

# Chapter 8

## Statistical analysis of transcript counts.

<sup>1</sup>Teacher / <sup>1,2</sup>Authors: Jan M Ruijter <sup>1</sup> and Gerben J Schaaf <sup>2</sup>

<sup>1</sup>Heart Failure Research Center, Dept. Anatomy & Embryology  
Academic Medical Center, Amsterdam

and

<sup>2</sup> Dept. of Cell Biology  
Erasmus MC, Rotterdam

j.m.ruijter@amc.uva.nl

*“The characteristics of an organism are determined by the genes expressed within it. A method was developed, ....., that allows the quantitative and simultaneous analysis of a large number of transcripts”.*

*(Velculescu et al. 1995)*

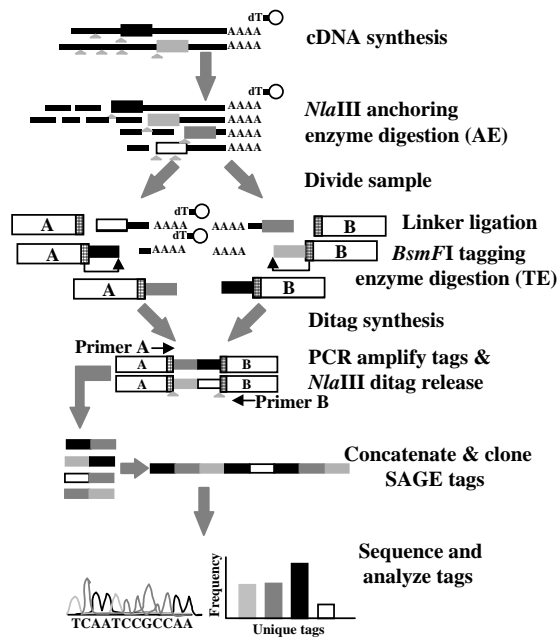
### 8.1. Transcript counts and analysis of gene Expression.

The frequency of a transcript, representing the steady state level of a mRNA, provides, within certain limitations, information about the level of gene expression and the amount of protein made. The total collection of transcript frequencies, determined for a certain tissue is dubbed a “library”. The information contained in such a library reflects the transcript abundances, “the transcriptome”, of that tissue. Comparison of transcriptomes yields information about the dynamics of total genome expression due to a change in environmental condition, state of development, differentiation or pathology (Ruan et al. 2004). In addition, it provides clues to determine the function of those genes whose contribution to these cellular processes is still unknown.

The SAGE technique (Fig. 1; Velculescu et al. 1995) samples short sequences of 10-14 nucleotides (called “tags”) from the most 3’ *NotI* restriction (or CATG) site of individual mRNAs. The sequence of these tags, together with the positional information, allows identification of the corresponding genes. The specificity of this identification has been improved by the introduction of LongSAGE, which uses 21 nucleotide-long tags (Saha et al. 2002).

Massive parallel DNA sequencing is an application of ‘next-generation sequencing’ that also results in transcript counts. The method is based on two concepts. Firstly, (part of) each individual transcript is PCR amplified in such a way that the amplification

product stays spatially clustered. Secondly, these clusters of PCR product are sequenced (Diverse authors, 2008).



**Figure 1.** Schematic outline of the SAGE protocol. All poly(A) species shown are double-stranded cDNA molecules after mRNA reverse transcription, which are bound to oligo(dT)-coated beads (circles). The anchoring enzyme (AE) exposes the 3'-most *Nla*III restriction sites (inverted arrowheads) of the bound cDNAs (shaded bars represent SAGE tags). Tagging enzyme (TE) *Bsm*FI cuts 10-bp 3' (indicated by the rectangular arrow) from its recognition site (chequered bar) in linkers A and B. The resulting linker-tag combinations are ligated to ditags and PCR-amplified with primers directed to the linkers. After removal of the linkers, the ditags are concatenated and cloned. Sequencing then reveals the identity of the tags. In the bar graph, the frequency of each of the sequenced tags is plotted. (reproduced from: Patino et al. 2002)

Increasing numbers of libraries with increasing numbers of unique tags become available. Whereas with the near completion of the human genome sequencing the number of genes is predicted to be between 30 000 and 40 000 (Lander et al. 2001; Venter et al. 2001), already over 400 000 unique SAGE tags are observed (Chen et al. 2002; Boheler and Stern 2003). On the basis of theoretical considerations and 'known' error rates, Akmaev and Wang (2004) calculated that about 17% of the LongSAGE tags contains at least one nucleotide mutation introduced by PCR or sequencing errors. On the other hand, Chen and coworkers (2002) already gave several arguments that even low abundant tags (count of 1 per library) may represent real and even novel transcripts. In contrast to common believe, they show that single nucleotide sequencing errors cannot be solely responsible for these singleton tags because 1) the frequency of such errors is much lower (only 2%) than assumed and 2) the number of unique singleton tags does not increase with increasing total tag numbers as it should when they were all just random errors. Therefore, the increasing numbers of, even very low abundant, novel tags may foreshadow an era of discovery of novel expressed genes (Patino et al. 2002).

In this sense, SAGE combines hypothesis-driven research ("Is my gene of interest affected by this treatment?") with discovery-driven research ("What genes are affected by this treatment?") although it has recently been predicted that such a role in genomic

research would be taken over by DNA arrays (Constans 2003). However, massive parallel sequencing has started to supplant micro-arrays (Editorial, 2008). Because the analysis methods described in this reader were developed for SAGE, the following text will refer to SAGE libraries. When libraries generated by massive parallel sequencing fulfill the basic assumption described in the section 8.2, the statistics are also applicable to those libraries.

## 8.2. Basic assumption in transcript count analysis.

The number of specific tags in a library per total number of tags in that library is assumed to be an estimate of the number of copies of a specific mRNA per cell as a fraction or *proportion* of the total number of mRNA transcripts in that cell. Thus, the proportion ( $p$ ) of a certain mRNA species is given by:

$$p = \frac{n_{\text{specific mRNA/cell}}}{N_{\text{total mRNA/cell}}} = \frac{n_{\text{specific tags}}}{N_{\text{total tags}}} \quad \text{Eq. 1}$$

In this approach the SAGE procedure can be described as sampling specific tags from a large population of tags: like picking a handful of colored marbles from a bucket full of marbles. Only this time the handful is a sample of 10 000 to 100 000 tags. Because of the amplification steps in the procedure, even such large samples can be assumed to be drawn from an infinitely large population (effectively: sampling with replacement). Because of this sampling, the results represent absolute expression levels, transcript count per total number of transcripts, which can be directly compared with any existing SAGE library (Velculescu et al. 2000).

## 8.3. All differences may be random error.

The aim of the statistical comparison of two SAGE libraries is to reject the null hypothesis that the observed tag counts in those libraries are equal. Testing of this hypothesis is hampered by the fact that a SAGE library is only ONE measurement. Although it is a large sample, each library is still only one sample: the necessary information on biological variation and experimental precision is not available in the data. So, it is possible that all differences between two libraries are just the result of random sampling from the same population. Therefore, before starting a pair-wise comparison of specific tags in two libraries, the null hypothesis that the differences between libraries result from random sampling from one population has to be rejected. In the context of SAGE research, only one reference to a test for this purpose has been published (Michiels et al. 1999). 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 the Chi-squared statistic for each simulated pair of libraries, a distribution of this statistic under the null hypothesis can be constructed. From this distribution and the Chi-squared value of the observed libraries, one can then determine the probability of obtaining the observed tag distributions at random. Rejection of the null hypothesis that all differences between SAGE libraries are just the result of random sampling then opens the way for pair-wise comparisons between the two libraries.

#### 8.4. Statistical analysis of tags in two libraries: Z-test.

The SAGE procedure can be described as taking a sample of tags from a large population of tags and counting the number of each specific tag. The number of specific tags is the result of the probability of each tag to be identified as being the specific mRNA or not and, therefore, can be assumed to be binomially distributed. For the large number of tags sequenced in one library this binomial distribution can be very well approximated by a normal distribution with mean  $p$  and standard deviation  $SD_p = \sqrt{p(1-p)}$ . The accuracy of the estimation of the proportion depends on the total number of tags ( $N$ ) sequenced. The standard error of this proportion is given by:

$$SE_p = SD / \sqrt{N}. \quad \text{Eq. 2}$$

The standard error of the difference  $p_1 - p_2$  of proportions  $p_1$  and  $p_2$ , observed in two libraries with sample sizes  $N_1$  and  $N_2$ , respectively, is given by (Altman, 1991):

$$SE_{p_1-p_2} = \sqrt{p_1(1-p_1)/N_1 + p_2(1-p_2)/N_2} \quad \text{Eq. 3}$$

The difference  $p_1 - p_2$  and the expected value of the above standard error when the null hypothesis is true can be used to calculate the test statistic Z (Eq. 4), which is approximately normally distributed and serves as a statistical test for the difference between the proportions  $p_1$  and  $p_2$  (Altman 1981; Kal et al. 1999). Note that because of the dependence of the  $SE_p$  on the total library size, in all SAGE data analysis the original tag counts and original library sizes have to be used. The use of "normalized" libraries will lead to erroneous conclusions.

#### 8.5. Some remarks on choosing a significance level.

The significance level  $\alpha$  of a hypothesis test is the chance that one is willing to accept that a true null hypothesis is rejected (Type I error or false positive). In *discovery-driven* research the choice of  $\alpha$  is not immediately important. The tags that are the most promising for future research are the ones which deviate most from the expected

proportion and these are the ones with the lowest  $P$ -value. So, putting the tags in ascending order of  $P$ -value gives you a list with the most interesting genes at the top. Just work your way down the list and it will be some time before you really need a significance level to decide whether or not it is sensible to continue.

A very different situation occurs when your research is *hypothesis-driven* or diagnostic. Then the choice of  $\alpha$  determines whether or not a tag is labeled as 'differentially expressed'. Which  $\alpha$  has to be chosen depends on the consequences of a false positive decision on the one hand and of a false negative decision on the other. The commonly used  $\alpha=0.05$  which gives a 5% chance that the decision is a false positive, should not be taken as a fixed value. It would not be acceptable when 1 in 20 leg amputations are due to a false positive test result but a lot more than 1 in 20 people take antibiotics without real need. On the other hand, the fear of overlooking a common flu does not justify putting the significance level on 0.1, but when it comes to Ebola one has to stay on the safe side: even half a chance ( $\alpha=0.5$ ) may then be reason enough to raise an alarm and avoid the consequences of a false negative decision.

Note that Type I errors accumulate when more than one hypothesis test is done within one experiment. When  $n$  pair-wise comparisons are done and each decision is based on a significance level  $\alpha$ , then the chance that at least one of the decisions is wrong is accumulated to  $1-(1-\alpha)^n$ , which for 10 comparisons and  $\alpha=0.05$  already increases to over 40 percent. One way to limit this accumulation of Type I error is to apply a Bonferroni correction (Altman, 1991): the significance level to use is calculated as  $\alpha$  divided by the number of comparisons that you plan to do. With the thousands of tests in one SAGE analysis the Bonferroni correction becomes much too stringent and therefore very conservative. Other methods have been proposed to remedy this. Van den Oord and Sullivan (2003) argue that it is better to base the corrected significance level on reducing false discoveries as well as the proportion of true detection and give an equation to this end (Van den Oord and Sullivan, 2003).

A common misconception is that the Type I error occurs when you do the test. When you have 10 tags with the same tags counts in two libraries, you will of course do only 1 test and copy the result for each of the 10 tags. However, when it comes to accumulation of Type I error you have in fact done all 10 tests. This is because the error does not occur during the test but during the sampling: each of the 10 samples may contain a random sampling error and, therefore, turn out to be a false positive.

