A COMBINATORIAL APPROACH TO THE ANALYSIS OF DIFFERENTIAL GENE EXPRESSION DATA

The Use of Graph Algorithms for Disease Prediction and Screening[*]

Michael A. Langston[1], Lan Lin[1], Xinxia Peng[2], Nicole E. Baldwin[1], Christopher T. Symons[1], Bing Zhang[3] and Jay R. Snoddy[3]

[1]Department of Computer Science, University of Tennessee, Knoxville, TN 37996-3450; [2]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996-0845; [3]Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6124.

Abstract: Combinatorial methods are studied in an effort to gauge their potential utility in the analysis of differential gene expression data. Patient and gene relationships are modeled using edge-weighted graphs. Two somewhat orthogonal algorithms are devised and implemented. One is based on finding optimal cliques within general graphs, the other on isolating near-optimal dominating sets within bipartite graphs. A main goal is to develop methodologies for training algorithms such as these on patient populations with known disease profiles, so that they can then be employed to classify and predict the likelihood of disease in patient populations whose profiles are not known in advance. These novel strategies are in marked contrast with Bayesian and other well-known techniques. Encouraging results are reported.

Key words: Combinatorial Methods; Discrete Mathematics; Disease Prediction and Screening; Graph Algorithms; Graph Theory; Microarray Analysis.

1. INTRODUCTION

A fundamental problem in cancer treatment is early and reliable detection. Identification of a set of genes whose expression levels serve as an accurate discriminator among normal and cancerous tissue samples would not only represent significant progress towards developing more reliable cancer diagnosis protocols, but might also identify novel therapeutic targets. With this motivation in mind, we investigate the hypothesis that only a modest number of genes may suffice for this task. We seek to develop algorithms and software for this purpose, and introduce a graph theoretical method of differential gene expression analysis. The goals of this method are to identify a set of genes useful in discriminating among tissue samples, and to use these genes in disease prediction and screening.

One of the important features of our algorithms is the computation of discrimination scores for each gene represented in an input microarray. These scores estimate a gene's relative ability to distinguish among sample tissue classes. We then select the highest-scoring genes, and use them to calculate a pairwise similarity metric between patients' tissue sample expression profiles. Genes that fail to discriminate among a defined percentage of the samples are eliminated using a dominating set algorithm as a high pass fil-

[*] This research is supported in part by the National Science Foundation under grants EIA-9972889, CCR-0075792 and CCR-0311500, by the Office of Naval Research under grant N00014-01-1-0608, by the National Institutes of Health under grant U01-AA013512-02, by the Department of Energy under contract DE-AC05-00OR22725, and by the Tennessee Center for Information Technology Research under award E01-0178-261.

ter. With this information, we construct a complete weighted graph, in which the vertices represent the tissue samples and the edges are weighted by the similarity metric between sample vertices. A user-defined threshold is next used to transform the complete weighted graph into an incomplete unweighted graph. The combination of these tools produces some very encouraging predictive results.

In the sequel, we describe the datasets we have chosen to study, the algorithms we have devised, and the results we have obtained. We also draw some conclusions from this effort.

2. DATA EMPLOYED

We use the Harvard [5], Michigan [4], and Stanford [10] datasets in this study. We do not include the Ontario dataset due to a lack of overlap in annotated genes with the other datasets. Since the log-expression image plots for Samples L54, L88, L89 and L90 in the Michigan dataset show large, round dark spots at the center of the arrays [13] indicative of poor data quality, they are removed from the dataset. This leaves us with 92 samples from the Michigan dataset. Because the Harvard and Michigan datasets were generated by different institutes using different Affymetrix array types (HG_U95A and HUGeneFL, respectively), the distributions of the two datasets may not be comparable. Thus, we choose to normalize the two datasets separately. The log-scale quantifications of the gene expression levels for each probe set are obtained by robust multi-array average (RMA) [15] using Bioconductor.

Since we intend to train and test our algorithms on different datasets, we need a mapping schema among the different datasets. However, the three datasets come from different array platforms using different gene identifiers; hence, direct mapping is not possible. We choose to use LocusLink IDs (LL_IDs) for gene mapping, because the NCBI LocusLink Database is both relatively reliable and stable. For the Harvard and Michigan datasets, we map each probe set ID to its corresponding LL_ID using array annotation files from Affymetrix. For the Stanford dataset, we map each UNIGENE ID to its corresponding LL_ID using our local database, GeneKeyDB. To construct a gene expression summary for each LL_ID, we average the values within each sample across the original gene identifiers that map to a common LL_ID. The final datasets used in this study include: the Harvard dataset, which has expression profiles for 8509 unique genes among 254 samples; the Michigan dataset, which has expression profiles for 4985 unique genes among 92 samples; and the Stanford dataset, which has expression profiles for 8829 unique genes among 73 samples.

3. A CLIQUE-BASED STRATEGY

3.1 The Clique Problem

Clique is a well-known *NP*-complete problem, and is typically formulated as in [11]:

Input: A graph $G=(V,E)$ and a positive integer $k \leq |V|$.

Question: Is there a subset $V' \subseteq V$ for which $|V'| \geq k$ and such that every pair of vertices in V' is joined by an edge in E.

Clique is rapidly becoming recognized for its relevance in bioinformatics. In our own work, for example, we use clique in the following ways. In [2], we devise and apply fast parallel algorithms for clique to extremely large microarray datasets in an effort to help identify putatively co-regulated genes in murine neural regulatory networks. In another application [3], we employ high performance implementations of clique in the study of *cis*-regulatory elements to discover putative motifs.

3.2 Scoring Method

Our goal in training is to develop graph-theoretic tools to help distinguish among sample groups (such as normal and adenocarcinoma). Ideally, we hope to be able to construct an unweighted graph in which edges connect mainly members of the same group. At that point, clique analysis would be an attractive approach for testing our methods against additional data.

In order to pinpoint a modest number of genes out of thousands from the original dataset, our first step in training is to determine which genes appear to discriminate best among sample types. To accomplish this, a discrimination score is calculated for each gene. Only the best genes (those with the highest scores) are retained for subsequent steps. Since the distributions of the expression values of these genes would be expected to be bimodal with respect to two distinct sample classes, the differences between class medians give us a general measure of the difference of expression between two classes. Subtracting the sum of the standard deviations of a gene within each group allows us to eliminate, or at least diminish, the importance of any gene whose expression levels vary excessively.

The data is obtained as in Section 2 as an $n \times m$ matrix, A, of expression values. Rows represent test samples, and columns denote genes. Our algorithm, as applied to discrimination between two sample groups, can be described in pidgin ALGOL as follows:

```
procedure gene-score-and-select
for j=1 to m
  normalize expression values in column j to the range [0, 1]
  compute median expression value (v_j) and standard deviation
     (σ_j) on group 1 sample data for gene j
  repeat computation on group 2 sample data for gene j
  set score(gene j) = | v_j(group 1) – v_j(group 2) | – | σ_j(group 1) +
     σ_j(group 2) |
delete genes with scores not exceeding some lower limit
return remaining genes and their scores
```

When training on the Michigan dataset in order to learn to distinguish between normal (group 1) and adenocarcinoma (group 2) samples and using a lower limit of zero, this procedure delivers a collection of 105 genes for further evaluation.

An assignment of inter-sample weights can help demonstrate the degree to which these genes and their respective scores delineate normal samples from adenocarcinoma. Here, the weight between samples i and j represents the degree of similarity in their respective expression profiles. We compute this weight as a sum over all genes selected in the previous step, because it is these genes that seem to have the greatest potential to serve as good discriminators. Accordingly, we set *weight(i,j)* to:

$$\sum \text{score}(\text{gene}_k) \bullet (1 - \left| \text{expression\_value}_{ik} - \text{expression\_value}_{jk} \right|)$$

As is shown in Figure 1, higher-weighted sample pairs tend to be homogenous. That is, either both tissue samples are normal or both are adenocarcinoma. Conversely, lower-weighted pairs tend to be heterogenous, where one sample is normal and the other is adenocarcinoma. While this seems to confirm our gene scoring and selection procedure, other scoring approaches appear to be viable as well. Therefore, we investigated several other alternatives before settling on this approach.

