RNA-seq: From (good) experimental design to (accurate) gene expression abundance.

Steve Munger
Narayanan Raghupathy

The Jackson Laboratory
21st Century Mouse Genetics
11 August 2016

Outline

General overview of RNA-seq analysis.

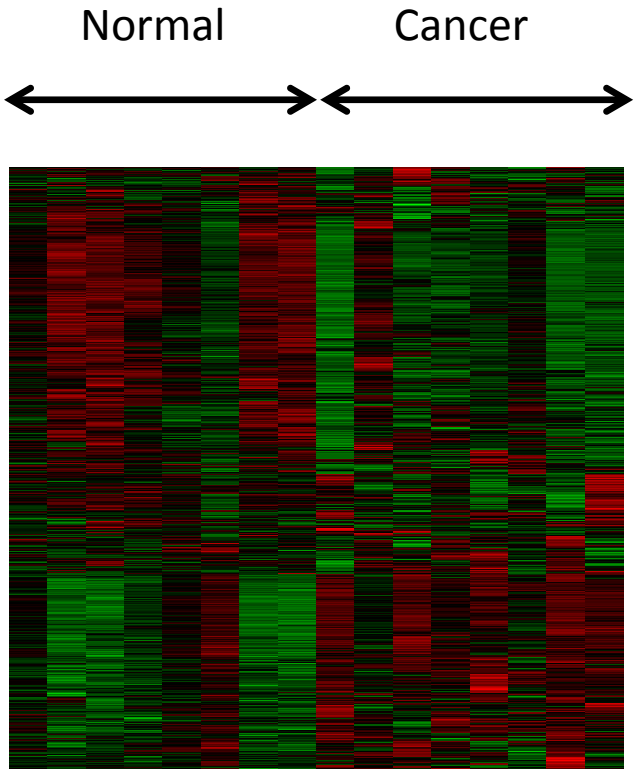
- Introduction to RNA-seq
- The importance of a good experimental design
- Quality Control
- Read alignment
- Quantifying isoform and gene expression
- Normalization of expression estimates

RNA-seq: Sequencing Transcriptomes

mRNA

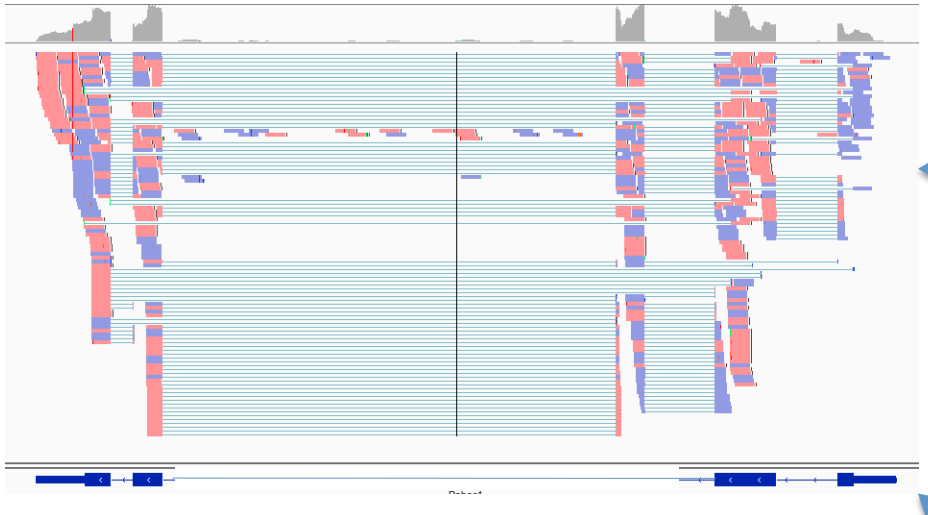
ATGCTCA AGCTA
TAGATGCTCAAGCTA
ATGCTCAAGCTAATC
ATGCTCAAGCTA
AGTAGATGCTCAAGCTA
ATGCTCAAGCTA
ATGCTCA AGCTA
ATGCTCAAGCTA
TAGATGCTCAAGCTAATC
CTCAAGCTAATCCTAG

Applications of RNA-seq Technology



Differential Gene expression analysis

Applications of RNA-seq Technology

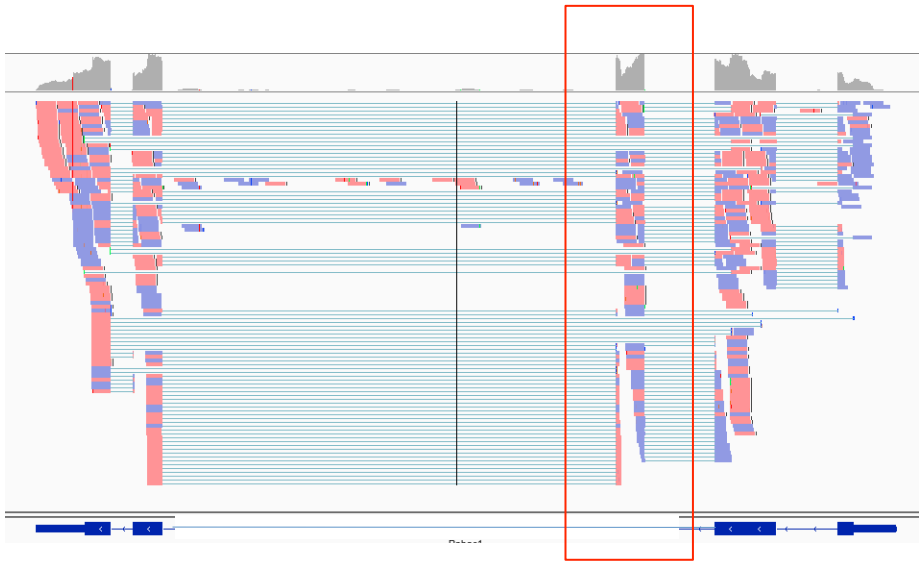


Evidence from RNA-seq

Annotated gene

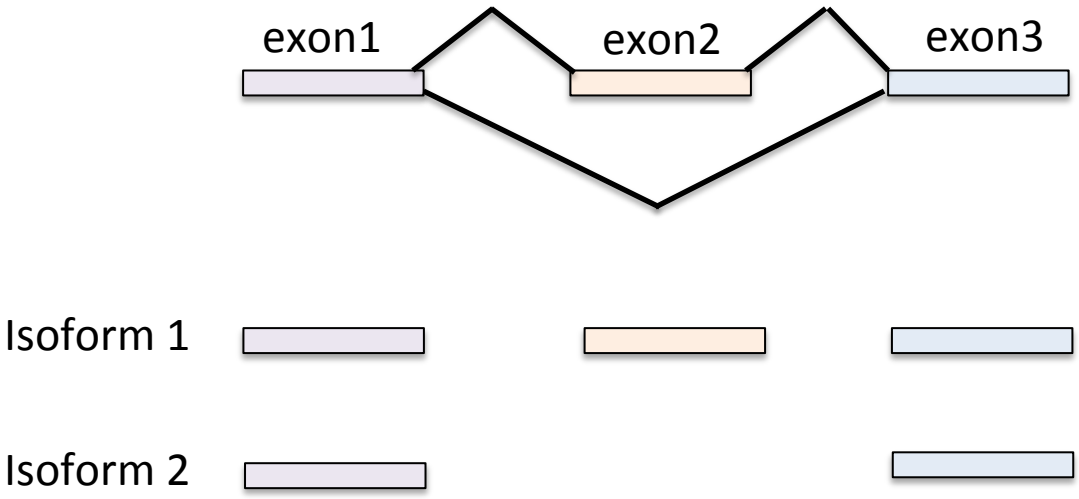
Novel exon discovery

Applications of RNA-seq Technology



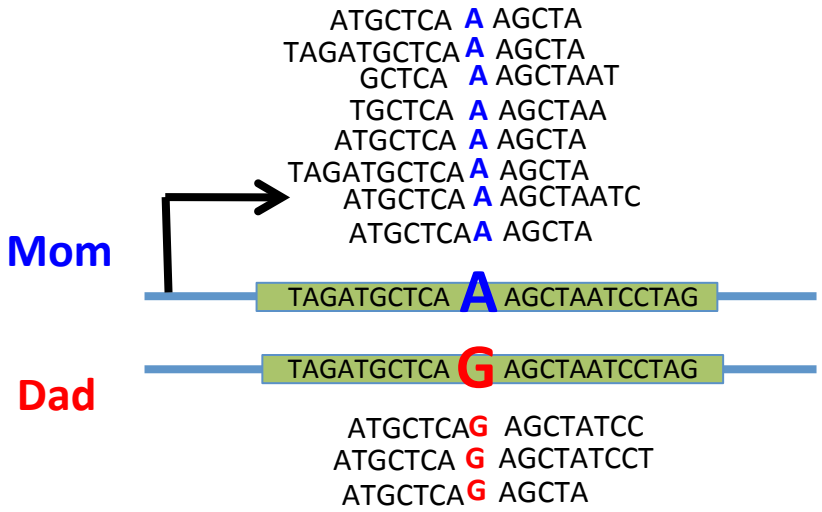
Novel exon discovery

Applications of RNA-seq Technology



Alternative splicing

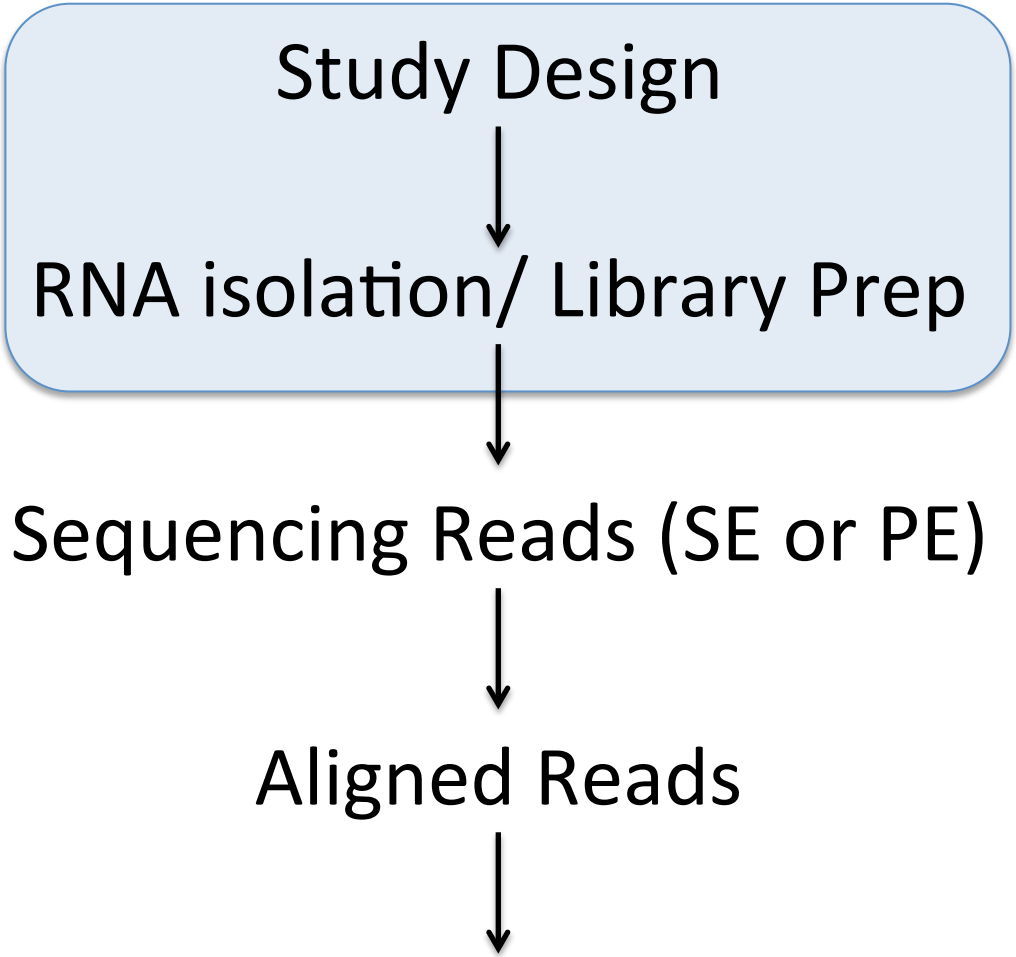
Applications of RNA-seq Technology



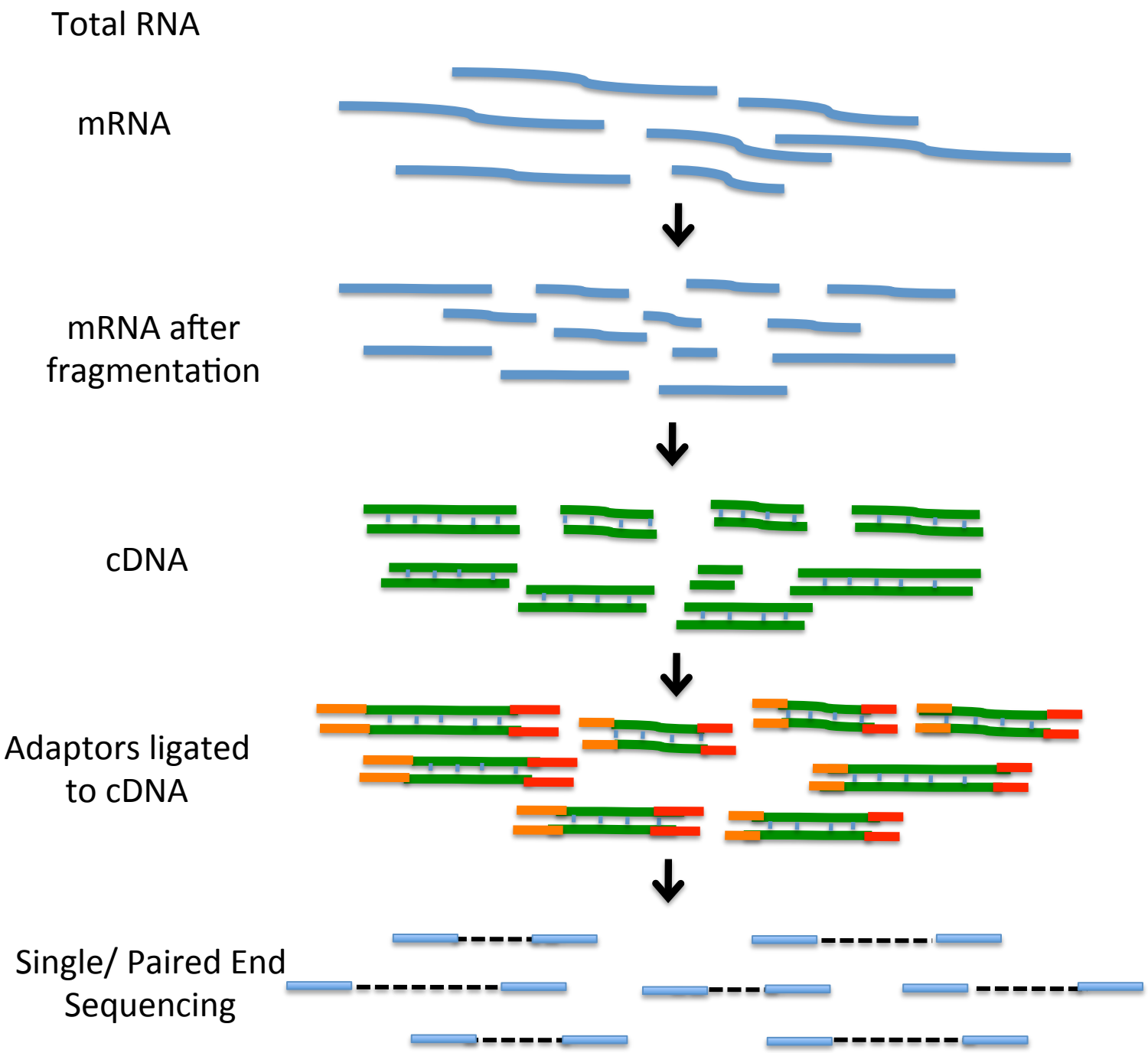
Allele-Specific gene Expression (ASE)

Preferential expression of one allele over the other.

RNA-seq Work Flow



RNA-Seq


Total RNA
mRNA
mRNA after fragmentation
cDNA
Adaptors ligated to cDNA
Single/ Paired End Sequencing

N

Know your application – Design your experiment accordingly

- How many reads? Read depth
- Single-end or Paired-end sequencing?
- Read length?
- How many samples?

N

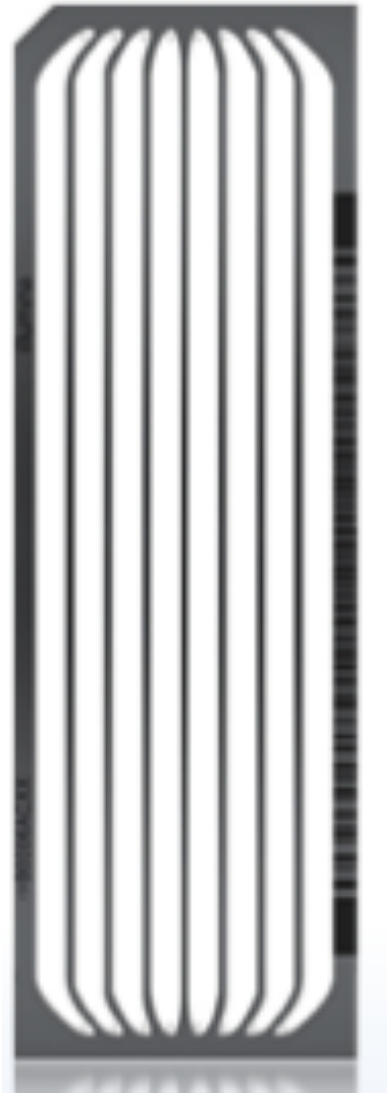
RNA-seq Experimental design

- Differential expression of highly expressed and well annotated genes?
  - Smaller sample depth; more biological replicates
  - No need for paired end reads; shorter reads (50bp) may be sufficient.
  - Better to have 20 million 50bp reads than 10 million 100bp reads.
- Looking for novel genes/splicing/isoforms?
  - More read depth, paired-end reads from longer fragments.