## 8.6. Comparison of tests for two libraries.

Since the introduction of SAGE several statistical tests have been published for the pair-wise comparison of two SAGE libraries. For all tests the null hypothesis is that there is no difference in tag numbers between the two libraries that are compared. 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 decision rules have to be formulated to result in a two-sided test. The significance level ( $\alpha$ ) can be set to 0.001 to safeguard against accumulation of false positive rate that may result from multiple testing (Bonferroni correction; Altman 1991). For a comparison of the differences that have to be present to reach a statistically significant result critical values can be determined. Critical values 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 pair-wise test is carried out. They can be determined by repeatedly and systematically testing simulated tag numbers until the resulting  $P$ -value leads to rejection of the null hypothesis at the required level of significance.

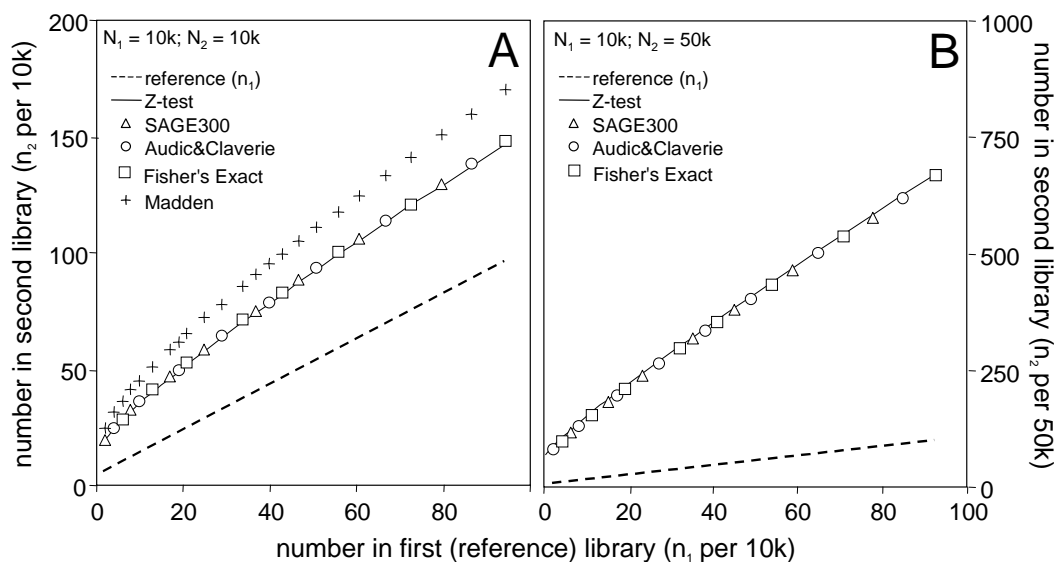
The Z-test proposed by Kal et al. (1999) focuses on the proportions of specific tags in each library and is based on the normal approximation of the binomial distribution (Altman 1991; Kal et al. 1999). As mentioned above, the test statistic  $Z$  is calculated as the difference in proportions divided by the standard error of this difference (Eq. 3) when the null hypothesis is true:

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)/N_1 + p_0(1-p_0)/N_2}} \quad \text{Eq. 4}$$

with  $p_1 = n_1/N_1$  and  $p_2 = n_2/N_2$ . The proportion  $p_0$ , the expected proportion when the null hypothesis is true, is calculated as  $p_0 = (n_1+n_2)/(N_1+N_2)$ .  $Z$  is approximately normally distributed and can be compared to  $Z_{\alpha/2}$ . The critical values of the Z-test are plotted as continuous lines in Fig. 2A and 2B.

In the original paper of Velculescu et al. (1995) tag numbers in different libraries were compared pair-wise with a test based on a Monte Carlo simulation of tag counts. This approach is included into the SAGE software package SAGE300 (Zhang et al. 1997). SAGE300 performs, in each pair-wise comparison, at least 100 with a maximum of 100 000 simulations to determine the chance of obtaining a difference in tag counts equal to or greater than the observed difference. This results in a one-sided  $P$ -value that has to be compared to  $\alpha/2$ . Since the Monte Carlo-based test of SAGE300 does not give the same  $P$ -value every time the same input is tested, each input was run 6

times and the mean  $P$ -value is used for the determination of the upper critical values. These critical values are plotted as triangles in Fig. 2A and 2B. They are all within 1.5% of those of the Z-test.



**Figure 2.** 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, Fisher's exact test, SAGE 300, and the tests of Madden et al and Audic and Claverie. 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. Madden's 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. (Reprinted from Ruijter et al. 2002)

The test suggested by Madden et al. (1997) is based on only the number of observed specific tags in each SAGE library and the test statistic is calculated as:

$$Z = \frac{n_1 - n_2}{\sqrt{n_1} + \sqrt{n_2}} \quad \text{Eq. 5}$$

with  $n_1$  and  $n_2$  as the number of specific tags in the first and second library, respectively. This test statistic is estimated to be normally distributed and can be compared to  $Z_{\alpha/2}$ . The test of Madden requires about 25% bigger differences than the Z-test and SAGE300 to reach statistical significance and is, therefore, more conservative (Fig. 2A). The origin of the test of Madden is not given in the original paper. However, the test seems to be based on the assumption that tag counts are Poisson-distributed. Since the variance of a Poisson distributed parameter is equal to its value, the square root of a tag count can be regarded to be the standard deviation of the tag count. How-

ever, when Eq. 5 represents an hypothesis test based on the difference of two Poisson-distributed tag counts divided by their standard error, the denominator of the test statistic should have been the square root of the sum of the tag counts, NOT the sum of the square roots (Ruijter et al. 2002). When the test statistic of Madden is 'corrected' accordingly, the critical values of this test, for large libraries, are similar to those of SAGE300 and other tests.

Audic and Claverie (1997) derived a new equation for the probability of finding  $n_2$  or more tags in one library given the fact that  $n_1$  tags have already been observed in the other library:

$$P(n_2|n_1) = \left(\frac{N_2}{N_1}\right)^{n_2} \frac{(n_1 + n_2)!}{n_1!n_2! \left(1 + N_2/N_1\right)^{(n_1+n_2+1)}} \quad \text{Eq. 6}$$

with  $N_1$  and  $N_2$  as the total number of tags in the first and second library, respectively. A summation of this probability over all  $n$  from  $n_2$  to  $\infty$  gives a one-sided  $P$ -value that can be compared to  $\alpha/2$ . The upper critical values for a significance level of 0.001 for Audic and Claverie's test are given in Fig. 2A and 2B as open circles. For libraries of equal and different size these critical values are all within 1.5% of those of the Z-test and SAGE300.

The Chi-square test can be used for comparing SAGE libraries (Michiels et al. 2000) after reorganizing the data in a 2 x 2 contingency table. This test is statistically equivalent to the Z-test on two proportions (Altman, 1991) and therefore will give the same  $P$ -values and have the same critical values as the Z-test.

Another test using 2 x 2 contingency tables is the Fisher's exact test (Altman 1991), which has also been applied to SAGE data (Man et al. 2000). Although the sampling design required by this test does not apply to SAGE (Conover 1980; Claverie 1999) and for the large number of tags involved in SAGE the Chi-square test is to be preferred (Altman 1991), the test has been included in comparisons of tests used for SAGE (Man et al. 2000; Ruijter et al. 2002). With the Fisher's exact test the  $P$ -value of the observed tag distribution is calculated as:

$$P(n_1, n_2) = \frac{\binom{N_1}{n_1} \binom{N_2}{n_2}}{\binom{N_1 + N_2}{n_1 + n_2}} \quad \text{Eq. 7}$$

which is the chance of finding the observed tag count distribution or an even less likely result when the null hypothesis is true (Altman 1991; the version of the Fisher's exact  $P$ -value given in Eq. 7 can be found at <http://quantrm2.psy.ohio-state.edu/kris/fisher/fisher.htm>). The  $P$ -value is compared to  $\alpha/2$ . The critical values of

the Fisher's exact test, when applied to SAGE data are very similar to those of the other tests (Fig. 2A and 2B, rectangles).

In the paper by Chen et al. (1998), a procedure, based on Bayesian statistics, is described to calculate the probability that the level of expression of a given mRNA is increased by at least X fold between libraries. Although this procedure can be used to statistically judge differences in tag numbers, its approach is clearly different from the classical approach of hypothesis testing and results of these test procedures cannot be directly compared. Similarly, the calculation of credibility intervals for the ratio of gene-expression between two libraries is conceptually very distinct and returns numbers with different interpretations (Vêncio et al 2003).

In conclusion, comparisons of tests for the differences between two SAGE libraries show that Monte Carlo simulation (SAGE300), Audic and Claverie's test, the Chi-square test, the Fisher's exact test and the Z-test, will all lead to the same decisions when applied for pair-wise comparison of SAGE libraries whereas Madden's test will behave considerably more conservative, unless the denominator of the test-statistic is 'corrected' (Ruijter et al. 2002). In a Monte Carlo comparison of the Chi-squared test, Fisher's exact test and Audic and Claverie's test it had previously been shown that the Chi-squared test, which is equivalent to the Z-test, has the best power and robustness (Man et al. 2000), especially at low expression levels.

## 8.7. Designing SAGE experiments.

An additional advantage of the Z-test is that this method of testing provides a way to calculate the number of tags that needs to be sequenced to detect a difference as significant (Fig. 3). In statistical testing, the relation between the difference  $p_1 - p_2$  that can be detected with a two-sided probability of a Type I error (incorrect rejection of the null hypothesis) of less than  $\alpha$ , as well as a probability of a Type II error (failure to detect a true difference  $p_1 - p_2$ ) of less than  $\beta$  can generally be expressed as:

$$difference > Z_{\alpha/2}SE(difference_{H_0}) + Z_{\beta}SE(difference_{H_1}) \quad \text{Eq. 8}$$

For the difference under  $H_0$ ,  $p_0=(n_1+n_2)/(N_1+N_2)$  should be substituted for  $p_1$  and  $p_2$  in Eq. 3. Thus, an equation for the calculation of the difference in proportions that can be detected as significant can be formulated by substitution of Eq. 3 into Eq. 8:

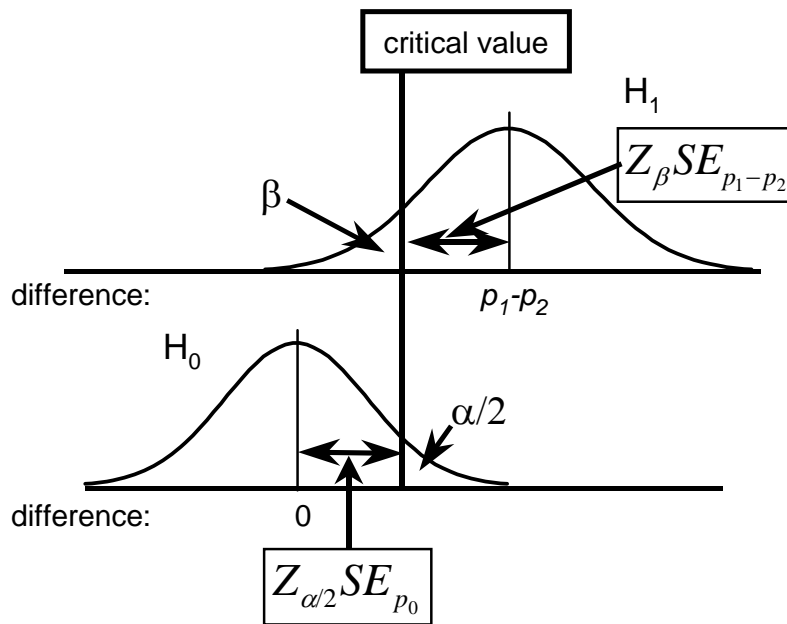
$$p_1 - p_2 > Z_{\alpha/2} \sqrt{p_0(1-p_0)/N_1 + p_0(1-p_0)/N_2} + Z_{\beta} \sqrt{p_1(1-p_1)/N_1 + p_2(1-p_2)/N_2} \quad \text{Eq. 9}$$

When  $N_1$  and  $N_2$  are equal, Eq. 9 can be rearranged into Eq. 10 which can then be used for the calculation of the number of tags needed to be sequenced in each of both

experimental conditions to detect a difference  $p_1 - p_2$  with a two-sided P-value of less than  $\alpha$  and a power greater than  $1-\beta$ :

$$N > \left( \frac{Z_{\alpha/2} \sqrt{2p_0(1-p_0)} + Z_{\beta} \sqrt{p_1(1-p_1) + p_2(1-p_2)}}{p_1 - p_2} \right)^2 \quad \text{Eq. 10}$$

(Armitage and Berry, 1987). However, when  $N_1$  and  $N_2$  are not equal, as is for instance the case when one wants to compare a SAGE library for an experimental condition with one for a control condition, which has already been published, such a rearrangement of Eq. 9 is not possible. Therefore, the sample size  $N_2$ , for a given  $N_1$ , significance level, power, a required difference  $p_1 - p_2$  can only be calculated by an iterative procedure based on Eq. 9.



