Chapter 11

Quantification of Variation in Expression Networks

Daniel J. Kliebenstein

Abstract

Gene expression microarrays allow rapid and easy quantification of transcript accumulation for almost transcripts present in a genome. This technology has been utilized for diverse investigations from studying gene regulation in response to genetic or environmental fluctuation to global expression QTL (eQTL) analyses of natural variation. Typical analysis techniques focus on responses of individual genes in isolation of other genes. However, emerging evidence indicates that genes are organized into regulons, i.e., they respond as groups due to individual transcription factors binding multiple promoters, creating what is commonly called a network. We have developed a set of statistical approaches that allow researchers to test specific network hypothesis using a priori-defined gene networks. When applied to *Arabidopsis thaliana* this approach has been able to identify natural genetic variation that controls networks. In this chapter we describe approaches to develop and test specific network hypothesis utilizing natural genetic variation. This approach can be expanded to facilitate direct tests of the relationship between phenotypic trait and transcript genetic architecture. Finally, the use of a priori network definitions can be applied to any microarray experiment to directly conduct hypothesis testing at a genomics level.

Key words: Microarray, network, quantitative, systems biology, hypothesis test.

1. Introduction

Phenotypic variation of animals and plants, including disease susceptibility and development, is controlled by quantitative trait loci (QTLs) whose underlying molecular mechanisms are typically studied in QTL mapping experiments (1–3). QTLs are regions of the genome where genetic diversity is associated with phenotypic variation in a specific trait or, if pleiotropic, a suite of traits. These regions may contain genes whose differential expression controls the associated phenotypic variation. Previous methods to link phenotypic variation to its genetic cause required intensive

Dmitry A. Belostotsky (ed.), *Plant Systems Biology, vol. 553*

DOI 10.1007/978-1-60327-563-7_11 Springerprotocols.com

fine-scale mapping experiments. Recently, the genomic technique of microarray-based transcriptomics has been applied to more quickly link phenotypic trait variation with transcriptome variation. This approach uses microarrays to measure global gene expression across a sample of individuals from a natural population. These gene expression values are then used to map expression QTLs (eQTLs) (4–11) or to assess association between transcript variation and phenotypic variation using association mapping style approaches (12–14). These genomics technologies may enable reverse (natural variation) genetics approaches to identify the genetic basis of quantitative traits and facilitate our understanding of network variation within plants (15–19).

The goal of global eQTL analysis is to quickly identify loci controlling the expression variation of gene networks that control distinct biological functions. One approach (4, 6) is to generate a mapping population, assess global gene expression using microarrays, and identify eQTLs controlling the expression of each gene via individual statistical analyses. The eQTL locations for all genes are then summed, "summation" approach, to identify common regions that control the expression of more genes than expected by random chance, frequently referred to as eQTL hotspots (4, 6, 10, 11, 20–22). This approach is complicated by the potential that individual transcript levels are potentially more variable than the network controlling them. As such, the statistical analysis of individual genes is likely to have significant false-positive and false-negative errors confounding attempts to interpret the biological meaning of any eQTL analysis.

A second complication of the summation is that this requires a posteriori tests to assess whether the genes controlled by an identified eQTL hotspot share a common biological function (e.g., a metabolic pathway, transcriptional co-regulation, similar gene ontology functional annotation) (23, 24). As such, this is descriptive and relies on the presence and absence of individual genes in the list of transcripts significantly controlled by the QTL in question. Hence, we desired to devise a quantitative approach that would allow for the generation of specific hypothesis about transcriptional networks and testing of these hypothesis using microarray analysis of natural genetic variation (25).

In our approach we define the gene networks prior to the statistical analysis allowing quantitative network testing or network eQTL mapping (25). To develop gene networks we rely on existing databases containing either gene co-expression values or predicted metabolic pathways. We define gene networks as a co-regulated set of genes involved in a common biological process. Once we define the networks, we obtain a quantitative measurement of the transcriptional activity of the network by averaging across the individual genes within the network. This single network activity metric can then be used to associate with phenotypic

variation or to map eQTL controlling biological networks. A benefit to this approach is that it is possible to predict a network and then identify the loci controlling the network. Further, it allows for rapid hypothesis generation about the biological impact of specific eQTL clusters. A final use of this approach is to apply standard statistical methodologies to test if networks are regulated in response to diverse inputs using standard experimental designs. In this chapter, we describe the approaches and tools required to generate and evaluate transcriptional networks using natural genetic variation.

2. Materials

2.1. Arabidopsis

An excellent model plant system for studying quantitative genetics is *Arabidopsis thaliana*. There is a rapidly developing set of both genomics tools and genetic variation populations that greatly aid development and testing of approaches to conduct quantitative network analysis of natural variation.

2.1.1. Natural Genetic Populations

Populations used to study natural genetic variation can be generally classified into structured populations or association populations. Structured populations have known parents allowing for accurate recombination measurements and the application of standard QTL mapping approaches (2). Recently, natural genetic variation in association populations has begun to be queried using linkage disequilibrium mapping approaches (26–28). Structured mapping populations have less genetic variation than association populations but it is unknown if this difference in genetic variation necessarily correlates to levels of phenotypic variation in the two population structures.

2.1.1.1. Structured Populations

In *Arabidopsis*, the main structured populations are made using the recombinant inbred line (RIL) structure where two parents are crossed and the progeny then undergo single seed descent for at least eight generations. After eight generations each resulting line is a homozygous mixture of the two parental genotypes. There are numerous RIL populations in existence in *Arabidopsis*, with the main populations being the Bay-0 × Sha, L*er* × Col-0 and L*er* × Cvi (29–33). These populations are of decently large size and have been phenotyped for innumerous diverse phenotypes. In addition to these populations, there are new populations in development or recently released (34–36). Important features of these populations are that they have already been genetically mapped and this information and the lines are or soon will be available from The Arabidopsis Resource Center (www.arabidopsis.org).

