

POWER_SAGE: comparing statistical tests for SAGE experiments

Michael Z. Man^{1,*}, Xuning Wang² and Yixin Wang

¹ Biostatistics and ² Bioinformatics, PGRD 2800 Plymouth Road Ann Arbor, MI 48105 USA

Received on December 22, 1999; accepted on June 7, 2000

Abstract

Motivation: The Serial Analysis of Gene Expression (SAGE) technology determines the expression level of a gene by measuring the frequency of a sequence tag derived from the corresponding mRNA transcript. Several statistical tests have been developed to detect significant differences in tag frequency between two samples. However, which one of these tests has the greatest power to detect real changes remains undetermined.

Results: This paper compares three statistical tests for detecting significant changes of gene expression in SAGE experiments. The comparison makes use of Monte Carlo simulation that, in essence, generates 'virtual' SAGE experiments. Our analysis shows that the Chi-square test has the best power and robustness. Since the POWER_SAGE program can easily run 'virtual' SAGE studies with different combinations of sample size and tag frequency and determine the power for each combination, it can serve as a useful tool for planning SAGE experiments.

Availability: The POWER_SAGE software is available upon request from the authors.

Contact: michael.man@pfizer.com

Introduction

Serial Analysis of Gene Expression (SAGE) is used to assess mRNA expression profiles and is based upon on the following two principles (Velculescu *et al.*, 1995). A short sequence tag derived from a defined position within a transcript contains sufficient information to uniquely identify the transcript. Concatenation of tags in a serial fashion allows high-throughput sequencing. As a counting-based technique, SAGE makes it easy to quantify expression levels of many transcripts simultaneously.

As SAGE offers a much more rapid and cost-effective method of identifying cDNAs than other methods, one can aspire to examine the entire transcriptome of 10 000 to 50 000 expressed genes in a cell. In a typical SAGE experiment, there are at least two samples (Chen *et al.*,

1998; Kal *et al.*, 1999; Zhang *et al.*, 1997). The usual question of interest is whether one sample has a significant change in expression relative to the other sample for each transcript (Audic and Claverie, 1997; Kal *et al.*, 1999). One of the most attractive applications of transcript profiling is to address the question of expression differences between normal and diseased samples or between samples with and without drug treatment. This approach will define the subset of genes that are potential diagnostic markers or therapeutic targets.

Over the past few years, several methods have been developed for determining the statistical significance of gene expression difference SAGE experiments. Zhang *et al.* use a simulation approach to determine the probability of obtaining the observed difference and more extreme ones (Zhang *et al.*, 1997). This method is part of the SAGE software (available at <http://www.sagenet.org/>). Because this method entails a large number of simulations (100 000), it is not suitable for fast and interactive applications, such as web access (Lal *et al.*, 1999). Madden *et al.* use a confidence interval approach based on Poisson distribution (Madden *et al.*, 1997). The difference in two samples is statistically significant if $(N_1 - kN_1^{1/2} - N_2 - kN_2^{1/2}) > 0$, where N_1 and N_2 are the tag numbers of one transcript from two samples, $N_1 \geq N_2$, $k = Z_{1-\alpha}$, α is the significance level. This method is not widely accepted because it is not appropriate for unequal sample sizes and lacks sensitivity (Kal *et al.*, 1999).

The Bayesian method proposed by Chen *et al.* (1998) and further modified by Lal *et al.* (1999), calculates the posterior probability of x , which is defined as $x = \frac{y}{y+z}$, where y and z are tag numbers of one particular transcript in two samples. This approach has a few shortcomings. Firstly, there is a difference on the choice of parameters for the prior distribution of x , which depends on the conditions or cell types in the comparison (Chen *et al.*, 1998; Lal *et al.*, 1999). Because of this difference, the cut-off for significant change is inconsistent. Secondly, the false positive rate of this method can not be assessed as this method is not based on hypothesis testing.

*To whom correspondence should be addressed.

