Workshop

GENCODE Annotation

How many transcripts are there for Tnnt2?

```
grep 'Tnnt2' gencode.gtf | cut -f9 | grep ENSMUST | cut -f2,8 -d ';' | sort | uniq |wc
```

How many transcripts are there for Tnnt2?

```
grep 'Tnnt2' gencode.gtf | cut -f9 | grep ENSMUST | grep  'transcript_type "protein_coding"' | cut -f2,8 -d ';' | sort | uniq | wc
```

GENCODE Annotation

How many lines in the GTF file?

```
wc -l gencode.gtf
```

How would you view the first 5 lines of the file?

```
head –n 5 gencode.gtf
```

What are the first 5 lines of the file?

##description: evidence-based annotation of the mouse genome (GRCm38), version M10 (Ensembl 85)
##provider: GENCODE
##contact: gencode-help@sanger.ac.uk
##format: gtf
##date: 2016-07-19

Why might this information be important?

Indicates the version of the annotation use and when it was generated.

RNA-seq Analysis

Sequence RNA transcripts (usually cDNA) to understand how gene regulation varies in different cell types, genetic backgrounds, or conditions.

Gene/Transcript Quantification

The number of reads that align to each gene is an estimate of how many RNA transcripts were present in the sample for that gene.

Problems:

- Paralogs (reads don't always map uniquely)
- Reads that span 2 or more exons
- Sampling error

http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Raw Read Counts per Gene/Transcript

Gene	Control	Experimental Condition
Gene A	27	27
Gene B	90	270
Gene C	280	640
Gene D	1003	3021
Gene E	3100	3342

Wow! Gene E is expressed at a 3X higher rate than Gene D!

Oooh! The experimental condition caused genes B,C,D, and E to be overexpressed!

FeatureCounts is a useful tool for generating counts:
https://www.ncbi.nlm.nih.gov/pubmed/24227677

Raw Read Counts per Gene/Transcript

Gene	Control	Experimental Condition
Gene A	27	27
Gene B	90	270
Gene C	280	640
Gene D	1003	3021
Gene E	3100	3342

Oooh! The experimental condition caused genes B,C, and D to be overexpressed!

Would you reach the same conclusion if you knew that the control experiment used 10 million reads while the experimental condition used 30 million reads?

Possible explanations:

1. 3 times as many transcripts are expressed for Gene E than Gene D.
2. Gene E is 3X as long as Gene D.
3. Gene D and E are the same length and produce the same number of transcripts, but Gene D has a close paralog that has acted as a "sponge" for ⅔ of its alignments.

Gene	Control	Experimental Condition
Gene A	27	27
Gene B	90	270
Gene C	280	640
Gene D	**1003**	3021
Gene E	**3100**	3342

Why we Normalize

- To smooth out technical variations amoung samples:
  - Sequencing depth: genes have more reads in a deeper sequenced library
  - Gene length: longer genes are likely to have more reads that the shorter reads

- **CPM (counts per million):** counts scaled by total number of reads. This method accounts for sequencing depth only.

- **TPM (transcripts per kilobase million):** counts per length of transcript (kb) per million reads mapped. This method accounts for both sequencing depth and gene length.

- **RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped):** similar to TPM, as this method also accounts for both sequencing depth and gene length as well

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control	Experimental Condition
Gene A	1 kb	27	27
Gene B	2 kb	90	270
Gene C	6 kb	280	640
Gene D	30 kb	1003	3021
Gene E	40 kb	3100	3342

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control	Experimental Condition
Gene A	1 kb	27	27
Gene B	2 kb	90	270
Gene C	6 kb	280	640
Gene D	30 kb	1003	3021
Gene E	40 kb	3100	3342
Total (w/ counts for ~19000 other genes)		**20,000,000**	**40,000,000**

Step 1: Normalize (i.e., adjust) gene counts by the total amount of sequences in the experiment

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control (RPM)	Experimental Condition (RPM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	4.5	6.75
Gene C	6 kb	14.0	16.0
Gene D	30 kb	50.15	75.525
Gene E	40 kb	155.0	83.55
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

Step 1: Normalize gene counts the total amount of sequences in the experiment

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control (RPM)	Experimental Condition (RPM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	4.5	6.75
Gene C	6 kb	14.0	16.0
Gene D	30 kb	50.15	75.525
Gene E	40 kb	155.0	83.55
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

