Power and Sample Size for RNA-seq Experiments

OCTRI BERD Research Forum

Jessica Minnier
OHSU

11/1/23

https://bit.ly/berd_pss_rnaseq

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

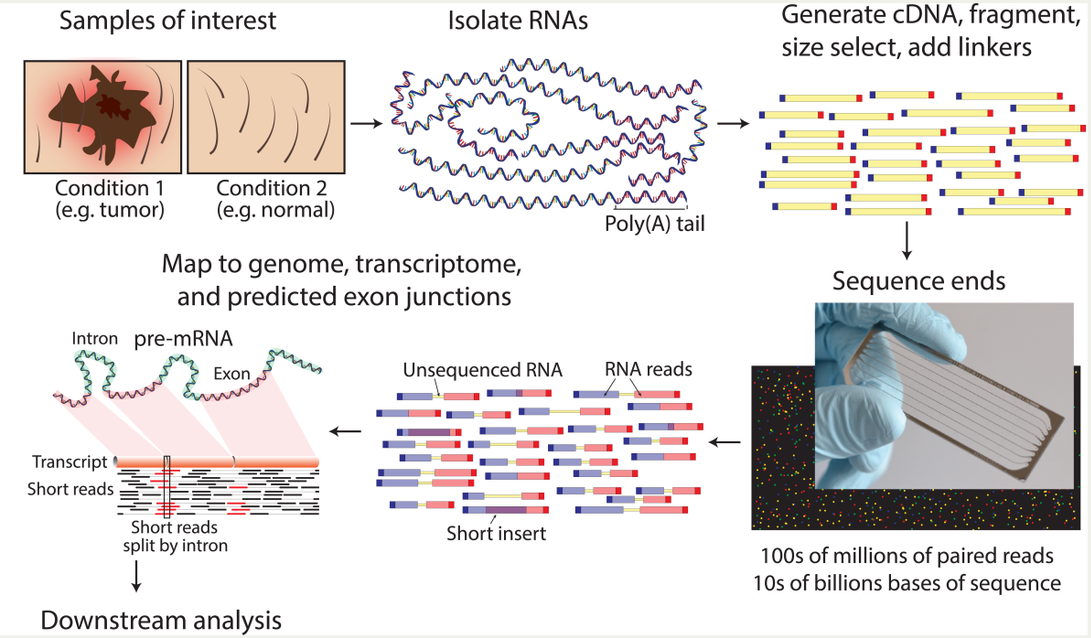
https://bit.ly/berd_pss_rnaseq

RNA-seq

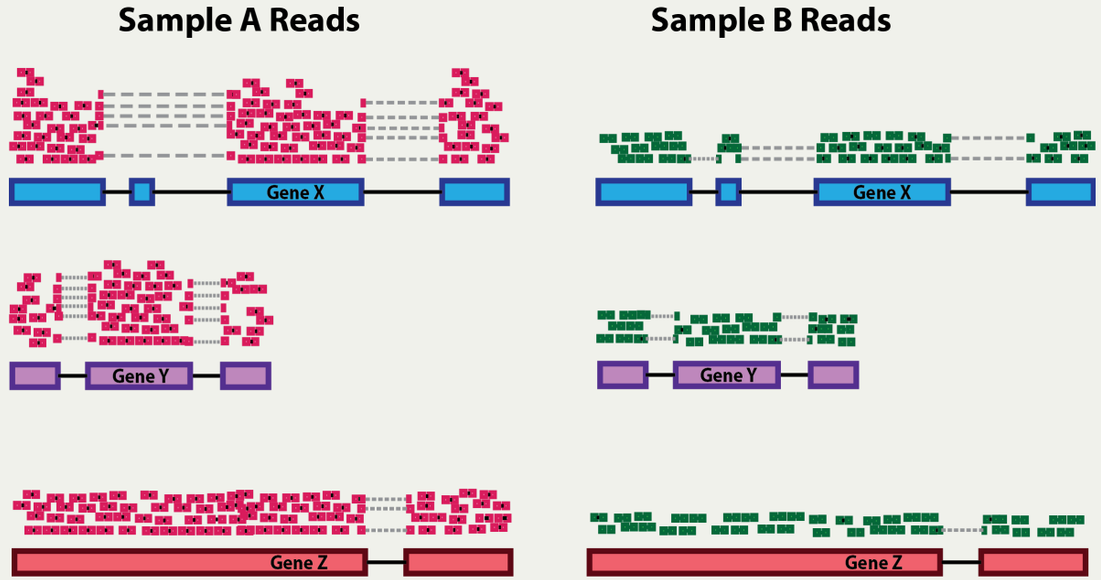
- Typically, transcriptomics: **gene expression profiling** of samples, using next-generation sequencing methods, reference genome
- RNA-seq can also examine alternative **gene spliced transcripts, mutations, levels of other RNA besides messenger RNA, etc**, but common aim is to evaluate gene expression 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)