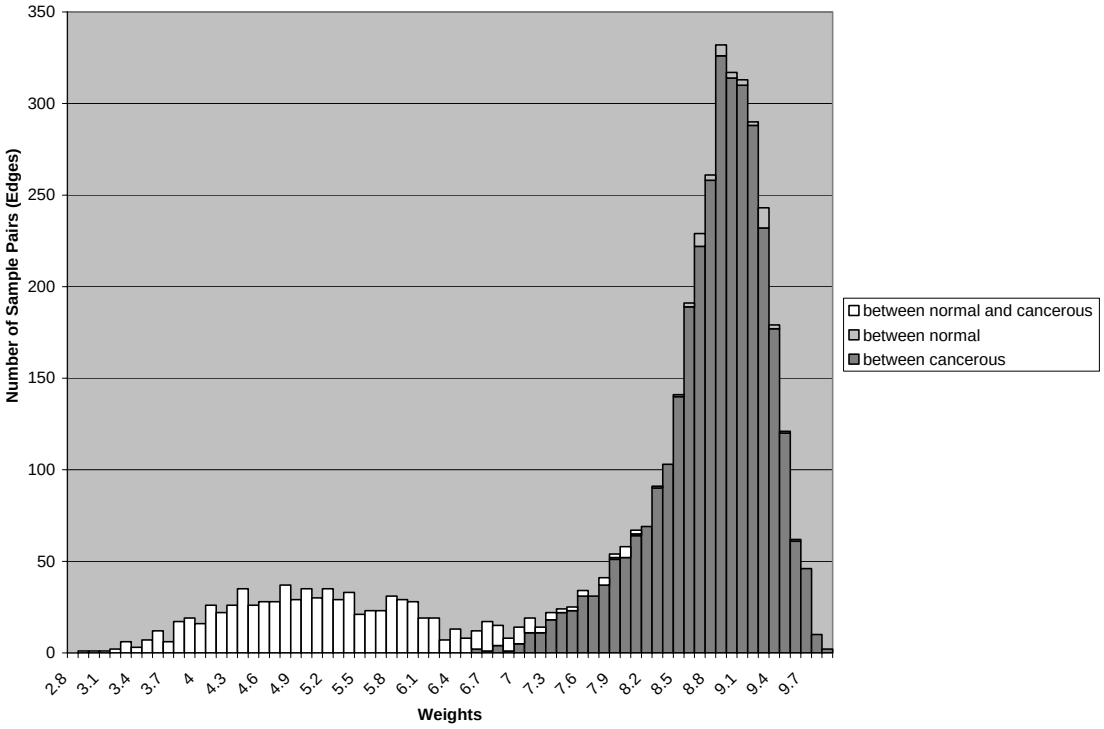

Figure 1. Weights between sample pairs using 105 genes from the Michigan dataset

Two of these alternative approaches are worthy of note in the computation of gene discrimination scores. One is the elimination of outliers before computing the scores, which is motivated by the fact that outliers might affect both the median and the standard deviation. To test this, we modified our approach by adding a screening phase, in which we first compute the medians and the standard deviations for each gene within each group, then check the expression values corresponding to that gene, discarding those at least three standard deviations away from the group median. We subsequently re-compute all medians and standard deviations using only the retained values. We describe this modified algorithm in pidgin ALGOL:

```
procedure gene-score-and-select2
for j=1 to m
  normalize expression values in column j to the range [0, 1]
  compute median expression value (v_j) and standard deviation
    (σ_j) on group 1 sample data for gene j
```

repeat computation on group 2 sample data for gene j
for i=1 to n
if sample i belongs to group 1
if its expression value (v_{ij}) satisfies $|v_{ij} - v_j(\text{group 1})| \geq 3\sigma_j(\text{group 1})$, delete v_{ij}
if sample i belongs to group 2
if its expression value (v_{ij}) satisfies $|v_{ij} - v_j(\text{group 2}| \geq 3\sigma_j(\text{group 2})$, delete v_{ij}
recompute median expression value (v_j) and standard deviation (σ_j) on group 1 sample data for gene j
repeat computation on group 2 sample data for gene j
set score(gene j) = $|v_j(\text{group 1}) - v_j(\text{group 2})| - |\sigma_j(\text{group 1}) + \sigma_j(\text{group 2})|$
delete genes with scores not exceeding some lower limit
return remaining genes and their scores

This modification does not appear to alter our original results appreciably, as illustrated in Figure 2.

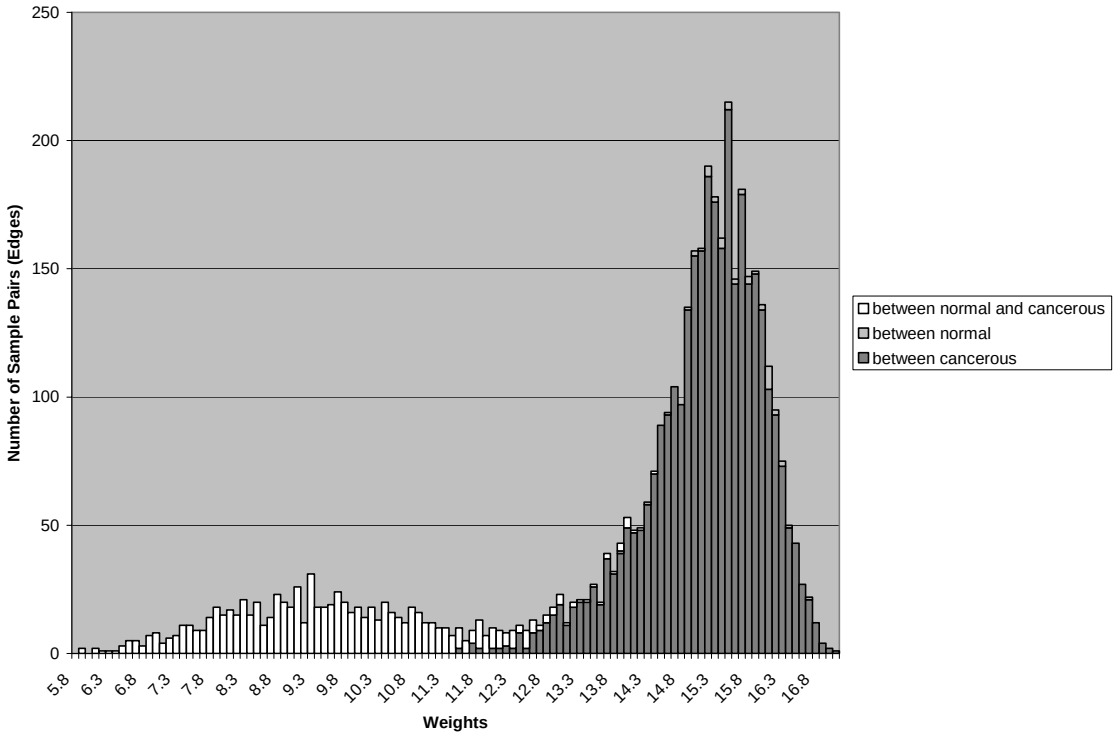

Figure 2. Weights between sample pairs after eliminating outliers using 105 genes from the Michigan dataset

The other noteworthy gene-scoring approach involves changing our original scoring function to a variant of the t-test function, a standard statistical measurement of population similarity. This test is realized using division as formulated in pidgin ALGOL:

procedure gene-score-and-select3
for *j=1* to *m*
  normalize expression values in column *j* to the range [0, 1]
  compute median expression value (v_j) and standard deviation (σ_j) on group 1 sample data for gene *j*
  repeat computation on group 2 sample data for gene *j*
  set score(gene *j*) = $|v_j(\text{group 1}) - v_j(\text{group 2})| \; / \; |\sigma_j(\text{group 1}) + \sigma_j(\text{group 2})|$
delete genes with scores not exceeding some lower limit
return remaining genes and their scores

As before, the results using the modified scoring function do not appear to improve upon our original results (Figure 3). We also experimented with Pearson's Correlation Coefficients and Spearman's Rank Correlation Coefficients, two popular methods of weighting. Neither of these methods were helpful. In fact, neither even revealed the bi-modal distribution we observed using our weight function.

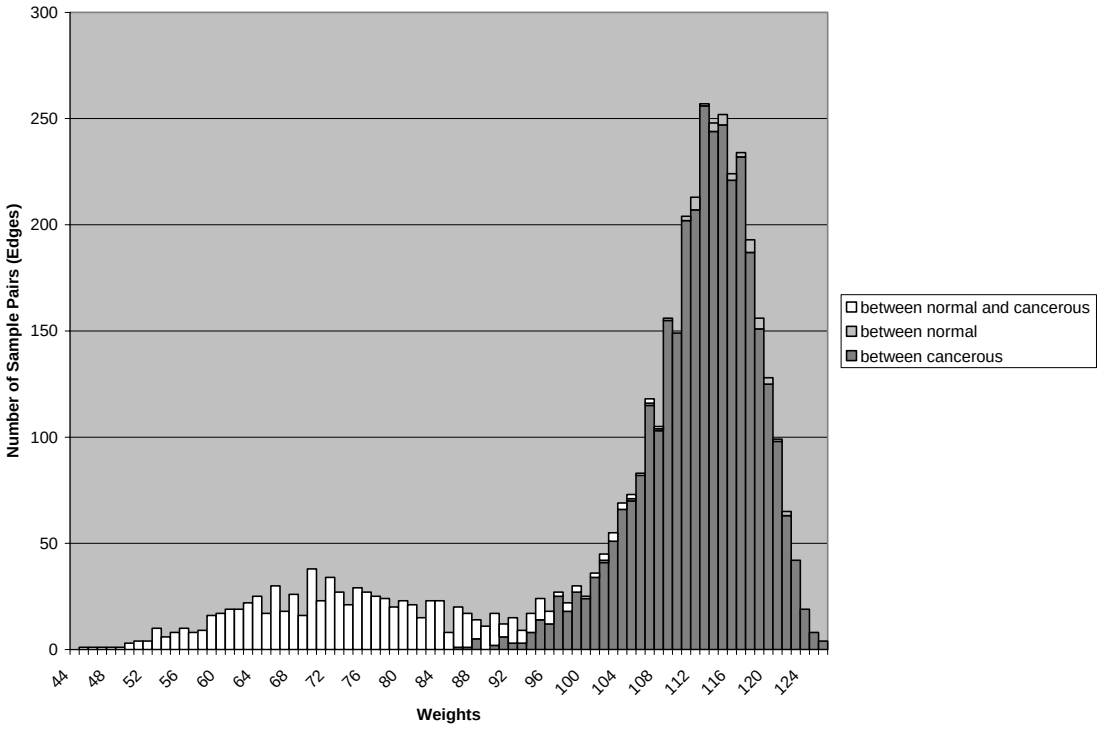

Figure 3. Weights between sample pairs using division-based scoring and 105 genes from the Michigan dataset

