DISSECTION OF COMPLEX GENE EXPRESSION USING THE COMBINED ANALYSIS OF PLEIOTROPY AND EPISTASIS

VIVEK M. PHILIP, ANNA L. TYLER, and GREGORY W. CARTER*

The Jackson Laboratory,
Bar Harbor, ME, 04609, USA
**E-mail: greg.carter@jax.org*

Global transcript expression experiments are commonly used to investigate the biological processes that underlie complex traits. These studies can exhibit complex patterns of pleiotropy when *trans*-acting genetic factors influence overlapping sets of multiple transcripts. Dissecting these patterns into biological modules with distinct genetic etiology can provide models of how genetic variants affect specific processes that contribute to a trait. Here we identify transcript modules associated with pleiotropic genetic factors and apply genetic interaction analysis to disentangle the regulatory architecture in a mouse intercross study of kidney function. The method, called the combined analysis of pleiotropy and epistasis (CAPE), has been previously used to model genetic networks for multiple physiological traits. It simultaneously models multiple phenotypes to identify direct genetic influences as well as influences mediated through genetic interactions. We first identified candidate *trans* expression quantitative trait loci (eQTL) and the transcripts potentially affected. We then clustered the transcripts into modules of co-expressed genes, from which we compute summary module phenotypes. Finally, we applied CAPE to map the network of interacting module QTL (modQTL) affecting the gene modules. The resulting network mapped how multiple modQTL both directly and indirectly affect modules associated with metabolic functions and biosynthetic processes. This work demonstrates how the integration of pleiotropic signals in gene expression data can be used to infer a complex hypothesis of how multiple loci interact to co-regulate transcription programs, thereby providing additional constraints to prioritize validation experiments.

Keywords: pleiotropy, genetic interaction, genetic network.

1. Introduction

The widespread adoption of genomic technologies has greatly increased the power and scope of genetic studies. One especially fruitful approach to understanding how genetic variation affects biological processes is the study of the genetics of gene expression.[1–5] In these studies, transcript levels are treated as panels of thousands of phenotypes that quantify the cellular composition and gene expression of a tissue sample that is related to a physiological phenotype such as disease. These data are commonly analyzed to identify expression quantitative trait loci (eQTL), which are specific chromosomal regions that associate with the expression level of a given transcript.

Associated eQTL are generally classified as local, *cis*-acting variants that affect the expression of a gene located near the associated variant, or remote, *trans*-acting variants that affect the expression of a gene located at a distance (*i.e.* outside of linkage disequilibrium (LD) or on another chromosome). The more common *cis* associations have the straightforward biological interpretation of a sequence variant directly affecting the self transcript production, stability, or splicing. However, *trans* associations are often more difficult to interpret. The structure of gene regulatory networks suggests that these *trans* associations are caused by transcription factors or other proteins that bind and regulate DNA or RNA. The co-regulatory structures

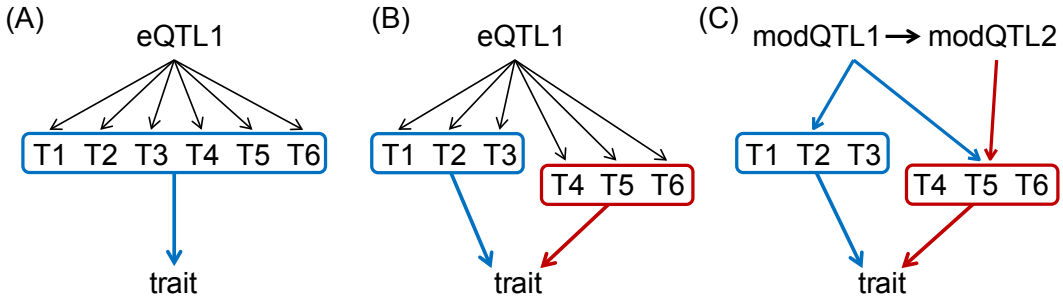

Fig. 1. Hypothetical regulatory architecture of transcripts ($T1, ..., T6$) that serve as an endophenotype for an organism-level trait. (A) Simple model in which all transcripts are associated with *trans*-acting *eQTL1* and part of a single underlying biological process affecting the trait. (B) Model with transcripts grouped into two modules that combine to affect the trait. Models (A) and (B) are indistinguishable using single-locus association. (C) Model obtained with co-expression clustering and CAPE analysis, in which the eQTL has been replaced by two multiple module QTL (modQTL). The genetic effects now map to the two modules distinctly, and the modQTL are linked by a directional influence mapping feed-forward regulation from *modQTL1* to the red module via *modQTL2*.

of these networks, in which proteins regulate multiple transcripts in complex hierarchies,[6] suggest that a genetic variation in one regulatory gene could have significant effects on the expression of multiple target transcripts. This would generate extensive pleiotropy as many redundantly regulated transcripts would associate with the variant. While this is pleiotropy in the sense that one genetic variant is influencing multiple traits, it is somewhat trivial in that the multiple traits are redundant outputs of the same regulatory module. This effect can be efficiently modeled by first finding modules of co-expressed transcripts that map to the common trans-acting module QTL (modQTL). Pleiotropy between modQTL, in which a single variant is associated with multiple distinct gene modules, is more informative in the sense of a single variant affecting multiple regulatory programs in a more complex genetic architecture (Figure 1). Distinguishing between trivial and informative pleiotropy can be difficult for complex regulatory networks in which multiple regulatory variants combine to affect hundreds of transcript outputs.