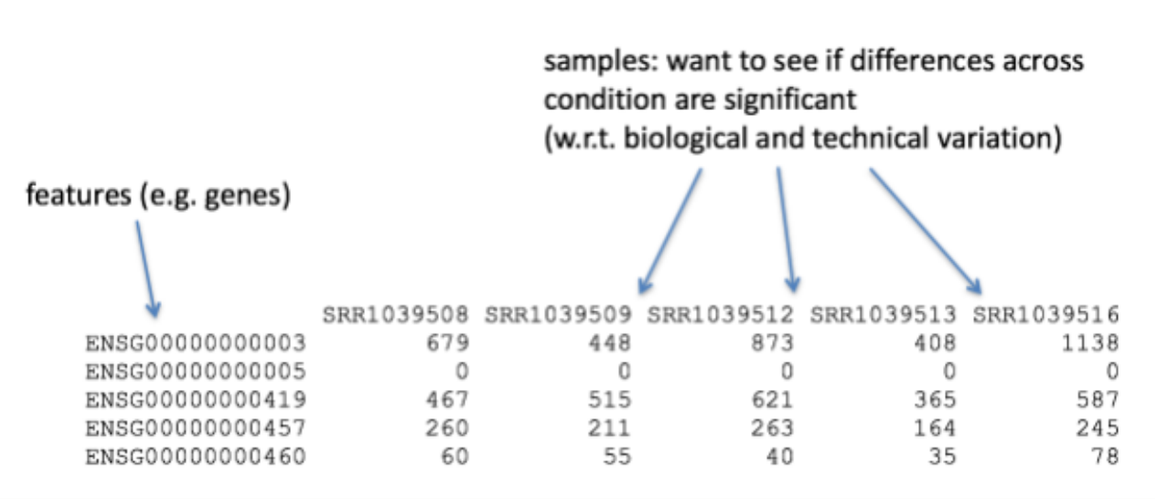
https://bit.ly/berd_pss_rnaseq

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



	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516
ENSG00000000003	679	448	873	408	1138
ENSG00000000005	0	0	0	0	0
ENSG00000000419	467	515	621	365	587
ENSG00000000457	260	211	263	164	245
ENSG00000000460	60	55	40	35	78

hbctraining Mistry et al. (2021)

https://bit.ly/berd_pss_rnaseq

Simplest Study: Two groups

Suppose we are studying two treatment groups (treatment vs. control), and we want to know which genes are deferentially 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 to 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

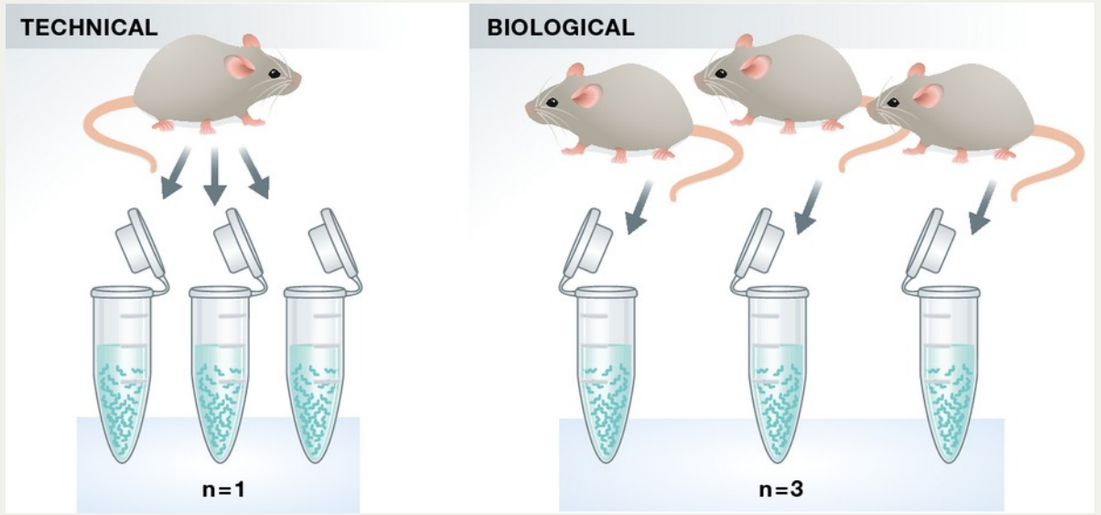
Reminder: Statistical Power Concepts

https://bit.ly/berd_pss_rnaseq

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)

