ABSTRACT

AYLOR, DAVID LAWRENCE. Not Just Another Trait: Methods for the Genetic Analysis of Gene Expression. (Under the direction of Zhao-Bang Zeng.)

Gene expression refers to the process by which DNA is transcribed to mRNA. It is now possible to measure genome-wide transcript abundance in many genetically distinct individuals. Genetical genomics refers to the application of quantitative genetic techniques to such data. We present two analyses of gene expression in distinct experimental populations.

We first present a method for a classical epistasis analysis that includes gene expression measurements. We propose a framework for estimating and interpreting epistasis that borrows from both classical and quantitative approaches. Regression analysis estimates the effects of gene deletions as well as interactions and significant effects are selected such that a reduced model describes each expression trait. We then show how the resulting models correspond to specific hierarchical relationships between two regulator genes and a target gene. These hierarchical relationships are the building blocks of systems diagrams and genetic pathways. Our framework can serve as a foundation for future epistasis analyses based on genomic data.

Secondly, we analyze expression quantitative trait locus mapping (eQTL) results in a segregating yeast population. We use prior information about yeast pathways to group expression measurements and ask questions about pathway regulation. We find that while many genes share quantitative trait loci, sharing is not prevalent within pathway groups. We

propose a possible explanation for our observations and describe how they fit in with previous interpretations of these data.

Lastly, we present a tool for manipulating sequence data within a population. Our software enables the user to pull out important features from a multiple alignment such as variable sites, unique haplotypes, and insertions or deletions. The output is compatible with a number of existing tools for population genetic analysis.

Not Just Another Trait: Methods for the Genetic Analysis of Gene Expression

by
David Lawrence Aylor

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Bioinformatics

Raleigh, NC

2008

APPROVED BY:

Dr. Zhao-Bang Zeng Committee Chair	Dr. Ignazio Carbone
Dr. Jeffrey Thorne	Dr. Philip Awadalla

Dedication

To my wife, Jessica

Biography

David Lawrence Aylor was born in Woodbridge, VA, about thirty miles south of Washington, D.C. His parents, Mrs. Bonita Satterfield Aylor and the late Mr. Herman Henry Aylor, Jr. raised their only son with an emphasis on family. Every Sunday after church they spent the afternoon at his grandmother's house with aunts, uncles and cousins. His favorite pastime was drawing, though most free time was spent outdoors.

It is a point of pride that David attended public schools throughout his academic career, from kindergarten to graduate school. At the University of Virginia, he received his Bachelors of Arts in Environmental Sciences in 1996. Summer courses at Mountain Lake Biological Station the year before were his first exposure to academic biology and were influential, yet he was eager to see the world and would not pursue additional formal education for another six years.

The last half of his graduation year and the first months of the next were spent in Anchorage, AK as an intern for The Nature Conservancy. His task was to develop maps of wildlife habitat using geographic information systems, a harbinger of the computing and biology combination yet to come. Computing would take the lead for the next five years, which were spent as a software developer in Charleston, SC and New York, NY. Media coverage of the genome projects led to new enthusiasm for computational biology and encouraged David to enroll in graduate school at North Carolina State in 2002. Soon afterwards, he met the former Jessica Banks Gilmour and they married in 2006.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Quantitative Genetics Concepts and Data

From the time of Mendel until the characterization of the genetic code, genes were physically unobserved. Mendel observed that offspring inherit traits from their parents in discrete units, and the term gene was later coined to describe this functional unit of heredity. Alleles are functionally distinct forms of a particular gene. Some traits vary continuously rather than discretely and at first did not appear to be inherited according to Mendel's rules. These traits were termed quantitative, meaning that they can be measured; they traits are also referred to as complex traits. Quantitative genetics is the study of such traits (Lynch and Walsh 1998). R.A. Fisher (1918) showed how combinations of multiple genes could explain quantitative variation in a population, and this has been a fundamental assumption of most subsequent quantitative genetic models. Non-genetic effects such as those imposed by different environments can also affect quantitative traits.

Two types of data are required for quantitative genetic analyses. These are the phenotypic variation and the genetic structure of the study population. Phenotypic variation or trait values can be measured in a variety of ways depending on the continuous trait being investigated. The degree of relatedness is usually a product of experimental design. Such analyses estimate such quantities as the proportion of heritable variation in a trait and its potential response to selection. These estimates are statistical in nature and describe the population of study organisms rather than any one individual. The methods used are predominately regression-based and rely on several assumptions about the properties of quantitative characters. Chief among these assumptions is that the effects of individual genes are additive and should be modeled as such. Deviations from additivity are interpreted as interactions between genes. Dominance refers to interactions between alleles within a locus. Epistasis refers to interactions between alleles at separate loci. Some models are extended to include such interaction effects.

Linkage analysis refers to a body of methods based on the observation that alleles do not always segregate independently. If two traits are observed together in offspring often but not always, then the allele combinations not observed in the parents are called recombinants, and recombination frequency is the fraction of these offspring. This concept was extended to create the first genetic map in fruit flies (Sturtevant 1919). To construct a linkage map, one must assume that genes exhibiting low frequency of recombination are located closer to each other on the chromosome, which had recently been validated as vectors of heredity (Morgan 1910). These frequencies are the genetic distance upon which the map is based. In eukaryotes, recombination is now known to be due to the biological phenomenon of crossing over during meiosis. It is notable that the original markers were phenotypes of mutants, and linkage maps predated many of the concepts we now associate with them.

The physical material of genes was unknown until the 1950s, when the structure of the DNA molecule was discovered and the gene concept expanded to describe something both functional and physical. A gene in this sense is a molecular region that is a template for an mRNA transcript. In the intervening years, the field of molecular genetics has provided details about how genes are transcribed into mRNA that is subsequently translated into proteins. These advances made it possible to link a function with a region of the genome.

Molecular Markers and QTL Mapping

Polymorphisms are individual differences in the DNA molecules within a population. In modern genetics, observed polymorphisms are called molecular markers and can be used to create linkage maps. Various types of markers exist including microsatellites, Restriction Fragment Length Polymorphisms (RFLPs), and Single Nucleotide Polymorphisms (SNPs). These are often associated with a particular technology for detecting variation.

Quantitative Trait Locus (QTL) mapping takes advantage of the gene's dual nature, by associating variation in observed traits with molecular markers. In short, the functional gene is mapped to a physical location. The goal of QTL mapping is to locate the physical regions of an organism's DNA that are linked to the trait of interest. Regions that are linked to a particular trait are called quantitative trait loci (QTLs). QTL mapping begins with observed trait measurements, a study population with a known degree of relatedness, and molecular markers from each individual in the study population. The relationship between individuals in the mapping population is usually a product of the experimental design. The molecular marker data should be sufficient to create a linkage map; it is common to have between a few dozen to a few hundred genetic markers spread throughout the genome of the study population. As mapping populations undergo more generations of recombination, they require more markers to study. This generally translates to increased resolution but less power to detect QTLs.

The simplest method of associating markers with trait values is the single marker analysis (SMA). For each marker, individuals are separated into two groups based on their marker allele and we test if the trait mean is different between the groups. The weaknesses of this approach are that it cannot identify multiple QTLs, cannot estimate QTL position, underestimates the QTL effect, and is not very powerful (Zeng 2000). A vast body of literature outlines more sophisticated methods for mapping the location of QTLs and the magnitude of their effect.

Interval mapping (IM) (Lander and Botstein 1989) extends the simple SMA test by testing association with the regions between two markers. This method is notable because it accounts for recombination between the markers and the QTL. The QTL is assumed to be located somewhere in the interval and at some genetic distance from each marker. A likelihood ratio (LR) test is used for each interval along the linkage map, and relates the hypothesis that a QTL in the given interval has an effect on the trait versus the hypothesis that it does not. The LR test statistic is usually reported as a log of odds (LOD) score, which is the same value arbitrarily scaled differently and historically used in human genetics.

Composite Interval Mapping (CIM) allows multiple QTLs to be evaluated in the same genetic model (Zeng 1993; Zeng 1994). This more closely agrees with the theory of quantitative traits, in which many genes combine to determine trait values. Multiple QTL models are accomplished in a sequential manner by first adding one QTL to the model in the manner of IM, then conditioning on that QTL when searching for additional QTLs. Multiple

Interval Mapping (MIM) (Kao et al 1999; Zeng et al 1999) refines CIM by virtue of an improved search procedure and allowance of more complex genetic models. The need for a QTL significance threshold is common to all these methods. This is an area of current development due to the computational burden of permutation testing, which has been the most successful approach (Churchill and Doerge 1994; Doerge and Churchill 1996).

Each QTL is assumed to contain one or more individual genes that contribute to the observed phenotype. However, QTL mapping is limited in resolution by the size of the mapping population, the amount of recombination in the population, and the density and distribution of markers. For these reasons, relatively few causative genes have been identified despite thousands of QTLs found for myriad traits. Generally, additional data must be incorporated to fine-map QTLs. However, following up on QTL results with additional genetic data is increasingly feasible because of advances in sequencing technology and the subsequent lower costs.

QTL mapping revolutionized genetic data analysis during an era in which phenotypic data was plentiful and genetic markers increased in availability. In model organisms, QTL mapping has led to discovery of new genes linked to well-studied phenotypes. It has allowed for advances in plant and animal breeding through identification of regions linked to traits of interest and marker-assisted selection. Most importantly, QTL mapping has provided new insights into the architecture of quantitative traits such as the distribution of QTL effects, the prevalence of heterosis, and the heritability of important traits.

The Genomic Era and Expression QTL Mapping

Again, advances in molecular biology are pushing the cutting edge of quantitative genetics. A genome is the entirety of the unique DNA belonging to a particular species, and the 1990s saw a flurry of genome projects for a variety of species. The goal of these projects is to create a physical map that contains the nucleotide sequence of an entire genome. Genes and other features are all placed on the map by processes referred to collectively as genome annotation. Some genes are located because they physically "look like" other genes. Others genes are mapped to specific functions through experimental means. Much of the progress in annotation was due to new technologies that allowed many simultaneous measurements. A litany of words ending in "omics" describes methods for measuring many molecules simultaneously – be they cellular proteins, metabolites, or mRNA transcripts. Gene expression refers to the relative amount of mRNA transcript produced by a particular gene, and the number of these experiments skyrocketed since the advent of cDNA microarrays and later manufactured oligonucleotide arrays for model organisms. These high throughput technologies have steadily grown in popularity, improved in quality, and lessened in cost over the past decade. During this time, procuring traditional phenotype data has not translated well to high-throughput methods and is relatively costly and time-consuming.

As gathering genomic data for segregating populations has become feasible, these data have been considered as continuous traits and candidates for QTL mapping in what has been called "genetical genomics." (Jansen and Nap 2001) Expression QTL (eQTL) mapping substitutes gene expression measurements for phenotype and uses essentially the same methods as traditional QTL mapping. However, rather than considering one or a few classical traits, many or even thousands of expression traits (e-traits) are considered. In practice, eQTL mapping is generally performed by iteratively applying traditional QTL mapping techniques to each individual expression measurement.

Expression traits have several appealing properties. Foremost, a physical location is associated with each trait, and one can distinguish between proximate *cis*-QTLs and *trans*-QTLs that locate elsewhere in the genome. *Cis*-QTLs indicate variation located in the same physical location as the gene itself. This may reflect variation in promoters or allelic variation. QTLs that regulate traits that are not collocated are called *trans*-QTL. *Trans*-regulatory polymorphisms may reside in regulatory proteins such as transcription factors, be indirect effects from genes upstream in regulatory pathways (Brem et al. 2002), or represent other mechanisms of regulatory feedback. Two patterns have been apparent in all eQTL studies to date. The first is a significant group of e-traits linked to a *cis*-QTL. The second pattern is multiple *trans*-QTLs that are shared by many traits, sometimes called hotspots. These two features have become starting points for how we describe expression QTL results and are central to how we interpret them.

Additionally, it is an important difference from phenotypic measurements that expression traits have an intrinsic molecular biological role (Schadt et al. 2003). Bioinformatics databases store wealth of information about specific genes. Therefore, expression traits can be associated with phenotypes and molecular biological functions without any additional experimentation. Likewise, e-traits can be grouped according to such data. This additional information can generate hypotheses regarding the gene underlying QTL.

Three Intergenic Relationships in Expression QTL Experiments

Traditional QTL mapping reflects a relationship between a physical region and a phenotype. Inasmuch as we assume a causative gene underlies the QTL, this can be considered a gene-trait relationship. Since in eQTL mapping the trait is itself a gene, the *trans*-QTL is a relationship between two genes. A polymorphism in one gene affects the expression of another. Epistasis can be modeled just as in traditional QTL mapping, as part of the genetic model. Based on the same assumption that a gene underlies each QTL, this is also a relationship between genes. Polymorphisms in two genes affect the expression of a third, or one of the genes themselves if one of the QTL is in *cis*-. Lastly, there are relationships between the expression measurements that represent a third relationship between genes. Expression traits can be correlated (co-expression of genes), share a QTL (co-regulation of genes), or represent genes of shared function. These three relationships allow us to interpret eQTL results in the context of regulatory networks. Depending on the experiment, additional relationships between genes and phenotype can also be explored.

We present several lines of research from the past five years in this context. Pioneering work in yeast helped describe the basic patterns how expression is regulated by genetic variation. Mice studies showed system level genetics in the context of different organs, and the community has made unique contributions to exploring the relationships between expression traits. Plant research has contributed the most in terms of integrating expression QTL mapping with phenotype and other biomolecular data.

Expression QTL in Yeast

In a series of papers, Leonid Kruglyak and colleagues presented expression QTL mapping results from yeast (Brem and Kruglyak 2005; Brem et al. 2005; Brem et al. 2002; Yvert et al. 2003). A wild isolate from a vineyard was crossed with a standard laboratory strain, and expression was measured on 112 haploid segregants. Yeast have a haploid phase in their life cycle, so these segregants are genetically comparable to a diploid F1. Both capture one generation of recombination. Preliminary results were released based on 40 segregants and later 86 segregants. Gene expression was measured for 6216 trait genes; 489 were eventually disregarded due to an independent survey of the yeast genome (Kellis et al. 2003), leaving 5727 genes. In a controlled environment, one quarter of the expression traits measured were differentially expressed between the parent strains at $P < 0.005$. They estimated a median heritability of 0.84 using mid-parent/offspring regression. Because there is no dominance in haploid organisms, this is heritability in both the broad and narrow sense. 3312 genetic markers were typed.

In the initial study using 40 individuals, 308 of the differentially expressed traits were linked to at least one marker in a single marker analysis. A third of these linkages were *cis*-acting. Transgressive segregation was common and may have complicated detecting linkage.

Eight shared *trans*-regulators accounted for over 40% of QTL, and each was characterized as being enriched for genes of a common function. For the shared regulators that also have a *cis*- linkage, the trait with the *cis*- linkage was inferred to be the gene harboring a causative polymorphism. Candidate regulator genes were supposed by prior knowledge of genes in the linkage region, but for which the corresponding expression trait had no *cis*-QTL.

In a follow-up paper, Yvert et al (Yvert et al. 2003) found many additional linkages using 86 segregants (2294 at $P < 3.4 \times 10^{-5}$, 992 at $P < 5 \times 10^{-7}$). The number of shared *trans*-regulators grew to thirteen. In this report they used correlation-based k-means clustering to group similar traits based on their expression profiles. Thirty percent of expression traits clustered with at least one other gene. They found 593 clusters of at least two genes and 205 larger clusters. This grouping approach is independent of the linkage analysis, but groups based solely on data generated from the experimental cross. These clusters were enriched for genes with shared annotations, and the mean expression of genes within each cluster linked to genetic variation for over half of clusters. They found that *trans*-QTLs regulating clusters were rarely transcription factors and noted diverse biological functions among putative *trans*-regulator genes. This may be due to strong evolutionary conservation among transcription factors.

With additional samples (Brem and Kruglyak 2005), they estimated slightly more expression traits with at least one QTL (2984 at FDR = 0.05). Most of these traits had high heritability of over 0.69. They explored the number of loci underlying these high heritability traits by fitting additive genetic models composed of one through ten QTLs of equal effect.

There was a direct relationship between increasing model complexity and the proportion of traits supporting each model, suggesting that expression traits are regulated by many QTL.