There are several hypothesis-testing based methods that are more rigorous. A Bayesian approach proposed by Audic and Claverie for examining EST data has been adopted to analyse SAGE data (Audic and Claverie, 1997). Kal *et al.* use a Z statistic, $Z \sim N(0, 1)$, to test the equality of two proportions in two experimental conditions (Kal *et al.*, 1999). This method has the additional advantage for calculating sample size that is necessary to detect a pre-specified difference in proportions (Kal *et al.*, 1999). Alternatively, the Chi-square test and Fisher's exact test can also be used. In fact, the Chi-square test for 2×2 table is exactly the same as the method proposed by Kal *et al.* because of the relation that exists between the Z statistic and the Chi-square statistic with one degree of freedom, $Z^2 = X_1^2$ (Fisher and van Belle, 1993; Hogg and Craig, 1995). Fisher's exact test is appropriate when sample size requirement (the expected cell count ≥ 5) is not met for the Chi-square test, but is generally more computationally expensive (Stokes *et al.*, 1995).

There has been no report in the literature on the systematic comparison of these hypothesis-testing based methods in analysing SAGE experiments. Such comparison is warranted for several reasons. Firstly, given the high cost of conducting SAGE experiments, we are compelled to use the most sensitive method. Secondly, we like to know whether a proposed SAGE study will have sufficient sensitivity to achieve an experimental goal. Thirdly, statistical rigorousness should be applied using established approaches that are recognized by both the scientific community and the regulatory agencies (FDA, 1998).

Different statistical tests can be compared in terms of specificity, power, and robustness (Fisher and van Belle, 1993). Specificity is measured by alpha, the type I error rate or the false positive rate, which is defined in statistical terms as the probability of erroneously rejecting the null hypothesis when it is true (Hogg and Craig, 1995). Smaller alpha means higher specificity. Alpha is controlled by the significance level (usually at 5%). In statistical terms, the sensitivity of a statistical test is defined more rigorously as power, which is the probability of correctly rejecting a null hypothesis when it is not true (Hogg and Craig, 1995). A robust statistical test should perform well for a broad range of data, or for a variety of probability distributions (Fisher and van Belle, 1993). A test is robust if the observed alpha is close to the pre-specified significance level for all ranges of data, or for different distributions. A robust test protects us from making too many, or too few type I errors (resulting in an over-conservative test). A statistical test that lacks robustness should be avoided because of the fluctuating type I error rate. Given the same significance level, a statistical test with higher power and robustness is more desirable because smaller differences can be detected with the same sample size, or the same difference can be detected with a smaller sample size

(Fisher and van Belle, 1993).

Since it is not practical to repeat SAGE experiments hundreds of times, we resort to the Monte Carlo approach to simulate SAGE experiments. Monte Carlo, in essence, performs statistical sampling trials on a computer using random number generators (Hammersley and Handscomb, 1964; Manly, 1991). It has been used in a wide variety of applications to calculate power and sample size (Collier and Baker, 1966; Sutradhar and Bartlett, 1993; Tanizaki, 1997).

This paper compares three hypothesis-testing based methods for determining statistical significance in SAGE studies, the Chi-square test (or test for equality of two proportions), Fisher's exact test, and Audic and Claverie's Bayesian method. Monte Carlo simulation is used for comparison of these tests in terms of specificity, power, and robustness.

Systems and methods

A SAGE experiment is analogous to taking a handful of variably colored beans from a big bag and then estimating the proportion of beans of each color. Mathematically, this can be modeled as sampling from multinomial distribution. Since the abundance of a transcript of a typical gene is low (less than 1%), one can use binomial distribution when a single gene is concerned.

$$X_i \sim B(N, p_i), \text{ where } X_i \text{ is the tag number of gene } i, \\ N \text{ the total number of tags sequenced,} \\ p_i \text{ the frequency of gene } i.$$

To carry out a Monte Carlo SAGE experiment, one needs to generate X_i s from the specified binomial distribution, $B(N, p_i)$. A simple example of generating a X_i from $B(10\,000, 0.5)$ would be flipping a coin 10 000 times and counting the number of times that the 'head' appears. A binomial deviate routine, based on the rejection method, is used to generate the X_i s (Press *et al.*, 1992).

