RNA-seq Data Analysis

Lecture 1: Raw data -> read counts;

Lecture 2: Statistics of RNA-seq data analysis

Lecture 3: Function Enrichment Analysis

Qi Sun, Jeff Glaubitz

Bioinformatics Facility
Biotechnology Resource Center
Cornell University

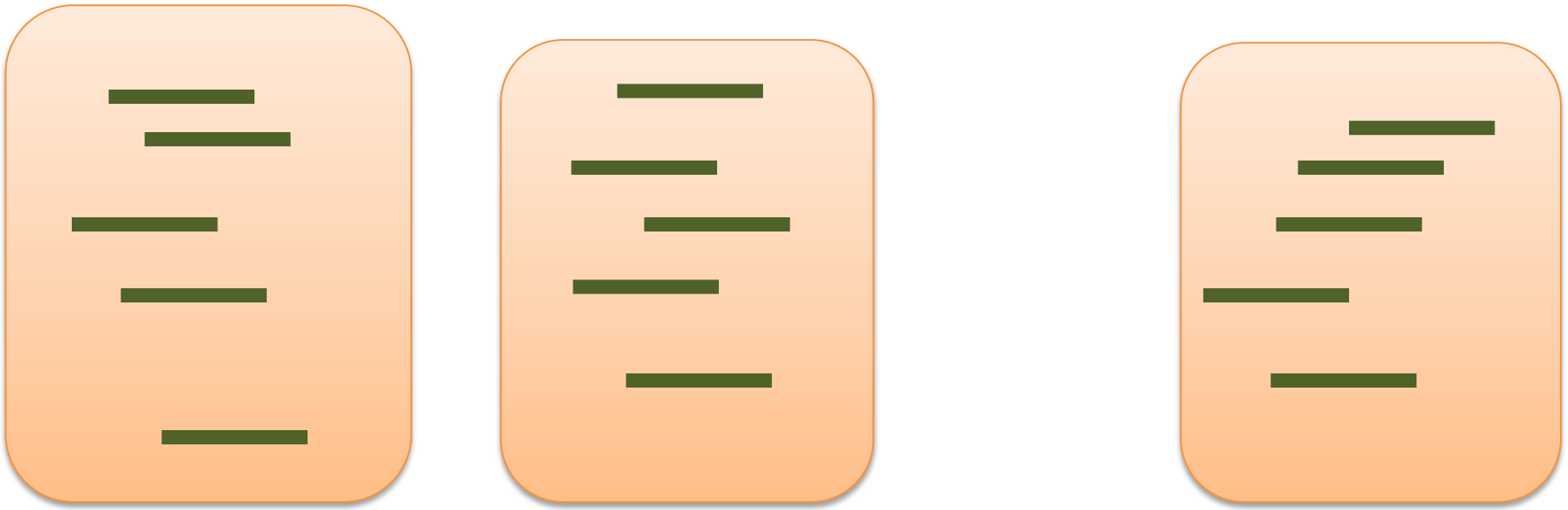
RNA-seq Experiment

Biological replicates

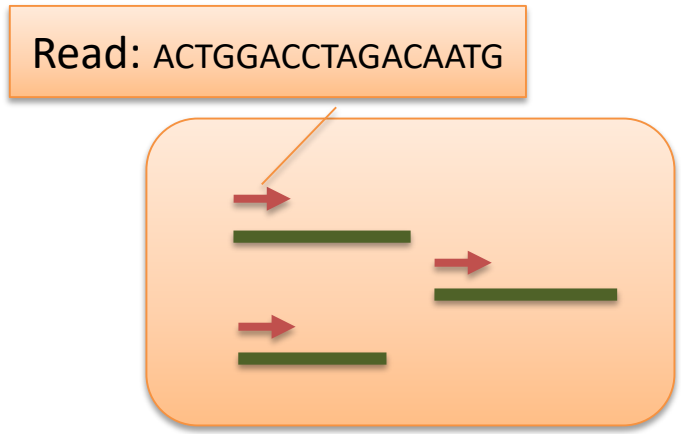
Sample 1 | Sample 2 | ... | Sample N

Extract RNA from samples

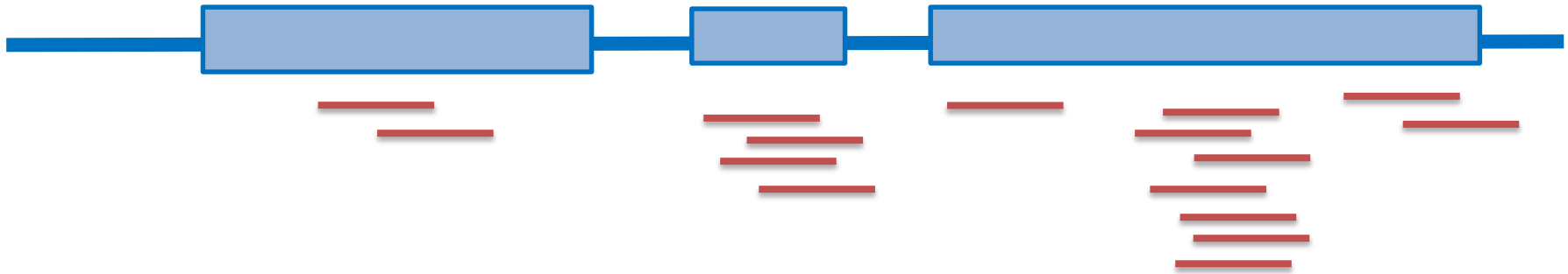
cDNA fragments
~100bp – 500bp range



Sequence the cDNA fragments



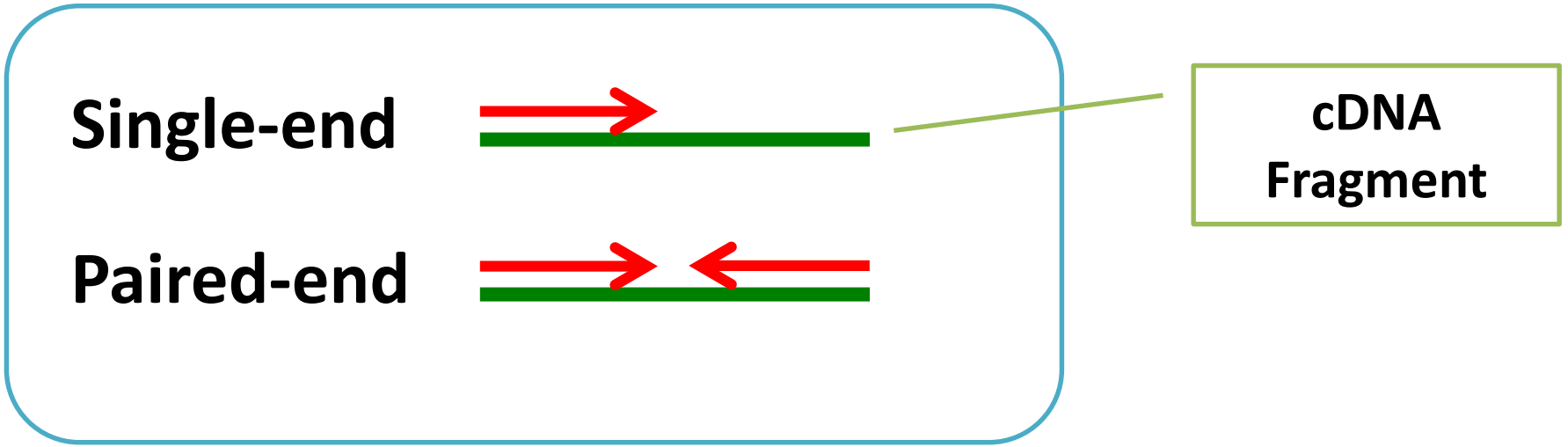
Map reads to Gene



Experimental Design

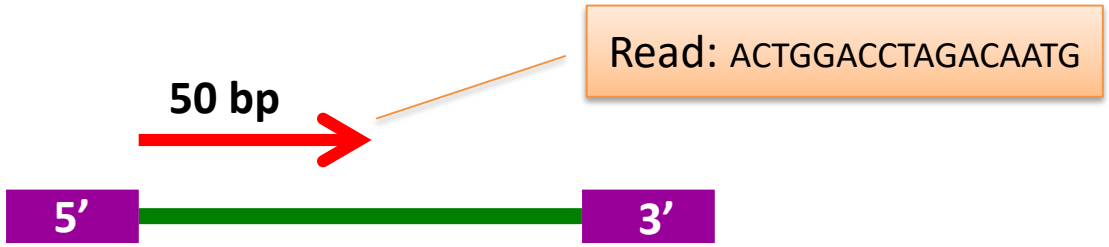
- **Single vs paired end;**
- **Read length (50bp, 75bp, …);**
- **Stranded vs non-stranded;**
- **Biological replicates;**

