Defining gene and QTL networks
Ritsert C Jansen1,2, Bruno M Tesson1, Jingyuan Fu1,2, Yajie Yang3 and Lauren M McIntyre3,4

Current technologies for high-throughput molecular profiling of large numbers of genetically different individuals offer great potential for elucidating the genotype-to-phenotype relationship. Variation in molecular and phenotypic traits can be correlated to DNA sequence variation using the methods of quantitative trait locus (QTL) mapping. In addition, the correlation structure in the molecular and phenotypic traits can be informative for inferring the underlying molecular networks. For this, new methods are emerging to distinguish among causality, reactivity, or independence of traits based upon logic involving underlying QTL. These methods are becoming increasingly popular in plant genetic studies as well as in studies on many other organisms.

Addresses
1 Groningen Bioinformatics Centre, University of Groningen, The Netherlands
2 Department of Genetics, University Medical Centre Groningen, University of Groningen, The Netherlands
3 Genetics Institute, University of Florida, USA
4 Department of Molecular Genetics and Microbiology, University of Florida, USA

Corresponding author: McIntyre, Lauren M (mcintyre@ufl.edu)


This review comes from a themed issue on Genome studies and molecular genetics
Edited by Masahiro Yano and Roberto Tuberosa

Available online 3rd February 2009
1369-5266/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
DOI 10.1016/j.pbi.2009.01.003

Introduction
Since the rediscovery of Gregor Mendel’s pioneering work on pea crosses, segregating populations have been used to explore the underlying genetic architecture. The motivation for such work ranges from basic understanding of how life is organized and has adapted to changing environments, to utilization of such knowledge for genetic improvement of crops. Quantitative traits were deconstructed into additive, dominant, and epistatic effects, without consideration for the underlying molecular components. Technological advances in the 1980s made comprehensive genotyping affordable and mapping the rough location of the underlying genetic contribution for a quantitative trait locus (QTL) became feasible [1,2]. To date, more than 1200 studies in plants have been published on mapping phenotypic QTL (phQTL). A new wave of technological advances makes it possible to profile segregating populations for thousands of gene expression phenotypes and map expression QTL (eQTL) [3]. New technology can be used for the parallel measurement of the abundance of 1000s of proteins and metabolites to map protein QTL (pQTL) and metabolite QTL (mQTL). Deep sequencing, chromatin, and methyl-DNA immunoprecipitation are just a few of the newest technologies that add to the impressive arsenal of tools available for the study of the genetic variation underlying quantitative phenotypes [4,5]. Mapping phenotypes for thousands of traits, ‘genetical genomics’ [6,7], is the first step in attempting to reconstruct gene networks. Methods for network reconstruction can be used within a particular level (intra-level analysis, i.e. transcript data only), to explain the relationship among traits [8] at that level. Alternatively, the focus can be on understanding relationships across levels (inter-level analysis, integrating transcript, protein, metabolite, and morphological phenotypic data). Prior knowledge from other experiments can also be incorporated to further develop the picture of the network. Figure 1 illustrates the challenges that can be encountered with real data.

