omputational and Statistical Approaches to Genomics

econd Edition

ei Zhang and Ilya Shmulevich

*mputational and Statistical Approaches to Genomics Second Edition* aims to p researchers deal with current genomic challenges. During the three years after the blication of the first edition of this book, the computational and statistical research in nomics has become increasingly important and indispensable for understanding llular behavior under a variety of environmental conditions and for tackling challenging ical problems. In the first edition, the organizational structure was: data à analysis à nthesis à application. In the second edition, the same structure remains, but the apters that primarily focused on applications have been deleted.

s decision was motivated by several factors. Firstly, the main focus of this book is mputational and statistical approaches in genomics research. Thus, the main emphasis on methods rather than on applications. Secondly, many of the chapters already lude numerous examples of applications of the discussed methods to current problems biology.

e range of topics has been broadened to include newly contributed chapters on topics ch as alternative splicing, tissue microarray image and data analysis, single nucleotide lymorphisms, serial analysis of gene expression, and gene shaving. Additionally, a mber of chapters have been updated or revised.

*mputational and Statistical Approaches to Genomics Second Edition* is for searchers in both academia and industry involved with genomic problems in fields such biology, computer science, statistics, and engineering. It can also be used as an vanced level textbook in a course focusing on genomic signals, information processing, genome biology.

ISBN 0-387-26287-3

ringer.com

Zhang / Shmulevich

Computational and Statistical Approaches to Genomics
Second Edition

Wei Zhang and Ilya Shmulevich

# Computational and Statistical Approaches to Genomics

## Second Edition

Springer

## 2. Biology and Bioinformatics Background

The SAGE method, described by Velculescu *et al.* (1995), is based on the isolation of unique sequence tags from individual transcripts. SAGE counts transcripts by sequencing a short 14-bp tag at the gene's 3' end adjacent to the last restriction enzyme site, usually *NlaIII*. The 4 bases of the 3'-most *NlaIII* restriction site (CATG) plus the following 10 variable bases define the transcript tag. After the concatenation of tags into long DNA molecules, sequencing of these concatemer clones allows quantification and identification of cellular transcripts. When dealing with SAGE data, it is important to bear in mind that this technology yields transcript counts, expressed as a fraction of the total amount of transcripts counted, and not results relative to another experiment or a particular housekeeping gene, such as in hybridization-based techniques. This advantage avoids error-prone normalization between experiments. Another advantage is that it determines expression levels directly from RNA samples and it is not necessary to have a gene-specific fragment of DNA arrayed to assay each gene. That is why SAGE is called an open system, justifying the analogy "Linux of the functional genomics" (SAGE2000 Conference, September 2000).

Transcript tags were extracted from the automatic sequencing machines output files, the chromatograms, using specific softwares. The steps are as follows (Lash *et al.*, 2000):

1 locate the *NlaIII* sites (i.e., CATG "punctuation signals") within the ditag concatemer;

2 extract ditags of 20-26 length bases, which fall between these sites;

3 remove repeat occurrences of ditags, including repeat occurrences in the reverse-complemented orientation;

4 define tags as the end-most 10 bases of each ditag, reverse-complementing the right-handed tag;

5 remove tags corresponding to linkers (e.g., TCCCCGTACA and TCCCTATTAA), as well as those with unspecified bases;

6 for each tag, count its number of occurrences.

A relevant problem in Bioinformatics is to assign the observed tags to its correspondent gene. Tag-to-gene mapping is accomplished by first orientating GenBank sequences using poly-adenylation signal (ATTAAA or AATAAA), poly-adenylation tail (minimum of 8 A's) and orientation annotation (3' or 5'). Tags for genes are defined from the 10-base sequence directly 3' adjacent to the 3'-most *NlaIII* site (CATG), and then linked to an UniGene cluster identifier. UniGene is a system for automatically partitioning GeneBank sequences, including ESTs, into a non-redundant set of gene-oriented clusters. Each UniGene cluster theoretically contains sequences that represent a unique gene (Schuler, 1997). When extracting tags, 5 arbitrary reliability "classes" are defined for tag-to-UniGene assignment, following the reliability order:

1 derived from well-characterized mRNA's cDNA sequences from GeneBank;

2 tags extracted from EST sequences with a polyadenylation signal and/or polyadenylation tail and annotated as 3' sequences;

3 derived from EST sequences with a polyadenylation signal and/or polyadenylation tail, but without a 3' or 5' annotation;

4 from EST sequences with a polyadenylation signal and/or polyadenylation tail, but annotated as 5' orientation;

5 from EST sequences without a polyadenylation signal or tail but annotated as having a 3' orientation.

For each tag, two other quality parameters are calculated: (i) the gene-to-tag assignment frequency (how many different genes, from the total of unique genes in the library, the tag is the best-match tag) and (ii) the tag-to-gene assignment frequency (how many different tags, from the total of unique tags in the library, the gene is the best-match gene).

The Cancer Genome Anatomy Project SAGE Project made available new Bioinformatics tools taking into account measurable reliability in tag-to-gene matching. This is accomplished in three major steps: (i) a confident tag list was distilled from ~6.8 million experimentally observed SAGE tags; (ii) virtual SAGE tags (predicted from cDNA transcript sequences) were obtained, parsed into databases reflecting the origin of the transcript sequence and ranked according to it; and (iii) custom programs were created to sift through the databases, choose the best tag-to-gene match, and present results online. Alternative transcripts, redundant tags, and internal priming were also considered for tag selection. This is the SAGE Genie interface (http://cgap.nci.nih.gov/SAGE) (Boon *et al.*, 2002).

Ideally, these tags are long enough to be unique to one transcript, and the abundance of a given tag is assumed to be proportional to the expression level of that transcript in the original pool of RNA. However, SAGE is a sampling method: some transcripts present in low abundance may be missing, and the number of copies of others may not accurately reflect

their true abundance in cells due to selection bias (Margulies *et al.*, 2001). In addition, there are errors in sequencing, possibility of non-unique tags or transcripts that produce no tag. A small number of transcripts are expected to lack the *NlaIII* enzyme anchoring site and therefore would be missed in the analysis. Transcripts that failed to be represented due to the lack of restriction enzyme anchoring site are estimated to be as low as 1% (Boon *et al.*, 2002). It is interesting to note that about 30% of the full length transcripts have a repetitive sequence inside its sequence. SAGE tags that fall inside repetitive elements will be counted many times for different transcripts and their frequency will be very high, compromising the gene-to-tag assignment (see web-Fig. 1). In the mouse SAGE web-site (http://mouse.biomed.cas.cz/sage/), for example, the tags with reliable associations to 12 or more UniGene clusters were labeled as "repetitive/low-complexity" to be easily distinguished. These problems, in time, may be disregarded by the increasing number of SAGE tags collected for future SAGE screens, and the use of longer SAGE tags or different anchoring enzymes.

## 3. Estimation

The objective of the *Estimation* process is to obtain numerical values for the unknown quantities in a sample or a population. The estimation is, along with hypothesis testing, a fundamental task of Statistics and is divided into *Point Estimation* and *Estimation by Interval*.

*Point Estimation* attributes the best possible scientific guess for a numerical unknown quantity. The Frequentist approach for Statistics claims that the "best" estimators have certain properties such as being non-biased, for example. The Bayesian radically disagrees since estimators are always biased by the *a priori* knowledge.

