Correction of sequence-based artifacts in serial analysis of gene expression

Viatcheslav R. Akmaev* and Clarence J. Wang

Genzyme Corporation, Framingham, MA 01701-9322, USA

Received on May 19, 2003; revised on October 27, 2003; accepted on October 18, 2003
Advance Access publication February 10, 2004

ABSTRACT

**Motivation:** Serial Analysis of Gene Expression (SAGE) is a powerful technology for measuring global gene expression, through rapid generation of large numbers of transcript tags. Beyond their intrinsic value in differential gene expression analysis, SAGE tag collections afford abundant information on the size and shape of the sample transcriptome and can accelerate novel gene discovery. These latter SAGE applications are facilitated by the enhanced method of Long SAGE. A characteristic of sequencing-based methods, such as SAGE and Long SAGE is the unavoidable occurrence of artifact sequences resulting from sequencing errors. By virtue of their low-random incidence, such tag errors have minimal impact on differential expression analysis. However, to fully exploit the value of large SAGE tag datasets, it is desirable to account for and correct tag artifacts.

**Results:** We present estimates for occurrences of tag errors, and an efficient error correction algorithm. Error rate estimates are based on a stochastic model that includes the Polymerase chain reaction and sequencing error contributions. The correction algorithm, SAGEScreen, is a multi-step procedure that addresses ditag processing. Estimation of empirical error rates from highly abundant tags, grouping of similar-sequence tags and statistical testing of observed counts. We apply SAGEScreen to Long SAGE libraries and compare error rates for several processing scenarios. Results with simulated tag collections indicate that SAGEScreen corrects 78% of recoverable tag errors and reduces the occurrences of singleton tags.

**Availability:** The SAGEScreen software is available for academic users from the first author.

**Contact:** slava.akmaev@genzyme.com

1 INTRODUCTION

The wide availability of human genome data has accelerated the use of computational tools to explore essential biological processes. A broad variety of methods from diverse research communities have produced a massive amount of data complementary to the core assembled genomic sequence. From this synthesis of sophisticated algorithms and expansive data sources, more accurate annotations and, consequently, better understanding of the genome emerge.

Here, we examine one such source of abundant genome-linked data, Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995). SAGE is a protocol for systematic, high-throughput generation of short expressed sequence tags (ESTs) from a cell sample, producing a global profile of gene expression. Briefly, SAGE generates short mRNA sequence tags from a specific position in transcripts. The tag position is defined by the location of the 3′-most anchoring enzyme restriction site. The most commonly used enzyme for this purpose is NlaIII. cDNA fragments from cleavage with an anchoring enzyme are further processed with the tagging enzyme, a Type II restriction endonuclease, typically BsmFI. Following amplification, cloning and sequencing, the end result of a SAGE experiment is a set of vector inserts from which ditag and, ultimately, tag sequences are extracted and counted. Each SAGE tag is prefixed by the anchoring enzyme restriction site and corresponds to the 10–11 bp extension of the 3′-most site in the cognate transcript. In theory, tags of this length are sufficiently specific to map the transcriptome. In fact, most human SAGE tags map uniquely to the UniGene clusters (Lash et al., 2000). Given such a bi-directional map, expression levels of the transcripts are inferred from observations of their SAGE tags. Recently, the SAGE protocol was enhanced with a new tagging enzyme. This enzyme, MmeI, cuts 21–22 bases downstream of the anchoring enzyme restriction site (Saha et al., 2002). The new protocol, Long SAGE, enhances the specificity of SAGE to transcriptome mapping and allows direct mapping of Long SAGE tags to the genome.

The predominant use of SAGE is for measurement of differential gene expression. In comparison with microarray methods, SAGE offers several advantages which have been reviewed elsewhere (e.g. Pollock, 2002). However, the SAGE protocol is subject to sequence errors introduced by the polymerase chain reaction (PCR) and sequencing steps. Sub-optimal fidelity in these procedures can introduce artifact tag sequences. Such ‘mutations’ occur infrequently for any individual transcript and have little effect on the quantification of differential expression of moderately expressed...
genes. Their consequence is greater on the measurement of rare transcripts and the identification of novel genes. In addition, accumulation of such spurious tags introduces noise into the overall profile of transcripts in a sample and obscures the characterization of transcriptome size. For example, classes of estimators developed for multinomial populations often rely heavily on the number of species observed at low levels (Bunge and Fitzpatrick, 1993). Application of these methods to SAGE transcriptome measurements is hindered by the presence of mutant tags. Herein, we present a detailed study of SAGE sequence errors and propose an efficient algorithm for detection and removal of spurious tags in SAGE libraries. We apply this methodology to a number of Long SAGE datasets and discuss representative results.

2 ACCUMULATION OF ERRORS IN THE SAGE PROTOCOL

Generation of SAGE tags involves serial execution of several biochemical protocols. Among them are biotinylation, reverse transcription–PCR (RT–PCR), enzymatic digestion, ligation, PCR, cloning and finally sequencing. Potentially, each of these reactions or protocols can introduce distortions in the observed tag populations. The present analysis estimates the potential number of artificial tags observed in a SAGE library due to errors introduced in the PCR and sequencing steps. These two procedures are expected to be the most prominent contributors of spurious tags.

