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

Kwangbom (KB) Choi
Narayanan Raghupathy

The Jackson Laboratory

Short Course on The Genetics of Addiction 2015

Next Generation Genome Sequencers

Illumina NextSeq, HiSeq and MiSeq

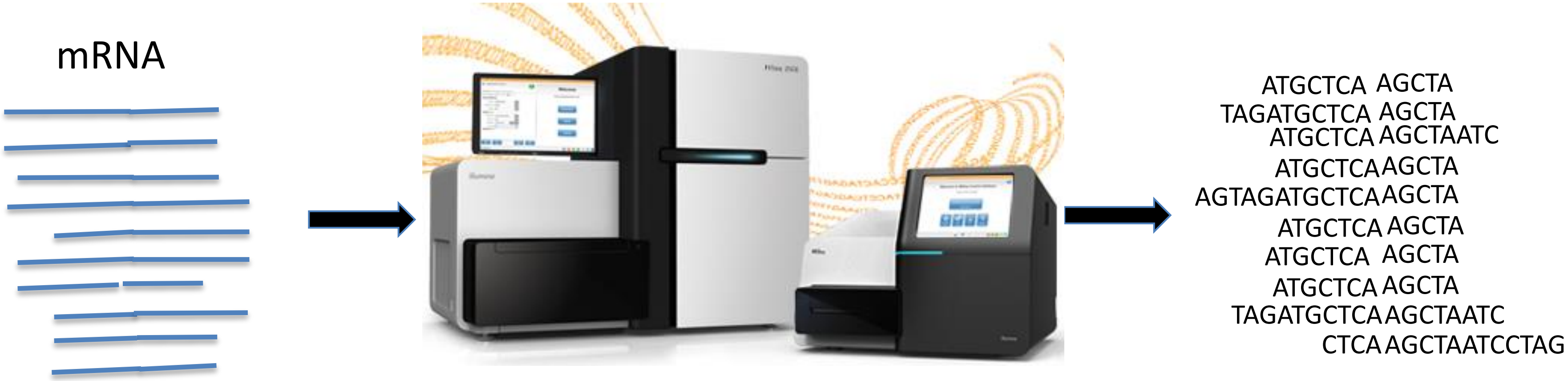
454 GS FLX

N

Ion Torrent Proton

Oxford Nanopore

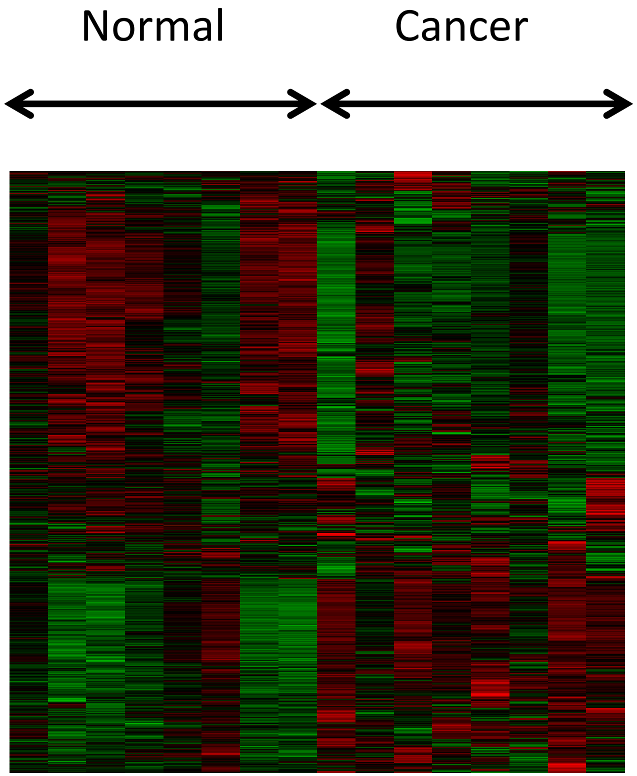
RNA-seq: Sequencing Transcriptomes



mRNA

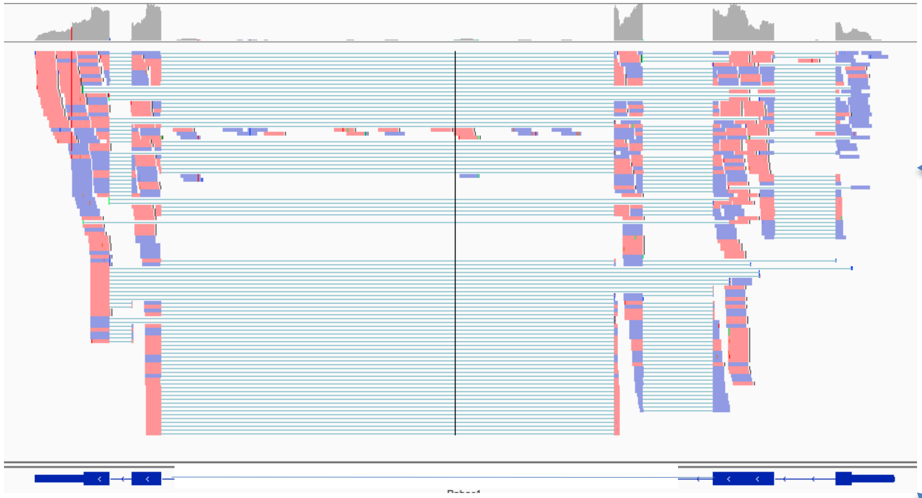
ATGCTCA AGCTA
TAGATGCTCA AGCTA
ATGCTCA AGCTAATC
ATGCTCA AGCTA
AGTAGATGCTCA AGCTA
ATGCTCA AGCTA
ATGCTCA AGCTA
ATGCTCA AGCTA
TAGATGCTCA AGCTAATC
CTCA AGCTAATCCTAG

Applications of RNA-seq Technology



Differential Gene expression analysis

Applications of RNA-seq Technology

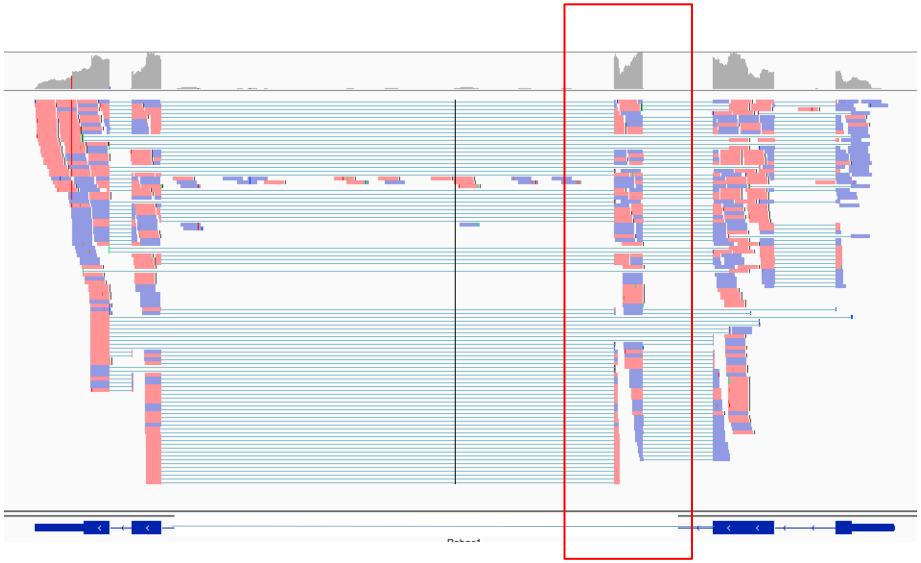


Evidence from RNA-seq

Annotated gene

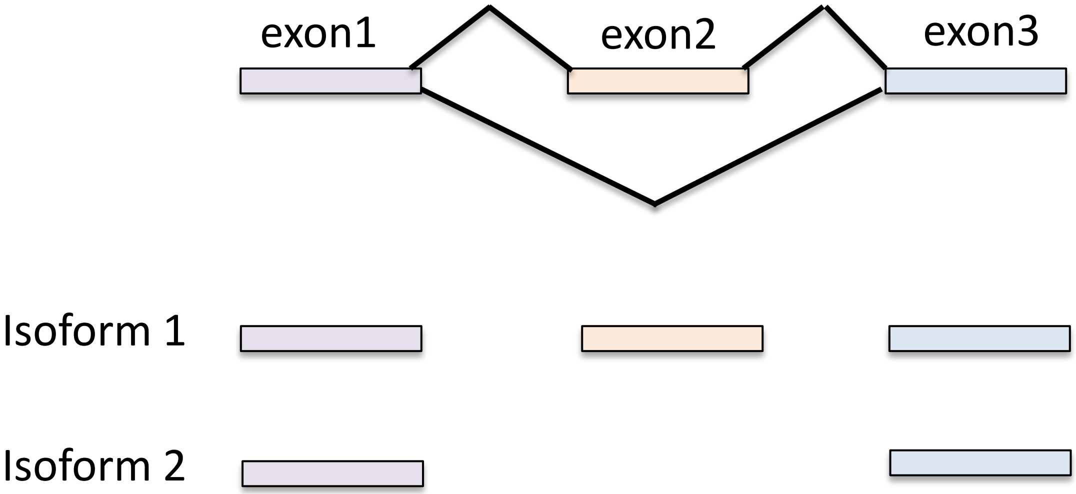
Novel exon discovery

Applications of RNA-seq Technology



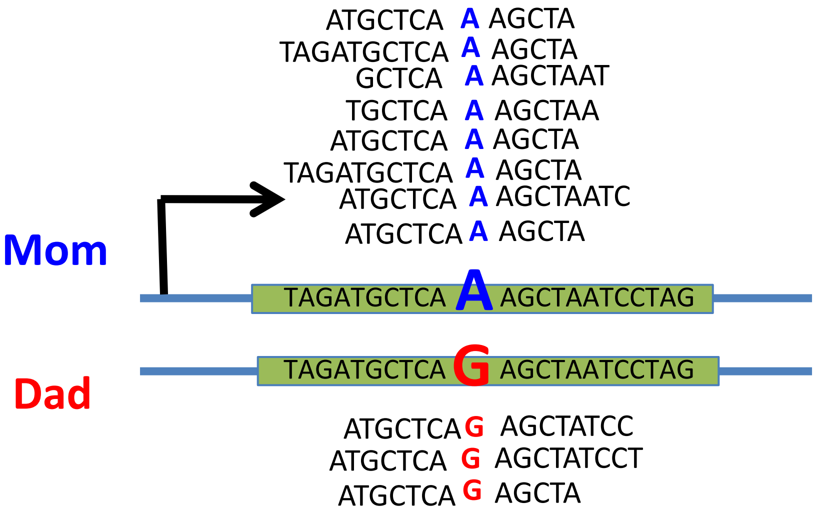
Novel exon discovery

Applications of RNA-seq Technology



Alternative splicing

Applications of RNA-seq Technology



ATGCTCA A AGCTA
TAGATGCTCA A AGCTA
GCTCA A AGCTAAT
TGCTCA A AGCTAA
ATGCTCA A AGCTA
TAGATGCTCA A AGCTA
ATGCTCA A AGCTAATC
ATGCTCA A AGCTA

Mom

TAGATGCTCA A AGCTAATCCTAG

TAGATGCTCA G AGCTAATCCTAG

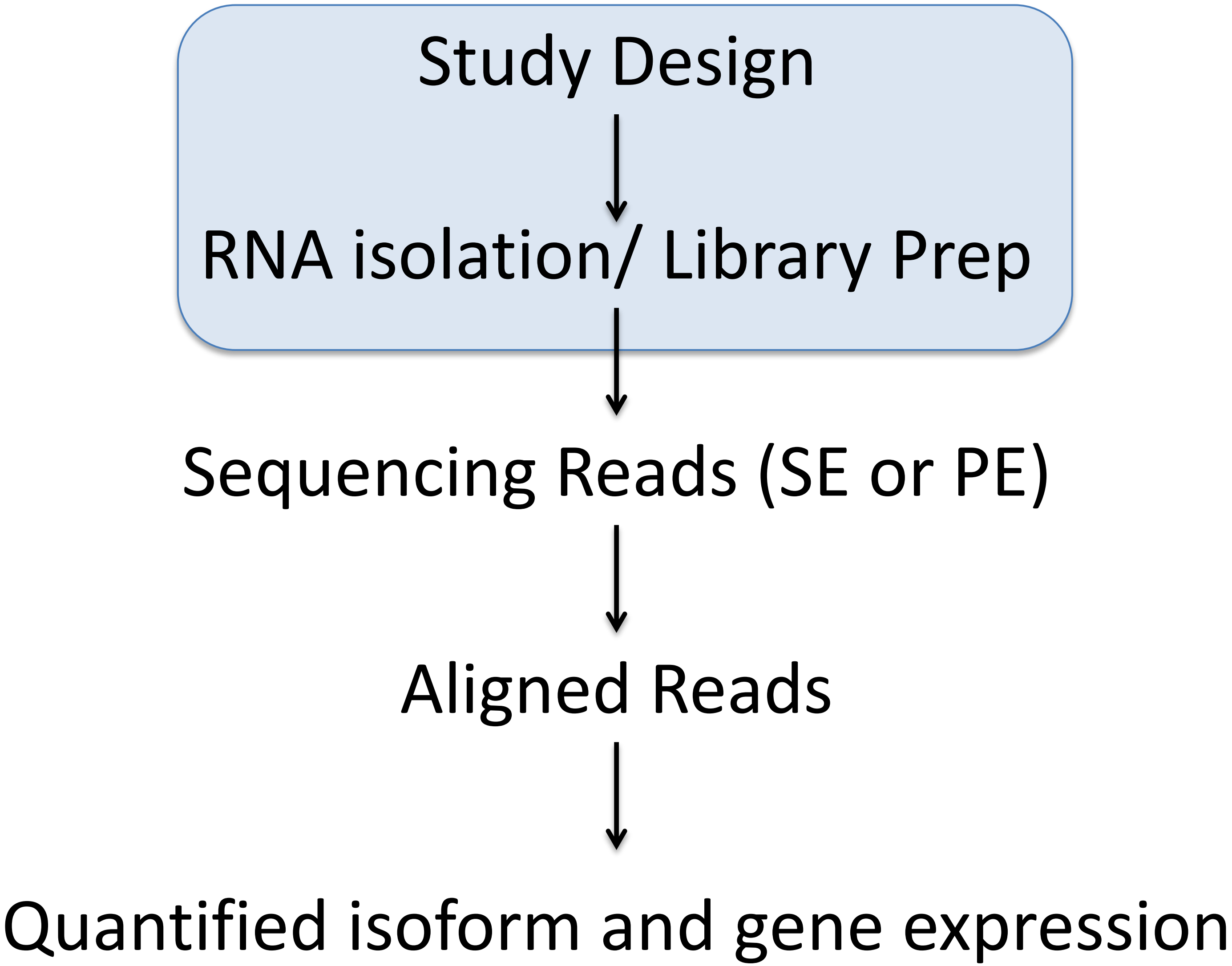
Dad

ATGCTCA G AGCTATCC
ATGCTCA G AGCTATCCT
ATGCTCA G AGCTA

Allele-Specific gene Expression (ASE)

Preferential expression of one allele over the other.