Step 2: Normalize gene counts RPM by gene length

RPKM (or FPKM): reads per kilobase of exon model per million reads

Gene	Gene Size	Control (RPKM)	Experimental Condition (RPKM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	2.25	3.375
Gene C	6 kb	2.33	2.67
Gene D	30 kb	1.67	2.52
Gene E	40 kb	3.875	2.09
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

Step 2: Normalize gene counts RPM by gene length

Which Genes is most expressed in each condition

Gene	Gene Size	Control (RPKM)	Experimental Condition (RPKM)
Gene A	1 kb	1.34	0.675
Gene B	2 kb	2.25	3.375
Gene C	6 kb	2.33	2.67
Gene D	30 kb	1.67	2.52
Gene E	40 kb	**3.875**	2.09
Total		**20,000,000**	**40,000,000**
Millions of reads		**20**	**40**

RPKM is best for *within-sample* comparisons of gene expression. TPM is best for inter-sample comparisons

Differential Expression Between Samples

METHOD | OPEN ACCESS

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love, Wolfgang Huber and Simon Anders

Genome Biology 2014 15:550 | DOI: 10.1186/s13059-014-0550-8 | © Love et al.; licensee BioMed Central. 2014
Received: 27 May 2014 | Accepted: 19 November 2014 | Published: 5 December 2014

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]
[1]Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010 and [2]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia

Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments*

Gordon K. Smyth
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3050, Australia

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/pdf/13059_2016_Article_881.pdf

Why you need replicates?

A true technical replicate repeats sample and library prep.

Library prep → sequencing → Technical replicate #1

Library prep → sequencing → Technical replicate #2

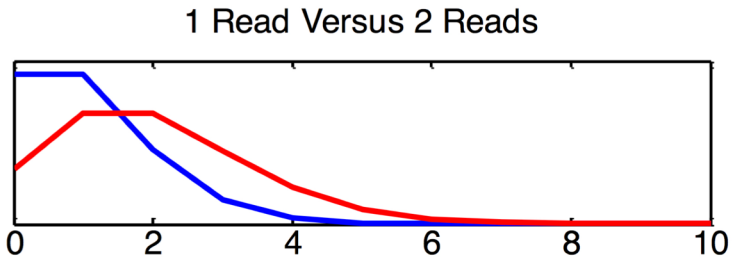
Library prep → sequencing → Biological replicate #1

Library prep → sequencing → Biological replicate #1

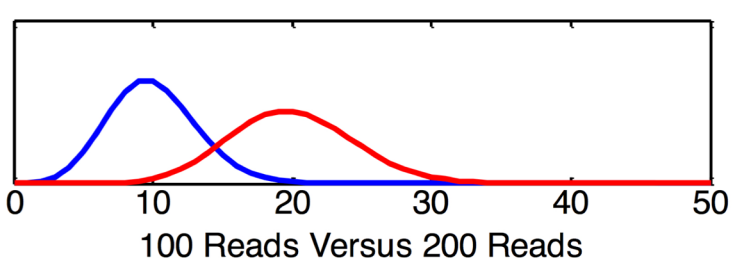
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Source of Variation. Poisson (counting) noise

