

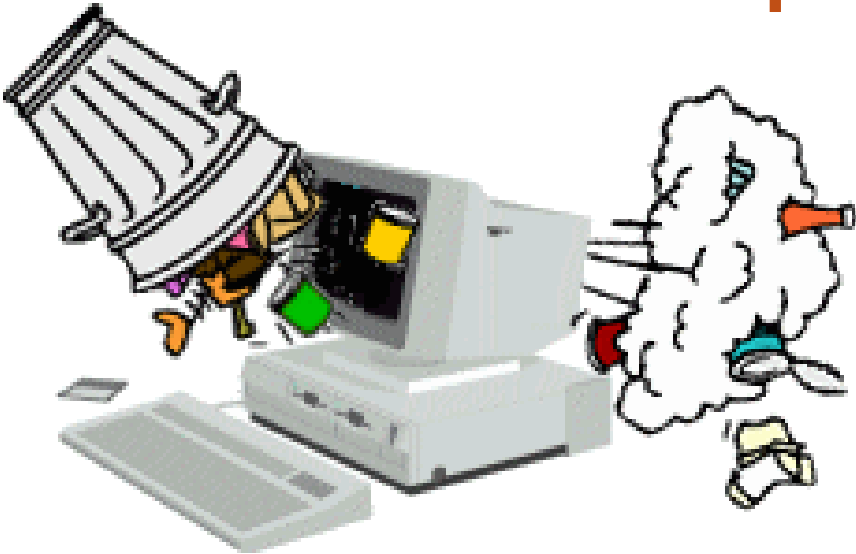
Data input and Quality Control

Bioplatforms Fungi Genomics Workshop 2024

Australian National University

Benjamin Schwessinger

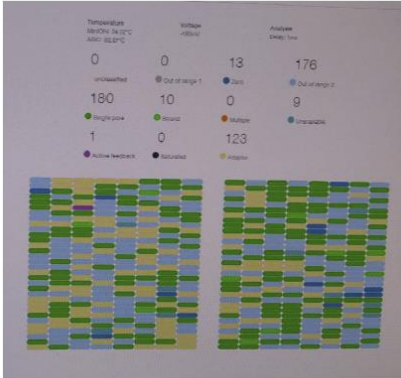
Data input and Quality Control



Garbage In



Garbage Out



<https://twitter.com/AaronPomerantz/status/826448809962020864>

✓ What is my research question?

✓ Genome evolution?

✓ Complete genome reconstruction?

✓ Care about TEs?



✓ What is already available in the public domain?

✓ How much data is enough?

✓ Gene discovery?

✓ Pop-gen?

✓ Phylogenomics?

✓ Else?

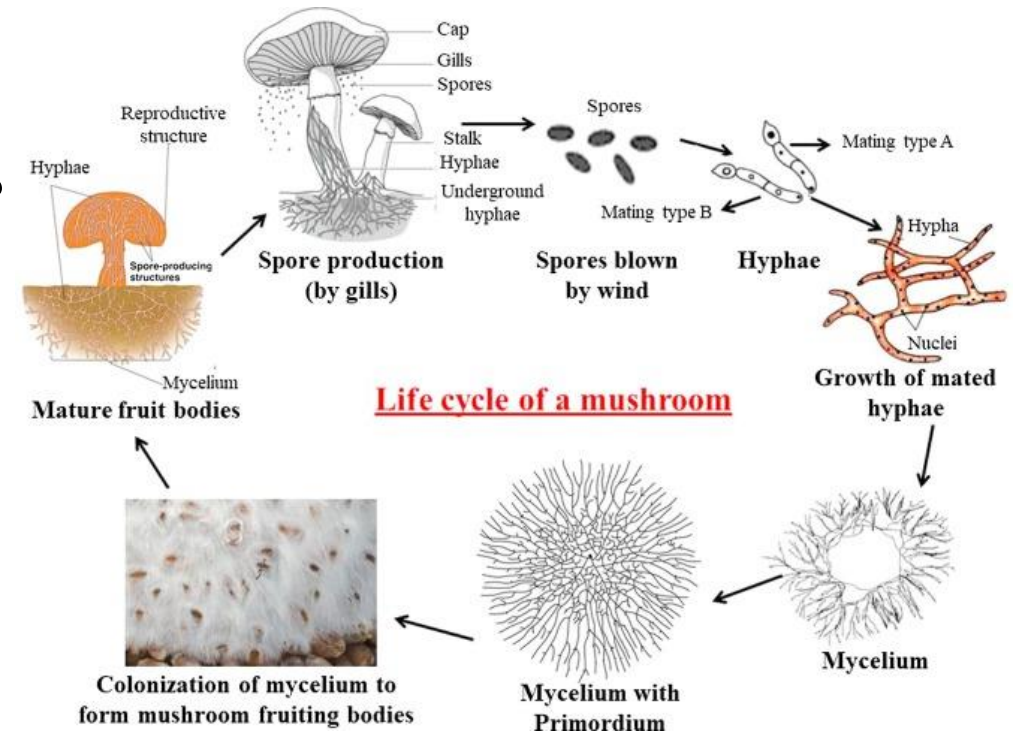
✓ Phasing of di- and ploykayrons?

