Analysis of High Throughput Genomic Datasets Across Species

Guy E. Zinman

May 2012
CMU-CB-12-102

Lane Center for Computational Biology
School of Computer Science
Carnegie Mellon University
Pittsburgh PA 15213

Thesis Committee:
Ziv Bar-Joseph, Advisor, Chair
Roni Rosenfeld
Panayiotis (Takis) Benos
Zoltan Oltvai

*Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy*

This research was sponsored by the National Science Foundation under contract no. DBI-0640543. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official polices, either expressed or implied, or any sponsoring institution, the U.S. government, or any other entity.

Keywords: dynamic network analysis, gene expression analysis, cross species analysis, protein-protein interactions, genetic interactions, soft clustering, active sub-networks, conservation and divergence

This dissertation is dedicated to my parents and friends
whose support and patience have carried me through

Abstract

Genes are highly conserved between closely related species, and biological systems often utilize the same genes across different organisms. This fact has allowed the study of various biological systems using model organisms and the development of many drugs for human diseases by first researching simpler model organisms. New high-throughput technologies have enabled researchers to use interactions and expression data to get a more precise view regarding the roles and functions of biological processes across species. However, combining and comparing these types of data across species is challenging due to several problems including homology assignments, coverage issues, and quality of the data in each of the species.

This thesis studies various aspects of cross species analysis in light of these obstacles and introduces new algorithms and computational tools that specifically address them. First, we performed a global analysis of conservation of interaction and expression data by developing a framework that integrated various data types from four model organisms. This analysis showed that while interactions are often not conserved at the protein level, they are conserved at a higher network organization level. These findings paved the way to developing three tools aimed at analyzing expression data from multiple species concurrently: 1) *ExpressionBlast*, a search engine for gene expression data, which provides the ability to query experimental results obtained in one species against all public expression studies conducted in the same or in a different species. 2) *SoftClust*, a new constrained clustering method which integrates expression data with sequence orthology information in a modified *k*-means model to jointly cluster expression data from several species. 3) *ModuleBlast*, an active sub-network search tool that makes use of both static interaction data and condition-specific expression data from multiple species to understand conservation and divergence of biological systems dynamics.

The tools introduced in this thesis were incorporated into a web-based expression analysis package with enhanced support for cross species analysis. We hope that these tools will have an impact in elucidating the underlying molecular mechanisms in a variety of organisms.

Acknowledgments

I am indebted to many who have helped me along the path to this dissertation. I would like to thank my advisor, Dr. Ziv Bar-Joseph, whose enthusiasm for my work has carried it forward from a few ideas to what it is today. Ziv has been a fantastic support and has allowed me to exploit my own strengths in research and to explore new fields and ideas. I have appreciated Ziv's knowledge of the field and his own excitement for research. I would also like to thank Shan Zhong, who was my 3-year long office mate and research partner that always had great research insights and gladly offered assistance whenever I needed one. I would like to thank all of those that I have collaborated with along the way: Zoltan Oltvai, Dwight Kuo, Trey Ideker, Takis Benos, Penelope Morel, Gerard Nau, Yariv Kanfi, Shoshana Naiman, and Haim Cohen. Each of these collaborators comes from a different field with different areas of expertise and introduced me to exciting new perspectives on biological and experimental research. My education would not be complete without the perspectives I have learned from our Systems Biology group past and present members: Jason Ernst, Yanjun Qi, Yong Lu, Henry Lin, Peter Huggins, Marcel Schultz, Saket Navlaka, Anthony Gitter, Shan Zhong, Hai-Son Le, and Aaron Wise. I would also like to thank the CMU-Pitt joint computational biology PhD program that provided the framework that enabled me to achieve my research goals. Lastly, I would like to thank the software developers that worked with me along the way, in particular Nishant Kumar who helped to develop the web interface for the tools presented here.

Table of Contents

List of Figures

List of Tables

1 Introduction

A landmark essay by Theodosius Dobzhansky, "Nothing in Biology Makes Sense Except in the Light of Evolution” [1] noted "The diversity and the unity of life are equally striking and meaningful aspects of the living world. Between 1.5 and 2 million species of animals and plants have been described and studied; the number yet to be described is probably as great. The diversity of sizes, structures, and ways of life is staggering but fascinating. The unity of life is no less remarkable than its diversity. Most forms of life are similar in many respects. The universal biologic similarities are particularly striking in the biochemical dimension".

This essay, although published almost 40 years ago, is still relevant for all the biological discoveries found since. Cell cycle complexes and mechanisms of operation can serve as an excellent example for the similarity and diversity of life. Most complexes that are part of the cell cycle including DNA replication, deoxynucleotide biosynthesis, and Anaphase-promoting utilize similar genes in human and in yeast. Nonetheless, a study on the in-time expression of the cell cycle proteins showed that while they are usually retained between two types of yeast and human they exhibit significant changes in their dynamics [2]. Understanding the similarities and mapping the differences is a major goal of the biological community in order to study intricate human diseases on simpler model organisms for which the costs and regulation are much simpler. Nurnberger *et al.* noted for example similarities and obvious differences in the innate immunity of plants and animals [3]. Indeed most drug discovery and development is conducted on simpler model organisms before it is applied to human.

Major use of model organisms in studying human diseases and drug discovery include:

1. Discovering new genes
2. Identifying genes causing human disease
3. Defining cellular pathways
4. Finding pathway perturbations leading to diseases

See Table 1-1 for a list of the major use of model organisms and discussion in [4].

These types of studies are possible due to the significant similarity in genes between different species. I list here only a few studies out of many showing specific applications for cross species studies to understand human diseases using model organisms.

- Foury Listed 105 yeast homologues of human disease-associated genes including Immunodeficiency, Anemia, Autism, Diabetes and Insulin resistance, Cataracts and Glaucoma, and Brain tumors [5].
- Yuel *et al.* found hundreds of cancer related genes that are conserved in yeast and showed many common synthetic lethal interactions among yeast CIN (Chromosome Instability) genes that have human homologs with known mutations leading to malignant tumors [6].
- Hariharan and Haber predicted that 60 to 80 percent of disease-causing genes in humans have orthologs in the fly genome [7].
- Bilen and Bonini showed that *Drosophila* is a good model organism for neurodegenerative diseases including polyglutamin diseases, Parkinson disease, noncoding trinucleotide repeat diseases, Alzheimer and related diseases [8].
- Potter showed that mosaic flies are a good model for patients with cancer predisposition syndromes such as those heterozygous for mutated tumor suppressor genes [9].
- Fontana *et al.* showed that nutrient signaling pathways regulating longevity are conserved in yeast, worms, flies, and mammals. These include conserved anti-aging transcription factors that are activated by dietary restrictions that lead to similar inhibition of nutrient sensing pathways including TOR signaling pathways, RAS-AC-PKA, and Insulin/Igf-like signaling [10].

Table 1-1: Major use of model organisms to study human diseases

Model Organism	Common Name	Research Applications
Saccharomyces cerevisiae	Yeast	Cell processes e.g. mitosis and diseases (e.g. cancer)
Drosophila melanogaster	Fruit fly	A wide variety of studies ranging from early gene mapping to mutant screens to identify genes related to specific biological functions
Caenorhabditis elegans	Nematode	Development of simple nervous systems and the aging process
Danio rerio	Zebra fish	Mapping and identifying genes involved in organ development
Mus musculus	House mouse	Used to study genetic principles and human disease
Rattus norvegicus	Brown rat	Used to study genetic principles and human disease

Many of the above studies were made possible through the development of new sequencing technologies and sequence analysis comparison algorithms including BLAST [11] that allow easy identifications of orthologs and their control regions. Orthologs are divergent copies of a single gene that were separated by a speciation event (See Figure 1-1). The key assumption is that sequence similarity implies functional similarity, making BLAST a rudimentary tool for every biology researcher for characterizing new genes and elucidating functions of known genes.

In recent years more high throughput datasets including expression data [12], Protein-Protein Interactions (PPI), Genetic Interactions (GI), and others became available for increasing number of species. These types of data provide the opportunity to capture functional attributes of genes and their overall role in the cell under various conditions in ways never before possible. Moreover, in many cases, genes with very high sequence similarity may have different roles under similar conditions; therefore, relying on sequence similarity alone to delineate gene functions might be misleading. The study mentioned earlier [2] exploring in-time expression of conserved cell cycle proteins, and proving that they exhibit significant changes in their dynamics is only one example for this discrepancy. Other examples, for such differences include a study by

Price *et al.* that showed that orthologous transcription factors in bacteria have different functions and regulate different genes [13]. Chapter 2 embodies a comprehensive discussion on this topic.

One interesting approach tried to directly compare disease models between model organisms and human [14]. In this study they tried to look for non-obvious equivalences between mutant phenotypes in different species based on overlapping sets of orthologous genes from human, mouse, yeast, worm, and plant. Using these orthologous phenotypes that were able to predict unique genes associated with diseases and suggest a yeast model for angiogenesis defects, a worm model for breast cancer, and a mouse model for autism, among others.

Gaining a better understanding of the conserved and divergent biological process across species is thus an important problem that will increase our knowledge on the basic operation of cells and aid the discovery of treatments for a large number of diseases. However the process of going from these large scale experimental data sources to new biological insights requires new methods that address a set of new computational challenges. These include early studies showing very low conservation rates for expression and interactions datasets (see Chapter 2), different measurement errors in each of the species, quality and coverage differences, orthologs assignment. All these make direct adaptation of many existing tools, designed to investigate only a single species, unfeasible. As a result, there is a lack of computational tools that are easy to use and can aid researchers performing cross species analysis.

This thesis presents new computational methods designed to better use high throughput data sources from different species to analyze dynamic conditions and biological processes in the cell.

1.1 Identifying Orthologs

Comparing and contrasting the high throughput datasets including expression and interactions data would not have worked without reliably identifying orthologs by comparing DNA and protein sequence information of multiple species. Orthologs are defined as genes that diverged in a speciation event and originated by vertical descent from a single gene of the last common ancestor. It is possible to have many-to-many orthologous relationships between genes from different species (see Figure 1-1). Orthologs tend to have a similar function but there are cases that do not hold to this assumption [13]. The traditional method for identifying orthologs was performing a reciprocal-best-BLAST-hit (RBH) query of one species proteins against the other species. Nonetheless, this approach may predict paralogs as orthologs based on the time of gene duplication. Genes that have duplicated after the speciation event (in-paralogs) are by definition orthologous, but genes that have duplicated before the speciation event (out-paralogs) are not orthologs.

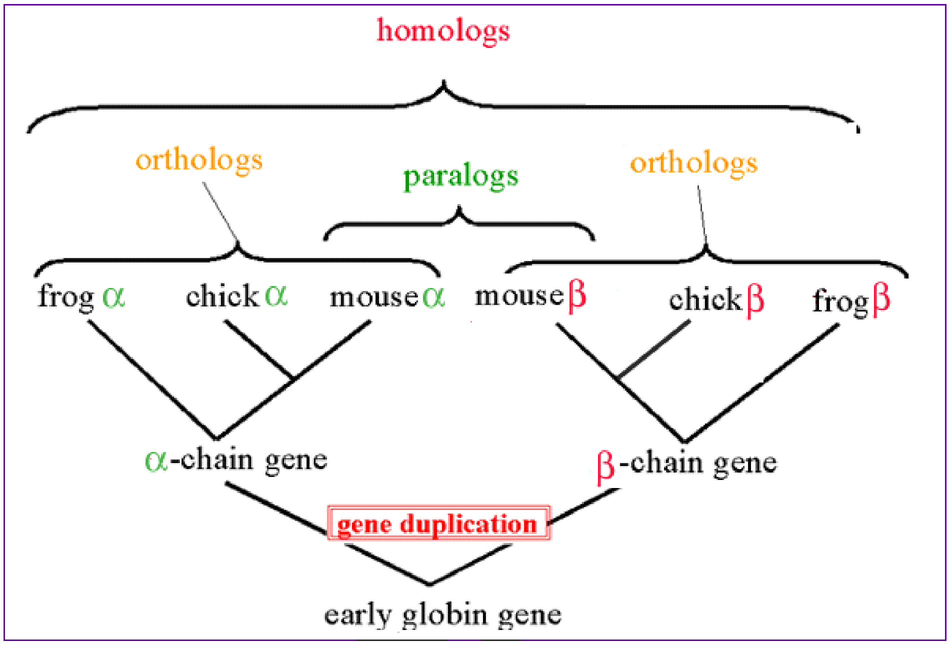


Figure 1-1: A visual representation of orthologs

Orthologs are defined as genes that diverged in a speciation event and originated by vertical descent from a single gene of the last common ancestor. It is possible to have many-to-many orthologous relationships between genes from different species. Image source: http://www.bio.davidson.edu

One approach to improve the identification of orthologs was suggested by Fulton *et al.* They developed a computational method named Orthologue that analyzes the phylogenetic distance ratios involving two comparison species and an outgroup species and identifies cases were relative gene divergence is atypical [15].

The state-of-the-art method today for identifying pairwise orthologs relationships and distinguish between in-paralogs and out-paralogs is Inparanoid [16], [17]. Inparanoid uses the pairwise similarity scores calculated by BLAST to construct core orthology groups that are expanded with other in-paralogs as long they are closer than any other sequence in any proteome.

Another approach for building orthology groups that pioneered orthology group definitions is the COG project [18], now extended and enhanced by EggNog [19]. The approaches employed by these projects include an all-against-all similarity search with a subsequent clustering step. Rather than calculating matches for each species-pairwise (like in Inparanoid), an orthology group is formed for multiple species. Another important project that needs to be mentioned in this respect is OrthoDB [20] that built a hierarchical catalog of orthologs.

Lastly, there are attempts aimed at improving orthology relations by combining sequence data with other high throughput datasets including protein-protein-interactions data [21].

1.2 Gene Expression Data

Gene expression refers to the quantity of mRNA produced from each gene. Expression measurements are one of the most popular methods to gain dynamic information on the state of cells under specific conditions. While biologically we are usually interested in the levels of proteins, mRNAs are chemically much easier to measure, compared to proteins, and the assumption is that there is a high correlation between the mRNA levels and proteins levels.

To date, the most popular type of experiments for measuring gene expression levels are microarray experiments. These experiments, measure the average mRNA expression level based on pre-define probes designed for specific genes. Microarrays can generally be divided into two categories; two-channel (or two color) microarrays and single-channel (one color) microarrays.

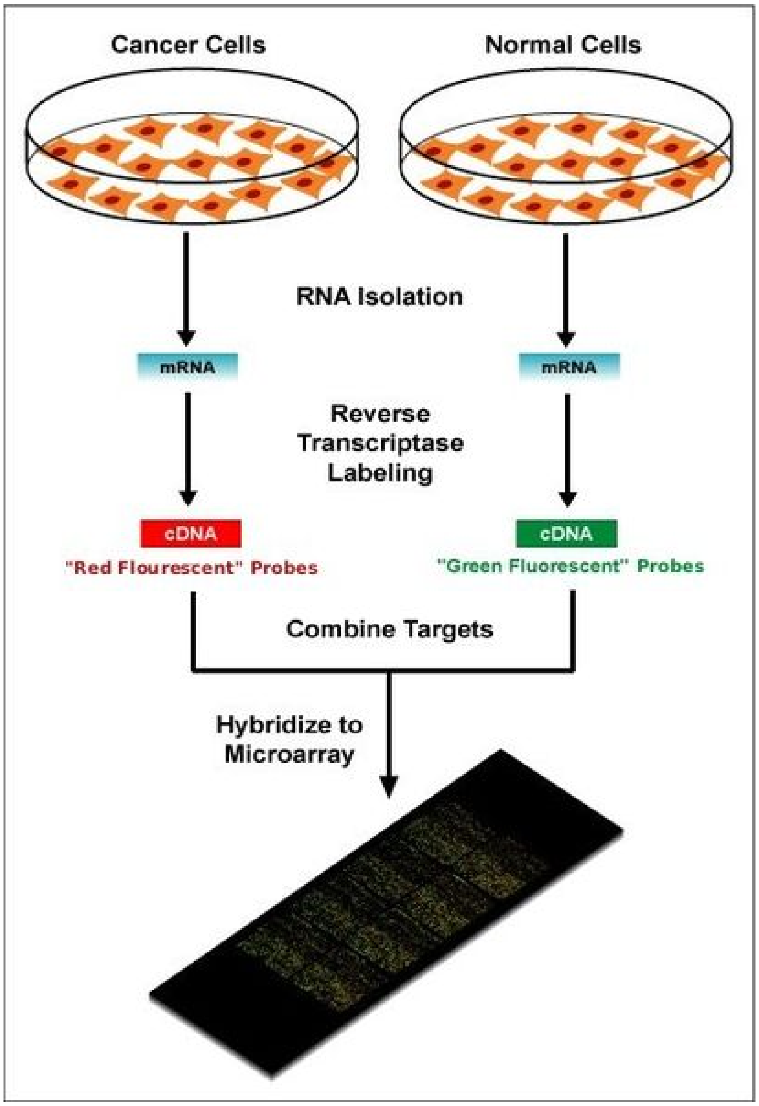


Figure 1-2: Outline of a cDNA Microarray experiment

The experiment begins with two different populations of cells (e.g. cell from normal tissue vs. cell from cancer tissue). mRNA is isolated from both populations and is then reverse transcribed into cDNA. cDNAs from each population are labeled with different fluoresce colors (simplistically 'red' and 'green'). In a process called hybridization the labeled cDNA is then placed on the microarray and the cDNA bind to probes on the microarray with a complementary sequence. The intensity and color of the spots indicate which population of cells has larger mRNA levels. Image source: http://en.wikipedia.org/wiki/DNA_microarray

The two-channel microarrays, that are oligonucleotide-based, are typically based on microscopic spots containing short sequences (~25 bp) that are directly synthesized onto the microarray. The outline for the experimental procedure (see Figure 1-2) is based on mRNA extraction from the two samples that we would like to compare (e.g., cells from healthy tissue vs. cells from cancer tissue). The mRNA is then converted to cDNA (complementary DNA) in a process called reverse transcription and is labeled with two different fluorescent dyes (commonly Cy3 and Cy5, which correspond to the green and red part of the light spectrum respectively). The two Cy-labeled cDNA samples are mixed and in a process called hybridization. Then the cDNA is placed on the microarray slide and binds to specific probes with complementary sequences. The hybridization process of two channel arrays is practically a competition between the two samples and both can bind to the same spot. The microarray is then scanned in a microarray scanner to determine the relative intensities of the two samples for each spot to determine up and down-regulated genes. In the single-channel microarrays, the intensity data is provided for only one sample, but they do not truly indicate the abundance levels of a gene, but rather the relative abundance compared to other samples processed in the same experiment. Microarrays tend to produce noisy measurements caused by different RNA extraction protocols, batch specific biases during amplification, labeling and hybridization phases thus making direct comparison of gene measurements to be uninformative. There are several companies that produce microarrays including Affymetrix, Agilent, Eppendorf, and TeleChem.

Expression studies are continuously gaining popularity as a result for the relatively cheap costs, ease of the procedures, and the valuable information gained. Each publication that uses expression study is required to upload the data to a public repository. The adoption trends of the technology are clearly seen in the exponential growth of the number of studies in Gene Expression Omnibus (GEO) [22] that has occurred over the past few years (see Figure 1-3). Expression data is also produced for more and more species (see Figure 1-4). In addition to GEO, there are other public databases collecting gene expression data, including databases with specialized focus on collecting expression data across species e.g., 4DXpress [23]. See Chapter 3 for more details on expression data repositories.

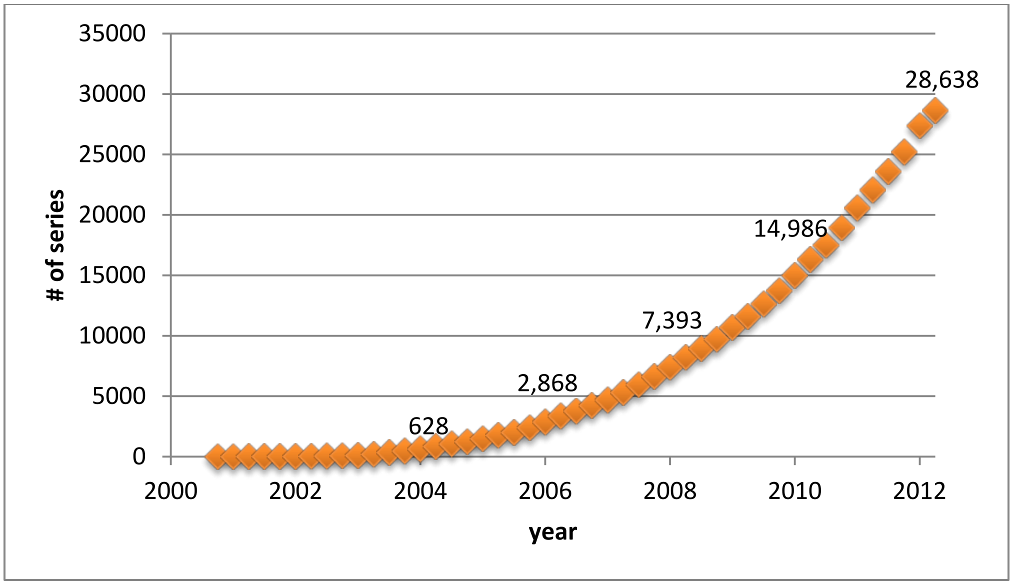
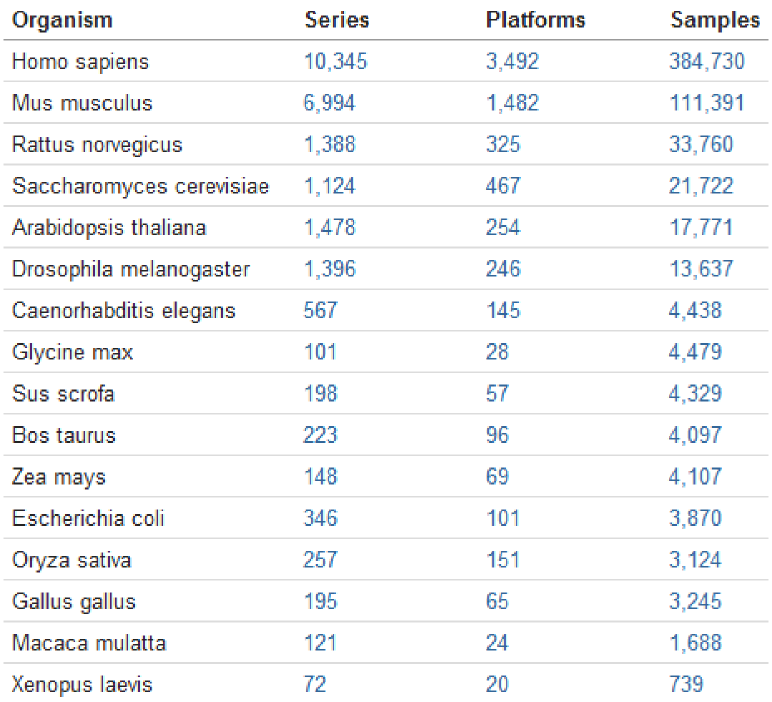


Figure 1-3: Expression series (GSE) accumulation

Exponential growth in the number of expression series (GSEs – a collection of individual samples (GSMs) that usually belong to one publication) in the Gene Expression Omnibus (GEO) [22] has occurred over the past few years. Specific counts are listed for round years. Data is based on GEO listings as of Feb 2012. Note: not all the entries counted correspond to microarray expression studies.

In recent years, a new expression measurement technology is gaining popularity is RNA-sequencing, which enables a much more accurate measurement of the number of transcripts and allows researchers to distinguish between different transcripts of the same gene including alternative splicing transcripts [24]. This technology is not based on a single probe for each gene, but rather identifies all the transcripts in the measured cell hence it is not restricted only to known genes. Nonetheless, this technology is more expensive, the generated output is less straight forward to use, the analysis requires expert knowledge, and there are still no gold standards for processing this type of data. Once these problems are solved, this technology will become the common standard for gene expression measurements.



Organism	Series	Platforms	Samples
Homo sapiens	10,345	3,492	384,730
Mus musculus	6,994	1,482	111,391
Rattus norvegicus	1,388	325	33,760
Saccharomyces cerevisiae	1,124	467	21,722
Arabidopsis thaliana	1,478	254	17,771
Drosophila melanogaster	1,396	246	13,637
Caenorhabditis elegans	567	145	4,438
Glycine max	101	28	4,479
Sus scrofa	198	57	4,329
Bos taurus	223	96	4,097
Zea mays	148	69	4,107
Escherichia coli	346	101	3,870
Oryza sativa	257	151	3,124
Gallus gallus	195	65	3,245
Macaca mulatta	121	24	1,688
Xenopus laevis	72	20	739

Figure 1-4: GEO public holdings by organism

Listings for the number of GEO series (GSEs), platforms (GPLs), and individual samples (GSMs) in the Gene Expression Omnibus (GEO) [22] by organism. Only key organisms are listed. Data is based on GEO listings as of Feb 2012.

1.3 Other High-throughput Experimental Methods and Sources

The ability to sequence entire genomes and proteomes resulted in a collection of genomic entities that serve as a 'parts list' of the intricate cell machineries. The gene expression technologies gave us the ability to measure the quantities of each entity in each condition, or rather get the 'count' of each item in the 'parts list'. However, genes and proteins do not operate in isolation but rather interact one with another to carry out their biological functions. Without understanding these interactions we are left with an ‘assembly instruction manual’ that contains only the 'parts list'. Several possible interactions can be formed between the genomic entities including protein-DNA interactions, direct protein-protein interaction (PPI), and indirect protein-protein interactions (genetic interactions). In recent years new high throughput methods were developed to measure these different types of interactions. We present here only a general overview of the most popular methods for each type.

1.3.1 Protein-DNA interactions

Proteins which we usually name Transcription Factors, bind to DNA molecules and activate or repress gene expression by binding to DNA motifs. The most common method to measure protein-gene interactions on a large scale is a technique called ChIP-chip that combines chromatin immunoprecipitation ('ChIP') and microarray technology ('chip'). This technology enables a whole-genome analysis to determine the location of binding sites for almost any protein of interest. For example, nowadays Affymetrix offers arrays with about 90 million probes spanning the complete non-repetitive part of the human genome with about 35bp spacing. These types of experiments are also conducted in an increasing number of species allowing for comparative analysis. For example, a study by Borneman *et al.* found divergence of transcription factor binding sites across related yeast species [25] and a study by Wilson *et al.* found species-specific transcription between mouse and human [26].

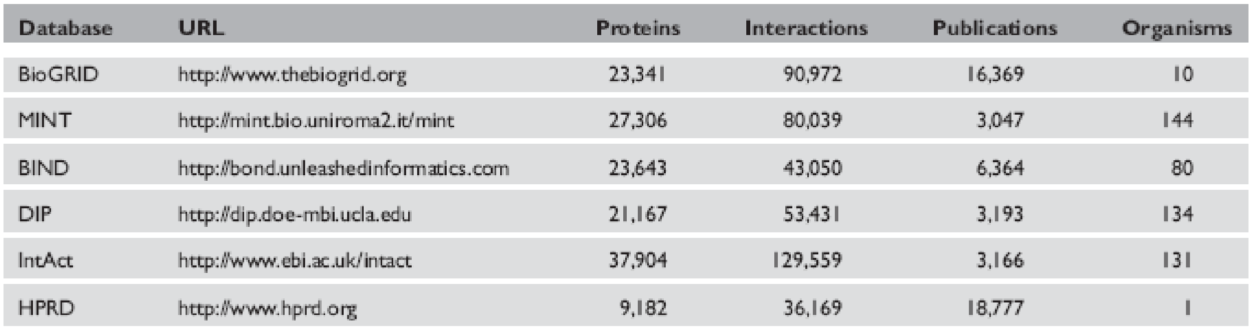
As with RNA-sequencing, Chip-sequencing is recently gaining popularity as an alternative to traditional ChIP-chip using massively parallel DNA sequencing. Chip-sequencing offers high resolution, less noise, and greater coverage as the precision is not limited to predetermined probes [27].

1.3.2 Direct protein-protein interactions

Proteins can bind physically to perform fundamental roles in numerous biological processes. In many cases a group of proteins establish long and stable interactions to form protein complexes. There are many methods to investigate physical protein-protein binding on a large scale, each with its own strengths and weaknesses and the reader is kindly referred to [28] for a review on methods for detection and analysis of protein-protein interactions. The most popular high-throughput method is Tandem Affinity Purification (TAP). It is based on a TAP-tag which is fused to a specific protein of interest and is then washed through two affinity columns and examined for binding partners. This method can be used to determine protein partners quantitatively in-vivo without prior knowledge on of complex composition. Nonetheless, it cannot readily detect transient protein-protein interactions. Two genome-wide TAP studies in *S.*

cerevisiae [29] and [30] processed over 4,500 different tagged yeast proteins and identified hundreds of proteins complexes.

There are several protein interaction databases encompassing thousands of proteins in hundreds of organisms that differ in scope and in content. See Figure 1-5 and [31] for a comparative review on protein-protein interaction databases. The most notable repositoreis are BioGRID [32], MINT [33], IntAct [34], DIP [35] and BIND [36].



Database	URL	Proteins	Interactions	Publications	Organisms
BioGRID	http://www.thebiogrid.org	23,341	90,972	16,369	10
MINT	http://mint.bio.uniroma2.it/mint	27,306	80,039	3,047	144
BIND	http://bond.unleashedinformatics.com	23,643	43,050	6,364	80
DIP	http://dip.doe-mbi.ucla.edu	21,167	53,431	3,193	134
IntAct	http://www.ebi.ac.uk/intact	37,904	129,559	3,166	131
HPRD	http://www.hprd.org	9,182	36,169	18,777	1

Figure 1-5: Protein-protein interaction databases

Listing for several major protein-protein interaction databases as of 2009. Source: [31]

1.3.3 Genetic interactions

Another way to examine indirect effect of one protein on another is through genetic (or epistatic) interactions. These interactions describe the extent that a mutation in one gene modulates the phenotype associated with altering a second gene. Genetic interactions (GI) can be mapped on a genome-wide scale using the Epistatic Miniarray Profile (E-MAP) platform. In E-MAP, double deletion strains are systematically constructed by cross a query strain, which carries a mutation of one gene with a library of test strains each one carrying a mutation of a second gene [37]. The double mutant strains are grown for a pre-determined period of time and the colony size of the double mutant strains is measured. The size of the double mutant colonies can then be compared with the size of the query gene mutant colonies to determine the epistatic relation between the two mutants. The genetic interactions can be classified to two categories; negative GIs which correspond to cases were the double mutant has a less severe phenotype than either single mutant, and positive GIs were the double mutant has a more severe phenotype than one predicted by the additive effects of the single mutants (see Figure 1-6 for depiction of these definitions).

To date, two large scale studies were conducted in *S. cerevisiae* [37] and *S. pombe* [38] and several small scale studies were conducted in other species [32].

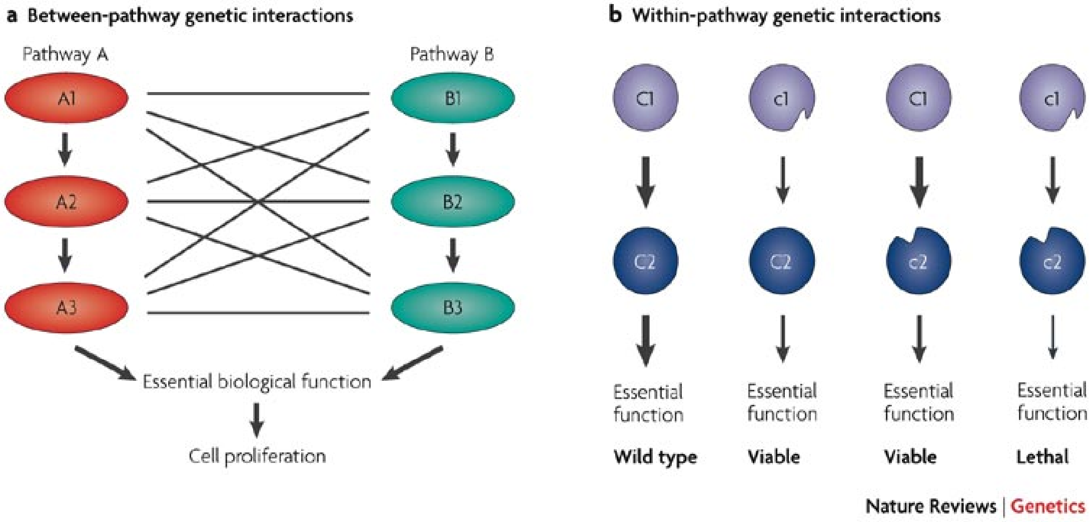


Figure 1-6: Outline for Genetic Interactions

Genetic interactions (GI) can be classified into two categories; **(a)** Negative GIs which correspond to cases were the double mutant has a less severe phenotype than either single mutant, and **(b)** Positive GIs were the double mutant has a more severe phenotype than one predicted by the additive effects of the single mutants. Source: [242].

1.4 Network Biology

The rich collection of various high throughput datasets mentioned in the previous section can be gathered to form association networks spanning all the genes and proteins and understand the flow of processes over entire cascades of interacting genes and proteins. In recent years the focus is shifting from understanding the network structure itself to understanding the networks underlying specific processes and diseases.

There are several papers that reviewed methods for the integration of interactions data into networks examining network properties, and their possible relations to known diseases. A review paper by Srinivasan *et al.* discussed how interactions data from different types can be integrated to allow for experimental prioritization [39]. Specifically, they explored how reference networks should be assembled in a similar manner to sequence references, which can be used as the lowest common denominator of interaction information to compare between species. In another review paper by Kann [40], he showed how effecting interactions between proteins and genes or production of undesirable interactions are the cause of many diseases. He also exemplified that genes may have roles in several diseases by sharing interaction sub-networks. An additional review paper by Ideker and Sharan [41] discussed the network properties of disease genes, showing that they tend to have higher network connectivity. They listed methods for identifying disease related sub-networks and demonstrated a specific protein interaction network for the Huntington disease.

Network biology is now further applied to drug discovery. Network Pharmacology [42] for example, describes the integration of network biology and polypharmacology. In one example, Fliri *et al.* examined how drugs affect cellular network structures and how resulting signals are translated into drug effects. In [43] they examined cause-effect relationships by determining protein network structures associated with the generation of specific *in-vivo* drug-effect patterns. Towards this goal they built drug-induced protein network toplogy maps by finding protein network positions that can be reached during drug treatment for 1320 medicines. Then they identifyed the average shortest paths for transferring drug-induced signals through the protein network. This method also allows for comparing different drugs by examining their protein reachability profiles [44]. See Figure 1-7 for illustration. Lastly, a recent study by Navlakha and Kingsford examined the performance of seven methods for determining gene-disease association using physical interaction networks [45] .

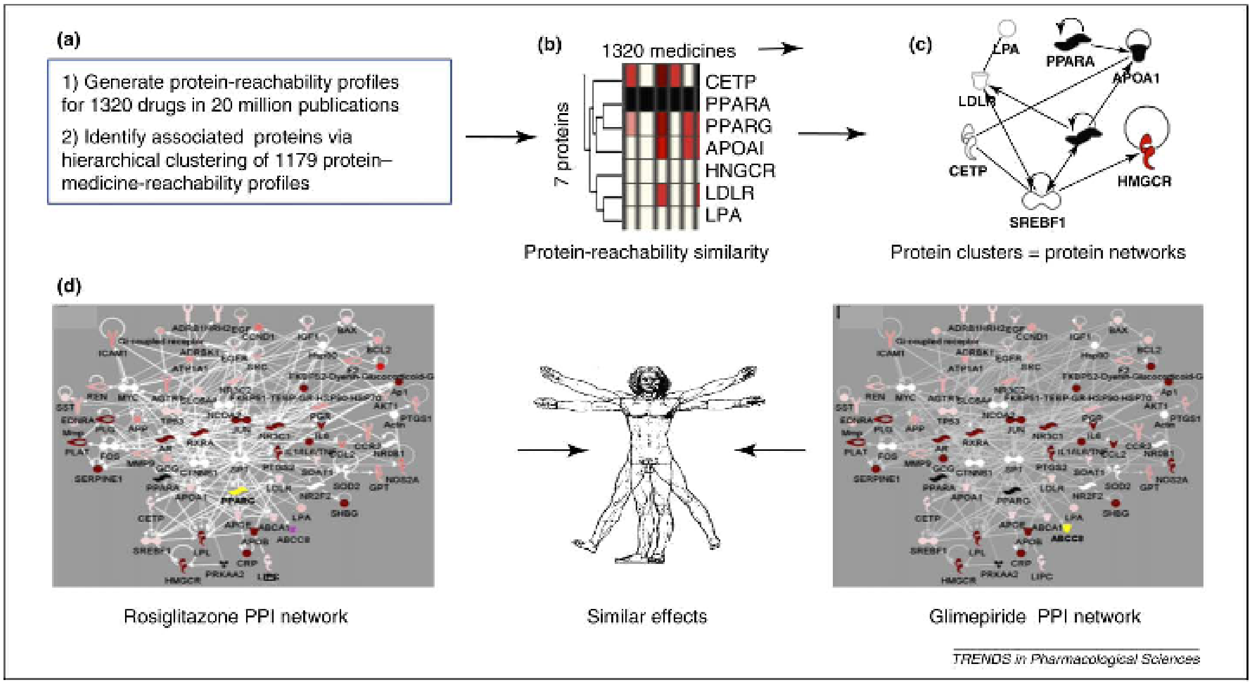

(a)
1) Generate protein-reachability profiles for 1320 drugs in 20 million publications
2) Identify associated proteins via hierarchical clustering of 1179 protein–medicine-reachability profiles
(b)
1320 medicines
7 proteins
CETP
PPARA
PPARG
APOAI
HNGCR
LDLR
LPA
Protein-reachability similarity
(c)
LPA
PPARA
APOA1
LDLR
CETP
HMGCR
SREBF1
Protein clusters = protein networks
(d)
Rosiglitazone PPI network
Similar effects
Glimepiride PPI network
TRENDS in Pharmacological Sciences

Figure 1-7: Protein interactions networks in drug discovery

(a) How often 1179 proteins are investigated with 1320 medicines using abstracts in the PubMed database was determined. **(b)** Protein associations were obtained by hierarchical clustering of 1179 protein–medicine profiles using 1320 medicines. Color coding (black = most; white = least) denotes how often proteins and medicines were co-investigated. The dendrogram of 7 proteins is an indicator that 1320 medicines view them as being highly associated (functionally coupled). **(c)** A protein network topology map was generated by adding the minimal number of neighbor proteins (in this case, one: SREBF1) identified using curated protein interaction databases to directly connect all the dendrogram proteins shown in b. **(d)** Examining drug–effect profiles of 1320 medicines revealed that rosiglitazone and glimepiride have similar effect profiles. The rosiglitazone and glimepiride protein network topology maps show similar protein network reachability. Their respective drug targets (PPARG and ABCC8) are shown in yellow. Source: [44].

1.5 Previous Studies analyzing High Throughput Data Across Species

There are numerous previous studies that analyzed various high throughput datasets across species. Yong *et al.* reviewed different strategies for cross species analysis of microarray expression data [12]. In this review they divided the cross species analysis strategies to three categories; 1) using the same array for all species, 2) expression meta-analysis, and 3) concurrent analysis of expression data. The approach employed in the first category can be used only to compare closely related species using multi-species arrays hence it is not commonly used today. The second category refers to studies that perform studies on single species and then combine species meta-analysis to leverage annotation in one species to improve the expression analysis in the less studied species. Several examples are listed in section 1.5.1. The third category refers to studies that employ a concurrent type of analysis. Several examples for this approach are listed in 1.5.2. Studies comparing other types of high throughput data including interactions and regulatory networks are listed in sections 1.5.3.

1.5.1 Expression meta-analysis

Examples for the meta-analysis approach include studies by Lu *et al* [46] compared mouse, rat, and human expression datasets to model human bladder cancer using carcinogen-induced rodent models and found a number of molecular pathways that were commonly activated leading to a conclusion that rodent models of bladder cancer represent well the human clinical disease.

Another study by Bergmann *et al.* [47] compared gene expression data from six evolutionary distant organisms and by linking genes whose expression profiles are similar, and found that the connectivity distribution follows a power-law and showed that the expression program is highly modular. Their approach demonstrated the potential of combining orthology information and expression information for improving gene annotation and expanding our understanding on how gene expression and diversity evolved.

Lastly, a study by Tirosh *et al.* [48] compared four closely related yeast species under a variety of environmental stresses and matched the expression response to regulatory elements on the promoter including the TATA box. They found enhanced expression divergence of TATA-containing genes in the yeast species and all eukaryotes including nematodes, fruit flies, plants, and mammals.

1.5.2 Concurrent analysis of expression data from multiple species

A popular approach to analyzing gene expression data is clustering the data in order to find groups or modules of co-expressed genes that biologically are likely to be part of the same biological process or are regulated by a similar set of transcription factors. In the cross species domain the idea is to find clusters of co-expressed genes that are show correlated expression across species. These can correspond to a regulatory program that is functionally important and thus resistant to evolutionary changes. Differences in the regulatory program across species may be interesting as well and can indicate on evolutionary divergence or novelty. See Chapter 4 for more details on cross species clustering approaches including a novel approach we present named *SoftClust* for clustering expression data across species that was published in [49]. I note below two additional recent studies published following [49] that present other approaches to cross species co-expression clustering.

Cai *et al.* [50] presented a probabilistic model and maximum likelihood approach called SCSC that enables a 'soft' assignment of orthologs similar to the *SoftClust* method presented in Chapter 4. Unlike the *SoftClust* method that performs a joint clustering, the SCSC clustering is first performed separately for each species, and then cluster labels are paired using orthologous relationships. The SCSC method was applied to human and mouse embryonic stem (ES) cell data and revealed several transcription factors and signaling proteins that were specifically expressed in either human or mouse ES cells, suggesting that the pluipotent cell identity can be established and maintained through more than one regulatory network.

In another recent study, Zarrineh *et al.* [51] presented a co-clustering approach named COMODO that constructs a module tree for each species separately by gradually decreasing the distance measure used by the clustering or distance approach. The modules generated by the most stringent thresholds are matched across species and are expanded simultaneously in both species by traversing up the two module trees and identifying the best matching pairs from all possible matching pairs. COMODO was applied *to Escherichia coli* and *Bacillus subtilis* and showed that despite the potential for extensive network rewiring in prokaryotes, some elementary pathways are extremely preserved, possibly due to the operon structure.

Another approach to analyzing cross species expression data concurrently in order to identify common and unique response patterns was introduced in a pair of studies [52], [53] which used probabilistic graphical models, in particular Markov random fields (MRFs) to combine data from different species. Nodes in the graph represent genes and edges in the graph represent sequence similarity. Belief propagation is applied on the graph in order to find a core set of genes with similar expression. This method was found to be useful for identifying cycling genes using cell cycle expression data from budding yeast and human [52], and for finding conserved and divergent key players in the innate immune system in mouse and human [53]. This approach was later expanded to cross species temporal graphical models by capturing casual relations between genes from time series microarray data using a hidden Markov random field regression [54].

Lastly, an approach for performing cross-species queries of expression data of large gene expression databases was presented by Le *et al.* [55]. This study defined a distance metric between the rankings of ortholgous genes in two species utilizing a training set that determines that the similarity between two experiments.

1.5.3 Interactions and regulatory networks studies

Chapter 2 summarizes a large body of previous work in assessing conservation rates of individual interactions. To complete the picture I list here a number of studies that compared full interactions networks and regulatory networks from multiple species.

Liang *et al.* [56] was among the first studies to examine conservation of protein interaction networks between orthologs. They compared networks from seven different species including bacteria and human using a fast graph isomorphism algorithm and looked for connected maximal common sub-graphs. They found conserved network substructures that correspond to basic cellular functions and substructures with different topology that infer potential species divergence.

Sharan *et al.* [57] integrated interaction data from three species by generating a three-way network alignment graph where each node in the graph consists of a group of orthologs, one from each species and links between the groups represent conserved interactions. Their method is a based on a search over the alignment using a probabilistic model to find linear paths that may represent conserved signal transduction pathways.

In another study Berg *et al.* [58] performed network alignment based on a scoring function measuring mutual similarity between networks using a Bayesian parameter inference. Their method was applied to compare human and mouse networks showing that most of the gene pairs have only average sequence similarity, hence the network alignment contains functional information beyond the corresponding sequence alignment.

Lastly, an approach gaining popularity in the last decade is to combine static interaction data and condition-specific expression data to identify biological pathways of interest. Chapter 5 of this thesis describes a novel method for identifying active sub-networks across species. Active sub-networks are connected group of genes that show high activation (up or down regulation) over the entire group. Only one study [59] examined this domain across species and is described in more details in chapter 5.

1.6 Overview of Thesis

In Chapter 2 of this thesis we examine conservation rates of interaction and expression data which were previously reported to be conserved at rates that are much lower than expected, thus raising the question whether these types of data can be used for cross species analysis. Towards this goal we built a robust analysis framework that compares various interactions and expression datasets across four model organisms (*S. cerevisiae, S. pombe*, *C. elegans,* and *D. melanogaster*). We looked for functional modules in interaction networks and compared conservation rates within and between modules. Our analysis shows that interactions and expression data are conserved at a higher network organization level rather than at the individual protein level and are suitable for cross species analysis. This chapter is based on our published paper [60].

In Chapter 3 we use these insights and focus on comparing expression data, the most comprehensive data type in terms of coverage, across different studies and species. Towards this goal we built a system we name *ExpressionBlast* that downloads, automatically parses and annotates all the expression experiments available at GEO, the largest repository of gene expression data [22]. The uniform treatment of the data across different studies and species allows it to be searchable and comparable. This is facilitated by a web interface that makes it possible for users to compare their own expression experiments against thousands of previous studies and gain new insights and leads for follow-up analyses. A follow-up to our recently published paper on lifespan extension in male mice [61] using *ExpressionBlast* led to far reaching hypotheses regarding the mechanisms leading to the gender specific lifespan extension and are now being follow up experimentally.

