Power and Sample Size for RNA-seq Experiments

OCTRI BERD Research Forum

Jessica Minnier
OHSU
11/1/23

Goals

- Review power and sample size statistical definitions
  - Power, sample size, effect size, type I error, type II error
- Overview of RNA-seq data generation
- Required components for power calculation
- Overview of tools available
- Examples of power calculations
- Brief overview of more complex designs and scRNA-seq issues

Prerequisites

- Some knowledge of basic power and sample size calculations and concepts
  - PSS 101 from BERD, worth reviewing and watching!
- Familiarity with RNA-seq experiments; good overview: (Conesa et al. 2016)
- Familiarity with statistical models and concepts such as regression, p-values, probability distributions (some review in context of RNA-seq: Harvard Chan Bioinformatics Core (HBC) training’s DGE and hypothesis testing lessons)
- Preferred: some experience with R/Bioconductor

RNA-seq

- Typically, transcriptomics: **gene expression profiling** of samples, using next-generation sequencing methods, reference genome

- RNA-seq can also examine alternative **gene spliced transcripts, mutations, levels of other RNA besides messenger RNA, etc**, but common aim is to evaluate gene expression measured by mRNA observed in sample tissue

- “**Bulk RNA-seq**” to distinguish from single cell RNA-seq, large populations of cells, mixed cell types

- **Alternative method: microarrays**, some of the following concepts still apply but distribution of expression measures is not count-based, but more related to continuous normal distribution (Lee and Whitmore 2002)

RNA-seq data generation

NGS or massively parallel sequencing: sample preparation, mRNA fragmentation, reverse transcription to complementary DNA, map cDNA to reference genome

(Griffith et al. 2015)

(Mistry et al. 2021)

What does the data look like?

Raw read sequencing data (large, fastq files) → Count data (gene ID x sample, matrix in .csv)

From a downstream statistical analysis standpoint:

- Level of measurement is the gene (or transcript) or “feature”
- Measurement/outcome is “gene count”: how many reads aligned to that gene’s RNA sequence after alignment to a reference genome? → count matrix
- Need a statistical model for count data that models the variability appropriately

Mistry et al. (2021)

Simplest Study: Two groups

Suppose we are studying two treatment groups (treatment vs. control), and we want to know which genes are differentially expressed between these two groups.

**Two groups**: Two experimental groups, multiple *biological* replicates within each group. The two groups contain different samples (i.e., not paired, not the same samples over time).

**Differentially expressed genes (DEGs)**: A gene is “differentially expressed” between the treatment vs. control if there is a difference observed in read counts or expression levels.

**Question of interest**: How many samples (biological replicates) in each group to we need to “detect a difference”?

- ... with some level of confidence (power), restricting false positives (type I error, FDR)?

Simplest Study: Two groups

Question of interest: How many samples do we need to “detect a difference”?

(Some) Information we need:

- Definition of “how different” (i.e. effect size, fold change)
- Desired level of power and type I error/p-value expectations (“significance level”)
- Information about sequencing depth, average read count
- Gene expression info: expected variability/dispersion of gene expression levels in each group, number of genes we expect to measure/observe
- Statistical model and test, method for multiple comparison adjustment
- “How much signal?” % genes differentially expressed above some fold change

Reminder: Statistical Power Concepts

Sample Size

More biological replicates leads to:

- better estimates of variation (gene, biological, sample-to-sample)
- identify outliers or possible sources of technical variation (batch effects)
- improve precision of estimates
- observe low abundance genes

Klaus (2015) and Mistry et al. (2021)

Components of Sample Size

In general, need to know 3 of the 4 to determine the 4th:

<table>
<thead>
<tr>
<th>Do We Know?</th>
<th>Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>??</td>
<td>Effect Size</td>
<td>Magnitude of difference or fold change</td>
</tr>
<tr>
<td>??</td>
<td>Sample Size</td>
<td>N total, n per group</td>
</tr>
<tr>
<td>0.05, 0.01</td>
<td>Type I Error / Significance level</td>
<td>$\alpha = \text{probability of rejecting null hypothesis when it is true}$ need to adjust for multiple comparisons!</td>
</tr>
<tr>
<td>0.9, 0.8</td>
<td>Power</td>
<td>$1 - \beta = \text{1 - Type II error = probability of rejecting null hypothesis when it is false}$ need to consider multiple comparisons!</td>
</tr>
</tbody>
</table>
Effect size $= \text{Mean Expression Ratio (FC)}$