In addition to confirming the validity of our approach, Figure 1 also suggests an initial threshold weight below which we delete edges in a subsequent step (to be described shortly). Call this threshold *T*. For example, based on the figure, we choose as a somewhat informed but still rather arbitrary starting value *T*=7.6. We use our restricted set of genes to build an edge-weighted graph. In this graph, samples are represented by vertices and the weight of an edge between a pair of samples is set using the simple summation formula already described. Any edge whose weight is less than *T* is removed. The re-

sulting unweighted graph is then searched for all maximal cliques. Our aim is to train our codes so that we can find appropriately-sized cliques to cover both all groups, while minimizing cliques that overlap these groups. This requires iteration, as detailed in the pidgin ALGOL algorithm:

```
procedure clique-analysis
initialize edge-weighted graph of order n
  for i=1 to n
     for j = 1 to n
        set the weight of each edge
for a user-specified number of iterations do
  use T to delete edges with low weight
  find in resulting undirected graph all maximal cliques, C
  analyze C to refine the choice of T
return T
```

Because we know which samples are normal and which are adenocarcinoma in the Michigan dataset, we are able to iterate our method until we have a reasonable set of covering cliques. The optimal threshold seems to be centered at around T=8.1. We are not completely satisfied, however, with the lingering presence of overlapping cliques. Additional experimentation with gene cutoff scores seems to indicate that the presence of genes with low scores is problematic. But neither raising the cutoff score nor additional modification of the threshold is of much use. What seems missing in our estimates of gene discrimination is a way to determine which genes impact the greatest number of samples. For this, we turn to another graph metric, dominating set.

4. REFINEMENT VIA DOMINATING SET

4.1 The Dominating Set Problem

Dominating Set, another well-known *NP*-complete problem, can be stated as follows.

Input: A graph $G=(V,E)$ and a positive integer $k \leq |V|$.
Question: Is there a subset $V' \subseteq V$ for which $|V'| \leq k$ and every vertex $v \in V - V'$ is joined to a vertex in V' by an edge in E.

Using the theory of fixed-parameter tractablility (FPT) [8], dominating set may be even more difficult than clique. This is because clique is W[1]-complete and can be solved using graph complementation and vertex cover. Practical, efficient kernelization techniques are known for vertex cover [1]. The same, however, may not hold for dominating set. In fact the dominating set version we address here is nonplanar red/blue dominating set, which is W[2]-complete. Although its complement problem is FPT, there are currently no practical kernelization techniques known for it. Thus, we will only approximate solutions to dominating set.

4.2 Scoring Method

We first assume a normal distribution of the expression values of each gene, and estimate for it the mean and standard deviation. We do this separately for each of the sample groups. Then, based on the estimated normal distribution, we calculate the p-values for the original individual expression values. It is perhaps easiest to formulate our approach by constructing a bipartite graph. In this graph, one set of vertices represents the genes, and the opposing set represents the samples. We place an edge between a gene and a sample if and only if the p-value of the expression value corresponding to that gene-sample combination is greater than 0.05. Following statistical convention, we consider a p-value below this cutoff to indicate an outlier.

In this setting, we want to identify the genes that dominate (or nearly dominate) all the samples. Therefore, we winnow out from consideration any gene vertex not adjacent to at least 90% of the sample vertices. For example, in the Michigan dataset, a gene is eliminated if it is connected to fewer than 74 of the adenocarcinoma samples or fewer than nine of the normal samples. The choice of 90% is arbitrary; it was selected only after extensive testing.

Next, in an effort to remove any remaining genes with a low possibility of discriminating between the two groups, we calculate the p-values for tests of equal means using both the Wilcoxon and t-test methods. We use both since the t-test assumes a normal distribution, while the Wilcoxon test does not. Only genes for which both p-values are less than 0.05 are retained.

For those genes that remain, we generate scores based on the previously calculated p-values from the Wilcoxon tests. We then filter out genes using an adjusted p-value cutoff by means of the Bonferroni method. Specifically, we choose a significance level of $\alpha = 0.01$ and only keep genes with a p-value less than α/N, where N is the total number of genes we begin with at this step. Since a smaller p-value indicates a greater probability that the groups' expression values are different for a given gene, we use -log10(p-value) for the gene score.

```
procedure dominating-set-winnow
initialize edge-weighted bipartite graph of order n+m
  for i=1 to m
     for j = 1 to n
        determine the p-value (weight) of each edge(i,j)
set threshold to 0.05 and eliminate edges of low weight
delete genes that dominate < 90% of cancer samples
delete genes that dominate < 90% of normal samples
n = n - |deleted genes|
for i=1 to n
  generate p-value of equal mean using Wilcoxon and t-test
delete genes with p-value greater than 0.05 for either test
n = n - |deleted genes|
delete genes with p-value greater than or equal to 0.01/n
n = n - |deleted genes|
for i=1 to n
   set gene score to -log10(p-value)
return remaining genes and their scores
```

Finally, and most importantly, we compute the intersection of the genes identified by the clique-based approach described in the last section with the genes chosen by the dominating set method as described in this section. We are left with a set of genes that have passed both the clique and the dominating set tests. We find that this refinement of our gene lists gives us improved results in the testing phase of our experiments.

5. RESULTS

Having completed the training phase, we proceed to testing on a new dataset under the assumption that we will not know sample classification in advance. We evaluate our approach with the following three experiments. First, we trained on the Michigan dataset as explained in section 3 in order to learn to distinguish between normal and adenocarcinoma samples. We proceed to test our ability to classify samples on the Harvard dataset. Second, we reverse this process, applying our training algorithms to the Harvard dataset to distinguish between cancerous and normal samples. We test our method on the Michigan dataset. Third, we train on the Harvard dataset to learn to separate adenocarcinoma from squamous samples, testing on the Stanford dataset.

5.1 Experiment One

Clique-based training on the Michigan dataset identifies 105 genes that distinguish between adenocarcinoma and normal samples. Our dominating- set-based refinement reduces this to 84 genes, 78 of which are available in the Harvard data. Functional classification of the selected 84 genes was performed using the web-based functional profiling tool Gene Ontology Tree Machine (GOTM) [20]. The results are shown in Figure 4.