2.1 Effects of the polymerase chain reaction

The PCR is an intrinsic part of the SAGE protocol. After digestion with a tagging enzyme, tag–linker combinations are ligated into ditag molecules. Ligated ditag molecules serve as templates for PCR amplification with primers specific to the linker sequences. A DNA polymerase will make errors during DNA synthesis. The rate of such errors varies among specific DNA polymerases and reaction conditions. The fidelity of PCR is a well-studied subject (e.g. Keohavong and Tilly, 1989). The error rates for several DNA polymerases were experimentally determined to vary from $10^{-6}$ to $10^{-4}$ mutations per base duplication. These numbers appear misleadingly small. When PCR protocols consist of up to 30 cycles of the reaction, mutations occurring at early stages propagate through the later cycles and may result in a substantial percentage of mutant base pairs. In order to estimate the number of mutant sequences, a mathematical model of the reaction is needed.

Theoretically, the PCR can be treated as a two-dimensional Galton–Watson branching process, $Z_n = \{Z_n^{(1)}, Z_n^{(2)}\}$ (Sun, 1995). Here, $n$ is the number of cycles, $Z_n^{(1)}$ is the number of correct sequences present in the DNA pool at cycle $n$ and $Z_n^{(2)}$ is the number of erroneous molecules in the pool. The initial state of the process is vector $(1, 0)$. If the amplification efficiency per cycle is fixed, then a general limit theorem implies (Athreya and Ney, 1972)

$$
\frac{Z_n}{(1 + f)^n} \rightarrow (0, w) \text{ as } n \rightarrow +\infty.
$$

Here, $f$ is the amplification efficiency per cycle, and $w$ is a random variable. This result shows that even for tiny mutation rates, the percentage of mutant sequences eventually approaches 100% with more duplications. In reality, the number of cycles routinely done in such experiments is reasonably small, and we are more interested in the statistics of this process for particular values of $n$ rather than the limit scenario. SAGE ditags are amplified with approximately 25–30 cycles of PCR. To get an idea of how many mutant ditags should be expected on average, we determined their expectation and variance. The expectation can be found in a closed form by solving a recurrence equation for the probability generating function of the Galton–Watson process $Z_n$. The resultant expression is quite simple:

$$
EZ_n^{(1)} = [(1 - p) \cdot (1 + f) + p]^n,
$$

where $p$ is the probability of inaccurate duplication of a single ditag. The variance can be found in the same fashion (Piau, 2002). For error rate estimates, it is more meaningful to consider the following expression:

$$
\frac{EZ_n^{(2)}}{(1 + f)^n} = 1 - \left(1 - p + \frac{p}{1+f}\right)^n. \quad (1)
$$

This quantity is the expected number of mutant molecules divided by the total expected number of molecules after $n$ cycles. If the initial number of molecules is large, the expression on the left is a good approximation of the expected error rate (Hayashi, 1990). Simulations for various values of $p$, $f$ and the initial number of molecules displayed good agreement with the theoretical approximation in Equation (1) (data not shown). The results changed little in simulations assuming variable amplification efficiency, $f$. The c.v. was estimated at 100% for simulations with a single seeded molecule, and it clearly decreases to zero with an increasing number of starting molecules. This indicates that the overall average error rate for ditags sequenced in a SAGE library is accurately approximated with the right-hand side expression in Equation (1). However, since it is quite possible for a specific ditag molecule to be present only once in the sample, the error rate may fluctuate considerably from one unique molecule to another. The overall average error rate will be assumed throughout this section.

We will apply the formulation above and look at the potential PCR error rate in SAGE libraries. The error rate is a composite of two parameters: the probability of a mutation and the amplification efficiency. The probability of a mutation, $p$, depends on the per base pair, per cycle error rate or, in other words, the fidelity of the amplification. As mentioned previously, several groups have reported base pair per
cycle error rates for DNA polymerases, inferred from the number of erroneous sequences observed after 20–30 cycles of PCR (Keohavong and Tilly, 1989). Such inferences were often based on simplistic mathematical models of the reaction and may be inaccurate. Tindall and Kunkel (1988) approached this question differently. They measured the fidelity of the DNA polymerase isolated from *Thermus aquaticus* (*Taq*) with two assays that scored errors produced during *in vitro* DNA synthesis of the *lacZα* complementation gene in M13mp2 DNA. These experiments allowed the authors to explicitly obtain the mutation rate per amplification cycle. The authors also compared the performance of the *Taq* polymerase with that of the large (Klenow) fragment of *Escherichia coli* DNA polymerase I. The *Taq* polymerase is industrially used in PCR, and is a standard in the SAGE protocol. Tindall and Kunkell’s estimates of the mutant frequency, i.e. the rate of faulty complementation and is a standard in the SAGE protocol. Tindall and Kunkell’s estimates of the mutant frequency, i.e. the rate of faulty complementation, varied with the temperature of the reaction from 81·10⁻⁴ to 130·10⁻⁴ with the background rate of 6.7·10⁻⁴ (Table III in Tindall and Kunkell). At the recommended temperature of about 70°C (e.g. the *Taq* DNA polymerase kit from Life Technologies), the average measured mutant frequency was estimated at 125·10⁻⁴. The error rate per base pair can be obtained using the formula (Tindall and Kunkel, 1988):