The additional information associated with expression traits can be used to identify the candidate genes underlying QTL. These methods have perhaps been developed best in yeast due to the wealth of biological information available on its molecular biology. Prior knowledge regarding the function of the genes in a QTL region has in several cases provided a clear candidate for positional cloning and functional assays. However, it is not feasible to conduct additional experiments for more than a few QTL. Computational approaches combine experimental data with bioinformatics. Bing and Hoeschle (2005) correlated the expression profiles of genes located in each QTL region with the expression profile of traits linked to that QTL. This approach assumes that related genes have strong correlations in their expression. It also indirectly assumes that the causative gene has a *cis*-QTL, since the marker would necessarily be correlated with both expression profiles if they were correlated with each other. However, they identified candidate genes for both *cis*- and *trans*-QTL, suggesting for the *trans*-QTLs that the underlying *cis*- regulation did not meet the significance threshold needed to declare a QTL in the linkage analysis.

Epistasis: Results from Yeast

The same yeast cross has also led to insights in the second intergenic relationship -- epistasis. Epistasis can be defined as interaction between two alleles at different loci.

Modeling epistasis in a segregating population involves statistically testing for interaction between each pair of loci or putative QTL. Testing all possible pair of loci is computationally demanding and has low power because of the enormous number of tests that must be performed. Epistasis is not always included in traditional QTL models, but such issues are compounded when thousands of expression traits are analyzed. Brem et al (2005) used a two-stage search strategy to find secondary loci only after conditioning on a primary locus, and then fitting a model including effects for both loci and an interaction. They found 65% of expression traits showed significant epistasis at $P < 0.05$. Creating yeast strains with each of the four allelic combinations validated a significant interaction between *MAT* and *GPA1* loci. The ability to manipulate yeast so easily makes them ideal for epistasis analyses at a range of genetic variation. Zou and Zeng applied Multiple Interval Mapping to the same data set, which also searches for models including epistasis (Zou and Zeng 2008).

Expression QTL in Mice

Schadt and colleagues showed that similar approaches could be used in mice (Schadt et al. 2003). They profiled 111 F2 mice from a cross of standard laboratory strains (C5BL/6J × DBA2J). They measured gene expression in liver cells because of the liver's role in obesity, which was their target phenotype. Using interval mapping, they observed that *cis*-QTLs were generally of larger effect than *trans*-QTLs, and found QTL associated with the major differences between parent strains that they expected. They combined their QTL analysis with phenotype by measuring fat pad mass on the mice. The expression data allowed them to identify two distinct expression profiles among the high fat mice, suggesting disease subtypes controlled by unique QTLs.

Two concurrently released papers extended this approach using the BXD recombinant inbred (RI) strains (Bystrykh et al. 2005; Chesler et al. 2005) to profile expression QTLs in two tissues. Using a 10% FDR they found that 83 of 88 QTLs were *cis*- in the brain. This is consistent with *cis*-QTLs having larger effects sizes than *trans*-QTLs; the more conservative the significance threshold is, the higher the ratio of *cis*- to *trans*- should be. Hotspots were prevalent but did not collocate between these three experiments, meaning that genetic control of transcription is tissue specific.

Relationships Between Expression Traits

It is intuitive to group expression traits that share a QTL and ask whether those traits may be related biologically. Another approach is to group related traits before the mapping analysis. Several methods have been explored that fall into two general types. The first type groups traits based on expression measurements from the experimental data. Clustering traits by expression profiles is a simple method that has been prevalent in the literature. Alternately, sets of genes can be defined *a priori*. Kliebenstein et al (Kliebenstein et al. 2006) mapped QTL for twenty networks in *Arabidopsis*; each network was pulled from previously published studies or cumulative experimental results. We present an *a priori* analysis for yeast in Chapter 3.

Composite trait mapping has been used in a variety of organisms (Ghazalpour et al. 2006; Kliebenstein et al. 2006; Yvert et al. 2003). This refers to mapping a trait derived from averaging or otherwise combining multiple grouped expression traits. In yeast and mice, the expression trait values within each cluster were averaged for every individual, and this average was treated as that individual's trait value for linkage analysis. In yeast, 304 of 593 clusters showed some linkage at $P < 3.4 \times 10^{-5}$. The idea behind composite traits mapping is to increase the power to identify shared regulators. However, shared regulators may not be present for every group and this could lead to incorrect conclusions.

A method for composite trait mapping is illustrated clearly in mice (Ghazalpour et al. 2006). They propose a four-step process in which traits are related by constructing a gene co-expression network, the biological significance of each module is determined from referencing bioinformatics data, a composite (average) trait is mapped to QTLs, and the relationship between QTL and module is reevaluated for biological relevance. They group traits based on pair-wise correlation of expression values, and refer to such groups as network modules. The biological significance of each module is determined by enrichment for Gene Ontology (GO) terms or pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG). QTLs linked to the composite trait for a given module are called module QTLs (mQTLs). Markers within each mQTL region were evaluated based on correlation with each trait within the module. This framework captures the essentials of most composite trait mapping methods.

Expression QTL in Plants

A recent study in *Arabidopsis* captures variation from 211 recombinant inbred lines (RIL) (West et al. 2007). Sixty-nine percent of 22,746 e-traits linked to QTLs at a $P < 0.05$ significance threshold. One third of these had a *cis*-QTL, which is slightly more than in yeast. One explanation for this is more variation was captured in the parental generation. In general, patterns in plants match those in other eukaryotes, including the relative proportion of *cis*- and *trans*- and the presence of hotspots.

While the techniques of eQTL mapping vary little between organisms, the plants community has contributed the most in terms of integrating the many levels of biomolecular variation including markers, expression, metabolites, and phenotypes. Jansen and Nap (2001) envisioned such an approach when they conceived genetical genomics and mentioned it explicitly. However, such data remain rare because of their complexity and expense. When gene expression correlates with a complex phenotype, the corresponding expression traits may reflect the molecular basis of that phenotype at a level intermediate between genotype and phenotype. The presumption is that expression QTL will collocate with metabolic and phenotypic QTL.

Results have in part confirmed this architecture. Expression QTLs for genes involved in flowering time were located at the same position as known QTLs for circadian period length and other related phenotypes (Keurentjes et al. 2007). For genes involved in glucosinolate metabolite production, all expression QTLs were in the same region as QTLs for the metabolite levels (Wentzell et al. 2007). However, metabolic QTL did not always collocate with eQTL. This provides important insight on how to integrate various levels of biological information and the mechanisms by which genetic variation affects an organism.

Network Models

Several studies suggest that genes with shared linkages and similar expression profiles are related within hierarchical genetic networks or pathways. The definition of network varies widely. For instance, groups of coregulated expression traits (referred to above as modules) may or may not represent a network, and we reserve the term for methods that go beyond simply grouping traits to constructing directed graphs that combine QTLs and e-traits.

Structural equation modeling (SEM) is an extension of multivariate regression that allows for hierarchical relationships. Variables are represented as nodes in the graph and the relationships between them as edges. Edges are directed from independent variables to dependent variables. In this context, that means edges are directed from QTLs to traits that are linked to those QTLs (Li et al. 2006). Additionally, expression e-traits can explain variation in other e-traits, implying that a QTL affects the expression of a gene that in turn affects other genes. This sort of regulatory cascade is expected biologically and these models provide a way of separating such downstream effects from direct QTL effects. A model can consist of many equations describing relationships between the variables, and individual variables may be present in multiple equations. The model is assessed for goodness-of-fit by comparing expected and observed covariance matrices. Using SEM for an traditional QTL analysis of obesity in mice revealed several QTLs linked to lean body weight and concluded lean body weight was causal to fat pad mass. In other words, several QTLs were affecting obesity only indirectly. SEM has not yet been used for expression QTL studies, but its ability to disentangle conditional relationships between QTLs and multiple traits makes it well suited for future network based analyses.

Bayesian network modeling is a similar method that has been used for eQTL analyses with some success. These networks also take the form of a directed graph and nodes and edges are the same as for the SEM approach. The main difference between the two is that SEMs are based on correlation structure while Bayesian networks emphasize conditional probabilities. The mouse liver eQTL data (Schadt et al. 2003) were used to demonstrate that causal relationships could be inferred by with eQTL information that could not be inferred using expression data alone (Zhu et al. 2004; Zhu et al. 2007). Further refinements include a likelihood-based causality model selection (LCMS) test (Schadt et al. 2005). One weakness of these methods is that the resulting graphs are acyclic, meaning there is no way to represent feedback in the model. Additionally, strong assumptions are made for *cis*-QTLs being causal nodes. This assumption could be tested by additional experiments targeting specific loci.

Conclusions

Seven years ago, eQTL Mapping was just a concept. We now have data from studies on yeast, mice, humans, nematodes, and plants, and a growing number of methodological articles with ideas for analyzing these data. Several insights into the genetic architecture of gene expression are now widely agreed upon (Gibson and Weir 2005). Expression traits generally are controlled by multiple QTLs, with *cis*-QTL being less abundant but of larger effect size than *trans*-QTL. Some eQTL are highly pleiotropic and are linked to many traits, forming hotspots.

However, much remains unknown about expression traits. Estimates of heritability have varied widely between studies, from less than 0.1 to 0.95. The role of epistasis has not been studied widely. No articles have yet reported estimates of gene × environment interaction on gene expression, though it is expected to be large. Ultimately, the challenge will be to integrate multiple levels of biomolecular data into these studies. A recent experiment in yeast suggested few relationships between QTL controlling protein levels with those controlling expression (Foss et al. 2007). More success has been found with metabolic QTL mapping (Wentzell et al. 2007). The data for these types of multi-tier studies are sparse, and future experiments will shed light on these open questions. Concurrently, new methods for gene networks should be refined to separate direct from indirect effects and shift results from lists of genes to graphical networks.

Currently, these experiments' biggest weakness is the expense involved in generating expression and sequence data in a quantity appropriate to a highly powered study. However, technology has improved and cheapened over just the past few years and we anticipate it will continue to do so. Already, the linkage/QTL approach is being complimented by genome-wide association studies (GWAS) due to the availability of SNP arrays for a number of organisms. Gene expression measurement has improved markedly due to standard platforms and analysis methods, but new methods for quantifying cellular transcripts are always in development. Whatever technological changes the future brings, the essence of genetical genomics will remain unchanged — associating genetic variation with variation in the level of other measurable biomolecules. Given the enthusiasm with which such inquiry has been embraced, it likely will remain at the forefront of biology in the coming decade or more.

References

Bing, N. and I. Hoeschele. 2005. Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170:** 533-542.

Brem, R.B. and L. Kruglyak. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* **102:** 1572-1577.

Brem, R.B., J.D. Storey, J. Whittle, and L. Kruglyak. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436:** 701-703.

Brem, R.B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296:** 752-755.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M.T. Pletcher, T. Wiltshire, A.I. Su, E. Vellenga, J. Wang, K.F. Manly, L. Lu, E.J. Chesler, R. Alberts, R.C. Jansen, R.W. Williams, M.P. Cooke, and G. de Haan. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37:** 225-232.

Chesler, E.J., L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H.C. Hsu, J.D. Mountz, N.E. Baldwin, M.A. Langston, D.W. Threadgill, K.F. Manly, and R.W. Williams. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37:** 233-242.

Churchill, G.A. and R.W. Doerge. 1994. Empirical threshold values for quantitative trait mapping. *Genetics* **138:** 963-971.

Doerge, R.W. and G.A. Churchill. 1996. Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142:** 285-294.

Fisher, R.A. 1918. The correlation between relatives on the supposition of mendelian inheritance. *Phil Trans Roy Soc Edin* **52:** 399-433.

Foss, E.J., D. Radulovic, S.A. Shaffer, D.M. Ruderfer, A. Bedalov, D.R. Goodlett, and L. Kruglyak. 2007. Genetic basis of proteome variation in yeast. *Nat Genet* **39:** 1369-1375.

Ghazalpour, A., S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E.E. Schadt, T.A. Drake, A.J. Lusis, and S. Horvath. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2:** e130.

Gibson, G. and B. Weir. 2005. The quantitative genetics of transcription. *Trends Genet* **21:** 616-623.

Jansen, R.C. and J.P. Nap. 2001. Genetical genomics: the added value from segregation. *Trends Genet* **17:** 388-391.

Kao, C.H., Z.B. Zeng, and R.D. Teasdale. 1999. Multiple interval mapping for quantitative trait loci. *Genetics* **152:** 1203-1216.

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:** 241-254.

Keurentjes, J.J., J. Fu, I.R. Terpstra, J.M. Garcia, G. van den Ackerveken, L.B. Snoek, A.J. Peeters, D. Vreugdenhil, M. Koornneef, and R.C. Jansen. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proc Natl Acad Sci U S A* **104:** 1708-1713.

Kliebenstein, D.J., M.A. West, H. van Leeuwen, O. Loudet, R.W. Doerge, and D.A. St Clair. 2006. Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics* **7:** 308.

Lander, E.S. and D. Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185-199.

Li, R., S.W. Tsaih, K. Shockley, I.M. Stylianou, J. Wergedal, B. Paigen, and G.A. Churchill. 2006. Structural model analysis of multiple quantitative traits. *PLoS Genet* **2:** e114.

Lynch, M. and B. Walsh. 1998. *Genetics and the Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Morgan, T.H. 1910. Chromosomes and heredity. *The American Naturalist* **44:** 449-496.

Schadt, E.E., J. Lamb, X. Yang, J. Zhu, S. Edwards, D. Guhathakurta, S.K. Sieberts, S. Monks, M. Reitman, C. Zhang, P.Y. Lum, A. Leonardson, R. Thieringer, J.M. Metzger, L. Yang, J. Castle, H. Zhu, S.F. Kash, T.A. Drake, A. Sachs, and A.J. Lusis. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* **37:** 710-717.

Schadt, E.E., S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, P.S. Linsley, M. Mao, R.B. Stoughton, and S.H. Friend. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422:** 297-302.

Sturtevant, A.H., C.B. Bridges, and T.H. Morgan. 1919. The spatial relations of genes. *Proc Natl Acad Sci* **5:** 168-173.

Wentzell, A.M., H.C. Rowe, B.G. Hansen, C. Ticconi, B.A. Halkier, and D.J. Kliebenstein. 2007. Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet* **3:** 1687-1701.

West, M.A., K. Kim, D.J. Kliebenstein, H. van Leeuwen, R.W. Michelmore, R.W. Doerge, and D.A. St Clair. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175:** 1441-1450.

Yvert, G., R.B. Brem, J. Whittle, J.M. Akey, E. Foss, E.N. Smith, R. Mackelprang, and L. Kruglyak. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* **35:** 57-64.

Zeng, Z.B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc Natl Acad Sci U S A* **90:** 10972-10976.

Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. *Genetics* **136:** 1457-1468.

Zeng, Z.B., C.H. Kao, and C.J. Basten. 1999. Estimating the genetic architecture of quantitative traits. *Genet Res* **74:** 279-289.

Zeng, Z.B. 2000. Statistical Methods for Mapping Quantitative Trait Loci.

Zhu, J., P.Y. Lum, J. Lamb, D. GuhaThakurta, S.W. Edwards, R. Thieringer, J.P. Berger, M.S. Wu, J. Thompson, A.B. Sachs, and E.E. Schadt. 2004. An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res* **105:** 363-374.

Zhu, J., M.C. Wiener, C. Zhang, A. Fridman, E. Minch, P.Y. Lum, J.R. Sachs, and E.E. Schadt. 2007. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLoS Comput Biol* **3:** e69.

Zou, W. and Z.B. Zeng. 2008. Multiple Interval Mapping for Gene Expression QTL Analysis.

Chapter 2

From Classical Genetics to Quantitative Genetics to Systems Biology: Modeling Epistasis

David L. Aylor and Zhao-Bang Zeng

Abstract

Epistasis has long had two slightly different meanings depending on the context in which it is discussed. The classical definition describes an allele at one locus completely masking the effect of an allele at a second locus. Such relationships can be interpreted as hierarchical, and they can be combined to infer genetic pathways. In quantitative genetics, epistasis encompasses a wide range of interactions and can be extended to more than two loci. These two definitions coexist because they are typically applied to different types of study populations and different types of traits.

The current trend is to treat gene expression as a trait in a variety of genetic backgrounds. Gene expression data has been used in lieu of phenotype in both classical and quantitative genetic settings. This provides a reason to revisit epistasis in this new context. We propose a framework for estimating and interpreting epistasis from a classical experiment that combines the strengths of each approach. We accommodate the continuous nature of gene expression using ideas from quantitative genetics. Regression analysis estimates the effects of gene deletions as well as interactions. Significant effects are selected such that a reduced model describes each expression trait. We show how the resulting models correspond to specific hierarchical relationships between two regulator genes and a target gene. These hierarchical relationships are the building blocks of systems diagrams and genetic pathways. This framework can serve as a foundation for future epistasis analyses based on genomic data.

