

RNA-seq Data Analysis

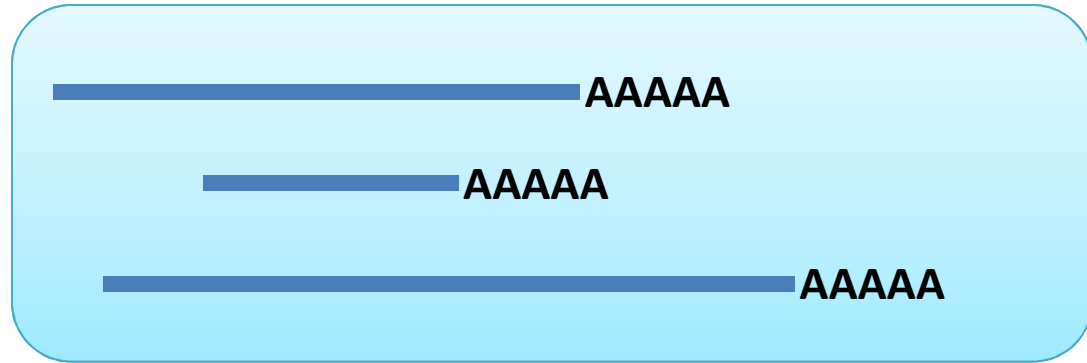
Qi Sun

Bioinformatics Facility
Biotechnology Resource Center
Cornell University

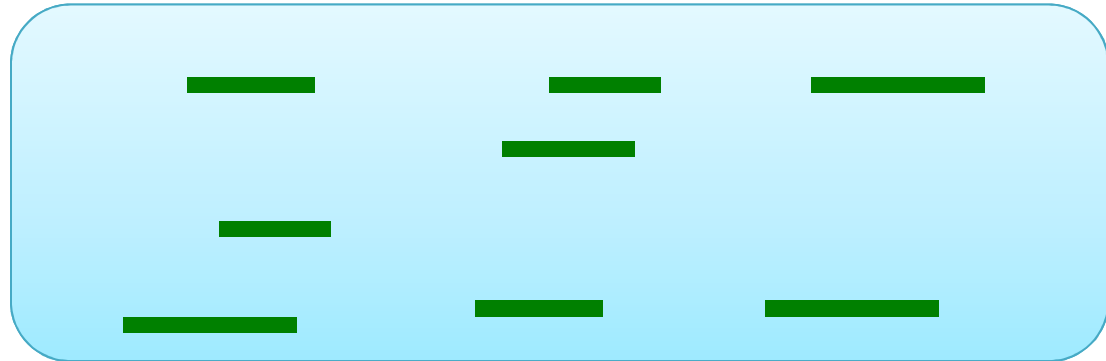
- **Lecture 1.** RNA-seq read alignment
- **Lecture 2.** Quantification, normalization & differentially expressed gene detection
- **Lecture 3.** Clustering; Function/Pathway Enrichment analysis

RNA-seq Experiment

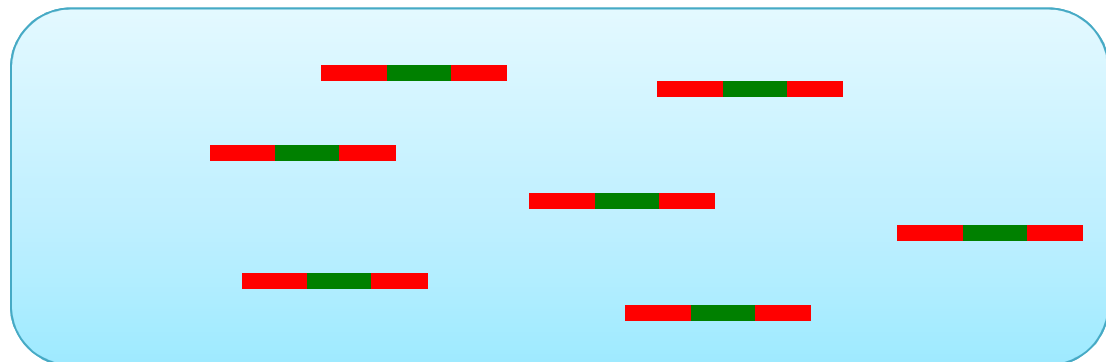
mRNA



cDNA Fragments
(100 to 500 bp)




