Introduction to RNA-Seq Data Analysis

Dr. Benilton S Carvalho
Department of Medical Genetics
Faculty of Medical Sciences
State University of Campinas

- Material:
- http://tiny.cc/rnaseq
- Slides:
- http://tiny.cc/slidesrnaseq

Tools of Choice

- R and BioConductor:
  - Both created by Robert Gentleman;
  - Open-source tools;
  - Easy to prototype;
  - Communicate with C/C++/Fortran;

About R

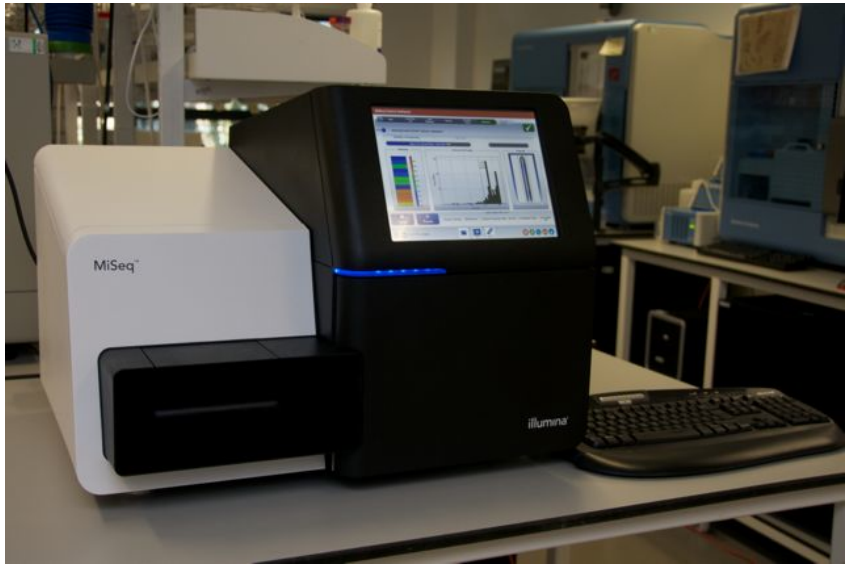
- Cross-plataform;
- Data analysis and visualization;
- Fast deployment to users;
- Able to interact with C/C++/Fortran;
- Thousands of packages:
  - Descriptive analyses;
  - Clustering and classification;
  - Regression Models and Trees;
  - Visualization;
  - Reproducible research;
  - Etc;

About Bioconductor

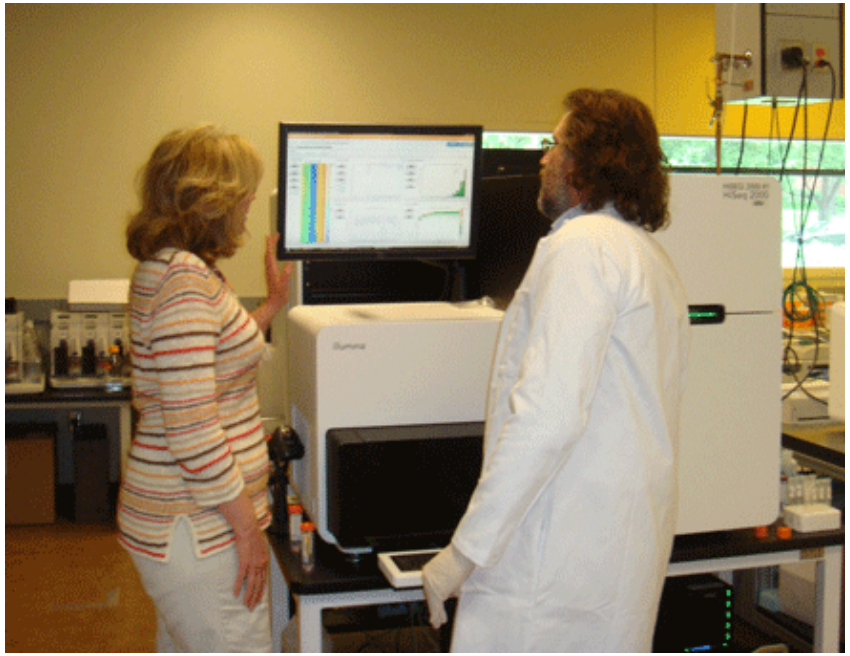
- Software infra-structure that uses R;
- Designed for biological data;
- Hundreds of packages:
  - Mass spectrometry;
  - Microarrays;
  - Next Generation Sequencing (NGS);
- Active community:
  - Heavily used by industry;
  - Releases in April and October;
  - Cutting-edge methods.

Illumina Products

MiSeq



HiSeq



Illumina Products

MiSeq

- 2 x 75bp ~ 24h : 3.8Gb
- 2 x 300bp ~ 65h : 15Gb

HiSeq

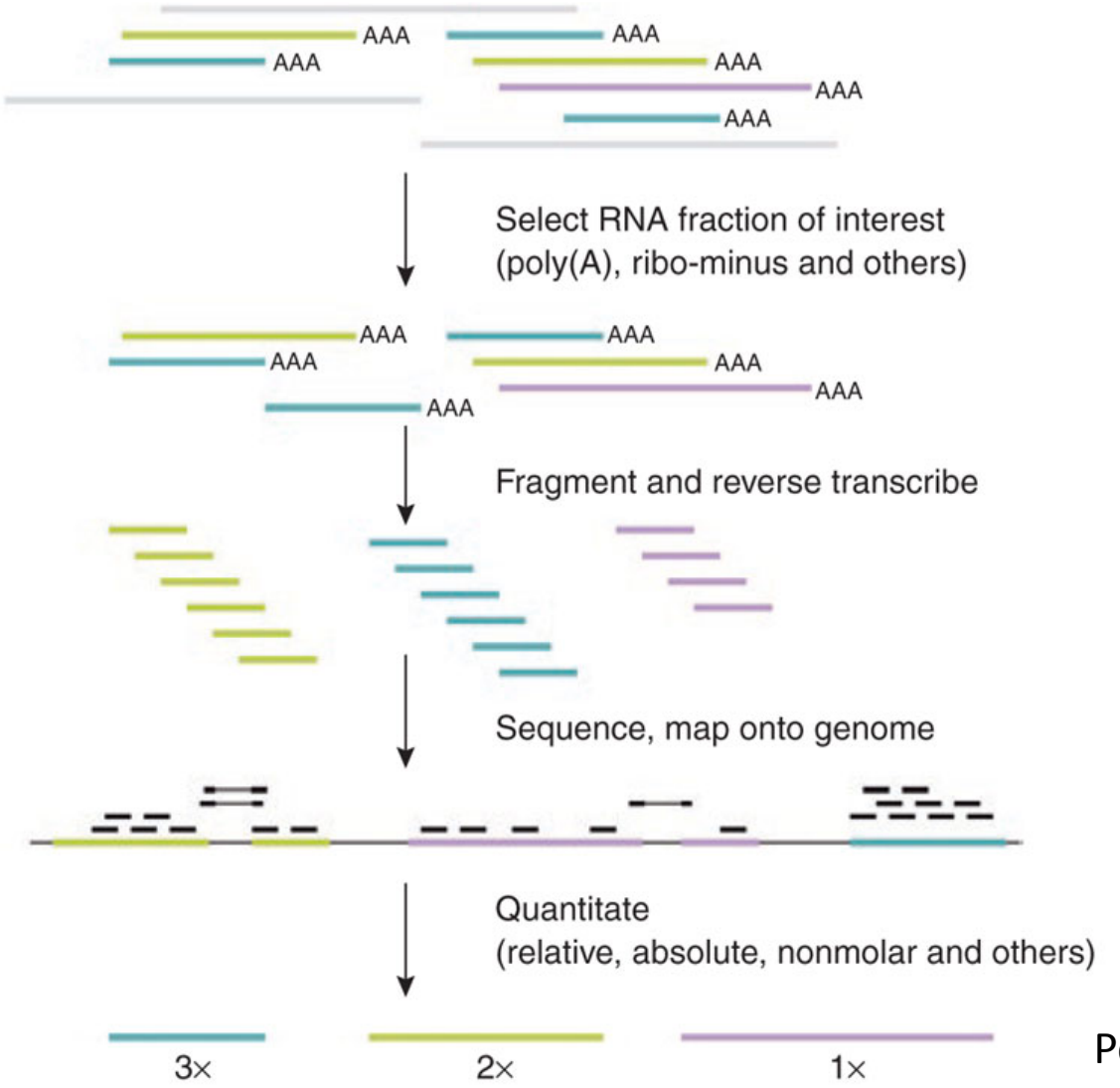
- 1 x 36bp ~ 29h : 144Gb
- 2 x 50bp ~ 60h : 400Gb
- 2 x 100bp ~ 120h : 800Gb
- 2 x 150bp ~ 144h : 1Tb

Illumina HiSeq X Ten

- Considering the Human Genome @ 30x;
- 320 Genomes per week;
- 1500 Genomes per month;
- 18000 Genomes per year;

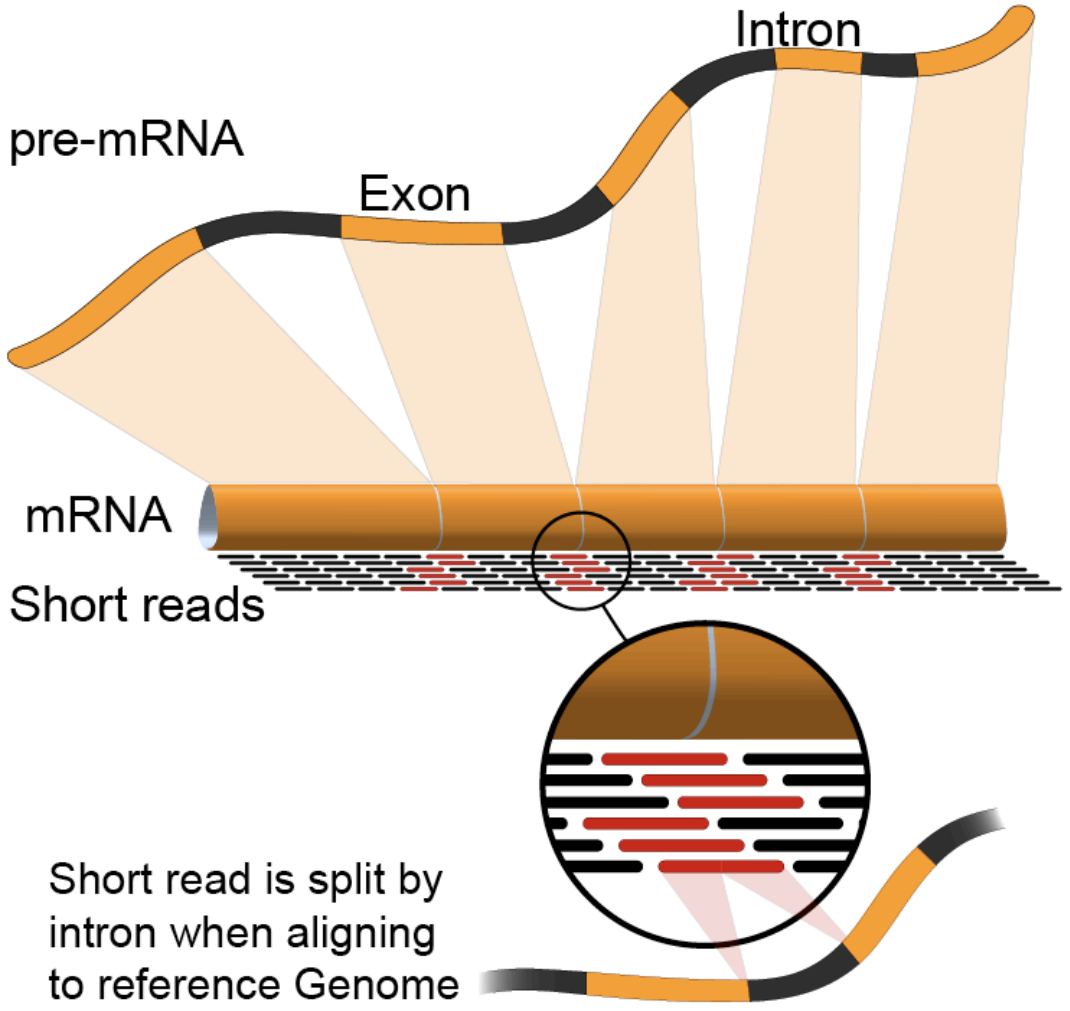
- Note: HiSeq 2500 ~ 10 Genomes per week

How does RNA-Seq work?

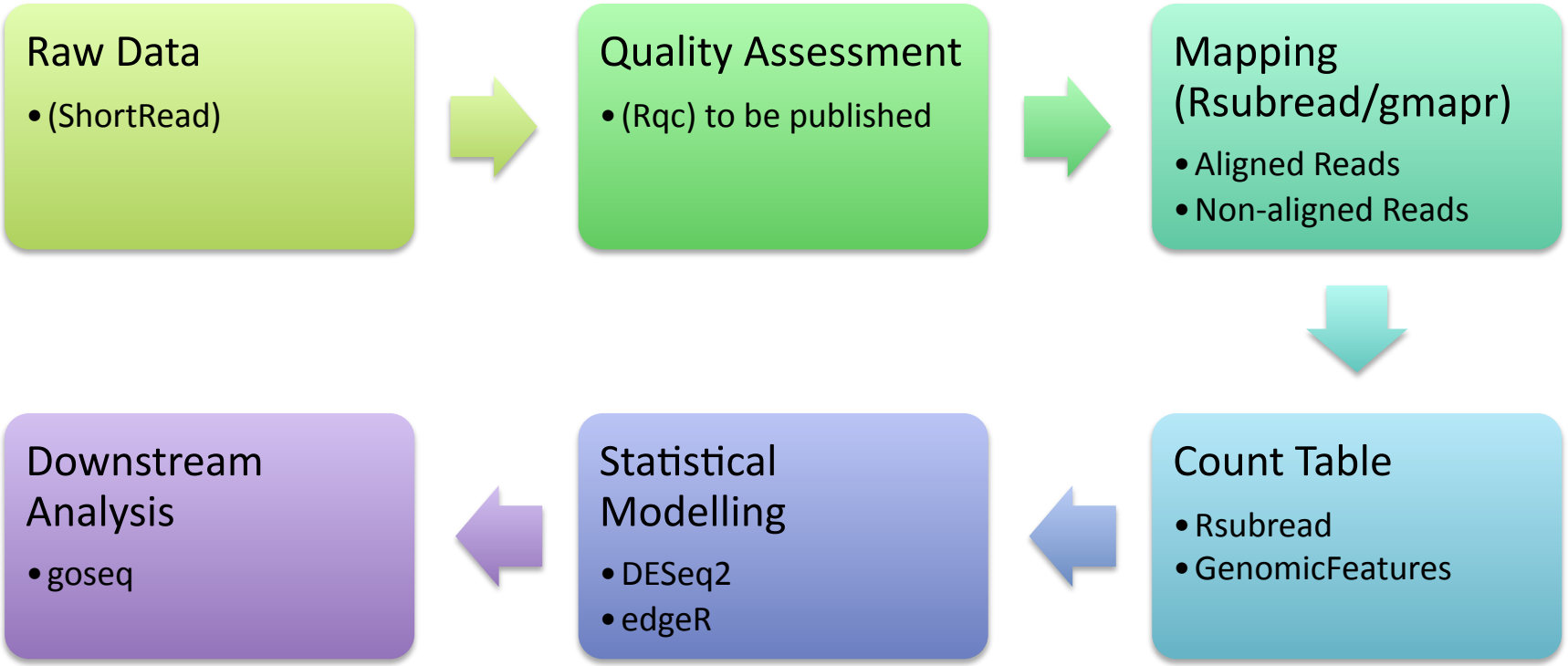


Pepke et. al. (2009)

How does RNA-Seq work?