RNA-seq Work Flow



Study Design

RNA isolation/ Library Prep

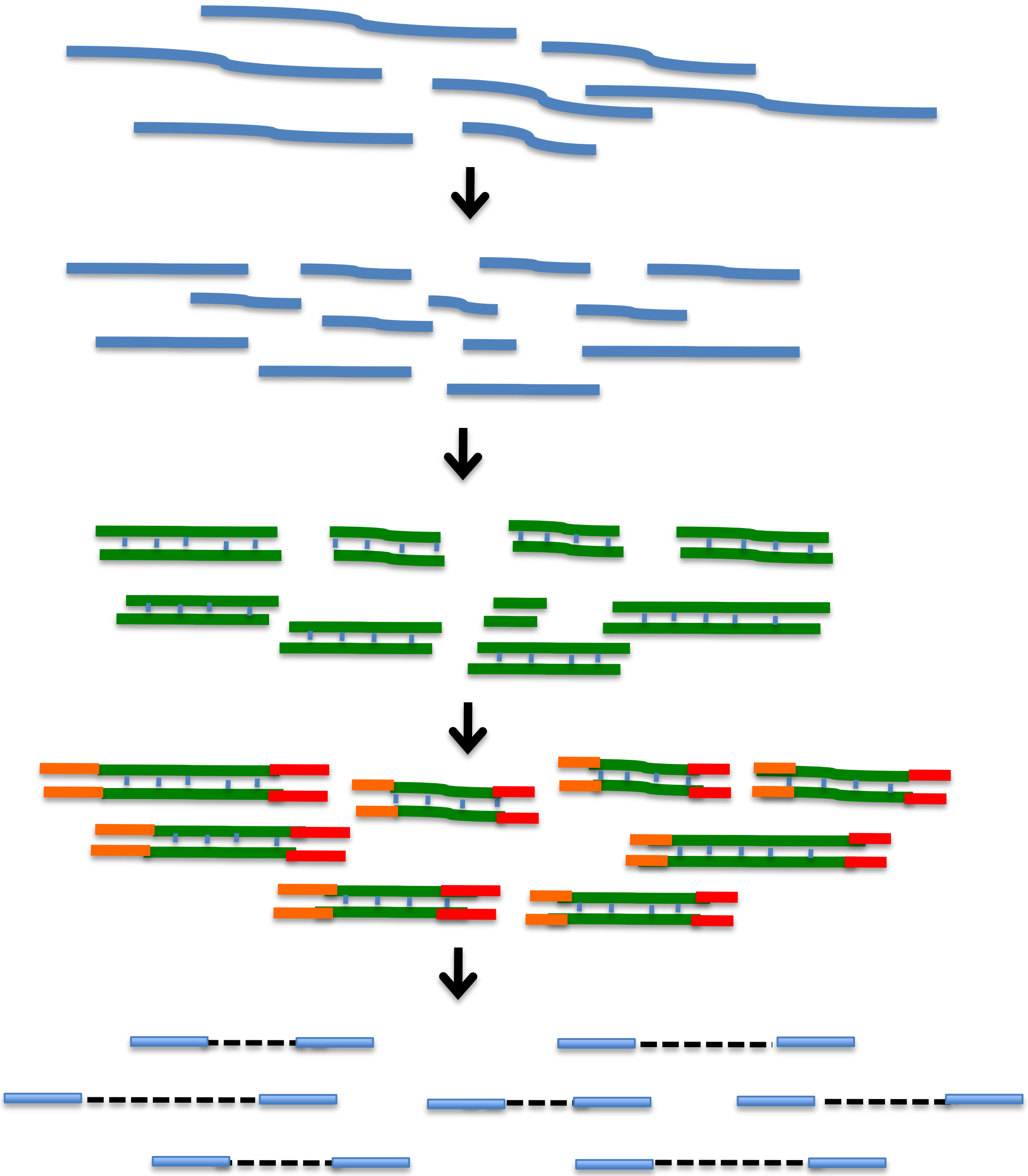
Sequencing Reads (SE or PE)

Aligned Reads

Quantified isoform and gene expression

N

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

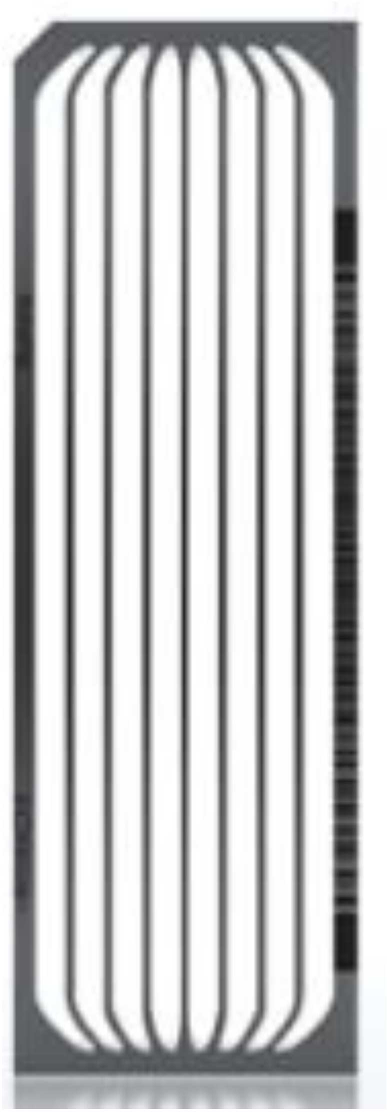
RNA-seq Experimental design

- 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

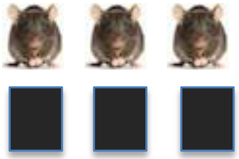
Good Experimental Design

Multiplexing
Replication
Randomization

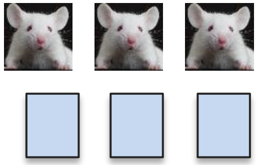


N

RNA-Seq Experimental Design: Randomization

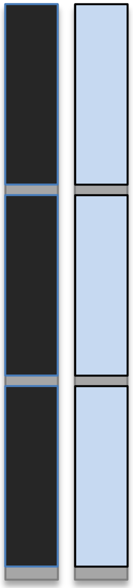


Experimental Group 1



Experimental Group 2

Two Illumina Lanes

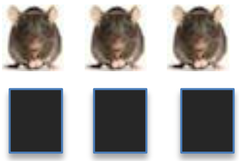


Bad Design

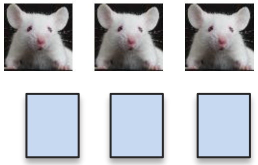
Random.org

N

RNA-Seq Experimental Design: Randomization

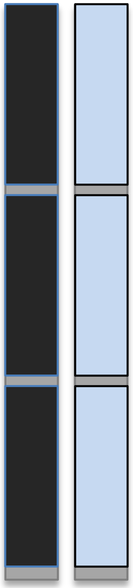


Experimental Group 1

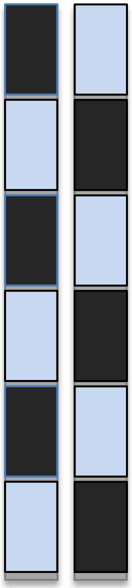


Experimental Group 2

Two Illumina Lanes



Bad Design



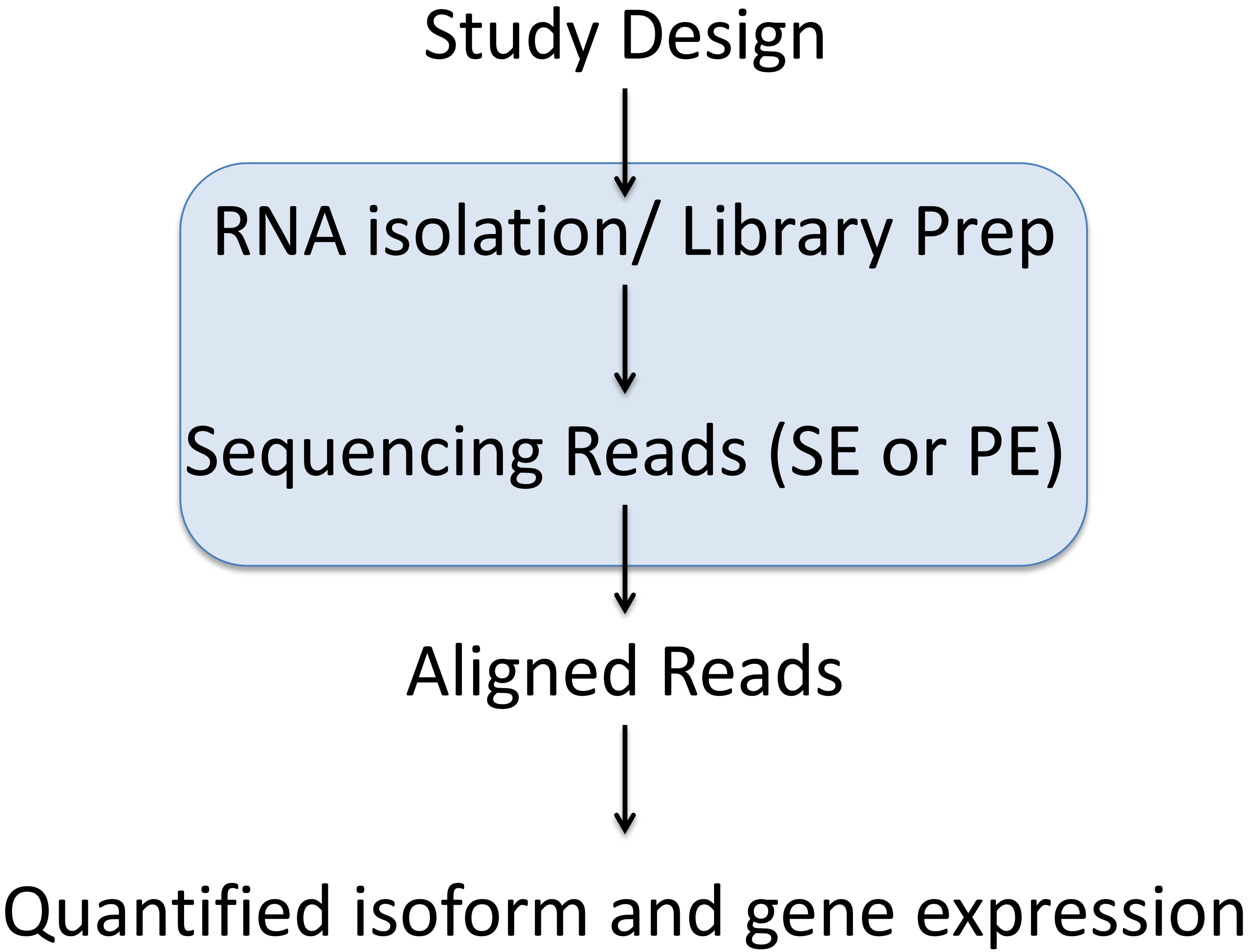
Better Design

Random.org

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

N

RNA-seq Work Flow



Study Design

RNA isolation/ Library Prep

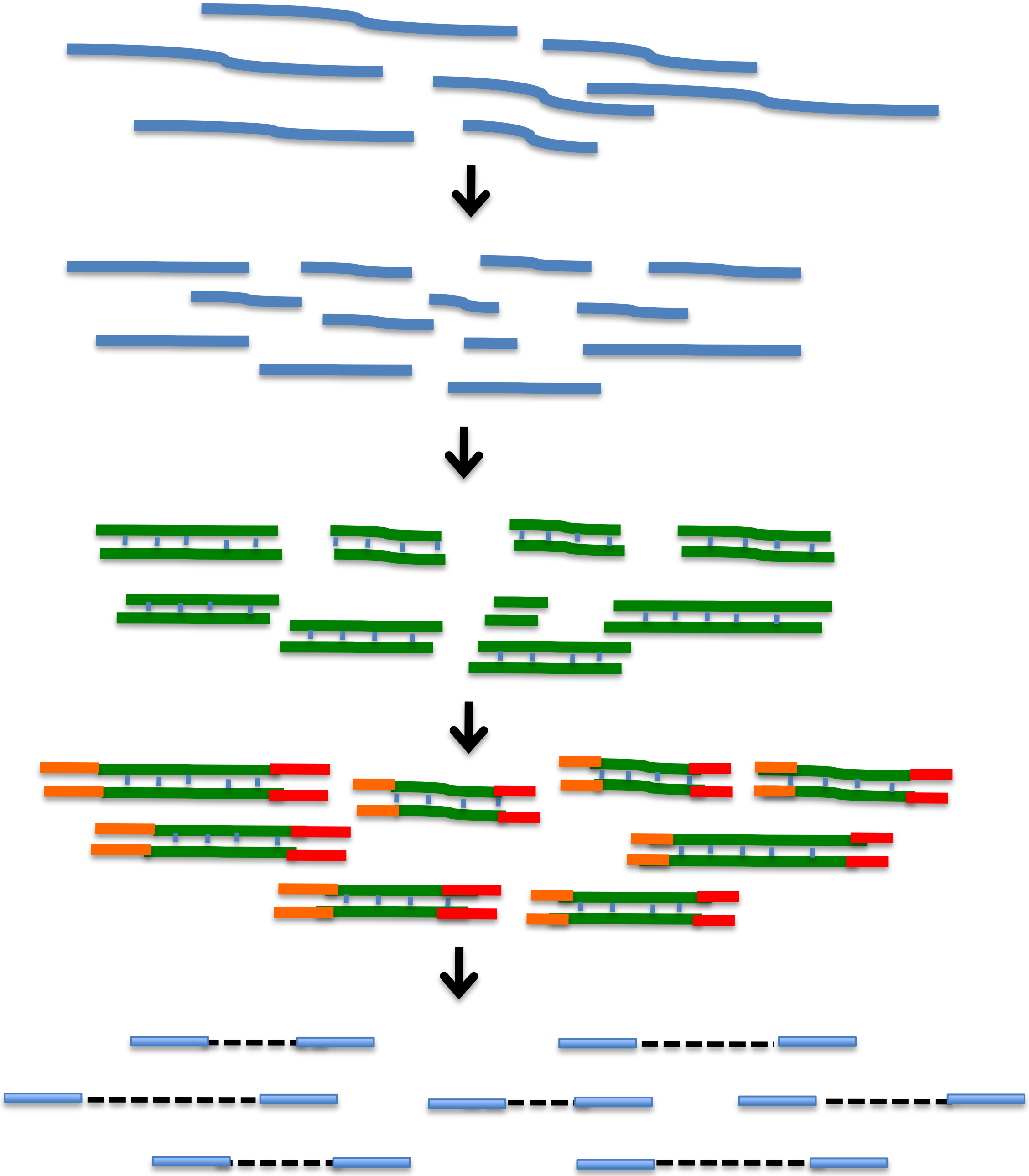
Sequencing Reads (SE or PE)