single-end vs paired-end



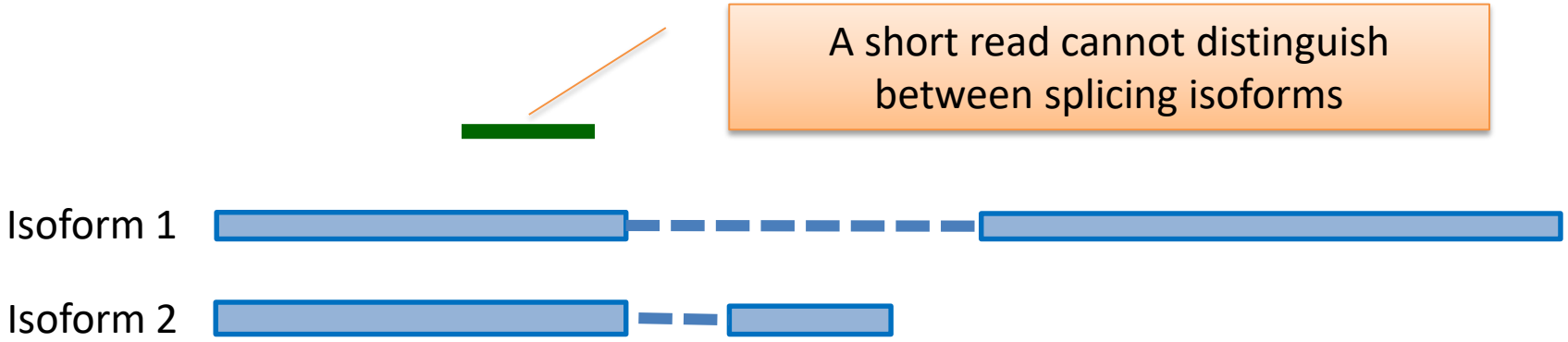
1. For expression quantification: single-end
2. For annotate genes on a new genome: paired-end

Read length (50 bp, 75 bp, …)

