From genomes to organisms: integrating genomic data

Cristian I. Castillo-Davis

June 30, 2005

1 Introduction

Biology has gone from being a data-poor science to a data-rich one and thus presents an exciting challenge not only for biologists but also for statisticians, computer scientists and other quantitative workers. The prodigious and ever-growing bounty of "-omic" data generated by technologies enabled by whole genome DNA sequencing projects is quickly out-pacing our ability to digest and meaningfully synthesize it. These data include transciptional, proteomic, and phenotypic data, to name but a few.

However, recent work has shown that a biologically and statistically thoughtful combination of different data types in either a hypothesis-driven or data-mining framework can lead to a deeper, more comprehensive understanding of biology. Post- genomic analysis, the interpretation and synthesis of thousands of data points from a chemical, clinical, evolutionary, or other perspective thus promises to be an area of great methodological and scientific development in this century.

For many genes, something is known about their molecular and biological function, pathway membership, physical chromosomal location, level of polymorphism, RNAi phenotype, disease phenotype, and rate of molecular evolution. For non-coding regions, data are often available concerning the presence of known or putative transcription factor binding sites, levels of DNA methylation or acetylation, and GC content. While freely available through public databases, these different kinds of biological data are often unexamined with respect to each other. One reason for this situation is a lack of conceptual and methodological tools for their analysis. The continual release of new genomic and proteomic datasets insures that this situation will only be exacerbated in the coming decades. At the same time, this problem offers an unprecedented opportunity for innovation and scientific discovery not only for biologist but for statisticians, computer scientists, and others.

Since there is no one solution to the problem of integrating high-throughput genomic data, and since the types of data available will undoubtedly change over time, I will concentrate on familiarizing the reader with specific examples where integrative post-genomic analysis has been successfully applied, and highlight key areas of investigation that are especially fertile for future contribution. In doing so, out of familiarity, I will use examples largely from my own work. My goal is not a comprehensive review of the literature but an illustration of some of the applications, challenging problems, and exciting possibilities of combining different types of genomic data toward biological ends.

2 Case-study I — Intron evolution

To illustrate a relatively straightforward case of hypothesis-driven post-genomic analysis that uses disparate data-types in its execution we will begin with an example involving the evolution of gene structure [CDMH$^+$02].

2.1 Background

Introns are intervening sections of DNA within protein-coding regions of genes that do not encode amino-acids (Figure 1) and are primarily made up of non-functional "junk DNA." These sections of DNA are nonetheless transcribed (copied) by the cell along with the protein-coding sections (known as exons) into messenger RNA (mRNA) as one long transcript. The introns in an mRNA transcript are subsequently cut out of the transcript (literally) and the exons spliced together (literally) to form the usable mRNA message.

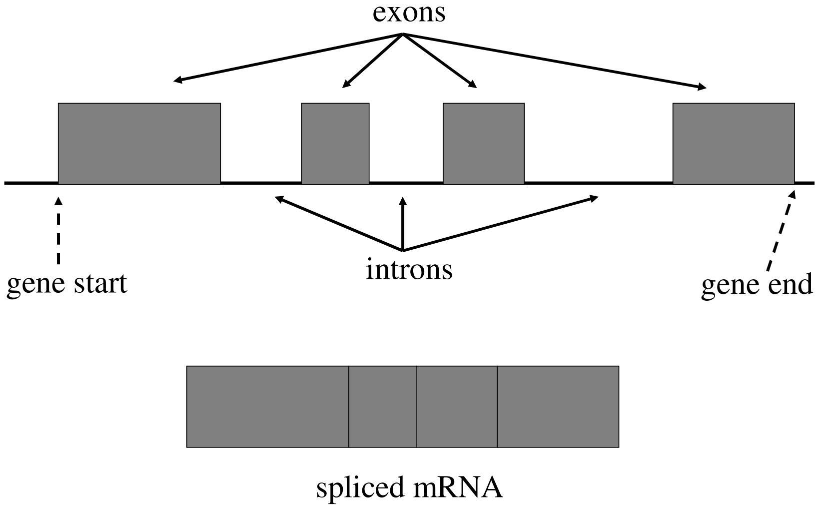


Figure 1: Exon-intron structure of a gene

This mRNA transcript is later translated into an amino-acid chain which then folds to make a protein. Some of the largest introns are found in the human genome, where the total length of intron sequences in a gene often reaches tens of thousands of nucleotides such that transcription of a single gene requires several minutes and thousands of ATP molecules (the energy currency of the cell).

2.2 Hypothesis

Because transcription is a slow and metabolically expensive process in eukaryotes, it was hypothesized that, at least for highly expressed genes, transcription of long introns, might be energetically costly. If so, in genes that are highly expressed, it is predicted that natural selection will favor shorter introns. To test this hypothesis requires at least two sets of data: 1) data on gene structure detailing the sizes of exons and introns making up all the genes in a genome, and 2) estimates of the expression level of each gene.

2.3 Methods

At the time of this study, sufficient information on both exon-intron structure and gene expression data were available only for two species: the nematode *Caenorhabditis elegans* and human. Gene structure information for each species was available through genome databases and consisted of coordinates listing exon and intron boundaries. In terms of expression data, for *C. elegans*, Affymetrix microarray expression data collected over development was available that provided absolute transcript abundance measures for each gene. Unfortunately, such microarray experiments were not available for human, and gene expression was instead estimated by expressed sequence tag abundance.

