

# Differential Expression in SAGE: Accounting for Normal Between-Library Variation

**Keith A. Baggerly\***

Department of Biostatistics  
UT M.D. Anderson Cancer Center  
Houston, TX 77030-4009

**Li Deng**

Department of Statistics  
Rice University  
Houston, TX 77005

**Jeffrey S. Morris**

Department of Biostatistics  
UT M.D. Anderson Cancer Center  
Houston, TX 77030-4009

**C. Marcelo Aldaz**

Department of Carcinogenesis  
UT M.D. Anderson Cancer Center  
Houston, TX 77030-4009

October 24, 2002

**Running Head:** Between-Library Variation in SAGE

\* To whom correspondence should be addressed:

Department of Biostatistics  
1515 Holcombe Blvd, Box 447

Houston, TX 77030-4009

Phone: (713) 745-5994

Fax: (713) 745-4940

Email: kabagg@mdanderson.org

## **Abstract**

### **Motivation**

In contrasting levels of gene expression between groups of SAGE libraries, the libraries within each group are often combined and the counts for the tag of interest summed, and inference is made on the basis of these larger “pseudolibraries”. While this captures the sampling variability inherent in the procedure, it fails to allow for normal variation in levels of the gene between individuals within the same group, and can consequently overstate the significance of the results. The effect is not slight: Between-library variation can be hundreds of times the within-library variation.

### **Results**

We introduce a beta-binomial sampling model that correctly incorporates both sources of variation. We show how to fit the parameters of this model, and introduce a test statistic for differential expression similar to a two-sample t-test.

### **Availability**

Matlab code for fitting the model is available from the first author.

### **Contact**

[kabagg@mdanderson.org](mailto:kabagg@mdanderson.org)

## Introduction

Most methods currently advanced for assessing differential expression in SAGE address the case where one library is contrasted with another, assuming a null hypothesis that there is no difference between the libraries being compared. Under this assumption, the chance of a single tag falling in one library or the other is roughly proportional to the library size. Differing approximations lead to modelling this behavior with binomial (Zhang et al., 1997) or Poisson distributions (Madden et al., 1997), normal approximations (Madden et al., 1997; Kal et al., 1999; Michiels et al., 1999; Man et al., 2000) or simulations involving permutation tests (Zhang et al., 1997). Bayesian approaches have been suggested by Audic and Claverie (1997), and by Chen et al. (1998) (the latter method was adapted by Lal et al. (1999) to accommodate unequal library sizes). Of these, the simulation approach of Zhang et al. (1997) and the Bayesian approach of Lal et al. (1999) (see also Lash et al. (2000)) are probably the most widely used, due to their implementation in easily accessible software (the SAGE 2000 software available from the Kinzler lab at Johns Hopkins, and the routine implemented in SAGEmap at the NCBI, respectively). As noted in the comparison conducted by Man et al. (2000) on several of the above methods, however, very similar results are obtained when the numbers of tags are large ( $> 20$ ); the authors contend that a normal approximation (Kal et al., 1999) or equivalently a chi-squared test has more power for detecting differences when the numbers of tags are small ( $< 15$ ). The validity of small tag comparisons is, however, questionable due to the presence of sequencing errors (Stollberg et al., 2000) though this may be somewhat ameliorated if there is some external measure of quality for the read, such as a *phred* score (Margulies and Innis, 2000; Margulies et al., 2001; Ewing et al., 1998). We note that the statistic suggested by Kal et al. (1999) is

$$Z = \frac{p_A - p_B}{\sqrt{\frac{p_0(1-p_0)}{N_A} + \frac{p_0(1-p_0)}{N_B}}}, \quad p_0 = \frac{X_A + X_B}{N_A + N_B}.$$

where the two library sizes are  $N_A$  and  $N_B$  and the two counts are  $X_A$  and  $X_B$ , respectively; the statistic we propose below will have a similar form.

When the number of libraries involved is more than two, so that there are two groups of libraries being compared, the most common approach is to reduce the number of effective libraries to two by pooling the libraries of like type, and reverting to the two-library comparison form. This is not universal, and cautionary statements have been made. Lash et al. (2000) recommend checking for a low coefficient of variation before applying the SAGEmap procedure to grouped libraries. Ryu et al. (2002) use a series of filters to deal with groups of pancreatic libraries, the first of which is a two-sample t-test applied to the proportions.

