Pharmaceutical Target Identification by Gene Expression Analysis

Michael G. Walker*

Incyte Genomics, 3174 Proter Dr., Palo Alto, California, USA

Abstract: The majority of newly-identified genes in the human genome show no significant sequence similarity to genes whose function is known, so they are not easily recognized as potential drug targets. Expression analysis is an alternative method to suggest possible functions of genes. We review statistical methods for gene expression analysis to identify potential pharmaceutical targets. Specifically, we illustrate the analysis of differential gene expression (using discriminant analysis, t-tests, and analysis of variance) and co-expression (using correlation, clustering, and chi-square). We present an example of the use of expression analysis to identify co-expressed cardiomyopathy-associated genes.

1. CANDIDATE DRUG TARGETS AND GENE EXPRESSION DATA

The DNA sequence of the human genome is now known. However, we still need to determine which genes are involved in disease. The majority of newly-identified genes in the human genome and in other genomes show little or no significant sequence similarity to genes with currently known function, so we need alternatives to sequence analysis. Gene expression data are available via expression microarrays; expression data may be readily collected for 10,000 genes with a single array [1, 2]. Expression data provide an alternative to sequence data to identify genes that may be candidate drug targets.

2. STATISTICAL METHODS TO IDENTIFY DISEASE-ASSOCIATED GENES

Expression data from experiments from even a single microarray require computational tools for analysis. A recent review on clustering methods for expression data may be found in [3]. General introductions to clustering, discriminant analysis, and other statistical methods may be found in texts or reviews on multivariate statistics [4]. Collections of articles on algorithms and statistics for gene expression analysis may be found at the web sites http://www.cgl.ucsf.edu/psb/ and http://industry.ebi.ac.uk/~alan/MicroArray/. Here we will illustrate the analysis of differential gene expression (using discriminant analysis, t-tests, and analysis of variance) and co-expression (using correlation, clustering, and chi-square). In a later section we present an example of co-expressed cardiomyopathy-associated genes.

2.1 Differential Expression (Discriminant Analysis and Analysis of Variance)

Researchers have long sought genes that are differentially expressed in disease versus non-disease states, and several groups have demonstrated the use of microarrays for this purpose [5-9].

An early approach to identifying differentially-expressed genes was simply to search for genes that were detected in one tissue or disease-state and not detected in the second tissue or disease-state. This method is sometimes useful, but suffers the obvious problem that expression may be altered with significant pathogenic effect without a gene necessarily being turned on or off. Soon after, the algorithms were modified to identify genes that show multi-fold changes in expression between the two disease states. This method is often useful, but may fail to identify interesting genes when the expression levels of the genes show large variance. Large variances are common in microarray expression measurements, because of variability in the sample preparation, in the arrays, and among patients. When expression levels show high variance, we may observe a seemingly large difference between two samples simply because of random error. In these situations, a more appropriate analysis method is discriminant analysis. Linear discriminant analysis considers both the difference in average expression between two groups and the variability of expression within each group. Specifically, linear discriminant analysis seeks to identify genes that have the best ratio of the difference in expression between two states to the variance of expression within each state. Figure 1 shows a hypothetical example of the expression level of a gene in two different groups, cancer versus non-cancer patients. Given the measured expression of the gene in a patient, linear discriminant analysis would give the probability that the individual belongs to one of the two groups.

*Address correspondence to this author at the Incyte Genomics, 3174 Proter Dr., Palo Alto, California, USA; e-mail: mwalker@incyte.com

