Physiol Genomics 25: 355–363, 2006.
First published March 22, 2006; doi:10.1152/physiolgenomics.00314.2004.

Invited Review

Microarray analysis of gene expression: considerations in data mining and statistical treatment

Joseph S. Verducci,[1,2,3] Vincent F. Melfi,[3,4] Shili Lin,[2,3] Zailong Wang,[3,5] Sashwati Roy,[1] and Chandan K. Sen[1]

[1]*Laboratory of Molecular Medicine and DNA Microarray Facility, Davis Heart and Lung Research Institute, Department of Surgery,* [2]*Department of Statistics, and* [3]*Mathematical Biosciences Institute, The Ohio State University, Columbus, Ohio;* [4]*Department of Statistics, Michigan State University, East Lansing, Michigan; and* [5]*Novartis Pharmaceuticals Corporation, East Hanover, New Jersey*

Submitted 30 December 2004; accepted in final form 23 February 2006

Verducci, Joseph S., Vincent F. Melfi, Shili Lin, Zailong Wang, Sashwati Roy, and Chandan K. Sen. Microarray analysis of gene expression: considerations in data mining and statistical treatment. *Physiol Genomics* 25: 355–363, 2006. First published March 22, 2006; doi:10.1152/physiolgenomics.00314.2004.—DNA microarray represents a powerful tool in biomedical discoveries. Harnessing the potential of this technology depends on the development and appropriate use of data mining and statistical tools. Significant current advances have made microarray data mining more versatile. Researchers are no longer limited to default choices that generate suboptimal results. Conflicting results in repeated experiments can be resolved through attention to the statistical details. In the current dynamic environment, there are many choices and potential pitfalls for researchers who intend to incorporate microarrays as a research tool. This review is intended to provide a simple framework to understand the choices and identify the pitfalls. Specifically, this review article discusses the choice of microarray platform, preprocessing raw data, differential expression and validation, clustering, annotation and functional characterization of genes, and pathway construction in light of emergent concepts and tools.

functional genomics; normalization; differential expression; false discovery rate; clustering; annotation; pathway construction

DNA MICROARRAY REPRESENTS a powerful tool in biomedical discoveries. This review article discusses the choice of microarray platform, preprocessing raw data, differential expression and validation, clustering, annotation and functional characterization of genes, and pathway construction in light of emergent concepts and tools (Fig. 1).

BENEFITS AND SHORTCOMINGS OF MICROARRAY ANALYSIS

The advent of the cDNA and oligonucleotide microarray accelerated the rate of discovery of genetic interplay by simultaneously monitoring thousands of genes in a single experiment (12, 63). This systemic approach is valuable to identify novel mechanisms in the regulation and production of proteins and to refine our understanding of known pathways in the context of proteomics and the metabolome (2, 9, 27). Microarray analysis also supports the discovery of drug-sensitive genes and the chemical substructures associated with specific genetic responses (20). Current clinical applications include the development of biomarkers for classification into disease subgroups and the monitoring of disease progression (33, 49, 62). On the other hand, attempts to reproduce expression values using different microarray platforms with the same samples, or the same platform with similar samples, or even pure technical replicates (the same platform with split samples), have demonstrated poor overall reliability (3, 4). Whatever information is embedded in a microarray experiment appears to be entangled in a complex mix of various types of noise. This has caused some researchers to call for establishing industrial manufacturing standards and further independent and thorough validation of the technology. Others welcome the diversity of platforms and analytic methods as complementary forms of discovery, relying on alternative PCR-based technologies for validation of expression levels. Lack of a robust and reliable data analysis platform represents the single most important limiting factor in microarray analysis. In the current environment, there are many choices and potential pitfalls for researchers who intend to incorporate microarrays as a tool to monitor global gene expression patterns. This review is intended to provide a framework to understand the choices and identify the pitfalls. Unless otherwise specified, all reference to microarrays in this article refers to oligonucleotide or cDNA microarrays.

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: C. K. Sen, 512 Davis Heart & Lung Research Institute, The Ohio State Univ. Medical Center, 473 W. 12th Ave., Columbus, OH 43210 (e-mail: chandan.sen@osumc.edu).

1094-8341/06 $8.00 Copyright © 2006 the American Physiological Society

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

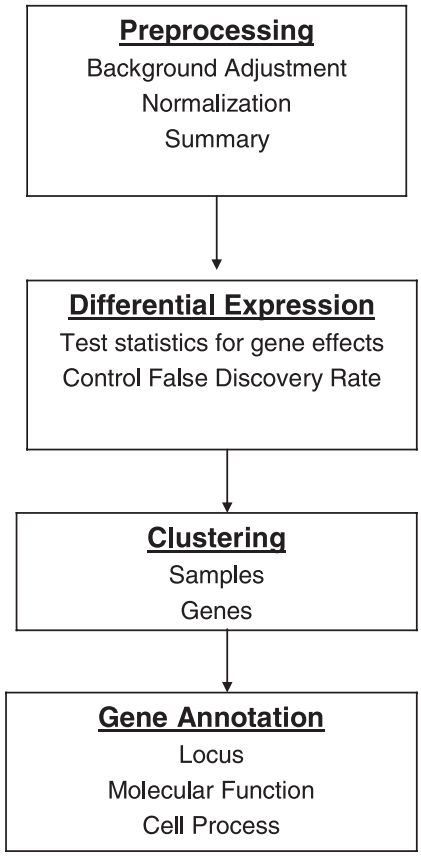

Fig. 1. Statistical analysis of microarray data: general approach.