Introduction

Epistasis has traditionally been discussed in two distinct contexts, corresponding to the disciplines of classical molecular genetics and quantitative genetics. In each case, the term describes an interaction between alleles at two or more loci. However, the methods for detecting epistasis and interpretations of the underlying biology have kept historical divisions in place despite calls for synthesis (Phillips 1998). This is largely because the two fields traditionally study different types of traits in different experimental populations.

The classical epistasis experiment compares a double-mutant with two associated single-mutants. Epistasis is present if the observed double-mutant phenotype is categorized as being the same as a single-mutant phenotype. This implies a specific type of interaction in which an allele at one locus masks the effect of variation at the second locus. This relationship is described as the first locus being epistatic to the second, and can be interpreted as one gene acting upstream of the other. This hierarchical interpretation has been used to construct biological pathways via a series of epistatic gene pairs. However, this approach is limited by the necessity of easily observed and categorized phenotypes (Hughes et al. 2000).

In contrast, quantitative genetics examines traits that vary continuously and cannot easily be categorized. Such trait distributions may result from the cumulative effects of many genes. Each additional gene increases the possible combination of alleles, and the number of possible phenotypes grows exponentially. An individual's phenotype is the sum of the allelic effects at each gene and the effect of the environment. Epistasis is defined as a deviation from these additive gene effects (Lynch and Walsh 1998). A quantitative genetic model can include multiple loci and multiple interactions. Epistasis in this sense describes a functional relationship between genes in the context of a trait, but it includes both hierarchical relationships and nonhierarchical relationships and there is no way to distinguish between these.

Any genetic effect is only relevant to the population being studied due to the presence of genetic background. Background is genetic variation that is unobserved in the population and cannot be modeled. The classical experiment is performed using genetically homogenous laboratory strains so there is no background. Quantitative genetics studies diverse populations and background variation is almost always present. The implication of this is that epistasis may be detected in one experiment but not in another. This has led to criticisms that epistasis in the quantitative genetic sense is a statistical construct rather than a true representation of biology.

In fact, both approaches seek to illustrate underlying molecular architecture and each has its strengths. A hierarchical interpretation of epistasis is attractive as increased focus is placed on genetic pathways and systems diagrams. However, quantitative approaches are necessary to accommodate continuous data types such as gene expression, metabolite concentrations, and fitness. Recent literature suggests that such approaches are being adopted. For example, while early large-scale fitness profiles in yeast deletion mutants (Tong et al. 2001; Tong et al. 2004) were scored categorically, St Onge et al (St Onge et al. 2007) measured fitness in 650 double-deletion yeast strains and employed a novel quantitative analysis.

The rise in genomic techniques has broken down one of the traditional barriers discussed above: the same traits are now being used in both classical and quantitative settings

(Jansen and Nap 2001). Gene expression is perhaps the most prevalent example. Instead of a single phenotypic trait value, a vector of expression measurements describes each individual. Expression profiling in single-deletion yeast strains found that 34% of mutants showed twenty or more differentially expressed genes (Hughes et al. 2000). Expression quantitative trait locus (eQTL) mapping uses a linear modeling approach to associate genetic variation with gene expression traits (Brem et al. 2002; Bystrykh et al. 2005; Chesler et al. 2005; Li and Burmeister 2005; Schadt et al. 2003). Storey et al. (Storey et al. 2005) found over thirty percent of traits were jointly linked to two loci in yeast. When gene expression correlates with a complex phenotype, the corresponding traits may reflect the molecular basis of that trait at a level intermediate between genotype and phenotype. Some studies suggest that epistasis is pervasive among expression traits (Auger et al. 2005; Gibson et al. 2004; Gibson and Weir 2005) and such traits may have more QTLs than classical traits (Brem and Kruglyak 2005; Storey et al. 2005). Since gene expression is being used in both classical and quantitative contexts, it is a valuable framework in which to compare the ability to detect epistasis and interpret the nature of relationships between genes.

We propose a framework for estimating and interpreting epistasis using expression traits. Our goal is to accommodate the continuous nature of the data, yet still preserve a hierarchical interpretation of epistasis. Such interpretations are well established for classical epistasis experiments (Avery and Wasserman 1992), but have only recently been studied for complex data (Li et al. 2006). We refine the classical interpretations by explicitly modeling gene expression. Gene effects and interactions are estimated using a linear model, in a manner comparable to eQTL mapping. Our method selects the best-fit regression model for each trait, which describe the order and the nature of gene function. Such relationships are the basic units of genetic pathways and systems biology. We specifically address how to use a continuous phenotype in a manner that is both statistically sound and consistent with the classical approach.

We illustrate our method with publicly available expression measurements from *Dictyostellium discoideum* wild type (Van Driessche et al. 2002) and deletion mutant strains (Van Driessche et al. 2005). This experiment is a classical epistasis analysis that targets the genes of the protein kinase (PKA) pathway and measures the gene expression profile of each strain.

Results

Modeling Epistasis for Continuously Variable Traits

In the classical epistasis analysis, triplets of deletion mutants combine with a wild type to form a contrast. Each contrast includes two single mutants and a double mutant. Each is described relative to the known wild type phenotype. A hypothetical example of a trait affected by two genes, *A* and *B*, can be described as follows, where *y* is the trait value, μ is the expected value of the wild type, β_A and β_B are the effects of deleting each gene, and ε is an error term.

$$\begin{aligned}
&A^+B^+ : y = \mu + \varepsilon \\
&A^-B^+ : y = \mu + \beta_A + \varepsilon \\
&A^+B^- : y = \mu + \beta_B + \varepsilon \\
&A^-B^- : y = \begin{cases} \mu + \beta_A + \varepsilon & \text{if } A \text{ is epistatic to } B \\ \mu + \beta_B + \varepsilon & \text{if } B \text{ is epistatic to } A \end{cases}
\end{aligned}$$

This adheres strictly to the classical definition, but there is a clear problem; there is no provision if the double mutant does not fall neatly into the same category as one of the single mutants. Gene expression traits fit poorly into the classical framework for this reason. Expression is continuous and intermediate levels are expected. Furthermore, even normalized trait values will inevitably include some measurement error. For these reasons, the double mutant observation is rarely the same as either of the single mutant observations or the wild type. Previous studies have attempted to circumvent this problem by relying on differences between the mutants to determine the most similar mutant pair. However, the assumption that expression is completely masked is poor. To address these issues, we move away from comparing trait values directly. Instead, we evaluate each deletion according to whether it significantly affects the expression of the target and associate unique patterns of significance with models of gene action.

We use a linear model to estimate the effect of each deletion. This is a general way to relate all mutants and the wild type without making any assumptions about the nature of the double mutant. We regress the trait value (e.g. expression) on indicator variables representing the presence or absence of each wild type allele and an interaction term. The interaction describes effects that are unique to the double mutant. The same example discussed above can be described as follows.

Trait value = Wild Type + Effect of deleting A + Effect of deleting B + Interaction + error

$$y = \mu + \beta_A x_A + \beta_B x_B + \beta_I x_A x_B + \varepsilon$$

$$x_A = \begin{cases} 0 \textit{ for } A^+ \\ 1 \textit{ for } A^- \end{cases} \qquad x_B = \begin{cases} 0 \textit{ for } B^+ \\ 1 \textit{ for } B^- \end{cases}$$

Various techniques can be used to fit such a linear model. We first fit a full model and then use stepwise backwards selection to drop model terms with coefficients that are not significant at a set level. The resulting reduced model is termed the best-fit model. For any trait, there are eight possible best-fit models. For clarity, we number the reduced models as follows:

$$\begin{aligned}
&\text{Model 1}: y = \mu + \beta_A + \varepsilon \\
&\text{Model 2}: y = \mu + \beta_B + \varepsilon \\
&\text{Model 3}: y = \mu + \beta_I + \varepsilon \\
&\text{Model 4}: y = \mu + \beta_A + \beta_B + \varepsilon \\
&\text{Model 5}: y = \mu + \beta_A + \beta_I + \varepsilon \\
&\text{Model 6}: y = \mu + \beta_B + \beta_I + \varepsilon \\
&\text{Model 7}: y = \mu + \beta_A + \beta_B + \beta_I + \varepsilon \\
&\text{Model 8}: y = \mu + \varepsilon
\end{aligned}$$

When the best-fit model has been determined, we estimate parameter values using that model for each trait. Thus, we have a best-fit model and coefficient estimates for each trait. The terms in each best-fit model represent the significant gene and interaction effects acting on that trait. Individual coefficients represent the estimated effect of deleting each gene. Model 7 corresponds to the classical model above when the interaction between the two deletions offsets the effect of one of them, either $\beta_I = -\beta_A$ or $\beta_I = -\beta_B$. Model 8 describes the case in which the deleted loci have no effect on the trait.

A best-fit model describes each gene expression trait. As such, we have dealt with the continuous variable problem. However, by embracing a quantitative genetic model we have lost the appealing feature of the classical experiment: the ability to interpret hierarchical relationships. In the following section we identify sixteen hierarchical relationships and propose that a specific best-fit model supports each.

Interpreting Hierarchical Epistasis

In quantitative genetics, the interaction term in the above model is considered epistasis. However, epistasis in this sense includes both hierarchical and nonhierarchical relationships. Conversely, while Model 7 can clearly be interpreted as hierarchical epistasis with the conditions described above, it does not apply to all possible hierarchies.

We considered all combinations of gene order and action within simple ON/OFF models and then predicted the hypothetical effect of deleting genes on each of them (Figures 2.1, 2.3, 2.4, 2.5). There are four points of variation to model for each gene pair relationship. The first is the identity of the upstream gene, i.e. the gene order. Secondly, the upstream gene will turn the downstream gene either on (enhance) or off (repress). Thirdly, the downstream gene can enhance or repress the expression of a target gene for which expression is observed. Lastly, we consider that the upstream gene itself will be enhanced or repressed by some initiating factor such as a developmental cue or environmental perturbation. Avery and Wasserman (Avery and Wasserman 1992) provide a general framework that has been widely used for interpreting epistasis in response to such signals, and note that the effect of a mutation is only observable for a specific signal state. However, knowing the signal state does not give any information about whether the upstream gene is enhanced or repressed in that state. In our models, we focus on the effect on the upstream gene. This model has sixteen possible variants describing hierarchical relationships between two genes and the target gene.

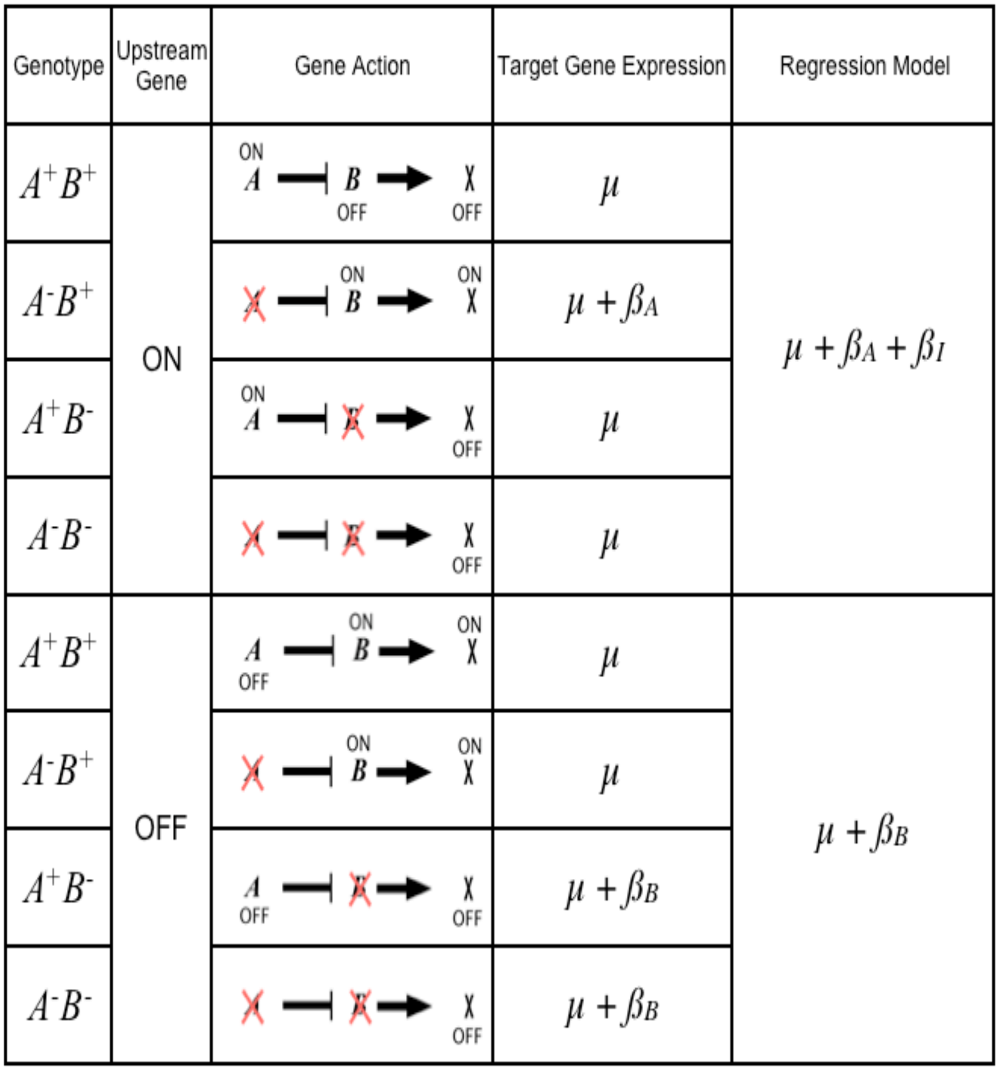


Figure 2.1 Modeling the relationship *A is an upstream repressor of B*. Gene *B* in turn enhances a target gene *X*. In this example, deleting *A* will change the state of the target gene from off to on. Therefore, we include *A*'s effect in the corresponding regression model. Deleting *B* leaves the target gene in the same state as the wild type and its effect is not included. The *AB* double mutant is also not expected to deviate from the wild type despite the significance of the *A* deletion. Since *A*'s effect is already included in the model for this contrast, it must be offset by the interaction term. We conclude that if *A* is enhanced by the signal, *A* represses *B*, and *B* enhances *X*, the corresponding best-fit regression model will include coefficients for *A* and an interaction term. Similar logic applies to the case in which the signal represses *A*. The signal represses *A*, thus deleting *A* has no downstream effects. We expect only the coefficient corresponding to the downstream gene in the best-fit model.

The key to our approach is connecting each of the sixteen hierarchical models to one of the eight possible best-fit regression models. If the deletion changes the state of a target gene relative to the wild type in a mutant, then that deletion is predicted to have a significant effect and it will be included in the regression model corresponding to that hierarchical model. Figure 1 gives an example of one possible model, in which *A* is enhanced by a signal; *A* is an upstream repressor to *B*; and *B* enhances a target gene *X*. We conclude that the corresponding best-fit regression model will include coefficients for *A* and an interaction term. Note that if the signal instead represses *A*, a different best-fit model represents the same relationship between *A* and *B*.

We applied the same approach to each of the sixteen cases and note several trends. First, the downstream gene's effect upon the target gene *X* does not influence the corresponding best-fit model. This allows us to reduce the model space to eight hierarchical relationships (Table 2.1a). This observation is convenient, because expression traits represent all the genes downstream of the deletions. Regardless of the downstream gene's direct effect, some traits will be enhanced while others are repressed. When the upstream gene is a repressor, four distinct regression models represent four unique hierarchical relationships. We can uniquely identify both the gene order and signal effect on the upstream gene. We cannot discern gene order if the upstream gene is an enhancer because the same best-fit model describes both hierarchies. If the upstream gene is merely enhancing the effect of the downstream gene, deleting either gene will affect the trait gene similarly. Six of the

Table 2.1. Correspondence between regression models and biological models
a. Six of the eight possible regression models represent hierarchical relationships between genes. If the upstream gene is a repressor we can identify gene order and the signal effect. If the upstream gene is an enhancer, we can identify only the signal effect. If the signal turns off an upstream enhancer, deleting either gene will have no effect. **b.** Non-hierarchical relationships can be distinguished if both genes are activated by the signal. Model 3 suggests buffering, while Model 4 suggests independent effects, i.e. no epistasis. If a potential regulator is turned off by the signal it has no effect on the target gene.

a. Hierarchical Relationships

	A upstream of B		B upstream of A	
Upstream Gene	ON	OFF	ON	OFF
Repressor	$\mu+\beta_A+\beta_I$ [5]	$\mu+\beta_B$ [2]	$\mu+\beta_B+\beta_I$ [6]	$\mu+\beta_A$ [1]
Enhancer	$\mu+\beta_A+\beta_B+\beta_I$ [7]	μ [8]	$\mu+\beta_A+\beta_B+\beta_I$ [7]	μ [8]

b. Non-hierarchical Relationships

<table>
<tr><th>State of A/B</th><th>ON/ON</th><th>ON/OFF</th><th>OFF/ON</th><th>OFF/OFF</th></tr>
<tr><td>Enhancer/Enhancer</td><td>$\mu+\beta_I$ [3]</td><td rowspan="3">$\mu+\beta_A$ [1]</td><td rowspan="3">$\mu+\beta_B$ [2]</td><td rowspan="3">μ [8]</td></tr>
<tr><td>Enhancer/Repressor
Or
Repressor/Enhancer</td><td>$\mu+\beta_A+\beta_B$ [4]</td></tr>
<tr><td>Repressor/Repressor</td><td>$\mu+\beta_I$ [3]</td></tr>
</table>

eight possible best-fit regression models correspond to the eight hierarchical relationships. It is notable that hierarchies can be indicated even without an interaction effect in the model.