https://bit.ly/berd_pss_rnaseq

Components of Sample Size

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

Do We Know?	Measure	Definition
??	Effect Size	Magnitude of difference or fold change
??	Sample Size	N total, n per group
0.05, 0.01	Type I Error / Significance level	α = probability of rejecting null hypothesis when it is true *need to adjust for multiple comparisons!*
0.9, 0.8	Power	1 - β = 1 - Type II error = probability of rejecting null hypothesis when it is false *need to consider multiple comparisons!*

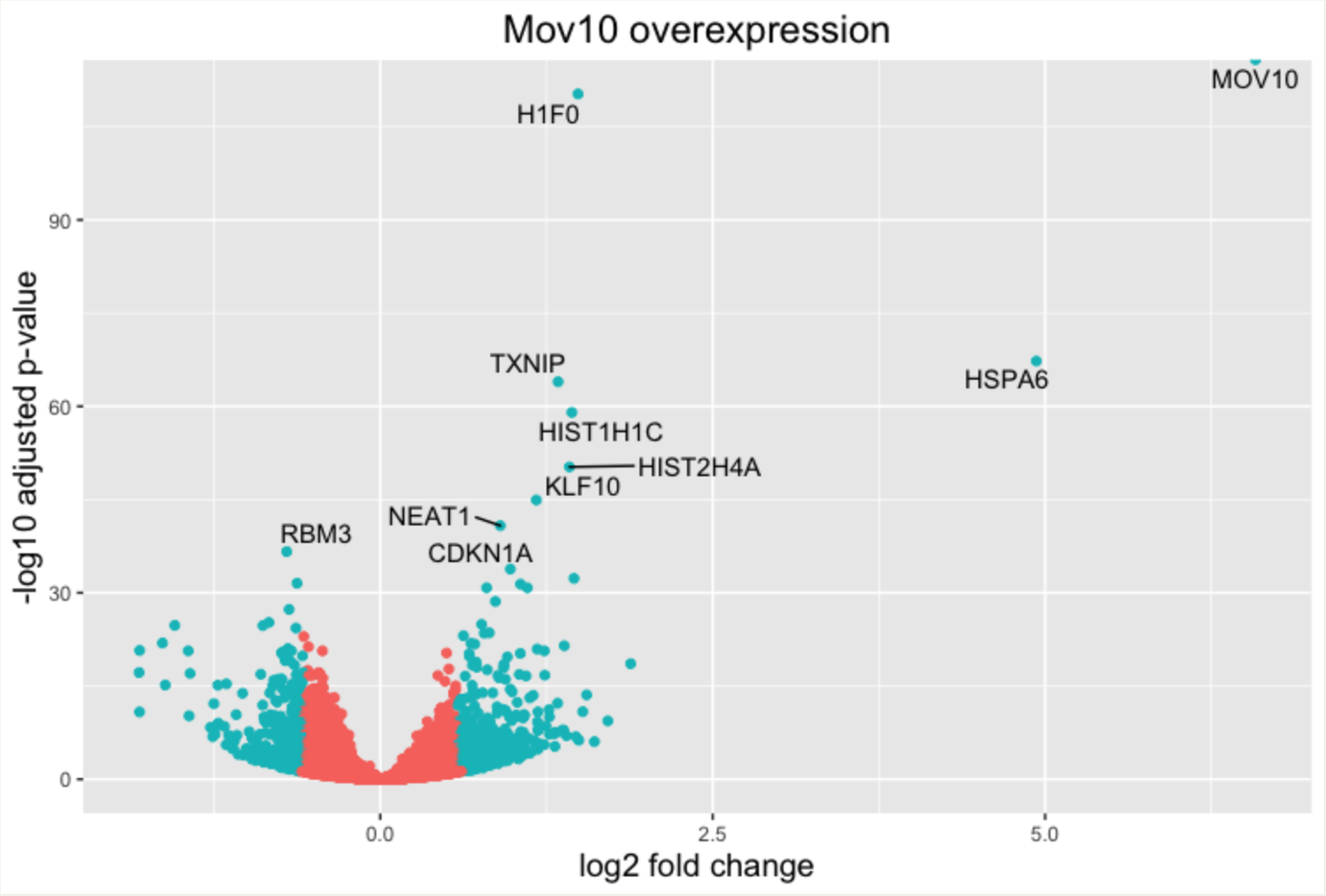
https://bit.ly/berd_pss_rnaseq

Effect size = Mean Expression Ratio (FC)