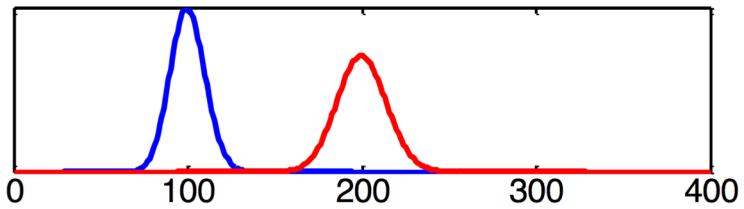
Gene A



Gene B



Gene C

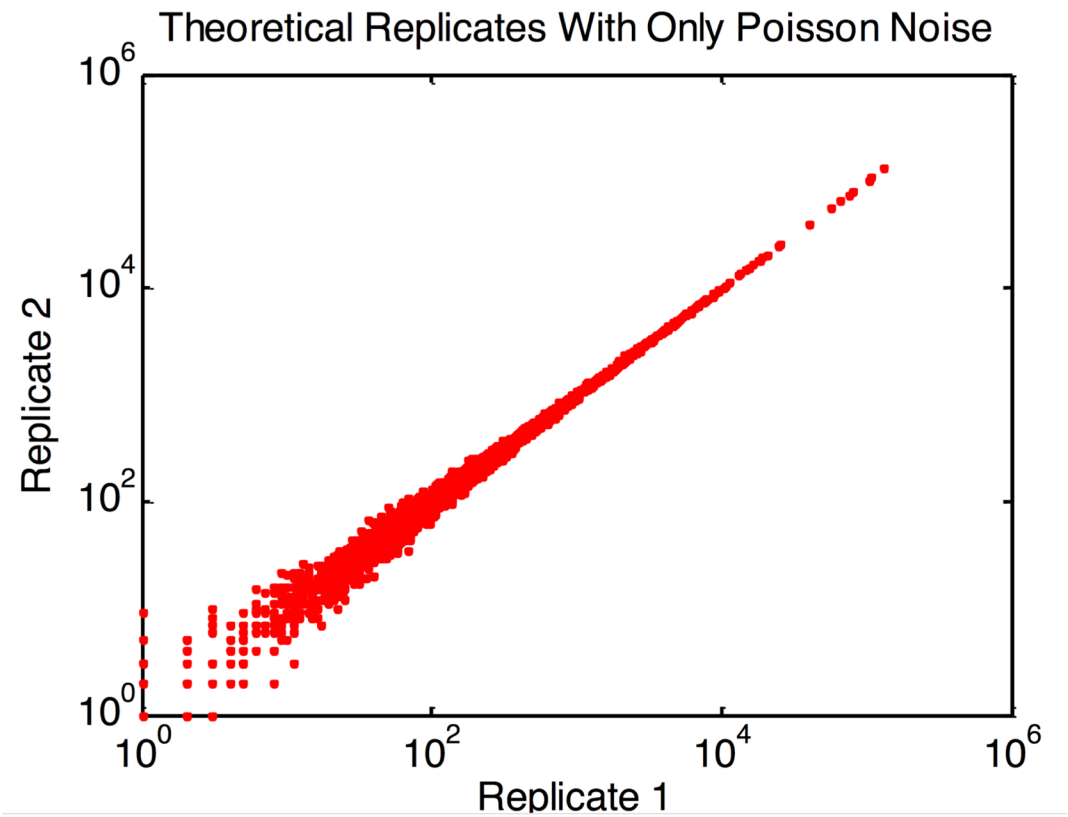


In each case, there is an apparent 2X difference in the mean read counts between control and experimental.

Yet poisson variance is higher relative to the total count when counts are low versus when they are high. For example, the difference in expression of a gene measured with one read versus two reads is inherently less certain than the differences in expression of a gene measured with 100 reads versus 200 reads, even though both differences are nominally a 2X fold change.

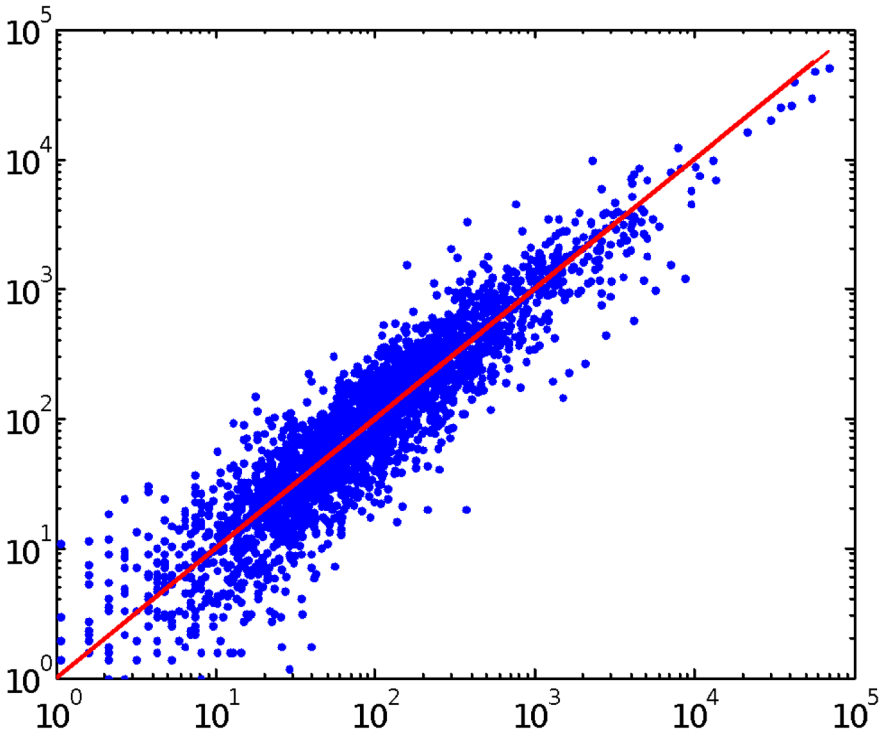
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Source of Variation. Poisson (counting) noise



http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Source of Variation. Non-Poisson technical noise



