Systems Genetics for Evolutionary Studies

Prins, Pjotr; Smant, Geert; Arends, Danny; Mulligan, Megan K.; Williams, Rob W.; Jansen, Ritsert C.

Published in:
Evolutionary Genomics

DOI:
10.1007/978-1-4939-9074-0_21

Chapter 21

Systems Genetics for Evolutionary Studies

Pjotr Prins, Geert Smant, Danny Arends, Megan K. Mulligan, Rob W. Williams, and Ritsert C. Jansen

Abstract

Systems genetics combines high-throughput genomic data with genetic analysis. In this chapter, we review and discuss application of systems genetics in the context of evolutionary studies, in which high-throughput molecular technologies are being combined with quantitative trait locus (QTL) analysis in segregating populations.

The recent explosion of high-throughput data—measuring thousands of RNAs, proteins, and metabolites, using deep sequencing, mass spectrometry, chromatin, methyl-DNA immunoprecipitation, etc.—allows the dissection of causes of genetic variation underlying quantitative phenotypes of all types. To deal with the sheer amount of data, powerful statistical tools are needed to analyze multidimensional relationships and to extract valuable information and new modes and mechanisms of changes both within and between species. In the context of evolutionary computational biology, a well-designed experiment and the right population can help dissect complex traits likely to be under selection using proven statistical methods for associating phenotypic variation with chromosomal locations.

Recent evolutionary expression QTL (*e*QTL) studies focus on gene expression adaptations, mapping the gene expression landscape, and, tentatively, define networks of transcripts and proteins that are jointly modulated sets of *e*QTL networks. Here, we discuss the possibility of introducing an evolutionary "prior" in the form of gene families displaying evidence of positive selection, and using that prior in the context of an *e*QTL experiment for elucidating host-pathogen protein-protein interactions.

Here we review one exemplar evolutionairy eQTL experiment and discuss experimental design, choice of platforms, analysis methods, scope, and interpretation of results. In brief we highlight how *e*QTL are defined; how they are used to assemble interacting and causally connected networks of RNAs, proteins, and metabolites; and how some QTLs can be efficiently converted to reasonably well-defined sequence variants.

Key words Systems genetics, Genetical genomics, QTL, eQTL, xQTL, R-genes, Evolution, R/qtl, LMM, GEMMA, NGS, Genomics, Metabolomics, Network inference, GeneNetwork

1 Introduction

Genetics concerns the study of heritably quantitative or complex traits. Many agricultural traits of interest, such as milk production in cattle and response to fertilizer in crops and most human, animal, and plant diseases, are complex traits. Associating, or linking,

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_21,

complex traits with certain positions on the genome is achieved through the mapping of the so-called quantitative trait loci (QTL).

Mapping QTL in experimental populations is possible when linkage and/or association information is available. When we have a population of individuals with known genotypes, it may be possible to link a phenotype with a certain genotype. To genotype individuals, first marker maps are created. A marker is a known genomic location, where the genotype of an individual can be determined. In the early days, the genotype was determined by visible chromosome features, later with restriction fragment length polymorphism (RFLP) and amplified fragment length polymorphism (AFLP, *see* also [1–3]), and, increasingly, with SNP/haplotype data [4]. When all individuals with genotype A at a marker location somewhere on the genome are susceptible to a disease and all other individuals with genotype B are not, there is linkage/association or a QTL. If it is clear cut, i.e., single QTL explains all phenotype variance, it is likely to be a single gene effect. Often it is not clear cut, and we need statistics to determine the strength of association between phenotype and genotype.

It is also possible to use linkage disequilibrium (LD) to map QTL in outbred and natural populations. LD occurs when certain stretches of the genome (haplotypes) show nonrandom behavior based on allele frequencies and recombination. Associating haplotype frequencies with phenotypes potentially renders QTL. Kim et al. describe the genome-wide pattern of LD in a sample of 19 *Arabidopsis thaliana* accessions using SNP microarrays [5]. LD is tested, for example, by Dixon et al., to globally map the effect of polymorphism on gene expression in 400 children from families recruited through a proband with asthma [6].

The use of terms "association" and "linkage" can be confusing, even in literature. In this text we use **association** with haplotypes in natural populations of unrelated individuals and **linkage** with markers in families and groups of families, often termed experimental populations. Note some genetic studies are hybrids of both methods, such as Dixon et al. [6], and individuals are related, i.e., some within-family linkage information is available for 400 children from 206 families which should be accounted for in the analysis.

Statistical power can be increased by using experimental crosses instead of natural populations. For example, each individual line in a set of recombinant inbred lines (RILs) is homozygous across the genome, doubling the genetic variance, simplifying genetic models, and increasing statistical power. For model organisms, such as *A. thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*, genotyped and even fully sequenced experimental crosses are available; i.e., for these species it is not necessary to generate a new cross, and for these crosses comprehensive SNP and sequence data may be available. One of the features of inbred model organisms is that they are "immortal" which means that

experiments conducted more than 10, even 30, years ago can still be compared with those today. Databases, such as GeneNetwork [7, 8], contain thousands of studies conducted on the same individual mouse strains.

Systems genetics combines genetics with high-throughput molecular technologies. Combining gene expression, as measured by microarray probes or RNA sequencing, with linkage leads to gene expression QTL (*e*QTL). Such *e*QTL studies elucidate how genotypic variation underlies, for example, morphological phenotypes, by using gene expression levels as intermediate molecular phenotypes. In other words, the expression level, as measured by a microarray probe or probe set, is treated as a phenotype, i.e., a gene expression trait. This phenotype is associated with the genome in the form of one or more *e*QTL. With microarrays, the genomic location of the probe is usually known. Therefore, expression phenotype and probe connect two types of genomic information: *e*QTL location(s) and gene location. It is usually assumed that *e*QTL loci represent *cis*- or *trans*-transcription regulators of the target gene [9]. If the *e*QTL is located close to the gene on the genome, the *e*QTL may point to a *cis*-regulator. If the *e*QTL is located far from the gene on the genome, the *e*QTL may point to a *trans*-regulator of a single gene or even *e*QTL *trans*-bands that regulate multiple genes (*see* Fig. 1a and [10, 11]).