2.1.1.2. Association Populations

Recent work is suggesting that association mapping populations are a complementary approach to using structured populations for quantitative analysis of networks (27). These populations consist of large collections of diverse *Arabidopsis* accessions with unknown ancestry (26–28). These populations are designed to contain the vast majority of genetic diversity within *Arabidopsis* providing a rich source of allelic diversity. This is done by sampling a very large population of accessions and then choosing a smaller experimental population that contains the maximal level of diversity within the larger population. The individual accessions have been genotyped at a large number genetic loci using genomics technologies including near complete genome resequencing (37–39) and this sequence or genotyping information is freely available (www.arabidopsis.org). The accessions in these populations are freely available from The Arabidopsis Resource Center (www.arabidopsis.org).

2.1.2. Microarray Data

2.1.2.1. Genetic Variation Data Sets

Microarrays have been utilized to survey transcript accumulation variation in structured *Arabidopsis* populations (10, 11) and small association populations (12–14, 40). The microarray data for the Bay × Sha RIL population and one small association population can be obtained from elp.ucdavis.edu (10, 12, 14). Alternatively, this data can be downloaded from ArrayExpress as data sets E-TABM126 and E-TABM62 (www.ebi.ac.uk/microarray-as/aer/?#ae-main[0]). This database will provide either the raw .CEL files or the normalized gene expression data. Replicated microarray data for another association population can be downloaded from www.weigelworld.org/resources/microarray/AtGenExpress (40). Currently, the microarray data on the L*er* × Cvi RIL population appears to be available via personal communication with the authors (11).

2.1.2.2. Co-expression Databases

The transcriptomic response of *Arabidopsis* to various environmental, genetic, and developmental perturbations has been intensively queried using microarrays. Most of this data is compiled into databases including www.genevestigator.org, www.Arabidopsis.leeds.ac.uk/ACT, and http://www.atted.bio.titech.ac.jp (41–46). These databases allow the researcher to enter a specific gene or set of genes to identify all other genes that show similar transcriptional variation within the whole database or a subset of the database. This provides an excellent data source for the generation of hypothetical networks as described in **Section 3.1.1**.

2.1.3. Metabolic Network Databases

Biosynthetic pathways are frequently co-regulated at the transcript level and as such are excellent sources of network hypothesis (47, 48). The Aracyc database for *Arabidopsis* contains an extensive list of enzyme encoding genes and their predicted or proven reactions. This database links enzymes and their corresponding genes

into predicted or proven metabolic pathways that can be treated as networks (49, 50). This includes both primary and secondary metabolic networks. This database is readily accessible or completely downloadable at the *Arabidopsis* webpage (www.arabidopsis.org) to aid in network generation as described in **Section 3.1.3**.

2.2. Barley

Barley (*Hordeum vulgare*) is the other plant species that has a large existing mapping population that has been intensively analyzed using genomic microarray data. These are both required to enable a network analysis of network eQTL.

2.2.1. Natural Genetic Populations

The main population for quantitative analysis of transcript networks in Barley is a doubled haploid population obtained from a cross of the Steptoe and Morex inbred parents. This doubled haploid population consists of 139 lines that have been high-throughput genotyped to create a dense marker map (51). This population also has extensive phenotypic information available for the lines across multiple environments with significant replication (wheat.pw.usda.gov/ggpages/SxM/phenotypes.html).

2.2.2. Microarray Data

2.2.2.1. Genetic Variation Data Sets

The microarray data for the Steptoe × Morex DH population is available from ArrayExpress as data set E-TABM-112 (http://www.ebi.ac.uk/microarray-as/aer/?#ae-main[0]) (9, 51). This database will provide either the raw .CEL files or the normalized gene expression data.

2.2.2.2. Co-expression Databases

Barleybase (www.barleybase.org) is a database containing numerous microarray experiments from Barley that can allow a researcher to query for co-expressed genes (52, 53). Additionally, microarray data can be downloaded to allow researchers to apply their own co-expression analysis or alter the statistical parameters at their desire. This can be done using a validated batch-learning self-organizing map approach as previously described (54). This provides an excellent data source for the generation of hypothetical networks as described in **Section 3.1.1**.

2.3. Other Species

While barley and *Arabidopsis* are currently the plant species that contain both the genetic populations and microarray analysis to allow for large-scale quantitative analysis of network variation, there are additional projects underway that will assuredly generate similar data for other species. For instance, maize and rice have large mapping populations available that only require the application of microarrays to generate the necessary transcript variation measures (55). Numerous other plants have had targeted microarray analysis of natural genetic variation to address specific questions showing the broad applicability of this technology (8, 56–63).

3. Methods

In the a priori approach to network analysis of gene expression, the hypothetical networks are defined prior to the analysis of the microarray data. The goal of this a priori network approach is to allow the researcher to develop hypotheses about gene sets using prior information and then test these hypotheses utilizing the gene networks and microarray data. For instance, a researcher could hypothesize that a set of genes are critical for defense against a given pathogen. The researcher can then use the following methods to identify pathogen response networks, map eQTL controlling these networks, and compare the resulting data to QTL controlling resistance against the pathogen. Alternatively, these same approaches can be used to directly test if two genotypes that differ in resistance also differ in the expression of their hypothetical defense network. The applications of this approach are only limited to a researcher's ability to generate hypothesis and conduct the experiment.

3.1. Network Assignment

The first step required in this method is to generate groups of genes for which the researcher thinks there is support to presume or hypothesize that the genes within the group are coordinately regulated. The evidence for gene network assignment can be generated from genes having coordinate regulation, having a similar biological function or from numerous existing and developing genomics databases.

3.1.1. Gene Co-expression

Numerous plant species have existing databases containing large collections of microarray analysis which allow for researchers to identify co-regulated genes. These co-regulated genes can function as a priori-defined gene networks that can be used for further analysis. There are two predominant avenues to querying gene expression databases for co-regulated gene networks, the "guide-gene" and "non-targeted" approaches (54, 64).