We must also consider that there is no hierarchical relationship between *A* and *B*, or that they do not affect the target gene (Table 2.1b). We can distinguish between two types of parallelism. Model 4, the two-gene additive model with no interaction, represents no epistasis. Model 3 represents buffering epistasis, in which both genes act on the target in the same direction, and the effect of deleting either is not apparent unless both genes are deleted. We refer to this as nonhierarchical epistasis since neither gene is upstream of the other. Deleting a deactivated regulator gene has no effect on the target gene, making it impossible to identify a biological relationship when regulators are deactivated.

The remainder of Table 2.1b represents cases in which one or both genes do not affect the target gene. Expression traits supporting Model 8 (no significant terms) may represent target genes that do not lie downstream of *A* or *B*, and are uninformative. The result is one-to-many relationships between best-fit regression Models 1, 2, and 8 and their corresponding gene expression models. If the upstream gene of a hierarchical pair is turned off, we cannot know whether it is upstream or uninvolved.

Typically, expression is measured from thousands of genes simultaneously and we do not expect them all to be informative. Even with clear interpretations for each trait individually, there is a challenge interpreting all traits together. We examine the distribution of all traits. Among informative traits associated with a best-fit model, the majority may represent the underlying biological relationship between the deleted genes.

Validating the Two-Step Modeling Framework

Van Driessche et al. used *Dictyostellium discoideum* wild type (Van Driessche et al. 2002) and deletion mutant strains (Van Driessche et al. 2005) to infer hierarchical epistasis among genes of the protein kinase (PKA) pathway. Each strain's gene expression profile was measured using cDNA microarrays with a common reference over 24 hours. These data are well suited for testing our methods for two reasons. First, the epistatic relationships between the deleted genes already have been characterized experimentally. Secondly, the mutant strains are genetically identical at all loci except the few being studied, i.e. there is no variation in their genetic background.

The PKA pathway is associated with the developmental aggregation response to nutrient deprivation, which initiated midway through the time course. Data before and after aggregation were considered separately so we can clearly interpret the deletion effects in each signal state. The data represented fold-change on a logarithmic scale, which made the distribution of expression measurements approximately normal; we consider the implications of this in the discussion. We studied 1553 expression traits. The genes we used were measured in both experiments and differentially expressed in the wild type during aggregation (Van Driessche et al. 2002). Five deletion strains target genes of the protein kinase A (PKA) pathway that is involved in the response to starvation and activates aggregation. This provided three contrasts: *pufA*/*pkaC*, *pufA*/*yakA*, and *regA*/*pkaR*. Although there are ten possible contrasts for these five genes, only these three double mutants were generated, presumably because these are known direct relationships.

For each contrast, some traits supported each model (Figure 2). Additionally, large number of traits showed no deletion effects (i.e. support Model 8). At a significance threshold of $p < 0.01$, a majority of traits supported Model 8 for every contrast pre-aggregation (Figure S4) and for the *regA*/*pkaR* contrast post-aggregation. According to our interpretive models, Model 8 can indicate three possibilities. The first two are hierarchical relationships in which an upstream enhancing gene is turned off during aggregation. The last possibility is that the genes are uninvolved in the expression of the target and the deletions have no effect.

Since not all target genes are downstream of the PKA pathway, it is logical that the deletions have no effect on these genes. Similarly, the PKA pathway is invoked during aggregation and it follows that the deletions may affect expression only after aggregation has begun. We assume that the target genes supporting Model 8 are not downstream of the pathway, and that the majority of the remaining target genes reflect the relationship within the pathway. To test this assumption, we looked at the overlap between the expression traits supporting Model 8 for each contrast. We found that all of the expression traits supporting Model 8 for the *pufA*/*yakA* contrast also supported Model 8 for the other two contrasts. These traits strongly support the assumption that they are not downstream of the PKA pathway.

When we looked at these genes for both the *pufA*/*pkaC* and *pufA*/*yakA* contrasts, there was strong support for one model over all others post-aggregation. Not only did these models explain more traits post-aggregation, but the models also fit better. On average, the best-fit model explained over half of the expression variation ($R^2 \geq 0.5$, adjusted for degrees of freedom in the model) for traits in the *pufA*/*pkaC* and *pufA*/*yakA* contrasts, and for both contrasts the R^2 increased post-aggregation (t-test with $p < 0.0001$).

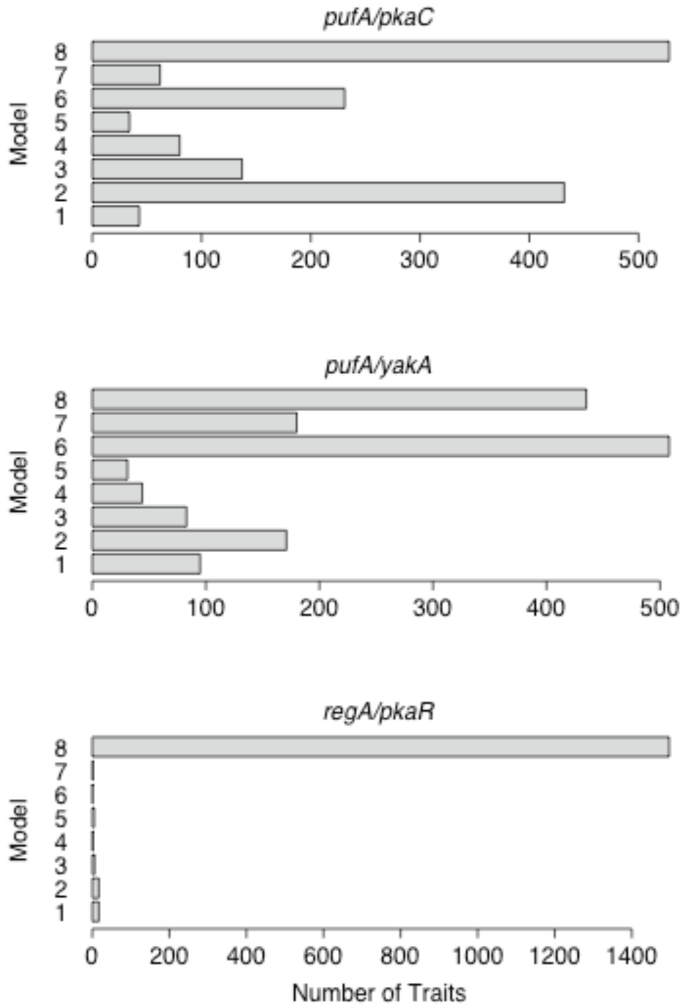


Figure 2.2 Post-aggregation distributions of best-fit models at p <0.01 significance thresholds

The frequency distribution of best-fit regression models can be interpreted as hierarchical relationships between genes. Model 8 corresponds to no deletion effects and is supported by a large number of traits in each contrast; these genes are likely not downstream of the deletions. The model supported by the majority of remaining traits is assumed to represent the true relationship.

For the *pufA*/*pkaC* contrast, Model 2 had the most support of the seven non-null models. Model 2 corresponds to two possible interpretations. The first is that *pkaC* is the downstream gene, that *pufA* is a repressor, and that the *pufA* is turned off in the presence of the aggregation signal. Alternately, we could interpret it to mean that only *pkaC* has an effect on the downstream targets and that *pufA* is unrelated. For the *pufA/yakA* contrast, Model 6 had the most support among non-null models. This model has a one-to-one correspondence to our interpretive models. It asserts that *yakA* is an upstream repressor of *pufA*, and that *yakA* is turned on at aggregation. These conclusions both agree with what has been determined previously about the roles these three genes play during development (Souza et al. 1999). *YakA* represses *pufA*, which then ceases to repress *pkaC*.

The *regA/pkaR* was problematic because almost all traits supported Model 8, the model with no effect terms. For the previous two cases, we assumed that these traits were not downstream of the pathway. Given this assumption, we could have concluded that *regA* and *pkaR* were not involved with aggregation. However, the other two contrasts had 435 and 528 traits supporting Model 8, while *regA/pkaR* has 1497. Because of this discrepancy, we suggest that some proportion of these genes support the hierarchical model corresponding to Model 8: that one gene is an enhancer of the other and is deactivated by aggregation. According to previously published results, *regA* and *pkaR* work together to repress *pkaC* pre-aggregation and are in fact deactivated post-aggregation (Shaulsky et al. 1998). This is consistent with the potential hierarchical relationship.

Because we are modeling nonadditive interactions, the logarithmic scale transformation on these data can potentially alter the results relative to untransformed data

(Frankel and Schork 1996; Lynch and Walsh 1998). To test this, we exponentiated the data and repeated our method. Despite dramatic changes to the shape of the data distribution, the resulting distribution of best-fit models agreed with the results presented above. Again, a majority of traits showed no deletion effects (i.e. support Model 8). Model 2 had the most support for the *pufA*/*pkaC* contrast, Model 6 had the most support for the *pufA*/*yakA* contrast, and Model 8 had near complete support for the *regA/pkaR* contrast using the post-aggregation data (Figure S5). Interestingly, this does not imply that each trait supports the same model regardless of the scale transformation. In fact, only 57% and 47% of traits support the same model with the untransformed data for the *pufA*/*pkaC* contrast and *pufA*/*yakA* contrast respectively. However, in both these cases the vast majority of changed traits support Model 8. This result amends our previous interpretation of the traits supporting Model 8; in addition to genes not downstream of the pathway, there may be some proportion of genes for which expression changes due to deletion is not detectable due to issues of scale. Fewer traits supported Model 8 using transformed data, suggesting that these data may be more informative using the logarithmic transformation.

Thus, in all three cases our best-fit regression models correspond to a set of interpretative models that includes the true relationship between the genes. Certain regression models have a one-to-many relationship with the interpretive models, but in these cases the number of candidate interpretive models is reduced to a few. Only one interpretation corresponds to Model 6, which makes the *pufA*/*yakA* contrast straightforward to describe. In evaluating *pufA*/*pkaC*, Model 2 corresponds to one hierarchical model and one single-gene model. Since the *pufA*/*yakA* contrast provides evidence that deleting *pufA* has an effect, the hierarchical model is a preferable interpretation to the *pkaC* only model. As we vary the significance threshold for model selection, our results are robust. The best-fit model among models 1-7 was the same for p-value thresholds from 0.05 to 0.001 (Figure S6). As the selection criterion becomes stricter we reject more effects as not significant, and more traits support Model 8.

Discussion

Measuring transcript abundance within a cell will remain a fundamental interest to biologists. Gene expression technologies have become popular over the past decade because of their ability to capture many genes simultaneously. Analyses that traditionally focused on a few genes now must be expanded to consider entire genomes. At this scale, the relationships between genes are of as much interest as the genes' individual effects. Many methods exist to infer gene networks or pathways from expression profiles (Bansal et al. 2007). Most of these require large datasets and result in large network diagrams that are difficult to interpret. These approaches are useful because they provide a genome scale view of transcription, and they are convenient because they can be applied to data from a variety of easily accessible sources.

However, there is a continuing need for experiments that allow us to infer pathways directly. The classical epistasis experiment we recount in our results (Van Driessche et al. 2005) is one such approach. Because it targets gene pairs directly, we can build pathways a relationship at a time. This local approach results in pathway diagrams that are easily comprehended and biologically relevant. Additionally, it associates genetic variation with expression variation. For these reasons, these types of experiments will be increasingly useful in constructing biological systems diagrams. While there are currently few experiments that measure expression in a genetically variable population, their number is increasing rapidly. Our motivation is to provide a conceptual framework in which these and related experiments can be interpreted. We have addressed the simplest genetically variable data structure for identifying epistasis, in which individuals vary at only two loci, but our ideas can be applied to a range of similar data.

Because expression data are continuous by nature, we must address them with quantitative methods. Regression analysis is a standard technique to relate continuous variables. Using a multiple regression model to estimate gene effects and interactions has several advantages. First, it allows us to consider information from all the deletion mutants and the wild type simultaneously. Additionally, it estimates an effect for each allele, allows for variance in allelic effects, and separates these effects from error variance. In a traditional epistasis analysis the double mutant is compared to each single mutant in a rule-based manner, and the two nearest trait values determine epistasis. In contrast to our method, this method does not take advantage of all the information from a given contrast, and it is difficult to distinguish signal from noise. Myriad sophisticated techniques exist for fitting multiple regression models, and these should be employed based on the distributional properties of particular data.

We consider individual expression traits rather than an expression profile. A gene expression model represents each trait, but we must infer the correct biological model through the results from the regression step. A corresponding regression model represents each possible gene expression model, but these relationships are not always one-to-one. Hierarchies in which an upstream gene is turned off by a signal are confounded with cases in which the gene has no effect. It makes sense that we cannot observe the effect of a deletion if the gene is already turned off in the wild type. Nonetheless, our framework was consistent with previous characterizations of the pathway in every case.

Scale transformations are common in genetics and genomics so that data meet statistical testing assumptions such as normality and homoscedascity (Lynch and Walsh 1998). Logarithmic transformations are ubiquitous in the literature for gene expression data such as those presented in our results. However, models with nonadditive interactions are subject to the scale of the data, and transformations can result in support for alternative models. This is a long-standing problem with describing epistasis for complex traits (Frankel and Schork 1996). Often it is difficult to know the most biologically appropriate scale, and the scale is instead often chosen arbitrarily based on the available measurement or statistical convenience. For gene expression traits the scale issue is even more complex. Since there are wide differences in the range of expression variation between genes, it is likely that no one scale will allow detection of the underlying biological interactions for all expression traits. The relationship between scale and epistasis is an area that demands further study, particularly in this era of genetics on biomolecular traits such as gene expression that have not been well studied in this context.

When we performed the same analysis on log-transformed and untransformed post-aggregation data, about half the traits supported a different best-fit model, yet the distribution of results led to the same conclusions regarding the underlying relationship between the deleted genes. This suggests our conclusions may be robust to scale effects that would affect single traits because they are based on the distribution of all traits. Those traits that are affected by scale trend toward having no detectable deletion effects with untransformed data. This further confounds the roughly one-third of traits supporting Model 8, which may also suggest an upstream enhancer or a trait truly unaffected by the deletions. While we do not discount scale effects, we assume most of these traits fit the last category because of the high percentage of these traits, the consistency of traits supporting Model 8 between contrasts, and the logic that deletions should affect only downstream genes. Whichever the case, these concerns make a strong argument for interpreting the distribution of results across expression traits. This contrasts with methods that consider all traits as an expression profile. These assume the profile as a whole supports one underlying pathway (Van Driessche et al. 2005).

Using our method, it is straightforward to interpret a range of experiments. The alleles being studied do not need to be null alleles, e.g. deletions. The same method could be applied to over-expressed genes, or any polymorphic locus. Additionally, the method can accommodate experiments investigating multiple loci and higher order interactions. Three-way and four-way epistasis models follow from the same principles as the two-way models we present. The regression model is very flexible and easy to extend by adding a parameter for each locus plus interaction terms. Connecting these statistical models to biological models follows the same process we have illustrated. The strengths of our approach are particularly apparent in multi-locus models because we provide a means for estimating effects using the entire population of mutants simultaneously. The number of genotypes increases by a power of two for each additional gene included in the experiment; with a three-locus experiment having eight genotypes. As the number of necessary pair-wise comparisons increases, they will contain more undetected error and become more difficult to interpret. Environmental effects can also be included in the model at the expense of increased complexity in interpretation. We considered observations before aggregation and after aggregation separately in our example for simplicity.

