Introduction to RNA-seq data analysis

Gladstone Institutes

Krishna Choudhary

Biostatistician @ Bioinformatics Core @ GIDB

March 28, 2019

Overall goals

- Demystifying RNA-Seq computational analysis.

- Enable informed conversations with computational biologists.

- Work with Galaxy.

Contents

- Introduction
- From sequencer output to differential analysis (Hands-on)
- Conclusion

Typical protocol

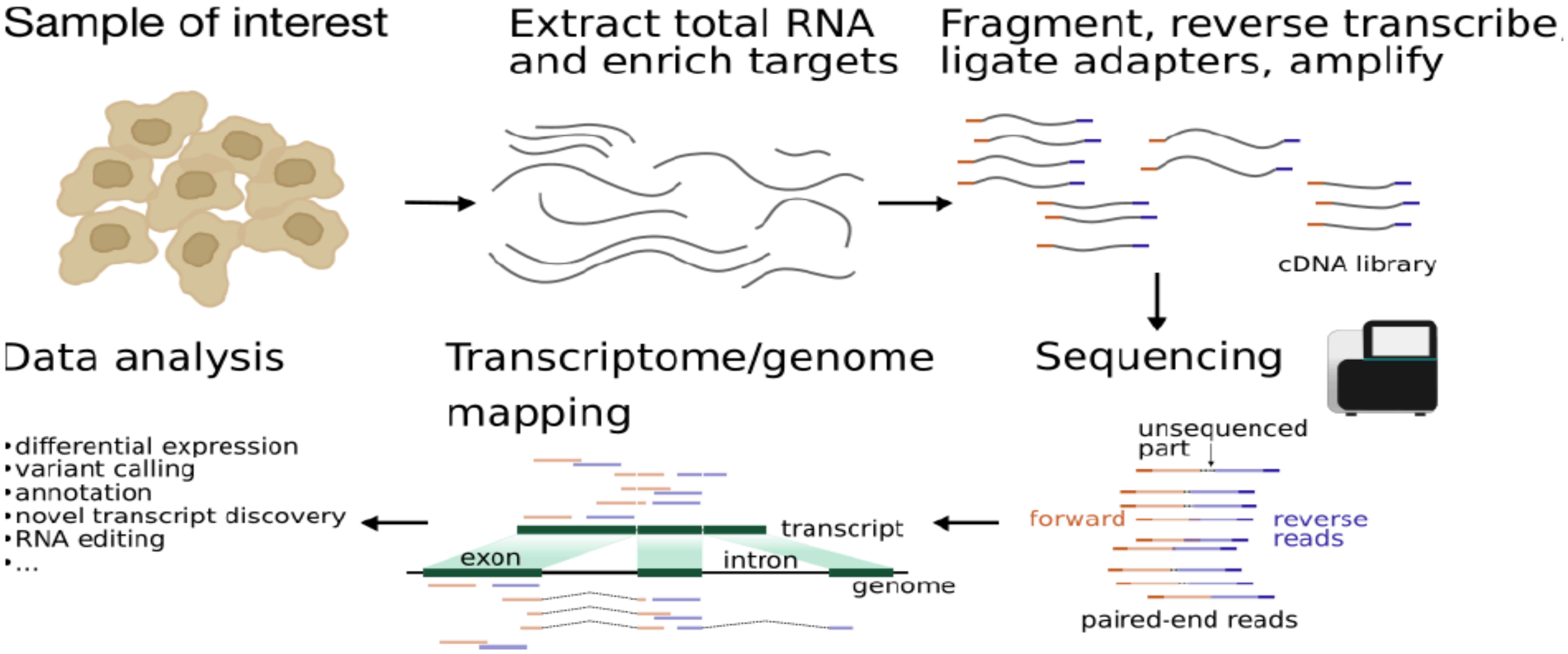
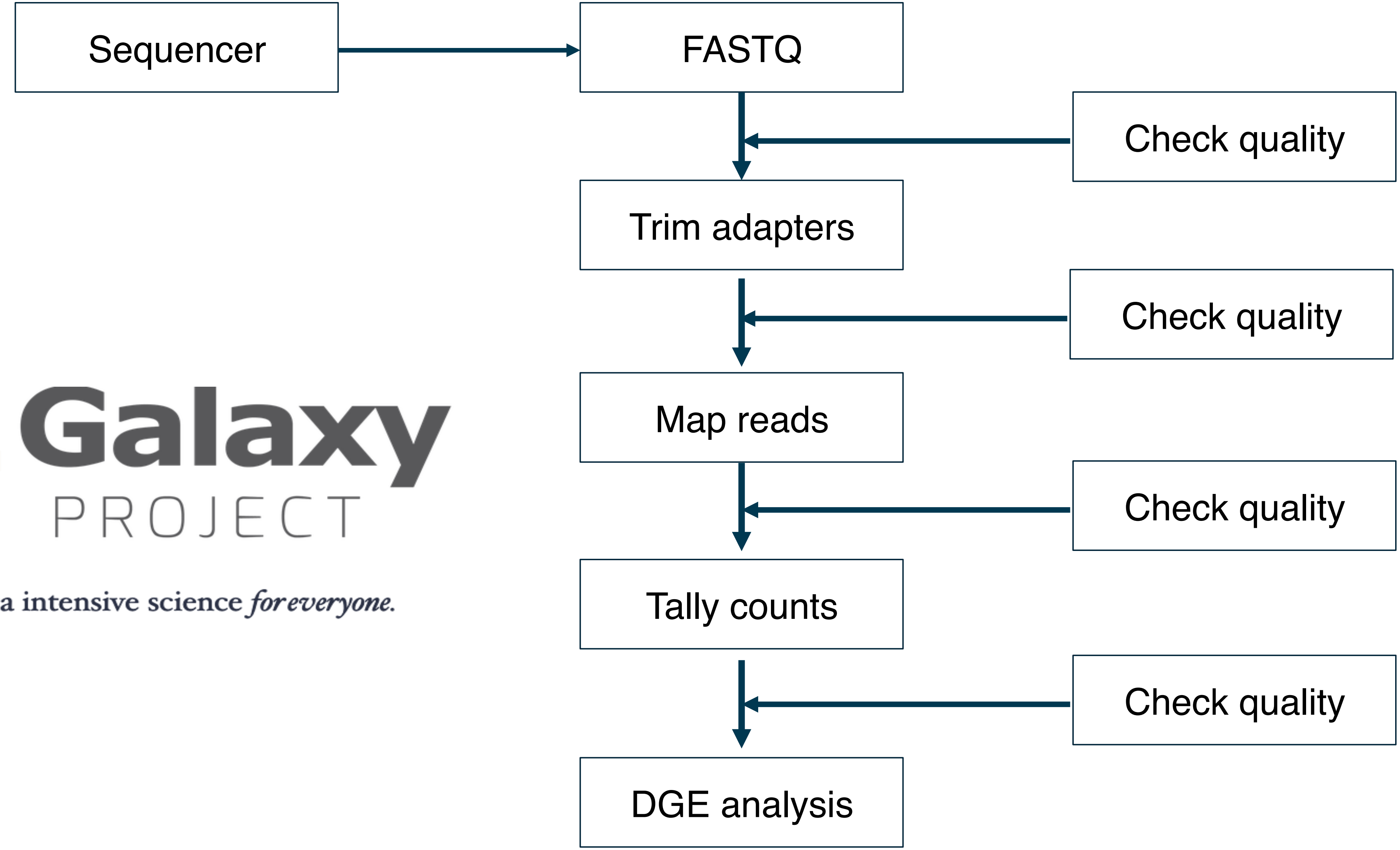


Figure: Berge et al., 2018, PeerJ Preprints.



Sequencer

FASTQ

Check quality

Trim adapters

Check quality

Map reads

Check quality

Tally counts

Check quality

DGE analysis

Galaxy PROJECT

Data intensive science *for everyone.*

Experiment design influences data analysis.
(should be planned to address relevant questions)

- What is the biological question that we seek to answer?
- How many tissue types and/or time points to compare?
- How deep should we sequence?
- Read length?
- Which sequencing platform?
- Single-end or paired-end?
- Pooling?
- Biological replicates?
- Technical replicates?
- Additional considerations?

Not the subject matter today!