In a two group comparison, effect size is usually \textit{fold change} (or logFC) for an individual gene:

$$\text{FC}_{\text{gene } X} = \frac{\text{mean expression of gene X in trt}}{\text{mean expression of gene X in control}}$$

May observe a large range of fold changes, or fold changes may be close to 1 for all genes

- FC measures the average size of the difference between groups, not variability
- Alternative effect size: biological coefficient of variation (BCV)

Volcano plot (hbctraining)

(Mistry et al. 2021)

Type I error (with multiple comparisons)

“The worst error” Truth = No difference, Test conclusion = there is a difference

- Typical power calculations are for one test, we could have tens of thousands of tests for each gene
- Multiple comparisons/multiple testing problem: If we use p-value < 0.05, for one test there is a 5% change we have a false positive, if we test 20,000 genes and use p<0.05, we would expect to detect 1000 false positives by chance; if we found 3000 DEGs total, one third are false positives
- Use “False Discovery Rate” = FDR as “adjusted” p-values
  - FDR = proportion of expected false positives in our set of DEGs, often controlled at 5% or 10%
  - Benjamini-Hochberg method common, alternative Storey/Q-value method, see (Mistry et al. 2021) for summary

Power (with multiple comparisons)

- With one test, power = 1 - type II error for a given effect size
  - Power = probability that our test is “significant” when the truth is that there is a difference
- With multiple tests/genes, could have many definitions
  - Probability we detect at least X% of genes that are truly different (at least some FC)
  - Probability that we detect all genes that are truly different (at least some FC)
    - Probability that we detect one specific gene that is truly different by some minimum FC
- Define “significant” as FDR < cutoff
Count data = discrete data (0, 1, 2…)

- Need **probability distribution** to match the type of data, such as: Poisson distribution, Negative Binomial distribution, Poisson-gamma mixture distribution

- Poisson model for count data assumes mean = variance, but RNA-seq has other **sources of variation** than the counting process $\rightarrow$ NB with dispersion parameter, other models

- Lower count genes can be harder to detect/measure/test

- Sequencing depth (total number of reads, i.e. 5-200 million) influences **average gene count**

(Mistry et al. 2021)
Variability

1. The mean is not equal to the variance (the scatter of data points does not fall on the diagonal).
2. For the genes with high mean expression, the variance across replicates tends to be greater than the mean (scatter is above the red line).
3. For the genes with low mean expression we see quite a bit of scatter. We usually refer to this as "heteroscedasticity". That is, for a given expression level in the low range we observe a lot of variability in the variance values.

(Mistry et al. 2021)

Factors affecting gene expression + power

- **Biological differences and variation**
- Sequencing depth/coverage
- Gene length
- Total sample RNA output, variation between samples within biological groups
- Variance/dispersion of gene abundance distributions
  - dispersion: parameter that defines how far we expect observed count will deviate from mean value (estimate from a model)

Sequencing depth vs. N

Typical analysis

- Common tools: DESeq2, edgeR, baySeq, and limma/voom packages in R/Bioconductor
- Normalization for gene length, sequencing depth (TMM, RPKM)
- DESeq2, edgeR, baySeq use generalized linear models (GLMs) with negative binomial (NB) distribution to fit the data:

\[ K_{ij} \sim \text{NB}(\lambda_{ij}, \alpha_i) \]

- Voom models mean-variance relationships to use normal based models in limma (linear model with Empirical Bayes variance smoothing)
- Hypothesis test options: Wald Test, Likelihood Ratio Test, and similar

Power calculation software/tools

- See review in Jeon et al. (2023) and comparisons in Poplawski and Binder (2018)
- RNASeqPower Bioconductor/R package
  - closed form equation based on Score statistic, NB model
- RnaSeqSampleSize Bioconductor/R package and web/Shiny app
  - NB model, gene-specific power function, can overestimate sample size
- ssizeRNA R package
  - NB model with normal-based test statistic via voom
- PROPER R package
  - NB model simulation based

Based on above reviews and simulation studies, ssizeRNA and PROPER were most recommended

Example R code, ssizeRNA

Adapted from ssizeRNA vignette and Suppl Material S1 Jeon et al. (2023)

Arguments for ssize_single() function, assuming all genes have the same characteristics:

- Total number of genes: \( G = 10000 \)
- Proportion of non-DE “null” genes: \( \pi_0 = 0.8 \)
- FDR level to control: \( \text{fdr} = 0.05 \)
- Desired average power to achieve: \( \text{power} = 0.8 \)
- Average read count for each gene in control group: \( \mu = 10 \)
- Dispersion parameter for each gene: \( \text{disp} = 0.1 \)
- Fold change for each gene: \( \text{fc} = 2 \)

```r
## Install ssizeRNA package if needed
# install.packages("ssizeRNA")

## Load package
library(ssizeRNA)

size_out <- ssizeRNA_single(nGenes = 10000, 
pi0 = 0.8, 
frd = 0.05, 
```
power = 0.8,
mu = 10,
disp = 0.1,
fc = 2)

# Output variables:
# ssize: sample sizes (for each treatment) at which desired power is first reached. size$ssize
size_out$ssize

pi0 ssize power
[1,] 0.8 14 0.8343262

# power: power calculations with corresponding sample sizes.
size_out$power

n 0.8
[1,] 3 0.0000000
[2,] 4 0.0000000
[3,] 5 0.0056349
[4,] 6 0.0953837
[5,] 7 0.2332468
[6,] 8 0.3697507
[7,] 9 0.4895573
[8,] 10 0.5898277
[9,] 11 0.6719140
[10,] 12 0.7383110
[11,] 13 0.7916434
[12,] 14 0.8343262
[13,] 15 0.8683667
[14,] 16 0.8954845
[15,] 17 0.9170581
[16,] 18 0.9341977
[17,] 19 0.9478082
[18,] 20 0.9586136
[19,] 21 0.9671815
[20,] 22 0.9739771
[21,] 23 0.9793624

Typical inputs

- Dispersion often depends on animal model, might be 0.1 for mouse data, 0.2-0.5 for human data
- Read count depends on sequencing depth and distribution of abundance, if using average can be low, typically 5-30, but in reality will vary widely by gene
- Proportion of non-DEG, varies greatly based on experiment/biological difference expectations
- Fold change, the magic number :) 1.5, 2, 2.5…
- Power 0.9 and FDR 0.05, more “relaxed” adjust power 0.8-0.85 and/or FDR 0.1

Interactive web app

- RnaSeqSampleSize Bioconductor/R package and web/Shiny app

Pilot data or public data

- Ideally, pilot data specific to the experiment would be used to determine expected variability/dispersion and effect sizes

- ssizeRNA gives examples on using existing data to estimate \( \mu \) and \( \text{disp} \) across all the genes, then use ssizeRNA\_vary() function, and see below from Suppl Material S1 Jeon et al. (2023) for R code

- RnaSeqSampleSize has TCGA data examples, see vignette, and also web app

```r
## Install other packages if needed.
# install.packages("BiocManager")

library(edgeR)
library(Biobase)
library(ssizeRNA)

## Example data saved in ssizeRNA package:
## Step 2-1. load hammer dataset (Hammer, P. et al., 2010)
## two group rat data
data(hammer.eset)

counts <- exprs(hammer.eset)[, phenoData(hammer.eset)$Time=="2 weeks"]
counts <- counts[rowSums(counts) > 0, ]
dim(counts)
```

```
1 18463 4
```

```
trt <- hammer.eset$protocol[which(hammer.eset$Time=="2 weeks")].
```
### After generating count data with column names of control and treatment, you may estimate the parameters of mu and disp using the following process.

#### mu: average read count in the control group

```
mu <- apply(counts[, trt == "control"], 1, mean)
```

```
<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>9.5</td>
<td>146.5</td>
<td>952.0</td>
<td>662.5</td>
<td>386345.0</td>
</tr>
</tbody>
</table>
```

#### disp: dispersion parameters estimates using the edgeR package with count data.

```
d <- DGEList(counts)
d <- calcNormFactors(d)
d <- estimateCommonDisp(d)
d <- estimateTagwiseDisp(d)
```

```
disp <- d$tagwise.dispersion
```

```
<table>
<thead>
<tr>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05566</td>
<td>0.06991</td>
<td>0.11093</td>
<td>0.13949</td>
<td>0.18397</td>
<td>1.01143</td>
</tr>
</tbody>
</table>
```

#### Output variables:

```
# ssize: sample sizes (for each treatment) at which desired power is first reached. size ssize
size_outssize
```

```
pi0 ssize power
[1,] 0.8 9 0.8434739
```

```
# power: power calculations with corresponding sample sizes.
size_outpower
```

```
n 0.8
[1,] 3 0.0000000
[2,] 4 0.1495363
[3,] 5 0.4275414
```

More complex studies

What are the additional experimental factors?