*Estimation by Interval* is expected to find numerical intervals that contain the unknown quantity, attaching some probability to this fact. An interval is also viewed with a very different interpretations depending if it is a Frequentist or a Bayesian approach. In the Frequentist view, obtaining an $\alpha\%$ *confidence interval* means just the application of a calculation procedure to our data that yields a numerical interval. This procedure claims that the calculated intervals, applied to virtual data, should contain the true parameter's value in at least $\alpha\%$ of the times that one uses it. This virtual data is assumed to be created by the same underlying *probability density function* (pdf) that created the observed data. This does not mean that the probability that the true value is contained in the interval is $\alpha$. In the Frequentist framework, the true value is a number and not a random variable, not allowing such a probabilistic assertion. However,



this kind of desired assertion is possible using a Bayesian framework. The Bayesians do not believe in the "data that could be observed but was not" (virtual data) concept and base their conclusions on the parameter's *a posteriori probability density function*. The probability that a parameter's unknown value is inside an interval can be calculated for any numerical interval. To match the intuitive notion of "error-bar", one calculates the smaller interval around *posteriori* pdf maximum that integrates $\alpha$ probability. This is the Bayesian *credibility interval*.

### 3.1 Point Estimation (Counts, Errors, Size)

The main tasks of *Point Estimation* are estimation of: error-rates of sequencing process, tag's counts, tag's abundance (also known as normalized counts, proportion, concentration, etc) and transcriptome size.

Transcriptome's size and distribution estimation, i.e., to find out how many different transcripts are expressed in a given cell condition and how transcripts are distributed among the expression levels, are very important problems. We let these problems aside because they are well discussed in Chapter 10. We remind the reader that transcripts' distribution is very skewed toward zero, i.e., there are several transcripts with low abundance and few highly abundant transcripts. Along with careful reading of Chapter 10, we recommend the works of Stollberg *et al.* (2000) and Stern *et al.* (2003) that show, by simulation approaches, the ill-posed nature of these kinds of estimation problems with the usual size of SAGE libraries. Perhaps these problems could be resolved in the near future by using alternative technologies such as Massively Parallel Signature Sequencing (MPSS) (Brenner *et al.*, 2000).

In the following, we will discuss error-rates estimation, the first issue to be considered in a SAGE statistical analysis pipe-line.

One can imagine that tag counting is clearly defined by sequencing machine output and it is sufficient to count the sequences identified in the chromatograms. However, the SAGE sequencing itself is subjected to stochastic processes like enzyme amplification errors or sequencing base miscalling. Therefore, a tag outcome could be modeled and estimated. The estimation of the error-rates are relevant issues.

The number of sequencing errors in a 10-base tag, assuming that base miscalling is equally probable for all nucleotides and independent of position inside the sequence, follows a Binomial(10,$\varepsilon$) distribution where $\varepsilon$ is the error-rate in errors per base units. The relevant quantities are $\mathbf{P}(\text{just 1 error}) = 10 \cdot \varepsilon^1(1-\varepsilon)^9$, $\mathbf{P}(\text{errors} \geq 1) = 1 - (1-\varepsilon)^{10}$ and $\mathbf{P}(\text{errors} \geq 2) = \mathbf{P}(\text{errors} \geq 1) - \mathbf{P}(\text{just 1 error})$. The first estimate of $\varepsilon$ came from the work of Velculescu *et al.* (1997), since they compared their

SAGE results with the yeast's complete sequenced genome. They found $\varepsilon \cong 0.007$ or $\mathbf{P}(\text{errors} \geq 1) \cong 0.068$. The use of 1% approximation for the error-rate is very common in SAGE analysis, imported from genome and EST sequencing projects, because it corresponds to a widely used quality score cutoff of 20 in the well-known *phred* sequencing chromatogram analysis software (Ewing and Green, 1998).

Colinge and Feger (2001) worked with $\varepsilon = 0.01$ and thus $\mathbf{P}(\text{errors} \geq 1) = 0.096 \gg \mathbf{P}(\text{errors} \geq 2) = 0.004$. This fact suggests that the majority of sequencing errors should be a single base change. A tag may have *neighbor* tags due to nucleotide substitution, insertion and deletion, i.e., tags that differ from each other by only one application of such editing operations. They constructed a transition matrix with $\{P\}_{ik} = p(j|k)$ elements as the probability that the observed $k$-th tag is counted as the observed $j$-th tag due to one-step substitution sequencing error. Note that a rare tag with no single-substitution distant *neighbor* tag has $p(j|j) = 1$, in spite of having $\mathbf{P}(\text{no errors}) = (1-\varepsilon)^{10} = 0.904$. A more refined approximation would consider insertions and deletions in the path from tag $k$ to tag $j$:

$$\begin{aligned} p(j|k) = \mathbf{P}(j \leftarrow k) = {} & \mathbf{P}(I) + \mathbf{P}(D) + \mathbf{P}(S) \\ & - \mathbf{P}(I)\mathbf{P}(D) - \mathbf{P}(I)\mathbf{P}(S) - \mathbf{P}(D)\mathbf{P}(S) \\ & + \mathbf{P}(I)\mathbf{P}(D)\mathbf{P}(S) \end{aligned} \tag{11.1}$$

where $I$, $D$ or $S$ are the transitions $\{j \leftarrow k\}$ due to one-step insertion, deletion or substitution errors, respectively.

Blades *et al.* (2004a) devised a simple procedure to estimate these first-order error-rates from each library. They call *shadows* the tags generated by process artifacts. Cleverly, they write the error-rate as the increase of *shadows* counts $s$ obtained with the increasing of true counts $x$ for a tag:

$$\varepsilon = \frac{\Delta s}{\Delta x} = \frac{\Delta s}{\Delta y + \Delta s} = \frac{1}{(\Delta s/\Delta y)^{-1} + 1} = \frac{1}{\theta^{-1} + 1} \tag{11.2}$$

where $y$ are the observed counts.

If all *neighbors* were *shadows*, we would be able to estimate $\theta$ as the angular coefficient of usual regression in observed counts vs. *neighbors* counts space. However, this is not true and therefore, we need to identify in this space a subset of $(y_j, s_j)$ points for which this assumption seems reasonable. Due to the highly skewed form of the transcriptome towards rare transcripts, one expects that this trend is only visible next to the region of abundant tags, i.e., higher $y_j$ points. Also, the error rate is expected to be small, thus one expects a regression line close to a horizontal line.

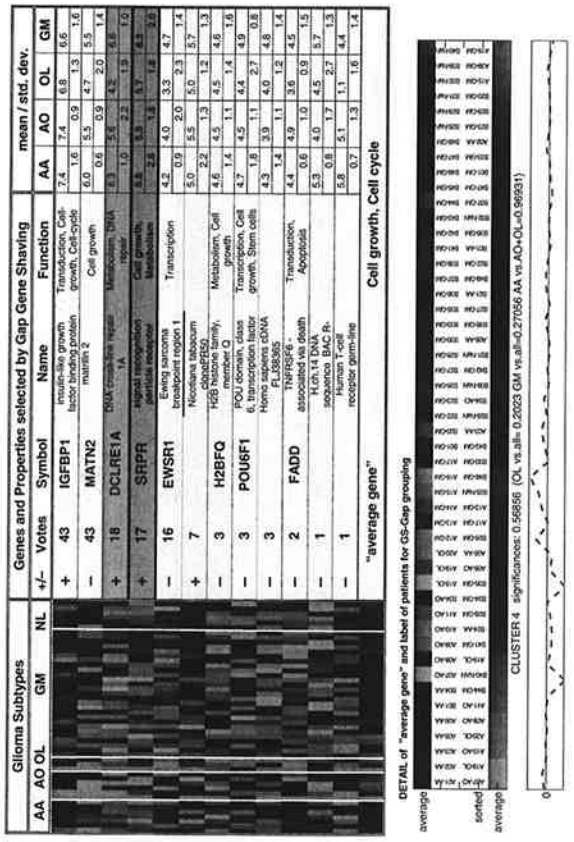


*Figure 7.3.* (see page 107)

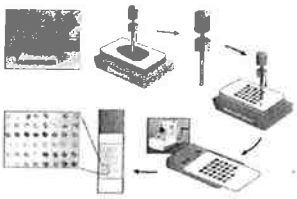

*Figure 18.1.* (see page 382)