- Workshop on April 2 by Reuben Thomas: *Intro to statistics and experimental design*.
- Reading material in Dropbox: *RNA sequencing data : hitchhiker's guide to expression analysis* by Berge *et al.*, 2018

Dataset

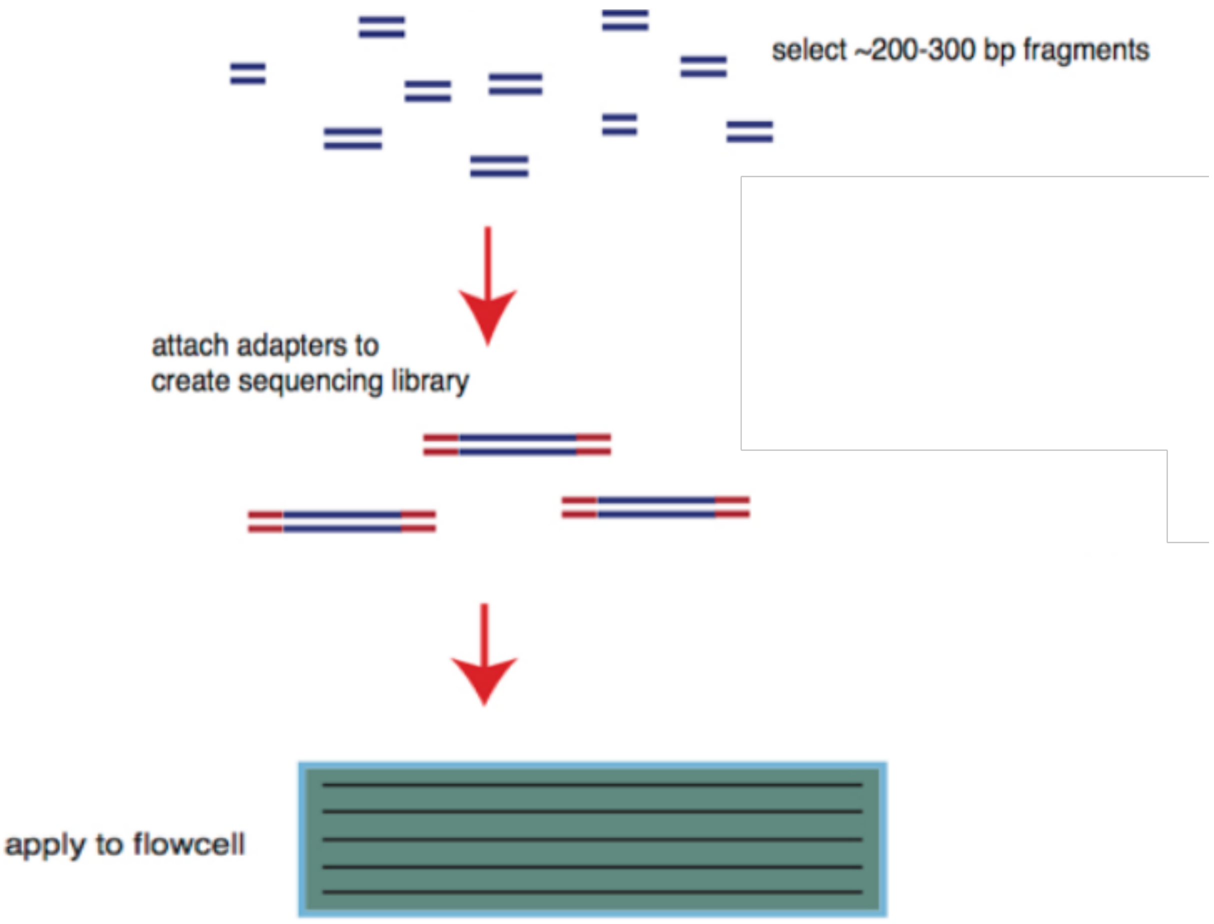
- Small dataset with 100k reads (for practice only).
  - FASTQ to tallying counts.

- Real counts data (GSE49712).
  - Use this for DGE analysis.
  - 5 replicates of two groups.
  - Group A: Strategene Universal Human Reference RNA
  - Group B: Ambion Human Brain Reference RNA

Sequencing centers provide FASTQ files. (~15 min)

Section goal: Understanding origin and contents of FASTQ file type.

cDNA library is applied to a flow cell.



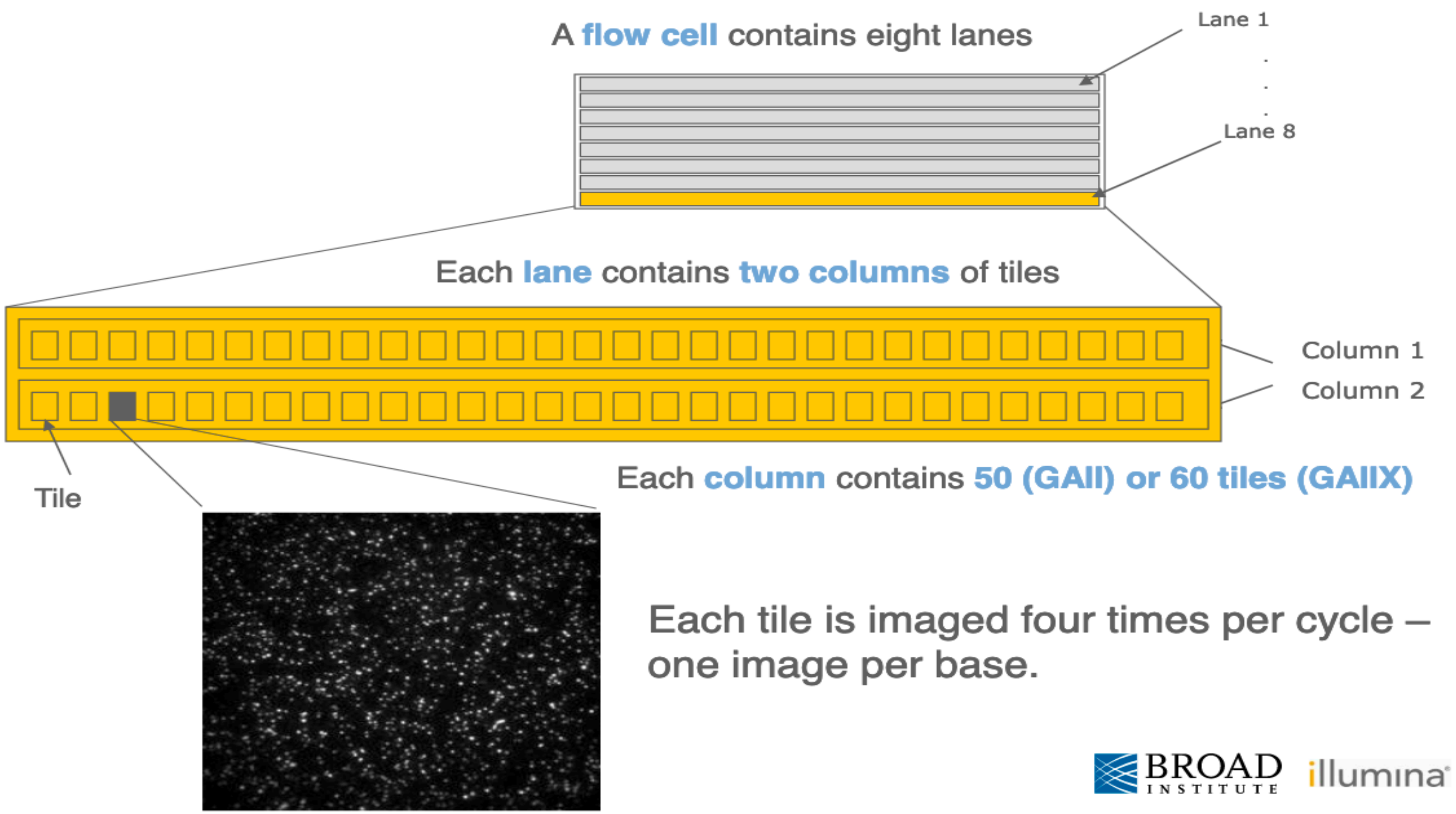
select ~200-300 bp fragments

attach adapters to
create sequencing library

apply to flowcell

Image adapted from a blog by Stuart M. Brown (link in description).

Flow cells are organized in lanes, columns and tiles.



A **flow cell** contains eight lanes

Each **lane** contains **two columns** of tiles

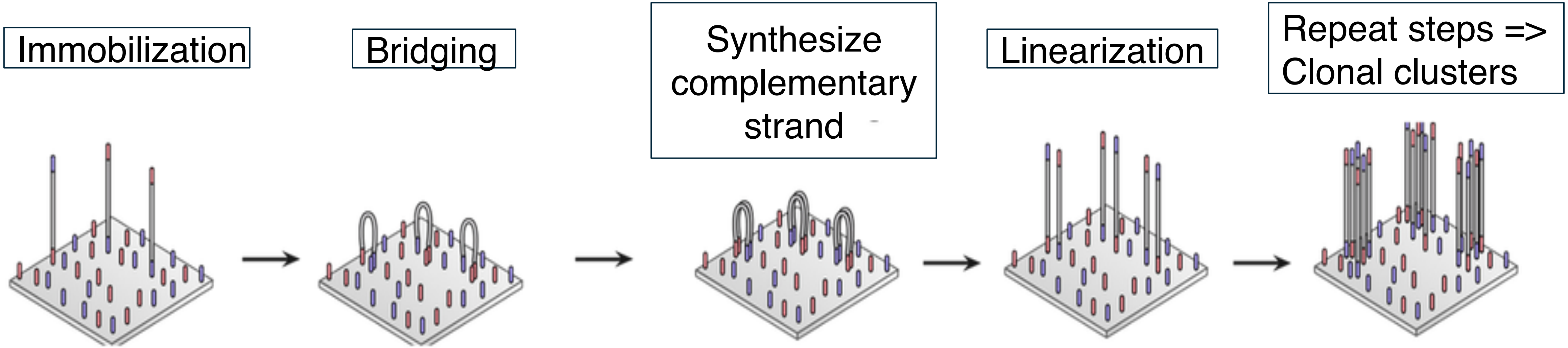
Each **column** contains **50 (GAII) or 60 tiles (GAIIX)**

Each tile is imaged four times per cycle – one image per base.

