BMC Bioinformatics

Methodology article

Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE)

Ricardo ZN Vêncio*[1,2], Helena Brentani[3,4], Diogo FC Patrão[3] and Carlos AB Pereira[1,2]

Address: [1]Statistics Department, Instituto de Matemática e Estatística – Universidade de São Paulo, Rua do Matão 1010, 05508-090 São Paulo, BRAZIL, [2]BIOINFO-USP – Núcleo de Pesquisas em Bioinformática da Universidade de São Paulo, Rua do Matão 1010, 05508-090 São Paulo, BRAZIL, [3]Ludwig Institute for Cancer Research – São Paulo Branch, Rua Prof. Antônio Prudente 109, 01519-010 São Paulo, BRAZIL and [4]Hospital do Câncer A.C. Camargo, Rua Prof. Antônio Prudente 109, 01519-010 São Paulo, BRAZIL

Email: Ricardo ZN Vêncio* - rvencio@vision.ime.usp.br; Helena Brentani - helena@lbc.ludwig.org.br; Diogo FC Patrão - djogo@lbc.ludwig.org.br; Carlos AB Pereira - cpereira@ime.usp.br

* Corresponding author

Published: 31 August 2004

BMC Bioinformatics 2004, **5**:119 doi:10.1186/1471-2105-5-119

Received: 14 May 2004
Accepted: 31 August 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/119

Abstract

Background: An important challenge for transcript counting methods such as Serial Analysis of Gene Expression (SAGE), "Digital Northern" or Massively Parallel Signature Sequencing (MPSS), is to carry out statistical analyses that account for the within-class variability, i.e., variability due to the intrinsic biological differences among sampled individuals of the same class, and not only variability due to technical sampling error.

Results: We introduce a Bayesian model that accounts for the within-class variability by means of mixture distribution. We show that the previously available approaches of aggregation in pools ("pseudo-libraries") and the Beta-Binomial model, are particular cases of the mixture model. We illustrate our method with a brain tumor vs. normal comparison using SAGE data from public databases. We show examples of tags regarded as differentially expressed with high significance if the within-class variability is ignored, but clearly not so significant if one accounts for it.

Conclusion: Using available information about biological replicates, one can transform a list of candidate transcripts showing differential expression to a more reliable one. Our method is freely available, under GPL/GNU copyleft, through a user friendly web-based on-line tool or as R language scripts at supplemental web-site.

Background

An important challenge in Serial Analysis of Gene Expression (SAGE) [1] analysis is the decision whether a gene is differentially expressed between two classes, for example tumoral vs. normal classes. In statistical terms, this essential step is to test the null hypothesis H_0: "gene has no differential expression between the two probed classes". A much more usual approach is to assign an index (*P*-value or Bayes factor, for example) that measures the confidence/significance of the hypothesis and let the biologists themselves to establish a cutoff of what they call significant.

This necessity arises because counting sequenced SAGE tags is a process prone to random and systematic errors that affect gene expression abundance estimates.

Systematic errors may come from various sources such as GC content bias [2], sequencing errors [3,4] as well as the possibility of non unique tags. This kind of error can be detected/corrected using some bioinformatics procedures such as quality control of automatic sequencing pipe-line [5], or statistical estimation procedures such as "denoising" [6,7]. Random errors are due to the inherent stochastic characteristic of SAGE data acquisition: sampling from automatic sequencing. Like colored balls in an urn, sampling and counting SAGE tags from a library is commonly modeled by a Bernoulli Process relying on an infinite population sampling approximation.

If an Expressed Sequence Tag (EST) library is non-normalized, its counting data, also known as "Digital-Northern", reflects the abundance of genes. Likewise, the Massively Parallel Signature Sequencing (MPSS) [8] technique counts tags to infer the transcriptome, but using a completely different strategy from traditional DNA sequencing methods, that allows augmented high-throughput capability. Therefore, all the results discussed here are readily applicable to "Digital-Northern" or MPSS context since, from a mathematical viewpoint, all represent the same bioinformatics problem: counting transcripts (as balls in urns).

Nowadays, the variability in SAGE abundance data is modeled only as due to sampling from sequencing, since almost all statistical procedures are performed after aggregation of observations from various libraries of the same class, creating a "pseudo-library". See [9-11] for good reviews on statistical techniques used in SAGE analysis. This extensively used trick tacitly ignores the within-class variability, i.e., the biological variability among individuals within a class (different patients having the same cancer diagnosis, for example), and could lead to overconfident conclusions.

Results

Here we propose a Bayesian model of mixtures to account for within-class variability as a generalization of the Beta-Binomial model [12]. We also show that the usual "pseudo-library" construction is a particular case of our mixture model. Finally, we propose the use of the Bayes Error Rate to intuitively rank the differential expression hypothesis under a Bayesian framework, avoiding several technicalities and difficulties such as: typeI and typeII error analysis, Bonferroni-like multiple testing correction, asymptotic results evocation, imposition of a test statistic and null probability density function (pdf), and so on.

Statistical model

The counting process from automatic sequencing of one single i-th library is often modeled as a Bernoulli Process and a fixed unknown tag abundance π_i is implicitly assumed. The pdf of the random variable of interest, "expression abundance" $\pi \in [0;1]$ among all n libraries is unknown, thus each library could be regarded as being created by a realization of π. These features lead naturally to mixture models [13,14]:

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

where: $f(\cdot)$ is the unknown pdf of the abundance among same-class libraries parameterized by a vector θ, $\mathbf{X} = (x_1,..., x_n)$ is the vector of counts in all n libraries of same class, $\mathbf{M} = (m_1,..., m_n)$ is the vector of library sizes and L is the likelihood of each i-th observation.

The common procedure of merging all observations from libraries of the same class, constructing a "pseudo-library" before statistical inference, is recognized as a particular case of this mixture model: just assume that all libraries have strictly the same abundance, with no biological variability. Mathematically, this is a function with infinite probability density over one single abundance value $\pi = \theta$ and zero over every other $\pi \neq \theta$, or a Dirac's Delta function. Using $f(\cdot)$ as a Dirac's Delta function constrained to [0;1], turn Eq.1 into the familiar and commonly used binomial distribution (see derivation in the Methods section).

We believe that Dirac's Delta is a naive description of real-life SAGE libraries. The Beta distribution is an alternative with non-zero within-class variance to account for intuitively expected biological differences among them. Using $f(\cdot)$ as a Beta in Eq.1, yields the so-called Beta-Binomial model (see derivation in the Methods section).

Given the parameter vector θ that describes the random variable π of some fixed gene G, we must decide if there is a difference between A and B classes (e.g, tumor vs. normal classes). We propose to consider genes as being differentially expressed based on non-superposition of the predictive Beta pdfs of both A and B classes. By "predictive" we mean that we use the *a posteriori* mode in the Beta pdfs. The "non-superposition" intuitive feature is mathematically written as the Bayes Error Rate E [15]:

$$E = \int_0^1 \min\left(f\left(\pi \mid \theta^A\right), f\left(\pi \mid \theta^B\right)\right) d\pi \qquad (2)$$

where $f(\cdot)$ is the Beta pdf and "hat" over the parameters indicate the values that lead to an *a posteriori* pdf maximum. The *a posteriori* distribution is obtained as usual from Bayesian Statistical Theory (*a priori* pdf choice and detailed derivation are in the Methods section).

