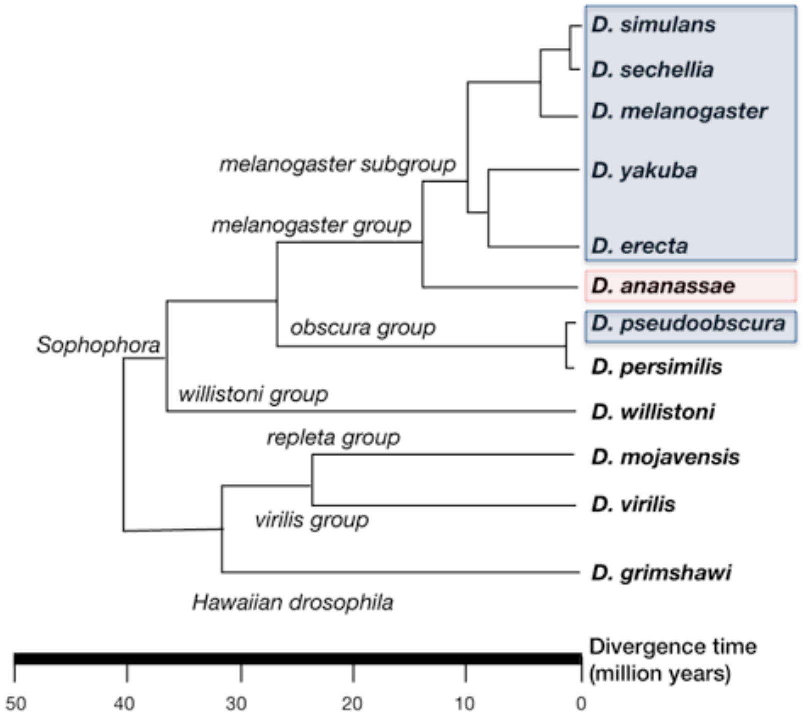


Figure 3. Twelve *Drosophila* genomes. *Drosophila* genomes that have been sequenced and an associated divergence timeline (http://rana.lbl.gov/drosophila). We subtracted genes of *D. ananassae* (red) from the subset of genes found in the genomes of the *melanogaster* subgroub and the outgroup *D. pseudoobscura* (blue).

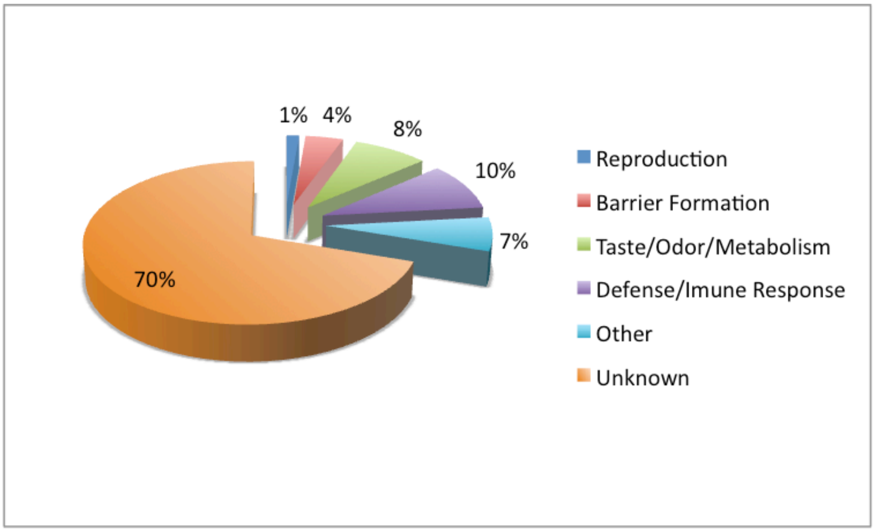


Figure 4. Genes lost selectively in *D. ananassae*. We identified over 73 genes that were lost during evolution of the *Drosophila ananassae lineage* that were retained in the sister melanogaster subgroup comprised of *D. melanogaster, D. simulans, D. sechelia, D. yakuba*, and *D. errecta* and in the outgroup *D. pseudoobscura*. Annotated lost genes fall into the same major functional classes as those that are found to be enriched among species-specific genes. The results of this query can be accessed at: http://booly.ucsd.edu/dana-lost.

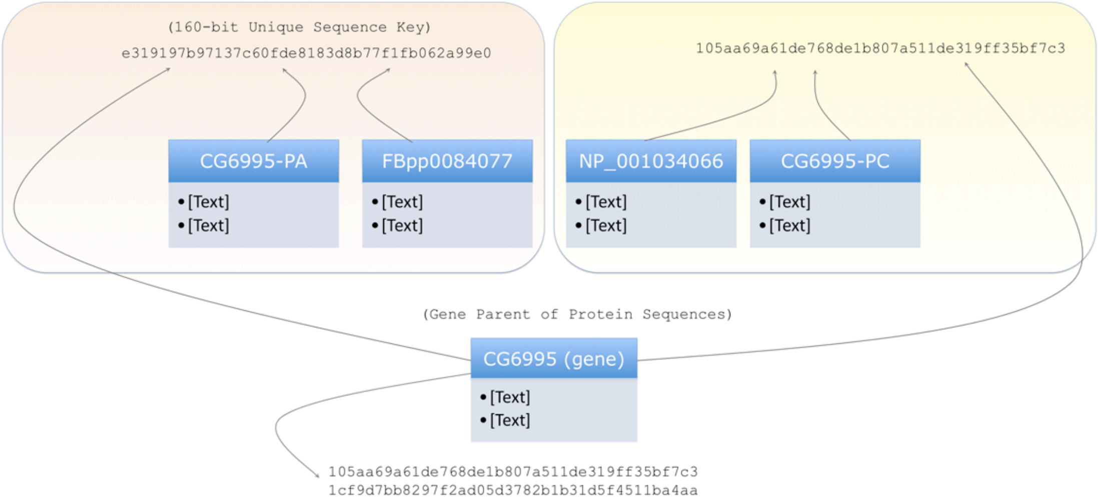

(160-bit Unique Sequence Key)
e319197b97137c60fde8183d8b77f1fb062a99e0
105aa69a61de768de1b807a511de319ff35bf7c3
CG6995-PA
• [Text]
• [Text]
FBpp0084077
• [Text]
• [Text]
NP_001034066
• [Text]
• [Text]
CG6995-PC
• [Text]
• [Text]
(Gene Parent of Protein Sequences)
CG6995 (gene)
• [Text]
• [Text]
105aa69a61de768de1b807a511de319ff35bf7c3
1cf9d7bb8297f2ad05d3782b1b31d5f4511ba4aa

Figure 5. Alias resolution of heterogeneous identifiers. In the following example, the protein variants CG6995-PA and CG6995-PC have aliases FPpp00084077 and NP_001034066, respectively. When joined by a Boolean operation, the variants are kept separate due to having different unique sequence keys. However, if the proteins are joined with a list containing the gene parent (CG6995), the entire group is merged together if the gene has aliases that point to the protein variants.