N

Good Experimental Design

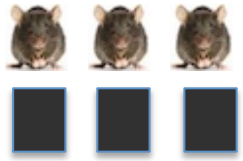
Multiplexing
Replication
Randomization



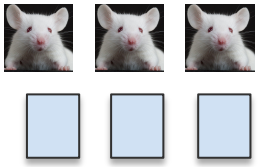
Illumina flowcell

N

RNA-Seq Experimental Design: Randomization

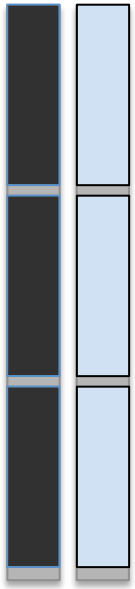


Experimental Group 1



Experimental Group 2

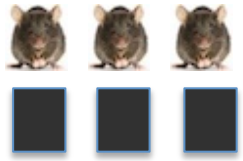
Two Illumina Lanes



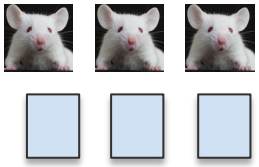
Bad Design

N

RNA-Seq Experimental Design: Randomization

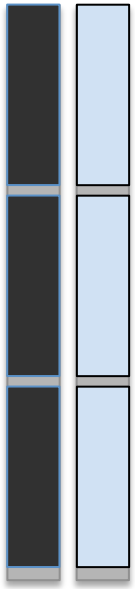


Experimental Group 1

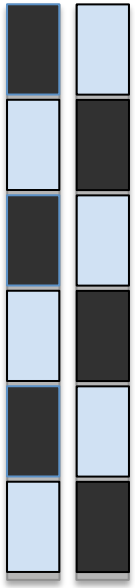


Experimental Group 2

Two Illumina Lanes



Bad Design

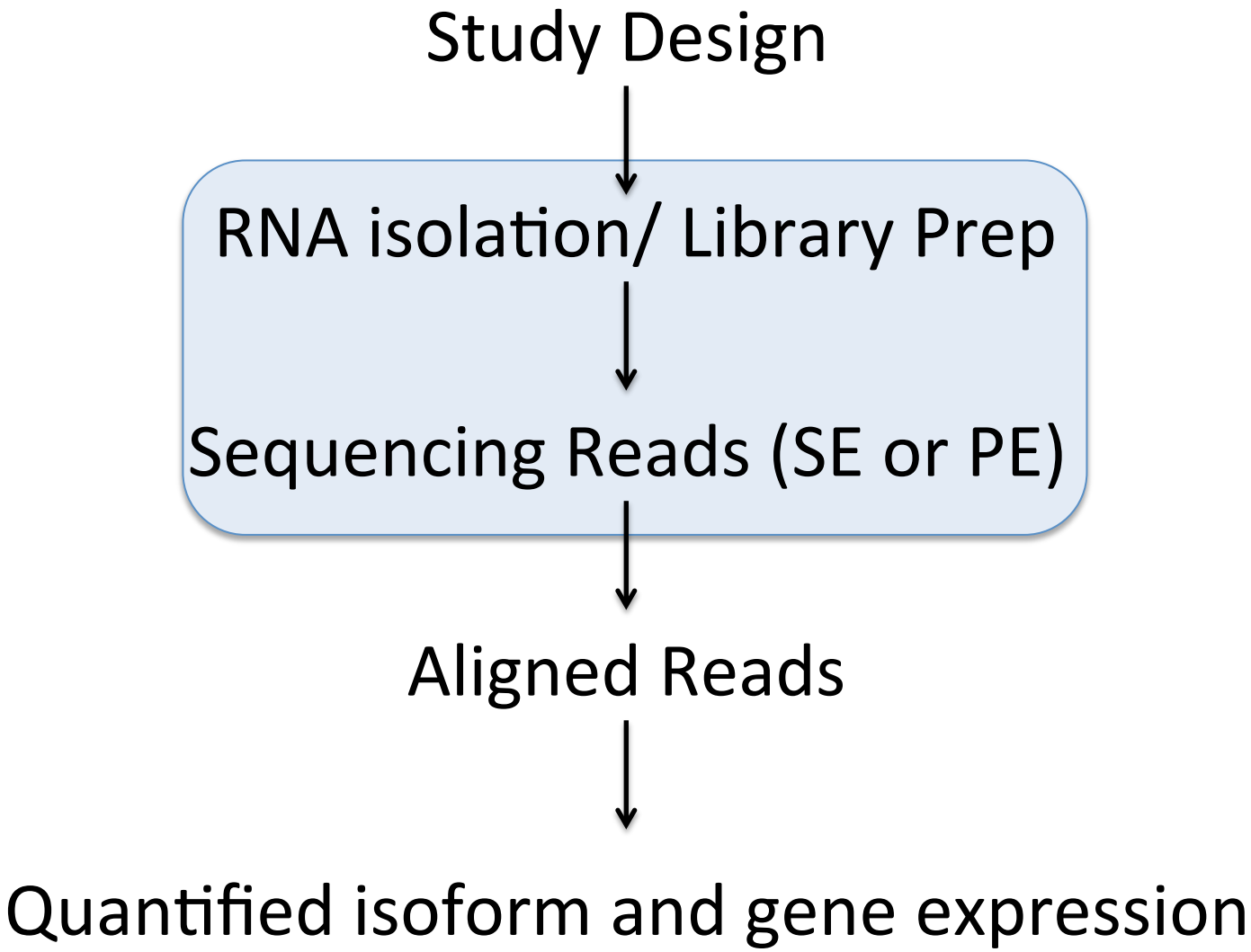


Better Design

Mouse ENCODE reanalysis: http://f1000research.com/articles/4-121/v1

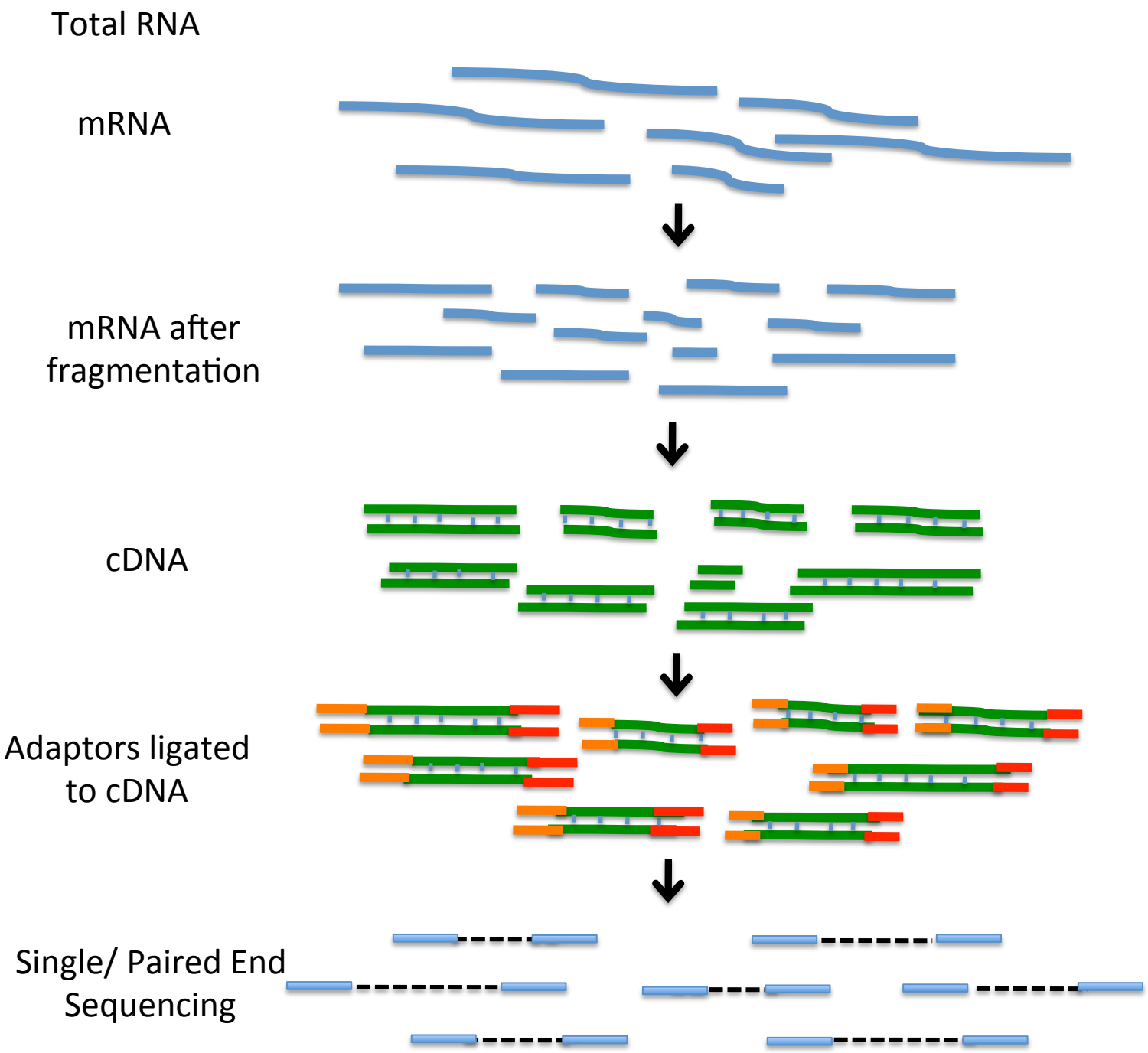
N

RNA-seq Work Flow



N

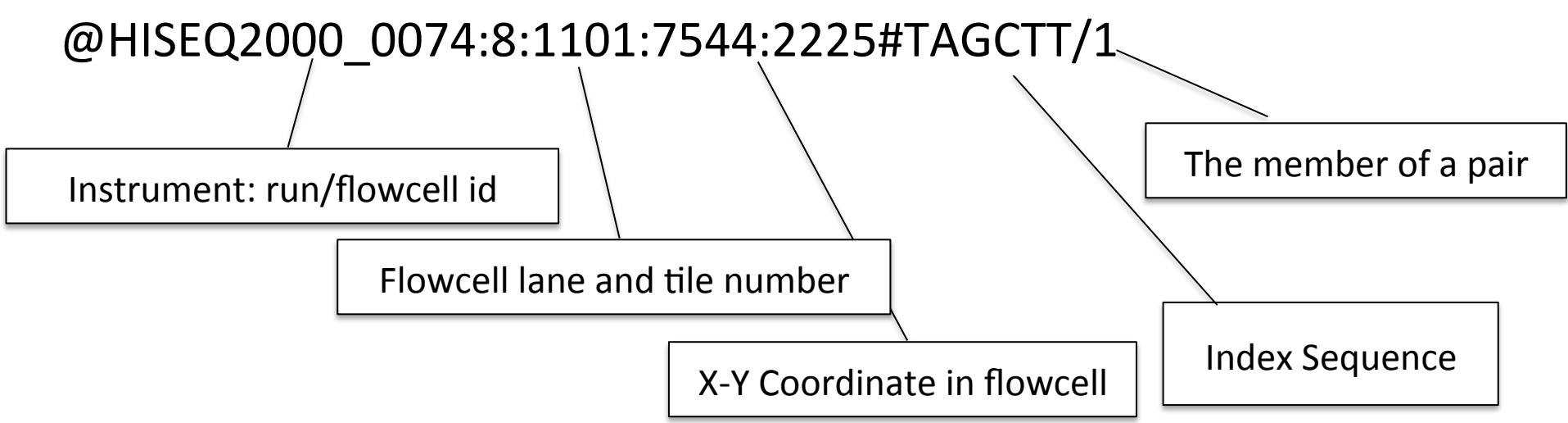
RNA-Seq


Total RNA
mRNA
mRNA after fragmentation
cDNA
Adaptors ligated to cDNA
Single/ Paired End Sequencing

N

Millions and millions of reads…

```
@HISEQ2000_0074:8:1101:7544:2225#TAGCTT/1
TCACCCGTAAGGTAACAAACCGAAAGTATCCAAAGCTAAAAGAAGTGGACGACGTGCTTGGTGGAGCAGCTGCATG
+
CCCFFFFFHHHHHDHHJJJJJJJJJIJJ?FGIIIJJJJJJIJJJJJJFHIJJJIJHHHFFFFD>AC?B??C?ACCAC>BB<<<>C@CCCACCCDCCIJ
```



Phred Score:

$Q = -10 \log_{10} P$

10 indicates 1 in 10 chance of error

20 indicates 1 in 100,

30 indicates 1 in 1000,

SN

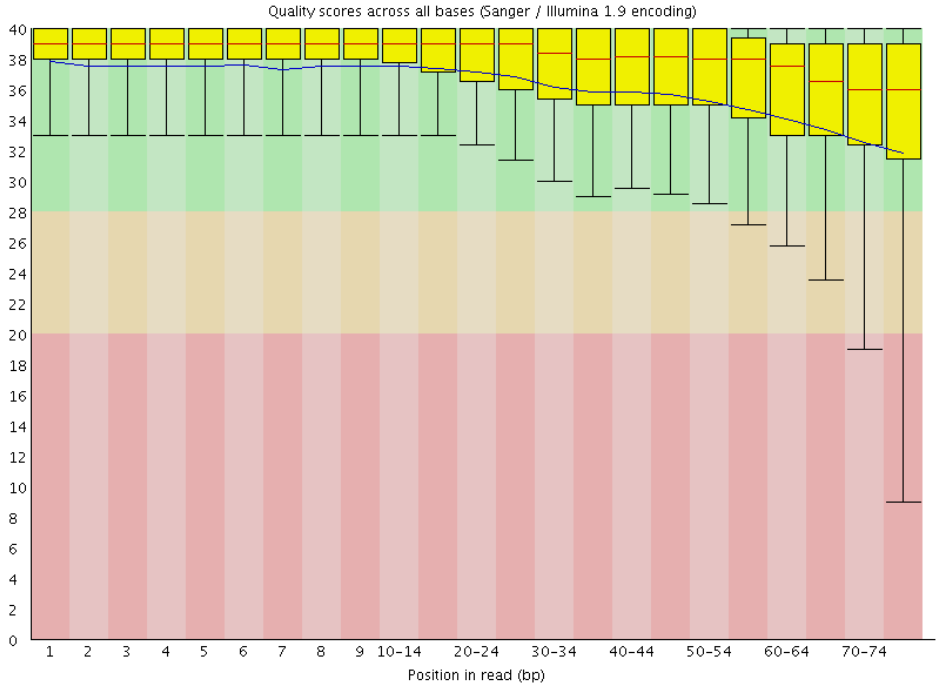
Quality Control: How to tell if your data is clean

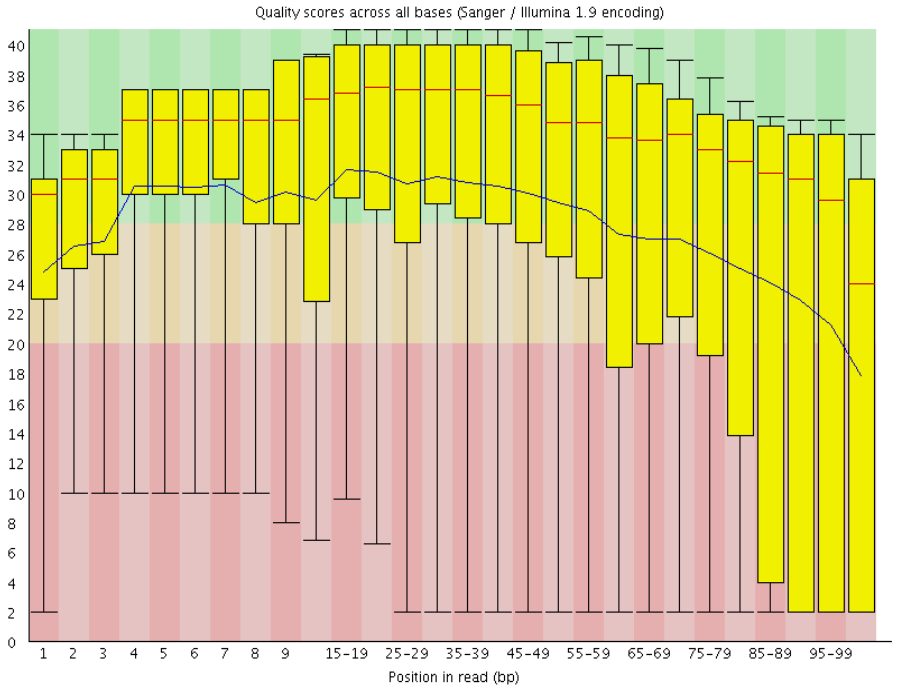
- FASTX-Toolkit
  - http://hannonlab.cshl.edu/fastx_toolkit/
- FastQC
  - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