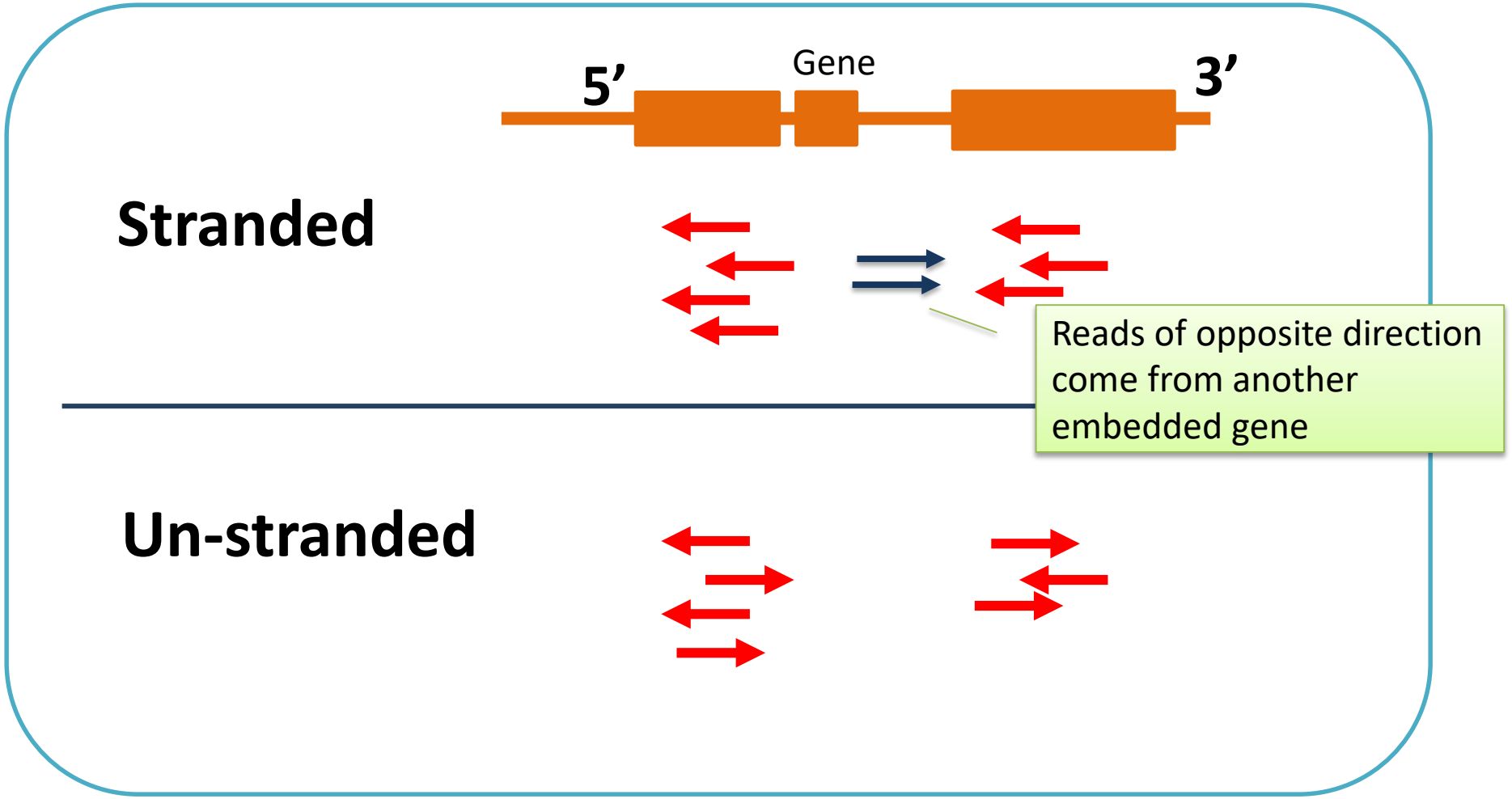


1. For gene expression level, 50 bp is good enough;

2. In some cases, longer reads are desired, eg.



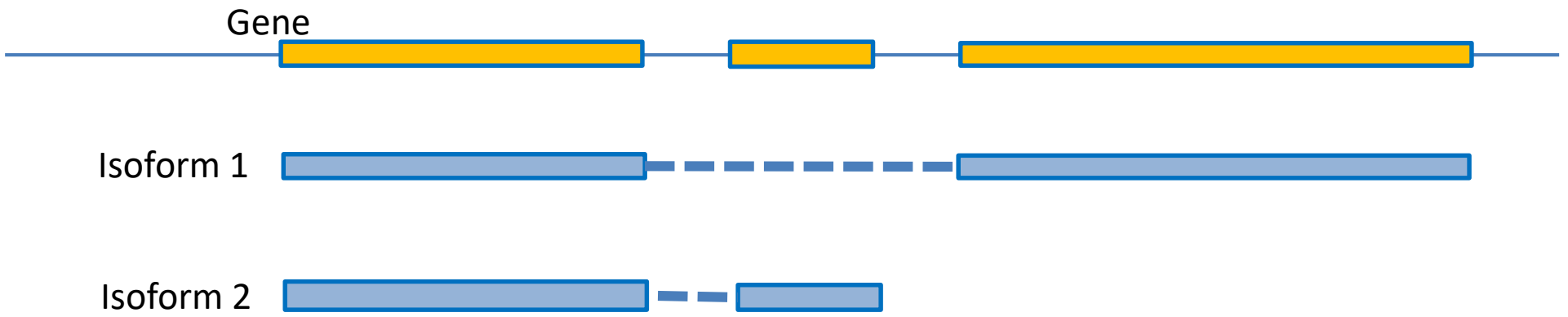
Stranded vs un-stranded



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

Data Analysis:
Gene level vs **Transcript Level** Quantification

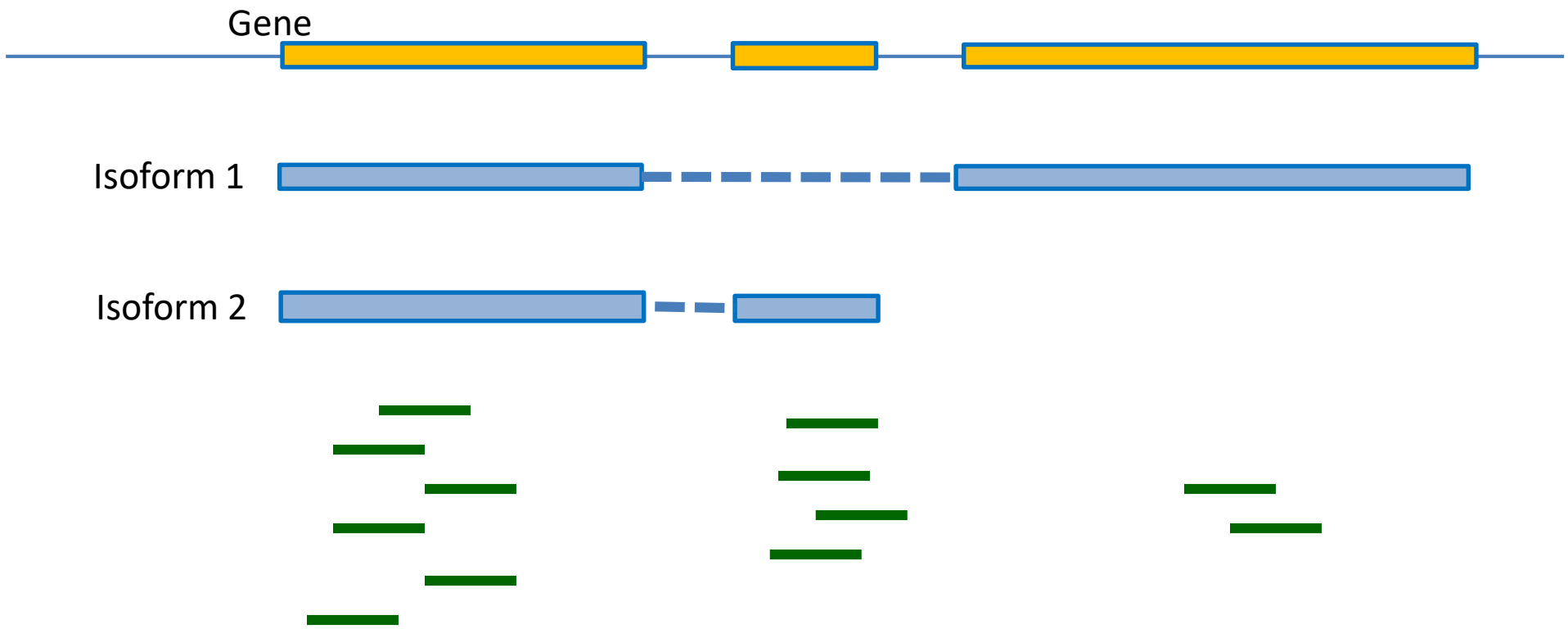


Gene level: Isoform 1 + Isoform 2

Transcript level: Isoform 1 and Isoform 2 counted separately

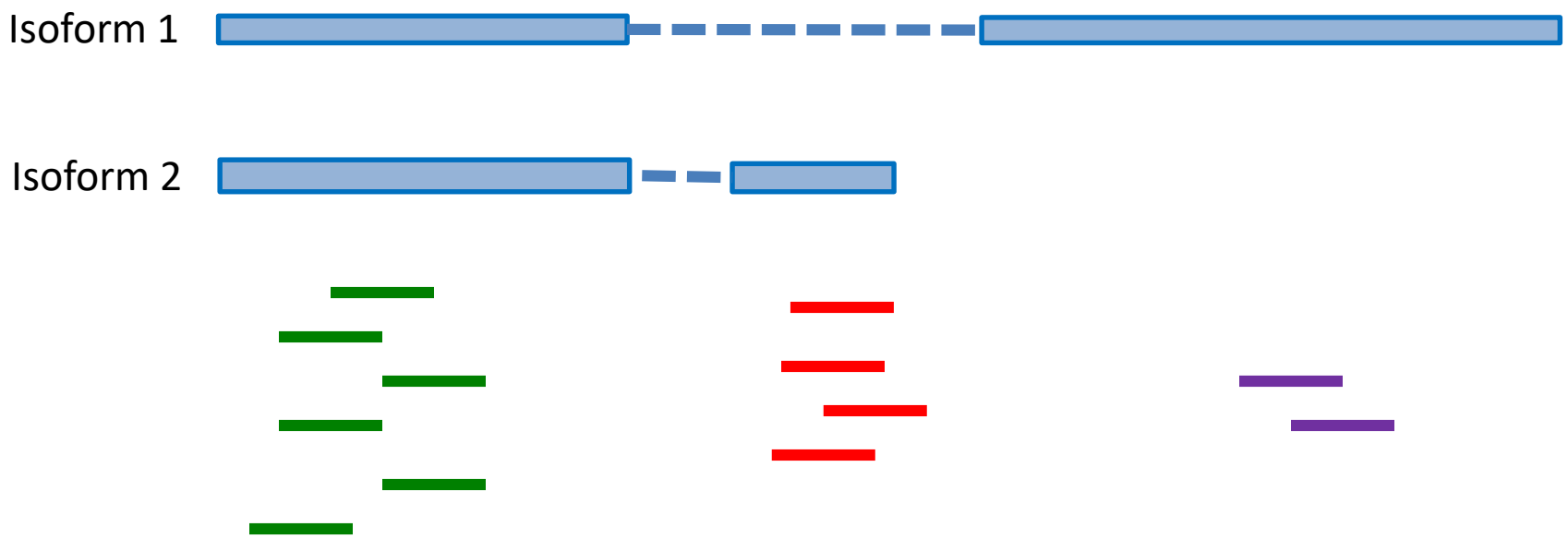
Gene level quantification:

- Count reads mapped to all exons of a gene

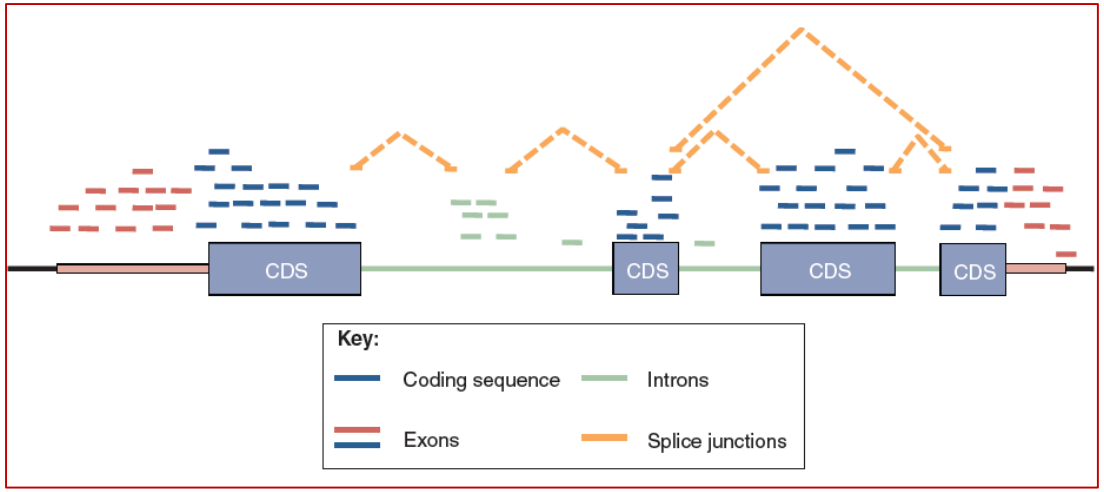


Transcript level quantification:

Count of ambiguous reads (in green) will be distributed to each isoform based on count of informative reads (in red and purple).