\[ r = \frac{mf(MF_0 - MF_0)}{f_0N_d}, \]  

where \( mf \) is the mutant fraction as determined by DNA sequence analysis, \( MF_0 \) is the observed mutation frequency, \( MF_0 \) is the background mutation frequency, \( f_0 \) is the frequency of expression of newly synthesized strand and \( N_d \) is the number of nucleotides known to yield a mutant. In their study, \( mf = 32/42 \) for substitutions and \( mf = 10/42 \) from frameshift mutations, \( f_0 \) has been measured to be 0.6 and \( N_d \) was 110 nt for substitutions and 160 for frameshifts. Applying formula 2 separately for substitutions and frameshift mutations, we get the following estimate:

\[ r \approx \frac{32 \cdot (125 - 6.7)}{42 \cdot 0.6 \cdot 110} \cdot 10^{-4} + \frac{10 \cdot (125 - 6.7)}{42 \cdot 0.6 \cdot 160} \cdot 10^{-4} = 1.66 \cdot 10^{-4}. \]  

The majority of observed Long SAGE ditags sequences without the flanking restriction sites vary between 32 and 34 nt, with more than 50% of them consisting of 33 nt. For simplicity, we will assume the ditag length of 33 nt. The probability of an erroneous duplication is then estimated as

\[ p = 1 - (1 - r)^{33} \approx 0.0055. \]  

The rate of mutant ditags in Equation (1) also depends on the number of PCR cycles, \( n \), and the amplification efficiency, \( f \). Typically, PCR amplification is optimized for exponential dynamics, and the reaction is stopped when the accumulation of DNA molecules plateaus. The number of cycles necessary for reaching such a limit is generally dependent on, but not limited to, the concentration and length of the amplified molecules. For the purpose of this demonstration, we will fix the number of cycles at \( n = 27 \). The efficiency of *Taq* polymerase has been estimated at \( f = 88\% \) (Keohavong and Tilly, 1989). With the moderate number of cycles, the amplification efficiency will be assumed constant. Given these estimates, the expected proportion of erroneous ditags in a Long SAGE library after a PCR amplification is approximated with the following expression [Equation (1)],

\[ \rho_{\text{ditags}} = 1 - \left( 1 - 0.0055 + \frac{0.0055}{1 + 0.88} \right)^{27} \approx 0.067. \]  

This shows that about 6.7% of Long SAGE ditags will have acquired mutations prior to ligation, cloning and sequencing. This value implies that about 3.5% of Long SAGE tags will have errors inherited from the PCR amplification (Supplementary data). Based on our c.v. estimate, the error rate may fluctuate considerably for low abundant tags. The error rate distribution is skewed towards zero and has a heavy, oscillating right tail, so that the majority of tags will have few mutant variants, while some of the tags will have a large number of variants. Detection of such abundant mutant tags is a challenging problem and, apparently, can be completely solved only through exhaustive search in the genome.

### 2.2 Sequencing effects

The percentage of polymerase errors observed in SAGE tags is substantial. Moreover, there are other sequence distortions downstream in the protocol. Amplified ditags are linked to form vector inserts. The length of the insert varies from 200 to 700 bp, or from 6 to 12 ditags flanked by anchoring enzyme restriction sites. The vectors are then cloned and sequenced. DNA sequencing in general is performed using the enzymatic dideoxy chain-termination method (Prober et al., 1987). When a clone of interest is purified, a reaction in which a DNA polymerase synthesizes a collection of labeled single-stranded fragments is performed. The fragments are then separated based on their length by gel electrophoresis. Through computational analysis, the exact sequence of the template DNA is deciphered from the relative lengths of the fragments and the identities of their final bases. The analysis consists of several steps: gel lane tracking, gel lane profiling, trace processing and base-calling. The latter two are mathematically and computationally involved and were thoroughly analyzed by several research groups (e.g. Giddings et al., 1993). One of these approaches, *phred*, combines canonical data processing and optimization algorithms with sophisticated *ad hoc* procedures into an automated base-calling program (Ewing et al., 1998). *phred* can be tuned to achieve lower error rates than the standard ABI software, *Sequencing Analysis*. In addition, *phred* calculates quality scores for...
called bases. The quality score is proportional to the logarithm of the error probability, where the distributional parameters were estimated by the authors from a large training dataset (Ewing and Green, 1998).