In this paper, we address this problem by modeling interacting *trans* associations for modules of co-expressed genes. We use kidney transcript data from a panel of F2 mouse intercross progeny to dissect the genetic regulation of multiple biological processes that affect overall kidney function in these genetically diverse mouse models. We use co-expression analysis to identify gene modules with correlated expression and common function and derive summary endophenotypes that describe transcriptional states. We next use a combined analysis of pleiotropy and epistasis (CAPE[7]) to simultaneously assess patterns of pleiotropy and statistical interactions between *trans* modQTL, in order to infer the variant-to-variant ordering of regulatory influences on the multiple processes. This approach improves the interpretation of genetic interactions in terms of directed QTL-to-QTL influences that map how a given locus suppresses or enhances the effects of a second locus. By integrating evidence of epistasis across multiple phenotypes, the CAPE method can improve power to detect modQTL interactions and assign directionality to the relationship. Furthermore, CAPE inherently parses

QTL-to-phenotype associations into direct effects and effects modified through genetic interactions, thereby separating the target transcripts into subsets that are influenced by distinct combinations of modQTL. In the case of transcript data, the result is a model of how multiple modQTL affect one another and, in turn, the regulation of multiple modules of co-expressed genes (Figure 1C). The resulting network model provides a clearer dissection of the nature of the observed pleiotropy and generates more specific hypotheses of variant activity and action.

2. Methods

We followed a multi-step strategy to systematically identify and model multiple gene modules that underlie kidney health and disease. The procedure is outlined in Figure 2, and consisted of three main steps: a preliminary eQTL analysis to identify transcripts affected by one or more genetic factors; clustering of the affected transcripts into co-expressed gene modules; and a network analysis to map how the gene modules are regulated by multiple interacting genetic loci. We began with a study of gene expression related to kidney function in a mouse intercross.[8] An F2 intercross population was derived from the kidney damage-susceptible SM/J inbred strain and the nonsusceptible MRL/MpJ inbred strain. Male SM/J mice exhibit kidney dysfunction, as measured by an increase in urinary albumin-to-creatine ratio (ACR). To identify causal genetic loci, ACR was measured in 173 male F2 progeny. Significant QTL were mapped on chromosomes (Chrs) 1, 4, and 15, with an additional suggestive QTL on Chr 17.[8] This established ACR as a trait affected by multiple QTL that vary between the SM/J and MRL/MpJ lines.

2.1. *Data*

To identify the biological pathways and processes underlying the ACR results, mRNA was collected from whole kidneys of the 173 F2 animals. Data generation and processing is de-

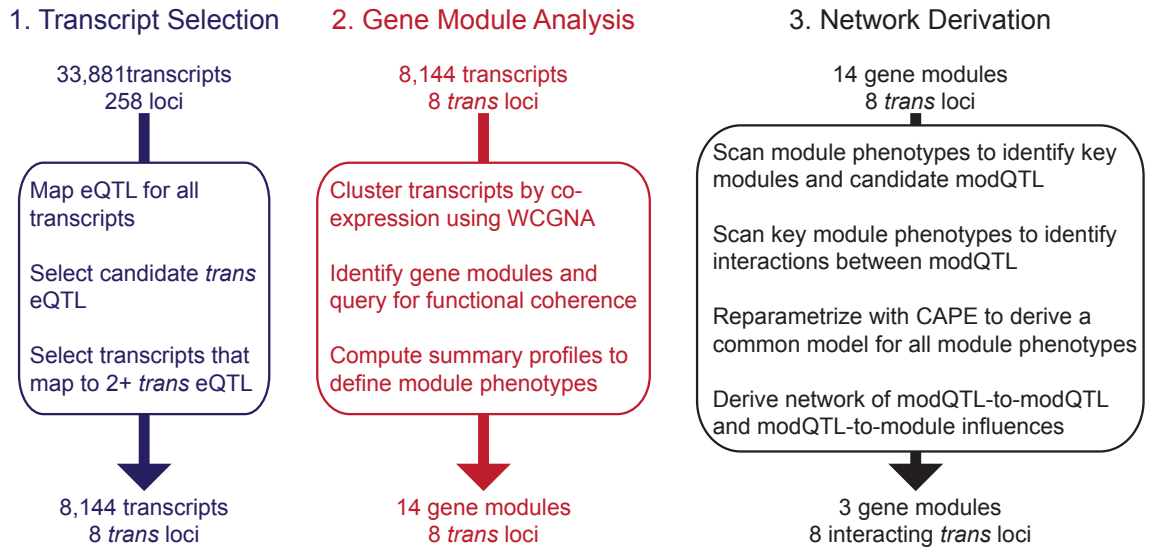

Fig. 2. Overview of analytical strategy.