3.1.1.1. "Guide-Gene" Approach to Co-expression Clustering

The simplest approach to using genomic expression databases for generating co-regulated gene networks is the "guide-gene" approach (54). The guide-gene approach involves researchers identifying their favorite gene, inputting it into the available databases, or using their own statistical analysis to identify all other genes in the genome that show a significant positive correlation across the available microarray data. This positive correlation suggests that these genes are controlled by the same regulatory network with the same directionality. These genes can then be classified as a co-regulated network. *See* **Section 3.1.6** for a discussion of the optimal size of co-regulated gene networks. *See* **Section 3.1.7** for a discussion of correlation thresholds and the potential ramification on the network's utility.

3.1.1.2. "Non-targeted" Approach to Co-expression Clustering

A more intensive and global approach to network definition using co-expression databases is to take the complete data set and compile all gene-to-gene correlations and then utilize this to conduct a complete clustering of all genes based on their correlation (42, 54, 65, 66). This approach will generate massive interconnected gene networks that can be utilized to create putative co-regulated gene networks (66). The genomic network requires dissection into discrete co-regulated gene networks that can then be handled individually. This dissection can be accomplished by deciding upon a correlation threshold required between genes to classify them as a co-regulated network. *See* **Section 3.1.7** for a discussion about correlation directionality and thresholds for calling co-regulated genes. An alternative to the hard correlational threshold is to visually inspect the networks and dissect them based on the density of clustering. Network diagrams typically are comprised of dense local gene clusters that are connected to other clusters via sparser interactions. A researcher could decide that they will dissect clusters based upon the frequency of interconnections within a cluster versus those between clusters. This would not require a hard correlational threshold and may yield more biologically relevant clusters (66).

3.1.2. Metabolic Pathway Network Definition

A useful method to define coordinated biological function is the cooperation of enzymes within a biosynthetic pathway. There are multiple databases containing both validated and predicted metabolic pathways present in *Arabidopsis* and other plant species (49, 50, 67, 68). As biosynthetic pathways exist to optimally transmute a beginning substrate to an end product, the genes in a metabolic pathway are frequently co-regulated (16, 25, 47, 48). As such, metabolic pathways provide an excellent beginning with which to predict coordinate gene expression networks. The available databases can be downloaded to generate a ready network list that can be further modified to the researcher's specific aims.

3.1.3. Protein Interaction Network Definition

Modern genomics technologies are providing a diverse array of data sets to allow gene networks to be defined and then tested. One such genomics data set allowing gene network prediction is protein interaction networks (69–71). These interaction networks predict the presence of protein complexes whose members are likely to be coordinately regulated to provide a common outcome (72, 73). There are two forms of protein interaction networks. In plants, the most common data currently available are for individual protein complexes (73). Another form of data that is coming is massive interactome maps attempting to illustrate all possible protein–protein interactions (69–71). While these interactome maps are highly complex, they do highlight local protein clusters that appear to function in protein complexes (72). A researcher could define the proteins/genes in a local cluster as likely to

function in a coordinate fashion and as such be a good candidate for a coordinately regulated gene network. *See* **Section 3.1.6** for a discussion of the optimal size of co-regulated gene networks.

3.1.4. Other Potential Biological Definitions

The above approaches to generating hypothetical gene networks for further testing are not meant to exclude other approaches. In fact, each approach to a priori network definition inherently limits and frames both the questions being tested and the answers obtained. For instance, gene networks defined a priori using metabolic pathways allow a research to test how their experimental variable X controls gene expression for the biosynthetic pathway. Similarly, the proteomics definition limits any test to addressing how the protein complex may be regulated. As such any approach can be used to define the networks and the specific approach to network definition should be chosen to maximize the precision and/or power of the future tests. For instance, if a researcher is interested in using microarray data to address natural variation in trichomes, then a network defined by genes exclusively or predominantly expressed in trichomes will be more powerful than a proteomic or metabolic pathway-defined network. Any data that can allow a researcher to generate a group of genes logically expected to be co-regulated is a valid approach to a priori gene network definition. As the network is simply a tool for hypothesis testing it does not have to be "correct"; future experiments will test the correctness of the original definition.

3.1.5. Duplicated Genes and Optimizing Network Definitions

One complexity of plant genomes is the vast amount of gene duplication that has occurred (74–77). This can lead to the duplication of entire gene networks allowing the duplicated networks to obtain similar but distinct biological functions that may not be co-regulated. For instance, in maize several tryptophan biosynthetic genes have been duplicated and recruited for 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) synthesis which is regulated differently from the tryptophan biosynthetic pathway in maize (78). If a researcher's network is defined using protein interaction or metabolic pathways, it is possible that there are duplicated copies of this network, each with its own regulation pattern. As such, the overlapping patterns would diminish the ability to identify a signal of network co-expression.

A simple approach for researchers to test their network for the presence of duplicated networks with opposing expression patterns is to obtain microarray measures of gene expression and conduct a correlational analysis among the genes within a network. If all members of the network are co-regulated, they will show a positive correlation. Genes that constitute a separate network will show no or negative correlation with the other network genes. An illustration of this principle comes from previous work in *Arabidopsis* that utilized metabolic pathway definitions to initially define

networks (25). Correlational analysis within these metabolically defined networks showed that each metabolic pathway typically had two different gene networks with opposing gene regulation (25). For instance, the genes predicted for lignin biosynthesis could be separated into two complete lignin sub-pathways that showed a positive correlation within each sub-pathway and negative correlation between the two sub-pathways. This correlational separation of duplicated networks should always be used to maximize the precision of any network definition before proceeding to specific network testing as described in **Section 3.2**.

