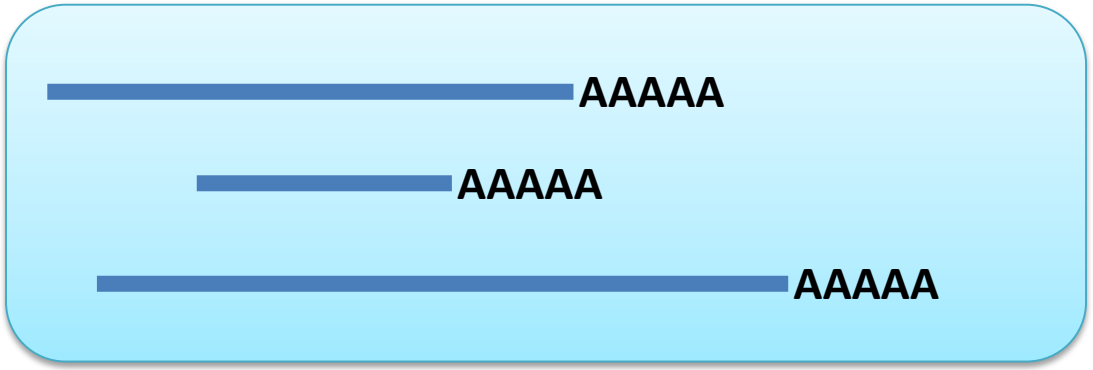
RNA-seq Data Analysis

Qi Sun
Bioinformatics Facility
Biotechnology Resource Center
Cornell University

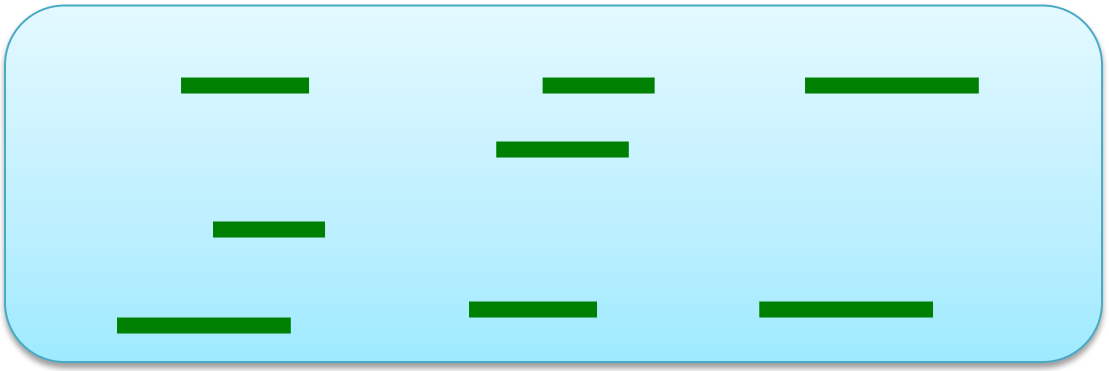
- **Lecture 1.** Mapping RNA-seq reads to the genome;
- **Lecture 2.** Quantification, normalization of gene expression & detection of differentially expressed genes;
- **Lecture 3.** Clustering; Function/Pathway Enrichment analysis

RNA-seq Experiment

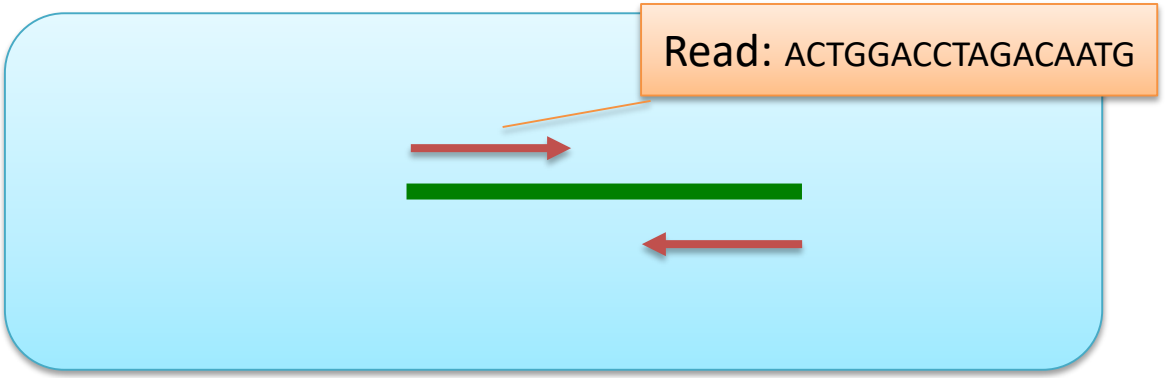
mRNA



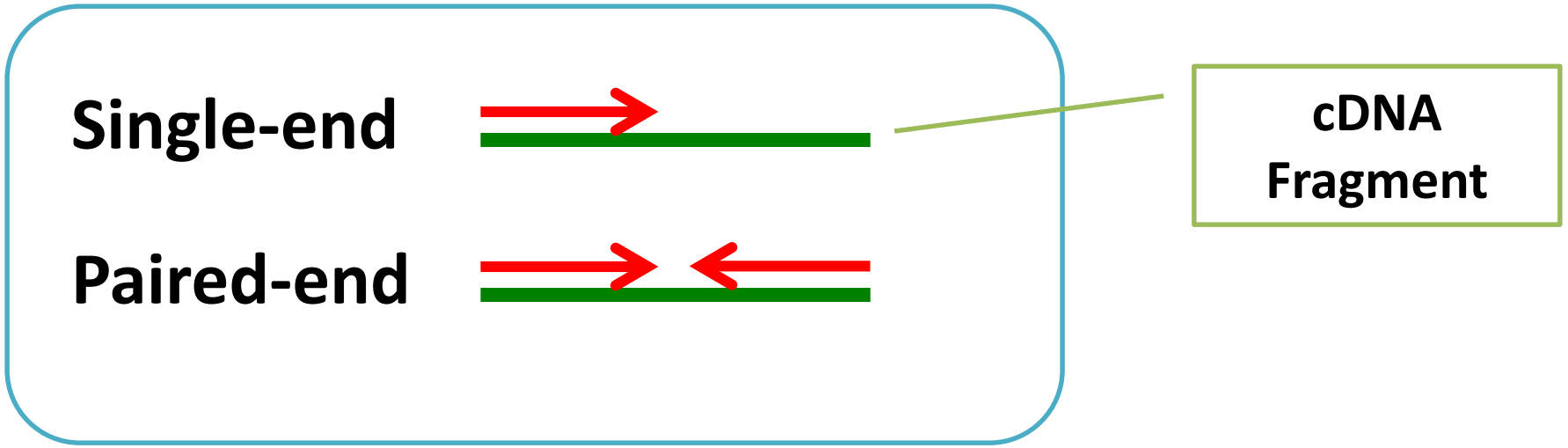
cDNA Fragments
(100 to 500 bp)