1389-5575/01 $20.00+.00

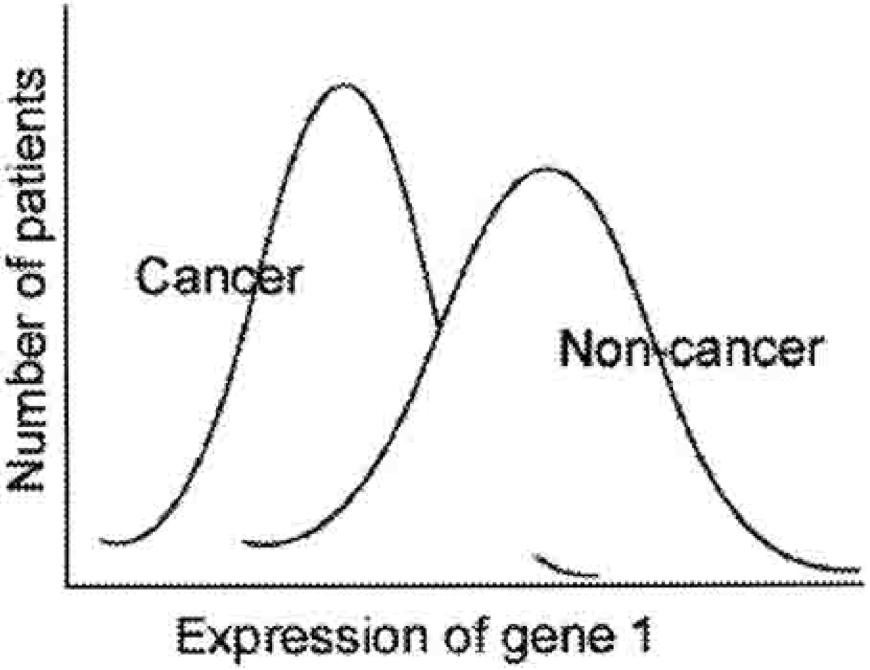

Fig. (1). Discriminant analysis: cancer vs. non-cancer.

It is often desirable to use more than one gene to classify a sample; Figure 2 is a hypothetical example showing the use of two genes (genes 1 and 2) to discriminate between two classes (classes A and B), where the classes might be cancer versus non-cancerous samples.

In a microarray experiment we may generate data on 10,000 or more genes, so it is not feasible to manually examine results for each gene one at a time to determine which single gene or which combination of genes best discriminates between the classes of interest. An alternative is stepwise discriminant analysis software, which can search through each of the thousands of genes to identify the single gene or combination of genes that best distinguishes between the disease states. There are numerous algorithms and software packages available for discriminant analysis; there are also algorithms that do not make the assumptions of linear partitions used in the examples given here [4]. Discriminant analysis is not limited to the case of two classes; it can readily analyze data from multiple classes, for example, non-cancerous, dysplastic, and cancer.

After we identify a gene or a small set of genes that appear to be differentially expressed in disease versus non-disease states, we should perform experiments to attempt to confirm the supposed differential expression. For these

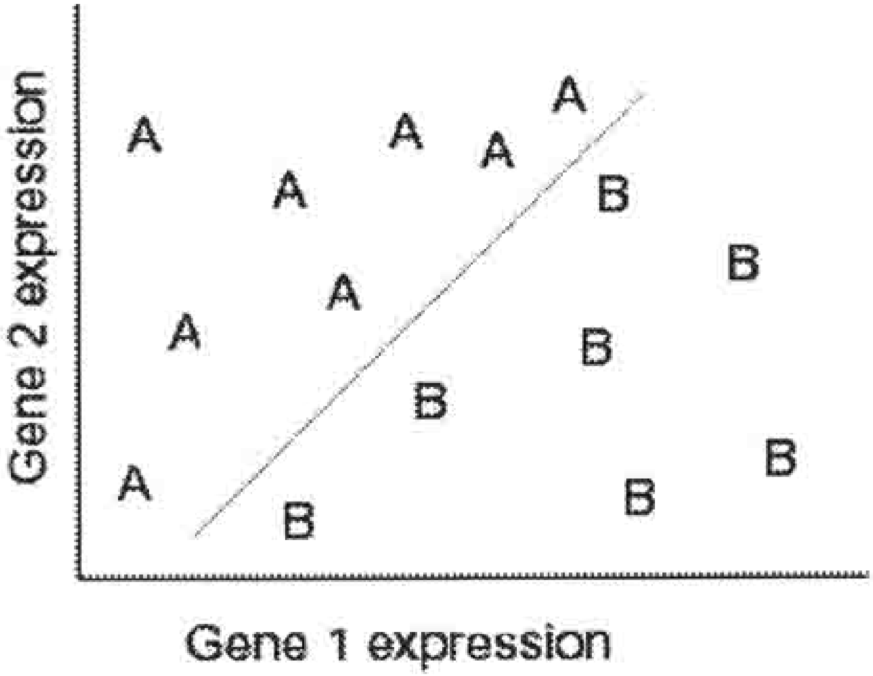

Fig. (2). Discriminant analysis using two genes.