RNA-seq Data: ftp://ftp.jax.org/dgatti/MouseGen2016/

- B6-100K.fastq and Cast-100K.fastq

Quality Control: How to tell if your data is clean





Good data

- **Consistent**
- **High Quality Along the reads**

Bad data

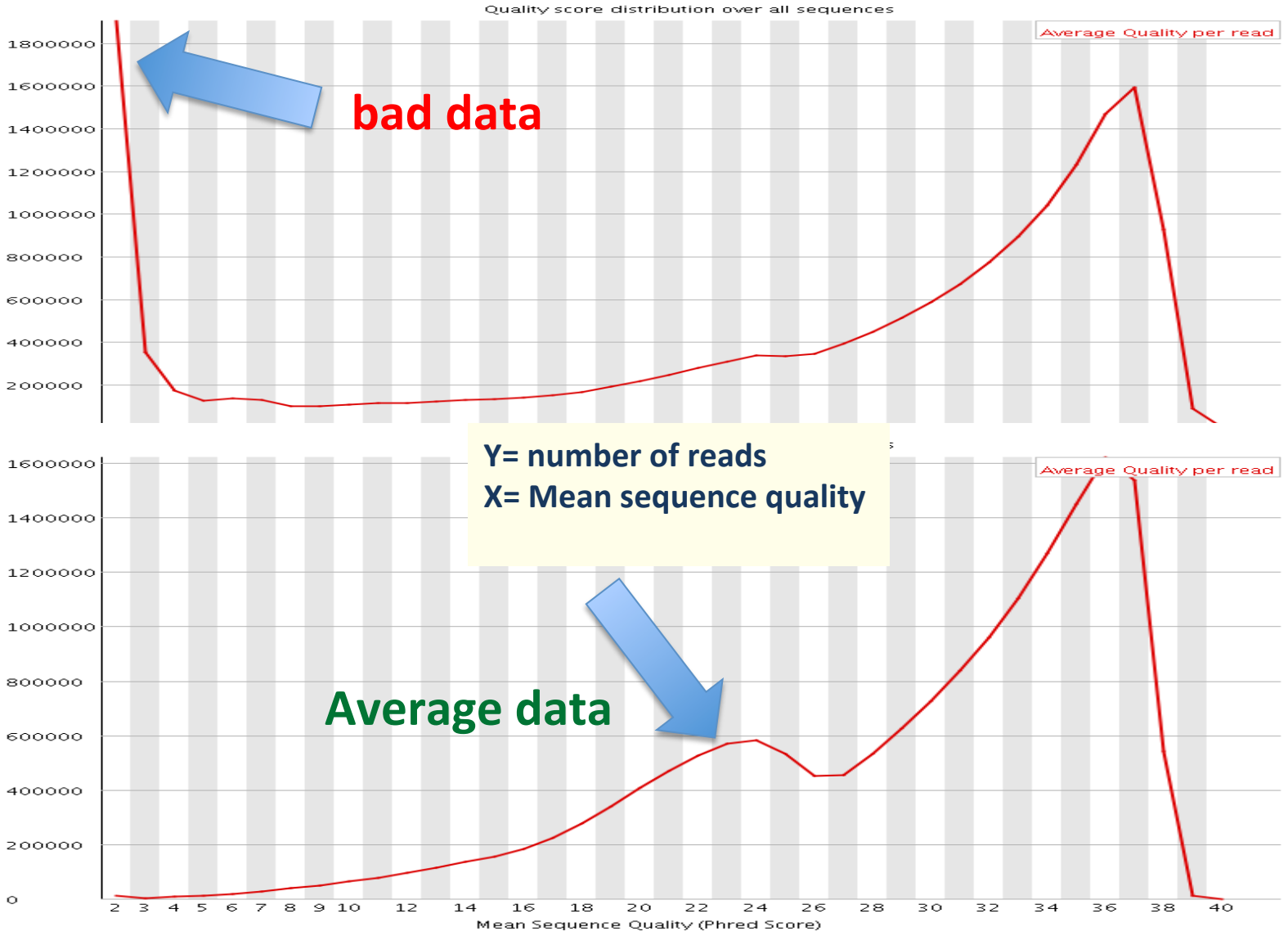
- **High Variance**
- **Quality Decrease with Length**

RNA-seq Data: ftp://ftp.jax.org/dgatti/MouseGen2016/

- B6-100K.fastq and Cast-100K.fastq

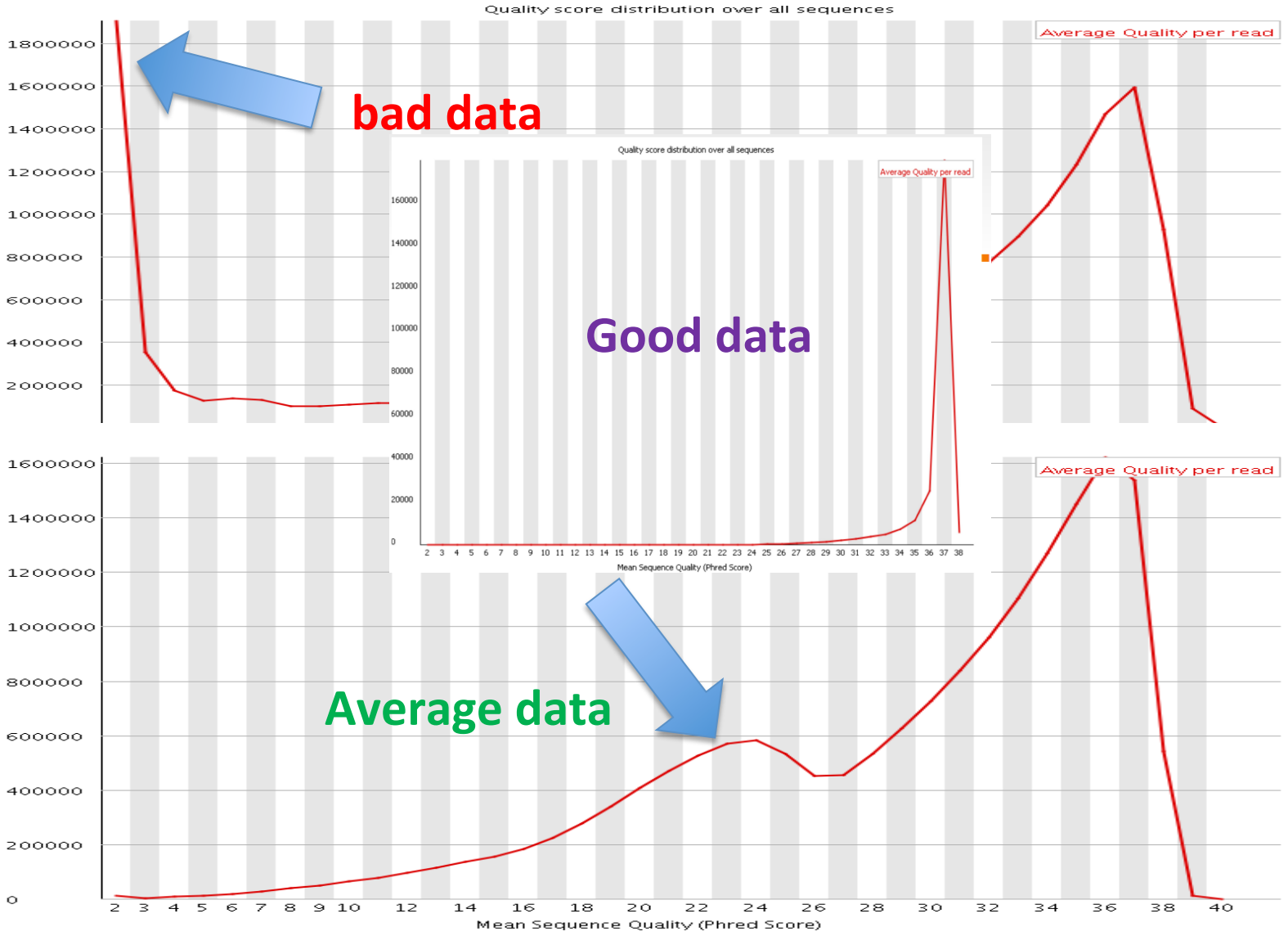
S

Per sequence quality distribution



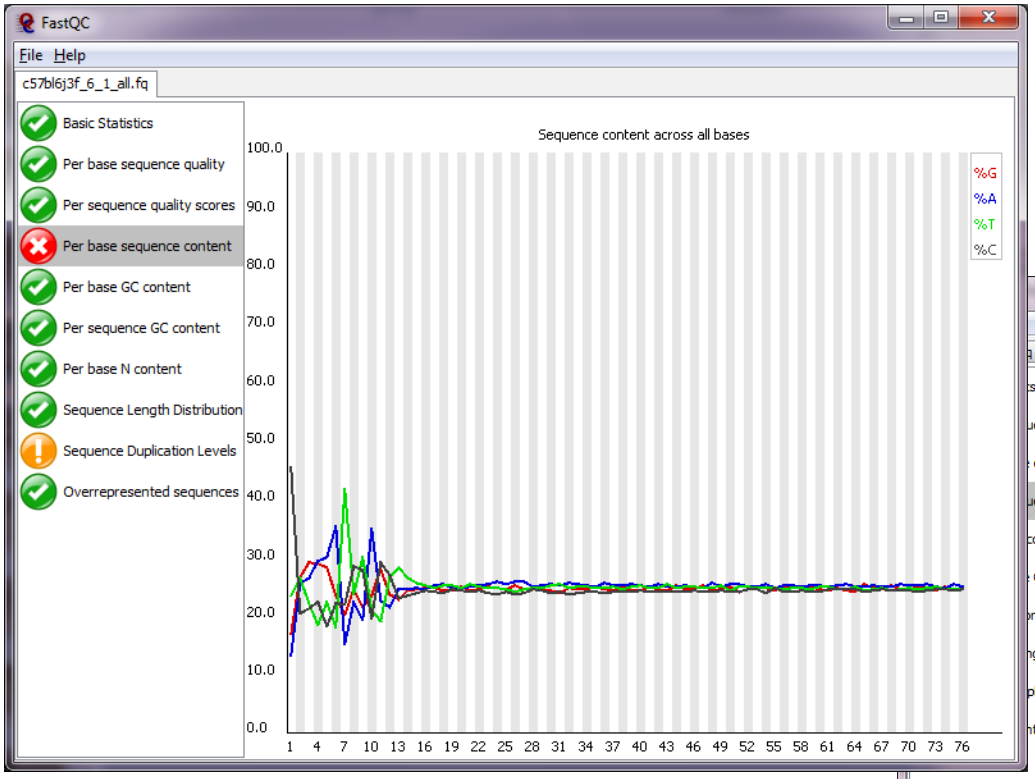
S

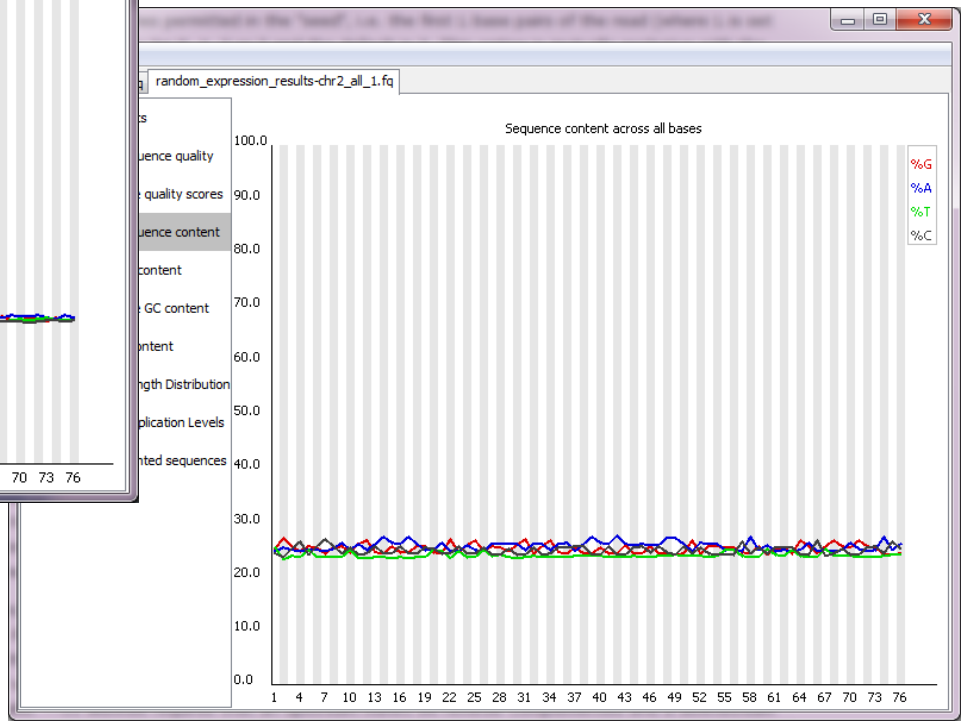
Per sequence quality distribution



S

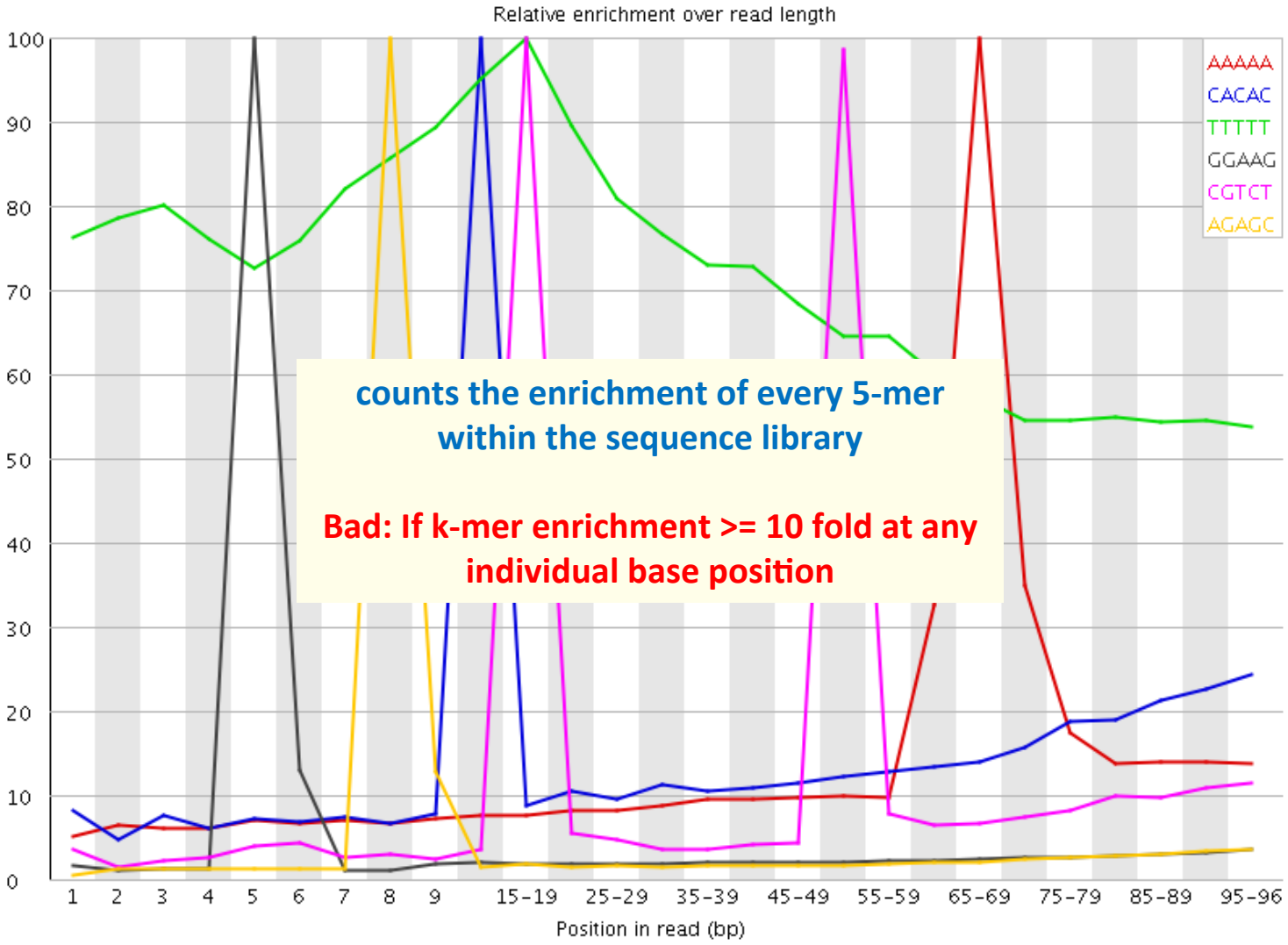
Quality Control: Sequence Content Across Bases



