Computational tools for discovery and interpretation of expression quantitative trait loci

Expression quantitative trait locus (eQTL) analysis is rapidly moving from a cutting-edge concept in genomics to a mature area of investigation, with important connections to genome-wide association studies for human disease, pharmacogenomics and toxicogenomics. Despite the importance of the topic, many investigators must develop their own code or use tools not specifically suited for eQTL analysis. Convenient computational tools are becoming available, but they are not widely publicized, and investigators who are interested in discovery or eQTL, or in using them to interpret genome-wide association study results may have difficulty navigating the available resources. The purpose of this review is to help investigators find appropriate programs for eQTL analysis and interpretation.

KEYWORDS: bioinformatics ▪ fast linear modeling ▪ gene expression

Fred A Wright[1] Andrey A Shabalin[1] & Ivan Rusyn*[2]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA
[2]Department of Environmental Sciences & Engineering, University of North Carolina, Chapel Hill, NC 27599, USA
**Author for correspondence: Tel.: +1 919 843 2596 iir@unc.edu*

Studies of the genetic basis of complex traits are contingent upon availability of the molecular biology data. Hence, it is not surprising that in the past decade, the pace of biomedical discovery has followed rapid advancements in technology, providing an unprecedented level of precision for studies of the genome and its products. It has long been suggested that the relationship between DNA sequence variation and variation in heritable phenotypes may be intimately dependent on the abundance of transcripts, proteins and other gene products. The foundational work on the genetics of natural variation in gene expression has a long history, dating back to the 1930s [1]. One of the first reports that provided large-scale experimental evidence for quantitative trait loci (QTL) underlying gene-product variation, was a study that analyzed the genetic determinism of the quantitative variation of 72 proteins separated by high-resolution 2D polyacrylamide gel electrophoresis in an F2 population of maize [2].

As a whole-genome expression-profiling technology reached maturity and became a mainstream technology, numerous genome-wide genetic analyses of gene expression have used the conceptual model of Damerval *et al.* to provide key information on the molecular underpinnings of the phenotypic variation among individuals [2]. Strain differences in gene expression in mammals [3] and plants [4], as well as segregation of gene-expression traits in crosses [5], have been reported. That the abundance of a transcript in a particular cell or tissue can be reliably quantified, and thus used as a quantitative trait, has led to the application of the linkage- and association-mapping tools to studies of gene expression on a population level in humans [6–8].

Jansen and Nap proposed a general concept of 'genetical genomics' as a merger of genetic mapping and gene-expression profiling based on data from microarray studies and marker-based fingerprinting of individuals in a segregating population, followed by a statistical analysis of QTL [9]. The first linkage study that mapped global gene expression in a yeast cross provided empirical evidence for the existence of heritable variation in genome-wide gene expression [10]. In the past decade, scores of studies have been published reporting expression QTL (eQTL)-mapping results for multiple species and organisms [1,11,12]. Information on the genetic control of gene expression is now an important component of the systems biology approaches that are being used to understand the molecular events underlying human disease, interindividual variability in susceptibility to environmental and lifestyle etiological factors, efficacy and potential toxicity of therapeutic interventions and disease prognosis [13–15].

Statistical methods for eQTL mapping

Early eQTL studies relied on segregating populations to take advantage of limited genotype density. The classical genetic methods for linkage or association mapping were adapted to eQTL analyses, with attendant challenges of scalability presented by thousands of expression quantitative traits [16]. Although the number of transcripts has remained relatively constant (in the order of ~10^4–10^5), higher-density genome

Future Medicine part of fsg

scan arrays can now have over 1 million SNPs. Thus, even a basic eQTL analysis becomes a computational challenge that requires billions of statistical tests to be performed. Moreover, sequencing technologies are further revolutionizing genomics by providing data on rare variants through high-throughput approaches [17]; however these create further strains on the analytic procedures. In summary, additional and more complex genetic and genomic data are becoming routine, creating a tremendous need for efficient eQTL analytic procedures.

The general steps and principles of eQTL analyses have been recently reviewed [1,18,19]. Issues of experimental design, sample size and considerations of the statistical power are considered in detail by Kendziorski and Wang [16]. While the processing of data collected in both genotyping and gene-expression experiments, including normalization and quality control, may also be challenging with some of the newer technologies [20], the major impediments of eQTL analyses lie in the underlying statistical assumptions and ensuring proper error control [21], as well as the need for computational efficiency [22]. Batch-effect correction – of either population stratification or technical artifacts – is not automatically performed by the eQTL packages reviewed here. However, batch-effect covariates derived using additional specialized statistical approaches can be handled by several of the eQTL packages (Table 1) [23,24]. Alternative approaches to handle latent batch effects via estimation of the data covariance structure are mostly unavailable with desktop software, and as such, covariate correction remains an active area of study [25].