Calculation of alpha The null hypothesis is that there is no difference in expression between two samples, $p_i^c = p_i^t$. Assume the total number of tags for each sample is N and fix p_i^c equal to p_i^t , generate X_i^c and X_i^t using the binomial distribution $B(N, p_i^c)$, then perform the statistical test and calculate the p -value. Iterate the process a large number of times (e.g. 1000). Alpha is calculated as the percentage of times that the null hypothesis is rejected at a prescribed significant level (e.g. 0.05).

Calculation of power Assume the total number of tags for each sample is N and fix $p_i^t = \text{fold} * p_i^c$ (fold = 0.5, 2, 3, ...), generate X_i^c and X_i^t using $B(N, p_i^c)$ and $B(N, p_i^t)$ respectively, then perform the statistical test and calculate the p -value. Iterate many times (e.g. 1000). Power is the

percentage of times that the null hypothesis is rejected at a prescribed significant level (e.g. 0.05).

Algorithm

When one particular gene (Gene A) is of interest, one can arrange the data in the following table.

	Sample 1	Sample 2	
Gene A	n_{11}	n_{12}	$N_{1.}$
Others	n_{21}	n_{22}	$N_{2.}$
	$N_{.1}$	$N_{.2}$	$N_{..}$

n_{ij} ($i = 1, 2; j = 1, 2$): the tag count for Gene A or others for condition j ;

$N_{i.}, N_{.j}$: the marginal total of tag count for the row or column;

$N_{..}$: the total number of tags in the two conditions.

The Chi-square test or test for equality of two proportions
The formula of the Chi-square statistic for 2×2 table is given below (Fisher and van Belle, 1993).

$$X^2 = \frac{N_{..}(n_{11}n_{22} - n_{12}n_{21})^2}{N_{1.}N_{2.}N_{.1}N_{.2}}. \quad (1)$$

The Z statistic, used by Kal *et al.* to test the equality of two proportions, is calculated as follows (Kal *et al.*, 1999).

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1 - p_0)(1/N_{.1} + 1/N_{.2})}} \quad (2)$$

$$\begin{aligned} p_1 &= n_{11}/N_{.1} \\ p_2 &= n_{12}/N_{.2} \\ p_0 &= (p_1 + p_2)/2. \end{aligned}$$

It can be shown quite easily that (1) and (2) are equivalent, $X^2 = Z^2$ (Fisher and van Belle, 1993; Hogg and Craig, 1995).

Fisher's exact test Fisher's exact test is based on hypergeometric distribution (Stokes *et al.*, 1995). Fisher and van Belle summarized Kennedy *et al.*'s study on the use of coronary artery bypass graft surgery in emergency cases vs. non-emergency cases (all other cases) (Kennedy *et al.*, 1981; Fisher and van Belle, 1993).

	Discharge status		
Surgical priority	Dead	Alive	
Emergency	n_{11}	n_{12}	$N_{1.}$
	1	19	20
Other	n_{21}	n_{22}	$N_{2.}$
	7	369	376
	$N_{.1}$	$N_{.2}$	$N_{..}$
	8	388	396

In Fisher's exact test, marginal totals ($N_{1.}$, $N_{2.}$, $N_{.1}$, and $N_{.2}$) are considered to be fixed. Following the hypergeometric distribution, the probability of observing each possible table configuration can be calculated (Stokes *et al.*, 1995).

$$P = \frac{N_{1.}!N_{2.}!N_{.1}!N_{.2}!}{N_{..}!n_{11}!n_{12}!n_{21}!n_{22}!} \quad (3)$$

n_{11}	n_{12}	n_{21}	n_{22}	Probability
0	20	8	368	0.658 06
1	19	7	269	0.28535*
2	18	6	270	0.051 29
3	17	5	271	0.004 98
4	16	4	272	0.000 28
5	15	3	273	0.000 01
6	14	2	274	0
7	13	1	275	0
8	12	0	276	0

The two-sided p -value is the sum of all the probabilities that are less than or equal to the observed probability (*). Since the p -value is 0.3419, the data fail to reject the null hypothesis of equal proportions and the observed difference is well within what is expected by chance alone.