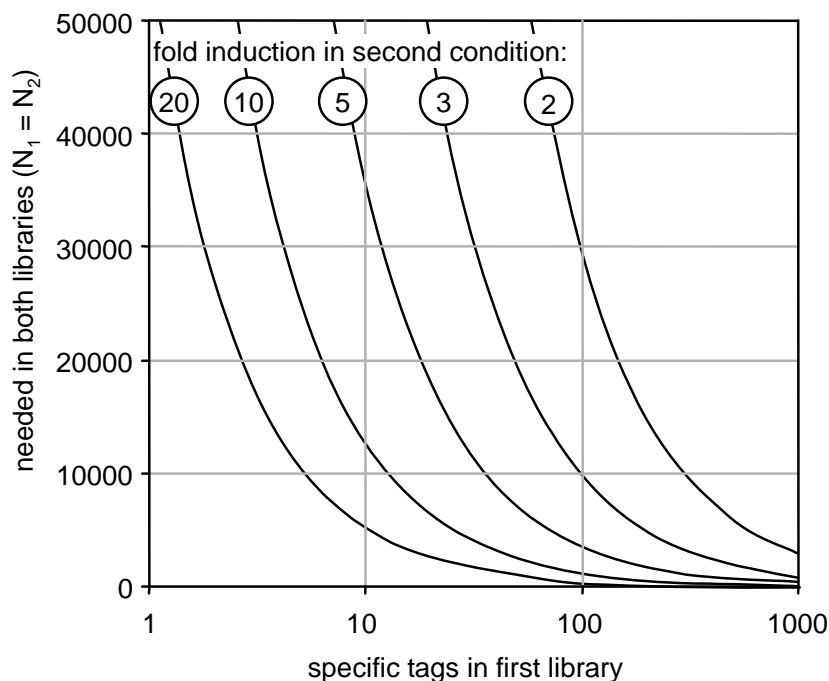
**Figure 3.** Illustration of the derivation of the sample size equation. When the null hypothesis ( $H_0$ ) is true the expected value of the difference  $p_1-p_2$  is 0. The distance between this expected value and critical value is  $Z_{\alpha/2}SE_{p_0}$ . When the alternative hypothesis is true the distance between the true  $p_1-p_2$  and the critical value is  $Z_{\beta}SE_{p_1-p_2}$ . Therefore, the distance from 0 to  $p_1-p_2$  is given by the sum of these two distances (Eq. 8).

Equation 9 can be used in different ways.

1. Given  $N_1$  and  $N_2$  (the SAGE libraries are compiled) the detectable difference can be calculated for a chosen significance level ( $\alpha$ ) and power ( $1-\beta$ ).
2. Given an observed difference, the total number of tags sequenced in both libraries and the chosen significance level, one can determine the power of the test.

3. 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 (Eq. 10; Fig. 4).
4. 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.

An example of the use of Eq. 10 is given in Fig. 4. From the nomogram in this figure it can be read that to detect a 5-fold increase in abundance for a transcript that occurs 10 times in the first library, two libraries of about 35 000 genes have to be assembled (Ruijter et al. 2002).



**Figure 4.** 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. (Reprinted from Ruijter et al. 2002)

### 8.8. Comparing more than two SAGE libraries.

The test for pair-wise comparison of two libraries cannot simply be used to test all possible pairs of libraries when more than two libraries have to be compared. The first objection to such an approach is that, although it will give you a matrix of differences between libraries, it does not tell you anything more: no information on subsets of libraries is gained. Additionally, multiple pair-wise testing would lead to an accumulation of Type I error. When two libraries are compared and the decision about rejecting the null hy-

pothesis is based on the significance level  $\alpha$  then the chance that a Type I error is made, which is the chance that a true null hypothesis is rejected, is equal to  $\alpha$ . For all possible pair-wise comparisons of  $k$  libraries the chance that at least one of the conclusions is wrong, would then accumulate to:  $1 - (1 - \alpha)^{\binom{k^2 - k}{2}}$ . For instance, for the 15 comparison needed to test all pairs of 6 libraries this would amount to a Type I error of 0.54, a chance of 54% that at least one of the conclusions is wrong. To avoid this accumulation of error, firstly the null hypothesis that all libraries have the same proportion of specific tags has to be rejected. Only after this overall null hypothesis has been rejected one can safely compare (subsets of) individual libraries. This reasoning is similar to the one that is used in the testing of the means of more than two samples: only after rejection of the overall null hypothesis with an overall analysis of variance, a multiple comparison of groups can be applied to test between-group differences.

library	tissue type	total N	specific n	proportion (x10 <sup>-6</sup> )
13	N	51280	39	761
30	MC	48554	2	41
37	A	80265	1	12
41	G	61886	3	48
42	G	70087	1	14
47	A	77004	1	13
56	M	38933	5	128
57	N	48545	56	1154
67	N	94876	47	495
68	N	58826	46	782
107	G	62675	2	32
112	N	77968	81	1039
122	NC	52261	2	38
125	N	63208	52	823
127	A	38634	3	78

**Table 1.** Overview of the example data-set used to illustrate the comparison of more than two SAGE libraries. The set consists of 15 SAGE libraries of brain tissue: 5 libraries of normal brain (N), a cell culture of normal brain tissue (NC) and 10 libraries of diverse brain tumors and cultures (A, G, M and MC). Specific tag counts are given for the tag AATAAAGCTA.

The overall null hypothesis that a specific tag has the same proportion in all libraries can be tested with the Chi-squared test on a  $k \times 2$  table. The Chi-squared statistic is calculated as:

$$Chi^2 = \sum_k \sum_{s+ns} \left\{ \frac{(n_{obs} - n_{exp})^2}{n_{exp}} \right\} \quad \text{Eq. 11}$$

in which  $n_{obs}$  is the observed number of specific or non-specific tags and  $n_{exp}$  is the expected number when the null hypothesis that all  $k$  libraries have the same proportion is true. Summation is over specific and non-specific tags and over all  $k$  libraries. The observed  $Chi^2$ -value is compared to the  $Chi^2$ -distribution with  $k-1$  degrees of freedom to determine its  $P$ -value. A disadvantage of the Chi-squared test is that the test is very

sensitive to low expected tag numbers that occur regularly in SAGE analysis. Also the test only allows you to make a decision about the overall null hypothesis. Likewise, the log-likelihood-ratio statistic derived by Stekel and co-workers (2000) only provides an overall test and offers no method to determine which libraries are deviating or to test differences between groups of libraries. It has been reported (Baggerly et al, 2003) that comparisons between two subsets of libraries have been carried out by pooling the tag counts and library sizes in both sets of libraries to obtain two 'large' libraries. Such a procedure completely ignores the between library variation in each of the subsets of libraries. Moreover, it adds a false sense of confidence to the tag frequencies because the tag counts in a large library have a smaller standard error (see also Eq. 2). Baggerly and co-workers (2003) propose the use of weighted proportions and variances in the calculation of a t-statistic between two subsets of libraries. Because the confidence in proportions depends on library size the weight factors that are included in this test depend on total and specific tag counts and may differ per tag. This test can only be used between a-priori-defined subsets of libraries.

To overcome the drawbacks of the above-mentioned approaches the preferred test for testing the overall null hypothesis that all  $k$  libraries have the same proportion of specific tags, is the Log-likelihood-ratio test or G-test (Sokal and Rohlf 1995). With this test the chance that the observed tag distribution is found when the null hypothesis is true is determined from the natural logarithm of the likelihood (or chance) ratio (Eq.12). The calculation of the G-statistic is based on a multinomial distribution, which is a generalization of the binomial distribution to the case where an attribute has more than two classes. For the calculation of the G-statistic the data for one tag are placed in a contingency table with the libraries as columns and two rows. The top row of this table contains the tag counts for the current tag in each library (specific counts) and the bottom row contains the number of other tags (non-specific counts). The numerator of the likelihood ratio is the probability of observing the frequencies in this contingency table. The denominator is this probability when the libraries and tag counts are independent. For a  $2 \times 2$  table (number of specific tags are  $a$  and  $b$ , nonspecific tags are  $c$  and  $d$ , in library 1 en 2 respectively;  $N$  is sum of  $N_1$  and  $N_2$ ), the G-statistic ( $G_{\text{intrinsic}}$ ) can be calculated as:

$$G = 2 \cdot \mathbf{Ln} \left( \frac{\frac{N!}{a!b!c!d!} \left(\frac{a}{N}\right)^a \left(\frac{b}{N}\right)^b \left(\frac{c}{N}\right)^c \left(\frac{d}{N}\right)^d}{\frac{N!}{a!b!c!d!} \left(\frac{(a+b)(a+c)}{N^2}\right)^a \left(\frac{(a+b)(b+d)}{N^2}\right)^b \dots \dots \left(\frac{(a+c)(c+d)}{N^2}\right)^c \left(\frac{(b+d)(c+d)}{N^2}\right)^d} \right) \quad \text{Eq. 12}$$

Note that  $a/N$  is the observed proportion of specific tags in the first library whereas  $(a+b)(a+c)/N^2$  is the expected proportion when both libraries have the same proportion. An important property of the G-statistic is that it is fully additive: G can be subdivided into different components, each representing a source of deviation from the overall null hypothesis. All these partial G's add up to the overall G that is calculated for the whole collection of libraries.

