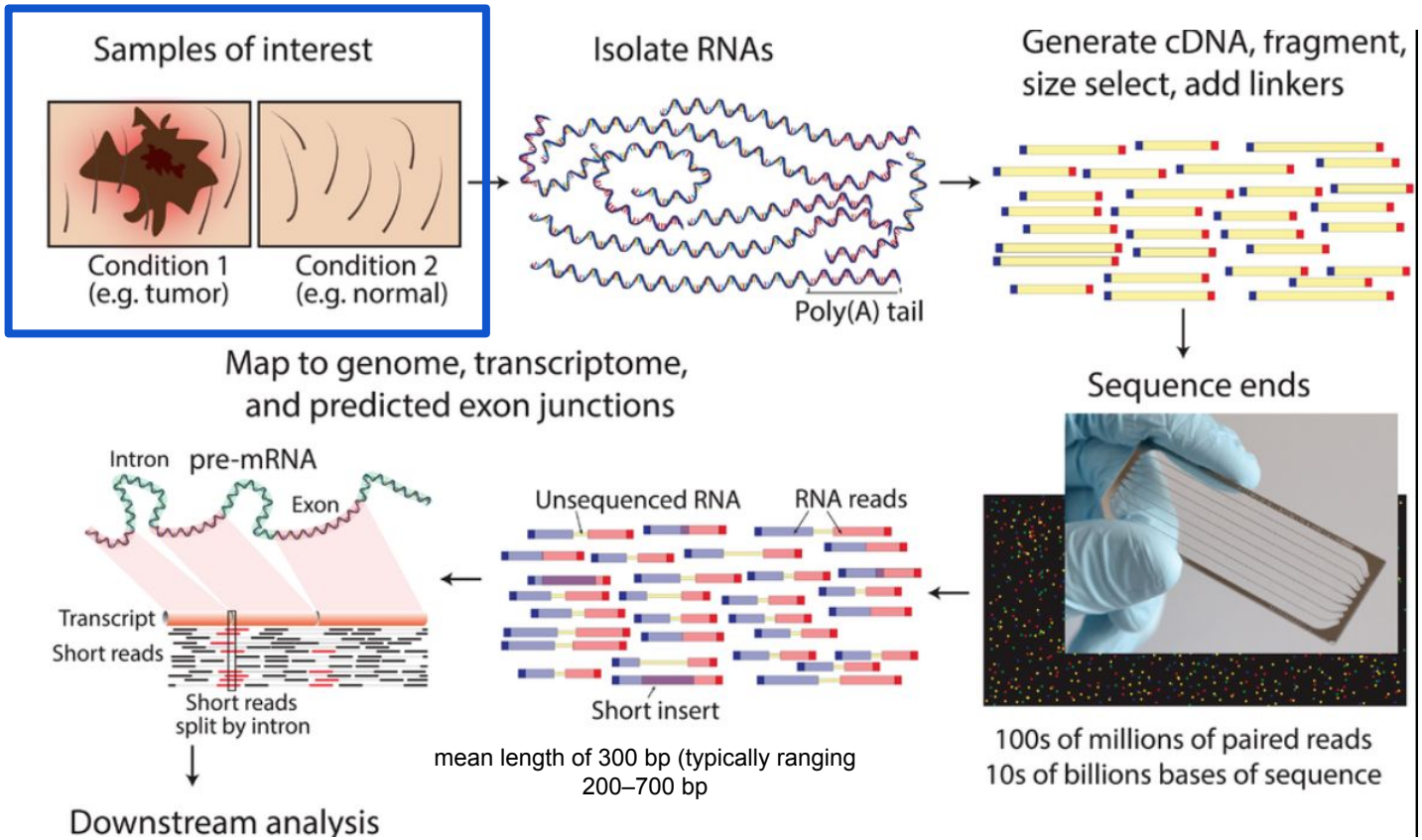
Library preparation

Please revise slides 2-8 before the workshop

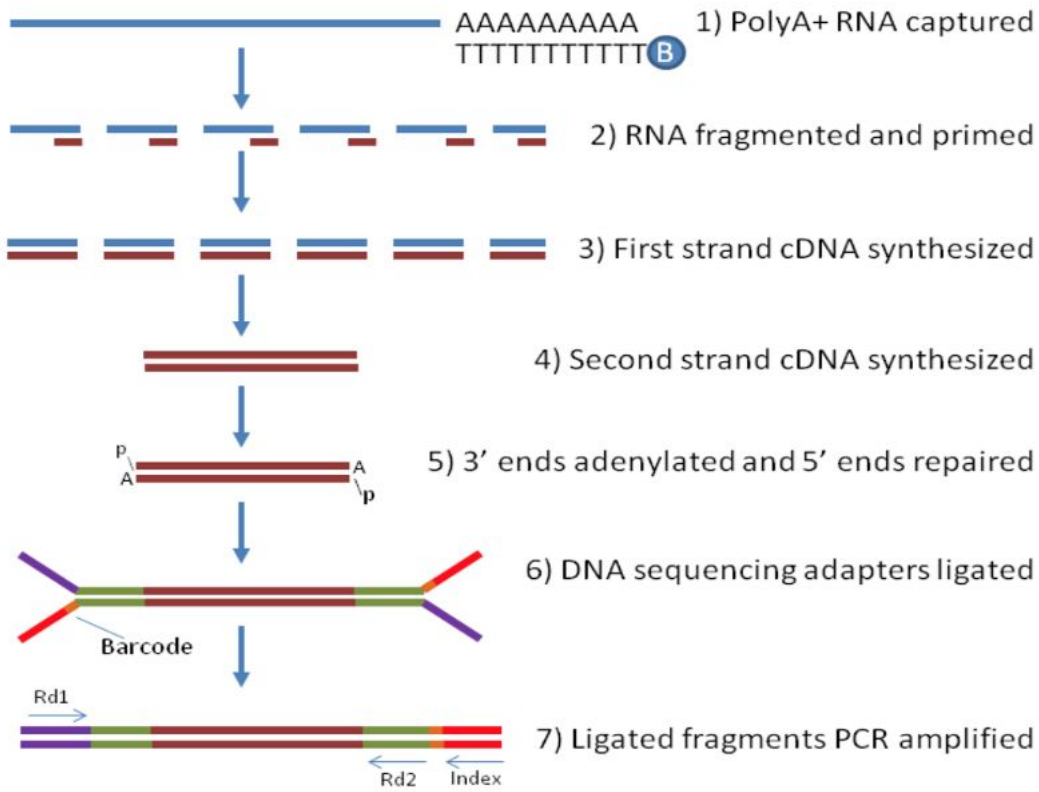
GLADSTONE
INSTITUTES

RNA-seq - experiment



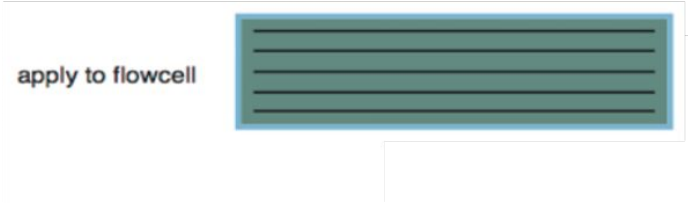
https://www.wikiwand.com/en/RNA-Seq

How does the bulk-RNA-seq technology work?



Flow cell organization for sequencing

cDNA library is applied to a flow cell



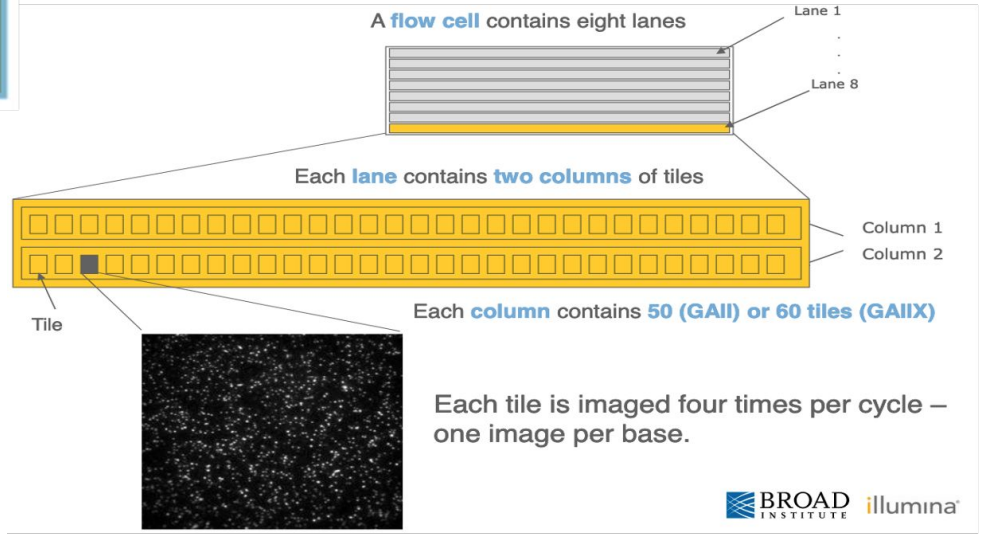
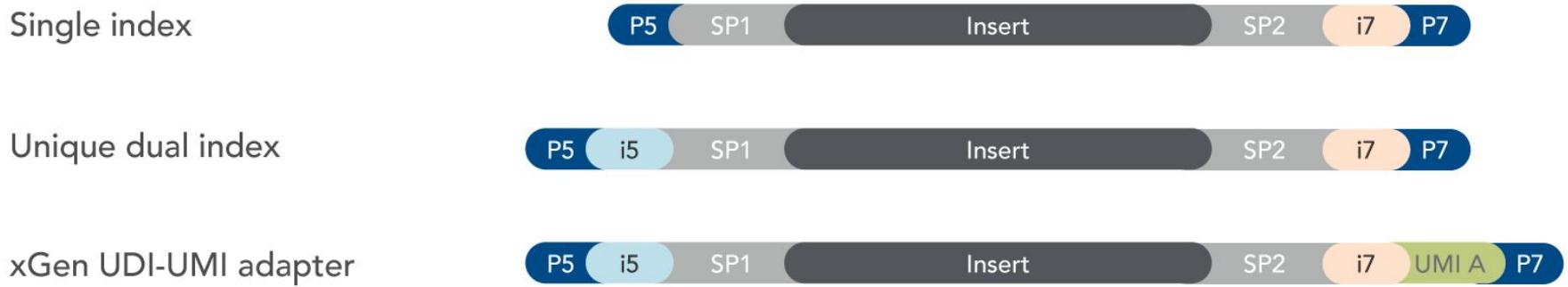


Image borrowed from slides shared by Broad Institute (link in description).

Fragments preparation for the sequencer



Flow cell binding sequence: Platform-specific sequences for library binding to instrument

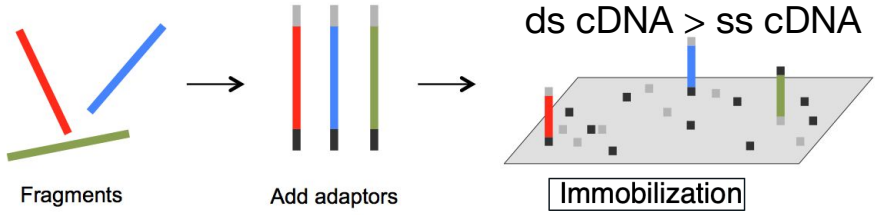
Sequencing primer sites: Binding sites for general sequencing primers

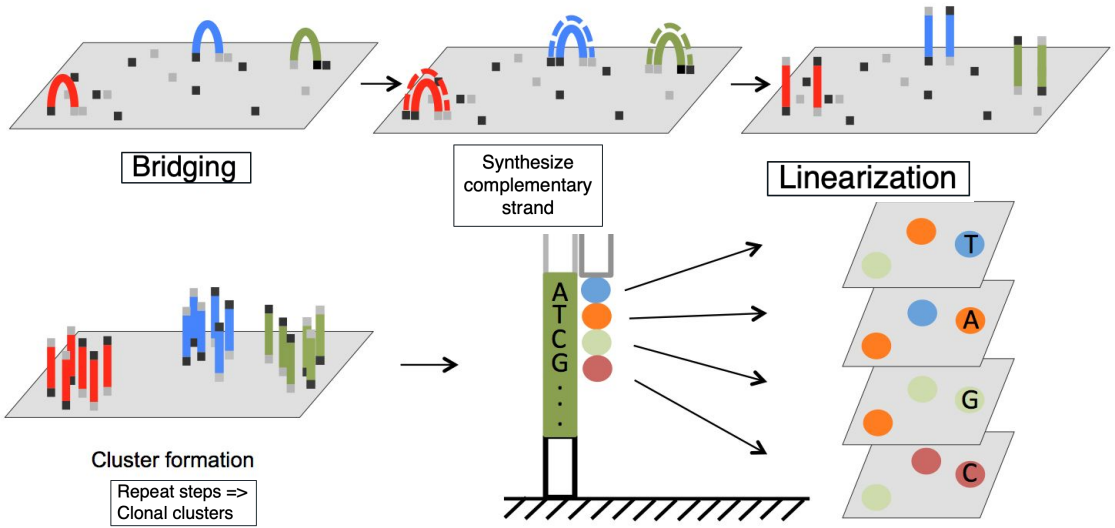
Sample indexes: Short sequences specific to a given sample library

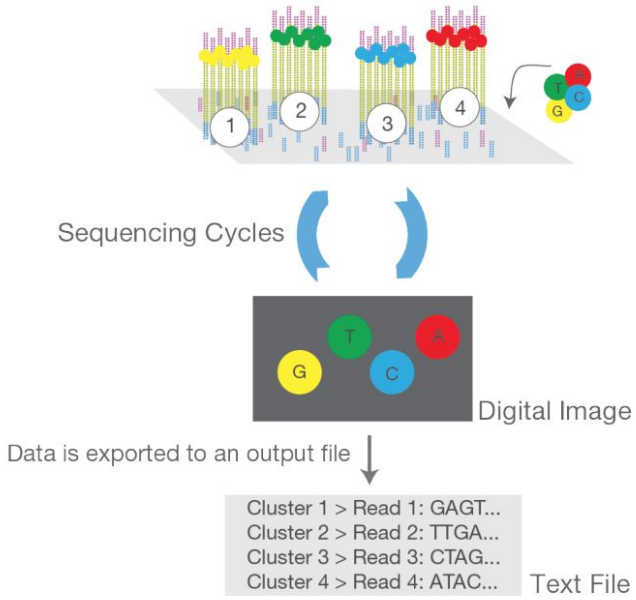
Molecular index/barcode: Short sequence used to uniquely tag each molecule in a given sample library

Insert: Target DNA or RNA fragment from a given sample library

DNA fragments immobilized on flow cell & amplified into clonal clusters







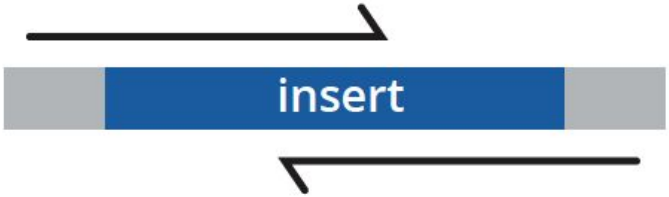
From DOI: 10.5772/61657 and
https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Types of sequencing

Single-end sequencing

- unidirectionally - sequencing from one end only
- less expensive, is typically reserved for quantifying gene expression in well-annotated genomes and analyzing small RNA molecules.