Audic and Claverie's Bayesian method Audic and Claverie's Bayesian method was originally developed to analyse EST data. In the following, 'tag' can be replaced with 'EST' if ESTs are used in the comparison.

The probability of observing n_{12} tags in second library given n_{11} tags observed in first library is expressed as:

$$P(n_{12} | n_{11}) = \binom{N_{.2}}{N_{.1}}^{n_{12}} \frac{(n_{11} + n_{12})!}{n_{11}!n_{12}!(1 + N_{.2}/N_{.1})^{(n_{11}+n_{12}+1)}}.$$

The p -value is computed from cumulated density function as

$$P = \min \left\{ \sum_{k=0}^{k \leq n_{12}} P(k | n_{11}), \sum_{k=n_{12}}^{\infty} P(k | n_{11}) \right\}.$$

(Audic and Claverie, 1997).

Implementation

A program, POWER_SAGE, was written in C language to perform Monte Carlo simulation and calculation of power and alpha. It was compiled and tested in an SGI IRIX environment. Since no special library was used, it should run on UNIX, Windows, NT, Linux, etc. It can be run as command-line driven, stand-alone program, or can be run via a web-interface (Figure 1).

a)

Power Analysis of Statistical Tests for SAGE

Input your specification as PLAIN TEXT here:

```

-19
2
1000
5
1 0.01 50000 50000
2 0.001 50000 50000
3 0.0001 50000 50000
4 0.0002 100000 200000
5 0.0005 200000 50000
    
```

A sample specification (don't enter the comments) :

```

-17                # Seed for random number generator
2                 # Difference Criteria. 1: 2 - STD, 2: 2 - fold, 3: both
100              # Iteration to get alpha and power
1                # Number of transcripts
1 0.01 5000 5000  # ID, frequency, N1, N2 (total # of tags)
    
```

To receive the output, please change to YOUR email address :

(b)

```

POWER_SAGE 1.2          July 29, 1999
Michael Z. Man         michael.man@wl.com
Biometrics
Parke-Davis Pharmaceutical Research
Seed      :             -19
Diff Crit. :             2
Replication :          10 000
K (elements):           1
ID & prob. :             3           0.000 200           50 000           50 000
    
```

	Group_1		Group_2		Audic and Claverie		Chi-sqr		Fisher's Exact	
	Diff	Level (log)	Expected	Expected	alpha	power	alpha	power	alpha	power
0.5Fold	-3.7		10.0	10.0	3.82	24.00	5.08	24.54	2.95	18.17
2Fold	-3.7		10.0	10.0	4.01	37.56	5.52	46.02	3.11	38.57
3Fold	-3.7		10.0	10.0	3.81	86.43	5.33	89.39	3.05	86.96
4Fold	-3.7		10.0	10.0	3.76	99.20	5.18	99.50	3.07	99.23
5Fold	-3.7		10.0	10.0	3.64	99.95	5.05	99.96	2.87	99.95
6Fold	-3.7		10.0	10.0	3.79	100.00	5.31	100.00	2.91	100.00
7Fold	-3.7		10.0	10.0	3.90	100.00	5.39	100.00	2.91	100.00
8Fold	-3.7		10.0	10.0	3.62	100.00	4.83	100.00	3.05	100.00
9Fold	-3.7		10.0	10.0	3.71	100.00	5.17	100.00	2.95	100.00
10Fold	-3.7		10.0	10.0	3.59	100.00	5.01	100.00	2.82	100.00

Fig. 1. The Web Interface for POWER_SAGE Program Specification of virtual SAGE experiments can be entered in the box (a). The output is delivered via email. An example of output data is given in (b).

