Introduction to RNA-seq data analysis

Gladstone Institutes

Krishna Choudhary
Bioinformatics Core, GIDB

October, 2020

Overall goals

- Demystify RNA-Seq data analysis
- Enable informed conversations with computational biologists
- Demonstrate how to analyze data
  - using a graphical user interface (Galaxy)
    - IMO: Good for dabbling but not if one needs to analyze large-scale data often
  - using a command line interface (HPC)
    - Most computational biologists use the command line

Contents

- Introduction
  - Typical RNA-seq protocol
  - Different platforms for computing
- From sequencer output to count matrix
  - Tools for today: fastqc, cutadapt, STAR, samtools, featureCounts
  - File formats: FASTQ, FASTA, BAM/SAM, GFF
- Conclusion & Overview of Session 2

Typical protocol

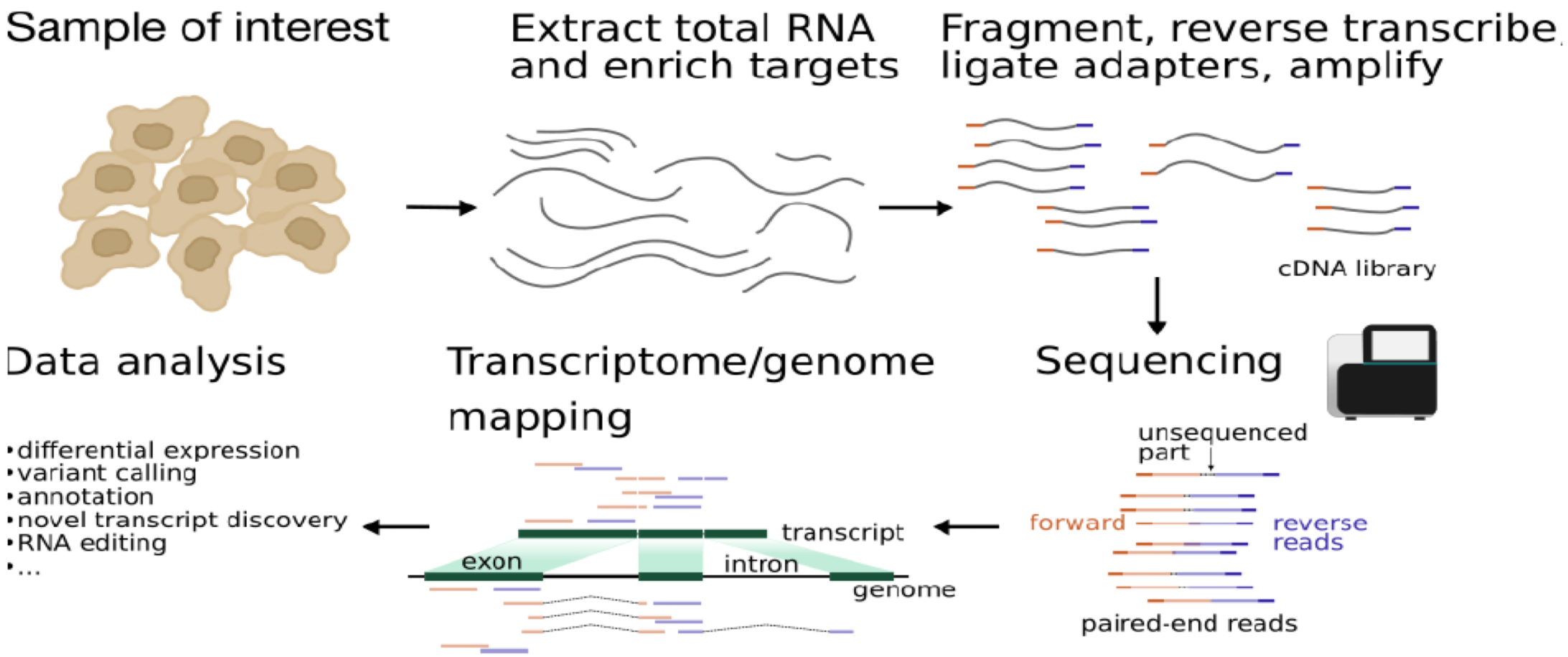
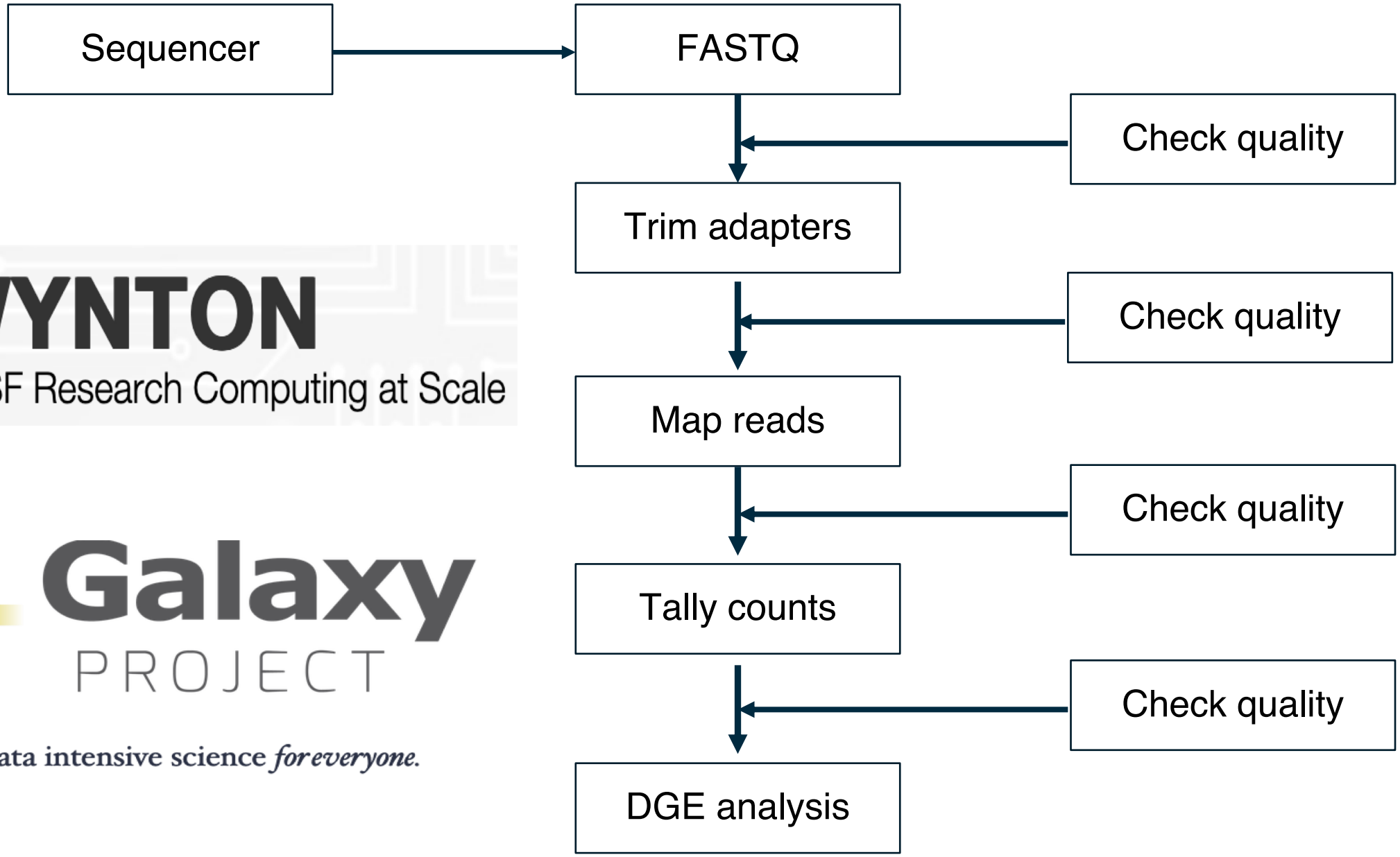


Figure: Berge et al., 2018, PeerJ Preprints.



WYNTON
UCSF Research Computing at Scale

Galaxy
PROJECT

Data intensive science *for everyone.*

Bioinformatics software ecosystem: Tools that “do one thing, and do it well”.

- Tools for today: fastqc, cutadapt, STAR, samtools, featureCounts
  - Available on Galaxy
  - 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

Galaxy provides a collection of tools to analyze variety of biomedical data

- https://usegalaxy.org/

Command line interfaces allow scripting

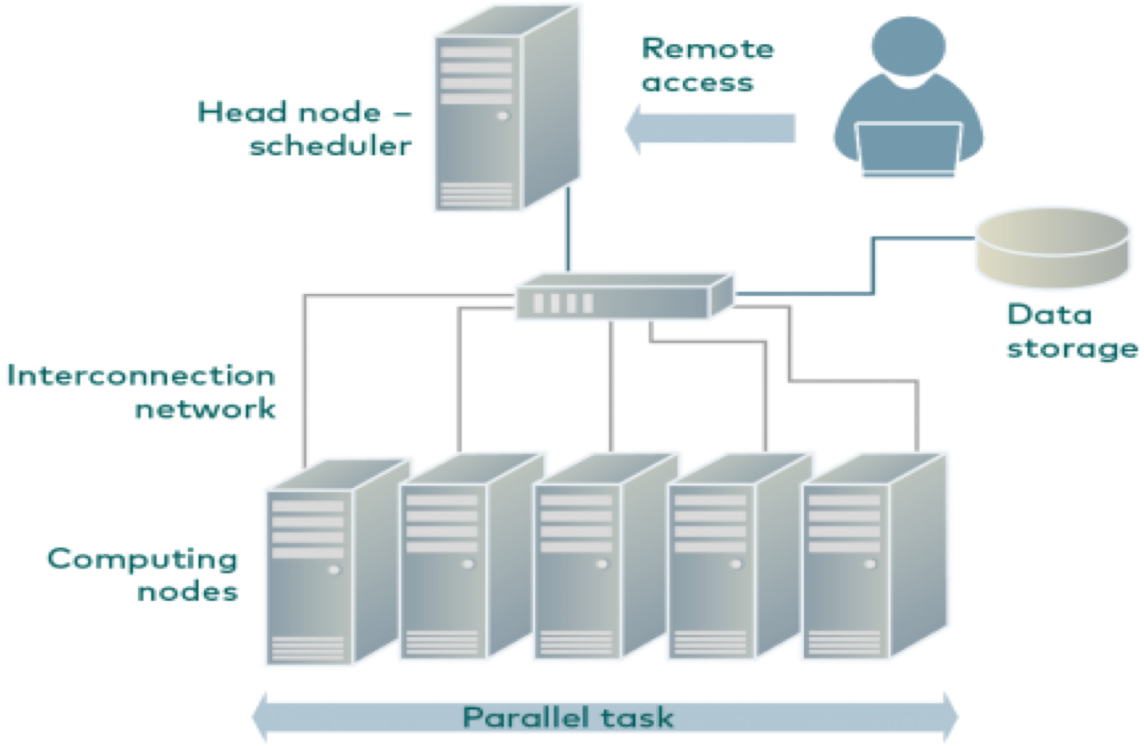
- Graphical User Interface
  - Consists of windows, icons, menus, pointers
  - Not always available for bioinformatics