In a similar fashion, proteins and metabolites can be measured to map protein QTL (*p*QTL) and metabolite QTL (*m*QTL). A remarkable study published in 1994 used two-dimensional protein electrophoresis and a restriction fragment length polymorphism map (RFLP) [12]. Deep sequencing, chromatin, and methyl-DNA immunoprecipitation are just a few of the latest technologies that add to the arsenal of tools available for the study of the genetic variation underlying quantitative phenotypes. Together, *e*QTL, *m*QTL, and *p*QTL are referred to as *x*QTL. Different *x*QTL appear to confirm each other, for example, with the *A. thaliana* glucosinolate pathway where *e*QTL, *m*QTL, and *p*QTL were mapped together and used to infer the underlying pathways [13]. Such causal inference can lead to dissecting pathways and gene networks which is an active field of research, e.g., [14–16] (*see* also Fig. 1).

1.1 Evolutionary xQTL Studies

From the perspective of evolutionary biology, systems genetics has been applied to elucidate evolutionary adaptations of transcript regulation. For example, Fraser et al. introduced a test for lineage-specific selection on gene expression and analyzed the directionality of microarray *e*QTL for 112 haploid segregants of a genetic cross between two strains of the budding yeast *Saccharomyces cerevisiae*, reanalyzing the two-color cDNA microarray data of Brem and Kruglyak [17]. They found that hundreds of gene expression levels have been subjected to lineage-specific selection. Comparing these findings with independent population genetic

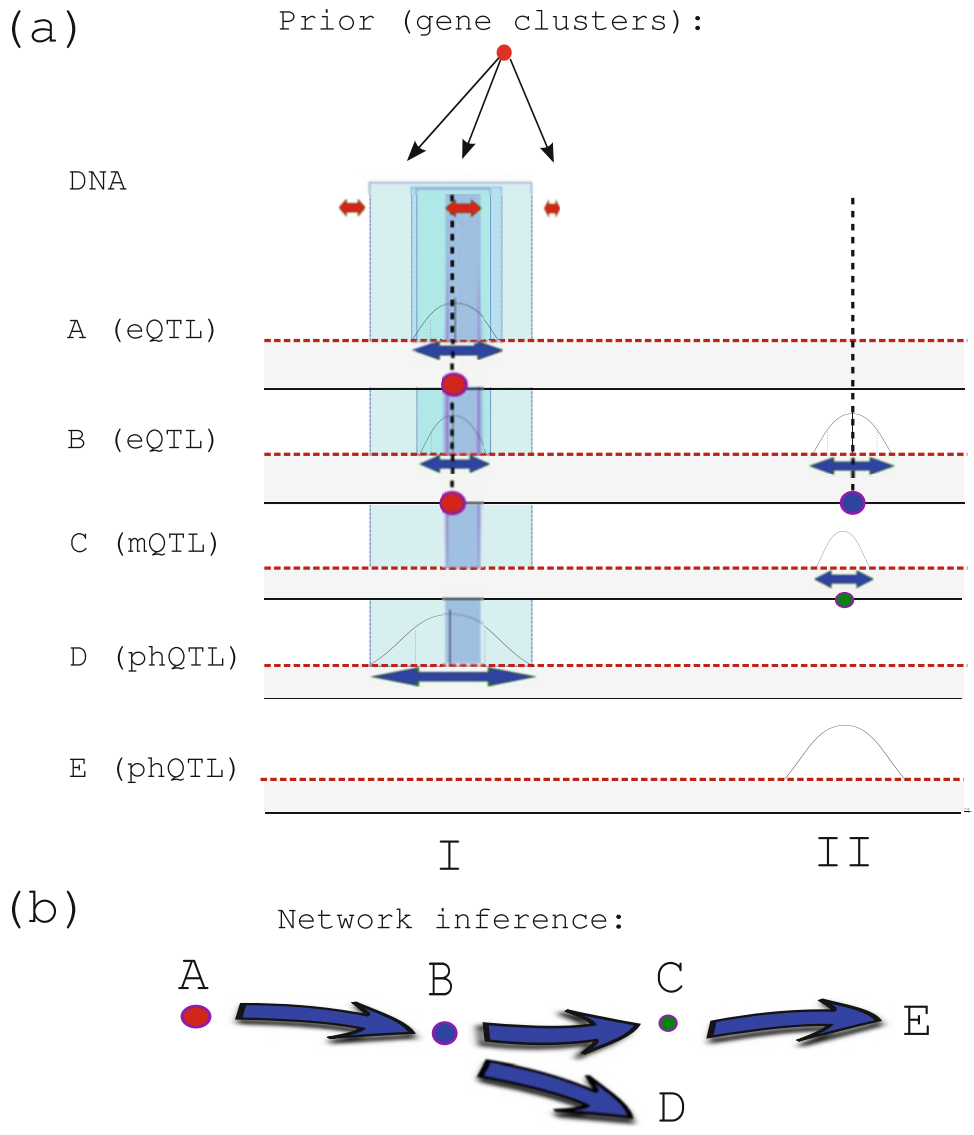

Fig. 1 In this hypothetical and schematic example related to mapped locations on a chromosome, prior information is combined with multiple phenotype-genotype QTL mappings to zoom in on genomic areas and to reason about causal relations between different layers of information. (**a**) The prior (red area on the chromosome) points out that certain sections are of interest; these sections consist of related genes with high homology showing evidence of positive selection, as discussed in the main text. The blue double arrow points out the confidence interval for each QTL, above the significance threshold (red dotted line). The accumulated evidence (light-blue areas) leads to a narrowed down section on the genome, where in this case the prior information is the most specific. In addition, expression phenotypes A and B point to exact gene locations (dotted line, based on exact probe information). (**b**) To infer causal relationships, network inference is possible. On the left (vertical I), traits A, B, and D map to one hot spot, where A may be a regulator of B because one QTL is shared. B causes metabolite phenotype C, again a shared QTL. Phenotype D matches A and B, and phenotype E matches A, B, and C. These causal relationships are drawn by arrows. The figure suggests that, even if individual QTL are not very informative, the accumulated evidence starts to paint a picture