Paired-end sequencing



- bidirectionally - both ends of DNA fragments to be sequenced
- usually recommended for most applications as it provides richer data and permits longer library insert sizes.

https://www.genewiz.com/

Library preparation affects the downstream analysis

- RNA quality is fundamental
- RNA extraction method can affect the transcript abundance
- Choice of platform and library preparation protocol


GLADSTONE
INSTITUTES

Working material for this workshop

1. Single_read.fastq
2. Bacteria_GATTACA_L001_R1_001.fastq
3. Adapter_Sequence.fasta
4. rDNA_sequence.fasta
5. rDNA.gtf
6. all_steps_docker_desktop_mac.sh
7. all_steps_docker_desktop_windows.sh
8. Slides

Please install Docker https://docs.docker.com/get-docker/

Upcoming workshop at Gladstone:

- **May 19 |** Intermediate RNA Seq Analysis Using R

Fall workshops schedule - Data Science Training Program

Introduction to RNA-seq data analysis

Michela Traglia, Ayushi Agrawal
Bioinformatics Core, GIDB
May 12-13, 2025

GLADSTONE INSTITUTES

Introductions

Michela Traglia

Senior Statistician

Ayushi Agrawal

Bioinformatician III

Workshop sessions

Session I
Monday - May 12
1-4p

Session II
Tuesday - May 13
1-4p

DEG analysis - Intermediate workshop
Monday - May 19
1-4p

Poll 1

Which is your expectation from this workshop?
I want to:

1) Analyze my RNA-seq raw data

2) Improve my existing analysis pipeline settings

3) Learn about RNA-seq analysis

Please select all that apply.

Goals

By the end of this workshop you should be able to:

- Demystify each step of the RNA-Seq data analysis
- Understand the bioinformatic pipeline from raw data
- Enable informed conversations with computational biologists
- Demonstrate how to analyze data using *docker*

Workshop outline

Session 1

- RNA-seq experiments and protocols overview
- Understanding the sequencer output
- From sequencer output to FastQC
- From FastaQC to Trimming
- Mapping to reference genome

Break

- Introduction to docker and setup

Session 2 (tomorrow)

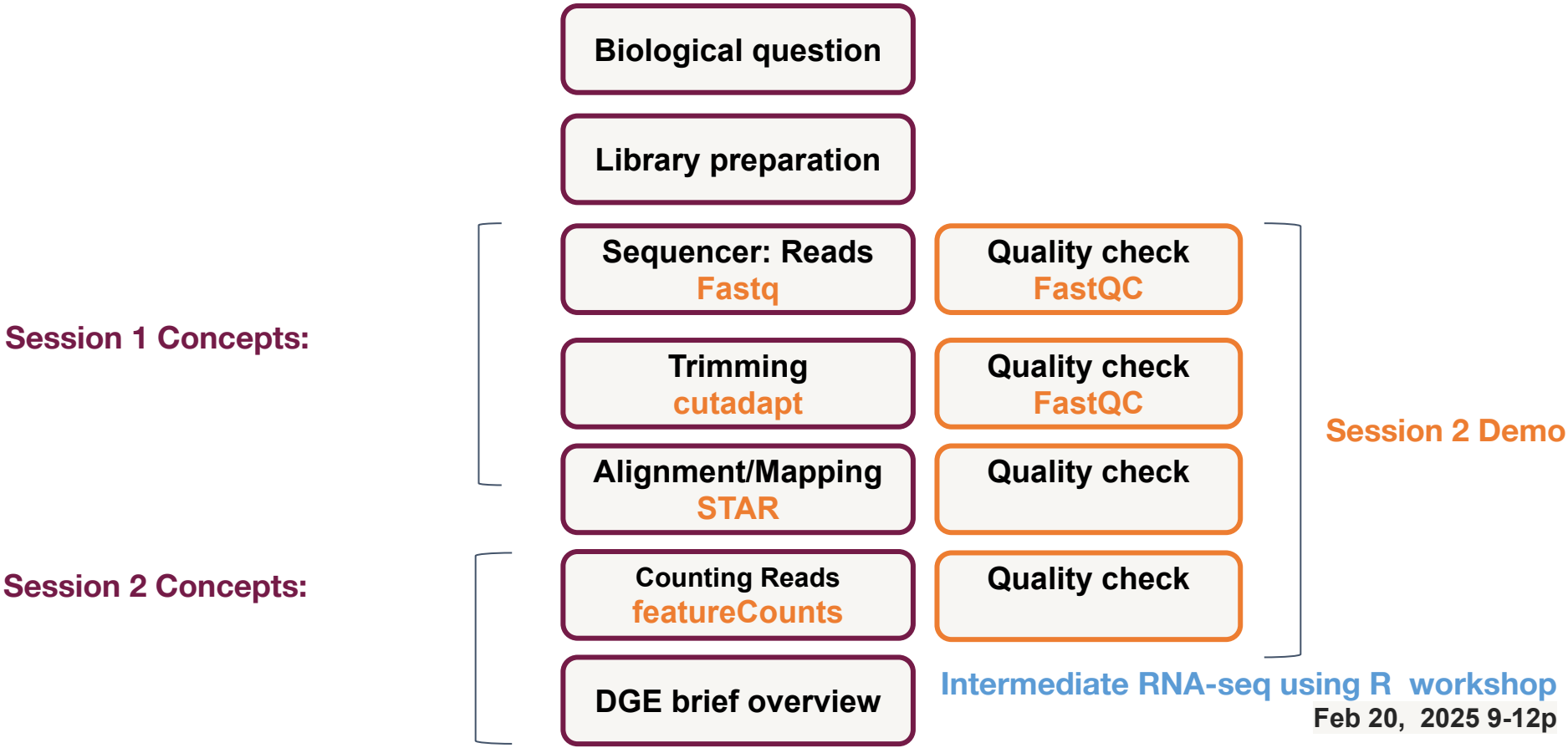
- Summarize steps so far
- From alignment to counting features

Break

- Demo
- Additional resources

Session 1

RNA-seq - analysis workflow



Bulk RNA-Seq

Averaged gene expression from a population of cells

Advantages and limitations

- A major breakthrough (replaced microarrays) in the late 00's and has been widely used since
- Useful for comparative transcriptomics, e.g. samples of the same tissue from different species
- Useful for quantifying expression signatures from ensembles, e.g. in disease studies
- Insufficient for studying heterogeneous systems, e.g. early development studies, complex tissues
- Does not provide insights into the stochastic nature of gene expression (fluctuations in mRNA)

Bulk RNA-Seq

Averaged gene expression from a population of cells

Applications

- **Qualitative:** identifying (**annotating**) or refining expressed transcripts, exon/intron boundaries, transcriptional start sites (TSS), and poly-A sites.

- **Quantitative:** measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation **between two or more treatments or groups.** (i.e. experiment to measure differential gene expression - DGE)

Experimental design is fundamental

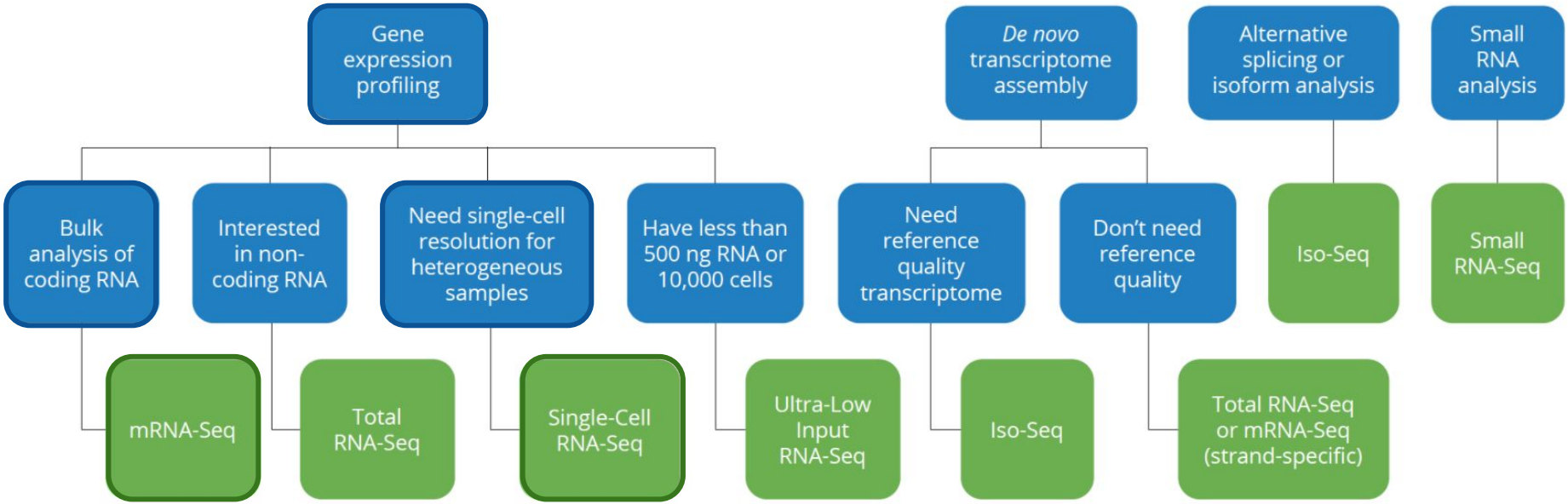
Which biological questions we want to answer

- *Coding* or *non coding* RNA?
- Why do you expect to find *differentially expressed genes* in the particular tissue?
- How many *tissue types* and/or *time points* to compare?
- What *types of genes* do you expect to find differentially expressed?
- What are the sources of *variability* from your samples?

More technical questions

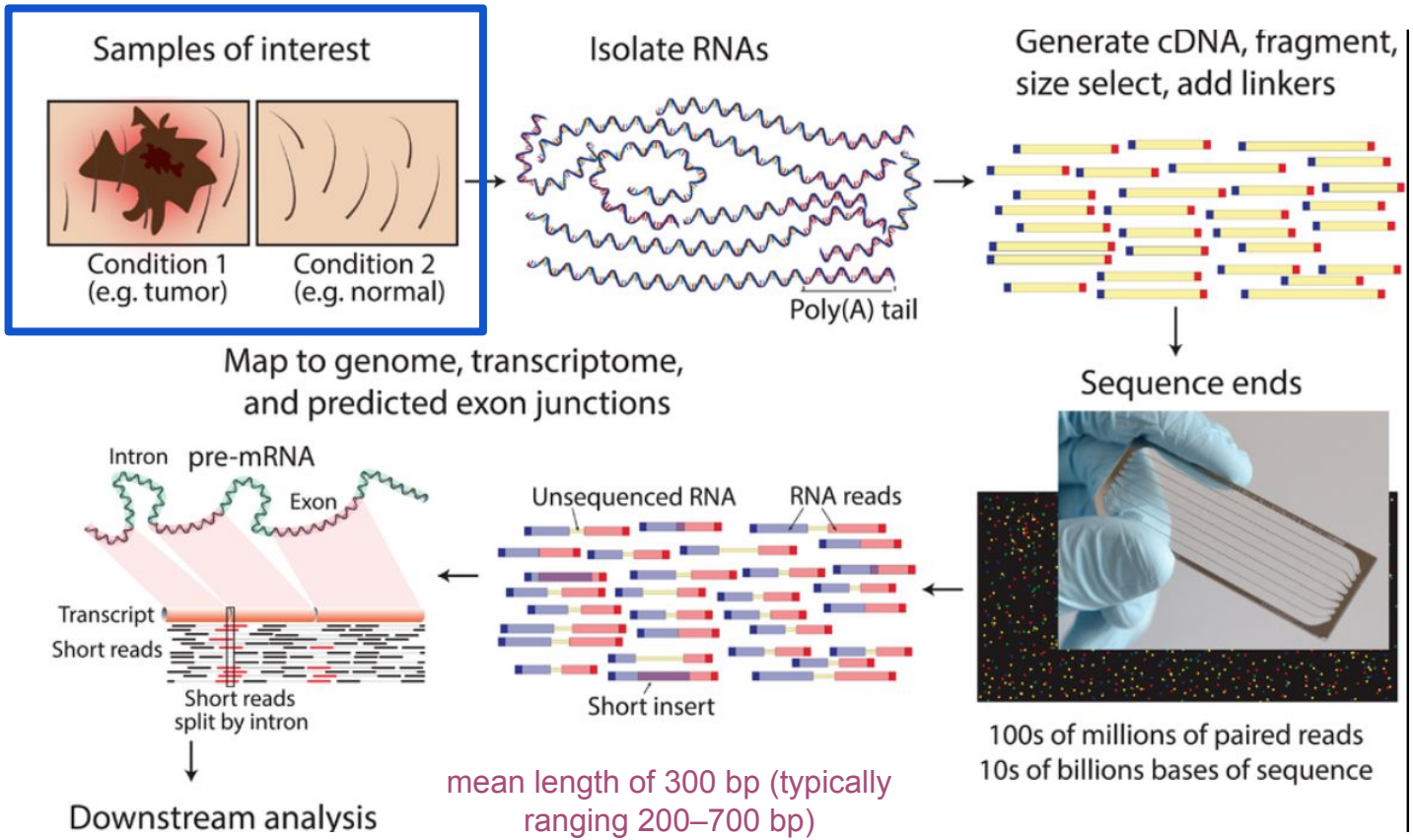
- Which sequencing platform?
- Depth of sequencing?
- Pooling samples?
- Biological replicates/technical replicates?
- Many others...

Which RNA-seq assay should I use?