- 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)

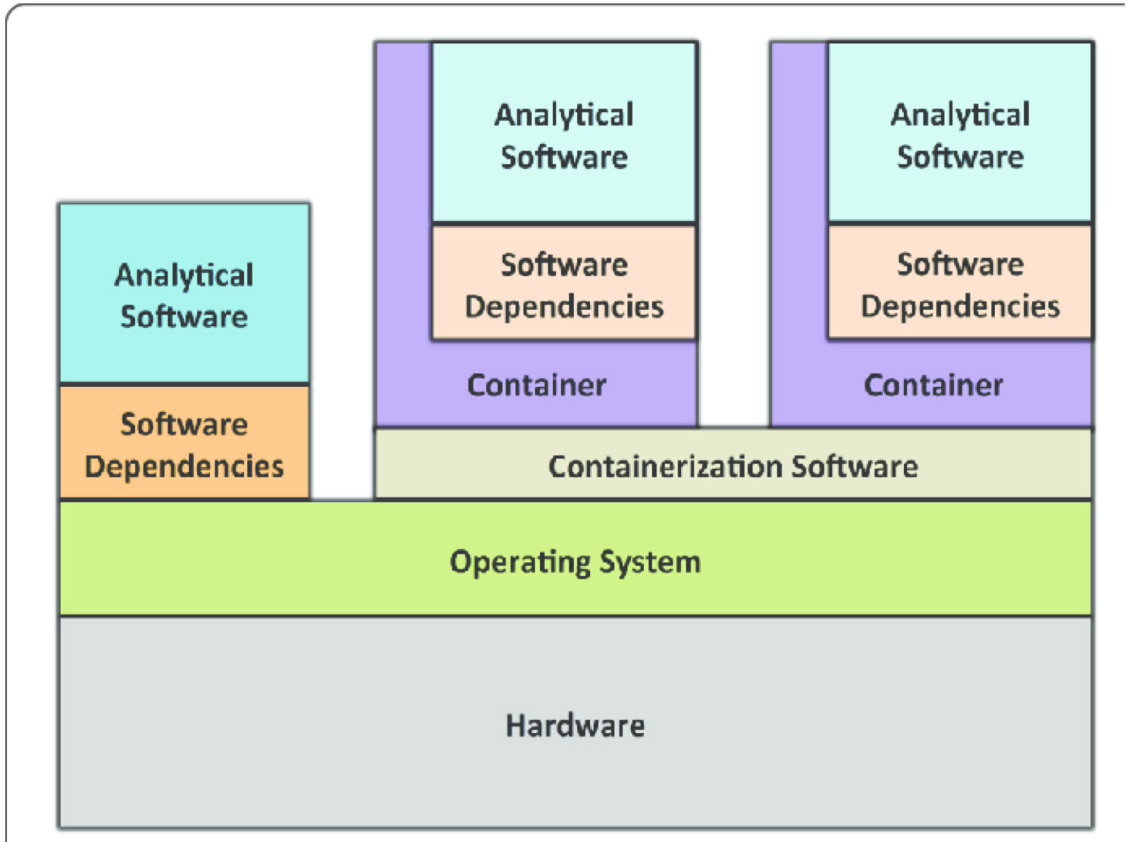
Containers enable reproducibility

- 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

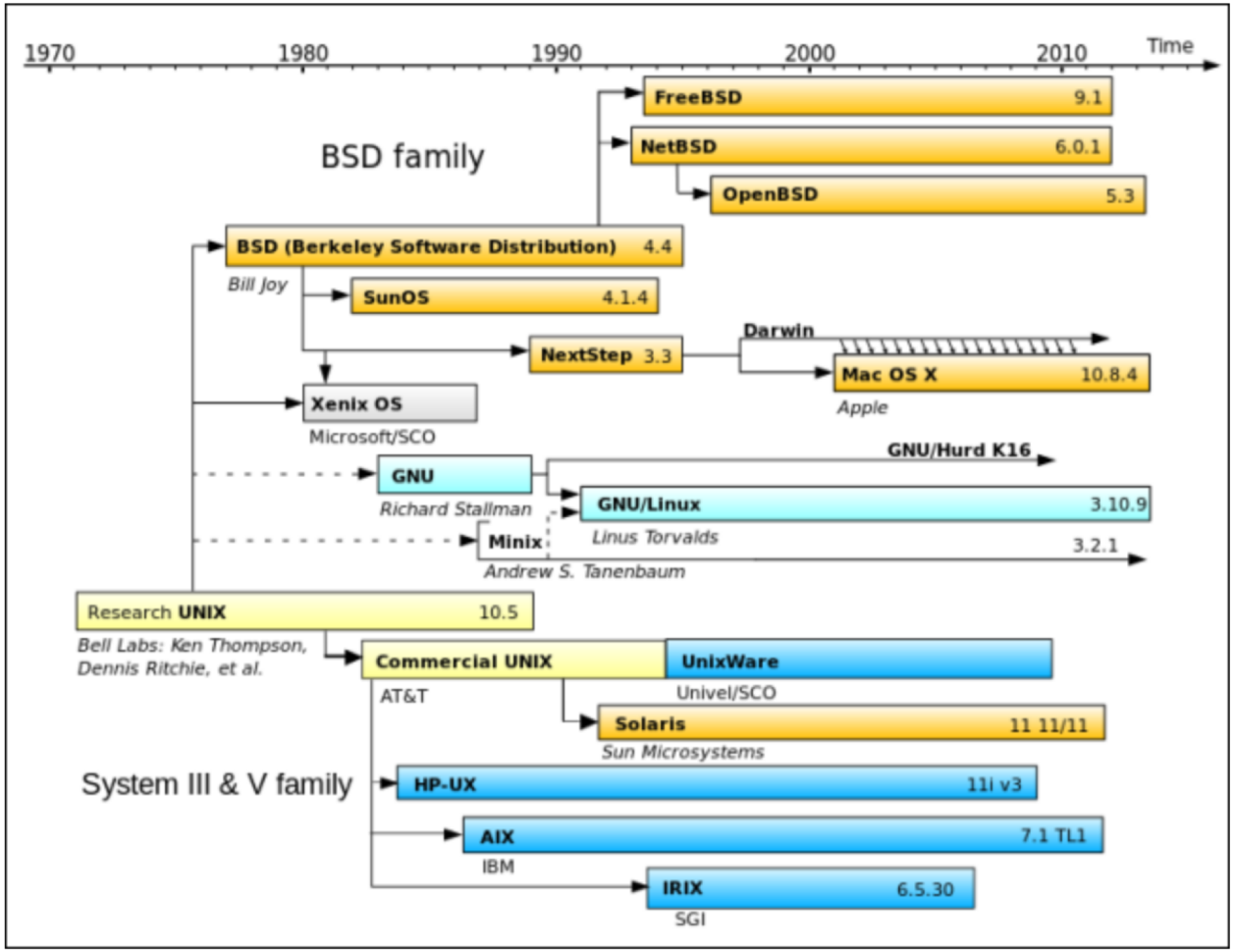
- Containerization software examples:
  - Singularity
  - Docker (has security issues in the context of HPC)
- Containers can be deployed on commercial cloud computing platforms, e.g., Amazon Web Services



(see Description for image source)

Best to use singularity with Linux (currently)

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



(see Description for image source)

Problem definition, shared data files

Problem: identifying differentially expressed genes

- Other applications
  - annotate novel transcriptional events, e.g., exon skipping, alternative 3’ acceptor or 5’ donor sites, intron retention
  - analysis of genetic variation among expressed genes
  - RNA editing events
  - characterization of long noncoding RNAs
  - …

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 by Reuben Thomas:
  Intro to statistics and experimental design.
- Reading material:
  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.

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.