At its essence, eQTL analysis involves correlating expression for each transcript with underlying genetic information at each locus, which can take the form of identical by descent information from parental strains [10], genotype [19] or copy number values [26]. Although the underlying transcriptional relationships may be complex and inter-related, most eQTL studies begin with single-locus, single-expression trait analysis, often using standard linear-regression models that are performed exhaustively. The use of normal error linear models is not as restrictive as it seems, as it includes three-genotype (i.e., analysis of variance) modeling as a special case, and has been applied to RNA-sequencing count data after variance-stabilizing transformations [27]. The inclusion of covariates is conceptually straightforward, although many of the software packages described below do not allow covariates, except by prior adjustment of the phenotypes (i.e., expression values). In addition, it is often convenient to handle genotypes on a continuous scale, so that genotype 'dosage' values (imputed values that represent an average over uncertain genotypes) may be used [28]. The results from the single eQTL analyses may then be used for downstream modeling, typically using only a small subset of DNA–transcript pairs achieving statistical significance.

While single-QTL mapping approaches have been useful in advancing the field of genetical genomics, several groups have reasoned that these methods have limitations in their ability to identify interactions and eQTLs of small effect [29]. As it is generally accepted that molecular pathways represent networks of interacting proteins, the ability to identify coregulated transcripts becomes an important feature for eQTL interpretation. Several recent statistical methods address the issue of multiple-eQTL analysis. Wang *et al.* proposed a clustering method that identifies groups of transcripts that may share similar biological functions by taking advantage of the underlying trait-correlation structure [29]. The latter approach identifies coexpression coregulation modules, an approach that facilitates combined analysis of expression-trait correlations and eQTL mapping results for discovery of, and improved confidence in, molecular networks that may be linked to a particular disease [30]. Additional work on correlated eQTL 'modules' indicates that they may be mediated by transcription factors [31,32].

Software tools for eQTL analysis

A number of groups have introduced standalone or web-based software tools that are either designed, or can be adapted for, eQTL analysis. Below we detail several software tools that have been used for eQTL analysis and provide a short description of the advantages and limitations of each package, with a focus on the ease of use for the wider scientific community (Table 1). Several of the packages were not specifically developed for eQTL analysis, and thus our speed comparisons should be viewed in that context. However, such packages may already be used in individual laboratories, and therefore we include those that, in our opinion, may be scaled for eQTL analysis on data from genome-wide array platforms.

R/qtl [101] is a comprehensive R package for QTL mapping, initially configured for individual traits (in eQTL studies, one trait = one transcript) [33]. It can perform single- and two-QTL

Table 1. Expression quantitative trait-loci tools.

Tool	Interface	Speed† (min)	Cross platform	Additive 1-df model	ANOVA 2-df model	Support for covariates	Multiple-testing correction‡	Main features	Ref.
R/qtl	R package	24	Yes	No	Yes	Yes	FDR (MQM only)	QTL in crosses, multiple-QTL mapping	[33,101]
eMap	R package	256	No	Yes	Yes	No‡	No	eQTL analysis, visualization	[102]
SNPster	Web-based	–§	Yes	No	Yes	No‡	gFWER	Haplotype-association testing	[34,103]
snpMatrix	R package	40	Yes	Yes	Yes	Yes	No	Genome-wide association testing	[35,104]
Plink	Command line	125	Yes	Yes	Yes	Yes	Bonferroni and FDR	Genome-wide association analysis toolset	[36,105]
Merlin	Command line	280	Yes	Yes	No	Yes	Simulations	Fast-pedigree analysis	[37,106]
Qxpak.5	Command line	–§	No	Yes	Yes	Yes	No	Mixed model-based eQTL analysis	[38,107]
FastMap	Graphical	6	Yes	Yes	Yes	No‡	Permutations and FDR	Fast eQTL analysis with a graphical user interface	[22,108]
Matrix eQTL	R and Matlab packages	0.11	Yes	Yes	Yes	Yes	FDR	Ultrafast eQTL analysis using linear models	[109]

†Speed of each tool was tested on a dataset of moderate size (57,000 SNPs, 2200 transcripts and 840 individuals), additive 1 df model, with no covariates and no permutations. R/qtl does not support 1 df model, so it was set to the analysis of variance model, which enables fitting of dominance effects.
‡All methods can perform approximate covariate correction by residualization of expression values.
§Speed could not be determined for SNPster and Qxpak.5.
ANOVA: Analysis of variance; df: Degrees of freedom; eQTL: Expression quantitative trait loci; FDR: False-discovery rate; gFWER: Generalized family-wide error rate; MQM: Multiple QTL mapping; QTL: Quantitative trait loci.

genome scans, interval mapping (typically most useful in a linkage context) and multiple QTL mapping. It can also perform imputation and genotype error-correction for experimental crosses using a hidden Markov model. Although R/qtl was originally designed for standard QTL analysis, it scales well for the massive task of eQTL analysis. In our tests, R/qtl showed generally fast performance. The R package 'eqtl' is a 'wrapper' for R/qtl, simplifying the process of analyzing eQTL data. R/qtl can estimate false-discovery rates to correct for multiple comparisons. However, this option is available for multiple QTL mapping only.

eMap [102] is an R package used for eQTL mapping that uses linear models. Its main features are model selection using backward selection, visualization and detection of eQTL hotspots. eMap was designed for eQTL analysis and shows average performance on large datasets. The package is partly written in C and requires GNU Scientific Library to compile, and does not readily support installation on Windows.