BROAD INSTITUTE illumina

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

DNA fragments immobilized on flow cell & amplified into clonal clusters.



Immobilization

Bridging

Synthesize complementary strand

Linearization

Repeat steps => Clonal clusters

cDNA fragments stick to complementary sequences on flow cell.

Fragments bend. Adapter on other end pairs with complementary sequence.

Image adapted from blog by Lauren Launen (link in description).

Sequencing by Synthesis

1. Adapters contain primer binding sites.
2. Nucleotide with reversible terminator & fluorophore added.
3. Image nucleotide added.
4. Remove terminator and fluorophore.
5. Repeat 2-4.

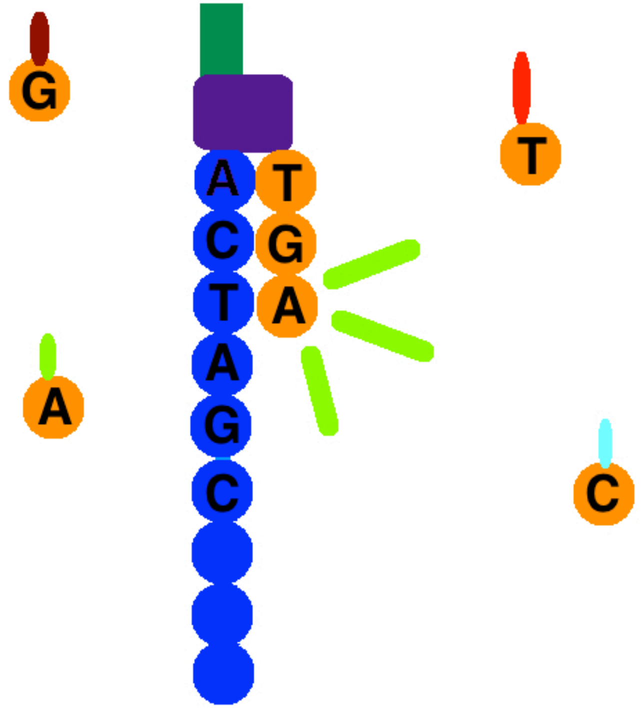
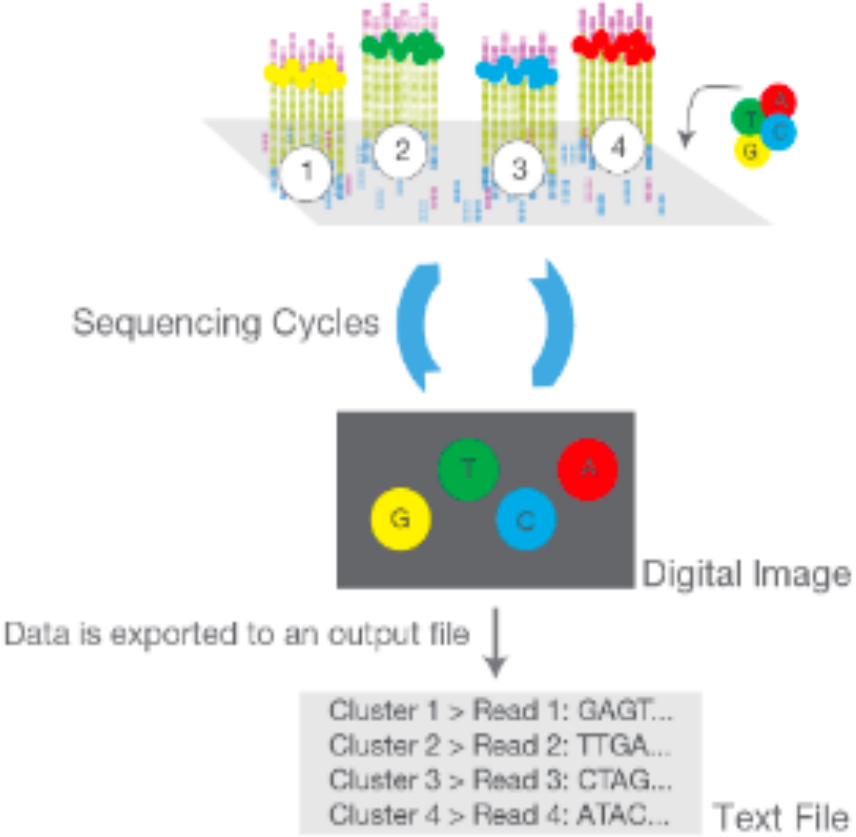


Image source: DMLapato [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]

Strong signal from monoclonal clusters.



Sequencing Cycles

Digital Image

Data is exported to an output file

Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...

Text File

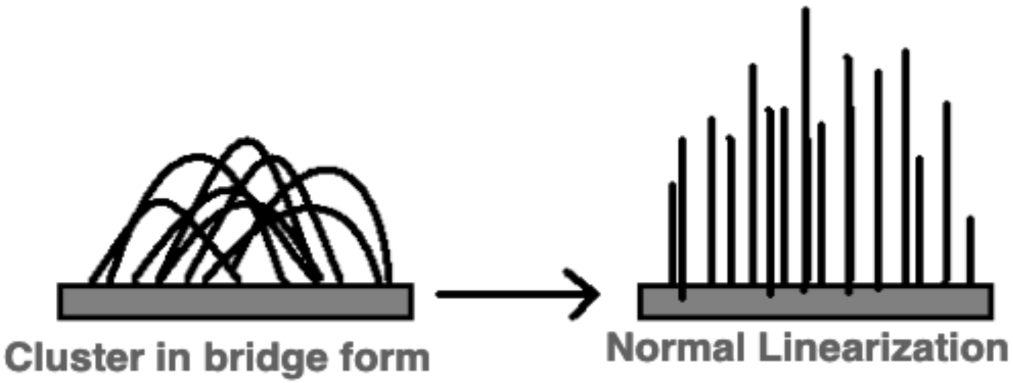
FASTQ files contain detailed information about each read.

- Read sequence.
- Instrument used, flow cell id, lane number, tile number, etc.
- Quality of each base call.

Base calling may not be accurate.

Various possible causes: Example

Ideal world



Real world

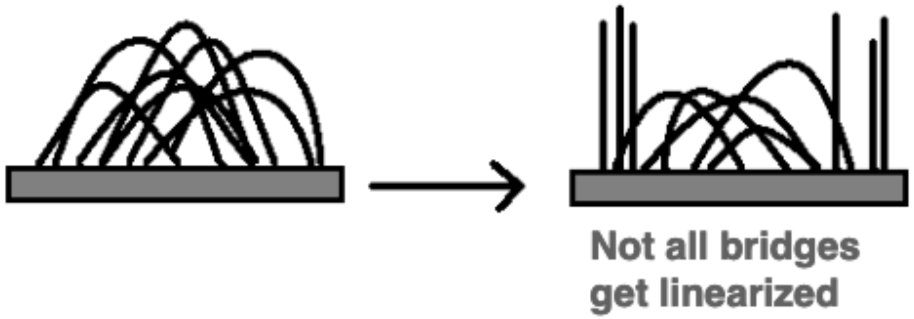


Image adapted from a Broad Institute presentation (link in description).

Base calling may not be accurate.

