REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

Public Reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188,) Washington, DC 20503.

1. AGENCY USE ONLY (Leave Blank)

2. REPORT DATE 02AUG13

3. REPORT TYPE AND DATES COVERED
Technical Report

4. TITLE AND SUBTITLE
Effectiveness of ranked t scores for identification of signaling genes: a simulation study

5. FUNDING NUMBERS
DAAD19-02-C-0045

6. AUTHOR(S)
Harry Hurd

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Harry L. Hurd Associates, Inc.
309 Moss Run
Raleigh, NC 27614
Tel. 919-846-9227
e-mail : harry_hurd@bellsouth.net

8. PERFORMING ORGANIZATION REPORT NUMBER TR02-1

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
43812.2-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Since some of the methods for identifying signaling genes in microarray experiments are hiearchical networks of often simple methods, it seems natural to use simulation to understand how well these methods perform under idealized circumstances. In this study we assume that expression levels for signaling genes are distributed $N(\mu,1),\:\mu>0$ and are distributed $N(0,1)$ for the non-signaling ones. Signaling genes are identified simply by taking the top N ranked t scores. Under this set-up we evaluate the probability that the top 10 scores will correctly identify at least M good genes as a function of the gene signaling level μ, the number of samples from the control and treatment populations and the number of genes that carry the signal.

In spite of these simplicity of the model we think some insight is gained about the relationships between the sample size and the signaling level at which some specified performance is obtained.The conclusion, under the assumption of equal signal strengths, is that there is considerable payoff for ``genefinding" in the first few doublings of sample size, say from 2 to 4 and perhaps to 8. The reduction of signal level required to give specified ``genefinding" performance continues and appears to agree with the anticipated asymptotic reduction by sqrt(2) for each doubling.

Effectiveness of ranked t scores for identification of signaling genes: a simulation study

August 13, 2002

Submitted to
U. S. ARMY RESEARCH OFFICE
Contract DAAD19-02-C-0045.

by

Harry L. Hurd Associates, Inc.
309 Moss Run
Raleigh, NC 27614
Tel. 919-846-9227
harry_hurd@bellsouth.net

APPROVED FOR PUBLIC RELEASE
DISTRIBUTION UNLIMITED

The views, opinions and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

Effectiveness of ranked t scores for identification of signaling genes: a simulation study

Harry L. Hurd

August 13, 2002

Abstract

Since some of the methods for identifying signaling genes in microarray experiments are hiearchical networks of often simple methods, it seems natural to use simulation to understand how well these methods perform under idealized circumstances. In this study we assume that expression levels for signaling genes are distributed $N(\mu, 1)$, $\mu > 0$ and are distributed $N(0, 1)$ for the non-signaling ones. Signaling genes are identified simply by taking the top N ranked t scores. Under this set-up we evaluate the probability that the top 10 scores will correctly identify at least M good genes as a function of the gene signaling level μ, the number of samples from the control and treatment populations and the number of genes that carry the signal.

In spite of these simplicity of the model we think some insight is gained about the relationships between the sample size and the signaling level at which some specified performance is obtained. The conclusion, under the assumption of equal signal strengths, is that there is considerable payoff for "genefinding" in the first few doublings of sample size, say from 2 to 4 and perhaps to 8. The reduction of signal level required to give specified "genefinding" performance continues and appears to agree with the anticipated asymptotic reduction by $\sqrt{2}$ for each doubling.

The purpose of computing is insight. R.W. Hamming

1 Introduction

Our simplistic view of the microarray-based *genefinding* process is the following. Genetic material from control and treatment experiments is applied to microarrays and the expression levels are read (we do not need to discuss the several technologies for doing this, although this simulation was motivated by thinking about the Affymetrix technology). Based on statistical tests comparing treatment and control data, we wish to identify candidate genes that signal the treatment. Of course some candidates may actually be biologically unrelated to the treatment due to randomness in the statistical decision process (false positives). Hence we are motivated to measure the performance of procedures for forming lists of candidate genes.

Since we do not know the real underlying distributions of expression level, we make the simplifying assumption they are normal and that the variations are independent from gene

to gene and that different samples are independent. We leave the interpretation "sample" to the reader, since it could be considered a replication for a single subject or possibly a single experiment for one of several subjects. Only the the mean levels of the signaling genes change during treatment. The following questions most certainly arise in any experimental program whose goal is to discover genes that signal a treatment : how many genes signal the treatment and how strong is the signaling? Without any biological experience we must be prepared to think that, depending on the treatment conditions, the number of signaling genes can be many or few and the signaling strengths can be strong or weak, or a mixture thereof.

To illustrate this point, Figure 1 presents simulated $|t|$ scores for (treatment - control) expression level differences of 100 genes in which there are 10 signaling genes randomly

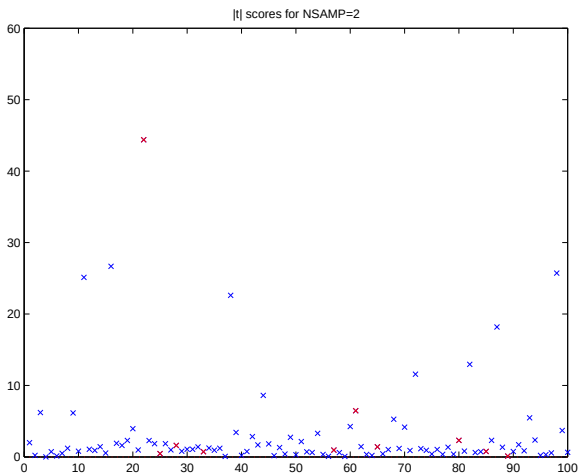


Figure 1: Simulated $|t|$ scores for 2 paired samples of control and treatment where control gene levels are $N(0,1)$ and signaling treatment genes are $N(\mu,1)$. The 10 signaling μ values are shown in Figure 2. Signaling genes shown in red. Only 2 signaling genes appear in the top 10 values of $|t|$.

placed within the 100 and having μ's as shown in Figure 2. We used the absolute value $|t|$ here for ease in plotting. For Figure 1 there are 2 samples (replications) per gene for both control and treatment. Gene expression levels under the control condition are $N(0,1)$ as they are also for non-signaling genes under the treatment condition. Expression levels are $N(\mu,1)$ for the signaling genes under the treatment condition. Only 2 of the 10 signaling genes appear in the top 10 values of $|t|$.