Illumina Sequencer can read single or both end(s) of each fragment

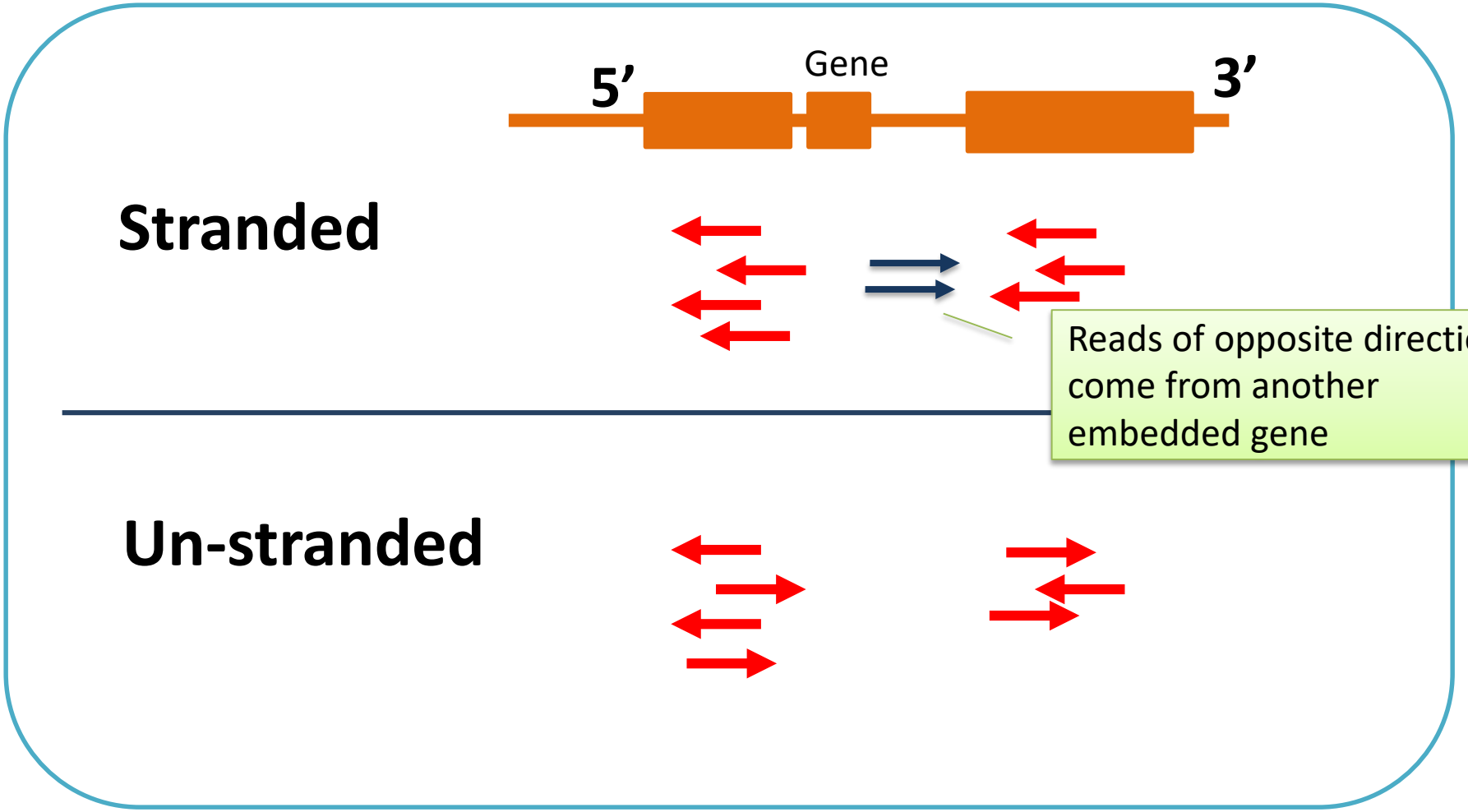


Experimental design: single-end vs paired-end



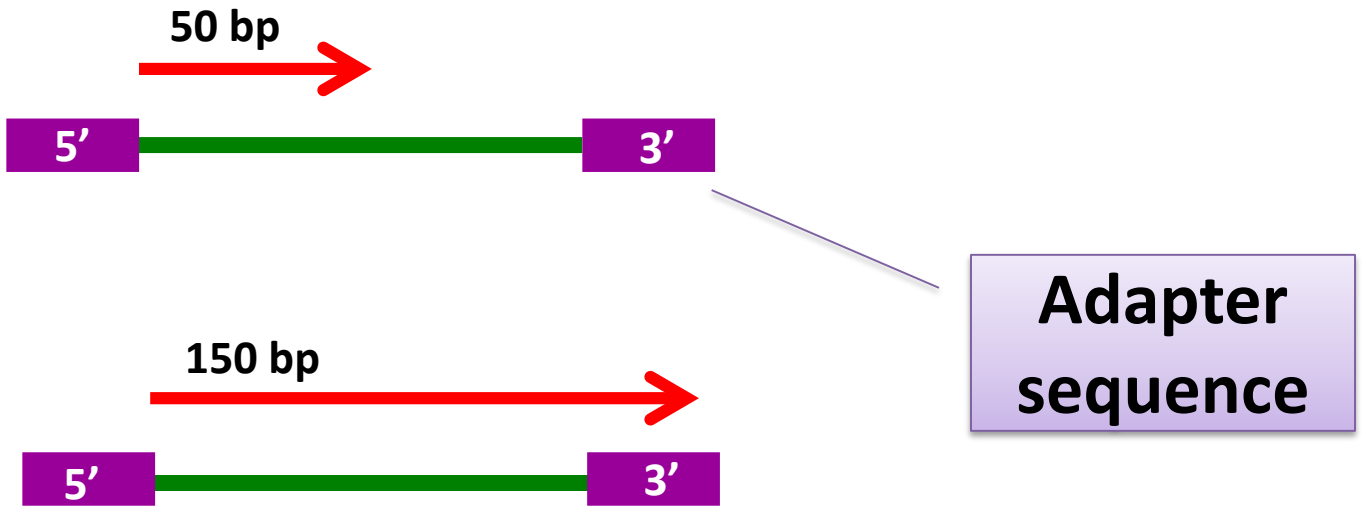
Single-end: one fastq file per sample
Paired-end: two fastq files per sample

Experimental design: stranded vs un-stranded



Experimental design: read length (50 bp, 100 bp, 150 bp, …)

Long sequence reads could read into the adapter:



Experiment design 1: for quantification of gene expression

- **Read length:** 50 to 100 bp
- **Paired vs single ends:** Single end
- **Number of reads:** >5 million per sample
- **Replicates:** 3 replicates

