Physiol Genomics 11: 37–44, 2002;
10.1152/physiolgenomics.00042.2002.

review

Statistical evaluation of SAGE libraries: consequences for experimental design

JAN M. RUIJTER,[1] ANTOINE H. C. VAN KAMPEN,[2] AND FRANK BAAS[3]
[1]*Department of Anatomy and Embryology,* [2]*Bioinformatics Laboratory, and* [3]*Neurogenetics Laboratory, Academic Medical Center, University of Amsterdam, 1105 AZ, Amsterdam, the Netherlands*

Submitted 12 April 2002; accepted in final form 5 September 2002

Ruijter, Jan M., Antoine H. C. van Kampen, and Frank Baas. Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol Genomics* 11: 37–44, 2002; 10.1152/physiolgenomics.00042.2002.—Since the introduction of serial analysis of gene expression (SAGE) as a method to quantitatively analyze the differential expression of genes, several statistical tests have been published for the pairwise comparison of SAGE libraries. Testing the difference between the number of specific tags found in two SAGE libraries is hampered by the fact that each SAGE library is only one measurement: the necessary information on biological variation or experimental precision is not available. In the currently available tests, a measure of this variance is obtained from simulation or based on the properties of the tag distribution. To help the user of SAGE to decide between these tests, five different pairwise tests have been compared by determining the critical values, that is, the lowest number of tags that, given an observed number of tags in one library, needs to be found in the other library to result in a significant *P* value. The five tests included in this comparison are SAGE300, the tests described by Madden et al. (*Oncogene* 15: 1079–1085, 1997) and by Audic and Claverie (*Genome Res* 7: 986–995, 1997), Fisher's Exact test, and the *Z* test, which is equivalent to the chi-squared test. The comparison showed that, for SAGE libraries of equal as well as different size, SAGE300, Fisher's Exact test, *Z* test, and the Audic and Claverie test have critical values within 1.5% of each other. This indicates that these four tests will give essentially the same results when applied to SAGE libraries. The Madden test, which can only be used for libraries of similar size, is, with 25% higher critical values, more conservative, probably because the variance measure in its test statistic is not appropriate for hypothesis testing. The consequences for the choice of SAGE library sizes are discussed.

critical values; hypothesis test; two-sided test; library size; power; serial analysis of gene expression

SERIAL ANALYSIS OF GENE EXPRESSION (SAGE; 17) was introduced as a method to quantitatively analyze the differential expression of genes. The method has since been applied successfully to cells and tissues obtained from different developmental stages or resulting from a variety of pathological processes. The SAGE procedure results in a library of short tags, each representing an expressed gene. The main assumption in the interpretation of the data in this library is that every mRNA copy in the tissue has the same chance of ending up as a tag in the library. This selection of a specific tag sequence from the total pool of transcripts can be well approximated as sampling with replacement (15).

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: J. M. Ruijter, Dept. of Anatomy and Embryology, Academic Medical Center, Meibergdreef 15, K2-283, 1105 AZ Amsterdam, the Netherlands (E-mail: j.m.ruijter@amc.uva.nl).

The aim of most SAGE studies is to identify genes of interest by comparing the number of specific tags found in two different SAGE libraries. In statistical terms, the aim is to reject the null hypothesis that the observed tag counts in both libraries are equal. Testing of this hypothesis is hampered by the fact that each SAGE library is only one measurement: the necessary information on biological variation and experimental precision is not available. Therefore, each of the published statistical tests for comparing SAGE libraries is based on its own assumptions about the statistical distribution of SAGE tags from which a measure of variance is obtained.

In comparing two SAGE libraries, a large number of pairwise tests, one for each specific tag, is performed. It is possible that most pairwise differences between two libraries are just the result of random sampling from two populations that do not differ. Therefore, before starting a pairwise comparison of specific tags in two libraries, the null hypothesis that the differences between libraries result from such a random sampling has to be rejected. A similar line of reasoning is applied in the comparison of the means of more than two groups: before a multiple comparison of groups can be carried out, an overall analysis of variance has to reject the null hypothesis that all groups originate from the same population (2). In the context of SAGE research, only one reference to such an overall test has been published (14). This overall test is based on a simulation of a large number of possible distributions of two libraries within the pooled marginal totals of the observed SAGE libraries. By calculating a chi-squared statistic for each simulated pair of libraries, a distribution of this statistic under the null hypothesis can be constructed. From this simulated distribution and the chi-squared statistic of the observed libraries, one can determine the probability of obtaining the observed tag distributions by chance. Rejection of the null hypothesis that all differences between SAGE libraries are just the result of random sampling then opens the way for pairwise comparisons.

In the seminal paper of Velculescu et al. (17), tag numbers in different libraries are compared pairwise with a test based on a Monte Carlo simulation of tag counts. This test has been included into the SAGE software package SAGE300 (19). SAGE300 determines for each pairwise comparison of tags the chance of obtaining a difference in tag counts equal to or greater than the observed difference from the number of trials it takes to simulate this difference 100 times. The resulting chance serves as P value in a one-sided test.

