Software for estimating and comparing rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys

Edward Susko and Wade Blanchard

Department of Mathematics and Statistics, Dalhousie University

Introduction

The main programs are

1. `cov_est`
2. `egene_est_single`
3. `egene_est_multiple`
4. `expr_est`
5. `equal_est`

The program `cov_est` does coverage estimation, the programs `egene_est_single` and `egene_est_multiple` give nonparametric estimates of the expected number of new genes, the program `expr_est` compares expression across libraries and the program `equal_est` tests whether two libraries are homogenous. All output is to the screen.

Installation

To create the directory that contains the source code

```
$ gzip -d est.tar.gz
$ tar xvf est.tar
```

This will create the directory `est`. To create the executables in this directory type

```
$ make
```

This will create the program files `cov_est`, `egene_est_single`, `egene_est_multiple`, `egene_est`, `expr_est` and `equal_est` which should be copied to a directory in your PATH. The file `est.tar.gz` and the directory `est` can then be removed.

Coverage

The program `cov_est` calculates the library coverage, the expected number of reads for a new gene and 95% confidence intervals for these quantities.

```
$ cov_est nlines nlib < infile
```

The file `infile` should contain rows of the form

$$x_1 \ \cdots \ x_m \qquad n_{x_1 \cdots x_m}$$

where $n_{x_1 \cdots x_m}$ is the number of genes that appeared x_j times in library j, $j = 1, \ldots, m$. The number `nlines` should be the number of lines in `infile` which should also be the number of distinct values of $x_1 \cdots x_m$. The number `nlib` should be the number of libraries to be compared; m in the above. An example `infile` for two libraries is given below

```
4 3     1
2 9     1
...
1 0   284
0 1   421
```

The first line indicates that a single gene had the property that it appeared 4 times in the first library and 3 times in the second library. The last line indicates that 421 genes appeared once in the second library but never in the first. An example of the output to the screen is

```
Coverages, standard errors and 95 percent CI
Multiple Library
1 0.703858 0.016484 (0.671550,0.736167)
2 0.565531 0.018231 (0.529798,0.601265)
Single Library
1 0.639208 0.019090 (0.601791,0.676624)
2 0.493292 0.020281 (0.453542,0.533043)


Expected number of reads for a new gene, standard errors and 95 percent CI
Multiple Library
1 3.376761 0.055662 (3.267662,3.485859)
2 2.301663 0.041962 (2.219417,2.383908)
Single Library
1 2.771676 0.052912 (2.667969,2.875384)
2 1.973523 0.040025 (1.895075,2.051972)
```

The expected number of new genes

Single library

The expected number of new genes, $\triangle(t)$, that will be found in a sample with $n_j t$ reads from the jth library, is obtained with

```
egene_est_single nlines nlib ilib ntval tvalfile < infile
```

where `tvalfile` contains the `ntval` choices of t for which $\triangle(t)$ is desired. The variable `ilib` is the library for which estimates are to be obtained; `ilib` $= 0, \ldots$ `nlib`-1. The form of `infile` and the variables `nlines` and `nlib` are as described for the `cov_est` program. The program outputs to the screen a file with rows

$$t \quad \triangle(t) \quad \text{se}(\triangle(t))$$

An example of the output to the screen is

```
0.010000 1.997914 0.114555
0.020000 3.991711 0.229110
...
0.990000 188.165068 12.841672
1.000000 190.000000 13.118237
```

Multiple libraries

The expected number of new genes, $\triangle(t_1, \ldots, t_m)$, that will be found in a sample with $n_j t_j$ reads from library j, $j = 1, \ldots m$, is obtained through

```
egene_est_multiple nlines nlib ntval tvalfile < infile
```

where `tvalfile` contains the `ntval` choices of $t_1, \ldots, t_m$ for which $\triangle(t_1, \ldots, t_m)$ is desired. The form of `infile` and the variables `nlines` and `nlib` are as described for the `cov_est` program. The program outputs to the screen a file with rows

$$t_1 \quad \ldots \quad t_m \quad \triangle(t_1, \ldots, t_m)$$

An example of the output to the screen is

```
0.010000 0.010000 7.039942
0.020000 0.020000 14.059932
....
0.990000 0.990000 613.569336
1.000000 1.000000 619.000000
```

Tests of expression for a gene

Differences of expression for genes can be tested with

```
expr_est nlines nlib < infile
```

The form of `infile` and the variables `nlines` and `nlib` are as described for the `cov_est` program. Output to the screen will have rows

$$x_1 \quad \ldots \quad x_m \quad n_{x_1 \cdots x_m} \quad \text{p-value} \quad \text{B-H cutoff} \quad \text{rejection}$$

In the special case of two libraries, additional output includes the estimated probability that a random hit from any gene will be to library 1 and the estimated probability for the gene of interest that a hit will be to library 1.

$$\text{gene} \quad x_1 \quad x_2 \quad n_{x_1,x_2} \quad \text{P(random)} \quad \text{P(gene)} \quad \text{p-value} \quad \text{B-H cutoff} \quad \text{rejection}$$

An example of the output to the screen is

```
...
582 5 5 18 2.248306e-01 5.000000e-01 1.042481e-01 7.381479e-03 0
583 9 14 2 2.248306e-01 3.913043e-01 1.084371e-01 7.455171e-03 0
584 24 120 1 2.248306e-01 1.666667e-01 1.087861e-01 7.463359e-03 0
585 10 61 1 2.248306e-01 1.408451e-01 1.097402e-01 7.467453e-03 0
...
```

Output is sorted from smallest to largest p-value. If the p-value is less than the B-H cutoff, the null hypothesis of equal expression would be rejected at the $\alpha = 0.05$ level even after adjusting for multiple comparisons, which would give an entry of 1 in the final column.

The overall test of equality of proportional representation

An overall test of equality of proportional representation available through

```
equal_est nlines nlib < infile
```

outputs the test statistic, standard error and p-value.The form of `infile` and the variables `nlines` and `nlib` are as described for the `cov_est` program. An example of the output to the screen is

```
182.244191 39.860379 2.414720e-06
```

Susko, E. and Roger, A.J. (2004). Estimating and comparing rates of gene discovery and expressed sequence tag (EST) frequencies in EST surveys. *Bioinformatics*, **20**:2279–2287.