Identifying Differentially Expressed Gene Categories

3/24/2011

Using Information about Genes to Interpret the Results of Microarray Experiments

- Based on a large body of past research, some information is known about many of the genes represented on a microarray.
- The information might include tissues in which a gene is known to be expressed, the biological process in which a gene's protein is known to act, or other general or quite specific details about the function of the protein produced by a gene.
- By examining this information in concert with the results of a microarray experiment, biologists can often gain a greater understanding of their microarray experiments.

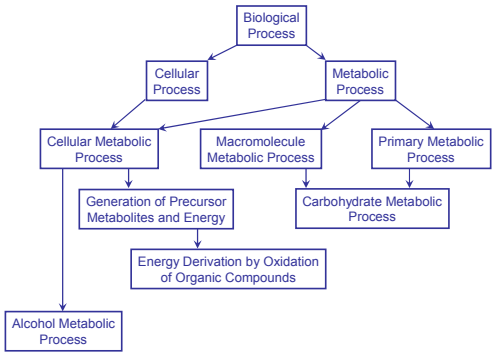
Gene Ontology (GO) Terms

- GO terms provide one example of information that is available about genes.
- The GO project provides three ontologies (structured controlled vocabularies) that describe a gene's
  1. Biological Processes,
  2. Cellular Components, and
  3. Molecular Functions.

Gene Ontology (GO) Terms

- Each gene may be associated with 0 or more GO terms in a given ontology.
- The GO terms in each ontology have varying levels of specificity.
- The GO terms in each ontology can be organized in a directed acyclic graph (DAG) where each node represents a term and arrows point from general terms to more specific terms.

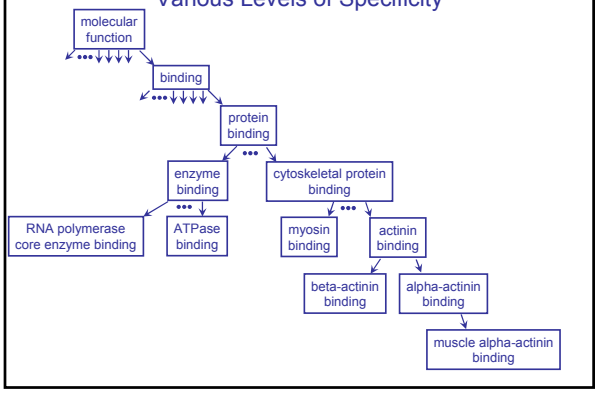
Part of the GO Biological Processes DAG



Constructing Gene Categories from GO Terms

- The set of genes associated with any particular GO term could be considered as a category or gene set of interest for subsequent testing.
- For example, we might ask if genes that are associated with the Molecular Function term *muscle alpha-actinin binding* are affected by a treatment of interest.
- We could simultaneously query many groups, general and specific, to better understand the impact of treatment on expression.

Simultaneous Testing of Multiple Categories with Various Levels of Specificity



Some Formal Methods for Testing Gene Categories with Microarray Data

- Fisher's exact test on lists of gene declared to be differentially expressed (DDE)
- Gene Set Enrichment Analysis (GSEA)
- Significance Analysis of Function and Expression (SAFE)
- Pathway Level Analysis of Gene Expression (PLAGE)
- Domain Enhanced Analysis (DEA)
- Efron and Tibshirani's Gene Set Analysis (GSA)

