A Bayesian framework for personalized design in alternative splicing RNA-seq studies

Camille Stephan-Otto Attolini
Biostatistics and bioinformatics Unit
IRB Barcelona

Workshop on experimental Design and Big Data

May 8th 2015

The University of Warwick

Introduction

RNAseq: Big Data, Big Questions

RNAseq is a technique to measure the abundance (expression) of mRNA in a cell.

The main application is to find differences in expression between samples in distinct biological conditions

RNAseq: Big Data, Big Questions

Q: What is the "Big" in the data?

A: Several million reads are generated from a single sample. These reads correspond to about 20,000 genes in the human genome and, according to the most conservative annotations, to more than 40,000 (mostly overlapping) transcripts

The question

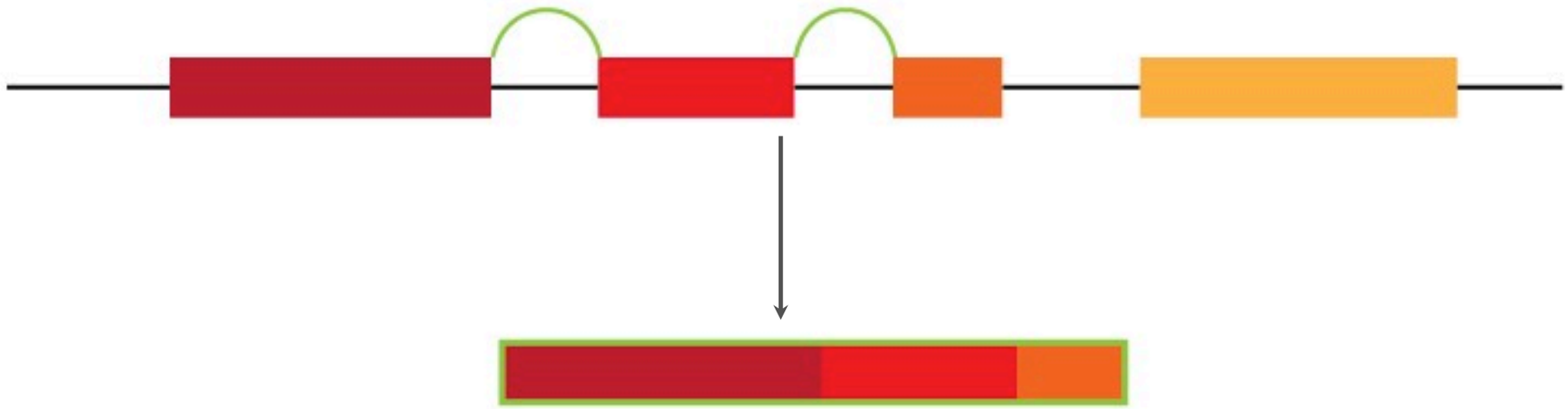
Find the optimal experimental design to better estimate transcript level expression and/or to find those transcripts differentially expressed between conditions while keeping the number of reads and samples (cost and time) as low as possible