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

$FC_{gene\ X}$ = (mean expression of gene X in trt) / (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)

https://bit.ly/berd_pss_rnaseq

Type I error (with multiple comparisons)

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

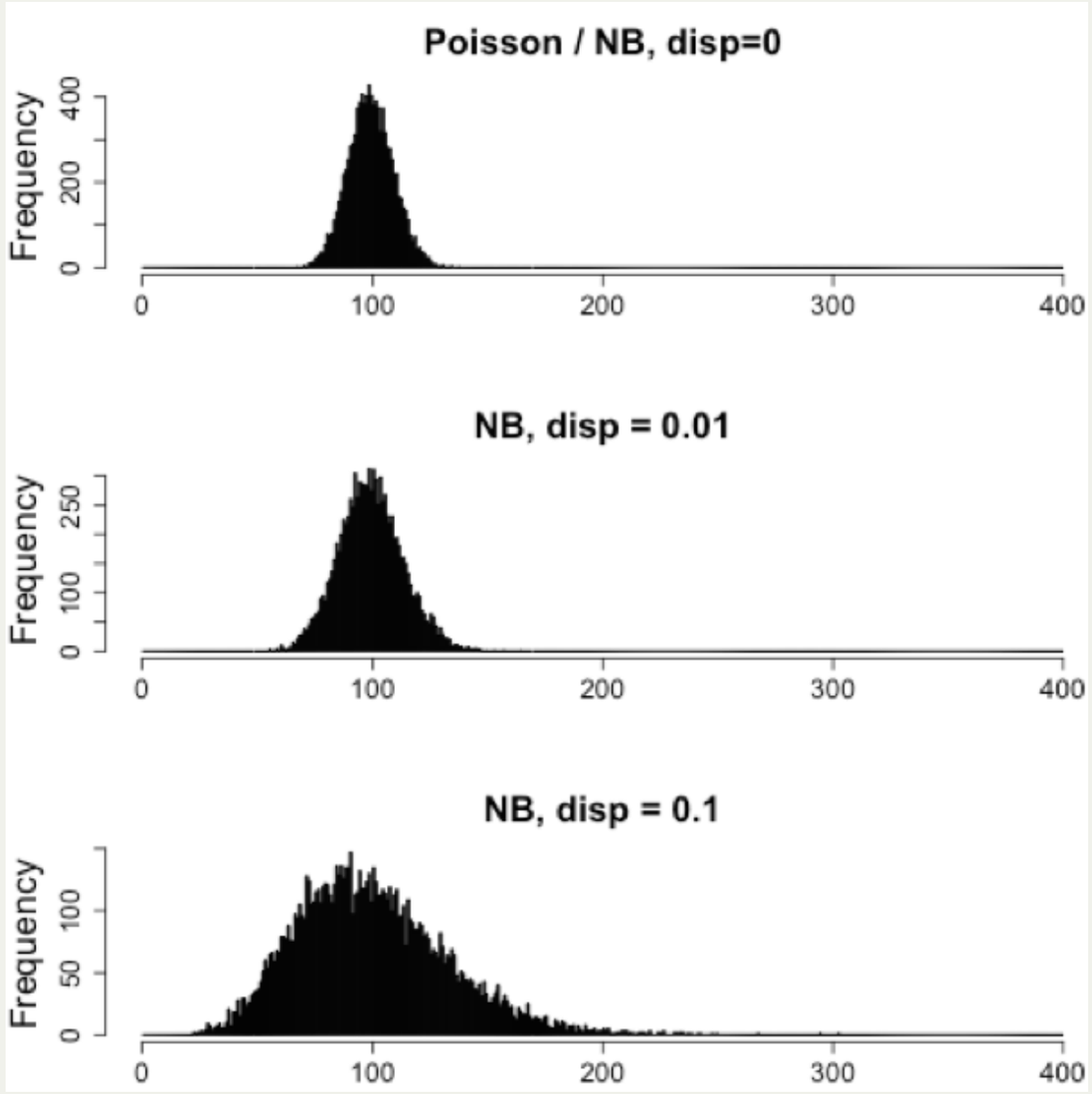
- Typical power calculations are for one test, we could have tens of thousands of tests for each gene
- Multiple comparisons/multiple testing problem: If we use p-value < 0.05, for one test there is a 5% change we have a false positive, if we test 20,000 genes and use p<0.05, we would expect to detect 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

