

Pharmaceutical target discovery using Guilt-by-Association: schizophrenia and Parkinson's disease genes

Michael G. Walker, Wayne Volkmuth, Tod M. Klingler

Incyte Pharmaceuticals
3174 Porter Drive, Palo Alto
California, USA, 94304

mwalker@incyte.com
voice: 650 845-5771
fax: 650 855-0572

Abstract

We wish to identify genes associated with disease. To do so, we look for novel genes whose expression patterns mimic those of known disease-associated genes, a method we call Guilt-by-Association (GBA). GBA uses a combinatoric measure of association that provides superior results to those from correlation measures used in previous expression analyses. Using GBA, we have examined the expression of 40,000 human genes in 522 cDNA libraries, and have identified several hundred genes associated with known cancer, inflammation, steroid-synthesis, insulin-synthesis, neurotransmitter processing, matrix remodeling and other disease genes. The majority of the genes thus discovered show no significant sequence similarity to known genes, and thus could not have been identified by homology searches. We present here an example of the discovery of five genes associated with schizophrenia and Parkinson's disease. Of the 40,000 most-abundant human genes, these five genes are the most closely linked to the known disease genes, and thus are prime targets for pharmaceutical intervention.

1.0 Previous related work

Genes that are differentially expressed in disease states are candidates for pharmaceutical intervention. Previous researchers have collected expression data for up to 10,000 genes simultaneously (Lockhart, Dong et al. 1996; Lashkari, DeRisi et al. 1997), have identified genes differentially expressed in cancer (DeRisi, Penland et al. 1996; Fannon 1996; Zhang, Zhou et al. 1997; Vasmatazis, Essand et al. 1998), and have identified clusters of co-expressed genes (Michaels, Carr et al. 1998; Wen, Fuhrman et al. 1998). Previous work has focussed on differential expression (for example, in healthy versus diseased tissue), but has rarely examined the joint

expression of novel genes with known disease genes. In addition, previous work has examined a small fraction of the total genome (typically 10,000 genes or less) and has used linear or monotonic measures of association, which are unsuitable for many known gene associations.

2.0 Prediction of gene function

To identify genes that are candidate therapeutic targets, we look for novel genes whose expression patterns mimic those of known disease-associated genes, using Guilt-by-Association (GBA). For the analyses presented here, we examined the expression of 40,000 human genes in 522 cDNA libraries, from a broad range of anatomic sites and pathologic states. In some cases, the libraries were normalized or subtracted to increase complexity.

For the purpose of this analysis, we consider a gene to be present (expressed) in a library if cDNA corresponding to that gene are detected in the sample taken from that library. We consider a gene to be absent (not expressed) in a library when no cDNA for that gene is detected in the library. To determine whether two genes, A and B, have similar expression patterns, we examine their occurrences in the 522 cDNA libraries, as shown in Table 1. A 0 indicates that the gene was not detected in the library; a 1 indicates that it was detected.

Table 1. Expression data for hypothetical genes A and B.

| Gene | Library 1 | Library 2 | ... | Library N |
|------|-----------|-----------|-----|-----------|
| A | 1 | 1 | ... | 0 |
| B | 1 | 0 | ... | 0 |

For a given pair of genes, the expression data in Table 1 can be summarized in a 2 by 2 contingency table. Table 2 presents such a co-expression contingency table for the hypothetical genes A and B in a total of 30 libraries; Table

3 presents the same data as variables that we will use shortly.

We determine the probability that the co-expression shown in Table 2 occurs by chance with a counting method, as follows. We take as our null hypothesis that there is no association between gene A and gene B. Under the null hypothesis, the marginal counts in Tables 2 and 3 are fixed, the expected count in each cell is a function of the marginals, and deviations from the expected count are random.

