Discovery of regulatory mechanisms from gene expression variation by eQTL analysis

Yang Huang*, Jie Zheng*, Teresa M. Przytycka

General terms: expression quantitative trait loci (eQTL), transcription, regulation, systems biology
*These authors contributed equally

1. Introduction

One of the big challenges in the post-genome era is the better understating of the gene regulation process. Recently developed high-throughput genotyping and gene expression platforms have enabled a powerful new tool: expression quantitative trait loci (eQTL) analysis, where loci or markers on the genomes are associated with variations in gene expression. In such studies, gene expression is treated as a quantitative trait. DNA polymorphism at the gene, binding site or a regulatory proteins, such as transcription factors (TFs), transcription association proteins and signaling proteins, is likely to affect the expression level of the gene in an inheritable way (Monks et al. 2004; Brem and Kruglyak 2005; Petretto et al. 2006b). Hence, a significant statistical association between a locus and the expression level of a gene suggests that the locus might regulate the gene. Since the early work of Jansen and Nap (Jansen and Nap 2001), eQTL has become a widespread technique to identify such regulatory associations and has been applied to a number of species including yeast (Yvert et al. 2003; Brem and Kruglyak 2005; Brem et al. 2005), *Eucalyptus* (Kirst et al. 2004), *Arabidopsis* (DeCook et al. 2006; West et al. 2007), barley (Potokina et al. 2008), maize (Schadt et al. 2003), Dropophila (Jin et al. 2001), killifish (Oleksiak et al. 2002), mouse (Klose et al. 2002; Bystrykh et al. 2005; Chesler et al. 2005), rat (Petretto et al. 2006a) and human (Monks et al. 2004; Cheung et al. 2005; Stranger et al. 2005; Stranger et al. 2007). Many of them used the strategy of genome-wide association study (GWAS), considering loci covering the whole genome and expression profiles of all or nearly all genes identified in the organism.

eQTL mapping is a variant of the classical quantitative trait loci (QTL) mapping, which discovers associations between genotypes and organism-level phenotypes, such as heritable

diseases, height, weight etc. Compared with classical QTL analysis, eQTL analysis presents unique opportunities and challenges. While classical QTL analysis considers traits at organism-level, eQTL analysis allows us to observe the most immediate consequences of genetic polymorphism, namely, its effects on gene expression through transcription and post-transcription control. Therefore eQTL-based approaches are often used identify various regulation mechanisms. Moreover, combined with QTL analysis, eQTL mapping also helps to unravel genetic architecture of classical traits like complex diseases. Probably the most striking potential of eQTL studies is the number of traits being considered. Focusing on a limited number of phenotypic traits, QTL analysis often reveals a partial picture of the genetic architecture related to particular traits. Simultaneous monitoring of thousands of gene expression traits provides unique and unbiased data and opens the possibility of constructing a global view of the regulation machinery. However, eQTL analysis also faces tremendous challenges because of the huge number of genes and genomic loci. Large scale eQTL analysis must deal with large computational demand and, more seriously, loss of statistical power due to multiple-testing issue.

This review is focused on the framework to study regulatory mechanisms using eQTL analysis. We start with introduction of the statistical methods used to discover eQTL association for single locus and multiple loci. Next, we review methods to infer co-regulated gene modules. Subsequently, we survey the methods for identifying causal relationship including the methods where eQTL data is combined with other biological information to identify regulator genes for differentially expressed gene(s). Finally, we discuss a computational method for discovering regulatory programs. In the concluding section, we also suggest some further readings. .

2. Mapping of eQTL

eQTL mapping requires two types of data: genetic variation data and gene expression data for the same set of individuals. For human study, samples of families or independent individuals are used. For the study of model organisms, two parental strains are often crossed and progeny samples (inbred, if applicable) are obtained. In fact, the methods described in this review, typically focus on analyzing the second type of data referred to as crosses. The chromosomes of the segregating individuals are genotyped (i.e. the alleles at genomic markers such as

microsatellites, single nucleotide polymorphisms (SNPs), etc. are determined using biological assays). Additionally, gene expression of the samples is measured and genes significantly differentially expressed in parental strains are selected. The expression levels of these genes comprise the list of expression traits.

Given above data, putative eQTL are obtained by associating the genotype of markers with the expression traits. The gene associated with an eQTL is often referred to as the target gene suggestive of putative regulation of the gene by the eQTL. Each trait-locus pair is assigned a score of linkage significance (e.g. a log of the odds ratio (LOD) score, or a *p*-value). Since there are many pairs of traits and loci in a genome, the next step is to correct for multiple-testing.

As a special type of QTL, eQTL mapping has borrowed many ideas from standard QTL mapping, and therefore we will introduce some classic QTL mapping approaches before addressing special challenges posed by eQTL.

We start by describing a statistical model that expresses the relation between phenotype and genotype. For simplicity, let us first look at a single pair of quantitative trait and putative QTL, and later we will address more challenging problems of multiple loci and multiple traits.
For the *i*th individual, the phenotype y_i is related with genotype x_i in the linear equation

$$y_i = a + bx_i + \varepsilon ,$$

where x_i is a variable indicating alleles of the locus, b measures the phenotypic effect of substituting the allele at a putative QTL, and ε is a random normal variable representing environmental contribution to the phenotype or noise, with mean 0 and variance σ^2. For the purpose of this review, we assume that the genotype at any marker takes only two values, 0 and 1. In this model, a, b and σ^2 are unknown parameters.

Within the above model, the traditional QTL mapping approach uses linear regression to test if the phenotypic effect is significantly different from 0. Since each time a single marker is analyzed, this approach is called *single-marker mapping.* This approach has a number of shortcomings. For instance, if the QTL does not lie at the marker, its phenotypic effect will be underestimated, and consequently more progenies may be required to identify such association.