Various possible causes: Example

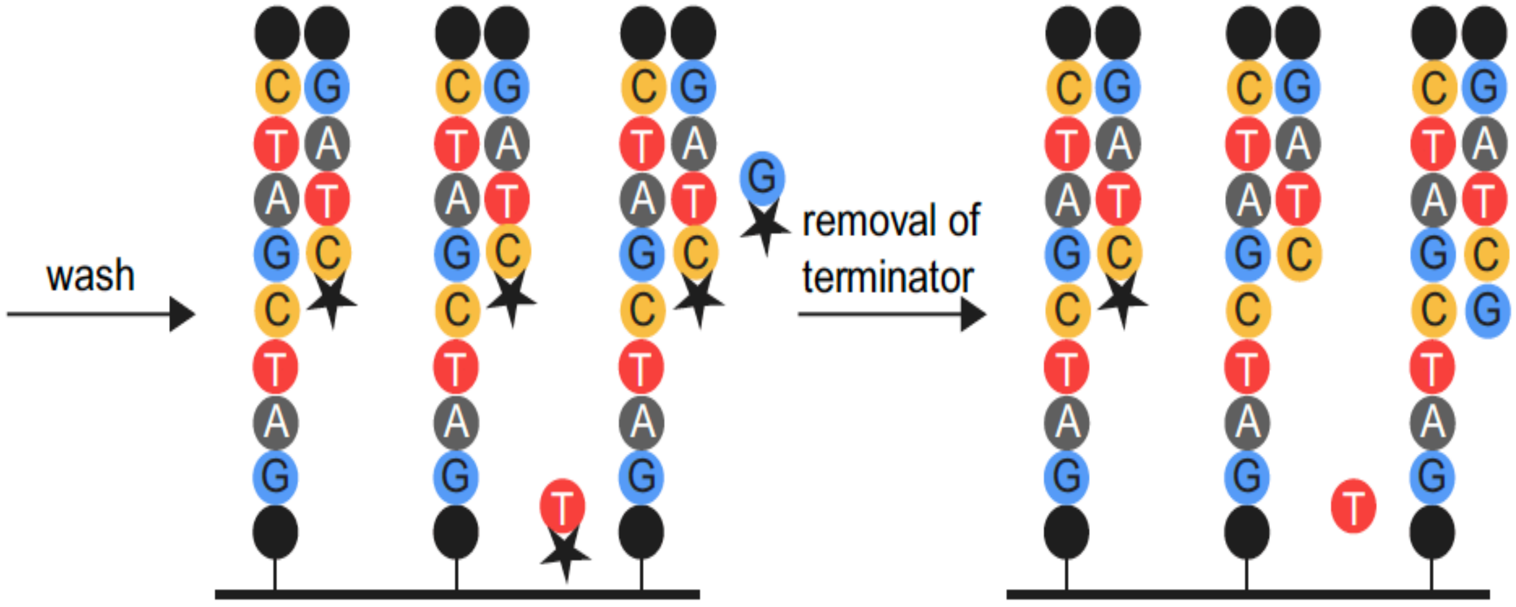
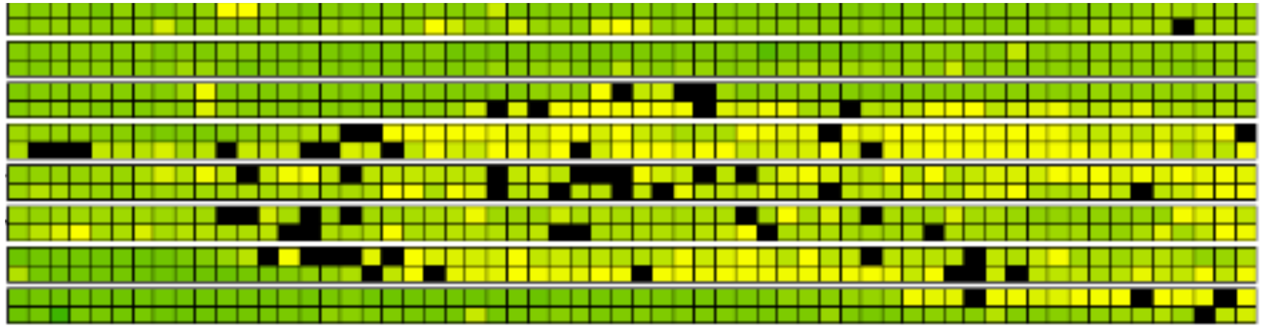


Image adapted from Pfeiffer et al., Scientific Reports, (2018) 8:10950.

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.

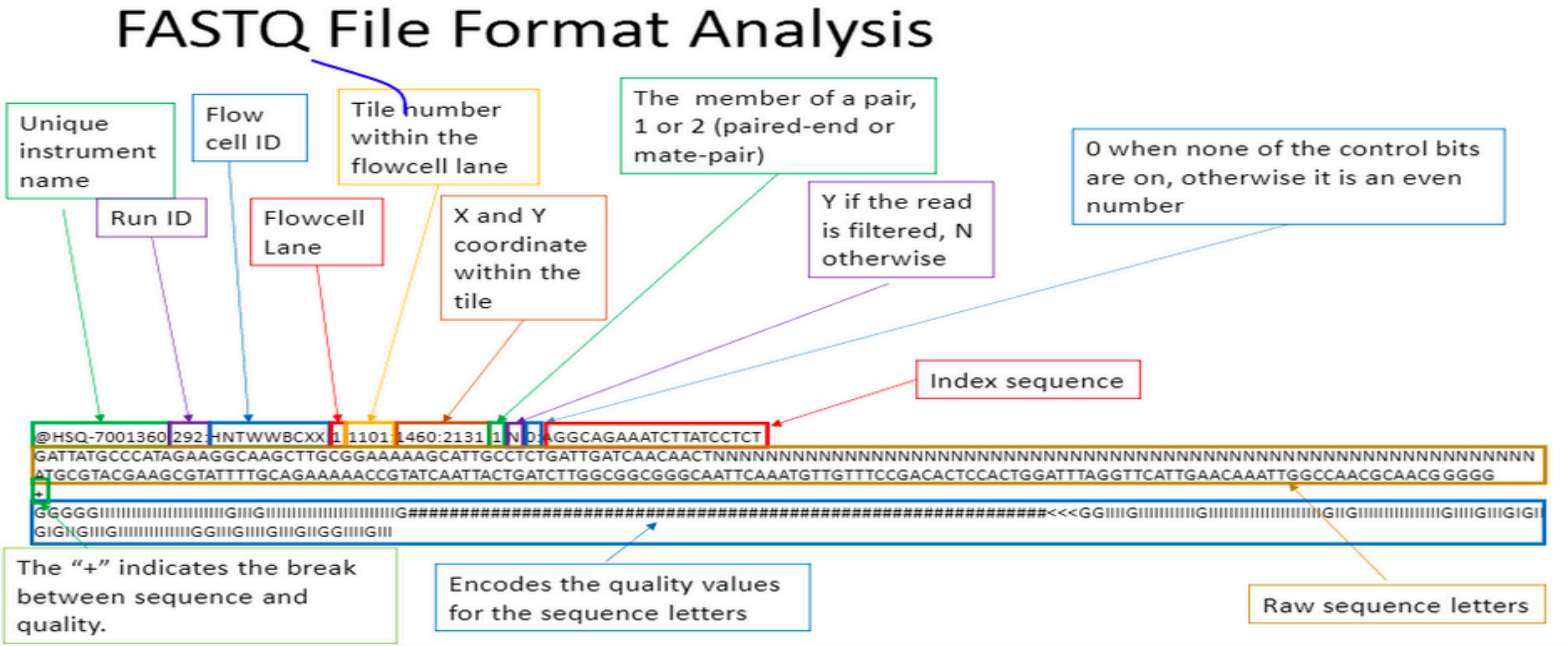


=> Need to record quality of each base call.

Example FASTQ file with one read only.

- Open Single_read.fastq

Four lines per read:
1. Read ID, 2. Sequence, 3. Space for optional info, 4. Quality.



