RNA seq: differential expression analysis

For INF-BIO 4121/9121
Fall semester 2016

Rebekah Oomen / Monica Hongrø Solbakken
r.a.oomen@ibv.uio.no / m.h.solbakken@ibv.uio.no

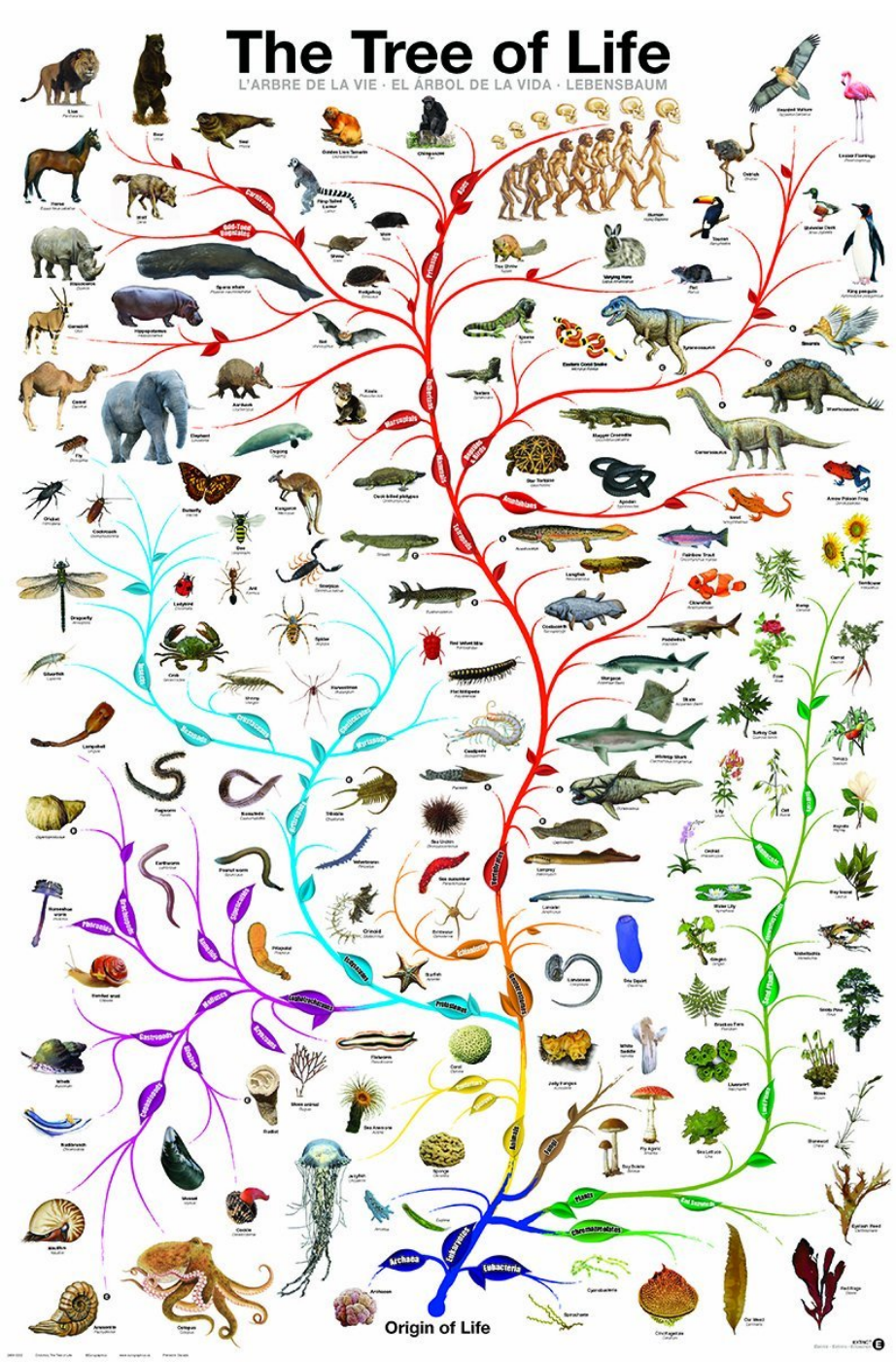
UiO : Centre for Ecological and Evolutionary Synthesis
University of Oslo

Aims I

- You should be able to tell us:
  - Overall:
    - What is RNAseq?
    - Are there different kinds of RNAseq and what separates them?
    - What is RNAseq used for?
  - In more depth:
    - How to design a RNAseq experiment for differential expression analysis
      - How to chose analysis strategy
      - Pitfalls in differential expression analysis (sequencing depth, batch effects, statistical approach etc.)

Aims II

- You should be able to perform:
  - A reference based differential gene expression analysis with a pair-wise comparison involving several biological replicates
  - Present the overall statistics from that analysis (mapping percentage, variance, potential outliers, number of differentially expressed genes etc.)
  - Extract and present the main biological result(s) based on the annotation of the differentially expressed genes



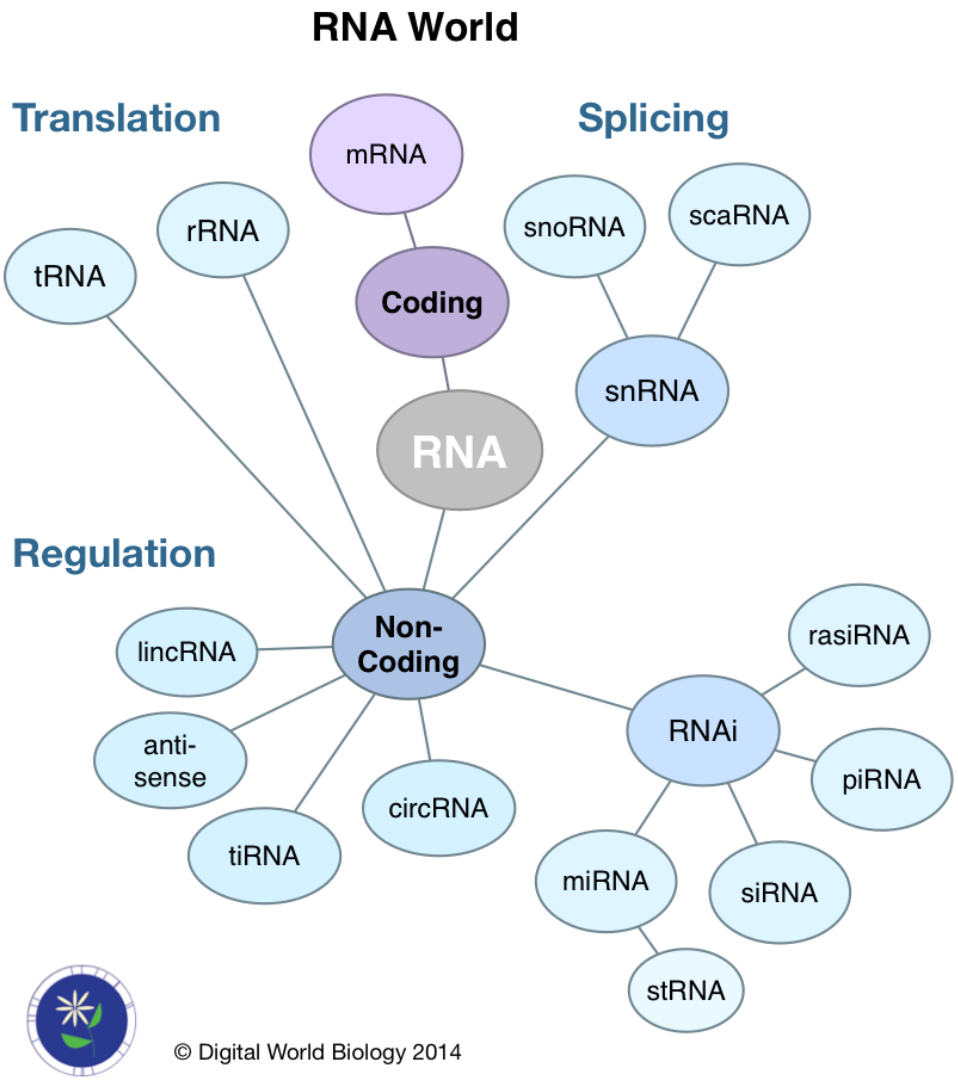
Non-model species

The tree is the limit...

https://www.thinglink.com/scene/645083259847311362

Transcription – all the RNAs

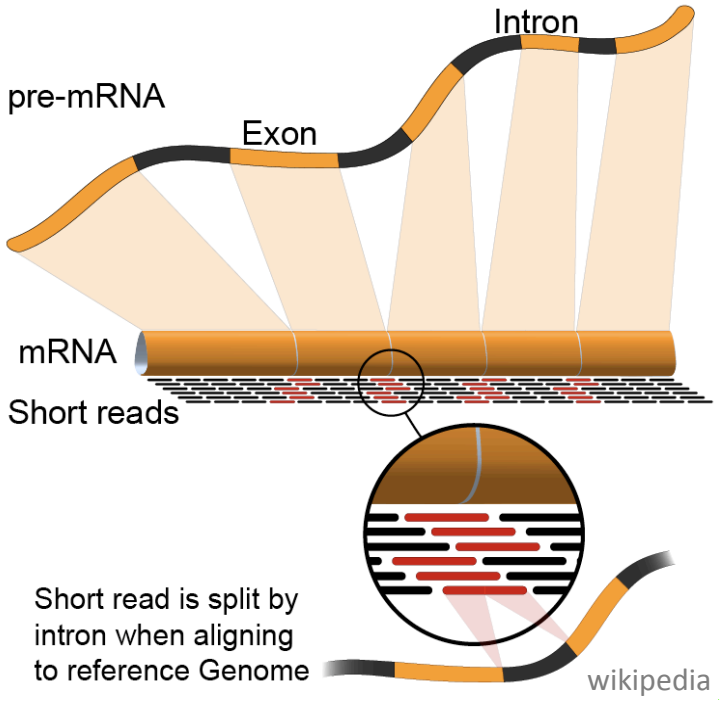
A transcriptome is a snapshot in time of **all RNAs** present in a sample isolated from a given cell, tissue or organism



Next generation transcriptomics

RNA sequencing

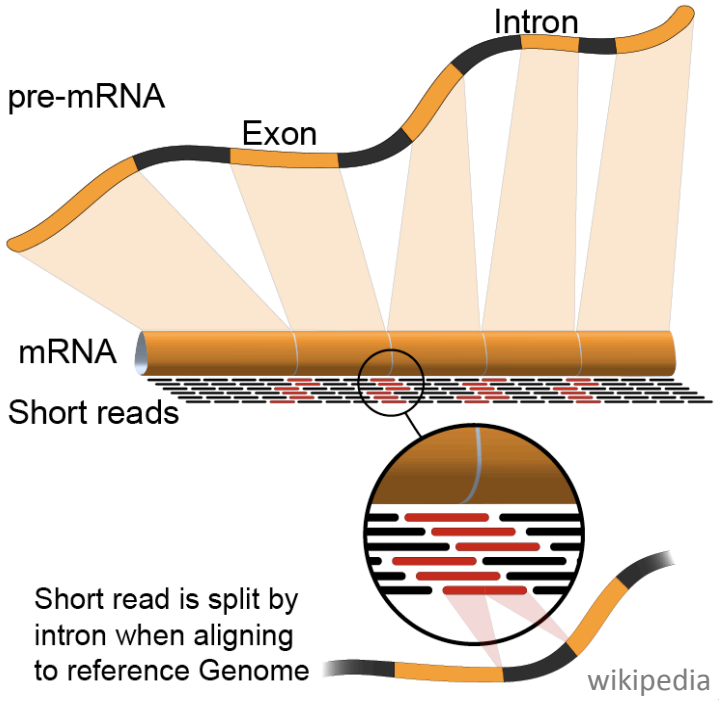
- Transcriptome and expression in one go
- No need for gene sequence information
- High throughput
- Can be outsourced
- Costly, but effective
- Expression results relative to all transcripts