Sequencing the
end(s) of cDNA
fragments




Some experimental aspects relevant to data analysis

Single End 

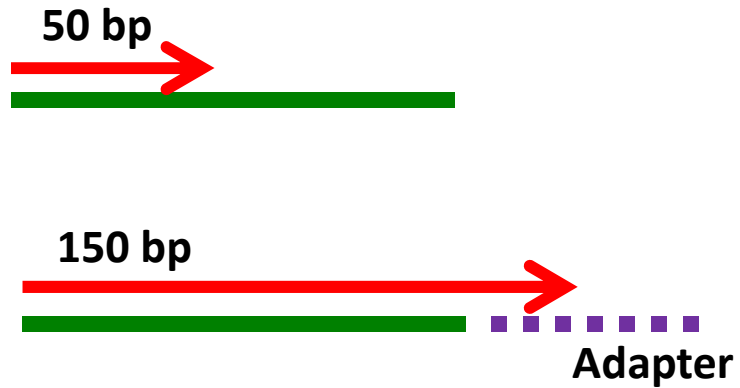
Paired End 

Stranded 

Unstranded 

Some experimental aspects relevant to data analysis

Long sequence reads



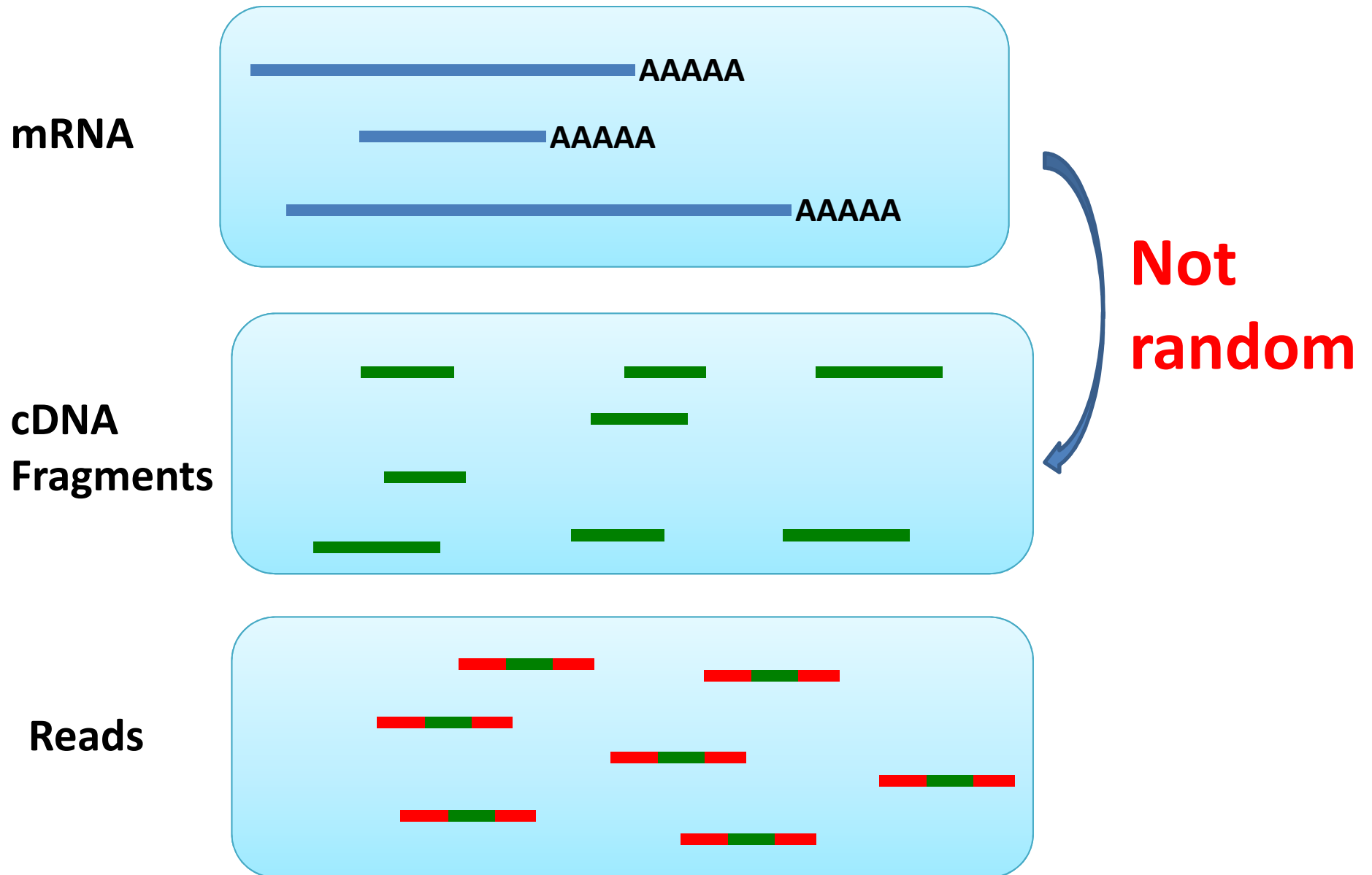
Experimental design with good reference genome