CHOICE OF PLATFORM

Gene microarray systems differ in terms of material used (short oligonucleotides, long oligonucleotides, or cDNA) and number of samples per array (single-channel or 2-channel). Oligonucleotide arrays typically are used for genome-wide (tens of thousands of genes) screening and cDNA arrays for the investigation of smaller sets of genes. Links to twenty-eight companies supplying off-the-shelf and custom array services may be found at St. George's University of London bioinformatics web site, **http://www.sgul.ac.uk/depts/medmicro/ArrayLinkDB.htm.** Companies differ widely in their recommendations for sample preparation, imaging conditions, and preprocessing of raw data. Caution is warranted not only for designing settings to compare different systems but also to ensure reliability of a single system. Use of a single system in multicenter trials is beginning to demonstrate reliable and generalizable findings. In 2002, Kuo et al. (34) compared mRNA measurements of 2,895 sequence-matched genes in 56 cell lines from the standard panel of 60 cancer cell lines from the National Cancer Institute (NCI 60) by calculating the correlation between matched measurements and calculating concordance between clusters from two high-throughput DNA microarray technologies, Stanford type cDNA microarrays and Affymetrix oligonucleotide microarrays. In general, corresponding measurements from the two platforms showed poor correlation. Clusters of genes and cell lines were discordant between the two technologies, suggesting that relative intratechnology relationships were not preserved. GC content, sequence length, average signal intensity, and an estimator of cross-hybridization were found to be associated with the degree of correlation. This suggests gene-specific or, more correctly, probe-specific factors influencing measurements differently in the two platforms, implying a poor prognosis for a broad utilization of gene expression measurements across platforms. Gene expression measurements generated from identical RNA preparations from human panc-1 pancreatic cancer cells that were obtained using three commercially available microarray platforms have been compared (57). Three biological replicates were prepared for each of two serum growth conditions, and three experimental replicates were produced for the first biological replicate. RNA was labeled and hybridized to microarrays from three major suppliers according to manufacturers' protocols, and gene expression measurements were obtained using each platform's standard software. For each platform, gene targets from a subset of 2,009 common genes were compared. Correlations in gene expression levels and comparisons for significant gene expression changes in this subset were calculated and showed considerable divergence across the different platforms (57).

Seven of the most popular platforms for microarray analysis, including Codelink, Affymetrix, Agilent, NimbleGen, Applied Biosystems, Febit, and custom-made cDNA spotted arrays, have been scrutinized side by side (24). The widely different approaches to measuring gene expression produced disparate estimates. Hardiman (24) recommends the use of meta-analysis techniques when attempting cross-platform integration of data. In a similar study (30) of four of these platforms, it was observed that

> clone errors [on the custom-made microarrays], annotation differences, and technical differences between the platforms may be so significant that they exceed the biological differences between gene expression patterns in samples whose expression profiles are relatively similar.

In a study of genes encoding transporters and ion channel, it was noted that the cDNA and Affymetrix arrays correlate well over those genes that are abundantly expressed, but there is very little agreement about differential expression when observing relatively low levels of expression (35). Recently, more positive results are beginning to appear. Short oligonucleotide (25- to 30-base), long oligonucleotide (50- to 80-base), and cDNA (highly variable in length) formats have been compared with test RNA samples from six different cell lines against a universal reference standard (47). The three platforms had 6,430 genes in common. The study noted that correlation of gene expression levels across the platforms was good if the criterion is the direction of change in gene expression and minimal emphasis is placed on the magnitude of change (47). Recently, lung adenocarcinoma expression data from four laboratories have been compared (17). To test the feasibility of combining data across laboratories, frozen tumor tissues, cell line pellets, and purified RNA samples were analyzed at each of the four laboratories. Samples of each type and several subsamples from each tumor and each cell line were blinded before being distributed. The laboratories followed a common protocol for all steps of tissue processing, RNA extraction, and microarray analysis using Affymetrix Human Genome U133A arrays. High within-laboratory and between-laboratory correlations were observed on the purified RNA samples, the cell

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

lines, and the frozen tumor tissues. Intraclass correlation within laboratories was only slightly stronger than between laboratories, and the intraclass correlation tended to be weakest for genes expressed at low levels and showing small variation (17). Hierarchical cluster analysis revealed that the repeated samples clustered together regardless of the laboratory in which the experiments were performed. These findings indicate that, under properly controlled conditions, it is feasible to perform complete microarray analysis, from tissue processing to hybridization and scanning, at multiple independent laboratories for a single study (17).

PREPROCESSING RAW DATA

Microarrays are imaged using an optical scanner. These images must then be subjected to background correction to adjust for nonspecific binding, fluorescence from other chemicals on the slide, and the like. In the next preprocessing steps, the background-corrected data are normalized to adjust for differences that are not biological in nature but are due to the technology (e.g., dye effects) and summarized, so that the normalized values of multiple probes for the same gene are combined into a single value representing the consensus level of expression for that gene. After preprocessing, additional steps are taken to determine which genes are differentially expressed, to search for clusters of genes (or subjects) with similar gene expression patterns, and to annotate the differentially expressed genes with a functional assessment. Often, relational databases are then used to identify pathway components compatible with the observed patterns of expression.

Background correction and normalization. DNA microarrays often contain multiple probes for each gene. The probes are typically scattered over the surface of the microarray hardware. Variations in intensity from probe to probe or chip to chip for the same sample need to be resolved into a reliable level of expression. Observed intensities are sometimes modified based on comparison with nearby background probes whose expression is theoretically known. For cDNA microarrays, background adjustment is controversial, since, although it can reduce bias, it can also increase variance. See Scharpf et al. (54) for discussion of the bias-variance tradeoff and Smyth et al. (56) for a description of some of the commonly used background adjustment methods. For Affymetrix arrays, each gene probe has a single-nucleotide mismatch probe mate. The Microarray Suite (MAS) 5.0 method of Affymetrix, which uses paired probes for adjustment, and the robust multichip average (RMA) method (8), which uses quantile adjustment, are both in common practice. The recently developed GC RMA method pools probes with comparable numbers of G-C bonds to achieve a stable mismatch adjustment (64).

