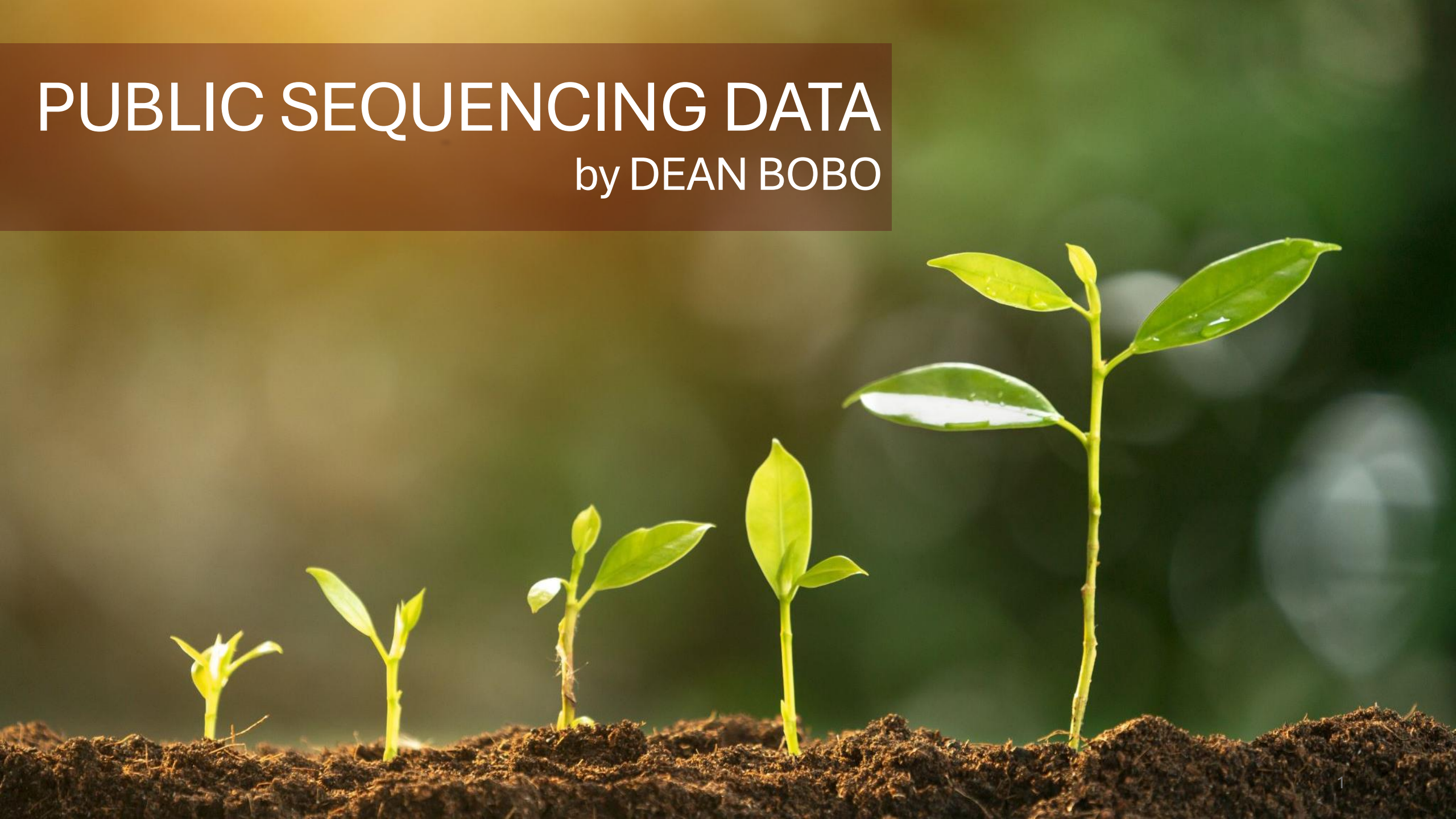
PUBLIC SEQUENCING DATA

by DEAN BOBO



slides available on

deanbobo.com

Topics & Objectives

Overview of public repositories available

searching for data

Review of next-generation sequencing data

Download an example dataset

Perform QC on data

Discussion: What to do with NGS data?

De novo assembly is next lecture

Live Google Doc

https://docs.google.com/document/d/13Ka7L8aIOOMhUYB_CngSQUn3GEHiRT4qvlkZ7Vb-1S0/edit?usp=sharing

Public Repositories

NCBI Sequence Read Archive (SRA)

European Nucleotide Archive

DNA Data Bank of Japan

NCBI SRA

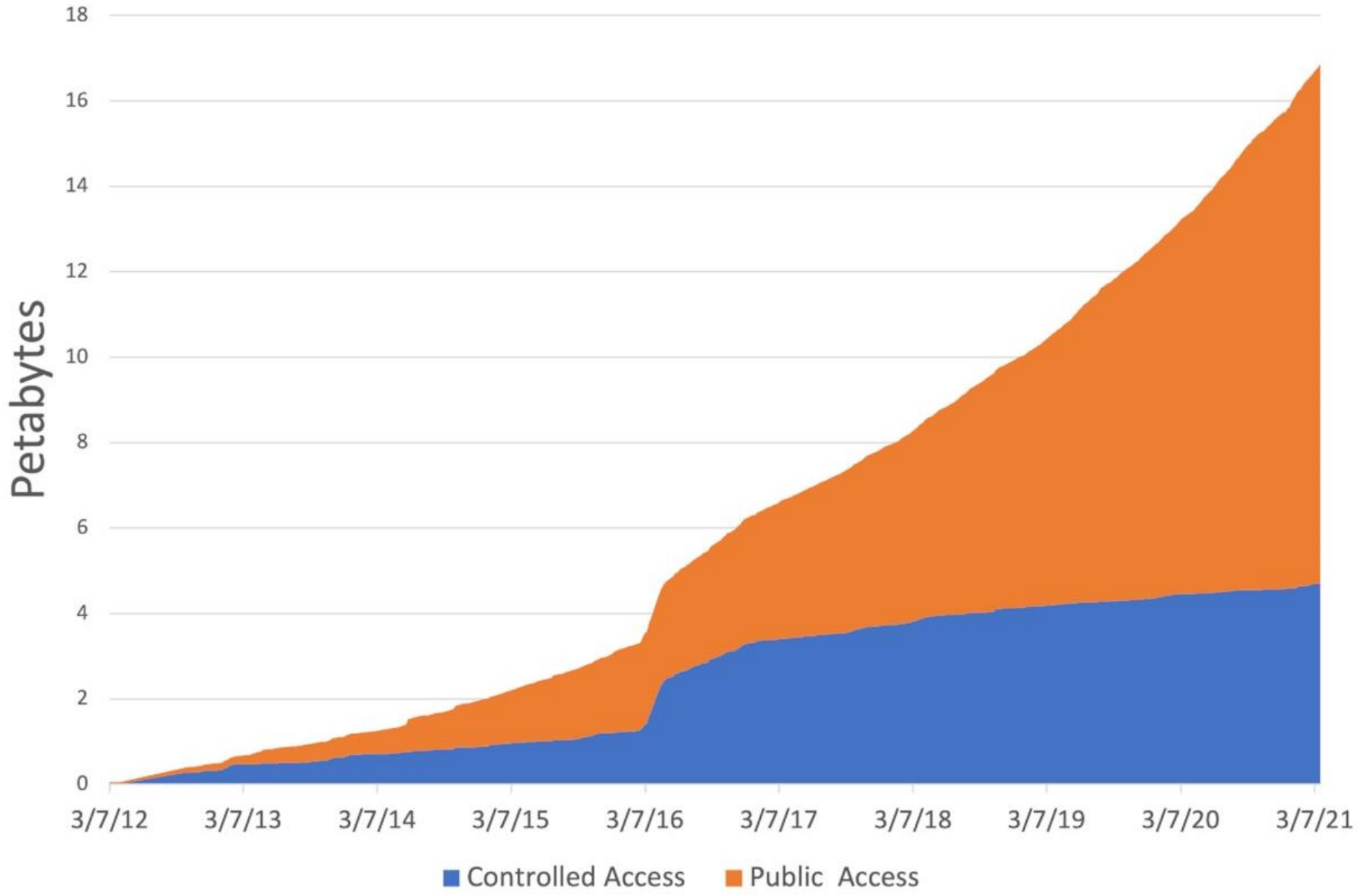
Largest repository of raw sequencing data