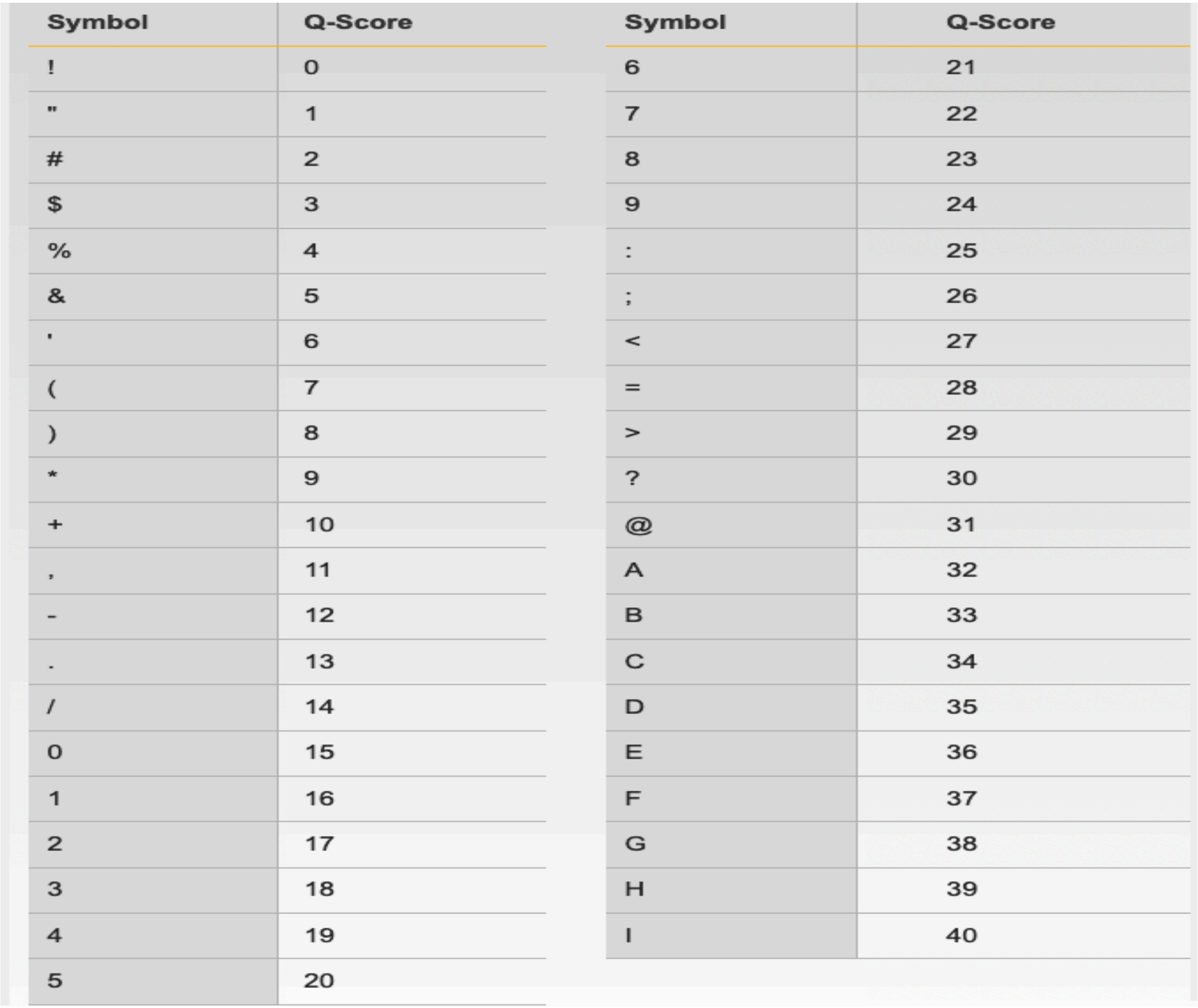
FASTQ File Format Analysis
Unique instrument name
Run ID
Flow cell ID
Flowcell Lane
Tile number within the flowcell lane
X and Y coordinate within the tile
The member of a pair, 1 or 2 (paired-end or mate-pair)
Y if the read is filtered, N otherwise
0 when none of the control bits are on, otherwise it is an even number
Index sequence
The "+" indicates the break between sequence and quality.
Encodes the quality values for the sequence letters
Raw sequence letters

```
@HSQ-7001360:292:HNTWWBCXX:1:1101:1460:2131 1:N:0:AGGCAGAATCTTATCCTCT
GATTATGCCCATAGAAGGCAAGCTTGCGGAAAAAGCATTGCCTCTGATTGATCAACAACTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
ATGCGTACGAAGCGTATTTTGCAGAAAAAACCGTATCAATTACTGATCTTGGCGGCGGGCAATTCAAATGTTGTTTCCGACACTCCACTGGATTTAGGTTCATTGAACAAATTGGCCAACGCAACGGGGG
+
GGGGGIIIIIIIIIIIIIIIIIIIIIIGIIIGIIIIIIIIIIIIIIIIIIIIIIIG###############################################################<<<GGIIIIGIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIIIIIIGIIGIIIIIIIIIIIIIIIIGIIIIGIIIGIGII
GIGIIGIIIGIIIIIIIIIIIIIIGGIIIGIIIIGIIIGIIGGIIIIGIII
```

Image adapted from blog by Lauren Launen (link in description).

Quality is encoded as symbols.

Quality measured in terms of Phred scores.



Symbol	Q-Score	Symbol	Q-Score
!	0	6	21
"	1	7	22
#	2	8	23
$	3	9	24
%	4	:	25
&	5	;	26
'	6	<	27
(	7	=	28
)	8	>	29
*	9	?	30
+	10	@	31
,	11	A	32
-	12	B	33
.	13	C	34
/	14	D	35
0	15	E	36
1	16	F	37
2	17	G	38
3	18	H	39
4	19	I	40
5	20		

Link for Illumina encoding of scores in description.

Adapters, primers, contaminants, target sequences, etc. represented in FASTQ files.

- Open Bacteria_GATTACA_L001_R1_001.fastq.

Length of insert < Length of reads ordered => Adapters included in reads.

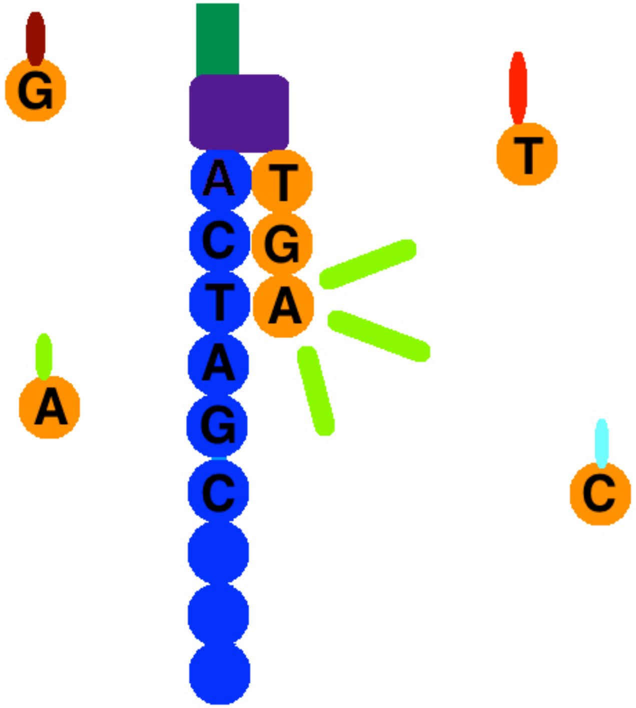


Image source: DMLapato [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)]

Naming conventions for fastq files.

- File names often follow a format.
  - SampleName_Barcode_LaneNumber_ReadNumber_SetNumber.fastq
  - Ex – Bacteria_GATTACA_L001_R1_001.fastq
- Paired-end reads named with R1 and R2 in file name.
  - Ex – 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.*

Quality control of sequencing files. (~ 30 mins)

Section goal: Running FastQC and interpreting results.

FastQC: Tool for quality control of sequencing data

- Summarizes quality of base calls.
- Checks for presence of know adapters.
- Any sequences more frequently observed than typical?
- Any sequence biases?
- Any GC biases?
- …

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!!!

- For protected or large data:
  - Setup local galaxy instance.
  - Run Galaxy on the cloud.

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 over-represented sequences?
  - Are they contaminants, e,g. adapers or PCR duplicates?
  - If so, clean up contaminants.
  - Could be attributed to highly abundant transcripts.

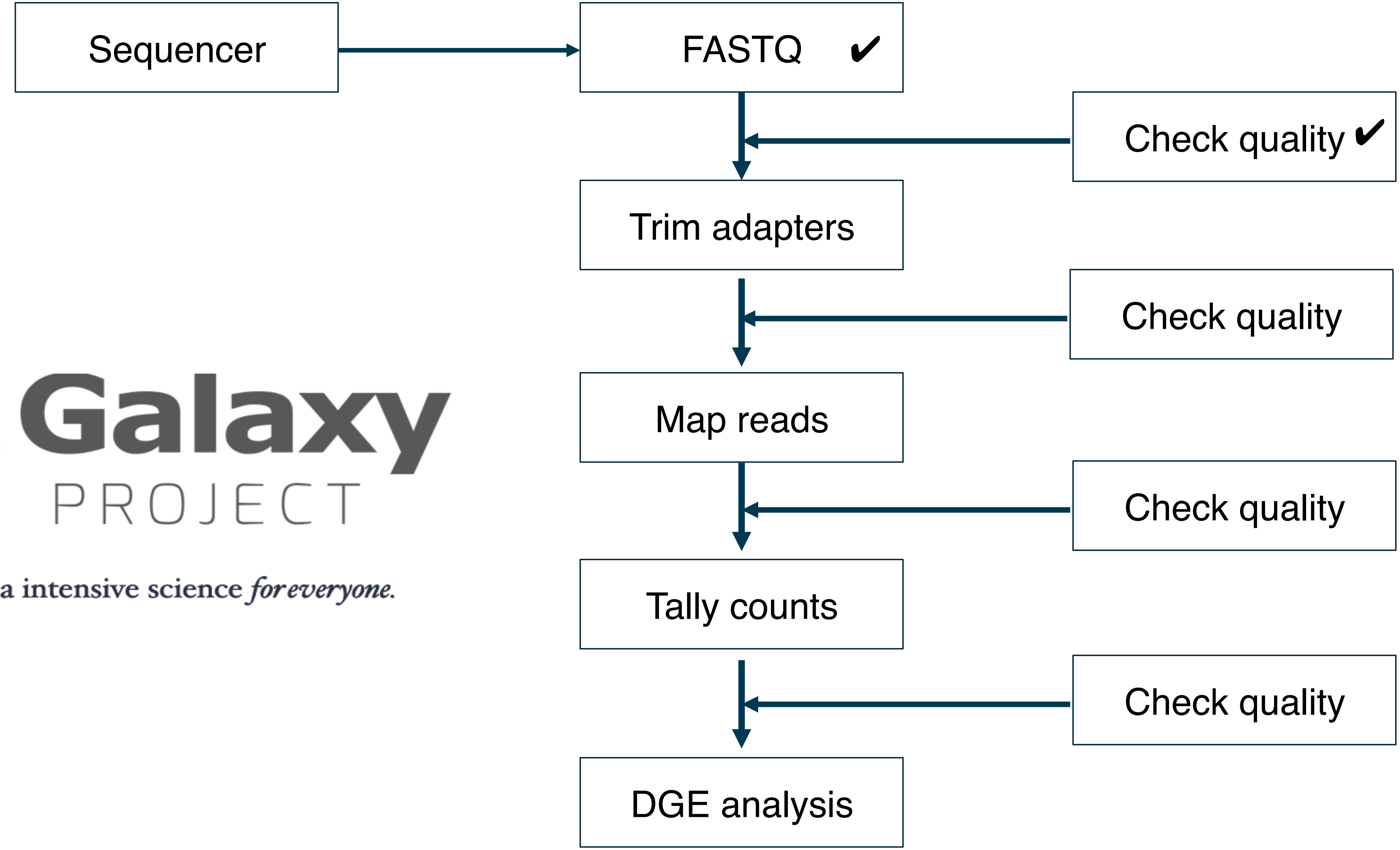
- Are sequence biases expected?