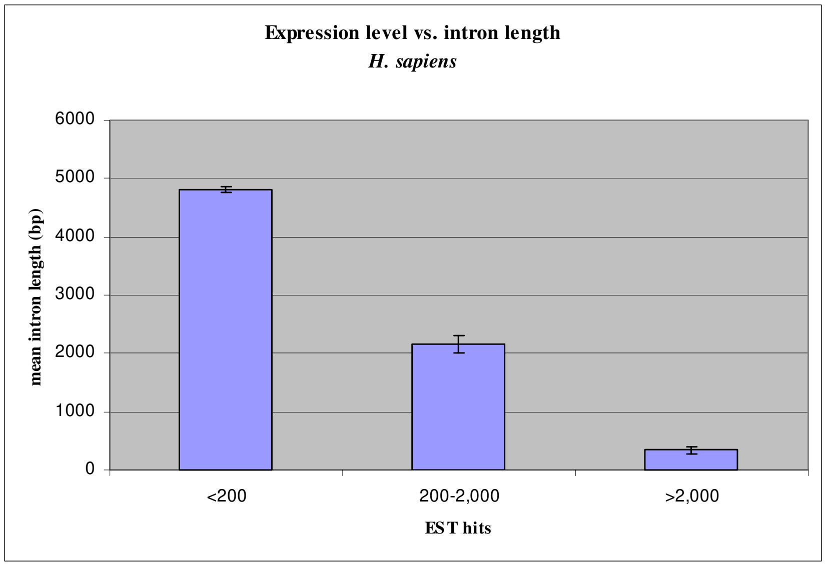
Expressed sequence tags (ESTs) are short stretches of DNA, randomly sequenced after reverse transcribing a pool of mRNA that is typically extracted from a tissue or organ. Since some mRNA transcripts are more abundant than others, these will be sequenced more often, and in turn will end up making up the bulk of sequences in an EST database. By aligning the known DNA sequence of a given gene with EST sequences in an EST database and counting the number of significant matches, one can estimate the expression level of that gene [BD99]. This was the approach taken in this study to estimate gene expression level in human, using BLAST [AMS+97] for sequence alignment and all available human EST sequences in GenBank [BKML+05].

2.4 Result

By combining information on intron size and the two types of expression data discussed above, it was found that introns in highly expressed genes were indeed substantially shorter than those of genes expressed at low levels in both in human and *C. elegans* (Figure 2).

2.5 Discussion

In this case study, the authors had a very specific hypothesis in mind and attempted to test its predictions using available data. No sophisticated modeling



(a)

(b)

Figure 2: Mean expression level versus intron size in (a) *H. sapiens* and (b) *C. elegans*

was used nor were high-level statistics necessary to obtain the biological results. This case study shows that the evaluation of important biological hypotheses is possible with a minimal amount of disparate genomic data (in this case, three types) if the data are combined in a biologically and statistically thoughtful manner. Many important biological questions remain unanswered even in the wake of an abundance of genomic data. I hope this inspires workers outside and on the periphery of biology to apply novel tools and fresh perspectives to genomic investigation. The opportunity for methodological and scientific contribution are great.

3 Case Study II — Functional genomics and protein evolution

To illustrate a case of post-genomic analysis that is more data-mining in spirit and that utilizes a number of different data types, we now turn to a study on protein conservation and function [CDKHK04]. This study is largely aimed at answering three basic questions: "What are the slowest evolving (most conserved) proteins in animal genomes and what do they do?" and "What are the fastest evolving (least conserved) proteins in animal genomes and what do they do?" And finally, "Are fast and slow evolving genes the same types of genes in different animals?"

3.1 Background

An important question in biology is how selective forces act on the genome in the evolution of different species. For example, does natural selection act similarly on proteins across lineages as distinct as phyla? Since most multicellular organisms contain a similar complement of genes and gene families owing, in part, to a common cellular biology, it might be expected that natural selection acts homogeneously across functionally similar genes in widely disparate taxa. However, this is not certain and there are many reasons why inhomogeneous levels of conservation across the proteome might be expected in different animals for example strong lineage-specific adaptation. To address this question we need first to determine the rate of evolution of all genes in two different animals and second, integrate this information with data on gene function.

3.2 Methods

Rates of evolution for two species pairs in two different animal phyla, Chordata (mouse/human) and Nematoda (*C. elegans*/*C. briggsae*) (Figure 3) were estimated by the maximum likelihood method of Yang and colleagues [GY94] [NY98]. This method calculates the estimated rate of nonsynonymous (amino-acid changing) substitutions between proteins d_N and the synonymous (non amino-acid changing) rate of substitution d_S between proteins.

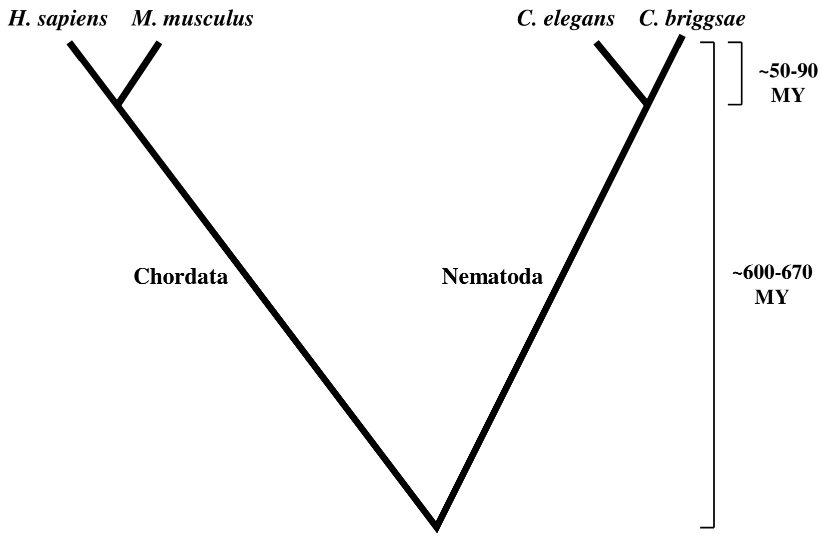


Figure 3: Evolutionary relationships and divergence times of species studied in case study II. MY = million years.

These data were subsequently examined with respect to gene function using two complementary approaches. In the first approach, a list of the top 10% fastest and slowest evolving proteins in each species pair (in terms of d_N were compared with known gene functions from the Gene Ontology (GO) database [ABB+00] (http://www.geneontology.org) and tested for statistical enrichment. In the second, tissue-specific expression of all genes in the mammalian dataset was estimated based on hits to EST sequence libraries and then rates of evolution for genes expressed in each tissue type were calculated.

The statistical enrichment of various functional classes among slow and fast evolving genes was evaluated using GeneMerge [CDH03] (http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html). Annotated gene functions from the Gene Ontology Consortium [ABB+00] for human and *C. elegans* were used as input for GeneMerge. GeneMerge related methods will be discussed in greater depth later in the chapter.

