Journal of Bioinformatics and Computational Biology
Vol. 4, No. 2 (2006) 259–274

ICP Imperial College Press
www.icpress.co.uk

ANALYSIS OF SNP-EXPRESSION ASSOCIATION MATRICES

ANYA TSALENKO

Agilent Technologies, 5301 Stevens Creek Blvd, MS23
Santa Clara, CA 95051, USA
anya_tsalenko@agilent.com

RODED SHARAN

School of Computer Science, Tel-Aviv University
Tel-Aviv 69978, Israel
roded@post.tau.ac.il

VESSELA KRISTENSEN*, HEGE EDVARDSEN†
and ANNE-LISE BØRRESEN-DALE‡

*,†,‡*Department of Genetics, Institute for Cancer Research,*
Rikshospitalet-Radiumhospitalet Medical Centre, Oslo, Norway

†,‡*Medical Faculty, University of Norway, Oslo, Norway*
**vessela@ulrik.uio.no*
†*hege.edvardsen@medisin.uio.no*
‡*a.l.borresen-dale@medisin.uio.no*

AMIR BEN-DOR

Agilent Technologies, 5301 Stevens Creek Blvd, MS23
Santa Clara, CA 95051, USA
amir_ben-dor@agilent.com

ZOHAR YAKHINI

Agilent Technologies, 94 Em Hamoshavot Rd
Petach-Tikva 49527, Israel
zohar_yakhini@agilent.com

Received 15 September 2005
Accepted 31 January 2006

High throughput expression profiling and genotyping technologies provide the means to study the genetic determinants of population variation in gene expression variation. In this paper we present a general statistical framework for the simultaneous analysis of gene expression data and SNP genotype data measured for the same cohort. The framework consists of methods to associate transcripts with SNPs affecting their expression, algorithms to detect subsets of transcripts that share significantly many associations with a subset of SNPs, and methods to visualize the identified relations. We apply our framework to SNP-expression data collected from 50 breast cancer patients. Our results demonstrate an overabundance of transcript-SNP associations in this data, and pinpoint

SNPs that are potential master regulators of transcription. We also identify several statistically significant transcript-subsets with common putative regulators that fall into well-defined functional categories.

Keywords: SNP genotype; gene expression; association analysis; biclusters.

1. Introduction

The development of high throughput techniques for expression profiling and genotyping enables the study of the genetic determinants of expression variation, both in humans and in other organisms. Expression levels are taken as quantitative phenotypes, of independent interest, as well as determinants or indications of endpoint clinical phenotypes. Much of our understanding of the genetic base of disease comes from identifying polymorphisms that affect protein structure or integrity. We know, however, that protein abundance and expression levels of mRNA also drive disease processes. It is therefore important to explore the genetic base of variation in gene expression, regarded as a quantitative phenotype.

Several studies that genetically analyzed expression phenotypes have been reported in the literature. In Brem *et al.*[1] the authors report the use of 3312 markers and genome wide linkage analysis to study the natural variation of expression in yeast. The expression levels of 570 genes were linked to one or more different loci, with most transcripts showing complex inheritance patterns. Loci that are linked to expression phenotypes are either cis-acting modulators of single genes (the polymorphism that is linked to the transcript is in the coding sequence of the gene or very close to it) or trans-acting modulators (the polymorphism is in a different part of the genome) of many genes. The study reports eight such trans-acting loci, each potentially affecting the expression levels of a group of 7 to 94 genes of related function. Jin *et al.*[2] investigated the genotypic contributions to transcriptional variance in Drosophila. They conclude that gene expression in adult flies is affected most strongly by sex, less so by genotype and only weakly by age (for 1- and 6-week flies). They also found that chromosome X genotype interactions may be present for as much as 10% of the Drosophila transcriptome.

Schadt *et al.*[3] describe genetic analysis of gene expression in three species, with an emphasis on mice. They produced a genetically diverse population of 111 mice by crossing two commonly used inbred strains and measured the expression levels of 23 574 mouse genes in blood samples from this population. They found that 7861 are differentially expressed in the two original strains. These were further used in studying the genetics of gene expression in this population, using a panel of more than 100 microsatellite markers. The study reports e-QTLs (expression quantitative trait loci): genetic regions (loci) that can account for variation in the levels of gene expression. Many e-QTLs were found at the same chromosomal location as the gene they hypothetically affect (*in-cis*). They also identified genomic regions with e-QTLs for an exceptionally large number of transcripts, suggesting the presence of regulatory elements.

Another approach to the study of the relationship between genetics and expression regulation is to study the differential expression between genetic variants. Sandberg *et al.*[4] studied genes that are differentially expressed between two inbred mouse strains at baseline and in response to seizure. They found that approximately 1% of expressed genes are differentially expressed between strains in at least one region of the brain and that the gene expression response to seizure is significantly different between the two strains. Hedenfalk *et al.*[5] study expression differences in breast tumor samples collected from patients of different *BRCA*1 and *BRCA*2 status to discover significant differences that attest to the effect of genetic differences on tumor expression patterns. In a related study[6] the authors suggested the use of differential expression and a class discovery process[7] to define more homogenous cohorts for genetic studies. A study of lymphoblastoid expression differences between carriers of ataxia telangiectasia, an autosomal recessive disease, and normal controls, further confirms the effect of genetic differences on the expression phenotype.[9] These results illustrated that heterozygous carriers, even of recessive conditions, can have a distinct phenotype.

