Abstract

Mining Biological Complexity: Cross integration of large-scale metagenomics, environmental, and chemical datasets

Tara Ann Gianoulis

2009

Annotation of complete genomes, community analysis of entire ecosystems (metagenomics), and comparative analysis of regulatory networks from multiple species, each of these experiments is emblematic of the high throughput data that is radically altering the scientific landscape. Moreover, so-called next generation sequencing has significantly increased the scope of questions being asked through sequencing making it crucial to understand how to interpret, decode, and integrate sequence data. Although each assay can provide only snapshots of the genes or proteins, through integration of multiple features across different conditions, time points, and species, the goal is to extract the dynamics from these static images and derive their emergent properties. Current integration schemas are constrained to single dimensional features and do not have the flexibility to integrate features not centered solely on genes or proteins. Here, we have developed a new type of integration, cross integration, where the goal is to integrate not to stack gene and protein features in a single dimension but to build spanning relationships (cross patterns) across multidimensional ones. We showed that fusing geography and metagenomics could illuminate microbial adaptations to environmental differences. We identified a number of metabolic components that co-vary with specific environmental features, which we term a metabolic footprint. Further, we speculate that analysis of these environmental dynamics could be used as a sensitive biosensor to detect chemical or other environmental perturbations. In addition, we developed a new formalism both to express and define cross integration and apply it to chemogenomics data. In this manner, we were able to

identify cross patterns between properties of drugs and their protein targets. Some of these were intuitive, such as the mirroring of physicochemical properties between drug and target, and others were subtler such as sensitivities to both environmental stress response and particular drug properties. Mining such biological complexity requires a robust infrastructure and new computational models. We have explored several methods to uncover subtle, indirect relationships between multidimensional features; many exciting discoveries remain.

Mining Biological Complexity: Cross integration of large-scale metagenomics, environmental, and chemical datasets

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Tara Ann Gianoulis

Dissertation Director: Mark Gerstein and Michael Snyder

May 2009

Dedicated to my grandfather

Charles P. Massa

in memoriam

Brief Contents

Detailed Contents

List of Figures

List of Tables

Acknowledgements

If "science is nothing more than long periods of boring conformist activity punctuated by outbreaks of irrational deviancy," then both of my advisors Mark Gerstein and Michael Snyder have mastered the cultivation of such deviancy. The sequencing of *A baumannii*, an amalgam of untested technology, important question, and thoroughly intimidating task, are emblematic of the types of projects we were encouraged to tackle. I feel truly privileged to have been a part of them.

Many thanks to my experimental collaborators: Michael Smith on *A. baumannii* and Anthony Borneman on the TF project and to my fellow denizens of late-night Nets, Andrea Sboner, Kevin Yip, Philip Kim, Haiyuan Yu, Prianka Patel, and Alberto Paccanaro and to Paul Bertone, Joel Rozowsky, Jan Korbel, Ashish Agarwal, Nicholas Carriero, Robert Bjornson, and Daniel Gelperin. Yale is a wonderful, supportive place to do science. Thanks to Nick Ornston, Avi Silberschatz, and Martin Schultz for many brainstorming sessions and to Peer Bork's group at EMBL for their help on the metagenomics project.

I would like to extend a special thank you to my CBB family, David Ballard, Kevin Keating, Jill Rubinstein, and Sara Nichols and to my R2C teammates: Brad Stanley, Robin Evans, Matthew Calabrese, Adam Fogel, Nathan Kucera, and Mary Stahley. When I found people as excited as me to run across NJ in August with a team name like PolymeRace, I knew I belonged. I would also like to thank Whitman Schofield and John Corwin for keeping me sane by and through miles of pedaling and to my parents, sister, cousin Nina, and long-time friend Kristy for their unflagging support.

I would like to dedicate this work to my grandfather Charles P. Massa. I was never happier than when watching him tinker in his workshop. His desire to take things apart to truly understand how they work and the joy derived from following that curiosity inspires me still.

Chapter 1

Introduction

1.1 Evolution of Computational Biology

1.1.1 A brief etiology of 'ome'

Despite its near ubiquity in referring to collections of large-scale biological data, the suffix *ome* has no discrete meaning and has only a loose claim to Greek roots through the coining of the word *chromosome*, derived from the Greek words for color and body, respectively. However, the, in some cases, overenthusiastic adaptation of *ome* by the biological community provides an apt example of the recent evolution taking place in the biological sciences.

1.1.2 Bridging disciplines and really big data

Technological advancements have resulted in a dramatic increase in the scale of biological data making it infeasible in many cases for a single scientist or even a group of scientists to interpret it. Thus, from database design to signal processing, machine learning to metadata, expertise from computer science, applied math, physics, statistics and engineering alike are being harnessed to decipher meaning from a virtual flood of the data. As a result, biological and computational vocabularies are undergoing an expansion, and context is often required to disambiguate words like *complexity*, a generic description of something with many inter-

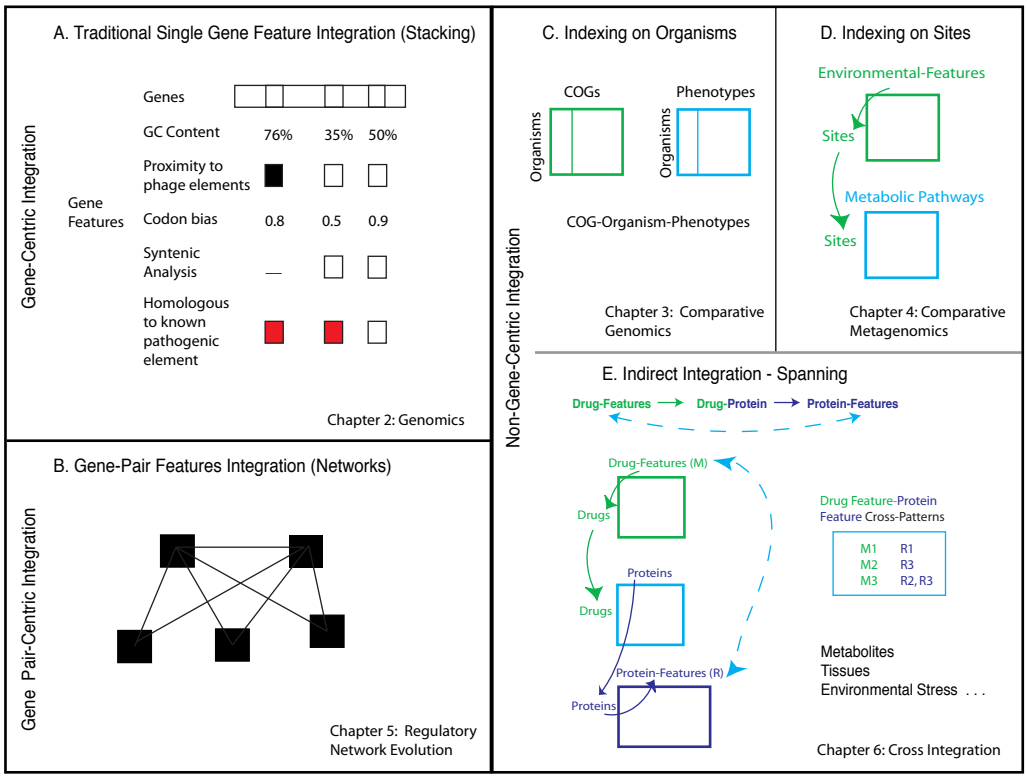


Figure 1.1: Graphical representation of major themes. (A) Identifying determinants of pathogenicity through integration of multiple genomic features. (B) Regulatory network evolution. (C) Prediction of phenotype through comparative genomics. (D) Integrating environmental and metabolic features to identify environmental adaptation in metagenomics datasets. (E) Spanning across differently indexed datasets in chemogenomics.

related parts or a more precise term in computer science used to characterize the amount of time or space required to execute a given algorithm. The increasing computational and biological complexity of the questions being asked requires a new discipline with the flexibility to bridge many. This in essence is the goal of computational biology to develop computational frameworks in the context of specific biological problems (Pevzner, 2004; Luscombe et al., 2001).

1.2 Scope of Dissertation: Two gradients of biological and computational complexity

The remainder of this text will present themes in computational biology along two gradients of biological and computational complexity: Genomics to Metagenomics (Figure 1.3) and Single to Multidimensional data integration (Figure 1.1). In chapter 2 Genomics, we describe our work on the first *de novo* sequencing of a genome using 454 sequencing and the identification of potential pathogenicity factors. In chapter 3 Comparative Genomics, we discuss a method using comparative genomics to associate microbial genotypes with specific phenotypic traits. In chapter 4 Network Dynamics, we move from comparisons of single microbial genomes to the comparison of entire microbial communities.

In this chapter, we develop a new algorithm DPM (discriminative partition matching) and adapt several computational methods including canonical correlation analysis (CCA) to identify pathways that showed strong co-variation with environmental features. We coin these environmental footprints. Further, we show that such footprints can be used to infer environmental adaptation of microbial metabolic pathways.

In chapter 5 Network Evolution, we explore the genetic basis for species variation by looking at changes in transcription factor binding in close yeast. In chapter 6 Cross Integration, we present an extension of the DPM algorithm presented in chapter 4 to allow for integration of data with different types of indices. We call this the ITeR algorithm (Identifying Transitive Relationships). We both formalize ITeR more generally and apply the method to search for relationships between sets of drug features and sets of protein features in chemogenomics datasets. Finally in chapter 7, we present a brief future outlook.

The rest of the introduction is structured as follows. Section 1.3 provides an overview of next-generation sequencing technologies and the anatomy of a genome sequencing project. Section 1.4 includes a brief literature survey to introduce metagenomics. Finally, section 1.5 motivates the rationale for cross integration (chapter 6) whose development was a natural outgrowth of the work presented in chapters 4 and 5.

1.3 One to Many: Genomics and Comparative Genomics

From the duck-billed platypus (Platypus, 2008) to the geyser-dwelling *Thermotoga maritima* (Nelson et al., 1999a), the genomes of 4700 species are currently available in NCBI's Entrez Genome Database (as of Sept 2008) including representatives of over 700 microbial species (Benson et al., 2008). Data generated from genome sequencing projects have proven themselves to be enormously versatile providing a wealth of insight into life style (e.g. the radiation-resistant bacteria *Deinococcus radiodurans R1*)(White et al., 1999), pathogenicity (Smith et al., 2007), evolutionary history (Dufresne et al., 2003), and metabolic capabilities (e.g. the cellulose degrading, ethanol producing fungus, *Trichoderma reesei*)(Martinez et al., 2008) as just a small sample of the enormous diversity of sequenced organisms.

1.3.1 Next generation sequencing technologies

Despite the manifold advantages of whole genome sequencing, the costliness of such endeavors has historically limited the execution of these types of projects to large sequencing centers and deep-pocketed consortiums. However, the advent of so-called next-generation sequencing platforms, such as, 454 pyrosequencing, Solexa, and SOLiD, has led to a more than exponential increase in sequencing capacity and a significant drop in cost opening the possibility of whole genome sequencing to single investigators and smaller groups (Mardis, 2008; Morozova and Marra, 2008). Indeed, since the introduction of such platforms in 2005, approximately 170 genomes including a range of eukaryotes (e.g. *C. remanei*, *D. mauritiana*, etc.) and microbial species (e.g. *A. bauamannii*, *B. thailandensis*, etc.) have been determined using next-generation sequencing platforms and more specifically (for whole genome sequencing) 454 pyrosequencing (Pop and Salzberg, 2008) and deposited in NCBI. In brief, in 454 pyrosequencing, genomic DNA fragments are clonally amplified in picotiter plates. The fragments are then "sequenced by synthesis" whereby the energy released when a base is incorporated is converted through a series of

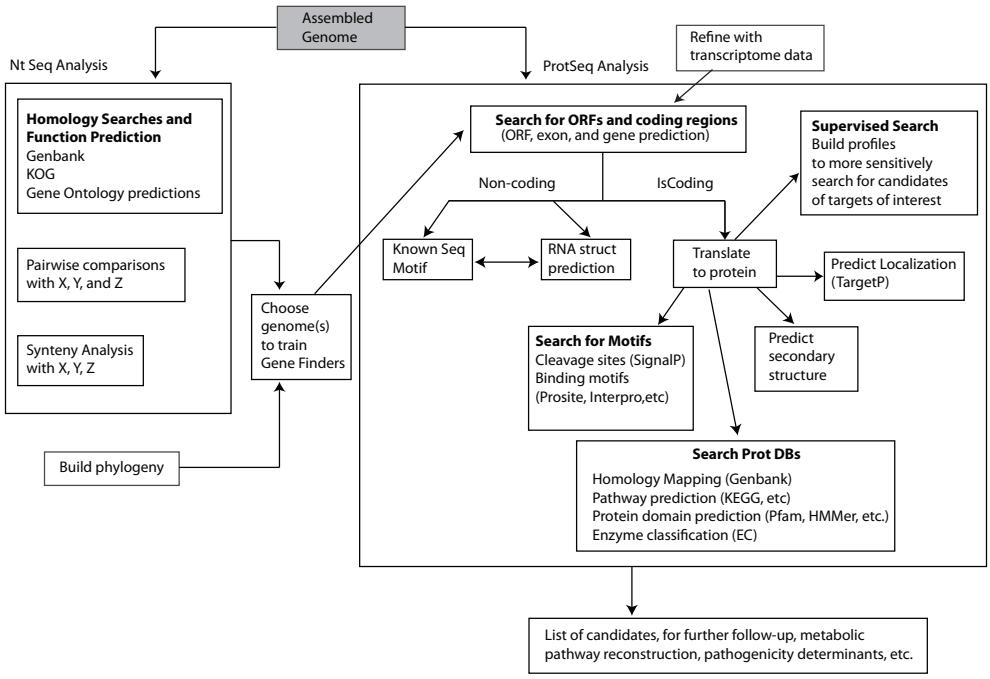


Figure 1.2: General annotation scheme, see text for details

enzymatic reactions to light. The intensity of which corresponds to the number of bases incorporated (Margulies et al., 2005). This technique, as do the other next generation technologies, results in massive parallelization and eliminates the need for cloning.

1.3.2 Assembly and Annotation

Genome sequencing projects can be split into two pieces: assembly and annotation. The raw output from a sequencing project is strings of contiguous letters called reads. As the genome was randomly sheared, the goal of assembly is to stitch together overlapping reads into contiguous blocks called contigs. The contigs can similarly be stitched together to form scaffolds. This process is often aided by using a paired end approach. The goal then is two-fold: (1) to identify particular features of interest (e.g. genes, non-coding RNAs, etc.) and (2) to provide a functional assignment for these features.

1.3.3 Three Levels: Genome composition, genome content, and comparative genomics

Such analyses can be further divided into three basic categories: genome composition, genome content, and comparative genomics (Figure 1.2). Genome composition measures properties such as GC content, codon usage, and amino acid bias across the genome and searches for deviations from what is expected(Karlin, 2001). Examined on a genome-wide scale, these features can elucidate genic structure (e.g. changes in GC content can be used to identify gene as well as exon/intron boundaries, etc.). More local changes can be used to predict secondary structure elements (e.g. identification of putative transporters) (Krogh et al., 2001). Genomic content analysis refers to the particular genes, non-coding RNAs, etc. that are encoded in the genome. Gene finding programs traditionally use hidden markov models to calculate the probability of a particular region being genic or not (Majoros et al., 2004). In addition to using genome composition statistics as above, gene finders are trained using a genome where the genic structure is already known (Salzberg et al., 1998). Thus, the performance of the gene finder is dependent on the evolutionary divergence of the training genome with new genome.

Finally, comparative genomics allows one to compare features of the new genome with other genomes through both synteny mapping (looking at conservation of gene order) and homology searches. These types of analyses have been shown to be extremely powerful, and a host of tools and databases have been developed to harness the collective knowledge derived from these millions of base pairs. Indeed, in the simplest case even the overall enrichment and depletion of particular functional categories, (e.g. those involved in lipid metabolism and secondary metabolite production) relative to other species that do not have the same capabilities or present the same phenotype has previously been informative in identifying candidates of a function of interest (Smith et al., 2007).

1.3.4 Unbiased vs. Targeted Search

All of the above approaches are considered unbiased. They either use information derived from the newly sequenced genome directly (such as composition statistics) or indirectly by comparison with other species. A complementary approach to unbiased search is to perform targeted searches. That is, to develop a list of candidates of potential interest in the process of interest and build models of what those targets look like in a host of other fungi. By building position weight matrices or hidden Markov models (HMM) from multiple alignments of these protein families, one can search the predicted ORFs to identify candidates of the process of interest that may have more distant homologies(Sonnhammer et al., 1998). Thus, one can leverage what is already known about the domains and other secondary structure characteristics of the targets of interest to search for candidates that based purely on amino acid identity maybe overlooked. This complementary analysis allows among others, the identification of novel metabolic components for in-silico metabolic pathway reconstruction or pathogenic determinants. By harnessing these three levels of analyses: genome composition, genome content, and comparative genomics in both a targeted and unbiased fashion, one can both learn about an organism's overall metabolic requirements and more specifically identify candidates for further experimentation that may explain observed phenotypes, morphologies, or metabolic capabilities. In addition, the approach described above is general enough for both prokaryotes and eukaryotes. We used such an approach, as described further in chapter 2, to perform the first De novo sequencing of a genome using 454 sequencing.

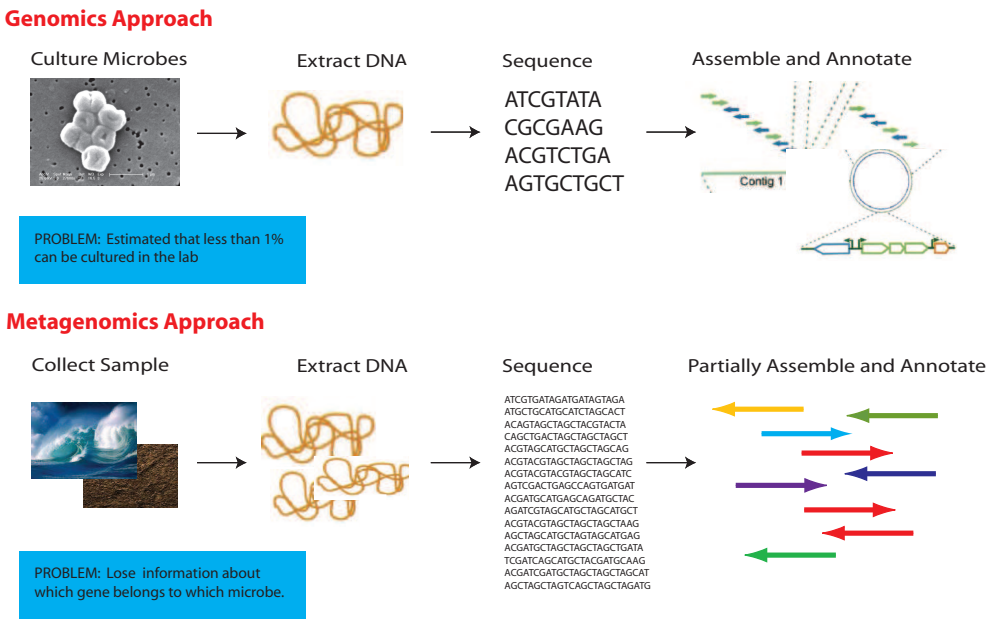


Figure 1.3: Comparison of Traditional Genomic and Metagenomics Approaches

1.4 Many to Even More: Metagenomics and Comparative Metagenomics

1.4.1 Introduction to Metagenomics

The approach described above can be considered the traditional genomics approach. A particular organism of interest is identified (e.g. in our case the bacteria *A. bauamannii*). It is cultured, DNA is extracted, and sequencing is performed on the single organism of interest (Figure 1.3). However, it is estimated that less than 1% of microbial species can be cultured under standard laboratory conditions (Whitman et al., 1998). Recently approaches have been developed to sample the genetic content in complex environments using sequencing (metagenomics). The power of metagenomics is that it is culture-independent. Almost anything from buckets of sea water (Venter et al., 2004) to spadefulls of soil (Tringe et al., 2005) to scoops of distal gut (Turnbaugh and Gordon, 2008; Turnbaugh et al., 2007) even flasks of air (Tringe et al., 2008) can serve as fodder for metagenomics projects. In this manner, one can begin to look at microbial communities as a whole rather

than studying them in isolation, which has profound implications particularly for microbial ecology.

1.4.2 Introduction to microbial ecology and comparative metagenomics

Marine microbes are estimated to account for 90% of the ocean's biomass (Sogin et al., 2006). However, the effect of microbial processes on their environment is perhaps even more staggering. Microbes are the main drivers of nutrient recycling and all geochemical cycles (Karl, 2002). Furthermore, while most marine habitats are oligotrophic (low in nutrients), they support a tremendous diversity of life. Traditionally, community function has been studied through cataloguing the environmental species diversity, typically by analyzing 16S rDNA sequence, or through detailed studies of the molecular processes exerted by only specific community members. However, unless the species identified have been previously studied, the overall biochemical functions and activities of both the individual species and the community as a whole remain unclear. In addition, even in well-studied species, function cannot always be directly extrapolated from phylogenetic classification (Sogin et al., 2006). Indeed, the recent advent of direct sequencing of environmental samples (i.e. metagenomics) has revealed an unprecedented mixture of both inter and intra-species genetic diversity (DeLong et al., 2006; Liles et al., 2003; Thompson et al., 2005; Tyson et al., 2004; Rusch et al., 2007) allowing the biochemical activities of the community as a whole to be inferred through its genetic content. These activities have been found to vary greatly from one environmental condition to another. However, how this diverse array of biochemical activities, and particularly metabolic versatility, reflects environmental differences is only beginning to be understood.

Comparative metagenomics approaches revealed significant variation in sequence composition (Foerstner et al., 2005), genome size (Raes et al., 2007), evolutionary rates (von Mering et al., 2007a), and metabolic capabilities (Tringe et al., 2005; DeLong et al., 2006; Tyson et al., 2004; Dinsdale et al., 2008) among qualitatively dissimilar environments (e.g. terrestrial vs. marine) providing evidence for genomic adaptations. Despite the wealth of

information generated by these studies, a quantitative description relating metagenomics sequences to the myriad of environmental features has generally been lacking. Such information is vital for relating specific environmental features to genomic sequences and thereby uncovering complex relationships for how microbes adapt to their environment.

Previously such an analysis was infeasible as in all of the above work (with the notable exception of (DeLong et al., 2006)), the datasets used provided no quantitative description of environmental variables. That is, the qualitative label of the environments (e.g. soil, ocean, etc.) could be used as an implicit means of separating them since no specific quantitative features of the environment were measured, nor since most of the environmental variables were changing simultaneously would such a study be relevant.

In order, to understand the contribution of a particular feature, the environments sampled must differ along a continuum. Venter and colleagues built such a dataset with the Global Ocean Survey (Rusch et al., 2007; Yooseph et al., 2007). By mining this dataset, it was shown that the well-known adaptation of microbes to different wavelengths of light could be explained by looking at specific amino acid changes of rhodopsin and further that such changes were dependent on the type of environment (Yooseph et al., 2007).

This example was found somewhat serendipitously. The ubiquity of rhodopsin in the ocean allowed for the high coverage of this particular family that is necessary in order to build multiple alignments and do this kind of in-depth analysis. The linking of specific metabolic capabilities to the environment is needed to understand oceanic processes and in particular human influence on them; however, this will require a methodological framework, which can systematically and explicitly account for quantitative differences in environments.

In chapter 4, we developed such a rigorous statistical approach for quantifying environmental adaptation to metabolic pathways through the integration of quantitative differences in both metabolism and the environment (Gianoulis et al., 2009b). We then applied this methodology to marine metagenomics datasets. The aim of this portion of the work was three-fold: (1) to develop a framework for explicitly integrating environmental

metadata with sequence data, (2) to use this framework in order to establish whether such a relationship exists, and (3) to explore (if goal 2 is true) what the major contributing factors to this relationship are.

1.5 Simple to Complex: Multidimensional data integration

1.5.1 "Stacking vs Spanning"

Each of these projects required a different schema for integration (Figure 1.1), and although there is a large corpus of work in the literature on mining gene or protein-based datasets, our questions required the integration of data not always indexable on a gene. Previous approaches revolve around a "gene or protein centric" view where individual datasets can be conceived of as data layers (Figure 1.1A), and positions within the layer are determined by referencing a gene or protein. In other words, the gene or protein (or pairs of genes or proteins) serves as the index to the individual data level and integration is then merely a matter of stacking the data levels via their index.

Despite the seeming simplicity of the concept, implementations of "stacking" have taken many non-trivial forms including functional coupling (Fraser and Marcotte, 2004), phylogenetic profiling (Marcotte et al., 1999; Pellegrini et al., 1999) and various machine learning approaches including decision trees (King et al., 2003), Bayesian networks, (Jansen et al., 2003; Troyanskya et al., 2003), unsupervised approaches (Flaherty et al., 2005; Bergmann et al., 2003), and many different kinds of kernel methods (Ben-Hur and Noble, 2005; Lanckriet et al., 2004; Tsuda and Noble, 2004). This type of integration has led to discoveries including general principles to predict gene essentiality (Seringhaus et al., 2006) to mechanisms involved in arsenic resistance (Kelley and Ideker, 2005) and DNA damage (Haugen et al., 2004; Begley et al., 2004) among many others. Further, by comparing these "stacks", it has been shown that genes or proteins that share similar properties (e.g. protein interaction partners) tend to share similar functional roles (Kelley and Ideker, 2005; Tasan et al., 2008; Parsons et al., 2004; Wong et al., 2004).

The major theme of "stacking" techniques is that the features are "indexable" by a single class of variables: gene or protein or pairs of genes or proteins (Figure 1.1A-B). This is an intuitive solution when all the data being stacked are of the same type and can thus be treated in a similar manner (e.g. stored in the same relational table and queried directly). The problem lies in capturing connections between associated metadata (e.g. structural properties of a drug and features of a protein that drugs targets); however, current data integration schemas lack the the flexibility to accommodate data that are not all indexed on the same type of variable (e.g. metabolites, tissues, and environmental conditions). Uncovering these kinds of indirect, complex connections requires the facility to systematically combine information from multiple tables, allowing one to not only stack features in a single dimension but also to span across multiple ones.

1.5.2 Illustrative Example

As an example, the data in the previous section can naturally be represented as two matrices where the rows are geographic locations (sites) and the columns are environmental and metabolic features, respectively. It is helpful to think of these as two relational tables where the index for both tables is the site name. In the metagenomics example, we then developed and adapted different techniques to inter-relate these matrices allowing us to infer relationships between environmental pressures and metabolic adaptations.

An obvious criticism of this analysis is *why not combine both types of information into a single relational table since they share the same index?* Intuitively, this would result in the simplification of the scheme necessary to answer the question posed above. However, there is no means of extracting information about sets of environmental features and sets of metabolic features from such a table (Please note: this example is used for illustrative purposes the question of data storage is outside the scope of this thesis; please see (Han and Kamber, 2000) for a review of multidimensional data storage, data warehousing, and slicing). In other words, there is no direct query that will return a set of environmental and metabolic features (columns) without first explicitly providing a set of sites (rows).

Thus, we experimented with several methods for integrating environment and metabolism including aggregating over the sites (partitioned by some type of labeling) to abstract *site-set* relationships (discriminative partition matching) and defining a change of basis to effectively create a new unified environmental-metabolic space (canonical correlation analysis) that would allow us to investigate these relationships further. In both instances, the only use of the index was to span across the other dimensions.

Many biological problems can be conceived in a similar manner as 2 or more matrices where the goal is to develop either a partitioning function or to define a change of basis to allow for integration. In earlier work, we described a simpler correlation method to associate microbial phenotypes with their corresponding genotypes (chapter 2). We provide a more general formalism for the case of more than two matrices (which do not share the same index) in chapter 6.

In this chapter, we present a method Identifying Transitive Relationships (ITeR) that uses the principle of transitivity to seamlessly integrate datasets with non-gene or protein centric indices. We apply this method to identify relationships spanning structural properties of a drug (e.g. molecular weight) and features of the target protein (e.g. target's localization). By integrating 1194 drug sensitivity profiles, six types of structural features, and seven types of target features including physicochemical properties, gene composition features, network topology statistics, localization, function and process, and environmental stress response, we identify numerous drug-feature target-feature relationships, which we term cross-patterns.

Some of these cross-patterns are intuitive (e.g. the charge of a drug and its target are complementary); however, we also find more subtle, less obvious cross-patterns (e.g. target's which were both sensitive to a particular environmental stress and a particular drug feature). Such connections suggest that there may be a set of physical properties underlying common stress responses.

Although currently, yeast represents a special case in terms of the depth and breadth of available system-wide data, this presages the considerable scale-up to humans and other

model organisms. Mining such complexity represents an exciting challenge, chapter 5 presents a more flexible data integration scheme, ITeR, that can be used to identify such indirect, complex relationships.

Chapter 2

Genomics: The first *de novo* sequencing of a microbial genome with next generation sequencing

2.1 Background on *A. baumannii*

Acinetobacter baumannii is a gram-negative, non-motile, obligate aerobic coccus that is commonly found in soil, water, sewage and in healthcare settings (Baumann et al. 1968a; Juni 1978). Difficulties in containing, controlling and eliminating the spread of *A. baumannii* have challenged clinicians and healthcare providers (Bergogne-Berezin and Towner 1996; Bernards et al. 2004; Koulenti and Rello 2006). Recently, drug-resistant *A. baumannii* was responsible for an outbreak of bacteremia in over 240 American troops in Iraq ((CDC) 2004; Abbott 2005), and there is significant concern of a major epidemic involving this organism. This versatile organism can utilize a variety of carbon sources and is able to grow in a range of temperatures and pH conditions (Juni 1978). La

Scola and Raoult (La Scola and Raoult 2004) isolated *A. baumannii* from human body lice and speculate that the bacteria may utilize the arthropod host as one means of transmission. This hardiness, combined with its intrinsic resistance to many antimicrobial agents, contributes to the organism's fitness and has enabled it to thrive in hospital settings worldwide. Mortality in patients suffering *A. baumannii* infections, can be as high as 75% (Chastre and Trouillet 2000).

Alarmingly, little is known about the virulence, antibiotic resistance or persistence strategies of *A. baumannii*. The pathogenic determinants which have been reported thus far for *A. baumannii* include a novel pilus assembly system involved in biofilm formation (Tomaras et al. 2003), an outer membrane protein (Omp38) which causes apoptosis in human epithelial cells (Choi et al. 2005) and a polycistronic siderophore-mediated iron-acquisition system conserved between *A. baumannii* and *Vibrio anguillarum* (Dorsey et al. 2003; Dorsey et al. 2004). This presumably comprises a small fraction of elements involved in *A. baumannii* pathogenesis, and thus, novel global approaches are essential to comprehensively understand the basic features of this organism in order to ultimately control the spread of *A. baumannii* infections and to develop effective countermeasures against this harmful pathogen.

In addition to its pathogenesis, the genus Acinetobacter is particularly interesting for other reasons. First, acinetobacters are capable of catabolizing a wide range of carbon sources and metabolites and as such were briefly classified as pseudomonads (Stanier et al. 1966). In fact, acinetobacters are among the most widely used microbes for petroleum remediation. Second, several strains of Acinetobacter, most notably *A. baylyi*, have an extraordinary ability to acquire foreign DNA(Young et al. 2005). It is currently unknown how pervasive natural competence is among acinetobacters. This trait is particularly important for microbial pathogens since it is one important mechanism by which they achieve genetic diversity. Pathogens which can rapidly acquire drug resistance and pathogenicity islands have a selective advantage over those with more static genomes.

Recently a new approach for high throughput DNA sequencing has been described using

pyrophosphate sequencing (Margulies et al. 2005). High-density pyrosequencing involves the clonal amplification of genomic DNA fragments followed by sequencing coupled to two enzymatic reactions (Margulies et al. 2005). The first enzyme, sulfurylase, regenerates ATP from the pyrophosphate released during base incorporation. The second enzyme, luciferase, converts the energy of the regenerated ATP into light. This procedure allows the simultaneous sequencing of hundreds of thousands of short DNA sequences (on average 100 bp; see below). However, significant challenges to using high-density pyrosequencing include the short DNA reads generated and a potential loss of accuracy due to long homopolymer stretches or low complexity DNA. High-density pyrosequencing has not yet been used effectively in resequencing efforts but has not been reported in de novo sequencing projects.

In this study we demonstrate that we can determine the DNA sequence of a microbe using pyrophosphate sequencing with reasonable accuracy. Analysis of the *A. baumannii* ATCC17978 DNA sequence revealed 28 putative alien islands predicted to be involved in virulence. Insertional mutagenesis of Acinetobacter coupled with a *C. elegans* virulence assay that we described previously (Smith et al. 2004) and a new assay using *Dictyostelium discoideum* described below, confirmed that at least 6 alien islands are involved in virulence including four that were not predicted to be involved in pathogenesis by sequence homology. While many of the genes in these alien islands were of previously unknown function, our analysis revealed potential functions for these genes and by inference, the entire operon. Thus, we were able to assign function to these previously uncharacterized operons. The combined DNA sequence and mutagenesis approach provide rapid and considerable insight into the pathogenicity of this microbe and is an approach that is generally applicable to any pathogenic microbe.

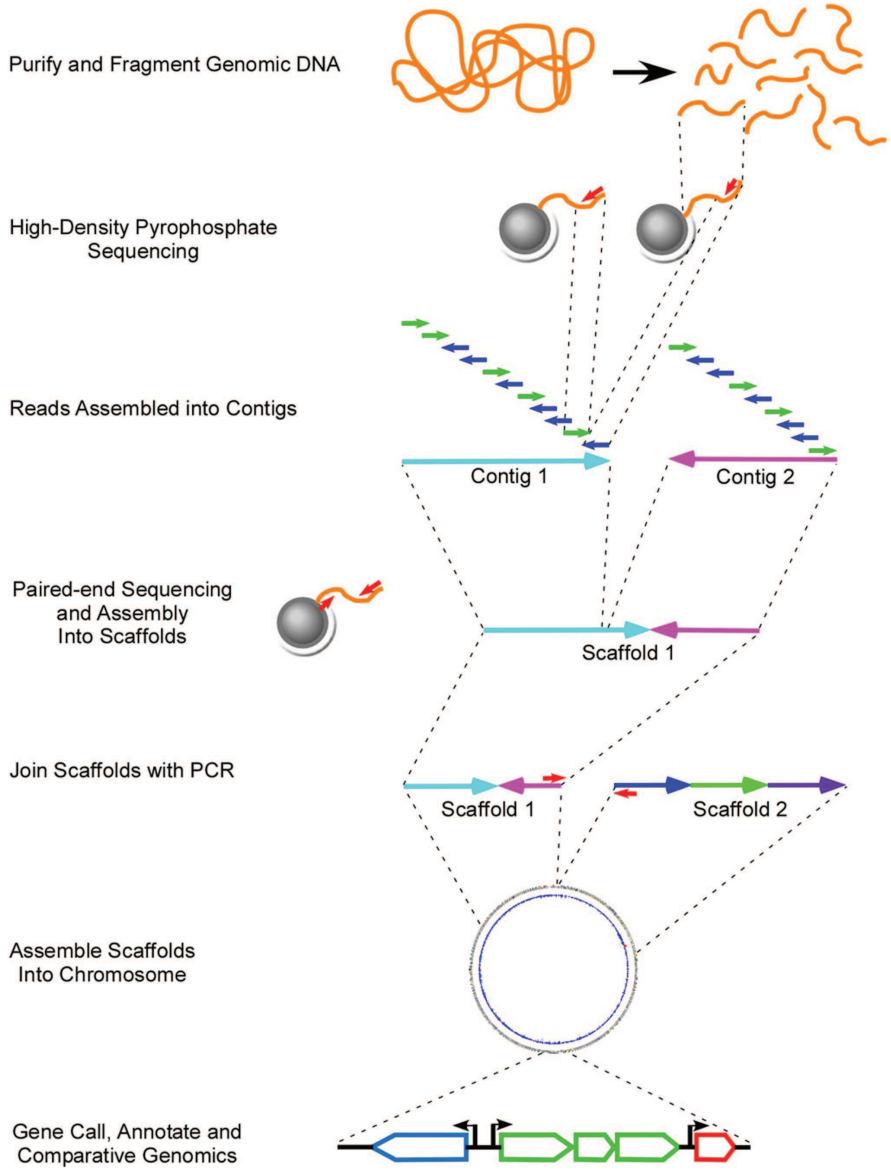


Figure 2.1: Illustrates step by step process for sequence, assembly, and annotation of the *Acinetobacter baumannii* genome.

2.2 Results

2.2.1 High-Density Pyrosequencing of the *A. baumannii* genome

We devised a strategy to determine the DNA sequence of *A. baumannii ATCC 17978* using high density pyrophosphate DNA sequencing (Figure 2.1). To overcome the limitation of short DNA reads, we sequenced a large excess of genomic DNA. From five DNA sequencing runs a total of 824,407 sequencing reads and 83,931,609 total base pairs of sequence were obtained; this depth of sequencing proved to be 21.1 fold coverage of the genome. The average read length was 101.8 base pairs and average G+C content 38.8%. The sequence was assembled into 139 contigs which ranged in length from 1,346 bp to 168,478 bp with an average length of 28,392 bp. We next performed paired end analysis in which sequence information was recovered from each end of 90,049 DNA fragments. Paired-end sequencing joined the 139 contigs into 22 supercontigs, or scaffolds, which ranged in length from 6,199 bp to 1,257,593 bp with an average of 179,384 bp. The predicted size of the *A. baumannii* chromosome based on these sequences is 3,946,442 bp. To complete the assembly, two PCR strategies were employed (Figure 2.1). First, several rounds of vectorette PCR were performed from the ends of contigs as described in the Materials and Methods (Riley et al. 1990; Kumar et al. 2002). Over 10,000 PCR and 2200 sequencing reactions were generated using conventional sequencing methods and capillary electrophoresis. This methodology was effective in filling in the gaps between contigs within scaffolds, as well as joining pairs of scaffolds. All remaining gaps were filled by direct PCR using primers that were designed to the ends of each of the remaining contigs. In this manner, the entire genome was sequenced and assembled. Gap filling added 30,304 bp to the genome assembly and indicates that high-density pyrosequencing was effective in returning 99.24% of the total chromosomal sequence. The final *A. baumannii* ATCC17978 genome sequence we determined is comprised of a single 3,976,746 base pair chromosome, not counting tandem repeats (such as might be expected for rDNA) which cannot be detected using the pyrosequencing method. In addition to the chromosomal DNA sequence, the shotgun

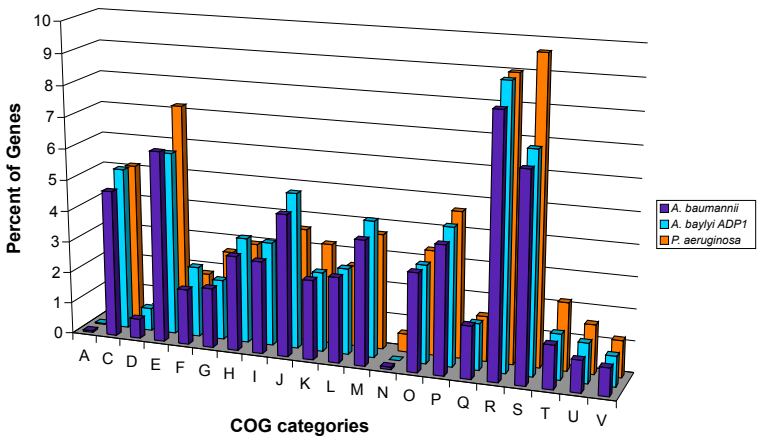


Figure 2.2: *A* Graphical depiction of the gene annotation. Genes products are represented by COG assignment (single letter code): A: RNA processing and modification; C: Energy production and conversion; D: Cell division and chromosome partitioning; E: Amino acid transport and metabolism; F: Nucleotide transport and metabolism; G: Carbohydrate transport and metabolism; H: Coenzyme metabolism; I: Lipid metabolism; J: Translation, ribosomal structure and biogenesis; K: Transcription; L: DNA replication, recombination and repair; M: Cell envelope biogenesis, outer membrane; N: Cell Motility. O: Post-translational modification, protein turnover, chaperones; P: Inorganic ion transport and metabolism; Q: Secondary metabolites biosynthesis, transport and catabolism; R: General function prediction only; S: Function unknown; T: Signal transduction mechanisms; U: Intracellular trafficking, secretion, and vesicular transport; V: Defense mechanisms.

approach also revealed the DNA sequence of two plasmids pAB1 (13,404 bp) and pAB2 (11,520 bp) from this organism. Interestingly, neither of these is identical to the previously sequenced *A. baumannii* plasmid, pMAC, although all three plasmids share the same replication protein, RepM (Dorsey et al. 2006). pAB1 shares a total of 3 ORFs with pMAC: RepM, a YPPCP.09C homologue, and a Cro-like protein however whereas pMAC contains ORFs involved in peroxide resistance, pAB1 bears genes involved in lysine metabolism. Thus, it is likely that plasmids can readily differ among *A. baumannii* species.

