

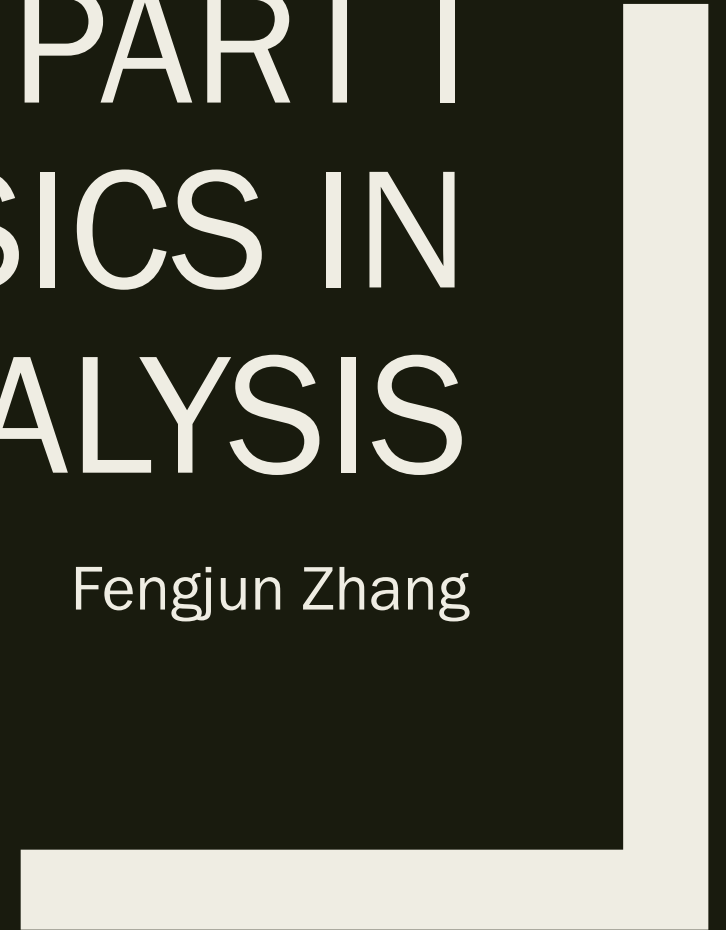
The image features two large, thick black L-shaped brackets. One is positioned in the top-left corner, and the other is in the bottom-right corner, framing the central text.

TRANSCRIPTOMICS

Daniel Dowling, Shrey Gandhi & Fengjun Zhang

PART I BASICS IN RNA-SEQ ANALYSIS

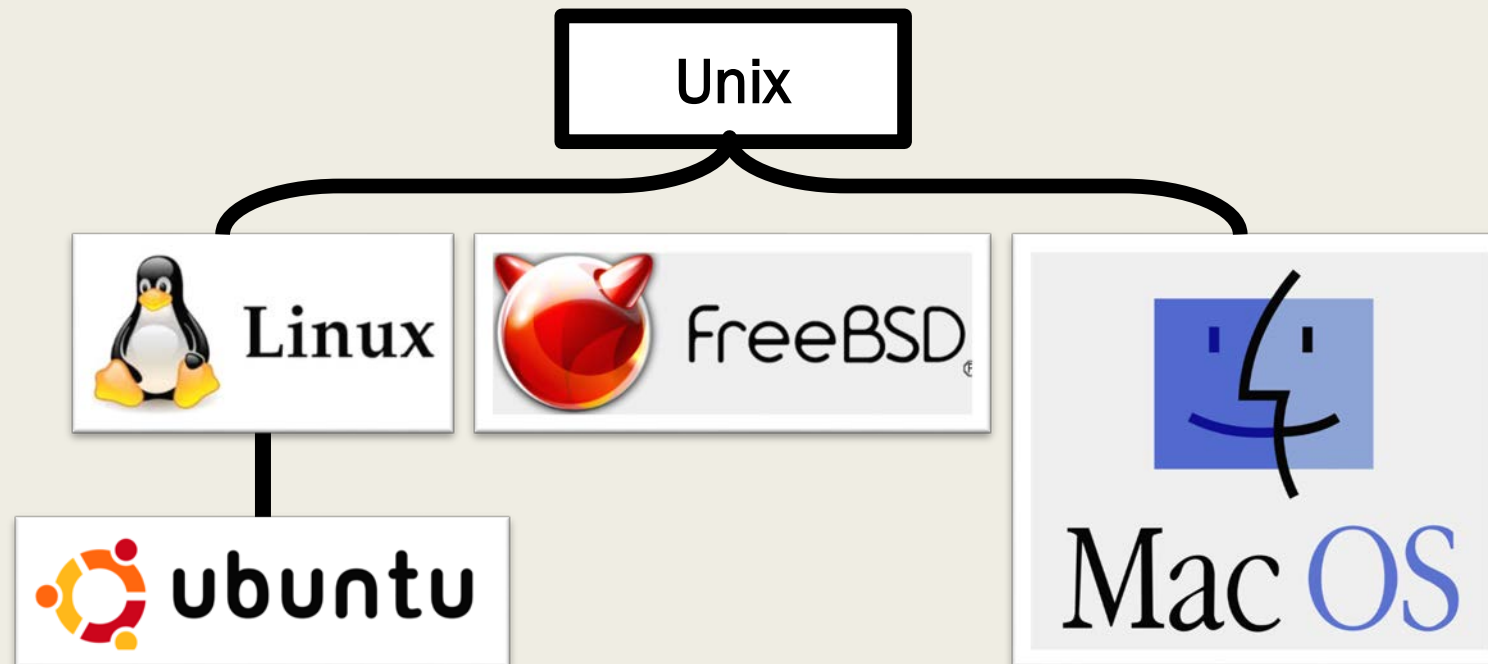
Fengjun Zhang



Before you start...

- What's your Operating System (OS)?
 - Most bioinformatic software work on Linux distribution

- What's your default shell?
 - `$ echo $SHELL`
 - bash or csh?
 - To call bash: `$ bash`



Summary of Basic command lines (1)

- \$ cd *PATH*
 - Change Directory
- \$ pwd
 - Print Working Directory
 - Variable \$PWD
- \$ ls [options] *PATH*
 - LiSt files
- Use ls in advanced way
 - \$ ls -laGh
 - Options:
 - -l : show full details
 - -a : all files include hidden ones
 - -G : make colorful
 - -h : size shown with human readable format
 - -S : sort by size
 - -t : sort by time

Summary of Basic command lines (1)

- `$ cd PATH`
 - Change Directory
- `$ pwd`
 - Print Working Directory
 - Variable `$PWD`
- `$ ls [options] PATH`
 - LiSt files
- Use ls in advanced way
 - `$ ls -laGh`
- Customized Abbreviation
 - `$ alias 'll'='ls -laGh'`
(temporary)
 - `$ echo \
"# alias for list\nalias 'll'='ls -laGh'" \
>> ~/.bash_profile && \
source ~/.bash_profile
(only for BASH)`

Summary of Basic command lines (2)