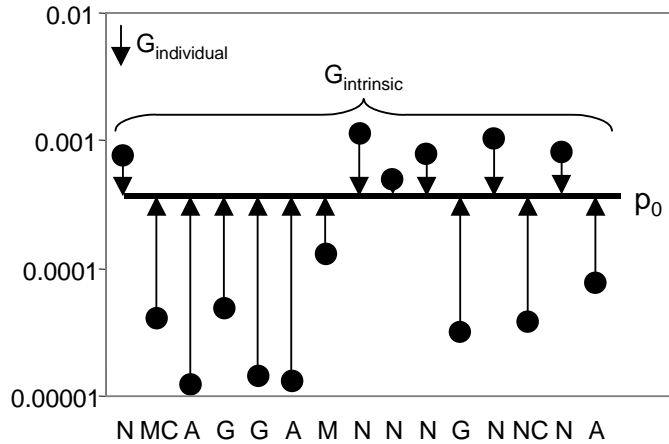
Fortunately, Eq. 12 can be simplified to an equation that can easily be implemented in computer programs or a spreadsheet:

$$G_{intrinsic} = 2 \sum_k \sum_{s+ns} \{n \cdot \mathbf{Ln}(n/n_0)\} \quad \text{Eq. 13}$$

in which  $n$  stands for the number of observed specific or non-specific tags in each library and  $n_0$  for the number of expected specific and non-specific tags when the null hypothesis is true:  $n_0 = p_0 N$  with  $p_0$  (the expected proportion when the null hypothesis is true) calculated as

$$p_0 = \sum_k n_k / \sum_k N_k \quad \text{Eq. 14}$$

Because this proportion is calculated from the current dataset, this overall G-statistic is called  $G_{intrinsic}$ . The summation in  $G_{intrinsic}$  (Eq. 13) is over specific and non-specific tags ( $s+ns$ ) in each library and over all  $k$  libraries. The calculation of  $G_{intrinsic}$ , and its relation to  $G_{individual}$  (G per library) is illustrated in Fig. 5. When the sample size is large, as is normally the case in SAGE, the G-statistic is approximately Chi<sup>2</sup>-distributed. The observed G-value is compared to the Chi<sup>2</sup>-distribution with  $k-1$  degrees of freedom (df) to determine its  $P$ -value. The degrees of freedom is one less than the number of libraries ( $k$ ) because the value of  $p_0$  is derived from the dataset. When the null hypothesis is rejected there are several ways to continue the analysis. These will be described in separate sections below.



**Figure 5.** Overall test: test of homogeneity of all libraries in the data set with the overall proportion  $p_0$  as expected value. The graph illustrates the additivity of the G-statistic: the G-value of the individual libraries add up to  $G_{intrinsic} = \sum G_{individual}$ .

### 8.9. Substitution of zero counts.

In the statistical analysis of SAGE libraries, tags that are not observed in one of the libraries included in the test are often excluded from the test. However, very low abundant transcripts, and thus zero tag counts still may reflect a true low transcript abundance, and their exclusion from the analysis may constitute a loss of valuable information. The  $n \cdot \ln(n/n_0)$  term that is part of the equations of all G-statistics cannot be calculated when the observed tag count  $n$  is zero. Therefore, a zero-substitution procedure has to be implemented. Replacing zeros by 1 might lead to high tag abundances in small libraries compared to the tag count of 1 in a larger library. Similarly, replacing zero tag counts by a number close to zero can lead to clearly false positive test results. The zero-substitution procedure that is implemented has to have a minimal effect on the test statistic and should not contribute to the decision about rejection of the null hypothesis. To obtain such an optimal zero substitution value, random tag counts of 0 and 1 are simulated for each of the libraries in the current collection of SAGE libraries. These simulated tag counts are generated in such a way that the tag count is always 0 in the smallest library and always 1 in the largest library. The other libraries are assigned either a 0 or 1, with the chance of getting a 1 depending on the library size. After each assignment of a 0 or a 1 tag count to each of the libraries, an iterative procedure determines a zero-substitution value that gives the minimal  $G_{intrinsic}$  for this simulated tag. The whole process is repeated 500 times and the mean zero-substitution value is used in the G-test procedure of all tags in the libraries. This way it is ensured that the zero-substitution, a value between 0 and 1, in itself will add only a minimal contribution to the G-statistic. This approach is on purpose conservative: genes that change from no to low expression and in which the tag count difference between 0 and 1 is real, will undeservedly not show up as statistically significant. However, the chosen

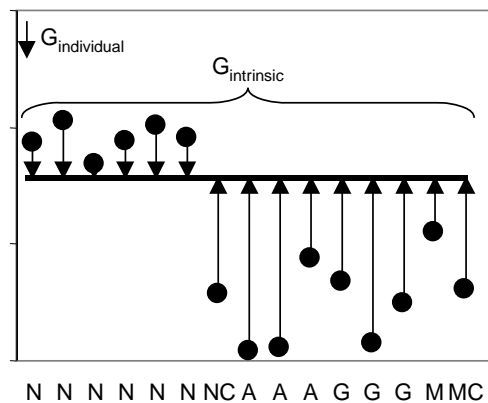
substitution procedure itself will not lead to a significant  $P$ -value and therefore avoids false positives.

### 8.10. After rejection of the null hypothesis.

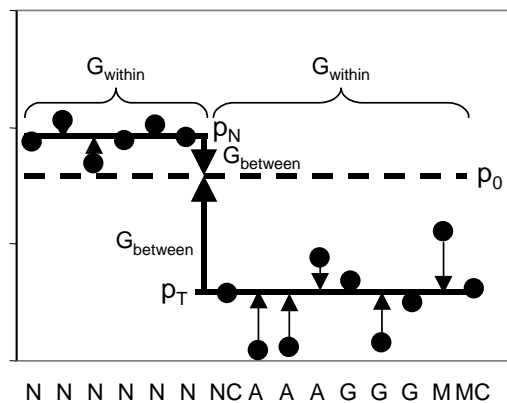
When the overall null hypothesis that the tag proportion in all libraries is equal is rejected, one can continue the analysis to determine which libraries deviate most from the expected proportion. To this end the  $G_{\text{individual}}$  for each of the  $k$  libraries is calculated as

$$G_{\text{individual}} = 2 \sum_{s+ns} (n_k \cdot \mathbf{Ln}\{n_k / (p_o N_k)\}) \quad \text{Eq.15}$$

with  $p_o$  according to Eq. 14.  $G_{\text{individual}}$  is two times the sum of  $n \cdot \mathbf{Ln}(n/n_o)$  for specific and non-specific tags. Note that, because of the additivity of  $G$ , the sum of  $G_{\text{individual}}$  is equal to  $G_{\text{intrinsic}}$  (Eq. 14; Fig. 5). Each  $G_{\text{individual}}$  is compared to the  $\text{Chi}^2$ - distribution with 1 df. This is a quick test to identify deviating libraries but its usefulness is restricted because the null hypothesis, and therefore the existence of a common proportion  $p_o$ , is already rejected. However, this test on  $G_{\text{individual}}$  can give useful information on which libraries deviate most from the others.



**Figure 6.** Test of two subsets with intrinsic  $p_o$  as reference proportion:  $G_{\text{intrinsic}}$  is the sum of  $G_{\text{individual}}$ . Note that this graph only differs from Fig 5. in the order of the libraries. The sum of  $G_{\text{individual}}$  per subset is  $G_{\text{within}}$ .



**Figure 7.** Test of two subsets with intrinsic  $p_o$  as reference proportion: the sum of  $G_{\text{individual}}$  per subset (Fig. 7A) can be divided into a  $G_{\text{within}}$  and a  $G_{\text{between}}$ . Each  $G_{\text{within}}$  is used for the test of the homogeneity within each subset. The sum of  $G_{\text{between}}$  can be used as a test for the difference between subsets.

A more revealing continuation of the test procedure is to test the deviation of each library from an expected proportion based on known information on the different libraries included in the data set. When you know that subsets of libraries share a common

property (tissue of origin or pathology) you can divide the libraries into subsets of libraries based on this knowledge and compare the average proportion of a specific tag in each of these subsets with the overall proportion of that tag. For each individual library  $G_{\text{individual}}$  is a measure for the deviation from the overall proportion (Fig. 6). However, for each set one can also calculate a G-value that sums up the deviations of the proportion of each library from the subset proportions ( $p_N$  and  $p_T$  in Fig 7). This ‘intrinsic’ G-value of each subset is called  $G_{\text{within}}$  and is calculated similar to  $G_{\text{intrinsic}}$  (Eq. 13) but with  $n_0$  based on the subset proportions  $p_N$  and  $p_T$ .  $G_{\text{within}}$  of a subset of  $k$  libraries has  $k-1$  df. The deviation the average subset proportion of a subset from the overall proportion  $p_0$  is called  $G_{\text{between}}$  and is calculated for each subset according to Eq. 16:

$$G_{\text{between}} = 2 \sum_{s+\text{ns}} \left( \sum_k n_k \cdot \mathbf{Ln} \left\{ \sum_k n_k / \left( p_0 \sum_k N_k \right) \right\} \right) \quad \text{Eq.16}$$

in which  $n$  and  $N$  of Eq. 15 have been replaced with their respective sums over the  $k$  libraries in a subset and the expected tag count,  $n_k$ , for the subset is given by  $p_0 \sum (N_k)$ . Each  $G_{\text{between}}$  is tested with 1 df.

When the researcher has defined  $m$  subsets, the additivity of the G-statistic leads to the following relation between the different G-values:

$$G_{\text{intrinsic}} = \sum_m G_{\text{within},m} + \sum_m G_{\text{between},m} \quad \text{Eq. 17}$$

in which  $m$  stands for the number of subsets. The  $\sum G_{\text{within}}$  can be considered a measure for the variation within subsets of libraries whereas the  $\sum G_{\text{between}}$  is a measure for the variation between subsets of libraries. The test of  $\sum G_{\text{between}}$  has  $m-1$  df because the overall proportion  $p_0$  was derived from the data set. Again, because of the fact that the overall test already rejected the notion of an overall proportion, the interpretation of  $G_{\text{between}}$  is difficult. Nevertheless,  $\sum G_{\text{within}}$  and  $\sum G_{\text{between}}$  play an important role in the search for homogeneous subsets of libraries that is described in section 8.13.

### 8.11. Supervised comparison of subsets.

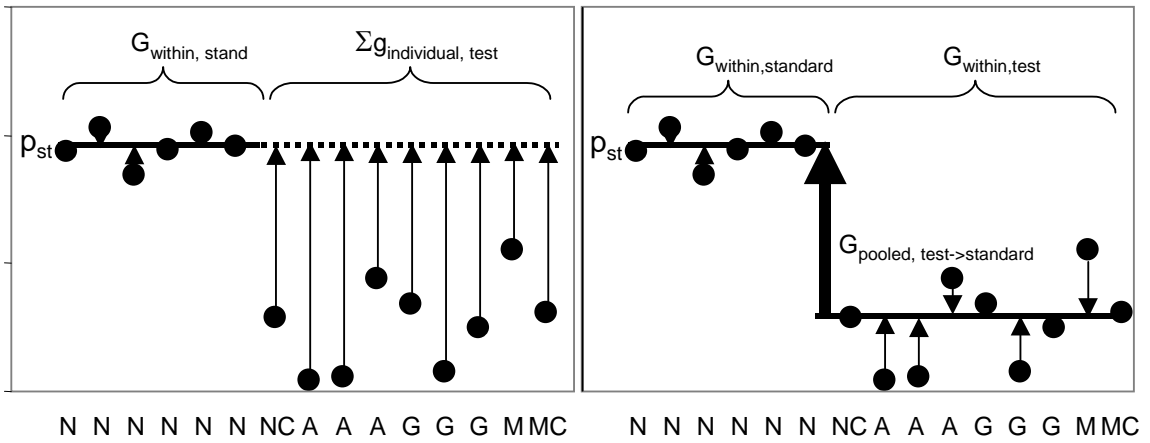
Instead of comparing each subset with the overall proportion of the data set, the researcher can decide to consider one of the subsets as the standard set (Fig 8). Because this comparison is based on the a-priori knowledge that subsets of libraries share some property (e.g. tissue of origin or pathology) and because the researcher

subdivides the libraries into two (or more) sets, this procedure has been dubbed a “supervised comparison of subsets”.