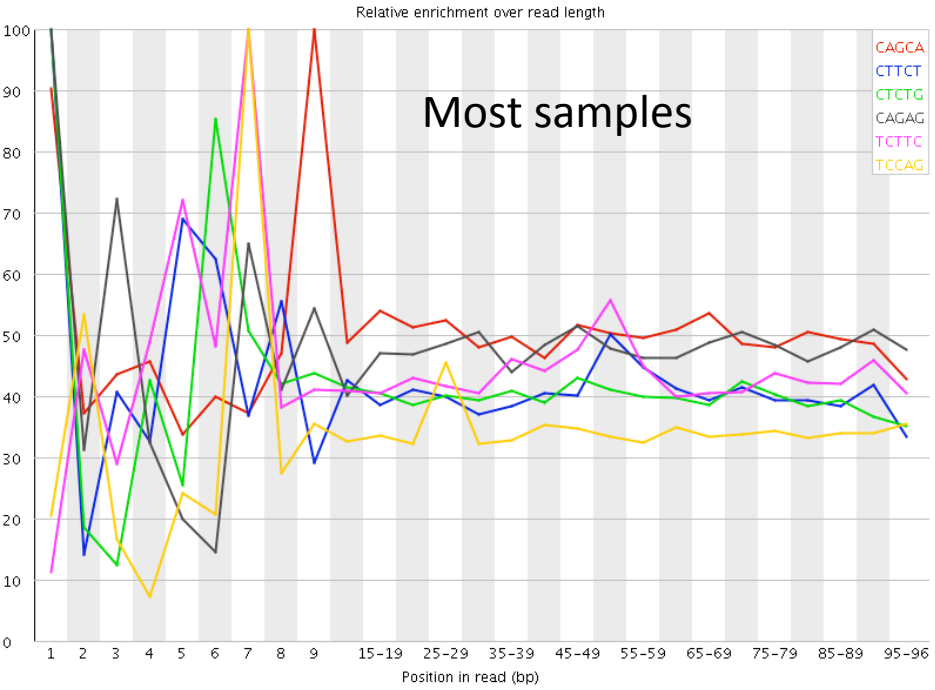


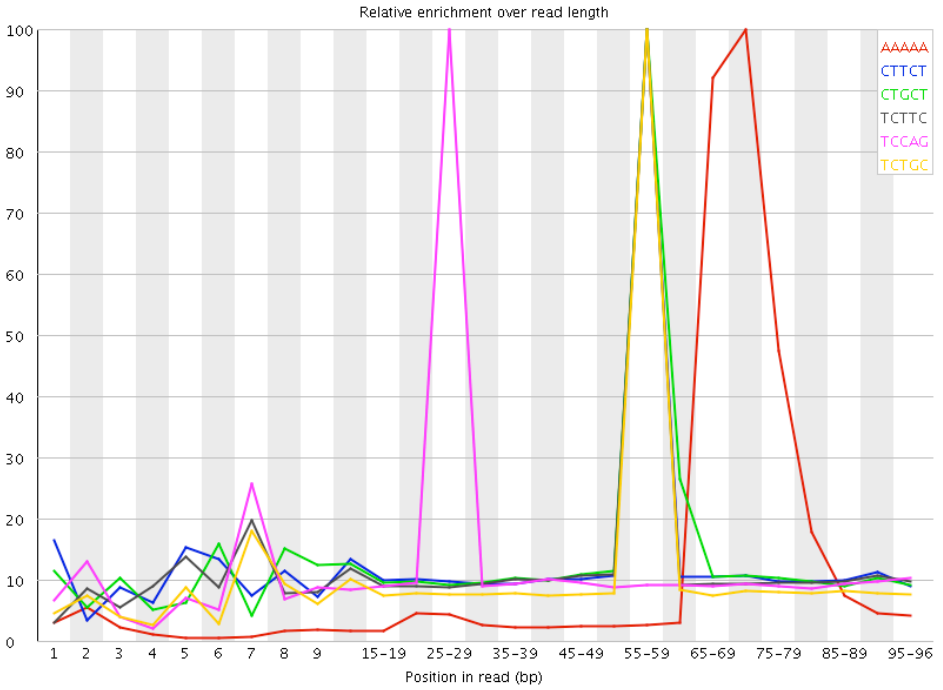
S

K-mer content

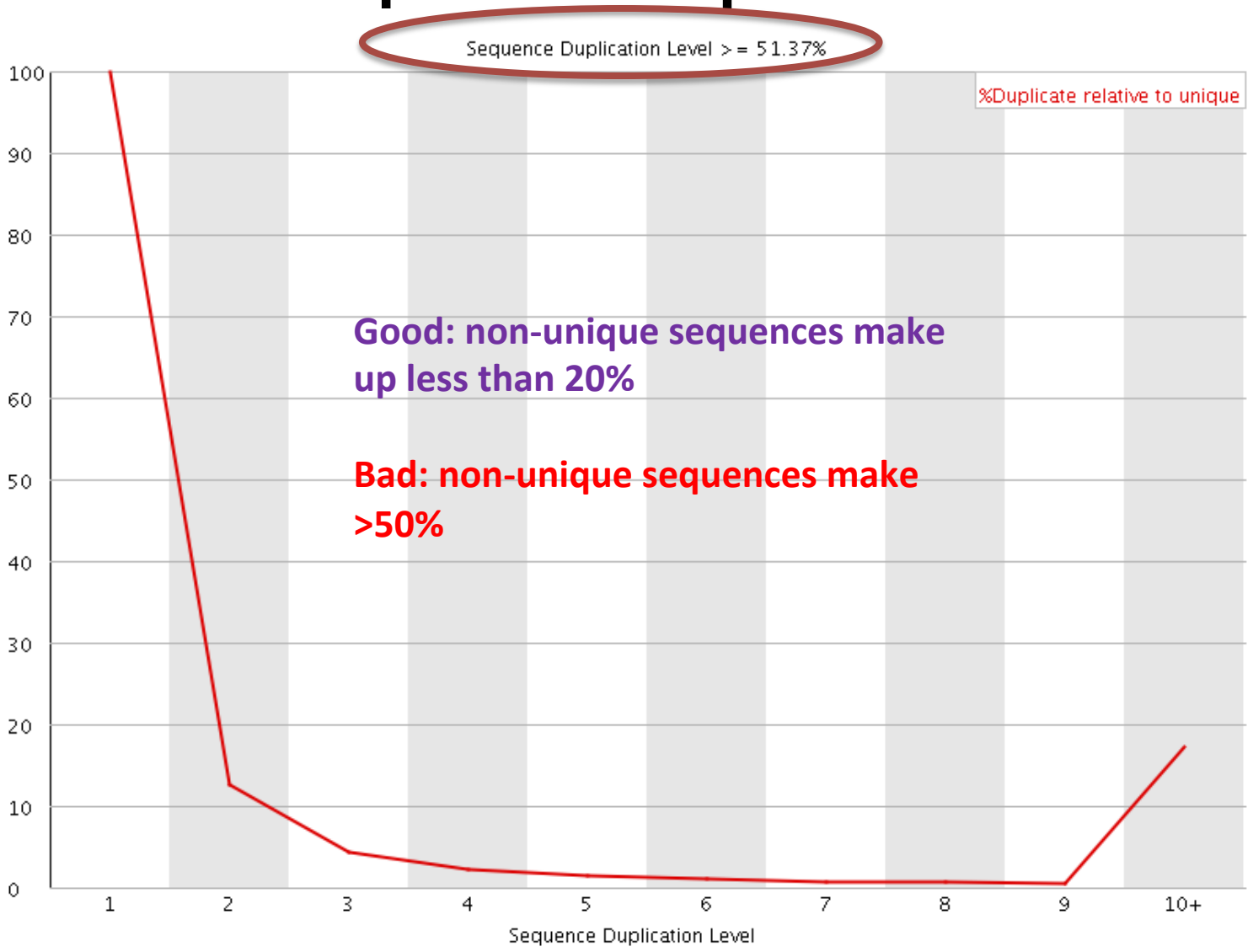


K-mer content





Duplicated sequences



Good: non-unique sequences make up less than 20%

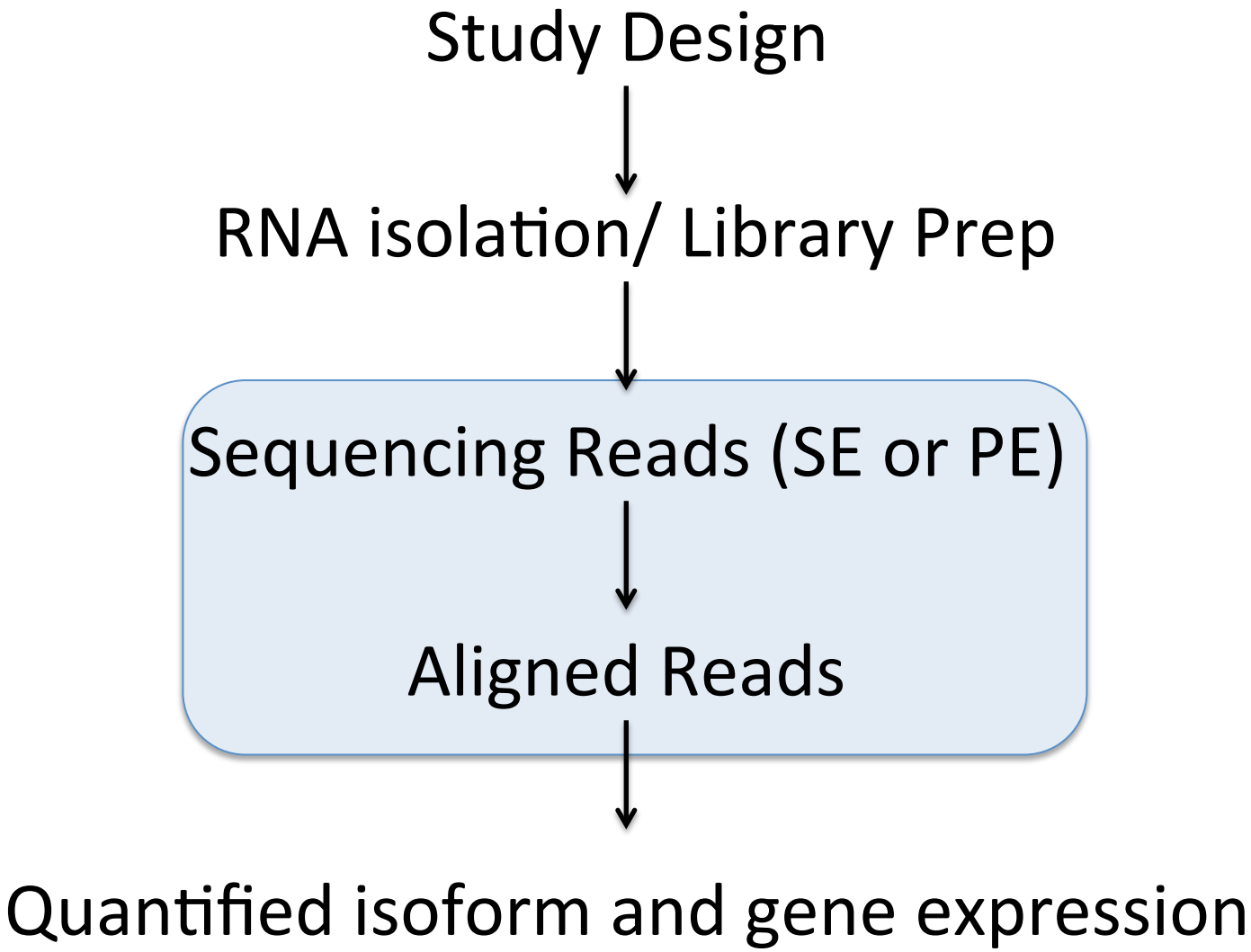
Bad: non-unique sequences make >50%

S

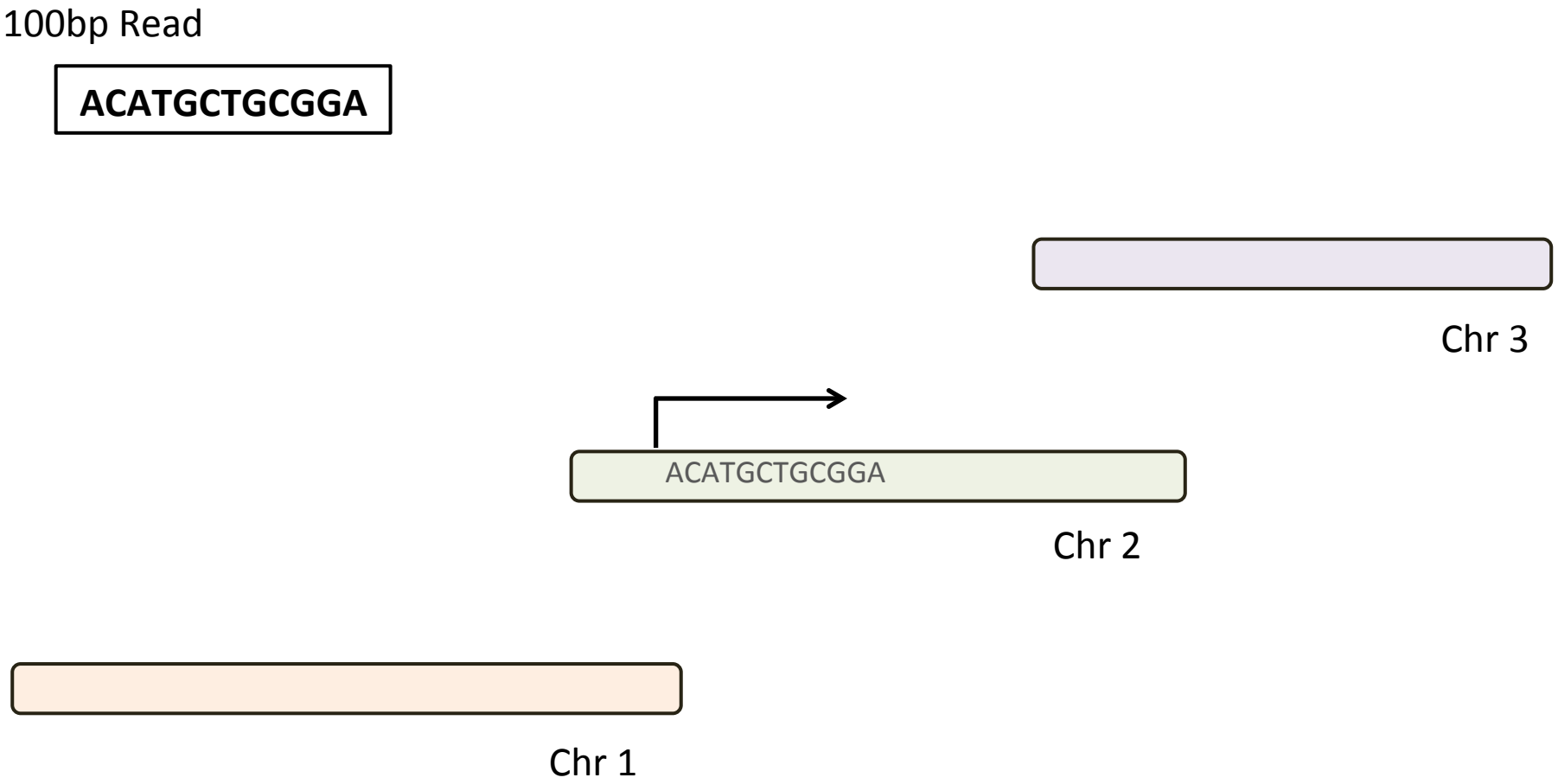
Tradeoffs to preprocessing data

- Signal/noise -> Preprocessing can remove low-quality “noise”, but the cost is information loss.
  - Some uniformly low-quality reads map uniquely to the genome.
  - Trimming reads to remove lower quality ends can adversely affect alignment, especially if aligning to the genome and the read spans a splice site.
  - Duplicated reads or just highly expressed genes?
  - Most aligners can take quality scores into consideration.
  - Currently, we do not recommend preprocessing reads aside from removing uniformly low quality samples.

