BIOINFORMATICS

Vol. 20 Suppl. 1 2004, pages i31–i39
DOI: 10.1093/bioinformatics/bth924

Statistical modeling of sequencing errors in SAGE libraries

Tim Beißbarth[1,], Lavinia Hyde[1], Gordon K. Smyth[1], Chris Job[2], Wee-Ming Boon[2], Seong-Seng Tan[2], Hamish S. Scott[1] and Terence P. Speed[1]*

[1]Walter and Eliza Hall Institute of Medical Research, Genetics and Bioinformatics, 1G Royal Parade, Parkville, Vic 3050, Australia and [2]Howard Florey Institute, Brain Development Laboratory, University of Melbourne, Parkville, Vic 3010, Australia

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: Sequencing errors may bias the gene expression measurements made by Serial Analysis of Gene Expression (SAGE). They may introduce non-existent tags at low abundance and decrease the real abundance of other tags. These effects are increased in the longer tags generated in LongSAGE libraries. Current sequencing technology generates quite accurate estimates of sequencing error rates. Here we make use of the sequence neighborhood of SAGE tags and error estimates from the base-calling software to correct for such errors.

Results: We introduce a statistical model for the propagation of sequencing errors in SAGE and suggest an Expectation-Maximization (EM) algorithm to correct for them given observed sequences in a library and base-calling error estimates. We tested our method using simulated and experimental SAGE libraries. When comparing SAGE libraries, we found that sequencing errors can introduce considerable bias. High abundance tags may be falsely called as significantly differentially expressed, especially when comparing libraries with different levels of sequencing errors and/or of different size. Truly, differentially expressed tags have decreased significance as 'true'-tag counts are generally underestimated. This may alter if tags near the threshold of differential expression are called significant. Moreover, the number of different transcripts present in a library is overestimated as false tags are introduced at low abundance. Our correction method adjusts the tag counts to be closer to the true counts and is able to partly correct for biases introduced by sequencing errors.

Availability: An implementation using R is distributed as an R package. An online version is available at http://tagcalling.mbgproject.org

Contact: beissbarth@wehi.edu.au

*To whom correspondence should be addressed.

1 INTRODUCTION

Serial Analysis of Gene Expression (SAGE) is a gene expression profiling technique that estimates the abundance of thousands of gene transcripts (mRNAs) from a cell or tissue sample in parallel (Velculescu *et al*., 1995). SAGE is based on the sequencing of short sequence tags that are extracted at defined positions of the transcript. As opposed to DNA microarray technology (Schena *et al*., 1995; Lockhart *et al*., 1996), SAGE does not require prior knowledge of the transcripts, and results in an estimate of the absolute abundance of a transcript. However, due to sequencing errors a proportion of the low-abundance tags do not represent real genes altering the ability of SAGE to estimate the number of transcripts that have been observed. Moreover, loss of 'true'-tags due to sequencing errors will result in altered numbers for the abundance of genuine transcripts.

Stollberg *et al*. (2000) have studied the effects of various sources of errors on SAGE results by simulating libraries. Previously sequencing errors have been minimized by removing low-abundance tags or tags with low sequence quality from the libraries (Margulies and Innis, 2000). Velculescu *et al*. (1999) first attempted to join low-abundance tags to their neighborhood. A more refined approach, that assumes constant error probabilities and uses matrix inversion to correct for sequencing errors, has been introduced by Colinge and Feger (2001). Another recently developed approach by Blades *et al*. (2004a,b) uses a linear relation between the copy number of observed tags and the number of neighbors with one-base substitutions to estimate the average rate of sequencing errors and eliminate unreliable tags. Akmaev and Wang (2004) use such error estimates to correct for sequencing errors and PCR based artifacts. They estimate that in LongSAGE libraries 3.5% of the tag sequences have errors resulting from PCR artifacts and 17.3% of the tag sequences have errors resulting from sequencing errors. These approaches, however, do not take into account estimates for sequencing errors of the individual bases.

The two main base-calling programs, the open source program Phred and the ABI KB basecaller, distributed with the ABI 3730 sequencing machines (Applied Biosystems), both assign a quality score to each sequenced base (Ewing and Green, 1998). The quality score is given as $-10\log_{10} P_e$, where P_e is the probability of a base-calling error.

Here, we introduce a novel method to correct for biases of sequencing errors in SAGE libraries, as well as an implementation to extract tags from the sequences of a library. Extraction of tags from the sequence runs and correction of biases resulting from sequencing errors are the basis for further analysis of SAGE libraries, such as the comparisons between different libraries (Man *et al*., 2000; Baggerly *et al*., 2003) and the assignment of known genes to their corresponding tag sequences (Lash *et al*., 2000). We show that our correction method has a significant effect on further analysis.