The number of ways that k occurrences of a gene can be distributed in r libraries is $(r \text{ C } k)$, that is, the combinatoric choose function. From Table 3, we can calculate the probability of observing n11 counts using the hypergeometric distribution, as in a Fisher Exact test (Agresti 1990). From the hypergeometric distribution, the probability of observing exactly n11 counts is $p(n11) = (n1. \text{ C } n11) \times (n2. \text{ C } n21) / (n.. \text{ C } n.1)$.

Table 2. Summary of co-expression for genes A and B from Table 1.

| 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 |

Table 3. Variables representing counts of gene occurrences.

| Number of libraries | Gene A present | Gene A absent | Total |
|---------------------|----------------|---------------|-------|
| Gene B present | n11 | n12 | n1. |
| Gene B absent | n21 | n22 | n2. |
| Total | n.1 | n.2 | n.. |

To determine if there is association (lack of independence) between the genes, we calculate the sum of all the (hypergeometric) probabilities for outcomes at least as extreme as the observed outcome. As a concrete example, consider the n11 count in the cell {Gene A present and Gene B present} in Table 2. We can calculate the probability of observing a count of exactly 8 using the hypergeometric distribution, that is, $p(n11 \text{ is } 8) = (10 \text{ C } 8) \times (20 \text{ C } 2) / (30 \text{ C } 10)$. To test the null hypothesis, we are interested not only in the case in which we observe a count of exactly 8 in the cell, but also the cases in which we observe more extreme values of n11, subject to the constraints of the marginals. Hence, we sum the probability of the observed count and of the more extreme possible counts (n11 = 8, 9, and 10) to determine the total

probability of counts at least as extreme as those observed. In the case of Table 2, the probability that the observed co-expression is due to chance is $p = 0.0003$.

This method of estimating the probability for co-expression of two genes makes several assumptions that do not hold strictly. Because more than one library may be obtained from a single patient (e.g., tumor and non-tumor tissue), libraries are not completely independent. In addition, because we perform multiple statistical tests on each gene, we require strict probabilities, using, for example, a Bonferroni correction to determine significance.

We next describe how we use GBA to identify genes associated with diseases of neurotransmitter processing.

3.0 Diseases associated with defects in neurotransmitter processing

A sufficient quantity of the neurotransmitter dopamine is ordinarily synthesized within the brain in the dopaminergic neurons of the substantia nigra. Dopamine functions in motor control; Parkinson's disease is triggered by death of the dopaminergic neurons (Jenner 1998). Insufficient dopamine leads to tremors and rigidity, which are the main clinical manifestations of Parkinson's disease (Birkmayer and Riederer 1986). Excess dopamine is associated with schizophrenia. Interactions among dopamine, norepinephrine and other neurotransmitters are involved in the disorder (Carlsson, Hansson et al. 1997).

4.0 Known neurotransmitter-processing genes

For the analysis described here, we examined a set of 10 known dopamine- and related neurotransmitter-processing genes: L-tyrosine hydroxylase (TH), AADC, dopamine β hydroxylase (DBH), nicotinic acetylcholine receptor $\alpha 3$ subunit precursor (nAChR- $\alpha 3$), secretogranin I and II, Rab3a, human cocaine and amphetamine-regulated transcript (hCART), vesicular monoamine transporter 1 (hVMAT1), and ARIX homeodomain protein.

Table 4 shows the five genes that have the strongest association with TH, an enzyme that participates in the synthesis and release of dopamine and norepinephrine, with the probability that the observed co-expression is due to chance. In Tables 4 through 6, the column headings have the following meanings.

- p*-value: The probability that the co-expression is due to chance.
- Associated gene: A gene that shows significant co-expression with the target gene.
- Occurs: The number of libraries in which the associated gene occurs.
- Both occur: The number of libraries in which both genes

occur.

Target only: The number of libraries in which only the target gene occurs.

Assoc only: The number of libraries in which only the co-expressed (associated) gene occurs.

Neither: The number of libraries in which neither of the genes occur.