In this workshop:
Gene level quantification



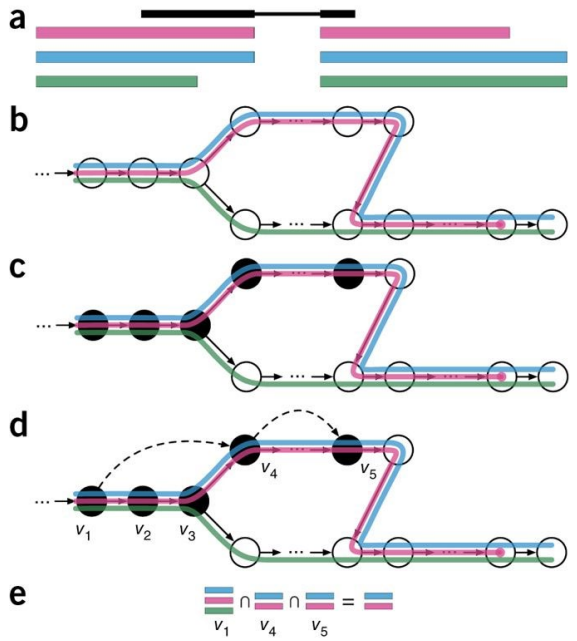
Three input files:

1. **Genome sequence;**
2. **RNA-seq data;**
3. **Gene annotation;**

Software for Transcript Level Quantification

Alignment based — **RSEM**

Alignment Free — Kallisto, Salmon



Alignment free approach of Kallisto:
Bray et al. 2016, *Nat. Biotech.*34, 525

Limitation of Alignment Free Methods

Kallisto Salmon

Pros:

- Very fast;
- Quantification of transcript isoforms;

Cons:

- Requires high quality gene annotation;
- Issues with lowly expressed genes.

File for RNA-seq Data Analysis

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
```

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";
...
```

Convert GFF3 to GTF file before you run STAR.

About the files

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

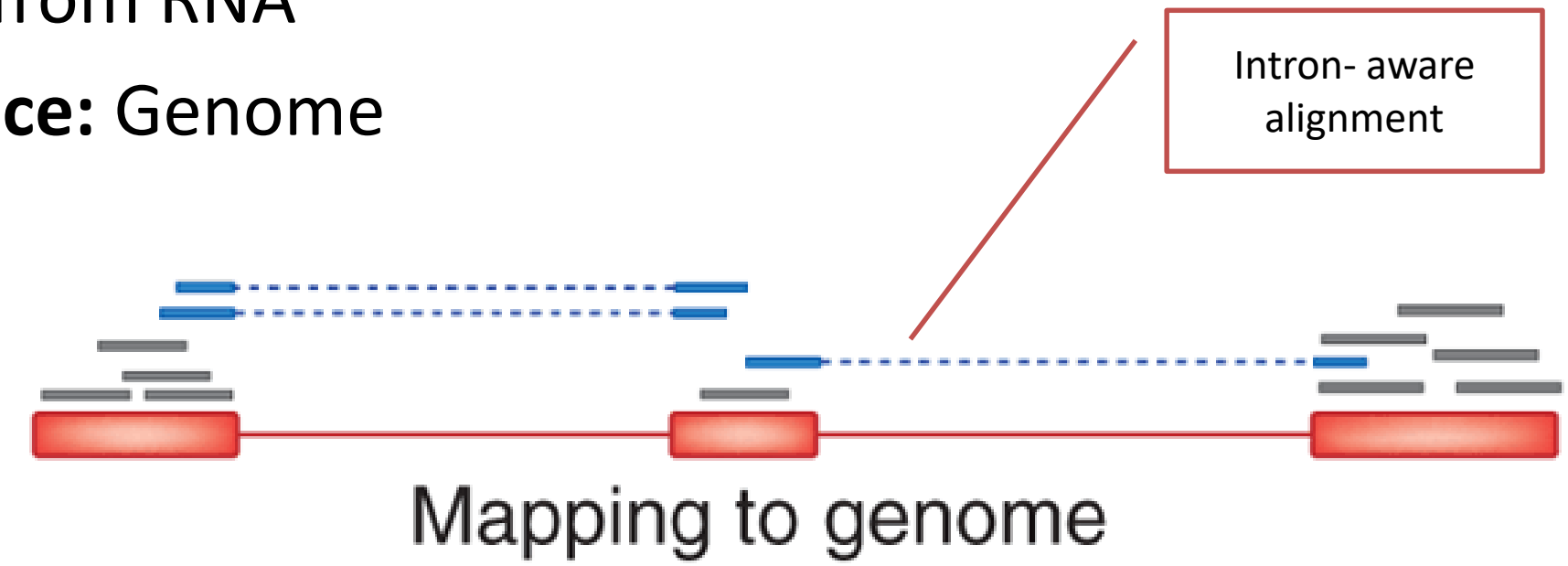
```
HWUSI-EAS525_0042_FC:6:23:10200:18582#0/1	16	1	10	40	35M	*	0	0	AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT	agafgfaffcfdf[fdcffcggggccfdffagggg	MD:Z:35 NH:i:1 HI:i:1 NM:i:0 SM:i:40 XQ:i:40 X2:i:0
HWUSI-EAS525_0042_FC:3:28:18734:20197#0/1	16	1	10	40	35M	*	0	0	AGCCAAAGATTGCATCAGTTCTGCTGCTATTTCCT	hghhghhhhhhhhhhhhhhhhghhhhhghhfhhh	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	hfhghhhhhhhhhghhhhhhhhhhhhhhhhhhg	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	CCCFFFFFFGFFHHJGIHHJJJFGGJJGIIIIIIGJJJJJJJJJIJJJJE	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	hhehhhhhhhhhghghhhhhhhhhhhhhhhhhh	MD:Z:35 NH:i:1 HI:i:1 NM:i:0 SM:i:40 XQ:i:40 X2:i:0
```

- SAM/BAM files contain information of read alignment to genome.
- BAM is a compressed format of SAM.

Map reads to genome: STAR

Read: from RNA

Reference: Genome



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

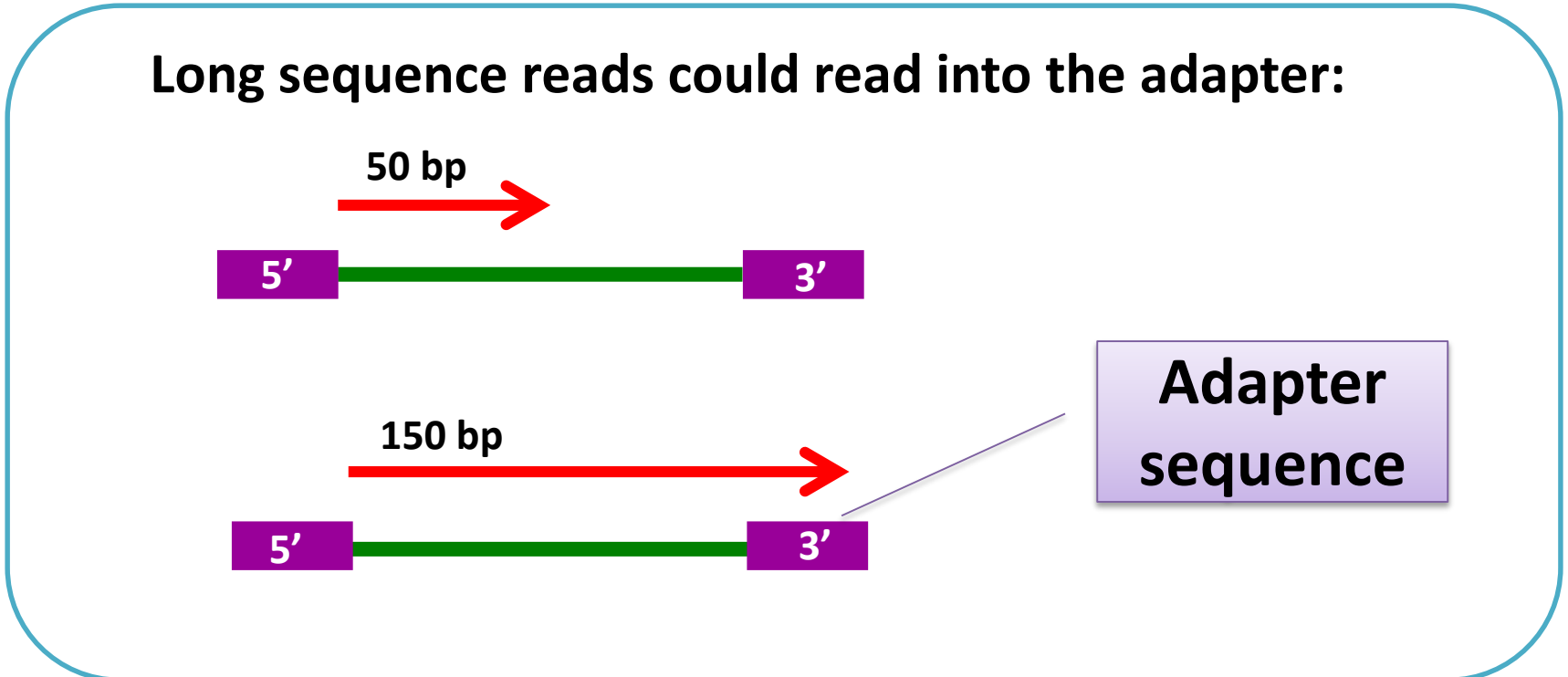
Issues in alignment

- Trim or not trim reads (remove adapters and low quality sequences);
- Novel Splicing junction discovery;
- Remove PCR duplicates?
- Remove rRNA tRNA?
- Ambiguity in alignment

