BIOINFORMATICS

Vol. 20 no. 14 2004, pages 2279–2287
doi:10.1093/bioinformatics/bth239

Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys

Edward Susko[1,*] and Andrew J. Roger[2]

[1]Genome Atlantic, Department of Mathematics and Statistics and [2]Genome Atlantic, Canadian Institute for Advanced Research, Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Received on November 5, 2003; revised on March 2, 2004; accepted on March 8, 2004
Advance Access publication April 1, 2004

ABSTRACT

Motivation: Expressed sequence tag (EST) surveys are an efficient way to characterize large numbers of genes from an organism. The rate of gene discovery in an EST survey depends on the degree of redundancy of the cDNA libraries from which sequences are obtained. However, few statistical methods have been developed to assess and compare redundancies of various libraries from preliminary EST surveys.

Results: We consider statistics for the comparison of EST libraries based upon the frequencies with which genes occur in subsamples of reads. These measures are useful in determining which one of several libraries is more likely to yield new genes in future reads and what proportion of additional reads one might want to take from the libraries in order to be likely to obtain new genes. One approach is to compare single sample measures that have been successfully used in species estimation problems, such as coverage of a library, defined as the proportion of the library that is represented in the given sample of reads. Another single library measure is an estimate of the expected number of additional genes that will be found in a new sample of reads. We also propose statistics that jointly use data from all the libraries. Analogous formulas for coverage and the expected numbers of new genes are presented. These measures consider coverage in a single library based upon reads from all libraries and similarly, the expected numbers of new genes that will be discovered by taking reads from all libraries with fixed proportions. Together, the statistics presented provide useful comparative measures for the libraries that can be used to guide sampling from each of the libraries to maximize the rate of gene discovery.

Finally, we present tests for whether genes are equally represented or expressed in a set of libraries. Binomial and χ^2 tests are presented for gene-by-gene comparisons of expression. Overall tests of the equality of proportional representation are presented and multiple comparisons issues are addressed. These methods can be used to evaluate changes in gene expression reflected in the composition of EST libraries prepared from different tissue types or cells exposed to different environmental conditions.

Availability: Software will be made available at http://www.mathstat.dal.ca/~tsusko

Contact: susko@mathstat.dal.ca

INTRODUCTION

Expressed sequence tag (EST) surveys are a powerful way to quickly characterize expressed genes from a given organism and are an efficient means for gene discovery in genomics projects (Gibson and Muse, 2002). In many cases, the redundancy of highly expressed transcripts makes it necessary to perform expensive and experimentally difficult ‘normalization’ protocols on cDNA libraries before large numbers of ESTs are gathered from an organism. Normalization is intended to make the frequencies of the genes in cDNA libraries more uniform so that random clone selection and sequencing approaches continue to yield sequences from genes that have not been previously sampled even after a large number of sequences has been obtained. However, currently, there are few, if any, rigorous methods available to assess the relative redundancy of various libraries prepared from the same organism or to evaluate whether protocols such as normalization have been successful. In this study, we present a number of statistical methods that can be used to estimate and compare the rate of gene discovery from clustered ESTs sampled from different cDNA libraries. These methods are ultimately useful for measuring the degree of redundancy of a library and in guiding the selection of the numbers of clones to be sampled from various cDNA libraries in the future to maximize the rate of gene discovery.

[*]To whom correspondence should be addressed.

For a single library, the problem is completely analogous to the problem of quantifying species frequencies, which has found applications in vocabulary word problems and artifact preservation models, based on observed samples of species. Here, the role of species is played by genes. We present a synthesis of some of the relevant single sample species literature. The main statistical measures for a single library that we review are coverage (Good, 1953) and estimates of the expected number of new genes (Good and Toumlin, 1956). Several interpretations can be given to coverage including the proportion of genes in the library that are represented in the sample of reads and the probability that a new read will already be represented in the library. As such it provides a measure of the uniformity of the library. For samples of the same sizes from multiple libraries, the library with the larger coverage probably has more redundant clones. An alternative single sample statistic of interest is an estimate of the expected number of additional genes that will be found in a new sample of reads. Both single sample statistics are non-parametric, requiring only independent sampling of reads. For estimates of the expected number of additional genes however, one obtains highly variable non-parametric estimates when the number of reads for which prediction is desired is larger than the number of reads in the dataset. Less variable estimates of the expected number of additional gene are obtainable through the negative binomial model of Fisher *et al*. (1943).

Analogous coverage and expected number of gene statistics can be constructed that adjust for overlap in the libraries. The coverage in this case would be the coverage of a single library from the samples of reads taken from all libraries. Coverage here can alternatively be interpreted as the probability that the next gene selected has already been sampled in any of the libraries. The expected number of new genes in the multiple library case is the expected number of new genes based on fixed numbers of new reads in all libraries. Such statistics allow a genomicist sequencing from two or more libraries to monitor the coverage and likelihood of new genes being found without having to separately consider the libraries.