These pooling approaches do catch differences, but they can overstate the significance of the results by ignoring the role of normal variation in expression levels between like

samples. As an example, we consider the case of a single prevalent tag in eight breast libraries that have been assembled in the Aldaz laboratory. This tag, AGGTCAGGAG, has multiple matches and hence is not immediately biologically informative, but it will serve to illustrate the point. All of these libraries are derived from breast tumors; the first five are from patients found to be lymph node positive (LN+), and the remaining three from patients found to be lymph node negative (LN-). The tag counts and proportions in the various libraries are given in Table 1. If we combine the five LN+ libraries and three LN- libraries and compare the resulting tag proportions, we are comparing 434/404105 to 799/284968, for which the  $\chi^2_1$  value (Michiels et al., 1999; Man et al., 2000) is 279.9752; the 95% cutoff for this distribution is 3.84, so this is obviously a “significant” result. The equivalence of the above tests noted by Man et al. (2000) for high count tags means that the other tests will catch the same genes. Checking the sign of the test statistic proposed by Kal et al. (1999) suggests that this tag is more strongly expressed in LN- tumors. However, if we follow Ryu et al. (2002) and compare the 7 proportions using a two-sample  $t$ -test,  $t = (p_A - p_B)/\sqrt{V_A + V_B}$ , with  $p_A$  being the average of the 5 proportions in group A,  $V_A$  being the sample variance of these 5 proportions, and  $p_B$  and  $V_B$  likewise defined, we get a test statistic value of -1.3174. The 95% cutoffs for a  $t_6$  distribution are  $\pm 2.4469$ , so this is a decidedly “insignificant” result. While the mean proportion is higher for the LN- tumors, this is mostly being driven by results from a single library (10T) so that the variability within the LN- group is high. The first approach fails to take into account the variability between like libraries, shown in Figure 1, which the  $t$ -test correctly captures.

Table 1  
here.

Shifting between test types ( $\chi^2$  and two-sample  $t$ ) gives a different view of which tags are important, as can be seen in Figure 2 (a) where the values of the two tests are plotted against each other for the high-count tags (tags with total counts of 40 or more when summed over all 8 libraries). Some of the most extreme  $\chi^2$  values correspond to  $t$ -values of marginal significance and vice-versa. If the two tests were in accord, we would expect to see a rough “U-shape” when the points were plotted — large  $\chi^2$  values matched to large  $t$ -values of either sign. As the extreme range of the  $\chi^2$  values makes some of the structure more difficult to discern, we restricted the range of display in Figure 2 (b); here the general U-shape (and deviations from it) are more apparent. Even here, though, the U-shape is more compressed than we might expect. Most of the tags being flagged as significant by the  $\chi^2$  test (values  $> 5$ ) are *not* significant according to the  $t$ -test (absolute values  $< 2$ ).

Figure 1  
here

Figure 2  
here

Between-library variability within a group can often be as large in magnitude as the within-library variability due to sampling. While the two-sample  $t$ -statistic applied to the different normalized library proportions illustrates the problem, it is too crude a tool to provide a solution in and of itself: it weights the proportions from all libraries equally, even though the estimates from larger libraries are less variable, and it is possible for the sample variance of the normalized proportions to be less than the known within-library variation. This can result in inappropriately large  $t$  values as the denominator of the test statistic goes to zero. (Most such cases occur when the total tag count is low, so these are not apparent

in Figure 2 due to our filtering.) In effect, the  $t$ -test is going to the opposite extreme and focusing on the between-library variability at the expense of the within-library variability.

In order to properly capture both types of variation, a compromise is needed. We introduce a beta-binomial model that includes both types of variation in a hierarchical fashion: The proportion of a gene within a library is selected from an underlying beta distribution representing the normal between library variation, and the count within that library is binomial with the chosen proportion as a parameter. Depending on the parameters of the beta model, this leads to estimates for the group proportions and associated variances that weight the different library proportions using values intermediate between equal weighting (all variation is between libraries) and weighting proportional to the library size (all variation is within libraries).

## Methods

For the sake of notational simplicity, we will focus on the case of modeling the counts of a specific tag within the first group. Let  $n_i$  denote the total tag count in library  $i$  of this group, and let  $p_i$  denote the true proportion of the tag of interest within library  $i$ . Finally, let  $X_i$  denote the corresponding count for the tag of interest. For the first part of our model, we assume that the true proportions may vary from library to library. A standard distribution for proportions is the beta distribution, and we shall assume this here. The particular distribution used doesn't matter a great deal. The main point is that the distribution is not necessarily degenerate: it can have a positive variance. We'll be focusing on the first two moments of the various distributions throughout, both for computational simplicity and out of an intent to invoke the central limit theorem to get an approximately normal test statistic. Here,

$$p_i \sim \text{Beta}(\alpha, \beta), \quad E(p_i) = \frac{\alpha}{\alpha + \beta}, \quad V(p_i) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

The second part of our model says that given the true proportion in a sample, the corresponding count will have a binomial distribution with the true proportion as a parameter:

$$X_i | p_i \sim \text{Binomial}(n_i, p_i).$$

Some straightforward algebra (details available from the author) shows that the unconditional mean and variance (integrating out the unseen true proportions  $p_i$ ) of the estimated proportion  $\hat{p}_i = X_i/n_i$  are

$$E(\hat{p}_i) = \frac{\alpha}{\alpha + \beta}, \quad V(\hat{p}_i) = \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \left[ \frac{1}{\alpha + \beta} + \frac{1}{n_i} \right].$$

There are two components to the variance of the proportion  $\hat{p}_i$  (in square brackets above), and only one of them (the within library variation) decreases as the library size is increased.