Microarray data can be quite noisy. Much of the variation in intensity levels can arise from technical rather than biological causes. Nonbiological sources of variation can be introduced during sample preparation (e.g., dye effects), array manufacture (e.g., probe concentration), and hybridization (e.g., amount of sample) and in the measurement process (e.g., scanner inaccuracies) (25). The possible sources of obscuring variation have been reviewed recently (25). Some biological sources of bias, such as comparison of in vivo with in vitro samples, may call for special adjustment. Hence, it is important to normalize data, as much as possible, to remove the technical variation while still retaining the informative biological variation. Normalization is performed both within each array and between arrays to make comparisons more meaningful. Although normalization is somewhat ad hoc, there are two basic ideas that are relevant to all microarray normalizations. First, the normalization method must be tailored to the microarray platform (8, 50, 52, 53, 66). Normalization of cDNA arrays is quite different from normalization of Affymetrix arrays, both in the sources of variation to be removed and in the algorithms. Second, linear normalization methods often miss obscuring variation that can be removed, so nonlinear methods should be used (8, 41, 66).

DIFFERENTIAL EXPRESSION AND VALIDATION

Differential expression. Identifying genes that are differentially expressed under two or more treatment conditions is a primary goal of most microarray studies. The two main issues in assessing differential expression are determining a method for assessing the extent of differential expression (e.g., fold change, *t*-test, ANOVA) and adjusting the method for the effects of multiple comparisons, since typically there are thousands of genes being studied. Differential expression is traditionally approached one gene at a time (e.g., fold change, *t*-test, ANOVA). One important point is the weakness of relying on fold change as the sole criterion, since fold change does not take into account the variability in the data. This can lead to two problems. First, genes with low expression levels yet large fold changes and high variability may be identified as differentially expressed. Second, genes that display small but reproducible (i.e., low variability) changes in gene expression may be missed. There have been some efforts to incorporate variability in methods that rely on fold change (41), but these still suffer from difficulties in assessing the error rates. Also, empirical Bayes methods that shrink individual estimates of variance toward a common value have been suggested for improving the behavior of *t*-statistics in the many gene settings (19). Recently, a number of high-dimensional methods have been proposed to use covariance structure to assist in identifying differentially expressed genes. These include elastic net (68), gradient-directed regularization (21), and multiple forward search (44). Shrunken centroid ordering by orthogonal projections (SCOOP) is a new method still under testing, with R code available from J. S. Verducci.

Multiple comparisons and false discovery rate. The issue of multiple comparisons is more complex. Ideally, the probability of a false positive (a gene incorrectly identified as differentially expressed) should be small, and the probability of correctly identifying genes that are differentially expressed should be large. Standard statistical methods are set up to balance these goals in the context of only one comparison, i.e., if the microarray contained only one gene. Without adjustment, standard statistical methods give incorrect results in the context of microarray data. For example, consider a microarray study with m genes, and suppose none is differentially expressed. For various values of m, the probability that a standard statistical tool set to reject the null hypothesis if a P value is <0.05 will yield at least one false positive is given in Table 1. Because most microarrays contain thousands of genes, standard statistical methods are clearly unacceptable.

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

Table 1. *Probability of at least one false positive increases rapidly as the no. m of hypotheses increases*

m	Probability of At Least One False Positive
1	0.05
10	0.40
50	0.92
100	0.994

The Bonferroni method is a simple method to correct for multiple testing that is still widely used in microarray data analysis (43). This method just divides the *P* value cutoff by the number of genes *m*. For example, if the probability of at least one false positive is to be limited to 0.01, and there are $m = 5{,}000$ genes on the array, the Bonferroni method would identify a gene as differentially expressed if its *P* value was $<0.01/5{,}000 = 0.000002$. Although this method is quite generally applicable, it is usually not a good choice for microarray studies because it has very low power, i.e., the probability of correctly identifying differentially expressed genes is very small, so many potentially interesting genes may be missed. For this and other reasons, different criteria than the probability of at least one false positive have been advocated. The most promising of these is the false discovery rate (FDR) (7, 65). FDR is the expected proportion of false positives among all rejected hypotheses. Instead of trying to avoid any false positives, the FDR controls the proportion of positive calls that are false positives. Designing procedures to control the FDR is challenging. The original technique of Benjamini and Hochberg (6), to control the FDR at level α, works as follows. First, *P* values are computed for each of the *m* genes, and the *P* values are ordered from smallest to largest. Second, the ordered *P* values are plotted vs. their rank along with the line with slope α/m and intercept zero. The last *P* value, say P^*, that lies below the line is noted. This value (P^*) is used to reject the hypotheses corresponding to all *P* values less than or equal to P^*. The Benjamini-Hochberg procedure has been shown to control the FDR under certain assumptions on the dependence structure of the genes' expression levels (6). The procedure is in wide use and is recommended by the American Physiological Society (13). Unfortunately, there are many microarray studies not covered by the assumptions underlying the Benjamini-Hochberg algorithm. Thus there is much work in the statistical community aimed at developing a method of controlling the FDR that is more generally applicable than the original Benjamini-Hochberg method. A promising method that relies on the bootstrap technique has been recently analyzed (48, 60, 61). However, this method achieves the FDR asymptotically. Thus it is not suitable for studies involving small numbers (e.g., 4–5) of arrays.

Determining sample size needed to control FDR. In planning an experiment, there are two major decisions to make about microarrays: *1*) the total number of microarrays that should be used and *2*) the proportion that will be used for biological vs. technical replication. The first decision is typically based on budget and the second on the reliability of the microarrays being used. The real question is whether a planned experiment has a realistic chance of detecting and identifying important biological processes. Recently, a decision theoretic procedure was introduced (46) where a typical loss function is a weighted sum of the FDR and its counterpart false negative rate (FNR). The idea is to plot the expected loss vs. sample size and judge whether a desired value can be achieved with a realistic sample size. The expected loss is estimated through simulating expression data and recording the behavior of the Benjamini-Hochberg method.