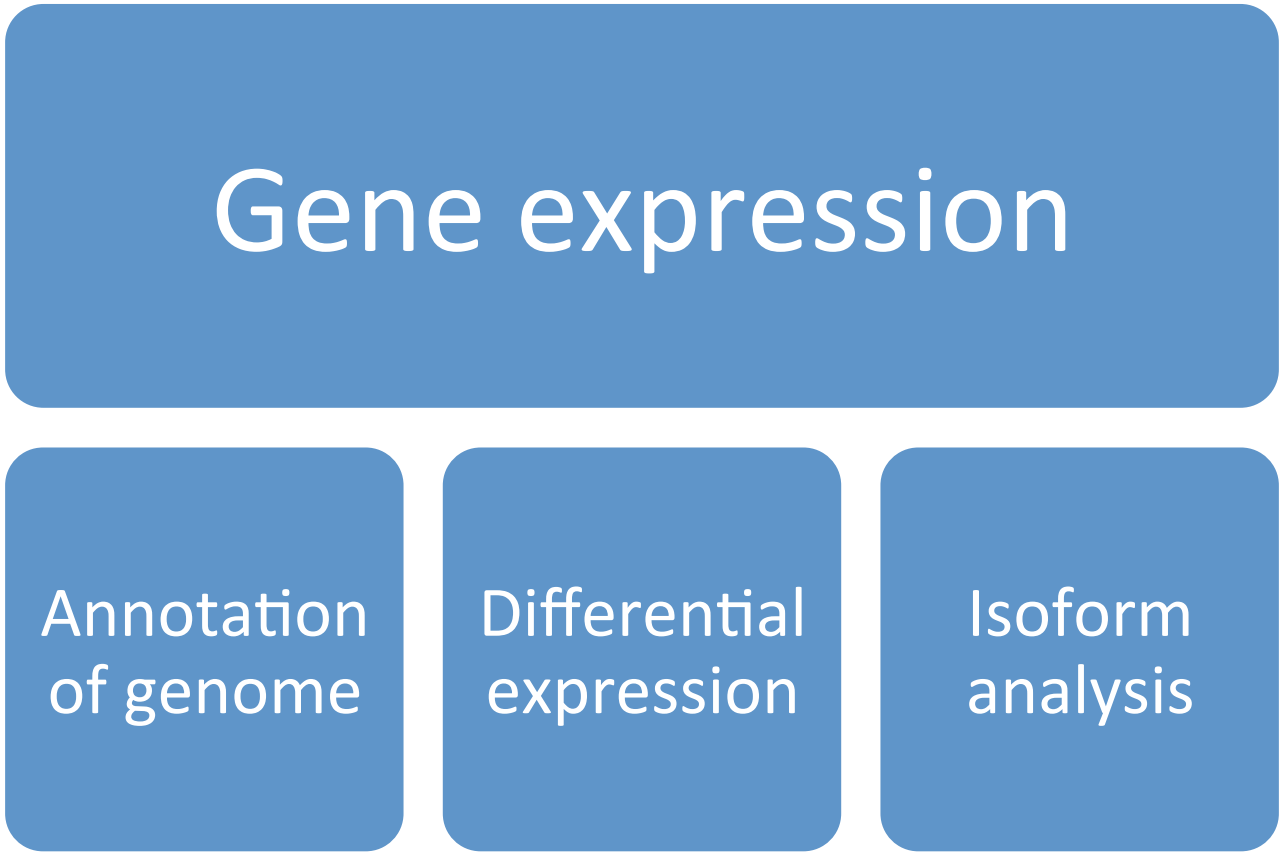
Next generation transcriptomics

RNA sequencing

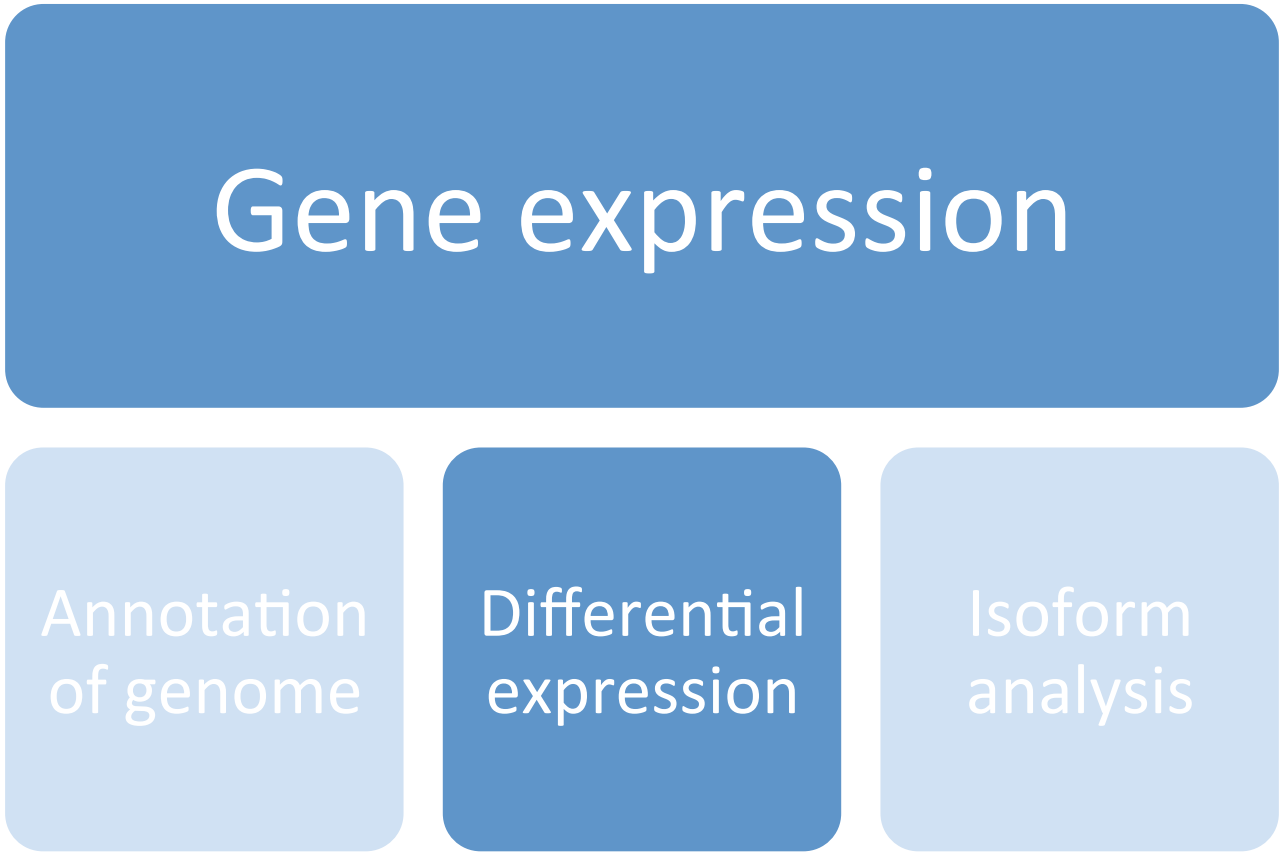
- Transcriptome and expression in one go
- No need for gene sequence information
- High throughput
- Can be outsourced
- Costly, but effective
- Expression results **relative** to all transcripts



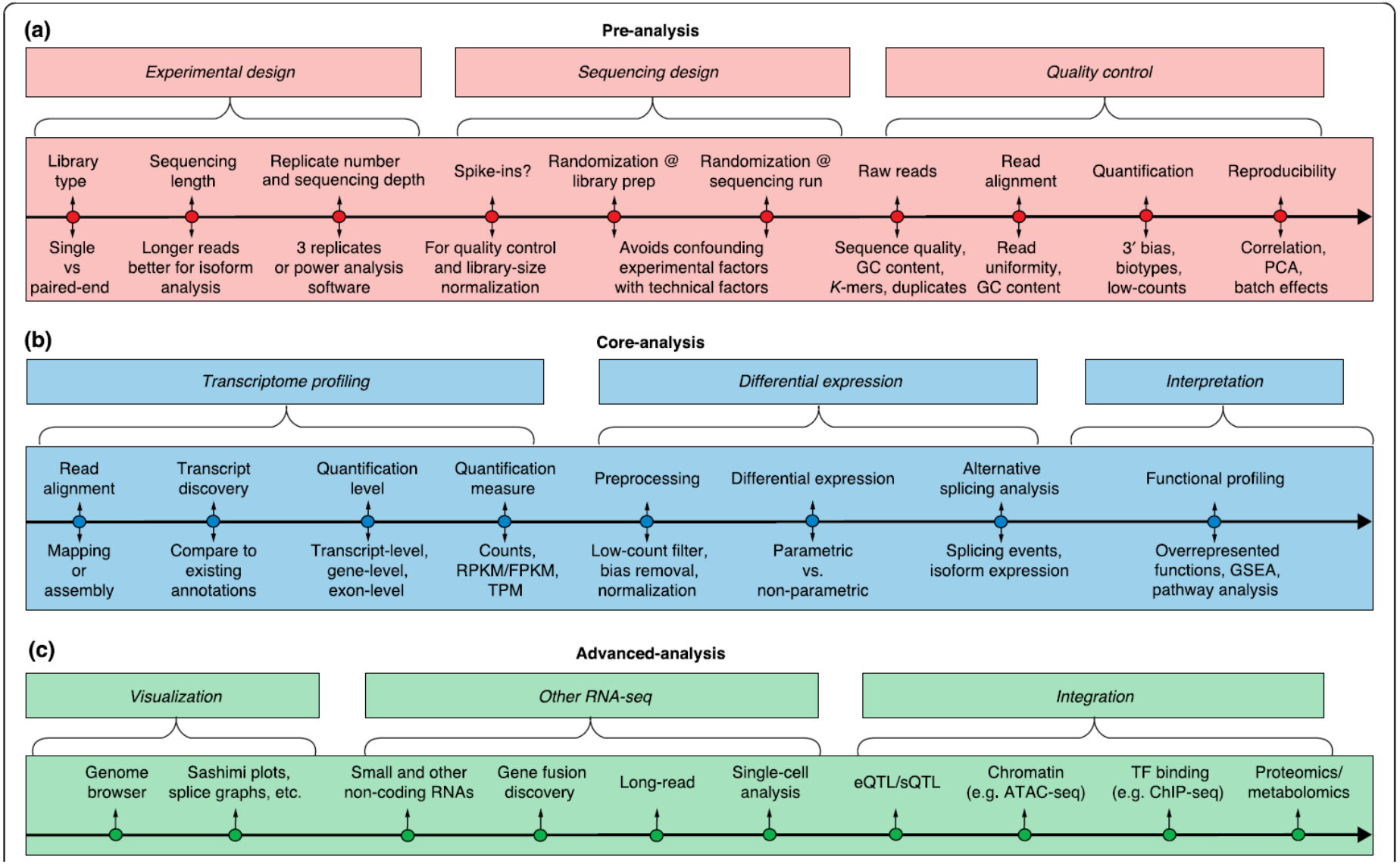
Uses of RNAseq



Uses of RNAseq



Generic pipeline



Experimental design recap

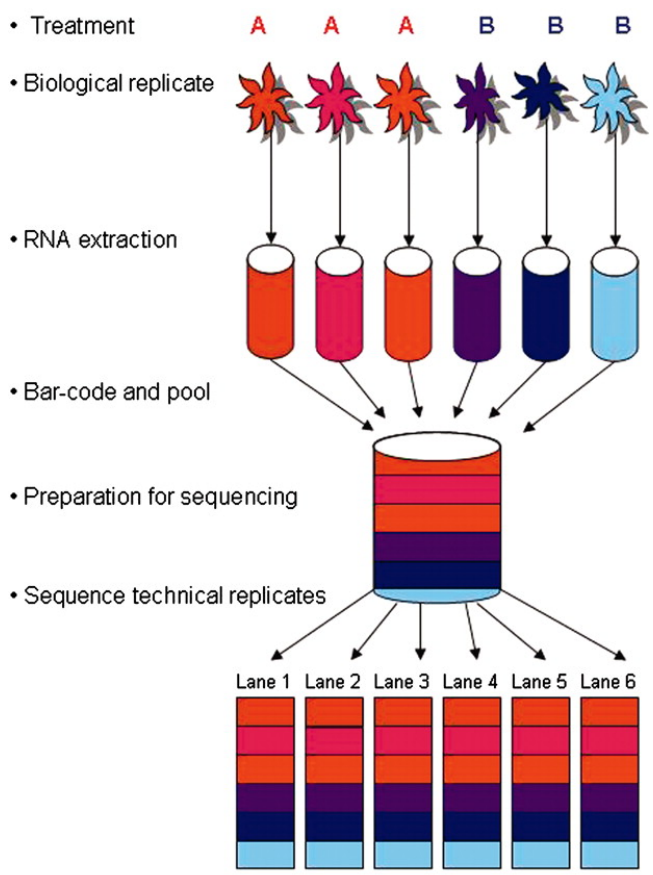
Experimental design considerations

- Research question/study species
  - Qualitative/quantitative?
  - Narrow/broad scope?
  - Model/nonmodel?
- Sources of variation
  1) Poisson counting error
  2) Technical variance
     - E.g. random sampling noise, PCR biases, lane biases
  3) Biological variance
     - Including that of interest!

"There's a flaw in your experimental design. All the mice are scorpios."

Controlling for variation

- Randomization and blocking
  - Of the factor of interest, among "nuisance" factors
- Replication
  - Technical reps = usually no longer necessary
  - Biological reps = extremely important, improves all variance estimates



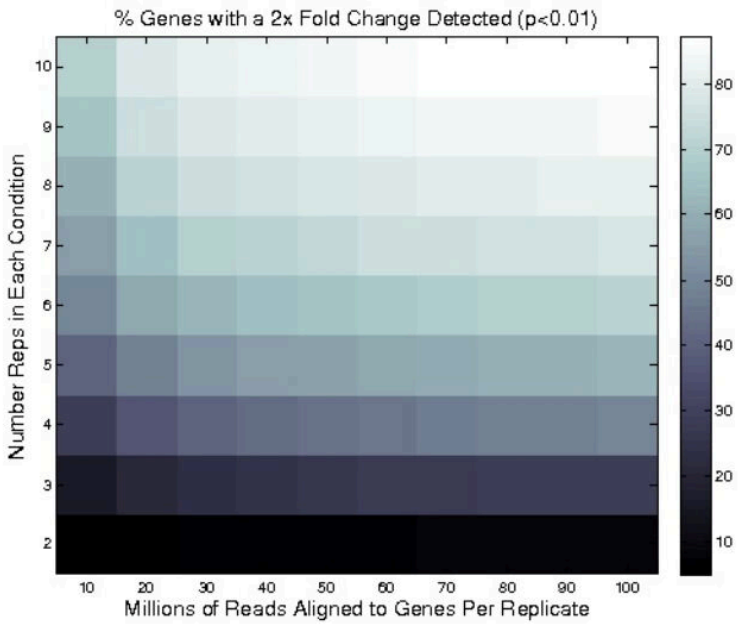
Auer & Doerge 2010

Sequencing depth

- Reduces Poisson noise and random sampling error, improving detection for lowly expressed transcripts and low fold-changes BUT:
  - Benefits plateau at an average of ~10 mapped reads per transcript (~5-20 million mapped reads)

Scotty

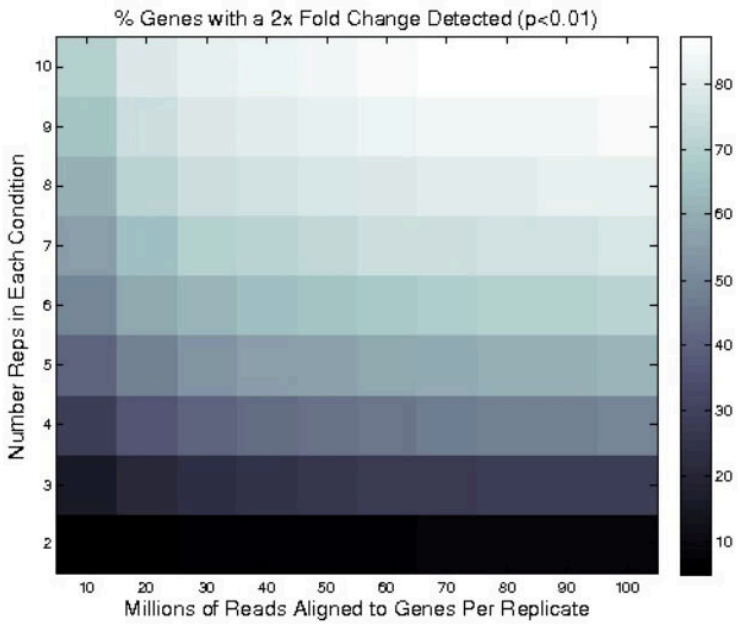
- Once minimum depth is achieved, allocate resources ($) to biological replicates



Model data

Scotty

- Once minimum depth is achieved, allocate resources ($) to biological replicates

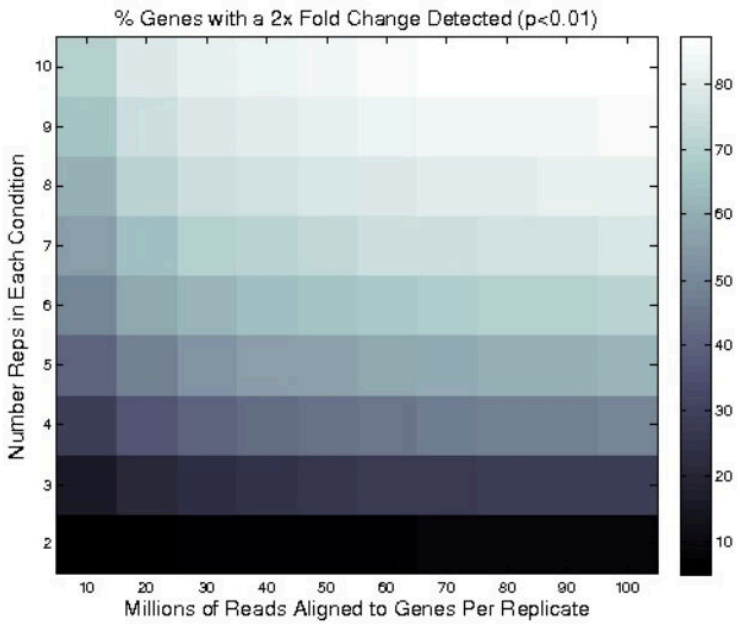


Model data

  - Power increases plateau around 10-20 M reads

Scotty

- Once minimum depth is achieved, allocate resources ($) to biological replicates