EST data were obtained from cDNA libraries available in GenBank (http://www.ncbi.nlm.nih.gov). More than 450,000 ESTs from 12 normal adult mouse tissues were collected and alignments evaluated against each mouse gene using BLASTN [AMS+97]. Genes with significant hits to ESTs were then normalized and clustered into tissue-specific groups by means of a Self-Organizing Tree Algorithm (SOTA) [HVD01]. Clusters represent genes that have similar expression patterns across tissues (Figure 4). Mean divergence estimates were then calculated for each cluster with confidence intervals estimated by means of nonparametric bootstrap resampling with 1,000 replicates.

3.3 Results

The 10% fastest evolving genes in mammals, according to the GO annotations, were largely involved in reproduction, immunity, and signal transduction (Table 2), whereas transcription factors were over-represented among fast evolving nematode proteins (Table 3).

GO Description	*Fraction*	*P-value*	*GO ID*
Immune response	100/577	3.77E-040	GO:0006955
Response to pest/pathogen/parasite	61/577	2.76E-023	GO:0009613
Antimicrobial humoral response	24/577	4.84E-013	GO:0019730
Response to wounding	27/577	2.68E-006	GO:0009611
Innate immune response	20/577	0.000357	GO:0045087
Inflammatory response	19/577	0.001230	GO:0006954
Lymphocyte activation	7/577	0.008820	GO:0046649
Pregnancy	8/577	0.009790	GO:0007565

Table 2. Functional overrepresentation of fast evolving mammal genes.

GO Description	*Fraction*	*P-value*	*GO ID*
DNA-dependent regul. of transcription	45/753	4.27E-5	GO:0006355
Regulation of transcription	45/753	5.72E-5	GO:0045449
Nucleic acid metabolism	53/753	0.03675	GO:0006139

Table 3. Functional overrepresentation of fast evolving worm genes.

Corroborating these results, the EST data (Figure 4) showed that genes co-expressed in the thymus and spleen (immune organs) in mouse evolved the fastest among all tissues $d_N = 0.142$. Additionally, an accelerated mean rate of evolution was seen in genes co-expressed in the ovary and uterus $d_N = 0.122$, organs with a reproductive role.

In contrast, the slowest-evolving genes in both nematodes and mammals were involved in *the same* basic cellular processes including protein biosynthesis, cell growth and GTP-mediated signal transduction (Table 4,5).

GO Description	*Fraction*	*P-value*	*GO ID*
Protein metabolism	140/699	5.76E-10	GO:0019538
Intracellular protein transport	44/699	5.89E-9	GO:0006886
Small GTPase mediated sign. transd.	30/699	1.85E-7	GO:0007264
Ubiquitin-dependent protein catabolism	25/699	4.81E-6	GO:0006511
Biosynthesis	69/699	0.000284	GO:0009058
Nucleocytoplasmic transport	13/699	0.000461	GO:0006913
Metabolism	265/699	0.001011	GO:0008152
mRNA splicing	10/699	0.039416	GO:0006371

Table 4. Functional overrepresentation of slow evolving mammal genes.

GO Description	*Fraction*	*P-value*	*GO ID*
Physiological processes	268/753	4.04E-12	GO:0007582
Protein biosynthesis	48/753	1.82E-11	GO:0006412
Cellular process	132/753	4.89E-11	GO:0009987
Biosynthesis	63/753	5.25E-11	GO:0009058
Small GTPase mediated sign. transd.	23/753	7.89E-11	GO:0007264
Metabolism	189/753	2.17E-5	GO:0008152
Protein metabolism	92/753	3.01E-5	GO:0019538
mRNA splicing	5/753	0.016747	GO:0006371

Table 5. Functional overrepresentation of slow evolving worm genes.

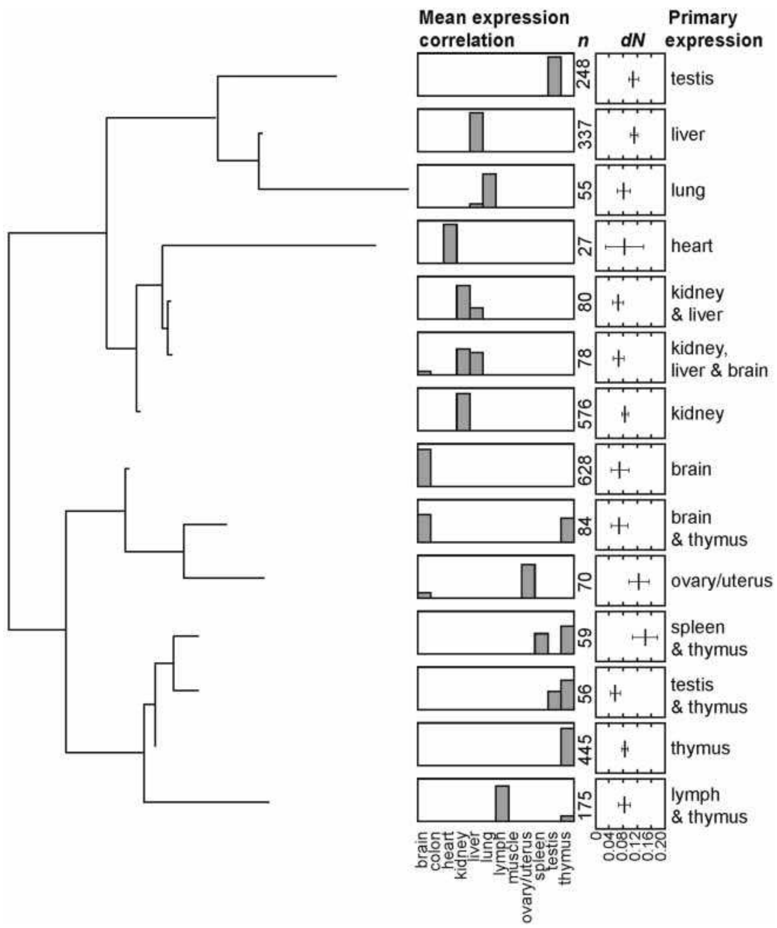


Figure 4: Tissue-specific gene expression and protein divergence. Histograms show the mean correlation coefficient for gene expressed in a cluster. Reproduced with permission from Cold Spring Harbor Laboratory Press [CDKHK04].