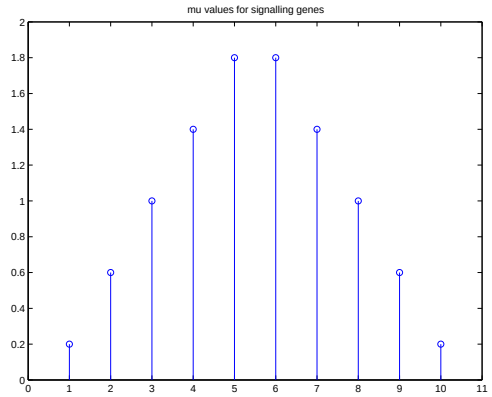


Figure 2: The 10 signaling μ values for the simulation experiment of Figure 1.

But as shown in Figure 3, by taking 8 paired samples of control and treatment and computing $|t|$ scores, more of the signaling genes are perceptible; now 6 signaling genes appear in the top 10.

This has demonstrated the main idea. In the general case, since we do not know a-priori where we are operating in μ-space, a reasonable strategy is to begin with a small number of samples and see if any (or how many) very significant t-scores are found. If there are many, and if enough of these make biological sense, then perhaps we are done. But if only a few are significant, taking adjustments for multiple hypotheses into account, then more samples would be called for. Taking more samples lets us perceive smaller values of μ in the noise. A major question is how much further down in μ can we see by an increase in sample size?

In this first simulation, which is motivated by Figures 1-3, we try to get an appreciation of the relationship between the correct identification of signaling genes and (1) the number of signaling genes, (2) the strength of the signaling genes and (3) the number of samples in the control and treatment groups. In all cases we assume the μ's for the signaling genes are all the same, which is yet a simpler case than that of Figure 1. We declare the signaling genes to be those having the highest 10 t scores. Choosing the top 10 is meant to represent the case in which only a few genes may be expected to signal the treatment condition. Other simulations are in progress for which many more genes are thought to signal the treatment condition. Also, in the remainder we use the highest t values rather than $|t|$, as in the previous paragraphs, because we have set $\mu > 0$ for the signaling genes.

2 The simulation

We provide estimates of two "genefinding" performance measures for the simple method of choosing N_{choose} genes from a large total N_G by taking those genes giving the top N_{choose} values of t scores computed (for each gene) from microarray expression level data. In this case it is assumed the control and treatment samples are paired so the t scores are based on the sample mean and variance of the *differences* between control and treatment. This

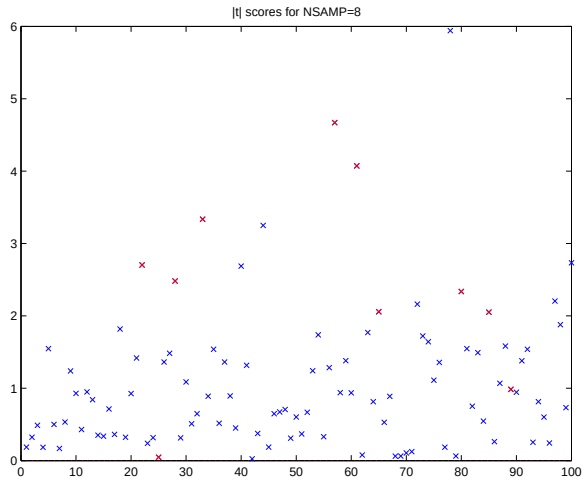


Figure 3: Simulated $|t|$ scores for 8 paired samples of control and treatment where control gene levels are $N(0,1)$ and signaling treatment genes are $N(\mu,1)$. The 10 signaling μ values are shown in Figure 2. Signaling genes shown in red. Now 6 signaling genes appear in the top 10 values of $|t|$.

assumption permits each sample to have a possibly random shift in mean that is common to the control and treatment.

The performance measures are:

1. Expected number of signaling genes found.

2. Probability that at least K signaling genes will be found.

Both of these quantities can be estimated from the empirical distribution of N_c, a random variable describing the number of correctly identified signaling genes found in an experiment.

The gene signaling model. All gene expression levels for the control group are made i.i.d normal with mean 0 and variance 1. We assume that only N_{good} genes carry the signal for the treatment and they are randomly chosen from all the genes. All gene expression levels for the treatment group are also independent and normal with variance 1, and all except the N_{good} signaling genes have mean 0 also. All the N_{good} signaling genes have mean $\mu \geq 0$, a parameter that may be interpreted as signal level.

In way of *criticism*, the normal constant variance model is much too simplistic although the results here would be unchanged if the genes had different variances provided the variances of control and treatment for each fixed gene were identical. That the signaling genes all have the same signal level $\mu \geq 0$ is also much too simplistic. It may be more realistic for the μ's for the signaling genes to be governed by a probability distribution, or even deterministically controlled as in Figure 2. The use of the t, which depends on normality,

can also be replaced with a non-parametric rank test at come cost in efficiency. But the main objective is to illustrate how well this process works in an idealized case, and to see the change in performance as a function of parameter values. So for now, these criticisms just motivate future improvements.

Gene identification. For each gene, the t statistic will be based on the sample mean of differences between control and treatment, and on the corresponding sample variance of the differences. This is a result of the *assumption that the control-treatment samples are paired.* True variances can also be dependent on the sample if we assume the shift in mean scales with the sigma. Since taking logs of the data removes scale-only effects of this sort, one can interpret the simulated variates to be logged data of this type.

Parameter definitions and values. Some of these were already defined but we list all here.

Symbol	Values	Description
N_G	12,000	Number of genes
N_{good}	5,10,20,40	Number of signaling genes
N_{choose}	10	Number of genes chosen by rank method
N_{samp}	2,4,8,16,32	Sample size of both control and treatment groups
N_{trials}	1000	Number of trials in the monte-carlo simulation
μ	as needed	Signal level of the signaling genes
N_c		Number of correctly identified genes

Quantities computed. For each setting of the parameters N_{good}, N_{samp} and μ, we compute the empirical distribution of

$$N_c = \text{No. correctly identified genes.}$$

Denote

$$\hat{p}_j = \frac{\#[N_c = j]}{N_{trials}}, \; j = 0, 1, \ldots, N_{choose}.$$

From these empirical distributions we plot the quantities

1. the mean of the empirical distribution, $\hat{m}_c = \sum_j j\hat{p}_j$;
2. $\widehat{P}[N_c \geq k] = \sum_{j \geq k} \hat{p}_j$, for $k = 1, 5, 9$

as a function of $log2(\mu)$ for each of the conditions $N_{good} = 5, 10, 20, 40$. This gives a total of 16 plots. We regret the large number of plots, but include them so because they may permit further analysis as in the next paragraph.