To remedy the problems of the single-marker approach, Lander and Botstein (Lander and Botstein 1989) proposed the approach of *interval mapping*, where instead of searching for a QTL at the position of a single marker, a putative QTL between two markers is located. We first formulate the linear regression in the single-marker approach as the *maximum likelihood estimates* (MLEs), and then we use it in interval mapping. The linear regression solutions of *a*, *b* and σ^2 are the values that maximize the likelihood of the full model (i.e. probability that the observed data would occur). The MLE of null model can be obtained under the restriction that *b* = 0. Let the MLEs of the full and null models be denoted L_A and L_N. The significance of linkage can be measured by the increase of likelihood assuming the presence of the linkage relative to the linkage assuming its absence. This is naturally summarized by the LOD score defined as $\log_{10}(L_A/L_N)$. If the LOD score is higher than a threshold then the putative QTL is claimed significant. The method of maximum likelihood and LOD score can be adapted to interval mapping as follows. In this case the QTL genotype is unknown but the genotypes of flanking markers are known. The probability that the QTL in the *i*th individual takes genotype *x*, denoted $G_i(x)$, can be estimated conditional on the observed genotypes of the markers and the recombination fraction between the QTL and the makers using a genetic map. Assuming the QTL has genotype *x,* the likelihood for the *i*th individual $L_i(x)$ can be estimated as in single-marker mapping. Given $L_i(x)$, the interval version of likelihood function is

$$L = \Pi_i \left[G_i(0) L_i(0) + G_i(1) L_i(1) \right].$$

This likelihood function can be plugged in the LOD score to assess the significance of linkage. Compared with single-marker approach, the interval mapping approach has a number of advantages: it is able to locate the QTL as an interval rather than a point as it takes into account of information from more markers, it is more powerful and requires fewer progenies; etc.

Compared with traditional QTL mapping, which usually involves only a few traits and markers, eQTL has much larger number of markers or traits (in the order of thousands). The large number of traits and loci poses challenges in both computational efficiency and statistical power. A prominent challenge is the *multiple-testing issue* (i.e. the chance of false positive in a family of multiple hypothesis tests is higher than that of a single test). A straightforward method to correct for multiple-testing is the well-known *Bonferroni* correction, which is to inflate the significance of an individual test by the total number of tests. In fact, it controls the family-wise error rate

(FWER), i.e. the probability of type I error in any test of a family under simultaneous consideration. For eQTL, in which we expect only a small fraction of pairs to be true positive, Bonferroni correction may often be too stringent.

Another approach is to threshold individual significance score of linkage with a critical value in the null distribution of all significance scores (Churchill and Doerge 1994). While the null distribution is hard to know, it is approximated by permutation tests. For each linkage between a trait and a locus, we shuffle the phenotype (i.e. randomly reassign the trait of each individual to a new individual but retain the individuals' genotypes), and we assess the significance of the association after shuffling. The results to a group of N such shuffled data provides an approximation to the null distribution for the hypothesis of no linkage. To control the overall type I error rate to no more than α, we derive the *experiment-wide* critical values as follows. For the results of each of the N shuffled data, we find the maximum test statistic over all loci; then we order these N selected values and their $100(1-\alpha)$ percentile is the critical value (i.e., a significant pair of trait and loci must have its test statistic more extreme than $N(1-\alpha)$ of these values from random shuffling).

Recently, the method of *False Discovery Rate* (FDR) has become frequently the method of choice for addressing the multiple-testing issue. By definition, FDR is the expected proportion of false positives in all the results claimed significant (Benjamini and Hochberg 1995). Since it focuses on testing results that are claimed significant and allows the rest to be false, FDR is more powerful than Bonferroni correction. A particular approach to control for FDR is q-value (Storey and Tibshirani 2003). Given a list of features each with a p-value to represent its significance, we calculate a q-value for each of the feature, which is equal to the FDR of the whole list when calling that feature significant. It has been shown that q-value is more powerful than the original FDR methodology.

The eQTL mapping methods described above have been successfully applied to real data analysis. To dissect the transcriptional regulation in budding yeast, Brem et al. (Brem et al. 2002) carried out eQTL mapping in a cross between a laboratory strain and a wild strain of Saccharomyces cerevisiae using single-marker mapping. Schadt et al. (Schadt et al. 2003)

detected microsatellite marker eQTL in maize, mouse and human, using standard interval mapping techniques and simple Bonferroni correction. Stranger et al. (Stranger et al. 2005) used HapMap data, where the expression traits are from cell-lines of HapMap human individuals and markers are dense SNPs. They used the three methods for multiple-testing correction (i.e. Bonferroni, permutation tests, and FDR) and observed significant overlap among them.

3. Multiple loci analysis

The eQTL mapping approaches described in the previous subsection identify association between a trait and a single locus. However, it has been shown that many gene expression traits have linkage to more than one locus (Brem et al. 2002; Yvert et al. 2003; Brem and Kruglyak 2005). Moreover, *epistatsis* (i.e. interaction among loci) is shown to be pervasive among expression traits. For instance, Brem et al. (Brem and Kruglyak 2005) approximated that 16% of heritable transcripts exhibit epistasis. Therefore, multiple locus models that explicitly consider epistasis are necessary to obtain valid estimates. For simplicity, let us assume here that we search for two loci for each trait; most approaches below can be extended to more than two loci.

A straightforward method for mapping two loci is exhaustive two-dimensional (2D) linkage scan where all pairs of loci are tested for linkage. However, this method is computationally costly and suffers from low statistical power due to the large number of tests. There are two other approaches for mapping multiple loci. One is *multiple interval mapping* which combines the previously described interval mapping with multiple regression (Zeng 1993; Jansen and Stam 1994; Zeng 1994; Kao et al. 1999). However, this method requires *a priori* models with a set of pre-chosen loci, which is difficult to formulate. Another approach is to use a *model selection* algorithm that aims to identify a subset of loci as parameters of the best model according to some optimality criterion (Zeng et al. 1999; Ball 2001; Broman and Speed 2002). This approach is computational demanding since it searches over a large number of potential models. In addition, none of the above methods provides a rigorous measure of the *joint* significance of multiple loci where *all* of the identified loci are truly linked.

To remedy the pitfalls of existing approaches for multiple loci, Storey et al. (Storey et al. 2005) proposed a sequential search approach to map two eQTLs for a gene expression trait. It includes

algorithms to assess joint significance of two loci and to measure evidence for epistasis. First we define a statistical model for multiple loci as follows.

(M0) Expression = baseline level + noise,

(M1) Expression = baseline level + locus1 + noise,

(M2) Expression = baseline level + locus1 + locus2 + locus1 × locus2 + noise,

where "locus1" denotes the effect of the primary locus, and "locs1 × locus2" represents the epistatic interaction between the primary and secondary loci. Clearly, this model can be extended to include the third, fourth loci and so on. Using this model, we perform the following sequential search algorithm. For each gene expression trait, we first select its primary locus that gives the greatest improvement in strength of linkage by comparing model M1 with model M0. The strength of linkage is measured by the Bayesian posterior probability **Pr**(locus 1 linked | Data). Then, conditional on the primary locus, we select the secondary locus that gives the biggest improvement in **Pr**(locus 2 linked | locus 1 linked, Data) when comparing model M2 with model M1. The joint significance (i.e. probability that both loci are linked with the trait) is

Pr(locus 1 and locus 2 are linked | Data)

= **Pr**(locus 1 linked | Data) × **Pr**(locus 2 linked | locus 1 linked, Data).

