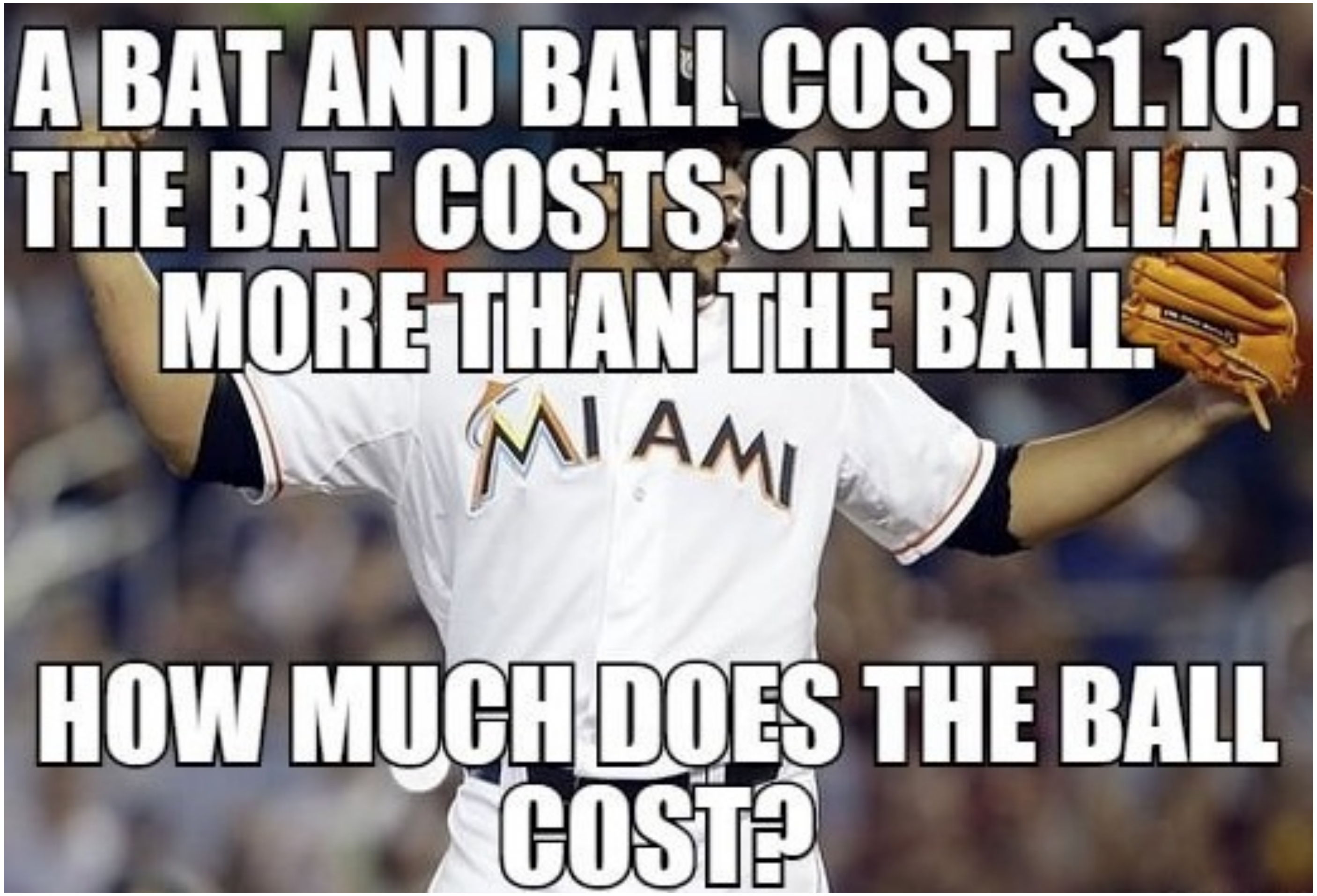
Introduction to RNA-Seq Data Analysis

Dr. Benilton S Carvalho

Department of Statistics

University of Campinas



A BAT AND BALL COST $1.10.
THE BAT COSTS ONE DOLLAR
MORE THAN THE BALL.

HOW MUCH DOES THE BALL
COST?

Tools of Choice

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

Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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

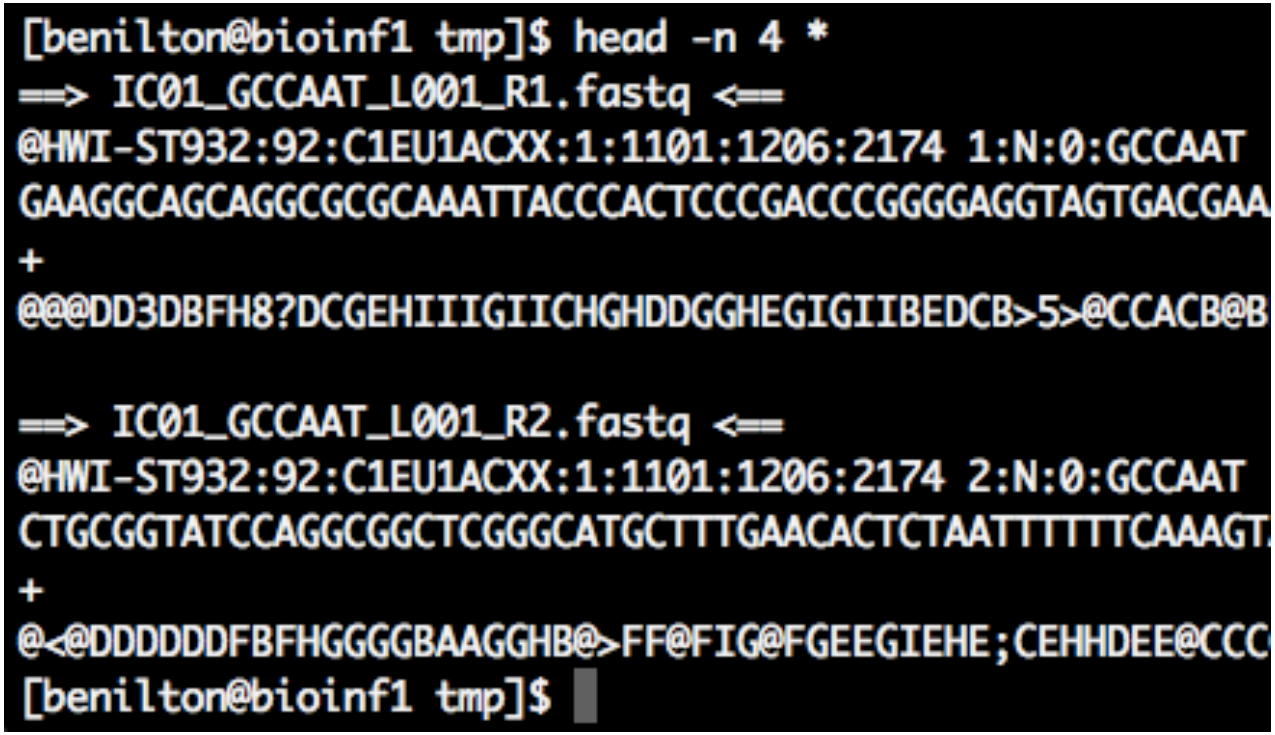
- Software infra-structure that uses R;
- Designed for biological data;
- Hundreds of packages:
  - Mass spectrometry;
  - Microarrays;
  - Next Generation Sequencing (NGS);
- Active community:
  - Heavily used by industry;
  - Releases in April and October;
  - Cutting-edge methods.

RAW DATA

Inside a FASTQ File

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

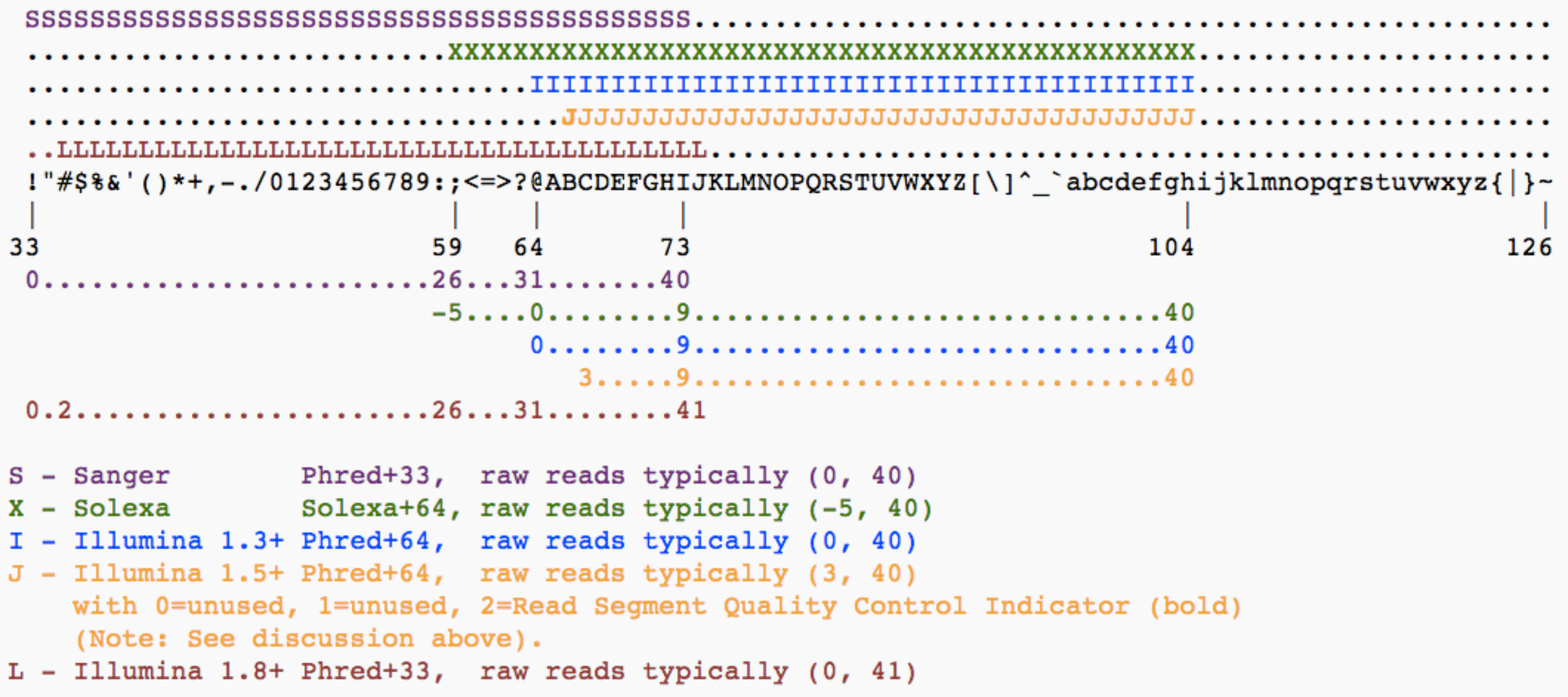
Mate
Fail filter
Control bits
Index seq



```
[benilton@bioinf1 tmp]$ head -n 4 *
==> IC01_GCCAAT_L001_R1.fastq <==
@HWI-ST932:92:C1EU1ACXX:1:1101:1206:2174 1:N:0:GCCAAT
GAAGGCAGCAGGCGCGCAAATTACCCACTCCCGACCCGGGGAGGTAGTGACGAA
+
@@@DD3DBFH8?DCGEHIIIGIICHGHDDGGHEGIGIIBEDCB>5>@CCACB@B

==> IC01_GCCAAT_L001_R2.fastq <==
@HWI-ST932:92:C1EU1ACXX:1:1101:1206:2174 2:N:0:GCCAAT
CTGCGGTATCCAGGCGGCTCGGGCATGCTTTGAACACTCTAATTTTTTCAAAGT
+
@<@DDDDDDFBFHGGGGBAAGGHB@>FF@FIG@FGEEGIEHE;CEHHDEE@CCC
[benilton@bioinf1 tmp]$ 
```