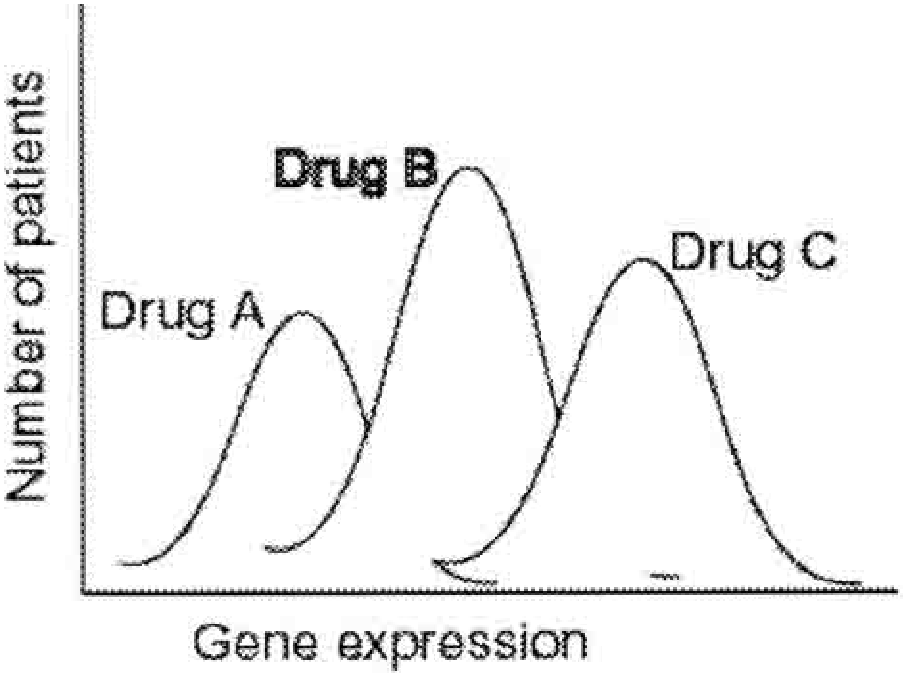

Fig. (3). Analysis of variance for three drugs.

experiments, we should measure the expression level of the gene(s) several times in each of the disease states. However, we should consider the variability in the measurements. To decide if the difference between the two groups is statistically significant, we should consider both the difference between the two groups in their mean expression, and the error in estimating the means. T-tests (for exactly two groups) and analysis of variance (for two or more groups) are suitable statistical tests in this situation. Figure 3 shows hypothetical data for an analysis of variance (ANOVA) of gene expression in response to three drugs. The t-test or ANOVA will indicate the probability the observed differences between the classes could have occurred by chance.

2.2 Co-expression (Correlation, Association, and Clustering)

The most widely-used expression analysis techniques, after differential expression, are based on examination of the co-expression of two or more genes. The commonly-used methods include correlation, clustering, and other measures of association [10-14], which we examine next.

2.2.1 Correlation

Correlation is a measure that describes a particular type of association between two genes. It is also the basis of several clustering methods. The most commonly used correlation statistic is Pearson’s linear correlation coefficient, r. It has a maximum value of one when two variables (such as the expression levels of two genes in various tissues) are exactly linearly proportional. Figure 4 shows possible data giving correlation coefficients near 1, 0, and –1. For perfect linear correlation, an increase in one variable occurs with an exactly proportional increase in the second variable. The Pearson correlation has a minimum value of negative one when the two variables are exactly inversely linear proportional (one increases as the other decreases). If two variables show no linear relation, then the Pearson correlation is zero.

Suppose we have data such as that shown in Table 1, showing the expression levels of three genes, A, B and C, in five different tissues. Genes A and B show very similar expression across the tissues, and have a correlation coefficient of 0.99, while gene C is quite different from the others, and has a correlation with both A and B near 0.0.

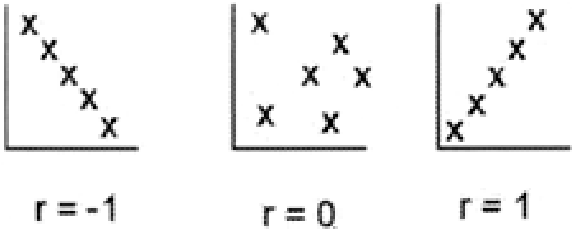

Fig. (4). Linear correlation near 1, 0, and –1.