As described in section 8.10, a  $G_{\text{within}}$  for each subset can be calculated as a  $G_{\text{intrinsic}}$  (Eq. 13) but with  $n_0$  based on the subset proportions  $p_{\text{standard}}$  and  $p_{\text{test}}$ . The deviation of a test subset from the standard set is reflected in a G-statistic called  $G_{\text{pooled, test} \rightarrow \text{standard}}$ . This  $G_{\text{pooled}}$  is calculated for each subset with Eq. 18. This calculation is very similar to that of  $G_{\text{between}}$  (Eq 16) but uses the tag proportion of the standard set as the expected proportion  $p_{\text{standard}}$ .

$$G_{\text{pooled}} = 2 \sum_{s+ns} \left( \sum_k n_k \cdot \text{Ln} \left\{ \sum_k n_k / \left( p_{\text{standard}} \sum_k N_k \right) \right\} \right) \quad \text{Eq.18}$$

Again  $G_{\text{within}}$  and  $G_{\text{pooled}}$  are measures for the variation between libraries within a subset and the variation between subsets, respectively. Note that  $\sum G_{\text{within}} + \sum G_{\text{pooled}}$  exceeds  $G_{\text{intrinsic}}$  because of the difference of  $p_{\text{standard}}$  and  $p_0$ .



**Figure 8.** Supervised G-test with the proportion of an user-defined standard subset ( $p_{\text{st}}$ ) as reference. Note that the sum of  $G_{\text{individual}}$  for the standard set adds up to  $G_{\text{within,standard}}$ , whereas this sum for the test set can be divided into  $G_{\text{within, test}}$  and  $G_{\text{pooled, test} \rightarrow \text{standard}}$ .

### 8.12. Decision rules in the supervised comparison.

The G-statistics described in section 8.11 can be used to plan a test strategy in the supervised comparison of subsets. When the purpose of the test is to find genes that may be involved in some biological or pathological pathway (discovery-driven research) the following decision rules may be applied (example from Schaaf, Ruijter et al, 2005). A tag, as this test comprises a per-tag procedure, can be considered to be differentially expressed between the standard and a test subset of libraries when:

1. The null hypothesis that all  $k$  libraries have the same tag proportion is rejected (overall test:  $G_{\text{intrinsic}} > \text{Chi}^2_{0.05, k-1 \text{ degrees of freedom}}$ ),

2. The tag proportion in the standard subset of  $n$  libraries is homogeneous ( $G_{\text{within, st}} < \text{Chi}^2_{0.05, n-1 \text{ df}}$ ),
3. The tag proportion in each individual library in the test subset is significantly different from the abundance of the standard set ( $G_{\text{individual, t->st}} > \text{Chi}^2_{0.05, 1 \text{ df}}$ ),
4. The average tag proportion of a test subset of libraries differs significantly from that of the standard set ( $G_{\text{pooled, t->st}} > \text{Chi}^2_{0.05, 1 \text{ df}}$ ),

And, to obtain a more stringent set of differentially expressed tags,

5. The tag proportion in the test subset of  $n$  libraries is homogeneous ( $G_{\text{within, t}} < \text{Chi}^2_{0.05, n-1 \text{ df}}$ ).

Note that rules 1 and 2 always have to be satisfied and that rules 3, 4, and 5 are applied for each test subset individually. Where rule 3 serves to show that all test libraries differ from the standard subset, rules 4 and, especially, 5 ensure that this difference is in the same direction and that the test libraries can be considered to be one group. An example of the application of these decision rules is shown in Table 2.

**Table 2.** Example of the application of the decision rules on 25 tags from four rhabdomyosarcoma (RMS) and two normal muscle SAGE libraries (Schaaf, Ruijter et al). The cells in the table give the P-values of the respective G-statistics. The decision rules are indicated at the bottom of the table. Grayed-out P-values do not satisfy the decision rules. Only the tags without gray cells are considered to be differentially expressed. All rules are tested with  $\alpha = 0.05$  as significance level.

FinalMC (UC 163)	Gene symbol	TAG	Gintrinsic	Gwithin standard	library > type >>				Muscle_old Normal	muscle_yng Normal	Gpooled t->st	Gwithin test
					ARMS36 RMS	ERMS112 RMS	ERMS12 RMS	ERMS102 RMS				
	MC	AAAAATAAAG	0.000	0.000	0.001	0.237	0.028	0.764	0.000	0.003	0.111	0.001
Hs.415722	LOC283120	AAAGAAATGG	0.000	0.047	0.000	0.160	0.590	0.840	0.258	0.102	0.000	0.000
	MC	AACCAAAAAA	0.014	0.008	0.013	0.934	0.442	0.426	0.036	0.102	0.585	0.070
Hs.436439	PTK9L	AACCTGGCCT	0.044	0.023	0.069	0.031	0.302	0.548	0.069	0.168	0.003	0.953
Hs.183435	NDUFB1	AAGAATCTGA	0.001	0.033	0.000	0.041	0.024	0.439	0.109	0.162	0.000	0.248
	MC	AAGACAGTGG	0.000	0.009	0.000	0.405	0.551	0.154	0.058	0.069	0.000	0.000
		AAAAAAAAAA	0.000	0.994	0.000	0.000	0.000	0.000	0.996	0.996	0.000	0.000
		AAAAAACATT	0.048	0.994	0.824	0.916	0.000	0.498	0.996	0.996	0.002	0.044
Hs.432491	ESD	AAAAACTCC	0.001	0.974	0.069	0.002	0.000	0.186	0.982	0.982	0.000	0.002
		AAAACAGTGG	0.017	0.202	0.000	0.409	0.338	0.906	0.307	0.446	0.004	0.033
		AAAACATTCT	0.000	0.051	0.000	0.000	0.000	0.000	0.165	0.171	0.000	0.424
Hs.387804	PABPC1	AAAAGAAACT	0.000	0.994	0.000	0.000	0.000	0.000	0.996	0.996	0.000	0.000
		AAAATACTGA	0.048	0.994	0.824	0.001	0.061	0.001	0.996	0.996	0.000	0.181
		AAAATGAAAA	0.007	0.994	0.824	0.916	0.000	0.498	0.996	0.996	0.000	0.009
		AAAATGTACT	0.008	0.994	0.009	0.000	0.002	0.498	0.996	0.996	0.000	0.446
Hs.63657	NGLY1	AAAATTATCT	0.026	0.994	0.824	0.000	0.613	0.001	0.996	0.996	0.000	0.104
Hs.133892	TPM1	AAAGTCATTG	0.000	0.201	0.000	0.000	0.000	0.000	0.363	0.370	0.000	0.886
Hs.255950	LOC154866	AAATGTGCTG	0.000	0.682	0.000	0.000	0.002	0.030	0.776	0.769	0.000	0.953
Hs.446354	TCEA3	AACAAGGTGA	0.000	0.942	0.000	0.000	0.041	0.017	0.959	0.959	0.000	0.512
	MC	AACCCAGGAG	0.000	0.329	0.000	0.000	0.000	0.050	0.518	0.465	0.000	0.753
	MC	AACCCGGGAG	0.000	0.320	0.000	0.000	0.000	0.000	0.467	0.497	0.000	0.488
		AAGCTGAGGT	0.000	0.485	0.000	0.000	0.000	0.000	0.664	0.585	0.000	0.004
Hs.375921	RPL31	AAGGAGATGG	0.000	0.804	0.000	0.000	0.000	0.000	0.862	0.859	0.000	0.000
Hs.405590	EIF3S6	AATATTGAGA	0.000	0.672	0.000	0.000	0.000	0.000	0.773	0.756	0.000	0.027
Hs.433394	TUBA3	AATGCTTTGT	0.000	0.994	0.009	0.000	0.000	0.000	0.996	0.996	0.000	0.000

Rule > 1 2 3 4 5

All tags in this selection meet the first rule (all  $G_{\text{intrinsic}}$  are significant) whereas the first 6 do not pass rule 2 because their  $G_{\text{within, st}}$  are significant ( $P < 0.05$ ). The third rule demands that all  $G_{\text{individual, t->st}}$  in the test subset of RMS libraries are significant,

which is not the case for 7 tags. All remaining tags pass rule 4 ( $G_{\text{pooled, t-st}}$ ) but another 6 tags fail to meet rule 5 ( $G_{\text{within, test}}$ ). This leaves 6 differentially expressed tags in this example (Schaaf, Ruijter et al. 2005). One of them is the tag AAAGTCATTG representing the gene TPM1 (tropomyosin). This is a muscle specific gene, coding for a protein that is associated with the actin filament and is involved in regulation of muscle contraction. The tag for TPM1 is expressed in the normal muscle libraries at an average tag count of 258 (238 and 278, respectively) per 100000, while an average tag count of 5.5 (6.2, 9.4, 6.6 and 0, respectively) per 100000 was observed in the RMS libraries. This shows that this gene plays an important role in differentiated healthy muscle and is significantly down regulated in rhabdomyosarcoma, a pediatric tumor with features (origin) of muscle tissue.

By customizing the significance level for each of the decision rules one can tune the chance that false positive or false negative test results occur.  $G_{\text{intrinsic}}$  and  $G_{\text{within}}$  are used to test for homogeneity in the total set or in each of the subsets of libraries, respectively. A high significance level (e.g.  $\alpha=0.05$  or more) for these statistics means that small differences between libraries already mark a tag as non-homogeneous. However, the effect of decreasing the significance level on the detection of differentially expressed tags differs for these two G-statistics. Decreasing the significance level for  $G_{\text{intrinsic}}$  will increase the occurrence of false negatives (differentially expressed tags may have been missed). For  $G_{\text{within}}$  it will mean that subsets will be considered homogeneous more quickly and more tags can pass rules 2 and 5. On the other hand, the rules based on  $G_{\text{pooled}}$  and  $G_{\text{individual}}$  test the deviation from a standard set of libraries for a test subset or an individual library, respectively. A low significance level (e.g.  $\alpha=0.01$  or less) for these statistics avoids false positives because tags have to differ strongly to reach such low P-values.

The choice of significance levels for the different G-statistics depends on the aim of the experiment. When the researcher has only a limited number of libraries and does not want to miss potentially interesting (groups of) genes, all significance levels should be set to 0.05. This setting avoids that genes are missed because of random heterogeneity between libraries or when the differences in the used samples are accidentally small. When one is not worried about missing interesting genes but only wants to select genes that have a strong differential expression between very homogeneous groups of libraries, the significance levels for  $G_{\text{intrinsic}}$ ,  $G_{\text{pooled}}$  and  $G_{\text{within}}$  should be set to e.g. 0.01, 0.001 and 0.05, respectively. Similarly, the significance level for  $G_{\text{individual}}$  can be set to adjust the accepted deviation of individual libraries from their subset. When more than one test set of libraries is compared with the standard set, the  $G_{\text{pooled}}$  and

$G_{within}$  statistics are computed for each test set. This way several subsets of libraries can be tested simultaneously.

### 8.13. Unsupervised identification of homogeneous subsets of libraries.

The aim of discovery-driven SAGE research is to assign functions to genes. In SAGE libraries a lot of tags are derived from unidentified transcripts or genes with as yet unknown functions (e.g tag AAAACAGTGG in Table 2). When libraries containing such tags can be classified into subsets that display similar expression levels for such a tag, the common properties of such libraries may point to a functional role of these transcripts. Classifying libraries can be done by looking for homogeneous subsets of libraries, which are sets of libraries with non-significant  $G_{within}$  values. Because this procedure is carried out without input of a-priori knowledge about the libraries, the method is dubbed an “unsupervised identification of homogeneous subsets”. The principal requirement for starting this procedure for a specific tag is again that the overall hypothesis that all libraries have the same proportion for this tag is rejected.