In a recent report Morley *et al.*[8] described a broad study of the genetic determinants of normal expression variation in humans. The authors used microarrays to measure the baseline expression levels of roughly 8500 genes, or transcripts, in immortalized B cells from members of CEPH Utah pedigrees.[9] They selected 3554 genes that varied more between individuals than between replicates and used these as quantitative traits, to be mapped into genomic locations. They used public genotype information to carry out linkage analysis for these expression phenotypes in 14 CEPH families. They found high linkage signals for 984 of the transcripts (at $p < 0.05$, leading to an FDR of about 0.2). Interestingly, they identified regions that show linkage signals to many of the transcripts, and proposed that these can point towards master regulators of baseline expression levels.

In this paper we address the methodology of interpreting genetic linkage and association data for expression level phenotypes. We focus on the case of association studies involving single nucleotide polymorphisms (SNPs). It is possible to modify the methods to address other types of polymorphisms or of genetic analysis, such as linkage studies in pedigrees and LOD scores. The first step in analyzing SNP and gene expression data consists of computing the *association graph* $G = (S, T, E)$, a bipartite graph where S represents SNP loci (or genomic loci in general), T represents the transcripts measured in the expression profiling study and E is a set of edges that represent the genetic association. An edge $e = (s, t) \in E$ can be labeled by the p-value (or other score) of the observed association between s and t or it can be binary, when using a threshold to determine edges and non-edges. An alternative mathematical representation of the graph is the *association matrix* P. The entries P_{st}, again, represent the level of association between the genotypes measured at SNP locus s and the expression levels of a gene or transcript t. Studying properties of the association graph enables the identification of meaningful biological signals, similar to the ones described in some of the literature cited above. We describe

methods and results related to the following structures in G:

- Overall overabundance of associated pairs: we assess the overall significance of the observed association between genotypes and expression phenotypes by comparing it to a null model.
- Potential master regulators: we seek vertices in S that have significantly high (or heavy, in the case of p-value labeled edges) degrees. Such structures represent the effect of this locus on many transcripts and suggest the presence of a regulation element. When several SNPs reside in the same gene these are lumped together for this analysis. We further study the transcripts that are potentially regulated by a gene or associated to a SNP locus, e.g. seeking functional enrichment.
- Biclusters: a *bicluster* is a subgraph of G. We adapt the methods of Tanay *et al.*[10] to find significantly dense biclusters, in which the participating SNPs share significantly many common associations with the participating transcripts. When several SNP loci, especially when not in LD, associate with the same transcript, this may be evidence of this expression phenotype being a complex trait, affected by several genetic events. When the same set of loci commonly associate with a large number of transcripts we have cross-confirmation of the individual associations as well as a possible multi-locus effect on a pathway or a biological process.
- To further validate and better understand vertices of high degree and dense biclusters we test the sets of participating transcripts for functional enrichment.

We exemplify our methods by applying them to SNP-expression data collected from 50 breast cancer patients.[11] The data consist of 578 SNPs in selected genes from the reactive oxygen species (ROS) biochemical and signaling pathways, and expression levels of 3351 transcripts, in tumor biopsies.[11,13] We identified SNPs that are associated to a significantly large number of transcripts, such as *PRKCA* which was found to be associated with 85 transcripts at $p < 0.001$. Several significant biclusters were found in the data, one of them consisting of 4 SNPs (in *IL1B*, *IER3* and *NOX3,* which are related to stress response) and 84 transcripts. The GO term "cytosolic small ribosomal subunit" was significantly over-represented in this set of transcripts (Bonferroni corrected p-value 0.002) together with other related GO terms.

The purpose of this paper is to present and demonstrate analysis tools that can be used in identifying significant relationships between genetic variation and differential expression. To fully understand the biological meaning of significant structures more experiments or analysis need to be designed and performed.

2. Computational Methods and Statistical Modeling

2.1. *Computing the SNP-transcript association matrix*

Let N and M denote the number of SNPs and transcripts, respectively. For each pair (s, t) of SNP and transcript, we compute an association score and a corresponding p-value P_{st} using one of the methods described below based of expression and genotype measurements of k samples in the study. The resulting matrix P is called

the *association matrix*. Given a threshold p, we say that s is *associated* with t if $P_{st} < p$.

Quantitative mutual information score. For a SNP locus s, let G be a partition of samples induced by the genotypic values at locus s. For a transcript t with expression vector $q = [q_1, q_2, \ldots, q_k]$ and a threshold α, let C_α be a partition of samples defined by the $q_i < \alpha$ and $q_i \geq \alpha$. The *mutual information score* (*MIS*) is the difference between the entropy of the partition C_α and the conditional entropy of C_α given G: $MIS(C_\alpha, G) = H(C_\alpha) - H(C_\alpha \mid G)$, where H is the entropy function. We define the *quantitative mutual information score* (*QMIS*) to be the maximum possible MIS over all possible α between minimal and maximal expression of transcript t, i.e.

$$QMIS(C, G) = \max_{\min(q) \leq \alpha \leq \max(q)} MIS(C_\alpha, G).$$

A p-value for the mutual information score under a null model of uniform distribution of genotype patterns of the same mixture can be computed exactly by an exhaustive approach.[14] These p-values comprise the entries of the association matrix P.