Alternative splicing

IN A NUTSHELL

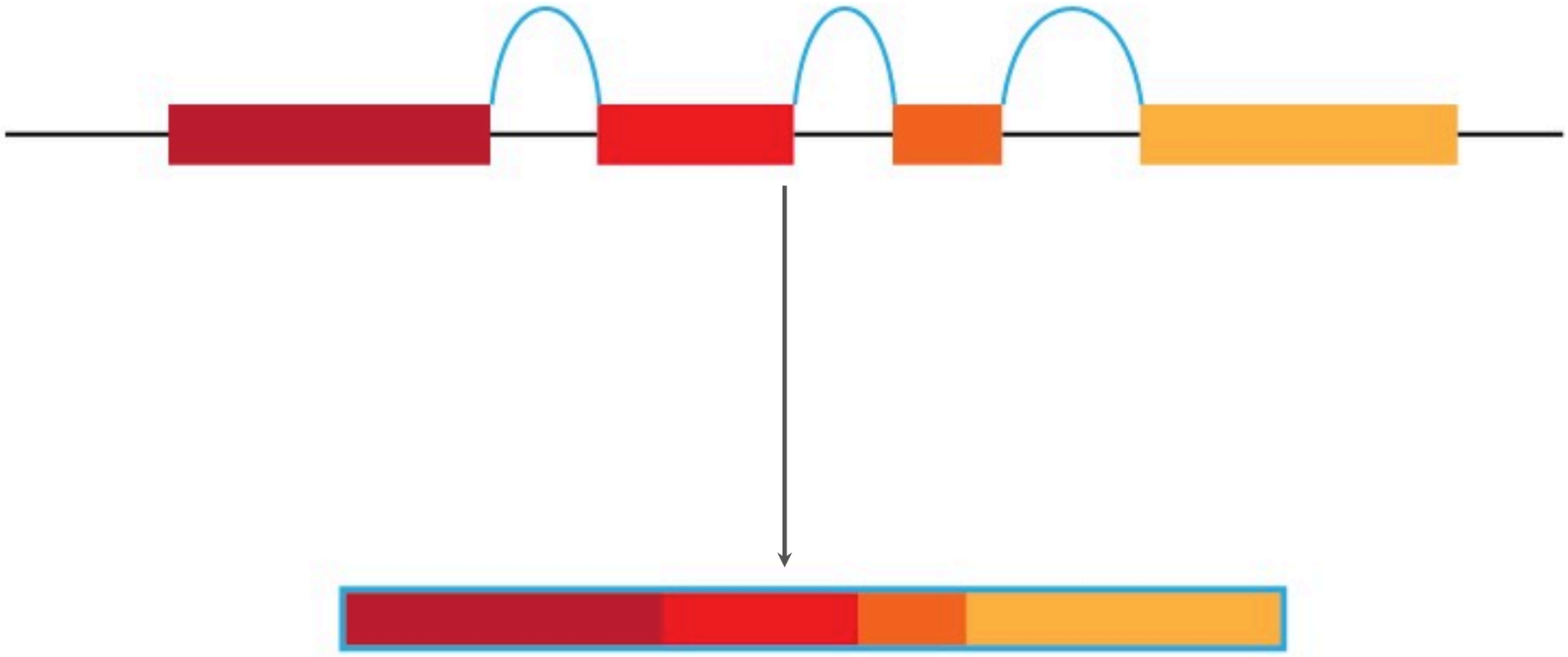
Alternative splicing

IN A NUTSHELL



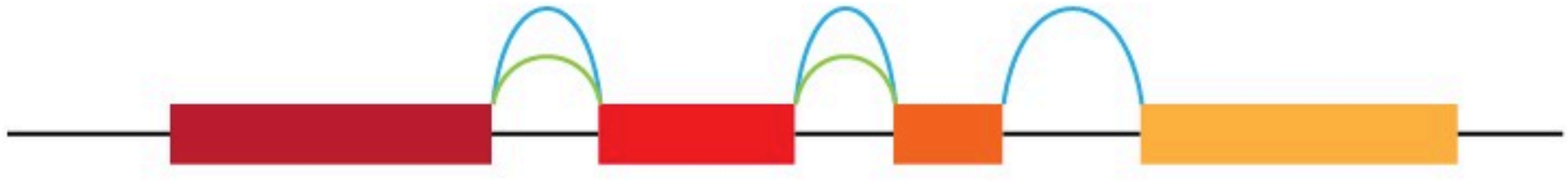
Alternative splicing

IN A NUTSHELL



Alternative splicing

IN A NUTSHELL



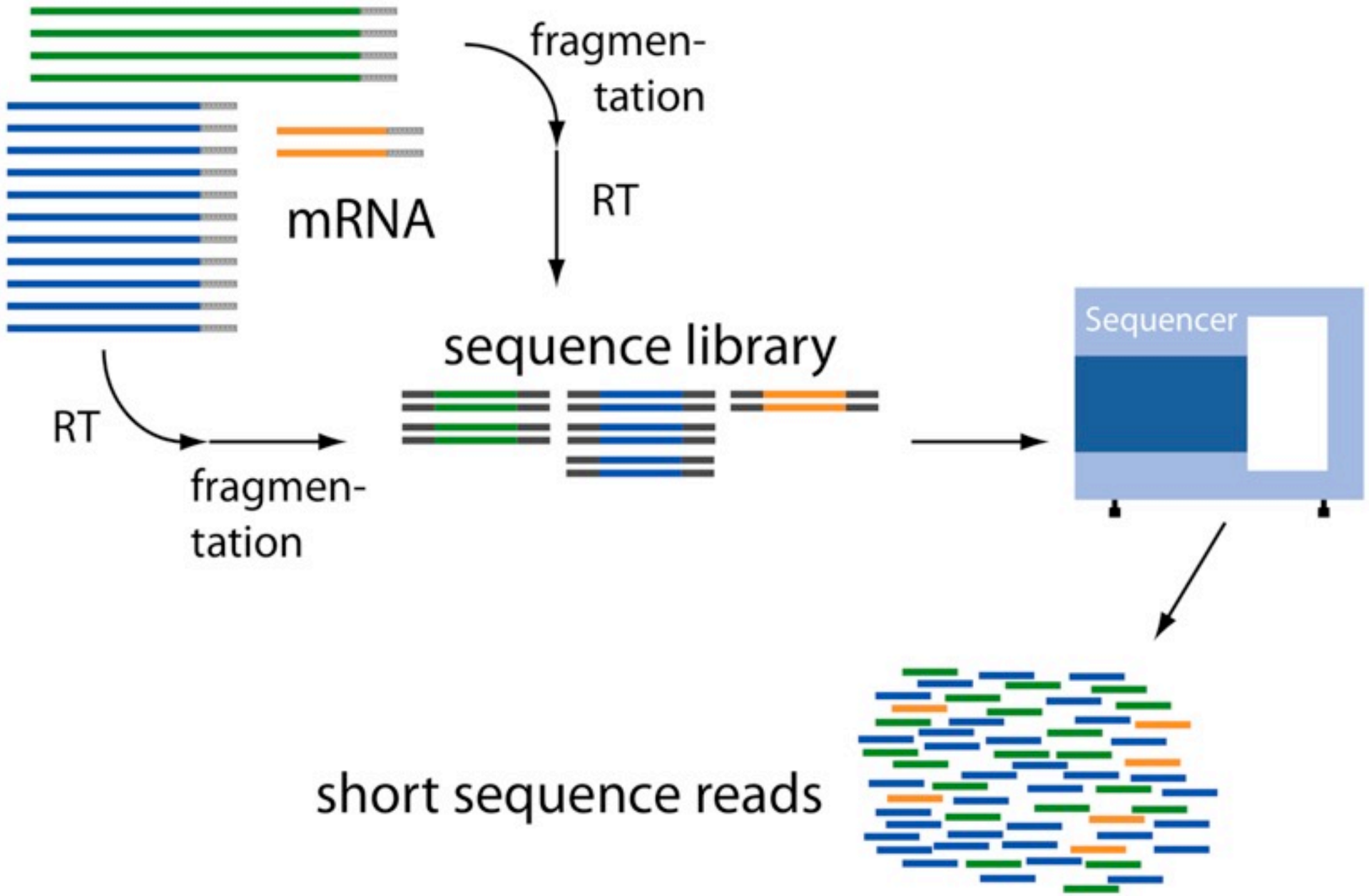
Relative expression

35%

65%

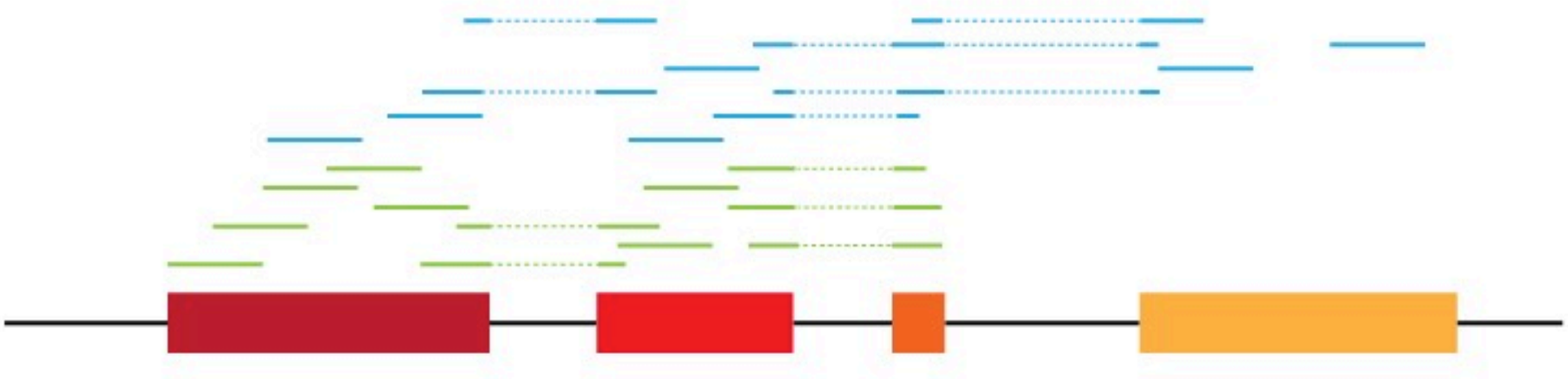
RNA sequencing

ALSO IN A NUTSHELL