Thus it appears that while fast-evolving genes tend to be lineage-specific, highly-conserved genes are the same in different types of animals and are mainly involved in core cellular functions.

3.4 Discussion

Leaving aside the biological implications of the study we will concentrate on the methods used to integrate the comparative and functional genomic data. Firstly, note that a two-pronged and complementary approach was taken to establish gene function. Database annotations are currently incomplete with upwards of 50% of genes having unknown function, even in model organisms. Thus, it was important in this study to complement the database annotation data with a method using all genes, even those with unknown functions. The EST-based tissue-specific expression analysis satisfied this goal. In general when combing genomic data— which are often noisy or incomplete— similar strategies of data complementation are often useful since certain data types can bolster deficiencies in others.

4 Toward general methods for data-combination and exploration

Having reviewed two case-studies involving the combination of disparate data types, we now turn to a more general discussion of methods to combine and analyze genomic data. Data associated with genes are many and varied and will undoubtedly grow as genomic and proteomic investigations accelerate. To deal with this explosion of data requires 1) a clear analytical framework and 2) the flexibility to examine new data as soon as they become available. To date, there are very few approaches that meet both these criteria.

However, one approach that has been quite fruitful is the so-called over-representation framework where investigators examine the overlap of particular attributes in a sample of genes drawn from a larger set of genes, often a genome. By far, the most common application of this approach is the examination of a list of genes that are found to be highly expressed in, say, breast cancer tissue versus normal breast tissue, for statistical over-representation of gene functions within the list. There are several programs that implement this general algorithm [MBR$^+$04] (Table 6) using functions provided by the Gene Ontology Consortium; the most commonly implemented statistic to assess overrepresentation is based on the hypergeometric distribution [MBR$^+$04].

$$\Pr(r|n,p,k) = \frac{\binom{pn}{r}\binom{(1-p)n}{k-r}}{\binom{n}{k}} \tag{1}$$

The hypergeometric distribution describes the discrete probability of selecting r items of one kind in a sample of size k from a population of size n, where p is equal to the proportion of r type items in the population, and sampling is without replacement. The hypergeometric thus gives quantification of the level of ones surprise at finding overrepresentation for a particular item in a given sample of size k drawn from the larger population, size n. k is typically a set of genes that are highly or lowly expressed and n is the population set, the set from which k is drawn, usually all genes on a particular microarray. P-values can be calculated by summing over the tail of the distribution for all less-likely cases.

$$\sum_{i=r}^{k} \Pr(i|n,p,k) \tag{2}$$

Since several to hundreds of gene attributes are usually tested for overrepresentation in a given analysis, correction for multiple hypothesis testing is important. For example, if we were to test whether a set of genes involved in brain cancer were found disproportionately on a particular chromosome by testing for overrepresentation on each of the 22 non-sex human chromosomes, we would have carried out 22 separate hypothesis tests. Thus, some kind of P-value correction must be made. The traditional Bonferroni correction is the most popular [MBR$^+$04], but, less severe corrections, such as those based on the False Discovery Rate (FDR) [BH95] [Sto02], are becoming more common.

Program	*Stat.*	*Mult. Test. Corr.*
CLENCH	Hypergeometric*	NA
FatiGO	Fisher exact test	FDR
FuncAssociate	Fisher exact test	P-value adjus.
FuncSpec	Hypergeometric	Bonferroni
GeneMerge	Hypergeometric	Bonferroni
GFINDer	Hypergeometric*	Bonferroni
GoMiner	Fisher exact test	NA
Gostat	Fisher exact test	Holm/FDR/Yekutieli
GO Term-Finder (CPAN)	Hypergeometric	Bonferroni/FDR
GOTM	Hypergeometric	NA
GOToolBox	Hypergeometric*	Bonferroni

Table 6. Overrepresentation tools that use Gene Ontology annotations; from [MBR$^+$04]; see references therein for associated publications). * indicates software is capable of other statistical tests as well.

The first general-purpose implementation of the overrepresentation approach to genomic data, GeneMerge [CDH03] was designed with the express purpose of combining many different types of data related to genes, and thus will be our focus here. In GeneMerge the study set k may be genes found to be significantly up or down-regulated in a microarray experiment or a list of genes deemed interesting *for any another reason*. Genes in the sample k are associated with particular identifiers, for example functions, processes, or states. The number of genes with a particular identifier is r. p is the fraction of genes in the population n associated with the particular identifier under investigation.

GeneMerge returns both descriptive information regarding the genes under investigation and Bonferroni corrected and uncorrected rank scores regarding overrepresentation of any number of different descriptors in a given set of genes. Functional or categorical descriptive data is associated with genes in *gene-association* files. These text files link each gene in a genome with a particular datum of information. For example, the name of a gene and its chromosomal location, its sensitivity to a particular small-molecule, or its identity as over-expressed in a particular type of cancer.

The use of overrepresentation techniques has been quite useful when applied to microarray data using GO gene functions (for example [RCDMH03] [PMM+02] and many, many more) and genetic pathway membership (for example [CTH00] and many, many more). Interestingly however, this method has been less often used for data exploration and hypothesis testing of more diverse gene-association data, for example, mutation phenotypes, microarray expression outcomes, and genetic interactions. A partial list of gene-association data potentially useful for different genomic analyses is given in Table 7. Unfortunately, most software implementations do not allow users to generate and utilize a wide range of gene-association data. One advantage of GeneMerge over other similar programs is that its gene-association files are simply tab-delimited text files that can be prepared using any spreadsheet program.

Gene-association Data
knock-out phenotype
disease phenotype
polymorphic / non-polymorphic locus
local recombination rate
expression phenotype under influence of chemical X
publication mention
transcription factor binding sites
protein-protein network connectivity (degree)
viability/inviability if deleted
acetylated under condition X
GC content
sex-specific expression
tissue-specific expression
has ortholog in clade X
rate of molecular evolution
genetic interactions with other genes
over/under-expressed in experiment X
alternatively spliced
RNAi phenotype
expressed in anatomical region X

Table 7. A partial list of gene-association data.