2.2.2 Sequence verification and annotation

Potential coding sequences were identified using two programs, GLIMMER and RBSfinder. GLIMMER identified 3830 open reading frames in the *A. baumannii* chromosome and an additional 10 and 9 ORFs on the two plasmids respectively (Delcher et al. 1999). Although GLIMMER has a high accuracy rate in calling genes, identifying the precise start site is problematic. Programs, such as, RBSfinder improve this accuracy by identifying potential ribosomal binding sites and shifting the starting coordinates of predicted open reading frames. It was shown that RBSfinder improves the accuracy of start site locations predicted by GLIMMER or Gene Mark from 67-77% to 90%, and these predictions were validated by N-terminal protein sequencing of representative E. coli proteins (Suzek et al. 2001).

RBSfinder was applied to the *A. baumannii* sequence, by first computing the consensus ribosomal binding site which was then used to identify the start codons in the 3830 ORFs. After the start sites were adjusted by RBSfinder only 76.6% of the genome encodes protein. While this number is several percent lower than that typically reported for bacteria, we believe that this is due to the use of RBSfinder which improves start site prediction accuracy. Indeed applying the GLIMMER and RBS finding method to the *A. baylyi* genome revealed a similar protein coding figure and gene homology .

subsubsectionAccuracy of the DNA Sequence One concern with using the high-density pyrosequencing method is its error rate. It was reported that high-density pyrosequencing is 99.96% accurate when compared to DNA sequenced by conventional sequencing methods and capillary electrophoresis (Margulies et al. 2005). To assess the accuracy of the *A. baumannii* sequence, two methods were used: 1) comparison of the originally assembled contigs to that of 50 PCR fragments sequenced by traditional methods and 2) determination of the frequency of split genes. To compare the sequence determined by pyroseqeuncing with that of traditional Sanger sequencing, we selected PCR products that did not reside near the ends of the assembled contigs and thus were not likely to contain low complexity DNA, transposons, rDNA or other repeated sections. These 50 PCR products

totalled 33,906 bp and had only 26 base pairs differences with the sequences assembled by pyrosequencing. Thus, we find that the genomic sequence is 99.92% accurate (Accuracy = 100-((26/33906)*100)). In all but one case, the base pair differences were in homopolymer run length, which is recognized as the most frequent type of error that high-density pyrosequencing creates (Margulies et al. 2005). In addition, in our efforts to join contigs over 750,000 bp of PCR products were generated and sequenced using traditional methods. These sequences corresponded to the lowest complexity, highest repeat-containing and most problematic sections of the genome. By incorporating the sequences generated by traditional methods into these problematic regions, the accuracy of the DNA sequence is likely to be more than 99.92% accurate.

Since the most likely sequencing error is a frameshift as a result of inappropriate base calling during homopolymer runs, a second measure of the error rate is the frequency of split genes. We used two criteria to determine the frequency of split genes in our assembly. First, these types of errors are distinguishable as a pair of tandem genes which return the same BLAST target. Second, the sum of the length of the tandem genes should closely approximate the length of the homologous gene. We found 30 instances of split genes in our assembly of the *A. baumannii* genome. An error rate calculation based on assumptions described in Materials and Methods indicates that the error rate is 0.0014%. Regardless of which accuracy measure is most precise, the sequence obtained from our approach is at least 99.92% accurate and there are relatively few (less than 1%) split genes obtained from this sequencing approach. subsubsectionAnnotation of the *A. baumannii* genome The annotated genes were assigned functions by a combination of BLAST analysis and KEGG annotation (Altschul et al. 1990; Altschul et al. 1997; Kanehisa et al. 2006) and then assigned to clusters of orthologous groups (COG) (Tatusov et al. 1997). Approximately 61% of the genes were assigned to a COG functional category (Figure2.2 and 2.3). The most represented classes of genes were those involved in translation, amino acid metabolism and energy production as would be expected from a member of the catabolically versatile pseudomonad family. We also identified 70 tRNA genes throughout the genome.

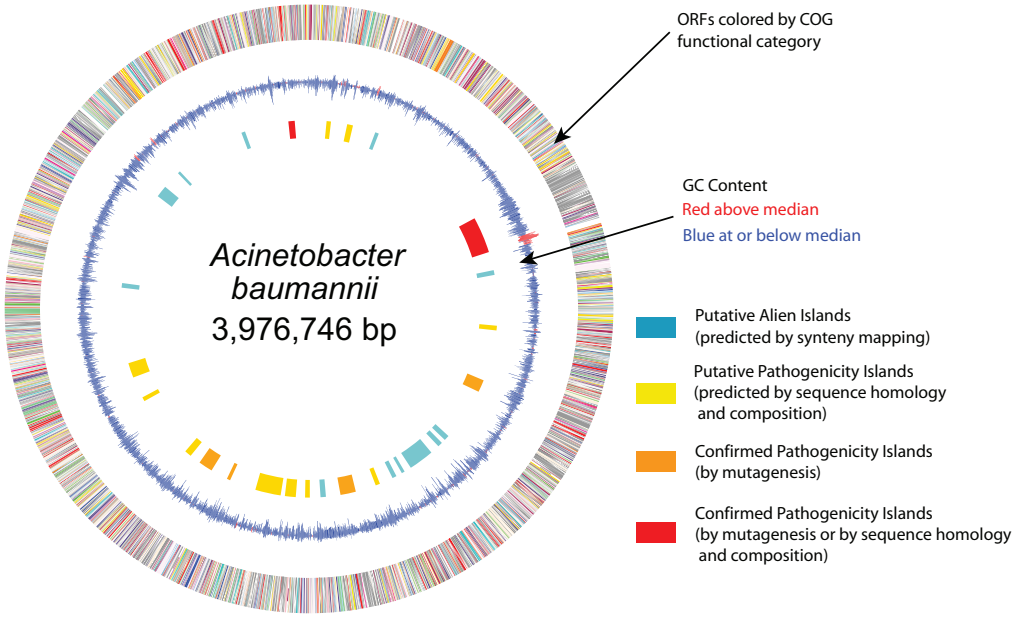


Figure 2.3: Circular Map of *A. baumannii* genome. The outermost circle shows genes color-coded by COG assignment: Translation,ribosomal structure and biogenesis (gold); RNA processing and modification (orange); Transcription (dark orange); DNA replication, recombination and repair (maroon); Cell division and chromosome partitioning (yellow); Defense mechanisms (pink); Signal transduction mechanisms (purple); Cell envelope biogenesis, outer membrane (peach); Cell motility and secretion (medium purple); Intracellular trafficking, secretion, and vesicular transport (pink); Posttranslational modification, protein turnover, chaperones (light green); Energy production and conversion (lavender); Carbohydrate transport and metabolism (blue); Amino acid transport and metabolism (red); Nucleotide transport and metabolism (green); Coenzyme metabolism (light blue); Lipid metabolism (cyan); Inorganic ion transport and metabolism (dark purple); Secondary metabolites biosynthesis, transport and catabolism (sea green); General function prediction only (light gray); Function unknown (ivory); Not in COGs (dark gray). The middle circle represents the G+C percentage, colored red for regions above median GC score (38%) and blue for regions less than or equal to the median. The circles were drawn with (cite genomeviz).

2.2.3 Acinetobacter synteny

The *A. baumannii* genome sequence was compared to that of its closest sequenced relative, *A. baylyi* using the Artemis Comparison Tool. This program uses BLAST (either blastn or tblastx) to compare two or more genomes for the arrangement of homologous genes (Carver et al. 2005). Although there are large number of local gene insertions, deletions and rearrangements, synteny mapping using the Web ACT comparison illustrates that a large amount of synteny exists between the two genomes and that large sections of the two genomes share similar orientations (Figure 2.4). Significant genomic similarities are also observed at the protein level in which 2137 of the 3830 (55.79%) predicted *A. baumannii* proteins returned their top BLAST scores as an *A. baylyi* gene product.

One of the most interesting features of the *A. baylyi* genome is the clustering of catabolic operons into an "archipelago of catabolic diversity" (Barbe et al. 2004). Of the genes found within the five islands described by Barbe et al., we found representatives of only three islands. Furthermore, these genes were not clustered into islands but were scattered throughout the chromosome. We also compared the clustering of the catabolic genes in *A. baylyi* and *A. baumannii* with the sequenced genomes of *Pseudomonas aeruginosa, Neisseria meningitidis, Bacillus subtilis* and *Escherichia coli* (Figure 2.4). The clustered organization of the catabolic archipelago is found only in *A. baylyi* and thus is likely to have occurred post-speciation. An interesting feature of the catabolic capacity of *A. baumannii* is its inability to catabolize glucose, a deficiency shared by many strains of Acinetobacter (Baumann et al. 1968b). We have identified the cause of this deficiency in *A. baumannii* as the absence of hexokinase, glucokinase or any other comparable enzyme that can transfer phosphate onto glucose. Thus, the first step of glycolysis cannot be completed.

2.2.4 Nucleic Acid Translocation

Examination of the *A. baumannii* genome revealed that it lacks two important genes involved in DNA uptake, comP and comA. However, it does have most others, such as

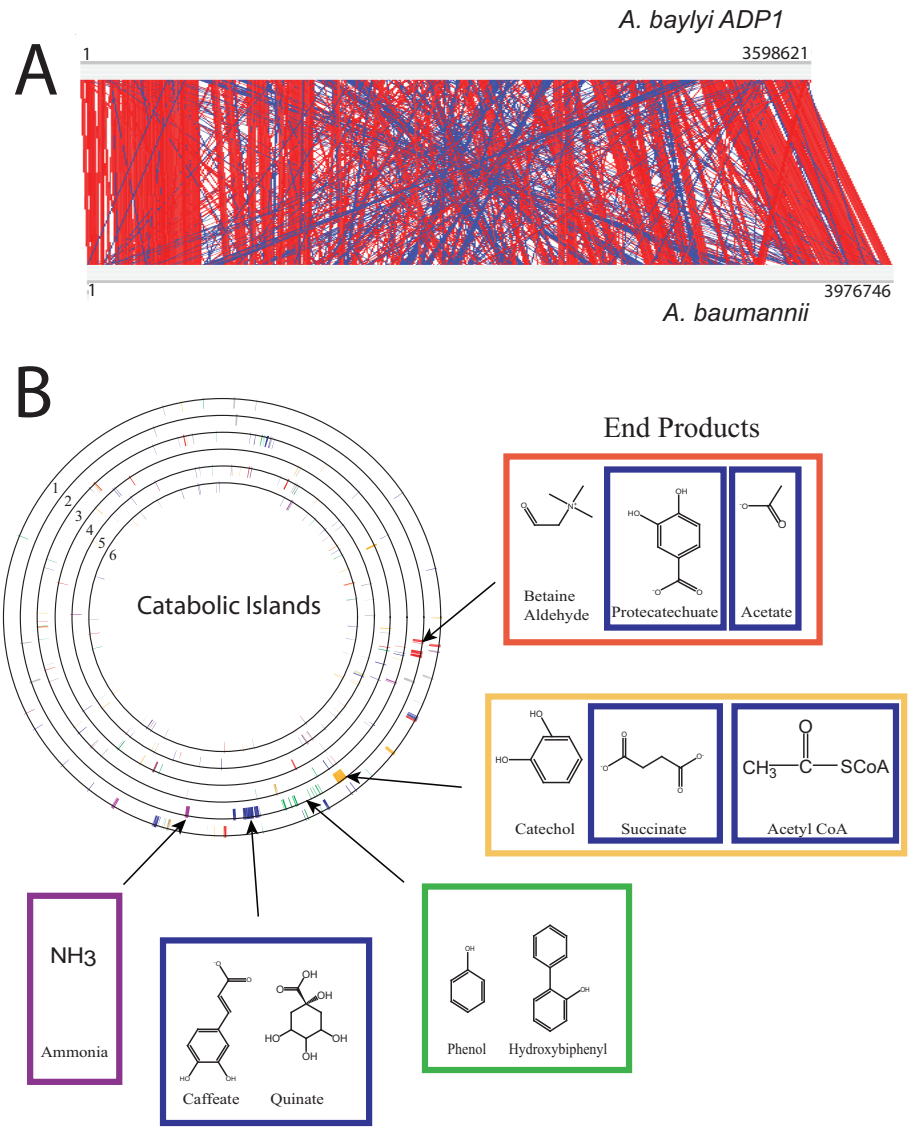


Figure 2.4: A) The genomes of *A. baylyi* (upper) and *A. baumannii* (lower) were compared by WebACT and visualized by the Artemis Comparison Tool (ACT) (Carver et al. 2005). Red indicates similar genomic organization, whereas blue indicates inversions. B) The genomes of six bacteria were compared for the distribution of key catabolic enzymes (from the outermost to the innermost ring): *A. baumannii* , *A. baylyi, Pseudomonas aeruginosa, Neisseria meningitidis, Escherichia coli K12*, and *Bacillus subtilis*. The Island clusters were defined by Barbe et al (Barbe et al. 2004) based on their location in the *A. baylyi* genome and are colored Grey (-); Red (I); Orange (II); Green (III); Blue (IV); Purple (V). The endproducts of the pathways encoded within each of the catabolic islands are depicted. The boxes surrounding the endproducts are colored to match the islands from which they were derived.

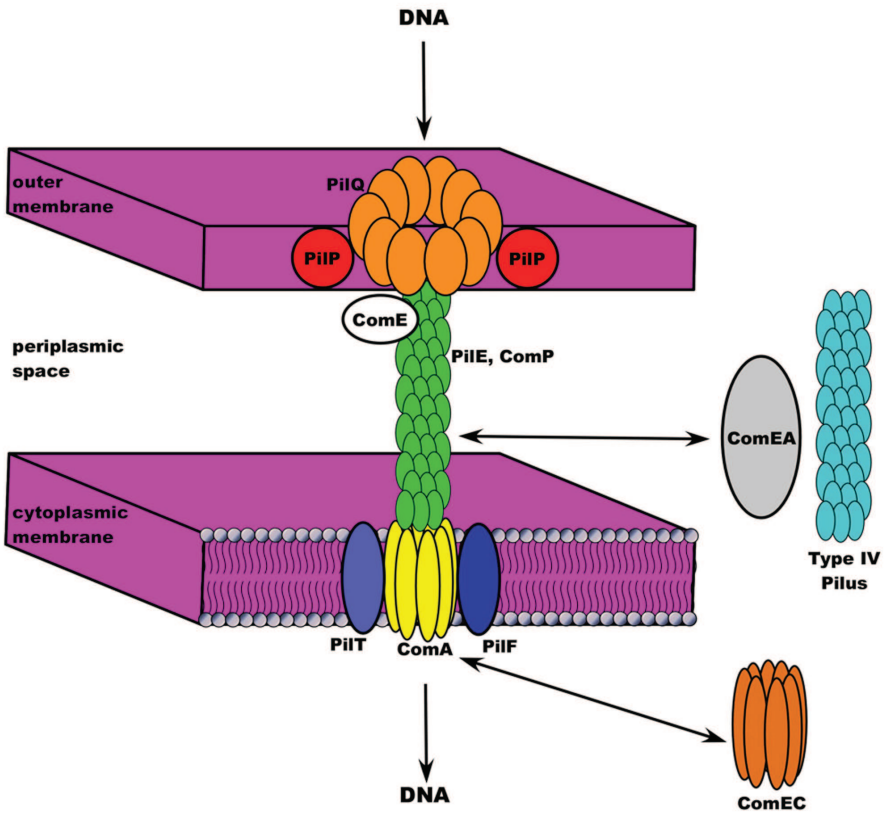


Figure 2.5: DNA Transport machinery. Adapted from Averhoff and Friedrich (Averhoff and Friedrich 2003). In *A. baylyi* and other gram-negative bacteria, foreign DNA is delivered to and through the outer membrane transporter PilQ. ComE-bound DNA is transported to the inner membrane transporter ComA (or ComEC) via ComP, PilE and/or the type IV pilus. ComEA may assist in this delivery.

pilQ, comE, and pilF among a total of over 20 pilus, fimbrial and competence genes. In addition, homologues of two genes required for DNA uptake, comEA and comEC, are found in *A. baumannii* genome but not in that of *A. baylyi*. Since comEA is a transmembrane protein whose role is to bind external DNA and deliver it to the comEC transporter, this suggests that *A. baumannii* can potentially compensate for the loss of comP with either comEA and/or its typeIV pilus. Secondly, *A. baumannii* may be able to compensate for the loss of comA with its homologue, comEC. In doing so it would build a novel nucleic acid transport machine. The presence of these genes may explain, in part, the large amount of foreign DNA found within the *A. baumannii* chromosome (Figure 2.5).

2.2.5 Putative Alien Islands (pAs)

Of particular interest are the potential virulence genes in *A. baumannii*. Sequence similarities with known virulence genes can reveal the identity of homologs, but contribute little new information. We used two approaches to identify *A. baumannii* virulence genes. Using comparative genomics, our first step was to directly compare the *A. baumannii* sequence with the non-pathogenic *A. baylyi* in order to identify *A. baumannii* specific genes. One of the benefits of the high degree of similarity between these two organisms is that their genomic differences are likely to contribute to their phenotypic differences. We defined our regions of interest as those regions greater than 10kB that had little homology with *A. baylyi*. The second approach we utilized was to examine the genome for sequence composition anomalies indicative of putative alien islands (pAs, (Karlin 2001)). Differences in G+C content, amino acid usage, dinucleotide frequency, and codon usage have successfully been used to identify microbial alien islands (Karlin 2001). While pAs may possess genes encoding any number of functions, those determined to be involved in virulence are termed pathogenicity islands (PAIs). There were twenty-eight clusters of genes that fit the criteria as alien islands. Twelve of these possess genes with sequence homology to genes with roles in pathogenesis (Table 2.1). The largest pA is 133,740 bp, contains several transposons and integrases and strikingly also contains eight genes homologous to the Legionella/Coxiella Type IV virulence/secretion apparatus and therefore may be a bonafide PAI. Since a separate full complement of genes involved in Type IV mediated conjugation are located elsewhere on the chromosome we speculate that if the eight Type IV secretion (virulence) genes found in the pA are indeed virulence factors, then in order to build a functional Type IV secretion apparatus for virulence, the structural elements of the conjugation pilus may be utilized during pathogenesis. In addition to the large island described above, 7 different pAs contain genes encoding drug resistance proteins. We also found pAs containing genes encoding heavy metal resistance, iron uptake and metabolism, fimbrial genes, autoinducer processing and cell envelope biogenesis

Putative Alien Islands				
pA #	Gene # start	Gene # end	General Function	Potential Role in Virulence? (evidence)
1	54	70	Cell Envelope Biogenesis	Yes (Sequence Homology)
2	119	130	Autoinducer Production	Yes (Sequence Homology)
3	213	226	No homology/Hypothetical Proteins	No
4	642	748	Type IV secretion	Yes (Sequence Homology and Genetic Screen)
5	796	809	Amino Acid Metabolism	No
6	981	995	Drug Resistance	Yes (Sequence Homology)
7	1164	1192	Amino Acid Metabolism	Yes (Genetic Screen)
8	1382	1399	Xenobiotic Degradation	No
9	1409	1426	Metabolism	No
10	1455	1542	Phenyl Acetic Acid Degradation	No
11	1566	1580	Amino Acid Metabolism	No
12	1602	1617	Arsenic resistance/Taurine metabolism	No
13	1665	1681	Pilus Biogenesis	Yes (Sequence Homology)
14	1755	1814	No homology/Hypothetical Proteins	Yes (Genetic Screen)
15	1863	1878	Vitamin B12 Metabolism	No
16	1919	1934	Drug/metabolite resistance	Yes (Sequence Homology)
17	1962	1999	Drug resistance	Yes (Sequence Homology)
18	2012	2103	Drug resistance/Metabolism	Yes (Sequence Homology)
19	2190	2201	Amino Acid Metabolism	Yes (Genetic Screen)
20	2261	2313	No homology/Hypothetical Proteins	Yes (Genetic Screen)
21	2349	2368	Iron transporters/metabolism	Yes (Sequence Homology)
22	2581	2594	Drug resistance	Yes (Sequence Homology)
23	2659	2704	Drug resistance	Yes (Sequence Homology)
24	2942	2960	Metabolism	No
25	3246	3276	Heavy metal resistance	No
26	3346	3355	Lipid Metabolism	No
27	3600	3616	No homology/Hypothetical Proteins	No
28	3760	3780	Drug resistance/Metabolism	Yes (Sequence Homology and Genetic Screen)

Table 2.1: The putative alien islands were numbered in order based on their chromosomal start site. The general function was based on an assessment of the ORFs with the pA. Genes unrelated to this general function may be found in any given pA. The potential role in virulence was determined by sequence homology to known virulence genes or by virtue of being recovered in the mutagenesis screen described in this manuscript.

(Table 2.1). Of the remaining 16 pAs, 12 islands contained genes involved in various aspects of metabolism, lipid metabolism, amino acid uptake and processing and xenobiotic degradation. Four of the pAs were difficult to characterize since they contained largely hypothetical genes and mobile elements.

Fournier et al. recently reported the identification of a hotspot in *A. baumannii* into which two different genomic islands inserted (Fournier et al. 2006). The larger of the two islands is an 86,190 bp drug resistance island containing 45 of the 52 drug resistance genes in *A. baumannii* strain AYE (Fournier et al. 2006). The smaller island, found

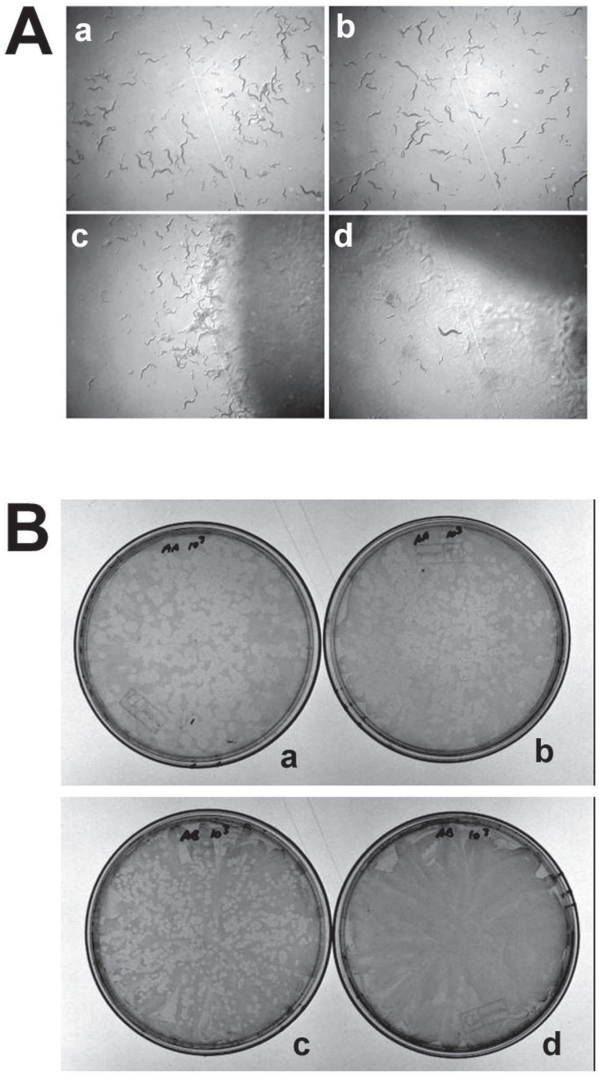


Figure 2.6: Ethanol-Stimulated Virulence of *A. baumannii.* *A.* Bacteria were incubated on NGM plates without (a and c) or with 1% ethanol (b and d). A single L4 stage worm was inoculated onto lawns of *E. coli OP50* (a-b) or *A. baumannii* (c-d) and allowed to proliferate for 4 days. The *E. coli OP50* lawns are completely consumed by this time (a-b) regardless of the presence of ethanol. The worm brood is considerably smaller on plates containing NGM + 1% ethanol and *A. baumannii* (d). *B.* Bacteria and *D. discoideum* amoebae were incubated on SM/5 plates without (a and c) or with 1% ethanol (b and d). In lawns containing *K. aerogenes* (a-b) amoebae plaques form within 4 days regardless of the presence of ethanol. Amoebae are able to form plaques in *A. baumannii* (c) but plaque formation is completely inhibited when 1% ethanol is added to the media (d).

in the drug sensitive strain SDF, is 19,362 bp and does not contain any drug resistance genes. We found an insertion similar to the one found in the drug-sensitive SDF strain. It is 13,277 bp, comprising 9 genes and is found between 5and 3 ends of a putative ATPase. The 9 genes include those encoding 4 hypothetical or unknown proteins, two transposases, one transposition helper, a universal stress protein and the sulphonamide resistance protein, sulI. Although only one drug resistance gene was found in this insertional hotspot, *A. baumannii* ATCC17978 possesses an additional 74 potential drug resistance genes, including 32 efflux pumps (19 RND type, 3 MFS and 9 others) and 11 permeases of the drug/metabolite transporter (DMT) superfamily. We also identified 26 genes encoding resistance to heavy metals including copper, cadmium, zinc, cobalt, tellurite and arsenic. Since these genes likely derive from mobile DNA and are at many locations in the genome, clearly the previously identified hotspot is not the only location into which genes have inserted in this organism.

2.2.6 Insertional Mutagenesis Reveals that Many Alien Islands are Important for *A. baumannii* Virulence

Although the genomic sequence analysis revealed many PAIs which might be involved in virulence, direct evidence is lacking. We therefore sought to determine whether any of these islands had a role in pathogenesis using a limited insertional mutagenesis approach. A library of 1324 *A. baumannii* mutants was generated using the EZ::TN $\langle$R6Kγori/KAN–2$\rangle$ Tnp Transposome system from Epicentre (Dorsey et al. 2002). These mutants were then tested for their ability to affect pathogenesis in two assays: reduced brood size of *C. elegans* and inhibition of *D. discoideum*. By using two different assays to study the pathogenesis of acinetobacters we expected to uncover general mediators of virulence, as well as factors that specifically affect a bacterium's ability to survive in either host.

The worm assays were performed by inoculating a lawn of bacteria grown on nematode growth medium (NGM) with a single L4 stage worm. The worm undergoes its final molt to the adult stage and begins laying as many as 300 eggs during the subsequent 24 hours.

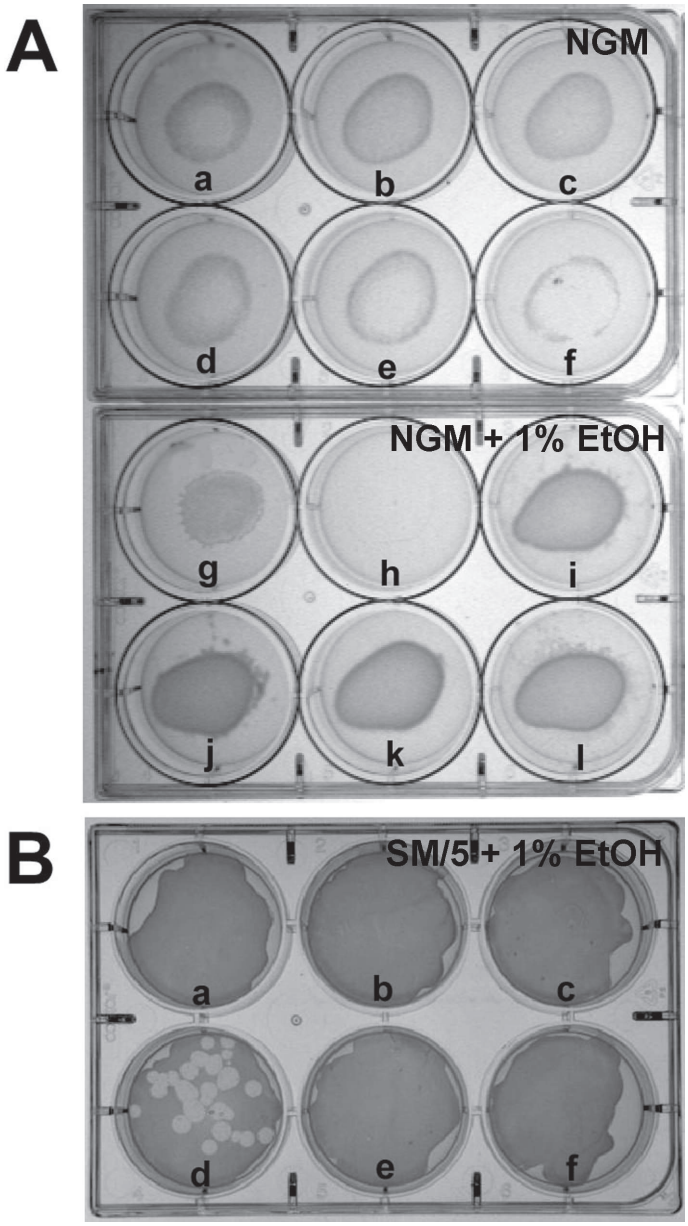


Figure 2.7: Ethanol-Stimulated Virulence Mutants of *A. baumannii* Mutants were generated according as described in Materials and Methods. A: Six individual *A. baumannii* mutants were incubated on NGM (wells a-f) or NGM + 1% ethanol (wells a-f). A single L4 stage worm was inoculated onto lawns of each mutant and allowed to proliferate for 4 days. Avirulent bacterial mutants were recovered as those which allowed the worms to consume the bacterial lawns as fast or faster on NGM + 1% ethanol as on NGM alone, e.g. mutant b (b and b) . B: Six individual *A. baumannii* mutants were mixed with . *discoideum* amoebae and incubated on SM/5 + 1% ethanol (wells a-f). These are not the same mutants as shown in A. Avirulent mutants were defined as those which allowed amoebae plaque formation, e.g. mutant d.

This rapid proliferation of worms results in a brood capable of exhausting the bacterial lawn within 4-5 days. However, when the worms are inoculated onto *A. baumannii* lawns that have grown on 1% ethanol (which induces virulence (Smith et al. 2004)), worm proliferation is slowed and brood sizes are dramatically diminished (Figure 2.6A). Ethanol does not affect the worms directly since worms feeding on *E. coli OP50* proliferate equally well on ethanol-containing and ethanol-free media. We screened our *A. baumannii* library for mutants that would allow worm proliferation at comparable rates on ethanol-containing and ethanol-free media. A total of 114 mutants were able to support the growth of worms on ethanol-containing media and were deemed avirulent to worms (Figure 2.7A).

We also tested for hostile interactions between *A. baumannii* and D. discoideum using a simple plating assay that was developed for *P. aeruginosa* (Pukatzki et al. 2002). Amoebae and bacteria were cultured on media with or without ethanol. Bacteria, due to their greater growth rate, form lawns with the slower-growing amoebae embedded within them. On media without ethanol, the amoebae were able to consume *A. baumannii* and form plaques in the bacterial lawn. On media supplemented with limited ethanol, no plaques were observed in *A. baumannii* lawns. As many as 1x105 amoebae were plated on a single plate and not a single plaque was formed on lawns of *A. baumannii* when ethanol was added to the media (Figure 2.6B). Ethanol does not appear to affect the amoebae directly since amoebae feeding on *K .aerogenes* form plaques equally well on ethanol-containing and ethanol-free media. Screening the mutant library for bacteria that allow plaque formation by the amoebae in the presence of ethanol, revealed 229 avirulent mutants out of the 1324 mutants tested (Figure 2.7B). Comparison of the results of the worm and amoebae screens revealed thirty mutants that tested positive in both assays. The identity of the gene disrupted by the insertion mutation in these thirty mutants was determined by rescuing the DNA adjacent to the transposon and DNA sequencing. In addition, we also rescued and sequenced 2 mutants which were robust in the worm assay, but failed in the amoebae assay and 3 mutants which did not affect amoebae growth, but did inhibit worm growth. In all, 35 mutants were sequenced and their genomic locations determined. 14 of the 35

mutants had no homology to *A. baylyi* seven of which localized to the pAs described above. Two of these seven pAs were predicted by sequence homology to encode genes involved in virulence. Thus, our functional assay has contributed 5 pAs as likely to be involved in the pathogenesis of *A. baumannii*. 3 of these 5 pAs are predicted to be involved in various aspects of metabolism, so their relation to virulence mechanisms is unclear. The remaining 2 pAs are particularly interesting as they may represent novel virulence effectors since BLAST analysis revealed little sequence similarity to any known genes and they may represent novel virulence factors.

2.2.7 Further Characterizatio of Ethanol-Stimulated Virulence Mutants

There are two likely mechanisms that would result in the reduction in the worm brood sizes observed: The worms are dying and/or they are not proliferating. To separate these two possibilities, we determined worm lifespans and also counted the number of eggs that worms laid when feeding on E. coli OP50, *A. baumannii* or 14 of the avirulent, ethanol-stimulated virulence (Esv) mutants. The results are summarized in Table 2.2. The lifespans of worms feeding on *A. baumannii* were 20% shorter when ethanol was added to the media. A more dramatic effect is observed when eggs are counted. Worms feeding on ethanol-stimulated *A. baumannii* lay 42.9% fewer eggs than those feeding on unstimulated bacteria. Thus, it appears that the combined effect of the shortened lifespan and the reduction in the reproductive capacity of the worms when feeding on ethanol-stimulated *A. baumannii* results in the diminished brood of worms illustrated in (Figure 2.6A). Several of the bacterial mutants are able to reverse the deleterious effects of ethanol-stimulated *A. baumannii* on worm proliferation by both increasing the worm lifespans and egg-laying capacity. EsvA, encodes a transcription factor in the AraC/XylS family. EsvC encodes the gamma subunit of urease. EsvI is a likely multidrug efflux transport protein. Some of the mutants were able to sustain prolonged worm lifespans but had lesser effects on egg production. Two such examples are EsvB, which encodes a protein in the MarR class of transcription factors and EsvD, which encodes an ABC-type membrane transporter. Two

of the mutants, EsvF and EsvM appeared to remain virulent and will not be studied further. To investigate whether the mutations caused a slow growth phenotype that might indirectly lead to avirulence, we measured the generation times of each mutant. Each mutant strain doubled at a rate equal to or faster than wild-type *A. baumannii* except for EsvF andEsvM (Table 2.2). Thus, slow growth is not responsible for the avirulent phenotypes observed in most of these mutants. Biofilm formation is a well characterized developmental pathway utilized by many bacteria in their pathogenesis. Wild-type *A. baumannii* form biofilms on abiotic surfaces (Tomaras et al. 2003). We examined the effect of ethanol on biofilm formation and the ability of our mutants to form biofilms on plastic surfaces. Ethanol has no apparent effect on biofilm formation; the cells adheres to plastic surfaces equally well in the presence or absence of ethanol (data not shown). Furthermore, all of the mutants tested retained the ability to form biofilms suggesting that, in this strain, biofilms are not essential for virulence (data not shown). Thus, these genes are likely involved in other virulence processes.

2.3 Discussion

This manuscript reports the first de novo sequencing effort using a novel high throughput method for genomic DNA sequence. Examination of this microbial genome revealed little nucleotide sequence identity between *A. baumannii* and *A. baylyi*, the only previously sequenced member of the Acinetobacter genus. Genomic similarities are seen at the protein level where in which 2137 of the 3830 (55.79%) predicted *A. baumannii* share homology with *A. baylyi* gene products. The most interesting differences between these two organisms lies in the 28 putative alien islands identified in *A. baumannii.* Many of the drug resistance and potential virulence factors found in the *A. baumannii* genome reside on these islands indicating that a large number of them are important for the pathogenesis of this organism.

High-density pyrophosphate sequencing is a rapid, cost-efficient method for large sequencing projects without the labor or potential bias of cloning steps. By virtue of

	Worm Lifespans			Egg Counts			Generation Time
	LT_{50} (NGM)	LT_{50} (NGM + EtOH)	% change	NGM	NGM + EtOH	% change	% change
E. coli OP50	232	224	-3.45	229.33	211.67	-7.70	
A. baumannii (wild-type)	210	168	-20.00	199.67	114.00	-42.90	
EsvA	228	276	21.05	186.67	209.67	12.32	13.4
EsvB	225	273	21.33	204.33	190.33	-6.85	16.9
EsvC	249	252	1.20	120.67	136.83	13.40	5.2
EsvD	220	280	27.27	163.00	161.00	-1.23	9.5
EsvE	245	236	-3.67	190.33	169.00	-11.21	7.9
EsvF	279	252	-9.68	137.33	110.33	-19.66	10.5
EsvF (isolate 2)	296	237	-19.93	250.33	63.67	-74.57	-3.6
EsvG	252	283	12.30	156.33	131.50	-15.88	1.6
EsvH	256	273	6.64	153.67	56.00	-63.56	4.2
EsvI	218	246	12.84	188.50	209.67	11.23	11.4
EsvJ	214	278	29.91	166.00	149.33	-10.04	7.5
EsvK	253	256	1.19	212.33	179.50	-15.46	3.4
EsvL	256	274	7.03	163.00	127.00	-22.09	1.9
EsvM	297	282	-5.05	206.00	55.33	-73.14	-38.6

Table 2.2: Analysis was performed as described in the Materials and Methods section.

large numbers of parallel sequencing runs, genomic sequence coverage is high, in our case 21.1x. This powerful method has been used effectively in genome resequencing efforts and for sequencing microbial strain variants (Margulies et al. 2005; Hofreuter et al. 2006). The primary drawbacks of this methodology are the short sequencing runs, which result in difficulties in assembly of sequences in low complexity or repeated regions, and errors in base calling in stretches of homopolymers, which can result in frameshifts. These limitations were overcome by not only the high sequence coverage, but also the use of paired end analysis and gap closures by traditional and random ended PCR. When combined with conventional gap filling, we find that the accuracy of the sequence and assembly are comparable to the whole-genome shotgun sequencing methods which have become the gold standard of genomic sequencing. In contrast to a recent report, which suggests that high-density pyrosequencing is unable to replace Sanger sequencing for de novo microbial genome projects(Goldberg et al. 2006), we find that high-density pyrosequencing can be a suitable replacement for the sequencing of microbial genomes. We found that an initial assembly of 21.1x high-density pyrosequencing data is sufficient to determine the genome size and build a working draft that could be used for almost any genomic analysis. Furthermore, our initial draft assembly predicted a genome of 3,946,442 base pairs, whereas the final assembly predicts a genome of 3,976,746 base pairs. Therefore the initial draft provided coverage for over 99.24% of the completed genome, and only 30,302 base pairs were missing. One factor that may have negatively impacted Goldman et als evaluation was the omission of paired end analysis (Goldberg et al. 2006). Our assembly was greatly facilitated by paired end sequencing and resulted in the joining of 139 contigs into a considerably more manageable 22 scaffolds. The ability to obtain longer reads will likely facilitate contig assembly; however paired end reads will still be useful for identification of large tandem repeats.