Data from Illumina, PacBio, Oxford Nanopore, etc.

Data accessible from:

- SRA Toolkit
- FTP
- web interface

Growth of SRA over the last decade



Metadata	Description
Study (SRP)	A study is a set of experiments and has an overall goal.
Experiment (SRX)	An experiment is a consistent set of laboratory operations on input material with an expected result.
Sample (SRS)	An experiment targets one or more samples. Results are expressed in terms of individual samples or bundles of samples as defined by the experiment.
Run (SRR)	Results are called runs. Runs comprise the data gathered for a sample or sample bundle and refer to a defining experiment.

SRA Web Interface

https://www.ncbi.nlm.nih.gov/sra

let’s go search around!

Another way to search

https://www.ncbi.nlm.nih.gov/taxonomy

let’s go search around again!

SRA Toolkit

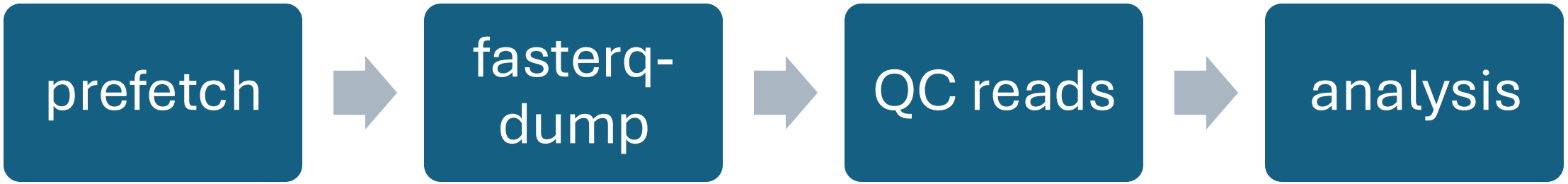
command-line tools to download and process sequencing data from SRA

prefetch : download SRA format for a run

fasterq-dump : convert SRA to fastq

vdb-config : configure cache and access settings.

Pipeline Overview



$ vdb-config

module load sratoolkit-3.1.1

vdb-config --help

vdb-config -i

then press c for cache

$ prefetch

module load sratoolkit-3.1.1

prefetch --help

prefetch SRA_ID

I recommend using -o or -O

-o [output_file]

-O [output_directory]

$ fasterq-dump

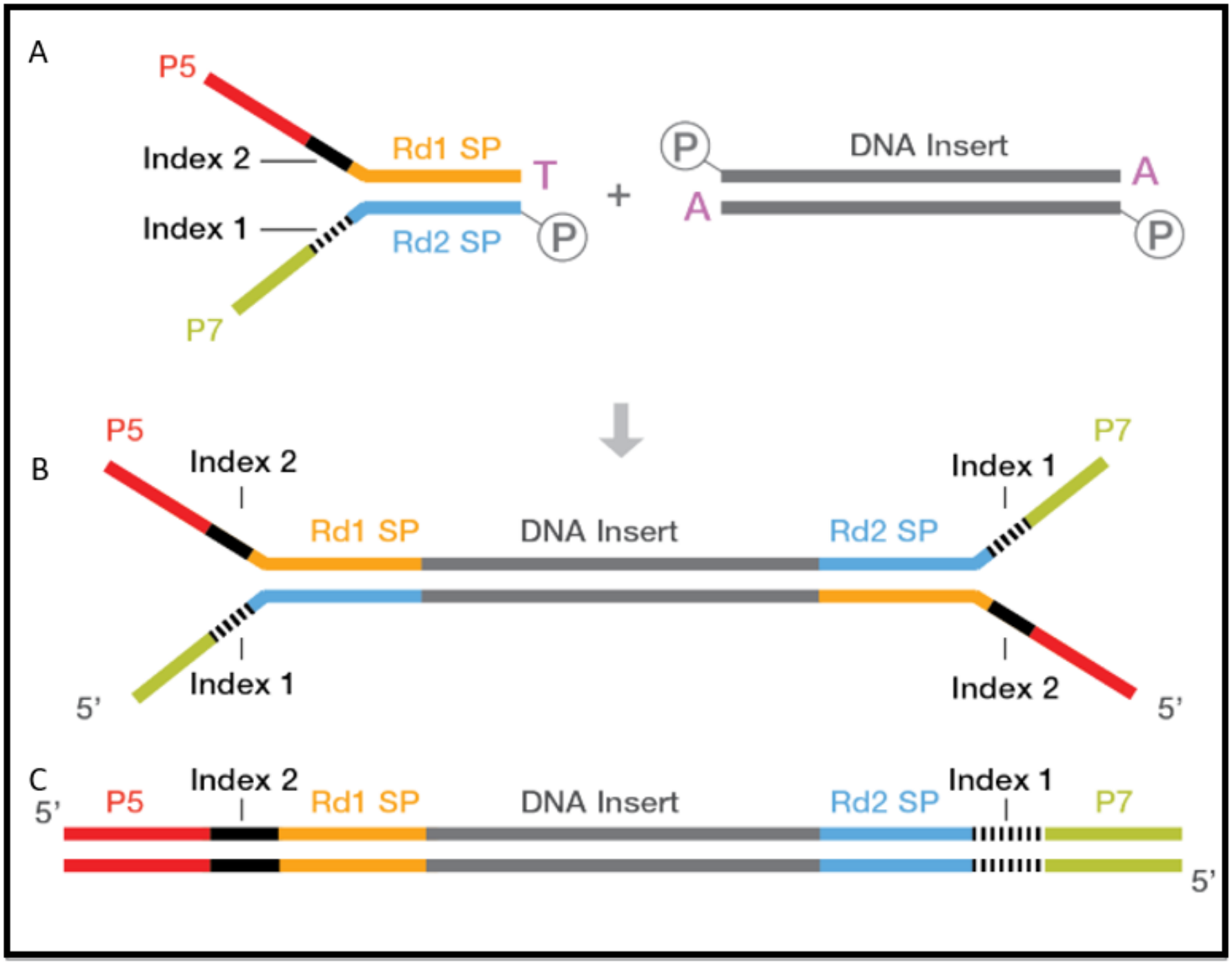
```
module load sratoolkit-3.1.1
fasterq-dump --help
fasterq-dump SRA_ID
```