ANOVA analysis. For each SNP locus and each transcript, we also computed one-way ANOVA p-value for the expression vector $[q_1, q_2, \ldots, q_k]$ and grouping of the samples based on SNP locus genotypes.[15] In this case, the entries of the association matrix P consist of p-values.

The two methods described above demonstrate possible parametric and nonparametric approaches to assess SNP-transcript association. ANOVA analysis assumes normal distribution of underlying measurements that may not hold for expression data in general. On the other hand ANOVA analysis is computationally efficient. QMIS does not make any distribution assumptions, but its p-value computation is rather slow, which limits its use in some of the applications. Alternatively, in cases where there is a concern about normality or uniform distribution assumptions for the null model, permutation based methods could be used to assess significance of SNP-transcript association.

2.2. *Bicluster identification*

A *bicluster* is a submatrix of the association matrix, that is, a subset of SNPs and transcripts. In graph terms, a bicluster is a subgraph of the association graph. The goal of the bicluster analysis is to identify significantly dense biclusters, in which the participating SNPs share significantly many common associations with the participating transcripts.

Let S be the set of SNPs and let T be the set of transcripts. To find biclusters in the association matrix P, we first transform P into a weighted bipartite graph $G = (S, T, E)$ as follows: For a given threshold p (e.g. 0.05) and any pair (s, t) of SNP-transcript, we define (s, t) to be an edge of weight $+1$ if $P_{st} < p$ and to be a non-edge of weight -1 otherwise. We now apply the biclustering algorithm of

Tanay *et al.*[10] to this graph, looking for the heaviest biclusters, where the weight of a bicluster is defined as the sum of the weights of its edges and non-edges. Briefly, the algorithm starts from small biclusters (complete bipartite subgraphs), serving as seeds, and expands them in a greedy fashion. The bicliques can be found efficiently using a hashing technique described in Tanay *et al.*[10] Each such seed is expanded iteratively, adding or removing a SNP or a transcript that contributes the most to the weight of the resulting bicluster, as long as this contribution increases the overall weight of the bicluster. The resulting biclusters correspond to dense subgraphs of the association graph, with edge density exceeding 0.5 due to the particular choice of edge weights.

2.3. *Assessment of statistical significance*

To assess the statistical significance of our findings, we compared our analysis results to those obtained on random data sets. Those were generated by randomly permuting the expression data, while leaving the genotype data intact. More precisely, in each instance of the random data samples were randomly assigned to expression data vectors. This randomization process ensures that we keep the structure of dependencies between SNP loci that exist in the original data, as well as between expression vectors. These permuted data were used in assessing overabundance of significant SNP-transcript association pairs as well as in assessing the significance of more complex structures.

2.4. *Visualization of biclusters*

We developed a method for visualizing a set of possibly overlapping biclusters by permuting the rows and columns of the association matrix so that rows and columns that belong to the same bicluster are ordered close to one another. To describe the algorithm we first note that the problem of determining whether there exist permutations, for which all biclusters appear consecutively in their rows and columns, can be reduced to two consecutive ones problems,[16] one in each dimension of the matrix. Even in the case of an imperfect solution (i.e. the consecutive property cannot be maintained) we shall assume that the two dimensions of the matrix are independent, and will optimize them separately.

For a given dimension d (rows or columns), our goal is to optimize over all biclusters the sum of their spreads, where the *spread* of a bicluster along dimension d is defined as the standard deviation of the locations of its elements along this dimension. We designed a greedy heuristic to this optimization problem that starts with a random permutation and iteratively tries to decrease its sum of spreads by exchanging pairs of elements.

2.5. *Analysis of functional enrichment*

Transcript sets that were associated to one or more SNPs were subjected to analysis of their functional enrichment based on their Gene Ontology (GO) annotations.

GO annotations for all genes were obtained using publicly available Biomolecule Naming Service (BNS),[17] http://openbns.sourceforge.net/, a high-speed directory service that resolves alias and official gene symbol differences, and links to publicly available databases. We say that gene g is *linked* to a GO term t, if t annotates g or is an ancestor of some GO annotation for g. For each GO term t, we tested whether t is overrepresented in any given set of genes using a hypergeometric distribution.[18,19]

3. Results

We analyzed SNP-expression data collected in the study of the effects of single nucleotide polymorphisms on genome-wide expression in tumors from breast cancer patients.[11] This data consisted of 725 SNPs in selected genes from the reactive oxygen species (ROS) biochemical and signaling pathways, and 8000 transcripts whose expression was measures in genome-wide study of tumor biopsies,[11,13] http://genome-www.stanford.edu/breast_cancer. 725 SNPs were divided among 203 genes distributed along all chromosomes with a wide spread of 1 to 19 SNPs per gene. Gene expression levels were measured on arrays manufactured at the Stanford Microarray Core Facility. Prior to the analysis, the expression data was filtered for signal quality; namely, we retained only transcripts whose ratio of feature intensity over background exceeded 1.5 in at least 80% of the experiments in each dye channel. 3351 transcripts passed this array signal filter. The SNP analysis was also restricted to 578 SNPs that had different genotypes among studied individuals. SNP and expression data was available for 50 breast cancer patients.