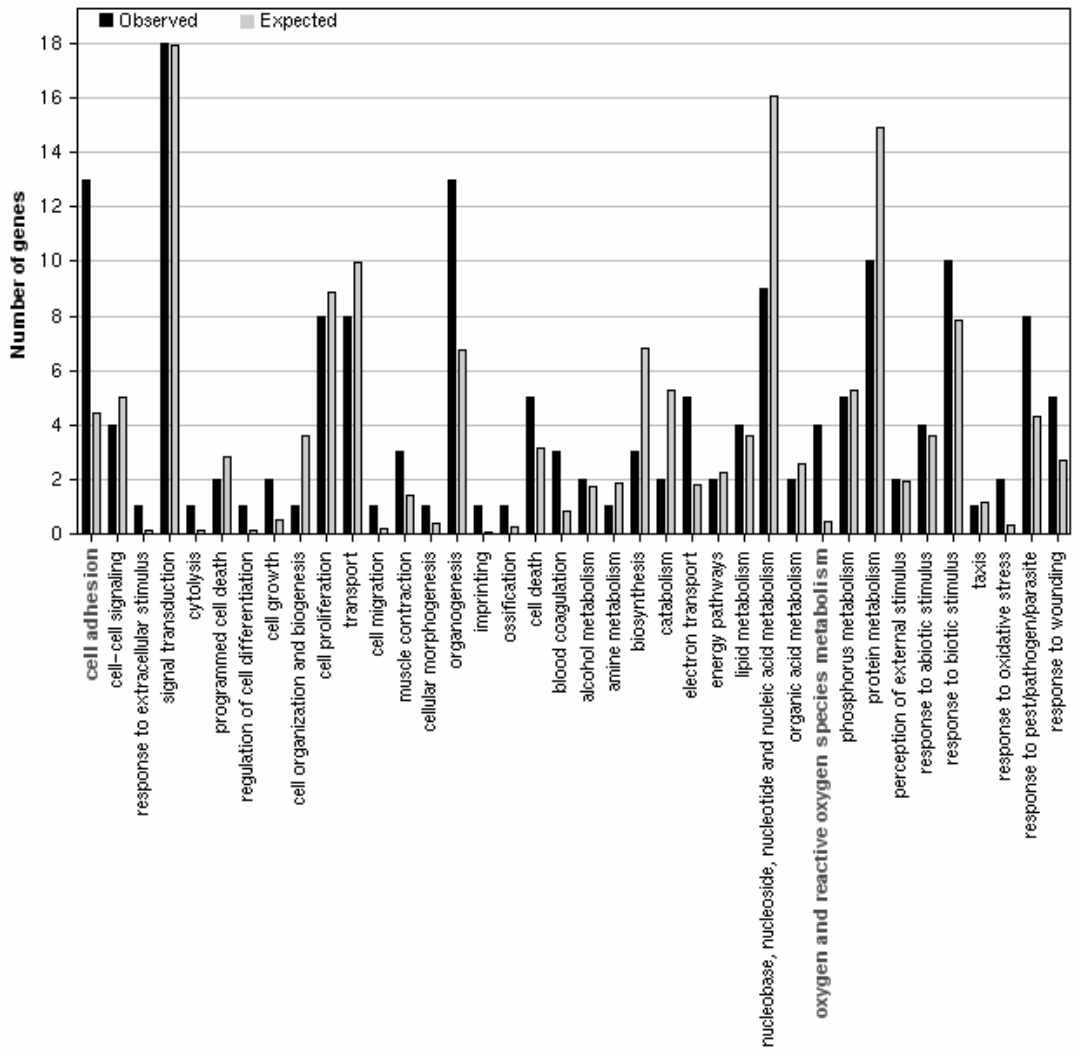

Figure 4. 84 genes (Michigan data) categorized under gene ontology. Black bars represent observed gene numbers. White bars represent expected gene numbers in the categories. The graph is derived from the fourth annotation level under biological process.

Figure 5 shows the distribution of the edge-weight scores generated using these genes on the normal and adenocarcinoma samples from the Harvard dataset. If our method is to be predictive, we expect to see something of a bimodal distribution, although peak height is dependent on the relative populations of the two groups. This is because weights between members of the same group are expected to be high, while weights between members of different groups are expected to be low. Such a distribution is in fact what we observe in Figure 5.

We exploit this property when carrying out threshold selection. We choose an initial threshold slightly to the right of the median edge-weight value. We then enumerate all maximal cliques in the unweighted graph, and check to see whether every sample is in at least one clique. If not, we choose lower and lower threshold values until we have full coverage (that is, until every sample is in at least one clique). If, on the other hand, our initial threshold gives us full coverage, we incrementally select higher and higher thresholds until we generate an unweighted graph for which there is at least one sample that is

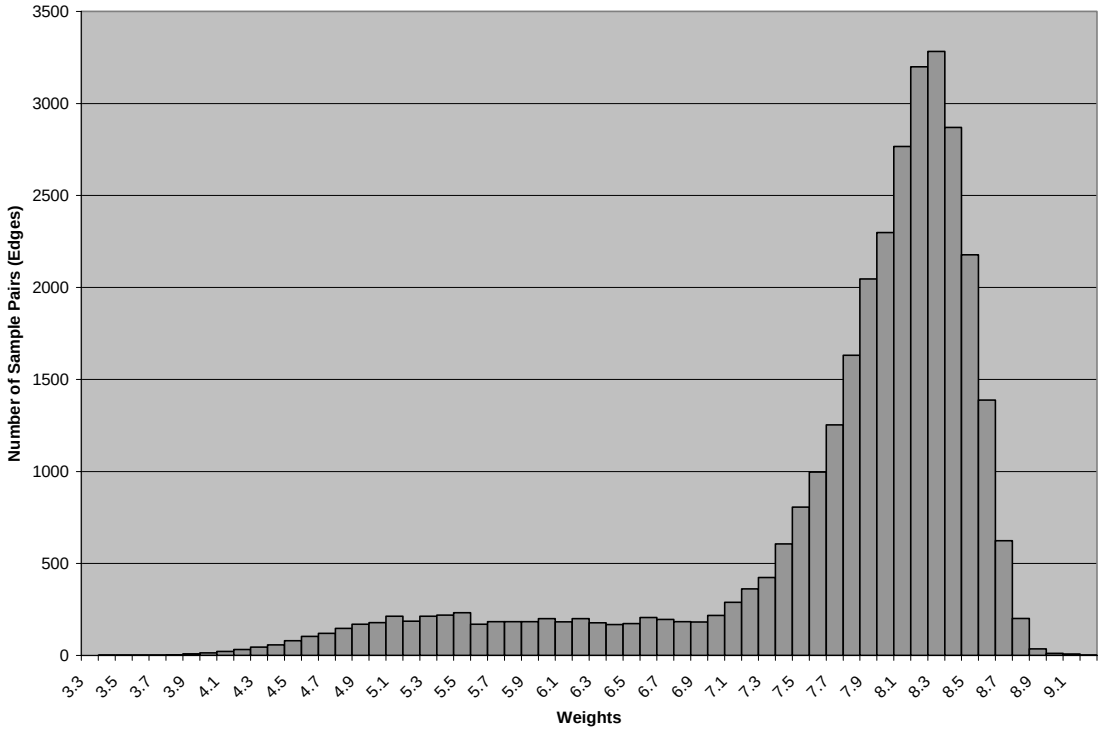

Figure 5. Weights between sample pairs using 78 genes (Harvard data)

missing from every maximal clique. At this point, we go back one step and use the highest threshold with full coverage. Naturally, this is only one possible method for selecting the threshold; other methods may work equally well. After a suitable threshold has been determined, we analyze the data by testing the supposition that all cliques of significant size are uniform in the sense that they contain samples from adenocarcinoma samples only or from normal samples only.

When this iterative process is carried out on the Harvard dataset without the use of any previous knowledge pertaining to its sample classifications, we are effectively able to separate the subjects into adenocarcinoma cliques and normal cliques. In fact, at our chosen threshold of 7.9, only one sample out of the 207 combined adenocarcinoma and normal samples would be misclassified according to the Harvard dataset using this approach. This sample is 2001032848AA.CEL. Because it was originally classified as adenocarcinoma but appeared in multiple normal cliques and no adenocarcinoma cliques, we suspect the original classification may be incorrect. The histogram of the enumerated cliques is shown in Figure 6. The largest mixed clique is of size six, and there are only five mixed cliques in total.

Of course, we are able to check the quality of our results because the tissue samples represented in the Harvard study have been previously classified. To use our methods in the absence of such information, one needs merely to examine the expression values of the highest-scoring genes directly to determine whether a clique represents a set of adenocarcinoma or normal samples.

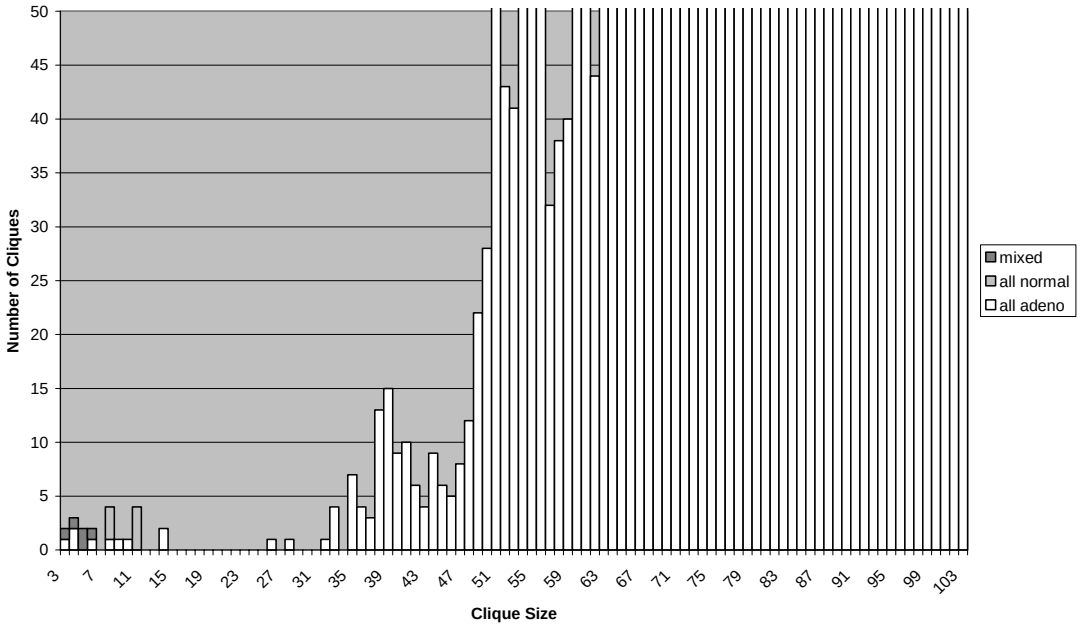

Figure 6. Clique frequency distribution from Harvard data set (adenocarcinoma and normal samples) using 78 genes and a threshold of 7.9

5.2 Experiment Two

In this case, we initially identify 195 genes that differentiate cancerous and normal samples. This is reduced to 180 (categorized by gene ontology in Figure 7) using our refinement technique, and 109 of these genes are available in the Michigan dataset.