The posterior probability is estimated by a nonparametric empirical Bayesian method that uses permutation tests to simulate null distributions (Churchill and Doerge 1994; Doerge and Churchill 1996). Then we rank the traits by the value of joint significance, and select those traits that have significance above a threshold. To get a reasonable threshold that controls the overall type I error and deals with multiple-testing issue, we deduce FDR from the posterior probabilities as follows. FDR is defined as the ratio of expected number of false positive traits divided by the number of traits called significant. Since the probability that a trait is false positive is 1 − **Pr**(locus 1 and locus 2 are linked | Data), then

FDR = ∑1 − **Pr**(locus 1 and locus 2 are linked | Data) / number of significant 2-locus linkage,

where the summation is over all traits called significant. The epistasis can be tested similarly by comparing the full model M2 to a purely additive model (i.e. with locus1 × locus2 equal to 0 in

model M2). Being applied to real gene expression traits in yeast (Brem et al. 2005; Storey et al. 2005), the sequential search approach is shown to be faster more powerful than exhaustive 2D linkage scan.

There are several limitations of the sequential search approach (Storey et al. 2005). It may miss those locus pairs with primarily epistatic effects in which neither single locus has significant effect. It may not be applicable to the case when the two loci are closely linked. Also, since this method is still limited to two loci; it is a challenging open problem to extend it to more loci without losing much computational efficiency and statistical power.

4. Basic properties of eQTL

Given a target gene and associated eQTL, it is convenient to distinguish two types of relationships: *cis*- and *trans*-acting eQTLs. A *cis*-acting eQTL locates in or close to the transcription region of the target gene. Such target gene is said being *cis*-regulated. The eQTL may exert effect on the target gene's expression in various ways. For example, DNA variation in the promoter sequence may affect TF binding. It is also possible that DNA variation in the coding sequence may affect mRNA sequence composition, splicing or secondary structure. A *trans*-acting eQTL can reside on the same chromosome with the target gene but distal to it or on a different chromosome. Such target gene is said being *trans*-regulated. For example, DNA variation in TF, transcription-associated proteins, such as activators and repressors, post-transcription regulation genes, can have various impacts on the expression level of the target gene. The differentiation between *cis*-and *trans*-acting eQTL is usually based on a distance threshold between an eQTL and its target gene. Note that, if only *cis*-acting eQTLs are to be identified, the multiple-testing issue discussed in section 2 will not be so severe since only loci close to a gene, whose expression is used as quantitative trait, are tested. As it will be discussed in the following sections, specific computational methods or models are used to discover causal relationship in *cis*-acting and *trans*-acting eQTL (Doss et al. 2005; Kulp and Jagalur 2006).

A *cis*-acting or *trans*-acting eQTL terminology is used to describe a single eQTL-target gene pair. A genomic region, which contains many eQTLs or to which many expression traits are mapped is referred to as an *eQTL hotspot.* eQTL hotspot is a conceptual term since there is no

precise requirement about how large a hot spot should be or at least how many genes need to be mapped to a hot spot. An eQTL hot spot can regulate tens to hundreds of genes (Morley et al. 2004; West et al. 2007; Zhu et al. 2008). Since an eQTL hotspot may point to a master regulator gene in the region, identification of possible eQTL hotspots is an important element of eQTL analysis. Due to the interest in eQTL hotspots, several methods were proposed for their identification. For example, Ghazalpour et al. (Ghazalpour et al. 2006) defined a module quantitative trait locus (mQTL) as the locus harboring significant number of eQTLs regulating genes within a given gene module. These mQTLs can be viewed as variants of eQTL hotspots as described above. Recently, Wang et al (Wang et al, 2007) observed, using a simulation experiment, that real eQTL hotspots may be sometimes difficult to discern from false positives resulting from possible artifacts caused by highly correlated gene expression or linkage disequilibrium..

In addition to the properties unique to eQTL, a genetic phenomenon observed in classical QTL, transgressive segregation is also observed in eQTL studies. Transgress segregation occurs, were a quantitative trait takes value more extreme relative to the values in either parent strain. Such transgressive segregation was also observed in some eQTL studies (Brem and Kruglyak 2005; Rowe et al. 2008). Brem et al. (Brem and Kruglyak 2005) used loci with alleles of opposite effect on expression traits to explain transgressive segregation. Genetic possibilities that could cause transgresive segregation were also described by Rieseberg et al. (Rieseberg et al. 1999)..

5. Inferring co-regulated modules

The expression level of genes in the same complex or pathway is often co-regulated. Conversely, a set of co-regulated genes is expected to be enriched for genes that share biological functions and/or canonical pathways. A set of co-regulated, presumably related, genes is therefore usually referred as a co-regulated module. The discovery of eQTL hotspots suggests that eQTL analysis can be helpful for uncovering such co-regulated modules and their regulators. In particular, it has been shown that the so called *trans* eQTL band, that is the set of genes linked to a common *trans*-acting eQTL has significant enrichment for genes with common annotations from GO

(Ashburner et al. 2000), KEGG (Kanehisa and Goto 2000) and ING (Ingenuity Systems, Redwood City, CA)(Wu et al. 2008).

The first step in inferring co-regulated modules from gene expression data is, typically, identification of clusters of co-expressed genes[1]. This can be done by a number of ways. For example, in their pioneering study on the yeast crosses, Yvert *et al.* (Yvert et al. 2003) used a simple hierarchical clustering based on similarity of expression patterns. In a more recent paper, Zhu et al, (Zhu et al. 2008) started with the so-called co-expression graph and used topological overlap distance (Ravasz et al. 2002) to identify co-regulated modules. There are a number of other clustering methods that could be applied in this context. For example Li et al. identified transcription modules in mouse by performing biclustering on two gene expression matrices (Li et al. 2006) where expression values from different parental strains were used as reference points. Namely, the (*i, j*) element each matrix was defined as the log ratio between the expression value of *i*th gene in *j*th progeny and the one of *i*th gene the corresponding parent strain.

The modules identified in these studies were, often shown to be enriched for genes sharing functional annotation such as GO category and/or linkage to a common chromosomal region. For example, in the study of Yvert *et al.* (Yvert et al. 2003) all but one of the modules were enriched for genes linked to a common region on a chromosome, the eQTL hotspot. Such linkage is consistent with the assumption that so constructed modules are indeed co-regulated.

