March 2, 2004

Supplementary material for
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*

[2] *Genome Atlantic, Canadian Institute for Advanced Research, Program in Evolutionary Biology,*

Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

This document contains material that was a part of the manuscript "Estimating and comparing the rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys" but was deleted in the interests of shortening the presentation. It provides useful additional material, including:

- Derivations and additional details associated with the results and methods presented.
- A table with an example data set. The results of clustering data from individual libraries for *Mastigamoeba* are shown in Table 1.
- A table with the estimated coverages and confidence intervals for the example data sets; Table 2.

- A table giving the negative binomial parameters for the example data sets.

- A figure giving the expected numbers of new genes as a function of the numbers of new reads from the *Mastigamoeba* libraries; Figure 1.

  As with the single library estimate $\triangle(t)$, $\triangle(t_1, \ldots, t_m)$ becomes highly variable when any $t_j > 1$. Since the sample size for the non-normalized *Mastigamoeba* library was larger than for the normalized library, the contours are plotted for a larger range of sample sizes from the non-normalized library. Because the counts n_{01} and n_{10} are much larger than most of the other counts the contours here are almost linear.

- A figure giving the histograms of simulated test statistics under both the null and alternative hypothesis for the overall test of the equality of proportional representation.

  The resulting test statistic, $t_d/\mathrm{se}_2(t_d)$, for the test of the equality of proportional representation is guaranteed to be positive but the standard error will be inflated. We conducted simulation studies to check whether this significantly affected the null and alternative distributions and found that it did not. Plots of the simulated $t_d/\mathrm{se}_2(t_d)$ for simulations conducted under the null and alternative distributions are given in Figure 2. One can see that a $N(0,1)$ curve provides a good approximation to the null distribution while the distribution is shifted to the right under the alternative distribution. In both simulations, 2 libraries were considered and 1000 reads were generated from each library. The proportions, p_i, of genes for the libraries were created by generating a set of numbers, λ_i, from the gamma distribution and then normalizing: $p_i = \lambda_i / \sum_i \lambda_i$. For the null simulations 5000 genes were in the common library. For the alternative simulation, 5000 genes were in one library with the other library containing 70% of these genes. For the alternative simulation, both libraries had p_i's that were separately generated from the gamma distribution as described above.

DERIVATIONS

Expected coverage - multiple libraries

We derive the result for the general case of multiple libraries. In the single library case, n would replace n_1 below and n_1 would replace $n_{10\cdots0}$ in what follows. The expectation of C_1 is

$$\begin{aligned} E[\sum_i p_{i1}\delta_i] &= E[1 - \sum_i p_{i1} I(X_{i1} = 0, \ldots, X_{im} = 0)] \\ &= 1 - \sum_i p_{i1} P(X_{i1} = 0, \ldots, X_{im} = 0) \\ &= 1 - \sum_i p_{i1} \prod_{j=1}^{m} (1 - p_{ij})^{n_j} \end{aligned}$$

If n_1 is large the expectation of C_1 will be approximately the same as the expectation of $\hat{C}_1$:

$$\begin{aligned} E[1 - n_{10\cdots0}/n_1] &= E[1 - n_1^{-1} \sum I(X_{i1} = 1, X_{i2} = 0, \ldots, X_{im} = 0)] \\ &= 1 - \sum_i n_1^{-1} P(X_{i1} = 1, X_{i2} = 0, \ldots, X_{im} = 0) \\ &= 1 - \sum_i p_{i1}(1 - p_{i1})^{n_1 - 1} \prod_{j=2}^{m} (1 - p_{ij})^{n_j} \end{aligned}$$

Expected new genes - multiple libraries

To see that (11) in the main text is approximately the expected number of new genes, note that

$$- \sum_{x_1 + \cdots + x_m \geq 1} \eta_{x_1 \cdots x_m} \prod_{j=1}^{m} (-t_j)^{x_j} = - \sum_{i=1}^{N} \prod_{j=1}^{m} (-t_j)^{X_{ij}} \delta_i \tag{1}$$