SNPster [103] is an online tool for eQTL analysis in mouse populations, and is designed for datasets with only relatively sparse genotyping density [34]. It tests for association using an analysis of variance model over k-SNP genotype windows. The significance of the findings can be estimated using parametric models or permutations. SNPster can estimate the generalized family-wise error rate to correct for multiple comparisons. Although SNPster is no longer in development and cannot handle modern large datasets, it remains potentially useful for smaller datasets.

snpMatrix [104] is a part of an R/Bioconductor toolset designed for genotype management and association analysis [35]. It supports several common input file formats, including Plink, PED

and HapMap files. In our tests, snpMatrix showed generally fast performance.

Plink [105] is a very popular command-line tool for human genome-wide association studies (GWAS), designed to perform a range of basic, large-scale analyses [36]. Many researchers find Plink extremely useful for data manipulation tasks, including data filtering, preprocessing and merging. Plink was designed for association analysis, including for quantitative traits, and it is not optimized for analysis of multiple traits. For our eQTL experiments with no covariates (--assoc), Plink demonstrated good performance; however, inclusion of even a single covariate (--linear) made it approximately ten-times slower. Plink supports a variety of adjustments for multiple testing, including the false-discovery rate, Bonferroni correction and Sidak's adjusted p-values.

Merlin [106] is a command-line tool for pedigree analysis [37]. Merlin can perform a range of tasks, including QTL analysis, linkage, error detection and haplotyping. QTL analysis in datasets with unrelated individuals can be performed by identifying each sample as belonging to a separate family. Merlin employs likelihood ratio and score-test statistics. Although Merlin was designed for pedigree analysis, it is reasonably well suited for eQTL analysis. When run in its 'fast'-association testing mode (--FastAssoc), Merlin had slower performance than other packages. In addition, Merlin can simulate datasets with characteristics of the original data, but with no associations, to estimate the number of false findings.

Qxpak.5 [107] is a command-line tool for various statistical genetic analyses using mixed models [38]. Its main features are linkage and association testing with fixed and mixed effects models using maximum likelihood and QTL analysis with single or multiple QTLs. Qxpak.5 has specific input-file format specifications not shared by other packages. In addition, Qxpak.5 does not scale well for the task of eQTL analysis. When multiple phenotypes are specified, Qxpak.5 creates a temporary file for each SNP, a total of 80 GB for just 57,000 SNPs of our test dataset. This tool only has 64-bit Linux and 32-bit Windows versions.

FastMap [108] is a Java-based eQTL tool that has a graphical user interface [22]. FastMap visualizes findings for any given transcript using Manhattan plots with a zoom function, and provides one-click connection to the University of California Santa Cruz (CA, USA) Genome Browser. Due to employed optimizations for discrete genotypes, FastMap cannot fully handle covariates in the analysis, except via preadjustment of the transcripts. FastMap is designed for fast eQTL analysis and was noticeably faster than all other tools in our tests, except Matrix eQTL. FastMap uses a two-step procedure to correct for multiple comparisons. For each gene it estimates the significance of the most strongly-associated SNP using permutations. Next, false-discovery rates are estimated for the obtained gene-level p-values.

Matrix eQTL [109] is a new R and Matlab package for fast eQTL analysis. Matrix eQTL can account for correlated and/or heteroskedastic errors if the error-covariance matrix is provided. Matrix eQTL accounts for multiple comparisons by estimation of the false-discovery rate, with the option of separate false-discovery rate calculations for *cis*- and *trans*-eQTLs. In our experiments, Matrix eQTL was two to three orders of magnitude faster than all other tools.

Scalability of the computational tools

The availability of faster and cheaper genotyping tools, coupled with considerable accuracy of imputation routines, has led to denser genotyping on more individuals. The dramatic increase in the number of gene–SNP pairs in eQTL analysis is thus creating serious impediments in computational and memory requirements, and it is infeasible to store the test statistics for all gene–SNP pairs. The challenge of potential scalability of the eQTL tools to bigger datasets thus becomes an important consideration. The R/qtl and snpMatrix packages store all test statistics in memory. Therefore, to analyze a large dataset, the user must apply the packages to one portion of the data at a time, manually filtering and aggregating the results. As we detailed above, Qxpak.5 and SNPster are not suitable for datasets of a large size. By contrast, we judge that Plink, Merlin, eMap, FastMap and Matrix eQTL can be more readily applied to large datasets, as they can filter the associations based on a user-defined significance threshold.