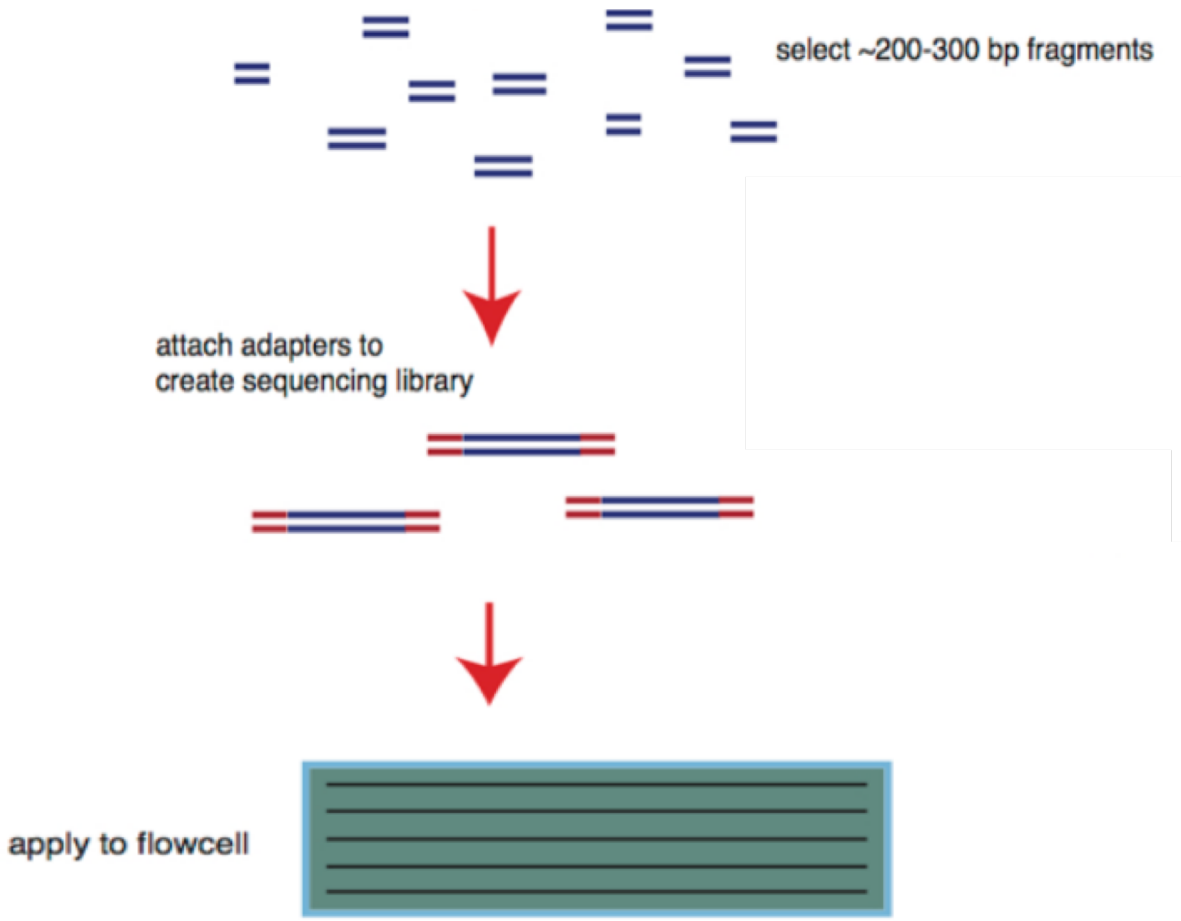


Image adapted from a blog by Stuart M. Brown (link in description).

Flow cells are organized in lanes, columns and tiles.

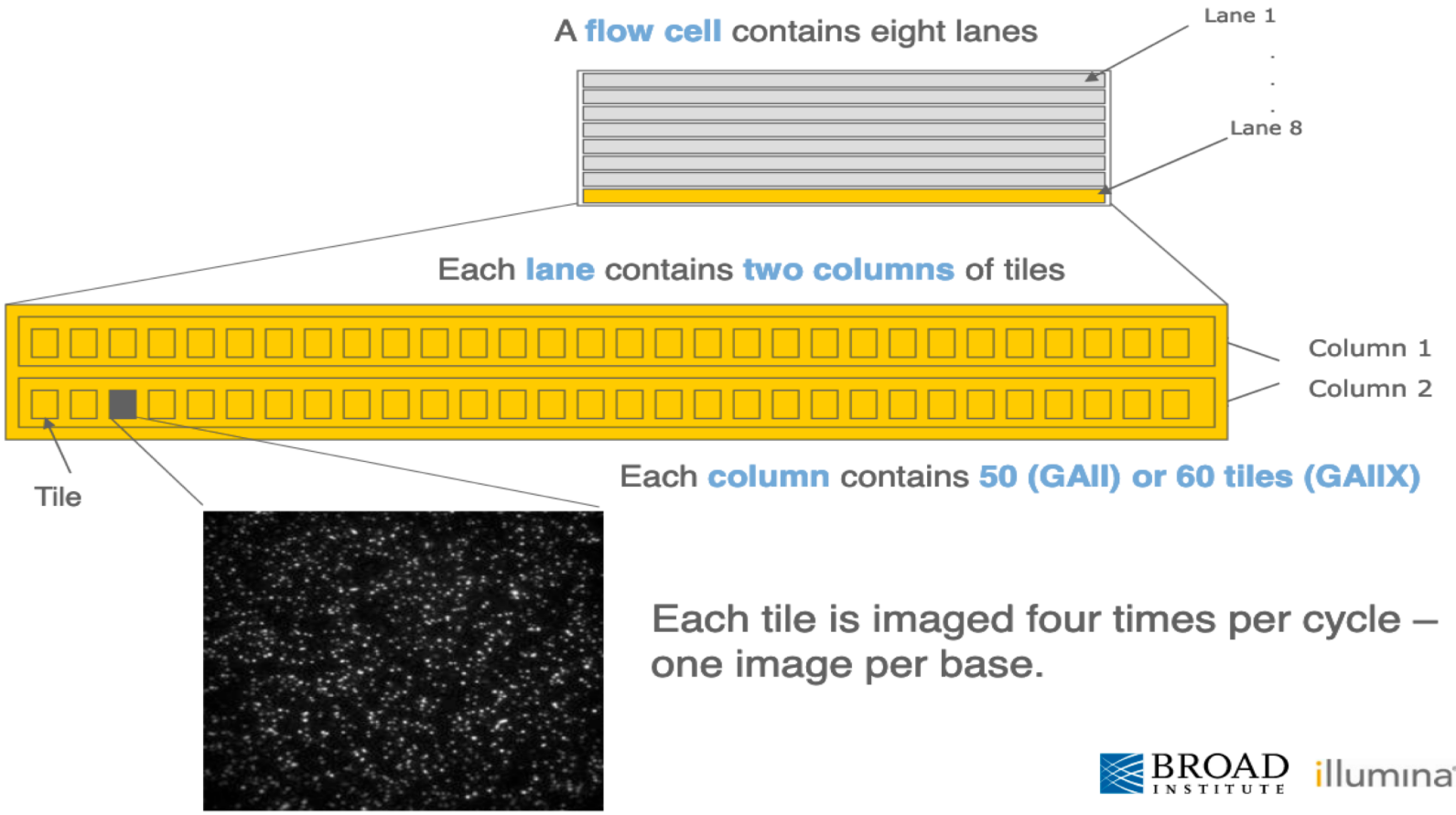


Image borrowed from slides shared by Broad Institute (link in description).

DNA fragments immobilized on flow cell

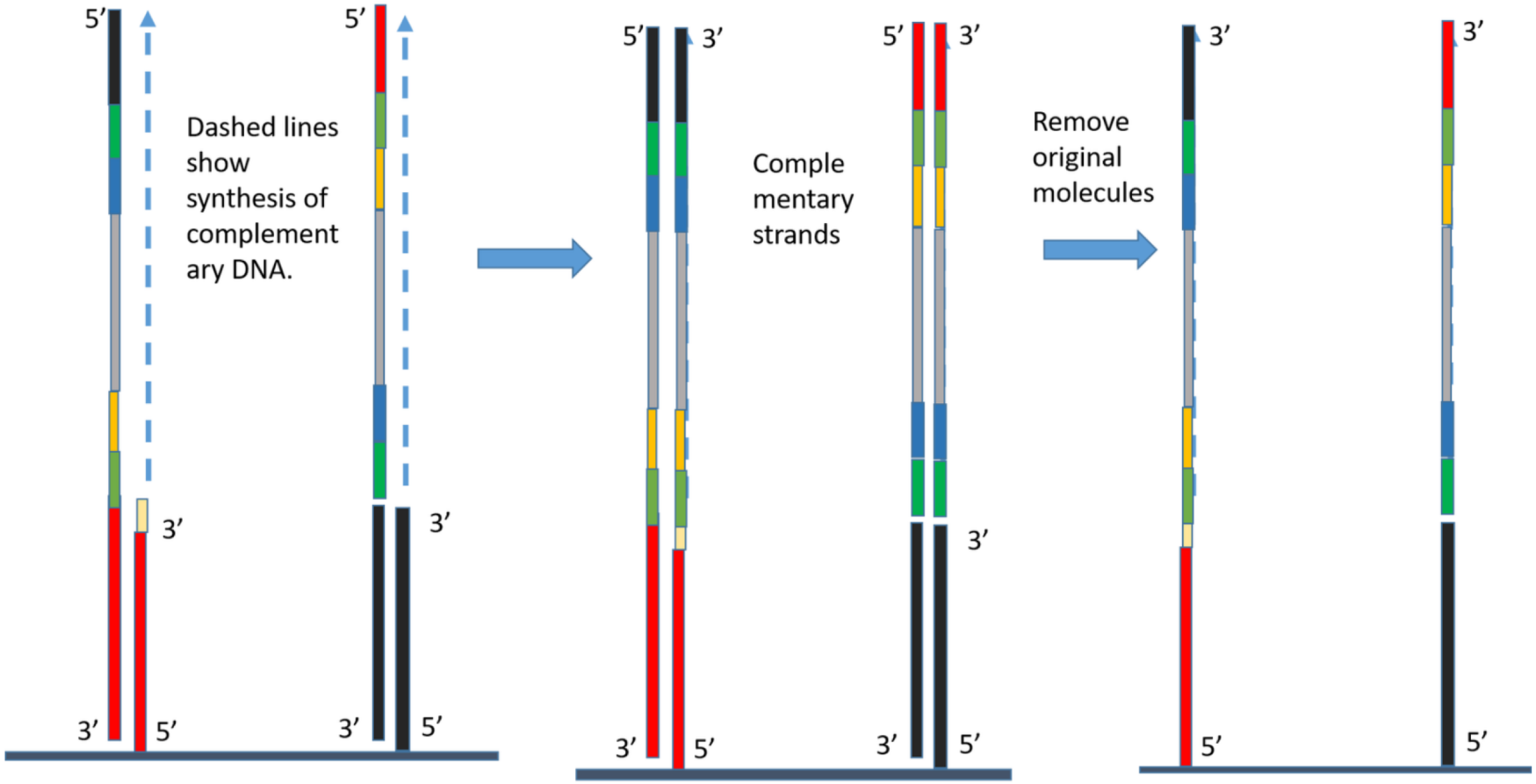


Image from blog by Lauren Launen (link in description).