✓ Species identification?

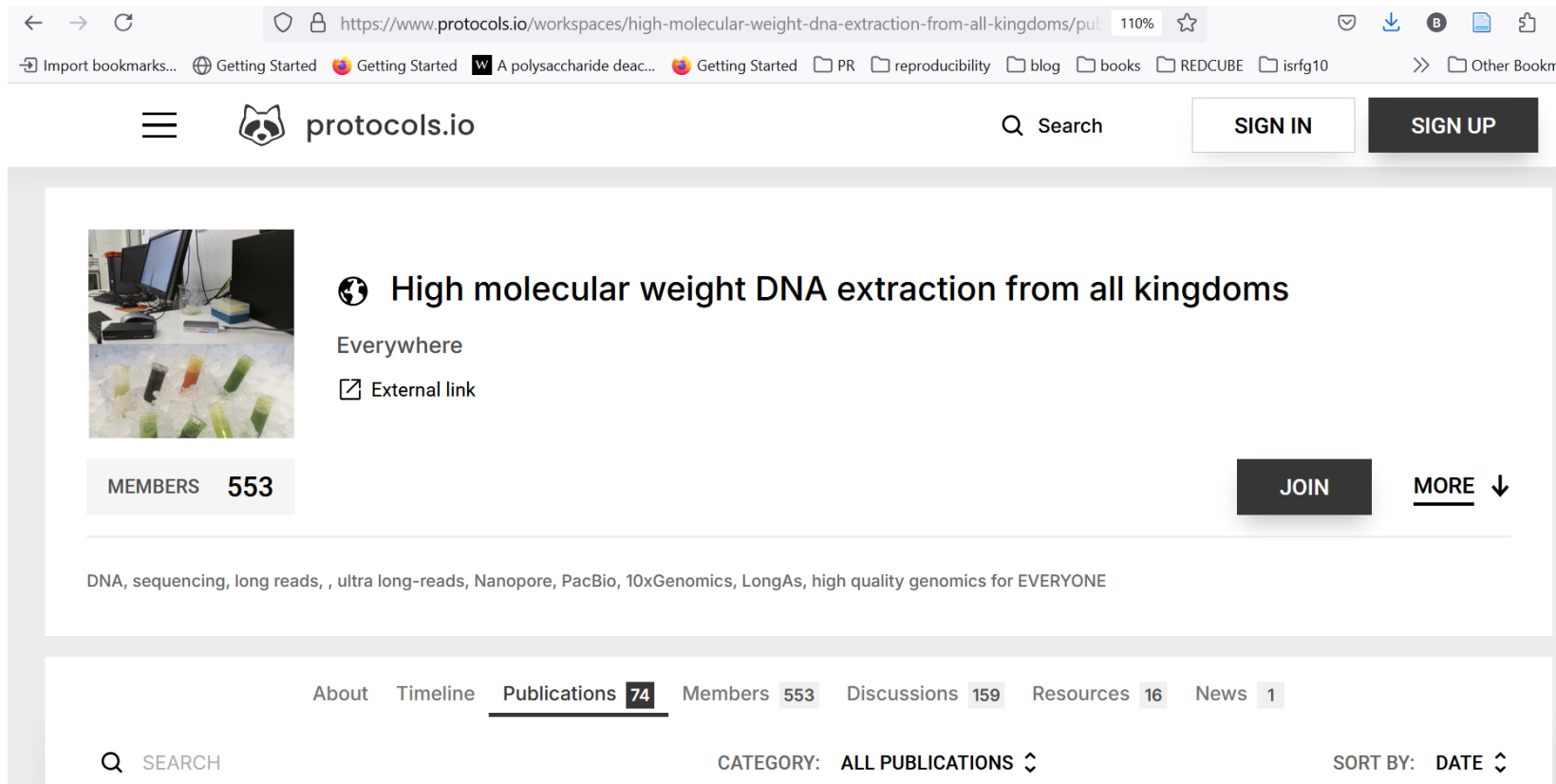
“What input material is (easily) available to me?”