Since $\delta_i = 1 - I(X_{i1} = 0, \ldots, X_{im} = 0)$, the expectation of any individual term in the sum is

$$- \prod_{j=1}^{m} E[(-t_j)^{X_{ij}}] + P(X_{i1} = 0, \ldots, X_{im} = 0)$$

which gives

$$- \sum_{x_1 + \cdots + x_m \geq 1} \eta_{x_1 \cdots x_m} \prod_{j=1}^{m} (-t_j)^{x_j} = \sum_{i=1}^{N} \{ \prod_{j=1}^{m} (1 - p_{ij})^{n_j} - \prod_{j=1}^{m} [1 - p_{ij}(1 + t_j)]^{n_j} \} \tag{2}$$

The number of new genes can be expressed in terms of indicator functions as $\sum_{i=1}^{N} \delta_i'$ where δ_i' is 1 if gene i did not appear in any of the libraries in the initial reads, but does appear in at least one of the libraries in the new reads. Since new and old reads are independent, the expectation of δ_i' or, equivalently, the probability $\delta_i' = 1$ is equal to $\prod_j (1 - p_{ij})^{n_j}$ times the probability that the gene does appear in the new reads for at least one of the libraries. This is calculated as $1 - \prod_j (1 - p_{ij})^{t_j n_j}$ and so the expected number of new genes is

$$\sum_{i=1}^{N} \{ \prod_{j=1}^{m} (1 - p_{ij})^{n_j} - \prod_{j=1}^{m} (1 - p_{ij})^{n_j(1+t_j)} \} \tag{3}$$

Using that $(1 + x/n)^n \approx \exp(x)$ for n large, we get that

$$\begin{aligned}(1 - p_{ij})^{n_j(1+t_j)} &\approx \exp[-n_j(1 + t_j)p_{ij}] \\ &\approx [1 - p_{ij}(1 + t_j)]^{n_j}\end{aligned}$$

which implies that the individual terms in both (2) and (3) are approximately the same. Even with n_j as small as 100, the approximation is quite good for $p_{ij}(1 + t_j) < 1$, which for $t_j \leq 1$ means that there should not be a single gene that appears more than 50% of the time in any library.

Standard Error for expected new genes - single library

To obtain standard error formulae, we consider characteristic function argmuents similar to those of Bartlett (1938), Holst (1979) and Esty (1983). The goal is to obtain a limiting distribution for a statistic of the form $\sum_{k=1}^{N} f_k(X_k)$. We are interested in

$$\hat{\triangle}(t) - \triangle(t) = N^{1/2} \sum_{k=1}^{N} f_k(X_k)$$

where

$$f_k(X_k) = N^{-1/2} \sum_{x} (-1)^{x+1} t^x [I(X_k = x) - p_k(x)] \tag{4}$$

Since the statistic of interest is a sum of random variables it might be expected that a central limit type result applies. The difficulty is that the X_k come from a multinomial distribution implying that they are dependent. This difficulty is overcome by using the relationship between the Poisson and multinomial distributions.

Theorem 1. Let Y_k be independent Poisson random variables with mean np_k, $k = 1, \ldots, N$. Suppose that conditions sufficient for the large sample normality,

$$[\sum f_k(Y_k), n^{-1/2} \sum (Y_k - np_k)]^T \rightarrow_d N(0, \Sigma)$$

as $N \rightarrow \infty$, hold. Then, for large n, $\sum f_k(X_k) \sim N(0, \Sigma_{11} - \Sigma_{12}^2)$.

Since the Y_k are independent, it can be expected that a central limit theorem will hold. Because the Y_k are not identically distributed, conditions under which this will be the case are difficult to specify exactly but at the least require that that most of the p_i be small. In other words, the approximations here will be useful when a large number of genes are present in the library.