In other papers dealing with SAGE, several pairwise test procedures have been proposed. Most of these tests have been incorporated into public database systems and analysis programs (5, 8, 10, 11, 13, 15). The test suggested by Madden et al. (11) is based on only the number of observed specific tags in each SAGE library, and the calculated statistic (Table 1) is compared with the normal distribution. Audic and Claverie (3) derived a new equation (Table 1) for the probability, $P(n_2|n_1)$, of finding n_2 tags in one library given the fact that n_1 tags have already been observed in the other library. The sum $\Sigma P(n_2|n_1)$ of this probability for n_2 or more tags then serves as a one-sided test. The test proposed by Kal and coworkers (7) focuses on the proportions of specific tags in each library. Since these proportions can be approximated to result from sampling with replacement, the probability of the resulting tag counts follows a binomial distribution (15). The proposed test is therefore based on the normal approximation of the binomial distribution (Z test; 7). The test statistic Z is calculated as the observed difference between proportions of specific tags in both libraries divided by the standard error of this difference when the null hypothesis is true (Table 1). This Z statistic is approximately normally distributed and can be compared with the

Table 1. *Reference, test statistic, and decision rule for each of the tests that have been compared in this study*

Reference	Statistic/P value	Decision Rule
Kal et al. 1999 (7) (Z-test)	$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$	reject H_0 when $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$
Madden et al. 1997 (11)	$Z = \frac{n_1 - n_2}{\sqrt{n_1} + \sqrt{n_2}}$	reject H_0 when $Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$
Audic and Claverie 1997 (3)	$P(n_2\|n_1) = \frac{(N_2/N_1)^{n_2}(n_1+n_2)!}{n_1!n_2!(1+N_2/N_1)^{(n_1+n_2+1)}}$	reject H_0 when $\sum_{i=n_2}^{\infty} P(i\|n_1) < \frac{\alpha}{2}$
Zhang et al. 1997 (19) (SAGE300)	P chance from Monte Carlo simulation	reject H_0 when $P < \alpha/2$
Fisher's Exact test	$P(n_1,n_2) = \binom{N_1}{n_1}\binom{N_2}{n_2} \Big/ \binom{N_1+N_2}{n_1+n_2}$	reject H_0 when $\sum_{i=0}^{n_1} P(i, n_1+n_2-i) < \frac{\alpha}{2}$

Since in serial analysis of gene expression (SAGE) experiments no a priori knowledge about the direction of the effects is available, all decision rules are formulated for a two-sided test of the null hypothesis (H_0). Total sample sizes of the two SAGE libraries are designated N_1 and N_2, the number of specific tags observed in these libraries are designated n_1 and n_2. The proportions of specific tags used in the Z-test (Kal et al. 1999, Ref. 7) are calculated as $p_1 = n_1/N_1$ and $p_2 = n_2/N_2$. The proportion p_0, the expected proportion when the null hypothesis is true, is calculated as $p_0 = (n_1 + n_2)/(N_1 + N_2)$. Note that the probability $P(n_2|n_1)$ given by Audic and Claverie's test has to be summed over i from to n_2 to infinity to give a one-sided P value. For this test only the situation where $n_1 < n_2$ is considered.

critical Z value for the two-sided significance level α (2).

The Fisher's Exact test has been proposed by the Cancer Genome Anatomy Project for comparison of specific tags between SAGE libraries (3, 12). Also, the chi-squared test has been used (14). Both tests are based on reorganizing the data per tag in a 2×2 contingency table (rows: specific and other tags; columns: *library 1* and *2*). Fisher's Exact test calculates the pooled probability of obtaining the observed table and all tables with a more extreme difference within the row and column totals (2). The use of Fisher's Exact test is controversial because the requirement that the row and column totals must be fixed and known prior to the experiment (5, 6) leads to conservative P values (1). However, the test can be used in situations such as SAGE, where the marginal totals are not naturally fixed, because the use of marginal totals is ancillary and does not lead to loss of information on the null hypothesis (1). Therefore, the Fisher's Exact test is included in the comparison of tests. For the large numbers of tags involved in SAGE, the chi-squared test is the preferred test for 2×2 contingency tables (2, 6). The chi-squared test is, however, not included in the current comparison because a chi-squared test on 2×2 tables is statistically equivalent to the Z test on two proportions (2, 12) and, therefore, gives exactly the same results as the Z test. Two approaches based on Bayesian statistics (4, 10) have been described to calculate the probability that the level of expression of a given mRNA is increased by at least x-fold between libraries. Although these procedures can be used to statistically judge differences in tag numbers, their approach is clearly different from the classic approach of hypothesis testing, and results of both test procedures cannot be directly compared.

Recently the chi-squared test, Fisher's Exact test, and the Audic and Claverie test were compared with respect to their power and robustness (12). The Madden test and SAGE300 were not included in this comparison, nor was there a comparison of the differences that are needed to lead to a statistically significant result. The latter hampers the comparison of test results in different papers. Therefore, and to further help the user of SAGE to decide between the available tests, the present review compares the critical values of five tests (excluding the chi-squared test). Critical values, sometimes called "first significant values" (3), are defined as the highest or lowest number of tags that, given an observed number of tags in one library, needs to be found in the other library to result in a P value below the significance level when the pairwise test is carried out.