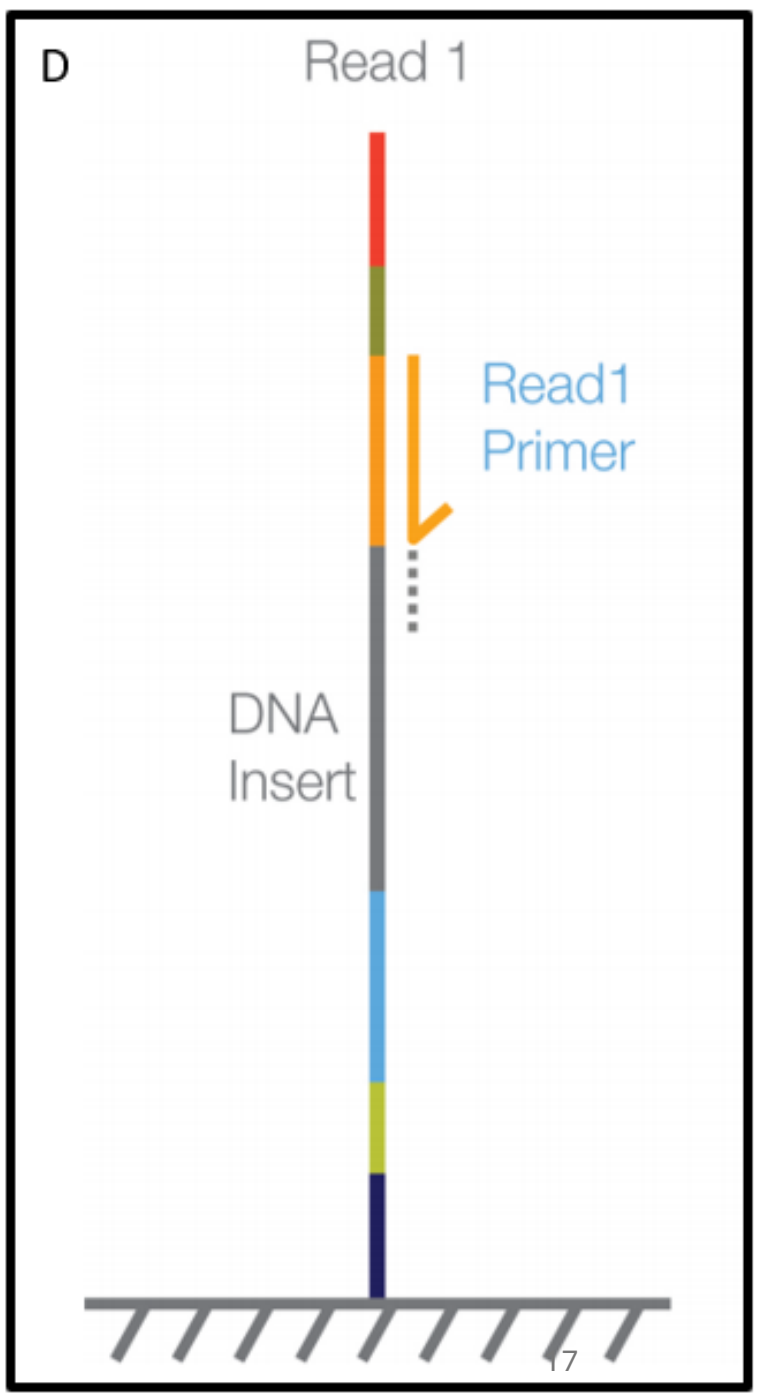
Supports multi-threaded processing and data streaming (so be careful on head node)

