

Testing Gene Class Enrichment in High-throughput Genomics

David L. Gold^{1, 2}, Kevin R. Coombes², Jing Wang², Bani Mallick¹

Department of Statistics, Texas A & M University, College Station, TX 77843-3143, and
²Department of Biostatistics and Applied Mathematics, The University of Texas M. D.
Anderson Cancer Center, Houston, TX 77030-4009.

*Address correspondence to:

David L. Gold

The University of Texas M.D. Anderson Cancer Center

Department of Biostatistics & Applied Mathematics

1515 Holcombe Boulevard, Box 447

Houston TX 77030-4009

dlgold@mdanderson.org

Running Title: Testing for Class Enrichment in High-throughput Genomic Experiments

Key words: GO Ontology, Functional Enrichment, Gene Set Enrichment Analysis,
Microarray

ABSTRACT

Motivation: Gene set enrichment analysis seeks to infer if any biological classes of genes are over represented in a set of differentially expressed genes in a microarray experiment. The conventional approaches assume independence between gene classes. Genes may be annotated with more than one biological classification, and thus the classes that share genes are dependent. Assuming independence may lead to faulty conclusions.

Method: We derive the null distribution of gene counts for GO classes with dependence and use a normal approximation to investigate the benefits of allowing for dependence in the null distribution. We illustrate with simulations and compare our approach against univariate Fisher Exact Test results with publically available data sets.

Results: Simulations show that the level of correlation in gene counts under the null is non-ignorable. Our hypothetical examples demonstrate the loss in power or level incorrectly assuming independence. In the public data sets we analyzed, though, the relative conclusions were similar, whether assuming independence or allowing for dependence in the null distribution.

Availability: R code is available form the corresponding author by e-mail.

Contact: dlgold@mdanderson.org

1. INTRODUCTION

High throughput genomics experiments generate exhaustive quantities of information for in many cases thousands of genes. A bioinformatician's role does not end with an analysis to find genes significantly associated with a phenotype or outcome. The results must be interpreted. One often-adopted strategy is to work with higher-level information of common regulatory relationships or shared pathways, i.e. biological 'classes' of genes, to shift the focus to biological events, which are easier to understand and may further generate hypotheses concerning specific genes. This has been referred to as functional enrichment, pathway analysis (Curtis 2005) or gene set enrichment analysis (Mootha 2003), but we refer to it as enrichment analysis (EA).

The Gene Ontology Consortium (GO) provides a useful biological classification of genes. GO is made up 17 member organizations involved in curating database gene annotations for a diverse collection of model organisms ranging from *Plasmodium falciparum* to *Caenorhabditis elegans* all the way to *Homo sapien*. Annotations may be obtained without charge by following the conditions listed at the GO website, <http://www.geneontology.org/GO.cite.shtml>. Gene annotations are evolving at a rapid rate, and monthly release notes detailing changes and updates are publically available from the GO web site.

GO curates the gene classifications for three distinct attributes: biological process (what), molecular function (how) and cellular component (where). As of 09/08/2005 there were 9,820 terms for biological process, 7,078 for molecular function and 1,575 for cellular component. For each class there is an increasing hierarchy of terms curated for each gene. For example, a gene annotated for *regulation of signal transduction*, a

member of the biological process attribute, may further be annotated for *negative regulation of signal transduction* and further for the type of signaling cascade. Figure 1 displays the path to the biological process term *myoblast maturation*. A gene annotated with *myoblast maturation* is also associated with *cell development*, *cell differentiation* and *cellular process* in increasing orders of generalization. GO offers evidence codes with annotations such as *Inferred by Curator*, *Inferred by Direct Assay*, and so on to weaker forms of evidence such as *Not Recorded*. These evidence codes have received very little attention for analytical purposes in the literature.

GO annotations are often formatted for array analysis. A common approach is to specify a maximum depth for each 'branch' of the GO tree, and replace all sub-terms below this depth with the new leaf node. Specifying the depth of the GO hierarchy for EA is somewhat of an art, and little evidence is available to suggest an optimal way to choose biological classes for EA. Formatted annotation files are publically available for a variety of microarray platforms, e.g. from dChip (<http://www.dchip.org/>) or Affymetrix (<http://www.affymetrix.com/analysis/index.affx>).