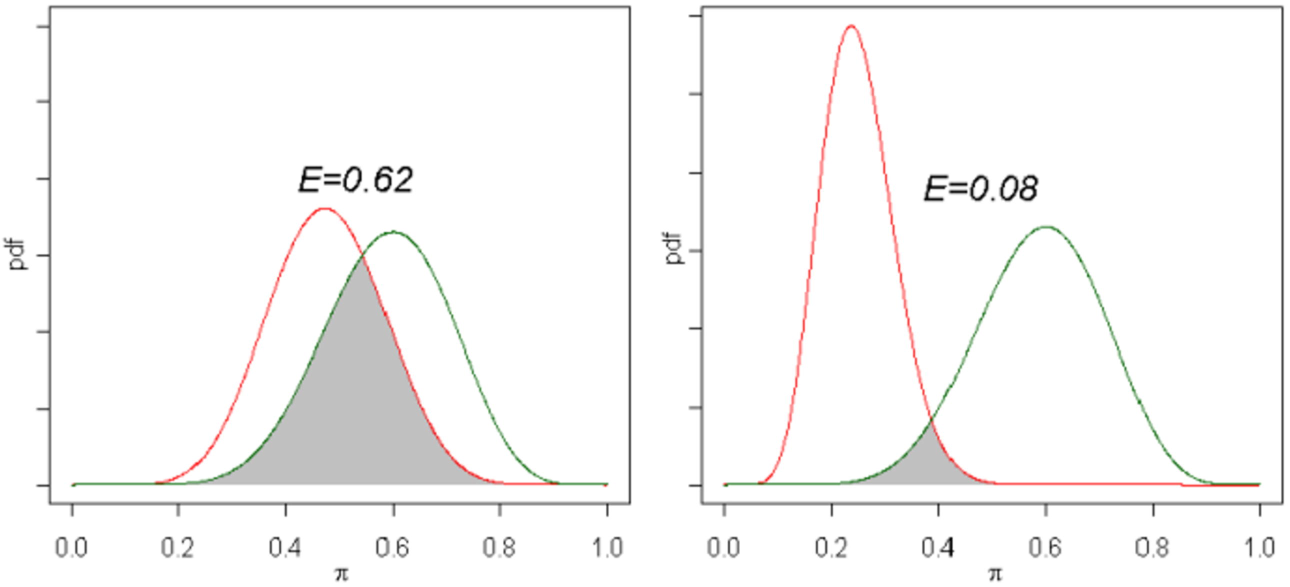


Figure 1
The Bayes Error Rate illustration. This figure shows two illustrations of the proposed use of Bayes Error Rate *E* to define differentially expressing genes based on pdf of expression abundance π. The left example shows an obvious superposition of classes' pdf, thus a gene having this profile does not present evidence of differential expression between classes. The right example shows two pdfs "far apart" and genes with this kind of behaviour should be safely considered differentially expressing between two classes.

Intuitively, if the pdfs are "far apart", the gene probably has reproducible differential expression between classes. In this case, rarely could one misclassify class A as B and vice-versa. Figure 1 gives some insight about this fact. Using our proposed approach, the "far apart" notion means a small Bayes Error Rate *E*. For adepts of the Frequentist Statistics, this evidence measure could resemble a typeI and typeII errors sum, however it is just an illustrative analogy.

As in any significance test method, the experimenter must define what is a high significance *E* value. This cutoff should be guided by external and independent confirmatory assays. To avoid crude decision boundaries, one could rank their significance results but there is no way to avoid some arbitrariness in any kind of statistical test.

In the classical Frequentist Statistics framework, it is common to call a result as significant if it presents a *P*-value ≤ 0.01 in a *t*-like-test, hoping that this could control the error at this level. However, due to technical difficulties such as lack of sensitivity of posterior confirmatory methods or high absolute expression (not differential expression) necessity, this apparent statistically sound results could be not useful in a pragmatic sense. That is why we prefer to rank the differential expression results and allow researchers to establish a cutoff compatible with their subsequent application for the selected genes, rather than split them based in assumption-derived error-rate cutoffs. People familiar with the Frequentist Statistics framework could miss multiple testing considerations, typeI/typeII error studies, and so on. However, in the Bayesian framework, several of these concerns are meaningless since we work with parameters space and not with sample space. The bayesians avoid statements about "data that could be observed but was not" and work only with available information (prior and experimental), extracting all possible information from data effectively observed.

For those genes classified as differentially expressed, one should aggregate intuitive information adding "error-bars" to expression ratios. Recently we have developed a method to add credibility intervals to gene expression ratio [16], which could improve posterior analyses such as clustering [17] or comparison with microarray data.

Comparison with available methods using publically available data

To show the model is usefulness, we applied it to a tumor vs. normal two-classes comparison problem. We chose a subset of brain tumor SAGE data from The Cancer Genome Anatomy Project's SAGE Genie public database web-site [18]. The SAGE Genie performs several bioinformatics protocols to assure the quality of its data with systematic errors cleaning/correction [19].

We used all 4 available libraries in SAGE Genie until Jan/2004 from astrocytoma grade III tumors and almost all (except the fetal library) normal brain libraries (see Methods section for details about libraries).

We want to stress 3 typical and important cases: (i) when our measure agreed with other evidence measures accepting null hypothesis H_0, i.e., there is no evidence of differential behavior between tumor and normal classes; (ii) when our method agreed with others rejecting H_0, i.e., there is evidence of differential expression; and (iii) when our method showed evidence in favor but other evidence measures showed evidence against the H_0. Case (iii) is the main motivation of our method since it reveals situations that researchers may call a gene differentially expressed and, in fact, it could be not so significant if biological replicates are taken into account. The other evidence measures used were: the Audic-Claverie bayesian evidence [20], the classical Fisher Exact Test *P*-value, and the classical χ^2 *P*-value, all obtained using the IDEG6 web-interface [21,22] (see Methods section).

A case (i) prototype is the TTTCAATAGA tag with $\mathbf{X}_T = (0, 2, 5, 8)$ and $\mathbf{X}_N = (1, 1, 0, 0, 0, 7, 2)$. The Audic-Claverie, Fisher and χ^2 methods yield *P*-values of 0.06, 0.44, 0.41, respectively, indicating low evidence against H_0 for all mystical significance level cutoffs ≤ 0.01, ≤ 0.05 or ≤ 0.1. The Bayes Error Rate evidence is $E = 0.61$, an intuitively unacceptable superposition level between the normal and tumoral predictive Beta pdfs, showing that there is no separable behavior between classes. Figure 2a shows an obvious superposition between pdf and observations of this two classes.

A case (ii) prototype is the AAAAGAAACT tag with $\mathbf{X}_T = (7, 11, 18, 10)$ and $\mathbf{X}_N = (7, 1, 2, 1, 2, 0, 3)$. All *P*-values are 0.00 (zero), significant at any cutoff level. Our evidence is $E = 0.03$, showing safely that this gene behaves differentially between normal brain and astrocytoma grade III patients. Figure 2b shows that two Betas are apart from each other and, even observing clear within-class variability, the expression is different.

A case (iii) prototype is the TTGGAGATCT tag with $\mathbf{X}_T = (7, 239, 244, 123)$ and $\mathbf{X}_N = (54, 27, 33, 21, 40, 196, 28)$. All *P*-values are 0.00 (zero), indicating significant difference between classes. On the other hand, our evidence $E = 0.73$ indicates high superposition between tumor and normal classes. Figure 2c shows that within-class variability for tumor class is not negligible. It is obvious that individual libraries confound their results with normal brain libraries, and the Betas have a relatively high intersection. Using a common "pseudo-library" approach, one is lead to call this gene as a strong discriminator between classes. We believe that this is a suspect conclusion.

There are several other obvious case (iii) examples, such as tag TACAGTATGT in Figure 2d, that received *P*-values < 0.01 from all other methods, and they are the main concern of our method since they may lead to waste of resources in clinical validation efforts of genes that, by SAGE itself, could be left behind in favor of other promising genes. All tag results are available as additional file and graphics for all tags are at the supplemental web-site [23].