Table 1 lists the five tests for pairwise testing of SAGE libraries that have been compared. It also gives the test statistic and the decision rule of each test. For details on the statistical basis of each of these tests, the reader is referred to the original papers. For all tests the null hypothesis (H_0) is that there is no difference in tag numbers between the two libraries. The five tests were compared by determining their critical values for a significance level of 0.001. Such a low significance level was chosen to safeguard against accumulation of type I error. The use of a significance level of 0.001 is equivalent to an overall significance level of 0.05 and a Bonferroni correction to allow for 50 hypothesis tests (2).

In this review only the upper critical values are considered. Critical values were determined by taking a fixed tag count in the first library and subsequently performing the statistical test for an increasing number of tags in the second library until the resulting P value leads to rejection of the null hypothesis at the required level of significance. Since the Monte Carlo-based test of SAGE300 does not give the same P value every time the same input is tested, for each input the test was run six times and the mean P value was used. Such an average P value based on three trials is also given by SAGE300 in its "analyze"-"entire project" option.

All critical values were determined for *1*) a total number of 10,000 tags in both SAGE libraries ($N_1 = N_2 = 10{,}000$) and *2*) a total of 10,000 tags in the first and 50,000 tags in the second library ($N_1 = 10{,}000$; $N_2 = 50{,}000$). The values for the number of specific tags observed in the first library (n_1) ranged from 1 to 100, effectively testing an abundance range of 0.0001 to 0.01. The critical values are the number of specific tags that have to be found in the second library (n_2) and are determined by systematic simulation of an increasing difference between the two libraries. It should be kept in mind that in most comparisons between specific tags in SAGE libraries, there is no a priori knowledge about the direction of the effect. Therefore, all pairwise tests have to be carried out as a two-sided test. To do this, the test statistic Z (7, 11) was compared with $Z_{\alpha/2}$, whereas the one-sided P values of SAGE300 as well as the integrated probabilities of the Audic and Claverie test and of the Fisher's Exact test were compared with $\alpha/2$ (Table 1).

The upper critical values for a 0.001 level of significance for the Z test of Kal et al. (7) are given in Fig. 1 for two SAGE libraries of equal size (Fig. 1*A*; both 10,000 tags) and for two SAGE libraries of different size (Fig. 1*B*; 10,000 and 50,000 tags, respectively) as continuous lines. Note that for a larger SAGE library the confidence level of an observed tag count is higher (7). Therefore, with a large second SAGE library, smaller differences in proportions can be detected as statistically significant. For two libraries of the same size (N) and relatively low tag counts ($n_1 + n_2$ less than 1% of $2N$) the test statistic Z of the Z test (Table 1) reduces to $Z = (n_1 - n_2)/\sqrt{(n_1 + n_2)}$. Thus, for low tag counts and two large libraries of the same size, the critical values of the Z test are independent of library size.

The critical values for the test of Madden et al. (11) for two libraries of the same size are plotted in Fig. 1*A*. Compared with the critical values of the Z test, the Madden test requires about 25% bigger differences to reach statistical significance and is, therefore, more conservative. Although the simple mathematics of this test (Table 1) make it very easy to use, its usefulness is

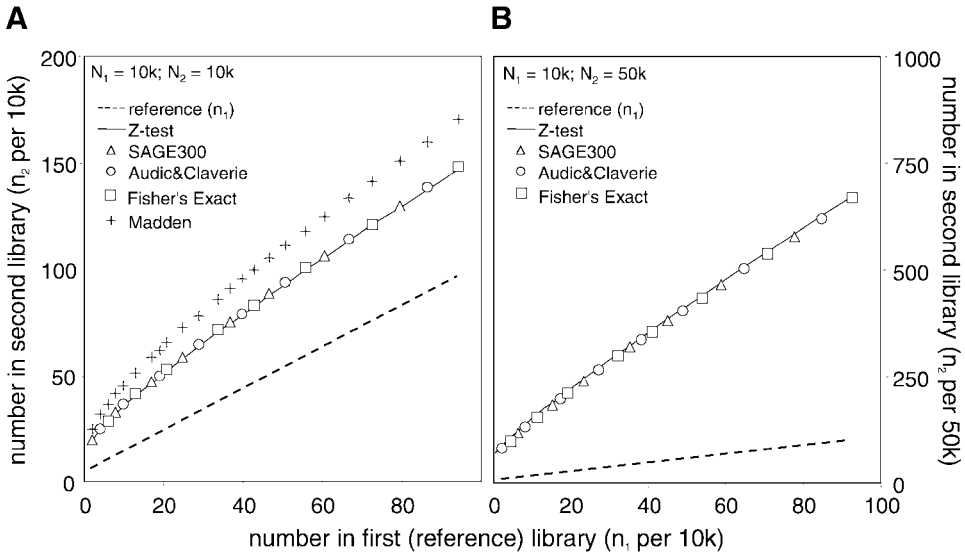

Fig. 1. Comparison of critical values of five tests for the comparison of SAGE libraries. Critical values are defined as the number of tags that needs to be found in the second SAGE library to be significantly different from the number of tags already found in the first SAGE library. Upper critical values for a 0.001 level of significance for the Z test (7), Fisher's Exact test (2), SAGE 300 (16), and the tests of Madden et al. (11) and Audic and Claverie (3). The critical values plotted in each graph are based on a first SAGE library with a total of 10,000 tags (reference values plotted as dotted lines) and a second library with a total of 10,000 tags (*A*; critical values plotted on the *left* y-axis) or a second library of 50,000 tags (*B*; critical values plotted on the *right* y-axis). In both graphs the continuous line represents the critical values of the Z test. The Madden test is only compared for a second library of 10,000 tags, because this test can only be used for libraries of similar size. Note that the number of tags in the first library starts at 1 tag per 10,000.