The second step of this approach is to discover eQTL(s) that regulate the co-regulated modules identified in the first step. This is usually done by using the mean expression value of the genes in a given cluster as a new quantitative phenotype to which eQTL analysis can be applied. Indeed, Yvert et al. showed that the majority of clusters retrieved by their study showed linkage to at least one locus.

Finally, the genomic region near the locus is hypothesized to contain regulatory elements that regulate target-genes in the cluster. Note that identification of co-regulated modules and eQTLs

[1] Unless stated otherwise, the gene expression data set is referred to the data set obtained from segregating population used in eQTL study.

associated with such modules is not equivalent to identification of the causal regulatory gene(s). Methods to identify such causal regulators will be discussed in the next subsection.

A different approach to find co-regulated modules has been proposed by Samson and Self (Sampson and Self 2008). In their method, they associate with each gene a LOD score profile – a vector of LOD association scores of the gene with all loci. Subsequently, they use the Pearson correlation between LOD score profiles of two expression traits, ρ_L, to test if these two genes share common regulatory locus. Finally, they apply a hierarchical clustering method to identify clusters of genes associated with common loci. They showed that uncovered modules were enriched with genes with common GO functional annotation.

6. From eQTL to Causal Regulatory Relations of Genes

Mapping of eQTLs is only the first step toward discovering regulatory mechanism. Our next goal is to identify regulators, that are regulatory genes near the eQTL, that are responsible for the expression traits of target genes mapped to the eQTL. This is a challenging problem due to several statistical issues. First, a putative eQTL region is usually big (a few centimorgans wide) and typically contains many genes. Thus we need to reduce the width of eQTL and the number of candidate regulators, a process called *fine mapping*. Second, neighboring markers tend to have high correlations due to linkage disequilibrium; as a result, a target gene may be linked to false positive eQTL close to the true eQTL. Third, in case that many target-genes map to an eQTL *hotspot*, it is challenging to accurately identify the complete set of co-regulated target genes for the eQTL hotspot. To deal with the statistical issues and solve the problem, we mainly rely on three types of information: (1) physical locations of eQTL and genes, (2) expression traits of regulators and target genes, (3) gene ontology (GO) information about functions of the genes.

The causal relations between genes are inferred using statistical models, and thus represent indirect connections. To fill in the gap, we further attempt to explain the causal relation with molecular interaction pathways inferred from various data, such as protein-protein interactions (PPI), TF-DNA binding sites (TFBS), protein phosphorylation, etc. These inferred pathways can

provide insight to the molecular mechanism of gene regulation. The identification of causal relations and the inference of molecular pathways are related to each other: a set of causal relations is a starting point for the inference of pathways; conversely, pathways from regulators to target genes can serve as evidence supporting the causality.

However, one has to keep in mind that correlation among gene expression values may not necessarily reflect all functional relationships among genes. A regulator can affect the expression of target genes without a change of its own expression

6.1 Fine-mapping and expression-correlation-based methods

As previously mentioned, a typical eQTL is large (up to several centimorgans) and often contains many genes, among which only a few have causal effects on the expression of target genes. To narrow down the candidate causal genes, the following two steps are frequently used (1) reduce lengths of the eQTL regions by fine-mapping techniques; (2) perform expression correlation analysis between candidate regulators in an eQTL region and the target genes affected by the eQTL. Step (2) is based on the assumption that genes in the same pathway have strong correlations between their expression values. The assumption has been used extensively in microarray analysis (Stuart et al. 2003). Such two-step analysis is used, for example, in Bing and Hoeschele (Bing and Hoeschele 2005), thus in the following we first describe their method and then mention other related work.

One way to reduce the length of eQTL regions, is to employ a bootstrap resampling method proposed in (Visscher et al. 1996). Bootstrap samples are created by sampling with replacement the given set of expression values with the set of marker genotypes. For each of 1000 bootstrap data sets, Bing and Hoeschele employ a conventional approach to identify an eQTL position with the highest test statistic. Then from the sorted physical positions the 1000 eQTL, they determine the 95% confidence interval of eQTL position by taking the largest and smallest values of the bottom and top 2.5% respectively of the eQTL positions. Genes physically located inside the confidence interval are selected as candidate regulators. However, eQTL confidence intervals may still be large, especially when eQTL effect is small. A likely reason for such large eQTL confidence interval is the presence of multiple eQTL. To resolve a large eQTL interval into

smaller subregions each with one eQTL, one can perform sliding three-marker regression based on theoretical properties of multiple interval mapping (Zeng 1993). For each subregion consisting of a pair of consecutive markers, they obtain a t-statistic associated with partial regression coefficients of the two markers (see (Thaller and Hoeschele 2000) for more details). If a candidate gene is found to be located in a subregion with significant t-statistic, it is identified as a strong causal candidate of regulator.

For every eQTL confidence interval containing a list of candidate regulatory genes, the list of candidate regulators can be further narrowed down using an expression correlation based approach. This correlation analysis is based on the assumption that genes belonging to the same pathway or network are more likely to have strong correlation between their expression values. First, among all genes in the confidence interval, the gene with the smallest significant p-value is identified as the primary causal gene (denoted G1). Then, for every candidate regulatory gene different than G1, the first-order partial correlation coefficients of expression profiles of that gene and the target gene, conditional on the primary regulator is computed.. If there is at least one significant correlation coefficient, the most significant gene is taken as the secondary regulator (G2). The process is continued by computing second-order partial correlation and conditional on the primary and secondary regulator, etc, and then computing higher-order partial correlations until there is no more significant partial correlation coefficient. In this way we obtain a list of candidate causal genes (G1, G2, …). Note that this method is different from Storey's method for identifying multiple loci (Storey et al. 2005) in that it selects regulators in a set of candidate genes using expression correlation.

The existence of multiple target genes, as for example in the case of an eQTL hotspot, can be the source of additional information. It is reasonable to assume that among all possible regulators, the best candidates are those that correlate with a large number of functionally related target genes. This assumption is, for example, explored by Keurentjes et al. (Keurentjes et al. 2007). They start by calculating all pair-wise rank-based Spearman correlations of expression between each of the functionally related target gene and candidate regulators in eQTL intervals. These correlation coefficients are then ranked so that more strongly correlated pairs of regulator-target genes are at the top of the list. Then they apply Iterative Group Analysis (iGA) to obtain the probability of change (PC-value) for each regulator candidate (Breitling et al. 2004). Here, the

PC-value of a pair (regulatory gene, target gene) is the p-value for testing "how likely it is to observe all target genes correlated with the particular candidate regulator at least as high on the list as the given pair, by chance". A candidate gene with a significant PC-value (adjusted for multiple hypothesis testing) is a putative regulator, and all genes contributing to the PC-value are putative target genes. In addition, we can extend the set of target genes by using expression correlation between the genes already identified and potential target genes outside the initial group of functionally related genes.