One could think about a case (iv) when considering within-class variability leads one to H_0 rejection, but considering "pseudo-libraries" leads to H_0 acceptance. This seems to be inconsistent since one expects that, once H_0 is accepted in a simplified model, it should also be accepted in the complete model. In fact, we do not observe such a situation, except by tags with *P*-values or Bayes Error Rate very close to arbitrarily defined cutoff values. We believe that these occurrences are just "edge effect" manifestations.

Discussion

In order to assure that we are dealing with a fundamental question in SAGE analysis, we show more insights analyzing the method's robustness using the same data but excluding "small" libraries. Also, we draw some parallels between our proposed method and the only available published solution for dealing with within-class variability, a *t*-test approximation [12].

We used our method with all available libraries but some of them are smaller than 50,000 tags (see Table 1). In the SAGE community, libraries smaller than this arbitrary limit are considered "small". Several researchers claim that these are non-representative and should be excluded from analysis. We observed several case (iii) tag examples which remain as case (iii) if we use libraries with size > 40,000 and > 50,000 (shown at the supplemental website only). Figure 3 shows a tag example analyzed in these tree setups and it is clear that inclusion of "small" libraries gave pretty much the same result, indicating robustness of our method against small class size variations and against "small" sized libraries. Moreover, these libraries are not always outliers from biological sampling but seem to be samples like any other. These results suggest that one can use the "small" libraries, jointly with non-"small" ones, because biological variability seems to be greater than binomial sampling variability.

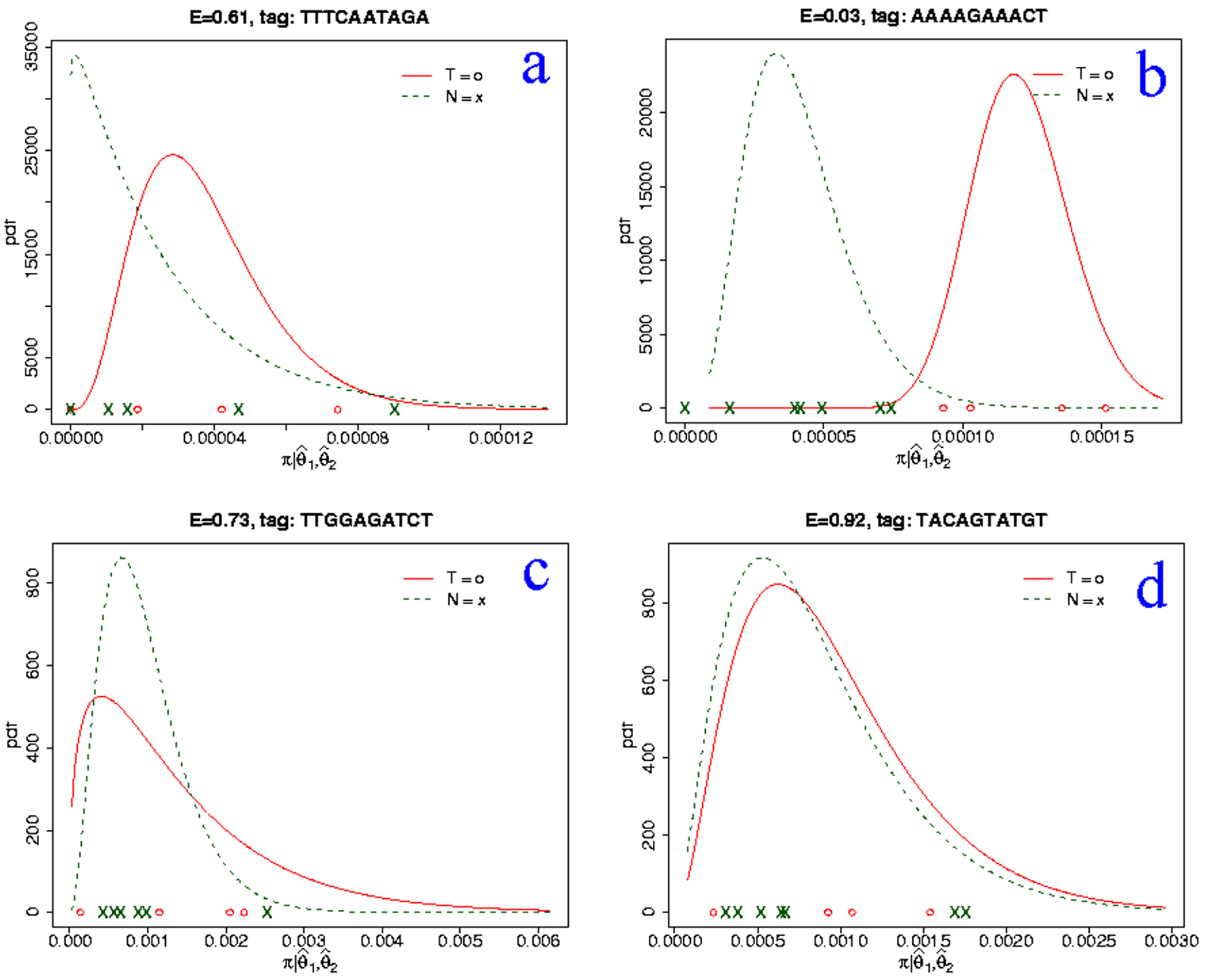


Figure 2

Maximum *a posteriori* Beta pdfs of classes in main prototype cases. The predictive pdf and the Bayes Error Rate *E* of both tumoral (T) and normal (N) classes, for examples tags contained in the three important prototype cases described in text, are shown in the figure. The 'x' and 'o' marks represent observed abundances in each tumoral and normal. Frame a) shows case (i) example when methods agree with "no differential expression" conclusion. Frame b) shows case (ii) example when methods agree with "differential expression" conclusion. Frame c) and d) show case (iii) examples when classical *P*-value method leads to significant differential expression between classes and our method indicates pdf superposition if one take within-class variability into account. Individual observations indicate that the classes are not clearly divided, casting doubt on "differential expression" conclusion.

Obviously, we are not recommending to use only "small" libraries in SAGE analysis, but suggesting that our method is relatively robust. For low expression genes, the binomial sampling variability should become more relevant as the library size decreases. Also, the results obtained using two/three libraries could be very different from using just one. These proprieties could be tag dependent since some tags could be much noisier than others for biological reasons. Some "denoising" procedure could be used before application of our method [7]. Therefore, our findings should be carefully interpreted.

To prove that the incoherence of using "pseudo-libraries" methods is not a prerogative of tags showing small fold-

Table 1: Brain tumor and normal libraries from SAGE Genie used as real data application.