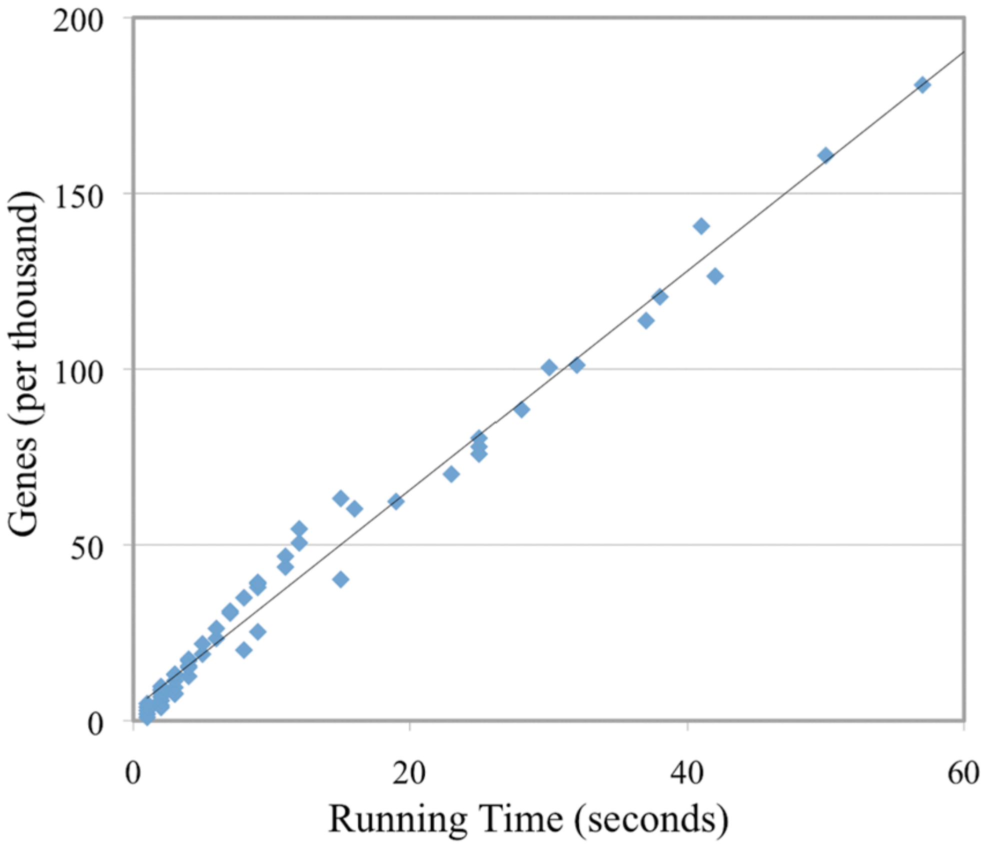
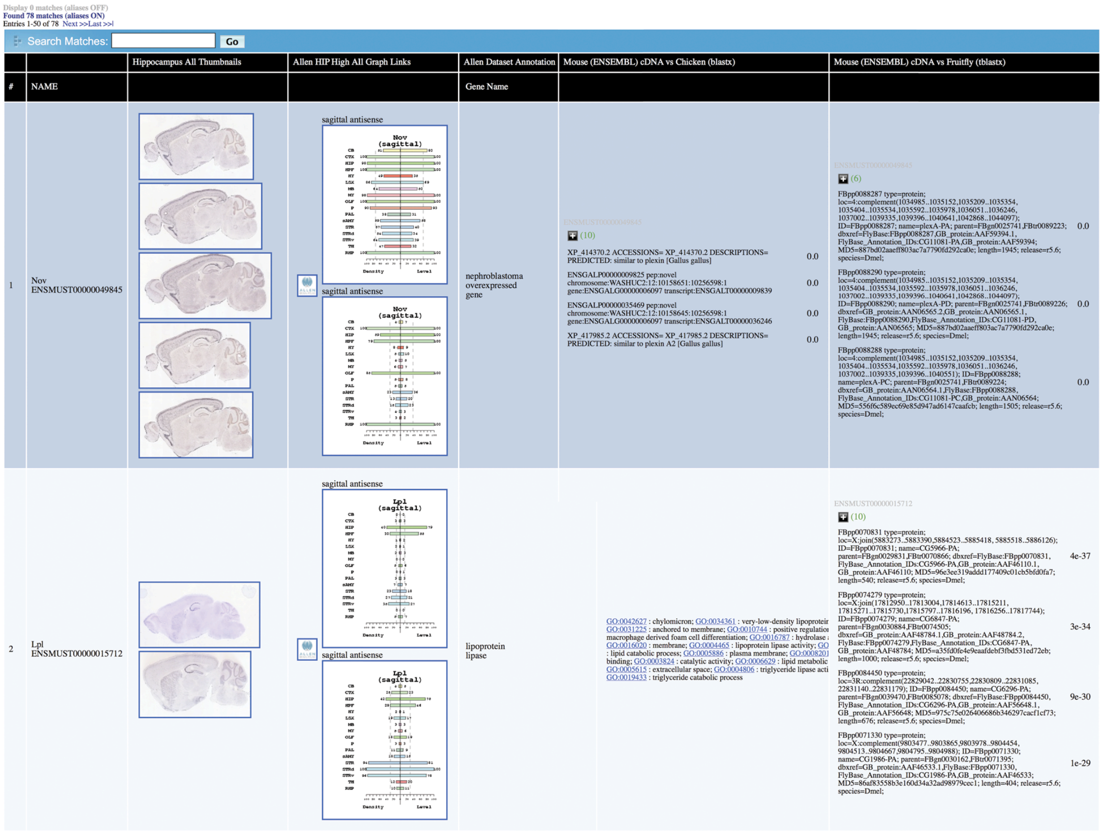


Figure 6. **Booly running time**. Approximate running time performing a Booly intersection with alias resolution. Y-axis contains the total number of combined identifiers (e.g. genes) for every dataset in a Booly merge. Plot shown represents 200,000 genes (10 intersections of datasets containing 20,000 genes apiece).



Display 0 matches (aliases OFF)
Found 78 matches (aliases ON)
Entries 1-50 of 78 Next >>Last >>|

Search Matches: Go

		Hippocampus All Thumbnails	Allen HIP High All Graph Links	Allen Dataset Annotation	Mouse (ENSEMBL) cDNA vs Chicken (blastx)		Mouse (ENSEMBL) cDNA vs Fruitfly (tblastx)	
#	NAME			Gene Name				

1	Nov
	ENSMUST00000049845
			sagittal antisense
			sagittal antisense
				nephroblastoma overexpressed gene	ENSMUST00000049845
					(10)
					XP_414370.2 ACCESSIONS= XP_414370.2 DESCRIPTIONS= PREDICTED: similar to plexin [Gallus gallus]
					ENSGALP00000009825 pep:novel chromosome:WASHUC2:12:10158651:10256598:1 gene:ENSGALG00000006097 transcript:ENSGALT00000009839
					ENSGALP00000035469 pep:novel chromosome:WASHUC2:12:10158645:10256598:1 gene:ENSGALG00000006097 transcript:ENSGALT00000036246
					XP_417985.2 ACCESSIONS= XP_417985.2 DESCRIPTIONS= PREDICTED: similar to plexin A2 [Gallus gallus]
						0.0
						0.0
						0.0
						0.0
							ENSMUST00000049845
							(6)
							FBpp0088287 type=protein; loc=4:complement(1034985..1035152,1035209..1035354, 1035404..1035534,1035592..1035978,1036051..1036246, 1037002..1039335,1039396..1040641,1042868..1044097); ID=FBpp0088287; name=plexA-PA; parent=FBgn0025741,FBtr0089223; dbxref=FlyBase:FBpp0088287,GB_protein:AAF59394.1, FlyBase_Annotation_IDs:CG11081-PA,GB_protein:AAF59394; MD5=887bd02aaeff803ac7a7790fd292ca0e; length=1945; release=r5.6; species=Dmel;
							FBpp0088290 type=protein; loc=4:complement(1034985..1035152,1035209..1035354, 1035404..1035534,1035592..1035978,1036051..1036246, 1037002..1039335,1039396..1040641,1042868..1044097); ID=FBpp0088290; name=plexA-PD; parent=FBgn0025741,FBtr0089226; dbxref=GB_protein:AAN06565.2,GB_protein:AAN06565.1, FlyBase:FBpp0088290,FlyBase_Annotation_IDs:CG11081-PD, GB_protein:AAN06565; MD5=887bd02aaeff803ac7a7790fd292ca0e; length=1945; release=r5.6; species=Dmel;
							FBpp0088288 type=protein; loc=4:complement(1034985..1035152,1035209..1035354, 1035404..1035534,1035592..1035978,1036051..1036246, 1037002..1039335,1039396..1040551); ID=FBpp0088288; name=plexA-PC; parent=FBgn0025741,FBtr0089224; dbxref=GB_protein:AAN06564.1,FlyBase:FBpp0088288, FlyBase_Annotation_IDs:CG11081-PC,GB_protein:AAN06564; MD5=556f6c589ec69e85d947ad6147caafcb; length=1505; release=r5.6; species=Dmel;
								0.0
								0.0
								0.0