scribed in depth in the initial publication[8] and will be summarized here. All mice were genotyped using an array that contained 258 polymorphisms that were informative between the MRL/MpJ and SM/J strains. RNA samples were labeled and hybridized to the mouse gene 1.0 ST microarray (Affymetrix, Santa Clara, California). Microarray data were imported in R (`http://www.r-project.org`) and processed using the *affy* package from Bioconductor (`http://bioconductor.org`). Normalization of the data was performed using robust multi-array average without any background subtraction. In total, 33,881 probe sets were considered.[8] Data were downloaded from the QTL Archive (`http://www.qtlarchive.org`).

2.2. *Transcript Selection*

Following the initial study, we performed eQTL scans using R/qtl[9] to test the association of every transcript with every marker. Transcript expression data were subjected to a Van der Waerden transformation[10] prior to eQTL mapping. Pseudo-markers were generated at 2 cM spacing for each chromosome and Haley-Knott regression was performed genome-wide for each transcript. To identify suggestive eQTL ($P < 0.63$), we followed the originally-reported LOD thresholds of 2.23 and 1.44 for autosomes and the X chromosome, respectively,[8] This comprised a set of candidate transcripts with at least one suggestive association, each potentially regulated by one or more genetic loci. Because we were interested in analyzing overlapping patterns of pleiotropy, we further reduced this list to a set of transcripts that were associated with at least two distinct suggestive eQTL.

2.3. *Co-Expression Modules*

Since the co-regulation of multiple genes is expected to be manifest as co-expression in array data, we next performed weighted gene correlation network analysis (WCGNA)[11] to identify gene modules. WGCNA has been widely and successfully used to parse sets of transcripts into co-expressed modules, particularly in genetic mapping populations.[12] A comprehensive list of tutorials on WGCNA can be found at `http://www.genetics.ucla.edu/horvath/CoexpressionNetwork`. WGCNA generates an adjacency matrix based on the underlying absolute values of Pearson correlations among all pairs of transcripts raised to a user-defined power β. Here, the β parameter was set to 6 in order to generate the scale-free topology criterion as defined by Zhang and Horvath.[13] For each module, we separately obtained the first principal component (termed "eigengenes" in WCGNA) to represent the summary expression pattern for that module. We hereafter refer to these quantitative expression vectors as *module phenotypes* since they represent composite phenotypes (and the term eigengene may be confused with our distinct concept of an eigentrait in Section 2.4). Modules were queried for coherent functions using the R package GOstats.[14] Both Gene Ontology annotations[15] and KEGG pathways[16] were queried for functional overrepresentation. GO enrichment significance scores were corrected for multiple tests using the decorrelation of GO graph structure.[17]

2.4. *CAPE Network Derivation*

The combined analysis of pleiotropy and epistasis (CAPE) is an approach to modeling two or more phenotypes across a population harboring genetic variation. Detailed explanations

of the method have been published elsewhere[7] and will be briefly summarized here. CAPE is designed to translate data from genetic studies with multiple traits into an integrated model that accounts for variance across all phenotypes. As input, the method requires two or more quantitative phenotypes and a matrix of genotype values at markers across the genome. Variants can be engineered mutations such as gene knockouts or amplifications, or natural variants that are commonly used to map QTL. In this work, the variants will be the modQTL associated with module phenotypes. The model of variants affecting phenotypes is obtained by multivariate linear regression followed by a novel reparametrization of the results.[7] For a given pair of genetic variants, this reparametrization recasts the set of interaction coefficients (one for each trait) in terms of two coefficients that describe how each variant suppresses or enhances the effects of the other. This procedure translates trait-specific interaction terms into trait-independent, directed edges between the two variants, providing a common model of gene action that consistently fits all traits. These quantitative, variant-to-variant influences can be readily interpreted as genetic suppression or enhancement. When combined with the variant-to-phenotype edges, the final output is a directed network of both direct and indirect effect of variants on multiple traits. CAPE is available as an R package (`http://cran.r-project.org/web/packages/cape`), which was used in our analysis.[18]

We first identified a subset of modules suitable for CAPE. Each module phenotype was first scanned for modQTL associations,[12] with candidate loci identified using a suggestive threshold ($P < 0.63$) based on a null distribution generated from 100 permutations. Genetic markers were used as loci for regression, with homozygous MRL/MpJ markers coded as 0, heterozygous markers as 0.5, and homozygous SM/J markers as 1. CAPE modules were then selected by identifying module phenotypes with a combination of candidate modQTL that included both shared and unique associations, and exhibited some degree of correlation (Figure 3). These criteria are essential to the CAPE method, given that it requires biologically related phenotypes (e.g. all modules related to kidney function) that also exhibit unique signals from which to draw functional distinctions.