Now, given that we know the variance of a single proportion, we turn to the mean and variance of a weighted linear combination of proportions to see how to combine the results from different libraries.

$$E\left(\sum w_i \hat{p}_i\right) = \sum w_i E(\hat{p}_i) = \frac{\alpha}{\alpha + \beta} \sum w_i = \frac{\alpha}{\alpha + \beta} \quad (1)$$

$$V\left(\sum w_i \hat{p}_i\right) = \sum w_i^2 \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \left[\frac{1}{\alpha + \beta} + \frac{1}{n_i}\right]. \quad (2)$$

As long as the weights sum to 1, the combination has the correct mean, so the focus shifts to choosing the weights so as to minimize the associated variance. This optimal combination can be found in a fairly straightforward fashion using Lagrange multipliers:

$$\begin{aligned} \frac{\partial}{\partial w_i} \left[ V\left(\sum w_i \hat{p}_i\right) + \lambda \left(1 - \sum w_i\right) \right] &= 2w_i \frac{\alpha\beta}{(\alpha + \beta)(\alpha + \beta + 1)} \left[\frac{1}{\alpha + \beta} + \frac{1}{n_i}\right] - \lambda = 0 \\ &\rightarrow w_i \propto \left[\frac{1}{\alpha + \beta} + \frac{1}{n_i}\right]^{-1}. \end{aligned}$$

At this point, we note that the optimal choice of weights is determined by a single relationship – the size of  $\alpha + \beta$  relative to  $n_i$ . If we consider the extremes of this type of arrangement, letting  $\alpha + \beta$  go to  $\infty$  implies both that the distribution of the  $p_i$ 's is degenerate, so that there is no change in the true proportion going from sample to sample, and that in this case the optimal weighting is proportional to the library size. If, on the other hand, the sum  $\alpha + \beta$  is very small relative to the  $n_i$  values, then the imprecision in our knowledge of the proportion in a given library is dwarfed by the imprecision due to library to library variability, and the optimal weights are roughly the same for all libraries. Thus, weighting by library size and weighting equally represent the two extremes, and the true optimum lies somewhere in between. Note that the optimum weighting may be different for different tags even if the same libraries are used!

Now, the form of the weighting vector gives us the estimated proportion for the group as

$$\hat{p} = \sum w_i \hat{p}_i.$$

Using the fact that the form of an expected squared proportion is

$$E(\hat{p}_i^2) = \mu^2 + \sigma_1^2 + \frac{\sigma_2^2}{n_i},$$

it can be shown that an unbiased estimator of the variance of this proportion is

$$\hat{V}_{unb} = \frac{\sum w_i^2 \hat{p}_i^2 - (\sum w_i^2) \hat{p}^2}{1 - (\sum w_i^2)}.$$

When all of the  $w_i$ 's are equal, this reduces to the standard unbiased estimator. This variance estimate is mostly right, but it can be too small — we know that the variance can never be less than the sampling variability. This lower bound follows in turn from the assumptions that the libraries are assembled independently, and that sampling within a library is also independent. These assumptions strike us as reasonable, and we make them here. Allowing for this lower bound suggests the modified estimator

$$\hat{V} = \max \left[ \hat{V}_{unb}, \frac{\sum \frac{X_i}{n_i} \left( 1 - \frac{\sum X_i}{\sum n_i} \right)}{\sum n_i} \right].$$

There are slightly different lower bounds that could be constructed, but they all have the same leading term,  $\sum X_i / (\sum n_i)^2$ . We revisit this point below.