8

The Dataset

Gene ID	Treatment 1					Treatment 2					p-value
1	4835.8	4578.2	4856.3	4483.7	4275.3	4170.7	3836.9	3901.8	4218.4	4094.0	p_1
2	153.9	161.0	139.7	173.0	160.1	180.1	265.1	201.2	130.8	130.7	p_2
3	3546.5	3622.7	3364.3	3433.6	2757.2	3346.9	2723.8	2892.0	3021.3	2452.7	p_3
4	711.3	717.3	776.6	787.5	750.3	910.2	813.3	687.9	811.1	695.6	p_4
5	126.3	178.2	114.5	158.7	157.3	231.7	147.0	102.8	157.6	146.8	p_5
6	4161.8	4622.9	3795.7	4501.2	4265.8	3931.3	3327.6	3726.7	4003.0	3906.8	p_6
7	419.3	555.3	509.6	515.5	488.9	426.6	425.8	500.8	347.8	580.3	p_7
8	2420.7	2616.1	2768.7	2663.7	2264.6	2379.7	2196.2	2491.3	2710.0	2759.1	p_8
9	321.5	540.6	471.9	348.2	356.6	382.5	375.9	481.5	260.6	515.7	p_9
10	1061.4	949.4	1236.8	1034.7	976.8	1059.8	903.6	1060.3	960.1	1134.5	p_{10}
11	1293.3	1147.7	1173.8	1173.9	1274.2	1062.8	1172.1	1113.0	1432.1	1012.4	p_{11}
12	336.1	413.5	425.2	462.8	412.2	391.7	388.1	363.7	310.8	404.6	p_{12}
13	325.2	278.9	242.8	255.6	283.5	161.1	181.0	222.0	279.3	232.9	p_{13}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
20000	249.6	283.6	271.0	246.9	252.7	214.2	217.9	266.6	193.7	413.2	p_{20000}

9

Are genes of functional category X overrepresented among the genes declared to be differentially expressed?

		Gene of Functional Category X? yes	no	
Declared to be Differentially Expressed?	yes	50	250	300
	no	50	19650	19700
		100	19900	20000

Highly significant overrepresentation according to a chi-square test or Fisher's exact test.

10

Problems with Chi-Square or Fisher's Exact Test for Detecting Overrepresentation

- The outcome of the overrepresentation test depends on the significance threshold used to declare genes differentially expressed.
- Functional categories in which many genes exhibit small changes may go undetected.
- Genes are not independent, so a key assumption of the chi-square and Fisher's exact tests is violated.
- Information in the multivariate distribution of genes in a category is not utilized.

11

Gene Set Enrichment Analysis (GSEA)

- Subramanian, et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*. **102**, 15545-15550.
- Compute a statistic for each gene that measures the degree of differential expression between treatments.
- Create a list L of all genes ordered according to these statistics.

12

GSEA (continued)

- Compute an Enrichment Score (ES) for each category (set) to determine whether the genes in a category are randomly distributed throughout L or primarily found at the top or bottom of L.
- Determine significance of ES by comparing the observed value to the values obtained when the scores are recomputed for each of all possible (or many randomly selected) permutations of the treatment labels on the microarrays.
- Normalize enrichment scores and use a permutation approach to identify significant categories while controlling FDR.

13

GSEA Details

$r(g_i) = r_i$ test statistic for gene g_i

$L = g_1, \ldots, g_N$ list of all genes ordered according to test statistics

$r_1 \leq \cdots \leq r_N$

$N_H = \sum_{g_j \in S} 1$ number of genes in gene set S

$N_R = \sum_{g_j \in S} |r_j|^p$ p is a user-specified value p=1 is recommended

14

GSEA Details (continued)

$$P_{\text{hit}}(S,i) = \sum_{g_j \in S, j \leq i} \frac{|r_j|^p}{N_R} \qquad P_{\text{miss}}(S,i) = \sum_{g_j \notin S, j \leq i} \frac{1}{N - N_H}$$

$$M(S) = \max\{P_{\text{hit}}(S,i) - P_{\text{miss}}(S,i) : i = 1, \ldots, N\}$$

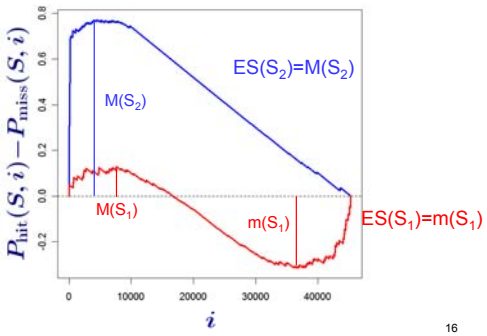
$$m(S) = \min\{P_{\text{hit}}(S,i) - P_{\text{miss}}(S,i) : i = 1, \ldots, N\}$$

$$ES(S) = \begin{cases} M(S) & \text{if } M(S) \geq -m(S) \\ m(S) & \text{if } M(S) < -m(S) \end{cases}$$

Enrichment score for gene set S

15

Enrichment Scores for Two Sets of 100 Genes Each



16

Permutation *p*-value for a Single Category S

- Randomly assign the treatment labels to the microarrays, preserving the number of arrays per treatment used in the experiment.
- Using the permuted data, compute ES(S,1).
- Repeat a total of 1000 times to get ES(S,1),...,ES(S,1000).
- If ES(S)>0 (<0), the approximate permutation *p*-value is given by the proportion of the positive (negative) ES(S,p) values (p=1,...,1000) that were greater (less) than or equal to ES(S).