The selected module phenotypes and sample genotypes were then used as input for the R implementation of CAPE.[18] As a first step in the analysis, CAPE decomposes all phenotypes into *eigentraits* using singular value decomposition (SVD). This procedure reorganizes the phenotypes into common and distinct signals that are expected to map to common and distinct genetic loci. Each eigentrait is scanned for its own QTL, and a user-defined number of eigentraits are selected for further analysis. This allows one to filter non-genetic signals in the data and maximizes efficiency in the analysis. In this case, the eigentraits were linear combinations of the module phenotypes. A suggestive threshold was used ($P < 0.63$, determined via 200 permutation tests) and the union of all suggestive markers comprised the set of markers to undergo pair-wise association tests.

Pair-wise regression models were derived and reparametrized following the CAPE method.[7,18] In all except specified instances, default CAPE parameters were used. To avoid effects due to LD, we omitted marker pairs with genotypes showing Pearson correlation above 0.6. Effects from QTL to eigentraits are then recomposed to map modQTL-to-phenotype influences. We performed 100,000 permutations to generate empirical P values for each parameter

in the model, and then performed a false discovery rate (FDR) correction[19] to compute q values. For the final network model, we used a significance cutoff of $q < 0.05$ on both variant-to-variant and variant-to-phenotype influences.

3. Results

3.1. *Selected Transcripts*

We performed eQTL scans on 33,881 probe transcripts across 254 independent genetic markers. This procedure yielded 53,134 suggestive associations for 26,097 transcripts, including both *cis*- and *trans*-acting loci (Table S1). In order to restrict our analysis to pleiotropic loci, we identified the number of *trans* eQTL per chromosome. This varied from 5977 transcripts associated with Chr 1 to 1101 transcripts associated with Chr 10. Since we were particularly interested in the loci associated with the ACR phenotype we concentrated our analysis on the top eight chromosomes, which comprised 60% of the associations. As in the previous study,[8] Chrs 1, 4, 15, and 17 were among the top *trans* chromosomes. With our weak significance cutoff, we also found four additional candidate chromosomes (Chrs 2, 6, 7, and 11). These patterns suggested widespread co-regulation of hundreds of genes by a few genetic loci. To explore potential pleiotropic effects, we selected the 8,144 transcripts associated with two or more of these chromosomes in order to analyze how these loci affect transcripts both jointly and distinctly. This provided us a large number of overlapping endophenotypes while maintaining focus on a tractable number of biological processes.

3.2. *Gene Modules Analysis*

WCGNA was performed on the 8,144 transcripts identified in the previous step. We obtained 14 distinct modules, which were automatically assigned color identifiers by the software. The number of genes per module ranged from 25 to 1299 (Table S2). We queried each module for functional overrepresentation and found GO and KEGG associations for nearly all modules at a significance of $P < 10^{-4}$ (Table S3). We observed a diversity of processes across modules, which included small organic molecule metabolism, macromolecule metabolism, immune processes, and structural development. However, the largest modules were concentrated in metabolic and transcriptional processes. These module results generally matched the KEGG pathways identified in the original analysis of the data,[8] which were obtained through a different analytical procedure.

We next assessed correlations between module phenotypes. Since the CAPE method relies on moderately correlated data, we sought pairs of modules with similar, but not redundant, profiles. The module phenotypes exhibited absolute Pearson correlations ranging from 0.001 to 0.8 (Figure S1).

3.3. *Single-Locus Genome Scans*

We performed single-locus scans on the 14 module phenotypes to assess common associations and pleiotropic loci (Figure S2). As expected, most (82%) of the suggestive ($P < 0.63$) modQTL were located on the eight chromosomes that were pre-selected for associations with individual

Table 1. Summary of gene modules used in CAPE analysis.

Module	Genes	Suggestive modQTL	Representative GO Function	Representative KEGG Pathway
blue	969	2,4,7,9,11,15	oxoacid metabolic process (6×10^{-13})	fatty acid metabolism (8×10^{-8})
grey	1299	1,4,9,11,17	oxidation-reduction process (1×10^{-4})	oxidative phosphorylation (3×10^{-3})
turquoise	1228	1,17	translational initiation (8×10^{-5})	cell cycle (5×10^{-5})

transcripts. Chrs 1, 4, 11, and 17 had the greatest number of associations, suggesting a strong biological overlap with the ACR phenotype. The number of suggestive modQTL ranged from one locus (magenta module) to eight loci (brown module).

