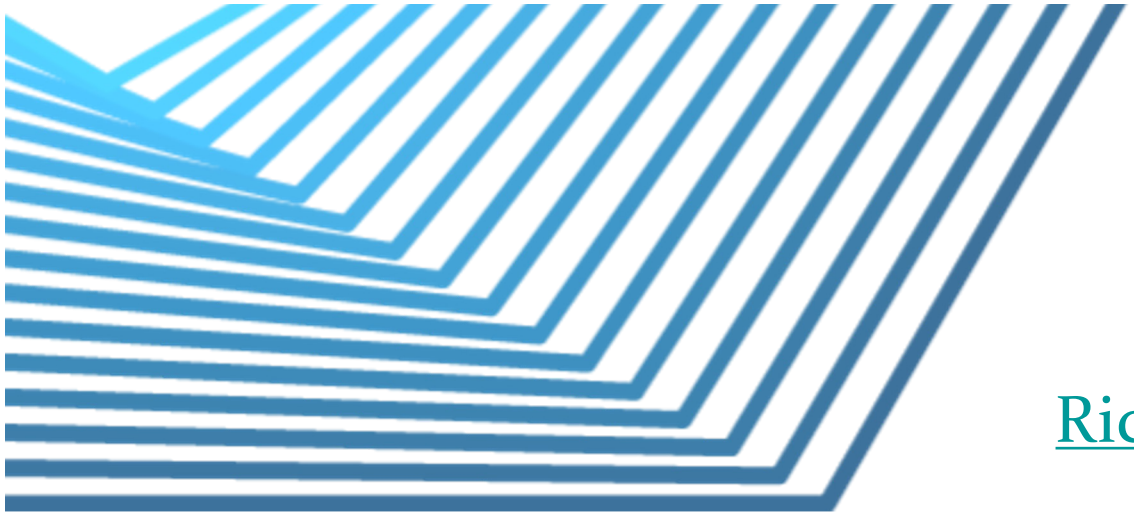
CVR
Medical Research Council
University of Glasgow
Centre for Virus Research