The conventional approach to EA is to perform a Fisher exact tests for each functional category assuming independence of the results. Other tests have been suggested in the literature. Curtis (2005) provides a list of some available softwares and methods for EA, such as GO Miner (<http://discover.nci.nih.gov/gominer/>) and GOsurfer, accessible in dCHIP software. Another novel methodology includes GO-based clustering of gene sets (Venezia 2005).

Deriving a null distribution for testing over enriched biological classes is complicated by interclass dependence. Statistics based on counts of genes between

classes may be positively correlated, since many genes share common GO classes and many GO functions share genes. Positive correlation may follow for other reasons, e.g. mutually exclusive classes of genes that are related biologically. We are typically not prepared to build this information into the null hypothesis. If genes are chosen *iid* at random without replacement it is reasonable to assume positive correlation in the counts of genes between GO classes follows exclusively from shared genes.

We derive the null distribution of gene counts for GO classes and offer an independent normal approximation. We illustrate the advantages of incorporating correlation into the null hypothesis with annotations of the Affymetrix U133 plus 2 Chip[®]. We look at some selected comparisons of public microarray experiments.

2. METHOD

Let N be the number of unique genes on the array, k the number of genes found to be differentially expressed for some meaningful biological contrast and N_i , $i = 1$ to B , the number of genes on the array annotated with biological class i . We denote the respective counts of genes annotated with each biological class from the k differentially expressed genes as $n_1, n_2, n_3, \dots, n_B$. We want to infer whether we found more genes than expected from each biological class, assuming genes were sampled *iid* at random without replacement from the array. Gene counts may be negatively correlated under this sampling scheme since, given k , the first gene is chosen with probability k/N , the next with probability $(k-1)/(N-1)$ and so on.

Suppose for each biological class i we derive a statistic z_i to perform a one-way test for functional enrichment, for example $H_0: E(z_i) = I_0$ versus $H_1: E(z_i) > I_0$. If the z_i 's are independent or may be treated as independent without much loss in inference, the test

for enrichment of biological class i tells us nothing about i' . If the tests are correlated, then inferences concerning one class do inform others. Testing involves determining z_i 's outside the $(1-\alpha)\%$ null joint density region. If two tests are correlated and this dependence is ignored a resulting loss of level or power is possible.

2.1 Deriving the Null Distribution

The null distribution for testing class enrichment is derived analytically for two biologically defined GO classes. Extension of these results to more classes is not difficult. Suppose we have gene sample space with each gene belonging to one or more of two biologically defined GO classes: $b_1, b_2, b_1 \cap b_2, (b_1 \cup b_2)^c$, i.e. each gene may be annotated with classes 1, 2, both or neither. Suppose there are N_1 of N genes annotated with just class 1, N_2 annotated with just class 2 and N_3 with both. From an *iid* random sample without replacement of k genes from N , we are formally interested in the sampling distribution of $z_1 = n_1 + n_3$ and $z_2 = n_2 + n_3$, where (n_1, n_2, n_3) are the respective quantities of k genes sampled from the events $b_1, b_2, b_1 \cap b_2$, for testing the null hypothesis that either quantity (z_1, z_2) is greater than expected, given some alpha level. The sample space is partitioned into 4 events, $b_1, b_2, b_1 \cap b_2, (b_1 \cup b_2)^c$. The probability distribution of (n_1, n_2, n_3) under *iid* random sampling from N is

$$P(n_1, n_2, n_3 | k, N_1, N_2, N_3, N) = \frac{\binom{N_1}{n_1} \binom{N_2}{n_2} \binom{N_3}{n_3} \binom{N - (N_1 + N_2 + N_3)}{k - (n_1 + n_2 + n_3)}}{\binom{N}{k}} \quad (1).$$

The marginal distribution of n_i given k, N_i and N is hypergeometric with mean and variance

$$E(n_i | k, N_i, N) = N_i \frac{k}{N} \quad (2)$$

$$\text{Var}(n_i | k, N_i, N) = k \frac{N_i}{N} \left(1 - \frac{N_i}{N} \right) \left(\frac{N-k}{N-1} \right) \quad (3)$$

The joint distribution of (z_1, z_2) is derived as,

$$P(z_1 = c_1, z_2 = c_2) = \sum_{n_3}^{c_1 \vee c_2} P(n_1 = c_1 - n_3, n_2 = c_2 - n_3, n_3). \quad (4)$$

Once we have the joint probability distribution of (z_1, z_2) , we can calculate moments.