EST surveys are not only useful for gene discovery, but are often conducted to evaluate differences in gene expression in different tissues or cells exposed to different conditions. For instance, the question of interest could be whether the expression levels of a particular gene differ in libraries prepared from different tissues as reflected by big differences in the number of EST reads corresponding to that gene obtained from the two libraries. Here the goals of EST surveys are similar to those of micro-array analysis (Gibson and Muse, 2002) and some of the same statistical issues arise. For a given gene, conditioning upon the total number of reads in all libraries, we present appropriate binomial and chi-squared test statistics for detecting differences in expression. An overall test of whether there are any differences in expression in the two libraries at all is constructed by aggregating the individual gene test statistics. Methods for comparing differences in expression have been presented in Audic and Claverie (1997). The methods presented here differ in that they do not require prior distributions and consequently have correct type I error probabilities regardless of the distributions of genes in the libraries.

An outline of the article is as follows. In the next section we describe the datasets that will provide running examples. In the following section we present the multinomial sampling framework and indicate what the single library measures are, including a new variance estimate expression for the expected number of new genes estimate. The statistics that utilize multiple library data are then presented. Finally methods for detecting differences in proportional representation of genes are presented. Proofs and more detailed derivations are given in Supplementary material.

EXAMPLE DATASETS

Data obtained from two libraries from each of two different organisms were utilized to test the proposed methods. ESTs were obtained by randomly selecting and sequencing clones from both non-normalized and normalized cDNA libraries from the amitochondriate protist *Mastigamoeba balamuthi*. In these cases, the normalized libraries were prepared from the non-normalized libraries and therefore the non-normalized library contains all the genes in the normalized library (but not vice versa). For the second organism, *Naegleria gruberi*, cDNA libraries were prepared from cells grown under different culture conditions: cells cultured aerobically and anaerobically. In this case, the libraries were separately prepared and will have some genes in common but will not necessarily contain all the same genes (i.e. some genes expressed under one culture condition may not be expressed under the other). The results of these EST studies will be published elsewhere and are used here only as test cases for the methods described.

Once EST data were obtained, the sequences were clustered into groups of identical sequence by individual library or by combining the two libraries using the base-calling and contig assembly programs phred, phrap and consed (Ewing *et al*., 1998; Ewing and Green, 1998; Gordon *et al*., 1998) using default parameter settings. These programs and detailed descriptions of the methods are available from http://www.phrap.org

In this study we have treated the EST clustering procedure as if there were no errors associated with it. Ideally, as long as full-length cDNA clones and long, high quality sequence reads are obtained all from the same end of the cDNA of the library, this assumption should not be problematic. However, as sequence read length and quality decreases and variation in the degree of truncation of cDNAs increases, it becomes possible for ESTs corresponding to the same genes to fail to be clustered together. In this case the error inherent in the clustering procedure could be significant and should be taken into account before further analysis of the clustering output. However, this issue is beyond the scope of this study.

SINGLE LIBRARY STATISTICS

If reads can be considered independent then the summary data for a single library is multinomial. Assuming that there are N genes in the library of interest, let p_i be the proportional representation of gene i. A uniform library would have each gene appearing with equal frequency, $p_i = 1/N$, which would simplify analysis greatly, but uniform libraries seldom arise. Usually there are at least a few genes that are over-represented. For a sample of n reads, let X_i be the number of times the i-th gene is represented. Then $(X_1, \ldots, X_N)$ is multinomially distributed:

$$P(X_1 = x_1, \ldots, X_N = x_N) = \frac{n!}{x_1! \cdots x_N!} p_1^{x_1} \cdots p_N^{x_N}.$$

What complicates analysis greatly is that any gene that is not represented in the sample of reads is not known to be in the library, or, in other words any x_i with $x_i = 0$ is unobserved. Indeed N is not known. An alternative description of the data that is useful is given by n_x, the number of genes that were represented x times in the sample of reads, $x = 1, \ldots$. Note that n_x is the size of a subset of genes and that $n_x = 0$ for $x > n$.

Coverage

The first single library statistic we consider is coverage. Coverage is the proportion of the library that appears in the sample of reads

$$C = \sum_{i=1}^{N} p_i I(X_i > 0). \tag{1}$$

Here $I(A)$ is an indicator of $A : I(A) = 1$ if A is true and 0 otherwise. For example, consider a sample of 10 reads with eight reads corresponding to a single gene, say gene 1, and 2 reads corresponding to another gene, gene 2. If the proportional representation of gene 1 in the library was $p_1 = 0.3$ and $p_2 = 0.1$ then $C = 0.4$. Note that since the p_i are unknown, coverage must be estimated and that even if the p_i were known, coverage would be a random quantity since it changes from sample to sample. An alternative interpretation for coverage comes from its expectation, $1 - \sum_i p_i(1 - p_i)^n$ which is the probability that the next read is for a gene that has already been represented in the family of reads; for a derivation see the section entitled 'Expected coverage' in the Supplementary material. Thus a large value of coverage implies that the likelihood of new discovery is small.

The approximately unbiased estimate of coverage that is most frequently used is

$$\hat{C} = 1 - n_1/n. \tag{2}$$

Recall that n is the number of genes that appear exactly once in the sample so that its expectation is the sum, over genes, of the probabilities that the genes appear exactly once which is approximately $1 - \sum_i p_i(1 - p_i)^n$. If the number of genes, N, in the library and the number of sampled genes, n is larger, then generally $C - \hat{C}$ is approximately normally distributed with mean 0 and standard error

$$\mathrm{se}(\hat{C}) = n^{-1/2}[(n_1/n) + (2n_2/n) - (n_1/n)^2]^{1/2}. \tag{3}$$

See the section 'Standard error for coverage' in the Supplementary material. A $(1 - \alpha) \times 100\%$ confidence interval is thus given by $\hat{C}^{\pm} z_{\alpha/2}\mathrm{se}(\hat{C})$.