Metagenomics



Richard Orton

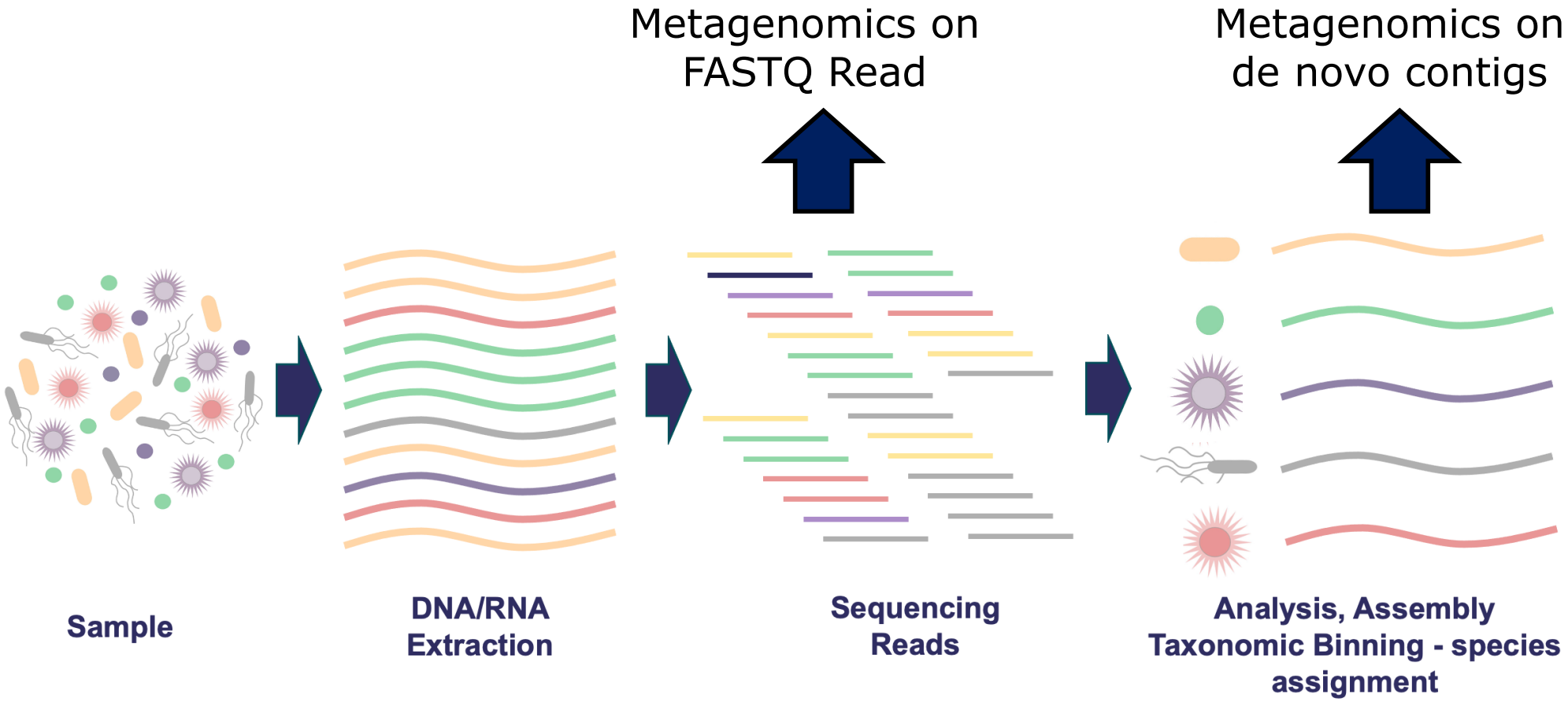
Richard.Orton@glasgow.ac.uk

MRC Medical Research Council

University of Glasgow

Overview

Medical Research Council
University of Glasgow
Centre for Virus Research



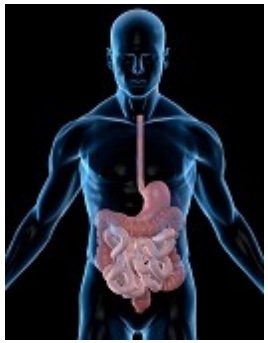
https://www.kennedy.ox.ac.uk/ocms/platforms/genomics

Metagenomics

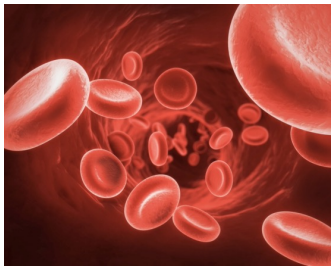
Medical Research Council
University of Glasgow
Centre for Virus Research

- **Metagenomics** can be defined as the sequenced-based analysis of the whole collection of genomes isolated directly from a sample
- The **advantage** is that isolation is not needed – only extraction and sequencing (although there's more to it than that!) – and enrichment can help