In order to come up with a concrete number for a test statistic, we need to estimate the beta parameters. This can be done quickly using the method of moments, applied to the unweighted sample proportions; this procedure can then be iterated as the weights provide revised estimates of the parameters. Consider the case of the LN+ proportions in the example given earlier.

$$\begin{aligned} w_i^{(0)} &= \frac{n_i}{\sum n_i} \\ \hat{p}^{(0)} &= \sum w_i^{(0)} \hat{p}_i = 0.00107 \\ \hat{V}^{(0)} &= \frac{\sum (w_i^{(0)})^2 \hat{p}_i^2 - \left( \sum (w_i^{(0)})^2 \right) (\hat{p}^{(0)})^2}{1 - \left( \sum (w_i^{(0)})^2 \right)} = 7.99514e - 06 \\ \hat{\beta}^{(1)} &= \frac{\hat{p}^{(0)} (1 - \hat{p}^{(0)}) \sum (w_i^{(0)})^2 - \hat{V}^{(0)}}{\hat{V}^{(0)} (1 - \hat{p}^{(0)})^{-1} - \hat{p}^{(0)} \left( \sum (w_i^{(0)})^2 / n_i \right)} = 3184.0592 \\ \hat{\alpha}^{(1)} &= \frac{\hat{p}^{(0)}}{1 - \hat{p}^{(0)}} \hat{\beta}^{(1)} = 3.4233 \\ w_i^{(1)} &\propto \frac{\left( \hat{\alpha}^{(1)} + \hat{\beta}^{(1)} \right) n_i}{\hat{\alpha}^{(1)} + \hat{\beta}^{(1)} + n_i} \propto (0.2051, 0.2048, 0.2045, 0.2047, 0.1808). \end{aligned}$$

(The expressions for  $\hat{\beta}$  and  $\hat{\alpha}$  were found by manipulating equations 1 and 2 to isolate the parameters as functions of the moments.) Empirically, convergence is quite rapid, as is shown for this example in Table 2.

Table 2  
here.

Here, the size of the sum of the beta parameters (about 3K) relative to the library sizes (about 100K) suggests that the between-library variability is roughly 30 times the within-library variability. Special note needs to be made of the case where the method of

moments fails: When the variability of the sample proportions is less than that known to be present due to sampling variability. In this case, it is instructive to look at the likelihood function. The likelihood function shows a ridge; the ratio  $\alpha/(\alpha + \beta)$  (corresponding to the mean proportion) is well characterized, but the sum  $\alpha + \beta$  diverges to  $\infty$  if we attempt to find a maximum. In this case, the underlying maximum likelihood beta distribution is a degenerate point mass, suggesting that the proper course of action is to ignore the between library variability and work just with the within library variability. This is precisely when we shift between different estimates of variance above; consequently the estimates do not become more precise (the variance doesn't drop below our working floor) as we attempt to account for additional variability.

The test statistic that we propose for comparing groups  $A$  and  $B$ , then, is

$$t_w = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{V}_A + \hat{V}_B}}$$

with  $\hat{p}$  and  $\hat{V}$  as defined above. For testing significance, it is useful to be able to specify the null distribution of a test statistic. This is somewhat difficult here in that the distribution of the above statistic depends on the relative sizes of the between and within variation. If the within-library variation is predominant, then the shape of the distribution is largely driven by the total counts within each group, and if these counts are reasonable (say 15 or more in each group) then the binomial distribution will be roughly normal and a  $Z$  distribution can be used. This follows from an appeal to the central limit theorem, the rough bell-shape of the binomial distribution, and the fact that the degrees of freedom used for estimating this component of the variation are very close to the total number of counts. Small counts in each group should force us to account for the asymmetric nature of the underlying distribution more directly, and in this region a test such as that proposed by Audic and Claverie (1997) seems reasonable. If, however, the between-library variation is dominant, then the dominant effect in many cases will be the small number of different libraries used to estimate this variance; in this case we are driven to a  $t$  distribution with  $n_A + n_B - 2$  degrees of freedom.

## Discussion

Between-library variability is nonnegligible for SAGE data. Indeed, for the higher count data, the between-library variability is the dominant part of the variation. We can see this for the LN+ group, by plotting the total (between + within) library variability as a multiple of the within library variability, with both quantities on the  $\log_2$  scale to make the structure more apparent. This is shown in Figure 3. The between-library variance is about the same size as the within-library variance at about a total count of 16; for the high count tags the between variance clearly wins. The worst case involves a ratio of nearly  $2^{9.8}$

Figure 3  
here.

or roughly 900-fold. Similar qualitative results hold for the LN- group (not shown). The final distribution of the test statistic values comparing LN+ and LN- is shown in Figure 4, with the counts  $\log_2$ -transformed to make the structure more apparent. The distribution appears largely stable as a function of tag count, so larger counts are not getting “more significant” just by default. The most extreme count tags (including our example tag) no longer appear as significantly different. There is some granularity at the low counts where the Normal approximation is breaking down.

Figure 4  
here.

We have proposed a method that accounts for between-library variability in addition to the within-library variability already well-treated by previous methods; indeed, our method reduces to that of Kal et al. (1999) in the special case when between-library variability can be neglected. Unlike the  $t$ -test, our method explicitly incorporates differences in library sizes when dealing with multiple proportions (note that one of our libraries is much smaller than the rest). There are, however, some additional points to note.