The Mistery of the Quality Scores

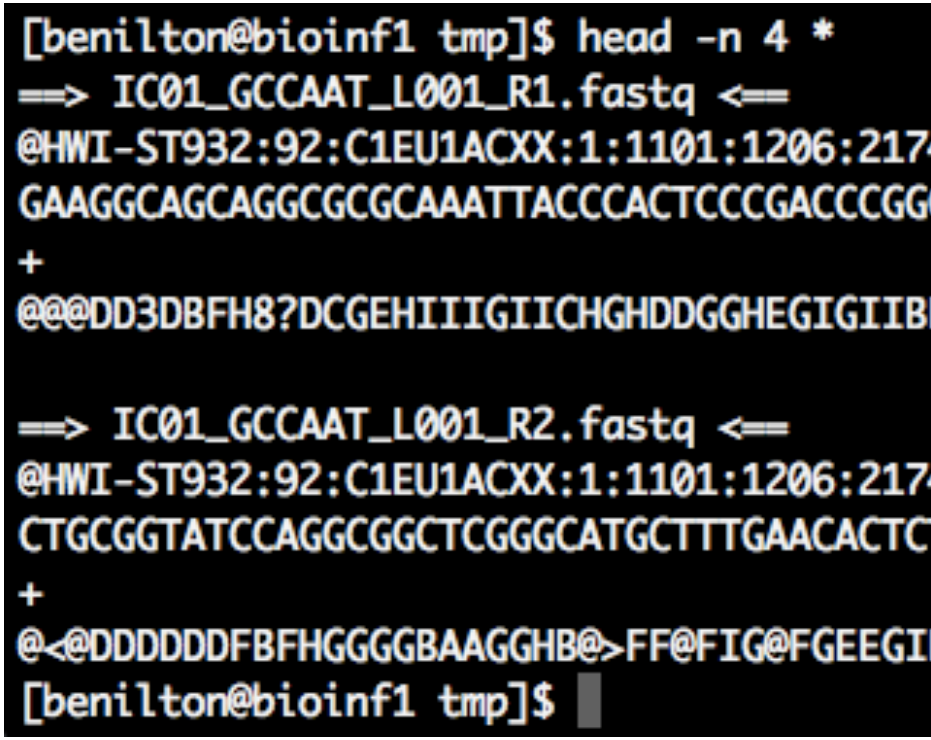


```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

The Mistery of Quality Scores

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



```
[benilton@bioinf1 tmp]$ head -n 4 *
==> IC01_GCCAAT_L001_R1.fastq <==
@HWI-ST932:92:C1EU1ACXX:1:1101:1206:217
GAAGGCAGCAGGCGCGCAAATTACCCACTCCCGACCCGG
+
@@@DD3DBFH8?DCGEHIIIGIICHGHDDGGHEGIGIIB