From: Genewiz

RNA-seq - experiment

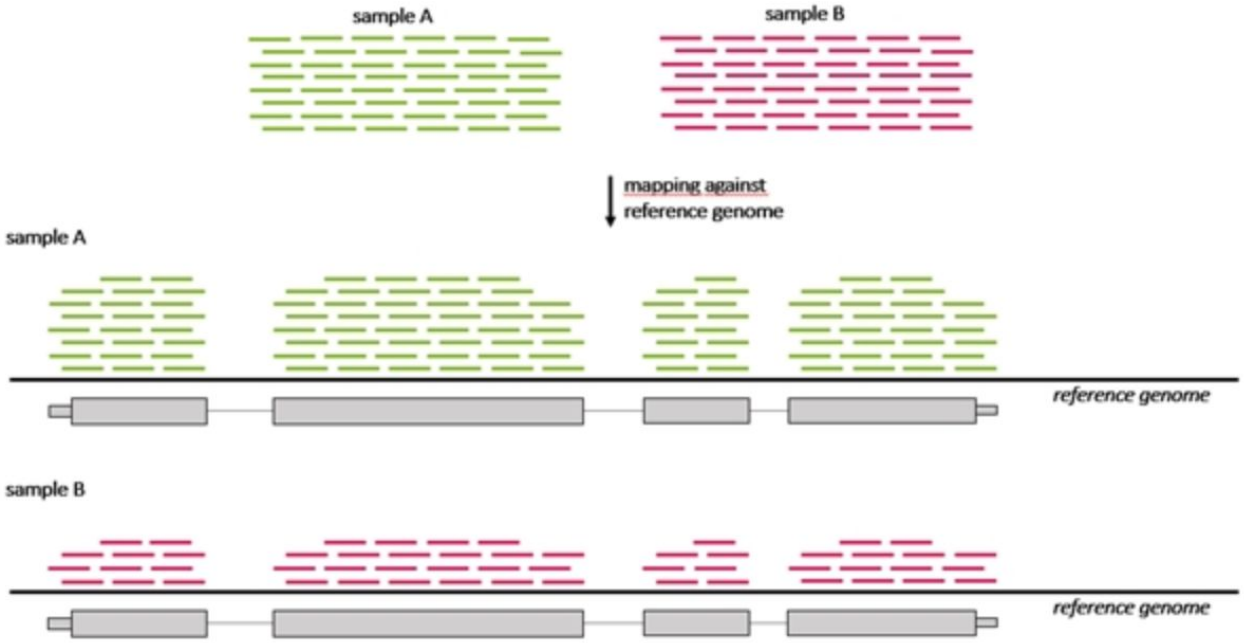


mean length of 300 bp (typically ranging 200–700 bp)

https://www.wikiwand.com/en/RNA-Seq

RNA-seq - differentially expressed genes (DGE)

Which genes are expressed at different levels between conditions (sample A and sample B)

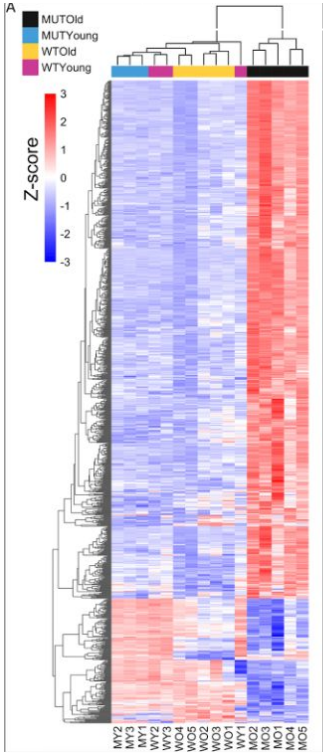


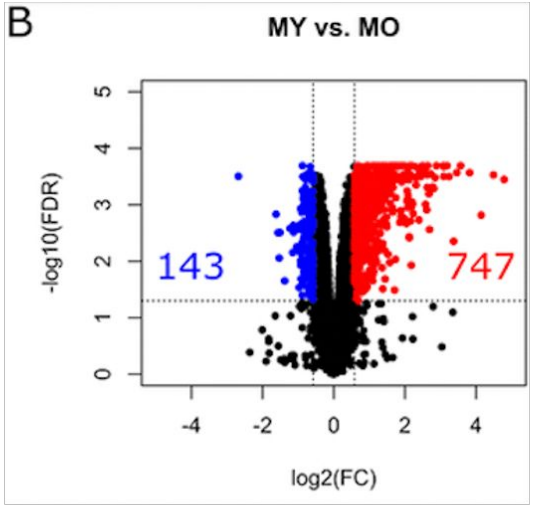
Source : seqme

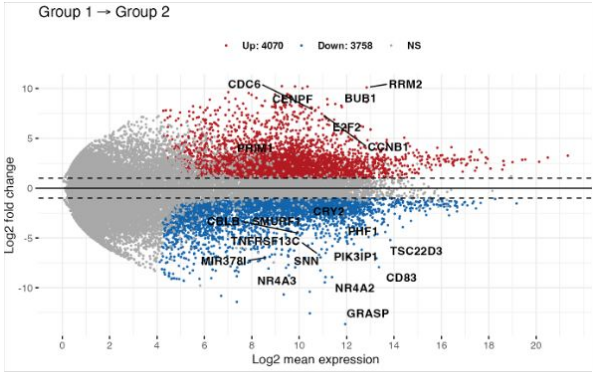
Many steps to calculate the DGE between sample A and B

Goal of the DEG analysis - Intermediate workshop

Identify genes (and molecular pathways) that are differentially expressed (DE) between two or more biological conditions







Bulk RNA-seq vs single cell (sc) RNA-seq

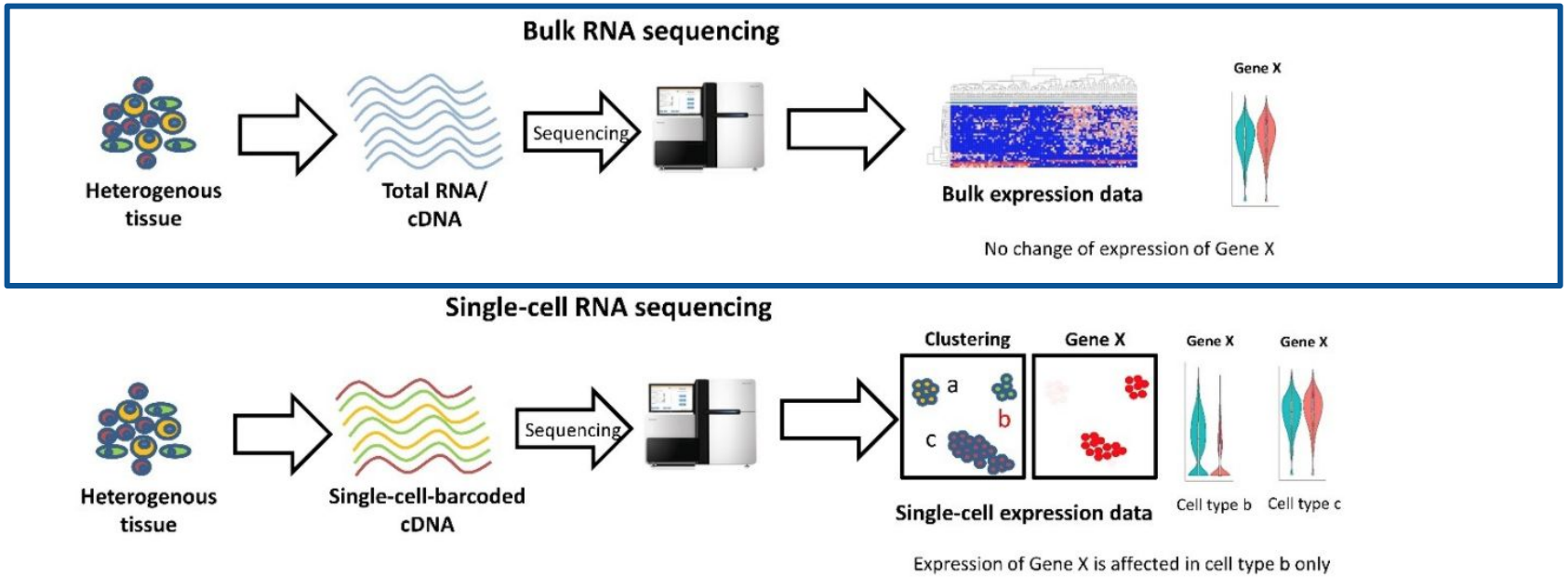


Figure 1. Bulk RNA sequencing vs Single-cell RNA sequencing. Image Credit: Dmitry Velmeshev.

Measures the ***average*** *expression level for each gene across a large population of input cells*

From: https://www.technologynetworks.com/genomics/articles/recent-advances-in-single-cell-genomics-techniques-324695

RNA-seq - analysis workflow

Biological question

Library preparation

Session 1 Concepts:

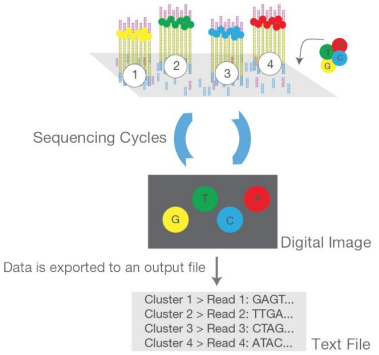
Sequencer: Reads Fastq	Quality check FastQC
Trimming cutadapt	Quality check FastQC
Alignment/Mapping STAR	Quality check

Session 2 Concepts:

Counting Reads featureCounts	Quality check
DGE brief overview	R

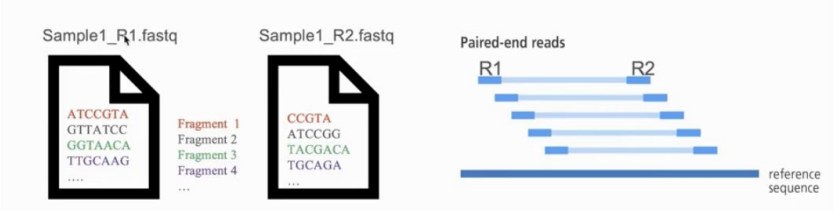
Demo

Naming conventions for fastq files



- File names often follow a format.
  - SampleName_SampleNumber_LaneNumber_ReadNumber_SetNumber.fastq
  - Eg – Bacteria_S1_L001_R1_001.fastq

- Paired-end reads named with R1 and R2 in file name.
  - Eg – Bacteria_GATTACA_L001_R1_001.fastq and Bacteria_GATTACA_L001_R2_001.fastq



- File extensions may be .fq or even .txt
- Often compressed using gzip.
  - gzip is free and open-source.
  - Resulting file names have .gz added. Example – .fq.gz.

FASTQ files are text files with detailed information about each read.

@A00564:60:HHJKFDMXX:1:1101:2031:1031 1:N:0:TGGCTTCA+CAACCACA
CGGACTGGTGGTATGCTGAGTACGTCCCAAGGGTATGGCTGTTCGCCATA
+
;;>@:=;=::@B;>A=<<=B?;;=@@?<?=B;A@=B?>=B:=B=@<<B:;

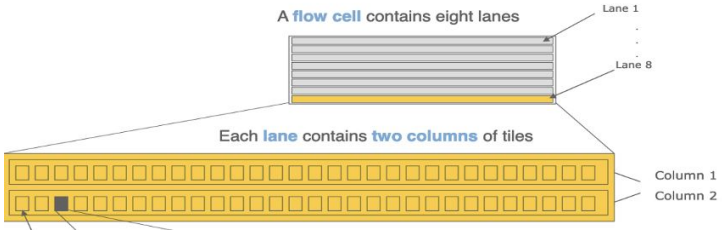
Label format in FastQ file

Sequence Header +Sequence ID

a b c d e f g h i j k

@HWI-ST486:166:C06K9ACXX:7:1101:1443:1995 1:N:0:ACAGTG

a. **unique instrument name**
b. run id
c. flowcell id
d. flowcell lane
e. tile number within the flowcell lane
f. x-coordinate of the cluster within the tile
g. y-coordinate of the cluster within the tile



h. the member of a pair, 1 or 2 (paired-end or mate-pair reads only)
i. Y if the read fails filter (read is bad), N otherwise
j. 0 when no control bits are on
k. index sequence

The more recent versions of Illumina software output a sample number (as taken from the sample sheet) in place of an index sequence (k).

Phred quality ccores

@A00564:60:HHJKFDMXX:1:1101:2031:1031 1:N:0:TGGCTTCA+CAACCACA
CGGACTGGTGGTATGCTGAGTACGTCCCAAGGGTATGGCTGTTCGCCATA
+
;;>@:=;=::@B;>A=<<=B?;;=@@?<?=B;A@=B?>=B:=B=@<<B:;

Quality score encoding based on ASCII table

Measurement of the likelihood that a base call in DNA sequencing is incorrect

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

Formula for getting PHRED quality from encoded quality:

$$Q = \text{ascii(char)} - 33$$

Example:

!+EJ

ASCII	33	43	69	74
-33	0	10	36	41

Phred quality score of each base

!	“	#	$	%	&	‘	(	)	*	+	,	-	.	/	0	1	2	3	4
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19

5	6	7	8	9	:	;	<	=	>	?	@	A	B	C	D	E	F	G	H	I
20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40

Each symbol is a probability p that the base call is incorrect.

The standard Sanger sequencing score to assess reliability of a base call is $Q=-10 \log_{10}(p)$

p, the probability that a given base is incorrectly called

Phred quality score calculation

$$Q = -10 * \log_{10}(P_{err})$$

Error probability (P_{err})	$\log_{10}(P_{err})$	Phred quality score
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40

Higher quality scores are better (>=20 is considered "good")

A quality score of 20 (Q20) represents an error rate of 1 in 100 (meaning every 100 bp sequencing read may contain an error), with a corresponding call accuracy of 99%.

FASTQ files are text files with detailed information about each read.

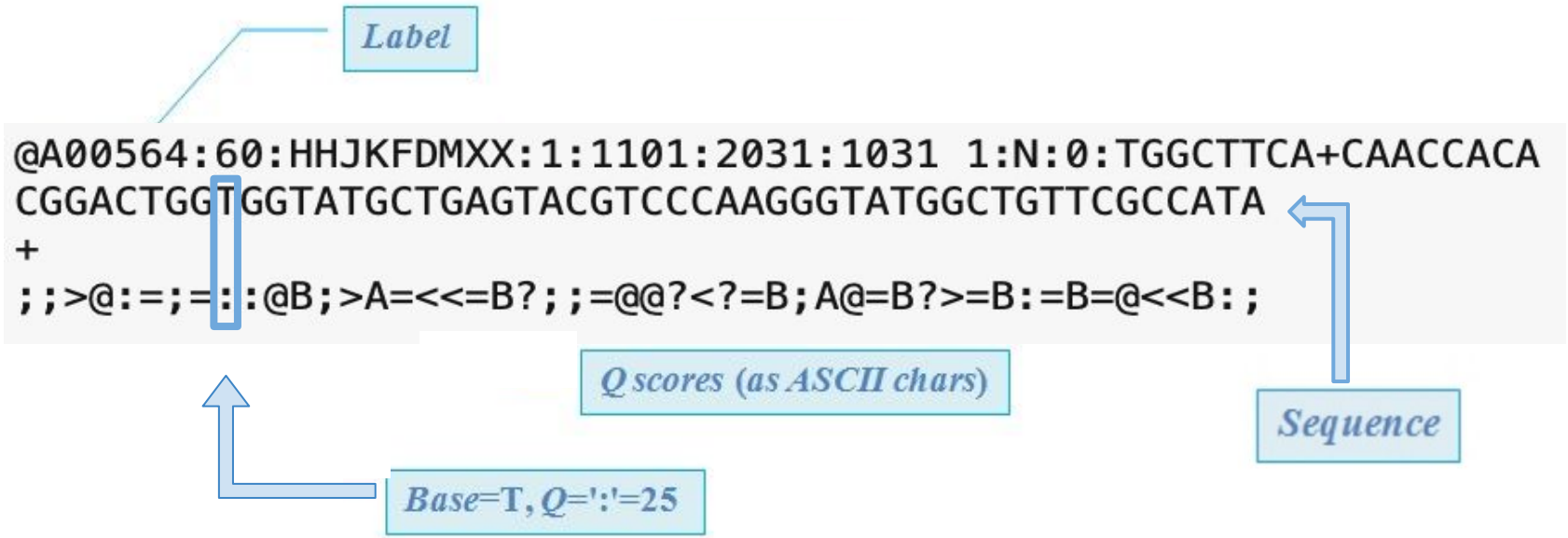
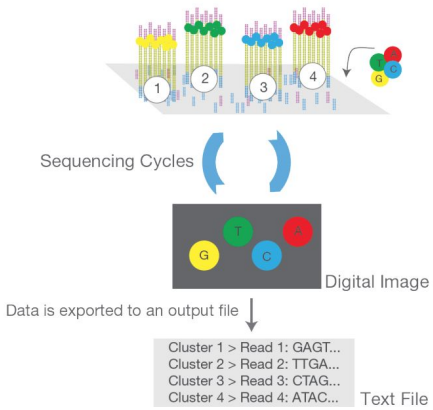


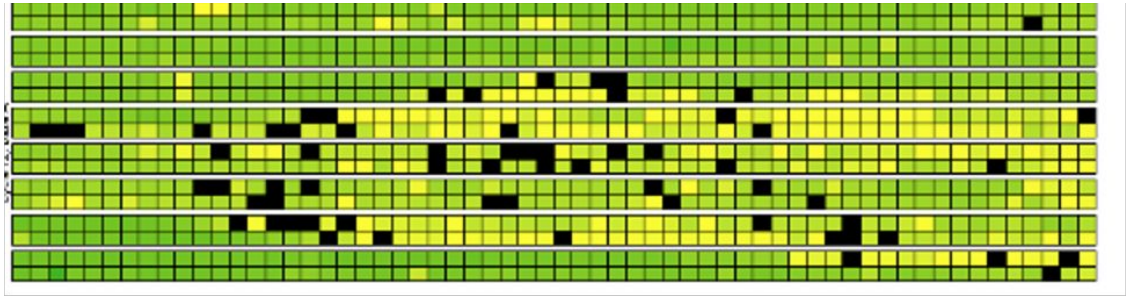
Image adapted from Robert Edgar's web site https://www.drive5.com/

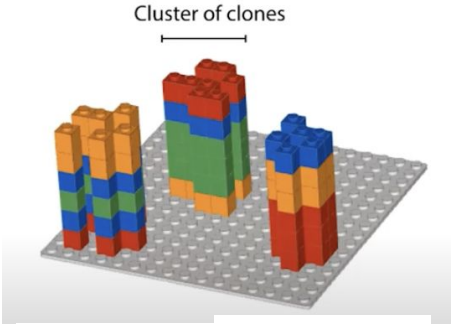
Base calling may not be accurate

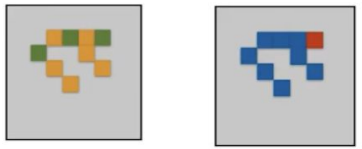
Possible causes

- Blocking of synthesis after one nucleotide addition may be inefficient.
- Clusters might not be monoclonal.
- A tile may be out of focus.
- Oil, reagent, etc. on flow cell or imaging component, etc.