Note that for the overlapping counts of any two classes $z_1 = n_1 + n_3$ and $z_2 = n_2 + n_3$ the covariance of (z_1, z_2) is

$$\text{Cov}(n_1 + n_3, n_2 + n_3) = \text{Cov}(n_1, n_2) + \text{Cov}(n_1, n_3) + \text{Cov}(n_2, n_3) + \text{Var}(n_3). \quad (5)$$

Finite population sampling requires the covariance terms, since these counts are negatively correlated which may be computed in closed form as

$$\text{Cov}(n_i, n_j) = -\frac{k(N-k)}{N^2(N-1)} N_i N_j. \quad (6)$$

Another case of interest is genes sampled from GO classes, which are subclasses of more general terms in the GO hierarchy. Consider the same classes b_1 and b_2 above and define the quantities: $z_1 = n_1$ and $z_2 = n_1 + n_2$. The probability distribution of (n_1, n_2) is

$$P(n_1, n_2 | k, N_1, N_2, N) = \frac{\binom{N_1}{n_1} \binom{N_2}{n_2} \binom{N - (N_1 + N_2)}{k - (n_1 + n_2)}}{\binom{N}{k}} \quad (7),$$

and the conditional distribution of z_2 given z_1 is

$$P(z_2 = c_2 | z_1 = c_1) = \frac{P(n_1 = c_1, n_2 = c_2 - n_1)}{P(n_1 = c_1)} \quad (8).$$

The marginal distribution of z_2 is

$$P(z_2 = c_2) = \sum_{n_1}^{c_2} P(n_1, n_2 = c_2 - n_1) \quad (9).$$

2.2 Simulating the Null Distribution

The joint distribution of n_1, n_2, \dots under the null requires knowledge of dependency. Computing exact results is very cumbersome. An empirical approximation to the null may be simulated given k, N_i and N . Simulation involves selecting k genes at random from N , and recording the counts of genes annotated with class i, n_i^r , for $r = 1$ to R reps.

2.3 Testing Significance of Over-Enrichment

The analytical approach to testing for functional enrichment involves computing for each set of biological terms, the joint probability of counts at least as extreme as observed. Given significant joint tests, conditional tests may be explored to determine the biological terms most enriched for genes. Care should be taken in evaluating how the terms are related on the GO hierarchy when testing. If many terms are successively tested from the same GO branch, much of the same information is tested repeatedly.

The analytical calculations for testing purposes, as demonstrated above, can be cumbersome requiring new numerical iterations for each platform, annotation and experiment. The simulation approach described here offers an empirical null distribution to compare with observed counts. The empirical p -value for a joint test is computed as the percent of simulated counts as or more extreme than those observed. The empirical p -value for a conditional test of observed counts $n_1, n_2, \dots, n_h | n_h, n_{h+1}, \dots, n_{h+m}$ is

computed similarly, given the levels of the observed counts $n_h, n_{h+1}, \dots, n_{h+m}$. One may need to simulate for $R \gg$ to compute empirical p -values to a desired level of accuracy, also presenting computational challenges since multiple testing requires controlling family-wise error rates.

Alternatively, the null distribution may be approximated as multivariate normal (MVN), following from the multivariate extension of the Central Limit Theorem for proportions,

$$\sqrt{n}(\underline{p} - \underline{P}) \sim AMVN(\underline{0}, \Sigma) \quad (10)$$

as $k/N \rightarrow P_D \in (0,1)$ and $N_i/N \rightarrow P_i \in (0,1)$ $i = 1, \dots, B$, where $p_i = n_i/k$ is the proportion of genes counts observed for the i^{th} biological class from k . For a reasonable B , \mathbf{S} may be acquired from (3) and (5); $B(B+1)/2$ quantities are required. In large samples, the simulated means and covariances may substitute for known quantities.