CLUSTERING GENES AND CASES

Heat map represents a common approach to present gene expression data. This is an array where, typically, genes index the rows, and chips index the columns (Fig. 2). Chips may represent either different subjects or the same subject under different conditions. The array itself is color coded to display the, usually normalized, level of expression. The variation of colors along any row is called the expression pattern of the associated gene. If the genes in the array are arbitrarily ordered, it is difficult to perceive patterns in the heat map. The simplest remedy is to sort the genes in such a way that two genes with similar expression patterns are close together. Hierarchical procedures begin by putting each individual into its own group, combining the two closest, combining the next two closest, and so on. This requires a measure of closeness between groups of individuals. Such measures are constructed by specifying a metric between individuals and a procedure of using this metric to induce a distance between two groups. For example, the Manhattan distance between two genes is the sum of absolute values of the difference in expression on each microarray. The complete linkage method of induction defines the distance between two groups as the largest metric distance between two individuals, one from each group. Different metrics and procedures may produce different blocks of patterns. The efficacy of standard clustering algorithms in identifying clear patterns has been reviewed (10, 11, 23, 42, 58).

A recently proposed method, called hierarchical ordered partitioning and collapsing hybrid (HOPACH), alternates the "top down" method of partitioning with the "bottom up" method of agglomeration to produce clusters that are reliably reproduced when subjects are resampled or experiments are replicated (51). The ultimate criterion for clustering is whether clustered genes tend to act in conjunction with each other. As an alternative to searching through dozens of possible clustering combinations, a new approach is to append additional information to the expression matrix before attempting to cluster genes. The additional information may be an important aspect of the expression matrix, for example the difference in mean expression between low and high oxygenation experiments, or external information such as gene functioning categories. This method of appending onto expression matrices may afford a stronger aid to identifying relevant gene families and/or pathways, or may just simplify the heat map. Although appending genes with functional categories may yield functionally interpretable groupings, it does not help all that much in identifying different pathways, since, for example, all active transcription factors tend to be clustered together even if they are operating in disjoint networks. Moreover, a single gene may be a functional part of more than one network or pathway, but traditional clustering methods allow a gene to be included in only one cluster. Plaid models (36) represent an attempt to handle multiple-group membership of genes. Regularized col-

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

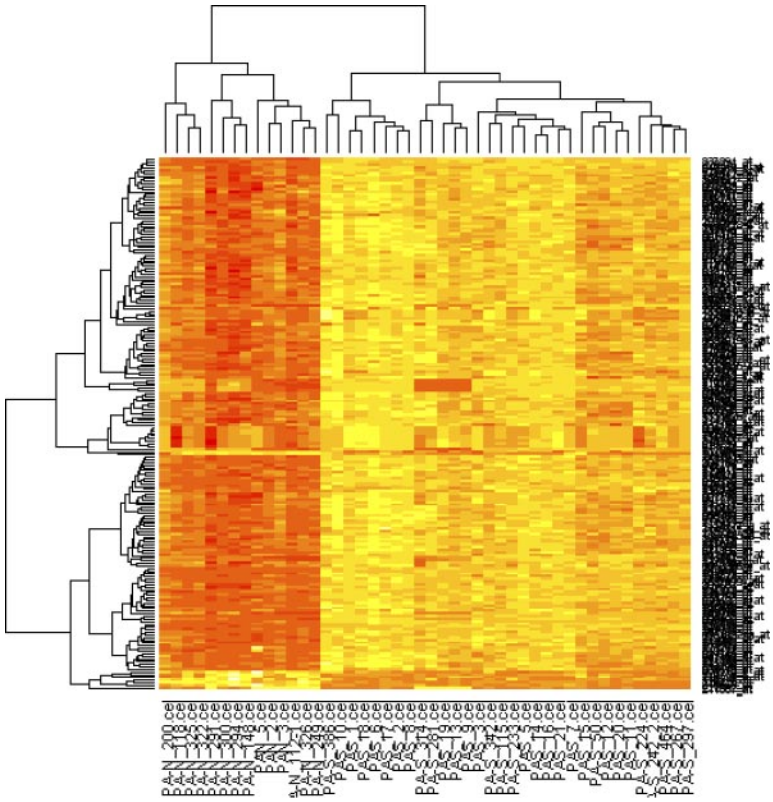

Fig. 2. Heat map of 500 genes for normal and ischemic human hearts (map generated based on data from **http://www.cardiogenomics.org**).

oring patterns are imposed in layers, which ideally represent independent networks. Each layer is color coded, with intensity of that color being proportional to the expression level attributable to activity in that layer. When the layers are overlaid, the result is a plaid pattern (39). To be successful, the method requires a precision and uniformity of variance in the microarray data that has not been achieved yet, but the method may be valuable as another tool in inferring pathways. A promising area of research is to guide the choice of layers using partial information about pathways.