Read Trimming: Low Quality and Adapter

- Not needed in most cases, especially for reads <=100bp;

- STAR soft clipping can remove some adapters;



Read Trimming: Low Quality and Adapter

Trimming software:

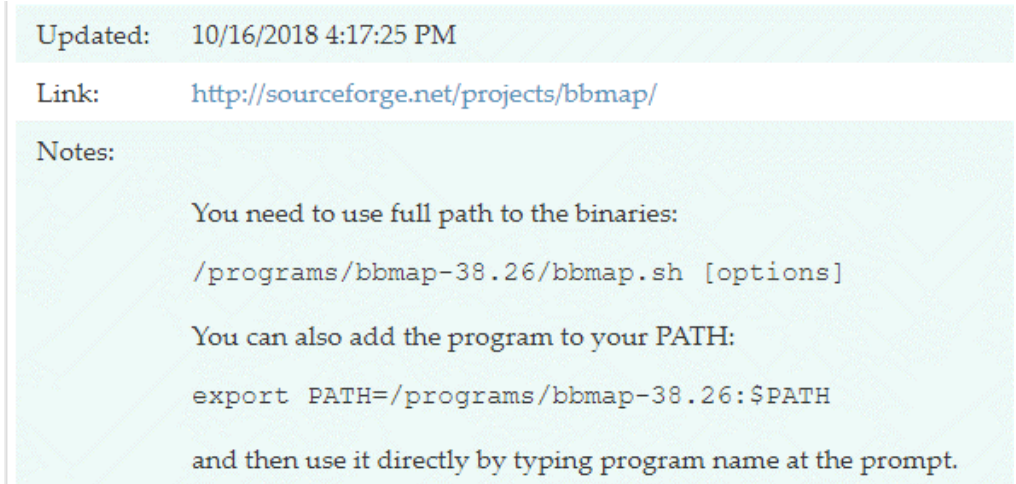
- BBDuk
- Trimmomatics
- Cutadapt

```
bbduk.sh in=reads.fq out=clean.fq
ref=adapters.fa t=8 ktrim=r k=23
mink=11 hdist=1 tpe tbo
```

If you want to run a software:

1. Read software manual;
2. Read instruction on BiopHPC

https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=248#c



Updated: 10/16/2018 4:17:25 PM

Link: http://sourceforge.net/projects/bbmap/

Notes:

You need to use full path to the binaries:

```
/programs/bbmap-38.26/bbmap.sh [options]
```

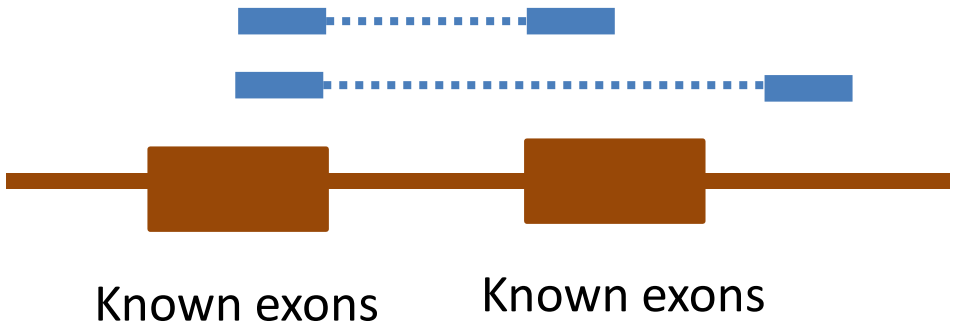
You can also add the program to your PATH:

```
export PATH=/programs/bbmap-38.26:$PATH
```

and then use it directly by typing program name at the prompt.

Novel splicing junction

- STAR always perform novel junction detection;
- Use two-pass if novel junctions are critical for your project;



Remove PCR Duplicates: Not needed;

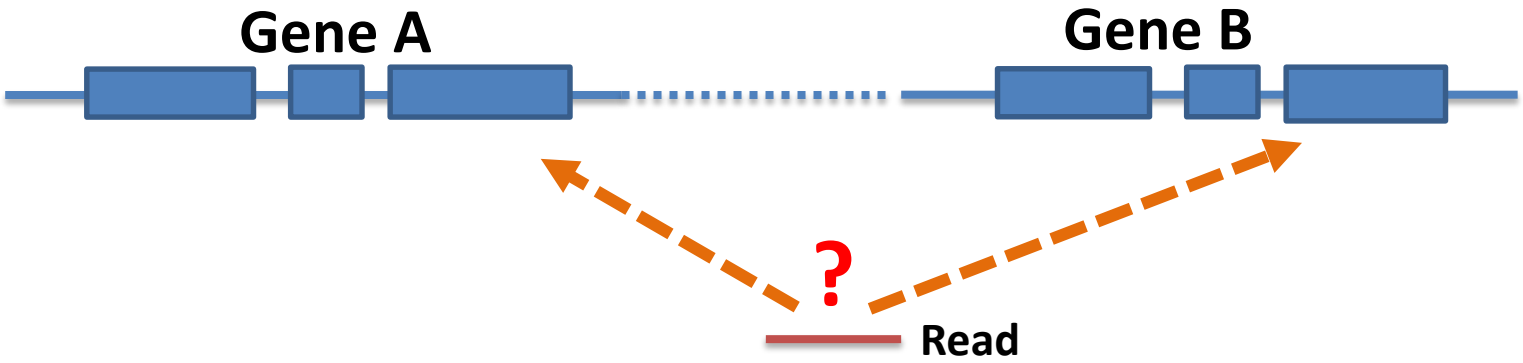
* Estimate library complexity based on detected gene numbers.

rRNA tRNA removal: Not needed.

They are mapped to genome, but not used in gene counts.

Short reads caused ambiguity in mapping

2. Ambiguity in paralogs



STAR and HTSeq

Discard multi-mapped reads

* This might not be desirable for some genes, e.g. duplicated genes in reference

Diagnose low mapping rate (<50% reads mapped)

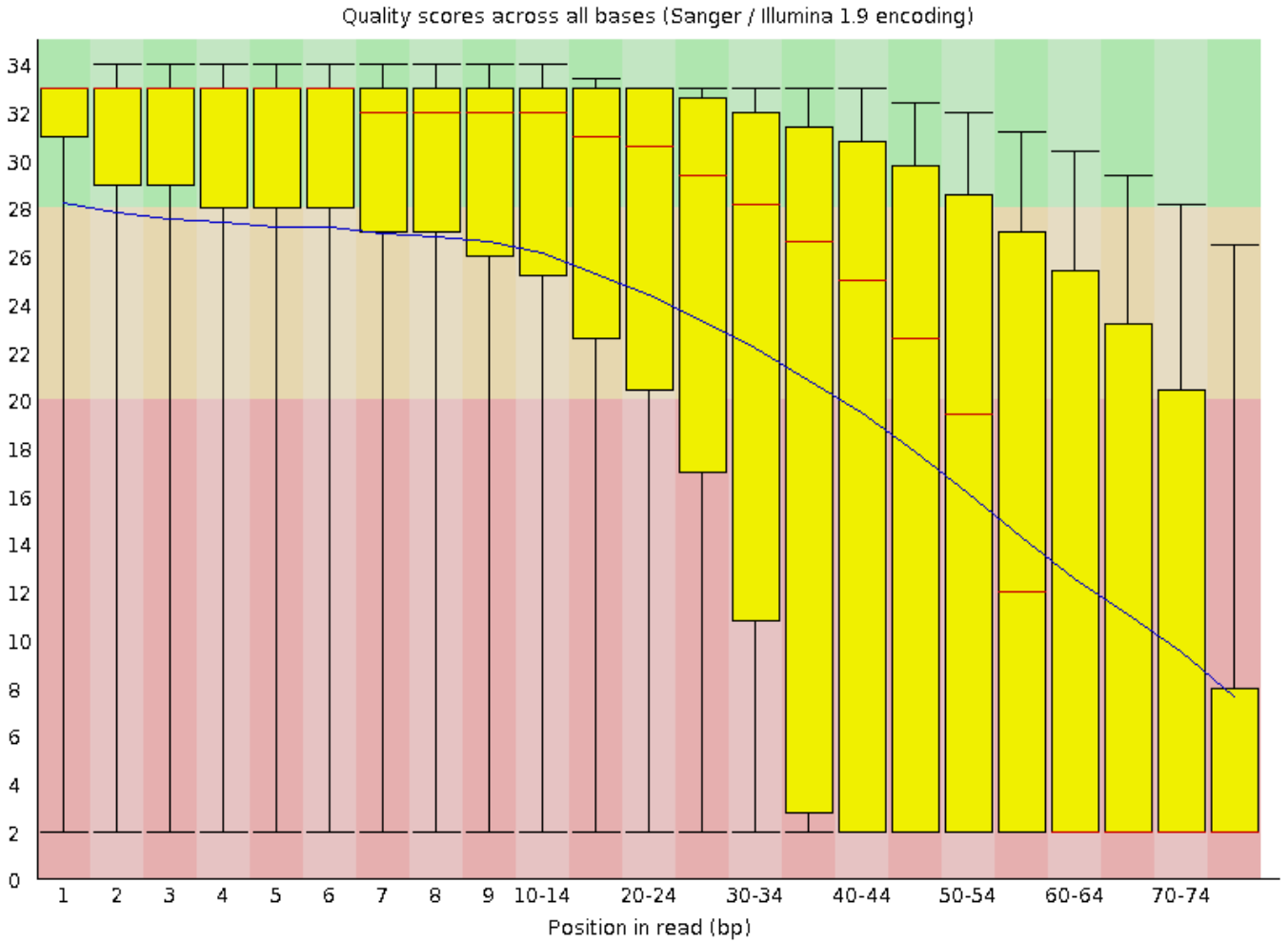
1. Low quality reads or reads with adapters *
   - Trimming tools (FASTX, Trimmomatic, et al.)