Our approach works one tag at a time. It may be possible to improve inference further by working with an ensemble approach that attempts to estimate parameters for all of the genes at once. This is the potential of “borrowing strength” across genes to improve estimates throughout. Such borrowing has been used to good effect in estimating variances associated with microarray readings (see Baggerly et al. (2001), Newton et al. (2001), and Hughes et al. (2000), among others). In the microarray context, this borrowing was achieved by grouping genes according to intensity. Likewise, SAGE tags could be grouped according to relative abundance and estimation of the between-library variation assessed for the group.

Our approach uses a beta-binomial model. Other models are of course possible. A natural alternative is the gamma-Poisson model, with

$$\begin{aligned} X_i &\sim \text{Poisson}(k_i \lambda_i) \\ \lambda_i &\sim \text{Gamma}(\alpha, \beta) \end{aligned}$$

where  $k_i$  is a rate parameter that adjusts for the differences in library sizes; if we were defining counts in terms of observations per 50K tags, and we had 3 libraries of sizes 10K, 40K, and 100K, the  $k_i$ 's would be 0.2, 0.8, and 2 respectively. Checking the moments,

$$\begin{aligned} E(X_i/k_i) &= \frac{\alpha}{\beta} \\ V(X_i/k_i) &= \frac{\alpha}{\beta} \left[ \frac{1}{\beta} + \frac{1}{k_i} \right], \end{aligned}$$

so again the variance has two components, only one of which changes with library size. A similar weighting process ensues and the results are qualitatively similar. Quantitatively, the lower bound for the variance (when scaled) reduces to the leading term for the beta-binomial variance found earlier.

Our approach dodges the small-variance problem by reverting to using just the sampling variance when the other estimate gives a value that is too small on its face. It is possible to treat this in a more rigorous and coherent fashion using a full-blown Bayesian approach that addresses the uncertainty in our estimates of  $\alpha$  and  $\beta$  by simulating draws from the posterior distribution (see Gelman et al. (1995) p.130ff for a discussion of how this might be done here). This approach, however, requires the additional specification of a prior and significant computational overhead (several orders of magnitude beyond that required here). An additional complexity is that a completely noninformative prior can lead to an improper posterior in this context. For genes of particular interest, the freely available BUGS software may be able to provide this type of approach without the need for much coding on the user’s part. We are exploring this.

The special case where each group contains just one library can be treated by setting the weight  $w_1$  to 1 in each group. As the within-group variance is smaller than the sampling variability, this would default to using the sampling variability only. This weighting approach suggests a difficulty with one vs. one comparisons if the between-library variability is suspected to be large. When we have just one library in each group, the degrees of freedom in our  $t$ -statistic formulation drop to zero, reflecting the fact that any differences we see could be due to either the biological change of interest or to normal variability, but we can’t tell which without an assessment of this variation. Useful inferences in this case rely on prior assumptions about the scale of change to be expected or on implicitly borrowing strength across genes by looking at which ones are “the most different”. This in turn treats the experimental results as supplying a ranking of interest rather than a straight significance value. This ranking viewpoint is reasonable in light of the multiple testing problems inherent in checking thousands of genes.

Finally, our approach treats all of the libraries within a group as similar enough that the observed variation can be described as “normal variation within the population of interest”. There may be additional known covariates that could account for much of this if they were included in modeling the data, but this leads to a more involved assessment of the variance structure, with pieces going to each of these included factors. Our approach provides perhaps the simplest way of incorporating between-library variability from a host of sources, and with the variance bound we have imposed is inherently at least as conservative as other tests. The number of genes identified as differentially expressed will drop using this method, but the false positive rate will drop at least as much.

## Acknowledgements

The authors gratefully acknowledge support from NIH-NCI Grant 1U19 CA84978-1A1.