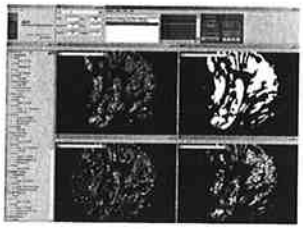

*Figure 18.3.* (see page 393)

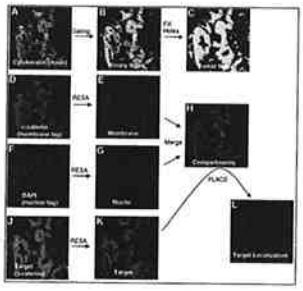

*Figure 18.4.* (see page 396)



To use all data $(y_j, s_j)$ points without choosing, it is necessary to carry out some very robust linear regression since there is superposition with $(y_j, s_j)$ pairs for which several rightful *neighbors* exists. Blades *et al.* (2004a) claim that *robust regression* performs better than regular *linear regression* or *resistant regression*. Figure 11.1 shows an illustration of their estimation method for one of the libraries utilized in their original work, the normal pancreatic HX library (GEO accession: GSM721). Using robust linear regression, they found an error-rate of $\varepsilon = 0.09$, with [0.07; 0.11] as 95% confidence interval, for base substitution error. The line corresponding to 20% error-rate is shown for illustration (see also web-Fig. 2).

In the following, we will discuss the counting estimation.

One can suspect that some of the rare tags occurrences are not real, being result from experimental errors. An example could be an abundant tag that suffered a change by sequencing error in one of the 10-base tag, yielding a non-existing unique tag or inflating the counts of other tags. Colinge and Feger (2001) proposed an estimation method that accounts for sequencing errors. They approximate the sequencing error effect to the expectations of first-order errors and build a system of linear equations:

$$\begin{cases} y_1 = p(1|1)x_1 + \cdots + p(1|t)x_t \\ \qquad\qquad \vdots \\ y_t = p(t|1)x_1 + \cdots + p(t|t)x_t \end{cases} \Rightarrow \mathbf{x} = \begin{bmatrix} p(1|1) & \cdots & p(1|t) \\ \vdots & \ddots & \vdots \\ p(t|1) & \cdots & p(t|t) \end{bmatrix}^{-1} \mathbf{y} \tag{11.3}$$

where $t$ is the number of observed unique tags, $p(j|k)$ is the probability that the observed $k$-th tag is counted as the observed $j$-th tag due to one-step sequencing error, $\mathbf{x}$ are the true unknown counts and $\mathbf{y}$ are the observed

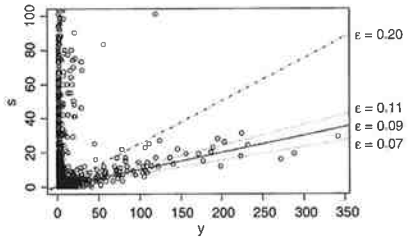


*Figure 11.1.* The substitution error-rate estimation. Normal pancreatic library HX (GEO accession: GSM721). Adapted from (Blades et al., 2004a).

counts. By first-order we mean that only the main effects are accounted for in the approximation, i.e., $p(j|k) = \mathbf{P}(j \leftarrow k) \leq \mathbf{P}(j \leftarrow \cdots \leftarrow k)$, where arrows denote the error process of insertion, deletion or substitution. In spite of the validity of the fundamental propriety $\Sigma x_\cdot = m = \Sigma y_\cdot$, their approximation forces the impossible continuity of counts, i.e., $y_j \in \{0, 1, \ldots, m\}$ but their solutions $x_j \in \Re$. Akmaev and Wang (2004) criticize this fact warning for lack of interpretability of true counts estimates, especially when yielding negative counts. A possible and conceptually correct formulation to approach error-free counting estimation should search solutions in constrained space $\Theta = \{(x_1, \ldots, x_t) : x_j \in \mathbf{Z}_+,\ \Sigma x_\cdot = m\}$. Such a kind of problem is called by computer scientists as an *Integer Programming* problem, and dealing with it as if it is a continuous *Linear Programming*, or even a simple *Linear Algebra* problem, is an inappropriate approach. Akmaev and Wang (2004) offered an alternative to the Colinge and Feger (2001) method to correct for sequencing errors. They use a multi-step heuristic approach very linked to SAGE process mechanics that preserves the data's discrete nature and uses information from chromatograms and *phred* scores (Ewing and Green, 1998). The statistical analysis is just one step of their Bioinformatics algorithm, available trough SAGEScreen software. Another recent alternative is the method developed by Beißbarth *et al* (2004) that is based on an EM-algorithm and on a rigorous and complete statistical modeling of sequencing errors, taking advantage of *phred* scores. It is important to notice that corrections of potential errors at counting level or *denoising* techniques (Blades *et al.*, 2004b) are promising approaches, but not yet widespread standard procedures in SAGE analysis.

In the following, we will discuss abundance estimation.

At first sight, the problem of estimating the abundance $\pi \in [0; 1]$ of a tag could be regarded as an easy and uninteresting problem: $p = x/m$ and that's all, where $x$ is the (pre-processed or not) number of counts for a given tag, $m$ is the total amount of sequenced tags and $p$ is the estimate for $\pi$. However, other elaborate options exist. In fact, $p$ is the *Maximum Likelihood* (ML) estimator of a Bernoulli Process. The Bernoulli or Poisson modeling, and not the Hypergeometric modeling for example, are widely used mathematical frameworks for gene expression counting data because "sampling from an infinite population approximation" is adequate (of course, $m$ is much smaller than the total amount of mRNA molecules in the harvested cells).

In the Bayesian framework, all parameters are unknown quantities and previous knowledge about it is quantified by means of *a priori* pdfs. If we believe in the Bernoulli Process description of SAGE, then, by Bayesian analysis:



$$\begin{cases} h(\pi) \propto (1-\pi)^{\beta-1}\pi^{\alpha-1} \\ L(x|\pi) \propto (1-\pi)^{m-x}\pi^{x} \end{cases} \Rightarrow h(\pi|x) = \frac{(1-\pi)^{m-x+\beta-1}\pi^{x+\alpha-1}}{B(x+\alpha, m-x+\beta)} \tag{11.4}$$

where $\alpha > 0$ and $\beta > 0$ are parameters that define the Beta *a priori* pdf, $x$ is the number of counts for a given tag, $m$ is the total of sequenced tags, $\pi$ is the tag abundance, $L(\cdot)$ is the likelihood function and $B(\cdot)$ is the beta special function. Beta is a standard *priori* choice in Bayesian analysis and it is very flexible to accommodate various kinds of prior knowledge. Note that Eq. 11.4 is equivalent to the result: if $\pi \sim \text{Beta}(\alpha, \beta)$ and $x|\pi,m \sim \text{Binomial}(\pi, m)$ then $\pi|x,m \sim \text{Beta}(\alpha + x, \beta + m - x)$, a basic result that will be used several times in this chapter. The *Posteriori Mode*, i.e., the value of parameter that lead *a posteriori* pdf to its maximum, is $p = (x + \alpha - 1)/(m + \alpha + \beta - 2)$. Therefore, the only way to get the same simple estimate obtained from the ML approach, is by using the uniform *a priori*, i.e., $\alpha = \beta = 1$. Other informative *priori* choices may have influence in the abundance estimate and could be derived from the transcript level distributions.