To illustrate how the overrepresentation approach can be applied to data beyond the traditional microarray expression/GO function paradigm, two examples are given below. These are followed by a discussion of the limitations and possible extensions of the method as well as potential new approaches for the combination and examination of disparate genomic data.

4.1 Overrepresentation methods — beyond microarray data

To illustrate the flexibility of overrepresentation techniques as applied to genomic data, I will present two unpublished examples, one involving a population genomics question and another involving literature mining.

4.1.1 Population genomics

Recent analysis of genomic data suggests that protein evolution is related to protein effects on organism fitness; specifically, it has been shown that proteins that cause lowered fitness when deleted in yeast, so called "non-dispensable" genes, tend to evolve more slowly [HF01]. While amino acid substitution rates tend to be higher in dispensable genes over long evolutionary distances [HF01], it is not known whether in natural populations these genes also tend to be more *polymorphic*, that is, show more inter-individual variation. Given that selection against deleterious mutations is also expected to operate at the population level, coupled with the observation that variation among natural populations is ultimately transformed into variation among species, we may predict that dispensable genes will be more polymorphic within populations. In other words, non-essential genes will show more variation than essential genes in a population.

Polymorphic genes were identified using genomic hybridizations to Affymetrix arrays containing 126,645 unique 25mer yeast probes among 14 strains of laboratory and wild yeast [WCDO$^+$03]. These arrays are sensitive to the detection of single nucleotide polymorphisms. Unfortunately, distinction between synonymous and nonsynonymous substitutions is not possible with these arrays. Genes with at least one detected polymorphism among the 14 strains were considered polymorphic. Genes with no polymorphism were considered non-polymorphic. Among the 2991 genes probed on the chip, 1874 (63%) were polymorphic by this criterion. To create a deletion viability gene association file, lists of genes that result in inviablity or are viable when deleted were obtained from the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces/) based on the data of [WSA$^+$99] and [GCN$^+$02]. 4713 genes were listed as having a deletion viable phenotype and 1115 genes an inviable deletion phenotype. 413 genes had no data available concerning deletion phenotype. The hypothesis that population level polymorphism is more likely to occur in dispensable genes appears to be supported by the data.

Among *S. cerevisiae* genes categorized as polymorphic, more are viable upon deletion than is expected by chance. Of the 1874 genes categorized as polymorphic, 1454 (77%) were deletion viable, representing an enrichment in this class of genes ($P < 0.006$) (Table 8).

Description	*Fraction*	*P-value*	*corr. P-value*	*ID*
deletion inviable	336/1874	0.3880	0.7760	del_inv
deletion viable	1454/1874	0.0025	0.0051	del_via

Table 8. Polymorphism and deletion viability in yeast.

Thus selection against deleterious mutations in potentially more important genes appears to result in visibly lower levels of polymorphism at the population level. While this result is preliminary, it provides one example of how overrepresentation approaches can be used to explore genomic hypotheses efficiently in data beyond the microarray/GO function paradigm.

4.1.2 Literature mining

Using word frequency to extract meaning from a corpus of literature is a mainstay of text-mining techniques. In terms of genomic analysis, for example, Jenssen and colleagues [JLKH01] used the frequency of co-occurrence of gene names in scientific abstracts to generate a gene-to-gene co-citation network that can be used in the analysis of microarray data. Conversely, others have used literature-mining techniques to asses whether clusters of particular genes share a common biological function [RSA02]. Since literature also constitutes a type of gene-association data, albeit of a more complex kind, it is possible to use overrepresentation-based approaches to mine literature as well.

One example of this strategy uses abstracts from scientific publications and extracted keywords to look for overrepresentation of keywords among publications associated with gene lists (Hong, Liu, Wong, Castillo-Davis, unpublished). In this work, approximately 10 million literature abstracts associated with all genes under analysis were extracted from PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) and filtered for keywords by excluding all non-technical words using a generic dictionary word list. Next, the overrepresentation of keywords in papers associated with the sample set of genes versus the population set was assessed using the hypergeometric distribution. This simple approach was effective when used to examine a set of genes with known enrichment in developmental functions in human (Table 9). This literature-based method generated much more detailed information on gene function and biological sub-processes, than, for instance, GO annotations (data not shown). These included specific gene names (BMP, NOTCH, MYC, NOGGIN), functional regions (enhancer, homeobox), and potentially interesting disease relationships (Parkinsonism).

Keyword	P-value
hif	6.14E-63
myc	8.86E-36
parkin	3.59E-33
gata	1.11E-20
bmp	2.71E-19
morphogenetic	2.46E-17
notch	2.46E-17
eph	3.18E-17
ephrin	8.99E-16
transcription	1.82E-15
malformation	3.90E-15
parkinsonism	8.12E-14
twist	3.96E-12
homeodomain	1.22E-10
hox	1.31E-10
differentiation	1.47E-09
noggin	1.64E-09
developmental	3.90E-09
homeobox	9.25E-09
enhancer	1.44E-08
morphogenesis	5.86E-08

Table 9. Literature-based over-representation results for developmental genes.

Different implementations of literature-based overrepresentation methods along these lines and others [MKS04] are likely to be increasingly useful for extracting biological meaning from genomic data.

5 Limitations and possible solutions

A major drawback of most over-representation methods is the discretization of data into binary categories. In case study II for example, a decision had to be made about what constituted slow and fast evolving genes. The authors chose to test for functional overrepresentation among the 10% slowest and then the 10% fastest evolving genes. While the results of the study were not particularly sensitive to the 10% cut-off (data not shown) such a cut-off may not be desired in other contexts. For example, for other purposes it might be interesting to comprehensively examine the relationship between evolutionary rate and gene function. One could choose *a priori* a set of biological functions and some level of granularity in the GO hierarchy and then calculate rates of evolution of genes in each category. Going a step further one could calculate rates of evolution for genes in every functional category at all levels in the GO hierarchy. While informative, note that interrogation of the data in this manner has moved us from a hypothesis-testing mode to a data-mining mode.

