Incorporating Mouse Genome
Mammalian Genome
Genes and Phenotypes

A review of statistical methods for expression quantitative trait loci mapping

Christina Kendziorski,[1] Ping Wang[2]

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin, 1300 University Avenue (6729 MSC), Madison, Wisconsin 53706, USA
[2]Department of Statistics, University of Wisconsin, 1300 University Avenue (1245 MSC), Madison, Wisconsin 53706, USA

Received: 23 December 2005 / Accepted: 23 February 2006

Abstract

With high-throughput technologies now widely available, investigators can easily measure thousands of phenotypes for quantitative trait loci (QTL) mapping. Microarray measurements are particularly amenable to QTL mapping, as evidenced by a number of recent studies demonstrating utility across a broad range of biological endeavors. The early success stories have impelled a rapid increase in both the number and complexity of expression QTL (eQTL) experiments. Consequently, there is a need to consider the statistical principles involved in the design and analysis of these experiments and the methods currently being used. In this article we review these principles and methods and discuss the open questions most likely to yield significant progress toward increasing the amount of meaningful information obtained from eQTL mapping experiments.

Expression QTL data

The data collected in an expression quantitative trait loci (eQTL) mapping experiment consist minimally of a genetic map, marker genotypes, and microarray data collected on a set of individuals. Normalization is done to remove systematic effects within and between arrays to obtain measurements (phenotypes) that ideally provide an accurate quantification of gene expression levels. Related clinical phenotypes are also often collected. Most eQTL studies have taken place in experimental populations and we focus here on methods used in those studies, noting that many of the same issues apply to human studies as well. Studies with experimental populations involve arranging a cross between two inbred strains. Segregating progeny are then typically derived from a backcross or intercross. Brother-sister mating after the F_2 generation can also be done to generate recombinant inbred (RI) lines. For each offspring of the cross, markers are genotyped and expression phenotypes are collected via microarrays.

Experimental design

Many of the questions of eQTL experimental design also arise in QTL mapping experiments, microarray experiments, or both; and experience within each of these areas can be used to guide developments specific to eQTL studies. The most relevant questions include: How many subjects and markers should be used to achieve a desired power? What type of cross is best? At what level should replication be done? Should biological samples be pooled? What should be used as a reference (for two-color arrays)? Should selective genotyping and/or phenotyping be done?

The power to identify loci affecting a single quantitative trait in an intercross population depends on many factors: the number of individuals phenotyped, the number genotyped (which may differ from the number phenotyped if selective genotyping is used); the number, effects of, and relationship among segregating QTL; the type I error rate tolerated; the magnitude of environmental and genetic variance components; and the method of analysis used. There is much literature on this subject (for a comprehensive review, see Dupuis and Siegmund 1999 and references therein). Dupuis and

Correspondence to: C. Kendziorski; E-mail: kendzior@biostat.wisc.edu

Siegmund (1999) provide power calculations for a univariate quantitative trait for different intercross designs, carefully accounting for each consideration above.

Power calculations for microarray experiments are less well developed. The core difficulties stem from the high-dimensional nature of the data. Error rate control becomes more complex and there are additional sources of variability arising from the microarray experimental procedure that must be considered. These sources of variability are gene dependent, as are effect sizes, and this further complicates the problem. Early work extending traditional power calculations to microarray data accounted for some of these issues (Black and Doerge 2002; Cui and Churchill 2003; Jung et al. 2005a; Lee and Whitmore 2002; Pan et al. 2002) but were based on control of gene-specific type I error or overall type I error (i.e., the family-wise error rate). Recent efforts have considered the false discovery rate (FDR) and allow for gene-specific variability and effect sizes (Dobbin and Simon 2005; Gadbury et al. 2004; Hu et al. 2005; Jung 2005b; Muller et al. 2004)

There currently are no power calculations available to guide eQTL studies. One could perform a QTL sample size calculation for each expression trait in isolation and choose the sample size that yields identification of some percentage of the traits. Alternatively, one could consider the problem of grouping animals by genotype at each marker and identifying transcripts differentially expressed across the groups. Sample size calculations for traditional microarray studies could be used to guide such a marker-specific analysis. Determining sample sizes based on the former does not directly account for multiple tests across transcripts, while the latter does not account for multiplicities across markers. Certainly, improved methods that combine ideas from both areas are possible. There will be added complexities: hot spot identifications, rather than individual transcripts, may be of primary interest; eQTL interactions could be both within and among transcripts; and there will be various ways to control FDR, depending largely on whether multiple linkages are considered (see the section *Generating a list of mapping transcripts*). Ideally, a power formula for eQTL studies will account for these factors and address questions related to resource allocations.