Coverage was first discussed in Good (1953) who attributed it to Turing which led estimates of this form to sometimes be referred to as Turing-type estimates. Since then coverage has received a considerable amount of additional attention notably in Good and Toumlin (1956), Robbins (1968) and Esty (1983). In particular, Esty (1983) established normality results and derived its standard error. For the *Naegleria* dataset the coverages were 0.64 (0.02) for the aerobic library and 0.49 (0.02) for the anaerobic library based on $n = 959$ and 969 reads, respectively; standard errors are indicated in brackets. For the normalized *Mastigamoeba* library the coverage was 0.45 (0.03) based on 363 reads. Surprisingly, since normalization is supposed to lead to more uniform libraries, the coverage for the non-normalized library is approximately the same, 0.47 (0.02), based on 715 reads.

The expected number of reads required to discover a new gene

There is a close relationship between coverage and the expected number of reads required to discover a new gene. Consider the probability that no new genes are discovered in the first $k - 1$ reads and a new gene is discovered on the k-th read. Independently at any one of the first $k - 1$ reads the probability that the next read corresponds to a new gene is equal to the coverage at the time of the previous read. However, since no new genes are found in the first $k - 1$ reads the coverage remains constant. Thus the probability that the first $k - 1$ reads result in no new genes is C^{k-1}. Since the coverage at the $k - 1$st read is still C, the probability that k reads are required to find a new gene is described by the well-known geometric distribution, $p(k) = C^{k-1}(1 - C)$, from which it follows that the expected number of reads required to discover a new gene is $1/(1-C)$. The standard error for the estimate, $1/(1-\hat{C})$, of this expectation can be derived from the standard error for $\hat{C}$ using the delta-method as $\mathrm{se}(\hat{C})/(1 - \hat{C})$. The expected number of reads required to discover a new gene can be thought of as a simple redundancy index for a library. The expected number of reads for a new gene and confidence intervals, based on single library data, for the two example datasets are given in the last three columns of Table 1.

The expected number of new genes

An alternative single library statistic is the expected number of new genes, first derived in Good and Toumlin (1956) and considered in Efron and Thisted (1976) as well. Let η_x denote

Table 1. The single and two library estimates of the expected number of new reads required to find a new gene

Library		Two library data		Single library data	
		Reads	95% CI	Reads	95% CI
Naegleria	Aerobic	3.38	(3.27, 3.49)	2.77	(2.67, 2.88)
Naegleria	Anaerobic	2.30	(2.22, 2.38)	1.97	(1.90, 2.05)
Mastigamoeba	Non-normalized	1.91	(1.83, 1.98)	1.89	(1.81, 1.97)
Mastigamoeba	Normalized	2.02	(1.90, 2.13)	1.82	(1.70, 1.93)

the expected value of n_x, the number of genes that will be represented x times in the a sample of n reads. Then the expected number of new genes, $\triangle(t)$, that will be found in a sample of size tn, where $t \in (0, \infty)$, is related to the η_x through

$$\triangle(t) \approx \sum_{x=1}^{\infty}(-1)^{x+1}t^x\eta_x. \tag{4}$$

The usual non-parametric estimate $\hat{\triangle}(t)$ is obtained by substituting n_x for its expectation η_x. As is shown in the section 'Expected number of new genes' in the Supplementary material, even with n as small as 100, the approximation for $t \leq 1$ is quite good as long as no single gene has proportional representation >50% in the library. As is shown in the section 'Standard error for expected new genes' in the Supplementary material, the asymptotic variance is given by

$$\sum_{x\geq 1} t^{2x}\eta_x - n^{-1}\Big\{\sum_{x\geq 1}(-1)^{x+1}t^x[x\eta_x - (x+1)\eta_{x+1}]\Big\}^2 - \sum_{x\geq 1}\eta_x(-1)^x[1 - 2(1+t)^x + (1+2t)^x]. \tag{5}$$

An example of the expected numbers of new genes with confidence bands for a *Mastigamoeba* library dataset is given for both normalized and non-normalized libraries in Figure 1. There is no overlap in the confidence regions and the expected number of genes for the non-normalized is larger than for the normalized library at the same new sample size indicating that sampling from this library is more likely to lead to new genes.

Since (4) is a sum of powers of t, for $t > 1$ the dominant contributions to the sum will come from n_x for large x. For instance, with $t = 2$, if most genes appeared in less then 10 reads but one gene appeared 30 times, the contribution to the sum from that one gene will be -2^{30} and $\hat{\triangle}(t)$ might quite possibly be negative as a consequence. If another read was taken and turned out to correspond to the gene that had appeared 30 times, the contribution would be 2^{31} resulting in an enormous change ($2^{31} + 2^{30}$) in $\hat{\triangle}(t)$. In short, for non-uniform libraries, where some large clusters of reads can be expected, $\hat{\triangle}(t)$ is usually highly variable and unstable for $t > 1$. A similar problem arises with the third sum of (5) since its terms are also powers of numbers that are greater than 1. One possible solution is to substitute less variable parametric estimates of η_x rather than the observed counts n_x. The possible shortcoming with this approach is that the parametric models might not provide good fits to the data. However chi-squared goodness of fit statistics can be calculated to check the parametric fits and for most of the datasets we considered, a truncated negative binomial model provided a good fit to the data. The bad behaviour of non-parametric estimates of $\triangle(t)$ for $t > 1$ and the improvements provided by parametric estimates is illustrated in Figure 2 where $\hat{\triangle}(t)$ is plotted as a function of t for both non-parametric and parametric estimates of $\triangle(t)$. One can see that while the estimates are in reasonable agreement for $t < 1$, they start to diverge for $t > 1$ with the non-parametric estimate eventually becoming negative.