Table 1. Hypothetical Expression Levels of Three Genes

	Brain	Heart	Muscle	Liver	Prostate
Gene A	0	4	18	7	25
Gene B	0	6	16	7	23
Gene C	6	4	6	9	6

A difficulty with Pearson linear correlation is that genes may be associated but not have a linear relationship. There are several alternatives to Pearson correlation to measure such non-linear associations. Spearman rank correlation, which uses the rank of each data point rather than its actual value, is less sensitive to outliers and extreme values than is Pearson linear correlation, and is useful for detecting monotonic (constantly increasing or decreasing) relationships. Another alternative is the chi-square test, in which we treat each gene as either expressed or not-expressed in a given sample, and measure the co-occurrence of the two genes. The chi-square test can detect relationships that are non-linear or non-monotonic, and that would therefore fail to be detected by Pearson linear or Spearman rank correlation. We present an example of the use of the chi-square test in the next section.

2.2.2 Categorical Measures: Chi-Square Test of Association

In some cases, linear or rank correlation will fail to detect associations among genes. Many genes that are known to be associated do not have the linear or monotonic relationships that these methods assume, and in some cases quantitative measurement may not be sufficiently accurate or reproducible. In such cases, we may choose to encode the level of gene expression as simply “on” or “off”, rather than use the quantitative information. We then look for a categorical association, in which both genes are turned on or both are turned off at the same time. This approach also reflects the usual situation in which most genes are not expressed in most tissues, and co-expression suggests related function.

If we examine the expression of two genes in cDNA libraries from, say, 30 tissue samples, we can summarize their co-expression as shown in Table 2. We use a chi-square test or a Fisher exact test (in the case of small expected values in any cell in the table) to determine if the observed co-occurrence of the genes differs significantly from that expected by chance; in this case the probability is 0.0003.

Table 2. Summary of Co-Expression for Genes A and B in 30 cDNA Libraries

Number of libraries	Gene A present	Gene A absent	Total
Gene B present	8	2	10
Gene B absent	2	18	20
Total	10	20	30

We have used this method of analysis to identify genes involved with a variety of diseases, including prostate cancer [13], Parkinson’s disease and schizophrenia [15] and others. In the final section of this paper, we give an example of genes associated with cardiomyopathy identified using this method. The chi-square test will usually have less power than Pearson or Spearman correlation to detect linear or monotonic relationships.

2.2.3 Cluster Analysis

Cluster analysis comprises a set of methods that help us to compare and visualize relationships among objects (such as sets of drugs or sets of genes) so that we can perceive which are similar and which are dissimilar from each other. Suppose that we examine the expression patterns of five genes in a variety of tissues (or when a tissue is treated with a variety of drugs at different doses). We can calculate the pairwise correlations of the five genes amongst themselves, and might get data such as that in Table 3.

Table 3. Pairwise Correlations of Expression Among Five Genes

	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Gene 1	1.0	.9	.5	.4	0
Gene 2		1.0	.5	.4	0
Gene 3			1.0	.8	0
Gene 4				1.0	0
Gene 5					1.0

In these hypothetical data, notice that genes 1 and 2 are highly correlated (r = 0.9), genes 3 and 4 are highly correlated (r = .8), and gene 5 is not correlated at all with the other genes (r = 0.0). To create a hierarchical (tree-structured) cluster such as that shown in Figure 5, we successively join the most-correlated pairs of genes, so that the cluster tree will indicate that genes 1 and 2 are highly correlated, genes 3 and 4 are highly correlated, and gene 5 is distant from the other genes.

There are many algorithms to produce clusters [4]. In addition, dimension-reduction methods such as multi-dimensional scaling and principal components analysis can provide visual displays that indicate similar and dissimilar genes identified under diverse experimental conditions in a way that clustering algorithms cannot. These methods are described in most texts on multivariate statistics.

2.2.4 Other Applications of Gene Expression Data in Drug Research

We can use correlation to indicate the likely mode of action of a drug. Suppose that we have expression data for

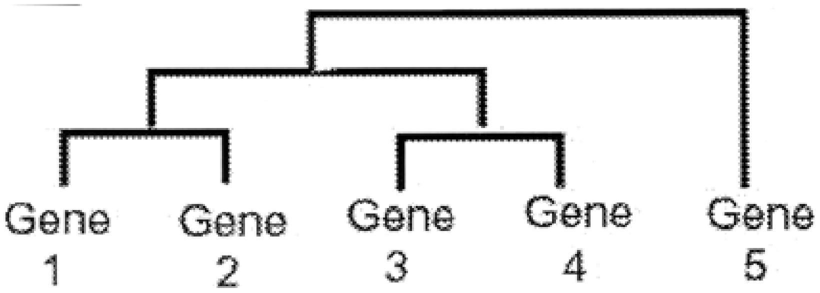

Fig. (5). Hierarchical clustering based on expression of five genes.

two antibiotic drugs A and C, each with a known mode action, such as that in Table 4. Drug A targets the bacterial cell membrane while drug C targets the ribosome. We want to know the mode of action of drugs B, D, and E.