Proof. Let $\phi(s)$ denote the characteristic function of the statistic of interest $\sum f_k(X_k)$. As indicated in Bartlett (1938) (see also Holst 1979 and Esty 1983), $\phi(s)$ is related to the characteristic function $\phi(s,t)$ for $\sum f_k(Y_k)$ and $n^{-1/2} \sum (Y_k - np_k)$ through

$$\phi(s) = [2\pi n^{1/2} P(\sum Y_k = n)]^{-1} \int_{-\pi n^{1/2}}^{\pi n^{1/2}} \phi(s,t) \; dt \tag{5}$$

Stirlings formula gives that $n^{1/2}P(\sum Y_k = n) \approx (2\pi)^{-1/2}$ so that

$$\phi(s) \approx [2\pi]^{-1/2} \int_{-\infty}^{\infty} \phi(s,t)\ dt \tag{6}$$

Since conditions for the central limit theorem are met, $\phi(t,s)$ converges to the characteristic function of a normal variate with mean 0 and covariance matrix Σ. The variance for $n^{-1/2}\sum(Y_k - np_k)$ is calculated as $\Sigma_{22} = 1$, and so we have that

$$\begin{aligned} \phi(s,t) &\approx \exp(-[\Sigma_{11}s^2 + 2\Sigma_{12}st + t^2]/2 \\ &= \exp(-[s^2(\Sigma_{11} - \Sigma_{12}^2) + (t + 2s\Sigma_{12})^2]/2) \end{aligned}$$

Substituting this into (6), which can be justified using arguments similar to those of theorem 2 of Holst (1979), we get

$$\phi(s) \approx \exp(-[s^2(\Sigma_{11}-\Sigma_{12}^2)]/2)\times[2\pi]^{-1/2}\int_{-\infty}^{\infty} \exp(-[(t+2s\Sigma_{12})^2/2)\ dt = \exp(-[s^2(\Sigma_{11}-\Sigma_{12}^2)]/2)$$

which is the characteristic function for a $N(0, \Sigma_{11} - \Sigma_{12}^2)$ variate. □

Calculation of standard errors now amounts to calculation of the variances and covariances of $n^{-1/2}\sum(Y_k - np_k)$ and $\sum f_k(Y_k)$ where the Y_k are independent Poisson random variables, rather than calculation under the original multinomial model for the X_k.

With $f_k(X_k)$ as in (4), calculation under the Poisson model gives

$$\Sigma_{11} = N^{-1}\sum_{x\geq 1} t^{2x}\eta_k - N^{-1}\sum_k[\sum_x(-1)^{x+1}t^x p_k(x)]^2$$

and

$$\Sigma_{12} = n^{-1/2}N^{-1/2}\sum_{x\geq 1}(-1)^{x+1}t^x(x\eta_x - (x+1)\eta_{x+1})$$

To obtain an expression that allows estimation of the second term in Σ_{11} we calculate

$$\sum_k[\sum_x(-1)^{x+1}t^x p_k(x)]^2 = \sum_k[\exp(-2np_k) - 2\exp(-np_k t) + \exp(-2np_k t)]$$

and use the fact that $E[a^{Y_k}] = \exp[-np_k(1+a)]$ to justify that

$$E\{(-1)^{Y_k}[1-2(1+t)^{Y_k}+(1+2t)^{Y_k}]I(Y_k>0)\} = \exp(-2np_k) - 2\exp(-np_k t) + \exp(-2np_k t)$$

Thus

$$\begin{aligned}\Sigma_{11} &= N^{-1}\sum_{x\geq 1}t^{2x}\eta_x - N^{-1}\sum_k E[(-1)^{Y_k}[1-2(1+t)^{Y_k}+(1+2t)^{Y_k}]I(Y_k>0)\} \\ &= N^{-1}\sum_{x\geq 1}t^{2x}\eta_x - N^{-1}\sum_{x\geq 1}\eta_x(-1)^x[1-2(1+t)^x+(1+2t)^x]\end{aligned}$$

Since the variance of $\hat{\triangle}(t)$ is N times the variance of $\sum f_k(X_k)$ we obtain (5) of the main text.