Experiment design 2: for RNA-seq without reference genome

- **Read length:** 100-150 bp

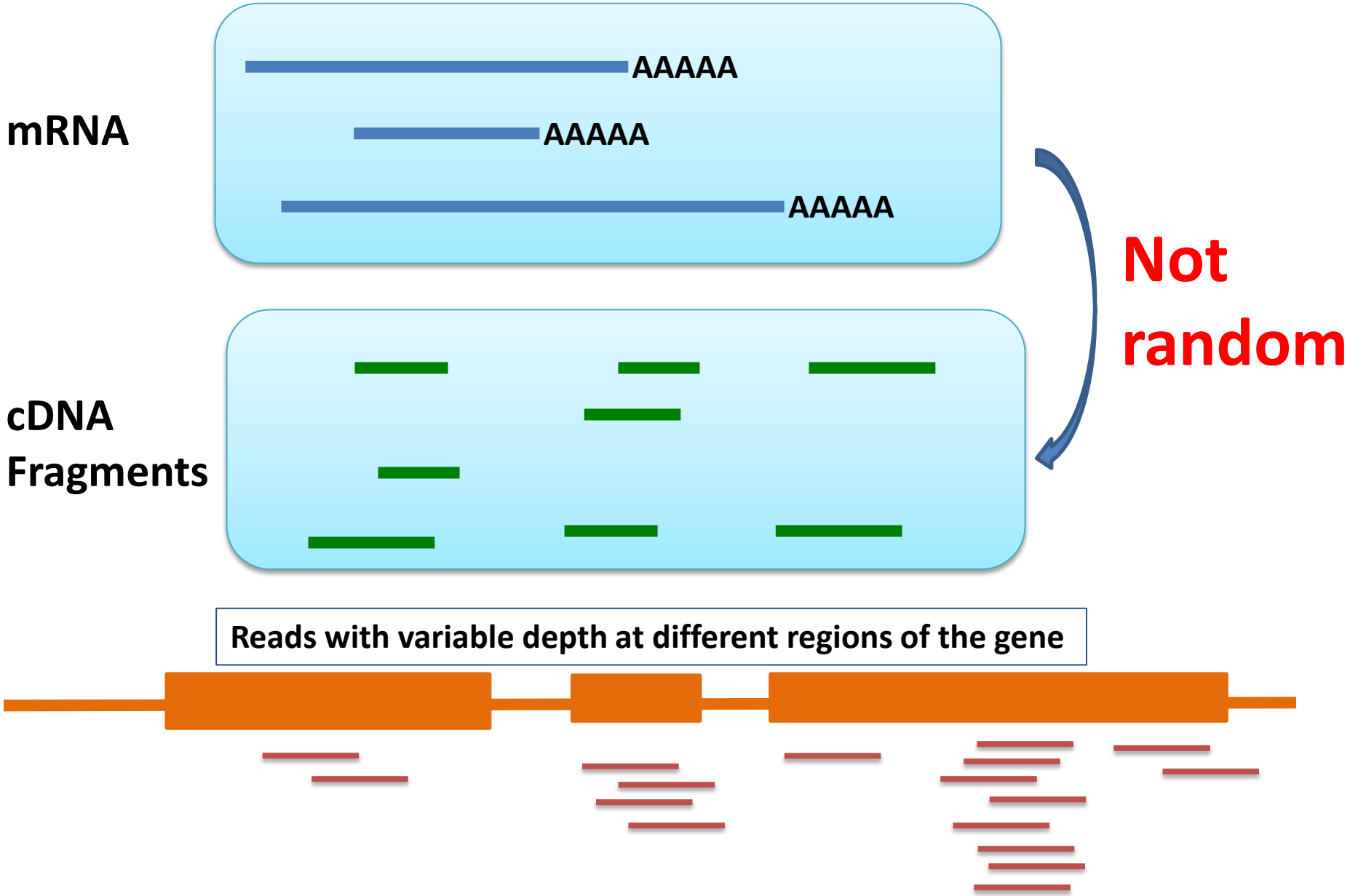
(longer reads are not always better, because of severe sequencing bias with longer fragments.)

- **Paired-end & stranded reads;**

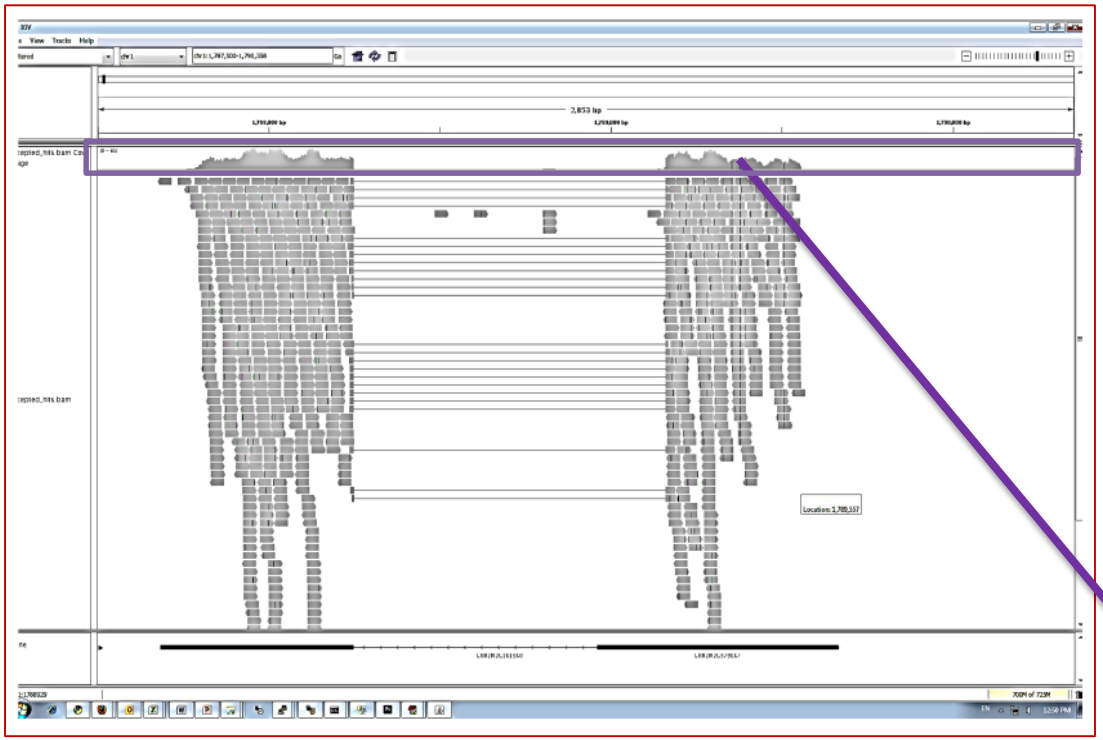
- **Higher read depth is necessary**

(Normally pooled from multiple samples)

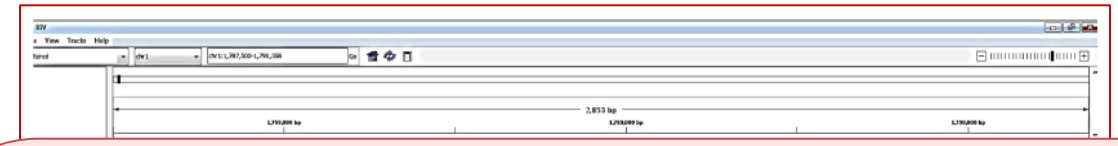
Limitation of RNA-seq: Sequencing bias



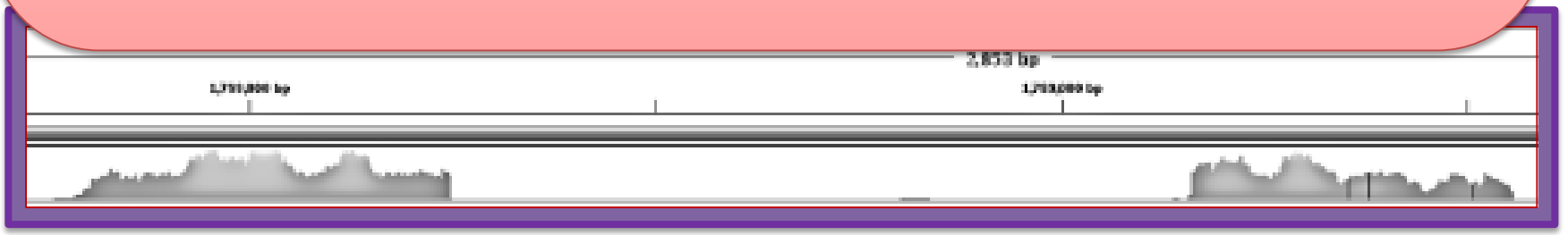
Read-depth are not even across the same gene



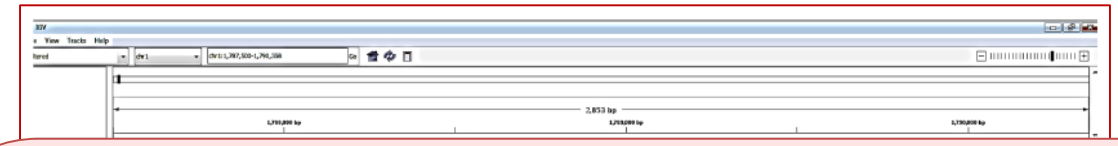
Read-depth are not even across the same gene