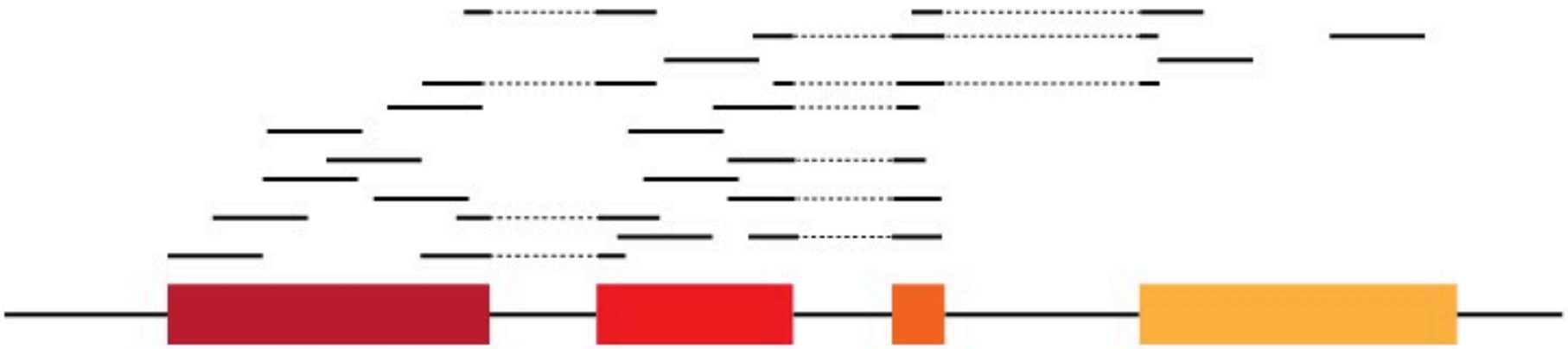
RNA sequencing

In a beautiful world



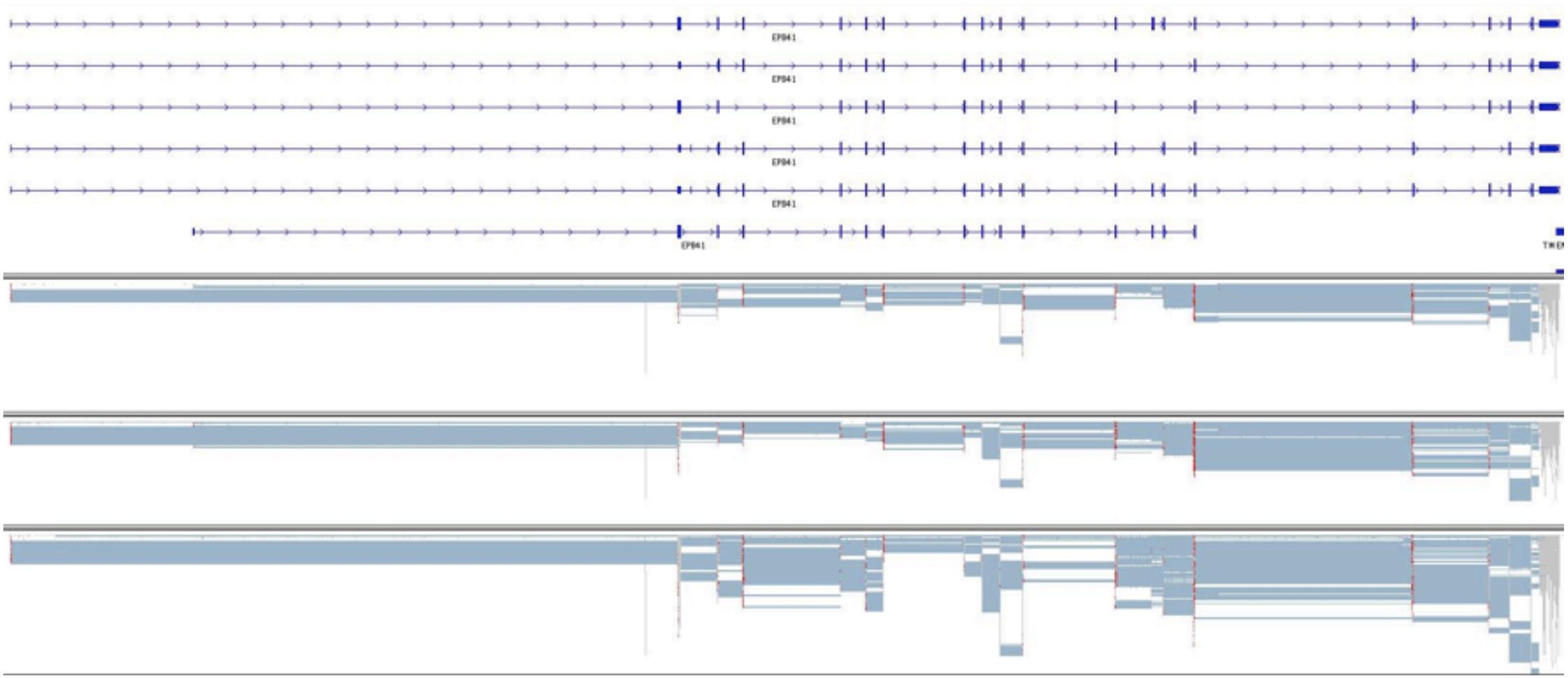
RNA sequencing

In the real world



RNAseq

Real data



Statistical framework

Statistical framework

The data

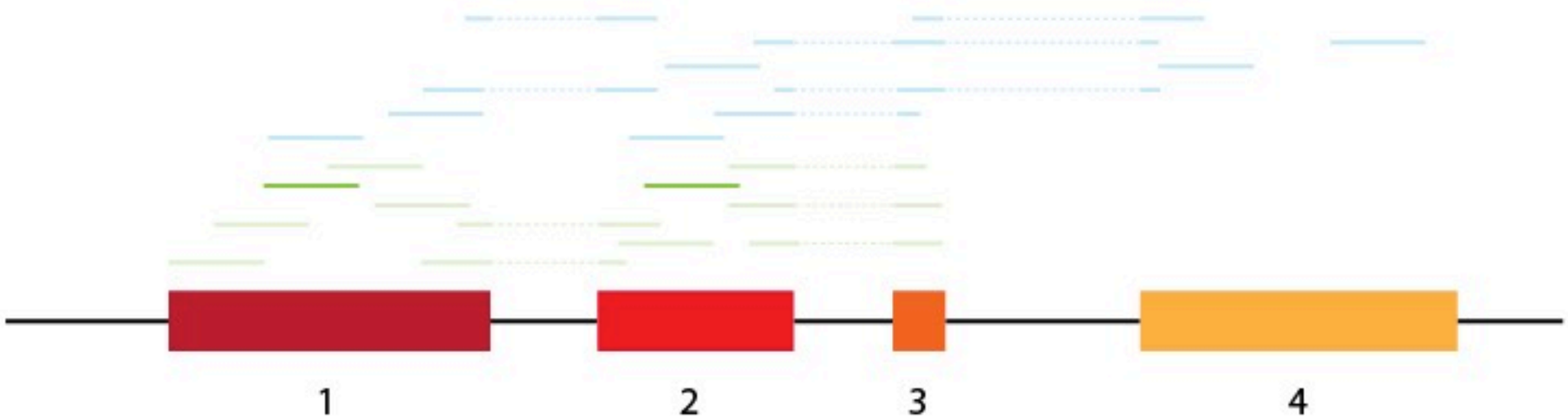
- Impossible to model full data (10^7 ~ 10^8 seqs. per experiment)

- Experimental biases such as uneven distributions of fragments along transcripts