3.1. *Overabundance analysis of the association matrix*

We used the framework for computing associations between SNPs and genes to test whether the number of associations detected in real data is larger than the random expectation. This overabundance analysis was performed using two measures: false discovery rate (FDR)[20] and binomial surprise rate.[7] FDR measures the ratio of expected and observed numbers of SNP-transcript association pairs with a given score or better. Binomial surprise rate as defined in Ref. 7 measures the probability of seeing larger or equal number of pairs with a given score or better when uniformly drawing random permutations of genotype data and assuming transcript independence. The number of associations in both cases was determined by counting the number of entries in the association matrix less than or equal to a given threshold. The results of this analysis are depicted in Figs. 1 and 2. For QMIS, 769 SNP-transcript association pairs with p-values $\leq$ 1.0e-04 were observed with FDR of 0.2 and binomial surprise of 9349. In random data you may expect to find only 150 such pairs. In this case the expected number of pairs is calculated based on QMIS p-value and size of the association matrix.

For ANOVA scores, 571 SNP-transcript association pairs with p-values $\leq$ 1.0e-04 were observed with FDR of 0.6 based on expected number of pairs estimated from permuted data. Figure 3 shows several examples of SNP-transcript

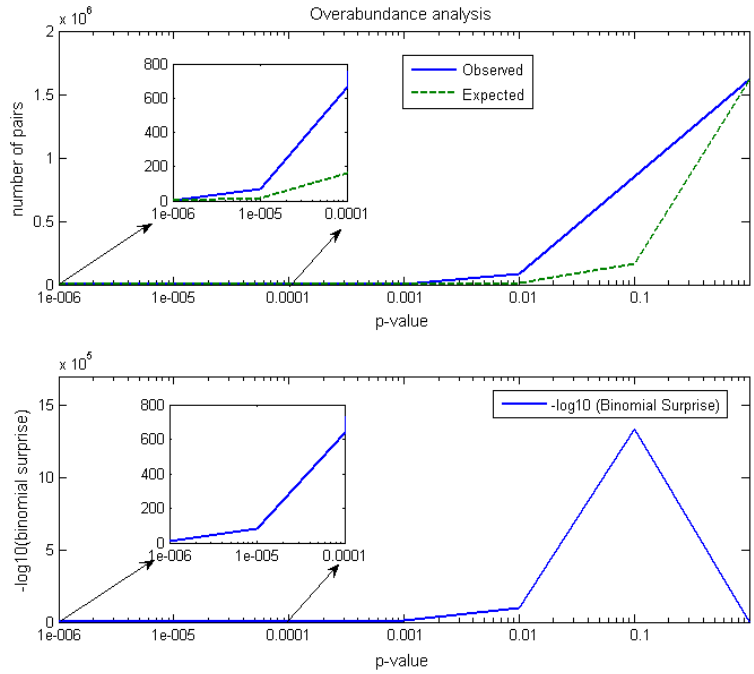

Fig. 1. Overabundance analysis for QMIS-based associations. The top plot shows comparison of distributions of observed and expected numbers of SNP-transcript pairs with a certain p-value or better. The bottom plot shows the corresponding $-\log 10$ (binomial surprise), see text and Ref. 7. Small insert plots show the same results restricted to p-values between 1.0e-06 and 1.0e-04.

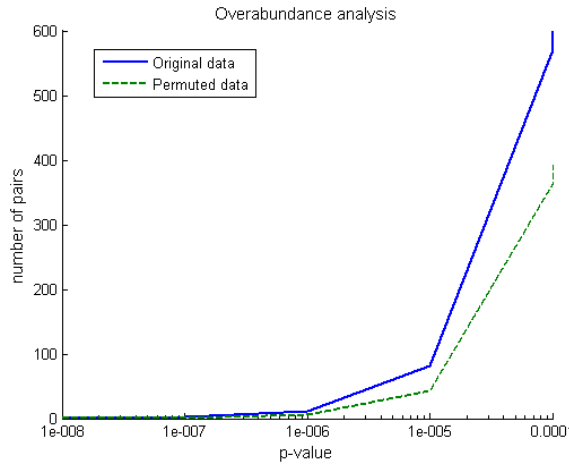

Fig. 2. Comparison of numbers of ANOVA-based association pairs in original and permuted data. The permuted data represents averages over 50 permuted data sets.