=> Need to record quality of each base call.

Knowledge check - Poll 2

What is the flow cell id in the fastq file below?

@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATA
+
<>;##=><9=AAAAAAAAA9#

1. 15
2. FCX
3. #
4. SIM

Knowledge check - Poll 3

What is the Q-score (ASCII) of the 3rd base?

@SIM:1:FCX:1:15:6329:1045 1:N:0:2
TCGCACTCAACGCCCTGCATA
+
<>;##=><9=AAAAAAAAA9#

1. G
2. >
3. ;
4. I
5. None of the above

FastQC: Quality check of sequencing data

- Summarizes quality of base calls
- Any sequences more frequently observed than expected?
- Any sequence biases?
- Any GC biases?
- Checks for presence of known adapters

Examples of FastQC reports

Good Illumina data:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Bad Illumina data:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

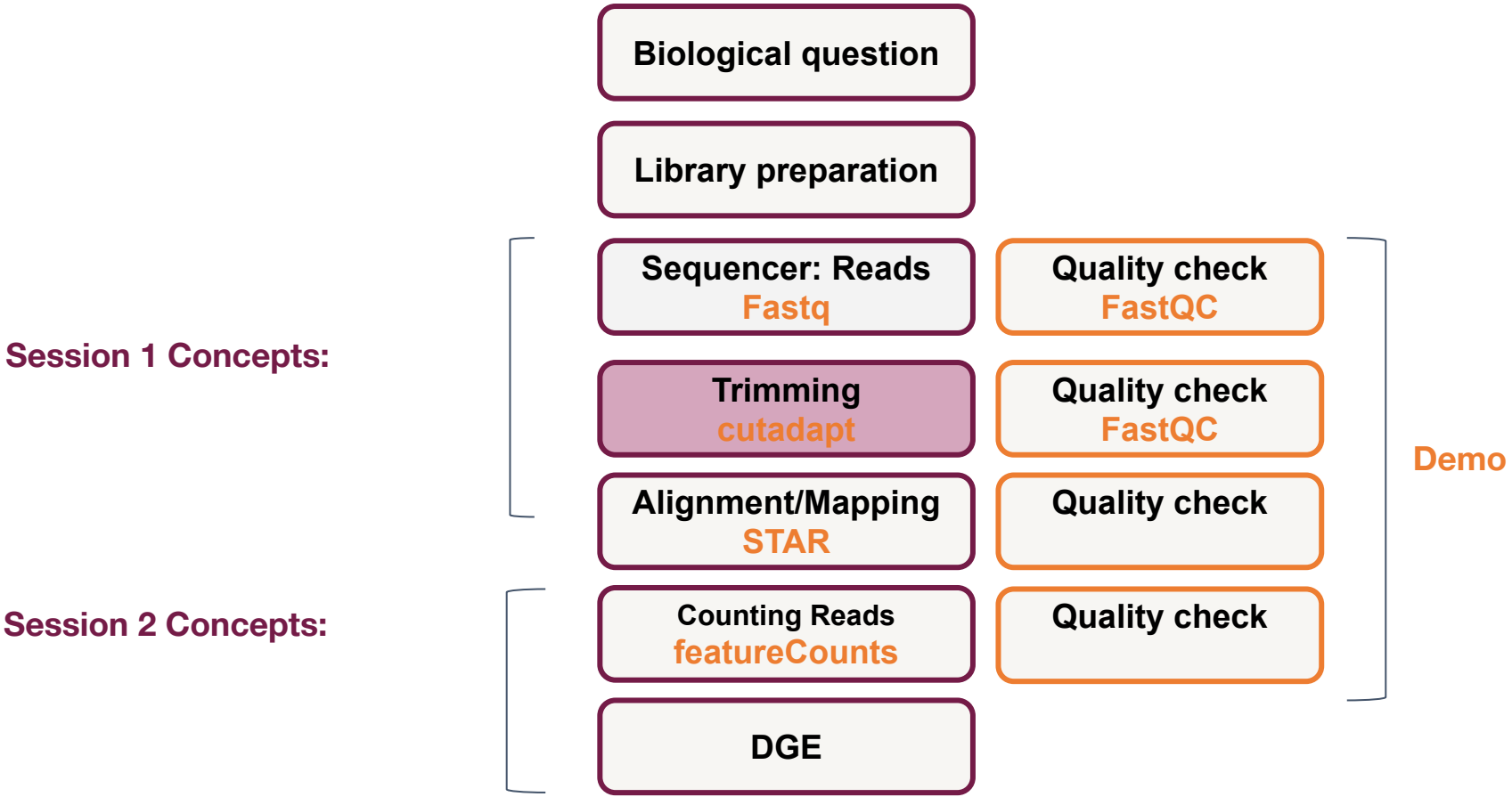
What if QC gives warn/fail flag?

- Non-normal GC content per read?
  - Normal expected for whole-genome shotgun sequencing.
  - RNA-seq might give different distributions.

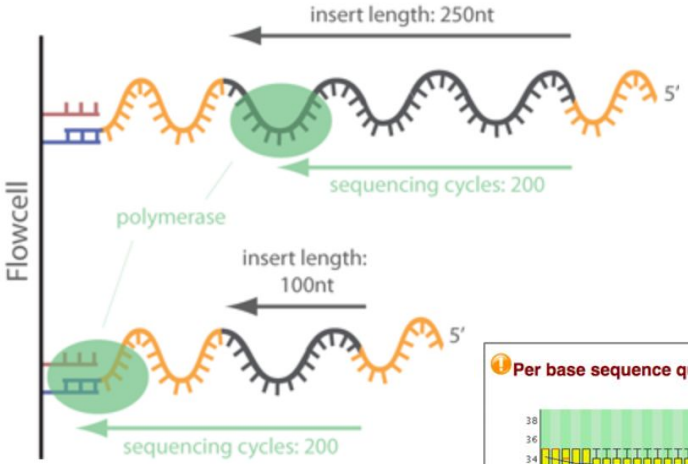
- Non-uniform sequence content per nucleotide?
  - First 10-15 nt in RNA-seq often non-uniform.

- High duplication levels or overrepresented sequences?
  - Are they contaminants, e,g. adapters or PCR duplicates?
  - If so, clean up contaminants.
  - Could be attributed to highly abundant transcripts.

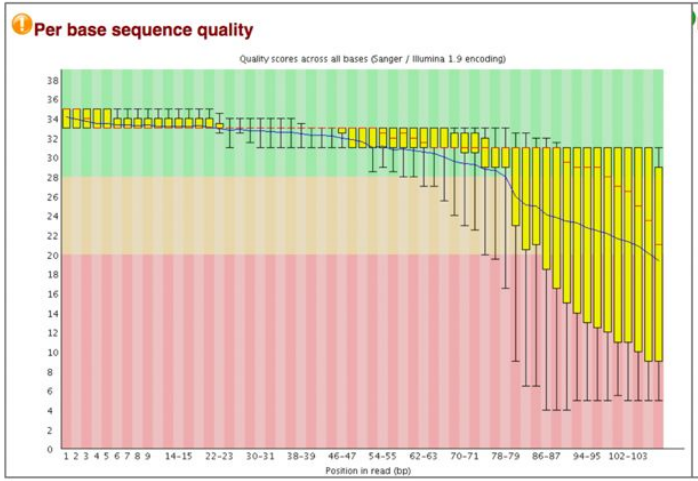
RNA-seq - analysis workflow



Trimming - Removing adapters



Before trimming



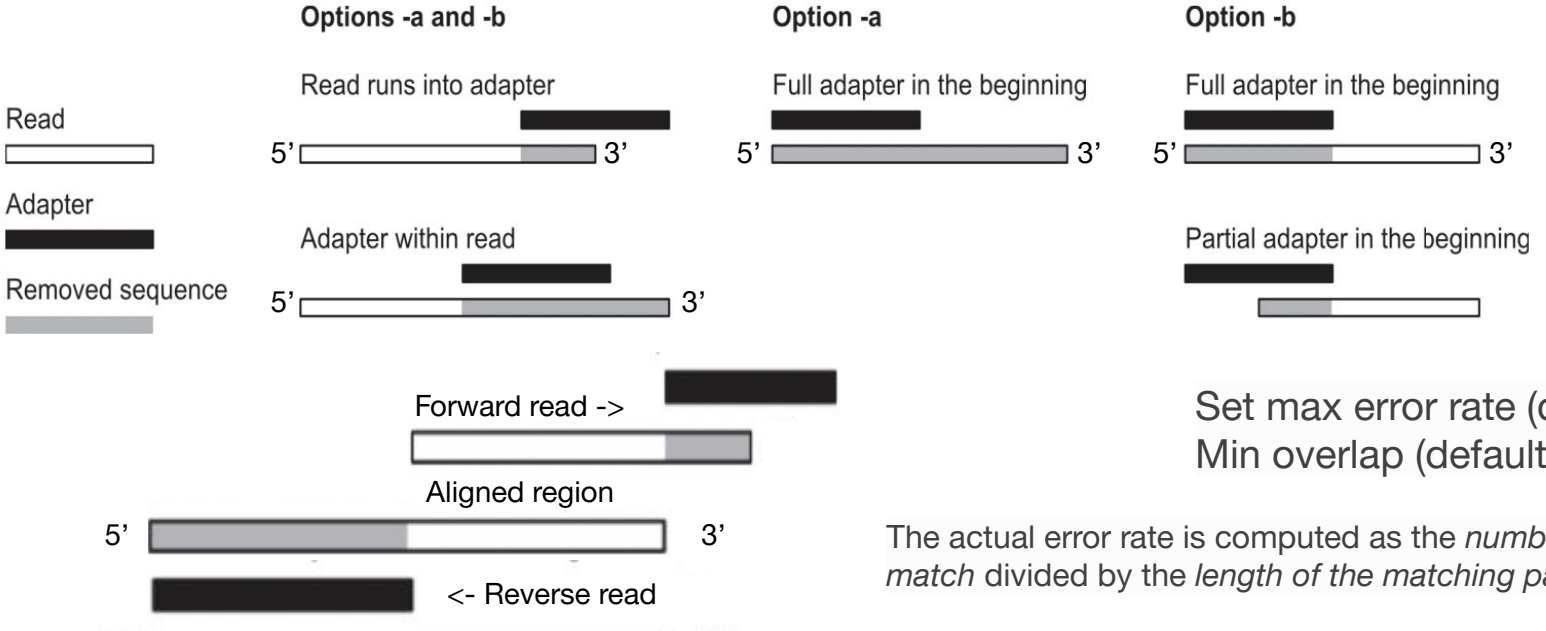
After trimming

Tools: cutadapt, Trimmomatic

cutadapt removes adapters

Full adapter sequence anywhere	acgtacgtADAPTERacgt
Partial adapter sequence at 3' end	acgtacgtacgtADAP
Full adapter sequence at 3' end	acgtacgtacgtADAPTER

- Search for adapter sequence in read.
- Allow for mismatches in sequence.
- If significant alignment, cut.



Set max error rate (default 1%)
Min overlap (default 3)

The actual error rate is computed as the *number of errors in the match* divided by the *length of the matching part of the adapter*.

Poll 4

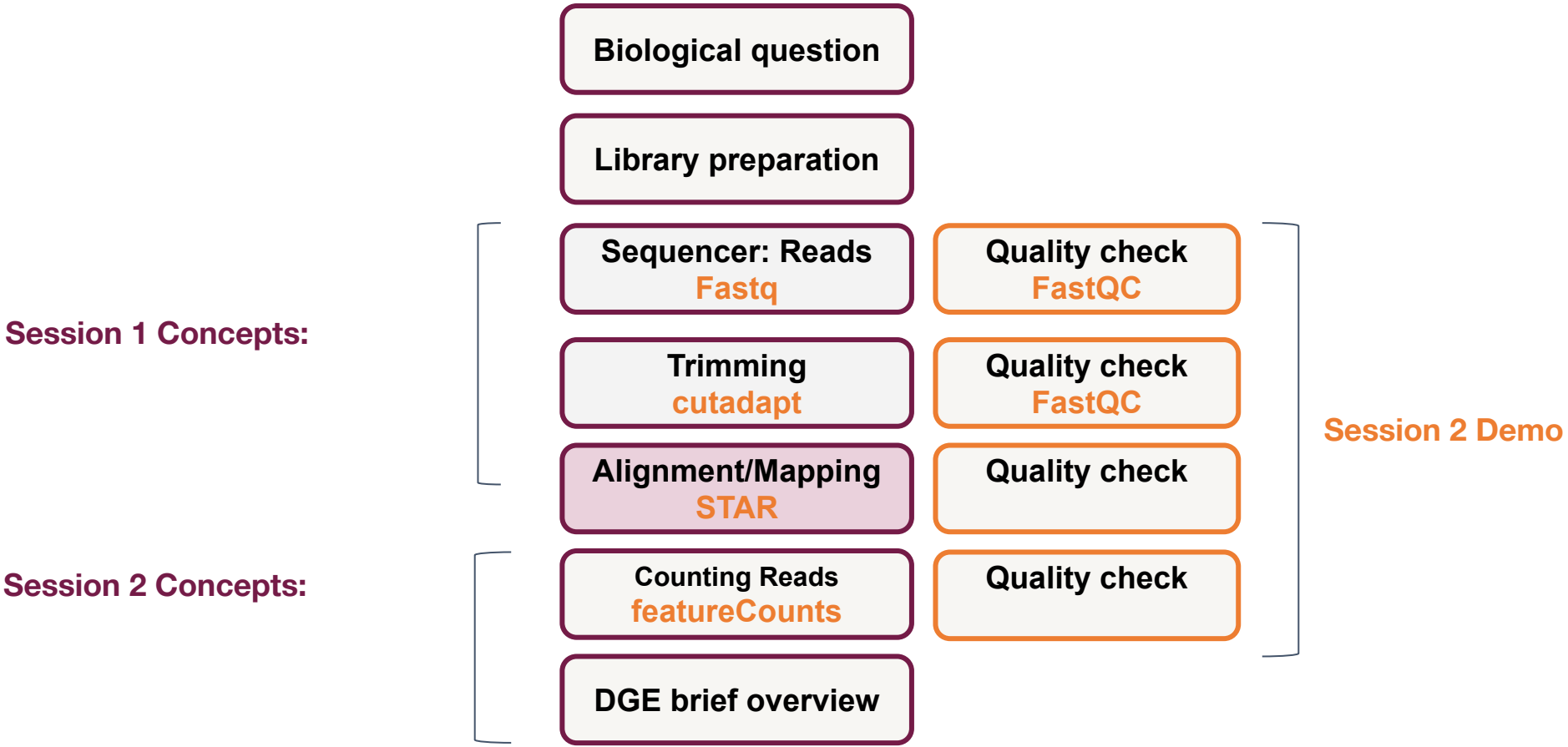
How are you feeling about the content covered so far?