- Fragment length distributions not always as reported by lab

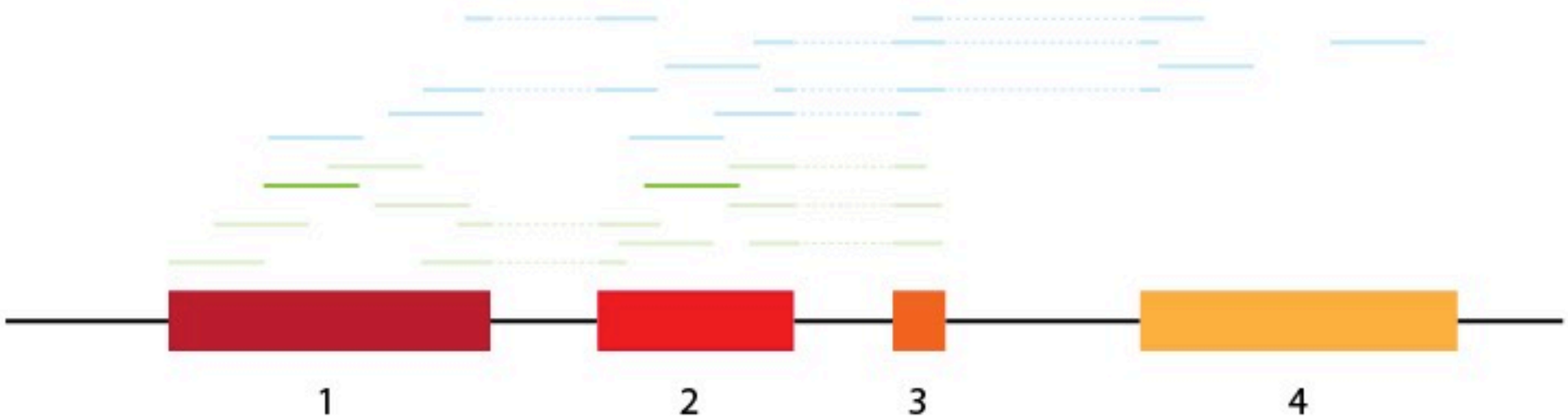
Statistical framework

Exon paths



Statistical framework

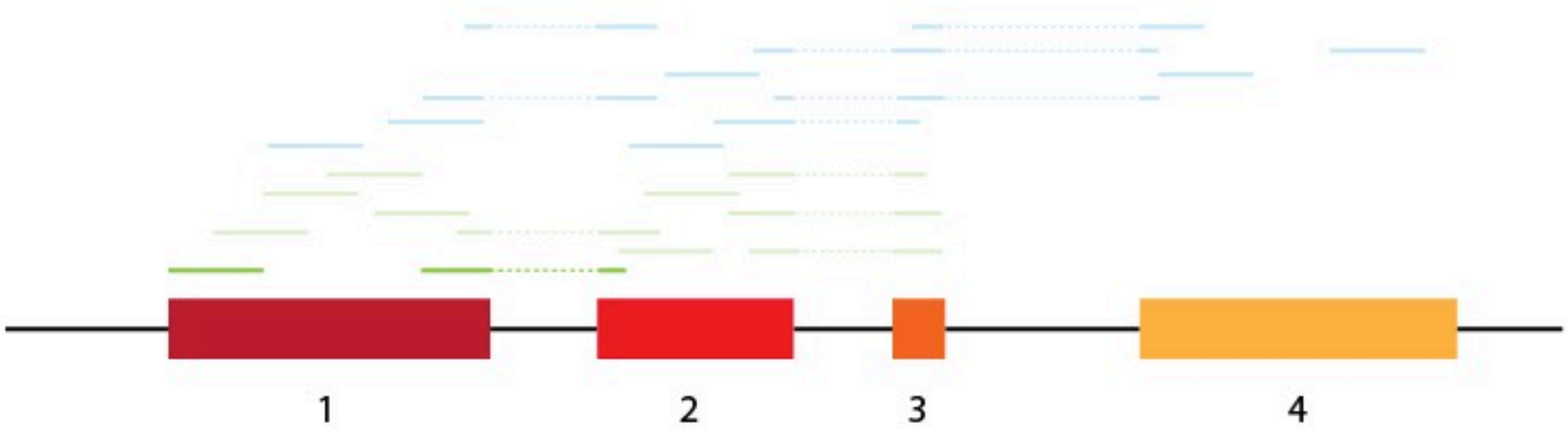
Exon paths



Path	Fragment Count
1 – 2	10

Statistical framework

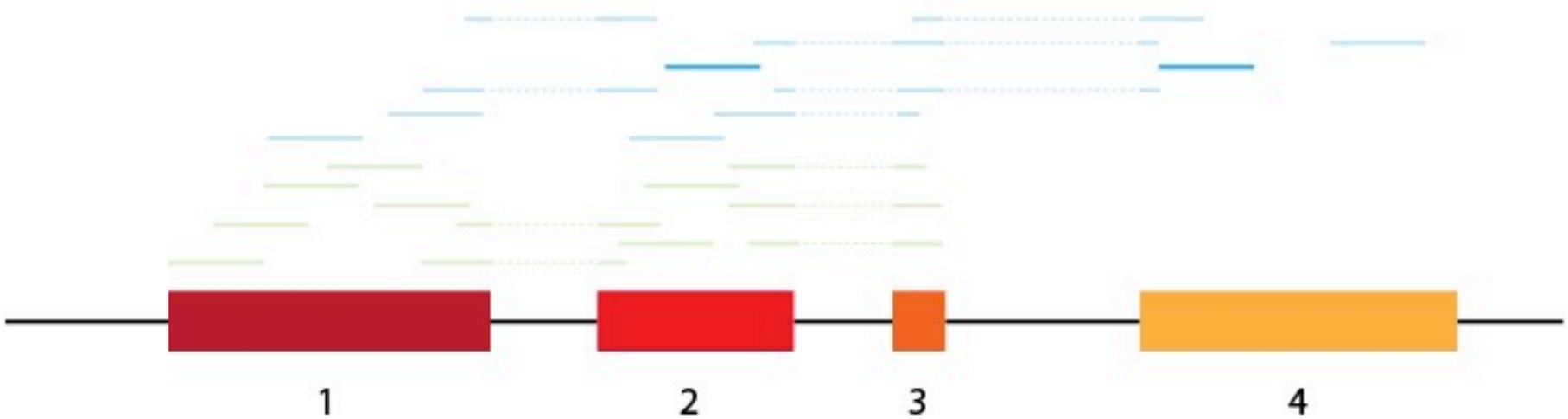
Exon paths



Path	Fragment Count
1 – 2	10
1 – 1.2	56

Statistical framework

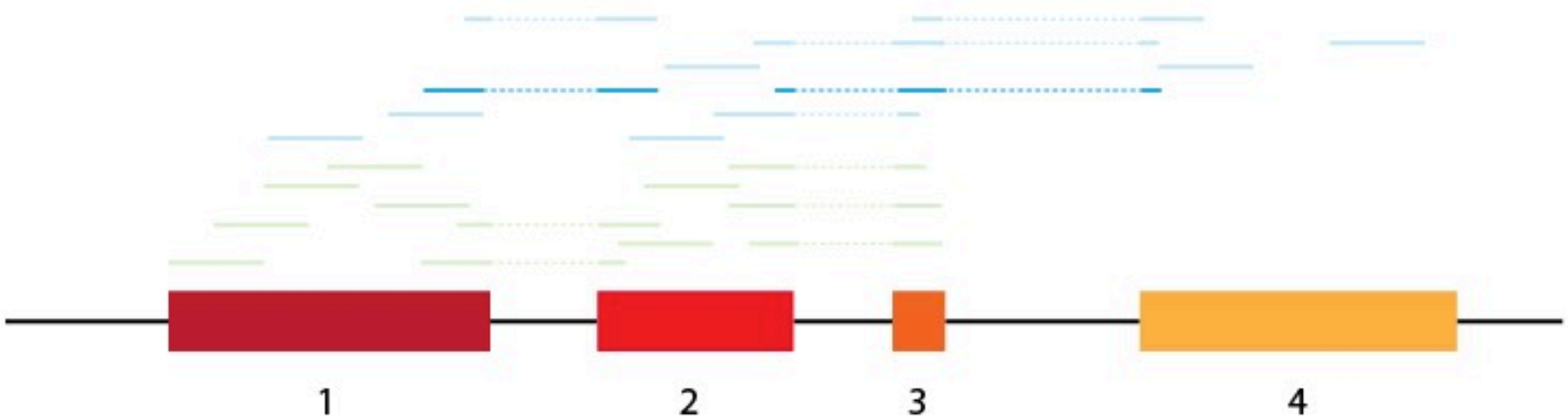
Exon paths



Path	Fragment Count
1 – 2	10
1 – 1.2	56
2 – 4	120

Statistical framework

Exon paths



Path	Fragment Count
1 – 2	10
1 – 1.2	56
2 – 4	120
1.2 – 2.3.4	8

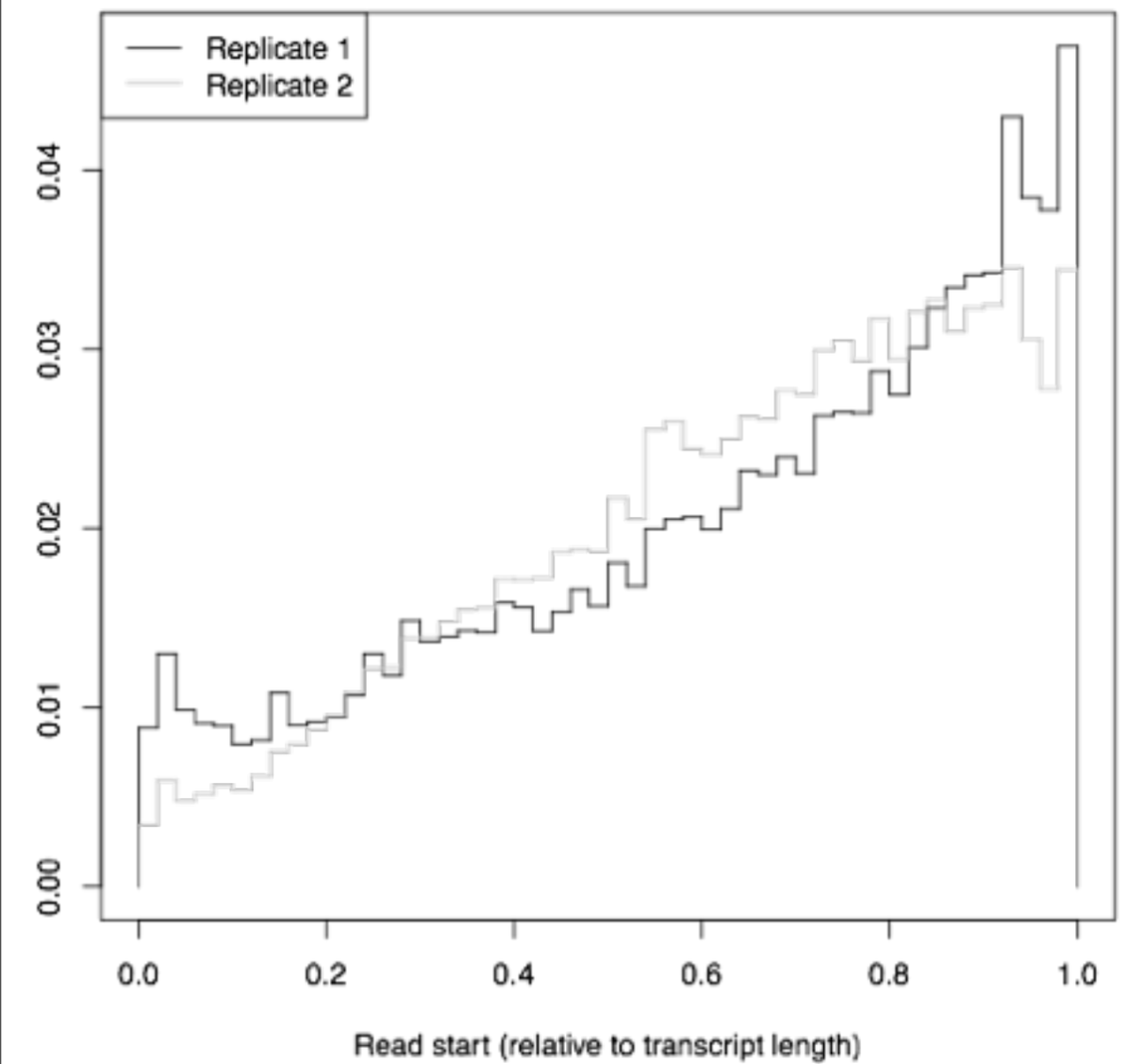
Statistical framework

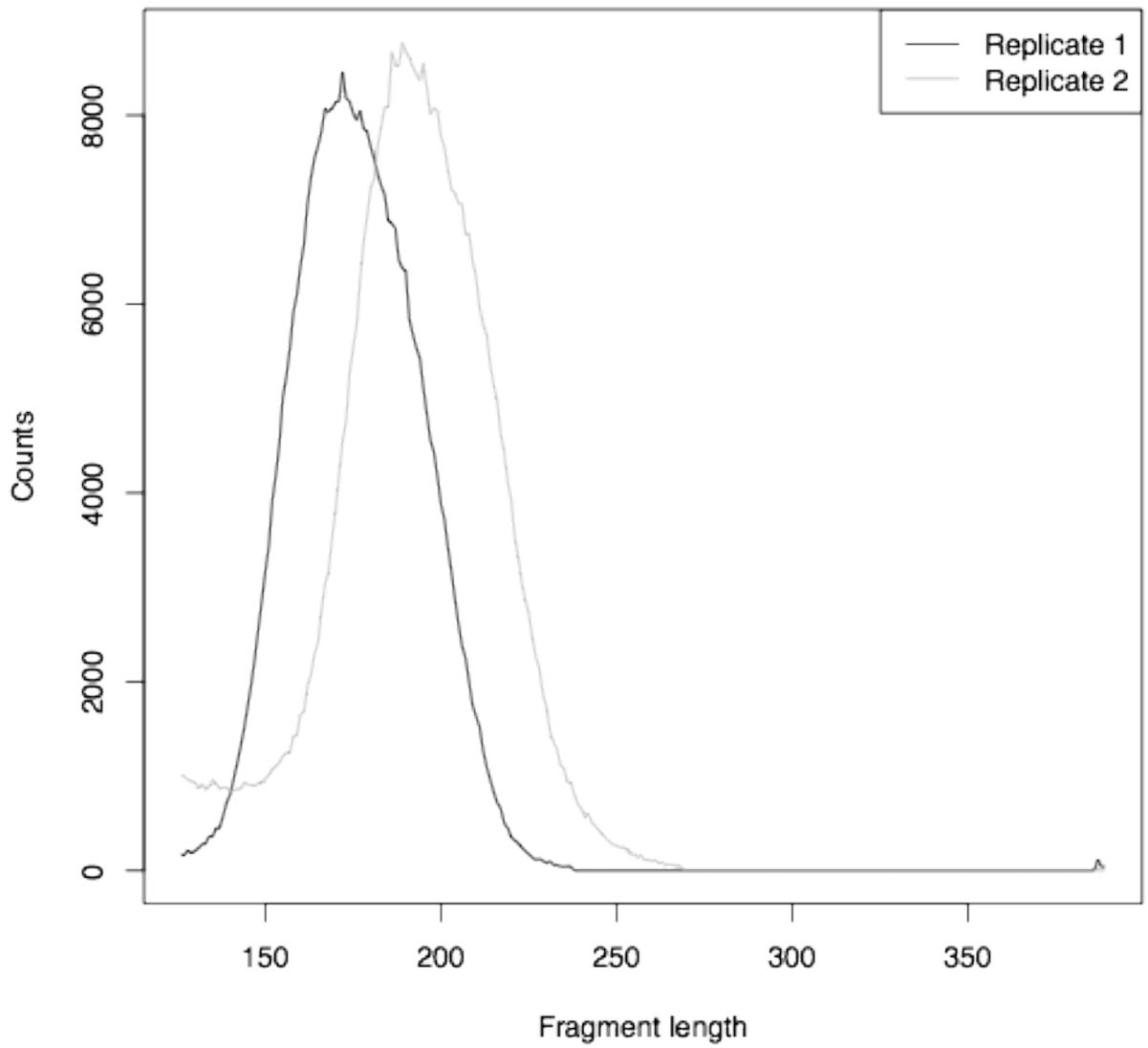