In Chapter 4 and Chapter 5 we show how a unified analysis can be performed on expression data from multiple species (e.g., through distinct studies that were found to match using *ExpressionBlast*). In Chapter 4 we introduce a novel clustering method we name *SoftClust* for clustering gene expression data. Clustering is one of the most popular analyses performed on gene expression data. Yet, available clustering methods lead to situations in which orthologous genes with similar expression patterns could be misplaced into different clusters due to factors such as measurement error and ambiguity of cluster boundaries. We developed a constrained soft clustering framework that can incorporate prior information and improve gene assignments to clusters when performing cross species analysis. We show a specific application of our approach to study the mechanisms by which three evolutionary distant yeasts respond to the anti-fungal medicine fluconazole overtime, revealing significant divergence among regulatory programs associated with fluconazole sensitivity. This chapter is based on our published paper [49]

In Chapter 5 we show a novel method we name *ModuleBlast* for combining static interaction data and condition-specific expression data to find active modules, connected groups of gene that show high differential expression, across multiple species and analyze their conservation patterns. This work is based on Chapter 2 insights that facilitated our approach in combining interaction networks from multiple species into a single weighted network whose nodes represent entire orthogroups. Our method looks for functionally active modules based on expression data from all species in the orthogroup. We applied our approach to examine the response of alveolar macrophages from mice and cynomolgus macaques to the highly infectious pathogen *F. tularensis*. We identified core modules that are active in at least one of the species and show similar or divergent responses between the species. Specifically, we identified several apoptotic and anti apoptotic modules that may explain how NFκB-mediated apoptosis is correlated with *F. tularensis* infection.

In Chapter 6 we present a comprehensive gene expression analysis package based on the work presented in Chapters 3, 4, and 5. One of the aims of this thesis is to provide tools that will facilitate the analysis of high throughput datasets, primarily expression data, and will include enhanced support for cross species analysis. The tools that were introduced in this thesis *ExpressionBlast, ModuleBlast,* and *SoftClust* were combined into a collection of user-oriented web tools. These tools provide the option for users to upload and store their data, save results, and share them. Furthermore, the tools are integrated and results of one tool can be easily directed to another tool for further analysis.

In Chapter 7 we conclude the contributions of this thesis and propose several directions for future work. We suggest a holistic view for the web tools development as a community-based open API approach to provide flexibility in incorporating new tools on three different layers: the data layer, the analysis layer, and the visualization layer.

To summarize the major contributions of this thesis, they are:

- A theoretical framework for cross species analysis of high throughput data showing that interactions are conserved on the module level (Chapter 2).
- *ExpressionBlast* - a novel search engine for gene expression data that operates within and across specie, which proved to suggest concrete hypotheses for follow-up analyses to explain gender specific life extension in mice (Chapter 3).
- *SoftClust* - a new clustering method specifically designed for cross species analysis that was able to identify divergent mechanisms between three yeasts treated with the anti-fungal medicine fluconazole (Chapter 4).
- *ModuleBlast* - a new method for identifying active modules (differentially expressed sub-networks) across species and classifying their conservation patterns. This method was applied to understand mice and macaques infected with the highly infectious bacteria *F. tulatrensis* and found possible explanations for mechanisms employed by *F. tularensis* during the infection (Chapter 5).
- A web-based expression analysis package, with enhanced support for cross species analysis, which integrates *ExpressionBlast*, *SoftClust*, and *ModuleBlast* (Chapter 6).

2 Studying the Conservation of Cross Species in High Throughput Data

Basic cellular systems and the proteins that participate in these systems are often conserved across many species. However, while sequence similarity often implies functional similarity, interaction data is not well conserved, even for proteins with high sequence similarity. Several recent studies comparing high throughput data including expression, protein-protein, protein-DNA, and genetic interactions between close species show conservation at a much lower rate than expected. Focusing on four model organisms for which high throughput datasets are available, we show that while interactions are often not conserved at the protein level, they are conserved at a higher network organization level that we term modules. Interactions within the same module are much more likely to be conserved than interactions between modules. This intermediate conservation level provides modularity allowing different species to use the same building blocks for different processes, mirroring basic sequence conservation patterns.

2.1 Introduction

Basic cellular systems including the cell cycle, innate immunity and mRNA translation operate in a similar manner across a large number of species. The proteins that participate in these systems are highly conserved across evolution [2]. This has led to many successful efforts to infer gene function using genes with similar sequence across species[62]. The availability of large sequence datasets and powerful computational methods, including BLAST [11], has further facilitated this process. Other applications of comparative genomics allowed the characterization of proteins, control regions [63], and micro RNAs [64] as well as other insights into function and development [65], [66].

While genes with very similar sequence often perform the same function, dynamic properties of conserved proteins, including expression and interactions, seem to differ substantially between species. In studies profiling similar tissues in mouse and human, researchers found a large divergence in expression profiles [67] Correlations from 0.17 to 0.37 were found for orthologous genes, depending on the tissue. The correlation of cell cycle expression between two yeasts was determined to be around 0.1 [68]. Similarly, in protein-DNA binding studies, researchers found that only 11% of binding interactions for a highly conserved transcription factor were conserved between human and mouse [69] . Studies of three yeast species with high sequence similarity identified only 20% conservation in binding targets [25] and similar results were obtained for binding in Bacteria [13] . Protein interactions were also found to overlap in very low rates [70–73] (Gandhi *et al.* reported rates that are as low as less than 1% of the interactions between four species [72]. Only an estimated 18% to 29% of negative genetic interactions between *S. cerevisiae* and *S. pombe* were found to be conserved [38], [74].

Early studies have mainly focused on pairwise comparisons based on a single genomic data type. While the results in these early papers indicated low overlap between species, no attempt was made to generalize observations to address reasons for the lower conservation of interaction data when compared to sequence data conservation. Recent high throughput experiments with better coverage [29], [30] made it possible to reassess the conservation of interaction data. A number of possible reasons have been proposed to explain the lack of conservation for specific types of interaction data. For example, Fox *et al.* [70] observed that interactions connecting hub proteins are more conserved when compared to interactions involving proteins with a lower degree of connectivity. As they show using PPI data from multiple species, there is a positive correlation between the average degree of a protein and the conservation of its interacting partners. Byrne *et al.* [75] studied the genetic interaction networks of *S.cerevisiae* and *C.elegans* and reported that while only little overlap is seen for individual interactions, the properties of their genetic interaction networks are conserved. They proposed that changes in individual genetic interactions might be a form of evolution. Another direction suggested by Roguev *et al.* [38] demonstrated that conservation of interactions within protein complexes is higher than that of other interactions. They compared genetic interactions between chromatin-related genes in two yeasts and determined that protein complexes and the evolution of a new biological mechanism (RNAi) can help explain the minimal overlap observed, hypothesizing that protein-protein interactions pose a constraint on functional divergence in evolution. Similarly, Jensen *et al.* [2] compared cell cycle expression of a number of species and discovered that while in-time expression was not conserved at the individual gene level, it was much more conserved at the protein complex level. Van Dam and Snel [76] showed that conservation rates for PPI within complexes in human and yeast are much higher than overall interaction conservation. On the other hand, Wang and Zhang [77] studied conservation of yeast, fly, and nematode PPI networks and determined that interactions in protein complexes are not conserved at levels that are higher than other interactions. Beltrao *et al.* [78] claimed that protein complexes are correlated with higher conservation only for stable interactions, while transient interactions, including phosphoregulation, are less conserved.

The experimental methods used to obtain expression data are large scale and produce measurements for the entire genome leading to a significantly better coverage of the interactome compared to the other data types. In addition, as there is no equivalent to protein complexes in expression data, early analysis of the conservation of dynamic properties in expression data focused on the identification of conserved expression modules across species [79],[47][58], [80] . While some important expression modules were conserved, many others were not.

The above discussion illustrates several (sometimes conflicting) trends observed for the conservation of interactions across species. One of the reasons for the disagreement between the results of these observations is the fact that each was only tested on a small dataset, often for only one type of interaction data (protein interaction, co-expression etc.), in one specific condition and between a single pair of species. To determine which of these trends hold more generally we performed a comprehensive analysis using four model organisms, and several genomic data types measured under a variety of conditions. As we show below, while all the proposed directions so far indeed explain part of the differences between species, none is enough to provide a comprehensive explanation. We have thus attempted to generalize these suggestions. Our findings suggest that while sequence and function are conserved at the individual protein level, interactions are conserved at a higher organizational level for which we use the term ‘functional modules’. These results indicate that while gene-gene interactions are not well conserved, the overall network, through the intermediate level of modules, is conserved to a much higher degree.

2.2 Methods

2.2.1 Network construction

Co-expression Network

All two-channel microarrays for *S. cerevisiae*, *C. elegans*, and *D. melanogaster* stored in Stanford Microarray Database (SMD)[81] were retrieved. Default filtering options for both arrays and genes were applied to all the three organisms, resulting in 788 arrays for *S. cerevisiae*, 332 arrays for *C. elegans*, and 164 arrays for *D. melanogaster*. All two-channel microarrays for *S. pombe,* were extracted from NCBI GEO [82] since SMD does not contain microarray data for *S. pombe*. For genes with several probes, the median log ratio of the probes was used as the value for the gene. The Spearman correlation coefficient (SCC) was computed for all pairs of genes in each of the four species. I.e., for each pair of genes (*x*,*y*) in the four species, the Spearman correlation coefficient (SCC) ρ was calculated as follows:

$$\rho = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}} \quad (2\text{-}1)$$

in which *n* is the number of arrays in the corresponding species, x_i and y_i are the ranks of the log ratio of gene *x* and y on the *i*th array respectively, and $\overline{x}$ and $\overline{y}$ are the average ranks of gene *x* and *y* respectively.

To generate the co-expression network, we used log likelihood score scheme, originally described in [83]. Log Likelihood Scores (LLS) were computed using a probabilistic approach that assigns a score to each interaction between two genes based on their likelihood of participating in the same biological process.

In this scheme, $LLS = \ln\left(\frac{p(L|E)/\,p(\neg L|E)}{p(L)/p(\neg L)}\right)$ where $p(L|E)$ and $p(\neg L|E)$ are the frequencies of linkages (L) observed in a given experiment (E) between annotated genes operating in the same pathway and in different pathways. $p(L)$ and $p(\neg L)$ are the total frequencies of linkages between annotated genes operating in the same pathway and different pathways. Genes sharing Biological Process annotation of GO level 5 or below were defined as in the same pathway. The final LLS score is defined by splitting the interaction to bins of 2000 interactions each, calculating the LLS for each group and thereafter building a regression line between the raw correlation values and the LLS that was obtained for each bin. The log likelihood scores were calculated for each set of expression experiment. All gene-pairs interactions with a positive score were connected in the co-expression network for that species. The maximal score was taken for an interaction if it was observed in more than one experiment. The maximal score is an effective way to avoid cases where the expression experiments are not independent.

Protein-protein and genetic interaction networks

We collected protein-protein interaction (PPI) data for the four species from the following databases: IntAct [34], MINT[33], DIP [35] and BioGRID [32]. We took the union of all the PPIs documented in these databases and represented them as networks for each of the four species. We collected the genetic interaction (GI) data for the four species from BioGRID [32]. For each species, one network for positive GIs and another for negative GIs were generated. LLS scores were calculated for all protein-protein and genetic interactions in all species, in a similar manner to the method described for the co-expression network. As protein-protein and genetic interactions are binary, no regression is needed and one LLS score is calculated for all edges per species.

Sequence network

Network representing paralogous genes within a species was generated by performing all-against-all BLASTP for each of the four organisms against itself. All genes that were matched with E-value less than 1E-25 divided by the number of genes in the species were considered as neighboring nodes. LLS scores were calculated for all genetic interactions in all species, in a similar manner to the method described for the PPI network. Regression lines were built for each of the sequence networks in a similar manner to the co-expression networks. Nonetheless the score variations between the bins were too little, effectively leading to one LLS score for all edges per species.

GO network

We generated a GO network for each species based on the Biological Process (BP) annotations in the Gene Ontology database[84].We used the semantic similarity measures developed by Wang *et al*. [85] for this purpose. Simply put, for each term A in GO:BP, let T_A represent all of *A*'s ancestor terms up the GO:BP tree plus *A* itself. An *S*-value[85] is calculated for each term *t* in T_A as follows:

$$S_A(t) = \begin{cases} 1 & t = A \\ \max\limits_{t':childrenof\,(t)} (w \times S_A(t')) & t \neq A \end{cases} \qquad (2\text{-}2)$$

in which *t*' represents all the children of *t* in T_A, and *w* is a weight-like semantic contribution factor and is set to default as described in [85] to 0.8 for is-a relations and 0.6 for part-of relations between *t* and *t*'. $S_A(t)$ represents the contribution of *t* to the semantics of *A*. Then the semantic similarities of each pair of GO terms *A* and *B* are calculated as

$$S_{GO}(A,B) = \frac{\sum\limits_{t \in T_A \cap T_B} \left(S_A(t) + S_B(t) \right)}{\sum\limits_{t \in T_A} S_A(t) + \sum\limits_{t \in T_B} S_B(t)} \qquad (2\text{-}3)$$

and the semantic similarities of each pair of genes *G1* and *G2* that have annotations $GO_1=\{go_{11},go_{12},...,go_{1m}\}$ and $GO_2=\{go_{21},go_{22},...,go_{2n}\}$ are calculated as

$$Sim(G1,G2) = \frac{\sum\limits_{1 \le i \le m} \max\limits_{1 \le j \le n} \left(S_{GO}(go_{1i}, go_{2j}) \right) + \sum\limits_{1 \le j \le n} \max\limits_{1 \le i \le m} \left(S_{GO}(go_{1i}, go_{2j}) \right)}{m+n} \qquad (2\text{-}4)$$

in which *m* and *n* are the number of GO:BP annotations for *G1* and *G2*, respectively. In calculating the gene-gene similarity scores, genes that are only annotated with large GO:BP (categories that contain more than 5% of the number of all genes in the corresponding species) were removed, since they are poorly characterized. A cutoff of 0.8 was applied for all the four species to convert the data into network representations.

The integrated network

The co-expression, PPI, positive GI, and sequence networks for each species were combined to generate an integrated weighted network by summing the log likelihood scores of an interaction from all networks. As the experiments from different genomic data are assumed to be independent, the summation should not create any bias for any edge in the integrated network.

2.2.2 Orthologs mapping

We identified one-to-one mappings of orthologs for each pair of the four species. For *S. cerevisiae* and *S. pombe*, we first started from a manually curated list of orthologs for these two species [86] , and extracted all the one-to-one mappings from this list. For cases of many-to-many mappings, all-against-all BLASTP was performed and pairs of genes that are each other's best reciprocal hit were assigned as additional one-to-one orthologs. For the other species, we directly used BLASTP to identify best reciprocal hits as one-to-one orthologs. Matches with an E-value below 1e-25 cutoff were considered as orthologs.

2.2.3 Module identification

The Markov Clustering algorithm (MCL) [87] was used to identify modules from each of the combined network for the four species with an inflation parameter of 3.5 that results in an intermediate granularity of the clustering. We also used the –pi option with a value of 5.0 which increases the constant on the edge weights to get a finer grained clustering. Modules with less than 3 genes were discarded from further analyses. MCL was shown to be robust to random edge addition or removal [88], a key issue for noisy genomics data.

2.2.4 Randomization

In order to evaluate the significance of our results, we generated randomized networks for each species and network type that preserve the degree distribution of the corresponding real networks. The randomized networks for each species were aggregated together into a combined randomized network for that species. We applied the same procedure that was used to analyze the real data on these randomized networks. Specifically, we ran MCL on each of the combined randomized network to get randomized modules for each species. Then, for each randomized network in species A, we compare it with the corresponding real network in species B using the randomized modules in A and the real modules in B, and we check how many WMI/BMI in A (randomized) are conserved directly in B (real), and how many edges in A are not directly conserved but their orthologs lie in the same module in B (extended module conservation). 1000 randomizations were performed and the mean and the standard deviation of each percentage were reported.

2.2.5 Matching modules across species

Modules between any two species were matched in the following way. First, the probability of finding *M* orthologs out of *N* genes in each module was calculated using hypergeometric test. In a second stage we calculated the probability of finding *m* genes that are included in the tested module in the other species, out of the *M* orthologs using hypergeometric test. Multiplying the two p-values represents the conditional probability of finding *m* matches between two modules from different species. The p-values were Bonferroni corrected by multiplying by the number of modules. If both of the reciprocal corrected conditional probabilities were below a cutoff of 0.01, we defined the modules as matching.

2.2.6 Matching *S. cerevisiae* modules with protein complexes

For each *S. cerevisiae* module we searched for known protein complexes [29], [30] that were found significantly corresponding in a hypergeometric test. .

2.2.7 Robustness analysis

Defining modules based on GO

In all species separately we defined genes as interacting if they shared at least one term in GO biological process level 7 or below and the GO annotation was defined based on a direct experimental evidence and not computationally. We ran MCL on the GO network in each species of the species separately to assign genes to module in a unique manner and calculated the WMI/BMI statistics in a similar manner to previous modules definitions.

Effect of sequence similarity on conservation patterns

For each of the obtained one-to-one orthologs between *S. cerevisiae* and *S. pombe*, we noted the %identity of the BLASTP match. Different orthology mappings were created by setting cutoffs on the %identity, reflecting increasing confidence in the orthology matching between the two species. The within/between/extended conservation patterns are kept for most mappings and data types. It is important to note that the population of *S. cerevisiae* genes that were still mapped to an *S. pombe* ortholog changed dramatically with the increase of sequence similarity matching, and most genes above a cutoff of 60% are ribosome related.

2.3 Results

2.3.1 Data collection and processing

We focused on four species for which large interaction datasets are available: the two yeasts *S. cerevisiae* and *S. pombe*, the nematode *C. elegans,* and the fruit fly *D. melanogaster*. We retrieved available sequence, expression, protein-protein interaction (PPI), and genetic interaction (GI) data as well as Gene Ontology (GO) annotations for all species. See Methods for details.

To facilitate the comparison of genomic datasets across species, we converted all datasets into network representation using a probabilistic approach that assigns a score to each edge (interaction) between two genes based on their likelihood of participating in the same biological process [89] (see Methods). This method was used in the past [83] to determine appropriate cutoffs for correlation networks in each species (for example the co-expression networks). From this point on, we refer to each data type as a network (e.g., the co-expression network). The co-expression, PPI, positive GI, and sequence networks were combined to create an integrated weighted network separately for each species (Figure 2-1). For each edge in the integrated networks, its score was calculated by summing up the log likelihood scores for that edge across the four individual network types. Integrating the individual data types to a single integrated network for the purpose of creating functional modules follows our hypothesis that interactions are conserved at the network level which may capture better the functional association between the genes or gene products. Therefore, the integrated network represents the most comprehensive functional association aggregation that we are able to achieve for each of the species in our study from the currently available experimental data. We determined orthology relationships using GeneDB [90] and reciprocal best BLASTP hits (Methods). (Results obtained using Inparanoid [16] to define orthology mapping were nearly identical). For a specific network in species *A* we extracted all pairs of genes $g_{A,1}$ and $g_{A,2}$ that are connected in that network. If both genes have orthologs in species *B* we define the interaction $g_{A,1}$- $g_{A,2}$ to be directly conserved if their orthologs ($g_{B,1}$ and $g_{B,2}$) have the same interaction in species *B*.

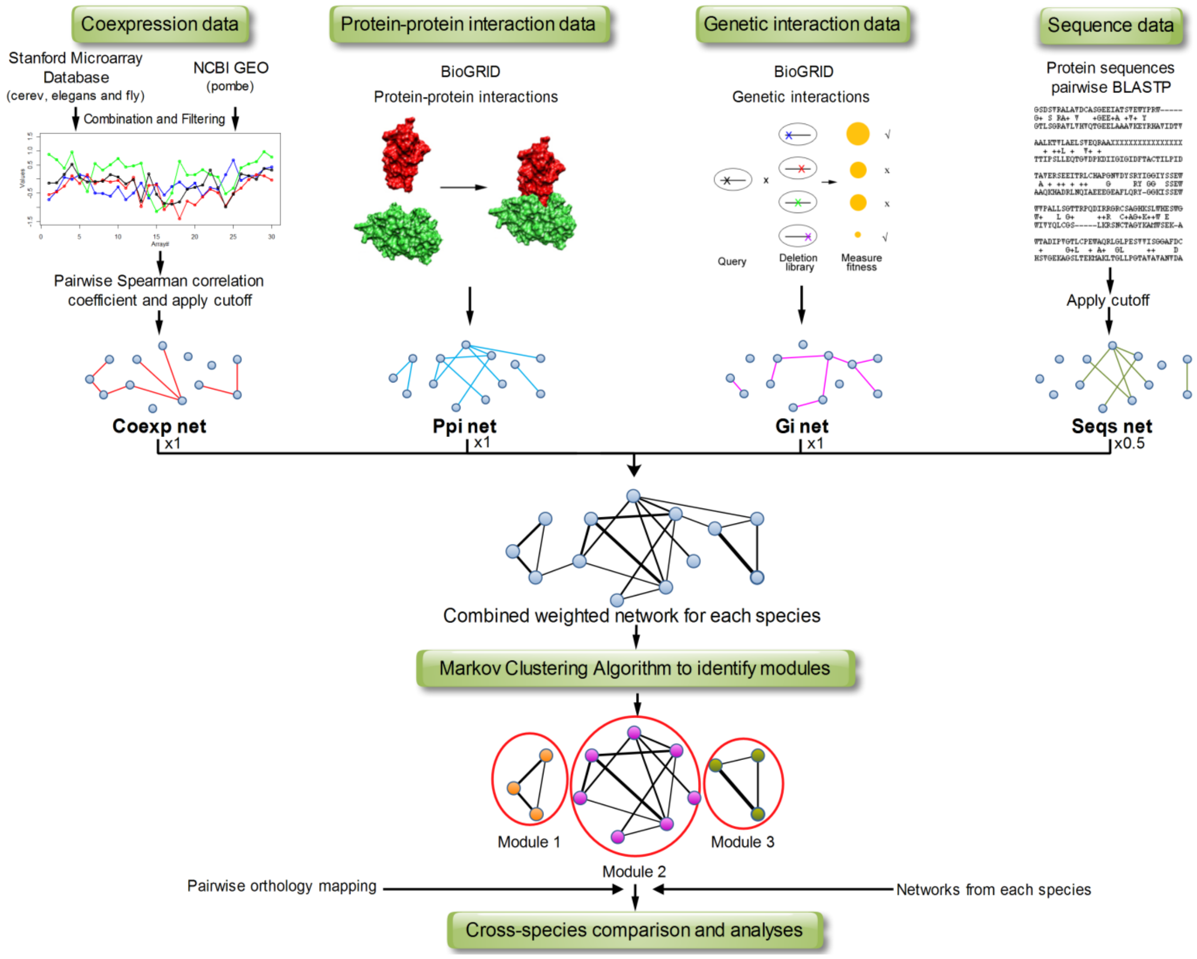

Coexpression data
Stanford Microarray Database (cerev, elegans and fly)
NCBI GEO (pombe)
Combination and Filtering
Values
Array#
Pairwise Spearman correlation coefficient and apply cutoff
Coexp net
x1
Protein-protein interaction data
BioGRID
Protein-protein interactions
Ppi net
x1
Genetic interaction data
BioGRID
Genetic interactions
Query
Deletion library
Measure fitness
Gi net
x1
Sequence data
Protein sequences pairwise BLASTP
Apply cutoff
Seqs net
x0.5
Combined weighted network for each species
Markov Clustering Algorithm to identify modules
Module 1
Module 2
Module 3
Pairwise orthology mapping
Networks from each species
Cross-species comparison and analyses

Figure 2-1: Overview of the modules identification procedure

For each species, available co-expression, PPI, GI, and sequence data were extracted and converted into networks. For PPI and GI the networks representation is straightforward. For co-expression, sequence, and GO we computed a similarity score between genes and used a cutoff to construct a network. Expression, PPI, positive GI, and sequence were combined to create a joint weighted network where the weight is a function of the number of edges connecting two genes. Next, the MCL algorithm was applied on the combined network to identify modules for each species separately. See Methods and Supplementary Methods for details.

We first computed conservation statistics directly from the networks for each species. Most interaction datasets are not well conserved across species, including networks that are fairly complete. The 'Baseline' column in Table 2-1 presents the overall conservation of interaction data (for the integrated networks and for the individual data types) between *S. cerevisiae* and *S. pombe*, the two closest species in our study (with an evolutionary distance estimated at ~400 Mya [91]). The overall conservation of the integrated gene network is 18.11% for *S. cerevisiae* with respect to *S. pombe*, and 22.18% for *S. pombe* with respect to *S. cerevisiae* (we denote this reciprocal comparison as 18.11% / 22.18% from this point on). Of all the types of datasets in our analysis, expression data is the most abundant. However, the co-expression interactions between these two yeasts are only conserved at a rate of 19.27% / 19.51% which is still low, although it is indeed higher than the other experimental data types. In contrast, we find a better agreement between GO edges of the two species (26.59% / 31.81%) despite the relatively low coverage of GO annotation for *S. pombe*.

2.3.2 Conservation of hub interactions

To test whether any of the previously suggested explanations can account for these low conservation rates we analyzed them using our integrated networks. We first checked whether interactions involving hub proteins are more likely to be conserved. In order to examine this, we binned the nodes according to their degrees in the integrated network, and for each bin, we calculated the conservation rates for interactions involving at least one node whose degree falls into that bin. We found a positive correlation between the degree of the nodes and the conservation rates of the interactions that connect them with their partners. Fewer than 15% of the interactions involving nodes with low degrees (up to 300), which include the vast majority of the interactions, are conserved in both *S. cerevisiae* and *S. pombe*, while for those interactions involving nodes with high degrees (600-800) , 24-26% are conserved. Therefore, we conclude that hub interactions are conserved at rates that are better than average, and the effect of hubs should be considered in subsequent analyses. Nonetheless, the conservation rates of hub interactions are still quite low and they provide only a limited explanation for the low conservation rates of all interactions.

2.3.3 Conservation of interactions within protein complexes

Protein complexes were previously shown [76] to have higher conservation rates. This analysis was limited to protein-protein interactions but interactions of other genomic data types that coincide with PPI were also shown to have higher conservation rates [38]. In our analysis, we checked conservation rates for protein complexes that were defined in two recent studies in *S. cerevisiae* [29], [30]. Interactions in the integrated network that were part of the complexes defined by Krogan *et al.* were conserved at a rate of 26.22% (out of 3738 possible interactions), while the 1930 interactions that were part of the complexes identified by Gavin *et al.* had a conservation rate of 35.49%. Note that this is only a one-way comparison, since the complexes are defined only for *S. cerevisiae*. These results show that while conservation rates for interactions within protein complexes are indeed higher, they still do not provide a complete explanation to the question of conservation.

Table 2-2: Conservation statistics between *S. cerevisiae* and *S. pombe*

Conservation rates for *S. pombe* with respect to *S. cerevisiae* are based on the integrated networks for the following categories: Baseline: the entire networks; Hubs: highest rate reported for any bin based on node degree; Complexes: complexes as defined by the Gavin and Krogran studies; Molecular function: highest rate reported for interactions with any GO molecular function; WMI: Within-Module Interactions; WMI – no hubs: WMI excluding interactions with hubs; Extended WMI: extended module interactions. See text for further details.

Baseline	Previous explanations			Module based explanations		
	Hubs	Complexes	Molecular function	WMI	WMI –no hubs	WMI ext.
18.11%	26%	26%/35%	26%	46.54%	42.87%	49.66%

2.3.4 Conservation of interactions by molecular activity

We further extended our analysis to check the hypothesis raised by Beltaro *et al.* [78], that stable interactions are more conserved than transient interactions. We exhaustively examined interactions linking proteins with all molecular function (MF) annotations in GO that contains more than 100 genes in cerevisiae. The average conservation rate for the *molecular function* term (GO:0003674, the root of the GO:MF tree) is similar to the baseline for the GO network (18% / 22% - see Table 2-1). Interestingly, there are big differences for conservation rates for the different MF terms. Interactions that link *transporters* (GO:0005215) exhibit significantly lower rates of conservation probably due to their dynamic nature (8% / 12%). A recent study on three yeast species [49] showed how differential expression of ABC transporters resulted in inherently different mechanisms for coping with an anti-fungal medicine. Interactions linking *RNA polymerase II transcription factor activity* (GO:0003702) also have lower conservation rates (9% / 9%), possibly due to the specific regulation in each of the species and the transient nature of the interaction [78]. However, unlike the proposed solution of Beltaro *et al.*, *kinases* did not show lower-than-average conservation rates, despite the transient nature of their interactions. Interactions connecting proteins annotated with *kinase activity* (GO:0016301), a category that consists of 222 proteins, are conserved at rates of 14% / 23% , but the sub category of *protein kinase activity* (GO:0004672) that contains 135 proteins are conserved at rates of 19% / 29% which is higher than the average. Interactions linking structural ribosome activity (GO:0003735) showed a significant higher-than-average conservation rate (25% / 34%) which is in accordance with previous findings [92]. It is important to note that the size of the molecular function terms did not have any effect on the conservation rates. To conclude, while the molecular function has an effect on the conservation rates of the interactions, we could not establish a clear trend showing that stable interactions are always more conserved than transient interactions. Moreover, even the most conserved category, *RNA binding activity* (GO:0003723), shows only moderate conservation levels (26% / 30%).

2.3.5 Extracting modules from diverse interaction datasets

Our analysis above indicates that the explanation for low conservation rates proposed so far (data type, hub status, protein complex, or protein activity) do not always generalize when applied to comprehensive data from four species. We thus hypothesized that a more general mechanism that combines elements from these proposed directions may be responsible for the low overlap between species. Specifically, we combined different types of interaction data to find gene modules, sets of highly interacting genes that often share similar function. Using these modules we studied the conservation of genomic interaction data at the network level rather than at the individual protein level. We used the Markov CLustering algorithm (MCL) [87] to search for modules in the integrated networks for each species (see Methods). MCL partitions a graph via a simulation of random walks effectively placing each node into exactly one module. Therefore, each module is a set of highly connected proteins and often contains different types of interactions. Since MCL can incorporate edge-weight information, edges that have higher linkage scores or are observed in more than one data type are more likely to be in the same module. MCL was also shown to be robust to random edge addition or removal [93], a key issue for noisy genomic data. Modules that did not include at least 3 nodes were discarded from further analyses. Module sizes follow exponential distribution with very few modules containing more than 100 nodes. As expected, many of the modules are significantly enriched with various functional GO categories. In addition, some of the modules in *S. cerevisiae* significantly overlap protein complexes derived from high throughput experiments [29], [30], though many modules are not related to protein complexes.

To evaluate the significance of our results, we created random networks for each of the real networks we studied for comparison. The random networks retained all the global network properties of the original networks including the distribution of in and out degrees, diameter etc (Methods). We used these random networks to identify random modules and to compare them across species in the same way real modules were identified and analyzed. 1000 random networks were generated for each data type and the results were averaged.

2.3.6 Conservation of functional genomics data on the module level

We divided all interactions into two sets. The first set is 'within-module interactions' (WMI). These interactions connect two nodes that reside in the same module in species *A*. The second set is 'between-modules interactions' (BMI). These interactions connect two nodes that reside in different modules in species *A*. Finally, we defined an interaction as 'extended module conservation' when the interaction itself is not directly conserved, but the orthologs of the two genes connected by the interaction reside in the same module in *B* (see Figure 2-2a). An 'extended module conservation' can indicate either a specific interaction that exists in the other species but so far has not been experimentally tested, or an interaction that is not conserved in the other species, but its functional effect is retained via the module structure (e.g., the interaction is replaced by two interactions that mediate indirectly the same functional effect through existing or new subunits in the module).

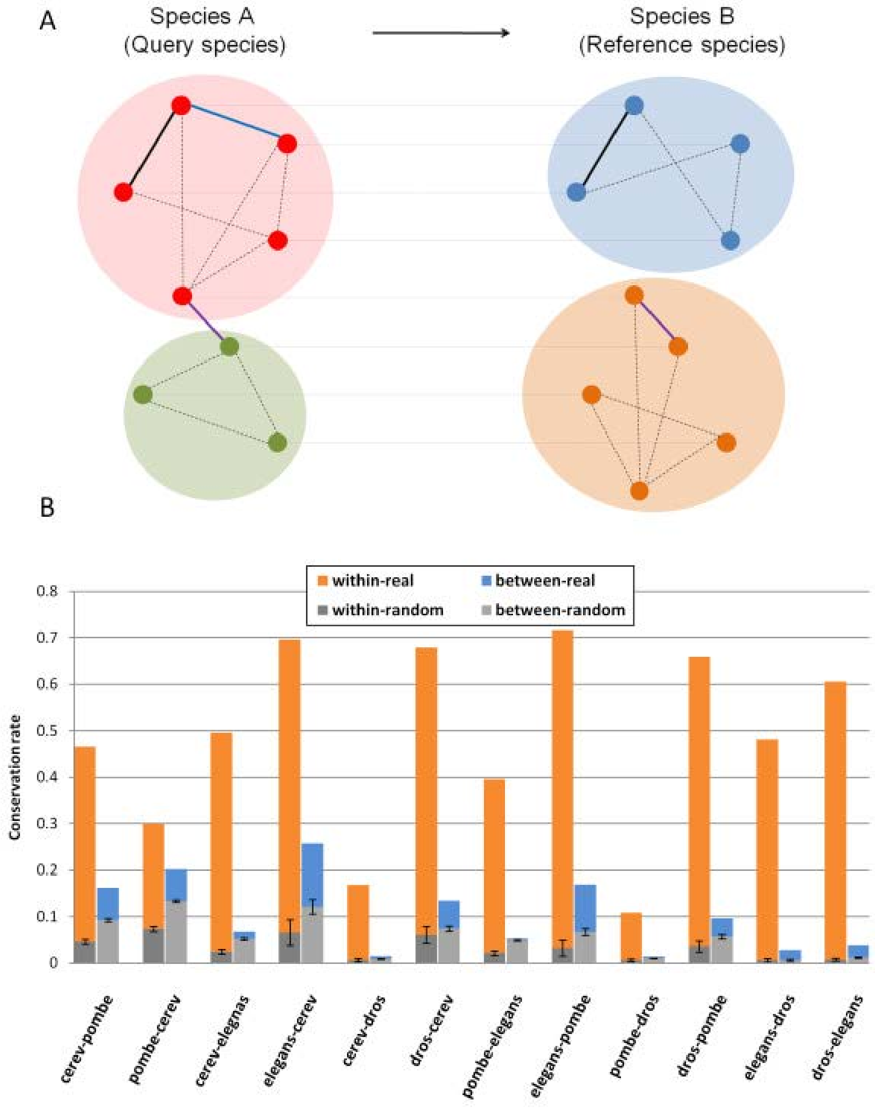


Figure 2-2: Edge conservation across species

(a) Types of conservation. We denote one species as the query species (species A, left) and the other as the reference species (B, right). Shaded groups of nodes represent modules. Nodes connected by a grey line between the species represent orthologous genes. The bold black edge in the upper module of both species is a within-module conservation edge. The purple edge connecting the two modules of species A is a between-modules conserved edge. The blue edge (upper module of species A) is an extended-module conserved edge as both proteins connected by this edge are in the same module in species B. **(b)** Conservation of the integrated network across all pairwise comparisons. Orange bars and blue bars represent within and between conservation rates respectively. Gray bars represent conservation statistics for random modules with error bars showing the standard deviation for 1000 random runs.

Recall that the overall interaction conservation rates between *S. cerevisiae* and *S. pombe* are 18.11% / 22.18%. However, using our modules we show that this is the result of two very different sets of interactions. The WMI conservation rates are much higher. 46.54% / 29.94% of WMIs are conserved between the two yeasts (more than twice the overall conservation for the *S. cerevisiae – S. pombe* comparison and 30% higher than *any* of the previously proposed explanations – see Table 2-1). In contrast, BMI conservation rates are lower than the overall conservation rates at 16.17% / 20.16%. To rule out the possibility that our results merely reflect the effect of hubs that might be more abundant in modules, we excluded hubs (nodes with degrees of 300 or higher) from our analysis. The WMI / BMI conservation statistic became even more distinct; while WMI conservation remained almost the same or better (42.87% / 33.31%), BMI conservation rates dropped (4.06% / 2.92%). These trends hold for almost all other types of genomic data as well (see Table 2-2).

Table 2-3: Conservation rates of edges in different types of networks between *S. cerevisiae* and *S. pomb*e

Conservation rates are listed for the following categories: Baseline: the entire networks; BMI: Between-Module Interactions; WMI: Within-Module Interactions; Extended WMI: extended module interactions. (no-seqs): statistics based on integrated network that does not include the sequence network. (exclude-para): in addition to 'no-seqs', all edges connecting paralogs (nodes with BLASTP E-value cutoff of 1e-25 or less) were removed.

		From *S. cerevisiae* to *S. pombe*				**From *S. pombe* to *S. cerevisiae***			
		Baseline	BMI	WMI	Extended WMI	Baseline	BMI	WMI	Extended WMI
Integrated	Real	**18.11**	**16.17**	**46.54**	**49.66**	**22.18**	**20.16**	**29.94**	**31.97**
	Rand	9.13±0.04	9.26±0.32	4.66±0.48	5.22±0.49	11.99±0.06	13.31±0.30	7.38±0.57	7.71±0.59
Integrated (no-seqs)	Real	**16.89**	**15.61**	**38.54**	**40.99**	**20.86**	**15.88**	**34.25**	**35.03**
	Rand	9.04±0.05	9.68±0.30	4.57±0.50	5.05±0.52	11.84±0.05	12.72±0.21	8.01±0.60	8.34±0.61
Integrated (exclude-para)	Real	**16.84**	**15.59**	**38.44**	**40.89**	**20.77**	**15.83**	**34.06**	**34.84**
	Rand	8.92±0.05	9.58±0.30	4.47±0.49	5.38±0.53	11.79±0.05	12.68±0.21	7.95±0.60	8.24±0.60
Coexpression	Real	**19.27**	**18.27**	**36.28**	**40.26**	**19.51**	**18.76**	**20.30**	**21.74**
	Rand	10.32±0.05	10.2±0.38	6.71±0.78	7.12±0.75	11.09±0.05	12.27±0.30	8.06±0.70	8.46±0.71
PPI	Real	**1.78**	**1.46**	**5.82**	**25.90**	**57.96**	**56.94**	**71.02**	**76.33**
	Rand	0.06±0.01	0.06±0.02	0.05±0.09	1.42±0.43	3.12±0.42	3.70±1.10	2.31±1.33	2.62±1.48
Positive GI	Real	**2.24**	**1.77**	**8.28**	**33.93**	**10.02**	**8.26**	**21.20**	**36.96**
	Rand	0.30±0.05	0.29±0.09	0.15±0.19	1.68±0.61	1.43±0.27	1.50±0.45	1.19±1.27	1.73±1.50
Negative GI	Real	**2.86**	**2.60**	**7.53**	**43.08**	**15.14**	**14.67**	**32.90**	**56.77**
	Rand	1.09±0.05	0.96±0.13	1.37±1.96	2.89±2.78	7.56±0.29	7.17±0.71	9.95±10.16	10.98±10.68
GO	Real	**26.59**	**26.41**	**45.87**	**61.69**	**31.81**	**31.47**	**39.70**	**57.81**
	Rand	2.23±0.08	2.16±0.13	2.27±2.12	3.78±2.96	4.05±0.11	4.28±0.15	4.11±2.58	5.22±2.88
Sequence	Real	**90.16**	**90.18**	**90.15**	**97.33**	**76.92**	**51.40**	**79.73**	**89.66**
	Rand	17.55±0.64	25.61±1.6	1.23±0.76	1.96±0.86	14.53±0.39	28.88±1.59	0.09±0.15	0.34±0.30

Random data does not display similar trends (Figure 2-2b). In fact, in clear contrast to the observations on the real modules, statistics for the modules based on the random networks showed that the averages of the BMI conservation ratios are higher than WMI conservation for all genomics data types and species comparison, indicating that results for real data are a function of strong non-random selection bias (Figure 2-3). None of the 1000 random networks we generated led to conservation rates seen in the real networks (p-value < 0.001). In fact, the rates obtained for all random networks were significantly lower than those observed for the real networks indicating that there is evolutionary pressure to maintain module conservation.

The conservation rates of extended-WMI are even higher (49.66% / 31.97%, Table 2-3), while extended-BMI rates have only moderately increased (16.91% / 20.79%), indicating that even if the specific interaction type is not observed in the other species, it may be that either it is actually present but was not measured, or that its effect is mediated indirectly through other members of the module.

We extended this analysis to all 12 pairwise species comparisons (note that the comparisons are not symmetric since the analysis depends on the query species, see Figure 2-2a). Figure 2-2b presents the results for all comparisons across the different data types (See also Figure 2-3). It can be seen that while the overall conservation rates change according to the distance between the species and the coverage of the specific data types, the overall trend is similar in all comparisons. Overall WMIs are more conserved than average, yet they are much less conserved in the random networks. Extended module conservation further increases the conservation rates. The only interaction type for which most comparisons do not show an improvement is negative GI. Indeed, negative GIs are often found between genes in parallel pathways rather than within the same pathway [38], so they are not expected to be conserved via modules.

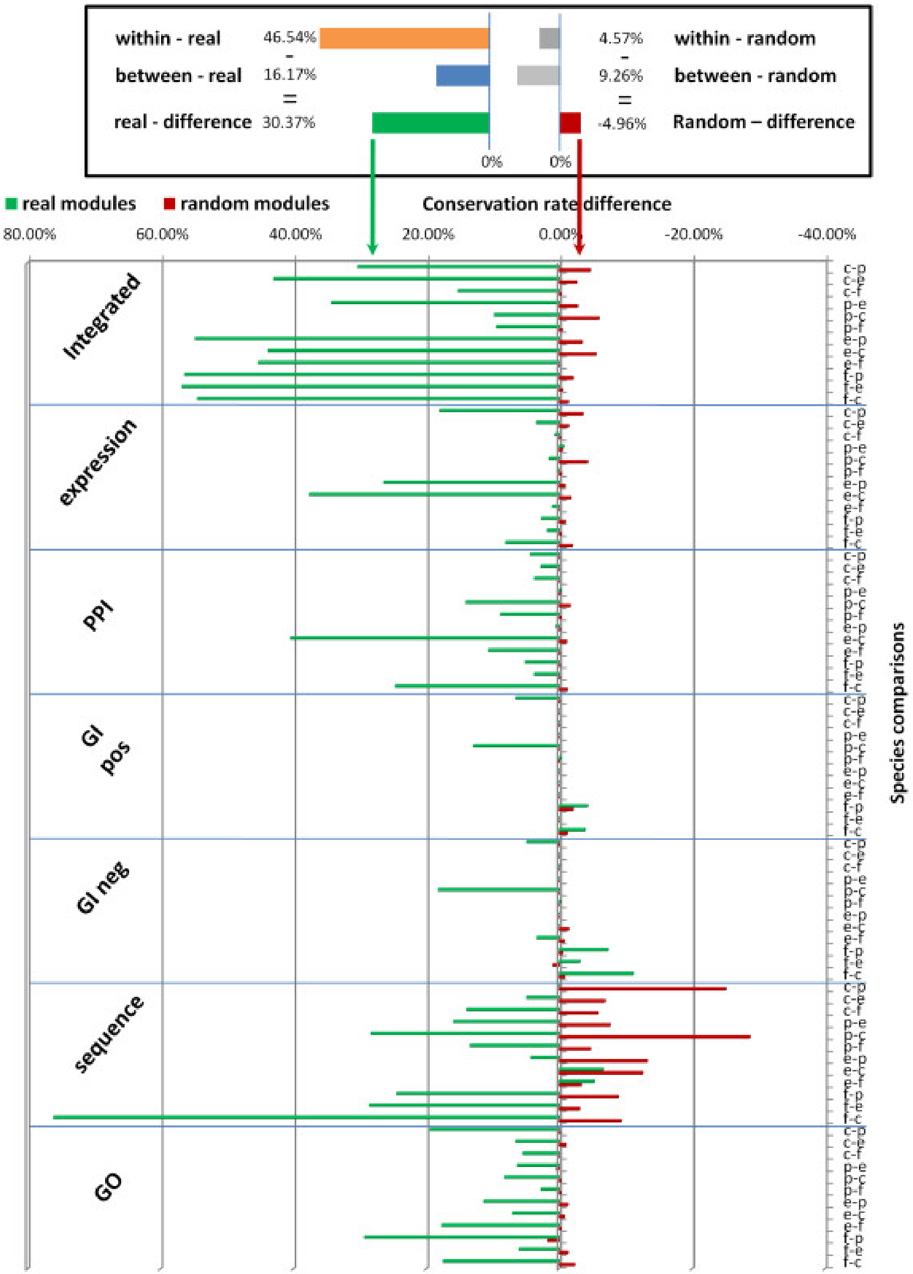


Figure 2-3: Differences between WMI and BMI conservation rates across all pairwise comparison Green bars and red bars represent conservation statistic for real and random modules respectively. The bars represent the difference between WMI and BMI conservation rates (darker green and red) and the difference between extended WMI and extended BMI conservation rates (brighter green and red). The species are indicated on the vertical axis as follows (c-*S.cerevisiae*, p-*S.pombe*, e-*C.elegans*, f-*D.melanogaster*). For most data types the improvement for the real networks is very large. In contrast, for random networks the within module edges are usually less conserved when compared to the overall conservation indicating that the within module conservation bias is even stronger.

2.3.7 Robustness analysis

In addition to using random networks as a control we carried out several other experiments to test the robustness of our findings and show that they are independent of the way the modules are defined, the amounts of data that are being used, or the orthology matching definitions.