limited by the fact that it does not include the total number of tags in the calculations and it can, therefore, only be used for SAGE libraries of the same, or very similar, size. The origin of the test statistic of this test is not given in the original paper (11), but when one considers tag counts to fit a Poisson distribution, the variance of a tag count can be estimated to be equal to this tag count (2). The denominator of the test statistic of the Madden test (Table 1) then contains the sum of the standard deviations of the tag counts n_1 and n_2. Statistics as applied by Madden effectively test the hypothesis that the difference in tag counts is zero. Therefore, one can argue that a denominator containing the standard deviation of this difference, that is, the square root of the sum of the tag counts, might be more appropriate. Note that this results in the same equation as is derived in the previous section from the test statistic of the Z test. For large libraries of very similar size it gives the same critical values as the Z test.

The test of Audic and Claverie (3), the Fisher's Exact test, and SAGE300 (19) all have critical values that are on average within 1.5% of those of the Z test, for libraries of equal size (Fig. 1*A*). This equivalence of these four tests holds for tag counts as low as 1 tag per 10,000 in the first library. Only for libraries of different size and low specific tag counts, the Z test needs slightly higher critical values (Fig. 1*B*). Also, for other levels of significance, the critical values of the Z test are almost the same as those published for the Audic and Claverie test (3). This comparison of tests shows that, apart from the test of Madden et al. (11), all tests perform with similar resolution in detecting differences between SAGE libraries. Also, except for Madden et al., all tests can handle SAGE libraries of equal as well as different size. Therefore, the tests published by Kal et al. (7), Audic and Claverie (3), and Zhang et al.

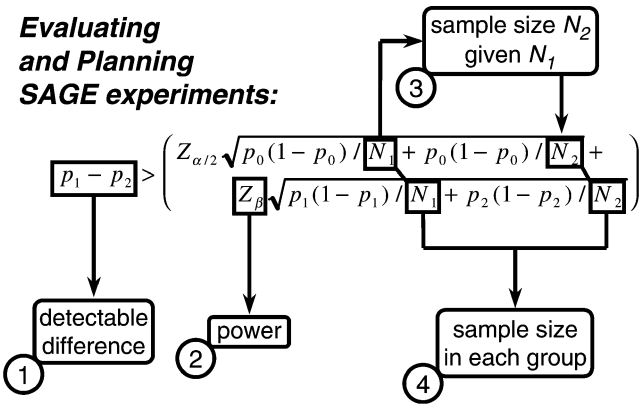

Fig. 2. Equation based on the normal approximation of the binomial distribution and its use for the planning and evaluation of SAGE experiments (7). The *left* side of the equation shows the observed or expected difference between proportions ($p_1 - p_2$). The *right* side includes the Z values for the significance level and the power ($Z_{\alpha/2}$ and Z_{β}, respectively) (Note: power = $1 - \beta$). The two square roots are the standard error of the difference between proportions under the null hypothesis and under the alternative hypothesis, respectively. In the equation, the proportion p_1 stands for the number of specific tags divided by the total number of tags in the first library (n_1/N_1). Under the null hypothesis, when $p_1 = p_2$, this proportion is indicated as p_0 and calculated as $p_0 = (n_1 + n_2)/(N_1 + N_2)$. The encircled numbers refer to the text.