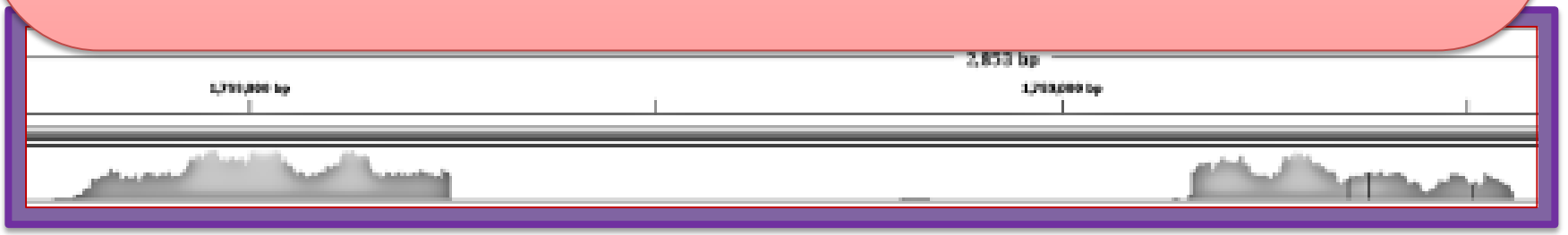
Consequence 1: batch effect (different batches or different protocols could have different ways of sequencing bias).



Read-depth are not even across the same gene

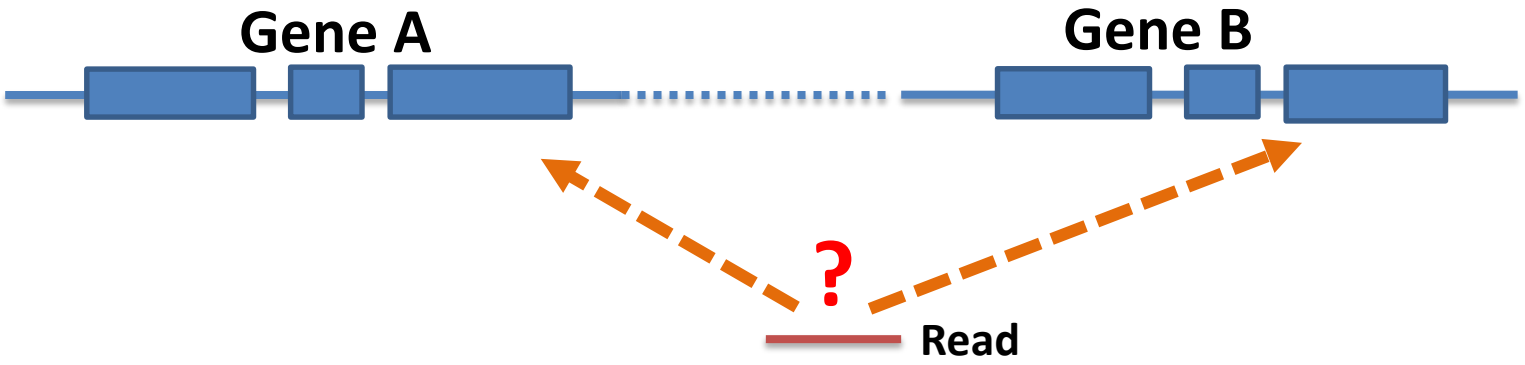


Consequence 2: RNA-seq is for comparing same gene across different samples, not for comparing different genes in the same sample.

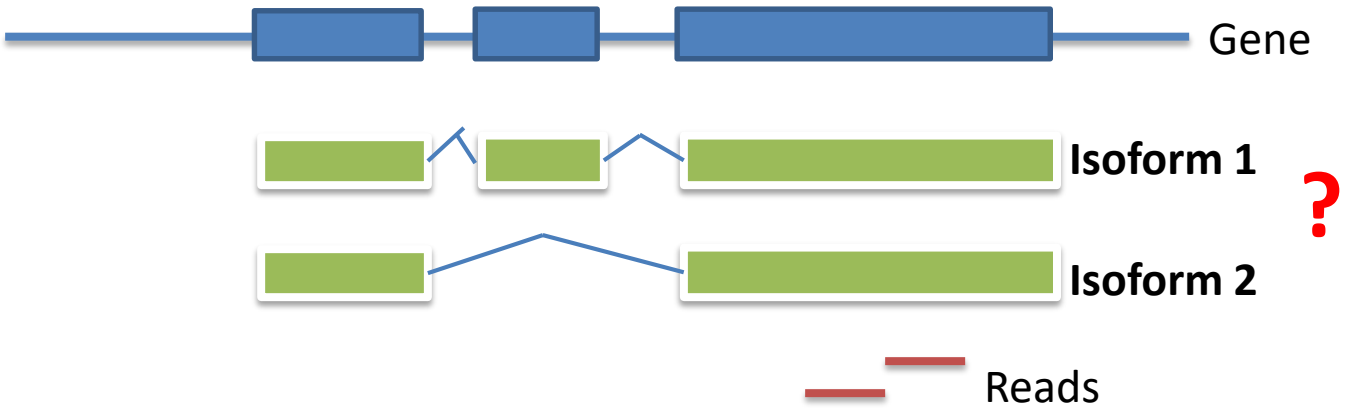


Short reads caused ambiguity in mapping

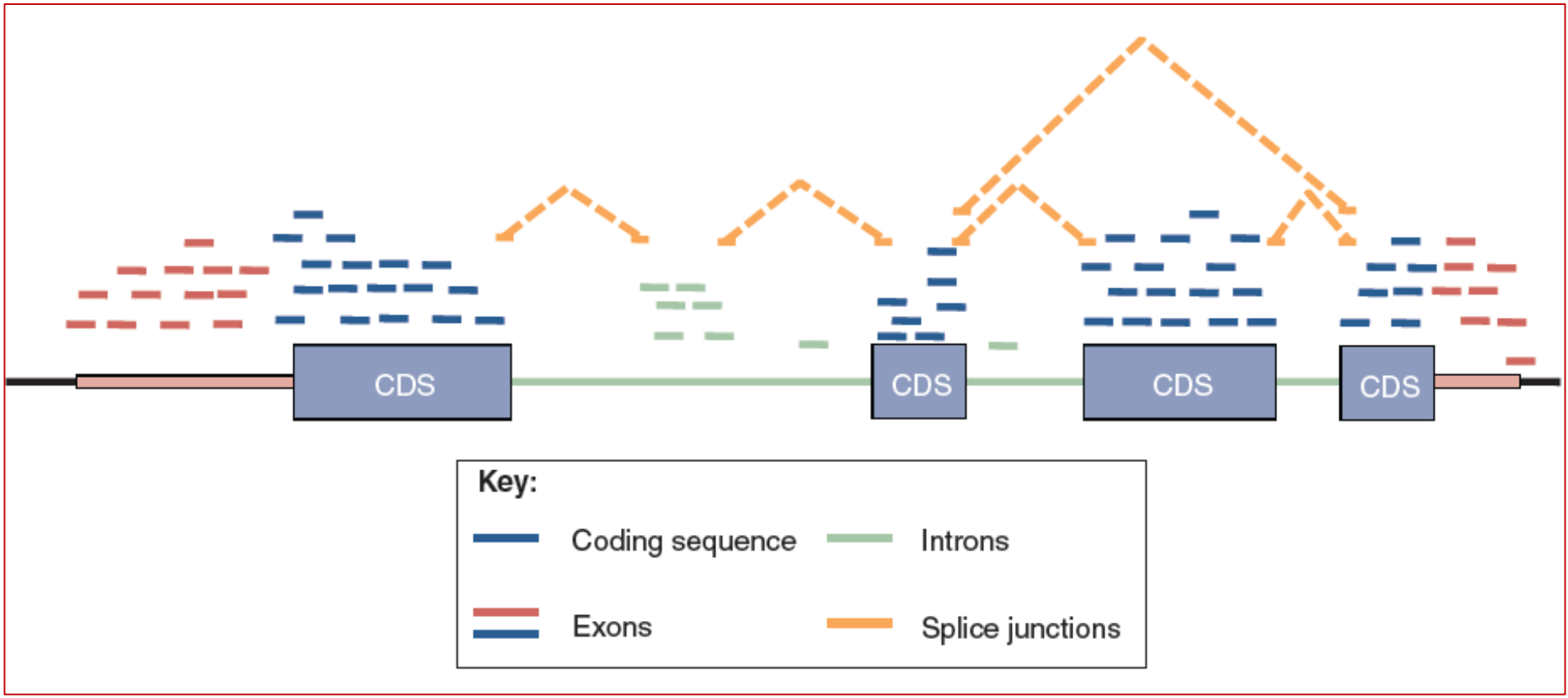
1. Reads from homologous genes;