#	Library Name	GEO accession	Total Tags
1	SAGE_Brain_astrocytoma_grade_III_B_H1020	GSM697	51573
2	SAGE_Brain_astrocytoma_grade_III_B_H970	GSM14763	106982
3	SAGE_Brain_astrocytoma_grade_III_B_R140	GSM14773	118733
4	SAGE_Brain_astrocytoma_grade_III_B_R927	GSM14766	107344
5	SAGE_Brain_normal_cerebellum_B_1	GSM761	50385
6	SAGE_Brain_normal_cerebellum_B_BB542	GSM695	40500
7	SAGE_Brain_normal_cortex_B_BB542	GSM676	94233
8	SAGE_Brain_normal_cortex_B_pool6	GSM763	62451
9	SAGE_Brain_normal_peds_cortex_B_H1571	GSM786	77554
10	SAGE_Brain_normal_substantia_nigra_B_1	GSM14796	42498
11	SAGE_Brain_normal_thalamus_B_1	GSM713	24015

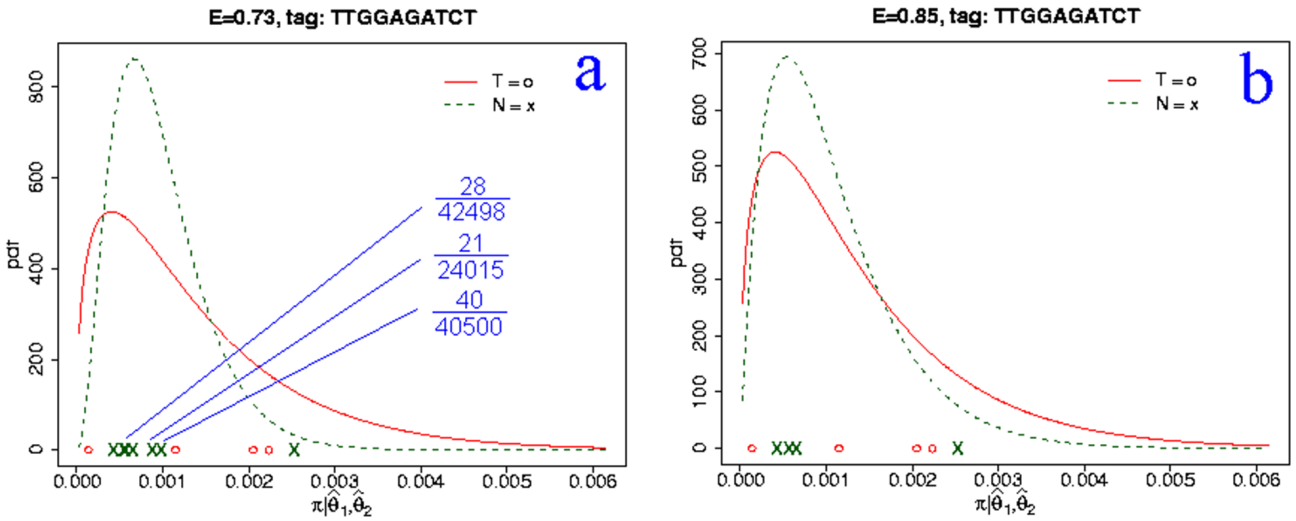


Figure 3
Effect of "small" size libraries on the final result. The predictive pdf and the Bayes Error Rate *E* of both tumoral (T) and normal (N) classes for examples tags contained in the three important prototype cases described in text are shown. The 'x' and 'o' marks represent observed abundances in each tumoral and normal libraries. Frame a) shows the result using all libraries. The "small" libraries (size < 50,000) are highlighted with their counts over library size. Frame b) shows results excluding those "small" libraries. It is clear that results are pretty much the same and that "small" libraries are not (necessarily) outliers of the sampling.

changes, we analyzed another three very illustrative examples: ATGGCAACAG, GGATGTGAAA, and GTATGGGCCC; which are case (iii) tags. These tags present high fold changes: 7.59, 8.15 and 25.80 fold-change respectively, augmented in pooled tumor libraries. Using the well-known Fisher Exact test, χ^2 classical test and the Audic-Claverie's method, we get 0.00 (zero!) for all *P*-values of the no differential expression null hypothesis. Using the conceptually different Bayesian *P*-value implemented at SAGE Genie [24,25] we obtain 0.01, 0.00 and 0.00 respectively for posterior probabilities of fold-changes greater than 4-fold. Finally, using our own proposed measure, applied to the pool, we get $E = 0.00$ meaning no superposition between the two classes pdfs. All these results indicate strong significance in differential expression of these tags.

However, if we consider within-class variability, the test proposed by Baggerly *et al.* [12] yields 0.08, 0.07 and 0.15 respectively for *t*-test *P*-values, and our method yields Bayes Error Rates of 0.38, 0.37, 0.43 respectively; indicating not so significant evidence in favor of the differentially expressed hypothesis. A closer look at the graphics of these tags induces one to believe that there is no reproducible differential expression because several observations of tumor and normal are superimposed (all graphics available at supplemental web-site [23]).

Since we show clearly that methods that use "pseudo-library" aggregation could be incoherent in some cases, a natural question is how our proposed method performs compared to the only published solution that accounts for within-class variability, the Baggerly *et al.* [12]*t*-test approximation. Without knowing the true state of all tags, it is impossible to carry out a serious benchmark. Since the interpretation of evidence measures is very different, the performance could be subjected to an arbitrary cutoff selection for each method. Figure 4 shows a scatter-plot of evidence measures obtained for each of the two methods.

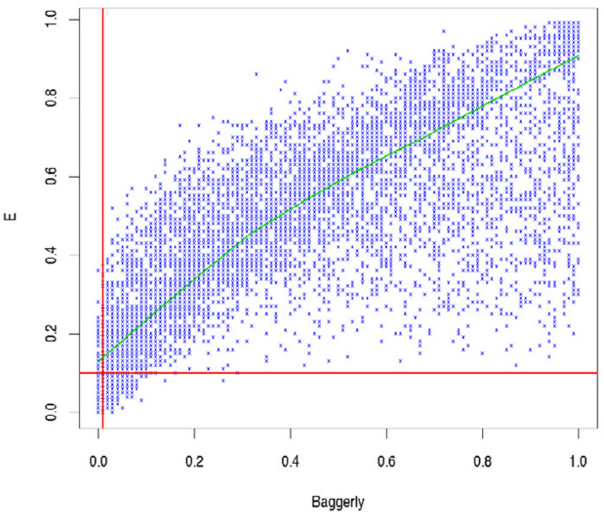


Figure 4
Qualitative comparison of Bayes Error Rate and *t*-test approximation. It is shown the Bayes Error Rate (*E*) versus the *t*-test approximation *P*-value (Baggerly) [12] for each tag. The red lines are arbitrary cutoffs that define significance regions E ≤ 0.1 and Baggerly ≤ 0.01. The green line is a LOWESS trend fit.

It is clear from this graphic that there are many more tags considered as differentially expressed using our method than the *t*-test approximation, considering $E \leq 0.1$ and *P*-value ≤ 0.01. There are also some tags selected by *t*-test and ignored by ours. It is impossible to know which method perform better without the true unknown status of those tags. Looking at individual libraries results, constructed as depicted in Figure 2 for example, could help in this analysis but this is a subjective procedure.

It is important to bear in mind that a difficulty is hidden in the Beta modeling imposed in the very first beginning. If Beta is not a good model for an unknown biological behavior, then some apparent inconsistency could appear in both Baggerly *et al.* [12] and our approaches. However, our general mixture model allows another propositions. Other simplex constrained pdfs, different from Beta, exist but the tractability is much more difficult [26]. We believe that to build a fully non-parametric approach to this problem is a very hard issue, but should be considered as a future challenge.

Conclusion

Until now, almost all statistical methods for SAGE data analysis tacitly ignore the within-class variability. To our knowledge, the firsts to formally address this issue was Baggerly *et al.* [12] who introduced the Beta-Binomial model as the correct way to model the probability of counting tags instead of Binomial models. They also proposed a *t*-like statistics, outlined a possible hypothesis test using the classical Frequentist Statistics framework and evoked some asymptotic results for *t* pdf justification.

In this work we presented the Bayesian alternative for this problem and defined a theoretical model that views Baggerly's Beta-Binomial approach or even the common Binomial approach as particular cases of mixture models. Other models are possible modifying the mixing distribution, such as Beta-Poisson [14], or using other simplex constrained pdf [26] to model expression abundance. At last, but not at least, we proposed a method for ranking differentially expressed genes between two classes using the Bayes Error Rate as an intuitive measure of separation between the classes pdfs, avoiding statistical test formalism and its conceptual/practical difficulties.