17

Simultaneous Testing for Multiple Categories

- Carry out the permutation procedure described on the previous slide simultaneously for all gene sets of interest (S_1,...,S_K).
- For all k=1,...,K; find $A^+(k)$=the average of the positive $ES(S_k,p)$ (p=1,...,1000). Also, find $A^-(k)$=the absolute value of the average of the negative $ES(S_k,p)$ (p=1,...,1000).
- If $ES(S_k)>0$, define the normalized enrichment score by $NES(S_k)=ES(S_k)/A^+(k)$. If $ES(S_k)<0$, define $NES(S_k)=ES(S_k)/A^-(k)$.
- Likewise, divide each $ES(S_k,p)$ by either $A^+(k)$ or $A^-(k)$, depending on its sign, to obtain $NES(S_k,p)$ (p=1,...,1000).

18

Simultaneous Testing for Multiple Categories (ctd.)

- FDR is estimated separately for positive and negative enrichment scores.
- For a positive normalized enrichment score NES*, FDR is estimated by the proportion of positive NES(S_k,p) (k=1,...,K; p=1,...,1000) that were greater than or equal to NES* divided by the proportion of positive NES(S_k) that were greater than or equal to NES*.
- The analogous calculation is carried out for negative normalized enrichment scores.

19

Potential Weakness of GSEA and Related Methods

- The enrichment score for a category compares the degree of differential expression among genes in a category relative to that of all other genes.
- The permutation testing procedure actually tests the null hypothesis of no differential expression across treatments.
- The method examines only the marginal distribution of each gene even though the question of interest involves a group of genes.

20

A Multivariate Approach

- Nettleton, D., Recknor, J., Reecy, J.M. (2008). Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis. *Bioinformatics*. **24** 192-201.

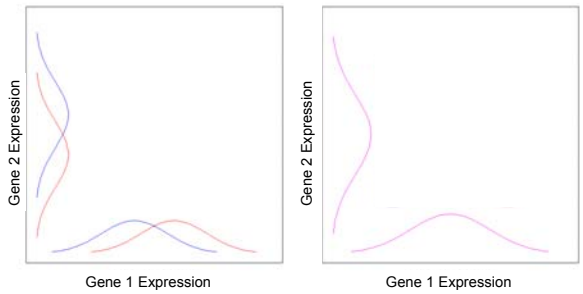
21

Rather then testing for enrichment, consider a different problem.

- Suppose X_c is a vector of expression measurements for genes in a category c of interest.
- Let F_{ci} denote the multivariate distribution of X_c under treatment i (i=1,...,T).
- We wish to test $H_{c0}: F_{c1} = ... = F_{cT}$ for each category c=1,...,C.
- This is the multivariate analog of the univariate testing problems that we have considered this semester.

22

Advantage of a Multivariate Approach



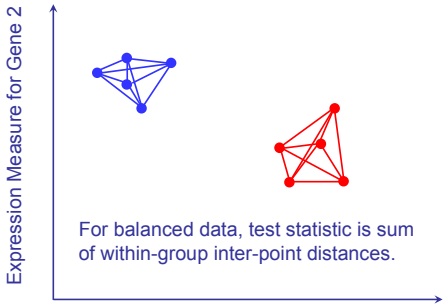
23

Multiresponse Permutation Procedure (MRPP)

- Mielke and Berry, (2001). *Permutation Methods: A Distance Function Approach*. Springer, N.Y.
- Nonparametric test for a difference among multivariate distributions
- Test statistic based on within-group inter-point distances
- P-value obtained by data permutation