Input data formats

Although there are some commonalities in the data formats, for the most part, each tool has its own specific file format requirements. Plink takes three files as input: .MAP, with the SNP location locations; .PED, with pedigree information and genotype; and a phenotype file with expression data. Merlin requires three different input files: .MAP, with SNP locations; .PED, with pedigree

information, genotype and phenotype; and a .DAT file with annotation for the .PED file. Matrix eQTL, FastMap, R/qtl and eMap can accept data in simple-matrix format with expression and genotype data in separate files. The files must have one SNP/gene per line. FastMap can also accept HapMap and transposed Plink format. R/qtl also supports transposed input files with one sample per line. Qxpak.5 input files include: .PAR, with model parameters; .MKR, with genotypes; .DAT, with expression; and .PED, with pedigree information. While most other tools accept tab-delimited files, Qxpak.5 requires the values to be space-delimited, and also treats zeros in input files as missing values, so that the user must replace zero values with a specific small value (0.000001).

The utility of eQTL data

Understanding the relationship between polymorphisms & gene expression

Early eQTL studies have contributed greatly to our understanding of the molecular events controlling gene expression. The traditional output of the eQTL analysis is a list of transcripts and loci that are associated with each other at a certain statistical threshold. eQTL that are identified are generally classified into two broad categories [1]: *cis*-eQTL, also frequently referred to as 'local eQTL' are loci that correlate with the expression of transcripts located in the same genomic region. As a biological phenomenon, the occurrence of *cis*-eQTL presumably reflects allele-specific expression [39]. However, unless special breeding designs have been used, or the expression platform is RNA-sequencing, *cis*-eQTL are 'defined' by proximity – the occurrence of an association within a specified region. The regions of declared *cis*-eQTL association vary by data type, for example, <1 Mb for human association or <5 Mb for some linkage studies. We emphasize that the association does not need to be in the gene itself, and may reflect linkage or linkage disequilibrium with the gene, or direct association in a regulatory region, and may affect expression levels during (e.g., transcription factor-binding efficiency) or after (e.g., mRNA stability) transcription. *Trans*-eQTL, also called 'distant eQTL' are any of the complementary set of eQTL associations more distant than the requisite *cis*-threshold or on a different chromosome. These types of associations may be direct, for example, due to a polymorphism in a transcription factor or a binding site that regulates expression of the gene, or indirect, for example, due to regulation of a nuclear-receptor ligand, further regulating expression of downstream genes [1].

In general, *cis*-eQTL have stronger associations, with more consistent evidence, than do *trans*-eQTL, and are thought to reflect greater biological plausibility [40]. However, hybridization-based expression technologies are also susceptible to the possibility of false *cis*-eQTL due to polymorphic probes [41]; therefore, careful prior filtering of expression may be necessary.

Identification of eQTL hotspots

Trans-eQTL can occur individually at a single-genomic locus, or they can occur collectively as part of eQTL *trans*-bands, also called 'hotspots' [42]. Although it has been suggested that most *trans*-eQTL hotspots may be spurious and attributable to correlation structure in expression data [43], some *trans*-eQTL have been biologically validated [44], or replicated between populations [45]. The presence of such hotspots has been used in several studies to bolster the argument for the biological plausibility and physiological importance of the genotype–gene expression associations.

For example, using eQTL mapping, potential key regulators of gene expression in a particular tissue can be discovered. Bystrykh *et al.* have identified candidate genes that control mouse hematopoietic stem cell proliferation, data that support the notion that stem cell turnover and organismal aging are genetically linked [46]. Kempermann *et al.* combined data on variation in neurogenesis with eQTL data to extract a set of genes with expression patterns that are also highly variable and covary with neurogenesis phenotypes [47]. For example, 12 neurogenesis-associated transcripts had significant *cis*-acting QTL, and of these, six had plausible biological association with adult neurogenesis (Prom1, Ssbp2, Kcnq2, Ndufs2, Camk4 and Kcnj9). Gatti *et al.* identified a mouse liver-specific master-regulatory locus on chromosome 12 that correlates strongly with expression level of approximately 100 genes [42]. Serpina3 and Dicer1 at this locus may act as master regulators of liver-gene expression [42,45].

In addition, tissue-specific patterns of genetic control of gene expression can be elucidated through eQTL analysis and can potentially form the basis for comparative examination of gene expression between tissues. For example, the comparison of the transcriptome maps between mouse liver and brain [42], or brain and hematopoietic stem cells [48] demonstrated tissue-specific differences in the regulation of gene expression.