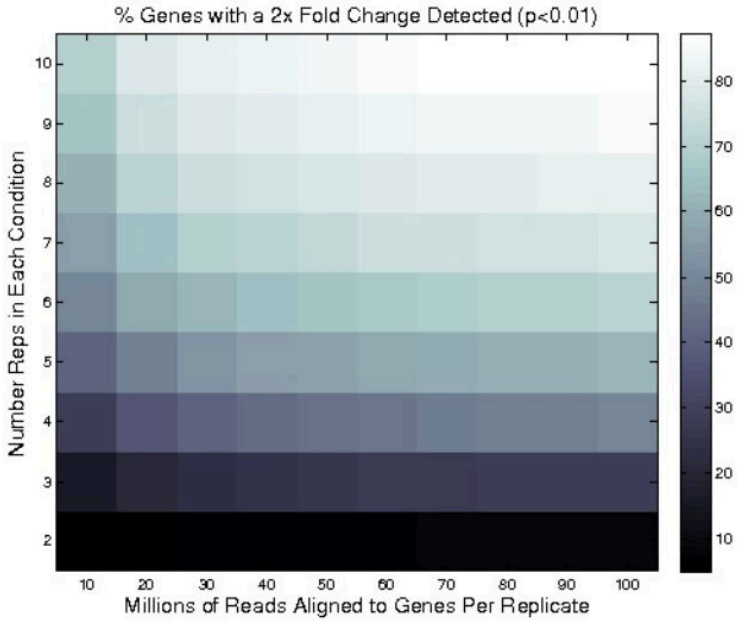
Model data

  - Power increases plateau around 10-20 M reads
  - Power increases more steadily with replicates

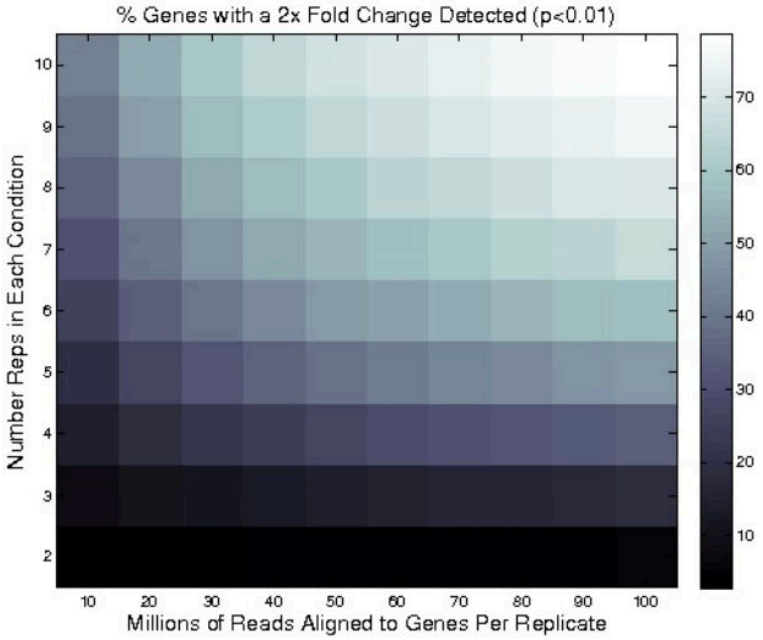
http://scotty.genetics.utah.edu

Scotty

- Pilot experiments improve experimental design by quantifying biological variation in study system



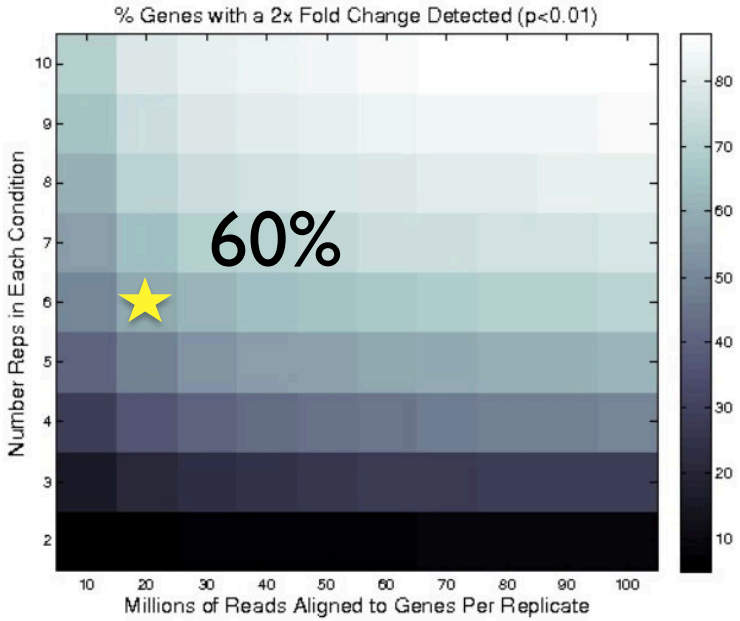
Model data



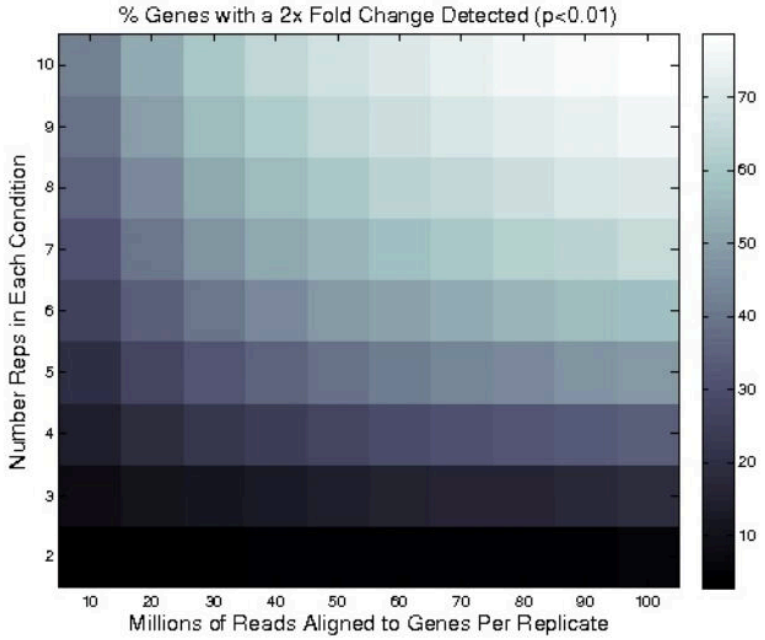
Nonmodel (pilot) data

Scotty

- Pilot experiments improve experimental design by quantifying biological variation in study system



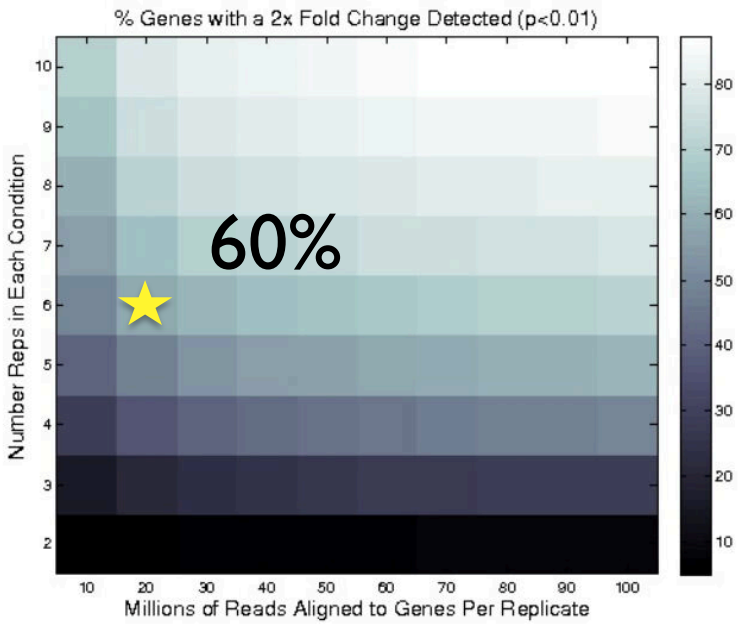
Model data



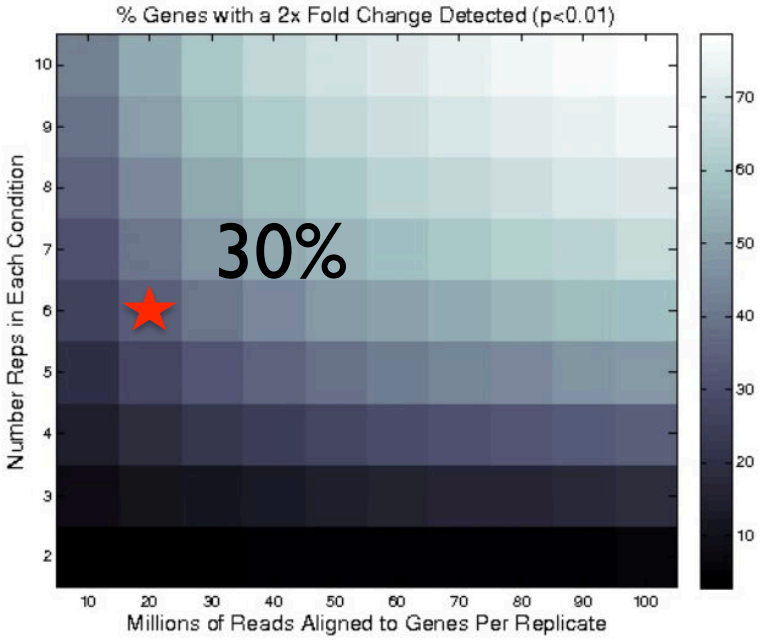
Nonmodel (pilot) data

Scotty

- Pilot experiments improve experimental design by quantifying biological variation in study system



Model data

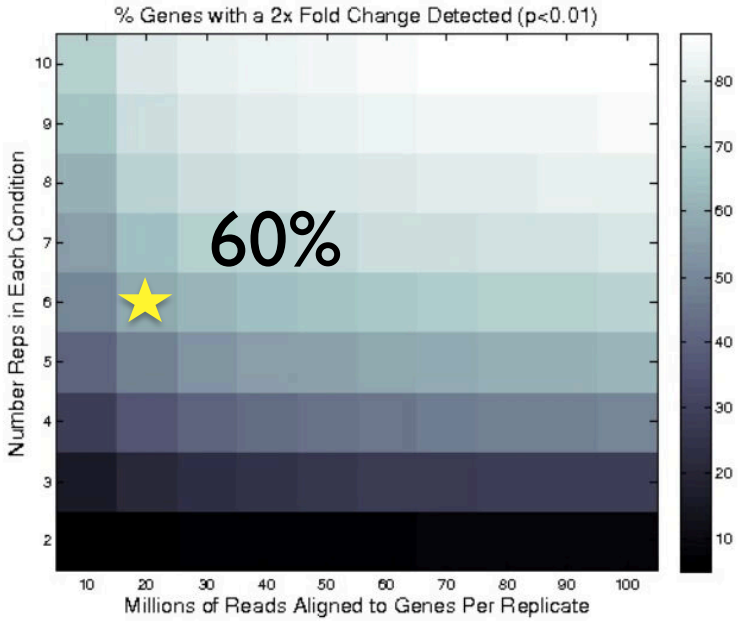


Nonmodel (pilot) data

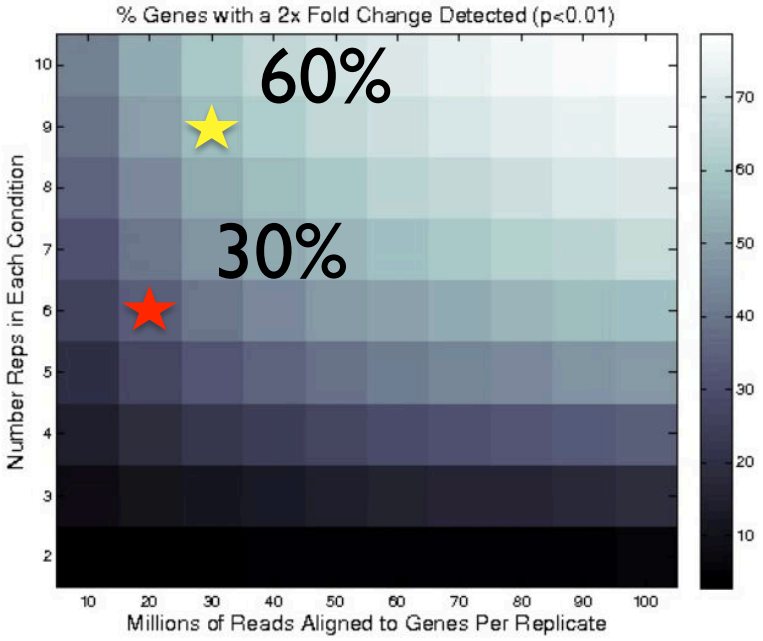
http://scotty.genetics.utah.edu

Scotty

- Pilot experiments improve experimental design by quantifying biological variation in study system



Model data



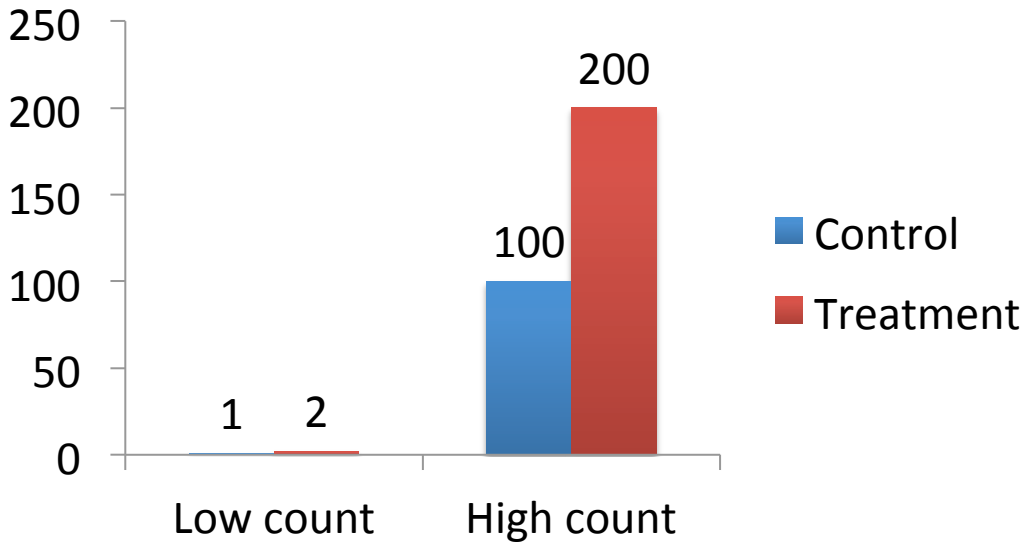
Nonmodel (pilot) data

http://scotty.genetics.utah.edu

DE statistics recap

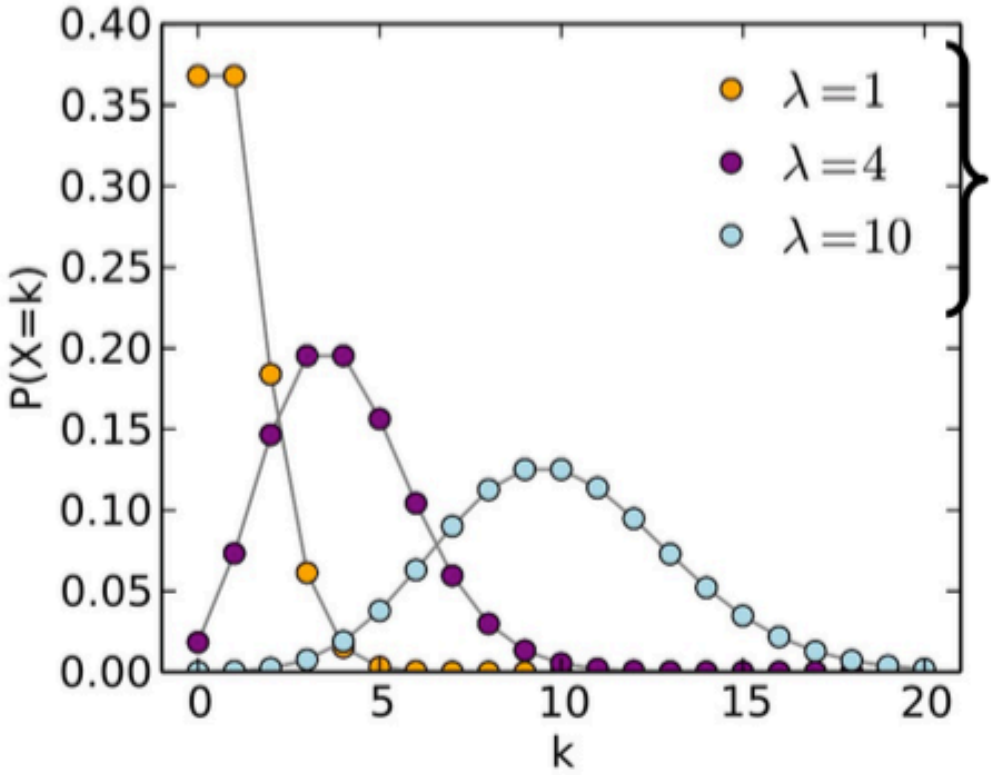
Poisson counting error

- Uncertainty in count-based measurements
- Disproportionately large for low-count data



The poisson distribution

The counts of technical replicates follow a **poisson distribution** (Marioni et al 2008). The Poisson distribution can be applied to systems with a large number of possible events, each of which is rare.