evidence of selective sweeps suggests that this lineage-specific selection has resulted in recent sweeps at over a hundred genes, most of which led to increased transcript levels. Fraser et al. suggest that adaptive evolution of gene expression is common in yeast, that regulatory adaptation can occur at the level of entire pathways, and that similar genome-wide scans may be possible in other species, including human [18].

In another *S. cerevisiae* study, Zou et al., by reanalyzing the same two-color cDNA microarray data, uncovered genetic regulatory network divergence between duplicate genes. They found evidence that the regulation of the ancestral gene diverged due to gene duplication [19].

Li et al. studied plasticity of gene expression in *C. elegans*, using a set of 80 RILs generated from a cross of N2 (Bristol) and CB4856 (Hawaii), representing two genetic and ecological extremes of *C. elegans*. While the overall level of polymorphism among wild isolates of *C. elegans* is relatively low, the genetic distance between N2 and CB4856 is high, representing millions of years of genetic drift. Differential expression induced in a RIL population by temperatures of 16 °C and 24 °C has a strong genetic component. With a group of transgenes, there was prominent evidence for a common master regulator: an *e*QTL trans-band of 66 coregulated genes appeared at 24 °C. The results suggest widespread genetic variation of differential expression responses to environmental impacts and demonstrate the potential of systems genetics for mapping the molecular determinants of phenotypic plasticity [11], leading to a more generalized systems genetics, where value is added from environmental perturbation [20].

Hager et al. determined that genetic architecture supports mosaic brain evolution and independent brain-body size regulation by a quantitative genetic approach involving over 10,000 BXD mouse RILs. The BXD family consists of over 100 lines derived from parental strains that differ at five million single nucleotide polymorphisms (SNPs), indels, transposons, and copy-number variants. This model system harbors naturally occurring genetic variation at a level approximating that of human populations. The study utilizes a high-density linkage analysis to map loci modulating phenotypic variation in overall brain size, body size, and the size of seven major brain parts: neocortex, cerebellum, striatum, olfactory bulb, hippocampus, lateral geniculate nucleus, and basolateral complex of the amygdala. Under the mosaic evolutionary hypothesis, the size of different systems evolves independently due to differential selective pressures associated with different tasks. They identified independent loci for size variation in seven key parts of the brain and observe that brain parts show low or no phenotypic correlation, as is predicted by a mosaic scenario. They also demonstrate that variation in brain size is independently regulated from body size [21].

Kliebenstein et al. detected significant gene network variation in 148 RILs originating from a cross between two *A. thaliana* accessions, Bay-0 and Shahdara. They were able to identify *e*QTL controlling network responses for 18 out of 20 *a priori* defined gene networks, representing 239 genes [22].

According to Gilad, *e*QTL studies show that (1) variation in gene expression levels is both widespread and highly heritable; (2) gene expression levels are highly amenable to genetic mapping; and (3) most strong *e*QTL are found near the target gene, suggesting that variation in cis-regulatory elements underlies much of the observed variation in gene expression levels [23]. Meanwhile, Alberts et al. suggest that sequence polymorphisms influencing the binding of microarray probes may cause many false cis *e*QTL, which should be accounted for [24].

1.2 Adding a Prior

QTL mapping links complex traits with one or more locations on the genome (*see* Fig. 1). Such a location is a wide measure because a QTL is a statistical estimate and rarely a precise indicator. On the genome, a single QTL may represent tens, hundreds, and even thousands of real genes. Combining the QTL with high-throughput technologies, such as microarrays, can add information. To zoom in on the genes underlying QTL, information from other sources has to be utilized. Such *a priori* knowledge (prior) could consist of results from traditional linkage studies or association studies of, for example, human disease. That way one can assign a specific regulatory role to polymorphic sites in a genomic region known to be associated with disease [23]. Other useful priors can be derived from existing information on gene ontology terms, metabolic pathways, and protein-protein interactions, which can be used to identify genes and pathways [25], provided these databases are sufficiently informative.

Zou et al., for example, used gene ontology as a **prior** and concluded that *trans*-acting *e*QTL divergence between duplicate pairs of genes is related to a fitness defect under treatment conditions, but not with fitness under normal condition [19].

Chen et al. identified strong candidate genes for resistance to leaf rust in barley and on the general pathogen response pathway using a custom barley microarray on 144 doubled haploid lines of the St/Mx population [26]. Fifteen thousand six hundred and eighty-five *e*QTL were mapped from 9557 genes. Correlation analysis identified 128 genes that were correlated with resistance, of which 89 had *e*QTL colocating with the phenotypic QTL (phQTL) or classic QTL. Transcript abundance in the parents and conservation of synteny with rice prioritized six genes as candidates for Rphq11, the phQTL of largest effect [26].

In this chapter we discuss the steps needed to design an *x*QTL experiment to make use of systems genetics in evolutionary studies more concrete. As the prior we add information on plant host genes showing evidence of positive selection.