The \textit{phred} error rate predictions were shown to be quite accurate, and we will use them here as a standard. For demonstration purposes, we will set the predicted sequencing error rate per base at \( r = 0.01; \) this is equivalent to the \textit{phred} quality score of 20. With this cut-off, about 70–80\% of called bases are retained. Stricter cut-offs lead to a decreased percentage of retained bases and, consequently, an increased cost of sequencing. Assuming again a ditag length of 33 nt and a binomial model of error accumulation, the expected rate of incorrectly sequenced ditags is found as

\[
R_{\text{ditags}}^{\text{Sequencing}} = 1 - (1 - 0.01)^{33} \approx 0.282. \tag{5}
\]

This shows 28.2\% of ditags in a Long SAGE library will have at least one erroneous base pair introduced by sequencing. In terms of Long SAGE tags, \( \sim15.6\% \) of them will have errors resulted from sequencing (Supplementary information). These are overall expected rates; observed rates would be close to these estimates for abundant tags and may vary greatly for rare tags.

The two sources of base pair distortions described above are dominant in the SAGE protocol. Other possible causes of errors may include the initial round of reverse transcription and cloning, and the general background DNA sequence distortions. These causes probably contribute to the error rates on the level of a few orders of magnitude lower. Therefore, assuming only the PCR and sequencing error contributions and combining Equations (4) and (5), we get the overall rate of mutant ditags,

\[
R_{\text{ditags}} = R_{\text{ditags}}^{\text{PCR}} + R_{\text{ditags}}^{\text{Sequencing}} - R_{\text{ditags}}^{\text{PCR}} \cdot R_{\text{ditags}}^{\text{Sequencing}} \approx 0.33. \tag{6}
\]

Approximately, one-third of the processed ditags in a Long SAGE library will have at least 1 bp mutation. This translates to a 17.3\% error rate for Long SAGE tags. This number is substantial: one-sixth of the Long SAGE tags in a library are represented by erroneous sequences. About 80\% of the spurious tags are one base variants of a real tag, and more than 95\% of the mutants are one or two base variants. These tags form clusters related to individual abundant tags (we will designate real tags or ditags with mutant variants as parent tags and ditags). Such structures are fertile soil for computational analysis. In the next section, we will describe an algorithm that locates mutant ditags and tags and corrects tag counts in a SAGE library.

3 \textbf{CORRECTION OF SEQUENCE-BASED ARTIFACTS}

It is evident from our theoretical exploration the volume of erroneous SAGE tags is substantial. The majority of mutant tags are observed at the low end of the expression range and will likely be of low significance for a differential expression analysis using appropriate statistical methodology (e.g. Lash et al., 2000). However, when SAGE libraries are mapped to the genome or the transcriptome, an error correction procedure is mandatory. Tag variants are rare and diverse, and their majority have observations in the single digit numbers. The error rate projections indicate the mutant tags will constitute approximately one-sixth of the total tag population. On the scale of unique tags, this percentage may very well reach 30–40\%. Such amounts of spurious tags will considerably hinder tag mapping efforts and a subsequent transcriptome analysis.

Colinge and Feger (2001) suggested an interesting correction approach for SAGE libraries. This approach embodies tags from a SAGE library into a network of related clusters, where the relationship is defined as the edit distance between the tag sequences. Tags with edit distance of one, or single base variants, are linked in the network. The authors then associate the observed and the error-free vectors of tag counts with a linear system of equations. The parameters of the system are based on an a priori defined sequencing error rate and expected count contributions between the tags. The error-free tag counts are found by solving this linear system. Major disadvantages of this scheme are the simplification of the error probability distribution to a deterministic factor expressed by its expectation and the forced continuity of the tag counts. As a result, the estimates of the error-free tag observations are in many cases negative and certainly non-integers. Such numbers are hard to interpret in the context of SAGE multinomial sampling.

We suggest a multi-step approach that exploits the intrinsic structure of SAGE data, utilizes empirical error rates and preserves the highly discrete and stochastic organization of the SAGE tag sampling. The algorithm, \textit{SAGEScreen}, starts with analysis of ditags. Observed duplicate ditags are generally discarded in the SAGE protocol to control for PCR amplification bias. \textit{SAGEScreen}, additionally, finds pairs of similar ditags and removes erroneous ditags if the evidence is sufficient. It continues until all pairs are processed. Then, \textit{SAGEScreen} estimates empirical error rates from expression patterns of abundant tags. Based on these rates, the algorithm finds clusters of tag variants related to one parent tag. Subsequently, the variants are deleted from the dataset, and their counts are added to the parent tag. At the end, only variant tags exhibiting sufficient evidence of veracity are retained in the set. A formal representation of the algorithm is shown in Figure 3, and a detailed description follows.

(1) \textit{Ditag processing}. \textit{SAGEScreen} takes as input a set of unique SAGE ditags analyzed with program \textit{phred}. Each ditag
Ditags pairs: a. 

<table>
<thead>
<tr>
<th>Ditag 1 - kept</th>
<th>Ditag 2 - discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>█████████████</td>
<td>□□□□□□□□□□□□□□□□□</td>
</tr>
<tr>
<td>□□□□□□□□□□□□□</td>
<td>□□□□□□□□□□□□□□□□□</td>
</tr>
</tbody>
</table>

Ditags b. 