==> IC01_GCCAAT_L001_R2.fastq <==
@HWI-ST932:92:C1EU1ACXX:1:1101:1206:217
CTGCGGTATCCAGGCGGCTCGGGCATGCTTTGAACACTC
+
@<@DDDDDDFBFHGGGGBAAGGHB@>FF@FIG@FGEEGI
[benilton@bioinf1 tmp]$
```

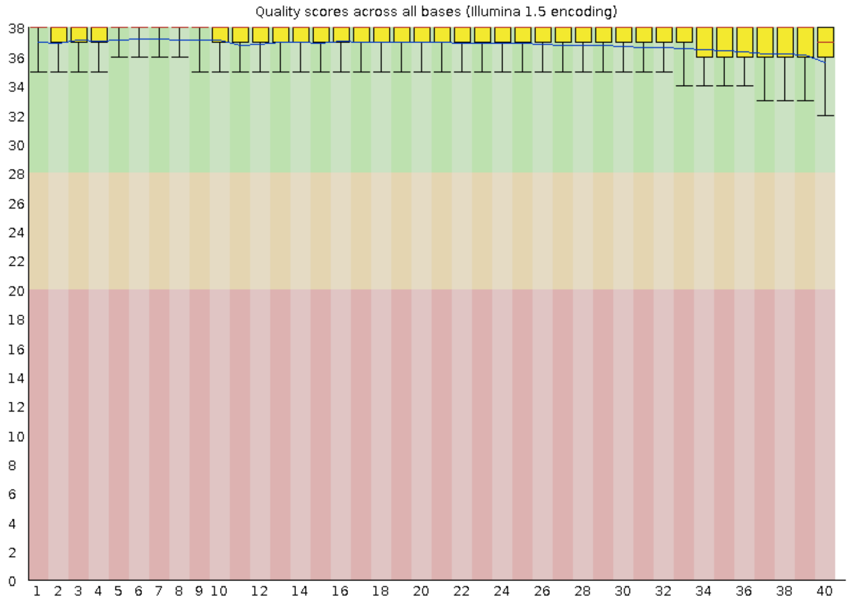
QUALITY ASSESSMENT

FastQC

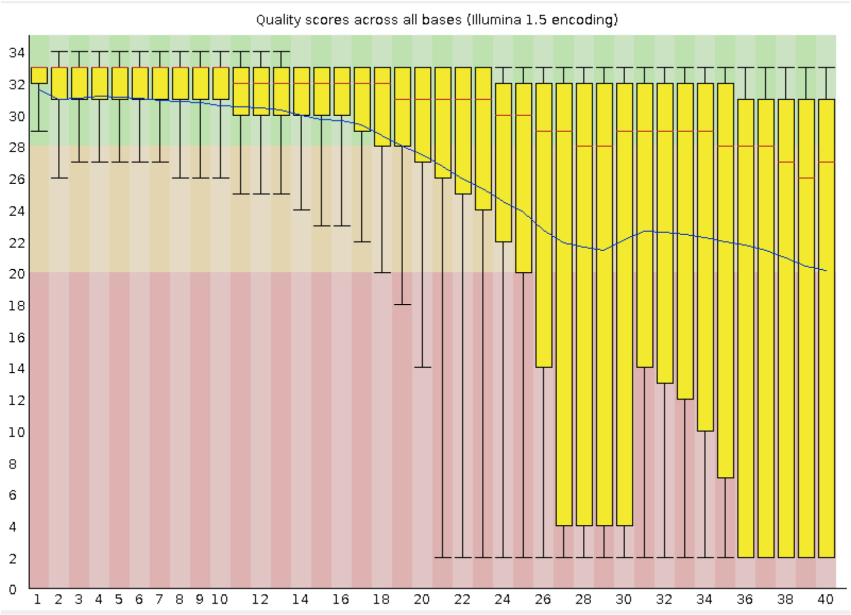
- We have experience with FastQC, but we are developing our own tool;
- FastQC is Java-based;
- Includes the option of pointing and clicking;
- http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/

FastQC – Per Base Seq Quality

Good

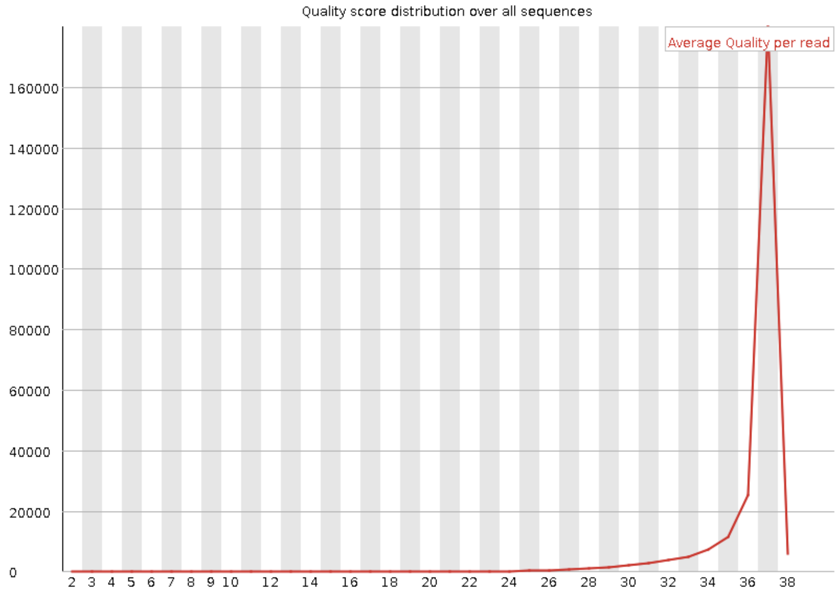


Poor

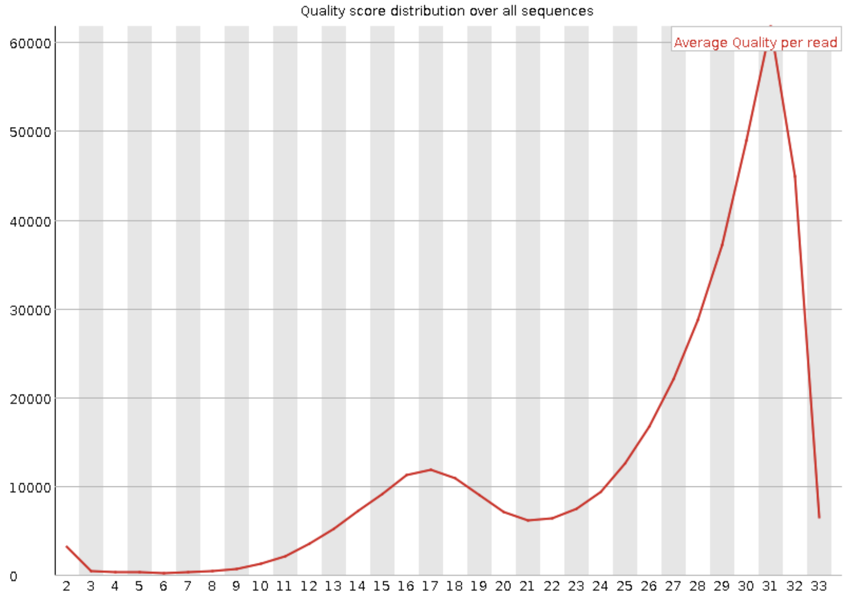


FastQC – Quality Score over All Seqs

Good

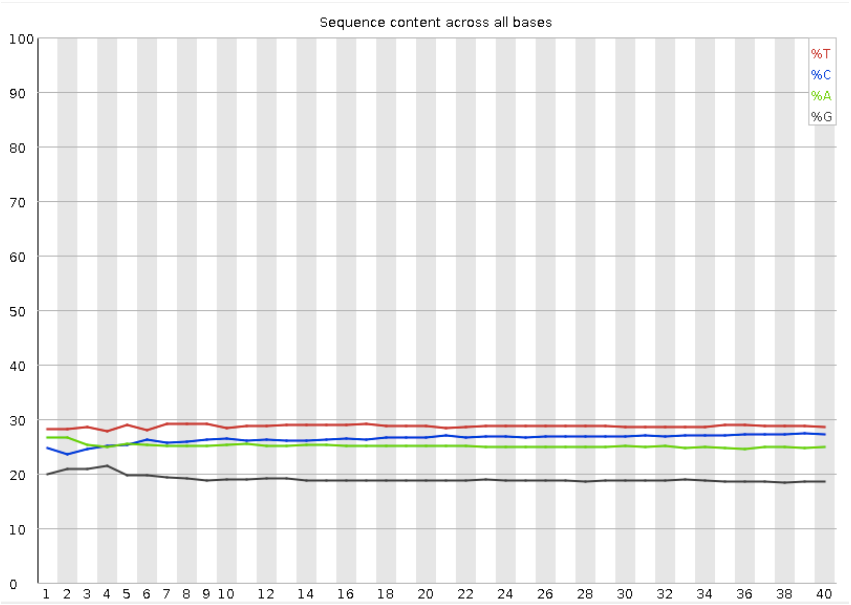


Poor

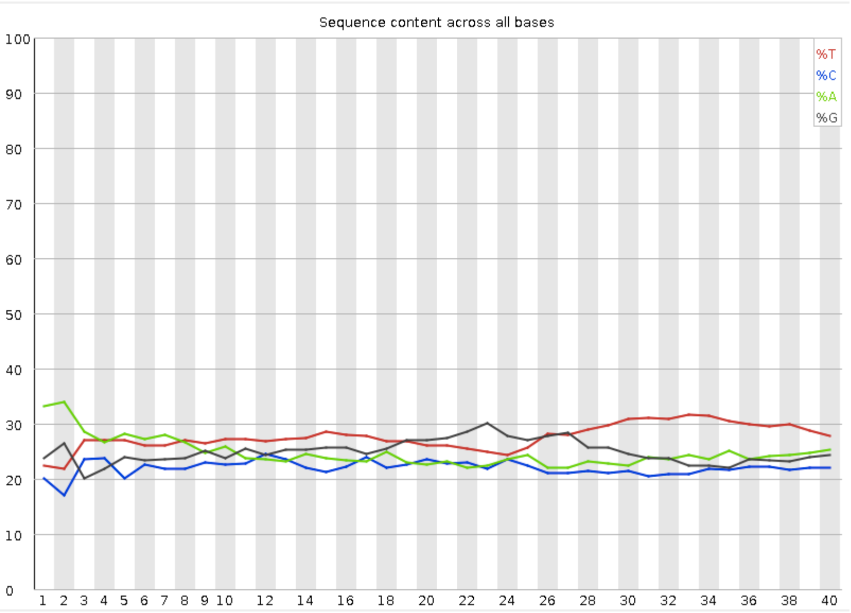


FastQC – Sequence Content

Good



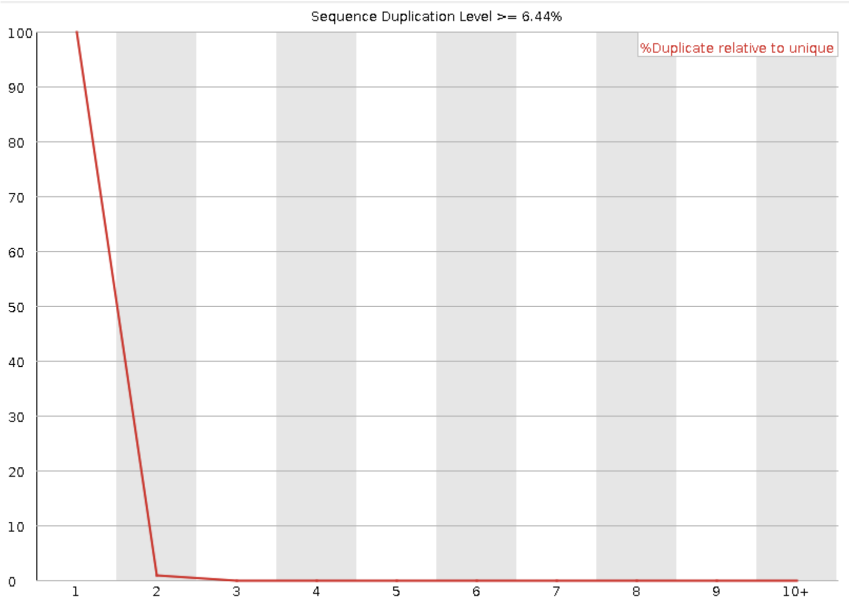
Poor

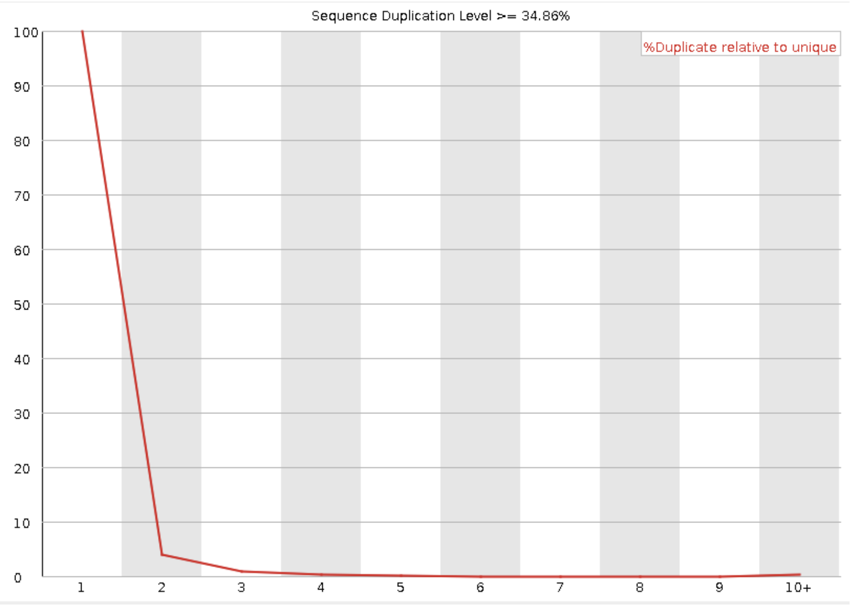


FastQC – Sequence Duplication

Good

Poor





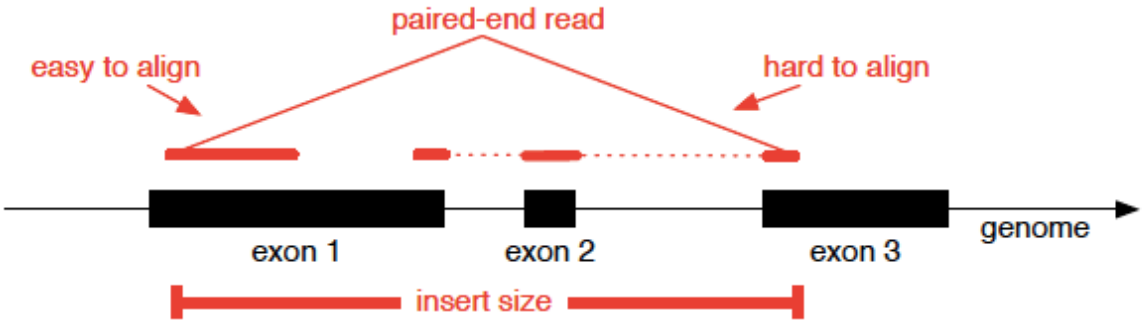
MAPPING

Principles of Mapping

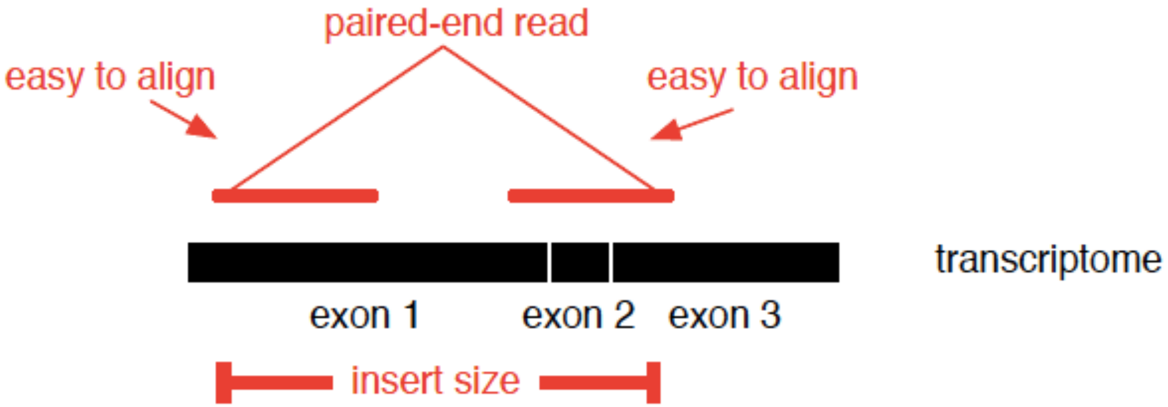
- Obtain the reference (genome or transcriptome) for the organism of interest:
- Mapping to the genome:
  - Allows for identification of novel genes/isoforms
  - Must allow for gaps (really hard)
- Mapping to the transcriptome:
  - Fast(er)
  - No need for spliced alignments
  - Can’t find novel genes/isoforms

Principles of Mapping

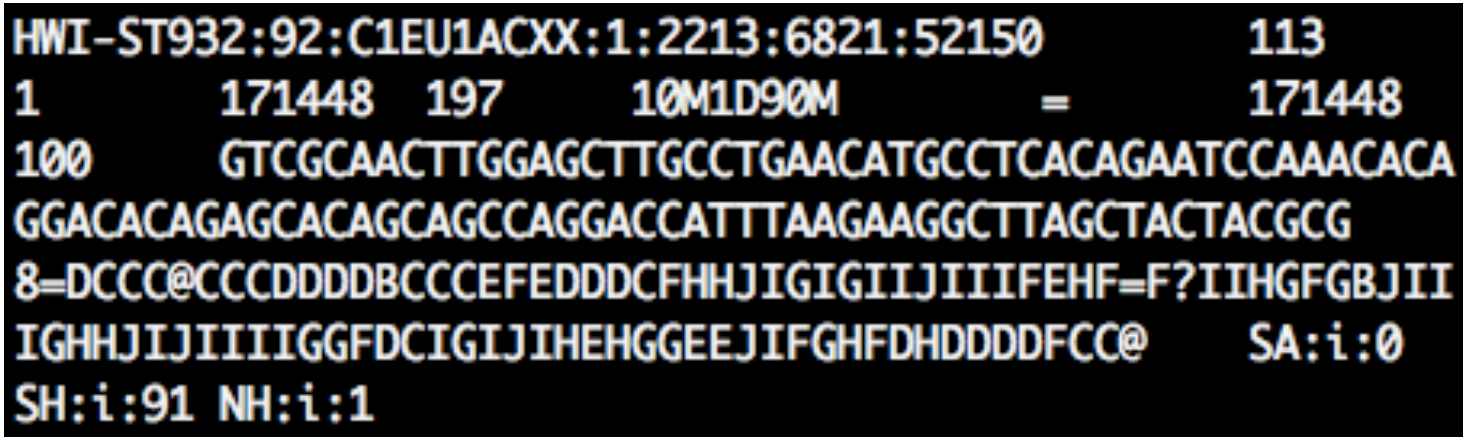
Genome alignment (e.g. align to 23 chromosomes):



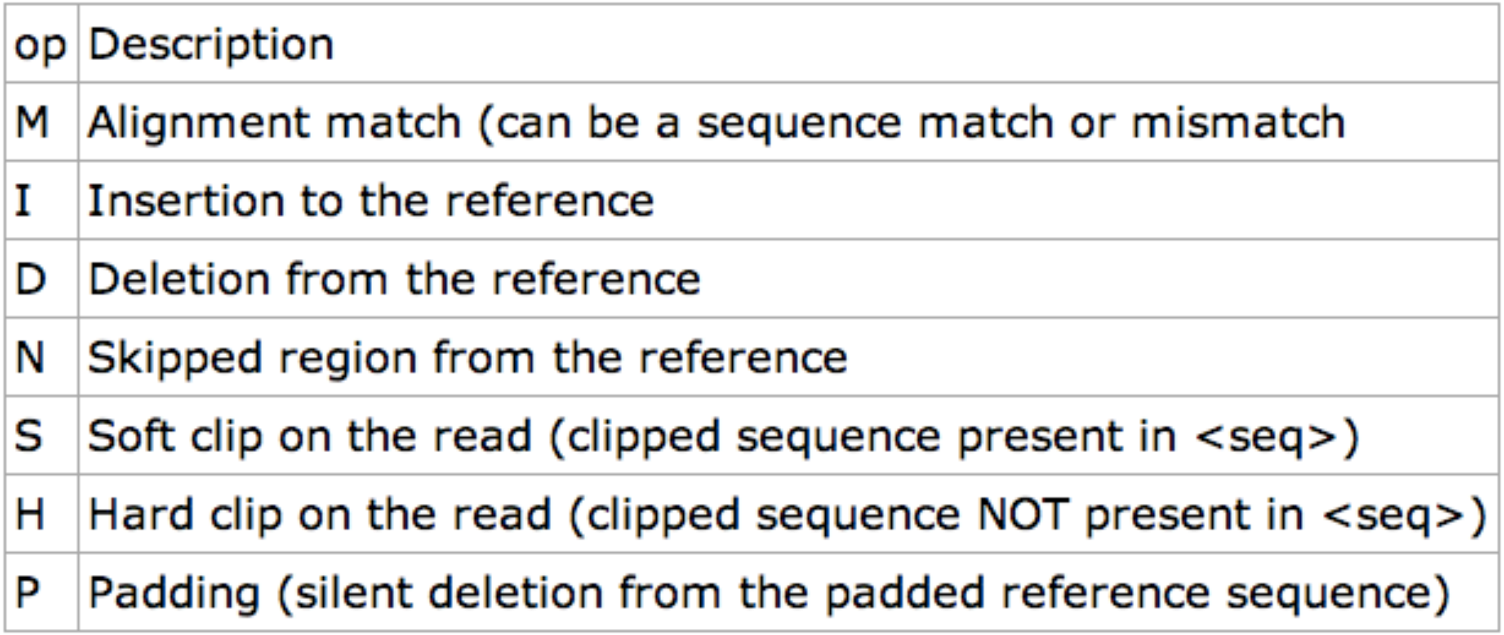
Transcriptome alignment (e.g. align to 150,000 *known* transcripts):



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
```