Results and discussion

The three statistical tests all have similar power with consistent alpha (5%) at high expression level (≥ 20 tags) (Figure 2a). The Chi-square test has higher power at the low expression level (≤ 15 tags) than the other two. When unequal sample sizes in two samples (e.g. 50 000 vs. 250 000, or 250 000 vs. 50 000) and different fold differences (e.g. 0.5-, 3-, 4-fold, etc.) are used, the Chi-square test consistently has the highest power (simulation results not shown). Since most transcripts (>95% of all transcripts) have low expression (tag number ≤ 15) (Figure 3), a closer examination at this low range is warranted. For transcripts with fewer than 15 tags, the Chi-square test has 5~10% higher power than both Fisher's exact test and Audic and Claverie's Bayesian method (Figure 2b). This increase does not seem to be a great difference at first glance. However, considering that the power is only from 0~60% at this range, a 5~10% improvement is significant. It is also worth noting that the decrease of power of Audic and Claverie's Bayesian method is accompanied by the decrease of alpha at this low expression range (1~15 tags). Our result on the decrease of alpha is in agreement with Audic and Claverie's observation of the more 'conservative behaviour' of their test at low expression range (Audic and Claverie, 1997). The Chi-square test, on the other hand, has an observed alpha consistently close to the pre-specified significance level (5%), even in the low expression range. This demonstrates the robustness of the Chi-square test.

In the original paper (Audic and Claverie, 1997), Audic and Claverie's Bayesian method and Fisher's exact test were compared. It was claimed that Fisher's exact test was more conservative than Audic and Claverie's Bayesian method. This seems to contradict our finding that the two tests are comparable in terms of power and alpha. A closer examination reveals that Fisher's exact test calculates a two-tail p -value and Audic and Claverie's Bayesian method calculates two one-tail cumulative probabilities, of which the smaller one is reported as the p -value (Audic and Claverie, 1997). Consequently, $\alpha/2$ (α being the significance level) should be used for Audic and Claverie's procedure to adjust for the overall false positive rate. In their example, the two-tailed p -value of Fisher's exact test (4.6%) should be compared against 5%, while the p -value from Audic and Claverie's procedure (1.6%) should be compared against 2.5%. Given that both p -values were statistically significant and only one example was used, one could not conclude that Audic and Claverie's Bayesian method is more powerful (less conservative) than Fisher's exact test.

There are arguments against aggregating genes together to construct a 2×2 table that the Chi-square test (or the

