Genetical genomics approaches for systems genetics
Tesson, Bruno Marie

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2011

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 13-11-2018
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 quantitative trait loci (QTL). These methods are becoming increasingly popular in plant genetic studies as well as in studies on many other organisms.

Originally published as:
Defining gene and QTL networks.
Jansen RC, Tesson BM, Fu J, Yang Y, McIntyre LM.
6.1 Introduction

Since the rediscovery of Georg Mendel’s pioneering work on pea crosses, segregating populations have been used to explore the underlying genetic architecture. 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 (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.

6.2 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 [8-13]. 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 $T_1$ maps to a subset of the QTL of trait $T_2$, 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 $T_1$ and $T_2$ 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 order-
ings among traits (Box 1 highlights approaches from [8, 10, 12]). Although these problems are ‘modern’ the groundwork for such analyses are evident in the earliest days of quantitative analysis [14].

Figure 1 - System-wide QTL analysis for aliphatic glucosinolates.

Data in this example are taken from [41] and demonstrate important network features that reconstruction methods should take into account. The colors in the QTL likelihood graphs (upper panel) and in the pathway (lower panel) correspond. The sign of the QTL effect is shown by plotting the QTL likelihood above the x-axis if the Cvi allele has higher average trait value or below the x-axis if the Ler allele has higher average trait value. The vertical dashed lines indicate the chromosome borders and the physical gene positions are shown as stars and grey vertical bars. (continued)
Figure 1 (continued) - Glucosinolates are important secondary metabolites in plants and are well-known for their toxic effect on insects. The aliphatic glucosinolate biosynthesis pathway is summarized in the lower panel. The MAM genes are involved in side chain elongation process: MAM1 mainly synthesizes C3-glucosinolates (in purple box) and MAM2 mainly synthesizes C4-glucosinolates (in green box). The parents Cvi and Ler carry different MAM genes. Cvi contains two MAM1 genes and Ler contains a functional MAM2 in addition to a truncated, non-functional MAM1 gene. The AOP genes are involved in the side-chain modification process: AOP2 and AOP3 generate different types of glucosinolates as described. Both AOP2 and AOP3 are present in Ler and Cvi. But AOP2 is only expressed in Cvi and AOP3 is only expressed in Ler. The top panel shows the QTL profiles at different levels (transcripts, proteins, metabolites, disease trait) from a Cvi x Ler recombinant inbred line population. To clearly demonstrate the QTL effects at different levels and along the pathway, the components are divided into two parts: the left part relates to the MAM1 gene and the metabolites produced by MAM synthesis (MAM2 is not measured); the right part refers to the AOP genes and the metabolites produced by the AOP synthesis. Cis-eQTL are detected for AOP2 and AOP3, cis-pQTL for MAM1, AOP2 and AOP3, and mQTL for the various aliphatic glucosinolates. These QTL have the same or opposite sign of QTL effect. To demonstrate whether QTL at molecular levels can propagate to phenotypic level, molecular QTL (eQTL, pQTL and mQTL) are also compared to the QTL of insect susceptibility (phQTL). The disease trait maps to ERECTA, a gene well known for its widespread pleiotropic effect, to AOP, and to third gene, but it does not map to MAM. This can be explained by the fact that the total glucosinolate content maps to AOP only.

6.3 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 the eQTL or pQTL co-localize with the gene, this effect may be due to cis regulatory effects (Fig 1). The caveat is that the detection of cis effects may be an artifact of differential probe hybridization due to sequence polymorphism [28, 29]. If gene expression at a particular locus is regulated by that locus (cis effect), and the abundance of the transcript in turn regulates additional loci (trans effect) then these expression traits should all map to the same locus. If the number of trans loci regulated by a single locus is large, as would be expected from a master regulator, or switch, a trans band will be observed at this location. All genes in the QTL are candidate regulators; their partial correlations with the regulated genes can be used to prioritize them [30, 31]. Importantly, genes without cis-eQTL can be regulators, manifesting only at the protein level [32]. If the number of transcript or protein traits mapping to a single location exceeds the number expected by chance, then a hotspot has been identified [33, 34]. The hotspot can be inferred to represent a possible master regulator or switch. However, as a cautionary note hotspots can be an artifact of improper permutation [34].
Box 1: Advanced causal reasoning

For two traits $T_1$ and $T_2$, denote $T_1 > T_2$ if $T_1$ affects $T_2$; denote $T_1 < T_2$ if $T_2$ affects $T_1$, and $T_1 - T_2$ if there is a correlation but the direction is unknown. 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 $T_2 = \alpha_2 + \beta_2 \text{QTL} + \epsilon_2$ and $T_1 = \alpha_1 + \beta_1 \text{QTL} + \epsilon_1$ and if $\epsilon_1$ and $\epsilon_2$ are uncorrelated, the QTL may be considered to have pleiotropic effects on the two traits, i.e. 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 $T_2 = \alpha_3 + \beta_3 T_1 + \epsilon_3$ and $T_1 = \alpha_4 + \beta_4 T_2 + \epsilon_4$. The residuals from these models can be used to infer the correct model. If QTL$>T_1>T_2$ is the true relation, then $\epsilon_3$ will not map to the QTL. In contrast, $\epsilon_4$ should have a residual signature of the QTL. In cases where $\epsilon_3$ and $\epsilon_4$ both map to the QTL, no direction can be given and $T_1 - T_2$, while if neither of the two models maps to the QTL, the results are inconclusive [12]. In addition, there are other competitive models such as QTL$>T_1>T_2$ and QTL$>T_2$; or a loop QTL$>T_1>T_2$ and $T_2>T_1$ that prevent clear (conclusive) inferences about the true network directions [8, 10]. As an important cautionary 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 [Fig 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].

### 6.4 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, i.e. eQTL at the gene expression level, pQTL for protein abundance or activity traits, mQTL for metabolite abundances and/or phQTL for morphological traits (Fig 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 the regressed 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 (phQTL) may also be a function of multiple underlying (perhaps interacting) sub networks (see the disease trait in Fig 1). Added complexity may be observed when DNA sequence variation directly affect 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 due to the potential impact of correlated measurement error (leading to false positive connections), and in inter level analyses due to the seeming lack of correlation of between levels (leading to false negative connections) [43].

### 6.5 Using a priori knowledge

Structural and functional data (gene sequence, gene localization, transcription factor
binding sites (TFBS), Gene Ontology (GO), metabolic pathway, 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].

### 6.6 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 structure 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 [46-48], 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 factors), 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 genetics hold for sure great promise for the field of quantitative genetics.
6.7 References


and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. BMC genomics 2007, 8:22.


41. Fu J, Keurentjes JJB, Bouwmeester H, America T, Verstappen FWA, Ward


52. Rosenberg NA, Nordborg M: A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations.