Network inference

Complex genetic interactions lie at the foundation of many diseases, and eQTL data have been used to reveal regulatory networks. The transcriptional network-inference approach of Bing and Hoeschele included: identification of eQTL confidence regions, fine mapping of identified eQTL, identification of regulatory candidate genes in each locus, correlation analysis of the expression of the genes in an eQTL with the gene(s) affected by the eQTL, drawing directional links between the regulator and regulated genes and statistical validation and refinement of the inferred network structure [49]. Emilsson *et al.* have demonstrated the utility of eQTL data for studies of the molecular underpinning of complex human diseases, such as obesity [50]. These authors showed that the transcriptional networks identified using human- and mouse-adipose data exhibit significant overlap, and that a core-network module found to be causally associated with obesity-related traits in both species is enriched for genes involved in the inflammatory and immune response.

Another set of widely used computational approaches for network inferences use Bayesian statistical methods. Although Bayesian networks may identify predictive relationships between genes from a combination of expression and eQTL data, this approach does not necessarily provide mechanistic information for predictive relationships [51]. Despite these inherent limitations, a recent study used Bayesian approach to prioritize targets for experimental intervention through the inference of causal-relationship networks from genotype, phenotype and eQTL data [52]. These networks can be used to make predictions of system-wide response to perturbations. Kidney gene-expression data from a genetically randomized population of MRL/MpJ×SM/J mouse intercross was used to predict a previously uncharacterized feedback loop in the local renin–angiotensin system.

Identification of disease subtypes

Gene-expression profiling provides invaluable molecular-level data that is used for the classification of various types of cancer into clinically relevant subtypes, predicting disease recurrence and response to treatments, and providing new insights into oncogenic pathways and the process of metastatic progression that can be used to devise novel therapies [53]. Schadt *et al.* have argued that combining gene expression, genotype and clinical data in a segregating population has the potential to objectively classify individuals according to disease subtypes and identify the drivers of the pathways or the causal factors underlying those disease subtypes [54]. The authors used eQTL data to demonstrate the complexity underlying complex diseases, such as obesity, and identify distinct disease subtypes. In their study, the authors showed that these subtypes were under the control of different loci in a mouse model, with QTL on chromosomes 2 and 19 each affecting only a subset of the mouse population with the disease [54].

Narrowing lists of candidate genes from GWAS

GWAS of complex diseases have yielded important and clinically relevant associations; however, the majority of variants identified to date confer only relatively small increments in risk, and, in general, have not led to greatly improved understanding of the biology underlying complex phenotypes [55]. In this regard, studies characterizing the genetics of gene expression can be leveraged to screen multiple genes within a GWAS-identified locus associated with a particular disease trait to provide a more objective assessment of genes in the region that may cause disease [11]. For example, if a SNP is found to be associated with a disease and it is also significantly associated with the expression of a gene in the vicinity of the SNP, this provides additional evidence that such a gene can be considered as a candidate-susceptibility gene.

A series of studies on the genetics of childhood asthma have used the approach of eQTL mapping to refine the results of GWAS. Specifically, *ORMDL3* was not only identified in a GWAS [56], but also confirmed to be a susceptibility gene for childhood and not adult-onset asthma in another population [57]. In addition, ORMDL3 expression in lymphoblastoid cell lines derived from children from families with asthma was found to be significantly associated with a nearby SNP that was also significantly associated with asthma [56]. Similarly, a study of liver gene expression in humans was used to provide further functional evidence for *SORT1*, *CELSR2* and *PSRC1* as candidate-susceptibility genes for cardiovascular disease and plasma low-density lipoprotein cholesterol levels from several large-scale GWAS [58].

The work of Nicolae *et al.* has taken advantage of the rapid accumulation of publicly available databases of complex trait-associated SNPs from GWAS, to generalize the concept that information on eQTLs has utility for understanding the genetic component of complex

traits [59]. The authors analyzed GWAS data to confirm that annotating SNPs with a score reflecting the strength of the evidence that the SNP is an eQTL can improve the ability to discover true associations and may further clarify the nature of the mechanism driving the associations. This raises the possibility that eQTL data may increase the proportion of heritability explained by identifiable genetic factors, and be used to gain a better understanding of the biology underlying complex traits.

Tools for interpreting eQTL results

For many genetic and clinical researchers, the main interest in eQTL analysis lies in its ability to suggest candidates for a follow-up investigation to association or other genetic analyses. For this purpose, the display and presentation of preanalyzed data can be very useful, even if it is derived from 'normal' (i.e., nondiseased) individuals, or from tissues of uncertain direct-disease relevance. To this end, here we describe several eQTL databases, browsers and other postprocessing tools that can provide assistance to those interested in incorporating eQTL data into their own research.

Webqtl is an online database [110] of linked datasets, including genotype and expression data, covering multiple species including mouse, macaque monkey, rat, drosophila, arabidopsis, plants and humans [60]. While this tool cannot be used to calculate eQTLs, it can be used to find and visualize eQTLs in different species, strains and tissues. It can perform single- and multiple-interval QTL mapping of up to 100 selected traits. Users can also upload their own trait data for populations included in the database. It can also calculate and display trait-correlation matrices and network graphs (also for up to 100 traits).