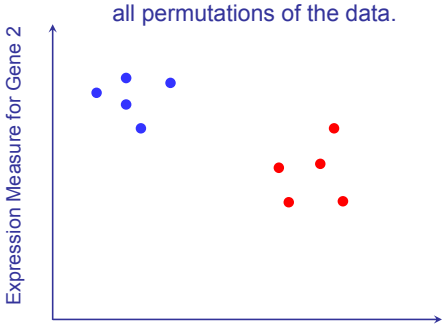
24

The MRPP Test: An Illustrative Example



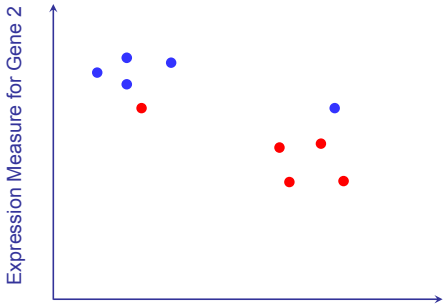
25

The test statistic is computed for all permutations of the data.



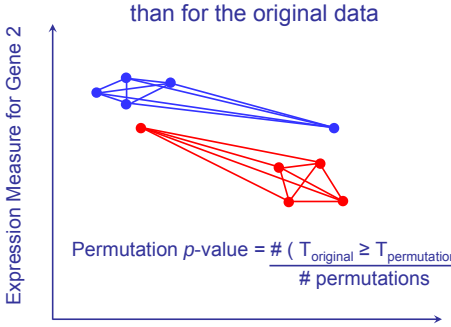
26

An Example Permutation



27

Test statistic will be larger for permutations than for the original data



Permutation p-value = $\frac{\#\,(\,T_{original} \geq T_{permutation}\,)}{\#\text{ permutations}} = \frac{2}{252}$

28

Application to the Myostatin Experiment

347 GO Molecular Function categories were tested.

Genes per Category:

min	Q1	med	Q3	max
40	60	115	234	18560

A p-value was obtained for each category using the MRPP test.

A resampling based method to control FDR at ~0.025 yielded 77 significant gene categories (see next slide).

29

Benjamini and Yekutieli (1999) Resampling-Based FDR

M=number of permutations
m=1 denotes original (unpermuted) data

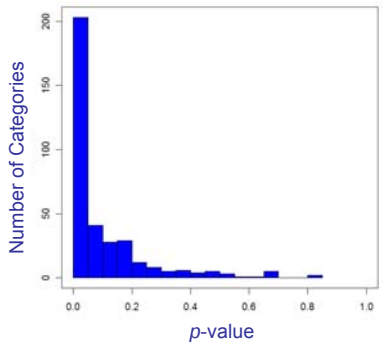
$$\widehat{\mathrm{FDR}}(p) = \min_{p':p' \geq p} \left(\frac{1}{M-1} \sum_{m=2}^{M} \frac{R_m(p')}{R_m(p') + S(p')} \right)$$

$$R_m(p') = \sum_{c=1}^{C} \mathbb{1}(p_{cm} \leq p')$$

$$S(p') = R_1(p') - \frac{1}{M-1} \sum_{m=2}^{M} R_m(p')$$

30

Histogram of *p*-values from the 347 MRPP tests



31

Potential Weakness of the Multivariate Method

- There is a known subset structure among the categories that imposes logical constraints about the state of each null hypothesis.
- The multiple testing method does not take these logical constraints into account when deciding which categories to declare significant.

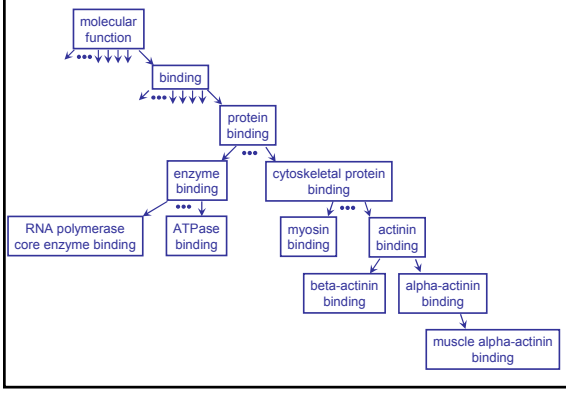
32

A New Method that Accounts for Logical Relationships among Null Hypotheses

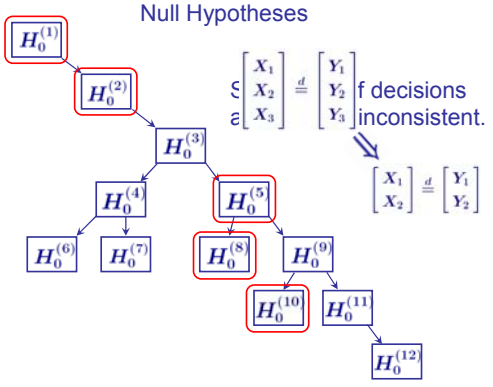
- Liang, K. and Nettleton, D. (2010). A hidden Markov model approach to testing multiple hypotheses on a tree-transformed gene ontology graph. *Journal of the American Statistical Association*. **105** 1444–1454.