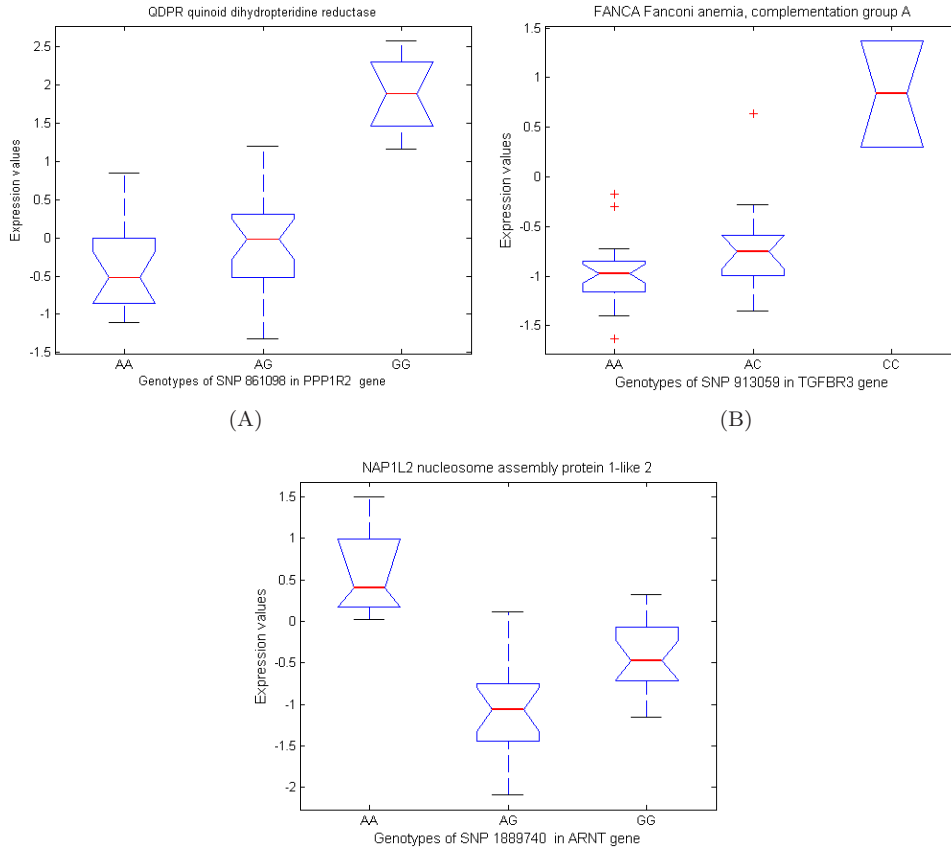

Fig. 3. Examples of SNP-transcript associations. (A) Difference between expression values in *QDPR* transcripts in samples with different genotypes at SNP rs 861098 (located within the gene *PPP1R2*). This association receives an ANOVA p-value of 2.1e-07. (B) Difference between expression values of the *FANCA* transcript in samples with different genotypes at SNP rs 913059 (in the gene *TGFBR3*); this association has an ANOVA p-value of 3.7e-06. (C) Difference between expression values of the *NAP1L2* transcript in samples with different genotypes at SNP rs 1889740 (in the gene *ARNT*); the association has an ANOVA p-value of 1.02e-06.

pairs with significant associations. Each graph shows a box-plot of expression data grouped by sample genotypes for each SNP. The boxes have lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the boxes to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. For SNPs in *PPP1R2* and *TGFBR3*, the presence of G or C alleles increases expression of *QDPR* and *FANCA* respectively. AA genotype in *ARNT* is associated with increased expression of the *NAP1L2* transcript.

3.2. *Master regulators*

Figure 4 depicts the differences between the total edge weights observed at the high end of the actual data and the corresponding results, averaged over 50 sets of simulated data (see Methods for simulation methodology). The total edge weight of SNP locus s is computed as sum $-\log(P_{st})$ over all transcripts t such that $P_{st} < 0.001$. SNP loci that have a set of edges with an exceptionally heavy weight are putative regulators of many transcripts. They potentially affect the expression levels or the mode of operation of transcription factors or of RNA binding proteins, directly or indirectly. Thus, these SNPs affect the transcription or degradation rates and hence expression levels of many transcripts. SNPs that had the highest weight belong to the genes *PRKCA*, *CYP2C19*, *IFG1R*, *IGF2R* and *XDH*. Genes *PRKCA*, *IGF1R*, *IGF2R* are part of large regulatory networks and potentially effect of many transcripts, *CYP2C19* and *XDH* are metabolizing enzymes and the biological explanation to the observed associations is less clear.

To further understand the association of SNPs to sets of transcripts we are interested in common properties of the sets of transcripts described above. To this end we performed GO enrichment analysis of these sets as described in the Methods. We considered sets of transcripts associated to SNPs that belong to the same gene, and looked for overrepresented GO terms in these sets. Several examples of our findings: 70 transcripts associated with gene *GCLM* contained all five transcripts that belong to term "fibrillar collagen" (Bonferroni corrected p-value 0.0002).