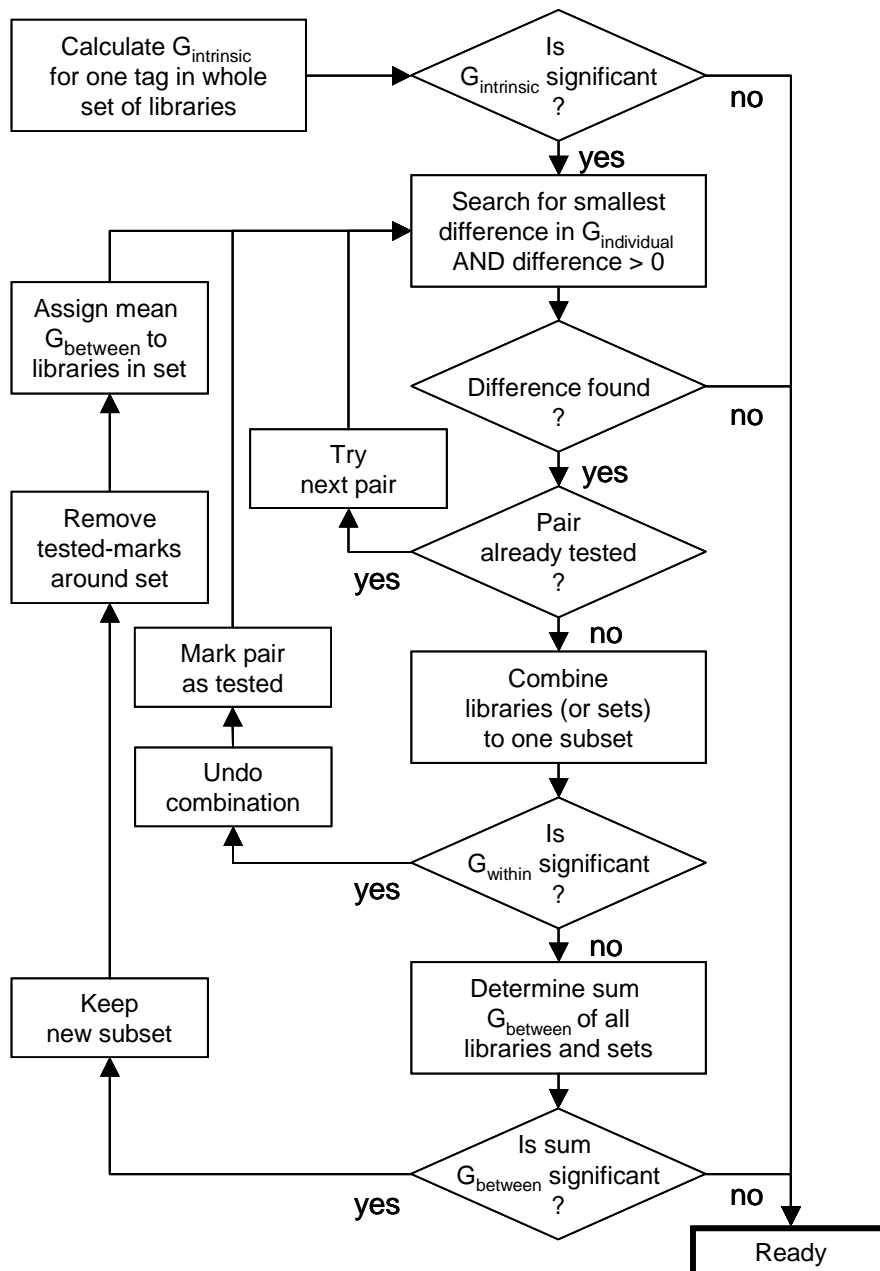
Assigning each library to a subset requires an algorithm that minimizes the number of statistical comparisons. To this end, the libraries are ranked on the basis of their  $G_{individual}$  (Eq. 15) from the overall proportion  $p_0$  (taking the direction of the difference into account) and only the joining of adjacent libraries or subsets in this ranking are considered and tested. The whole procedure is illustrated in the flow chart in Fig. 9. It is based on  $G_{individual}$  and  $G_{between}$ , which are both calculated with respect to the overall proportion  $p_0$  (Eq. 15 and Eq. 16, respectively) and serve as a measure for the difference between subsets. For a collection of  $k$  libraries the sum of  $G_{individual}$  and  $G_{between}$ , is always compared to the  $\text{Chi}^2$  value with  $k-1$  df. Additionally  $G_{within}$  (see 8.10) is used as a measure for the homogeneity within the subset resulting from each step. To test this homogeneity the  $G_{within}$  of the new subset of  $n$  libraries is compared to the  $\text{Chi}^2$  value with  $n-1$  df.

The procedure is iterative with two stopping criteria. The search for homogeneous subsets in a data set of  $k$  libraries stops when:

1. The sum of  $G_{individual}$  and  $G_{between}$  becomes smaller than the critical  $\text{Chi}^2$  value based on  $k-1$  df, in other words, when the sum of  $G_{individual}$  and  $G_{between}$  is no longer significant for the given alpha.
2. All combinations of libraries and subsets have been tested, but some may have been rejected because of the occurrence of a significant  $G_{within}$ .

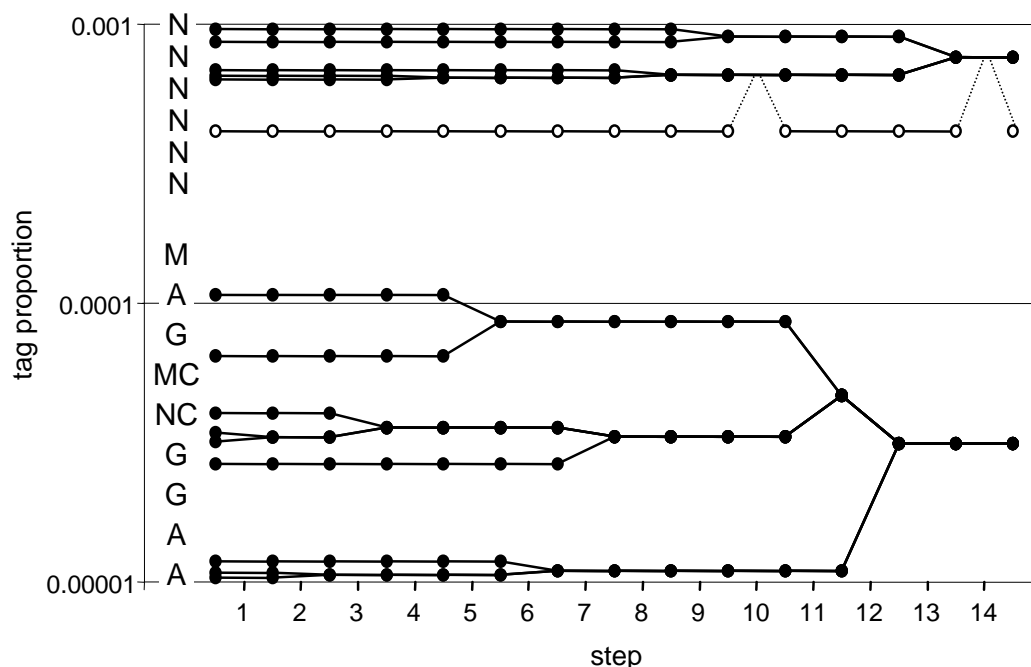
Because the P-value of  $G_{within}$  depends on the number of libraries in the subset, libraries flanking a newly formed subset that were rejected in a previous step are tested

again when their  $G_{\text{individual}}$  is close to the mean  $G_{\text{between}}$  of the new subset. Note that this procedure divides the original  $G_{\text{intrinsic}}$  of the whole data set into a sum of  $G_{\text{between}}$  that is maximized and a sum of  $G_{\text{within}}$  that is minimized. Therefore, this algorithm results in a collection of subsets that are each homogenous and that differ maximally from each other.



**Figure 9.** Flow chart of unsupervised search for subsets of libraries. The search process starts when  $G_{\text{intrinsic}}$  is significant. Note that libraries will be added to a subset as long as its  $G_{\text{within}}$  is not significant. The process stops when the sum of  $G_{\text{individual}}$  and  $G_{\text{between}}$  is no longer significant or when all additions to each subset are tested.

When applied to the example data set in Table 1, this algorithm finds two homogeneous subsets of 5 normal brain libraries and 9 brain tumors, respectively, whereas 1 library cannot be included in either of these sets (Fig. 10). On closer inspection of the example data set, this normal brain library turned out to be the only white matter library among the normal, gray matter, brain libraries.



**Figure 10.** Illustration of the unsupervised search for homogeneous subsets of libraries. The example data set consists of 15 brain-tissue libraries and the counts for tag AATAAAGCTA (see also Table 1). The libraries are sorted based on their proportion (first column on the left). The difference between the  $G_{\text{individual}}$  of each neighboring pair of libraries is used to decide which libraries to combine. When this combination does not lead to a significant  $G_{\text{within}}$ , and when the sum  $G_{\text{between}}$  is still significant the resulting set is accepted. The procedure iterates and another pair of libraries, or a library and an earlier subset are combined. Several subsets thus grow simultaneously. Note that in step 10 a significant  $G_{\text{within}}$  occurs and the combination is undone (dotted lines). This combination is again tried and rejected in step 15. The procedure results in two non-overlapping subsets and one library that cannot be included in those sets.

#### 8.14. Concluding remarks.

In the above text a survey is given on the statistical comparison of the data obtained for one tag in two or more SAGE libraries. In the near future, the increasing availability of SAGE libraries enables the extension of these tests in several directions even when we restrict ourselves to the test of one tag at the time. The additivity of the G-statistic also allows the test of more complicated experimental designs in sets of SAGE libraries. E.g. when a collection normal and pathologic SAGE libraries are available of different tissues one can aim a study at identifying genes that show tissue spe-

cific changes that occur in a common pathology. To accomplish that, an algorithm can be used that subdivides the  $G_{\text{intrinsic}}$  into a  $G_{\text{between tissues}}$  and a  $G_{\text{between pathologies}}$  component that reflect the difference between tissues irrespective of pathology and between pathology irrespective of the tissue, respectively. However, an extra component that 'measures' the dependence of the pathological changes on the tissue, dubbed a  $G_{\text{interaction}}$  in accordance with analysis of variance, will then be present. The latter component may be biologically and clinically of more importance because it enables the test of tissue specific disease related gene activation or vice versa.

Another extension that the G-statistic allows is the test for more than one tag at the time. Whereas in the above approach a  $k \times 2$  table was used to find subsets of libraries with a similar proportion of one tag, converting the design to a  $2 \times k$  table would allow the search of tags with a similar behavior in two libraries can easily be accomplished. In principle the unsupervised identification of subsets can be applied to a  $k \times 2 \times \text{number of tags}$  design. It should then be possible to expand this discovery-driven approach to a search for combined subsets of libraries and tags.

### 8.15. Computer programs

A computer program that performs the Z-test between two libraries, as well as the calculations needed to plan the required library size in a SAGE study, is available through automatic request. Send an e-mail to [bioinfo@amc.uva.nl](mailto:bioinfo@amc.uva.nl) with the subject: **SAGEstat**. Other programs for the comparison of two libraries have been made available by the authors (Audic and Claverie 1997; Man et al. 2000).

A program that uses the G-statistic to test differences between more than two libraries has been made available through a similar link: [biolab-services@amc.uva.nl](mailto:biolab-services@amc.uva.nl); subject Gtest (Schaaf, Ruijter et al. 2005).