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

Step 1. Quality Control (QC) using FASTQC Software

1. Sequencing quality score



A BioHPC tool for detecting read contamination

fastq_species_detector

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

STAR is one of the most commonly used reads mapping software

- **Fast;**
- **Requires a computer of lots of memory**
  - **30G for human genome;**
  - **10G for 500 MB genome.**
- Use HISAT2 if you do not have access to a computer with large RAM.

Index the genome:

```
STAR --runMode genomeGenerate \
--runThreadN 2 \
--genomeDir STARgenome \
--genomeFastaFiles testgenome.fa \
--sjdbGTFfile testgenome.gtf  \
--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
```

First step of running STAR: Index the genome:

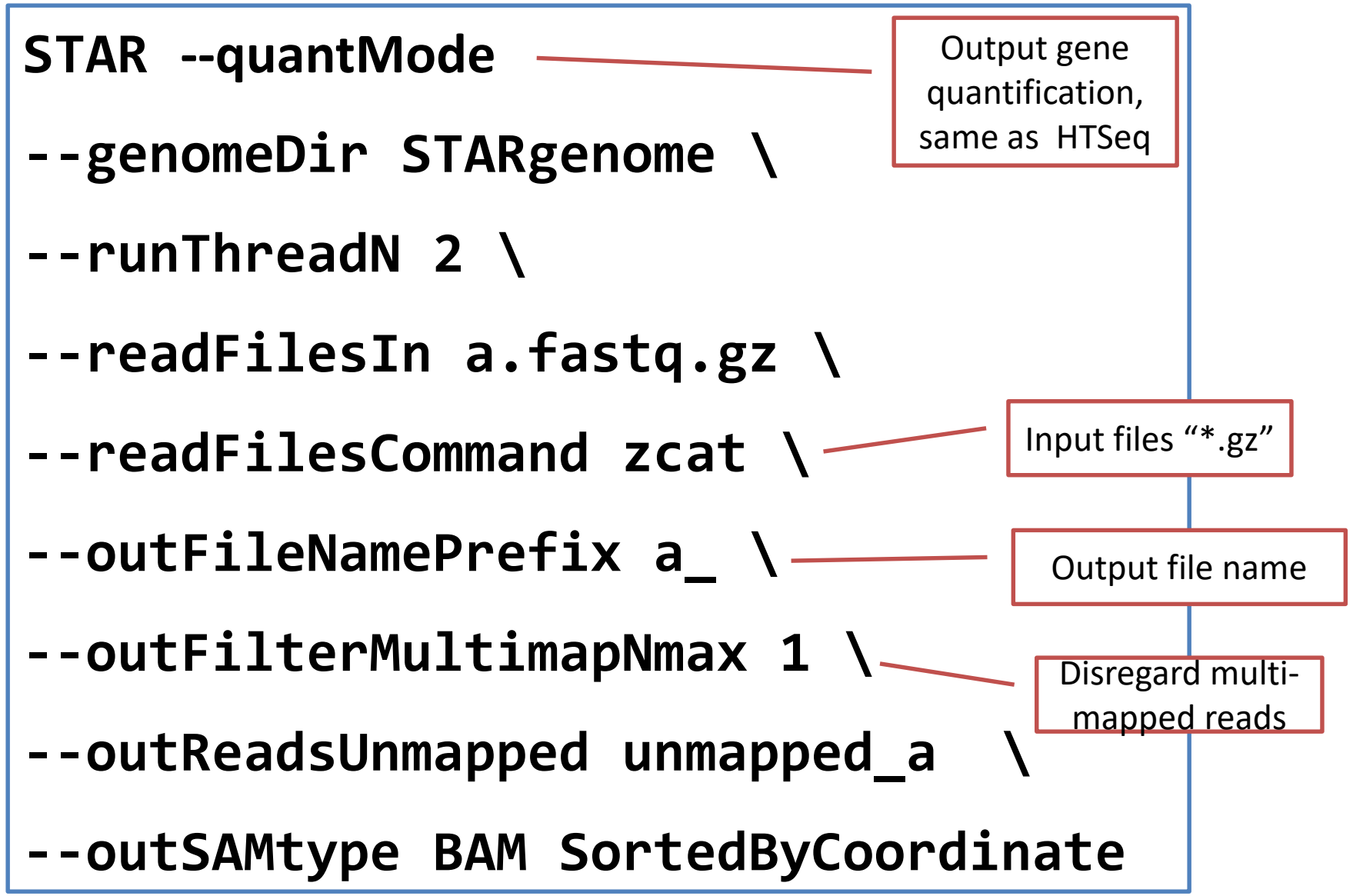
```
STAR --runMode genomeGenerate \
--runThreadN 2 \
--genomeDir STARgenome \
--genomeFastaFiles testgenome.fa \
--sjdbGTFfile testgenome.gtf  \
--sjdbOverhang 49
```

Use GTF, not gff3.
The STAR manual offers an option to use gff3, but in our experience, it is better to convert gff3 to gtf first with “gffread” tool.

Read length - 1

Second step of running STAR: Map/count reads

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



Setting parameters

```
STAR --quantMode GeneCounts --genomeDir genomedb  --
runThreadN 2 --outFilterMismatchNmax 2 --readFilesIn
WTa.fastq.gz --readFilesCommand zcat --outFileNamePrefix
WTa --outFilterMultimapNmax 1 --outSAMtype BAM
SortedByCoordinate
```

Some other parameters:

--outFilterMismatchNmax : max number of mismatch (Default 10)

--outFilterMultimapNmax 1: do not output multi-mapped reads

Manual:https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf

Using a shell script to batch process many samples

Using a text editor to put all commands in one file:

```
STAR --quantMode GeneCounts --genomeDir genome --runThreadN 2 --readFilesIn
ERR458493.fastq.gz --readFilesCommand zcat --outFileNamePrefix wt1_ --
outFilterMultimapNmax 1 --outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate

STAR --quantMode GeneCounts --genomeDir genome --runThreadN 2 --readFilesIn
ERR458494.fastq.gz --readFilesCommand zcat --outFileNamePrefix wt2_ --
outFilterMultimapNmax 1 --outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate

STAR --quantMode GeneCounts --genomeDir genome --runThreadN 2 --readFilesIn
ERR458495.fastq.gz --readFilesCommand zcat --outFileNamePrefix wt3_ --
outFilterMultimapNmax 1 --outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate

STAR --quantMode GeneCounts --genomeDir genome --runThreadN 2 --readFilesIn
ERR458500.fastq.gz --readFilesCommand zcat --outFileNamePrefix mu1_ --
outFilterMultimapNmax 1 --outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate

STAR --quantMode GeneCounts --genomeDir genome --runThreadN 2 --readFilesIn
ERR458501.fastq.gz --readFilesCommand zcat --outFileNamePrefix mu2_ --
outFilterMultimapNmax 1 --outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate

STAR --quantMode GeneCounts --genomeDir genome --runThreadN 2 --readFilesIn
ERR458502.fastq.gz --readFilesCommand zcat --outFileNamePrefix mu3_ --
outFilterMultimapNmax 1 --outFilterMismatchNmax 2 --outSAMtype BAM SortedByCoordinate
```