- \$ mv *FILEorPATH PATH/*
 - MoVe files/folders
- \$ cp [option] *FILE PATH/*
 - CoPy files
 - Use cp in advanced way
 - \$ cp -r *FOLDER PATH/*
 - option -r : recursively
- \$ rm -r *FILEorPATH*
 - ReMove files/folders recursively
- \$ mkdir [option] *PATH/*
 - MaKe DIRectory
 - Use mkdir in advanced way
 - \$ mkdir *PATH/* && cd \$_
 - && : and execute
 - \$_ : current temporary variable (*PATH/* in this case)

Summary of Basic command lines (3)

- Archive files under Unix-like system
 - *.zip: zipped files
 - To compress: \$ zip *NEW_ARCHIVE.zip* *FILE*
 - To decompress: \$ unzip *ARCHIVE.zip*
 - *.gz: G(NU)-zipped files
 - To compress: \$ gzip [-k] *NEW_ARCHIVE.gz* *FILE*
 - To decompress: \$ gunzip [-k] *ARCHIVE.gz*
 - *.tar.gz: tar G(NU)-zipped files
 - To compress: \$ tar -zcf *NEW_ARCHIVE.tar.gz* *FILE(s)/FOLDER(s)*
 - To decompress: \$ tar -zxf *ARCHIVE.tar.gz*

Installing bioinformatic software

- Installations via package managers
 - Ubuntu (& other Debian-based Linux distribution) : \$ sudo apt-get install *PROGRAM*
 - Mac OS : \$ brew install *PROGRAM*
([Homebrew installation](#))
- Follow instructions on its official website
- Get git clone ([GitHub guidelines](#))

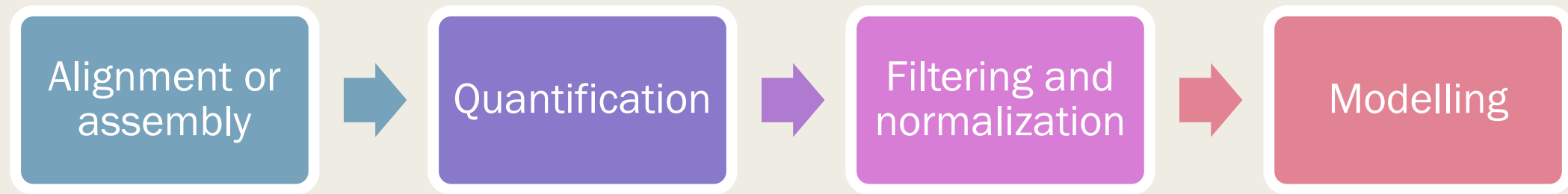
Installing bioinformatic software

- Customized installations (Linux and Mac)
 - Download binary files (usually *.tar.gz) to home folder
 - `$ mkdir ~/PROGRAM_NAME && cd $_; wget URLtoFILE.tar.gz`
 - Decompress
 - `$ tar -zxf FILE.tar.gz`
 - Set environmental variables (optional)
 - `$ export PATH=$PATH:$PWD/bin`
(only for BASH)

Installing bioinformatic software

- Exercise 1: install wget via package managers
- Exercise 2: install seqkit
 - <https://bioinf.shenwei.me/seqkit/download/>
- Exercise 3: install samtools (optional)
 - <https://www.biostars.org/p/328831/>
 - <http://www.htslib.org/download/>
- Try the following one (optional) 😊 (credit @Shrey)
 - Install cowsay
 - \$ cowsay Holy Cow

RNA-seq data analysis overview



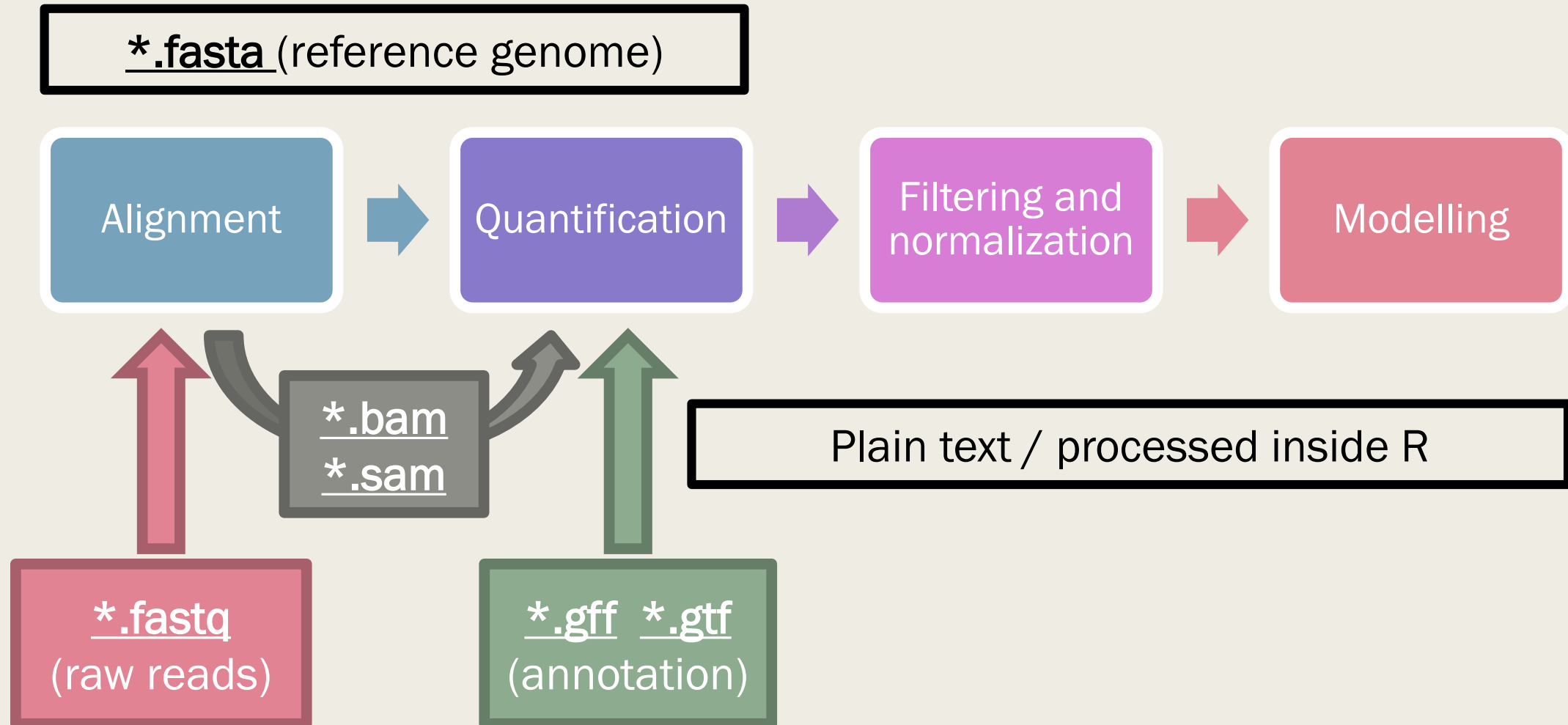
- TopHat
- STAR
- HISAT
- StringTile
- SOAPdenovo-Trans