Let the counts be transformed linearly as

$$Z = (\underline{p} - \underline{P})C \sim MVN(\underline{0}, I) \quad (11)$$

for some matrix C . The elements of Z are distributed as standard normal variates (see Appendix). The normal approximation depends on large k and N_i . Consequently, we do not advocate the normal approximation for small k or nodes on the GO hierarchy with small total gene counts N_i . Simulating data from the null distribution will be helpful in determining appropriate cutoff values of k and N_i for the normal approximation to maintain its level.

3. RESULTS

3.1 Simulation

The joint distribution of counts under the null was simulated with biological process annotations for $k = 200$ differentially expressed genes and $R = 1,000$ repetitions for genes on the Affymetrix U133 plus 2 Affymetrix Chip[®]. The distribution was approximated as multivariate normal with exact analytic moments and transformed as in (11). All of the annotations presented here were downloaded from Affymetrix in October 2005.

Figure 2 shows the empirical density of simulated correlation coefficients between counts of 73 biological process pairs, each with at least 100 genes. The mass is predominantly symmetric about zero with a long right tail, demonstrating the degree of negative and positive correlation. Table 1 shows the correlations, simulated and analytically derived, for selected pairs. Many pairs of terms with high positive correlation are related such as *protein transport* and *intracellular protein transport*, a subset of the former. The simulated correlation coefficients generally agree with truth.

We illustrate discrepancies between independent Fisher Exact Tests and *MVN* approximations with, for simplicity, two functional categories: (1) *protein transport* and (2) *intracellular protein transport*, (row 8 in Table 1) for several sets of plausible combined gene counts. Table 2 shows the results for the two hypothetical count pairs. We report the independent marginal Fisher one-way p -values, the joint Fisher test (summed over the product of marginals for all possible counts at least as extreme as observed) and the one-way and joint *MVN* p -values. The simulated counts are plotted in Figure 3.

For count pair (12, 7) the joint Fisher test p -value indicates some evidence of enrichment, disagreeing with the *MVN* p -value, which is down weighted allowing for correlation. For hypothetical counts (5, 7), the marginal Fisher tests lack evidence in

favor of enrichment, while the *MVN* tests identify the *intracellular protein transport* as functionally enriched. These hypothetical results are indicative of plausible errors, loss of power *or* level, when ignoring dependence.

3.2 Breast Cancer Metastasis (2004) Experimental Results

We examined the functional results from Wang (2004), using 60 genes detected for change by metastasis in the ER+ patient pool. Affymetrix U133A Chips[®] were hybridized with lymph node negative breast tumor samples from 286 patients. Patients were analyzed for differential gene expression by metastatic outcome given ER status. For our analysis, genes without biological process annotations were not counted towards the total, $N = 9,403$. Among the 60 genes reported by Wang for differential expression, $k = 45$ had annotations. We narrowed the analysis presented here to those biological processes with at least 100 genes. Table 3 lists the p -values from the *MVN* approximation and Fisher's Exact Test. The test yield p -values with similar order, although the *MVN* approximation tends to have smaller p -values at the top of the list. The five most significant categories are the same for both methods: *Cell cycle*, *cell proliferation*, *immune response*, *cell adhesion*, *proteolysis* and *peptidolysis*. Both methods find *Cell cycle* as highly significant. The relative rankings of the functions are similar in both, but the *MVN* approximation yields smaller p -values overall.

3.3 Leukemia Treatment Response (2003) Experimental Results

In vivo treatment response in leukemia cells (Cheok 2003) was used to identify a 150-probeset profile to discriminate between treatment specific changes. The comparisons were made on with Affymetrix U95A Chips[®] for 60 pre- and post- treatment samples from children newly diagnosed with ALL. There were $N = 7,201$ genes with

biological annotations, and of the 150 probes they identified as discriminating treatment outcome, $k = 111$ genes had annotations. Table 4 lists the p -values from the *MVN* approximation and Fisher's test for the top 10 functions from the *MVN* test. *Cell cycle* and *immune response* have the smallest p -values in both methods. The orders of the p -values are similar, but the *MVN* approximation shows smaller p -values in 8 of the ten terms.