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

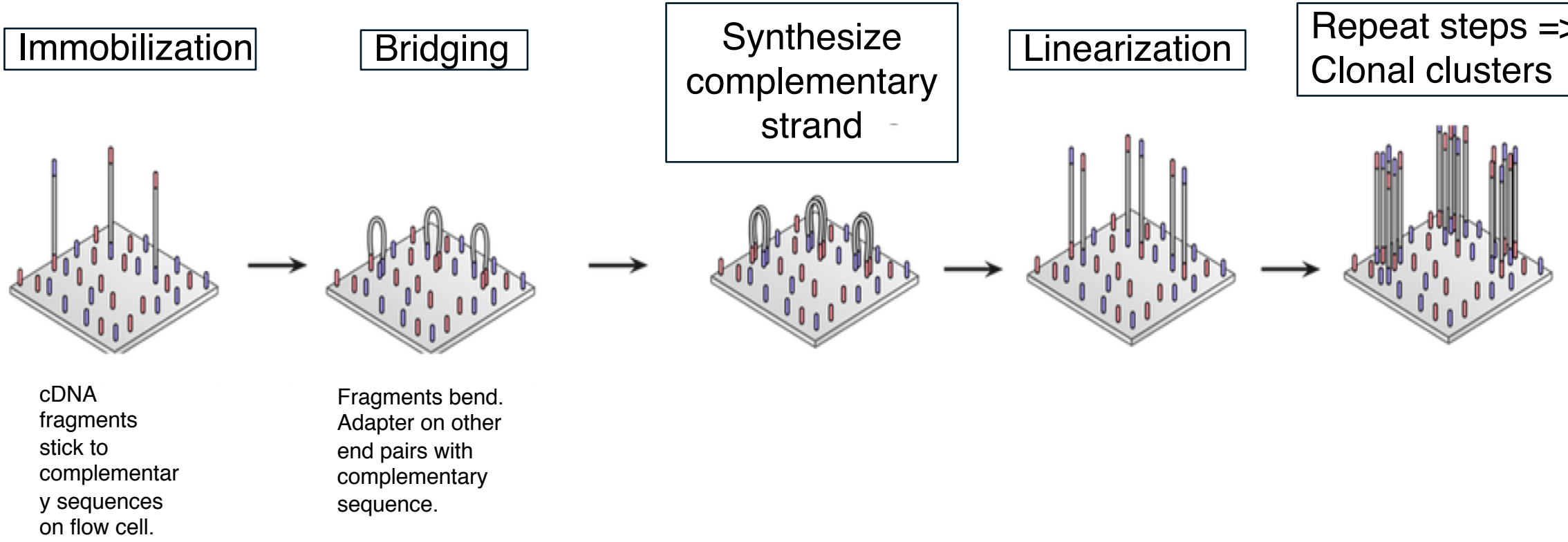
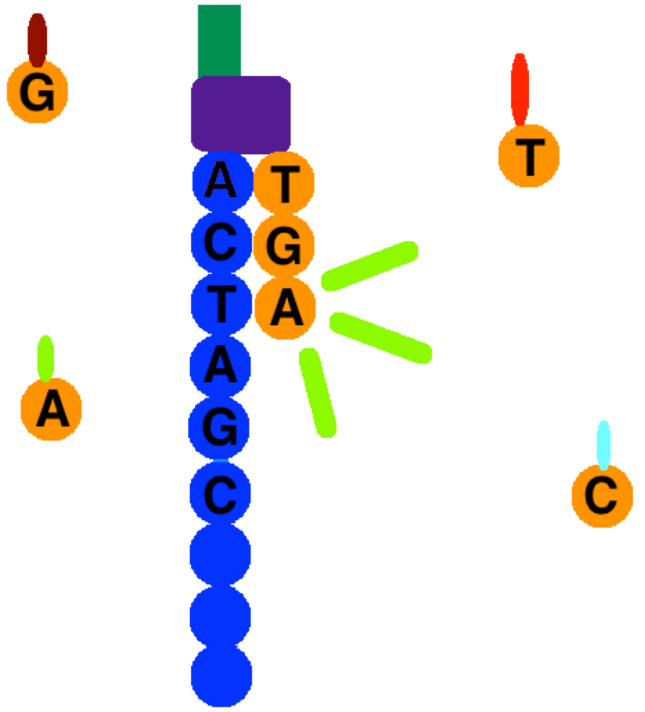


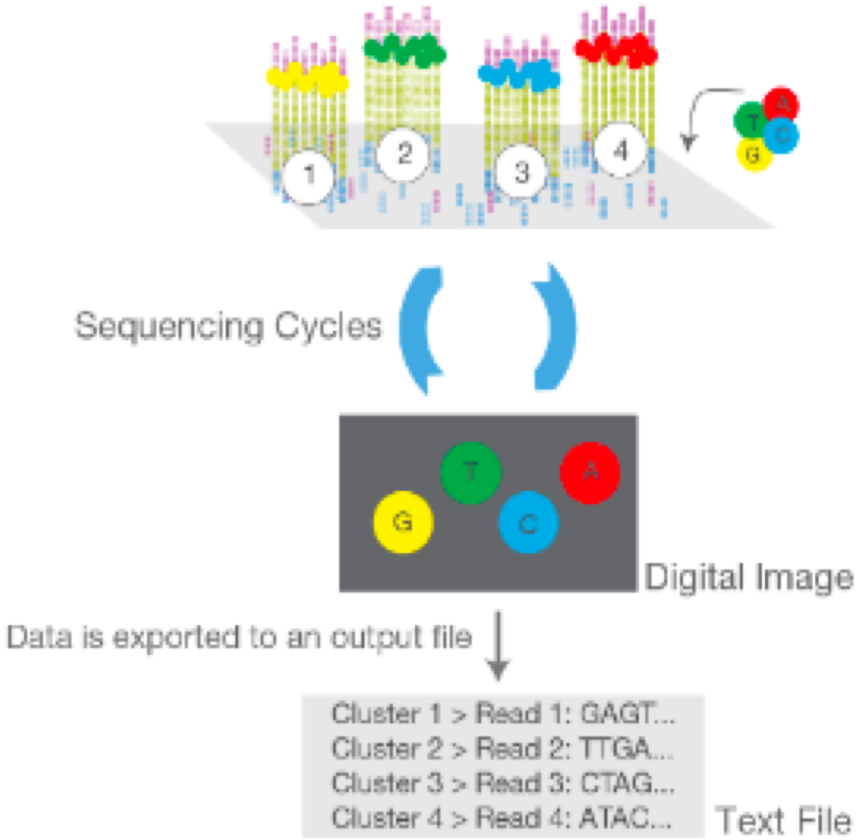
Image adapted from blog by Lauren Launen (link in description).

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.



Strong signal from monoclonal clusters.



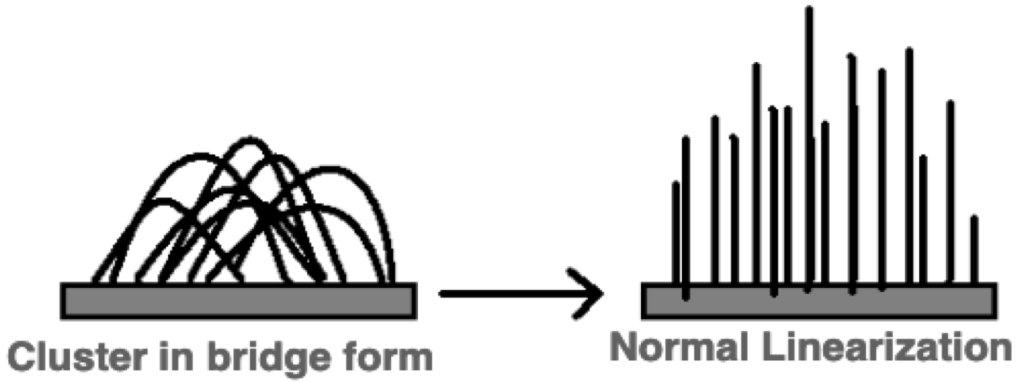
FASTQ files contain detailed information about each read.

- 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

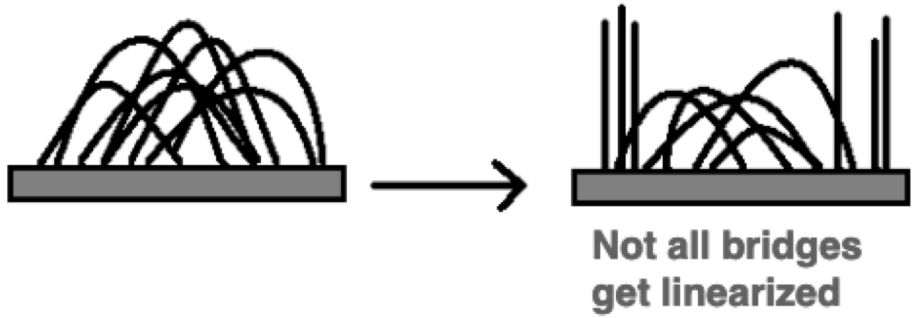


Image adapted from a Broad Institute presentation (link in description).

Base calling may not be accurate.