From Wikipedia. Can be 3 different genes, each with their own poisson distribution. Lambda is the mean of the gene's distribution, with a certain number of reads.

Y=axis: chance to pick that number of reads.

Poisson statistics

- Mean = count
- Variance = mean = count
- Standard deviation = sqrt(count)
  - Absolute error estimation
- Coefficient of variation = sqrt(count)/count
  - Relative error estimation

Poisson statistics

- Mean = count
- Variance = mean = count
- Standard deviation = sqrt(count)
  - Absolute error estimation
- Coefficient of variation = sqrt(count)/count
  - Relative error estimation

RNA-seq is a relative measurement,

not an absolute one.

Poisson statistics

Count	SD = sqrt(count)	CV = sqrt(count)/count
1	1.00	1.00
2	1.41	0.71
3	1.73	0.58
4	2.00	0.50
5	2.24	0.45
6	2.45	0.41
7	2.65	0.38
8	2.83	0.35
9	3.00	0.33
10	3.16	0.32

Poisson statistics

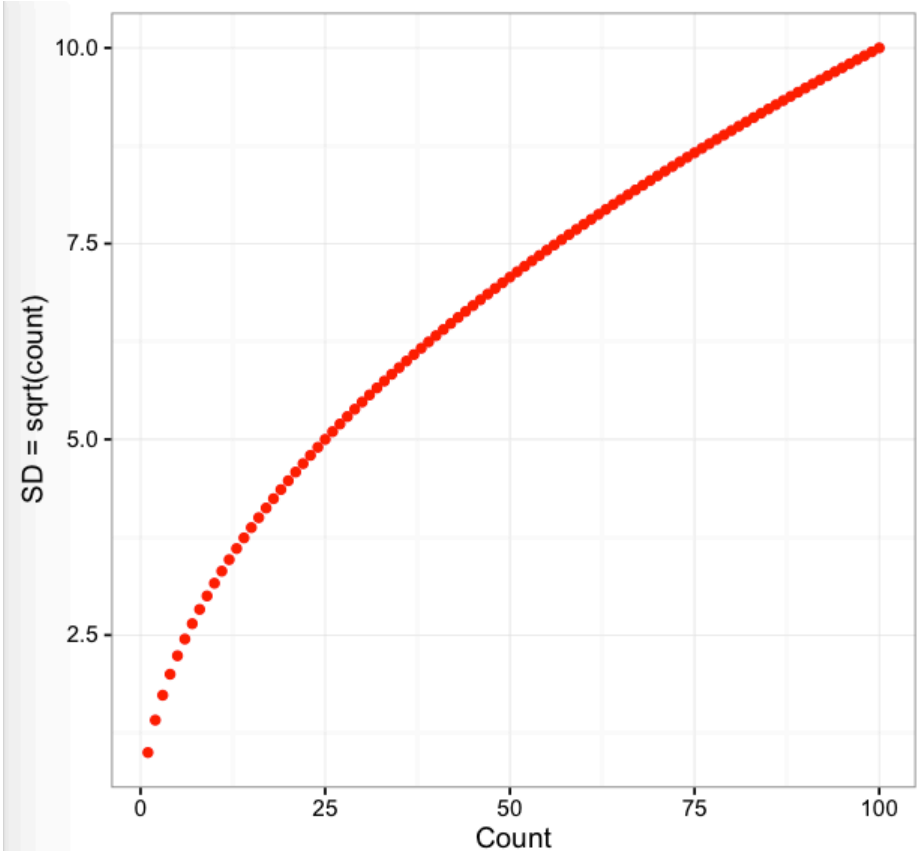
Count	SD = sqrt(count)	CV = sqrt(count)/count
1	1.00	1.00
2	1.41	0.71
3	1.73	0.58
4	2.00	0.50
5	2.24	0.45
6	2.45	0.41
7	2.65	0.38
8	2.83	0.35
9	3.00	0.33
10	3.16	0.32

Poisson statistics

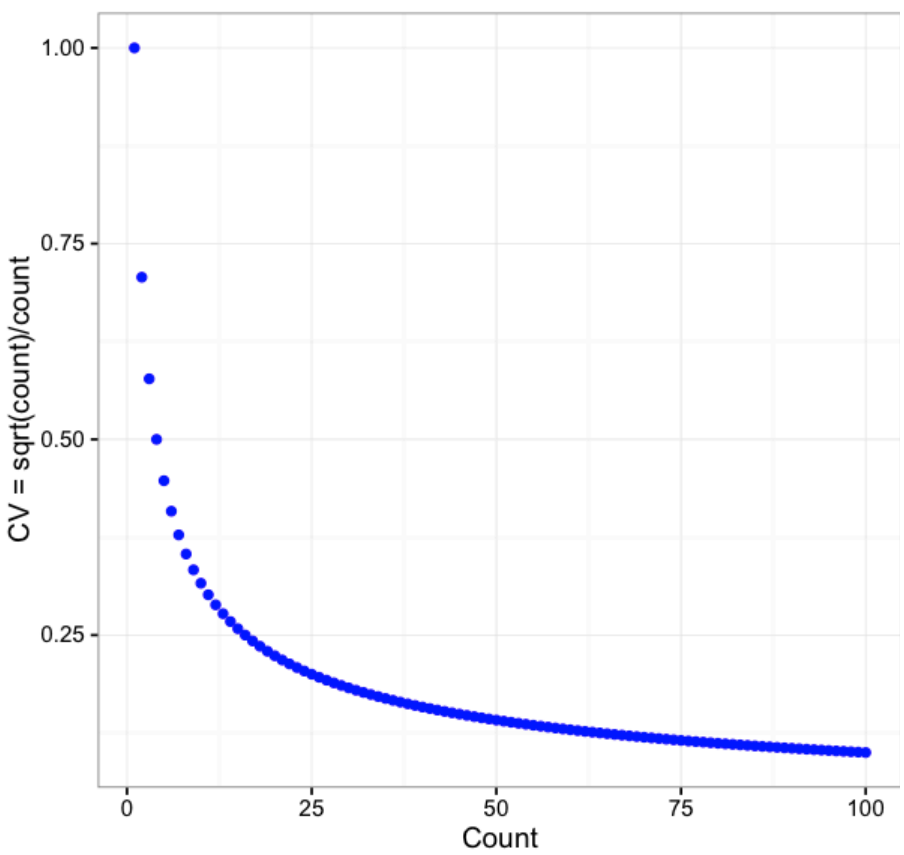
Count	SD = sqrt(count)	CV = sqrt(count)/count
1	1.00	1.00
2	1.41	0.71
3	1.73	0.58
4	2.00	0.50
5	2.24	0.45
6	2.45	0.41
7	2.65	0.38
8	2.83	0.35
9	3.00	0.33
10	3.16	0.32

Poisson statistics

Standard deviation



Coefficient of variation



Poisson statistics

- It's the relative error (CV) that's important:

Count	SD	CV	Random sample (n=5)	Fold-change
1	1.00	1.00	1, 0, 2, 1, 0	1, -, 2, 1, -
10	3.16	0.32	5, 8, 4, 13, 9	0.5, 0.8, 0.4, 1.3, 0.9
100	10	0.1	105, 97, 77, 111, 104	1.05, 0.97, 0.77, 1.11, 1.04

Poisson statistics

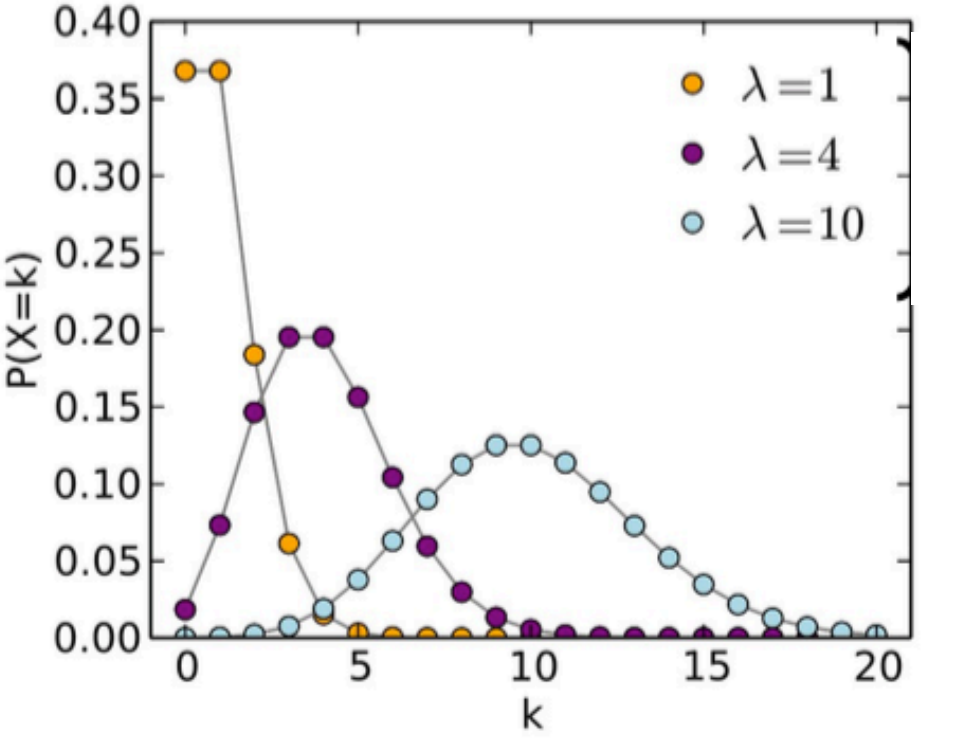
- It's the relative error (CV) that's important:

Count	SD	CV	Random sample (n=5)	Fold-change
1	1.00	1.00	1, 0, 2, 1, 0	1, -, 2, 1, -
10	3.16	0.32	5, 8, 4, 13, 9	0.5, 0.8, 0.4, 1.3, 0.9
100	10	0.1	105, 97, 77, 111, 104	1.05, 0.97, 0.77, 1.11, 1.04

Greater fold-change uncertainty at lower counts!

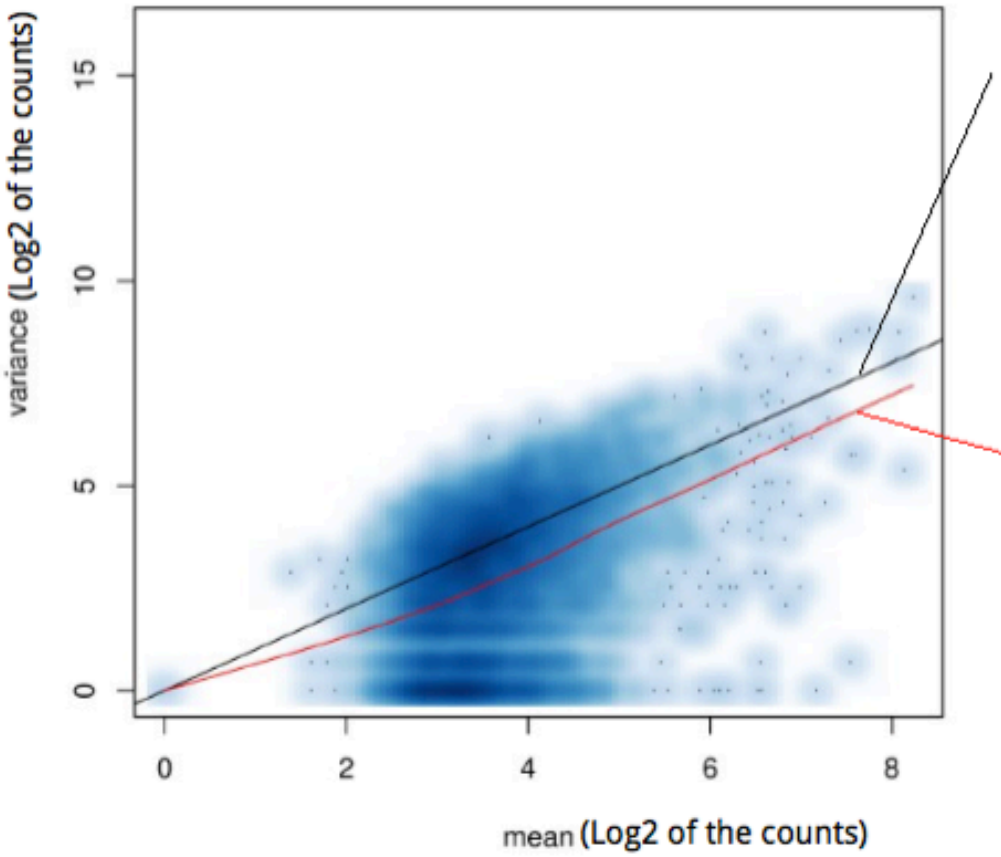
Improving power

- Increasing replication and sequencing depth fill in the distribution, providing a better estimate of the mean
  - For both low and high counts
    - Though the gap in uncertainty remains…



Wikipedia

Comparing technical replicates



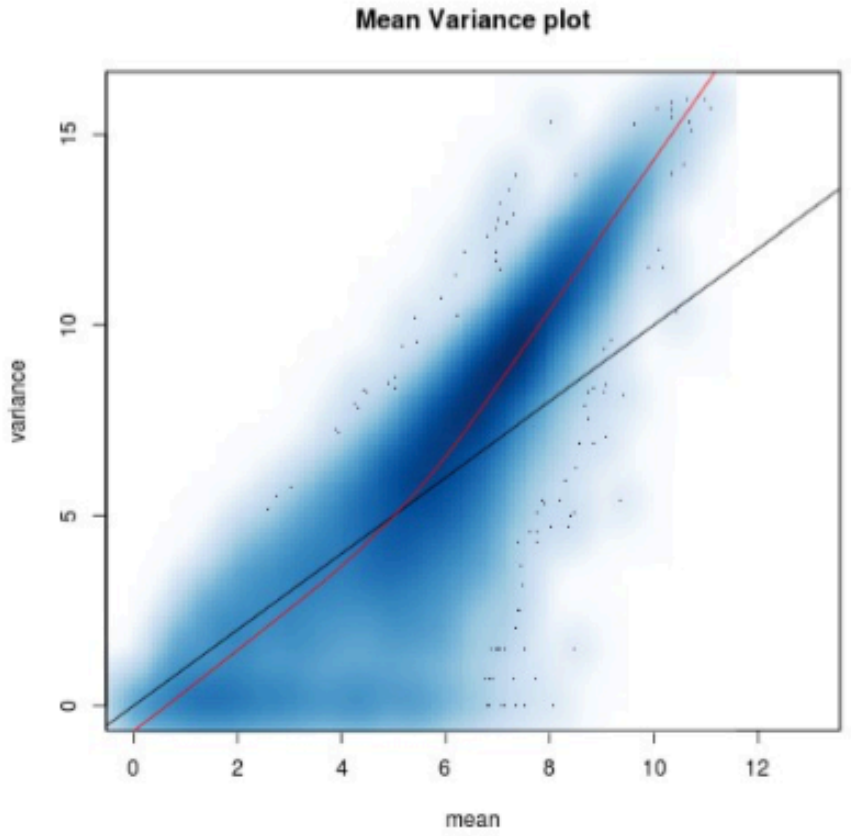
Correlation between mean and variance according to Poisson

Lowess fit through the data

Risso et al. "GC-Content Normalization for RNA-Seq Data"
BMC Bioinformatics 2011, 12:480
http://www.biomedcentral.com/1471-2105/12/480 - EDASeq package (R)

But poisson does not seem to fit

Extending the samples to real biological samples, this mean variance relationship does not hold...