## References.

1. Akmaev VR, Wang CJ. Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics*, xx-xx, 2004.
2. Altman DG. *Practical statistics for medical research*. London: Chapman-Hall, 1991, p. 161-167 and 253-258.
3. Armitage P, Berry G. *Statistical methods in medical research*. Oxford: Blackwell Scientific Publications. 1987. p. 179-185.
4. Audic S and Claverie J-M. The significance of digital gene expression profiles. *Genome Research* 7: 986-995, 1997.
5. Baggerly KA, Deng L, Morris JS, Aldaz CM. Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* 19: 1477-1483, 2003
6. Boheler KR, Stern MD. The new role of SAGE in gene discovery. *Trends Biotech* 21, 55-57. 2003.
7. Chen H, Centola M, Altschul SF and Metzger H. Characterization of gene expression in resting and activated mast cells. *J Exp Medicine* 188: 1657-1668, 1998.
8. Chen J, Sun M, Lee S, Zhou G, Rowley J, Wang SM. Identifying novel transcripts and novel genes in the human genome by using novel SAGE tags. *PNAS* 17 12257-12262, 2002
9. Claverie J-M. Computational methods for the identification of differential and coordinated gene expression. *Human Mol Genetics* 8: 1821-1832, 1999.
10. Conover WJ. *Practical nonparametric statistics*. New York: Wiley, 1980, p. 162-167.
11. Constans A. The state of the microarray. *The Scientist* 17: 34, 2003.
12. Diverse authors. Method of the year 2007. *Nature Methods* 5: 1, 11-21, 2008.
13. \* 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.
14. Lander, ES et al. Initial sequencing and analysis of the human genome. *Nature* 409: 860 – 921, 2001.
15. Madden SL, Galella EA, Zhu JS, Bertelsen AH and Beaudry GA. Sage transcript profiles for P53-dependent growth regulation. *Oncogene* 15: 1079-1085, 1997.
16. Man MZ, Wang X and Wang Y. POWER\_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* 16: 953-959, 2000.
17. Michiels, EMC., Oussoren E, Van Groenigen M, Pauws E, Bossuyt PMM, Voûte PA and Baas F. Genes differentially expressed in medulloblastoma and fetal brain. *Phys Genomics* 1: 83-91, 1999.
18. Patino WD, Mian OY, Hwang PM. Serial Analysis of gene expression. Technical considerations and applications to cardiovascular research. *Circulation Research* 91: 565-569, 2002.
19. Ruan Y, Le Ber P, Ng HH Liu ET. Interrogating the transcriptome. *Trends in Biotechnol* 22: 23-30, 2004.
20. \* Ruijter JM, Van Kampen AHC, Baas F. Statistical evaluation of Serial Analysis of Gene Expression (SAGE) libraries: consequences for experimental design. *Phys Genomics* 11: 37-44, 2002.
21. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. Using the transcriptome to annotate the genome. *Nature Biotech* 19: 508-512, 2002.
22. \* Schaaf GJ, Ruijter JM, van Ruissen F, Zwijnenburg DA, Waaijer R, Valentijn LJ, Benit-Deekman J, van Kampen AHC, Baas F, Kool M. Comparative gene expression profiling of rhabdomyosarcomas and normal skeletal muscle: statistical com-

- parison of multiple SAGE libraries. *FASEB J* 19,404-406, 2005 (with online additional material on the G-test)
23. Sokal RR, Rohlf FJ. *Biometry, the principles and practice of statistics in biological research*. Chapter 17. Analysis of frequencies. New York, WH Freeman and company, 1995, p 685-789.
  24. Stekel, D.J., Y. Git, and F. Falciani. The comparison of gene expression from multiple cDNA libraries. *Genome Res* 10: 2055-2061, 2000.
  25. Van den Oord EJCG, Sullivan, PF. False discoveries and models for gene discovery. *Trends Genetics* 19: 537-542, 2003.
  26. \* Van Kampen AHC, Ruijter JM, van Schaik BDC, Caron HN, Versteeg, R. Gene expression informatics and analysis. In: *Bioinformatics for geneticists*, Barnes MR, Gray IC (Eds), London, John Wiley & Sons, 2003, p 319-344.
  27. Vêncio RZN, Brentani H, Pereira CAB. Using credibility intervals instead of hypothesis tests in SAGE analysis. *Bioinformatics* 19, 2461-2464, 2003.
  28. Venter JC et al. The Sequence of the Human Genome. *Science* 16: 291: 1304-1351, 2001.
  29. Velculescu VE, Zhang L, Vogelstein B and Kinzler KW. Serial analysis of gene expression. *Science* 270: 484-487, 1995.
  30. Velculescu VE, Vogelstein B and Kinzler KW. Analyzing uncharted transcriptomes with SAGE. *Trends Genetics* 16: 423-425, 2000.
  31. 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.

\* Contributions to these papers have been used for this reader (JR).

## 8.15 Practical exercises statistics on SAGE libraries.

The practical exercises in this part of the course are meant to give you some hands-on experience with the tests for the comparison of two and more SAGE libraries that have been discussed in the previous sections. In these exercises you will be learning how to use Excel to do these test. But you will also use SAGEstat to perform the Z-test between two libraries, and to calculate the number of tags that you need to sample when you want to detect an expected difference as statistically significant. The third program you will use is G-test. This is a program to perform the G-test between more than two libraries. You will find an Excel file with a sample data set and some pre-arranged sheets in the user directory of the computers in the exercises room. The programs can be found under Bioinformatica in the Start menu. The exact location will be given at the start of the exercises.

Please note:

1. That SAGEstat, as well as G-test, expect the Excel file to be opened before they are started.
2. Microsoft Office programs, like Excel, in the exercise room 'think' Dutch: you will have to enter data with decimal commas instead of decimal points.

### Exercise 1: Z-test and SAGEstat

#### Z-test with Excel

- Start Excel
- Open the file 'exercises\_SAGEstat\_course.xls'
- Activate the sheet 'exercise\_1\_Z\_test'

This sheet contains parts of two SAGE libraries. The annotation of these libraries is given in the first 8 rows. Row 9 gives the total library size and row 10 till 52 give the specific tag counts of 43 different tags. You can use Excel to perform a Z-test per tag:

- Label the columns: enter 'p1', 'p2', 'p0', 'Z', and 'P' into cells F9, G9, H9, I9, and J9, respectively.
- Calculate p1: enter  $=B10/B\$9$  into cell F10
- Calculate p2: enter  $=C10/C\$9$  into cell G10
- Calculate p0: enter  $=(B10+C10)/(B\$9+C\$9)$  into cell H10
- Calculate Z: enter  $=ABS(F10-G10)/SQRT(H10*(1-H10)*(1/B\$9+1/C\$9))$  into cell I10
- Calculate P: enter  $=2*(1-NORMSDIST(I10))$  into cell J10
- Copy cells F10 till J10 down to row 52

Refer to Eq. 4 in section 8.6 for details about the formula you have just used.

Why is there a multiplication by 2 in the calculation of the P-value in column J?

### **Z-test with SAGEstat**

SAGEstat can be used to perform a Z-test on the difference in tag counts in two libraries: per tag and for a series of tags.

- Start SAGEstat.

SAGEstat opens with a screen that gives you a choice of 3 testing procedures and 6 planning procedures.

#### **One tag:**

- Press 'difference between two tags'.

Each screen in SAGEstat consists of an input and an output part. For this test SAGEstat needs the number of specific tags and the total number of tags in each library.

- Give as total number of tags 52261 and 63208 for library 1 and 2, respectively.
- Give as number of specific tags 12 and 22 for library 1 and 2, respectively.
- Press 'calculate'

The results of the test are displayed in the output part of the screen. This gives you:

- the proportions in each library
- the difference in proportions ( $p_1 - p_2$ )
- the confidence interval of the difference for the given significance level

What is the P-value of the difference?

- Exchange the numbers of specific tags into 22 for library 1 and 12 for library 2.

What is now the P-value?

Can you explain why the P-value is different while you are testing the same difference in specific tag numbers?

#### **Series of tags:**

Instead of entering all specific tag numbers by hand, SAGE stat can test all pairs of tags in two libraries.

- Make sure that the sheet "exercise\_1\_Z\_test" is active in Excel.
- Go back to SAGEstat.
- Press the Tab 'between two libraries'
- Press 'connect'

SAGEstat connects to the active Excel sheet and displays the dialog of Fig. 11.

book:	exercises_SAGEstat_course.xls			
sheet:	exercise_1_Z_test			
Read from:		Write to:		
	column:	row #:	column:	row #:
N1	B	9	N2	C
n1	B	10	n2	C
	trhu:	52	trhu:	52
			column:	row #:
			D	10
			trhu:	52

**Figure 11.** Input interface that appears when SAGEstat is used for the comparison of all tags in two libraries.

- Make sure the book and the sheet box display the right book and sheet.
- Fill in the edit fields as in the figure. The easiest way is to start with the upper left B and press Tab to go to the next field.

With this you tell SAGEstat from which cells it should read the library totals (N1 and N2) and from which rows the numbers of specific tags (n1 and n2). Check the Excel sheet to see which cells are referred to by the entries in Fig. 10.

- Press 'Calculate P-values'.
- Activate Excel

The P-value for each tag is written to column D of the Excel sheet. Label this column with P(Z-test). Column D only displays 3 decimals but you can extend this.

- Compare the values in column D with those you just calculate in column J

### Calculation of required library size.

SAGEstat can be used to determine the number of tags that have to be sequenced to detect an expected difference between libraries as significant. You can do this for one specific difference or, when you have no prior idea of the difference you expect, for a matrix of differences.

### One specific difference

- In SAGEstat: press 'procedures'
- Press 'library size for both libraries'

The expected difference has to be given as expected abundance per total number of mRNA transcripts in the cell:

- Give total: 100000
- Give tissue 1: 50

- Give tissue 2: 100
- Press 'calculate'

How many tags do you have to sequence per library to be able to conclude that the difference between 50 and 100 transcripts per 100000 is statistically significant?

- Change the significance level alpha to 0.01
- Press 'calculate' again

Can you explain why you need to sequence fewer tags per library?

- Change beta to 0.2
- Press 'calculate' again

At what expense have you managed to reduce the number of required tags?

### **Comparison with an existing library**

Suppose someone else has already sequenced a library of 25 000 tags of the control condition of the tissue you are studying and made this library public. In that case you can plan to compare your library with that already available library.

- Press the 'N for second library' Tab.
- Give total transcripts: 100000
- Give 50 specific transcripts in tissue 1 and 100 in tissue 2
- Give 25000 for the total tags in the control library
- Press 'calculate'

What does the result (\*\*\*\*\*) mean?

### **Library size matrices**

SAGEstat can also be used to calculate a matrix of required library sizes. You can use this option to plan your SAGE experiments when you do not have a fixed idea about the differences you expect to find.

- Click the 'N for both libraries' Tab.
- Choose the radio button 'matrix'.
- Choose the radio button 'big matrix'.
- Give total: 100000
- Press 'calculate'

The left gray column of the output gives the number of copies in tissue 1 (to be interpreted as transcripts per 100000 for this input). The numbers in the gray top line give

the fold difference between tissue 1 and tissue 2. The numbers in the table are the required number of tags per library.

- Press show graph

The X-axis of the graph is the abundance of transcripts in tissue 1 (per 100000 in this case). The Y-axis gives the number of tags that has to be sequenced in both libraries for a chosen fold difference to be significantly different (choose a fold difference in the list box on the top-right). The graph can be saved to clipboard and pasted into your presentation program.

### **Exercise 2. Accumulation of Type I error**

The accumulation of Type I error that results from multiple testing can easily be demonstrated with a calculation in Excel.

- Activate Excel.
- Go to sheet "exercise\_2\_Type\_I\_error"

The sheet already contains a template for the graph that we want to make.

When you have  $k$  libraries and you compare each library with each of the others in a test that compares two libraries, you will do  $k(k-1)/2$  comparisons.