6.2 Likelihood-based model selection

Methods described in the previous subsection typically consist of two steps: (1) map eQTL confidence intervals, (2) from genes physically located inside the eQTL intervals, identify candidate regulators. Kulp and Jagalur (Kulp and Jagalur 2006) proposed an alternative approach that unified the two steps into one step. Instead of finding loci first, they directly look at the genotype in a candidate regulatory gene *and* the expression traits of the regulator, simultaneously. The genotype in the candidate regulator corresponds to putative eQTL in previous methods. Substituting loci with genes, the method is called quantitative trait gene (QTG) mapping. In QTG mapping, one can express the causal relation between a candidate regulatory gene (with genotype Q_j and expression T_j) and a target gene (with expression T_i) using a Gaussian model:

$$P(T_i \mid Q_j, T, \theta) = N(\beta_0 + \beta_1 T_j + \beta_2 Q_j + \beta_3 T_j Q_j, \sigma)$$

where θ represents the parameters β and σ. Note that the term $\beta_3 T_j Q_j$ is used to model the interaction between genotype and expression value of the regulator. This conditional probability is similar to the likelihood function of the standard eQTL interval mapping. Indeed, the maximum likelihood estimation of parameters Q_j and θ can be obtained by expectation maximization (EM) algorithm. Analogous to LOD score, the strength of the causal relationship can be assessed by comparing full model with null model, as

$$\log_{10} \frac{P(T_i \mid Q_j, T_j, \theta)}{P(T \mid Q_j, T_j, \theta : \beta_1 = \beta_2 = \beta_3 = 0)}.$$

Compared to the standard two-step approach of eQTL mapping and regulator identification, the QTG method has the advantage that all evidence is integrated in a single model. Moreover, it provides us with the flexibility of choosing different null models by setting different subsets of $\{\beta_1, \beta_2, \beta_3\}$ to zero. For example, setting $\beta_1 = \beta_3 = 0$ in the null model measures the contribution of the expression values of candidate regulators to the expression traits of target genes.

After identification of regulatory genes and their causal relation with target genes, a natural question is: how does a regulator affect the expression levels of target genes? A simple situation is that the regulator is a transcription factor (TF). However, it has been shown that TFs are not enriched in eQTL hotspots (Yvert et al. 2003). Nonetheless, even if the regulator is not a TF, the propagation of genetic perturbation from the regulator to target genes is likely to be mediated by TF activities. This scenario is considered by Sun et al. (Sun et al. 2007) They estimate TF activities using approach of Yu and Li (Yu and Li 2005), combining data from literature and from genome-wide TF-binding study (Harbison et al. 2004). Given an eQTL module (an eQTL hotspot together with target genes), let *GC* denote the expression level of one *cis*-linked gene, *TA* denote a TF's activity, and *GT* denote the expression level of any gene other than *GC*. Then they consider three models:

- Causal model: $GC \rightarrow TA \rightarrow GT$
- Reactive model: $GC \rightarrow GT \rightarrow TA$
- Conditional independent model: $TA \leftarrow GC \rightarrow GT$

where an arrow indicate the direction of the pairwise relation. Such idea of the three models was considered earlier by Schadt et al. to study relationships between two clinical traits (Schadt et al. 2005; Sieberts and Schadt 2007). Each pairwise relation can be modeled by simple linear regression. From the regression parameters, conditional densities could be derived, which are further used to compute the likelihood of each model. Finally, likelihood ratio tests can be used to identify one model that is significantly better than the other two models.

Model selection methods have also been used to infer the causal relationship between expression traits and clinical traits (Schadt et al. 2005; Aten et al. 2008). Although these methods were used for a purpose different from our goal, they may be adapted to the context of finding causal relation between regulatory genes and downstream expression traits of target genes.

6.3 Pathway-based methods

The methods discussed in the previous section can only provide indirect casual relations but cannot explain such relationships on a molecular level. Realizing this problem, Tu et al. (Tu et al. 2006) integrated eQTL mapping results with a biological network built from protein-protein interaction, protein phosphorylation and TF-DNA interaction data. In this network, they search for pathways from candidate regulatory genes in an eQTL region to target genes. The inferred pathways are subsequently used to both identify most likely causal genes in the eQTL. This approach is based on two assumptions: (1) causal genes regulate target gene by affecting the activities of TFs for the target gene; (2) activities of genes on the likely pathway correlate with target gene's expression. Based on these assumptions, the computational problem is to find the pathway from a causal gene to TFs of the target gene so that expression levels of genes on the pathway correlate with the target gene. Tu et al. designed a stochastic backward search algorithm based on random walk starting from the target gene g_t. In the first step the algorithms picks one of the TFs regulating gene g_t from which it starts the random walk. The probability of moving from gene u, to previously unvisited gene v is proportional to the Pearson correlation coefficient of expression levels of v and the target gene. Each node can be visited only once, making sure the path is non-cyclic. The random walk is stopped when it reached a candidate regulator or when there were no more reachable unvisited nodes (hit a "dead end") or when it had reached the maximum allowed number of steps. The probability that a candidate regulator g_j is a causal gene can be approximated by $V_{t_k}(g_t)/N$, where $V_{t_k}(g_t)$ is the number of times g_j was visited and N is the number of random walks. In other words, the causal effect of g_j on g_t is modeled as the probability that g_j could be reached from g_t via random walks in the network. If there is only one TF for g_t, the candidate regulator with the largest $V_{t_k}(g_t)/N$ is taken as the best candidate. If g_t

has more than one TF, a linear combination of $V_{t_k}(g_t)/N$ for different t_k is used to determine the best candidate. Then a backwards search from the regulator gene toward t_k for nodes with largest count identified the most probable pathway from the regulator gene to the target gene.

The random walk method, while very intuitive, is computationally intensive because large number of random walks have to be generated. Furthermore many walks might end up in a dead end without reaching any candidate regulator gene. Thus in practice, this approach is unlikely to effectively explore the search space. Suthram et al. (Suthram et al. 2008) proposed a more efficient approach to identify regulator genes and the pathways from regulator genes to target genes. The approach is based on the well-established analogy between random walks and electric networks. It can be proved that when a unit current flows into an electric network, the amount of current through any intermediary node or edge is proportional to the expected number of times a random walker will pass through that node or edge (Doyle and Snell 1984). The algorithm of Suthram et al., called eQTL electrical diagrams (eQED), replaces the random walk model with current flows in electric circuits. Instead of assigning weights to nodes as in Tu et al., eQED assigns a weight to each edge (u, v), which is equal to the average of correlation coefficients of u and v with the target gene. The weights on the edges are modeled as conductances in the electric circuit. The *p*-value of the eQTL is used as the amount of current flowing from the locus to the target gene. Then, the problem is formulated with linear programming, where the optimal path has the maximum total sum of currents flowing. After computing current on each edge using a linear programming approach, eQED predicted the best candidate regulator gene to be the one with the highest current flowing through it.