1. Great - I am getting it
2. Okay - I am hanging in there
3. I am learning a lot
4. I need help

Break

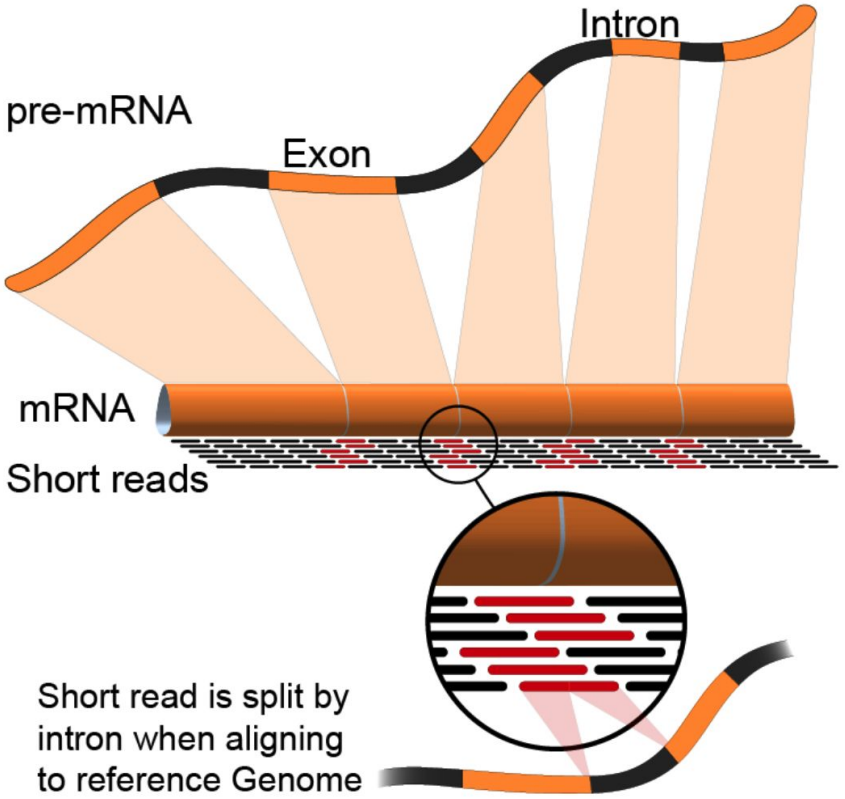
Please drop a question in the chat or speak up

RNA-seq - analysis workflow



Alignment to genome : challenges

- Reads from junctions: one part of it maps to one exon and the other half maps to the other exon
- Reference sequences can be very long (~3 billion bp for humans).
- Order of 100 million reads to be mapped.
- Need to account for splicing.
- Allow for PCR artifacts/sequencing errors.



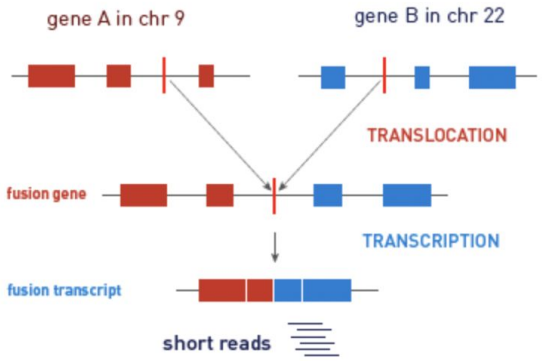
Alignment/Mapping: STAR tool

STAR (Spliced Transcripts Alignment to a Reference (STAR))

is popular for RNA-Seq data because it

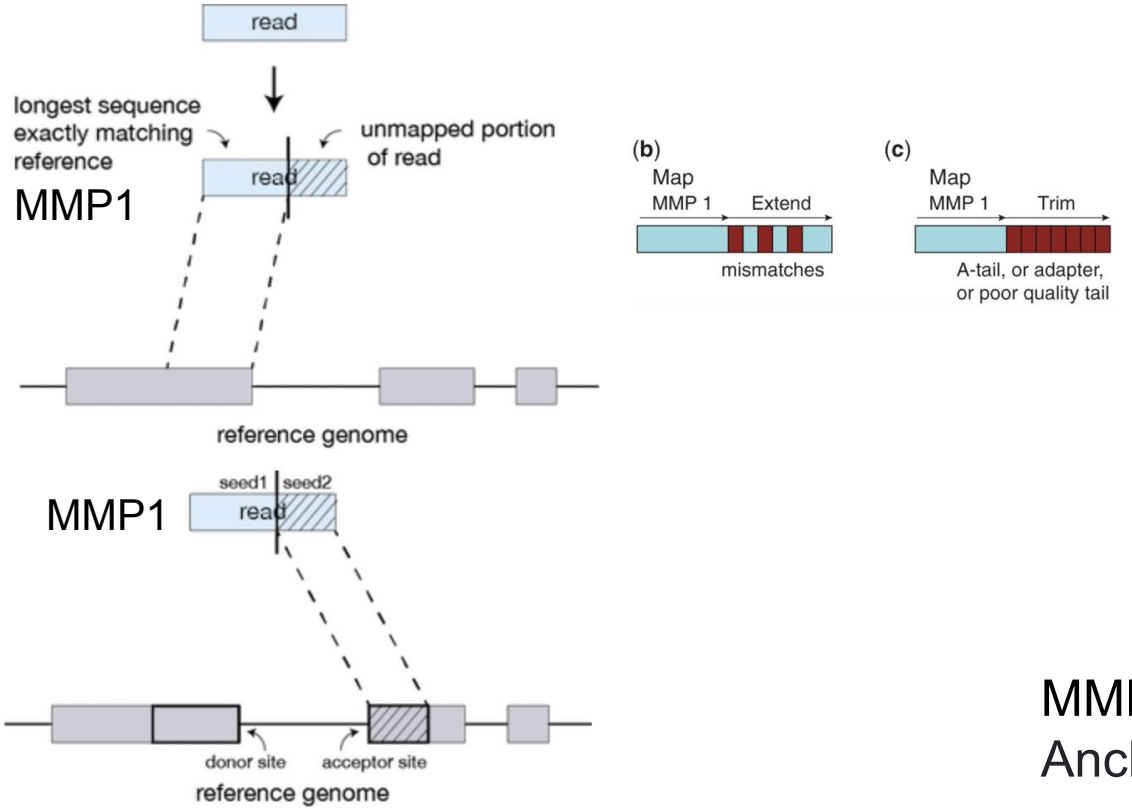
- does unbiased de novo detection of canonical junctions
- can discover non-canonical splices
- can discover chimeric (fusion) transcripts

Tools: STAR, HISAT2, TopHat2, bowtie2, salmon, kallisto

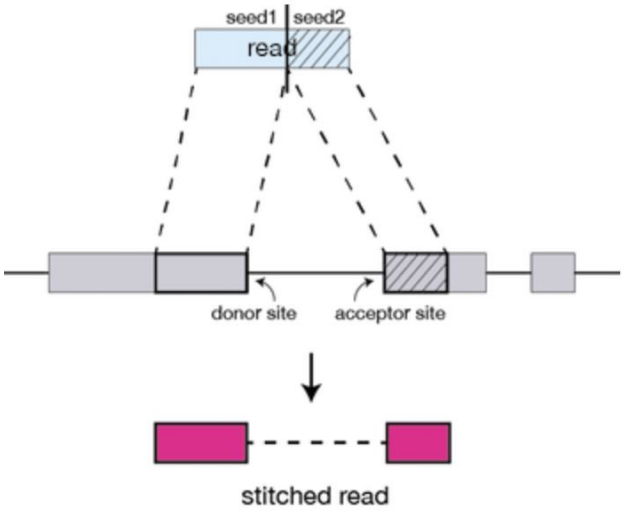


Alignment/Mapping: How does STAR work?

Step 1: Seed search



Step 2: Stitching and scoring



MMP = Maximum Mappable Prefix
Anchor seeds in a genomic region

Image source: https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

Alignment/Mapping: Inputs needed

- Reads to align
  - FASTQ file after cleaning (trimming adapters).

- Reference sequence to align to
  - Example – “rDNA_sequence.fasta”
  - FASTA format. Two lines per sequence.
    - Starting with “>”, followed by sequence name/identifier.
    - Sequence.
  - File extensions: .fasta, .fa, .txt.
  - Examples of acceptable genome sequence files for STAR:
    - ENSEMBL: files marked with *.dna.primary.assembly*
    - GENCODE: files marked with *PRI* (primary). Strongly recommended for mouse and human.

Understanding STAR output

1. Alignments in SAM format

2. Summary of mapping statistics
    - How many reads mapped?
    - How many unmapped?
    - ...

Sequence Alignment/Map (SAM) format

- Open with Excel.
- First few lines contain metadata about alignments.
  - These lines start with “@”.
  - Example – version of file format, sorting order of alignments, grouping, etc.
- After header, a table of alignments of each read to the genome.
- Alignment reports often very large files.
- Binary Alignment/Map (BAM) extension used for compressed SAM files.
- Indexed BAM -> BAI

From the alignment to SAM format

Alignment

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGGCAT
```

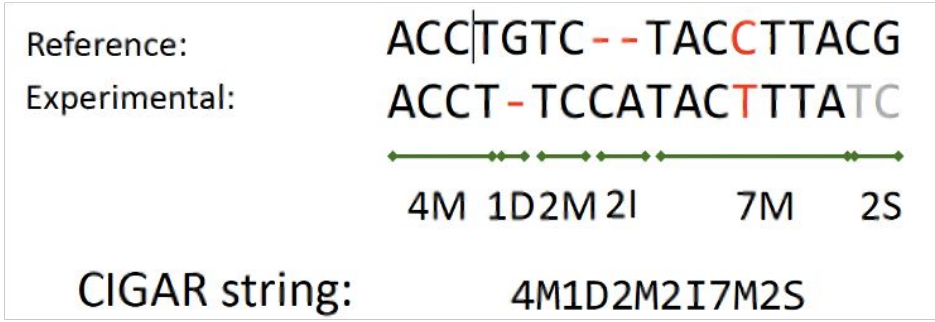
$$MAPQ = -10*\log_{10}(P_{map\_loc\_wrong})$$

SAM

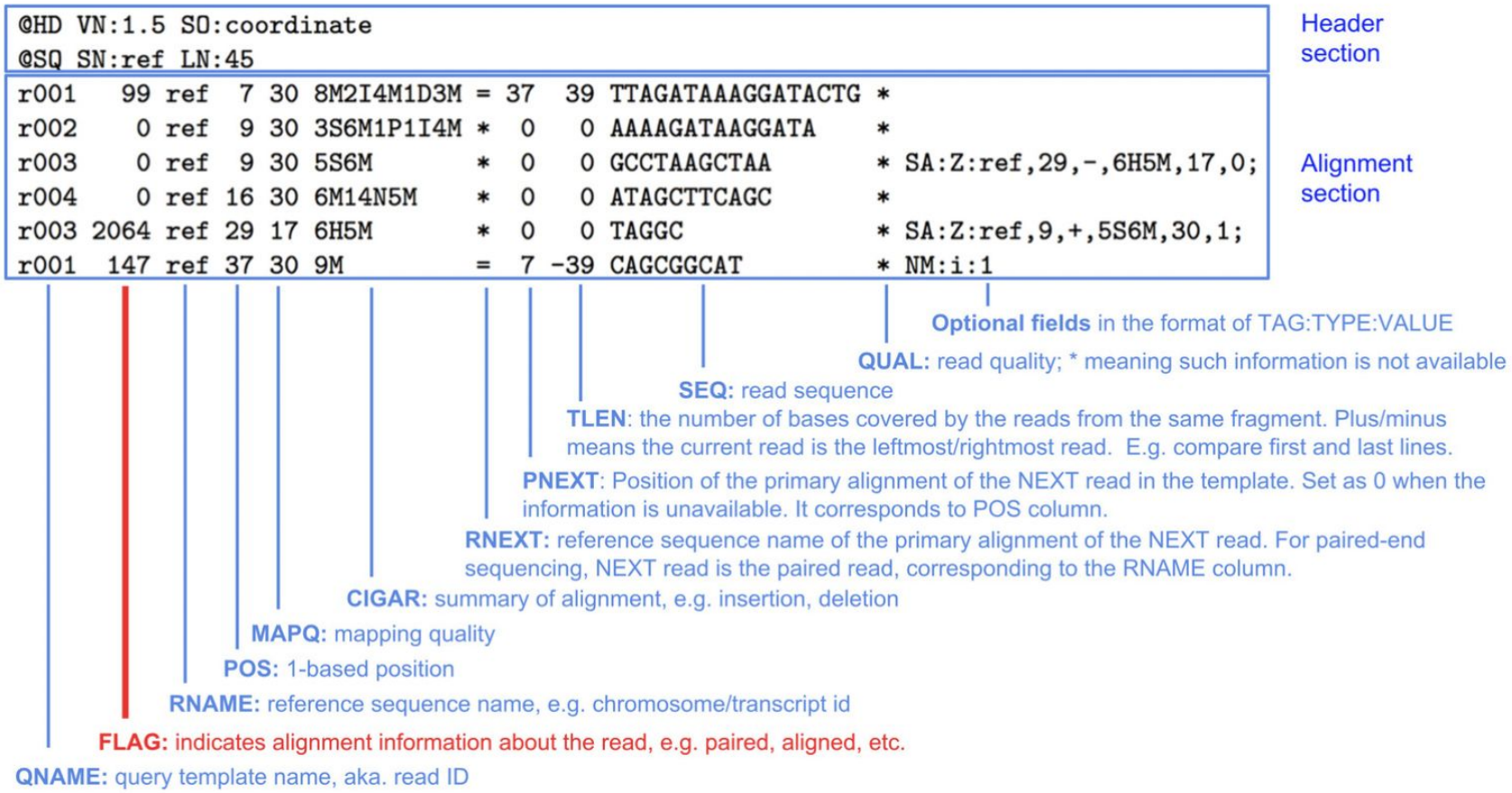
```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC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

The CIGAR string: encode the details of the alignment

Operation	Meaning
M	Match
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding



SAM format


@HD VN:1.5 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 section
Alignment section
Optional fields in the format of TAG:TYPE:VALUE
QUAL: read quality; * meaning such information is not available
SEQ: read sequence
TLEN: the number of bases covered by the reads from the same fragment. Plus/minus means the current read is the leftmost/rightmost read. E.g. compare first and last lines.
PNEXT: Position of the primary alignment of the NEXT read in the template. Set as 0 when the information is unavailable. It corresponds to POS column.
RNEXT: reference sequence name of the primary alignment of the NEXT read. For paired-end sequencing, NEXT read is the paired read, corresponding to the RNAME column.
CIGAR: summary of alignment, e.g. insertion, deletion
MAPQ: mapping quality
POS: 1-based position
RNAME: reference sequence name, e.g. chromosome/transcript id
FLAG: indicates alignment information about the read, e.g. paired, aligned, etc.
QNAME: query template name, aka. read ID

Alternative tools

- Many. Example – HISAT2, TopHat2, bowtie2, salmon, kallisto, etc.
- Differences in speed and memory requirement.
- Pros and cons of each:
  - Example: Some handle spliced alignment, others do not.
  - …

Comparison across tools

Pseudo-aligners

- Much, much faster
- Less memory intensive than STAR
- They are called "pseudo" aligners because they do not perform the full alignment of each sequencing read to a reference genome or transcriptome.
- Pseudoaligners use a reference transcriptome to create a set of potential transcript sequences or exonic regions, respectively. Then, they map the reads to these sequences without attempting to align each read to its exact position in the reference.
- Examples of tools: Salmon, Kallisto

Alignment tool summary

Be aware of the downstream analysis that you need!
Convert genomic to transcriptomic coordinates - https://github.com/OceanGenomics/mudskipper

Adapted from https://www.nature.com/articles/s41598-020-76881-x/figures/1

RNA-seq - analysis workflow

Session 1 Concepts:

Session 2 Concepts:

Biological question		
Library preparation		
Sequencer: Reads Fastq	Quality check FastQC	Demo
Trimming cutadapt	Quality check FastQC	
Alignment/Mapping STAR	Quality check	
Counting Reads featureCounts	Quality check	
DGE		

Bioinformatics software ecosystem

- Tools that “do one thing, and do it well”.
- Tools for this workshop: fastqc, cutadapt, STAR, featureCounts
  - Available via docker hub
  - Some are pre-installed on Wynton; others we can install ourselves
  - Download a *container* with all tools installed; use anywhere
  - Everything can be installed on a laptop

Different platforms for computing

- Web-based platforms
- Graphical User Interface
- Command Line Interface
- etc..

Galaxy: Open source, web-based platform that integrates many tools.

- Free, public, internet accessible resource.
  - https://usegalaxy.org/
- Data transfer and data storage are not encrypted.
  - DO NOT UPLOAD PROTECTED DATA!!!

Command line interfaces allow scripting

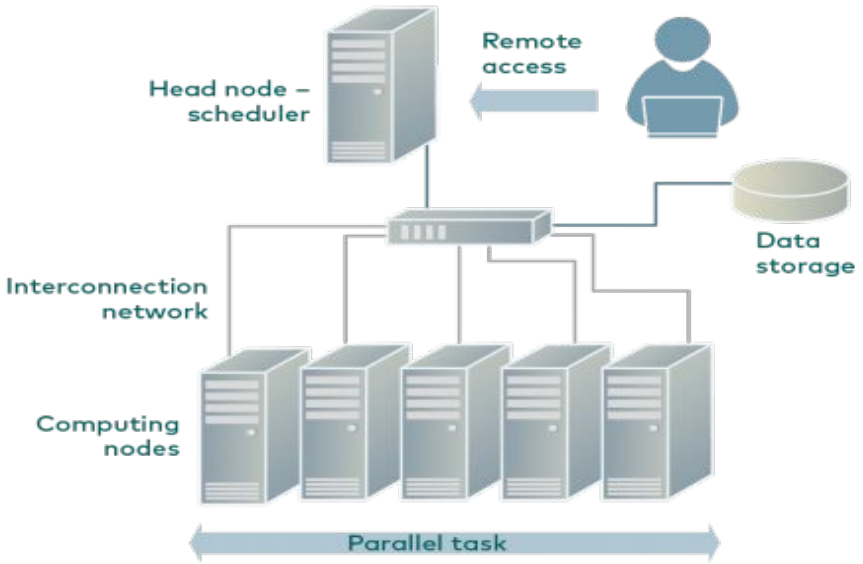
- **Graphical User Interface**
  - Consists of windows, icons, menus, pointers
  - Not always available for bioinformatics

- **Command Line Interface**
  - Text based
  - Allow automation by scripting
  - Examples
    - Wynton CLI
    - MacOS: Terminal
    - Windows: Command Prompt, PuTTY

Wynton is a high-performance computing (HPC) system for UCSF affiliates

- How to access Wynton? Visit: https://wynton.ucsf.edu/hpc/get-started/access-cluster.html

- Most universities/ institutions doing data-intensive science have HPC cluster on campus.



(see Description for image source)

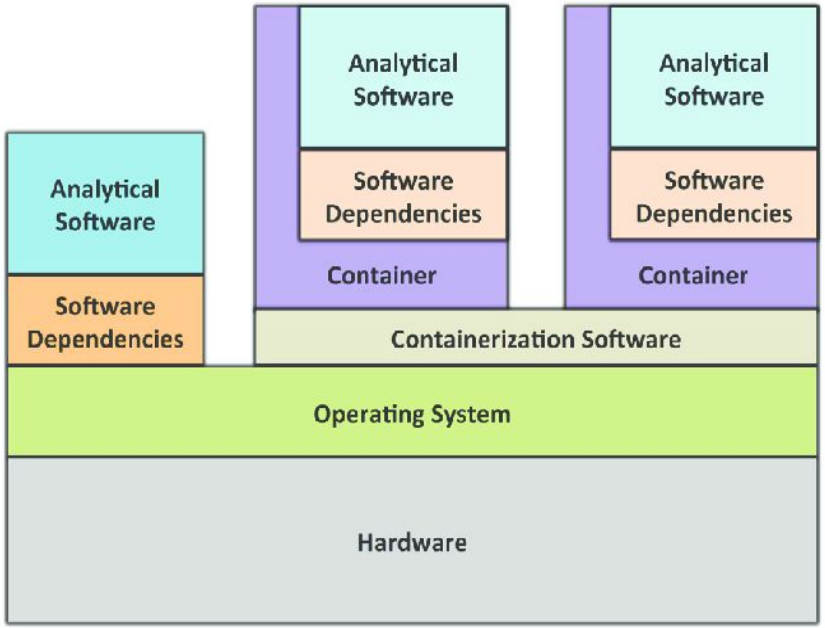
Containers enable reproducibility

- Tools are constantly under development
  - => Many versions around
- Dependencies complicate installations
  - Dependencies are also constantly under development
  - => Many versions around
- Different labs use different programming languages

"A little big, but they'll shrink after a few washings."

A container is like a computer within a computer

- Containerization software examples:
  - Singularity
  - Docker (has security issues in the context of HPC)
- Containers can be deployed on commercial cloud computing platforms, e.g., Amazon Web Services



(see Description for image source)

Best to use singularity with Linux (currently)

- Limited support for MacOS
- Even more limited support for Microsoft Windows

Docker Container Lifecycle

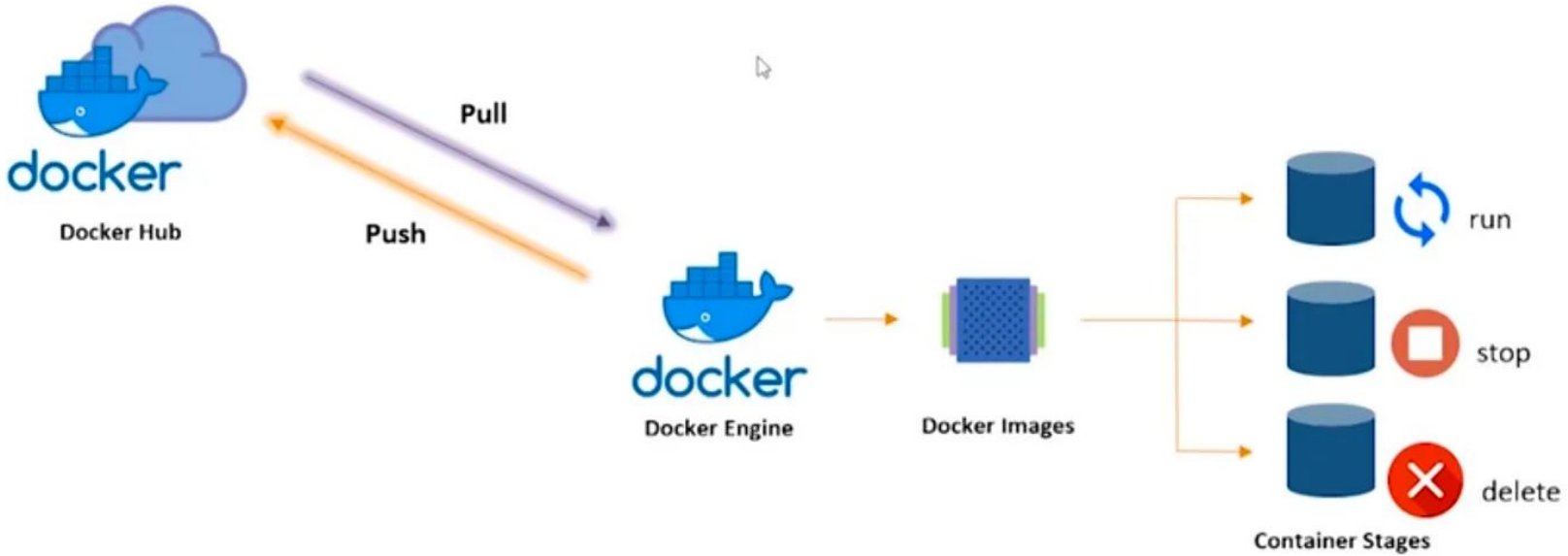


Image source: https://pradeep-sg406.medium.com/introduction-to-docker-5212421782df

Poll 5

Which operating system will you use for the hands-on section?

1. Mac - Intel chip
2. Mac - Apple chip
3. Windows
4. Linux

- **Docker desktop download and installation**
  (https://www.docker.com/products/docker-desktop/)
  (https://docs.docker.com/desktop/install/windows-install/)

- **How to access public datasets?**
  - Downloading data from GEO
    - https://bioinformaticsworkbook.org/dataAcquisition/fileTransfer/sra.html#gsc.tab=0
    - https://erilu.github.io/python-fastq-downloader/

Session 1: Take-home messages

1) Define your hypothesis and datasets before planning RNA-seq experiment

2) Each steps of the analysis can be affected by some kind of bias - Check the quality after each step!

3) Be familiar with file formats

Tomorrow:

- Start with summarizing the steps so far
- Understand the alignment output
- Feature counts
- Demo step by step

If you are not able to attend tomorrow, please take the survey:
https://www.surveymonkey.com/r/F75J6VZ

Introduction to RNA-seq data analysis

Michela Traglia, Ayushi Agrawal
Bioinformatics Core, GIDB
May 12-13, 2025

GLADSTONE INSTITUTES

Working material for this workshop

1. Single_read.fastq
2. Bacteria_GATTACA_L001_R1_001.fastq
3. Adapter_Sequence.fasta
4. rDNA_sequence.fasta
5. rDNA.gtf
6. all_steps_docker_desktop_mac.sh
7. all_steps_docker_desktop_windows.sh
8. Slides

Please install Docker https://docs.docker.com/get-docker/

Workshop outline

Session 1

- RNA-seq experiments and protocols overview
- Understanding the sequencer output
- From sequencer output to FastQC
- From FastaQC to Trimming
- Mapping to reference genome

Break

- Introduction to docker and setup

Session 2 (day2)