A common question related to allocation is whether it is better to consider an F_2 of some size, or perhaps two backcrosses each half as big as the F_2. The optimal way in which samples should be allocated to arrays is also a question. Experience from QTL mapping and microarray studies gives some ideas. In particular, Dupuis and Siegmund (1999) indicate that intercross designs, in addition to being able to estimate dominance effects, are usually more powerful than backcross designs. Liu and Zeng (2000) concur. This result will most likely extend to eQTL studies. In studies of microarray experimental design, it has been shown that when identifying differentially expressed transcripts is of interest, resources should be spent on biological, not technical, replicates (Churchill 2002; Kendziorski et al. 2003; Kerr 2003). No eQTL study to date has used technical replicates and most studies would not benefit from doing so. It has also been suggested in the context of microarray studies that pooling subject samples can reduce the effects of biological variability and thereby reduce the number of arrays required in a given experiment (Churchill 2002; Kendziorski et al. 2003; Simon and Dobbin 2003). For this, subject samples within group are pooled before hybridization, providing an expression profile averaged across that group. In the context of eQTL studies, one would not want to pool across some subset of individuals from a backcross or an F_2. In these cases, each subject has a unique genotype profile and there exists a one-to-one correspondence between genotype and phenotype. This correspondence would not be preserved by pooling. For RI lines, there is a one-to-many correspondence between genotype and phenotype that allows pooling within genetically identical groups. Chesler et al. (2005) and Li et al. (2005a) considered brain tissue pooled across groups of three mice from an RI population. If samples are available, pooling across larger groups might provide further advantage, particularly if biological variability is much larger than technical for most genes (Kendziorski et al. 2003, 2005). The ratio of biological to technical variability will depend on the samples under study as well as the types of arrays being used.

In a standard microarray experiment using two-color arrays, an investigator must decide if a reference design is appropriate and if dye swaps are needed. It has been shown that a direct comparison is more efficient than a reference design when comparing two (or a few) groups (Churchill 2002; Yang and Speed 2002). For example, consider a simple experiment comparing treatments A and B. If, say, four chips are available, an investigator can hybridize both A (labeled in color 1) and B (in color 2) to the same chip and replicate once. The labels could be switched and the process repeated. Alternatively, a reference sample R can be used. In this case, A (color 1) is hybridized with R (color 2) and then repeated with the dyes swapped; the last two arrays compare B with R. Comparisons between A and B for this last case are made indirectly via the reference,

which inflates the variance of the ratio of interest [log(A/B)]. Despite this, reference (indirect) designs are useful when multiple groups are being compared, because the distance between any two samples is always two steps and all comparisons are made with equal efficiency. In eQTL experiments, it is not of primary interest to compare any two particular members of an intercross population, as in a direct comparison, but rather to make comparisons among subgroups of the population defined by marker genotypes. As a result, eQTL investigations to date have used a reference design. Brem et al. (2002) and Yvert et al. (2003) used parental RNA (BY4716) as a reference; Schadt et al. (2003, 2005) used a pool constructed from individual samples. In general, the reference should be plentiful, homogeneous, and stable over time (Churchill 2002). Each of these two-color eQTL experiments used a dye swap. An alternative to this that would reduce the number of arrays required would be to include dye swaps of randomly selected animals, or animals selected based on genetic distance.

Investigators have pointed out in the context of microarray studies that dye swaps are not always necessary (Dobbin et al. 2003a). It is well known that without the dye swap, treatment and dye effects are confounded (Kerr and Churchill 2001; Yang and Speed 2002); however, when a reference design is used, it has been argued that it may not be necessary to estimate and adjust for a potential dye effect since comparisons are made between samples labeled with the same dye. The idea is that any gene-specific dye bias between the samples of interest (nonreference samples) and the reference would cancel out when the nonreference samples are compared (Dobbin et al. 2003b). It has been shown empirically that this is not always the case, and inference can change depending on dye orientation, even when using a reference design (Dombkowski et al. 2004). For these reasons, we currently recommend that dye swaps be used for eQTL studies (at least on a selected set of samples).