To rule out the possibility that the WMI:BMI statistics are a result of the way the modules definition and parameter selection, we used an alternative graph clustering method, SPICi [94], to partition the networks into modules and ran the same analyses. SPICi uses heuristic approach to greedily build clusters from selected seeds. Also under this graph clustering scheme, WMIs are shown to be conserved at higher rates than BMIs for almost all species comparison and data types. In addition, we tested conservation rates for modules that are based on previous knowledge rather than clustering the interaction data. We created modules based on gene ontology terms that are defined based on direct experimental evidence only (precluding annotations that are defined by sequence similarity to avoid bias in the reported results, see Methods). While the resulting networks and modules are smaller and less comprehensive compared to our interactions data, the conservation trends for the GO-based modules are similar to the modules based on interaction data. All together, these results show that our conclusions hold and are independent of the way the modules are defined, as long as there is a strong functional relationship within the module.

We also studied the effect of insufficient data coverage on our results. Missing data is the most common reason for differences between the true biological networks and our integrated networks. This is more likely to be the case for species other than *S. cerevisiae*, as fewer experiments for all data types were conducted. To this end, we randomly removed edges from the *S. cerevisiae* network and generated modules that are based on the trimmed networks. Calculating the conservation rates against *S. pombe* showed that in all cases our results regarding the large increase in WMI and extended-WMI conservation still hold (see Figure 2-4a). Also, many of the modules from the full *S. cerevisiae* network were significantly retained in the trimmed networks (see Figure 2-4b).

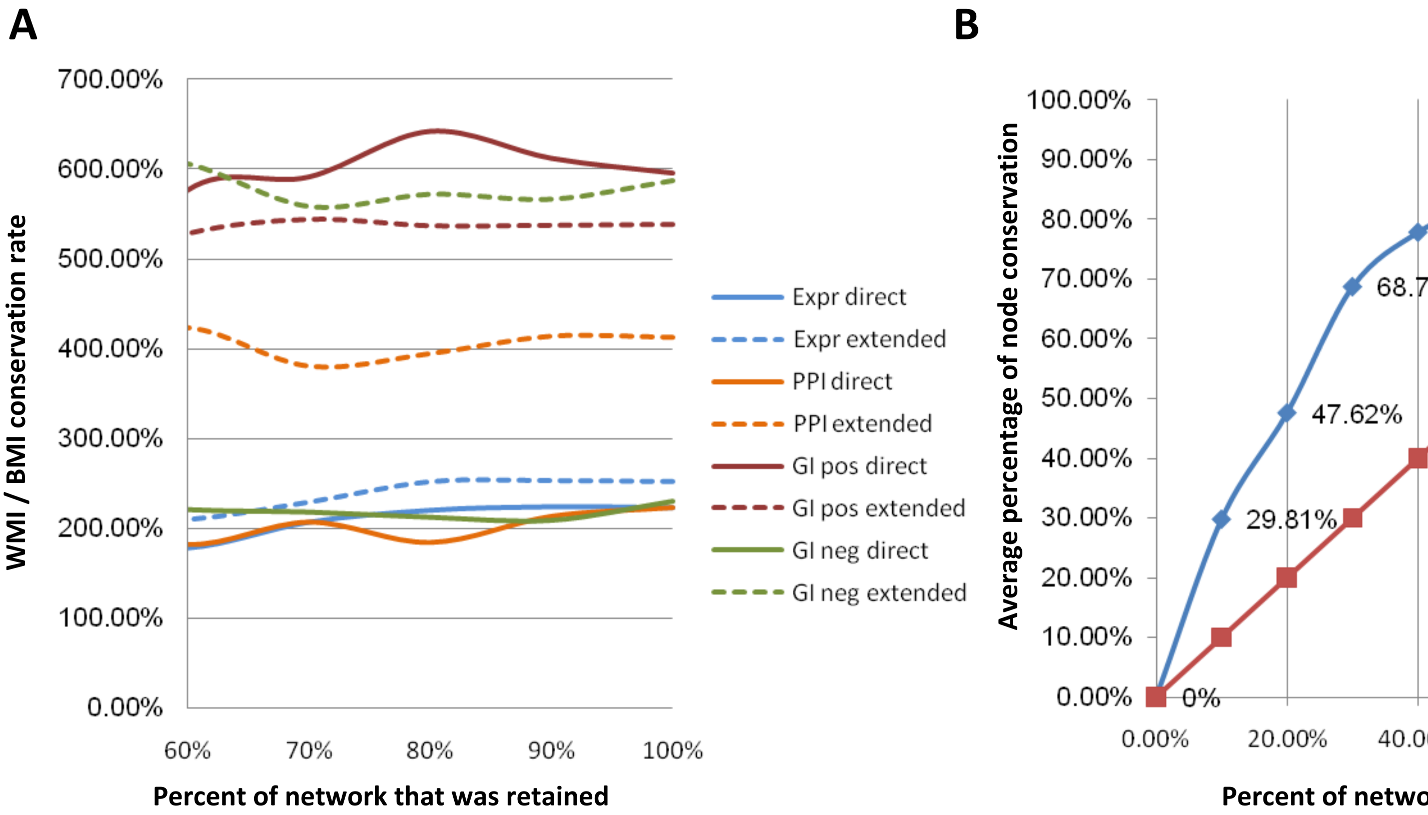


Figure 2-4: Robustness analysis

(a) Ratio of within-module / between-module edge conservation results of *S. cerevisiae* modules that are based on varying sizes of the interaction data compared to *S. pombe* as reference. The X-axis indicates the percent of randomly picked *S. cerevisiae* edges out of the entire network that were used for the modules search. **(b)** The average percentage (Y-axis) of node conservation between the *S. cerevisiae* modules that were constructed based on the full interaction network compared with modules constructed over varying sizes of interaction data as a percentage out of the entire network (X-axis).

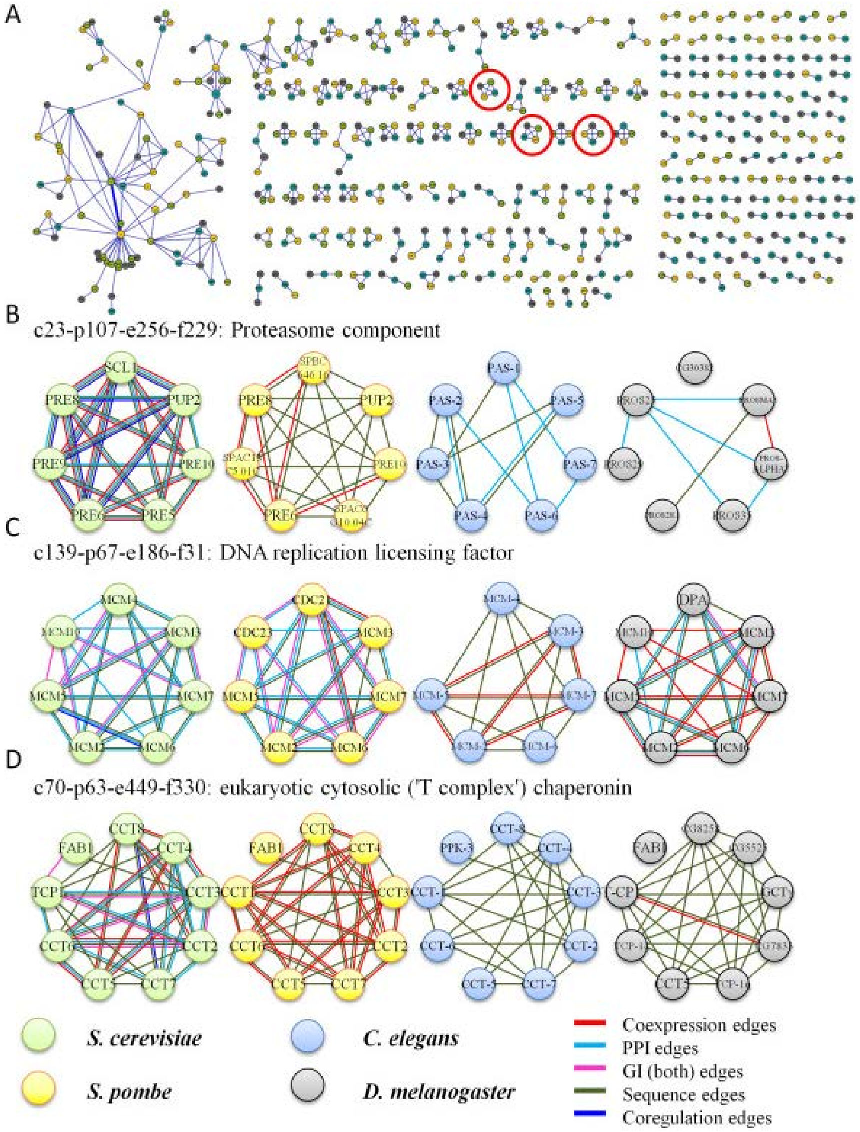
To rule out the possibility that our results are affected by the orthology definition we repeated the analysis using Inparanoid [16] mapping. Very similar results to the ones presented above were achieved for the one-to-one mappings generated from Inparanoid (not shown). Furthermore, we checked whether using many-to-many (*M*:*N*) Inparanoid mapping would change our results. Conservation definitions are slightly changed under *M*:*N* mapping definitions. We marked an edge as conserved in the query species if any edge between possible orthologous nodes in the reference species was conserved. While conservation statistics for both WMI and BMI in almost all species and data types naturally increased using the new definitions, WMI higher conservation rates trend is retained for most comparisons.

We further evaluated the effect of stricter orthology mappings on the conservation patterns. We tried various orthology mappings between *S. cerevisiae* and *S. pombe* by keeping only high confidence orthology matching between the two species (Methods). Stricter orthology mapping corresponded to fewer interactions whose functions are known to be more conserved (e.g., the ribosome complex), and showed similar or higher WMI / BMI conservation rate patterns for most comparisons.

Lastly, we evaluated our results using an integrated network that included only the co-expression, PPI, and GI positive and did not include the sequence networks to rule out the possibility that our results are driven by paralog conservation. The trends we observed for our original analysis remained the same for this smaller network indicating that our module based conservation result is robust to the type of data used (see the “no-seqs” row in Table 2-3). Moreover, we created an additional network in which we further excluded all interactions (regardless of their type) connecting two nodes (genes) with BLASTP E-value cutoff of 1e-25 or less in all species. We observed the same trends for this network as for the other networks we analyzed (see the “exclude-para” row in Table 2-3) indicating that module-based conservation is a general trend that is independent of sequence conservation.

2.3.8 Conservation of modules across species

Having established the within-modules conservation trend, we asked whether the modules themselves are conserved (in terms of membership) across the species. For this we extracted all modules with at least three members resulting in 741 modules for *S. cerevisiae*, 523 for *S. pombe*, 1484 for *C. elegans* and 1237 for *D. melanogaster*. For each such module we computed the significance of its overlap with all modules in the other three species (see Methods). For *S. cerevisiae*, 131 modules were found to match *S. pombe* modules, with a reciprocal p-value < 0.05 (based on hypergeometric test and corrected for multiple hypothesis testing, see Methods). This number, which is 25% of all *S. pombe* modules, is high considering coverage limits. A total of 562 matches were found for all species comparisons. Figure 2-5a shows a graph with significant reciprocal matches between the modules. We next examined modules that are conserved among all species in our analysis, and 33 such groups were found, spanning various functional categories like signal transduction, protein folding, metabolic processes and many others. Figures 2-5b,c,d present some examples of such modules. The module matches are based on the nodes, nevertheless these examples show that relatively little rewiring (especially in the integrated network) had occurred between orthologous proteins that participate in these modules. Modules may also contain other proteins that do not have an ortholog.



c23-p107-e256-f229: Proteasome component

c139-p67-e186-f31: DNA replication licensing factor

c70-p63-e449-f330: eukaryotic cytosolic ('T complex') chaperonin

Figure 2-5: Matching modules between species

(a) Module matching. Green, yellow, blue, and grey nodes correspond to modules in *S. cerevisiae*, *S. pombe*, *C. elegans*, and *D. melanogaster* respectively. The size of a node corresponds to the number of genes in the module. The width of an edge connecting two nodes reflects the p-value of the reciprocal match between two modules, when more significant matches correspond to wider edges. **(b-d)** Examples for matched modules across the four species. Each row contains modules that significantly overlap based on orthology for all pairwise comparison. The examples are marked in a red circle in Figure 2-5a. The nodes are colored with the same color scheme of a. The edges are colored based on the interaction type (see legend – note that GI edges refer to both positive GI and negative GI edges), and multiple edges between two nodes are allowed. For clarity, only genes that have orthologs in at least one of the other modules are shown. See text for details on the matched modules.

Figure 2-5b shows orthologous proteins from modules c23-p107-e256-d229. These modules were significantly enriched for proteolysis and are part of the proteasome complex. *S. cerevisiae*, the most extensively studied organism in our study, shows many interactions from the various networks like co-expression, PPI, and sequence, and even other types of interactions like genes that are co-regulated by the same transcription factor [95], which were not used in the module construction process. Many of the PPI interactions in the *S. cerevisiae* module are retained in the matched *C. elegans* module, and we can suspect that similar interactions should be experimentally found in *S. pombe.* The many similar co-expression edges observed for *S. pombe* indicate that these proteins are probably present at the same time in the cell, which make them likely to form PPI. Similarly, Figure 2-5c shows orthologous proteins from modules c139-p67-e186-d31 that are all enriched for DNA replication in the S phase of the mitotic cell cycle. *S. cerevisiae* and *S. pombe* exhibit very similar patterns of PPI and GI, which were not measured for *C. elegans*.

Nonetheless, the co-expression and sequence edges indicate that it is likely that the PPI and GI edges should be present in *C. elegans* as well. Figure 2-5d shows an example for modules enriched for protein folding. *S. pombe* exhibits many co-expression edges, especially with TCP1/CCT1 that are absent in *S. cerevisiae*. Nonetheless, many of these edges are present in *S. cerevisiae* as PPI edges, a fact that might indicate that these modules operate in a similar manner in both species, as PPI are more likely to be co-expressed.

2.4 Discussion

Our results indicate that while overall interactions at the node (protein) level are conserved at low rates, interactions within modules are conserved to a much greater degree. This raises the intriguing possibility that interactions are conserved on a level different from that of the individual genes. In other words, while there is a strong selective pressure to maintain interactions within a module, there is less pressure to maintain between-module interactions.

The within-module conservation statistics that are presented in this study are probably an underestimate for the real conservation rates due to the incompleteness of interaction data [72]. Our results are robust considering varying the amount of available data (and coverage), when compared to random interaction networks, across all four species we studied. Many of the modules we discover independently in each species are significantly conserved across more than one species, and we expect this number to grow once additional data becomes available. This refined understating of conservation may lead to better cross species search tools that can utilize the network context in addition to sequence similarity.

Our results also shed new light on some recent discoveries about the relationships between genes associated with very different phenotypic outcomes in close species[96]. The results suggest that while modules are conserved, interactions between modules may change more rapidly, allowing modules involved in a specific function in one species to become involved in a different function in another species through interactions with other modules.

A possible analogy to our proposed view for module conservation is sequence conservation (Figure 2-6). When looking at the sequence similarity between close species, we see that the overall similarity is lower than the similarity of the coding regions, as there is less evolutionary pressure to preserve intergenic regions. Similarly, the overall network similarity is lower than the similarity of the modules, as there is less evolutionary pressure to preserve between-modules interactions. There are also cases where some nucleotide substitutions in coding regions result in functionally similar proteins (e.g., synonymous mutations or mutations that retain the physical properties of the amino acids). Likewise, changes in within-module interactions can result in functionally similar modules, and can be explained by redundancy or indirect interactions via a third protein, as long as the two proteins remain in the same module. This network organization structure allows both robustness (as modules often stay the same across species) and flexibility (by changing the interactions between modules) which may confer advantages in evolving species.

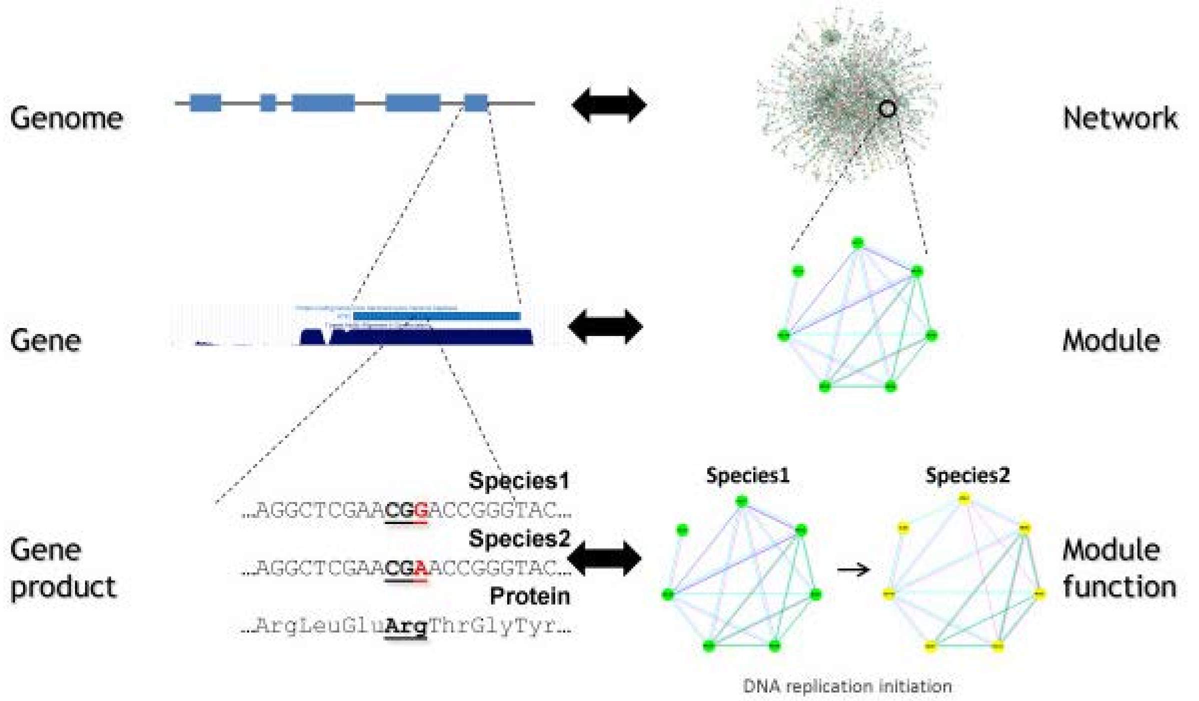

Genome
Network
Gene
Module
Species1
...AGGCTCGAACGGACCGGGTAC...
Species2
...AGGCTCGAACGAACCGGGTAC...
Protein
...ArgLeuGluArgThrGlyTyr...
Gene product
Species1
Species2
Module function
DNA replication initiation

Figure 2-6: Module conservation is analogous to sequence conservation

For sequences (left) coding regions are usually much more conserved than the genome as a whole. Similarly, in the network setting, modules are more conserved than the entire network. In addition, coding regions can often tolerate synonymous mutations that change the DNA sequence itself but do not alter the protein product. Similarly, modules may be able to tolerate loss of specific interactions as long as the two interacting orthologs remain in the same module (often through redundant interactions or interactions with other module members).

Our results indicate that although individual interactions in one species are generally conserved at low levels when compared directly with a closely related species, interactions within functional modules are much more likely to be conserved. Therefore, the networks are still conserved at the functional module level and biological processes they rely on such modules including the cell cycle and protein synthesis, are also well conserved. In contrast, interactions between functional modules are usually conserved at a lower rate than the general case. This may introduce flexibility in the evolution of networks since such between-module interactions can change more rapidly, allowing modules involved in a specific function in one species to become involved in a different function in another species through interactions with other modules.

3 *Expression Blast* – Comparing Expression Data Within and Across Species

In the previous chapter we have shown that expression data that is part of a higher network organization level is conserved across species. In this chapter we present a method for utilizing this observation to aid the comparison of expression data from multiple species. We developed a computational approach and a web portal that allow querying and comparing expression data against a large compendium of expression experiments within and across species. This method enables researchers to easily identify experiments with correlated and anti-correlated expression signature when compared to a query expression set. The method supports several comparison metrics, utilizes text analysis, is integrated with additional databases and provides an easy to use GUI. We used this tool to study expression data from SIRT6 transgenic mice and found several experiments that seem to trigger parallel and / or related pathways including PPARα and LXRα. Comparison of the mice data with human revealed that human studies that had a correlated profile were related to female tumors include ovarian cancer and breast cancer while human studies that showed an anti-correlated profile were related to male tumors including prostate cancer. These female tumors are tightly related to the female hormone estrogen and may suggest a possible regulation of estrogen by SIRT6.

3.1 Introduction

Expression studies, both using microarrays and RNA-Seq, are among the most popular methods for measuring dynamic, condition-specific responses of complex biological systems.

The number of gene expression datasets uploaded to public databases is growing exponentially (see Figure 1-3), in part due to requirements imposed by journals and funding agencies. One of the biggest expression repositories, GEO [22] (Gene Expression Omnibus), contains hundreds of thousands of expression experiments grouped into dozen of thousands of series. Other notable repositories for expression data are ArrayExpress [97], and SMD [98], but there are also several smaller expression datasets, usually dedicated to specific species, tissues, or diseases. In addition, several large pharmaceutical companies maintain large proprietary expression databases. See Table 3-1 for further details.

Table 3-1: Expression data repositories

Various public repositories for gene expression data divided into three categories: general, species specific, and tissue / disease specific.

Name	URL	Comments
General		
GEO [22]	http://www.ncbi.nlm.nih.gov/geo/	
Array Express [97]	http://www.ebi.ac.uk/arrayexpress/	
SMD [98]	http://smd.stanford.edu/	
Species Specific		
MGI	www.informatics.jax.org	Mouse database
SGD	www.yeastgenome.org/	Yeast database
PlexDB	http://www.plexdb.org	Plants database
NASCarrays [99]	http://affymetrix.arabidopsis.info/	Arabidopsis database
M3D	http://m3d.bu.edu/cgi-bin/web/array/index.pl?section=home	Microbes database
Tissues / Diseases Specific		
Allen Brain Atlas	http://www.brain-map.org/	Human and mouse brain tissues
4DXpress [23]	http://4dx.embl.de/4DXpress	Development dataset
TCGA	http://tcga-data.nci.nih.gov/tcga/	Dedicated to cancer
Oncomine	www.oncomine.org	Dedicated to cancer
caArray	https://array.nci.nih.gov/caarray/	Dedicated to cancer

While accessible, this data is not well organized. Even within a single species there are many different platforms with different probe identifiers and different value scales. This situation prevents researchers from mining this wealth of data to conduct large scale analysis across different expression studies performed in multiple labs. In addition, many datasets provide only cryptic annotations and there is no systematic way to distinguish treatment and controls (e.g., cancer tissue vs. healthy tissue). There are several tools aiming to integrate data from several platforms usually in order to identify differentially expressed or co-expressed genes across multiple experiments. The ease of use, clarity of output, and the amount of data vary considerably; see Table 3-2 for details. Please note that Table 3-2 does not list tools involved in other phases of gene expression analysis including data normalization, searching for differentially expressed genes, identifying and partitioning expression patterns, gene annotation, pathway analysis, and network analysis. See review by Olson [100] for relevant tools regarding these analysis phases.

Existing tools also do not allow users to compare their new expression results to previous expression experiments. Such analysis can provide both validation of specific hypothesis regarding the underlying causes of a specific treatment and a method for raising further hypotheses regarding the underlying response mechanisms. Only two previous tools that we are aware of, ProfileChaser [101] and cellMontage [102], allow users to compare expression values to other experiments taken from large scale repositories. However, both tools are quite limited in their offerings. ProfileChaser allows comparing only against curated GEO datasets (GDSs) and only using all the expression data at hand using a technique introduced in [103] that is based on a reduced set of gene expression features to reduce the dimensionality of the data. cellMontage also does not integrate expression data from different platforms and allows only platform-specific queries. Since there are over ten thousand platforms as of April, 2012, this type of querying is quite cumbersome to work with. Moreover, treatment and control values were not identified and users can query only against the raw values of the expression experiments. Therefore, comparison of differentially expressed genes cannot be done using the most prevalent 1-color arrays.

In addition, there is no current tool that supports cross species queries. Cross species plays a crucial rule in drug development with pre-clinical studies often performed on model organisms. The ability to quickly predict whether a certain experiment conducted on lower mammals is likely to have a similar effect on human has the potential to greatly expedite the drug discovery process. Nonetheless, comparing expression results across species is even more complicated due to orthology considerations.

Table 3-2: Large scale expression querying tools

Various methods currently available to inspect various aspects of expression data from large scale repositories.

Name	URL	Description
GEO profiles [22]	http://www.ncbi.nlm.nih.gov/geoprofiles	Finds profile neighborhoods for specific genes using GEO data.
EBI – Atlas [97]	http://www.ebi.ac.uk/gxa/	Allows identifying strong differential expression candidates in conditions of interest using ArrayExpress data.
Gemma	http://www.chibi.ubc.ca/Gemma/	Allows searching for differential expression and co-expression. Allows input for only one gene against a predefined set of experiments chosen by keywords. Based only on curated data for only six species.
Genevestigator [104]	https://www.genevestigator.com/gv/biomed.jsp	Query conditions and find genes affected by them or query genes by name and find which conditions affect their expression. Based only on curated data for fifteen species.
NextBio	http://www.nextbio.com	Find conditions for specific genes by name using a pre-defined correlation matrix
ProfileChaser [101]	http://profilechaser.stanford.ed	Allows searching for similar gene expression profiles but only against curated GEO datasets and using the entire expression data.
Cell Montage [102]	http://cellmontage.cbrc.jp	Allows searching for similar gene expression profiles by query genes with expression numerical values but only for per platform.
MADtools	http://cardioserve.nantes.inserm.fr/madtools/home/	Finds genes that are co-expressed with a list of input genes
SPELL (yeast) [105]	http://spell.yeastgenome.org/	Finds co-expressed genes with query input for yeast only
Sigma-Aldrich	http://www.sigmaaldrich.com/life-science/your-favorite-gene-search.html	Finds relevant microarray based on GEO data for a specific gene input
CMap [106]	http://www.broadinstitute.org/cmap	Finds correlation between gene input and specific predefined expression datasets of human cells treated with bioactive small molecules.

In this study, we developed a computational tool that gathers expression data from GEO, processes it uniformly across all experiments, species, and platforms. The framework identifies technical replicates and classifies the experiments to control and treatment. The uniform processing of the data allows users to compare their own new results within and across species through a web interface. A summary and visualization of the closest database matches to the query input are provided together with detailed information for each match.

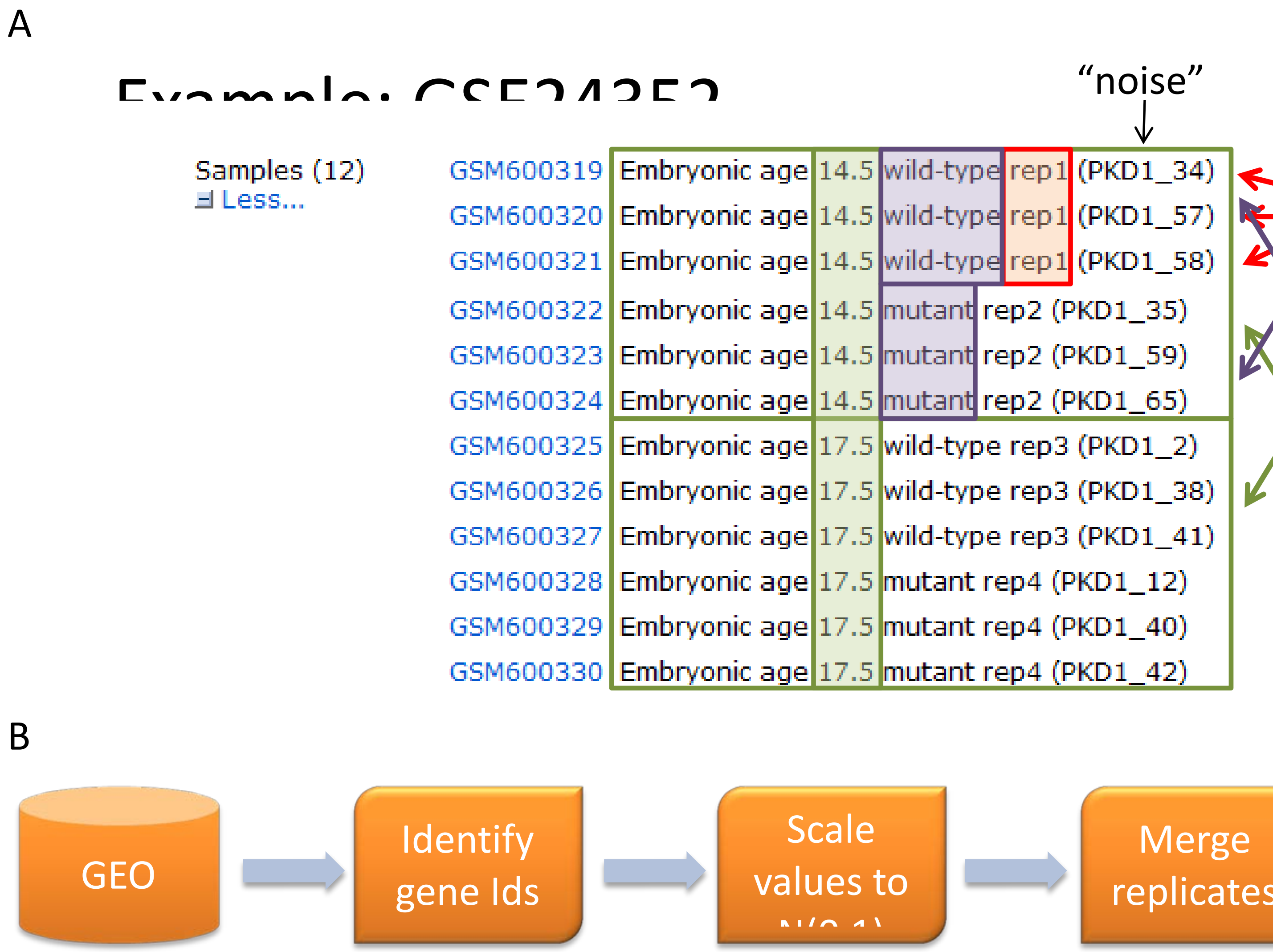
We have applied *ExpressionBlast* to mouse expression data from our previous study that demonstrated increased life-span for male mice following SIRT6 over-expression [61]. We found several studies that trigger related pathways including LXRα and PPARα. Follow up experiments revealed high correlation on downstream genes in both SIRT6 over-expression and LXRα activation. Earlier studies showing similar effects on LDL cholesterol levels independently by SIRT6 over-expression and LXRα activation together with dependency tests, may indicate on a mechanism of regulation of LXRα by SIRT6 in a similar manner to the regulation of LXRα by SIRT1 [107].

Comparison of the mice data with human revealed the studies that were correlated with our mouse results were enriched with female tissue tumors, while studies that were anti-correlated with out mouse results were enriched with male tissue tumors. This observation led to a hypothesis suggesting that SIRT6, through LXRα and PPARα, can regulate estrogen, the female hormone, levels. This may explain why the lifespan extension phenomena was observed only in male mice and may tie SIRT6 as a possible treatment for breast cancer.

3.2 Methods

3.2.1 Data Processing

We built an automatic system for parsing and processing expression data collected from GEO [22]. A uniform processing of all the GSMs includes matching probe identifiers to a single gene id reference, log 2 transformations of raw expression values, and normalization of the values to the same mean and variance (see Figure 3-1). This uniform treatment of the data enables comparisons across platforms and species. The metadata of entire series (GSE) and each experiment (GSM) is analyzed to identify relevant annotations in two steps. First, technical replicates are identified using text analysis and are merged for increasing the confidence in the results. Second, the expression experiments are classified to treatment and control cases (e.g., cancer tissue vs. healthy tissue) for understanding condition specific expression patterns, and the relevant treatment and control experiments are paired to enable the identification of differentially expressed genes on each experiment. A special consideration is given to cases where multiple controls and treatments are present in the same GSE. In addition, time series studies are being identified and the first time point in each series is marked as the control case.



Samples (12)
Less...

GSM600319	Embryonic age 14.5 wild-type rep1 (PKD1_34)
GSM600320	Embryonic age 14.5 wild-type rep1 (PKD1_57)
GSM600321	Embryonic age 14.5 wild-type rep1 (PKD1_58)
GSM600322	Embryonic age 14.5 mutant rep2 (PKD1_35)
GSM600323	Embryonic age 14.5 mutant rep2 (PKD1_59)
GSM600324	Embryonic age 14.5 mutant rep2 (PKD1_65)
GSM600325	Embryonic age 17.5 wild-type rep3 (PKD1_2)
GSM600326	Embryonic age 17.5 wild-type rep3 (PKD1_38)
GSM600327	Embryonic age 17.5 wild-type rep3 (PKD1_41)
GSM600328	Embryonic age 17.5 mutant rep4 (PKD1_12)
GSM600329	Embryonic age 17.5 mutant rep4 (PKD1_40)
GSM600330	Embryonic age 17.5 mutant rep4 (PKD1_42)

Figure 3-1: Data sample and processing

(a) An example for the difficulties in processing the free text descriptions attached to each GSM. GSE24352 contains 12 GSMs that can be divided to two sub series, each having three replicates for the WT and mutant. The text analysis system needs to know how identify all the different components in these descriptions, and ignore the unrelated identifier in the end of each description. **(b)** The steps applied to the expression data in order to enable querying across experiments, platforms, and species.

3.2.2 Data input

The query requires the following mandatory inputs:

1. Name of the query (input) species
2. Namespace or nomenclature schemes of the input genes
3. Type of the input expression values (e.g., treatment/control ratio or the raw unclassified values)
4. Query set gene names and their corresponding values
5. Name of the reference (output) species - the species for which database matches will be returned
6. The distance function heuristic to be used for the query (See Search engine matching process for details).

The query set for the search engine consists of a set of gene identifiers and their corresponding values which are assumed to come from a value distribution of mean of zero and standard deviation of one. For speed of processing, this set is recommended to be of a limited size (~40 genes), though the algorithm can handle efficiently larger input query sets as well. Gene names can be of various namespaces and are converted to Unigene identifiers.

In addition, the following inputs are optional:

1. Filtering the results by keywords
2. Limits for the number of matches to be returned as output
3. Limit for the minimal number of genes in the query set that should appear in the output
4. Various parameters that control the weighted Euclidean distance function

See 'Search engine matching process' for further details on the role of the optional parameters. See Chapter 6 for user interface details and screen shots.

3.2.3 Search engine matching process

To explain the search procedure, we define all experiments in the database that are considered for comparison with the input set as *Possible Matches* (*PM*s). Due to platform differences, many PMs do not have measurements for all the genes in the query set. We therefore define a user parameter to control the minimal number of genes in the query set that should be matched for each PM (set to 65% by default).

We support several distance functions to identify the optimal matches for a query set:

- Weighted Euclidean Distance:

$$Weighted\ Euclidean(input, PM) = \frac{\sum_{g\ \in input \cap PM} \mathrm{W}(g_i) * (\mathrm{TRIM}(g_i) - \mathrm{TRIM}(g_{PM}))^2}{\sum W_g} \quad (3\text{-}1)$$

  where g is a gene in the intersection of the input set and a PM, g_i and g_{PM} are the expression measurements for gene g in the input set and PM correspondingly, TRIM is a function based on TRIM_CUTOFF, a user defined parameter intended to mitigate the effect of unusually high or low measurements in the input set or PM (i.e., TRIM(g) = min(g, TRIM_CUTOFF) if g is positive and max(g, TRIM_CUTOFF) if g is negative, TRIM_CUTOFF is set to 3 by default), and W is a scaling function based on two user defined weight parameters: W_MAGNITUDE and W_SIGN_PENALTY. W_MAGNITUDE is intended to give a higher importance, and thus better match, for the extreme measurements in the input set, and W_SIGN_PENALTY is intended to penalize cases where the measurements for gene g in the input set and a PM differ in the sign (i.e., g is up-regulated in the input set and down-regulated in the PM and vice versa). The distance over all genes is divided by a sum over all weights $W_{g,}$, calculated for each gene g.

- Correlation Distance:

$$Correlation(input, PM) = 1 - \mathrm{Pearson}(input \cap PM) \quad (3\text{-}2)$$

where Pearson is the Pearson correlation function calculated over the set of genes in the intersection between the input set and a PM.

- Anti-correlation Distance:

$$Anti - correlation(input, PM) = 1 + \text{Pearson}(input \cap PM) \quad (3\text{-}3)$$

where Pearson is the Pearson correlation function calculated over the set of genes in the intersection between the input set and a PM.

To facilitate cross species queries, each gene identifier is matched to the corresponding orthology group in the output species using Inparanoid [17]. As multiple orthologs can be mapped for a single input gene, for each PM the orthologous gene with the minimal Euclidian distance measurement to the input gene is selected and used for the comparisons.

In many cases several PMs from the same GSE exhibit similar expression patterns. Therefore, in order to show a more heterogeneous output set, only the best PM from each GSE is returned.

In order to assess the quality of a PM match we produce 10,000 random gene vectors of the same size drawn from a N(0,1) distribution for each PM, and calculate the distance of each random vector to the input query using the chosen distance function. To determine a *p*-value for the match we calculate how many random vectors obtain a distance smaller than the one obtained for the evaluated PM.

3.2.4 Output matches set analysis

The output set of matched expression experiments are expected to share biological functions. In order to identify common biological processes we search for keywords that are enriched among the output set. The keywords are selected from a pre-defined compendium of words assembled from GO terms definitions. We collected the abstract attached to each GSE, either directly from GEO or from the associated publication when a pubmedID is provided, and counted the number of appearances of each keyword in the output set. These counts are compared with the number of appearances of each keyword over all GSEs in the output species using a hypergeometric test. A higher weight for the final ranking of the results is given to keywords that are found in the GSM experiment titles.

In addition to analyzing the descriptions attached to each experiment, we can use the expression data directly to understand the biological processes associated with the output set of matches. The top differentially expressed genes for each PM in the database are pre analyzed for GO terms enrichment using a hypergeometric test, and enriched terms are saved for each GSE. Enriched GO terms associated with the set of output matches are listed and tested for significance using hypergeometric test.

Lastly, one of the aims of *ExpressionBlast* is to provide concrete hypotheses regarding specific genes. The abstracts were scanned for known gene names among the set of output matches. The output gene names are listed based on their number of appearances.

3.2.5 Experimental procedures

AML12 cells (normal mouse hepatocyte cell line) were cultured in a 1:1 (v/v) mixture of DMEM and Ham's F12 medium supplemented with 10% fetal calf serum, 5 µg/ml insulin, 5 µg/ml transferrin, 5 ng/ml selenium, 0.1 µM dexamethasone, 100 ng/ml streptomycin, and 100 U/ml penicillin.

To generate AML12 cells overexpressing SIRT6, the cells were transiently transfected with pcDNA-SIRT6 or with a control (GFP) vector using Metafectene transfection reagent (Biontex), according to the manufacturer's protocol. The cells were harvested 48 hours post transfection. For LXR activation, cells were treated for 48 h with either LXR agonist GW3965 (1 µM) or DMSO (vehicle control).

Total RNA was extracted using TRI-reagent (Sigma) according to the manufacturer’s protocol. cDNA was generated using First Strand cDNA Synthesis Kit (Fermentas). Quantitative real-time PCR was performed using a SYBR Green mix (Roche) in a StepOnePlus thermocycler (Applied Biosystems). Ct values were normalized to actin.

3.3 Results

3.3.1 Amounts of data collected and parsed by *ExpressionBlast*

The automatic parsing of microarray series in GEO introduced in the methods was able to create the largest collection of computationally annotated expression data currently available. Table 3-3 lists the number of series for which we were able to automatically identify replicates and treatment vs. control (TC) cases. The number of TC cases identified by our method vary between the species and we are currently working on improving the gene identifier matching methods for some of the species, which will enable us to increase the number of annotated expression studies.

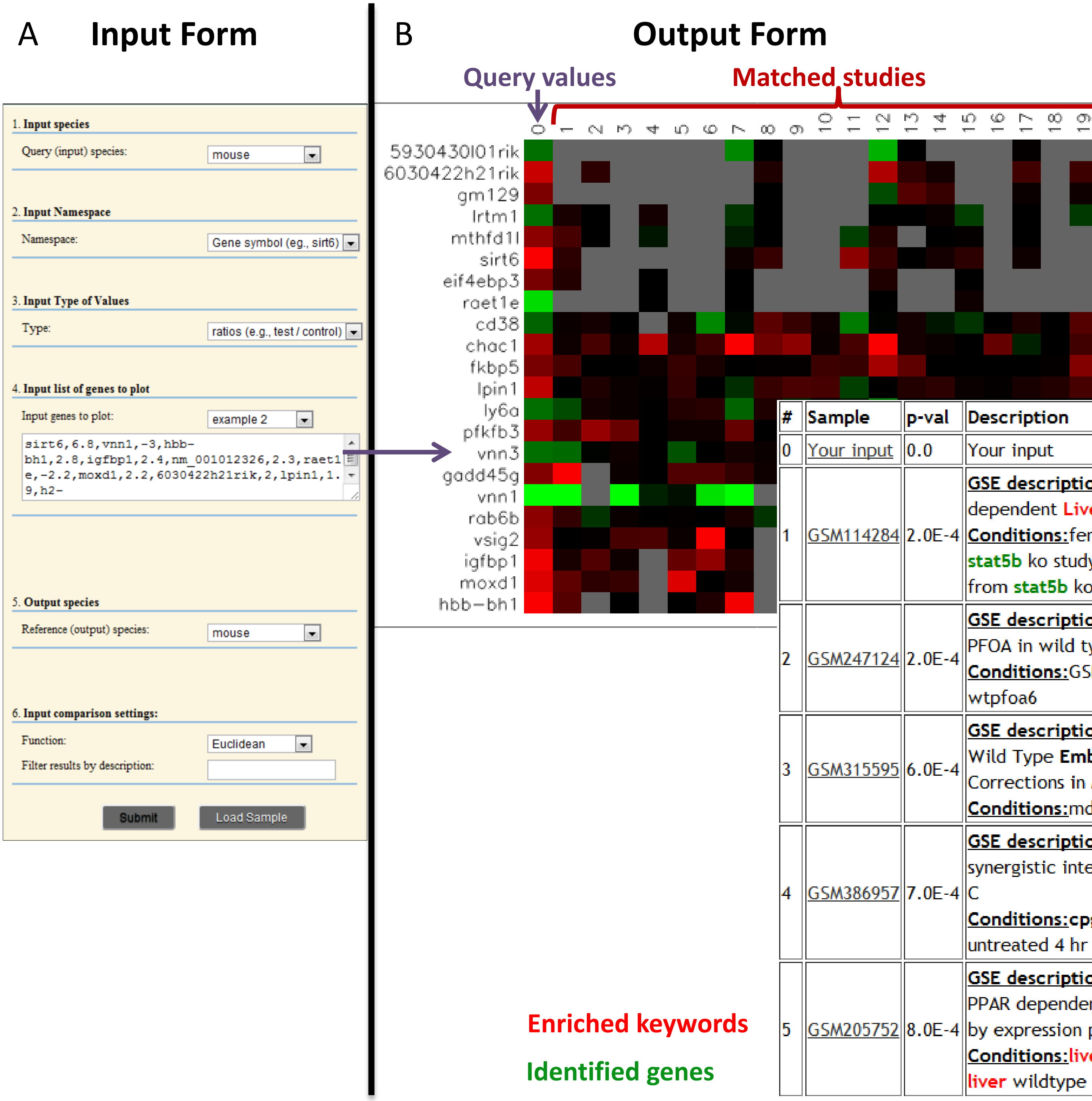
Table 3-3: Amounts of data collected and parsed by *ExpressionBlast*

Amounts of parsed expression data available on *ExpressionBlast* for key species as of April, 2012. Columns definition: species – scientific name; Tax ID – official taxonomy ID of a species; Expression series – the number of microarray expression series for that species in GEO (notice these numbers differ from Figure 1-4 that lists all the studies available for a certain species); Replicates identified – the number of series for which replicates were identified and merged; Treatment / Control (TC) identified – the number of series for which treatment and controls were identified and mapped; Treatment / Control percent – the percent of the studies with identified TC compared to the total number of expression series.

Species	Tax ID	Expression series	Replicates identified	Treatment / Control identified	Treatment / Control percent
Homo sapiens	9606	6302	6301	3165	50%
Mus musculus	10090	4383	4363	2578	59%
Arabidopsis thaliana	3702	1418	917	520	37%
Drosophila melanogaster	7227	1290	505	241	19%
Saccharomyces cerevisiae	4932	939	804	459	49%
Rattus norvegius	10116	821	821	402	49%
Caenorhabditis elegans	6239	541	285	122	23%
Escherichia coli	562	237	225	107	45%
Danio rerio	7955	222	190	104	47%
Schizosaccharomyces pombe	4896	215	174	120	56%
Oryza sativa	4530	211	155	76	36%
Bos taurus	9913	189	170	82	43%
Zea mays	4577	132	114	46	35%
Macaca mulatta	9544	63	54	21	33%
Total		**17056**	**15139**	**8075**	**47%**

3.3.2 Web tool for performing comparisons

Please refer to Figure 3-2 for a sample screenshot for how the search engine can be used. Chapter 6 provide details and screen shot on the various capabilities offered by the *ExpressionBlast* framework.