Table 4. Known genes most strongly co-expressed with the TH gene.

| p-value | Associated gene | Occurs | Both | Target only | Assoc only | Neither |
|----------|-----------------|--------|------|-------------|------------|---------|
| 1.47E-13 | DBH | 12 | 9 | 6 | 3 | 503 |
| 6.11E-11 | nAchR-a3 | 13 | 8 | 7 | 5 | 501 |
| 6.22E-10 | Secretogranin I | 45 | 11 | 4 | 34 | 472 |
| 3.51E-09 | Human dlk mRNA | 52 | 11 | 4 | 41 | 465 |
| 1.01E-08 | Homolog to ECE | 13 | 7 | 8 | 6 | 500 |

TH occurs in 15 of 522 libraries studied, and shows strong co-expression with genes known to be related to neurotransmitter processing, including DBH, nAchR, secretogranin I, and a homolog of endothelin converting enzyme (ECE). The endothelins (ET's) are a class of secreted peptides that are synthesized from inactive precursor peptides, three of which are known. ET-1 and ET-3 induce dopamine release (Horie, Morita et al. 1995). The third and final processing step leading to the active ET is catalyzed by a member of the class of endothelin converting enzymes (ECE's).

We observed similar results to those for TH for the other nine known genes; in particular, the known dopamine-processing genes exhibit clear co-expression. Several known genes, not previously reported to be involved in neurotransmitter processing, were also significantly co-expressed with one or more of the 10 known genes. We next examine the application of co-expression analysis to predict the function of novel genes.

5.0 Identification of novel neurotransmitter-associated genes

We identified five novel genes by testing for the co-expression of their RNA with that from known neurotransmitter-processing genes. Table 5 shows the known genes that are most closely associated with NTP-1, one of the five novel genes identified here.

NTP-1 occurred in 14 of 522 cDNA libraries studied and showed strongest co-expression with the known neurotransmitter-processing genes DBH, nAchR, TH type

4, and secretogranin II, as shown in Table 5.

Table 5. Co-expression results for gene NTP-1.

| p-value | Associated gene | Occurs | Both | Target only | Assoc only | Neither |
|---------|------------------|--------|------|-------------|------------|---------|
| 1.3E-9 | DBH | 12 | 7 | 7 | 5 | 502 |
| 2.8E-9 | nAchR | 13 | 7 | 7 | 6 | 501 |
| 1.0E-8 | TH type 4 | 15 | 7 | 7 | 8 | 499 |
| 2.5E-7 | Secretogranin II | 33 | 8 | 6 | 25 | 482 |

NTP-2 occurred in three of 522 cDNA libraries studied and showed strongest co-expression with the known neurotransmitter-processing genes DBH, nAchR, TH type 4, and AADC, as shown in Table 6.

Table 6. Co-expression results for gene NTP-2.

| p-value | Associated gene | Occurs | Both | Target only | Assoc only | Neither |
|----------|-----------------|--------|------|-------------|------------|---------|
| 9.39E-06 | DBH | 12 | 3 | 0 | 9 | 509 |
| 1.22E-05 | nAchR | 13 | 3 | 0 | 10 | 508 |
| 1.94E-05 | TH type 4 | 15 | 3 | 0 | 12 | 506 |
| 3.48E-05 | AADC | 18 | 3 | 0 | 15 | 503 |

We found similar patterns of association for the remaining three genes, NTP-3, 4, and 5, which, to conserve space, we do not report here. None of the five genes show significant sequence similarity to known genes.

6.0 Comparison of GBA to correlation methods

For the method of analysis described here, we reduce each expression datum to a binary variable (present or absent), rather than analyzing expression as a continuous variable using linear or rank correlation. Before we chose this binary-encoding method to identify co-expressed genes, we evaluated Pearson linear correlation and Spearman rank correlation using continuous values. These correlation methods were less sensitive than the Fisher exact test in detecting known relationships, in many cases. Specifically, we observed that genes with known biological relationships commonly had correlation coefficients comparable to or less than genes with no known or plausible relationship; this observation made it difficult to be confident in the biological significance of correlations between previously uncharacterized genes and genes with known function.