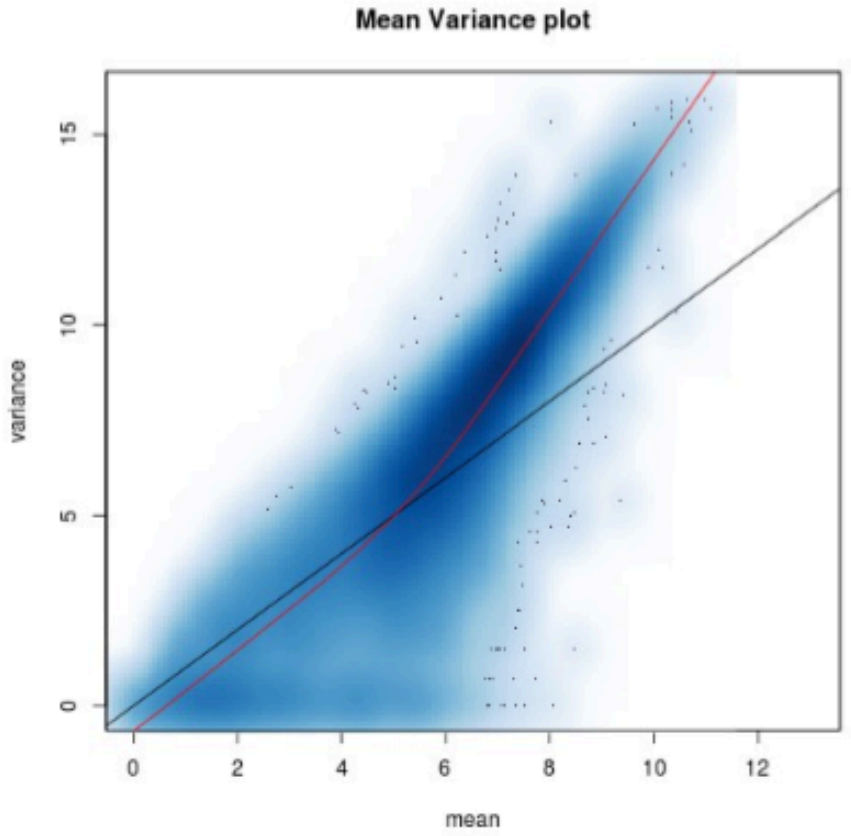


Plotted using EDASeq Package in R.

standard

But poisson does not seem to fit

Extending the samples to real biological samples, this mean variance relationship does not hold...



Plotted using EDASeq Package in R.

standard

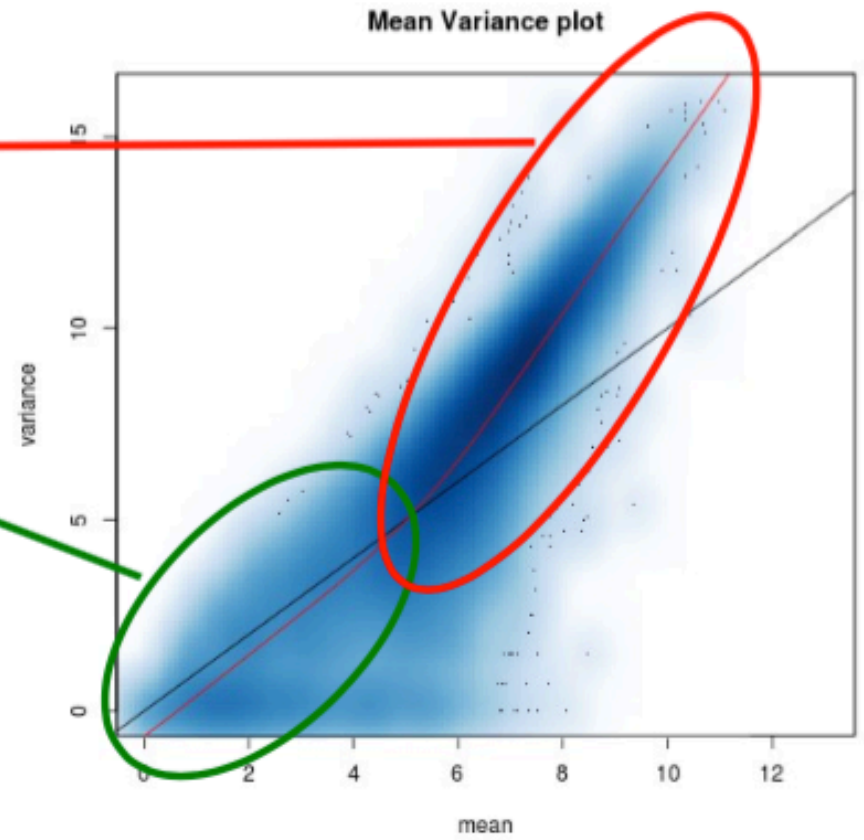
But poisson does not seem to fit

Extending the samples to real biological samples, this mean variance relationship does not hold!

Something is going on!

Reasonable fit

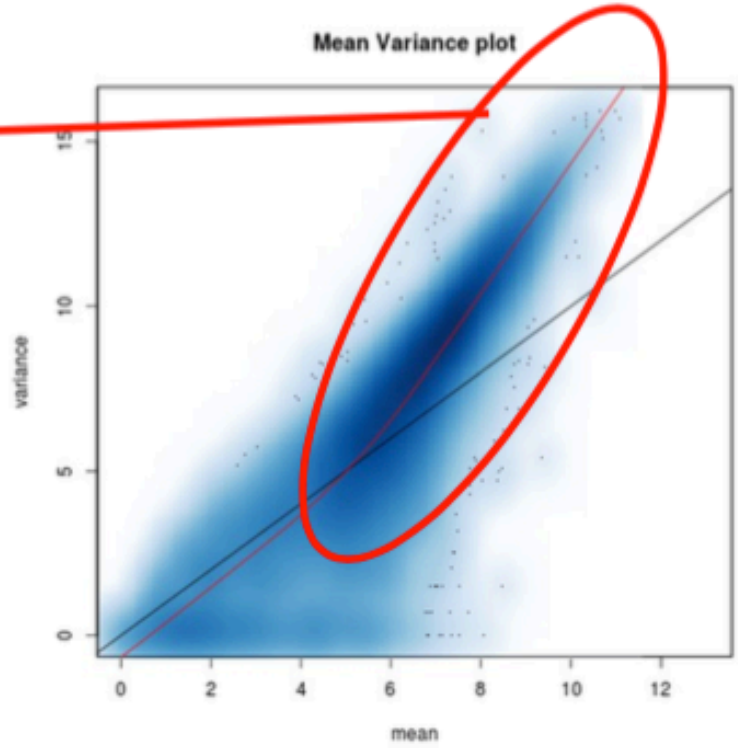
Plotted using EDASeq Package in R.



An extra source of variation

The Poisson distribution has an '**overdispersed**' variance: the variance is bigger than expected for higher counts between biological replicates.

Something is going on!



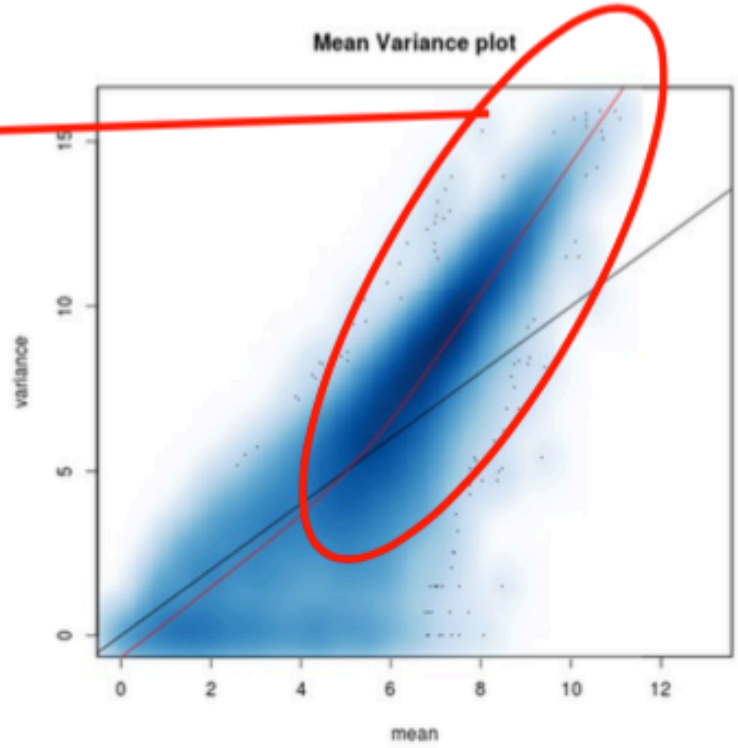
Plotted using EDASeq Package in R.

An extra source of variation

data have

The ~~Poisson distribution has~~ an '**overdispersed**' variance: the variance is bigger than expected for higher counts between biological replicates.

Something is going on!



Plotted using EDASeq Package in R.

An extra source of variation

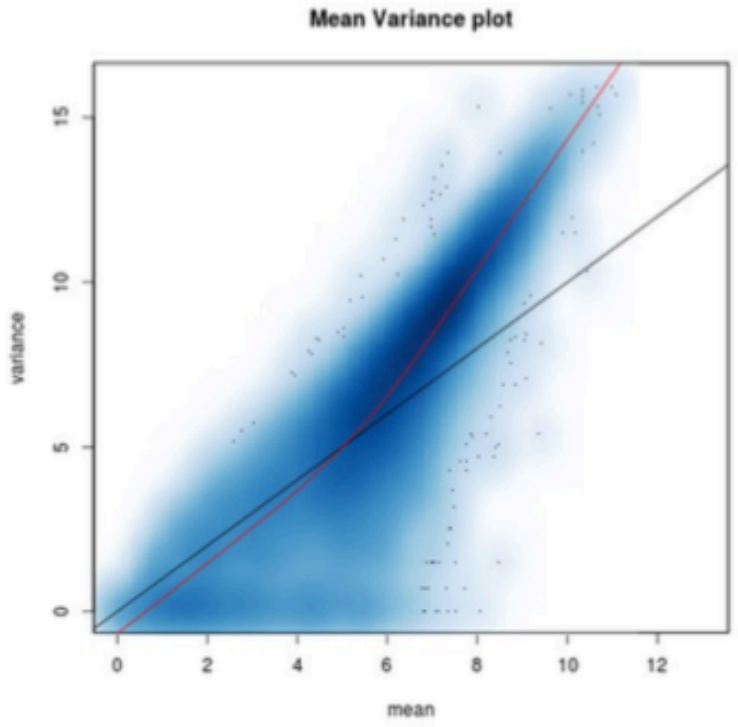
Where Poisson: CV = std dev / mean => $CV^2 = 1/\mu$
If an additional distribution is involved (also dependent on π, the fraction of the gene in the cDNA pool), we have a

mixture of distributions:

$CV^2 = 1/\mu + \varphi$

Low counts!

dispersion



Generalization of Poisson with this extra parameter: the **Negative Binomial Model** fits better!

VIB

An extra source of variation

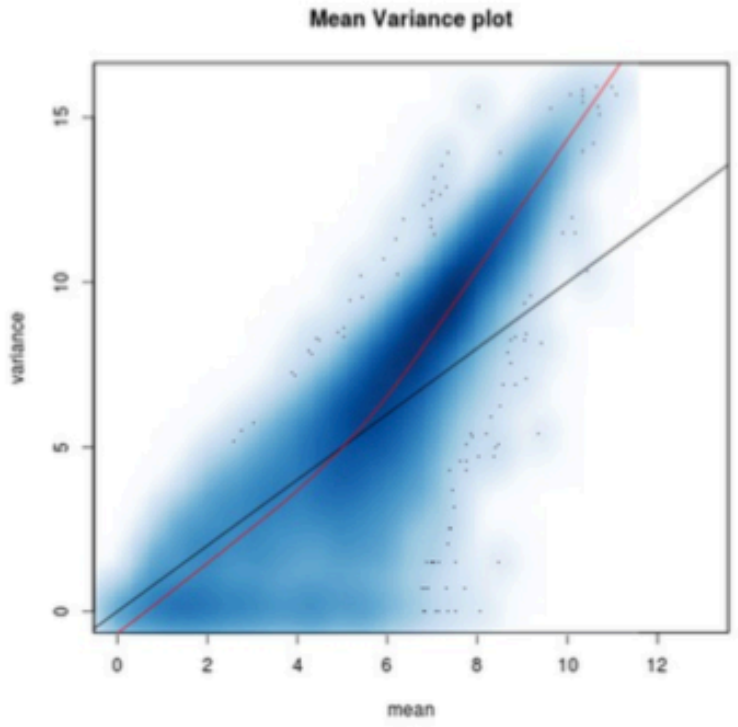
Where Poisson: CV = std dev / mean => $CV^2 = 1/\mu$
If an additional distribution is involved (also dependent on π, the fraction of the gene in the cDNA pool), we have a

mixture of distributions:

$CV^2 = 1/\mu + \varphi$

Low counts!

dispersion



Generalization of Poisson with this extra parameter: the **Negative Binomial Model** fits better!