2 SYSTEMS AND METHODS

2.1 Generation of SAGE libraries

Briefly, SAGE works as follows: RNAs from either cells or tissues are converted into double-stranded cDNA, which is anchored to a solid phase at the 3′ end. The double-stranded cDNA is then cleaved with a restriction endonuclease at a 4-bp recognition sequence, most commonly CATG. The 3′ ends of these cDNA fragments are collected and are then divided into two populations and ligated to linkers containing a type IIS restriction endonuclease recognition sequence, where the enzyme cleaves up to 20 bp away from their recognition site. The two populations are ligated together and amplified by PCR, resulting in two tags orientated tail to tail with an anchoring enzyme recognition site at either end. Two types of SAGE libraries are commonly used, generating tags of different length, i.e. 10 and 17 base tags, respectively, depending on the enzyme used. All libraries were obtained as described (http://www.sagenet.org/sage_protocol.htm) on different mouse tissues using NlaIII as the anchoring enzyme. The E15 library was generated from posterior cortex of embryonic C57/BL6 mice at stage E15.5. The B6Hypothal library was generated from hypothalamus of 8-week-old C57/BL6 mice.

2.2 Base-calling and extraction of SAGE tags

SAGE libraries were generated from between 1000 and 5000 sequenced clones, with each sequence run consisting of up to 40 tags. Automated sequencers generate a four-color chromatogram showing the results of the sequencing gel. These chromatograms are read by the Phred or ABI software to call bases and assign an error estimate for each base. The resulting Phred or ABI files are read by functions implemented in R which subsequently extract the ditags and tags between the anchoring enzyme sites (CATG) in the sequence, keeping the error scores with each base. Ditags have to be within a specified length range, e.g. 20–24 bases for 10 base tags or 32–38 bases for 17 base tags. Duplicate ditags are removed to reduce possible PCR bias, keeping the ditag with the highest average sequencing quality. Tag sequences with a low average sequence quality ($\leq$10) are also removed. From experimental SAGE libraries usually 20 000–100 000 tag sequences are generated.

2.3 Simulation of SAGE libraries

In order to test our method, we simulated SAGE data with sequencing errors. We set the number of possible transcripts to 100 000 and assign a random SAGE tag to each of them out of all 4^{10} or 4^{17} possible SAGE tags. For each SAGE tag a random proportion p within the library is generated from a log-normal distribution, and the proportions are then adjusted to have a sum of 1. The true counts of a tag are simulated by sampling from Poisson distributions with parameters $p\lambda$, where p is the proportion of the tag in the library and λ is a parameter for setting the size of the library. The simulation of the sequencing errors is done on each individual occurrence of a tag sequence. For each tag sequence, a mean sequencing quality value is generated from a log–normal distribution. The individual quality values for each base are then generated from log–normal distributions with means equal to the simulated sequencing quality values for the tag sequences. We have noticed that with experimentally generated data the within tag sequence variation of sequencing quality values is usually about one-fifth of the between tag sequence variation. From each true tag sequence one observed tag sequence is generated using the simulated quality values of the true sequence as the multinomial probabilities, i.e. replacing each base with either one of the three other bases with the probability specified by the sequencing quality value of that base. The counts of these generated tags are then summed to represent the observed tags. When generating several simulated libraries for comparisons, we use the same proportions of the genes for all libraries, replacing up to one-third of the proportions by proportions with a known differential factor. We used a variety of parameters for the distributions for testing and in order to closely resemble our experimental data.

2.4 Implementation in R

We have implemented all our functions in R (http://www.r-project.org). For efficient computation we use the SparseM package by Roger Koenker available at CRAN (http://cran.r-project.org). In a sparse matrix the non-zero values are represented by a vector ra, the column indexes of the values are stored in a vector ja and the relative start positions of each row of the matrix in vector ra are stored in a vector ia. A collection of our functions is available as an R package and will be distributed with the Bioconductor bundle (Gentleman and Carey, 2002). Furthermore, we have generated a web interface using Perl scripts and the Apache web server, which is accessible at http://tagcalling.mbgproject.org

2.5 Comparison of SAGE libraries

SAGE tags are assessed for differential expression between two SAGE libraries by computing Fisher's Exact test for each unique tag. If a particular tag has count n_A in library A and count n_B in library B, and if the total number of sequences counted is t_A for library A and t_B for library B, then Fisher's Exact test is computed to test for independence in the 2×2 contingency table with counts n_A, n_B, $t_A - n_A$ and $t_B - n_B$. This results in a P-value for the null hypothesis of no differential expression for each gene. Since the tests for different tags are almost independent, the method of Benjamini and Hochberg (1995) was used to control the false discovery rate (fdr). Fisher's Exact test has been found to be slow to compute but an exact binomial test proved to be an excellent approximation when t_A and t_B are large and large relative to n_A and n_B, as they are for typical SAGE libraries. This test is defined similarly to Fisher's Exact test but with binomial probabilities replacing the hypergeometric probabilities. We used a vectorized version of the binomial exact test to allow rapid computation for complete libraries. By analogy with microarray analysis the relative difference of a tag between two libraries is summarized by an M-value, which is calculated as $\log_2(n_A + 0.5) + \log_2(t_B - n_B + 0.5) - \log_2(n_B + 0.5) - \log_2(t_A - n_A + 0.5)$, and the mean absolute expression is summarized as an A-value, which is calculated as $0.5[\log_2(n_A(t_A + t_B)/2t_A + 0.5] + \log_2[n_B(t_A + t_B)/2t_B + 0.5)]$. We call changes with a fdr of less than 0.1 significant.