3.3 Multiple Myeloma / Normal Comparison (2002) Experimental Results

Affymetrix HuGeneFL Chips[®] hybridized with plasma cells from 74 patients with multiple myeloma (MM) were compared with plasma from 31 healthy volunteers (Cheek 2004). There were $N = 4,845$ genes with biological annotations on the array, and of the 114 genes identified by Cheek as discriminating between groups, $k = 93$ had annotations. Table 5 lists the results. Neither method finds any of these methods to be statistically enriched, but the orders of the p -values are similar. *Cell cycle* and *regulation of transcription, DNA-dependent*, are at the top of both lists.

4. DISCUSSION

We demonstrate with simulated class counts that inter-class correlation under the null hypothesis may be non-ignorable. While calculating exact significance levels to account for correlation is computationally burdensome, the normal approximation is quick to compute. The public experimental data results show that the *MVN* and the Fisher Exact levels agree for functions with large observed gene counts, and the relative orders of the p -values are similar in both. In high dimensional spaces with sparse covariance matrices our method is not expected to show dramatic differences, but rather to raise or lower the levels of some of the more highly correlated terms. In several

datasets, the normal approximation produced smaller p -values. The normal approximation requires large k and N_i . Simulating data from the null distribution was helpful in determining appropriate cutoff values of k and N_i for the normal approximation to maintain its level. Several drawbacks of the approximation are that some of the more interesting functions may have small total gene counts and tail probabilities may be understated.

Care should be had in testing counts as one moves up a branch of the GO tree, as the same information is tested repeatedly. Our methodology very flexibly allows the biologists to select the classes, or sub-classes, of gene functions that interest them. Some overlapping classes are rich in common genes, such as regulation of metabolism or transcription, but there are also many classes that are direct subclasses of more general terms in the GO hierarchy. The *MVN* approximation to the joint distribution of class counts transformed to identity covariance does account for correlation and weight appropriately. Future work includes making use of the evidence codes offered by GO to decide how believable a given annotation really is for testing purposes.

REFERENCES

1. Anderson T.W. (1984) An Introduction to Multivariate Statistical Analysis 2nd Ed., John Wiley and Sons, Inc.
2. Curtis R.K. et al. (2005) Pathways to the analysis of microarray data, *Trends in Biotechnology* Vol.23 No.8 August
3. Johnson N.L. et al. (1997) Discrete Multivariate Distributions, John Wiley and Sons, Inc.
4. Mootha V.K. et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273
5. The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25: 25-29.
6. Venezia T.A. et al. (2004) Molecular Signatures of Proliferation and Quiescence in Hematopoietic Stem Cells, *PLoS Biology* Vol. 2, No. 10, e301 DOI: 10.1371/journal.pbio.0020301
7. Wang Y., et al. (2005) Gene-expression to predict distant metastasis of lymph-node-negative primary breast cancer, *Lancet*; 365: 671–79.

Appendix

We begin with the case of a bivariate normal random variate X , of mean 0 and covariance matrix Σ with nonzero off diagonal terms. Let the principal components of X be rotated onto the major axes and scaled to variance 1, denoted $Y = XTD$, where without loss of generalization

$$T = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

is the matrix projection of the components in X to the major axes, and $D = \text{diag}\{T'\Sigma T\}^{-1/2}$ is a diagonal matrix of rescaling constants to the variables in Y . Y is a matrix of *iid* standard normal random variables. Having been rescaled along the axes, linear transformations of Y back onto the principal components will not change the off diagonal covariance structure. Let $Z = YTH$ be the resulting matrix after rotating the axes of Y back to the principal components where $H = \text{diag}\{TD'T'\Sigma TDT'\}^{-1/2}$. Z is bivariate normal with mean vector $\underline{0}$ and covariance equal to the identity, with axes corresponding to the original axes in X . For the general K -variate normal case, the projection matrix T may be obtained as the inverse of the Eigen vector matrix of the singular value decomposition of the covariance matrix and the diagonal elements of D from root $-1/2$ of the Eigen values.