A final design consideration is whether selective genotyping or phenotyping should be used. As genotyping costs continue to decrease, selective genotyping is used less frequently. No eQTL studies to date have used this approach. On the other hand, phenotyping is relatively expensive and selective phenotyping strategies have been developed (Jannink 2005; Jin et al. 2004). We used the method of Jin et al. (2004) in our own eQTL studies (Kendziorski et al. 2006; Lan et al. 2006). The approach identifies the group of animals with maximum recombination among pairs of markers. All markers can be used, or markers can be selected from genomic regions of interest. Jin et al. (2004) show via simulation that power is maximized in regions guided by prior information, while power in remaining regions is similar to that observed in the case of random sampling. The simulations use the same data for selection and mapping. As noted in their article, it might be useful to consider a two-stage approach where mapping results from a sparse map are used to guide selection. Additional genotyping could then be done in interesting areas. Further simulations to investigate and develop the two-stage procedure would be worthwhile as would further study of the theoretical properties of different selective phenotyping strategies, provided to some extent by Jin et al. (2004) and Sen et al. (2005).

Generating a list of mapping transcripts

Results from eQTL studies have been used for identifying hot spots (Brem et al. 2002; Bystrykh et al. 2005; Chesler et al. 2005; Hubner et al. 2005; Lan et al. 2006; Morley et al. 2004; Schadt et al. 2003), constructing gene networks (Bing and Hoeschele 2005; Chesler et al. 2005; Li et al. 2005a; Schadt et al. 2005; Zhu et al. 2004), elucidating subclasses of clinical phenotypes (Bystrykh et al. 2005; Schadt et al. 2003), and narrowing down lists of candidate genes (Bystrykh et al. 2005; Hubner et al. 2005; Schadt et al. 2003). Each of these tasks relies largely on the ability to generate a list of mapping transcripts and the genomic locations to which these transcripts map. As a result, the statistical methods used for list construction deserve some attention.

Since eQTL studies differ from traditional QTL studies only in number of phenotypes, it is perhaps not surprising that most efforts to localize eQTL use standard QTL mapping methods. Typically, a LOD profile is constructed for each transcript and then calibrated to adjust for multiple tests across markers. Multiplicities across transcripts are often not considered. Ideally, a statistical method for eQTL identification would properly account for multiplicities across the genome, multiplicities across transcripts, and correlations among transcripts. A repeated application of traditional QTL mapping to each transcript in isolation respects the first consideration. Some groups have attempted to deal in a second stage with the multiplicities across transcripts (Brem and Kruglyak 2005; Chesler et al. 2005; Hubner et al. 2005). There is currently no standard approach for doing this. Efforts have generally involved controlling FDR at some step in the analysis, but particular implementations have differed in detail.

Approaches to control FDR have relied on calculation of q-values as described in Storey and

Tibshirani (2003) (Chesler et al. 2005; Hubner et al. 2005) or on permutations (Brem and Kruglyak 2005). *Q*-Values are transformed versions of *p*-values that were developed to address the multiple-testing problem. The transformation uses the full distribution of *p*-values across all tests and allows for estimation of an overall FDR. For example, a list of tests for which the corresponding *q*-values are less than or equal to α controls the FDR for that list at α (provided some relatively mild conditions on the dependence of the *p*-values hold; see Storey and Tibshirani 2003 for details).

To implement this in eQTL studies, *p*-values corresponding to the peak LOD scores from each transcript have been used. Using these trait-specific *p*-values controls the FDR for a list of transcripts mapping to at least one location. Because only peak LOD scores are considered, this approach gives misleading information for transcripts mapping to multiple locations. Simply using multiple *p*-values per transcript in a *q*-value calculation is not recommended as this potentially violates dependence assumptions (Storey and Tibshirani 2003). A stepwise procedure would also be flawed because biases could be introduced (Storey et al. 2005). Similar considerations apply to the permutation-based estimation of FDR implemented by Brem and Kruglyak (2005).