3 MODEL AND ALGORITHM

A SAGE library consists of short sequence tags. These tags are usually either 10 or 17 bases long, this means depending on the type of the library either 4^{10} or 4^{17} different sequences of the tag-length exist which could contribute a tag. We enumerate all the possible tags with $1, \ldots, N$. Each tag sequence can be observed 0, 1 or several times in a given SAGE library, we call the observed counts $n_0, \ldots, n_N$. The true counts for the tags, i.e. without sequencing errors, are denoted by $m_0, \ldots, m_N$. For illustration see Figure 1. During the sequencing process each true tag generates sequences, which are either the same as the original or including sequencing errors. At first we assume that the error rates are given and that true tag j generates tag i with probability α_{ij}. This results in a table of independent counts n_{ij}, the rows of this table summing up to the observed counts.

In order to form a probability model, we assume the true tag counts follow Poisson distributions, i.e. given we have the true proportion p_j of tag j in the library, the true count is m_j with probability $[e^{-p_j\lambda}(p_j\lambda)^{m_j}]/m_j!$ for a fixed λ. The individual counts of n_{ij} then result from multinomial thinning, and also follow Poisson distributions, i.e. Poisson$(\alpha_{ij} p_j \lambda)$, which can be shown to all be independent. Further, this results in the observed counts $n_0, \ldots, n_N$ to also follow

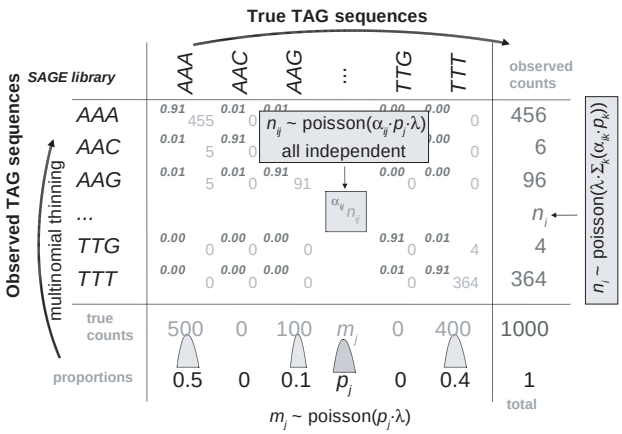

Fig. 1. True SAGE tags generate sequences not equal to the true tags due to sequencing errors. The generated counts are visualized in a contingency table. The marginals represent the true and observed tag counts.

independent Poisson distributions with probability of n_i being Poisson$\left(\sum_{k=1,\ldots,N} \alpha_{ik} p_k \lambda\right)$.

We devise an Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977) to estimate the true counts given the observed counts and estimates for the sequencing error rates. The parameters θ we want to estimate are p_j and λ. The observed data are $n_0, \ldots, n_N$ and estimates for error rates α_{ij}. The complete data are n_{ij} with $i = 1, \ldots, N$, $j = 1, \ldots, N$. The probability of the complete data is thus

$$pr(n_{00}, \ldots, n_{NN}) = \prod_{j=1,\ldots,N} \left(\frac{e^{-p_j\lambda}(p_j\lambda)^{m_j}}{m_j!} \cdot \frac{m_j!}{\prod_{i=1,\ldots,N} n_{ij}!} \cdot \prod_{i=1,\ldots,N} \alpha_{ij}^{n_{ij}} \right). \quad (1)$$

Taking the logarithm, dropping all terms which do not contain parameters we want to estimate, and replacing the complete data by the expected values of the complete data $\hat{m}_0, \ldots, \hat{m}_N$ given the observed data and the estimated parameters, we get

$$Q(\theta, n_{00}, \ldots, n_{NN}) = -\lambda + \sum_{j=1\ldots N} \hat{m}_j \log(p_j\lambda). \quad (2)$$

The EM algorithm cycles between two steps.