Samples

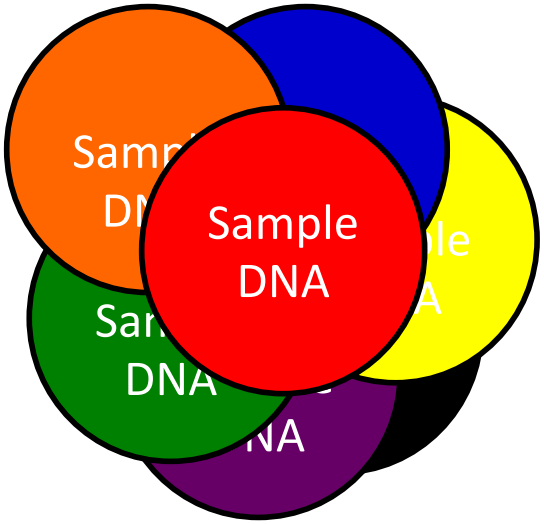



1cm
2cm



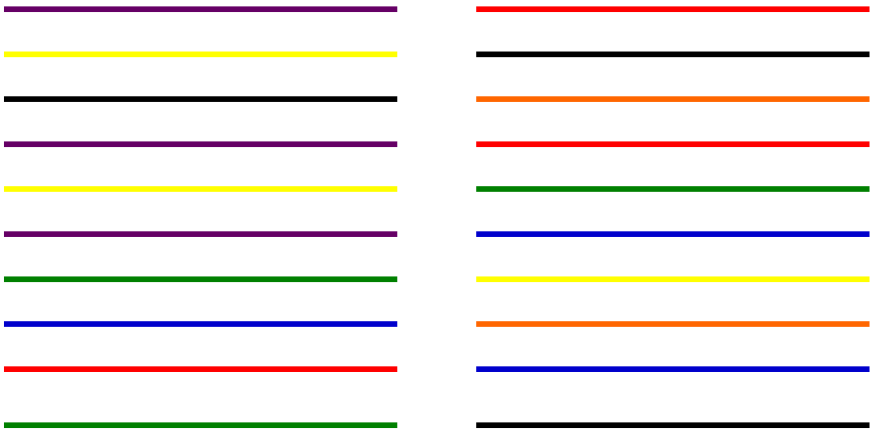

Accident & Emergency
Primary Care Out of Hours Centre

Metagenomics methods



RNA/DNA extraction

Sequencing quality control



De novo assembly

BLAST

Species A

Species B

Species C

Species D

Species E

Species F

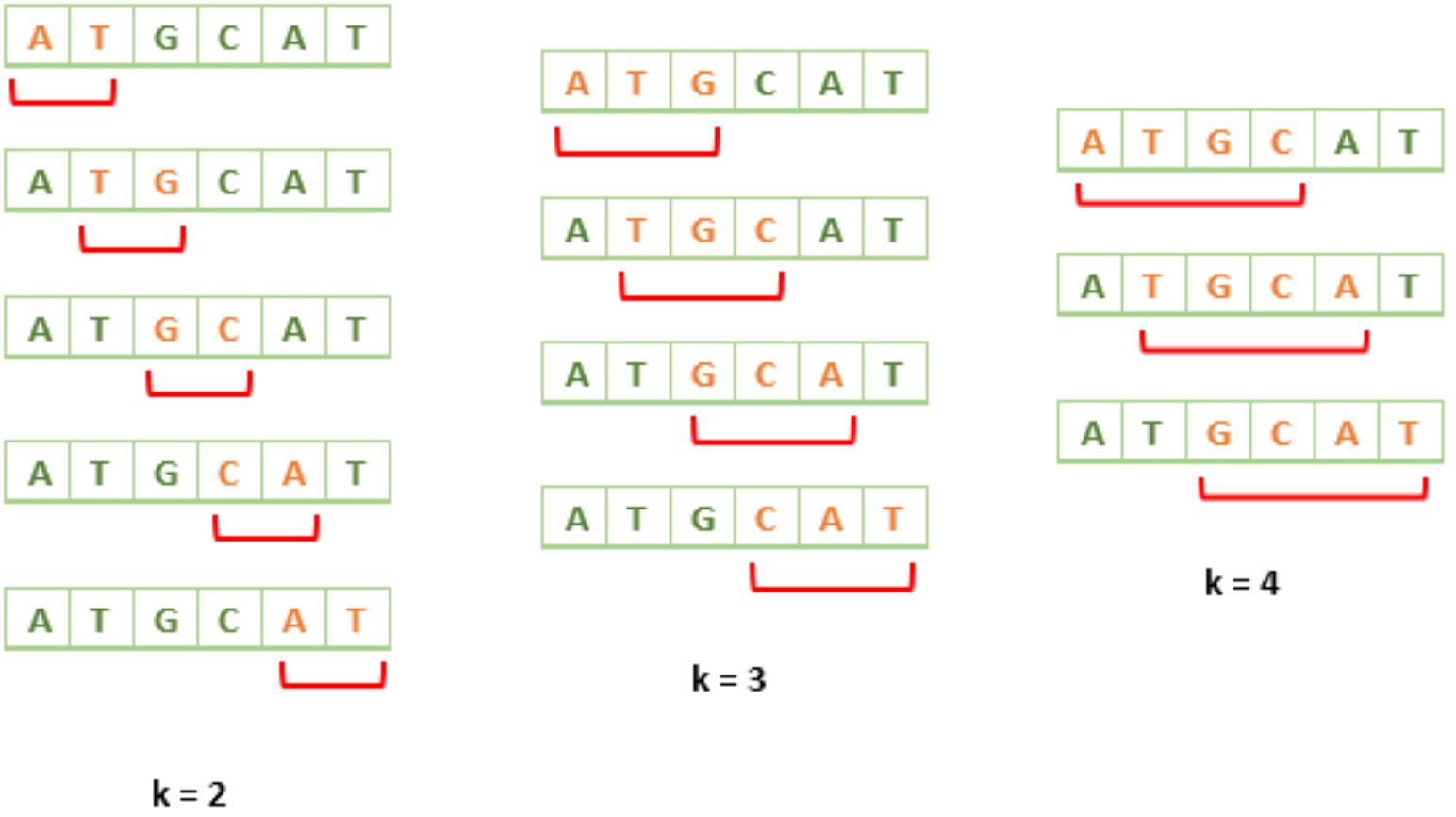
Species G

CVR
Medical Research Council
University of Glasgow
Centre for Virus Research

Kmer metagenomics (fast)

What is a k-mer?