- For more: https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/

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



Sequencer

FASTQ ✔

Check quality ✔

Trim adapters

Check quality

Map reads

Check quality

Tally counts

Check quality

DGE analysis

Galaxy PROJECT

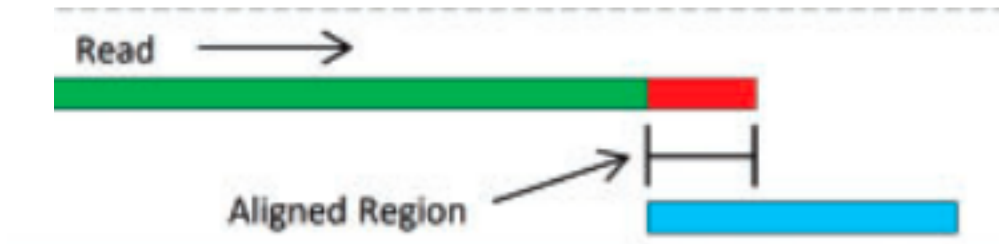
Data intensive science *for everyone.*

Cleaning up contaminants (20 mins)

Section goal: Run cutadapt on fastq files to remove adapters.

cutadapt removes adapters.

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

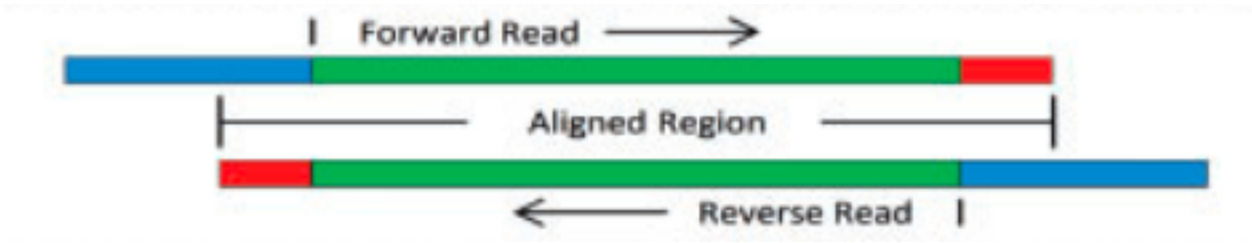


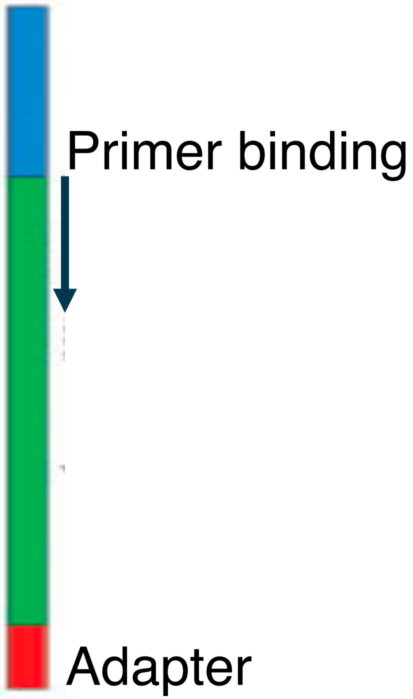
(Image adapted from Bolger et al., 2014, Bioinformatics. Link in description.)

Alternative approach: Trimmomatic

- Say adapter sequence in read is very short.
  - Can we still identify it?

- Yes for paired-end reads.





(Image adapted from Bolger et al., 2014, Bioinformatics. Link in description.)

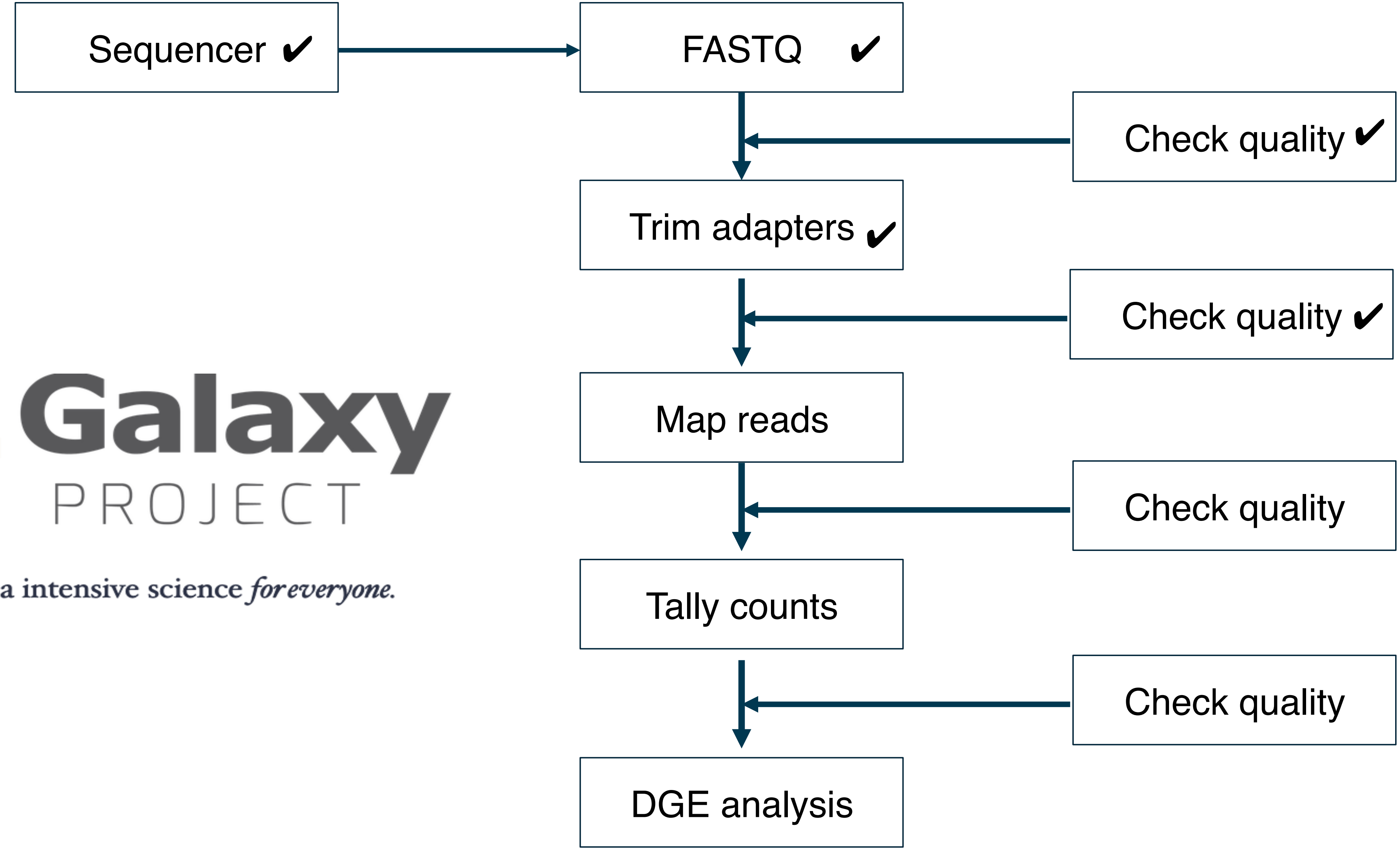
What else to clean?

- PCR primers?
- Unique molecular identifiers?
- Poor quality base calls?
- …

Redo QC to ensure satisfactory quality.

- Run FastQC.

- Are over-represented sequences gone?



Galaxy PROJECT

Data intensive science *for everyone.*

Mapping reads (20 mins)

Section goal: Understand alignment method

Mapping := Aligning reads to regions of reference DNA.

- After cleaning, reads from real sample only. (Assumption)

- Mapping := Aligning reads to regions of reference DNA.

- Challenges:
  - Reference sequences can be very long (~3 billion bp for humans).
  - Order of 100 million reads to be mapped.
  - Sometimes, need to account for splicing.
  - Allow for PCR artifacts/sequencing errors.