2. Assignment to correct splicing isoform;



RNA-seq Data Analysis



Data analysis procedures

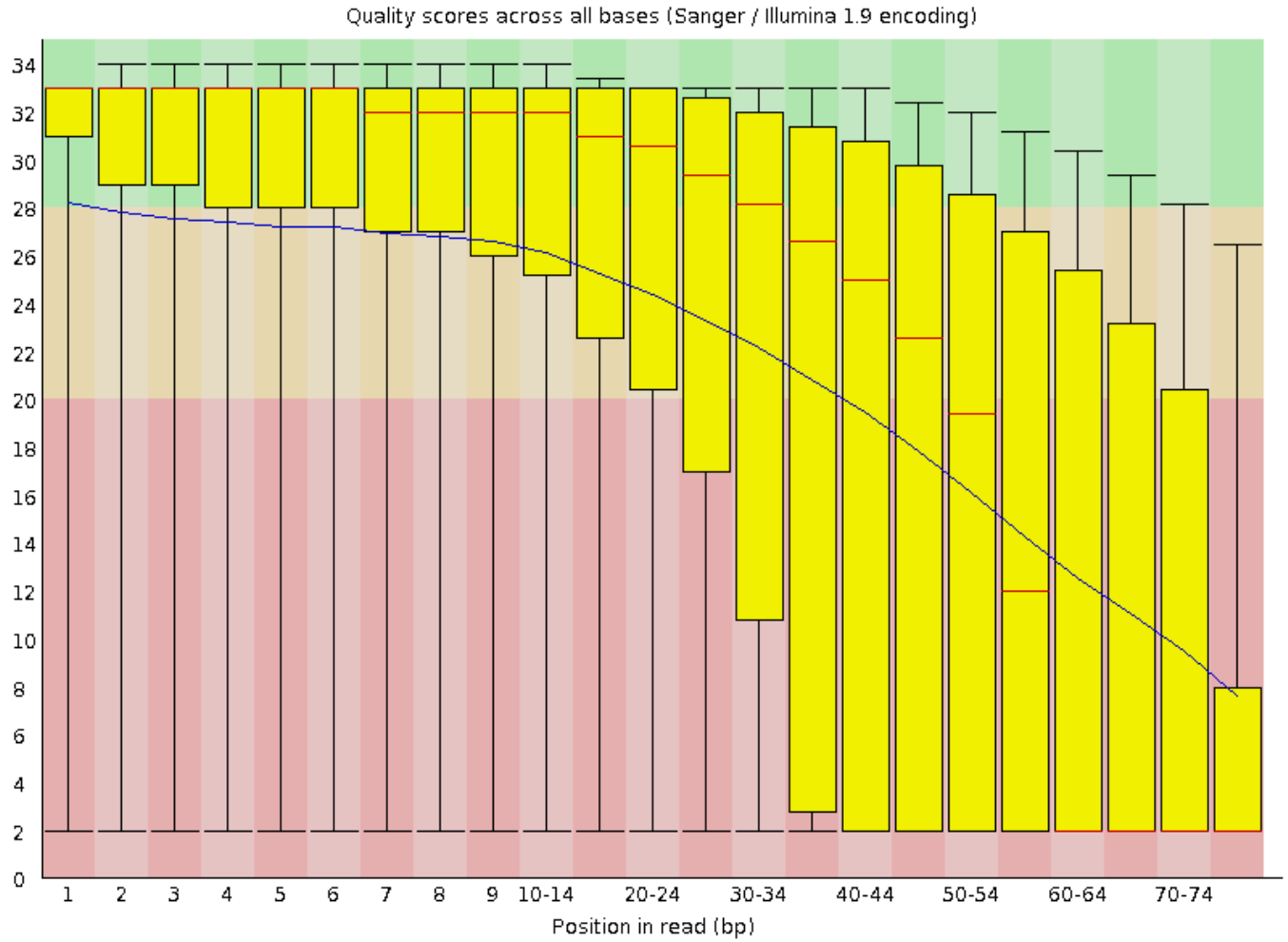
Step 1. **Check quality of the reads (optional);**

Step 2. **Map reads to the genome;**

Step 3. **Estimate expression levels by counting reads per gene.**

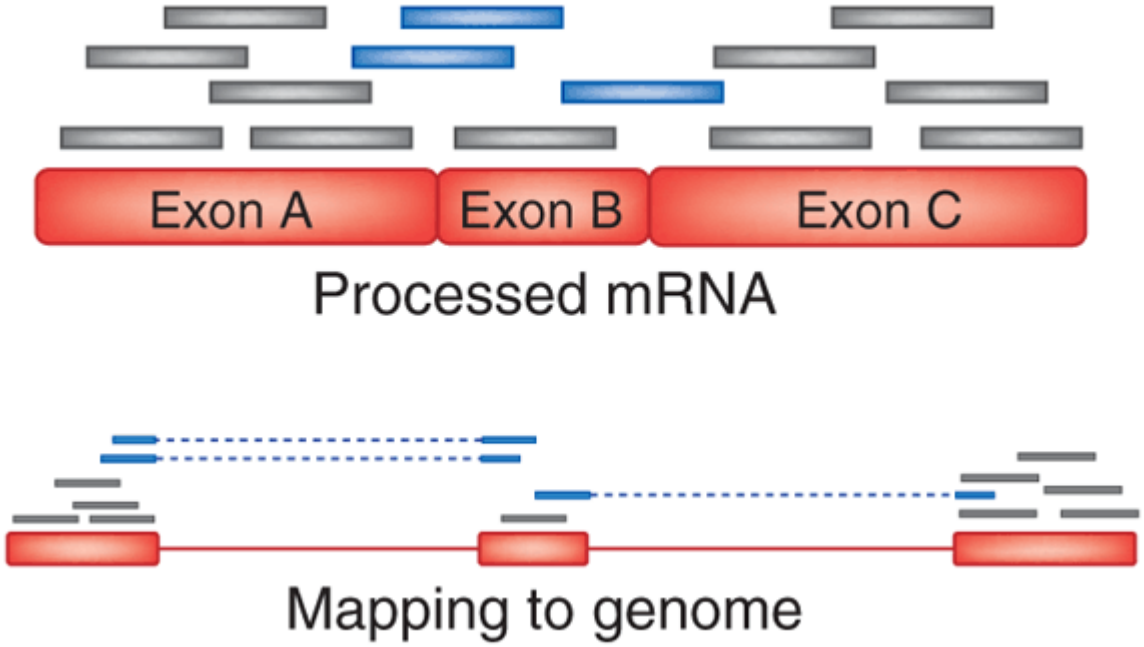
Step 1. Quality Control (QC) using FASTQC Software