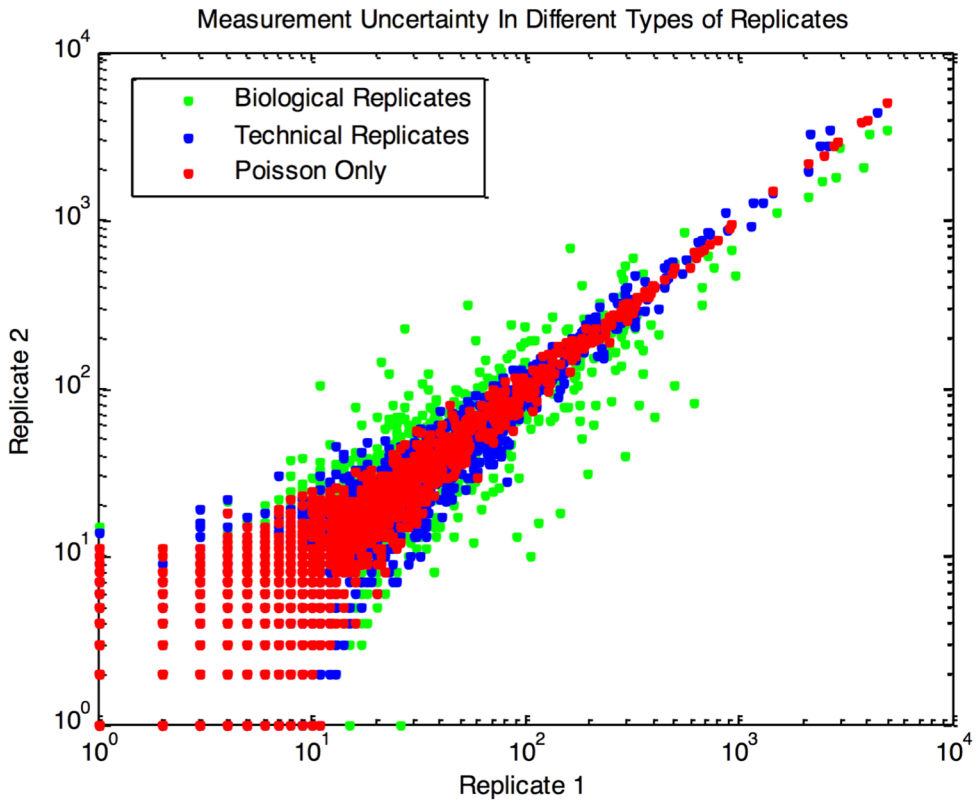
"Non-Poisson Technical Variance is measurement imprecision that stems from the inability of RNA-Seq measurements to measure expression perfectly.

This imprecision is seen when expression from the same sample is measured twice. The expression measurements will not match exactly, and have error greater than what is expected from Poisson noise alone.

Sources of measurement imprecision may include PCR amplification errors during library preparation or machine errors."
-- Michele Busby

http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Sources of variance. Biological Variance