Inputs needed.

1. Reads to align.
   - FASTQ file after cleaning.

2. Reference sequence to align to.
   - Example – “rDNA_sequence.fasta”
   - FASTA format. Two lines per sequence.
     I. Starting with “>”, followed by sequence name/identifier.
     II. Sequence.
   - File extensions: .fasta, .fa, .txt.

Indexing reference sequence speeds up mapping.

- Use bowtie2 to build index.

- Use cleaned reads and index of reference sequence to map.

Output =>
1. Alignments in SAM format, 2. Summary of mapping statistics.

- SAM format:
  - For each read, mapped where, in what orientation?

- Summary statistics:
  - How many reads mapped?
  - How many unmapped?
  - …

Binary Alignment/Map (BAM) format

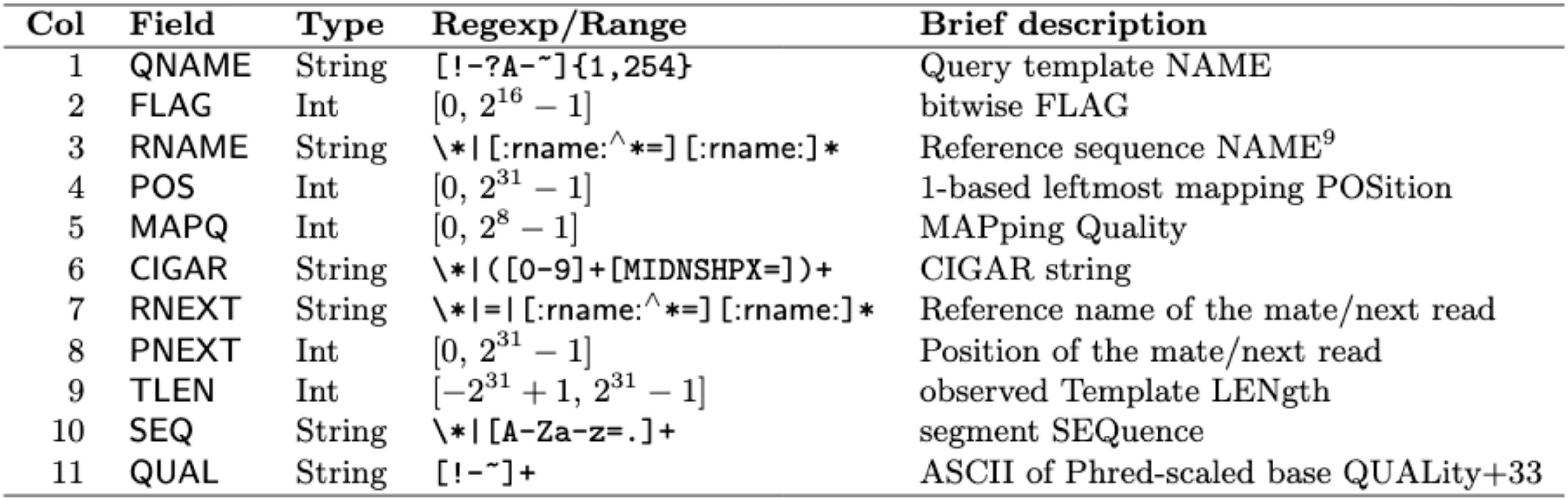
- Alignment reports often very large files.

- BAM extension used for compressed SAM files.

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.

11 fields for each alignment (per row).



Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[9]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Alternatives

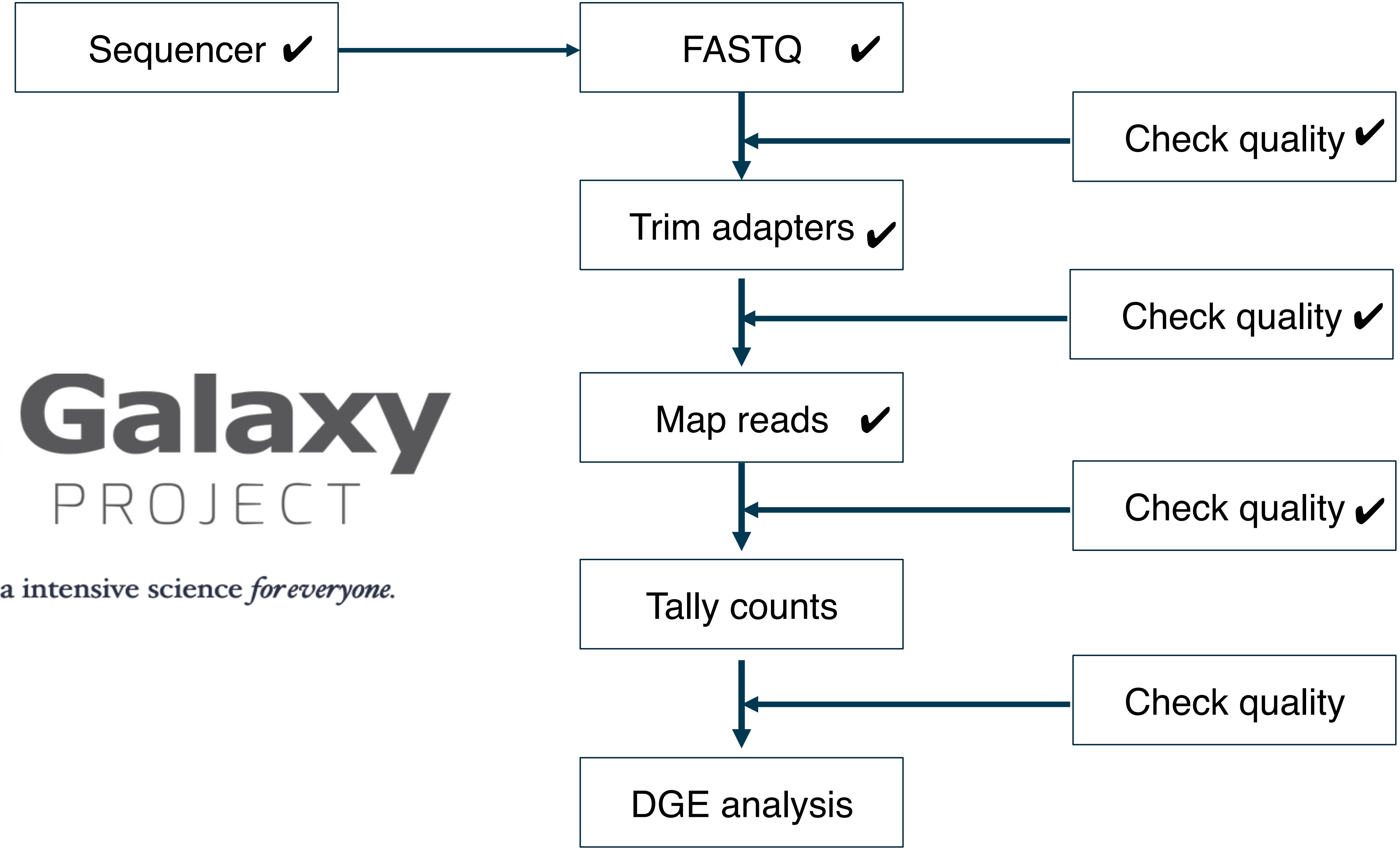
- Several. Example – bowtie2, BWA, subread, etc.
- Differences in speed and memory requirement.
- Pros and cons of each:
  - Example: Some handle spliced alignment, others do not.
  - …

Online resources for sequencing data analysis

- http://seqanswers.com/forums/
- https://www.biostars.org/
- https://www.rna-seqblog.com/
- …

Tools to manipulate files are available.