The visualization of the results of such a comprehensive analysis also present a difficult problem. How can one visualize the discrete and structured functional relationships inherent in the GO hierarchy and the continuous evolutionary rate information or other such variables, perhaps several, all at the same time? A bubble graph illustrated in Figure 5 is one possible solution. In this figure, each node represents a particular GO function and edges connect functions in accordance with relationships of the GO function tree (a so-called directed acyclic graph, or DAG). The size of each node indicates the mean rate of amino-acid substitution (d_N) for genes within the node— that is, the rate of evolution of genes with a particular function. In this hypothetical example, genes with known transcription factor (TF) activity exhibit a faster rate of evolution than other types genes.

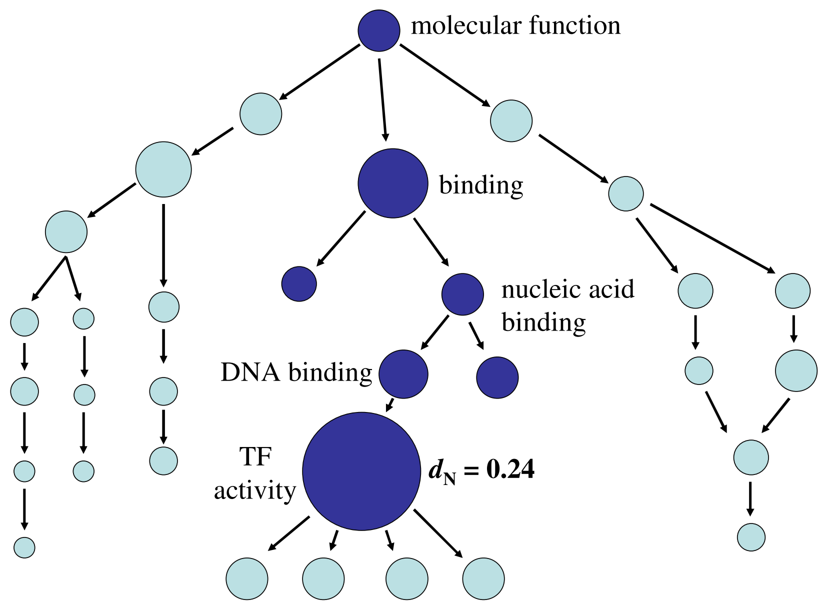


Figure 5: Bubble graph representation of the relationship between a continuous variate (rate of evolution, d_N) and a graphical structure (the GO functional hierarchy).

Note that the number of genes within each node may be different and genes may belong to multiple nodes. Thus, while we have improved our understanding of relationship between evolutionary rate and gene function in the data, we have done so at the expense of statistical power; all possible relationships between evolutionary rate and function have been explored. Such trade-offs are likely to be common and must be weighed by the aims of the study, specifically, whether the ultimate goal is hypothesis-testing or data-mining.

While some gene attributes are discrete, such as on which chromosome a gene resides, others are continuous, such as the expression level of a gene in a particular tissue or its relative level of evolutionary conservation. In these cases the usefulness of categorical statistical tests such as those based on the hypergeometric distribution are called into question. Suppose, for example, that instead of rigidly assigning gene function in a boolean manner, one could assign probabilities concerning gene function to genes. How might one design a test for overrepresentation in this case? While regression-based techniques have been applied on a case-by-case basis to particular problems (for instance [LBL02] there is as of yet no general algorithm available for the interrogation and comparison of disparate data types with continuous values or with a mixture of continuous and categorical values.

The development of such a framework will be challenging, in particular because the categorical structure of some types of biological data can be complex, such as the GO DAG. Interestingly however, directed and undirected graphs are often extremely natural representations of gene functions and gene interactions, for example, protein-protein interaction networks [UGC+00] [ICO+01] [LAB+04] [GBB+03] and metabolic and developmental pathways [KGKN02]. The addition of weights to graph edges or variance measurements for individual nodes will only increase the complexity of analyzing such data. The development of statistical tests and data-exploration methods, perhaps akin to overrepresentation techniques, will be critical in exploiting these types of data.

Equally important in the analysis of disparate genomic data is data visualization. How best to visualize several dimensions of the data simultaneously, some of which may have complex structures? Some overrepresentation-based tools have begun to address this issue by creating dynamic output that maps, for instance, overrepresentation P-values onto the GO hierarchy, for example, the"GO Term Finder" of the Saccharomyces Genome Database [BWG+04] (Figure 6, or the up- or down-regulation of genes onto a metabolic pathway using the KEGG database [KGKN02], for example, Pathway Processor [GTHC02]. Unfortunately, these visualization solutions are species- or gene-association-specific and have not yet been generalized.

A particularly important challenge will be the analysis of high-throughput phenotypic data in combination with genomic data. Phenotypic data, including anatomical sections [oMUBoR90], three-dimensional CT scans, and MR images have even more complex structures than graphs and network diagrams. Incorporating genomic and proteomic data with the mixture of continuous and discontinuous spatial information inherent in morphological data will be challenging. Flexible visualization and statistical techniques that allow for the input and processing of standardized structural information (in the form of graphs, network diagrams, or 3-dimensional volumes) along with the requisite gene lists and relevant gene-association data is desperately needed. Contribution from many different disciplines, including computer imaging, scientific visualization, and biostatistics, especially areas related to morphometrics, will be required to achieve a comprehensive understanding of how genotype and organism phenotype are related.