RNA-seq Work Flow

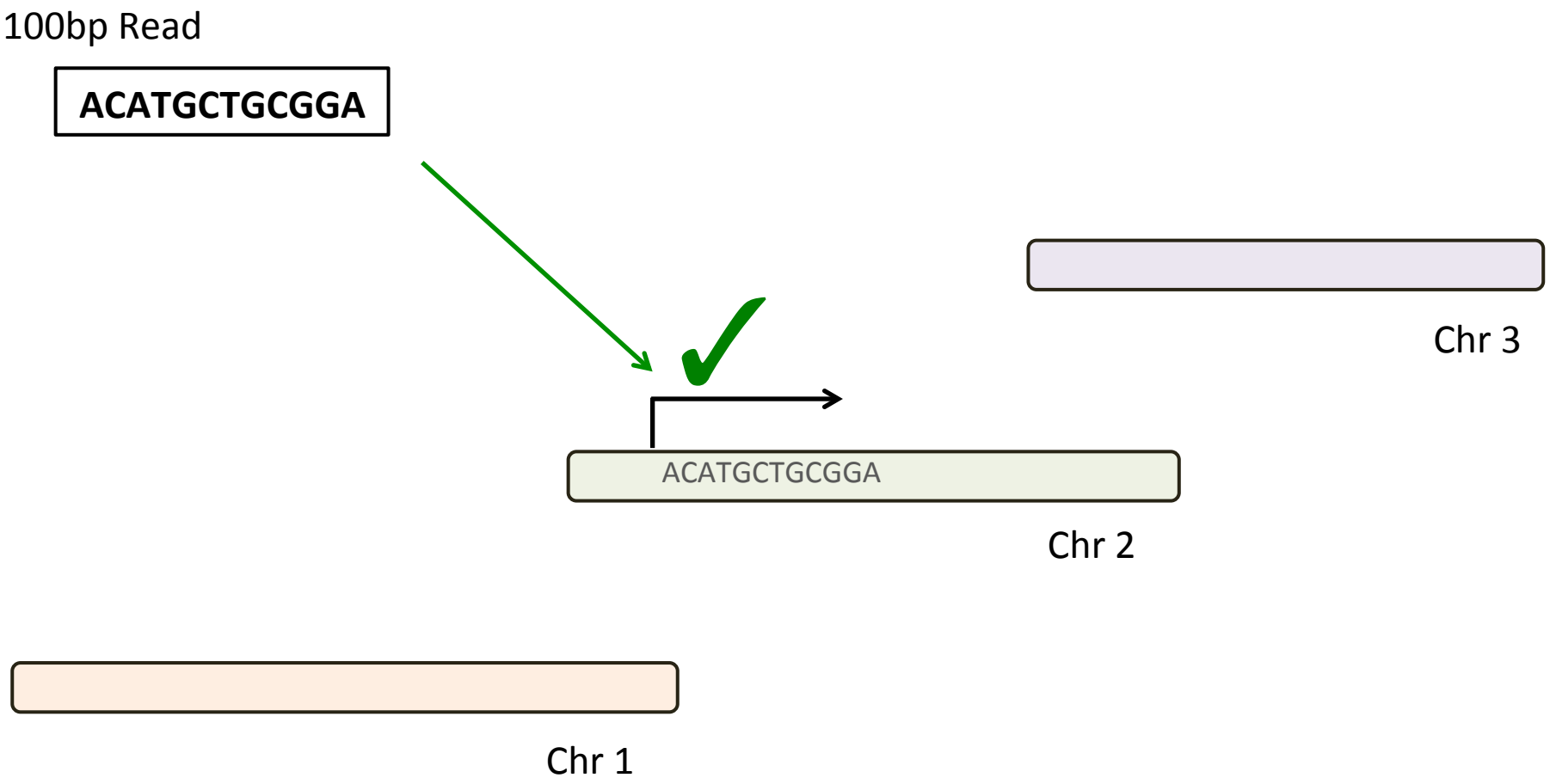


Alignment 101



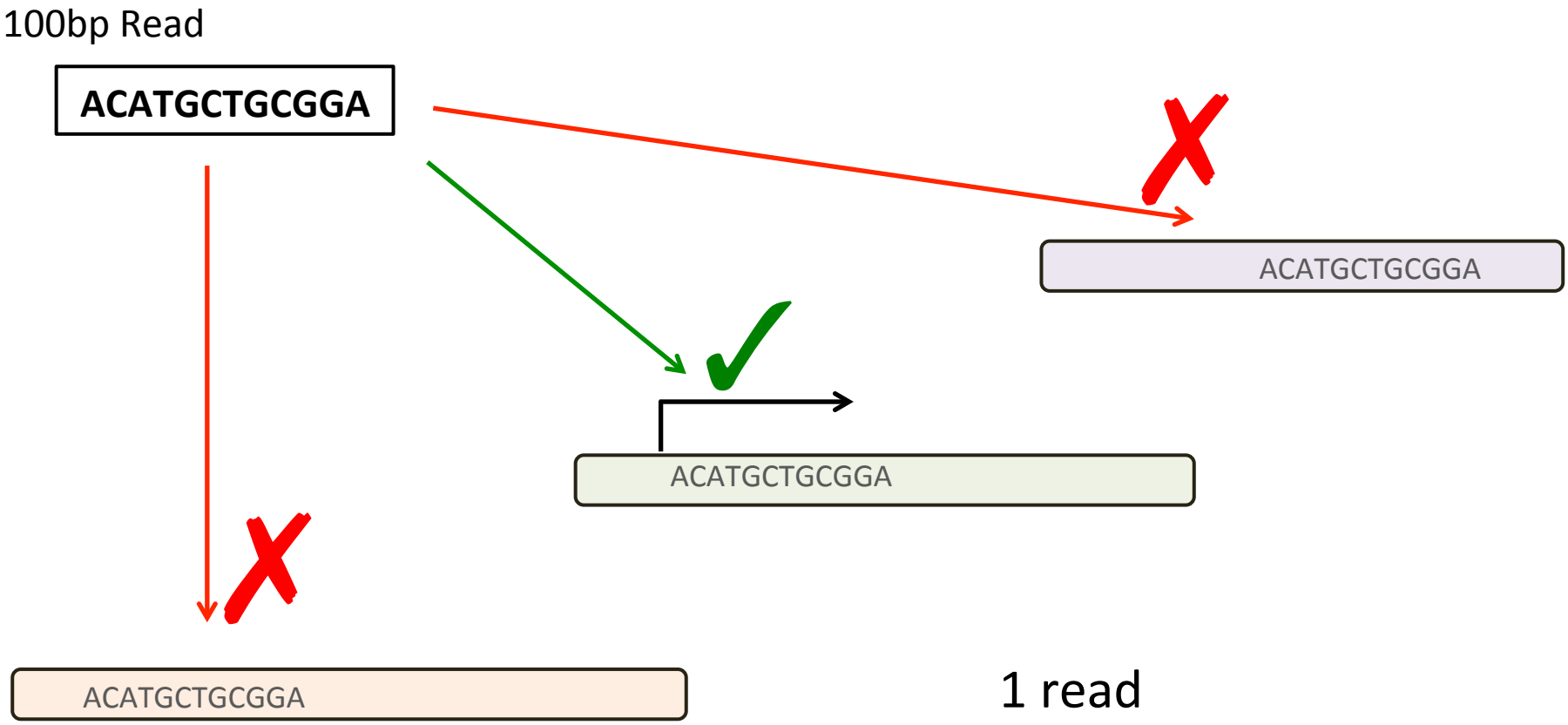
S

The perfect read: 1 read = 1 unique alignment.



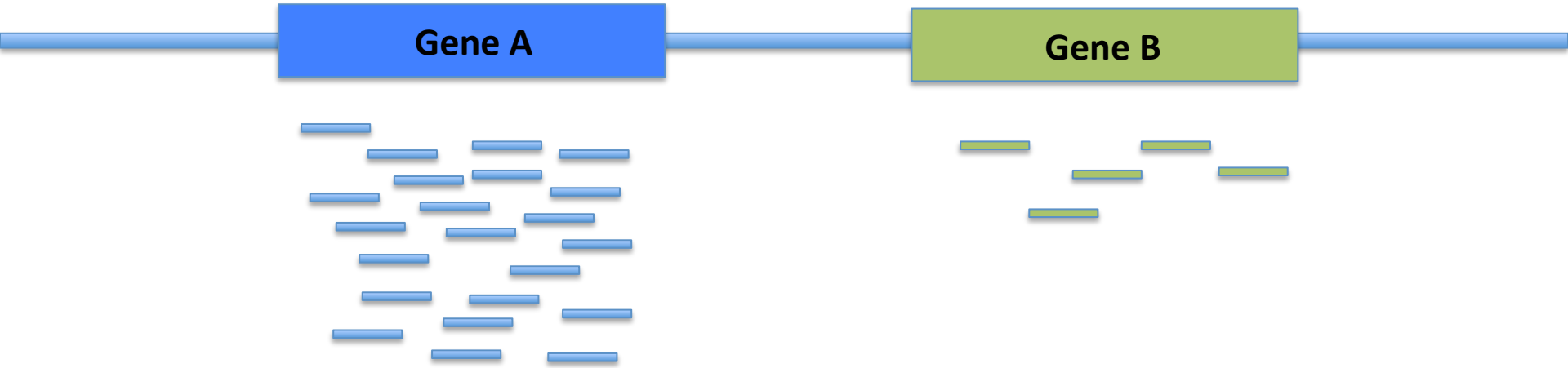
S

Some reads will align equally well to multiple locations. “Multireads”



1 read
3 valid alignments
Only 1 alignment is correct

Aligning Millions of Short Sequence Reads

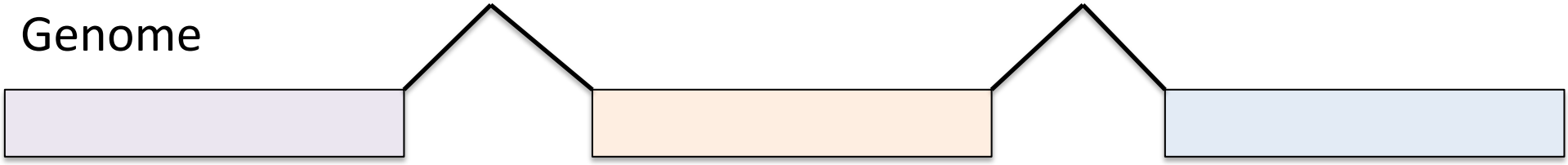


Aligners: Bowtie, GSNAP, STAR, BWA, BLAT, HISAT2, Bowtie2, Kallisto, Salmon

N

Align to Genome or Transcriptome?

Genome



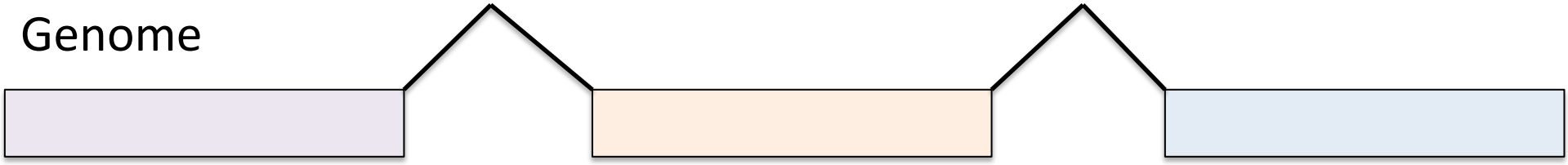
Advantages: Can align novel isoforms.
Disadvantages: Difficult, Spurious alignments, spliced alignment, gene families, pseudo genes

Transcriptome

N

Align to Genome or Transcriptome?

Genome



Advantages: Can align novel isoforms.
Disadvantages: Difficult, Spurious alignments, spliced alignment, gene families, pseudo genes

Transcriptome

Advantages: Easy, Focused to the part of the genome that is known to be transcribed.
Disadvantages: Reads that come from novel isoforms may not align at all or may be misattributed to a known isoform.

N

Output of most aligners: Bam/Sam file of reads and genome positions

```
HISEQ2000:113:D0636ACXX:1:2308:15958:82100      409     1       3035058 0       100M    *       0       0
       AACCTGGGTATGCCTCGTAGTTAAAACATTCCTGGGAACATCTTGACCATAAGATAAAGGGGACTGTGAAGACATAGCAGGGCTATATTATCTAAGTCAA
    5:?<DC?C<CBADDDECEA=IGEHA@JJJIHEIGGIIHJIHEJJJJIIIIIIJJIIJJJIIGFJJIJJIJIHJIIJJIJIJJJJIGJHHHFHFFFFFCCC
NM:i:0  NH:i:5  CC:Z:10 CP:i:4143337    HI:i:0
HISEQ2000:113:D0636ACXX:1:2206:6975:110266      163     1       3035206 3       100M    =       3035297 191
     GTAAAAGTCACACATCAACTGGTTGCTATGTGAACAAAGATAAGCCCCCAGCCCACAGGAACAAAGTCCTGATGCACTGTGTTCTTTCTGTTAATGTTTG
    @BCFFFFDHHHHHJJIJJJJJJIJJJJJJJIJJJJIIJJJJJJJIJJJIJJJJJJJJJJIJIJJHHHHHHFFFFFEEEEEECDDEDDDDEDDDEDEEE<
NM:i:0  NH:i:2  CC:Z:4  CP:i:118529266  HI:i:0
HISEQ2000:113:D0636ACXX:1:2206:6975:110266      83      1       3035297 3       100M    =       3035206 -191
    TAATGTTTGAATAAGCCAATAGTGTGTTGCTATGCTGAATTCCACACCCCTAAGCCCCGTACCCCATAAAAGCCCCTGGCTTTCGAGCCTCGTGGCCGGC
CCC?:DEDEEDDDDDDEDDCDD@?@DDDDDDDDDC@DDDCB?55,DDDDCDDDDFFHHB;JJJJIGJJJJIJJJJIJGIIGIJJJJHHHHHFFFFFCCC    NM:i
:2  NH:i:2  CC:Z:4  CP:i:118529175  HI:i:0
HISEQ2000:113:D0636ACXX:1:1204:3972:146753      329     1       3044627 1       100M    *       0       0
       ATTCATGGCCCATGCCGACTTTGTTTCTAGAGGACAAACAGTTTCAAGGGCTCCTGGATACCGGGGCAGATGTGACAGTAATTTCCTCAACACATTGGCC
    BCCFFFDFHHGHHJIJJFIIJJFIIIGIIGGGGGIIGIGBG0BFGIG?DHGCGHGHIIIBEHEEDDDBBBCDDDC>@C3>>BDDDECDD<?@??CC@:>:
NM:i:1  NH:i:4  CC:Z:15 CP:i:9154226    HI:i:0
HISEQ2000:113:D0636ACXX:1:2201:2762:178840      355     1       3045593 0       100M    =       3045612 119
     ATAAATTAAAAGTTTTAGATGCTTGCCAGAAACTGTTAGAAAATTTTGGATTTTAATCTTGGTTTGACAAGCTACCTCTTCTTACAAGCAGGAAAGGAAA
    CC@FFFFFHHGHDHHJJIJJJJJIJIIIJJIGIIJIIIGIIJIHIIJIIDCGHIGIJIIJJDIIIGGIIIGIIJJJEFEHFFEFFFDD?ED?7=CDA?CD
NM:i:0  NH:i:7  CC:Z:=  CP:i:16822458   HI:i:0
```

N

Visualization of alignment data (BAM/SAM)

Genome browsers – IGV and UCSC

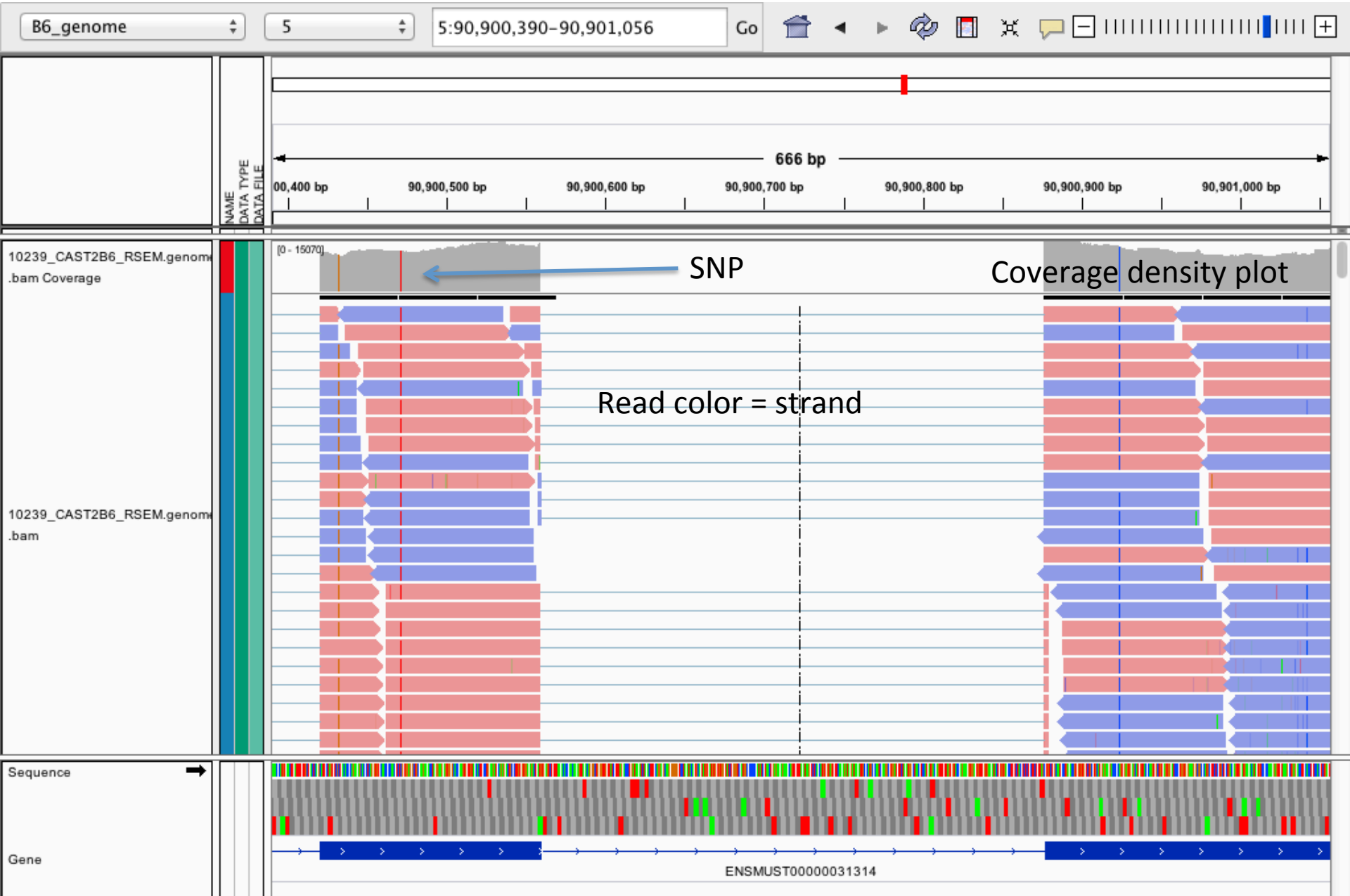
Integrative Genome Viewer (IGV)

http://software.broadinstitute.org/software/igv/download

RNA-seq Data: ftp://ftp.jax.org/dgatti/MouseGen2016/

- DO.chr1XY.sorted.bam and DO.chr1XY.sorted.bam.bai

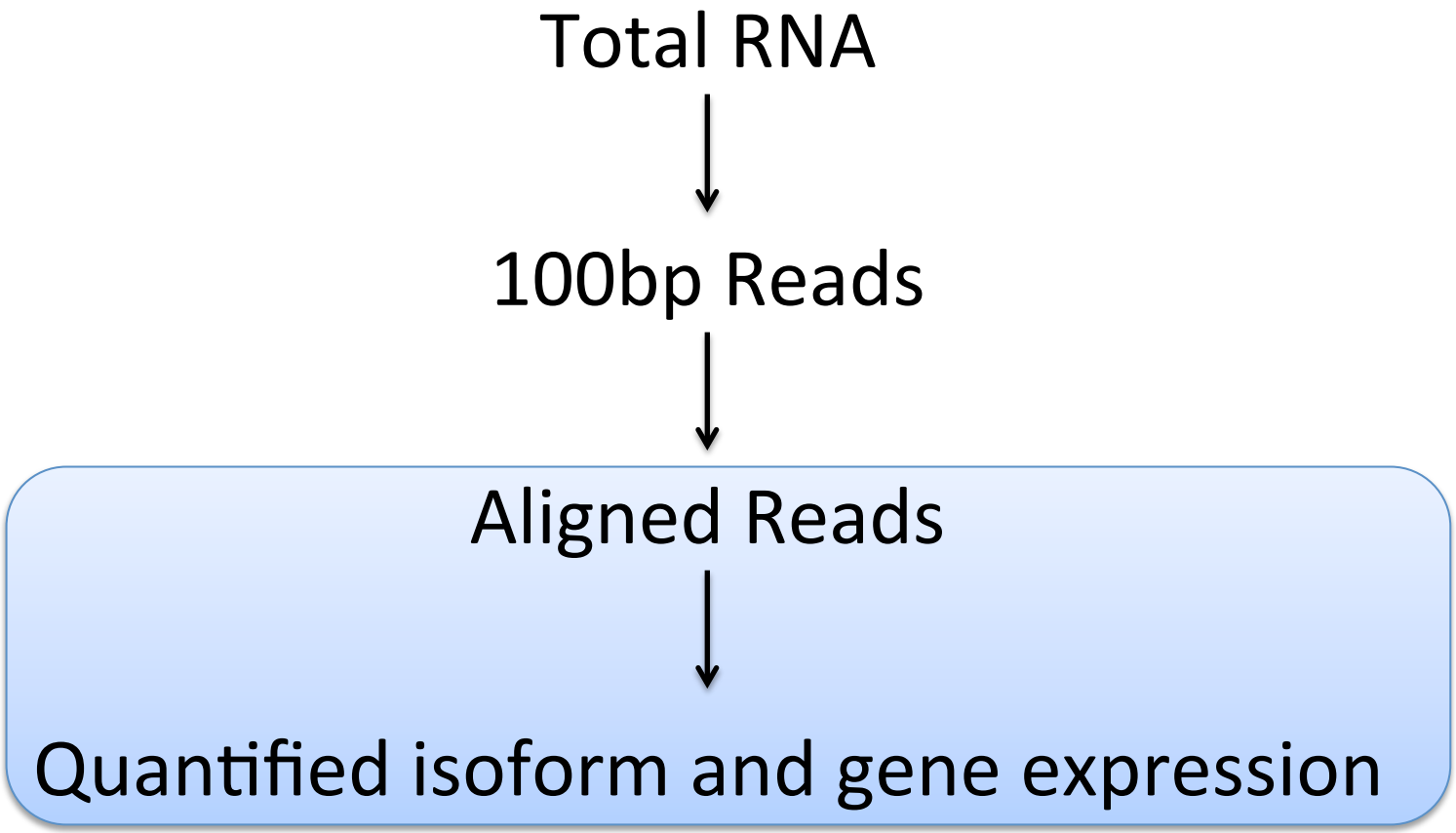
IGV is your friend.