In other words…

- The Negative Binomial model is an extension of the Poisson model, incorporating both Poisson error and overdispersion

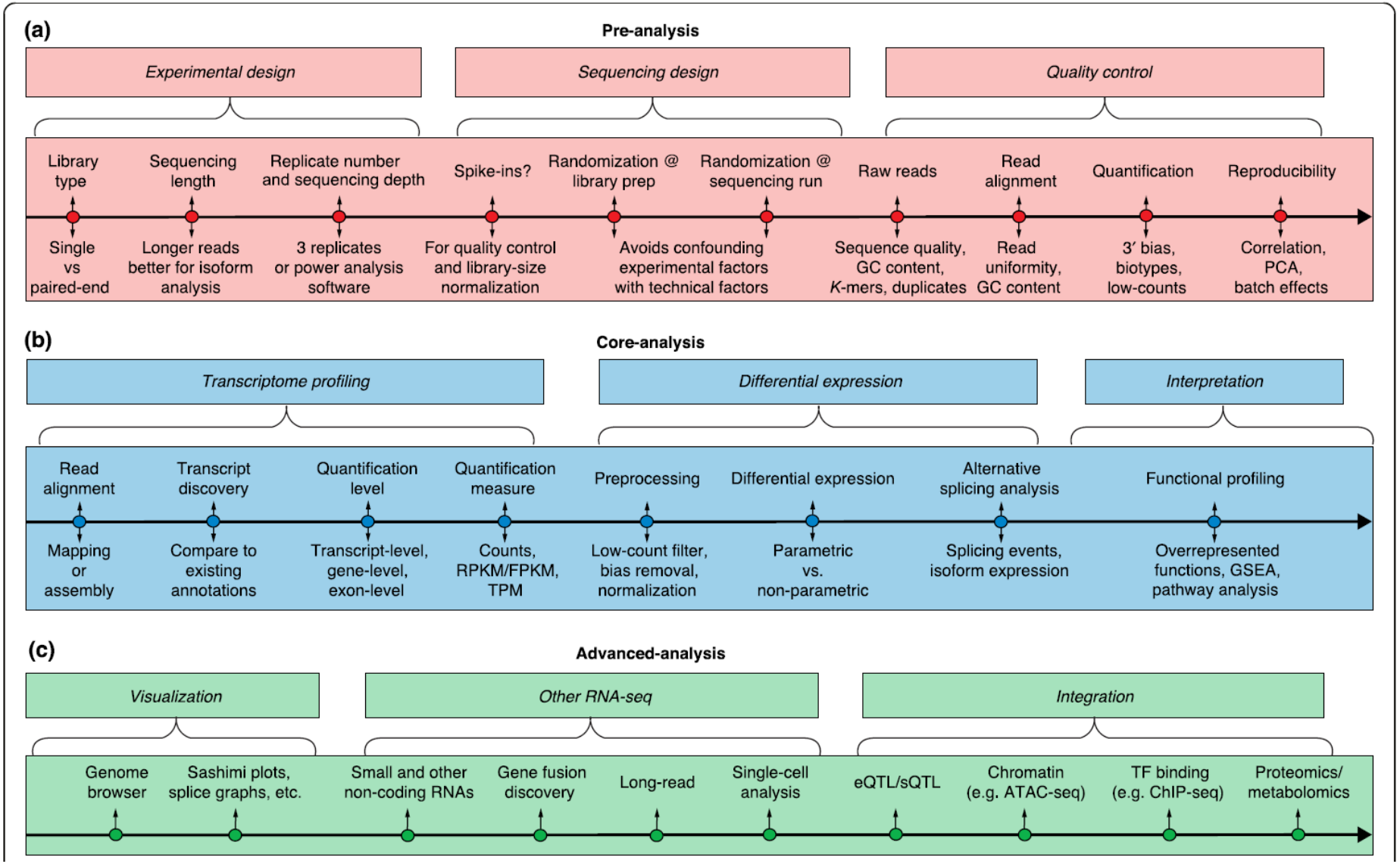
$$\text{Total } CV^2 = \text{Technical } CV^2 + \text{Biological } CV^2$$

Dominant source of variation for low counts, best explained by Poisson error

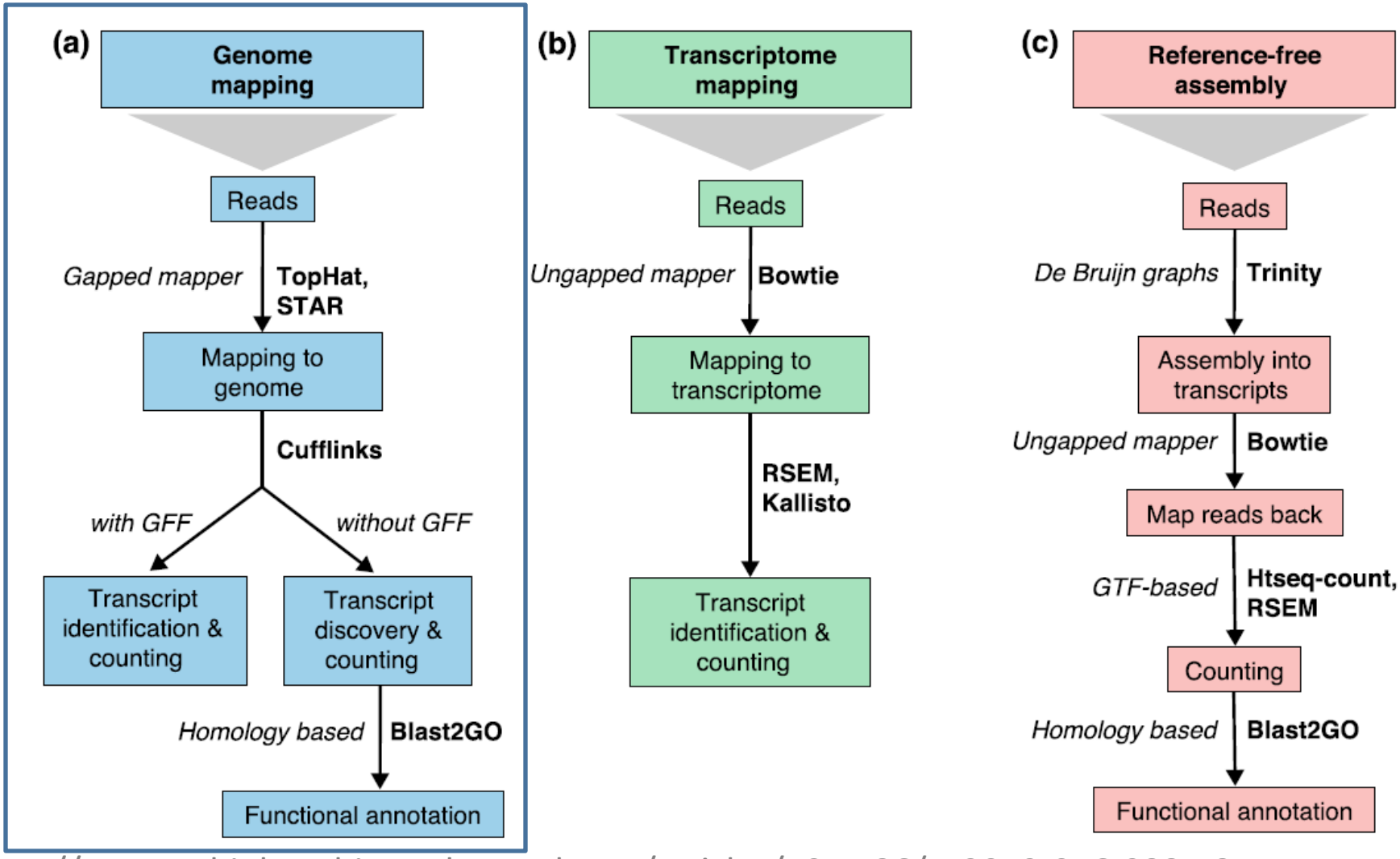
Dominant source of variation for high counts, best explained by dispersion parameter

Carrying on with the pipeline…

Generic pipeline



Three ways to «Rome» – Differential gene expression



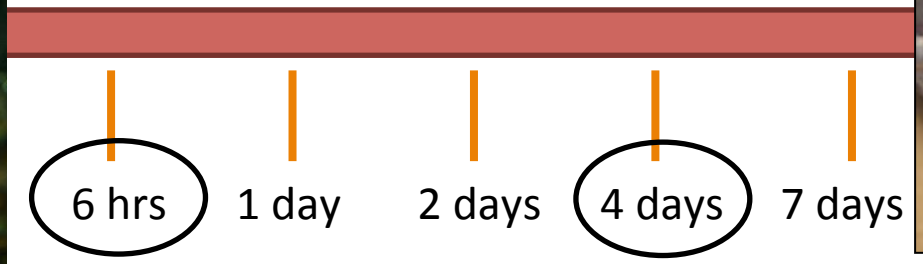
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

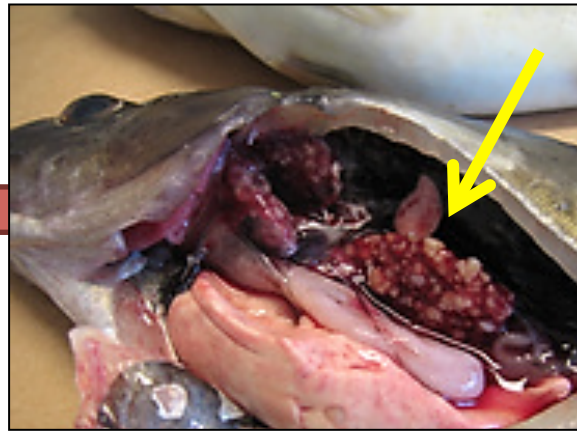
The strategies I

- Reference based (*ab initio*)
  - Maps RNAseq reads back towards reference genome and builds transcripts
  - Needs a certain amount of splice-junction covering reads
- *De novo* (with/without genome guiding)
  - Assembly of RNAseq reads only
  - Guided reads are clustered according to chromosome / scaffold prior to *de novo* assembly
- Mixed approach
  - Merging several assemblies to one