Morris *et al.* (2003) raised a series of criticism against the use of simple ML estimator $p = x/m$ in SAGE analysis. In spite of the very common use of Binomial to model the outcome of a given tag, they have reminded that SAGE is an *incomplete multinomial* sampling:

$$L(\mathbf{x}|\boldsymbol{\pi}, m) = m! \prod_{j=1}^{t} \frac{\pi_j^{x_j}}{x_j!} \tag{11.5}$$

where $\pi_j$ is the abundance and $x_j$ are the counts of the $j$-th tag, $\boldsymbol{\pi} \in \{(\pi_1, \ldots, \pi_t) : \pi_\cdot > 0,\ \Sigma\pi_\cdot = 1\}$, $t$ is the number of unique tags and $m$ is the total amount of sequenced tags. By "incomplete" we mean that $t$ is unknown, but to go further with this modeling, one must assume that it is known in advance. Note that $\pi_j = 0$ is not allowed, but for an existent and non-observed tag, the ML estimator is $p = 0$. Since we (assume to) know that there must be $t$ transcripted tags, $j$-th tag's $x_j = 0$ means that $\pi_j < 1/m$. These tags are called *underrepresented* and, since $\Sigma\pi_\cdot = 1$, the others are *overrepresented*. They suggest that this is not a minor effect based on the skewness of gene expression distributions and propose a "*Robin Hood*" non-linear shrinkage estimator for abundances. It is important to note that Stern *et al.* (2003) warned about usual SAGE studies' inability to estimate the number of unique transcripts $t$, arguing that $m$ should be larger than is nowadays available in SAGE studies. This could be a major drawback for the use of Morris *et al.* (2003) estimator in general cases.

## 3.2 Estimation by Interval (Error-Bars)

The second and complementary approach to obtain quantitative insights about unknown parameters is by means of *Estimation by Interval*, or using informally terms, to define error-bars.

The SAGE process is analogous to the well-known "balls and urns" statistical problem. In practice, this means that a lot of theoretical framework is already available and proposed statistical models have solid underlying physical basis. For microarray data, for example, this is not true since competitive hybridization is a physical phenomenon much more harder to model, requiring assumption-prone analysis from statisticians. Given the simple "counting" nature of SAGE data, it is easy to report tag abundance with some error-bar.

Using basic Bayesian statistics, as in Section 3.1, and choosing a non-informative uniform *a priori* we saw (Eq. 11.4) that tag abundance is $\pi|x,m \sim \text{Beta}(1+x, 1+m-x)$, where $x$ are the counts and $m$ the total size of the library. Once a credibility level $\alpha$ is defined, it is only necessary to integrate around the *posteriori*'s peak until this probability is reached (Fig. 11.2).

A tag with $x = 16$ counts in $m = 80{,}000$ has an abundance of $2.0 \cdot 10^{-4}$ and its 68% and 95% credibility intervals are $[1.5 \cdot 10^{-4}; 2.5 \cdot 10^{-4}]$ and $[1.2 \cdot 10^{-4}; 3.1 \cdot 10^{-4}]$, respectively. A tag with $x = 32$ and $m = 160{,}000$ has also an abundance of $2.0 \cdot 10^{-4}$, however, its 68% and 95% credibility intervals are more precise, respectively $[1.7 \cdot 10^{-4}; 2.4 \cdot 10^{-4}]$ and $[1.4 \cdot 10^{-4}; 2.7 \cdot 10^{-4}]$, as one intuitively expects.

Sometimes, people prefer to model such kinds of rare counting data, like SAGE, using Poisson random variables. In this case, the parameter in focus is $\lambda$, the number of counts per $m$ (note that $\lambda = m\pi$). Audic

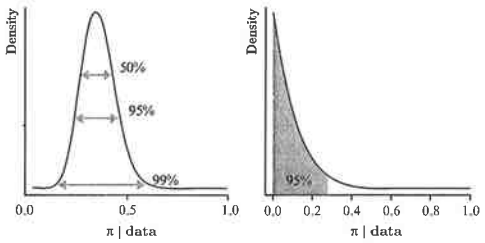


*Figure 11.2.* The error-bar construction. Left: three examples of credibility intervals. Right: posteriori peak can coincide with interval boundaries.



and Claverie (1997) remind the Frequentist alternative for the problem, the Ricker's formula, to find an $\alpha$% confidence interval $[\lambda_1; \lambda_2]$ for $\lambda$. Using this different approach with, again, $x = 16$ and $m = 80{,}000$ we get $\lambda = 16.0$ with [12.0; 21.1] and [9.1; 26.0] for 68% and 95% confidence intervals, respectively, or using abundance as result $\lambda/m = 2.0 \cdot 10^{-4}$ with $[1.5 \cdot 10^{-4}; 2.6 \cdot 10^{-4}]$ and $[1.1 \cdot 10^{-4}; 3.2 \cdot 10^{-4}]$. Note that in this example the results were very similar to Bayesian analysis but this is not a general fact. There are several examples in statistical literature in which Bayesian and Frequentist analysis disagree.

In Gene Expression analysis, hybridization-based techniques such as cDNA microarray or traditional northern blot give, by construction, relative results, e.g., expression ratios. Given two gene abundances obtained by SAGE, it is easy to transform them into expression ratios but the opposite is impossible. Until now, the only solution that we know for the *Estimation by Interval* of expression ratio in the SAGE community is a Bayesian solution (Vêncio *et al.* 2003). As viewed before, $\pi|x,m$ follows a Beta pdf, but for two classes A and B the pdf of $R = (\pi_A|x_A,m_A)/(\pi_B|x_B,m_B)$ could be hard to obtain analytically. Vêncio *et al.* (2003) sampled pseudo-random numbers from Beta pdfs that describe each class and estimated the expression ratio $R$ pdf calculating the quotient for every pair-wise simulated observations. In fact, the estimated distribution is the pdf of a re-parametrization of expression ratios $Q = 1/(1+R) \in [0;1]$ because it is better suited for non-parametric *Kernel Density Estimators* since $R \in [0; \infty]$. Once obtained the $\alpha$% credibility interval, it is easy to go back to $R$ space solving $Q$ for $R$. For a given tag, the final result could be presented as $R = 5.8$ with the possible scenarios for 95% credibility intervals: [5.5;6.0], [1.2;15.3], [2.2; $\infty$] or [0.3;25.5], instead of simple 5.8-fold change. The first scenario is the ideal situation with an intuitively small interval indicating a relatively precise result for differential expression ratio; the second scenario suggests that, although differentially expressed, the quantitative aspect of the ratio is poor since ratio possibilities are widespread in a wide interval; the third shows that the only safe conclusion, with this level of credibility, is that the ratio is greater than 2.2-fold; and the last scenario indicates that the ratio is very wide, crossing the non-differential expression ratio $R = 1$ barrier, and should not be seriously considered in spite of 5.8-fold indication of an apparent significant change.

## 4. Differential Expression Detection

Although SAGE is innovative from a biological viewpoint, it is, from statistical viewpoint, a very old and known problem: to draw balls from urns. We will discuss the statistical viewpoint of the comparison between

libraries and its possibilities, but one should always have in mind that we are working with difficult biological data. This means that sometimes we do not have enough samples to be compared, because the disease is very rare or because obtaining the samples is a hard technical issue, for example. For this cases, several single-library pair-wise comparison methods are available. Sometimes, in the Statistics viewpoint, this comparison sounds inappropriate. As SAGE, "Digital Northern" or MPSS are not easy and cheap techniques, we would like to stress that in this field it is very important to release data into public databases to ameliorate this problem. Only recently SAGE community learned how to account for the within-class, or between-library, variability in SAGE analysis (Baggerly *et al.*, 2003; Vêncio *et al.*, 2004). Until that, differential expression detection methods merged all count observations from libraries that compose a class in "pseudo-libraries", in order to use previously available pair-wise comparison techniques.

## 4.1 Single-Library or "Pseudo-library"

There are several methods for dealing with single library in each class, i.e., methods that consider only variability due to sampling error. Even when biological replicates are available, it is very common in SAGE studies to construct a "pseudo-library" aggregating the counts of biological replicates and use these single-library methods. There are 3 clear distinct groups of methods: simulation based, Frequentist and Bayesian.