Exon paths

Path	Fragment Count
1 – 2	10
1 – 1.2	56
2 – 4	120
1.2 – 2.3.4	8
…	…

Statistical framework

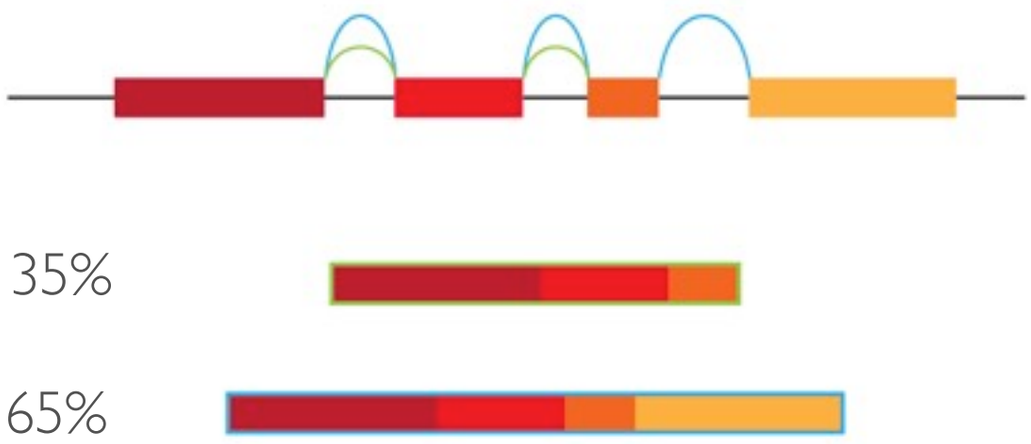
Fragment start and length distributions

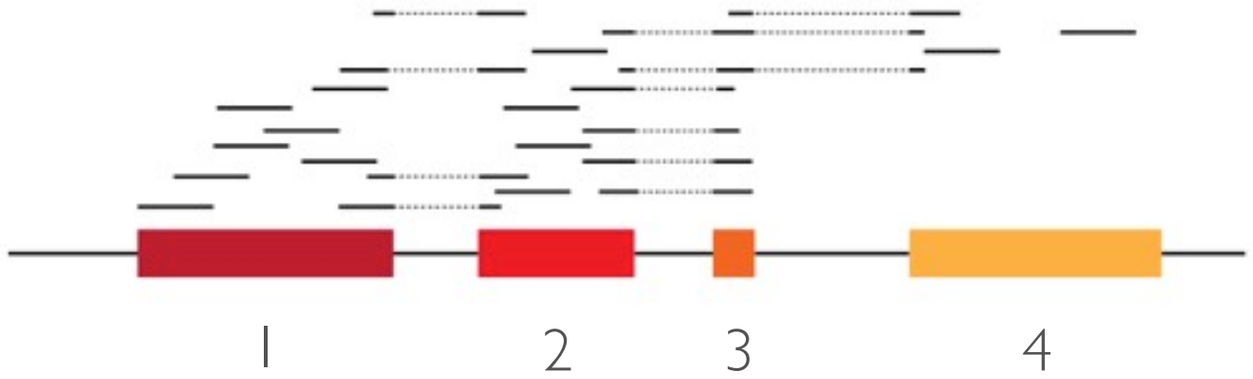




Bringing all together

The model

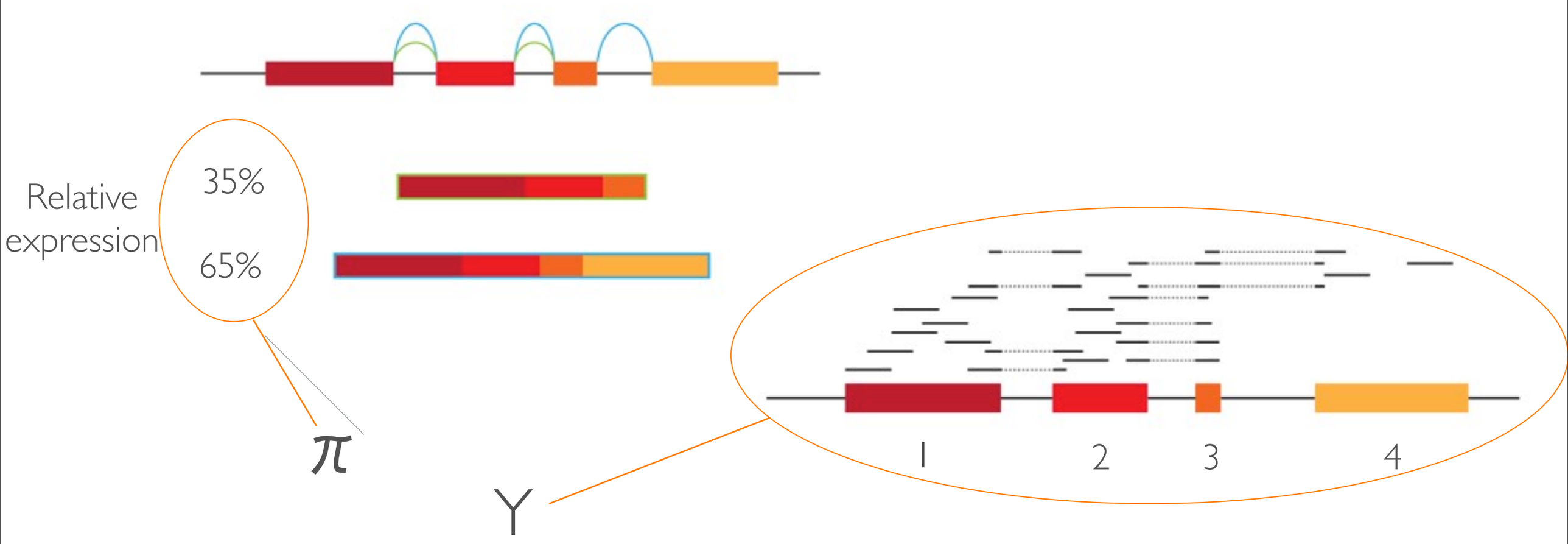




Path	Fragment Count
1 – 2	10
1 – 1.2	56
2 – 4	120
1.2 – 2.3.4	8

Bringing all together

The model



Path	Fragment Count
1 – 2	10
1 – 1.2	56
2 – 4	120
1.2 – 2.3.4	8

Statistical framework

Bayes rule rules!

$$P(\pi|Y) \propto P(Y|\pi)P(\pi)$$

Statistical framework

Bayes rule rules!

$$P(\pi|Y) \propto P(Y|\pi)P(\pi)$$

Posterior probability

Statistical framework

Bayes rule rules!