Standard Error for coverage - multiple libraries

For coverage we are interested in the distribution of

$$C_1 - \hat{C}_1 = N^{1/2}\sum_{k=1}^{N} f_k(X_{k1},\ldots,X_{km})$$

where

$$f_k(X_{k1},\ldots,X_{km}) = N^{-1/2}[n_1^{-1}I(X_{k1}=1,X_{k2}=0,\ldots,X_{km}=0) - p_{k1}I(X_{k1}=0,\ldots,X_{km}=0)] \tag{7}$$

Similarly as in the single library case we obtain

Theorem 2. Let $Y_{k1}, \ldots, Y_{km}$ *be independent Poisson random variables with means* $n_j p_{kj}$, $k = 1, \ldots, N$. *Suppose that conditions sufficient for the large sample normality,*

$$[\sum f_k(Y_{k1}, \ldots, Y_{km}), n_1^{-1/2} \sum (Y_{k1} - n_1 p_{k1}), \ldots, n_m^{-1/2} \sum (Y_{km} - n_m p_{km})]^T \to_d N(0, \Sigma)$$

as $N \to \infty$, *hold. Then, for large* $n_1, \ldots, n_m$, $\sum f_k(X_{k1}, \ldots, Y_{km}) \sim N(0, \Sigma_{11} - ||\Sigma_{12}||^2)$.

Here Σ_{12} is the vector of covariances between $\sum f_k(Y_{k1}, \ldots, Y_{km})$ and the other components.

Proof. A relationship between the characteristic function $\phi(r)$ of $\sum f_k(X_{k1}, \ldots, X_{km})$ and the characteristic function $\phi(r, s_1, \ldots, s_m)$ for $\sum f_k(Y_{k1}, \ldots, Y_{km})$ and the other components can be derived:

$$\phi(r) = (2\pi)^{-m} [\prod_{j=1}^{m} n_j^{1/2} P(\sum Y_{kj} = n_j)]^{-1} \int_{-\pi n_m^{1/2}}^{\pi n_m^{1/2}} \cdots \int_{-\pi n_1^{1/2}}^{\pi n_1^{1/2}} \phi(r, s_1, \ldots, s_m)\, ds_1 \ldots ds_m \quad (8)$$

as a generalization of the case $m = 1$ given in Bartlett (1938), Holst (1979) and Esty (1983). As in the single library case, $n_j^{1/2} P(\sum Y_{kj} = n_j) \approx (2\pi)^{-1/2}$, so

$$\phi(r) \approx (2\pi)^{-m/2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \phi(r, s_1, \ldots, s_m)\, ds_1 \ldots ds_m \quad (9)$$

Using the central limit theorem assumption, $\phi(r, s_1, \ldots, s_m)$ converges to the characteristic function of a normal variate with mean 0 and covariance matrix Σ. The covariance matrix for $[n_1^{-1/2} \sum (Y_{k1} - n_1 p_{k1}), \ldots, n_m^{-1/2} \sum (Y_{km} - n_m p_{km})]$ is calculated as the identity matrix and so, letting $s = [s_1, \ldots, s_m]^T$, we have that

$$\begin{aligned}\phi(t, s_1, \ldots, s_m) &\approx \exp[-(\Sigma_{11} r^2 + 2r\Sigma_{12}^T s + s^T s)/2] \\ &= \exp[-r^2(\Sigma_{11} - ||\Sigma_{12}||^2)/2] \exp[-(s + r\Sigma_{12})^T (s + r\Sigma_{12})]\end{aligned}$$

Similarly as in the single library case, substituting this expression into (9) gives that

$$\phi(r) \approx \exp[-r^2(\Sigma_{11} - ||\Sigma_{12}||^2)/2]$$

□

With $f_k(X_{k1}, \ldots, X_{km})$ as in (7) calculation under the Poisson model gives

$$\Sigma_{11} = N^{-1} n_1^{-2} (\eta_{10\cdots0} + 2\eta_{20\cdots0})$$

and

$$\Sigma_{12} = N^{-1/2} [n_1^{-3/2} \eta_{10\cdots0}, 0, \ldots, 0]^T$$

Since the variance of $\hat{C}_1 - C_1$ is N times the variance of $\sum_k f_k(X_{k1}, \ldots, X_{km})$ we obtain (10) of the main text.