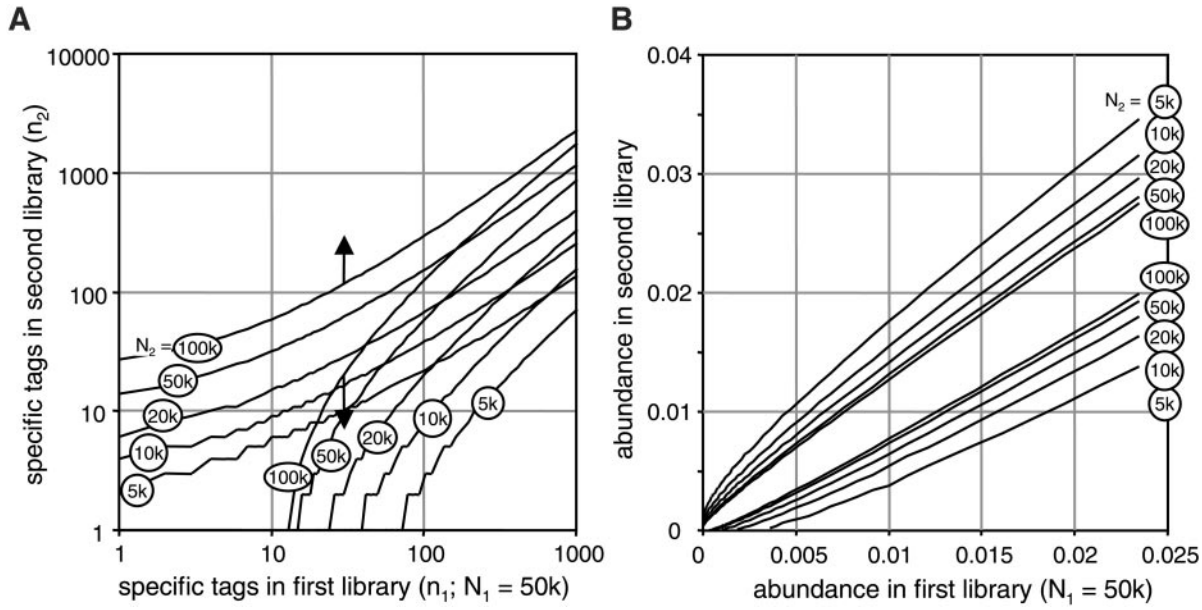

Fig. 3. *A*: evaluation of SAGE experiments: Nomogram of critical values for the statistical comparison of a first library of 50,000 tags with libraries of 5,000 to 100,000 tags. To be statistically significant at $\alpha = 0.001$, the observed number of tags in the second library should be above or below the appropriate pair of lines. The arrows indicate that to be statistically different from 30 tags per 50,000, below 20 or above 117 specific tags have to be found in a library of 100,000 tags. *B*: planning of SAGE experiments. Relation between library size and the detectable difference between two SAGE libraries. The abundance of the transcript in the first (reference) library of 50,000 tags is plotted on the *x*-axis. The abundance that has to be present in the second population to be found as statistically significant at $\alpha = 0.001$ and power = 0.9 is plotted on the *y*-axis and is given for second libraries ranging in size from 5,000 to 100,000 tags (different lines). Note that the gain in resolution decreases when the library size increases.

(19), as well as the Fisher's Exact test, will all give the same test results when applied for pairwise comparison of SAGE libraries.

In addition, a recent paper by Man and coworkers (12) compared the chi-squared test, the test of Audic and Claverie (3), and the Fisher's Exact test. This comparison was based on Monte Carlo simulations of SAGE libraries. The specificity, power, and robustness of the tests were determined for simulated SAGE libraries of various size and at severalfold difference. This comparison showed that the chi-squared test has consistently a higher power and is more robust than the other tests, especially at low expression levels (<15 tags/50,000). Therefore, the chi-squared test, which is equivalent to the *Z* test, was concluded to be the preferred choice for evaluating SAGE experiments (12).

The normal approximation of the binomial distribution that forms the basis of the *Z* test can also be used

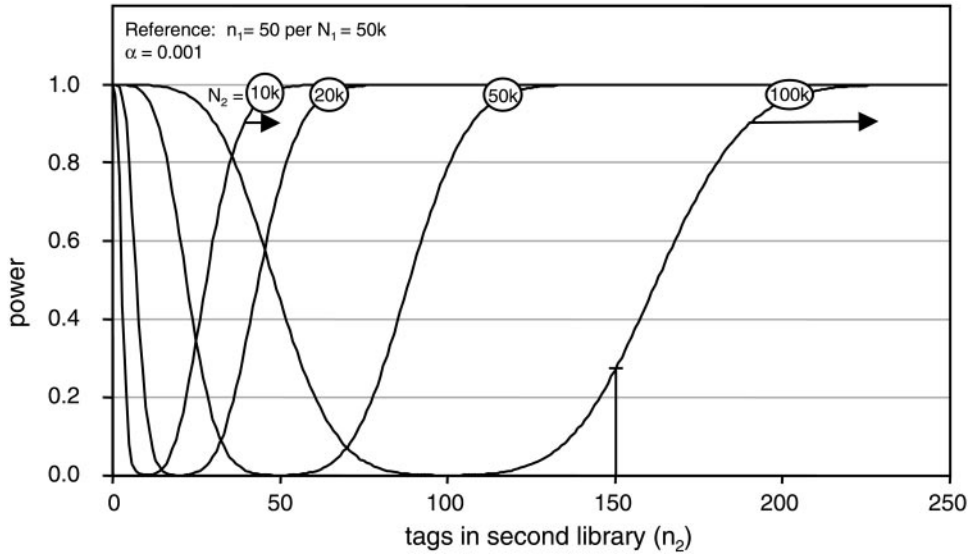

Fig. 4. Relation of library size and power of the *Z* test. The power of the statistical comparison of two libraries was calculated for the occurrence of 50 specific tags per 50,000 (abundance 0.001) in the first library and increasing number of specific tags in second libraries ranging from 10,000 to 100,000 tags. Note that the power is low when the difference in abundance is low. The number of tags that is required to reach a power of 0.9 can be found by tracing the power curve to the value 0.9 and then dropping a line to the *x*-axis. The arrows indicate that for a power higher than 0.9 in a comparison with the reference condition (50 tags per 50,000), at least 190 tags have to be found in a library of 100,000 tags, whereas for the same power in the comparison with a library of 10,000 tags, at least 40 specific tags are required.