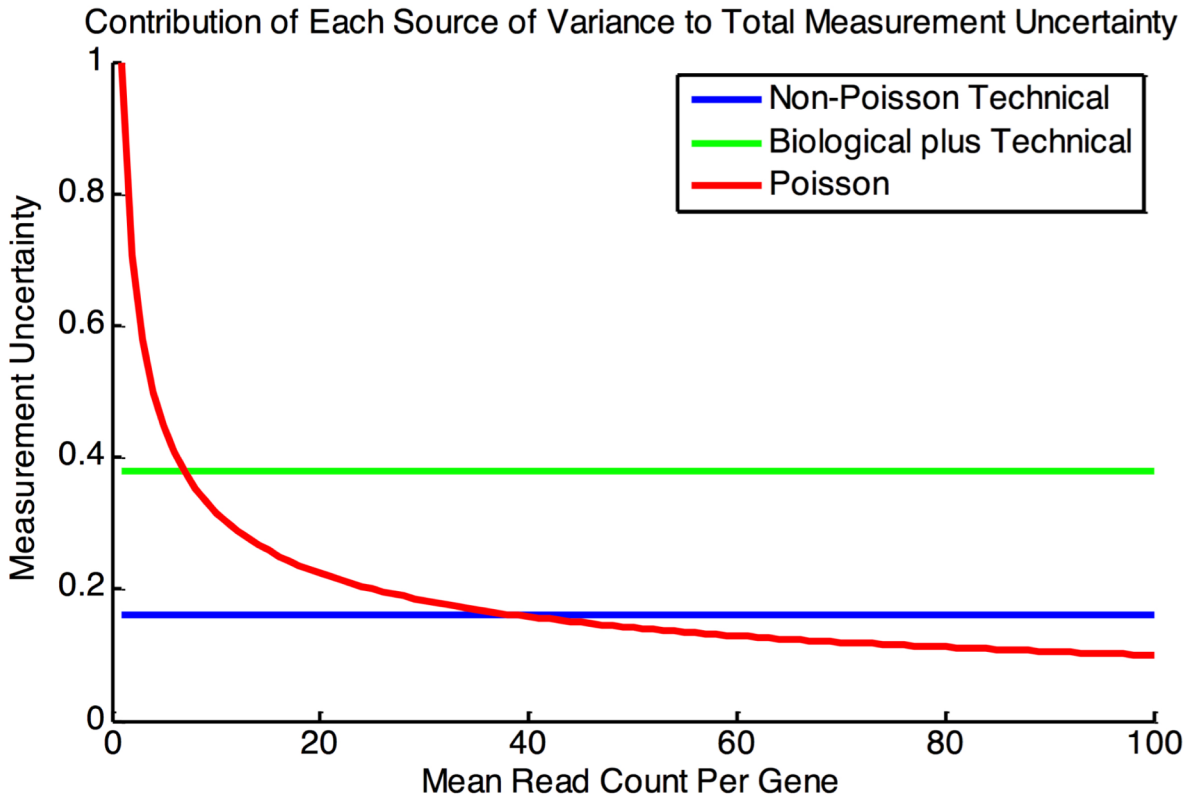
Biological variance is variance that naturally occurs within the samples under investigation. This variance stems from the fact that the expression of any given gene is likely to naturally fluctuate within the cells themselves, and between samples of the same condition. Sources of biological variance include genetic differences among samples and gene expression responses to the environment.

Green dots: biological replicates from S. cerevisiae

Blue dots: technical replicates from S. cerevisiae

Red dots: simulated replicates with only Poisson noise

Relative contribution of sources of variance



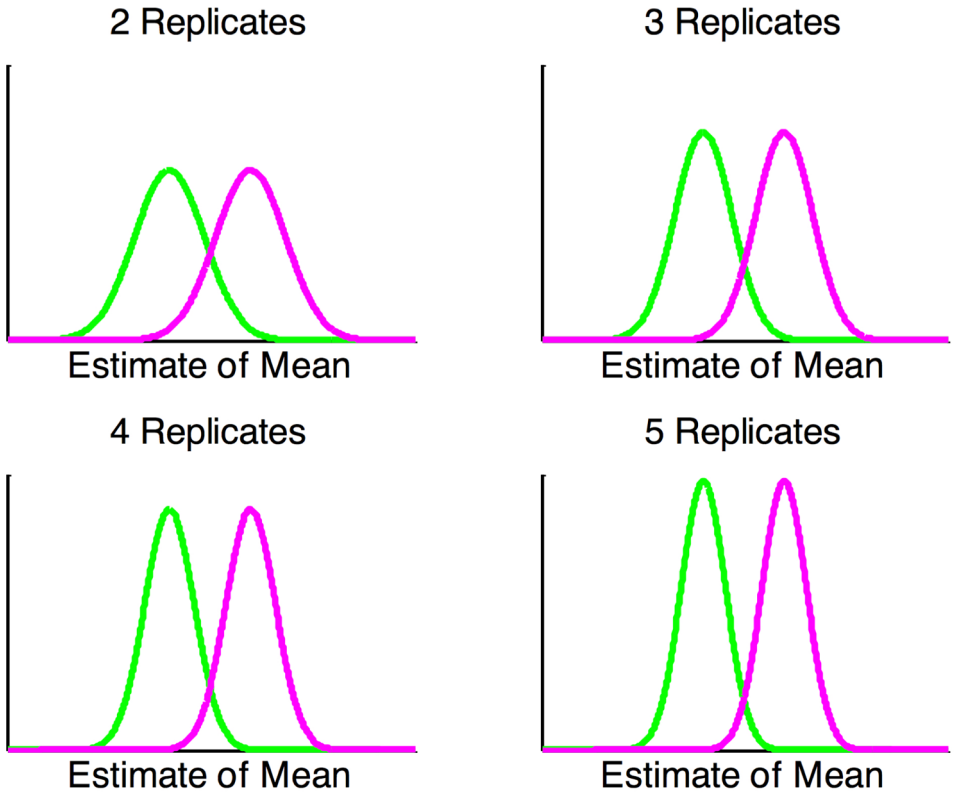
How to decrease uncertainty? Increase Reads