The simulation based method is well-known because it was the chosen method in the original *Science* paper describing differential expression analysis with SAGE (Zhang and Zhou *et al.*, 1997). Some details of their algorithm are just available inside simulation software's source code (Prof. Kinzler, Johns Hopkins University School of Medicine, personal communication). To find differentially expressed tags between two libraries A and B, simulated data sets were generated using a Monte Carlo technique, or in other words, by creating each $k$ simulated data set distributing the $(x_{jA} + x_{jB})$ counts of the $j$-th tag according to a rule, repeating for all $t$ unique tags:

$$\begin{cases} x_{jA}^{(k)} = \sum_{z=1}^{x_{jA}+x_{jB}} \mathbf{1}_{\{u_z \le m_A(m_A+m_B)^{-1}\}} \\ x_{jB}^{(k)} = x_{jA} + x_{jB} - x_{jA}^{(k)} \end{cases} \qquad u \sim U_{[0;1]} \tag{11.6}$$

where $u_z$ are uniform pdf realizations, $\mathbf{1}_{\{\cdot\}}$ is the indicator function and $m$ are the total sequenced tags of each library. To quantify the evidence of



the tag, they defined a measure called $P$-chance $K$:

$$K_j = \frac{\min\left(100, \sum_{k=1}^{100000} \mathbf{1}_{\left\{|x_{jA}^{(k)} - x_{jB}^{(k)}| \ge |x_{jA} - x_{jB}|\right\}} \mathbf{1}_{\left\{(x_{jA}^{(k)} - x_{jB}^{(k)})\cdot(x_{jA} - x_{jB}) \ge 0\right\}}\right)}{\min(tot,\ 100000)} \tag{11.7}$$

where *tot* is:

$$\sum_{k=1}^{tot} \mathbf{1}_{\left\{|x_{jA}^{(k)} - x_{jB}^{(k)}| \ge |x_{jA} - x_{jB}|\right\}} \mathbf{1}_{\left\{(x_{jA}^{(k)} - x_{jB}^{(k)})\cdot(x_{jA} - x_{jB}) \ge 0\right\}} = 100$$

This is only a formal way to say that they simulate samples until 100 occurrences of the event "difference in simulated data set is equal/greater than actual observed difference", or 100,000 runs limit is reached, counting the occurrences until this limit. The selection of a significant $K$ is done by comparing it with those obtained from an artificial data set that represent the null hypothesis of no differential behavior between libraries:

$$x_{jAN} = x_{jBN} = \frac{x_{jA} + x_{jB}}{2} \tag{11.8}$$

Repeating the same procedure of Eq. 11.6 and Eq. 11.7 for this "null data set" one can choose a suitable $K$ for his observed experimental differences. SAGE300 and SAGE2002 softwares implement this method using 40 "null data sets" to rank $K$.

The Frequentist approach is based on the proposition of a function of the data and the discovery of the function's pdf when there should not exist differences between classes, i.e., the so-called null pdf. From this Classical framework come the meaning of $p$-value, power of the test, size of the test, etc. It is always based on fact that there exists some (assumed) distribution from which the data was generated and one tries to estimate mistaken conclusions (false positive and false negative) facing this underlying pdf. There is considerable Statistics literature comparing methods for proportion testing. Comparisons in SAGE context are also available elsewhere (Man *et al.*, 2000; Romualdi *et al.*, 2001; Ruijter *et al.*, 2002). Some studies show small advantages of one method in relation to others, but Ruijter *et al.* (2002) remind that the differences are technical and negligible face the drastic approximation of dealing with no biological replicates or, as they called, "one measurement" framework. We discussed replication based methods in Section 4.2. The main Frequentist methods are the well-known Classical Fisher's exact test for contingency tables and Z or $\chi^2$ based methods that use asymptotic results to get $p$-values. For large $m$ values, the combinatorial computation needed in Fisher's test becomes

hard, but the standard approximations become very accurate. The approximate test mostly known in the SAGE community was suggested by Kal *et al.* (1999). For the significance level $\alpha$, we say that there is differential expression if:

$$\left| \frac{\frac{\sum_A x_i}{\sum_A m_i} - \frac{\sum_B x_i}{\sum_B m_i}}{\sqrt{\left(\frac{\sum_A x_i + \sum_B x_i}{\sum_A m_i + \sum_B m_i}\right)\left(1 - \frac{\sum_A x_i + \sum_B x_i}{\sum_A m_i + \sum_B m_i}\right)\left(\frac{1}{\sum_A m_i} + \frac{1}{\sum_B m_i}\right)}} \right| > Z_{\frac{\alpha}{2}} \tag{11.9}$$

where $Z_{\alpha/2}$ is the standard normal $\alpha/2$ quantile, $x_i$ is the number of counts of a particular tag in the $i$-th library and $m_i$ is the sequenced total of the $i$-th library.

The Bayesian approaches follow the Bayesian Statistics framework, using Bayes' rule to go from previous information, the so-called *a priori* pdf, to the information updated by the observations, the so-called *a posteriori* pdf. They work at the parameter space instead of sample space and do not admit the existence of "data that could be observed but was not". Therefore, the Bayesian $p$-value has not the same interpretation of the Frequentist one. It is easy to rewrite the Eq. 11.4 to accommodate several separate replicates from a Bernoulli Process, only generalizing the Likelihood function:

$$\Lambda(\mathbf{X}|\pi) = \prod_{i=1}^{n} L(x_i|\pi) \propto (1-\pi)^{\sum_{i=1}^{n} m_i - \sum_{i=1}^{n} x_i} \pi^{\sum_{i=1}^{n} x_i} \tag{11.10}$$

This means that $\pi$ |data $\sim$ Beta($\alpha+\Sigma x.$, $\beta+\Sigma m.-\Sigma x.$) for A or B classes. There are several ways to rank the "equality of abundances" hypothesis, ranging from simple Bayes Error Rate (Duda *et al.*, 2000) to the well-known Jeffreys' test for precise hypothesis (Jeffreys, 1961) or the genuine Bayesian test for precise hypothesis presented in Madruga *et al.* (2003).

Analysis of equality of abundances $\pi_A = \pi_B$ is not the only paradigm that could be used. It is also possible to carry out significance ranking using absolute counts $\Sigma x.$ or using expression ratio fold-changes $\pi_A/\pi_B$. The most known Bayesian methods using these alternative paradigms are Audic and Claverie's (1997) method and the method implemented by SAGEmap and SAGE Genie (Lal and Lash *et al.*, 1999).



Supposing that the sampling process is well approximated by a Poisson distribution, Audic and Claverie (1997) write the probability of observing the data from one class, given the data observed in the other class. They say that there is differential expression, with some pre-defined probability $\alpha$, if:

$$\sum_{k=\Sigma_B x_i}^{\infty} \left[ \frac{(k+\sum_A x_i)!}{(\sum_A x_i)!k!} \left(\frac{\sum_B m_i}{\sum_A m_i}\right)^k \left(1 + \frac{\sum_B m_i}{\sum_A m_i}\right)^{-k-\Sigma_A x_i - 1} \right] < \frac{\alpha}{2}. \tag{11.11}$$

where again $m_i$'s and $x_i$'s are summed over each class A and B to create the pool and $B(\cdot)$ is the beta special function. Another well-know Bayesian method is the method adapted by Lal and Lash *et al.* (1999) from Chen *et al.* (1998) to accommodate classes with different total counts. This method is implemented in the important SAGE public database tools: the National Center for Biotechnology Information's SAGEmap (http://www.ncbi.nlm.nih.gov/SAGE), and the National Cancer Institute's Cancer Genome Anatomy Project - SAGE Genie (http://cgap.nci.nih.gov/SAGE). They determine the *posterior probability* of fold-changes in expression ratio $R$ greater than an arbitrary value $R^*$. The *a posteriori* pdf used is:

$$h(q|\mathbf{X}, \mathbf{M}, c) \propto q^{c+\sum_A x_i} (1-q)^{c+\sum_B x_i} \left(1 + \frac{\sum_A m_i}{\sum_B m_i} q - q\right)^{-\left(\sum_A x_i + \sum_B x_i\right)} \tag{11.12}$$

where $q$ is a convenient reparameterization of expression ratio $q = R/(R+1)$ and $c$ is a constant that came from an *a priori* pdf being modeled by previous knowledge of researchers. In SAGEmap and SAGE Genie tools, for example, it is assumed $c = 3$ (Lal and Lash *et al.*, 1999). The Eq. 11.12 holds for fold-change $R$ of class A relative to class B and for estimated abundance in A class greater than in class B, $p_A \geq p_B$. If the contrary occurs, just permute the classes labels in Eq. 11.12, for simplicity. It is important to note that a crucial difference/difficulty arise because the differential expression conclusion depends on a pre-defined fold-change. The test answers the question about this fold-change and it is the user's responsibility to define what is a fold-change that means differential expression. Using SAGE Genie's method with $\Sigma_A m = 80{,}000$, $\Sigma_A x = 60$, $\Sigma_B m = 50{,}000$, $\Sigma_B x = 10$, $p_A/p_B = 3.75$ fold change and $R^* = 2$, for example, we obtain by integration of Eq. 11.12: $\mathbf{P}(R \geq 2) = \mathbf{P}(0.666 \leq q \leq 1) = 0.93$. For another cutoff example $R^* = 4$ we obtain: $\mathbf{P}(R \geq 4) = \mathbf{P}(0.8 \leq$

$q \leq 1) = 0.19$. Therefore, there is little chance that the true fold-change is greater than 4-fold but considerable chance that it is greater than 2-fold (see web-Fig. 3). However, if one defines that differential expression occurs simply if an abundance is greater than the other, then $\mathbf{P}(R > 1) = \mathbf{P}(0.5 < q \leq 1) = 0.999$. This highlights the user's responsibility in defining a suitable $R$.

## 4.2 Replicated Libraries in One Class

Suppose that one is sampling colored balls from several urns. Also, before choosing an urn, one choose at random which one will be sampled. To make general conclusions about blue balls, one must weight the sampling with the probability of choosing a particular urn, especially if it is known that each urn could have different abundance of blue balls. A very similar situation occurs when dealing with biological replicates in SAGE analysis. Like the different urns in the above illustration, is intuitive to accept that different biological replicates have distinct abundances for a given gene.

For a given tag, the counting process of an $i$-th library is commonly modeled as a Bernoulli Process with a fixed unknown abundance $\pi_i \in [0; 1]$. The pdf of this abundance among all $n$ libraries is unknown and $\pi_i$ is the $i$-th realization of $\pi$. For a fixed tag, the likelihood of a particular count $x_i$ in total of $m_i$ sequenced tags is often modeled by the Binomial, weighted by the possible $\pi$ outcome:

$$L(x_i|m_i, \boldsymbol{\theta}) = \int_0^1 f(\pi|\boldsymbol{\theta}) \binom{m_i}{x_i} (1-\pi)^{m_i - x_i} \pi^{x_i} \, d\pi \qquad (11.13)$$

The function $f(\cdot)$ is the unknown pdf of tag abundance $\pi$, and is all we want/need to know. This function is parameterized by vector $\boldsymbol{\theta}$. This is a *mixture model*, with the Binomial being the *mixing distribution*, but others, such as Poisson, could be used (Bueno *et al.*, 2002). Since we do not know in advance the stopping rule, the pdf could not be a Binomial but differ only by a multiplicative constant. To reach the common *Binomial model*, used by almost all SAGE methods, it is enough to assume that $f(\cdot)$ is a degenerate pdf over some scalar value $\theta$, i.e., a Dirac's Delta function constrained to [0;1]. With this assumption, we tacitly ignore the possible variability between libraries, due to any reason other than sampling, since only $\pi = \theta$ has positive density. Using a much more realistic approximation, one could assume that the tag abundance in different libraries $f(\cdot)$ is described by a Beta random variable, with non-zero variance. This leads to the well-known *Beta-Binomial model*,



and appears in the SAGE differential expression context introduced by Baggerly *et al.* (2003), derived as a hierarchical model in a Frequentist framework instead of a particular case of a mixture model. For a fixed tag, given the vector of counts $\mathbf{X} = (x_1, \ldots, x_n)$ and the vector of sequenced totals $\mathbf{M} = (m_1, \ldots, m_n)$ in all $n$ libraries of the same class, it is necessary to fit the Beta-Binomial model parameters and to test for differential expression.

In the Frequentist framework first proposed by Baggerly *et al.* (2003), $p_i = x_i/m_i$ is used as an estimator of $\pi_i$ and a linear combination of these abundances is proposed as the correct way to combine results from different libraries:

$$p = \sum_{i=1}^{n} w_i p_i, \quad V_u = \frac{\sum_{i=1}^{n} w_i^2 p_i^2 - p^2 \sum_{i=1}^{n} w_i^2}{1 - \sum_{i=1}^{n} w_i^2}, \quad w_i \propto \frac{m_i(\alpha+\beta)}{\alpha+\beta+m_i} \qquad (11.14)$$

where $w_i$ are the weights that yield an unbiased minimum variance estimator $V_u$ for weighted proportion's variance and $\boldsymbol{\theta} = (\alpha, \beta)$ are the Beta pdf parameters. However, this unbiased variance could be unrealistically small when it becomes smaller than the sampling variability. We know that the variance of this model cannot be smaller than the variance eventually obtained if we do not consider within-class variability. Therefore, they propose the final *ad hoc* estimator:

$$V = \max\left[V_u; V_{pseudo\text{-}lib}\right] \qquad (11.15)$$

where:

$$V_{pseudo\text{-}lib} = \frac{1}{\sum_{i=1}^{n} m_i} \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i} \left(1 - \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} m_i}\right)$$

The max(·) function assure that $V$ is not unrealistic small when $V_u$ is unrealistic small. $V_{pseudo\text{-}lib}$ is exactly the natural estimator for variance if one considers $f(\cdot)$ as a Dirac's Delta instead of a Beta pdf as the underlying model. Note that the $\Sigma x./\Sigma m.$ term is equivalent to the abundance if one merges all libraries into a "pseudo-library". To *fit* all these parameters, they used the computationally practical *Method of Moments*. Once $p_A$, $p_B$, $V_A$ and $V_B$ are found for classes A and B, it is necessary to test if the proportions are significantly different. Evoking asymptotic results they propose

the use of a $t_w$ statistics as following a Student's $t_{df}$ pdf:

$$t_w = \frac{p_A - p_B}{\sqrt{V_A + V_B}}, \quad df = \frac{(V_A + V_B)^2}{\frac{V_A^2}{\sum_A m_i - 1} + \frac{V_B^2}{\sum_B m_i - 1}} \tag{11.16}$$