2 Designing an Evolutionary *x*QTL Experiment

An experimental design based on systems genetics can highlight sections of the genome showing correlation with an evolutionary trait. One such evolutionary trait of interest is plant resistance against pathogens. Plants have developed mechanisms to defend themselves against pests. When a pathogen, such as potato blight *Phytophthora infestans*, or a nematode, such as *Meloidogyne hapla*, infects a plant, it uses a battery of so-called effectors to help invade the plant. Some of these effector molecules act to dissolve cellulose [27]. Intriguingly, other molecules are involved in actively reprogramming plant cells. Such plant-pathogen effectors have been shown to mimic plant transcription factors [28] and switch on genes that help the pathogen [29]. A susceptible plant allows the pathogen to suppress defense mechanisms and to change cell configuration. For example, the nematodes *M. hapla* and *Globodera rostochiensis* transform plant cells, so they become elaborate feeding structures. The genetics of this plant-pathogen interaction is potentially even relevant for human medicine, as an increased understanding of host-pathogen relationships may help understand the workings of the innate immune system and nematode immunomodulation [30, 31]. The innate immune system, through plant resistance genes (R-genes, *see* Box 1), influences susceptibility to infections in all multicellular organisms and is a much older evolutionary mechanism than the advanced adaptive immune system found in higher organisms.

Box 1: Adaptive evolution in R-genes
Plant resistance genes (R-genes) are a homologous family of genes, formed by gene duplication events and hypothesized to be involved in an evolutionary arms race with pathogen effectors. R-genes are involved in recognizing specific pathogens with cognate avirulence genes and initiating defense signaling that results in disease resistance [32]. R-genes are characterized by a molecular gene-for-gene interaction [33] in which a specific allele of a disease resistance gene recognizes an avirulence protein or pathogen allele. This specificity is often encoded, at least in part, in a relatively fast-evolving leucine-rich repeat (LRR) region [34], which consists of a varying number of LRR modules. Activation of at least some

(continued)

Box 1: (continued)
of these proteins is regulated in trans, as has been shown for RPM1 and RPS2 [35].

A single *A. thaliana* plant has about 150 R-genes, representing a subset of R-genes in the overall population. The protein products of R-genes are involved in molecular interactions. They generally have a recognition site which can dock against, i.e., recognize, one or more specific molecule(s). The proteins encoded by the largest class of R-genes carry a nucleotide-binding site LRR domain (NB-LRR, also referred to as NB-ARC-LRR and NBS-LRR). NB-LRR R-genes can be further subdivided based on their N-terminal structural features into TIR-NB-LRR, which have homology to the *Drosophila* Toll and mammalian interleukin-1 receptors and CC-NB-LRR, which contain a putative coiled-coil motif [36]. The LRR domain appears to mediate specificity in pathogen recognition, while the N-terminal TIR, or coiled-coil motif, is likely to play a role in downstream signaling [34]. When a molecule is docked, the R-protein is able to activate pathways in the cell, resulting in, for example, a hypersensitive response causing apoptosis and preventing spread of infection.

Meanwhile, one single R-protein only recognizes one type of invading molecule. Therefore, through its R-genes, one individual plant only recognizes a limited number of strains of invading pathogens, as the individual pathogens have variation in effectors too. When a pathogen evolves to use nonrecognized effectors, the plant becomes susceptible. The success of plant defense is determined by both evolution and the variation of specificity in a population. Unlike the evolved mammal immune system, which can change in a living organism and learn about invasions "on the fly" [37], plant R-genes depend on the variation inside a gene pool to provide the resistance against a pathogen; *see*, for example, Holub et al. [38]. Even so, many genes involved in pathogen recognition undergo rapid adaptive evolution [39], and studies have found that *A. thaliana* R-genes show evidence of positive selection, e.g., [40–42].

In this chapter we do not limit ourselves to (known) R-genes. Plants have evolved a complex array of chemical and enzymatic defenses, both constitutive and inducible, that are not involved in pathogen detection but whose effectiveness influences pathogenesis and disease resistance. The genes underlying these defenses comprise a substantial portion of the host genome. Based on

genomic sequencing, it is estimated that some 14% of the 21,000 genes in *A. thaliana* are related to defense against pathogens [43]. Most of these genes are not involved in direct pathogen detection, but their protein products interact directly with pathogen proteins or protein products at the molecular level. Among these proteins, for example, are chitinases and endoglucanases that attack and degrade the cell walls of pathogens and which pathogens counterattack with inhibitors. Such systems of antagonistically interacting proteins provide the opportunity for molecular coevolution of individual systems of attack and resistance [39].

In this chapter we design an experiment to look for all gene families showing evidence of positive selection. This evidence of positive selection is the prior for *e*QTL analysis: combining known genomic locations of gene families with *e*QTL locations derived from gene expression variation in a host-pathogen interaction experiment, which hopefully results in zooming in on gene families involved in plant resistance. The prior adds statistical power in locating putative gene families involved in host-pathogen coevolution (Fig. 1). Note that, in this chapter, the term "interaction" is used in two ways. The first is for QTL interaction, where two QTL on the genome interact **statistically**. The second is for host-pathogen gene-for-gene interaction, where gene products from different species interact **physically**.

2.1 Create a Prior with PAML

To create the prior, we use Ziheng Yang's codeml implementation of phylogenetic analysis by maximum likelihood (PAML) [44]. PAML can find amino acid sites which show evidence of positive selection using d_N/d_S ratios, which is the ratio of non-synonymous over synonymous substitution (ω, *see* [44]). The calculation of maximum likelihood for multiple evolutionary models is computationally expensive, and executing PAML over an alignment of a hundred sequences may take hours, sometimes days, on a PC. The software for generating the prior is prepackaged and makes up the workflow in Chap. 25, which includes BLAST [45], Clustal Omega [46], pal2nal [47], PAML [44], and BioRuby [48].