A pathway based approach has been also used by Zhu et al. to identify casual regulator genes of a set of genes associated with an eQTL hotspot (Zhu et al. 2008). They used a gene network constructed using Bayesian network method (Zhu et al. 2004) by combining eQTL, PPI and TFBS data (see also Section 7 of this chapter). First, genes that could be *cis*-regulated by within eQTL hotspot region were identified. Next, for each gene in this set, the set of genes that could be reached by a path in the network starting from that gene were found. This set was, in turn, intersected with the set of the genes linked to the corresponding hotshot and the significance of the overlap was estimated using Fisher test corrected for multiple testing. If the overlap was significant, the corresponding gene was considered to be a regulator of the module associated

with the given hotspot. Using their method the authors recovered previously indentified (Yvert et al. 2003) as well as some novel hotspot regulators demonstrating added power of the integrative network reconstruction.

7. Using eQTL for inferring regulatory networks

It may seem to be straightforward to apply the approaches discussed in the previous section to construct regulatory networks. As a first approximation, one can put a directed edge between a regulator gene and one of its target genes (Bing and Hoeschele 2005; Li et al. 2005; Keurentjes et al. 2007). However approaches that infer network structure locally by adding one edge at a time may miss sophisticated regulation pattern involving large number of genes. Therefore, method considering the data set as a whole should be applied to construct regulatory networks.

One of such methods is the Bayesian network approach, which is a probabilistic graphical model, represented by a directed acyclic graph. In the graph, nodes represent variables and edges represent conditional probabilistic dependence between variables. There are many important application of Bayesian network in genomics study. For example, Bayesian network can be inferred from gene expression data, where each node represents the expression level of a gene and an edge indicates two genes' expression is conditionally dependent (Friedman 2004; Jensen 2007). It is natural to interpret such a Bayesian network as a regulatory network: an edge $g_j \rightarrow g_i$ (g_j is a parent of g_i) indicates the gene g_j regulates g_i. The Bayesian networks have been successfully used to infer complex causal relationship pattern among hundreds of genes and is ideal for integrating multiple data sources.

A big limitation of Bayesian network approach is its demand of large computational power. Hence it is a common practice to reduce the search space of possible networks through various biological or computational heuristics (Zhu et al. 2004; Li et al. 2005). For example, in their early work, Zhu et al. (Zhu et al. 2004) reconstruct a Bayesian network using gene expression data. In the reconstruction method, they reduce the number of possible edges in the network by assuming that one gene could have at most three regulator genes and by considering only a

subset of all genes as candidate regulator genes. To select this subset they use a mutual information threshold and a measure of a correlation between LOD scores similar to the one described in Section 5. To further reduce the computational burden, they also infer directly some causal relationship from the number of overlapping eQTL of two genes by defining

$$prob(g_j \text{ regulates } g_i) = r(g_i, g_j)\, N(g_i) \Big/ \big(N(g_i) + N(g_j)\big),$$

where $N(g_i)$ is the number of significant eQTL linked to g_i. If g_i and g_j have common eQTL, but the g_j has more eQTL than g_i, a prior would be set to indicate g_j is a candidate regulator gene of g_i. Subsequently, the authors construct a consensus network from 1000 Bayesian networks. Finally, data processing inequality was applied to check if an edge $g_i \to g_k$ was over fit when there were edges $g_j \to g_i$ and $g_j \to g_k$. If mutual information (MI) between g_i and g_k is less than MI between g_i and g_j or g_j and g_k then the edge $g_i \to g_k$ would be removed. Later, Lum et al. (Lum et al. 2006) applied Zhu et al's Bayesian network method to construct a regulatory network of murine brain.

More recently, Zhu et al. combined their Bayesian network with PPI data and Transcription Factor Binding Site (TFBS) data to construct yeast regulatory network (Zhu et al. 2008). First, using the Bayesian network method (Zhu et al. 2004), they construct a Bayesian network from the expression data. Then eQTL data was added to extend the network. Namely, genes with *cis* eQTL are assumed to be parents of genes with coincident *trans* eQTL. Genes derived from the same eQTL are subsequently used to infer casual/reactive or independent reaction as described above (Schadt et al. 2005). If casual/reactive relationship could not be determined in this way, the authors break the ties using a complexity criterion that intuitively was considering a gene with simpler and stronger eQTL signature as the causal gene (Zhu et al. 2008). Finally, PPI data is added by considering protein complexes and their regulators. Namely, if at least half of genes in a protein complex carries a given TFBS, then all gens in the complex were included as being under the control of the corresponding transcription factor. For additional details on including TFBS in the network, we refer readers to the original paper (Zhu et al. 2008).Their construction showed that integrating eQTL data with other data sources can improve the reconstruction of regulatory networks

A different approach has been developed by Liu et al. (Liu et al. 2008), who constructed a regulatory gene network using the SEM method (Stein et al. 2003; Kline 2004). SEM is an extension of general linear model and is used to test and estimate causal relationships. In general, it consists of a measurement model and a structural model. The measurement model describes the relationships between latent variables, which are not directly measured, and their indicators, various measurements obtained from experiments. The structural model describes causal relationships among latent variables. In the SEM model used by Liu et al., all variables are measured, hence no measurement model is needed and the structural model describes casual relation between measured variables. Algebraically, the structural model is represented as $y_i = By_i + Fx_j + e_i; i \in [1..N]$ where N is the number of progenies, y_i is the expression vector, x_j is the vector of eQTL genotypes, and e_i is a vector of error terms. Matrix B models the causal relationships between gene expression traits and matrix F models the causal relationships of eQTL on gene expression traits. Here, y_i and x_j are observed variables and B, F, and e_i are model parameters. The network representing the model contains a node for every x_i and every y_i. There is an edge between nodes corresponding to kth and mth gene expression trait (that is between y_k and y_m.) if $B[k, m]$ is non-zero. Similarly there is an edge between nodes corresponding kth eQTL and mth gene expression trait (that is between x_k and y_m) if $F[k, m]$ is non-zero. To estimate the model, Liu et al. applied an approach combing likelihood maximization and network topology search. The quality of the model was evaluated using a criterion which considered both: a maximized likelihood function used to optimize the parameters of a given network and a penalty term for the number of free parameters to optimize over various network topologies. To reduce the search space for the network topology, the network constrained to edges of a predefined network, which authors referred to as an encompassing directed network (EDN). To construct the EDN several criteria were applied to include the edges are most likely to be a part of the final model.. More details about their eQTL analysis and regulator identification are beyond the scope of this review. We refer readers to the original paper. One of the advantages of the SEM method is that it allows cyclic structures, which are not allowed in Bayesian networks. However, similar to Bayesian network method, it cannot process the network of large size and is computationally demanding.