- ✓ Type?
- ✓ Culturable?
- ✓ Ploidy and nuclear state?
- ✓ Purity?
- ✓ Reproducibility?



“How do I (bloody) get some clean high-quality high-molecular DNA out of fungi?”



The screenshot shows a web browser displaying a Protocols.io workspace page. The browser's address bar shows the URL: <https://www.protocols.io/workspaces/high-molecular-weight-dna-extraction-from-all-kingdoms/publications>. The page features a navigation bar with a search icon, a search input field, and buttons for 'SIGN IN' and 'SIGN UP'. Below the navigation bar, there is a workspace header for 'High molecular weight DNA extraction from all kingdoms', which is marked as 'Everywhere' and an 'External link'. A 'MEMBERS 553' badge is visible, along with 'JOIN' and 'MORE' buttons. The main content area displays a list of publications, with a navigation bar at the bottom showing 'About', 'Timeline', 'Publications 74', 'Members 553', 'Discussions 159', 'Resources 16', and 'News 1'. A search bar and filters for 'CATEGORY: ALL PUBLICATIONS' and 'SORT BY: DATE' are also present.

<https://www.protocols.io/workspaces/high-molecular-weight-dna-extraction-from-all-kingdoms/publications>

[20221129_AshJMorning.pdf](#)

A wryly wind simplified intro to DNA sequencing techs

Illumina

- Relatively forgiving (length and quality)
- Little DNA input
- High-quality short-reads
- Good for haploid gene space analysis
- Good for pop-gen including reduced representation like DArT
- Used for many different applications
- Cheap at scale

<https://www.youtube.com/watch?v=fCd6B5HRaZ8>

PacBio

- Not forgiving (length and quality)
- Large DNA input
- High-quality mid-range reads (up to 20kb)
- Build for human genomes
- Requires multiplexing to be cost effective which risks failing of all samples on the run