$$P(\pi|Y) \propto P(Y|\pi)P(\pi)$$

Likelihood

$$P(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\delta}) = \prod_k \left(\sum_{d \in \boldsymbol{\delta}} P(\text{path } k|d)\pi_d \right)^{y_k}$$

where $P(\text{path } k|d) = \int \int I(\text{path } k|s, l) d\hat{P}_L(l) d\hat{P}_s(s)$

- P_L: fragment length distrib.
- P_s: read start distrib. (e.g. 3' bias)

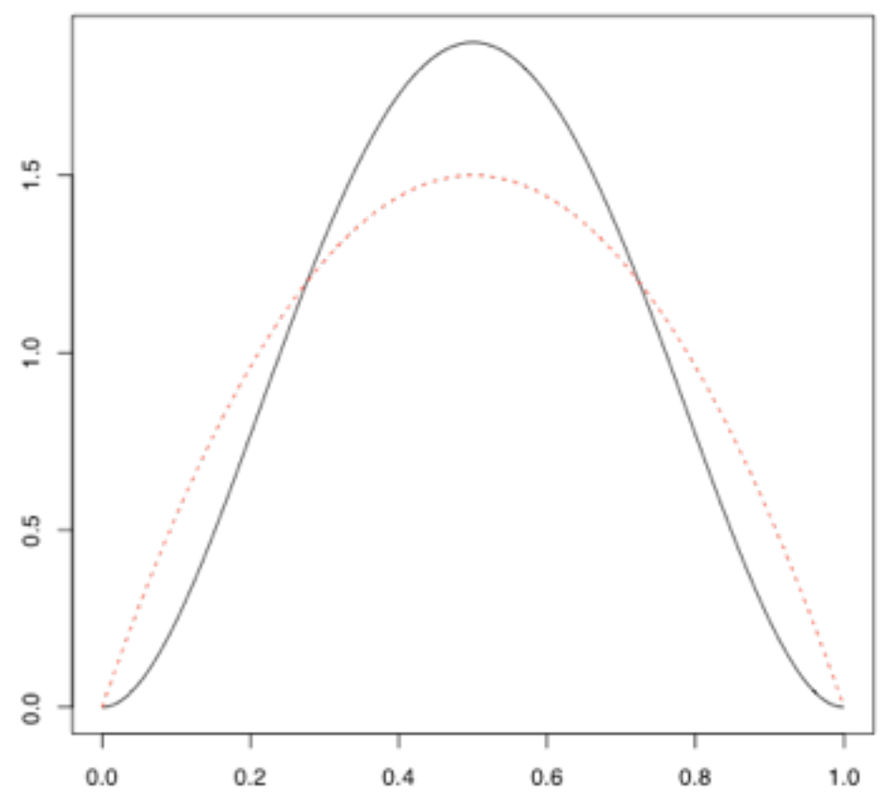
Statistical framework

Bayes rule rules!

$$P(\pi|Y) \propto P(Y|\pi)P(\pi)$$

Prior

$P(\pi) = \text{Dir}(q)$ with $q = 2$ or 3



$$P(\mathbf{y}|\pi, \delta) = \prod_k \left(\sum_{d \in \delta} P(\text{path } k|d)\pi_d \right)^{y_k}$$

where $P(\text{path } k|d) = \int \int I(\text{path } k|s, l) d\hat{P}_L(l) d\hat{P}_s(s)$

- P_L: fragment length distrib.
- P_s: read start distrib. (e.g. 3' bias)

The Model

Path counts arise from a mixture (one component from each variant):

$$P(\mathbf{y}|\boldsymbol{\pi}, \boldsymbol{\delta}) = \prod_k \left(\sum_{d \in \boldsymbol{\delta}} P(\text{path } k|d)\pi_d \right)^{y_k}$$

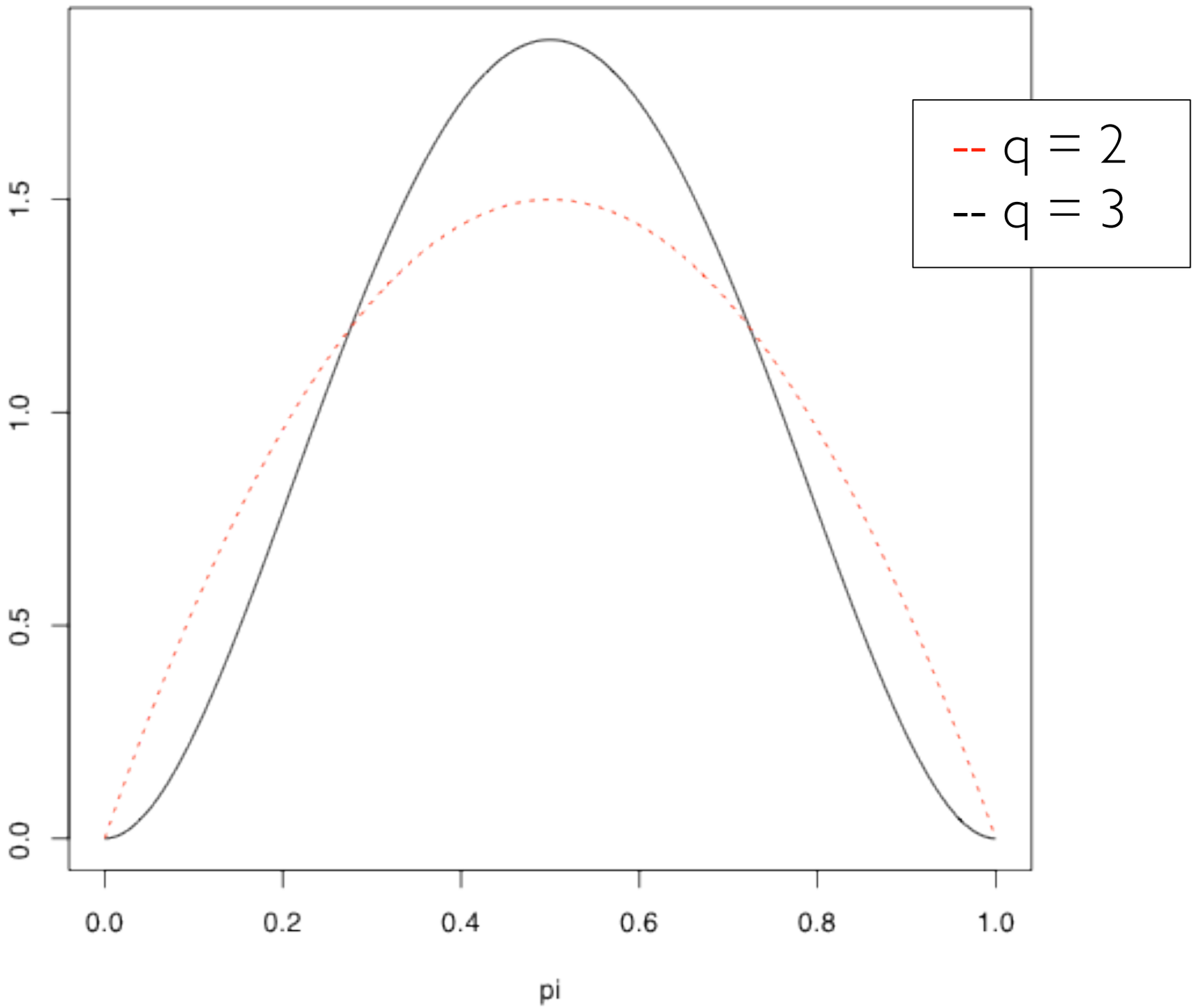
where $P(\text{path } k|d) = \int \int I(\text{path } k|s, l) d\hat{P}_L(l) d\hat{P}_s(s)$

- P_L: fragment length distrib.
- P_s: read start distrib. (e.g. 3' bias)

David Rossell et al. Annals of Applied Statistics, 2012

Statistical framework

Prior



Statistical framework

Model fitting

- EM algorithm to maximize posterior probability

- Find point estimates of relative expressions, asymptotic credibility intervals and exact posterior samples

- The algorithm converges to a single maximum

Reads simulation

Following the same model, we can simulate reads from a given set of expressions, reproducing the same technical biases observed in the data

Sample size calculation

Question I : Coverage in a single sample

What is the number of reads that need to be sequenced in order to control the estimation error below a certain threshold?