Causal, reactive, or independent?
The examination of pairwise correlation between traits, or principal components summaries of these traits, can lead to the hypothesis of a functional relationship if that correlation is high [8,9,10,11–13]. Incorporating QTL information, allows the inference of a functional relationship if two traits share multiple QTL, something that it is unlikely to happen at random. Going beyond the detected QTL, the correlation between residuals among traits, after accounting for QTL effects, or correlations between traits conditional on other traits, is further evidence for a network connection. To infer directional effects, it is necessary to analyze the correlations among pairs of traits in detail. If trait T1 maps to a subset of the QTL of trait T2, then the common QTL can be taken as evidence for their network connection, while the distinct QTL can be used to infer the direction [6,8]. If traits T1 and T2 have common QTL, without QTL that are distinct, then the inference is complicated and further analysis is needed to discriminate pleiotropy from any of the possible orderings among traits (Box 1 highlights approaches from [8,9,10]). Although these problems are ‘modern’ the groundwork for such analyses are evident in the earliest days of quantitative analysis [14].
Intra-level analysis
In reference organisms, such as *Arabidopsis*, and in a growing list of plants, the location of the genes producing the transcript or protein studied is known. This added information provides a layer of interpretation for eQTL and pQTL. In *Arabidopsis*, eQTL and pQTL networks have been defined [15–21]; in barley, eucalyptus, and maize eQTL networks have been defined [22–27]. When...
For traits $T_1$ and $T_2$, $T_1 > T_2$ denotes that $T_1$ affects $T_2$; $T_1 < T_2$ denotes that $T_2$ affects $T_1$, and $T_1 - T_2$ denotes a correlation of unknown direction. If traits $T_1$ and $T_2$ have one common QTL, without QTL that are distinct, then the inference of causality is complicated and further analysis needed to discriminate pleiotropy from any of the possible orderings among traits. In this case there are at least three possible models: $QTL > T_1 > T_2$, $QTL > T_2 > T_1$, $QTL > T_1$, and $QTL > T_2$. If we write the simple regression models $T_2 = \alpha_2 + \beta_2QTL + e_2$ and $T_1 = \alpha_1 + \beta_1QTL + e_1$, and if $e_1$ and $e_2$ are uncorrelated, the QTL may be considered to have pleiotropic effects on the two traits, that is with no direct link between $T_1$ and $T_2$. Alternatively, if there is no evidence for pleiotropy, then the following models can be considered $T_2 = \alpha_3 + \beta_3T_1 + e_3$ and $T_1 = \alpha_4 + \beta_4T_2 + e_4$. The residuals from these models can be used to infer the correct model. If $QTL > T_1 > T_2$ is the true relation, then $e_3$ will not map to the QTL. In contrast, $e_4$ should have a residual signature of the QTL. Similarly, if $QTL > T_2 > T_1$ is the true relationship, then $e_4$ will not map to the QTL, and $e_3$ should have a residual signature of the QTL. In other cases, that is $e_3$ and $e_4$ together map or do not map to the QTL, the directionality is not clearly indicated [9]. In addition, there are other competitive models such as $QTL > T_1 > T_2$ and $QTL > T_2$; or a loop $QTL > T_1 > T_2 > T_1$ that prevent clear (conclusive) inferences about the true network directions [8, 10]. As an important caution note, the above conditional models are based on various assumptions, and violation of these assumptions may lead to an increase in error rates for inferences about network structure.

At the metabolite level, mQTL for traits connected in a network may show complex patterns of correlation. For example, the mQTL for the precursor and product of an enzymatic step with differential activity should have opposite signs — indicating that the sign of the mQTL effect also conveys valuable information [35–37]. The effect of an mQTL may be visible on the precursor, the corresponding product and downstream products (Figure 1). As the number of steps grows, the complexity of the network increases, and network reconstruction based purely on correlation coefficients is challenging. Epistatic interactions among enzymes may further complicate the effort to map and deconstruct their unique patterns, as in the cases where some allelic combinations can be found in offspring which will then produce metabolites not found in either parent. Although such epistasis may be a rare phenomenon of complex traits [38], it is potentially abundant in secondary metabolism [36, 37].

**Inter-level analysis**

Inter-level inferences have been made between eQTL and mQTL in Arabidopsis [39] and between mQTL and phenotypic traits in tomato [40], and between eQTL, pQTL, mQTL, and phenotypic traits in Arabidopsis [41]. A system-wide analysis can reveal the impact of DNA sequence variation across multiple levels, that is eQTL at the gene expression level, pQTL for protein abundance or activity traits, mQTL for metabolite abundances and/or pQTL for morphological traits (Figure 1). Some DNA sequence variation will induce strong effects to be detected as hotspots or master regulators of many molecular and phenotypic traits, while others induce effects that are more subtle or are buffered in the network to ensure robustness of the system [41]. Correlations among traits from different levels can be used to generate hypotheses about network connections in inter-level analyses. Principal components may be used to summarize a network on one level and then regress on traits on another level [42]. The complexity of the system is such that two adjacent levels (i.e. transcript and protein) may not be linearly related. For example, DNA sequence variation may not affect expression level (no eQTL) while it does affect protein abundance or activity (pQTL). The ‘higher’ level traits (pQTL) may also be a function of multiple underlying (perhaps interacting) subnetworks (see the disease trait in Figure 1). Added complexity may be observed when DNA sequence variation directly affects higher level traits that — through feedback loops — affect other traits at the same or lower levels [39]. These examples indicate that caution is warranted given the intrinsic complexity in real networks.

Correlation analyses will only reveal the linear relationships among levels. Interpreting the correlation structure ‘beyond’ the common and shared QTL, using methods such as those described in Box 1, may generate hypotheses about system-wide networks. However, extreme caution is advisable in these interpretations in intra-level analyses owing to the potential impact of correlated measurement error (leading to false positive connections), and in inter-level analyses owing to the seeming lack of correlation between levels (leading to false negative connections) [43].