By proceeding to add genetic and environmental complexity, it is apparent how the classical epistasis framework connects to the quantitative genetic paradigm. An additional benefit of our method is that it enables comparisons between any population-based expression analyses. Whether study populations consist of deletion mutants, experimentally designed crosses, inbred lines, chromosome substitution strains, or natural populations, each expression trait is the same. For this reason, comparing these results is highly desirable. Estimating the allelic effects and interactions for each expression trait allows direct comparison across a variety of genetic backgrounds. By embracing a common interpretive framework to a range of experiments that use gene expression as a trait, we can integrate results and form clearer insights into the genetic control of systems.

Data and Methods

Dictyostellium gene expression data

We used data originally presented by Van Driessche et al. We use data from *Dictyostellium discoideum* wild type (Van Driessche et al. 2002) and eight deletion mutant strains ($pufA^{-}$, $pkaC^{-}$, $pufA^{-}pkaC^{-}$, $yakA^{-}$, $pufA^{-}yakA^{-}$, $regA^{-}$, $pkaR^{-}$, $regA^{-}pkaR^{-}$) (Van Driessche et al. 2005). They measured each strain's gene expression profile over a time course using cDNA microarrays and a common reference that was pooled from all time points. Expression was measured thirteen times over 24 hours and captured the developmental aggregation response to nutrient deprivation, which initiated midway through the time course. We grouped observations before (hours 0,2,4,6) and after (hours 14,16,18,20) aggregation. Expression at these time points is highly correlated (Figure 2 in (Van Driessche et al. 2002)) and consistent with the regulatory changes previously reported. This data pooling increased the sample size for our regression analysis. Observations during the transitional period (hours 8,10, and 12) were disregarded, as were observations in the late stages of development that were less correlated (hours 22 and 24). The data represented fold-change on a logarithmic scale. We studied 1553 genes that were measured in both experiments and differentially expressed in the wild type during aggregation (Van Driessche et al. 2002).

Regression analysis

We fit models in the R statistical environment (R Development Core Team). Stepwise backwards selection entails fitting a fully parameterized model, then eliminating model terms that do not meet a specified significance threshold. The model is refit with the remaining terms until no further terms can be dropped.

References

Auger, D.L., A.D. Gray, T.S. Ream, A. Kato, E.H. Coe, Jr., and J.A. Birchler. 2005. Nonadditive gene expression in diploid and triploid hybrids of maize. *Genetics* **169:** 389-397.

Avery, L. and S. Wasserman. 1992. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* **8:** 312-316.

Bansal, M., V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo. 2007. How to infer gene networks from expression profiles. *Mol Syst Biol* **3:** 78.

Brem, R.B. and L. Kruglyak. 2005. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A* **102:** 1572-1577.

Brem, R.B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296:** 752-755.

Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M.T. Pletcher, T. Wiltshire, A.I. Su, E. Vellenga, J. Wang, K.F. Manly, L. Lu, E.J. Chesler, R. Alberts, R.C. Jansen, R.W. Williams, M.P. Cooke, and G. de Haan. 2005. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37:** 225-232.

Chesler, E.J., L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H.C. Hsu, J.D. Mountz, N.E. Baldwin, M.A. Langston, D.W. Threadgill, K.F. Manly, and R.W. Williams. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* **37:** 233-242.

Frankel, W.N. and N.J. Schork. 1996. Who's afraid of epistasis? *Nat Genet* **14:** 371-373.

Gibson, G., R. Riley-Berger, L. Harshman, A. Kopp, S. Vacha, S. Nuzhdin, and M. Wayne. 2004. Extensive sex-specific nonadditivity of gene expression in Drosophila melanogaster. *Genetics* **167:** 1791-1799.

Gibson, G. and B. Weir. 2005. The quantitative genetics of transcription. *Trends Genet* **21:** 616-623.

Hughes, T.R., M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109-126.

Jansen, R.C. and J.P. Nap. 2001. Genetical genomics: the added value from segregation. *Trends Genet* **17:** 388-391.
Li, J. and M. Burmeister. 2005. Genetical genomics: combining genetics with gene expression analysis. *Hum Mol Genet* **14 Spec No. 2:** R163-169.

Li, R., S.W. Tsaih, K. Shockley, I.M. Stylianou, J. Wergedal, B. Paigen, and G.A. Churchill. 2006. Structural model analysis of multiple quantitative traits. *PLoS Genet* **2:** e114.

Lynch, M. and B. Walsh. 1998. *Genetics and the Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Phillips, P.C. 1998. The language of gene interaction. *Genetics* **149:** 1167-1171.

R Development Core Team (2007) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

Schadt, E.E., S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, P.S. Linsley, M. Mao, R.B. Stoughton, and S.H. Friend. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422:** 297-302.

Shaulsky, G., D. Fuller, and W.F. Loomis. 1998. A cAMP-phosphodiesterase controls PKA-dependent differentiation. *Development* **125:** 691-699.

Souza, G.M., A.M. da Silva, and A. Kuspa. 1999. Starvation promotes Dictyostelium development by relieving PufA inhibition of PKA translation through the YakA kinase pathway. *Development* **126:** 3263-3274.

St Onge, R.P., R. Mani, J. Oh, M. Proctor, E. Fung, R.W. Davis, C. Nislow, F.P. Roth, and G. Giaever. 2007. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39:** 199-206.

Storey, J.D., J.M. Akey, and L. Kruglyak. 2005. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3:** e267.

Tong, A.H., M. Evangelista, A.B. Parsons, H. Xu, G.D. Bader, N. Page, M. Robinson, S. Raghibizadeh, C.W. Hogue, H. Bussey, B. Andrews, M. Tyers, and C. Boone. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294:** 2364-2368.

Tong, A.H., G. Lesage, G.D. Bader, H. Ding, H. Xu, X. Xin, J. Young, G.F. Berriz, R.L. Brost, M. Chang, Y. Chen, X. Cheng, G. Chua, H. Friesen, D.S. Goldberg, J. Haynes, C. Humphries, G. He, S. Hussein, L. Ke, N. Krogan, Z. Li, J.N. Levinson, H. Lu, P.

Menard, C. Munyana, A.B. Parsons, O. Ryan, R. Tonikian, T. Roberts, A.M. Sdicu, J. Shapiro, B. Sheikh, B. Suter, S.L. Wong, L.V. Zhang, H. Zhu, C.G. Burd, S. Munro, C. Sander, J. Rine, J. Greenblatt, M. Peter, A. Bretscher, G. Bell, F.P. Roth, G.W. Brown, B. Andrews, H. Bussey, and C. Boone. 2004. Global mapping of the yeast genetic interaction network. *Science* **303:** 808-813.

Van Driessche, N., J. Demsar, E.O. Booth, P. Hill, P. Juvan, B. Zupan, A. Kuspa, and G. Shaulsky. 2005. Epistasis analysis with global transcriptional phenotypes. *Nat Genet* **37:** 471-477.

Van Driessche, N., C. Shaw, M. Katoh, T. Morio, R. Sucgang, M. Ibarra, H. Kuwayama, T. Saito, H. Urushihara, M. Maeda, I. Takeuchi, H. Ochiai, W. Eaton, J. Tollett, J. Halter, A. Kuspa, Y. Tanaka, and G. Shaulsky. 2002. A transcriptional profile of multicellular development in Dictyostelium discoideum. *Development* **129:** 1543-1552.

Genotype	Upstream Gene	Gene Action	Target Gene Expression	Regression Model
$A^{+}B^{+}$	ON	$\overset{\text{ON}}{A} \dashv \underset{\text{OFF}}{B} \dashv \overset{\text{ON}}{X}$	μ	$\mu + \beta_A + \beta_I$
$A^{-}B^{+}$		$\cancel{A} \dashv \overset{\text{ON}}{B} \dashv \underset{\text{OFF}}{X}$	$\mu + \beta_A$	
$A^{+}B^{-}$		$\overset{\text{ON}}{A} \dashv \cancel{B} \dashv \overset{\text{ON}}{X}$	μ	
$A^{-}B^{-}$		$\cancel{A} \dashv \cancel{B} \dashv \overset{\text{ON}}{X}$	μ	
$A^{+}B^{+}$	OFF	$\underset{\text{OFF}}{A} \dashv \overset{\text{ON}}{B} \dashv \underset{\text{OFF}}{X}$	μ	$\mu + \beta_B$
$A^{-}B^{+}$		$\cancel{A} \dashv \overset{\text{ON}}{B} \dashv \underset{\text{OFF}}{X}$	μ	
$A^{+}B^{-}$		$\underset{\text{OFF}}{A} \dashv \cancel{B} \dashv \overset{\text{ON}}{X}$	$\mu + \beta_B$	
$A^{-}B^{-}$		$\cancel{A} \dashv \cancel{B} \dashv \overset{\text{ON}}{X}$	$\mu + \beta_B$	

Figure 2.3 Modeling the relationship *A is an upstream repressor of B*, *which represses a target gene*.

Genotype	Upsteam Gene	Gene Action	Target Gene Expression	Regression Model
$A^{+}B^{+}$	ON	$\overset{ON}{A} \rightarrow \overset{ON}{B} \dashv \underset{OFF}{X}$	μ	$\mu + \beta_A + \beta_B + \beta_I$
$A^{-}B^{+}$		$\cancel{A} \rightarrow \underset{OFF}{B} \dashv \overset{ON}{X}$	$\mu + \beta_A$	
$A^{+}B^{-}$		$\overset{ON}{A} \rightarrow \cancel{B} \dashv \overset{ON}{X}$	$\mu + \beta_B$	
$A^{-}B^{-}$		$\cancel{A} \rightarrow \cancel{B} \dashv \overset{ON}{X}$	$\mu + \beta_B$	
$A^{+}B^{+}$	OFF	$\underset{OFF}{A} \rightarrow \underset{OFF}{B} \dashv \overset{ON}{X}$	μ	μ
$A^{-}B^{+}$		$\cancel{A} \rightarrow \underset{OFF}{B} \dashv \overset{ON}{X}$	μ	
$A^{+}B^{-}$		$\underset{OFF}{A} \rightarrow \cancel{B} \dashv \overset{ON}{X}$	μ	
$A^{-}B^{-}$		$\cancel{A} \rightarrow \cancel{B} \dashv \overset{ON}{X}$	μ	

Figure 2.4 Modeling the relationship ***A is an upstream enhancer of B, which represses a target gene.***

Genotype	Upstream Gene	Gene Action	Target Gene Expression	Regression Model
A^+B^+	ON	$\overset{ON}{A} \rightarrow \overset{ON}{B} \rightarrow \overset{ON}{X}$	μ	$\mu + \beta_A + \beta_B + \beta_I$
A^-B^+		$\cancel{A} \rightarrow \underset{OFF}{B} \rightarrow \underset{OFF}{X}$	$\mu + \beta_A$	
A^+B^-		$\overset{ON}{A} \rightarrow \cancel{B} \rightarrow \underset{OFF}{X}$	$\mu + \beta_B$	
A^-B^-		$\cancel{A} \rightarrow \cancel{B} \rightarrow \underset{OFF}{X}$	$\mu + \beta_B$	
A^+B^+	OFF	$\underset{OFF}{A} \rightarrow \underset{OFF}{B} \rightarrow \underset{OFF}{X}$	μ	μ
A^-B^+		$\cancel{A} \rightarrow \underset{OFF}{B} \rightarrow \underset{OFF}{X}$	μ	
A^+B^-		$\underset{OFF}{A} \rightarrow \cancel{B} \rightarrow \underset{OFF}{X}$	μ	
A^-B^-		$\cancel{A} \rightarrow \cancel{B} \rightarrow \underset{OFF}{X}$	μ	

Figure 2.5 Modeling the relationship *A is an upstream enhancer of B, which enhances a target gene.*

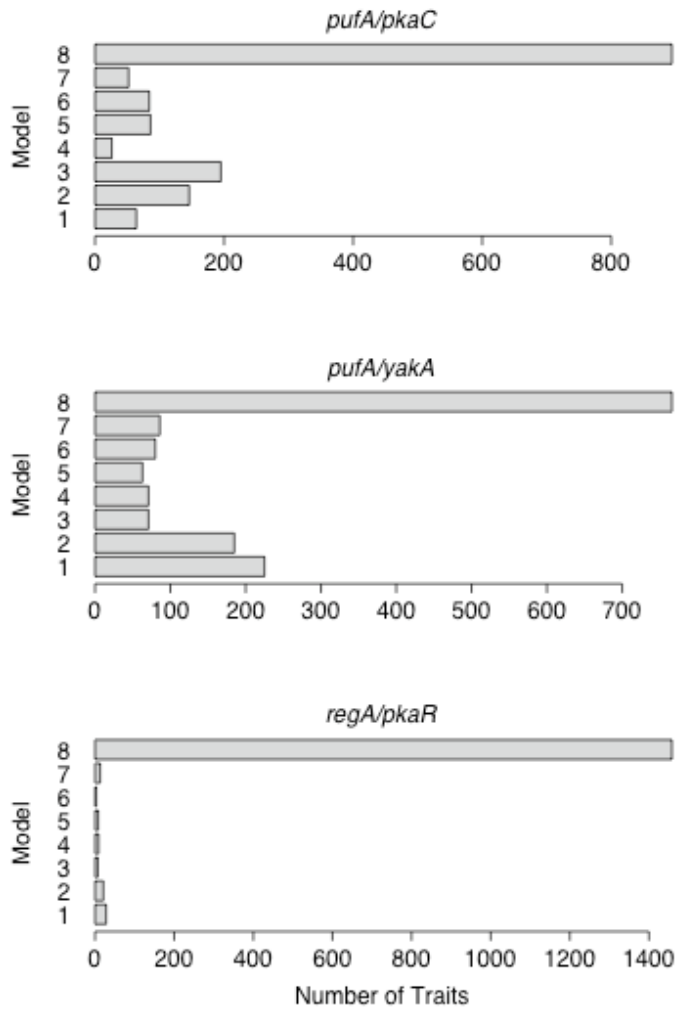


Figure 2.6 Distribution of best-fit models pre-aggregation.