- Short word(s) generated from short reads
- Sub-strings of the length k



http://zerocool.is-a-geek.net/wp-content/uploads/2012/10/counting_kmers.png

What is a k-mer?

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

Reverse complement and redo as well!

Unique kmers

- Viral taxonomy
  - Order > Family > Genus > Species

- Kmers of length 31
  - Some are only seen in specific virus species
  - Some are common to a genus
  - Some are common to a family
  - Some are common to many families

- What virus is this 31-kmer from?
  - AGTTGTTAGTCTACGTGGACCGACAAGAACA

Unique kmers in Books (130 million)

Mr. and Mrs. Dursley, of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much.

It is a truth universally acknowledged that a single man in possession of a good fortune must be in want of a wife

It was the best of times, it was the worst of times.

When Mr. Bilbo Baggins of Bag End announced that he would shortly be celebrating his eleventyfirst birthday with a party of special magnificence, there was much talk and excitement in Hobbiton

To be, or not to be: that is the question

And may the odds be ever in your favor

Get busy living or get busy dying

A mouse took a stroll through the deep dark wood.

Unique kmers throughout taxonomy

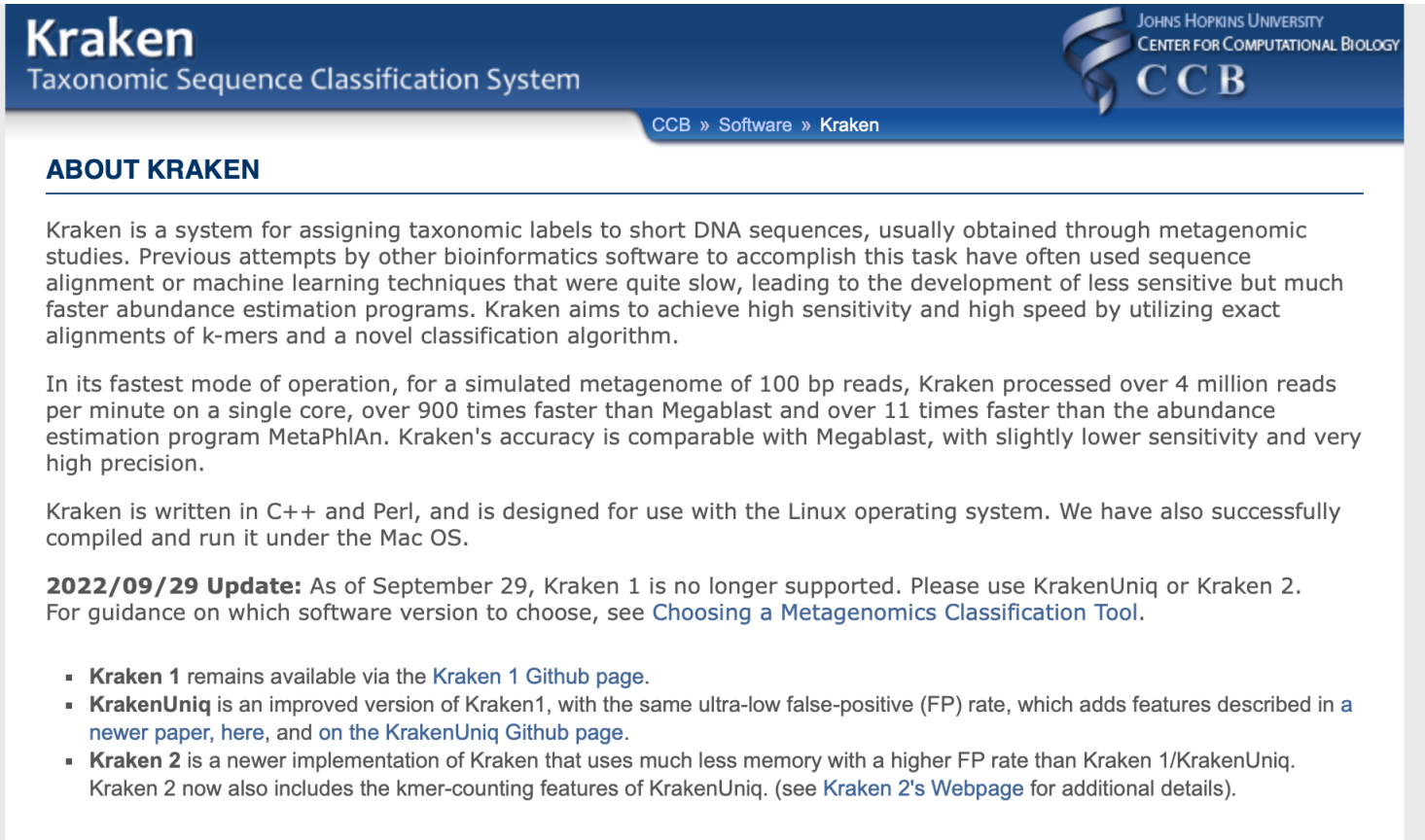
- FASTQ Read/Pair or a Contig
  - Split the sequence up into kmers of length 31