## References

- Audic, S. and Claverie, J.-M. 1997. The significance of digital gene expression profiles. *Genome Research* 7,986–995.
- Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V., and Zhang, W. 2001. Identifying differentially expressed genes in cDNA microarray experiments. *Journal of Computational Biology* 8(6),639–659.
- Chen, H., Centola, M., Altschul, S. F., and Metzger, H. 1998. Characterization of gene expression in resting and activated mast cells. *Journal of Experimental Medicine* 188(9),1657–1668.
- Ewing, B., Hillier, L., Wendl, M. C., and Green, P. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research* 8,175–185.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. 1995. *Bayesian Data Analysis*. Chapman and Hall.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraborty, K., Simon, J., Bard, M., and Friend, S. H. 2000. Functional discovery via a compendium of expression profiles. *Cell* 102,109–126.
- Kal, A. J., van Zonneveld, A. J., Benes, V., van den Berg, M., Koerkamp, M. G., Albermann, K., Strack, N., Ruijter, J. M., Richter, A., Dujon, B., Ansorge, W., and Tabak, H. F. 1999. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Molecular Biology of the Cell* 10,1859–1872.
- Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. 1999. A public database for gene expression in human cancers. *Cancer Research* 59,5403–5407.
- Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J., and Altschul, S. F. 2000. SAGEmap: A public gene expression resource. *Genome Research* 10,1051–1060.
- Madden, S. L., Galella, E. A., Zhu, J., Bertelsen, A. H., and Beaudry, G. A. 1997. SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* 15,1079–1085.
- Man, M. Z., Wang, X., and Wang, Y. 2000. POWER\_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* 16(11),953–959.
- Margulies, E. H. and Innis, J. W. 2000. eSAGE: managing and analyzing data generated with serial analysis of gene expression (SAGE). *Bioinformatics* 16(7),650–651.
- Margulies, E. H., Kardia, S. L. R., and Innis, J. W. 2001. A comparative molecular analysis of developing mouse forelimbs and hindlimbs using serial analysis of gene expression (SAGE). *Genome Research* 11,1686–1698.

- Michiels, E. M. C., Oussoren, E., van Groenigen, M., Pauws, E., Bossuyt, P. M. M., Voûte, P. A., and Baas, F. 1999. Genes differentially expressed in medulloblastoma and fetal brain. *Physiological Genomics* 1,83–91.
- Newton, M. A., Kendzierski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 8(1),37–52.
- Ryu, B., Jones, J., Blades, N. J., Parmigiani, G., Hollingsworth, M. A., Hruban, R. H., and Kern, S. E. 2002. Relationships and differentially expressed genes among pancreatic cancers examined by large-scale serial analysis of gene expression. *Cancer Research* 62,819–826.
- Stollberg, J., Urschitz, J., Urban, Z., and Boyd, C. D. 2000. A quantitative evaluation of SAGE. *Genome Research* 10,1241–1248.
- Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. 1997. Gene expression profiles in normal and cancer cells. *Science* 276,1268–1272.

Name	1T+	3T+	4T+	6T+	8T+	7T-	9T-	10T-
Tag Count	129	167	71	61	6	43	247	509
Library Size	100474	96631	92510	95785	18705	95155	91593	98220
Proportions (%)	0.13	0.17	0.08	0.06	0.03	0.05	0.27	0.52

Table 1: Counts and proportions of tag AGGTCAGGAG in 8 breast tumor libraries, 5 lymph node positive and 3 lymph node negative.

$i$	1	2	3	4	5
$\alpha^{(i)}$	3.4233	2.8977	2.9039	2.9036	2.9036
$\beta^{(i)}$	3184.0592	3007.2053	3015.9134	3015.4784	3015.5001

Table 2: Convergence of moment-based estimates of the beta parameters for the LN+ values given in Table 1.