Gene Expression/Abundance Distributions

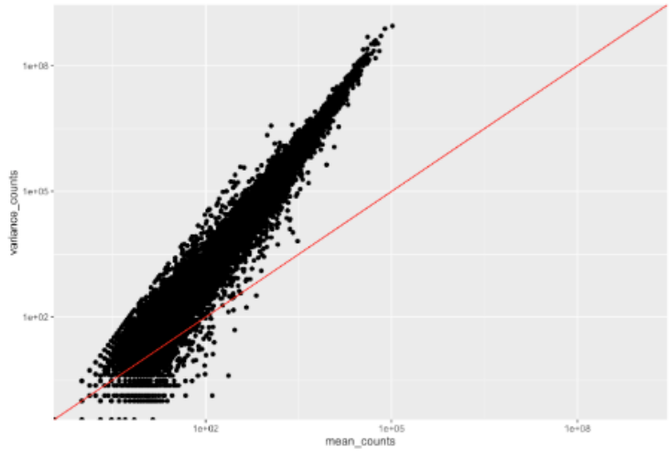
- Count data = discrete data (0, 1, 2…)
- Need **probability distribution** to match the type of data, such as: Poisson distribution, Negative Binomial distribution, Poisson-gamma mixture distribution
- Poisson model for count data assumes mean = variance, but RNA-seq has other **sources of variation** than the counting process → 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**



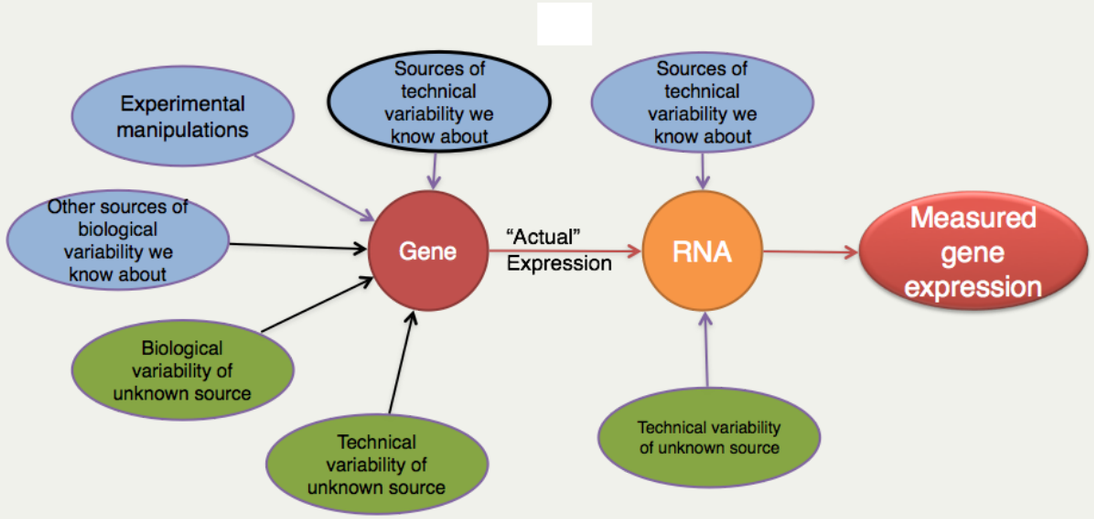
(Mistry et al. 2021)

https://bit.ly/berd_pss_rnaseq

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.



Courtesy of Paul Pavlidis, UBC

(Mistry et al. 2021)

https://bit.ly/berd_pss_rnaseq

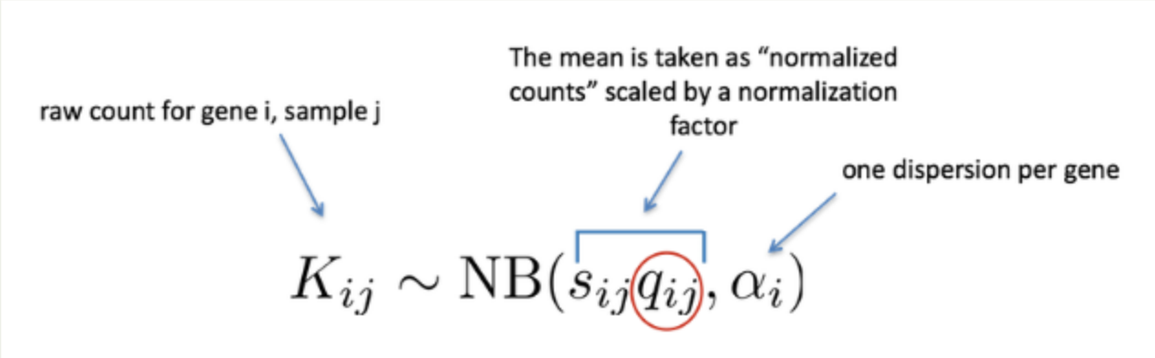
Factors affecting gene expression + power

