

Web Supplement of

## Improved prediction of treatment response using microarrays and existing biological knowledge

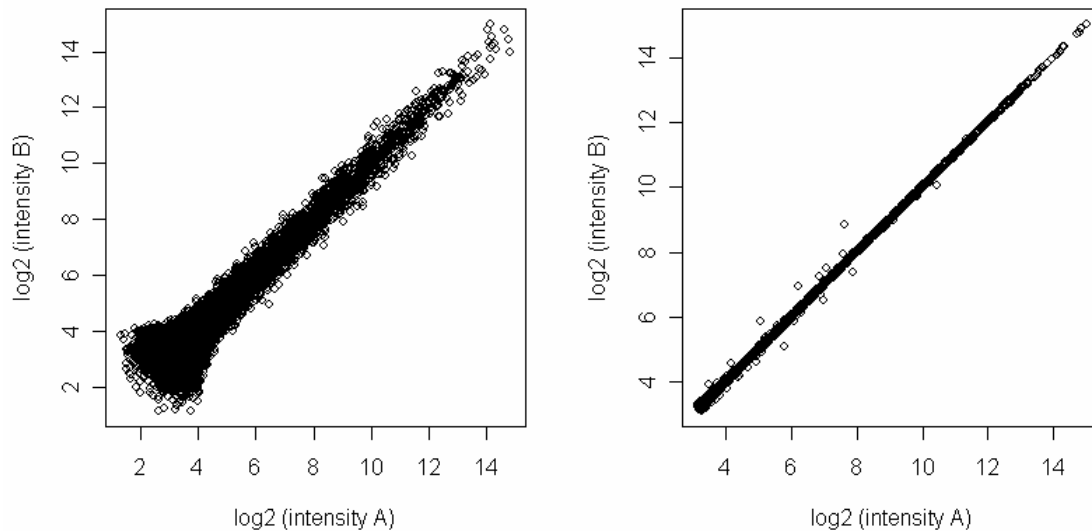
Simon M. Lin<sup>1</sup>, Jothi Devakumar<sup>2</sup>, Warren A. Kibbe<sup>1,\*</sup>,<sup>1</sup> Robert H. Lurie Cancer Center, Northwestern University, Chicago, IL 60611, and <sup>2</sup>Jubilant Biosys Ltd, Devasandra, 80 ft road, RMV Extn II stage, Bangalore, India, 560094

\* corresponding author

<http://basic.northwestern.edu/publications/topdown/>

### Details of the Microarray Data Simulation

The microarray data was simulated with the following script written in R. The simulated data mimics the characteristics of Affymetrix array which is noisier at the low intensity region (Figure 1a). Only 200 genes out of 16000 genes are differentially expressed, and most of the differentially expressed genes have a smaller fold change (see the average of 200 arrays in each group in Figure 1b).



**Supplemental Figure 1. Scatter plot of simulated microarray data.** Two groups each containing 200 arrays were simulated. a) A representative pair of arrays. b) An average of all arrays in each group.

R script for this simulation:

```

# Truth: 200 genes changed
# Total number of genes: up to 16000
#
# 200 cases in group A, 200 cases in group B

n.genes<- 16000
n.subjectToChange<- 200
n.patientsInEachGroup<- 200

# Distribution of true hybridization signal
# most hybridization is of low intensity
hyb<- rbeta(n.genes, shapel=0.5, shape2=3)*16
# histogram plot
hist (hyb, xlab= "log2 (Intensity)")

# Most of the fold change is small
# It can be either up (+) or down (-)
fold<- rbeta(n.subjectToChange, shapel=0.5, shape2=30)*10
fold<- fold * sample (c(+1, -1), length (fold), replace=T)
# histogram plot
hist (fold, xlab= "log2 (Fold Change)")

classLabel<- rep (as.factor(c(-1, 1)), n.patientsInEachGroup)

# True expression level, without noise
a<- hyb; b<- hyb
b[1:200]<- b [1:200] + fold

# Observed expression level, with noise, 400 arrays
mIntensity<- matrix (NA, nrow= n.genes, ncol=
n.patientsInEachGroup *2)
k<- 1
for (i in 1:n.patientsInEachGroup) {
  # random noise added
  x<- a + rnorm (n.genes, mean=0, sd=0.3)
  y<- b + rnorm (n.genes, mean=0, sd=0.3)

  x<- (2^x) + rbeta(n.genes, shapel=3, shape2=3)*16+1
  y<- (2^y) + rbeta(n.genes, shapel=3, shape2=3)*16+1

  mIntensity[,k]<- log2(x); mIntensity[,k+1]<- log2(y)
  k<- k+2
}

# Supplemental Figure 1a: a pair of arrays
par (mfrow= c(1,2))
plot (mIntensity[,1],mIntensity[,2],
      xlab= "log2 (intensity A)", ylab="log2 (intensity B)")

# Supplemental Figure 1b: average of arrays
avg.x<- apply (mIntensity[, classLabel =="-1"], 1, mean)
avg.y<- apply (mIntensity[, classLabel =="1"], 1, mean)
plot (avg.x, avg.y,
      xlab= "log2 (intensity A)", ylab="log2 (intensity B)")

```

## Detecting Differentially Expressed Genes in the Simulated Data Set

To further characterize the simulated data set, we conducted Significance Analysis of Microarrays (SAM) to find differentially expressed genes. SAM is an empirical Bayesian method modifying the t-test [1]. The variance of each gene is estimated by both the gene itself and contributions from other genes on the array.

With a false discovery rate of 5%, SAM analysis reported a total of 44 genes differentially expressed genes with 6 falsely identified ones (Supplemental Table 1). It suggests that truly differentially expressed genes (a total of 200) can be reasonably detected (a total of 44) with a small type I error (6 false claims), although the type II error (156 missing detections) is relatively high.

Classification and biomarker discovery (differential detection) are two major types of microarray data analysis [2]. This simulated data set with 200 arrays poses a significant challenge for classification, but not as difficult for detecting differentially expressed genes.

**Supplemental Table 1. Result of SAM analysis.**

Strength of Evidence	Gene Row ID	Estimated FDR %	Truth
1	184	0	Correct
2	70	0	Correct
3	31	0	Correct
4	60	0	Correct
5	122	0	Correct
6	25	0	Correct
7	84	0	Correct
8	190	0	Correct
9	58	0	Correct
10	148	0	Correct
11	21	0	Correct
12	43	0	Correct
13	34	0	Correct
14	127	0	Correct
15	1	0	Correct
16	82	0	Correct
17	96	0	Correct
18	5	0	Correct
19	13	0	Correct
20	125	0	Correct
21	27	0	Correct
22	136	0	Correct
23	74	0	Correct
24	116	0	Correct
25	123	0	Correct
26	39	0	Correct
27	28	0	Correct
28	73	0	Correct
29	178	0	Correct
30	76	0	Correct
31	64	0	Correct
32	164	0	Correct
33	118	0.002	Correct
34	20	0.003	Correct
35	3543	0.01	Wrong
36	11	0.012	Correct
37	10563	0.018	Wrong
38	11821	0.018	Wrong
39	15367	0.02	Wrong

40	130	0.022	Correct
41	3426	0.024	Wrong
42	9040	0.043	Wrong
43	126	0.047	Correct
44	86	0.06	Correct

---

## References

1. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*. Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
2. Simon, R.M., *Design and analysis of DNA microarray investigations*. Statistics for biology and health. 2003, New York: Springer. x, 199.