SeeQTL is an online database [111] of human eQTL associations, with a common display of false-discovery rate-based evidence for all of the data [61]. The database is built on the Gbrowse2 platform, and populated with eQTL results for several tissues and lymphoblastoid cell line results for several HapMap populations. In addition to results for individual datasets, seeQTL displays a consensus meta-analysis of eQTL evidence for several HapMap studies. Novel features include the use of so-called 'segment plots' that visually connect each transcript to *cis*-associated nearby SNPs, as well as Manhattan plots of association for both *cis*- and *trans*-associations.

SCAN is an online database [112] with expression, genotype and eQTL information for human data from multiple public sources including HapMap and the National Center for Biotechnology Information (NCBI) [62]. It supports queries by gene name, SNP name, copy number variations and their genomic locations, and can report variables (i.e., genes, SNPs and single nucleotide variations) that are strongly associated with the selected end point. SCAN also conveys HapMap allele-frequency information, and is among the few browsers that displays expression associations with copy number variation. In addition, SCAN places an emphasis on text queries rather than visual display.

The University of Chicago (IL, USA) eQTL Browser is an online Gbrowse database [113] with eQTL information from multiple sources, including HapMap lymphoblasts and several other tissues [8]. The browser includes data on liver and T cells not currently represented in other databases, although this browser and many other databases are expanding rapidly. The eQTL browser, with an emphasis on visual display and 'clickable' links, is the most similar to seeQTL among the databases described here.

Genotype–Tissue Expression (GTEx) eQTL browser is an online database [114] in support of the NIH GTEx roadmap project. The database is currently populated with data from seven non-GTEx human studies at the time of writing, including several datasets represented in the other browsers. As the GTEx project progresses, the database will be populated with new eQTL results for multiple human tissues. One unique feature of the current browser is an attempt to merge the eQTL functionality with searches by phenotypic traits that have been mapped via GWAS. This effort follows a common belief that the ultimate utility of eQTL analysis will be to strengthen the connections between disease phenotype and the underlying eQTL architecture.

Future perspective

eQTL methods have rapidly evolved in the last few years, following a rapid expansion of available datasets and attempts to functionally relate eQTLs to phenotypic outcome. The initial discoveries of the basic architecture of genetical genomics have given way to much more challenging questions involving SNP–transcriptional networks, tissue- and developmentally specific transcription, and paths of true causation. Many of the published human datasets have only used one tissue – often lymphocyte-derived – and comparisons across datasets have been made more challenging due to differences in expression platforms. In the next several years, new and larger datasets will become available.

The GTEx project [115] will enable true cross-tissue comparisons in humans, and 'omics profiling in rodent models will reach a new standard in studies such as the Collaborative Cross [63].

In addition, RNA-sequencing data will likely replace hybridized arrays, providing much richer information of sequence context, splice variation and allele-specific expression. Several recent studies have begun to explore of the power of these new approaches to uncover the full spectrum of regulatory variation (e.g., alternative transcription starts, alternative splicing and alternative polyadenylation) and its effect on gene expression [27,64]. The results suggest that the regulatory effects of genetic variation in a normal human population are far more complex than previously believed. Still, the data-analysis methods for this novel information are not yet entirely mature, or accompanied by user-friendly software packages.

The ever-increasing complexity of the systems biology-level molecular information will create opportunities for further development and refinement of the computational tools. eQTL analysis will thus need to add extra dimensions to the analysis. Of particular importance will be methods and software that efficiently summarize the tissue specificity of overall eQTL expression, splice variation and allele-specific expression. Thus, improved speed will become even more important, along with a greater attention to interpretability of these inherently complex phenomena. Transcriptional pathways and modules, already believed to be crucial for the interpretation of study results, may become the unit of functional study in a similar manner that individual genes/transcripts are studied now. The added dimensions in the data will produce additional challenges in the visualization of the results and effective-query interfaces will be needed for easy interpretation by end-users.

Financial & competing interests disclosure

This work was supported, in part, by grants from the NIH (R01 MH090936 and P42 ES005948) and US EPA (STAR RD83382501). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

Statistical methods for expression quantitative trait-loci mapping

- Most expression quantitative trait loci (eQTL) analyses to date have been performed by individual laboratories, using purpose-built software.
- Several eQTL analysis tools are available, covering a wide range of speed and platform portability.

Tools for interpreting eQTL results

- Several tools are available for the display and visualization of eQTL results in a genomic context.
- There is a great need for further improved tools to display data with additional dimensions, conveying tissue, species and disease-specific information.

References

Papers of special note have been highlighted as:
- ▪ of interest
- ▪▪ of considerable interest

1 Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat. Rev. Genet.* 7(11), 862–872 (2006).

▪▪ **Key review describing major concepts concerning expression quantitative trait loci (eQTL) and genetical genomics.**

2 Damerval C, Maurice A, Josse JM, de Vienne D. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. *Genetics* 137(1), 289–301 (2011).