2	Lpl
	ENSMUST00000015712
			sagittal antisense
			sagittal antisense
				lipoprotein lipase	GO:0042627 : chylomicron; GO:0034361 : very-low-density lipoprotei GO:0031225 : anchored to membrane; GO:0010744 : positive regulatio macrophage derived foam cell differentiation; GO:0016787 : hydrolase GO:0016020 : membrane; GO:0004465 : lipoprotein lipase activity; GO : lipid catabolic process; GO:0005886 : plasma membrane; GO:000820 binding; GO:0003824 : catalytic activity; GO:0006629 : lipid metabolic GO:0005615 : extracellular space; GO:0004806 : triglyceride lipase acti GO:0019433 : triglyceride catabolic process		ENSMUST00000015712
							(10)
							FBpp0070831 type=protein; loc=X:join(5883273..5883390,5884523..5885418, 5885518..5886126); ID=FBpp0070831; name=CG5966-PA; parent=FBgn0029831,FBtr0070866; dbxref=FlyBase:FBpp0070831, FlyBase_Annotation_IDs:CG5966-PA,GB_protein:AAF46110.1, GB_protein:AAF46110; MD5=96e3ee319addd177409c01cb5bfd0fa7; length=540; release=r5.6; species=Dmel;
							FBpp0074279 type=protein; loc=X:join(17812950..17813004,17814613..17815211, 17815271..17815730,17815797..17816196, 17816256..17817744); ID=FBpp0074279; name=CG6847-PA; parent=FBgn0030884,FBtr0074505; dbxref=GB_protein:AAF48784.1,GB_protein:AAF48784.2, FlyBase:FBpp0074279,FlyBase_Annotation_IDs:CG6847-PA, GB_protein:AAF48784; MD5=a35fd0fe4e9eaafdebf3fbd531ed72eb; length=1000; release=r5.6; species=Dmel;
							FBpp0084450 type=protein; loc=3R:complement(22829042..22830755,22830809..22831085, 22831140..22831179); ID=FBpp0084450; name=CG6296-PA; parent=FBgn0039470,FBtr0085078; dbxref=FlyBase:FBpp0084450, FlyBase_Annotation_IDs:CG6296-PA,GB_protein:AAF56648.1, GB_protein:AAF56648; MD5=975c75e026406686b346297cacf1cf73; length=676; release=r5.6; species=Dmel;
							FBpp0071330 type=protein; loc=X:complement(9803477..9803865,9803978..9804454, 9804513..9804667,9804795..9804988); ID=FBpp0071330; name=CG1986-PA; parent=FBgn0030162,FBtr0071395; dbxref=GB_protein:AAF46533.1,FlyBase:FBpp0071330, FlyBase_Annotation_IDs:CG1986-PA,GB_protein:AAF46533; MD5=86af83558b3e160d34a32ad98979cec1; length=404; release=r5.6; species=Dmel;
								4e-37
								3e-34
								9e-30
								1e-29

Figure 7. A Booly query combining disparate datasets utilizing on-demand alias resolution. Datasets merged in this Boolean query include annotation data, mouse brain expression summary graphs and *in-situ* thumbnails (Allen Mouse Brain Atlas), and BLAST summaries of mouse against the chicken, fish, and fruitfly. The results of this query can be accessed at: http://booly.ucsd.edu/hippocampus and a higher resolution version of the image can be downloaded at: http://booly.ucsd.edu/figures/Allen_results.jpg.

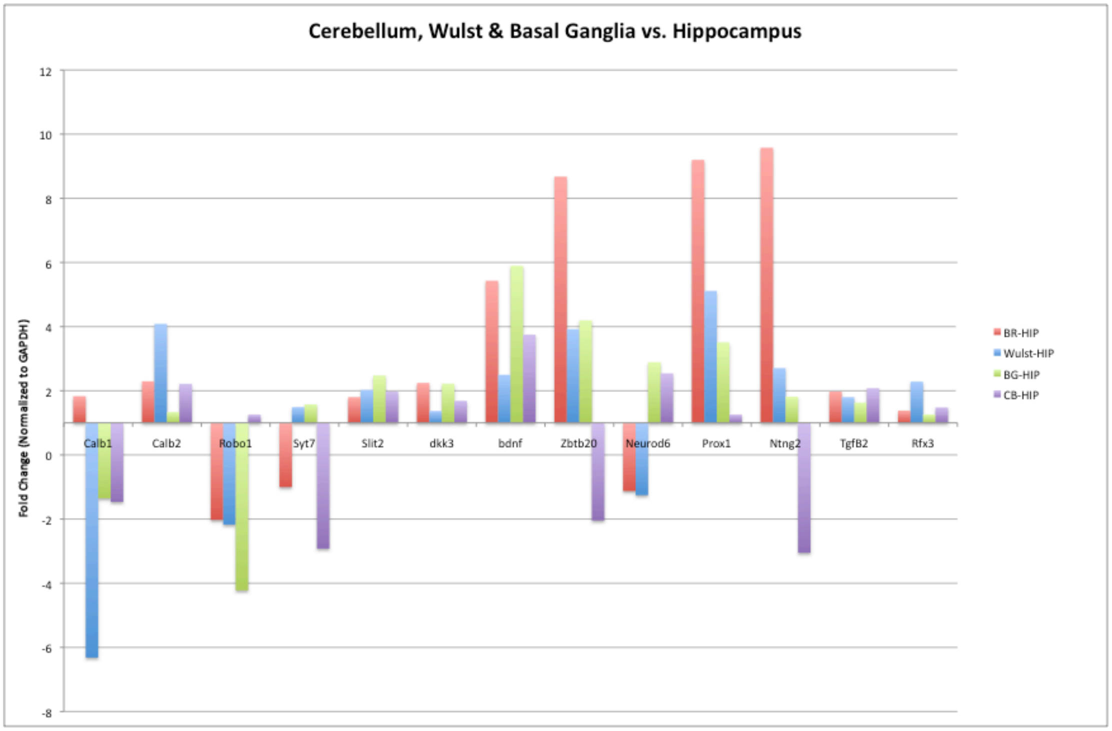


Figure 8. RT-PCR analysis of genes predicted to be enriched in the avian hippocampus. Relative fold change of selected genes in the Hippocampus and other areas of the Chick Brain, based on two or more individual independent experiments, which were highly concordant. Housekeeping genes GAPDH and actin were used as controls for normalization as described in the Methods section.