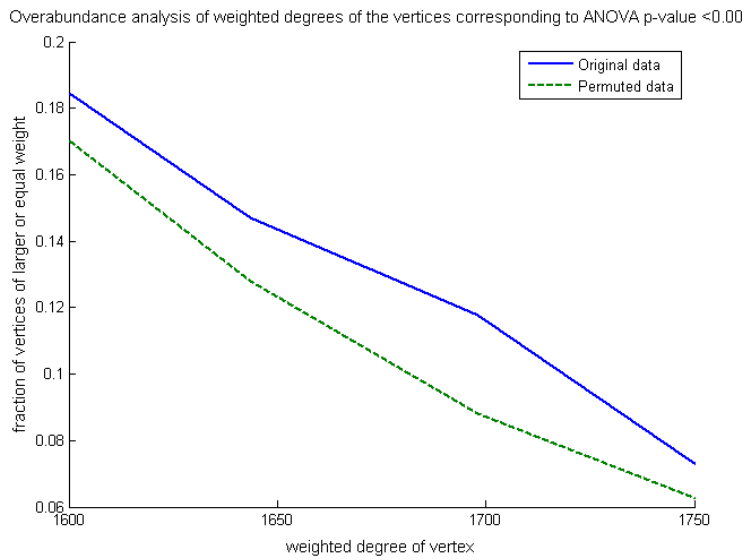

Fig. 4. Comparison of vertex weights in original and permuted data. For each SNP locus s, total edge weight was computed as sum $-\log(P_{st})$ over all transcripts t such that $P_{st} < 0.001$. For permuted data, 50 instances of randomized expression data (as described in Sec. 2.3) were used in the calculation of weighted vertex degree in permuted plot.

90 transcripts associated with gene *PPP3CA* contained 6 out of 12 transcripts that belong to term "antigen processing" and "antigen presentation" (Bonferroni corrected p-value of 0.002 and 0.005 respectively). Transcripts associated with genes *IER3* and *NOX3* included all three members of the term "cytosolic small ribosomal subunit" (Bonferroni corrected p-value of 0.008 and 0.04) as well as members of other groups related to ribosome functions (Fig. 6). As *IER3* and *NOX3* are part of the stress response system our findings are consistent with stress response activation of protein synthesis and how it may be affected by genetic variants in these three genes.

3.3. *Bicluster analysis*

In addition to the overabundance analyses described above, we conducted several bicluster analyses of the association matrix, searching for these more complex structures in the data. First, we analyzed the association graph obtained from a QMIS-based association matrix using a p-value threshold of 0.02. Our algorithm identified 28 biclusters in the graph whose size (product of number of SNPs and number of transcripts) exceeded 32. We analyzed the gene set contained in each bicluster for functional enrichment. One of the biclusters had four SNPs in genes *IL1B*, *IER3* and *NOX3* and 84 transcripts. The GO term "cytosolic small ribosomal subunit" was significantly overrepresented in this set of transcripts (Bonferroni corrected p-value of 0.002) together with other related GO terms. Figure 6 shows the part of the GO cellular component hierarchy connecting the term "cytosolic small ribosomal subunit" to the root of the hierarchy. Node sizes are proportional to the significance of the corresponding GO terms. Figure 7 shows a visualization of the biclusters that were identified in the QMIS analysis.

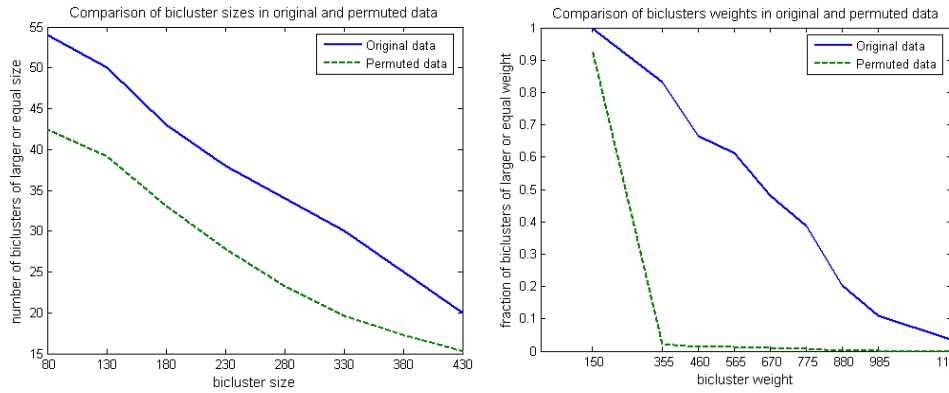

Fig. 5. Comparison of bicluster sizes and bicluster weights in original and permuted data. Biclusters in original and permuted data were identified using the corresponding association matrices computed using ANOVA p-values ≤ 0.01. Results for the permuted data represent averages over 50 random data sets.