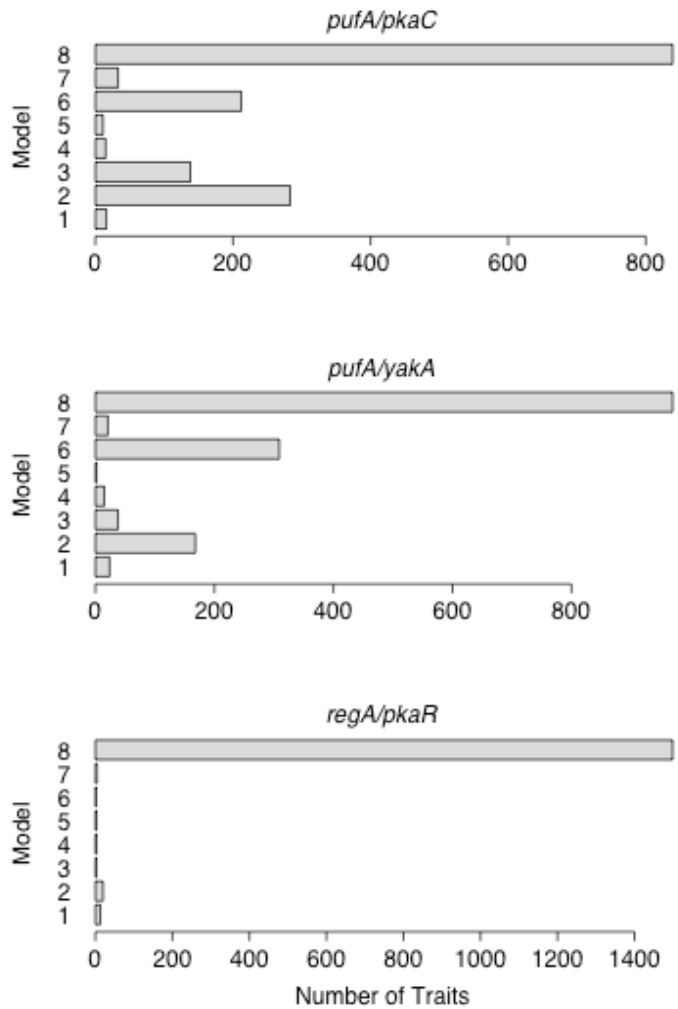
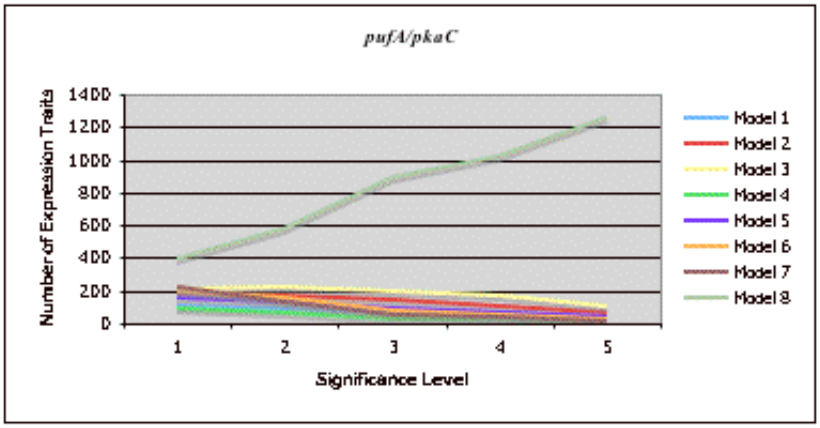
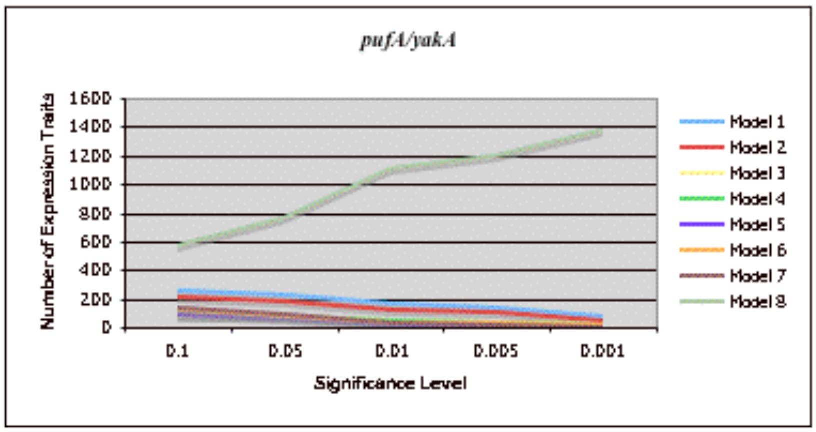


Figure 2.7 Distribution of best-fit models post-aggregation (untransformed data)





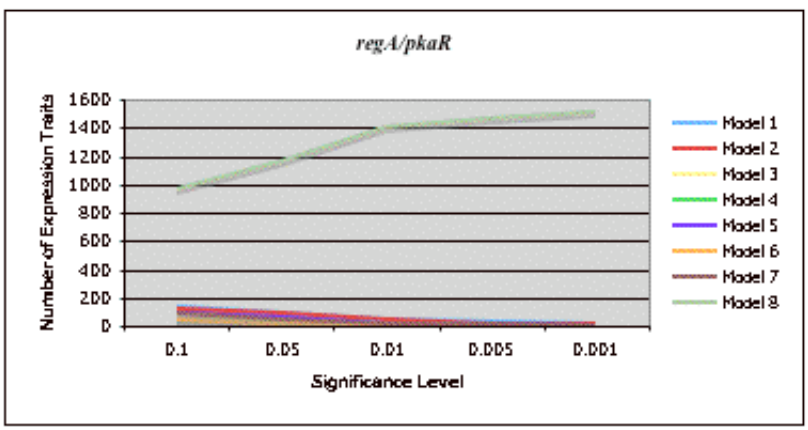
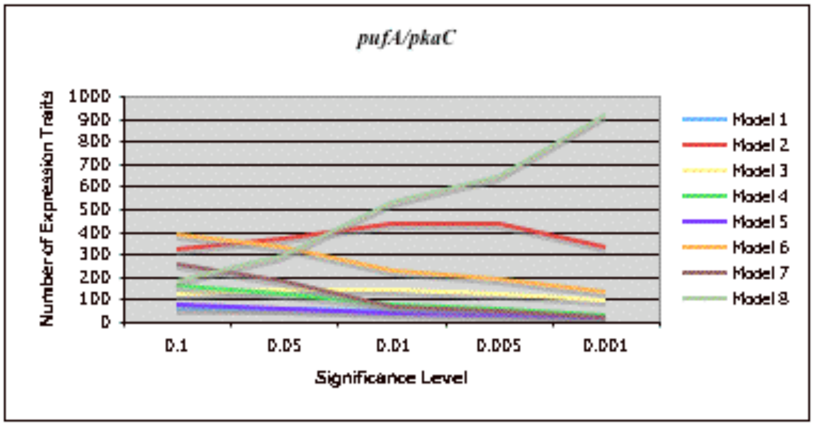
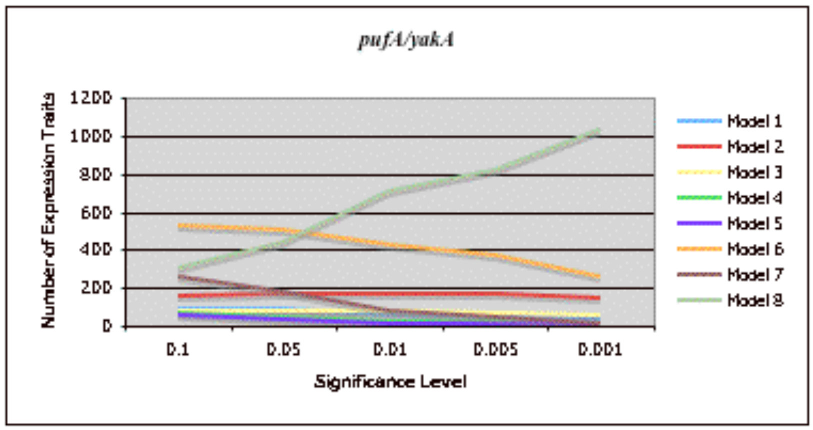


Figure 2.8 Distribution of best-fit models at varying significance thresholds pre-aggregation.





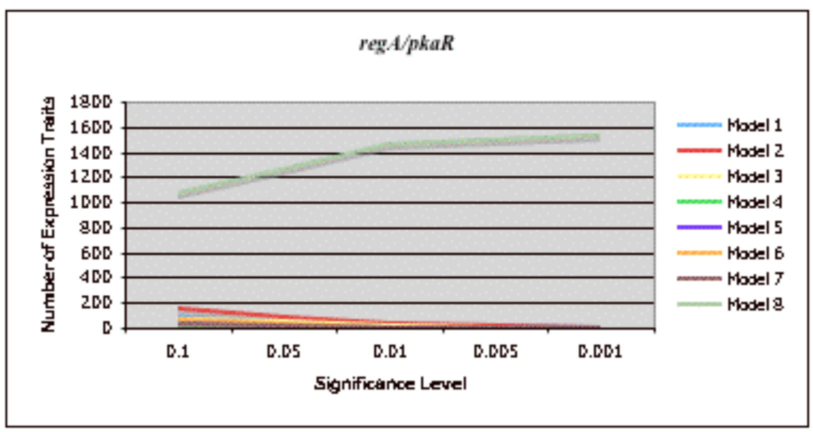


Figure 2.9 Distribution of best-fit models at varying significance thresholds post-aggregation.

Chapter 3

An Expression QTL Survey of Yeast Pathways

David L. Aylor and Zhao-Bang Zeng

Abstract

Expression QTL mapping is among the latest strategies for untangling complex traits. Since hundreds of expression traits can be measured simultaneously, one goal is to assess patterns of co-regulation among traits. If a common biological process relates a set of trait genes, their shared QTLs may be regulators for that process. One approach to finding such regulators is to define related gene sets a priori and then evaluate these sets for shared QTLs. We survey QTL results for all yeast pathways in the Saccharomyces Genome Database and report the prevalence of shared QTLs, the degree of sharing between pathway genes, and the relative proportion of cis- versus trans- pathway regulation. Genes within yeast pathways often share QTLs, but QTLs are rarely shared by a majority of the pathway genes. Hotspot QTL are linked to the expression of genes in many different pathways. We also present a web-based browser for visualizing our results.

Introduction

Background

Linking genetic variation to changes in gene expression is among the latest strategies for untangling complex traits. Such methods are a new application of existing techniques for mapping quantitative trait loci (QTL). QTL mapping is a genome-wide inference of the relationship between quantitative phenotypes and genotypes of QTL. In expression QTL (eQTL) mapping, some measure of transcript abundance is substituted for a traditional phenotypic measure. However, we must make several special considerations for eQTL mapping.

Microarray technology has enabled the expression of thousands of genes to be measured simultaneously. Typically, linkage mapping for each expression trait is done individually. However, the results from all traits are often interpreted together. The basic assumption in many eQTL studies is that gene expression measurements capture an intermediate level between genotype and a complex phenotype. Implicit to this assumption is that gene expression traits are most meaningful in groups that work together biologically. The challenge is to group traits appropriately to explain the molecular basis of the phenotype. Even when no particular phenotype is under study, we assume that regulatory networks govern the expression of any given gene and expression traits are best studied in groups for that reason. Methods for grouping expression traits and analyzing groups of traits are the subject of much of the current eQTL literature.

The complete set of all expression traits is the largest group for any experiment. Two genome-wide patterns have been apparent in all eQTL analyses reported to date. First, some portion of traits have been linked to a *cis*-QTL, which reflects variation located in the same physical location as the gene itself. This is interpreted as a polymorphism in a gene's coding region or promoter region that regulates that gene's expression. QTLs that regulate traits that are not collocated are called *trans*-QTL. The second common pattern is multiple *trans*-QTLs that are shared by many traits, sometimes called hotspots. These two features have become starting points for how we describe expression QTL results and central to how we interpret them.

Early studies suggested that shared *trans*-QTLs affect groups of genes having similar functions (Brem et al. 2005; Brem et al. 2002; Yvert et al. 2003), and they have been interpreted as controlling regulatory networks related to these functions. More recent reports (West et al. 2007) have suggested that while many functions are affected by the hotspot QTL, no functions are overrepresented relative to what is expected by chance. A complementary approach is to group genes of shared function and map QTLs shared within the group.

Two types of methods have been employed to group expression traits. The first type groups traits based on data or results from the cross itself, generally by some clustering method. An intuitive approach to this is to group traits with shared QTL. Brem et al (2002) identified eight shared *trans*- regulators and grouped the traits accordingly. Each group was characterized by shared annotations in the yeast genome database, and the trait with *cis*-linkage was inferred to be the gene harboring a causative polymorphism. In a follow-up paper, Yvert et al (2003) used correlation-based k-means clustering to group similar traits based on their expression profiles. This approach is independent of the linkage analysis, but nonetheless groups based solely on data generated from the experimental cross. The clusters also were enriched for genes with shared annotations. Yvert et al (2003) was also the first appearance of linkage analysis of a composite trait derived from multiple grouped expression traits. The expression trait values within each cluster were averaged for every individual, and this average was treated as a trait. Composite trait mapping is now common in eQTL literature. A similar clustering approach was used to group expression traits in mice (Ghazalpour et al. 2006). QTLs linked to grouped trait averages were called module QTLs (mQTLs).

Alternately, sets of genes can be defined *a priori*. Kliebenstein et al (2006) mapped QTLs for twenty networks in *Arabidopsis*. These networks were largely inferred from coordinate expression, but separate experiments determined these groupings, not expression data from the QTL mapping population. In a conceptually similar approach, the mammalian phenotype ontology (MPO) was used to systematically pair genes associated with similar phenotypes (Bao et al. 2006). These gene pairs were compared to eQTL results to uncover potential regulatory relationships. Human eQTLs were linked to Gene Ontology terms, canonical pathways, and disease using Ingenuity pathway analysis software (Wessel et al. 2007). Myriad data sources exist for grouping genes, and each provides a new lens through which to interpret linkage results.

The composite trait or trait averaging strategy can be used with previously defined gene groups or groups clustered from the data. Composite trait mapping assumes that a few shared QTLs affect most or all of the individual traits in a group, and that pooling these traits will increase the ability to find these QTLs. Two issues potentially complicate these analyses. First, QTLs having a large effect on a single trait within the group can be inferred as module QTL. On the other hand, QTLs of smaller effect are likely to go undetected if they are not shared. Averaging trait values has been prevalent in the literature, but principal components have also been mapped as composite traits.

Motivation

In a series of papers, Kruglyak and colleagues presented expression QTL mapping results from yeast (Brem et al. 2005; Brem et al. 2002; Yvert et al. 2003). Additionally, yeast has a well-characterized molecular biology and many bioinformatics resources are available. As such, several natural *a priori* groups of gene sets exist for yeast. However, no study to date has used such an approach with these data. Our survey begins with the mapping results from an eQTL analysis. We chose to use Multiple Interval Mapping results from Zou and Zeng (2008) based on familiarity and MIM's power to detect multiple QTL models. However, this choice is arbitrary and our method can be applied to linkage results produced by a range of methods.

We survey all yeast genetic pathways in the Saccharomyces Genome Database, with a focus on QTL shared within pathways. Because we used results from individually mapped expression traits, we captured both shared QTLs and those linked to only one trait. We report the proportion of shared QTLs and the degree of sharing between QTLs. Module QTLs presumably represent individual QTLs shared by multiple related traits (Ghazalpour et al. 2006), but we avoid that term because it has been previously defined in the context of a composite (averaged) trait. We instead refer to such QTLs as putative pathway regulators.

However, this approach has important implications for composite trait mapping, since the degree of QTL sharing determines its success.

Data and Methods

Yeast Linkage Data and Analysis

A wild isolate was crossed with a standard laboratory strain, and expression was measured on 112 haploid segregants (Brem et al. 2005). Gene expression was measured for 6216 trait genes and 2957 genetic markers were typed. Expression measurements in the segregants were referenced against the wild parent strain. In a controlled environment, one quarter of the expression traits measured were differentially expressed between the parent strains at $P < 0.005$ (Brem et al. 2002).

In a single marker analysis (SMA), 2984 traits had at least one QTL (FDR = 0.05). Approximately a third of these linkages were *cis*-acting (Brem et al. 2002). Fitting multiple QTL models suggested that the majority of highly heritable traits had more than one QTL. Zou and Zeng (2008) applied Multiple Interval Mapping (MIM) to the same data. MIM uses a sequential procedure to find the best genetic model for each trait. Whereas SMA tests each marker separately, MIM conditions on QTLs already in the model as it searches for additional QTLs. This provides power to detect additional QTLs with smaller effects. *Cis*-QTLs tend to be of larger effect than *trans*-QTLs, thus more *trans*-QTLs are detected with MIM. We use the MIM results for our analysis. The start and end of the physical QTL regions are determined by a 1.5 log odds (LOD) interval from the QTL peak.

Saccharomyces Genome Database Pathways

The Saccharomyces Genome Database (SGD) contains a wide variety of information on the molecular biology of yeast. It primarily contains 6608 genes or predicted genes (open reading frames) and their annotations, but also has complementary data (Hong et al. 2005). The Yeast Biochemical Pathways database was built using the Pathway Tools PathoLogic module, software developed at SRI Laboratories (Karp et al. 2002). It is then manually curated and corrected based on the relevant literature. It contains reactions, enzymes, genes, and compounds. We used the gene list for each pathway as an *a priori* grouping of expression traits.

There are 135 pathways in the database. Only 468 genes of the 6608 total are found in the pathway database, meaning that the majority of known genes have not yet been associated with a specific pathway. Nonetheless, we expect pathway genes to be closely related biologically. As with any study using *a priori* data to group genes, our conclusions are dependent on this assumption.

On average, five genes are associated with each pathway. The pathway with the largest gene list has 22 genes. Sixteen pathways have only one gene listed and an additional 21 have only two genes.

Visualizing Results

In order to visualize QTL distributions within and across pathways, we created a variety of graphs using Support Vector Graphics (SVG). Each is a variation on the popular plot found in most expression QTL literature, which features expression traits on the y-axis and

QTL physical location on the x-axis. First, we created the standard plot for just the 468 genes found in the SGD Pathway database. Next, we created a plot for each of the 135 yeast pathways (two examples are shown in Figure 3.1). These plots are a resource for additional characterization of individual pathways. Lastly, we provide a heat map summarizing QTL sharing (Figure 3.2). All plots are available at http://statgen.ncsu.edu/aylor.