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



Sequencer ✔

FASTQ ✔

Check quality ✔

Trim adapters ✔

Check quality ✔

Map reads ✔

Check quality ✔

Tally counts

Check quality

DGE analysis

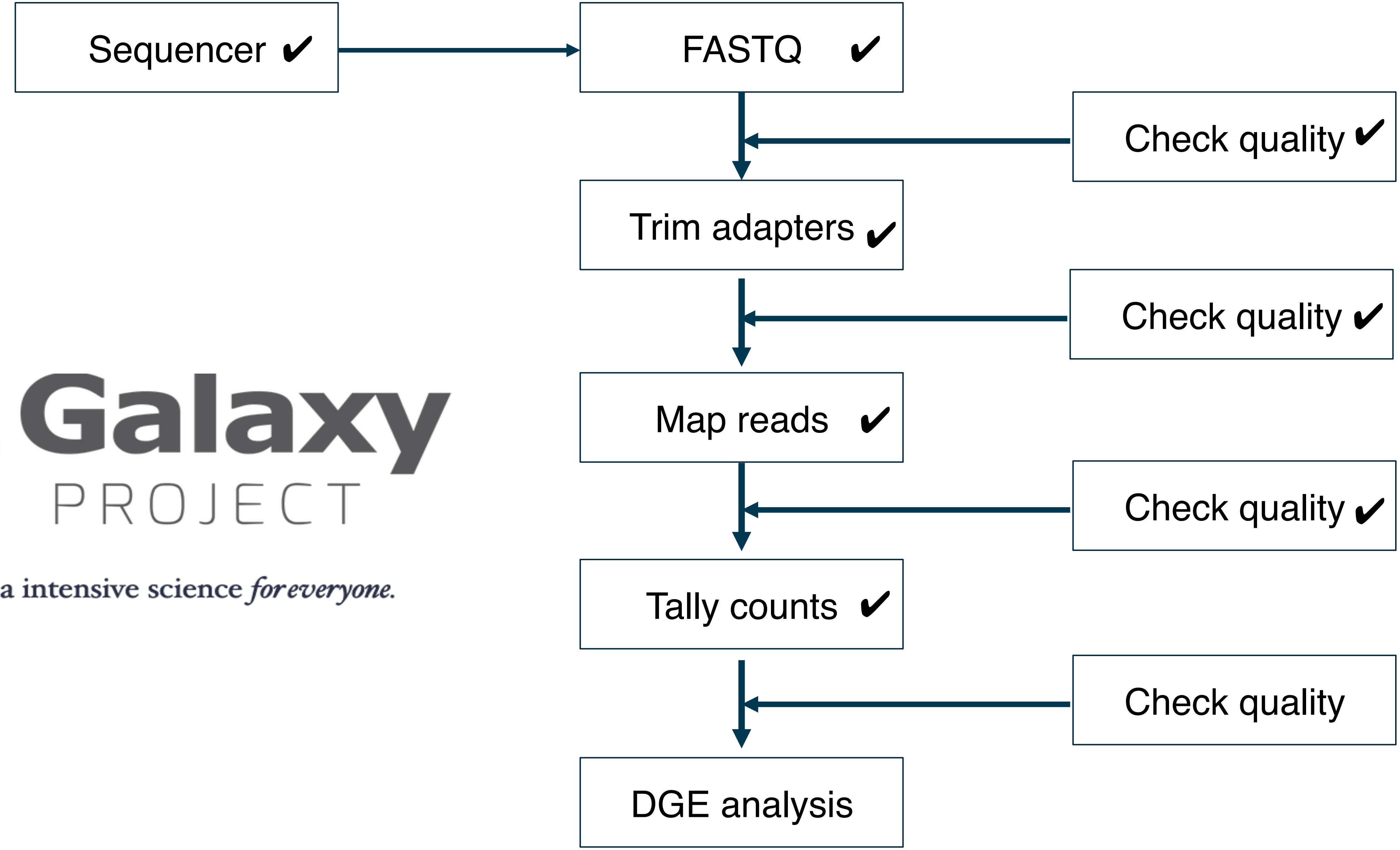
Galaxy PROJECT

Data intensive science *for everyone.*

Tally counts (~15 mins)

How many reads overlap annotated regions?

- Need annotation information.
- Need alignment information.
- Use featureCounts.



Sequencer ✔

FASTQ ✔

Check quality ✔

Trim adapters ✔

Check quality ✔

Map reads ✔

Check quality ✔

Tally counts ✔

Check quality

DGE analysis

Galaxy PROJECT

Data intensive science *for everyone.*

Downstream analysis (~15 mins)

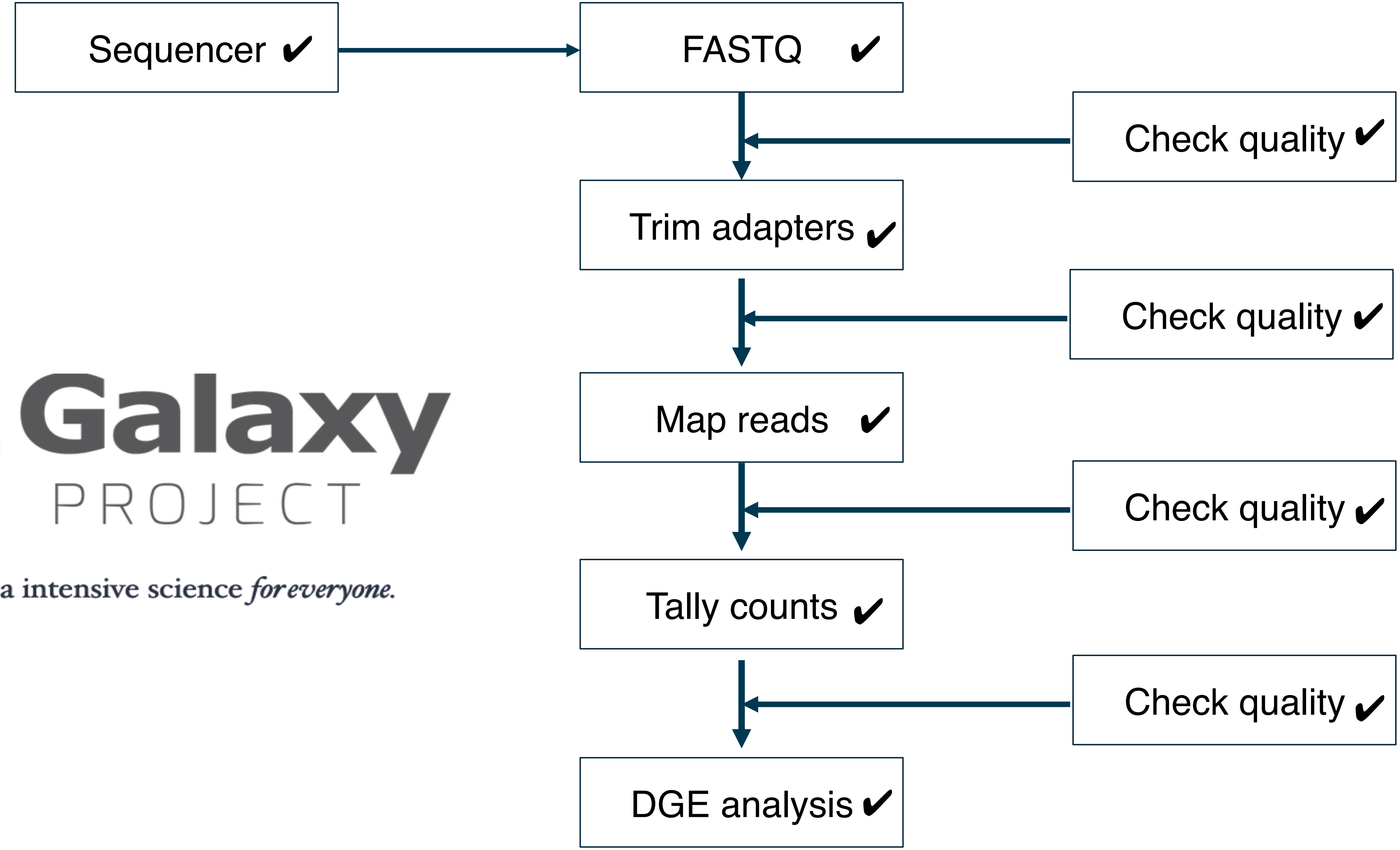
No. 1: Differential gene expression analysis.

Gene-wise counts should be normalized before comparing between samples.

- Counts can differ because of different library sizes.
- Mapping statistics might be different for samples.
- Real change in expression level of a gene.
- …
- Need to factor out differences due to non-biological reasons.

Counts may differ due to inherent noisiness of biological systems.

- Identical individuals may give different counts.
- Inherent variation used as benchmark to call out interesting variation.
- Need to estimate inherent variation or dispersion.



Sequencer ✔

FASTQ ✔

Check quality ✔

Trim adapters ✔

Check quality ✔

Map reads ✔

Check quality ✔

Tally counts ✔

Check quality ✔

DGE analysis ✔

Galaxy PROJECT

Data intensive science *for everyone.*

Your feedback is important to us!

- https://bioinformatics-course-feedback.questionpro.com/
- ~5 min.

Conclusions (~5 min)

Topics covered

- Steps of analysis.
- Common tools, e.g., cutadapt, fastqc, bowtie2, edgeR, etc.
- Common file formats, e.g., FASTQ, FASTA, SAM, GFF, etc.
- Analysis with Galaxy.

Additional information: Sources of data

- Sequence read archive
  - https://www.ncbi.nlm.nih.gov/sra

- Download and install SRA toolkit

- Step-by-step guide:
  - https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/#download-sequence-data-files-usi

More tools

- Quality control: RSeQC, MultiQC, etc.
- Mapping: STAR, BWA, etc.
- File manipulation: bedtools, samtools, fastx-toolkit, etc.
- Visualization: UCSC Genome Browser
- …

Upcoming Workshops

- Intermediate RNA-Seq analysis
  - April 17
- Single cell analysis (Symposium)
  - May 14
- Pathway analysis
  - Oct 7

Thank you!

STARBUCKS COFFEE

GLADSTONE
INSTITUTES

Applications

- Genome annotation
- Gene regulation
- Clinical applications, e.g., molecular sub-classification of cancer
- Meta-transcriptomics
- Spatial transcriptomics
- …