Pipeline for Analysis



Relatively Large Files

- In our pilot experiment (per sample):
  - FastQ: 20GB per strand;
  - BAM: 8GB;
  - Counts: 250KB;
  - Temporary Files: 2 x 20GB per strand;
  - Total: ~ 130GB!
- The example above: RNA-Seq on Rats;
- For Human samples, when sequencing DNA, files are in average 10x bigger;

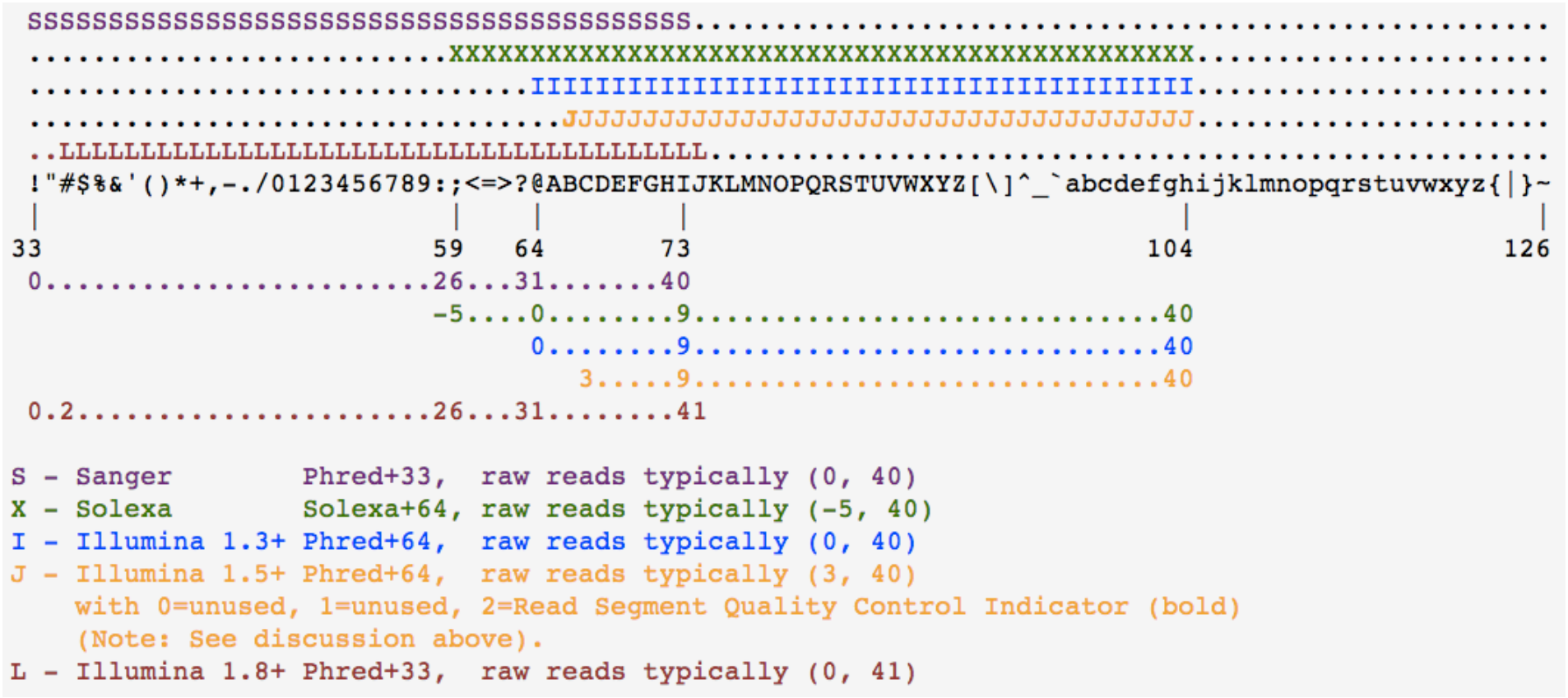
RAW DATA

Inside a FASTQ File

Instrument
Run ID
Flowcell ID
Lane
Tile number
X in tile
Y in tile

Mate
Fail filter
Control bits
Index seq

The Mistery of the Quality Scores



The Mistery of Quality Scores

- Base 1:
  - G/@
- @ = 31
- PHRED = 31
- -10*log10(1-P) = 31
- P = 0.9992057

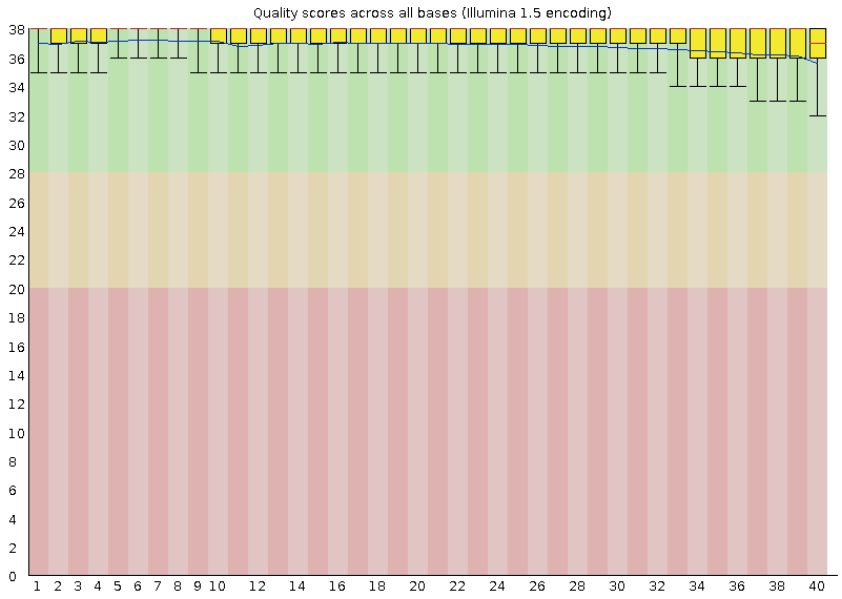
QUALITY ASSESSMENT

FastQC

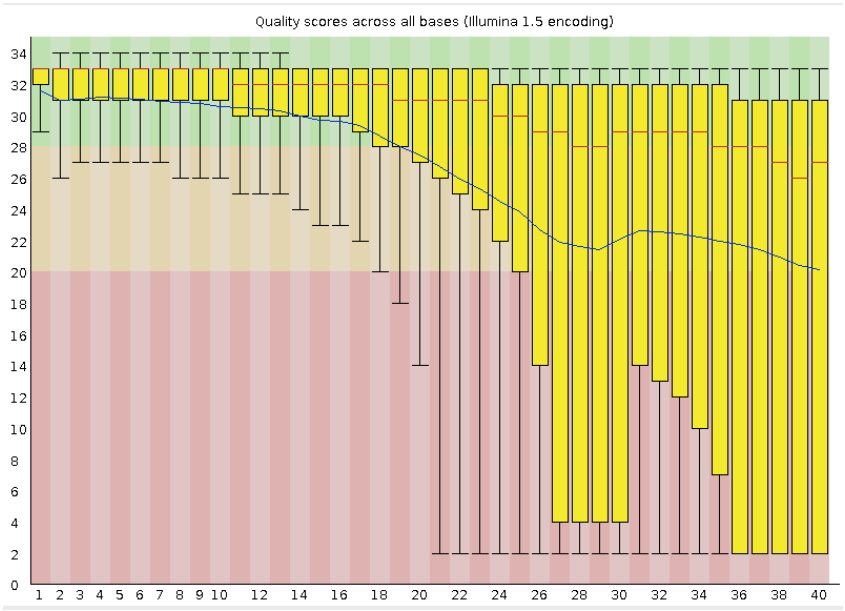
- We have experience with FastQC, but we are developing our own tool;
- FastQC is Java-based;
- Includes the option of pointing and clicking;
- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