A different approach that accounts for within-class variability, uses the Bayesian Statistics framework and does not rely on asymptotic results was recently presented by our group (Vêncio *et al.*, 2004). Considering our likelihood as obtained by the Beta-Binomial model, it is easy to write the *a posteriori* pdf:

$$g(\theta_1, \theta_2 | \boldsymbol{X}, \boldsymbol{M}) \propto \mathbf{1}_{[\theta_2 \geq \sigma]} \cdot \prod_{i=1}^{n} \frac{B(\alpha_\theta + x_i, \beta_\theta + m_i - x_i)}{B(\alpha_\theta, \beta_\theta)} \tag{11.17}$$

where the indicator function is our *a priori* pdf. Note that we made a re-parameterization since now $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is the mean and standard deviation (stdv) of the Beta pdfs that describe each class. The new parameter space $\Theta = \{(\theta_1, \theta_2)\colon 0 \leq \theta_1 \leq 1, 0 \leq \theta_2^2 < \theta_1(1 - \theta_1) \leq 1/4\}$, is more intuitive than the common $(\alpha, \beta)$ one and is bound (much more amenable to numerical computations). We use the sub-index $\boldsymbol{\theta}$ in $\alpha$ and $\beta$ to remind that they are functions of a new parameterization, easily obtained from Beta mean and stdv expressions. The *a priori* is an uniform pdf over $\Theta$, but constrained to variances greater than the variance obtained by "pseudo-library" construction. The variance "working floor" $\sigma^2$ come from the Beta pdf obtained using the Eq. 11.10 generalization into Eq. 11.4. We find the two Beta pdfs that describe each class taking the *Posteriori Mode*, i.e., $(\theta_1, \theta_2) \in \Theta$ that lead Eq. 11.17 to its maximum (see web-Fig. 4). Finally, to test if a tag is differentially expressed between the two classes, we propose an evidence measure other than the $p$-value. We use the intuitive Bayes Error Rate (Duda *et al.*, 2000):

$$E = \int_0^1 \min(f(\pi|\hat{\theta}^A), f(\pi|\hat{\theta}^B))\, d\pi \tag{11.18}$$

where $\hat{\theta} = \arg_\theta(\max_\Theta(posteriori))$. Small Bayes Error Rate $E$ values indicate that the whole Beta pdfs are "far apart", thus with high evidence of differential expression (web-Fig. 5). We rank tags by $E$ evidence and let biologists say what they intuitively think that is an unacceptable level of superposition (classification error) between the two classes. An indispensable tool for checking intuitive consistency of the results obtained with any method is the graphic representation of all individual observations, like in Fig. 11.3 of Section 5. Using this simple tool one can easily note the inconsistency of "pseudo-libraries" methods in several cases.



### 4.3 Multiple Libraries Outlier Finding

There is a third type of comparison analysis using counting data that is to find outlier libraries in a multiple libraries context. The methods reviewed in previous section deal with pair-wise comparison of classes, having one or more libraries, accounting or not for biological replication. On the other hand, multiple libraries comparison is not pair-wise but rather search for tags with a "non-usual" behavior in a set of libraries. The concept of class is non longer used, or in other words, all libraries are regarded as an unique class.

Probably, the most known method for outlier detection was the method presented by Stekel *et al.* (2000). It came to improve the previously available method introduced by Greller and Tobin (1999) which only detects an outlier very different than others in a set of libraries. Stekel *et al.* (2000) proposed a flexible solution that tries to detect if a transcript has the same abundance across several libraries simultaneously. For example, the input could be several tissues and our aim could be to detect tissue-specific transcripts. They skip from $p$-value pitfalls and simply rank their proposed $R$ statistics. For a given tag and fixed some cutoff $R^*$, we say that there is differential expression in at least one library if:

$$\sum_{i=1}^{n} \left[ x_i \ln \left( \frac{x_i}{m_i} \frac{\sum_{j=1}^{n} m_j}{\sum_{j=1}^{n} x_j} \right) \right] = R > R^* \tag{11.19}$$

where $x_i$ are the counts for some tag in $i$-th library and $m_i$ is the total of sequenced tags in $i$-th library. To help the user to define a $R^*$ cutoff, they propose an alternative evidence measure called *believability*. This measure is obtained by an usual randomization strategy or from asymptotic appeal since $2R \rightarrow \chi^2_{n-1}$.

## 5. Illustration of Methods Application

In order to gain intuition about the methods presented in this chapter, we applied some of them to a relevant publicly available data set. Our analysis, along with R language (Ihaka and Gentleman, 1996) scripts, are available in detail at the chapter's supplemental web-site: http://www.vision.ime.usp.br/~rvencio/CSAG/. Our aim is to search for genes differentially expressed between grade II and grade III astrocytoma from bulk material collected from different patients. The data is available at SAGE Genie web-site at (http://cgap.nci.nih.gov/SAGE). Table 11.1 shows the libraries used in this illustration. Here we do not want to focus on biology of the analysis but rather in the fundamental difference of methods

*Table 11.1.* Brain tumor library from SAGE genie.

| # | Library Name – Class A | Total Tags |
|---|---|---|
| 1 | SAGE_Brain_astrocytoma_grade_III_B_H1020 | 51573 |
| 2 | SAGE_Brain_astrocytoma_grade_III_B_H970 | 106982 |
| 3 | SAGE_Brain_astrocytoma_grade_III_B_R140 | 118733 |
| 4 | SAGE_Brain_astrocytoma_grade_III_B_R927 | 107344 |
| | grade III merged "pseudo-library" | 384632 |
| # | Library Name - Class B | Total Tags |
| 5 | SAGE_Brain_astrocytoma_grade_II_B_H563 | 88568 |
| 6 | SAGE_Brain_astrocytoma_grade_II_B_H359 | 105764 |
| 7 | SAGE_Brain_astrocytoma_grade_II_B_H388 | 106285 |
| 8 | SAGE_Brain_astrocytoma_grade_II_B_H530 | 102439 |
| | grade II merged "pseudo-library" | 403056 |

for differential expression detection. However, we take care to define each class with libraries with the same histopathological grade and only from bulk material, excluding cell lines.

Following the previous sections we applied a pipe-line for statistical analysis. We used some methods of *Estimation* section and concentrated our attention on *Differential Expression Detection* section.

First we tried to estimate the sequencing substitution, insertion and deletion error-rates of our data-set. We applied the Blades *et al.* (2004a) error-rate estimation method described in Section 3.1 but was difficult to carry out the line fitting, similar to those used in Fig. 11.1 line fitting, due to lack of points at higher expression level. On the other hand, the method proposed by Akmaev and Wang (2004) can be applied only to original output data from sequencing machines, the chromatograms, thus it was impossible to use it here since public databases have the raw counting data and not the original chromatograms. Therefore, we moved further without counting corrections.

Second, given the tags' counts, we used the *Posteriori Mode* for abundance estimation with non-informative uniform *a priori* pdf (Eq. 11.04) to match with *Maximum Likelihood* estimator. We want to focus on the differential expression detection issue.

Third, as an initial approach to differential expression detection, we merged all libraries of each class summing their observations and creating the so-called "pseudo-libraries", as usual in SAGE analysis. To perform the Fisher's Exact Test, the $\chi^2$, and the Audic and Claverie (1997) methods discussed in Section 4.2, we used an user-friendly freely available software called IDEG6 (Romualdi *et al.*, 2001). It has an on-line web-version and



also perform the Stekel *et al.* (2000) or Greller and Tobin (1999) methods for multi-library analysis. To perform the Lash and Lal et al. (1999) Bayesian method we used SAGE Genies' on-line tool but we have implemented their method as a R script to allow the use in any data-set.

Finally, to perform the same analysis in a replicated library context, discussed in Section 4.1, we used the only two available solutions: our SAGEbetaBin method (Vêncio *et al.*, 2004) and the first published solution, the Baggerly *et al.* (2003) $t$-test approximation.

As previously and emphatically announced in Section 4 introduction, the results could be very different using the common "pseudo-libraries" methods or these two that account for within-class inherent variability. A tag considered differentially expressed using replicated library methods should always appear as differentially expressed using "pseudo-library" methods. However, the opposite is not always true. An example of such an effect is obtained for the AATAGAAATT tag, corresponding to secreted phosphoprotein 1 (osteopontin, bone sialoprotein I, early T-lymphocyte activation 1) gene. Using any available "pseudo-library" method, one is lead to believe that this tag is differentially expressed with high significance. As calculated by IDEG6 software, all methods give 0.00 (zero!) $p$-values. The SAGE Genie's method gives 0.00 (zero) $p$-value for a difference greater than 2-fold or 0.01 for a difference greater than 4-fold. Also, our error-bar method (Vêncio *et al.*, 2003) shows that 95% credibility interval is [4.3;6.7] for $R = 5.3$-fold change of A relative to B class and does not surpass ratio equal to 1. All of these results indicate a very high level of confidence in the differential expression conclusion. However, if one plots the individual abundance results for each library, it is easy to note that the conclusion of all these methods is suspicious.