Consider the following hypothetical experiment. We grow the bacteria of interest in a test tube, take three samples from the tube, and treat each of the three samples with one of the three drugs. We then measure the expression levels of each of, say, 1000 genes. We calculate the pairwise correlation of gene expression among the five drugs. If the expression pattern of drug B is most similar to that of drug A, as in this example (the correlation, r, is near 1.0), then the mode of action of drug B is most like that of drug A. Of course, in practice, we would prefer to use multiple drugs with the same mode of action for each class, and to replicate the experiment, to have confidence in the results. We may wish to display the correlations among several drugs, in which case we may use cluster analysis, as shown in Figure 6.

Table 4. Expression Levels of Genes in a Tissue Treated with Five Drugs

	Mode of action	Gene 1	Gene 2	Gene 3	…	Gene 1000
Drug A	Cell membrane	0	100	65	…	0
Drug B	?	0	98	63		1
Drug C	Ribosome	80	0	70	…	100
Drug D	?	81	0	75	…	120
Drug E	?	18	4	44		7

We can use correlation to examine the effect of a compound or drug on the expression levels of a gene. Consider the possible effects of the compound benzene on the expression levels of two genes, as shown in Figures 7 and 8.

In Figure 7, we see that as we increase the concentration of benzene, there is an increase in the level of expression of the first gene (high correlation). In Figure 8, we see that as we increase the concentration of benzene, there no clear pattern in the level of expression of the second gene (low correlation). The first result would indicate that benzene

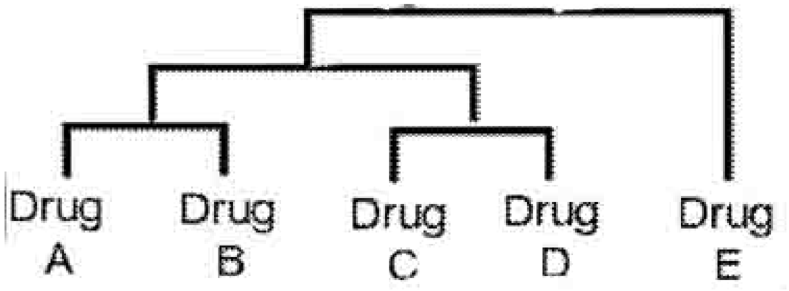

Fig. (6). Clustering of drugs to indicate mode of action.

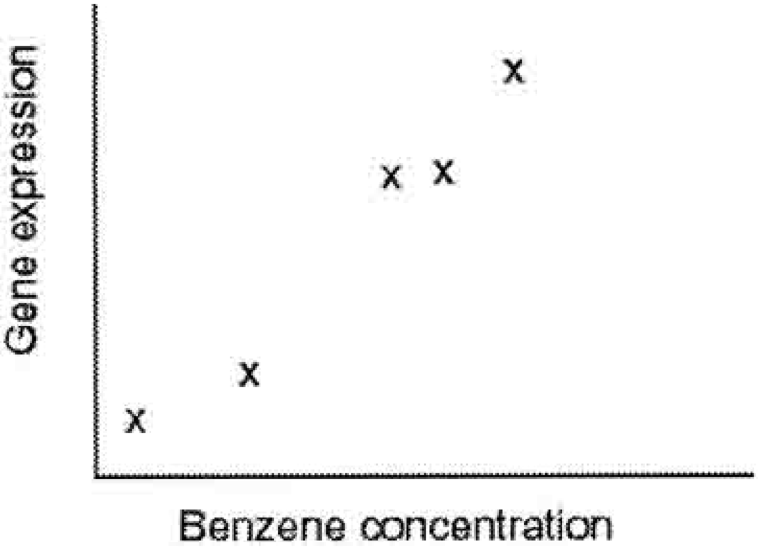

Fig. (7). Gene expression correlated with benzene concentration.

affects the first gene, while the second result would indicate that benzene has little effect on the expression of the second gene. While correlation indicates that there is a relationship between two genes, or between a gene's expression level and the concentration of a compound such as benzene, it does not tell us quantitatively how much the gene expression increases per unit of benzene. Regression analysis provides us with such quantitative information, and might indicate, for example, that for every unit increase in benzene there is a doubling of the gene expression (regression slope = 2). Quantitative information from regression analysis guides decisions on how much drug to give to achieve the desired response.

