Sample Size for RNA-Seq and similar Studies

Steven Hart Terry Therneau

April 30, 2018

1 Framing the problem

Sample size computations for an experiment that involves high throughput sequencing consist of a number of separate steps. The first, and most important, is to ask "what is the scientific question". This drives the entire process, and answers to the later steps cannot be addressed satisfactorily out of the context.

Here is a particular chain as illustration.

1. What is the scientific question?
   - At what sites do a set of tumor and normal samples differ in RNA expression?
2. How many samples will be needed?
   - Define an "important" difference.
     - Any site where the relative expression differs by 2 fold and is of a worthwhile amount (we don't want 2 transcripts vs 5 to be noticed).
     - At least 85% power at $\alpha = .01$ for any given site.
   - How much difference is there within group?
3. What is the trade off between coverage and biological samples?
   - A new biological sample = 3 sequencing runs (opportunity cost + time + dollars)
4. What depth of sequencing coverage will be required?
   - What is the pattern of coverage?
   - What fraction of sites can be lost due to uneven coverage?
     - 20%

Perhaps the hardest question concerns heterogeneity within a group, which can change by an order of magnitude as we go from cell lines to inbred animals to human subjects. If the true expression at a particular site varies by $\pm 30\%$ between control subjects, then using deep sequencing to measure the value for each individual subject to within $\pm .01\%$ is clearly a waste of resources. (Is a milligram scale necessary in a weight loss study?)

2 Usage

A formal sample size calculation for comparison of two groups will involve five factors, each of which is an argument to the rnapower function.

- The depth of sequencing and consequent expected count μ for a given transcript, argument `depth`.
- The coefficient of variation of counts within each of the two groups, argument `cv`.
- The relative expression that we wish to detect Δ, argument `effect`.
- The target false positive rate α and false negative rate β desired (or power $= 1 - \beta$), arguments `alpha` and `power`.
- The number of samples n in each group, argument `n`

Here is an example where we assume an average depth of coverage of 20, equal within group coefficient of variation of .4, and a range of effect sizes. The result is the number of subjects per group that are required. When only 1 CV value is given the program assumes the same value for both groups.

```
> library(RNASeqPower)
> rnapower(depth=20, cv=.4, effect=c(1.25, 1.5, 1.75, 2),
         alpha= .05, power=c(.8, .9))
           0.8       0.9
1.25 66.204618 88.629200
1.5  20.051644 26.843463
1.75 10.526332 14.091771
2     6.861294  9.185326
```

The resulting table has one row for each of the four effect sizes and one column for each power value. It requires a large sample size to detect a 25% increase with high power (89 per group), but the sizes fall off rapidly as the true biologic impact grows larger. The following calls vary the depth of sequencing and the CV.

```
> rnapower(depth=100, cv=.4, effect=c(1.25, 1.5, 1.75, 2),
         alpha= .05, power=c(.8, .9))
           0.8      0.9
1.25 53.594215 71.74745
1.5  16.232283 21.73042
1.75  8.521316 11.40762
2     5.554381  7.43574
> rnapower(depth=1000, cv=.4, effect=c(1.25, 1.5, 1.75, 2),
         alpha= .05, power=c(.8, .9))
           0.8       0.9
1.25 50.756874 67.949053
1.5  15.372927 20.579988
1.75  8.070188 10.803691
2     5.260326  7.042083
```

```
> rnapower(depth=20, cv=.3, effect=c(1.25, 1.5, 1.75, 2),
         alpha= .05, power=c(.8, .9))
           0.8       0.9
1.25 44.136412 59.086133
1.5  13.367763 17.895642
1.75  7.017554  9.394514
2     4.574196  6.123551
```

Increasing the depth to 100 gives some decrease in sample size, but a further increase to 1000 results in only minor further gain. Reduction in the within-group CV has a more substantial effect. We have found that in most cases the between subject variability and not assay variation is the major component of error.

When calling the function any one of parameters except the depth of sequencing can be omitted, the function will then solve for that parameter. So for instance to compute the power of a run with a known sample size we could use

```
> rnapower(depth=8, n=10, cv=0.1, effect=c(1.5, 1.75, 2),
          alpha=.05)
      1.5      1.75         2
0.6941394 0.9258762 0.9880395
```

The function has two other optional arguments `n2` and `cv2`, which allow calculations for the case that the sample size or coefficient of variation differs between the two groups. The default is to assume that they are the same.

Values for the statistical parameters are traditionally set to α=.05 and power of .8 or .9. As a guide for depth and CV Hart et al [1] found that over a range of experiments 85%–95% of targets had coverage of .1 per million mapped, i.e. that a total depth of 40 million would give a coverage of ≥ 4 for the majority of targets. They also found an average within group CV of ≤ 0.4 for 90% of the genes in a range of human studies, with a lower CV of 0.1 for inbred animals. There will always be a few transcripts with either very low coverage or high within-group variation and for these sample sizes and/or depth would need to be very large. Power calculations would normally be targeted to the majority of targets that are better behaved.

3 Derivation

The set of counts y_{ij} for a particular transcript i and a set of subjects j often closely follows a negative binomial distribution with variance

$$\mathrm{var}(y_{ij}) = \mu_i + (\mu_i \psi)^2 \tag{1}$$

$$\mathrm{var}\log(y_i) \approx 1/\mu_i + \psi^2 \tag{2}$$

where μ_i is the average expected count at that locus and ψ is the expected coefficient of variation between subjects. The first term in the variance represents simple Poisson counting variation in the sequencer, the second is due to variation from subject to subject within the group. The value μ_i is the expected count for the feature, $E(y_i) = \mu_i$. For most studies the final analysis will focus on relative changes in the counts, e.g., "group 2 has a 55% greater expression of gene z than group 1", which is equivalent to a difference on log scale. Thus equation (2) is the more

appropriate one, and we see that the error for any given sample is proportional to the subject to subject variation plus approximately $1/\mu$ variation due to sequencing.

Hart et al [1] carry these derivations through formally in terms of a score test and give illustrative examples of coverage and within-group variation for several studies. The `rnapower` function is based on their resultant formula:

$$(z_\alpha + z_\beta)^2 = \frac{\Delta^2}{(1/\mu + \psi_1^2)/n_1 + (1/\mu + \psi_2^2)/n_2}$$

References

[1] Steven N Hart, Terry M Therneau, Yuji Zhang and Jean-Pierre Kocher, *Calculating Sample Size Estimates for RNA Sequencing Data*, submitted.