Discussion of the plots and formation of Table 1. The plots are presented in Figures 5 through 12. First note that the bottom plot of Figure 6 ($P[N_c \geq 9] = .5$) is completely void because the event $N_c \geq 9$ is impossible if $N_{good} = 5$. From these figures we can estimate the effect of sample size for a fixed number of good genes (N_{good}), or we can see the performance as a function of N_{good} for fixed sample size. Of course in real experiments we (the experimenter) have control of sample size but the values of N_{good} are unknown to us. Simulation may help us parametrically study the effect of N_{good} on the outcomes. To illustrate the use of these plots, we will investigate the effect of *sample size.*

For this simulation we have chosen the parameters $N_{good} = 5, 10, 20, 40$ and $N_{samp} = 2, 4, 8, 16, 32$ to be finite sequences that increase by a factor of two. This permits us to estimate the change in signal level μ required, as a function of doubling sample size N_{samp}, to meet some performance specification. For example, in the top of Figure 9, let us examine the changes in μ corresponding to the increasing of N_{samp} from 4 to 8. We denote $\hat{\mu}_{m,5}$ as the empirical solution to $\hat{m}(\mu) = 5$. Thus $\hat{\mu}_{m,5}$ moves from about $log2(\mu) = 2.6$ to $log2(\mu) = 1$, or a factor of $2^{1.6} = 3.03$. Then to increase N_{samp} from 8 to 16 decreases the required $log2(\mu)$ by 1, or a factor of 2 (table 1 gives a factor of 1.8). Note that the $log2(\mu)$ required for $N_{samp} = 2$ is not available on this scale. We can also determine the ratio of μ's required to maintain the probabilities $\widehat{P}[N_c \geq k] = \sum_{j \geq k} \hat{p}_j = P_0$ where we use $P_0 = .5$. Table 1 results from the application of this procedure to all of the figures.

Discussion of Table 1. Note first that the top part of Table 1 has many asterisks, each of which indicates that the quantity could not be determined from the computed curves. In some cases where μ values were not initially chosen well, additional simulations were done to supply the needed parameter values. However, the many asterisks in the section labeled $N_{good} = 5$ occur because $N_c \geq 9$ is impossible if $N_{good} = 5$ and achieving $\hat{m}_c = 5$ or $P[N_c \geq 5] = .5$ is not to be expected for finite μ.

The main observation to be made from the table is that for all of the performance measures used, the sample size doubling of 2 to 4 has a much larger effect on the decrease in discernable signal level than does the one from 4 to 8 and especially the latter doublings, say from 16 to 32. Here we take discernable signal level as the value of μ that gives some specified level of performance. Note the μ ratios for the 16 to 32 doubling are all near 1.5 whereas we anticipate a diminishing of $\sqrt{2}$ in the limit as the sample size tends to infinity. This asymptotic value of $\sqrt{2}$ would occur, for example, whenever the sampling distributions for $\hat{m}_c$ and $\hat{p}_j$ have means that are constant with respect to sample size and are symmetrically distributed about those means. Since we can expect the estimators $\hat{m}_c$ and $\hat{p}_j$ to be asymptotically normal, the preceding condition would hold asymptotically. The $\sqrt{2}$ dependence comes simply from the diminishing of sample variance due to doubling sample size.

To be a little more explicit, suppose the sample size N is sufficiently large so that $\hat{m}_c(2^k N)$ is normal ($k \geq 1$) with mean m_0 and variance $\sigma^2/2^k$. Denote z_p as the $100 \times p$th

$\Delta\ N_{samp}$	μ ratio for $\hat{m}_c = 5$	μ ratio for $P[N_c \geq 1] = .5$	μ ratio for $P[N_c \geq 5] = .5$	μ ratio for $P[N_c \geq 9] = .5$
$N_{good} = 5$				
2 to 4	*/*	*/4.94	*/*	*/*
4 to 8	*/*	4.9/1.71=2.86	*/*	*/*
8 to 16	*/*	1.71/0.97=1.76	*/*	*/*
16 to 32	*/*	.97/.61=1.59	*/*	*/*
$N_{good} = 10$				
2 to 4	*/7.38	45/3.65=12.3	*/8.03	*/25.20
4 to 8	7.38/2.24=3.29	3.65/1.38=2.64	8.03/2.38 = 3.37	25.20/4.71=5.35
8 to 16	2.24/1.22=1.83	1.38/0.78=1.77	2.38/1.26 = 1.89	4.71/2.17=2.17
16 to 32	1.22/0.78=1.56	0.78/0.51=1.53	1.26/0.81 = 1.55	2.17/1.30=1.67
$N_{good} = 20$				
2 to 4	240/5.18=46.3	39.42/2.75=14.3	200/5.58=35.8	*/13.90
4 to 8	5.18/1.75=2.96	2.75/1.09=2.52	5.58/1.83=3.05	13.90/3.05=4.56
8 to 16	1.75/0.97=1.80	1.09/0.64=1.70	1.83/1.01=1.81	3.05/1.49=2.05
16 to 32	0.97/0.62=1.56	0.64/.41=1.56	1.01/0.64=1.58	1.49/0.91=1.64
$N_{good} = 40$				
2 to 4	120/3.87=31.0	20.77/2.06=10.08	100/4.17=24.0	*/ 10.04
4 to 8	3.87/1.42=2.73	2.06/0.87=2.37	4.17/1.49=2.80	10.04/2.47=4.21
8 to 16	1.42/0.79=1.80	0.87/0.51=1.71	1.49/0.83=1.80	2.47/1.23=2.01
16 to 32	0.79/0.51=1.55	0.51/.33=1.55	0.83/0.53=1.57	1.23/0.75=1.64

Table 1: Ratios of signaling level μ required to achieve $\hat{m}_c = \sum_j j\hat{p}_j = 5$, and $\widehat{P}[N_c \geq k] = \sum_{j \geq k} \hat{p}_k = .5$, for $k = 1, 5, 9$. An asterisk indicates that the quantity could not be determined from the computed curves. In some cases where μ values were not initially chosen well, additional simulations were done for needed parameter values. However, the many asterisks in the section labeled $N_{good} = 5$ occur because achieving $\hat{m}_c = 5$ or $P[N_c \geq 5] = .5$ is not to be expected for finite μ.

percentile of a standard normal. Then

$$Pr\left[\frac{\hat{m}_c(N) - m_0}{\sigma} > z_p\right] = 1 - p$$

$$Pr\left[\frac{\hat{m}_c(2^k N) - m_0}{\sigma/\sqrt{2^k}} > z_p\right] = 1 - p.$$

Thus the change in the $100 \times p$th percentile threshold for increasing from N to $2N$ is

$$m_0 + z_p\sigma - m_0 - z_p\sigma/\sqrt{2} = z_p\sigma(1 - 1/\sqrt{2})$$

and for increasing from $2N$ to $4N$ it is

$$m_0 + z_p\sigma/\sqrt{2} - m_0 - z_p\sigma/2 = z_p\sigma(1 - 1/\sqrt{2})/\sqrt{2}.$$

This produces a ratio of changes in the $100 \times p$th percentiles due to doubling sample sizes of $1/\sqrt{2}$.