Example genes to look at in IGV

1. Tsn
2. Gorab
3. Fmo1, Fmo2, Fmo3, Fmo4, Fmo6
4. Ids
5. Zfx
6. Ssty1, Ssty2

Aligned Reads to Gene Abundance



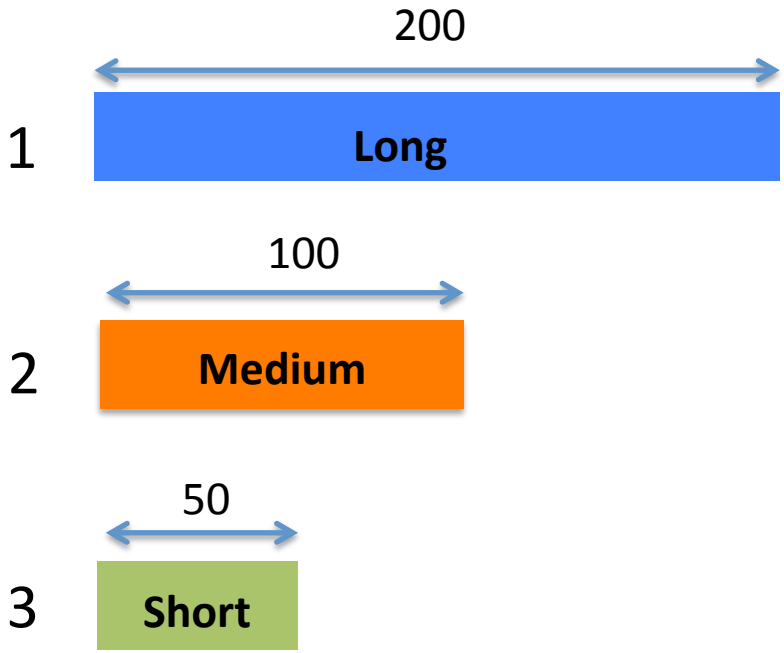
N

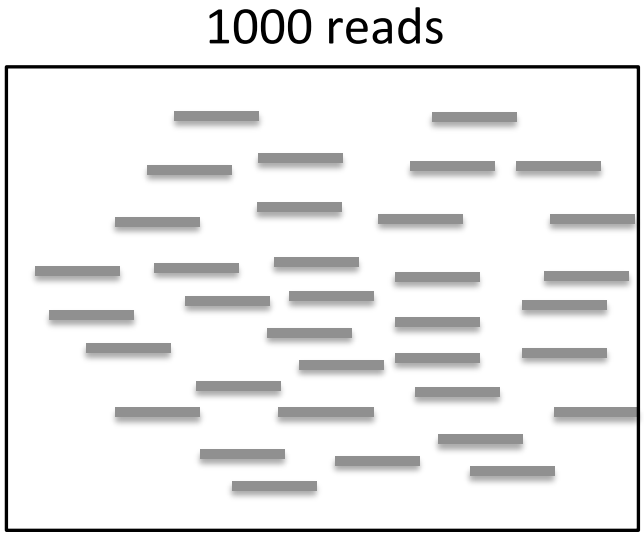
Aligned Reads to Gene Abundance: Challenges

Many approaches to quantify expression abundance

N

Aligned Reads to Gene Abundance: Challenges

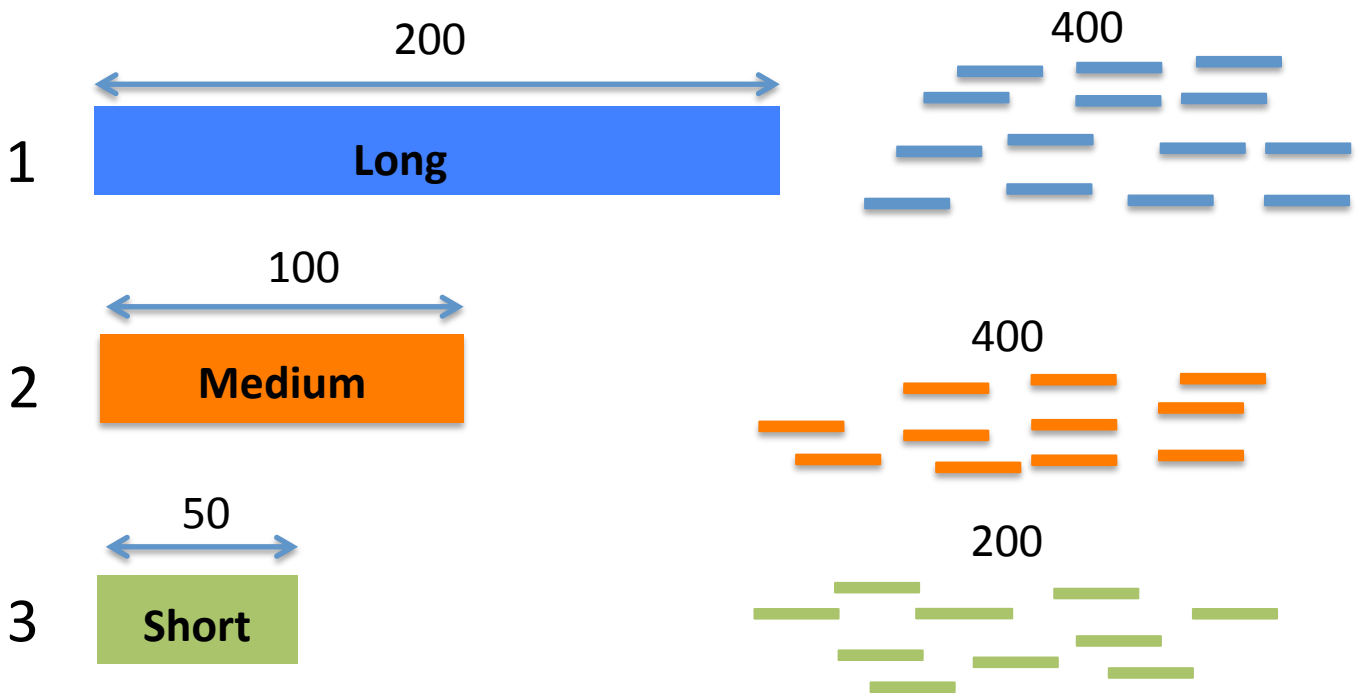




Relative abundance for these genes, f_1, f_2, f_3

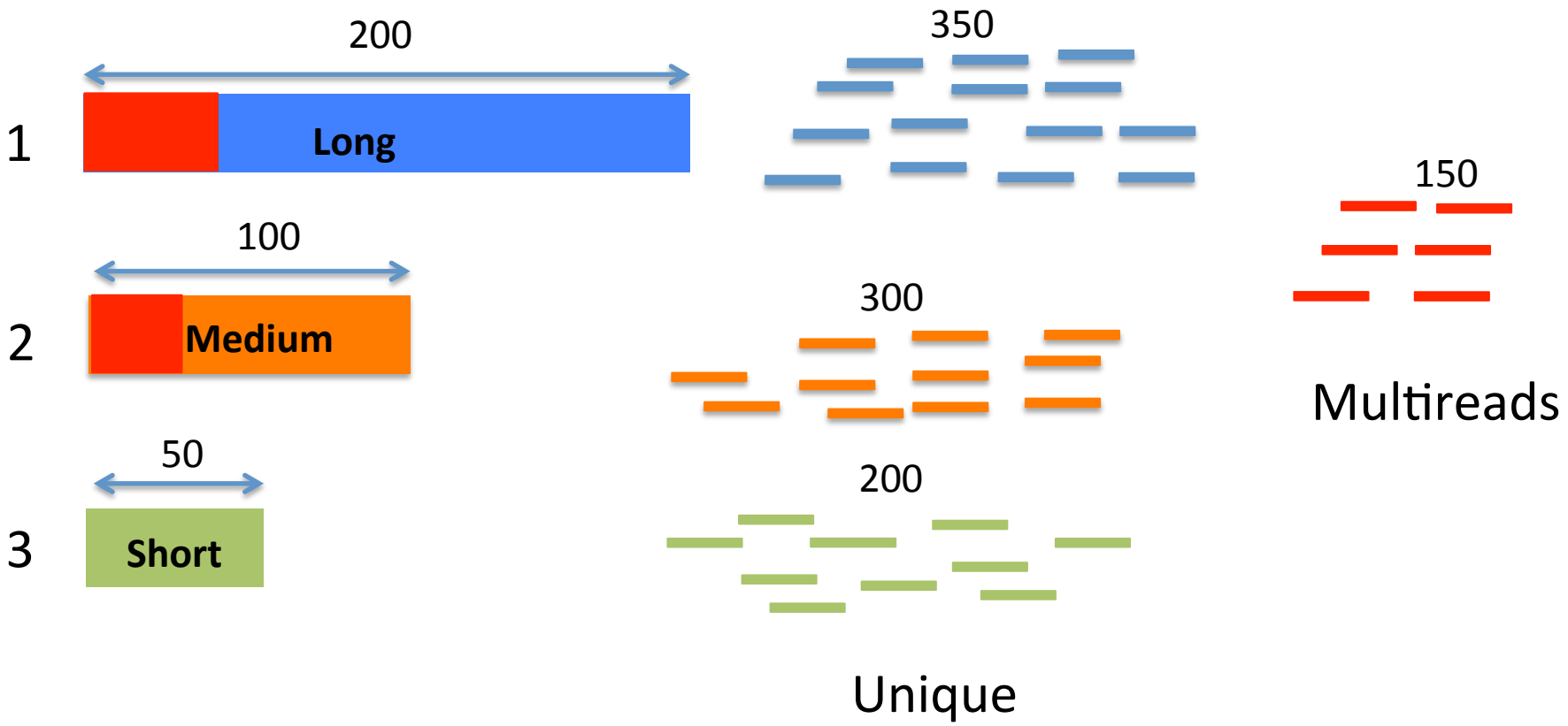
N

Aligned Reads to Gene Abundance: Challenges



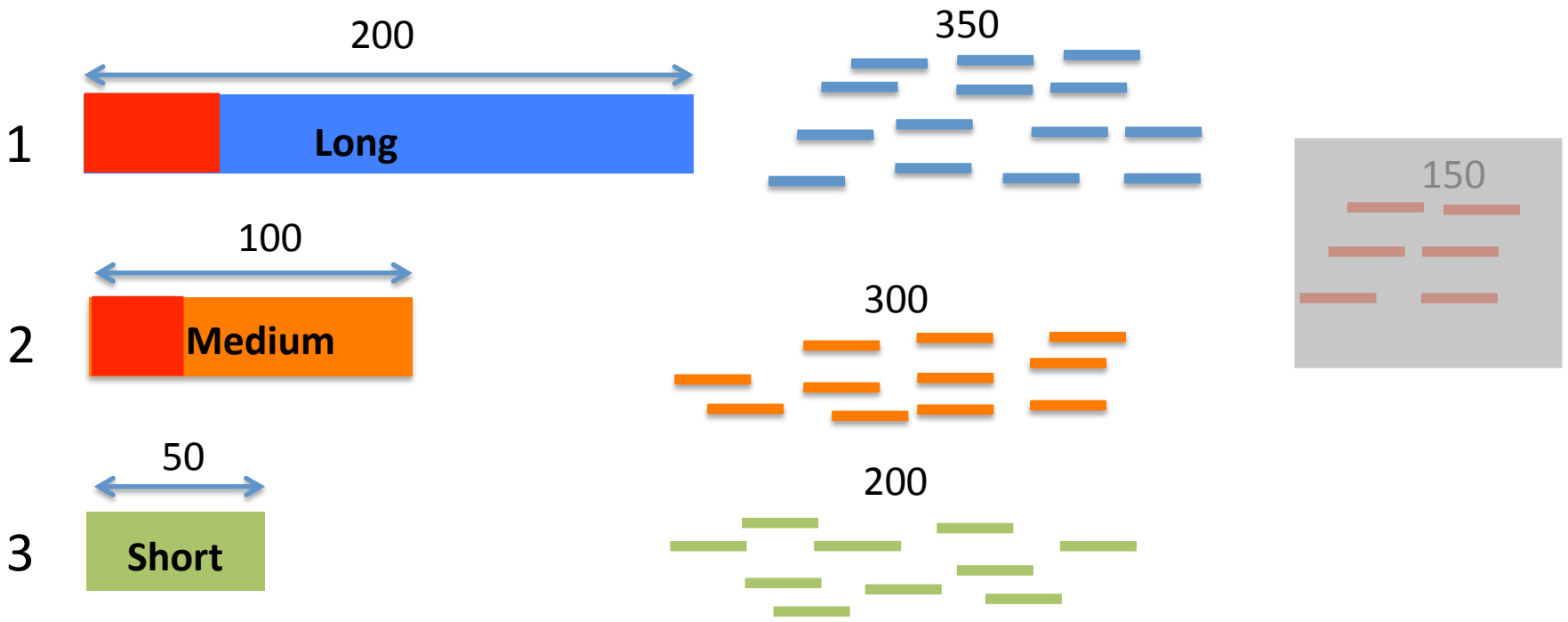
Relative abundance for these genes, f_1, f_2, f_3