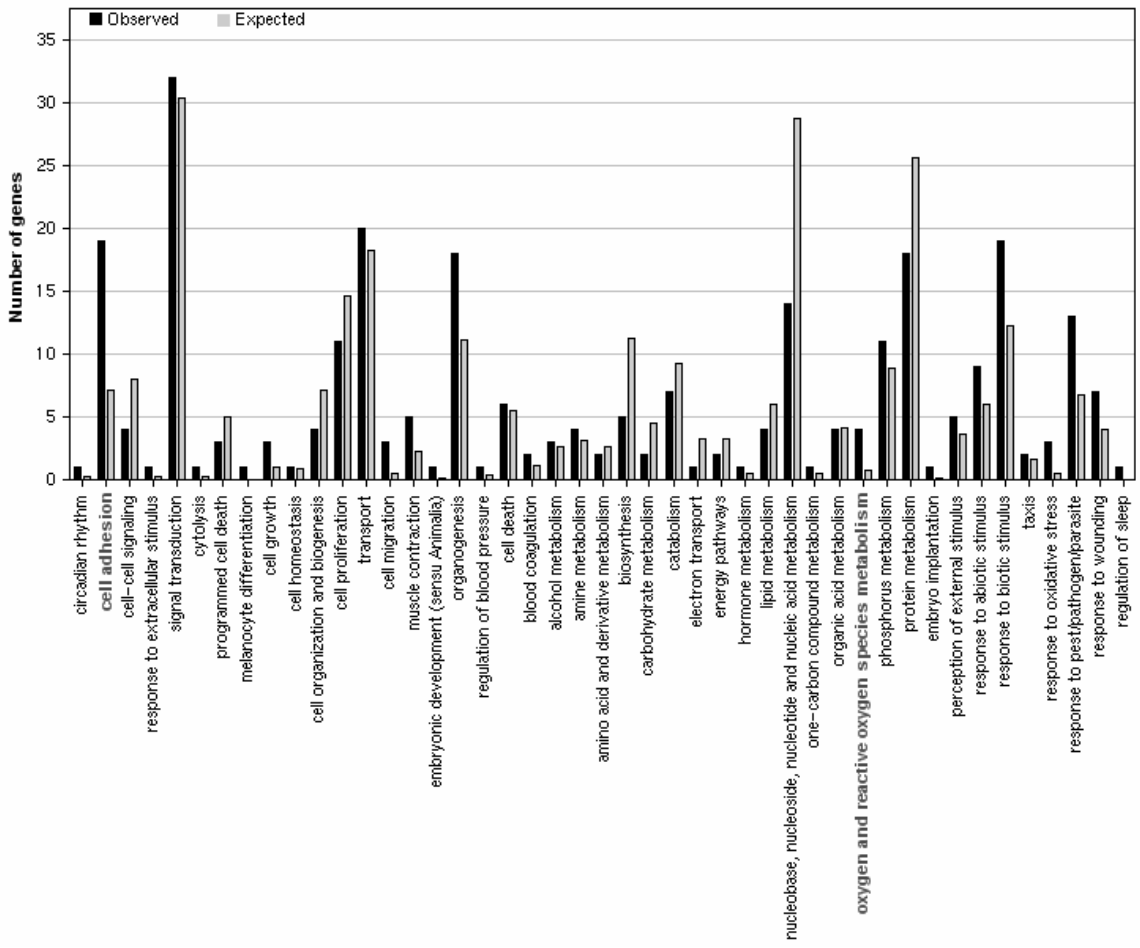

Figure 7. 180 genes (Harvard data) categorized by gene ontology. Black bars represent observed gene numbers. White bars represent expected gene numbers in the categories.

After following the process we have detailed, we select a threshold of 8.7, and enumerate maximal cliques on the resulting unweighted graph. Our methods are able to sort the samples into cancerous and normal cliques almost flawlessly. In fact, out of the 235 cliques of size 3 or greater in the resulting graph, only one clique has both cancerous and normal samples, and this is very small (size 3). The resultant frequency distribution of these cliques is depicted in Figure 8.

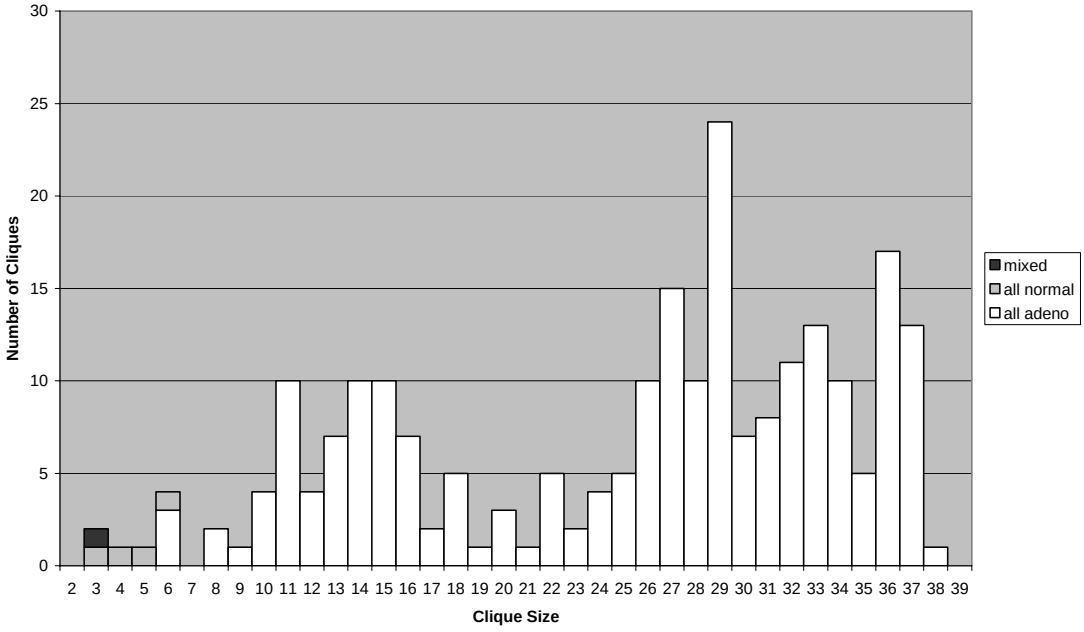

Figure 8. Clique distribution from Michigan data set using 109 genes and a threshold of 8.7

5.3 Experiment Three

Training on the Harvard dataset to discriminate between adenocarcinoma and squamous cell carcinoma initially gives us 37 genes. After refinement, 35 are left (Figure 9), 26 of which are found in the Stanford data set.

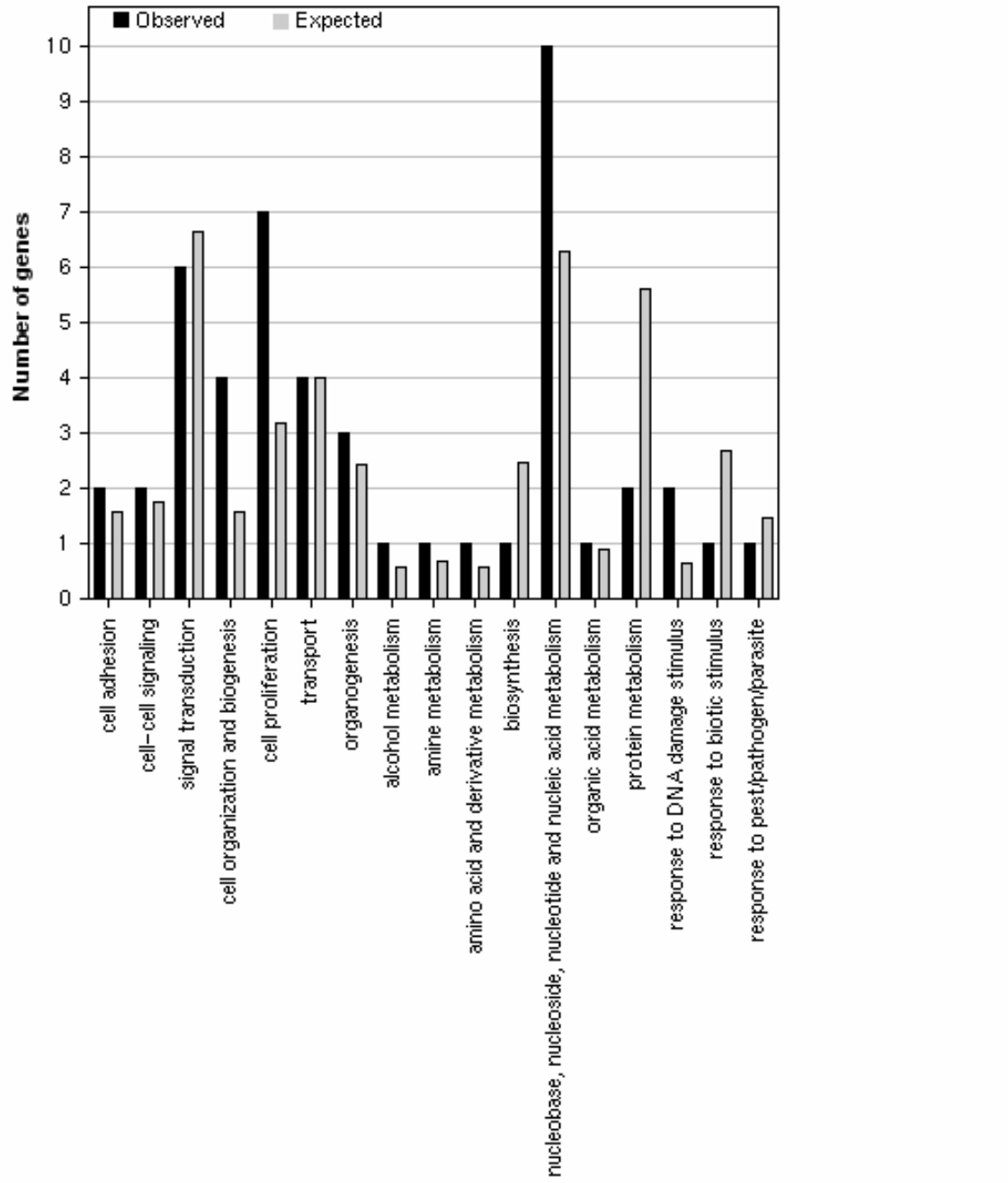

Figure 9. 35 genes (Harvard data) categorized under gene ontology. Black bars represent observed gene numbers. White bars represent expected gene numbers in the categories.