Various possible causes: Example

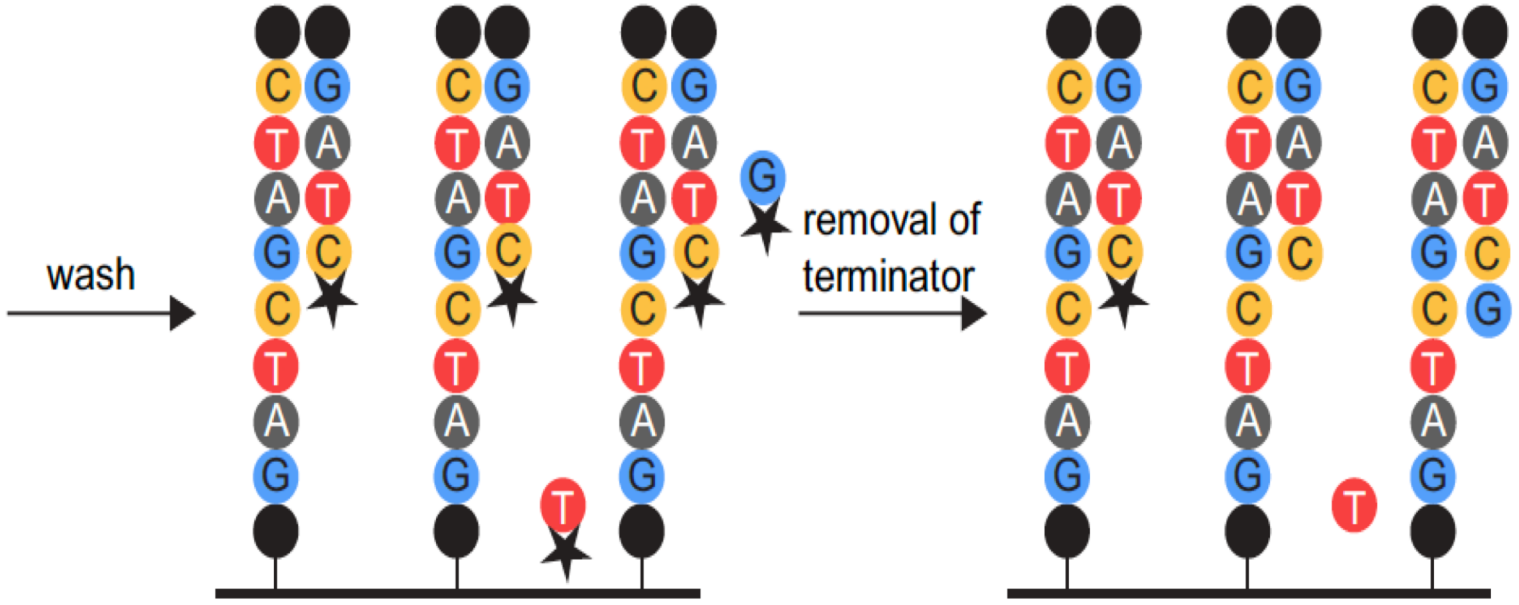
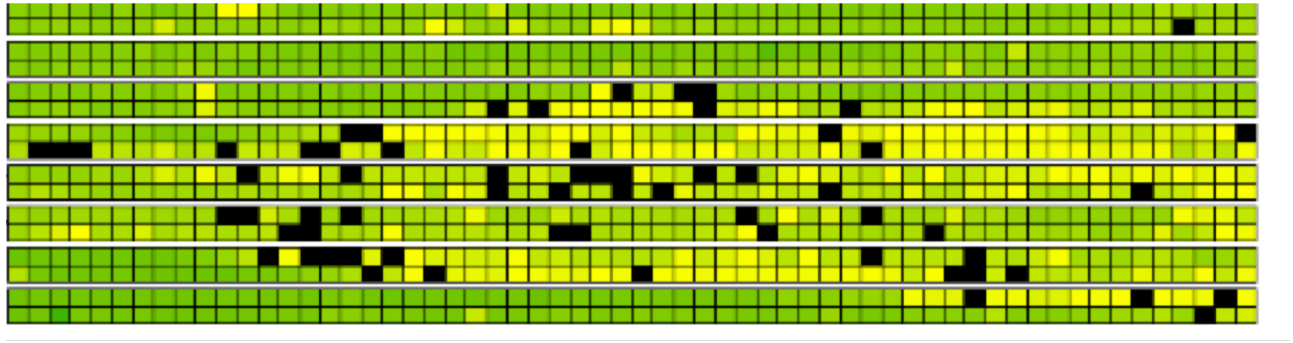


Image adapted from Pfeiffer et al., Scientific Reports, (2018) 8:10950.

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.

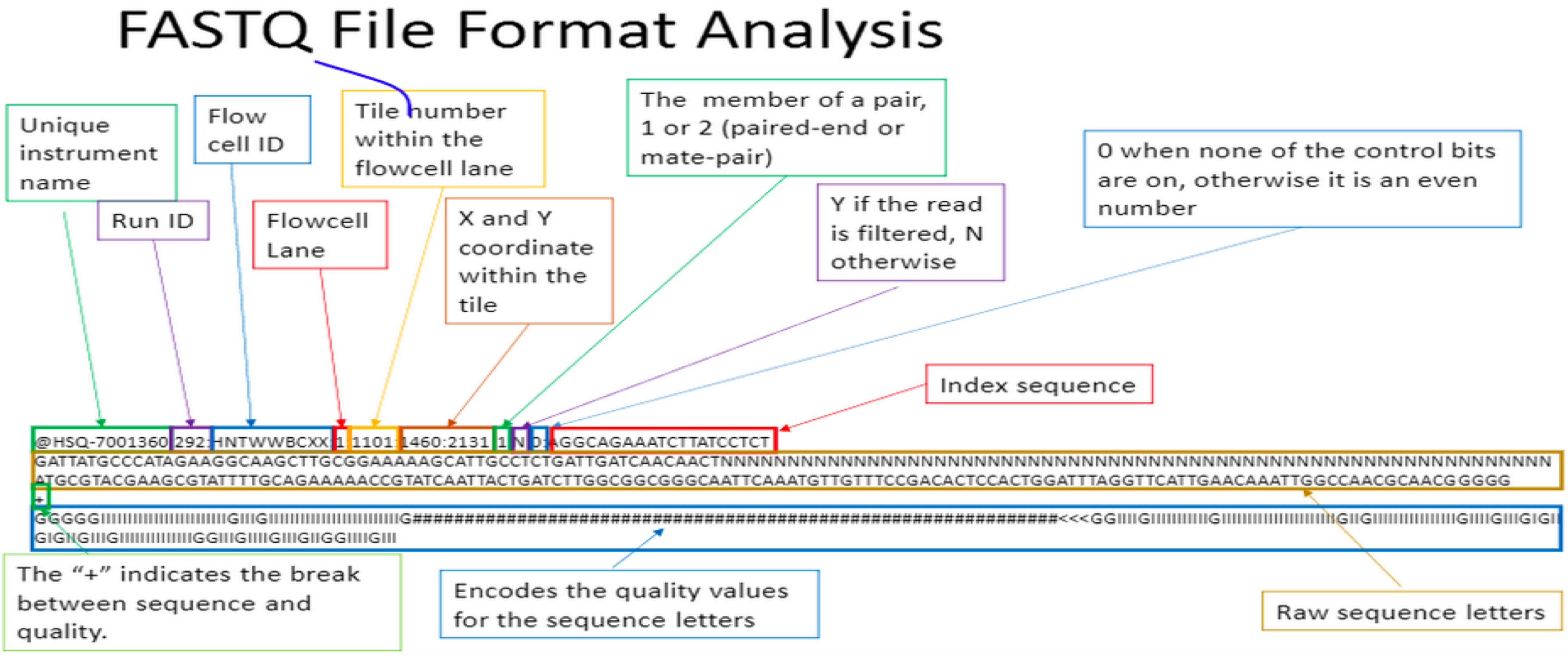


Image adapted from blog by Lauren Launen (link in description).

Quality is encoded as symbols.

Quality measured in terms of Phred scores.

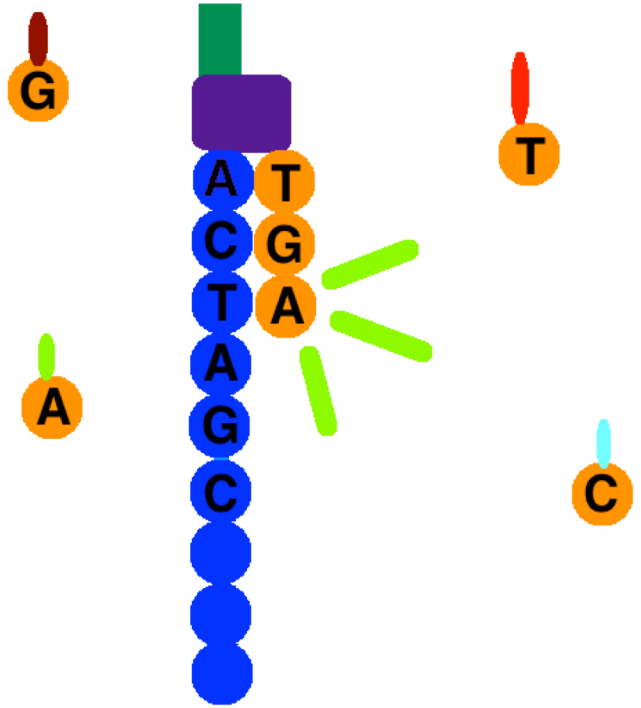
Symbol	Q-Score	Symbol	Q-Score
!	0	6	21
"	1	7	22
#	2	8	23
$	3	9	24
%	4	:	25
&	5	;	26
'	6	<	27
(	7	=	28
)	8	>	29
*	9	?	30
+	10	@	31
,	11	A	32
-	12	B	33
.	13	C	34
/	14	D	35
0	15	E	36
1	16	F	37
2	17	G	38
3	18	H	39
4	19	I	40
5	20		

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.



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 known adapters.
- Any sequences more frequently observed than expected?
- 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.