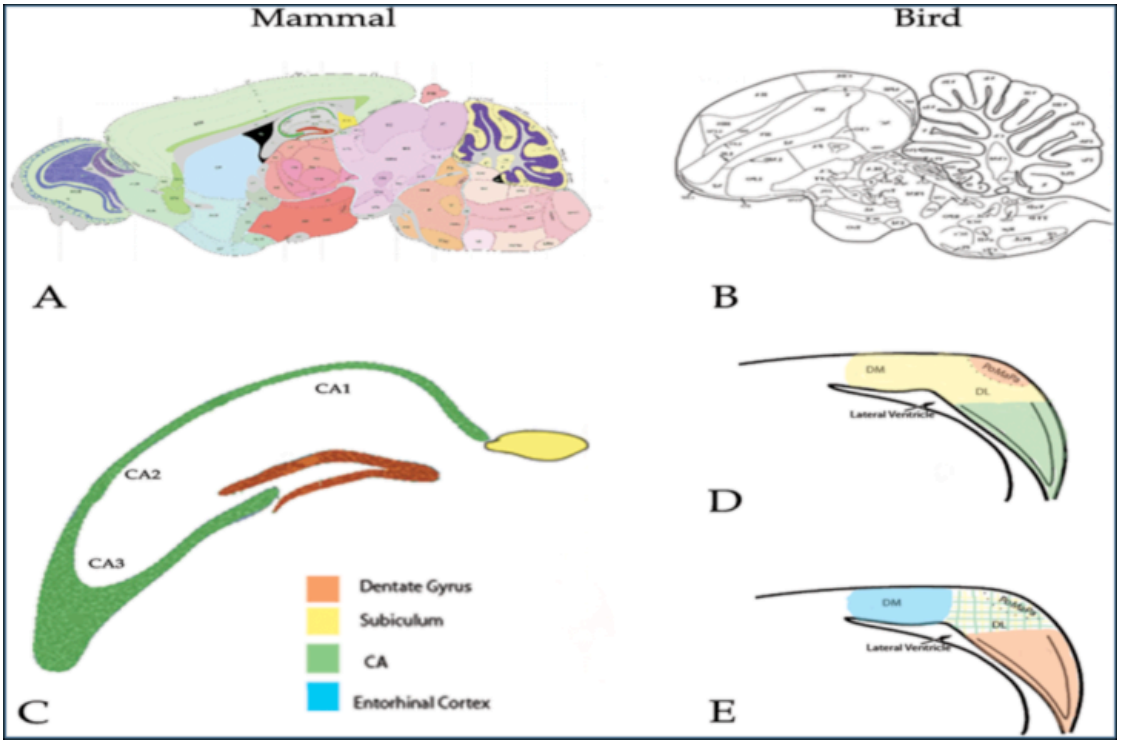


Figure 9. Hippocampus of birds and mammals. In mammals (A,C) the hipocampus can be divided into two broad sub regions the CA fields and the dentate gyrus based on citoarchitecture alone, while in birds (B,D,E) this is not the case since there are no prominent citoarchitectural features and in fact, the organization of the subdomains in the avian hippocampus remains unclear with opposed views (D, E).

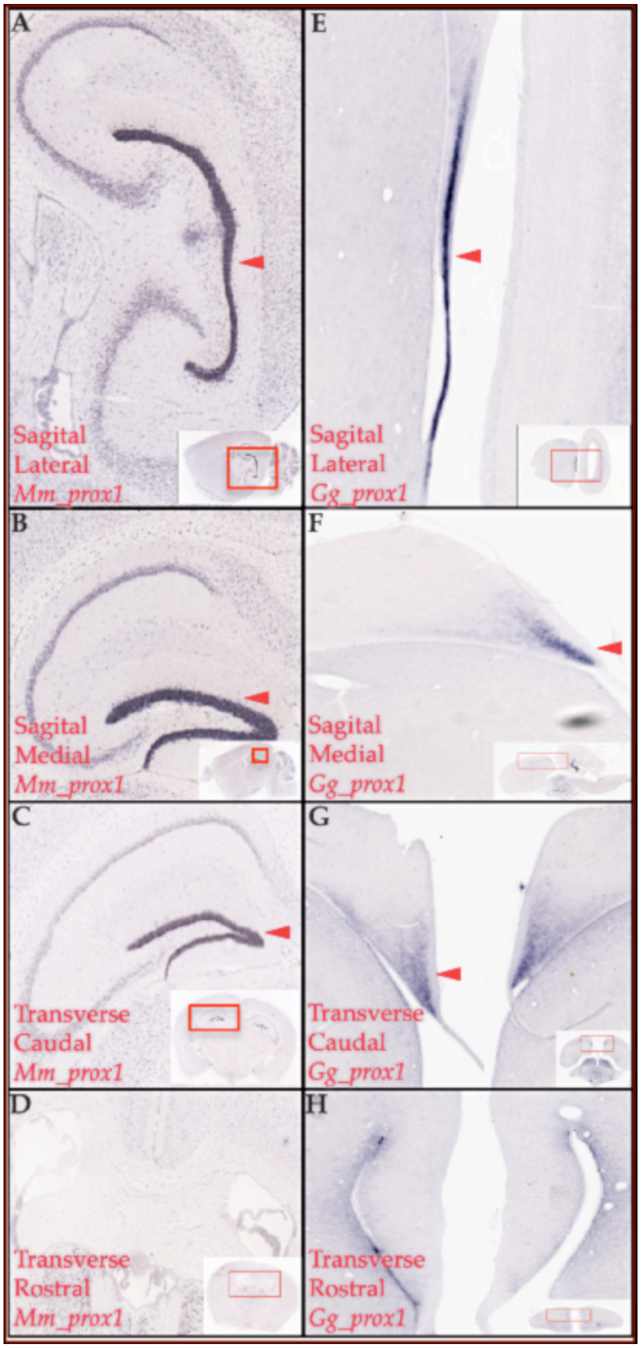
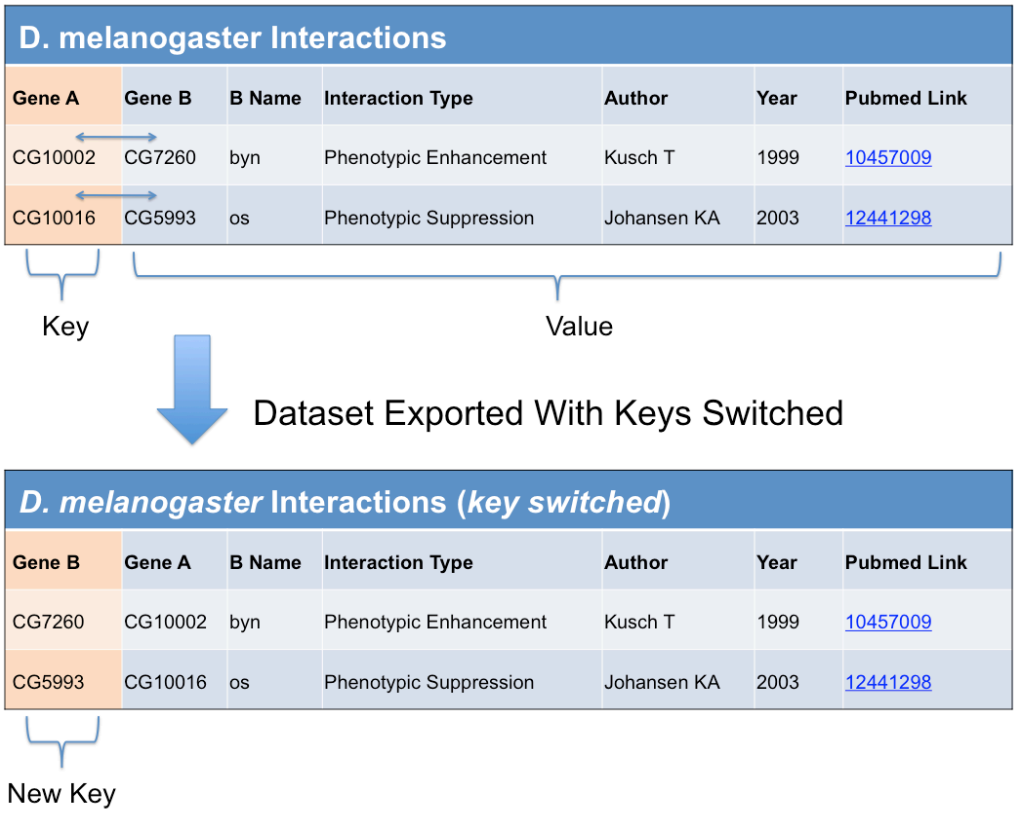