3 Sandberg R, Yasuda R, Pankratz DG *et al.* Regional and strain-specific gene expression mapping in the adult mouse brain. *Proc. Natl Acad. Sci. USA* 97(20), 11038–11043 (2000).

4 Primig M, Williams RW, Winzeler EA *et al.* The core meiotic transcriptome in budding yeasts. *Nat. Genet.* 26(4), 415–423 (2011).

5 Karp CL, Grupe A, Schadt E *et al.* Identification of complement factor 5 as a susceptibility locus for experimental allergic asthma. *Nat. Immunol.* 1(3), 221–226 (2000).

6 Cheung VG, Spielman RS. The genetics of variation in gene expression. *Nat. Genet.* 32(Suppl.), 522–525 (2002).

7 Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 437(7063), 1365–1369 (2005).

8 Veyrieras JB, Kudaravalli S, Kim SY *et al.* High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* 4(10), e1000214 (2008).

9 Jansen RC, Nap JP. Genetical genomics: the added value from segregation. *Trends Genet.* 17(7), 388–391 (2001).

10 Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296(5568), 752–755 (2002).

11 Schadt EE, Zhang B, Zhu J. Advances in systems biology are enhancing our understanding of disease and moving us closer to novel disease treatments. *Genetica* 136(2), 259–269 (2009).

12 Li H, Deng H. Systems genetics, bioinformatics and eQTL mapping. *Genetica* 138, 915–924 (2010).

13 Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. *Nat. Rev. Genet.* 10(9), 595–604 (2009).

14 Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10(3), 184–194 (2009).

15 Liu C. Brain expression quantitative trait locus mapping informs genetic studies of psychiatric diseases. *Neurosci. Bull.* 27(2), 123–133 (2011).

16 Kendziorski C, Wang P. A review of statistical methods for expression quantitative trait loci mapping. *Mamm. Genome* 17(6), 509–517 (2006).

17 Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum. Mol. Genet.* 19(R2), R227–R240 (2010).

18 Tesson BM, Jansen RC. eQTL analysis in mice and rats. *Methods Mol. Biol.* 573, 285–309 (2009).

19 Franke L, Jansen RC. eQTL analysis in humans. *Methods Mol. Biol.* 573, 311–328 (2009).

20 Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* 11(9), 647–657 (2010).

21 Kendziorski CM, Chen M, Yuan M, Lan H, Attie AD. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* 62(1), 19–27 (2006).

22 Gatti DM, Shabalin AA, Lam TC, Wright FA, Rusyn I, Nobel AB. FastMap: fast eQTL mapping in homozygous populations. *Bioinformatics* 25(4), 482–489 (2009).

▪ **Fast eQTL analysis tool with a graphical user interface.**

23 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38(8), 904–909 (2006).

24 Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3(9), 1724–1735 (2007).

25 Listgarten J, Kadie C, Schadt EE, Heckerman D. Correction for hidden confounders in the genetic analysis of gene expression. *PLoS Genet.* 3(9), 1724–1735 (2007).

26 Stranger BE, Forrest MS, Dunning M *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315(5813), 848–853 (2007).

27 Pickrell JK, Marioni JC, Pai AA *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289), 768–772 (2010).

28 Zheng J, Li Y, Abecasis GR, Scheet P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet. Epidemiol.* 35(2), 102–110 (2011).

29 Wang P, Dawson JA, Keller MP *et al.* A model selection approach for expression quantitative trait loci (eQTL) mapping. *Genetics* 187(2), 611–621 (2011).

30 Lan H, Chen M, Flowers JB *et al.* Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet.* 2(1), e6 (2006).

31 Sun W, Yu T, Li KC. Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics* 23(17), 2290–2297 (2007).

32 Ayroles JF, Carbone MA, Stone EA *et al.* Systems genetics of complex traits in *Drosophila melanogaster*. *Nat. Genet.* 41(3), 299–307 (2009).

33 Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19(7), 889–890 (2003).

▪ **R-based package for QTL mapping in crosses and multiple QTL mapping.**

34 Pletcher MT, McClurg P, Batalov S *et al.* Use of a dense single nucleotide polymorphism map for *in silico* mapping in the mouse. *PLoS Biol.* 2(12), e393 (2004).

35 Clayton D, Leung HT. An R package for analysis of whole-genome association studies. *Hum. Hered.* 64(1), 45–51 (2007).

▪ **A tool for genome-wide association testing.**

36 Purcell S, Neale B, Todd-Brown K *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81(3), 559–575 (2007).

▪ **A widely used tool set for genome-wide association analysis.**

37 Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* 30(1), 97–101 (2002).

▪ **A tool for fast pedigree analysis.**

38 Perez-Enciso M, Mosztal I. Qxpak.5: old mixed model solutions for new genomics problems. *BMC Bioinformatics* 12, 202 (2011).

39 Sun W. A Statistical framework for eQTL mapping using RNA-seq data. *Biometrics* doi:10.1111/j.1541-0420.2011.01654.x (2011) (Epub ahead of print).