Statistical methods designed specifically to control an overall FDR for single and multiple linkages are beginning to emerge. Kendziorski et al. (2006) proposed an empirical Bayesian approach to eQTL mapping that shares information across transcripts to determine a posterior probability that each transcript maps to each marker. The primary goal of their approach is to identify mapping transcripts; multiple eQTL are identified in a second stage using the posterior probabilities. Specifically, a genome region is considered linked to a trait if the associated posterior probability of linkage is in the upper 100 * $(1 - \alpha)$% of all probabilities for that trait (α is often taken to be 5%). These highest posterior density (HPD) regions allow for multiple-eQTL identification of mapping transcripts. Storey et al. (2005) propose an approach specifically designed to identify multiple eQTL per transcript (they focus on two) and estimate the FDR associated with the multiple identifications. They also allow for epistatic effects.

A common feature of these approaches is that adjustments for multiple tests across *both* markers and transcripts are considered. Furthermore, both approaches demonstrate an increase in power obtained when information is shared across transcripts. An advantage of Storey et al. (2005) is that FDR is estimated precisely for multiple linkages. However, the approach of Storey et al. (2005) is not designed to identify a relatively large number of loci with small effects; the HPD regions of Kendziorski et al. (2006) are better suited for this task. For example, in Kendziorski et al. (2006), approximately 20% of identified mapping transcripts showed only moderate effects, most of which were not statistically significant on their own. Further investigation of these approaches and additional developments along these lines should prove useful.

Identifying hot spots

Whatever statistical method is used for list generation, once obtained, ''hot spots'' are often of primary interest. Hot spots are genomic regions where an abundance of transcripts map, and to date they have been found in a straightforward way. At each genome region, the total number of mapping transcripts is tallied. Hot spot candidates are those regions with the highest totals. Although intuitive, it is not clear that this is the best way to define hot spots, particularly if there are numerous loci with moderate effects that perhaps do not reach the level of statistical significance.

Kendziorski et al. (2006) considered five statistical methods that could be used to generate lists of mapping transcripts. They then identified hot spot candidates in two ways: by counting the number of mapping transcripts, as described above, and by summing evidence in favor of mapping (e.g., as assessed by LOD score) across every transcript whether it exceeded a significance threshold or not. They found increased agreement among methods when all transcripts were used. This is consistent with a system in which most transcripts are affected by multiple eQTL with moderate effect size. Brem and Kruglyak (2005) provide evidence for this in yeast. Each of the methods considered in Kendziorski et al. (2006) uses the mapping information provided by individual transcripts for hot spot identification. Profiles averaged across correlated transcripts (Yvert et al. 2003) or profiles from sets of correlated transcripts that are functionally related (Lan et al. 2006) could also be used. This last approach has been shown to improve the power for eQTL hot spot identification.

Whatever becomes the best way to define potential hot spots, statistical tests are required to determine with some confidence which spots are truly hot. Perez-Enciso (2004) considers a number of scenarios that can lead to spurious identifications or ''ghost'' hotspots. These considerations are addressed to some extent by the statistical test provided by Brem et al. (2002). They propose a Poisson-

based test that calculates the probability that a particular genome region would have at least n transcripts linked to it if in fact there were no hot spots. This test should prove useful in similar studies that define hot spot candidates by number of transcripts exceeding some significance threshold. New statistical tests are required to address the case in which hot spot candidates are defined by summing the evidence of linkage across all transcripts.

Tests for enrichment have also been done in an attempt to validate hot spot candidates. Most of these tests rely on hypergeometric calculations that compare the proportion of transcripts with a particular biological function to the proportion with that function that map to the region under question. The appropriate thresholds for these tests is not obvious since there are thousands of functional groups that are tested, and, furthermore, small p-values can result when testing functional groups with many, or just a few, transcripts (Gentleman 2005). Statistical tests for enrichment are being improved in the context of standard microarray studies (Barry et al. 2005; Subramanian et al. 2005) and should prove useful in eQTL studies as well, perhaps with modifications to address the increased number of tests at multiple-genome regions.

Networks

The identification of hot spots provides lists of comapping transcripts and often leads to the inspection of putative candidates controlling the collection. The idea is that comapping is the result of comembership in a biological pathway, an idea similar to that put forth in Eisen et al. (1998), where functional information was inferred using temporally correlated transcripts. Jansen and Nap (2001) were perhaps the first to formally recognize how hot spot lists could be used to construct networks.