https://www.youtube.com/watch?v=_ID8JyAbwEo

Nanopore

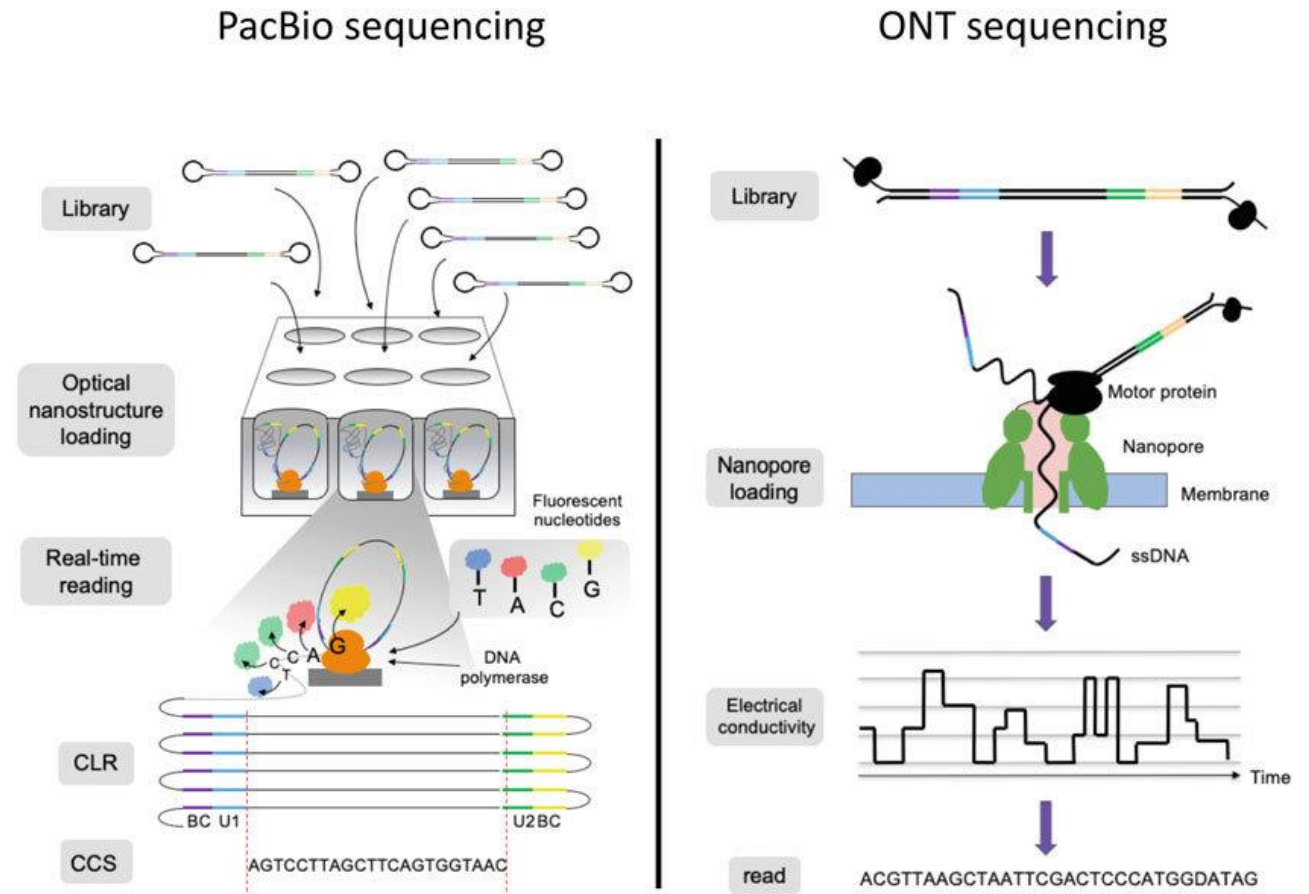
- Sometimes forgiving (length and quality)
- medium DNA input
- High-quality reads of any size. Super high-quality reads of >10kb
- Very (too) flexible platform
- Good for microbial genomes including fungi
- Different scales available
- Great for lots of applications

<https://www.youtube.com/watch?v=sv9fFeSd3kE>

Output data types

Signal level data

- PacBio - *.bam
- Nanopore - *.pod5
- Important of DNA modification analysis otherwise not
- Allows reanalysis of data with different basecallers



Output data types

Per base data and per base quality

- Fasta < sequence and name only
- Fastq < sequence and quality value estimates
- The most important what most people ever need.
- These are plain text file and Windows might do weird stuff to them

Format of a FASTA definition line

>Seq1 [organism=Carpodacus mexicanus] [clone=6b] actin (act) mRNA, partial CDS
CCTTTATCTAATCTTTGGAGCAYGAGCTGGCATAGTTGGAACCGCCCTCAGCCTCCTCATC

The diagram illustrates the format of a FASTA definition line. It shows a sequence identifier followed by a description in square brackets, separated by spaces. The first space is highlighted with a callout box labeled "space". The second and third spaces are also highlighted with callout boxes labeled "space". The description ends with a hard return character, highlighted with a callout box labeled "hard return".