1. Sequencing quality score



Step 2. Map reads to genome using TOPHAT Software

- Alignment of genomic sequencing vs RNA-seq



Cole Trapnell & Steven L Salzberg, *Nature Biotechnology* **27**, 455 - 457 (2009)

Diagnose low mapping rate

1. Low quality reads or reads with adapters *
   - Trimming tools (FASTX, Trimmomatic, et al.)

2. Contamination?
   - fastq_species_detector (Available on BioHPC Lab. It identifies species for reads by blast against Genbank)

* Trimming is not needed in majority of RNA-seq experiments except for de novo assembly

fastq_species_detector

A BioHPC tool for detecting contaminants

Commands:

```
mkdir /workdir/my_db
cp /shared_data/genome_db/BLAST_NCBI/nt*  /workdir/my_db
cp /shared_data/genome_db/BLAST_NCBI/taxdb.*  /workdir/my_db
/programs/fastq_species_detector/fastq_species_detector.sh  my_file.fastq.gz  /workdir/my_db
```

Sample output:

Read distribution over species:

Species	#Reads	%Reads
Drosophila melagaster	254	35.234
Cyprinus carpio	74	10.529
Triticum aestivum	12	2.059
Microtus ochrogaster	3	1.765
Dyella jiangningensis	3	1.765

About the files

1. Reference genome (FASTA)
2. FASTQ
3. GFF3/GTF
4. SAM/BAM

```
>chr1
TTCTAGGTCTGCGATATTTCCTGCCTATCCATTTTGTTAACTCTTCAATG
CATTCCACAAATACCTAAGTATTCTTTAATAATGGTGGTTTTTTTTTTTT
TTTGCATCTATGAAGTTTTTTCAAATTCTTTTTAAGTGACAAAACTTGTA
CATGTGTATCGCTCAATATTTCTAGTCGACAGCACTGCTTTCGAGAATGT
AAACCGTGCACTCCCAGGAAAATGCAGACACAGCACGCCTCTTTGGGACC
GCGGTTTATACTTTCGAAGTGCTCGGAGCCCTTCCTCCAGACCGTTCTCC
CACACCCCGCTCCAGGGTCTCTCCCGGAGTTACAAGCCTCGCTGTAGGCC
CCGGGAACCCAACGCGGTGTCAGAGAAGTGGGGTCCCCTACGAGGGACCA
GGAGCTCCGGGCGGGCAGCAGCTGCGGAAGAGCCGCGCGAGGCTTCCCAG
AACCCGGCAGGGGCGGGAAGACGCAGGAGTGGGGAGGCGGAACCGGGACC
CCGCAGAGCCCGGGTCCCTGCGCCCCACAAGCCTTGGCTTCCCTGCTAGG
GCCGGGCAAGGCCGGGTGCAGGGCGCGGCTCCAGGGAGGAAGCTCCGGGG
CGAGCCCAAGACGCCTCCCGGGCGGTCGGGGCCCAGCGGCGGCGTTCGCA
GTGGAGCCGGGCACCGGGCAGCGGCCGCGGAACACCAGCTTGGCGCAGGC
TTCTCGGTCAGGAACGGTCCCGGGCCTCCCGCCCGCCTCCCTCCAGCCCC
TCCGGGTCCCCTACTTCGCCCCGCCAGGCCCCCACGACCCTACTTCCCGC
GGCCCCGGACGCCTCCTCACCTGCGAGCCGCCCTCCCGGAAGCTCCCGCC
GCCGCTTCCGCTCTGCCGGAGCCGCTGGGTCCTAGCCCCGCCGCCCCCAG
TCCGCCCGCGCCTCCGGGTCCTAACGCCGCCGCTCGCCCTCCACTGCGCC
CTCCCCGAGCGCGGCTCCAGGACCCCGTCGACCCGGAGCGCTGTCCTGTC
GGGCCGAGTCGCGGGCCTGGGCACGGAACTCACGCTCACTCCGAGCTCCC
GACGTGCACACGGCTCCCATGCGTTGTCTTCCGAGCGTCAGGCCGCCCCT
ACCCGTGCTTTCTGCTCTGCAGACCCTCTTCCTAGACCTCCGTCCTTTGT
```

About the files

1. FASTA
2. **RNA-seq data (FASTQ)**
3. GFF3/GTF
4. SAM/BAM

```
@HWUSI-EAS525:2:1:13336:1129#0/1
GTTGGAGCCGGCGAGCGGGACAAGGCCCTTGTCCA
+
ccacacccacccccccccc[[cccc_ccaccbbb_
@HWUSI-EAS525:2:1:14101:1126#0/1
GCCGGGACAGCGTGTTGGTTGGCGCGCGGTCCCTC
+
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWUSI-EAS525:2:1:15408:1129#0/1
CGGCCTCATTCTTGGCCAGGTTCTGGTCCAGCGAG
+
cghhchhgchehhdffccgdgh]gcchhcahWcea
@HWUSI-EAS525:2:1:15457:1127#0/1
CGGAGGCCCCCGCTCCTCTCCCCCGCGCCCGCGCC
+
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWUSI-EAS525:2:1:15941:1125#0/1
TTGGGCCCTCCTGATTTCATCGGTTCTGAAGGCTG
+
SUIF\_XYWW]VaOZZZ\V\bYbb_]ZXTZbbb_b
@HWUSI-EAS525:2:1:16426:1127#0/1
GCCCGTCCTTAGAGGCTAGGGGACCTGCCCGCCGG
```

About the files

1. FASTA
2. **RNA-seq data (FASTQ)**
3. GFF3/GTF
4. SAM/BAM

```
@HWUSI-EAS525:2:1:13336:1129#0/1
GTTGGAGCCGGCGAGCGGGACAAGGCCCTTGTCCA
+
ccacacccacccccccccc[[cccc_ccaccbbb_
@HWUSI-EAS525:2:1:14101:1126#0/1
GCCGGGACAGCGTGTTGGTTGGCGCGCGGTCCCTC
+
```

Single-end data: one file per sample

Paired-end data: two files per sample

```
+
SUIF\_XYWW]VaOZZZ\V\bYbb_]ZXTZbbb_b
@HWUSI-EAS525:2:1:16426:1127#0/1
GCCCGTCCTTAGAGGCTAGGGGACCTGCCCGCCGG
```

About the files

1. FASTA
2. FASTQ
3. **Annotation (GFF3/GTF)**
4. SAM/BAM

```
chr12   unknown exon    96066054        96067770
.       +       .       gene_id "PGAM1P5"; gene_name
"PGAM1P5"; transcript_id "NR_077225"; tss_id "TSS14770";
chr12   unknown CDS     96076483        96076598
.       -       1       gene_id "NTN4"; gene_name
"NTN4"; p_id "P12149"; transcript_id "NM_021229"; tss_id
"TSS6395";
chr12   unknown exon    96076483        96076598
.       -       .       gene_id "NTN4"; gene_name
"NTN4"; p_id "P12149"; transcript_id "NM_021229"; tss_id
"TSS6395";
chr12   unknown CDS     96077274        96077487
.       -       2       gene_id "NTN4"; gene_name
"NTN4"; p_id "P12149"; transcript_id "NM_021229"; tss_id
"TSS6395";
...
```