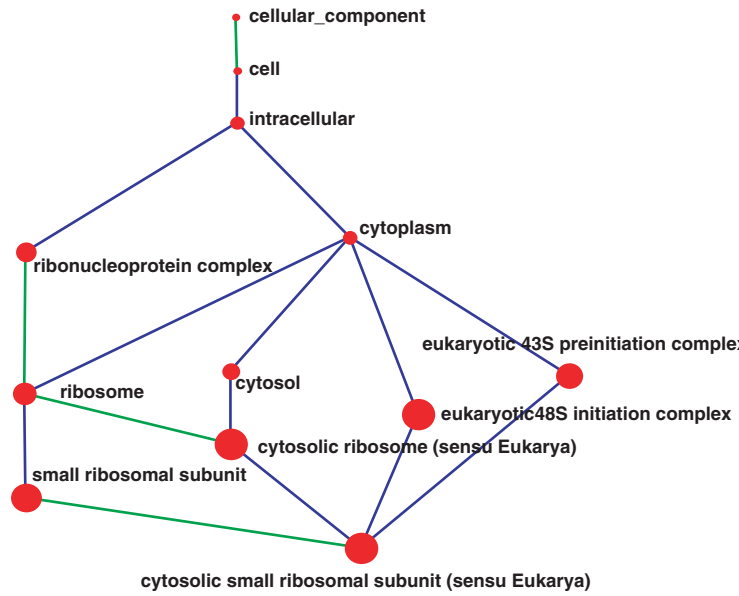

Fig. 6. Part of the cellular component hierarchy connecting overrepresented terms in the set of transcripts associated with SNPs in genes *IL1B*, *IER3* and *NOX3* to the root of the hierarchy. Each node correspond to GO term, sizes of the nodes are proportional to term significance.

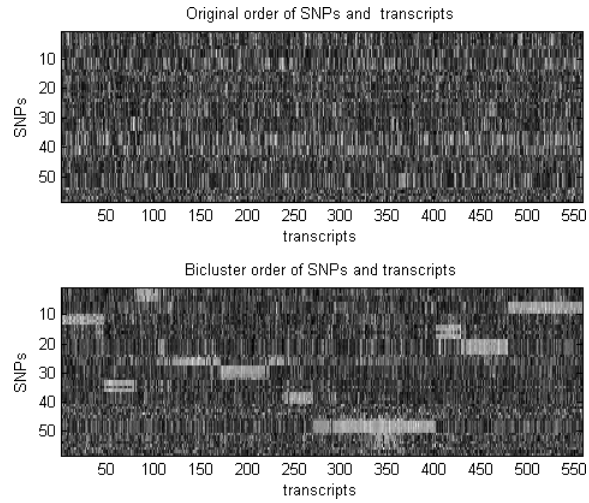

Fig. 7. Visualization of the association matrix computed using the QMIS-based analysis. Top plot shows a subset of SNPs and a subset of transcripts in their original order in the data. Each SNP and transcript shown belongs to one of the biclusters. Bottom plot shows the same subset of SNPs and transcripts reordered to highlight biclusters identified in the data.

We also analyzed an ANOVA-based association matrix with a threshold of 0.01, which corresponds to the best 1% of scores observed in the data. In total, we identified 54 biclusters of size exceeding 32. Since ANOVA scores are efficient to compute, we were able to simulate association matrices for random data (see Methods) and evaluate the significance of the biclusters detected. To this end, we associated with each bicluster B a weight $W(B)$ that measures the extent to which the SNPs in the bicluster are associated with the genes in the bicluster. $W(B)$ is defined to be the sum over all pairs of SNP s and transcript t in the bicluster B of $(m - \log(P_{st}))$, where $m = \frac{1}{NM} \sum_{s \in g, t \in T} \log P_{st}$.

The significance of a bicluster is then estimated as the probability of observing a bicluster with score $W(B)$ or better in the permuted data. For 22 of the 54 biclusters identified in the ANOVA-based association matrix, the probability of observing a bicluster of this weight in permuted data was less than 0.01. Figure 5 shows a comparison of bicluster sizes and weights in the original and permuted data.

4. Conclusions

We present a framework for the integrated analysis of genotype and expression data. The analysis relies on the computation of an association matrix whose entries represent the extent to which each SNP is correlated with each transcript. We develop methods for the performing statistical overabundance analysis of the entire matrix and of each of its rows or columns. We also adapted methods for identifying high-scoring biclusters in this matrix, representing a set of SNPs that share significantly many associations. We applied our framework to analyze SNP-expression data from breast cancer patients, discovering novel gene sets that may be associated with the disease. Possible extensions to our method include (1) improved weighting schemes to the entries of the association matrix; (2) a more refined bicluster analysis that does not require the discretization of the association data; and (3) methods for further evaluating the identified gene sets, e.g., by using additional expression data from independent sources.

Acknowledgments

Roded Sharan was supported by an Alon fellowship. Vessela Kristensen, Hege Edvardsen and Anne-Lise Børresen-Dale were supported by the Norwegian Cancer Society Grants D-99061 and D-03067, National Program for Research in Functional Genomics Grant 152004/150, Norwegian Research Grant 155218/300, and a Swiss Bridge Award. Hege Edvardsen is a Fellow of the Norwegian Cancer Society.

References

1. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast, *Science* 26,296(5568), p. 7525, 2002.

