

Significance Analysis of Microarrays (SAM)

3/12/2009

Copyright © 2009 Dan Nettleton

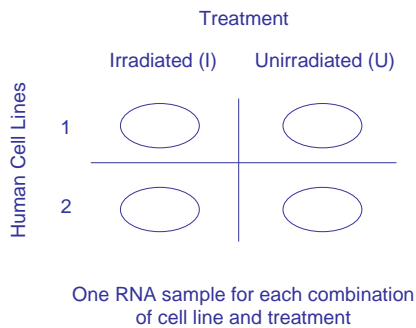
1

Significance Analysis of Microarrays (SAM)

- Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98(9)**, 5116-5121.
- See <http://www-stat.stanford.edu/~tibs/SAM/> for more recent developments.

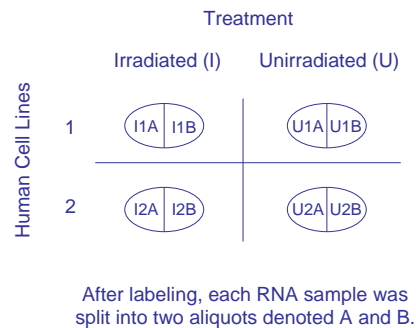
2

Motivating Experiment



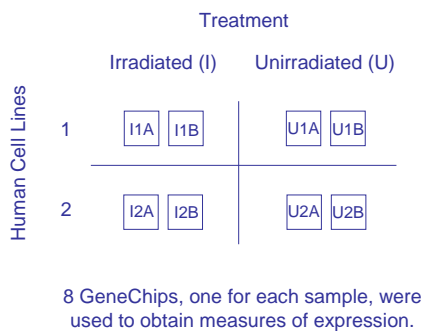
3

Motivating Experiment



4

Motivating Experiment



5

The Measure of Expression

- The Tusher et al. paper appeared before much work had been done on computing expression measures from PM and MM intensities.
- It appears that they used something similar to an average of PM-MM intensities over the probe pairs in a probe set as their expression measure.
- They conducted a regression-based normalization in which each GeneChip was normalized to the average over all GeneChips.
- Furthermore they worked with normalized data on the original rather than the log scale.

6

Test Statistic for the j^{th} Gene

Average of 4 normalized
measures from
irradiated samples

→

Average of 4 normalized
measures from
unirradiated samples

→

$$d(i) = \frac{\bar{X}_I(i) - \bar{X}_U(i)}{S(i) + s_0}$$

←

The usual standard
error in the denominator
of a two-sample t-stat

←

A constant common to all
genes that is added to make
variation in $d(i)$ similar across
genes of all intensity levels

7

Selecting the constant s_0

- At low expression levels, variance in $d(i)$ can be high because of small values of $s(i)$.
- To stabilize the variance of $d(i)$ across genes, a small positive constant s_0 was used in the denominator of the test statistic.
- "The coefficient of variation of $d(i)$ was computed as a function of $s(i)$ in moving windows across the data. The value for s_0 was chosen to minimize the coefficient of variation."
- s_0 was chosen to be 3.3 for the ionizing radiation data.

8

More Detail on Selecting s_0

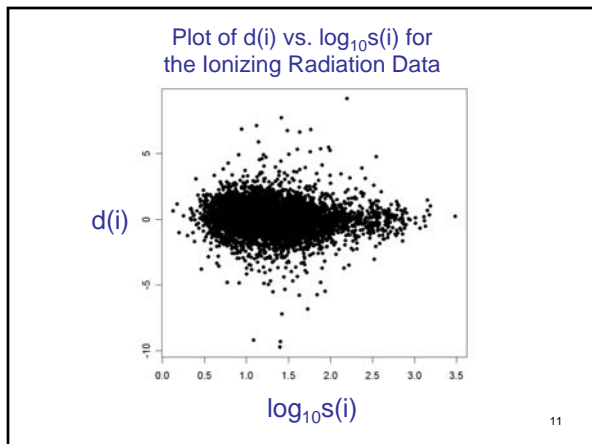
- The $d(i)$ are separated into approximately 100 groups. The 1% of the $d(i)$ values with the smallest $s(i)$ values are placed in the first group, the 1% of the $d(i)$ values with the next smallest $s(i)$ are placed in the second group, etc.
- The median absolute deviation (MAD) of the $d(i)$ values is computed separately for each group.
- The coefficient of variation (CV) of these 100 MAD values is computed.