Aligned Reads

Quantified isoform and gene expression

N

RNA-Seq



Total RNA

mRNA

mRNA after fragmentation

cDNA

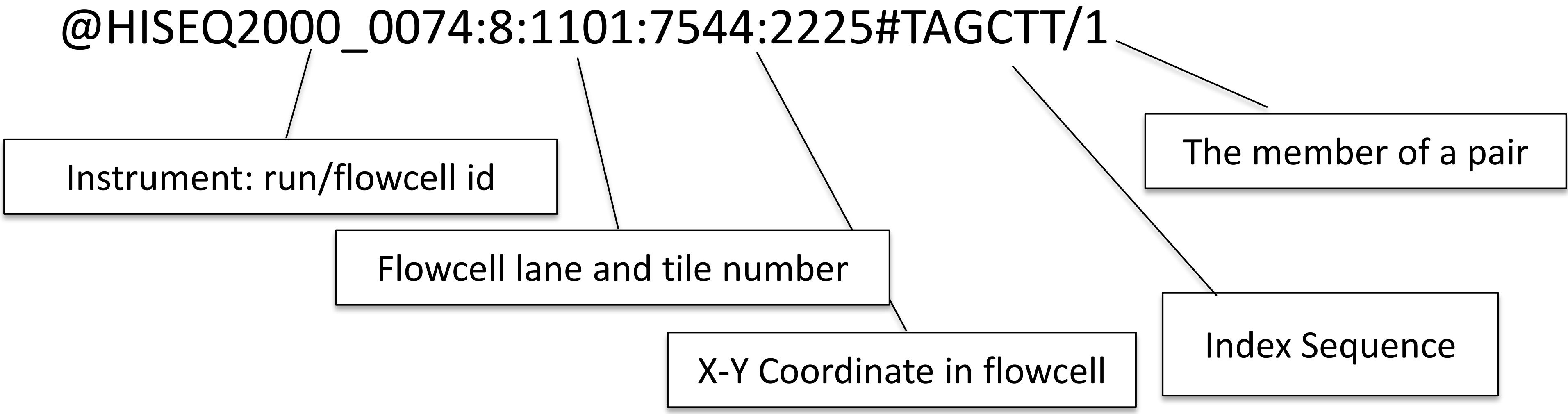
Adaptors ligated to cDNA

Single/ Paired End Sequencing

N

Millions and millions of reads…

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



@HISEQ2000_0074:8:1101:7544:2225#TAGCTT/1

Instrument: run/flowcell id

Flowcell lane and tile number

X-Y Coordinate in flowcell

Index Sequence

The member of a pair

Phred Score:

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

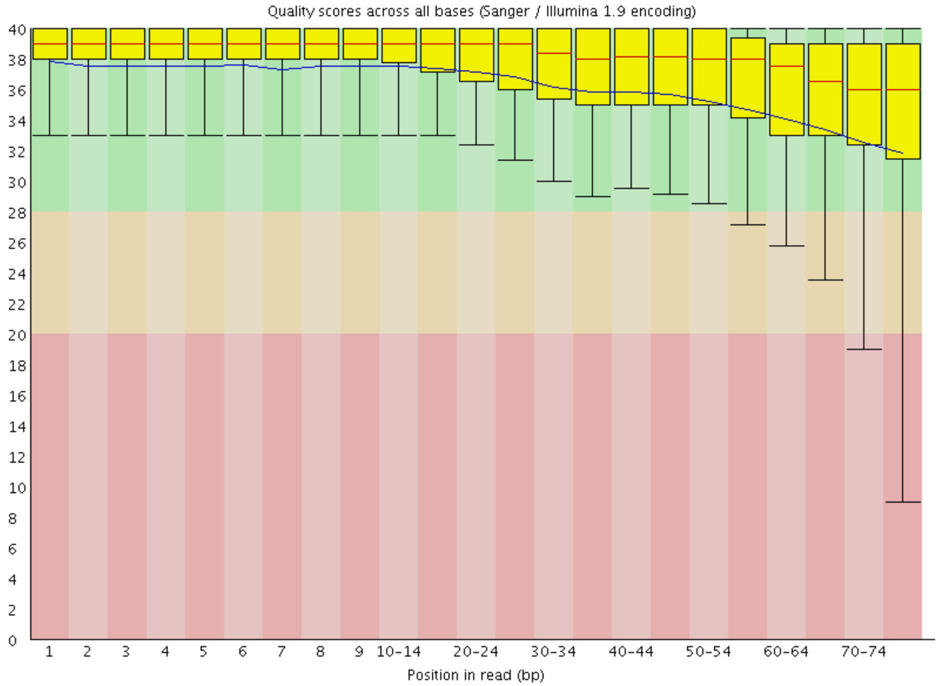
10 indicates 1 in 10 chance of error

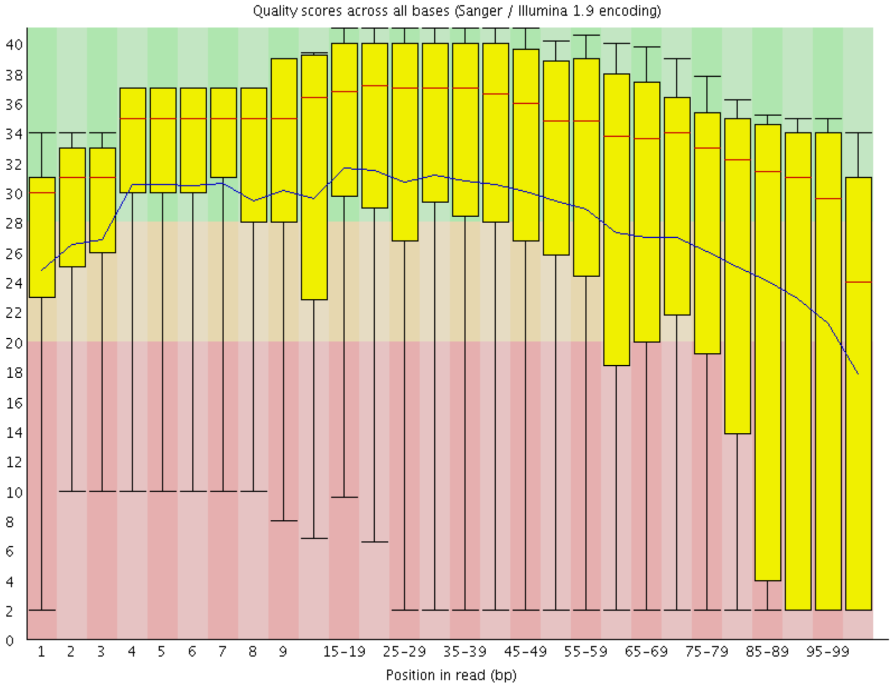
20 indicates 1 in 100,

30 indicates 1 in 1000,

SN

Quality Control: How to tell if your data is clean





Good data

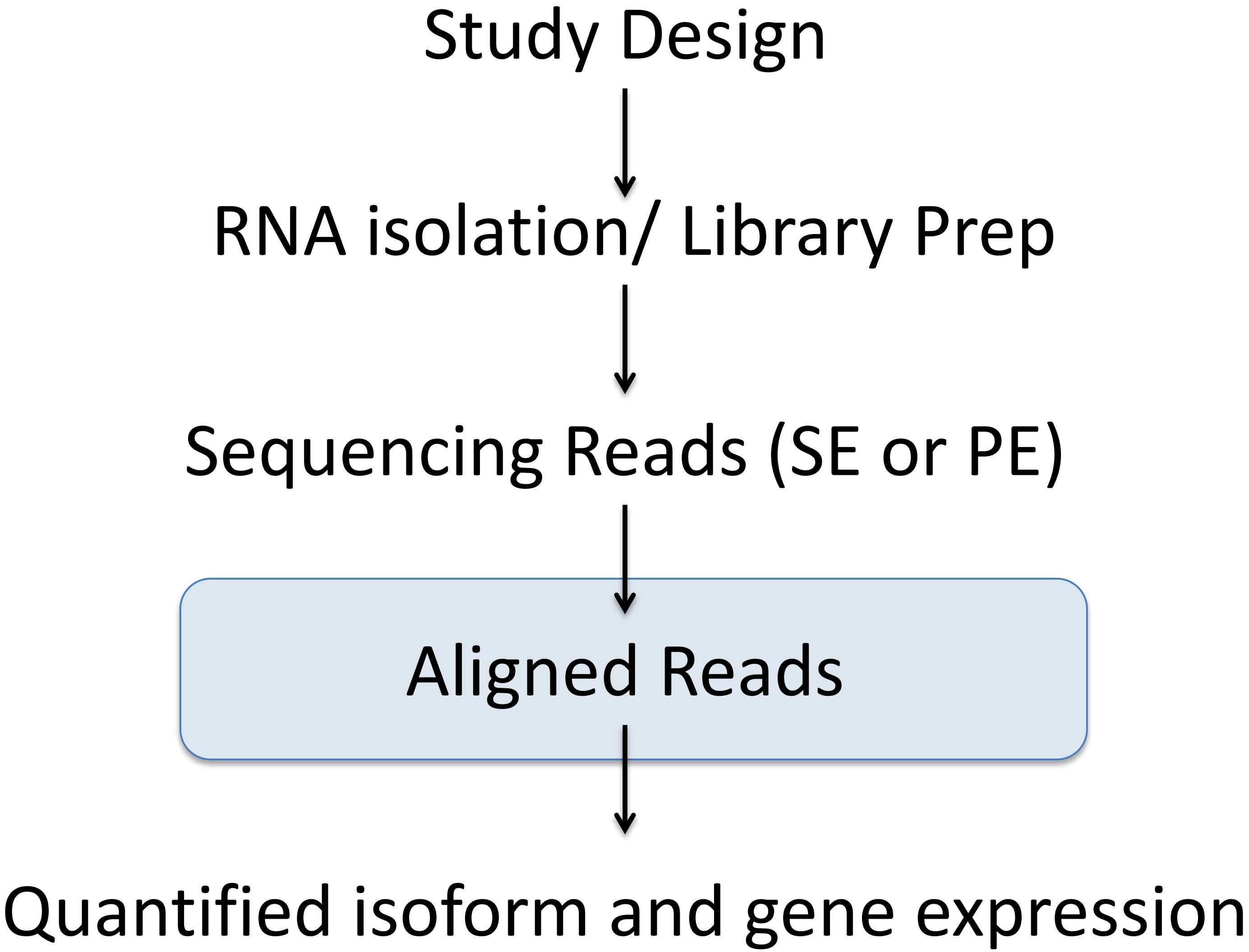
- **Consistent**
- **High Quality Along the reads**

Bad data

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

S

RNA-seq Work Flow



Study Design

RNA isolation/ Library Prep

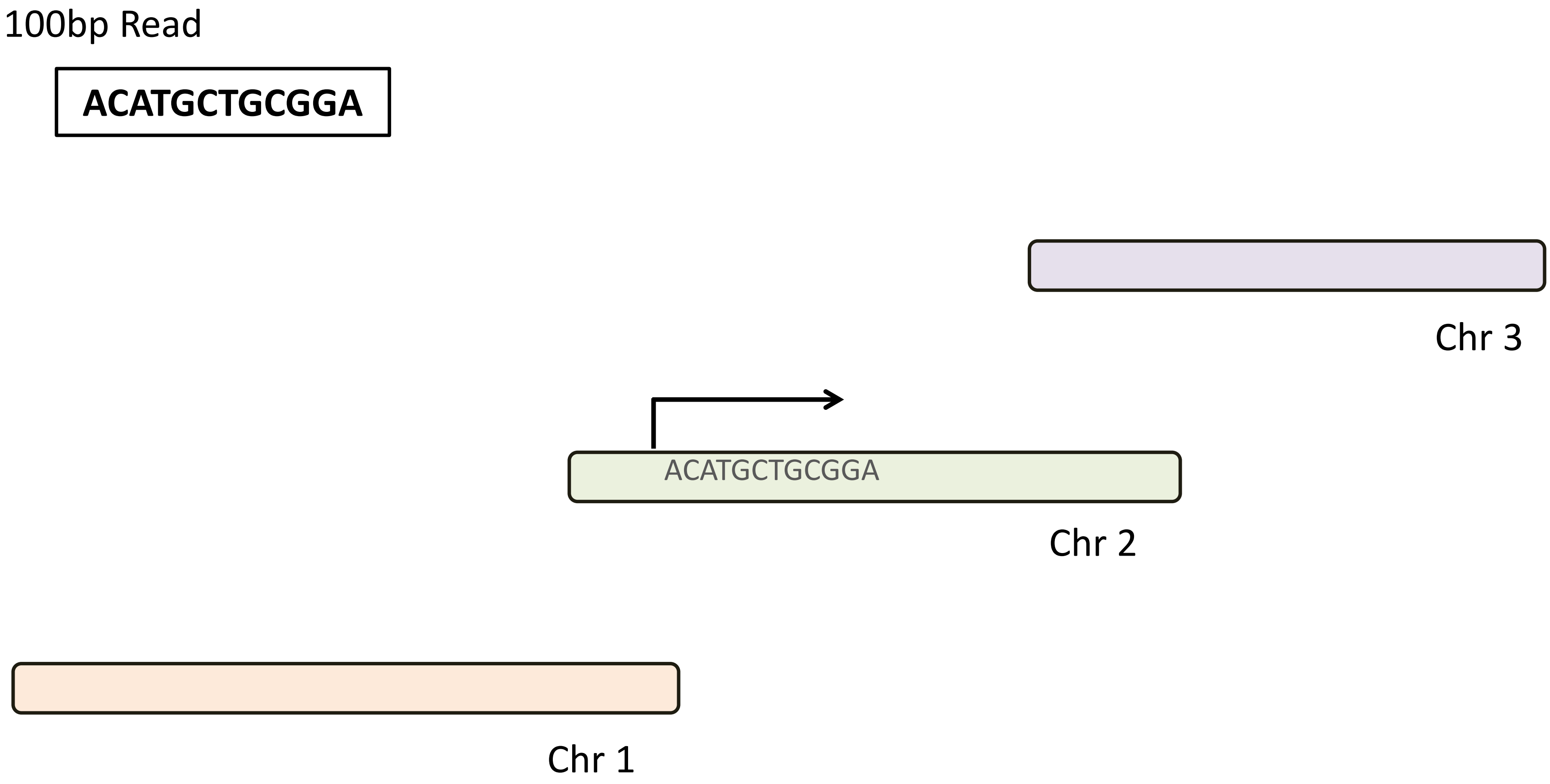
Sequencing Reads (SE or PE)