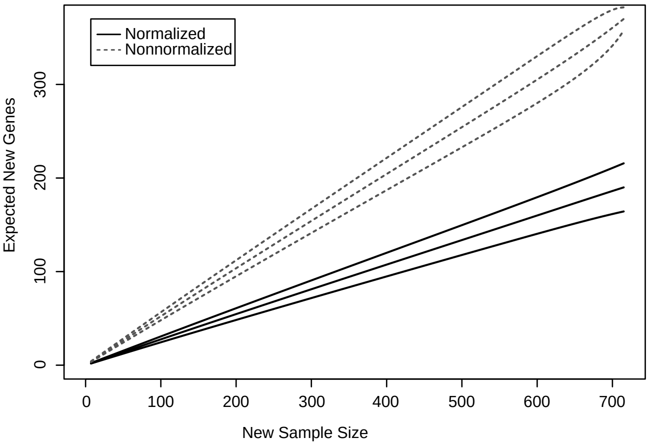

Fig. 1. Estimates of the expected numbers of new genes as a function of the size of a sample of new genes. The central lines give the estimates and the surrounding lines give 95% confidence bands. These plots are for the *Mastigamoeba* library dataset; normalized and non-normalized libraries.

The truncated negative binomial model considered is the one of Fisher *et al.* (1943). It is defined by specifying the probability that a randomly selected gene appears x times,

$$p(x) \propto \frac{\Gamma(x+\alpha)}{x!\Gamma(1+\alpha)}\gamma^{x-1}, \quad x = 1, \ldots. \tag{6}$$

If the model is fitted to all available data, $\alpha > 0$ is required, however as in Fisher *et al.* (1943) and Efron and Thisted (1976), more flexible models with $\alpha > -1$ and $\gamma > 0$ can be fit by restricting attention to genes that appeared between 1 and x_0 times. In the examples we considered we used $x_0 = 10$ since genes with larger numbers of reads than this are rare. In this case, the model can be fit with maximum likelihood to the data $(n_1, \ldots, n_{x_0})$ treating the individual $p(x)$ as multinomial probabilities for $x = 1, \ldots, x_0$. The constant of proportionality for $p(x)$ is obtained by rescaling to ensure $\sum_{x=1}^{x_0} p(x) = 1$. The expected number of new genes $\triangle(t)$

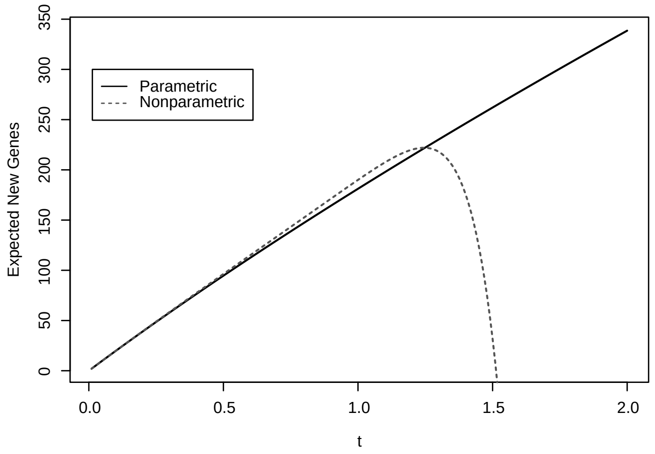

Fig. 2. Estimates of the expected numbers of new genes in a new sample of size tn as a function of the multiple t of the original sample size n. The estimated numbers are given for both parametric and nonparametric estimation for the *Mastigamoeba* library dataset. Note the difference in behaviour for $t > 1$.

can then be calculated by substituting expressions for η_x calculated under (6) with the estimated α and γ. The resulting sum $\triangle(t)$ simplifies to

$$\triangle(t) = \eta_1 \alpha^{-1} \gamma^{-1} \{1 - (1 + \gamma t)^{-\alpha}\}, \tag{7}$$

so that a parametric estimate of $\triangle(t)$ can be obtained by substituting maximum likelihood estimates for α and γ; which determines η_1 as well.

The truncated negative binomial model was fitted to the example datasets with $x_0 = 10$. The smallest p-value for a chi-square goodness-of-fit test of the parameteric model was 0.615 indicating that for each dataset the paramteric model was reasonable. Indeed, some of the p-values were quite large because of a very good fit of the expected counts under the parametric model to what was observed. The truncated negative binomial model described above, that allows $\alpha < 0$, was required however, as was indicated by the small α estimates for each of the datasets; each was less than -0.5. The big advantage with the parametric model is that the expected numbers of new genes can be calculated for new sample sizes greater than the original sample size ($t > 1$). This is illustrated in Figure 2 where reasonable parameter estimates are obtained for $0 \leq t \leq 2$. Because of the small α parameter, one can see from (7) that the expected number of new genes will be almost linear as a function of t as is evident in Figure 2.