Figure Legends

Figure 1. Cutaway view from the CGAP Browser, GO Biological Process *myoblast maturation*, August 2005.

Figure 2. Histogram of Pearson correlation coefficients between 73 biological process gene class pairs of > 100 Genes, simulated for $r = 1000$ reps.

Figure 3. Scatter plot of simulated null gene counts, jittered, for *protein transport* and *intracellular protein transport*.

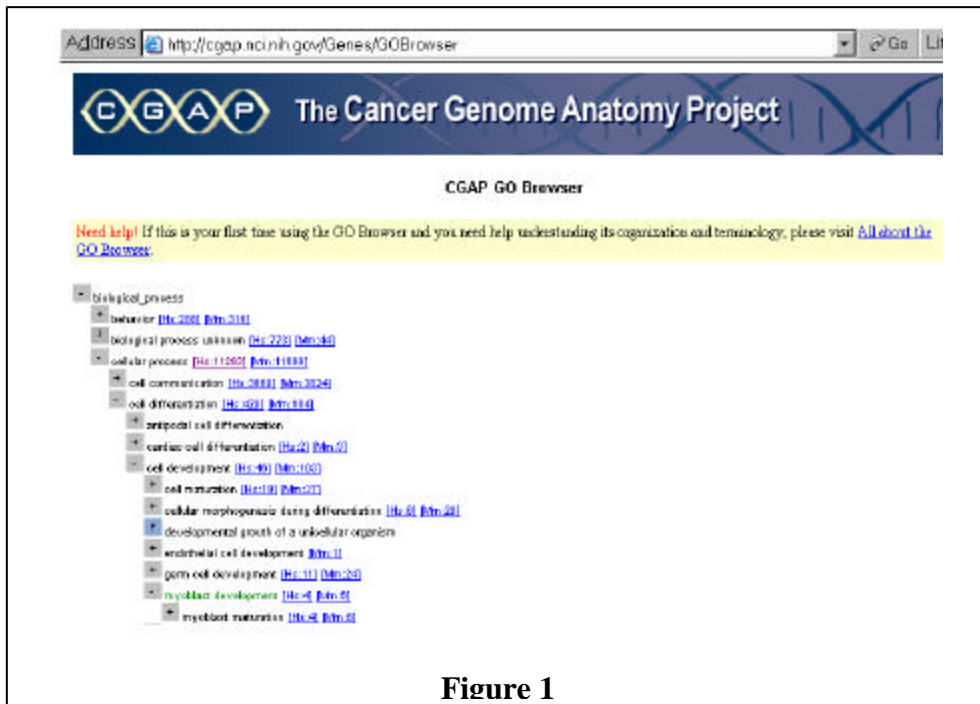
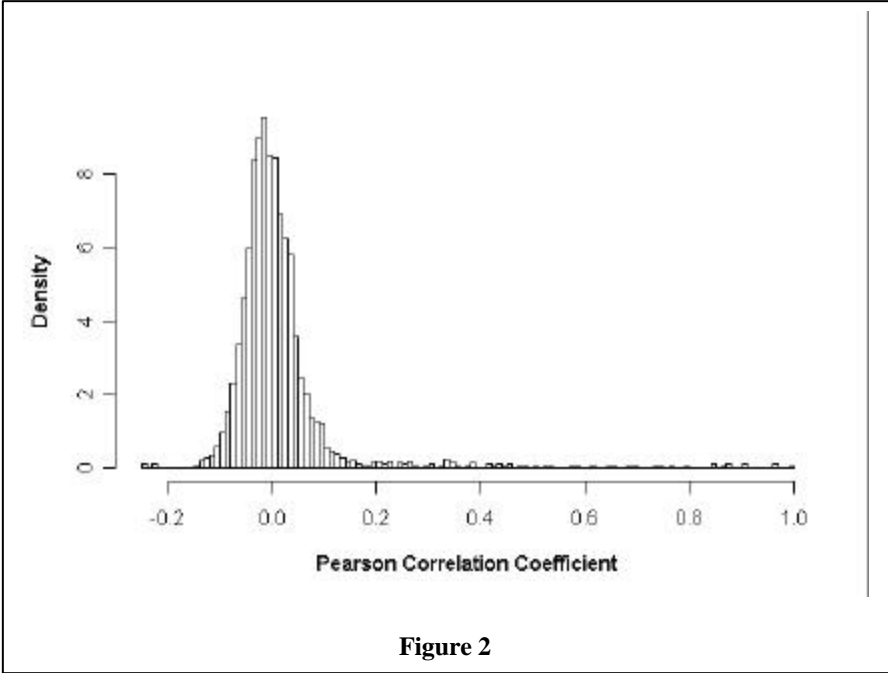