E-step: Assuming current parameters and using the observed data, compute expected values for the complete data and calculate the likelihood. The expected values for $m_0, \ldots, m_N$ are given by

$$\hat{m}_j = \sum_{i=1,\ldots,N} \left(\frac{\alpha_{ij} p_j}{\sum_{k=1,\ldots,N} \alpha_{ik} p_k} \cdot n_i \right) \quad (3)$$

M-step: Maximize the likelihood of the complete data given the expected values and recompute new estimates for

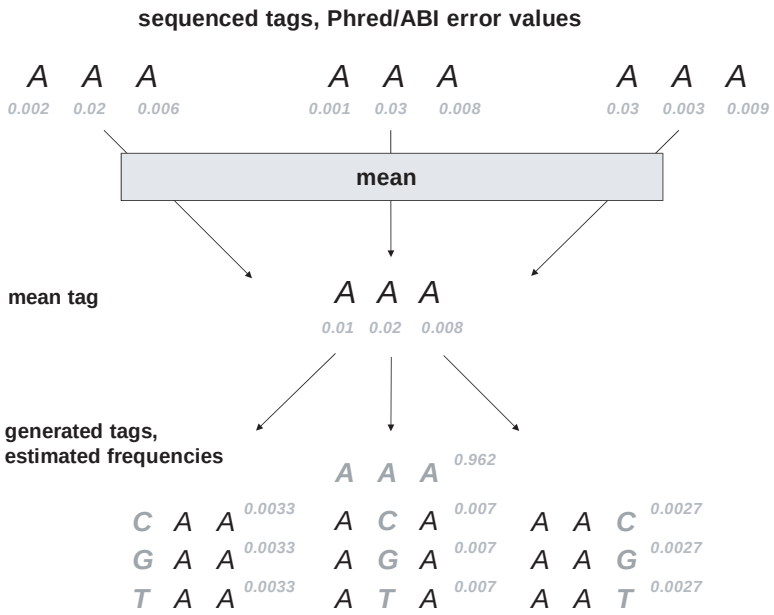

Fig. 2. Estimating the error rates from the observed tags and the Phred/ABI base-calling error estimates. For illustration purposes only one tag with three bases and a count of three is shown.

the parameters.

$$\frac{\partial Q(\theta, n_{00}, \ldots, n_{NN})}{\partial \lambda} = 0 \Rightarrow \hat{\lambda} = \sum_{k=1,\ldots,N} \hat{m_k} = n. \quad (4)$$

$$\frac{\partial Q(\theta, n_{00}, \ldots, n_{NN})}{\partial p_j} = 0 \Rightarrow \hat{p_j} = \frac{\hat{m_j}}{n}. \quad (5)$$

These steps are iterated and the parameter estimates should converge after a few cycles. In our implementation based on simulated libraries the expected values for the true counts $m_0, \ldots, m_N$ are initially set to the observed (or simulated) counts $n_0, \ldots, n_N$ and converge to the true counts in less than 50 cycles. This remains very stable even with large simulated error rates. One problem left open is, where to get expectations for the error rates from.

We use the error measures P_e that are provided by the Phred or ABI base-callers, to estimate the rates that tag j generates tag i. In order to simplify this we only look at the neighborhood of each tag that has one base exchange, assuming that the rates of individual two base exchanges will be very small. For an individual occurrence of a tag sequence s the rate that the sequence corresponds to the true base of the tag at a position b is $1 - P_{e(s,b)}$, we assume all base exchanges are equally likely with probability $P_{e(s,b)}/3$. For each observed tag i we take the means of the estimated error probabilities from the individual occurrences of the tag sequences and assign these to the sequence neighbors j of the observed tag. Figure 2 illustrates this process. We use the so calculated error rates as estimates for α_{ij} in our EM algorithm.

4 IMPLEMENTATION AND EVALUATION

We implemented the procedures described for correcting sequencing errors in SAGE libraries and for simulating SAGE libraries in R using sparse matrices. We run a fixed number of 50 cycles, which led to convergence of the estimates in all the cases we tested. Our implementation has a runtime of a few minutes for a typical SAGE library on a fast PC with reasonable memory. For further analysis of a library we use our estimates $\hat{m}_0, \ldots, \hat{m}_N$ as adjusted counts. We have constructed an R package containing functions for extraction of tags from the sequences, error correction and comparison of libraries (see Systems and methods section). To make these methods easier to access, we have also constructed a web interface.

4.1 Results on simulated data

In order to test our methods we simulated library data as described in Systems and Methods. True counts for tags and randomly distributed sequencing errors were simulated and compared to the resulting observed and adjusted counts for each SAGE library. A variety of parameters were used for simulation in order to test our methods under different conditions. Knowing the true, as well as the observed counts, allows us to study the systematic effects of sequencing errors in a library and for library comparison and to test whether our method is able to correct such effects.