1. **Input species**

Query (input) species: mouse

2. **Input Namespace**

Namespace: Gene symbol (eg., sirt6)

3. **Input Type of Values**

Type: ratios (e.g., test / control)

4. **Input list of genes to plot**

Input genes to plot: example 2

```
sirt6,6.8,vnn1,-3,hbb-
bh1,2.8,igfbp1,2.4,nm_001012326,2.3,raet1
e,-2.2,moxd1,2.2,6030422h21rik,2,lpin1,1.
9,h2-
```

5. **Output species**

Reference (output) species: mouse

6. **Input comparison settings:**

Function: Euclidean

Filter results by description:

#	Sample	p-val	Description
0	Your input	0.0	Your input
1	GSM114284	2.0E-4	**GSE descriptio** dependent **Live** **Conditions:**fer **stat5b** ko study from **stat5b** ko
2	GSM247124	2.0E-4	**GSE descriptio** PFOA in wild ty **Conditions:**GS wtpfoa6
3	GSM315595	6.0E-4	**GSE descriptio** Wild Type **Emb** Corrections in **Conditions:**md
4	GSM386957	7.0E-4	**GSE descriptio** synergistic inte C **Conditions:cp** untreated 4 hr
5	GSM205752	8.0E-4	**GSE descriptio** PPAR depender by expression **Conditions:live** **liver** wildtype

Figure 3-2: ***ExpressionBlast*** **input and output forms**

(a) The input panel (shown on left) contains the following mandatory inputs required for the query: 1. Name of the query (input) species. 2. Namespace or nomenclature schemes of the input genes. 3. Type of the input expression values (e.g., treatment/control ratio or the raw unclassified values). 4. Query set gene names and their corresponding values. 5. Name of the reference (output) species - the species for which database matches will be returned. 6. The distance function heuristic to be used for the query (See Search engine matching process for details). **(b)** The matched expression experiments set is shown in a heatmap format with the query genes as rows and the matched expression experiments as columns where the first column depicts the user input values. The information on the columns is available in the Matches or Abstracts tabs. Keywords that are present in the GO based keyword compendium are shown in bold. Enriched keywords are highlighted in red, and gene names are highlighted in green. See full input and output form listing in the supplementary information.

3.3.3 Application to Mice Aging Data

We applied *ExpressionBlast* is used to study aging data from SIRT6 (sirtuin 6) transgenic mice. SIRT6 deficiency in mice results in premature aging phenotypes and metabolic defects and was implicated in a calorie restriction response. Kanfi *et al.* [108] explored SIRT6 role in metabolic stress by feeding wild type and transgenic (TG) mice over-expressing SIRT6 with high fat diet. The SIRT6 TG mice accumulated significantly less visceral fat, LDL-cholestrol, and triglycerides compared to the WT mice. Expression analysis showed reduced expression of selected peroxisome proliferator-activated receptor PPARγ regulated genes affecting lipid homestatsis. These results showed that SIRT6 over-expression has protective roles in disorders induced by high-fat diet. In a follow-up study we examined the effect of SIRT6 over-expression on mice lifespan [61]. We found out that male TG mice, but not female TG mice, were shown to have a significantly longer lifespan. Two lines of transgenic mice showed increased lifespan by 14.8% and 16.9% respectively on average. To understand better the mechanisms of the gender-specific lifespan extension in SIRT6-TG mice, we performed a microarray analysis to examine differential expression in the livers of animals of both sexes. Differential expression analysis using SAM [109] showed that the most extensive gene expression differences occurred between the genders. ANOVA analysis uncovered a subset of genes whose expression differed significantly between genotypes and that were gender-specific. Comparing this differentially expressed gene set with the set of genes that was differentially expressed between male and female WT mice revealed that 50% (41 of 82) of the genes that were differentially expressed in the SIRT6-TG males were also differentially expressed between male and female WT mice (*P*=0) (see Figure 3-3). These results were confirmed for 11 differentially expressed genes using quantitative PCR. GO analysis over the differentially expressed genes found categories significantly enriched for categories related to metabolism and cellular responses. A key factor in the regulation of lifespan is the IGF1 signaling pathway and high levels of IGFBP1, one of the most differentially expressed genes in our study, is correlated with protection against metabolic disorders [110] and mice with fat-specific insulin receptor gene knockout have been shown to have increased mean lifespan of similar magnitude to the transgenic mice in our study [111]. Nonetheless, most genetic modifications of the IGF1 or insulin signaling pathway affect the lifespan of both genders or show stronger effect in females.

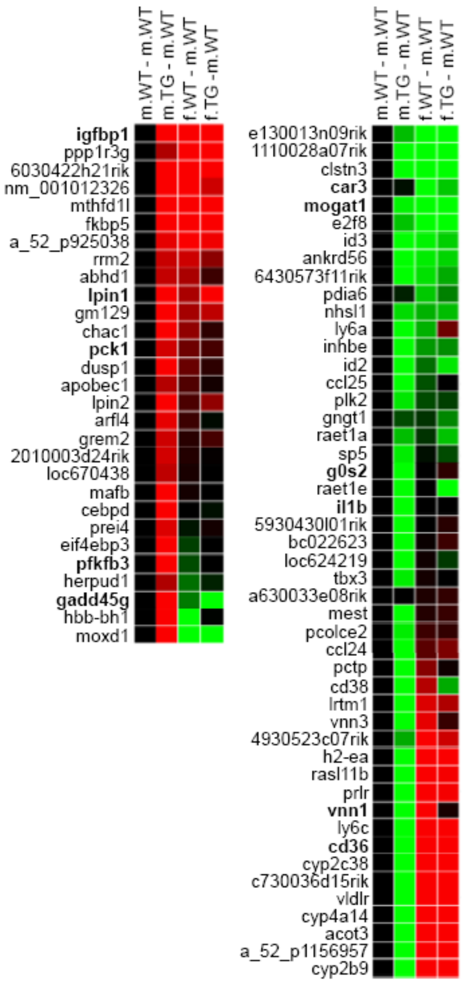


Figure 3-3: Expression profile of differentially expressed genes in male Sirt6-transgenic mice
Heat maps displaying the significantly upregulated (red) and downregulated (green) genes in Sirt6-transgenic males (m.TG) compared with WT males (m.WT). The expression profile of these genes in WT females (f.WT) or Sirt6-transgenic females (f.TG) compared with WT males is also illustrated. Statistical analysis was performed using all 24 arrays. Marked in bold are genes that were validated using quantitative PCR.

In order to further investigate the mechanisms of the genes regulated by SIRT6 and understand whether the effects of SIRT6 are blocked in females rather than enhanced in males we queried the 22 top differentially expressed genes in *ExpressionBlast* against mouse and other species. The *ExpressionBlast* analysis results show the returned set of 20 matches to be enriched for the keywords 'liver' (7/20), 'hepatic' (3/20), and 'lipid metabolism' (3/20) which corresponds well with the query expression experiment. Specifically, among the top most similar studies identified by *ExpressionBlast*, three studies involved PPARα (peroxisome proliferator-activated receptor alpha) knockout mice [112–114] and one study involved LXRα [115]. PPARα is a ligand activated transcription factor involved in the regulation of nutrient metabolism and is known to be negatively regulated by SIRT1, a different homolog of the sirtuin family. LXRα (liver X receptor alpha) is a nuclear receptor that controls transcriptional programs involved in lipid homeostasis and was shown to be associated with lifespan [116]. LXR is also known to play roles in Glucocorticoid, Estrogen, and Androgen homeostasis [117]. Estrogens are the primary female sex hormones and as discussed below may explain some of the lifespan extension differences observed between male mice and female mice. The LXR study that is matched to the SIRT6 experiment showed a perfectly positive correlation over the 22-gene signature with fold change below 0.5 or above 2.0 (see Figure 3-4). This study examines the role of hepatic LXRα on diet-dependent cardiovascular lipid metabolism and shows that LXRα can be selectively modulated to control specific pathways including cholesterol metabolism [115].

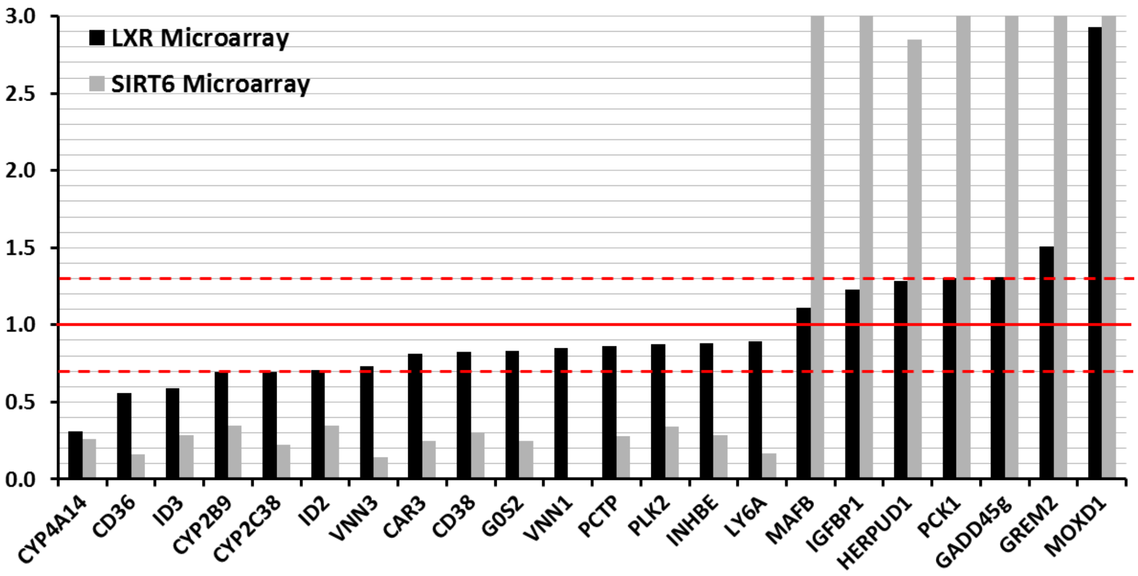
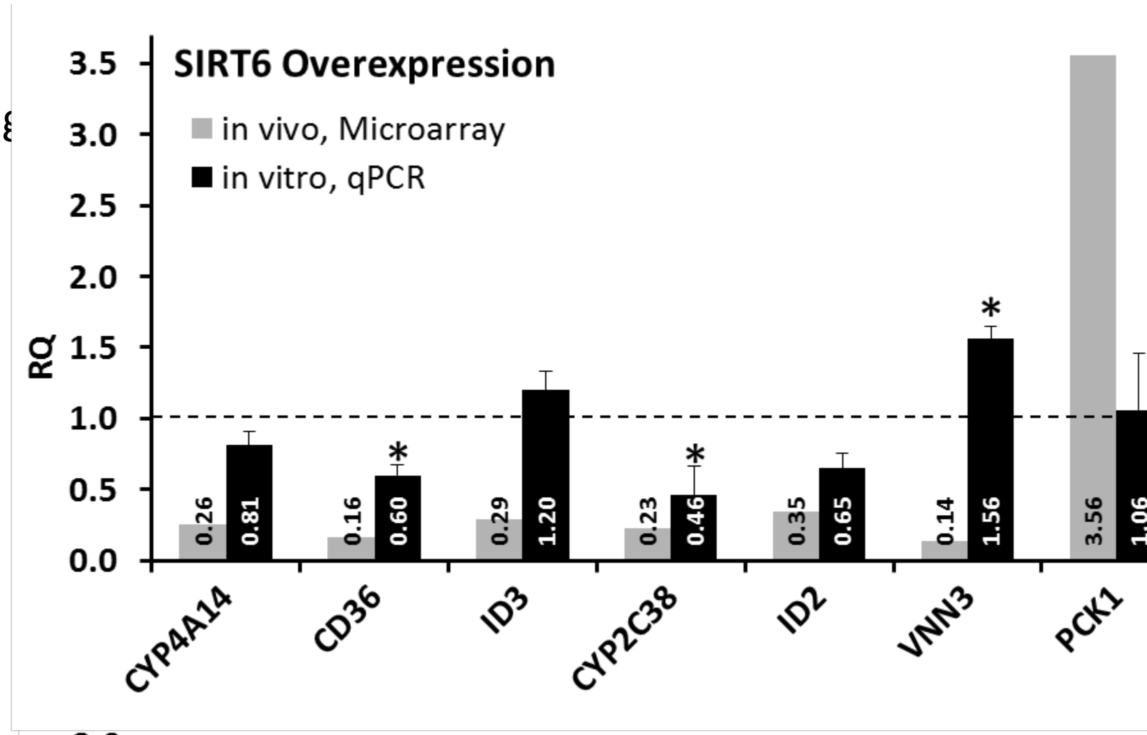


Figure 3-4: Perfect correlation between SIRT6 microarray result and LXR microarray study

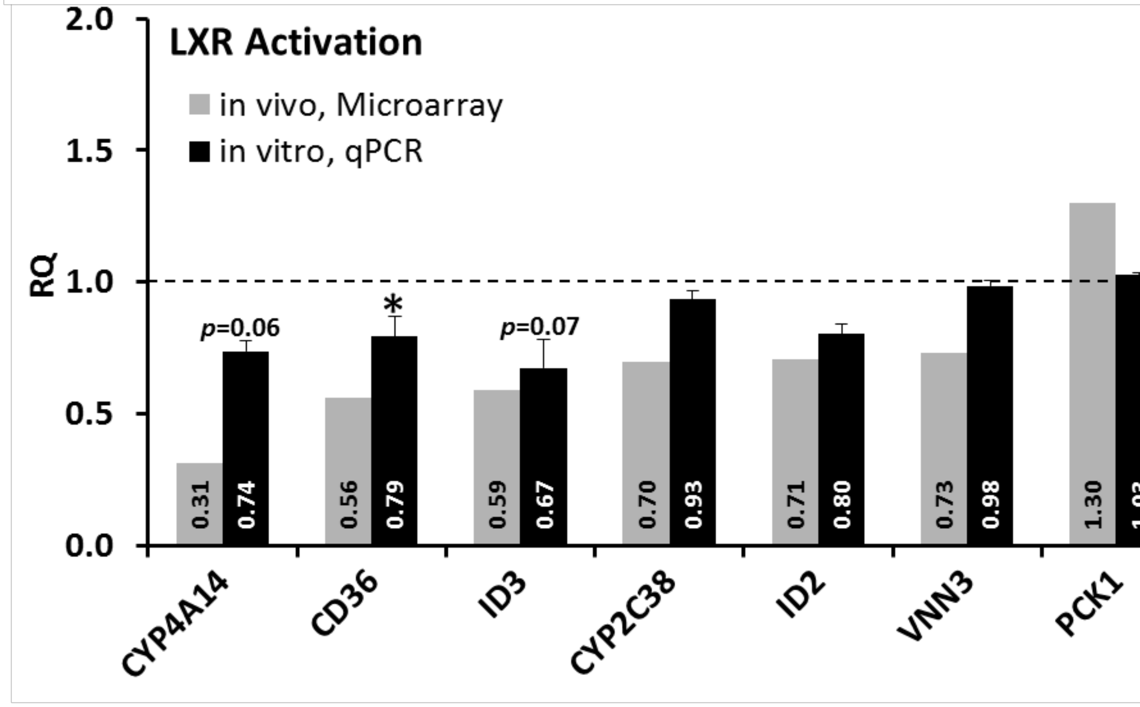
Comparing a 22-gene expression signature of genes with FC above 2 or below 0.5 found a perfect correlation with LXR Study from GEO (GSM50805). Light grey bars correspond to the SIRT6 study results, while the dark bars correspond to the LXR study. Ratio values are shown; values below 1.0 are down-regulated and values above 1.0 are up-regulated.

These *ExpressionBlast* results pose LXRα and PPARα as possible candidates to explain the mechanisms by which SIRT6 operate. We examined the similarity in expression patterns *in-vitro* systems using AML12 hepatocyte cell line to test whether they mimic the expression results obtained for the in-vivo whole animal microarray results for both SIRT6 over-expressed cells and cells treated with the LXR agonist GW965. The qPCR experiments were performed *in-vitro* for eight of the differentially expressed genes identified by the *in-vivo* microarray results. These results showed high correlation between the activation pattern of the over-expressed SIRT6 and the LXRα agonist (see Figure 3-5). CD36 was found to be significantly different both in the SIRT-over expressed cells and in the LXR activated cells. CD36 is a known target gene of both LXRα [118] and PPARα [119], and was shown to have a physiological function in oxidizing LDL [120]. This supports earlier evidences that shows that both SIRT6 over-expression can lower LDL-cholesterol levels [108], results similar to a study that showed that LXRα activation increased LDL-receptor levels, hence mediated LDL efflux, lowering LDL cholesterol levels [121]. A recent review on LXR control of cholesterol homeostasis is available here [122] .

AML12 cells
Overexpressing
SIRT6



AML12 cells
treated with
GW3965



AML12 cells = normal mouse hepatocyte cell line

Figure 3-5: Comparison between SIRT6 over-expressed cells and LXR activated cells

Eight genes showing differential expression in a SIRT6 over-expression in the *in-vivo* whole animal microarray experiments are selected for in-vitro qPCR experiments on hepatocyte AML12 cell line. The qPCR experiments are contrasting over-expressed SIRT6 cells with cells treated with LXR agonist GW3965. Ratio values are shown; values below 1.0 are down-regulated and values above 1.0 are up-regulated.

Another independent study showed that SIRT1, the most studied member of the sirtuin family, deacytelates and positively regulates LXR [107]. SIRT1, SIRT6, and SIRT7 are considered the most similar proteins of the sirtuin family functioning as nuclear proteins and are known to be enriched in the nucleoplasm, the heterochromatin, and the nucleoli. Taken together, these results show the possibility of regulatory mechanism of operation for SIRT6 by modulating LXRα activity through affecting the formation of the LXR-co-activator complex or the formation of the LXR-co-repressor complex possibly through deacetylation (see Figure 3-6). This hypothesis is currently being followed-up experimentally.

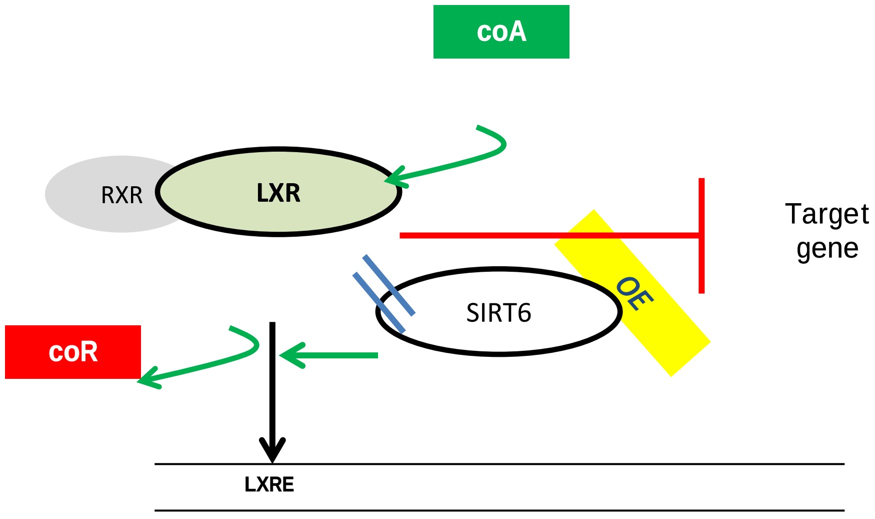


Figure 3-6: A possible mechanism for SIRT6 modulation of LXR

Sirt6 may modulate LXR activity by affecting the formation of the LXR-coA or LXR-coR complexes possibly through deacetylation.

3.3.4 *ExrpressionBlast* Cross-species comparison

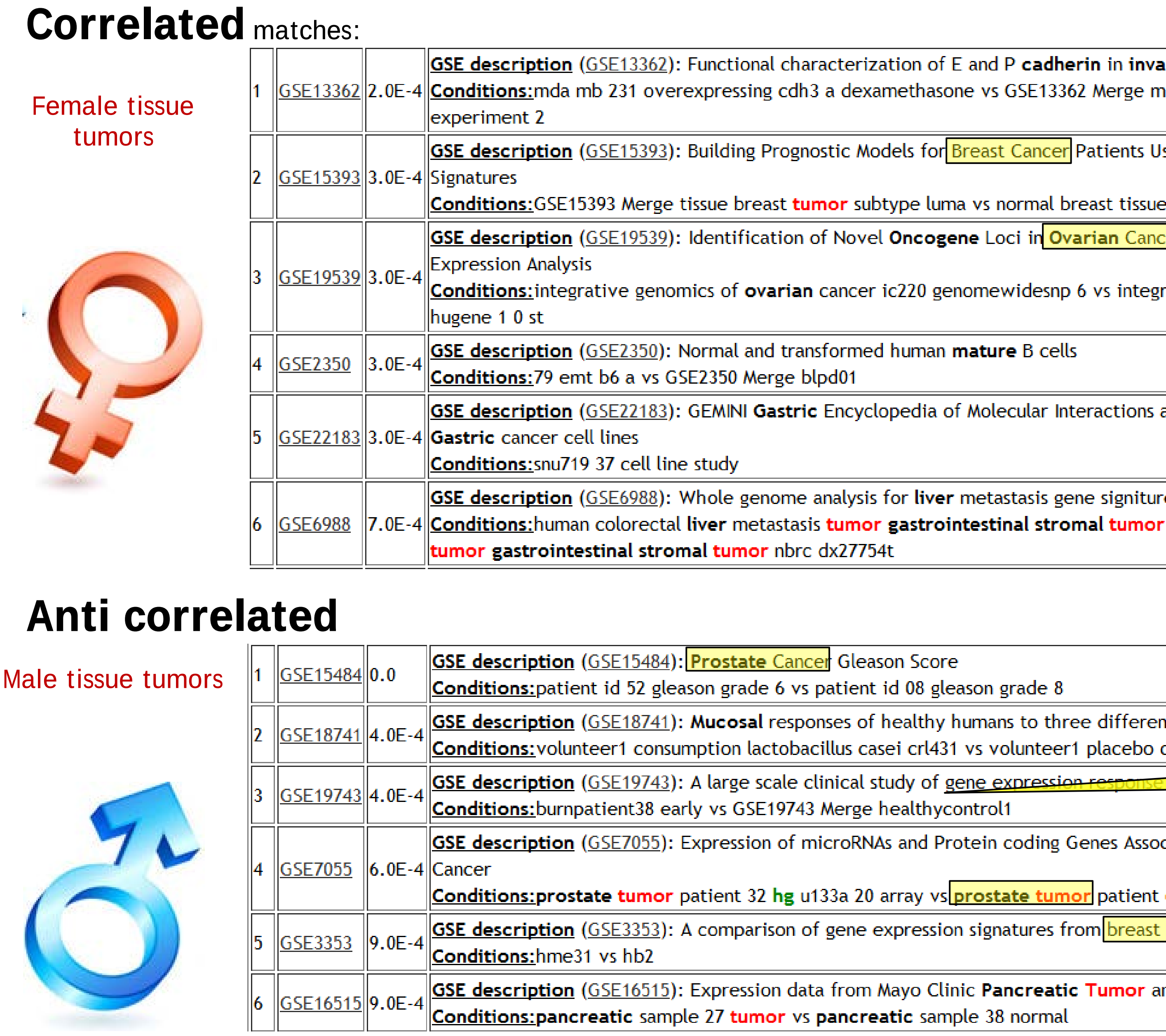
Performing queries across species can help understand the broader story to gain new aspects of the investigated mechanism and facilitate quick knowledge by looking at experiments that were conducted in other species. Finding experiments in other species that show similar expression signature helps in two directions: 1) it can increase the confidence in the experimental results of the expression experiment that was performed and 2) it may lead to insights on conserved evolutionary mechanisms and pinpoint to key follow-ups to be performed. Cross species queries may also reduce drug development time. Most drugs are first tested on lower mammals before being applied to human. Gaining insights for whether the desired phenotypic outcome was observed at human early on in the drug discovery process may increase the confidence in the research direction.

Rat

Comparing the mice expression data to rat studies in the compendium fell into the first category of information that can be gained from a cross species query, and increased the confidence in our experimental results. We found many rat studies that were performed in liver cells. These studies were enriched with several relevant keywords including 'weight' and 'xenobiotic metabolism'. Specifically, among the top studies in rat that were found to have a similar expression signature to the over-expressed SIRT6 mouse expression signature were a study that analyzed xenobiotic metabolizing enzyme gene expression in aging male rats [123], and a study that measures the hypothalamic responses to reduced food intake [124] strengthening the role of Sirt6 in modulating metabolisms across species.

Human

Comparing the mice expression data to human studies in the compendium fell into the second category of information that can be gained from a cross species query, and was able to suggest new research directions. We performed two types of queries, one to find human studies that are correlated with our mouse results and the other for human studies that are anti-correlated with our mouse results. We found that the top matches among the correlated set of human studies involved female tissues tumors including breast cancer [125] [126] and ovarian cancer [127] and some of the top matches in the anti-correlated set of human studies involved male tissues tumors including prostate cancer [128] [129] (see Figure 3-7).



1	GSE13362	2.0E-4	**GSE description** (GSE13362): Functional characterization of E and P **cadherin** in **inva** **Conditions:**mda mb 231 overexpressing cdh3 a dexamethasone vs GSE13362 Merge m experiment 2
2	GSE15393	3.0E-4	**GSE description** (GSE15393): Building Prognostic Models for Breast Cancer Patients Us Signatures **Conditions:**GSE15393 Merge tissue breast **tumor** subtype luma vs normal breast tissue
3	GSE19539	3.0E-4	**GSE description** (GSE19539): Identification of Novel **Oncogene** Loci in **Ovarian Canc** Expression Analysis **Conditions:**integrative genomics of **ovarian** cancer ic220 genomewidesnp 6 vs integr hugene 1 0 st
4	GSE2350	3.0E-4	**GSE description** (GSE2350): Normal and transformed human **mature** B cells **Conditions:**79 emt b6 a vs GSE2350 Merge blpd01
5	GSE22183	3.0E-4	**GSE description** (GSE22183): GEMINI **Gastric** Encyclopedia of Molecular Interactions a **Gastric** cancer cell lines **Conditions:**snu719 37 cell line study
6	GSE6988	7.0E-4	**GSE description** (GSE6988): Whole genome analysis for **liver** metastasis gene signitur **Conditions:**human colorectal **liver** metastasis **tumor gastrointestinal stromal tumor tumor gastrointestinal stromal tumor** nbrc dx27754t

1	GSE15484	0.0	**GSE description** (GSE15484): Prostate Cancer Gleason Score **Conditions:**patient id 52 gleason grade 6 vs patient id 08 gleason grade 8
2	GSE18741	4.0E-4	**GSE description** (GSE18741): **Mucosal** responses of healthy humans to three differen **Conditions:**volunteer1 consumption lactobacillus casei crl431 vs volunteer1 placebo c
3	GSE19743	4.0E-4	**GSE description** (GSE19743): A large scale clinical study of gene expression response **Conditions:**burnpatient38 early vs GSE19743 Merge healthycontrol1
4	GSE7055	6.0E-4	**GSE description** (GSE7055): Expression of microRNAs and Protein coding Genes Assoc Cancer **Conditions:prostate tumor** patient 32 **hg** u133a 20 array vs **prostate tumor** patient
5	GSE3353	9.0E-4	**GSE description** (GSE3353): A comparison of gene expression signatures from breast **Conditions:**hme31 vs hb2
6	GSE16515	9.0E-4	**GSE description** (GSE16515): Expression data from Mayo Clinic **Pancreatic Tumor** ar **Conditions:pancreatic** sample 27 **tumor** vs **pancreatic** sample 38 normal

Figure 3-7: Correlated and anti-correlated human matches

The mouse expression results were queried again human studies using *ExpressionBlast* to find correlated and anti-correlated studies. The set of studies that were correlated with the mouse values was enriched with tumors related to female tissue tumors including breast cancer and ovarian cancer. The set of studies that was anti-correlated with the mouse values was enriched with male tissue tumors including prostate cancer. This may explain some of the feminization effect seen in the original queried mouse study, possibly through changes in the estrogen levels.

The matched breast cancer study by Sarrio *et al.* tried to characterize E and P-caderin in invasive breast cancer cells [125]. Among the genes that are the most highly differentially expressed in this study are SERPINA1 which is known to be correlated with SIRT1 and is known to be highly expressed in the liver [130] and HERPUD1 a gene that was found to be highly differentially expressed also in the mouse LXR study. P-cadherin positive expression in breast cancer carcinomas is associated with unfavorable prognostic factors including lack of estrogen receptor [131]. The matched breast cancer study by Fan *et al.* also showed a negative regulation between estrogen receptor (ER) and survival rate of breast cancer patients [126]. Among the highly amplified genes found in the ovarian cancer study the cancer driver HER2/ERBB2 was found [127], a pathway that has long been implicated in breast cancer aetiology and was found to be directly regulated by o-estrogen receptor (ER) [132].

The most anti-correlated human study to our mouse expression results was a study showing that differential expression of apoptotic genes PDIA3 and MAP3K5 distinguishes between low and high risk prostate cancer [128]. The MAPK pathway was previously shown to be activated by testosterone [133], high levels of which (and low levels of estradiol/estrogen) were shown to be correlated with prostate cancer [134].

The distinction seen in the correlated and anti-correlated human studies showing enrichment for female tissue tumor and male tissue tumors respectively may possibly be a result of common factors that drove the observed life extension phenomena only in male mice in our queried study. Estrogen, the female sex hormone, is an obvious candidate to explain the observations in both species. About 80% of breast cancers, once established, continue to grow as long as enough estrogen is available and estrogen deprivation is one of the treatments used for breast tumors. The human matched breast cancer studies found using *ExpressionBlast* clearly showed decreased estrogen receptor levels in breast cancer patients. Estrogen was also shown to down-regulate the levels of testosterone [135], and under certain conditions may be used for treatment of prostate cancer [136]. Estrogen is also known to play an active role in aging processes having beneficial effect on healthy life and lifespan and helping to prevent age-related conditions. For example, estrogen was shown to delay memory loss, regulated liver production of cholesterol thus decreasing atherosclerosis, and preserve bone density. After menopause, estrogen levels drop suddenly and it is common today for woman to take estrogen supplements to control the symptoms of menopause and reduce aging symptoms. In addition, a recent study on men with chronic heart failure showed that men in the lowest quantile of estradiol (the primary type of estrogen) were 317% more likely to die during a 3-year follow up [137].

Estrogen is also tightly linked to the PPARα and LXRα pathways found in the mouse vs. mouse comparisons. He *et al.* [117] reviewed how LXRα controls estrogen homeostasis. In particular, Gong *et al.* [138] showed that orphan LXRα activation promote estrogen deprivation and inhibition of breast cancer growth *in-vivo* by regulating basal and inducible hepatic expression of EST/SULT1 [139]. LXRα [140] was also shown to regulate androgens including testosterone through binding and activating the androgen receptor promoting benign hyperplasia and prostate cancer. Androgen deprivation is a key treatment of hormone-dependent prostate cancer and this process can be mediated through SULT-mediated sulfonation including SULT2A1 that was reported to be an LXRα target. LXR was also shown [141] to play role in inhibiting proliferation of human breast cancer cells. Estrogen-dependent gallbladder carcinogenesis was also reported [142] in LXRβ -/- female mice.

The other direction is also present. Estrogen plays an important role in many of processes regulated by LXRα; Estrogen was shown to affect LDL and vLDL metabolism [143], reduce LDL accumulation [144], protect against LDL cholesterol oxidation [145] [146], and decrease atherosclerosis in monkeys [147].

3.4 Discussion

We presented a novel tool, *ExpressionBlast*, for comparing expression results against a large compendium of expression experiments. BLAST is commonly used for comparing sequence information within and across species, and we aim that *ExpressionBlast* brings this ability to dynamic, condition-specific expression data. We expect that comparing expression data will become a routine in exploratory expression analysis, in addition to other basic analyses including clustering and GO enrichment. In addition, the ability to perform expression queries across species is important for basic cell research as well as drug discovery, since most drugs are developed and first tested on lower mammals before being approved for human experiments.

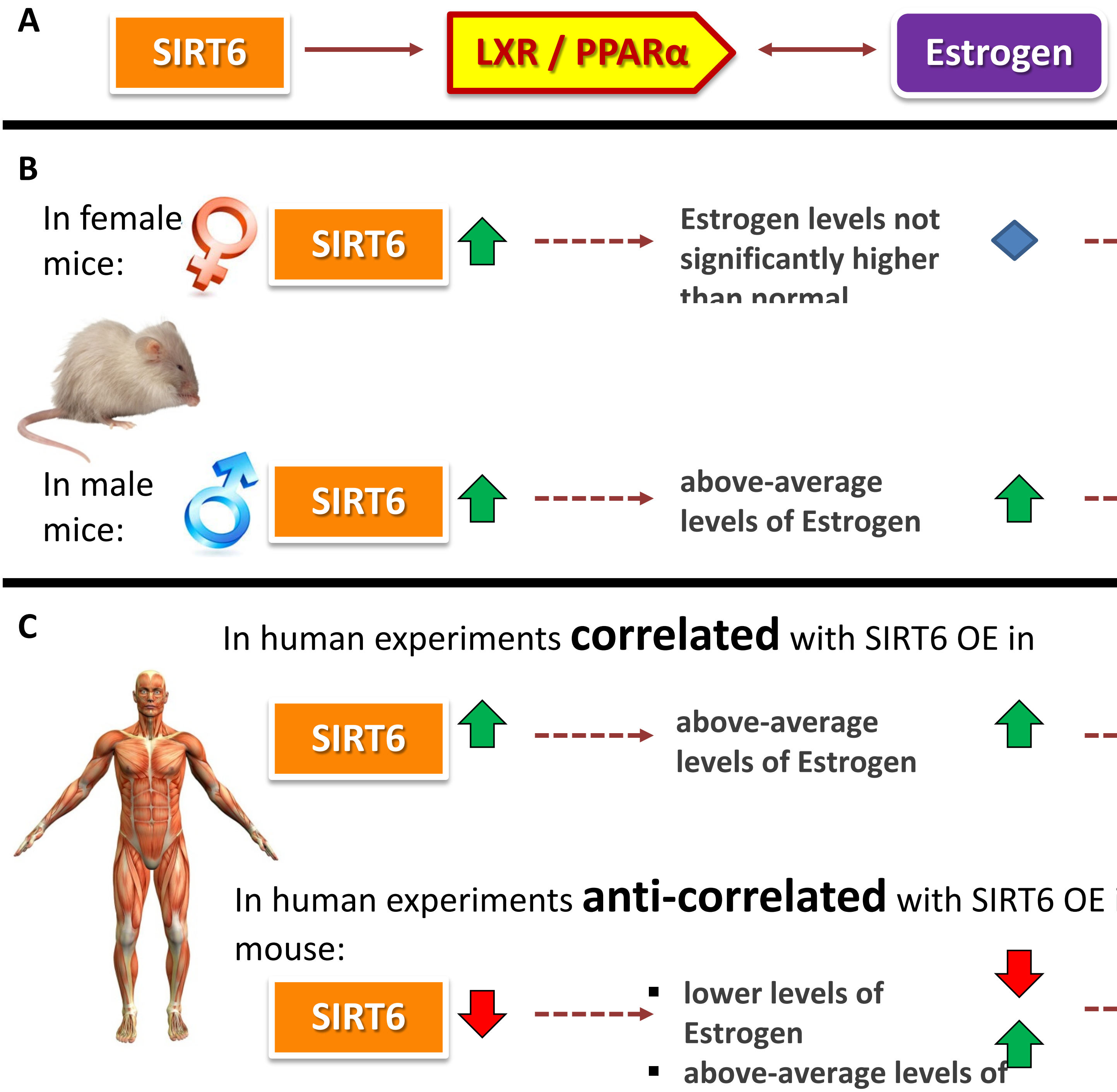
3.4.1 Use of studies meta-data to improve significance analysis of highlighted terms

A great deal of knowledge can be gained from the meta data attached to each study in order to improve the highlighted terms for the set of matched studies. For example, one problem that may arise is that in many cases academic groups publish several studies in one area, examining several similar variants of the same condition and using similar lab protocols that may result in similar measurements. This can cause a bias in the highlighted keywords towards this area of research. One possible option to mitigate this bias is to find heuristics to identify studies that are published by the same group and reduce the score given to these studies. Four simple solutions to identify these studies include (1) identifying studies that were uploaded to the repository by the same contact name, (2) share contact's university and department, (3) have a high overlap among contributors of different studies, or (4) share the last contributor (presumably the principal investigator). It may also be informative to highlight these cases to users, hence pointing out to key researchers and departments conducting a large body of research in this area.

3.4.2 SIRT6 may regulate estrogen levels through PPAR and LXR

ExpressionBlast identified PPARα and LXRα as possible mechanisms for SIRT6 regulation on downstream genes in the mouse vs. mouse queries, showing similar effects on LDL-cholesterol levels. The queries comparing our mouse results to human studies revealed female tissue tumors enriched among the correlated studies and male tissue tumors enriched among the anti-correlated studies. A large body of studies demonstrates an interplay effect between LXRα, PPARα and the female sex hormone estrogen on breast cancer and other aging symptoms including cholesterol homeostasis. A possible effect of SIRT6 on estrogen levels through LXRα and PPARα (Figure

3-8a) may explain the feminization effect seen for the over-expressed SIRT mice and the results that were obtained in *ExpressionBlast* for the cross-species queries with human.



A

SIRT6 → LXR / PPARα ↔ Estrogen

B

In female mice: SIRT6 ⬆ - - - - → Estrogen levels not significantly higher than normal ◆ - -

In male mice: SIRT6 ⬆ - - - - → above-average levels of Estrogen ⬆ - -

C

In human experiments **correlated** with SIRT6 OE in

SIRT6 ⬆ - - - - → above-average levels of Estrogen ⬆ - -

In human experiments **anti-correlated** with SIRT6 OE i mouse:

SIRT6 ⬇ - - - - →

- lower levels of Estrogen ⬇
- above-average levels of ⬆ - -

Figure 3-8: Simplified hypothesis on SIRT6 effect on life span and cancer through Estrogen
(a) Possible hypothesis of how SIRT6 may affect estrogen levels which can explain the feminization effect observed in mice **(b)** where only male mice had a significantly longer lifespan while female mice did not. **(c)** The effect of SIRT6 on downstream genes may be similar to the way higher or lower levels of estrogen affect them to drive female and male tissue tumors respectively.

Our hypothesis suggests that estrogen levels may be correlated with SIRT6 levels in mice. Female mice with boosted levels of SIRT6 may have resulted in only a minor increase on the estrogen levels and no significant effect on the lifespan. On the other hand, male mice with boosted levels of SIRT6 may had significant above-than-average levels of estrogen, a demonstrated indicator for aging, showing a significant increase in lifespan (Figure 3-8b). The estrogen levels in SIRT6 over-expressed mice are currently being experimentally tested. This hypothesis may explain also the *ExpressionBlast* results obtained for the comparison with the human studies. The effect of SIRT6 over-expression on downstream genes may be similar to their levels with higher-than-average levels of estrogen which may drive female tissue tumors enriched in the correlated studies. The anti-correlated studies which correspond to a reverse effect of SIRT6 on downstream genes may be similar to lower-than-average levels of estrogen, hence higher-than-average levels of testosterone which may drive male tissue tumors (Figure 3-8c). This hypothesis is only one example of how *ExpressionBlast* can help formulate new hypotheses using cross-species queries.

3.4.3 How to select genes for a query

One question that may arise when performing queries with *ExpressionBlast* is which genes are the best to use for the query. We suggest here three options for selecting genes for a query that are integrated into the *ExpressionBlast* system.

1. Top differentially expressed (DE) genes. The DE genes can be viewed as the signature of the expression experiments capturing the genes that are affected the most by the specific condition that is measured. The *ExpressionBlast* framework supports uploading full expression datasets, identifying differentially expressed genes, and automatically storing the top differentially expressed genes for future *ExpressionBlast* query analysis.

2. Gene expression clustering. Clusters with coherent expression pattern can indicate on specific biological processes that are activated in a similar manner, which are relevant to the experiment. One possibility is to use the clustering method presented in Chapter 4.
3. Active sub-networks / modules identification. Active sub-networks are connected groups of genes (usually protein-protein interactions) that show higher expression activation over the entire group. One possibility is to use the method presented in Chapter 5. See Chapter 6 for more details on our two-step paradigm for experimental expression results validation, using both tools.

3.4.4 Support for RNA-sequencing experiments

RNA-sequencing data is growing rapidly but pose many challenges in data processing and analysis compared to the more traditional microarray technology. Please refer to the discussion in Section 7.2.4 for further details.

4 *Soft Clust* - Clustering Expression Data Across Species

In the previous chapter we presented a method for comparing expression data across species and finding studies with similar expression signature. In this chapter we discuss how multiple studies from different species (or conditions) that were found to be related (for example through *ExpressionBlast*) can be analyzed together to identify core conserved or divergent processes. Clustering is one of the most popular analyses for expression data performed by experimental biologists. A unified and concurrent analysis for multiple species [12] has advantages over independent analyses for each species. However, available clustering methods are not suited to analyze expression data from multiple species concurrently and lead to situations in which orthologous genes with similar expression patterns could be misplaced into different clusters due to factors such as measurement error and ambiguity of cluster boundaries. We developed a constrained soft clustering method that we name *SoftClust* that can incorporate prior information and improve gene assignments to clusters when performing cross species analysis. In this chapter we show a specific application of our approach to understanding the mechanisms by which three evolutionary distant yeasts respond to the anti-fungal medicine fluconazole overtime, revealing significant divergence among regulatory programs associated with fluconazole sensitivity.

4.1 Introduction

Performing cross species clustering analysis is a powerful tool to identify core genes and processes that have similar dynamic behavior and may indicate on mechanisms that play important roles that remained protected in speciation events. Cases were similar dynamics are seen across species can be informative for understanding the mechanisms of operation in less studied species. However, cases were orthologs, that are expected to show similar dynamics, are actually different are equally interesting any may indicate on new roles and functions developed through evolution.

The key insight in clustering gene expressions is that cluster membership is often influenced by small changes that can be attributed to noise (especially when many clusters are used which is the appropriate thing to do when clustering thousands of genes). In the cross species realm, this may lead to cases were orthologs are assigned to different clusters even though their expression profiles are pretty similar. To overcome this we developed a soft constraint clustering framework we name *SoftClust*, which is based on k-means clustering [148–150], and that can incorporate various data sources, including sequence information, to influence the resulting clusters. The idea behind this algorithm is to reward assignments of constraints (e.g., orthologs, or genes regulated by the same TF) to the same cluster, which is controlled by user defined weight parameters that influence the significance of the constraint on the clustering. For clustering data from multiple species we use orthology relationships as a constraint.

The term "soft clustering" has also previously been used in other clustering methods to define cases in which a gene can belong to more than one cluster rather than any constraint used to identify clusters [151], [152]. For our method, "soft clustering" refers to the prior we use as a weight to encourage co-clustering of orthologous genes. In these cases, "soft" refers to the assignment of genes to clusters. Other methods which focus on the analysis of expression levels across species are limited to simultaneous analysis of two species or require assumptions regarding the distribution of expression data [50], [153], [154].

We applied *SoftClust* to understand the range of mechanisms by which yeasts can respond to anti-fungals. Fungal infections are an emerging health risk, especially those involving yeast that are resistant to antifungal agents. We compared gene expression patterns across three evolutionarily distant species - *Saccharomyces cerevisiae* (*Sc*), *Candida glabrata* (*Cg*), and *Kluyveromyces lactis* (*Kl*) - over time following fluconazole exposure. Mucosal and invasive mycoses are a major world health problem leading to morbidity [155], [156] and a mortality rate of up to 70% in immunocompromised hosts [157]. The most common treatment for fungal infections is the family of chemical compounds known as the azoles, which interfere with formation of the cell membrane by inhibiting synthesis of ergosterol [158]. However, the use of azoles to treat a broad spectrum of fungal infections has led to widespread azole resistance [158–163], and resistance is also emerging against the limited number of secondary compounds that are currently available[164], [165]. It is likely that the full range of anti-fungal resistance pathways is even greater, thus, an important goal moving forward is to better understand the entire pool of genotypic variation underlying fungal stress responses, particularly as they relate to antifungal agents.

Using *SoftClust* we were able to identify conserved and diverged expression patterns that suggested complementary strategies for coping with ergosterol depletion by azoles - *Saccharomyces* imports exogenous ergosterol, *Candida* exports fluconazole, while *Kluyveromyces* does neither, leading to extreme sensitivity. In support of this hypothesis we found that only *Saccharomyces* becomes more azole resistant in ergosterol-supplemented media; that this depends on sterol importers *AUS1* and *PDR11*; and that transgenic expression of sterol importers in *Kluyveromyces* alleviate its drug sensitivity.

4.2 Methods

4.2.1 Soft clustering algorithm

We developed a constrained clustering method based on the *k*-means algorithm, but using a revised objective function. Like regular *k*-means, the objective function considers the similarity of each gene's expression profile to the center of its assigned class. However, it also rewards class assignments in which orthologs are co-clustered. The reward (*W*) is a user-defined parameter that serves as a tradeoff between cluster expression coherence and percentage of co-clustered orthologs: each gene, $x \in X$, is assigned to cluster *h** such as to minimize the objective function:

$$h^* = \arg\min_{h} (\sum (D(x, C_h\ ) - W) \quad (4\text{-}1)$$

where $\sum(D(x,C_h)-W)$ refers to all possible partitions of genes in the same orthology group, *D()* refers to a user defined distance function, and C_h denotes the center of cluster *h*.