TABLES AND FIGURES

Table 1: The numbers of clusters (N_k) of sequences that were read k times from the non-normalized and normalized *Mastigamoeba* libraries.

	Non-normalized	Normalized
k	N_k	N_k
1	378	200
2	33	21
3	21	14
4	9	4
5	6	3
6	1	3
7	3	1
8	1	0
9	1	1
10	1	0
13	1	0
14	0	1
15	5	0

Table 2: The single library and two library coverages for several overlapping libraries.

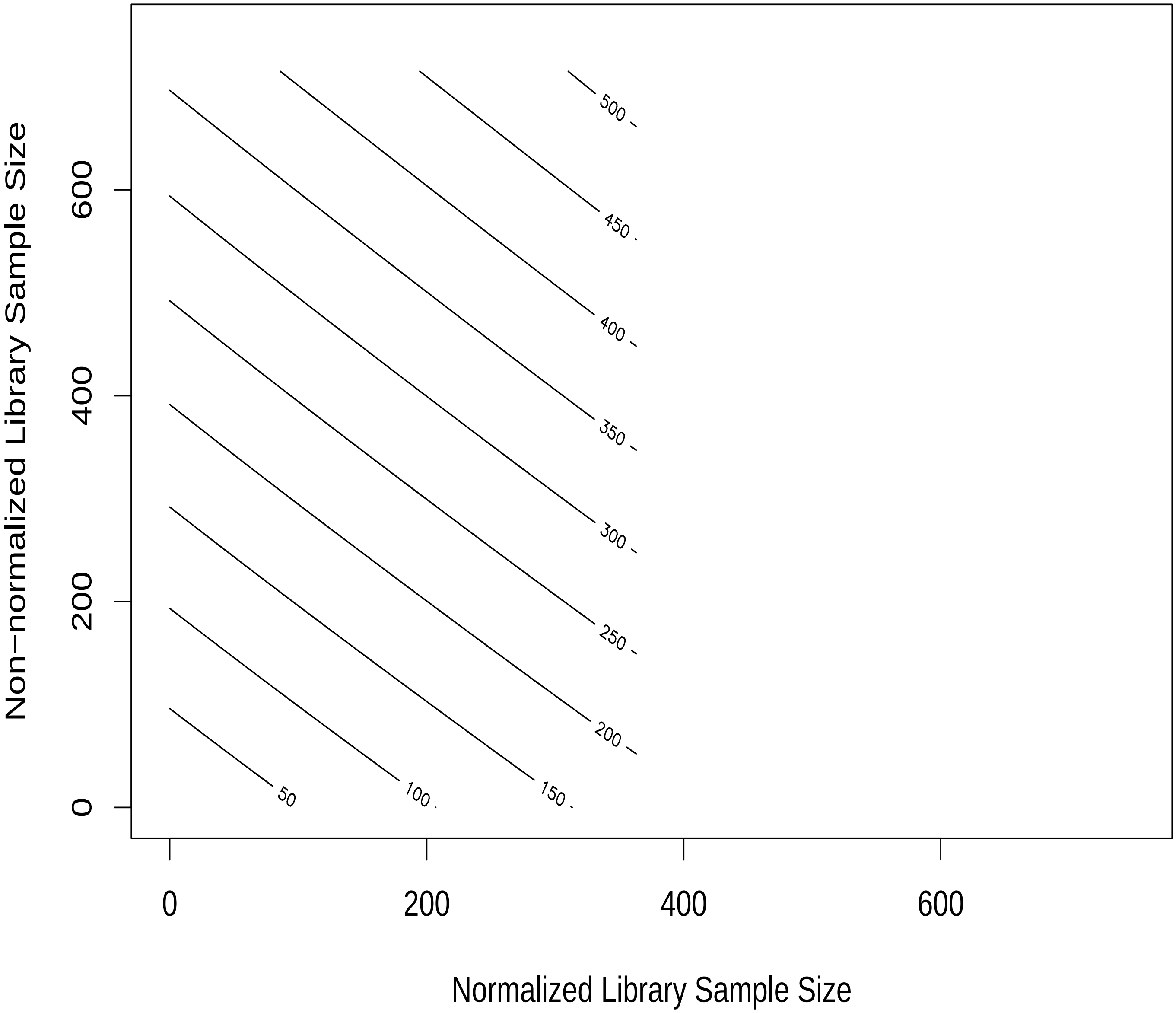
Library			Two Library Data		Single Library Data	
		# Reads	Coverage	95% CI	Coverage	95% CI
Naegleria	Aerobic	959	0.70	(0.67, 0.74)	0.64	(0.60, 0.68)
Naegleria	Anaerobic	969	0.57	(0.53, 0.60)	0.49	(0.45, 0.53)
Mastigamoeba	Non-normalized	715	0.48	(0.43, 0.52)	0.47	(0.43, 0.51)
Mastigamoeba	Normalized	363	0.50	(0.45, 0.56)	0.45	(0.39, 0.51)