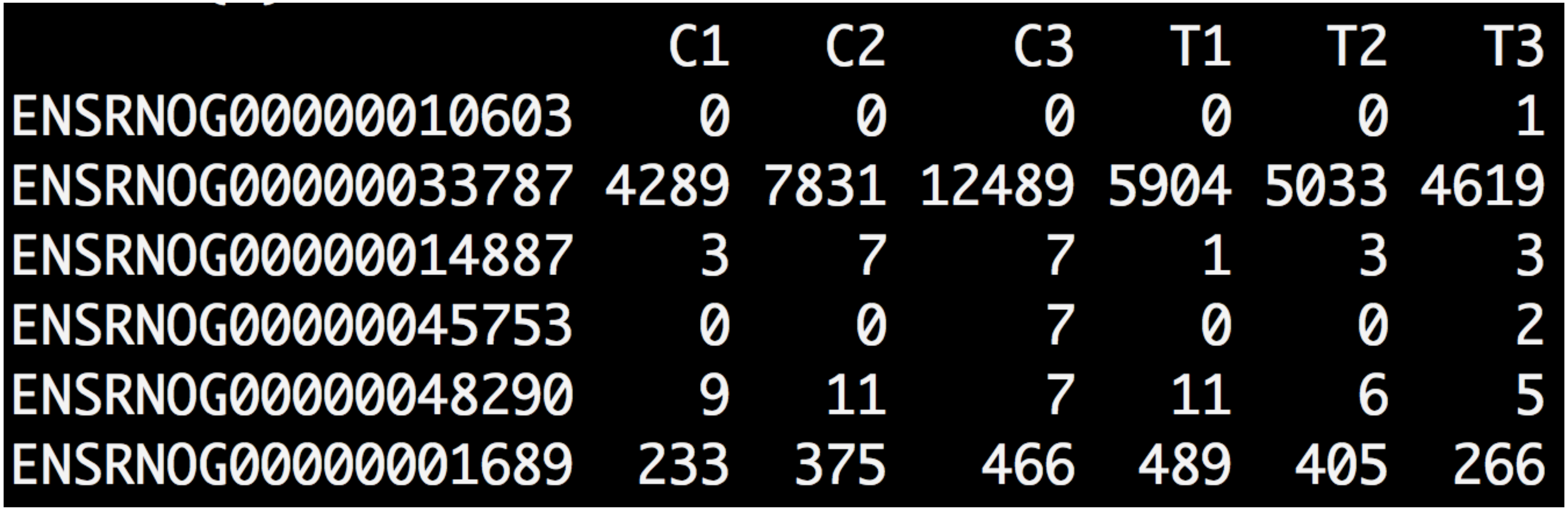
op	Description
M	Alignment match (can be a sequence match or mismatch
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent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