$ fasterq-dump

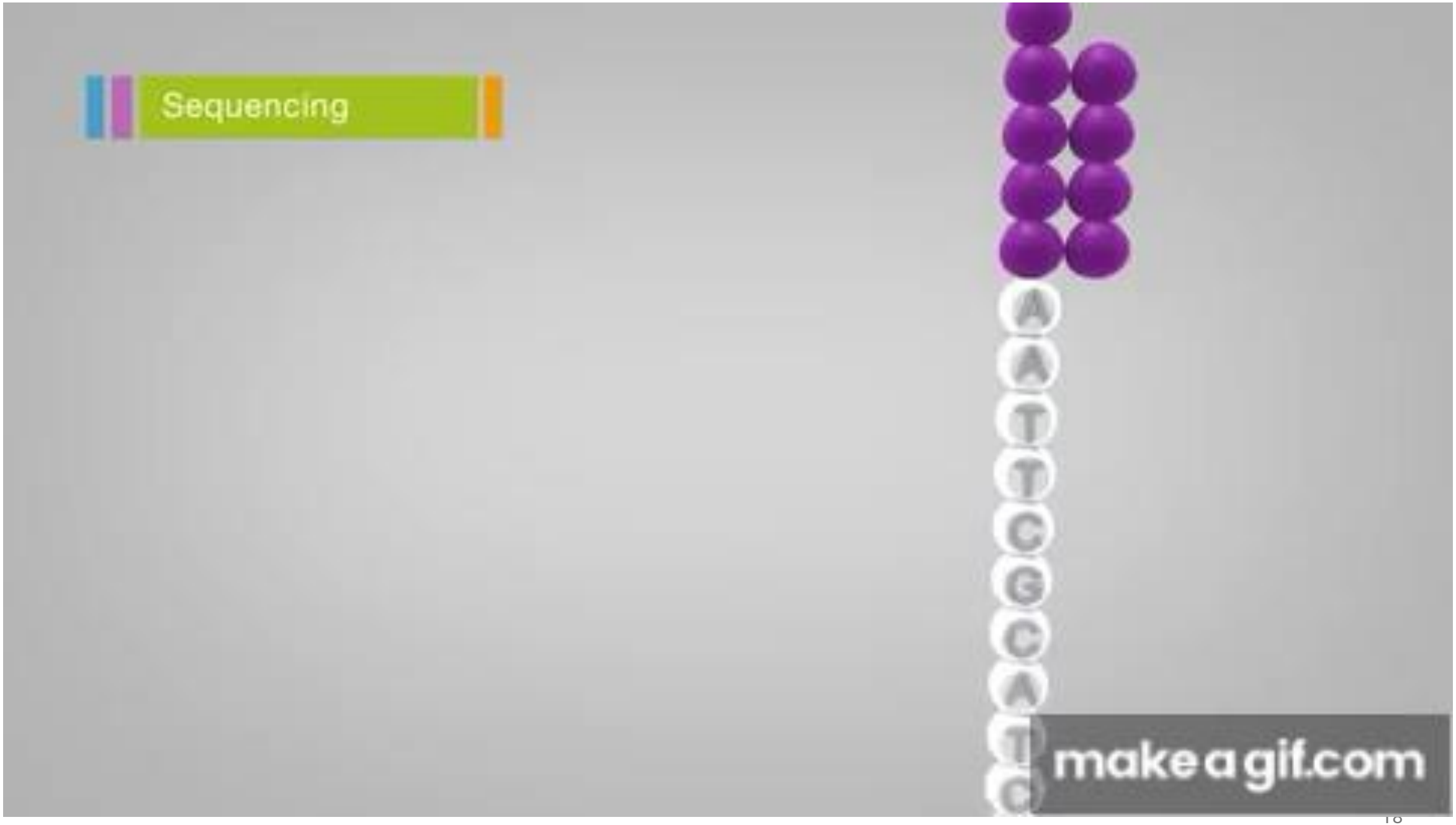
limiting the number of reads

```
fastq-dump --split-files -X 10000 SRR1553607
```

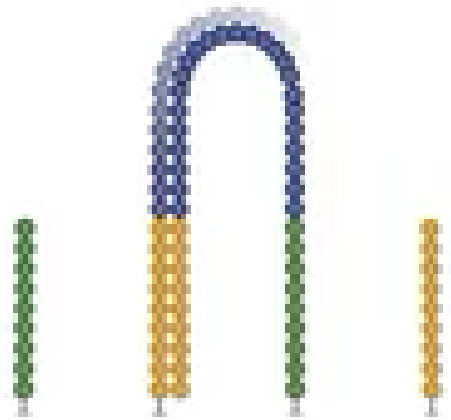

A
P5
Index 2
Rd1 SP
T
Index 1
Rd2 SP
P
P7
+
P
DNA Insert
A
A
P
B
P5
Index 2
P7
Index 1
Rd1 SP
DNA Insert
Rd2 SP
Index 1
Index 2
5'
5'
C
5'
P5
Index 2
Rd1 SP
DNA Insert
Rd2 SP
Index 1
P7
5'