Figure 10. ***prox1* is a reliable marker for the Dentate Gyrus**. In the mouse 8 week old brain (B,C,D), *prox1* is easily distinguished in the Dentate Gyrus and predictably disappears when a rostral section missing the hippocampus is probed (D). Equivalently, the 1-2 day old chick brain shows homologous *prox1* patterns in the cell dense "V" shaped area of the presumed avian hippocampus. F. Esteves *et al.* (unpublished).



D. melanogaster Interactions

Gene A	Gene B	B Name	Interaction Type	Author	Year	Pubmed Link
CG10002	CG7260	byn	Phenotypic Enhancement	Kusch T	1999	10457009
CG10016	CG5993	os	Phenotypic Suppression	Johansen KA	2003	12441298

D. melanogaster Interactions (*key switched*)

Gene B	Gene A	B Name	Interaction Type	Author	Year	Pubmed Link
CG7260	CG10002	byn	Phenotypic Enhancement	Kusch T	1999	10457009
CG5993	CG10016	os	Phenotypic Suppression	Johansen KA	2003	12441298

Figure 11. Exporting Results and Switching Keys. Example of exporting and saving as a new dataset in Booly. A new key is assigned for each row by taking an identifier within the "value" field.

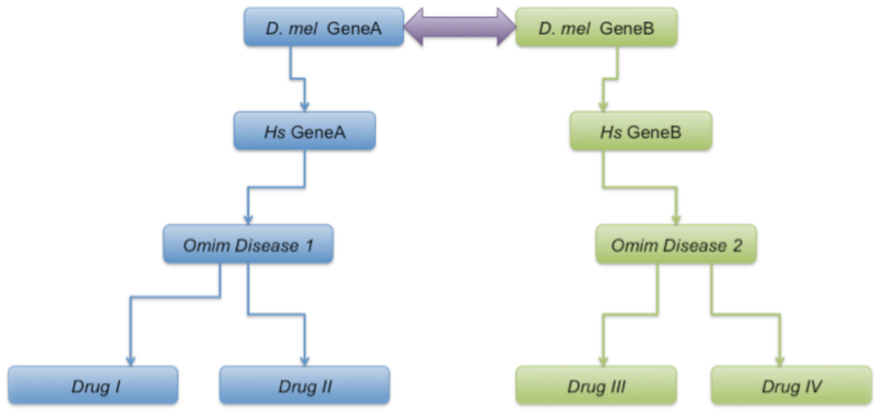


Figure 12. Linking Drugs to Interaction Networks. An example of a complex, chained Boolean comparison is the identification of new diseases that might be treated by FDA approved drugs currently used to treat a different disease. The idea is to first link a list of FDA approved drugs to diseases they can treat, then to associate genes with these diseases based on mutations in these genes causing phenotypes similar to the diseases treated by drugs, then linking these human disease genes to homologous genes in the fruit fly, then to broaden this list of genes to those interacting genetically with mutations in the fly gene homolog, then to ask whether any of the interacting fly genes have human homologs that also lead to disease, and finally to ask whether these potentially related human diseases might also be treatable with drugs used for the first disease, and vice-versa. The results of this query can be accessed at: http://booly.ucsd.edu/drug-networks.

Figure 13. Switching Keys and Chaining Boolean Queries. An example of switching “touch-points” so that two separate diseases and their associated drugs can be integrated within an interaction network found in *D. melanogaster* (*fkh* and *bkn*).

Dataset 1. Disease and drugs mined from FDA

Dataset 2. Human genes with allelic variants that have association with disease, mined from OMIM

Results of Booleome DB intersection (AND) from Datasets 1 and 2.

Dataset 3. We switch the identifier to the human gene and save as a new dataset (3).

Dataset 4. A list of BLAST hits from human genes to genes from the fruitfly.

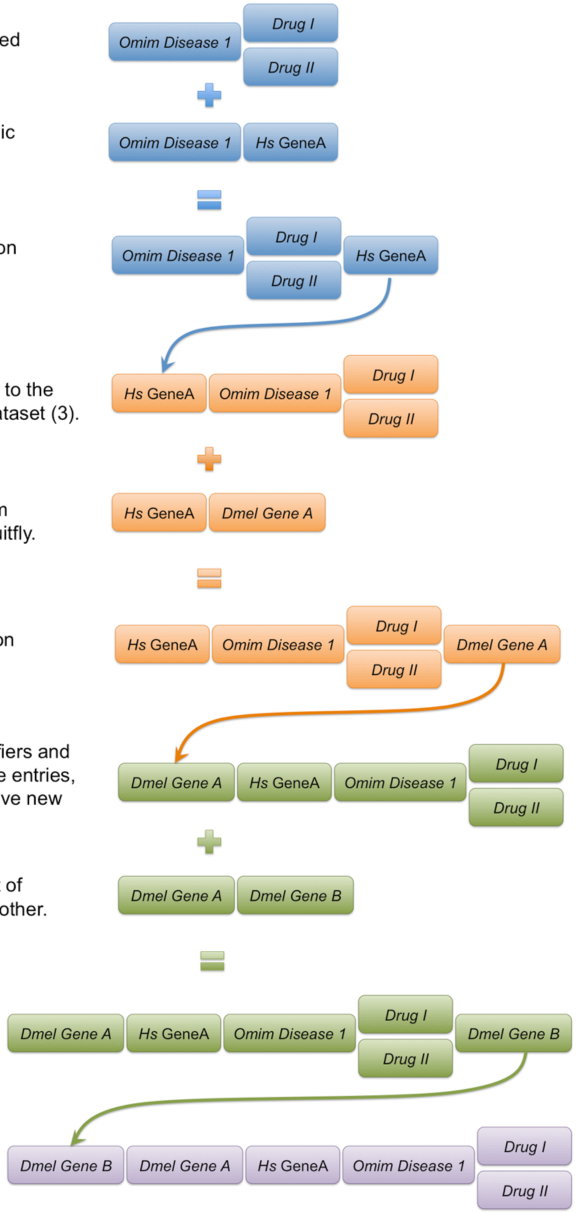
Results of Booleome DB intersection from Datasets 3 and 4.