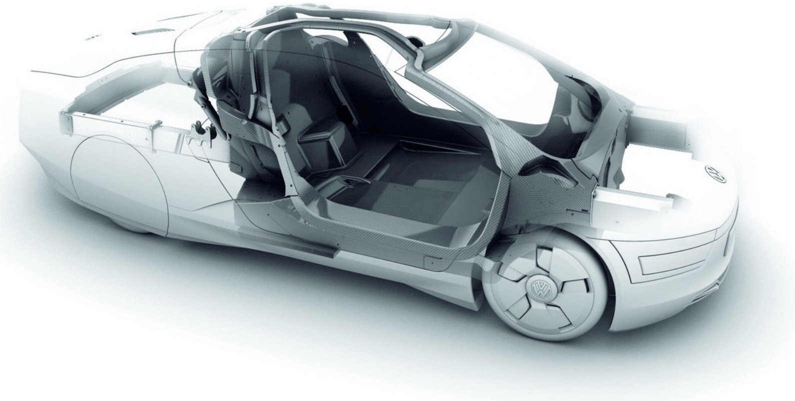
Count-table Example



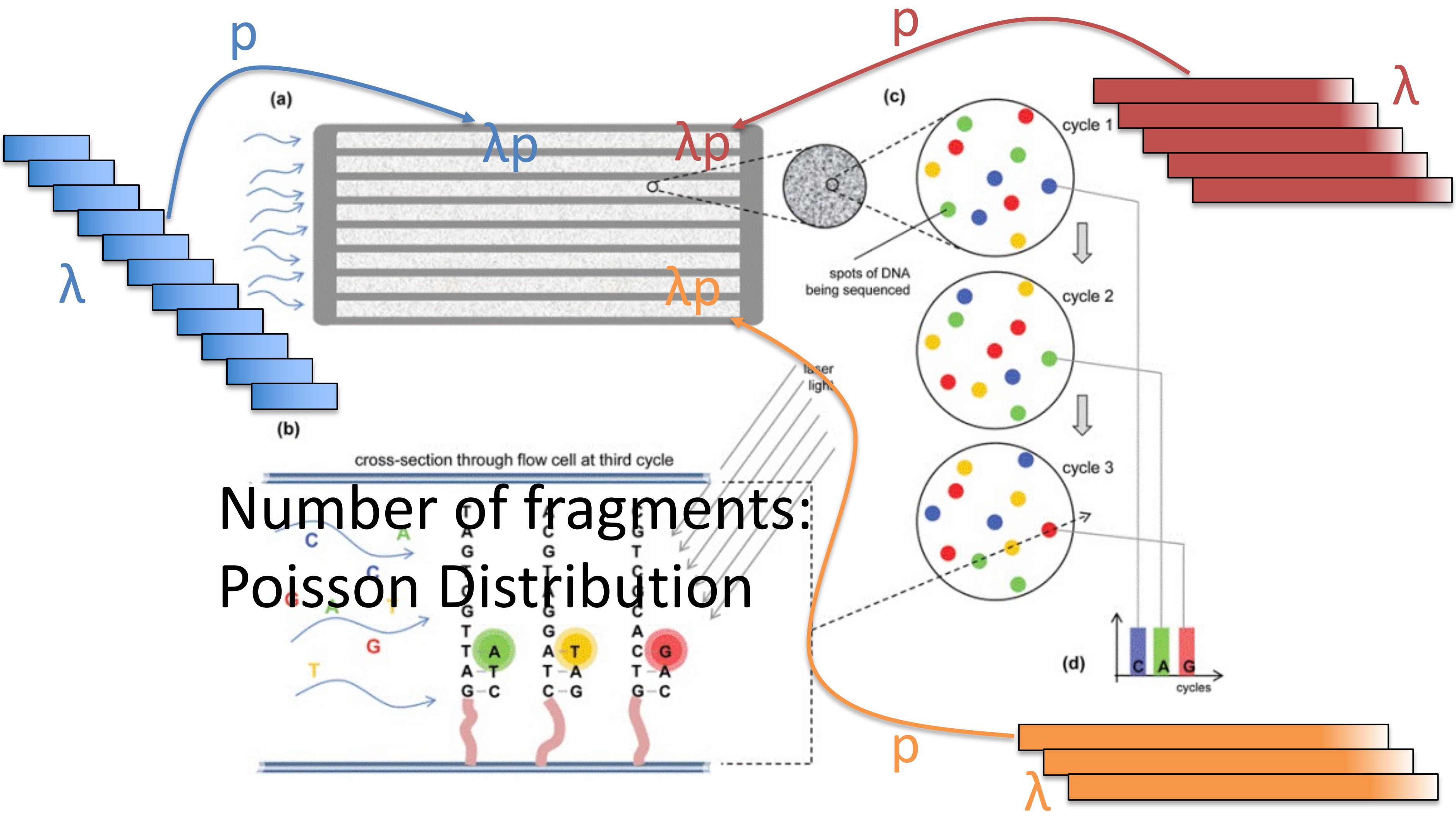
```
                     C1   C2    C3   T1   T2   T3
ENSRNOG00000010603    0    0     0    0    0    1
ENSRNOG00000033787 4289 7831 12489 5904 5033 4619
ENSRNOG00000014887    3    7     7    1    3    3
ENSRNOG00000045753    0    0     7    0    0    2
ENSRNOG00000048290    9   11     7   11    6  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

STATISTICAL MODELING

What is a model?

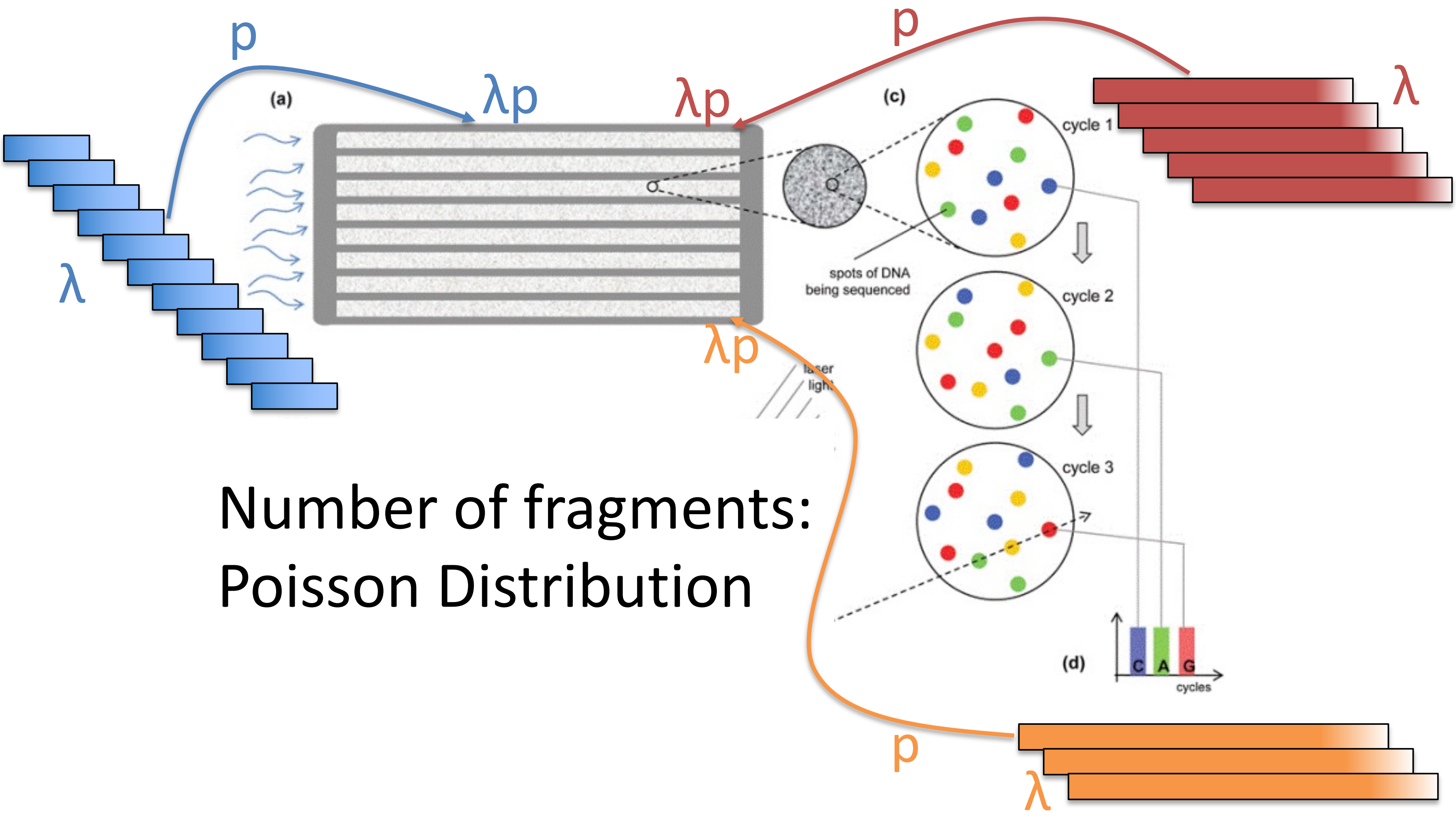


Different Transcripts, Rates and Probabilities


p
λ
λp
p
λ
λp
λp
p
λ
(a)
(b)
(c)
(d)
cycle 1
cycle 2
cycle 3
spots of DNA being sequenced
laser light
cross-section through flow cell at third cycle
C
A
G
cycles

Number of fragments:
Poisson Distribution

Different Transcripts, Rates and Probabilities



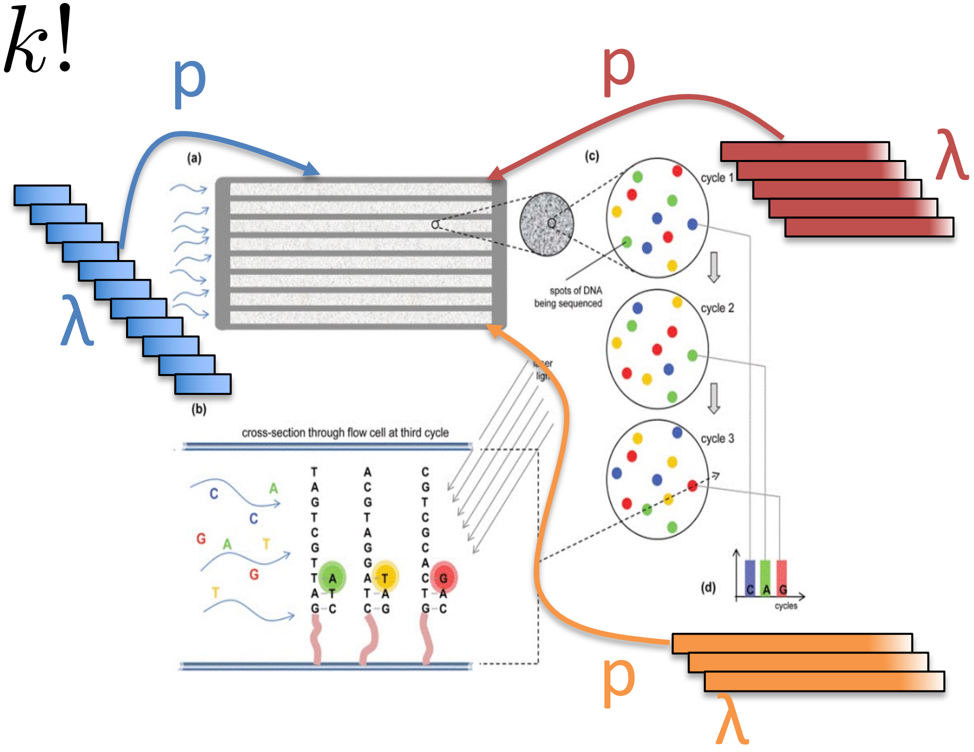
Number of fragments:
Poisson Distribution

Characteristics of a Poisson Distribution

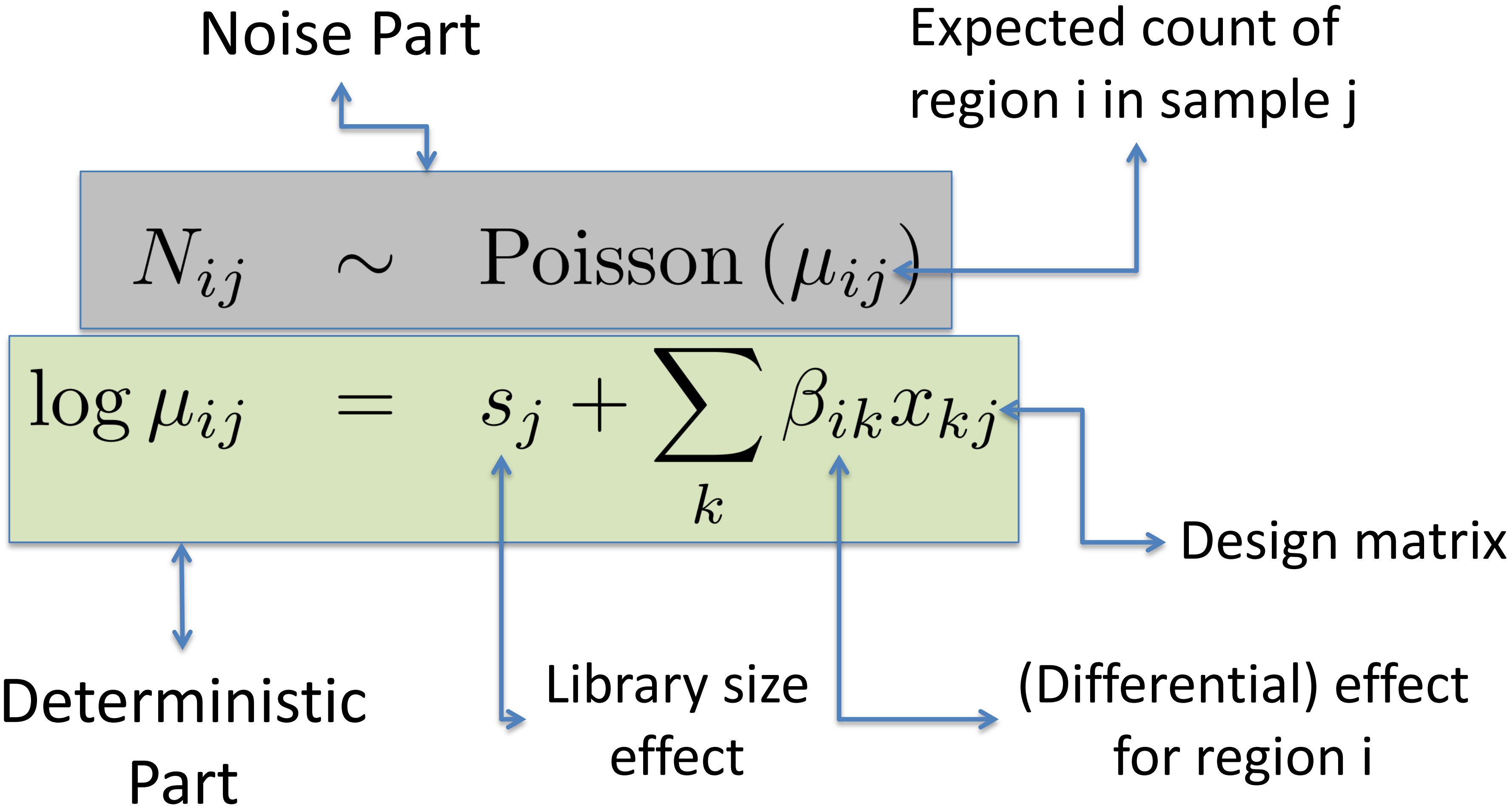
- X ~ Poisson(λp)

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

- Mean: λp
- Variance: λp


p
p
λ
λ
p
λ
(a)
(b)
(c)
(d)
cycle 1
cycle 2
cycle 3
spots of DNA being sequenced
cross-section through flow cell at third cycle
laser light
C
A
G
cycles

Analysis method: GLM



Noise Part