D
Read 1
Read1
Primer
DNA
Insert

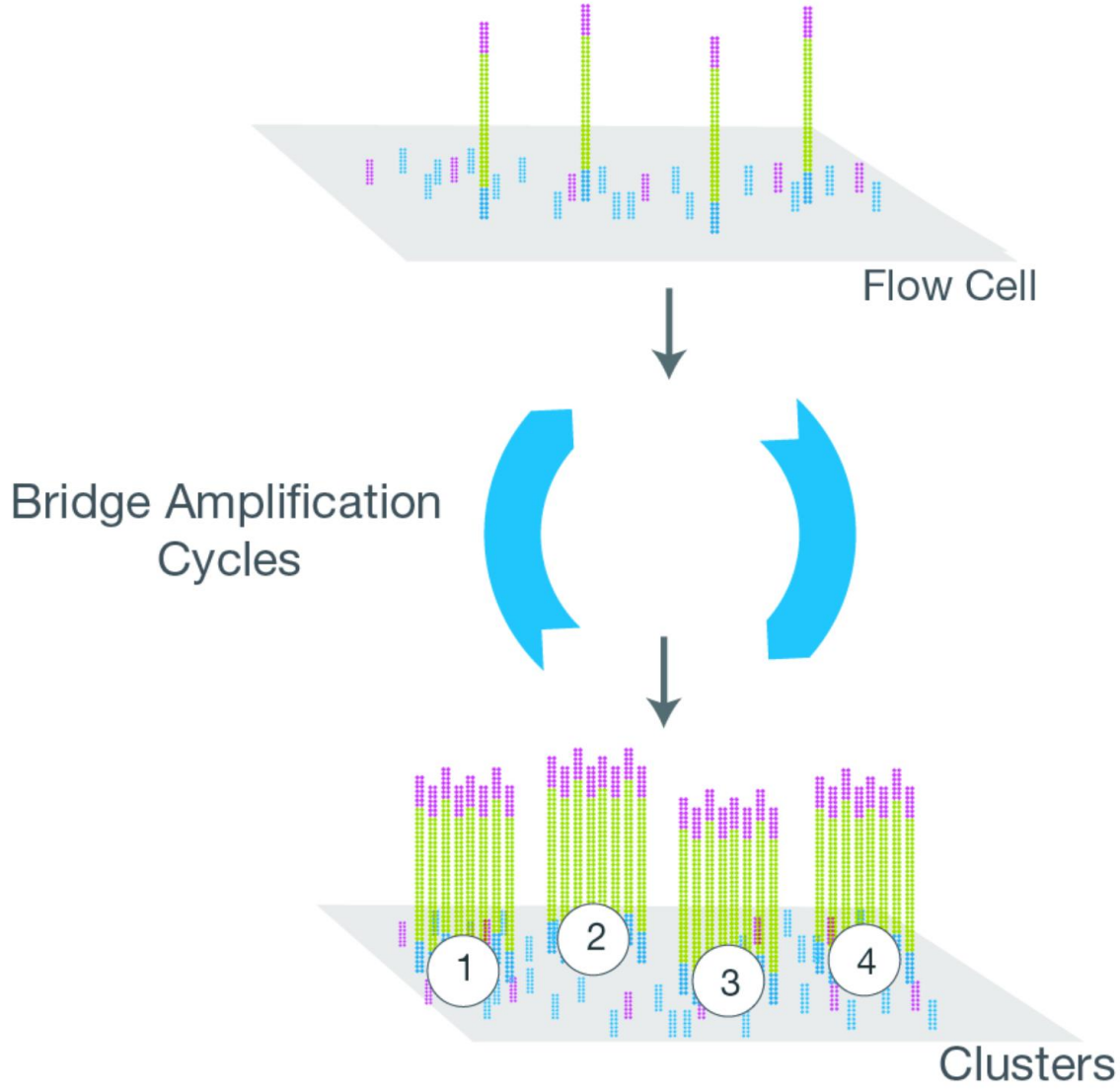
Sequencing



Cluster Generation



Cluster Amplification

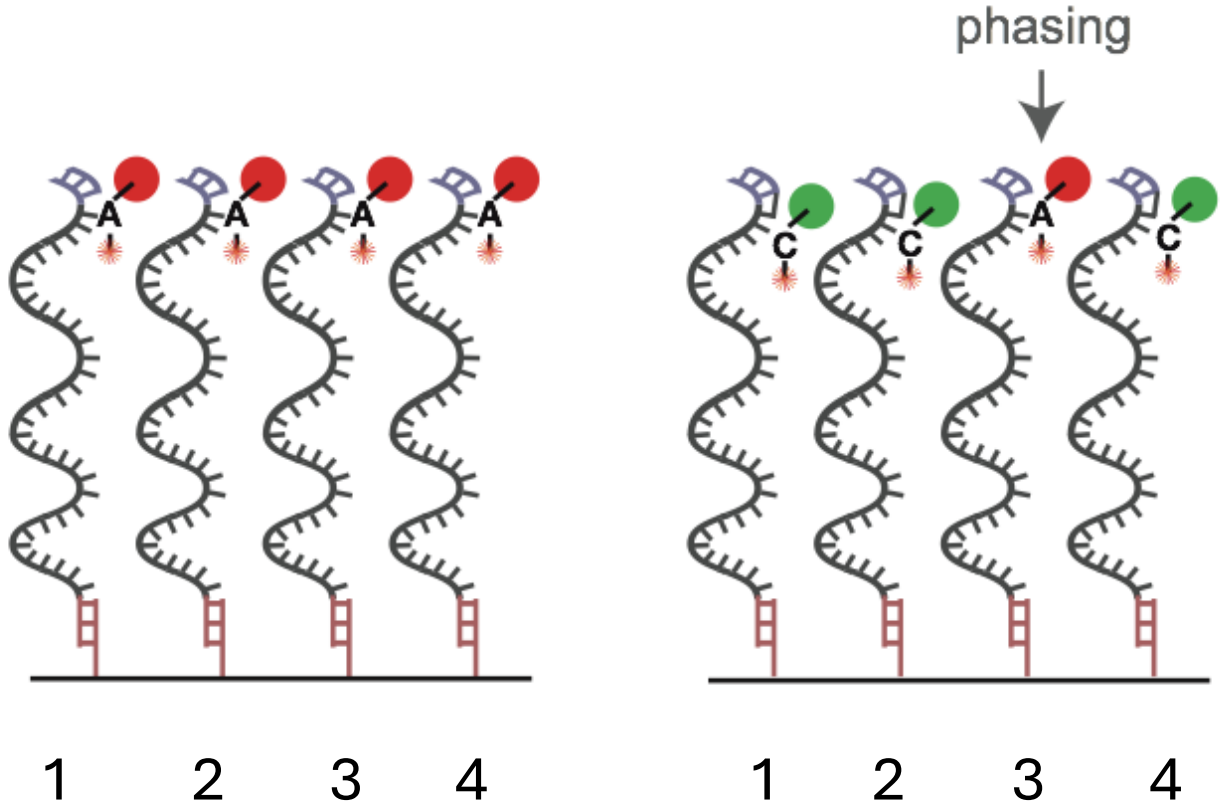


Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Phasing

Phasing means that the blocker of a nucleotide is not correctly removed after signal detection.

Molecule 3 on the right is one cycle behind the rest and will pollute the light emitted from the cluster.



FASTQs

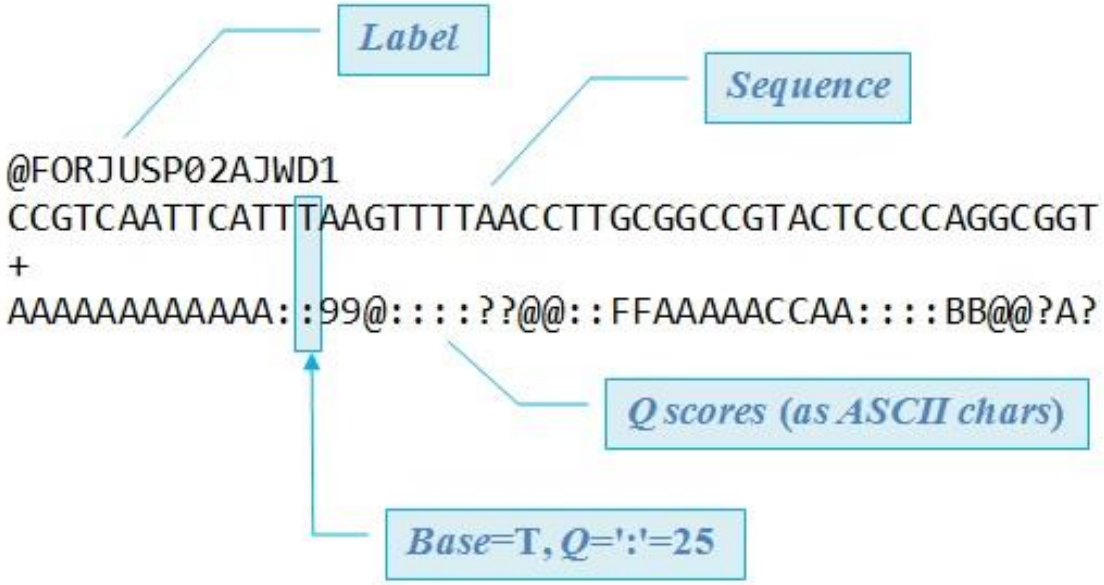
The FASTQ format is the *de facto* standard by which many sequencing instruments represent data.