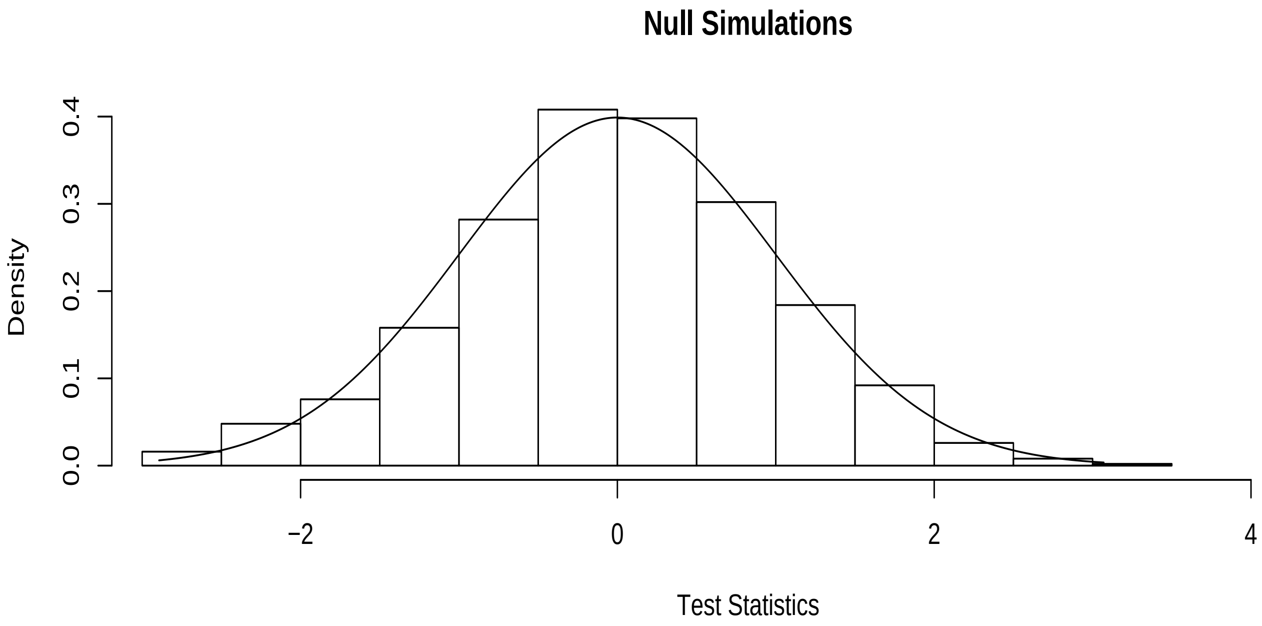
Table 3: The estimated negative binomial parameters for the example data sets and the p-values for a chi-square goodness of fit test.