It is possible to find nonoverlapping large gene families by using BLASTCLUST, a tool that is part of the BLAST tool set [45]. After fetching the *A. thaliana* cDNA sequences from the Arabidopsis Information Resource (TAIR) [49], convert the sequences to a protein BLAST database format. Based on a homology criterion, the identity score and genes are clustered into putative gene families by running BLASTCLUST with 70% amino acid sequence identity. Note that the percentage identity may not render all families and will leave out a number of genes. It is used here for demonstration purposes only. For *A. thaliana* such a genome-wide search finds at least 60 gene families, including some R-gene families.

After aligning all family sequences, use PAML's codeml to find evidence of positive selection in the gene families. Clustal Omega is used to align the amino acid sequences and create a phylogenetic tree. Next, pal2nal creates codon alignments, which can be used by PAML. Finally run PAML's codeml M0-M3 (one ratio vs. nearly neutral) tests and M7-M8 (beta vs. beta + ω) tests in a computing cluster environment as shown in Chap. 25.

An M0-M3 $\chi 2$ test finds that 43 gene families (out of 60) show significant evidence of positive selection. M7-M8, meanwhile, finds 35 gene families. Therefore, based on the described procedure, approximately half the families show significant evidence of positive selection and can be considered candidate gene families involved in host-pathogen interactions. Note that this number contains false positives because the evolutionary model may be too simplistic; *see* also [50]. Nevertheless, these candidate gene families can be used as an effective filter for further research.

When a gene family displays evidence of positive selection, the genome locations can be used as a prior for systems genetics (*see* Fig. 1). With the full genome sequence of *A. thaliana* available, the location of gene families showing evidence of positive selection is known. For example, in the *Columbia* (Col-0) ecotype, the majority of the 149 R-genes are combined in clusters spreading 2–9 loci; the remaining 40 are isolated. Clusters are organized in so-called superclusters [36, 51]. Phylogenetic analysis shows that such clusters are the result of both old segmental duplications and recent chromosome rearrangements [36, 52].

2.2 Select a Suitable Experimental Population

To select a suitable experimental population, the choice of parents is key. Because we want a descriptive evolutionary prior based on gene families with known genome locations, we also need a sequenced genome, from one parent and ideally from both of the parental strains. The choice of parents for QTL analysis is normally based on large (classical) phenotypic differences. For testing pathogen resistance, the choice would ideally be one susceptible parent and one resistant (nonsusceptible) parent. For *e*QTL, phylogenetic distance can be used, when there is no obvious phenotype. In general, it is a good idea to choose one or both parents from common library strains based on, for example, *Columbia* (Col-0), *Landsberg erecta* (Ler-0), *Wassilewskija* (Ws-0), or *Kashmir* (Kas-1). This is because a great number of experimental resources and online information will be available. In addition, a reference genetic background is provided in this way, which allows the comparison of the effects of QTL and mutant alleles [53]. A number of RIL populations can be found through TAIR, a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials, and community [49].

2.3 Select an xQTL Technology

A large part of published *x*QTL studies is based on gene expression *e*QTL partly because gene expression probe provides a direct genomic link. When it comes to selecting single-color or two-color arrays, one consideration may be that two-color arrays have higher efficiency when using a distant pair design [54].

Deep sequencing technology (RNA-seq, [55]) is affordable for *e*QTL studies. The main advantage over microarrays is improved signal-to-noise ratios and possibly improved coverage depending on the reference genome. Microarrays are noisy partly due to cross hybridization, e.g., [56], and have limited signal on low-abundance transcripts or expressors; both facts are detrimental to significance. Deep sequencing is no *panacea*, however, since it accentuates the high expressors. High expressors are expressed thousands of times higher than low expressors. Low expressors may lack significance for differential expression. Worse because deep sequencing is stochastic, many low expressors may even be absent. Another point to consider is that currently at least 1 in 1000 nucleotide base pairs is misread, which makes it harder to disentangle error from genetic variation. Only when a sequence polymorphism is measured many times (say 20×), it can be considered to represent genetic variation.

Also a choice for a certain *e*QTL technology should take into account that, when looking at differential gene expression analysis, different microarray platforms agree with each other, but overlap between microarray and deep sequencing is much lower, suggesting a technical bias [57].

For an example of a metabolite *m*QTL study, *see* Keurentjes et al. [58] and Fu et al. [59]. For a study integrating *e*QTL, *p*QTL, *m*QTL, and classical phenotypic QTL, *see* Fu et al. [60] and Jansen et al. [13].

2.4 Sizing the Experimental Population

The size of the experimental population should be large enough to give informative results. For classical QTL analysis, the sizing may be assisted using estimates of total environmental variance and the total genetic variance derived from the accessions, selected as parents. Roughly, population sizes of 200 RILs, without replications, will allow detection of large-effect QTL with an explained variance of 10% in confidence intervals of 10–20 cM. Detection of small-effect QTL or mapping accuracy below 5% requires increasing the population size to at least 300 RILs [53]. It is important to note that QTL mapping accuracy is a function of marker density and population size. The number of strains to use differs between inbred lines. The promise of extreme dense marker maps, such as delivered by SNPs, does not automatically translate to higher accuracy. It is the number of recombination events in the population for a particular genomic region that limits QTL interval size. In fact, current marker maps, in the order of thousands of (evenly spread) markers per genome, suite population sizes of a few hundred RILs.

It is a fallacy, for example, to expect higher mapping power when combining an ultradense SNP map with just 20 individuals.