The sequence of *A. baumannii* reveals that this organism has acquired a number of genes from its environment and that these genes likely play a direct role in its virulence. We identified 28 putative alien islands (pAs) based on sequence characteristics and also

sequence comparisons with a non-pathogenic relative. One-fourth of these islands appear to be involved in drug resistance. The strain used for these studies is resistant to β-lactams, but shows only weak resistance to tetracycline and is aminoglycoside (kanamycin) sensitive (unpublished observations). Thus, the presence of so many drug resistance islands is surprising. This strain was isolated in or around 1951 (Piechaud and Second 1951; Baumann et al. 1968b), prior to the development of the macrolides (erythromycin), glycopeptides (vancomycin), and cephalosporins and other latter generations of β-lactams. It is possible that this strain was never exposed to chloramphenicol, which was first used as a therapeutic in 1949. For these reasons, it would be interesting to subject ATCC17978 to a full panel of antibiotics and assess its resistance capacity. Potentially more interesting will be comparative analysis with more recent isolates which can assess the evolution of antibiotic resistance over the last 50-60 years. In this respect a recent clinical isolate, resistant to several β-lactams, aminoglycosides, fluoroquinolones, chloramphenicol, tetracycline, and rifampin was found to contain a single large drug resistance island containing 45 of the 52 drug resistance genes in its genome (Fournier et al. 2006). This impressive clustering of drug resistance genes was not observed in the ATCC17978 strain that we sequenced. Given the differences in the resistance capacities of these two strains, it appears that the more recent isolate has dispensed with the many less effective resistance genes in favor of a single, highly potent cassette of drug resistance genes. Other pAs that appear to play a role in virulence were identified using a random insertion mutagenesis protocol. Seven different islands were identified as virulent by two different pathogenesis screens. Of these seven, two have genes thought to be involved in virulence, namely the type IV secretion apparatus and drug resistance genes. Three are predicted to be involved in metabolism. The fact that they were identified in this screen may be related to the use of ethanol as a virulence stimulus and will be investigated further. The final pair of pAs identified by the mutagenesis contained many hypothetical genes and nucleic acid mobility related genes such as transposons, integrases and phage proteins. Thus their potential function assigned by sequence homology is speculative. However, since insertion

in these sites results in avirulent mutants, the function of these islands is important for pathogenesis. Thus, the combination of genome sequencing and mutagenesis is a powerful approach for identifying and validating pathogenic genes in a. baumannii and will be a useful general method for finding such genes.

Many host-pathogen interactions have evolved in the environment where bacteria interact with predators and competitors. *C. elegans* and *D. discoideum* have been successfully used as host models for bacterial infection (Pukatzki et al. 2002; Alegado et al. 2003). These two organisms consume their bacterial prey in different ways. *C. elegans D. discoideum* has a digestive tract in which the bacteria are crushed, lysed, enzymatically digested and the subsequent nutrients are absorbed by the cells comprising the intestine (Avery and Thomas 1997). This process is almost entirely extracellular. Conversely, the unicellular D. discoideum uses phagocytosis, followed by vesicle fission and fusion resulting in a phagolysosome to digest bacteria; this process is largely intracellular (Cardelli 2001). Escape from either the worm gut or amoebae vesicles presents two distinct challenges to the bacteria, and while some of the tools required may overlap, many must be different. This is demonstrated by the fact that only 30 of the mutants overlapped between the two screens. Therefore, using both assays to study pathogenesis will uncover global regulators of virulence required in both instances, as well as factors specific to either host. This approach was validated by the recovery of two transcription factors that are potentially global regulators of virulence. Future studies will identify their downstream effectors. One interesting gene that was uncovered by our screen was urease. The enzymatic function of urease is to produce ammonia and carbon dioxide from urea. However, recent studies on plant ureases suggest that they play a role in defense mechanisms and have biological function independent of their enzymatic activity (Olivera-Severo et al. 2006). The *A. baumannii* urease gene may play a similar role.

Seven different alien islands were identified as virulent by two different pathogenesis screens. Of these seven, two have genes thought to be involved in virulence, namely the type IV secretion apparatus and drug resistance genes. Three are predicted to be involved

in metabolism. The fact that they were identified in this screen may be related to the use of ethanol as a virulence stimulus and will be investigated further. The final pair of pAs identified by the mutagenesis contained many hypothetical genes and nucleic acid mobility related genes such as transposons, integrases and phage proteins. Thus their function, by sequence homology is speculative. However, since insertion in these sites results in avirulent mutants, the function of these islands is apparently pathogenesis related. Future studies will be aimed at understanding how these genes play a role in *A. baumannii* virulence and what role ethanol contributes to their function.

2.4 Materials and Methods

2.4.1 Whole-genome sequencing

Genomic DNA from *A. baumannii* (ATCC 17978) was extracted from cells grown overnight at 30C in LB liquid cultures using the MasterPure DNA kit (Epicentre, Madison, WI). 50 μg DNA was then sequenced by 454 Life Sciences (Branford, CT) using high-density pyrosequencing methodology (Margulies et al. 2005).

2.4.2 Sequence assembly and validation

Initial assembly efforts were performed by 454 Life Sciences and resulted in the formation of 139 contigs assembled into 22 scaffolds. Contigs were joined using Vectorette PCR (Riley et al. 1990; Kumar et al. 2002). Briefly, the primers ABP1 and ABP2 (ABP1 : 5-GAAGG AGAGG ACGCT GTCTG TCGAA GGTAA GGAAC GGACG AGAGA AGGGA GAG-3; ABP2 : 5-GACTC TCCCT TCTCG AATCG TAACC GTTCG TACGA GAATC GCTGT CCTCT CCTTC-3) were denatured for 5minutes at 95oC and annealed together at 5nM each in annealing buffer (10 mM Tris pH 8.0, 10 mM MgCl2, 50 mM NaCl). The primers were then ligated to genomic DNA previously digested with a blunt-ended restriction endonuclease. The ligation mixture is then used as a template for HotStart PCR (Qiagen, Hilden, Germany). The PCR primer UV was used in all reactions (UV: 5 CGAAT

CGTAA CCGTT CGTAC GAGAA TCGCT 3) while the reverse primer was designed to each of the two ends of the 139 contigs. In all, 9 different restriction enzymes/enzyme combinations were used (DraI, HaeIII, HincII, PsiI, ScaI, SspI, XmnI, HincII/EcoRV and MscI/HpaI/PvuII). PCR products were gel purified and sequenced using the same primers from which they were amplified and standard methods and capilary electrophoresis using an ABI 3730 DNA sequencing instrument. Contigs and PCR products were assembled using the Sequencher assembly program (Gene Codes, Ann Arbor, MI).

2.4.3 Genome analysis and annotation

A combined gene prediction strategy was applied on the assembled sequences using GLIMMER and GeneMark. Putative ribosomal binding sites and tRNA genes were identified with RBSFinder (Suzek et al. 2001) and tRNAscan-SE (Schattner et al. 2005), respectively. Prior to the manual annotation of each predicted gene, an automatic functional annotation was computed based on different analyses. Similarity searches were performed against different databases, including BLASTnr and KEGG. Finally, each gene was functionally classified by assigning a cluster of orthologous groups (COG) number and corresponding COG category.

2.4.4 Calculating error frequency

To calculate an error frequency, we divided the number of errors resulting in a split gene by the fraction of protein coding genes regions whose frameshift is likely to produce two ORFs (greater than or equal to 100 bp) that can be detected by our BLAST searches. This amounts to 2,853,379 bp of total protein coding sequence minus 200 bp for each gene (723,181 bp), since insertions in this regions will not generated new ORFs. This calculation assumes that an error will result in a nearby translation termination. Thus, the final error rate from this method is 30/2,130,198 bp or 0.0014%.

2.4.5 Genomic comparison

For comparative analyses, the annotated genome sequence of *A. baylyi* (GenBank accession no. CR543861) was accessed. Homology searches were conducted on the nucleotide and amino acid sequence level using BLAST [cite BLAST]. Comparisons of chromosomal sequences were carried out with the Artemis Comparison Tool (ACT) [cite ACT].

2.4.6 Detection of regions with atypical G+C content

Genomic regions with atypical G+C content were identified using a sliding window technique with a window size of 1,000 bp. For this, the G+C content was treated as a Gaussian distribution, and regions with differences of at least 1.5 standard deviations from the mean were calculated.

2.4.7 Detection of putative pathogenicity islands

The four methods outlined by S. Karlin (Karlin 2001) were implemented for detection of anomalous regions. These included atypical G+C content, dinucleotide signature, codon bias, and amino acid bias. In each case, a sliding window technique with a window size of 1000 bp was used. Assuming a Gaussian distribution, those regions with a difference of at least 1.5 standard deviations from the mean were identified as anomalous regions. Homology searches on these regions were carried out to identify putative pathogenicity islands.

2.4.8 *A. baumannii* mutagenesis

Mutagenesis of *A. baumannii* was performed as described previously (Dorsey et al. 2002). Briefly, electrocompetent cells were generated as follows: Overnight cultures of *A. baumannii* were diluted 1:100 and grown to a cell density of 0.5-0.8 x 108 cells/ml. The cells were collected by centrifugation and washed with ice cold ddH2O three times. The cells were washed once in ice cold 10% glycerol and resuspended in 10% glycerol to a final

concentration of 2.0-2.5 x 1010 cells/ml. These electrocompetent cells were electroporated with the EZ::TN ⟨R6Kγori/KAN–2⟩ Tnp Transposome (Epicentre, Madison, WI) and kanamycin-resistant transformants were selected. The electroporation was performed using a Gene Pulser II (Bio-Rad Laboratories, Hercules, CA) at settings of 25 FD, 200 and 1.8kV. Mutants were rescue cloned according the Epicentres protocol. Briefly, mutant bacteria were grown, and their genomic DNA purified. The DNA was cut using EcoRI, ligated to itself and transformed into pir-116 electrocompetent *E. coli*. Kanamycin-resistant colonies were harvested, and plasmids containing the insertion cassette and flanking *A. baumannii* sequence were purified and sequenced using primers complementary to the insertion cassette.

2.4.9 *A. baumannii* generation times

Three colonies of each mutant as well as the wild-type *A. baumannii* were inoculated into LB an incubated overnight at 37 degrees with shaking. Cultures were diluted to a cell density of 1.0 x 107 cells/ml in fresh LB and incubated at 37 degrees with shaking. Cell densities were measured by optical density at 600nm at 1-2 hour intervals for 36 hours. Generation times were determined during the exponential phase of growth for each isolate and the average for each strain determined. Generation times of the mutants were then compared to the wild-type strain.

2.4.10 *C. elegans* killing assay

Escherichia coli (strain OP50) grown on NGM media(Sulston and Hodgkin 1988) were fed to *Caenorhabditis elegans* (strain N2). L3/L4 stage worms were placed onto lawns of *A. baumannii* . grown on NGM or NGM + 1% ethanol. Plates were incubated at 25oC. Viability was tested every 24 hours by visual examination. Worms were considered dead if they no longer moved nor responded to touch. For each strain of bacteria tested, 60 worms were assayed. The LT50 is defined as the time it takes for half of the worms to die.

2.4.11 *C. elegans* egg count assay

1 L4 stage worm was placed onto lawns of *E. coli OP50* or *A. baumannii.* grown on NGM or NGM + 1% ethanol. Each day for 5 consecutive days, the worm was moved to a fresh bacterial lawn. The plates were examined for the presence of eggs/L1 stage worms for each of the 5 plates. Egg production typically peaked at day 2 or 3 and was exhausted by day 5. Plates were incubated at 25∘. For all strains of bacteria tested, the egg production of 9 worms were assayed.

2.4.12 *D. discoideum* plaque assay

Performed essentially as described by Pukatzki et al.(Pukatzki et al. 2002). Briefly, D. discoideum AX3 were grown axenically in HL/5 media (Sussman 1987) at 20 degrees. D. discoideum cells from mid-logarithmic cultures were collected by centrifugation(1,000 x g; 4 min), washed once with SM/5 medium (Sussman 1987), and added to the bacterial suspensions at a final concentration of 5 x 102cells/ml suspension; 0.2 ml of this mixture was plated out on SM/5 plates. Plates were incubated for 35 days and examined for plaques formed by amoebae.

2.4.13 Database submission

The nucleotide sequences of the chromosome of *Acinetobacter baumannii* ATCC17978 and its two plasmids, pAB1 and pAB2, are being submitted to GenBank.

Chapter 3

Comparative Genomics: Integation of curated databases to identify genotype-phenotype associations

3.1 Introduction to microbial phenotype prediction

The ability to rapidly characterize an unknown microorganism is critical for both responding to infectious disease and biodefense. Such characterization requires a method for anticipating or predicting an organism's phenotype based on the molecules encoded by its genome. However, the link between molecular composition (i.e. genotype) and phenotype for microbes is not obvious. Traditionally, microbes have been identified on the basis of their response to a battery of phenotypic assays, for example, survival on a particular type of growth media or morphological characteristics. However, with the advent of high throughput sequencing efforts, over 300 microbes have been completely sequenced (Benson, 2005). By integrating complex phenotypic data with sequence information, an approach can be developed to identify new phenotype-genotype relationships. Such relationships will both increase our understanding of the mechanisms of the phenotype and perhaps provide more sensitive assays.

The underpinnings for this work can be found in Marcotte *et al.* where phenotype was defined in terms of pathway membership, which was used to predict protein function (Marcotte, 1999). In addition, previous studies have proposed comparative genomic methods to predict characteristics such as hyperthermophily (Makarova, 2003; Jim, 2004), flagellar motility (Jim, 2004; Levesque, 2003; Korbel, 2005), plant degradation (Korbel, 2005), and pili assembly (Jim, 2004). However, most of these studies focused on a few specific phenotypes within certain organisms (Makarova, 2003; Jim, 2004; Levesque, 2003). Korbel *et al.* proposed an automated method to make word-species associations retrieved from Medline abstracts (Korbel, 2005); however, here we present a systematic approach to discover genotype-phenotype associations that combines phenotypic information from a biomedical informatics database, GIDEON, with the molecular information contained in National Center for Biotechnology Information's Clusters of Orthologous Groups database (NCBI COGs).

Using this approach, we identified phenotype-COG association pairs and verified these findings in the literature. Finally, this analysis suggested possible phenotype-genotype pairs that have not yet been experimentally determined. By integrating a clinical microbiological database, GIDEON, with a molecular database, COGs, we can make inferences between the presence of a protein and the protein's function in a large-scale fashion.

3.2 Results

3.2.1 Identifying Phenoype-Genotype Associations

To identify associations between the presence or absence of a particular gene (COG) in a microbial genome and that microbial species' expressed phenotype (GIDEON), we computed the correlation between the measured expression of a certain phenotype to the absence or presence of COGs (genomic content) in that microbial species and filtered for significant correlations ($P < .01$, figure:fig-gideon-schematic).

We generated two separate result sets ($r_2 = 0.8$ and 0.9, containing 290 and and

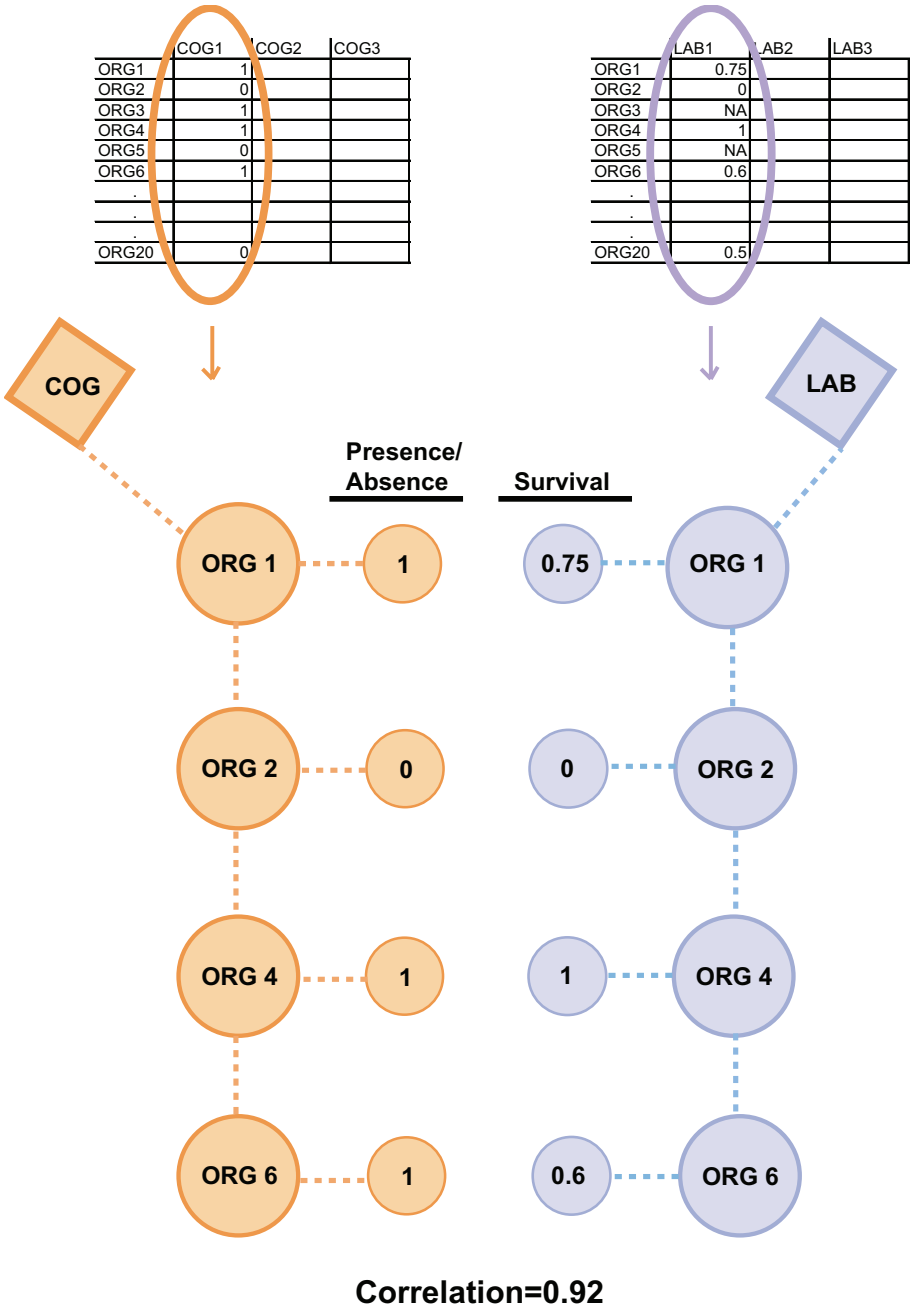


Figure 3.1: Diagram of correlation analysis for associating COGs to lab condition phenotypes. The correlation analysis measures the association between a COGs organism profile (presence or absence of an organism) and a lab conditions organism survival profile. Organisms that have a COG (red) are mapped to the organisms response to adverse growth conditions (blue) creating two vectors that are used for the correlation calculation.

Data Set	Total Number of Associated Pairs	Number of COGs with Known Function	Number of Pairs Randomly Selected for Literature Search	Number of Pairs with Confirmed Association in the Literature	% Confirmed Pairs
Corr Scores >=0.8	290	154	100	66	66%
Corr Scores>=0.9	74	36	36	31	86%

Table 3.1: Number of validated associations at the 0.8 and 0.9 threshold

74 association pairs, respectively). One hundred and thirty-six random data pairs were selected from the 0.8 and 0.9 result set, respectively for literature validation. We refer to verified association pairs as annotated pairs. Sixty percent for the 0.8 set and 86% of the 0.9 set were confirmed (Table 3.1). Below, we give several examples of these annotated pairs (Table 3.3). For simplicity, the laboratory conditions are referred to by their GIDEON identifier/phenotypic description, and a COG with a known function is defined as characterized.

3.2.2 Examples of Annotated Association Pairs

A) *B01/Gram-negative* - Sixteen of the annotated COG-phenotype pairs (Table 3.2) are involved in the B01/Gram-negative phenotype. This resulted in 94% accuracy for determining Gram-negative organisms. Gram negative bacteria differ from Gram-positive in the composition of their cell wall (Beveridge, 1999). Perhaps unsurprisingly, the confirmed pairs found with the Gram-negative phenotype contained proteins involved with lipid A and lipopolysaccharide biosynthesis and other proteins belonging to the outer membrane of Gram-negative bacteria (Table 3.3).

B) *B02/Gram-positive* - More interestingly, annotated pairs with the B02/Gram-positive phenotype were not just specific to the specialized Gram-positive membrane but also to a variety of conserved genes found only in the Gram-positive bacteria. Of the six proteins with known function found in the 0.8 correlation score data set, 3 pairs were positively confirmed by the scientific literature. In the 0.9 correlation result set, 100%

(2/2) of the characterized pairs were corroborated.

C) *B29/Growth on MacConkey Agar* - Growth on MacConkey agar is indicative of Gram-negative bacteria that can ferment lactose (MacConkey, 1905). Sixteen (52%) of the associated pairs from the 0.8 correlation data were confirmed, as were 4 (100%) of the associated pairs from the 0.9 correlation set Organisms that were able to grow on MacConkey agar contained proteins involved with the outer membrane of Gram-negative bacteria. Given the use of this test in specifically differentiating those Gram-negatives that can ferment lactose, one would expect this result; however, the proteins associated with growth on MacConkey agar do not overlap with those proteins most associated with the Gram-negative test. This suggests that this method of building associations can be specific to a particular condition.

D) *B30/Oxidase* - Two characterized COGs were positively associated ($P < 7.48x10_{-6}$) with oxidase activity. Both are components of Cbb3-type cytochrome oxidase, which is unsurprising since the goal of the oxidase test is to detect the presence of this enzyme. Although this is not a novel finding, it does illustrate the ability of the method to recaptitulate known relationships.

E) *B31/Catalase* - In the catalase test, hydrogen peroxide is added to the media. Those microbes that do not contain the catalase enzyme are unable to break the hydrogen peroxide into oxygen and water and die. As would be expected, the COGs associated to the B31/Catalase test were usually enzymes that belong to similar regulation pathways as the catalase enzyme. For example, human acyl-CoA hydrolase, one of the COGs found to be positively associated to the catalase phenotype, upregulates peroxisome biogenesis and, in turn, activates catalase activity (Alexson, 1989). The highest scoring pair was a member of the catalase protein family. For both the 0.8 and 0.9 correlation result sets, the confirmation percentages were 64% and 63% respectively.

F) *FAC/L-Arabinose* - With a high correlation score of 0.97, 5-keto 4-deoxyuronate isomerase was the only characterized protein family associated with the ability to assimilate arabinose. 5-keto 4-deoxyuronate isomerase, or kduI, is an enzyme involved with pectin

degradation and shares the same regulator protein, crp or CAP protein, as the L-Arabinose catabolism pathway (Bankaitis, 1981; Nasser, 1997).

G) *FAT/Trehalose* - The COGs most associated with the capability to metabolize the sugar trehalose were several maltose-related proteins with correlation scores of 0.94. It has previously been shown that addition of trehalose to growth media induces the maltose system verifying both of these associations (Boos, 1998).

H) *G03/Motile* - Finally, the pairs related to the G03/Motile phenotype contain proteins involved with chemotaxis and flagella. The result set with a correlation score above 0.8 contained 17 such proteins, and 100% of these were verified by the literature. Similarly, all 5 proteins from the result set with a 0.9 threshold were also confirmed.

Additionally, the 0.8 and 0.9 correlation score threshold data sets for motility were compared with the KEGG database (Kanehisa, 1997; Kanehisa, 2000). This analysis revealed that 100% of the proteins found to be associated with motility were also annotated as part of the Cell Motility functional classification in the KEGG pathway database.

3.2.3 Prediction of genes associated to phenotypes

After analyzing the accuracy of the data sets, it is also possible to make reasonable hypotheses for COG-phenotype pairs that are characterized but have not yet been confirmed by the biological literature. These COG-phenotype pairs are listed using their GIDEON identifier/description-COG description/protein name. One example is the B31/Catalase-COG1651/Protein-disulfide isomerase (DsbG) pair with a correlation score of 0.91. Dsb proteins are known to oxidize the sulfhydryl groups of periplasmic proteins to disulfide bonds, donating electrons to ubiquinone, and thereby making the electron transport chain the primary source of oxidizing power for sustaining periplasmic sulfhydryl oxidation (Rietsch, 1998; Bader, 1999). During the stationary phase, electron transport to oxygen is reduced. Bandyopadhyay et al. suggest a possible complementary role between catalase and the Dsb proteins in maintaining periplasmic sulfhydryl oxidation. It is possible that catalase may be critical in peroxidatically oxidizing ubiquinol or another periplasmic

Lab/Condition	Correlation Above 0.8			Correlation Above 0.9		
	Total Characterized Pairs	Confirmed Pair Associations	Percent Confirmed	Total Characterized Pairs	Confirmed Pair Associations	Percent Confirmed
B01/Gram-negative	17	16	94%	12	12	100%
B02/Gram-positive	6	3	50%	2	2	100%
B28/ Growth on Ordinary Blood Agar	5	0	0%	NA	NA	NA
B29/Growth on MacConkey Agar	31	16	52%	4	4	100%
B30/Oxidase	2	2	100%	NA	NA	NA
B31/Catalase	11	7	64%	8	5	63%
FAC/L-Arabinose	1	1	100%	1	1	100%
FAJ/Lactose	1	0	0%	NA	NA	NA
FAL/D-Mannitol	1	0	0%	NA	NA	NA
FAM/D-Mannose	2	2	100%	NA	NA	NA
FAP/L-Rhamnose	1	0	0%	2	0	0%
FAT/Trehalose	2	2	100%	2	2	100%
FAU/D-Xylose	1	0	0%	NA	NA	NA
G03/Motile	17	17	100%	5	5	100%
G14/ Nitrate to Nitrite	1	0	0%	NA	NA	NA

Table 3.2: Accuracy of associations confirmed by literature broken down by individual condition. Characterized are those pairs where the COG has a known function. Confirmed are those associations that were verified in the literature.Number of validated associations at the 0.8 and 0.9 threshold

or inner membrane component using H2O2 as an electron acceptor during the stationary phase when the oxidizing capacity of the electron transport is diminished (Bandyopadhyay, 2000).

With a correlation score of 0.95, other possible associations can be made for the FAP/L-Rhamnose phenotype with various phosphotransferase system sorbitol-specific component proteins. Some microbes such as the *Klebsiella I-174* make exopolysaccharides with a high rhamnose content (Morin, 1990). Farres *et al.* showed that the addition of sorbitol increased the production and growth of rhamnose over other carbon sources such as sucrose (Farres, 1997). This study suggests that proteins involved with sorbitol metabolism and utilization could be linked to rhamnose production.

3.3 Discussion

Based on the breakdown of total number of associated pairs for each laboratory condition (Figure 3.2) for the 0.8 correlation data set, the phenotypes that have 10 or more associated COGs have a more likely chance of containing confirmed literature hits. This is roughly 3% of the total number of phenotype-COG pairs. However, there are labs such as B30/Oxidase, FAM/Mannose, and FAT/Trehalose with only 2 results, but all are confirmed at 100%. The 0.9 correlation data set has 86% confirmed associations out of all the characterized pairs, while the 0.8 correlation data set has 66%.

This study reports a percentage of confirmed associations in order to approximate the accuracy of these results. However, this number is most likely a lower bound, since it is possible that some of the predicted associations mentioned in this paper will be experimentally corroborated in the future, raising these percentages.

In addition, although we used the literature as a means of verifying associations, in essence, it is those associations which we were unable to verify that are perhaps the most interesting because these represent new testable hypotheses. By uncovering these novel relationships, it is possible to make inferences about the interrelatedness of what

Lab/Condition	Protein Name	r^2	P	Protein Function
B01/Gram-negative	Lipid A disaccharide synthetase	0.95	2.E-09	Involved in Lipid A biosynthesis
	UDP-3-O-acyl-N-acetylglucosamine deacetylase	0.95	2.E-09	Involved in Lipid A biosynthesis
	CMP-2-keto-3-deoxyoctulosonic acid synthetase	0.95	2.E-09	Involved in lipopolysaccharide biosynthesis
	3-deoxy-D-manno-octulosonic acid 8-phosphate synthase	0.95	2.E-09	Involved in lipopolysaccharide biosynthesis
	Biopolymer transport protein	0.95	2.E-09	Outer membrane transporters
	Outer membrane protein	0.84	2.E-09	Outer membrane protein
B02/Gram-positive	Sortase	1	3.E-08	Plasma membrane protein
	AT-rich DNA-binding protein	0.92	8.E-07	Transcriptional regulator
	D-alanine transfer protein	0.84	1.E-05	Cell wall/membrane component protein
B29/ Growth on MacConkey Agar	Outer membrane cobalamin receptor protein	0.99	8.E-09	Outer membrane protein
	Flagellar basal body rod protein	0.97	2.E-07	Periplasmic protein
	Tfp pilus assembly protein	0.83	1.E-05	Outer membrane proteins
B30/Oxidase	Cbb3-type cytochrome oxidase, 1 and c	0.85	8.E-06	Oxidase protein subunit
	Acyl-CoA hydrolase	0.97	8.E-06	Enzyme involved in lipid metabolism
B31/Catalase	Catalase	0.97	8.E-06	Peroxisomal Marker Enzyme
FAC/L-Arabinose	5-keto 4-deoxyuronate isomerase	0.97	4.E-05	Enzyme involved in carbohydrate metabolism
FAM/D-Mannose	Mannitol-1-phosphate/altronate dehydrogenases	0.85	5.E-05	Oxidizes mannitol to mannose
FAT/Trehalose	Maltose-binding periplasmic proteins/domains	0.94	3.E-05	Maltose-related protein
G03/Motile	Chemotaxis signal transduction protein	0.94	5.E-09	Chemotaxis-related protein
	Flagellar capping protein	0.94	5.E-09	Flagella-related protein
	Flagellin-specific chaperone FliS	0.94	5.E-09	Flagella-related protein

Table 3.3: Accuracy of associations confirmed by literature broken down by individual condition. Characterized are those pairs where the COG has a known function. Confirmed are those associations that were verified in the literature.

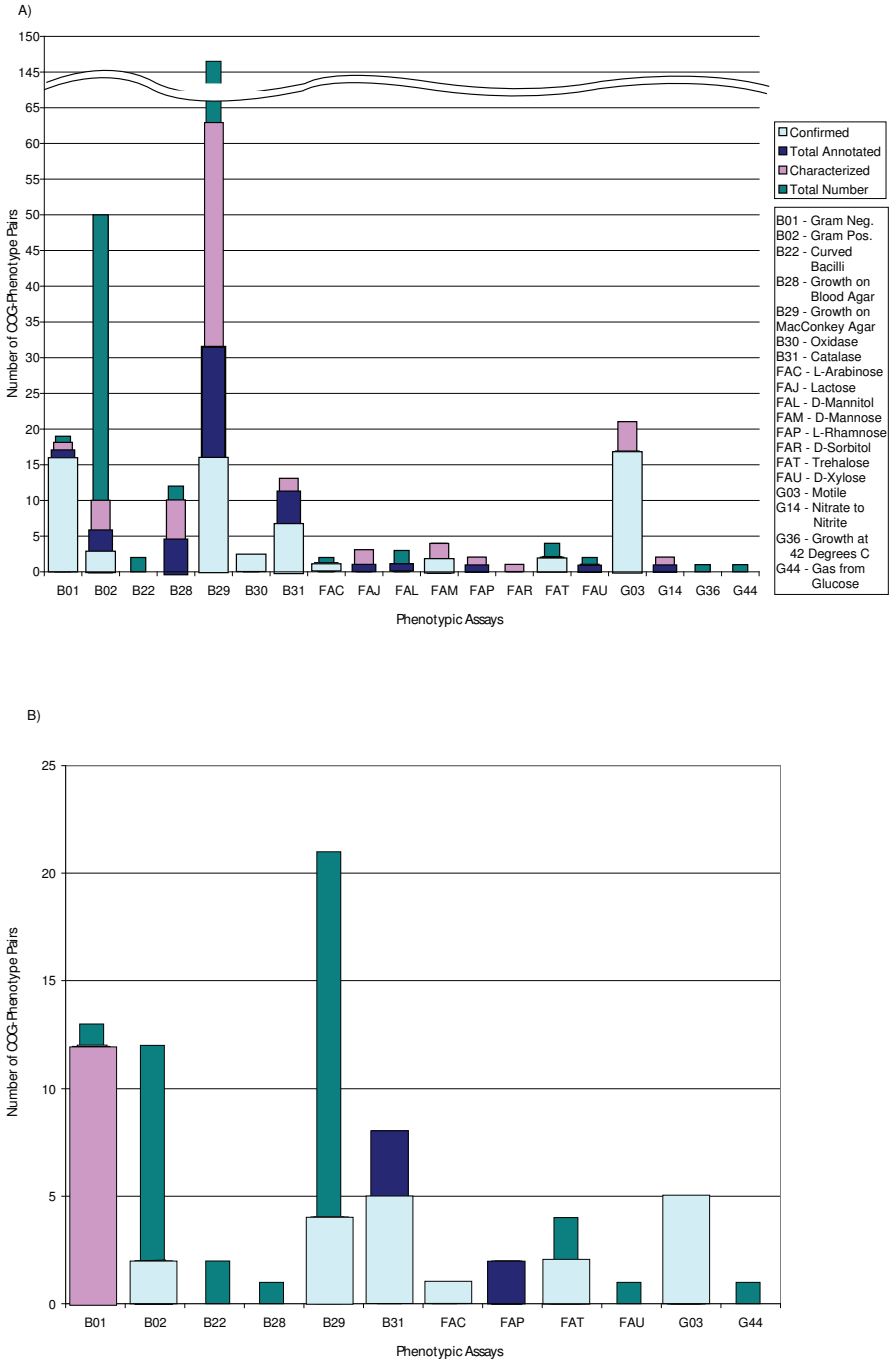


Figure 3.2: Number of COG-phenotype associated pairs in each subset of the 0.8 and 0.9 threshold correlation score data sets. The resulting data sets of the (a) 0.8 correlation threshold and the (b) 0.9 correlation threshold are broken down into four different subsets. Total number (dark blue) is the total number of COG-phenotype associated pairs found at the 0.8 and 0.9 thresholds respectively. Characterized (light purple) refers to those pairs where the COG has a known function. Annotated (blue-green) are those pairs which were selected for literature verification. Finally, confirmed (light blue) are the associations which were validated in the literature. This is shown for each lab indicated with by its GIDEON identifier.

at the outset seem disparate processes. In a similar fashion, for the purpose of assessing our method we were unable to include the COGs with unknown function, but ideally we would like to extend this method to make predictions regarding possible functions of these uncharacterized COGs on the basis of the phenotypes they are most associated with. Finally, while the data in the GIDEON database is extensive, not all assays were performed on all microbes resulting in some missing data.

3.4 Conclusions

This analysis shows that the integration of biological and biomedical information databases can augment and enhance biological understanding. This approach is an introduction to resources that are yet to be fully utilized. Here we describe the combination of a manually annotated phenotype database, GIDEON, with the well-documented COG database to find new associations between a certain phenotype and a microbial genotype. We have demonstrated that the method is able to detect known phenotype-COG relationships, as well as, discover new ones.

These results suggest a new direction for inferring either the phenotype or genotype of an uncharacterized organism. This approach can further be applied to discovering relationships between the pathogenicity of these organisms to functionally related proteins. Moreover, this type of analysis could be extended beyond phenotype-genotype to phenotype-drug design by associating molecules to their phenotypic effects. By integrating clinical and biological databases, additional studies can be developed to further the understanding of phenotypic relationships and, in turn, augment the medical community's ability to rapidly identify infectious agents.

3.5 Materials and Methods

3.5.1 GIDEON and COG Database Descriptions

The Global Infectious Diseases and Epidemiology Network (GIDEON) is an expert system used primarily by physicians to aid in the diagnosis of infectious diseases (Berger, 1993). This database was chosen because it provides an exhaustive, hand-curated categorization of microbial phenotypes. GIDEON catalogs the results of 93 different microbiological assays for 1147 microbial taxa providing a wealth of phenotypic data.

NCBI's Cluster of Orthologous Groups of proteins (COGs) database currently consists of 138,458 proteins, which form 4873 COGs (Tatusov, 1997; Tatusov, 2000; Tatusov, 2003). This database uses orthology to group proteins from completely sequenced prokaryotes into COGs. All the newly classified COGs and new members of pre-existing COGs are manually curated.

3.5.2 Mapping organisms between databases

The laboratory results in the GIDEON database are primarily used for identifying bacterial species for medical diagnostics. Since different strains of the same bacterial species are often sequenced, NCBI's taxonomic annotation is sometimes at the subspecies level. In contrast, the GIDEON phenotypes do not achieve such a high resolution, and for this reason GIDEON taxonomic annotation is established at the level where the phenotype is consistent in all descendants of the phylogenetic tree (generally the species level).

This presented a complication in integrating the two data sources. To overcome this, we assumed that phenotypes from the microbiological database for one species are valid for every subsumed subspecies and strain listed in the COGs database. This is a valid assumption since the GIDEON dataset provides microbiologists with relevant tests designed to distinguish between organisms according to their phenotypes. Thus if the phenotype is specific to a subspecies, it will be annotated at the level of the subspecies, but if the phenotype is common to all subspecies, it is recorded at the level of the species.

Following this principle, we first identified the taxonomic level for the fully sequenced bacteria in the COGs dataset, and then used text string matching followed by manual examination to map the species in GIDEON (Lussier, 2004). As a result, we have mapped the 37 microorganisms present in both GIDEON and the COGs. Of the 37 mappings, 23 have identical species annotations in GIDEON and COGs, and 9 have a species annotation in GIDEON mapped to one or more subspecies in COGs.

There were several COGs species including *H. pylori*, *E. coli*, *M. tuberculosis*, and *N. meningitides* which had more than one subspecies with complete genome sequences. In these cases, the subspecies were merged to the single GIDEON species by selecting only the COGs common to all subspecies. In this manner, we eliminated the subspecies specific differences that the phenotypic assays would have been unable to resolve. We generated a matrix showing the presence and absence of COGs across these 37 species.

3.5.3 Associating genes to phenotypes

We employed a correlation analysis to quantify the association between a given COG and a GIDEON phenotype. Two matrices were constructed. We defined X as a two-dimensional matrix indicating the presence or absence of organisms within a COG (X was constructed as an MxN matrix, where M is equal to the number of COGs and N is equal to the number of organisms within a COG). For the corresponding GIDEON lab conditions, a similar distance matrix, Y, was constructed as an NxL matrix, where N is equal to the percent survival of organisms subjected to each lab condition and L is equal to the number of lab conditions. X_{ij} is the presence or absence of a COG m_i within an organism n_j, and Y_{kj} signifies the percent response of an organism n_j under a certain lab condition l_k. We computed an MxL matrix of linear correlation coefficients r_{ik}; the hypergeometric distribution was used to test for significance.