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

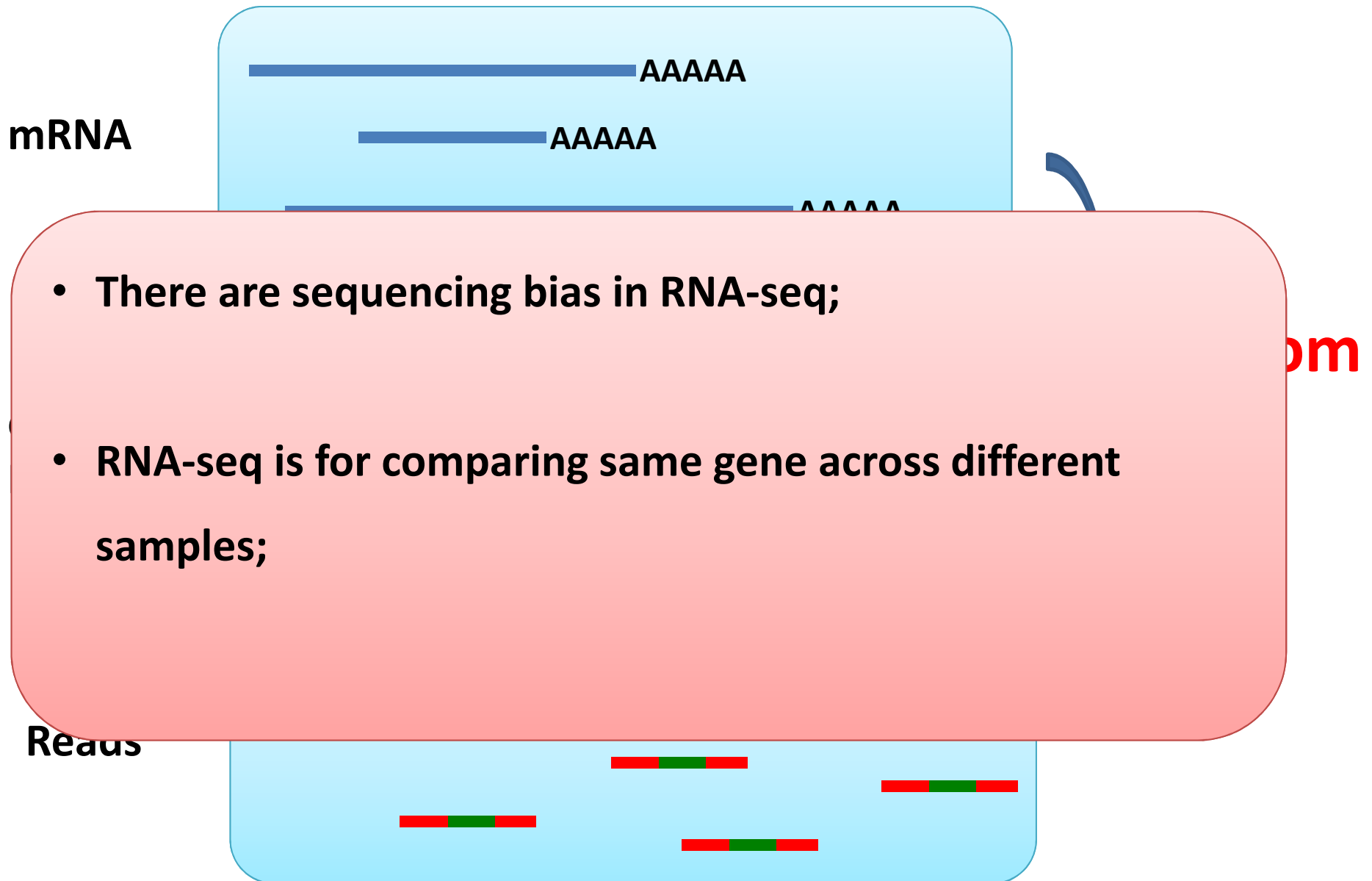
RNA-seq Experiments with NO reference genome

- Longer reads (150 bp or longer)
- Paired-end & stranded
- More reads (pooled from multiple samples)