Mathematically, a network is a collection of nodes (or vertices) and edges. Here, nodes are genes or transcripts and an edge exists between nodes when there is some relationship between them (oftentimes measured via a correlation coefficient). Elucidating the precise actions of and interactions between nodes is a difficult challenge, but promising strategies are beginning to emerge in the context of eQTL mapping experiments.

Chesler et al. (2005) use pairwise correlations among all transcripts to identify cliques, i.e., sets of transcripts completely connected by edges. The cliques themselves provide information about the relationship among members in that two transcripts are connected by an edge that indicates extent of correlation. Mapping regions common to clique members, and perhaps also to clinical traits, are studied further to identify candidates likely affecting the pathway. Utility of the approach is demonstrated in a study of neural synapse function.

Complementary approaches that allow for the elucidation of potentially causal relationships among transcripts are also being developed. Bing and Hoeschele (2005) identify eQTL confidence regions and narrow down the number of candidates within a region by requiring high correlation between the candidates and affected transcripts. Networks are constructed by drawing directional edges between retained candidates and downstream transcripts. This type of approach assumes that transcripts belonging to the same network have strong correlations between their expression values, as indicated above. For most cases, the assumption is likely necessary but not sufficient. In other words, it will often be the case that genes within a network are correlated; however, other scenarios such as independent control by closely linked loci can give rise to correlated traits that are in distinct pathways. As noted in their article, the approach of Bing and Hoeschele (2005) can be extended to incorporate multitrait mapping methods so that this issue can be resolved.

The issue is one of many addressed in the context of Bayesian networks, where effective algorithms exist for finding the ''best'' model in some model space (for an introduction to Bayes nets, see Jensen 2001). ''Best'' can be defined in different ways, but often the definition involves calculation of a penalized likelihood that balances the goodness of fit of the model to data and number of model parameters; the model space must be moderately sized to make the problem computationally feasible. Reducing the model space for eQTL mapping often starts with considering only those transcripts that map to at least one location. In Li et al. (2005a), the number of transcripts considered for network reconstruction was reduced to approximately 200, and the model space was further reduced by using SNP genotype information to narrow down the number of possible regulatory nodes. In Zhu et al. (2004), the information provided by the eQTL map is used to reduce both the number of possible nodes and model complexity. In particular, measures on pairs of transcripts are considered. Only those pairs of transcripts with highly correlated LOD profiles and large mutual information measures are considered (the latter provides some evidence for pleiotropy as opposed to multiple tightly linked loci). This narrows down the list of transcripts considerably ($\sim$1000) and provides selection for pairs that may be causally related. For a pair of transcripts, to distin-

guish between independent control by a common eQTL and a causal relationship where an eQTL affects one transcript that in turn affects the other, Zhu et al. (2004) use eQTL overlap information, as initially prescribed by Jansen and Nap (2001). In short, they assume that if, say, X maps to a number of locations and Y maps to some subset of those (perhaps with higher LODs), then it is likely that Y controls X. Schadt et al. (2005) further develop this approach to allow for incorporation of clinical data and elucidation of both causal and reactive relationships among transcripts and clinical traits. The power of this approach is maximized when relatively homogeneous clinical traits can be measured, or derived.

Augmenting traditional QTL studies

Schadt et al. (2003) provide a beautiful example of how eQTL data can be used to define clinical traits more precisely, thereby reducing the challenges imposed by genetic heterogeneity. By clustering a collection of genes differentially expressed between mice varying in fat pad mass (fpm), they identified two distinct groups within the high fpm group, confirming some degree of heterogeneity within the fpm trait. QTL mapping of the groups separately (low fpm vs. high fpm 1 and low fpm vs. high fpm 2) identified two nonoverlapping genomic regions, suggesting independent control of subsets of the fpm trait. When using only fpm, a second peak was missed and the primary peak had reduced LOD.