We show that there are cases in which approaches that ignore within-class variability will lead to high significance in differences between tumor and normal classes, but looking carefully at individual observations jointly, one should not attribute such high significance to them since abundance probability density functions have considerable superposition.

In conclusion, we recommend that within-class variability must be taken into account in any statistical analysis of SAGE data if replicates are available. We suggest that biological replication should be considered in planning new SAGE experiments.

Methods

General bayesian model

To start generically, suppose that the probability density function (pdf) for the random variable of interest "expression abundance" $\pi \in [0;1]$ of some gene *G* is indexed in a model family by means of a parameter vector θ. Therefore, following the usual Bayesian framework, the *a posteriori* pdf that describes the class is:

$$g(\theta \mid \mathbf{X},\mathbf{M}) = \frac{g(\theta)\cdot\prod_{i=1}^{n} L(x_i \mid m_i,\theta)}{\int_{\Theta} g(\theta')\cdot\prod_{i=1}^{n} L(x_i \mid m_i,\theta')\,d\theta'} \qquad (3)$$

where: $\mathbf{X} = (x_1,..., x_n)$ is the vector of counts in all n libraries of same class, $\mathbf{M} = (m_1,..., m_n)$ is the vector of total observations in all n libraries of same class, $g(\cdot)$ is the *a priori* pdf, and L is the likelihood of each i-th observation. Note that the product of all likelihood functions over all observations is the so-called Likelihood Function.

The counting process from automatic sequencing is often modeled as a Binomial. Since the sample size and the stopping rule are not known in advance the model is not strictly Binomial. We do not need the combinatorial constant in the model, but we write it just because it is commonly used and will vanish in *posteriori* expression anyway.

"Pseudo-Library" method as particular case

Merging all observations from the same class libraries and constructing "pseudo-libraries", with the sum of their components, is the standard procedure to use replicates. Our general model is reduced to this (unrealistic) one if one uses $f(\cdot)$ as a Dirac's Delta in Eq.1:

$$L(x_i \mid m_i,\theta) = \int_{-\infty}^{\infty} 1_{\{0\le\pi\le1\}}\,\delta(\pi-\theta)\binom{m_i}{x_i}(1-\pi)^{m_i-x_i}\,\pi^{x_i}\,d\pi = \binom{m_i}{x_i}(1-\theta)^{m_i-x_i}\,\theta^{x_i} \qquad (4)$$

where: $\mathbf{1}\{\cdot\}$ is the indicator function.

Using Eq.4 in Eq.3 yield:

$$g(\theta \mid \mathbf{X},\mathbf{M}) = \frac{(1-\theta)^{\sum_{i=1}^{n} m_i - \sum_{i=1}^{n} x_i}\,\theta^{\sum_{i=1}^{n} x_i}}{\int_0^1 (1-\theta')^{\sum_{i=1}^{n} m_i - \sum_{i=1}^{n} x_i}\,\theta'^{\sum_{i=1}^{n} x_i}\,d\theta'} \qquad (5)$$

where: $g(\theta) = 1$, the non-informative uniform *a priori* distribution.

The expert recognizes that $\theta \sim \mathrm{Beta}(1 + \Sigma x_i, 1 + \Sigma m_i - \Sigma x_i)$, and the sum of observations is the mathematical translation of "pseudo-libraries" construction.

Beta-Binomial method as particular case

The only published solution that allows non-zero within-class variance in SAGE analysis is the Beta-Binomial model [12]. Using $f(\cdot)$ as a Beta in Eq.1 we get the Beta-Binomial model as a particular case of general model:

$$L(x_i \mid m_i,\theta) = \int_0^1 \frac{(1-\pi)^{\beta_\theta-1}\,\pi^{\alpha_\theta-1}}{B(\alpha_\theta,\beta_\theta)}\binom{m_i}{x_i}(1-\pi)^{m_i-x_i}\,\pi^{x_i}\,d\pi = \binom{m_i}{x_i}\frac{B(\alpha_\theta + x_i,\beta_\theta + m_i - x_i)}{B(\alpha_\theta,\beta_\theta)} \qquad (6)$$

where: $B(\cdot)$ is the beta special function, and:

$$\begin{cases} \alpha_\theta = \dfrac{\theta_1^2(1-\theta_1)}{\theta_2^2} - \theta_1 \\[2ex] \beta_\theta = \dfrac{\theta_1(1-\theta_1)}{\theta_2^2} - 1 - \dfrac{\theta_1^2(1-\theta_1)}{\theta_2^2} + \theta_1 \end{cases} \Leftrightarrow \begin{cases} \theta_1 = \dfrac{\alpha_\theta}{\alpha_\theta + \beta_\theta} \\[2ex] \theta_2 = \sqrt{\dfrac{\alpha_\theta\beta_\theta}{(\alpha_\theta + \beta_\theta)^2(\alpha_\theta + \beta_\theta + 1)}} \end{cases} \qquad (7)$$

Again, an expert recognizes $\theta = (\theta_1, \theta_2)$ as the mean and standard deviation (stdv) of a Beta random variable. We prefer this parameterization of Beta distributions instead of the common (α, β) one because: (i) it is much more intuitive to biologists to deal with mean and stdv than with abstract α and β, and (ii) as $\alpha, \beta > 0$, the domain $\Theta = \{(\theta_1, \theta_2)\colon 0 \le \theta_1 \le 1,\ 0 \le \theta_2^2 < \theta_1 (1 - \theta_1) \le 1/4\}$ is bounded and much more amenable to perform the necessary numerical computations.

Using Eq.6 in Eq.3 yield:

$$g(\theta_1,\theta_2 \mid \mathbf{X},\mathbf{M}) = \frac{g(\theta_1,\theta_2)\cdot\prod_{i=1}^{n}\dfrac{B(\alpha_\theta + x_i,\beta_\theta + m_i - x_i)}{B(\alpha_\theta,\beta_\theta)}}{\int_0^1\int_0^{\sqrt{\theta_1'(1-\theta_1')}} g(\theta_1',\theta_2')\cdot\prod_{i=1}^{n}\dfrac{B(\alpha_{\theta'} + x_i,\beta_{\theta'} + m_i - x_i)}{B(\alpha_{\theta'},\beta_{\theta'})}\,d\theta_2' d\theta_1'} \qquad (8)$$

where: $g(\theta_1, \theta_2)$ is the *priori* pdf.

A Priori *distribution definition*

To complete a Bayesian model, it is necessary to choose the *a priori* pdf. We use an uniform distribution over Θ. On the other hand, we know in advance that variance of this model cannot be smaller than the variance eventually obtained if we do not consider within-class variability. Even if the within-class variability is very small, it cannot be estimated as being smaller than the simple sampling error because they are inseparable, and sampling error is the lower limit [12]. As an illustration, the same situation could occur if one takes several diameter measurements of a folded paper ball and a perfect sphere using a common ruler. In the first case, the intrinsic nature of the measured object dominates the measurement variability but, in the second case, we cannot know the diameter of the perfect sphere with better precision than our ruler can measure.

This kind of knowledge is naturally incorporated in Bayesian statistics by means of *a priori* distributions. To match our desired features, it is sufficient to define an uniform over the Θ parameter space but constrained at a minimum stdv σ, obtained from the result of no within-class variance model:

$$g(\theta_1,\theta_2) \propto 1_{\{\theta_2 \geq \sigma\}} \qquad (9)$$

over domain $\Theta = \{(\theta_1, \theta_2): 0 \leq \theta_1 \leq 1, 0 \leq \theta_2{}^2 < \theta_1 (1 - \theta_1) \leq 1/4\}$.