Aligned Reads

Quantified isoform and gene expression

S

Alignment 101



100bp Read

ACATGCTGCGGA

Chr 3

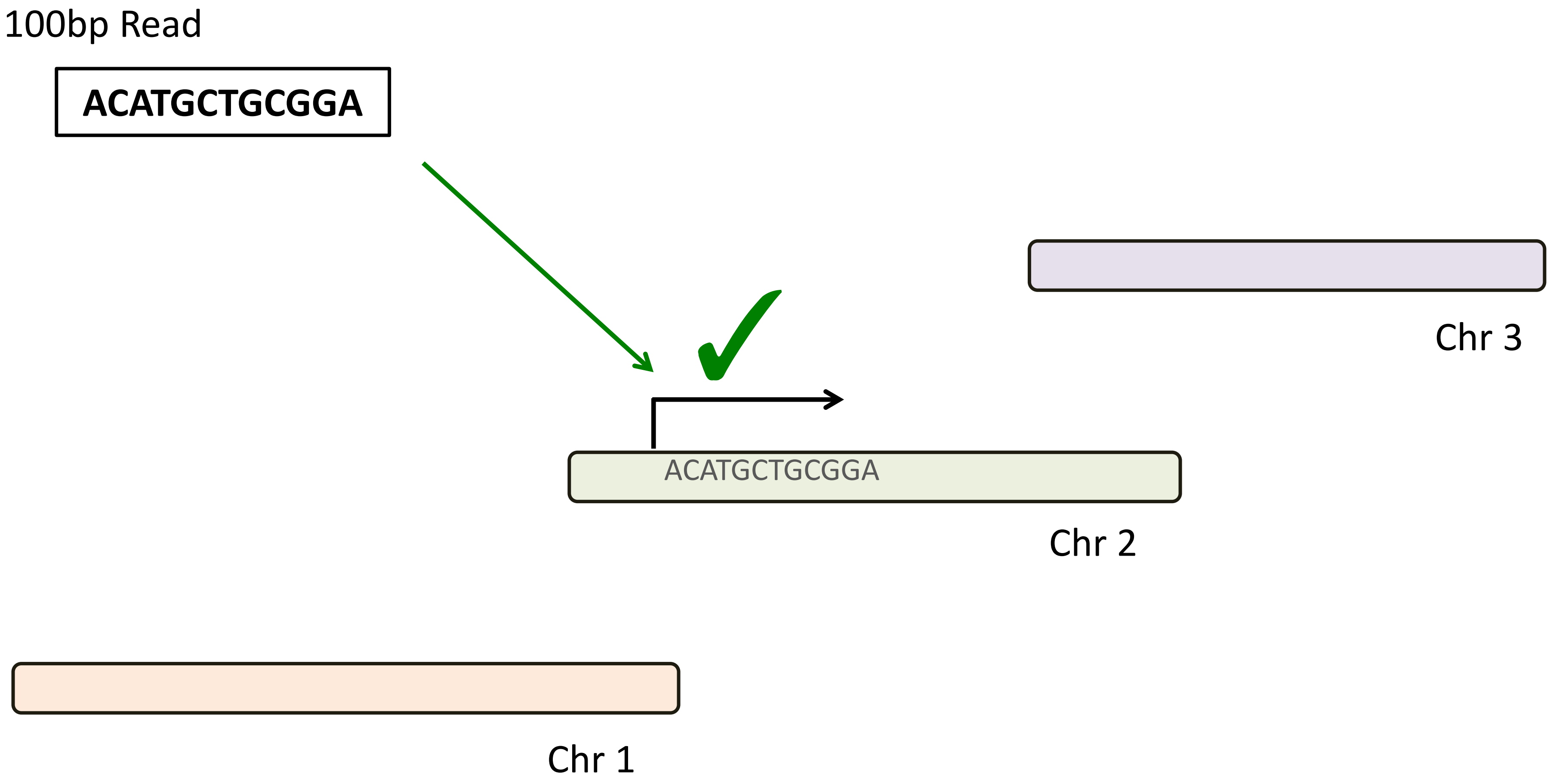
ACATGCTGCGGA

Chr 2

Chr 1

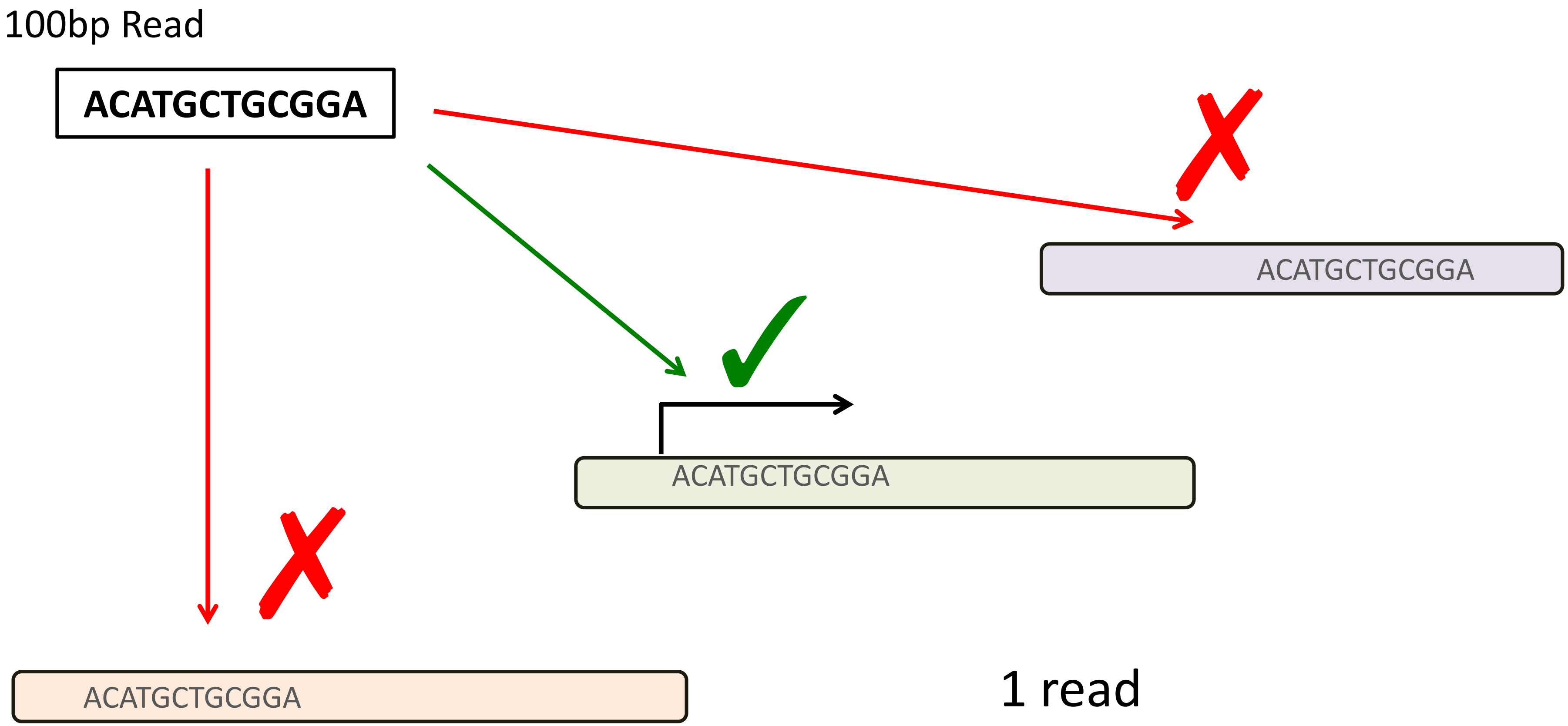
S

The perfect read: 1 read = 1 unique alignment.



S

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



100bp Read

ACATGCTGCGGA

ACATGCTGCGGA

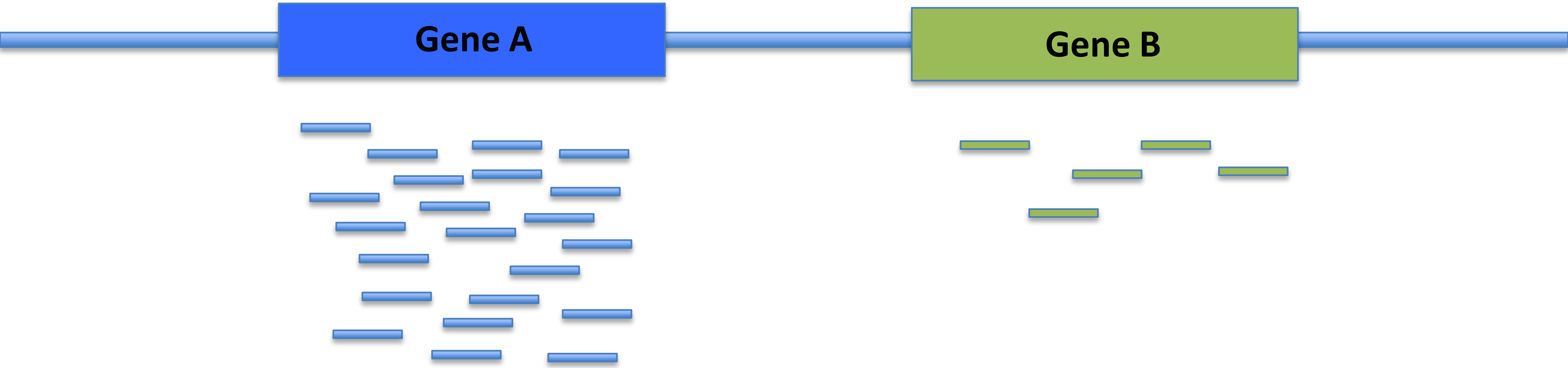
ACATGCTGCGGA

ACATGCTGCGGA

1 read
3 valid alignments
Only 1 alignment is correct

S

Aligning Millions of Short Sequence Reads



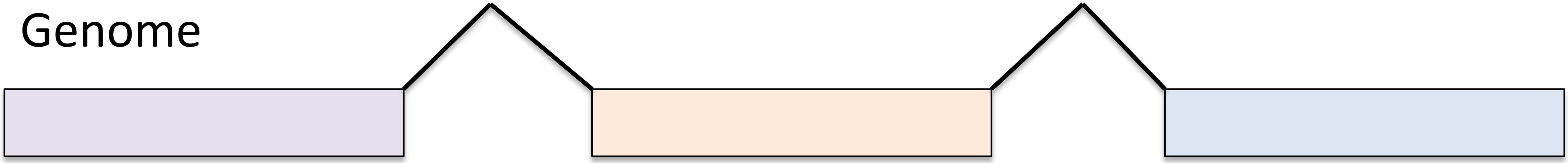
Aligners: Bowtie, GSNAP, STAR, BWA, BLAT, HISAT, Bowtie2, Bowtie10000

Designed to align the short reads fast, but not accurate

N

Align to Genome or Transcriptome?

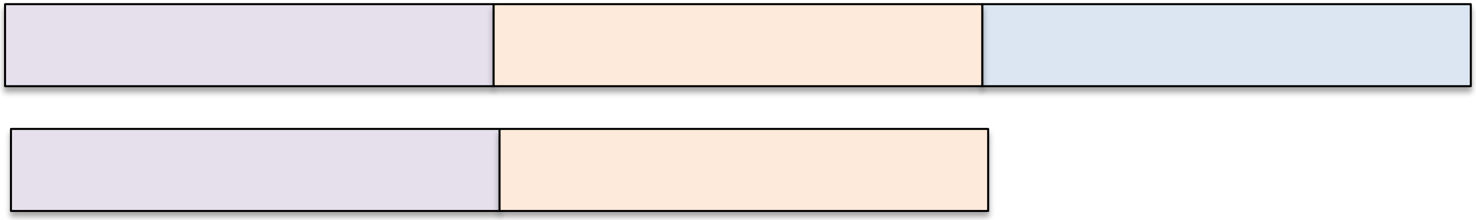
Genome



Advantages: Can align novel isoforms.

Disadvantages: Difficult, Spurious alignments, spliced alignment, gene families, pseudo genes

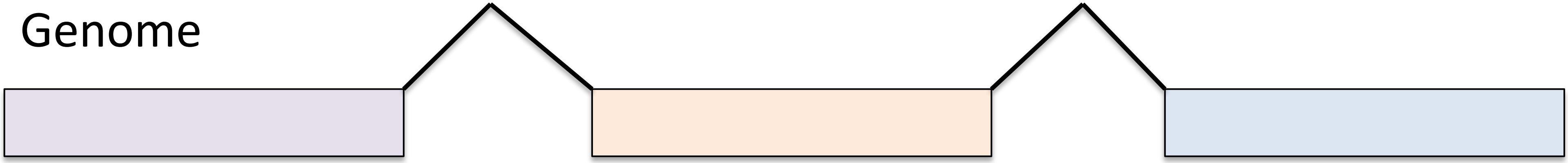
Transcriptome



N

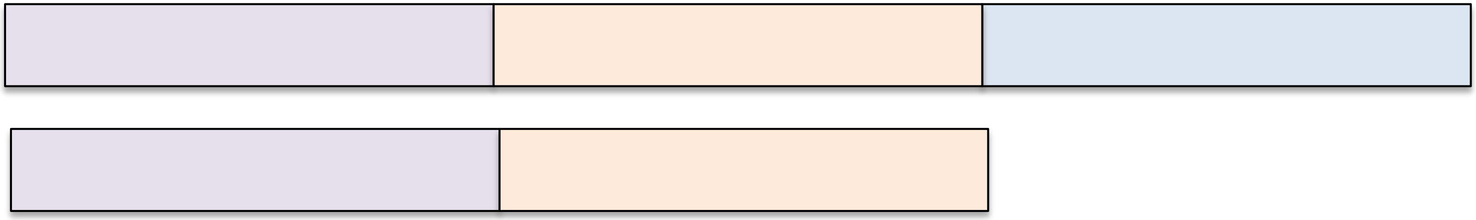
Align to Genome or Transcriptome?