Since the quantities of interest, $\hat{m}_c$ and $\hat{p}_j$ both are with respect to the random variable N_c (no. good genes identified), we show empirical distributions of N_c from the simulations in the four plots of Figure 4. These plots show that the empirical distributions of N_c are far from symmetric (and far from Gaussian) for the cases $N_s = 4, 8, 16$ but looks much more Gaussian for $N_s = 32$.

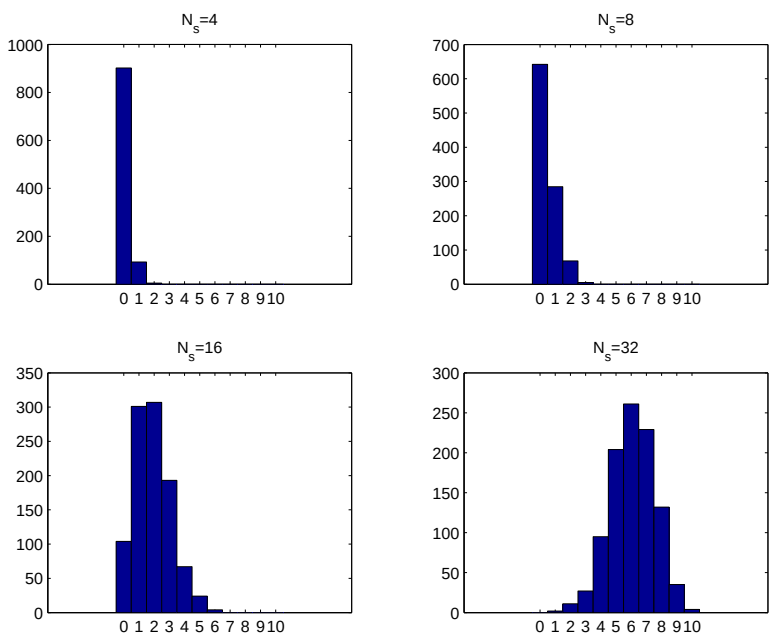


Figure 4: Empirical distributions of N_c for $\mu = .8, N_{good} = 20$ and $N_s = 4, 8, 16, 32$.

We remind that these results are based on equal signal strengths, but under this assumption, the simulation implies that there is considerable payoff in the reduction of discernable signal level in going from 2 to 4 and perhaps to 8 experiments. And although expensive to go from 2 to 4 or to 8, the expense is much less than from 16 to 32.

Confidence limits The 95% confidence interval for estimating a probability $P_0 = .5$ with a sample of 1000 is $[.469, .531]$ or roughly $.5 \pm .03$. This may be transformed back to a statement about μ using the experimentally determined curves (in the 16 Figures).

The confidence interval for $\hat{m}_c$ is estimated using the sample variances $\hat{\sigma}_c^2 = \sum_j (j - \hat{m}_c)^2 \hat{p}_j$ which were found to be at most 4.5 at μ's that gave $\hat{m}_c = 5$. Then using asymptotic normality of

$$\hat{m}_c = \sum_j j\hat{p}_j = \frac{1}{N_{trials}} \sum_{n=1}^{N_{trials}} N_{c,n}$$

where the number of correctly found in each trial, $N_{c,n}$ are considered independent random variables, the estimate of standard error for $\hat{m}_c$ is $\sqrt{4.5 \times 10^{-3}} = .067$. From the normality assumption, the 95% confidence interval around $\hat{m}_c = 5$ is within the interval $\hat{m}_c \pm .13$.

Comments, Suggestions for improvement. The purpose of this exercise was to see if simulation could help our understanding about the interplay of the parameters N_{samp}, N_{good} and N_{choose} in the simple genefinding algorithm of choosing the top N_{choose} t scores. The author welcomes comments, suggestions and new questions that arise from this small effort. The following items of improvement seem clearly interesting.

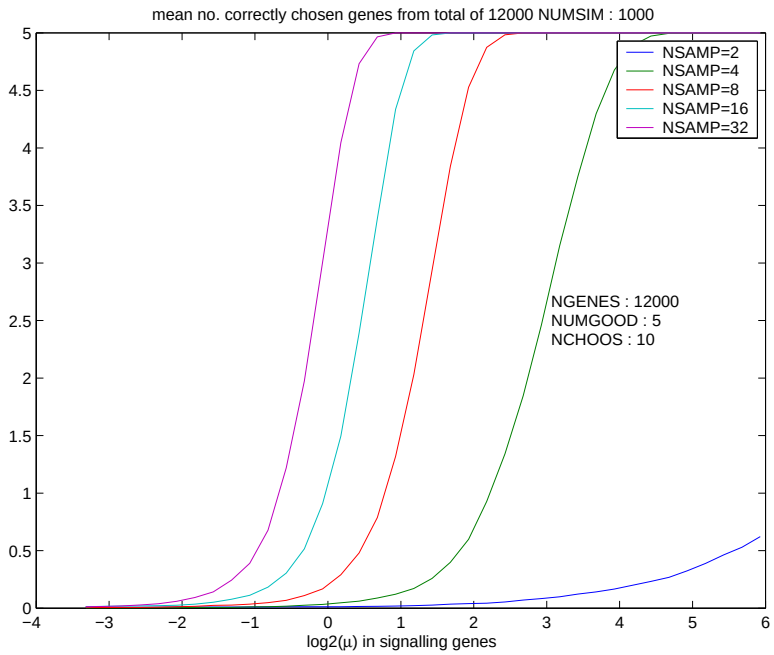
1. Evaluate the effect of other signal strength distributions (here it is constant, i.e., uniform).

2. Base the random expression levels on empirical distributions from observed data or on distributions whose parameters are determined by observed data.