40 Doss S, Schadt EE, Drake TA, Lusis AJ. *Cis*-acting expression quantitative trait loci in mice. *Genome Res.* 15(5), 681–691 (2005).

41 Verdugo RA, Farber CR, Warden CH, Medrano JF. Serious limitations of the QTL/microarray approach for QTL gene discovery. *BMC Biol.* 8, 96 (2010).

42 Gatti D, Maki A, Chesler EJ *et al.* Genome-level analysis of genetic regulation of liver gene expression networks. *Hepatology* 46(2), 548–557 (2007).

43 Breitling R, Li Y, Tesson BM *et al.* Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 4(10), e1000232 (2008).

44 Wu C, Delano DL, Mitro N *et al.* Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genet.* 4(5), e1000070 (2008).

45 Gatti DM, Harrill AH, Wright FA, Threadgill DW, Rusyn I. Replication and narrowing of gene expression quantitative trait loci using inbred mice. *Mamm. Genome* 20(7), 437–446 (2009).

46 Bystrykh L, Weersing E, Dontje B *et al.* Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat. Genet.* 37(3), 225–232 (2005).

47 Kempermann G, Chesler EJ, Lu L, Williams RW, Gage FH. Natural variation and genetic covariance in adult hippocampal neurogenesis. *Proc. Natl Acad. Sci. USA* 103(3), 780–785 (2006).

48 Chesler EJ, Lu L, Shou S *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* 37(3), 233–242 (2005).

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

50 Emilsson V, Tholeifsson G, Zhang B *et al.* Genetics of gene expression and its effect on disease. *Nature* 452(7186), 423–428 (2008).

51 Bumgarner RE, Yeung KY. Methods for the inference of biological pathways and networks. *Methods Mol. Biol.* 541, 225–245 (2009).

52 Hageman RS, Leduc M, Korstanje R, Paigen B, Churchill GA. A Bayesian framework for inference of the genotype-phenotype map for segregating populations. *Genetics* 187(4), 1163–1170 (2011).

53 Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat. Rev. Cancer* 7(7), 545–553 (2007).

54 Schadt EE, Monks SA, Drake TA *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422(6929), 297–302 (2003).

▪▪ **First experimental study addressing the issue of genetics of gene expression.**

55 Manolio TA, Collins FS, Cox NJ *et al.* Finding the missing heritability of complex diseases. *Nature* 461(7265), 747–753 (2009).

56 Dixon AL, Liang L, Moffatt MF *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* 39(10), 1202–1207 (2007).

57 Moffatt MF, Gut IG, Demenais F *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* 363(13), 1211–1221 (2010).

58 Schadt EE, Molony C, Chudin E *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6(5), e107 (2008).

59 Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* 6(4), e1000888 (2010).

60 Wang J, Williams RW, Manly KF. WebQTL: web-based complex trait analysis. *Neuroinformatics* 1(4), 299–308 (2003).

61 Xia K, Shabalin AA, Huang S *et al.* seeQTL: a searchable database for human eQTLs. *Bioinformatics* doi:10.1093/bioinformatics/btr678 (2011) (Epub ahead of print).

62 Gamazon E, Zhang W, Konkashbaev A *et al.* SCAN: SNP and copy number annotation. *Bioinformatics* 26(2), 259–262 (2010).

63 Aylor DL, Valdar W, Foulds-Mathes W *et al.* Genetic analysis of complex traits in the emerging collaborative cross. *Genome Res.* 21(8), 1213–1222 (2011).

64 Kwan T, Benovoy D, Dias C *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* 40(2), 225–231 (2008).

Websites

101 A QTL mapping environment. www.rqtl.org

102 eMap. http://bios.unc.edu/~wsun/software.htm

103 SNPSTER tool. http://snpster.gnf.org

104 snpMatrix. www.bioconductor.org/packages/2.3/bioc/html/snpMatrix.html

105 Plink. http://pngu.mgh.harvard.edu/~purcell/plink

106 Center for Statistical Genetics. www.sph.umich.edu/csg/liang/RepeatedMeasures/Merlin.html

107 Qxpak.5. www.icrea.cat/Web/OtherSectionViewer.aspx?key=485

108 FastMap 2.0: fast association mapping in heterozygous populations. http://comptox.unc.edu/resources.html

109 Matrix eQTL: ultra fast eQTL analysis via large matrix operations. http://bios.unc.edu/research/genomic_software/Matrix_eQTL

110 WebQTL. http://webqtl.org

111 seeQTL: a searchable database for human eQTLs. www.bios.unc.edu/research/genomic_software/seeQTL

112 SCAN. SNP and CNV annotation database. www.scandb.org

113 Resources regarding gene regulation from the Pritchard laboratory. http://eqtl.uchicago.edu

114 Genotype–Tissue Expression eQTL Browser. www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi

115 Genotype–Tissue Expression project. http://commonfund.nih.gov/GTEx