Optimizing and automating this or a similar approach on a large scale will surely help elucidate the genetic basis of complex traits. Doing so requires addressing a number of questions: Given a population of animals for which clinical phenotypes are available, which animals will prove most powerful for identifying the differentially expressed genes that will then be used to detect heterogeneity? How many differentially expressed genes should be used? Can and should the process of identifying subgroups be automated (Schadt et al. 2003 discovered the subgroups by visual inspection of a cluster plot)? These remain important open questions.

Discussion

In this review we have summarized information on the statistical methods currently available for the design and analysis of eQTL mapping studies, the ways in which these methods can best be used, and the most promising avenues for the development of new statistical methods. Much of what we have learned from traditional microarray and QTL mapping experiments has given insights into addressing the questions posed by eQTL mapping experiments. For example, studies of microarray experimental design suggest that for an eQTL mapping experiment using two-color arrays, dye swaps should be used on at least some selected animals so that any dye effects can be estimated. Pooling can also be useful in reducing the effects of biological variability within genetically identical subgroups of RI lines. A design consideration specific to the eQTL mapping setting is selective phenotyping; and recent studies indicate that the practice is a useful one. Further evaluation and methodologic extensions should further increase the utility of selective phenotyping approaches. Another experimental design question requiring attention is the calculation of power. There are currently no power calculations available for eQTL studies. Until appropriate power calculations are available, calculations from standard QTL mapping (or microarray) experiments can be applied transcript by transcript (or marker by marker) and used as suggested here in the section on experimental design. This will provide at least ballpark estimates on sample sizes.

For eQTL experiments that are not well powered, strict thresholding to control FDR can yield few interesting results. It should be remembered that there are different methods used to control FDR and these methods are often applied at different steps in an eQTL data analysis. Both the q-value approach and estimation via permutations have been used to estimate false positive rates across maximum LODs per transcript or across multiple identifications within and among transcripts. It is good to remember that these approaches provide an estimate, and the assumptions upon which the estimation is made should always be considered. In addition, it is well known but not widely appreciated that FDR estimates can have high variance (Efron 2005). These reasons provide good justification for lowering FDR significance thresholds in many cases. For example, Chesler et al. (2005) tolerate an FDR estimate of 25% (calculated across a set of heritable transcripts). When automated *in silico* approaches are used to validate and further refine lists, relaxing an FDR measure perhaps poses little trouble. On the other hand, a close consideration of the assumptions made in estimation, the properties of the estimators, and the implications of associated thresholds should be given when expensive followup experiments are planned on a list of identified transcripts (a rigorous consideration of choosing thresholds in light of competing goals is given in Muller et al. 2004).

Fortunately, statistical methods are beginning to emerge that precisely define the models used in FDR estimation and the conditions under which FDR

control is obtained for eQTL identifications (Kendziorski et al. 2006; Storey et al. 2005). An advantage of Storey et al. (2005) is that FDR is estimated in the context of multiple linkages. However, the approach of Storey et al. (2005) uses statistically significant eQTL identified in a first step of the analysis and is therefore not amenable to identifying a relatively large number of loci with moderate effects. The approach of Kendziorski et al. (2006) does not require significant linkage at any one location and is therefore better suited for this task. Further investigation of these approaches and additional developments along these lines should prove useful. It would also be useful to consider more carefully correlations across transcripts, which are most likely imposed not only by the biology but also by the technology. Both Storey et al. (2005) and Kendziorski et al. (2006) assume that sufficient preprocessing has been done so that correlations from the latter source have been minimized. Statistical methods that uncover and exploit the former are sure to improve inferences.

Statistical methods are also being applied for eQTL network construction. Most approaches are extensions of existing methods with modifications made to best use the information in eQTL mapping data. Identifying the network that best describes the actions of and interactions among large sets of transcripts is computationally and algorithmically challenging; many of the current approaches require that the number of transcripts be greatly reduced. A straightforward approach is to consider mapping transcripts, perhaps further restricting the set to those that are correlated, colocalized, and functionally related to a traditional quantitative trait of interest. Further consideration of optimal methods for sifting through candidate nodes and reducing model complexity is required.

It has been suggested that some biological networks are scale free (SF) (Jeong et al. 2000). If in fact this is the case, incorporating properties of SF networks could be used to help reduce model complexity. For example, the distribution of the numbers of edges connecting a random node in a SF network is a power law and, consequently, there are a few nodes that are highly connected. This is just one property inherent to an SF network. Li et al. (2005b) precisely consider all properties of SF networks and provide rigorous definitions of the necessary and sufficient conditions. At this point, incorporating properties of SF networks into eQTL studies may be premature because Li et al. (2005b) provide good reason to question the relevance of SF networks in many biological systems.