3.1.6. Number of Genes in a Network

An important consideration in any network definition is how the number of genes within a network may affect future tests of that network. If a network has too few genes, then any statistical test using that network will be sensitive to variation in individual genes. This could create or destroy network significance due to error or variation in an individual gene within the network. Conversely if a network has too many genes, then these genes are likely integrating diverse and independent regulatory inputs and any desired biological specificity may be lost. As such, expression across very large gene networks may act as a measure of the plant's physiological status complicating the ability to resolve and specific biological phenomena (10, 25, 63, 79). Thus, to maximize the statistical power in terms of error potential and to increase the precision on the biological questions being asked, networks must be of a moderate gene membership.

In practice, the minimal gene membership within a network should be no fewer than five, with ten genes being a more optimal limit (12, 16, 25). The upper boundary of a network gene population is harder to define as this is dependent upon the co-regulation among members of a gene network. If the network members are absolutely co-regulated with no other influences separating them, then the network can be of any size. In practice, an analysis of eQTL in *Arabidopsis* showed that gene networks with more than 50 genes typically identified a limited set of eQTL hotspots whereas gene networks of 25 were more specific (Kliebenstein, unpublished data). This suggests that somewhere between 25 and 50 is likely the upper bound of the optimal gene network in *Arabidopsis* for network expression analysis. However, if the physiological measurements are the desired outcome of any analysis, then larger networks are valid uses of this a priori network approach.

3.1.7. Strictness of Network Definition

It is important for the ensuing network analysis that only those genes showing a positive correlation are considered as a co-regulated gene network. Admittedly, many regulatory networks have both positive and negative consequences on gene expression. However, the inclusion of negatively regulated genes would cause

the "signal" from the co-regulated gene network to be diminished because these genes' negative changes would erase the positive regulation in the other genes within the co-regulated network. If the researcher feels that the negatively regulated genes are of sufficient interest to merit inclusion the solution is to create a separate negatively co-regulated network for analysis. If in fact the two gene networks are controlled by the same regulatory machinery in different directions, then the two co-regulated gene networks will identify the same factors in the ensuing experiment and can strengthen the researcher's interpretations.

Another important factor in generating gene groups via the co-expression analysis is the level of correlation between the input gene and the other genes that is used as the threshold for calling genes as a co-regulated network. This threshold will impact the results obtained from any network analysis of these genes. While there are no absolute thresholds that can be universally applied, in general the tighter the correlation required to call a group of genes a co-regulated network, the more likely that they will be regulated by a single transcriptional network. The lower the correlation between the genes in the network, the more likely they are regulated by a mixture of transcriptional networks. In this case, the gene network may actually function as more of a measure of some specific physiological condition such as drought or general stress level. Thus, the choice of the correlation level for defining networks by the guide-gene approach will likely alter the results from any network analysis.

3.2. Network Testing with Natural Variation Data

The above approaches to defining gene networks provide the opportunity to test a networks quantitative response to natural genetic variation. This can be in the form of a network eQTL analysis which only requires small changes to the standard single trait methodologies with which most laboratories are familiar. Below, we present a discussion of approaches to analyze a priori networks using eQTL analysis.

3.2.1. Network eQTL Analysis

After previous microarray data from the desired population is obtained (*see* **Sections 2.1.2.2** and **2.2.2.2**) or microarray data from a new population has been generated and gene networks have been defined, the next step is to identify eQTLs controlling these a priori-defined gene networks for which there are two basic approaches readily available to most labs. These are the average z score approach and the multi-trait approach as described below.

3.2.1.1. Average z Score Approach to A Priori Network eQTL

One approach to map network eQTLs for a priori-defined gene networks is to use standard software packages such as QTL Cartographer (80, 81). This requires the generation of a single metric describing the expression of the gene network. In traditional QTL mapping, a single metric for the trait is measured and entered into

the QTL algorithm, for example the accumulation of a metabolite. The development of a single metric for a priori-defined gene networks is complicated by the genes having widely varying expression ranges (25). If this difference between genes is not corrected, variation in any single metric for the network will be dominated by those genes with higher expression and defeat the ability of an a priori network to encapsulate the information provided by all genes within the network. One solution to this complication is to conduct a simple mean centering. In this approach, the average expression across the different lines for each gene is set to a preordained value, say 0. The actual value for each gene is independently normalized by subtracting the measured gene expression value in that line by that gene's average expression measured across all lines. This is similar to the RMA adjustment for microarrays where the average gene expression per microarray is set to a constant and the transcript accumulation within each microarray is normalized accordingly (82). While a simple mean-centering approach does normalize the means, it does not compensate for genes with large expression ranges also having larger variances.

Simultaneously compensating for differences in variance and mean expression requires the use of the z scores for each gene within the network (25). This requires standardizing the expression of each gene in each line to its z score. This is accomplished by first subtracting the expression of each gene in each line by the average expression of that gene across all lines. This value is then divided by the standard deviation of that gene's expression across all lines. This forces all genes within the network to have an average expression of 0 and a standard deviation of 1 across all lines. Once the z score for each gene in each line has been determined, the average z score across the genes in the a priori gene network is measured in each line. This provides a single metric or number for the a priori gene networks expression that can be entered into a lab's favorite QTL mapping package to identify network eQTL using all appropriate significance determinations as would be conducted for any other trait (83–86).

3.2.1.2. Multi-trait Approach to A Priori Network eQTL

Multi-trait mapping algorithms provide a second approach to mapping eQTLs for a priori-defined gene networks. These algorithms were initially developed to test for QTLs across multiple environments (87–89). In the standard approach to multi-trait mapping, the same trait is measured in multiple environments and QTLs are mapped in each environment and across the environments. The multi-trait algorithms can be adapted to map gene network QTLs by treating each gene in the network as a separate measure of the gene network's response, hence treating each network as a different "environment" measure of the trait (90). The genes can then be entered into the multi-trait algorithms and

eQTLs that map across the genes (environments) are the network eQTL for that specific a priori-defined gene network. An advantage to this approach is that gene-specific eQTLs can be rapidly identified in the ensuing QTL analysis. Additionally if any genes obviously behave differently than the other genes in the network in the multi-trait analysis, they can be dropped from the network and the eQTL analysis repeated to test if this better refines the a priori network. This approach can likely be extended into the more complex Bayesian QTL approaches being developed (90–93).