9

More Detail on Selecting s_0 (continued)

- This process is repeated for values of s_0 equal to the minimum of $s(i)$ over i , the 5th percentile of the $s(i)$ values, the 10th percentile of the $s(i)$ values, ..., the 95th percentile of the $s(i)$ values, and the maximum of the $s(i)$ values.
- The value of s_0 that minimizes the CV of the 100 MAD values over candidate s_0 described above is selected as the constant s_0 .

10

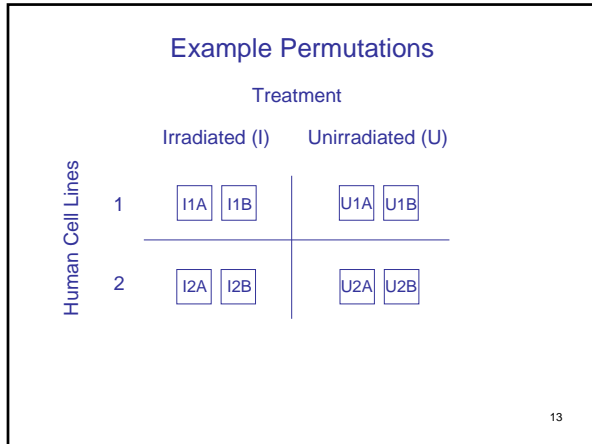


A Permutation Procedure for Assessing Significance

1. The irradiated and unirradiated GeneChips were shuffled within each cell line.
2. The $d(i)$ statistic was computed for each gene and ordered across genes from smallest to largest to obtain $d_p(1) < d_p(2) < \dots < d_p(g)$ where g denotes the number of genes.
3. Steps 1 and 2 were repeated for all possible data permutations described in step 1 to obtain $d_p(1) < d_p(2) < \dots < d_p(g)$ for $p=1, \dots, 36$.

$\begin{pmatrix} 4 & 4 \\ 2 & 2 \end{pmatrix}$

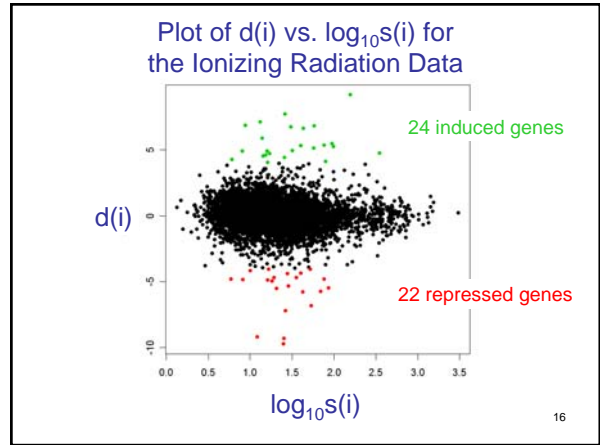
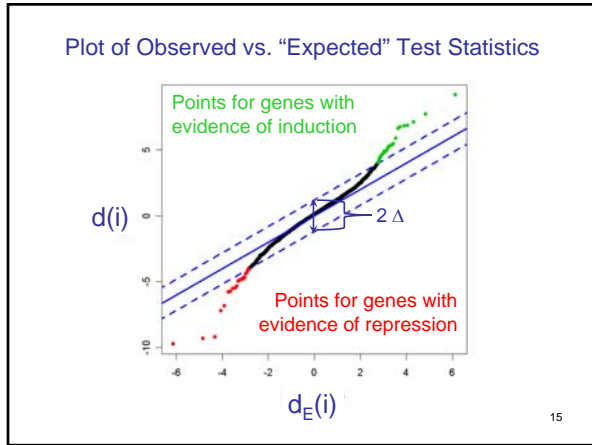
12



A Permutation Procedure for Assessing Significance (continued)

4. For each i , $d_1(i), \dots, d_{36}(i)$ were averaged to obtain $d_E(i)$, the "expected relative difference."
5. The original $d(i)$ statistics were also sorted so that $d(1) < d(2) < \dots < d(g)$.
6. Genes for which $|d(i) - d_E(i)| > \Delta$ were declared significant, where Δ is a user specified cutoff for significance.