In this case, the results given by our method are not as compelling as in the previous two experiments. By using the largest clique containing each sample, we classify 41 out of 47 samples correctly according to the Stanford classifications. Nevertheless, there are still too many mixed cliques. This is not unexpected. Our methods isolate a set of 35 genes as a good discriminator. However, with only 26 of these available in the test data-set, their use provides at best a crude classification tool.

6. CONCLUSIONS

There is no apparent consensus as to the best approach for mining microarray data. Popular methods in current use include Bayesian analysis [9,18], hierarchical clustering, and scale-free networks [17], to name just a few. We believe that the novel methodology we have described here can be used to complement these techniques, and also be of independent interest. Deliverables accompanying this effort include the algorithmic framework of our overall strategy, the software tools we have developed and implemented, and of course the resultant gene sets themselves.

A key feature of our approach is the use of two distinct gene-scoring systems, each coupled with a different combinatorial algorithm. One is based on finding optimal cliques within general graphs, the other on isolating near-optimal dominating sets within bipartite graphs. Used in tandem, these algorithms appear to provide an effective means for identifying and ranking predictive genes whose expression levels serve as an accurate discriminator between adenocarcinoma and normal tissues. We emphasize that the use of clique and dominating set together seems to produce better results than would be possible with either approach alone.

The high fidelity with which the resulting cliques partition cancerous and normal samples, as illustrated in Figures 6 and 8, prompts us to posit that our methodology has the potential to become the basis for a highly reliable tool for cancer prediction. No *a priori* knowledge of the number of classes contained in the dataset is required. Moreover, it is known that tumor tissue samples are frequently a mixture of multiple types of cells, and that the exact ratio of this mixture is not necessarily consistent, even among samples from the same tumor. Therefore, it is expected that tissue samples might have significant similarity to more than one class, such as adenocarcinoma and normal. This is, in fact, what is observed. Using our method, the classification of the sample is not limited to one class. Nor is the classification based on the highest similarity score. Instead, it is based on a significant degree of similarity to the greatest number of samples that also are significantly similar to each other. In other words, classification is based on the largest (maximal) clique to which the sample belongs. This should result in a higher degree of confidence in our classification.

As a further proof of principle, several of the genes we have identified as discriminators in the Michigan data are known or suspected to play a role in oncogenesis. Among these are: CYP4B1, a cytochrome P450 enzyme that has been implicated in both bladder and lung cancer in humans [6,14]; FHL1, shown to have cytotoxic effects on melanoma cell lines and to possibly play a role in cellular differentiation[19]; the p85 alpha subunit of phosphoinositide-3-kinase, which plays a role in human breast cancer [7,16]; and tetranectin, which has already been shown to have prognosticative value for survival rates at certain stages of ovarian cancer [12]. A list of all the genes we have identified is in the Appendix in tables 1 and 2.

A number of opportunities for future research beckon. For example, the formula we are currently using to assign edge weights relies only on the gene scoring algorithm of our clique-based strategy. This can perhaps be refined by incorporating into it the gene scores computed during our dominating set analysis. Another idea we believe holds promise relies on the use of clique intersection graphs. These are computed as follows. Suppose we are given a filtered, unweighted sample similarity graph, G. The vertices of its associated clique intersection graph are the maximal cliques in G. Each pair of vertices in the clique intersection graph is connected by an edge if and only if the intersection of the two respective cliques they represent is nonempty. Thus, a clique intersection

graph may help to discern the overall structure of relationships contained within sample data. Moreover, cliques within a clique intersection graph may serve to tighten the focus on discriminating factors and act as an aid in quantifying the salient characteristics of archetypical diseased or healthy tissues.

REFERENCES

1. Abu-Khzam FN, Collins RL, Fellows MR, Langston MA, Suters WH, Symons CT. Kernelization algorithms for the vertex cover problem. *Proceedings, Workshop on Algorithm Engineering and Experiments (ALENEX)*, New Orleans, LA, January, 2004.
2. Abu-Khzam FN, Langston MA, Shanbhag P. Scalable Parallel Algorithms for Difficult Combinatorial Problems: A Case Study in Optimization. *Proceedings, International Conference on Parallel and Distributed Computing and Systems*, Los Angeles, CA, 563-568, November, 2003.
3. Baldwin NE, Collins RL, Langston MA, Leuze MR, Symons CT, Voy BR. High performance computational tools for motif discovery. *Proceedings, IEEE Workshop on High Performance Computational Biology*, Santa Fe, NM, April, 2004.
4. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB, Hanash S. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine* 9 (816), 816-824, 2002.
5. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA.* 98 (24), 13790-13795, 2001.
6. Czerwinski M, McLemore TL, Gelboin HV, Gonzalez FJ. Quantification of CYP2B7, CYP4B1, and CYPOR messenger RNAs in normal human lung and lung tumors. *Cancer Res*. 54(4):1085-91, 1994.
7. Das R, Mahabeleshwar GH, Kundu GC. Osteopontin stimulates cell motility and nuclear factor kappaB-mediated secretion of urokinase type plasminogen activator through phosphatidylinositol 3-kinase/Akt signaling pathways in breast cancer cells. *J Biol Chem*. 278(31):28593-606, 2003.
8. R. G. Downey and M. R. Fellows. Parameterized Complexity. Springer-Verlag. 1999.
9. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 7(3-4):601-20, 2000.
10. Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, Petersen I. Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A*. 98(24):13784-13789, 2001.
11. Garey MR, Johnson DS. Computers and Intractability. W. H. Freeman, New York, 1979.
12. Hogdall CK, Norgaard-Pedersen B, Mogensen O.The prognostic value of pre-operative serum tetranectin, CA-125 and a combined index in women with primary ovarian cancer. *Anticancer Res*. 22(3):1765-8, 2002.
13. Hu JH, Yin GS, Morris JS, Zhang L, Wright FA. Entropy and survival-based weights to combine Affymetrix array types in the analysis of differential expression and survival. *Critical Assessment of Microarray Data Analysis "CAMDA'03": Oral and Poster Presenters Abstracts*, 78-82, 2003.
14. Imaoka S, Yoneda Y, Sugimoto T, Hiroi T, Yamamoto K, Nakatani T, Funae Y. CYP4B1 is a possible risk factor for bladder cancer in humans. *Biochem Biophys Res Commun*. 277(3):776-80, 2000.
15. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2): 249-264. 2003.
16. Mahabeleshwar GH, Kundu GC. Syk, a protein-tyrosine kinase, suppresses the cell motility and nuclear factor kappa B-mediated secretion of urokinase type plasminogen activator by inhibiting the phosphatidylinositol 3'-kinase activity in breast cancer cells. *J Biol Chem*. 278(8):6209-21, 2003.
17. del Rio G, Bartley TF, del-Rio H, Rao R, Jin KL, Greenberg DA, Eshoo M, Bredesen DE. Mining DNA microarray data using a novel approach based on graph theory. *FEBS Letters* 509(2):230-4, 2001.
18. Sok JC, Kuriakose MA, Mahajan VB, Pearlman AN, DeLacure MD, Chen FA. Tissue-specific gene expression of head and neck squamous cell carcinoma in vivo by complementary DNA microarray analysis. *Arch Otolaryngol Head Neck Surgery* 129(7):760-70, 2003.

19. de Vries JE, Meyering M, van Dongen A, Rumke P. The influence of different isolation procedures and the use of target cells from melanoma cell lines and short-term cultures on the non-specific cytotoxic effects of lymphocytes from healthy donors. *Int J Cancer*. 15(3):391-400, 1975.
20. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. To appear in *BMC Bioinformatics*, 2004; http://genereg.ornl.gov/gotm.

APPENDIX

Table 1. Functional annotation of genes from the Harvard (H) and Michigan (M) datasets that our methods identify as discriminators among cancerous and normal samples. *Y* means that the gene was identified. *N* means that the gene was not identified, but was present in the dataset. *N/A* means the gene was not available in the dataset.