ANNOTATION AND FUNCTIONAL CHARACTERIZATION OF GENES

Detection and clustering of differentially expressed genes, as described above, are just the first steps toward learning about gene function and genetic networks. There are many situations when limited gene expression data are available but existing gene networks or functional classes of genes are known. In this case, one can try to relate. Perhaps the single most important source of information for relating newly acquired gene expression level data to known functional and partial pathway information is the Gene Ontology (GO) database. The GO project (**http://www.geneontology.org/**) is a collaborative effort to address the need for a consistent description of gene products in different databases. GO produces a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. GO provides three structured networks of defined terms to describe gene product attributes (37, 59). The three organizing principles of GO are molecular function, biological process, and cellular component. A gene product has one or more molecular functions and is used in one or more biological processes; it might be associated with one or more cellular components. As an example to demonstrate the potential usage of the GO tools in cardiovascular medicine, a sample functional analysis of the list of genes identified to be differentially expressed in ischemic cardiomyopathy patients vs. "normal" organ donors (**http://www.cardiogenomics.org**) is described. The data set consists of a total of 46 Affymetrix microarrays of 32 ischemic cardiomyopathy patients and 14 normal donors. Figure 2 illustrates the expression patterns for 500 genes screened as differentially expressed using RMA preprocessing and a multivariate filter. Values from normal donors comprise the *left* 14 columns. From the set of 500 genes, 66 genes were chosen using FDR criteria. Among these genes, 47 of them were upregulated in the ischemic hearts, while 19 were downregulated. To discern whether the genes selected as differentially expressed are meaningful biologically, we utilized the Gene Ontology Tree Machine (GOTM; **http://genereg.ornl.gov/gotm**) to annotate their functions and to classify them into functional categories. Using all genes in the human genome as our reference gene set, we were interested in identifying GO categories that are being enriched in our set of 66 genes. In other words, we sought to identify functional categories in which there are more genes in our list belonging to them than expected if the genes were randomly selected from the human genome. For a specific given category, under the null hypothesis of random selection, the number of genes from our list falling into that particular category follows a hypergeometric distribution, leading to a simple test for the hypothesis. All GO categories that were identified to be significantly enriched (raw $P < 0.01$; with category names in red), together with their ancestral categories (up to the top level with 3 main categories: biological process, molecular function, and cellular component), are displayed as a directed acyclic graph (see Fig. 4). The numbers below or next to a category are the observed/

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

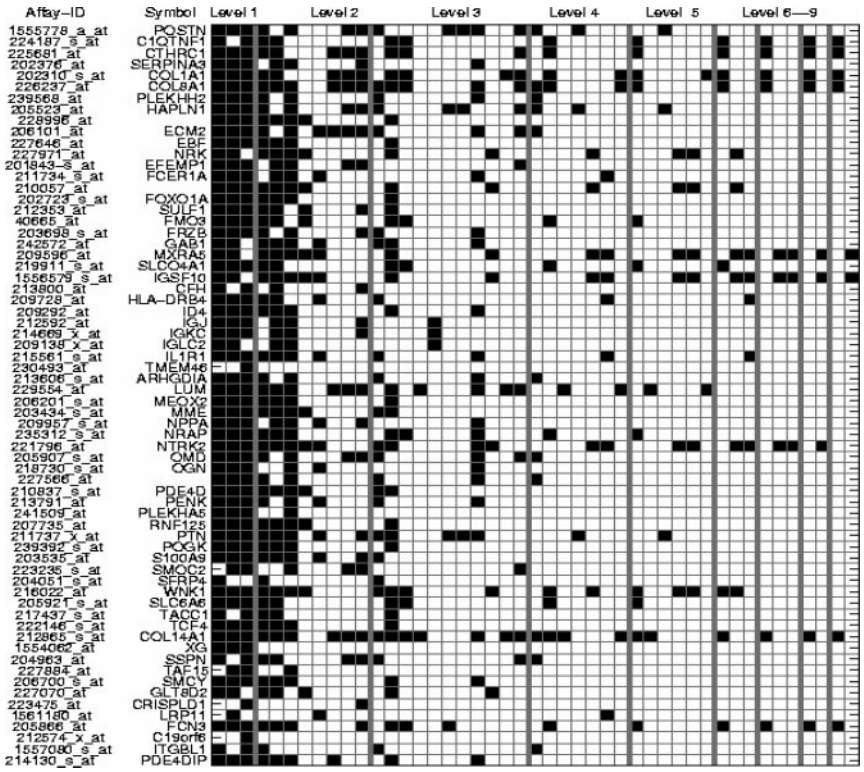

Fig. 3. List of genes involved in each functional category of Fig. 4.