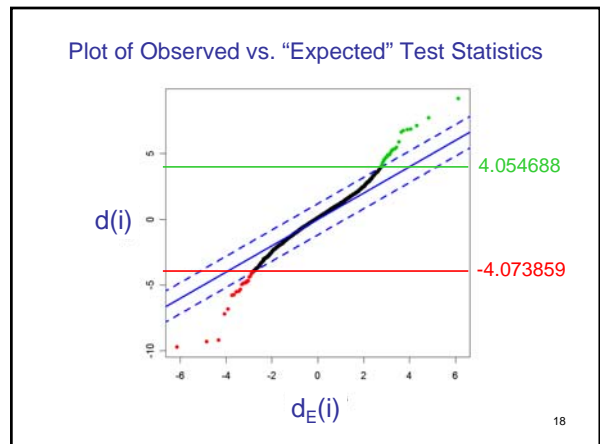
14

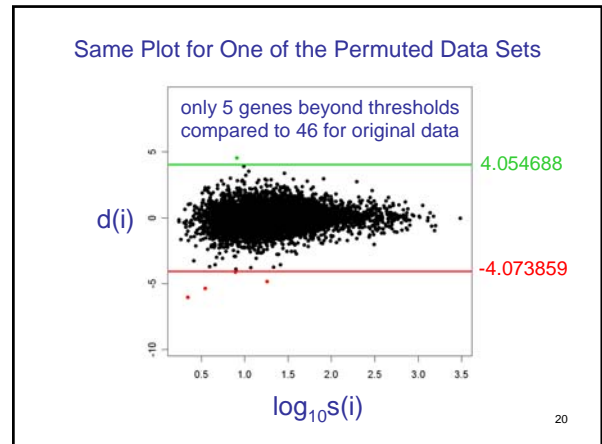
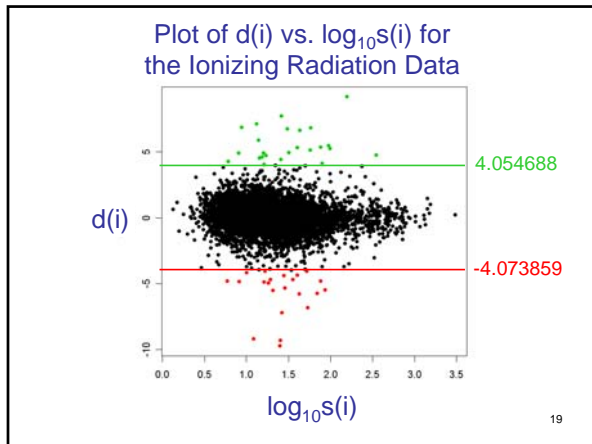


Estimating FDR for a Selected Δ

1. Find the smallest $d(i)$ among those $d(i)$ for which $d(i) - d_E(i) > \Delta$ and call it d_{up} .
2. Find the largest $d(i)$ among those $d(i)$ for which $d(i) - d_E(i) < -\Delta$ and call it d_{down} .
3. For each permuted data set, find the number of genes with $d(i) \geq d_{up}$ or $d(i) \leq d_{down}$ and denote these counts by n_1, \dots, n_{36} .
4. FDR is estimated by \bar{n} / n where \bar{n} is the average of n_1, \dots, n_{36} and n is the number of genes identified as significant in the original data.

17





Counts of Genes beyond the Threshold for Each Permutation

Perm	Count	Perm	Count	Perm	Count
1	45	13	4	25	4
2	5	14	1	26	2
3	2	15	3	27	1
4	3	16	9	28	1
5	4	17	12	29	5
6	11	18	31	30	9
7	8	19	31	31	11
8	5	20	12	32	4
9	1	21	9	33	3
10	1	22	3	34	2
11	3	23	1	35	5
12	4	24	4	36	46

21

Mean Count = 8.472 FDR Estimate = $8.472/46 = 18.4\%$

Perm	Count	Perm	Count	Perm	Count
1	45	13	4	25	4
2	5	14	1	26	2
3	2	15	3	27	1
4	3	16	9	28	1
5	4	17	12	29	5
6	11	18	31	30	9
7	8	19	31	31	11
8	5	20	12	32	4
9	1	21	9	33	3
10	1	22	3	34	2
11	3	23	1	35	5
12	4	24	4	36	46

22

- ### Concluding Remarks
- Ideally the method would be applied in situations with more true biological replications than were available in the ionizing radiation data.
 - Can be extended in many ways.
 - Other choices for test statistics are possible.
 - Should permute in such a way that all permutations are equally likely under the null hypothesis of interest.
 - The properties of SAM regarding error rate control are not well known.
- 23