- HTSeq
- featureCounts
- RSEM
- MMSEQ
- CuffLinks

- edgeR
- DESeq2
- CuffDiff2
- limma+voom

(Rory Stark et al. 2019)

RNA-seq data analysis overview



Common file formats in RNA-seq analysis:

- fasta & fastq

- *.fasta *.fna *.fa

- Header (name of the sequence) starts with “>”
- Sequence itself could be one line or multiple lines

- *.fastq *.fq

- Header usually starts with “@”
- 4 lines as a group
 - 1st: header (format differs among sequencers)
 - 2nd: sequence
 - 3rd: separator
 - 4th: quality of the sequencing
(scoring system might differ by suppliers: [Phred 33 or 64](#))

Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 1, mRNA

NCBI Reference Sequence: NM_001112.4

[GenBank](#) [Graphics](#)

>NM_001112.4 Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 1, mRNA

Header

```
GAGGCGCTGAGGCGGCCGTGGCGCGGCCGGCGGGCGGGCGGGCGGGCAGCGGGCCAAAGCGGCCAGGTTGGCG
GCCGGGGCTCCGGGCGCGCGAGGCCACGGCCACGCCCGCGCCGCTGCGCACAAACCAACGAGGCAGAGCGC
CGCCCGGCGCGAGACTGCGGCCGAAGCGTGGGGCGCGCGTGCAGGAGACCAGGCGCGGCCGGCTGCGGC
TGAGAGTGGAGCCTTTCAGGCTGGCATGGAGAGCTTAAGGGGCAACTGAAGGAGACACACTGGCCAAGCG
CGGAGTTCTGCTTACTTCAGTCTCTGCTGAGATACTCTCTCAGTCCGCTCGCACCGAAGGAAGCTGCCTTG
GGATCAGAGCAGACATAAAGCTAGAAAAATTTCAAGACAGAAACAGTCTCCGCCAGTCAAGAAACCTCA
AAAGTATTTTGGCATGGATATAGAAGATGAAGAAAAACATGAGTTCCAGCAGCACTGATGTGAAGGAAAAAC
CGCAATCTGGACAACGTGTCCCCAAGGATGGCAGCACACCTGGGCCTGGCGAGGGCTCTCAGCTCTCCA
ATGGGGGTGGTGGTGGCCCCGGCAGAAAAGCGGCCCTGGAGGAGGGCAGCAATGGCCACTCCAAGTACCG
CCTGAAGAAAAGGAGGAAAACACCAGGGCCCCGTCTCCCCAAGAACGCCCTGATGCAGCTGAATGAGATC
AAGCCTGGTTTTGCAGTACACACTCTGTCCCAGACTGGGCCCGTGCACGCGCCTTTGTTTGTCAATGCTG
TGGAGGTGAATGGCCAGGTTTTTGGAGGGCTCTGGTCCCACAAAAGAAAAGGCCAAAACCTCCATGCTGCTGA
GAAGGCCTTGAGGTCTTTCGTTTTCAGTTTCCCTAATGCCTCTGAGGCCACCTGGCCATGGGGAGGACCTG
TCTGTCAACACGGACTTCACATCTGACCAGGCCGACTTCCCTGACACGCTCTTCAATGGTTTTGAAACTC
CTGACAAGGGCGGAGCCTCCCTTTTACGTGGGCTCCAATGGGGATGACTCCTTCAGTTCAGCGGGGACCT
CAGCTTGTCTGCTTCCCCGGTGCCTGCCAGCCTAGCCCAGCCTCCTCTCCCTGTCTTACCACCATTCCCA
CCCCCGAGTGGGAAGAATCCCCGTGATGATCTTGAACGAACTGCGCCCAGGACTCAAGTATGACTTCCTCT
CCGAGAGCGGGGAGAGCCATGCCAAGAGCTTCGTGCATGTCTGTGGTCTGGATGGTCAGTTCCTTTGAAGG
CTCGGGGAGAAAACAAGAAGCTTGCCAAGGCCCGGGCTGCGCAGTCTGCCCTGGCCGCCATTTTTAACTTG
CACTTGGATCAGACGCCATCTCGCCAGCCTATTCAGTGGGGTCTTCAGCTGCATTTACCGCAGGTTT
TAGCTGACGCTGTCTCACGCCCTGGTCTGGGTAAGTTTGGTGCCTGACCGACAACCTCTCCTCCCCTCA
CGCTCGCAGAAAAGTGTGCTGGCTGGAGTCGTGCATGACAACAGGCACAGATGTTAAAGATGCCAAGGTGATA
AGTGTTCACAGGAACAAAATGTATTAATGGTGAATACATGAGTGCATCGTGGCCTTGCATTAATGACT
GCCATGCAGAAATAATATCTCGGAGATCCTTGTCTCAGATTTCTTTATACACAACCTTGAGCTTTACTTAAA
TAACAAAGATGATCAAAAAAGATCCATCTTTTCAGAAATCAGAGCGAGGGGGTTTAGGCTGAAGGAGAAT
GTCCAGTTTCATCTGTACATCAGCACCTCTCCCTGTGGAGATGCCAGAATCTTCTCACCACATGAGCCAA
TCCTGGAAGAACCAGCAGATAGACACCCAAAATCGTAAAGCAAGAGGACAGCTACGGACCAAAATAGAGTC
TGGTGAAGGGGACGATTCAGTGCCTCCAATGCGAGCATCCAAACGTGGGACGGGGTGTGCAAGGGGAG
CGGCTGCTCACCATGCTCTGCAGTGACAAGATTGCACGCTGGAACGTGGTGGGCATCCAGGGATCCCTGC
TCAGATTTTCGTGGAGCCATTTACTTCTCGAGCATCATCCTGGGCAGCCTTTACCACGGGGACCACT
TTCCAGGGCCATGTACCAGCGGATCTCCAACATAGAGGACCTGCCACCTCTCTACACCTCAACAAGCCT
TTGCTCAGTGGCATCAGCAATGCAGAAGCACGGCAGCCAGGGAAAGGCCCCCAACTTCAGTGTCAACTGGA
CGGTAGGCGACTCCGCTATTGAGGTCATCAACGCCACGACTGGGAAGGATGAGCTGGGCCGCGCGTCCCCG
```