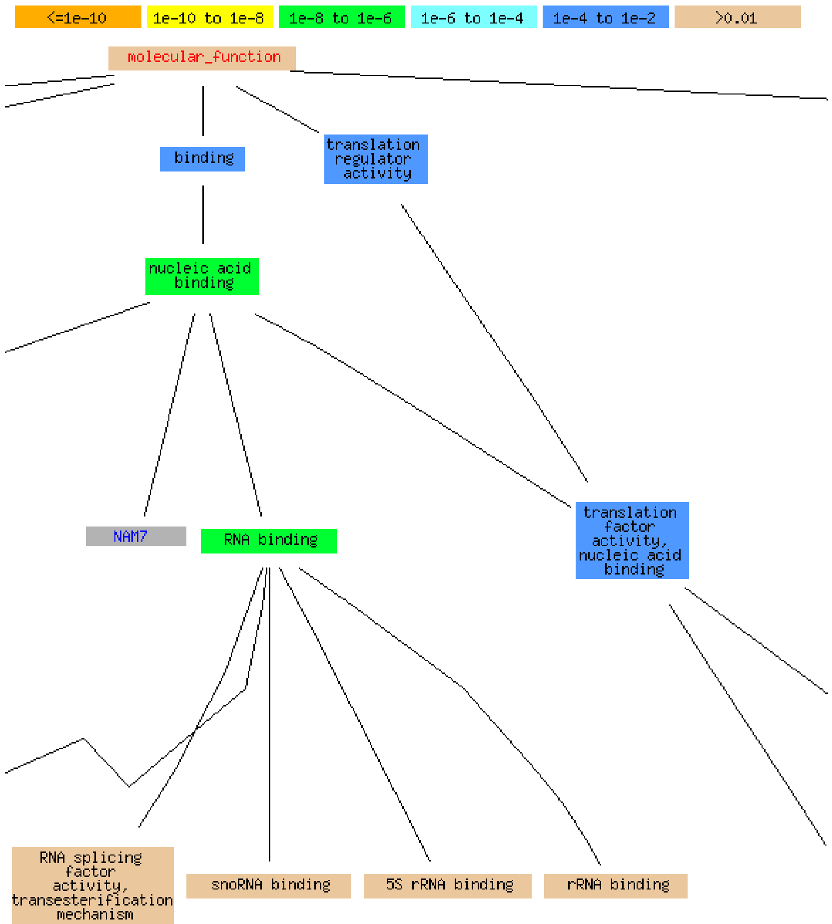


Figure 6: Partial graphical output of the SGD GoTermFinder [BWG+04].

6 Summary

Despite an abundance of genomic, proteomic, and increasingly, gross phenotypic data, many straightforward biological questions remain difficult to answer due to the complex and varied nature of these data. As we have seen, overrepresentation techniques and related methods, when applied creatively and critically, hold some promise in helping shed light on this tangled surfeit of biological information. In particular, the use of more and varied gene-association data with these methods promises to be quite powerful for data-mining and cursory hypothesis testing applications. At the same time, the limitations of these approaches are many; highly structured data in the form of gene networks, morphological data, protein-protein interactions, and simply the growing dimensionality of biological measurements in genome-wide studies strain the conceptual and statistical limits of the overrepresentation framework. Many challenges remain in assimilating complex biological data structures into current statistical and data-mining approaches. Data visualization will be an additional challenge. Progress will likely require heavy cross-disciplinary collaboration amongst statisticians, biologists, and computer scientists, among others. The expansion and application of statistical and graphical approaches to the analysis of genomic data presents numerous, rich opportunities for intellectual contribution. With luck, these advances will help expedite the larger goal of deciphering nature's profound complexity.

7 Acknowledgements

I wish to thank Jun S. Liu for support during my post-doctoral fellowship at Harvard, Pengyu Hong for collaborating on the literature mining work discussed, Bjarki Eldon, Gordon Kindlmann, and Rima Izem for help with LaTeX, and Ping Ma, Lei Shen, Gordon Kindlmann, Christian Landry, and Rima Izem for valuable comments on the manuscript.

References

[ABB+00] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.

[AMS+97] S F Altschul, T L Madden, A A Schaffer, J Zhang, Z Zhang, W Miller, and D J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, Sep 1997.

[BD99] S Bortoluzzi and G A Danieli. Towards an in silico analysis of transcription patterns. *Trends Genet*, 15(3):118–119, Mar 1999.

[BH95] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995.

[BKML+05] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and David L Wheeler. GenBank. *Nucleic Acids Res*, 33(Database issue):34–38, Jan 2005.

[BWG+04] Elizabeth I Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J Michael Cherry, and Gavin Sherlock. GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, Dec 2004.

[CDH03] Cristian I Castillo-Davis and Daniel L Hartl. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, May 2003.

[CDKHK04] Cristian I Castillo-Davis, Fyodor A Kondrashov, Daniel L Hartl, and Rob J Kulathinal. The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint. *Genome Res*, 14(5):802–811, May 2004.

[CDMH+02] Cristian I Castillo-Davis, Sergei L Mekhedov, Daniel L Hartl, Eugene V Koonin, and Fyodor A Kondrashov. Selection for short introns in highly expressed genes. *Nat Genet*, 31(4):415–418, Aug 2002.

[CTH00] D Cavalieri, J P Townsend, and D L Hartl. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. *Proc Natl Acad Sci U S A*, 97(22):12369–12374, Oct 2000.

[GBB+03] L Giot, J S Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, Y L Hao, C E Ooi, B Godwin, E Vitols, G Vijayadamodar, P Pochart, H Machineni, M Welsh, Y Kong, B Zerhusen, R Malcolm, Z Varrone, A Collis, M Minto, S Burgess, L McDaniel, E Stimpson, F Spriggs, J Williams, K Neurath, N Ioime, M Agee, E Voss, K Furtak, R Renzulli, N Aanensen, S Carrolla, E Bickelhaupt, Y Lazovatsky, A DaSilva, J Zhong, C A Stanyon, R L Jr Finley, K P White, M Braverman, T Jarvie, S Gold, M Leach, J Knight, R A Shimkets, M P McKenna, J Chant, and J M Rothberg. A protein interaction map of Drosophila melanogaster. *Science*, 302(5651):1727–1736, Dec 2003.