These are several possible reasons why correlation measures may identify associations with less confidence than does GBA for these data. We note several here. Many

genes that are known to be associated do not exhibit the simple linear or monotonic relationships assumed by these methods. To increase complexity, libraries may in some cases be normalized and/or subtracted, which may confound linear association measures. Quantitative measurement of expression has sufficiently high variability (inability to accurately distinguish two-to-three fold changes, particularly at low expression levels) that correlation measures may vary dramatically with resampling of the same experiment. Finally, most genes are not expressed in most tissues (or are expressed at levels below the detection limit of current assays); when a pair of genes have measured expression of zero across several hundred libraries, and non-zero values in a few score libraries, they show high correlation regardless of their joint expression in libraries where they have non-zero values.

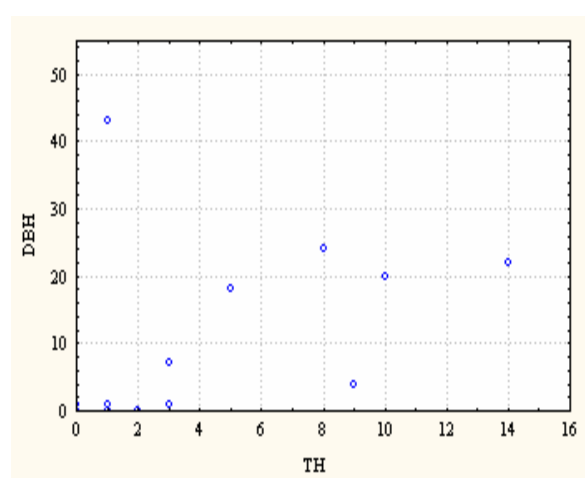


Figure 1. Scatterplot of the expression of TH versus DBH in the 522 libraries. Both genes have zero measured expression in 503 of 522 libraries, hence most points are at (0,0).

The problem then becomes to distinguish spurious correlation from biologically meaningful correlation. We illustrate this problem using TH, AADC and DBH, three of the genes from the neurotransmitter data set. Recall from Table 1 that TH occurs in 15 libraries and DBH occurs in 12 libraries. Table 1 also indicates that that TH and DBH occur together in 9 libraries, and both are absent (zero expression) in 503 libraries. Because both genes have zero expression in the vast majority of libraries (503 of 522), they will inevitably have high correlation. Figure 1 is a scatterplot of the expression of TH and DBH in the 522 libraries.

The Pearson correlation coefficient for TH and DBH is 0.66; the Spearman rank correlation is 0.67, and the GBA p -value is 1.5 E-13. Which of these values is the best detector of the known biological relationship between the

two genes? To determine the answer to this question, we calculated the pairwise correlations and the Fisher p -values for 1040 randomly-selected genes (540,280 pairs), and compared these values to those for TH, DBH, and AADC. Table 7 shows the results of this comparison.

Table 7. Comparison of GBA (Fisher) to correlation measures for TH, DBH & AADC.

| Gene pair | TH-DBH | TH-AADC | AADC-DBH |
|------------------|---------|---------|----------|
| GBA (Fisher) p | 1.5E-13 | 8.9E-8 | 5.6E-7 |
| % < Fisher p | 3.7E-6 | 5.9E-5 | 1.5E-4 |
| Pearson r | 0.66 | 0.6 | 0.5 |
| % > Pearson r | 5.4E-3 | 8.7E-3 | 2.4E-3 |
| Spearman r | 0.67 | 0.41 | 0.4 |
| % > Spearman r | 1.8E-6 | 2.8E-4 | 3.9E-4 |

For each pair of genes, Table 7 indicates the following.

The pair of genes examined.

The Fisher p -value for the pair.

The percent of the 540,280 randomly selected pairs that had a Fisher p -value less than that of the gene pair

The Pearson linear correlation coefficient, r , for the pair.

The percent of the 540,280 randomly selected pairs that had a Pearson correlation coefficient greater than that of the pair.