Limitation of RNA-seq 1. Sequencing bias



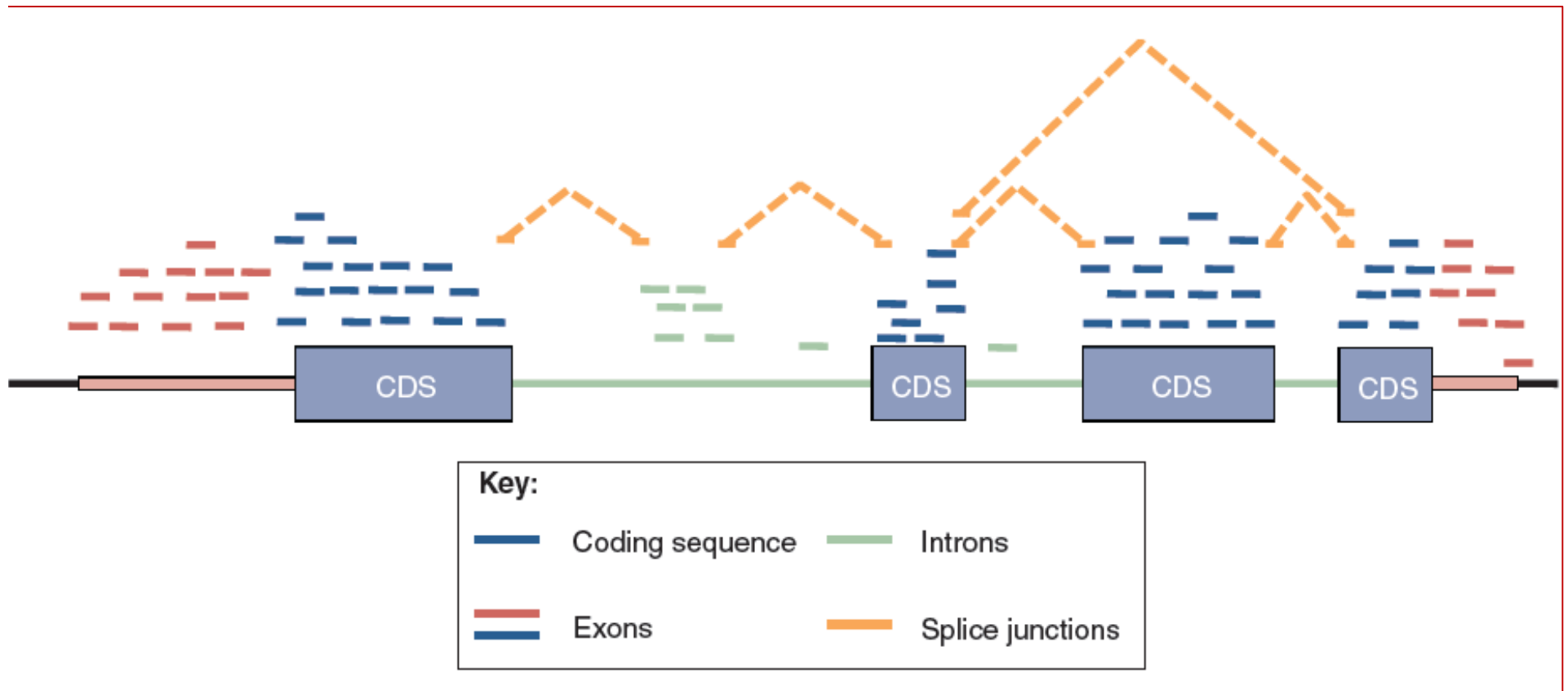
Limitation of RNA-seq 1. Sequencing bias



RNA-seq Data Analysis

Step 1. Map reads to gene

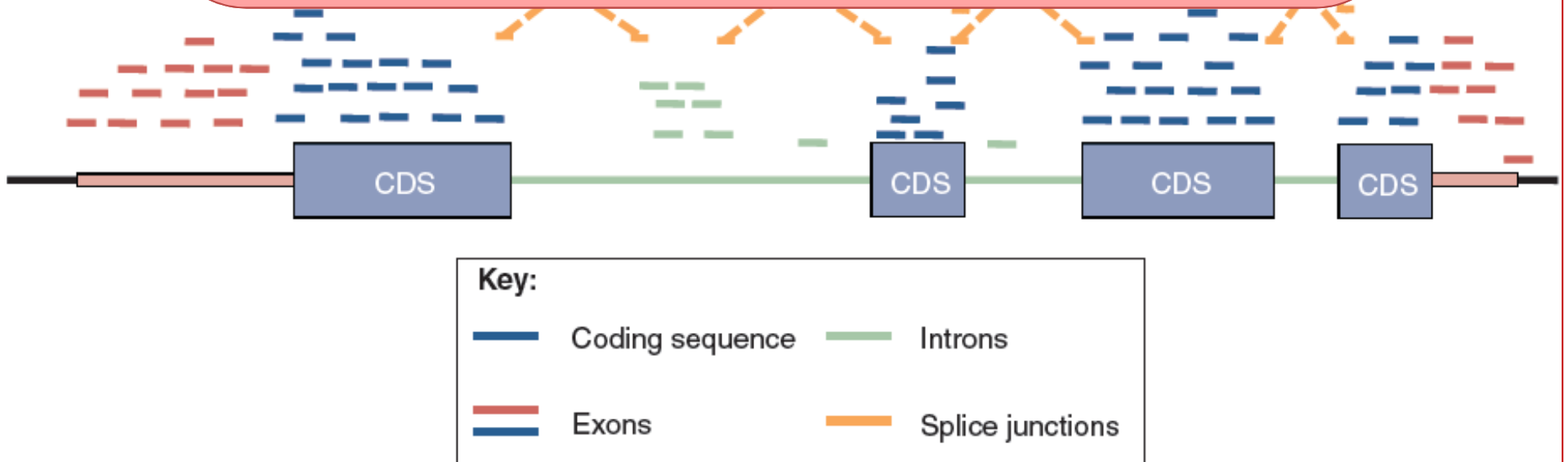
Step 2. Count reads per gene, estimate the transcript abundance