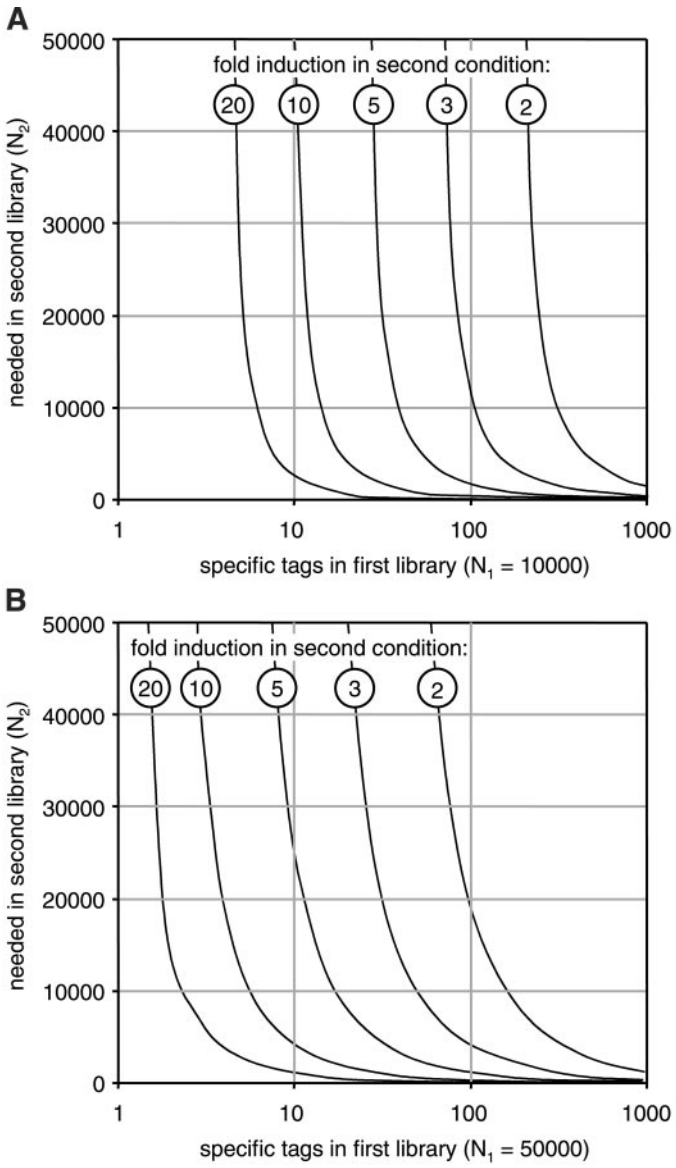

Fig. 5. Number of tags that need to be sequenced in the second library (N_2) to detect a 2- to 20-fold difference as significant (at α = 0.001 and power = 0.9) when a first library of 10,000 (*A*) or 50,000 (*B*) tags is already sequenced. To allow direct comparison, in both graphs the abundance of the specific gene in the first population is taken as the *x*-axis values per 50,000 transcripts. The almost vertical lines at the *top* of *A* indicate that for low abundances in the first library it is not possible to sequence enough tags in the second library to reach a statistically significant difference. The standard error associated with the proportion in the first library is already too large to ever reach the significance level. The shift of all lines to the *left* of the graph in *B* illustrates that with a higher number of tags in the first library, the same fold difference can be detected for a less abundant transcript.

to easily construct confidence intervals for the observed proportion of specific tags as well as for the difference in proportions between two SAGE libraries (7). This approximation also enables the determination of the statistical power of the comparison of two SAGE libraries and the calculation of the sample size needed to detect an expected difference, both of which are essential in the planning of future SAGE analyses. A similar decision about sample size can be reached with a Monte Carlo-based program that calculates the power of a test for a given difference and sample size (POWER_SAGE; 12). Figure 2 shows a rearrangement of the equation of the *Z* test in such a way that it can be used for the evaluation and planning of SAGE experiments. In this form this equation can be used in several ways.

1) Given N_1 and N_2 (the SAGE libraries are compiled), the critical values (Fig. 3*A*) or the detectable differences (Fig. 3*B*) can be calculated for a chosen significance level (α) and power ($1 - \beta$).

2) Given an observed difference, the total number of tags sequenced in both libraries, and the chosen significance level, the power of the test can be determined (Fig. 4).

3) Given an expected difference, a significance level, a power, and the number of tags already sequenced in an existing SAGE library (N_1), the number of tags that is needed in a new library (N_2) can be calculated (Fig. 5).

4) Given an expected difference, a chosen significance level, and a required power, the number of tags that is needed in each library ($N_1 = N_2$) can be calculated (Fig. 6).