[GCN+02] Guri Giaever, Angela M Chu, Li Ni, Carla Connelly, Linda Riles, Steeve Veronneau, Sally Dow, Ankuta Lucau-Danila, Keith Anderson, Bruno Andre, Adam P Arkin, Anna Astromoff, Mohamed El-Bakkoury, Rhonda Bangham, Rocio Benito, Sophie Brachat, Stefano Campanaro, Matt Curtiss, Karen Davis, Adam Deutschbauer, Karl-Dieter Entian, Patrick Flaherty, Francoise Foury, David J Garfinkel, Mark Gerstein, Deanna Gotte, Ulrich Guldener, Johannes H Hegemann, Svenja Hempel, Zelek Herman, Daniel F Jaramillo, Diane E Kelly, Steven L Kelly, Peter Kotter, Darlene LaBonte, David C Lamb, Ning Lan, Hong Liang, Hong Liao, Lucy Liu, Chuanyun Luo, Marc Lussier, Rong Mao, Patrice Menard, Siew Loon Ooi, Jose L Revuelta, Christopher J Roberts, Matthias Rose, Petra Ross-Macdonald, Bart Scherens, Greg Schimmack, Brenda Shafer, Daniel D Shoemaker, Sharon

Sookhai-Mahadeo, Reginald K Storms, Jeffrey N Strathern, Giorgio Valle, Marleen Voet, Guido Volckaert, Ching-yun Wang, Teresa R Ward, Julie Wilhelmy, Elizabeth A Winzeler, Yonghong Yang, Grace Yen, Elaine Youngman, Kexin Yu, Howard Bussey, Jef D Boeke, Michael Snyder, Peter Philippsen, Ronald W Davis, and Mark Johnston. Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, 418(6896):387–391, Jul 2002.

[GTHC02] Paul Grosu, Jeffrey P Townsend, Daniel L Hartl, and Duccio Cavalieri. Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. *Genome Res*, 12(7):1121–1126, Jul 2002.

[GY94] N Goldman and Z Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11(5):725–736, Sep 1994.

[HF01] A E Hirsh and H B Fraser. Protein dispensability and rate of evolution. *Nature*, 411(6841):1046–1049, Jun 2001.

[HVD01] J Herrero, A Valencia, and J Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–136, Feb 2001.

[ICO+01] T Ito, T Chiba, R Ozawa, M Yoshida, M Hattori, and Y Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574, Apr 2001.

[JLKH01] T K Jenssen, A Laegreid, J Komorowski, and E Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21–28, May 2001.

[KGKN02] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30(1):42–46, Jan 2002.

[LAB+04] Siming Li, Christopher M Armstrong, Nicolas Bertin, Hui Ge, Stuart Milstein, Mike Boxem, Pierre-Olivier Vidalain, Jing-Dong J Han, Alban Chesneau, Tong Hao, Debra S Goldberg, Ning Li, Monica Martinez, Jean-Francois Rual, Philippe Lamesch, Lai Xu, Muneesh Tewari, Sharyl L Wong, Lan V Zhang, Gabriel F Berriz, Laurent Jacotot, Philippe Vaglio, Jerome Reboul, Tomoko Hirozane-Kishikawa, Qianru Li, Harrison W Gabel, Ahmed Elewa, Bridget Baumgartner, Debra J Rose, Haiyuan Yu, Stephanie Bosak, Reynaldo Sequerra, Andrew Fraser, Susan E Mango, William M Saxton, Susan Strome, Sander Van Den Heuvel, Fabio Piano, Jean Vandenhaute, Claude Sardet, Mark Gerstein, Lynn Doucette-Stamm, Kristin C Gunsalus, J Wade Harper, Michael E Cusick, Frederick P Roth, David E Hill, and Marc Vidal. A map of the interactome network of the metazoan C. elegans. *Science*, 303(5657):540–543, Jan 2004.

[LBL02] X Shirley Liu, Douglas L Brutlag, and Jun S Liu. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–839, Aug 2002.

[MBR+04] David Martin, Christine Brun, Elisabeth Remy, Pierre Mouren, Denis Thieffry, and Bernard Jacq. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, 5(12):R101, 2004.

[MKS04] Hans-Michael Muller, Eimear E Kenny, and Paul W Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2(11):e309, Nov 2004.

[NY98] R Nielsen and Z Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, Mar 1998.

[oMUBoR90] National Library of Medicine (U.S.) Board of Regents. Electronic imaging: Report of the board of regents. U.S. department of health and human services, public health service, national institutes of health. NIH Publication 90-219, 1990.

[PMM+02] Scott D Pletcher, Stuart J Macdonald, Richard Marguerie, Ulrich Certa, Stephen C Stearns, David B Goldstein, and Linda Partridge. Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster. *Curr Biol*, 12(9):712–723, Apr 2002.

[RCDMH03] Jose M Ranz, Cristian I Castillo-Davis, Colin D Meiklejohn, and Daniel L Hartl. Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science*, 300(5626):1742–1745, Jun 2003.

[RSA02] Soumya Raychaudhuri, Hinrich Schutze, and Russ B Altman. Using text analysis to identify functionally coherent gene groups. *Genome Res*, 12(10):1582–1590, Oct 2002.

[Sto02] John D Storey. A direct approach to false discovery rates. *J. R. Statist. Soc. B*, 64:479–498, 2002.

[UGC+00] P Uetz, L Giot, G Cagney, T A Mansfield, R S Judson, J R Knight, D Lockshon, V Narayan, M Srinivasan, P Pochart, A Qureshi-Emili, Y Li, B Godwin, D Conover, T Kalbfleisch, G Vijayadamodar, M Yang, M Johnston, S Fields, and J M Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403(6770):623–627, Feb 2000.

[WCDO+03] Elizabeth A Winzeler, Cristian I Castillo-Davis, Guy Oshiro, David Liang, Daniel R Richards, Yingyao Zhou, and Daniel L Hartl. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics*, 163(1):79–89, Jan 2003.

[WSA+99] E A Winzeler, D D Shoemaker, A Astromoff, H Liang, K Anderson, B Andre, R Bangham, R Benito, J D Boeke, H Bussey, A M Chu, C Connelly, K Davis, F Dietrich, S W Dow, M El Bakkoury, F Foury, S H Friend, E Gentalen, G Giaever, J H Hegemann, T Jones, M Laub, H Liao, N Liebundguth, D J Lockhart, A Lucau-Danila, M Lussier, N M'Rabet, P Menard, M Mittmann, C Pai, C Rebischung, J L Revuelta, L Riles, C J Roberts, P Ross-MacDonald, B Scherens, M Snyder, S Sookhai-Mahadeo, R K Storms, S Veronneau, M Voet, G Volckaert, T R Ward, R Wysocki, G S Yen, K Yu, K Zimmermann, P Philippsen, M Johnston, and R W Davis. Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, Aug 1999.