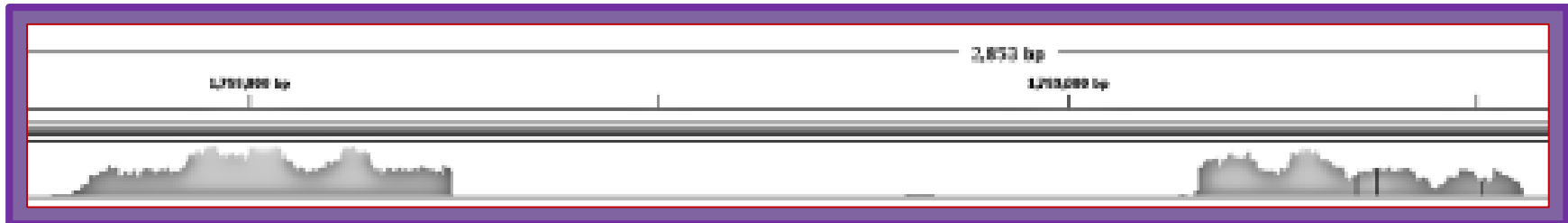
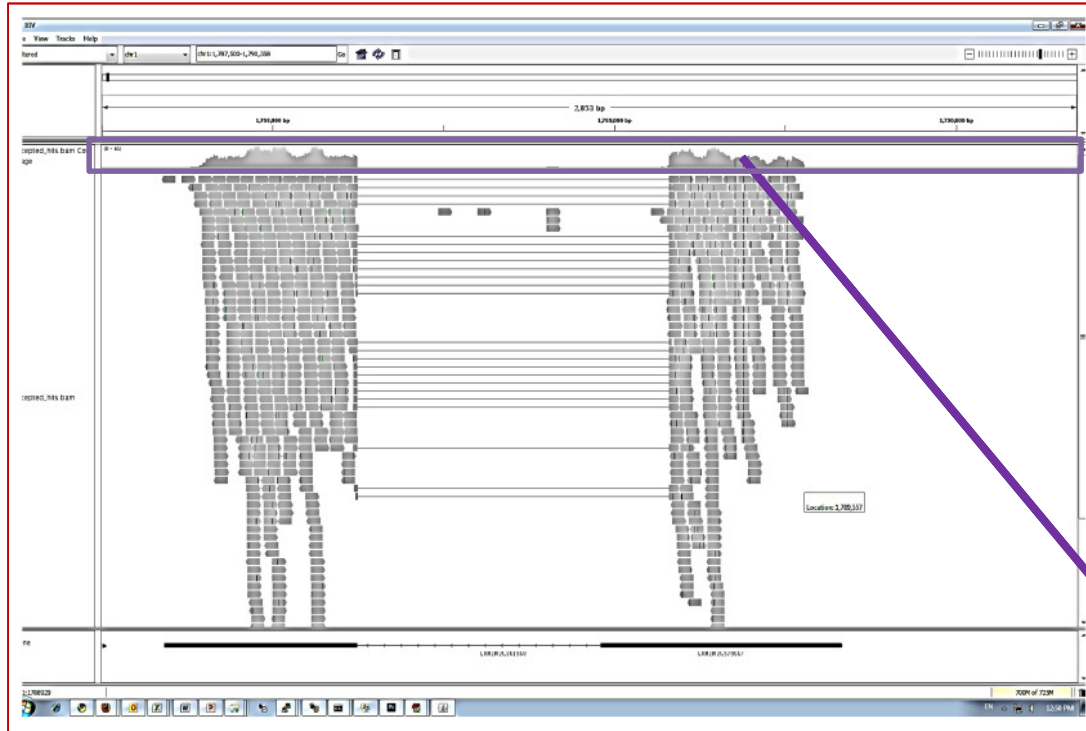
RNA-seq Data Analysis