Sequence

```
[fengjun@retrogenomics3-work] [(·θ·)]
```

```
└─[~/rnaseq/human/ori-fq] head -50 ERR2598363.fastq
```

```
@ERR2598363.1 WINDU:52:C6NM6ANXX:7:1101:1125:69999/1
```

```
GTGTGTAGCTATTGTAAGGTTTCATTTCTTATAATGATCTTATACACTATCATTGGAAAGTACTTTAGAGCAGGCAAGGGATTTGTGGTATGGTGAGAN
```

```
+
```

```
BBB/ << /B / / / <F / <FB / <B#####
```

```
@ERR2598363.2 WINDU:52:C6NM6ANXX:7:1101:1125:70375/1
```

```
CCAAGAATTAAGAAGCTCTAAAGTCTACAAACATCTTACTTCACCAAGACTAACTATAATTGAAGGGTTACTATTTGTTAATAAAAAATCACACATCAN
```

```
+
```

```
/ <BB / / <FFFBF / BFFBFFFF / <FF / <FFF / BFFFFB <FFBFFBFFFFBFFF <FFFFFF << / <F <FFF BFFFFF / < / FF / F <F <BF BFBFF <FFFF #
```

```
@ERR2598363.3 WINDU:52:C6NM6ANXX:7:1101:1125:71003/1
```

```
GGGAGGGGAAAAAAGGATATACAGGGGCAGGTGTATTCTCTGTACAGAGGTGCAGAGAAAATTTACATAGCTTTAGAGAATGCCTTGTGGAAAAAAAAN
```

```
+
```

```
BBBBBFFBFFFFF#####
```

```
@ERR2598363.4 WINDU:52:C6NM6ANXX:7:1101:1125:72223/1
```

```
CCCACAATAAAGTATGCATGTTACGTCACCATCTGTCATTATTCAAATATTCCAAATACAAACATAGAGCATTAAACAAACAGGTTAAAAACGTTCTCAN
```

```
+
```

```
B / BBBF / BFFFFF << FF / BF / <FFF << FFBFFB <FF / / << FFFB < / < / << / B / F <BBF / <BFFB#####
```

```
@ERR2598363.5 WINDU:52:C6NM6ANXX:7:1101:1125:73531/1
```

```
CTGGGCCTCCAGTTCAGGAATACATATTAGCTGCAGCCACTCCTGGATGTTACCTGGATTTGTCCAGTGCCATCTTTGTCAAGAGATTTGAAGGCACGN
```

```
+
```

```
<BB / < / / B <FF / FF <F / FBFBFFBFF / <FFFFB / / <BB <FFFFF BFFFFF BFFFFF / F <BF / <FFF BFFFF / FFFFF / / <FBB / <BF / FFBF #
```

```
@ERR2598363.6 WINDU:52:C6NM6ANXX:7:1101:1125:88402/1
```

```
GCGGTCAATAGGCACACTCTCTTTGCATCAGTTCCTTGTGTCCCCTGTAGCCATGCAGCCGGAGCGGGTTGTGTGCTCCTACGGACTTGGCGCTGGGCTN
```

```
+
```

```
< / / / / <F / B / / B / B <F#####
```

```
@ERR2598363.7 WINDU:52:C6NM6ANXX:7:1101:1126:19004/1
```

```
GGCGGCATTTTCTTTGCAGCCTTCTCATTCTTTTCCGATTCTGTCATTTTAAAGATCTTGAATATGATCTTCCAGGCTGTGAACATTCTCAATCTCGT
```

```
+
```

```
BB / <BFFFB <FBFFB / / <FF#####
```

Header

Sequence

Quality

Common file formats in RNA-seq analysis:

· fasta & fastq

■ From *.fastq to *.fasta

- Simple shell script : slow but no additional program installations

- `$ sed -n '1~4s/^@/>/p;2~4p' FILE.fq > FILE.fa`

- Other programs : parallel computing (quick) and no need to decompress archive files

- `$ seqkit fq2fa FILE.fq.gz -o FILE.fa.gz`


- Other programs like FastQC is also available

Common file formats in RNA-seq analysis:


- gff & gtf































- Both *.gff3 and *.gtf are files for annotation of a reference genome
 - The two are in different structures but usually contain the same information
 - ([Detailed explanation](#))
- Used for quantification tools, along with fasta files from the reference genome
- Latest version in Ensembl: <https://www.ensembl.org/info/data/ftp/index.html>

★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations
Y	Human <i>Homo sapiens</i>	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	GFF3 TSV RDF JSON
Y	Mouse <i>Mus musculus</i>	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON
Y	Zebrafish <i>Danio rerio</i>	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON



Index of /pub/release-97/gff3/homo_sapiens

 [parent directory]

Name	Size	Date Modified
 CHECKSUMS	1.6 kB	31/05/2019, 19:22:00
 Homo_sapiens.GRCh38.97.abinitio.gff3.gz	6.2 MB	27/05/2019, 10:36:00
 Homo_sapiens.GRCh38.97.chr.gff3.gz	39.0 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chr_patch_hapl_scaff.gff3.gz	42.5 MB	27/05/2019, 10:35:00
 Homo_sapiens.GRCh38.97.chromosome.1.gff3.gz	3.6 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.10.gff3.gz	1.5 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.11.gff3.gz	2.3 MB	27/05/2019, 10:32:00
 Homo_sapiens.GRCh38.97.chromosome.12.gff3.gz	2.2 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.13.gff3.gz	711 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.14.gff3.gz	1.4 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.15.gff3.gz	1.5 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.16.gff3.gz	1.8 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.17.gff3.gz	2.3 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.18.gff3.gz	736 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.19.gff3.gz	2.3 MB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.2.gff3.gz	2.9 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.20.gff3.gz	934 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.21.gff3.gz	480 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.22.gff3.gz	887 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.3.gff3.gz	2.4 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.4.gff3.gz	1.6 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.5.gff3.gz	1.8 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.6.gff3.gz	1.8 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.7.gff3.gz	1.9 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.8.gff3.gz	1.9 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.9.gff3.gz	1.7 MB	27/05/2019, 10:32:00
 Homo_sapiens.GRCh38.97.chromosome.MT.gff3.gz	3.3 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.chromosome.X.gff3.gz	1.2 MB	27/05/2019, 10:34:00
 Homo_sapiens.GRCh38.97.chromosome.Y.gff3.gz	136 kB	27/05/2019, 10:33:00
 Homo_sapiens.GRCh38.97.gff3.gz	39.0 MB	27/05/2019, 10:35:00
README	11.7 kB	27/05/2019, 02:11:00

How to download from FTP site

- Downloading large files via browser is not recommended
- `$ cd PATH/ ; wget URLtoFILE`

Common file formats in RNA-seq analysis:

- bam & sam

- Both are common output files from aligner programs
- *.bam is the binary format for *.sam. *.bam files are smaller, suitable for storage
- SAM stands for Sequence Alignment/Map format

```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Header: starts with '@'

Alignment: tab-delimited text

Alignment: tab-delimited text

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```



Reads name



Reference name

Alignment: tab-delimited text

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```



FLAG

- FLAG: bitwise numbers indicate results of the alignment
 - [Simple FLAG explanation](#)
 - Useful for filtering

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

Alignment: tab-delimited text

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```



MAPQ

- MAPQ: MAPping Quality
 - Used for quality control
 - Careful: score might differ among aligners
- ([further reading about MAPQ](#))

Common file formats in RNA-seq analysis:

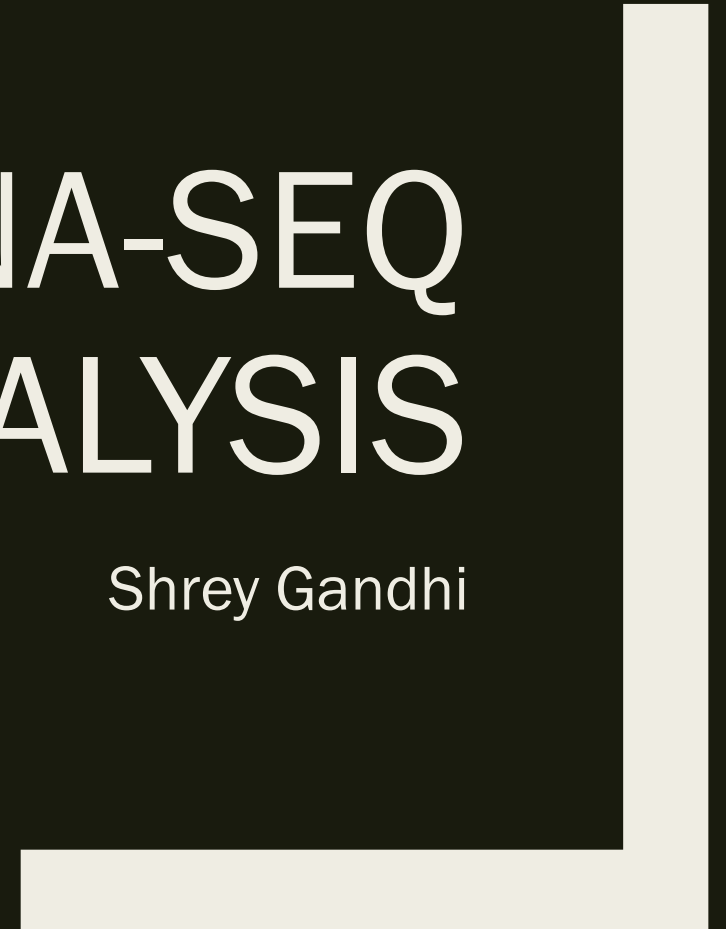
· bam & sam

- From *.bam to *.sam with samtools
 - Basic conversion
 - `$ samtools view [-h] FILE.bam > FILE.sam`
 - Filter out unmapped reads
 - `$ samtools view [-h] -F 4 FILE.bam > FILE.sam`
 - Extract unique reads (only for TopHat)
 - `$ samtools view [-h] -q 50 FILE.bam > FILE.sam`

PART 2

RNA-SEQ
DATA ANALYSIS

Shrey Gandhi



Transcriptome Sequencing (RNA-Seq)

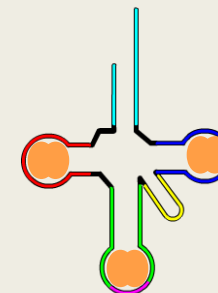
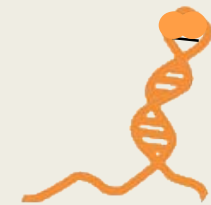
- ❑ Differential Gene/Transcript Expression
 - Quantitative evaluation and comparison of transcript levels across different groups
 - Functional studies
- ❑ Transcriptome Assembly
 - Build new or improve gene assemblies/models of the genome
 - Novel gene identification
- ❑ Splice variant analysis
- ❑ SNP detection
- ❑ Meta-transcriptomics
 - Profiling of community-wide gene expression (e.g., gut bacteria, soil)
 - Gene activity diversity
 - Gene expression abundance

Types of RNA

- ❑ Ribosomal RNA (rRNA)
 - Responsible for protein synthesis
 - up to 95% of total RNA in a cell
- ❑ Messenger RNA (mRNA)
 - Translated into proteins and have Poly-A tail in eukaryotes