Run jobs in serial:

sh runSTAR.sh

Run jobs in parallel:

parallel -j 4 < runSTAR.sh

Windows: Notepad++ ; Mac: BBEdit

Parallelization (run in “screen”)

```
parallel -j 4 < ~/runSTAR.sh   >& mylog &
```

Monitoring a job
top
top -o %MEM
ps -fu myUserID
ps -fu myUserID | grep STAR

Kill a job:
kill PID ## you need to kill both shell script and STAR alignment that is still running
kill -9 PID
killall userID

when do parallelization:

```
parallel -j 10 < ~/runSTAR.sh >& mylog &
```

CPU core: number of jobs x number of threads per job not exceed total CPU cores;

Memory: number of jobs x memory per job not exceed total RAM;
(STAR has option to share memory between jobs)

Hard drive: Not too many jobs simultaneously (-j setting <=10 for large memory machine)

e.g. BioHPC large memory gen2: 80-core cpu, 512 GB RAM.
Parallele (-j 10), STAR (runThreadN 8) , use shared memory,

Three Output files from STAR

File 1: *Log.final.out

```
                 Number of input reads |       13547152
             Average input read length |       49
                           UNIQUE READS:
            Uniquely mapped reads number |       12970876
                 Uniquely mapped reads % |       95.75%
                   Average mapped length |       49.32
                Number of splices: Total |       1891468
    Number of splices: Annotated (sjdb) |       1882547
                Number of splices: GT/AG |       1873713
                Number of splices: GC/AG |       15843
                Number of splices: AT/AC |       943
       Number of splices: Non-canonical |       969
```

Three Output files from STAR

File 2: *ReadsPerGene.out.tab

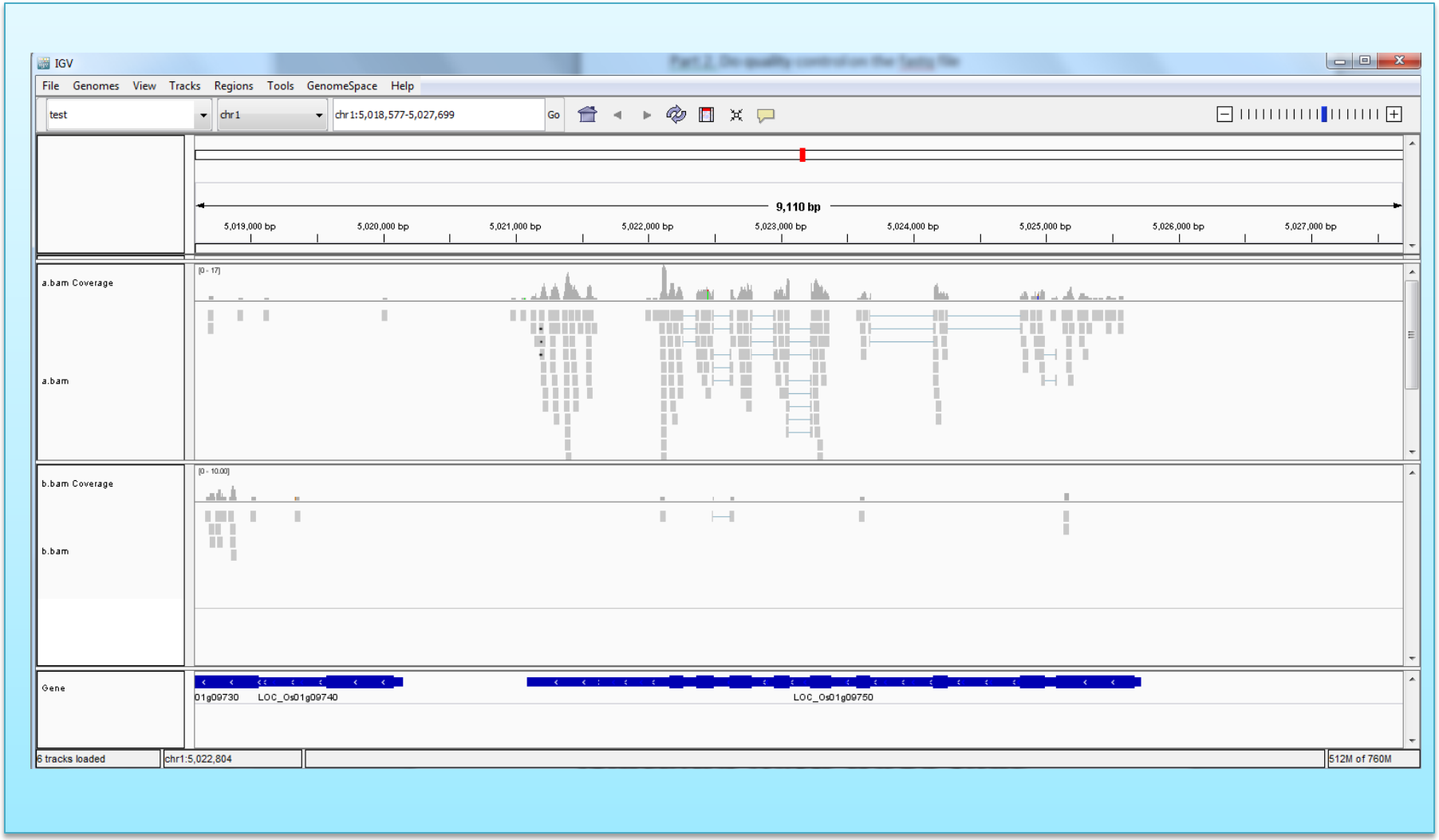
N_unmapped	1860780	1860780	1860780
N_multimapping	0	0	0
N_noFeature	258263	13241682	375703
N_ambiguous	461631	9210	17159
gene:AT1G01010	50	1	49
gene:AT1G01020	149	1	148
gene:AT1G03987	0	0	0
gene:AT1G01030	77	0	77
gene:AT1G01040	583	41	669
…			

column 1: gene ID
column 2: counts for unstranded RNA-seq
column 3: counts for the 1st read strand aligned with RNA (htseq-count option -s yes)
column 4: counts for the 2nd read strand aligned with RNA (htseq-count option -s reverse)

File 3. *.bam

Visualizing BAM files with IGV

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



STAR output: one file per sample

STAR output:

Sample1

N_unmapped	1860780	1860780	1860780
N_multimapping	0	0	0
N_noFeature	258263	13241682	375703
N_ambiguous	461631	9210	17159
gene:AT1G01010	50	1	49
gene:AT1G01020	149	1	148
gene:AT1G03987	0	0	0
gene:AT1G01030	77	0	77
gene:AT1G01040	583	41	669
…			

Sample2

N_unmapped	1637879	1637879	1637879
N_multimapping	0	0	0
N_noFeature	224759	11828019	354396
N_ambiguous	445882	8133	14924
gene:AT1G01010	57	0	57
gene:AT1G01020	174	2	172
gene:AT1G03987	1	1	0
gene:AT1G01030	91	3	88
gene:AT1G01040	516	27	594
gene:AT1G03993	0	81	2

…

Downstream analysis: all samples in one file

Merged file:

gene	Sample1	Sample2	Sample3	Sample4
AT1G01010	57	49	36	40
AT1G01020	172	148	197	187
AT1G03987	0	0	0	0
AT1G01030	88	77	74	101
AT1G01040	594	669	504	633
AT1G03993	2	1	0	0
…	…	…	…	…

```
paste file1 file2 file3 | \
cut -f1,4,8,12 | \
tail -n +5 \
> tmpfile
```

3’ mRNA sequencing for RNA quantification (QuantSeq)

(Quantseq pipeline: https://www.lexogen.com/quantseq-data-analysis/)