Ambiguous reads placements

1. Between paralogous genes;
2. Between splicing isoforms;



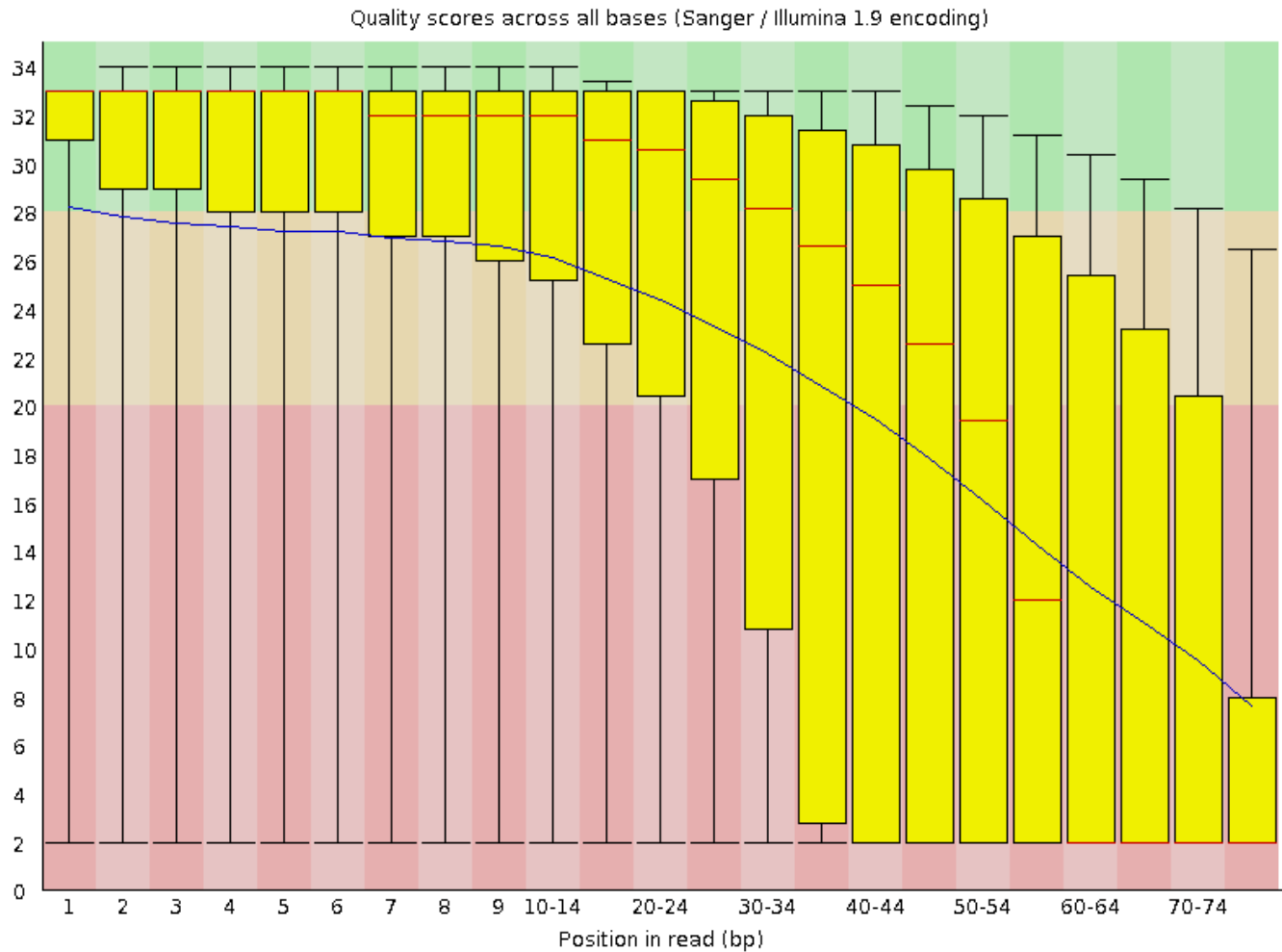
Read-depth are not even across the same gene



Data analysis procedures

Step 1. Quality Control (QC) using FASTQC Software

1. Sequencing quality score



Diagnose low quality data

1. Low quality reads & reads with adapters

- Trimming tools (FASTX, Trimmomatic, et al.)

2. Contamination (BLAST against Genbank)

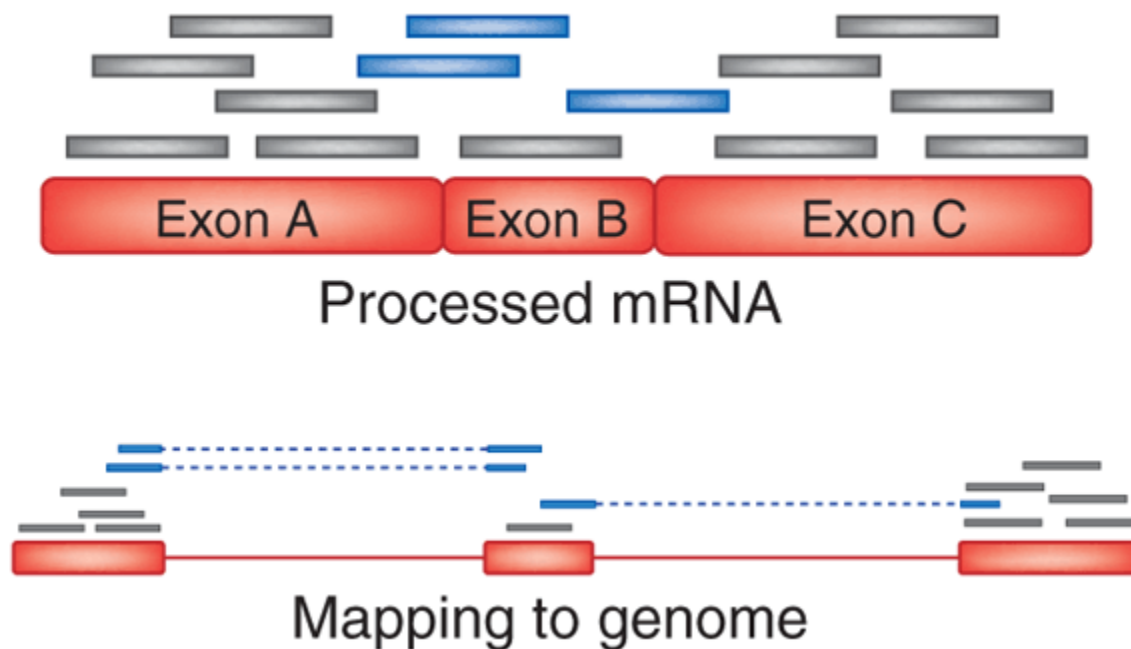
- Tool in bioHPC: fastq_species_detector