FastQC – Per Base Seq Quality

Good

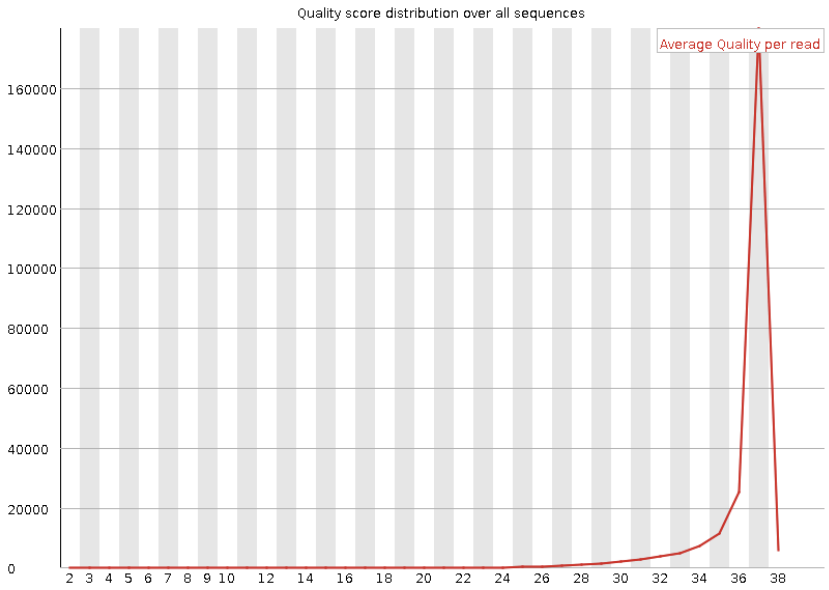


Poor

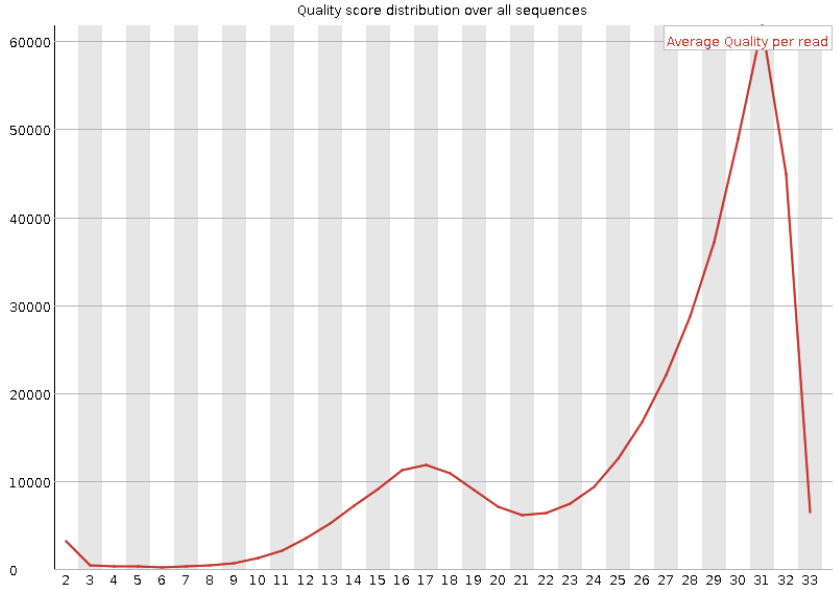


FastQC – Quality Score over All Seqs

Good

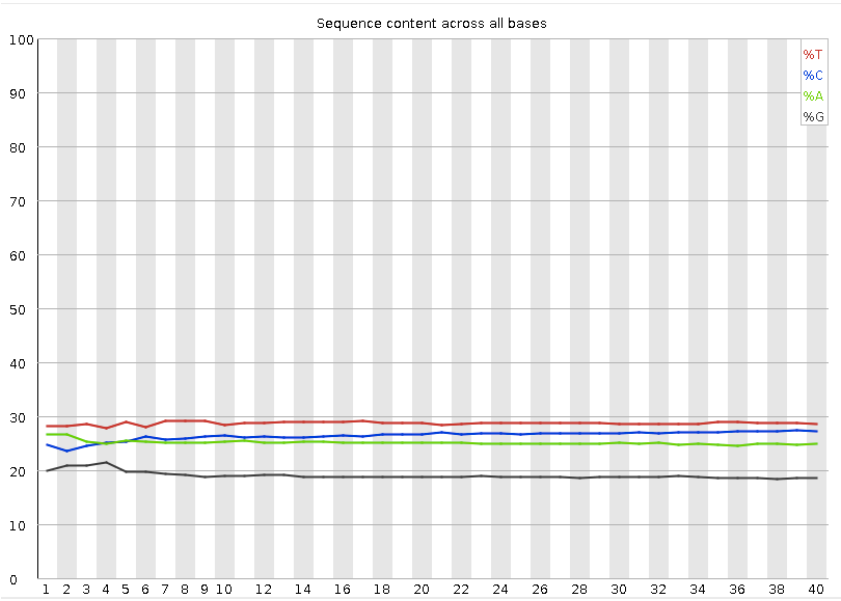


Poor

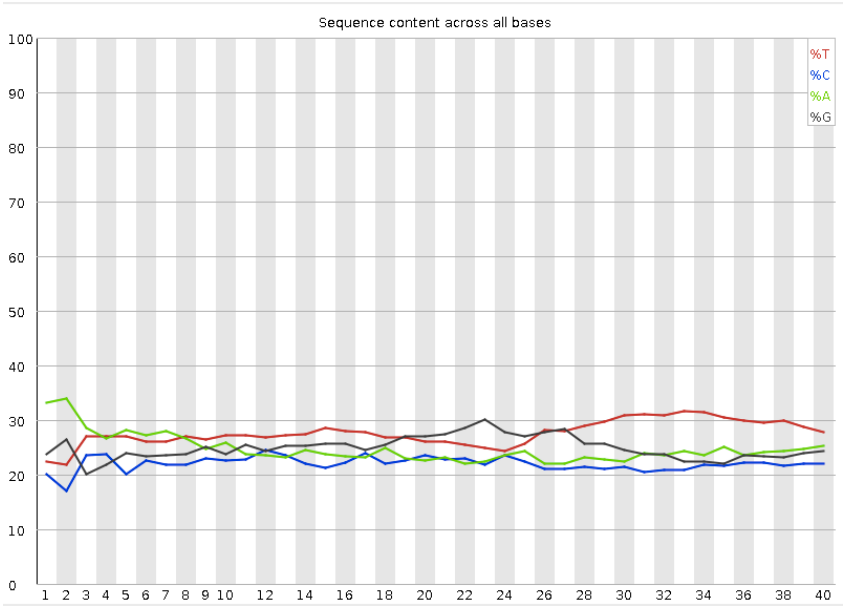


FastQC – Sequence Content

Good

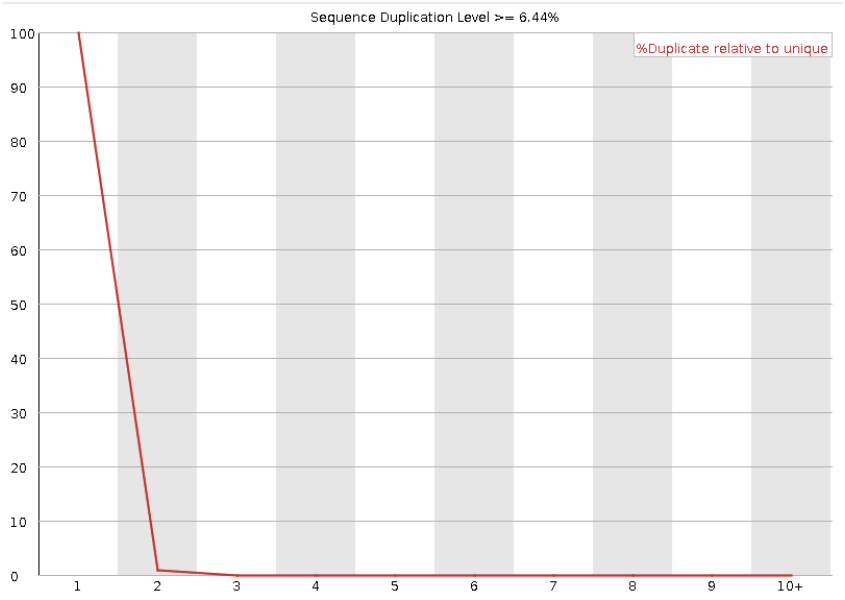


Poor

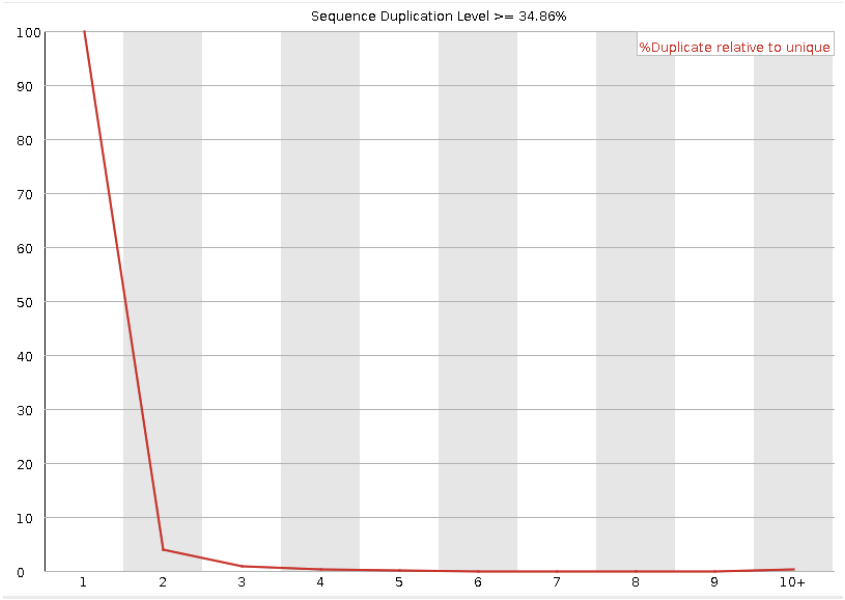


FastQC – Sequence Duplication

Good



Poor



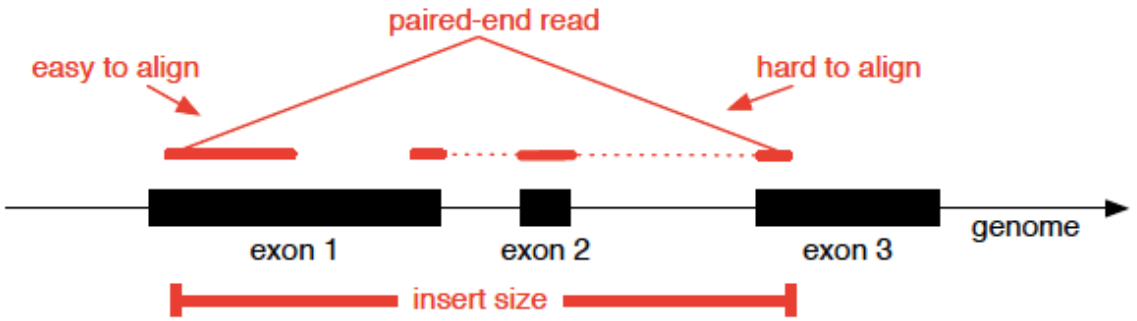
MAPPING

Principles of Mapping

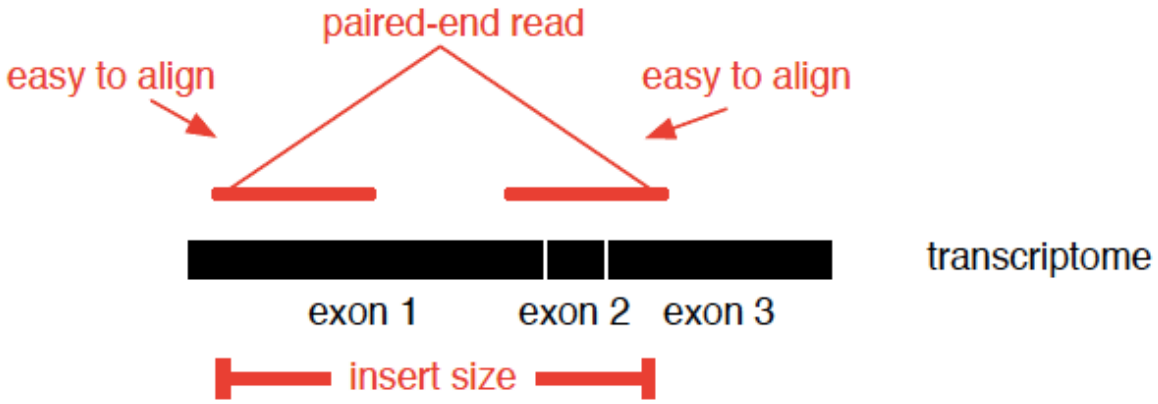
- Obtain the reference (genome or transcriptome) for the organism of interest:
- Mapping to the genome:
  - Allows for identification of novel genes/isoforms
  - Must allow for gaps (really hard)
- Mapping to the transcriptome:
  - Fast(er)
  - No need for spliced alignments
  - Can't find novel genes/isoforms

Principles of Mapping

Genome alignment (e.g. align to 23 chromosomes):



Transcriptome alignment (e.g. align to 150,000 *known* transcripts):



Result of Mapping: SAM/BAM

```
HWI-ST932:92:C1EU1ACXX:1:2213:6821:52150        113
1       171448  197     10M1D90M        =       171448
100     GTCGCAACTTGGAGCTTGCCTGAACATGCCTCACAGAATCCAAACACA
GGACACAGAGCACAGCAGCCAGGACCATTTAAGAAGGCTTAGCTACTACGCG
8=DCCC@CCCDDDDBCCCEFEDDDCFHHJIGIGIIJIIIFEHF=F?IIHGFGBJII
IGHHJIJIIIIGGFDCIGIJIHEHGGEEJIFGHFDHDDDDFCC@    SA:i:0
SH:i:91 NH:i:1
```

op	Description
M	Alignment match (can be a sequence match or mismatch
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

COUNT TABLE

The BAM isn't the final file

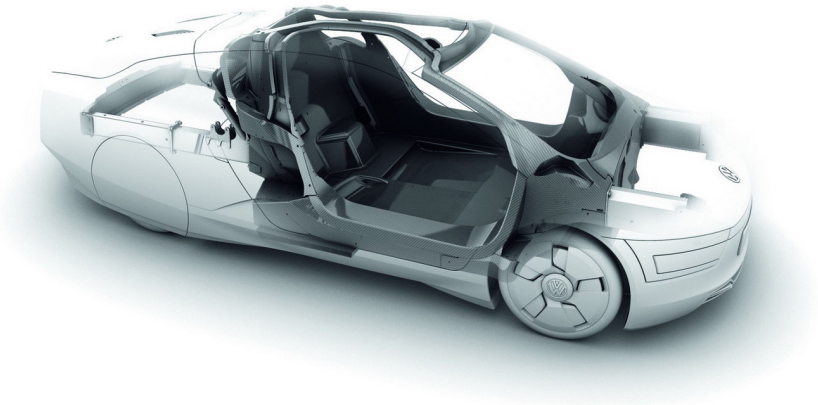
- BAM files give the location of mapped reads;
- But, per individual, how many reads should be considered as from any particular gene?
- The count table represents this;
- It can be obtained through *GenomicAlignments*, *HTSeq*, *Rsubread* and *EasyRNASeq;*

Count-table Example

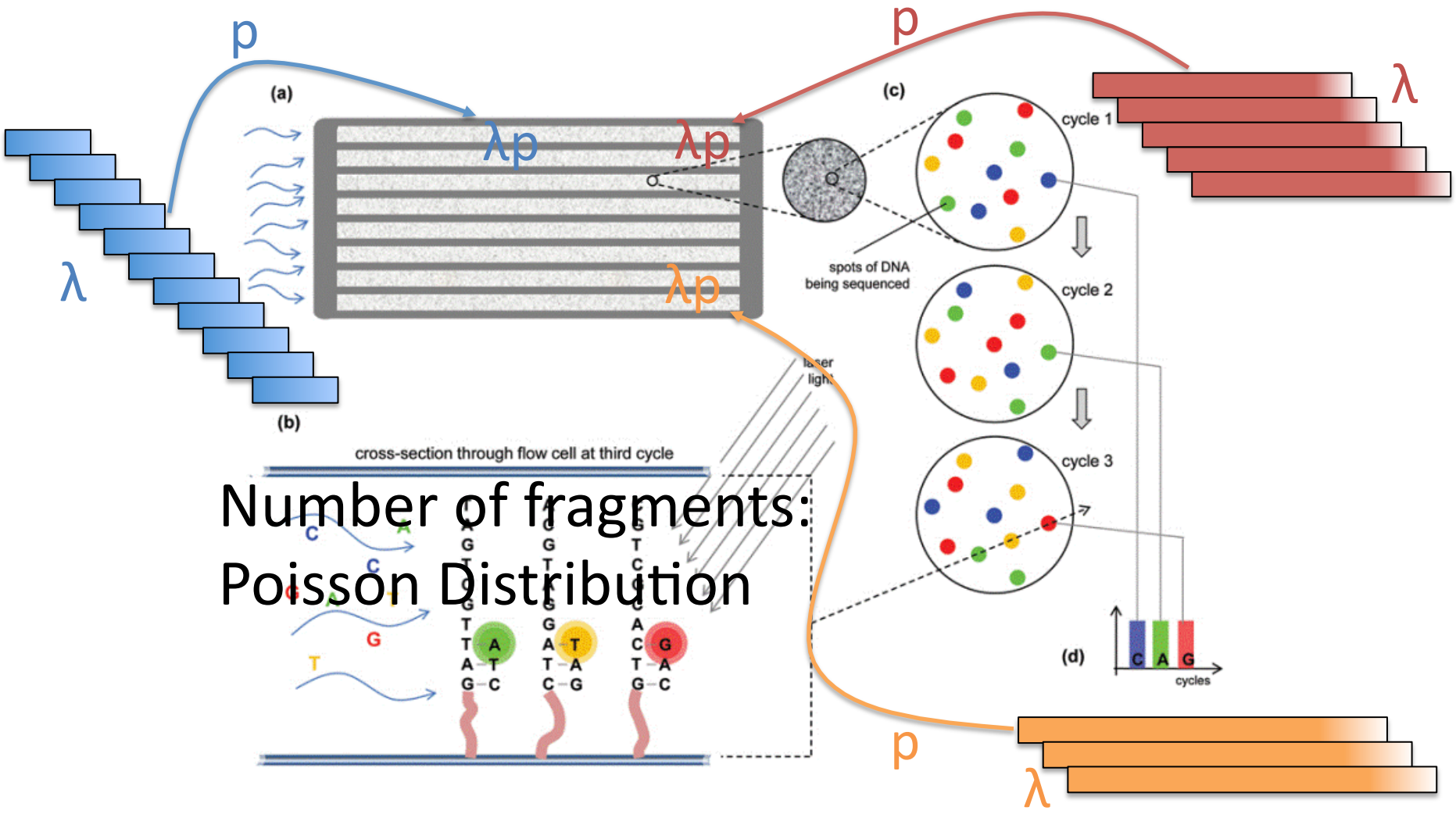
	C1	C2	C3	T1	T2	T3
ENSRNOG00000010603	0	0	0	0	0	1
ENSRNOG00000033787	4289	7831	12489	5904	5033	4619
ENSRNOG00000014887	3	7	7	1	3	3
ENSRNOG00000045753	0	0	7	0	0	2
ENSRNOG00000048290	9	11	7	11	6	5
ENSRNOG00000001689	233	375	466	489	405	266

STATISTICAL MODELING

What is a model?