Expected count of region i in sample j

$$N_{ij} \quad \sim \quad \mathrm{Poisson}\left(\mu_{ij}\right)$$

$$\log \mu_{ij} \quad = \quad s_j + \sum_k \beta_{ik} x_{kj}$$

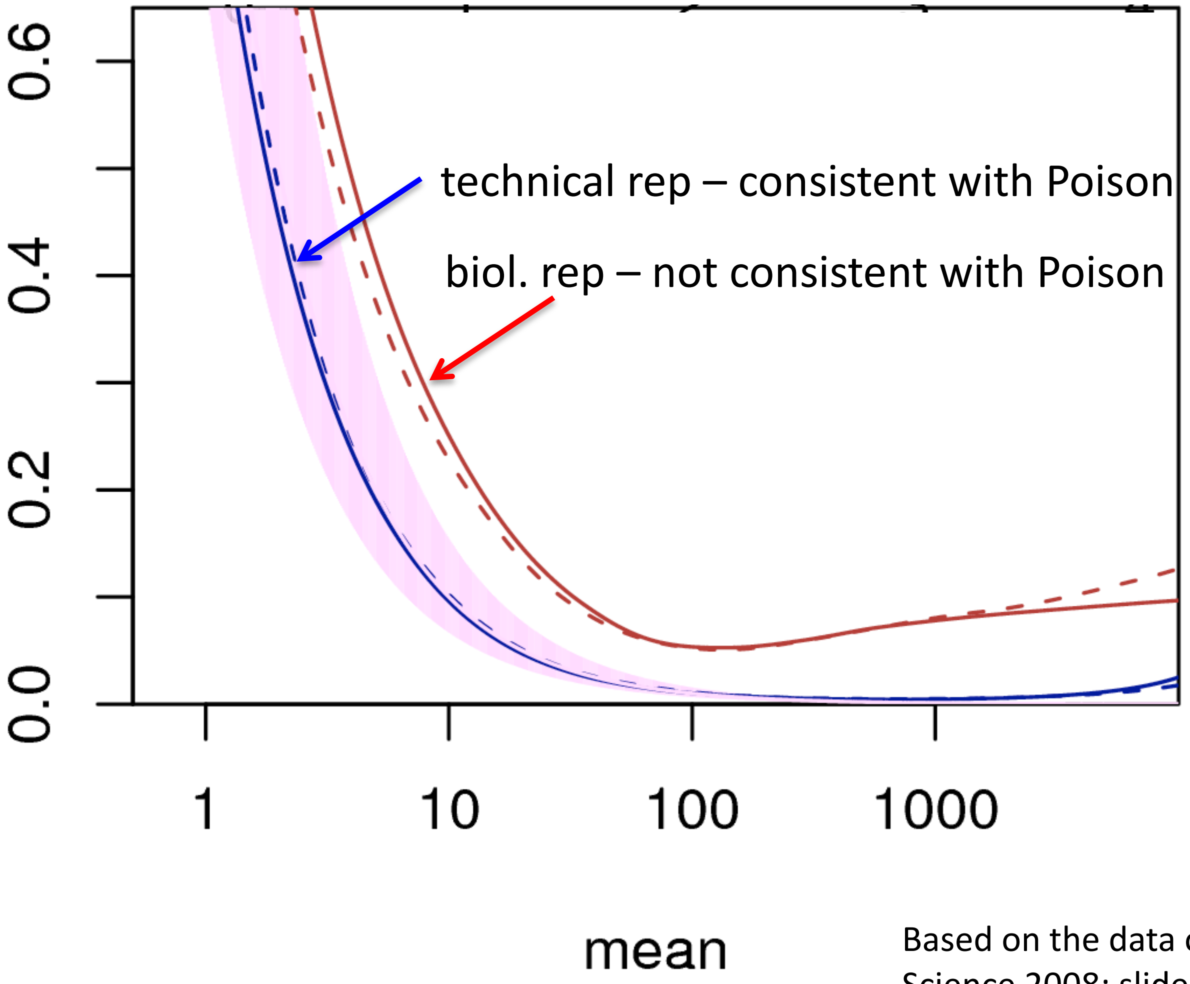
Design matrix

Deterministic Part

Library size effect

(Differential) effect for region i

Need to account for extra variability



technical rep – consistent with Poison

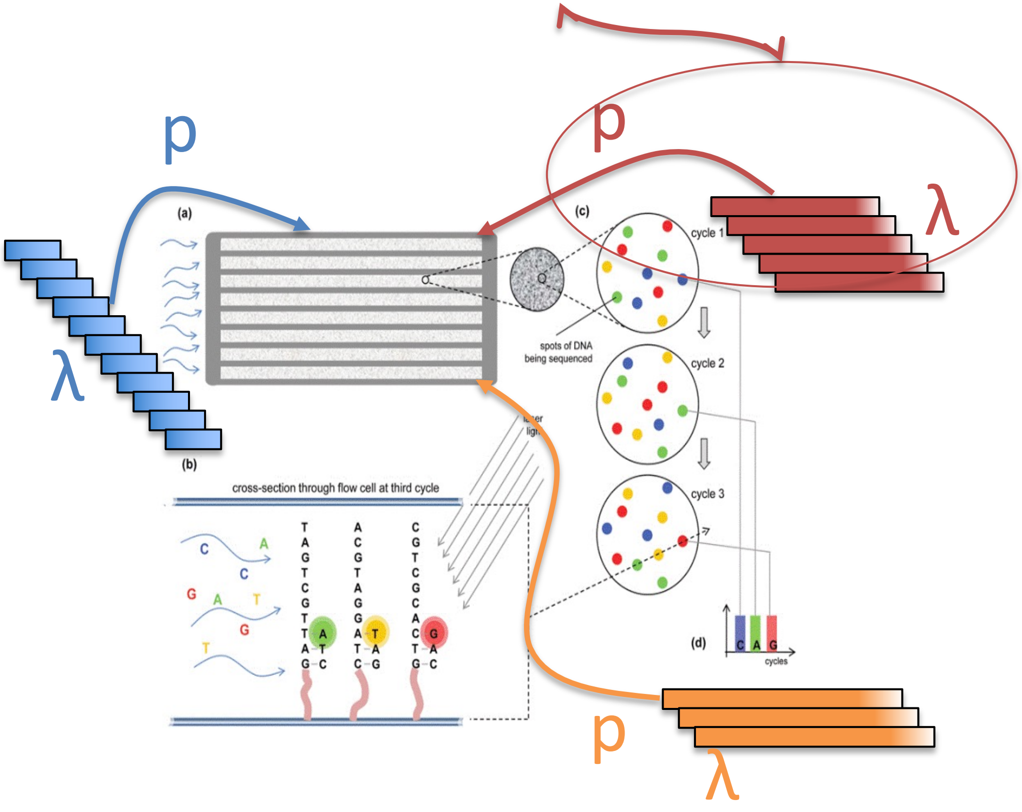
biol. rep – not consistent with Poison

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!!!


p
λ
p
λ
p
λ
(a)
(b)
(c)
(d)
spots of DNA being sequenced
cycle 1
cycle 2
cycle 3
laser light
cross-section through flow cell at third cycle
cycles
C A G

Current methods for DE use NB model!

Sequencing – Rationale
Biological Replicates

- For subject j, on transcript i:

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

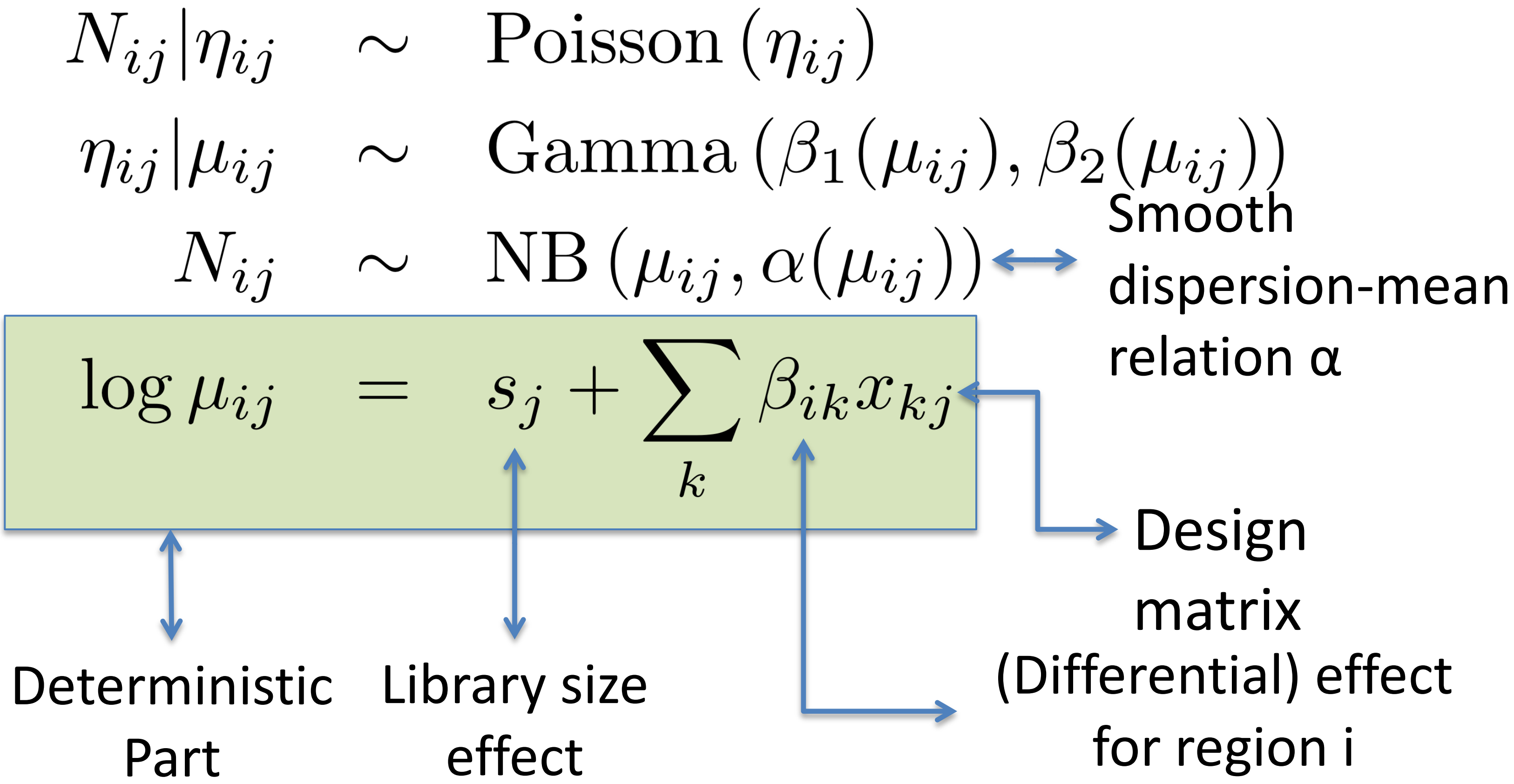
- Different subjects have different rates, which we can model through:

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

- This hierarchy changes the distribution of Y:

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

An additional source of variation



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

Smooth dispersion-mean relation α

Design matrix

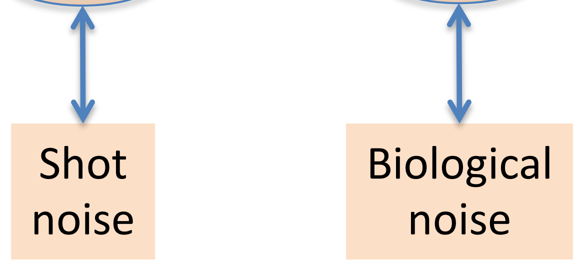