3.4. *Pair-Wise Scans and Interaction Network*

We next performed two-locus interaction scans and CAPE reparametrization to derive a network of pleiotropic effects on gene modules. We selected modules with partial pleiotropy and correlation for further analysis, since modules with simpler genetic associations would not require genetic dissection with CAPE. We selected the three largest modules for CAPE analysis, summarized in Table 1. These modules met the criteria of exhibiting moderate correlations (Figure 3A) and had suggestive associations with one or more pleiotropic modQTL (Table 1). They comprised 78% of the annotated genes in all modules together, thereby accounting for the vast majority of expression variance in the data set. All modules had multiple significantly enriched annotations (Table S3). The blue module contained specific acid metabolic processes and transport genes. The grey module was concentrated in metabolic processes, programmed cell death, and catabolism. Although WCGNA assigns the grey color to transcripts that do not belong to any other module based on correlated expression, and therefore might not be co-expressed in some cases, our pre-selection of transcripts based on eQTL associations generated a grey module phenotype with sufficient common signal to generate modQTLs and a gene set with common functional annotations. Genes in the turquoise module were associated with gene expression and RNA metabolism, and other cell cycle processes. While it would have been feasible to include additional modules in the analysis, many of the modules had relative weak associations and poor correlation with other modules (Figures S2 and S3), suggesting CAPE analysis would provide little additional information. Furthermore, the addition of phenotypes associated with non-pleiotropic modQTL will likely have distinct genetic etiology, and thus can weaken significance of CAPE results by adding genetically independent variance.[7]

We performed SVD on the three selected module phenotypes to obtain three eigentraits (Section 2.4), which represent linear combinations of the three module phenotypes (Figure 3B). We scanned each eigentrait for QTL associations and found that most of our candidate modQTL were associated with the first and/or second eigentrait, suggesting that the genetically-driven variance is captured by these two composite phenotypes. Additionally, the first two eigentraits are of comparable weight and together account for 87% of the global variance. We therefore used these two eigentraits in our analysis, which is the default for CAPE.[18] A total of 54 candidate markers were identified by pooling those markers with suggestive ef-

fects, leading to 1303 marker pairs tested after removing pairs in LD. After performing the interaction analysis (Section 2.4) we transformed the eigentraits back to the original module phenotypes. This transformation does not change modQTL-to-modQTL influences.[7] An adjacency matrix of significant results for all marker pairs is shown in Figure 4. This non-symmetric matrix maps directed edges from each source marker to each target marker or target phenotype (rightmost columns).

A summary interaction network is shown in Figure 5. To avoid redundant interactions and nodes due to adjacent markers within a given modQTL, each modQTL-containing chromosome is represented by a single node. Although the pleiotropic modQTL and genetic interactions consistently map to the same regions on the indicated chromosomes (Figures 4 and S4), the relatively large intervals preclude reliable identification of candidate genes and therefore we simply represent the modQTL with chromosome names. Network nodes represent the effect of the SM/J allele at each modQTL. Thus the modQTL-to-phenotype edges represent the effects of a SM/J allele at the modQTL, and negative modQTL-to-modQTL interaction represents the presence of a SM/J variant at one locus suppressing another SM/J variant at a second locus. All interactions between modQTL were negative, consistent with the vast majority of findings in intercross experiments.[20] This may be due to functional redundancy between modQTL, suggesting that variants within pathways underlie the interactions.[21–23] In sum, we detected six significant modQTL-to-modQTL interactions between chromosome pairs and 15 significant modQTL-to-phenotype interactions.

Our interaction network most prominently detected interactions between Chr 1, 4, and 15. These correspond to QTL previously associated with ACR and kidney health,[8] and also comprised the most significant influences in our analysis. The co-suppression observed between Chr 1 and Chr 15 and between Chr 4 and Chr 15 suggest candidate genes of similar function underlie these modQTL. This genetic co-suppression was frequently observed for knockdowns of genes in the same pathways in a previous study of fly cell proliferation,[23] and is a consequence of highly redundant effects when SM/J alleles are present at both loci. We also note