For high-throughput *x*QTL, the experimental population should be sized against an acceptable false discovery rate (FDR), minimizing for type I and type II errors. This can be achieved using a permutation strategy to assess statistical significance, maintaining the correlation of the expression traits while destroying any genetic linkages or associations in natural populations: marker data is permuted while keeping the correlation structure in the trait data, such as presented by Breitling et al. [61]. Unfortunately, this information differs for every experiment and is only available afterward. Analyzing a similar experiment, using the same tissue and data acquisition technology, may give an indication [60], but when no such material is available, a crude estimate may be had by taking the thresholds of a (classic) single-trait QTL experiment and adjusting that for multiple testing by the Bonferroni correction (minimize type I errors) or Benjamini-Hochberg correction (minimize type II errors). Note that Bonferroni results in a very conservative estimate.

2.5 Analyzing the *x*QTL Experiment with R/qtl

R/qtl is extensible, interactive free software for the mapping of *x*QTL in experimental crosses. It is implemented as an add-on package for the widely used statistical language/software R. Since its introduction, R/qtl has become a reference implementation with an extensive guide on QTL mapping [62].

R/qtl includes multiple QTL mapping (MQM), as described in [10], an automated procedure, which combines the strengths of generalized linear model regression with those of interval mapping. MQM can handle missing data by analyzing probable genotypes. MQM selects important marker cofactors by multiple regression and backward elimination. QTL are moved along the chromosomes using these preselected markers as cofactors. QTL are interval mapped using the most informative model through maximum likelihood. MQM for R/qtl brings the following advantages to QTL mapping: (1) higher power, as long as the QTL explain a reasonable amount of variation; (2) protection against overfitting, because MQM fixes the residual variance from the full model; (3) prevention of ghost QTL detection (between two QTL in coupling phase); and (4) detection of negating QTL (QTL in repulsion phase) [10].

MQM for R/qtl brings additional advantages to systems genetics data sets with hundreds to millions of traits: (5) a pragmatic permutation strategy for control of the FDR and prevention of locating false QTL hot spots, as discussed above; (6) high-performance computing by scaling on multi-CPU computers, as well as clustered computers, by calculating phenotypes in parallel, through the message passing interface (MPI) of the parallel package for R; and (7) visualizations for exploring interactions in a genomic

circle plot and cis- and trans-regulation. MQM comes with a 40-page tutorial for MQM and is part of the software distribution of R/qtl [10, 63].

2.6 Matching the Prior

After detecting *e*QTL, we have a map of gene regulation in the form of a cis-trans map. When taking *a priori* information into account, i.e., genomic locations derived through other methods, we can potentially match the genomic locations of genes and gene families with the *e*QTL cis-trans map. Until now, there has been no combined QTL and evolutionary study, involving PAML, for host-pathogen relationships in plants, though they have been conducted separately.

2.7 Combining xQTL Results: Causality and Network Inference

In addition to identifying *e*QTL or *x*QTL, it is possible to think in terms of grouping related traits by correlations. Molecular and phenotypic traits can be informative for inferring underlying molecular networks. When two independent non-correlated traits share multiple QTL, inference of a functional relationship is possible (Fig. 1b). Thus, distinguishing trait causality, reactivity, or independence can be based upon logic involving underlying QTL. This was the basic idea in Jansen and Nap 2001 [64]. Later, people started to use biological variation as an extra source for reasoning because if A affects B, biological variation in trait A is propagated to B and not vice versa. This assumes there is no hidden trait C affecting both A and B; *see* also Li et al. [15].

Mapping QTL for thousands of molecular phenotypes is the first step in attempting to reconstruct gene networks. Not only can network reconstruction be used within a particular layer, say within *e*QTL analysis, i.e., transcript data only, but also across layers. Such interlevel (system) analysis integrates transcript *e*QTL, protein *p*QTL, metabolite *m*QTL, and classical QTL [13].

The examination of pairwise correlation between traits can lead to the hypothesis of a functional relationship when that correlation is high. 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 A maps to a subset of the QTL of trait B, then the common QTL can be taken as evidence for their network connection, while the distinct QTL can be used to infer the direction (Fig. 1b), unless all the common QTL have widespread pleiotropic effects, which is when a single gene influences multiple traits. If traits A and B have common QTL, without QTL that are distinct, then the inference is more complicated, and further analysis is needed to discriminate pleiotropy from any of the possible orderings among traits [13, 15].

Li et al. [15] point out that, despite the exciting possibilities of correlation analysis, extreme caution is advised, especially in

intralevel analyses, owing to the potential impact of correlated measurement error (leading to false-positive connections). By introducing a prior, however, causal inference becomes feasible for realistic population sizes [15]. The outcome of a causal inference on two traits sharing a common QTL may be either that one is causal for the other or that they are independent. In the first case, QTL-induced variation is propagated from one trait to the other, while in the latter case, the two traits may even be regulated by different genes or polymorphisms within the QTL region, and their apparent relationship (correlation) is explained by linkage disequilibrium and not by a shared biological pathway [15].

3 Discussion

A QTL is a statistical property connecting genotype with phenotype. In this chapter, we reviewed studies which, with various degrees of success, combine some type of prior information with *x*QTL. We propose that a search for genome-wide evidence of positive selection can produce a valid and interesting prior for *x*QTL analysis. This is achieved by combining information of genomic locations of putative gene families, possibly involved in plant-pathogen interactions, with QTL locations derived from a systems genetics experiment. Both the *e*QTL example and the search for genome-wide evidence of positive selection pressure are essentially exploratory and result in a list of putative genes, or gene families, with known genomic locations. The combined information yields candidate genes and pathways that are under positive selection pressure and, potentially, involved in host-pathogen interactions. We explain that it is possible to design an *e*QTL experiment using existing experimental populations, e.g., using an *A. thaliana* RIL population, and analyze results with existing free and open-source software, such as the R/qtl tool set.

