Power calculations for RNAseq experiments

The problem with power calculations

"Some individuals use statistics as a drunk man uses lamp-posts — for support rather than for illumination."
(Andrew Lang, 1937)

The main problem with doing a power calculation for an RNAseq experiment is that the calculation requires values that you can't be sure about until you've actually done your proposed experiment.

The other problem is that a power calculation is concerned with the probability of observing a particular result in isolation of other experimental observations. So a network of interlinked genes in a pathway whilst observably as clearly changed, may present at a value that would be under your defined threshold.

Caveats aside, this doesn't mean that a power calculation is meaningless, just one that needs to be through about carefully. And just because your proposed experiment "fails" a power calculation doesn't mean it is worthless.

What do I need to think about before I can do a calculation?

Before you can perform a power calculation there a number of things to think about that will inform the input values you need for your calculation. If you've got previous datasets then these can be helpful to determine these values, or you may be able to look at public datasets in your field to assist in your best guesses at these figures.

Number of samples / required level of power
Starting at the end, do you want your calculation to output the number of samples required to achieve a particular power, or are you interested in what power the samples and money you've got available can provide.

Significance level
How sure do you want to be in your finding once corrections have been made for multiple testing? Are you going to play safe with a $p<0.05$ threshold (so a 1:20 chance your finding is a false positive, or are you going to tighten things up with an awareness this will require more samples and cost more money?

How many genes are you going to be testing and how many are likely to be prognostic?
Most likely you'll be looking at all of the annotated genes in your genome, or maybe you're going to focus on just a cancer transcriptome? How many genes do you expect to be changed, counter intuitively the more genes you expect to be changed the higher the power obtained for the same set of values?

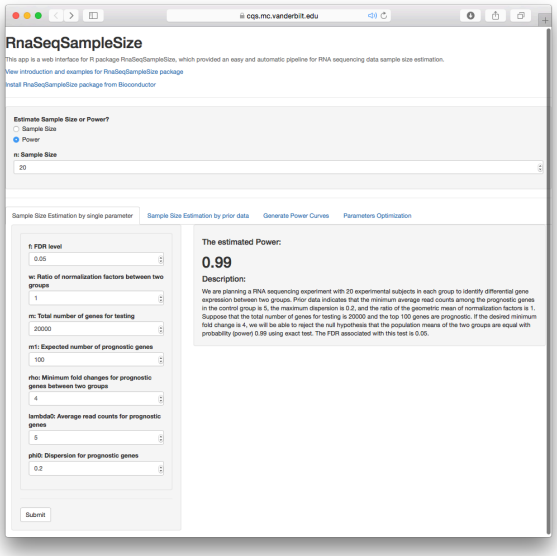
Fold-change
What magnitude of change are you hoping to identify?

Dispersion
How variable are the measurements you make for each gene? Do you have a highly heterogeneous patient cancer samples or a will defined cell line? The authors of the Scotty tool comment that in their experience, real world data dispersion varied between 0.2 and 0.4.

Tools for power calculations

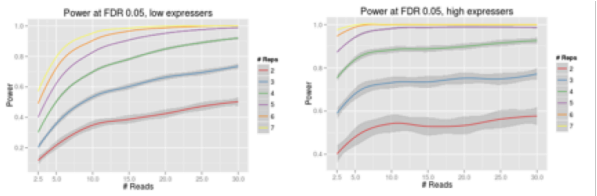
There are a number of different tools available for power calculations, but as a simple starting point for your explorations, the RnaSeqSampleSize tool is an easy to use online tool with a number of different options which can be tweaked to explore different experimental options.

```
https://cqs.mc.vanderbilt.edu/shiny/RnaSeqSampleSize/
```



Coverage vs samples

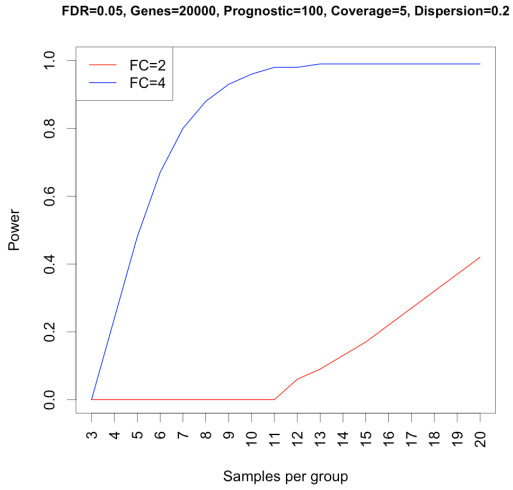
Work by Lui et al. (2014) shows that typically more samples is preferable to more reads and a series of graphs in the supplemental data show how reads affect power at a number of different samples sizes (2-7) for low, medium and highly expressed genes.

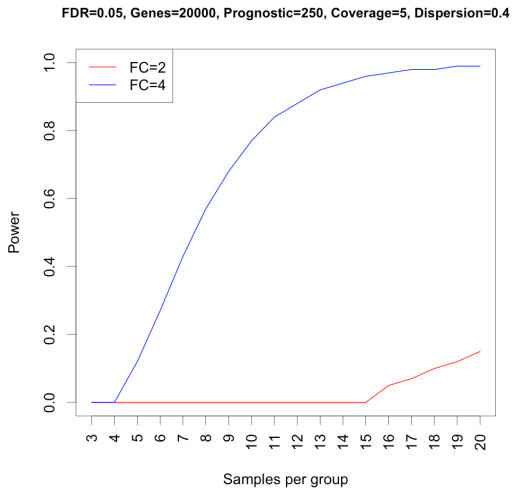


How many reads?

The power calculations presented here very much focused on the number of reads and samples required for differential expression profiling (typically 10-30M reads). If however you have other ambitions for your data analysis then you may need to look at greater coverage of say 50-100M reads for alternative splicing or 100M+ for de novo assembly or a transcriptome.

Examples





References

Liu Y et al. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30(3):301-304.

Busby MA et al. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics. 2013;29(5):656-657.

Guo Y et al. A Web Tool for Estimating Sample Size and Power for RNAseq Experiment. Cancer Informatics. 2014;13(Suppl 6):1-5.