3. AN EXAMPLE: CO-EXPRESSION OF KNOWN CARDIOMYOPATHY-ASSOCIATED GENES

In this section, we present an example, using the chi-square method described above, of the co-expression of genes associated with cardiomyopathy. In this analysis, we examined the expression pattern of human genes in cDNA

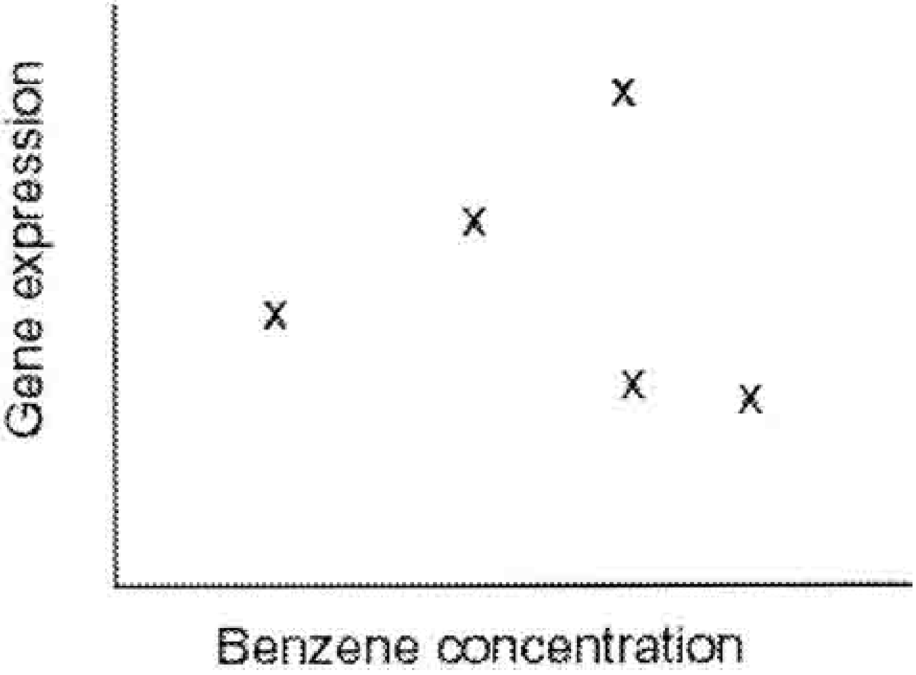

Fig. (8). Gene expression not correlated with benzene concentration.

Table 5. Known Cardiomyopathy-Associated Genes

#	Gene description
1	Atrial regulatory myosin. Regulatory isoform in atrial muscle. Differentially expressed in cardiovascular development and disease. [16, 17]
2	Cardiac alpha-myosin heavy chain. Altered expression in heart failure. Mutation in myosin heavy chain causes hypertrophic cardiomyopathy. [18-20]
3	Cardiac myosin alkali (essential) light chain (ventricular) Differentially expressed in myocardial hypertrophy. [16, 18, 19, 21, 22]
4	Cardiac troponin. Marker of cardiac injury. [23-25]
5	Cardiac ventricular myosin. Expressed in remodelling after infarction. [16, 21, 26]
6	Cardiodilatin (atrial natriuretic factor). Induces vasorelaxation. Differentially expressed following myocardial infarction. [27, 28]
7	Creatine kinase M. Marker of cardiac injury. [23-25]
8	Myoglobin. Marker of cardiac injury. [23-25]
9	Natriuretic peptide precursor. See cardiodilatin. [27, 28]
10	Sarcomeric mitochondrial creatine kinase. Essential enzyme in energy metabolism, particularly in tissue with high energy requirements such as heart. [29, 30]
11	Telethonin. Sarcomeric protein of heart and skeletal muscle. [31, 32]
12	Titin. Temporal and spatial control of sarcomere assembly. Differentially expressed after atrial fibrillation. [32, 33]
13	Troponin C. Troponins are markers of cardiac injury. [23-25]

prepared from 522 libraries of diverse anatomic and pathologic origin. Genes selected at random in this data set typically show a probability of co-expression due to chance of 10E-3 or greater, as measured using the Fisher exact test. For example, if we examine the co-expression of two genes with no known relationship, myosin and elongation factor 1-alpha, there is no evident pattern to their co-occurrences, and the probability that their (seemingly random) co-expression is due to chance is about 0.1. By contrast, genes with known relationships usually have p-values less than 10E-6.