In an interesting application related to the identification of casual relationship using eQTL data, Schadt's group (Chen et al. 2008) has recently successfully identified a gene sub-network having crucial causal relationship with complex metabolic syndromes. They first obtained several co-expression sub-networks by clustering the gene expression data set of a mouse eQTL experiment. Such sub-network was then considered as the gene network controlling a metabolic trait if it had significant enrichment of expression traits having causal relationship with the metabolic trait.

8. Inferring regulatory programs

In the previous sections, we discussed methods to identify co-regulated modules as well as approaches to identify regulators of such modules. Importantly, co-regulated modules are often controlled by more than one regulator. In such situation the expression of genes from the module is a function of the combination of states (expression level, phosphorylation, etc.) of the corresponding regulators. The combinations of states that leads to the observed expression patterns is referred to as regulatory program. Uncovering such regulatory programs is one the major challenges on the way to the complete understanding of regulatory networks. To address this problem, Lee et al. proposed a method that is directed toward detecting co-regulated modules together with their regulatory programs (Lee et al. 2006). The method, implemented as software called Geronemo, builds on their previous work on extracting network modules (Segal et al. 2003b; Segal et al. 2003a).

Regulatory programs targeted by Geronemo consist of TF, signaling molecules, chromatin factors and SNPs. Each such program is assumed to have hierarchical, tree-like structure, where each internal node is associated with a "regulator" and splits the gene expression of the module genes into two distinct behaviors. The two submodules correspond to two subsets of strains. Geronemo allows for two types of regulators: g-regulators and e-regulators. g-regulators correspond to a split along a SNP and have two clearly defined split values corresponding to the two progenitor alleles. I contrast, e-regulators define a split on the continuous set of possibilities based on their expression level. Note, however, that e-regulators still allow for splitting the gene

expression into two distinct behaviors (e.g. up-regulated and not up-regulated). In Gerenemo, the set of possible e-regulators is restricted to known TF, signaling molecules and chromatin factors. One can think of the regulatory program as a decision tree with regulators in the decision nodes. The goal of such decision tree is to sort the progenies into groups so that, for each group, all genes in the modules are coherently expressed. As a special case, a one-gene module with one g-regulator corresponds, roughly, to a traditional eQTL and a multi-gene module with one g-regulator may correspond to an eQTL hotspot. In contrast, a multi-gene module with one e-regulator contains genes whose expression is correlated or anti-correlated with the expression of this regulator.

The co-regulated modules and their regulators are being discovered through an iterative learning approach. The learning procedure is initialized with a certain number of modules obtained by *k* means clustering and then iterated over two phases: (*i*) assigning each gene into some regulatory module; and (*ii*) learning the regulation program for each module. The program is learned in a recursive way by choosing, at each point, the regulator that best splits the gene expression of the module genes into two distinct behaviors. When considering a potential split, Geromeno evaluates all candidate regulators and splits values and selects the one that leads to the highest improvement in score. The procedure is iterated until convergence.

When applied to the yeast crosses, Geronemo identified a large number of modules that have both chromosomal characteristics and are regulated by chromatin modification proteins.

Compared with eQTL methods described earlier, the approach applied by Gerenomo is unique - it not only finds the co-regulated modules but also identifies regulatory programs. While other approaches were focusing on finding putative regulators, Gerenomo, also predicts, the states of these regulators (e.g. up / down regulation of aTF, SNP variant, etc). Namely, for every expression value in a gene in such module, one can trace back the decision tree defining the regulatory program and, for an internal node on such path, read the state (e.g. up/down) of the regulator associated with this node. However, unlike the methods described in the previous section, the regulators are selected from a pre-defined set of genes.

9. Conclusions and further reading

The discoveries of the last decade brought an apperception of the complexity of regulatory networks. High-throughout data opens, for the first time, the possibility of their preliminary reconstruction. The eQTL analysis provides an important step in this direction. First, it allows addressing the network reconstruction in a non-biased genome-wide fashion. More importantly, genetic data can be used in a natural way to infer causal relation. However, even with eQTL analysis discovering causal relationships remains non-trivial. A formidable challenge is related to multiple hypothesis testing. Another difficulty is posed by challenges related to uncovering combinatorial regulation and epistasis.

In this review, we focused our attention on applications of eQTL analysis to uncovering regulatory mechanisms. Consequently many other aspects of eQTL studies, as well methods that do not apply directly eQTL data, have not been covered. For a retrospective historical account of Genome Wide Association Study (GWAS), we refer the readers to the article by Kruglyak (Kruglyak 2008). Rockman and Kruglyak (Rockman and Kruglyak 2006) provide a review focusing on genetics of gene expression. Methods of combining eQTL with QTL to study complex disease traits are discussed in (Schadt 2005; Sieberts and Schadt 2007). Kendziorski and Wang's review discusses various statistical methods for eQTL in more details (Kendziorski et al. 2006).

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1): 25-29.

Aten JE, Fuller TF, Lusis AJ, Horvath S (2008) Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Syst Biol 2: 34.

Ball RD (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. Genetics 159(3): 1351-1364.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 57(1): 289-300.

Bing N, Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170(2): 533-542.

Breitling R, Amtmann A, Herzyk P (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. BMC bioinformatics 5: 34.