Output data types

FASTQ file sample:

```
@SRR6407486.1 1 length=100
CCTCGTCTACAGCGACAACGTCCAGACCCGCGAACGGGTGATGCGGGCCCTGGGCAAACGGTTCACCCGGATCTGCCCGATTTGACCTACGTCGAAGTG
+SRR6407486.1 1 length=100
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFF7FFFF<FF
```

@SRR6407486.1 1 length=100

CCTCGTCTACAGCGACAAC ... GATTTGACCTACGTCGAAGTG

+SRR6407486.1 1 length=100

BBBBBFFFFFFFFFFFFFFFF ... FFFFFFFFFFFFFFFFF7FFFF<FF

Sequence name

DNA sequence

Quality line break

Quality scores

Base: T
Quality: 7

Quality scores as ASCII characters:

!"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJK

Q:	0	5	15	30	40
P _{error} :	1.0	0.32	0.032	0.001	0.0001

$$Q = -10 \log_{10} P_{\text{error}}$$

Read Quality

Quality score	Base calling error probability	Base calling accuracy
10	10^{-1}	90%
20	10^{-2}	99%
30	10^{-3}	99.9%
40	10^{-4}	99.99%

Read length

Illumina

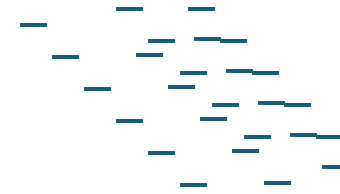
- Fixed based on sequencer and sequencing approach
- Most often paired end with 100, 150, 300bp each

PacBio

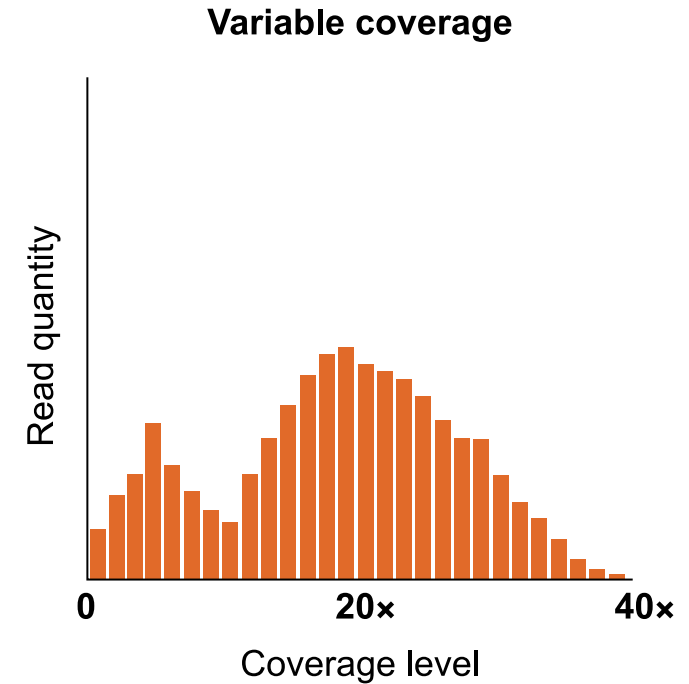
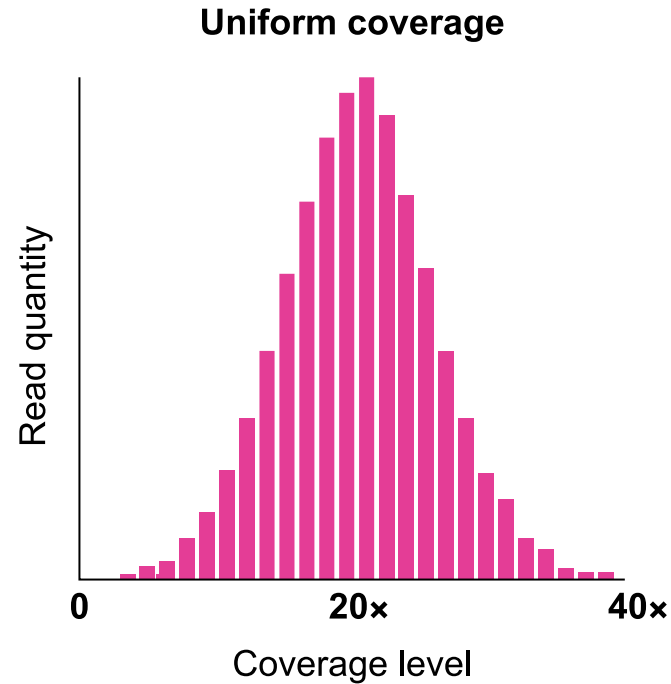
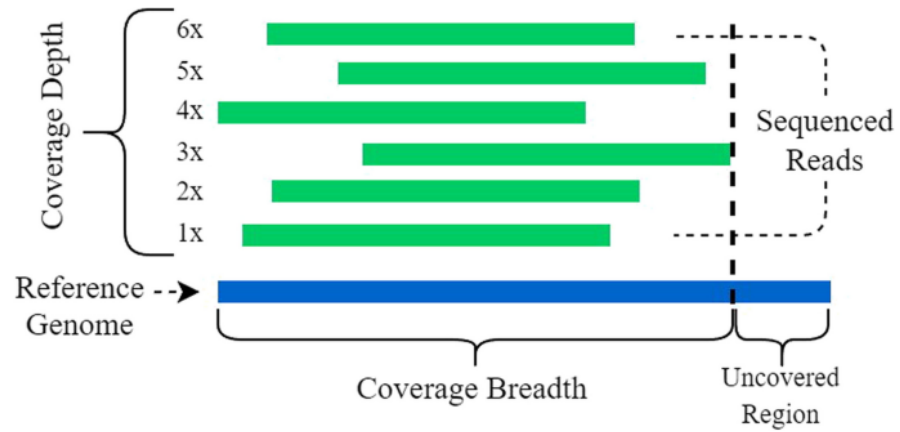
- Variable between runs but mostly fixed within run
- Current sweet spot 15-20kb per sequencing well