Locus-Link ID	SYMBOL	GENE_NAME	Identified	
			H	M
21	ABCA3	ATP-binding cassette, sub-family A (ABC1), member 3	Y	Y
104	ADARB1	adenosine deaminase, RNA-specific, B1 (RED1 homolog rat)	Y	Y
124	ADH1A	alcohol dehydrogenase 1A (class I), alpha polypeptide	Y	Y
125	ADH1B	alcohol dehydrogenase IB (class I), beta polypeptide	Y	Y
284	ANGPT1	angiopoietin 1	Y	Y
361	AQP4	aquaporin 4	Y	Y
687	BTEB1	basic transcription element binding protein 1	Y	Y
730	C7	complement component 7	Y	Y
1003	CDH5	cadherin 5, type 2, VE-cadherin (vascular epithelium)	Y	Y
1043	CDW52	CDW52 antigen (CAMPATH-1 antigen)	Y	Y
1359	CPA3	carboxypeptidase A3 (mast cell)	Y	Y
1465	CSRP1	cysteine and glycine-rich protein 1	Y	Y
1675	DF	D component of complement (adipsin)	Y	Y
1910	EDNRB	endothelin receptor type B	Y	Y
2013	EMP2	epithelial membrane protein 2	Y	Y
2014	EMP3	epithelial membrane protein 3	Y	Y
2167	FABP4	fatty acid binding protein 4, adipocyte	Y	Y
2273	FHL1	four and a half LIM domains 1	Y	Y
2294	FOXF1	forkhead box F1	Y	Y
2313	FLI1	Friend leukemia virus integration 1	Y	Y
2719	GPC3	glypican 3	Y	Y
2791	GNG11	guanine nucleotide binding protein (G protein), gamma 11	Y	Y
2869	GPRK5	G protein-coupled receptor kinase 5	Y	Y
2878	GPX3	glutathione peroxidase 3 (plasma)	Y	Y
3043	HBB	hemoglobin, beta	Y	Y
3730	KAL1	Kallmann syndrome 1 sequence	Y	Y
4005	LMO2	LIM domain only 2 (rhombotin-like 1)	Y	Y
4360	MRC1	mannose receptor, C type 1	Y	Y
4638	MYLK	myosin, light polypeptide kinase	Y	Y
4688	NCF2	neutrophil cytosolic fact. 2 (65kDa, autosomal 2)	Y	Y
5376	PMP22	peripheral myelin protein 22	Y	Y
5627	PROS1	protein S (alpha)	Y	Y
6711	SPTBN1	spectrin, beta, non-erythrocytic 1	Y	Y
7010	TEK	TEK tyrosine kinase, endothelial	Y	Y
7048	TGFBR2	transforming growth factor, beta receptor II (70/80kDa)	Y	Y
7049	TGFBR3	transforming growth factor, beta receptor III	Y	Y
7123	TNA	tetranectin (plasminogen binding protein)	Y	Y
7450	VWF	von Willebrand factor	Y	Y

Locus-Link ID	SYMBOL	GENE_NAME	Identified	
			H	M
8404	SPARCL1	SPARC-like 1 (mast9, hevin)	Y	Y
8516	ITGA8	integrin, alpha 8	Y	Y
8613	PPAP2B	phosphatidic acid phosphatase type 2B	Y	Y
8639	AOC3	amine oxidase, copper containing 3	Y	Y
9459	ARHGEF6	Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6	Y	Y
9934	GPR105	G protein-coupled receptor 105	Y	Y
10398	MYL9	myosin, light polypeptide 9, regulatory	Y	Y
10974	APM2	adipose specific 2	Y	Y
154	ADRB2	adrenergic, beta-2-, receptor, surface	Y	N
195	AHNAK	AHNAK nucleoprotein (desmoyokin)	Y	N
358	AQP1	aquaporin 1 (channel-forming integral protein, 28kDa)	Y	N
762	CA4	carbonic anhydrase IV	Y	N
858	CAV2	caveolin 2	Y	N
947	CD34	CD34 antigen	Y	N
1066	CES1	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	Y	N
2022	ENG	endoglin (Osler-Rendu-Weber syndrome 1)	Y	N
2078	ERG	v-ets erythroblastosis virus E26 oncogene like (avian)	Y	N
2192	FBLN1	fibulin 1	Y	N
2202	EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	Y	N
2219	FCN1	ficolin (collagen/fibrinogen domain containing) 1	Y	N
2597	GAPD	glyceraldehyde-3-phosphate dehydrogenase	Y	N
2615	GARP	glycoprotein A repetitions predominant	Y	N
2701	GJA4	gap junction protein, alpha 4, 37kDa (connexin 37)	Y	N
2771	GNAI2	guanine nucleotide binding prot, alpha inhibit activity polypep 2	Y	N
2824	GPM6B	glycoprotein M6B	Y	N
3133	HLA-E	major histocompatibility complex, class I, E	Y	N
3340	NDST1	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1	Y	N
3373	HYAL1	hyaluronoglucosaminidase 1	Y	N
3575	IL7R	interleukin 7 receptor	Y	N
3936	LCP1	lymphocyte cytosolic protein 1 (L-plastin)	Y	N
4035	LRP1	low density lipoprotein-related protein 1	Y	N
4091	MADH6	MAD, mothers against decapentaplegic homolog 6 (Drosophila)	Y	N
4239	MFAP4	microfibrillar-associated protein 4	Y	N
4286	MITF	microphthalmia-associated transcription factor	Y	N
4332	MNDA	myeloid cell nuclear differentiation antigen	Y	N
4502	MT2A	metallothionein 2A	Y	N
4628	MYH10	myosin, heavy polypeptide 10, non-muscle	Y	N
4629	MYH11	myosin, heavy polypeptide 11, smooth muscle	Y	N
4855	NOTCH4	Notch homolog 4 (Drosophila)	Y	N
4881	NPR1	natriuretic peptide receptor A/guanylate cyclase A	Y	N
4973	OLR1	oxidised low density lipoprotein (lectin-like) receptor 1	Y	N
5225	PGC	progastricsin (pepsinogen C)	Y	N
5730	PTGDS	prostaglandin D2 synthase 21kDa (brain)	Y	N
5787	PTPRB	protein tyrosine phosphatase, receptor type, B	Y	N
5797	PTPRM	protein tyrosine phosphatase, receptor type, M	Y	N

Locus-Link ID	SYMBOL	GENE_NAME	Identified	
			H	M
5831	PYCR1	pyrroline-5-carboxylate reductase 1	Y	N
6237	RRAS	related RAS viral (r-ras) oncogene homolog	Y	N
6403	SELP	selectin P (granule membrane protein 140kDa, antigen CD62)	Y	N
6556	SLC11A1	solute carrier fam. 11, memb. 1	Y	N
6709	SPTAN1	spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)	Y	N
6909	TBX2	T-box 2	Y	N
7122	CLDN5	claudin 5	Y	N
7133	TNFRSF1B	tumor necrosis factor receptor superfamily, member 1B	Y	N
7148	TNXB	tenascin XB	Y	N
7433	VIPR1	vasoactive intestinal peptide receptor 1	Y	N
7704	ZNF145	zinc finger protein 145	Y	N
8692	HYAL2	hyaluronoglucosaminidase 2	Y	N
9034	CCRL2	chemokine (C-C motif) receptor-like 2	Y	N
9124	PDLIM1	PDZ and LIM domain 1 (elfin)	Y	N
9173	IL1RL1	interleukin 1 receptor-like 1	Y	N
9341	VAMP3	vesicle-associated membrane protein 3 (cellubrevin)	Y	N
9413	X123	Friedreich ataxia region gene X123	Y	N
9638	FEZ1	fasciculation and elongation protein zeta 1 (zygin I)	Y	N
9647	PPM1F	protein phosphatase 1F (PP2C domain containing)	Y	N
9748	SLK	Ste20-related serine/threonine kinase	Y	N
10129	13CDNA73	hypothetical protein CG003	Y	N
10609	SC65	nucleolar autoantigen sim to rat synaptonemal complex protein	Y	N
10979	PLEKHC1	pleckstrin homology domain containing, fam. C memb. 1	Y	N
23294	ANKS1	ankyrin repeat and SAM domain containing 1	Y	N
25802	LMOD1	leiomodin 1 (smooth muscle)	Y	N
54861	SNRK	SNF-1 related kinase	Y	N
115	ADCY9	adenylate cyclase 9	Y	N/A
177	AGER	advanced glycosylation end product-specific receptor	Y	N/A
241	ALOX5AP	arachidonate 5-lipoxygenase-activating protein	Y	N/A
306	ANXA3	annexin A3	Y	N/A
409	ARRB2	arrestin, beta 2	Y	N/A
847	CAT	catalase	Y	N/A
2277	FIGF	c-fos induced growth factor	Y	N/A
2532	FY	Duffy blood group	Y	N/A
3384	ICAM2	intercellular adhesion molecule 2	Y	N/A
4008	LMO7	LIM domain only 7	Y	N/A
4282	MIF	macrophage migration inhibitory factor	Y	N/A
5175	PECAM1	platelet/endothelial cell adhesion molecule (CD31 antigen)	Y	N/A
5348	FXYD1	FXYD domain containing ion transport regulator 1	Y	N/A
5420	PODXL	podocalyxin-like	Y	N/A
6279	S100A8	S100 calcium binding protein A8 (calgranulin A)	Y	N/A
6440	SFTPC	surfactant, pulmonary-associated protein C	Y	N/A
6867	TACC1	transforming, acidic coiled-coil containing protein 1	Y	N/A
6943	TCF21	transcription factor 21	Y	N/A
7134	TNNC1	troponin C, slow	Y	N/A