Genome



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

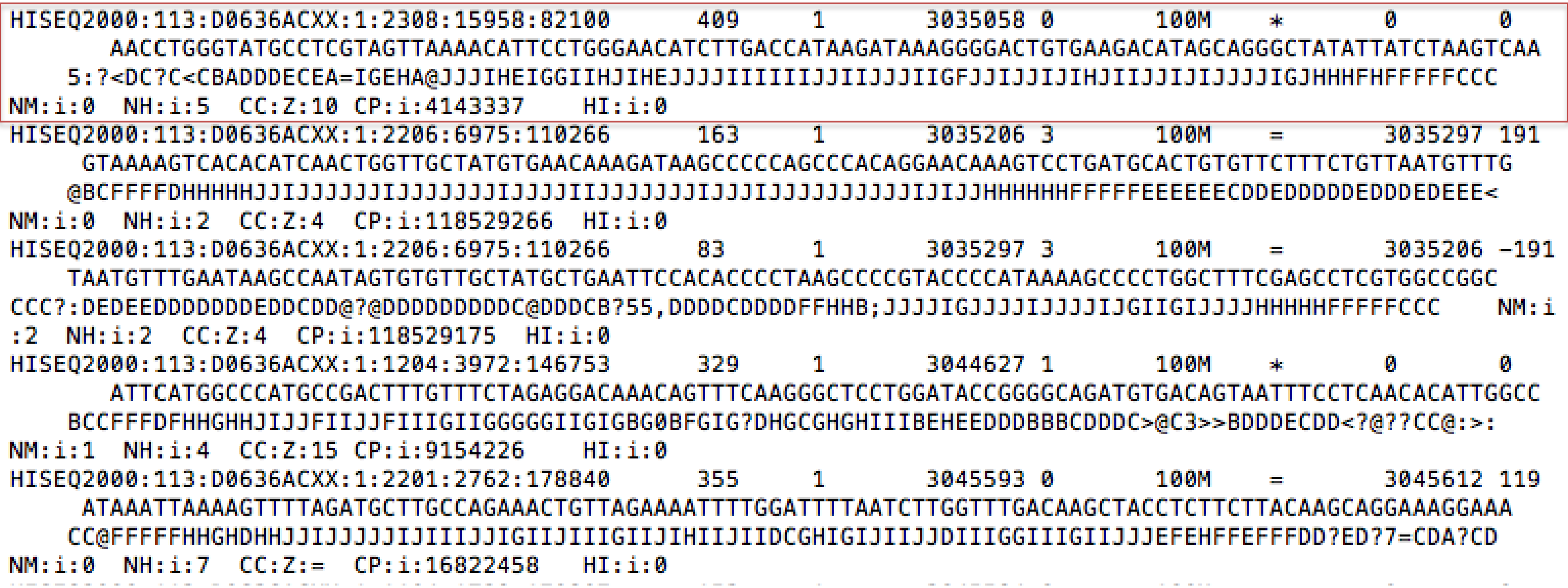
Transcriptome



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

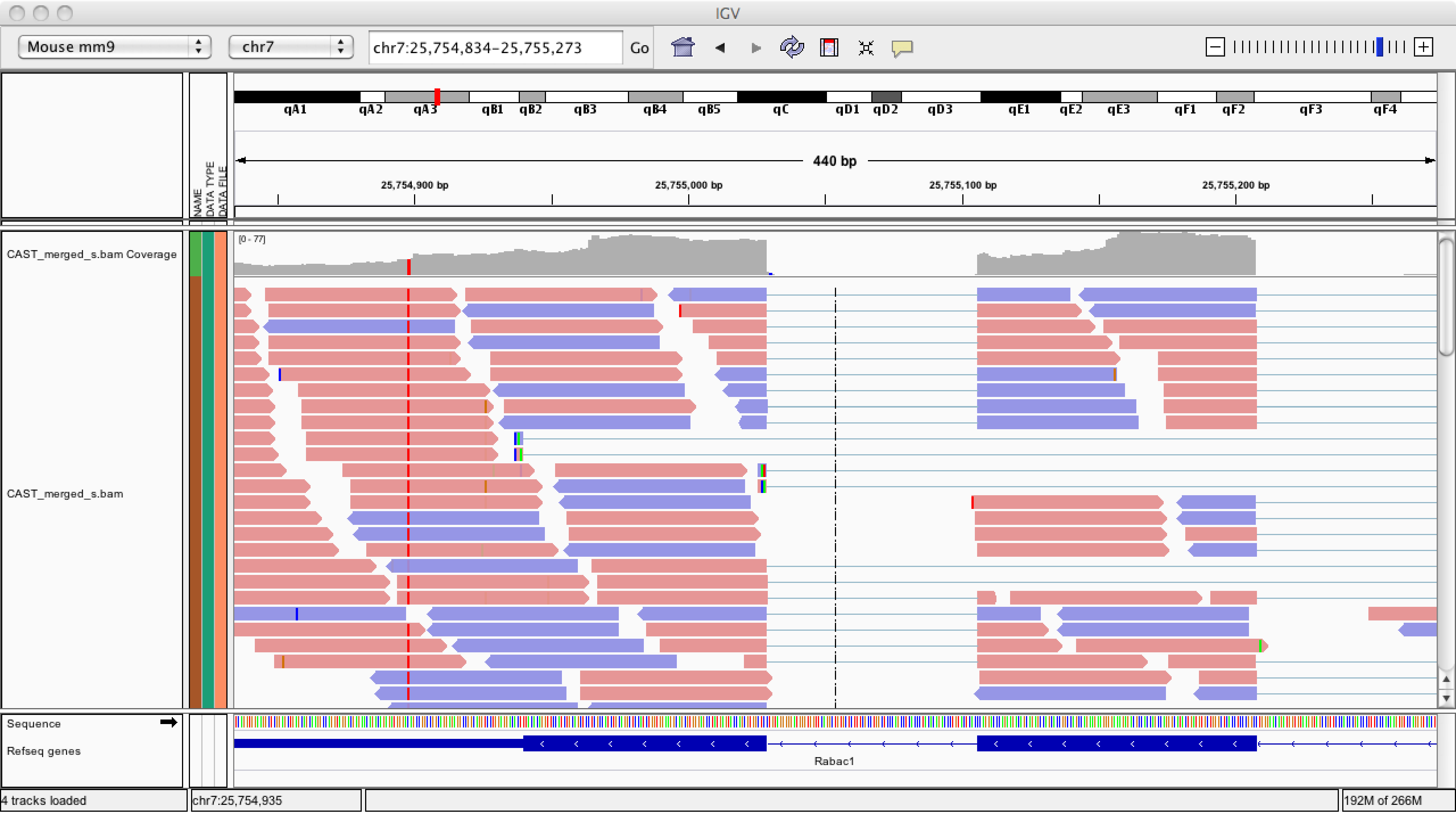


```
HISEQ2000:113:D0636ACXX:1:2308:15958:82100      409     1       3035058 0       100M    *       0       0
       AACCTGGGTATGCCTCGTAGTTAAAACATTCCTGGGAACATCTTGACCATAAGATAAAGGGGACTGTGAAGACATAGCAGGGCTATATTATCTAAGTCAA
    5:?<DC?C<CBADDDECEA=IGEHA@JJJIHEIGGIIHJIHEJJJJIIIIIJJIIJJJIIGFJJIJJIJIHJIIJJIJIJJJJIGJHHHFHFFFFFCCC
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
    @BCFFFFDHHHHHJJIJJJJJJIJJJJJJJIJJJJIIJJJJJJJIJJJIJJJJJJJJJJIJIJJHHHHHHFFFFFEEEEEECDDEDDDDDEDDDEDEEE<
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?:DEDEEDDDDDDDEDDCDD@?@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
```

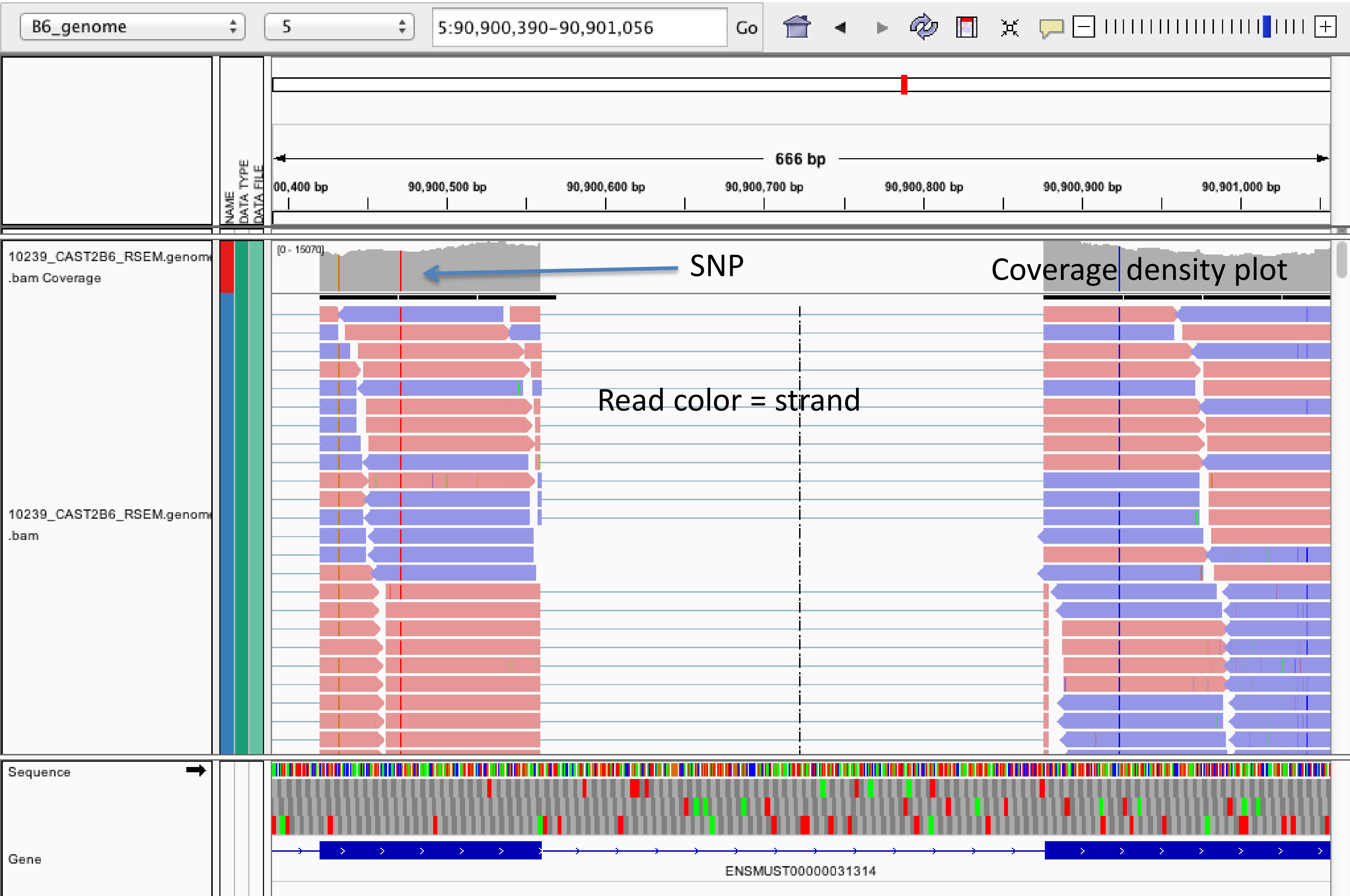
N

Visualization of alignment data (BAM/SAM)

- Genome browsers – UCSC, IGV, etc.



IGV is your friend.