Systems genetics bridges the study of quantitative traits with molecular biology and gives new momentum to QTL population studies. 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 and propagation of biological variation, which can be used to infer network causality, reactivity, or independence [15]. Unexplained biological variation can be used to infer direction between traits that share a common QTL and have no distinct QTL, though it may be difficult to separate biological from technical variation. 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 [65].

Evolutionary systems genetics can help dissect the underlying genetics of pathogen susceptibility in plants. Where "evolutionary genetics" describes how evolutionary forces shape biodiversity, as observed in nature, "evolutionary systems genetics" describes how phenotype variation in a population is formed by genotype variation between, for example, host and pathogen involved in an evolutionary arms race.

For purpose of online analysis we created GeneNetwork.org (GN) [7], a free and open-source (FOSS) framework for web-based genetics that can be deployed anywhere. GN allows biologists to upload high-throughput experimental data, such as expression data from microarrays and RNA-seq, and also classical phenotypes, such as disease phenotypes. These phenotypes can be mapped interactively against genotypes using embedded tools, such as R/QTL [10] for model organisms and FaST-LMM [66] and GEMMA [67] which are suitable for human populations and outbred crosses, such as the mouse diversity outcross. Interactive D3 graphics are included from R/qtl charts, and presentation-ready figures can be generated. Recently we have added functionality for phenotype correlation [68], correlation trait loci [16], and network analysis [14]. For examples on using GeneNetwork, *see* also Mulligan et al. [8].

If you want to know more about *e*QTL, we suggest the review by Gilad et al. [23], which also discusses *e*QTL in genome-wide association studies (GWAS), useful in situations where experimental crosses are not available (such as with many pathogens and humans). For further reading on R-gene evolution, we recommend Bakker et al. [34]. For R/qtl analysis, we recommend the R/qtl guide [62] and our MQM tutorial online [63]. For integrating different *x*QTL methods and causal inference, we recommend Li et al. [15] and Jansen et al. [13].

4 Questions

1. What is an *e*QTL, and why does it present two genomic locations?
2. Can a prior, as used here, really add statistical power, or is it no more than circumstantial evidence?
3. When designing an evolutionary systems genetics experiment, what are the steps to consider?
4. How can causality be inferred from QTL networks?

References

1. Qin L, Prins P, Jones JT et al (2001) Genest, a powerful bidirectional link between cdna sequence data and gene expression profiles generated by cdna-aflp. Nucleic Acids Res 29 (7):1616–1622
2. Mckeown PC, Laouielle-duprat S, Prins P et al (2011) Identification of imprinted genes subject to parent-of-origin specific expression in arabidopsis thaliana seeds. BMC Plant Biol 11:113
3. Nandi S, Subudhi PK, Senadhira D et al (1997) Mapping QTLs for submergence tolerance in rice by AFLP analysis and selective genotyping. Mol Gen Genet 255(1):1–8
4. Meaburn E, Butcher LM, Schalkwyk LC, Plomin R (2006) Genotyping pooled DNA using 100K SNP microarrays: a step towards genomewide association scans. Nucleic Acids Res 34 (4):e27
5. Kim S, Plagnol V, Hu TT et al (2007) Recombination and linkage disequilibrium in arabidopsis thaliana. Nat Genet 39(9):1151–1155
6. Dixon AL, Liang L, Moffatt MF et al (2007) A genome-wide association study of global gene expression. Nat Genet 39(10):1202–1207
7. Sloan Z, Arends D, Broman KW et al (2016) Genenetwork: framework for web-based genetics. JOSS 1(2):25
8. Mulligan MK, Mozhui K, Prins P, Williams RW (2017) Genenetwork: a toolbox for systems genetics. Methods Mol Biol 1488:75–120
9. Gibson G, Weir B (2005) The quantitative genetics of transcription. Trends Genet 21 (11):616–623
10. Arends D, Prins P, Jansen RC, Broman KW (2010) R/qtl: high-throughput multiple QTL mapping. Bioinformatics 26 (23):2990–2992
11. Li Y, Alvarez OA, Gutteling EW et al (2006) Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet 2(12):e222
12. Damerval C, Maurice A, Josse JM, De Vienne D (1994) Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics 137(1):289–301
13. Jansen RC, Tesson BM, Fu J, Yang Y, Mcintyre LM (2009) Defining gene and QTL networks. Curr Opin Plant Biol 12(2):241–246
14. Langfelder P, Horvath S (2008) Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics 9:559
15. Li Y, Tesson BM, Churchill GA, Jansen RC (2010) Critical reasoning on causal inference in genome-wide linkage and association studies. Trends Genet 26(12):493–498
16. Arends D, Li Y, Brockmann G et al (2016) Correlation trait loci (ctl) mapping: phenotype network inference subject to genotype. JOSS. 1 (6):87
17. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A 102(5):1572–1577
18. Fraser HB, Moses AM, Schadt EE (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. Proc Natl Acad Sci U S A 107(7):2977–2982
19. Zou Y, Su Z, Yang J, Zeng Y, Gu X (2009) Uncovering genetic regulatory network divergence between duplicate genes using yeast eqtl landscape. J Exp Zool B Mol Dev Evol 312 (7):722–733
20. Li Y, Breitling R, Jansen RC (2008) Generalizing genetical genomics: getting added value from environmental perturbation. Trends Genet 24(10):518–524
21. Hager R, Lu L, Rosen GD, Williams RW (2012) Genetic architecture supports mosaic brain evolution and independent brain-body size regulation. Nat Commun 3:1079
22. Kliebenstein DJ, West MA, Van leeuwen H et al (2006) Identification of QTLs controlling gene expression networks defined a priori. BMC Bioinformatics 7:308
23. Gilad Y, Rifkin SA, Pritchard JK (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 24 (8):408–415
24. Alberts R, Terpstra P, Li Y et al (2007) Sequence polymorphisms cause many false cis eqtls. PLoS One 2(7):e622
25. Franke L, Bakel van H, Fokkens L et al (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78(6):1011–1025
26. Chen X, Hackett CA, Niks RE et al (2010) An eqtl analysis of partial resistance to puccinia hordei in barley. PLoS One 5(1):e8598
27. Qin L, Kudla U, Roze EH et al (2004) Plant degradation: a nematode expansin acting on plants. Nature 427(6969):30
28. Saijo Y, Schulze-lefert P (2008) Manipulation of the eukaryotic transcriptional machinery by