The methods that eventually prove most useful for eQTL network construction will benefit tremendously by considering the errors involved in estimation. As discussed in this review, a relationship identified as associative or causal is not guaranteed to be so. Unfortunately, there are currently no statistical methods for assessing confidence in large-network predictions. Ideas for assessing estimation errors in phylogenetic inference (Larget and Simon 1999) might guide future developments in eQTL applications. Of course, biological validation studies such as the single-gene perturbation experiments done in Schadt et al. (2005) and Mehrabian et al. (2005) are the gold standard for testing network predictions, and this is always recommended when possible.

In addition to biological validation, we recommend practices that allow for statistical validation. We now know that results from microarray studies can be misleading, as evidenced by several high-profile articles reporting results that have been impossible to reproduce (see PLoS Medicine Editors 2005 and references therein). Three articles in the May 2005 issue of *Nature Methods* (Larkin et al., Irizarry et al., and Weis et al.) find that differences in microarray platforms are most likely not the *main* reason for discrepancies in results. They show that with standardized protocols for sample preparation *and* data analysis, results for most genes are reliable across platforms and experiments.

The Microarray Gene Expression Data Society has made a laudible effort to address the reproducability problem at the sample preparation and processing level by establishing a set of guidelines (MIAME) for reporting microarray data experimental protocols and requiring that all such data be deposited in public databases. This will certainly benefit the eQTL mapping community. Those developing statistical methods and performing data analysis should follow suit. To this end, Ruschhaupt et al. (2004) proposed an R-based system (R Development Core Team 2004) to facilitate not only publication of raw data but also the detailed statistical methods, computer code, documentation, and derived data associated with a study. Whatever system is used for data analysis, this type of public availability will allow for more rapid testing and facilitate comparative studies across methods.

Acknowledgments

The authors thank Alan Attie, Meng Chen, Michael Newton, and Brian Yandell for useful discussions and two anonymous reviewers for comments that improved the manuscript. They also thank Stephanie Ciatti for extra help at home.

References

1. Barry WT, Nobel AB, Wright FA (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 21, 1943–1949
2. Bing N, Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170, 533–542
3. Black MA, Doerge RW (2002) Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments. Bioinformatics 18, 1609–1616
4. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5700 gene expression traits in yeast. Proceedings of the National Academy of Sciences 102, 1572–1577
5. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752–755
6. 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, 225–232
7. 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, 233–242
8. Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nat Genet 32, 490–495
9. Cui X, Churchill GA (2003) How many mice and how many arrays? Replication in mouse cDNA microarray experiments In: *Methods of Microarray Data Analysis III*, Johnson KF, Lin SM (eds.) (Norwell MA: Kluwer Academic Publishers) pp 139–154
10. Dobbin K, Simon R (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 6(1), 27–38
11. Dobbin K, Shih JH, Simon R (2003a) Statistical design of reverse dye microarrays. Bioinformatics 19(7), 803–810
12. Dobbin K, Shih JH, Simon R (2003b) Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. J Natl Cancer Inst 95(18), 1362–1369
13. Dombkowski AA, Thibodeau BJ, Starcevic SL, Novak RF (2004) Gene-specific dye bias in microarray reference designs. FEBS Lett 560, 120–124
14. Dupuis J, Siegmund D (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151, 373–386
15. Efron B (2005) Local False Discovery Rates. Available at http://www-stanford.edu/~brad/papers/. Last accessed April 21 2006
16. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95, 14863–14868
17. Gadbury GL, Page GP, Edwards JW, Kayo T, Prolla TA, et al. (2004) Power and sample size estimation in high dimensional biology. Stat Methods Med Res 13, 325–338
18. Gentleman R (2005) Using GO for Statistical Analyses, Bioconductor vignette http://www.bioconductor.org
19. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, et al. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 37, 243–253
20. Hu J, Zou F, Wright FA (2005) Practical FDR-based sample size calculations in microarray experiments. Bioinformatics 21(15), 3264–3272
21. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2, 345–350
22. Jannink JL (2005) Selective phenotyping to accurately map quantitative trait loci. Crop Sci 45, 901–908
23. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17, 388–391
24. Jensen FV (2001) Bayesian Network and Decision Graphs. In *Statistics for Engineering and Information Science* (New York: Springer-Verlag)
25. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL (2000) The large scale organization of metabolic networks. Nature 407, 651–653
26. Jin C, Lan H, Attie AD, Bulutuglo D, Churchill GA, et al. (2004) Selective phenotyping for increased efficiency in genetic mapping studies. Genetics 168, 2285–2293
27. Jung S-H, Bang H, Young S (2005a) Sample size calculation for multiple testing in microarray data analysis. Biostatistics 6(1), 157–169
28. Jung S-H (2005b) Sample size for FDR-control in microarray data analysis. Bioinformatics 21(14), 3097–3104
29. Kendziorski C, Zhang Y, Lan H, Attie AD (2003) The efficiency of mRNA pooling in microarray experiments. Biostatistics 4, 465–477
30. Kendziorski C, Irizarry RA, Chen K, Haag JD, Gould MN (2005) On the utility of pooling biological samples in microarray experiments. Proc Natl Acad Sci USA 102(12), 4252–4257
31. Kendziorski C, Chen M, Yuan M, Lan H, Attie AD (2006) Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics 62, 19–27
32. Kerr K (2003) Design considerations for efficient and effective microarray studies. Biometrics 59(4), 822–828
33. Kerr K, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2, 183–201
34. Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, et al. (2006) Combined expression trait correlations and expression quantitative trait locus mapping. PLoS Genet 2, e6
35. Larget B, Simon D (1999) Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 16, 750–759