Column B in the sheet gives the number of libraries. Calculate the number of pair-wise comparisons:

- Enter  $=B6*(B6-1)/2$  into cell C6
- Copy this formula downward in column C

This is also the number of conclusions that you will draw from the resulting P-values. The chance that all of these conclusions are right is (1 minus the significance level) to the power the number of tests. Calculate this chance in column D:

- Enter  $=(1-C\$3)^{C6}$  into cell D6
- Copy this formula downward in column D

The chance that one or more conclusions were wrong is then 1 minus the above-calculated chance. To calculate this chance:

- Enter  $=1-D6$  into cell E6
- Copy this formula downward in column E

Excel automatically updates the graph. The graph shows the accumulated Type I error. What is the chance of a wrong conclusion when you do all pair-wise comparisons between 10 libraries?

What was the chance of a wrong conclusion that you were willing to accept?

- Change the value of alpha (cell C3) in =C2/10

What is now the chance of a wrong conclusion when you do 10 pair-wise comparisons?

### **Exercise 3: G-test with intrinsic $H_0$**

To avoid accumulation of Type I error you first have to reject the overall null hypothesis that all libraries have the same proportion of specific tags. This hypothesis can be tested with, among others, the G-test. This exercise illustrates how this can be done with Excel.

- Activate Excel.
- Go to sheet "exercise\_3\_G\_test\_intrinsic\_Ho"

The sheet shows 35 SAGE libraries (35 rows) of brain tissue. The annotations of the libraries are in columns A to H. Column I gives the library size and column J the counts of the specific tag (in this example AAGATCCCCG) in each library.

When the  $H_0$  that all libraries have the same proportion of specific tags is true, the best estimate for this proportion can be calculated from the sum of the library totals and the sum of the specific tags counts:

- Calculate the sum of all library totals in cell I38: enter =SUM(I2:I36)
- Calculate the sum of the specific tags counts in cell J38: enter =SUM(J2:J36)
- Calculate the overall proportion of specific tags in cell K38: enter =J38/I38

To do a G test we need the number of non-specific tags. Calculate these numbers in column K:

- Enter =I2-J2 into cell K2
- Copy downward in column K

Calculate the expected specific tags numbers when the  $H_0$  is true in column M:

- Enter =K\$38\*I2 in cell M2
- Copy downward in column M

And also calculate the expected number of non-specific tags

- Enter =I2-M2 in cell N2
- Copy downward in column N

The G-statistic is defined as 2 times the sum over all libraries of the sum of  $n \cdot \ln(n/n_0)$  for specific and non-specific tags in which  $n$  is the observed number and  $n_0$  is the expected number when the null hypothesis is true. G can be calculated as follows:

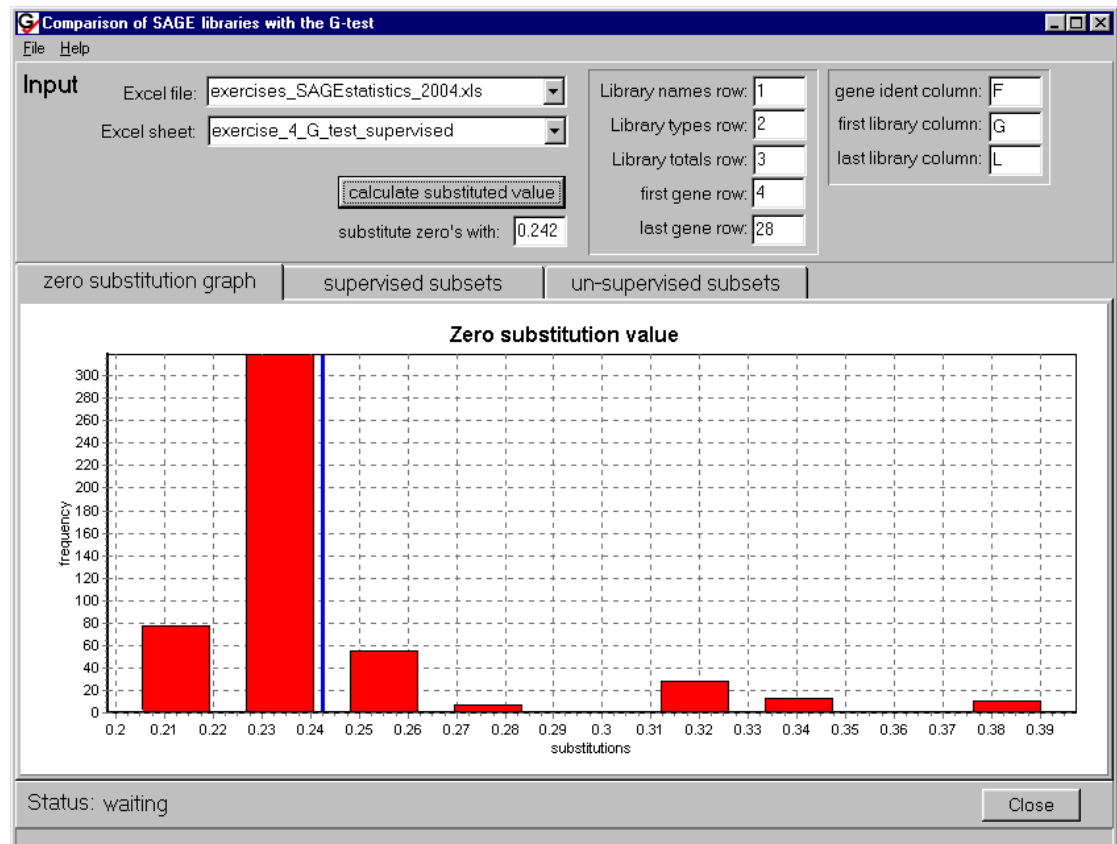
- Enter  $=J2*LN(J2/M2)$  into cell P2
- Copy this formula to cell Q2
- Copy downward in both columns P and Q
- Enter  $=SUM(P2:P36)$  into cell P38
- Copy this formula to cell Q38
- Enter  $=2*(P38+Q38)$  into cell R38

The G-statistic is Chi-square distributed. So we can calculate the probability of finding this value of G:

- Enter  $=CHIDIST(R38;COUNT(I2:I36)-1)$  into cell R40

The “COUNT(I2:I36)-1” part in this formula stands for the degrees of freedom, which in this case is the number of libraries minus 1.

What do you conclude from the P-value in cell R40?



**Figure 12.** Interface of the G-test program after calculation of the zero substitution value.

**Exercise 4: G-test between a subset of rhabdomyosarcoma (RMS) libraries a normal muscle standard subset.**

The G-test can be used to test whether one or more subsets of libraries differ from a subset of libraries that has been defined as standard set. Although it is possible to do this with Excel, it is easier to use a dedicated program and, therefore, we developed a program to do this so-called supervised G-test. The theory behind the program is discussed in the previous sections.

- Activate Excel and open the sheet "exercise\_4\_G\_test\_subsets"

This sheet contains 6 SAGE libraries (4 RMS and 2 normal muscle libraries) in columns G thru L. The sheet contains the selected data of 25 tags (rows 4 to 28), however, the program is not limited in the number of tags it can process.

- Start the G-test program (G\_test\_projectXX.exe)

The program interface is shown in Fig. 12.

- Make sure the book and sheet box of G-test refer to the names given in Fig. 12.
- Make sure the edit fields for rows and columns are filled as they are in Fig. 12.
- Check the Excel sheet to see which rows and columns the cells refer to.

The Tab-pages are disabled until you have calculated a zero-substitution value. To determine the best zero-substitution value for this set of libraries:

- Press the 'calculate substitution value' button

The graph will appear in the bottom part of the screen and the best substitution value is written into the edit field in the top part of the interface.

When you, because of the tissue of origin of your SAGE libraries, know which libraries to combine into subsets you can use this information to do a supervised comparison of subsets.

Go to the Tab-page 'supervised subsets' (Figure 13). To do a comparison of subsets you have to:

- Define the subsets
  - Define the decision rules
- Choose your output.

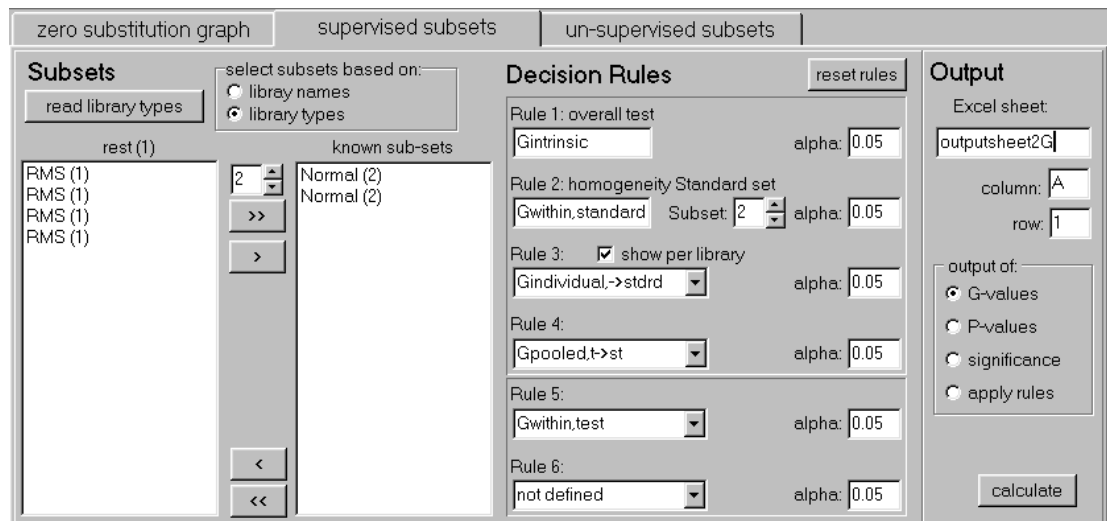
Define subsets

- With the radio buttons in the Subsets part of the window select 'library types'
- Press the read library types button
- In the list on the left, select all 'normal muscle' libraries and press the right arrow button

These libraries are transferred to the list on the right and are now defined as subset 2, the standard set. The remaining libraries stay in the list on the left and form together the subset 1 and constitute the test set.

Define decision rules.

- Make sure subset 2 is defined as standard set in Rule 2.
- Leave the decision rules as they are, you can experiment with them later if time allows.



**Figure 13.** Interface of the G-test program for defining a supervised comparison of subsets .

Choice of output.

- Change the output sheet name to “outputsheet2G”
- Choose ‘output of G-values’ with the radio buttons

You have now given all required input.

- Press calculate

The program now compares for each tag the test subset with the standard set and writes the G-statistics of this comparison to Excel.

- Do the same for the P-values, significances and apply rules choices and send the output to outputsheet2P, outputsheet2S, and outputsheet2AR, respectively.

- Switch to Excel

What information is displayed in the header of the output?

Which columns in the output refer to the different decision rules?

On first glance the output is confusing. Therefore concentrate first on the 'apply rules' output. For each tag you see a series of 1's and 0's: a 1 means that the tag meets the rule. The first two rules ( $G_{\text{intrinsic}}$  and  $G_{\text{within,standard}}$ ) always have to be satisfied. When also rules 3 thru 5 are met you can conclude that the tag is differentially expressed in the test subset, compared to the standard set.

To find out why for a certain tag a rule is not met, look at the other output sheets for the G, and P values or look at the output per library. The G and P values per library give information on the individual deviations of each library from its subset.

What is the relation between G-values and P-values?

What is the relation between P-values and the significance?

What is the relation between the significance and the 'rules applied' output?

Why does the latter relation differ per rule?