Dataset 5. Again, we switch identifiers and save as a new dataset (5). Multiple entries, i.e., multiple BLAST hits, also receive new separate entries for each identifier.

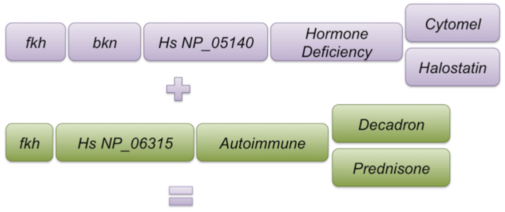
Dataset 6. We now introduce a list of genes known to interact with each other.

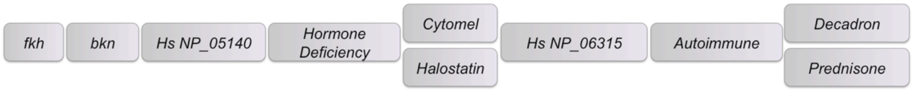
Intersect 5 & 6.

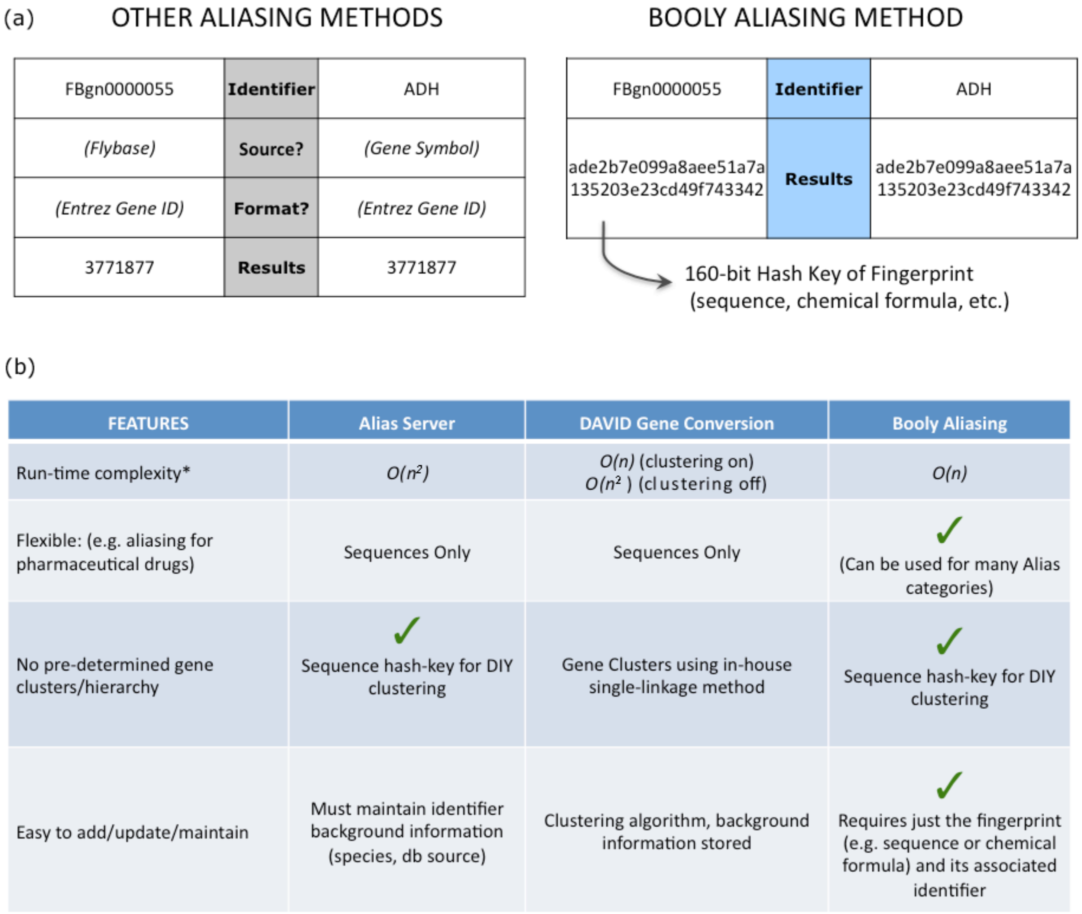
Dataset 7. Switch and Save.



Intersect Datasets 5 & 7. We have replaced the toy names with an example placing two diseases and their associated drugs within an interaction network found in *D. melanogaster* (*fkh* and *bkn*).






(a)
OTHER ALIASING METHODS
BOOLY ALIASING METHOD
160-bit Hash Key of Fingerprint
(sequence, chemical formula, etc.)
(b)

OTHER ALIASING METHODS

FBgn0000055	**Identifier**	ADH
(Flybase)	**Source?**	*(Gene Symbol)*
(Entrez Gene ID)	**Format?**	*(Entrez Gene ID)*
3771877	**Results**	3771877

BOOLY ALIASING METHOD

FBgn0000055	**Identifier**	ADH
ade2b7e099a8aee51a7a 135203e23cd49f743342	**Results**	ade2b7e099a8aee51a7a 135203e23cd49f743342

FEATURES	Alias Server	DAVID Gene Conversion	Booly Aliasing
Run-time complexity*	$O(n^2)$	$O(n)$ (clustering on) $O(n^2)$ (clustering off)	$O(n)$
Flexible: (e.g. aliasing for pharmaceutical drugs)	Sequences Only	Sequences Only	✓ (Can be used for many Alias categories)
No pre-determined gene clusters/hierarchy	✓ Sequence hash-key for DIY clustering	Gene Clusters using in-house single-linkage method	✓ Sequence hash-key for DIY clustering
Easy to add/update/maintain	Must maintain identifier background information (species, db source)	Clustering algorithm, background information stored	✓ Requires just the fingerprint (e.g. sequence or chemical formula) and its associated identifier

Figure 14. Booly aliasing resource. (a) Difference between other aliasing approaches and the Booly-hashing method. The single question we wish to answer efficiently is, whether two identifiers (e.g., FBgn0000055 and *ADH*) are one and the same? Booly-hashing utilizes a 160-bit SHA-1 hash key to generate unique fingerprints of sequences and their identifiers represented as a 40 character hexadecimal number. Identifiers with the same hash-keys are considered as aliases of each other. Other approaches require knowledge of the source of the original identifier or knowledge of a conversion format. **(b)** Comparison of two commonly used aliasing tools in bioinformatics (AliasServer and DAVID Gene Conversion Tool) against the Booly-hashing resource.

*Run-time analysis of a worse case scenario to iterate through a given input list of identifiers and the total number of database source or output formats.