- Summarize steps so far
- From alignment to counting features

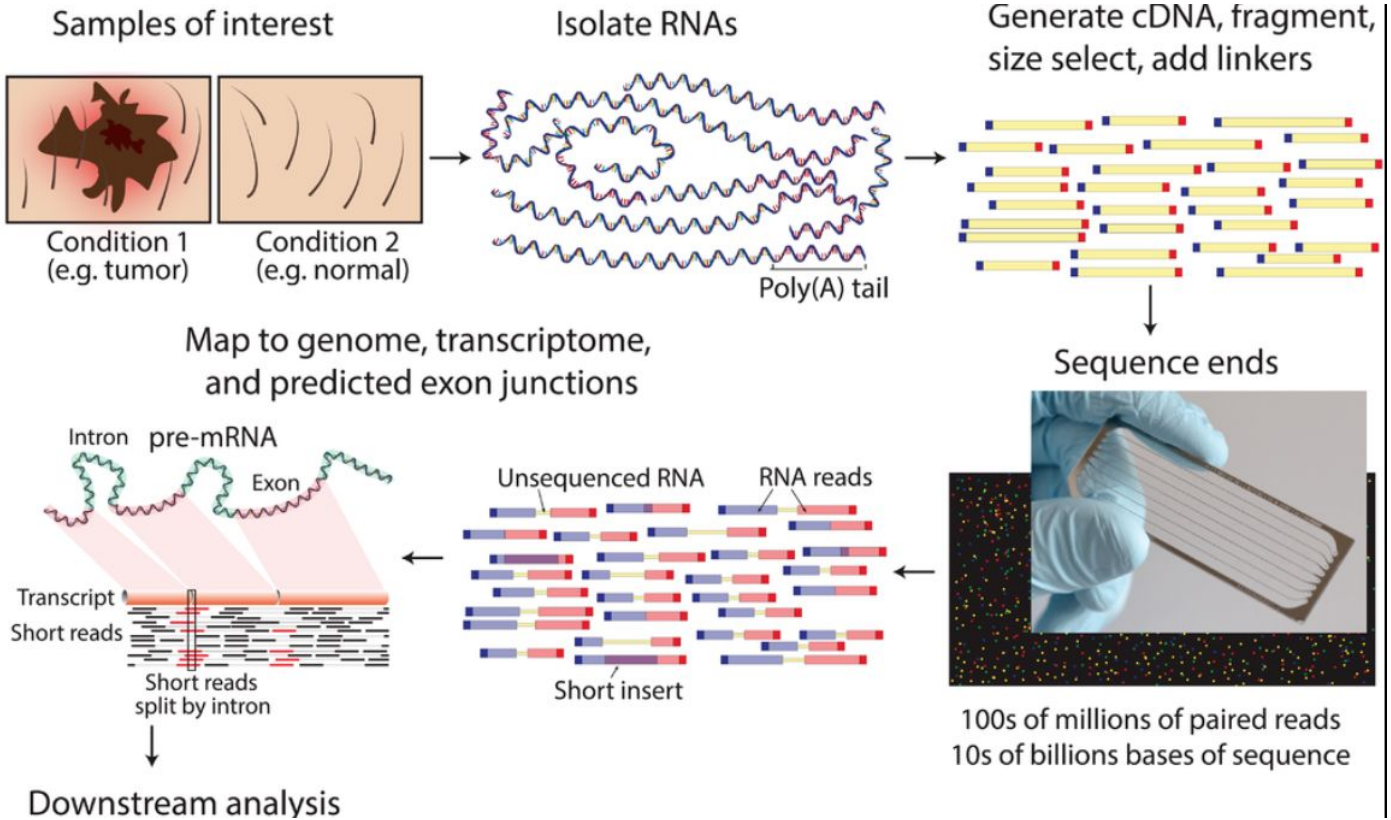
Break

- Demo
- Additional resources

Session 2

Tools: Docker, FastaQC, cutadapt, STAR, featureCounts

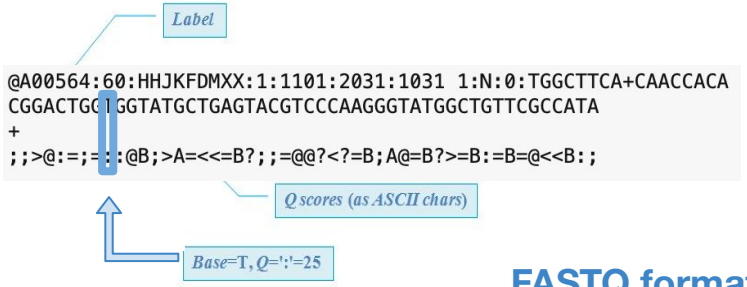
RNA-seq - summary



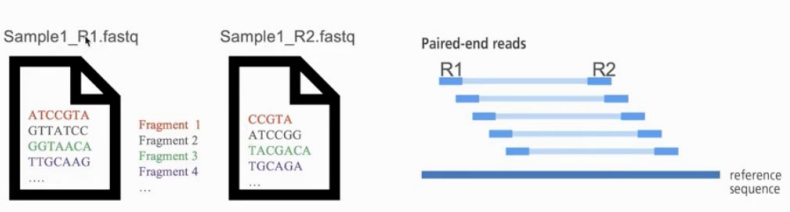
Single_read.fastq

Bacteria_GATTACA_L001_R1_001.fastq

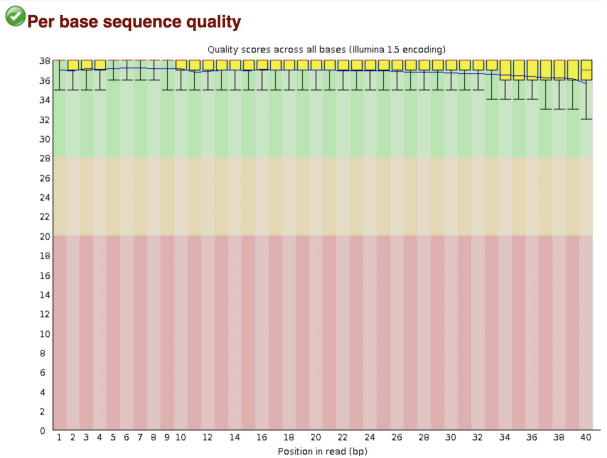
Bioinformatic pipeline - summary



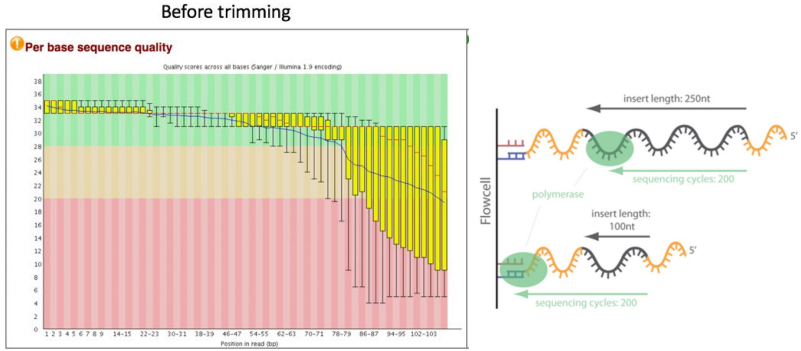
FASTQ format



FASTQ file

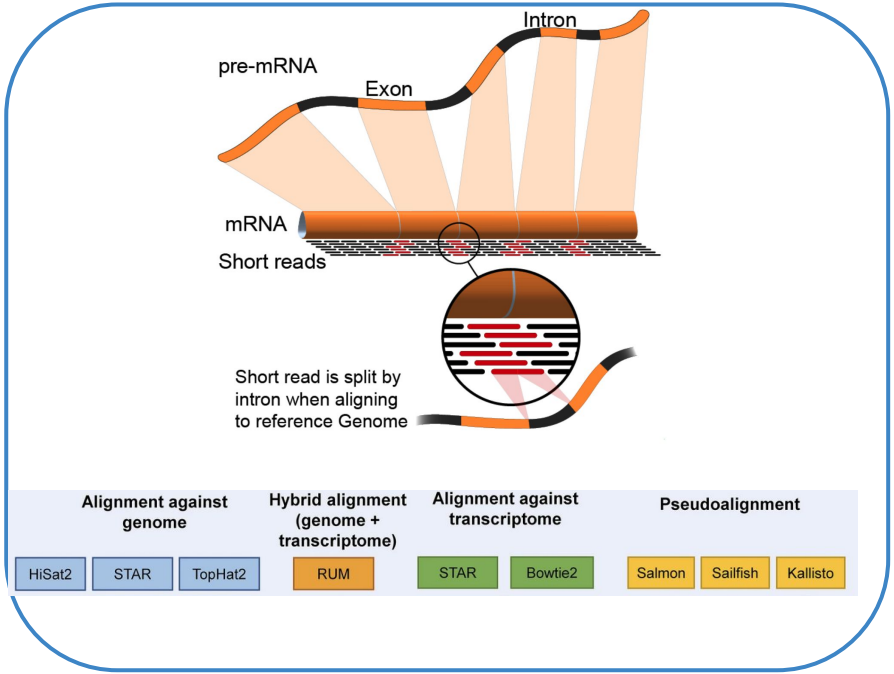


FastQC - good experiment

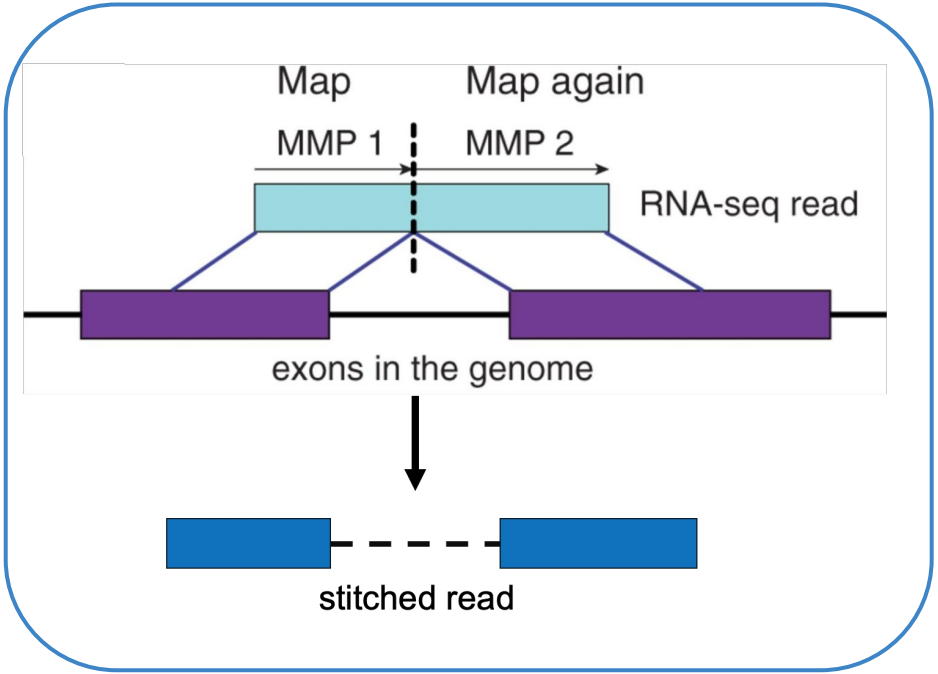


FastQC - bad experiment needs trimming - cutadapt

Bioinformatic pipeline - summary



Alignment and pseudo-alignment tools



How STAR works

Knowledge check

Which information do you find in a SAM/BAM file?

1. Sequences, like a FASTQ file
2. Location of the read on the chromosome
3. Mapping quality
4. All of the above

featureCounts

- Input:
  - Alignment BAM file
  - Annotation SAF/GFF/GTF file

GeneID	Chr	Start	End	Strand
497097	chr1	3204563	3207049	-
497097	chr1	3411783	3411982	-
497097	chr1	3660633	3661579	-
100503874	chr1	3637390	3640590	-

SAF

GTF

```
chr1 unknown  exon         113217048  113217252  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon         113217048  113217351  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon         113217470  113217671  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS          113217535  113217671  .  +  0  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

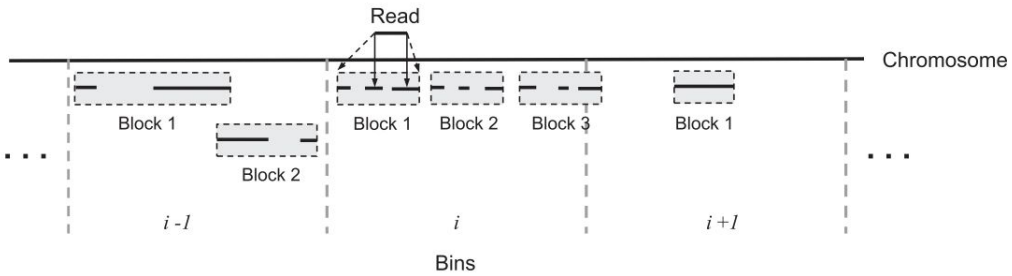
feature type

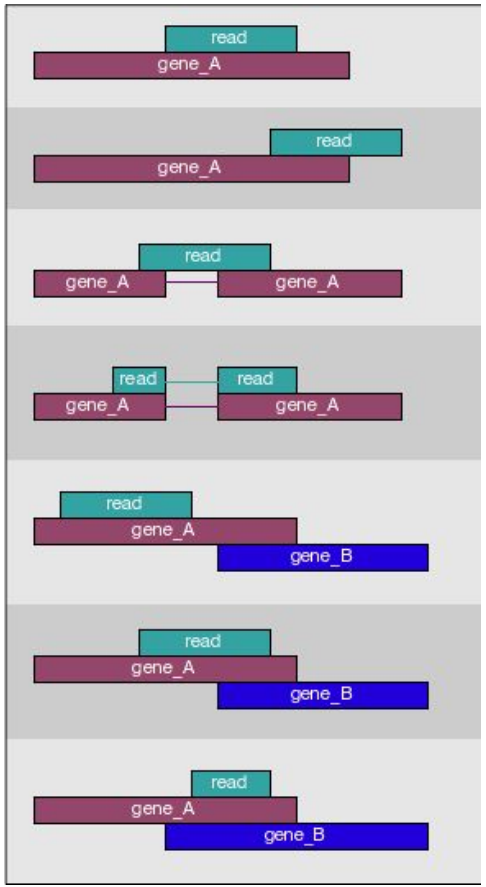
feature

meta-feature is the aggregation of a set of features with the same gene_id

How many reads overlap annotated regions?

Use [featureCounts]





Narrow down the genomic region that could contain features overlapping with the query read

Hierarchical data structure - features within blocks within bins

Estimation of the strandness - directionality of the RNA molecule

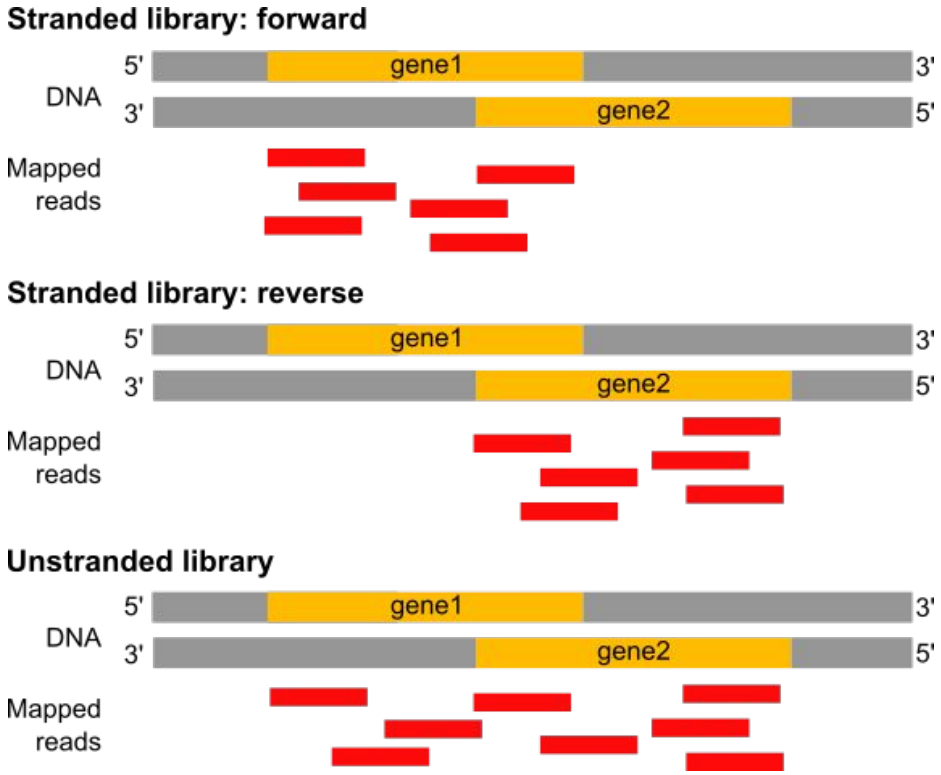
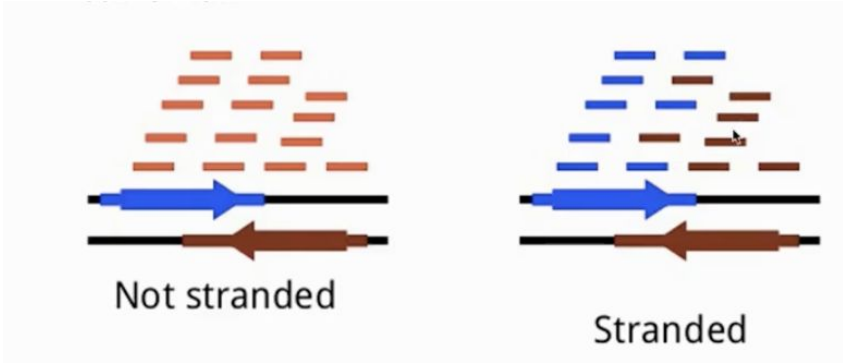


Image source: https://training.galaxyproject.org/training-material/

- In a stranded forward library, reads map mostly on the genes located on forward strand (here gene1).
- With stranded reverse library, reads map mostly on genes on the reverse strand (here gene2).
- With unstranded library, reads maps on genes on both strands.



featureCounts output

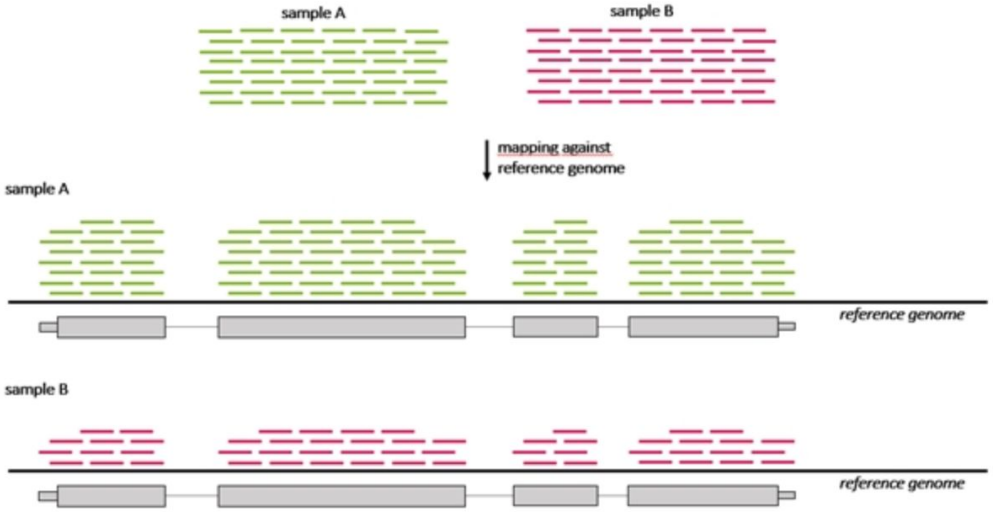
- count matrix (text file)
- a summary file

Each column is a sample

Each row is a gene

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Gene-wise counts should be normalized before comparing between samples



- Library size: counts can differ because of different library sizes (Sample A and Sample B)
- Real change in expression level of a gene vs non-biological reasons
- Need to estimate variability (dispersion)
- Several steps and specific tools

RNA-seq - analysis workflow

Session 1 Concepts:

Session 2 Concepts:

Biological question

Library preparation

Sequencer: Reads
Fastq

Trimming
cutadapt

Alignment/Mapping
STAR

Counting Reads
featureCounts

DGE brief overview

Quality check
FastQC

Quality check
FastQC

Quality check

Quality check

Session 2 Demo

Typical command line syntax of bioinformatics tools