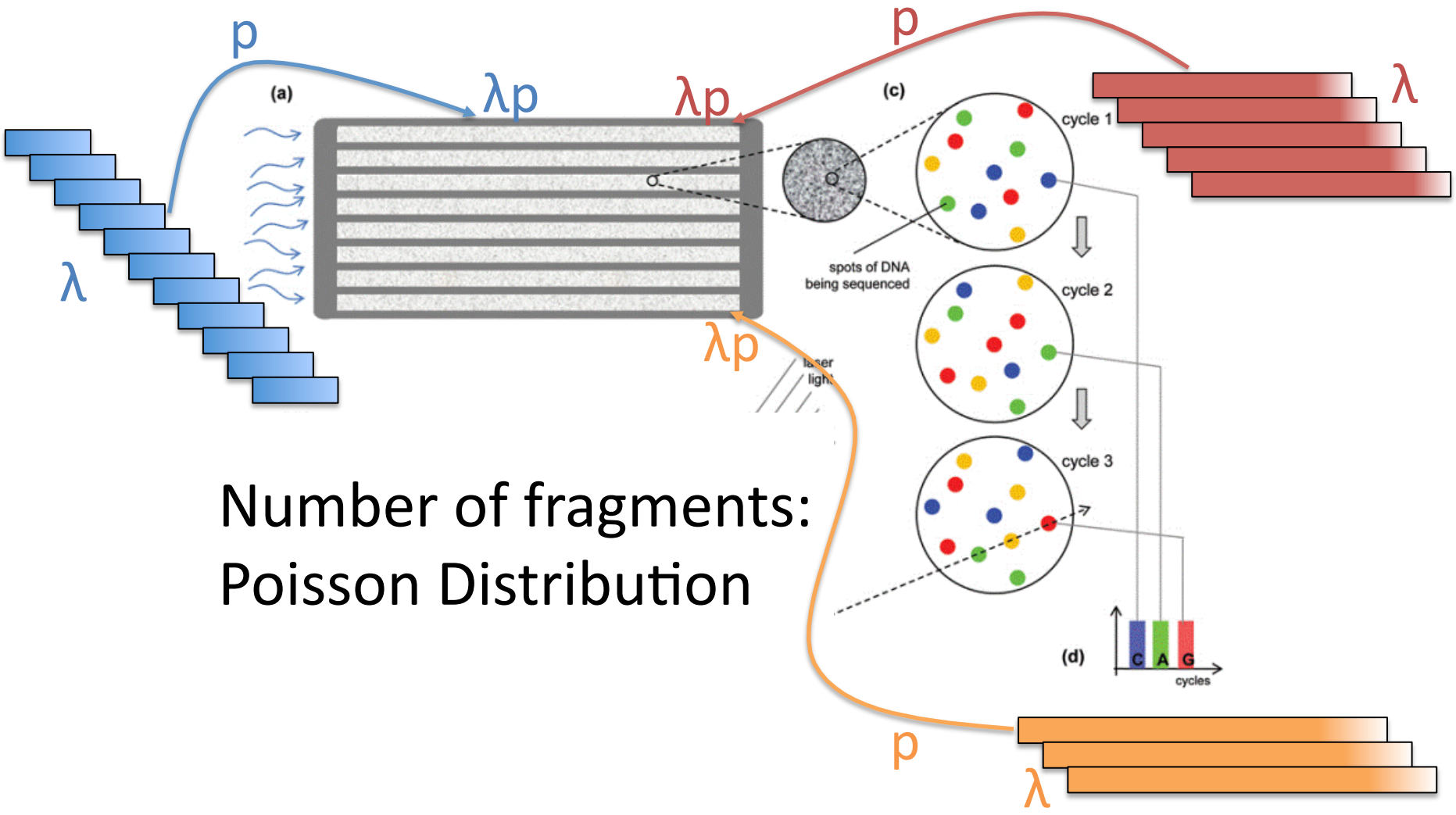


Different Transcripts, Rates and Probabilities



Number of fragments: Poisson Distribution

Different Transcripts, Rates and Probabilities



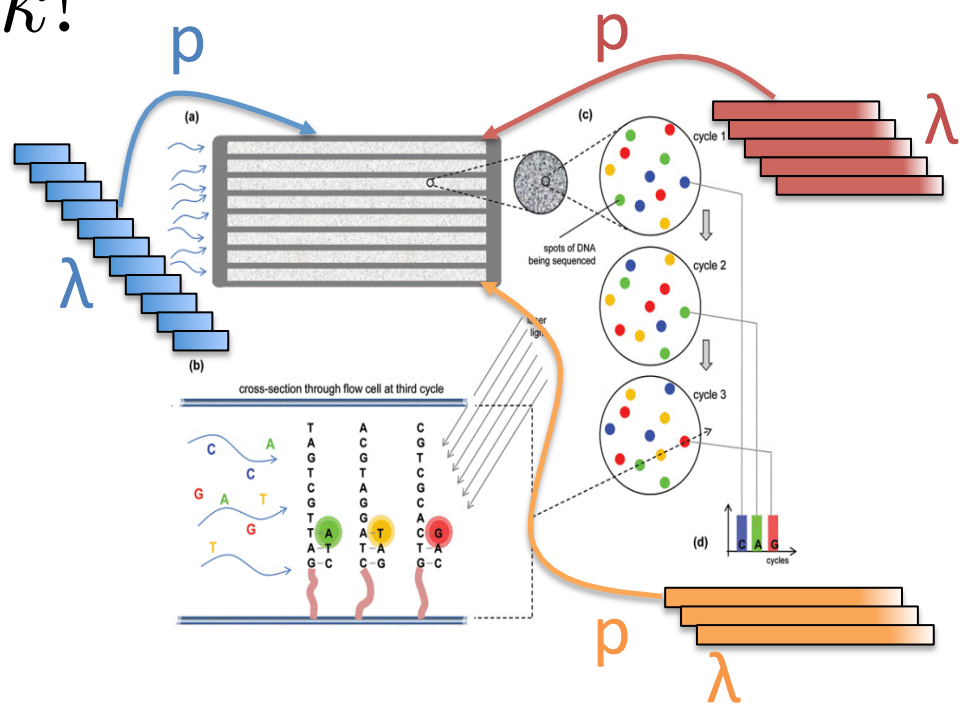
Number of fragments:
Poisson Distribution

Characteristics of a Poisson Distribution

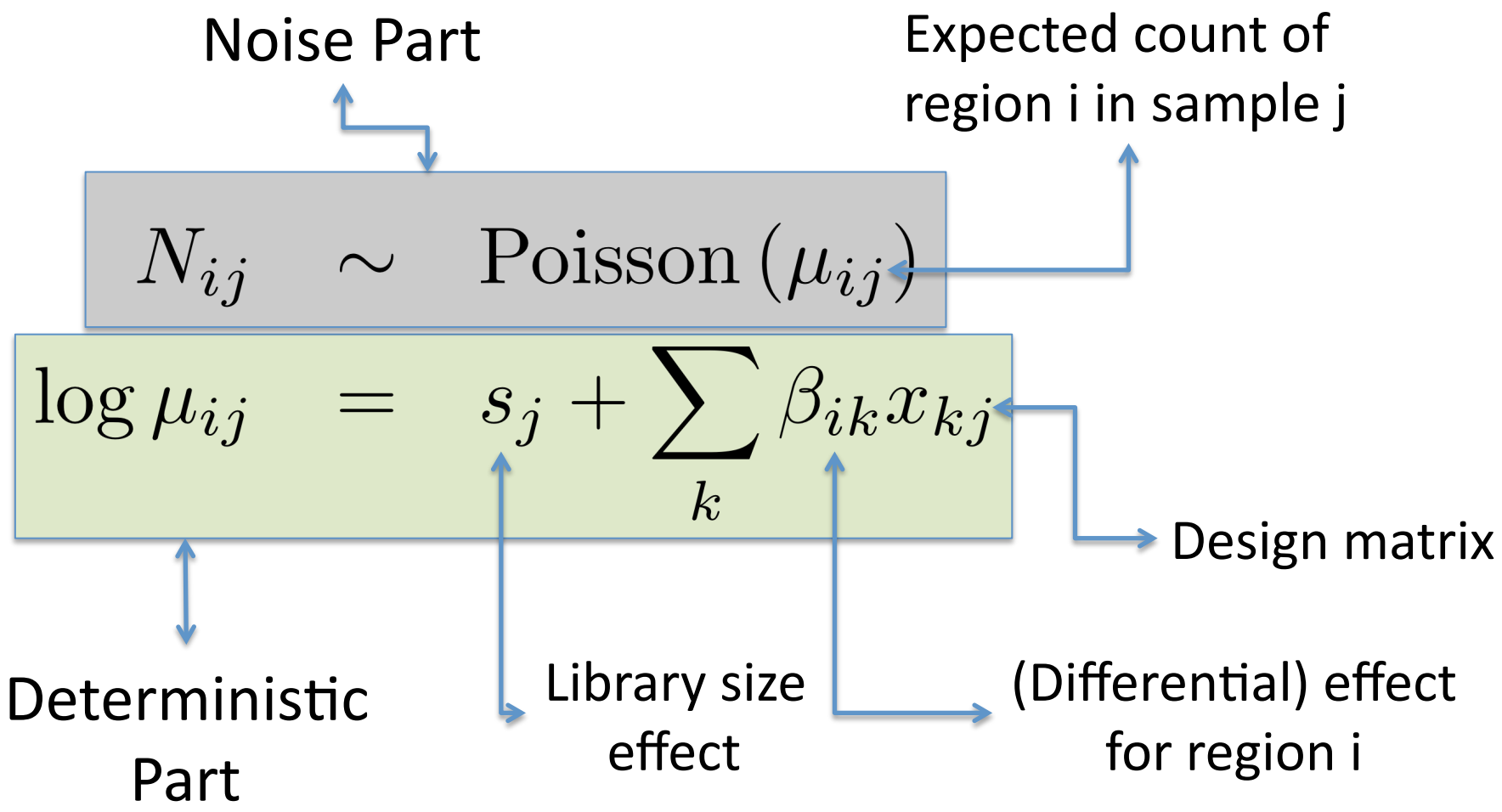
- X ~ Poisson(λp)