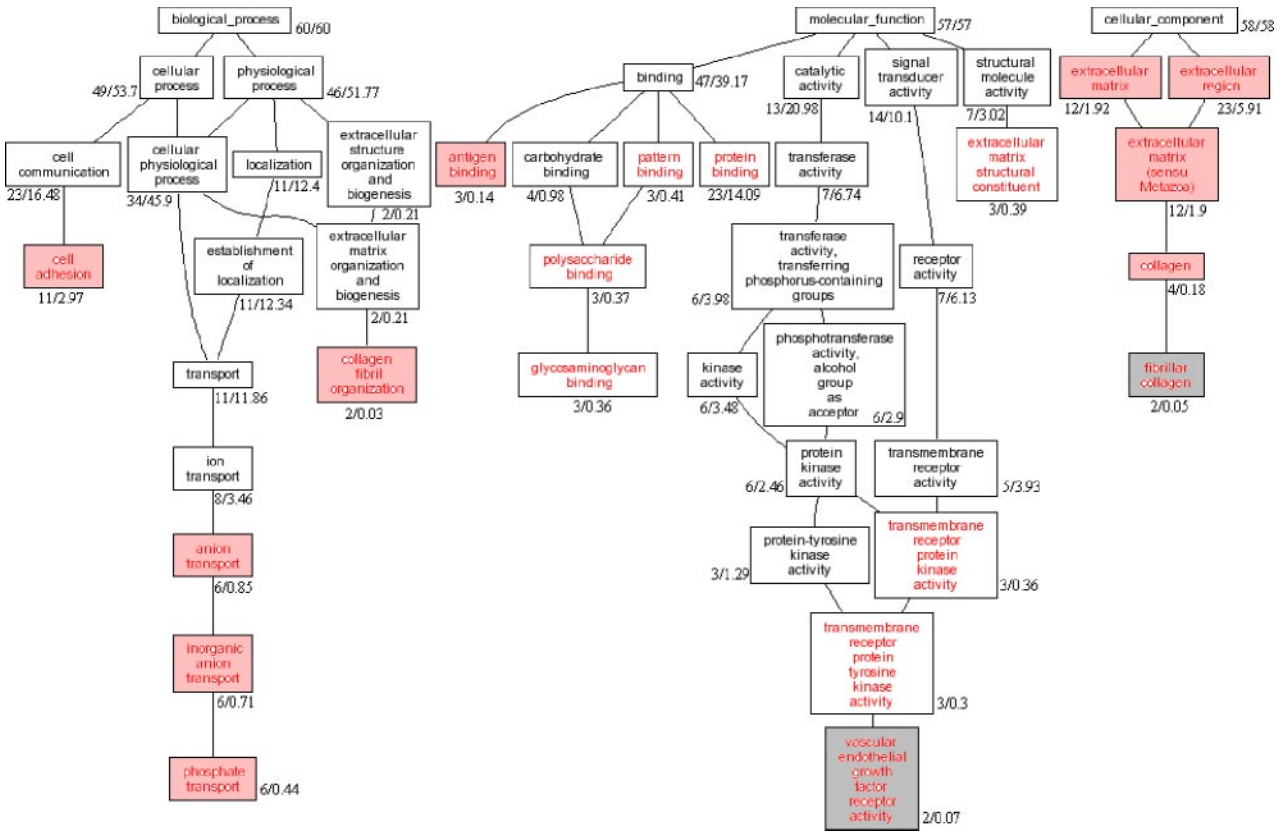

Fig. 4. Directed acyclic graph view of the significantly enriched Gene Ontology categories (names in red) and their ancestor categories.

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

expected gene numbers for that category. Displayed in Fig. 3 are the genes involved in each GO category shown in Fig. 4. Each row represents one gene, with a black square denoting the involvement of the gene in the corresponding column (category). Each column stands for one category ordered first by levels. Within each level, the categories are ordered from *left* to *right* according to the display. From the raw *P* values of GOTM, we calculated the adjusted *P* values to correct for multiple testing using the FDR method (6). A cutoff of 0.05 for the adjusted *P* values led to a number of enriched categories no longer being enriched (red categories without shading in Fig. 3). More specifically, those with FDR-adjusted *P* values < 0.01 are shaded in pink, whereas those with FDR-adjusted *P* values between 0.05 and 0.01 are shaded in gray. We note that most of the enriched categories are involved in "transport," "binding," "extracellular," and "transmembrane" activities. In particular, the two genes involved in the category "vascular endothelial growth factor receptor activity" are MXRA5 and IGSF10, both of which are upregulated. These findings are consistent with those reported by Lee et al. (38).

PATHWAY CONSTRUCTION

The Kyoto Encyclopedia of Genes and Genomes, or KEGG (31, 32), is a suite of databases and associated software integrating current knowledge on molecular interaction networks in biological processes (the PATHWAY database), on the universe of genes and proteins (the GENES/SSDB/KO databases), and on the universe of chemical compounds and reactions (the COMPOUND/GLYCAN/REACTION databases). Several methodologies have been proposed for constructing gene networks based on gene expression data, such as the Boolean networks (1) or differential equation models (15). Another way of addressing this question is the Bayesian networks framework (22, 26), where the expression level of each gene is treated as a random variable and each regulatory interaction as a probabilistic dependency between such variables. Bayesian networks are graph-based models of joint multivariate probability distributions that capture properties of conditional independence between variables. Such models are attractive for their ability to describe complex stochastic processes and provide a tool for learning from noisy observations. In addition, bootstrap methods can be used for estimating confidence in the learned structures (19). The main idea is to sample, with replacement, observations from the given data set and learn for them. In this way, many networks are generated, all of which are reasonable models reflecting the effect of small perturbations in data on the learning process (22). However, because of limited expression data typically available for any particular system in a given state, network reconstruction processes typically result in the identification of multiple networks that explain data equally well. In most cases, causal relationships cannot be reliably inferred from gene expression data alone, since, for any particular network, changing the direction of the edge between two nodes has little effect on the model fit. To reliably infer causal relationships, additional information is required. Biological knowledge (29), including protein-protein and protein-DNA interactions (28), binding site sequences, and transcription factors (40), is needed. More recently, pathway reconstruction associated with complex disease traits was obtained by integrating genotype, transcription, and clinical trait data (67). In this approach, gene expression data were treated as quantitative trait loci (QTL). Patterns of colocalization between disease trait QTL and gene expression QTL are indicative for causal inference. There are several network/pathway reconstruction and analysis software packages that implement these ideas. One example is Genetic Network Analyzer (GNA), which is a computer tool for the modeling and simulation of genetic regulatory networks (5, 16). The aim of GNA is to assist biologists and bioinformaticians in constructing a model of a regulatory network using knowledge about regulatory interactions in combination with gene expression data. Another software tool is Gene MicroArray Pathway Profiler (GenMAPP), a free computer application designed to visualize gene expression data on maps representing biological pathways and groups of genes (14, 18). Another useful software tool is Cytoscape, an open-source software project for integrating biomolecular interaction networks with high-throughput expression data and other molecular states into a unified conceptual framework (45, 55). Although applicable to any system of molecular components and interactions, Cytoscape is most powerful when used in conjunction with the large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans and model organisms.

In conclusion, the promise of gene expression studies using microarray technology has inspired much new hope for finding complex disease genes. The majority of the initial technical challenges of conducting experiments are being resolved only to be replaced with new informatics hurdles, including statistical analysis, data visualization, and interpretation. Advances in microarray technology have necessitated parallel mining of large volumes of biological data. Progress in the genomics revolution is limited by our ability to transform such large amounts of raw data into reliable and meaningful biological sense. Emergent software and statistical tools as well as web resources address the multidimensional complexities faced by investigators while making sense of their microarray data. Academic core facilities are the likely medium of distilling that interdisciplinary information and carrying it to the end user who is seeking to employ microarrays as a tool to generate or address hypotheses. The robust ability to reconstruct signaling pathways based on microarray data requires tighter interplay and integration between bioinformatics and systems biology.