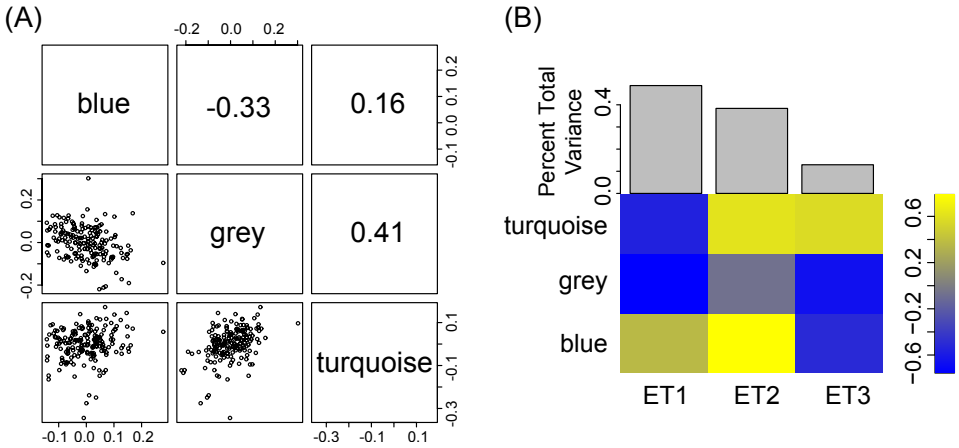

Fig. 3. Correlation structure of the three module phenotypes selected for CAPE analysis. (A) Pearson correlations and scatter plots of each pair of module phenotypes. (B) The three module phenotypes decomposed into orthogonal eigentraits, showing phenotype composition and global variance fraction for each eigentrait.

that upon conditioning on interaction effects, these modQTL are pleiotropic, with each significantly influencing both the blue and grey modules. Interestingly, the turquoise module is primarily influenced by a network of interactions between modQTL on Chrs 7, 9, and 17. The Chr 9 modQTL suppression of the Chr 17 modQTL is an example of how the CAPE method can identify indirect effects between loci, in that the Chr 9 SM/J-derived effects on the turquoise and grey modules are mediated via the presence of an SM/J allele at the Chr 17 locus. The hypothesis is that Chr 9 allele indirectly acts to suppress the Chr 17 allele, and this conditional dependence on the Chr 17 modQTL renders the Chr 9 modQTL only marginally significant when considered in isolation (Figure S2).

4. Discussion and Conclusions

The CAPE method has been developed to map networks of how multiple genetic variants interact to affect multiple phenotypes, thereby identifying shared and distinct genetic etiology of complex traits. Here, we have applied this approach to address the regulation of kidney gene expression in an inbred mouse intercross. This required a focused approach to identifying patterns of co-expressed genes, followed by an application of the CAPE algorithm that separated the co-regulation of those genes in a network of causal genetic loci.

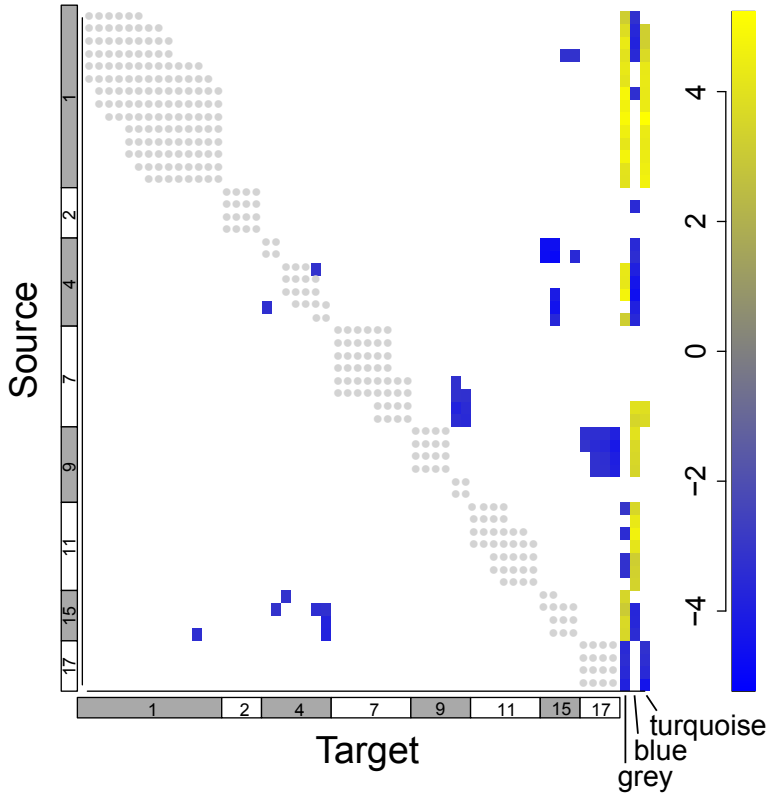

Fig. 4. Adjacency matrix of interactions derived with CAPE (FDR $q < 0.05$). Markers are designated as sources or targets of directed interactions, and marker-to-phenotype influences are in the rightmost columns. Only candidate markers are shown with chromosome locations labeled, and grey dots marking pairs that were not tested due to LD. Standardized effects (effect divided by standard error) are shown to reflect significance.