Figure 1



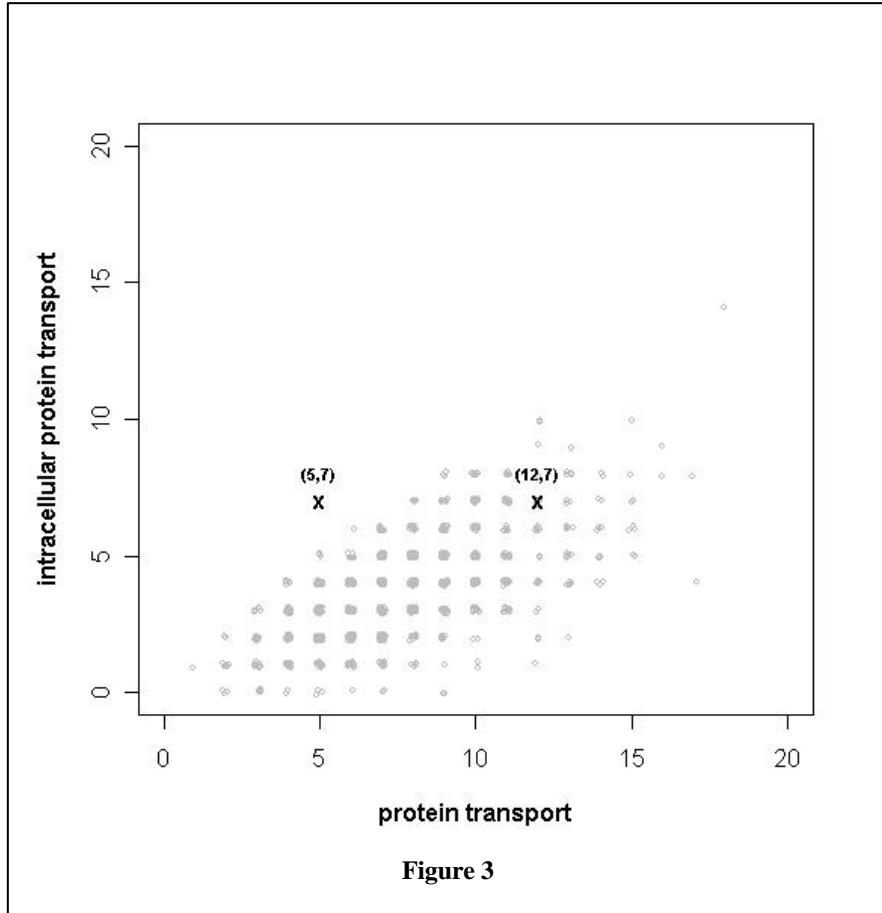


Table 1. Analytic and Simulated Correlation of Selected Pairs of Biological Processes of > 100 Genes

	Function 1	Function 2	N₁	N₂	N₁∩N₂	Simulated	Analytic
1.	<i>regulation of transcription, DNA-dependent</i>	<i>transport</i>	1625	2111	31	-0.2263	-0.1697
2.	<i>regulation of transcription</i>	<i>signal transduction</i>	1892	1670	119	-0.1303	-0.1013
3.	<i>regulation of cell proliferation</i>	<i>visual perception</i>	253	160	1	0.0024	-0.0125
4.	<i>positive regulation of cell proliferation</i>	<i>regulation of cell cycle</i>	105	298	20	0.1097	0.0996
5.	<i>transcription from RNA polymerase II promoter</i>	<i>transcription, DNA-dependent</i>	434	1631	237	0.2015	0.2302
6.	<i>regulation of transcription from RNA polymerase II promoter</i>	<i>negative regulation of transcription</i>	250	142	72	0.3401	0.3722
7.	<i>regulation of cell proliferation</i>	<i>positive regulation of cell proliferation</i>	253	105	105	0.6424	0.6401
8.	<i>protein transport</i>	<i>intracellular protein transport</i>	440	213	213	0.6589	0.6888
9.	<i>protein catabolism</i>	<i>ubiquitin-dependent protein catabolism</i>	164	119	119	0.8459	0.8502
10.	<i>dephosphorylation</i>	<i>protein amino acid dephosphorylation</i>	136	127	127	0.9610	0.9660