- ***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:



- `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: `pi0 = 0.8`
- FDR level to control: `fdr = 0.05`
- Desired average power to achieve: `power = 0.8`
- Average read count for each gene in control group: `mu = 10`
- Dispersion parameter for each gene: `disp = 0.1`
- Fold change for each gene: `fc = 2`

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

## Load package
library(ssizeRNA)

size_out <- ssizeRNA_single(nGenes = 10000,
                      pi0 = 0.8,
                      fdr = 0.05,
```

https://bit.ly/berd_pss_rnaseq

```
                      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
```

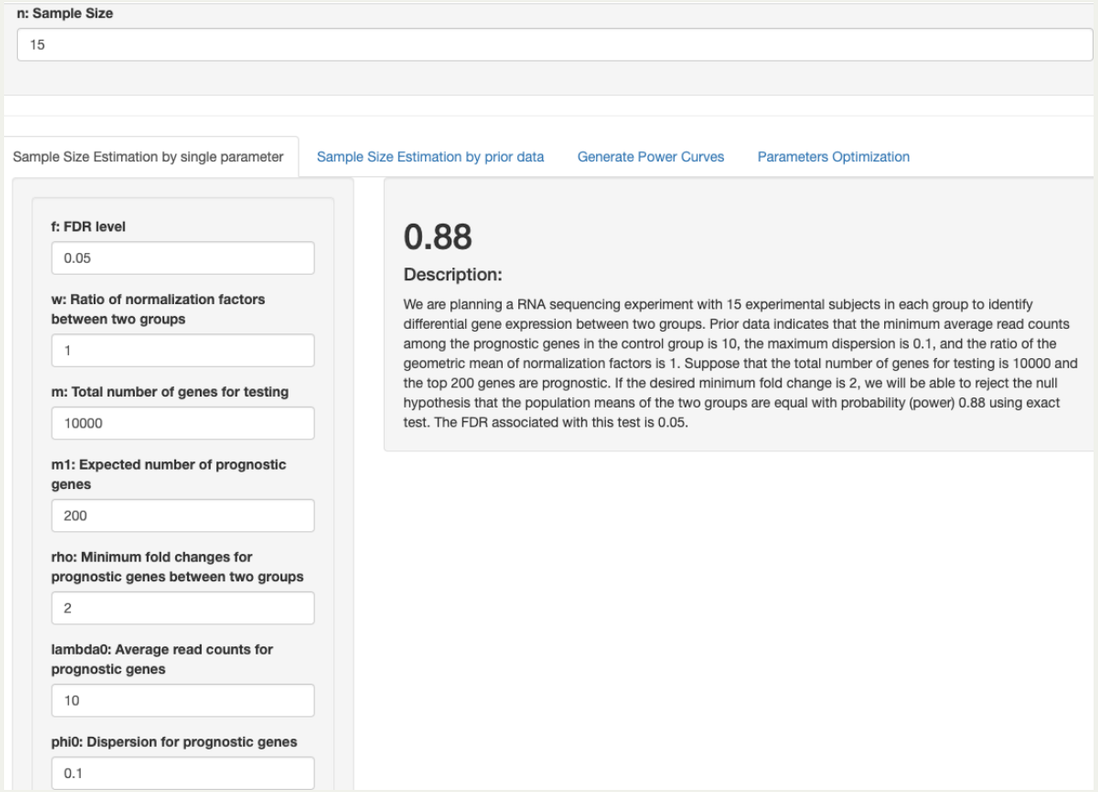
```
       n       0.8
 [1,]  3 0.0000000
 [2,]  4 0.0000000
 [3,]  5 0.0056349
 [4,]  6 0.0953837
 [5,]  7 0.2332468
 [6,]  8 0.3697507
 [7,]  9 0.4895573
 [8,] 10 0.5898277
 [9,] 11 0.6719140
[10,] 12 0.7383110
[11,] 13 0.7916434
[12,] 14 0.8343262
[13,] 15 0.8683667
[14,] 16 0.8954845
[15,] 17 0.9170581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