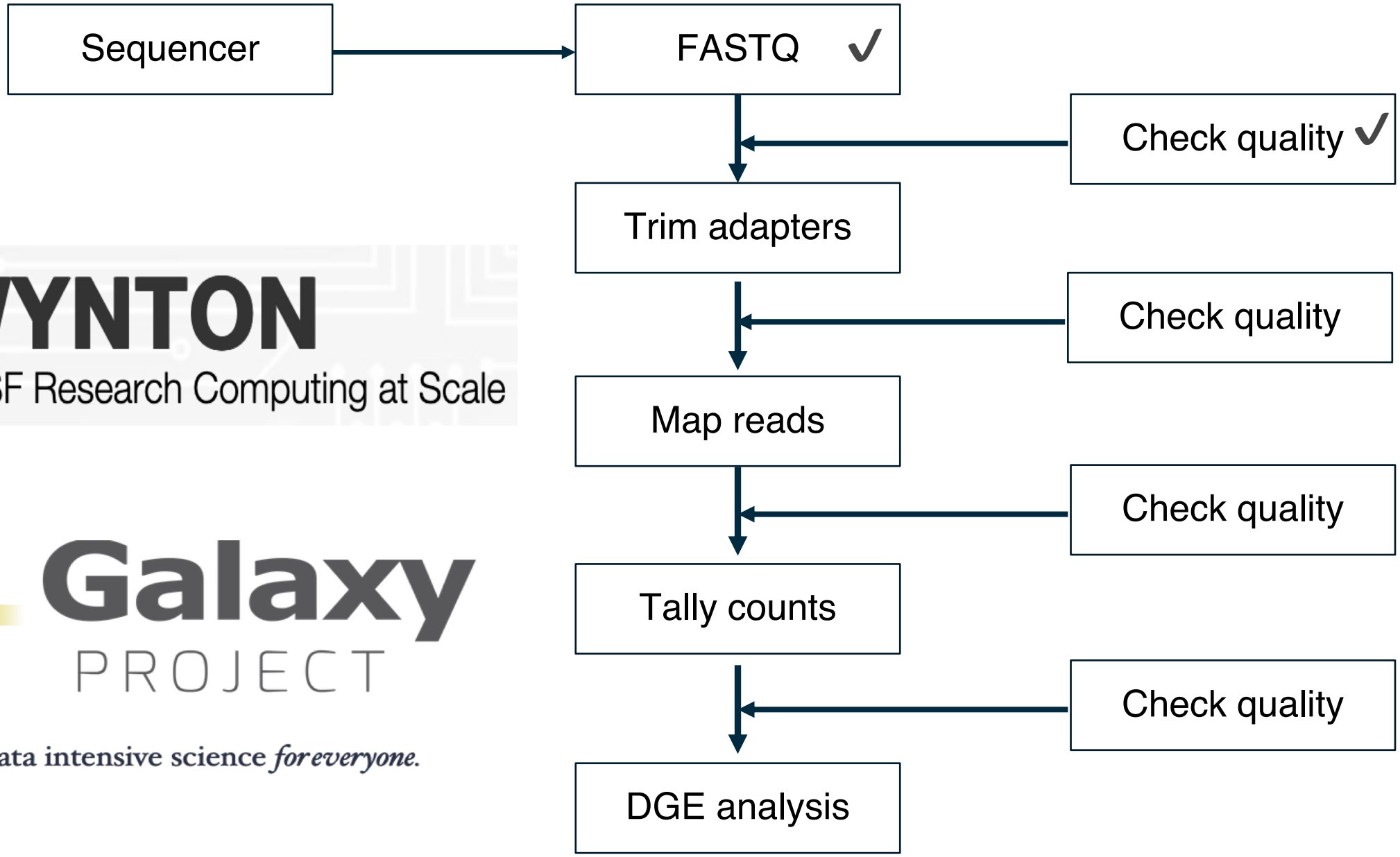
Examples of FastQC reports

- 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 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/



WYNTON
UCSF Research Computing at Scale

Galaxy
PROJECT

Data intensive science *for everyone.*

Conclusion: Session 1

- Bioinformatics software ecosystem: Tools that “do one thing, and do it well”.
  - Multiple tools are available for the same task.
- HPC clusters enable data-intensive science.
- Containerization facilitates reproducibility.
- Always check “quality” after each step.
- Tools and file formats used:
  - fastqc
  - FASTQ

Contents: Session 2

- Resources for after the workshop
- Mapping trimmed reads
  - cutadapt, STAR or bowtie2, samtools
  - FASTA, BAM/SAM, GFF
- Tallying gene-wise counts
- Maybe:
  - Shell scripting
  - Submitting jobs to compute nodes
  - Nextflow

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
- …

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

How much programming skills do we need for bioinformatics?

- Minimum essential
  - Introductory R
  - Introductory command line (link to a cheatsheet in speaker notes)
- Available at
  - Gladstone Data Science Training program
  - Data Science workshops from the UCSF library
    - https://www.library.ucsf.edu/ask-an-expert/classes-catalog/
- For RNA-seq data analysis beyond tallying gene-wise counts:
  - Intermediate RNA-seq
  - Pathway analysis

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

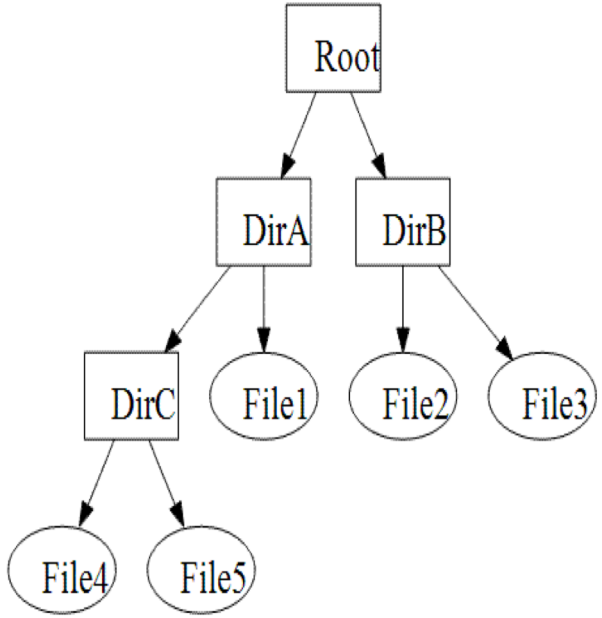
- Trimming adapters

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

- Similarly for other tools

File paths := Location of file on computer

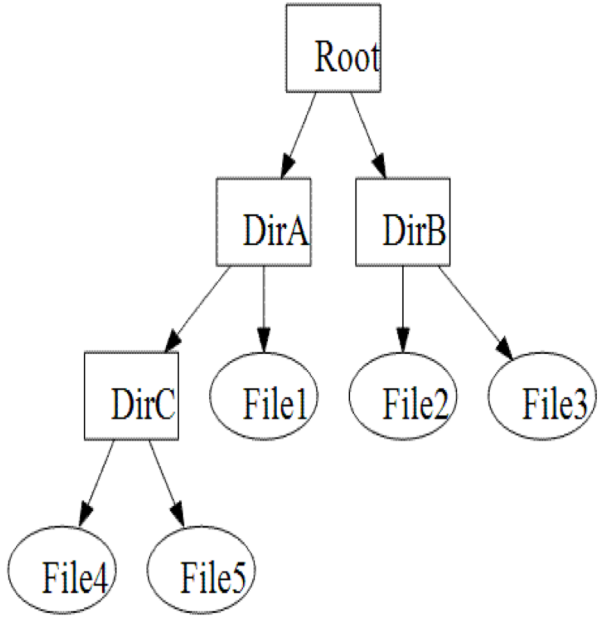
- Lots of files on a computer
- Organized in directories which may contain sub-directories
- Bioinformatics tools may need file inputs and may output files
  - Where are the input files located on the computer?
    - Searching entire computer not practical
    - What if multiple files have the same name?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



Submitting jobs for execution on compute nodes: Same syntax as that for bioinformatics tools

```
$ qsub -cwd -pe smp 4 -l mem_free=2G -l
scratch=50G -l h_rt=00:20:00 script.sh
```

- For details visit:
  - https://wynton.ucsf.edu/hpc/scheduler/submit-jobs.html

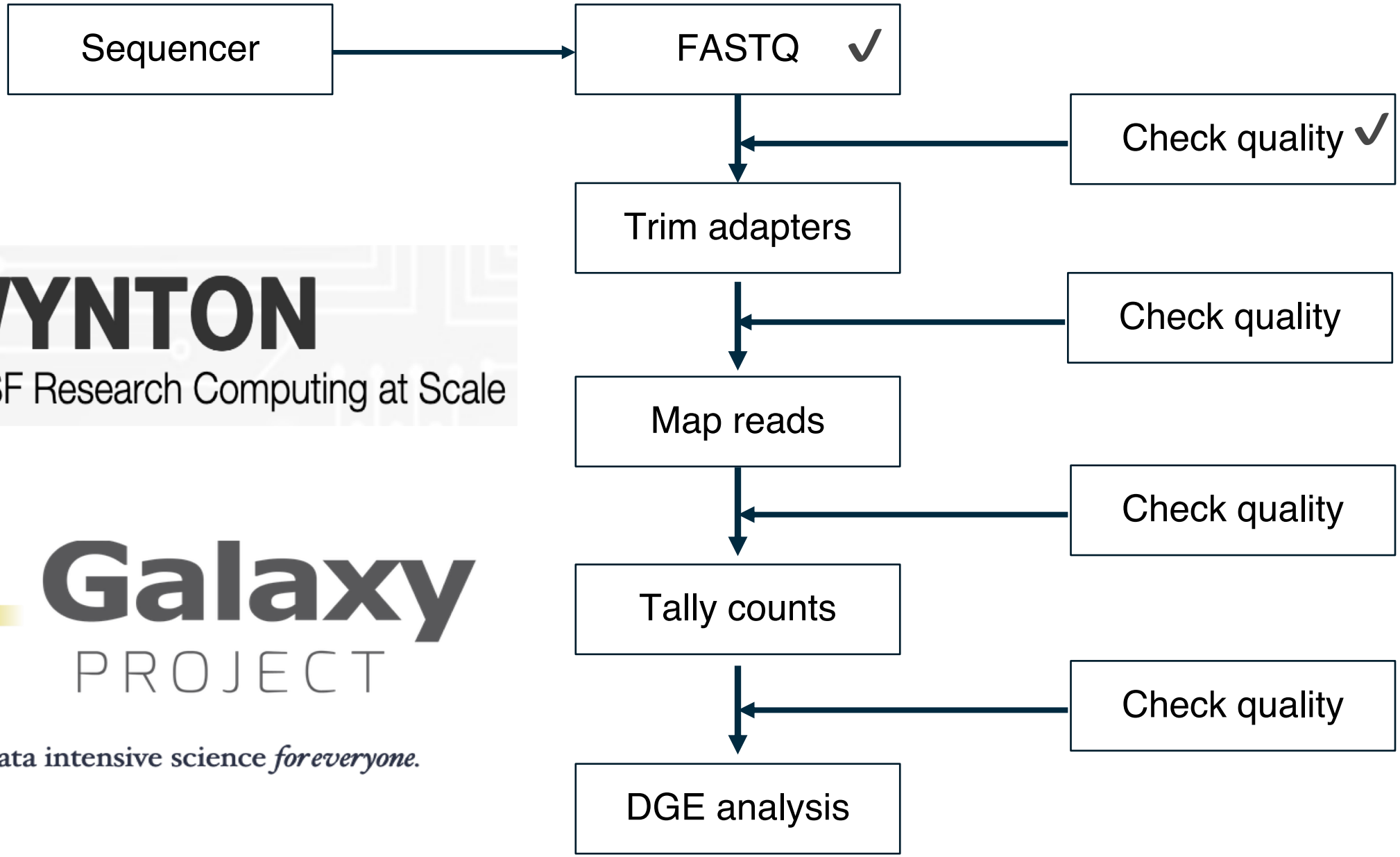
Schedulers maintain a queue of jobs

- Many users are submitting jobs to Wynton
  - Different resource requirements
  - In what order should the jobs run?
- HPC administrators decide which scheduler to use
- Popular schedulers
  - SGE
  - Slurm
  - PBS Pro

Workflow managers: One command with all required inputs for a pipeline

- Examples:
  - Nextflow
  - Snakemake
  - Cromwell
- See link in notes for how to run a nextflow pipeline on Wynton
- Example usage:

```
$ nextflow run nf-core/rnaseq \
> --reads '*_R{1,2}.fastq.gz' \
> --genome GRCh37 \
> -profile singularity
```



WYNTON
UCSF Research Computing at Scale

Galaxy
PROJECT
Data intensive science *for everyone.*

Important!