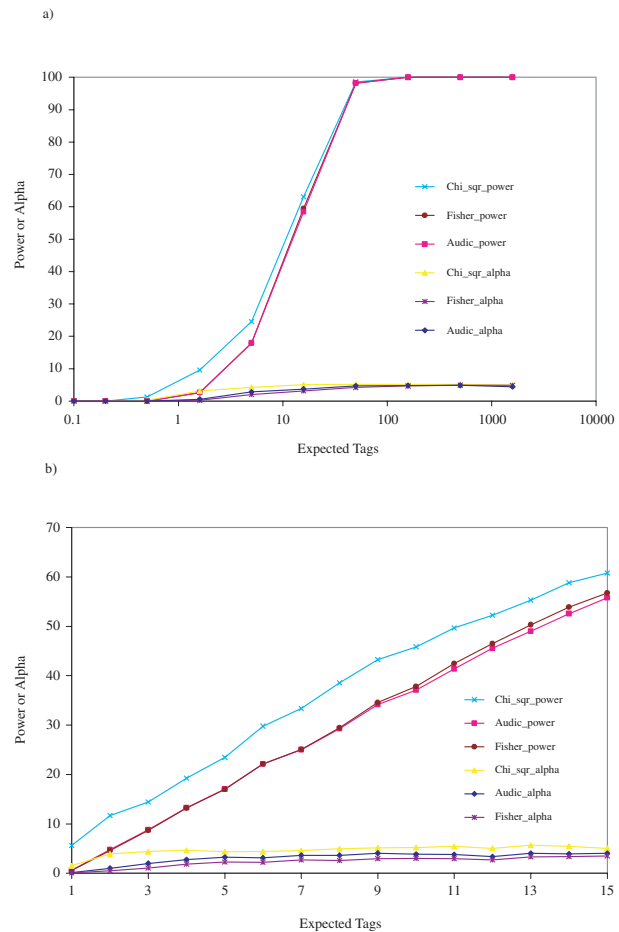


Fig. 2. Comparison of the Chi-square test, Fisher's exact test, and Audic and Claverie's Bayesian method. (a) The total number of tags is the same for each sample, $N.1 = N.2 = 50\,000$. The significance level is set at 5%. The total number of iterations used to get power and alpha is 10 000. The tag frequency of a transcript in the control sample is p^c and that in the treatment sample $2p^c$. (b) Comparison results at low expression levels. The specifications are the same as those in (a).

Z-statistic based test of equality of two proportions) and Fisher's exact test use (see **Algorithm** section) (Audic and Claverie, 1997; Claverie, 1999). Claverie argues that "the definition of the 'all others' aggregated gene category is logically inconsistent, as it implies that the genes expressed and observed in conditions A and B are the same, which might be a largely incorrect assumption" (Claverie, 1999). Using this argument, the study described in the **Algorithm** section (emergency cases vs. non-emergency cases—aggregated all others) would then be incorrect. SAGE or EST experiments are essentially sampling processes from a population of all transcripts and are hence probabilistic in nature. The failure to observe a

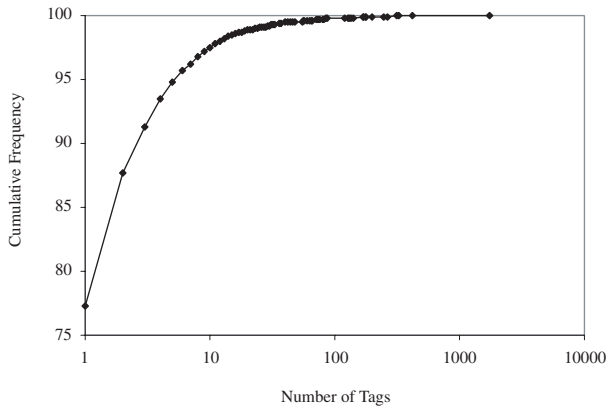


Fig. 3. Cumulative distribution of transcripts in one SAGE experiments. The total number of tags collected from the sample is 20 562, representing 7707 transcripts. The majority of transcripts have less than 15 tags. Cumulative frequency is plotted against the number of tags in log scale.

transcript in one sample does not prove that the transcript is not present in the population. With a larger sample size, the transcript may be easily found. Even with a very large sample size, not observing the transcript does not prove the absence of the transcript.

The Chi-square test requires random sampling and a large sample size (expected cell counts ≥ 5) (Fisher and van Belle, 1993). Combined tag number from two samples greater than 10 will meet the sample size requirement if the sizes of the two samples are roughly equal. Fisher's exact test is more restricted (fixing both row and column margins) and is used when the sample size is too small for the Chi-square test (Fisher and van Belle, 1993). In practice, biologists view any combination of tag numbers from two groups less than 10 with reservation (personal communication). In the range that represents expression levels of most genes in SAGE experiment (≤ 15 tags), the Chi-square test is more powerful than Audic and Claverie's test and Fisher's exact test. Therefore, the Chi-square test is preferred for its higher power and robustness. It can be easily calculated using many statistical software. An option in POWER_SAGE can be used to perform Chi-square tests for SAGE experiments.