<table>
<thead>
<tr>
<th>Ditag 1 - kept</th>
<th>Ditag 2 - discarded</th>
</tr>
</thead>
<tbody>
<tr>
<td>□□□□□□□□□□□□□</td>
<td>□□□□□□□□□□□□□□□□□</td>
</tr>
<tr>
<td>□□□□□□□□□□□□□</td>
<td>□□□□□□□□□□□□□□□□□</td>
</tr>
</tbody>
</table>

Fig. 1. Illustration of the Ditag processing selection criterion. Parent tags are emphasized in boldface. (a) Shows situations with an obvious parent ditag in the pair. (b) An example where parent tags are observed in both ditags.

has an indicator if one or both of its tags passed the phred score cut-off. SAGEScreen builds a first pass SAGE library; the count of zero is assigned to the tags that do not satisfy the phred score criterion. The ditags are loaded and analyzed in the order they appear in the data file. The algorithm looks at all pair-wise (direct and reverse compliment) comparisons of the ditags. For a particular pair, SAGEScreen retrieves the sequences of the four tags, T_{1}^{l}, T_{1}^{r}, T_{2}^{l}, and T_{2}^{r}, where l and r denote left and right tags, respectively. If T_{1}^{l} = T_{2}^{l} and T_{1}^{r} = T_{2}^{r}, SAGEScreen removes the shorter ditag from the dataset and proceeds to the next pair. Such situations occur because of the variable length of the ditags and could be protocol attributable artifacts. When the tags are not the same, the algorithm checks if the ditags are one or two base variants of each other. If so, SAGEScreen calculates indices of the maximal and minimal tag counts:

\[ i_{1}^{\max} = \arg \max \ C_{1}^{i}, \quad i_{1}^{\max} = \arg \max \ C_{1}^{r} \]

\[ i_{1}^{\min} = \arg \min \ C_{1}^{i}, \quad i_{1}^{\min} = \arg \min \ C_{1}^{r} \]

where \( C_{1}^{i} \) is the observed abundance of tag \( T_{1}^{i}, \ i = 1, 2, j = r, l \). If \( i_{1}^{\max} = i_{1}^{\min} \) and either of the maxima is strictly bigger than the corresponding minimum, ditag \( i_{1}^{\min} \) is removed from the dataset (Fig. 1a). This is the case where the cumulative expression of one ditag is bigger, and, consequently, the tags of the other ditag are assumed spurious. If \( i_{1}^{\max} \neq i_{1}^{\max} \) and both maxima are strictly bigger than the corresponding minima, a new ditag is created such that its left tag is \( T_{1}^{i} \) and its right tag is \( T_{2}^{r} \) (Fig. 1b). The new ditag is added to the dataset, while both original ditags are deleted. In this situation, the errors possibly occurred on different sides of both ditags. In all other cases, SAGEScreen retains both ditags and moves on to the next pair. Once all the pairs are processed, SAGEScreen retrieves tags from the processed set of ditags and builds a second pass SAGE library.

In general, duplicate ditags are removed from SAGE libraries. One might argue that for highly abundant tags the probability of sampling more than one ditag of the same kind is large enough for the event to occur. In such cases, we sample duplicate ditags due to their high frequency, and their removal from the dataset skews the observed counts of abundant SAGE tags. This deficiency of the SAGE data preparation routine can be overcome with simple and accurate adjustment techniques that are beyond the scope of this work (Snyder and St.-Amand, 2000). We emphasize that mandatory removal of the duplicate ditags in SAGE data analysis does bias the observed counts of abundant tags, but this bias is well understood and may easily be compensated for computationally. Indeed, not only exact duplicates but also distorted duplicates are to be filtered out in the presence of errors. The ditag processing in SAGEScreen achieves two goals: elimination of the PCR bias by rigorous discarding of duplicate ditags and a more specific detection of mutant tags. The effect of false positives (real ditags removed because of their sequence similarity) is minuscule: when a real ditag is removed, the observations of its tags are reduced by one. This reduction is only noticeable for tags with the expected abundance of one or two. The situation when two such tags with high sequence similarity appear linked to the same highly abundant tag is a rare event. Roughly, we would expect to see fewer than 10 such ditag pairs in a library of 50K tags (Supplementary information). The loss of the corresponding tags is greatly outweighed by the number of erroneous ditags removed.

(II) Error rate estimation. Here, SAGEScreen examines a SAGE library of tags. The algorithm looks at all tags with counts above a certain threshold. The threshold is chosen to allow for reasonable expectations of erroneous tags with the three types of mutations: substitutions, insertions and deletions. In the current implementation, the threshold was set at 50 tag counts. For each such tag, SAGEScreen finds all one or two base variants in the library. The sum of the parent tag count and the counts of all its variants is defined as the hypothetical error-free parent tag count. Next, SAGEScreen sorts all one base variants into three bins distinguished by the mutation type (if the classification of the mutation is ambiguous, the substitution has precedence) and calculates the sum for each bin. The sums are converted to frequencies of the mutation types. When all parent tags are processed, SAGEScreen has three sets of numbers: the substitution, insertion and deletion frequencies. If the set size is sufficient for estimation (i.e. ≥10 data points), the program chooses the median of the set as a frequency estimate; otherwise, SAGEScreen uses preset estimates obtained from a large, high quality Long SAGE library. Finally, the algorithm calculates the mutation rates per nucleotide by solving the equation \( l \cdot r \cdot (1 - r)^{l-1} = f \), where
l is the tag length, r is the per base mutation rate and f is the corresponding frequency.