Results and Discussion

Summary of Pathway QTL Distributions

The QTL distribution for the 468 expression traits in the pathway database closely resembles the QTL distribution over all traits. Seventy-three percent of traits have at least one QTL and seventeen percent of those traits have *cis*-QTL. This compares to 54 percent of 6216 traits measured overall having at least one QTL, and 22 percent of those being *cis*-. The 468 traits linked to 573 QTLs, which matches the estimate of slightly more than one QTL per trait in the complete data. In other words, the expression traits found in the pathway database constitute a representative sample of the entire data set.

Ninety-five percent of pathways contain at least one trait that is linked to one or more QTLs with an average of slightly more than four per pathway. Half of these pathways have at least one shared QTL. *Cis*-QTLs are found in 60 pathways. We asked if any pathway had more *cis*-QTLs than expected by chance. To test this, we treated the number of *cis*-QTLs per pathway as a binomial random variable distributed with n pathway genes and success probability equal to the proportion of all traits with *cis*-QTLs. We conclude that *cis*-QTLs are randomly distributed between pathways.

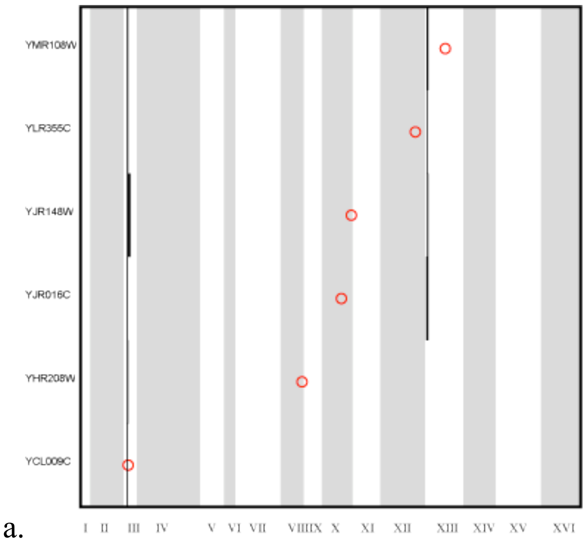
These results suggest that the yeast cross has captured regulatory variation affecting a broad range of biological functions and that some regulators are indeed affecting groups of related genes. However, only 15 percent of traits share a QTL with a member of the same pathway. The discrepancy between the high incidence of sharing overall and the paucity of sharing within individual pathways bears further examination. The implication is that few traits that share each QTL. There was no consistent pattern between individual pathways. Within the valine biosynthesis pathway, all QTLs were shared by a majority of traits (Figure 3.1a). However, the tryptophan degradation pathway had several QTLs but almost no sharing among twelve traits (Figure 3.1b). We asked which pattern was prevalent over all pathways.

Finding Pathway Regulators

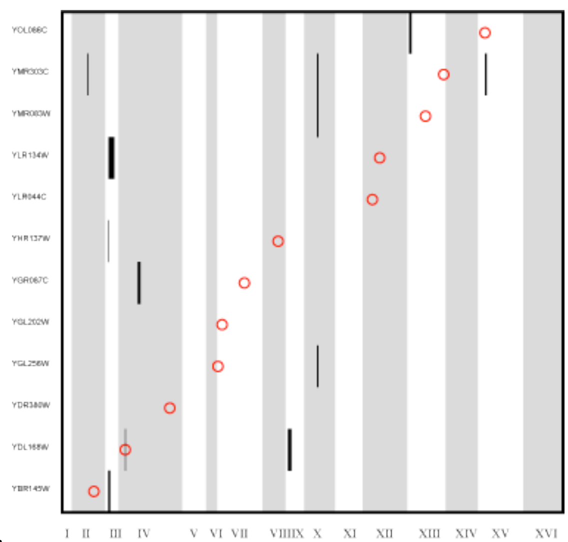
To report on the degree to which individual QTLs are shared within pathways, we eliminated pathways with two or fewer genes. No sharing is possible if a pathway only has one gene, and pathways with two genes make the degree of sharing hard to assess since a QTL is shared among all genes or none. This left 486 QTLs linked to 98 pathways. Twenty-six percent (126) of these QTLs are shared. Importantly, only six percent of shared QTLs are *cis*-. This is evidence that pathway regulation is not due to downstream effects of *cis*-regulated pathway genes as some have suggested. Instead, this observation supports a model in which *trans*- acting factors initiate changes in gene expression.

We calculated the percentage of traits in each pathway affected by each shared QTL. The median degree of sharing was 33 percent. Because of the small size of SGD pathways, this means that more than half of shared QTLs affect only two expression traits. This explains the discrepancy introduced in the previous section. Over all pathways, 96 percent of QTLs affect fewer than 60 percent of traits within the pathways to which they are linked (Figure 3.2). Since extensive sharing is so rare, we conclude that most pathways are not regulated as units in this cross. Individual pathway genes appear to be regulated independently and have few downstream effects on expression of other genes in the same pathway.

Thirty QTLs regulated greater than 60 percent of pathway traits. We consider these QTL as putative pathway regulators (Table 3.1). These QTLs affect 23 pathways, with seven pathways having two QTLs regulating a high percentage of traits.



a.



b.

Figure 3.1 Contrasting Pathway eQTL Plots
Two pathways are shown above. Each row along the y-axis represents an expression trait corresponding to a gene in the pathway. The x-axis is QTL location. Black bars are QTL linked to that trait. Red circles mark the physical location of the trait gene. a. Valine Biosynthesis has two clear pathway QTL. The QTL on Chromosome III affects all the traits in the pathway and is *cis-* for *ILV6* (YCL009C). The associated gene product catalyzes the first step of valine biosynthesis, so the colocalized *trans*-QTL may be interpreted as downstream effects of the *cis*-QTL. b. The Tryptophan Degradation pathway is linked to several QTL but almost none are shared.

Pathway Enrichment of trans-Hotspots

We next asked if the putative pathway regulator QTLs were unique to individual pathways or regulated groups of related pathways. Nine biosynthesis pathways all have a putative regulator QTL on chromosome III. The affected traits are not the same in every pathway, suggesting a higher order control of biosynthesis. However, this same region on chromosome III is also the largest hotspot region when considering all traits independently and is linked to 87 traits in 64 different pathways.

Of the thirty putative pathway regulators, twenty-five are located in the five largest genome-wide hotspot regions. Since these regions affect so many traits, this presents the question of whether the high percentage of traits affected is due to chance. To test this, we treated the number of linked traits per pathway as a binomial random variable distributed with *n* pathway genes and success probability equal to the proportion of all traits from the SGD pathway database with linkages to that QTL region. We calculated a p-value by summing the probabilities of the observed data and more extreme data (additional linkages with pathway genes).

$$P(K \geq k) = \sum_{k}^{n} \binom{n}{k} {p_Q}^{k} (1 - p_Q)^{n-k}$$

n number of traits in pathway

k number of pathway traits linked to QTL Q

p_Q proportion of all traits linked to QTL Q

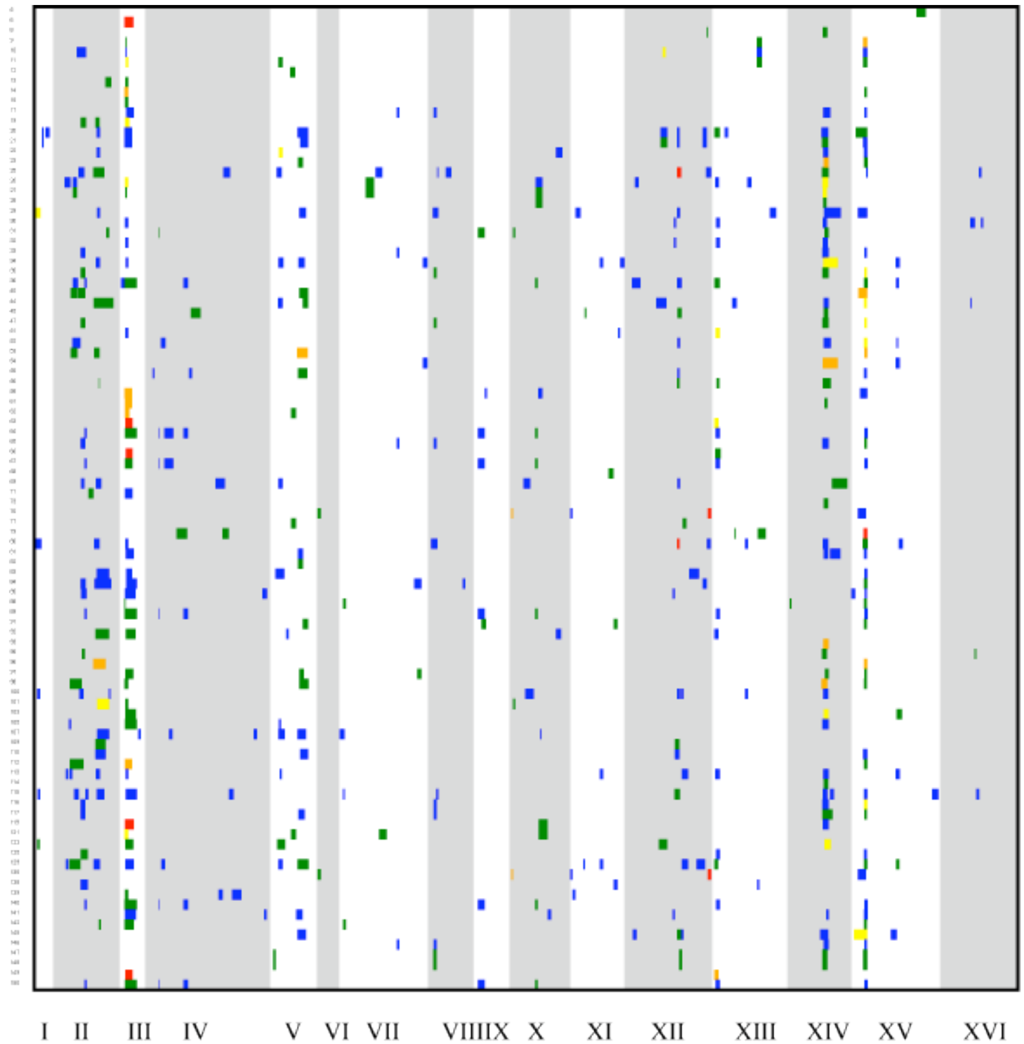


Figure 3.2 Most QTLs are shared by Less Than 40% of Pathway Genes
Ninety-eight pathways of at least three genes each are listed along the y-axis above. The x-axis is the location of QTL. Each row illustrates the QTLs linked to traits in that pathway. A QTL affecting one or few genes would appear as a blue bar, while warmer colors illustrate a QTL that controls a high percentage of pathway genes. About half of pathways have at least one shared QTL. However, these QTL are typically shared by only a few expression traits. In the plot above, 30 of 486 QTL are linked to over 60% of the expression traits within a pathway (red or orange bars). These putative pathway regulators affect 23 pathways.

On chromosome III, five of nine biosynthesis pathways were highly significant ($P < 0.0035$), as was the sulfate assimilation pathway. However, several putative pathway regulators were rejected with a $P < 0.01$ threshold (Table 3.1), especially for smaller pathways. None of the previously suggested pathway regulators on chromosomes II or XV were significant at that level. However, two pathways were overrepresented in the hotspot QTLs despite linkages with fewer than 60% of traits in those pathways. Eight of fifteen gluconeogenesis genes were linked to the hotspot on chromosome XIV ($P < 1.5 \times 10^{-4}$) and eight of nineteen TCA cycle gene were linked to the hotspot on chromosome II ($P < 3 \times 10^{-3}$). Only seven putative regulators are significant with a very conservative Bonferroni correction ($P < 4 \times 10^{-3}$).

Linkage analysis is limited in resolution and consequently QTLs often span large physical regions containing many genes. We cannot exclude the possibility that multiple causative polymorphisms underlie hotspot QTL and that each affects a subset of expression traits. For instance, the linkages observed in valine biosynthesis pathway may indeed be due to the effects of a *cis*- polymorphism in the upstream gene *ILV6*, yet be unrelated to the linkages to expression traits in other pathways.

Table 3.1 Putative Pathway Regulator QTL
Thirty QTLs were linked to 60 percent or more of the traits within a pathway. However, many of these QTL co-locate with genome-wide hotspots. For these shared QTL, we tested each pathway to see if the number of pathway traits linked to that QTL were greater than expected by chance. Two pathways were overrepresented for linkages to hotspot QTL despite fewer than 60 percent of traits being linked.

Pathway Definition	Traits in Pathway	QTL Chromosome	Traits Sharing QTL (%)	P-value
polyamine degradation	3	II	66.7	0.056
TCA cycle, aerobic respiration	19	II	42.1	0.003
sulfate assimilation pathway II	6	III	100.0	4.14×10^{-5}
valine biosynthesis	6	III	100.0	4.14×10^{-5}
leucine biosynthesis	6	III	100.0	4.14×10^{-5}
isoleucine biosynthesis	7	III	85.7	2.44×10^{-4}
arginine biosynthesis	7	III	85.7	2.44×10^{-4}
serine biosynthesis	4	III	75.0	0.022
chorismate biosynthesis	4	III	75.0	0.022
histidine biosynthesis	7	III	71.4	0.0033
homoserine biosynthesis	3	III	66.7	0.091
homoserine methionine biosynthesis	3	III	66.7	0.091
glycogen catabolism	3	V	66.7	9.6×10^{-3}
m-cresol degradation	6	X	66.7	1.24×10^{-6}
toluene degradation, via catechol	6	X	66.7	1.24×10^{-6}
mevalonate pathway	8	XII	87.5	6.35×10^{-7}
ergosterol biosynthesis	14	XII	85.7	5.91×10^{-11}
m-cresol degradation	6	XII	66.7	0.0012
toluene degradation, via catechol	6	XII	66.7	0.0012
valine biosynthesis	6	XIII	66.7	1.31×10^{-4}
glycine degradation	5	XIII	60.0	1.58×10^{-3}
phospholipid biosynthesis	3	XIV	66.7	0.042
proline biosynthesis	3	XIV	66.7	0.042
glycolysis	14	XIV	64.3	7.69×10^{-6}
gluconeogenesis	15	XIV	53.3	1.6×10^{-4}
asparagine biosynthesis	4	XV	75.0	0.013
polyamine degradation	3	XV	66.7	0.064
glutamate degradation I	3	XV	66.7	0.064
methylglyoxal catabolism	3	XV	66.7	0.064
glycogen catabolism	3	XV	66.7	0.064
starch and cellulose biosynthesis	5	XV	60.0	0.029
glycogen biosynthesis	5	XV	60.0	0.029

Conclusions

Combined with findings in other expression QTL studies, these results give new insights into the genetic control of expression and the role of pathways. While most pathways appear to be under genetic control, the control points are distributed throughout the pathway and tend to affect the expression of only one or two genes. Downstream expression effects within a pathway are limited. However, the impact of differential expression in one or few genes could affect the output of the pathway as a whole, specifically through the effects on protein levels, metabolites, or other molecules. Our results for yeast pathways will inform future studies investigating QTL linked to metabolite production. Integrating such studies could elucidate the underlying mechanisms of pathway control. Just as interesting as knowing which genes can be upregulated or downregulated within a pathway to affect metabolite production, is knowing which genes are robust to genetic perturbations.

Every published eQTL study has found many genes co-expressed and linked to a genetic hotspot region, and we found these linkages span genes from many different pathways. We conclude that modules of co-expressed genes are not the same thing as pathways, but that there is an important relationship that bears further examination. We conjecture that such modules are a higher order of transcriptional control with phenotypic consequences. In this study, perhaps these module QTLs are adaptations related to the wild yeast strain's ability to survive in its natural environment. This adaptation encompasses changes in the function of many pathways. Our findings suggest that this may be accomplished by differentially regulating very few genes from each pathway.

This model of transcription control has important implications for future studies. Currently, regulatory modules are generally examined for biological similarity via external data sources such as Gene Ontology or even pathway databases such as SGD. In our model, modules can be viewed primarily as a collection of intermediate regulators, which in turn govern the production of other biomolecules. Strong similarity within modules is not necessary. The analytical challenge in testing such a model is separating downstream expression effects from primary pathway regulation. A hotspot QTL may directly affect relatively few genes corresponding to a few key pathways, resulting in a cascade that results in a particular phenotype. However, this cascade is likely to have secondary effects on the expression of many genes, which also appear linked to the hotspot QTL.