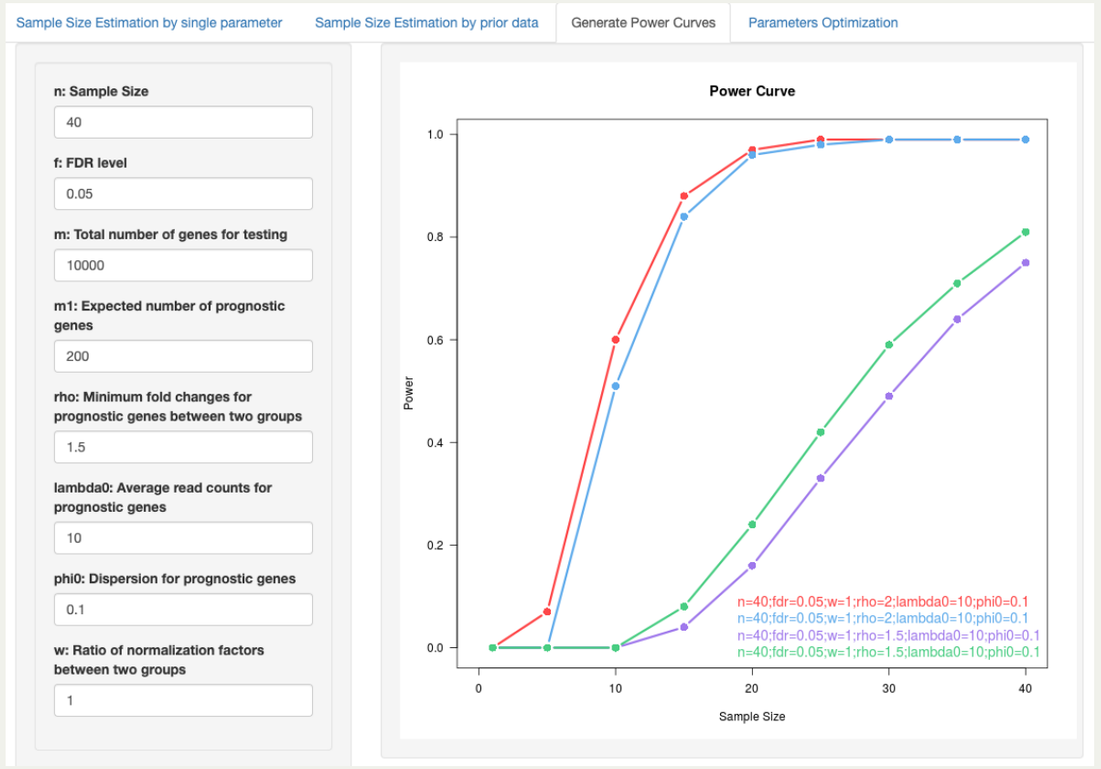
Typical inputs

- 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





https://bit.ly/berd_pss_rnaseq

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 `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")
# BiocManager::install("edgeR") # cf. https://bioconductor.org/packages/release/bioc/html/edge R.html
# BiocManager::install("Biobase") # cf. https://bioconductor.org/packages/release/bioc/html/Biob ase.html

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
```

```r
trt <- hammer.eset$protocol[which(hammer.eset$Time=="2 weeks")]

```

https://bit.ly/berd_pss_rnaseq

```r
## After generating count data with column names of control and treatment, you may estimate the parameters of mu and disp using the f
## mu: average read count in the control group
## The following apply function averages the count values for each gene in the control group.
mu <- apply(counts[, trt == "control"], 1, mean)
summary(mu)
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    0.0      9.5    146.5    952.0    662.5 386345.0
```

```r
## 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
summary(disp)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.05566 0.06991 0.11093 0.13949 0.18397 1.01143
```

```r
size_out <- ssizeRNA_vary(nGenes = 10000,
                          mu = mu, # estimated above
                          disp = disp, # estimated above
                          fc = 2,
                          up = 0.5,
                          fdr = 0.05,
                          power = 0.8,
                          maxN = 35)
```

```r
# 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     9 0.8434739
```

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

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