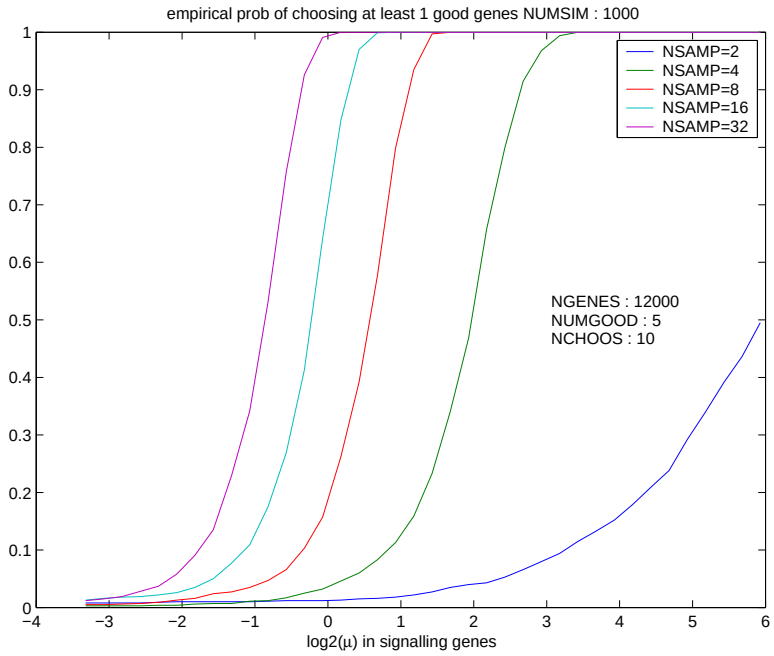
3. Use simulation to help understand how the elements of the *list* change as sample size is increased. For example, given a realization from $N_{samp} = 2$, what should we expect from our gene list, t scores, etc, when we go to $N_{samp} = 4$?

References

[1] A. D. Whalen, *Detection of Signals in Noise*, Academic Press, New York, 1971.

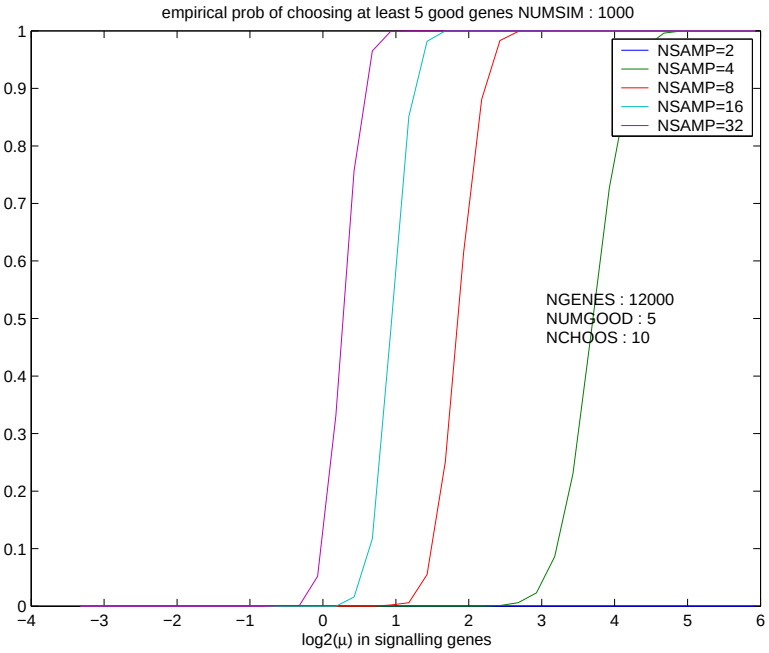


(a) Average number of identified good genes.

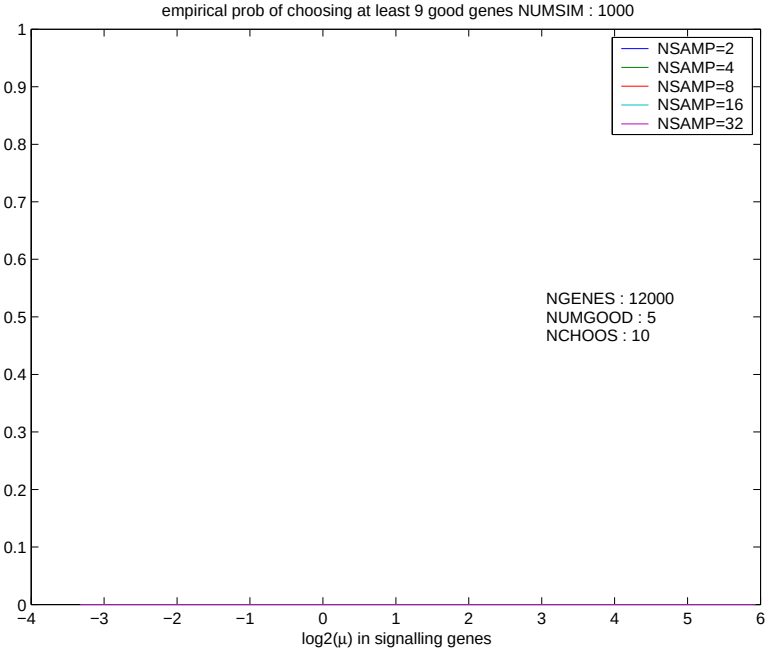


(b) Probability of identifying at least 1 good gene.

Figure 5: Results of genefinding simulation, $N_{choose} = 10$, $N_{good} = 5$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.

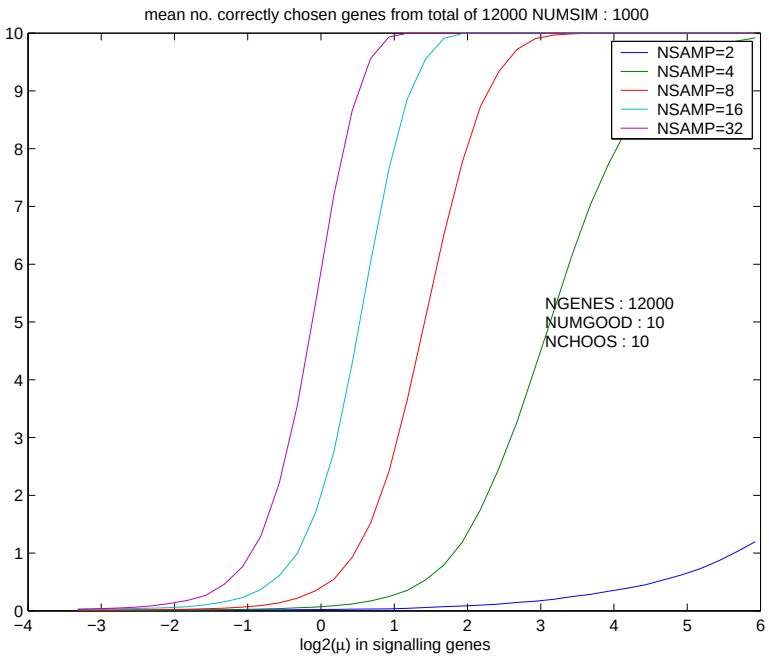


(a) Probability of identifying at least 5 good genes.