$$P(X = k) = \frac{(\lambda p)^k e^{-\lambda p}}{k!}$$

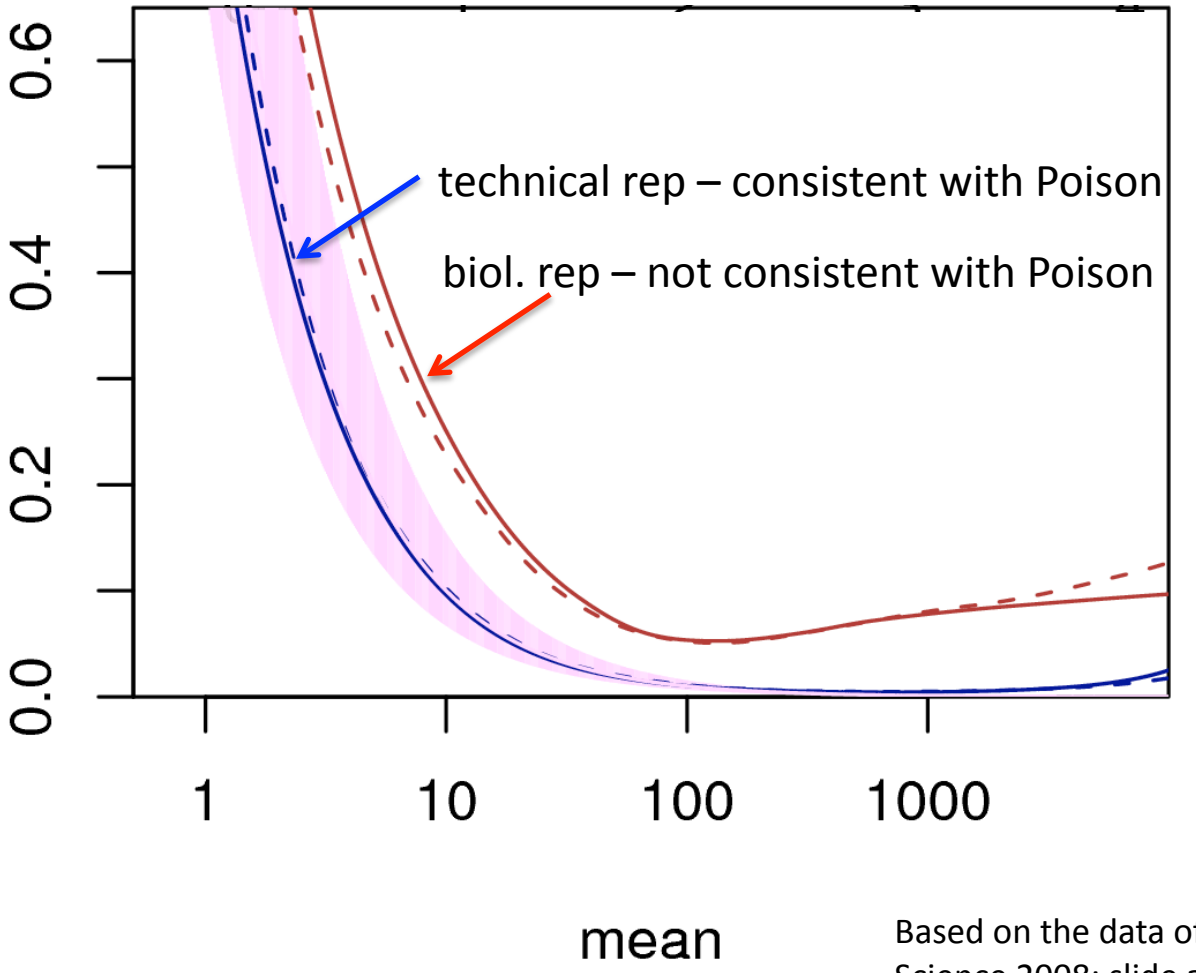
- Mean: λp
- Variance: λp



Analysis method: GLM



Need to account for extra variability

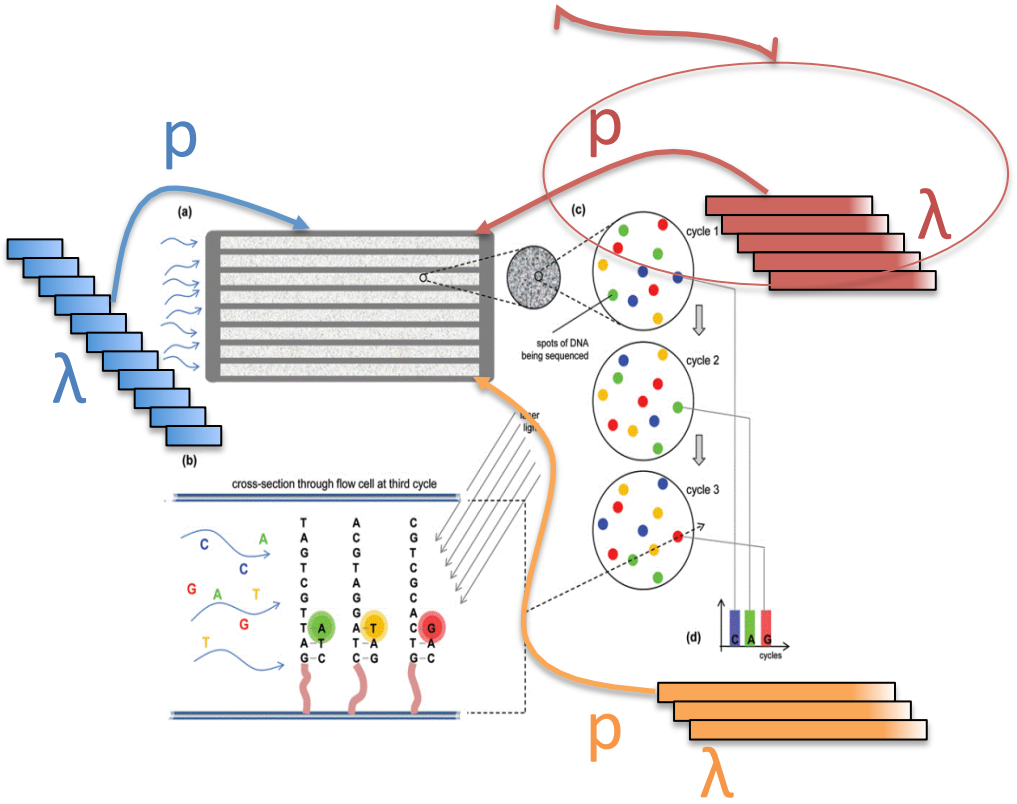


Based on the data of Nagalakshmi et al. Science 2008; slide adapted from Huber;

Characteristics of a Negative Binomial (NB) Distribution

- $X \mid \lambda p \sim \text{Poisson}(\lambda p)$
- $\lambda p \sim \text{Gamma}(a, b)$
- Mean: μ
- Variance: μ/ν
  $0 < \nu < 1$

Allow these to change!!!



Current methods for DE use NB model!

Sequencing – Rationale
Biological Replicates

- For subject j, on transcript i:

$$Y_{ij}|\lambda_{ij} \sim P(\lambda_{ij})$$

- Different subjects have different rates, which we can model through:

$$\lambda_{ij} \sim \Gamma(\alpha, \beta)$$

- This hierarchy changes the distribution of Y:

$$Y_{ij} \sim \mathrm{NB}\left(\alpha, \frac{1}{1+\beta}\right)$$

An additional source of variation

$$
\begin{aligned}
N_{ij}|\eta_{ij} &\sim \text{Poisson}\,(\eta_{ij}) \\
\eta_{ij}|\mu_{ij} &\sim \text{Gamma}\,(\beta_1(\mu_{ij}), \beta_2(\mu_{ij})) \\
N_{ij} &\sim \text{NB}\,(\mu_{ij}, \alpha(\mu_{ij})) \\
\log \mu_{ij} &= s_j + \sum_k \beta_{ik} x_{kj}
\end{aligned}
$$

Smooth dispersion-mean relation α

Design matrix

(Differential) effect for region i

Deterministic Part

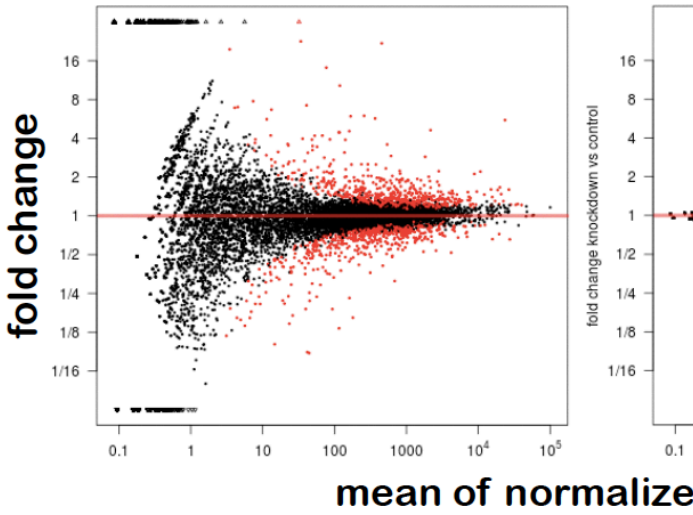
Library size effect

Summary of the Poisson and Negative Binomial Models

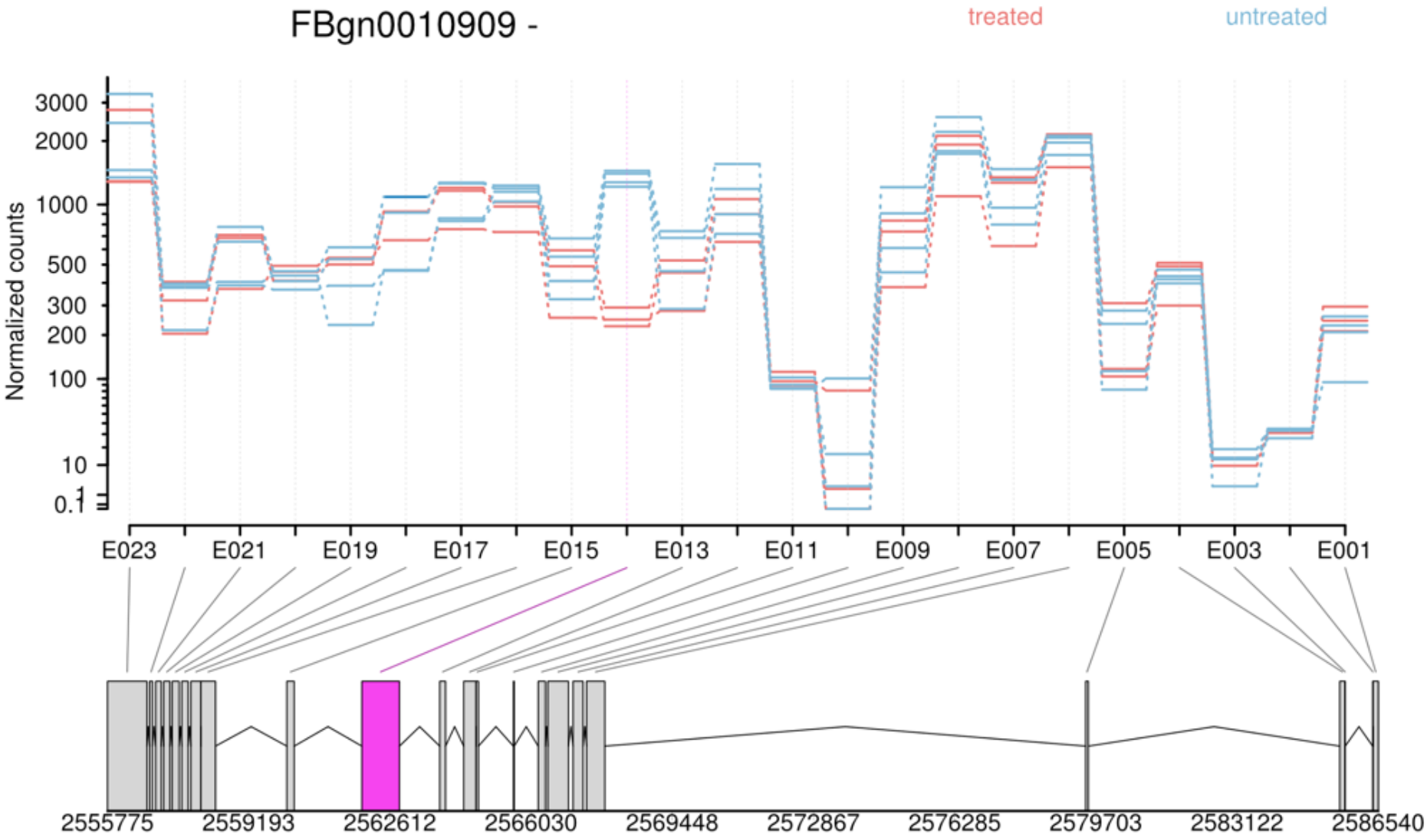
- Poisson(λ):
  - Mean: λ
  - Variance: λ
- Negative Binomial (α, 1/(1+β)):
  - Mean: α/β
  - Variance: α(1+β)/β²
    - = α/β + α/β² = mean + 1/α * mean²

Shot noise

Biological noise



Example: DE / DEU



Summary of Models
Treatment (x_j) as Covariate

Gene Expression / DESeq

$$N_{ij} \sim NB(s_j \mu_{ij}, \alpha(\mu_{ij}))$$

$$\log \mu_{ij} \sim \beta_i^0 + \beta_i^T x_j^T$$

Expression in control

Change for treatment

Alternative Exon Usage / DEXSeq

$$N_{ijl} \sim NB(s_j \mu_{ijl}, \alpha(\mu_{ijl}))$$

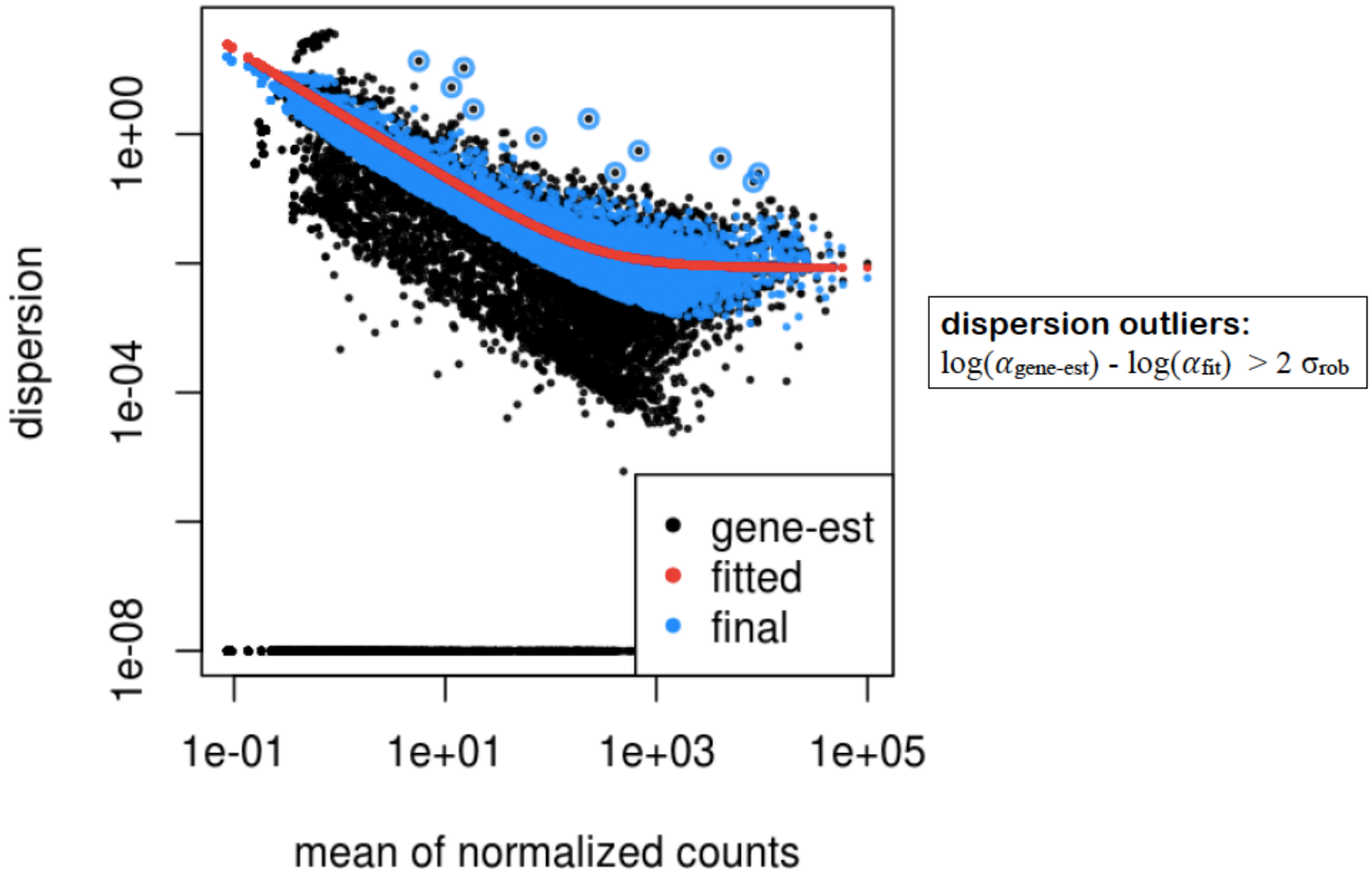
$$\log \mu_{ijl} \sim \beta_i^0 + \beta_{il}^E x_j^E + \beta_{ij}^T x_j^T + \beta_{ijl}^{ET} x_l^E x_j^T$$