GRANTS

This work is supported by National Heart, Lung, and Blood Institute Grant RO1-073087 to C. K. Sen. In addition, microarray-related research in the laboratory is supported by National Institutes of Health Grants GM-069589 and NS-42617 to C. K. Sen. This work was also supported in part by the National Science Foundation (Agreement No. 0112050).

REFERENCES

1. **Akutsu T, Miyano S, and Kuhara S.** Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*: 17–28, 1999.
2. **Arakawa K, Kono N, Yamada Y, Mori H, and Tomita M.** KEGG-based pathway visualization tool for complex omics data. *In Silico Biol* 5: 0039, 2005.
3. **Asyali MH and Alci M.** Reliability analysis of microarray data using fuzzy c-means and normal mixture modeling based classification methods. *Bioinformatics* 21: 644–649, 2005.

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

4. **Asyali MH, Shoukri MM, Demirkaya O, and Khabar KS.** Assessment of reliability of microarray data and estimation of signal thresholds using mixture modeling. *Nucleic Acids Res* 32: 2323–2335, 2004.
5. **Batt G, Ropers D, de Jong H, Geiselmann J, Mateescu R, Page M, and Schneider D.** Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in *Escherichia coli*. *Bioinformatics* 21, *Suppl* 1: i19–i28, 2005.
6. **Benjamini Y and Hochberg Y.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 57: 289–300, 1995.
7. **Benjamini Y and Yekutieli D.** The control of the false discovery rate in multiple testing under dependency. *Ann Statist* 29: 1165–1188, 2001.
8. **Bolstad BM, Irizarry RA, Astrand M, and Speed TP.** A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19: 185–193, 2003.
9. **Bono H and Okazaki Y.** The study of metabolic pathways in tumors based on the transcriptome. *Semin Cancer Biol* 15: 290–299, 2005.
10. **Bryan J, Pollard KS, and van der Laan MJ.** Paired and unpaired comparison and clustering with gene expression data. *Statist Sinica* 12: 87–110, 2002.
11. **Chen G, Jaradat SA, Banerjee N, Tanaka TS, Ko MSH, and Zhang MQ.** Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data. *Statist Sinica* 12: 241–262, 2002.
12. **Clarke JD and Zhu T.** Microarray analysis of the transcriptome as a stepping stone towards understanding biological systems: practical considerations and perspectives. *Plant J* 45: 630–650, 2006.
13. **Curran-Everett D and Benos DJ.** Guidelines for reporting statistics in journals published by the American Physiological Society. *Physiol Genomics* 18: 249–251, 2004.
14. **Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, and Conklin BR.** GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet* 31: 19–20, 2002.
15. **de Hoon MJ, Imoto S, Kobayashi K, Ogasawara N, and Miyano S.** Inferring gene regulatory networks from time-ordered gene expression data of *Bacillus subtilis* using differential equations. *Pac Symp Biocomput*: 17–28, 2003.
16. **de Jong H, Geiselmann J, Hernandez C, and Page M.** Genetic Network Analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics* 19: 336–344, 2003.
17. **Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, Conley B, Buetow KH, Heiskanen M, Simon RM, Minna JD, Girard L, Misek DE, Taylor JM, Hanash S, Naoki K, Hayes DN, Ladd-Acosta C, Enkemann SA, Viale A, and Giordano TJ.** Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin Cancer Res* 11: 565–572, 2005.
18. **Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, and Conklin BR.** MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol* 4: R7, 2003.
19. **Efron B and Tibshirani RJ.** *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.
20. **Favis R, Gerry NP, Cheng YW, and Barany F.** Applications of the universal DNA microarray in molecular medicine. *Methods Mol Med* 114: 25–58, 2005.
21. **Friedman J and Popescue B.** Gradient directed regularization [Online]. http://www-stat.stanford.edu/~jhf/ftp/pathlite.pdf [2004].
22. **Friedman N, Linial M, Nachman I, and Pe'er D.** Using Bayesian networks to analyze expression data. *J Comput Biol* 7: 601–620, 2000.
23. **Goldstein DR, Ghosh D, and Conlon EM.** Statistical issues in the clustering of gene expression data. *Statist Sinica* 12: 219–240, 2002.
24. **Hardiman G.** Microarray platforms—comparisons and contrasts. *Pharmacogenomics* 5: 487–502, 2004.
25. **Hartemink A, Gifford D, Jaakkola TS, and Young RA.** Maximum likelihood estimation of optimal scaling factors for expression array normalization. In: *Microarrays: Optical Technologies and Informatics*, edited by Bittner M, Chen Y, Dorsel A, and Doubgherty E. Bellingham, WA: SPIE-International Society for Optical Engineering, 2001, p. 132–140.
26. **Heckerman D.** A tutorial on learning Bayesian networks. In: *Learning in Graphical Models*, edited by Jordan M. Cambridge, MA: MIT Press, 1999, p. 301–354.
27. **Howbrook DN, van der Valk AM, O'Shaughnessy MC, Sarker DK, Baker SC, and Lloyd AW.** Developments in microarray technologies. *Drug Discov Today* 8: 642–651, 2003.
28. **Ideker T, Ozier O, Schwikowski B, and Siegel AF.** Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18: S233–S240, 2002.
29. **Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, and Miyano S.** Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *J Bioinform Comput Biol* 2: 77–98, 2004.
30. **Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, and Monni O.** Are data from different gene expression microarray platforms comparable? *Genomics* 83: 1164–1168, 2004.
31. **Kanehisa M.** A database for post-genome analysis. *Trends Genet* 13: 375–376, 1997.
32. **Kanehisa M and Goto S.** KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30, 2000.
33. **Kittleson MM and Hare JM.** Molecular signature analysis: using the myocardial transcriptome as a biomarker in cardiovascular disease. *Trends Cardiovasc Med* 15: 130–138, 2005.
34. **Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, and Kohane IS.** Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* 18: 405–412, 2002.
35. **Landowski CP, Anderle P, Sun D, Sadee W, and Amidon GL.** Transporter and ion channel gene expression after Caco-2 cell differentiation using 2 different microarray technologies. *AAPS J* 6: e21, 2004.
36. **Lazzeroni L and Owen A.** Plaid models for gene expression data. *Statist Sinica* 12: 61–86, 2002.
37. **Lee JS, Katari G, and Sachidanandam R.** GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics* 6: 189, 2005.
38. **Lee SH, Wolf PL, Escudero R, Deutsch R, Jamieson SW, and Thistlethwaite PA.** Early expression of angiogenesis factors in acute myocardial ischemia and infarction. *N Engl J Med* 342: 626–633, 2000.
39. **Lee SI and Batzoglou S.** Application of independent component analysis to microarrays. *Genome Biol* 4: R76, 2003.
40. **Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, and Young RA.** Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799–804, 2002.
41. **Li C and Wong WH.** Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci USA* 98: 31–36, 2001.
42. **Liang J and Kachalo S.** Computational analysis of microarray gene expression profiles: clustering, classification, and beyond. *Chemometrics Intel Lab Sys* 62: 199–216, 2002.
43. **Lin DY.** An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 21: 781–787, 2005.
44. **Lu Y, Liu PY, Xiao P, and Deng HW.** Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics* 21: 3105–3113, 2005.
45. **Maere S, Heymans K, and Kuiper M.** BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21: 3448–3449, 2005.
46. **Muller P, Parmigiani G, Robert C, and Rousseau J.** Optimal sample size for multiple testing: the case of gene expression microarrays. *J Am Stat Assoc* 99: 990–1001, 2004.
47. **Petersen D, Chandramouli GV, Geoghegan J, Hilburn J, Paarlberg J, Kim CH, Munroe D, Gangi L, Han J, Puri R, Staudt L, Weinstein J, Barrett JC, Green J, and Kawasaki ES.** Three microarray platforms: an analysis of their concordance in profiling gene expression. *BMC Genomics* 6: 63, 2005.
48. **Pollard KS and van der Laan MJ.** Choice of a null distribution in resampling-based multiple testing. *J Stat Plann Inference* 125: 85–100, 2004.
49. **Pusztai L and Hess KR.** Clinical trial design for microarray predictive marker discovery and assessment. *Ann Oncol* 15: 1731–1737, 2004.
50. **Quackenbush J.** Microarray data normalization and transformation. *Nat Genet* 32, *Suppl*: 496–501, 2002.
51. **Salomonis N, Cotte N, Zambon AC, Pollard KS, Vranizan K, Doniger SW, Dolganov G, and Conklin BR.** Identifying genetic networks underlying myometrial transition to labor. *Genome Biol* 6: R12, 2005.

