**To**: Don Delker PhD and Deb Neklason PhD

**From**: Kenneth M Boucher PhD

**RE**: Power for RNA-seq experiments

**Date**: June 11, 2012

Additional calculations added June 23, 2014

More additions August 3, 2017

**Statistical Methods**

As noted in Fang and Cui (2011), there is very little literature on the topic of sample size or power estimation for RNA-seq experiments. We used ideas from Fang and Cui (2011) and Liu and Hwang (2007) to simulate power for a two-sample test using negative binomial distributions to model the count data. Assumptions in the power calculations were drawn from the preliminary data from genes with mean counts ≥ 10, and included the following:

* The proportion of genes for which the null hypothesis is true (π0) was set at 0.90
* We used a shifted exponential distribution to model the fold change. The fold change for the alternative hypothesis was varied from 2.0-8.4, with a mean of 2.55 and a standard deviation of 0.56.
* The mean and variance of the counts were allowed to vary by gene, as in the pilot data. The following assumptions were used for the count data:
  + The mean for log of the counts was modeled as a normal distribution with mean 4.43 and standard deviation 1.31. With these assumptions the median mean count is 84.
  + The variance of the count from a negative binomial distribution may be written as (1 + over-dispersion)\*mean. The log of the over-dispersion was modeled as a normal distribution with mean 3.49 and standard deviation 1.57.

Additional assumptions for the calculations included the following:

* A 0.05 false discovery rate (FDR) threshold was used to determine statistical significance.
* The Anscombe transformation was used to transform the data (Anscombe, 1948). This is known to transform a Poisson distribution to approximately a standard normal distribution.
* Welch’s two-sample *t-*statistic was then applied to the transformed count data and used as the test statistic. The significance level of the *t*-statistic was estimated empirically.

Briefly, the simulation method proceeded as follows. For each simulated gene, a mean was chosen in the reference group, a standard deviation in the reference group, and a fold change were chosen using the distributional assumptions described above. Then N counts were simulated using the negative binomial distribution assumptions in each group, where N is the sample size. This was repeated M= 100,000 times per sample size. The test statistic was then applied to data for each of the M simulated genes. The empirical significance level of the test (obtained by applying the same method to null data) was adjusted and the process repeated iteratively to maintain FDR at 0.05. It can be shown that the process converges. The empirical power was calculated as the proportion of tests that were rejected. The simulations were performed using R 2.15.0 (R Development Core Team, 2011).

**Paired Comparisons**

A similar method was used for paired comparisons. Paired t tests were used. Positive correlation between the paired samples was incorporated. The mean and SD of the correlation were 0.48 and 0.22 respectively.

**Results – Independent samples**

The figure below presents the estimated average power as a function of the sample size.

This is for FDR = 0.05

O:\Consulting\Hagedorn, Curt\RNA-seq sample size for grant\Power for RNAseq.tif

**Second figure for FDR = 0.1**

**O:\Consulting\Hagedorn, Curt\RNA-seq sample size for grant\Power for RNAseq, FDR = 0.1.tif**

|  |  |
| --- | --- |
| **Power for Comparison of Two Independent Samples at FDR = 0.05** | |
| **N per group** | **Power** |
| 7 | 0.45996 |
| 8 | 0.50704 |
| 9 | 0.55512 |
| 10 | 0.59079 |
| 11 | 0.6222 |
| 12 | 0.64563 |
| 13 | 0.66892 |
| 14 | 0.69357 |
| 15 | 0.70394 |
| 16 | 0.71687 |
| 17 | 0.73239 |
| 18 | 0.74594 |
| 19 | 0.75648 |
| 20 | 0.77046 |
| 21 | 0.77725 |
| 22 | 0.78568 |
| 23 | 0.79248 |
| 24 | 0.7997 |
| 25 | 0.80601 |

**Results – Paired Samples**

|  |  |
| --- | --- |
| **Power for Comparison of Paired Samples at FDR = 0.05** | |
| **N** | **Power** |
| 7 | 0.513 |
| 8 | 0.584 |
| 9 | 0.630 |
| 10 | 0.662 |
| 11 | 0.685 |
| 12 | 0.721 |
| 13 | 0.740 |
| 14 | 0.757 |
| 15 | 0.774 |
| 16 | 0.786 |
| 17 | 0.801 |
| 18 | 0.811 |
| 19 | 0.822 |
| 20 | 0.831 |
| 21 | 0.837 |
| 22 | 0.840 |
| 23 | 0.849 |
| 24 | 0.855 |
| 25 | 0.860 |

**References**

Anscombe FJ (1948). The transformation of Poisson, binomial and negative binomial data. Biometrika **35** (3-4) 246-254.

Fang Z and Cui X (2011). Design and validation issues in RNA-seq experiments. Briefings in Bioinformatics **12** (3) 280-287.

Liu P and Hwang JTG (2007) Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. Bioinformatics 23 (6) 739-746.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0