On the other hand, the two methods that account for within-class variability do not claim a high significance for this tag, as one intuitively suspects looking at the superposition of crosses (class B) and circles (class A) in Fig. 11.3. The Baggerly *et al.* (2003) $t$-test approximation gives 0.21 for $p$-value and our SAGEbetaBin evidence is 0.48, indicating great superposition between pdfs that describe each class (curves in Fig. 11.3), as discussed in Section 4.1. These results do not support this tag as being differentially expressed in general terms. This is an illustrative example because just one class A library leads traditional analysis to a mistaken conclusion, but there are other more subtle cases (see supplemental website data and web-Fig. 6). In this illustration we consider a tag differentially expressed if it has Bayes Error Rate, arbitrarily defined, $E \leq 0.05$.

Sometimes this finding could bring difficulties in the differential expression validation by other techniques. It is important to note that when we are dealing with high variability between samples, and we are not taking

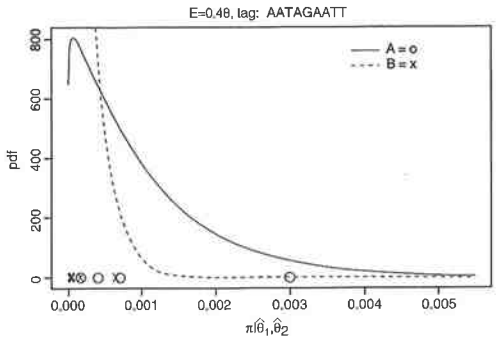


*Figure 11.3.* Example of tag regarded as differentially expressed by "pseudo-library" methods but discarded by replicated-library.

this variability into account to select our differentially expressed candidates, the validation process could become completely arbitrary. It will rely on the samples chosen. The final list of differentially expressed tags between grade II and grade III astrocytomas, their values, and fold change error-bars are available at a supplemental web-site.

## 6. Conclusions

In this chapter we aimed to give a guide to the state-of-art in statistical methods for SAGE analysis. We just scratch some issues for the sake of being focused in differential expression detection problems, but we hope that main ideas could be useful to track the original literature. We saw that estimation of a tag abundance could not be simpler than observed counts divided by sequenced total, but rather can receive sophisticated treatments such as multinomial estimation, correction of potential sequencing errors, *a priori* knowledge incorporation, and so on. Given an (assumed) error-corrected data set, one could search for differentially expressed tags among conditions. Several methods for this were mentioned, but we stress the importance of using biological replication designs to capture general information. Finally, we want to point out that only accumulation of experimental data in public databases, with biological replication, and use of good statistics could improve usefulness of SAGE, MPSS or EST counting data in general terms, helping to elucidate basic/applied gene expression questions.



## Acknowledgments

RZNV received FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) support. We thank Prof. Carlos A. B. Pereira (IME-USP) for teaching us most of what we learned about Statistical Science and Dr. Sandro J. de Souza (Ludwig Institute for Cancer Research) for teaching us most of what we learned about SAGE. We give a special acknowledgement to all colleges that carefully revised our manuscript.

## References

Akmaev, V.R. and Wang, C.J. (2004) Correction of sequence based artifacts in serial analysis of gene expression. *Bioinformatics* **20**, 1254–1263.

Audic S. and Claverie J. (1997) The significance of digital gene expression profiles. *Genome Research* **7**, 986–995.

Baggerly, K.A., Deng, L., Morris, J.S. and Aldaz, C.M. (2003) Differential expression in SAGE: accounting for normal between-library variation. *Bioinformatics* **19**, 1477–1483.

Beißbarth, T., Hyde, L., Smyth, G.K., Job, C., Boon, W., Tan, S., Scott, H.S. and Speed, T.P. (2004) Statistical modeling of sequencing errors in SAGE libraries. *Bioinformatics* **20**, i31-i39.

Blades, N., Velculescu, V.E. and Parmigiani, G. (2004a) Estimation of sequencing error rates in SAGE libraries. *Genome Biology* **in press**.

Blades, N., Jones, J.B., Kern, S.E. and Parmigiani, G. (2004b) Denoising of data from serial analysis of gene expression. *Bioinformatics* **in press**.

Boon, K., Osório, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., Morin, P.J., Buetow, K.H., Strausberg, R.L., Souza, S.J. and Riggins, G.J. (2002) An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 11287–11292.

Brenner, S., Johnson, M., Bridgham, J., *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology* **18**, 630–634.

Bueno, A.M.S., Pereira, C.A.B., Rabello-Gay, M.N. and Stern, J.M. (2002) Environmental genotoxicity evaluation: Bayesian approach for a mixture statistical model. *Stochastic Environmental Research and Risk Assessment* **16**, 267–278.

Chen, H., Centola, M., Altschul, S.F. and Metzger H. (1998) Characterization of gene expression in resting and activated mast cells. *J. Exp. Med* **188**, 1657–1668.

Colinge, J. and Feger, G. (2001) Detecting the impact of sequencing errors on SAGE data. *Bioinformatics* **17**, 840–842.

Duda, R.O., Hart, P.E. and Stork, D.G. (2000) in *Pattern Classification - 2nd Edition*, (Wiley- Interscience Press)

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research* **8**, 186–194.

Greller, L.D. and Tobin, F.L. (1999) Detecting selective expression of genes and proteins. *Genome Research* **9**, 282–296.

Ihaka, R. and Gentleman, R. (1996) R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Jeffreys, H. (1961) in *Theory of Probability*, (Oxford University Press).

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. *Mol. Biol. 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* **21**, 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.

Madruga, M.R., Pereira, C.A.B. and Stern, J.M. (2003) Bayesian evidence test for precise hypotheses. *Journal of Planning and Inference* **117**, 185–198.

Man, M.Z., Wang X. and Wang Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinfomatics* **16**, 953–959.

Margulies, E.H., Kardia, S.L. and Innis, J.W. (2001) Identification and prevention of a GC content bias in SAGE libraries. *Nucleic Acids Res.* **29**, e60.

Morris, J.S., Baggerly, K.A. and Coombes, K.R. (2003) Bayesian shrinkage estimation of the relative abundance of mRNA transcripts using SAGE. *Biometrics* **59**, 476–486.

Romualdi, C., Bortoluzzi, S. and Danieli, G.A. (2001) Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. *Human Molecular Genetics* **10**, 2133–2141.



Ruijter, J.M., Kampen, A.H.C. and Baas F. (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol Genomics* **11**, 37–44.

Schuler, G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75**, 694–698.

Stekel, D.J., Git, Y. and Falciani, F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Research* **10**, 2055–2061.

Stern, M.D., Anisimov, S.V. and Boheler, K.R. (2003) Can transcriptome size be estimated from SAGE catalogs?. *Bioinformatics* **19**, 443–448.

Stollberg, J., Urschitz, J., Urban, Z. and Boyd, C.D. (2000) A Quantitative Evaluation of SAGE. *Genome Research* **10**, 1241–1248.

Vêncio, R.Z.N., Brentani H. and Pereira, C.A.B. (2003) Using credibility intervals instead of hypothesis tests in SAGE analysis. *Bioinformatics* **19**, 2461–2464.

Vêncio, R.Z.N., Brentani, H., Patrão, D.F.C. and Pereira, C.A.B. (2004) Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). *BMC Bioinformatics* **5**, 119.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science* **270**, 484–487.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai M.A., Bassett, D.E., Hieter, P., Vogelstein, B. and Kinzler, K.W. (1997) Characterization of the yeast transcriptome. *Cell* **88**, 243–251.

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.