Multireads: Reads Mapping to Multiple Genes/Transcripts



Relative abundance for these genes, f_1, f_2, f_3

Approach 1: Ignore Multireads

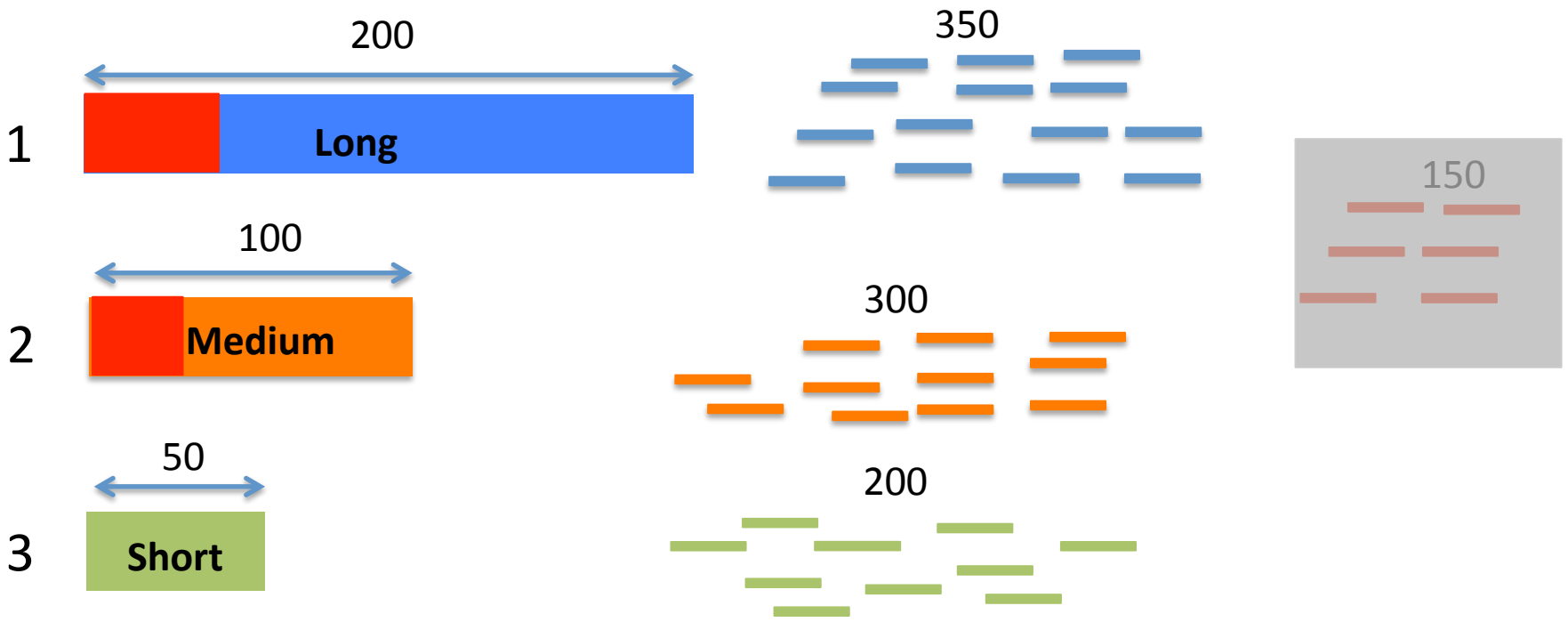


Relative abundance for these genes, f_1, f_2, f_3

Nagalakshmi et. al. Science. 2008
Marioni, et. al. Genome Research 2008

N

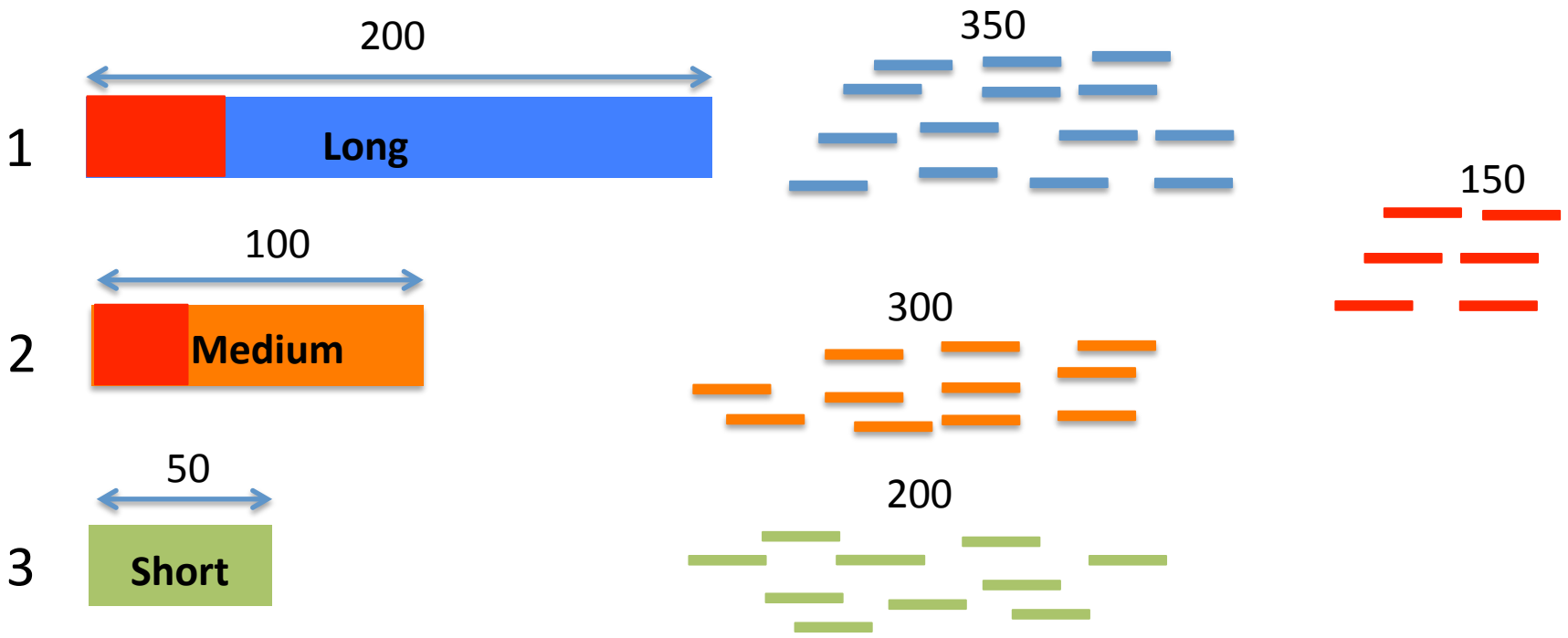
Approach 1: Ignore Multireads



- Over-estimates the abundance of genes with unique reads
- Under-estimates the abundance of genes with multireads
- Not an option at all, if interested in isoform expression

N

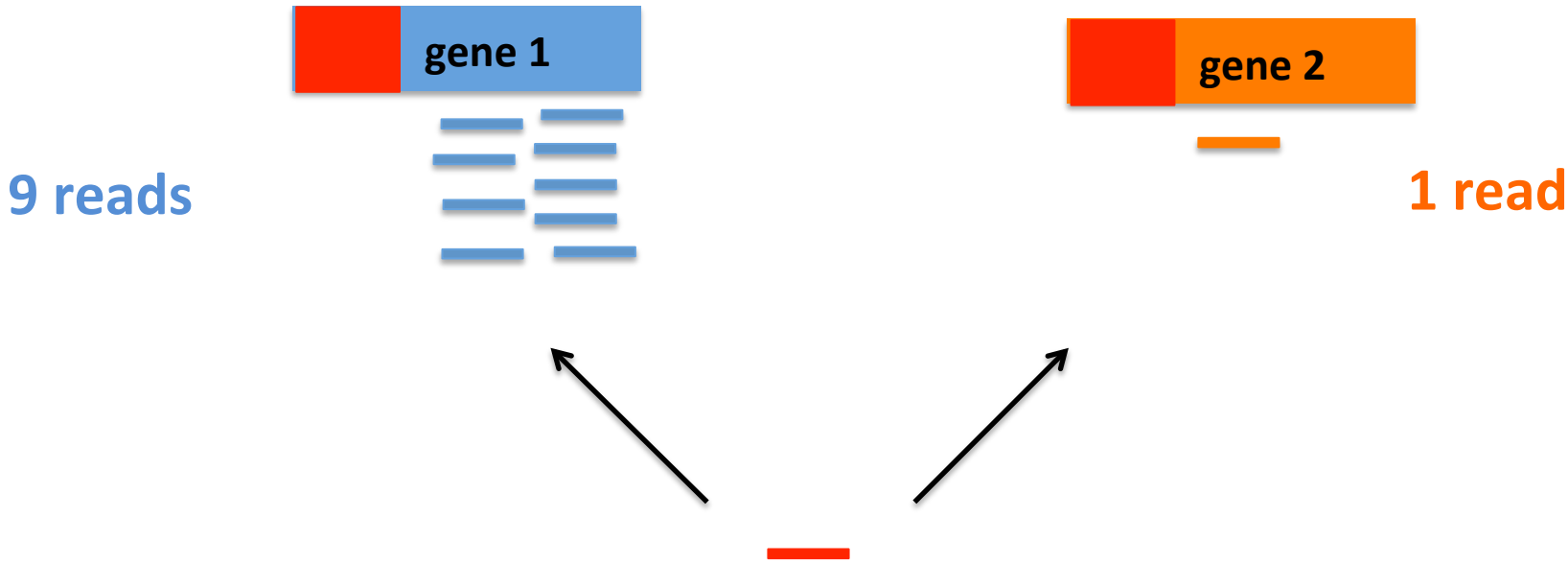
Approach 2: EM algorithm based allocation of Multireads



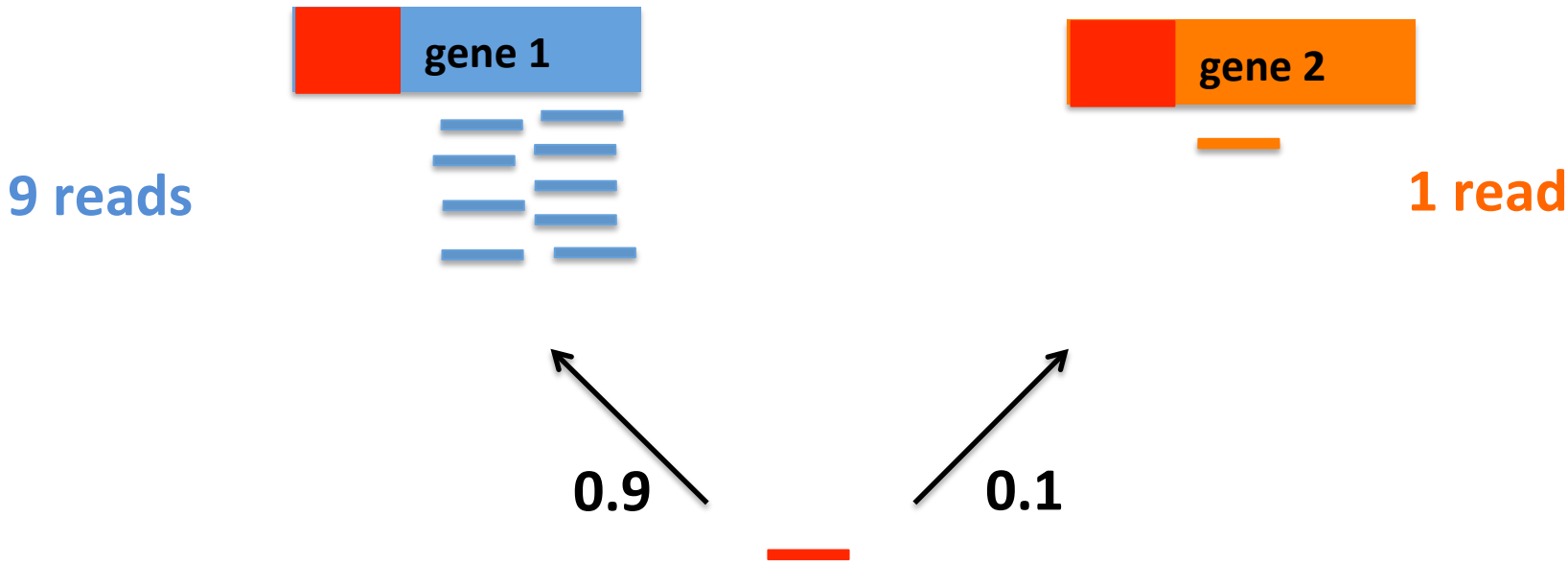
Relative abundance for these genes, f_1, f_2, f_3

RSEM, Cufflinks, isoEM, MMSEQ & eXpress

Approach 2: EM algorithm based allocation of Multireads



Approach 2: EM algorithm based allocation of Multireads



The rise of Pseduo-alignment a.k.a alignment-free methods

100bp Read

ACATGCTGCGGA

K-mers

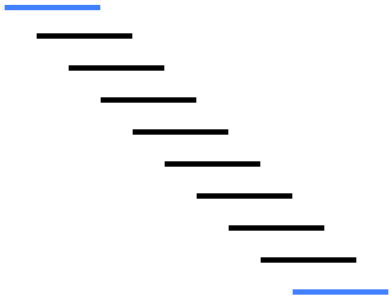
Sailfish, Salmon, and Kallisto

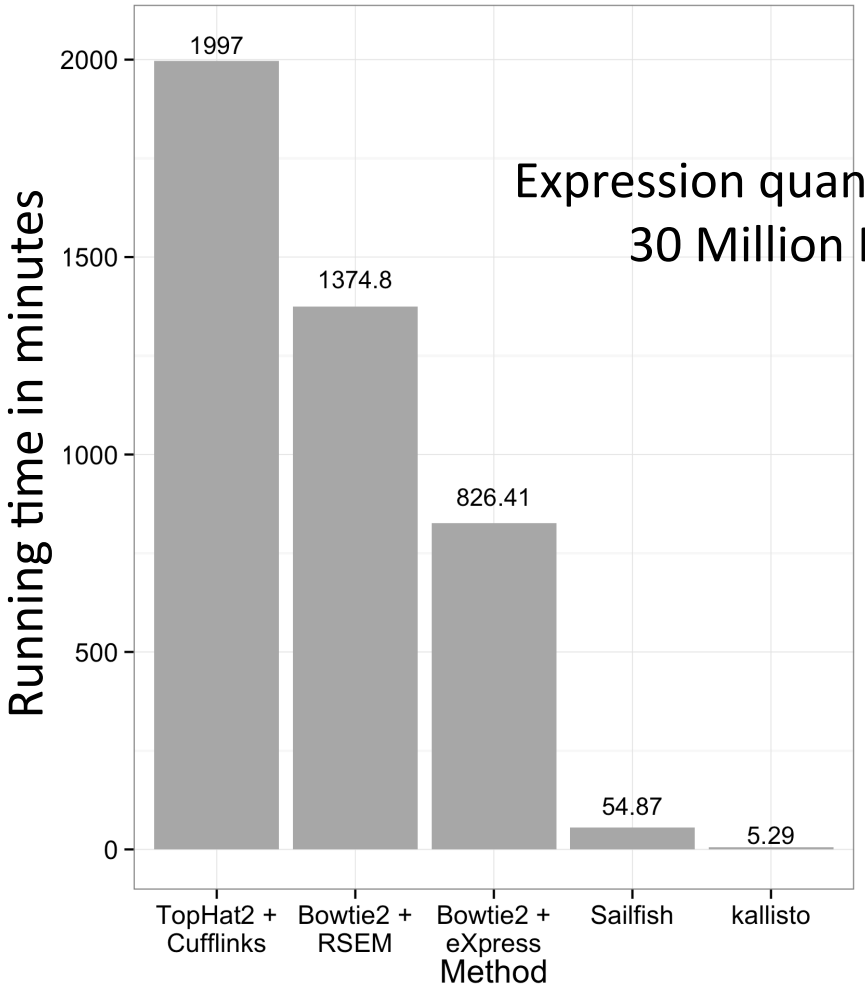
Transcriptome

Kallisto: K-mer based pseudo-alignment

100bp Read

ACATGCTGCGGA





Conclusions for quantitation

- EM approaches are currently the best option.
- Isoform-level estimates are stilll challenging and will become easier as read length increases.
- K-mer counting methods (Salmon, Kallisto) are very fast – they can be run easily on your own PC – and are reasonably accurate.

N

Expression Abundance: Counts, RPKM/FPKM, TPM

Sample 1

Sample 2

$$\textbf{FPKM} = \frac{\text{Number of Fragments Matched to a Gene / Kilo base}}{\text{Total matched reads in Millions}}$$

A speed bump on the road from raw counts to differential expression.

NORMALIZATION

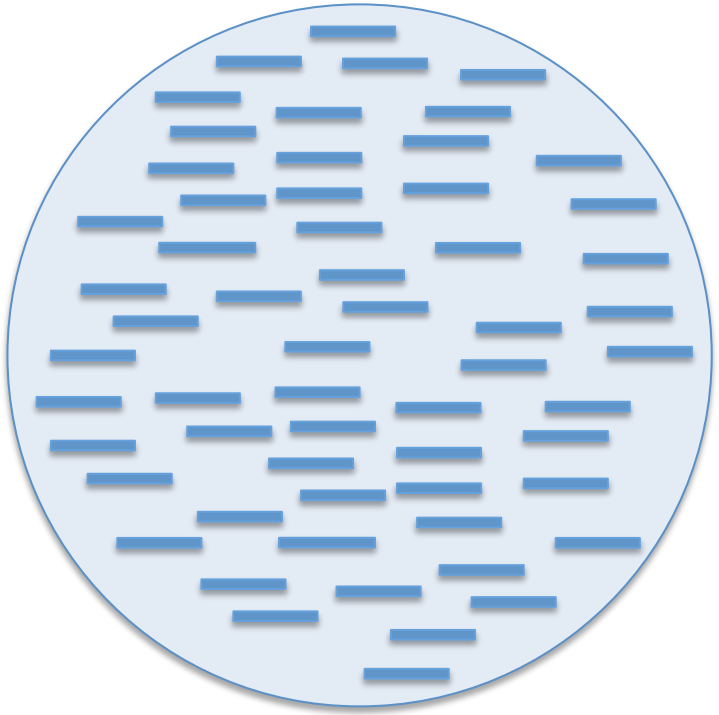
Large pool, small sample problem

- Typical RNA library estimated to contain 2.4 x 10^{12} molecules. McIntyre et al 2011
- Typical sequencing run = 25 million reads/sample.
- This means that only 0.00001 (1/1000th of a percent) of RNA molecules are sampled in a given run.
- High abundance transcripts are sampled more frequently.

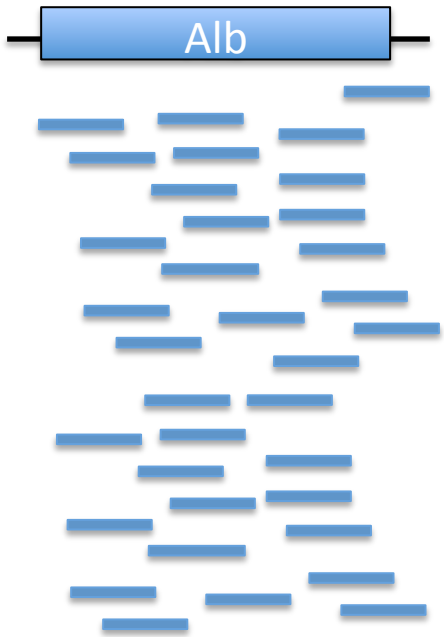
  Example: Albumin = 13% of all reads in liver RNA-seq samples.
- Sampling errors affect low-abundance transcripts most.

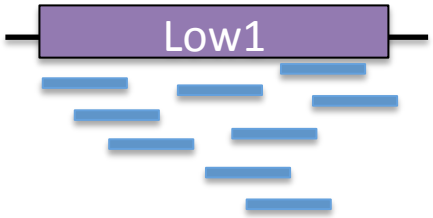
S

A finite pool of reads.