The nomogram in Fig. 3*A* can be used to quickly evaluate the differences between two SAGE libraries from the laboratory or the literature. To be statistically significant (at $\alpha = 0.001$) from the first library, the number of specific tags found in the second library should be above or below the appropriate pair of lines. For example, when 30 specific tags are found in the first library of 50,000 tags, a second library of 100,000 tags should yield below 20 or above 117 specific tags to be significantly repressed or upregulated, respectively. The graph of the detectable differences (Fig. 3*B*) can

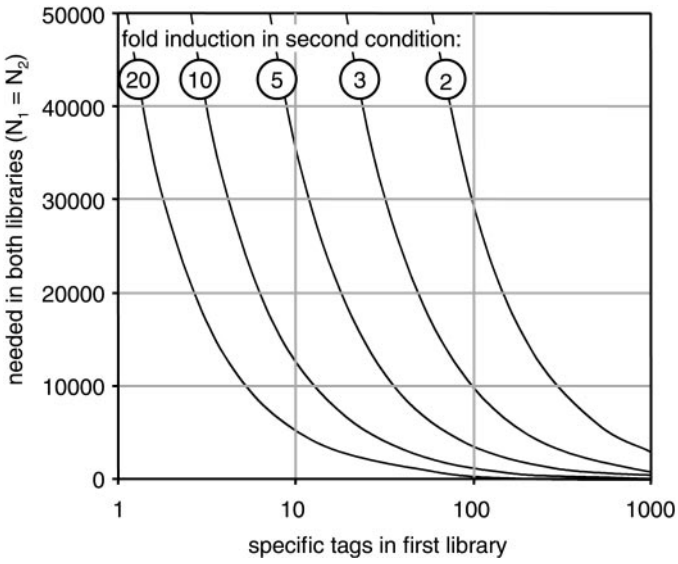

Fig. 6. Number of tags that need to be sequenced in each of the libraries to detect a 2- to 20-fold difference in abundance at a significance level of 0.001 and a power of 0.9. Comparison with the required tag numbers in Fig. 5 shows that the total number of tags that need to be sequenced is always lower when both libraries have the same size.

help the reader to plan a SAGE experiment. Consider a SAGE user who has already (from the laboratory or the literature) the information on the tag counts in a library of 50,000 tags (N_1) and plans the assembly of a second library. It can be seen from Fig. 3*B* that with increasing number of tags in this new library (N_2), smaller differences can be detected as significant. However, whereas upgrading N_2 from 20,000 to 50,000 tags still gives a substantial increase in resolution, the sequencing of another 50,000 tags (N_2 = 100,000) does not seem to pay off statistically. However, the chances of picking up very low abundant transcripts keep increasing with library size.

The power of a performed test tells the user how big the chance is that a real difference has been overlooked, or, in statistical terms, that a false null hypothesis is not rejected. The effect of the differences between libraries on the power of the statistical comparison of these libraries is illustrated in Fig. 4. Figure 4 shows this power as a function of the difference between a first library with 50 specific tags per 50,000 tags and second libraries of various sizes and with different numbers of specific tags. Note that the power is at its lowest when the differences in abundance are low. From this graph it can be read that when the abundance increases 1.5 times, the maximum power of the significance test will only be about 0.25: even when a second library of 100,000 tags is generated, a real 1.5-fold increase would be missed 75% of the time. To reach an acceptable power of 0.9, at least 190 tags per 100,000 should be observed. A smaller library requires relatively larger differences: at least 40 specific tags have to be observed in a library of 10,000 tags to reach the same power.

Instead of looking at detectable differences and power, one can also calculate the number of tags (N_2) needed to detect a 2- to 20-fold difference between the new library and a library known from previous work or the literature (6). The number of tags needed to observe an *x*-fold difference as significant increases exponentially with decreasing abundance of the transcripts in the first library (Fig. 5, *A* and *B*, *x*-axis) and with decreasing difference between conditions (Fig. 5, separate lines) making the detection of small differences for low abundant transcripts a practical impossibility. When the number of tags in the first library is low, differences for the low abundant transcripts may never be detectable. Because the standard error of a proportion is a function of the proportion and the library size [$SE = \sqrt{(p(1-p)/N)}$; Ref. 2] a small difference may never exceed the critical value. In such a case one also has to increase the size of the first library. A comparison of Fig. 5 with Fig. 6 shows that, when no prior knowledge on transcript abundance is available, the most efficient way to set up a SAGE study is to compile two SAGE libraries of equal size. For example, detecting a 10-fold difference for a gene that occurs 10 times in a library of 10,000 tags would take a second library of at least 50,000 tags (Fig. 5*A*), whereas two new libraries of both 14,000 tags would be sufficient (Fig. 6).

Other tests for pairwise comparison of SAGE libraries may be proposed in the future. The usefulness of such tests will be limited by the fact that each SAGE library, no matter how large, only represents one experimental measurement. Consequently, one has no information about the biological variation and the precision of the observed tag counts. Such a measure of experimental variance is crucial for hypothesis testing. In the currently available tests, this measure of variance is obtained from simulation (19) or based on the putative properties of the tag distribution (3, 7, 11). The test results will be dependent on the validity of these assumptions. However, the above comparison shows that the test results of SAGE300, Fisher's Exact test, the *Z* test, and the Audic and Claverie test differ only marginally. Additional tests will, therefore, only be a significant addition to SAGE statistics when these issues of experimental variance and accuracy are addressed. Probably the modeling of the sampling error, sequencing error, and other aspects of SAGE experiments (15) may play a role in the development of such hypothesis tests and the calculation of more accurate *P* values.