Fraction of reads falling onto exon *l* in control

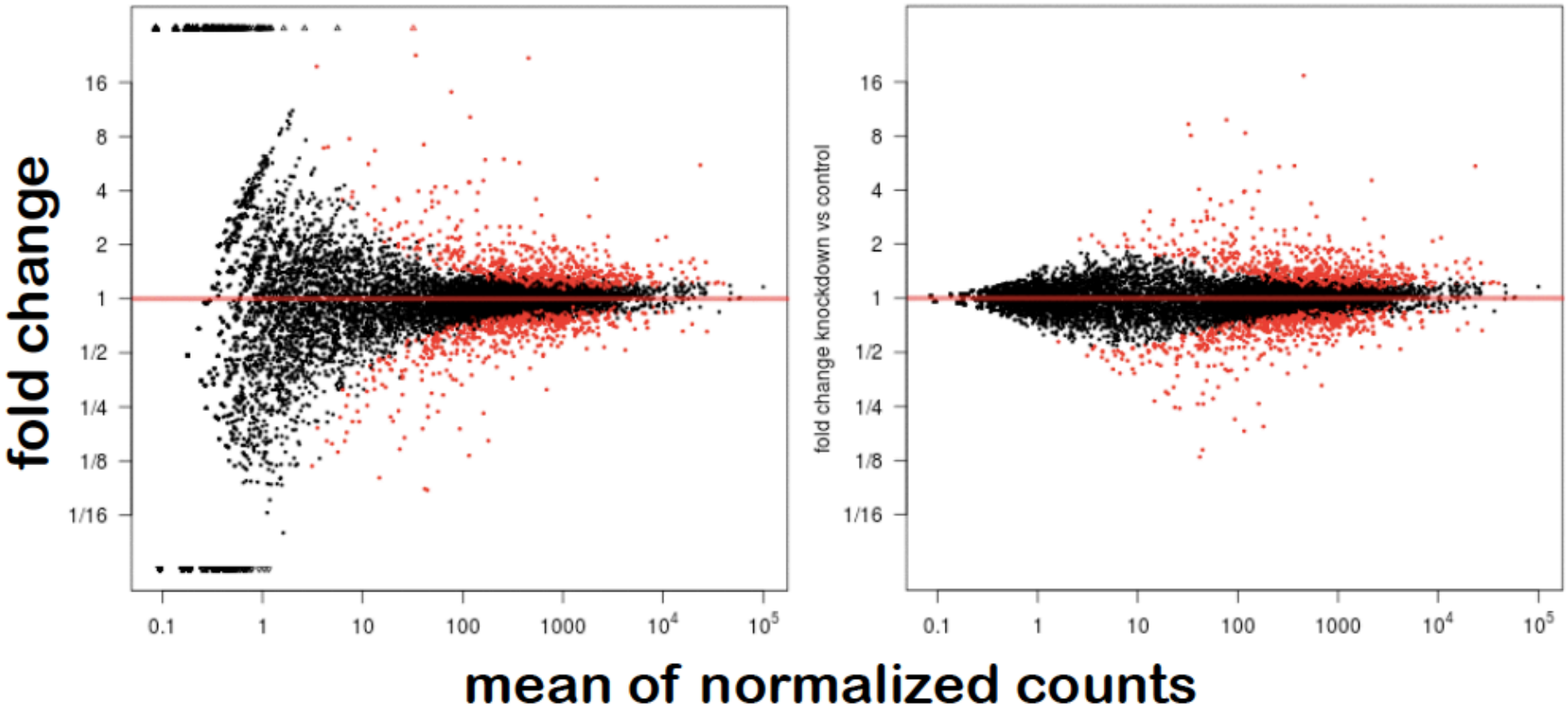
Change to fraction of reads for exon *l* due to treatment

Variance Shrinkage

Dispersion estimation: shrinkage

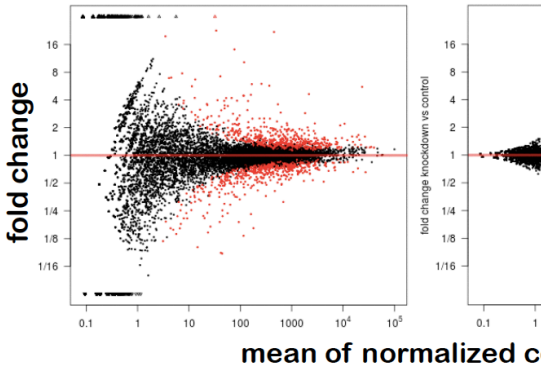


Downstream Effect of Shrinkage

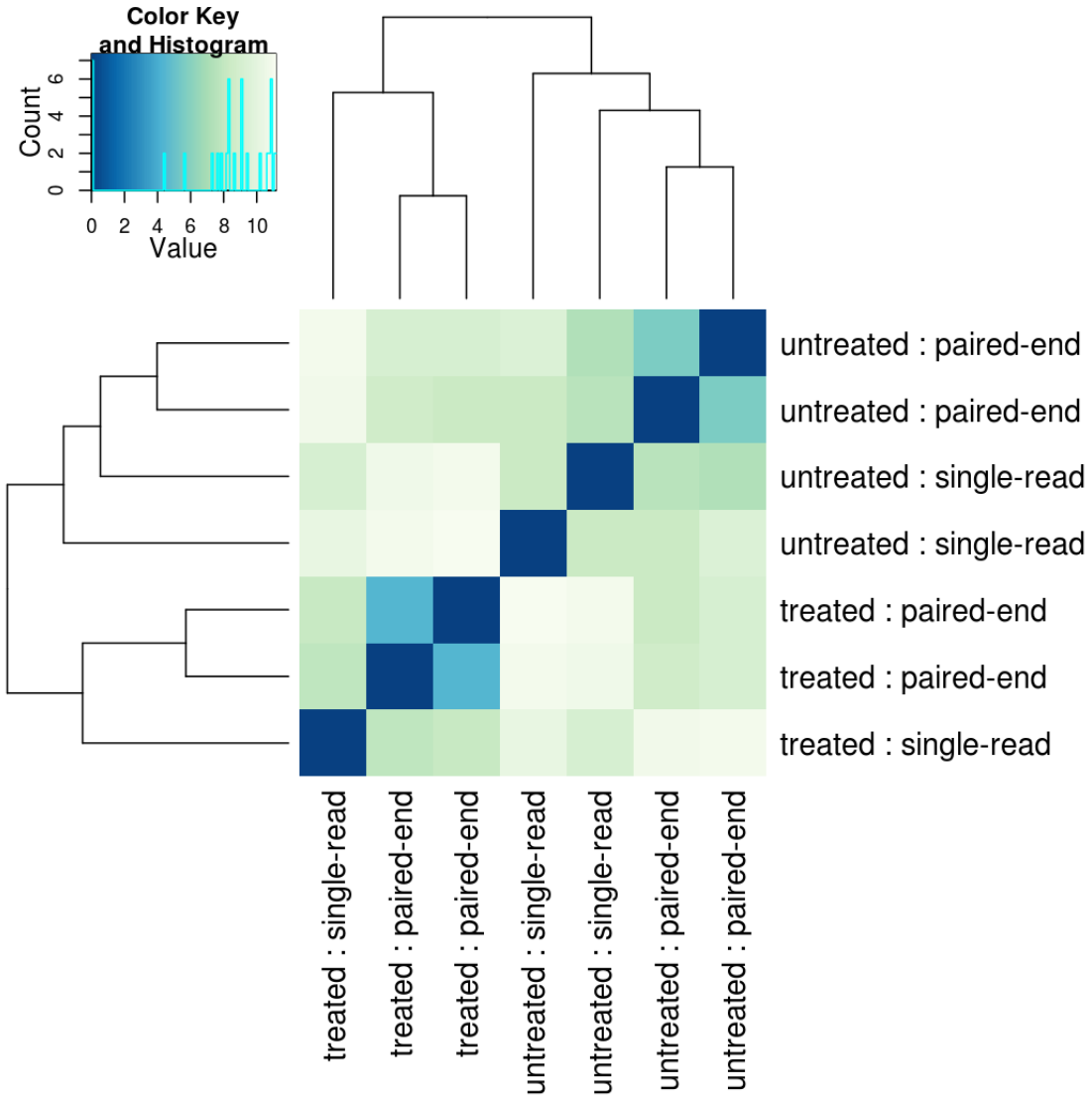


Remember the variance effect!

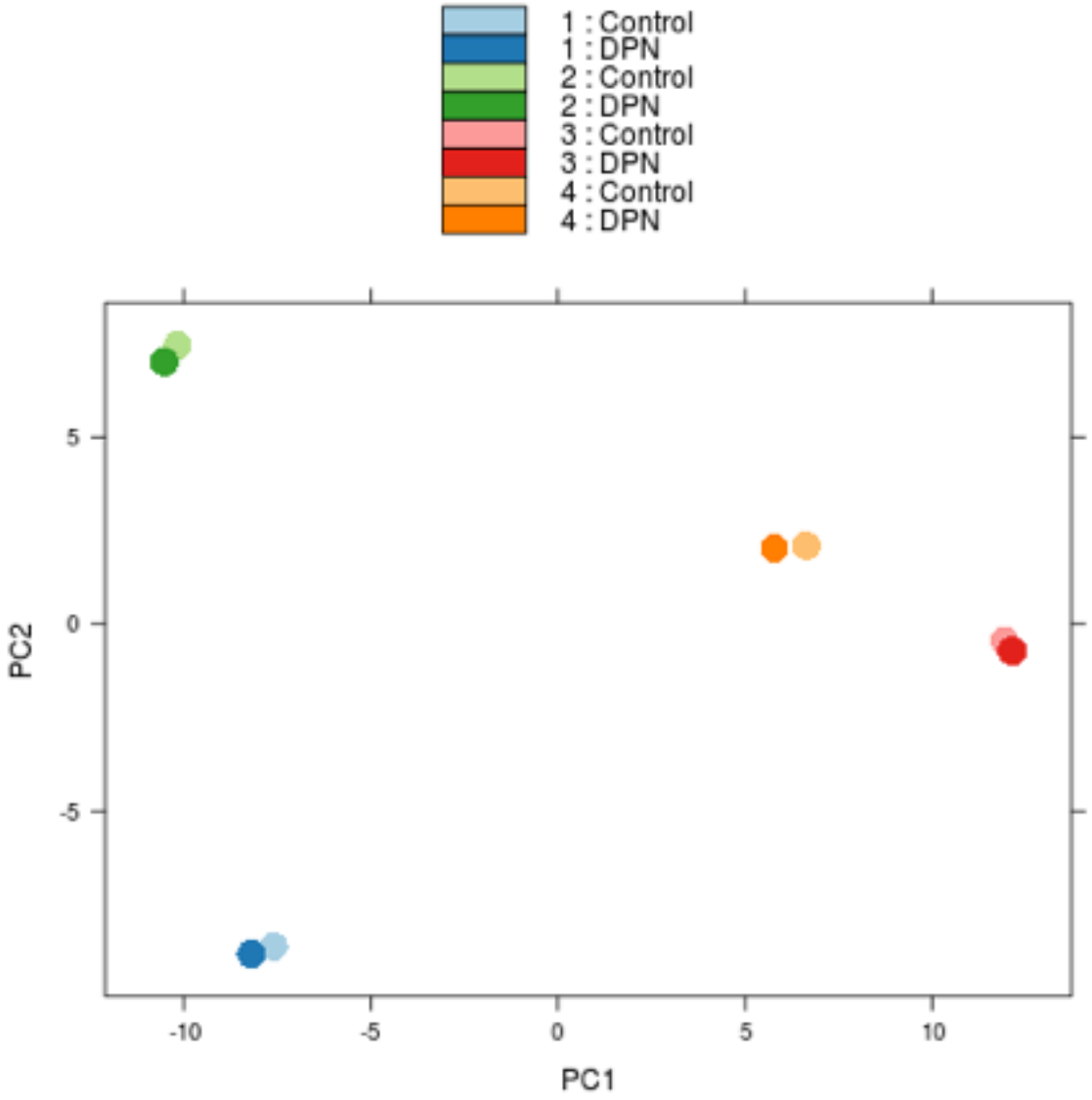
- Variance changes as mean changes…
- This seriously affects visualization;
- It also interferes with comparisons;
- One needs to adjust variance before performing clustering, visualization, PCA;
- DESeq2 has a “regularized log-transformation” method designed for that.



Clustering


Color Key
and Histogram
Count
0
2
4
6
0
2
4
6
8
10
Value
untreated : paired-end
untreated : paired-end
untreated : single-read
untreated : single-read
treated : paired-end
treated : paired-end
treated : single-read
treated : single-read
treated : paired-end
treated : paired-end
untreated : single-read
untreated : single-read
untreated : paired-end
untreated : paired-end

PCA


1 : Control
1 : DPN
2 : Control
2 : DPN
3 : Control
3 : DPN
4 : Control
4 : DPN
PC2
PC1
5
0
-5
-10
-5
0
5
10

The Truth Statistical Models

- There is no “correct model”;
- Models are approximations of the truth;
- There is a “useful model”;
- Understand the mechanisms of the system for better choices of model alternatives;

THINGS THAT STATISTICIAN SAYS…

The Experiment

- A procedure used to answer the questions;
- Comprised of multiple items:
  - Population;
  - Sample;
  - Hypotheses;
  - Test statistic;
  - Rejection criteria;

Population

- Superset of subjects of interest;
- Ideally, every subject in the population is surveyed;
- Issues with the “census approach”;

Sample

- Select some subjects from the population;
- We refer to this subset as sample;
- Subject in a sample can be called replicate;
- Replicate: technical vs. biological;

Hypotheses

- Sets that define the “underlying truth”;
- Null Hypothesis (H0): default situation.
  - Cannot be proven;
  - Reject (in favor of H1) vs. fail to reject;
- Alternative Hypothesis (H1): alternative (duh!)
  - Complements H0 on the parametric space;
  - Assists on the definition of the rejection criteria.