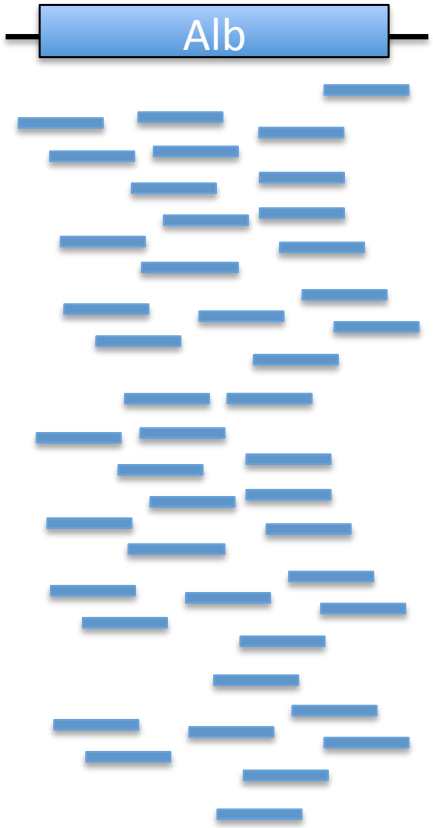
Sample 1

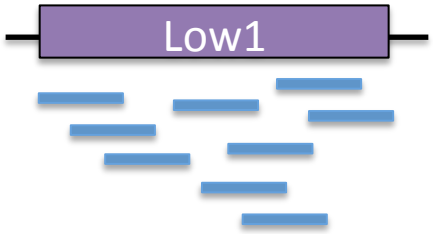




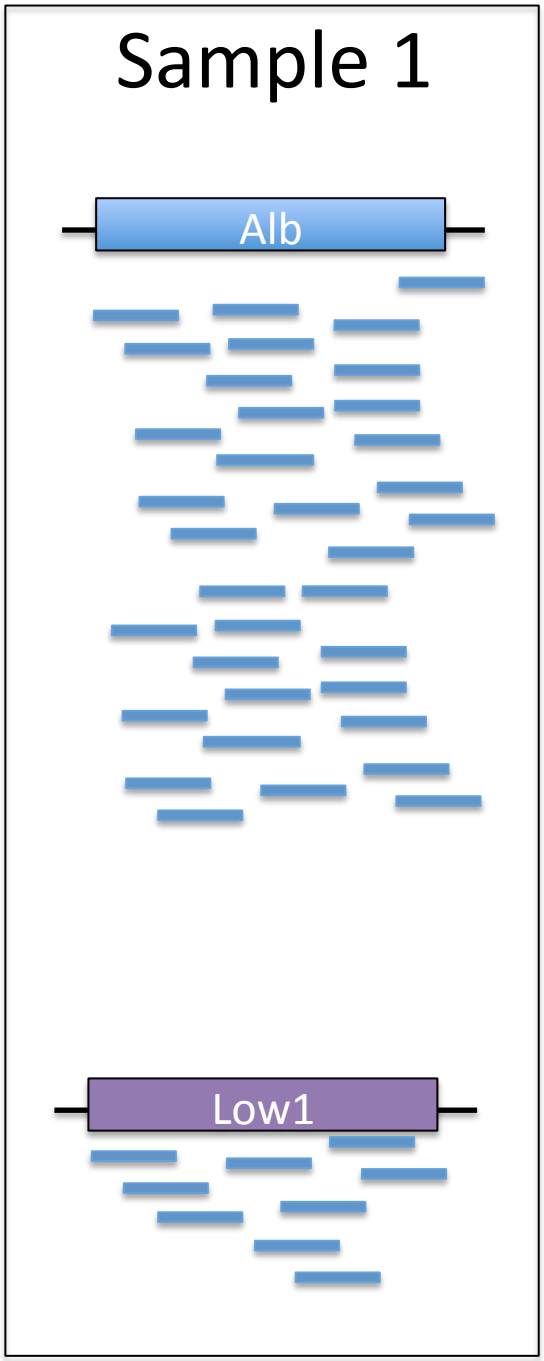
Perfect world:
All transcripts counted.

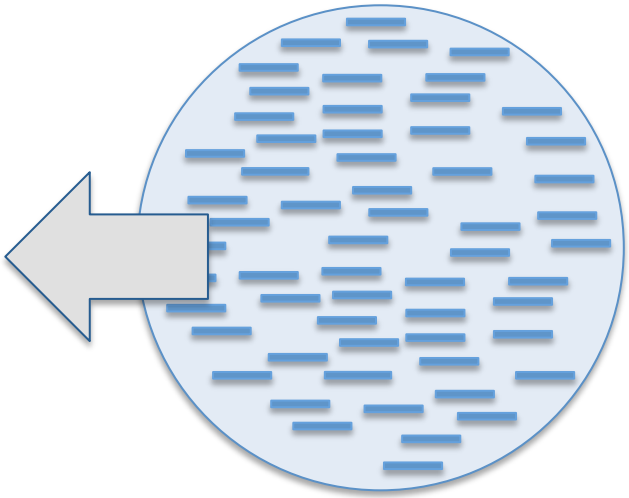
Sample 2



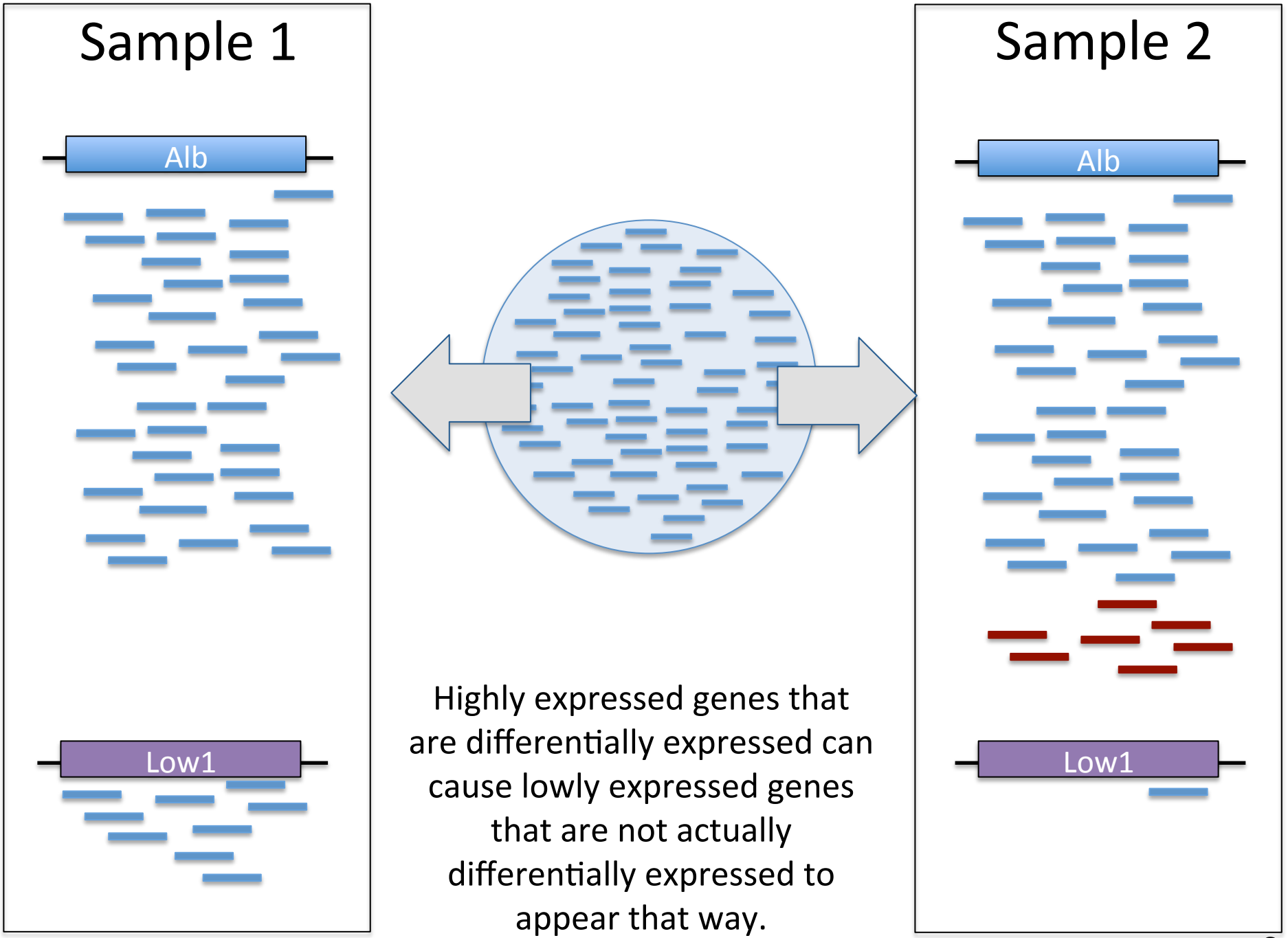


S





Real world: More reads taken up by highly expressed genes means less reads available for lowly expressed genes.

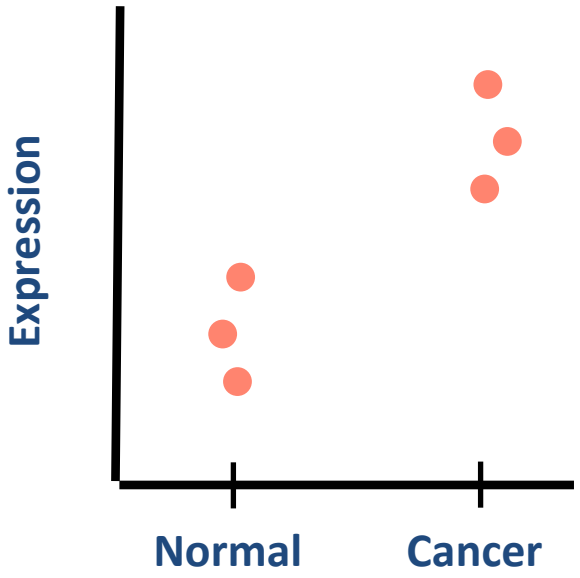

Sample 1
Alb
Low1
Sample 2
Alb
Low1
Highly expressed genes that are differentially expressed can cause lowly expressed genes that are not actually differentially expressed to appear that way.

Normalization of raw counts

- Wrong way to normalize data
  - Normalizing to the total number of mapped reads (e.g. FPKM). Top 10 highly expressed genes soak up 20% of reads in the liver. FPKM is widely used, and problematic.
- Better ways to measure data
  - Normalize to upper quartile (75^{th} %) of non-zero counts, median of scaled counts (DESeq), or the weighted trimmed mean of the log expression ratios (EdgeR).

S

Differential Expression Analysis



T-test

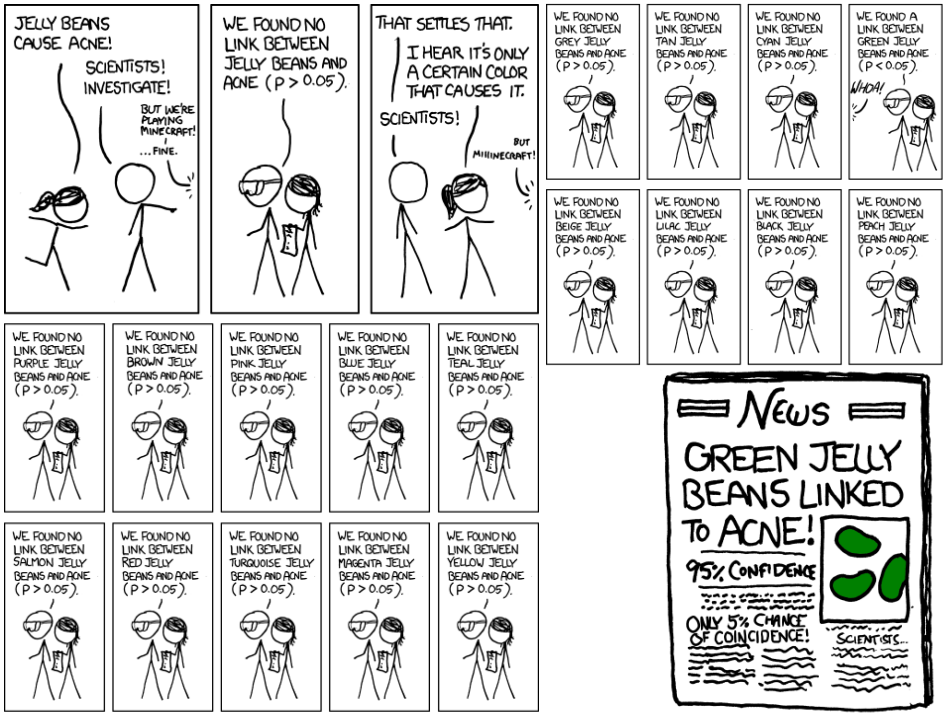
$$t_g = \frac{\hat{\mu}_{g,1} - \hat{\mu}_{g,2}}{\sqrt{\frac{\hat{\sigma}^2_{g,1}}{N_1} + \frac{\hat{\sigma}^2_{g,2}}{N_2}}}$$

Over-estimation of $\hat{\sigma}^2_g$ → Too conservative

Under-estimation of $\hat{\sigma}^2_g$ → Too sensitive
(Many false positives)

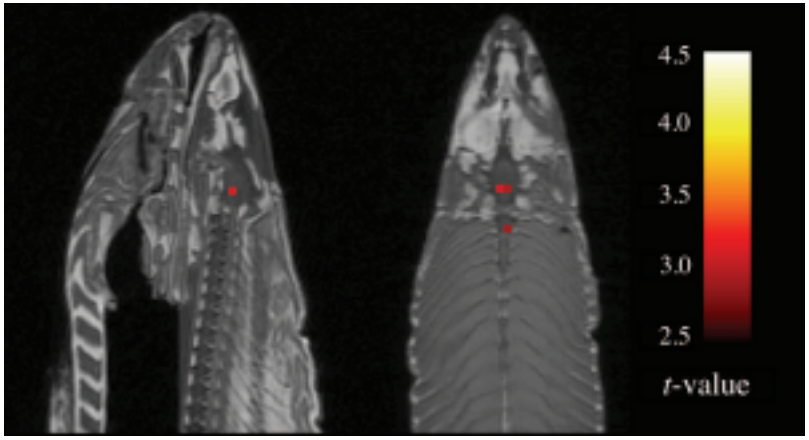
DESEQ2, edgeR, Voom, & CuffDiff

Multiple Testing Correction and False Discovery rate



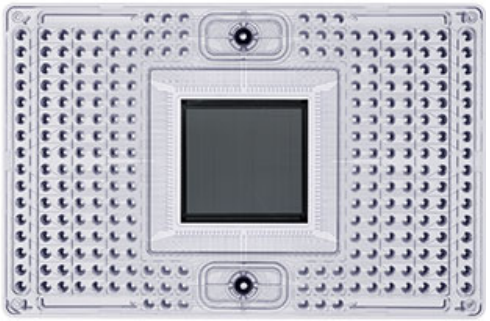
XKCD Significant

2012 IgNobel prize in
Neuroscience for “finding
Brain activity signal in dead salmon using fMRI”



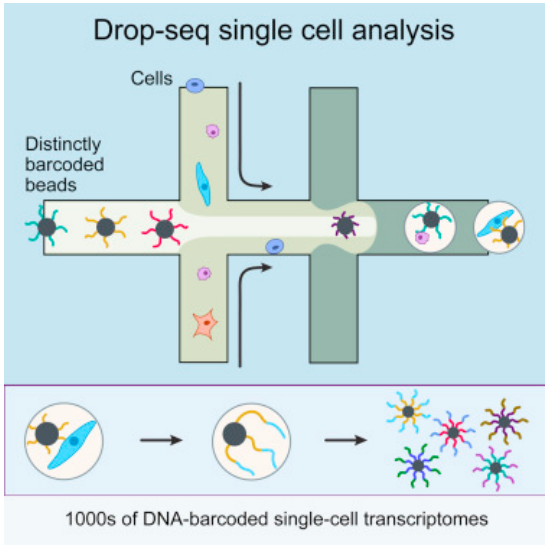
N

Single Cell RNA-seq Technologies



Fluidigm C1 Chip
96 cells / 800 Cells

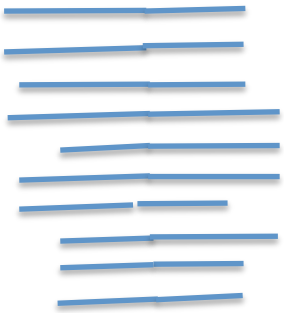
DropSeq: 40,000 cells



10X Genomics:
48,000 cells

Summary

RNA



ATGCTCA AGCTA
TAGATGCTCAAGCTA
ATGCTCAAGCTAATC
ATGCTCAAGCTA
AGTAGATGCTCAAGCTA
ATGCTCAAGCTA
ATGCTCA AGCTA
ATGCTCAAGCTA
TAGATGCTCAAGCTAATC
CTCAAGCTAATCCTAG

Summary

As sequences get longer, alignment and isoform quantitation becomes easier!

Resources

Aligner

- Bowtie 2 http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- GSNAP http://research-pub.gene.com/gmap/

Transcript Discovery/Annotation

- STAR https://github.com/alexdobin/STAR/releases
- Tophat http://tophat.cbcb.umd.edu/

Transcript Abundance

- Kallisto http://pachterlab.github.io/kallisto/
- RSEM http://deweylab.biostat.wisc.edu/rsem/
- EMASE https://github.com/churchill-lab/emase

Differential Expression

- DESeq http://www-huber.embl.de/users/anders/DESeq/
- edgeR http://bioconductor.org/packages/release/bioc/html/edgeR.html
- EBSeq https://www.biostat.wisc.edu/~kendzior/EBSEQ/

Example 1

Differential expression in my mutant mouse compared to wild type. What genes are up- or down-regulated?

Things to consider…

- Differential expression of highly expressed and well annotated genes?
  - Smaller sample depth; more biological replicates
  - No need for paired end reads; shorter reads (50bp) may be sufficient.
  - Better to have 20 million 50bp reads than 10 million 100bp reads.
- Looking for novel genes/splicing/isoforms?
  - More read depth, paired-end reads from longer fragments.

N

Example 2

- How to quantify gene expression in a species that has not been sequenced or annotated?
  - Multistep strategy using multiple sequencing technologies.

Example 3

- How to quantify single cell gene expression in a heterogeneous human tumor?

Any other applications you are interested in?

Steve Munger
steven.munger@jax.org

Narayanan Raghupathy
narayanan.raghupathy@jax.org

Acknowledgements

- KB Choi
- Gary Churchill
- Ron Korstanje/ Karen Svenson/ Elissa Chesler
- Joel Graber
- Doug Hinerfeld
- Anuj Srivastava
- Churchill Lab – Dan Gatti
- Al Simons and Matt Hibbs
- Lisa Somes, Steve Ciciotte, mouse room staff at JAX
- Gene Expression Technologies group at JAX