The INFBIO case

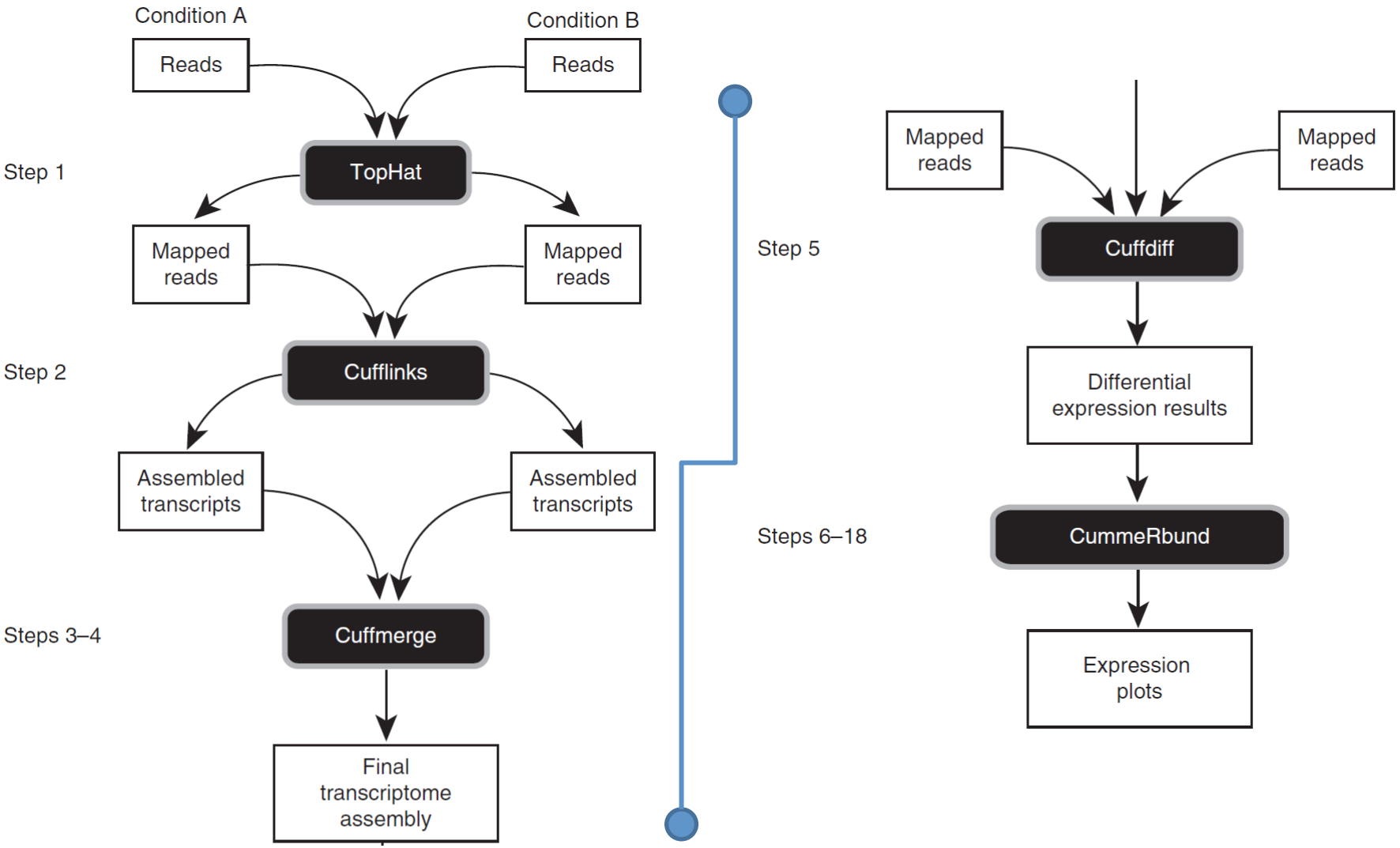
An infection over time

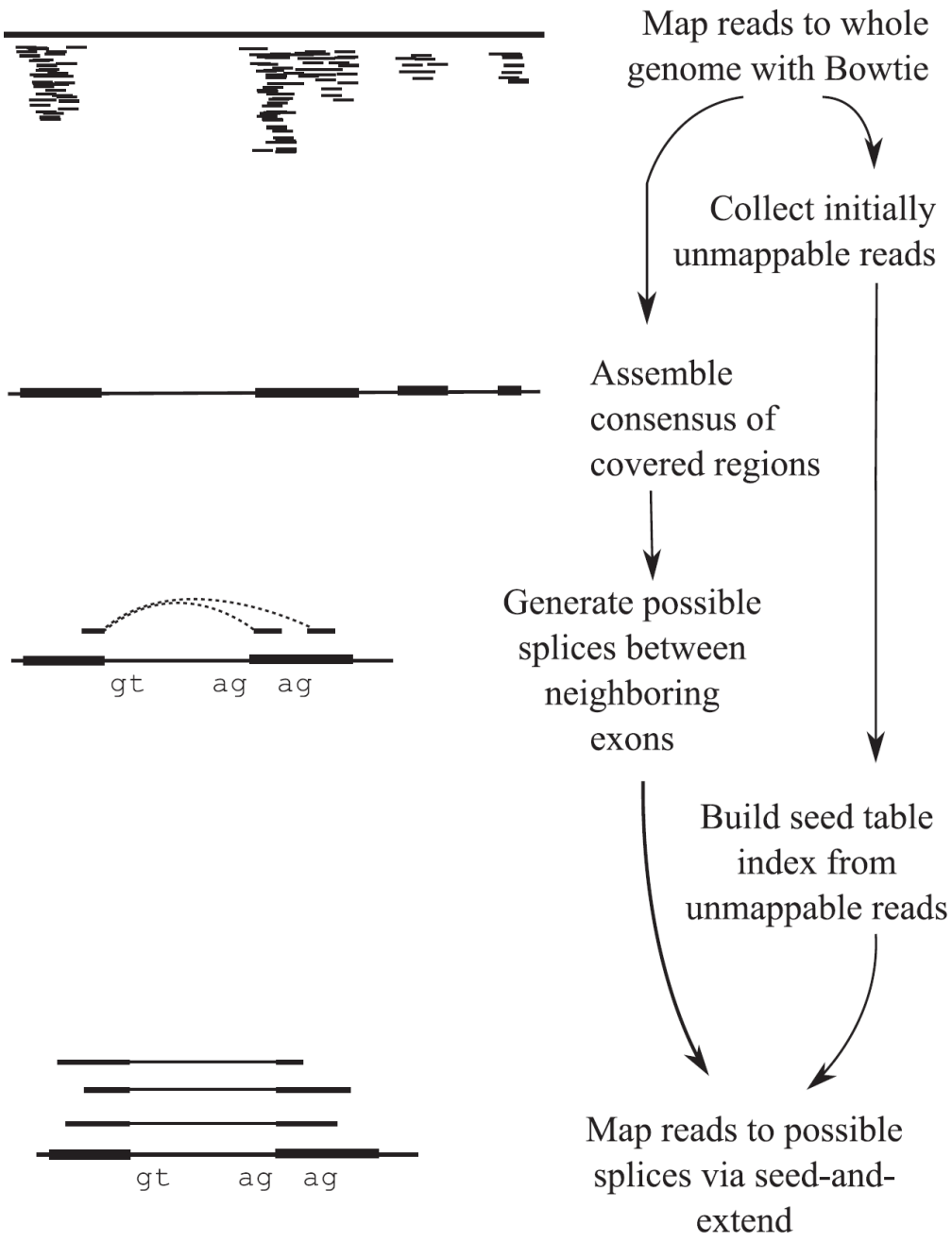




Two time-points, 6 treated and 6 controls
In total 12 samples

The TUXEDO pipeline



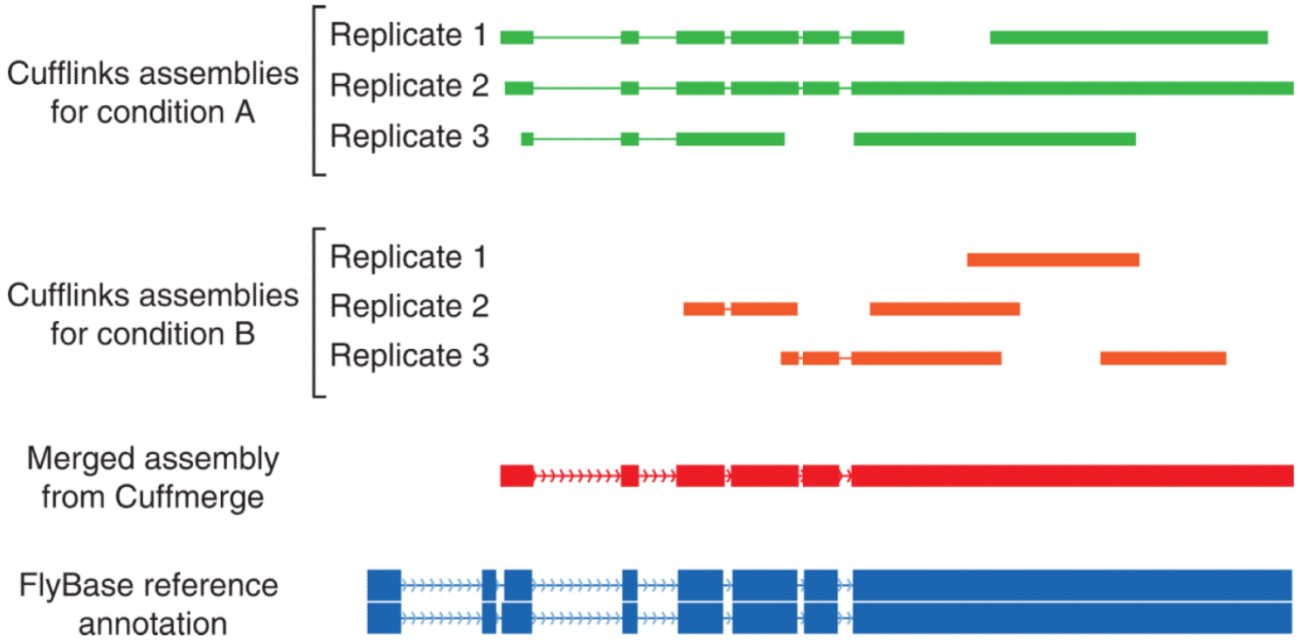

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
Build seed table index from unmappable reads
Map reads to possible splices via seed-and-extend
gt ag ag
gt ag ag

Cufflinks

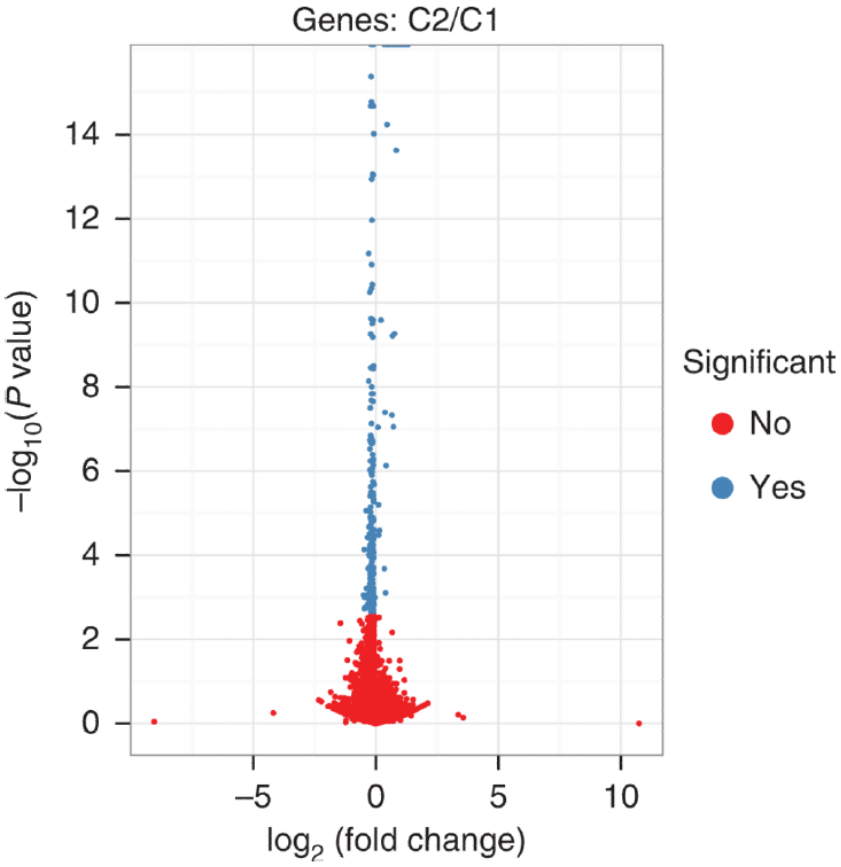
- Transcript assembly
  - A parsimonious strategy to resolve isoforms
- First level transcript quantification
  - Immature vs mature transcripts

Cuffmerge

- Pooling of cufflinks data per sample to ensure proper overall experiment ”present transcripts” overview

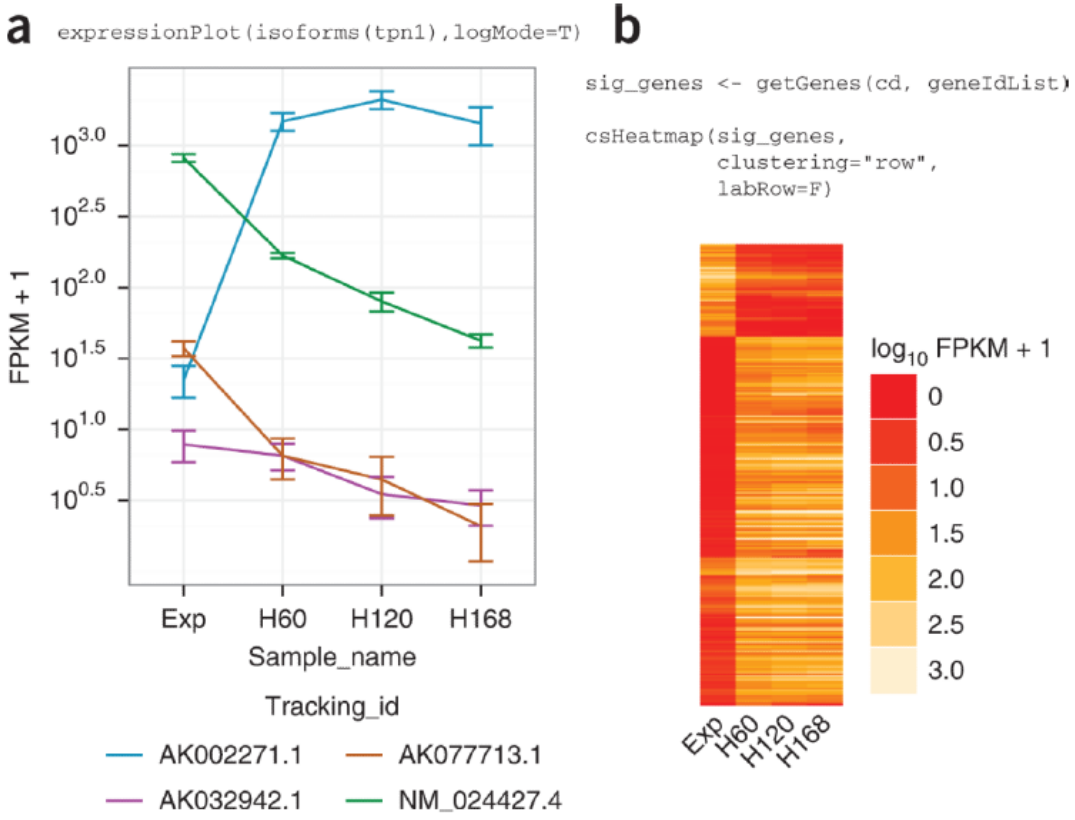


Cuffdiff



Cuffdiff ”learns the variation for each gene across replicates” to calculate differential expression

cummeRbund



All about the visuals…

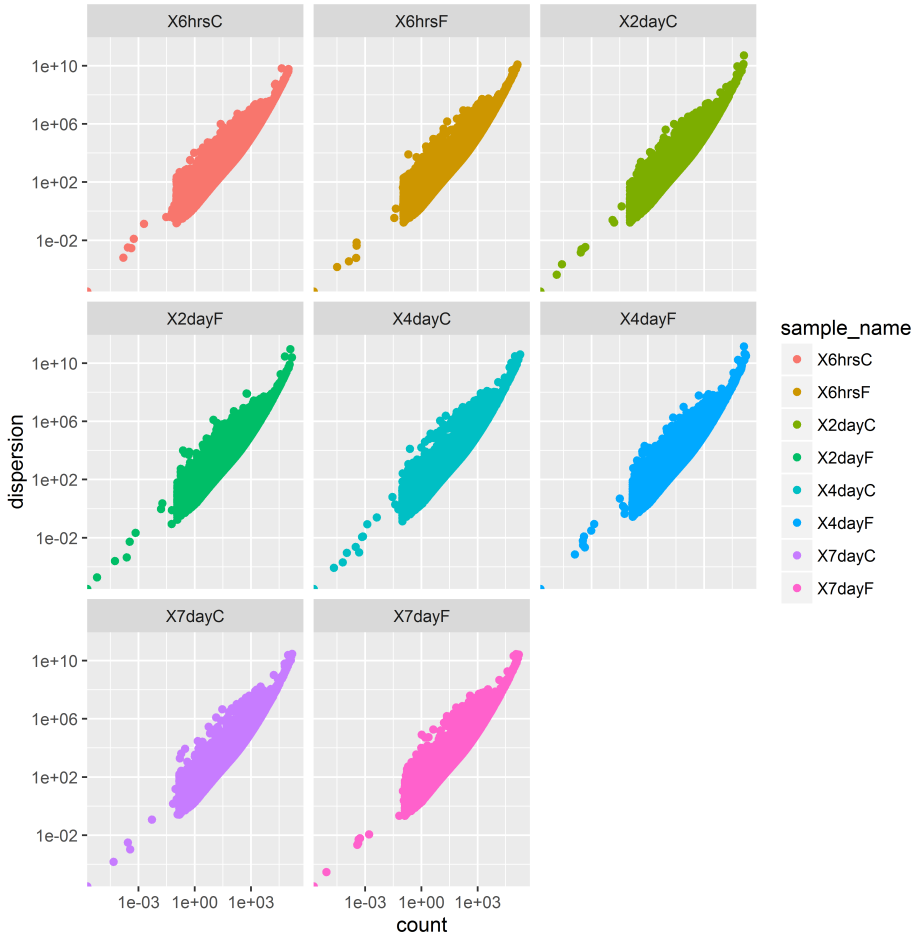
FPKM – a normalized read count measure taking into account gene length

- Fragments Per Kilobase of transcript per Million mapped reads
- Used for paired end data (RPKM is for single end data – reads per…)
- FPKM = 10^9 * C / (N * L)
  - C is the transcript-specific number of mappable fragments
  - N is the total number of mappable fragments
  - L is the number of basepairs in the exon (exon because isoforms – resolved later with a probabilistic algorithm)

The cummeRbund manual

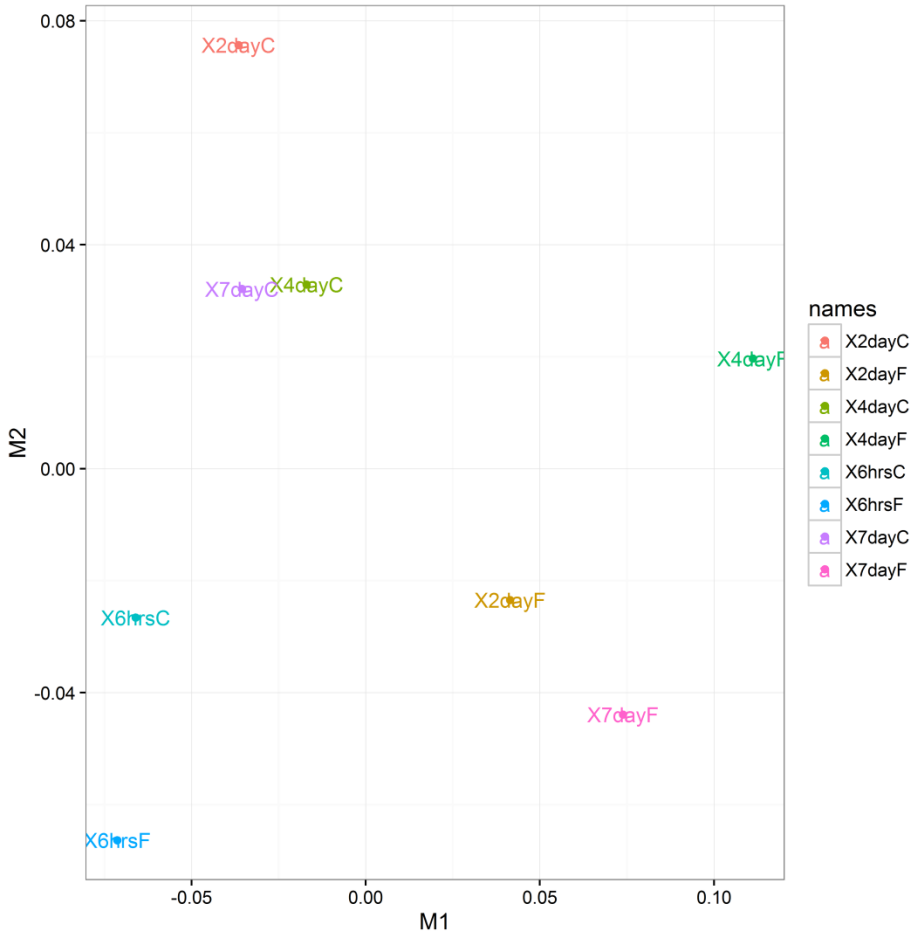
- http://bioconductor.org/packages/release/bioc/html/cummeRbund.html