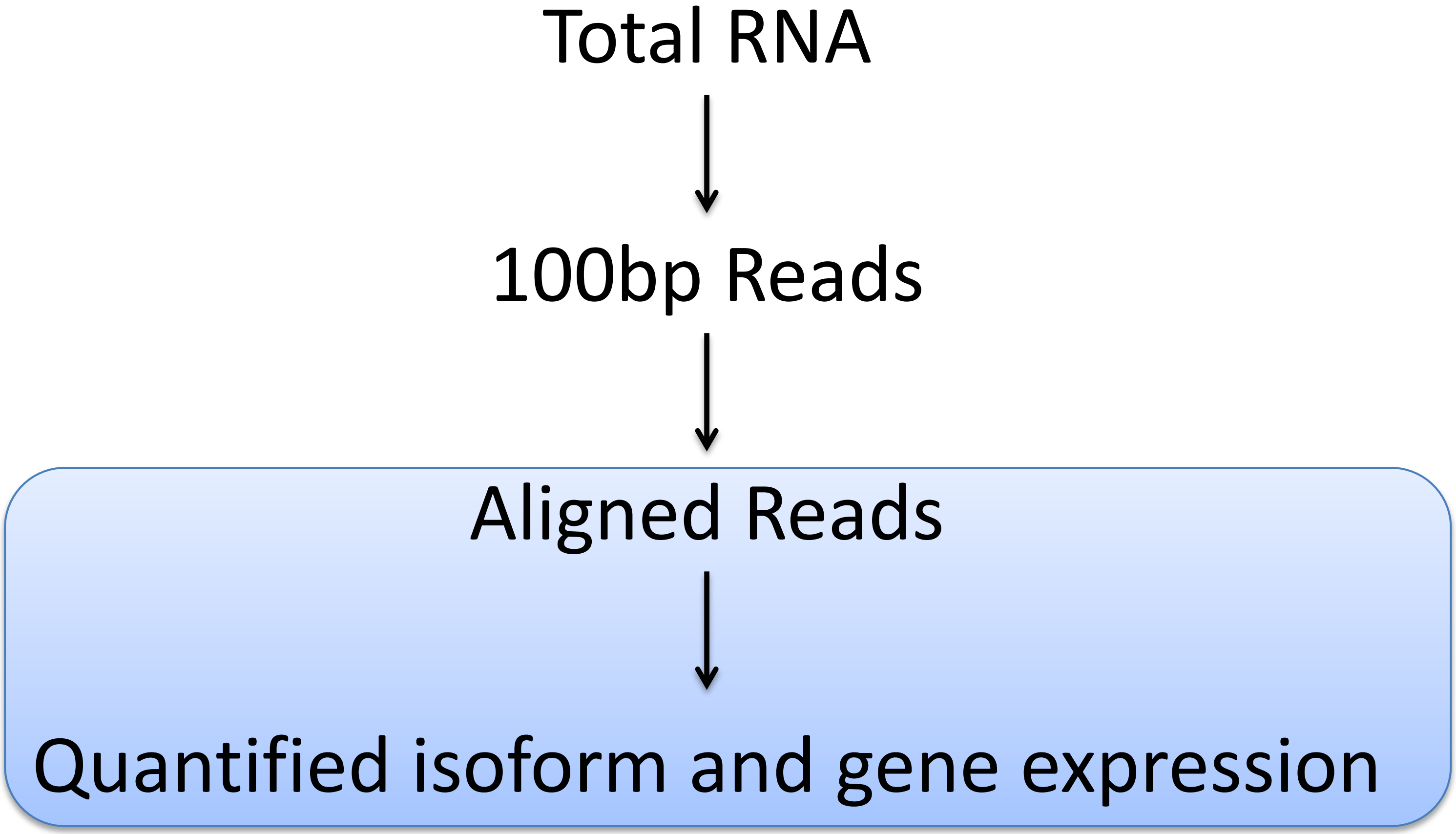


SNP

Coverage density plot

Read color = strand

Aligned Reads to Gene Abundance



Total RNA

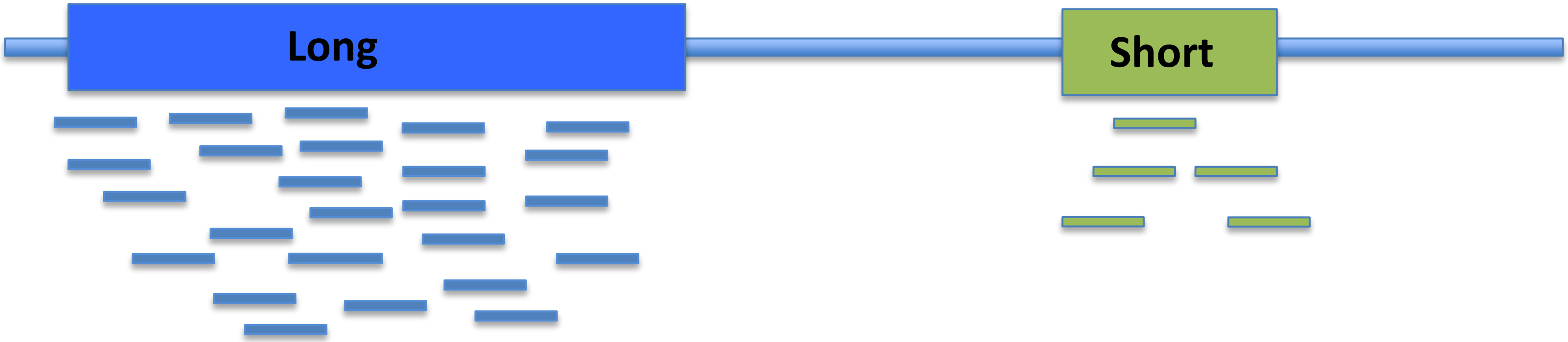
100bp Reads

Aligned Reads

Quantified isoform and gene expression

N

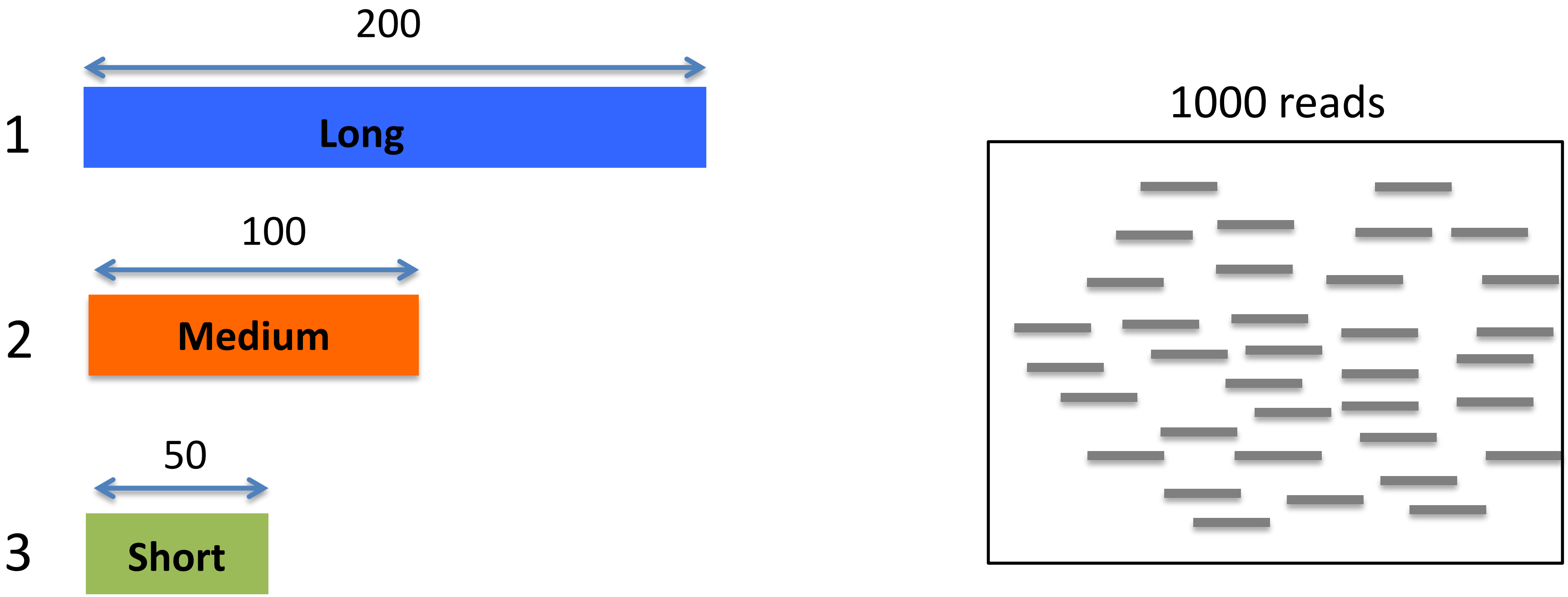
Aligned Reads to Gene Abundance: Challenges



Many approaches to quantify expression abundance

N

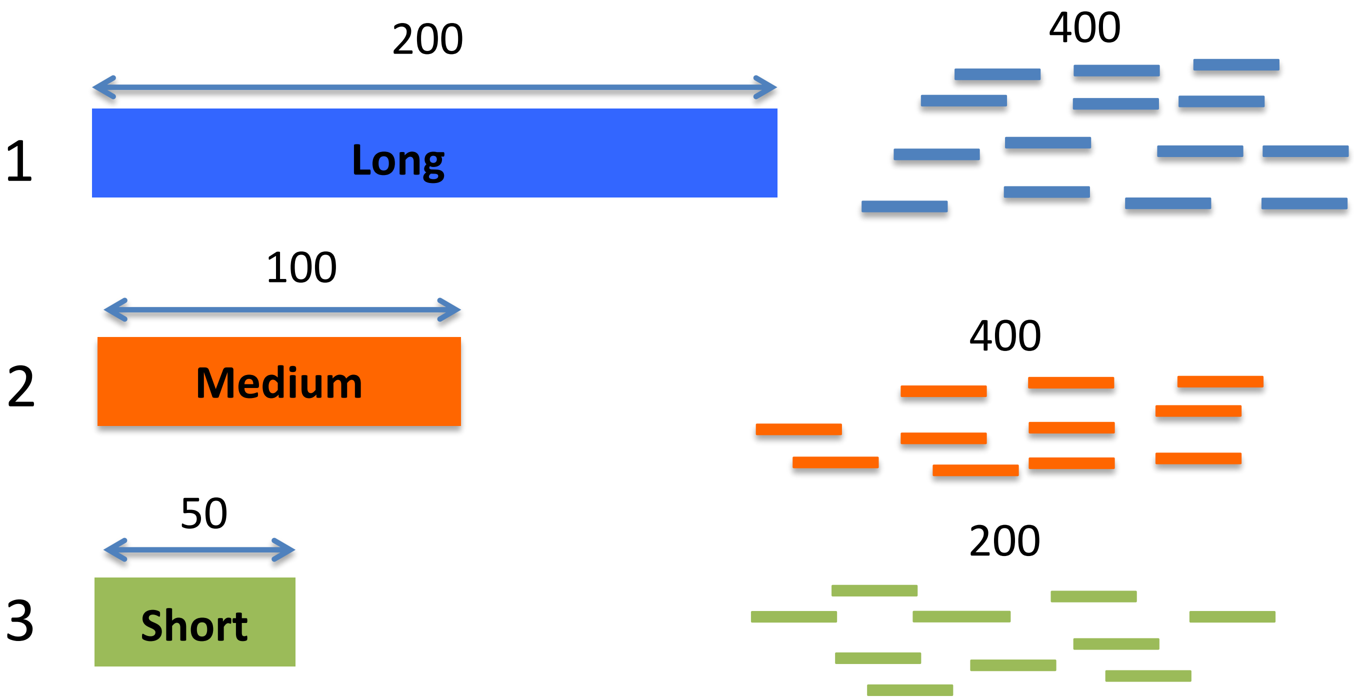
Aligned Reads to Gene Abundance: Challenges



Relative abundance for these genes, f_1, f_2, f_3

N

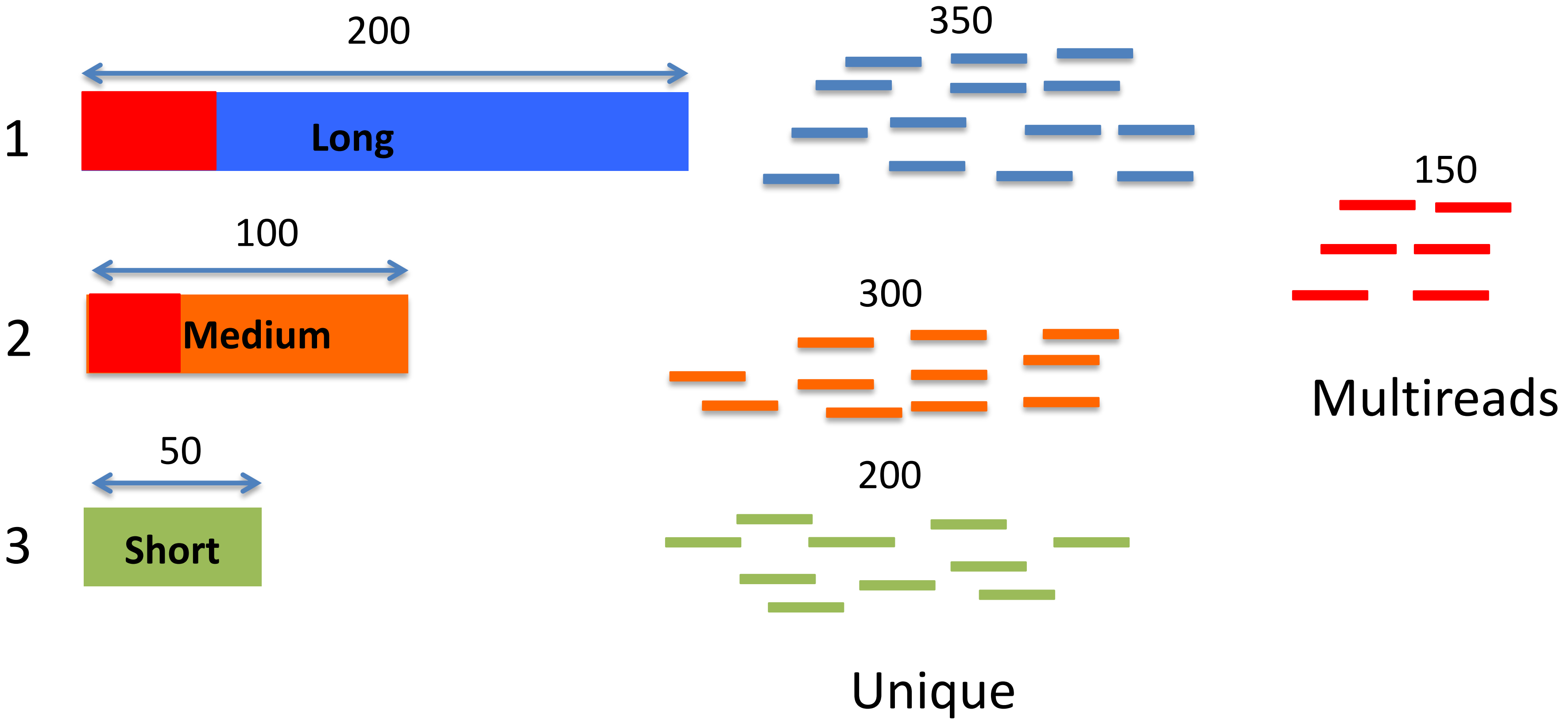
Aligned Reads to Gene Abundance: Challenges



Relative abundance for these genes, f_1, f_2, f_3

N

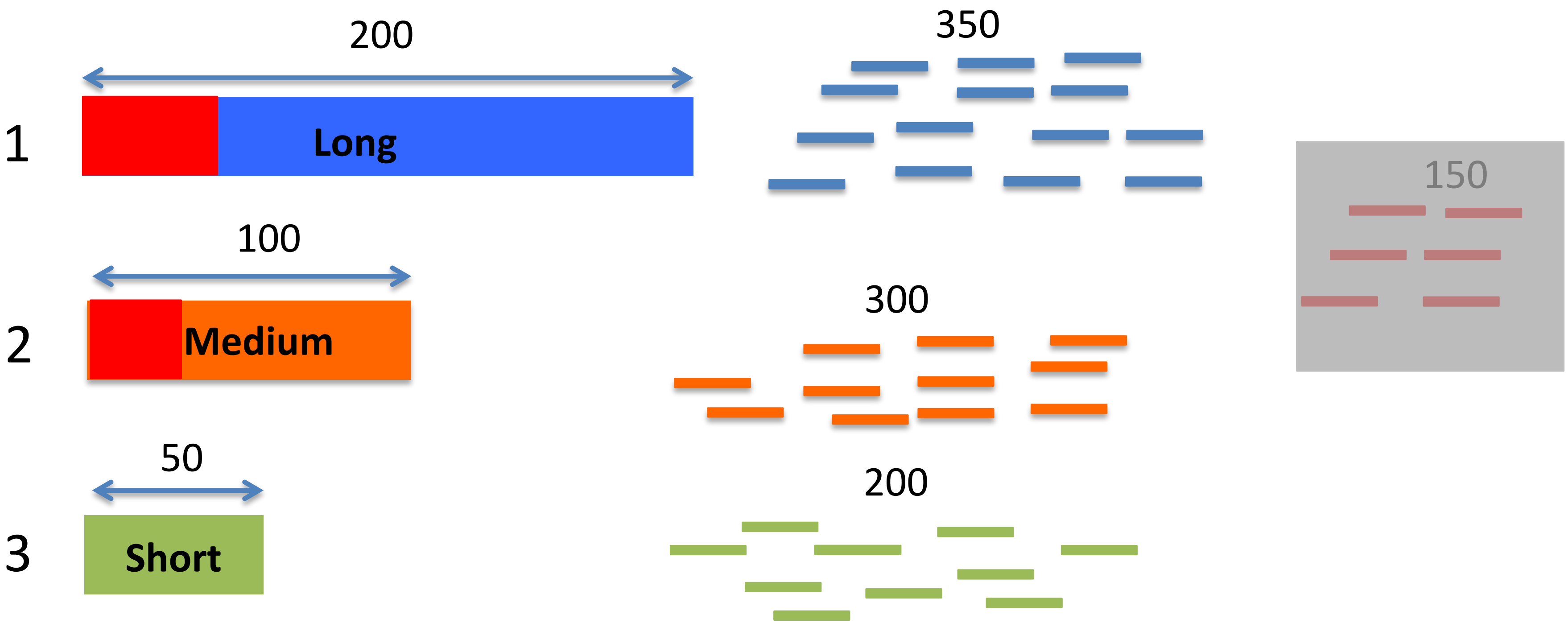
Multireads: Reads Mapping to Multiple Genes/Transcripts