36. Larkin JE, Frank BC, Gavras H, Sultana R, Quackenbush J (2005) Independence and reproducibility across microarray platforms. Nat Methods 2, 337–344
37. Lee MT, Whitmore GA (2002) Power and sample size for DNA microarray studies. Stat Med 21, 3543–3570
38. Li H, Lu L, Manly KF, Chesler EJ, Bao L, et al. (2005a) Inferring gene transcriptional modulatory relations: a genetical genomics approach. Hum Mol Genet 14(9), 1119–1125
39. Li L, Alderson D, Doyle JC, Willinger W (2005b) Towards a theory of scale-free graphs: definition, properties, and implications. Internet Mathematics 2(4), 431–523
40. Liu Y, Zeng ZB (2000) A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genet Res 75, 345–355
41. Mehrabian M, Allayee H, Stockton J, Lum PY, Drake TA, et al. (2005) Integrating genotypic and expression data in a segregating mouse population to identity 5-lipoxygenase as a susceptibility gene for obesity and bone traits. Nat Genet 37, 1224–1233
42. 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, 743–747
43. Muller P, Parmigiani G, Robert C, Rousseau J (2004) Optimal sample size for multiple testing: the case of gene expression microarrays. J Am Stat Assoc 99, 990–1001
44. Pan W, Lin J, Le CT (2002) How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. Genome Biol 3(5), research0022
45. Perez-Enciso M (2004) In silico study of transcriptome genetic variation in outbred populations. Genetics 166, 547–554
46. R Development Core Team (2004) R: *A language and environment for statistical computing* (Vienna, Austria: R Foundation for Statistical Computing)
47. Ruschhaupt M, Huber W, Poustka A, Mansmann U (2004) A compendium to ensure computational reproducibility in high-dimensional classification tasks. Statistical Applications in Genetics and Molecular Biology 3(1), article 37
48. 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, 297–302
49. 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, 710–717
50. Sen S, Satagopan J, Churchill GA (2005) QTL study design from an information perspective. Genetics 170, 447–464
51. Simon RM, Dobbin K (2003) Experimental design of DNA microarray experiments. BioTechniques Suppl, 16–21
52. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100(16), 9440–9445
53. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. PLoS Biol 3(8), e267
54. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 102, 15545–15550
55. Weis BK, Members of the Toxicogenomics Research Consortium (2005) Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods 2(5), 351–356
56. Yang YH, Speed TP (2002) Design issues for cDNA microarray experiments. Nat Rev Genet 3, 579–588
57. 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, 57–64
58. 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, 363–374