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.

52. **Schadt EE, Li C, Ellis B, and Wong WH.** Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J Cell Biochem Suppl*, *Suppl* 37: 120–125, 2001.
53. **Schadt EE, Li C, Su C, and Wong WH.** Analyzing high-density oligonucleotide gene expression array data. *J Cell Biochem* 80: 192–202, 2000.
54. **Scharpf RB, Iacobuzio-Donahue CA, Sneddon JB, and Parmigiani G.** When should one subtract background fluorescence in two color microarrays? (July 2005). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working paper 50. http://www.bepress.com/jhubiostat/paper50.
55. **Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T.** Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498–2504, 2003.
56. **Smyth GK, Yang YH, and Speed T.** Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* 224: 111–136, 2003.
57. **Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, and Cam MC.** Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* 31: 5676–5684, 2003.
58. **Tibshirani R, Hastie T, Narasimhan B, Eisen M, Sherlock G, Brown P, and Botstein D.** Exploratory screening of genes and clusters from microarray experiments. *Statist Sinica* 12: 47–59, 2002.
59. **Tuikkala J, Elo L, Nevalainen OS, and Aittokallio T.** Improving missing value estimation in microarray data with gene ontology. *Bioinformatics* 22: 566–572, 2006.
60. **van der Laan MJ, Dudoit S, and Pollard KS.** Multiple Testing Part II: Step-down procedures for control of the family-wise error rate. *Stat Appl Genet Mol Biol* 3, 2004.
61. **van der Laan MJ, Dudoit S, and Pollard KS.** Multiple Testing Part I: Single-step procedures for control of general type I error rates. *Stat Appl Genet Mol Biol* 3, 2004.
62. **Wadlow R and Ramaswamy S.** DNA microarrays in clinical cancer research. *Curr Mol Med* 5: 111–120, 2005.
63. **West RB and van de Rijn M.** The role of microarray technologies in the study of soft tissue tumours. *Histopathology* 48: 22–31, 2006.
64. **Wu Z and Irrizarry R.** A model based background adjustment for oligonucleotide arrays. *J Am Stat Assoc* 100: 909–917, 2005.
65. **Yang JJ and Yang MC.** An improved procedure for gene selection from microarray experiments using false discovery rate criterion. *BMC Bioinformatics* 7: 15, 2006.
66. **Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP.** Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30: e15, 2002.
67. **Zhu J, Lum PY, Lamb J, GuhaThakurta D, Edwards SW, Thieringer R, Berger JP, Wu MS, Thompson J, Sachs AB, and Schadt EE.** An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet Genome Res* 105: 363–374, 2004.
68. **Zou H and Hastie T.** Regularization and variable selection via the elastic net. *JR Stat Soc Ser B Stat Methodol* 67: 301–320, 2005.

Downloaded from journals.physiology.org/journal/physiolgenomics at Ohio State Univ HSL (140.254.087.149) on December 10, 2020.