MULTIPLE LIBRARY STATISTICS

Additional subscripts are required to discuss the multiple library case. Here p_{ij}, n_j and X_{ij} are the p_i, n and X_i values for the j-th library, $j = 1, \ldots, m$; here i indexes the N genes that are present in any library and, for instance, $p_{i1} = 0$ if a gene is not present in library 1 but is present in at least one of the other libraries. The notation $n_{x_1 \cdots x_m}$ differs however, denoting the number of genes that appeared x_j times in library j, $j = 1, \ldots, m$; $\eta_{x_1 \cdots x_m}$ denotes the expected value of $n_{x_1 \cdots x_m}$. Since the labels of the libraries are arbitrary, it suffices to consider statistical measures for library 1.

COVERAGE

Coverage for the first library is

$$C_1 = \sum_i p_{i1} \delta_i, \tag{8}$$

where $\delta_i = 1$ if gene i is represented among reads from any of the libraries and $\delta_i = 0$ otherwise. The estimate of C_1 is

$$\hat{C}_1 = 1 - n_{10\cdots0}/n_1. \tag{9}$$

As indicated in the appendix, generally, $C_1 - \hat{C}_1$ can be expected to be approximately normally distributed with mean 0 and standard error.

$$\mathrm{se}(\hat{C}_1) = n_1^{-1/2} \Big[(n_{10\cdots0}/n_1) + (2n_{20\cdots0}/n_1) - (n_{10\cdots0}/n_1)^2 \Big]^{1/2}. \tag{10}$$

Because two sample coverage is an estimate of the coverage of a given library from the reads taken from any of the libraries, it will be larger than coverage based on reads from a single library. Similarly, the expected number of reads from a given library needed to discover a new gene for any of the libraries, which is calculated as $1/(1 - \hat{C}_1)$, will be larger than the expected number of reads from a single library. This is indicated in Table 1 which gives the expected numbers of reads for two libraries based on both single and two library data. The estimated decrease in the expected number of reads due to using two library data is most noticeable for the *Naegleria* libraries.

The expected number of new genes

The expected number of new genes, $\triangle(t_1, \ldots, t_m)$, that will be found in a sample of $n_j t_j$ reads from library j, $j = 1, \ldots, m$, is related to the $\eta_{x_1 \cdots x_m}$ through

$$\triangle(t_1, \ldots, t_m) \approx - \sum_{x_1 + \cdots + x_m \geq 1} \eta_{x_1 \cdots x_m} \prod_{j=1}^{m} (-t_j)^{x_j}. \tag{11}$$

An estimate, $\hat{\triangle}(t_1, \ldots, t_m)$, is obtained by substituting $n_{x_1 \cdots x_m}$ for $\eta_{x_1 \cdots x_m}$. As with the single library estimate $\triangle(t)$, $\triangle(t_1, \ldots, t_m)$ becomes highly variable when any $t_j > 1$.

An alternative useful function in the two library case that one can obtain from (11) is the expected number of genes from a fixed number of new reads from both libraries. Given a fixed number n of new reads, the expected number of new genes can be obtained as a function of the number of new reads

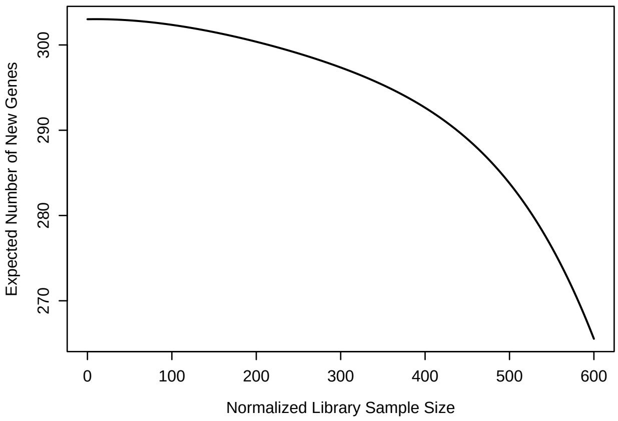

Fig. 3. Estimates of the expected numbers of new genes for the *Mastigamoeba* library data set for a fixed number, $n = 600$, new reads, as a function of the number of reads (n'_1) from the normalized library. In this case, for any given n'_1, the corresponding number from the non-normalized library would be $n'_2 = 600 - n'_1$.

from library 1, n'_1, by solving $n'_1 = t_1 n_1$ and $t_1 n_1 + t_2 n_2 = n$ to obtain the t_1 and t_2 that should be substituted in (11). An example plot of such a function is given in Figure 3 with $n = 600$. Since the function is decreasing it suggests that all new reads should be taken from the non-normalized library, however, the slow rate of decrease in the range of 0–300 reads indicates that taking as much as an even split of reads from both libraries will not appreciably affect the expected number of new genes.