3.5.4 Assessment of predicted results

The following criteria were applied to the correlated data set. The intersection between a specific COG and a phenotype had to contain at least 3 organisms, and for any intersection, 30% of the microbes had to share the COG. The scores were adjusted using the standard Bonferroni error correction for multiple testing.

For the 0.8 correlation threshold, 290 total associations were obtained. We identified a subset of these data (154) that contained only COGs that have a known function. Of these 154 pairs, we performed detailed literature searches on 100 randomly selected pairs to confirm the validity of the positive COG-phenotype associations.

There were 74 associations found in the 0.9 correlation data set. Thirty-six of these pairs contained a COG of known function. Literature searches were performed on all 36 associations.

3.5.5 Data Deposition

All data files and additional tables are available from http://gersteinlab.org/proj/phenome

Chapter 4

Network Dynamics: Quantifying environmental adaptation of metabolic pathways in metagenomics

4.1 Introduction

Microbes function as highly interdependent communities. Fundamental to the maintenance of the ecosystem's energy balance, the recycling of nutrients, and the neutralization and degradation of toxins and other detritus (Karl, 2002), microbial community processes are intimately intertwined with ecosystem functioning. Thus, it is critical to understand the complex interplay between the environment's influence on microbial communities and microbe's reshaping of their environment.

Until recently the tools to systematically study global community function and environment at the molecular level were not available, since complex microbial communities are generally not amenable to laboratory study(Allen, 2005). The recent advent of direct

sequencing of environmental samples (i.e. metagenomics) has allowed the first large-scale insights into the function of these complex microbial communities.

Comparative metagenomics approaches have revealed significant variation in sequence composition (Foerstner, 2005), genome size (Raes, 2007), evolutionary rates (von Mering, 2007), and metabolic capabilities (Tringe, 2005; Dinsdale, 2008; Rodriguez-Brito, 2006) among qualitatively dissimilar environments (e.g. terrestrial vs. marine) providing evidence for genomic adaptations. Further, variation in specific community biological processes have been shown for different water column zones at a single geographic site (DeLong, 2006), different climatic regions in the ocean (Rusch, 2007), and more recently, among nine ecosystems (Dinsdale, 2008).

The wealth of information generated from these studies emphasizes the importance of investigating relative differences in biological processes among qualitatively different environments. However, to date none of them have directly incorporated multiple, specific measurements of the environment. By treating the environment explicitly as a set of complex, continuous features rather than relying on an implicit subjective classification, one can build models to determine how a diverse array of biochemical activities, and particularly metabolic versatility, reflect sets of or specific environmental differences.

Providing an ideal dataset for exploring these environmental-biochemical links, the Global Ocean Survey (GOS) collected quantitative environmental features and metagenomic sequences from over 40 different aquatic sites (Rusch, 2007). Here, we used GOS data to investigate and develop multivariate approaches to systematically relate metabolic pathway usage directly to quantitative environmental differences. These approaches allowed us to address multiple relationships simultaneously as well as to relate specific environmental features to metabolic processes at different levels of resolution including 14 broad functional categories, 111 pathways, 141 modules (sections of pathways), 191 operons, and 15554 orthologous groups. By identifying environmentally-dependent pathways involved in energy conversion, amino acid metabolism, and cofactor synthesis, among others, we were able to define metabolic footprints of distinct environments. Our

study provides an analytical framework for uncovering ways in which microbes adapt to (and perhaps even) how they change their environment.

4.2 Results

4.2.1 Quantitative approach for footprint detection

We mapped thirty-seven size-filtered GOS sites (Table 4.4) to their respective environmental and metabolic features at several levels of complexity (pathways, modules, operons; Figure 4.1A-B). These data can naturally be represented as matrices where the rows are geographic sites and the columns are either environmental or metabolic features. We inter-related these matrices to examine how pathway usage across different sites is related to environmental parameters. The simplest and most direct approaches for performing such operations are correlation and regression (see Materials and Methods for comparisons with other types of methods). Thus, we examined the first order relationships by computing the pairwise correlation between each metabolic pathway and each environmental feature (e.g. photosynthesis and temperature). Note that for clarity, we use the word *pathway* to refer to the usage of the pathway, as in photosynthesis as opposed to *usage of photosynthesis*, in the remainder of the text. This analysis revealed a number of significant correlations (environmentally-dependent pathways). Such pathways were used to build linear models (LM) of each environmental feature. Although these models performed well in predicting single environmental features (Figure 4.2), there are limitations to viewing each environmental measurement in isolation, as there are hidden dependencies among the environmental features.

To discover the complex, higher order interactions between and within environmental features and metabolism, we used a second complementary approach, regularized canonical correlation analysis (CCA). CCA has two primary functions: (1) to determine if a global relationship between two types of features (here environmental and metabolic) exists and (2) to calculate the relative contribution of each feature to the global relationship (e.g.

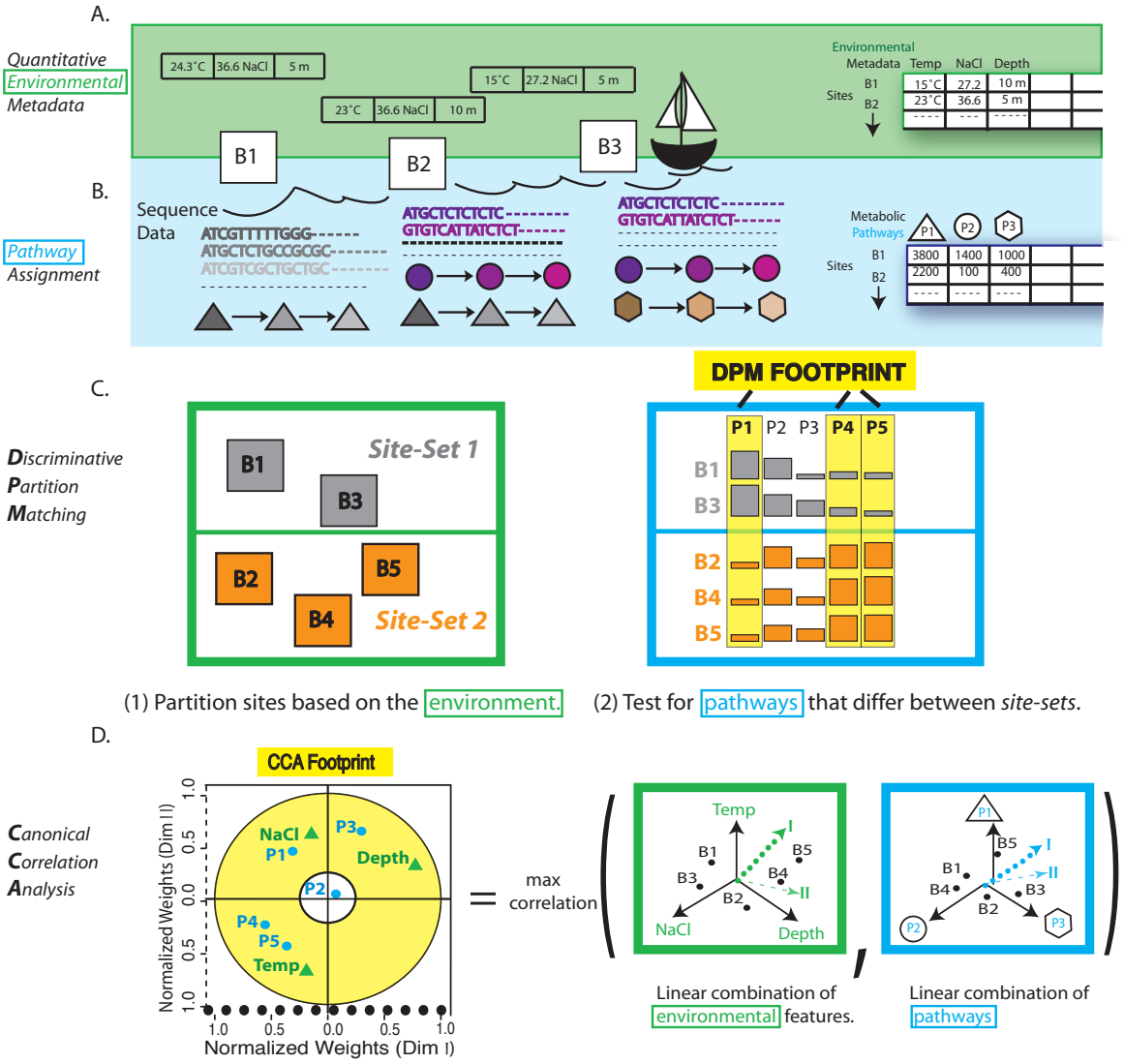


Figure 4.1: Illustrated schematic of approach. The large squares labeled B1, B2, etc. represent the geographic sites (buckets). Each bucket has sequence and environmental feature data associated with it. (A) Panel A illustrates the mapping of quantitative environmental features (*salinity (ppt), sample depth (position in water column from which the sample was collected), water column depth (measured from surface to floor), and chlorophyll)*, (B) Panel B shows the mapping of reads to pathways (see Materials and Methods) Reads are color coded according to their corresponding pathway elements (shapes). Different pathways are represented by different shapes (square, circle, etc). All the instances of a particular pathway are summed and normalized to compute the pathway score. (C) Schematic of discriminative partition matching (see details in text). (D) Schematic of canonical correlation analysis (see details in text).

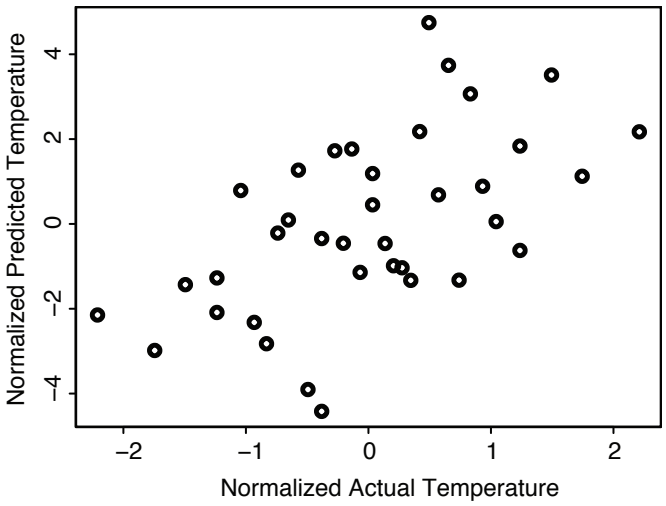


Figure 4.2: Predicting specific environmental parameters from subsets of metabolic pathways. Linear model for temperature built from subsets of highly correlated pathways including N-acetylglucosamine biosynthesis, many components of amino acid metabolism, and fatty acid biosynthesis. Axes are normalized actual and predicted temperature for x and y, respectively.

temperature or photosynthesis) by weighting both sets of features simultaneously. In brief, CCA computes a linear combination for each feature set and simultaneously attempts to maximize the correlation between the two feature vectors (Figure 4.1D). Thus, CCA is able to simultaneously assess relationships both between and among the environmental features and metabolic pathways. Since the sites are quite similar, we developed a more robust but less sensitive method called discriminative partition matching (DPM). DPM first partitions the sites into site-sets on the basis of their environmental parameters, then tests which pathways give the greatest discriminatory power among the site-sets (Figure 4.1C).

4.2.2 Footprint Characteristics

The goal of DPM and CCA is to simultaneously explore the relationship between metabolism and the *quantitative* environmental parameters by identifying environmentally-dependent or co-varying metabolic pathways (footprints). The main difference between DPM and CCA is that DPM identifies those pathways that discriminate the best between

site-sets, but when defining the *site-sets*, all the environmental variables are considered equally important. Thus, although robust to noise, DPM is more coarse-grained and at this resolution the individual differences among sites and their relationship to the environment can be lost. In contrast, CCA can highlight these individual differences by weighting each environmental feature and each metabolic pathway independent of any partitioning making it both more sensitive but also more susceptible to noise (Figure 4.1D).

4.2.3 DPM Footprint

Applying DPM, the sites were partitioned into two different site-sets that can loosely be classified as open ocean and coastal. We found the distribution of those COGs and KEGG maps annotated as having a role in metabolism were significantly different between *site-sets* ($P < 9x10^{-3}$ and $4x10^{-14}$, respectively); however, no statistically significant difference was found for control matrices that were composed of translational/transcriptional machinery (see Materials and Methods).

Further, we find 10 KEGG maps, 24 modules, 61 operons, and 98 gene families were significantly different (FDR-corrected $q < .05$) between the two *site-sets*. These pathways together form the DPM footprints. By examining the broader trends of these footprint pathways, we found that secondary metabolite biosynthesis, lipid transport and metabolism, amino acid metabolism, and energy production and conversion were significantly different between site-sets. Finally, we showed the cluster similarity between the environment-based site partitioning and metabolic footprint-based site partitioning was quite high (normalized mutual information = 0.46, rand index = 0.76, $P < .001$) suggesting that footprints have predictive power in recapturing features of the environment based purely on pathways identified as significant in DPM.

4.2.4 CCA Footprint

Next, we applied regularized CCA to measure the strength of each metabolic pathway's co-variation with environmental features (Figure 4.3). We identified 22 KEGG maps, 53

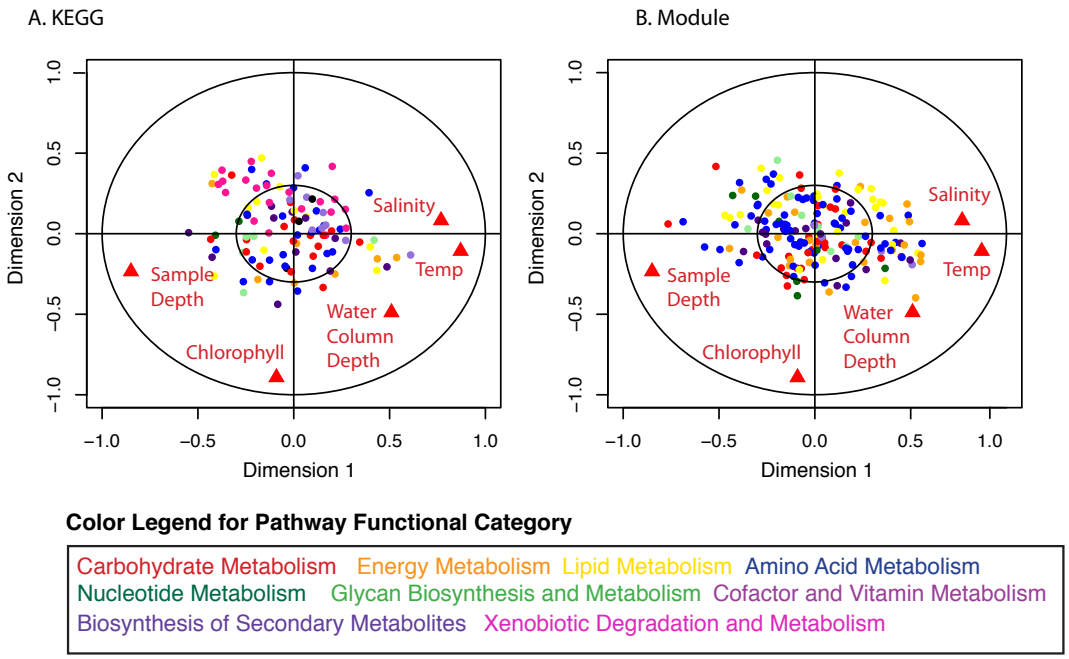


Figure 4.3: Bullseye plot of CCA-derived structural correlations. Results from CCA for (a) KEGG and (b) module. The x and y-axes represents the structural correlation coefficients (normalized weights) in the first and second dimension, respectively. The closer either environmental features (red triangles) or metabolic pathways (circles color coded by functional category) are to the perimeter of the outer circle, the better they fit the model. In addition, the closer an environmental feature is to a metabolic pathway the stronger the co-variation between them. The inner circle (radius 0.3) represents those features that did not fit the model (see Gonzalez for further explanation). Those pathways in the inner circle can be thought of as environmentally invariant, and those outside this circle as environmentally variant.

modules, and 35 operons as being environmentally-dependent (absolute value of structural correlations > 0.3; Figure 4.4). These pathways form footprints that can be investigated for subtler environment-based changes in metabolic capabilities. In this manner, we identified diverse functional processes that co-varied significantly with the environment including xenobiotic degradation, energy conversion, lipid metabolism, and amino acid metabolism.

4.2.5 Adaptation of energy conversion strategies to specific environmental challenges

Many of the environmentally-dependent pathways were associated with energy conversion. The diversification in energy conversion strategies is reasonable given that a primary challenge to all microbial communities is how to maintain adequate energy reserves despite challenging conditions in their specific environment.

Our results demonstrate ample diversification in energy conversion strategies linked to such quantitative environmental differences. In particular, we show that proteins involved in (photo)autotrophic processes, such as photosynthesis, oxidative phosphorylation, and carbon and nitrogen fixation, are strongly influenced by variation in environmental parameters (Figure 4.4). This link is seen at all functional levels and reinforced by multiple methodologies (Table 4.1). The module-level analysis showed that only photosynthetic modules involved in light capture and electron transport (photosystem I, II and the cytochrome b6/f complexes) correlated with the environment. In contrast, the abundance of the module for the ATP synthase complex, whose function is independent of the particular energy conversion strategy, does not change significantly (Figure 4.4A). A similar trend can be seen for oxidative phosphorylation, albeit not as strongly (Figure 4.4B). The seeming lack of environmental constraint on the ATP synthases probably reflects their role in coupling energy to a proton gradient (e.g. oxidative phosphorylation, etc) that are required regardless of which specific energy conversion strategy is employed. Furthermore, in some cases our approach allows the three-way linking of functional, phylogenetic and environmental patterns. For instance, in respiratory complex I, the module covering the

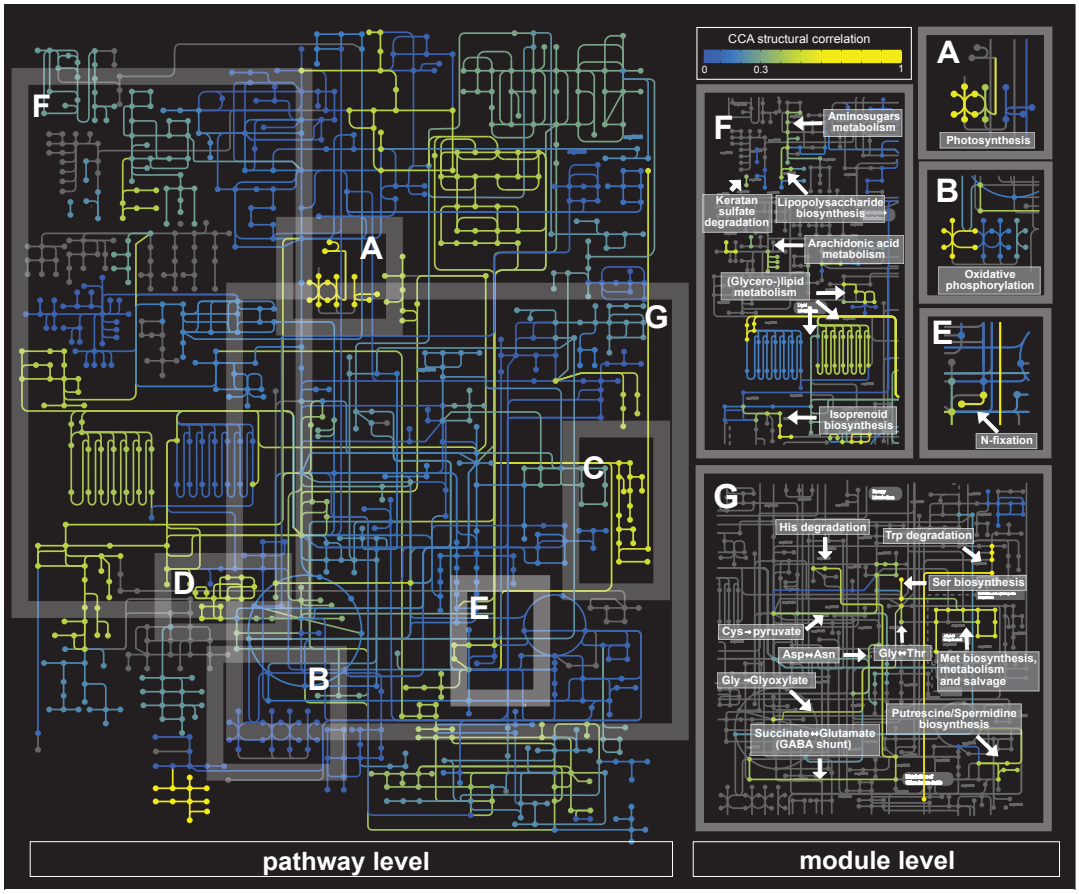


Figure 4.4: Map of Structural Correlations. Central panel is a plot of the environmental features (triangles) and pathways (circles) where the x-axis and y-axis are the structural correlation coefficients (normalized weights) derived from CCA for the first and second dimension, respectively (see Figure 4.1D). The remainder of the figure depicts the strength of both pathways' (KEGG, right) and sections of pathways' (modules, left) environmental co-variation as measured by the absolute value of normalized weights (color-coded yellow strongest to blue weakest) (see interactive version of this map http://pathways.embl.de/metagenomics). Nodes symbolize compounds, and lines connecting nodes are enzymes. All enzymes (lines) corresponding to a single KEGG map or a single module will have the same color. Shaded gray boxes (A-G) for pathways and corresponding boxes for modules (no modules available for C and D) denote examples from the text: energy conversion (A-E), amino acid metabolism (G), and lipid synthesis and glycan metabolism (F). Photosystem I and II modules (box A, bright yellow to green) show significant co-variation with the environment, but the ATPase is invariant (blue). Similar pattern observed for oxidative phosphorylation (B, see text for more details). Box C highlights pieces of the photosynthetic machinery (including heme/porphyrin synthesis) and D shows carbon fixation. Glycerophospholipid pathways (F) shows only the "pipe" leading to or from the citrate acid cycle co-varies. G highlights amino acid metabolic pathways discussed in the text. Map generated using (Letunic, 2008)

.

cyanobacterial NADH dehydrogenases (i.e. most likely those from *Prochlorococcus*-like species) co-varies positively with temperature and other photosynthesis modules. However, the module covering the proteobacterial NADH dehydrogenase (i.e. most likely from SAR11-like species) varies inversely with the temperature gradient. Such observations can be associated with their respective geographic distributions. Photosynthetic *Prochlorococci* are mostly absent in the northern, temperate sites but dominate in tropical waters (Rusch, 2007; Johnson, 2006); whereas, SAR11-like proteobacteria, which do not rely on the classical photosynthetic machinery to collect energy, dominate the northern, temperate regions (Giovannoni, 2005). Thus, although variation in the reliance on autotrophic processes is not unexpected, these observations illustrate the potential of the proposed methodology to detect biologically relevant co-variation.

4.2.6 Balancing amino acid synthesis vs. import: adapting to nutrient-limited conditions

We observed that metabolic pathways associated with amino acid and cofactor transport and metabolism varied significantly with the environmental features. Given the oligotrophic nature of the oceans (Stocker, 2008), this observation may reflect the variability in amino acid uptake and recycling pathways as an alternative nutrient source in the various environments sampled; a strategy used by many of the dominating species in ocean surface waters (Mary, 2008). Lending further support to this hypothesis, operons with significant structural correlations consisted of both amino acid metabolism pathways and transporters necessary for exogenous uptake (Table 4.2). Amino acid uptake is sensitive to light availability (Mary, 2008) which, given the north to south sample collection gradient, could be an additional factor in their variation. The strength of this co-variation is further reflected by the positioning of many of the amino acid metabolism maps along the same principal axis as temperature and chlorophyll in the positive direction (Figure 4.3).

Level	Id	Description	Footprint		
			CCA	DPM	LM
K	00710	Carbon fixation (including dark reaction)	✓	✓	✓
M	10297		✓		✓
M	10274	Complex I NADH dehydrogenase	✓		✓
M	10284	Cytochrome b6f/c complexes	✓		✓
M	10291		✓		✓
S	01937	Ferrodoxin oxidoreductase	✓		
K	00941	Flavonoid biosynthesis	✓		
M	10625		✓		
S	03655	Heme/quinol biosynthesis	✓		
M	10305	Nitrogen fixation (plus transporter)		✓	
S	04345		✓		
K	00190	Oxidative Phosphorylation	✓		✓
K	00195	Photosynthesis (including Photosystem I and II)			✓
M	10292		✓		✓
M	10290		✓		✓
K	00860	Porphyrin and chlorophyll metabolism	✓		✓
K	00130	Ubiquinone/Thiamine biosynthesis	✓		✓
S	02892		✓		

Table 4.1: Pathways involved in energy conversion with significant environmental covariation.

4.2.7 Environmentally variant/invariant amino acid pathways differ by cofactor cost

One of the most striking aspects of our findings is that amino acid biosynthetic pathways could be divided into those that vary with the environment (high structural correlation coefficient) and those that do not. Interestingly, co-variation of amino acid biosynthesis with the environment was unrelated to the energetic cost of synthesizing a particular amino acid (e.g. metabolic optimization). This simple result is seemingly counter-intuitive as one would expect that those pathways that used the most energy might vary the most with the energetic potential of their environment. However, we observe a significant positive correlation ($P < 0.05$) between the structural correlation of the amino acid pathways (strength of environmental co-variation) and their dependence on potentially limiting cofactors (e.g. thiamin, tetrahydrofolate, cobalamin; see methods), corroborated by concordant variation in the cofactors' ABC transporters.

This result suggests that the "cost" of obtaining trace metals for use in cofactors could be more expensive than the energetic cost of synthesizing transport machinery and degradative components that would allow for import of exogenous amino acids reducing the need for cofactor. The relationship among an amino acid's environmental co-variation, cofactor dependency, and transporters suggests the idea of "synthesis versus import" as an adaptive strategy in aquatic environments. That is, the import of exogenous amino acids may be more favorable than direct synthesis in environments where the manufacture of the cofactors required for their synthesis is limiting.

4.2.8 Environment-driven variation in methionine (-dependent) pathways

Methionine, a central amino acid in oceanic micro-organisms, presents a particularly interesting example of this phenomenon and further illustrates the importance of a complex network of metabolic adaptations to limiting factors. Reduction in the use of

Level	Id	Description	Footprint		
			CCA	DPM	LM
M	10086	Cysteine degradation/taurine biosynthesis	✓	✓	
M	10087		✓	✓	
M	10064		✓	✓	
M	10063			✓	
M	10022	Asparagine biosynthesis	✓	✓	✓
M	10046	Asparagine degradation		✓	
S	03285	Aspartate-arginosuccinate shunt (including cobalamin-dependent step)	✓		
S	03791		✓		
S	04029		✓		✓
M	10040	GABA shunt (Glutamate)			✓
M	10246	Glutamate degradation/(Heme, proline, siderophore) synthesis		✓	
M	10020		✓		
S	10020		✓		
M	10079	Glutamate synthesis/Histidine degradation	✓		
K	00471	D-Glutamate (D-Glutamine) metabolism	✓		✓
M	10048	Lysine biosynthesis	✓		
S	03459		✓		✓
M	10028	Leucine biosynthesis		✓	✓
M	10024	Methionine biosynthesis (including S-adenosyl-methionine and cobalamin synthesis pathways)	✓		✓
S	03056		✓		✓
S	04163		✓		✓
M	10056		✓		✓
M	10054	Methionine degradation & salvage pathways	✓		✓
M	10055		✓		✓
M	10260	Polyamine biosynthesis (including spermidine putrescine transporters)	✓		✓
M	10261		✓		
S	04453			✓	✓
S	04298	Selenoamino acid biosynthesis	✓		
M	10030	Serine biosynthesis	✓		
M	10057	Threonine biosynthesis	✓		
M	10027	Valine biosynthesis		✓	✓
M	10068	Tryptophan degradation	✓		
S	03812		✓	✓	✓
M	10078	Tyrosine degradation	✓	✓	✓
K	00350	Tyrosine metabolism	✓		✓
M	10169	Other Amino Acid Metabolism	✓	✓	
M	10165		✓	✓	

Table 4.2: Pathways involved in amino acid synthesis, degradation, salvage, and transport with significant environmental co-variation.

methionine in nutrient limited environments has been noted previously (Mazel, 1989). Our results suggest this reduction may stem from cofactor (and perhaps more specifically metal) cost optimization rather than (or in addition to) energetic constraints. We find environmentally-linked variation throughout methionine metabolism including methionine synthesis, salvage, and degradation reinforced at multiple levels of pathway resolution. More specifically, we note that synthesis of both methionine and its cofactor cobalamin (contains cobalt) both decrease as methionine degradation and amino acid transporters (e.g. spermidine and putrescine) increase. Oceanic micro-organisms have been shown to take extreme measures to conserve limited metals (e.g. iron Palenik, 2003); these observations suggest an analogous adaptive response to cobalt limitation.

If such a limitation exists, one would expect to find equally wide-spread changes throughout methionine- and thus cobalamin-dependent pathways - in particular, in those which are dependent on the cofactor S-adenosylmethionine (SAM) such as methylation and secondary metabolites biosynthesis. Indeed, we do find evidence for environmental dependence for a whole suite of methionine processes including cobalamin biosynthesis, as well as variation in many of the SAM-dependent processes (e.g. polyamines, ubiquinone, chlorophyll, and heme) hinting that methionine plays a significant role in shaping downstream environmental adapatations. These observations provide evidence in support of a "synthesis versus import" theory.

4.2.9 Modulating lipid and glycan metabolism as an adaptation to physicochemical conditions.

Lipids and glycans are important components of the microbial outer membrane and thus would be expected to be particularly responsive to environmental conditions. We do find strong environment-linked variation in a plethora of lipid and glycan metabolism-related processes (Table 4.3; Figure 4.4). Indeed, modification of the cell wall is a known adaptive

Level	Id	Description	Footprint		
			CCA	DPM	LM
M	10120	Cell wall synthesis maintenance	✓		✓
S	03937		✓		✓
M	10228	Cholesterol degradation			✓
M	10227				✓
S	03167	Extracellular polysaccharide synthesis		✓	
K	00061	Fatty acid biosynthesis (initiation and elongation)	✓		✓
M	10172		✓		✓
M	10159		✓		✓
K	00071	Fatty acid degradation	✓		✓
M	10177		✓		✓
M	10181	Glycerophospholipid degradation (Triacylglycerol biosynthesis)	✓		✓
K	00561	Glycerophospholipid biosynthesis (diacylglycerol degradation)			✓
M	10202			✓	✓
K	00564				✓
K	00600				✓
M	10204	Glycerophospholipid degradation (CDP-diacylglycerol biosynthesis)	✓		✓
M	10215	Glycerophospholipid metabolism (ether lipid biosynthesis)		✓	✓
M	10197	Isoprenoid biosynthesis (mevalonate and non-mevalonate pathway)			✓
M	10191		✓		
M	10192		✓		✓
M	10155	Keratan sulfate degradation	✓		✓
K	00540	Lipopolysaccharide biosynthesis (including lipid A)	✓		✓
M	10156			✓	
M	10114		✓		
M	10101	N-GlycNac synthesis (including GlycNac transporter)			✓
M	10103		✓		
S	03722		✓		
S	04359			✓	✓
K	00550	Peptidoglycan biosynthesis			✓

Table 4.3: Pathways involved in glycan and lipid metabolism with significant environmental co-variation.

mechanism (e.g. for membrane fluidity) (Morgan-Kiss, 2006; DeLong, 2006), and the variation of pathways involved in extracellular polysaccharide synthesis, lipopolysaccharide synthesis, cell wall maintenance, and glycerophospholipid synthesis (Table 4.3) along the salinity, sample depth, and temperature gradients sampled in the GOS sites could be a reflection of such an adaptation. In addition, significant contributions of lipid metabolism modules in the construction of a linear model for sample depth may illustrate an adaptation strategy to maintain buoyancy for optimal growth conditions (e.g. to optimally profit from light scavenging machinery adaptations for certain wavelengths (Johnson, 2006). Alternatively, it could reflect an adaptation of heterotrophic prokaryotes to the varying composition of phytoplankton-produced dissolved organic matter with depth. Due to the diversity of these pathways' roles without further experimentation one can only speculate on the validity of these particular interpretations. However, undoubtedly, the extreme variation and flexibility of these pathways indicate their central importance in metabolic adaptation to the environment.

4.3 Discussion

As different evolutionary strategies are required to cope with the unique set of challenges specific to each geographic site sampled, our results suggest how environmental pressures shaped these pathway differences. The detailed analysis of three case studies revealed particular pathway adaptations that provide numerous testable hypotheses for linking metabolic versatility to the environment.

Recently, Dinsdale, et al. demonstrated that functional differences can be used to discriminate among nine qualitatively categorized, discrete ecosystems (Dinsdale, 2008). However, as in genome wide association studies where methods using binned data have been supplemented by more sensitive methods that make use of continuous measurements (Sanna, 2008), we have demonstrated the utility of a similar transition in microbial ecology by using comparative metagenomics. Our methods associate microbial community

functions with quantitative, continuous features of the environment allowing for an objective, data-driven framework to classify sites both on the basis of their metabolism and environmental parameters. We show evidence for widespread environmentally-dependent metabolic versatility even in seemingly similar sites (sharing same habitat classification). The methods implemented here also provide a valuable and sensitive assay for simultaneously assessing a number of environmental parameters allowing us to predict both individual and groups of environmental features (Figure 4.2). In reverse, we also predict the usage of a particular metabolic pathway given a set of environmental conditions. Thus, our results suggest that metabolic footprints can be used as the basis of biosensors in situations where no clear measurable environmental factors are available (e.g. monitoring water quality, predicting health state from clinical samples). Indeed, such biosensors would provide more information than the current practice based on species composition (Carignan, 2002), which measure downstream effects (e.g. marker species in pollution) instead of focusing on the molecular processes of the ecosystem as a whole.

Like all current metagenomics datasets, the GOS dataset provides only a snapshot of a site's total genomic content. However, by quantifying the difference in pathway usage along different environmental gradients, one can see the environmental (spatial) dynamics of pathways – analogous to the temporal dynamics in usage of pathways between different cellular states (Luscombe et al., 2004).

Although we have taken precautions to ensure the coverage across sites is the same (see Materials and Methods), the potential remains for important but rare components of metabolic adaptation to be overlooked. Similarly, although we were able to map 74% of proteins to STRING orthologous groups, there is still a fraction of hypothetical proteins that may harbor unknown and thus "unmapped" metabolic components. Indeed, environmental covariation may provide contextual clues to annotate this uncharacterized portion. Novel techniques to functionally characterize this fraction represent a significant challenge and an avenue of active research (Schloss, 2008; Harrington, 2007). Additionally, the five features reported do not fully encapsulate environmental complexity, and the

integration of more environmental measurements will likely reveal many new and exciting discoveries. Despite the data's inherent limitations, they do not compromise the ideas or the conceptual framework presented. Indeed, although the available datasets have constrained us to an analysis of aquatic habitats, the same methodology could readily be applied to investigate the specific metabolic capabilities for any ecosystem in which (physical) environmental parameters are collected including e.g. different anatomical locations, which form "microbial habitats" in humans.

The potential contribution of large viruses as a reservoir for microbial diversity has recently been shown (Monier, 2008; Ghedin, 2005). However, less than 0.3% of proteins in our set can be characterized as viral suggesting a negligible impact on our reported findings (see Materials and Methods). Repeating this analysis on just the viral sample represents an interesting avenue for future research.

4.4 Conclusion

It is clear that microbial communities play a critical role in shaping our world from aiding in global climate regulation (Watson, 1998) and geochemical cycles to degrading hazardous byproducts; however, the complicated, intertwined nature between microbial communities and the environments they inhabit and influence remains poorly understood. We have presented a methodological framework that provides a roadmap to explore these questions in a systematic and statistically rigorous fashion.

4.5 Materials and Methods

4.5.1 GOS data collection and preprocessing

For this study, we filtered the data from the first phase of the GOS expedition to keep only those sites that used a 0.1 to 0.8 μm filter size (with the exception of the Sargasso Sea station 11, which was excluded as it is suspected of contamination; Mahenthiralingam, 2006) thus

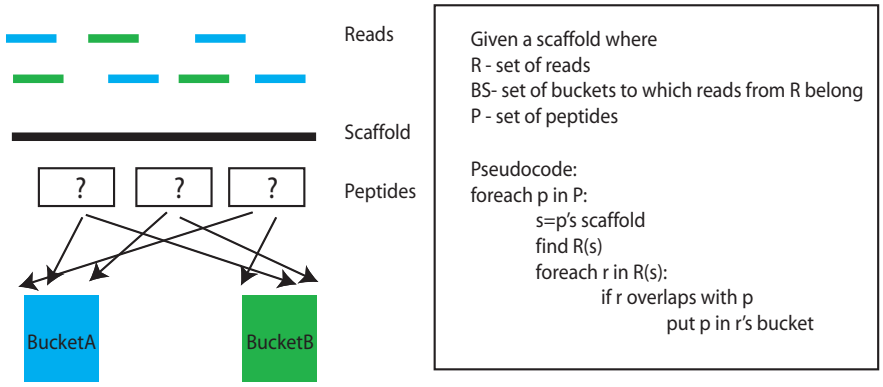


Figure 4.5: Schematic and pseudocode for mapping of peptides to a geographic site(s). Given a set of reads (color coded blue or green depending on which bucket (site) they were recovered from), a set of peptides (boxes), and the coordinates from the scaffold (long black line), the algorithm returns which buckets the peptides (boxes) belong to.

only prokaryotes are part of this analysis. For the remaining 37 sites (Table tabletable-meta-sites), the site metadata was downloaded from the CAMERA database (Seshadri, 2007). For this study, the measurements for temperature, sample depth, water depth, salinity and monthly average chlorophyll level were used. As ten salinity measurements were missing, we averaged the salinity for all non-zero (excluded freshwater site) salinity measurements. In some cases, we were able to corroborate the missing measurements validity through extrapolating from the World Ocean Database (Boyer, 2006). For the protein sequence data, the 6.1 million predicted proteins (Yooseph, 2007) were downloaded from CAMERA.

4.5.2 Mapping peptides to sites

Peptides were mapped to sites based on the read-to-scaffold and orf-to-scaffold mappings available at CAMERA (Seshadri, 2007). Thus, to assign these peptides to a particular site, we used a mapping algorithm that cross-referenced between reads, scaffolds, and peptides based on predicted gene coordinates (Figure 4.5). Therefore, there were instances in which reads that formed part of a single peptide originated from two different sites; as this allowed peptides to be "present" in multiple sites; we term these "multi-site" peptides (see below

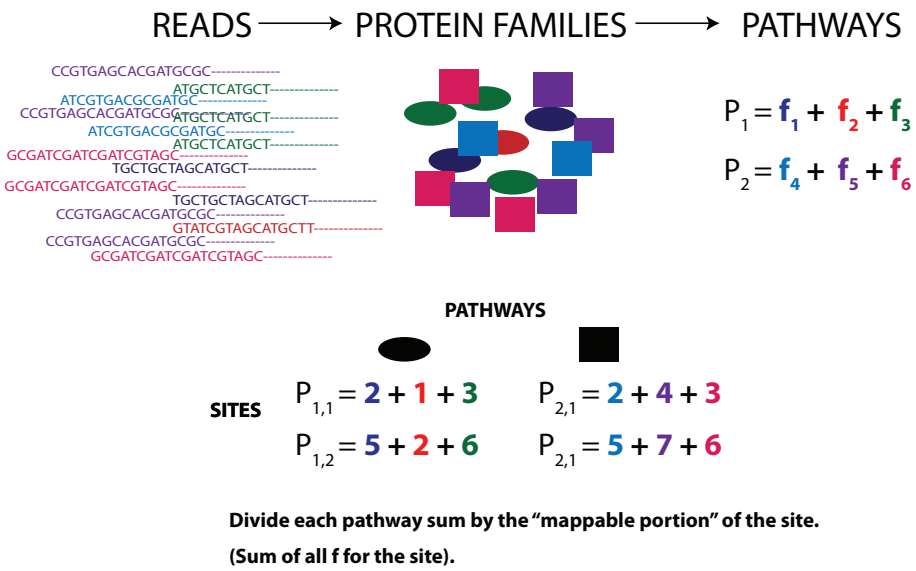


Figure 4.6: Pathway score schematic (see text for additional details)

for additional details).