	Replicate 1	Replicate 2	Replicate 3	Mean
Control	12	14	19	15
Test	12	41	7	20

	Replicate 1	Replicate 2	Replicate 3	Mean
Control	22	26	38	30
Test	23	82	15	41

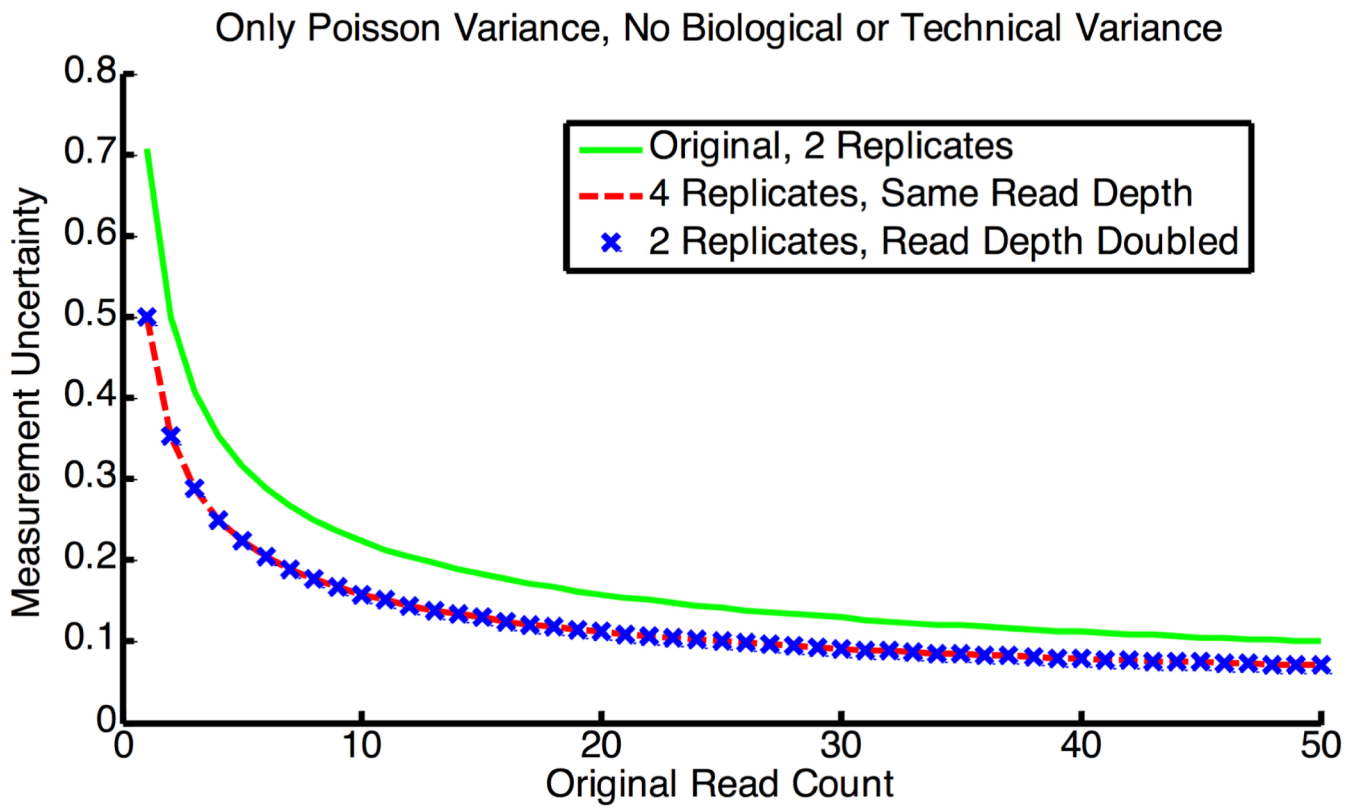
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