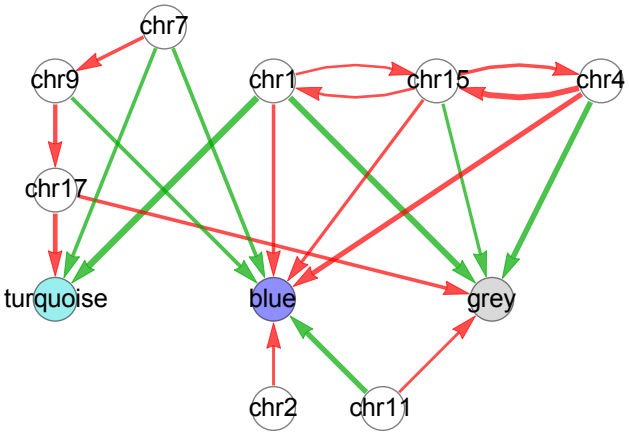

Fig. 5. Summary interaction network derived with R/cape, with interacting modQTL labeled by chromosome location on white nodes and gene modules on nodes colored by WCGNA assignments. Width of positive (green) and negative (red) edges represent significance in terms of standardized effect size.

4.1. *Co-Expressed Gene Modules as Complex Pleiotropy*

By clustering transcripts into modules, we efficiently identified common *trans*-acting modQTL that regulate multiple co-expressed genes.[12] Although this strategy will not detect the majority of *cis* eQTL, which can be readily detected through direct associations of each individual transcript, it quickly identifies *trans* modQTL that exhibit pleiotropy by affecting multiple gene modules. Furthermore, the coherent expression patterns within each gene module were used as summary traits representing the activity levels of multiple biological processes. This allowed the use of the CAPE approach to map an interacting network of causal gene variants, providing an enhanced view of how multiple genetic variants commonly and differentially affected multiple gene expression patterns in the kidneys of genetically diverse mice.

4.2. *How Genetic Interactions Modify Pleiotropic Effects*

By simultaneously analyzing genetic interactions across multiple module phenotypes, we were able to identify cases in which pleiotropic modQTL are directly associated with a module and cases in which the modQTL was indirectly affecting a module via interaction with a second modQTL. This separation provides an improved genetic model of how the modQTL might affect overall kidney health through two or more processes. The interaction cascade observed for Chrs 7, 9, and 17 suggests a series of co-dependent effects from the SM/J variant at these loci (Figure 5). When all three modQTL are inherited from SM/J, the model implies that the Chr 9 and Chr 17 modQTL are suppressed and ineffective, leading to an overall Chr 7 positive effect on the expression of the turquoise and blue modules. However, changing this scenario with an MRL/MpJ allele at the Chr 9 modQTL implies the Chr 17 modQTL counteracts the effect of the Chr 7 modQTL on the turquoise module, leaving the primary effect of the Chr 7 modQTL on the blue module only and therefore diminishing its pleiotropic effect. Examples of epistasis-dependent pleiotropy are a key element of hypotheses generated from CAPE, and their inference requires a systematic integration of both epistasis and pleiotropy in a single model of genetic effects.

4.3. *Overlapping Patterns of Pleiotropy to Model Complex Traits*

At the core of the CAPE method is the use of multiple QTL with partially overlapping patterns of pleiotropy over a panel of complex traits. The information coded in these patterns is used to constrain models of genetic interactions and, at the same time, map pleiotropic effects as either independent or dependent on other QTL. Thus the appropriate choice of phenotypes in analysis is essential. The most direct method is to perform single-locus scans for all phenotypes to identify shared QTL regions, with the assumption that the causal variant is common to all phenotypes. However, the sensitivity of QTL significance on limited sample numbers can rarely preclude that a QTL that falls slightly below a significance threshold is in fact causal.

In this work, we have surmounted this problem by allowing highly permissive significance thresholds for pre-selection of potentially interacting loci. Nevertheless, some of our modules exhibited few suggestive modQTL or unique loci, such as the distal Chr 6 locus that dominates the magenta module scan (Figure S2). An alternative, related approach is to select phenotypes with moderately correlated values across all samples, such as Pearson correlations of 0.3-0.8. Excessive correlation among phenotypes generates redundant genetic associations, which are ineffective for the CAPE approach, while a lack of sufficient correlation between phenotypes introduces too many conflicting signals to arrive at a common genetic model. Finally, we note that an excess of complex phenotypes can reduce the ability of CAPE to find a common genetic model. While the number of phenotypes that can be co-analyzed is theoretically unlimited, the core of the analysis is based on a dimensional reduction of a series of epistasis coefficients (one for each phenotype) to two influence parameters describing how a pair of QTL influence each other in either direction.[7] While the method maximizes the amount of phenotype information in two degrees of freedom independently for each locus pair, conflicting data can weaken the interaction signal. Indeed, in an earlier study of global transcript data that directly modeled principal components instead of more focused co-expression modules, it was found that simultaneously modeling more than three components diluted the power to detect interactions.[7] This finding applies whether the additional components are interpreted as experimental noise or additional biological signal.