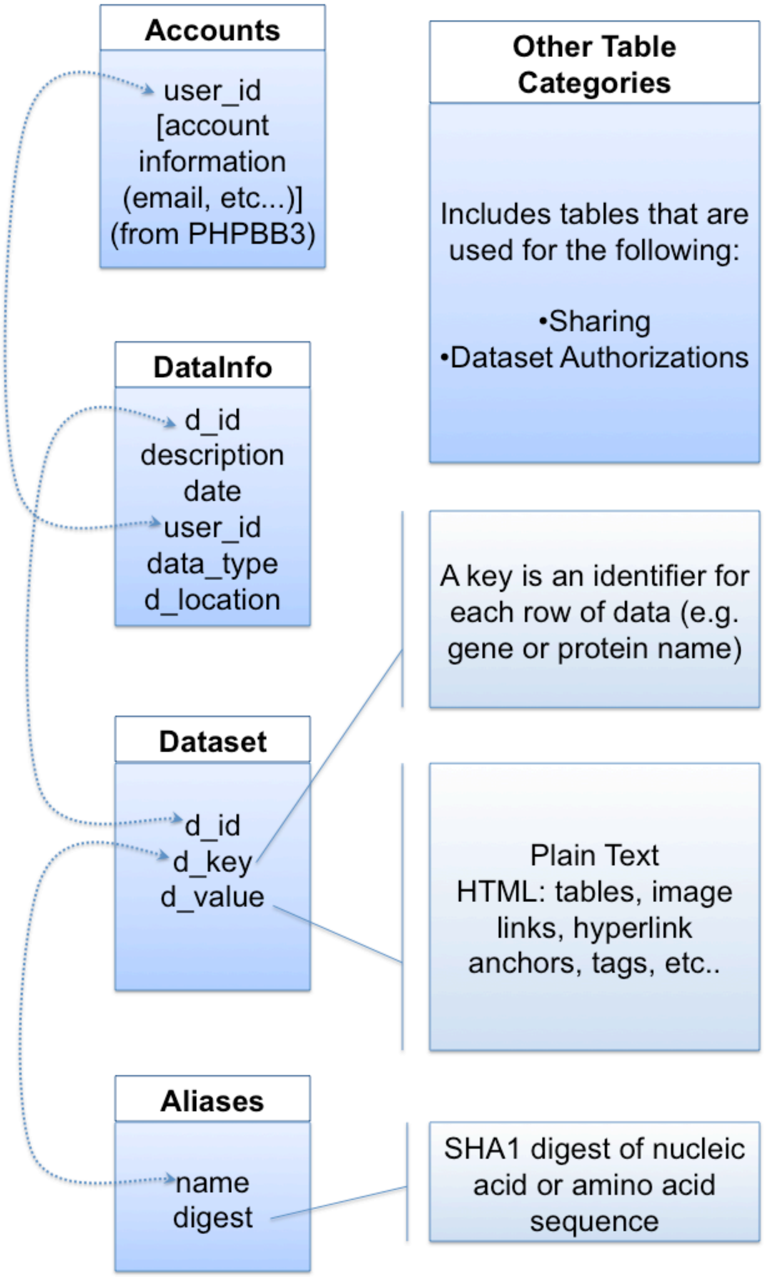

Accounts
user_id
[account information (email, etc...)]
(from PHPBB3)
Other Table Categories
Includes tables that are used for the following:
•Sharing
•Dataset Authorizations
DataInfo
d_id
description
date
user_id
data_type
d_location
A key is an identifier for each row of data (e.g. gene or protein name)
Dataset
d_id
d_key
d_value
Plain Text
HTML: tables, image links, hyperlink anchors, tags, etc..
Aliases
name
digest
SHA1 digest of nucleic acid or amino acid sequence

Figure 15. Database Schema for Booly. Booly is an account based web tool which utilizes a relational MySQL database and custom scripts to perform Boolean merges between different datasets. The "Dataset" table consists of similarly structured tables horizontally partitioned across multiple servers (d_location). Each row of data contains a key, value pair. The key is the identifier for the value (text or html). Booly has integrated aliasing to group the same genes or proteins together.

Table 1. Genes lost selectively in *D. ananassae*. We identified over 73 genes that were lost during evolution of the *Drosophila ananassae lineage* that were retained in the sister melanogaster subgroup comprised of *D. melanogaster, D. simulans, D. sechelia, D. yakuba*, and *D. errecta* and in the outgroup *D. pseudoobscura*.

Function	***D. mel* Gene**	**D. mel CG**	***D. mel* Name**	***D. mel* Gene Ontology**	***D. mel* Prot. Lngth**
Defense					
	FBgn0044811	CG31691	**TotF**	humoral defense mechanism (sensu Protostomia)	125
	FBgn0031701	CG14027	**TotM**	humoral defense mechanism (sensu Protostomia)	131
	FBgn0044810	CG31193	**TotX**	humoral defense mechanism (sensu Protostomia)	142
	FBgn0053117	CG33117	**Victoria**	extracellular, "humoral defense mechanism (sensu Protostomia)"	134
	FBgn0004240	CG12763	**Dpt**	extracellular, "antibacterial humoral response (sensu Protostomia)", "defense response to bacteria", "innate immune response", "NOT defense response to Gram-negative bacteria"	106
Barrier Formation					
	FBgn0000357	CG6517	**Cp18**	structural constituent of chorion (sensu Insecta), "insect chorion formation", "chorion"	172
	FBgn0041252	CG15573	**Femcoat**	structural constituent of chorion (sensu Insecta), "cytoplasm", "insect chorion formation"	201
Chemosensation					
	FBgn0041232	CG32395	**Gr65a**	taste receptor activity	408
	FBgn0038203	CG14360	**Or88a**	olfactory receptor activity, "odorant binding", "perception of smell", "NOT integral to membrane"	401
	FBgn0034509	CG13421	**Obp57c**	odorant binding, "transport", "cellular_component unknown"	149
	FBgn0030103	CG12665	**Obp8a**	odorant binding, "transport"	163
Reproduction					
	FBgn0010401	CG3250	**Os-C**	pheromone binding	131
	FBgn0000246	CG17604	**c(3)G**	synaptonemal complex, "structural constituent of cytoskeleton", "protein targeting", "cytoskeleton organization and biogenesis", "mitosis", "meiotic recombination", "microtubule binding"	744
Metabolism					
	FBgn0025809	CG8962	**Paf-AHalpha**	1-alkyl-2-acetylglycerophosphocholine esterase activity, "phospholipid metabolism"	225
	FBgn0044051	CG14173	**Ilp1**	insulin receptor binding, "hormone activity", "extracellular", "physiological process"	154
Transcription					
	FBgn0033010	CG3136	**Atf6**	DNA binding, "nucleus", "regulation of transcription, DNA-dependent", "protein homodimerization activity"	741
	FBgn0033459	CG12744	**CG12744**	nucleic acid binding, "nucleus", "zinc ion binding"	160
	FBgn0037183	CG14451	**CG14451**	nucleic acid binding, "nucleus", "zinc ion binding"	264
Translation					
	FBgn0039739	CG15527	**RpS28a**	nucleic acid binding, "structural constituent of ribosome", "cytosolic small ribosomal subunit (sensu Eukarya)", "protein biosynthesis"	64
	FBgn0011824	CG4038	**CG4038**	small nucleolar ribonucleoprotein complex, "rRNA processing", "35S primary transcript processing", "ribosome biogenesis", "rRNA binding"	237
Proteolysis					
	FBgn0033875	CG6357	**CG6357**	cysteine-type endopeptidase activity, NOT cathepsin L activity	439
	FBgn0051704	CG31704	**CG31704**	serine-type endopeptidase inhibitor activity, "proteolysis and peptidolysis"	68

Table 2. Redundancy of common reference databases. The DAVID Gene Conversion tool creates clusters of gene groups analogous to Entrez Gene. Gene clusters from DAVID are labeled with numerical identifiers that are reused and recycled after each update, thus not optimal for use as a reference identifier of a gene. A common approach is to convert aliases into a single source database (REFSEQ, Entrez, etc.) identifier for comparison. The above table shows the lack of complete redundancy across multiple reference databases. Only 29% (37081/127749) of gene clusters identified by DAVID (v6.7) are found to be present in all five reference databases from the three organisms.