Locus-Link ID	SYMBOL	GENE_NAME	Identified	
			H	M
7356	SCGB1A1	secretoglobin, family 1A, member 1 (uteroglobin)	Y	N/A
7466	WFS1	Wolfram syndrome 1 (wolframin)	Y	N/A
8425	LTBP4	latent transforming growth factor beta binding protein 4	Y	N/A
8547	FCN3	ficolin (collagen/fibrinogen domain containing) 3	Y	N/A
8612	PPAP2C	phosphatidic acid phosphatase type 2C	Y	N/A
8630	RODH	3-hydroxysteroid epimerase	Y	N/A
8685	MARCO	macrophage receptor with collagenous structure	Y	N/A
8727	CTNNAL1	catenin (cadherin-associated protein), alpha-like 1	Y	N/A
9056	SLC7A7	solute carrier fam. 7, memb. 7	Y	N/A
9079	LDB2	LIM domain binding 2	Y	N/A
9353	SLIT2	slit homolog 2 (Drosophila)	Y	N/A
9411	PARG1	PTPL1-associated RhoGAP 1	Y	N/A
9452	ITM2A	integral membrane protein 2A	Y	N/A
9467	SH3BP5	SH3-domain binding protein 5 (BTK-associated)	Y	N/A
9535	GMFG	glia maturation factor, gamma	Y	N/A
9732	DOCK4	DOCK4	Y	N/A
10266	RAMP2	receptor (calcitonin) activity modifying protein 2	Y	N/A
10268	RAMP3	receptor (calcitonin) activity modifying protein 3	Y	N/A
10351	ABCA8	ATP-binding cassette, sub-family A (ABC1), member 8	Y	N/A
10395	DLC1	deleted in liver cancer 1	Y	N/A
10516	FBLN5	fibulin 5	Y	N/A
10908	NTE	neuropathy target esterase	Y	N/A
11025	LILRB3	leukocyte immunoglobulin-like receptor, subfamily B, memb. 3	Y	N/A
11142	PKIG	protein kinase (cAMP-dependent, catalytic) inhibitor gamma	Y	N/A
11170	TU3A	TU3A protein	Y	N/A
11197	WIF1	WNT inhibitory factor 1	Y	N/A
11217	AKAP2	A kinase (PRKA) anchor protein 2	Y	N/A
11309	SLC21A9	solute carrier family 21 (organic anion transporter), member 9	Y	N/A
11326	Z39IG	Ig superfamily protein	Y	N/A
22885	KIAA0843	KIAA0843 protein	Y	N/A
22939			Y	N/A
22998	KIAA1102	KIAA1102 protein	Y	N/A
23037	PDZK3	PDZ domain containing 3	Y	N/A
23266	LPHN2	latrophilin 2	Y	N/A
23328	SASH1	SAM and SH3 domain containing 1	Y	N/A
23371	TENC1	tensin like C1 domain-containing phosphatase	Y	N/A
23499	MACF1	microtubule-actin crosslinking factor 1	Y	N/A
23673	STX12	syntaxin 12	Y	N/A
23710	GABARAPL1	GABA(A) receptor-associated protein like 1	Y	N/A
25777	UNC84B	unc-84 homolog B (C. elegans)	Y	N/A
27074	LAMP3	lysosomal-associated membrane protein 3	Y	N/A
27253	PCDH17	protocadherin 17	Y	N/A
57188	KIAA1233	KIAA1233 protein	Y	N/A
57493	KIAA1237	KIAA1237 protein	Y	N/A
64116	BIGM103	BCG-induced gene in monocytes, clone 103	Y	N/A

Locus-Link ID	SYMBOL	GENE_NAME	Identified	
			H	M
79602	FLJ21432	hypothetical protein FLJ21432	Y	N/A
83604	BCMP1	brain cell membrane protein 1	Y	N/A
84724			Y	N/A
91851	NRLN1	likely ortholog of mouse neuralin 1	Y	N/A
115207	LOC115207	hypothetical protein BC013764	Y	N/A
126393	FLJ32389	hypothetical protein FLJ32389	Y	N/A
203317			Y	N/A
240	ALOX5	arachidonate 5-lipoxygenase	N	Y
857	CAV1	caveolin 1, caveolae protein, 22kDa	N	Y
894	CCND2	cyclin D2	N	Y
948	CD36	CD36 antigen (collagen type I recept, thrombospondin receptor)	N	Y
976	CD97	CD97 antigen	N	Y
1318	SLC31A2	solute carrier family 31 (copper transporters), member 2	N	Y
1346	COX7A1	cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)	N	Y
1410	CRYAB	crystallin, alpha B	N	Y
1580	CYP4B1	cytochrome P450, family 4, subfamily B, polypeptide 1	N	Y
1601	DAB2	disabled homolog 2, mitogen-responsive phosphoprotein	N	Y
1808	DPYSL2	dihydropyrimidinase-like 2	N	Y
1901	EDG1	endothelial dif., sphingolipid G-protein-coupled receptor, 1	N	Y
2268	FGR	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog	N	Y
2327	FMO2	flavin containing monooxygenase 2	N	Y
2823	GPM6A	glycoprotein M6A	N	Y
2995	GYPC	glycophorin C (Gerbich blood group)	N	Y
4192	MDK	midkine (neurite growth-promoting factor 2)	N	Y
5295	PIK3R1	phosphoinositide-3-kinase, regulatory subunit, polypeptide 1	N	Y
6275	S100A4	S100 calcium binding protein A4 (murine placental homolog)	N	Y
6404	SELPLG	selectin P ligand	N	Y
6414	SEPP1	selenoprotein P, plasma, 1	N	Y
6595	SMARCA2	SWI/SNF rel., mat. assoc., actin dep. reg. of chromatin, sfm a2	N	Y
7262	TSSC3	tumor suppressing subtransferable candidate 3	N	Y
7264	TSTA3	tissue specific transplantation antigen P35B	N	Y
8406	SRPX	sushi-repeat-containing protein, X chromosome	N	Y
9770	RASSF2	Ras association (RalGDS/AF-6) domain family 2	N	Y
9806	SPOCK2	sparc/osteonectin, cwcv and kazal-like doms proteoglycan 2	N	Y
9936	DCL-1	type I transmembrane C-type lectin receptor DCL-1	N	Y
10203	CALCRL	calcitonin receptor-like	N	Y
10216	PRG4	proteoglycan 4	N	Y
10806	SDCCAG8	serologically defined colon cancer antigen 8	N	Y
26578	OSTF1	osteoclast stimulating factor 1	N	Y
2	A2M	alpha-2-macroglobulin	N/A	Y
316	AOX1	aldehyde oxidase 1	N/A	Y
2171	FABP5	fatty acid binding protein 5 (psoriasis-associated)	N/A	Y
5468	PPARG	peroxisome proliferative activated receptor, gamma	N/A	Y
6435	SFTPA1	surfactant, pulmonary-associated protein A1	N/A	Y
84099	ID2B	striated muscle contraction regulatory protein	N/A	Y

Table 2. Functional annotation of 35 genes that our methods identify as discriminators among adenocarcinoma and squamous cell carcinoma

Locus-Link ID	SYMBOL	GENE_NAME
10057	ABCC5	ATP-binding cassette, sub-family C (CFTR/MRP), member 5
1173	AP2M1	adaptor-related protein complex 2, mu 1 subunit
131	ADH7	alcohol dehydrogenase 7 (class IV), mu or sigma polypeptide
1365	CLDN3	claudin 3
1475	CSTA	cystatin A (stefin A)
1606	DGKA	diacylglycerol kinase, alpha 80kDa
1830	DSG3	desmoglein 3 (pemphigus vulgaris antigen)
1854	DUT	dUTP pyrophosphatase
22824	APG-1	heat shock protein (hsp110 family)
23250	ATP11A	ATPase, Class VI, type 11A
23299	BICD2	coiled-coil protein BICD2
23650	TRIM29	tripartite motif-containing 29
244	ANXA8	annexin A8
2817	GPC1	glypican 1
2956	MSH6	mutS homolog 6 (E. coli)
3655	ITGA6	integrin, alpha 6
3852	KRT5	keratin 5
3853	KRT6A	keratin 6A
3872	KRT17	keratin 17
4171	MCM2	MCM2 minichromosome maintenance deficient 2, mitotin (S. cerevisiae)
4680	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6
483	ATP1B3	ATPase, Na+/K+ transporting, beta 3 polypeptide
5111	PCNA	proliferating cell nuclear antigen
5268	SERPINB5	serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 5
54107	POLE3	polymerase (DNA directed), epsilon 3 (p17 subunit)
5905	RANGAP1	Ran GTPase activating protein 1
5984	RFC4	replication factor C (activator 1) 4, 37kDa
6273	S100A2	S100 calcium binding protein A2
6657	SOX2	SRY (sex determining region Y)-box 2
7080	TITF1	thyroid transcription factor 1
8323	FZD6	frizzled homolog 6 (Drosophila)
86	BAF53A	BAF53
8714	ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
8893	EIF2B5	eukaryotic translation initiation factor 2B, subunit 5 epsilon, 82kDa
9982	HBP17	heparin-binding growth factor binding protein