- Multiple groups, multiple treatment combinations
- Interaction effects
- Continuous predictor rather than groups
- Time or repeated measure component
- Technical replicates
- Batch effects or technical variation

→ Likely need to use simulation based methods for power/sample size

Pathway analyses?

Even more complex to calculate power with simulations and many assumptions

## Single-cell RNA-seq or spatial transcriptomics

<table>
<thead>
<tr>
<th></th>
<th>Bulk RNA-Seq</th>
<th>Single-cell RNA-seq</th>
<th>High-throughput Spatial Transcriptomics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samples</td>
<td>C1</td>
<td>A, B, C, D</td>
<td>D1, D2, D3</td>
</tr>
<tr>
<td>Bulk Expression Profile</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cell</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Data Structure</strong></td>
<td>Subject x Gene Expression Count Data</td>
<td>Cell x Gene Expression Count Data</td>
<td>Cell/Spot x Gene Expression Count Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cell/Spot 2-dimensional Coordinates</td>
</tr>
<tr>
<td><strong>Detection Target</strong></td>
<td>Differentially Expressed Genes</td>
<td>Differentially Expressed Genes</td>
<td>Spatially Variable Genes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1.** Comparison of bulk RNA-seq, single-cell RNA-seq, and high-throughput spatial transcriptomics technologies in terms of the profiling resolution (level), data structure, and target discoveries.  
scRNA-seq

• scRNA-seq have more 0 counts than bulk which affects dispersions/variation, require “zero-inflated” model

• Power analysis for DEG (1) between biological groups/experimental conditions for a specific cell type or (2) between cell types or sub-populations within a sample

• Power analyses for identifying cell sub-populations, either within a single tissue, or proportional differences across experimental conditions

Table 2. A table with information about different software tools for scRNA-seq power analysis with two distinct detection targets. Experimental Factors: cell number (1), individual number (2), Sequencing depth (3).

<table>
<thead>
<tr>
<th>Detection Target</th>
<th># of Samples</th>
<th>Tool Name</th>
<th>Experimental Factor</th>
<th>Software</th>
<th>Model</th>
<th>Power Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell sub-population</td>
<td>Single sample</td>
<td><code>SCOPIT</code> [37]</td>
<td>(1)</td>
<td>R package &amp; Web application</td>
<td>Multinomial</td>
<td>Analytical</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>howmanycells</code></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>Sense</code> [38]</td>
<td>(1), (2)</td>
<td>Web application</td>
<td>Negative Binomial</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>scPOST</code> [39]</td>
<td></td>
<td></td>
<td>Beta Binomial</td>
<td></td>
</tr>
<tr>
<td>Multi sample</td>
<td></td>
<td><code>scPower</code> [40]</td>
<td>(1), (2), (3)</td>
<td>R package</td>
<td>Linear mixed model</td>
<td>Simulation-based</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>hierarchical</code></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEG</td>
<td>Single sample</td>
<td><code>powsimR</code> [42]</td>
<td>(1)</td>
<td>R package</td>
<td>A mixture of zero-inflated Poisson and log-normal Poisson distributions</td>
<td>Simulation-based</td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>POWSC</code> [43]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>scDesign</code> [44]</td>
<td>(1), (3)</td>
<td></td>
<td>Gamma-Normal mixture model</td>
<td></td>
</tr>
</tbody>
</table>

Conclusions

Important Factors (bulk RNA-seq DEG)

- Distributions: average (or distribution of) read counts, as well as average (or distribution of, or maximum) dispersion
- Proportion of differentially expressed genes
- Fold change effect size
- False discovery rate and power
- Experimental design complexity (beyond two groups, consult statistician)

Primary resources

- Jeon et al. (2023) and Mistry et al. (2021) tutorial

Thank you

- Recording will be available at OCTRI BERD Research Forum Website
- email: minnier [at] ohsu.edu
- Thanks for consultation/tips from ONPRC BBC Director Suzi Fei and all I’ve learned from IGL/MPSSR Directors Bob Searles and Chris Harrington
- Please fill out the survey that Amy Laird will send you!
- More “omics” PSS series to come
- Need stats help? Contact BDP or Knight BSR

References