(Differential) effect for region i

Deterministic Part

Library size effect

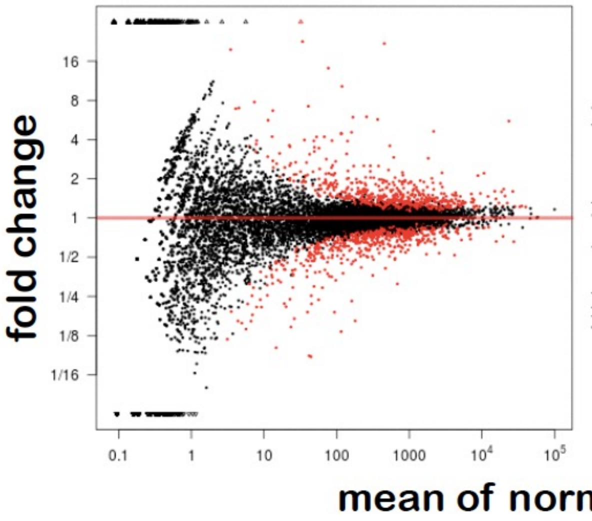
Summary of the Poisson and Negative Binomial Models

- Poisson(λ):
  - Mean: λ
  - Variance: λ
- Negative Binomial (α, $1/(1+\beta)$):
  - Mean: α/β
  - Variance: $\alpha(1+\beta)/\beta^2$
    - $= \alpha/\beta + \alpha/\beta^2 = \text{mean} + 1/\alpha * \text{mean}^2$

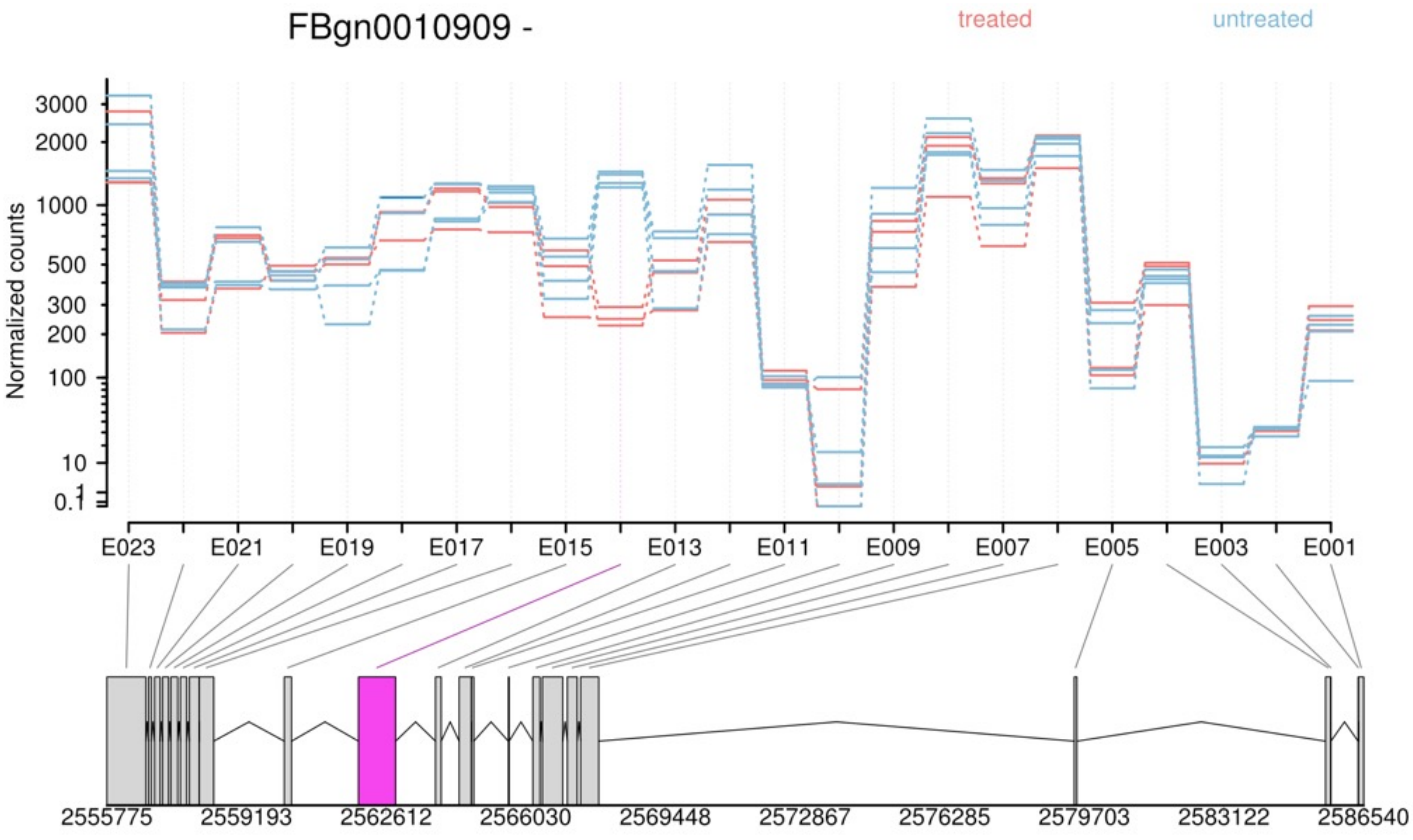


Shot noise

Biological noise



Example: DE / DEU



Summary of Models
Treatment (x_j) as Covariate

Gene Expression / DESeq

$$
\begin{aligned}
N_{ij} &\sim NB(s_j \mu_{ij}, \alpha(\mu_{ij})) \\
\log \mu_{ij} &\sim \beta_i^0 + \beta_i^T x_j^T
\end{aligned}
$$

Expression in control

Change for treatment

Alternative Exon Usage / DEXSeq

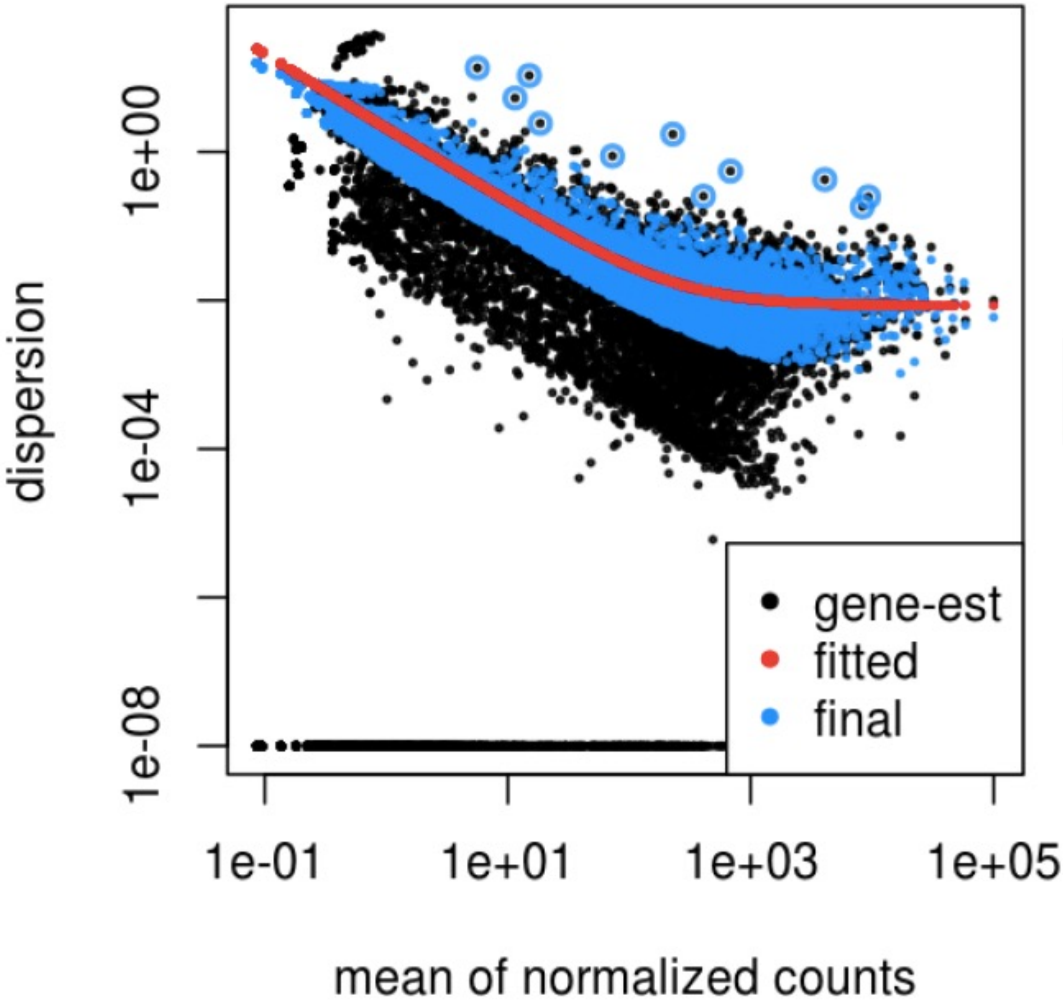
$$
\begin{aligned}
N_{ijl} &\sim NB(s_j \mu_{ijl}, \alpha(\mu_{ijl})) \\
\log \mu_{ijl} &\sim \beta_i^0 + \beta_{il}^E x_j^E + \beta_{ij}^T x_j^T + \beta_{ijl}^{ET} x_l^E x_j^T
\end{aligned}
$$

Fraction of reads falling onto exon *l* in control

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

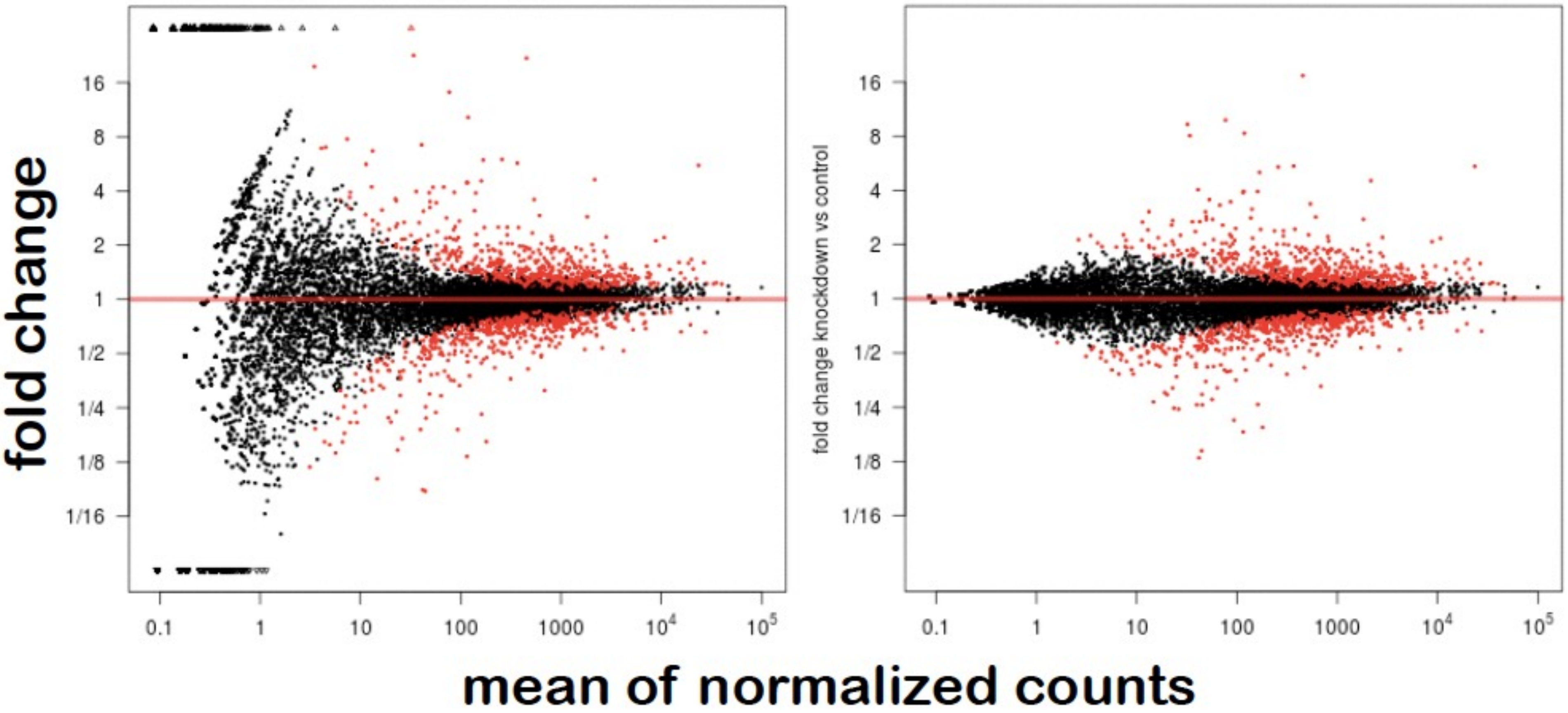
Variance Shrinkage

Dispersion estimation: shrinkage

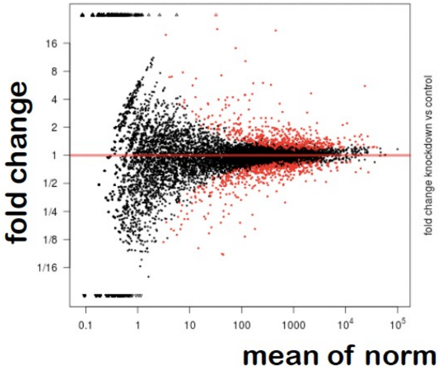


dispersion outliers:
$\log(\alpha_{\text{gene-est}}) - \log(\alpha_{\text{fit}}) > 2\ \sigma_{\text{rob}}$

Downstream Effect of Shrinkage

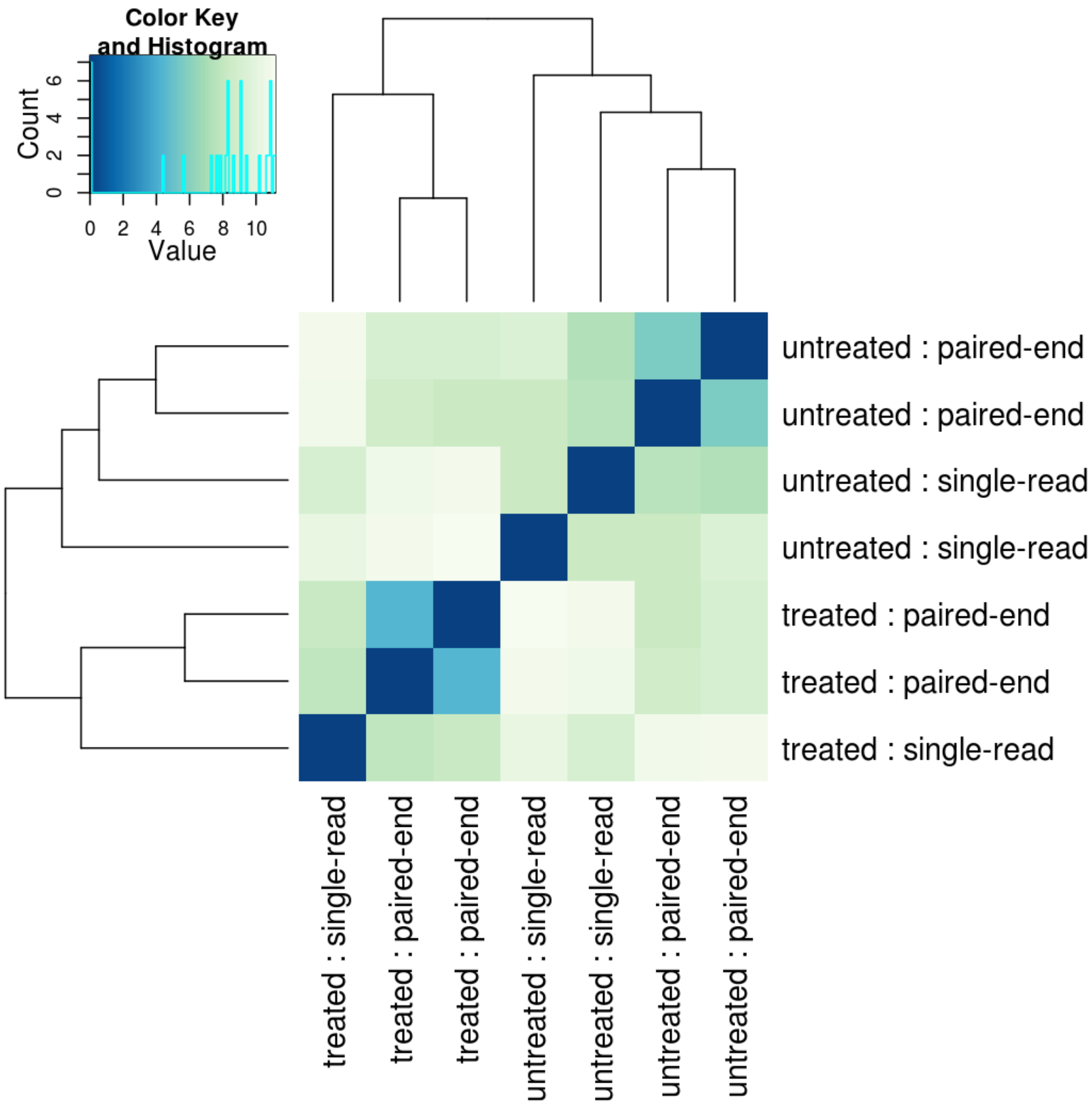


Remember the variance effect!

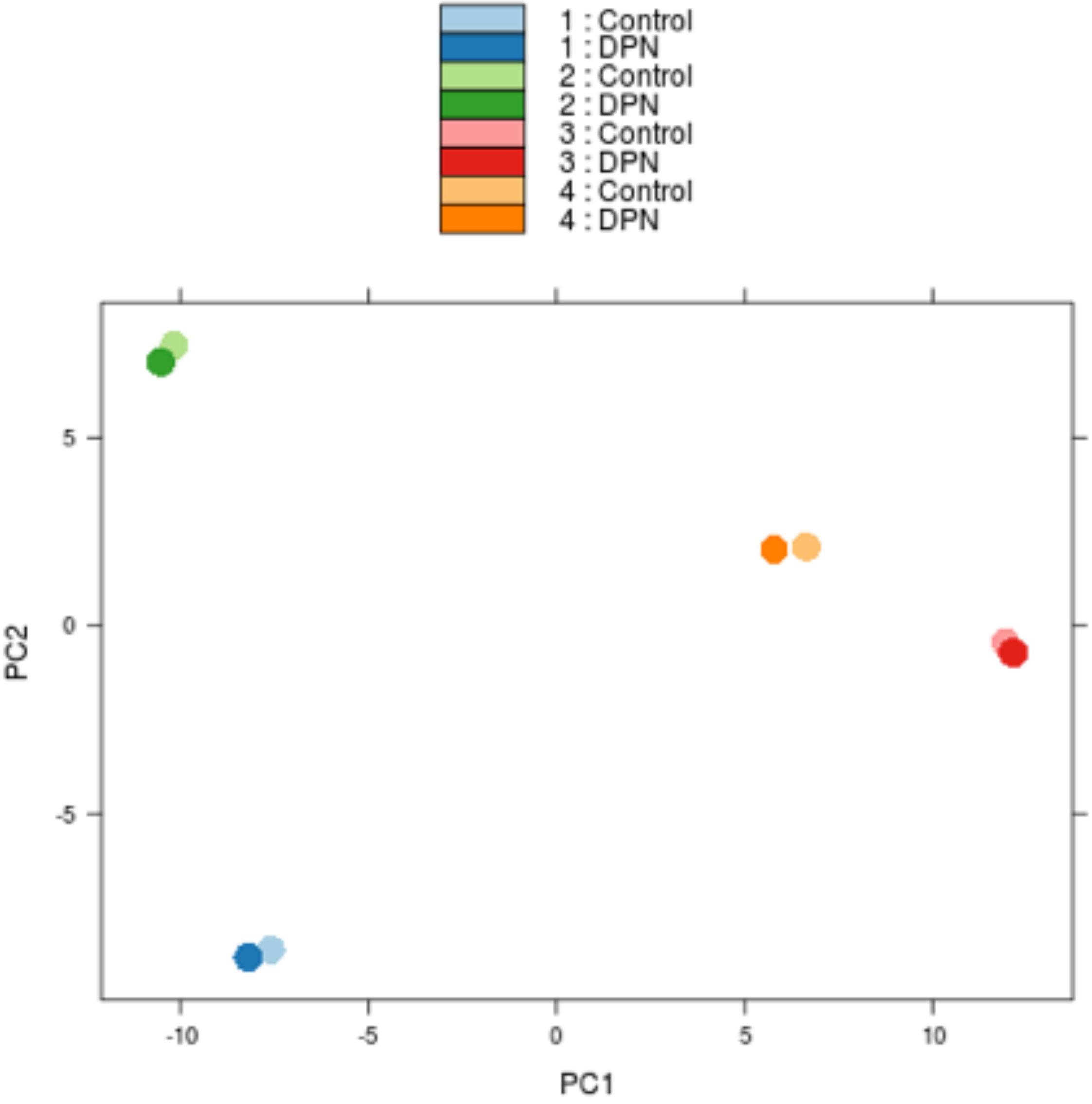


- Variance changes as mean changes…
- This seriously affects visualization;
- It also interferes with comparisons;
- One needs to adjust variance before performing clustering, visualization, PCA;
- DESeq2 has a “regularized log-transformation” method designed for that.

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
5
0
-5
-10
-5
0
5
10
PC1

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;

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;

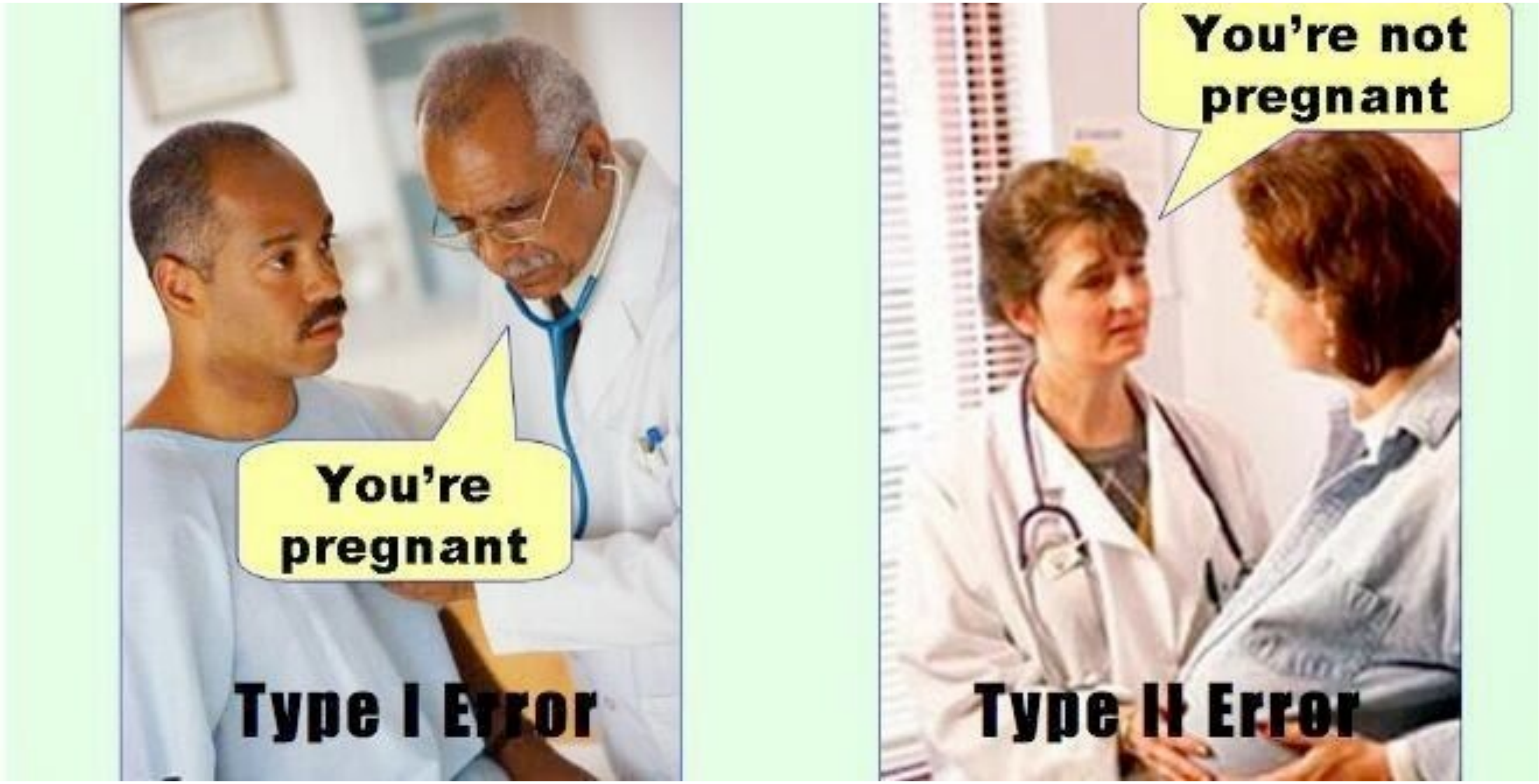
Multiple Testing

- 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

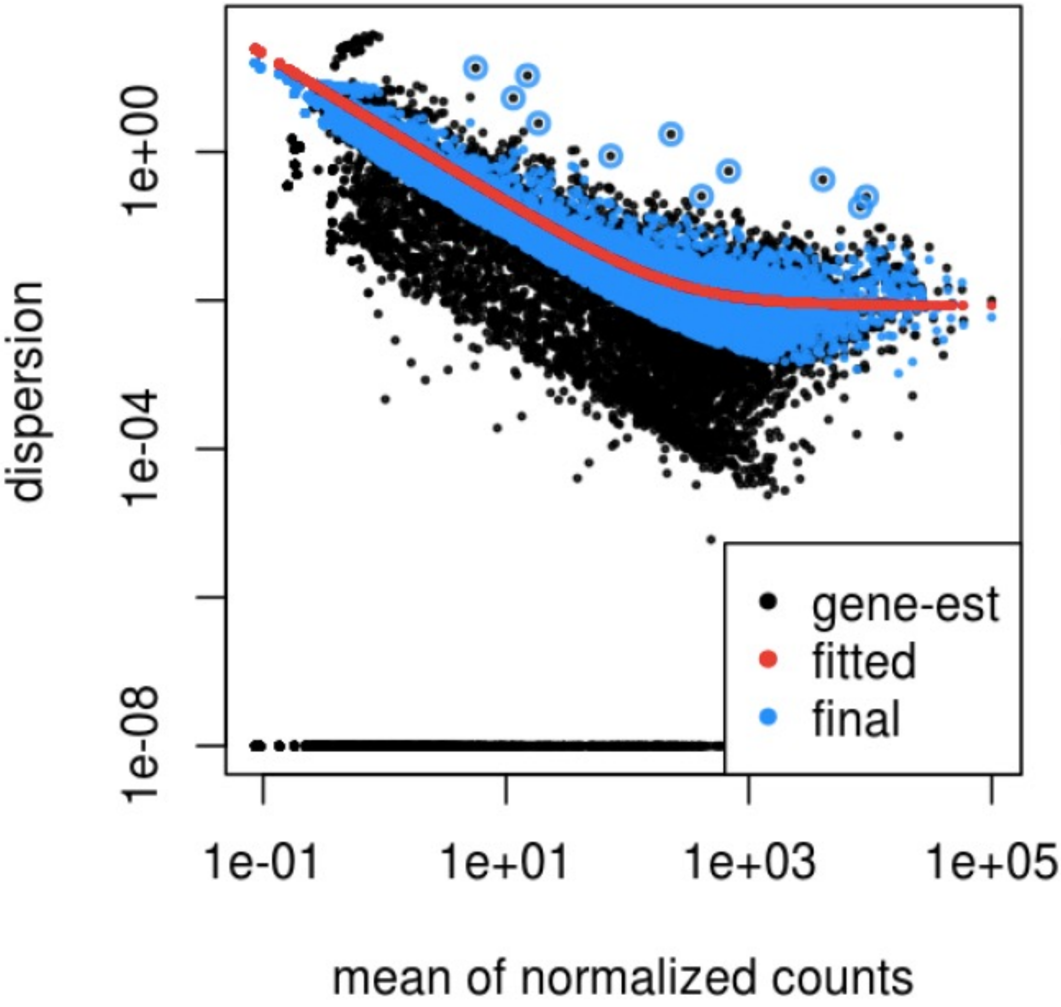


Non-Specific Filtering

- 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 Mean Expression as a Filter

Dispersion estimation: shrinkage

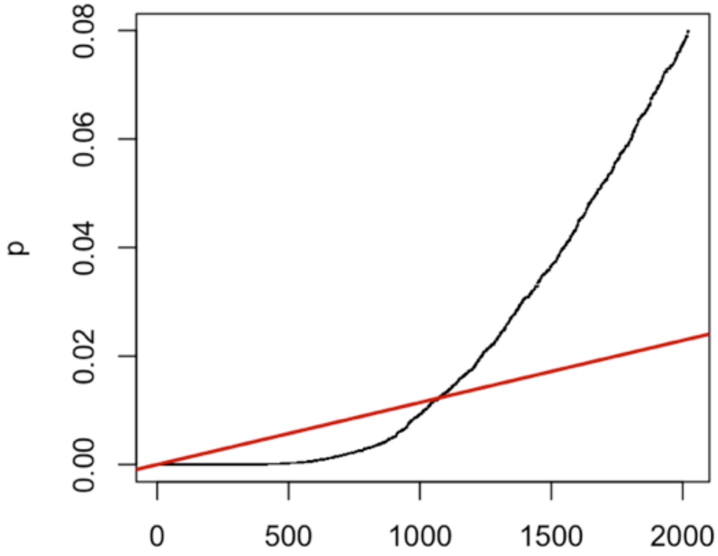


dispersion outliers:
$\log(\alpha_{\text{gene-est}}) - \log(\alpha_{\text{fit}}) > 2\,\sigma_{\text{rob}}$

FDR – Benjamini Hochberg (BH)

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

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


p
0.00
0.02
0.04
0.06
0.08
0
500
1000
1500
2000