 - 2-3% of total RNA in a cell
- ❑ Long non-coding RNA (lncRNA)
 - > 200 bases long and not translated into proteins
 - May or may not have poly-A tail
 - Can be circular as well (Circular RNA)
- ❑ Micro RNA (miRNA)
 - ~22 bases long involved in expression regulation
- ❑ Transfer RNA (tRNA)
 - Bring specific amino acids for protein synthesis
- ❑ Others (shRNA, snRNA, siRNA, snoRNA etc)



AAAAAAAAAA



Experimental considerations and challenges

Experimental Design Considerations:

- ❑ Biological question?
- ❑ Genome Availability
- ❑ RNA quality: RIN values
- ❑ Biological replicates:
 - Measurement of variation between samples
 - At least 3 biological replicates for statistical power
 - More are better
- ❑ Batch Effect
 - Best to sequence everything for an experiment at the same time
 - Consistency in library preparation
 - If unavoidable – Distribute labelled libraries for different groups equally across batches

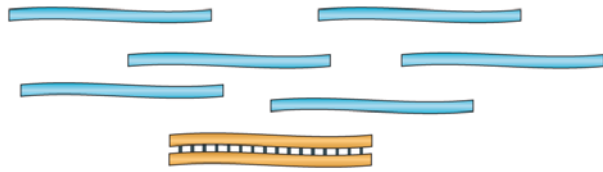
Illumina Sequencing Technology



Library Preparation

a Data generation

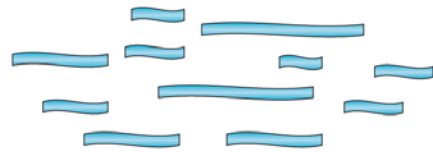
① mRNA or total RNA



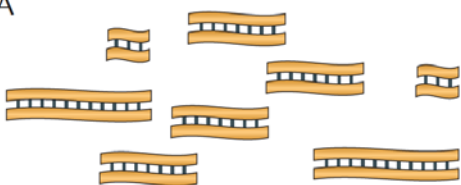
② Remove contaminant DNA



③ Fragment RNA
Remove rRNA?
Select mRNA?

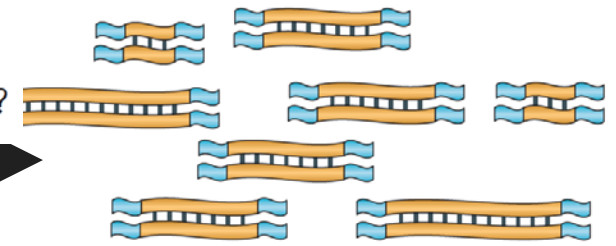


④ Reverse transcribe into cDNA



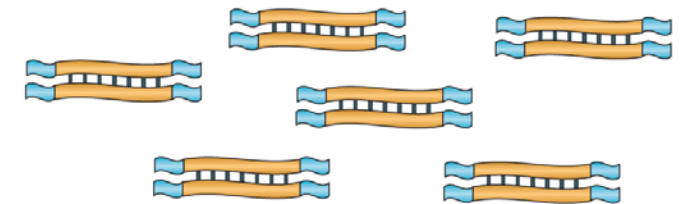
⑤ Ligate sequence adaptors

Strand-specific RNA-seq?

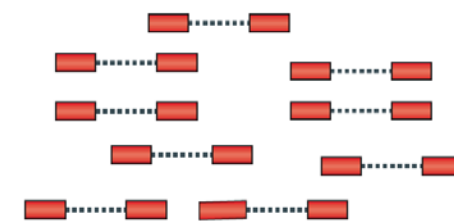


PCR amplification?

⑥ Select a range of sizes



⑦ Sequence cDNA ends

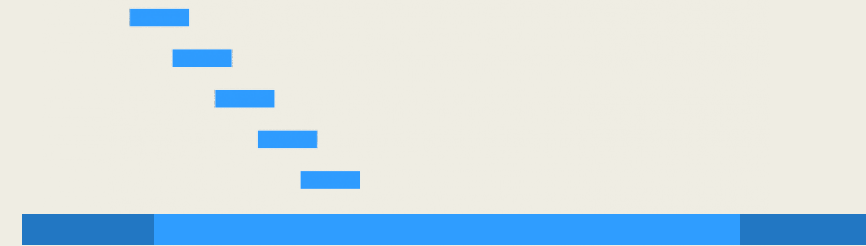


Experimental considerations and challenges

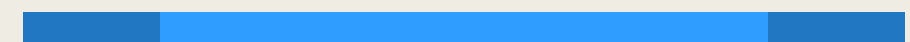
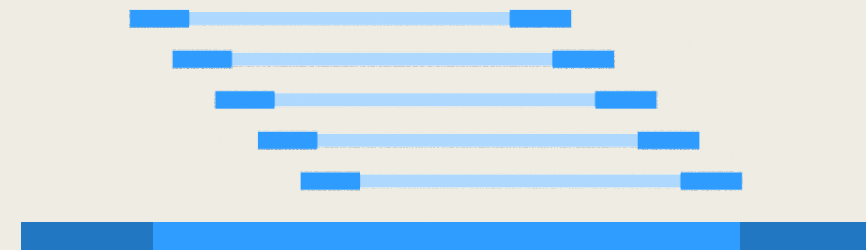
Sequencing Considerations:

- Single-read or paired-end sequencing
- Stranded or un-stranded libraries
 - Can identify which strand of DNA was transcribed
 - Strandedness is preferred for all applications
- Read length
- Sequencing coverage and depth
- Types of RNA Selection:
 - rRNA removal
 - Poly-A selection (eukaryotes) – mRNA Sequencing
 - rRNA depletion – Total RNA Sequencing
 - Size selection – small RNA Sequencing

Single-end reads



Paired-end reads