Simulation setup

1. Compute relative expressions and distributions for pilot data (human and mouse)

2. For a range of total bp sequenced simulate under 5 parameter scenarios:

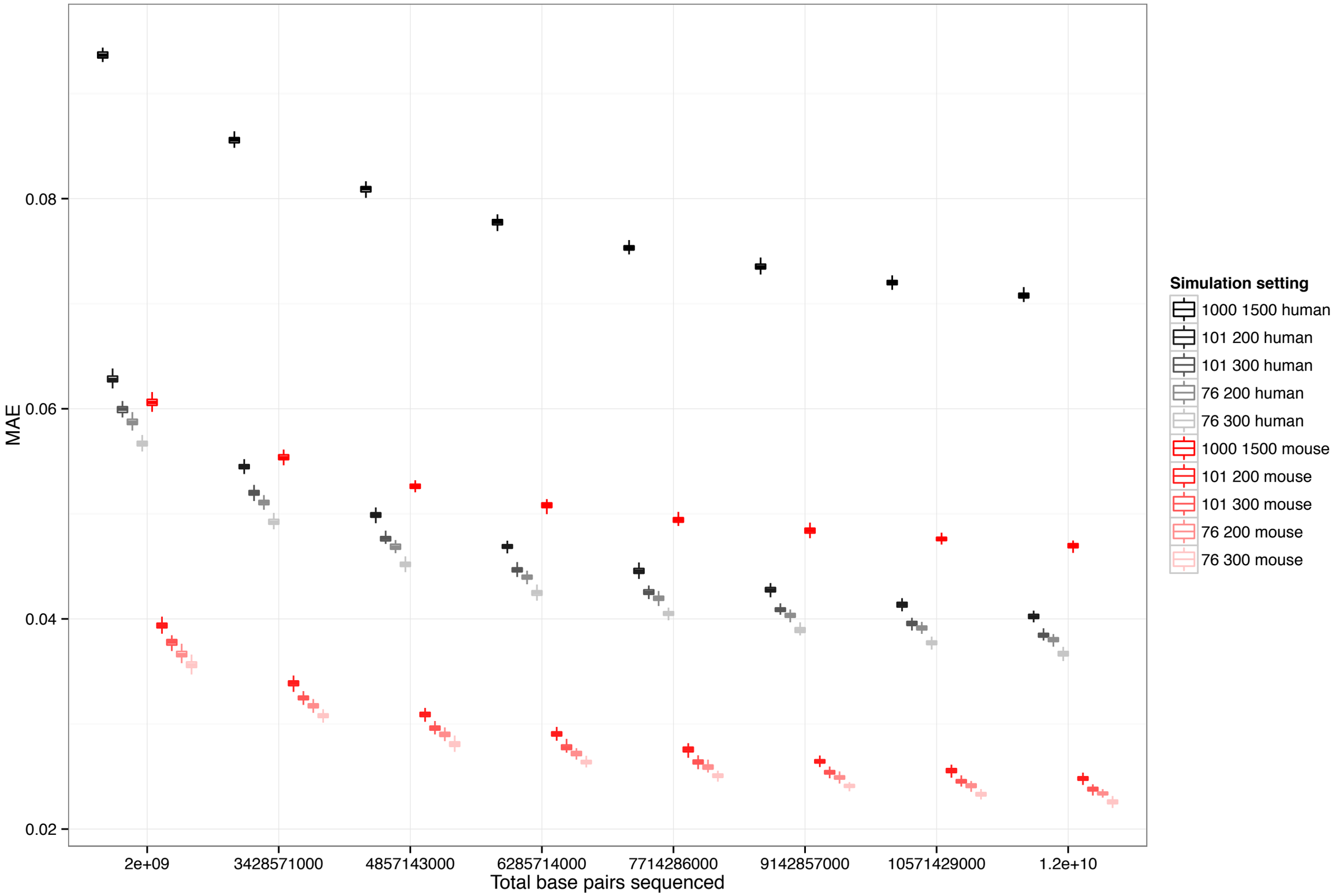
Mean of fragment's length	Read length
200	76
200	101
300	76
300	101
1500	1000

Simulation setup

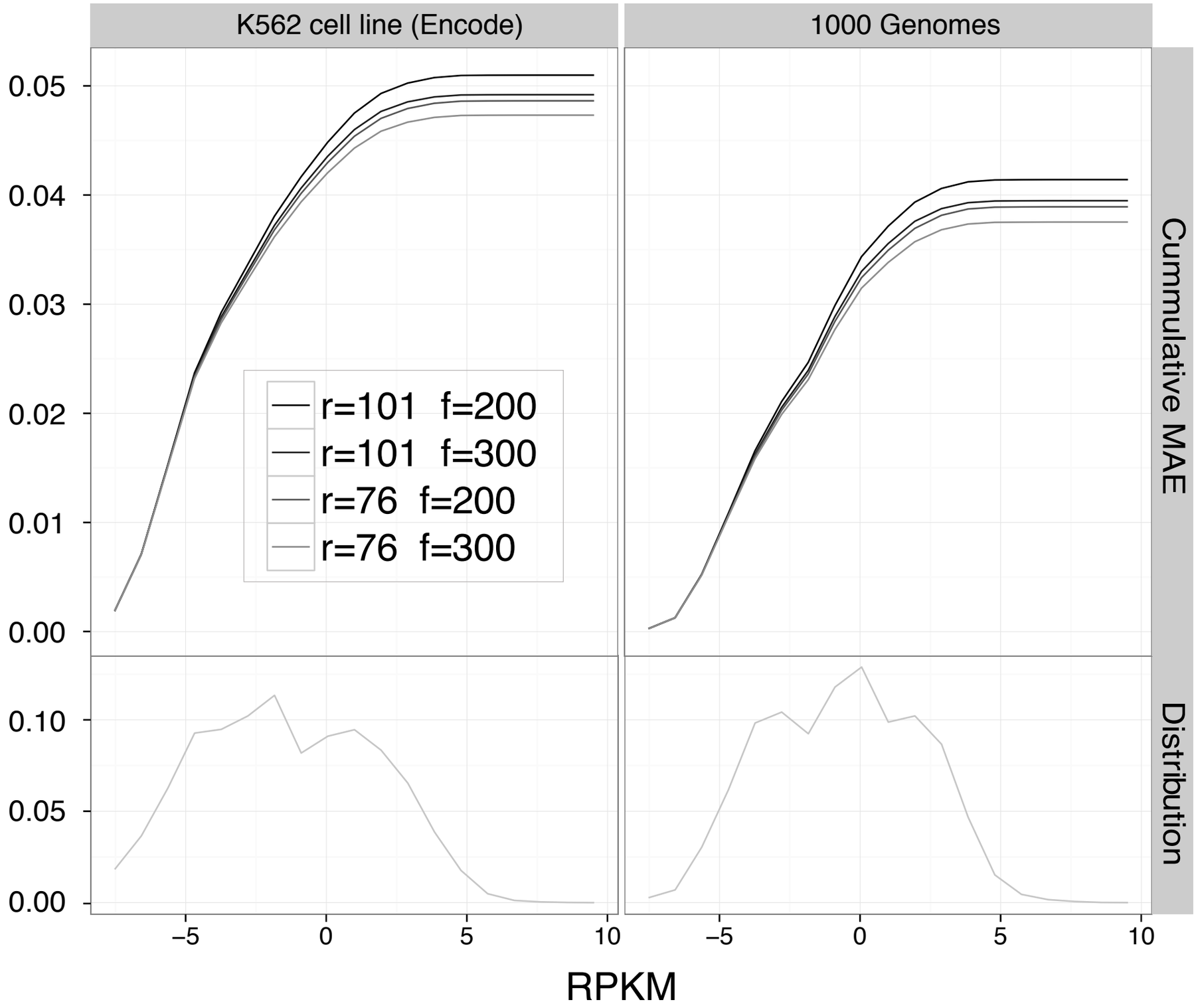
3. Compute relative expression from simulated data

4. Compute mean absolute error between original expressions and computed expressions

Results: One sample problem



Results: One sample problem



Question I I: Differential expression

How many samples per group and at what coverage should I sequence to get a desired number of differentially expressed genes(DEG)?

Publicly available data

Dataset A

We found a dataset consisting of 6 samples (3 for control and 3 for condition) sequenced at low coverage in a Mlseq Illumina sequencer.

A follow up dataset was generated with high coverage from a Hlseq Illumina sequencer for the same conditions

Publicly available data

Dataset B

Well know MAQC samples (human brain vs pooled human tissues) were generated in 5 technical replicates with high coverage in Hlseq Illumina sequencer.