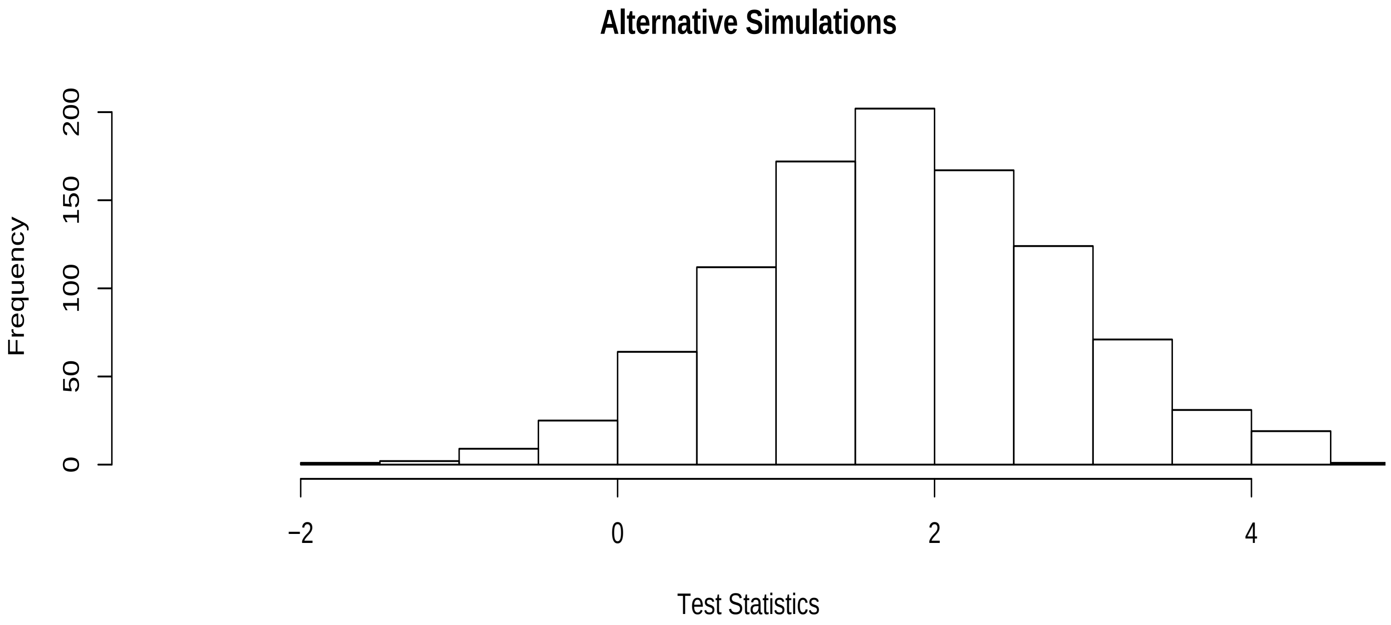
Library		α	γ	p-value
Naegleria	Aerobic	-0.699	1.000	0.822
Naegleria	Anaerobic	-0.638	0.834	0.615
Mastigamoeba	Non-normalized	-0.778	0.944	0.760
Mastigamoeba	Normalized	-0.715	0.889	0.813

Figure 1: Estimates of the expected numbers of new genes for the *Mastigamoeba* library data set for samples from both normalized and non-normalized libraries. Each contour line has indicated on it the estimate of the expected number of new genes for any pair of sample sizes from the libraries that falls on that line.

Figure 2: Histograms of the normalized test statistics for the test of the equality of proportional representation, simulated under both the null and alternative hypotheses.


Non−normalized Library Sample Size
Normalized Library Sample Size
0
200
400
600
0
200
400
600
50
100
150
200
250
300
350
400
450
500


Null Simulations
Density
0.0
0.1
0.2
0.3
0.4
−2
0
2
4
Test Statistics


Alternative Simulations
Frequency
0
50
100
150
200
−2
0
2
4
Test Statistics