Examples of Hypotheses

- Comparing expression: Tumor vs. Normal:
  - Expressions on tumor and normal are the same;
  - Expressions on tumor and normal are different;

$$H_0 : \mu_T = \mu_N$$
$$H_1^a : \mu_T > \mu_N$$
$$H_1^b : \mu_T < \mu_N$$

$$H_0 : \mu_T = \mu_N$$
$$H_1 : \mu_T \neq \mu_N$$

Test Statistic

- Summary of the data;
- Built “under H0”;
- Independent of unknown parameters;
- Known distributions;
- Compatibility between data and H0;

Test Statistic

- What the statistician see…

$$X_{T,i} \sim N(\mu_T, \sigma^2) \qquad \bar{X}_T \sim N(\mu_T, \sigma^2/n)$$
$$X_{N,i} \sim N(\mu_N, \sigma^2) \qquad \bar{X}_N \sim N(\mu_N, \sigma^2/n)$$

If $H_0 : \mu_T = \mu_N$

Then $Z = \dfrac{\bar{X}_T - \bar{X}_N}{\sqrt{2\sigma^2/n}} \sim N(0,1)$

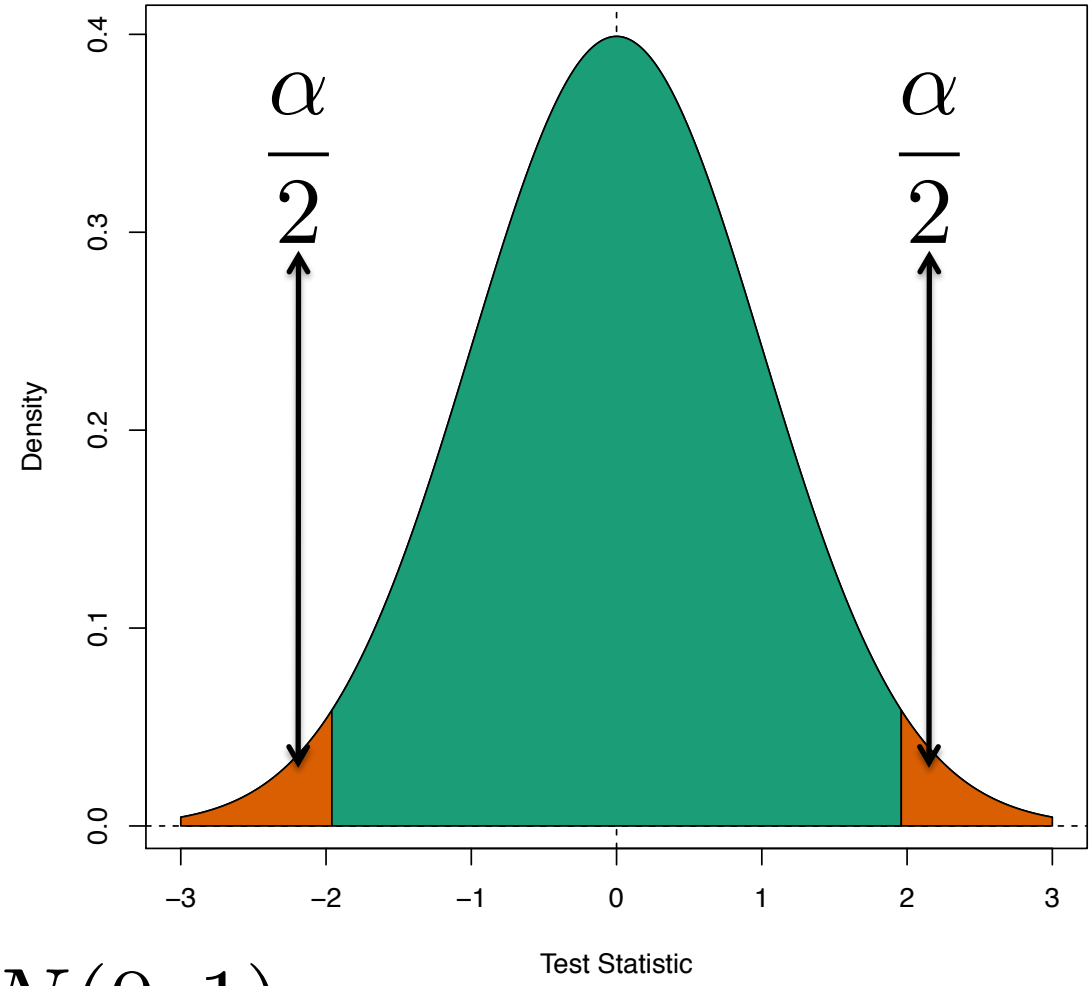
Rejection Criteria

- Function of three factors:
  - Test statistic;
  - Hypotheses;
  - Type I Error (False Positive), α;
- Determines thresholds used to reject H0:
- Defines what is “extreme” for the experiment;

Rejection Criteria

$$H_0 : \mu_T = \mu_N$$

$$H_1 : \mu_T \neq \mu_N$$



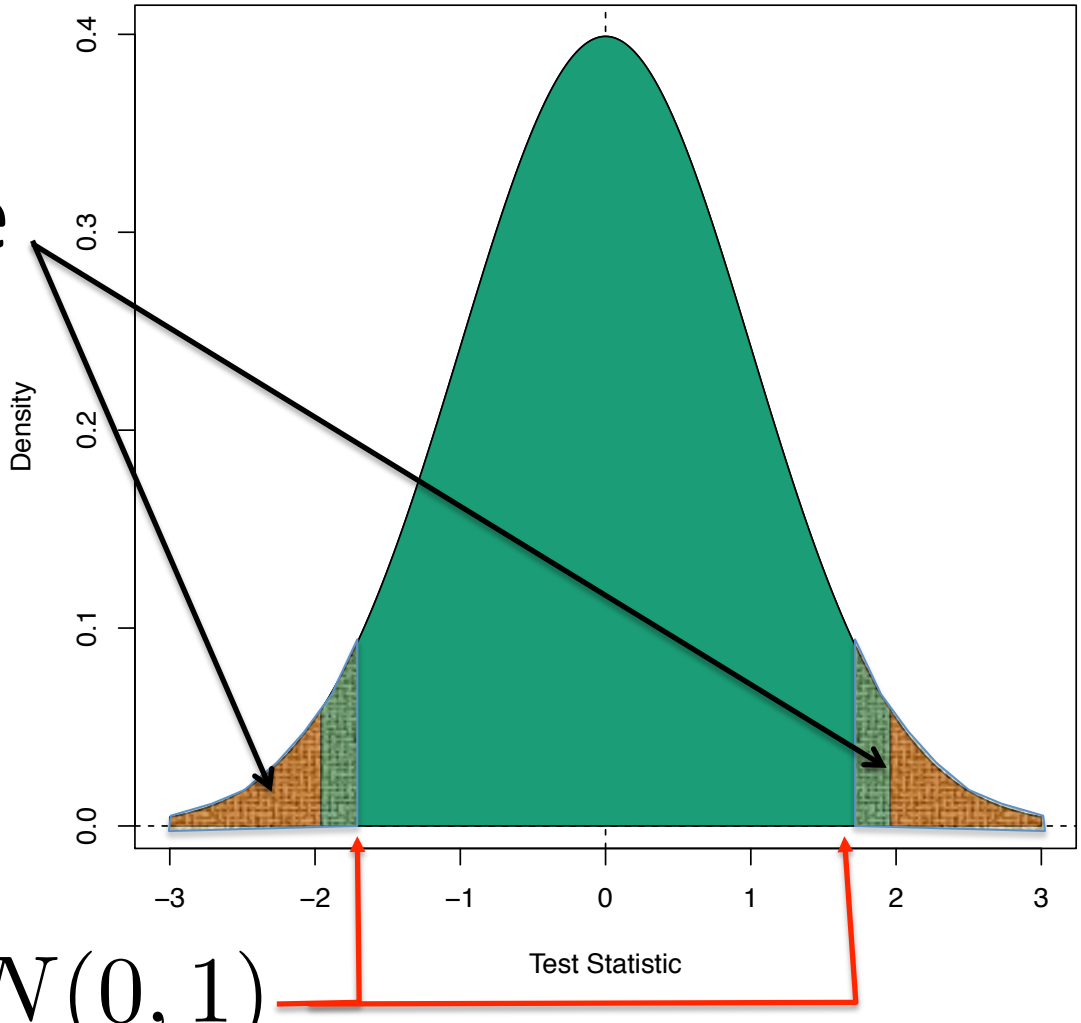
$$Z = \frac{\bar{X}_T - \bar{X}_N}{\sqrt{2\sigma^2/n}} \sim N(0,1)$$

From Rejection Criteria to P-value!

p-value

$$H_0 : \mu_T = \mu_N$$

$$H_1 : \mu_T \neq \mu_N$$



$$Z = \frac{\bar{X}_T - \bar{X}_N}{\sqrt{2\sigma^2/n}} \sim N(0,1)$$

What if we look at multiple p-values at a time?

- On a Gene Expression study, we test often 20K genes for differential expression;
- Each test leads to one p-value;
- Should we trust the p-values in order to make decisions?

What if we look at multiple p-values at a time?

- Can we simulate this?
- Choose an α–level;
- Generate two populations with the same pars;
- Run t-test;
- Is the result smaller than α?
  - Yes: reject;
  - No: don't reject;

Multiple Testing

- We are doing high-throughput experiments;
- Comparing thousands of units simultaneously;
- At this scale, we can observe several instances of rare events **just by chance**:
  - Event A: 1 in 1000 chance of happening;
  - Event B: 999 in 1000 chance of happening;
  - And the experiment is tried 20,000 times;
  - We expect 20 occurrences of Event A to be observed, although Event B is much more likely;

Multiple Testing

- Similar scenario, for example, with DE;
- Most genes are not differentially expressed;
- High-throughput experiments;
- Differential expression is tested for 20K genes;
- Need to protect against false positives;
- Suggestion:
  - use non-specific filtering;
  - use adjusted p-values;

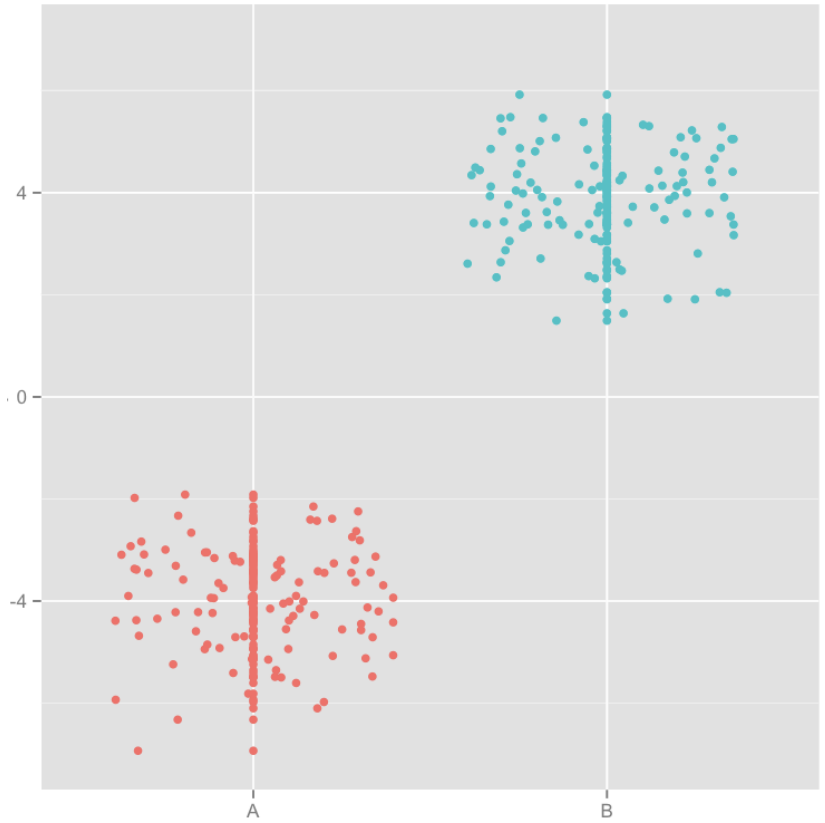
Type I and Type II Errors

Non-Specific Filtering

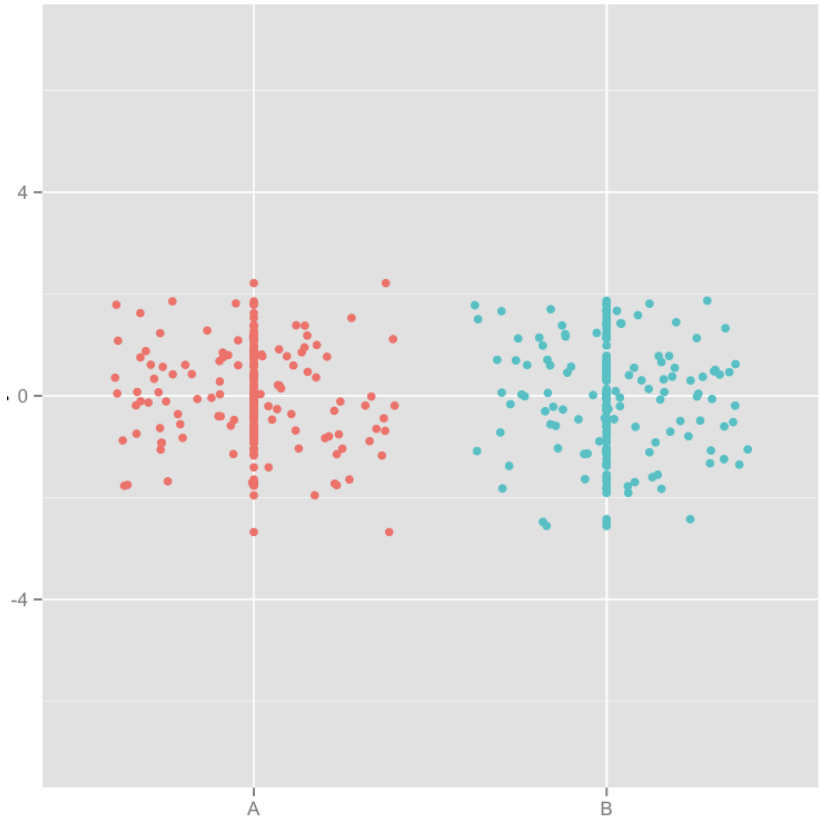
- The majority of the genes are not differentially expressed – this is the basic hypothesis for normalization;
- If we reduce the number of genes to be tested, the chance of making a wrong decision is reduced;
- Non-Specific filtering refers to removing genes that are clearly not DE without looking at the phenotypic information of the samples;

Using Variance as a Filter

Differentially Expressed



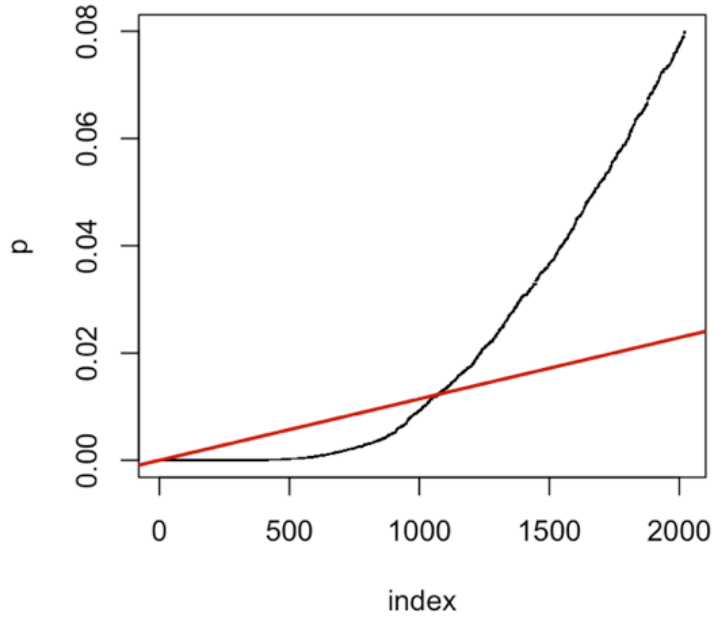
Not-Differentially Expressed



FDR – Benjamini Hochberg (BH)

- Sort the p-values by magnitude;
- Get the adjusted values by

$$j^* = \max\left\{ j : p_j \leq \frac{j}{m}\alpha \right\}$$



ADDITIONAL STUFF TO REMEMBER!