Contains a sequence with an associated quality measurement for to each sequence base

FASTA with QUALITIES

Example FASTQ

comes from
Illumina sequencer




```
@GWNJ-0957:537:GW2001112798th:6:1101:5051:1379 2:N:0:GTAAGGTG+GCAATTCG
GAAAGTCTTCTTTCTTTTTTCTCTGATCTTGAACATCATTTTCAAATAAGGTTACATTATTTGAGTTAAGA
+
AAAAFF--AA-AF--7A-<-AJJFJFJ<JJJFFFF<FAFJJJFJJFJFA-<----<FF--7--<-F7JFFF
```

FASTQ quality scores

“Encoded” numerical values

each character represents a Phred score

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|    |    |    |    |    |    |    |    |
0....5...10...15...20...25...30...35...40
|    |    |    |    |    |    |    |    |
worst.............................best
```

Phred Scores

A Phred score Q is used to compute the probability of a base call being incorrect by the formula: $P=10^{(-Q/10)}$.

Q	Error	Accuracy
0	1 in 1	0%
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

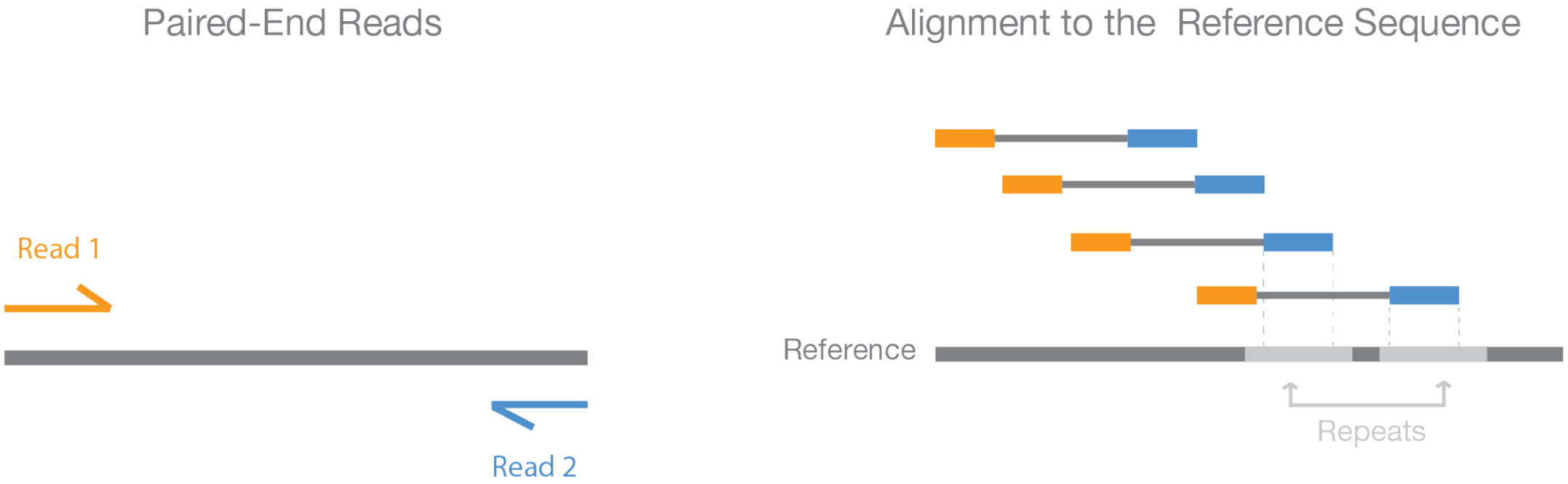
Are FASTQ error values accurate?

Not really.

“In our observation the numbers are quite unreliable - treat them as an advisory rather than accurate measurements.”

-BioStarHandbook 2020

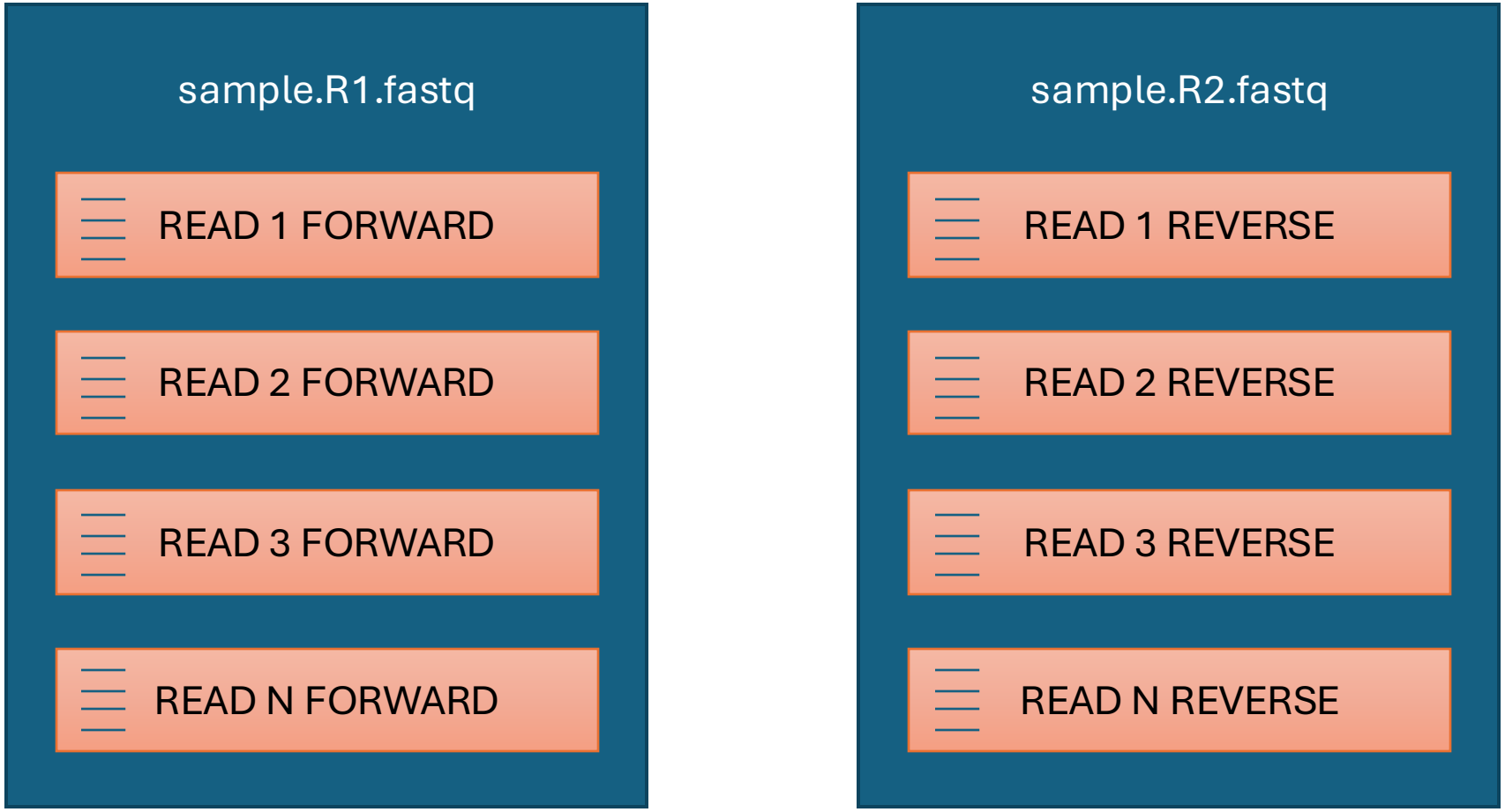
Paired-End Sequencing



Paired-end sequencing enables both ends of the DNA fragment to be sequenced.

Paired-end reads help resolve ambiguous alignments

Paired End FASTQ files



NGS QC options

1: Fastqc for data quality visualization

example fastqc report:

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

2: Fastp for data quality visualization AND trimming

- example multiqc report: https://opengene.org/fastp/fastp.html

$ fastqc