Simulation setup

1. Compute relative expressions and distributions for a set of samples from pilot data (control vs condition)

2. Simulate 3 combinations of parameters:

Mean of fragment's length	Read length	Total reads
300	101	16M
300	101	32M
150	750	1.8M

Simulation setup

3. Simulate 3 and 6 samples for each group *

4. Compute number of DEG genes for pilot and simulated data combined

5. Compare to number of DEG genes from real data

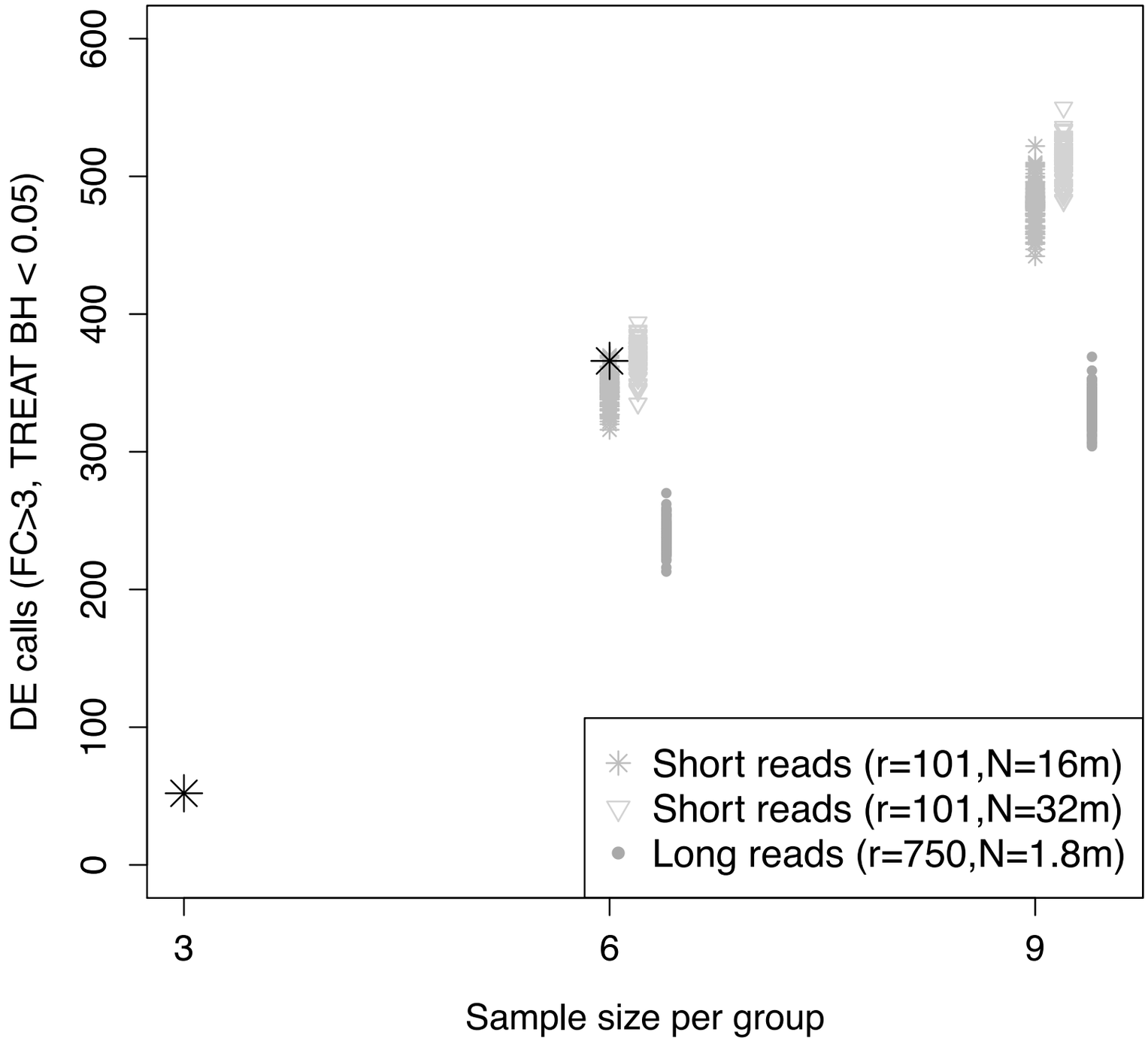
(*) We simulate new samples following a LNNMV model for each of the group's expression per transcript.

Simulation setup

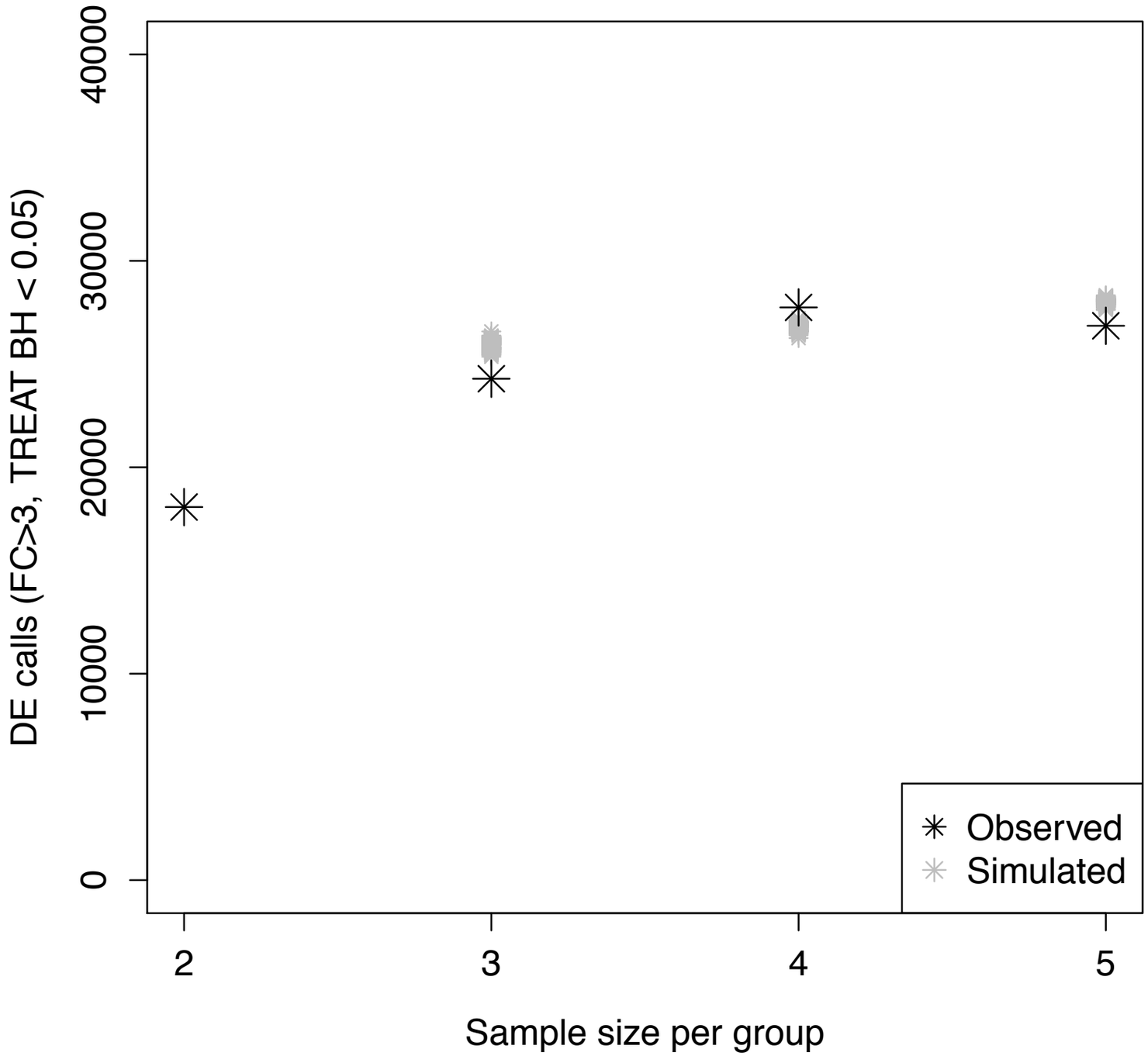
(*) To simulate the expressions of new samples we use the lognormal-normal with modified variance model (LNNMV, Yuan and Kendiorski (2006)).

This model correctly fits the data as was seen from whole genome quantile plots and asymmetry checks

Results: DEG genes dataset A



Results: DEG genes dataset B



Implications

Prediction	Consequence	Benefits
Accuracy in the expression estimation depends on on the organism and tissue	no general recommendations are valid	Don't be fooled
Shorter reads (but more) are better to estimate expression	The tendency of increasing read length may not be optimal for RNAseq	Get better results at the same cost
Fold changes between groups are unknown	sample size calculation is impossible without a pilot	Don't be fooled!
DEG genes are gained when adding more samples to an existing dataset and not by increasing the number of reads per sample	Sequence more samples, not reads	Spend money for a benefit
No more DEG are gained when adding more samples to an existing dataset	Do not sequence more	Do not spend money

Conclusions

- With the advances in technology in biology little importance has been given to the optimal use of resources
- Almost no research has been done on sample size calculation
- Researchers believe that money=results... oh my my
- Batch and technical effects are rarely taken into account leading to no or wrong results

Bioconductor package: casper

http://www.bioconductor.org/packages/2.12/bioc/html/casper.html

Rossell D., Stephan-Otto Attolini C., Kroiss M., Stöcker A. (2014) **Quantifying alternative splicing from paired-end RNA-seq data**. Annals of Applied Statistics, 8:1, 309-330.

Size calculations:

C. Stephan-Otto Attolini, V. Peña and D. Rossell. (submitted)

Joint work with:

David Rossell, IRB
Manuel Kroiss, Almond Stocker, Ludwig Maximilians Universitaet
Victor Peña, University of Duke

Thank you