4.5.3 Mapping cofactors for modules

Cofactors were mapped to each module via EC numbers using the BRENDA database (Barthelmes, 2007). In order to normalize the effects of module size, the fraction of chemical reactions requiring certain cofactors per module is regarded as the cofactor-dependence of module. We then used a goodness of fit test (K-S test) to compare the distribution of CCA structural correlation coefficient between the amino acids that have no cofactors (score=0) and those with cofactors (score>0)($p < .05$).

4.5.4 Assignment and Pathway score

The 111 KEGG maps, 141 modules and 191 operons were assigned as in (Tringe, 2005). For clarity, in the remainder of the text we use the term pathway to refer to all of these levels. Module definitions were downloaded from KEGG (Kanehisa, 2006) and operons were constructed as in (von Mering, 2007). In brief, protein sequences were searched against the extended database of proteins assigned to orthologous groups (OGs) in STRING 7.0 (von

Mering, 2007) using BLASTP (Altschul, 1997), and a pathway was called present when a hit matching one of its proteins occurred (with a BLAST score of at least 60 bits). All results described were also manually scrutinized to reduce artefactual assignments.

The pathway frequency for each site was assigned by summing the total number of instances of that pathway for a particular site and normalizing by total number of assignments for that site to compensate for sample coverage differences. For all analyses, pathways for which the summed count over all sites constituted less than or equal to 0.01% of the total count were removed to avoid artifacts (Figure 4.6).

In addition, we calculated a mismatch rate where we looked to see how many times the top 5 BLAST hits for each peptide mapped to the same pathway. We find that 80% of the time all the top-5 hits will map to the same pathway with a corresponding drop at less stringent bit scores suggesting our results are threshold-independent. A second source of miscalling could be cross hitting of pathways by more "generalist" enzymes. Therefore, we have manually checked the assignments and sought confirmation at multiple levels of resolution (map-module-operon-orthologous group) for all the case stories reported in this manuscript.

4.5.5 Pairwise Correlations and Linear Regression

We computed pairwise Spearman correlations between each pathway frequency vector and each environmental metadata vector for the same sample set – p-values corrected for multiple testing using the Benjamini-Hochberg false discovery rate (Benjamini, 1995) – . Linear models were constructed in two directions. (1) The environmental factor was treated as the response variable and predicted from a subset of pathway frequencies), and (2) the inverse model where pathway frequency was treated as the response variable and predicted from environmental factors. To identify the subset of predictive variables, we used a stepwise regression analysis based on Akaike's information criterion (implementation in R stats package). To avoid overfitting in (1), we used only the top 20 pathways that showed the highest pairwise correlation (as measured by uncorrected p-value) with

the environmental feature modeled. As in many feature selection methods, one is not guaranteed the "best" subset, and we acknowledge that there can be multiple suboptimal solutions. Linear models were considered significant at $p < 0.05$ for both the total model and the estimate of the variable coefficients. For regressions in both directions, the pathway frequencies were standardized to a mean of zero and a standard deviation of one. For (1) we used the centered, quantile-normalized environmental data transformed into percentiles to ensure a truly normal distribution and thus accurate p-values.

4.5.6 Discriminative Partition Matching (DPM)

To analyze whether groupings of sites based on similar environmental features also shared functional similarities, we clustered the sites based on their quantitative environmental metadata resulting in two distinct clusters or site-sets. Next, we partitioned the sites in the metabolism matrices (see Figure 4.1A) into the same two site-sets and calculated the mean normalized frequency for each pathway in each site-set (see below for generalized approach). If the means of the pathway frequency between the two site-sets were not significantly different, this would suggest that the environment-based partitioning does not reflect functional differences. If the distributions do differ significantly, it would imply that the environmental features are related to the specific aspect of metabolism. Further, we computed the two-sample t-test for each individual map, module, operon, and COG. Those pathways that were significantly different (Benajamini-Hochberg corrected $p < 0.05$) were combined to form the DPM footprint.

4.5.7 Canonical Correlation Analysis (CCA)

The goal of canonical correlation analysis is to identify the set of projections that maximally correlate two sets of variables (Wichern, 2003) (for a more detailed description of the relations of CCA to other common techniques including principal components analysis and least squares regression; see Borga, 1998).

Due to the large number of dimensions and small number of data points the

solution can be unstable, thus we applied a variant of CCA, regularized CCA (Eaton, 1973) implementation in (Gonzalez, 2008). We estimated regularization parameters λ_1 and λ_2 (penalty to covariance matrices) via a leave-one out cross-validation procedure (implementation in Gonzalez, 2008). Because of interdependencies between metabolic pathways, canonical weights must be interpreted with caution. For this reason, we also calculated the structural correlation coefficient, which is the correlation between the original variable and the canonical variate. This allows one to specifically answer the question how important is this one variable (metabolic pathway) relative to all the other variables (metabolic pathways) (see below for additional evaluation metrics). Those pathways, which had a structural correlation coefficient greater than 0.3, formed the CCA footprint. In addition, we investigated the effect of changing this threshold. Principal components analysis and the resultant biplot on the environmental features show these features to be basically orthogonal (Figure 4.8).

4.6 Additional Evaluation Metrics and Controls

4.6.1 Construction and Results from Control Matrices

To control for relative differences in metabolic pathways among the geographic locations simply reflecting sampling bias, we constructed two control matrices composed of proteins that would not be expected to change among sites, such as those involved in basal transcription or translational machinery. The first is composed of those COG categorizes as information processing, and the second those involved in cellular processes. We used Student's t-test and found that although the distributions of the means for the control matrices (composed of those COGs annotated as belonging to either information or cellular processing) are not significantly different between the two environmental site-sets ($p < .07$ and $p < .08$, respectively), there are significant differences in metabolism ($p < 9x10^{-3}$ and $p < 4x10^{-14}$, COG and KEGG annotated metabolism definitions). However, we do see the same asymmetry as originally noted in the GOS paper for DNA polymerase,

topoisomerase, and gyrase (Yooseph, 2007) by aggregating across the basal machinery this effect is minimized. Thus, DPM's greatest strength is as a means of evaluating the functional significance of a particular partitioning and in controlling for potential sampling bias through the testing of control matrices (expected to be environmentally invariant) alongside matrices that are suspected of being environmentally variant.

4.6.2 More detailed CCA Evaluation Metrics

As in PCA, there are a number of metrics that can be used to determine the number of dimensions, in this case canonical variates, that should be included in the analysis (Borga, 1998). The overall canonical correlations for both dimension 1 and dimension 2 are high for KEGG maps, module, and operons; however, there is a significant drop in average redundancy between dimension 1 and dimension 2 and further dimension 2 and dimension 3 making it appropriate to use only these two dimensions in the overall analysis. We can also measure the amount of information the environmental variate is able to "cover" from metabolism and vice versa by calculating the average variance of the dimensions and the redundancy (Wichern, 2003). These measurements are both high for the environment but lower for the metabolic pathways. This suggests that there are many weaker signals coming from the metabolic matrices as opposed to a few strong ones.

4.6.3 Generalization of DPM

We provided a specific use of DPM in the text; however, DPM can be generalized. There are three basic steps to DPM. (1) The sites from the first matrix are partitioned to create site-sets. (2) The second matrix is partitioned in accordance with these site-sets. (3) A t-test (or *ANOVA* for more than two site-sets) is performed to test whether the *site-sets* are statistically different in the attributes of the second matrix.

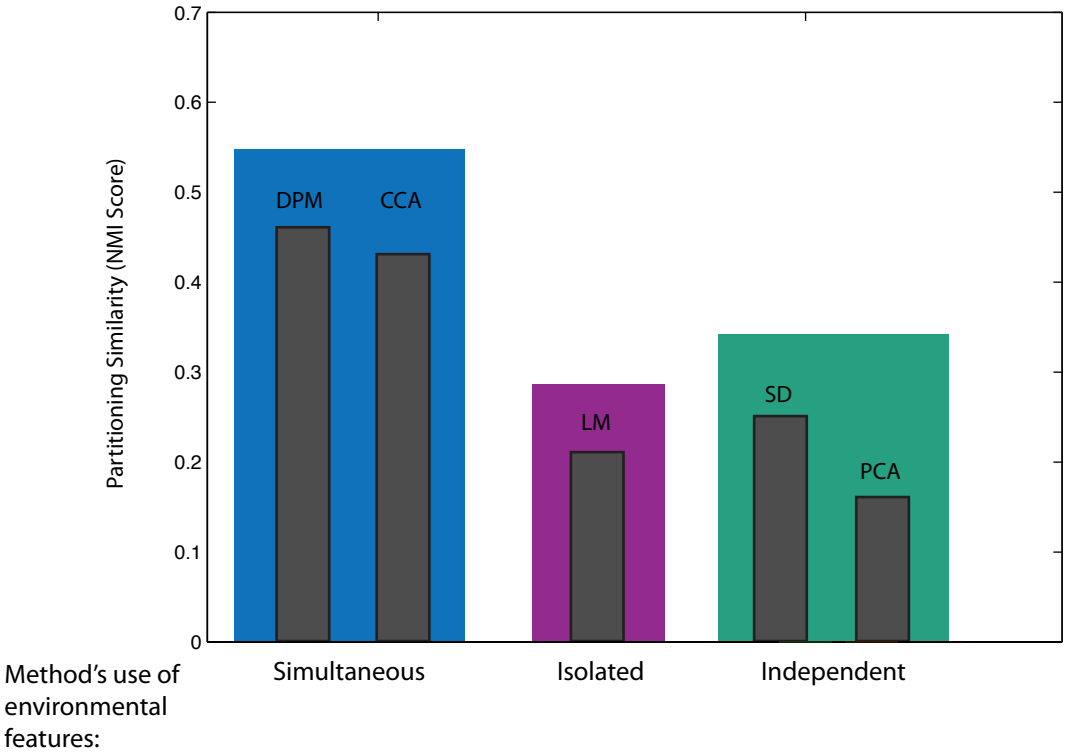


Figure 4.7: Comparison of different classes of methods. We evaluated the efficacy of three different classes of methods based on their explicit use of the quantitative environmental data, which we term independent, isolated, and simultaneous. Independent methods include no environmental description (green), isolated only one environmental feature at a time (purple), and simultaneous methods incorporate all environmental features simultaneously (blue). For clarity, we refer to the highly-weighted set of pathways generated for each method as a footprint. Each of the five methods was used to generate a metabolic footprint, and each bar represents the normalized mutual information (NMI) score for that method's footprint. No statistically significant difference was observed between scores within each particular category ($p > .05$).

4.7 Comparison with Variance-Maximization Approaches

4.7.1 Compare/Contrast with other Methods

An entirely different approach to the one presented in the text assumes that the inherent variability of the environments could be directly observed by examining the global variance in the metabolic dataset. That is, one identifies the pathways with greatest variance without directly measuring whether they co-vary specifically with the environment. First, we performed a simple standard deviation (SD) calculation to find pathways that changed the most. We also used a principal component analysis (PCA) to identify the pathways that encapsulate the greatest proportion of variance. We then assessed the performance of these methods to identify metabolic adaptation to environmental parameters based on their ability to recapture the environmental-based partitioning using only the metabolic pathways identified as significant for each method by measuring cluster similarity (see below). Simply identifying the metabolic pathways with the greatest variance did not always reflect changes in the environmental parameters (Figure 4.7). Indeed, both methods that simultaneously incorporate environmental and metabolic data significantly outperform the variance-based, independent methods, and perhaps, unsurprisingly, the linear models, which are more appropriate for investigating single relationships than looking at global context. These results were consistent despite varying the number of pathways using a variety of different thresholds for all methods SD, PCA, LM, DPM, and CCA.

4.7.2 Clustering

The environmental data matrix was first standardized to mean of zero and standard deviation of one. We evaluated distances using 1-correlation and used average linkage hierarchical clustering. The clustering procedure was repeated using spectral k-means without significance differences (data not shown).

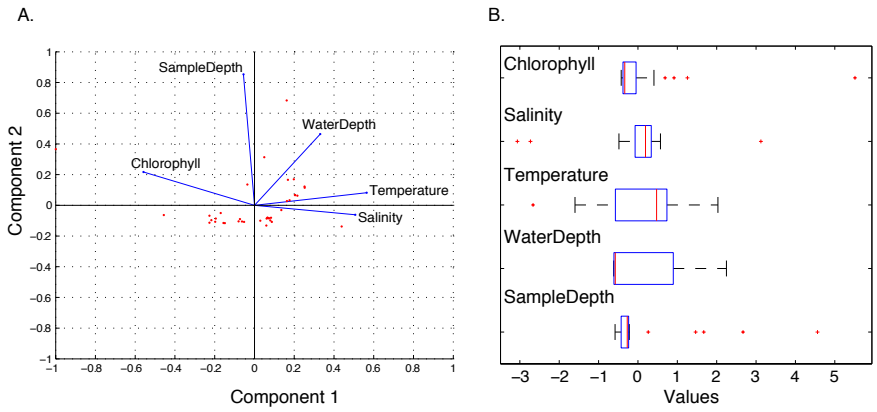


Figure 4.8: Biplot and boxplot of standardized environmental variables. To examine possible dependencies between the variables we performed principal component analysis. We next plotted component 1 and 2 (A). One can see that the variables with the exception of temperature and salinity are basically orthogonal to one another. (B) Boxplot of standardized variables.

4.7.3 Metrics to Compare Cluster Similarity

Cluster similarity was measured by computing both a normalized mutual information score, which measures the amount of information lost if one applies the classification "clustering" from the first partition to the second (Forbes, 1995), and the rand index (Hotelling, 1936), which computes the number of "correct" pairwise interactions between the two sets of clusters. The closer the normalized mutual information (NMI) score is to 1.0 the better the metabolic footprint generated from the method performed in recapitulating the structure of the environmental data. For each set, a significance value for the rand index was computed by randomly shuffling the clustering assignment, recalculating the index after each iteration, and counting the number of times the index computed from the random data exceeded the index computed from the real data after 10,000 iterations.

Hydrostation S, Sargasso Sea	Browns Bank, Gulf of Maine
Outside Halifax, Nova Scotia	Bedford Basin, Nova Scotia
Bay of Fundy, Nova Scotia	Northern Gulf of Maine
Newport Harbor, RI	Block Island, NY
Cape May, NJ	Delaware Bay, NJ
Chesapeake Bay, MD	Off Nags Head, NC
South of Charleston, SC	Off Key West, FL
Gulf of Mexico	Yucatan Channel
Rosario Bank	Northeast of Colon
Lake Gatun	Gulf of Panama
250 miles from Panama City	30 miles from Cocos Island
134 miles NE of Galapagos	Devil's Crown, Floreana Island
Coastal Floreana	North James Bay, Santigo Island
Warm seep, Roca Redonda	Upwelling, Fernandina Island
Mangrove on Isabella Island	Punta Cormorant, Hypersaline Lagoon, Floreana Island
North Seamore Island	Wolf Island
Cabo Marshall, Isabella Island	Equatorial Pacific TAO Buoy
201 miles from F. Polynesia	Rangirora Atoll

Table 4.4: Size selected GOS sites

4.8 Future Challenges and Current Limitations

4.8.1 Need for rigorous, quantitative descriptions of the environment

Taxonomic classification underwent a paradigm shift with the introduction of molecular features to a system that had been based purely on morphological characteristics. Many species were discovered to have been mistakenly lumped together under the scrutiny of comparative sequence analysis. Similarly, the terms currently used to describe marine habitats need to be grounded in rigorous quantitative measurements of oceanic features. While this idea is certainly not new to oceanographers, in order to truly exploit the potential of metagenomics, this same rigor must be brought to bear. There is a great human influence on marine microbial communities. To better understand this human-microbial interaction, we must have more comprehensive ways of characterizing these environments both in space, as well as, through time. One of the major limitations of the current study is that without measurements of nutrient gradients and fluxes, one is forced to rely on indirect and thus somewhat artificial measurements of these variables. Oceanographic institutes such as

NOAA have gridded the entire earth for the collection of oceanographic measurements including nitrate, phosphate, oxygen levels, etc which reach back over two decades and are often accompanied by exhaustive censuses of biomass subdivided into bacterioplankton, phytoplankton, zooplankton, and even larger marine creatures. While oceanographic tools can measure explicit variables in the ocean (nitrate, phosphate, etc), it is the microbes and particularly their amazingly versatile metabolism that are reshaping them. It would make sense then to try to couple some of the oceanographic measurements with a metagenomics initiative. Studies such as the Hawaiian Ocean Time series (DeLong, 2000) and the planned Monterrey Bay Time Series are employing such a strategy by exhaustively collecting oceanographic data along with performing metagenomics profiling. However with a large number of oceanographic measurement projects beginning particularly in the Arctic over the next couple of years; it would be wise to consider these sites or at the least the availability of such data in deciding on metagenomics project locations. The increase in such data collections will need to be accompanied with the development of more analytical tools for connecting the genic content with a particular environment.

4.8.2 Computational challenges

Working with this type of data is necessarily compute-intensive. This study consumed 0.5 cpu-years and was only plausible because of access to a high performance compute cluster. However, in order for such data to be used it needs to be presented and stored in an accessible manner. It is not current practice to release raw data such as blast scores for each read, etc resulting in widespread and unnecessary duplication of this type of work which tends to be some of the most compute intensive steps in these types of analysis. To this end, all of the raw data including the original blast hits and many of the utility scripts are included in the supplementary material and the website http://networks.gersteinlab.org/metagenomics.

Database construction and computational resources

The scale and complexity of the data presented several obstacles. We built a mysql database to facilitate the storage and retrieval of analysis results. The mysql server used ran on a dedicated Dell fileserver with four 3.2 Ghz cpus and 8 GB RAM. The database contained approximately 328 Million records and totaled 5 GB on disk. In order to facilitate high performance computing on the data, the read sequences themselves were not stored in the tables; rather, metadata and a pointer to a position within a fasta file was stored. We found this to be a very useful division, keeping the mysql database to a manageable size. The blasting between the various databases was performed on two compute clusters (bulldogc and bulldogi) at the Yale Life Sciences High Performance Computing Center. Bulldogc consists of 130 dual cpu Dell PowerEdge1855 nodes, totalling 260 3.2 Ghz cores. Bulldogi consists of 170 dual/dual Dell PowerEdge 1955 nodes, totalling 680 3.0 Ghz cores. Lustre, a high performance, parallel filesystem, was used to manage the I/O traffic. Each blast job was run in parallel, typically on 10-100 cpus, depending on availability. The computations were parallelized by splitting the input sequence set into many disjoint pieces, blasting each piece against the database, and combining the results. We have found that this technique scales very well, so long as there are enough input sequences and the I/O system sustain the required rates. Approximately 4500 cpu-hours (more than 0.5 cpu-year) were consumed during the blasting.

Chapter 5

Network Evolution: Divergence of gene regulation in close yeast

5.1 Introduction

Differences in related individuals are generally attributed to changes in gene composition and/or alterations in their regulation. Previous efforts to examine divergence of regulatory information have relied on the analysis of conserved sequences in putative promoter regions (Cliften, 2003; Kellis, 2003). However, these approaches are limited as transcription factor (TF) binding sites are often short and degenerate making their computational detection difficult (Tompa, 2005), while requiring the conservation of motifs across species precludes detection of sequences which are evolutionarily divergent.

In an effort to measure the divergence in transcriptional networks across related species, we used chromatin immunoprecipitation followed by DNA microarray analysis (chIP chip) (Iyer, 2001; Ren, 2000) to directly monitor the binding site distribution of orthologous transcription factors in four yeast species. The targets of the transcription factors Ste12 and Tec1 were mapped in diploid strains of *S. cerevisae, S. mikatae, S. bayanus* and *C. albicans* (Ste12 only). *S. cerevisiae, S. mikatae* and *S. bayanus*, which are all members of the *Saccharomyces sensu stricto* group, are estimated to have diverged between 5 and 20

millions years ago; *C. albicans*, an important human pathogen, is thought to have diverged over 200 million years ago (Kellis, 2003 ; Wolfe, 2006).

In diploid strains of *S. cerevisiae*, Ste12 and Tec1 act cooperatively to regulate genes during the formation of pseudohyphae (branching chains of elongated cells formed during growth in low nitrogen environments (Madhani, 1997; Gavrias, 1996; Liu, 1993); whereas, in haploid cells, Ste12 regulates mating genes through association with a different factor (Fields, 1985). Highly conserved orthologs of Ste12 and Tec1 are present in *S. mikatae* and *S. bayanus* (80-88% identity); these are more divergent in *C. albicans* (50% and 44% identity, respectively) where they are required for regulating dimorphic growth (related to pseudohyphal growth) and pathogenicity in this species (Stoldt, 1997; Liu, 1994; Leberer, 1996; Lane, 2001). Our mapping of Ste12 and Tec1 across the genome of these species has revealed remarkable diversity in transcription factor binding site locations, even among closely related organisms.

5.2 Results

5.2.1 Identification of Ste12 and Tec1 Binding Sites in *S. cerevisiae*, *S. mikatae* and *S. bayanus*

The detection of binding sites using chromatin immunoprecipitation and microarray analysis (chIP chip) (Iyer, 2001; Ren, 2000) offers the ability to globally map transcription factor (TF) binding locations experimentally, rather than computationally. For species such as yeast, where genome sequences of numerous related species are available (Piskur, 2004), this approach can allow for the evolutionary comparison of binding sites of conserved TFs across species.

We have used this approach to investigate evolutionary divergence in the targets of two developmental regulators in the *Saccharomyces sensu stricto* yeasts *S. cerevisae, S. mikatae, S. bayanus.* In *S. cerevisiae* diploids, Ste12 and Tec1 act cooperatively to regulate genes during pseudohyphal development (Madhani, 1997; Gavrias, 1996;

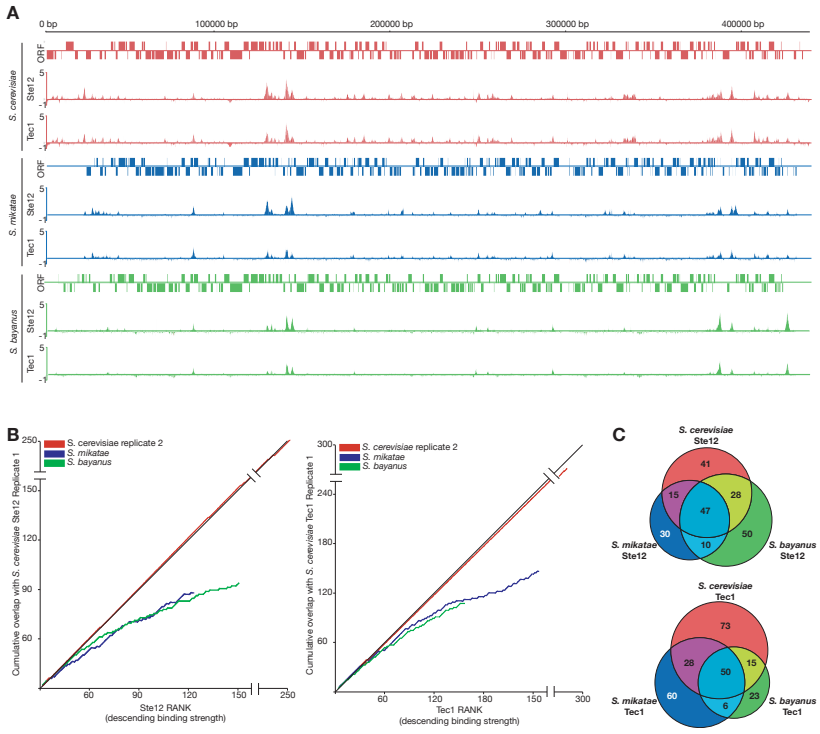


Figure 5.1: Ste12 and Tec1 binding overlap. Ste12 and Tec1 bind to discrete regions of chromosome IX of *S. cerevisiae* and to orthologous regions of *S. mikatae* and *S. bayanus*. chIP chip enrichment by Tec1 and Ste12 (log2 ratios) are shown relative to ORFs of *S. cerevisiae* (red), *S. mikatae* (blue) and *S. bayanus* (green) (B) Rank order analysis of Ste12 and Tec1 chIP chip in *S. cerevisiae* (red) *S. mikatae* (blue) and *S. bayanus* (green) (C) Gene target overlap across the Saccharomyces species.example caption

Liu, 1993), whereas in haploid cells, Ste12 regulates mating genes (Fields, 1985). The binding sites of Ste12 and Tec1 were mapped in all three species under low-nitrogen (pseudohyphal) conditions using triplicate chIP chip experiments and species-specific high-density oligonucleotide tiling microarrays (Figure 5.5) (Borneman, Submitted).

5.2.2 Extensive Divergence of Binding Sites in the *Saccharomyces sensu stricto* Species

Ste12 bound to 380, 167, and 250 discrete sites in S. cerevisiae, *S. mikatae* and *S. bayanus*, respectively, whereas Tec1 bound 348, 185 and 126. For each species the two factors bound to a high proportion of common regions (86%, 80% and 87% for *S. cerevisiae*, *S. mikatae* and *S. bayanus*, respectively) suggesting that the cooperative interaction observed between Ste12 and Tec1 in *S. cerevisiae* is conserved across the three *Saccharomyces* species.

Analysis of the signal tracks allowed for global comparisons in TF binding to be made between the species revealing qualitative and quantitative differences in ChIP binding regions (Figure 5.1A). To systematically perform inter-species comparisons, regions that were not represented across all three yeast genomes were removed. Comparison of the overlap in binding across species as a function of rank order revealed significant binding differences throughout the rank order indicating that even strong targets from one species may not be bound in the others (Figure 5.1B). As a control, replicate experiments from *S. cerevisiae* displayed over 98% concordance in binding.

5.2.3 Three classes of TF binding events

Overall, three classes of TF binding events were observed: those conserved across all three species, those present in two of three species and species-specific binding events (Figure 5.3). Of the 221 and 255 targets bound in total by Ste12 and Tec1 respectively, only 47 (Ste12, 21%) and 50 (Tec1, 20%) were conserved across all three species (Figure 5.1C, reffig:regnet-profilesA). The conserved binding events were present throughout the rank order indicating that both highly occupied and less occupied regions are conserved. To ensure that these

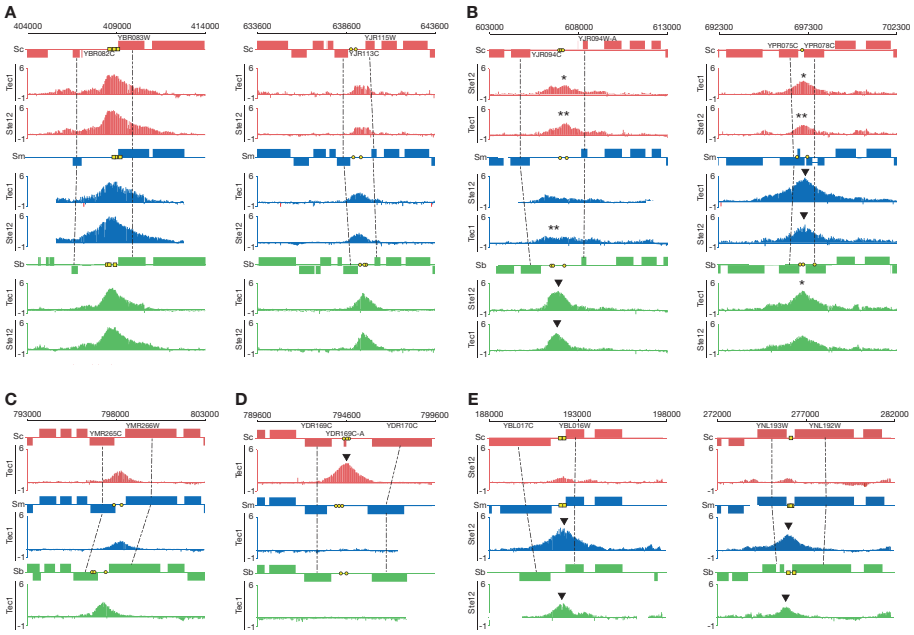
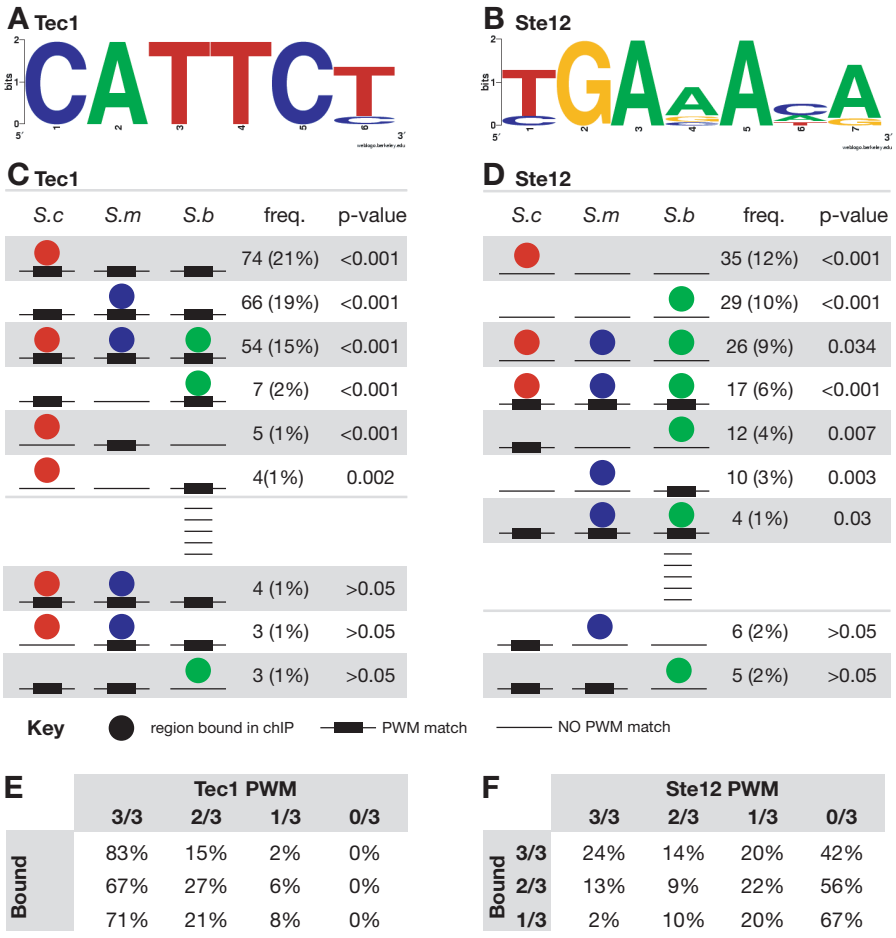


Figure 5.2: Ste12 and Tec1 binding patterns. Comparison of binding by Ste12 and Tec1 across *S. cerevisiae* (red), *S. mikatae* (blue) and *S. bayanus* (green). (A) Conserved binding. (B) Conserved binding with quantitative signal differences (C) Conserved binding with loss of consensus sequences in one species. (D) Species-specific binding despite conserved consensus sequences. (E) Binding only in *S. mikatae* and *S. bayanus*. Significant binding peaks are indicated by arrows.

binding differences were not due to the scoring threshold used, signal distributions for non-bound orthologs of target regions were calculated. Of the unbound orthologous regions, 80% had signals similar to background, indicating that the vast majority will be unaffected by threshold changes (Figure reffig:fig-regnet-threshold). Even when identical binding regions were utilized, 23% differed in their intensity by at least 1.5 fold between species (0% between *S. cerevisiae* replicates), suggesting quantitative differences exist in site occupation or binding strength between species (Figure 5.2B). Thus, the majority of target genes were bound in only one or two of the three species indicating considerable divergence in binding sites across these yeasts (Figure 5.2C). As the fraction of non-conserved genes between *S. cerevisiae*, *S. mikatae* and *S. bayanus* is less than 0.05% (Kellis, 2003), the amount of variation in TF binding is substantially larger than that of gene variation.

5.2.4 Comparison of Binding Sites with Conserved Sequences Reveals Significant Differences.

One possible cause for the inter-species differences in the chIP binding locations is divergence in binding site sequences, whereby the loss of a regulatory motif results in the concomitant loss of transcription factor binding. To examine this possibility, sequence motifs in both bound and orthologous unbound regions were investigated across the three *Saccharomyces* species. Position weight matrices (PWM) representing the putative binding motifs for Ste12 and Tec1 were generated from the chIP chip data (Liu, 2002). Analysis of the Tec1 targets of the three species revealed an over-represented sequence motif which matched the known Tec1 consensus (Madhani, 1997) (Figure 5.3A), while the targets of Ste12 in *S. cerevisiae* and *S. mikatae* revealed a motif that was similar to the known binding sequence (Dolan, 1989) (Figure 5.3B). This sequence was not over-represented in *S. bayanus*. Using the PWM sequences, chIP bound regions and orthologous unbound regions from each species were then scored for the presence of each motif (Bailey, 1998). There were several significant classes of TF binding events, with those genes bound by all three factors present near the top of both the Tec1 (all bound, motif in all) and Ste12



E

Bound	Tec1 PWM 3/3	2/3	1/3	0/3
	83%	15%	2%	0%
	67%	27%	6%	0%
	71%	21%	8%	0%

F

Bound	Ste12 PWM 3/3	2/3	1/3	0/3
3/3	24%	14%	20%	42%
2/3	13%	9%	22%	56%
1/3	2%	10%	20%	67%

Figure 5.3: Motif analysis of chIP binding targets. Logo representations of the PWM for (A) Ste12 and (B) Tec1 (12). (C, D) Significant classes of binding targets following classification by both the conservation of chIP binding and the presence or absence of consensus motifs. (E, F) Compiled proportions of binding targets and PWM matches for Ste12 and Tec1.

(all bound, with and without motif) lists (Figure 5.3C, D). For promoter regions that displayed evolutionarily-conserved chIP binding in all three Saccharomyces species, 83% (Tec1) and 24% (Ste12), contained at least one significant occurrence of the PWM motif for that factor in each species (Figure 5.3E, F). In contrast, 2% and 62% of the promoters that displayed conserved chIP binding did not contain a match to the PWM in at least two of the three species. Thus, the Ste12 motif is not present in a high proportion of pseudohyphal-responsive genes implying Tec1 may target Ste12 to these regulatory regions (Chou, 2006).

In contrast to the previous results in which experimentally determined binding correlated with the presence of predicted motifs, there were many examples where a species-specific loss of binding and/or a loss of sequence have occurred. There were 48 (Tec1, 14% of total binding events) and 35 (Ste12, 10%) experimentally-bound regions, which contained PWM match where the orthologous region in at least one other species was neither bound nor contained a motif match. For these loci, loss of chIP binding is concordant with loss of the motif for this factor, providing clear examples of regions where network evolution occurred through the gain or loss of regulatory sequences.

Furthermore, there were 45 (Tec1, 12%) and 9 (Ste12, 3%) instances where a PWM match occurred in all three species but where that region was experimentally-bound in only two (Figure 5.2D). Either these loci are occupied at other times in the life cycle or they are not functional. Conversely, in 11 (Tec1, 3%) and 22 (Ste12, 6%) instances, genomic regions displayed conserved chIP binding but at least one species was missing a PWM motif match (Figure 5.2E. Thus, sequence conservation does not predict binding.

To further examine the role of conserved versus non-conserved chIP binding events and motifs, these results were compared with expression microarray studies of pseudohyphal formation in *S. cerevisiae* (Prinz, 2004). Of the chIP binding targets which had significantly altered expression ($\tilde{2}0\%$ of the chIP targets), there was no significant enrichment for genes with conserved binding (11% bound versus 14% unbound) or PWM matches (12% with motif versus 16% without). Thus, in this case, sequence-based motif analyses in the absence

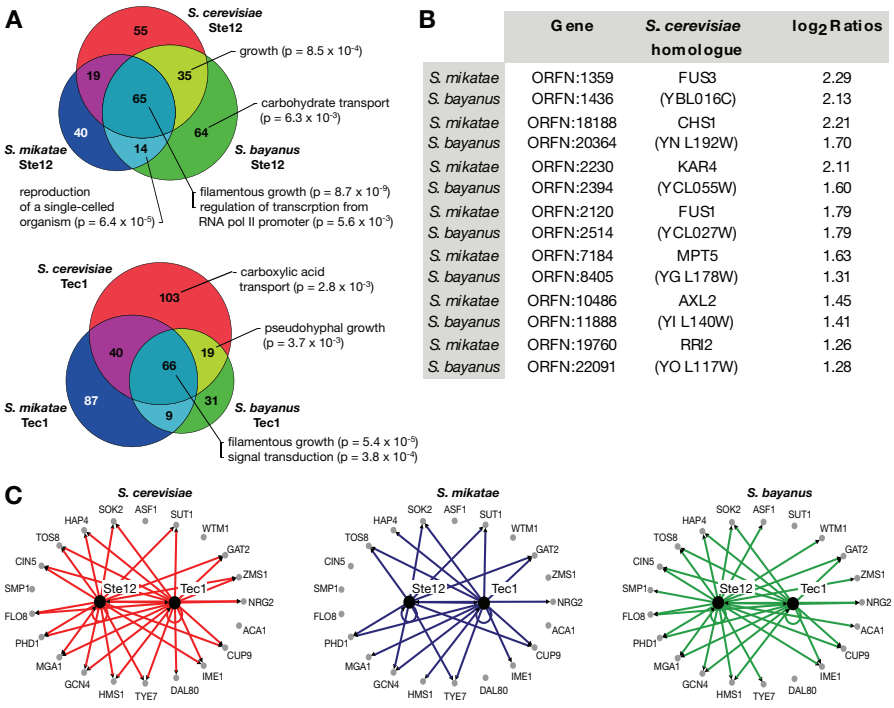
of experimentally-determined binding data were not sufficient for the accurate prediction of TF binding profiles and gene function.

5.2.5 Conserved Classes of Targets and Regulatory Networks Across Related Yeasts

To elucidate the biological significance of both the conserved and species-specific gene targets, each bound region was mapped to its downstream target genes by identifying ORFs which were 3' of and directly flanking each chIP binding event. Conserved Ste12 and Tec1 gene targets displayed enrichment for two GO (Boyle, 2004) categories, "filamentous growth" and "regulation of transcription from RNA polymerase II promoters" (Figure 5.4A). As the majority of the genes from within the second category encode TFs, the predicted downstream TF networks of *S. bayanus* and *S. mikatae* were compared to those of *S. cerevisiae* (Borneman, 2006) to determine which connections had been maintained during the evolution of the *Saccharomyces sensu stricto* group (Figure 5.4C). The binding of Ste12 and Tec1 to downstream TFs was shown to be highly conserved (73% across the three species). The network of *S. mikatae* was most diverged and had several key regulatory omissions including Flo8 (not bound by either Ste12 or Tec1) and Mga1 (Ste12). Thus, although important differences can be found, TF binding to the promoters of other TFs was highly conserved between species relative to the level of conservation observed for other genes.