This file can be opened in Excel

About the files

1. FASTA
2. FASTQ
3. GFF3/GTF
4. **Alignment (SAM/BAM)**

```
HWUSI-EAS525_0042_FC:6:23:10200:18582#0/1  16  1  10  40  35M
*  0  0  AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT
agafgfaffcfdf[fdcffcggggccfdffagggg  MD:Z:35 NH:i:1  HI:i:1  NM:i:0  SM:i:40
XQ:i:40 X2:i:0
HWUSI-EAS525_0042_FC:3:28:18734:20197#0/1  16  1  10  40  35M
*  0  0  AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT
hghhghhhhhhhhhhhhhhhhghhhhhghhfhhh  MD:Z:35 NH:i:1  HI:i:1  NM:i:0
SM:i:40 XQ:i:40 X2:i:0
HWUSI-EAS525_0042_FC:3:94:1587:14299#0/1  16  1  10  40  35M
*  0  0  AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT
hfhghhhhhhhhhhghhhhhhhhhhhhhhhhhhhg  MD:Z:35 NH:i:1  HI:i:1  NM:i:0
SM:i:40 XQ:i:40 X2:i:0
D3B4KKQ1:227:D0NE9ACXX:3:1305:14212:73591  0  1  11  40  51M
*  0  0  GCCAAAGATTGCATCAGTTCTGCTGCTATTTCCTCCTATCATTCTTTCTGA
CCCFFFFFFGFFHHJGIHHJJJFGGJJGIIIIIIGJJJJJJJJJIJJJJJE  MD:Z:51 NH:i:1  HI:i:1
NM:i:0  SM:i:40 XQ:i:40 X2:i:0
HWUSI-EAS525_0038_FC:5:35:11725:5663#0/1  16  1  11  40  35M
*  0  0  GCCAAAGATTGCATCAGTTCTGCTGCTATTTCCTC
hhehhhhhhhhhghghhhhhhhhhhhhhhhhhhh  MD:Z:35 NH:i:1  HI:i:1  NM:i:0
SM:i:40 XQ:i:40 X2:i:0
```

Running TOPHAT

- **Required files**
  - Reference genome. (FASTA file indexed with bowtie2-build software)
  - RNA-seq data files. (FASTQ files)

- **Optional files**
  - Annotation file * (GFF3 or GTF)

* If not provided, TOPHAT will try to predict splicing sites;

Running TOPHAT

```
tophat -G myAnnot.gff3 myGenome myData.fastq.gz
```

Bowtie2 indexed

Some extra parameters

- **--no-novel** : only using splicing sites in gff/gtf file
- **-N** : mismatches per read (default: 2)
- **-g**: max number of multi-hits (default: 20)
- **-p** : number of CPU cores (BioHPC lab general: 8)
- **-o**: output directory

* TOPHAT manual: http://ccb.jhu.edu/software/tophat/manual.shtml

Running TOPHAT

```
tophat -G myAnnot.gff3 myGenome myData.fastq.gz
```

Some extra parameters

- **--no-novel** : only using splicing sites in gff/gtf file
- **-N** : mismatches per read (default: 2)
- **-g**: max number of multi-hits (default: 20)
- **-p** : number of
- **-o**: output dire

- **In majority of the cases, it is recommended to use this parameter;**
- **Tophat is very slow without this option. You might want to use an alternative aligner like STAR.**

* TOPHAT manual: http://cc

What you get from TOPHAT

- **A BAM file per sample**
  File name: accepted_hits.bam

- **Alignment statistics**
  File name: align_summary.txt

Input: 9230201

Mapped: 7991618 (86.6% of input)

of these: 1772635 (22.2%) have multiple alignments (2210 have >20)

86.6% overall read alignment rate.

STAR is becoming more commonly used than TOPHAT

- Much faster;
- Requires more memory
  - 30G for human genome;
  - 10G for 500GB genome.

Index the genome:

```
STAR --runMode genomeGenerate \
--runThreadN 2 \
--genomeDir STARgenome \
--genomeFastaFiles testgenome.fa \
--sjdbGTFfile testgenome.gff3  \
--sjdbGTFtagExonParentTranscript Parent \
--sjdbOverhang 49
```

Map reads:

```
STAR --genomeDir STARgenome \
--runThreadN 2 \
--readFilesIn a.fastq.gz \
--readFilesCommand zcat \
--outFileNamePrefix a_ \
--outFilterMultimapNmax 1 \
--outReadsUnmapped unmapped_a  \
--outSAMtype BAM SortedByCoordinate
```

STAR is becoming more commonly used than TOPHAT

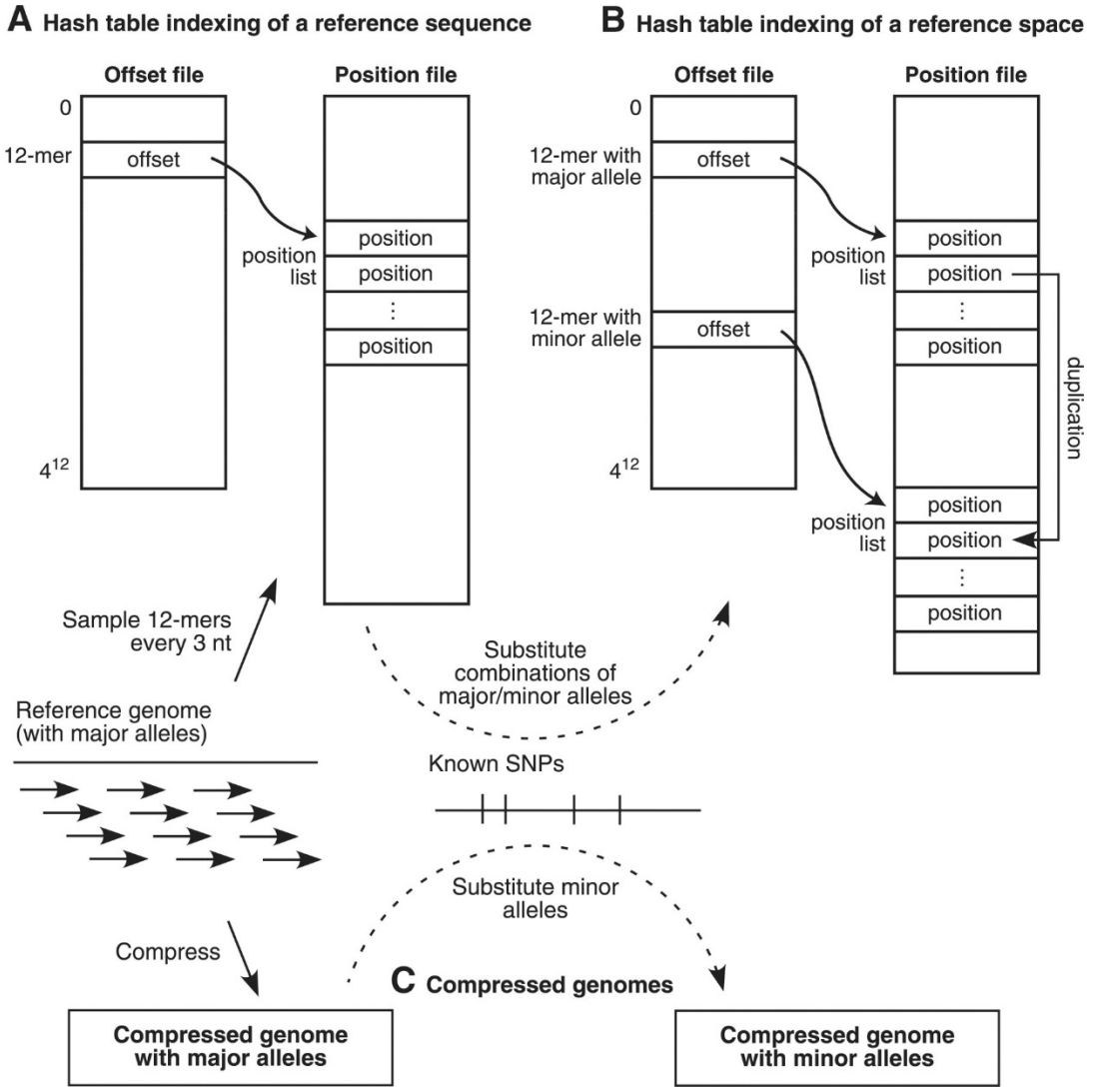
2-pass mapping of STAR is recommended for:

1. **Novel splicing junction discovery or genome annotation;**

2. **SNP calling;**

GSNAP: A SNP-tolerant alignment

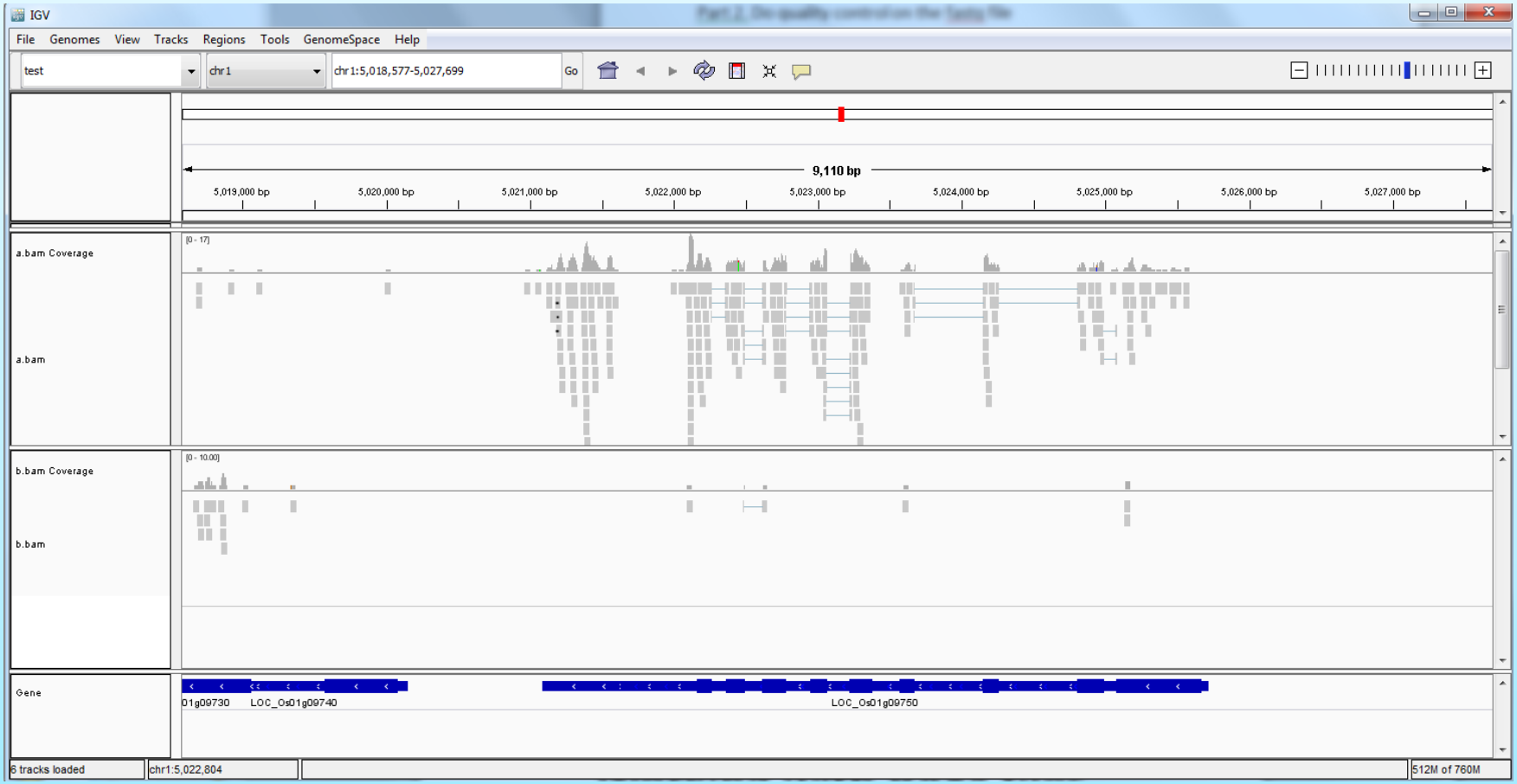
* It is critical for allele species expression analysis. GSNAP limit to single splicing per read.



Thomas D. Wu, and Serban Nacu Bioinformatics 2010;26:873-881

Visualizing BAM files with IGV

* Before using IGV, the BAM files need to be indexed with “samtools index”, which creates a .bai file.



Exercise 1

- Run TOPHAT to align RNA-seq reads to genome;
- Visualize TOPHAT results with IGV;
- Learn to use Linux shell script to create a pipeline;

QuantSeq 3′ mRNA sequencing for RNA quantification

Genomics Diversity Facility:
$23/library construction

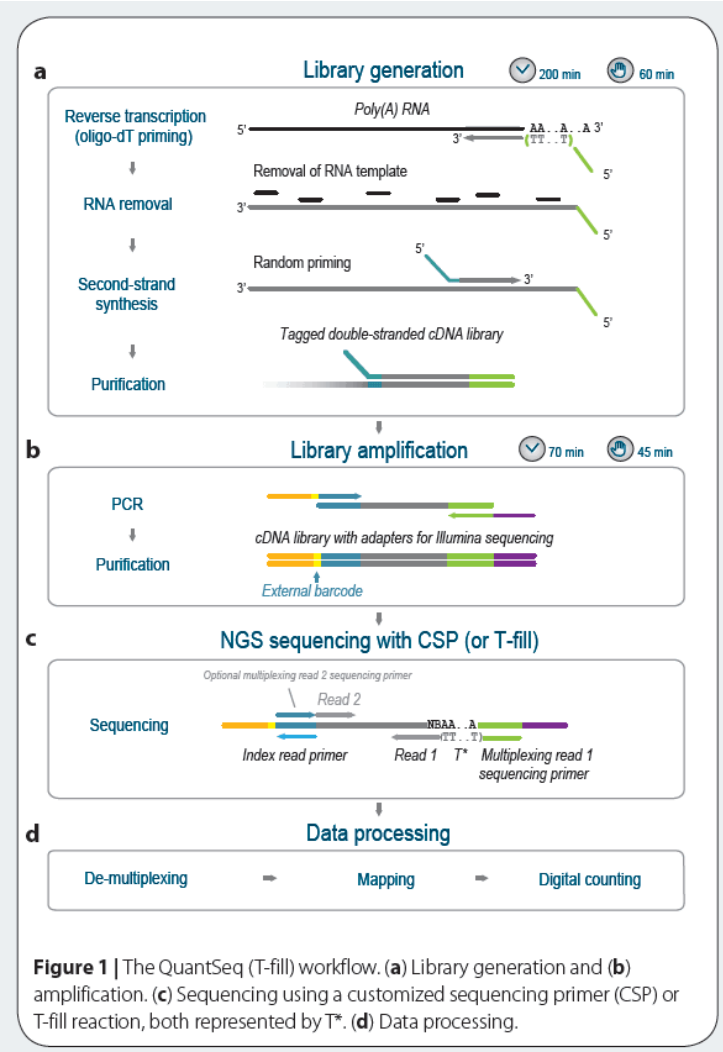


Figure 1 | The QuantSeq (T-fill) workflow. (**a**) Library generation and (**b**) amplification. (**c**) Sequencing using a customized sequencing primer (CSP) or T-fill reaction, both represented by T*. (**d**) Data processing.

BioHPC Lab office hours

Time: 1-3 pm, every Monday & Thursday

Office: 618 Rhodes Hall

Sign-up: https://cbsu.tc.cornell.edu/lab/office1.aspx

- General bioinformatics consultation/training is provided;
- Available throughout the year;