The error rates are quite important for accurate detection of mutant sequences. SAGE preparations are perturbed by numerous environmental factors. Moreover, data collection is affected by the choice of technology and methodology. As we show in the Results section, the error rate difference is significant between SAGE libraries. This observation clearly points out the necessity of empirical estimates.

(III) Tag processing. Given estimated error rates and a set of tags, SAGEScreen considers each tag with a count large enough for a one base variant to satisfy a P-value criterion as a potential parent tag. Parent tags are processed in the order they appear in the dataset. For each parent tag, the algorithm finds all one or two base variants with observations less than they appear in the dataset. For each parent tag, the algorithm derives the P-values from a Poisson model based on the error rates. The child tags’ P-values are modified with a Bonferroni-like adjustment: the P-values are multiplied by the number of possible one or two base variants. SAGEScreen tests if the cluster’s P-value is greater than a P-value cut-off; if not, child tags with minimal P-values are dropped until the condition is met. Then, the program applies the same criterion for each adjusted individual P-value in the cluster. Child tags with adjusted P-values less than the cut-off are also dropped from the cluster. SAGEScreen stores child–parent relationship information from processed clusters in a separate data array until all parent tags are processed. At that moment, the algorithm cycles through all the tags in the library and retrieves the child–parent relationships. If a tag is a child tag, its count is added to the parent tag’s count, and the child tag is deleted. This cycle completes the computations.

Figure 2 illustrates P-value thresholding for three hypothetical tag clusters. The P-values shown were calculated based on the default set of error rates with the overall error percentage of about 8.3%. The cluster P-values are equal for the three situations and satisfy the cut-off criterion. In the first case, the individual adjusted P-values fit the criterion. Consequently, the child tags are removed, with their counts added to the parent tag’s count. In the second case, the tag with an insertion has the adjusted P-value lower than the threshold of 0.1, and it is retained in the dataset, while the other two tags are deleted. No change is observed in the last case for the same reason: the only child tag does not satisfy the cut-off condition.

The P-values calculated in the algorithm are the probabilities of observing mutant tags at the particular or larger abundances. We approximated binomial probabilities with a Poisson model, where the parameter was estimated as the product of the overall number of tags in the cluster and the frequency of the particular mutation. In this setting, the P-values control the type II error, namely, the fraction of mutant tags detected as genuine, and it is only semi-indicative of the fraction of genuine tags detected as mutants. These two fractions move in reverse directions; therefore, we set a large type II P-value cut-off, in hope of reducing the number of deleted false positives. If the P-value is small, the tag has either a large count or the mutation is rare. In both cases, the evidence suggests the tag is genuine, and it is not considered further. Control of the type I error here is practically impossible without rather accurate knowledge of the transcriptome and, consequently, the collection of all possible SAGE tags.

SAGEScreen is implemented in C++ and was successfully compiled with Microsoft Visual C++ 6.0. SAGEScreen requires O(n^2) ditag comparisons. Computationally, the algorithm is optimized to allow quick processing of clearly non-similar ditags, while potential ditag pairs are thoroughly checked for variation. The number of such interesting ditag pairs is substantially smaller than the number of all pair-wise comparisons in an average SAGE library. The computing burden of the error estimation sub-block is negligible due to the small number of abundant tags. The final step is quick for the same reason: the amount of parent tags is small compared with the total number of tags. Overall, SAGEScreen takes less than 10 min to scan through a 50K SAGE library on an average PC workstation.

4 RESULTS AND DISCUSSION

4.1 SAGEScreen experimental performance

We have looked at five brain endothelial Long SAGE libraries to evaluate SAGEScreen performance. Two of these libraries were derived from surgically removed non-neoplastic brain tissue, and three were derived from glioma tumor tissue.
We will use the authors’ nomenclature and refer to the libraries as $N_1, N_2$ and $T_1, T_2, T_3$, respectively. Most of the project clones were sequenced by the ABI 3700 machine, although some of the clones were sequenced by ABI 377 and 3100 machines. We have obtained two versions of ditag sets for each library, processed with either phred software or Sequencing Analysis software from ABI. The results of SAGEScreen analysis are outlined in Table 1. The columns correspond to ditag set information, and the rows list various SAGEScreen statistics. The table includes two sets of numbers: numbers in bold relate to the results obtained following all the steps of the algorithm and the numbers not in bold display the results of the algorithm with the Ditag processing step omitted. The table starts with the number of ditags taken for the analysis at step two of SAGEScreen, followed by the number of equivalent ditags—ditags with redundant tag combinations but distinctive by their length, if the Ditag processing is omitted. Lines three and four have the number of unique Long SAGE tags and the total number of Long SAGE tags derived from each ditag set. The ‘Estimated error rate’ line shows the empirical percentage of mutant tags estimated in the SAGEScreen second step. The remaining rows of the table are related to the Tag processing part of the algorithm. In this step, we used a conservative type II $P$-value cut-off of 0.1. With this cut-off, line six has the number of unique tags SAGEScreen had classified as spurious and removed from the
Table 1. SAGEScreen results for five brain endothelial Long SAGE libraries