5.2.6 Genes Important for *S. cerevisiae* Mating are Bound Under Pseudohyphal Conditions in *S. mikatae* and *S. bayanus*

From those groups of genes which did not display conserved binding across the three species, one interesting class was bound by Ste12 specifically in *S. mikatae* and *S. bayanus* and was enriched in genes involved in mating (GO: reproduction in single-celled organisms) in *S. cerevisiae* (Figure 5.4A, B). Unlike the diploid cells used in this study, these genes are targets of Ste12 in haploid *S. cerevisiae* cells (Harbison, 2004; Zeitlinger, 2003) and this



B

	Gene	*S. cerevisiae* homologue	log_2 Ratios
S. mikatae	ORFN:1359	FUS3	2.29
S. bayanus	ORFN:1436	(YBL016C)	2.13
S. mikatae	ORFN:18188	CHS1	2.21
S. bayanus	ORFN:20364	(YN L192W)	1.70
S. mikatae	ORFN:2230	KAR4	2.11
S. bayanus	ORFN:2394	(YCL055W)	1.60
S. mikatae	ORFN:2120	FUS1	1.79
S. bayanus	ORFN:2514	(YCL027W)	1.79
S. mikatae	ORFN:7184	MPT5	1.63
S. bayanus	ORFN:8405	(YG L178W)	1.31
S. mikatae	ORFN:10486	AXL2	1.45
S. bayanus	ORFN:11888	(YI L140W)	1.41
S. mikatae	ORFN:19760	RRI2	1.26
S. bayanus	ORFN:22091	(YO L117W)	1.28

Figure 5.4: (A) Ste12 and Tec1 bind to common and distinct genes across the *Saccharomyces sensu stricto* lineage. Over-represented GO terms are listed for each combinatorial category. (B) Mating genes bound specifically by Ste12 in *S. mikatae* and *S. bayanus*. (C) TF network conservation in *S. cerevisiae* (red), *S. mikatae* (blue) and *S. bayanus* (green).

differential binding occurs despite the presence of conserved Ste12 binding motifs. Thus, Ste12 binding targets may be occupied under different conditions across related species. In *S. cerevisiae*, Ste12 binds to these sites only during mating, while in *S. mikatae* and *S. bayanus* Ste12 binds to these same regions in diploid cells.

5.2.7 The Ste12 homolog of *C. albicans* also binds upstream of mating genes

To extend this study outside of *Saccharomyces* yeasts, the binding of the *C. albicans* Ste12 ortholog, Cph1, was also mapped (Jones, 2004). Cph1 functions in the dimorphic switch of this yeast, a process which shares many genetic components with pseudohyphal growth (Sanchez-Martinez, 2001). A total of 52 significant Cph1 chIP binding events were detected under dimorphic growth conditions, with many residing upstream of known

pathogenicity determinants (Birse, 1993; Braun, 2000; Braun, 2000; Lane, 2001; White, 1993). From these gene targets, 33 have recognizable orthologs in *S. cerevisiae* and of these, 10, 10 and 13 displayed conserved binding with *S. cerevisiae*, *S. mikatae* and *S. bayanus*, respectively. While the majority of gene targets of Cph1 in *C. albicans* are not conserved with the *Saccharomyces* species, the *C. albicans* orthologs bound by Ste12, like those from *S. mikatae* and *S. bayanus* included a significant number of genes that function during reproduction and mating in *S. cerevisiae* ($P = 4x10^{-3}$) (Boyle, 2004) including CEK1 (FUS3), FUS1 and SST2. Thus, in *C. albicans*, like *S. mikatae* and *S. bayanus*, the Ste12 ortholog also binds to genes required for mating in *S. cerevisiae* under filamentous growth conditions raising the possibility that these genes have become more specialized in *S. cerevisiae*.

5.3 Discussion

We find that extensive regulatory changes can exist in closely related species, which is consistent with a recent study which showed that distinct regulatory circuits can produce similar regulatory outcomes in *S. cerevisiae* and *C. albicans* (Tsong, 2006). Furthermore, while *S. cerevisiae* and *S. mikatae* are quite similar to one another at the nucleotide sequence level, they are equally different to each other and *S. bayanus* in their TF profiles. We expect that the extensive binding site differences observed in this study reflect the rapid specialization of these organisms for their distinct ecological environments and that differences in transcription regulation between related species may be responsible for rapid evolutionary adaptation to varied ecological niches.

5.4 Materials and Methods

5.4.1 Yeast strains, growth conditions and epitope tagging

Yeast strains used in this study were the S. bayanus NRRL Y-11845, *S. mikatae* IFO 1815 and *C. albicans* BWP17 (Wilson, 1999). S. bayanus, *S. mikatae* were both transformed by a PCR-based approach used for *S. cerevisiae*. As each strain is diploid, sequential transformations were performed using G418 resistance (kanMX) (Wach, 1994) as a marker, with marker conversion to nourseothricin (nat) resistance (natMX) (Goldstein, 1999) performed to allow reuse of the kanMX marker prior to tagging of the second allele. For *S. mikatae*, G418 and nat were used at 200 μg/ml and 100 μg/ml respectively for selection of transformants, whereas for S. bayanus, 50 μg/ml and 20 μg/ml were used. *C. albicans* was also transformed using a PCR based approach, with sequential rounds of tagging performed using modified versions of pFA6a-13myc-kanMX6 (Longtine, 1998) where kanMX was replaced by *URA3* and *ARG4* as selectable markers. For each strain, protein expression was examined by immunoblot analysis and tagged proteins were produced of the expected size. The tagged strains appeared functional as cell elongation in the tagged strains appeared similar to that of wild type strains under conditions that induce pseudohyphal or dimorphic growth.

5.4.2 Array design

Arrays were designed to the available genome sequences of *S. mikatae*, S. bayanus (Kellis, 2003) by selecting 50 bp oligonucleotides every 60 bp on both strands of each sequencing contig, with top and bottom strand oligonucleotides offset by 30bp (see Figure 5.5). For *C. albicans*, 50 bp oligonucleotides were also designed every 60 bp across the published genome sequence (Jones, 2004), although due to microarray feature constraints, tiling of the bottom strand was limited to one 50 bp oligonucleotide every 120 bp.

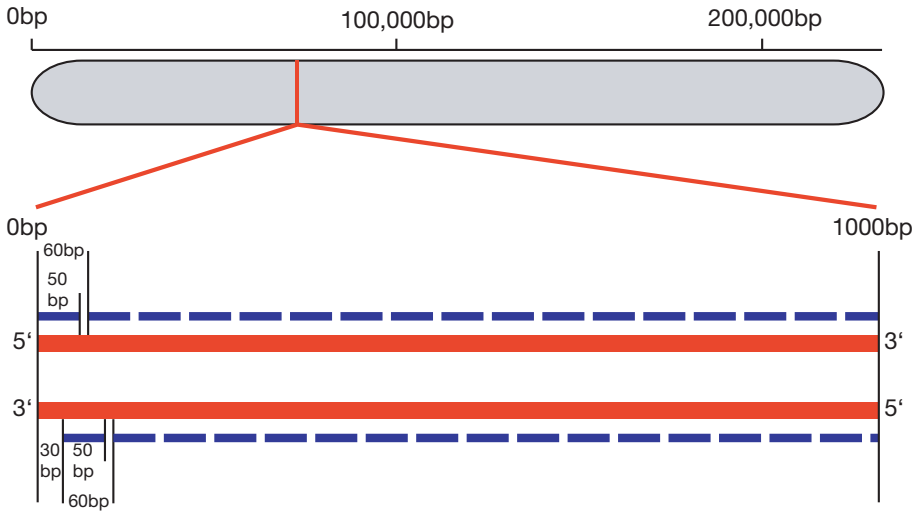


Figure 5.5: Tiling array designs used in S. mikatae, S. bayanus and C. albicans. For each species, oligos were deisgned to tile the genome sequence contigs with a 50bp oligonucleotide every 60 bp on the Watson strand. For S. mikatae and S. bayanus, this same spacing was used on the Crick strand offset 30 bp compared to those oligonucleotide on the Watson strand. For C .albicans, limitations on the number a features allowed on the arrays resulted in a 50 bp oligonucleotide being spaced every 120 bp, offset 30 bp.

5.4.3 Immunoprecipitations, DNA labelling and hybridisation

For immunoprecipitations, S. bayanus and *S. mikatae* were grown using conditions similar to those used for *S. cerevisiae* (Borneman, 2006), except with the time in nitrogen starvation medium altered to reflect the differences in doubling time of *S. mikatae* (3 hrs induction) and S. bayanus (6hrs induction). For *C. albicans*, cells were grown at $25°C$ to an OD_{600} of 0.3 in Lee's medium (Lee, 1975) prior to being induced for 4 hrs at $37°C$ Lee's medium. Cells were fixed with formaldehyde and immunoprecipitations, DNA labelling and array hybridisations were carried out as described elsewhere (Borneman, Submitted; Borneman, 2006).

5.4.4 Microarray analysis and scoring

Following scanning, the two files corresponding to each channel (in .pair file format) were uploaded to the Tilescope pipeline for high-density tiling array data normalization and scoring (http://Tilescope.gersteinlab.org (Zhang, submitted). Tilescope processes the data

in a sequential fashion. These steps can be approximately grouped into three stages: normalization, tile scoring, and feature identification. We describe some of the key steps in our system below in detail.

Normalization. For each array in an experimental set, the relative contributions of the test and reference signals are compared. Ideally, if nucleic acid probes have equal concentration in the test and reference samples, the signals of the two dyes should be relatively equal (i.e. the ratio of the two signals should be close to one for probes hybridizing to an equal degree in both fluorescence channels). In practice, the signals can be rather different due to different chemical properties of dyes and non-specific or incomplete hybridization to the array. Normalization is used to compensate for these effects by either applying a scale factor to equalize signals from probes with unchanged concentration or imposing the same empirical distribution of signal intensities. Tilescope uses Quantile normalization. This not only normalizes data between channels and across arrays simultaneously but also removes the dependency of the log-ratio on the intensity in one step. It imposes the same empirical distribution of intensities to each channel of every array. Quantile normalization is fast and has been demonstrated to outperform other normalization methods)(Bolstad, 2003). *Tile scoring.* Tilescope pools the normalized log-ratios of all tiles on every array into a matrix and sorts them based on the tiles' genomic locations regardless of which strand they come from. At the tile scoring step, the program identifies tiles that exhibit differential hybridization. These tiles ultimately correspond to the locations of transcription factor binding sites. Instead of considering each tile across array replicates separately, a sliding window around each tile that incorporates the hybridization intensity of its neighboring tiles is used. For each tile, given its neighboring tiles across replicates, Tilescope calculates its signal, the pseudo-median log-ratio value

$S = median(logratio(i) + logratio(j))/2)$

from all $(i,\ j)$ pairs of tiles in the sliding window across arrays. Due to the small sample size in each sliding window, whether the intensity distribution is normal or not in a given window cannot be reliably assessed. Without making the normality assumption

about the intensity distribution, Tilescope uses the nonparametric Wilcoxon signed-rank test (Troyanskaya, 2002) to compare the test with the reference signal intensities and quantifies the degree of significance by which the former consistently deviates from the latter across the window. At the scoring step, Tilescope generates two tile maps, the signal map and the p-value map. Two values are calculated for each tile position: the pseudo-median of log-ratios, the signal, as a measure of the hybridization difference between test and reference samples at this genomic location and the probability, the p-value, that the null hypothesis (the local intensities of the test and the reference samples are the same) is true. *Feature identification.* Given the tile map annotated with pseudo-medians and p-values, Tilescope filters away tiles that are below user-specified thresholds. Retained tiles are used to identify binding sites. Based on the observation that a tile is usually too short to constitute a feature alone, the Max-gap and min-run method, modified from the scoring scheme used in Cawley et al. (Cawley, 2004), groups together qualified tiles that are close to each other along the genomic sequence into 'proto-features' and then discards any proto-features that are too short. To use this method, a user needs to specify the maximum genomic distance ('max-gap') below which two adjacent qualified tiles can be joined and the minimum length ('min-run') of a proto-feature for it to be qualified as a feature. *Experimental scoring variables.* For all experiments, max-gap was set at 60 bp and min-run at 120bp. p-value cut offs were set at $\leq 1X10^{-4}$, with pseudo-median cut offs of $\leq$1.25 (*Sc* Tec1 and Ste12, *Sm* Tec1 and Ste12), $\leq$1.10 (*Sb* Tec1) and $\leq$1.00 (*Ca* Cph1) used. Independent confirmation of the chIP chip procedure was performed by qPCR for binding targets from across the range of binding strengths, plus two non-enriched controls; all positives targets which gave PCR signals were enriched compared to the negative controls.

5.4.5 Array Reproducibility

To determine the reproducibility of the chIP chip and high-density microarray methodology, duplicate Ste12 and Tec1 binding experiments (each consisting of three additional biological

replicates) were performed in *S. cerevisiae* at the beginning and end of the study. In this second set of samples, Ste12 bound 290 targets, while Tec1 bound 357 targets. The two duplicate datasets showed a high degree of congruence, with 97% (Ste12) and 95% (Tec1) of the targets from the smaller dataset contained within the larger set such that nearly all of the observed differences were due to variations near the signal threshold used (Figure reffig:fig-regnet-tracksB.

5.4.6 Species-specific arrays and sequence independence

In order to compare the results of the species-specific arrays, we perform tests to ensure that the actual sequences printed on the arrays did not significantly affect the hybridization results and any subsequent binding sites scoring. To calculate the sequence-independent array reproducibility, three biological replicates were chosen and alternating probes from each replicate were separated into two new result files. These new files were then scored independently for binding events using Tilescope. Comparison of the files showed that 93% of the total binding peaks arrays were shared, with each pair of peaks differing in average signal enrichment by ±5% and starting and stopping on average ±68bp from each other.

5.4.7 Genome Alignment and Standardization

As the *S. mikatae* and *S. bayanu*s genome sequences are in draft form, difficulties arose in directly comparing results from different species as regions from one species may not necessarily be present in all. To guard against a lack of sequence representation influencing our results, the genomes of all three species were aligned to ensure that orthologous sequences were present in all three species for any bound region. This was performed using conserved gene sequences and chromosomal synteny to position sequence contigs from *S. mikatae* and *S. bayanus* onto the *S. cerevisiae* genome. Over 250 instances were identified in which differences in binding between at least two of the three species were attributed to species-specific gene annotation, missing sequences or contig breaks (often due to the presence of repetitive Ty elements, which are bound strongly by both Ste12

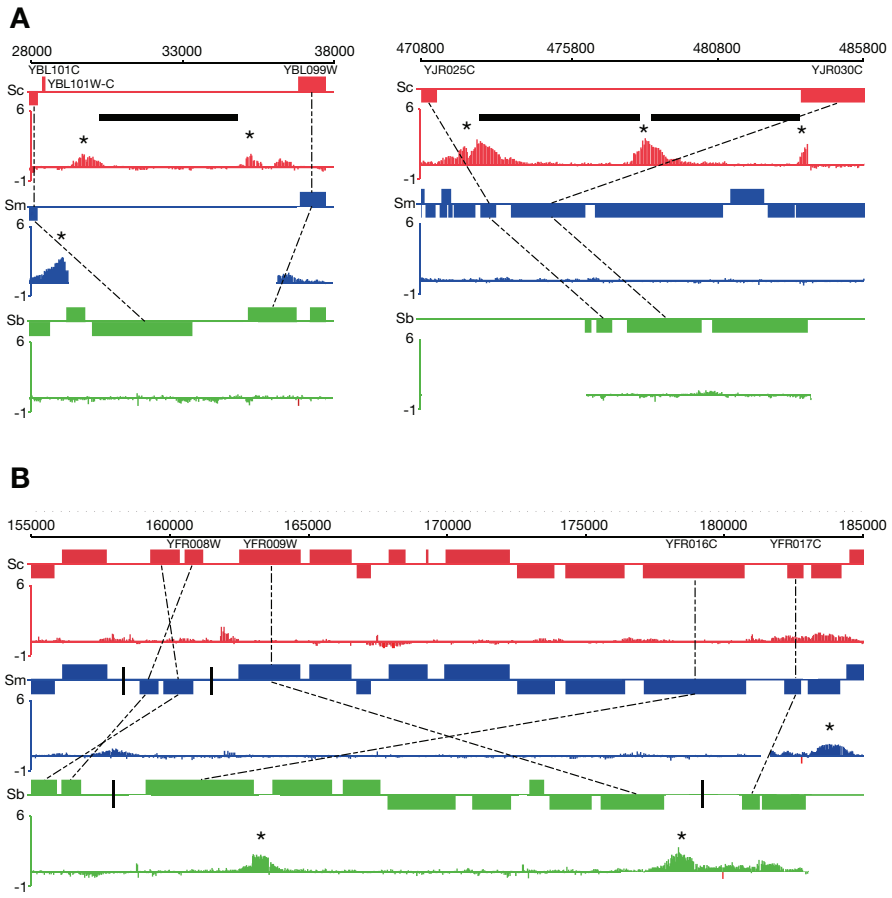


Figure 5.6: Differences in binding caused by (A) Ty elements (black bars) and (B) genomic rearrangements (inversion pictured here, with the break points indicated by the vertical black lines). In each case, the binding signal (log2 tagged versus untagged) of Ste12 in S. cerevisiae (red), S. mikatae (blue) and S. bayanus (green). In each case, the positions of homologous ORFs between species are indicated by the dashed lines.

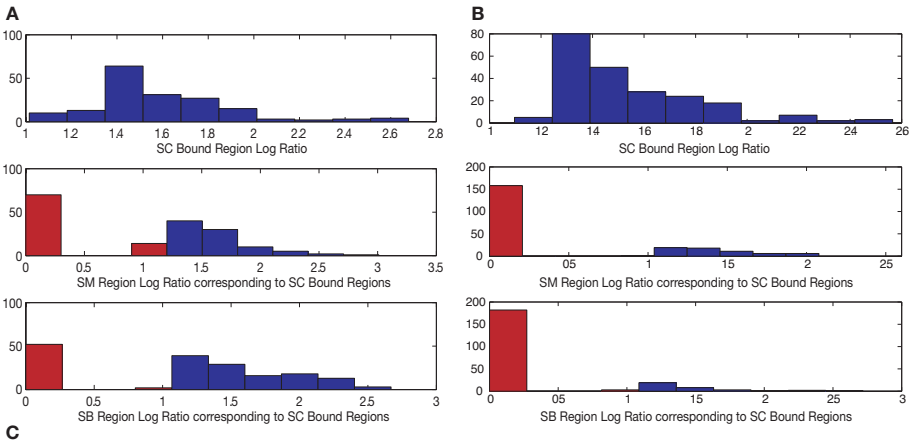
and Tec1, Figure 5.6). To prevent these ambiguous regions from affecting further analyses, they were excluded from any subsequent calculations.

5.4.8 Measuring Threshold Effects

To examine the effect of setting different thresholds on the amount of binding conservation, we set one species as the reference and examined the signal distribution of all regions in the remaining two species (Figure 5.7). We next determined the number of orthologous unbound regions in each of these two remaining species where the intensity was just below threshold but still above background. In the case where 2 of 3 species had binding, we showed that 6, 9, and 6 for *S. cerevisiae*, *S. mikatae*, and *S. bayanus* respectively had regions that fit these criteria (Figure 5.7).

5.4.9 Motif Discovery and Scoring

MDscan (Liu, 2002) was used to generate the position weight matrices for both Tec1 and Ste12. In all cases, input data to MDscan included the central 250 bp of each bound region (corresponding to the center of the binding peak). For Tec1, the entire list of bound regions from all species was sorted by signal intensity with the top 20 sequences used to seed the algorithm. Employing the same strategy as listed above, failed to elicit a significant match to the known Ste12 consensus sequence; however, upon restricting the search in a species-specific manner, a suitable PWM was obtained for both *S. cerevisiae* and *S. mikatae*, but not for *S. bayanus*. The *S. mikatae* PWM was used for all subsequent analysis. Logos were prepared using Weblogo (http://weblogo.berkeley.edu/logo.cgi). To compare between the bound and unbound orthologous regions, 1 kb regions corresponding to the peak of each chIP hit for the bound regions and 1 kb regions directly upstream of unbound homologs were selected for motif searching. 1 kb was selected to ensure all of the potential regulatory space was searched; however, given that the bulk of the PWM matches were 200 bp - 500 bp upstream of start, 1 kb maybe somewhat larger than necessary. The program MAST (Bailey, 1998) was then used to score both the bound and unbound regions



C

Bound in Two of Three Species	0<X<1.2
Bound in SC, SM NOT SB	12
Bound in SC, SB NOT SM	18
Bound in SM, SB NOT SC	12
Total below threshold but above background	42

Bound in One of Three (at least one below theshold but above background)	0<X<1.2
Bound in SC, NOT SM or SB	17
Bound in SM, NOT SC or SB	19
Bound in SB, NOT SM or SC	27
Total below threshold but above background	64

Bound in One of Three (both below threshold but above background)	0<X<1.2
Bound in SC, NOT SM or SB	6
Bound in SM, NOT SC or SB	5
Bound in SB, NOT SM or SC	1
Total below threshold but above background	18

*where X refers to the target's log ratio signal intensity

Figure 5.7: (A) Distribution of orthologs of Ste12 bound regions from S. cerevisiae. (B) Distribution of orthologs of Tec1 bound regions from S. cerevisiae. Blue bars, enrichment signals higher than the threshold, red bars, below the cut off threshold. Background intesity is equal to 0 (C) Total numbers of binding events which were below the signal enrichment threshold, but which had signals which were detectable above background levels.

(Ste12, $P < 0.0001$; Tec1, $P < 0.001$).

5.4.10 Testing Significance of the Relationship between Binding, Motif Matching, and Conservation

To test the significance of the configuration of binding and sequence motif matching, each chIP hit for each of the three species was assigned a two bit code, where the first bit represented binding/no binding and the second bit motif match/no motif match. This gave a 6 bit representation, so for example 1 1 1 1 1 1 means that the region was bound in all three species, and there was a motif match in all; whereas, 101010 indicates the region is bound in all three but that there are no matches to the motif. We next shuffled each column of this table, which preserves the overall distribution but scrambles the relationship between binding, conservation, and motif matching. We calculated a p-value by counting the number of times that the frequency of the class i in each of the 1000 random datasets exceeded the frequency of class i in the scored data.

5.4.11 Data Deposition

All array designs have been deposited in GEO under the accession numbers GPL4033, GPL4034 and GPL4037 for *S. bayanus*, *S. mikatae* and *C. albicans* respectively. All array results were deposited in GEO under the series accession number GSE5421. Detailed lists of scored binding regions, conservation information, and motif scores are available from http://www.gersteinlab.org/proj/regnetdiverge.

Chapter 6

Cross Integration: A framework for identifying cross patterns in systems biology

6.1 Summary

The scale and complexity of available yeast systems-data presages the considerable scale-up to humans and other organisms, and mining such complexity represents an exciting challenge. Although numerous scientific discoveries have resulted from the integration of these datasets, current schemas only allow integration in a single dimension lacking the flexibility to accommodate data that do not share the same index. As an example, understanding the relationship between the conservation score of target genes and their associated transcription factor (TF) binding sites may elucidate factors in gene regulatory evolution; however, as binding sites and gene targets do not share the same index, the two types of objects features can neither be correlated nor "stacked.

Here, we introduce cross patterns to describe more complex, multidimensional relationships that spans across differently indexed features. Further, we develop a method Identifying Transitive Relationships (ITeR) to identify them. The key of ITeR is a connector

matrix that maps one type of feature to another. Construction of a connector between target genes and their associated TF binding sites is straightforward; however, we can also formulate more complex connectors.

As an example, we used a connector matrix mapping small molecules to their associated target proteins and applied ITeR to an available chemogenomics dataset. By integrating 1194 drug sensitivity profiles, 6 types of structural features, and 7 types of protein properties, we identified a number of cross patterns spanning structural properties of a drug and features of their target proteins. Some were intuitive, such as, the charge of drugs and protein targets are often complementary. Others were less obvious, such as, a shared sensitivity to both a particular type of environmental stress and to a particular structural parameter of a drug. This finding suggests that one may be able to track sets of physical properties underlying common stress responses.

6.2 Introduction

From gene regulation to kinase specificity to protein-protein interactions, the growth of systems-wide *S. cerevisiae* datasets has led to rapid advances in our understanding of yeast biology. The experimental innovations and computational techniques developed in yeast presage the considerable scale-up coming for human from efforts such as the ENCODE consortium (ENCODE, 2007). To fully take advantage of the depth and breadth of such system-wide data, questions will increasingly rely on integrating heterogeneous forms of data. Current integration schemes revolve around a "gene or protein centric" view where individual datasets can be conceived of as data layers, and positions within the individual layer are determined by referencing a gene or protein. In other words, the gene or protein serves as the index to all of the individual data layers allowing the gene or protein to be represented by a single data vector (Lan et al., 2002) or "stack" of all its features. Integration then is a matter of performing operations on these stacks.

We can use "stacking" as an explanatory term to describe the creation of such data

vectors; however, operations to combine these vectors or "stacks" have taken many non-trivial forms including functional coupling (Fraser and Marcotte, 2004), phylogenetic profiling (Marcotte et al., 1999; Pellegrini et al., 1999) and various machine learning approaches including decision trees (King et al., 2003), Bayesian networks(Jansen et al., 2003; Troyanskya et al., 2003), unsupervised approaches (Flaherty et al., 2005; Bergmann et al., 2003), and many different kinds of kernel methods (Ben-Hur and Noble, 2005; Lanckriet et al., 2004; Tsuda and Noble, 2004). These have provided a wealth of insights into biological processes from gene essentiality (Seringhaus et al., 2006) to arsenic resistance (Kelley and Ideker, 2005) and DNA damage (Haugen et al., 2004; Begley et al., 2004) among many others. Further, through such integration, it has been shown that genes or proteins that share similar properties (e.g. protein interaction partners) tend to share similar functional roles (Kelley and Ideker, 2005; Tasan et al., 2008; Jansen et al., 2003; Parsons et al., 2004; Wong et al., 2004).

6.2.1 Limitations of Current Techniques

The major theme of "stacking" techniques is that the features are "indexable"by a single class of variables: gene or protein. This is an intuitive solution when all the data can be stacked. That is, when they are of the same type and can thus be treated in a similar manner (e.g. stored in the same relational table and queried directly). The limitation in stacking is in capturing connections between features that are indexed on different kinds of objects and thus cannot be stacked (Figure 6.1). To give a simple example, we pose the following question: *Is there a relationship between the conservation scores of genes and conservation scores of their transcription factor (TF) binding sites?*. It is clear that properties of TF binding sites cannot simply be stacked on top of properties of targets as they do not share the same index. However, despite the dissimilarity of object types, such integration could potentially identify principles governing gene regulatory evolution that would not be observable from just looking at the patterns of a single gene or single set of binding sites. Similarly, identification of associations between properties of a small molecule

(e.g. charge) and properties of the small molecule's target protein (e.g. protein's charge) could provide additional details about general mechanisms underlying such interactions. The means to identify this type of indirect, complex connection remains an open question.

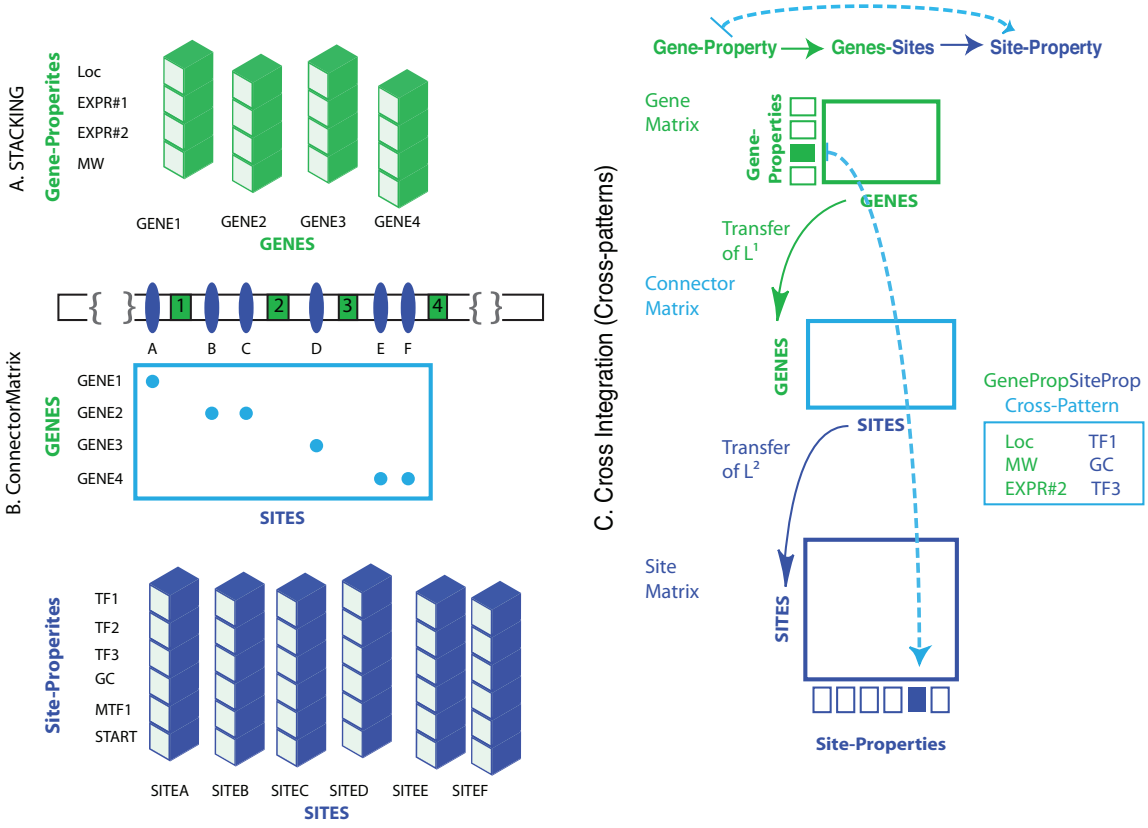


Figure 6.1: Graphical representation and comparison of ITeR algorithm. (A) Each stack represents a single gene (green), and each slot in a stack is a feature of the gene or its associated protein product. (B) This panel gives an example of connector matrix (cyan) that maps gene targets (numbered green boxes) to their associated transcription factor binding sites (lettered blue ellipses). The rows of the matrix are the genes (represented as stacks in panel A) and the columns are binding sites (represented as stacks in B). A dot denotes a mapping between a gene and a known binding site. The bottom panel shows stacks of site-properties (blue) analogous to A. (C) Schematic of the ITeR algorithm (see text for more details).

6.2.2 Connector Matrix

Here, we introduce the concept of a connector matrix to map between two differently indexed datasets. In the transcription factor example, we can easily derive such a matrix

where the rows are the gene targets and the columns would be the binding sites identified through ChIP-Seq or other kinds of transcription factor mapping datasets. We developed a new algorithm, Identifying Transitive Relationships or ITeR, to leverage this connector matrix in order to systematically combine information from multiple tables with different indices. This allows one to not only stack features in a single dimension but also to span across multiple ones. Thus, ITeR captures a new type of relationship between different types of data (e.g. gene and binding site properties) called a cross pattern. It is important to note that this is not a correlation; there is no obvious way to correlate two differently indexed objects. Rather, cross patterns are a means of achieving consilience between different objects, and we formally define both cross patterns and the algorithm itself in the text.

6.2.3 Applying ITeR to Chemogenomics

Finally, we applied ITeR to the earlier example of identifying relationships between small molecule properties and properties of their target features. We first selected six molecular descriptors frequently used in computational chemistry to characterize small molecules (Todeschini and Consonni, 2000; Tetko et al., 2005): molecular weight, charge, the number of aromatic bonds and rings, hydrophilicity, and hydrophobicity (MlogP), and seven categories of systems data as features of the target proteins including physicochemical properties, localization, known function and process categories from GO, topological properties of the regulatory and protein-protein interaction networks, gene composition features, and response to environmental stress. The connector matrix was derived from a recent chemogenomics dataset generated by Hillenmeyer *et. al.* (Hillenmeyer et al., 2008) that provides a potential mapping between over 300 drugs and their effect on almost all potential target *S. cerevisiae* proteins.

For the purposes of this work, we assume this data allows us to infer a direct physical connection between a drug and a target protein; there are some limitations to this assumption that we discuss in considerable detail later. Despite these caveats, given a direct

physical connection between a drug and a particular protein, it is perhaps intuitive that physicochemical properties of a drug and the protein it targets would be complementary (e.g. charge of drug and protein). Indeed, a large body of literature describing the relationship of structural properties of small molecules and individual proteins exists. However, the structural diversity of the compounds assayed in the Hillenmeyer dataset provides the opportunity to investigate how changes in a structural parameter influence the drug's effect on a proteome-wide scale.

As described above, these relationships cannot be ascertained directly as molecular descriptors cannot be indexed on a particular protein. Rather such comparisons must be done indirectly through abstracting features of the small molecules and similarly features of the proteins themselves. Here, we applied ITeR for such indirect integration to test the hypothesis that the subset of proteins affected by a structural parameter may also share physicochemical or other types of properties. This allowed us to pose questions in the form: *Does a property of a small molecule relate to properties of its protein? For instance, Do charged proteins exhibit a tendency to interact with charged compounds?*

We integrated the response of 1194 proteins that were subjected to 253 different small molecules, six properties of these small molecules, and 7 different types of protein properties to identify numerous drug property-protein property cross patterns comprising 10 types of environmental stress, 2 gene composition features, 2 physicochemical properties, 1 network topological property, and many function and process categories from GO along with strong enrichment for particular localization categories.

6.3 Results

6.3.1 Identifying Transitive Relationships (ITeR)

We applied ITeR to identify cross patterns between properties of drugs and properties of their target proteins. We first give a general, high-level overview of the ITeR algorithm before describing the method more formally.

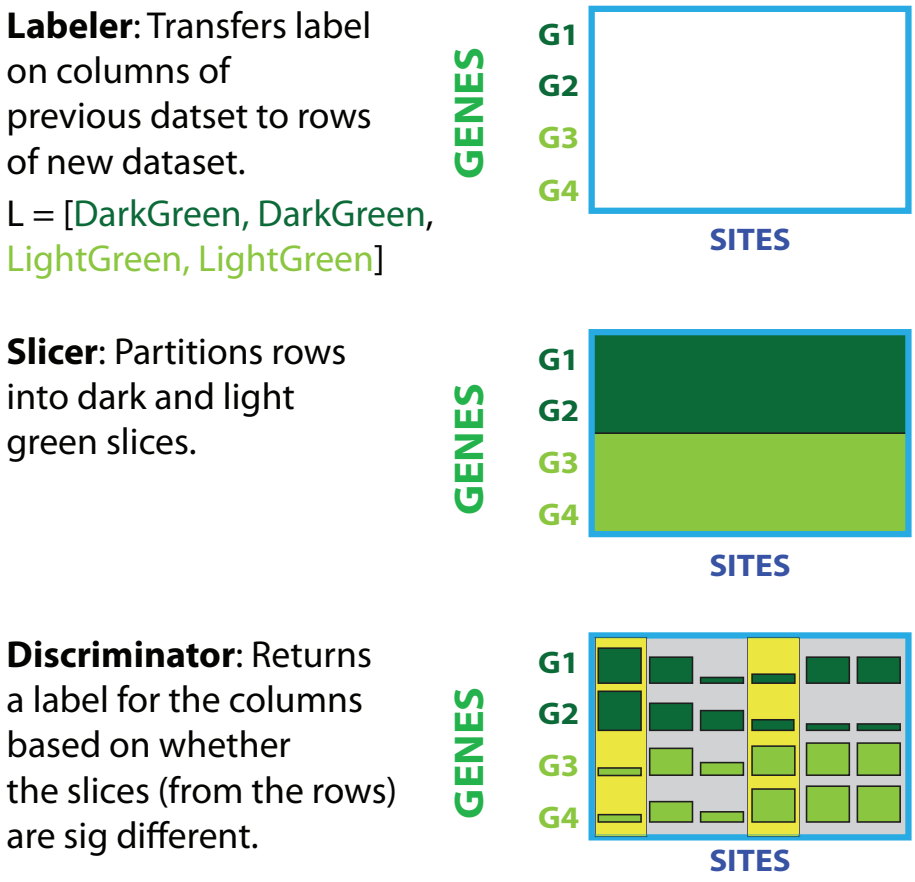


Figure 6.2: More details of the labeler, slicer, and discriminator.

ITeR has three generic types of functions: a labeler, a slicer, and a discriminator. The labeler transfers a label from one dataset to another (rows to columns or the reverse). As an example, drugs (rows) maybe classified as either high or low molecular weight, and the labeler transfers this classification (high or low) to the columns of the new dataset (e.g. drug response dataset). The slicer partitions this new dataset into separate "slices" on the basis of the label generated in the previous step (e.g. treatments with high versus low molecular weight drugs).

Finally, the discriminator applies some statistical test to the slices to generate a new set of labels. In our example, proteins that had disparate responses to large and small drugs were labeled the "sensitive" proteins and those that did not were labeled the "insensitive"

proteins. More generally, the discriminator determines if there are any features in the second dataset that "discriminate" among the labeled slices based on the parameter in the first dataset. The entire process is iterated until all of the matrices have been used (Figure 6.2). As we will show more formally, such relationships can be thought of as transitive.

Basic Notation

We apply our algorithm to three matrices M^1, M^2, and M^3, which contain drug properties, a mapping between drugs and their effect on proteins, and protein properties, respectively. It will suffice in this section to think of these matrices more abstractly. In the following, we define some basic notation to support the labeler, slicer, and discriminator described conceptually above, and in the next section we define the operations themeselves.

Let us consider a matrix M whose rows and columns are indexed by sets I and J respectively. Compactly, this specification is written $M : \mathbb{R}^{I \times J}$, where the colon is read "is of type". Elements are accessed using brackets: each element $M[i, j]$, where $i \in I$ and $j \in J$, has a value $x \in \mathbb{R}$. The set of indices are just named values, and in particular there is no assumption that they be contiguous.

The labeling we provide also requires vectors of categorical values, which can be defined in a manner similar to matrices. For example, if matrix M's columns are going to be labeled either a or b, we can define the set of labels $\mathbb{L} = \{a, b\}$. Then the notation $L : \mathbb{L}^J$ declares a vector L indexed by $j \in J$ and whose values are elements of $\mathbb{L}$. For example, we might have

$$L = \begin{bmatrix} a & a & b & a \end{bmatrix}, \tag{6.1}$$

where we have assumed J has 4 elements.

We can also invert L to get the subset of columns with a particular label. Let $\hat{J}_a = \{j \in J \mid L_j = a\}$ be the subset of columns J that were labeled a. Similarly, $\hat{J}_b = \{j \in J \mid L_j = b\}$. For labeling vector (6.1), we have $\hat{J}_a = \{j_1, j_2, j_4\}$ and $\hat{J}_b = \{j_3\}$.