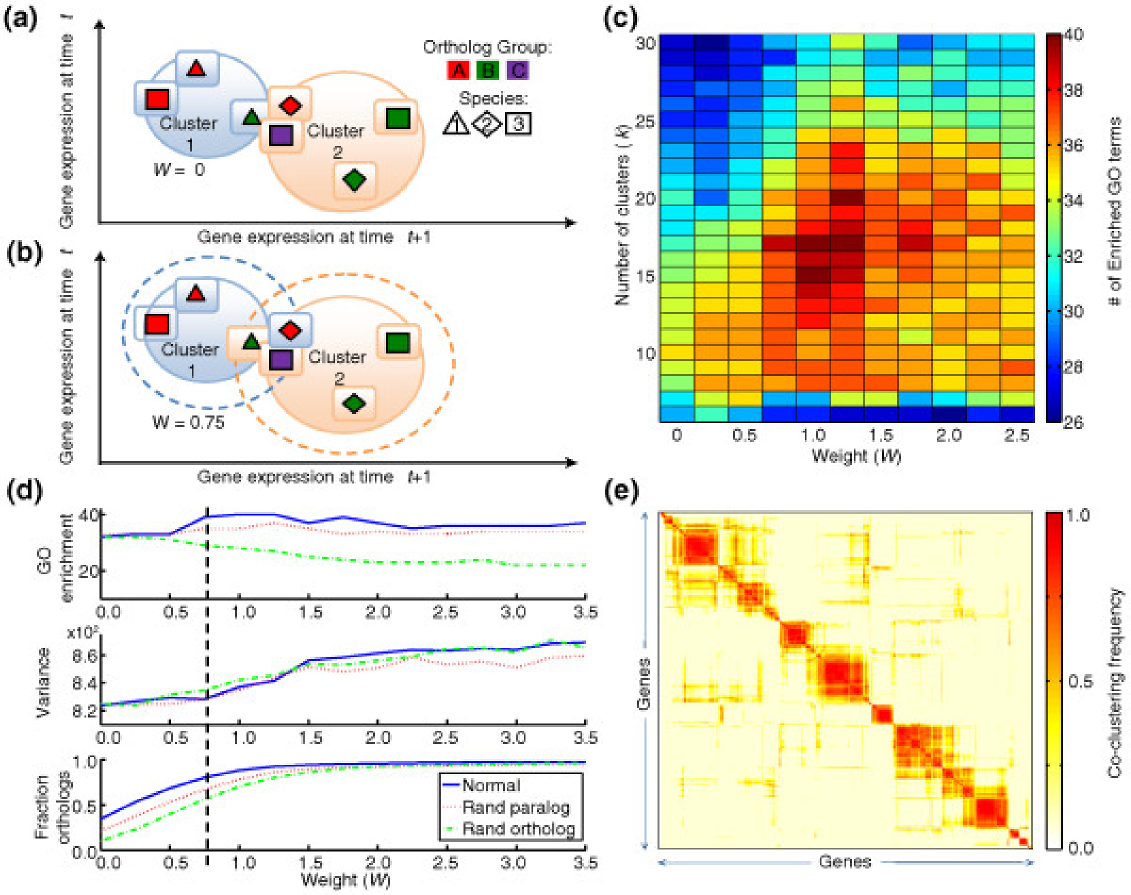


Figure 4-1: Soft clustering method

(a) Standard clustering based on expression only: two sets of orthologs are depicted (color represents orthology, shape represents species) where orthologs are split between clusters 1 and 2. For illustrative purposes, only two time points (*t* and *t* + 1) are shown. **(b)** Soft clustering based on expression and orthology: dashed circles denote regions where orthologs will be co-clustered. Since the purple square has no orthologs in cluster 1, it remains assigned to cluster 2. **(c)** Effect of number of clusters *k* and orthology weight *W* on GO term enrichment. **(d)** The number of enriched GO terms, variance, and fraction of co-clustered orthologs for *k* = 17 as a function of *W* in comparison to randomized paralogs/orthologs. Randomization was performed as described in Additional file 1: Randomizing the Orthology Mapping. **(e)** Since *k*-means is non-deterministic, to ensure robustness we performed 50 runs of the algorithm recording the fraction of times each gene pair was co-clustered (including all genes from all species). This matrix was hierarchically clustered.

An illustration of how *SoftClust* is different from traditional *k*-means clustering is depicted in Figures 4-1a,b. For illustrative purposes, we show a comparison using orthologous genes measured in two time points (t and $t + 1$). Orthologous genes that are within the 'extended orthology boundary' based on the reward W and the number of corresponding orthologs in each cluster, will change their cluster assignment. the appropriate value of the reward, W, can be determined using complementary information. Here, it was tuned to maximize the GO enrichment of the clusters (see Section 4.2.3 and Figure 4-1c).

The new objective function also leads to changes in the search algorithm for determining the optimal cluster assignments: for each group of orthologs across the three species, we search for the partitions that result in the minimum total distance between all pairs of group members. Since there are 2^m possible subgroups, where m is the size of the orthology group (here, most orthology groups are of size $m = 3$), and each subgroup is checked for all possible k clusters, the search complexity for each group is $O(2^m * k)$. Since m is small, the running time of the algorithm is typically very fast.

4.2.2 Algorithm Pseudo code and Implementation

SoftClust (data set X, # of clusters k, distance metric D, orthology relations Rel_{Orth}, orthology weight W_{orth})

1. Let $C_1 \ldots C_k$ be the k initial cluster centers.
2. Each gene $x \in X$ is assigned to cluster h^* (i.e., to set $X_{h^*}^{(t+1)}$) for:

$$h^* = \arg\min_h \Big(\sum_{I(x,x_i,x_j)\in \mathrm{Re}\, l_{Orth}} (D(x, C_h^{(t)}) - W_{orth}) \Big)$$

where $\sum_{I(x,x_i,x_j)\in \mathrm{Re}\, l_{Orth}} (D(x, C_h^{(t)}) - W_{orth})$ refers to all possible partitions of genes in the same orthology group into clusters based on the distance metric, D, by calling the recursive function: *PartitionSet(G)*.

3. For each cluster h, update its center by averaging all gene profiles assigned to it in step 2.
4. Iterate between (2) and (3) until convergence.
5. Return $\{C_1 \ldots C_k\}$.

```
**PartitionSet** (orthology set $G$)

1. If $|G|$ is 1, calculate distance of the gene to all clusters. Store the distance value and the best assignment.
2. Otherwise, for each possible partitioning ($P^i$) of $G$ to $j$ sub groups, $j$>1
      i. For each sub group $g^j$ in $P^i$
            a. If optimal partitioning of $g^j$ was already calculated, use the stored partitioning and distance value.
            b. Otherwise PartitionSet ($g^j$).
     ii. *Distance value of* $P^i$ = sum of optimal distance values for all sub groups $g^j$.
    iii. If the *Distance value of* $P^i$ is minimal, keep distance value and partitioning.
3. Calculate the reward for clustering $G$ together. Find the cluster that minimizes the rewarded assignment of $G$ when all genes are clustered together.
4. Store the minimal distance value (step 2 & 3) for corresponding partitioning of $G$
```

4.2.3 Selecting Parameters for the Constrained Clustering Method

Similar to the standard *k*-means clustering we need to specify the number of clusters (*k*). In addition we need to choose an appropriate reward weight (W_{orth}). We tested various number of clusters and reward weights using the total number of enriched GO terms (Bonferonni-corrected $p \leq 0.05$) as an external objective measure for selecting the values of these parameters. Figure 4-1c shows the median number of enriched GO terms for 50 runs at different k and W_{orth}. A clear enrichment of GO terms is seen for reward weights between 0.75 and 1.5. We chose the conservative parameters of $k = 17$ and $W_{orth} = 0.75$. While $W_{orth} = 1.0$ obtains a higher number of enriched GO terms than our choice of $W_{orth} = 0.75$, the associated variance from $W = 0.75$ to 1.0 increases by approximately 1.2% whereas the increase in variance from $W = 0$ to 0.75 is twenty fold smaller (0.06%). As we already observe large increases in ortholog co-clustering with a miniscule increase in cluster variance from $W = 0$ to 0.75, we believe that the more conservative choice of $W = 0.75$ is appropriate.

4.2.4 Constrained Clustering Leads to only a Small Increase in Inter-Class Variance

Our clustering method imposes a delicate balance between achieving noise reduction, more biologically meaningful clusters, and forcing divergent expression profiles to co-cluster. We assessed this balance by measuring the increase of the within-cluster-variance between rewarded and standard *k*-means runs. Increasing the reward weight results in increased cluster variance which are affected by orthology relationships. We computed the cumulative cluster variance for various reward weights. As can be seen (Figure 4-1d), the variance only marginally changes when using the reward selected for this set (W_{orth} = 0.75). A marked increase in the variance is seen for larger reward weights. This indicates that at the selected reward level expression profiles still plays a major role in cluster assignments. The reward (which is based on orthology relationships) is used only to move genes to clusters that provide a good fit in terms of expression profiles (even if not optimal). Thus, the reward achieves its intended goal: Identifying co-clustered orthologs while not dramatically affecting the resulting cluster profiles.

4.2.5 Randomizing Orthology Assignments Significantly Decreases Co-clustered Orthologs

The underlying assumption when using the orthology information to aid the clustering of gene expression, is that we expect orthologs to have similar expression profiles. To test this assumption we randomized the orthology mappings by assigning genes in one species to random genes in another while keeping the same number of orthology relationships. Next we applied our soft constraints clustering method using the randomized mapping. Two randomizations were employed; ortholog randomization and paralog randomization. For ortholog randomization, each gene was randomly substituted with a different gene from the same species, thus keeping orthology relationships intact. For paralog randomization, one of the paralogs was selected to be part of the orthology set. The remaining paralogs were randomized while keeping the true orthology relationships intact. It should be noted that only 443 orthology groups contained paralogs out of 4275 ortholog sets (10.3%).

As can be seen (Figure 4-1d, top graph) in the randomized mapping, the number of enriched GO terms is markedly reduced and the variance of the randomized mapping is larger for our selected reward value of W_{orth} = 0.75. The increase in variance is a direct result of the decrease in expression similarity between genes considered orthologs. This indicates that the true orthology relationships are indeed between similarly expressed genes. Moreover, the fraction of co-clustered orthologs is about 50% lower for the randomized mapping when using W_{orth} = 0.75 since this reward is not large enough to encourage distinct expression profiles to cluster together

(Figure 4-1d, bottom graph). Combined, these results support our assumptions regarding the similarity in expression of orthologous genes.

Lastly, as the *k*-means algorithm is initialized with random selection of the centroids, different clusters are produced for each run. We ran the clustering algorithm with the selected parameters for 50 times, noted the fraction that each pair of genes were clustered together out of the 50 runs, gathered the co-clustered genes in a matrix, and applied hierarchical clustering to the co-clustering matrix. A clear clustering structure is revealed (Figure 4-1e).

4.2.6 Analysis of Doubling Time Points vs. Absolute Time Points

We assessed the impact of using the number of doubling times in lieu of absolute times when choosing points for the time course. Since *Cg* had the shortest doubling time, we linearly interpolated the time courses of *Sc* and *Kl* to match that of *Cg*. We re-ran our clustering algorithm with the same parameters ($k = 17$, $W = 0.75$) and compared the number of co-clustered orthologs between species. We found that using doubling time greatly increased the number of co-clustered orthologs as compared to using absolute (interpolated) time points.

4.2.7 Motif Analysis

Four DNA-motif finding methods were used on each cluster: AlignACE [166], MEME [167], Weeder [168], and Consensus [169]. Default parameters were used for each method. The resulting position weight matrices (PWMs) from each method were used to scan species-specific promoter regions (using Patser [169]) for enrichment via the hypergeometric test. A promoter region was considered bound for MotifScore ≥ 0.7. MotifScore is calculated as the fraction of the maximum possible information content for the motif. All species intergenic regions were used as background for calculation of information content. Enriched PWMs were compared to known *S. cerevisiae* PWMs [170], [171] using the STAMP software package [172].

As an alternative to *de novo* DNA motif finding, we also used pre-defined PWMs [170], [171] to directly scan promoter regions using Patser [169]. Promoters were considered bound as previously described. This approach permits the calculation of an enrichment score for each PWM using the hypergeometric enrichment test followed by multiple test correction [173].

4.2.8 Species-specific Motifs

Scanning the promoter regions of *Cg* and *Kl* genes predisposes us to elucidating regulation by transcriptional regulators with close orthology to their *Sc* counterparts. In order to identify putative transcription factor binding sites (TFBS) unique to *C. glabrata* and *K. lactis* (either novel TFs or TFs with highly diverged DNA binding domains) we performed *de novo* motif search on each of the clusters previously described. In addition, we verified that each discovered motif was species-unique by calculating its hypergeometric enrichment in both co-clustered genes and orthologs. We discovered two *Kl* TFBS in clusters 15 and 8 ($p = 6.51 \times 10^{-4}$ and $p = 5.80 \times 10^{-19}$) showing high similarity (E ≈ 0) and a lack of enrichment in *Sc* and *Cg* clusters ($p \geq 0.109$). *Kl* genes in clusters 15 and 8 possessing this putative motif lack *Sc* and *Cg* orthologs. These promoters are also enriched for the ScHac1p TFBS (*Sc*: $p = 1.0$, *Cg*: $p = 0.158$, *Kl*: $p = 7.71 \times 10^{-7}$). A search of known TFBSs from TRANSFAC [171] shows the *Kl* motif has high similarity to the MEF-3 TFBS in mouse.

4.2.9 Expression Conservation of the General Stress Response

To examine the evolution of the transcriptional regulatory mechanisms of the conserved stress-response genes, we used the technique of phylogenetic profiling. 19 fungal genomes were selected from the Fungal Orthgroups Repository [174]. Nine orthologous differentially expressed genes that possess both RRPE and PAC motifs in *S. cerevisiae*, *C. glabrata*, and *K. lactis* were used as a reference for the entire orthogroup. The PAC and RRPE PWMs were searched against the orthogroup fungal sequences using Patser [169], and sequences in which the motif was found with MotifScore > 0.7 were determined as containing the motif. Previous work showed the PAC motif emerged during the *S. cerevisiae* - *C. albicans* divergence [175]. In contrast, our analysis suggests that the PAC motif first emerged by *Y. lipolytica* lineage and became well established in the fungal phylogeny by *A. gossypii*.

4.2.10 Identifying highly conserved and divergent pathways

We first ranked GO processes categories [176] based on their significance of overlap with differentially expressed orthologous groups [177]. An orthologous group was considered differentially expressed if at least one member was differentially expressed. We used the top 20 ranked GO processes for identifying conserved and divergent pathways. Conserved pathways were defined as those with the highest ‘full co-clustering’ fraction of genes known to be involved in the process and divergent pathways were defined as those with the highest ‘no co-clustering’ fractions.

4.2.11 Strains and growth conditions

Standard laboratory strains with known genomic sequence [178] were used: *Sc* BY4741, *Cg* CBS138 (ATCC 2001), and *Kl* NRRL Y-1140 (ATCC 8585). Cultures were grown in rich media (YPD) from OD_{600} of 0.05 to 0.2 at 30°C and 225 rpm. Cells were treated with fluconazole at species-specific sub-inhibitory concentrations, and harvested at 0, 1/3, 2/3, 1, 2 or 4 doubling times as measured for untreated cells.

4.2.12 Microarray expression profiling

RNA was isolated by hot phenol/chloroform extraction and enriched for mRNA via poly-A selection (Ambion 1916, Austin, TX, USA). mRNA from untreated cells was combined in equal amounts from all time points to form a species-specific reference sample. Six replicates per time point were dUTP labeled (three biological replicates by two technical replicates) with Cy3 and Cy5 dyes (Invitrogen SKU11904-018, Carlsbad, CA, USA) creating a dye-swapped reference design. Samples were hybridized to Agilent expression arrays using the protocol recommended by Agilent. Differential expression was called using the VERA error model [179] and false discovery rate multiple-test correction [173].

4.2.13 Insertion of ScAUS1/ScPDR11 into Kl

To facilitate insertion of *ScAUS1* and *ScPDR11* into *Kl*, open reading frames were placed under control of the strong $P_{LAC4\text{-}PBI}$ promoter by cloning into plasmid pKLAC2 (NEB N3742S), which possesses approximately 2-kb homology to the *Kl.LAC4* locus. Open reading frames were amplified with a *Sac*I restriction site (3′ end), which was used to ligate a kanamycin marker from pCR-Blunt (Invitrogen K-2800-20). *Xho*I (5′ end) and *Sbf*I (3′ end) restriction sites were added by PCR for ligation into pKLAC2. Modified plasmids were transformed into *Escherichia coli* and screened on Luria-Bertani media containing ampicillin and kanamycin. Plasmids were mini-prepped (GE Healthcare #US79220-50RXNS, Piscataway, NJ, USA) and verified by PCR and *Sac*II digestion. All restriction enzymes were obtained from New England Biolabs (Ipswich, MA, USA). *Sac*II-linearized plasmids were transformed into *Kl* NRRL Y-1140 by electroporation, thereby inserting *ScAUS1* and *ScPDR11* non-disruptively at the *Kl.LAC4* locus. Colonies were selected on YCB + 5 mM acetamide (New England Biolabs N3742S) and verified by PCR. mRNA expression of *ScAUS1* and *ScPDR11* was validated by quantitative RT-PCR.

4.3 Results

4.3.1 Comparative expression profiling of Sc, Cg, and Kl

We applied *SoftClust* to study three yeast species *Sc*, *Cg*, and *Kl* and examined the phenotypic response of these species to a range of concentrations of fluconazole, a triazole antifungal drug commonly used in the treatment and prevention of superficial and systemic fungal infections [158]. We found that *Kl* was approximately 70 times more sensitive to fluconazole than *Sc* and *Cg*. Cross-species differences in sensitivity could be due to a variety of factors, including differences in membrane permeability or drug transport, divergence in sequence or regulation of the drug target *Erg11*, or in any of the pathways previously linked to azole resistance.

While it is possible that complementary strategies might be observed at different fluconazole dosages [180], we exposed each species to fluconazole at its 50% inhibitory concentration to facilitate direct comparison of the transcriptional response between species. We then monitored global mRNA expression levels at 1/3, 2/3, 1, 2, and 4 population doubling times (Figure 4-2a). We also found that sampling based on the doubling time of each species, as opposed to absolute time measurements, led to greater coherence in the expression profiles across species. Selected mRNA measurements were validated using quantitative RT-PCR against six genes. We also found significant overlap of the *Sc* differentially expressed genes with several previous microarray studies and some overlap with gene deletions conferring fluconazole sensitivity.

To compare expression profiles across species, orthologous genes were defined using MultiParanoid [177]. As might be expected based on known phylogenetic distances [181], *Cg* shared more differentially expressed genes with *Sc* than with *Kl* (Figure 4-2b). We also found some overlap with previously published *C. albicans* microarray data, especially with the functions of the responsive genes such as those involved in ergosterol biosynthesis and oxido-reductase activity.

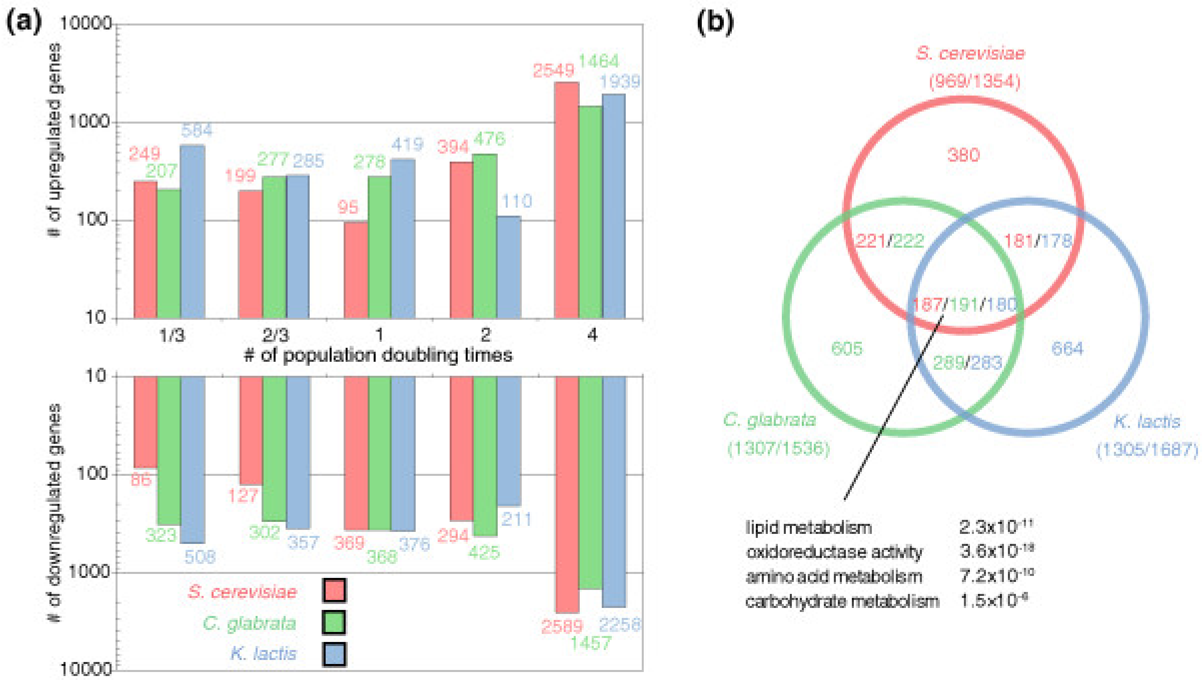


Figure 4-2: Differentially expressed genes

(a) Number of differentially expressed (up- and down-regulated) genes by species versus the number of cell doublings. **(b)** Venn diagram showing the overlap in the sets of differentially expressed genes selected in each species at a false discovery rate of $q \leq 0.1$. The number of differentially expressed genes in each region of the Venn diagram is not identical across species, since the number of genes that a species contributes to an orthologous group (that is, number of paralogs) can vary. Ratios in parentheses indicate the number of differentially expressed orthologs by the total number of differentially expressed genes (not all genes possess orthologs).

4.3.2 *SoftClust* analysis

The *SoftClust* analysis of the differentially expressed genes from the three yeast species resulted in 17 cross-species gene expression clusters (see Figures 4-3a,b). The number of clusters and appropriate reward W for orthologs co-clustering was determined by scanning the parameters over a range of values. We selected $W = 0.75$ and $k = 17$ as choices that approximately optimized the enrichment for Gene Ontology (GO) terms (See Figure 4-1c and Methods).

We compared our soft clustering approach to additional standard clustering methods. In comparison to classical k-means (equivalent to $W = 0$), the fraction of co-clustered orthologs increased from approximately 35% to 70%, with a negligible increase in within-cluster variance

(Figure 4-1d). For $W > 0.75$, we saw no improvement in the number of enriched GO terms, a marked increase in total cluster variance, and little improvement in the fraction of co-clustered orthologs.

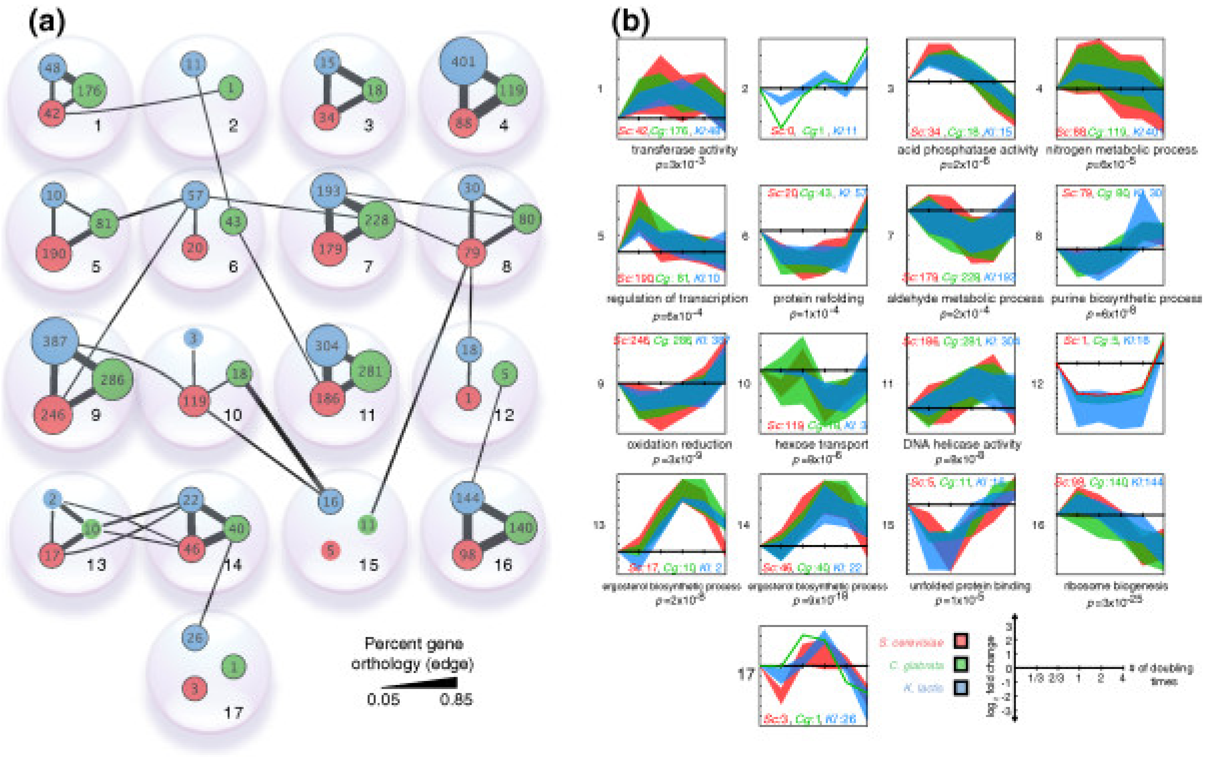


Figure 4-3: Cluster structure and dynamics

(a) Each of the 17 clusters appears as a bubble containing up to three colored nodes whose sizes represent the number of genes contributed by each species. Edge thickness denotes the percent of gene orthology shared within or between clusters, measured using the size of the intersection divided by the size of the union of the sample sets. Only significant edges ($P < 0.01$) are shown. Several clusters show conserved orthology but not dynamics (for example, cluster 10 *Sc*, *Cg* with cluster 15 *Kl*). Note that clusters were ordered to minimize orthology edge crossings. **(b)** Expression dynamics of the 17 soft clusters over time following fluconazole exposure. The width of each band corresponds to ± one standard deviation about the mean. A selection of enriched GO terms is shown for different clusters. The number of genes for each species in each cluster is also shown.

4.3.3 Conservation of cis-regulatory motifs across clusters

We analyzed the resulting clusters using the GO Biological Process and found that two cross-species clusters (13 and 14) were highly enriched for ergosterol biosynthetic genes ($P \leq 10^{-8}$) and were coherently up-regulated in all three species - likely in response to ergosterol depletion. Both clusters were also enriched for the upstream DNA-binding motif of the sterol biosynthesis regulators Ecm22 and Upc2 [182]. Interestingly, Upc2 has also been implicated in increased fluconazole resistance in the fungal pathogen *C. albicans*[183]. Rox1 motifs were enriched in *Sc* and *Cg* but not *Kl*. A likely explanation for this divergence is that Rox1 is a repressor of hypoxia-induced genes, and *Kl* both lacks a Rox1 ortholog and the capacity for anaerobic growth.

Beyond the clusters representing ergosterol biosynthesis, we found two additional clusters (9 and 16) in which high conservation of expression patterns, sequence orthology, and *cis*-motif conservation were observed across species. Cluster 9 was regulated by the general stress-response transcription factors Msn2p and Msn4p ($q < 10^{-5}$) and showed GO enrichment for oxido-reductase activity ($q < 10^{-8}$) and carbohydrate metabolism ($q < 10^{-7}$). Cluster 16 was enriched for ribosomal biogenesis and assembly ($q < 10^{-13}$) with upstream PAC [184] and RRPE motifs previously implicated in regulating genes involved in the general stress response and ribosomal regulation [184], [185].

For other clusters, conserved motifs were absent, suggesting divergence across species. This lack of motif conservation was particularly surprising for clusters 3, 4, 7, and 11, which contained large numbers of co-expressed orthologous genes. On the other hand, this finding is consistent with previous studies finding low motif conservation. We also found no significant enrichment for binding sites of orthologous transcription factors (Tac1, Mrr1, Crz1) known to mediate fluconazole-resistance in the evolutionarily diverged pathogen *C. albicans*[186].

4.3.4 Co-clustering implicates both highly conserved and divergent pathways

Another advantage *SoftClust* has over standard clustering methods, which focus solely on cluster coherence, is that it can simultaneously detect both similar and divergent behavior between orthologs. For instance, when orthologs are not co-clustered despite the addition of a reward, one can be assured that their dynamic profiles truly differ. Some clusters shared significant gene orthology (but not expression) with other clusters, such as clusters 10 and 15 in Figure 4-3a. In these cases, we also found no conserved motifs between these clusters, indicating both promoter and expression divergence among orthologs in addition to species-specific motifs.

We further analyzed the clusters to identify pathways for which the fluconazole response is either highly conserved or strikingly divergent. For this purpose, differentially expressed pathways were identified using the GO Biological Process (see Methods). For each pathway, we computed the number of orthologous gene groups for which: 1) all three species were in the same cluster (full co-clustering); 2) two species were in the same cluster (partial co-clustering); or 3) no two species were in the same cluster (no co-clustering). The pathways with the highest percentage of orthologs with full co-clustering are shown in Figure 4-4a. The pathways with the highest percentage of orthologs that do not co-cluster are shown in Figure 4-4b. By this analysis, the most conserved pathway was ergosterol biosynthesis, which is consistent with our study of conserved motifs (above). Fluconazole directly inhibits ergosterol synthesis by targeting of *Erg11*, and all species appear to respond strongly to this reduction in ergosterol by up-regulating the enzymes required for its novel biosynthesis. *ERG11* was up-regulated early in both *Sc* and *Cg* and later in *Kl*. Since *ERG11* over-expression is one mechanism by which yeast can overcome fluconazole-induced growth inhibition [187], delays in its induction could contribute to *Kl*'s greater fluconazole sensitivity.

The first stages of ergosterol biosynthesis are carried out by a subset of enzymes of the isoprenoid pathway. While most ergosterol genes were coordinately up-regulated in all three species, the expression levels of isoprenoid biosynthesis genes were strikingly divergent (see Figures 4-4b,d). In all eukaryotes, regulation of isoprenoid biosynthesis is known to be complex with multiple levels of feedback inhibition [188]. Thus, the extensive divergence in isoprenoid biosynthesis expression suggests that the regulation of this pathway has also diverged between species.

Extensive expression divergence was also observed in methionine biosynthesis and amino acid transport (see Figure 4-4b). Curiously, many *Cg* methionine biosynthesis orthologs were strongly down-regulated early in the time-course (see Figure 4-4e). This strong down-regulation was not mirrored in *Sc* and *Kl*, which displayed divergent expression responses that were not co-clustered. Interestingly, it has been previously suggested that differences in methionine biosynthesis may alter azole susceptibility in *C. neoformans* [189] and *C. albicans* [190].

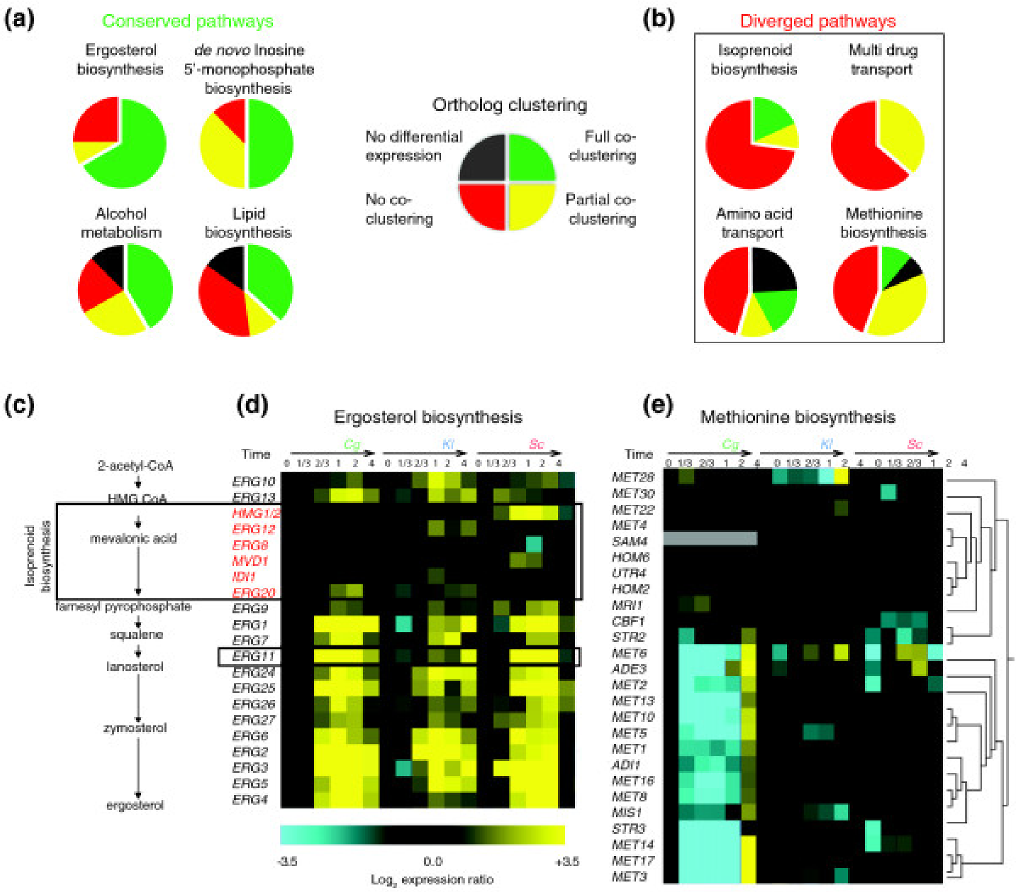


Figure 4-4: Pathway expression conservation and divergence

(a) Top conserved and **(b)** diverged pathway responses as revealed by the soft clustering approach. Each pathway is represented by a pie with four slices - green, yellow, red, and black - denoting the percentage of orthologs in that pathway for which all three species co-clustered, two species co-clustered, no two species co-clustered, and no species' orthologs were differentially expressed, respectively. Pathways were defined using GO biological process annotations. **(c)** Schematic of ergosterol biosynthesis, the most conserved pathway response. Interestingly, this pathway includes isoprenoid biosynthesis, for which the response was one of the most divergent. **(d)** mRNA expression responses of ergosterol pathway genes are shown in order of occurrence in the pathway. Expression levels of genes 3 to 8 (boxed, and red) corresponding to isoprenoid biosynthesis are strikingly divergent. The fluconazole target Erg11 is boxed. **(e)** Hierarchically clustered mRNA expression responses of methionine biosynthesis genes show extensive divergence across species. Grey expression values denote a gene for which the species lacks an ortholog.

4.3.5 Major divergence in mRNA expression of transporters

A final pathway for which we observed striking expression divergence was multi-drug transport (see Figure 4-4b). Most genes in this pathway were covered by clusters 8, 11, 16 (see Figure 4-5a,b). Multi-drug transporters are divided into two classes: ATP-binding cassette (ABC) and major facilitator superfamily (MFS) transporters [159]. We examined the expression patterns of these transporters and found at least two types of divergent behaviors. First, the fraction of differentially expressed *Sc* MFS transporters was low compared to *Cg* and *Kl* (Fisher exact test, one-tailed *P* = 0.025 and 0.020, respectively). Second, the timing of MFS gene expression differed, with *Sc* up-regulated late and *Cg* up-regulated early (Figure 4-5b). In *SC*, several ABC and MFS transporters have been shown to bind fluconazole as a substrate [191], [192]. Of these, we found that the *PDR5*/*10/15* family of ABC transporters was up-regulated in *Cg* and *Sc* but not *Kl*. Another fluconazole transporter, *SNQ2*, was up-regulated in *Cg* only.

We also found strong differences in the expression of other multi-drug transporters that have not been previously linked to fluconazole: *PDR12* was strongly down-regulated in *Sc* and *Cg* but up-regulated in *Kl*; *ATR1* and *YOR378W* were up-regulated in *Cg* and *Kl* but not *Sc*; *HOL1* was up-regulated in *Sc* and *Kl* but not *Cg*. Some transporters also showed differences in expression timing (*YOR1*, *PDR12*).

Additionally, two ABC transporters, *AUS1* and *PDR11*, which uptake sterol under anaerobic conditions [193], were up-regulated in *Sc* but were not differentially expressed in *Cg* (*Cg* does not possess a *PDR11* ortholog). This suggests that *Sc* but not *Cg* increases sterol transport during fluconazole exposure. Intriguingly, since the direct effect of fluconazole is to inhibit sterol synthesis, increased sterol transport could be a mechanism for increased fluconazole tolerance. In support of this hypothesis, we found that the normally repressed cell wall mannoprotein DAN1, whose expression is required for sterol uptake [194], was up-regulated in *Sc* but not *Cg*. Since *Kl* lacks sterol transporters, it cannot import sterol and only grows aerobically [195], [196]. As a possible explanation for this divergent behavior, we found that the promoter regions of *ScAUS1*, *ScPDR11*, and *ScDAN1* contain binding motifs for ergosterol biosynthesis and/or sterol transport regulators Ecm22p, Rox1p and Sut1p, all of which were absent upstream of *CgAUS1* and *CgDAN1*.

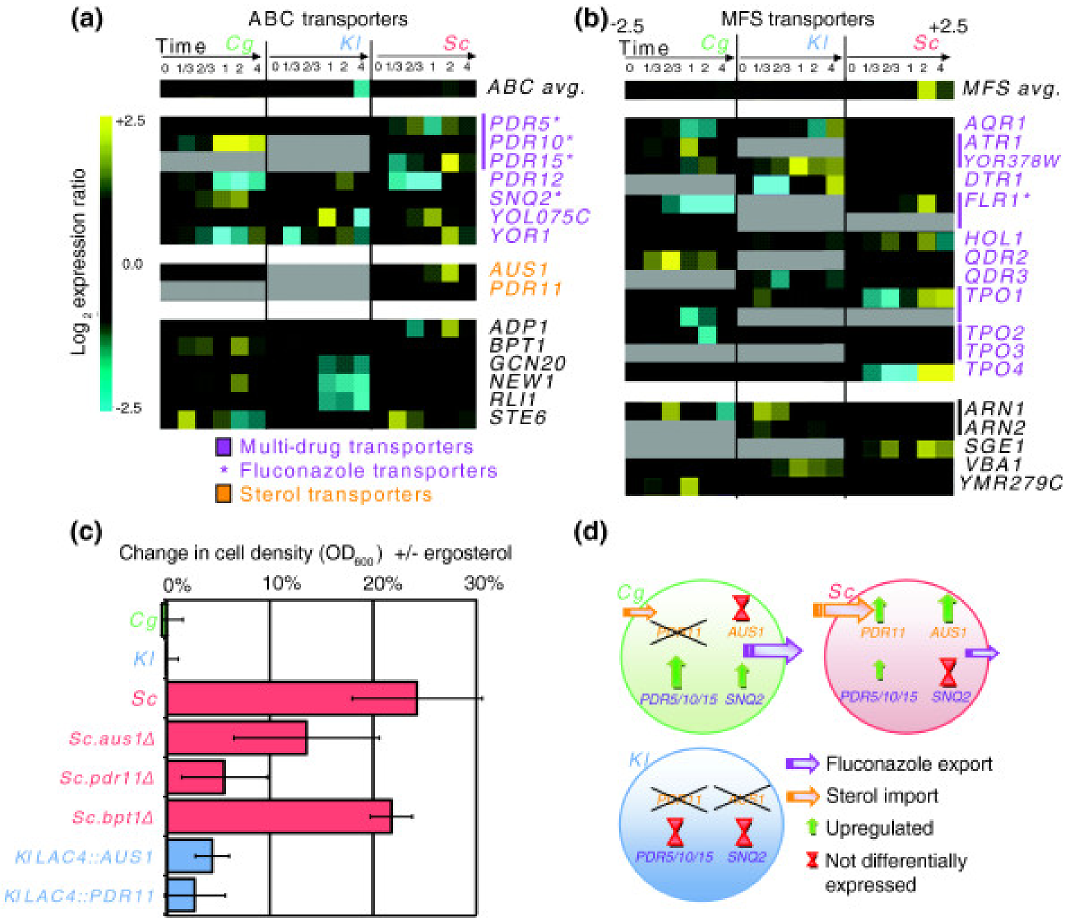


Figure 4-5: Divergence in transporter usage

Cross-species expression profiles of **(a)** ATP-binding cassette (ABC) and **(b)** major facilitator superfamily (MFS) transporters are shown. Grey expression values denote a gene for which the species lacks an ortholog. **(c)** Change in cell density with addition of exogenous ergosterol at the fluconazole 50% inhibitory concentration across different mutant backgrounds. *Sc.bpt1Δ* is a gene knockout unrelated to fluconazole response and is included as a control. Error bars indicate one standard deviation. **(d)** Model for differential usage of transporters among *Sc*, *Cg*, and *Kl*.

Therefore, the striking divergence in expression of fluconazole export and sterol import pathways suggests differing strategies in the azole response: following fluconazole exposure, *Sc* appears to activate sterol influx through up-regulation of *PDR11* and *AUS1*; in contrast, *Cg* may activate fluconazole efflux through strong up-regulation of *SNQ2* and a *PDR5/10/15* ortholog (Figure 4-5a).

4.3.6 Sterol import increases fluconazole tolerance in Sc, but not Cg or Kl

To investigate these hypotheses, we grew wild-type *Sc* and *Cg* along with deletion mutants *Sc.aus1Δ* and *Sc.pdr11Δ* under fluconazole treatment in the presence or absence of exogenous ergosterol (4 µg/ml). As shown in Figure 4-5c, we found that addition of ergosterol had no effect on growth of *Cg* but led to an increase in growth of *Sc* (*P* = 0.018). This increase was attenuated in *Sc.aus1Δ* and *Sc.pdr11Δ* (*P* = 0.033), which lack sterol import genes, but not in an unrelated control knockout, *Sc.bpt1*. Thus, *Sc* but not *Cg* is aided by adding ergosterol to the environment, and this process is likely dependent on *AUS1* and/or *PDR11*.

Three additional lines of evidence support the hypothesis that *Sc* prefers sterol import while *Cg* prefers fluconazole export in response to fluconazole treatment. A retrospective analysis of deletion mutant fitness in *Sc* revealed that a greater proportion of gene deletions involved in the sterol pathway lead to fluconazole sensitivity than deletion of fluconazole transporters themselves (Fisher exact test, one-tailed *P* = 0.043). This suggests a role for sterol transporters in the *Sc* fluconazole response. Second, fluconazole tolerance in *Cg* has been shown to be unaffected when constitutively expressing *CgAUS1* in the presence of exogenous free cholesterol (though not in the presence of serum) [197]. Third, deletion of the *Cg* orthologs of fluconazole transporters *PDR5* (*CgCDR1*) [162] or *SNQ2* [198] both resulted in increased fluconazole sensitivity.

4.3.7 Expression of sterol importers in Kl increases fluconazole tolerance

Since *Kl* neither up-regulates drug exporters nor encodes sterol importers, we considered that this lack of a transport response might be responsible for the higher drug sensitivity we observed for *Kl* in relation to the other species. Consistent with this hypothesis, we found that *Kl* growth was unaffected by addition of exogenous ergosterol (Figure 4-5c), similar to *Cg* but in sharp contrast to *Sc*. We also predicted that transgenic expression of sterol importers ScAus1 or ScPdr11 in *Kl* might increase fluconazole tolerance in the presence of exogenous ergosterol. To test this prediction, we chromosomally integrated *ScAUS1* and *ScPDR11* into *Kl* non-disruptively at the *KlLAC4* locus under control of the strong constitutive *Kl* $P_{LAC4\text{-}PBI}$ promoter (Materials and methods). Transformed *Kl* strains were grown under fluconazole treatment with and without exogenous ergosterol (4 µg/ml). We observed that transgenic expression of sterol importer *AUS1* in *Kl* significantly increased fluconazole tolerance (*P* = 0.012; Figure 4-5c) in an ergosterol-dependent manner. Thus, it appears that differences in sterol import and drug export are responsible for a component of the anti-fungal response, and of the observed functional divergence across the three yeast species.

4.4 Conclusions

This study introduced a novel clustering method, *SoftClust*, which cluster expression data from several species concurrently. This approach is distinct from other methods for cross-species expression analysis [153], [154], [199], [200] in several important ways. Chief among these, it integrates sequence orthology with gene expression patterns to produce accurate orthologous clusters. This integration is accomplished by a symmetric process that does not require the designation of one species as a reference. In addition, *SoftClust* handles data from more than two species and can, in principle, analyze any number of species simultaneously. Lastly, orthology groups that were not assigned to the same cluster despite the reward are a good indication for a divergent biological process.

We showed here a specific application of *SoftClust* to compare the dynamic transcriptional responses of three diverse yeast species to fluconazole treatment, revealing significant divergence in their regulatory programs. Analyzing the clusters produced by *SoftClust* we revealed specific biological processes showing conserved and divergent mechanisms. Specifically *SoftClust* found several different mechanisms of azole tolerance, depending on the species (Figure 4-5d). The *Sc* response depends on sterol influx, through up-regulation of *PDR11* and *AUS1*. In contrast, the *Cg* response relies on fluconazole efflux through strong up-regulation of *SNQ2* and a *PDR5*/*10/15* ortholog. Neither of these responses has evolved in *Kl*, leading to its severe drug sensitivity. These conclusions are supported by follow-up experiments demonstrating that growth in ergosterol increases the fluconazole tolerance of *Sc*, but not other species, in a *PDR11*- and *AUS1*-dependent fashion. They are also supported by the finding that transgenic expression of *AUS1* in *Kl* increases the fluconazole tolerance of this species.

The *SoftClust* approach can be applied to a wide range of species, conditions, and stimuli to reveal conserved and divergent molecular response pathways. For example, *SoftClust* was also used in a study comparing mice and macaques response to two bacterial (*M. tuberculosis* and *F. tularensis* Schu S4) and two viral (Influenza A/PR8 and A/Fuj/02) infections and found similarities and species-specific response patterns. See Zinman *et al.* [201] for details.

5 *Module Blast* – Finding Active Subnetworks Across Species

In the previous chapter we showed how co-clustering can provide both global and local biological insights. However, it is difficult to delineate the molecular mechanisms driving the responses from the clustering results alone. In this chapter we show that by combining the condition-specific expression data with gene association networks we can find connected groups of genes (or sub-networks) that can model better particular biological processes of interest. Finding an active sub-network is a hard problem and applying it across species requires further considerations with regard to orthology information, expression data, and networks from different sources. We devised a novel approach, named *ModuleBlast*, for facilitating cross species comparison of network activation patterns. *ModuleBlast* uses both expression and network topology to search for highly relevant subnetworks and to classify them as either conserved or divergent. We have applied *ModuleBlast* to data from mouse and macaque macrophages infected with *Francisella tularensis*, a highly virulent and often deadly bacterium. Several relevant modules were identified, consistent with recent findings on apoptosis and NFκB activation following infection. We performed follow up biological experiments to support these results, which highlight the advantages of cross species analysis of interaction networks. *ModuleBlast* is implemented as a web tool and offers easy-to-use web interface with built-in support for several species.