4.1.1 One library Figure 3A shows the effect of sequencing errors and our adjustment to the counts on an example simulated library. It is apparent that the observed counts are

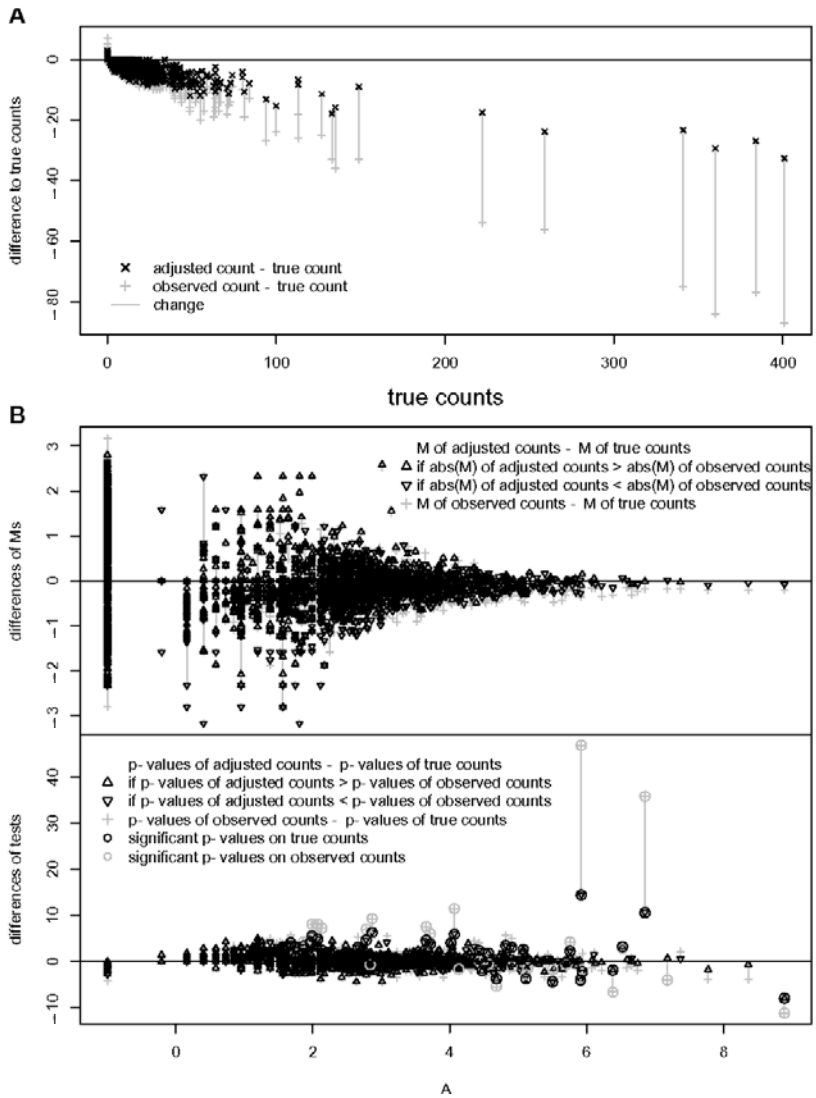

Fig. 3. Results on simulated data. Part **A** shows the changes to the tag counts on a simulated library of 50 000 17 base tags with a mean sequencing error rate of 1% per base. The plots in (**B**) show the comparison of two simulated libraries, the one from (**A**) is compared to a library of 30 000 tags with a mean error rate of 0.5%. The M value represents the $\log_2$ fold change of a tag between the two libraries. A statistical test is performed to assign a P-value to these differences. Shown are the differences of the observed to the true M- and P-values and the changes after adjustment.

underestimated, as they lose numbers to sequencing errors linearly to their abundance. Tags with a low count can be either higher or lower than their true count, especially tags with an observed count of 1 often have a true count of 0. The adjusted counts resulting from our EM algorithm are closer to the true counts in most cases, although often not yet identical. This observation is typical for short-tag (10-base) libraries, as well as for long-tag (17-base) libraries. The mean absolute difference of true and observed counts increases with percentage of sequencing errors, length of the tag sequences and library size. This applies also to the number of introduced artificial tags, which have a true count of 0. A library of 50 000 10 base tags with a mean sequencing error of 1% has typically 20 000–25 000 unique tags and 6000–7000 tags with a true count of 0. More than 99% of these false tags have an observed count of 1 or 2. Due to our error adjustment 30–40% of these false tags get a reduced count and ∼20% end up with a count of less than 0.5. Libraries of 50 000 17 base tags and 1% error rate typically have 25 000–30 000 unique tags of which 10 000 are false tags.

4.1.2 Two libraries In Figure 3B it is noticeable that with libraries of different size or with different mean sequencing error rates there is a systematic bias in the M-values.