Next we define a notation for slicing matrices. Assume we have a matrix $M : \mathbb{R}^{I \times J}$,

and consider a subset of rows $I' \subseteq I$ and a subset of columns $J' \subseteq J$. Then, we let $M[I', J']$ denote the slice of matrix M such that $M[I', J'] : \mathbb{R}^{I' \times J'}$, where the values for the $(i, j)^{\text{th}}$ element are equal to those in the original matrix. Note $M[I', J']$ has no values for $i \notin I'$ or $j \notin J'$. Often we need to extract a single row i which could be done with the notation $M[\{i\}, J]$, but we allow omission of the curly braces and write simply $M[i, J]$. Analogously, column j is extracted by writing $M[I, j]$ instead of $M[I, \{j\}]$. When both sets are singletons, this coincides with the notation for element access, $M[i, j]$.

Formal Definition of ITeR

Thus far our notation has assumed a single matrix, but our algorithm iterates over a sequence of datasets $M^1, M^2, \ldots, M^n$. Furthermore, it is required that the columns of each matrix are indexed over the same set as the rows of the next. Thus, we refer to the n^{th} matrix's rows as I^{n-1} and its columns as I^n, instead of I and J as above. The $(n+1)^{\text{th}}$ matrix's rows would then be I^n, giving the desired correspondence between the columns and rows of adjacent matrices. In summary, we have

$$M^1 : \mathbb{R}^{I^0 \times I^1}$$

$$M^2 : \mathbb{R}^{I^1 \times I^2}$$

$$\ldots$$

$$M^n : \mathbb{R}^{I^{n-1} \times I^n}$$

Similarly, the labeling of the n^{th} matrix's columns will be denoted L^n instead of just L, and the set of columns labeled a in the n^{th} matrix is denoted $\hat{I}^n_a$, instead of $\hat{J}_a$ as above.

Labeling is the first step in our algorithm. On the n^{th} iteration, we operate on matrix $M^n : \mathbb{R}^{I^{n-1} \times I^n}$, and use the labeling $L^{n-1} : \mathbb{L}^{I^{n-1}}$ of the previous matrix. Assuming $\mathbb{L} = \{a, b\}$, we can get the sets $\hat{I}^{n-1}_a$ and $\hat{I}^{n-1}_b$, which give the columns of M^{n-1} that were labeled a and b, respectively. This labeling can be transferred directly to the current matrix M^n because its rows are equal to the columns of M^{n-1}.

The next step is to slice M^n along its rows such that each resulting partition has only the rows with one label. These slices are $M^n[\hat{I}_a^{n-1}, I^n]$ and $M^n[\hat{I}_b^{n-1}, I^n]$. For simplicity of discussion, we have assumed the specific labels a and b, but in general there can be many labels, leading to more than just two slices.

Finally, let f^n denote the discriminator employed to label the columns of M^n. Various statistical tests might be used (e.g. t-test or ANOVA for more than two slices), but the general idea is that f^n tests whether the values of each column in M^n differ amongst the slices generated in the previous step. The output of f^n is a labeling vector $L^n : \mathbb{L}^{I^n}$. (Again for simplicity, we have assumed the columns of M^n are assigned the same labels $\mathbb{L}$ as those of M^{n-1}, but in general different sets of labels can be used.) For each column $j \in I^n$, the discriminator f^n will compare the j^{th} columns of the slices generated in the previous step, $M^n[\hat{I}_a^{n-1}, j]$ and $M^n[\hat{I}_b^{n-1}, j]$, and assign a label to column j based on the result. For example, if the values in column j differ significantly between the two slices, $L[j]$ might be set to a and otherwise it might be set to b.

The net input to the discriminator f^n is the current matrix M^n and the labeling of the previous matrix L^{n-1}. Its output is L^n. In other words, $f^n : \mathbb{R}^{I^{n-1} \times I^n} \times \mathbb{L}^{I^{n-1}} \to \mathbb{L}^{I^n}$, which means f^n is a function taking two arguments, the first a matrix of type $\mathbb{R}^{I^{n-1} \times I^n}$ and the second a labeling vector of type $\mathbb{L}^{I^{n-1}}$, and returns a labeling of type $\mathbb{L}^{I^n}$. Precisely then, we have $L^n = f^n(M^n, L^{n-1})$, which makes the transitive nature of our algorithm apparent.

The final output of the algorithm defines a new type of relationship which we call a cross pattern. A cross pattern defines a relationship between a row $i \in I^0$ of the initial matrix and a column $j \in I^n$ of the final matrix such that j is labeled as being *interesting* (according to the particular application) through the propagation of labelings from L^0 through L^n. We notate the set of cross patterns between all the rows of the initial matrix and all the columns of the final matrix by $I^0 \mapsto I^n$. The specific cross pattern would be defined as $i \mapsto j$.

On the first iteration, numbered 1, an initial labeling L^0 must be obtained from an

external procedure. In the next section, we show that our specific application of ITeR does not require this, or alternatively that it consists trivially of a single label. Thus, the initial discriminator f^1 differs slightly in that it does not compare values between multiple slices, but uses another test to assign labels to the first set of columns.

Application of ITeR

We apply the ITeR algorithm to three datasets: drug property measurements $M^1 : \mathbb{R}^{I^{\text{dprop}} \times I^{\text{did}}}$, the Hillenmeyer drug response dataset $M^2 : \mathbb{R}^{I^{\text{did}} \times I^{\text{pid}}}$, and protein property measurements $M^3 : \mathbb{R}^{I^{\text{pid}} \times I^{\text{pprop}}}$. Note these are structured as required, with the column indices of each equaling the row indices of the next where the abbreviations are as follows: did - drug IDs, dprop - drug property names, pid - protein IDs, and pprop - protein property names. The goal then is to link M^1 to M^3 through M^2 by testing for the presence of transitive relationships resulting in cross patterns of the form drug-property/protein-property.

We actually apply the algorithm to each row of M^1, one at a time. Thus, the initial matrix for each application of ITeR is $M^1\left[m, I^{\text{did}}\right]$, where m is the current drug property under consideration. A specific m is implicitly assumed in the subsequent discussion.

The discriminator for $M^1\left[m, I^{\text{did}}\right]$ determines whether each drug has a high or low value for drug property m. It simply tests whether the value for each drug is above or below the median over all the drugs. This produces the labeling $L^{\text{did}} : \{\text{lo}, \text{hi}\}^{I^{\text{did}}}$, a vector assigning the label lo or hi to every drug ID in I^{did}. Recall that L^{did} can be inverted to give $\hat{I}^{\text{did}}_{\text{lo}}$ and $\hat{I}^{\text{did}}_{\text{hi}}$, the drug IDs assigned the label lo and hi.

The second dataset is M^2, which gives the degree of growth defect of each target subjected to each drug. The labeling from the first matrix is applied to M^2 by slicing it into the two matrices $M^2[\hat{I}^{\text{did}}_{\text{lo}}, I^{\text{pid}}]$ and $M^2[\hat{I}^{\text{did}}_{\text{hi}}, I^{\text{pid}}]$, a partitioning of M^2 into the lo- and hi-labeled drugs.

For each protein in I^{pid}, the columns of M^2, we consider whether the protein's affect on growth is significantly different when subjected to the lo- versus hi-labeled drugs.

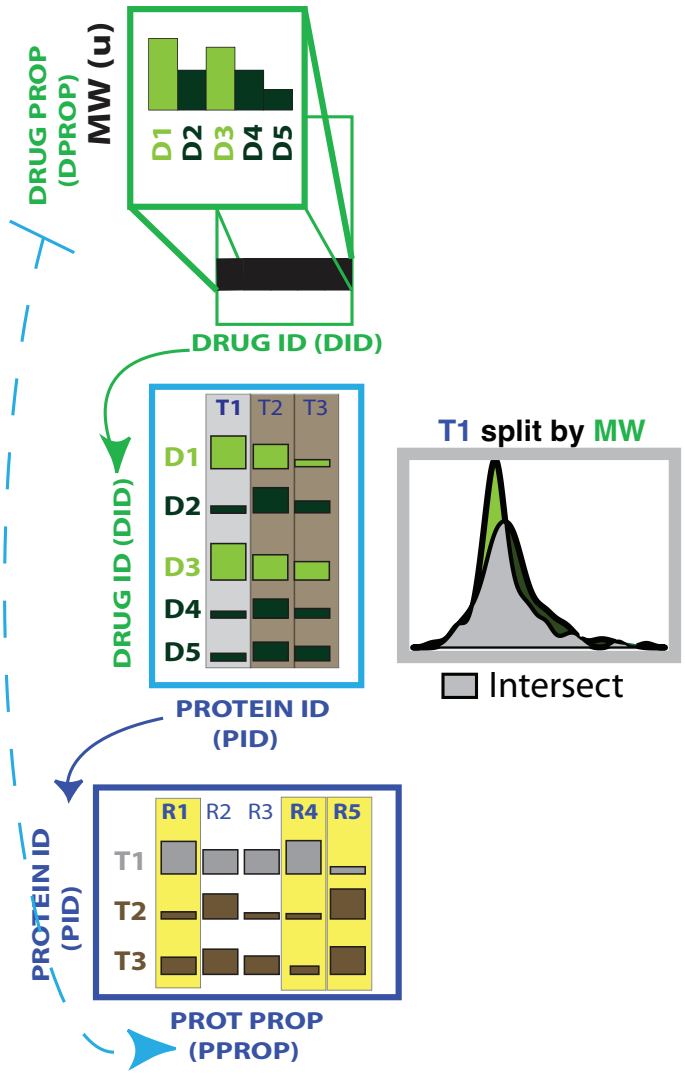


Figure 6.3: Illustration of ITeR algorithm. Drug property matrix (green) with the drugs color-coded by molecular weight (light green high and dark green low. Connector matrix (cyan) with a zoom in panel on column T1 showing the disparate response to high and low molecular weight compounds (peak height corresponds to how affected the protein was by the drug). Protein property matrix (blue) showing the sensitivity of R1, R4, and R5 to the labeling on the target proteins.

This produces the labeling $L^{\text{pid}} : \{\text{sensitive}, \text{insensitive}\}^{I^{\text{pid}}}$. In other words, we classify proteins as either DP-sensitive or DP-insensitive, where DP is replaced with the specific drug property m under consideration.

Analogously, the third dataset M^3 gives a measurement of many protein features for every protein. As with M^2, we slice M^3 into the two matrices $M^3[\hat{I}^{\text{pid}}_{\text{sensitive}}, I^{\text{pprop}}]$ and $M^3[\hat{I}^{\text{pid}}_{\text{insensitive}}, I^{\text{pprop}}]$, containing the rows with just the sensitive- and insensitive-labeled proteins.

As in M^2, for each protein property in I^{pprop}, we consider whether the protein property's measurement is significantly different between the sensitive and insensitive targets. This produces the labeling $L^{\text{pprop}} : \{\text{yes}, \text{no}\}^{I^{\text{pprop}}}$. Now we can conclude through transitivity that any yes-labeled protein property is sensitive to the original drug property m under consideration as shown in Figure 6.3. We term this transitive relationship between each drug property and each protein property a cross pattern. This then is the heart of the algorithm. There is no means to directly correlate these two types of data. Further there is no matrix that directly involves both indices; however, by using the Hillenmeyer dataset as a connector matrix and employing ITeR, one can begin to explore the relationship between these two disparate and differently indexed data types.

We seeded the algorithm with six different drug properties. By changing the initial I^{dprop} and repeating the entire procedure, we can find cross patterns between multiple drug properties and protein properties, $I^{dprop} \mapsto I^{pprop}$. An individual cross pattern would be represented as $d \mapsto p$ where $d \in I^{dprop}$ and $p \in I^{pprop}$.

6.3.2 DP-sensitive Proteins

A summary of the number of the proteins found to be sensitive to each type of drug property is provided in Table 6.1. Cross patterns are classified as either: direct, secondary or complex, based on the type of protein property they refer to. Direct cross patterns arise from both physicochemical properties, such as charge and primary sequence features, such as, codon bias. Cross patterns derived from secondary characteristics include localization,

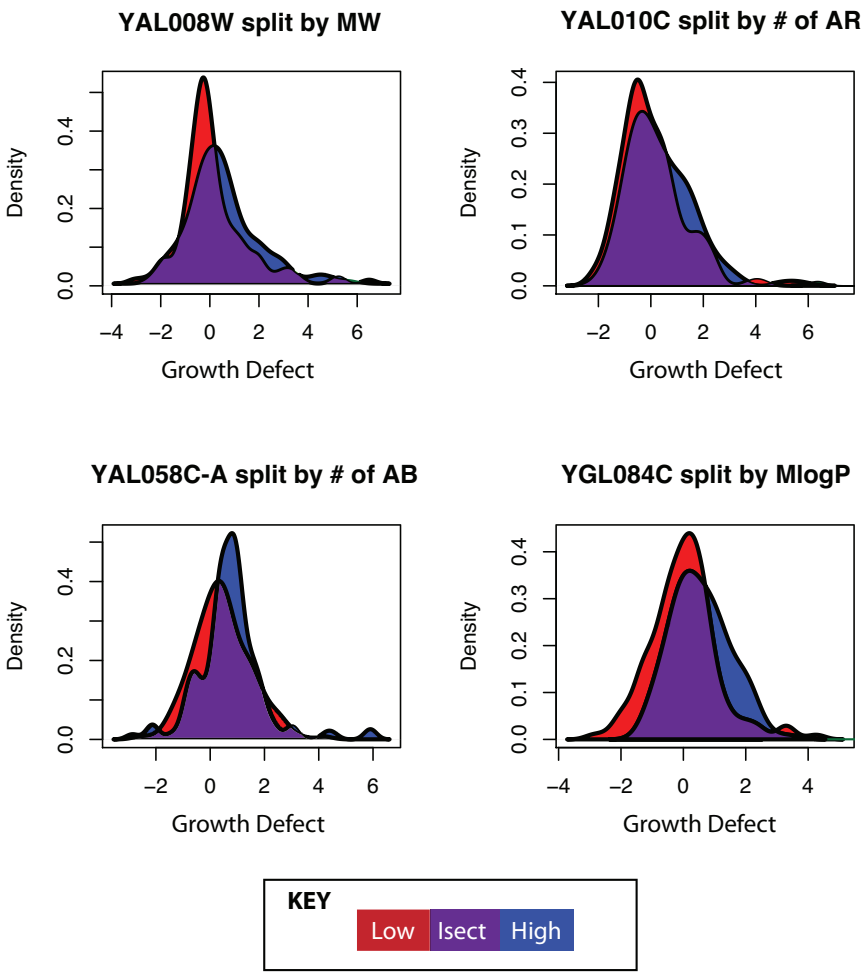


Figure 6.4: Plots of DP-sensitive proteins. Plots were generated by first segmenting small molecule treatments into high (above median, blue) and low (equal to or below median, red.) The x-axis is the growth defect score of the particular protein after treatment with a small molecule and the y-axis is the density plot. The purple region shows the overlap between the two distributions. The smaller this overlap the greater the difference in response to high and low-labeled drug treatments and the more "sensitive" the protein is to the value of the particular drug property.

	MW	Ms	nAB	ARR	Hy	MlogP
MW	77	170	119	143	153	261
Ms	9	102	136	162	158	253
nAB	5	13	47	95	126	229
ARR	4	10	22	70	145	253
Hy	6	26	3	7	82	249
MlogP	12	45	14	13	29	196

Table 6.1: Matrix showing the total number of proteins sensitive to each drug property. For each drug property pair (row, column), we report both the number of proteins that are sensitive to both properties (lower triangle, intersection) and the total number of proteins sensitive to either property (upper triangle, union). The diagonal is the total number of proteins that were sensitive to the particular drug property.

interaction network topology, functional categories, and finally complex cross patterns include more subtle properties such as stress response. Below, we show several examples of cross patterns that both recapitulate expected observations and suggest new indirect relationships between molecular descriptors and protein properties (Figure 6.5).

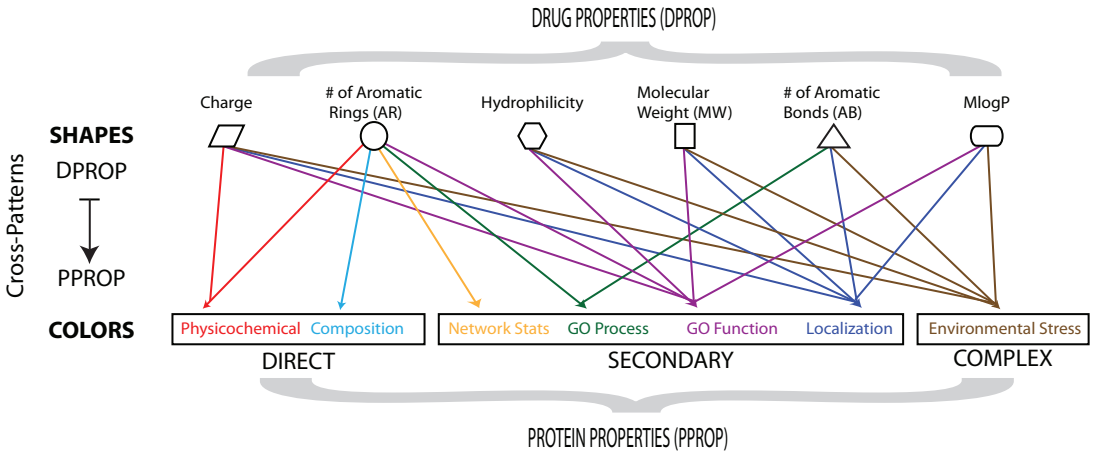


Figure 6.5: Graphical representation of cross patterns. The top portion of the tree are the drug features represented by different shapes. The bottom portion of the tree corresponds to the different types of protein properties: direct (physicochemical properties and gene feature composition), secondary (localization, network statistics, and GO terms), and complex features (environmental stress response). A line connecting the top and bottom portion are cross patterns.

6.3.3 Direct properties of small molecules are sometimes mirrored by those of their protein targets

Although cross patterns are identified indirectly, there is a physical basis underlying why physicochemical properties of small molecules may mirror physicochemical properties of their protein. In order to disrupt a protein's function, a small molecule must either bind directly to the protein or act indirectly by interfering with another component up or downstream. In the former case, there is a logical intuition that the composition of the small molecule would constrain or limit the types of proteins that it could affect or more positively that certain properties of a small molecule would be more favorable in disrupting a particular flavor of target proteins.

As an example, membrane proteins have a distinct polarity moving from the hydrophilic head group to hydrophobic tails. This unique composition disallows the entrance of most polar compounds while encouraging passive diffusion of hydrophobic compounds. Thus, it might be expected that membrane proteins may be more affected by the hydrophobicity of a compound, and indeed we identified a cross pattern, explained in more detail below, which recapitulated this observation.

A standard means of identifying membrane proteins is through measuring a protein's hydrophobicity (GRAVY score). In this scoring function, each amino acid has been assigned a hydropathy value based on its free energy change when moved from a hydrophobic solution to water. The highly polar arginine has the lowest score, and isoleucine has the highest. By summing over all residues, GRAVY scores provide a measure of hydrophobicity and are used to predict membrane-spanning domains and exposed regions (Doolitte, 1982). We find that the 102 charge-sensitive proteins had a large proportion of high GRAVY scores indicative of membrane protein enrichment and were more affected by compounds of low charge than highly charged compounds. Since low charge compounds would be expected to more easily interact and thus more easily disrupt the function of membrane proteins, this finding is concordant with membrane protein physiology.

Aromatic-ring (AR) sensitive proteins also exhibited a complementary physicochemical property. The seventy AR-sensitive proteins had a higher degree of aromaticity than the AR-insensitive set. Although we do not know enough about the placement of the aromatic residues on the protein itself, it is true that aromatic packing is of particular importance in aromatic proteins. Aromatic compounds would be particularly effective in disrupting aromatic protein function because of their ability to disrupt these stacking interactions.

AR-sensitive proteins had both a higher frequency of aromatic residues and a higher frequency of two sequence composition scores: optimal codon usage and codon bias scores. Primary sequence features can often serve as proxies for a physicochemical property. Using the above as an example, aromaticity can be estimated by counting the number of aromatic residues; however, it can also be measured indirectly by looking at the optimal codon usage and codon bias scores. The conceptual basis behind both scores is that areas of higher codon adaptation may represent optimizations or high selective pressure. Aromatics represent a special case as the aromatic amino acids have fewer possible codons. Tryptophan has just one codon, so it is always optimal, and phenylalanine and tyrosine both have only two. The lack of degeneracy itself has been thought to reflect the importance of aromatic interactions for biological function as the importance of these types of optimizations was vividly illustrated in a recent paper where the polio virus was almost completely crippled by recoding its genome with a non-preferred set of codons (Coleman et al., 2008).

6.3.4 Secondary Characteristics

We next examined cross patterns related to secondary characteristics of the protein. These include localization, interaction network topologies, and functional categories. Since a small molecule must be able to reach its protein to disrupt function, physiological aspects of a protein's cellular compartment can constrain physicochemical properties of a small molecule. Thus, the localization of the protein will have a profound effect restricting the entrance of compounds with one set of physicochemical characteristics and enhancing favorable access of others. Likewise, topological properties of the networks, such as

Protein Features	Drug Features	MW	Charge	# of Aromatic Bonds	# of Aromatic Rings	Hydrophilicity	MlogP
Direct	Physicochemical and Composition	-	GRAVY	-	CodonBias FOP Aromaticity	-	-
	Localization	Nuclear	Nuclear	Mitochondrion	-	Cytoplasm	Cytoplasm
		Other	Vacuole	Vacuole		Nuclear	Nuclear
			Golgi				Vacuole
	GO Process	-	-	RNA metabolism	RNA metabolis	-	Protein catabolism
	GO Function	-	Transferase activity	-	DNA binding	DNA binding	Protein binding
							Transcriptional regulator activitiy
Secondary	Network Features						
	Environmental Stress	Hyper-osmotic shock	Heat Shock	DDT	-	DDT	Hydrogen peroxide
				Hypo-osmotic shock		Mild Heat Shock with Hypo-osmotic Shock	DTT
				Amino-acid starvation			Alternative Carbon Source (Galactose)
Complex				Steady State			Alternative Carbon Source (Raffinose)

Table 6.2: Summary of cross patterns between drug and protein features

degree and betweenness, can be used to infer additional constraints on the physicochemical property of the drugs (Kim et al., 2007). In particular, hubs in interaction networks have been shown to be more disordered as this allows more potential interaction partners (Kim et al., 2008). Using ITeR, we identified global cross patterns between the physiological conditions encountered in the proteins' compartment and the compound's corresponding physicochemical properties. As an example, we observed that proteins, which responded differently to drugs that were charged as opposed to those that were uncharged, are more likely to localize to the Golgi (highly hydrophobic) or the nucleus than proteins which were as affected or unaffected by charged as with uncharged drugs (charge-insensitive proteins). Similarly, MW-sensitive proteins showed an enrichment to be localized to the nucleus, AB-sensitive to the mitochondria, and both AB-and MlogP-sensitive to the vacuole. Further, we found that AR-sensitive proteins had higher degree in the regulatory interaction network reinforcing the importance of disrupting aromatic interactions in this class of proteins.

It is sometimes important to identify how to disrupt a particular functional class (e.g. cell wall synthesis) irrespective of a specific protein (Begley et al., 2004). Thus, by calculating the GO enrichment within the drug sensitivity sets, we could see whether disruption of a specific functional class could be related to the compounds' physicochemical

properties. We found enrichment in RNA metabolism for both AR and AB-sensitive proteins, in DNA binding for AR and hydrophilicity-sensitive proteins, and protein binding for MlogP-sensitive proteins. In addition, charge-sensitive proteins showed an enrichment in transferase activity and MlogP in transcriptional regulator activity and protein catabolism.

We have summarized the individual cross patterns; however, we also searched for more general trends amongst all the cross patterns. As an example, the AR-sensitive set was enriched in aromatics (physicochemical), had higher scores for frequency of optimal codon usage and codon bias (composition), had a high degree in the regulatory network (topology), and finally was enriched in DNA-binders (function). The flexibility of the aromatic residues partially explains the higher degree observed in the regulatory network, and further integrating across all cross patterns suggests that to disrupt DNA-binders (such as transcription factors) compounds should most likely be aromatic. Thus, by looking at these cross pattern profiles a picture begins to emerge of the overall relationship between drug-features and protein-features.

6.3.5 Complex Protein Properties: Environmental Stress Response

For physicochemical properties, localization, network topology, and functional classes, the cross patterns are easy to interpret from a purely physical or physiological basis. To see if we could use these cross patterns in a more a subtle way, we searched for transitive relationships with a more complex and less obvious protein feature: environmental stress responses. Although it is well-known that disparate types of stress can result in a similar response, the mechanistic reasoning is unclear. We hypothesized that perhaps there is an underlying set of molecular properties unifying shared portions of the stress response, and moreover that the cross patterns derived from ITeR could be useful in identifying some of these properties. By applying ITeR to a dataset that measured the change in gene expression of yeast subjected to a number of different types of stress (Gasch et al., 2000), we identified cross patterns comprising ten different types of stress including amino-acid

starvation, heat, hypo, and hyper-osmotic shock, which we describe below in more detail.

6.4 Discussion

In this study, we presented a method to identify cross patterns between small molecule descriptors and seven classes of systems data. We showed that physicochemical properties of drugs often complement physicochemical properties of their proteins or physiological properties of the protein's compartments and further overlap with particular protein functions or processes.

As an example, we identified forty-seven proteins that were sensitive to compounds containing aromatic bond (AB-sensitive proteins) and showed that these proteins have a tendency to be localized to mitochondrion and the vacuoles. From this cross pattern, one could infer that access to mitochondrial or vacuolar proteins and thus to mitochondrial or vacuolar function is partially determined by the aromatic nature of the compound. Interestingly, a recent drug screen was performed to search for compounds that decrease radical oxygen species production and concomitantly increase mitochondrial oxidative phosphorylation (Wagner et al., 2008). After testing 2500 compounds for this activity, they identified six highly aromatic compounds as being particularly effective in modulating these mitochondrial functions. Cross-referencing with the Hillenmeyer set showed that three of these: nocodazole, mebendazole, and paclitaxel had been profiled on the deletion collection. Further, all three were in the top 20% for number of aromatic bonds within the Hillenmeyer set. This experiment thus provides support for the utility in identifying structural parameter sensitivities of functional classes.

6.4.1 Implications of Responses in Environmental Stress

In addition, to these perhaps intuitive cross patterns, we identified environmental stress response cross patterns requiring a more subtle interpretation. By looking at these cross patterns, we investigated whether molecular properties can tease out hidden similarities

that unify common stress responses or conversely provide a more mechanistic reasoning for the observed specificities (dissimilarities) in responding to stress. These specificities were observed in the early nineties when Ramoter and Masson deleted genes known to be involved in generalized environmental stress response (ESR) (Masson and Ramotar, 1996). Previously, it had been thought that such perturbation would then result in wide-spread sensitivity to stress; however, they noted particular deletions corresponded to specific types of stresses. Gasch *et. al.* (Gasch et al., 2000) have since performed a more comprehensive study subjecting yeast to 14 different types of environmental stress (e.g. heat shock, oxidative, etc) and profiling genome-wide changes in gene expression. The Gasch study showed that although there is a "core" of yeast ESR-regulated genes that respond in a characteristic manner to a diverse array of stresses, there are also "physiological themes" that regulate condition and gene-specific expression programs. In total over 900 genes were shown to be activated or repressed in response to environmental stresses. Despite the enormous quantity of data collected open questions remain regarding the commonalities and specificities of stress response, their mechanistic underpinnings, and regulation, and we hypothesized that an underlying physical basis may be observed. Below, we provide an interpretation of these environmental stress response (ESR) cross patterns within the context of three general principles established by Gasch et al in their landmark set of experiments (Gasch et al., 2000). (1) Specific sets of genes have coordinated behaviors in response to disparate typs of stress, (2) isozymes are often involved in different types of stress-response, and (3) different treatments resulting in the same type of stress often work through the same stress response pathway (Hohmann and Mager, 2003) (see Table 6.3.

ESR-regulated proteins exhibit specific drug feature-sensitivities

Analogous to "physiological themes," there are several cases where ESR-regulated genes including TOR1, CYC7, GPM2, and SSA3 also exhibited strong preferences for a particular structural determinant. Further, each of these proteins is known to play a stereotypical role in one or more stress responses. As an example, TOR1 (protein of rapamycin) is a

Gasch Results	**Our Results**	**Implications**	**Example**
Specific sets of genes have coordinated behaviors in response to disparate types of stress (ESR-regulated).	ESR-regulated genes exhibit **specific** DP-sensitivities.	DP-sensitivities may allow the tracking of an underlying molecular reasoning for similarities and dissimilarities in ESR. *That is, they may form Equivalence Groups with Env Stress.*	TOR1, CYC7, GPM2, SSA3 stereotypical stress response and DP-sens.
Isozymes are often involved in **different** types of stress response.	Isozymes exhibit **different** DP-sensitivities.	Subtle differences in amino acids *may render one isozyme more suitable than another* under a given set of conditions. DP-sensitivities may allow tracking of underlying biochemical differences.	GGT1 exhibited charge sensitive, but GTT2 showed no specificity
Different treatments resulting in the same type of stress often work through the **same** stress response pathway.	Different treatments resulting in the same type of stress exhibit **different** DP-sensitivities.	Different mechanisms of similar ESR can be detected vi differential DP-sensitivities.	H202 and meniadione have different DP-sensitivities and subtle differences in their response to osmotic stress have since been identified.

Table 6.3: Implications of complex cross patterns for environmental stress response

kinase that controls response to amino acid starvation, and we show that it also exhibits a sensitivity to a compound's charge. Similarly, SSA3 is involved in protein unfolding and heat shock response and is MlogP-sensitive. Thus, one intriguing possibility is that structural determinants could form equivalence groups with environmental stresses. That is, one can use the connection with specific drug features to track an underlying molecular reasoning for similarities and conversely dissimilarities in stress response. Here, we have only begun to identify such relationships, and future work will be required to unravel the mechanistic reasoning underlying the stress response specificity and structural determinant sensitivity.

Isozymes exhibit different DP-sensitivities

Isozymes have different amino acid sequences, but they perform the same chemical reaction. Interestingly, one of the hallmarks of the general environmental stress response (ESR) in yeast is differential regulation of isozymes (Hohmann and Mager, 2003). That is, only one of a pair of isozymes has a known role in a stress response, or both may have roles but each under a different set of conditions. As an example, GPD1 is important in responding to osmotic stress; whereas, GPD2 seems to be activated under aerobic conditions. The mechanism governing this differential regulation remains unknown. One intriguing possibility is that the isozyme's subtly different amino acid sequence results in dissimilar biochemical properties that may render one isozyme more suitable than another under a given set of conditions. In keeping with this theory, we do observe differential drug property sensitivities between several pairs of isozymes. As an example, the non-ESR regulated glutathione transferase, GTT2, exhibits charge-sensitivity, but GTT1 showed extremely low variance and there was almost no-specificity in its response to drug-treatments. This suggests that differential drug sensitivity may prove useful in tracking these underlying biochemical differences and how they impact stress response regulation.

Different mechanisms of similar ESR can be detected via FD-sensitivities

It has been shown that different perturbations can sometimes induce the same type of stress (O'Rourke et al., 2002). As an example, oxidative stress can be triggered in yeast through the application of either hydrogen peroxide or meniadione among others (Hampsey, 1997). Further, despite the different reactive oxygen species (ROS) generated by each of these oxidants, both menadione and hydrogen peroxide seemed to show almost identical expression profiles suggesting a remarkably similar response despite the disparate properties of the ROS generated (Gasch et al., 2000). We identified a cross pattern between the drug property MlogP and hydrogen peroxide treatment; however, we found no significant cross pattern between the MlogP and the meniadione profile. Interestingly, differences in response were identified among hydrogen peroxide, menadione, and two other types of oxidants were reported in *S. pombe* Mutoh et al. (2005). One potential explanation then is that differences in structural parameter sensitivities may reflect the specific requirements in responding to each of the different types of reactive species generated. This suggests that cross patterns may prove useful in teasing apart differences between closely related stress responses.

6.4.2 Guilt-by-association to predict function or mechanism of compound action

Akin to building a compendium of a protein's response to small molecules, the cross patterns described can be aggregated to generate a profile of a protein's sensitivity to structural parameters across a number of different small molecule applications (drug property-sensitivity profiles). There are numerous ways to characterize small molecules; nevertheless, using just 6 well-characterized molecular descriptors, we see evidence that proteins whose sensitivity profiles overlapped were also functionally similar. Thus, it is likely that by applying traditional "guilt-by-association" rules using these profiles (Pellegrini et al., 1999), we can generate hypotheses about the role of uncharacterized

proteins.

A large number of the DP-sensitive proteins are of unknown function, and one of the most intriguing is YCR101C, which is both molecular weight and aromatic-bond sensitive. Five proteins had a similar DP-sensitivity profile to YCR101C including GUP1, which localizes to the membrane and is suspected of being a glycerol transporter and a known GPI remodeler. The shared DP-sensitivities also mapped to osmotic stress response and a proclivity to be localized to the vacuoles. The physiological role of the vacuole during osmotic stress is unclear; however, it is known that phosphoinositides quickly accumulate stimulating actin patch-formation and that disruption of this pathway causes abnormal vacuole morphology. Based on these observations, we would suggest that YCR101C plays some role in cytoskeletal reorganization in the vacuole. Indeed, YCR101C also has several known interacting partners including GPI1, a membrane protein involved in N-GlycNac synthesis and ARP1, a cytoskeletal element. Finally, in MIPS, YCR101C is reportedly similar to a vacuole-localized glycoprotein making the inference from the sensitivity profile seem likely.

Although the number of MW-sensitive proteins that localized to the mitochondria was too small to pick up a general trend, we did observe that the mitochondrial protein YAL008W was markedly more affected by smaller compounds ($P < .01$) than larger ones (Figure 6.4). YAL008W's function is unknown although interactions with two other mitochondrial proteins: YML086C, which plays an important role in protecting against oxidative stress, and YPL186C, which has some as yet uncharacterized role in ubiquitination, have been identified in large-scale screens. All three of these mitochondrial proteins are localized to the outer mitochondrial membrane. These other interactions are particularly interesting given that MW-sensitive proteins also have a tendency to be affected by hyper-osmotic shock.

Although we used only 6 molecular descriptors, including additional features of these small molecules can allow structure-based profiles to be built. Such profiles can be used both to infer new knowledge regarding mechanism of action, specificity of response, and

through a nearest neighbor approach the function of currently uncharacterized genes and the discovery of novel functions of pre-existing ones

6.4.3 Connector Matrix Interpretation

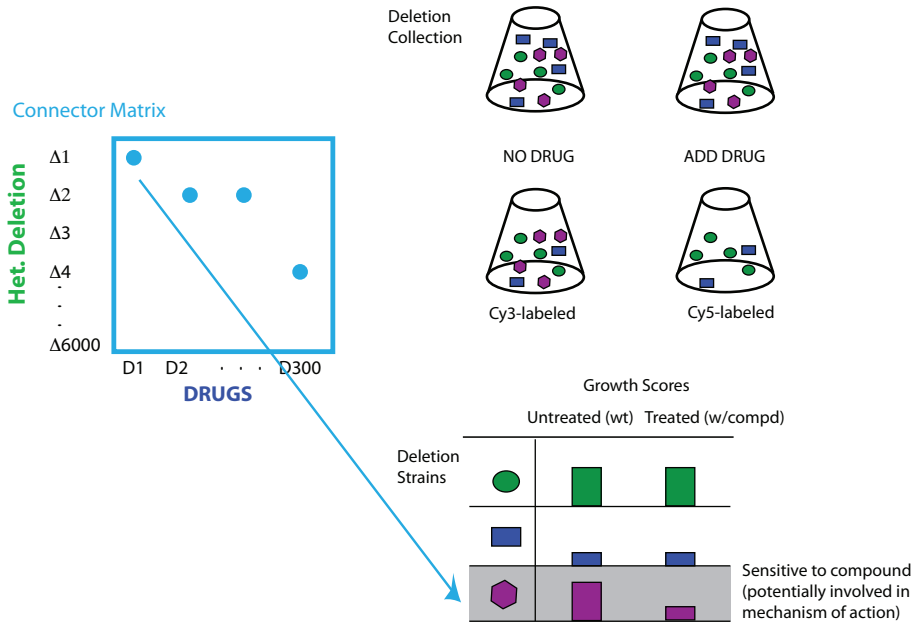


Figure 6.6: Schematic of a chemogenomics profiling experiment, see text for details

Connector matrices can range from the simple and straightforward to subtler connections. In order to use the Hillenmeyer dataset (Hillenmeyer et al., 2008) as a connector matrix, we have made several simplifying assumptions in the mapping of drugs to their associated target proteins. In particular, we have treated ORFs as being synonymous with their protein product and for clarity used the terms protein and protein property as opposed to ORF and ORF property. However, the actual measurements in the drug-response datasets are not based on proteins. Precisely, rows of this matrix represent a heterozygous deletion strain and each column is a separate small molecule treatment. Each element of the matrix is the fitness defect of that particular strain treated with a particular compound where fitness defect is quantified as the difference between growth under wild type conditions and in the presence of the compound. Hillenmeyer *et. al* (Hillenmeyer et al., 2008) subjected the heterozygous yeast deletion collection (only one

copy of ORF is deleted) and homozygous collection (both copies of ORF deleted) to over 300 structurally diverse compounds and quantified the difference between growth under wild type conditions and in the presence of the compounds. We used only treatments of the heterozygous collection as fitness defects of heterozygotes have been found to serve as an indicator of the ORF's involvement in the compound's mechanism of action (Lum et al., 2004; Giaever et al., 1999, 2004; Hillenmeyer et al., 2008). The simplest interpretation stems from a phenomenon termed the haploinsufficiency effect (Giaever et al., 1999; Deutschbauer et al., 2005). As heterozygous deletions (only one copy remaining) natively produce less gene product than the wild type strain, it is expected that adding a compound targeted to that gene product should result in a more severe loss of function (in this case measured by growth) than in the wild type (Deutschbauer et al., 2005). However, it is important to note that fitness defects alone do not provide sufficient evidence to infer whether the mechanism of action is physical binding between the small molecule and target protein. Target disruption could also occur through more downstream processes, or the target itself could be lipid, DNA, or RNA rather than a protein. We performed extensive filtering on both the drug and protein set to remove drugs that acted on too few or too many proteins and correspondingly proteins that were affected by large numbers of drugs (see Materials and Methods) to reduce problems arising from non-specificity of drug-protein interactions.

6.4.4 Generality of ITeR

The amount of available multidimensional data (unstackable features) will continue to grow. A number of current datasets could be formulated in terms of connector matrices and thus be amenable to an ITeR type of approach. The complexity of the connector matrix can range from the straightforward, such as mapping transcription factors to their binding sites, to the more complicated derivation, such as chemogenomics datasets to tissue or tumor-specific expression surveys. Whereas, direct integration only allows for identification of tissue-specific or tumor-specific expression, ITeR can connect such tissue properties potentially to sets of gene properties or metabolites. One can even potentially

integrate features of disease or clinical state alongside metagenomics data cataloguing a particular person's microbial community structure perhaps uncovering additional instances where microbes are responsible for things like chronic inflammatory response or other microbial-induced disease progression.

6.5 Conclusion

At the moment, yeast represents a special case in terms of the range of available system-wide datasets; however, yeast is a harbinger for other systems. Technological and computational advances are leading to a dramatic increase in system-wide datasets for many different model organisms. The unprecedented scale and diversity of these datasets present both opportunities for new discoveries and interesting computational challenges. Straightforward integration, as currently done in genomics, does not provide enough flexibility when the dataset can no longer be indexed on a gene or protein or even a single class of variable. We have introduced a method to discover transitive relationships between differently indexed metadata and have used this formalism to identify cross patterns connecting small molecule descriptor sensitivities to disparate types of systems-wide features. Further, we showed that this type of integration can reveal novel and non-obvious connections between many different and not necessarily gene-centric types of data. In a broader context, to fully leverage the coming deluge of systems-wide datasets will require the development of new types of spanning techniques as more and more model organisms join the ranks of yeast in terms of both quantity and diversity of data. Mining such complexity requires a robust infrastructure and new computational models; many rich and exciting discoveries remain.