TESTING EQUALITY OF PROPORTIONAL REPRESENTATION

Answers to questions about the expected numbers of new genes, coverage and the probability of finding new genes are of particular interest when dealing with differing libraries, often normalized versus non-normalized, obtained under similar conditions. However, often the libraries under comparison are prepared from different tissues of the same organism or cells exposed to different environmental conditions. A primary question of interest in these cases is whether there are differences in gene expression that are reflected in the composition of the different libraries. An example is provided by the two *Naegleria* libraries which were prepared from cells cultured aerobically and anaerobically. For these two conditions one expects different gene expression profiles that indicate how this primarily aerobic organism responds biochemically to oxygen deprivation. Although the general goals here are the same as those of a micro-array based analysis, the nature of the data gathered is quite different. For instance, in contrast to micro-array data where gene expression is measured as continuous variation in intensities of spots, in EST surveys, the expression level of a given gene is described as the numbers of times, $x_1, \ldots, x_m$ that the gene was encountered in samples of size $n_1, \ldots, n_m$ from the m libraries. Genes for which the x'_is are highly variable can be expected to be differentially expressed. Because the x_is are discrete there may be several genes with the same observed 'expression'. For instance, with two library data, 10 genes might share the property that reads for them came up five times in one library and eight times in the other. Below we discuss comparison of expression on a gene-by-gene basis, adjustments for the large numbers of comparisons that are being made and give an overall test of whether there are any differences in expression at all.

Tests of expression for a gene

We will fix a gene for which an expression comparison is of interest and let $X_1, \ldots, X_m$ denote the numbers of reads of the gene amongst the m libraries. Then, independently, each X_i has a binomial distribution with size parameter n_i and probability parameter p_i, the proportional representation of the gene in library i. Because n_i is usually large and p_i is small, the X_i have an approximate Poisson distribution with rate parameter $n_i p_i$. The null hypothesis of interest is $H_0 : p_1 = \cdots = p_m$.

Two library case In the two library case ($m = 2$), which is the main case that has been of interest here, an optimal test in the sense of being uniformly most powerful and unbiased can be calculated (Lehmann, 1991, pp. 145–156). It is a conditional test that rejects the null hypothesis when X_1 is less than c_1 or greater than c_2. The constants c_1 and c_2 must be calculated for an α level test so that

$$P(X_1 < c_1 | X_1 + X_2 = t) + P(X_1 > c_2 | X_1 + X_2 = t) \leq \alpha. \tag{12}$$

In order for the test to be optimal, the constants c_1 and c_2 must be chosen to satisfy the Equations (18) and (19) of Lehmann (1991, p. 147). In practice, c_1 and c_2 are usually taken as the $\alpha/2$ and $(1 - \alpha/2)$-th quantile of the conditional distribution of X_1 given $X_1 + X_2 = t$ which often coincides with, and approximately agrees with, the c_1 and c_2 that one would obtain solving the equations giving the optimal test. The distribution of $X_1 | X_1 + X_2 = t$ is hypergeometric but for most EST applications, t and x for any given gene are small relative to n_1, n_2. In this case, the hypergeometric distribution is well approximated by a binomial distribution with size parameter t and probability parameter $n_1/(n_1 + n_2)$.

Methods for comparing expression across two libraries have been discussed previously in Audic and Claverie (1997). They derive a form of a 'conditional distribution' of $p(X_1|X_2)$ for testing and reject the null if X_1 is large or small relative to what would be expected under $p(X_1|X_2)$. Audic and Claverie's derivation of $p(X_1|X_2)$ assumes a prior distribution for the proportional representations of genes. Although they favour a uniform prior, different priors can be used and give different critical regions. Assuming a uniform prior in this case is akin to assuming that in both libraries genes are uniformly distributed which need not be the case even in the null hypothesis of equality of representation for a given gene is true. Surprisingly, given the very different derivations, we found the critical regions for the two library test considered in were not very different in at least some cases. In particular, the critical regions given in Table 1 of Audic and Claverie (1997) are in reasonable agreement with the critical regions resulting from (12). Nevertheless the test presented here, as a uniformly most powerful unbiased test, is more soundly supported by theory.

Multiple libraries For multiple libraries, a test can be constructed using the conditional distribution of $(X_1, \ldots, X_m)$, given that $\sum X_i = t$, which under the null hypothesis is approximately multinomial with size parameter t and probability parameters $n_1/n., \ldots, n_m/n.$ where $n. = \sum n_i$. Thus, if the null hypothesis is true, the expected value of X_i is $t \times n_i/n.$. A chi-square statistic provides a natural statistic for comparing observeds and expecteds:

$$\chi^2(x_1, \ldots, x_m) = \sum_{i=1}^{m} (x_i - tn_i/n.)^2/(tn_i/n.). \qquad (13)$$

The p-value for this test is obtained as the probability, under the null hypothesis, that a random chi-square statistic is larger than the observed statistic. For large t, the distribution of a random chi-square statistic is approximately chi-square with $m-1$ degrees of freedom. Unfortunately t is often not large for the comparisons considered. There is a large body of experience with χ^2 tests, which in the present context indicates that use of the chi-square distribution gives reasonable approximations even for $tn_i/n. \geq 5$. For smaller values of t the number of possible $(x_1, \ldots, x_m)$ that sum to t becomes small so that the p-value can be computed by summing multinomial probabilities, from the multinomial distribution described above, of all choices of $x_1, \ldots, x_m$ that yield larger values of the chi-square statistic than the one that was observed.