To illustrate the co-expression of functionally related genes, consider the set of 13 genes known to be involved in cardiomyopathy listed in Table 5. The co-expression of these genes is shown in Table 6. Each entry in this table is the negative log of the probability that the observed association is due to chance (for example, a *p*-value of 0.00001 yields an entry of –log(0.00001) = –log(10E-5) = 5. Thus, large values in the table indicate very small probability. From this table, we see that the analysis readily identifies that these known genes are co-expressed, and thus likely to be related in function, when compared to the p-

Table 6. Co-Expression of 13 Known Cardiomyopathy-Associated Genes (Numbered 1 Through 13). Table Entries are Negative Log of the Probability that the Observed Association is Due to Chance

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	36	16	16	16	14	15	14	16	15	16	15	14	14
2	16	35	14	15	14	12	15	16	13	16	17	13	16
3	16	14	56	16	26	9	25	26	10	19	26	22	28
4	16	15	16	41	16	12	18	16	9	14	13	11	19
5	14	14	26	16	85	8	39	37	8	22	29	30	28
6	15	12	9	12	8	46	9	11	9	10	8	9	8
7	14	15	25	18	39	9	80	41	8	26	32	37	33
8	16	16	26	16	37	11	41	85	10	27	35	31	35
9	15	13	10	9	8	9	8	10	22	9	9	7	10
10	16	16	19	14	22	10	26	27	9	68	22	21	23
11	15	17	26	13	29	8	32	35	9	22	63	27	27
12	14	13	22	11	30	9	37	31	7	21	27	79	25
13	14	16	28	19	28	8	33	35	10	23	27	25	85

values observed for unrelated genes. Five previously uncharacterized genes (Genbank AW755250 to AW755254) are co-expressed with the 13 known genes and may also be associated with cardiomyopathy.

CONCLUSIONS

Statistical analysis of gene expression data provides a method to identify disease-associated genes, to indicate mode of action of a compound, and to identify and quantify the effects of a compound or drug on the expression levels of a gene. The methods we examined include differential expression (discriminant analysis and analysis of variance) and co-expression (correlation, association, and clustering). Expression analysis can be useful to find previously-uncharacterized disease associated genes, even if those genes show no significant sequence similarity to known genes, and thus are first functionally characterized by statistical analysis of expression data. Such genes are potentially useful as diagnostic or prognostic markers, as drug targets or therapeutic proteins, or in gene therapy.

ACKNOWLEDGEMENTS

Colleagues at Incyte Genomics contributed significantly to the development of the co-expression algorithm described here, in particular, Wayne Volkmuth, Einat Sprinzak, Tod Klingler, and David Hodgson.

REFERENCES

[1] Lashkari, D. A.; DeRisi, J. L.; McCusker, J. H.; Namath, A. F.; Gentile, C.; Hwang, S. Y., et al. *Proc. Natl. Acad. Sci. USA*, **1997**, *94*, 13057-62.

[2] Lockhart, D. J.; Dong, H.; Byrne, M. C.; Follettie, M. T.; Gallo, M. V.; Chee, M. S., et al. *Nat. Biotechnol.*, **1996**, *14*, 1675-80.

[3] D'Haeseleer, P.; Liang, S., Somogyi, R. *Bioinformatics*, **2000**, *16*, 707-26.

[4] Gnanadesikan, R. *Statistical Science*, **1989**, *4*, 34-69.

[5] DeRisi, J.; Penland, L.; Brown, P. O.; Bittner, M. L.; Meltzer, P. S.; Ray, M., et al. *Nat. Genet.*, **1996**, *14*, 457-60.

[6] Fannon, M. R. *Trends Biotechnol.*, **1996**, *14*, 294-8.

[7] Vasmatzis, G.; Essand, M.; Brinkmann, U.; Lee, B., Pastan, I. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 300-4.

[8] Zhang, L.; Zhou, W.; Velculescu, V. E.; Kern, S. E.; Hruban, R. H.; Hamilton, S. R., et al. *Science*, **1997**, *276*, 1268-72.

[9] Greller, L. D. and Tobin, F. L. *Genome Res.*, **1999**, *9*, 282-296.

[10] Eisen, M. B.; Spellman, P. T.; Brown, P. O., Botstein, D. *Proc. Natl. Acad. Sci. USA*, **1998**, *95*, 14863-8.