5.1 Introduction

In recent years increasing importance is given to understanding condition-specific responses by using gene expression studies and the number of microarray experiments is growing exponentially [55]. Traditional microarray analysis focused mainly on finding differentially expressed genes between treatment and control to find the genes that are involved in the condition being studied. While such studies led to useful results, genes and proteins usually operate in groups or cascades and are often post transcriptionally regulated, so in many cases important genes may not be differentially expressed and are missed when using only expression data. Interaction data is useful for identifying such genes [202] and an increasing number of studies attempted to integrate static gene and protein interaction data with dynamic expression data in order to find 'active subnetworks'. Such subnetworks are represented by connected regions within the gene interaction network that contains several genes that are differentially expressed between treatments and controls. The active sub-network approach was used to generate concrete testable hypotheses for the underlying mechanisms governing the changes in gene expression [202–210]. Many researchers refer to active subnetworks as 'modules' underscoring the ability of these subnetworks to capture coherent functionality. In this manuscript we will use the terms 'subnetworks' and 'modules' interchangeably.

Several studies have utilized cross species expression data for studying the same condition in multiple species [48], [55], [201]. Such analyses highlight the similarities and differences in key mechanisms between the species, improving our biological understanding both from an evolutionary point of view [47] and for the activity under specific conditions including drug response [49]. These two types of analyses (sub-networks and cross species comparisons) have primarily remained separate with researchers either using one or the other in each study. While the active subnetworks approach can be an excellent tool for analyzing expression pattern differences between species and tracking the origins of these differences, to do so we need to overcome several challenges. These include orthology assignments, comparison of expression patterns across platforms and species, and differences in the association networks.

Optimally finding active sub-networks or modules in a general graph was shown to be NP-hard [202] leading to several heuristic and approximation algorithms suggested for this problem. Previous approaches for finding such networks in a single species included using pre-defined groups [211], [212], simulated annealing [202–204], greedy search approaches [205–208], optimization algorithms [209], [210], and methods based on graph theory [213], [214]. See Table 5-1 and [215] for further details. In practice greedy search approaches were shown to obtain good results with significantly shorter runtime compared to other approaches [206]. We are aware of only one study that attempted to combine active sub-network discovery with cross species analysis [59]. In that paper, the authors tried to search for active subnetworks across species with the goal of finding connected components that present similar expression patterns across both species.

While similarities are important for identifying common mechanisms, differences are also important, for example, when trying to determine if a model organism is suitable for human studies. We have thus developed a novel method, *ModuleBlast,* which addresses the need for an integrated analysis of network data across species allowing for the identification of both conserved and divergent subnetworks. Another important aspect missing from prior work on cross species sub-network analysis which we address with our new method is using temporal data. *ModuleBlast* can link modules over different time points to identify causal effects leading to expression changes. Our method is based on integrating expression and interaction data from two or more species and searching for active modules using the absolute activation in all species. Module search expands highly activated seeds into modules as long as the overall activation of the modules is maximized. Comparing the expression differences between the species and the overall activation in each to random data we can classify the modules to conserved (CM), species specific (SP), and divergent in opposing patterns (OP).

We used *ModuleBlast* to study the temporal response of murine and macaque macrophages to *Francisella tularensis* subsp. *tularensis* Schu S4, a gram-negative bacteria that is highly virulent in humans [216]. The resulting modules provide an overview of the dynamic response in these two species and were used to generate hypotheses regarding the cell response to the infection. Functional analysis indicates that several of the modules we identified were related to immune response. One of the conserved modules we identified is significantly enriched with apoptosis and anti-apoptosis related genes and, using additional protein-DNA interaction data, we determined that it is associated with NFκB. We performed follow up experiments and were able to correlate NFκB and this module, leading to new perspectives into the complexities between *F. tularensis* and host cell interactions.

Table 5-1: Methods for searching for active subnetworks

Summary of the main approaches for searching for active subnetworks and key studies using each approach.

Method	Paper	Description
Simulated annealing	[202][203] [204]	Defines scoring function over connected components. Search is slow, produces large sub-networks. Variations: combinations with conventional gene set analysis, or combination of PPI and co-expression.
Greedy search	[205]	Algorithm based on BFS followed by attempts to merge sub-networks, and then pruning of insignificant nodes. This algorithm is faster and produces smaller sub-graphs.
	[206]	Tried different scoring functions based on a greedy search. Fast and proved to produce good results.
	[207], [208]	Search based on probabilistic seeds (higher probabilities to vertices with high score). Runtime is slower compared with other greedy methods.
	[59]	The only cross species algorithm. Works in pa1rallel on two species trying every node as a seed. Runtime is very slow
Optimization algorithm (Exact approach)	[209] [210]	Based on Maximum weight connected sub-graph problem and Prize-collecting Steiner tree problem. Solution based on Integer programming. Supposedly finds an optimal solution, but requires many assumptions and parameters. Problematic with large networks.
Graph theory	[213]	Graph theory using q-connected modules and dynamic-minimum-cut problem. Covering using Shortest paths algorithm with heuristic rules.

5.2 Methods

We first discuss our method for generating a cross species network that can be used to identify both similarities and differences between species. Once we have networks that integrate information across species, we use a novel target function that takes into account both activity and connectivity to search for active sub-networks. Our target function is useful for identifying not just the active / repressed genes but also proteins that may be affecting the expression of these genes even if they are post-transcriptionally regulated. We present a greedy search method for finding modules (sub-networks) that maximize the target function and discuss how such modules can be connected in time when using time series expression data to identify the progression of information within cells. A general outline for searching method is shown in Figure 5-1.

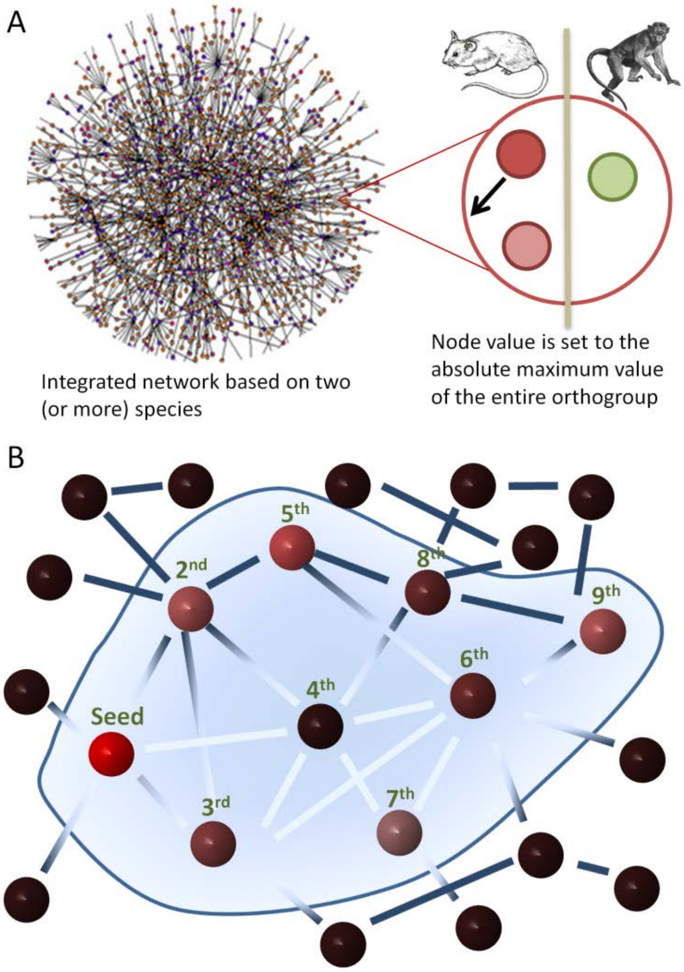

A
Integrated network based on two (or more) species
Node value is set to the absolute maximum value of the entire orthogroup
B
Seed
2nd
3rd
4th
5th
6th
7th
8th
9th

Figure 5-1: Method for searching for active subnetworks
(**a**) An integrated network from two (or more) species is assembled in the following manner; nodes represent entire orthogroups and the node value is set to the absolute maximum value in the orthogroup (from either species). Edges are the union of all individual species edges that connect genes in any two orthogroups. Edges observed in more than one species have a higher score. **(b)** Searching for active modules. Node colors represent their absolute activation. Module search starts by expanding initial good seeds (seed shown on left). Additional nodes are added based on activity and connectivity to current module members. Modules are extended as long as the overall module score improves and up to a user defined limit.

5.2.1 Generating cross species gene association networks

We assembled gene association networks using various genomic data types including protein-protein interactions and genetic interactions from BioGRID [32], version 2.0.63. To combine multiple species we created networks in which nodes correspond to entire orthogroups containing gene orthologs from all the relevant species based on Inparanoid [17] orthogroup definitions. We are interested in finding subnetworks with high activation, regardless of the species, therefore the node score is set to be the absolute of the most extreme value in the orthogroups (Figure 5-1). Edges between nodes in our network include interactions connecting any of the genes in the two orthogroups. Interactions can be weighted according to the confidence in the interaction if one is provided (e.g., the log likelihood score (LLS) [83], see Chapter 2 for more details) and summed across various data types and species. As these weights are comparable across species, it is possible to merge association networks from several species into a single association network. This construction is supported by the analysis performed in Chapter 2 showing that interactions are conserved at a network organization level. Edges that have evidence in more than one species have a higher weight indicating the increased confidence in seeing this edge across species. As in this study, we used for both species interactions only from BioGRID, which does not provide confidence measurement for the interactions, hence the normalized assigned weights could practically get a weight of either 1 or 2. Integrating data from several species into a single gene association network allows us to overcome some of these issues related to missing data. We note that current interaction data, especially of higher organisms including mammals, is incomplete, and in this study human interactions contributed

~90% of the edges. However, over 50% of the mouse interactions were also found in human. Our final joint network contained 6188 nodes and 21655 unique edges.

5.2.2 Scoring sub-networks

Given a connected sub-network, early studies [202] scored it by summing the node scores (which were assumed to be drawn from a standard normal distribution) with respect to background distribution setting, e.g.:

$$S_j = \frac{1}{\sigma_Z}\frac{\sum_i Z_i - \beta_Z \mu_Z}{\sqrt{M}} \qquad \text{(5-1)}$$

Where Z_i is a node score for a node in module j (based on the absolute of the most extreme value of the orthogroup and was shown to follow a normal distribution with parameters (μ_Z ,σ_Z). The subscript Z refers to parameters pertaining to nodes. M is the number of nodes in sub-network j. β_Z is an empirical parameter designed to produce fewer nodes with a positive score, hence creating smaller sub-networks [205]. Practically, S_j is a sum over normal standard variables, which can be estimated as a normal variable with mean μ/M and standard deviation $\sigma/\sqrt{M}$. Note that scores S_j of randomized subnets are guaranteed to have a mean of 0 and standard deviation of 1.

While node scores are useful, they can only identify transcriptionally regulated genes. To allow the identification of proteins that are post-transcriptionally regulated, but may still be affecting activated or repressed genes, we integrated node scores with interaction scores by looking at the density and overall score of edges in a sub-network. We thus extended the node score objective function leading to a target function that is a weighted sum of two components; nodes and edges scores:

$$S_j = \frac{}{\sigma_Z}\frac{\sum_i (Z_i - \beta_Z \mu_Z)}{\sqrt{M}} + W\frac{(\sum_i E_i) - \mu_{E(M)}}{\sigma_{E(M)}} \qquad \text{(5-2)}$$

The first part is the same as Eq. 5-1 above. The second uses a similar idea to score the edges in each sub-network. E_i is the edge score (see previous section) defined for edge i for an edge in module j, $\mu_{E(M)}$, $\sigma_{E(M)}$ are respectively the mean and SD of edge scores calculated for a module of size M. Computing the background statistics for the edges score component is less straight forward as it is a function of the number of nodes. The number of nodes in a sub-network sets minimum and maximum limits on the number of edges in this sub-network. To learn the conditional distribution of edge scores (as a function of size, $\mu_{E(M)}$, $\sigma_{E(M)}$) we used an iterative approach using randomized modules. We calculated mean and standard deviation of the edge score distribution over random modules for every possible module size *M* between 1 and 150.

In addition, the score of the edge component in Equation 5-2 is dependent on the topology and density of the network. To balance the impact of edge component on the overall score of the sub-network, we combine the node score and edge score components using a tunable weight parameter *W*. The optimal value of *W* is determined based on the parameter selection criterion described below.

5.2.3 Searching for high scoring sub-networks

We followed a greedy search approach that allows us to quickly evaluate several potential modules. Greedy search was previously shown to produce good results when searching for sub-networks [206]. Our search procedure starts by selecting seed nodes (e.g., highly activated nodes) that are expanded using breadth first search by evaluating the objective function described above. In each step, we add the node that maximizes the objective function for the subset of nodes that were previously selected. Nodes can be added to the component as long as the overall score of the component increases. We set an optional minimal active module size to five nodes which allows the algorithm to grow seeds with high initial scores. An optional maximal size limit (before module merging) ensures the coherency of the modules over a range of parameters. As nodes can appear in more than one module, highly intersecting modules are merged by keeping only edges that are appear in a high percentage of the modules. Edges that are found to be part of several modules starting from different seeds are more likely to be relevant to the analysis. This is evaluated by calculating the number of appearances of an edge *e*(*a*,*b*) between node *a* and node *b* in any module, divided by the maximum number of appearances of node *a* or node *b* in any module, and comparing this calculation to some pre-defined cutoff. In set notation this can be written as:

For each $e(a,b) \in E$, $e(a,b)$ is kept if

$$\frac{\sum_j I(M_j,a) \wedge I(M_j,b)}{\max(\sum_j I(M_j,a), \sum_j I(M_j,b))} > cutoff \qquad (5\text{-}3)$$

where *E* is the set of all edges, *e*(*a*,*b*) is an edge connecting nodes *a* and *b*, M_j is a module identified by iterator *j* for all possible modules (1..*n*), and I(M_j,*x*) is an indicator function of whether node *x* is part of module M_j, for nodes *a* and *b* respectively. In experiments we performed we observed that a cutoff value of 0.5 works well, as lower cutoffs often result in more edges and large modules, and higher cutoffs result in few edges and small modules. Varying the parameters within a reasonable range [0.2-5] had little effect on the assessment criteria (see Parameter selection Criterion). Modules with less than 3 nodes are omitted. To conclude, our method requires the setting of three parameters: The weight *W* of the edge score component in Equation 5-2 (which can be learned, see below), the maximal size of a module before merging, and the cutoff from equation 5-3.

5.2.4 Parameter Selection Criterion

As with any data integration method, we need to determine the weight assigned to each data type (*W* in our model). The best method to determine parameter values is by using a gold standard set (e.g., known modules in the condition) and, in a training procedure, choosing values that lead to the best recovery of such known modules. However, in our case little is known about modules that are activated during different types of infection, and we expect this to be a general problem for other studies as well. Thus, to determine the value of this parameter we searched for values that optimize the following three general criterions (so that it is applicable across a wide range of conditions being studies): 1) the percentage of the number of modules that contain uniquely enriched GO biological processes terms [176]. (Alternatively the KEGG biological pathways [217] can be used, although the granularity level of KEGG is less refined compared to GO). This criterion examines the ability of the algorithm to capture distinct biological processes. 2) The percentage of the number of differentially expressed nodes out of total nodes in the selected modules. This criterion aims to explain as many of the observed expression changes as possible. 3) The total number of modules. This criterion maximizes the number of distinct processes that are captured by the algorithm. Taken together the three criteria aim to produce the most relevant set of modules for the given expression data.

5.2.5 Assessing module conservation

As mentioned above, for each node in our network we have at least two scores (one from each species). While we use the maximum absolute value when searching for subnetworks, once these are found, we can study and compare the activation of nodes from the two species in each module. In order to evaluate the convergence or divergence of modules we calculated for each module the Euclidean distance over all the nodes in the module using the difference between the most extreme values in the genes represented by each node in the two species. Specifically, we compute:

$$Diff(S_j) = \frac{\sum_i |Z_{iA} - Z_{iB}|}{\sqrt{M}} \qquad \text{(5-4)}$$

where S_j is a sub-network of size M, Z_{iA} and Z_{iB} are node scores for all nodes i in S_j from species A and B respectively. The distance obtained for each module was compared to distances calculated for 10000 random modules with the same number of nodes, using nodes that are part of some module.

In addition, we assessed the overall activation / repression of the modules in each of the species using a similar randomization method over the sum of values in nodes for each of the species separately, i.e.:

$$Active(S_j, X) = \frac{\sum_i |Z_{iX}|}{\sqrt{M}} \qquad \text{(5-5)}$$

Where S_j is a sub-network of size M and Z_{iX} is a node score for all nodes i in S_j from species X. In a similar method to the *Diff* calculations, we compared the obtained distances to distances calculated for 10000 random modules with the same number of nodes, using nodes that are part of some module.

Using these measurements we can classify the modules into three categories:

- CM - Conserved modules. These modules show little difference between the species compared to random modules (*Diff(S_j)* <0.1), i.e., at least 90% of the random modules are more divergent than the inspected module.
- SP - Modules that are species specific. Modules that are different between the species (*Diff(S_j)*>0.9) and show high activation (*Active(S_j,A)* > 0.9) in one of the species and low activation in the other (*Active(S_j,B)* < 0.1).
- OP - Divergent modules that are divergent in opposing patterns (e.g., one is up-regulated and the other is down-regulated). These modules are highly different between the species (*Diff(S_j)*> 0.9) and show high activation in both species, i.e., *Active(S_j,A)* > 0.9 and *Active(S_j,B)* > 0.9).

Note that several modules fall outside all three categories (divergence score between 0.1 and 0.9). While such modules are still very relevant to the condition being studied, for these modules we do not make a call regarding conservation.

5.2.6 Matching modules through time

In order to identify cascades of activated modules, we generated a separate modules set for each time point using a search procedure that is similar to the one described above. We next tested the overlap of module sets between time points using a hypergeometric test. If reciprocal tests were found to be significant (*p*-value < 0.01), we defined these modules as matching. In many cases clear chains are identified throughout the time series indicating a module that is preserved through time. Nonetheless, usually in earlier time points where the overall activation of modules is lower there may be several modules that are matched to later time points creating a fan-in structure.

5.2.7 Cell and Bacterial Culture

Human monocyte-derived macrophages and murine bone marrow-derived macrophages were generated using standard protocols [218], [219]. All work involving *F. tularensis* was performed under biosafety level 3 conditions with approval from the Centers for Disease Control and Prevention Select Agent Program at the University of Pittsburgh. *F. tularensis* subsp. *tularensis* Schu S4 was obtained from the Biodefense and Emerging Infections Research Repository (Manassas, VA). Bacteria were grown on chocolate II agar for 1 to 3 days at 37°C and 5% CO_2. Muller Hinton broth supplemented with 0.1% glucose, 0.025% ferric pyrophosphate (Sigma), and IsoVitaleX (Becton Dickinson) was used for overnight liquid cultivation of Schu S4 prior to infection. Bacteria were pelleted and washed with tissue culture media prior to dilution to the appropriate multiplicity of infection (MOI). Eukaryotic cells were cocultured with an MOI of 10 bacteria for 24 hrs to generate supernatants to assess cytokine production. Eukaryotic cells were incubated for 2 hrs with an MOI of 500 bacteria then the cells were washed twice with Hanks Balanced Salt Solution (HBSS) and the media was replaced until well contents were harvested for protein analysis.

5.2.8 ELISA and Western Blot

Cytokines were measured by ELISA using murine CXCL11/ITAC (R&D Systems), murine CCL2/MCP-1 (eBioscience), human CXCL11/ITAC (R&D Systems) and human CCL2/MCP-1 (R&D Systems) kits. For Western blots, media was removed and the monolayer was washed with HSBB. Cells were lysed in RIPA Buffer (Cell Signaling) with 2% SDS, protease/phosphatase inhibitor cocktail (Cell Signaling), and 1mM PMSF. Samples were boiled for 5 min and sonicated before being used. Antibodies used in these studies include anti- NFκB p65 (C-20, Santa Cruz), anti-phospho-NFκB-p65 (93H1, Cell Signaling), and anti-rabbit IgG (A6154, Sigma). Proteins were separated on a 12% polyacrylamide gel and transferred to a PVDF membrane for blotting. Proteins were visualized using ECL chemiluminescence (Amersham Biosciences) and film. Pixel intensity was quantified using ImageJ.

5.3 Results

Gene association networks are assembled as discussed in Methods using various genomic data types from multiple species. We next search the resulting network for active sub-networks (modules) using a target function that combines node and edge scores (see Figure 5-1) using a greedy search approach that starts with high scoring nodes (based on expression data) as seeds to further expand the sub-network. Highly overlapping sub-networks are merged based on the quality of the overlap (see Methods).

5.3.1 Comparing mice and macaques infected by *Francisella tularensis* Schu S4

We applied *ModuleBlast* to study the response of alveolar macrophages (AM) from mice and cynomolgus macaques to *Francisella tularensis* Schu S4. *F. tularensis* causes a wide range of infections, but pneumonias of the lower respiratory tract are major concern for mortality and morbidity in the setting of an intentional release. AM cells are the first cells of the innate immune system to respond to invading pathogens. In previous experiments, AM cells were harvested from C57BL/6 mice and cynomolgus macaques by bronchoalveolar lavage (BAL) and were exposed to pathogens [201]. Cells were collected at 6 time points (0, 1, 2, 6, 12, and 24hrs) to determine changes in gene expression over time. Expression values for the infected arrays are calculated as the log2 difference from any time point to time point zero followed by a subtraction of a mock infection from the corresponding time series. For the initial module searches we set the score of each node as the most extreme value of the entire time series for any of the orthologous genes in each node (see below the changes for temporal module analysis). Orthology relations were defined using Inparanoid [17]. Interaction data for both species was downloaded from BioGRID [32].

An underlying assumption in our analysis is that paralogous genes are likely to show similar expression patterns and their measurements can be summarized by taking the most extreme value over the group indicating the maximum possible activation for all paralogs. Nonetheless, this assumption is not always realistic and paralogous genes may exhibit quite distinct behavior. One option to mitigate this problem is to repeat the analysis using the mean expression value over the group which should result in similar modules being identified. In order to understand how prevalent these cases are, we calculated statistics for the number of members in each orthogroups between mouse and macaque. 80.1% (11368 out of 14200 orthogroups) have only 2 members corresponding to a 1:1 orthology match. 16.5% (2337) have 3 members (corresponding to 1:2 or 2:1 matches). Only 0.3% of the orthogroups have 6 members or more, indicating that paralogs are not likely to cause a large shift in the results.

Using *ModuleBlast* we obtained 46 modules containing 443 unique nodes, out of which 22 modules were enriched for unique GO terms. These results highlight the ability of *ModuleBlast* to identify distinct mechanisms triggered by the infection in both species. 125 unique GO terms were identified for all modules, including modules that are enriched for chemotaxis, cytokine activity, TAP complex, apoptosis, and anti-apoptosis, all relevant to the strategies employed by *F. tularensis* upon infection and immune response of the cell (see further discussion below). Of the 433 nodes, 65% had a fold change difference of 2 or more in either of the species. 35% of the nodes are not differentially expressed, highlighting the importance of using both expression and interaction data. The modules contain 1233 unique edges (a 2.85 edges to nodes ratio) indicating a high connectivity in the resulting modules.

To test whether cross species analysis improves our ability to identify relevant modules, we compared the results for the combined mouse-macaque analysis with analyses that were conducted for each of the species separately using species-specific expression data and association networks. As can be seen in Table 5-2, cross species analysis leads to more modules, spanning a larger number of nodes and a significantly larger number of unique enriched GO terms. Importantly, none of the enriched modules that were identified using the individual species data are enriched with the TAP complex or the apoptotic / anti-apoptotic processes which play significant role in the *F. tularensis* infection (see below).

Table 5-2: Comparing the combined species analysis to species-specific analyses

Modules and nodes counts and enrichment information for *ModuleBlast* using a combined mouse and macaque network (first column), using only mouse expression data and network (second column), and macaque expression data and network (third column). Rows definitions: number of modules, number of modules that have at least one uniquely enriched GO term, number of nodes, number of activated nodes (fold change > 2), the total number of unique GO terms. In all categories the combined analysis improves upon species specific analyses.

	Combined analysis	**Mouse analysis**	**Macaque analysis**
# of modules	**46**	17	41
# of unique enriched modules	**22**	9	19
# of nodes	**433**	125	371
# of unique activated nodes	**282**	54	183
# of unique GO terms	**125**	30	51

5.3.2 Comparing ModuleBlast to other methods

We compared the results of *ModuleBlast* to *NeXus* [59] and *GXNA* [206]. *NeXus* is the only previous study that combined expression and interaction data across species. Unlike *ModuleBlast*, *NeXus* is only focused on conserved modules and may thus miss divergent modules. Unlike *ModuleBlast* and *NeXus*, *GXNA* was developed for the analysis of single species data and we included it in our comparison for completeness (once the networks are formulated, a single species method can be used as well). Since *NeXus*'s default settings are limited to mouse-human comparison based on an old Inparanoid format, we limited the dataset to include only genes that were found in the *NeXus* orthology information. The results of this comparison are summarized in Table 5-3. As can be seen, for this data *NeXus* identifies many more modules than *ModuleBlast* (54 vs. 15). However, many of these extra modules are highly overlapping and 80% of the nodes appear in more than one module. This may present a problem when searching for a few distinct modules for follow up experimental analysis. Indeed, while the number of modules obtained by *NeXus* is almost four times the number of modules identified by *ModuleBlast*, the total number of active nodes in *NeXus* modules is actually lower (67 vs. 106) and the GO statistics are very similar. One of the biggest downside of using *NeXus* is its runtime. Even for the limited dataset, *NeXus* requires 10 days to run whereas *ModuleBlast* terminates in few seconds on this dataset (all tests were performed on a standard dual core desktop workstation). *GXNA* did not perform well on this dataset. Using its default settings, *GXNA* resulted in 153 modules each with 15 genes. These modules are highly overlapping and thus difficult to use for practical follow up analysis. Moreover, permutation tests performed on *ModuleBlast* and *GXNA* showed a significantly decreased statistics for *ModuleBlast* and almost the same statistics for *GXNA*. An algorithm with high sensitivity and accuracy should detect more modules from the true network compared to the random one.

Table 5-3: Comparing *ModuleBlast* to *NeXus* and *GXNA* on limited data

See Table 5-2 legend for definition of information in rows. The comparison was conducted on a limited set of data to accommodate *NeXus* default orthology settings (see text for details). The best value in each category is marked in bold.

	ModuleBlast	***NeXus***	***GXNA***
# of modules	**15**	54	153
% unique enriched modules	**46.67%** (7/15)	16.67% (9/54)	17.65% (27/153)
% unique enriched KEGG modules	**40.00%** (6/15)	5.56% (3/54)	6.54% (10/153)
% of unique nodes	**100%** (138/138)	20.47% (164/801)	32.83% (743/2263)
% of unique active nodes	**76.81%** (106/138)	40.85% (67/164)	34.86% (259/743)
# of unique GO terms	29	32	**79**
# of unique KEGG pathways	**24**	7	19
Running time	**seconds**	10 days	**seconds**

5.3.3 Evaluating divergence and conservation

We next assessed each of the modules in the full dataset analysis to determine if they are conserved or divergent across the two species. In general, a single gene or interaction can be either conserved or not. Conservation becomes multifaceted when examining larger components. The three options for conservation and divergence we considered are: 1) conserved modules (CM), 2) divergent modules that are species-specific (SP), i.e., active in only one of the species, and 3) divergent modules that show opposite expression patterns in the two species (OP), e.g., up regulated in one and down regulated in the other (See Methods). In order to evaluate the convergence or divergence of the modules we calculated the sum of differences between the various species over all the nodes in the module and compared these differences to 10000 random modules with the same number of nodes (see Methods). In addition, we assessed the overall activation of the modules in each of the species using a similar randomization method. Out of the 46 modules, we classified 8 modules as highly conserved (CM), 5 modules as species specific (SP), and 7 modules as divergent in opposing patterns (OP).

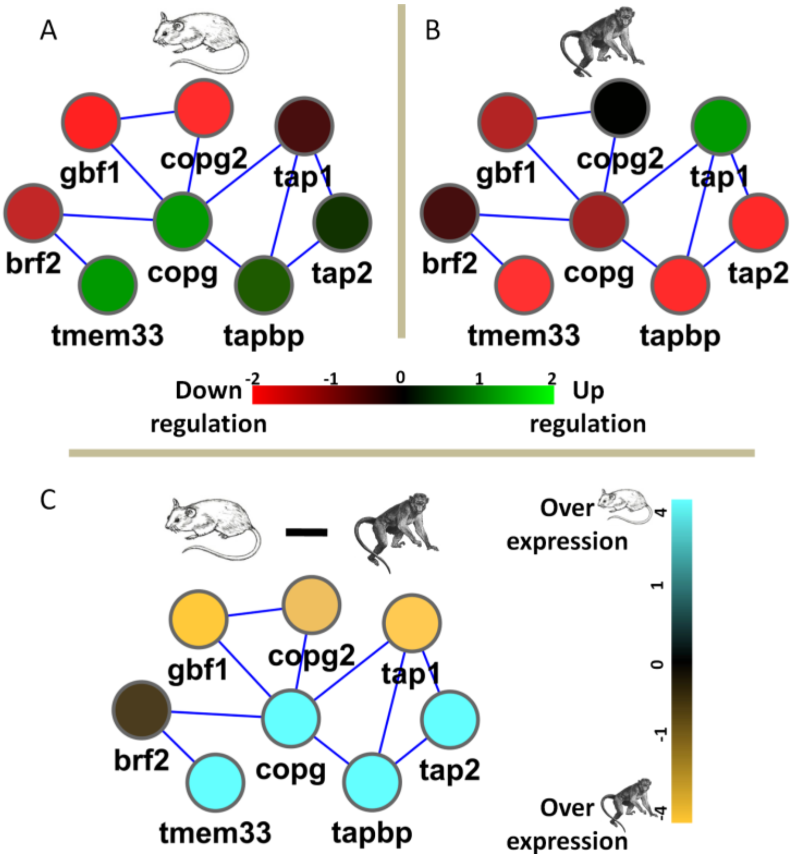


Figure 5-2: Capturing similarities and differences between the species

As modules are generated based on the absolute most extreme value in each orthogroup, each module has the potential of showing significant differences between the species. Module 34 is one of the divergent modules in opposing directions (OP) and involves the transfer of antigenic peptides (TAP) complex, which was previously shown to be transcriptionally active after *F. tularensis* infection in human cells. This module is depicted using A. changes in mouse expression values, B. changes in macaque expression values, and C. macaque expression values subtracted from the mouse expression values, highlighting similarities and differences between the species. Note that the species-difference scale is set to a minimum and maximum of 4-fold change compared to a 2-fold change for each species separately. Mouse gene names are noted.

GO enrichment analysis of the conserved modules indicated that they were mostly related to immune response and apoptosis suggesting that both mice and macaques activate similar pathways in response to *F. tularensis* infection. The SP and OP modules showed enrichment for various processes; one of the species-specific modules was enriched for a receptor complex that include IL6ST, a signal transducer shared by many cytokines, and OSMR, a member of type I cytokine receptor family that heterodimerizes with IL6ST to activate STAT3, and possibly STAT1 and STAT5. These transcription factors that were found to be enriched in an independent TF enrichment analysis for several modules including module 113 (see below). One of the divergent modules, module 34, was enriched for Transfer of Antigenic Peptides (TAP) complex (Figure 5-2). This complex is known to be involved in the transport of antigens from the cytoplasm to the ER for association with MHC class I molecules and TAP1 was previously shown to be transcriptionally active after *F. tularensis* infection of human cells [220], [221]. However, TAP1 was not modulated in murine macrophage-like cells infected with *F. tularensis* [220].

5.3.4 F. tularensis induces changes in apoptotic and anti-apoptotic gene expression, and is associated with NFκB activation

Module 113 (Figure 5-3a), is highly conserved across both species, and is highly enriched for apoptosis (1E-15), NFκB regulation (1E-13), anti-apoptosis (1E-5), and other apoptotic related terms. TF enrichment analysis for this module found RelA-p65, an active form of NFκB to be the most enriched TF regulating this module (1E-3). Given the conserved activation identified by

ModuleBlast in both species, we decided to further test NFκB-p65 involvement. Because the AM used to generate the microarray data were no longer available to us, we first verified that human monocyte-derived macrophages (hMDMs) and murine bone marrow-derived macrophages (mBMDMs) recapitulated responses seen previously with AM. We found that CXC11 and CCL2 chemokine gene expression were specific for the murine and macaque AMs, respectively. Protein levels measured by ELISA showed our alternative macrophage sources behaved similar to the AM: human macrophages produced CXC11 and murine macrophages made CCL2 (data not shown). We next measured the protein levels of phosphorylated RelA-p65 (phospho-p65) compared to total protein levels in macrophages 2hrs and 6 hrs following infection (See Methods).

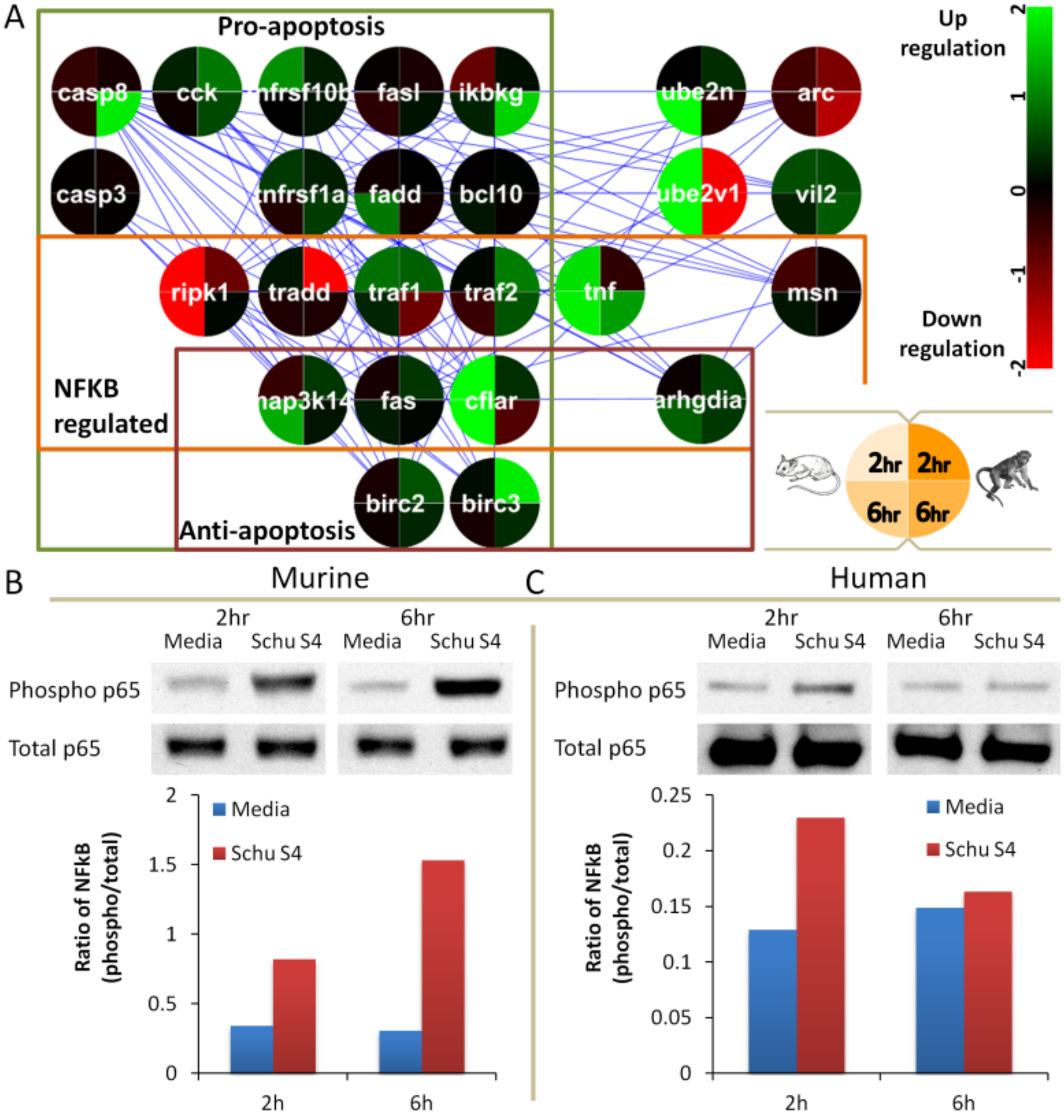


Figure 5-3: *F. tularensis* induces pro-and anti-apoptotic response through NFκB
Module 113 is significantly enriched with apoptotic and anti-apoptotic terms and is highly regulated by NFκB, an anti-apoptosis inducer. **(a)** Each node is colored according to the expression levels in mouse (left) and macaque (right) in 2hrs (up) and 6hrs (down) after infection. Pro and anti-apoptotic genes as well as NFκB regulated genes are highlighted. All the genes that are pro or anti-apoptotic except for CASP3, CASP8, CCK, BIRC2, and BIRC3 are part of the NFκB cascade (not shown). Mouse gene names are noted. **(b)** Western blot of phosphorylated and total NFκB-p65 from mBMDMs demonstrating an increased level of phosphorylated NFκB-p65 2hrs and 6hrs post infection with Schu S4. **(c)** Western blot of NFκB-p65 from hMDMs demonstrating an increased level of phosphorylated NFκB-p65 2hrs post infection that returns to baseline by 6hrs post infection.

The results indicate that the ratio of activated NFκB to total NFκB is higher 2hr following infection in human macrophages compared to 6hrs and to control (Figure 5-3b, c). In contrast, the level of phospho-p65 remained elevated in murine macrophages 6hrs after infection. Therefore, aberrant NFκB activation by *F. tularensis* coupled with differential activation of NFκB in murine versus primate cells could contribute to differences observed in Figure 5-3 and influencing apoptosis-related genes and possibly apoptosis.

In support of this, Module 113 contains transcript for a number of pro- and anti-apoptotic molecules including TNF, TRAF1, TRAF2, TRADD, BIRC2, and BIRC3. All of these show significant expression changes between 2hrs and 6hrs after F. tularensis infection in both species. NFκB is regulated by the heterodimeric TRAF1/2 complex that interacts with the inhibitor-of-apoptosis proteins (IAPs) and TRADD to mediate an anti-apoptotic signal from the TNF receptors. The TRAF1/2 complex also interacts with anti-apoptotic BIRC2 and BIRC3 E3-ubiquitin ligases, further supporting an interrelationship between NFκB regulation and apoptotic/anti-apoptotic pathways during infection. Independent evidence also supports this interrelationship. A related bacterium, Francisella novicida, was recently shown to block staurosporin-induced apoptosis in macrophages, which correlated with activation of nuclear transcription factor B (NFκB) [222].

5.3.5 Response progression over time

The above analysis was conducted by summarizing the entire time series for each gene into a single value. While useful for finding relevant functional modules, we sought to identify the entire cascade of events that occur following *F. tularensis* infection over time. We therefore constructed a module set for each time point separately and matched the resulting module sets in each time point to all other time points using a reciprocal hypergeometric test (See Methods). General trends through time show that the size of the modules and the number of enriched GO terms significantly increase as the response progresses.

In one example, module 124, in time point 24hr module set, is enriched with transcription regulation and is reciprocally significantly matched to modules in all earlier time points. Figure 5-4 plots the genes in module 124 in five columns based on the earliest time point a gene was part of a matched module. Each node is colored by the expression level in all time points in a counter-clockwise fashion. It is easy to see that in this module the number of genes and the overall activation of the genes (up or down regulation) increased over time. Note that this depiction focuses on genes that are part of mechanisms that are transcriptionally active 24 hours after infection, and may miss genes that peak in early time points and are not active after 24 hours. However, additional analyses can be performed to evaluate genes that peak earlier. Module 124 is enriched with important transcription regulators, including p53, Jun, RelA (NFκB p65), CREB binding protein, and histone deacetylases (HDAC) 1, 3, 4, and 5. TP53 plays role in apoptotic and anti-apoptotic processes, and its expression was increased in 6 hrs. RelA, a part of the NFκB complex, increased at 12 hrs and is a pro-inflammatory transcription factor that also triggers anti-apoptotic responses. HDACs, which are shown to be transcriptionally active in late time points can regulate the function of NFκB and TP53 [223], [224]. Pro and anti-apoptotic processes may play a significant role during *F. tularensis* infection (see above).

Module 124 also contains several genes that were found in matched modules in early time points. Specifically, AKT1 expression is elevated 1hr after exposure to bacteria, during the *F. tularensis* penetration, but not in later time points. This result is similar to previous observations that *F. tularensis* Schu S4 infection reduces AKT1 gene and protein expression, thereby reducing cytokine response and host defenses against infection (Butchar et al., 2008). Another gene that is active in early time points is androgen receptor (AR), a nuclear receptor that is regulated by TLR stimulation and IFN-γ in macrophages [227]. In addition, we found immune modules enriched with chemotaxis, cytokine activity, and chemokine activity. Specifically, for module 34 in the 24 hours module set, we observed early activation of chemokine ligands CCL20 and CCL8. CCL8 interacts with CCR10, a chemokine receptor that is activated at later time points when both are assigned to the same module. Taken together, these results indicate that the temporal matching method can identify relevant associations when integrating expression and interaction data.

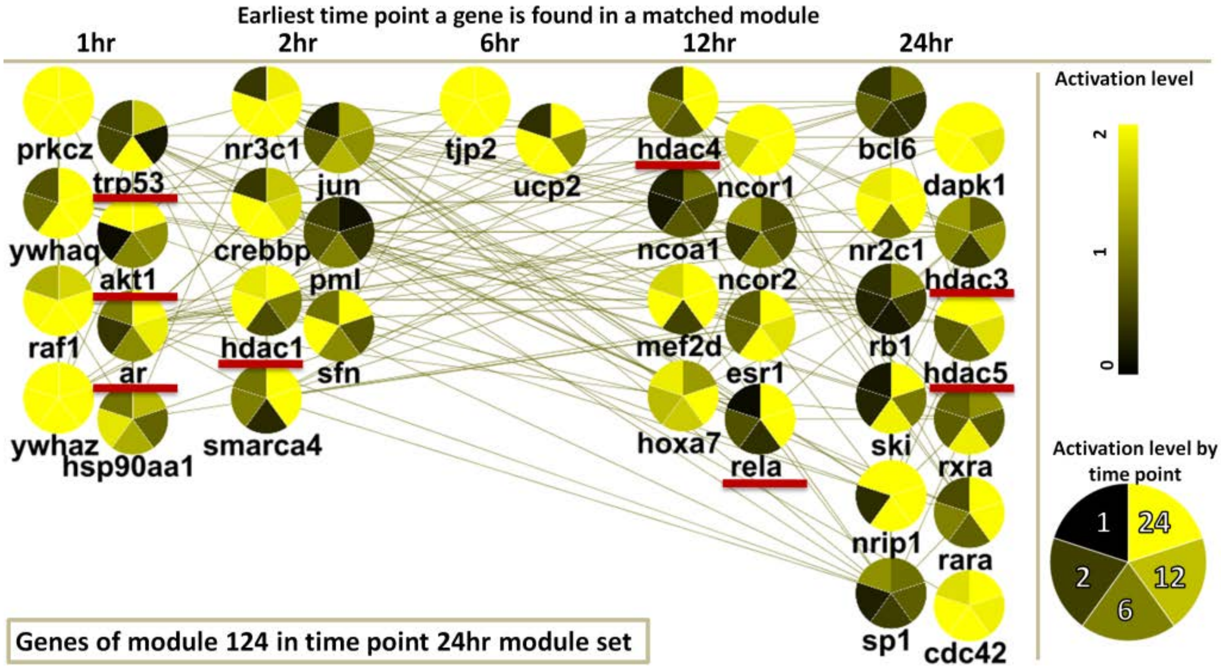


Figure 5-4: Response progression over time

Modules were created for each time point and were matched using reciprocal hypergeometric tests. Module 124, in time point 24 hrs module set, is shown in a layout that depicts the module expansion through time; genes that are part of matched modules in earlier time points are placed in columns from left to right based on the earliest time point that they were found to be part of a matched module. The node coloring is based on the overall activation (up or down regulation) of the orthogroups in either of the species in all time points in a counter-clockwise fashion. The number of nodes and their activation level increase through time. This module is enriched with transcription regulation genes, including TP53, relA, and AKT1 (underlined) that are involved with the *F. tularensis* infection. HDACs can reverse NFκB and TP53 inactivation. Mouse gene names are noted.

5.3.6 Web tool

ModuleBlast is offered as a web tool that allows users to find active modules within and across species. Please refer to Chapter 6 for details.

5.4 Discussion

We developed a novel method to find active sub-networks within and across species. *ModuleBlast* integrates expression data and gene interaction data from multiple species, and unlike many previous methods, directly utilizes both types of data in order to find highly relevant modules. *ModuleBlast* provides enhanced cross species capabilities by classifying the modules to several conservation types and identifying modules that show interesting activation patterns. Identifying conserved and divergent response patterns in the context of connected groups of genes is becoming increasingly important with applications for both basic science and drug discovery research by highlighting biological mechanisms that are likely to be affected similarly or differently to a specific drug or treatment.

We have applied *ModuleBlast* to a time series of expression data from mouse and macaque AMs infected with *F. tularensis* Schu S4 and found several modules with high relevance to the response progression over time and immune response mechanisms. One of the modules was enriched with apoptosis and anti-apoptosis genes that show significant expression changes between 2hrs and 6hrs post infection. These genes are associated with NFκB, an anti-apoptosis inducer, based on TF enrichment analysis and experimental evidence. In addition, performing the same analysis for each time point separately, and matching the resulting module sets across different time points, identified several modules that were consistently found in multiple time points. One of these modules was enriched with transcription regulators that are active at different time points and may play role in apoptosis modulation. Taken together, the summarized modules and the temporal matching method indicated that *F. tularensis* can promote and antagonize apoptosis at different stages of the infection and lead to concrete testable hypotheses that can elucidate the mechanisms governing the observed changes in gene expression.