Table 1. Simulated data: examples from simulated libraries (50 000 17-base tags with 1% error rate versus 30 000 tags with 0.5% error)

Tag	True			Observed			Adjusted		
	Count1	Count2	fdr	Count1	Count2	fdr	Count1	Count2	fdr
tcaatatatgatcggtt	32	1	0.01	24	1	0.17	27	1	0.05
tcaacatatgatcggtt	0	0	1	1	0	1	0.8	0	1
tcaatatatgatccgtt	0	0	1	1	0	1	0.8	0	1
gcccacggattgtctct	149	132	0.5	116	117	0.09	140	122.7	0.85
gcgcacggattgtctct	0	0	1	2	0	1	0	0	1
gcccacgcattgtctct	0	0	1	1	2	1	0.04	1.7	1

The columns 'fdr' display the false discovery rate, which is the result of the significance test for differential expression adjusted for multiple testing.

Especially abundant tags appear to be relatively higher in the library with the lower rate of sequencing errors. These small differences in the M-values can lead to quite significant P-values if these tags have a high abundance. After correction most of the M-values move closer to the M-values calculated with the true counts and the P-values of the tests for differential expressions move closer to the P-values calculated with the true counts. Simulated libraries having the same total number of tags and the same mean sequencing error rate show only little change to the M-values after correction. However, P-values of truly differentially expressed tags frequently become more significant and closer to the P-values calculated from the true counts as these are computed at a higher abundance level.

Table 1 shows some examples of tags, where sequencing errors result in either falsely differential tags or insignificant P-values for real differential tags.

4.2 Results on experimental libraries

We have tested our method on eight of our experimentally generated SAGE libraries, of which four were generated using the LongSAGE method. These libraries have a total of more than 380 000 tag sequences. The mean sequencing error estimate for bases present in tag sequences ranges between 0.5 and 1.5% in different libraries. Tags with a count of less or equal than 3 on average get a decreased adjusted count in each of the libraries. Tags with a count of greater than 3, on the other hand, on average get an increased adjusted count. Each library contains between 30 000–70 000 tag sequences and 13 000–35 000 unique tags. Between 300 and 2300 of the unique tags get an adjusted count of less than 0.5, these have average estimated base-calling error rates of 2.5–7.2% per library. This shows that counts of tags with low abundance and high sequencing error estimates are generally reduced, while counts of tags with high abundance are more often increased.

Figure 4 shows an example of a comparison of two experimental libraries. In Figure 4A, the changes to the counts in comparison to the abundance of the tags are shown for a mouse hypothalamus library. It can be seen that the counts are increased more relative to the abundance of a tag. Counts of low-abundant tags can be decreased. This picture is similar for the embryonic mouse cortex library as well as for the other libraries tested. The comparison of the two libraries shows that the P-values of tests on the adjusted counts are often lower than those calculated on the observed counts, marking some new tags as significantly differential. The M-values change predominantly for the low-abundance tags, where they are unstable, and in a region of moderately-abundant tags, where they may represent real differences of the libraries. Some examples of possibly differentially expressed tags along with all their sequence neighbors are shown in Table 2.

5 DISCUSSION AND CONCLUSION

We have developed and implemented a method to correct the bias of sequencing errors in SAGE libraries. We have shown in simulations that sequencing errors result in a bias that might change the counts of all tags with a linear relation to the abundance, as well as introducing low-abundance sequence tags, which do not correspond to the sequence of any genes. This is in line with the previous observation that the abundance of a tag shows a linear relation to the number of direct neighbors, which can be used to estimate the prevalence of sequencing errors in a library (Blades *et al.*, 2004a,b). We have shown that with simulated data our method is able to move the estimated counts closer to the true counts and reduce the number of 'false' tags, which do not correspond to any genes. It is easy to construct cases where the bias due to sequencing errors leads to misleading results, when comparing SAGE libraries in order to find differential genes. For example, in libraries of different size and with differing mean sequencing error rates, highly abundant tags appear to be differential with significant P-values even though they are not. Differentially expressed tags appear to be less significant without error correction, because the tag counts are decreased in both libraries. The scale of the P-values is a concern as it is difficult to find differential tags at a low level of expression, especially when applying a rigorous multiple testing correction. On the other hand even tiny differences on highly expressed tags may lead to very significant P-values. We could show in examples, that our method reduces the effects