How to decrease uncertainty? Increase Replicates



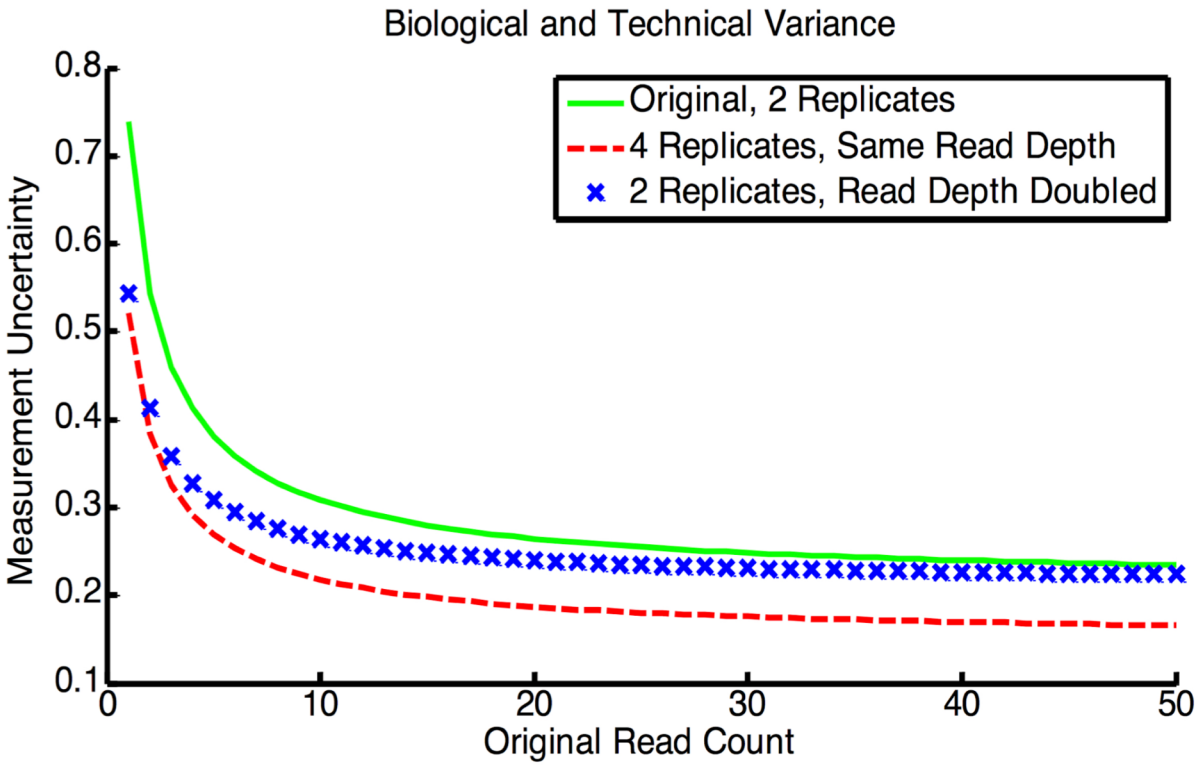
http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Which is better (Depth or Replicates)?



http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Which is better (Depth or Replicates)?



http://michelebusby.tumblr.com/post/26913184737/thinking-about-designing-rna-seq-experiments-to

Correcting for Multiple Testing

1. If you test 30,000 genes for differential gene expression, and you use a significance cut off of $p<0.05$, then you should expect to call approximately 1500 (i.e., 5% of 30000) genes to exhibit differential expressed solely random chance.
2. Thus, if your list of differentially expressed genes at $p<0.05$ is about 1500 genes long, then either there are no genes differentially expressed between the two conditions, or your experiment is underpowered.

False Discovery Rate

It is informative to report a False Discovery Rate (FDR) as well as a p-value. This value is the expected proportion of false positives among all of the significant results (100% in the previous example).

Usually we strive for a FDR at 5% or lower. Importantly, the FDR can only be calculated after an experiment is run, because it requires knowing how many genes were called differentially expressed.

Workshop