```
module load fastqc-0.11.9
fastqc --help
```

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

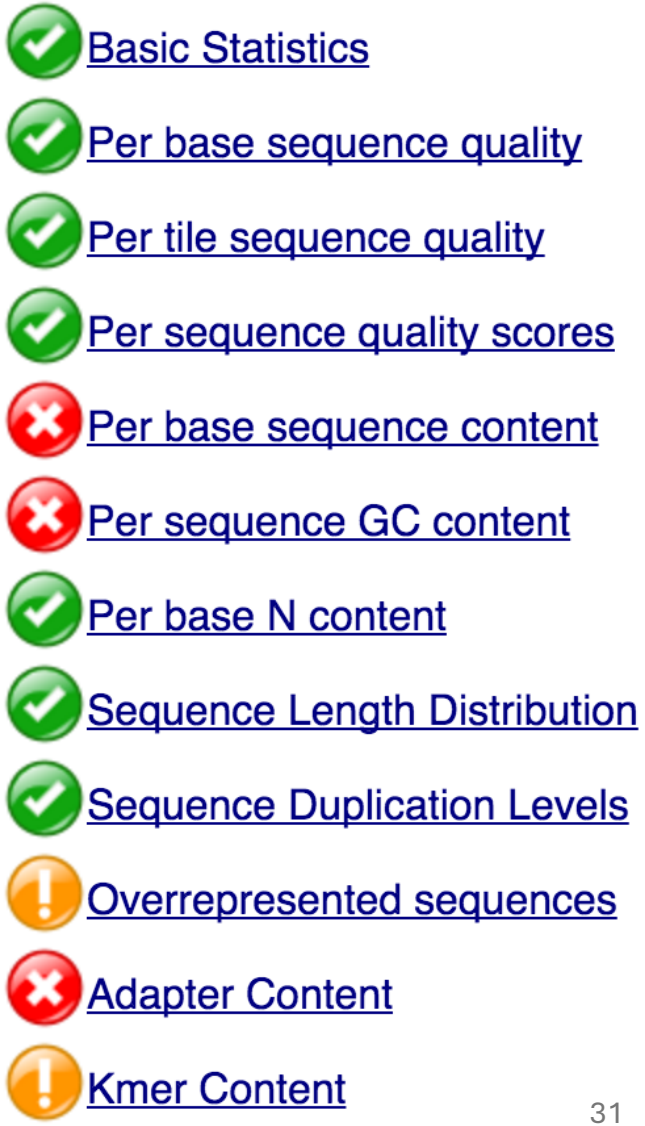
Bad illumina data:

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

Should I be worried about the “stoplight” symbols?

Usually not.

They were developed for only a particular class of samples and library preparation methods and just for certain types of instruments.



$ fastp

```
module load fastp-0.23.4
fastp --help
```

all-in-one tool for quality control and preprocessing of NGS data:

trimming, filtering, and detailed QC reports

```
$ multiqc
```

```
module load multiqc-1.25.1
multiqc --help
```

Aggregates and summarizes outputs from various bioinformatics tools (e.g. FastQC, FastP) into a single, comprehensive report.

Remove Adapters

- pseudocode:

```
cutadapt -a AGATCGGAAGAG \                  #Illumina universal adapter
-o R1.trimmed.cutadapt.fastq.gz \           #output  forward
-p R2.trimmed.cutadapt.fastq.gz \           #output reverse
in.R1.trimmed.fastq.gz \                    #input forward
in.R2.trimmed.fastq.gz                      #input reverse
```

- code example:

```
cutadapt -a AGATCGGAAGAG -o Groth-07C-
JG2_R1_001_trimmed.cutadapt.fastq.gz -p Groth-07C-
JG2_R2_001_trimmed.cutadapt.fastq.gz Groth-07C-
JG2_R1_001_trimmed.fastq.gz Groth-07C-JG2_R2_001_trimmed.fastq.gz
```

Author's note: By the way, if you ever end up writing a QC software tool, please do us all a service and find a memorable and simple name for it. Call it `speedyQC`, call it `monsterQC` call it `sevenQC`, but please, please don't make it an awkward variation of an existing, already awkward name. While we are on this topic, here are more suggestions:

- Crac: Funny And Weird Names For Bioinformatics Tools

$ srapath

...can list the location of the file whether it has already been downloaded locally or is still on the web.

srapath SRR1553607

prints:

https://sra-downloadb.be-md.ncbi.nlm.nih.gov/sos1/sra-pub-run-5/SRR1553607/SRR1553607.1

QC tools to know about

- A considerable number of QC tools have been published.

Others in alphabetical order:

- BBDuk part of the BBMap package
- BioPieces a suite of programs for sequence preprocessing
- CutAdapt application note in Embnet Journal, 2011
- Fastp published in Bioinformatics 2018
- fastq-mcf published in The Open Bioinformatics Journal, 2013
- Fastx Toolkit: collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing - one of the first tools
- FlexBar, Flexible barcode and adapter removal published in Biology, 2012
- NGS Toolkit published in Plos One, 2012
- PrinSeq application note in Bioinformatics, 2011
- Scythe a bayesian adaptor trimmer
- SeqPrep - a tool for stripping adaptors and/or merging paired reads with overlap into single reads.
- Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads.
- TagCleaner published in BMC Bioinformatics, 2010
- TagDust published in Bioinformatics, 2009
- Trim Galore - a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries
- Trimmomatic application note in Nucleic Acid Research, 2012, web server issue

$ srapath

but after performing a:

prefetch SRR1553607

the same srapath SRR1553607 might print:

/nas4/dbobo/ncbi/public/sra/SRR1553607.sra

GNU screen

$ screen allows users to run multiple terminal sessions simultaneously, even if they disconnect from the system.

cheat sheet:
https://kapeli.com/cheat_sheets/screen.docset/Contents/Resources/Documents/index

BioStar Handbook

https://www.biostarhandbook.com/

if RGGS doesn’t have a subscription, I’ll share my credentials

AMNH Bioinformatics Sharepoint Site

https://amnh.sharepoint.com/sites/Bioinformatics

e.g. PBS scripting on Huxley:

https://amnh.sharepoint.com/sites/Bioinformatics/SitePages/PBS.aspx

In trouble?
E-mail the bioinformatics core

bioinformatics@amnh.org

after troubleshooting and
suffering on your own, of course :-P

We need several things when you contact us

1) command being run
2) error message
3) location of any scripts, software, or virtual environments
4) location of data or working directory

Bioinformatics Working Group

meets every other Wednesday at 11am
in bioinformatics suite.
(but we also open zoom).

Working Group: we brainstorm and troubleshoot together.

i.e. very informal discussions are welcome!

Exercise

search SRA for data of interest

download data from SRA

inspect quality with fastqc

QC data (i.e. trim_galore, trimmomatic, fastp)

re-inspect

checklist for one sample

1. find relevant data on SRA
2. make a list of accession
3. ssh into Huxley
4. start screen session
5. create directory structure for project data
6. module load necessary software
   sratoolkit, fastqc, etc. module avail to search for software
7. prefetch data
8. use PBS job to fasterq-dump data
9. use PBS job to generate fastqc report
10. download HTML report to local machine and inspect
    use scp or rsync
11. trim as needed and repeat steps 8 - 10.

Where to go from here?

De novo assembly and annotation (phylogenomics)

Referenced-based approaches (population genetics)

Appendix

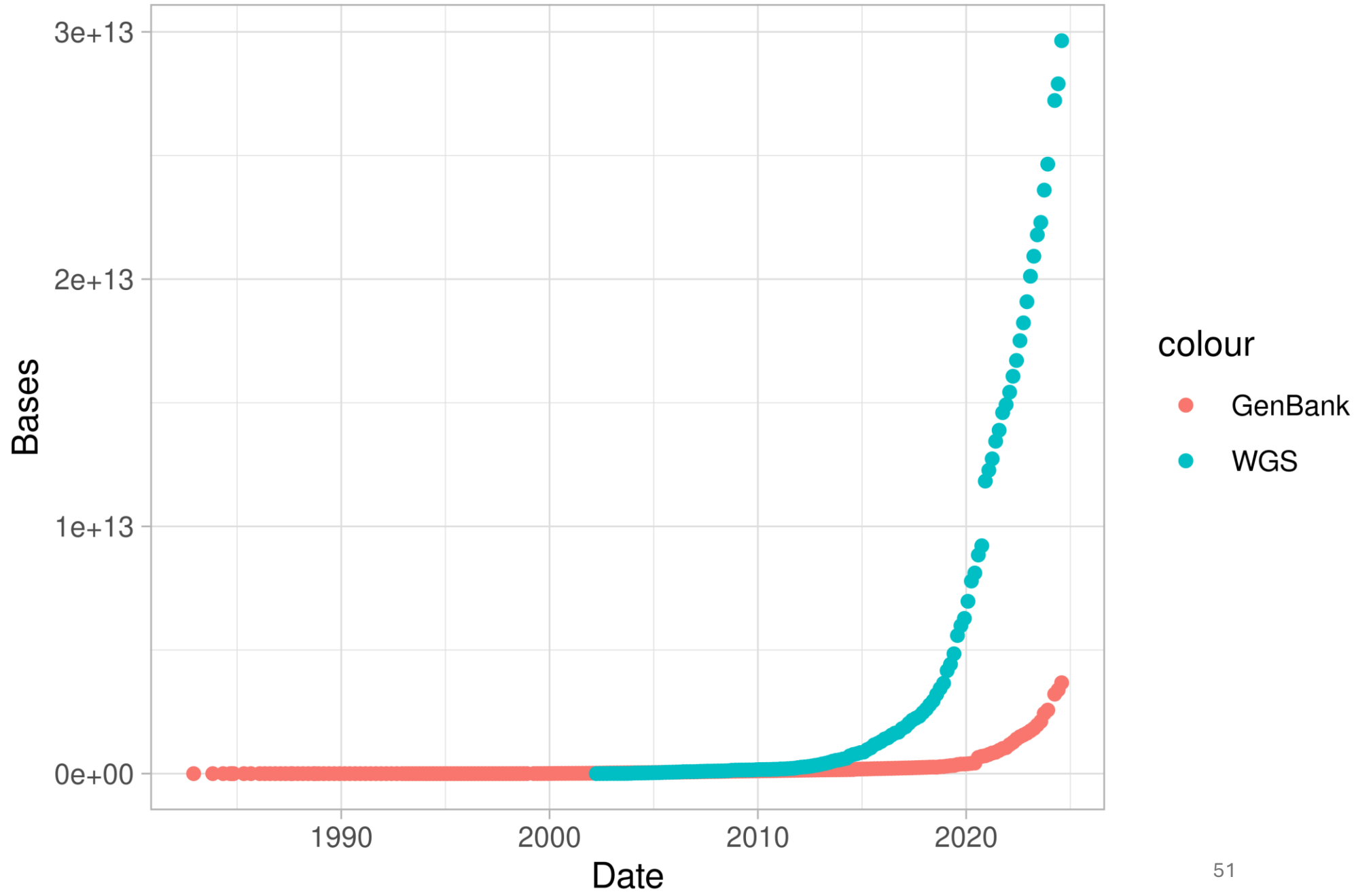
NGS Tutorial (Reference Based)

https://github.com/deanbobo/amnh-ngs-workshop

Micromamba

```
export software=/nas4/$USER/software

mkdir -p $software

curl -Ls https://micro.mamba.pm/api/micromamba/linux-64/latest | tar -xvj -C $software bin/micromamba

export PATH=$software/bin:$PATH

micromamba shell init -s bash -r $software/mamba
```

GenBank growth over time



GenBank growth over time