Relative abundance for these genes, f_1, f_2, f_3

N

Approach 1: Ignore Multireads

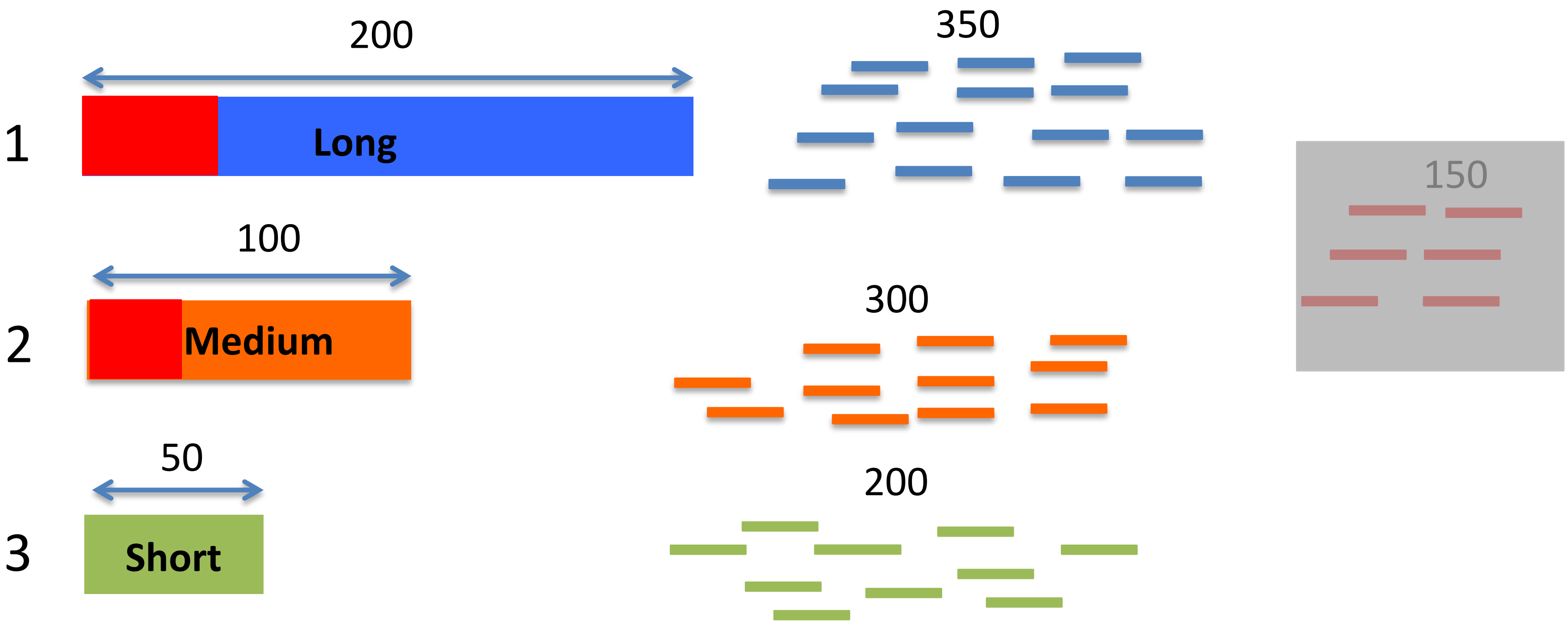


Relative abundance for these genes, f_1, f_2, f_3

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

N

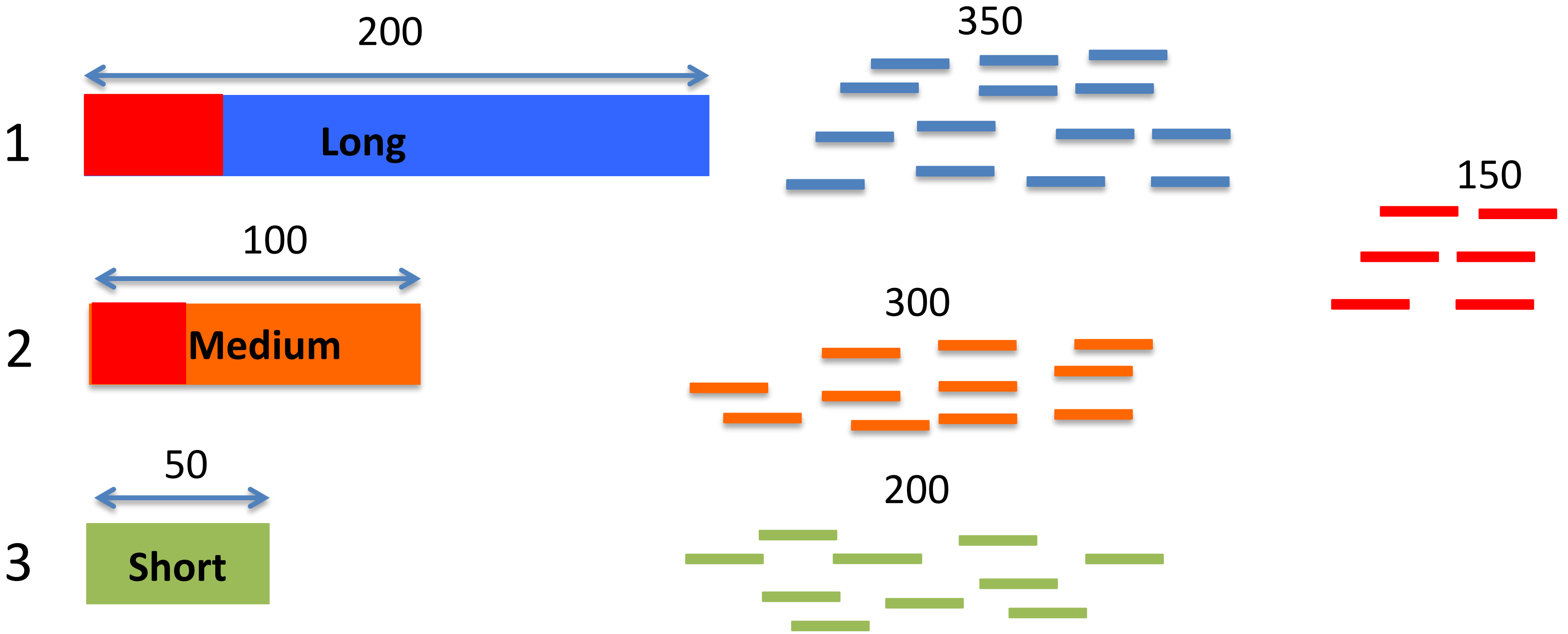
Approach 1: Ignore Multireads



- 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

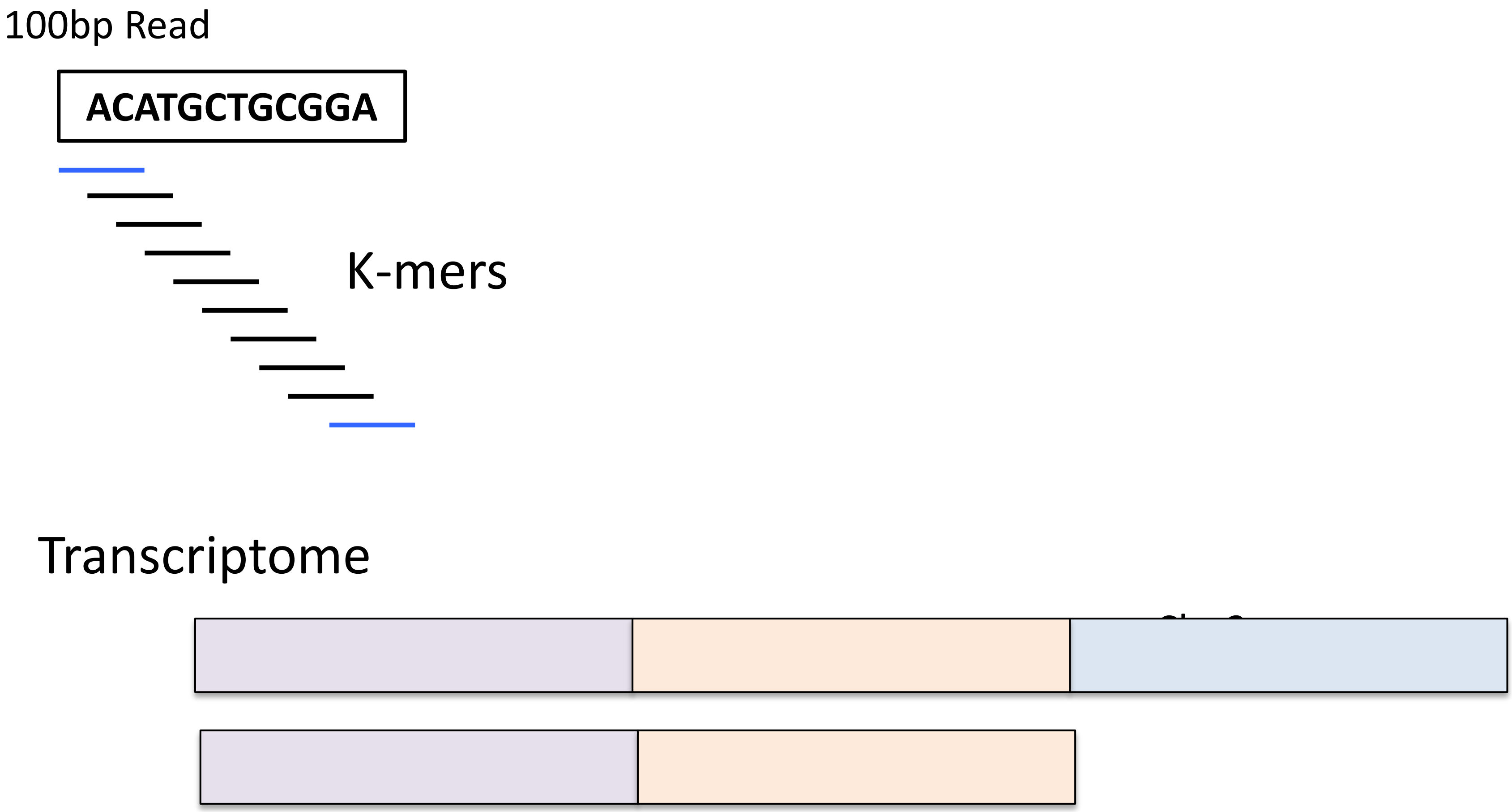


Relative abundance for these genes, f_1, f_2, f_3

RSEM, Cufflinks, isoEM, MMSEQ & eXpress

N

Sailfish, Salmon, and Kallisto: The rise of Pseduo-alignment a.k.a alignment-free methods