2. Jin W, Riley RM, Wolfinger RD, *et al.*, The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster, *Nature Genet* **29**(4):389–395, 2001.
3. Schadt EE, Monks SA, Drake TA, *et al.*, Genetics of gene expression surveyed in maize, mouse and man, *Nature* **422**(6929):297–302, 2003.
4. Sandberg R, Yasuda R, Pankratz DG, *et al.*, Regional and strain-specific gene expression mapping in the adult mouse brain, *Proc Natl Acad Sci USA* **97**:11038–11043, 2000.
5. Hedenfalk I, Duggan D, Chen Y, *et al.*, Gene-expression profiles in hereditary breast cancer, *N Engl J Med* **344**(8):539–548, 2001.
6. Hedenfalk I, Ringner M, Ben-Dor A, *et al.*, Molecular classification of familial non-BRCA1/BRCA2 breast cancer, *Proc Natl Acad Sci USA* **100**(5):2532–2537, 2003.
7. Ben-Dor A, Friedman N, Yakhini Z, Class discovery in gene expression data, *Proc Fifth Int Conf Computational Biol. RECOMB'01*, pp. 31–38, 2001.
8. Morley M, Molony C, Weber T, *et al.*, Genetic analysis of genome-wide variation in human gene expression, *Nature*, **430**, 2004.
9. Watts JA, Morley M, Burdick JT, *et al.*, Gene expression phenotype in heterozygous carriers of ataxia telangiectasia, *Am J Hum Genet* **71**(4):791–800, 2002.
10. Tanay A, Sharan R, Kupiec M, Shamir R, Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genome-wide data, *Proc Natl Acad Sci USA* **101**:2981–2986, 2004.
11. Kristensen V, Edvardsen H, Tsalenko A, *et al.*, Genetic variation in putative regulatory loci controlling gene expression in breast cancer, to appear in *PNAS*.
12. Sørlie T, Perou CM, Tibshirani R, *et al.*, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proc Natl Acad Sci USA* **98**:10869–10874, 2001.
13. Sørlie T, Tibshirani R, Parker J, *et al.*, Repeated observation of breast tumor subtypes in independent gene expression data sets, *Proc Natl Acad Sci USA* **100**:8418–8423, 2003.
14. Tsalenko A, Ben-Dor A, Cox N, Yakhini Z, Methods for analysis and visualization of SNP genotype data for complex diseases, *Pac Symp Biocomput* **8**:548–561, 2003.
15. Rice J, *Mathematical Statistics and Data Analysis*, Duxbury Press, 1995.
16. Setubal J, Meidanis J, *Introduction to Computation Molecular Biology*, Boston: PWS Publishing Company, 1997.
17. Kincaid R, Kleusing D, Vailaya A, BNS: An LDAP-based Biomolecule Naming Service, *OiBC*, Washington DC, 2002.
18. Gao F, Foat B, Bussemaker H, Defining transcriptional networks through integrative modeling of mRNA expression and transcriptional factor binding data, *BMC Bioinformatics* **5**:31, 2004.
19. Martin D, Brun C, Remy E, *et al.*, GOToolBox: Functional analysis of gene datasets based on Gene Ontology, *Genome Biol* **5**:R101, 2004.
20. Benjamini Y, Hochberg Y, Controlling the false discovery rate- a practical and powerful approach to multiple testing, *J Roy Stat Soc B Met* **57**(1):289–300, 1995.

Anya Tsalenko is an Expert Scientist in Agilent Laboratories, working in the area of computational biology. Her research is focused on analysis of gene expression, array CGH and SNP genotype data generated in various complex disease studies. Prior to joining Agilent, Anya was a post-doc at the University of Chicago. Anya has a Ph.D. in Mathematics from Stanford University.

Roded Sharan is a Senior Lecturer at the School of Computer Science, Tel Aviv University (since 2005), working in the areas of bioinformatics and systems biology. Prior to joining Tel Aviv University he was a post-doc at the University of California, Berkeley. He has a Ph.D. in Computer Science from Tel Aviv University, Israel.

Vessela Kristensen is a Professor in the Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Centre. Professor Kristensen leads cancer genome variation group, working on different projects related to how genetic variation affects occurrence of somatic alterations, gene expression patterns and genome wide copy number alterations in human breast and ovarian tumors.

Hege Edvardsen is a Ph.D. student in the Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Centre. Under the supervision of Professor Vessela N. Kristensen and Professor Anne-Lise Børresen-Dale she studies how genetic factors influence the response to radio- and chemotherapy in breast cancer patients and to what extent they affect the level of adverse side effects of treatment.

Anne-Lise Børresen-Dale is a Professor in the Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Centre. Professor Børresen-Dale leads molecular genetics studies of breast and ovarian cancer group, working on the molecular biology of breast and ovarian cancer with emphasis on identification of genotypes and gene expression profiles contributing to elevated cancer risk, radiation sensitivity, tumor aggressiveness and therapy resistance.

Amir Ben-Dor is a Master Scientist in Agilent Laboratories, working on developing data workflow, analysis and visualization tools for highly multiplexed biological measurements, e.g. array CGH and gene expression. Prior to joining Agilent, Amir was a post-doc at the University of Washington. He has a Ph.D. in Computer Science from The Technion, Israel.

Zohar Yakhini is doing active research in computational biology focusing on algorithmic and statistical aspects of genomic and proteomic measurement techniques, including assay design and data analysis. He is a Master Scientist in Agilent Laboratories and a Visiting Professor at the Technion CS Department. He has a Ph.D. in Mathematics from Stanford University.