<table>
<thead>
<tr>
<th>SAGE libraries</th>
<th>N1 Phred</th>
<th>N1 ABI</th>
<th>N2 Phred</th>
<th>N2 ABI</th>
<th>T1 Phred</th>
<th>T1 ABI</th>
<th>T2 Phred</th>
<th>T2 ABI</th>
<th>T3 Phred</th>
<th>T3 ABI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ditags</td>
<td>14 914</td>
<td>16 667</td>
<td>17 782</td>
<td>24 251</td>
<td>16 594</td>
<td>23 103</td>
<td>17 638</td>
<td>22 910</td>
<td>18 226</td>
<td>27 779</td>
</tr>
<tr>
<td>Equivalent ditags</td>
<td>591</td>
<td>531</td>
<td>767</td>
<td>798</td>
<td>487</td>
<td>379</td>
<td>674</td>
<td>707</td>
<td>1031</td>
<td>1090</td>
</tr>
<tr>
<td>Unique tags derived</td>
<td>12 831</td>
<td>13 342</td>
<td>13 477</td>
<td>18 781</td>
<td>14 927</td>
<td>20 918</td>
<td>13 294</td>
<td>18 021</td>
<td>13 636</td>
<td>21 228</td>
</tr>
<tr>
<td>Tags derived</td>
<td>28 057</td>
<td>30 678</td>
<td>32 879</td>
<td>45 395</td>
<td>30 365</td>
<td>43 587</td>
<td>33 393</td>
<td>43 586</td>
<td>33 440</td>
<td>51 841</td>
</tr>
<tr>
<td>Estimated error rate</td>
<td>9.5%</td>
<td>13.4%</td>
<td>8.7%</td>
<td>11.3%</td>
<td>8.4%</td>
<td>9.7%</td>
<td>11.4%</td>
<td>14.0%</td>
<td>7.6%</td>
<td>11.3%</td>
</tr>
<tr>
<td>Unique tags removed</td>
<td>968</td>
<td>1485</td>
<td>1151</td>
<td>2765</td>
<td>962</td>
<td>2130</td>
<td>1513</td>
<td>3057</td>
<td>1009</td>
<td>3120</td>
</tr>
<tr>
<td>Tags redistributed</td>
<td>113 76</td>
<td>1966</td>
<td>2234</td>
<td>1307</td>
<td>3126</td>
<td>1040</td>
<td>2244</td>
<td>1870</td>
<td>3643</td>
<td>1124</td>
</tr>
<tr>
<td>Redistributed (%)</td>
<td>4.0%</td>
<td>6.4%</td>
<td>4.9%</td>
<td>6.9%</td>
<td>3.4%</td>
<td>5.1%</td>
<td>5.6%</td>
<td>8.4%</td>
<td>3.4%</td>
<td>6.9%</td>
</tr>
<tr>
<td>Unique tags left</td>
<td>11 863</td>
<td>11 857</td>
<td>15 359</td>
<td>12 326</td>
<td>16 016</td>
<td>13 965</td>
<td>18 788</td>
<td>11 781</td>
<td>14 964</td>
<td>12 627</td>
</tr>
</tbody>
</table>

*Phred* and ‘ABI’ indicate if the data were obtained with *phred* or the ABI software. The quality score of 20 (1% theoretical per base error rate) was the nucleotide quality criterion for the *phred* processed datasets.

The analysis of ditags greatly enhances the efficiency of the algorithm. On average, spurious ditags contribute 5–10% to the tag observations. The greater part of this contribution affects highly abundant tags, inflating their observations by as much as 30%. The amount of redundant ditags of different lengths contributes 3–5% to the counts; such ditags need to be removed from the dataset unequivocally. If we look at distinct SAGE tags left after the analysis, the numbers for the standard and short procedures are comparable—the discrepancy is <0.5%. This balance depends on the type II $P$-value threshold set in the program. The short algorithm is less effective in detecting mutant sequences if the threshold is more conservative, i.e. when the type II $P$-value cut-off is large. It is more effective with small cut-offs due to generally larger error estimates.

The majority of the estimated error rates are below the predicted value of 17.3%. We expected that the empirical error rates in the *phred* processed datasets would be closer to the predicted value, given the same assumption of the 1% per base sequencing error rate. A plausible explanation for the disparity is the *phred* quality scores were trained on a set of cosmids generated almost entirely on the ABI 373 machine (Ewing and Green, 1998). The sequencing technology has greatly improved since, and, at present, *phred* quality scores probably overestimate the error rate of nucleotide sequencing. On another note, the percentage of mutant tags detected by SAGEScreen is far less than the expected percentage estimated on a subset of highly abundant tags. The sample size for the better part of moderately expressed tags is not sufficient to control type II error at the set level; therefore, spurious tags related to these parent tags are not detected with the required confidence.