Since we showed (Eq.5) that the no within-class variance model is $\theta \sim \text{Beta}(1 + \Sigma x_i, 1 + \Sigma m_i - \Sigma x_i)$, it is easy to obtain our lower stdv boundary from Beta variance:

$$\sigma = \sqrt{\frac{(1+\sum_{i=1}^{n} x_i)(1+\sum_{i=1}^{n} m_i - \sum_{i=1}^{n} x_i)}{(2+\sum_{i=1}^{n} m_i)^2 (3+\sum_{i=1}^{n} m_i)}} \qquad (10)$$

Therefore, using Eq.9 and Eq.10 in Eq.8, our *posteriori* is completely defined.

Differential expression detection

We detect a tag as differentially expressed using the Bayes Error Rate E [15] in both predictive Beta pdfs:

$$E = \int_0^1 \min\left(f\left(\pi \mid \theta^A\right), f\left(\pi \mid \theta^B\right)\right) d\pi$$

where:

$$f(\pi \mid \theta) = \frac{(1-\pi)^{\beta_\theta - 1}\, \pi^{\alpha_\theta - 1}}{B(\alpha_\theta, \beta_\theta)} \qquad (11)$$

Note that $f(\cdot)$ is the Beta pdf, as in Eq.6 development. The "hat" over $\theta = (\theta_1, \theta_2)$ indicates values that leads Eq.8 to maximum. As usual, the maximization, subject to constrain Θ defined previously, is made upon logarithm of *posteriori's* core since it gives the same estimates as the *posteriori* itself:

$$C = \ln(1_{\{\theta_2 \geq \sigma\}}) - n\ln(B(\alpha_\theta, \beta_\theta)) + \sum_{i=1}^{n} \ln(B(\alpha_\theta + x_i, \beta_\theta + m_i - x_i)) \qquad (12)$$
$$\hat{\theta} = \arg_\theta\left(\max_\Theta(C)\right)$$

Figure 5 shows an example of this process. See Results section to get an intuitive notion of this evidence measure.

Implementation – numerical analysis

The method was implemented as R language [27,28] scripts which are freely available under GPL/GNU copyleft license at supplemental web site [23]. At this web page there are details on how to run it locally.

Our method is computer-intensive mainly because some numeric maximization and integration are needed. We used efficient R built-in routines to perform such numerical tasks. Remember that maximization needed in Eq.12 is constrained, thus we used simply auxiliary re-parameterization to obtain linear constrains and used the constrOptim R routine. For numerical integration we used the 1-dimensional gaussian quadrature integrate R built-in function. Although numerical integration of Eq.9 should be performed in all [0;1] support, the relevant contribution for this integral is concentrated in a much smaller region. Integrating over the formal limits will cause serious numerical errors, and to avoid this problem we approximate our integration region to an interval delimited by 0.005 and 0.995 quantile of each Beta pdf since the relevant density lie in there.

The credibility intervals ("error-bars") for the expression ratio of interesting tags were obtained as described in our recent work [16]. We chose arbitrarily 68% credibility intervals.

Implementation – Web based interface

We have also developed an easy-to-use web-based service that performs all calculations at our server and provides password-protected results. Although desirable, for the sake of automatic web hyperlink with SAGE Genie database, it is not necessary to explicitly identify the tags analyzed but rather any (custom) i.d. string. This could increase privacy or make our web-interface useful for "Digital-Northern", MPSS or any mathematically related problem of mixtures from binomial sampling. Figure 6 shows snapshots of the interface.

Publically available data

The Table 1 list the SAGE Genie's library name, Gene Expression Omnibus (GEO) [29] accession code and size of all used libraries.

For our aims, it is sufficient to focus the analysis at the tag level. Thus, we process the tag counts and let the identification of tag's best gene match as a posterior question that could be carefully done only to really interesting tags. We choose not to process tags whose counts appear only in libraries of one class. It is important to note that all libraries are from bulk material, without cell-lines, and came from patients with similar disease description. The normal libraries came from different normal regions of the brain.