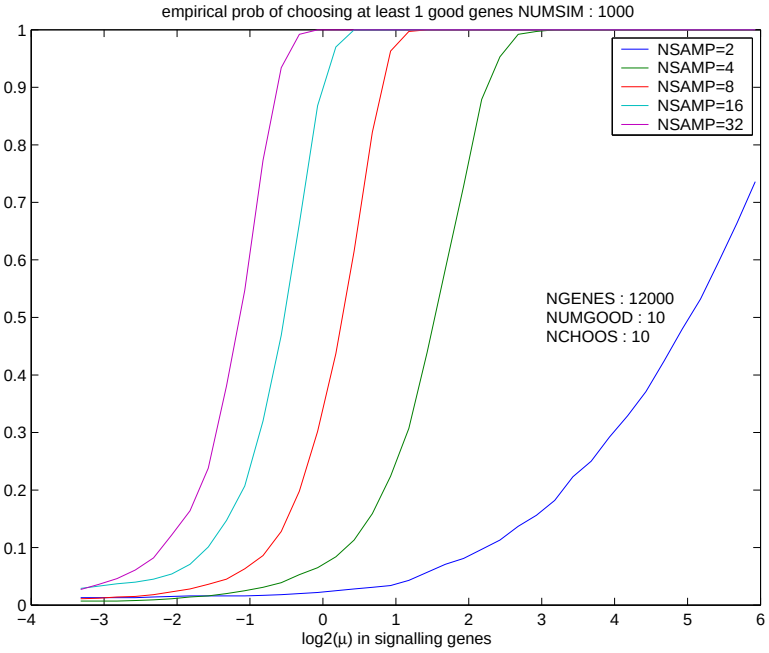


(b) Probability of identifying at least 9 good genes.

Figure 6: Results of genefinding simulation, $N_{choose} = 10$, $N_{good} = 5$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations. Note $N_c \geq 9$ is impossible for $N_{good} = 5$.

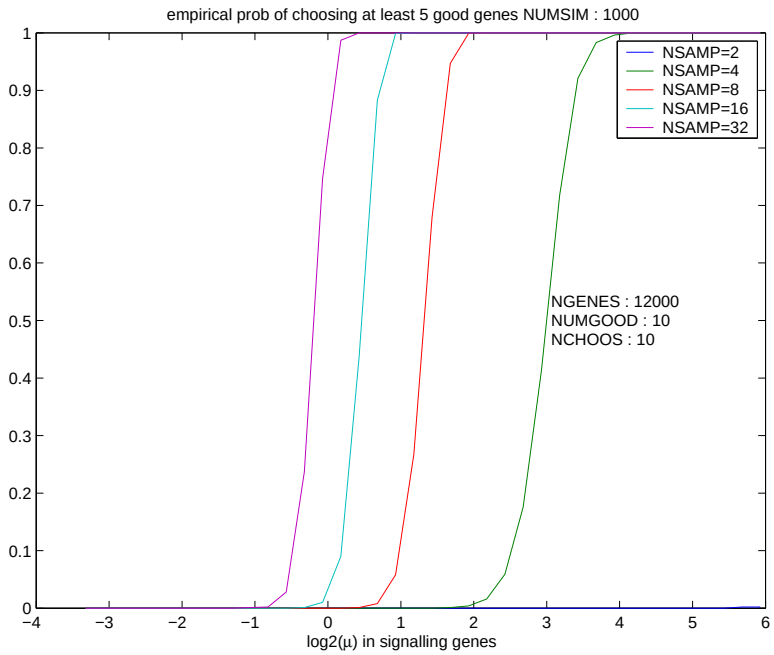


(a) Average number of identified good genes.

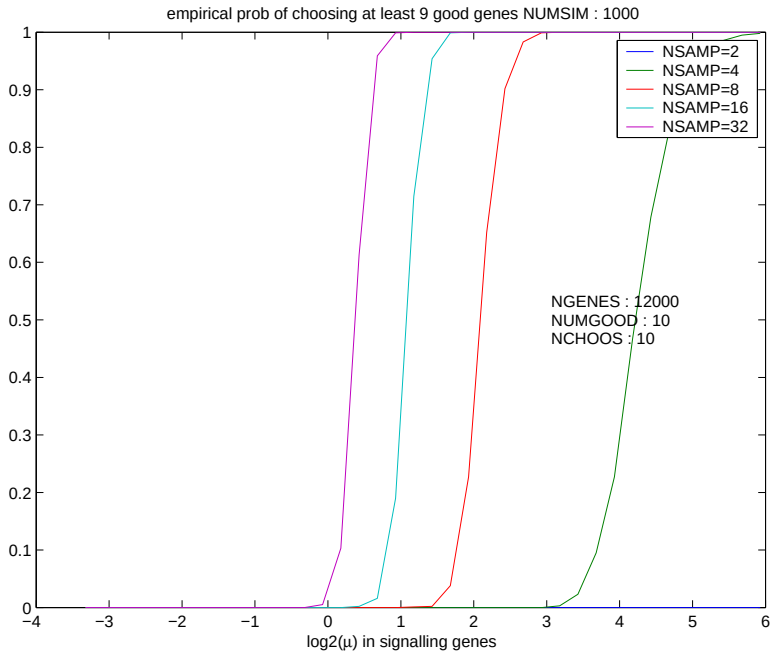


(b) Probability of identifying at least 1 good gene.

Figure 7: Results of genefinding simulation, $N_{choose} = 10$, $N_{good} = 10$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.

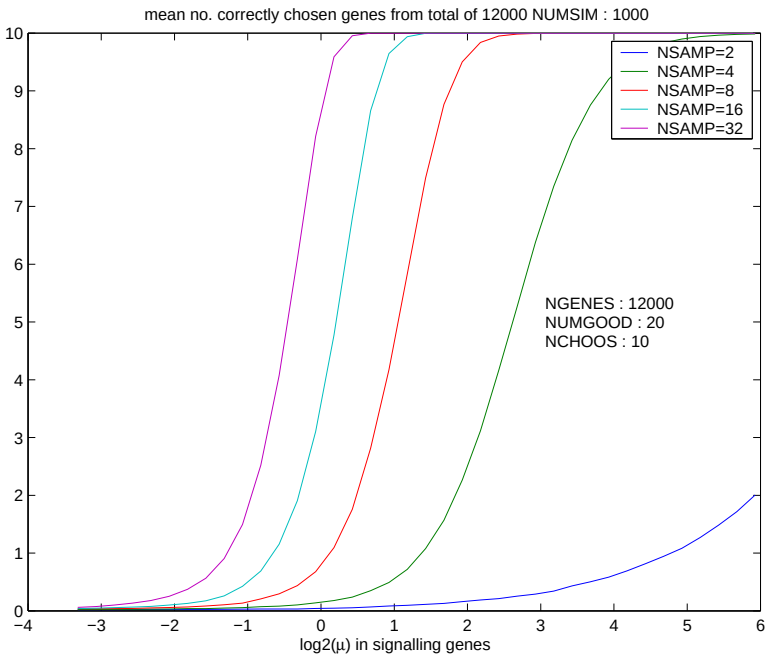


(a) Probability of identifying at least 5 good genes.