Lastly, we would like to point out differences between *phred* and *Sequencing Analysis* processed datasets. To our knowledge, the ABI software does not have a control over the rate of introduced errors. From the results in Table 1, the datasets quantified with the *phred* quality score of 20 have significantly lower error rates (two-way analysis of variance $P$-value < 0.005) than the *Sequencing Analysis* processed datasets. The differences are more striking for the empirical rates of spurious insertions ($P$-value < 0.0001, data not shown). Apparently, *Sequencing Analysis* either tends to insert extraneous bases near specific nucleotide combinations or
poorly handles band spacing artifacts. On the other hand, Sequencing Analysis provides, on average, 30% more data that may be fixed with an error correction procedure like SAGEscreen. Given the cost of sequencing, this consideration is certainly viable, and the choice is project-specific and up to the investigator.

### 4.2 Assessment of the SAGEScreen efficacy

Initial conclusions on SAGEScreen efficacy may be drawn from Table 1. The numbers in rows ‘Estimated error rate’ and ‘Redistributed (%)’ indicate SAGEScreen finds and repairs ~50% of mutant sequence tags, with the assumption that the number of falsely identified mutants is negligible. To gauge the specificity/sensitivity balance, we have performed a simulation study. We treated the phred and SAGEScreen processed N1 library as the model transcriptome and drew 10 000 ditags according to their frequency distribution. This collection of 20 000 tags represented an error-free Long SAGE library. The same set of 10 000 ditags were mutated according to the error rates observed in the N1 library. As a result, the mutant library accumulated 2511 erroneous tags, 2344 of which were distinct tag sequences. The N1 library and, consequently, the error-free library are good examples of Long SAGE libraries dominated by low or moderately abundant tags. In the error-free library, the fraction of tags (in tag counts) with expression below five counts was close to 50%. Such a dataset is a suitable test for SAGEScreen as we suspect the algorithm will be less efficient in the situation of a close-to-flat transcript distribution. We started the analysis of the simulated data with the ditags. Ditag processing found 332 child ditags in the error-free library, 299 of which were exact matches; this lead to 33 misclassified ditags or 66 erroneously removed tags. In the mutant library, we ended up with 326 child ditags, 208 of which were exact matches. Given it is unlikely the mutations generated new pairs of sequence related ditags, Ditag processing has detected more than 93% of redundant mutant ditags. To simplify tracking of tag relationships, we applied the short version of SAGEScreen in the analysis of the mutant library. Results are summarized in Figure 4. The rates shown in the figure were calculated with respect to tag sequences as opposed to tag counts. The dotted line shows the rate of correctly detected mutant tag sequences plotted against the genuine tag abundance in the error-free library. As expected, SAGEScreen does not detect mutants with original abundances below 4. The algorithm corrects ~85% of errors for tags with parent counts between 4 and 10. Above 10, the algorithm corrects more than 96% of errors. The dashed line graphs the rate of false negatives, which behaves opposite to the rate of correct detections. The rate of false positives is graphed on the dash–dot line; it peaks at 8% and drops down to 0.5% for highly abundant tags. In addition to these rates, we plotted the fraction of detected erroneous tags that were assigned to wrong parent tags. This fraction does not exceed 2% in any of the categories. Overall, SAGEScreen identified almost 50% or 1175 mutant tags. It is worth mentioning that the set of mutants with original abundances 1 or 2, we will call them unrecoverable tags, constituted 36% of the erroneous tags. If the unrecoverable tags are excluded, the algorithm detected and repaired 78% of errors with the rate of false positives at 4%.

### 5 CONCLUSIONS

SAGE is a powerful technology that allows systematic harvesting of short ESTs. SAGE libraries are voluminous collections of transcriptional information and rich data sources for posterior computational and laboratory analyses. In this work, we present an extensive study of sequence errors introduced in the PCR and sequencing steps of the SAGE protocol. The impact of mutant sequence tags is expected to be especially important in the characterization of complete transcriptomes using Long SAGE. With this in mind, we have developed a correction algorithm, SAGEScreen. The program detects and removes spurious ditags and tags by grouping similar-sequence tags into clusters and analyzing their observations with a stochastic error model. SAGEScreen is applicable to standard SAGE and Long SAGE datasets. We demonstrate the performance of SAGEScreen on a number of Long SAGE datasets and assess its efficiency in simulations. Even under conservative testing criteria, the program greatly reduces the amount of erroneous sequences in the data. The simulations also indicate that any major improvement in performance of
the presented algorithm would only be possible with complete knowledge of the transcriptome under study.

ACKNOWLEDGEMENTS
We thank Stephen L. Madden for access to the Long SAGE libraries, Xiaohong Cao for helpful discussions, Peter Roberts for help with phred processing and the referees for valuable comments and suggestions.

SUPPLEMENTARY DATA
Supplementary data for this paper are available on Bioinformatics online.

REFERENCES