3. Correlation of biological replicates

- MDS plot

Step 2. Map reads to genome using TOPHAT Software

- Alignment of genomic sequencing vs RNA-seq



About the files

1. Reference genome (FASTA)

2. FASTQ

3. GFF3/GTF

4. SAM/BAM

```
>chr1
TTCTAGGTCTGCGATATTTCTGCCTATCCATTTTGTTAACTCTTCAATG
CATTCCACAAATACCTAAGTATTCTTTAATAATGGTGGTTTTTTTTTTTT
TTTGCATCTATGAAGTTTTTTCAAATTCTTTTTAAGTGACAAAACCTTGTA
CATGTGTATCGCTCAATATTTCTAGTCGACAGCACTGCTTTCGAGAATGT
AAACCGTGCACTCCCAGGAAAATGCAGACACAGCACGCCTCTTTGGGACC
GCGGTTTATACTTTTGAAGTGCTCGGAGCCCTTCTCCAGACCGTTCTCC
CACACCCCGCTCCAGGGTCTCTCCCGGAGTTACAAGCCTCGCTGTAGGCC
CCGGAACCCAACGCGGTGTCAGAGAAGTGGGGTCCCCTACGAGGGACCA
GGAGCTCCGGGCGGGCAGCAGCTGCGGAAGAGCCGCGCAGGCTTCCCAG
AACCCGGCAGGGGCGGGAAGACGCAGGAGTGGGGAGGCGGAACCGGGACC
CCGAGAGCCCGGGTCCCTGCGCCCCACAAGCCTTGCTTCCCTGCTAGG
GCCGGGCAAGGCCGGGTGCAGGGCGCGGCTCCAGGGAGGAAGCTCCGGGG
CGAGCCAAGACGCCTCCCGGGCGGTGCGGGCCCAGCGGCGGCGTTGCA
GTGGAGCCGGGCACCGGGCAGCGGCCGCGGAACACCAGCTTGGCGCAGGC
TTCTCGGTCAGGAACGGTCCCGGGCCTCCCGCCCGCTCCCTCCAGCCCC
TCCGGGTCCCCTACTTCGCCCCGCCAGGCCCCACGACCCTACTTCCCGC
GGCCCCGGACGCCTCCTCACCTGCGAGCCGCCCTCCCGAAGCTCCCGCC
GCCGCTTCCGCTCTGCCGGAGCCGCTGGGTCTAGCCCCGCCGCCCCAG
TCCGCCCGCGCTCCGGGTCTAACGCCCGCTCGCCCTCCACTGCGCC
CTCCCCGAGCGCGGCTCCAGGACCCCGTCGACCCGGAGCGCTGTCTGTC
GGGCCGAGTCGCGGGCCTGGGCACGGAACCTCACGCTCACTCCGAGCTCCC
GACGTGCACACGGCTCCCATGCGTTGTCTTCCGAGCGTCAGGCCGCCCT
ACCCGTGCTTTCTGCTCTGCAGACCCTTCTTAGACCTCCGTCCTTTGT
```


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
+
ccacaccaccccccccccc[[cccc_ccaccbbb_
@HWUSI-EAS525:2:1:14101:1126#0/1
GCCGGGACAGCGTGTGGTTGGCGCGCGGTCCCTC
+
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWUSI-EAS525:2:1:15408:1129#0/1
CGGCCTCATTCTTGCCAGGTTCTGGTCCAGCGAG
+
cghhchhgchehhdffccgdgh]gcchhcahWcea
@HWUSI-EAS525:2:1:15457:1127#0/1
CGGAGGCCCCGCTCCTCTCCCCCGCGCCGCGCC
+
^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWUSI-EAS525:2:1:15941:1125#0/1
TTGGGCCCTCCTGATTCATCGGTTCTGAAGGCTG
+
SUIF\_XYWW]VaOZZZ\V\bYbb_]ZXTZbbb_b
@HWUSI-EAS525:2:1:16426:1127#0/1
GCCCGTCTTAGAGGCTAGGGGACCTGCCCGCCGG
```

About the files

1. FASTA

2. RNA-seq data
(FASTQ)

3. GFF3/GTF

4. SAM/BAM

```
@HWUSTI-EAS525:2:1:13336:1129#0/1
GTTGGAGCCGGCGAGCGGGACAAGGCCCTTGTCCA
+
ccacaccaccccccccccc[[cccc_ccaccbbb_
@HWUSTI-EAS525:2:1:14101:1126#0/1
GCCGGGACAGCGTGTGGTTGGCGCGCGGTCCCTC
+
```

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