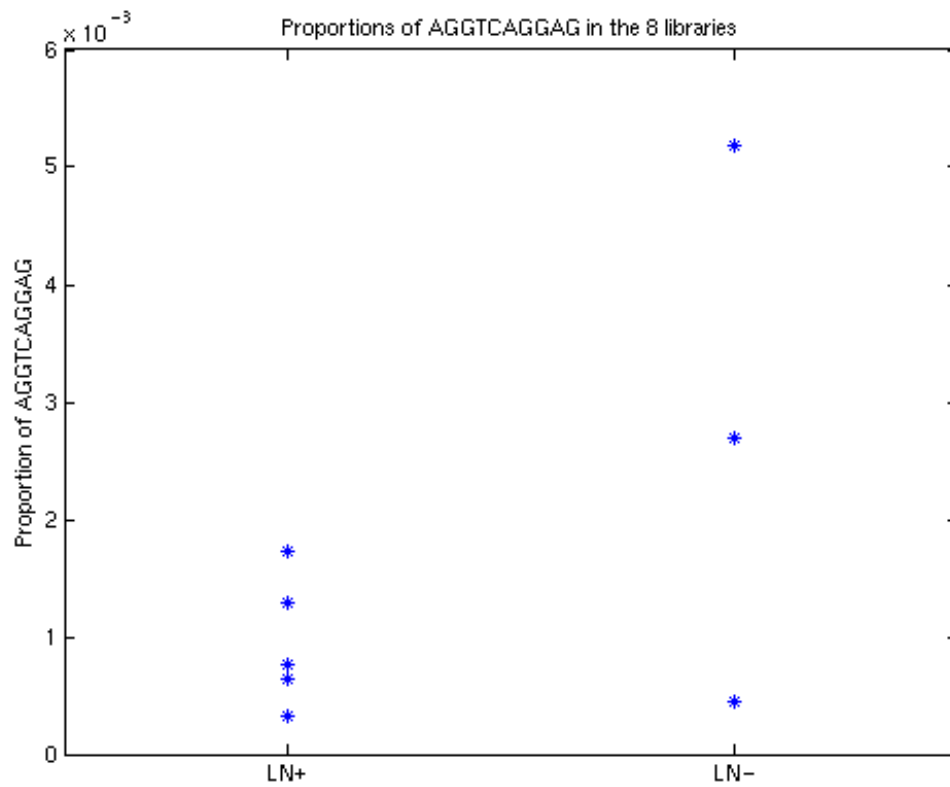


Figure 1: Tag counts of AGGTCAGGAG as a proportion of total library size for the LN+ tumors and the LN- tumors. While the mean levels between LN+ and LN- are different, this difference is not significant given the level of count variability within LN status, in particular the wide spread in counts between the most extreme LN- values.

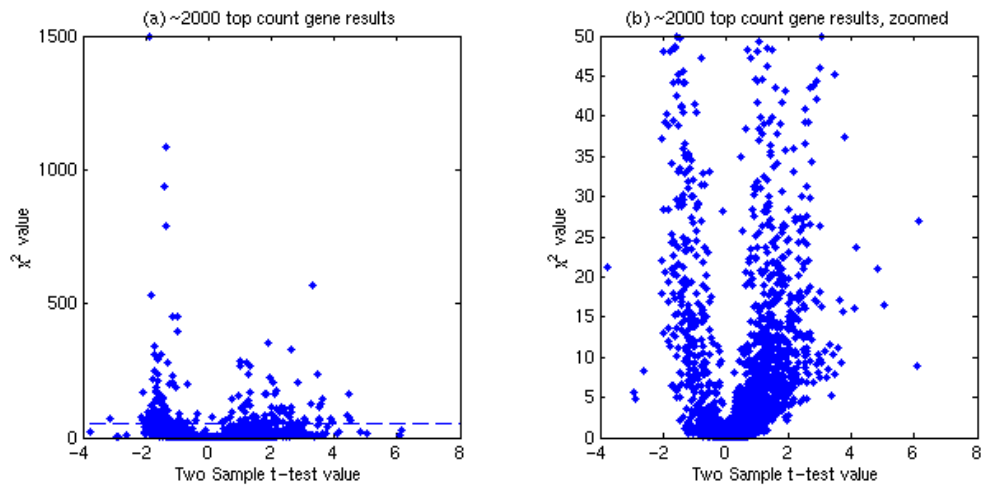


Figure 2: (a) Two-sample  $t$ -test and  $\chi^2$  values for all high count tags (40 or more total counts across all 8 libraries). If the two tests were in agreement, we would see a “U-shape” corresponding to genes being found equally extreme by both. Here, some of the most extreme values by one test (large chi-square values, or large absolute  $t$ -test values) are associated with at best weakly significant values of the other. (b) Zoom on points with chi-squared values below 50, indicated with a dashed line in (a). While the U-shape is clearly evident, it is more compressed than agreement would indicate. Most of the tags being flagged as significant by the  $\chi^2$  test (values  $> 5$ ) are *not* significant according to the  $t$ -test (absolute values  $< 2$ ).