Adjusting for multiple comparisons Tests of differences in expression for a single gene can be conducted using the chi-square or binomial test statistics described above. For a single gene, the probability of a false positive can be controlled to be small but when a large number of genes are tested, the overall probability that a false positive will be found becomes large unless a multiple comparisons adjustment is made. The adjustment that we use is described in Benjamini and Hochberg (1995) and controls the probability of experiment-wise false positives at less than α. It is simply described:

(1) Order the distinct p-values from smallest to largest: $p_{(1)} < p_{(2)} \ldots$, and let N_j denote the number of genes that gave p-value $p_{(j)}$.
(2) Reject the hypotheses corresponding to the k smallest p-values, where k is the largest i for which

$$p_{(i)} < \alpha \sum_{j=1}^{i} N_j/N.$$

Here N is total number of distinct genes. The Benjamini and Hochberg correction ensures that the probability of experiment-wise false positives is less than α. The actual probability of experiment-wise false positives can be much less than α. In addition the probability of false positives on a gene-by-gene basis will be much smaller than α. Thus it provides conservative decision rules about rejection that can be contrasted with the liberal decision rules implied by comparison of p-values to α.

An overall test of the equality of proportional representation

Whether one is interested in differences in expression or in comparing normalized and non-normalized libraries, one of the questions of interest is whether the frequencies of appearance, $p_{i1}, \ldots, p_{im}$ of the genes that were sampled are the same in each library across all genes; i.e. $H_0 : p_{i1} = \cdots = p_{im}$ for all i. In an analysis of gene expression differences, this hypothesis would provide the appropriate initial question, namely, are there any differences in expression at all? In the case that the answer to this question is yes, the follow-up question of interest is which genes have differences in expression? The methods of the previous subsections are applicable to this question. The reason that we have chosen to describe tests of this hypothesis after describing tests for gene-by-gene comparisons is that the chi-square tests used on a gene-by-gene basis can be aggregated to produce an overall test statistic.

Under the null hypothesis that $p_{i1} = \cdots = p_{im}$,

$$t_d := \sum_{x_1 \cdots x_m} n_{x_1 \cdots x_m} \chi^2(x_1, \ldots, x_m) - (m-1), \qquad (14)$$

has a mean of 0. A standard error for this test statistic, under the null hypothesis can be shown to be given by

$$\begin{aligned} \mathrm{se}(t_d)^2 = &\sum_{x_1 \cdots x_m} n_{x_1 \cdots x_m} [\chi^2(x_1, \ldots, x_m) - (m-1)]^2 \\ &- \sum_{j=1}^{m} \Big\{ \sum_{x_1 \cdots x_m} n_{x_1 \cdots x_m} [\chi^2(x_1, \ldots, x_m) - (m-1)] x_j \Big\}^2 \Big/ n_j. \end{aligned} \qquad (15)$$

Under the null hypothesis, $t_d/\mathrm{se}(t_d)$ has a large sample $N(0,1)$ distribution and so a p-value for a test of $H_0 : p_{i1} = \cdots = p_{im}$, all i $Pr(Z > t_d/\mathrm{se}(t_d)$ where Z has a $N(0,1)$ distribution.

While $\mathrm{se}(t_d)$ is a valid standard error when the null hypothesis is true and is be expected to be positive with large samples, it can turn out to be negative. To provide a conservative approach that is less likely to reject, one can ignore the second two sums in (15), resulting in the standard error estimate

$$\mathrm{se}_2(t_d)^2 = \sum_{x_1 \cdots x_m} n_{x_1 \cdots x_m} [\chi^2(x_1, \ldots, x_m) - (m-1)]^2.$$

The resulting test statistic, $t_d/\mathrm{se}_2(t_d)$ is guaranteed to be positive but the standard error will be inflated. We conducted

Table 2. The results of the binomial tests of equality of proportional representation for *Naegleria* library ($n_1 = 959$, $n_2 = 969$)

N	x	y	p-value	Reject Null
1	55	14	5.57e−07	Yes
1	0	13	2.61e−04	No
1	12	0	4.59e−04	No
1	18	4	4.03e−03	No
1	11	1	6.02e−03	No
1	17	4	6.71e−03	No
1	0	8	8.14e−03	No
1	10	1	1.12e−02	No
2	6	0	3.03e−02	No
3	10	2	3.69e−02	No
1	16	6	4.95e−02	No

Here N indicates the number of genes that appeared x times in the aerobic library and y times in the anaerobic library with the p-value being the next entry in the row. Also listed is the Benjamini and Hochberg decision rule when the false positive rate is controlled at $\alpha = 0.05$.

Table 3. The results of the binomial tests of equality of proportional representation for the *Mastigamoeba* library ($n_1 = 363$, $n_2 = 715$)

N	x	y	p-value	Reject Null
1	6	0	0.002	No
1	9	3	0.008	No
1	7	2	0.017	No
2	4	0	0.025	No
8	3	0	0.076	No
1	0	7	0.113	No
1	14	15	0.147	No
1	0	6	0.170	No
2	3	15	0.193	No
12	2	0	0.226	No

For each gene, x gives the number of reads in the normalized library and y the number in the non-normalized library.

simulation studies to check whether this significantly affected the null and alternative distributions and found that it did not.