```
+
SUIF\_XYWW]VaOZZZ\V\bYbb_]ZXTZbbb_b
@HWUSTI-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";
chr12 unknown exon 96077274 96077487 . - .
gene_id "NTN4"; gene_name "NTN4"; p_id "P12149"; transcript_id
"NM_021229"; tss_id "TSS6395";
chr12 unknown CDS 96104219 96104407 . - 2
gene_id "NTN4"; gene_name "NTN4"; p_id "P12149"; transcript_id
"NM_021229"; tss_id "TSS6395";
chr12 unknown exon 96104219 96104407 . - .
gene_id "NTN4"; gene_name "NTN4"; p_id "P12149"; transcript_id
"NM_021229"; tss_id "TSS6395";
```

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 AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCTT
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 AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCTT
hghhghhhhhhhhhhhhhhhhhhhghhhhhghhfhhh 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 AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCTT
hfhghhhhhhhhhhhghhhhhhhhhhhhhhhhhhhhg 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 GCCAAAGATTGCATCAGTTCTGCTGCTATTTCTTCTCCTATCATTCTTTCTGA
CCCCFFFFFFGFFHHJGIHHJJJFGGJJGIIIIIGJJJJJJJJJJJE 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 GCCAAAGATTGCATCAGTTCTGCTGCTATTTCTTCT
hhehhhhhhhhghghhhhhhhhhhhhhhhhhhhhh 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
```

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>

What you get from TOPHAT

- **A BAM file per sample**
File name: accepted_hits.bam
- **Alignment statistics**
File name: align_summary.txt

Input: 9230201

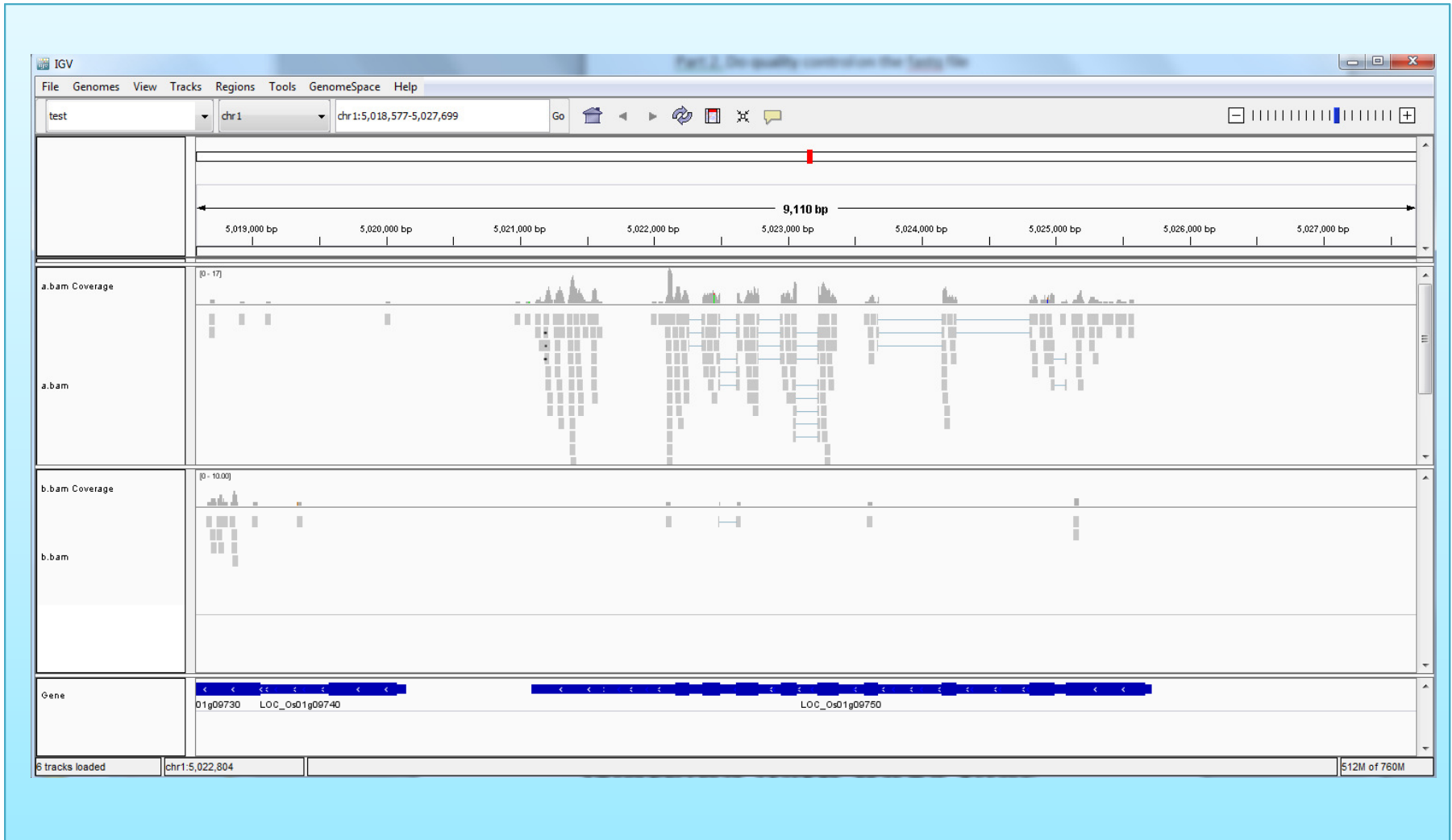
Mapped: 7991618 (86.6% of input)

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

86.6% overall read alignment rate.

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