Useful Facts

- The Law of the Large Numbers guarantees that the larger the sample size is, the closer the sample average is to the actual mean;
- Normality assumption isn't that important with large sample size;
- The Central Limit Theorem states that the **average** is asymptotically normal;

Useful Facts

- The Z-score depends on the precise knowledge of the variance term:

$$Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

- Estimating the variance changes the distribution of the test statistic:

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \sim t_n$$

Useful Facts

- The Student's *t* distribution is similar to the Normal distribution, but has heavier tails;
- Larger sample size, more d.f.;
- More d.f., closer to Normal;

DO I REALLY NEED A STATISTICIAN BEFORE I EVEN RUN MY EXPERIMENT?

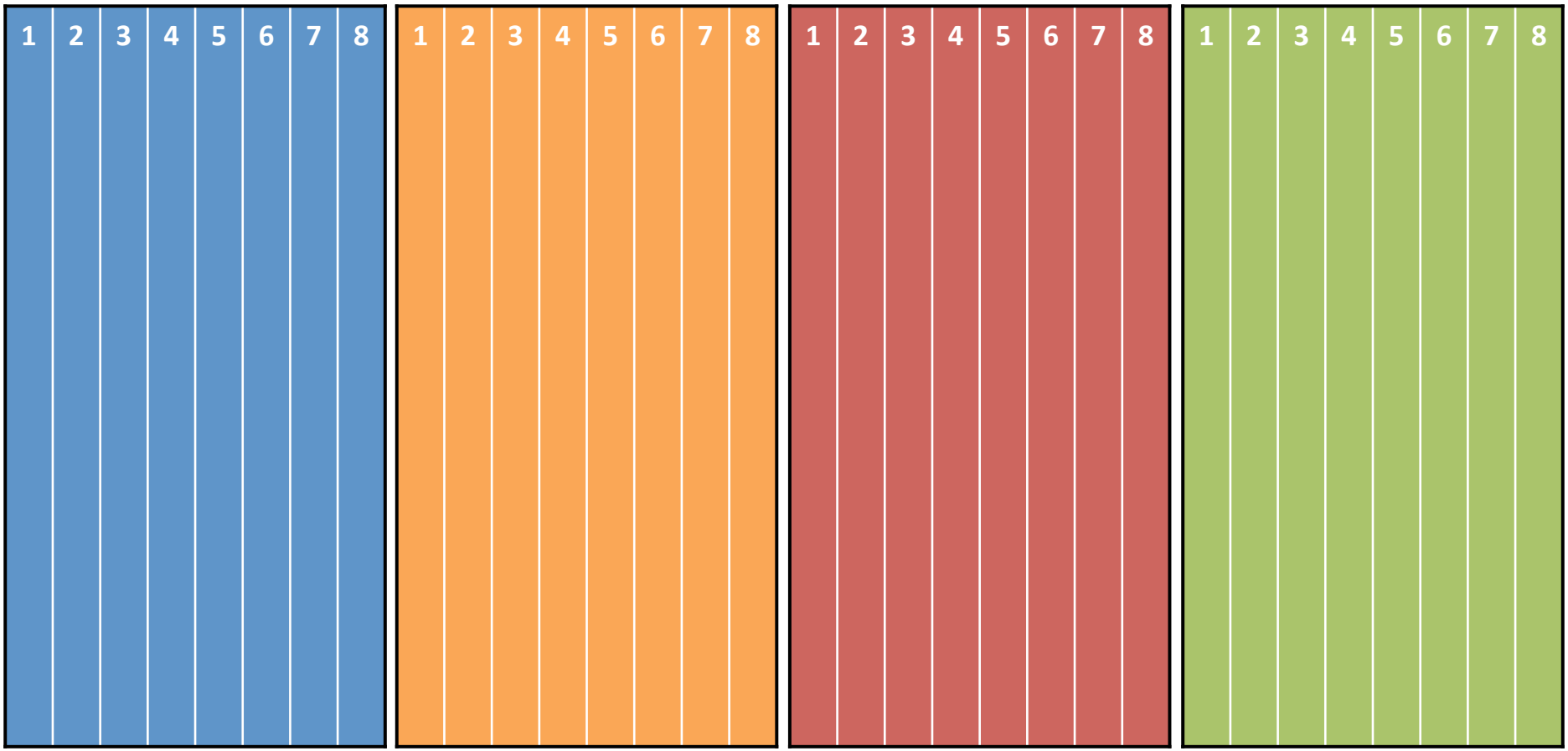
Sample size is crucial

- The larger, the better;
- Ideal N = ($$ I have) / ($$ it costs)
- With differential expression, one can observe this more easily;
- RNASeqPower BioConductor package;

About Technology

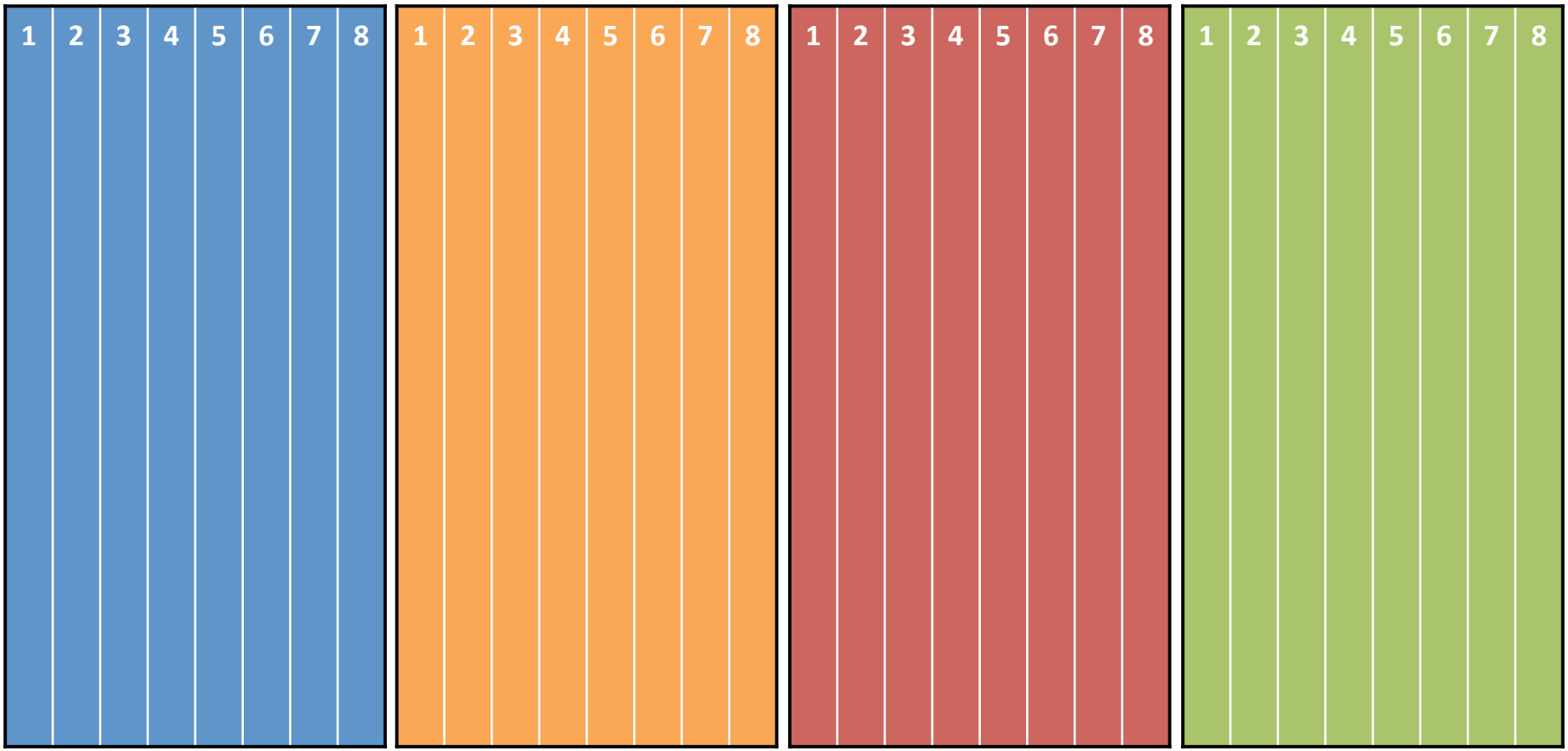
- Is RNA-Seq really worth it when we consider:
  - Cost,
  - Strategies for analysis, and
  - Technical requirements?

Can my experiment answer the question of interest?



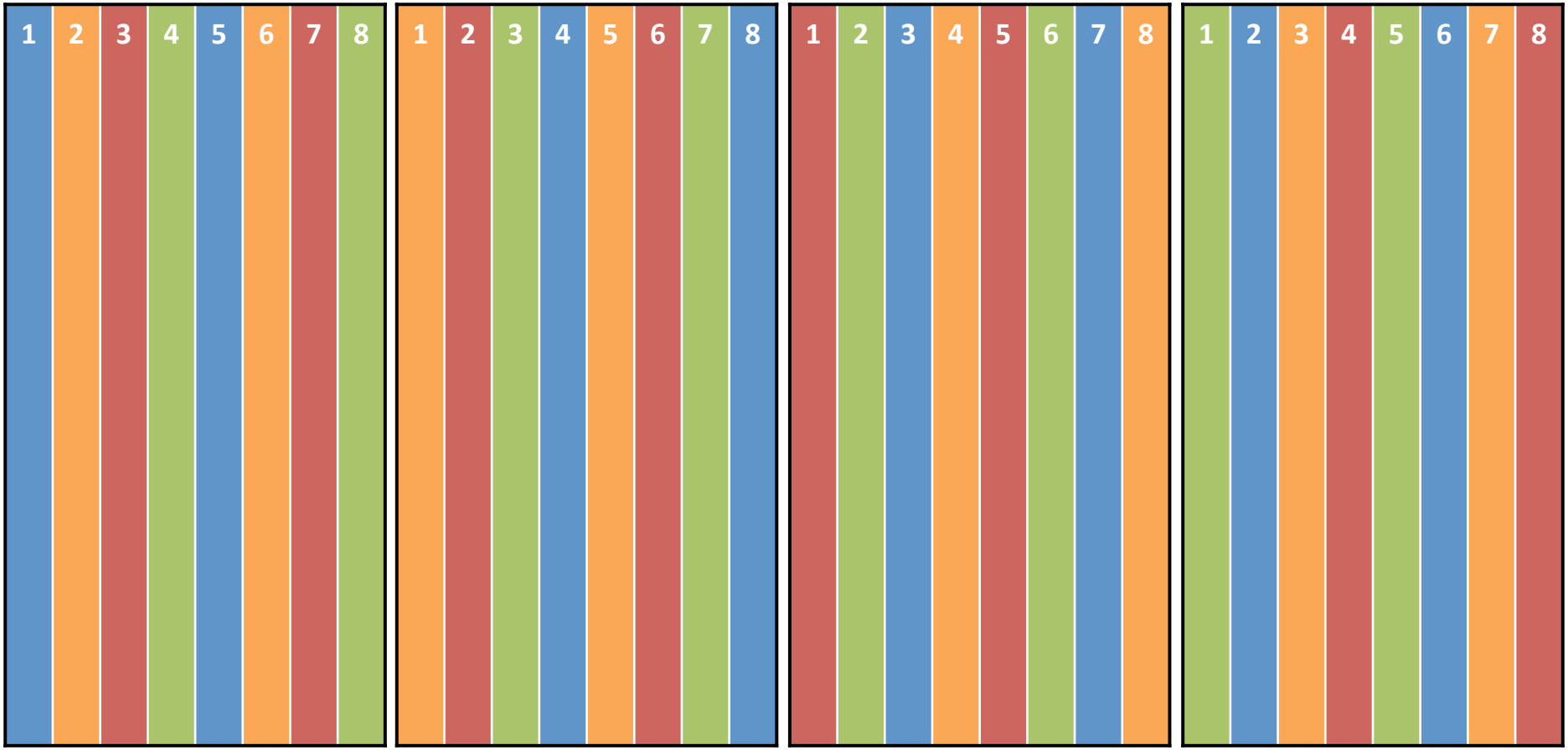
Differential Expression Across Groups

Flow Cell Confounded With Group



Differential Expression Across Groups

Randomize Samples wrt Flow Cell



Differential Expression Across Groups

Barcoding vs. Lane Effect