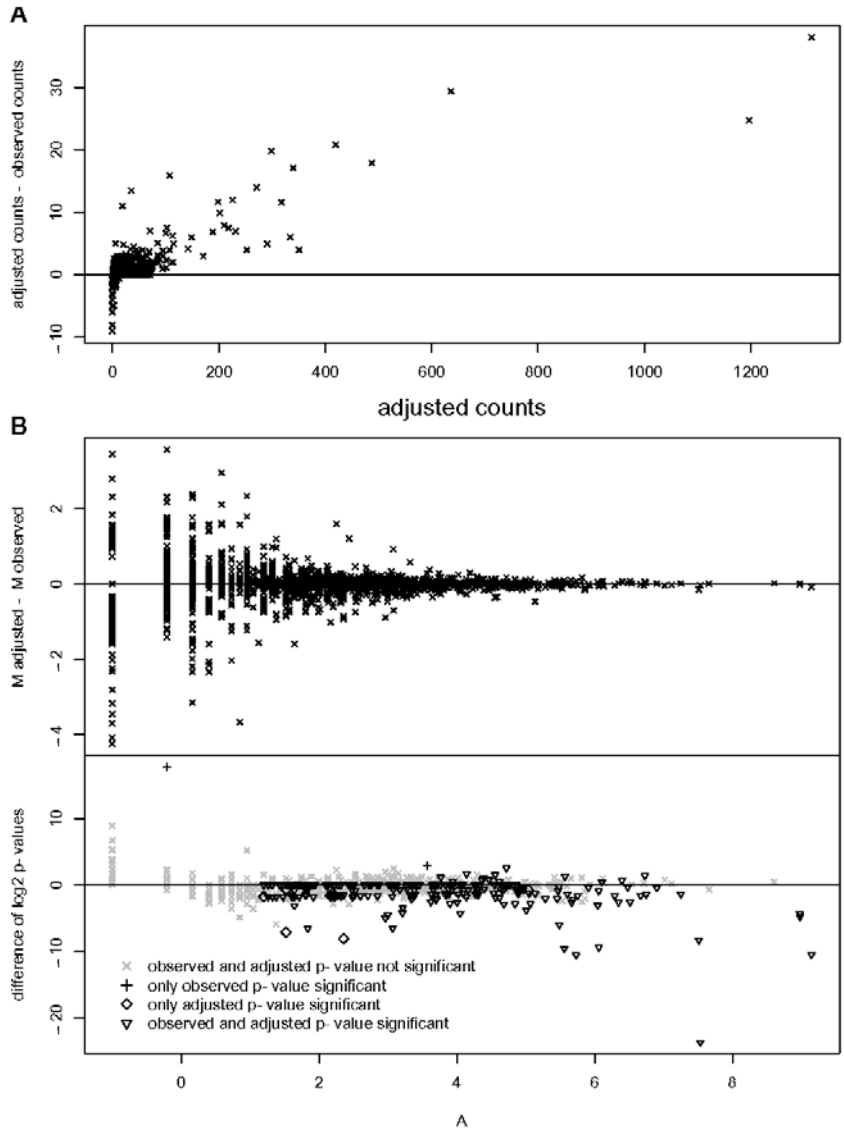

Fig. 4. Results on experimental data. (**A**) Shows the changes to the tag counts on a SAGE library from adult mouse hypothalamus having 42 775 sequenced 10-base tags with a mean sequencing error rate of 1.1% per base. The plots in (**B**) show the comparison of this library to a library from embryonic mouse cortex consisting of 27 349 sequenced tags with an average error rate of 1.4%. The M-value represents the $\log_2$ fold change of a tag between the two libraries. Statistical tests are performed to assign P-values to these differences. A P-value of 0.1 after adjustment for multiple testing is considered as significant. Shown are the differences of the M- and P-values calculated from the observed and adjusted counts.

of sequencing errors and renders some truly differential tags as significant while correcting for biases introducing falsely differential tags.

Experimental data displays a less predictable distribution of sequencing errors, a single library may consist of sequences ranging from very high to very poor quality. Different libraries also display significant differences in sequencing quality, as they may be run on different sequencing machines, at different times and by different laboratories. It is also common to merge and compare libraries coming from very diverse sources. It is therefore of great value to use the error scores provided by the base-calling software (Ewing and Green, 1998). We showed that using our method on experimental library data changes the counts of tags in the way one would expect. The counts of moderately to highly abundant tags are generally increased and low-abundance tags with low quality scores are removed or reduced in count. In a library comparison this is seen to have an influence on the P-values, and examples could be shown where tags that would otherwise have been considered as noise receive significant P-values.

Table 2. Experimental data: examples from a comparison of mouse adult hypothalamus and embryonic cortex libraries (42 775 and 27 349 tags, respectively)

Tag	Mean error		Observed				Adjusted			
	Hypo	Cortex	Count1	Count2	*M*	fdr	Count1	Count2	*M*	fdr
caggactccg	1.9%	0.8%	32	5	1.92	0.134	33.0	5.0	1.96	0.090
ccggactccg	11%		1	0	0.94	1	0.0	0.0	−0.65	1
gccaagggtc	0.8%	0.9%	6	17	−2.07	0.117	6.0	18.9	−2.44	0.04
accaagggtc		3.0%	0	1	−2.23	1	0.0	0.1	−0.67	1
tccaagggtc		0.6%	0	1	−2.23	1	0.0	1.0	−2.25	1
ggcaagggtc		12%	0	1	−2.23	1	0.0	0.0	−0.65	1
gcctggggtc		0.004%	0	1	−2.23	1	0.0	1.0	−2.23	1

The columns 'fdr' display the false discovery rate, which is the result of the significance test for differential expression adjusted for multiple testing. The columns '*M*' indicate the fold changes of a tag. The mean error of a tag is calculated from the error rates estimated in base-calling for all instances of the tag in the library.