Dispersion

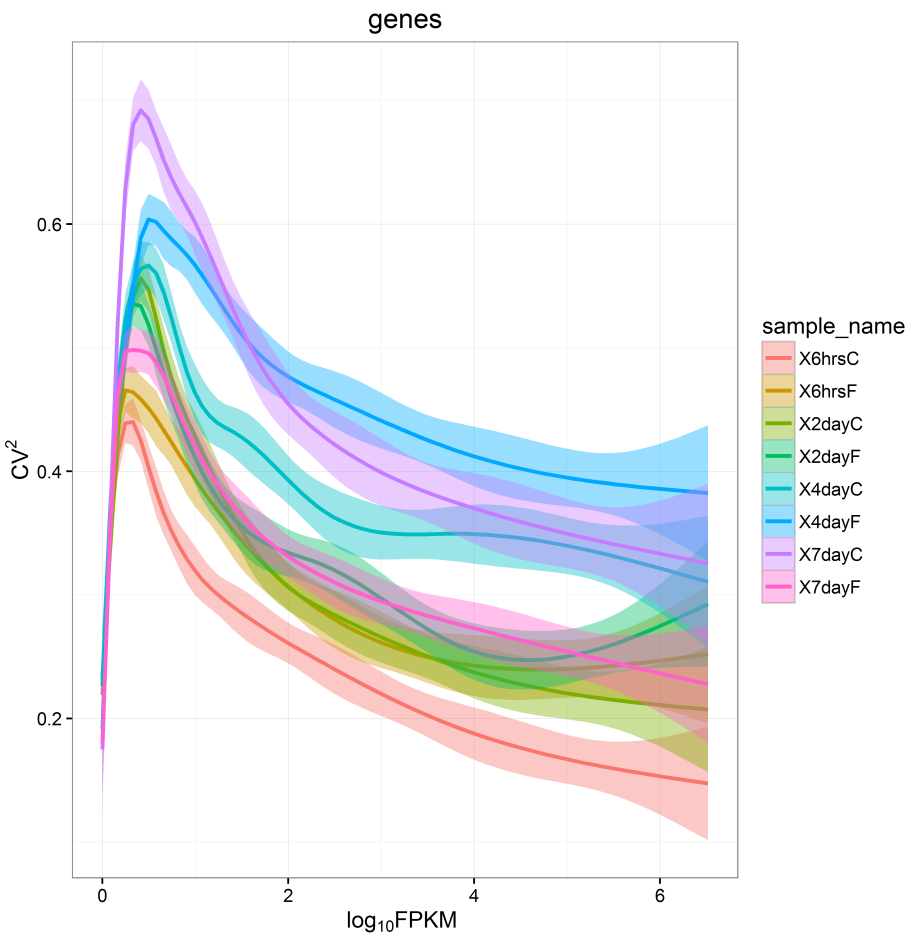


- Evaluate the model fitting (any overdispersion?)
- A scatter plot comparing the mean counts against the estimated dispersion for a given level of features from a cuffdiff run
- Default is all genes

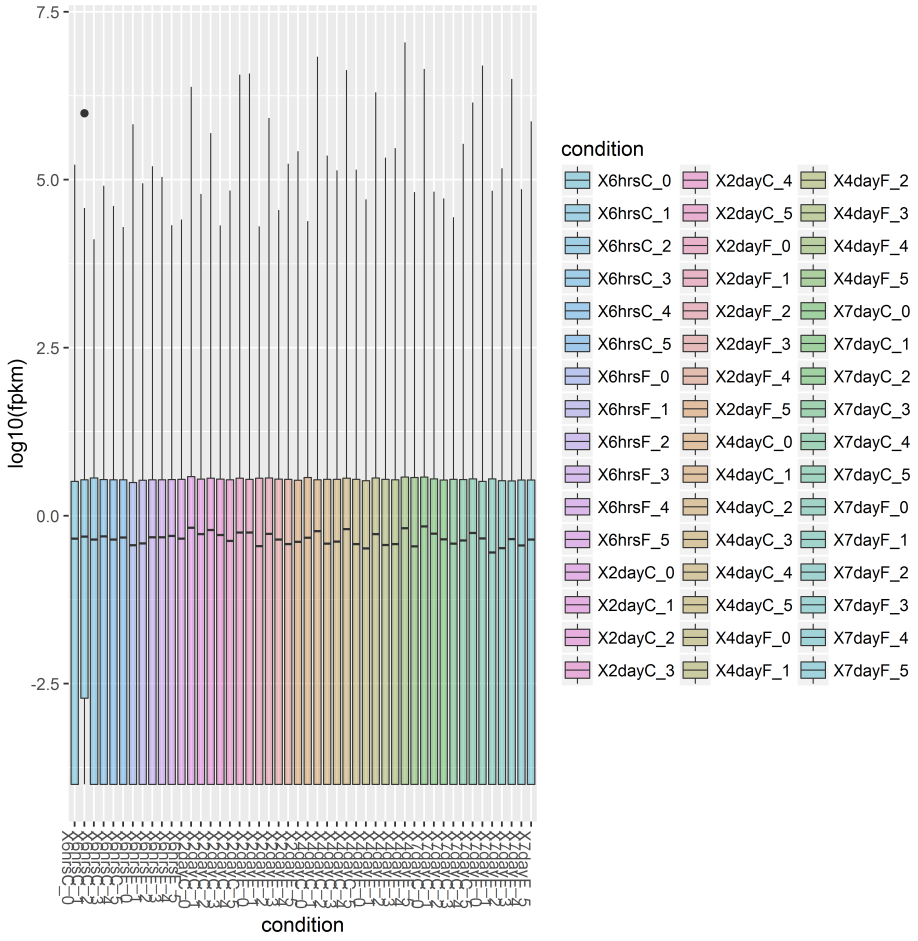
MDS (alt. PCA)


0.08
0.04
0.00
-0.04
M2
-0.05
0.00
0.05
0.10
M1
X2dayC
X7dayC
X4dayC
X4dayF
X6hrsC
X2dayF
X7dayF
X6hrsF
names
X2dayC
X2dayF
X4dayC
X4dayF
X6hrsC
X6hrsF
X7dayC
X7dayF

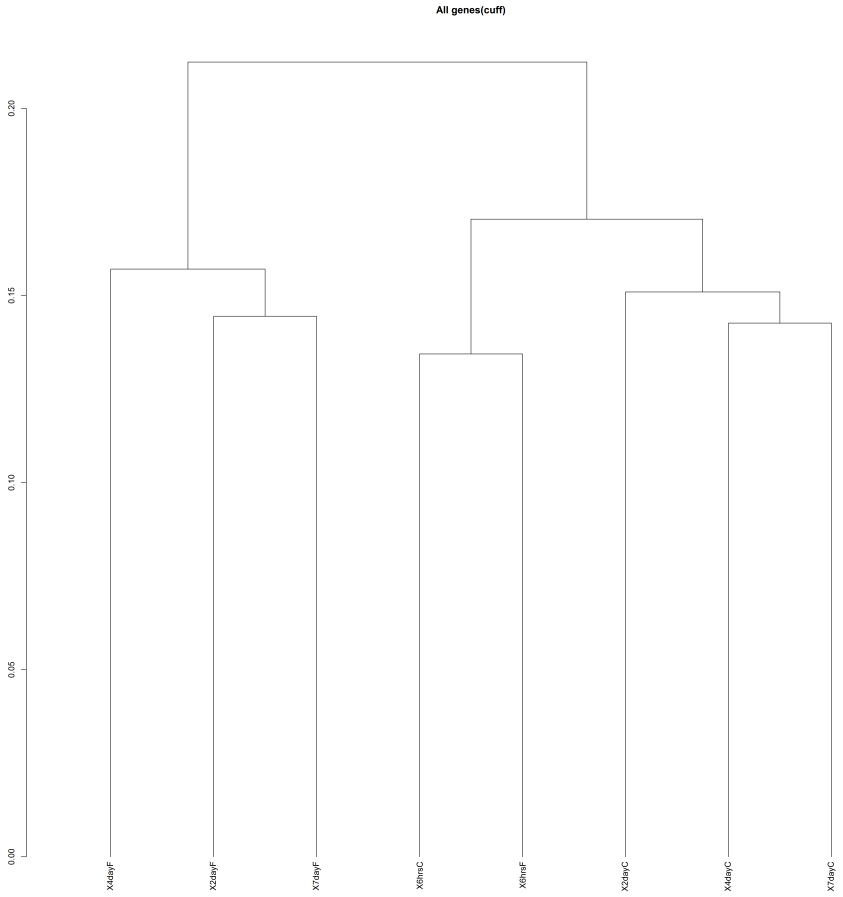
Squared coefficient of variation



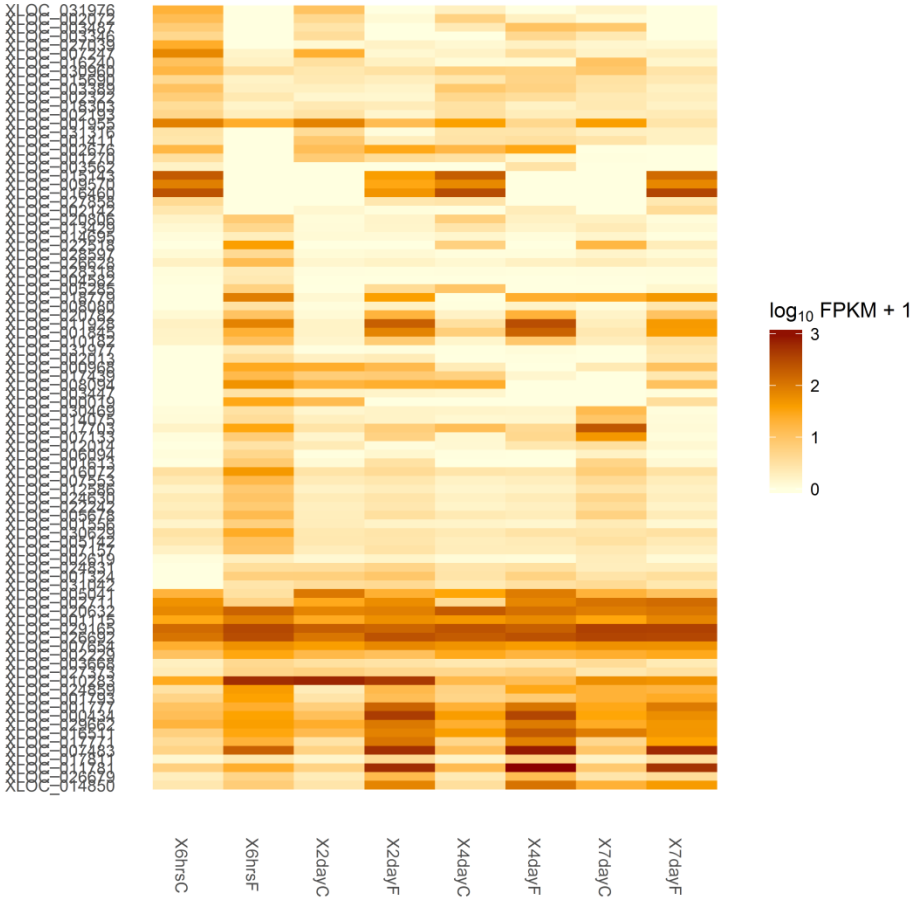
Boxplot (replicates shown)



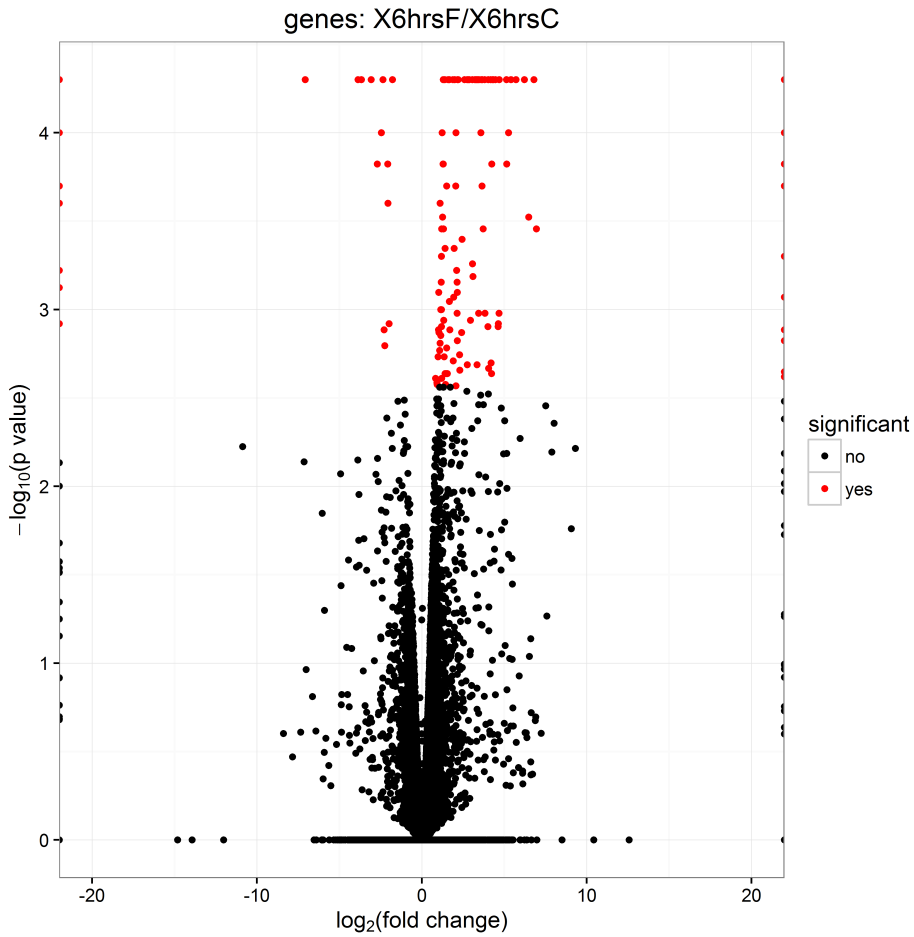
Dendrogram


All genes(cuff)
0.00
0.05
0.10
0.15
0.20
X4dayF
X2dayF
X7dayF
X5hrsC
X5hrsF
X2dayC
X4dayC
X7dayC

The genes in 6 hrs over time heat-map



Volcano 6hrs



Partitioning - 6hrs co-variance