The Spearman rank correlation coefficient, r , for the pair.

The percent of the 540,280 randomly selected pairs that had a Spearman correlation coefficient greater than that of the pair.

For these three genes, which are known members of the neurotransmitter synthesis pathway, there are fewer of the random genes with a more extreme GBA Fisher p -value than there are random genes that have a more extreme Pearson correlation coefficient. Thus, in this case, the Fisher test is a better detector of the known association than is the Pearson coefficient. For these genes, the Spearman coefficient is consistently superior to the Pearson, while the Fisher test is superior to the Spearman in two out of three. We have not yet carried out a comprehensive test with known genes using this method of comparison, but our empirical experience to date indicates that Fisher is superior to the correlation methods sufficiently often that it should, at least, be used in conjunction with those methods.

6.0 Conclusions

We have analyzed the expression patterns of over 40,000 genes in over 500 libraries (the largest such expression analysis reported to date), and have identified several hundred disease-associated genes using a novel co-expression algorithm, Guilt-by-Association. We describe here the discovery of five genes associated with neurotransmitter processing. These genes are potentially useful for the diagnosis and treatment of neurotransmitter-processing related diseases such as Parkinson's disease and schizophrenia. Of the 40,000 most-abundant human genes, these genes are the most closely linked to the known disease genes, and thus are prime targets for pharmaceutical intervention.

Acknowledgements

We thank our colleagues at Incyte for their support and assistance in this research, and our editor, Lyn Dupre, and the reviewers for their many helpful suggestions.

References

- Agresti, A. (1990). Categorical data analysis. New York, Wiley.
- Birkmayer, W. and P. Riederer (1986). Understanding the Neurotransmitters : Keys to the Workings of the Brain. New York, Springer-Verlag.
- Carlsson, A., L. O. Hansson, et al. (1997). "Neurotransmitter aberrations in schizophrenia: new perspectives and therapeutic implications." Life Sci **61**(2): 75-94.
- DeRisi, J., L. Penland, et al. (1996). "Use of a cDNA microarray to analyse gene expression patterns in human cancer [see comments]." Nat Genet **14**(4): 457-60.
- Fannon, M. R. (1996). "Gene expression in normal and disease states --identification of therapeutic targets." Trends Biotechnol **14**(8): 294-8.
- Farde, L. (1997). "Brain imaging of schizophrenia--the dopamine hypothesis." Schizophr Res **28**(2-3): 157-62.
- Horie, K., A. Morita, et al. (1995). "Endothelin-1 and endothelin-3 modulate dopaminergic neurons through different mechanisms." Life Sci **57**(8): 735-41.
- Jenner, P. (1998). "Oxidative mechanisms in nigral cell death in Parkinson's disease [In Process Citation]." Mov Disord **13**(Suppl 1): 24-34.
- Lashkari, D. A., J. L. DeRisi, et al. (1997). "Yeast microarrays for genome wide parallel genetic and gene expression analysis." Proc Natl Acad Sci U S A **94**(24): 13057-62.
- Lockhart, D. J., H. Dong, et al. (1996). "Expression monitoring by hybridization to high-density oligonucleotide arrays [In Process Citation]." Nat Biotechnol **14**(13): 1675-80.
- Michaels, G. S., D. B. Carr, et al. (1998). Cluster analysis and data visualization of large-scale gene expression data. Pacific Symposium on Biocomputing. R. Altman, A. Dunker, L. Hunter and T. Klein. Singapore, World Scientific: 42.
- Vasmatzis, G., M. Essand, et al. (1998). "Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis." Proc Natl Acad Sci U S A **95**(1): 300-4.
- Wen, X., S. Fuhrman, et al. (1998). "Large-scale temporal gene expression mapping of central nervous system development." Proc Natl Acad Sci U S A **95**(1): 334-9.
- Zhang, L., W. Zhou, et al. (1997). "Gene expression profiles in normal and cancer cells." Science **276**(5316): 1268-72.