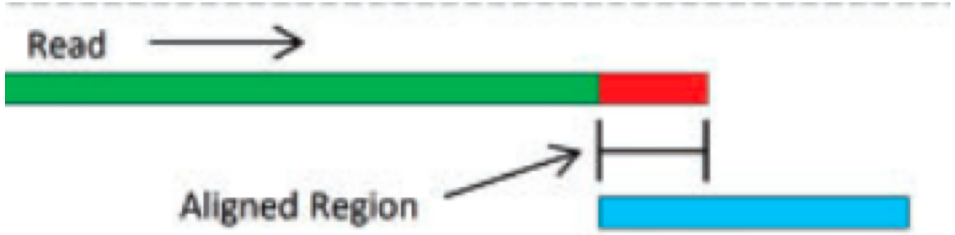
- Know the tools you are using
- Know what each tool is doing
- Know the best practices for analysis
  - Read benchmarking papers
  - Read papers providing practical guidelines
  - Read papers reviewing common practices

Cleaning up contaminants (20 mins)

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

cutadapt removes adapters.

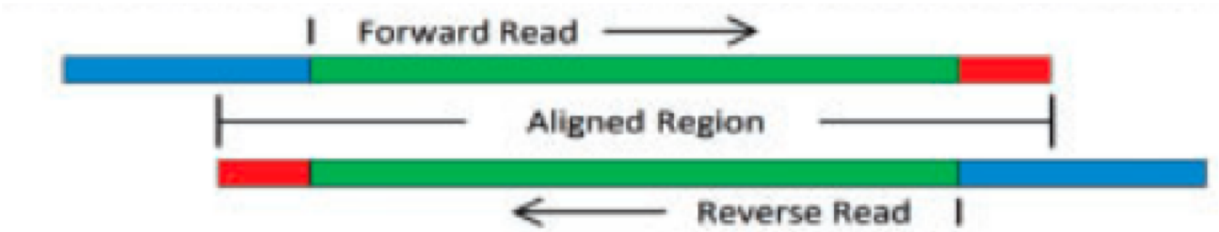
- Search for adapter sequence in read.
- Allow for mismatches in sequence.
- If significant alignment, cut.

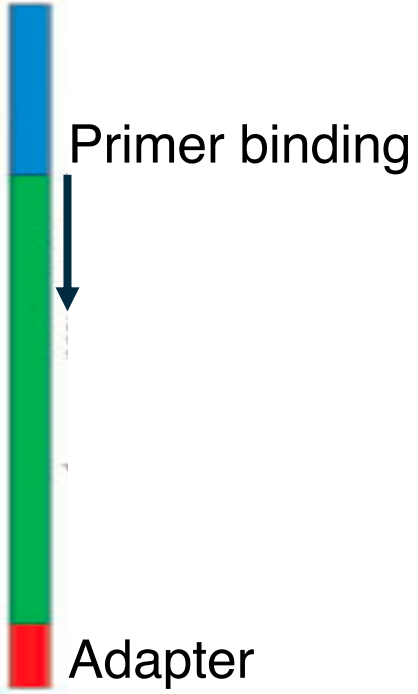


(Image adapted from Bolger et al., 2014, Bioinformatics. Link in description.)

Alternative approach: Trimmomatic

- Say adapter sequence in read is very short.
  - Can we still identify it?
- Yes for paired-end reads.





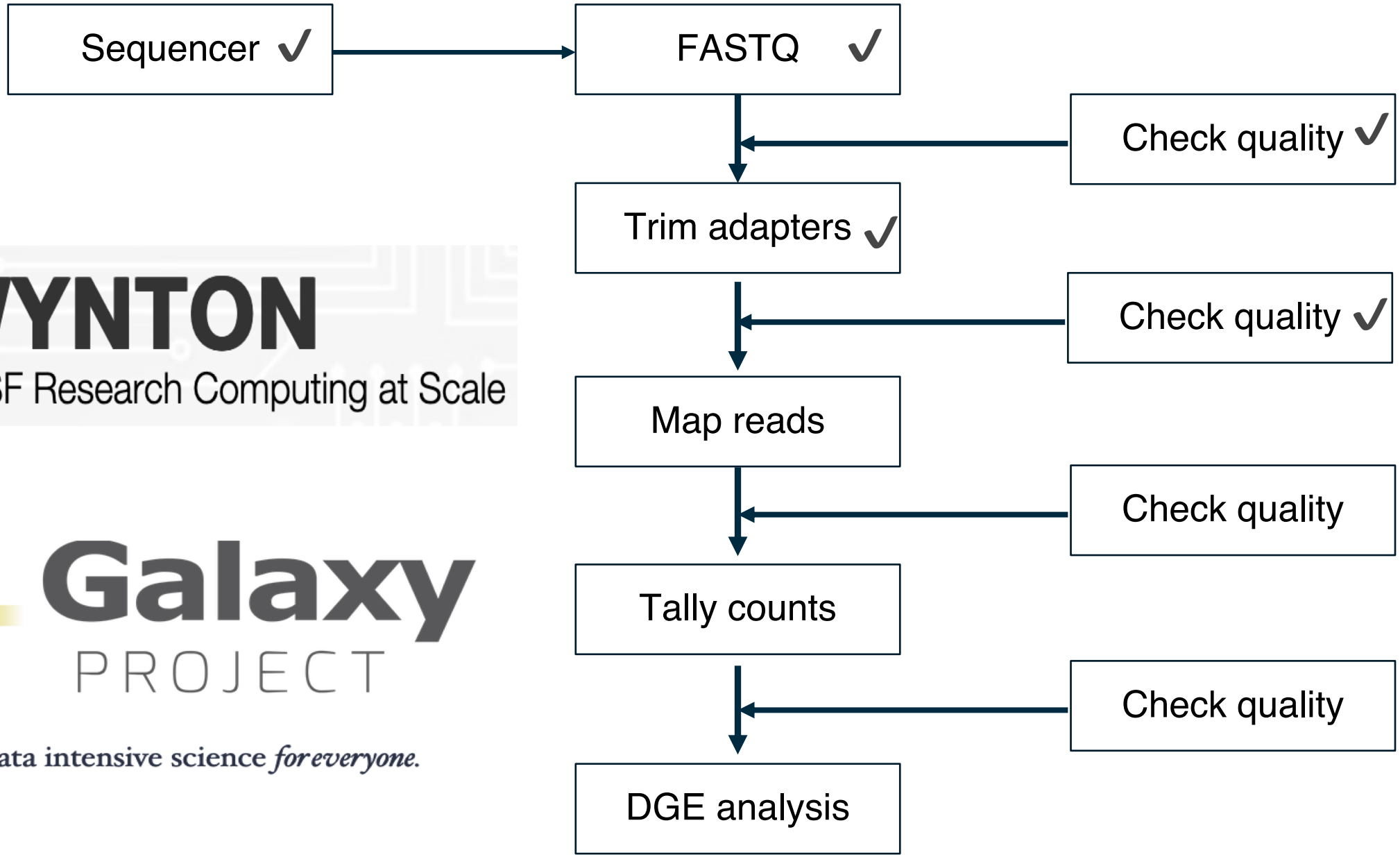
(Image adapted from Bolger et al., 2014, Bioinformatics. Link in description.)

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?



WYNTON
UCSF Research Computing at Scale

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.
  - Need to account for splicing.
  - Allow for PCR artifacts/sequencing errors.

Inputs needed.

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

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 STAR or bowtie2 to build index.
- 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

- 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).