3.3. Network Testing of Experimental Data

In addition to allowing for analysis of natural variation in gene expression networks, the a priori definition approaches also provide the opportunity to test the network's quantitative response to more traditional experimental variation. This experimental variation could be in the form of environmental or genetic perturbation of the plant. Further, the a priori network analysis only requires small changes to the standard single gene methodologies with which most laboratories are familiar. This approach should be applicable to network testing of metabolomics data (*see* **Notes 1 and 2** for brief discussion).

3.3.1. Experimental Design

If the a priori network is being used to test existing microarray data sets for a network's regulation, then the researcher is limited to what the existing experiments allow. However, the researcher can utilize this a priori network approach to test a network's response to new experimental variables that were not a factor in the network's definition (12, 17). In this case, standard experimental designs should be followed to maximize the statistical power just as if the researcher was focusing on a single gene rather than a network. There is some thought that a network analysis may not require as much replication as an individual gene. However, as the basis of the a priori network approach is that there is a single underlying biological mechanism for the gene's co-regulation, it is possible that the variation present in this biological mechanism is similar to the variation identified in a single gene. This is shown by the lack of increased genetic heritability for the aliphatic glucosinolate network in comparison to the average heritability for the underlying genes (16). Further, individual genes and the networks within which they reside appeared to control similar levels of variation across *Arabidopsis* accessions, suggesting that gene networks and individual genes require similar levels of replication (25). As such, it is advisable to conduct sufficient replication with an experimental design meant to control for and minimize error as much as possible.

3.3.2. Nested ANOVA of Experimental Variables

One key aspect of the a priori network definition is that it facilitates the direct testing of gene network responses to experimental perturbation. This can be done using any standard experimental

design meant to query gene expression responses to biotic, abiotic, or genetic perturbations. For this analysis, the gene networks are designed as described (**Section 3.1**), the appropriate experiment conducted, and data collected. The experiment can be a microarray analysis of a wild-type plant versus a mutant, plants grown in normal versus drought conditions, or a factorial experiment combining different experimental factors. An a priori network analysis of this data only requires a modification of the traditional ANOVA that many laboratories already utilize. In this modification, gene and gene network membership for each gene are both entered into the statistical analysis as separate variables. The data are then analyzed as a nested ANVOA whereby gene is nested under the gene network term (25). For instance, genes A, B, and C are considered members of network X and genes D, E, and F are members of network Y. This allows the data for each gene's expression data to be used by the model but only within the specific gene network in which that gene resides. This allows the model to compare expression variation between genes within a network to that between specific networks. For example, variation within the genes A, B, and C for network X is analyzed separately to the variation for genes D, E, and F in network Y. Finally, the variation between network X and Y is analyzed. Additionally, a nested ANOVA can compare the level of variation controlled by each component of the model. For instance, an analysis of natural variation in *Arabidopsis* gene expression suggested that network variation was on a similar order of individual gene variation (25). The ANOVA can then be extended to directly test for effects of different experimental perturbations upon the networks.

An example of this nested ANOVA approach is an analysis of how modifying three MYB transcription factors within *A. thaliana* altered the expression of sulfur utilization networks. In this experiment, WT and the different MYB expression lines were measured with replicated microarrays. The nested ANOVA tested if the introduction of the MYBs into *Arabidopsis* predominantly altered individual genes or the sulfur utilization networks within which the genes reside (17). This found a significant effect of the transcription factors upon the different networks, showing that the MYBs control distinct sulfur networks (17). The nested ANOVA can be easily implemented in any statistical package. However for very large data sets containing numerous genes and networks, the R platform is likely better due to more efficient matrix inversion algorithms. Smaller more discrete tests are feasible in any statistical package.

3.4. Conclusions

Genomics experiments are sometimes thought of as limited to generating hypothesis that are then tested by other methodologies. This leaves a need for developing approaches to allow for hypothesis testing using genomics-scale experiments. In this methods description, we relay one approach to using genomics

data, specifically microarray data, to directly test hypothesis and map genetic variation for a priori-defined gene networks. This a priori approach has been mostly used for the analysis of eQTL controlling gene networks but can be extended to nearly any experimental approach. The methods described in this chapter are readily accessible to any laboratory with basic statistical programs such as Excel, R, SAS, or Systat and do not require any special programming. As such, these methods should allow any researcher to being treating gene networks as testable hypothesis using existing or new microarray data. This should allow for an increase in specific biological inference to be derived from transcriptomics data and experiments in any species. Finally, the approaches described here can be adapted to any genomics platform such as metabolomics whereby quantitative measurements of network members can be conducted and networks can be defined.

4. Notes

1. Applying the a priori network approach to metabolomics would be feasible to compare the network responses of biosynthetic pathways, i.e., TCA cycle, to the responses of the individual metabolites within the pathway.
2. A caveat to applying any expression analysis approaches to metabolite analysis is that metabolites can be interconverted from one to another. In contrast, the transcript for one gene cannot be directly converted into the transcript for another gene. As such, this difference in the relationship between metabolites and the relationship between transcripts may generate different variance properties in the two genomics data sets.

Acknowledgments

Funding for this methods development was obtained by a National Science Foundation grants DBI 0642481 to DJK.

References

1. Zeng, Z.-B., Kao, C.-H., and Basten, C.J. (1999) Estimating the genetic architecture of quantitative traits. *Genetic Research* **75**, 345–355.
2. Mackay, T.F.C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* **35**, 303–339.
3. Lander, E.S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.
4. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G.,

Linsley, P.S., Mao, M., Stoughton, R.B., and Friend, S.H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302.