Table 2. Hypothetical Comparison of P-Value Test Results

Method	<i>protein transport</i> (n₁=12)	<i>intracellular protein transport</i> (n₂ = 7)	Joint Test
Fisher	0.0921	0.0851	<.0001
<i>MVN</i>	0.0997	0.0942	0.1848
Fisher	(n₁=5) 0.8695	(n₂ = 7) 0.0851	<.0001
<i>MVN</i>	0.9875	0.0005	0.0003

Table 3. Enrichment Analysis, Top 10 Biological Processes (>300 Genes) by MVN P-value of, Breast Cancer Metastasis ER+ Group, ($k=45$)

Biological Process	MVN	Fisher	E(n_i)	n_i
<i>cell cycle</i>	9.72E-06	0.0013	2.35	9
<i>cell proliferation</i>	0.0035	0.0122	2.27	7
<i>immune response</i>	0.0054	0.0319	1.65	5
<i>cell adhesion</i>	0.0221	0.0662	2.04	5
<i>proteolysis and peptidolysis</i>	0.0382	0.0895	1.61	4
<i>protein transport</i>	0.2744	0.5321	1.76	2
<i>phosphorylation</i>	0.3112	0.5442	2.82	3
<i>G-protein coupled receptor protein signaling pathway</i>	0.3850	0.5178	1.71	2
<i>biosynthesis</i>	0.4095	0.4788	3.51	4
<i>regulation of transcription, DNA-dependent</i>	0.5354	0.9297	5.94	3

Table 4. Enrichment Analysis, Top 10 Biological Processes (>300 Genes) by MVN P-value of, Leukemia Treatment Response Study ($k=111$)

Biological Process	MVN	Fisher	E(n_i)	n_i
<i>cell cycle</i>	0.0060	0.0271	6.74	13
<i>immune response</i>	0.0146	0.0561	4.66	9
<i>regulation of transcription</i>	0.0415	0.2642	17.59	21
<i>apoptosis</i>	0.1194	0.1677	5.19	8
<i>protein amino acid phosphorylation</i>	0.1673	0.3701	5.78	7
<i>transcription from RNA polymerase II promoter</i>	0.4610	0.3629	5.73	7
<i>transcription</i>	0.5387	0.3406	18.57	21
<i>transport</i>	0.5399	0.6523	19.39	18
<i>cell adhesion</i>	0.6088	0.6704	5.67	5
<i>metabolism</i>	0.6934	0.7294	13.93	12

Table 5. Enrichment Analysis, Top 10 Biological Processes (>300 Genes) by MVN P-value of, Multiple Myeloma / Normal Comparison (k=93)

Biological Process	MVN	Fisher	E(n_i)	n_i
<i>cell cycle</i>	0.1146	0.1606	5.91	9
<i>regulation of transcription, DNA-dependent</i>	0.1749	0.1522	11.59	16
<i>regulation of transcription</i>	0.2727	0.1620	14.28	19
<i>transcription</i>	0.2872	0.1697	15.26	20
<i>metabolism</i>	0.3163	0.4130	12.71	14
<i>transcription, DNA -dependent</i>	0.3252	0.1575	11.67	16
<i>cell proliferation</i>	0.4822	0.5106	6.68	7
<i>phosphorylation</i>	0.5361	0.5742	7.12	7
<i>biosynthesis</i>	0.7926	0.8047	8.00	6
<i>development</i>	0.8341	0.8016	9.16	7