1. Trim adapter, poly-A and poly-G;

/programs/bbmap-38.45/bbduk.sh in=input.fastq.gz out=output.fastq.gz ref=/programs/bbmap-38.45/resources/polyA.fa.gz,/programs/bbmap-38.45/resources/truseq_rna.fa.gz k=13 ktrim=r useshortkmers=t mink=5 qtrim=r trimq=10 minlength=20

2. Alignment with STAR;

Increase the mismatch cutoff, e.g. “--outFilterMismatchNmax 10 “

3. Quantification using forward strand counts;

4. If annotation is poor, you might need to extend 3’ UTR (use the extend_gtf.py tool on BioHPC)

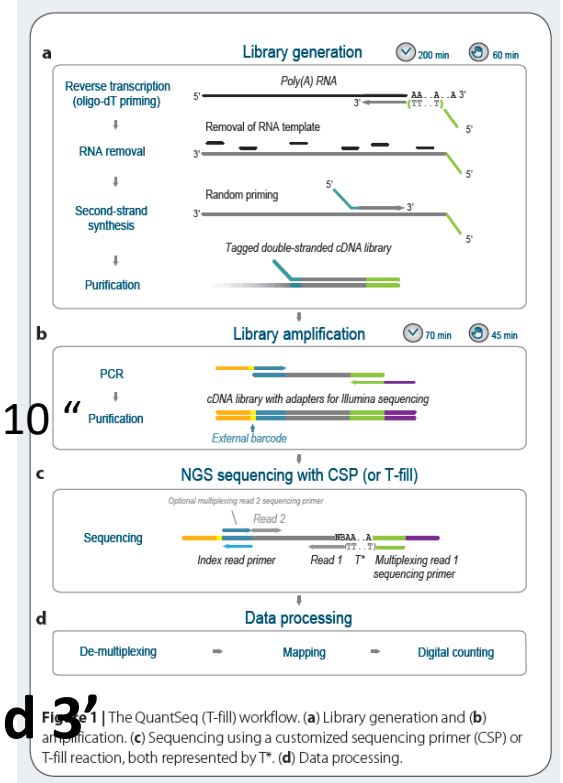
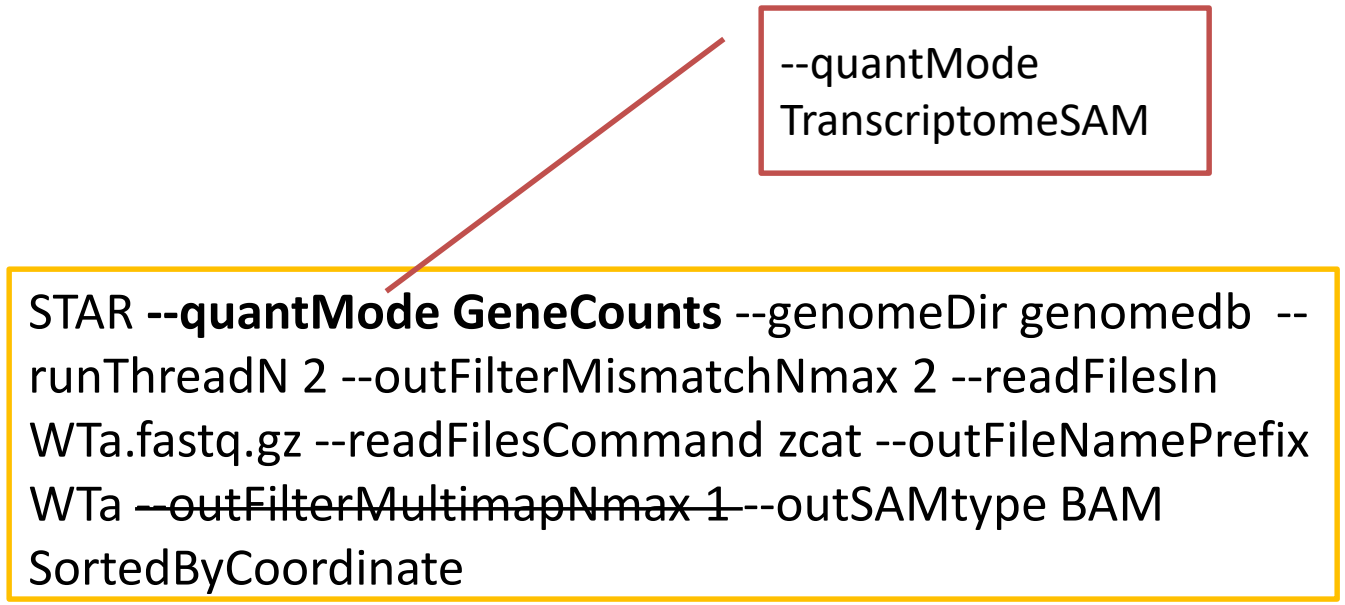


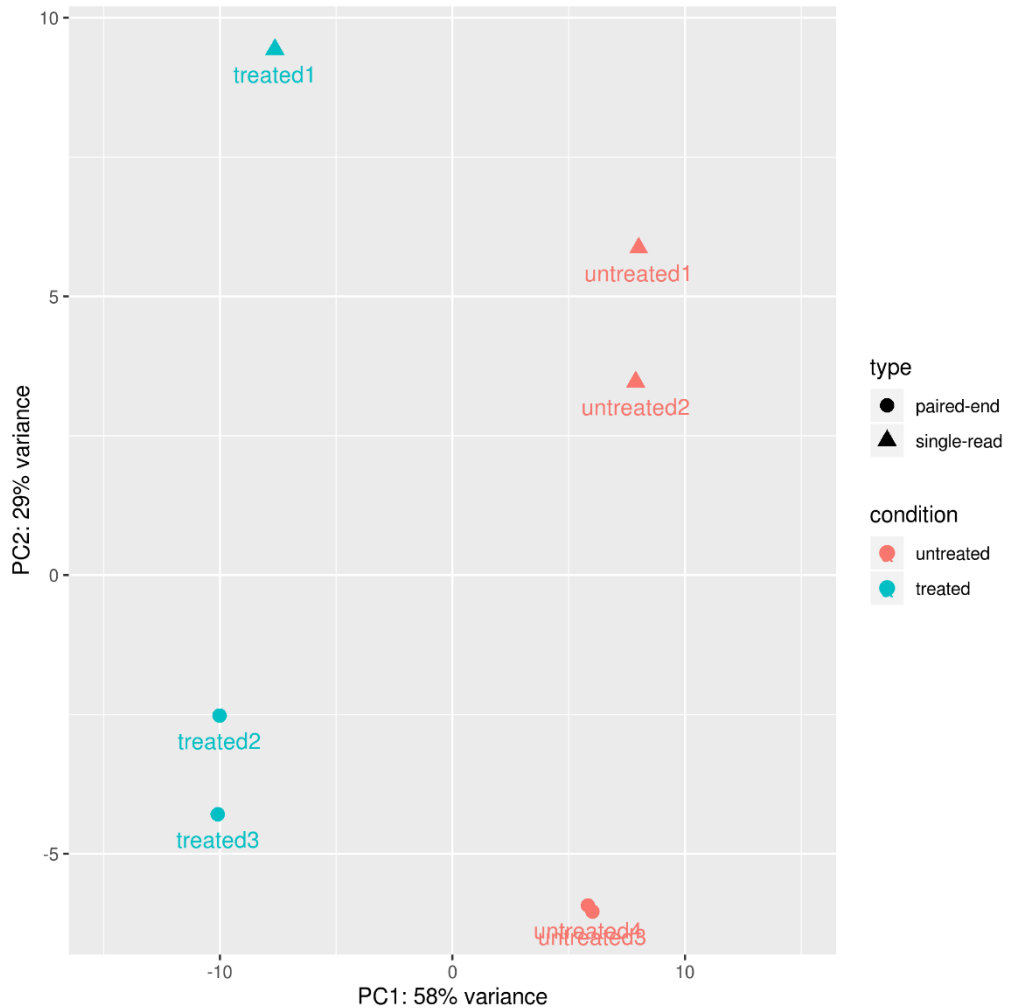
Figure 1 | The QuantSeq (T-fill) workflow. (**a**) Library generation and (**b**) amplification. (**c**) Sequencing using a customized sequencing primer (CSP) or T-fill reaction, both represented by T*. (**d**) Data processing.

Using STAR for transcript level quantification



The output transcript SAM can be used as input for RSEM.

Use PCA to check the sample variations and identify outliers.



BioHPC Lab office hours

Qi Sun, Jeff Glaubitz, Melissa Hubisz

Time: Every Monday, Tuesday & Thursday

Location: Zoom

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

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