33

Accounting for the DAG Structure



Accounting for Logical Relationships among the Null Hypotheses



Model for the *p*-values that accounts for the DAG structure

Notation:

$\mathcal{G}_c = \{g : \text{gene } g \text{ in category } c\}$ (genes in category c)

$\mathcal{P}_c = \{c' : \mathcal{G}_c \subset \mathcal{G}_{c'} \text{ and } \nexists\, k \text{ such that } \mathcal{G}_c \subset \mathcal{G}_k \subset \mathcal{G}_{c'}\}$

(parental categories of category c)

$$S_c = \begin{cases} 0 & \text{if } H_0^{(c)} \text{ true} \\ 1 & \text{if } H_0^{(c)} \text{ false} \end{cases}$$ (state of category c)

36

Markov Model for Unobserved States

Probability that initial category is non-null:

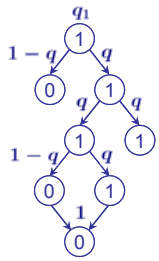
$$P(S_1 = 1) = q_1$$

Probability that category c is non-null given that all its parents are non-null:

$$P(S_c = 1 | S_{c'} = 1 \; \forall \; c' \in \mathcal{P}_c) = q$$

Probability that category c is null given that at least one parent is null:

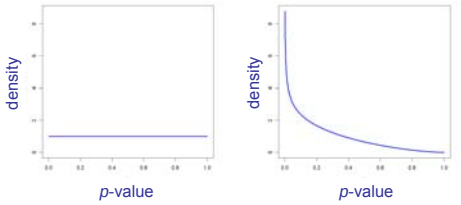
$$P(S_c = 0 | S_{c'} = 0 \text{ for some } \; c' \in \mathcal{P}_c) = 1$$



37

Distribution of *p*-value Given the State

$(p_c | S_c = 0) \sim \mathrm{u}(0,1)$ $(p_c | S_c = 1) \sim \mathrm{b}(\alpha, \beta)$



38

Prior Distributions

$q_1, q \sim \mathrm{b}(0.5, 0.5)$ ← Jeffreys prior on probability of a category being non-null, given that all ancestors are non-null

$\alpha \sim \mathrm{u}(0,1)$

$\beta \sim u(1, 1000)$

← Priors on the parameters of the beta distribution that ensure a decreasing density for non-null *p*-values.

39

Inference

- Gibbs sampling is used to obtain draws from the posterior distribution of states given the *p*-values.
- Because of our Markov model for the states, all draws of $S_1, \ldots, S_C$ from the posterior distribution honor the logical relationships among the null hypotheses.
- Categories for which $P(S_c = 1 | p_1, \ldots, p_C)$ is large are declared to be non-null.
- The resulting set of decisions is logically consistent.

40

Application to the Myostatin Experiment

347 GO Molecular Function categories were tested.

Genes per Category:

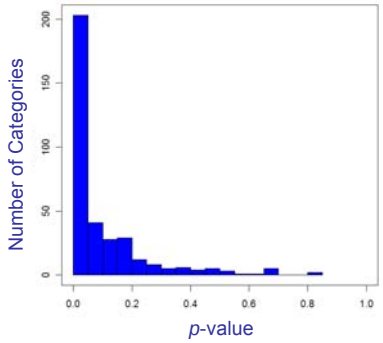
min	Q1	med	Q3	max
40	60	115	234	18560

A *p*-value was obtained for each category using the MRPP test.

A resampling based method to control FDR at ~0.025 yielded 77 significant gene categories (Nettleton et al., 2008).