100bp Read

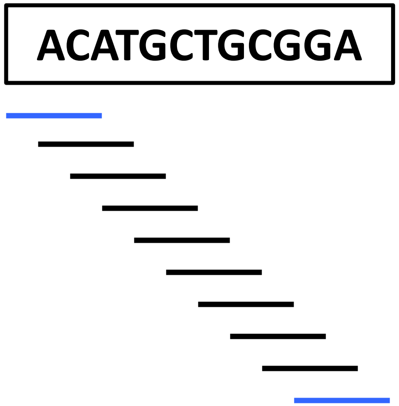
ACATGCTGCGGA

K-mers

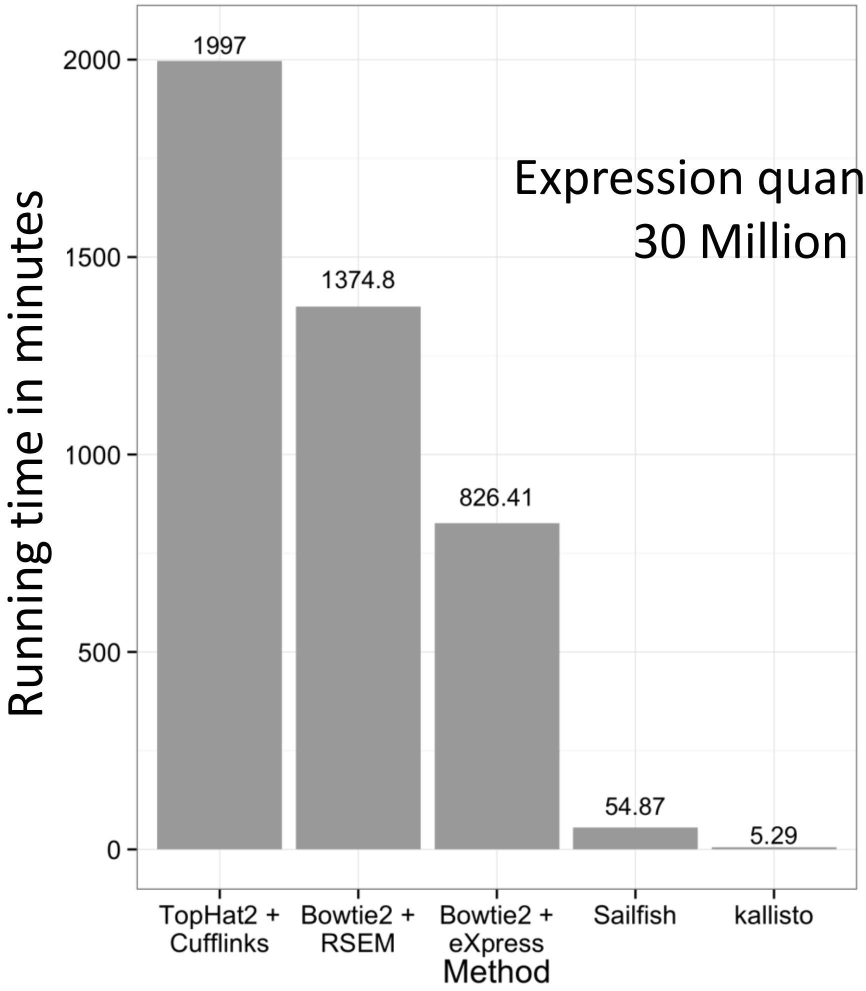
Transcriptome

Kallisto: K-mer based pseudo-alignment

100bp Read



ACATGCTGCGGA



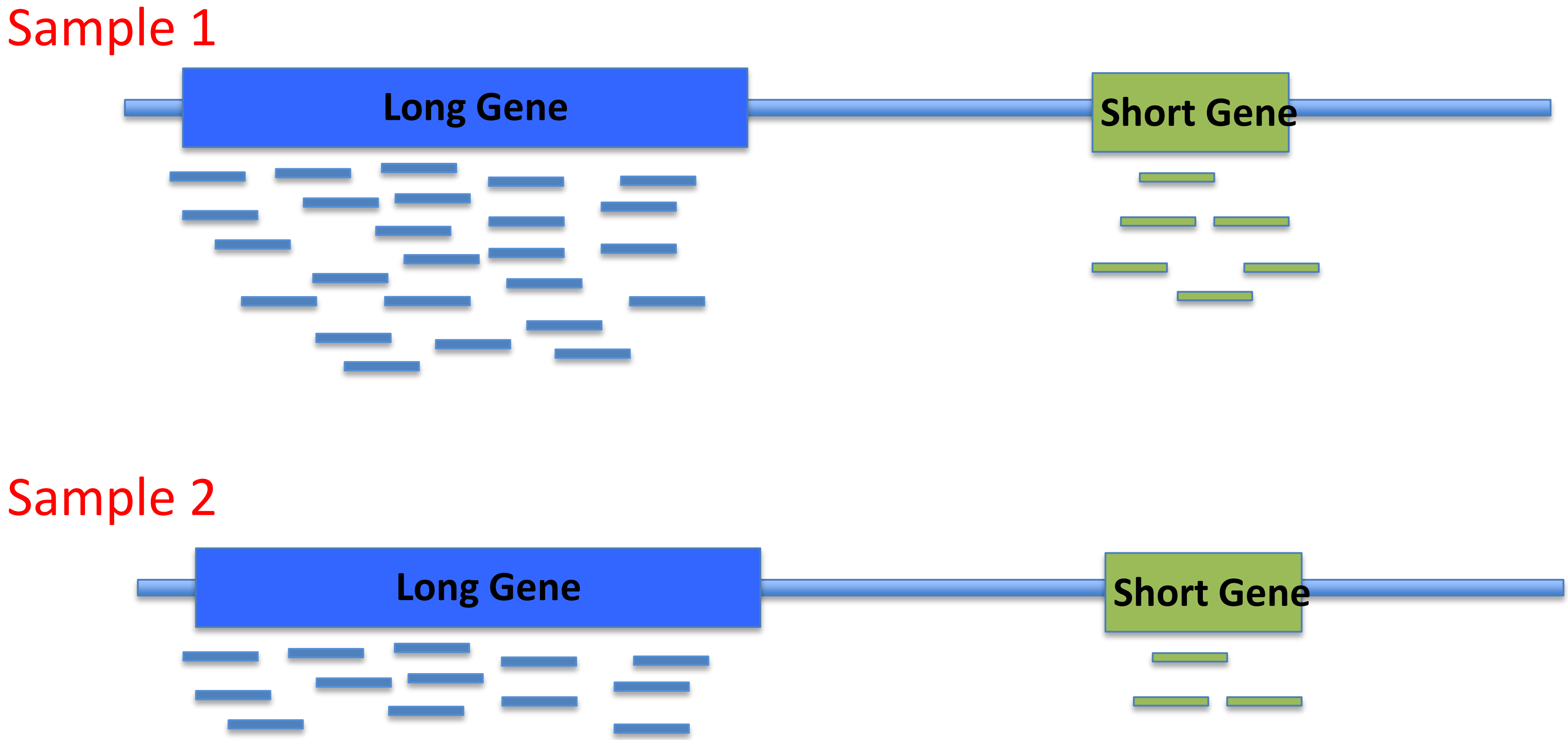
Expression quantitation for 30 Million Reads

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.

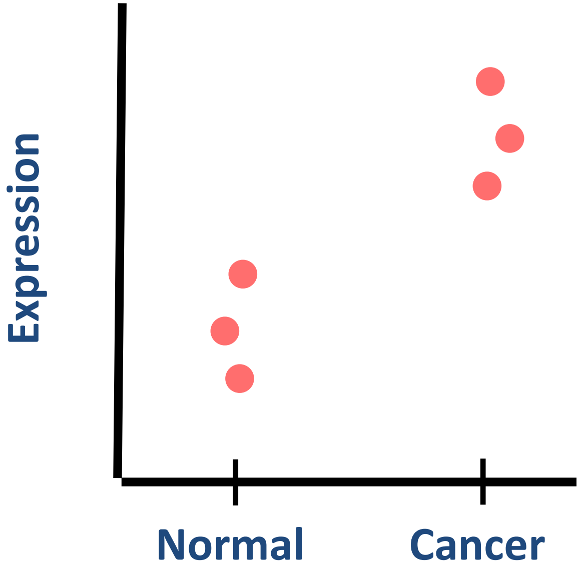
N

Expression Abundance: Counts, RPKM/FPKM, TPM



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

Differential Expression Analysis



T-test

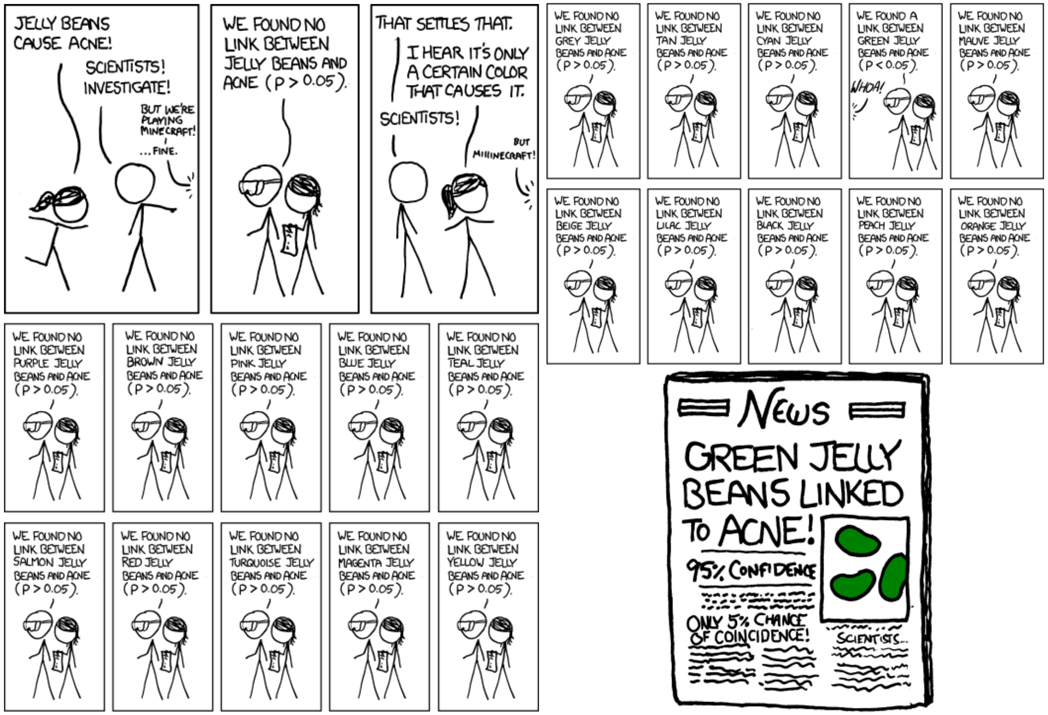
$$t_g = \frac{\hat{\mu}_{g,1} - \hat{\mu}_{g,2}}{\sqrt{\frac{\hat{\sigma}^2_{g,1}}{N_1} + \frac{\hat{\sigma}^2_{g,2}}{N_2}}}$$

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

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

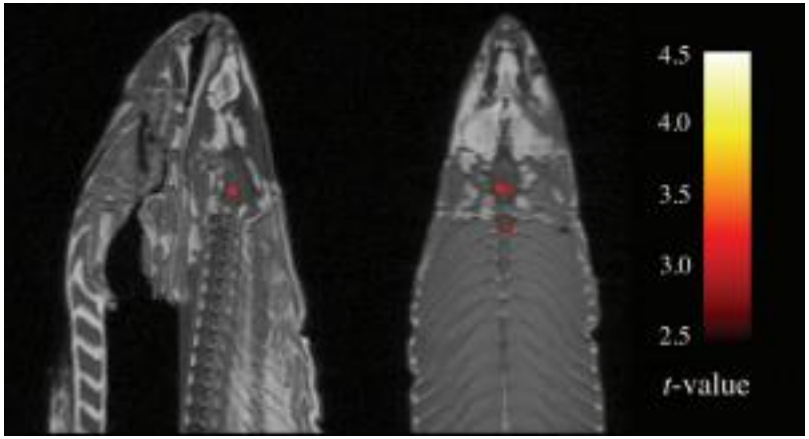
DESEQ2, edgeR, Voom, & CuffDiff

Multiple Testing Correction and False Discovery rate


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

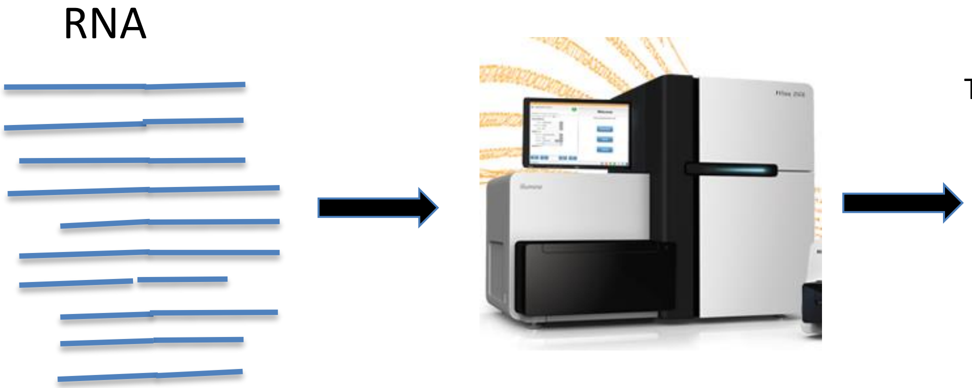
XKCD Significant

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



N

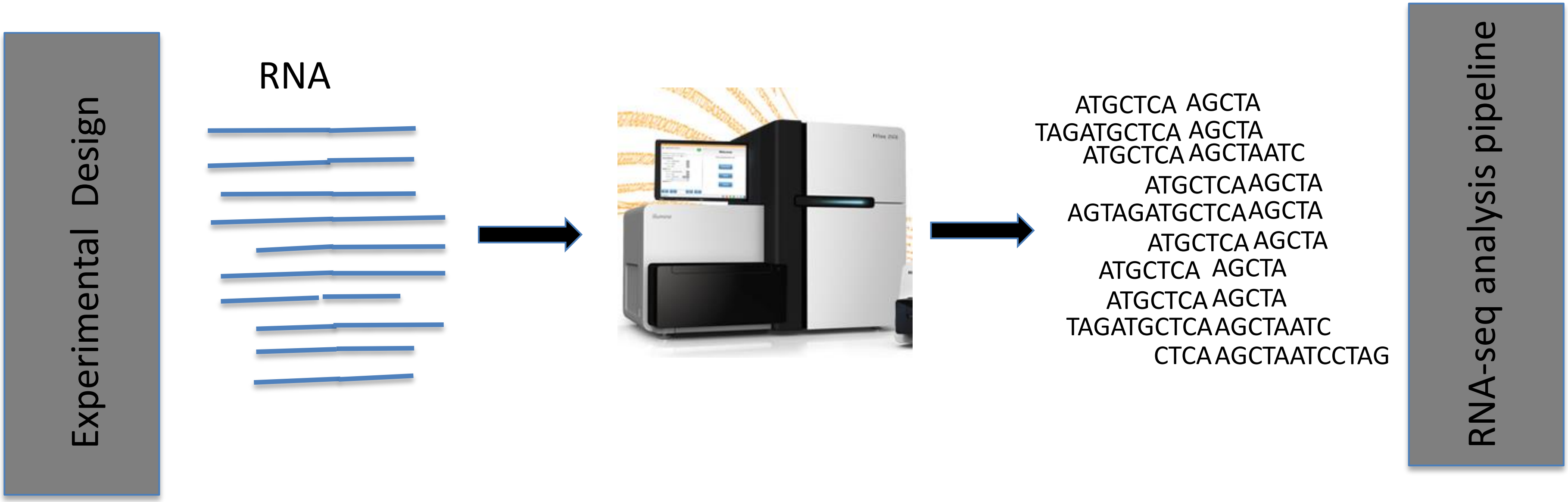
Summary



RNA

ATGCTCA AGCTA
TAGATGCTCA AGCTA
ATGCTCA AGCTAATC
ATGCTCA AGCTA
AGTAGATGCTCA AGCTA
ATGCTCA AGCTA
ATGCTCA AGCTA
ATGCTCA AGCTA
TAGATGCTCA AGCTAATC
CTCA AGCTAATCCTAG

Summary



Experimental Design

RNA

ATGCTCA AGCTA
TAGATGCTCA AGCTA
ATGCTCA AGCTAATC
ATGCTCA AGCTA
AGTAGATGCTCA AGCTA
ATGCTCA AGCTA
ATGCTCA AGCTA
ATGCTCA AGCTA
TAGATGCTCA AGCTAATC
CTCA AGCTAATCCTAG

RNA-seq analysis pipeline

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/

Acknowledgements

- Steve Munger
- Gary Churchill
- Elissa Chesler
- Ron Korstanje/ Karen Svenson/ 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