Because SAGE is an expensive and laborious process, studies should be carefully planned to ensure that the scientific question could be addressed with sufficient sensitivity and minimal resource. The determination of sample size is therefore very important. Actually, sensitivity/power and sample size are intimately related. R.A.Fisher, father of modern statistics, stated: 'By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the

Table 1. An example using POWER_SAGE to select optimal sample size. Assume that the total numbers of tags from each of the two samples are the same (= sample size). The tag frequency of the transcript in the control sample (p^c) is set at 0.1% or 0.01% and that in the treatment sample (p^t) is $2p^c$. The significant level is 5%. The total number of iteration is 1000 (10 000 iteration gives similar results). The simulation results from POWER_SAGE are validated using nQuery Advisor®, which is a more general package for sample size and power calculations (Elashoff, 1999). The results are in good agreement except that there is slight overestimate of power using nQuery Advisor when the large sample requirement (expected cell count ≥ 5) is not met. For example, when 10 000 tags are collected for a transcript at 0.01% in the control and 0.02% in the treatment, the expected cell count for each of the corresponding cells is $1.5 \left(\frac{0.01\% + 0.02\%}{2} * 10\,000 \right)$. In this case, the large sample estimation of power used in nQuery Advisor is no longer appropriate

Sample size	Power (Chi-square)	
	@0.1% expression level (%)	@0.01% expression level (%)
5 000	24.8	1.4
10 000	44.7	5.6
25 000	84.1	11.3
50 000	98.4	23.8
100 000	100	44.1
200 000	100	73.6
300 000	100	88.4

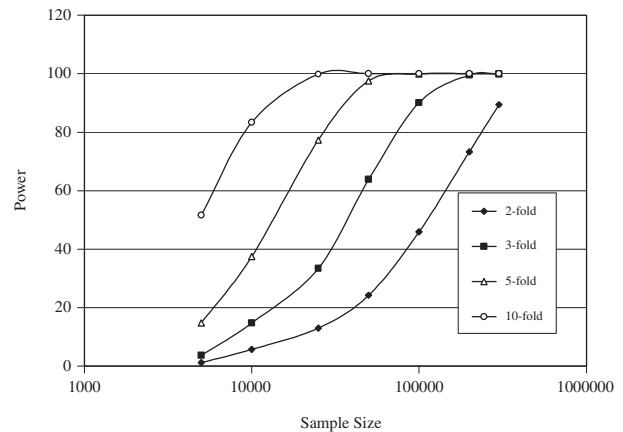


Fig. 4. Simulation results from POWER_SAGE. The tag frequency of the transcript in the control sample is set at 0.01% (p^c) and that in the treatment sample is set at fold* p^c . The sample size (5000, 25 000, 50 000, 100 000, 200 000, or 300 000) represents the total number of tags from each sample. The total number of iteration used is 10 000 and the significance level is 5%.

detection of a lower degree of sensory discrimination, or, in other words, of a quantitatively smaller departure from the null hypothesis' (Fisher, 1951). Fisher's point can be illustrated with simulation results in Figure 4. For large differences (5- and 10-fold), small sample size (total tags collected from one sample) is quite sufficient

to achieve high power. For small differences (2- and 3-fold), much larger sample size is needed to get the same level of power. In practice, if too few tags were sequenced, there would be insufficient power to detect changes in gene expression. Conversely, if more than necessary were sequenced, valuable resources would be wasted. POWER_SAGE can be used to run 'virtual' SAGE experiments to get a 'feel' of the proposed study (Figure 1). An example is also provided (Table 1) in which we assume that the genes of interest are expressed at 0.1% level. One wants to determine the number of tags to sequence in order to have sufficient power (e.g. $\geq 80\%$) to detect a 2-fold change (p in control, $2p$ in treatment). Instead of sequencing 50 000 tags by default, one can do a series of 'virtual' SAGE with different sample sizes. From Table 1, a sample size of 25 000 tags, half the usual size (50 000), will guarantee sufficient power and, hence, results in 50% cost reduction. On the other hand, when the expression level is 0.01%, a sample size of 25 000 tags has little power (11.3%). One needs to look beyond 50 000 tags even if a modest power is desired. Therefore, POWER_SAGE can be useful in helping researchers select the optimal number of tags to sequence, given some prior knowledge about expression levels of the genes of interest. POWER_SAGE is quite flexible, capable of accommodating a large number of transcripts and different sample size combinations (including unequal sample sizes, Figure 1). Another program that is based on Z -statistic for testing equality of two proportions (= Chi-square test), SAGEstat, can do sample size calculation for pre-specified alpha, power, p^c , and p^t (Kal *et al.*, 1999). The two programs are complementary to each other, focusing on power and sample size, respectively.