We observe, however, that our method still underestimates the changes occurring due to sequencing errors: it reduces the effect, but does not quite reach the correct values. Our method only takes into account sequence neighbors of a tag that have one base exchange. Further, we are still working on improving our estimates of the error rates α_{ij}. Here we assumed a direct relation of Phred error scores P_e with the error rates α_{ij}, which might be inaccurate. We have done some tests that indicate that the estimation can be improved by weighting P_e with the relative proportion of a sequence neighbor among the other possible sequence neighbors, instead of taking $P_e/3$ for each. Our EM approach already works well, however. It also works well, when used with constant estimates for the error rates α_{ij} set for each sequence neighbor as used, for example, in Colinge and Feger (2001) and Akmaev and Wang (2004).

We are also working on extending our approach to take into account single-base insertion and deletion errors. To be able to do this we still need to define the probability of the occurrence of a one-base insertion or deletion based on the Phred or ABI error scores. In our implementation for extracting the tags and quality values from the sequence file we extract an extra base, following directly to the tag. As the ditags are frequently longer than two times the tag length, this might still be the true base in many cases and can assist in dealing with deletions as well as mapping tags to the correct genes.

Careful calling of tags in a SAGE library and management of the information is a prerequisite for any further analysis of the data. We believe that our methods and implementation can assist in this process. We found that, in practice, sequencing errors can introduce a bias rendering high-abundance tags as falsely differential when comparing libraries with different amounts of sequencing errors and different size. This may be of particular importance in detecting significantly differentially expressed genes of low abundance when comparing a single SAGE library of moderate size (e.g. 40 000 tags) to a pool of many SAGE libraries with a large number of tags (e.g. 300 000) or when comparing SAGE libraries from different sources with a wide range of sequencing error rates. Moreover, sequencing errors result in decreased significance of truly differential tags as in general 'true' counts are underestimated. On the other hand, the number of different transcripts present in a library is overestimated as false tags are introduced at low abundance. Our correction method moves true tags to higher abundance levels and towards significance when comparing libraries, and it reduces the number of low-abundance tags, which do not have any matches to genes.

ACKNOWLEDGEMENTS

We thank Steven Evans for helpful suggestions. We thank Nick Tan and the WEHI IT department for IT support and members of WEHI Bioinformatics for discussions. Most DNA sequencing of SAGE libraries was performed by the Australian Genome Research Facility (AGRF) which was established through the Commonwealth-funded Major National Research Facilities program. This work was supported by a fellowship from the Deutsche Forschungsgemeinschaft and the WEHI NHMRC Transitional Institute Grant 215499 to T.B.; NHMRC fellowship 171601 to H.S.S.; NHMRC grants (219176, 257501, 215201, 257529) to H.S.S., G.K.S., T.P.S. and S.-S.T.; the WEHI Nossal Leadership Award to H.S.S.

REFERENCES

Akmaev,V. and Wang,C. (2004) Correction of sequence-based artifacts in serial analysis of gene expression. *Bioinformatics* (in press).

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

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS-B*, **57**, 289–300.

Blades,N., Jones,J., Kern,S. and Parmigiani,G. (2004a) Denoising of data from Serial Analysis of Gene Expression. *Bioinformatics* (in press).

Blades,N., Velculescu,B. and Parmigiani,G. (2004b) Estimation of sequencing error rates in SAGE libraries. *Genome Biol.* (in press).

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

Dempster,A., Laird,N. and Rubin,D. (1977) Maximum likelihood from incomplete data using the EM algorithm. *J. R. Stat. Soc.*, **39**, 1–38.

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

Gentleman,R. and Carey,V. (2002) Bioconductor. *R News*, **2**, 11–16.

Lash,A., Tolstoshev,C., Wagner,L., Schuler,G., Strausberg,R., Riggins,G. and Altschul,S. (2000) SAGEmap: a public gene expression resource. *Genome Res.*, **10**, 1051–1060.

Lockhart,D., Dong,H., Byrne,M., Follettie,M., Gallo,M., Chee,M., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

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

Margulies,E. and Innis,J. (2000) eSAGE: managing and analysing data generated with serial analysis of gene expression (SAGE). *Bioinformatics*, **16**, 650–651.

Schena,M., Shalon,D., Davis,R. and Brown,P. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Stollberg,J., Urschitz,J., Urban,Z. and Boyd,C. (2000) A quantitative evaluation of SAGE. *Genome Res.*, **10**, 1241–1248.

Velculescu,V., Madden,S., Zhang,L., Lash,A., Yu,J., Rago,C., Lal,A., Wang,C., Beaudry,G., Ciriello,K. *et al*. (1999) Analysis of human transcriptomes. *Nat. Genet.*, **23**, 387–388.

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