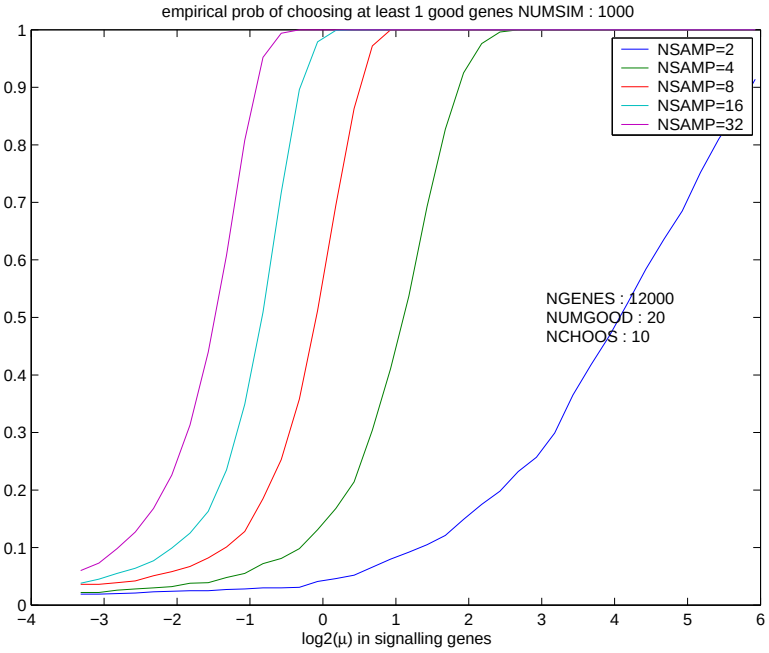


(b) Probability of identifying at least 9 good genes.

Figure 8: Results of genefinding simulation, $N_{choose} = 10$, $N_{good} = 10$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.

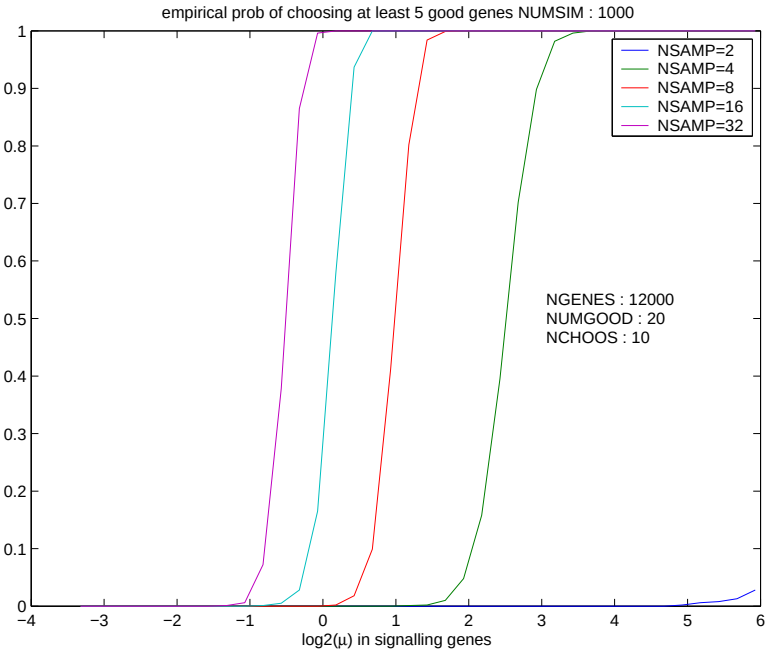


(a) Average number of identified good genes.

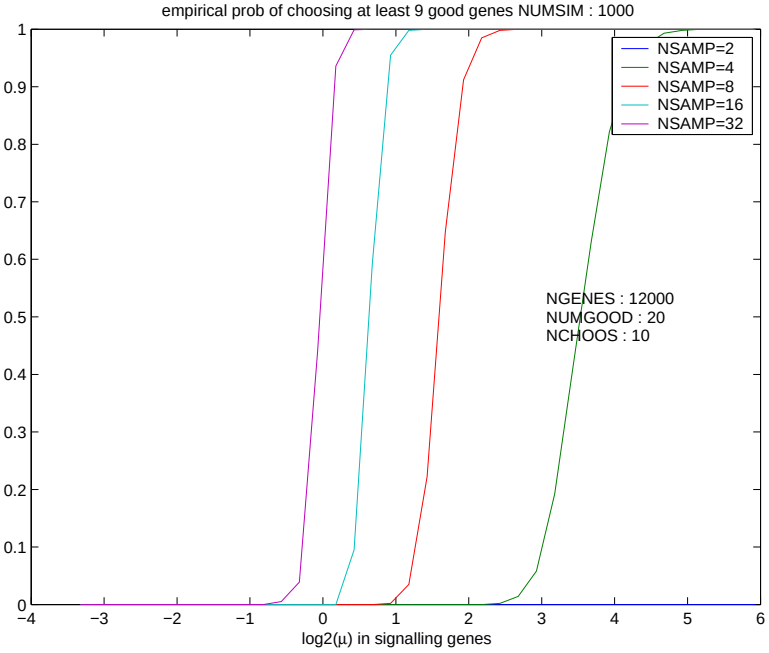


(b) Probability of identifying at least 1 good gene.

Figure 9: Results of genefinding simulation, $N_{choose} = 10$, $N_{good} = 20$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.