5. Craig, B.A., Black, M.A., and Doerge, R.W. (2003) Gene expression data: the technology and statistical analysis. *Journal of Agricultural Biological and Environmental Statistics* **8**, 1–28.
6. Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755.
7. Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics* **17**, 388–391.
8. Kirst, M., Basten, C.J., Myburg, A.A., Zeng, Z.B., and Sederoff, R.R. (2005) Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. *Genetics* **169**, 2295–2303.
9. Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsey, M. (2007) Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant Journal* doi: 10.1111/j.1365-313X.2007.03315.x.
10. West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., Doerge, R.W., and St. Clair, D.A. (2007) Global eQTL mapping reveals the complex genetic architecture of transcript level variation in Arabidopsis. *Genetics* **175**, 1441–1450.
11. Keurentjes, J.J.B., Fu, J.Y., Terpstra, I.R., Garcia, J.M., van den Ackerveken, G., Snoek, L.B., Peeters, A.J.M., Vreugdenhil, D., Koornneef, M., and Jansen, R.C. (2007) Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 1708–1713.
12. Van Leeuwen, H., Kliebenstein, D.J., West, M.A.L., Kim, K.D., van Poecke, R., Katagiri, F., Michelmore, R.W., Doerge, R.W., and St. Clair, D.A. (2007) Natural variation among *Arabidopsis thaliana a*ccessions for transcriptome response to exogenous salicylic acid. *Plant Cell* **19**, 2099–2110.
13. Van Poecke, R.M.P., Sato, M., Lenarz-Wyatt, L., Weisberg, S., and Katagiri, F. (2008) Natural variation in RPS2-mediated resistance among Arabidopsis accessions: correlation between gene expression profiles and phenotypic responses. *Plant Cell* **19**, 4046–4060.
14. Kliebenstein, D.J., West, M.A.L., Van Leeuwen, H., Kyunga, K., Doerge, R.W., Michelmore, R.W., and St. Clair, D.A. (2006) Genomic survey of gene expression diversity in *Arabidopsis thaliana*. *Genetics* **172**, 1179–1189.
15. Flint, J., Valdar, W., Shifman, S., and Mott, R. (2005) Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Reviews Genetics* **6**, 271–286.
16. Wentzell, A.M., Rowe, H.C., Hansen, B.G., Ticconi, C., Halkier, B.A., and Kliebenstein, D.J. (2007) Linking metabolic QTL with network and *cis*-eQTL controlling biosynthetic pathways. *PLoS Genetics* **3**, e162.
17. Sønderby, I.E., Hansen, B.G., Bjarnholt, N., Ticconi, C., Halkier, B.A., and Kliebenstein, D.J. (2007) A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *PLoS ONE* **2**, e1322.
18. Hansen, B.G., Kliebenstein, D.J., and Halkier, B.A. (2007) Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in Arabidopsis. *Plant Journal* **50**, 902–910.
19. Zhang, Z.-Y., Ober, J.A., and Kliebenstein, D.J. (2006) The gene controlling the quantitative trait locus *EPITHIOSPECIFIER MODIFIER1* alters glucosinolate hydrolysis and insect resistance in Arabidopsis. *Plant Cell* **18**, 1524–1536.
20. Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R., and Kruglyak, L. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nature Genetics* **35**, 57–64.
21. Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., Su, A.I., Vellenga, E., Wang, J.T., Manly, K.F., Lu, L., Chesler, E.J., Alberts, R., Jansen, R.C., Williams, R.W., Cooke, M.P., and de Haan, G. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics' *Nature Genetics* **37**, 225–232.
22. Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsey, M. (2008) Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant Journal* **53**, 90–101.
23. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub,

T.R., Lander, E.S., and Mesirov, J.P. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–15550.

24. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M.J., Patterson, N., Mesirov, J.P., Golub, T.R., Tamayo, P., Spiegelman, B., Lander, E.S., Hirschhorn, J.N., Altshuler, D., and Groop, L.C. (2003) PGC-1 a-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267–273.
25. Kliebenstein, D., West, M., van Leeuwen, H., Loudet, O., Doerge, R., and St. Clair, D. (2006) Identification of QTLs controlling gene expression networks defined a priori. *BMC Bioinformatics* **7**, 308.
26. Zhao, K.Y., Aranzana, M.J., Kim, S., Lister, C., Shindo, C., Tang, C.L., Toomajian, C., Zheng, H.G., Dean, C., Marjoram, P., and Nordborg, M. (2007) An Arabidopsis example of association mapping in structured samples. *PLoS Genetics* **3**, e4.
27. Weigel, D. and Nordborg, M. (2005) Natural variation in arabidopsis. How do we find the causal genes? *Plant Physiology* **138**, 567–568.
28. Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., Stahl, E.A., and Weigel, D. (2002) The extent of linkage disequilibrium in Arabidopsis thaliana. *Nature Genetics* **30**, 190–193.
29. Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., and Daniel-Vedele, F. (2002) Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theoretical and Applied Genetics* **104**, 1173–1184.
30. El-Assal, S.E.D., Alonso-Blanco, C., Peeters, A.J.M., Raz, V., and Koornneef, M. (2001) A QTL for flowering time in Arabidopsis reveals a novel allele of *CRY2*. *Nature Genetics* **29**, 435–440.
31. Koornneef, M., Alonso-Blanco, C., and Vreugdenhil, D. (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annual Review of Plant Biology* **55**, 141–172.
32. Clarke, J., Mithen, R., Brown, J., and Dean, C. (1995) QTL analysis of flowering time in *Arabidopsis thaliana*. *Molecular and General Genetics* **248**, 278–286.
33. Lister, C. and Dean, D. (1993) Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant Journal* **4**, 745–750.
34. Perchepied, L., Kroj, T., Tronchet, M., Loudet, O., and Roby, D. (2006) Natural variation in partial resistance to *Pseudomonas syringae* is controlled by two major QTLs in *Arabidopsis thaliana*. *PLoS ONE* **1**, e123.
35. Symonds, V.V., Godoy, A.V., Alconada, T., Botto, J.F., Juenger, T.E., Casal, J.J., and Lloyd, A.M. (2005) Mapping quantitative trait loci in multiple populations of *Arabidopsis thaliana* identifies natural allelic variation for trichome density. *Genetics* **169**, 1649–1658.
36. El-Lithy, M.E., Bentsink, L., Hanhart, C.J., Ruys, G.J., Rovito, D.I., Broekhof, J.L.M., van der Poel, H.J.A., van Eijk, M.J.T., Vreugdenhil, D., and Koornneef, M. (2006) New Arabidopsis recombinant inbred line populations genotyped using SNPWave and their use for mapping flowering-time quantitative trait loci. *Genetics* **172**, 1867–1876.
37. Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., Jakobsson, M., Kim, S., Morozov, Y., Padhukasahasram, B., Plagnol, V., Rosenberg, N.A., Shah, C., Wall, J.D., Wang, J., Zhao, K., Kalbfleisch, T., Schulz, V., Kreitman, M., and Bergelson, J. (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology* **3**, e196.
38. Borevitz, J.O., Hazen, S.P., Michael, T.P., Morris, G.P., Baxter, I.R., Hu, T.T., Chen, H., Werner, J.D., Nordborg, M., Salt, D.E., Kay, S.A., Chory, J., Weigel, D., Jones, J.D.G., and Ecker, J.R. (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12057–12062.
39. Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., Chen, H.M., Frazer, K.A., Huson, D.H., Schoelkopf, B., Nordborg, M., Raetsch, G., Ecker, J.R., and Weigel, D. (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* **317**, 338–342.
40. Lempe, J., Balasubramanian, S., Sureshkumar, S., Singh, A., Schmid, M., and Weigel, D. (2005) Diversity of flowering responses