6.6 Materials and Methods

6.6.1 Preprocessing ORFs

Yeast strains with defects in transport machinery, lipid permeability, and drug efflux pumps, etc. (Bauer et al., 1999) will be sensitive to a wide range of drugs; however, their effect is non-specific and any connections we observe with this class is likely to be spurious (DeRisi et al., 2000). Thus, we first filtered multi-drug resistant ORFs using identical criteria as in (Hillenmeyer et al., 2008). Analogously,if the variance of a single protein's growth scores across all small molecule perturbations is too low, one would only be correlating noise. Therefore, we computed the variance of growth scores for each ORF and selected those with a variance greater than 1.5. There were 1194 ORFs remaining after this two-step filtering. After removal of ORFs not in the protein-feature datasets (see below), the final set is 1170. Finally, there were a few cases where the ORF grew better in the presence of a particular drug suggesting resistance; however, we set the value of these treatments to 0 effectively excluding this effect from the analysis.

6.6.2 Preprocessing Small Molecules and Calculating Molecular Descriptors

Hillenmeyer et al. tested 291 unique compounds on the heterozygous deletion collection under a number of different concentrations (Hillenmeyer et al., 2008) . As the concentration of a drug increases, its effects can become less specific as it approaches toxicity. Since we are most interested the specific response, we selected profiles generated using the minimum drug concentration. We then converted the small molecules to text strings using the SMILES format (James et al., 2005). Although many different properties of small molecules can be calculated from SMILES, we chose 6 for their interpretability, diversity of measurements, and wide-spread use in computational chemistry (Leach and Gillet, 2003); the molecular weight, charge, number of aromatic bonds, number of aromatic rings, hydrophilicity, and MlogP of each compound was calculated (Tetko et al., 2005). Only compounds with no missing values were kept. Thus, the final dimensions are 281x1170

and 6x281, respectively. However, this analysis could easily be expanded to include other parameters.

6.6.3 Significance Testing

As shown formally in the text, we performed two rounds of significance testing. First, we calculated a sensitivity score for each protein and similarly, we computed a sensitivity score for each protein property. For the protein sensitivity, the drugs were first partitioned into two classes based on the median (e.g. high and low molecular weight drugs). Then the sensitivity score S is then calculated as follows:

$$S = \frac{\hat{X}_H - \hat{X}_L}{S_{\hat{X}_H - \hat{X}_L}}$$

where $\hat{X}_H$ is the mean growth score for a protein after all treatments with drug labeled as high for the particular feature and $S_{\hat{X}H}$ is the standard error. Similarly, $\hat{X}_L$ is the mean growth score for a protein after all treatments with drug labeled as low for the particular feature and $S_{\hat{X}L}$ is the standard error. Since the protein propertiess encompassed both categorical and continuous measurements, a slight modification was required to calculate a sensitivity score for the categorical variables. These were transformed into a series of binary variables (e.g. localization to nucleus, cytoplasm, etc). We then used these new features to calculate the expected frequency from the background distribution and compared this expected frequency with the distribution for each of the six drug property sensitivity classes. To determine the significance of enrichment, we used the hypergeometric distribution. Any number of tests could be used to compute significance for the continuous features; here we use the Welch's t-test or Wilcoxon for bimodal distributions.

6.6.4 Protein-features

We used seven different types of systems-data. Physicochemical properties of the ORFs were obtained from SGD including molecular weight, isoelectric point, protein length,

GRAVY (hydropathicity index), and aromaticity (Nash et al., 2007) as were the gene composition features as follows: codon adaptation index (CAI) and frequency of optimal codons (FOP) and GO categories (Ashburner et al., 2000). The localization was taken from (Huh et al., 2003). Compartments were aggregated into nine categories, and binarized as described above. We used two types of networks: protein-protein interactions and gene regulatory (Stark et al., 2006) (genetic interaction and phosphorylome (Ptacek et al., 2005) had too little overlap with the protein set to be able to determine significance; results not shown). All topological statistics (degree, clustering coefficient, betweenness, eccentricity, shortest path) were computed for each node in the network using tYNA (Yip et al., 2006). The environmental stress response data were from (Gasch et al., 2000). If there were fewer than 5 missing values in a given row, they were imputed by computing the row mean. If there were more than 5 missing values, the ORF was deleted from the analysis.

Chapter 7

Future Outlook: Mining Biological Complexity

Sheer brute force sequencing is allowing questions to be asked at a scale unimaginable even several years ago. With so-called second generation sequencing technologies, sequencing capacity will only continue to increase (Mardis, 2008; Salzberg, 2008). Such capabilities have allowed groups to explore new applications for sequencing including identifying structural variants (Korbel, 2007), profiling transcription (Nagalakshmi, 2008; Sultan, 2008), mapping transcription factor binding sites (Robertson, 2007; Rozowsky, 2008), and cataloguing microbial and viral genetic content of entire ecosystems (Dinsdale, 2008). The theme reinforced in all of these studies is that from simple components emerges unimaginable complexity. The ability to decipher it requires a deep understanding of the context and combinatorial interactions from which it emerges. Mining such complexity offers rewards commensurate with the difficulty of the challenge; for as we continue to develop the capacity to read the genetic code, we also acquire more tools to edit pieces of it and even write some it ourselves.

Bibliography

Abbott, A. 2005. Medics braced for fresh superbug. Nature 436:758.

Alegado, R. A., M. C. Campbell, W. C. Chen, S. S. Slutz, and M. W. Tan. 2003. Characterization of mediators of microbial virulence and innate immunity using the Caenorhabditis elegans host-pathogen model. Cell Microbiol 5:435–44.

Allen, E. E. and J. F. Banfield. 2005. Community genomics in microbial ecology and evolution. Nat Rev Micro 3:489.

Allen, J. E., M. Pertea, and S. L. Salzberg. 2004. Computational Gene Prediction Using Multiple Sources of Evidence. Genome Res. 14:142–148.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J Mol Biol 215:403–10.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–402.

Ashburner, M., 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, and G. M. Rubin. 2000. Gene Ontology: tool for the unification of biology. Nature Genetics 25:25.

Averhoff, B. and A. Friedrich. 2003. Type IV pili-related natural transformation systems: DNA transport in mesophilic and thermophilic bacteria. Arch Microbiol 180:385–93.

Avery, L. and J. H. Thomas. 1997. Feeding and Defecation. Pages 679–716 *in* C. elegans II (D. Riddle, T. Blumenthal, B. J. Meyer, and J. Priess, eds.). Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bailey, T. L. and M. Gribskov. 1998. Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14:48–54.

Barbe, V., D. Vallenet, N. Fonknechten, A. Kreimeyer, S. Oztas, L. Labarre, S. Cruveiller, C. Robert, S. Duprat, P. Wincker, L. N. Ornston, J. Weissenbach, P. Marliere, G. N. Cohen, and C. Medigue. 2004. Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. Nucleic Acids Res 32:5766–79.

Barthelmes, J., C. Ebeling, A. Chang, I. Schomburg, and D. Schomburg. 2007. BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Res 35:D511–4.

Bauer, B. E., H. Wolfger, and K. Kuchler. 1999. Inventory and function of yeast ABC proteins: about sex, stress, pleiotropic drug and heavy metal resistance. Biochimica et Biophysica Acta (BBA) - Biomembranes 1461:217–236.

Baumann, P., M. Doudoroff, and R. Y. Stanier. 1968a. A study of the Moraxella group. II. Oxidative-negative species (genus Acinetobacter). J Bacteriol 95:1520–41.

Baumann, P., M. Doudoroff, and R. Y. Stanier. 1968b. Study of the Moraxella group. I. Genus Moraxella and the Neisseria catarrhalis group. J Bacteriol 95:58–73.

Begley, T. J., A. S. Rosenbach, T. Ideker, and L. D. Samson. 2004. Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. Mol Cell 16:117–25.

Ben-Hur, A. and W. S. Noble. 2005. Kernel methods for predicting protein-protein interactions. Bioinformatics 21:i38–46.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57:289–300.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler. 2008. GenBank. Nucleic Acids Res 36:D25–30.

Bergmann, S., J. Ihmels, and N. Barkai. 2003. Iterative signature algorithm for the analysis of large-scale gene expression data. Physical Review E 67:031902.

Bergogne-Berezin, E. and K. J. Towner. 1996. Acinetobacter spp. as nosocomial pathogens: microbiological, clinical, and epidemiological features. Clin Microbiol Rev 9:148–65.

Bernards, A. T., H. I. Harinck, L. Dijkshoorn, T. J. van der Reijden, and P. J. van den Broek. 2004. Persistent Acinetobacter baumannii? Look inside your medical equipment. Infect Control Hosp Epidemiol 25:1002–4.

Birse, C. E., M. Y. Irwin, W. A. Fonzi, and P. S. Sypherd. 1993. Cloning and characterization of ECE1, a gene expressed in association with cell elongation of the dimorphic pathogen Candida albicans. Infect Immun 61:3648–55.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–93.

Borga, M. 1998. Learning Multidimensional Signal Processing. Ph.D. thesis Linkoping University.

Borneman, A. R., J. A. Leigh-Bell, H. Yu, P. Bertone, M. Gerstein, and M. Snyder. 2006. Target hub proteins serve as master regulators of development in yeast. Genes Dev 20:435–48.

Borneman, A. R., Z. Zhang, J. S. Rozowsky, M. Seringhaus, M. Gerstein, and M. Snyder. 2007. Divergence of Transcription Factor Binding Sites Across Related Yeast Species. Science 317:815–819.

Boyer, T., J. Antonov, H. Garcia, D. Johnson, R. Locarnini, A. Mishonov, M. Pitcher, O. Baranova, and I. Smolyar. 2006. World Ocean Database 2005. Tech. rep. US Government Printing Office.

Boyle, E. I., S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. 2004. GO:TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20:3710–5.

Cardelli, J. 2001. Phagocytosis and macropinocytosis in Dictyostelium: phosphoinositide-based processes, biochemically distinct. Traffic 2:311–20.

Carignan, V. and M. A. Villard. 2002. Selecting indicator species to monitor ecological integrity: a review. Environ Monit Assess 78:45–61.

Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill. 2005. ACT: the Artemis Comparison Tool. Bioinformatics 21:3422–3.

Cawley, S., S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116:499–509.

CDC. 2004. Acinetobacter baumannii infections among patients at military medical facilities treating injured U.S. service members, 2002-2004. Morbidity and Mortality Weekly Report 53:1063–66.

Chastre, J. and J. L. Trouillet. 2000. Problem pathogens (Pseudomonas aeruginosa and Acinetobacter). Semin Respir Infect 15:287–98.

Choi, C. H., E. Y. Lee, Y. C. Lee, T. I. Park, H. J. Kim, S. H. Hyun, S. A. Kim, S. K. Lee, and J. C. Lee. 2005. Outer membrane protein 38 of Acinetobacter baumannii localizes to the mitochondria and induces apoptosis of epithelial cells. Cell Microbiol 7:1127–38.

Chou, S., S. Lane, and H. Liu. 2006. Regulation of mating and filamentation genes by two distinct Ste12 complexes in Saccharomyces cerevisiae. Mol Cell Biol 26:4794–805.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71–6.

Coleman, J. R., D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, and S. Mueller. 2008. Virus Attenuation by Genome-Scale Changes in Codon Pair Bias. Science 320:1784–1787.

Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27:4636–41.

DeLong, E. F., C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N. U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm, and D. M. Karl. 2006. Community genomics among stratified microbial assemblages in the ocean's interior. Science 311:496–503.

DeRisi, J., B. van den Hazel, P. Marc, E. Balzi, P. Brown, C. Jacq, and A. Goffeau. 2000. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. FEBS Lett 470:156–60.

Deutschbauer, A. M., D. F. Jaramillo, M. Proctor, J. Kumm, M. E. Hillenmeyer, R. W. Davis, C. Nislow, and G. Giaever. 2005. Mechanisms of Haploinsufficiency Revealed by Genome-Wide Profiling in Yeast. Genetics 169:1915–1925.

Dinsdale, E. A., R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White, and F. Rohwer. 2008. Functional metagenomic profiling of nine biomes. Nature 452:629–32.

Dolan, J. W., C. Kirkman, and S. Fields. 1989. The yeast STE12 protein binds to the DNA sequence mediating pheromone induction. Proc Natl Acad Sci U S A 86:5703–7.

Doolitte, J., RF Kyte. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132.

Dorsey, C. W., M. S. Beglin, and L. A. Actis. 2003. Detection and analysis of iron uptake components expressed by Acinetobacter baumannii clinical isolates. J Clin Microbiol 41:4188–93.

Dorsey, C. W., A. P. Tomaras, and L. A. Actis. 2002. Genetic and phenotypic analysis of Acinetobacter baumannii insertion derivatives generated with a transposome system. Appl Environ Microbiol 68:6353–60.

Dorsey, C. W., A. P. Tomaras, and L. A. Actis. 2006. Sequence and organization of pMAC, an Acinetobacter baumannii plasmid harboring genes involved in organic peroxide resistance. Plasmid 56:112–23.

Dorsey, C. W., A. P. Tomaras, P. L. Connerly, M. E. Tolmasky, J. H. Crosa, and L. A. Actis. 2004. The siderophore-mediated iron acquisition systems of Acinetobacter baumannii ATCC 19606 and Vibrio anguillarum 775 are structurally and functionally related. Microbiology 150:3657–67.

Dufresne, A., M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. Tandeau de Marsac, J. Weissenbach,

P. Wincker, Y. I. Wolf, and W. R. Hess. 2003. Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci U S A 100:10020–5.

Eaton, M. and M. Perlman. 1973. The Non-Singularity of Generalized Sample Covariance Matrices. The Annals of Statistics 1:710–717.

ENCODE. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Flaherty, P., G. Giaever, J. Kumm, M. I. Jordan, and A. P. Arkin. 2005. A latent variable model for chemogenomic profiling. Bioinformatics 21:3286–3293.

Foerstner, K. U., C. von Mering, S. D. Hooper, and P. Bork. 2005. Environments shape the nucleotide composition of genomes. EMBO Rep 6:1208–13.

Forbes, A. 1995. Classification-algorithm evaluation: Five performance measures based on confusion matrices. Journal of Clinical Monitoring and Computing 11:189–206.

Fournier, P. E., D. Vallenet, V. Barbe, S. Audic, H. Ogata, L. Poirel, H. Richet, C. Robert, S. Mangenot, C. Abergel, P. Nordmann, J. Weissenbach, D. Raoult, and J. M. Claverie. 2006. Comparative Genomics of Multidrug Resistance in Acinetobacter baumannii. PLoS Genet 2:e7.

Fraser, A. G. and E. M. Marcotte. 2004. A probabilistic view of gene function. Nat Genet 36:559–564.

Gasch, A. P., P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. 2000. Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Mol. Biol. Cell 11:4241–4257.

Ghedin, E. and J. M. Claverie. 2005. Mimivirus relatives in the Sargasso sea. Virol J 2:62.

Giaever, G., P. Flaherty, J. Kumm, M. Proctor, C. Nislow, D. F. Jaramillo, A. M. Chu, M. I. Jordan, A. P. Arkin, and R. W. Davis. 2004. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. Proc Natl Acad Sci U S A 101:793–8.

Giaever, G., D. D. Shoemaker, T. W. Jones, H. Liang, E. A. Winzeler, A. Astromoff, and R. W. Davis. 1999. Genomic profiling of drug sensitivities via induced haploinsufficiency. Nat Genet 21:278–283.

Gianoulis, T. A., A. Agarwal, M. Snyder, and M. Gerstein. 2009a. Mining Complexity in Systems Biology: Identifying Transitive Relationshipsin Chemogenomics Data. (in preparation) .

Gianoulis, T. A., J. Raes, P. Patel, R. Bjornson, J. O. Korbel, T. Yamada, I. Letunic, A. Paccanaro, L. J. Jensen, M. Snyder, P. Bork, and M. Gerstein. 2009b. Quantifying environmental adaptation of microbial metabolic pathways. PNAS (in press).

Giovannoni, S. J., H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, M. S. Rappe, J. M. Short, J. C. Carrington, and E. J. Mathur. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–5.

Goldberg, S. M., J. Johnson, D. Busam, T. Feldblyum, S. Ferriera, R. Friedman, A. Halpern, H. Khouri, S. A. Kravitz, F. M. Lauro, K. Li, Y. H. Rogers, R. Strausberg, G. Sutton, L. Tallon, T. Thomas, E. Venter, M. Frazier, and J. C. Venter. 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A 103:11240–5.

Goldstein, A. L. and J. H. McCusker. 1999. Three new dominant drug resistance cassettes for gene disruption in Saccharomyces cerevisiae. Yeast 15:1541–53.

Gonzalez, I., S. Dejean, P. G. Martin, and A. Baccini. 2008. CCA: A R Package to Extend Canonical Correlation Analysis. Journal of Statistical Software 22:1–14.

Hampsey, M. 1997. A Review of Phenotypes in Saccharomyces cerevisiae. Yeast 13:1099–1133.

Han, J. and M. Kamber. 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann Series in Data Management Series Morgan Kaugmann Publishers.

Harbison, C. T., D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104.

Harrington, E. D., A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, and P. Bork. 2007. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A 104:13913–8.

Haugen, A., R. Kelley, J. Collins, C. Tucker, C. Deng, C. Afshari, J. M. Brown, T. Ideker, and B. Van Houten. 2004. Integrating phenotypic and expression profiles to map arsenic-response networks. Genome Biology 5:R95–R95.

Hillenmeyer, M. E., E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St.Onge, M. Tyers, D. Koller, R. B. Altman, R. W. Davis, C. Nislow, and G. Giaever. 2008. The Chemical Genomic Portrait of Yeast: Uncovering a Phenotype for All Genes. Science 320:362–365.

Hofreuter, D., J. Tsai, R. O. Watson, V. Novik, B. Altman, M. Benitez, C. Clark, C. Perbost, T. Jarvie, L. Du, and J. Galan. 2006. Unique Features of a Highly Pathogenic Campylobacter jejuni Strain. Infection and Immunity 74:4694–4707.

Hohmann, S. and W. H. Mager. 2003. Yeast Stress Responses.

Hotelling, H. 1936. Relations between two sets of variates. Biometrika 28:321–377.

Huh, W. K., J. V. Falvo, L. C. Gerke, A. S. Carroll, R. W. Howson, J. S. Weissman, and E. K. O'Shea. 2003. Global analysis of protein localization in budding yeast. Nature 425:686–91.

Iyer, V. R., C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409:533–8.

James, C., D. Weininger, and J. Delany. 2005. Daylight Theory Manual.

Jansen, R., H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. 2003. A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science 302:449–453.

Johnson, Z., E. R. Zinser, A. Coe, N. P. McNulty, E. Malcolm, S. Woodward, and S. W. Chisholm. 2006. Niche Partitioning Among Prochlorococcus Ecotypes Along Ocean-Scale Environmental Gradients. Science 311:1737–1740.

Jones, T., N. A. Federspiel, H. Chibana, J. Dungan, S. Kalman, B. B. Magee, G. Newport, Y. R. Thorstenson, N. Agabian, P. T. Magee, R. W. Davis, and S. Scherer. 2004. The diploid genome sequence of Candida albicans. Proc Natl Acad Sci U S A 101:7329–34.

Juni, E. 1978. Genetics and physiology of Acinetobacter. Annu Rev Microbiol 32:349–71.

Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34:D354–7.

Karl, D. M. 2002. Nutrient dynamics in the deep blue sea. Trends Microbiol 10:410–8.

Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol 9:335–43.

Kelley, R. and T. Ideker. 2005. Systematic interpretation of genetic interactions using protein networks. Nat Biotech 23:561–566.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423:241–54.

Kim, P. M., J. O. Korbel, and M. B. Gerstein. 2007. Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context. PNAS 104:20274.

Kim, P. M., A. Sboner, Y. Xia, and M. Gerstein. 2008. The role of disorder in interaction networks: a structural analysis. Mol Syst Biol 4.

King, O. D., R. F. Foulger, S. S. Dwight, and F. P. Roth. 2003. Predicting Gene Function From Patterns of Annotation. Genome Res. 13:896–904.

Korbel, J. O., A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, B. E. Taillon, Z. Chen, A. Tanzer, A. C. Saunders, J. Chi, F. Yang, N. P. Carter, M. E. Hurles, S. M. Weissman, T. T. Harkins, M. B. Gerstein, M. Egholm, and M. Snyder. 2007. Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–6.

Koulenti, D. and J. Rello. 2006. Gram-negative bacterial pneumonia: aetiology and management. Curr Opin Pulm Med 12:198–204.

Krogh, A., B. r. Larsson, G. von Heijne, and E. L. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. Journal of Molecular Biology 305:567–580.

Kumar, A., S. Vidan, and M. Snyder. 2002. Insertional mutagenesis: transposon-insertion libraries as mutagens in yeast. Methods Enzymol 350:219–29.

La Scola, B. and D. Raoult. 2004. Acinetobacter baumannii in human body louse. Emerg Infect Dis 10:1671–3.

Lan, N., R. Jansen, and M. Gerstein. 2002. Toward a systematic definition of protein function that scales to the genome level: Defining function in terms of interactions. IEEE 90.

Lanckriet, G. R. G., T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. 2004. A statistical framework for genomic data fusion. Bioinformatics 20:2626–2635.

Lane, S., C. Birse, S. Zhou, R. Matson, and H. Liu. 2001. DNA array studies demonstrate convergent regulation of virulence factors by Cph1, Cph2, and Efg1 in Candida albicans. J Biol Chem .

Leach, A. and V. Gillet. 2003. An Introduction to Chemoinformatics. Kluwer Academic Publishers.

Lee, J., H. Yun, A. Feist, B. Palsson, and S. Lee. 2008. Genome-scale reconstruction and in silico analysis of the Clostridium acetobutylicum ATCC 824 metabolic network. Applied Microbiology and Biotechnology .

Lee, K. L., H. R. Buckley, and C. C. Campbell. 1975. An amino acid liquid synthetic medium for the development of mycelial and yeast forms of Candida Albicans. Sabouraudia 13:148–53.

Letunic, I., T. Yamada, M. Kanehisa, and P. Bork. 2008. iPath: interactive exploration of biochemical pathways and networks. Trends Biochem Sci 33:101–3.

Liles, M. R., B. F. Manske, S. B. Bintrim, J. Handelsman, and R. M. Goodman. 2003. A Census of rRNA Genes and Linked Genomic Sequences within a Soil Metagenomic Library. Appl. Environ. Microbiol. 69:2684–2691.

Liolios, K., N. Tavernarakis, P. Hugenholtz, and N. C. Kyrpides. 2006. The Genomes On

Line Database (GOLD) v.2: a monitor of genome projects worldwide. Nucleic Acids Res 34:D332–4.

Liu, X. S., D. L. Brutlag, and J. S. Liu. 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20:835–9.

Longtine, M. S., r. McKenzie, A., D. J. Demarini, N. G. Shah, A. Wach, A. Brachat, P. Philippsen, and J. R. Pringle. 1998. Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. Yeast 14:953–61.

Lum, P. Y., C. D. Armour, S. B. Stepaniants, G. Cavet, M. K. Wolf, J. S. Butler, J. C. Hinshaw, P. Garnier, G. D. Prestwich, A. Leonardson, P. Garrett-Engele, C. M. Rush, M. Bard, G. Schimmack, J. W. Phillips, C. J. Roberts, and D. D. Shoemaker. 2004. Discovering Modes of Action for Therapeutic Compounds Using a Genome-Wide Screen of Yeast Heterozygotes. Cell 116:121–137.

Luscombe, N. M., M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. B. Gerstein. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. Nature 431:308–312.

Luscombe, N. M., D. Greenbaum, and M. Gerstein. 2001. What is bioinformatics? A proposed definition and overview of the field. Methods of Information in Medicine 40:346–58.

Mahenthiralingam, E., A. Baldwin, P. Drevinek, E. Vanlaere, P. Vandamme, J. J. LiPuma, and C. G. Dowson. 2006. Multilocus sequence typing breathes life into a microbial metagenome. PLoS ONE 1:e17.

Majoros, W. H., M. Pertea, and S. L. Salzberg. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20:2878–9.

Marcotte, E. M., M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. 1999. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science 285:751–753.

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

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–80.

Martinez, D., R. M. Berka, B. Henrissat, M. Saloheimo, M. Arvas, S. E. Baker, J. Chapman, O. Chertkov, P. M. Coutinho, D. Cullen, E. G. J. Danchin, I. V. Grigoriev, P. Harris, M. Jackson, C. P. Kubicek, C. S. Han, I. Ho, L. F. Larrondo, A. L. de Leon, J. K. Magnuson, S. Merino, M. Misra, B. Nelson, N. Putnam, B. Robbertse, A. A. Salamov, M. Schmoll, A. Terry, N. Thayer, A. Westerholm-Parvinen, C. L. Schoch, J. Yao, R. Barbote, M. A. Nelson, C. Detter, D. Bruce, C. R. Kuske, G. Xie, P. Richardson, D. S. Rokhsar, S. M. Lucas, E. M. Rubin, N. Dunn-Coleman, M. Ward, and T. S. Brettin. 2008. Genome sequencing and analysis of the biomass-degrading fungus Trichoderma reesei (syn. Hypocrea jecorina). Nat Biotech 26:553–560.

Mary, I., G. A. Tarran, P. E. Warwick, M. J. Terry, D. J. Scanlan, P. H. Burkill, and M. V.

Zubkov. 2008. Light enhanced amino acid uptake by dominant bacterioplankton groups in surface waters of the Atlantic Ocean. FEMS Microbiol Ecol 63:36–45.

Masson, J. and D. Ramotar. 1996. The Saccharomyces cerevisiae IMP2 gene encodes a transcriptional activator that mediates protection against DNA damage caused by bleomycin and other oxidants. Mol. Cell. Biol. 16:2091–2100.

Mazel, D. and P. Marliere. 1989. Adaptive eradication of methionine and cysteine from cyanobacterial light-harvesting proteins. Nature 341:245–8.

Monier, A., J. M. Claverie, and H. Ogata. 2008. Taxonomic distribution of large DNA viruses in the sea. Genome Biol 9:R106.

Morgan-Kiss, R. M., J. C. Priscu, T. Pocock, L. Gudynaite-Savitch, and N. P. Huner. 2006. Adaptation and acclimation of photosynthetic microorganisms to permanently cold environments. Microbiol Mol Biol Rev 70:222–52.

Morozova, O. and M. A. Marra. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–264.

Mutoh, N., M. Kawabata, and S. Kitajima. 2005. Effects of Four Oxidants, Menadione, 1-Chloro-2,4-Dinitrobenzene, Hydrogen Peroxide and Cumene Hydroperoxide, on Fission Yeast Schizosaccharmoyces pombe. J Biochem 138:797–804.

Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–9.

Nash, R., S. Weng, B. Hitz, R. Balakrishnan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, M. S. Livstone, R. Oughtred, J. Park, M. Skrzypek, C. L. Theesfeld, G. Binkley, Q. Dong, C. Lane, S. Miyasato, A. Sethuraman, M. Schroeder, K. Dolinski, D. Botstein, and J. M. Cherry. 2007.

Expanded protein information at SGD: new pages and proteome browser. Nucl. Acids Res. 35:D468–471.

Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser. 1999a. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329.

Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser. 1999b. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. Nature 399:323–329.

Olivera-Severo, D., G. E. Wassermann, and C. R. Carlini. 2006. Ureases display biological effects independent of enzymatic activity: is there a connection to diseases caused by urease-producing bacteria? Braz J Med Biol Res 39:851–61.

O'Rourke, S. M., I. Herskowitz, and E. K. O'Shea. 2002. Yeast go the whole HOG for the hyperosmotic response. Trends in Genetics 18:405–412.

Palenik, B., B. Brahamsha, F. W. Larimer, M. Land, L. Hauser, P. Chain, J. Lamerdin, W. Regala, E. E. Allen, J. McCarren, I. Paulsen, A. Dufresne, F. Partensky, E. A. Webb, and J. Waterbury. 2003. The genome of a motile marine Synechococcus. Nature 424:1037–42.

Parsons, A. B., R. L. Brost, H. Ding, Z. Li, C. Zhang, B. Sheikh, G. W. Brown, P. M. Kane, T. R. Hughes, and C. Boone. 2004. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. Nat Biotech 22:62–69.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. S. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. PNAS 96:4285–4288.

Pevzner, P. A. 2004. Educating biologists in the 21st century: bioinformatics scientists versus bioinformatics technicians. Bioinformatics 20:2159–61.

Piechaud, M. and L. Second. 1951. [Studies of 26 strains of Moraxella Iwoffi.]. Ann Inst Pasteur (Paris) 80:97–9.

Piskur, J. and R. B. Langkjaer. 2004. Yeast genome sequencing: the power of comparative genomics. Mol Microbiol 53:381–9.

Platypus. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453:175–183.

Pop, M. and S. L. Salzberg. 2008. Bioinformatics challenges of new sequencing technology. Trends in Genetics 24:142–149.

Prinz, S., I. Avila-Campillo, C. Aldridge, A. Srinivasan, K. Dimitrov, A. Siegel, and T. Galitski. 2004. Control of yeast filamentous-form growth by modules in an integrated molecular network. Genome Res 14:380–90.

Ptacek, J., G. Devgan, G. Michaud, H. Zhu, X. Zhu, J. Fasolo, H. Guo, G. Jona, A. Breitkreutz, R. Sopko, R. R. McCartney, M. C. Schmidt, N. Rachidi, S.-J. Lee, A. S. Mah, L. Meng, M. J. R. Stark, D. F. Stern, C. D. Virgilio, M. Tyers, B. Andrews, M. Gerstein, B. Schweitzer, P. F. Predki, and M. Snyder. 2005. Global analysis of protein phosphorylation in yeast. Nature 438:679–684.

Pukatzki, S., R. H. Kessin, and J. J. Mekalanos. 2002. The human pathogen Pseudomonas aeruginosa utilizes conserved virulence pathways to infect the social amoeba Dictyostelium discoideum. Proc Natl Acad Sci U S A 99:3159–64.

Raes, J., J. O. Korbel, M. J. Lercher, C. von Mering, and P. Bork. 2007. Prediction of effective genome size in metagenomic samples. Genome Biol 8:R10.

Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306–9.

Riley, J., R. Butler, D. Ogilvie, R. Finniear, D. Jenner, S. Powell, R. Anand, J. C. Smith, and A. F. Markham. 1990. A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic Acids Res 18:2887–90.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–7.

Rodriguez-Brito, B., F. Rohwer, and R. A. Edwards. 2006. An application of statistics to comparative metagenomics. BMC Bioinformatics 7:162.

Rozowsky, J. S. 2009. Peak-Seq. Nat Biotech (in press).

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt,

E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5:e77.

Salzberg, S., A. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated Markov models. Nucl. Acids Res. 26:544–548.

Sanchez-Martinez, C. and J. Perez-Martin. 2001. Dimorphism in fungal pathogens: Candida albicans and Ustilago maydis– similar inputs, different outputs. Curr Opin Microbiol 4:214–21.

Sanna, S., A. U. Jackson, R. Nagaraja, C. J. Willer, W. M. Chen, L. L. Bonnycastle, H. Shen, N. Timpson, G. Lettre, G. Usala, P. S. Chines, H. M. Stringham, L. J. Scott, M. Dei, S. Lai, G. Albai, L. Crisponi, S. Naitza, K. F. Doheny, E. W. Pugh, Y. Ben-Shlomo, S. Ebrahim, D. A. Lawlor, R. N. Bergman, R. M. Watanabe, M. Uda, J. Tuomilehto, J. Coresh, J. N. Hirschhorn, A. R. Shuldiner, D. Schlessinger, F. S. Collins, G. Davey Smith, E. Boerwinkle, A. Cao, M. Boehnke, G. R. Abecasis, and K. L. Mohlke. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198–203.

Schattner, P., A. N. Brooks, and T. M. Lowe. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33:W686–9.

Schloss, P. D. and J. Handelsman. 2008. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. BMC Bioinformatics 9:34.

Seringhaus, M., A. Paccanaro, A. Borneman, M. Snyder, and M. Gerstein. 2006. Predicting essential genes in fungal genomes. Genome Res. Page gr.5144106.

Seshadri, R., S. A. Kravitz, L. Smarr, P. Gilna, and M. Frazier. 2007. CAMERA: a community resource for metagenomics. PLoS Biol 5:e75.

Smith, M. G., S. G. Des Etages, and M. Snyder. 2004. Microbial synergy via an ethanol-triggered pathway. Mol Cell Biol 24:3874–84.

Smith, M. G., T. A. Gianoulis, S. Pukatzki, J. J. Mekalanos, L. N. Ornston, M. Gerstein, and M. Snyder. 2007. New insights into Acinetobacter baumannii pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. Genes Dev 21:601–14.

Sogin, M. L., H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103:12115–20.

Sonnhammer, E. L., S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26:320–2.

Stanier, R. Y., N. J. Palleroni, and M. Doudoroff. 1966. The aerobic pseudomonads: a taxonomic study. J Gen Microbiol 43:159–271.

Stark, C., B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. 2006. BioGRID: a general repository for interaction datasets. Nucl. Acids Res. 34:D535–539.

Stocker, R., J. R. Seymour, A. Samadani, D. E. Hunt, and M. F. Polz. 2008. Rapid chemotactic response enables marine bacteria to exploit ephemeral microscale nutrient patches. Proc Natl Acad Sci U S A 105:4209–14.

Sulston, J. and J. Hodgkin. 1988. Methods. Pages 587–606 *in* The Nematode Caenorhabditis elegans (W. B. Wood, ed.). Cold Spring Harbor Press, Cold Spring Harbor, New York.

Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas,

M. Vingron, H. Lehrach, and M. L. Yaspo. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–60.

Sussman, M. 1987. Cultivation and synchronous morphogenesis of Dictyostelium under controlled experimental conditions. Methods Cell Biol 28:9–29.

Suzek, B. E., M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg. 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics 17:1123–30.

Tasan, M., W. Tian, D. Hill, F. Gibbons, J. Blake, and F. Roth. 2008. An en masse phenotype and function prediction system for Mus musculus. Genome Biology 9:S8.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. Science 278:631–7.

Tetko, I., J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V. Palyulin, E. Radchenko, N. Zefirov, A. Makarenko, V. Tanchuk, and V. Prokopenko. 2005. Virtual Computational Chemistry Laboratory – Design and Description. Journal of Computer-Aided Molecular Design 19:453–463.

Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D. L. Distel, and M. F. Polz. 2005. Genotypic Diversity Within a Natural Coastal Bacterioplankton Population. Science 307:1311–1313.

Todeschini, R. and V. Consonni. 2000. Handbook of Molecular Descriptors. Wiley-VCH, New York.

Tomaras, A. P., C. W. Dorsey, R. E. Edelmann, and L. A. Actis. 2003. Attachment to and biofilm formation on abiotic surfaces by Acinetobacter baumannii: involvement of a novel chaperone-usher pili assembly system. Microbiology 149:3473–84.

Tompa, M., N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23:137–44.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2005. Comparative metagenomics of microbial communities. Science 308:554–7.

Tringe, S. G., T. Zhang, X. Lui, Y. Yu, and e. a. Lee, WH. 2008. The Airborne Metagenome in an Indoor Urban Environment. PLoS ONE 3:e1862.

Troyanskaya, O. G., M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18:1454–61.

Troyanskya, O., K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein. 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). PNAS 100:8348–8353.

Tsong, A. E., B. B. Tuch, H. Li, and A. D. Johnson. 2006. Evolution of alternative transcriptional circuits with identical logic. Nature 443:415–20.

Tsuda, K. and W. S. Noble. 2004. Learning kernels from biological networks by maximizing entropy. Bioinformatics 20:i326–333.

Turnbaugh, P. J. and J. I. Gordon. 2008. An invitation to the marriage of metagenomics and metabolomics. Cell 134:708–13.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon. 2007. The human microbiome project. Nature 449:804–10.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304:66–74.

von Mering, C., P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. 2007a. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science 315:1126–30.

von Mering, C., L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork. 2007b. STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35:D358–62.

Wach, A., A. Brachat, R. Pohlmann, and P. Philippsen. 1994. New heterologous modules for classical or PCR-based gene disruptions in Saccharomyces cerevisiae. Yeast 10:1793–808.

Wagner, B. K., T. Kitami, T. J. Gilbert, D. Peck, A. Ramanathan, S. L. Schreiber, T. R. Golub, and V. K. Mootha. 2008. Large-scale chemical dissection of mitochondrial function. Nat Biotech 26:343–351.

Watson, A. J. and P. S. Liss. 1998. Marine biological controls on climate via the carbon and sulphur geochemical cycles. Philos Trans R Soc Lond B Biol Sci 353:41–51.

White, O., J. A. Eisen, J. F. Heidelberg, E. K. Hickey, J. D. Peterson, R. J. Dodson, D. H. Haft, M. L. Gwinn, W. C. Nelson, D. L. Richardson, K. S. Moffat, H. Qin, L. Jiang, W. Pamphile, M. Crosby, M. Shen, J. J. Vamathevan, P. Lam, L. McDonald, T. Utterback, C. Zalewski, K. S. Makarova, L. Aravind, M. J. Daly, K. W. Minton, R. D. Fleischmann, K. A. Ketchum, K. E. Nelson, S. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser. 1999. Genome sequence of the radioresistant bacterium Deinococcus radiodurans R1. Science 286:1571–7.

White, T. C., S. H. Miyasaki, and N. Agabian. 1993. Three distinct secreted aspartyl proteinases in Candida albicans. J Bacteriol 175:6126–33.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences 95:6578–6583.

Wichern, R. and D. Johnson. 2003. Applied Multivariate Statistical Analysis. Fifth ed. Prentice Hall.

Wicker, T., E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein. 2006. 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7:275.

Wilson, R. B., D. Davis, and A. P. Mitchell. 1999. Rapid hypothesis testing with Candida albicans through gene disruption with short homology regions. J Bacteriol 181:1868–74.

Wommack, K. E., J. Bhavsar, and J. Ravel. 2008. Metagenomics: Read Length Matters. Appl. Environ. Microbiol. 74:1453–1463.

Wong, S. L., L. V. Zhang, A. H. Tong, Z. Li, D. S. Goldberg, O. D. King, G. Lesage, M. Vidal, B. Andrews, H. Bussey, and F. Roth. 2004. Combining biological networks to predict genetic interactions. PNAS 101:15682–15687.

Yip, K. Y., H. Yu, P. M. Kim, M. Schultz, and M. Gerstein. 2006. The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. Bioinformatics 22:2968–2970.

Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter. 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5:e16.

Young, D. M., D. Parke, and L. N. Ornston. 2005. Opportunities for genetic investigation afforded by Acinetobacter baylyi, a nutritionally versatile bacterial species that is highly competent for natural transformation. Annu Rev Microbiol 59:519–51.

Zeitlinger, J., I. Simon, C. T. Harbison, N. M. Hannett, T. L. Volkert, G. R. Fink, and R. A. Young. 2003. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell 113:395–404.

Zhang, Z., J. Rozowsky, H. Lam, J. Du, M. Snyder, and M. Gerstein. 2007. Tilescope: online analysis pipeline for high-density tiling microarray data. Genome Biology 8:R81–R81.