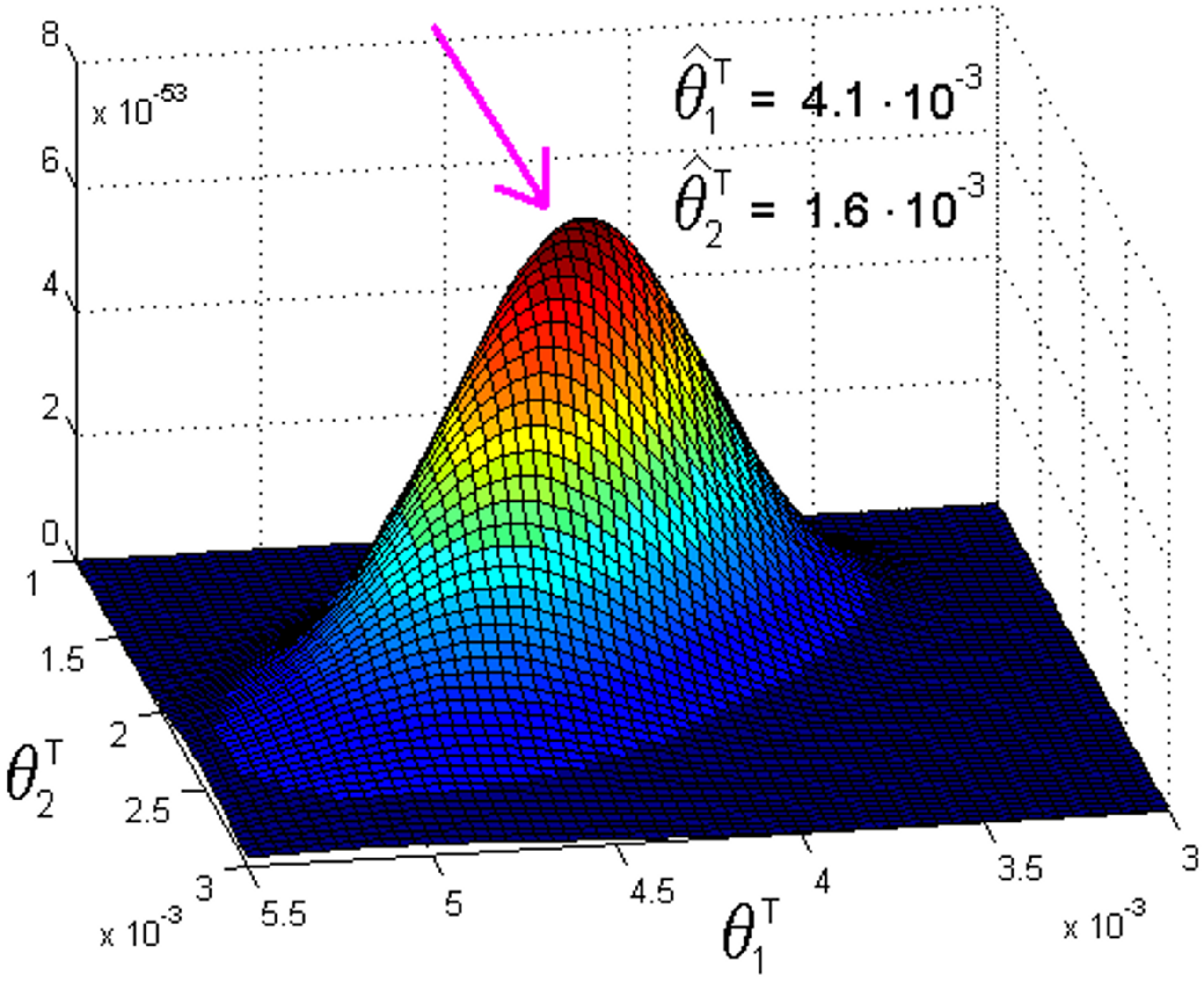


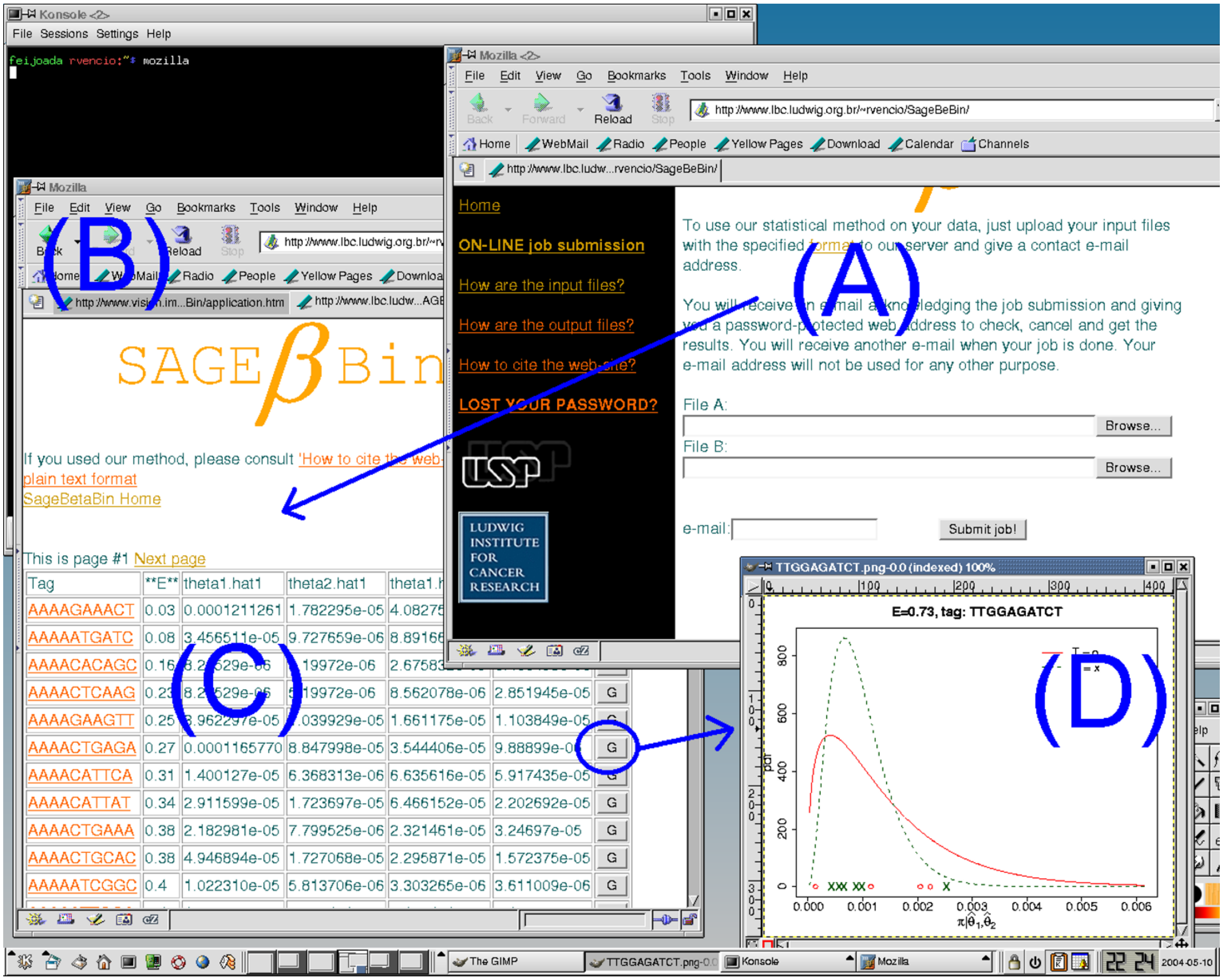
Figure 5
Illustration of Maximum *a Posteriori* parameter estimation. The maximum a posteriori parameters of Eq.12 in an artificial example it is shown. The bi-dimensional pdf from which "hat" (pointed by the arrow) parameters are extracted is proportional to that described in Eq.8.

We think that this data set is very illustrative since there are biological replicates in the tumor class allowing clear verification of within-class biological variability. On the other hand, taking only one kind of disease, astrocytoma grade III, instead of all brain tumors in the database, leads one to believe that the within-class variability is in fact due to biological diversity of the patients and not due to very distinct molecular profile of distinct brain tumors stored in SAGE Genie's database.

Therefore, we believe that this *in silico* comparison is well-suited to demonstrate the necessity of dealing with within-class effect, although it is not our aim here to make a detailed or biological analysis of brain tumor data.

Comparison with other methods

To bring some intuition about our differential expression evidence measure, we tabulated evidence measure obtained from the famous Fisher Exact Test, the classical Pearson's χ^2 proportion test and the bayesian Audic-Claverie's method. All these tests were performed using the easy-to-use web-interface IDEG6 [21,22].



```
feijoada rvencio:~$ mozilla
```

http://www.lbc.ludwig.org.br/~rvencio/SageBeBin/

http://www.lbc.ludw...rvencio/SageBeBin/

http://www.vision.im...Bin/application.htm

http://www.lbc.ludw...AGE

SAGEβBin

Home

ON-LINE job submission

How are the input files?

How are the output files?

How to cite the web-site?

LOST YOUR PASSWORD?

To use our statistical method on your data, just upload your input files with the specified format to our server and give a contact e-mail address.

You will receive an e-mail acknowledging the job submission and giving you a password-protected web address to check, cancel and get the results. You will receive another e-mail when your job is done. Your e-mail address will not be used for any other purpose.

File A: Browse...

File B: Browse...

e-mail: Submit job!

If you used our method, please consult 'How to cite the web-
plain text format
SageBetaBin Home

This is page #1 Next page

Tag	**E**	theta1.hat1	theta2.hat1	theta1.h		
AAAAGAAACT	0.03	0.0001211261	1.782295e-05	4.08275		
AAAAATGATC	0.08	3.456511e-05	9.727659e-06	8.89166		
AAAACACAGC	0.16	8.2[illegible]529e-06	[illegible]19972e-06	2.67583		
AAAACTCAAG	0.23	8.2[illegible]529e-06	5[illegible]19972e-06	8.562078e-06	2.851945e-05	G
AAAAGAAGTT	0.25	[illegible].962297e-05	[illegible].039929e-05	1.661175e-05	1.103849e-05	G
AAAACTGAGA	0.27	0.0001165770	8.847998e-05	3.544406e-05	9.88899e-0[illegible]	G
AAAACATTCA	0.31	1.400127e-05	6.368313e-06	6.635616e-05	5.917435e-05	G
AAAACATTAT	0.34	2.911599e-05	1.723697e-05	6.466152e-05	2.202692e-05	G
AAAACTGAAA	0.38	2.182981e-05	7.799525e-06	2.321461e-05	3.24697e-05	G
AAAACTGCAC	0.38	4.946894e-05	1.727068e-05	2.295871e-05	1.572375e-05	G
AAAAATCGGC	0.4	1.022310e-05	5.813706e-06	3.303265e-06	3.611009e-06	G

Figure 6

Snapshot of the web-interface for our SAGEbetaBin method. An illustration of on-line tool implemented to make our method easily available it is shown. Researchers submit their data (A) and receive, by e-mail, an alert when we finished the job along with instructions to get results in a password-protected web-page (B). If the ID supplied is a human tag (C), then results are linked with SAGE genies' tag-to-gene map. The individual observations graphics, as utilized in this work, is available on-demand (D).