```
$ Toolname –a 10 –b file.txt –c xyz.fq –o pqrs.tuv
```

```
$ Toolname --paramA 10 –b file.txt –c xyz.fq –outFile pqrs.tuv
```

- Examples:

```
$ fastqc –t 16 –o ./ Bacteria_GATTACA_L001_R1_001.fastq
```

```
$ fastqc *.fastq
```

```
$ cutadapt –a GATTACA –o ./trimmed.fastq input.fastq
```

Examples of commands to execute various steps

- Trimming adapters

```
$ cutadapt \
> –a file:Adapter_Sequence.fasta \
> –o ./trimmed.fastq Bacteria_GATTACA_L001_R1_001.fastq
```

- Similarly for other tools

Typical command line syntax with singularity container

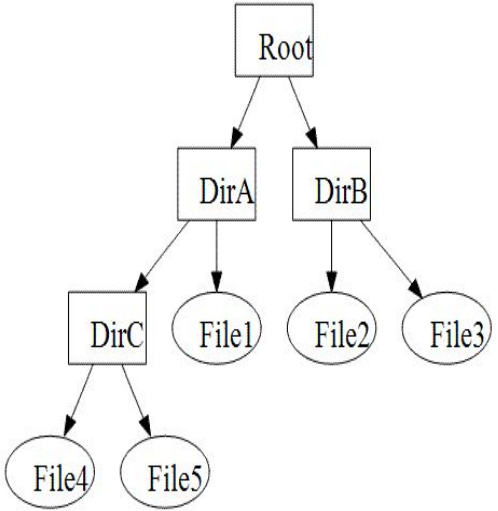
```
$ singularity exec containername Toolname –a 10 –b file.txt –c xyz.fq –o pqrs.tuv
```

- Examples:

```
$ singularity exec  rna_seq_container.sif fastqc –t 16 –o ./ Bacteria_GATTACA_L001_R1_001.fastq
```

File paths := Location of file on computer

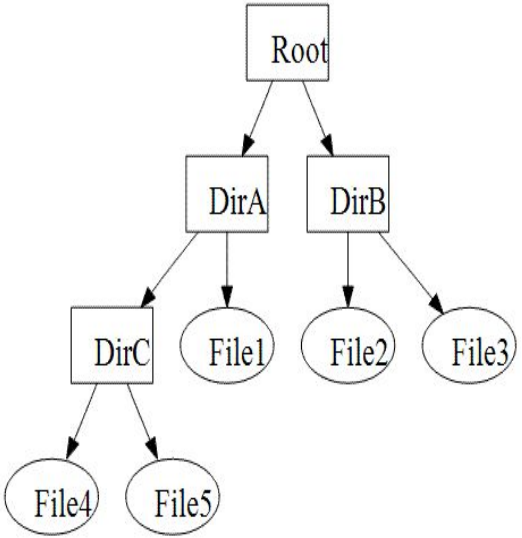
- Lots of files on a computer
- Organized in directories which may contain sub-directories
- Bioinformatics tools may need file inputs and may output files
  - Where are the input files located on the computer?
    - Searching entire computer not practical
    - What if multiple files have the same name?
  - Where should the output files be saved?



File paths

- **./** is for current working directory
- **../** is for parent directory of current working directory
- **../../** is for parent directory of parent directory of current working directory

- **/Root/DirA/DirC/File4** is the path of File4 in the image to the right.



RNA-seq - analysis workflow

Session 1 Concepts:

Session 2 Concepts:

Biological question

Library preparation

Sequencer: Reads
Fastq

Trimming
cutadapt

Alignment/Mapping
STAR

Counting Reads
featureCounts

DGE brief overview

Quality check
FastQC

Quality check
FastQC

Quality check

Quality check

Demo

Dataset

- Small dataset with 100k reads (for practice only).
  - FASTQ to tallying counts.

- Analysis of real datasets can take time. Real sequenced read libraries can be found online https://www.ncbi.nlm.nih.gov/sra/.

- The practice dataset has a low number of reads, which is not meaningful for differential gene expression analysis. Hence, for differential gene expression analysis use a real counts matrix, an example of which can be found at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49712.

Demo

- Start Docker Desktop
- Uploading FASTQ files to Docker container
- Running FASTQC
- Checking the adapter content

Knowledge check - Poll 6

What is the sequence length reported by the FASTQC?

1. 100000
2. 50
3. 0

Redo QC to ensure satisfactory quality

- Run FastQC.
- Is the adapter content gone?

Demo using Docker

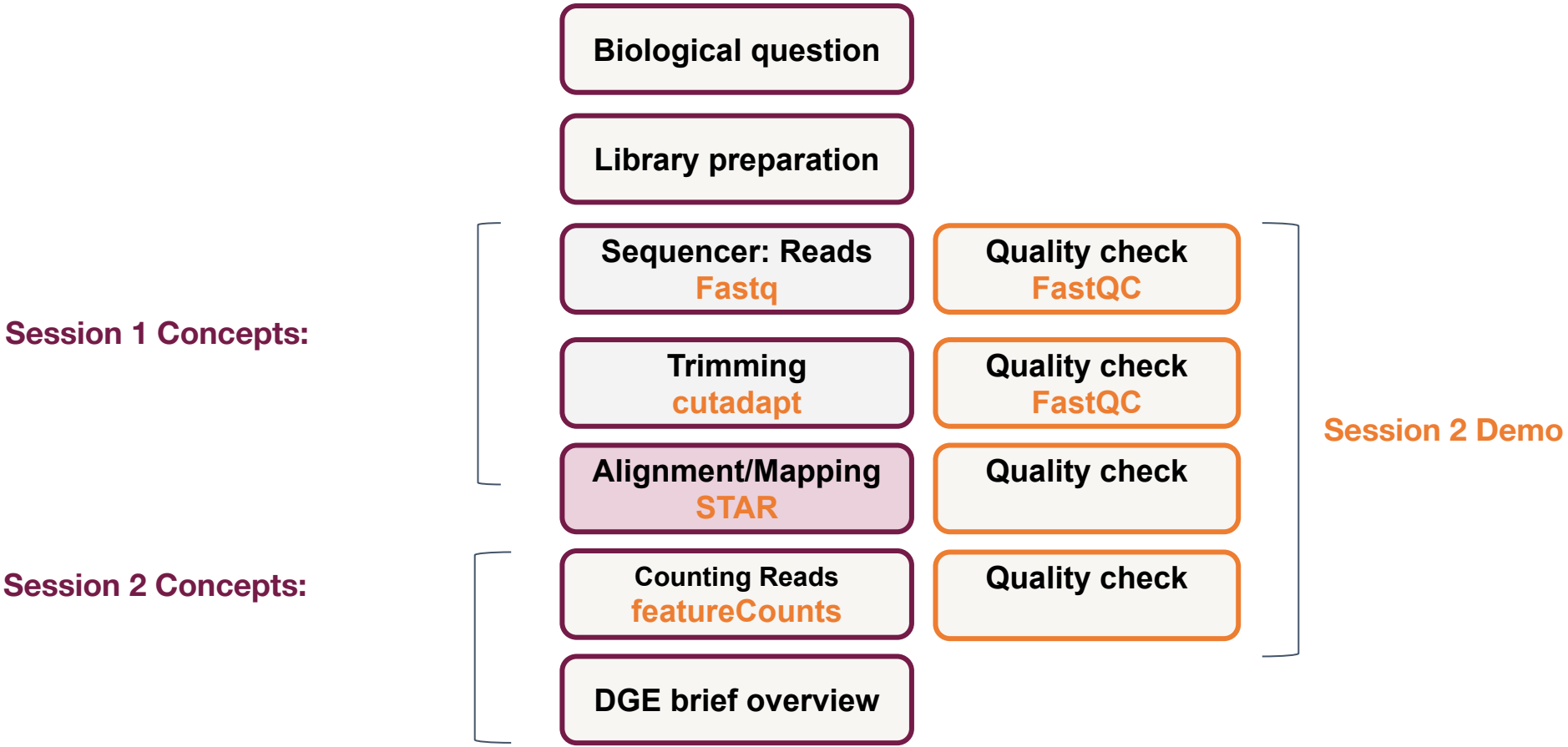
- Run STAR to align the reads to the reference genome

Break (5 min)

Need help? Please drop a question in the chat or speak up

Please take the survey:
https://www.surveymonkey.com/r/F75J6VZ

RNA-seq - analysis workflow



Poll 7

How is the pace?

1. Too fast - please go slower
2. Just right - keep going at the same pace
3. Too slow - please go faster

Demo using Docker

- Inspect the STAR output
- How do the output files look?

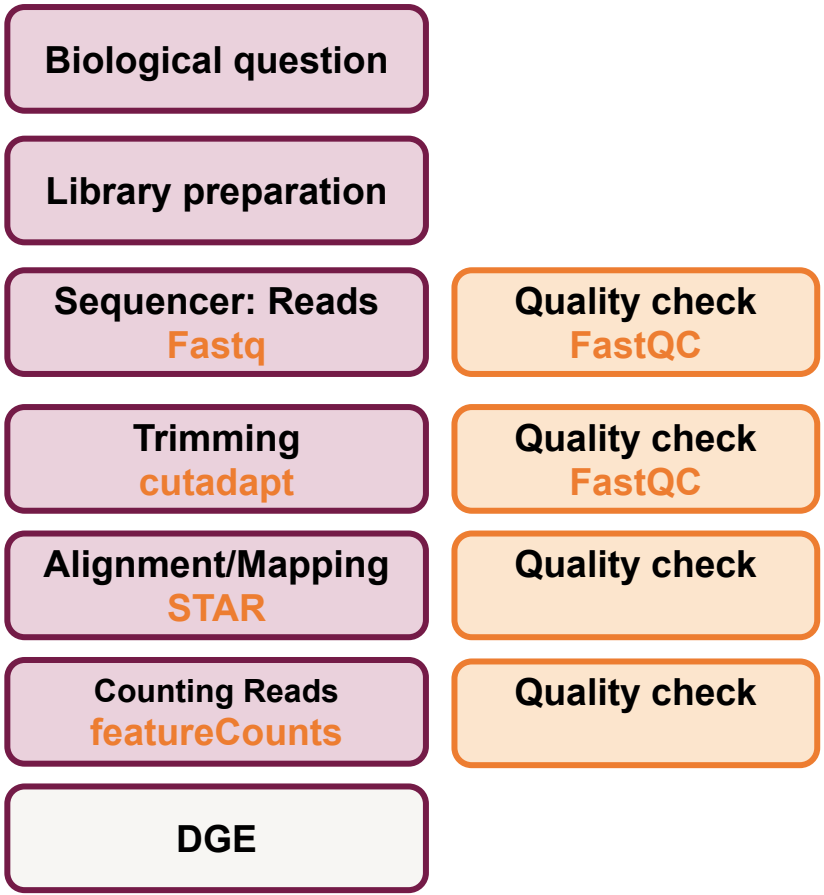
Tools to manipulate files are available

- Need to sort alignment report?
  - samtools
- Need to convert FASTQ to FASTA?
  - fastx-toolkit
- …
- Google!

Demo using Docker

- Run featureCounts to tally counts

RNA-seq - analysis workflow



Session 2 - Take-home messages

1) Each steps of the analysis can be affected by some kind of bias - Check the quality after each step!

2) Be aware of possible batch effects, non biological variability.

3) Know the tools and how they work. Use best practices for the analysis.

This workshop covered:

1) Common tools, e.g., fastqc, cutadapt, STAR

2) Common file formats, e.g., FASTQ, FASTA, SAM, GTF, BAM

3) Analysis on your computer using docker and on Wynton using singularity

Real data is complex

- This workshop provides an introduction to typical RNA-seq analysis steps using an artificial dataset. Real data might need additional analyses choices.

- Possible challenges with real data: (What we did not cover today)
  - What to do you if only 50%of reads align to reference?
  - What to do if FastQC reports unusual GC content?
  - What to do if the reference genome is incomplete?
  - How to deal with more complicated experimental designs?
  - How many replicates to use for RNA-seq?

- Some analysis choices need experience. Consult with the Gladstone Bioinformatics core for such scenarios and data.

How to publish your code?

Tailored Training Sessions on Publishing Your Code with GitHub

Maximize the impact of your research by mastering GitHub with **Gladstone's Bioinformatics Core training.** The team of experts will help you easily share your code, collaborate with colleagues, publish a function or library, build a website, and more. The core can tailor a training session to meet your lab's specific needs and bring your research to the next level. **Reach out to the core** to learn more or set up a training session.

Helpful resources

- Wynton slack channel
  - ucsf-wynton.slack.com
- Gladstone Bioinformatics Core slack channel
  - https://gladstoneinstitutes.slack.com/archives/C0145F1L7QS
- Wynton tutorials
  - https://github.com/ucsf-wynton/tutorials/wiki

Helpful resources

- Bacterial reference genome:
  - https://bacteria.ensembl.org/index.html
  - STAR and HISAT2 are more commonly used for eukaryotic RNA-seq. For bacterial RNA-seq data, Bowtie2 and BWA can be used as these are non-spliced reads. STAR can also be used for bacterial RNA-seq data but there some additional STAR options that should be specified when doing so.

- How to add reporter sequence to the reference:
  - https://groups.google.com/g/rna-star/c/FGQRotrCB1Q

- Cutadapt possible adapters types:
  - https://cutadapt.readthedocs.io/en/stable/guide.html#adapter-types

When I need help, which I do need on a daily basis, I visit:

- Slack channel for Wynton users
  - ucsf-wynton.slack.com
- http://seqanswers.com/forums/
- https://www.biostars.org/
- https://www.rna-seqblog.com/
- https://stackexchange.com/
- Google groups for specific tools
- GitHub issues
- …

Thank you!

Please take the survey:

https://www.surveymonkey.com/r/F75J6VZ

Other resources

- RNA-seq and analysis
- RNA-seq review
- RNA-seqlopedia
- Galaxy tutorial

Our upcoming workshops


GLADSTONE
INSTITUTES