Unique DAVID Id's Mapped	Fruitfly	Mouse	Human	Total
DAVID-Refseq mRNA	13978	16262	16060	46300
DAVID-Entrez Gene ID	21227	58530	40959	120716
DAVID-Ensembl ID	13945	18307	18370	50622
DAVID-Genbank GI	14253	48480	37281	100014
DAVID-Gene Symbol	20571	44859	32100	97530
Total DAVID ID Overlap	11525	13231	12325	37081
Total Unique DAVID IDs	23569	59881	44299	127749

References

1. Ashby WR: **An Introduction to Cybernetics**. London: Chapman and Hall; 1957.

2. Bertalanffy L: **General Systems Theory**. Harmondsworth: Penguin; 1973.

3. Mesarovic MD: **System theory and biology -- view of a theoretician**. New York: Springer-Verlag; 1968.

4. Wolkenhauer O: **Systems biology: the reincarnation of systems theory applied in biology?** *Brief Bioinform* 2001, **2**(3):258-270.

5. Noble D: **The Music of Life: Biology beyond the genome**: Oxford University Press; 2006.

6. Goble C, Stevens R: **State of the nation in data integration for bioinformatics**. *J Biomed Inform* 2008, **41**(5):687-693.

7. Stein LD: **Integrating biological databases**. *Nat Rev Genet* 2003, **4**(5):337-345.

8. Gopalacharyulu PV, Lindfors E, Bounsaythip C, Kivioja T, Yetukuri L, Hollmen J, Oresic M: **Data integration and visualization system for enabling conceptual biology**. *Bioinformatics* 2005, **21 Suppl 1**:i177-185.

9. Lee TJ, Pouliot Y, Wagner V, Gupta P, Stringer-Calvert DW, Tenenbaum JD, Karp PD: **BioWarehouse: a bioinformatics database warehouse toolkit**. *BMC Bioinformatics* 2006, **7**:170.

10. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart--biological queries made easy**. *BMC Genomics* 2009, **10**:22.

11. Davidson SB, Overton CG, Tannen V, Wong L: **BioKleisli: a digital library for biomedical researchers**. *Int J on Digital Libraries* 1997, **1**:36-53.

12. Chung SY, Wong L: **Kleisli: a new tool for data integration in biology**. *Trends Biotechnol* 1999, **17**(9):351-355.

13. Donelson L, Tarczy-Hornoch P, Mork P, Dolan C, Mitchell JA, Barrier M, Mei H: **The BioMediator system as a data integration tool to answer diverse biologic queries**. *Stud Health Technol Inform* 2004, **107**(Pt 2):768-772.

14. Birkland A, Yona G: **BIOZON: a hub of heterogeneous biological data**. *Nucleic Acids Res* 2006, **34**(Database issue):D235-242.

15. Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks**. *Methods Enzymol* 1996, **266**:114-128.

16. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A: **TAMBIS: transparent access to multiple bioinformatics information sources**. *Bioinformatics* 2000, **16**(2):184-185.

17. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems**. *J Biomed Inform* 2008, **41**(5):706-716.

18. Consortium TB: **Interoperability with Moby 1.0–It's better than sharing your toothbrush!** *Briefings in Bioinformatics* 2008, **9**(3):220-231.

19. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ: **The HUGO Gene Nomenclature Database, 2006 updates**. *Nucleic Acids Res* 2006, **34**(Database issue):D319-321.

20. Berners-Lee T, Hendler J, Lassila O: **The Semantic Web**. *Scientific American* 2001(May 1, 2001).

21. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.

22. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN *et al*: **Evolution of genes and genomes on the Drosophila phylogeny**. *Nature* 2007, **450**(7167):203-218.

23. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN *et al*: **Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures**. *Nature* 2007, **450**(7167):219-232.

24. Huang da W, Sherman BT, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID gene ID conversion tool**. *Bioinformation* 2008, **2**(10):428-430.

25. Berriz GF, Roth FP: **The Synergizer service for translating gene, protein and other biological identifiers**. *Bioinformatics* 2008, **24**(19):2272-2273.

26. Iragne F, Barre A, Goffard N, De Daruvar A: **AliasServer: a web server to handle multiple aliases used to refer to proteins**. *Bioinformatics* 2004, **20**(14):2331-2332.

27. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, Bernard A, Boe AF, Boguski MS, Brockway KS, Byrnes EJ *et al*: **Genome-wide atlas of gene expression in the adult mouse brain**. *Nature* 2007, **445**(7124):168-176.

28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

29. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L *et al*: **Ensembl 2009**. *Nucleic Acids Res* 2009, **37**(Database issue):D690-697.

30. Jarvis ED, Gunturkun O, Bruce L, Csillag A, Karten H, Kuenzel W, Medina L, Paxinos G, Perkel DJ, Shimizu T *et al*: **Avian brains and a new understanding of vertebrate brain evolution**. *Nat Rev Neurosci* 2005, **6**(2):151-159.

31. Lein ES, Zhao X, Gage FH: **Defining a molecular atlas of the hippocampus using DNA microarrays and high-throughput in situ hybridization**. *J Neurosci* 2004, **24**(15):3879-3889.

32. Macphail EM: **The role of the avian hippocampus in spatial memory**. *Psicológica* 2002, **23**:93-108.

33. Reiter LT, Potocki L, Chien S, Gribskov M, Bier E: **A systematic analysis of human disease-associated gene sequences in Drosophila melanogaster**. *Genome Res* 2001, **11**(6):1114-1125.

34. Hu G, Agarwal P: **Human disease-drug network based on genomic expression profiles**. *PLoS One* 2009, **4**(8):e6536.

35. Babnigg G, Giometti CS: **A database of unique protein sequence identifiers for proteome studies**. *Proteomics* 2006, **6**(16):4514-4522.

36. Smith M, Kunin V, Goldovsky L, Enright AJ, Ouzounis CA: **MagicMatch--cross-referencing sequence identifiers across databases**. *Bioinformatics* 2005, **21**(16):3429-3430.

37. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**(Database issue):D13-21.

38. Rivest RL: **The MD4 Message Digest Algorithm, Advances in Crytology, CRYPTO '90**. *Proceedings, Springer-Verlag* 1991:303-311.

39. Rivest RL: **RFC 1321: The MD5 message-digest alorithm.** *Internet Engineering Task Force* 1992.

40. Do LH, Esteves FF, Karten HJ, Bier E: **Booly: a new data integration platform**. *BMC Bioinformatics* 2010, **11**(1):513.

41. Radack S: **The Cryptographic Hash Algorithm Family: Revision Of The Secure Hash Standard And Ongoing Competition For New Hash Algorithms**. In: *ITL NIST Bulletin.* 2009.

42. **The free open source bulletin board** [http://www.phpbb.com/]

43. Albert I, Thakar J, Li S, Zhang R, Albert R: **Boolean network simulations for life scientists**. *Source Code Biol Med* 2008, **3**:16.

44. **Meeting Report: A Workshop to Plan a Deep Catalog of Human Genetic Variation**. In*: 2007*; 2007.

45. Hoffmann R: **A wiki for the life sciences where authorship matters**. *Nat Genet* 2008, **40**(9):1047-1051.

46. Zhang Z, Cheung KH, Townsend JP: **Bringing Web 2.0 to bioinformatics**. *Brief Bioinform* 2009, **10**(1):1-10.

47. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V *et al*: **Advancing translational research with the Semantic Web**. *BMC Bioinformatics* 2007, **8 Suppl 3**:S2.