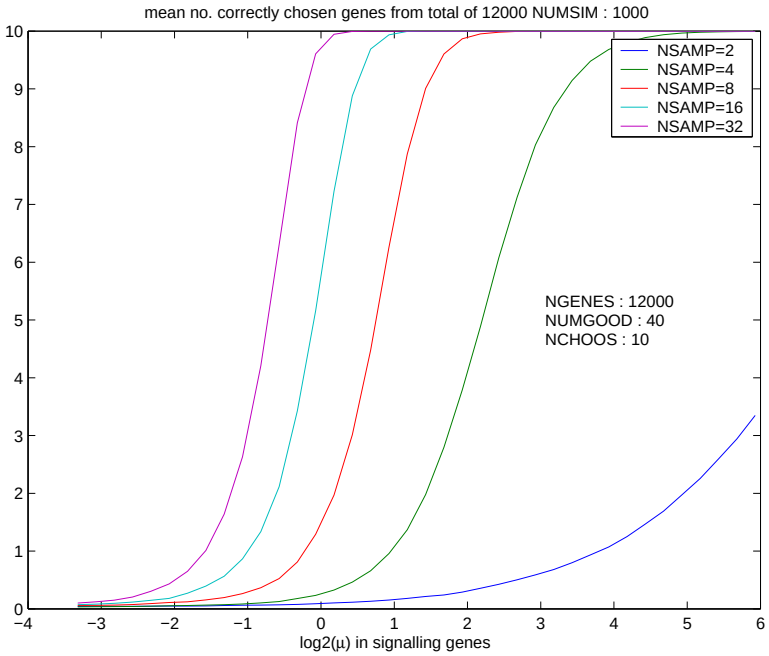


(a) Probability of identifying at least 5 good genes.

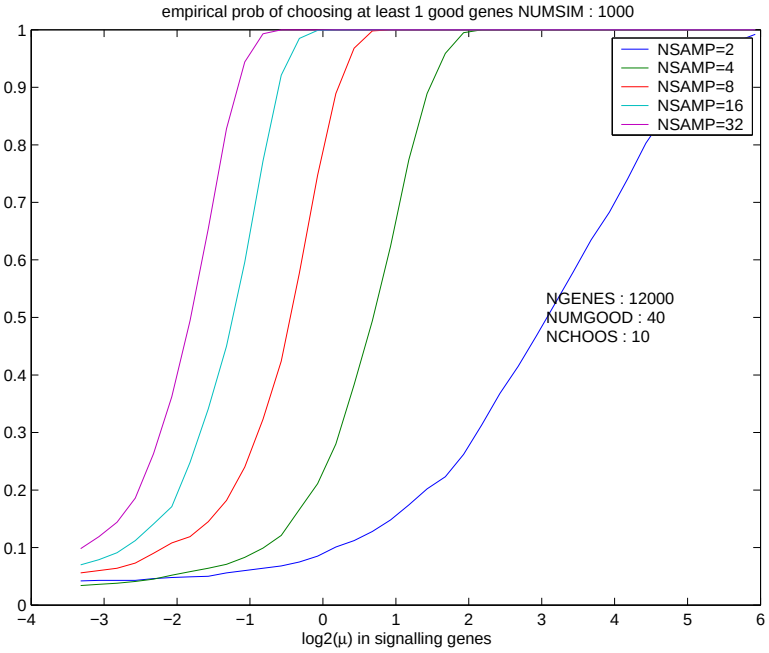


(b) Probability of identifying at least 9 good genes.

Figure 10: Results of genefinding simulation, $N_{choose} = 10$, $\mathrm{N}_{good} = 20$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.

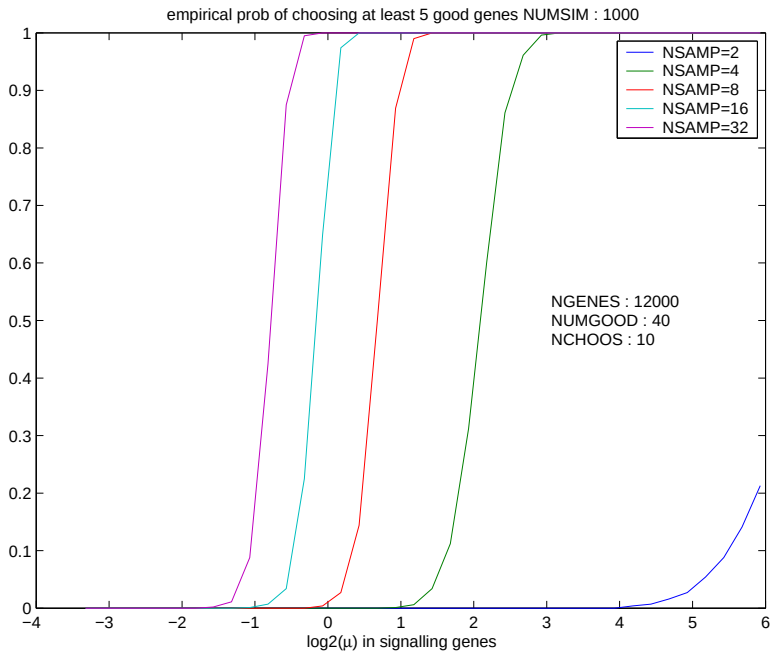


(a) Average number of identified good genes.

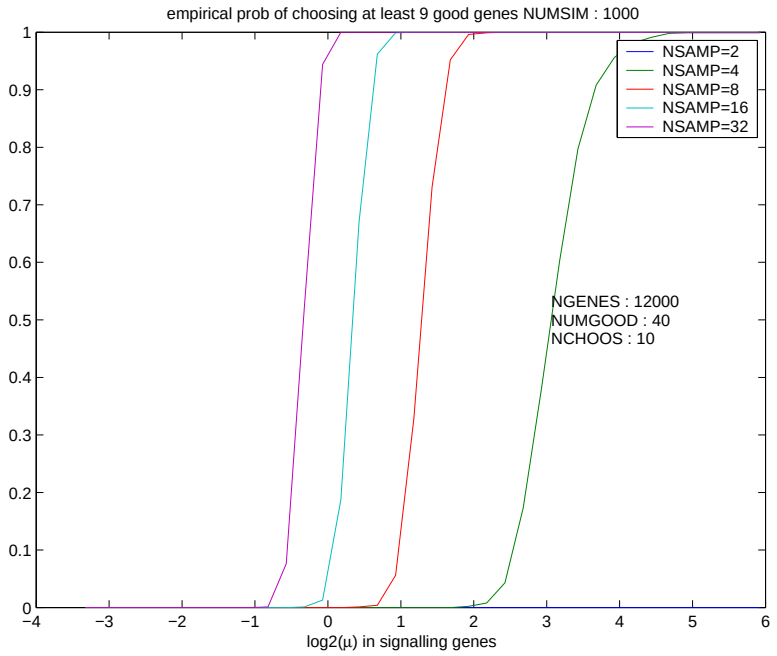


(b) Probability of identifying at least 1 good gene.

Figure 11: Results of genefinding simulation, $N_{choose} = 10$, $\mathrm{N}_{good} = 40$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.



(a) Probability of identifying at least 5 good genes.



(b) Probability of identifying at least 9 good genes.

Figure 12: Results of genefinding simulation, $N_{choose} = 10$, $\mathrm{N}_{good} = 40$, $N_{trials} = 1000$. Genes chosen from top 10 t-scores based on paired observations.