[11] Michaels, G. S.; Carr, D. B.; Fuhrman, S.; Wen, X., Somogyi, R., Cluster analysis and data visualization of large-scale gene expression data, in Pacific Symposium

on Biocomputing, R. Altman, *et al.*, Editors. 1998. World Scientific: Singapore. 42.

[12] Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E., et al. *Proc. Natl. Acad. Sci. USA,* **1999**, *96*, 2907-2912.

[13] Walker, M. G.; Volkmuth, W.; Sprinzak, E.; Hodgson, D., Klingler, T. *Genome Res.*, **1999**, *9*, 1198-203.

[14] Wen, X.; Fuhrman, S.; Michaels, G. S.; Carr, D. B.; Smith, S.; Barker, J. L., et al. *Proc. Natl. Acad. Sci. USA,* **1998**, *95*, 334-9.

[15] Walker, M. G.; Volkmuth, W., Klingler, T. in Intelligent Systems in Molecular Biology. 1999: AAAI Press, Menlo Park, CA.

[16] Fewell, J. G.; Hewett, T. E.; Sanbe, A.; Klevitsky, R.; Hayes, E.; Warshaw, D., et al. *J. Clin. Invest.,* **1998**, *101*, 2630-9.

[17] Hailstones, D.; Barton, P.; Chan-Thomas, P.; Sasse, S.; Sutherland, C.; Hardeman, E., et al. *J. Biol. Chem.*, **1992**, *267*, 23295-300.

[18] Sakai, S.; Miyauchi, T.; Kobayashi, T.; Yamaguchi, I.; Goto, K., Sugishita, Y. *J. Cardiovasc. Pharmacol*., **1998**, *31*, S302-5.

[19] Swynghedauw, B. Molecular cardiology for the cardiologist. 2 ed. **1998**, Boston: Kluwer.

[20] Epstein, N. D. *Adv. Exp. Med. Biol.*, **1998**, *453*, 105-14.

[21] Morano, I.; Hadicke, K.; Haase, H.; Bohm, M.; Erdmann, E., Schaub, M. C. *J. Mol. Cell Cardiol.,* **1997**, *29*, 1177-87.

[22] Schneider, M. D. and Parker, T. G. in Molecular basis of cardiology, R. Roberts, Editor. **1993**, Blackwell Scientific: Boston. 113-134.

[23] Feng, Y. J.; Chen, C.; Fallon, J. T.; Lai, T.; Chen, L.; Knibbs, D. R., et al. *Am. J. Clin. Pathol.*, **1998**, *110*, 70-7.

[24] Luscher, M. S.; Ravkilde, J., Thygesen, K. *Cardiology*, **1998**, *89*, 222-8.

[25] Kost, G. J.; Kirk, J. D., Omand, K. *Arch .Pathol. Lab. Med.*, **1998**, *122*, 245-51.

[26] Trahair, T.; Yeoh, T.; Cartmill, T.; Keogh, A.; Spratt, P.; Chang, V., et al. *J. Mol. Cell Cardiol.*, **1993**, *25*, 577-85.

[27] Gidh-Jain, M.; Huang, B.; Jain, P.; Gick, G., El-Sherif, N. *J. Mol. Cell Cardiol.,* **1998**, *30*, 627-37.

[28] Magga, J.; Vuolteenaho, O.; Tokola, H.; Marttila, M., Ruskoaho, H. *Ann. Med.*, **1998**, *30* Suppl 1, 39-45.

[29] Klein, S. C.; Haas, R. C.; Perryman, M. B.; Billadello, J. J., Strauss, A. W. *J. Biol. Chem.*, **1991**, *266*, 18058-65.

[30] Qin, W.; Khuchua, Z.; Klein, S. C., Strauss, A. W. *J. Biol. Chem.*, **1997**, *272*, 25210-6.

[31] Valle, G.; Faulkner, G.; De Antoni, A.; Pacchioni, B.; Pallavicini, A.; Pandolfo, D., et al. *FEBS Lett.*, **1997**, *415*, 163-8.

[32] Mayans, O.; van der Ven, P. F.; Wilm, M.; Mues, A.; Young, P.; Furst, D. O., et al. *Nature*, **1998**, *395*, 863-9.

[33] Ausma, J.; Wijffels, M.; van Eys, G.; Koide, M.; Ramaekers, F.; Allessie, M., et al. *Am. J. Pathol.*, **1997**, *151*, 985-97.