```
 [4,]  6 0.6056247
 [5,]  7 0.7184589
 [6,]  8 0.7927465
 [7,]  9 0.8434739
 [8,] 10 0.8792088
 [9,] 11 0.9050462
[10,] 12 0.9241417
[11,] 13 0.9385395
[12,] 14 0.9495731
[13,] 15 0.9581587
[14,] 16 0.9649286
[15,] 17 0.9703306
[16,] 18 0.9746886
[17,] 19 0.9782380
[18,] 20 0.9811584
[19,] 21 0.9835790
[20,] 22 0.9856014
[21,] 23 0.9873035
```

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

https://bit.ly/berd_pss_rnaseq

Single-cell RNA-seq or spatial transcriptomics

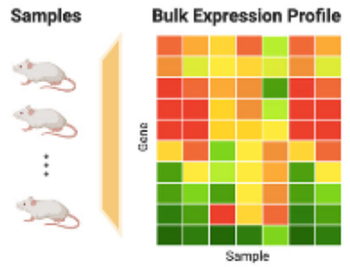
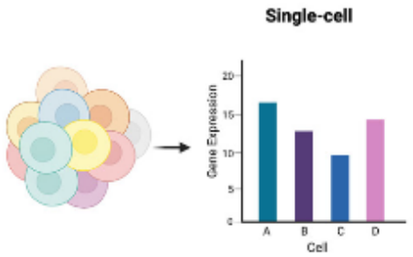
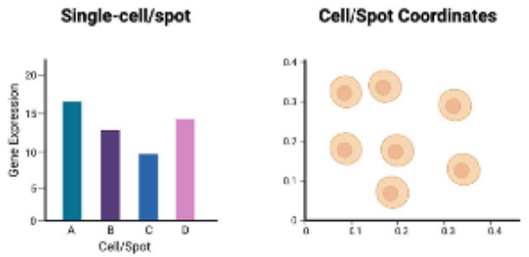
	Bulk RNA-Seq	Single-cell RNA-seq	High-throughput Spatial Transcriptomics
Level			
Data Structure	Subject x Gene Expression Count Data	Cell x Gene Expression Count Data	Cell/Spot x Gene Expression Count Data Cell/Spot 2-dimensional Coordinates
Detection Target	Differentially Expressed Genes	Differentially Expressed Genes Cell Sub-populations	Spatially Variable Genes Tissue Architecture Cell-Cell Communication

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.

https://bit.ly/berd_pss_rnaseq

scRNA-seq

https://bit.ly/berd_pss_rnaseq

- 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).

Detection Target	# of Samples	Tool Name	Experimental Factor	Software	Model	Power Assessment
Cell sub-population	Single sample	'SCOPIT' [37]	(1)	R package & Web application	Multinomial	Analytical
		'howmanycells'		Web application	Negative Binomial	
	Multi sample	'Sensei' [38]	(1), (2)		Beta Binomial	
		'scPOST' [39]		R package	Linear mixed model	Simulation-based
DEG		'scPower' [40]	(1), (2), (3)	R package & Web server	Negative Binomial	Pseudobulk
		'hierarchicell' [41]		R package		Simulation-based
	Single sample	'powsimR' [42]	(1)			
		'POWSC' [43]	(1), (3)		A mixture of zero-inflated Poisson and log-normal Poisson distributions	
		'scDesign' [44]			Gamma-Normal mixture model	

https://bit.ly/berd_pss_rnaseq

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

https://bit.ly/berd_pss_rnaseq

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

Conesa, Ana, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, et al. 2016. "A Survey of Best Practices for RNA-Seq Data Analysis." *Genome Biology* 17 (1): 1–19.

Griffith, Malachi, Jason R Walker, Nicholas C Spies, Benjamin J Ainscough, and Obi L Griffith. 2015. "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud." *PLoS Computational Biology* 11 (8): e1004393.

https://bit.ly/berd_pss_rnaseq

Jeon, Hyeongseon, Juan Xie, Yeseul Jeon, Kyeong Joo Jung, Arkobrato Gupta, Won Chang, and Dongjun Chung. 2023. “Statistical Power Analysis for Designing Bulk, Single-Cell, and Spatial Transcriptomics Experiments: Review, Tutorial, and Perspectives.” *Biomolecules* 13 (2): 221.

Klaus, Bernd. 2015. “Statistical Relevance—Relevant Statistics, Part i.” *The EMBO Journal* 34 (22): 2727–30.

Lee, Mei-Ling Ting, and George Alex Whitmore. 2002. “Power and Sample Size for DNA Microarray Studies.” *Statistics in Medicine* 21 (23): 3543–70.

Liu, Yuwen, Jie Zhou, and Kevin P White. 2014. “RNA-Seq Differential Expression Studies: More Sequence or More Replication?” *Bioinformatics* 30 (3): 301–4.

Mistry, Meeta, Mary Piper, Jihe Liu, and Radhika Khetani. 2021. “hbctraining/DGE_workshop_salmon_online: Differential Gene Expression Workshop Lessons from HCBC (first release).” Zenodo.

Poplawski, Alicia, and Harald Binder. 2018. “Feasibility of Sample Size Calculation for RNA-Seq Studies.” *Briefings in Bioinformatics* 19 (4): 713–20.

https://bit.ly/berd_pss_rnaseq