in wild *Arabidopsis thaliana* strains. *PLoS Genetics* **1**, 109–118.

41. Manfield, I.W., Jen, C.H., Pinney, J.W., Michalopoulos, I., Bradford, J.R., Gilmartin, P.M., and Westhead, D.R. (2006) Arabidopsis co-expression tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Research* **34**, W504–W509.
42. Obayashi, T., Kinoshita, K., Nakai, K., Shibaoka, M., Hayashi, S., Saeki, M., Shibata, D., Saito, K., and Ohta, H. (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. *Nucleic Acids Research* **35**, D863–D869.
43. Grennan, A.K. (2006) Genevestigator: facilitating web-based gene-expression analysis. *Plant Physiology* **141**, 1164–1166.
44. Jen, C.H., Manfield, I.W., Michalopoulos, I., Pinney, J.W., Willats, W.G.T., Gilmartin, P.M., and Westhead, D.R. (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. *Plant Journal* **46**, 336–348.
45. Zimmermann, P., Hennig, L., and Gruissem, W. (2005) Gene-expression analysis and network discovery using Genevestigator. *Trends in Plant Science* **10**, 407–409.
46. Obayashi, T., Okegawa, T., Sasaki-Sekimoto, Y., Shimada, H., Masuda, T., Asamizu, E., Nakamura, Y., Shibata, D., Tabata, S., Takamiya, K.I., and Ohta, H. (2004) Distinctive features of plant organs characterized by global analysis of gene expression in arabidopsis. *DNA Research* **11**, 11–25.
47. Gachon, C.M.M., Langlois-Meurinne, M., Henry, Y., and Saindrenan, P. (2005) Transcriptional co-regulation of secondary metabolism enzymes in Arabidopsis: functional and evolutionary implications. *Plant Molecular Biology* **58**, 229–245.
48. Wei, H.R., Persson, S., Mehta, T., Srinivasasainagendra, V., Chen, L., Page, G.P., Somerville, C., and Loraine, A. (2006) Transcriptional coordination of the metabolic network in Arabidopsis. *Plant Physiology* **142**, 762–774.
49. Zhang, P.F., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D., and Rhee, S.Y. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Physiology* **138**, 27–37.
50. Mueller, L.A., Zhang, P.F., and Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiology* **132**, 453–460.
51. Luo, Z.W., Potokina, E., Druka, A., Wise, R., Waugh, R., and Kearsey, M. J. (2007) SFP genotyping from Affymetrix arrays is robust but largely detects cis-acting expression regulators. *Genetics* **176**, 789–800.
52. Richardson, A., Boscari, A., Schreiber, L., Kerstiens, G., Jarvis, M., Herzyk, P., and Fricke, W. (2007) Cloning and expression analysis of candidate genes involved in wax deposition along the growing barley (Hordeum vulgare) leaf. *Planta* **226**, 1459–1473.
53. Shen, L.H., Gong, J., Caldo, R.A., Nettleton, D., Cook, D., Wise, R.P., and Dickerson, J.A. (2005) BarleyBase – an expression profiling database for plant genornics. *Nucleic Acids Research* **33**, D614–D618.
54. Saito, K., Hirai, M., and Yonekura-Sakakibara, K. (2008) Decoding genes with coexpression networks and metabolomics – 'majority report by precogs'. *Trends in Plant Science* **13**, 36–43.
55. Kearsey, M.J. and Farquhar, A.G.L. (1998) QTL analysis in plants; where are we now? *Heredity* **80**, 137–142.
56. Jordan, M.C., Somers, D.J., and Banks, T.W. (2007) Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnology Journal* **5**, 442–453.
57. DeCook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006) Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics* **172**, 1155–1164.
58. Juenger, T.E., Wayne, T., Boles, S., Symonds, V.V., McKay, J., and Coughlan, S.J. (2006) Natural genetic variation in whole-genome expression in *Arabidopsis thaliana*: the impact of physiological QTL introgression. *Molecular Ecology* **15**, 1351–1365.
59. Street, N.R., Skogstrom, O., Sjodin, A., Tucker, J., Rodriguez-Acosta, M., Nilsson, P., Jansson, S., and Taylor, G. (2006) The genetics and genomics of the drought response in Populus. *Plant Journal* **48**, 321–341.
60. An, C.F., Saha, S., Jenkins, J.N., Scheffler, B.E., Wilkins, T.A., and Stelly, D.M. (2007) Transcriptome profiling, sequence characterization, and SNP-based chromosomal assignment of the EXPANSIN genes in cotton. *Molecular Genetics and Genomics* **278**, 539–553.