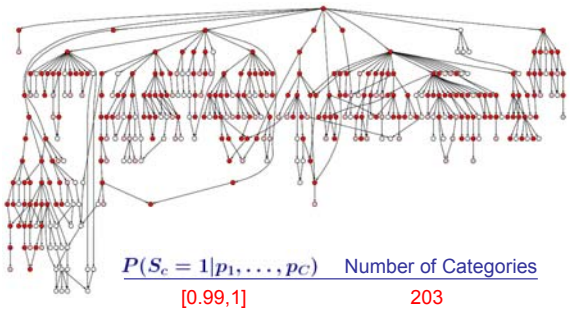
41

Histogram of *p*-values from the 347 MRPP tests



42

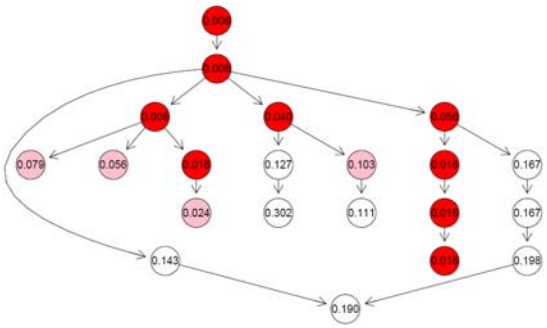
Results of our Hidden Markov model approach



$P(S_c = 1 \mid p_1, \ldots, p_C)$	Number of Categories
[0.99,1]	203
[0.95,0.99)	59

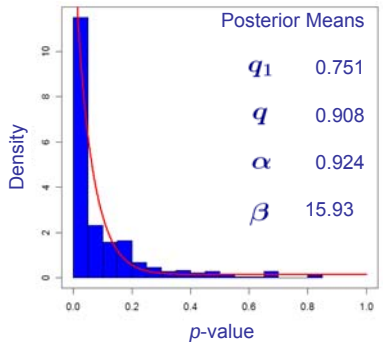
43

Zooming in on part of the DAG



44

Histogram of *p*-values from the 347 MRPP tests



45

ALL Data

- B- and T-cell Acute Lymphocytic Leukemia (ALL) data set described in part by Chiaretti et al. (2004).
- Publicly available in the Bioconductor ALL package at www.bioconductor.org.
- 12,625-dimensional expression profiles from the Affymetrix HGU95aV2 GeneChip for each of 128 ALL patients.
- 95 B-cell samples and 33 T-cell samples.

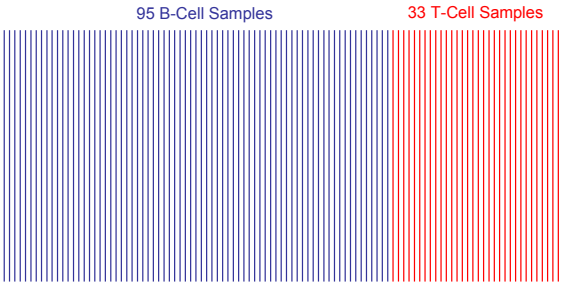
46

ALL Data

- Based on previously published results, we identified the 2 GO Biological Process categories with strong evidence for differential expression between B-cell and T-cell cancer types.
- The union of those categories involved 53 genes of the 12,625 genes.