While the analysis of expression data and integration with gene association information within and across species is very much in demand by the experimental community, there are only a few tools that can do such an analysis efficiently and obtain useful results. Many of the non-commercial tools are either not supported any longer, are not operating cross platform, or require deep technical knowledge for the configurations of file formats and parameters. Moreover, these tools do not support cross species analysis and often result in modules that are highly overlapping, hence less practical for follow up analyses. We have implemented *ModuleBlast* as an easy-to-use web tool with built-in support for various gene identifier names spaces, orthology information, and underlying networks. The web tool requires only gene identifiers and values to operate and offers extensive analysis options of the results (see Chapter 6 for further details). We hope that our tool will be a useful addition to the current set of analysis packages used by the experimental and computational communities.

6 Online Expression Analysis Package with Enhanced Support for Cross Species Analysis

Several methods targeting various aspects of expression data analysis were introduced in Chapter 3 (*ExpressionBlast*), Chapter 4 (*SoftClust*), and Chapter 5 (*ModuleBlast*). One of the main goals of this thesis is to support experimental biologists in their analysis of expression data, especially in the cross-species domain. Towards this goal we sought to provide a comprehensive expression analysis package that will encompass all the methods that were introduced in this thesis and will allow integration options between them. The expression analysis package is available at http://www.expression.cs.cmu.edu/.

6.1 Introduction

Surveying the current landscape of expression analysis tool revealed that there is a need for a comprehensive, reliable, and easy-to-use analysis package. Unfortunately, many of the previous tools are unusable due to one of the following reasons:

- They were retired and were no longer supported / available.
- They usually require software download and installation that in several cases is limited to specific operating systems. Cumbersome installations process often thwart experimental biologist from using these tools.

- Frequently they require installation of additional software frameworks and packages (e.g., R) that may be too complicated to use for experimental biologists. Moreover, in several cases they rely upon legacy versions of the software frameworks and packages, making the installation process even more challenging.
- They make use of old data.
- They have software bugs that prevent them from having proper operation.
- They assume strong hardware for proper operation and have long runtimes otherwise.
- They are not user friendly and may be too complicated to use for experimental biologists.

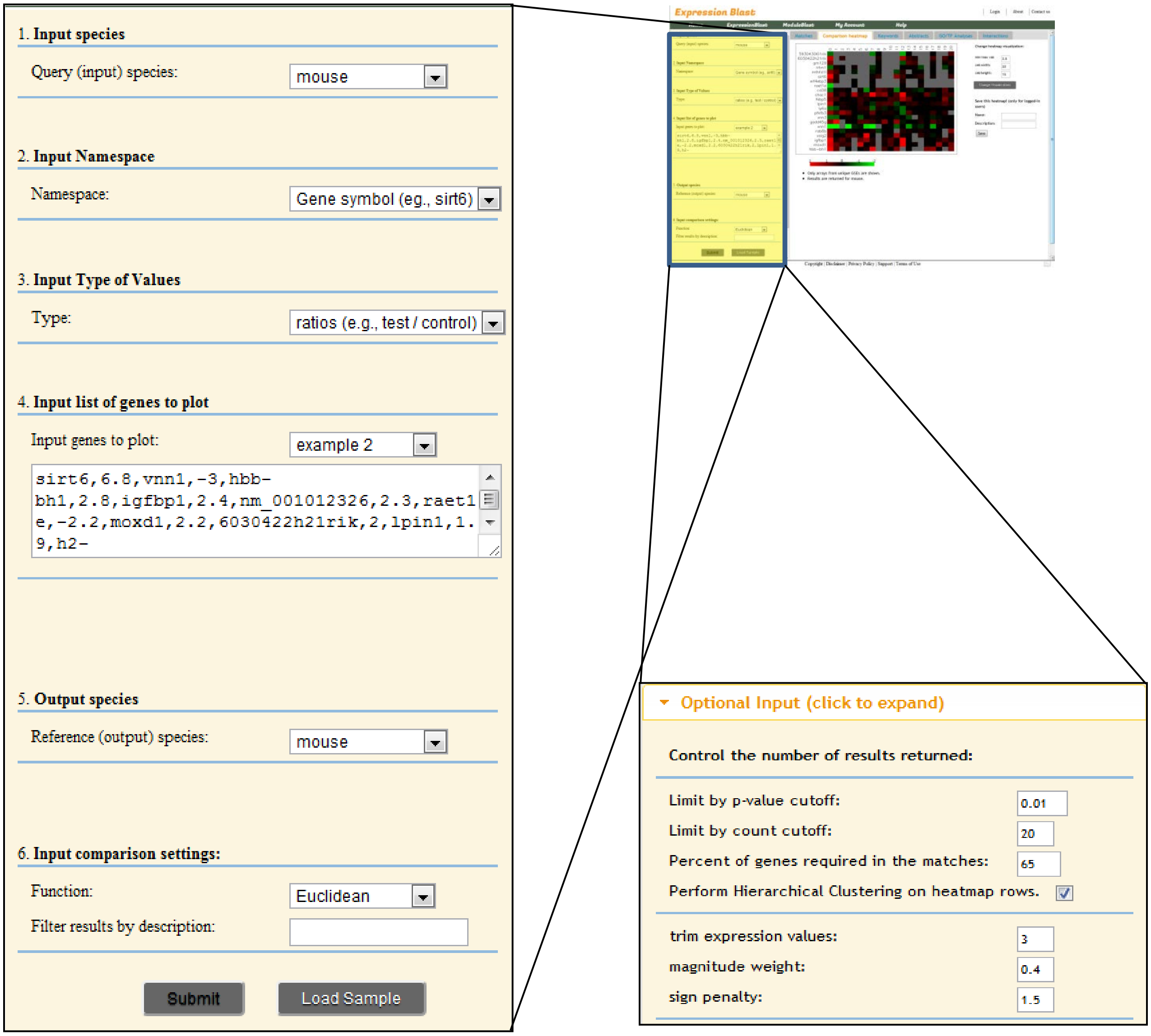
During the work on this thesis we have learned valuable lessons on how experimental biologists use computational tools and what do they need in terms of computational tools through collaborations with various experimental groups. The problems mentioned above together with the input obtained from our collaborators made us formulate the following guidelines for our approach to computational tools.

- Web-based: there is a clear advantage to web-based tools that are not reliant on any specific operating system, do not require installation, and can be updated easily making sure that users are always up-to-date with the latest release without the need for any further installations.
- Saving results: users need to be able to save their results together with the parameters used to achieve them for future references.
- Sharing: many studies today are being conducted in collaboration between several people and even between groups, that may be physically distant.
- Updated data: new data is constantly added to the major public repositories and this should be reflected in the online tools.
- Integrative: users often perform various analyses on the same dataset and need the ability to integrate results obtained from different tools together.

We hope that the tools that were developed in this thesis and are detailed below, together with the future development plans for the expression analysis package listed in Section 7.2, will create a real impact to aid expression analysis and cross species research by the entire experimental community.

6.2 *ExpressionBlast*

The web portal allows researchers to easily compare their own expression results to all publicly available expression data, even across multiple species by specifying a query (input) species and reference (output) species. See Figure 6-1 for input details. The queries run time depend on the requested output species and the size of the input gene set. Using the current server, typical runtimes are under 10 seconds for a query against mouse, and under 50 seconds for a query against human.



1. Input species

Query (input) species: mouse

2. Input Namespace

Namespace: Gene symbol (eg., sirt6)

3. Input Type of Values

Type: ratios (e.g., test / control)

4. Input list of genes to plot

Input genes to plot: example 2

```
sirt6,6.8,vnn1,-3,hbb-
bh1,2.8,igfbp1,2.4,nm_001012326,2.3,raet1
e,-2.2,moxd1,2.2,6030422h21rik,2,lpin1,1.
9,h2-
```

5. Output species

Reference (output) species: mouse

6. Input comparison settings:

Function: Euclidean

Filter results by description:

Optional Input (click to expand)

Control the number of results returned:

Limit by p-value cutoff: 0.01

Limit by count cutoff: 20

Percent of genes required in the matches: 65

Perform Hierarchical Clustering on heatmap rows. ☑

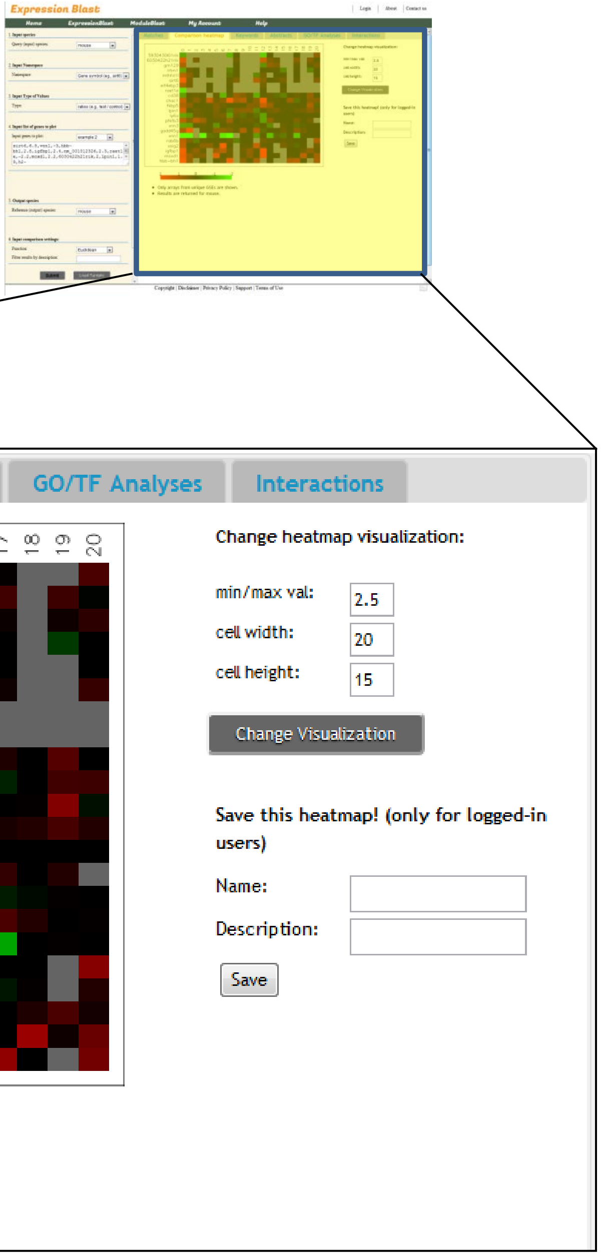
trim expression values: 3

magnitude weight: 0.4

sign penalty: 1.5

Figure 6-1: *ExpressionBlast* input form

The input panel (shown on left) contains the following mandatory inputs required for the query: 1. Name of the query (input) species. 2. Namespace or nomenclature schemes of the input genes. 3. Type of the input expression values (e.g., treatment/control ratio or the raw unclassified values). 4. Query set gene names and their corresponding values. 5. Name of the reference (output) species - the species for which database matches will be returned. 6. The distance function heuristic to be used for the query (See Search engine matching process for details). Another optional parameter is to filter the results by keywords that appear in the matches abstracts. The optional parameters (shown on bottom right) include options to control the number of matches that are returned (by p-value and/or count), limit for the minimal number of genes in the query set that should appear in the output, and various parameters that control the weighted Euclidean distance function. See Chapter 3 for how the input parameters are used.



Change heatmap visualization:

min/max val: 2.5

cell width: 20

cell height: 15

Save this heatmap! (only for logged-in users)

Name:

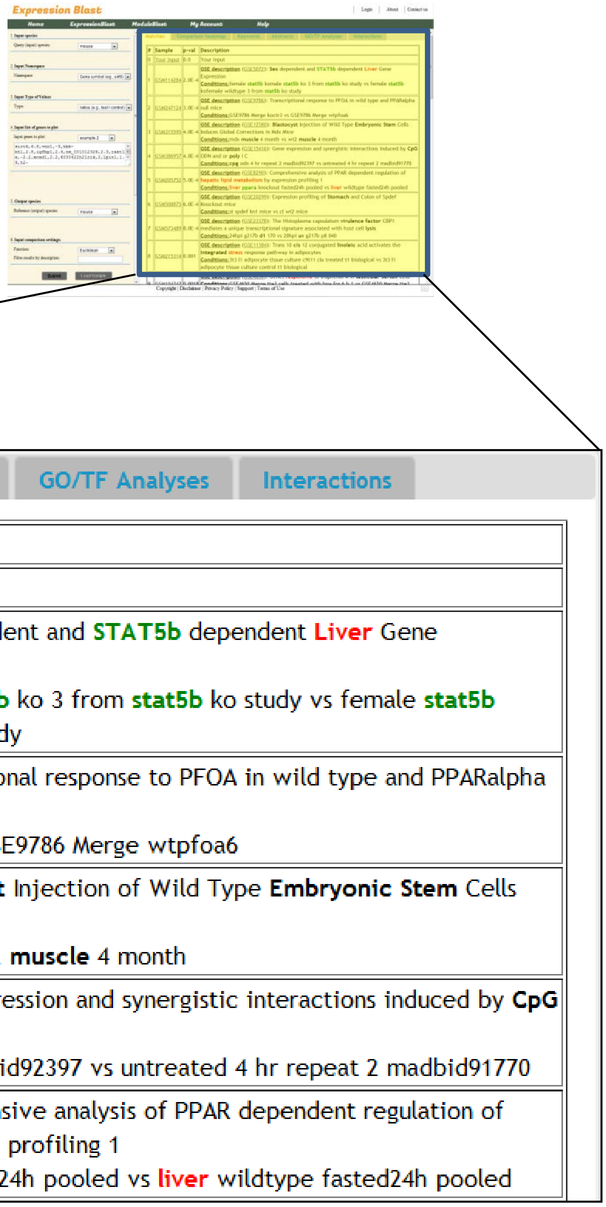
Description:

- Only arrays from unique GSEs are shown.
- Results are returned for mouse.

Figure 6-2: *ExpressionBlast* output form - comparison heatmap tab

The matched expression studies set is shown in a heatmap format with the query genes as rows and the matched expression experiments as columns where the first column depicts the user input values. The information on the columns is available in the Matches or Abstracts tabs. Options to control the heatmap visualization including color intensity and cell dimensions are available to the right of the heatmap.

The matched expression studies set is shown in a heatmap format with the query genes as rows and the matched expression experiments as columns where the first column depicts the input values (See Figure 6-2). The heatmap is complemented with short descriptions for each match and full abstracts taken from the GEO database or from Pubmed whenever Pubmed identifier is available for the experiment. In addition, the abstracts and GO terms enrichment of the matched output set are analyzed for keywords and terms. Over represented keywords, terms, and gene names are highlighted in the output (see Figure 6-3). See Chapter 3 for details.



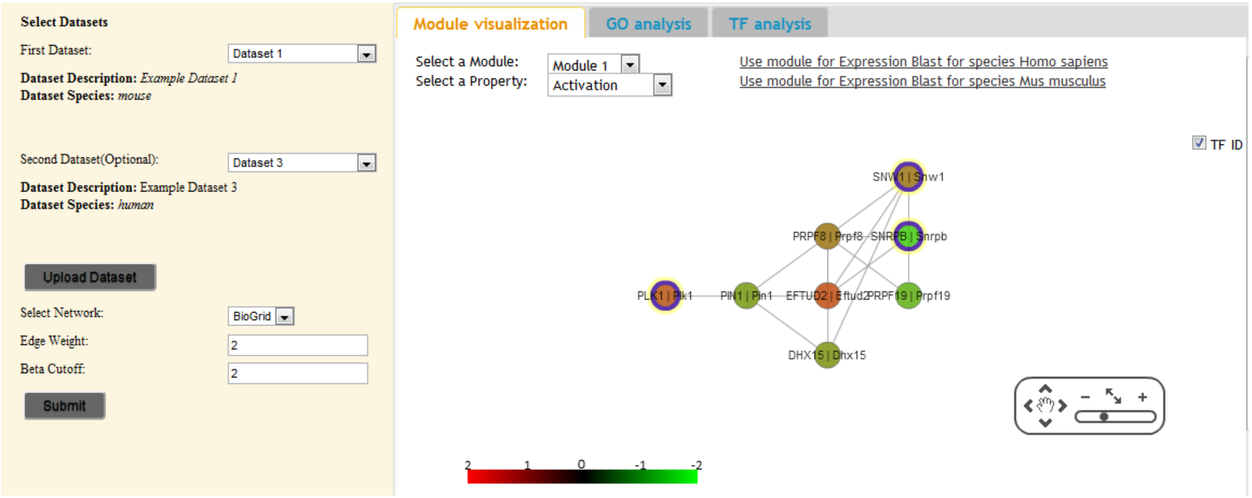
#	Sample	p-val	Description
0	Your input	0.0	Your input
1	GSM114284	2.0E-4	**GSE description** (GSE5072): **Sex** dependent and **STAT5b** dependent **Liver** Gene Expression **Conditions:**female **stat5b** komale **stat5b** ko 3 from **stat5b** ko study vs female **stat5b** kofemale wildtype 3 from **stat5b** ko study
2	GSM247124	3.0E-4	**GSE description** (GSE9786): Transcriptional response to PFOA in wild type and PPARalpha null mice **Conditions:**GSE9786 Merge koctr3 vs GSE9786 Merge wtpfoa6
3	GSM315595	4.0E-4	**GSE description** (GSE12580): **Blastocyst** Injection of Wild Type **Embryonic Stem** Cells Induces Global Corrections in Mdx Mice **Conditions:**mdx **muscle** 4 month vs wt2 **muscle** 4 month
4	GSM386957	4.0E-4	**GSE description** (GSE15416): Gene expression and synergistic interactions induced by **CpG** ODN and or **poly** I C **Conditions:cpg** odn 4 hr repeat 2 madbid92397 vs untreated 4 hr repeat 2 madbid91770
5	GSM205752	5.0E-4	**GSE description** (GSE8290): Comprehensive analysis of PPAR dependent regulation of **hepatic lipid metabolism** by expression profiling 1 **Conditions:liver ppara** knockout fasted24h pooled vs **liver** wildtype fasted24h pooled

Figure 6-3: *ExpressionBlast* output form - comparison matches tab

Short descriptions that are based on the GSM titles of the matched studies are shown to provide a short summary. Keywords that are present in the GO based keyword compendium are shown in bold. Enriched keywords are highlighted in red, and gene names are highlighted in green. A p-value based on randomization tests is provided for each match. See Chapter 3 for details.

6.3 *ModuleBlast*

ModuleBlast is offered as an integrated web tool in the analysis package that allows users to find active modules within and across species and provides an easy-to-use web interface with built-in support for several species, various gene identifier namespace inputs, orthology information, and underlying networks based on BioGRID [32] or STRING [228] (see Figure 6-4). Data input for one or two datasets (possibly from different species) is taken based on previously uploaded datasets (see Section 6.5) as *ModuleBlast* is reliant on having expression data for all possible gene entries. The output includes an overall analysis of the modules in terms of activation and divergence, and graphical representation for the modules using Cytoscape Web [229] (see Figure 6-4). GO enrichment and transcription factor analysis are noted for each module and allow dynamic highlighting of the annotated / regulated genes, by GO annotation or TF.



Select Datasets

First Dataset: Dataset 1

Dataset Description: *Example Dataset 1*
Dataset Species: *mouse*

Second Dataset(Optional): Dataset 3

Dataset Description: Example Dataset 3
Dataset Species: *human*

Select Network: BioGrid

Edge Weight: 2

Beta Cutoff: 2

Select a Module: Module 1
Select a Property: Activation

Use module for Expression Blast for species Homo sapiens
Use module for Expression Blast for species Mus musculus

☑ TF ID

Figure 6-4: ***ModuleBlast*** **input and output**

The input panel (on left) contains options to select up to two possible datasets that were pre-uploaded to the registered user account. The information on the selected datasets is shown below the dropdown selection. A button to upload a new dataset to the account, a dropdown to select the underlying protein interaction network, and required parameters are available.

The output panel - module visualization tab (on right) shows the interaction network for the chosen module (module selection as a dropdown at the top). Node colors are based on values that can be chosen out of four possible options in the two datasets query based on user selection in a dropdown control: overall activity (most extreme value in the orthogroup), divergence (subtracting the second dataset from the first dataset, set 1 activity, and set 2 activity. Nodes that are controlled by a specific transcription factor (TF ID for example), are highlighted with bold border. The selected module can be sent to *ExpressionBlast* query by clicking on one of the links at the top.

6.4 Integration between Tools

The ability to integrate analysis tools together creates an added value which is greater than the value of tools alone. Specifically, we found that allowing easy transitions between the 'functional groups tools' including *ModuleBlast* and *SoftClust* and the expression query tool *ExpressionBlast* can create valuable synergies.

6.4.1 Two step paradigm for experimental expression results validation

ModuleBlast → ExpressionBlast

In one example, we integrated *ModuleBlast* and *ExpressionBlast* to allow querying the groups identified by *ModuleBlast* to all other public expression experiments. This integration practically allows a two-step validation paradigm for novel expression analysis; the first step (*ModuleBlast*) can be considered as an internal validation of the experimental expression results to identify relevant groups of genes that are changing in response to the experiment at hand; the second step (*ExpressionBlast*) can be considered as an external validation of the experimental results to increase the confidence that similar results are achieved for similar conditions by other experimental labs.

ExpressionBlast* → *ModuleBlast

The other direction is also useful. One can first identify close or relevant studies to his own experimental results in the database using *ExpressionBlast* and then use *ModuleBlast* in order to find similar and divergent modules, pointing out to specific biological processes of interest.

6.5 Data management and Collaboration

This section briefly lists other functions provided by the expression analysis package to enhance the usability of the tools.

6.5.1 Users

Many functions (including storing data and results) require user registration. Users information is password protected and no one can access the user data without specific permissions. Users can also log in using their Google account. Registered users can also work in groups on the same Project in order to facilitate collaboration.

6.5.2 Projects

Projects are a collection of datasets and results that were uploaded by a user or were shared with him/her. Projects allow collaboration with different groups on distinct datasets.

6.5.3 Data upload

There are several formats of expression data that are accepted for upload. The uploaded data is also summarized to one column value for comparison in *ModuleBlast* or *ExpressionBlast* in case it contains several replicates or a time series. Replicates are merged using median value and time series are merged by taking the most extreme value over the series.

6.5.4 Saving results

Registered users can save the tools output and the input parameters that were used to generate the results. This also facilitates quick retrieval of the results without the need to perform the queries again.

6.5.5 Sharing results

Projects can be shared with other registered users by inviting them to collaborate on a project. There are two types of permissions that can be granted to collaborators: editor permissions and viewer permissions. Editors can add datasets and results to a project while viewers can only view existing datasets and results. Comments can be added to every dataset or result to facilitate the collaboration.

7 Conclusions and Future Work

7.1 Grand vision

Recent developments of new experimental techniques allow us to explore the genome, transcriptome, proteome, and interactome at scales that are much greater than those possible in the past. To fully utilize the volumes of data generated by these high throughput methods would require the development of novel computational methods and software tools. This is especially true for cross species studies, which by definition require the integration of high throughput data from several sources and pose new challenges on the analysis. Such studies are commonly performed, both in academia and by the pharmaceutical industry, to explore issues ranging from development to various responses and diseases. Yet, to date, very few computational tools specifically designed for cross species analysis of functional genomic data are available.

This thesis addresses this challenge by developing a number of methods and software tools which enable experimental biologists to analyze their high-throughput data within and across species. These methods target various types of high throughput data that are collected by biologists. For gene expression, which is one of the predominant experimental methods used to study condition specific responses, we developed *ExpressionBlast*. *ExpressionBlast* enables researchers to perform queries across thousands of studies and close to a million microarrays experiments that were deposited in public databases. We have built a unique automated annotation framework that integrates data from different platforms, labs, and species, and is accessible via an easy-to-use web interface. *ExpressionBlast* analysis provides validation for experimental results and can suggest concrete hypothesis and follow-up experiments by pointing to specific genes of interest. *ExpressionBlast* was already found to be effective in providing novel insights and research directions to study cancer and aging. We believe that *ExpressionBlast* will become a widely used tool for transcriptomic analysis.

In addition to ExpressionBlast, this thesis develops methods and easy-to-use web tools for other analyses routinely conducted by experimental researchers, including finding co-expressed clusters and identifying active sub-networks. For clustering, we developed constrained clustering methods that can utilize priors to improve grouping of genes across species. For the sub-network discovery and analysis, we expended previous approaches that worked only in single species settings so that can be used in a cross species setting. These tools were shown to identify core conserved and divergent modes of operation. These lead to specific experiments that elucidated how different organisms employ distinct biological processes in response to changes in the environment. We hope that these tools will aid biologists performing cross species analyses and will have an impact in elucidating molecular mechanisms in a variety of conditions and organisms.

7.2 Summary

This thesis aims to both highlight concepts relevant to cross species analysis and develop new algorithms and tools that are specifically designed for cross species analysis of high throughput data. We approached this problem from four related directions. Our first direction was to develop a framework that will lay the foundations of cross-species analysis of interactions and expression data. We then focused on methods for analyzing expression data within and across species. We developed a search engine we name *ExpressionBlast* for gene expression data than can take expression input in one species and find related studies in the same or a different species. We then focused on how to conduct a unified analysis on data from multiple species. We showed a method we name *SoftClust* that could perform gene expression clustering and use orthology information to reduce the inherent noise in expression data measurements to get more meaningful biological results. We also showed a method we name *ModuleBlast* that integrates static interaction data and condition-specific data from multiple species to find connected groups of genes that show high differential expression in one or more species and identifies modules that are conserved or divergent between the species. All these tools were integrated to an easy-to-use web interface that could facilitate cross species analysis of expression data across species for the benefit of the entire scientific community. In Section 7.1 of this chapter we summarize the contributions presented in this thesis and some conclusions reached. We end this chapter with a discussion of possible future extensions of the work presented here (Section 7.2). Specifically, a holistic approach for the future development of the expression analysis package is being proposed. We also discuss what it would take in order to incorporate next generation RNA-sequencing data into the expression analysis package.

7.3 Conclusions

This thesis explored various aspects of the use of high throughput data for comparing multiple species. Many researchers are now performing identical experiments in more than one species trying to contrast the responses in the different species and to understand the molecular mechanisms driving these responses. This is especially true for expression experiments that allow the investigation of dynamic and condition-specific responses, at lower costs than ever before. As a result, the number of expression experiments is growing exponentially. While the amounts of expression and interactions data are growing fast, thus far, only few algorithms and tools were specifically designed for cross species analysis. As mentioned throughout this text, cross species analysis poses many problems including differences in the quality and coverage of the different data types between species, differences in the measurements and the experimental methods, and issues related to orthology assignment. One should also remember that in the current time point, the coverage for some species and data types is still low and requires careful attention to make sure that the conclusions that are drawn are robust and will hold when more data becomes available.

Chapter 2 of this study directly addresses this reservation and shows that while simplistically the conservation rates of high throughput data were measured to be low, on average, interactions that are part of the same functional unit show a much higher degree of conservation than previously reported. These results paved the way to using high throughput interactions data in cross species comparison studies and combining multiple data types together to increase the confidence in the results. The tools developed in Chapters 3, 4, and 5 are based on this similarity of interactions and expression data within the same functional context.

The gene expression search engine, *ExpressionBlast*, presented in Chapter 3 allows to directly compare expression data collected in one species to all other previous expression studies conducted in the same or in a different species. This may lead to new targeted hypotheses that may explain the investigated mechanism. The cross species application of *ExpressionBlast* to our recent study on lifespan extension in male mice was indeed able to lead us to a new hypothesis for why the life extension is gender-specific. If these hypotheses prove to be correct, it may open new treatment possibilities for human to increase the right form of estrogen levels in specific tissues and to the specific levels, rather than the whole body single-dose quantity offered today to females in their menopause period. *ExpressionBlast* thus opens new possibilities for drug discovery where researchers can quickly examine whether a drug being developed in lower mammals may have been shown to have the desired effect in human.

Chapter 4 expands on comparing expression data across species by introducing a new tool, we call *SoftClust*, that allows prior knowledge to be incorporated into the popular and widely used clustering method, *k*-means. This has two main advantages 1) Overcoming some of the experimental error that might lead to orthologs not being co-clustered in cases of ambiguous cluster boundaries. The better assignment of orthologs in clusters can lead to more biologically meaningful clusters. 2) Increasing the confidence that if two orthologs are not co-clustered despite the reward than they are indeed divergent. Analyzing such divergent orthologous groups between clusters, revealed significant divergence among regulatory programs associated with fluconazole sensitivity in three diverse yeast species. In the future, such approaches might be used to survey a wider range of species, drug concentrations, and stimuli to further discover conserved and divergent molecular response pathways.

Chapter 5 further expands on this concept, and rather than relying only on expression data to find groups of interesting genes it introduces a new framework we call *ModuleBlast*, that combines both static interaction data and condition-specific expression data to understand conservation and divergence of biological systems dynamics. This is done by searching for active sub-networks, groups of connected genes that show high differential expression, and examining the conservation patterns of each module. This study also expands on previous similar studies in the single species domain by using both the differential expression and the connectivity to guide the search for modules. In addition, it shows how time series data can be incorporated to gain further insights. Exploring modules obtained for mice and monkeys infected by *F. tularensis* led to hypotheses for how apoptotic and anti-apoptotic mechanisms are employed by *F. tularensis* during the infection. Ultimately, we hope that these types of studies, inspired by conserved and divergent modules will lead to the design of novel therapeutic agents.

We note that there are limitations to the set of tools we introduce in this thesis. In many cases there are differences between species that cannot be measured correctly, are difficult to account for, and may affect the obtained results. Moreover, proteins can be in several states within the cell and these states are in most cases not distinguished by the high throughput experimental methods. Lastly, as we have observed throughout this thesis, there are big differences in the coverage for various data types from different species. The results obtained by the different tools may change as more data becomes available for more species. For example, more studies can be matched to a specific query in *ExpressionBlast* and the cluster and module assignment discovered by *SoftClust* and *ModuleBlast* may change. However, we do not expect to see changes in the results that were experimentally validated.

A large effort was put in this thesis to make the different tools that were developed publicly available for the benefit of the experimental community. Easy-to-use user-oriented web interfaces were developed, which also allows integration between the tools, as described in Chapter 6. We hope that these tools will aid biologists performing cross species analysis and will have an impact in elucidating molecular mechanisms in a variety of organisms.

7.4 Robustness of the results

Throughout this thesis, we were faced with variable and noisy data. Whenever applicable we aimed to use repeats in the data to increase the confidence in the measured values. Specifically, a major emphasis in *ExpressionBlast* is on correctly identifying and merging technical replicates. In general, when dealing with expression data it is important to normalize the data properly. In addition, if several measurements are available for a gene over a time course, it is possible to identify spikes, i.e., a significantly large increases or decreases in a time point compared to the two adjacent time points. These spikes can be corrected by taking the mean measurement of all three time points. For interactions data we aimed to combine data from multiple sources with the assumption that interactions that were observed in multiple sources, data types, or species are more reliable. Interactions from different sources were combined based on a log likelihood score (LLS) scheme, originally described in [83]. The idea is that each interaction is assigned a score based on the likelihood of the two connected genes to participate in the same biological process. As the databases that collect interactions information are not independent, and may list interactions measured in the same experiment, we assigned the maximal score for an interaction if it was observed more than once.

While all the tools we presented in this thesis analyze sets of genes (and so are less likely to be affected by individual incorrect measurements), there are several evaluation methods we used to assess the effect of individual measurements on the overall conclusions drawn by our analyses. One common possibility was applying a permutation test, in which algorithms are applied to discover modules on random data that is based on same distribution. An algorithm with high sensitivity and accuracy should detect more modules from the true network than the random ones. There are two randomization methods that we applied; edge switching and node shuffling. One of the services that should be offered by the proposed framework is an API access to randomized networks that can be used by developers to evaluate their results.

A second possibility to evaluate the robustness of the results is cross validation. The idea is that holding out a small part of the data should not affect much the final outcome. For example, in Chapter 2 we randomly removed increasing proportions of S. cerevisiae interaction network while showing that the conclusions still hold. A cross validation approach can also be applied for assessing the *ExpressionBlast* results. A very small proportion of the genes can be omitted from the comparison but the top matches should still be present. A similar tactic can be employed in *ModuleBlast* or *SoftClust*, where the overall modules membership and ontology enrichment should be kept.

A third possibility for assessing the robustness of the results is repeating the analysis several times with slightly different configurations. As mentioned before, minimal variations to the parameters should create only a minor effect on the results. One option is to add a programmatic definition for each parameter of the algorithm that will state a range of values or changepercentages that can be sent to each tool to test the robustness of the application under different configurations.

7.5 Future Work

A number of future research directions extending the work presented in this thesis are possible in order to support further analysis of gene expression data and integration of other high throughput datasets, in the cross species domain in particular. In this section we discuss some of these using a suggested framework shown in Figure 7-1. The main idea is to build a framework that is flexible in incorporating new tools, offering them standard services in the expression analysis domain on multiple levels. This requires standard API for methods to utilize various high throughput data types (Section 7.2.1), defining a standard output structures (Section 7.2.2), and offering standard display components (7.2.3). This will allow researchers to develop new analysis tools quickly without the need to worry about how to get, access, and update the data. The researchers will also be able to use standard evaluation services (e.g., GO and transcription factor analysis) to quickly evaluate their new methods, compare their results to other similar methods using standard benchmarks, and use standard visualization component (e.g., heatmap or interactions display) hence reducing development time and quickly reaching a large user base that could test their method. Specific discussion on how to support next generation RNA-sequencing technologies is mentioned in sections 7.2.4.

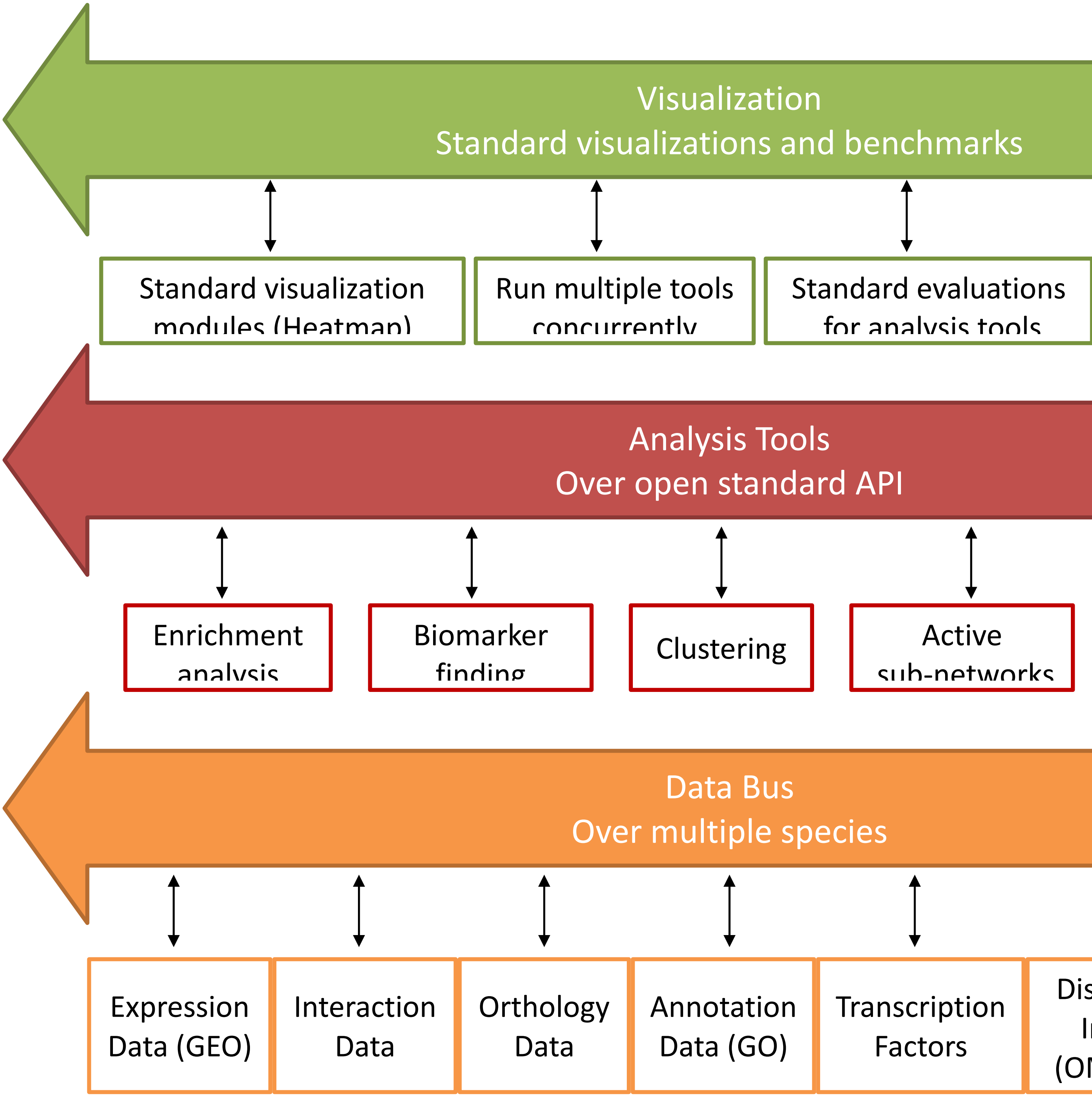


Figure 7-1: Framework for integrated and collaborative expression analysis and visualization

A schematic view of a future framework for the *Expression Blast* website that supports integration of additional analysis and visualization tools. The Scheme is based on three layers: Data Bus, Analysis Tools, and Visualization tools (bottom to top). See text for details.

7.5.1 Data Layer

One theme that was clear throughout this thesis is that integration of multiple data sources can improve the reliability of the predictions and lead to new insights and research directions. The work done in this thesis relied upon various data sources including a parsed, better annotated version of the largest expression database (GEO), interactions data from multiple species, orthology relations between numerous species, annotation data for multiple species (GO, KEGG), and transcription factor and miRNA information. All this data is stored locally in database tables that can be accessed through well defined Java access objects that support easy transition of gene names and annotations between species. The purpose of the Data Bus component in the framework (see Figure 7-1) is to allow easy utilization of the data by various analysis tools. Additional information is planned to be added soon including Disease information (e.g., OMIM [230] or the Comparative toxicogenomics database [231]), drug-target relationships (e.g., Drugbank [232] or BindingDB [233]), phenotypic information [234], [235], and other types of data.

7.5.2 Application layer

The analysis tools layer of the framework (see Figure 7-1) is intended to provide an easy integration of analysis tools, ideally with enhanced support for cross species analysis. This will allow researchers and developers to focus on the algorithmic part and utilize standard components for data access and visualization as well as standard services including GO annotation enrichment analysis, TF / miRNA enrichment analysis, and others. Several components are already implemented that utilize the framework including *ExpressionBlast* (Chapter 3), *SoftClust* (Chapter 4), and *ModuleBlast* (Chapter 5). Additional analysis tools (e.g., other clustering methods, biomarker finding, and disease / drug target finding) can be easily incorporated as long as they implement the necessary API. This can be used as an open source system for many algorithm developers that want to quickly try out their algorithms over various datasets and take advantage of the a large user base to test them. The integration will be carried out in two phases. New analysis tools will first be incorporated to a development version of the web-framework and be announced only to early adopters. After administrator approval the tools can be incorporated to the official version.

7.5.3 Visualization layer

The visualization layer of the framework (see Figure 7-1) is intended to provide standard user interface components for the analysis tools (e.g., a heatmap display, interactions display, GO and transcription factors enrichment analysis display, and others). This layer also has the ability to combine different tools together, to enhance their added values. For example, active networks that were identified by *ModuleBlast* or clusters identified by *SoftClust* can be easily directed to be queried in *ExpressionBlast* (see Section 6.4 for details). This layer can also provide standard benchmarks to various algorithm developers to evaluate their methods against other algorithms. Lastly, the user interface can allow for multiple tools to be run concurrently, hence increasing the confidence in the predictions.

7.5.4 Support for RNA-sequencing data

Processing RNA-sequencing (in short, RNA-seq) data pose new challenges in terms of the need to align the reads to a reference genome or transcriptome, summarize the expression of the mapped reads (tables of counts showing how many reads are in a coding region), and normalize the data in new ways. The size of the output is also significantly larger (~3000 times larger than the output produced by a microarray experiment) and requires significantly more computational power to analyze. There are several tools developed so far to handle the different steps in the RNA-seq process [236]. These tools employ intrinsically different approaches for each step, require many input parameters, and produce quite distinct results one from another. There are still no gold standards for processing RNA-seq data making it hard to choose the right method to use and to find an automatic way for adjusting the input parameters for each sequencing experiment. There are already few attempts to build complete pipelines from the raw sequence files to the transcripts measurements including ArrayExpressHTS [237] and Myrna [238]. In addition, there already several repositories that try to aggregate RNA-seq experiments including the RNA-Seq Atlas [239] that collected studies from several tissues from eleven human donors, ReCount [240] that contains data from 18 different published studies, and RecountDB [241] that stores batch data on transcripts for 45 organisms.

As the number of RNA-seq experiments is accumulating quickly, an effort should be made to integrate this type of data as part of the expression analysis package, by either creating a batch process to analyze RNA-seq data uploaded to GEO using a pipeline of the type of

ArrayExpressHTS [237] and Myrna [238] or by incorporating the data in the already analyzed transcript repositories.

Bibliography

[1] T. Dobzhansky, “Nothing in biology makes sense except in the light of evolution,” *The american biology teacher*, vol. 35, pp. 125-129, 1973.

[2] L. J. Jensen, T. S. Jensen, U. de Lichtenberg, S. Brunak, and P. Bork, “Co-evolution of transcriptional and post-translational cell-cycle regulation.,” *Nature*, vol. 443, no. 7111, pp. 594-7, Oct. 2006.

[3] T. Nurnberger, F. Brunner, B. Kemmerling, and L. Piater, “Innate immunity in plants and animals: striking similarities and obvious differences,” *Immunological Reviews*, vol. 198, no. 1, pp. 249-266, Apr. 2004.

[4] D. Simmons, “Genetic inequality: Human genetic engineering,” *Nature Education*, no. 1, 2008.

[5] F. Foury, “Human genetic diseases: a cross-talk between man and yeast.,” *Gene*, vol. 195, no. 1, pp. 1-10, Aug. 1997.

[6] K. W. Y. Yuen, C. D. Warren, O. Chen, T. Kwok, P. Hieter, and F. A. Spencer, “Systematic genome instability screens in yeast and their potential relevance to cancer,” vol. 104, no. 10, pp. 3925-3930, 2007.

[7] I. K. Hariharan, D. Ph, and D. A. Haber, “occasional notes Yeast , Flies , Worms , and Fish in the Study of Human Disease,” *October*, pp. 2457-2463, 2003.

[8] J. Bilen and N. M. Bonini, “Drosophila as a model for human neurodegenerative disease.,” *Annual review of genetics*, vol. 39, pp. 153-71, Jan. 2005.

[9] C. J. Potter, “Drosophila in cancer,” *Genetic Analysis*, vol. 9525, no. 1996, 2000.

[10] L. Fontana, L. Partridge, and V. D. Longo, “Extending healthy life span--from yeast to humans.,” *Science (New York, N.Y.)*, vol. 328, no. 5976, pp. 321-6, Apr. 2010.

[11] S. F. Altschul et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,” *Nucleic acids research*, vol. 25, no. 17, pp. 3389-402, Sep. 1997.

[12] Y. Lu, P. Huggins, and Z. Bar-Joseph, “Cross species analysis of microarray expression data.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 12, pp. 1476-83, Jun. 2009.

[13] M. N. Price, P. S. Dehal, and A. P. Arkin, “Orthologous transcription factors in bacteria have different functions and regulate different genes.,” *PLoS computational biology*, vol. 3, no. 9, pp. 1739-50, Sep. 2007.

[14] K. L. McGary, T. J. Park, J. O. Woods, H. J. Cha, J. B. Wallingford, and E. M. Marcotte, “Systematic discovery of nonobvious human disease models through orthologous phenotypes.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 14, pp. 6544-9, Apr. 2010.

[15] D. L. Fulton, Y. Y. Li, M. R. Laird, B. G. S. Horsman, F. M. Roche, and F. S. L. Brinkman, “Improving the specificity of high-throughput ortholog prediction.,” *BMC bioinformatics*, vol. 7, p. 270, Jan. 2006.

[16] K. P. O’Brien, M. Remm, and E. L. L. Sonnhammer, “Inparanoid: a comprehensive database of eukaryotic orthologs.,” *Nucleic acids research*, vol. 33, no. Database issue, pp. D476-80, Jan. 2005.

[17] G. Ostlund et al., “InParanoid 7: new algorithms and tools for eukaryotic orthology analysis.,” *Nucleic acids research*, vol. 38, no. Database issue, pp. D196-203, Jan. 2010.

[18] Roman L. Tatusov, Eugene V. Koonin, and David J. Lipman, “A Genomic Perspective on Protein Families,” *Science*, vol. 278, no. 5338, pp. 631-637, Oct. 1997.

[19] S. Powell et al., “eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges.,” *Nucleic acids research*, vol. 40, no. Database issue, pp. D284-9, Jan. 2012.

[20] R. M. Waterhouse, E. M. Zdobnov, F. Tegenfeldt, J. Li, and E. V. Kriventseva, “OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011.,” *Nucleic acids research*, vol. 39, no. Database issue, pp. D283-8, Jan. 2011.

[21] N. Yosef, R. Sharan, and W. S. Noble, “Improved network-based identification of protein orthologs.,” *Bioinformatics (Oxford, England)*, vol. 24, no. 16, pp. i200-6, Aug. 2008.

[22] T. Barrett et al., “NCBI GEO: mining tens of millions of expression profiles--database and tools update.,” *Nucleic acids research*, Jan-2007. .

[23] Y. Haudry et al., “4DXpress: a database for cross-species expression pattern comparisons.,” *Nucleic acids research*, vol. 36, no. Database issue, pp. D847-53, Jan. 2008.