The "*P*-values" are conceptually very different from our evidence measure but are the most used evidence measures. Although numbers cannot be compared, the conclusions obtained from these methods should be since graphical representation of each library observation gives clear indication of incoherence of "pseudo-library" methods. The results of the significance measures for all tags are available as additional Excel© or OpenOffice© interactive files in which the user can set cutoffs for the significance measures, and explore the differences in conclusions.

We carry out a qualitative comparison of our method with Baggerly *et al.* [12]*t*-test approximation in a graphical way since it is impossible to judge them without the unknown true status of analyzed tags, given the too different interpretation of numeric values returned. In their Frequentist framework, the estimator $p_i = x_i/m_i$ is used for π_i and a linear combination of these abundances is proposed as the correct way to combine results from different libraries:

$$p = \sum_{i=1}^{n} w_i p_i \ , \ V_\mathrm{u} = \frac{\sum_{i=1}^{n} w_i^2 p_i^2 - p^2 \sum_{i=1}^{n} w_i^2}{1 - \sum_{i=1}^{n} w_i^2} \ , \ w_i \propto \frac{m_i(\alpha+\beta)}{\alpha+\beta+m_i} \qquad (13)$$

where w_i are the weights that yield an unbiased minimum variance estimator V_u for weighted proportion's variance and $\theta = (\alpha, \beta)$ 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:

$$V = \max\,[V_u; V_{pseudo\text{-}lib}] \qquad (14)$$

where:

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

The max(·) function assure that V is not unrealistic small when V_u is unrealistic small. 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, these authors test if the proportions are significantly different proposing the use of a t_w statistics as following a Student's t_{df} pdf:

$$t_w = \frac{p_A - p_B}{\sqrt{V_A + V_B}} \ , \ df = \frac{(V_A + V_B)^2}{\frac{V_A^2}{\sum_A m_i - 1} + \frac{V_B^2}{\sum_B m_i - 1}} \qquad (15)$$

List of Abbreviations

SAGE: Serial Analysis of Gene Expression

MPSS: Massively Parallel Signature Sequencing

EST: Expressed Sequence Tag

pdf: probability density function

GEO: Gene Expression Omnibus

Authors' Contributions

RV conceived and executed this work. HB helped with all biological issues. DFCP helped in differential expression detection methods and implemented the on-line web-based tool. CABP helped with Bayesian statistics and proposed the mixture ideas.

Additional material

Additional File 1

Results for all evidence measures*. This file allows the user to interactively define significance cutoffs for ranked tags. The ranks are based on evidence measures against "no differential expression" hypothesis, i.e., evidences closer to 0 (zero) denote higher confidence in differential expression and closer to 1 (one) denote no evidence of differential expression.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-5-119-S1.xls]

Acknowledgements

RV is supported by FAPESP 02/04698-8 fellowship. We thank Tie Koide for critical reading of the manuscript and BIOINFO-USP/Rede-Vision for computational support.

References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.
2. Margulies EH, Kardia SL, Innis JW: **Identification and prevention of a GC content bias in SAGE libraries.** *Nucleic Acids Res* 2001, **29:**e60.
3. Stern MD, Anisimov SV, Boheler KR: **Can transcriptome size be estimated from SAGE catalogs?** *Bioinformatics* 2003, **19:**443-448.
4. Stollberg J, Urschitz J, Urban Z, Boyd CD: **A Quantitative Evaluation of SAGE.** *Genome Research* 2000, **10:**1241-1248.
5. Akmaev VR, Wang CJ: **Correction of sequence based artifacts in serial analysis of gene expression.** *Bioinformatics* 2004, **20:**1254-1263.
6. Morris JS, Baggerly KA, Coombes KR: **Bayesian shrinkage estimation of the relative abundance of mRNA transcipts using SAGE.** *Biometrics* 2003, **59:**476-486.
7. Blades NJ, Jones JB, Kern SE, Parmigiani G: **Denoising of data from serial analysis of gene expression.** *Bioinformatics* in press.
8. Brenner S, Johnson M, Bridgham J: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.** *Nature Biotechnology* 2000, **18:**630-634.
9. Man MZ, Wang X, Wang Y: **POWER_SAGE: comparing statistical tests for SAGE experiments.** *Bioinfo matics* 2000, **16:**953-959.
10. Romualdi C, Bortoluzzi S, Danieli GA: **Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests.** *Hum Mol Genet* 2001, **19:**2133-2141.
11. Ruijter JM, Kampen AHC, Baas F: **Statistical evaluation of SAGE libraries: consequences for experimental design.** *Physiol Genomics* 2002, **11:**37-44.
12. Baggerly KA, Deng L, Morris JS, Aldaz CM: **Differential expression in SAGE: accounting for normal between-library variation.** *Bioinformatics* 2003, **19:**1477-1483.
13. Aitchison J, Dunsmore IR: *Statistical Prediction Analysis* Cambridge: Cambridge University Press; 1975.
14. Bueno AMS, Pereira CAB, Rabello-Gay MN, Stern JM: **Environmental genotoxicity evaluation: Bayesian approach for a mixture statistical model.** *Stochastic Environmental Research and Risk Assessment* 2002, **16:**267-278.
15. Duda RO, Hart PE, Stork DG: *Pattern Classification* 2nd edition. New York: Wiley-Interscience Press; 2000.
16. Vêncio RZN, Brentani H, Pereira CAB: **Using credibility intervals instead of hypothesis tests in SAGE analysis.** *Bioinformatics* 2003, **19:**2461-2464.

17. Yeung KY, Medvedociv M, Bumgarner RE: **Clustering gene-expression data with repeated measurements.** *Genome Biology* 2003, **4:**R34.1-R34.17.
18. **SAGE Genie Home Page** [http://cgap.nci.nih.gov/SAGE]
19. Boon K, Osório EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, Souza SJ, Riggins GJ: **An anatomy of normal and malignant gene expression.** *Proc Natl Acad Sci USA* 2002, **99:**11287-11292.
20. Audic S, Claverie J: **The significance of digital gene expression profiles.** *Genome Research* 1997, **7:**986-995.
21. Romualdi C, Bortoluzzi S, D'Alessi F, Danieli GA: **IDEG6: a web tool for detection of differentially expressed genes in multiple tag sampling experiments.** *Physiol Genomics* 2003, **12:**159-162.
22. **IDEG6 Home Page** [http://telethon.bio.unipd.it/bioinfo/IDEG6_form/]
23. **SAGEbetaBin Home Page** [http://www.vision.ime.usp.br/~rvencio/SAGEbetaBin]
24. Lal A, Lash AE, Altschul SF, Velculescu V, Zhang L, McLendon RE, Marra MA, Prange C, Morin PJ, Polyak K, Papadopoulos N, Vogelstein B, Kinzler KW, Strausberg RL, Riggins GJ: **A public database for gene expression in human cancers.** *Cancer Research* 1999, **21:**5403-5407.
25. Chen H, Centola M, Altschul SF, Metzger H: **Characterization of gene expression in resting and activated mast cells.** *J Exp Med* 1998, **188:**1657-1668.
26. Aitchison J: **A General class of distributions on the simplex.** *J R Statistical Soc B* 1985, **47:**136-146.
27. Ihaka R, Gentleman R: **R: A language for data analysis and graphics.** *Journal of Computational and Graphical Statistics* 1996, **5:**299-314.
28. **R project Home Page** [http://r-project.org]
29. **Gene Expression Omnibus Home Page** [http://www.ncbi.nlm.nih.gov/geo]