4.4. *Potential Extensions and Validation*

The genetic models obtained by CAPE are formulated in terms of inferred influences that quantify the associated effects of variants on (1) all phenotypes; and (2) the effective weight of other variants on the phenotypes. The resulting networks structure provides a hypothesis of regulatory architecture, but does not provide any direct evidence of molecular binding. When available, the network can be used as a template for the integration of complementary molecular interaction data, with candidate regulatory interactions limited by the sign and direction of each variant-to-variant influence.[24] In systems lacking existing molecular interaction data, the inferred networks can serve to direct experimental validation to specific combinations of loci. For example, the binding sites of a candidate transcription factor may be predicted to be modified by the presence of a second *trans*-acting variant. This could be directly assayed with chromatin immunoprecipitation experiments performed with and without the second variant. This framework can guide follow-up investigations by providing additional constraints to

prioritize candidate regulators.

Supplementary Material

Tables S1-S3 and Figures S1 and S2 are located at `http://carterdev.jax.org/psb2014`.

Acknowledgments

We thank Ron Korstanje for assistance with the data. This work was supported by NIGMS grants P50 GM076468 and K25 GM079404, and NCI grant CA034196. The content does not necessarily represent the official views of NIGMS, NCI, or NIH.

References

1. R. Jansen and J.-P. Nap, *Trends Genet* **17**, 388 (2001).
2. R. Brem, G. Yvert, R. Clinton and L. Kruglyak, *Science* **296**, 752 (2002).
3. Schadt, E.E., S. Monks, T. Drake, A. Lusis, N. Che, V. Colinayo, T. Ruff, S. Milligan, J. Lamb, G. Cavet, P. Linsley, M. Mao, R. Stoughton and S. Friend, *Nature* **422**, 297 (2003).
4. E. Chesler, L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. Hsu, J. Mountz, N. Baldwin, M. Langston, D. Threadgill, K. Manly and R. Williams, *Nat Genet* **37**, 233 (2005).
5. J. Keurentjes, J. Fu, I. Terpstra, J. Garcia, G. van den Ackerveken, L. Snoek, A. Peeters, D. Vreugdenhil, M. Koornneef and R. Jansen, *Proc Natl Acad Sci U S A.* **104**, 1708 (2007).
6. H. Yu and M. Gerstein, *Proc Natl Acad Sci U S A.* **103**, 14724 (2006).
7. G. W. Carter, M. Hays, A. Sherman and T. Galitski, *PLoS Genet* **8**, p. e1003010 (2012).
8. R. S. Hageman, M. S. Leduc, C. R. Caputo, S.-W. Tsaih, G. A. Churchill and R. Korstanje, *J Am Soc Nephrol* **22**, 73 (2011).
9. K. W. Broman, H. Wu, S. Sen and G. A. Churchill, *Bioinformatics* **19**, 289 (2003).
10. B. L. van der Warden, *Koninklijke Nederlandse Akademie van Wetenschappen* **55**, 453 (1952).
11. P. Langfelder and S. Horvath, *BMC Bioinformatics* **9**, 559 (2008).
12. A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis and S. Horvath, *PLoS Genet* **2**, p. e130 (2006).
13. B. Zhang and S. Horvath, *Stat Appl Genet Mol Biol* **4**, p. 17 (2005).
14. S. Falcon and R. Gentleman, *Bioinformatics* **23**, 257 (2006).
15. M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin and G. Sherlock, *Nat Genet* **25**, 25 (2000).
16. M. Kanehisa and S. Goto, *Nucleic Acids Res* **28**, 27 (2000).
17. A. Alexa, J. Rahnenführer and T. Lengauer, *Bioinformatics* **22**, 1600 (2005).
18. A. L. Tyler, W. Lu, J. J. Hendrick, V. M. Philip and G. W. Carter, *PLoS Comp Bio* (2013), in press.
19. Y. Benjamini and Y. Hochberg, *J. R. Stat. Soc. Ser. B* **57**, 289 (1995).
20. W. Huang, S. Richards, M. A. Carbone, D. Zhu, R. R. H. Anholt, L. D. Julien F. Ayroles, K. W. Jordan, F. Lawrence, M. M. Magwire, K. B. Crystal B. Warner, Y. Han, M. Javaid, J. Jayaseelan, S. N. Jhangiani, D. Muzny, L. P. Fiona Ongeri, Y.-Q. Wu, Y. Zhang, X. Zou, E. A. Stone, R. A. Gibbs and T. F. C. Mackay, *Proc Natl Acad Sci U S A.* **109**, 15553 (2012).
21. L. Avery and S. Wasserman, *Trends Genet* **8**, 312 (1992).
22. D. Segré, A. Deluna, G. Church and R. Kishony, *Nat Genet* **37**, 77 (2005).
23. G. W. Carter, *G3* **3**, 807 (2013).
24. G. W. Carter, S. Prinz, C. Neou, J. P. Shelby, B. Marzolf, V. Thorsson and T. Galitski, *Mol Syst Biol* **3**, p. 96 (2007).