When only *P* values are published, it should be noted that SAGE300 and the Audic and Claverie test, as well as the conversion from the *Z* statistic to a *P* value for the Kal test and the Madden test, will result in a one-sided *P* value. The authors should be aware of this and should mention whether a one-sided or a two-sided *P* value is tabulated (see, for instance, Ref. 8). However, since in SAGE experiments no a priori knowledge about the direction of the effects is available, the publication of two-sided *P* values would be the most appropriate and should be encouraged. This would enable the direct comparison of published *P* values with the required level of significance and simplify the comparison of different papers on the same tissues. However, the significance of the *P* value of the observed difference between tag counts should not be overemphasized: the rank order of the *P* values may well be all the information the reader needs to pinpoint important genes and to plan future research.

We thank Drs. Arnoud Kal, Henk Tabak, and Patrick Bossuyt for help in locating the different statistical tests and Drs. Wout Lamers and Antoon Moorman for critical comments on the manuscript.

SAGE300 is available from **http://www.sagenet.org**. The test of Audic and Claverie (3) is available from **http://igs-server.cnrs-mrs.fr/~audic/significance.html**. SAGEstat, for the application of the *Z* test (7) as well as the calculation of critical values and the number of tags needed to detect an assumed difference, is available on request (E-mail: bioinfo@amc.uva.nl; subject, SAGEstat). An R (S-plus) implementation of SAGEstat, with the possibility to compare public domain SAGE libraries and to plot graphs of the required number of SAGE tags is incorporated in USAGE (16), which can be reached at **http://www.cmbi.kun.nl/usage/**. Another program that will calculate the number of required tags and perform a chi-squared test between SAGE libraries is POWER_SAGE (E-mail: michael.man@pfizer.com; Ref. 12), which is based on Monte Carlo simulations.

REFERENCES

1. **Agresti A.** A survey of exact inference for contingency tables. *Stat Sci* 7: 131–177, 1992.

2. **Altman DG.** *Practical Statistics for Medical Research*. London: Chapman-Hall, 1991, p. 161–167 and 253–258.
3. **Audic S and Claverie JM.** The significance of digital gene expression profiles. *Genome Res* 7: 986–995, 1997.
4. **Chen H, Centola M, Altschul SF, and Metzger H.** Characterization of gene expression in resting and activated mast cells. *J Exp Med* 188: 1657–1668, 1998.
5. **Claverie JM.** Computational methods for the identification of differential and coordinated gene expression. *Hum Mol Genet* 8: 1821–1832, 1999.
6. **Conover WJ.** *Practical Nonparametric Statistics*. New York: Wiley, 1980, p. 162–167.
7. **Kal AJ, Van Zonneveld AJ, Benes V, Van den Berg M, Groot Koerkamp M, Albermann K, Strack N, Ruijter JM, Richter A, Dujon B, Ansorge W, and Tabak HF.** Dynamics of gene expression revealed by comparison of SAGE transcript profiles from yeast grown on two different carbon sources. *Mol Biol Cell* 10: 1859–1872, 1999.
8. **Kenzelmann M and Mühlemann K.** Transcriptome analysis of fibroblast cells immediate-early after human cytomegalovirus infection. *J Mol Biol* 304: 741–751, 2000.
9. **Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, Papadopoulos N, Vogelstein B, Kinzler KW, Strausberg RL, and Riggins GJ.** A public database for gene expression in human cancers. *Cancer Res* 59: 5403–5407, 1999.
10. **Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, and Altschul SF.** SAGEmap: a public gene expression resource. *Genome Res* 10: 1051–1060, 2000.
11. **Madden SL, Galella EA, Zhu JS, Bertelsen AH, and Beaudry GA.** Sage transcript profiles for P53-dependent growth regulation. *Oncogene* 15: 1079–1085, 1997.
12. **Man MZ, Wang X, and Wang Y.** POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* 16: 953–959, 2000.
13. **Margulies EH and Innis JW.** eSAGE: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics* 16: 650–651, 2000.
14. **Michiels EMC, Oussoren E, van Groenigen M, Pauws E, Bossuyt PMM, Voûte PA, and Baas F.** Genes differentially expressed in medulloblastoma and fetal brain. *Physiol Genomics* 1: 83–91, 1999.
15. **Stollberg J, Urschitz J, Urban Z, and Boyd CD.** A quantitative evaluation of SAGE. *Genome Res* 10: 1241–1248, 2000.
16. **Van Kampen AHC, Van Schaik BDC, Pauws E, Michiels E, Ruijter JM, Caron HN, Versteeg R, Heisterkamp SH, Leunissen JAM, Baas F, and Van der Mee M.** USAGE: a web-based approach towards the analysis of SAGE data. *Bioinformatics* 16: 899–905, 2000.
17. **Velculescu VE, Zhang L, Vogelstein B, and Kinzler KW.** Serial analysis of gene expression. *Science* 270: 484–487, 1995.
18. **Vingron M and Hoheisel J.** Computational aspects of expression data. *J Mol Med* 77: 3–7, 1999.
19. **Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, and Kinzler KW.** Gene expression profiles in normal and cancer cells. *Science* 276: 1268–1272, 1997.