Acknowledgements

The authors would like to thank colleagues at PGRD for their support. David Payne provided stimulating discussion and encouragement. Ron Emaus and 'Dak' Atipat Rojnuckarin gave much needed technical support. Brad Evans, Paul Juneau, Bill Dougherty, and Jeffrey Thomas carefully reviewed the paper and gave many insightful suggestions. The authors are also indebted to three anonymous reviewers who provided constructive criticism and suggested several references.

References

- Audic, S. and Claverie, J. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
- Chen, H., Centola, M., Altschul, S.F. and Metzger, H. (1998) Characterization of gene expression in resting and activated mast cells. *J. Exp. Med.*, **188**, 1657–1668.
- Collier, R.O. and Baker, F.B. (1996) Some Monte Carlo results on the power of the F -test under permutation in the simple randomized block design. *Biometrika*, **53**, 199–203.
- Claverie, J.M. (1999) *Hum. Mol. Genet.*, **8**, 1821–1832.
- Elashoff, E.D. (1999) nQuery Advisor[®] Version 3.0 User's Guide, Los Angeles, CA.
- FDA (1998) *International Conference on Harmonisation; Guidance on Statistical Principles for Clinical Trials; Availability*. Federal Register, Vol 63, No. 179.
- Fisher, L.D. and van Belle, G. (1993) *Biostatistics: a Methodology for the Health Sciences*. John Wiley & Sons, New York.
- Fisher, R.A. (1951) *The Design of Experiments*. Hafner, New York.
- Hammersley, J.M. and Handscomb, D.C. (1964) *Monte Carlo Methods*. Wiley, New York.
- Hogg, R.V. and Craig, A.T. (1995) *Introduction to Mathematical Statistics*. Prentice-Hall, Englewood Cliffs, New Jersey.
- Kal, A.J., van Zonneveld, A.J., Benes, V., van den Berg, M., Koerkamp, M.G., Albertmann, K., Strack, N., Ruijter, J.M., Richter, A., Dujon, B., Ansorge, W. and Tabak, H.F. (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast growth on two different carbon sources. *Mol. Biol. Cell*, **10**, 1859–1872.
- Kennedy, J.W., Kaiser, G.W., Fisher, L.D., Fritz, J.K., Myers, W., Mudd, J.G. and Ryan, T.J. (1981) Clinical and angiographic predictors of operative mortality from the collaborative study in coronary artery surgery (CASS). *Circulation*, **63**, 793–802.
- Lal, A., Lash, A.E., Altshul, S.F., Velculescu, V., Zhang, L., McLendon, R.E., Marra, M.A., Prange, C., Morin, P.J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K.W., Strausberg, R.L. and Riggins, G.J. (1999) A public database for gene expression in human cancers. *Cancer Res.*, **59**, 5403–5407.
- Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. and Beaudry, G.A. (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene*, **15**, 1079–1085.
- Manly, B.E. (1991) *Randomization and Monte Carlo Methods in Biology*. Chapman and Hall, London.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in C: the Art of Scientific Computing*. Cambridge University Press.
- Stokes, M.E., Davis, C.S. and Koch, G.G. (1995) *Categorical Data Analysis Using the SAS System*. SAS Institute, Cary, NC.
- Strausberg, R.L., Dahl, C.A. and Klausner, R.D. (1997) New opportunities for uncovering the molecular basis of cancer. *Nature Genet.*, **16** (suppl.), 415–516.
- Sutradhar, B.C. and Bartlett, R.F. (1993) Monte Carlo comparison of Wald's likelihood ratio and Rao's tests. *J. Stat. Comp. Simu.*, **46**, 23–33.
- Tanizaki, H. (1997) Power comparison of non-parametric tests: small-sample properties from Monte Carlo experiments. *J. Appl. Stat.*, **24**, 603–632.
- Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
- Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B. and Kinzler, K.W. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.