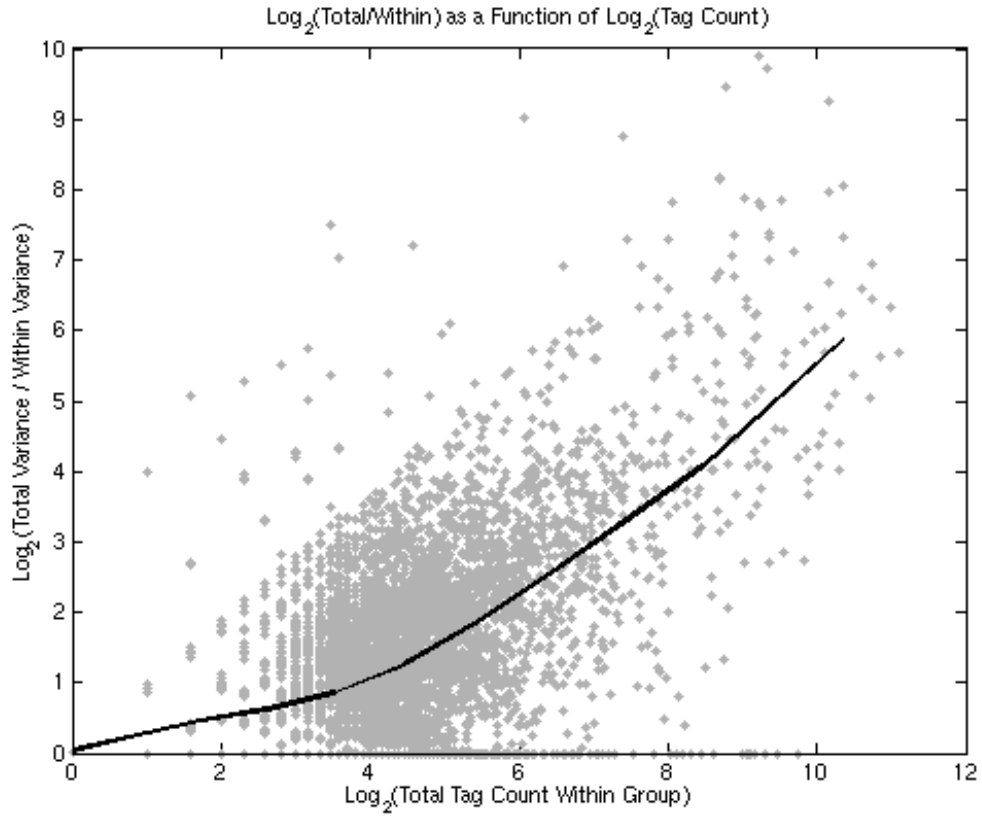


Figure 3: The  $\log_2$  ratio of total (within plus between) variation to within variation as a function of  $\log_2$  of the total tag count for the LN+ group. The smooth line was fit by binning along the x-axis one unit at a time, taking the mean  $(x, y)$  point within that bin, and fitting a loess smooth with span 5 to the resultant 12 points. Note the line crosses 1 (between is equal to within) at about 4 on the x-axis, corresponding to a raw count of 16. For larger count tags, the between-library variation is clearly dominant, with the biggest multiple being about  $2^{9.8}$ , or roughly 900-fold.

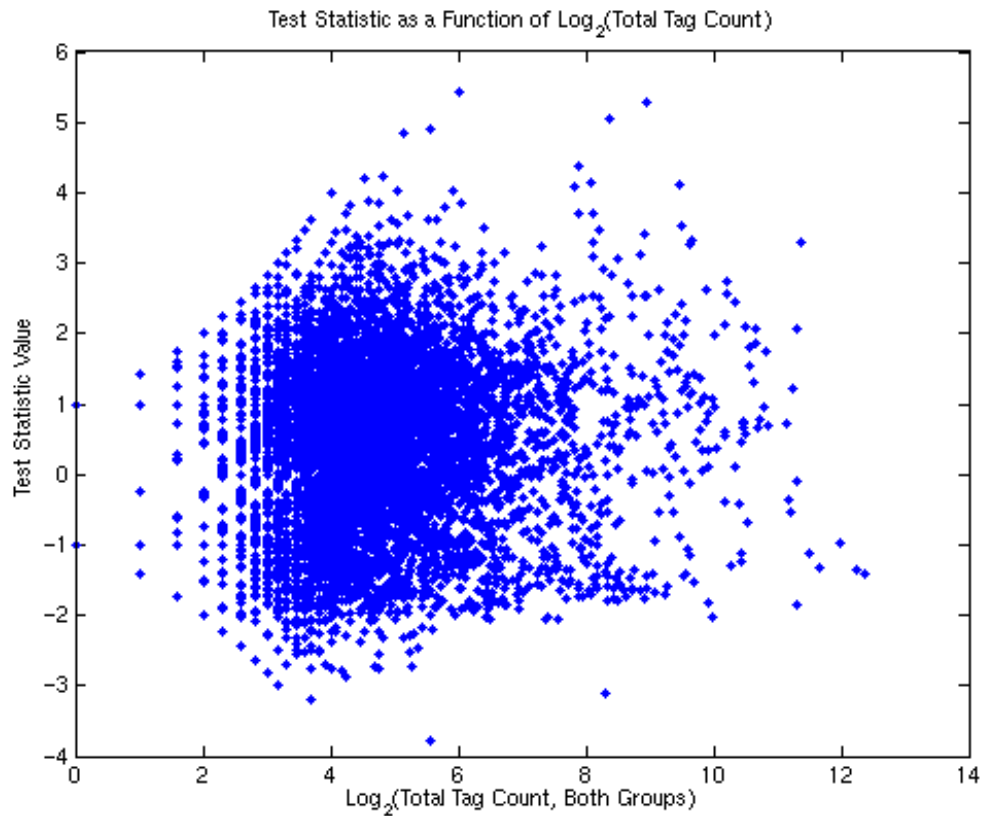


Figure 4: Our test statistic as a function of  $\log_2$  tag count for the LN+ vs LN- comparison. Unlike the  $\chi^2$  test, this test is not overly biased towards giving larger values to larger count tags. The distribution appears roughly uniform, with granularity creeping in at the low counts as the normal approximation fails.