Nanopore

- Read length agnostic (bp to Mbp)
- Works best when DNA length in same range (order of magnitude)
- Great for amplicons
- For genome assembly should be > 10kb



Coverage/Sequencing depth



15-20x coverage per haplotype is a good start

Data purity



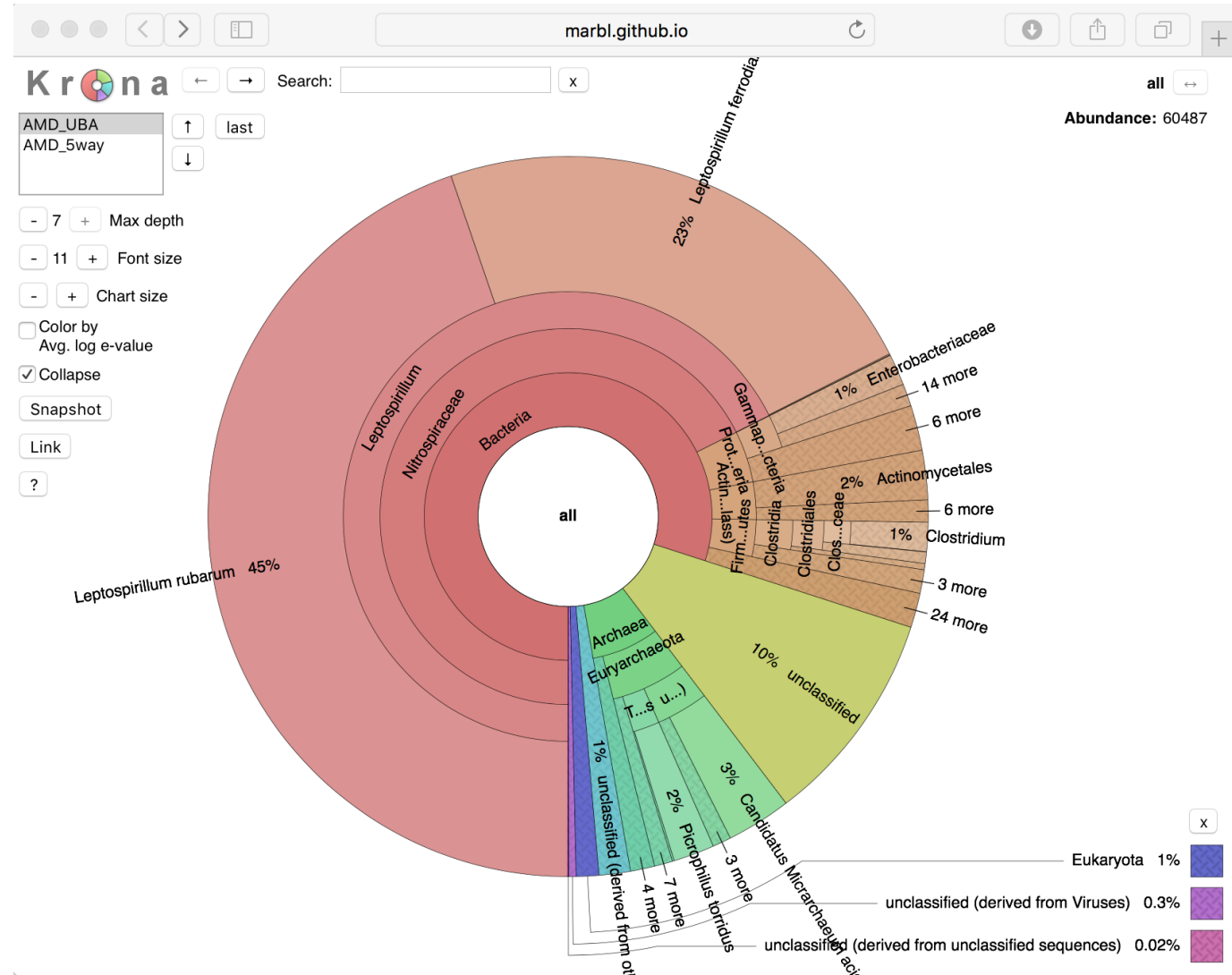
Data purity

K-mer based

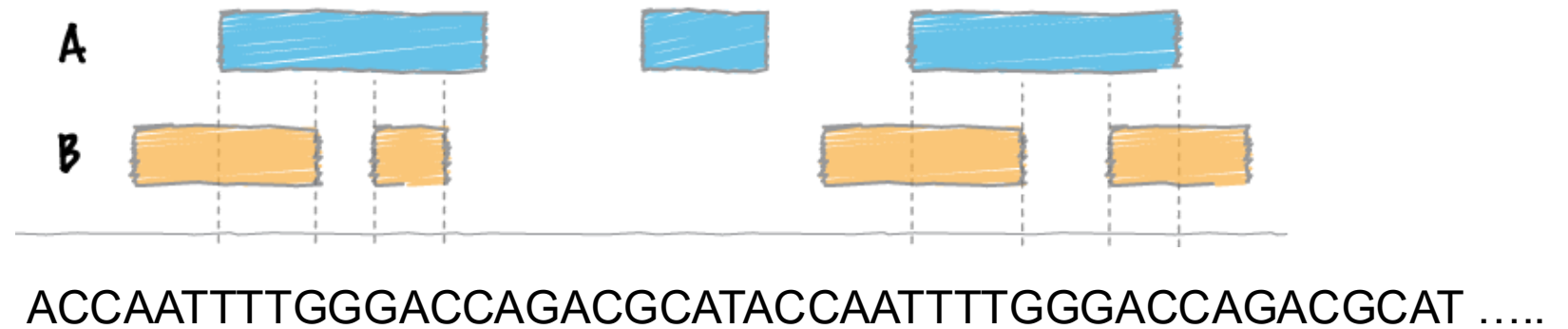
- Kraken2, Bracken
- Krona and Pavian

Alignment based

- Blast, reference mapping



File formats > a coordinate based system



A1	6	14
B1	3	9