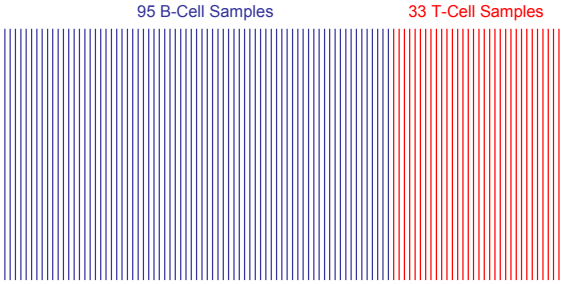
47

Data-Based Simulation Study



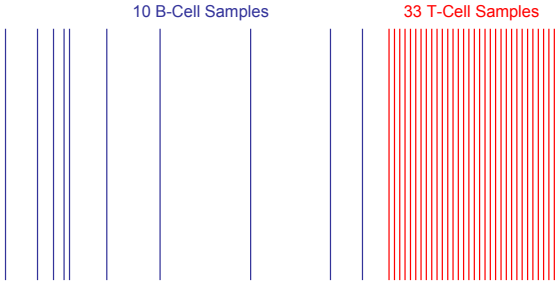
48

Randomly Pick 2n B-Cell Samples



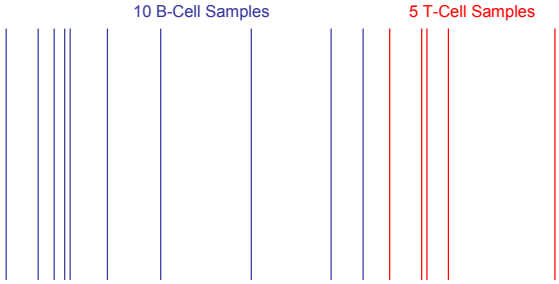
49

Randomly Pick 2n B-Cell Samples



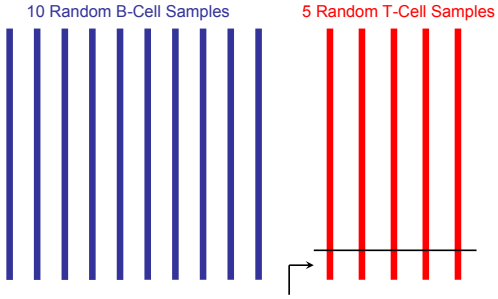
50

Randomly Pick n T-Cell Samples



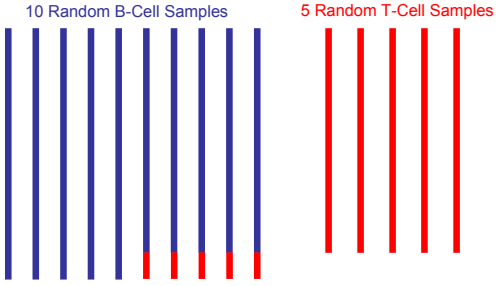
51

Create Differential Expression for Genes in the Top 2 Categories



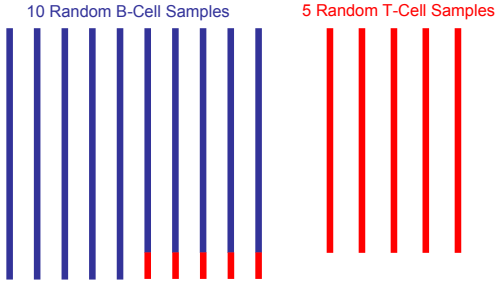
52

Create Differential Expression for Genes in the Top 2 Categories



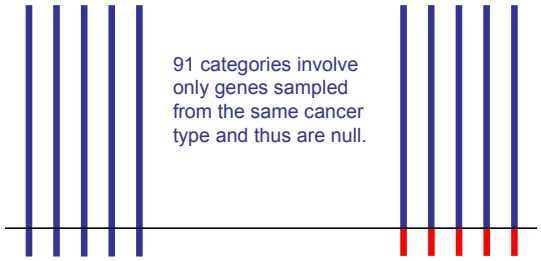
53

Create Differential Expression for Genes in the Top 2 Categories



54

Test 324 Categories for Expression Differences between These Two Groups



The remaining 233 categories include one or more of the genes simulated from different cancer types and thus are non-null.

55

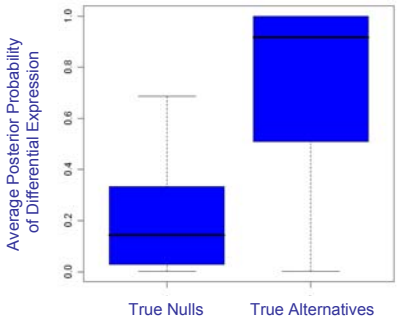
Results from 10 Simulation Runs

$$P(S_c = 1 | p_1, \ldots, p_C) \geq \theta$$

	$\theta = 0.99$		$\theta = 0.95$		$\theta = 0.90$	
Run	Total	Null	Total	Null	Total	Null
1	104	0	112	0	119	0
2	136	0	142	0	143	0
3	145	1	167	3	189	8
4	116	0	136	0	150	0
5	120	0	138	1	146	2
6	138	0	151	0	160	0
7	131	0	144	0	149	0
8	82	0	98	0	108	0
9	109	0	120	0	123	0
10	121	0	137	0	145	0

56

Simulation Results Averaged over 10 Simulations



57