Application to the example data

The p-values for the test of $H_0 : p_{i1} = \cdots = p_{im}$, all i were small for each of the libraries: $<10^{-5}$ for the *Naegleria* library and equal to 0.0004 for the *Mastigamoeba* libraries, which perhaps not surprisingly gave the most similar coverage values. Since there appear to be differences of proportional representation of genes in the libraries, the follow-up question is which genes those differences correspond to. Tables 2 and 3 give the 10 smallest distinct p-values and the Benjamini and Hochberg decision rule when the false positive rate is controlled at less than $\alpha = 0.05$. Among each of the libraries it is apparent that there are a small number of genes that have different proportional representations. For the *Mastigamoeba* libraries, none of the hypotheses about differences in proportion representation is rejected even though the overall test rejected the hypothesis of equality of proportional representation. This reflects the conservative nature of the multiple comparisons adjustment. For the *Naegleria* libraries (Table 2) one gene displays a significant difference in proportional representation between the two libraries. In this case, the gene was sampled 55 times in the aerobic and 14 in the anaerobic library. Further investigation showed that this gene codes for the translation elongation factor 1α, a protein that is a key component of the translation apparatus in all cells. The significant difference in expression of this gene between aerobic and anaerobic conditions makes sense since cells growing aerobically are rapidly dividing and need to produce proteins quickly and hence have a high expresion of translational proteins. On the other hand, anaerobically cultured cells grow very slowly (if at all) and are stressed and therefore produce much less protein. This case illustrates the potential usefulness of this statistical test in identifying genes that are differentially expressed in EST surveys.

CONCLUSIONS

We have presented two measures for the comparison of libraries: coverage or, equivalently, the expected number of reads for a new gene, and the expected number of genes for a given number of reads. Coverage is a simple measure of uniformity and most directly monitors the information about the library contained in a given collection of reads. The expected number of new genes is easily interpretable and allows one to ascertain the performance of the library for a larger extrapolation of new reads. Difficulties in extrapolation occur when the new number of reads exceeds the current number of reads, in which case parametric models like the negative binomial model considered here may be useful. The measures have been presented for both single and multiple library cases. The advantage with the multiple library formulas are that they take into account the additional information about overlap between libraries and therefore allow one to optimize future sequencing efforts by recommending the relative proportions one should sample from the libraries to maximize gene discovery.

Curiously, in the two cases examined here where normalized libraries had been constructed to decrease the redundancy of the original library, it appears that the normalization procedure was not successful. Indeed, a maximum rate of gene discovery would be achieved in both cases by sampling ESTs only from the non-normalized libraries. This kind of information is extremely invaluable to the researcher in order to assess the quality of cDNA libraries and to avoid wasting time and money on producing redundant ESTs from poor quality libraries.

The other class of EST library comparison problem that we have considered is gene-by-gene and overall comparisons of proportional representation of genes in the libraries. The tests presented are conditional tests, conditioning upon the total numbers of reads for each of the genes in all the libraries. The use of conditional statistics is important. First, in the two library case, it yields a uniformly most powerful, unbiased test. Second, it avoids some of the subtle difficulties that might arise with other tests of the equality of several proportions. These difficulties include the use of large sample distributions that fit poorly when the numbers of reads for a given gene comparison are small as well as what might be referred to as a truncation issue: the fact that we never make comparisons between libraries for genes which did not yield reads in any of the libraries.

ACKNOWLEDGEMENTS

The authors thank Å.Sjörgren and E.Gill for preparation and sequencing of the Naegleria libraries. The methods described in this paper were developed as a collaboration between the Prokaryotic Genome Diversity and Evolution and the Protest EST program projects funded by Genome Atlantic/Genome Canada. E.S. and A.J.R. are supported by the Natural Sciences and Engineering Research Council of Canada. A.J.R. thanks the Canadian Institute for Advanced Research Program in Evolutionary Biology, the Canadian Institutes for Health Research for and the Alfred P. Sloan Foundation for fellowship support.

SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

REFERENCES

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

Bartlett,M.S. (1938) The characteristic function of a conditional statistic. *J. Lond. Math. Soc.*, **13**, 62–67.

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

Efron,B. and Thisted, R. (1976) Estimating the number of unseen species: How many words did Shakespere know? *Biometrika*, **63**, 435–447.

Esty,W.W. (1983) A normal limit law for a nonparametric estimator of the coverage of a random sample. *Ann. Stat.*, **11**, 905–912.

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

Ewing,B., Hillier,L., Wendl,M., and Green,P. (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, **8**, 175–185.

Fisher,R.A., Corbet,A.S. and Williams,C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, **12**, 42–58.

Gibson,G. and Muse,S.V. (2002) A Primer of Genome Science. Sinauer Associates, Sunderland MA, USA

Good,I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Good,I.J. and Toumlin,G.H. (1956) The number of new species, and the increase in population coverage when a sample is increased. *Biometrika*, **43**, 45–63.

Gordon,D., Abajian,C. and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.

Holst, L. (1979) A unified approach to limit theorems for urn models. *J. Appl. Prob.*, **16**, 154–162.

Lehmann,E.L. (1991) Testing Statistical Hypotheses. Wadsworth and Brooks, California.

Mao,C. and Lindsay,B.G. (2001) A Poisson model for coverage problems with an application in genomic research. Technical report #01-06-01, Center for Likelihood Studies, Department of Statistics, Pennsylvania State University.

Robbins,H.E. (1968) Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Stat.*, **39**, 256–257.

Starr, N. (1979) Linear estimation of the probability of discovering a new species. *Ann. Stat.*, **7**, 644–652.