The repercussion for composite trait mapping is significant. Such strategies are primarily used to gain power to detect module QTLs and to save computation by mapping fewer traits overall. They are undoubtedly effective in both respects. However, they must be used with caution because they assume shared regulators within groups. If applied to groups like yeast pathways the resulting QTLs could be incorrect. One must also consider what is lost in such an approach. We found the majority of QTLs are unique to specific traits and may not show up in a composite averaging approach.

References

Bao, L., L. Wei, J.L. Peirce, R. Homayouni, H. Li, M. Zhou, H. Chen, L. Lu, R.W. Williams, L.M. Pfeffer, D. Goldowitz, and Y. Cui. 2006. Combining gene expression QTL mapping and phenotypic spectrum analysis to uncover gene regulatory relationships. *Mamm Genome* **17:** 575-583.

Brem, R.B., J.D. Storey, J. Whittle, and L. Kruglyak. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436:** 701-703.

Brem, R.B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296:** 752-755.

Ghazalpour, A., S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E.E. Schadt, T.A. Drake, A.J. Lusis, and S. Horvath. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genet* **2:** e130.

Hong, E., R. Balakrishnan, K. Christie, M. Costanzo, S. Dwight, S. Engel, D. Fisk, J. Hirschman, M. Livstone, R. Nash, J. Park, R. Oughtred, M. Skrzypek, B. Starr, C. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, B. Hitz, S. Miyasato, M. Schroeder, A. Sethuraman, S. Weng, K. Dolinski, D. Botstein, and J. Cherry. 2005. Saccharamyces Genome Database.

Karp, P.D., S. Paley, and P. Romero. 2002. The Pathway Tools software. *Bioinformatics* **18 Suppl 1:** S225-232.

Kliebenstein, D.J., M.A. West, H. van Leeuwen, O. Loudet, R.W. Doerge, and D.A. St Clair. 2006. Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics* **7:** 308.

Wessel, J., M.A. Zapala, and N.J. Schork. 2007. Accommodating pathway information in expression quantitative trait locus analysis. *Genomics* **90:** 132-142.

West, M.A., K. Kim, D.J. Kliebenstein, H. van Leeuwen, R.W. Michelmore, R.W. Doerge, and D.A. St Clair. 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. *Genetics* **175:** 1441-1450.

Yvert, G., R.B. Brem, J. Whittle, J.M. Akey, E. Foss, E.N. Smith, R. Mackelprang, and L. Kruglyak. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* **35:** 57-64.

Zou, W. and Z.B. Zeng. 2008. Multiple Interval Mapping for Gene Expression QTL Analysis. *(in press)*

Chapter 4

SNAP: Combine and Map Modules for Multilocus Population Genetic Analysis

David L. Aylor, Eric W. Price, and Ignazio Carbone

Abstract

We have added two software tools to our application suite for working with DNA sequences sampled from populations. SNAP Map collapses DNA sequence data into unique haplotypes, extracts variable sites, and manipulates output into multiple formats for input into existing software packages for evolutionary analyses. Map includes novel features such as recoding indels, including or excluding variable sites that violate an infinite-sites model and the option of collapsing sequences with corresponding phenotypic information, important in testing for significant haplotype-phenotype associations. SNAP Combine merges multiple DNA sequence alignments into a single multiple alignment file. The resulting file can be the union or intersection of the input files. SNAP Combine currently reads from and writes to several sequence alignment file formats including both sequential and interleaved formats. Combine also keeps track of the start and end positions of each separate alignment file allowing the user to exclude variable sites or taxa, important in creating input files for multilocus analyses. SNAP Combine and Map are freely available at (http://snap.cifr.ncsu.edu/). These programs can be downloaded separately for Mac, Windows and Unix operating systems or bundled in SNAP Workbench. Each program includes online documentation and a sample dataset. A description of system requirements and installation instructions can be found at http://snap.cifr.ncsu.edu/.

Introduction

Recent advances in theoretical approaches for exploring population processes from DNA sequence variation within populations have resulted in a surge of new software tools (Beerli, 2006; Coop and Griffiths, 2004; De Iorio and Griffiths, 2004; De Iorio and Griffiths, 2004; De Iorio, et al., 2005; Hey and Nielsen, 2004; Lyngsø, et al., 2005; Song, et al., 2005). These tools are often designed and validated using simulated data with unique input file formats and rarely make provisions for converting data into those formats. As a result, biologists with real data of varying complexity must create input files manually. Large multilocus data sets make this increasingly complex, and necessitate the development of software tools to make multiple DNA sequence alignments accessible to new evolutionary methods. We have developed two such tools that we report here. SNAP Combine and Map will help researchers to visualize the distribution of DNA sequence variation within populations, extract and merge information from one or multiple sequence alignments, and enable further analysis by creating input files for several population genetic analysis programs.

Systems and Methods

Previously we developed a workbench program that can manage and coordinate a Suite of Nucleotide Analysis Programs (SNAP; Price and Carbone, 2005). SNAP Map will manipulate raw DNA sequence data from population samples into a variety of useful formats. The utility was conceived both to extract and characterize sequence variation in multiple sequence alignments and to serve as a bridge between existing applications requiring dissimilar input file formats. Variation includes both single nucleotide polymorphisms (SNPs) and insertions or deletions (indels). An indel is defined as one or more contiguous sites in a multiple sequence alignment that contain gaps in at least one sequence. Beyond identifying variable sites, SNAP Map provides the option to include or exclude indels, missing data, or infinite-sites violations. The infinite-sites model is based on the assumption that few polymorphic sites will have more than two nucleotides present (Hartl and Clark, 1997), and recent software such as Beagle (Lyngsø, et al., 2005) and Shrub (Song, et al., 2005) requires that sites violating this assumption be eliminated from input data. Additionally, we have included the ability to merge biogeographic or other phenotypic information with genetic sequence data to enable association-based analyses (Dean, et al., 2005). Our goal was to maximize flexibility of the program so that it may be used to catalog sequence variation or simply convert multiple sequence alignments into specific file formats (Table 4.1).

A key feature of SNAP Map is the ability to collapse individual sequences into unique haplotypes, and to keep track of the count of each haplotype in the population sample. This is a necessary step for analyses that assume an infinite sites model, and is a requirement for several of the software implementations we support (Table 4.1). A haplotype is a specific sequence of alleles or SNPs. Haplotypes are a useful way of grouping individuals according to genotype and are part of a powerful framework for testing significant associations with phenotype (Carbone, et al., 2004; Dean, et al., 2005; Phillips, et al., 2002).

A novel extension of the collapse functionality is the option to collapse indels to unique integers. Indels are often removed from multiple sequence alignments because of the difficulty in modeling the mutation process at these sites. Our software provides the user with the option to extract indels and recode each unique indel with a one-digit integer. The appropriate integer is reinserted into each individual sequence, yielding alignments in which gaps are recoded as a single polymorphic site. By recoding indels, we can take full advantage of variation at these sites in parsimony analyses and identify those sites that are compatible with an infinite-sites model. For example, the recoding of multilocus microsatellite and fingerprint data has important applications in phylogenetics and allows us to combine rapidly evolving markers with more slowly evolving base substitutions when reconstructing patterns of descent (Carbone, et al., 1999; Dettman and Taylor, 2004).

SNAP Combine is designed to facilitate multilocus analyses. Since most existing software has no provisions for multilocus sequences, we developed a tool that could seamlessly merge sequence data for each individual/locus within the population. The merging operation performs a union of the input loci by default but intersection is also supported. The intersection is important to accommodate loci with missing sequence data for some taxa, thereby allowing researchers to start on data analysis while continuing the work to fill-in missing data. SNAP Combine merges multiple, potentially heterogeneously formatted, input files into an output file of specified format. Combine supports the following interleaved sequence formats for input and output: PHYLIP, NEXUS, FASTA and CLUSTAL. PHYLIP, NEXUS, and FASTA are also supported as sequential sequence formats. SNAP Combine can be used to extract sequence subregions or taxon subsets and create new alignment files with specific combinations of loci. This is a powerful functionality and has been useful for examining patterns of sequence variation within and among loci (Charles, et al., 2005).

The lack of standard file formats for population analysis software is a ubiquitous problem and the process of manually converting between different file formats can be tedious and problematic. Both SNAP Map and Combine will simplify this process. The input file for SNAP Map is a sequential PHYLIP formatted sequence alignment, a standard output file option in sequence alignment programs, such as CLUSTAL W (Thompson, et al., 1994), Sequencher Version 4.5 (Gene Codes Corporation, Ann Arbor, MI) and phylogeny inference packages, such as PHYLIP (Felsenstein, 2004) and PAUP (Swofford, 1998). To facilitate the conversion, SNAP Combine has the added functionality of converting CLUSTAL W and NEXUS-formatted alignment files into the sequential PHYLIP-formatted files for SNAP Map and vice versa. The conversion of combined PHYLIP files to CLUSTAL format is especially important when excluding strains from multiple alignments; this may result in suboptimal alignments with unnecessary alignment gaps that can easily be removed by realigning with CLUSTAL W.

Implementation

Our current implementation of SNAP Map generates more than ten distinct output formats (Table 4.1). These particular formats were included from necessity as they are the required input file formats for the various analysis tools we use frequently in our laboratory and centre (Carbone, et al., 2004; Dean, et al., 2005). Each tool requires a specific input format that is not generated by other programs; this can be a source of frustration and error if generating these files manually. These are primarily applications developed for analyzing population structure and history within a nonparametric, coalescent or Bayesian framework but are also useful in multilocus macro-evolutionary analyses (Charles, et al., 2005; Geml, et al., 2006). Several of these tools can include geographical location or other phenotypic data in their analyses of population processes, and SNAP Map can merge specific individuals or haplotypes with corresponding phenotypic data. This ability allows us to take full advantage of both nonparametric and parameter-rich sequence-based models in testing for significant genotype-phenotype associations.

SNAP Map generates a summary table output that provides a visual overview of sequence variation in the population sample. This serves as a convenient reference and as a tool for exploring evolutionary processes at specific variable sites. The summary table numbers each site and gives the position of the variable site in the original sequence alignment. Each site is further labeled as a transition/transversion and informative/uninformative polymorphism. If the sequences have been collapsed to haplotypes, the summary table includes the frequency of each haplotype and provides a haplotype consensus sequence.

Framework

SNAP Combine is written in Java and Map in ANSI C. They were developed on Apple's OS X operating system, but can be compiled on any platform. Both are part of the SNAP suite of software tools developed in the Carbone laboratory at North Carolina State University (http://snap.cifr.ncsu.edu). SNAP WorkBench provides an event-driven graphical user interface for integrating SNAP Map, Combine and other command-line tools. Several program calls to SNAP Map, each including different options, can be included in the SNAP Workbench menus; we recommend using Map with the Workbench for maximum ease of use. These and other SNAP tools can be downloaded from our web site.

References

Beerli, P. (2006) Comparison of Bayesian and maximum-likelihood inference of population genetic parameters, *Bioinformatics*, **22**, 341-345.

Beerli, P. and Felsenstein, J. (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach., *Genetics*, **152**, 763-773.

Carbone, I., Anderson, J.B. and Kohn, L.M. (1999) Patterns of descent in clonal lineages and their multilocus fingerprints are resolved with combined gene genealogies, *Evolution*, **53**, 11-21.

Carbone, I., Liu, Y.C., Hillman, B.I. and Milgroom, M.G. (2004) Recombination and migration of Cryphonectria hypovirus 1 as inferred from gene genealogies and the coalescent, *Genetics*, **166**, 1611-1629.

Charles, L., Carbone, I., Davies, K.G., Bird, D., Burke, M., Kerry, B.R. and Opperman, C.H. (2005) Phylogenetic Analysis of *Pasteuria penetrans* by Use of Multiple Genetic Loci, *The Journal of Bacteriology*, **187**, 5700-5708.

Coop, G. and Griffiths, R.C. (2004) Ancestral inference on gene trees under selection, *Theor Popul Biol*, **66**, 219-232.

De Iorio, M. and Griffiths, R.C. (2004) Importance sampling on coalescent histories. I, *Adv. in Appl. Probab.*, **36**, 417–433.

De Iorio, M. and Griffiths, R.C. (2004) Importance sampling on coalescent histories. II: Subdivided population models, *Adv. in Appl. Probab.*, **36**, 434–454.

De Iorio, M., Griffiths, R.C., Leblois, R. and Rousset, F. (2005) Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models, *Theoretical Population Biology*, **68**, 41-53.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., Read, N.D., Lee, Y.H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E. and Birren, B.W. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*, *Nature*, **434**, 980-986.

Dettman, J.R. and Taylor, J.W. (2004) Mutation and evolution of microsatellite loci in Neurospora, *Genetics*, **168**, 1231-1248.

Felsenstein, J. (2004) PHYLIP (Phylogeny Inference Package). *Distributed by the author, Department of Genomic Sciences, University of Washington, Seattle*.

Geml, J., Laursen, G.A., O'Neill, K., Nusbaum, H.C. and Taylor, D.L. (2006) Beringian origins and cryptic speciation events in the fly agaric (*Amanita muscaria*), *Molecular Ecology*, **15**, 225-239.

Griffiths, R.C. and Marjoram, P. (1996) Ancestral inference from samples of DNA sequences with recombination, *J Comput Biol*, **3**, 479-502.

Griffiths, R.C. and Tavaré, S. (1994) Ancestral inference in population genetics, *Statistical science*, **9**, 307-319.

Hartl, D. and Clark, A. (1997) *Principles of Population Genetics*. Sinauer Associates, Inc., Sunderland, MA.

Hein, J. (1993) A Heuristic Method to Reconstruct the History of Sequences subject to Recombination, *Journal of Molecular Evolution*, **36**, 396-406.

Hey, J. and Nielsen, R. (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of Drosophila pseudoobscura and D. persimilis, *Genetics*, **167**, 747-760.

Hudson, R.R. (2000) A new statistic for detecting genetic differentiation, *Genetics*, **155**, 2011-2014.

Hudson, R.R., Boos, D.D. and Kaplan, N.L. (1992) A statistical test for detecting geographic subdivision, *Molecular Biology and Evolution*, **9**, 138-151.

Lyngsø, R., Song, Y.S. and Hein, J. (2005) Minimum recombination histories by branch and bound, *Proceedings of Workshop on Algorithms in Bioinformatics 2005, Lecture Notes in Computer Science*, **3692**, 239-250.

Maddison, D.R., Swofford, D.L. and Maddison, W.P. (1997) NEXUS: an extensible file format for systematic information, *Syst Biol*, **46**, 590-621.

Myers, S.R. and Griffiths, R.C. (2003) Bounds on the minimum number of recombination events in a sample history, *Genetics*, **163**, 375-394.

Nielsen, R. and Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach, *Genetics*, **158**, 885-896.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A*, **85**, 2444-2448.

Phillips, D.V., Carbone, I., Gold, S.E. and Kohn, L.M. (2002) Phylogeography and genotype–symptom associations in early and late season infections of canola by *Sclerotinia sclerotiorum*, *Phytopathology*, **92**, 785-793.

Price, E.W. and Carbone, I. (2005) SNAP: workbench management tool for evolutionary population genetic analysis, *Bioinformatics*, **21**, 402-404.

Song, Y.S., Wu, Y. and Gusfield, D. (2005) Efficient computation of close lower and upper bounds on the minimum number of recombinations in biological sequence evolution, *Bioinformatics*, **21 Suppl 1**, i413-i422.

Swofford, D.L. (1998) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0.* Sinauer Associates, Sunderland, Massachusetts.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, **22**, 4673-4680.

Table 4.1 File formats currently generated using SNAP Combine and Map
[1]IM (Hey and Nielsen, 2004) file format is also supported.
[2]Refers to the file formats for the programs Seqtomatrix, Permtest, Permchi, and Snn developed by R. Hudson.
[3]Can also combine multiple single locus MIGRATE files into one multilocus file.

File Format	Includes Phenotypic Information	Reference
NEXUS	No	Maddison, et al., 1997
CLUSTAL	No	Thompson, et al., 1994
FASTA	No	Pearson and Lipman, 1988
PHYLIP	No	Felsenstein, 2004
MDIV[1]	Yes	Nielsen and Wakeley, 2001
GENETREE	Optional	Griffiths and Tavaré, 1994
RECOM58	No	Griffiths and Marjoram, 1996
RECMIN	No	Myers and Griffiths, 2003
RECPARS	No	Hein, 1993
HUDSON[2]	Yes	Hudson, 2000 Hudson, et al., 1992
MIGRATE[3]	Yes	Beerli, 2006 Beerli and Felsenstein, 1999
SHRUB and HAPBOUND	No	Song, et al., 2005
BEAGLE	No	Lyngsø, et al., 2005