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

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296(5568): 752-755.
Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. Nature 436(7051): 701-703.
Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J R Stat Soc [Ser B] 64: 641-656.
Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT et al. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 37(3): 225-232.
Chen Y, Zhu J, Lum PY, Yang X, Pinto S et al. (2008) Variations in DNA elucidate molecular networks that cause disease. Nature 452(7186): 429-435.
Chesler EJ, Lu L, Shou S, Qu Y, Gu J et al. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat Genet 37(3): 233-242.
Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M et al. (2005) Mapping determinants of human gene expression by regional and genome-wide association. Nature 437(7063): 1365-1369.
Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138(3): 963-971.
DeCook R, Lall S, Nettleton D, Howell SH (2006) Genetic regulation of gene expression during shoot development in Arabidopsis. Genetics 172(2): 1155-1164.
Doerge RW, Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142(1): 285-294.
Doss S, Schadt EE, Drake TA, Lusis AJ (2005) Cis-acting expression quantitative trait loci in mice. Genome Res 15(5): 681-691.
Doyle P, Snell J (1984) Random walks and electric networks.
Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science (New York, NY 303(5659): 799-805.
Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C et al. (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. PLoS Genet 2(8): e130.
Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431(7004): 99-104.
Jansen RC, Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136(4): 1447-1455.
Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17(7): 388-391.
Jensen FV (2007) Bayesian Networks and Decision Graphs. New York: Springer-Verlag.
Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G et al. (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 29(4): 389-395.
Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1): 27-30.
Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 152(3): 1203-1216.
Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD (2006) Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics 62(1): 19-27.
Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G et al. (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. Proc Natl Acad Sci U S A 104(5): 1708-1713.

Kirst M, Myburg AA, De Leon JP, Kirst ME, Scott J et al. (2004) Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiol 135(4): 2368-2378.
Kline RB (2004) Principles and Practice of Structural Equation Modeling New York: The Guilford Press.
Klose J, Nock C, Herrmann M, Stuhler K, Marcus K et al. (2002) Genetic analysis of the mouse brain proteome. Nat Genet 30(4): 385-393.
Kruglyak L (2008) The road to genome-wide association studies. Nat Rev Genet 9(4): 314-318.
Kulp DC, Jagalur M (2006) Causal inference of regulator-target pairs by gene mapping of expression phenotypes. BMC Genomics 7: 125.
Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121(1): 185-199.
Lee SI, Pe'er D, Dudley AM, Church GM, Koller D (2006) Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci U S A 103(38): 14062-14067.
Li H, Lu L, Manly KF, Chesler EJ, Bao L et al. (2005) Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet 14(9): 1119-1125.
Li H, Chen H, Bao L, Manly KF, Chesler EJ et al. (2006) Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. Hum Mol Genet 15(3): 481-492.
Liu B, de la Fuente A, Hoeschele I (2008) Gene network inference via structural equation modeling in genetical genomics experiments. Genetics 178(3): 1763-1776.
Lum PY, Chen Y, Zhu J, Lamb J, Melmed S et al. (2006) Elucidating the murine brain transcriptional network in a segregating mouse population to identify core functional modules for obesity and diabetes. J Neurochem 97 Suppl 1: 50-62.
Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P et al. (2004) Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 75(6): 1094-1105.
Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG et al. (2004) Genetic analysis of genome-wide variation in human gene expression. Nature 430(7001): 743-747.
Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. Nat Genet 32(2): 261-266.
Petretto E, Mangion J, Pravanec M, Hubner N, Aitman TJ (2006a) Integrated gene expression profiling and linkage analysis in the rat. Mamm Genome 17(6): 480-489.
Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK et al. (2006b) Heritability and tissue specificity of expression quantitative trait loci. PLoS Genet 2(10): e172.
Potokina E, Druka A, Luo Z, Wise R, Waugh R et al. (2008) Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. Plant J 53(1): 90-101.
Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science (New York, NY 297(5586): 1551-1555.
Rieseberg LH, Archer MA, Wayne RK (1999) Transgressive segregation, adaptation and speciation. Heredity 83 (Pt 4): 363-372.
Rockman MV, Kruglyak L (2006) Genetics of global gene expression. Nat Rev Genet 7(11): 862-872.
Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell 20(5): 1199-1216.
Sampson JN, Self SG (2008) Identifying trait clusters by linkage profiles: application in genetical genomics. Bioinformatics 24(7): 958-964.
Schadt EE (2005) Exploiting naturally occurring DNA variation and molecular profiling data to dissect disease and drug response traits. Curr Opin Biotechnol 16(6): 647-654.

Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N et al. (2003) Genetics of gene expression surveyed in maize, mouse and man. Nature 422(6929): 297-302.
Schadt EE, Lamb J, Yang X, Zhu J, Edwards S et al. (2005) An integrative genomics approach to infer causal associations between gene expression and disease. Nat Genet 37(7): 710-717.
Segal E, Yelensky R, Koller D (2003a) Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics 19 Suppl 1: i273-282.
Segal E, Shapira M, Regev A, Pe'er D, Botstein D et al. (2003b) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34(2): 166-176.
Sieberts SK, Schadt EE (2007) Inferring causal associations between genes and disease via the mapping of expression quantitative trait loci. In: Balding DJ, Bishop M, Cannings C, editors. Handbook of Statistical Genetics: John Wiley & Sons, Ltd.
Stein CM, Song Y, Elston RC, Jun G, Tiwari HK et al. (2003) Structural equation model-based genome scan for the metabolic syndrome. BMC genetics 4 Suppl 1: S99.
Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. Proceedings of the National Academy of Sciences 100: 9440-9445.
Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3(8): e267.
Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1(6): e78.
Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP et al. (2007) Population genomics of human gene expression. Nat Genet 39(10): 1217-1224.
Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science (New York, NY 302(5643): 249-255.
Sun W, Yu T, Li KC (2007) Detection of eQTL modules mediated by activity levels of transcription factors. Bioinformatics 23(17): 2290-2297.
Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. Mol Syst Biol 4: 162.
Thaller G, Hoeschele I (2000) Fine-mapping of quantitative trait loci in half-sib families using current recombinations. Genet Res 76(1): 87-104.
Tu Z, Wang L, Arbeitman MN, Chen T, Sun F (2006) An integrative approach for causal gene identification and gene regulatory pathway inference. Bioinformatics 22(14): e489-496.
Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. Genetics 143(2): 1013-1020.
West MA, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW et al. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics 175(3): 1441-1450.
Wu C, Delano DL, Mitro N, Su SV, Janes J et al. (2008) Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. PLoS Genet 4(5): e1000070.
Yu T, Li KC (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis. Bioinformatics (Oxford, England) 21(21): 4033-4038.
Yvert G, Brem RB, Whittle J, Akey JM, Foss E et al. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 35(1): 57-64.
Zeng ZB (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proceedings of the National Academy of Sciences of the United States of America 90(23): 10972-10976.
Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136(4): 1457-1468.

Zeng ZB, Kao CH, Basten CJ (1999) Estimating the genetic architecture of quantitative traits. Genetical research 74(3): 279-289.
Zhu J, Zhang B, Smith EN, Drees B, Brem RB et al. (2008) Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 40(7): 854-861.
Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW et al. (2004) An integrative genomics approach to the reconstruction of gene networks in segregating populations. Cytogenet Genome Res 105(2-4): 363-374.