---

**Box 1  Advanced causal reasoning**

> For traits $T_1$ and $T_2$, $T_1 > T_2$ denotes that $T_1$ affects $T_2$; $T_1 < T_2$ denotes that $T_2$ affects $T_1$, and $T_1 - T_2$ denotes a correlation of unknown direction. If traits $T_1$ and $T_2$ have one common QTL, without QTL that are distinct, then the inference of causality is complicated and further analysis needed to discriminate pleiotropy from any of the possible orderings among traits. In this case there are at least three possible models: $QTL > T_1 > T_2$, $QTL > T_2 > T_1$, $QTL > T_1$, and $QTL > T_2$. If we write the simple regression models $T_2 = \alpha_2 + \beta_2QTL + e_2$ and $T_1 = \alpha_1 + \beta_1QTL + e_1$, and if $e_1$ and $e_2$ are uncorrelated, the QTL may be considered to have pleiotropic effects on the two traits, that is with no direct link between $T_1$ and $T_2$. Alternatively, if there is no evidence for pleiotropy, then the following models can be considered $T_2 = \alpha_3 + \beta_3T_1 + e_3$ and $T_1 = \alpha_4 + \beta_4T_2 + e_4$. The residuals from these models can be used to infer the correct model. If $QTL > T_1 > T_2$ is the true relation, then $e_3$ will not map to the QTL. In contrast, $e_4$ should have a residual signature of the QTL. Similarly, if $QTL > T_2 > T_1$ is the true relationship, then $e_4$ will not map to the QTL, and $e_3$ should have a residual signature of the QTL. In other cases, that is $e_3$ and $e_4$ together map or do not map to the QTL, the directionality is not clearly indicated [9]. In addition, there are other competitive models such as $QTL > T_1 > T_2$ and $QTL > T_2$; or a loop $QTL > T_1 > T_2 > T_1$ that prevent clear (conclusive) inferences about the true network directions [8]. As an important caution note, the above conditional models are based on various assumptions, and violation of these assumptions may lead to an increase in error rates for inferences about network structure.
Using a priori knowledge
Structural and functional data (gene sequence, gene localization, transcription factor binding sites (TFBS), Gene Ontology (GO), metabolic pathway, and protein–protein interaction (PPI)) as well as independent experimental data gleaned from secondary sources (i.e., Gene Expression Omnibus (GEO)) can be used post hoc to verify the defined gene and QTL networks. For example, if a disease maps to multiple QTL, then the candidate genes in each of the QTL can be analyzed and prioritized using known functional interactions [44]. As another example, particular eQTL trans bands may be identified as significantly enriched for a functional GO category [45] or as more likely to represent binding sites for transcription factors [46]. Prior knowledge can also be integrated in analysis. For example, a set of pathway-related genes may not show significant eQTL in gene-by-gene tests, while the set of genes can show such significance in a group-wise test [18].

Future directions
Genetic variation at multiple loci in combination with environmental factors can induce molecular or phenotypic variation. Variation may manifest itself as linear patterns among traits at different levels that can be deconstructed. Correlations can be attributed to detectable QTL and a logical framework based on common and distinct QTL can be used to infer network causality, reactivity, or independence. Unexplained variation can be used to infer direction between traits that share a common QTL and have no distinct QTL. Unexplained variation originates from other minor or modifier QTL, epigenetic factors, and biological, environmental and unfortunately, technical factors. Correlation structures present in the molecular data may reflect technical artifacts, in which case the models used to infer causality are potentially invalid and the inference is potentially erroneous. Additional studies are needed to understand and quantify the level of sensitivity of these network reconstruction methods to technical errors. Further research is also needed to develop and evaluate experimental designs other than the current biparental line crosses: for example, multiple line crosses [47–49], advanced intercrosses [47,48], or populations of natural ecotypes [49–54]. Prior knowledge and complementary experiments such as deletion mapping followed by independent gene expression studies between parental lines may validate or disprove implicated network connections [55].

The trend of genetic studies to go deeper (more levels) and broader (larger scale and more factors, including environmental ones) brings challenges to develop methodology that can reconstruct networks more efficiently and more accurately. Despite the obvious limitations of gene and QTL network reconstruction methods, these and other future developments in biotechnology and genomics hold for sure great promise for the field of quantitative genetics.

Acknowledgment
Netherlands Organization for Scientific Research (NWO) and NSF (0821954).

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:
- of special interest
- of outstanding interest

This paper provides the statistical framework for causal inference on the basis of common and distinct QTL. The relationship between the graphical network and the statistical model is elucidated.
This paper describes causal reasoning in QTL studies.
This paper clearly enumerates all the possible causal relationships between a QTL and two phenotypes.


This is the first paper to show that eQTL data can be used to prioritize genes for phQTL.