[24] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics.,” *Nature reviews. Genetics*, vol. 10, no. 1, pp. 57-63, Jan. 2009.

[25] A. R. Borneman et al., “Divergence of transcription factor binding sites across related yeast species.,” *Science (New York, N.Y.)*, vol. 317, no. 5839, pp. 815-9, 2007.

[26] M. D. Wilson et al., “Species-specific transcription in mice carrying human chromosome 21.,” *Science (New York, N.Y.)*, vol. 322, no. 5900, pp. 434-8, Oct. 2008.

[27] P. J. Park, “ChIP-seq: advantages and challenges of a maturing technology.,” *Nature reviews. Genetics*, vol. 10, no. 10, pp. 669-80, Oct. 2009.

[28] T. Berggård, S. Linse, and P. James, “Methods for the detection and analysis of protein-protein interactions.,” *Proteomics*, vol. 7, no. 16, pp. 2833-42, Aug. 2007.

[29] N. J. Krogan et al., “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae.,” *Nature*, vol. 440, no. 7084, pp. 637-43, 2006.

[30] A.-C. Gavin et al., “Proteome survey reveals modularity of the yeast cell machinery.,” *Nature*, vol. 440, no. 7084, pp. 631-6, 2006.

[31] B. Lehne and T. Schlitt, “Protein-protein interaction databases: keeping up with growing interactomes.,” *Human genomics*, vol. 3, no. 3, pp. 291-7, Apr. 2009.

[32] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “BioGRID: a general repository for interaction datasets.,” *Nucleic acids research*, vol. 34, no. Database issue, pp. D535-9, Jan. 2006.

[33] A. Chatr-aryamontri et al., “MINT: the Molecular INTeraction database,” *Nucleic Acids Research*, vol. 35, no. Database, p. D572-D574, Jan. 2007.

[34] S. Kerrien et al., “IntAct--open source resource for molecular interaction data.,” *Nucleic acids research*, vol. 35, no. Database issue, pp. D561-5, Jan. 2007.

[35] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, “DIP: the database of interacting proteins.,” *Nucleic acids research*, vol. 28, no. 1, pp. 289-91, Jan. 2000.

[36] G. D. Bader, I. Donaldson, C. Wolting, B. F. F. Ouellette, T. Pawson, and C. W. V. Hogue, “BIND--The Biomolecular Interaction Network Database,” *Nucleic Acids Research*, vol. 29, no. 1, pp. 242-245, Jan. 2001.

[37] S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman, “A strategy for extracting and analyzing large-scale quantitative epistatic interaction data.,” *Genome biology*, vol. 7, no. 7, p. R63, Jan. 2006.

[38] A. Roguev et al., “Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast.,” *Science (New York, N.Y.)*, vol. 322, no. 5900, pp. 405-10, Oct. 2008.

[39] B. S. Srinivasan, N. H. Shah, J. a Flannick, E. Abeliuk, A. F. Novak, and S. Batzoglou, “Current progress in network research: toward reference networks for key model organisms.,” *Briefings in bioinformatics*, vol. 8, no. 5, pp. 318-32, Sep. 2007.

[40] M. G. Kann, “Protein interactions and disease: computational approaches to uncover the etiology of diseases.,” *Briefings in bioinformatics*, vol. 8, no. 5, pp. 333-46, Sep. 2007.

[41] T. Ideker and R. Sharan, “Protein networks in disease.,” *Genome research*, vol. 18, no. 4, pp. 644-52, Apr. 2008.

[42] A. L. Hopkins, “Network pharmacology: the next paradigm in drug discovery.,” *Nature chemical biology*, vol. 4, no. 11, pp. 682-90, Nov. 2008.

[43] A. F. Fliri, W. T. Loging, and R. a Volkmann, “Drug effects viewed from a signal transduction network perspective.,” *Journal of medicinal chemistry*, vol. 52, no. 24, pp. 8038-46, Dec. 2009.

[44] A. F. Fliri, W. T. Loging, and R. a Volkmann, “Cause-effect relationships in medicine: a protein network perspective.,” *Trends in pharmacological sciences*, vol. 31, no. 11, pp. 547-55, Nov. 2010.

[45] S. Navlakha and C. Kingsford, “The power of protein interaction networks for associating genes with diseases.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 8, pp. 1057-63, Apr. 2010.

[46] Y. Lu et al., “Cross-species comparison of orthologous gene expression in human bladder cancer and carcinogen-induced rodent models.,” *American journal of translational research*, vol. 3, no. 1, pp. 8-27, Jan. 2010.

[47] S. Bergmann, J. Ihmels, and N. Barkai, “Similarities and differences in genome-wide expression data of six organisms.,” *PLoS biology*, vol. 2, no. 1, p. E9, Jan. 2004.

[48] I. Tirosh, A. Weinberger, M. Carmi, and N. Barkai, “A genetic signature of interspecies variations in gene expression.,” *Nature genetics*, vol. 38, no. 7, pp. 830-4, Jul. 2006.

[49] D. Kuo, K. Tan, G. Zinman, T. Ravasi, Z. Bar-Joseph, and T. Ideker, “Evolutionary divergence in the fungal response to fluconazole revealed by soft clustering.,” *Genome biology*, vol. 11, no. 7, p. R77, Jul. 2010.

[50] J. Cai et al., “Modeling co-expression across species for complex traits: insights to the difference of human and mouse embryonic stem cells.,” *PLoS computational biology*, vol. 6, no. 3, p. e1000707, Mar. 2010.

[51] P. Zarrineh, A. C. Fierro, A. Sánchez-Rodríguez, B. De Moor, K. Engelen, and K. Marchal, “COMODO: an adaptive coclustering strategy to identify conserved coexpression modules between organisms.,” *Nucleic acids research*, vol. 39, no. 7, p. e41, Apr. 2011.

[52] Y. Lu, R. Rosenfeld, and Z. Bar-Joseph, “Identifying cycling genes by combining sequence homology and expression data.,” *Bioinformatics (Oxford, England)*, vol. 22, no. 14, pp. e314-22, Jul. 2006.

[53] Y. Lu, R. Rosenfeld, G. J. Nau, and Z. Bar-Joseph, “Cross species expression analysis of innate immune response.,” *Journal of computational biology : a journal of computational molecular cell biology*, vol. 17, no. 3, pp. 253-68, Mar. 2010.

[54] Y. Liu, A. Niculescu-Mizil, A. Lozano, and Y. Lu, “Temporal graphical models for cross-species gene regulatory network discovery.,” *Journal of bioinformatics and computational biology*, vol. 9, no. 2, pp. 231-50, Apr. 2011.

[55] H.-S. Le, Z. N. Oltvai, and Z. Bar-Joseph, “Cross Species Queries of Large Gene Expression Databases.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 19, pp. 2416-2423, Aug. 2010.

[56] Z. Liang, M. Xu, M. Teng, and L. Niu, “Comparison of protein interaction networks reveals species conservation and divergence.,” *BMC bioinformatics*, vol. 7, p. 457, Jan. 2006.

[57] R. Sharan et al., “Conserved patterns of protein interaction in multiple species.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1974-9, 2005.

[58] J. Berg and M. Lassig, “Cross-species analysis of biological networks by Bayesian alignment,” *Proc Natl Acad Sci USA*, vol. 103, pp. 10967-10972, 2006.

[59] R. Deshpande, S. Sharma, C. M. Verfaillie, W.-S. Hu, and C. L. Myers, “A Scalable Approach for Discovering Conserved Active Subnetworks across Species.,” *PLoS computational biology*, vol. 6, no. 12, p. e1001028, Jan. 2010.

[60] G. E. Zinman, S. Zhong, and Z. Bar-Joseph, “Biological interaction networks are conserved at the module level,” *BMC Systems Biology*, vol. 5, no. 1, p. 134, 2011.

[61] Y. Kanfi et al., “The sirtuin SIRT6 regulates lifespan in male mice,” *Nature*, Feb. 2012.

[62] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan, “Predicting function: from genes to genomes and back.,” *Journal of molecular biology*, vol. 283, no. 4, pp. 707-25, Nov. 1998.

[63] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander, “Sequencing and comparison of yeast species to identify genes and regulatory elements.,” *Nature*, vol. 423, no. 6937, pp. 241-54, May 2003.

[64] A. Stark et al., “Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes,” *Genome Res.*, Nov. 2007.

[65] J. L. Riechmann et al., “Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes,” *Science*, vol. 290, no. 5499, pp. 2105-2110, Dec. 2000.

[66] W. Miller, K. D. Makova, A. Nekrutenko, and R. C. Hardison, “Comparative genomics.,” *Annual review of genomics and human genetics*, vol. 5, pp. 15-56, Jan. 2004.

[67] B.-Y. Liao and J. Zhang, “Evolutionary conservation of expression profiles between human and mouse orthologous genes.,” *Molecular biology and evolution*, vol. 23, no. 3, pp. 530-40, Mar. 2006.

[68] G. Rustici et al., “Periodic gene expression program of the fission yeast cell cycle.,” *Nature genetics*, vol. 36, no. 8, pp. 809-17, Aug. 2004.

[69] D. T. Odom et al., “Tissue-specific transcriptional regulation has diverged significantly between human and mouse.,” *Nature genetics*, vol. 39, no. 6, pp. 730-2, 2007.

[70] A. Fox, D. Taylor, and D. K. Slonim, “High throughput interaction data reveals degree conservation of hub proteins.,” *Pacific Symposium on Biocomputing*, vol. 402, pp. 391-402, Jan. 2009.

[71] S. Suthram, T. Sittler, and T. Ideker, “The Plasmodium protein network diverges from those of other eukaryotes.,” *Nature*, vol. 438, no. 7064, pp. 108-12, Nov. 2005.

[72] T. K. B. Gandhi et al., “Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets.,” *Nature genetics*, vol. 38, no. 3, pp. 285-93, Mar. 2006.

[73] H. Yu et al., “Annotation Transfer Between Genomes : Protein – Protein Interologs and Protein – DNA Regulogs,” *Genome Research*, pp. 1107-1118, 2004.

[74] S. J. Dixon et al., “Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes,” *Proc Natl Acad Sci USA*, vol. 105, pp. 16653-16658, 2008.

[75] A. B. Byrne et al., “A global analysis of genetic interactions in Caenorhabditis elegans,” *J Biol*, vol. 6, p. 8, 2007.

[76] T. J. P. van Dam and B. Snel, “Protein complex evolution does not involve extensive network rewiring.,” *PLoS computational biology*, vol. 4, no. 7, p. e1000132, Jan. 2008.

[77] Z. Wang and J. Zhang, “In search of the biological significance of modular structures in protein networks,” *PLoS Comput Biol*, vol. 3, p. e107, 2007.

[78] P. Beltrao et al., “Evolution of phosphoregulation: comparison of phosphorylation patterns across yeast species.,” *PLoS biology*, vol. 7, no. 6, p. e1000134, Jun. 2009.

[79] E. Segal, N. Friedman, N. Kaminski, A. Regev, and D. Koller, “From signatures to models: understanding cancer using microarrays.,” *Nature genetics*, vol. 37, no. June, pp. S38-45, 2005.

[80] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, “A gene-coexpression network for global discovery of conserved genetic modules,” *Science*, vol. 302, pp. 249-255, 2003.

[81] J. Demeter et al., “The Stanford Microarray Database: implementation of new analysis tools and open source release of software.,” *Nucleic acids research*, vol. 35, no. Database issue, pp. D766-70, Jan. 2007.

[82] R. Edgar, M. Domrachev, and A. E. Lash, “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.,” *Nucleic acids research*, vol. 30, no. 1, pp. 207-10, Jan. 2002.

[83] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, “A probabilistic functional network of yeast genes.,” *Science (New York, N.Y.)*, vol. 306, no. 5701, pp. 1555-8, Nov. 2004.

[84] M. Ashburner et al., “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.,” *Nature genetics*, vol. 25, no. 1, pp. 25-9, May 2000.

[85] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, “A new method to measure the semantic similarity of GO terms,” *Bioinformatics*, vol. 23, no. 10, pp. 1274-1281, May 2007.

[86] V. Wood, “Schizosaccharomyces pombe comparative genomics; from sequence to systems.,” in *Comparative Genomics Using Fungi as Models (Series: Topics in Current Genetics)*, vol. 15, P. Sunnerhagen and J. Piskur, Eds. Berlin, Heidelberg: Springer Berlin, 2006, pp. 233-285.

[87] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, “An efficient algorithm for large-scale detection of protein families,” *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575-1584, Apr. 2002.

[88] S. Brohée and J. van Helden, “Evaluation of clustering algorithms for protein-protein interaction networks.,” *BMC bioinformatics*, vol. 7, p. 488, Jan. 2006.

[89] I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte, “A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans.,” *Nature genetics*, vol. 40, no. 2, pp. 181-8, Feb. 2008.

[90] C. Hertz-Fowler et al., “GeneDB: a resource for prokaryotic and eukaryotic organisms,” *Nucleic Acids Res*, vol. 32, p. D339-D343, 2004.

[91] M. Sipiczki, “Where does fission yeast sit on the tree of life?,” *Genome Biology*, vol. 1, no. 2, p. reviews1011.1-reviews1011.4, 2000.

[92] P. Smits, J. A. M. Smeitink, L. P. van den Heuvel, M. A. Huynen, and T. J. G. Ettema, “Reconstructing the evolution of the mitochondrial ribosomal proteome.,” *Nucleic acids research*, vol. 35, no. 14, pp. 4686-703, Jan. 2007.

[93] S. Brohee and J. van Helden, “Evaluation of clustering algorithms for protein-protein interaction networks,” *BMC Bioinformatics*, vol. 7, p. 488, 2006.

[94] P. Jiang and M. Singh, “SPICi: a fast clustering algorithm for large biological networks,” *Bioinformatics*, vol. 26, pp. 1105-1111, 2010.

[95] C. T. Harbison et al., “Transcriptional regulatory code of a eukaryotic genome.,” *Nature*, vol. 431, no. 7004, pp. 99-104, Sep. 2004.

[96] B.-Y. Liao and J. Zhang, “Null mutations in human and mouse orthologs frequently result in different phenotypes.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 19, pp. 6987-92, 2008.

[97] H. Parkinson et al., “ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression.,” *Nucleic acids research*, vol. 37, no. Database issue, pp. D868-72, Jan. 2009.

[98] J. Hubble et al., “Implementation of GenePattern within the Stanford Microarray Database.,” *Nucleic acids research*, vol. 37, no. Database issue, pp. D898-901, Jan. 2009.

[99] D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May, “NASCArrays: a repository for microarray data generated by NASC’s transcriptomics service.,” *Nucleic acids research*, vol. 32, no. Database issue, pp. D575-7, Jan. 2004.

[100] N. E. Olson, “The microarray data analysis process: from raw data to biological significance.,” *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*, vol. 3, no. 3, pp. 373-83, Jul. 2006.

[101] J. M. Engreitz, R. Chen, A. a Morgan, J. T. Dudley, R. Mallelwar, and A. J. Butte, “ProfileChaser: searching microarray repositories based on genome-wide patterns of differential expression.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 23, pp. 3317-8, Dec. 2011.

[102] W. Fujibuchi, L. Kiseleva, T. Taniguchi, H. Harada, and P. Horton, “CellMontage: similar expression profile search server.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 22, pp. 3103-4, Nov. 2007.

[103] J. M. Engreitz et al., “Content-based microarray search using differential expression profiles.,” *BMC bioinformatics*, vol. 11, no. 1, p. 603, Jan. 2010.

[104] P. Zimmermann, O. Laule, J. Schmitz, T. Hruz, S. Bleuler, and W. Gruissem, “Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases.,” *Molecular plant*, vol. 1, no. 5, pp. 851-7, Sep. 2008.

[105] M. A. Hibbs, D. C. Hess, C. L. Myers, C. Huttenhower, K. Li, and O. G. Troyanskaya, “Exploring the functional landscape of gene expression: directed search of large microarray compendia.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 20, pp. 2692-9, Oct. 2007.

[106] J. Lamb et al., “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.,” *Science (New York, N.Y.)*, vol. 313, no. 5795, pp. 1929-35, Sep. 2006.

[107] X. Li, S. Zhang, G. Blander, J. G. Tse, M. Krieger, and L. Guarente, “SIRT1 deacetylates and positively regulates the nuclear receptor LXR.,” *Molecular cell*, vol. 28, no. 1, pp. 91-106, Oct. 2007.

[108] Y. Kanfi et al., “SIRT6 protects against pathological damage caused by diet-induced obesity.,” *Aging cell*, vol. 9, no. 2, pp. 162-73, Apr. 2010.

[109] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116-21, Apr. 2001.

[110] B. B. Yeap et al., “IGF1 and its binding proteins 3 and 1 are differentially associated with metabolic syndrome in older men.,” *European journal of endocrinology / European Federation of Endocrine Societies*, vol. 162, no. 2, pp. 249-57, Feb. 2010.

[111] M. Blüher, B. B. Kahn, and C. R. Kahn, “Extended longevity in mice lacking the insulin receptor in adipose tissue.,” *Science (New York, N.Y.)*, vol. 299, no. 5606, pp. 572-4, Jan. 2003.

[112] M. B. Rosen et al., “Toxicogenomic dissection of the perfluorooctanoic acid transcript profile in mouse liver: evidence for the involvement of nuclear receptors PPAR alpha and

CAR.," *Toxicological sciences : an official journal of the Society of Toxicology*, vol. 103, no. 1, pp. 46-56, May 2008.

[113] M. Rakhshandehroo et al., "Comprehensive analysis of PPARalpha-dependent regulation of hepatic lipid metabolism by expression profiling.," *PPAR research*, vol. 2007, p. 26839, Jan. 2007.

[114] L. M. Sanderson et al., "Effect of synthetic dietary triglycerides: a novel research paradigm for nutrigenomics.," *PloS one*, vol. 3, no. 2, p. e1681, Jan. 2008.

[115] M. Lehrke et al., "Diet-dependent cardiovascular lipid metabolism controlled by hepatic LXRalpha.," *Cell metabolism*, vol. 1, no. 5, pp. 297-308, May 2005.

[116] S. P. Mooijaart et al., "Liver X Receptor Alpha Associates With Human Life Span," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 62, no. 4, pp. 343-349, Apr. 2007.

[117] J. He, Q. Cheng, and W. Xie, "Minireview: Nuclear receptor-controlled steroid hormone synthesis and metabolism.," *Molecular endocrinology (Baltimore, Md.)*, vol. 24, no. 1, pp. 11-21, Jan. 2010.

[118] J. H. Lee, J. Zhou, and W. Xie, "PXR and LXR in hepatic steatosis: a new dog and an old dog with new tricks.," *Molecular pharmaceutics*, vol. 5, no. 1, pp. 60-6, 2008.

[119] I. Jedidi et al., "Cholesteryl ester hydroperoxides increase macrophage CD36 gene expression via PPARalpha.," *Biochemical and biophysical research communications*, vol. 351, no. 3, pp. 733-8, Dec. 2006.

[120] A. C. Nicholson, S. Frieda, A. Pearce, and R. L. Silverstein, "Oxidized LDL Binds to CD36 on Human Monocyte-Derived Macrophages and Transfected Cell Lines : Evidence Implicating the Lipid Moiety of the Lipoprotein as the Binding Site," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 15, no. 2, pp. 269-275, Feb. 1995.

[121] K. Ishimoto et al., "Identification of human low-density lipoprotein receptor as a novel target gene regulated by liver X receptor alpha.," *FEBS letters*, vol. 580, no. 20, pp. 4929-33, Sep. 2006.

[122] C. Zhao and K. Dahlman-Wright, "Liver X receptor in cholesterol metabolism.," *The Journal of endocrinology*, vol. 204, no. 3, pp. 233-40, Mar. 2010.

[123] J. S. Lee et al., "Coordinated changes in xenobiotic metabolizing enzyme gene expression in aging male rats.," *Toxicological sciences : an official journal of the Society of Toxicology*, vol. 106, no. 1, pp. 263-83, Nov. 2008.

[124] L. Pourtau et al., “Hormonal, hypothalamic and striatal responses to reduced body weight gain are attenuated in anorectic rats bearing small tumors.,” *Brain, behavior, and immunity*, vol. 25, no. 4, pp. 777-86, May 2011.

[125] D. Sarrió, J. Palacios, M. Hergueta-Redondo, G. Gómez-López, A. Cano, and G. Moreno-Bueno, “Functional characterization of E- and P-cadherin in invasive breast cancer cells.,” *BMC cancer*, vol. 9, no. 1, p. 74, Jan. 2009.

[126] C. Fan et al., “Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures.,” *BMC medical genomics*, vol. 4, no. 1, p. 3, Jan. 2011.

[127] M. Ramakrishna et al., “Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis.,” *PloS one*, vol. 5, no. 4, p. e9983, Jan. 2010.

[128] N. C. Pressinotti et al., “Differential expression of apoptotic genes PDIA3 and MAP3K5 distinguishes between low- and high-risk prostate cancer.,” *Molecular cancer*, vol. 8, no. 1, p. 130, Jan. 2009.

[129] R. L. Prueitt et al., “Expression of microRNAs and protein-coding genes associated with perineural invasion in prostate cancer.,” *The Prostate*, vol. 68, no. 11, pp. 1152-64, Aug. 2008.

[130] K. A. Moynihan et al., “Increased dosage of mammalian Sir2 in pancreatic beta cells enhances glucose-stimulated insulin secretion in mice.,” *Cell metabolism*, vol. 2, no. 2, pp. 105-17, Aug. 2005.

[131] J. Paredes, A. Albergaria, J. T. Oliveira, C. Jerónimo, F. Milanezi, and F. C. Schmitt, “P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation.,” *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 11, no. 16, pp. 5869-77, Aug. 2005.

[132] A. Hurtado et al., “Regulation of ERBB2 by oestrogen receptor-PAX2 determines response to tamoxifen.,” *Nature*, vol. 456, no. 7222, pp. 663-6, Dec. 2008.

[133] C. Fix, C. Jordan, P. Cano, and W. H. Walker, “Testosterone activates mitogen-activated protein kinase and the cAMP response element binding protein transcription factor in Sertoli cells.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 30, pp. 10919-24, Jul. 2004.

[134] P. H. Gann, C. H. Hennekens, J. Ma, C. Longcope, and M. J. Stampfer, “Prospective Study of Sex Hormone Levels and Risk of Prostate Cancer,” *JNCI Journal of the National Cancer Institute*, vol. 88, no. 16, pp. 1118-1126, Aug. 1996.

[135] E. P. Stover, A. V. Krishnan, and D. Feldman, “Estrogen Down-Regulation of Androgen Receptors in Cultured Human Mammary Cancer Cells (MCF-7),” *Endocrinology*, vol. 120, no. 6, pp. 2597-2603, Jun. 1987.

[136] M. S. Kurzer, “Hormonal Effects of Soy in Premenopausal Women and Men,” *J. Nutr.*, vol. 132, no. 3, p. 570S-573, Mar. 2002.

[137] E. A. Jankowska et al., “Circulating estradiol and mortality in men with systolic chronic heart failure.,” *JAMA : the journal of the American Medical Association*, vol. 301, no. 18, pp. 1892-901, May 2009.

[138] H. Gong et al., “Estrogen deprivation and inhibition of breast cancer growth in vivo through activation of the orphan nuclear receptor liver X receptor.,” *Molecular endocrinology (Baltimore, Md.)*, vol. 21, no. 8, pp. 1781-90, Aug. 2007.

[139] J. Gao et al., “Sex-Specific Effect of Estrogen Sulfotransferase on Mouse Models of Type 2 Diabetes.,” *Diabetes*, pp. db11-1152-, Mar. 2012.

[140] H. Uppal et al., “Activation of LXRs prevents bile acid toxicity and cholestasis in female mice.,” *Hepatology (Baltimore, Md.)*, vol. 45, no. 2, pp. 422-32, Feb. 2007.

[141] L.-L. Vedin, S. A. Lewandowski, P. Parini, J.-A. Gustafsson, and K. R. Steffensen, “The oxysterol receptor LXR inhibits proliferation of human breast cancer cells.,” *Carcinogenesis*, vol. 30, no. 4, pp. 575-9, Apr. 2009.

[142] C. Gabbi, H.-J. Kim, R. Barros, M. Korach-Andrè, M. Warner, and J.-A. Gustafsson, “Estrogen-dependent gallbladder carcinogenesis in LXRbeta-/- female mice.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 33, pp. 14763-8, Aug. 2010.

[143] H. Campos, “Effect of Estrogen on Very Low Density Lipoprotein and Low Density Lipoprotein Subclass Metabolism in Postmenopausal Women,” *Journal of Clinical Endocrinology & Metabolism*, vol. 82, no. 12, pp. 3955-3963, Dec. 1997.

[144] B. A. Walsh, A. E. Mullick, C. E. Banka, and J. C. Rutledge, “17{beta}-Estradiol acts separately on the LDL particle and artery wall to reduce LDL accumulation,” *J. Lipid Res.*, vol. 41, no. 1, pp. 134-141, Jan. 2000.

[145] C. Subah Packer, “Estrogen protection, oxidized LDL, endothelial dysfunction and vasorelaxation in cardiovascular disease: New insights into a complex issue.,” *Cardiovascular research*, vol. 73, no. 1, pp. 6-7, Jan. 2007.

[146] M. Florian and S. Magder, “Estrogen decreases TNF-alpha and oxidized LDL induced apoptosis in endothelial cells.,” *Steroids*, vol. 73, no. 1, pp. 47-58, Jan. 2008.

[147] K. Kavanagh et al., “Estrogen decreases atherosclerosis in part by reducing hepatic acyl-CoA:cholesterol acyltransferase 2 (ACAT2) in monkeys.,” *Arteriosclerosis, thrombosis, and vascular biology*, vol. 29, no. 10, pp. 1471-7, Oct. 2009.

[148] S. Basu, A. Banerjee, and R. J. Mooney, “Active Semi-Supervision for Pairwise Constrained Clustering,” *Proceedings of the SIAM International Conference on Data Mining (SDM-2004)*, pp. 333-344, 2004.

[149] K. L. Wagstaff, “Value, cost, and sharing: open issues in constrained clustering,” in *Proceedings of the 5th international conference on Knowledge discovery in inductive databases*, 2007, pp. 1–10.

[150] Ian Davidson and S. S. Ravi, “Clustering with Constraints: Feasibility Issues and the k-Means Algorithm,” *SDM*, May 2005.

[151] M. E. Futschik and B. Carlisle, “Noise-robust soft clustering of gene expression time-course data.,” *J Bioinform Comput Biol*, vol. 3, pp. 965-988, 2005.

[152] A. P. Gasch and M. B. Eisen, “Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.,” *Genome Biol*, vol. 3, p. RESEARCH0059, 2002.

[153] D. Banerjee et al., “Responses of pathogenic and nonpathogenic yeast species to steroids reveal the functioning and evolution of multidrug resistance transcriptional networks.,” *Eukaryot Cell*, vol. 7, pp. 68-77, 2008.

[154] G. Lelandais, V. Tanty, C. Geneix, C. Etchebest, C. Jacq, and F. Devaux, “Genome adaptation to chemical stress: clues from comparative transcriptomics in Saccharomyces cerevisiae and Candida glabrata.,” *Genome Biol*, vol. 9, p. R164, 2008.

[155] A. Paulitsch, W. Weger, G. Ginter-Hanselmayer, E. Marth, and W. Buzina, “A 5-year (2000-2004) epidemiological survey of Candida and non-Candida yeast species causing vulvovaginal candidiasis in Graz, Austria.,” *Mycoses*, vol. 49, pp. 471-475, 2006.

[156] M. Sanguinetti, B. Posteraro, B. Fiori, S. Ranno, R. Torelli, and G. Fadda, “Mechanisms of azole resistance in clinical isolates of Candida glabrata collected during a hospital survey of antifungal resistance.,” *Antimicrob Agents Chemother*, vol. 49, pp. 668-679, 2005.

[157] L. S. Wilson, C. M. Reyes, M. Stolpman, J. Speckman, K. Allen, and J. Beney, “The direct cost and incidence of systemic fungal infections.,” *Value Health*, vol. 5, pp. 26-34, 2002.

[158] J. A. Maertens, “History of the development of azole derivatives.,” *Clin Microbiol Infect*, vol. 10, no. 1, pp. 1-10, 2004.

[159] J. B. Anderson, “Evolution of antifungal-drug resistance: mechanisms and pathogen fitness.,” *Nat Rev Microbiol*, vol. 3, pp. 547-556, 2005.

[160] A. J. Carrillo-Munoz, G. Giusiano, P. A. Ezkurra, and G. Quindos, “Antifungal agents: mode of action in yeast cells.,” *Rev Esp Quimioter*, vol. 19, pp. 130-139, 2006.

[161] A. Lupetti, R. Danesi, M. Campa, M. Del Tacca, and S. Kelly, “Molecular basis of resistance to azole antifungals.,” *Trends Mol Med*, vol. 8, pp. 76-81, 2002.

[162] D. Sanglard, F. Ischer, D. Calabrese, P. A. Majcherczyk, and J. Bille, “The ATP binding cassette transporter gene CgCDR1 from Candida glabrata is involved in the resistance of clinical isolates to azole antifungal agents.,” *Antimicrob Agents Chemother*, vol. 43, pp. 2753-2765, 1999.

[163] L. E. Cowen, “The evolution of fungal drug resistance: modulating the trajectory from genotype to phenotype.,” *Nat Rev Microbiol*, vol. 6, pp. 187-198, 2008.

[164] L. E. Cowen and S. Lindquist, “Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi.,” *Science*, vol. 309, pp. 2185-2189, 2005.

[165] G. Jansen et al., “Chemogenomic profiling predicts antifungal synergies.,” *Mol Syst Biol*, vol. 5, p. 338, 2009.

[166] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, “Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.,” *J Mol Biol*, vol. 296, pp. 1205-1214, 2000.

[167] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, “MEME: discovering and analyzing DNA and protein sequence motifs.,” *Nucleic Acids Res*, vol. 34, pp. W369-373, 2006.

[168] G. Pavesi, G. Mauri, and G. Pesole, “An algorithm for finding signals of unknown length in DNA sequences.,” *Bioinformatics*, vol. 17, no. 1, pp. S207-214, 2001.

[169] G. Z. Hertz and G. D. Stormo, “Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.,” *Bioinformatics*, vol. 15, pp. 563-577, 1999.

[170] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel, “An improved map of conserved regulatory sites for Saccharomyces cerevisiae.,” *BMC Bioinformatics*, vol. 7, p. 113, 2006.

[171] V. Matys et al., “TRANSFAC: transcriptional regulation, from patterns to profiles.,” *Nucleic acids research*, vol. 31, no. 1, pp. 374-8, Jan. 2003.

[172] S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D. S. Rokhsar, “Transcription factor binding site identification using the self-organizing map.,” *Bioinformatics*, vol. 21, pp. 1807-1814, 2005.

[173] J. D. Storey and R. Tibshirani, “Statistical significance for genomewide studies.,” *Proc Natl Acad Sci USA*, vol. 100, pp. 9440-9445, 2003.

[174] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev, “Natural history and evolutionary principles of gene duplication in fungi.,” *Nature*, vol. 449, pp. 54-61, 2007.

[175] A. Tanay, A. Regev, and R. Shamir, “Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast.,” *Proc Natl Acad Sci USA*, vol. 102, pp. 7203-7208, 2005.

[176] G. O. Consortium, “The Gene Ontology project in 2008.,” *Nucleic Acids Res*, vol. 36, pp. D440-444, 2008.

[177] A. Alexeyenko, I. Tamas, G. Liu, and E. L. L. Sonnhammer, “Automatic clustering of orthologs and inparalogs shared by multiple proteomes.,” *Bioinformatics*, vol. 22, pp. e9-15, 2006.

[178] D. J. Sherman, T. Martin, M. Nikolski, C. Cayla, J. L. Souciet, and P. Durrens, “Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes.,” *Nucleic Acids Res*, vol. 37, pp. D550-554, 2009.

[179] R. Kelley, H. Feizi, and T. Ideker, “Correcting for gene-specific dye bias in DNA microarrays using the method of maximum likelihood.,” *Bioinformatics*, vol. 24, pp. 71-77, 2008.

[180] J. B. Anderson et al., “Mode of selection and experimental evolution of antifungal drug resistance in Saccharomyces cerevisiae.,” *Genetics*, vol. 163, pp. 1287-1298, 2003.

[181] K. H. Wolfe and D. C. Shields, “Molecular evidence for an ancient duplication of the entire yeast genome.,” *Nature*, vol. 387, pp. 708-713, 1997.

[182] B. S. Davies and J. Rine, “A role for sterol levels in oxygen sensing in Saccharomyces cerevisiae.,” *Genetics*, vol. 174, pp. 191-201, 2006.

[183] P. M. Silver, B. G. Oliver, and T. C. White, “Role of Candida albicans transcription factor Upc2p in drug resistance and sterol metabolism.,” *Eukaryot Cell*, vol. 3, pp. 1391-1397, 2004.

[184] C. Zhu et al., “High-resolution DNA binding specificity analysis of yeast transcription factors.,” *Genome Res*, vol. 19, pp. 556-566, 2009.

[185] D. H. Nguyen and P. D’Haeseleer, “Deciphering principles of transcription regulation in eukaryotic genomes.,” *Mol Syst Biol*, vol. 2, p. 2006.0012, 2006.

[186] J. Morschhauser, “Regulation of multidrug resistance in pathogenic fungi.,” *Fungal Genet Biol*, vol. 47, pp. 94-106, 2009.

[187] A. Selmecki, A. Forche, and J. Berman, “Aneuploidy and isochromosome formation in drug-resistant Candida albicans.,” *Science*, vol. 313, pp. 367-370, 2006.

[188] D. Dimster-Denk et al., “Comprehensive evaluation of isoprenoid biosynthesis regulation in Saccharomyces cerevisiae utilizing the Genome Reporter Matrix(TM).,” *J Lipid Res*, vol. 40, pp. 850-860, 1999.

[189] R. C. Pascon, T. M. Ganous, J. M. Kingsbury, G. M. Cox, and J. H. McCusker, “Cryptococcus neoformans methionine synthase: expression analysis and requirement for virulence.,” *Microbiology*, vol. 150, pp. 3013-3023, 2004.

[190] K. C. Ha and T. C. White, “Effects of azole antifungal drugs on the transition from yeast cells to hyphae in susceptible and resistant isolates of the pathogenic yeast Candida albicans.,” *Antimicrob Agents Chemother*, vol. 43, pp. 763-768, 1999.

[191] S. Tenreiro, P. C. Rosa, C. A. Viegas, and I. Sa-Correia, “Expression of the AZR1 gene (ORF YGR224w), encoding a plasma membrane transporter of the major facilitator superfamily, is required for adaptation to acetic acid and resistance to azoles in Saccharomyces cerevisiae.,” *Yeast*, vol. 16, pp. 1469-1481, 2000.

[192] N. Broco, S. Tenreiro, C. A. Viegas, and I. Sa-Correia, “FLR1 gene (ORF YBR008c) is required for benomyl and methotrexate resistance in Saccharomyces cerevisiae and its benomyl-induced expression is dependent on pdr3 transcriptional regulator.,” *Yeast*, vol. 15, pp. 1595-1608, 1999.

[193] L. J. Wilcox, D. A. Balderes, B. Wharton, A. H. Tinkelenberg, G. Rao, and S. L. Sturley, “Transcriptional profiling identifies two members of the ATP-binding cassette transporter superfamily required for sterol uptake in yeast.,” *J Biol Chem*, vol. 277, pp. 32466-32472, 2002.

[194] P. Alimardani et al., “SUT1-promoted sterol uptake involves the ABC transporter Aus1 and the mannoprotein Dan1 whose synergistic action is sufficient for this process.,” *Biochem J*, vol. 381, pp. 195-202, 2004.

[195] F. Bussereau, S. Casaregola, J. F. Lafay, and M. Bolotin-Fukuhara, “The Kluyveromyces lactis repertoire of transcriptional regulators.,” *FEMS Yeast Res*, vol. 6, pp. 325-335, 2006.

[196] I. S. Snoek and H. Y. Steensma, “Why does Kluyveromyces lactis not grow under anaerobic conditions? Comparison of essential anaerobic genes of Saccharomyces cerevisiae with the Kluyveromyces lactis genome.,” *FEMS Yeast Res*, vol. 6, pp. 393-403, 2006.

[197] H. Nakayama et al., “The Candida glabrata putative sterol transporter gene CgAUS1 protects cells against azoles in the presence of serum.,” *J Antimicrob Chemother*, vol. 60, pp. 1264-1272, 2007.

[198] R. Torelli et al., “The ATP-binding cassette transporter-encoding gene CgSNQ2 is contributing to the CgPDR1-dependent azole resistance of Candida glabrata.,” *Mol Microbiol*, vol. 68, pp. 186-201, 2008.

[199] J. Ihmels, S. Bergmann, J. Berman, and N. Barkai, “Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program.,” *PLoS Genet*, vol. 1, p. e39, 2005.

[200] O. Alter, P. O. Brown, and D. Botstein, “Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.,” *Proc Natl Acad Sci USA*, vol. 100, pp. 3351-3356, 2003.

[201] G. Zinman et al., “Large scale comparison of innate responses to viral and bacterial pathogens in mouse and macaque.,” *PloS one*, vol. 6, no. 7, p. e22401, Jan. 2011.

[202] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, “Discovering regulatory and signalling circuits in molecular interaction networks.,” *Bioinformatics (Oxford, England)*, vol. 18 Suppl 1, no. 1, pp. S233-40, Jan. 2002.

[203] M. Liu et al., “Network-based analysis of affected biological processes in type 2 diabetes models.,” *PLoS genetics*, vol. 3, no. 6, p. e96, Jun. 2007.

[204] Z. Guo et al., “Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction sub-network.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 16, pp. 2121-8, Aug. 2007.

[205] D. Rajagopalan and P. Agarwal, “Inferring pathways from gene lists using a literature-derived network of biological relationships.,” *Bioinformatics (Oxford, England)*, vol. 21, no. 6, pp. 788-93, Mar. 2005.

[206] S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes, “Gene expression network analysis and applications to immunology.,” *Bioinformatics (Oxford, England)*, vol. 23, no. 7, pp. 850-8, Apr. 2007.

[207] F. Sohler, D. Hanisch, and R. Zimmer, “New methods for joint analysis of biological networks and expression data.,” *Bioinformatics (Oxford, England)*, vol. 20, no. 10, pp. 1517-21, Jul. 2004.

[208] R. Breitling, A. Amtmann, and P. Herzyk, “Graph-based iterative Group Analysis enhances microarray interpretation.,” *BMC bioinformatics*, vol. 5, no. 1, p. 100, Jul. 2004.

[209] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. Müller, “Identifying functional modules in protein-protein interaction networks: an integrated exact approach.,” *Bioinformatics (Oxford, England)*, vol. 24, no. 13, pp. i223-31, Jul. 2008.

[210] Y.-qing Qiu, “Uncovering differentially expressed pathways with protein interaction and gene expression data,” *Systems Biology*, pp. 74-82, 2008.

[211] L. Cabusora, E. Sutton, A. Fulmer, and C. V. Forst, “Differential network expression during drug and stress response.,” *Bioinformatics (Oxford, England)*, vol. 21, no. 12, pp. 2898-905, Jun. 2005.

[212] K. Faust, P. Dupont, J. Callut, and J. van Helden, “Pathway discovery in metabolic networks by subgraph extraction.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 9, pp. 1211-1218, Mar. 2010.

[213] I. Ulitsky and R. Shamir, “Identifying functional modules using expression profiles and confidence-scored protein interactions.,” *Bioinformatics (Oxford, England)*, vol. 25, no. 9, pp. 1158-64, May 2009.

[214] I. Ulitsky and R. Shamir, “Identification of functional modules using network topology and high-throughput data.,” *BMC systems biology*, vol. 1, no. 1, p. 8, Jan. 2007.

[215] Z. Wu, X. Zhao, and L. Chen, “Identifying responsive functional modules from protein-protein interaction network.,” *Molecules and cells*, vol. 27, no. 3, pp. 271-7, Mar. 2009.

[216] M. Santic, S. Al-Khodor, and Y. Abu Kwaik, “Cell biology and molecular ecology of Francisella tularensis.,” *Cellular microbiology*, vol. 12, no. 2, pp. 129-39, Feb. 2010.

[217] M. Kanehisa et al., “From genomics to chemical genomics: new developments in KEGG.,” *Nucleic acids research*, vol. 34, no. Database issue, pp. D354-7, Jan. 2006.

[218] B. C. Russo et al., “A Francisella tularensis locus required for spermine responsiveness is necessary for virulence.,” *Infection and immunity*, vol. 79, no. 9, pp. 3665-76, Sep. 2011.

[219] P. E. Carlson, J. A. Carroll, D. M. O’Dee, and G. J. Nau, “Modulation of virulence factors in Francisella tularensis determines human macrophage responses.,” *Microbial pathogenesis*, vol. 42, no. 5-6, pp. 204-14, 2007.

[220] H. Andersson et al., “Transcriptional profiling of the peripheral blood response during tularemia.,” *Genes and immunity*, vol. 7, no. 6, pp. 503-13, Sep. 2006.

[221] C. Paranavitana, P. R. Pittman, M. Velauthapillai, E. Zelazowska, and L. Dasilva, “Transcriptional profiling of Francisella tularensis infected peripheral blood mononuclear cells: a predictive tool for tularemia.,” *FEMS immunology and medical microbiology*, vol. 54, no. 1, pp. 92-103, Oct. 2008.

[222] M. Santic, G. Pavokovic, S. Jones, R. Asare, and Y. A. Kwaik, “Regulation of apoptosis and anti-apoptosis signalling by Francisella tularensis.,” *Microbes and infection / Institut Pasteur*, vol. 12, no. 2, pp. 126-34, Feb. 2010.

[223] Chen Lin-feng, Mu Yajun, and Greene Warner C., “Acetylation of RelA at discrete sites regulates distinct nuclear functions of NF-kappaB,” *The EMBO Journal*, vol. 21, no. 23, pp. 6539-6548, Dec. 2002.

[224] C. Prives and J. L. Manley, “Why is p53 acetylated?,” *Cell*, vol. 107, no. 7, pp. 815-8, Dec. 2001.

[225] J. P. Butchar et al., “Microarray analysis of human monocytes infected with Francisella tularensis identifies new targets of host response subversion.,” *PloS one*, vol. 3, no. 8, p. e2924, Jan. 2008.

[226] M. V. S. Rajaram, L. P. Ganesan, K. V. L. Parsa, J. P. Butchar, J. S. Gunn, and S. Tridandapani, “Akt/Protein kinase B modulates macrophage inflammatory response to Francisella infection and confers a survival advantage in mice.,” *Journal of immunology (Baltimore, Md. : 1950)*, vol. 177, no. 9, pp. 6317-24, Nov. 2006.

[227] G. D. Barish et al., “A Nuclear Receptor Atlas: macrophage activation.,” *Molecular endocrinology (Baltimore, Md.)*, vol. 19, no. 10, pp. 2466-77, Oct. 2005.

[228] D. Szklarczyk et al., “The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.,” *Nucleic acids research*, vol. 39, no. Database issue, pp. D561-8, Jan. 2011.

[229] C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris, and G. D. Bader, “Cytoscape Web: an interactive web-based network browser.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 18, pp. 2347-8, Sep. 2010.

[230] 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*, vol. 33, no. Database issue, pp. D514-7, Jan. 2005.

[231] C. J. Mattingly, M. C. Rosenstein, A. P. Davis, G. T. Colby, J. N. Forrest, and J. L. Boyer, “The comparative toxicogenomics database: a cross-species resource for building chemical-gene interaction networks.,” *Toxicological sciences : an official journal of the Society of Toxicology*, vol. 92, no. 2, pp. 587-95, Aug. 2006.

[232] D. S. Wishart et al., “DrugBank: a knowledgebase for drugs, drug actions and drug targets.,” *Nucleic acids research*, vol. 36, no. Database issue, pp. D901-6, Jan. 2008.

[233] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, “BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities.,” *Nucleic acids research*, vol. 35, no. Database issue, pp. D198-201, Jan. 2007.

[234] P. Groth et al., “PhenomicDB: a new cross-species genotype/phenotype resource.,” *Nucleic acids research*, vol. 35, no. Database issue, pp. D696-9, Jan. 2007.

[235] P. Groth, I. Kalev, I. Kirov, B. Traikov, U. Leser, and B. Weiss, “Phenoclustering: online mining of cross-species phenotypes.,” *Bioinformatics (Oxford, England)*, vol. 26, no. 15, pp. 1924-5, Aug. 2010.

[236] A. Oshlack, M. D. Robinson, and M. D. Young, “From RNA-seq reads to differential expression results.,” *Genome biology*, vol. 11, no. 12, p. 220, Jan. 2010.

[237] A. Goncalves, A. Tikhonov, A. Brazma, and M. Kapushesky, “A pipeline for RNA-seq data processing and quality assessment.,” *Bioinformatics (Oxford, England)*, vol. 27, no. 6, pp. 867-9, Mar. 2011.

[238] B. Langmead, K. D. Hansen, and J. T. Leek, “Cloud-scale RNA-sequencing differential expression analysis with Myrna.,” *Genome biology*, vol. 11, no. 8, p. R83, Jan. 2010.

[239] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel, “RNA-Seq Atlas – A reference database for gene expression profiling in normal tissue by next generation sequencing,” *Bioinformatics*, Feb. 2012.

[240] A. C. Frazee, B. Langmead, and J. T. Leek, “ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets.,” *BMC bioinformatics*, vol. 12, no. 1, p. 449, Jan. 2011.

[241] E. Wijaya, M. C. Frith, K. Asai, and P. Horton, “RecountDB: a database of mapped and count corrected transcribed sequences.,” *Nucleic acids research*, vol. 40, no. Database issue, pp. D1089-92, Jan. 2012.

[242] C. Boone, H. Bussey, and B. J. Andrews, “Exploring genetic interactions and networks with yeast.,” *Nature reviews. Genetics*, vol. 8, no. 6, pp. 437-49, Jun. 2007.