bacterial pathogens. Cell Host Microbe 4 (2):96–99

29. Chen LQ, Hou BH, Lalonde S et al (2010) Sugar transporters for intercellular exchange and nutrition of pathogens. Nature 468 (7323):527–532
30. Hewitson JP, Grainger JR, Maizels RM (2009) Helminth immunoregulation: the role of parasite secreted proteins in modulating host immunity. Mol Biochem Parasitol 167(1):1–11
31. Bird PI, Trapani JA, Villadangos JA (2009) Endolysosomal proteases and their inhibitors in immunity. Nat Rev Immunol 9 (12):871–882
32. Dangl JL, Jones JD (2001) Plant pathogens and integrated defence responses to infection. Nature 411(6839):826–833
33. Flor H (1956) The complementary genic systems in flax and flax rust*. Adv Genet 8:29–54
34. Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell 18 (8):1803–1818
35. Mackey D, Belkhadir Y, Alonso JM, Ecker JR, Dangl JL (2003) Arabidopsis rin4 is a target of the type iii virulence effector avrrpt2 and modulates rps2-mediated resistance. Cell 112 (3):379–389
36. Richly E, Kurth J, Leister D (2002) Mode of amplification and reorganization of resistance genes during recent arabidopsis thaliana evolution. Mol Biol Evol 19(1):76–84
37. Medzhitov R, Janeway CA Jr (1997) Innate immunity: impact on the adaptive immune response. Curr Opin Immunol 9(1):4–9
38. Holub EB (2001) The arms race is ancient history in Arabidopsis, the wildflower. Nat Rev Genet 2(7):516–527
39. Bishop JG, Dean AM, Mitchell-olds T (2000) Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci U S A 97 (10):5322–5327
40. Xiao S, Emerson B, Ratanasut K et al (2004) Origin and maintenance of a broad-spectrum disease resistance locus in Arabidopsis. Mol Biol Evol 21(9):1661–1672
41. Mondragon-palomino M, Meyers BC, Michelmore RW, Gaut BS (2002) Patterns of positive selection in the complete nbs-lrr gene family of Arabidopsis thaliana. Genome Res 12 (9):1305–1315
42. Sun X, Cao Y, Wang S (2006) Point mutations with positive selection were a major force during the evolution of a receptor-kinase resistance gene family of rice. Plant Physiol 140 (3):998–1008
43. Bevan M, Bancroft I, Bent E, Chalwatzis N (1998) Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of Arabidopsis thaliana. Nature 391(6666):485–488
44. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13(5):555–556
45. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
46. Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 7:539
47. Suyama M, Torrents D, Bork P (2006) Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34(Web Server): W609–W612
48. Goto N, Prins P, Nakao M et al (2010) BioRuby: bioinformatics software for the Ruby programming language. Bioinformatics 26 (20):2617–2619
49. Rhee SY, Beavis W, Berardini TZ et al (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Res 31(1):224–228
50. Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164(3):1229–1236
51. Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant arabidopsis thaliana. Nature 408 (6814):796–815
52. Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 8(11):1113–1130
53. Salinas J, Sanchez-serrano J (2006) Arabidopsis protocols. Humana Press Inc, Totowa, NJ
54. Fu J, Jansen RC (2006) Optimal design and analysis of genetic studies on gene expression. Genetics 172(3):1993–1999
55. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods 5(7):621–628
56. Eklund AC, Turner LR, Chen P et al (2006) Replacing cRNA targets with cDNA reduces microarray cross-hybridization. Nat Biotechnol 24(9):1071–1073
57. Hoen PA, Ariyurek Y, Thygesen HH et al (2008) Deep sequencing-based expression

analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res 36 (21):e141

58. Keurentjes JJ, Sulpice R, Gibon Y et al (2008) Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in Arabidopsis thaliana. Genome Biol 9(8): R129
59. Fu J, Swertz MA, Keurentjes JJ, Jansen RC (2007) Metanetwork: a computational protocol for the genetic study of metabolic networks. Nat Protoc 2(3):685–694
60. Fu J, Keurentjes JJ, Bouwmeester H et al (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat Genet 41(2):166–167
61. Breitling R, Li Y, Tesson BM et al (2008) Genetical genomics: spotlight on QTL hotspots. PLoS Genet 4(10):e1000232
62. Broman K, Sen S (2009) A guide to QTL mapping with R/qtl. Springer, New York, NY
63. Arends D, Prins P, Broman KW, Jansen RC (2010) Tutorial - multiple-QTL mapping (MQM) analysis. http://www.rqtl.org/tutorials/MQM-tour.pdf
64. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. Trends Genet 17(7):388–391
65. Wayne ML, Mcintyre LM (2002) Combining mapping and arraying: an approach to candidate gene identification. Proc Natl Acad Sci U S A 99(23):14903–14906
66. Lippert C, Listgarten J, Liu Y et al (2011) Fast linear mixed models for genome-wide association studies. Nat Methods 8(10):833–835
67. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44(7):821–824
68. Wang X, Pandey AK, Mulligan MK et al (2016) Joint mouse-human phenome-wide association to test gene function and disease risk. Nat Commun 7:10464