61. Venu, R.C., Jia, Y., Gowda, M., Jia, M.H., Jantasuriyarat, C., Stahlberg, E., Li, H., Rhineheart, A., Boddhireddy, P., Singh, P., Rutger, N., Kudrna, D., Wing, R., Nelson, J.C., and Wang, G.L. (2007) RL-SAGE and microarray analysis of the rice transcriptome after Rhizoctonia solani infection. *Molecular Genetics and Genomics* **278**, 421–431.
62. Kiani, S.P., Grieu, P., Maury, P., Hewezi, T., Gentzbittel, L., and Sarrafi, A. (2007) Genetic variability for physiological traits under drought conditions and differential expression of water stress-associated genes in sunflower (*Helianthus annuus* L.). *Theoretical and Applied Genetics* **114**, 193–207.
63. Shi, C., Uzarowska, A., Ouzunova, M., Landbeck, M., Wenzel, G., and Lubberstedt, T. (2007) Identification of candidate genes associated with cell wall digestibility and eQTL (expression quantitative trait loci) analysis in a Flint x Flint maize recombinant inbred line population. *BMC Genomics* **8**, 22.
64. Aoki, K., Ogata, Y., and Shibata, D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology* **48**, 381–390.
65. Hirai, M.Y., Klein, M., Fujikawa, Y., Yano, M., Goodenowe, D.B., Yamazaki, Y., Kanaya, S., Nakamura, Y., Kitayama, M., Suzuki, H., Sakurai, N., Shibata, D., Tokuhisa, J., Reichelt, M., Gershenzon, J., Papenbrock, J., and Saito, K. (2005) Elucidation of gene-to-gene and metabolite-to-gene networks in Arabidopsis by integration of metabolomics and transcriptomics. *Journal of Biological Chemistry* **280**, 25590–25595.
66. Ma, S.S., Gong, Q.Q., and Bohnert, H.J. (2007) An Arabidopsis gene network based on the graphical Gaussian model. *Genome Research* **17**, 1614–1625.
67. Urbanczyk-Wochniak, E. and Sumner, L.W. (2007) MedicCyc: a biochemical pathway database for Medicago truncatula. *Bioinformatics* **23**, 1418–1423.
68. Caspi, R., Foerster, H., Fulcher, C.A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S.Y., Tissier, C., Zhang, P.F., and Karp, P.D. (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Research* **34**, D511–D516.
69. Li, S.M., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D.J., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q.R., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H.Y., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., van den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E., and Vidal, M. (2004) A map of the interactome network of the metazoan C-elegans. *Science* **303**, 540–543.
70. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403.
71. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 4569–4574.
72. Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H., and Geisler, M. (2007) A predicted interactome for Arabidopsis. *Plant Physiology* **145**, 317–329.
73. Wei, N., Chamovitz, D.A., and Deng, X.W. (1994) Arabidopsis Cop9 is a component of a novel signaling complex mediating light control of development. *Cell* **78**, 117–124.
74. Vision, T.J., Brown, D.G., and Tanksley, S.D. (2000) The origins of genomic duplications in Arabidopsis. *Science* **290**, 2114–2117.
75. Blanc, G., Hokamp, K., and Wolfe, K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Research* **13**, 137–144.
76. Wolfe, K.H. and Shields, D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713.
77. Blanc, G. and Wolfe, K.H. (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16**, 1667–1678.
78. Ober, D. (2005) Seeing double: gene duplication and diversification in plant secondary metabolism. *Trends in Plant Science* **10**, 444–449.
79. Meyer, R.C., Steinfath, M., Lisec, J., Becher, M., Witucka-Wall, H., Törjék, O., Fiehn, O., Eckardt, A., Willmitzer, L., Selbig, J., and Altmann, T. (2007) The

metabolic signature related to high plant growth rate in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 4759–4764.

80. Basten, C.J., Weir, B.S., and Zeng, Z.-B. (1999) QTL Cartographer, Version 1.13, Department of Statistics, North Carolina State University, Raleigh, N.C.
81. Wang, S., Basten, C.J., and Zeng, Z.-B. (2006) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.
82. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264.
83. Churchill, G.A. and Doerge, R.W. (1994) Empirical threshold values For quantitative trait mapping. *Genetics* **138**, 963–971.
84. Bogdan, M. and Doerge, R.W. (2005) Biased estimators of quantitative trait locus heritability and location in interval mapping. *Heredity* **95**, 476–484.
85. Doerge, R.W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* **3**, 43–52.
86. Doerge, R.W. and Churchill, G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.
87. Gilbert, H. and Le Roy, P. (2003) Comparison of three multitrait methods for QTL detection. *Genetics Selection Evolution* **35**, 281–304.
88. Knott, S.A. and Haley, C.S. (2000) Multitrait least squares for quantitative trait loci detection. *Genetics* **156**, 899–911.
89. Ronin, Y.I., Kirzhner, V.M., and Korol, A.B. (1995) Linkage between loci of quantitative traits and marker loci – multi-trait analysis with a single marker. *Theoretical and Applied Genetics* **90**, 776–786.
90. Chen, M. and Kendziorski, C. (2007) A statistical framework for expression quantitative trait loci mapping. *Genetics* **177**, 761–771.
91. Ball, R.D. (2007) Quantifying evidence for candidate gene polymorphisms: Bayesian analysis combining sequence-specific and quantitative trait loci colocation information. *Genetics* **177**, 2399–2416.
92. Hoti, F. and Sillanpaa, M.J. (2006) Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity* **97**, 4–18.
93. Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T.K., Flowers, M.T., Schueler, K.L., Manly, K.F., Williams, R.W., Kendziorski, C., and Attie, A.D. (2006) Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **2**, 51–61.