Adapter

cDNA

Adapter

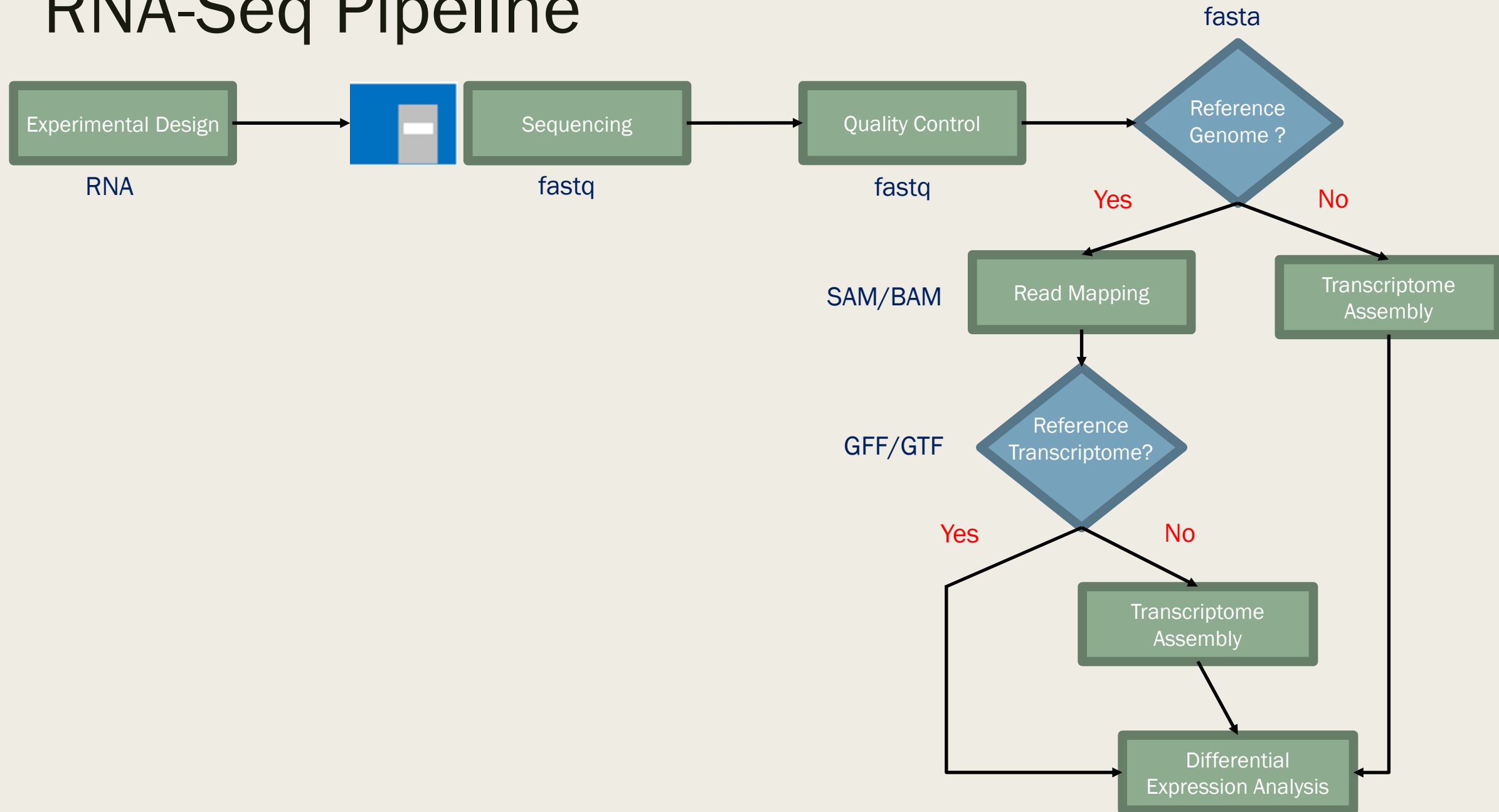
~ 60 bases

~ 100 - 800 bases

~ 60 bases

Insert Size

RNA-Seq Pipeline



File Formats

Sequence formats

- FASTA
- FASTQ

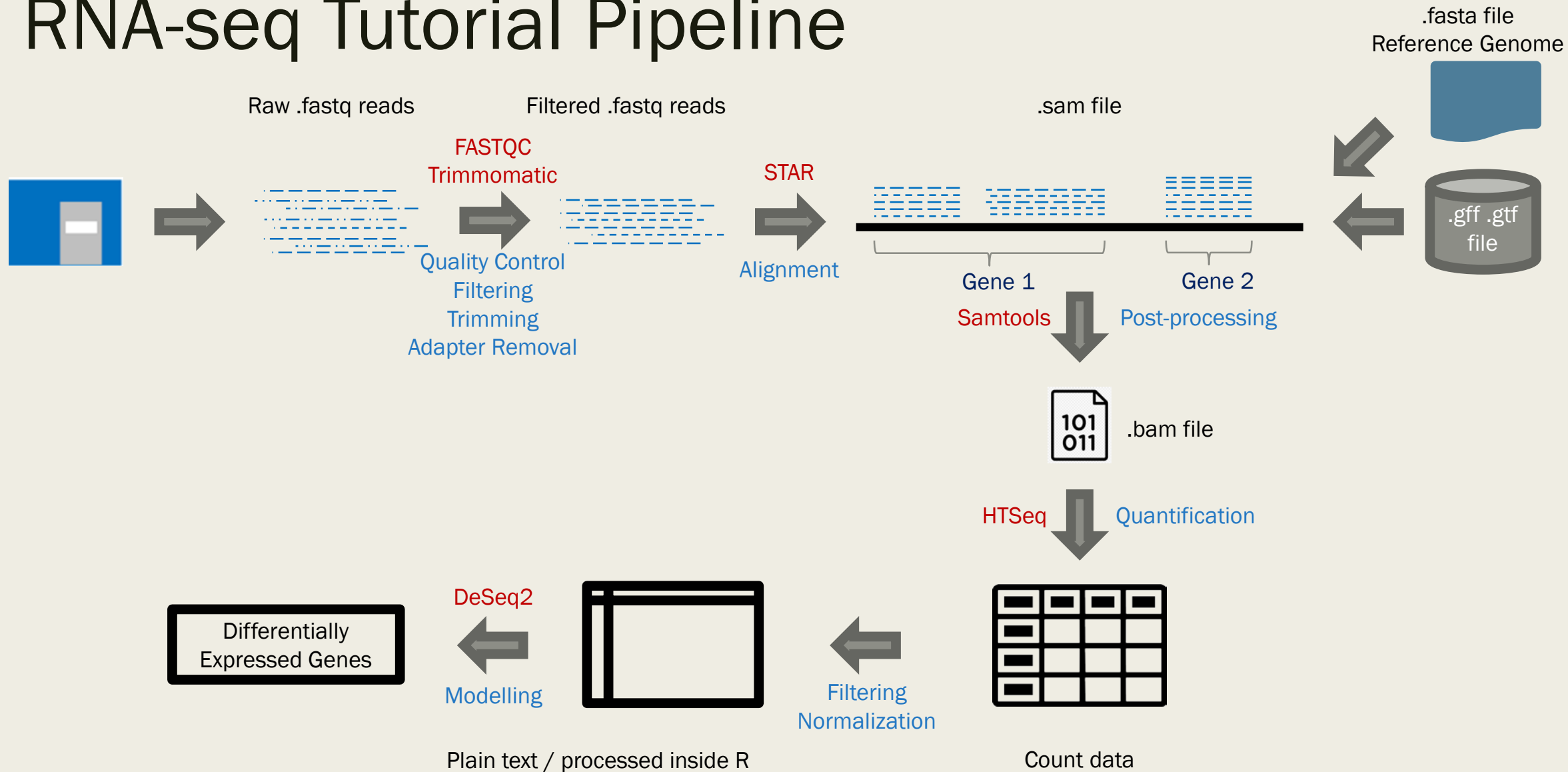
Annotation formats

- GFF
- GTF

Alignment formats

- SAM
- BAM

RNA-seq Tutorial Pipeline



Step 1: Quality Control

FASTQC:

- Tool to analyse Fastq sequence quality
- Gives an overview of the sequencing quality associated with fastq files
- Execute:

fastqc

- Base quality and content, read length, k-mer content, presence of ambiguous bases, over-represented sequences, and duplicates.
- A good sequence report -
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html
- A bad sequence report -
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Step 1: Quality Control

Trimmomatic:

- Trimmomatic allows dynamic read filtering, trimming and adapter removal
- Execute:

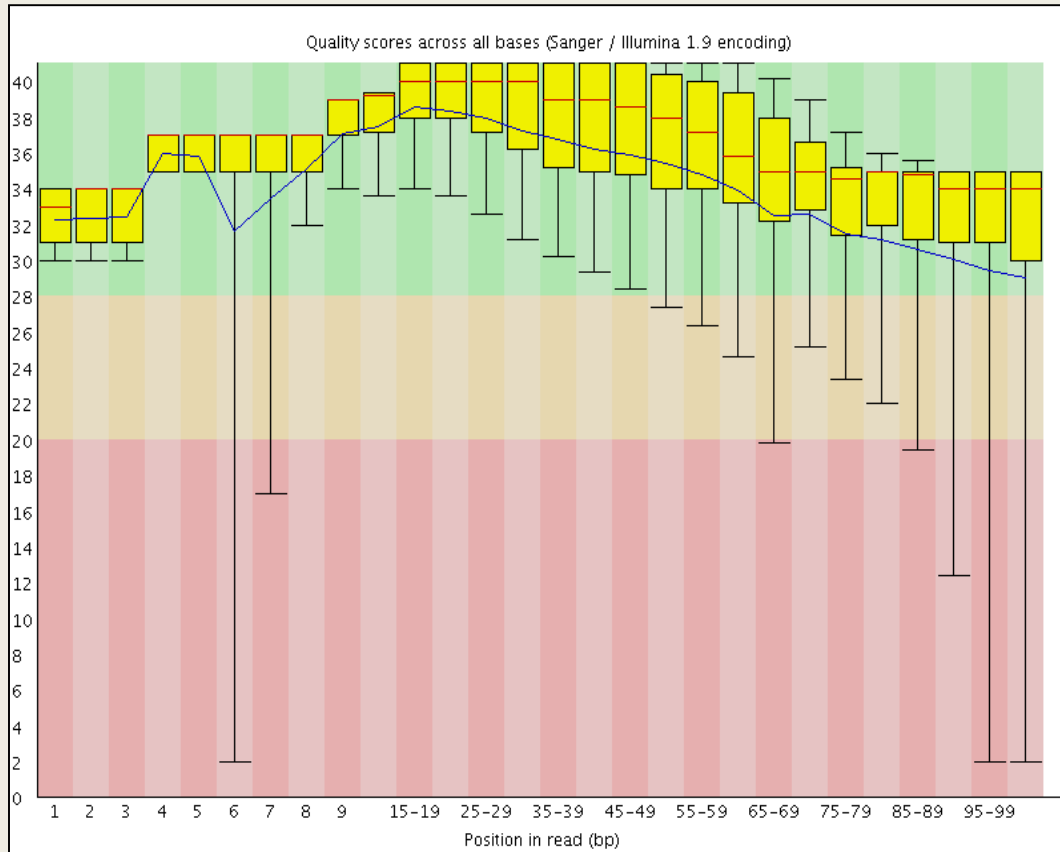
```
java -jar /software/Trimmomatic-0.39/trimmomatic-0.39.jar PE -threads 4  
filename_1.fastq filename_2.fastq trimmed_filename_1.fastq  
unpaired_filename_1.fastq trimmed_filename_2.fastq unpaired_filename_2.fastq  
AVGQUAL:20 SLIDINGWINDOW:5:20 MINLEN:50
```

- AVGQUAL – Average Read quality
- SLIDINGWINDOW – Checks Reads for trimming
- MINLEN – Minimum Read length
- ILLUMINACLIP – Adapter Removal

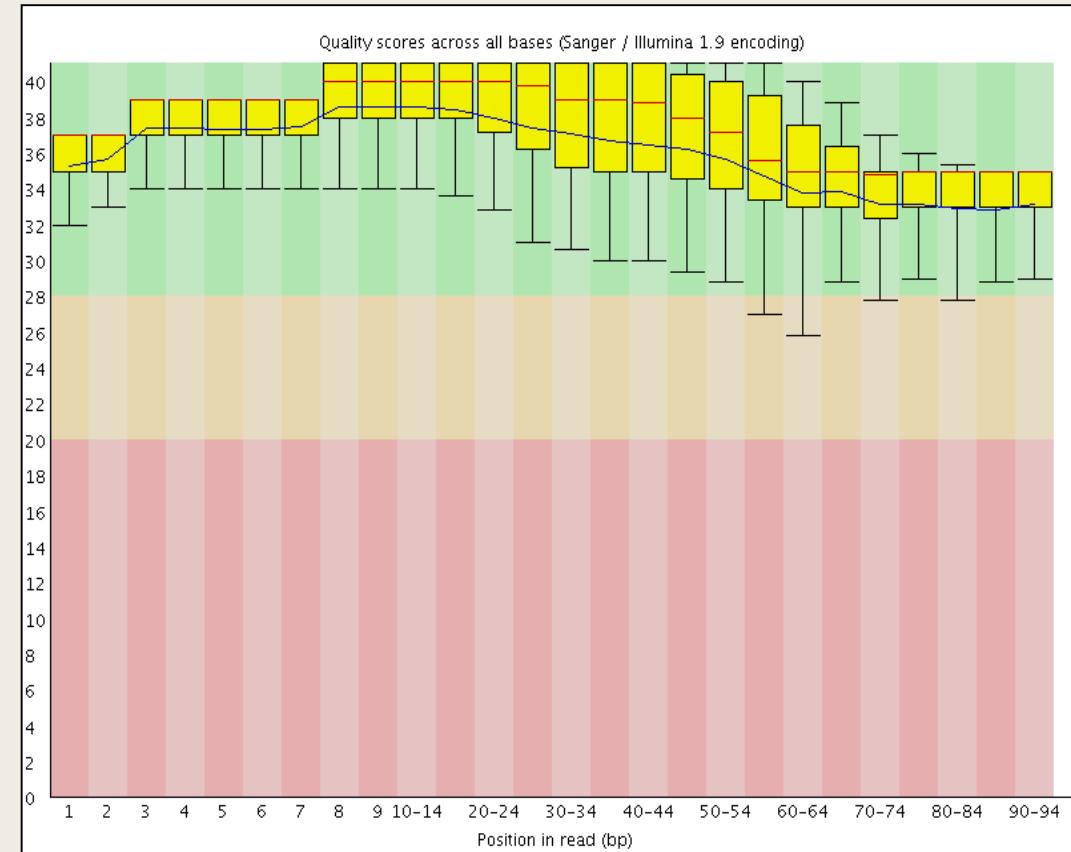
Step 1: Quality Control

FastQC Quality Reports:

Before quality trimming

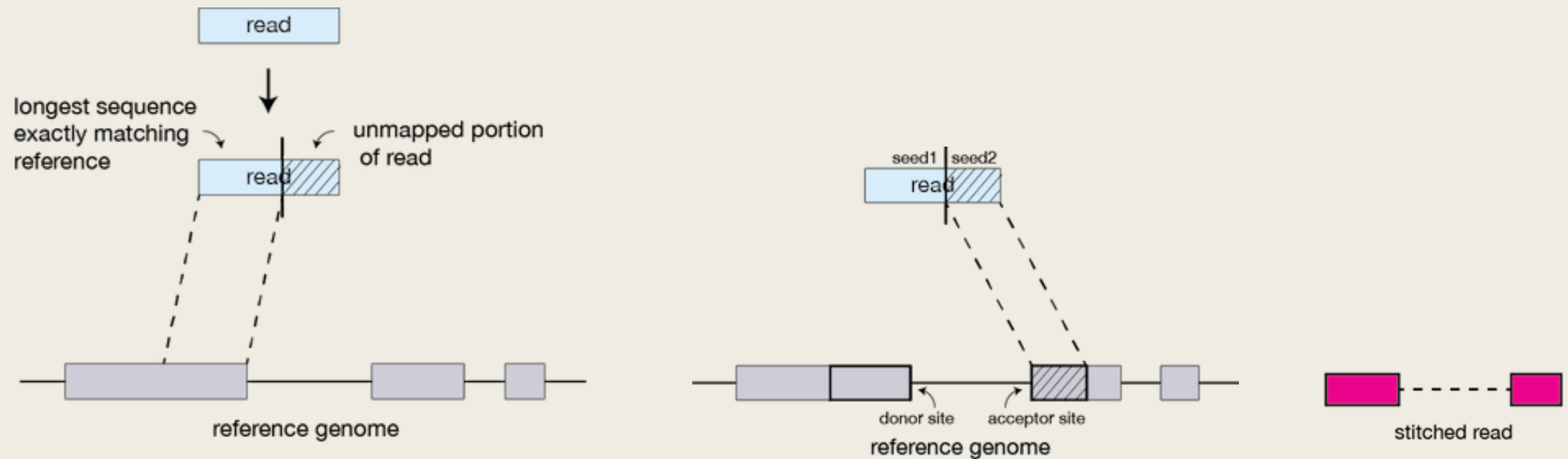


After quality trimming



Splice-aware mapping

- ❑ RNA-Seq reads might span large introns which are not represented in the cDNA
- ❑ Splice aware aligners are needed to align reads back to the reference genome
- ❑ Reads can also be aligned directly to reference transcriptome
 - Recommended only for well annotated transcriptomes
 - Novel transcripts isoforms can't be detected



Step 2: Aignment



STAR:

- Fast, accurate and splice aware aligner
- Drawback : Memory intensive
- Reference Genome : Genome assemblies can be downloaded from NCBI, Ensembl, Gencode and UCSC genome browser websites.
- Generating index:

```
STAR --runMode genomeGenerate --genomeDir genome/ --  
genomeFastaFiles genome/chr10.fa --sjdbGTFfile genome/chr10.gtf --  
sjdbOverhang 74
```

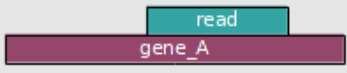
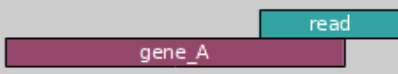


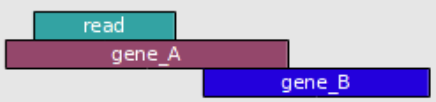
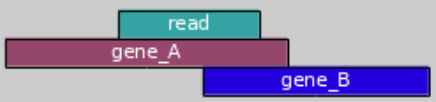
- Align reads to the genome:

```
STAR --genomeDir genome/ --readFilesIn trimmed_LA_R1.fastq  
trimmed_LA_R1.fastq --outFileNamePrefix leftatrium
```

Step 3: Gene quantification

HTSeq-count:

- Three modes of overlap resolution:
 - Union
 - Intersection-strict
 - Intersection-nonempty
- Outputs a table with counts for each feature
- Drawback:
 - Simple counting based method
 - Quantifying the abundances of individual transcripts not possible
- Execute:

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A

```
htseq-count -f sam rightatriumAligned.out.sam ../chr10.gtf >  
RA_count_data
```

PART 3

DIFFERENTIAL
EXPRESSION ANALYSIS

Daniel Dowling