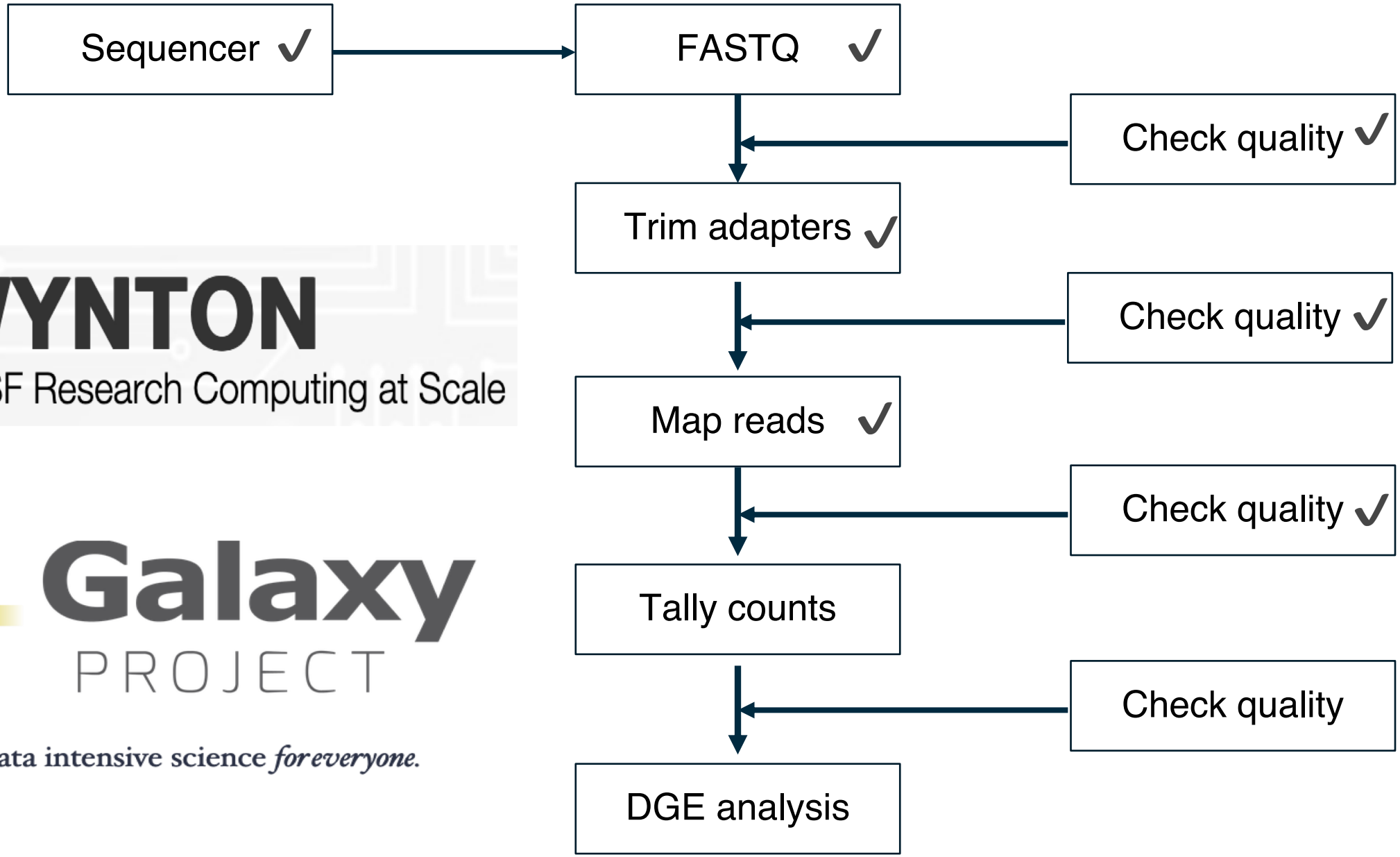
Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[9]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	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

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

Tools to manipulate files are available.

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



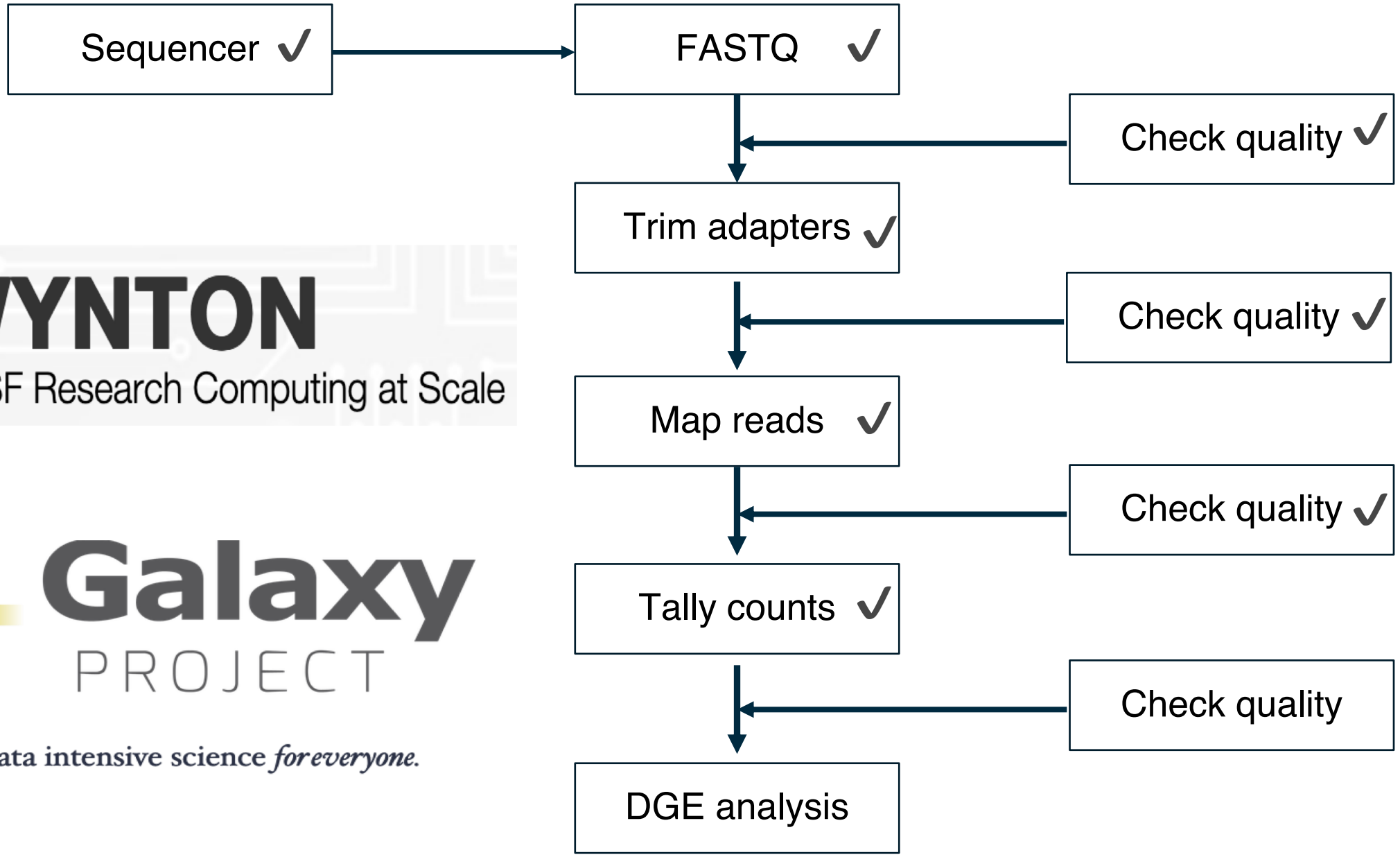
WYNTON
UCSF Research Computing at Scale

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.



WYNTON
UCSF Research Computing at Scale

Galaxy
PROJECT
Data intensive science *for everyone.*

Downstream analysis (~15 mins)

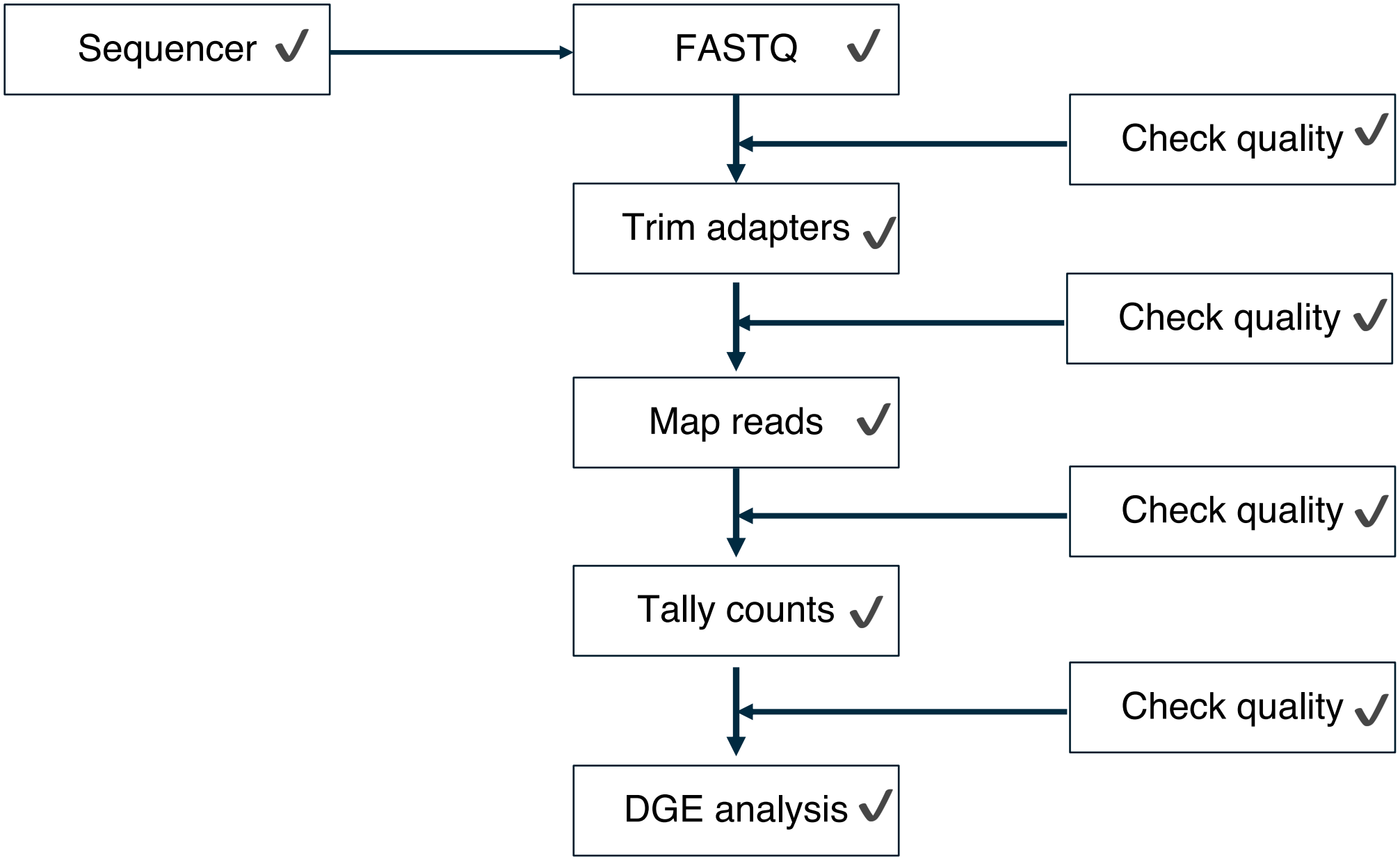
No. 1: Differential gene expression analysis.

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 ✓

Your feedback is important to us!

- https://www.surveymonkey.com/r/RRTZPTC
- ~3 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 and Wynton.

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
- Consider parallel-fastq-dump if SRA toolkit is too slow

More tools

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

Upcoming Workshops

- Single cell RNA-seq data analysis
- Machine Learning

Thank you!


GLADSTONE
INSTITUTES