- **Compare to a database of known kmers**
  - k-mers unique to a virus species
  - k-mers unique to a virus genus
  - k-mers unique to a virus family
  - Non-specific k-mers e.g. AAAAAAAAAAAAAAAAAAA

- Same for bacteria, eukaroytes etc

Kraken

- Kraken is a system for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomic studies.
- **Analyse FASTQ files rapidly and report what is in there**



Kraken
Taxonomic Sequence Classification System

Johns Hopkins University
Center for Computational Biology
CCB

CCB » Software » Kraken

ABOUT KRAKEN

Kraken is a system for assigning taxonomic labels to short DNA sequences, usually obtained through metagenomic studies. Previous attempts by other bioinformatics software to accomplish this task have often used sequence alignment or machine learning techniques that were quite slow, leading to the development of less sensitive but much faster abundance estimation programs. Kraken aims to achieve high sensitivity and high speed by utilizing exact alignments of k-mers and a novel classification algorithm.

In its fastest mode of operation, for a simulated metagenome of 100 bp reads, Kraken processed over 4 million reads per minute on a single core, over 900 times faster than Megablast and over 11 times faster than the abundance estimation program MetaPhlAn. Kraken's accuracy is comparable with Megablast, with slightly lower sensitivity and very high precision.

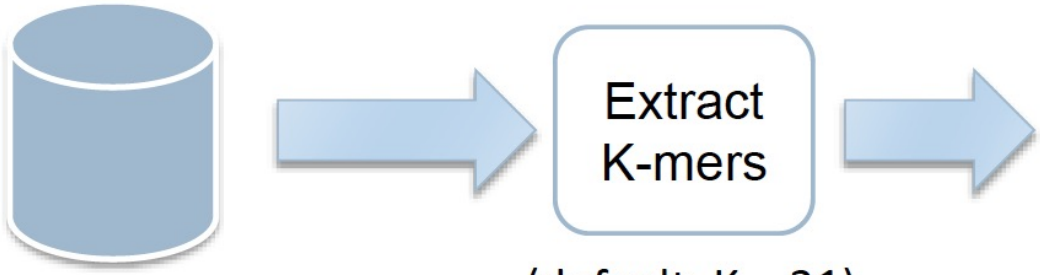
Kraken is written in C++ and Perl, and is designed for use with the Linux operating system. We have also successfully compiled and run it under the Mac OS.

2022/09/29 Update: As of September 29, Kraken 1 is no longer supported. Please use KrakenUniq or Kraken 2. For guidance on which software version to choose, see Choosing a Metagenomics Classification Tool.

- **Kraken 1** remains available via the Kraken 1 Github page.
- **KrakenUniq** is an improved version of Kraken1, with the same ultra-low false-positive (FP) rate, which adds features described in a newer paper, here, and on the KrakenUniq Github page.
- **Kraken 2** is a newer implementation of Kraken that uses much less memory with a higher FP rate than Kraken 1/KrakenUniq. Kraken 2 now also includes the kmer-counting features of KrakenUniq. (see Kraken 2's Webpage for additional details).

Kraken: K-mer LCA database

Reference Genomes
(2,256 RefSeq Genomes)

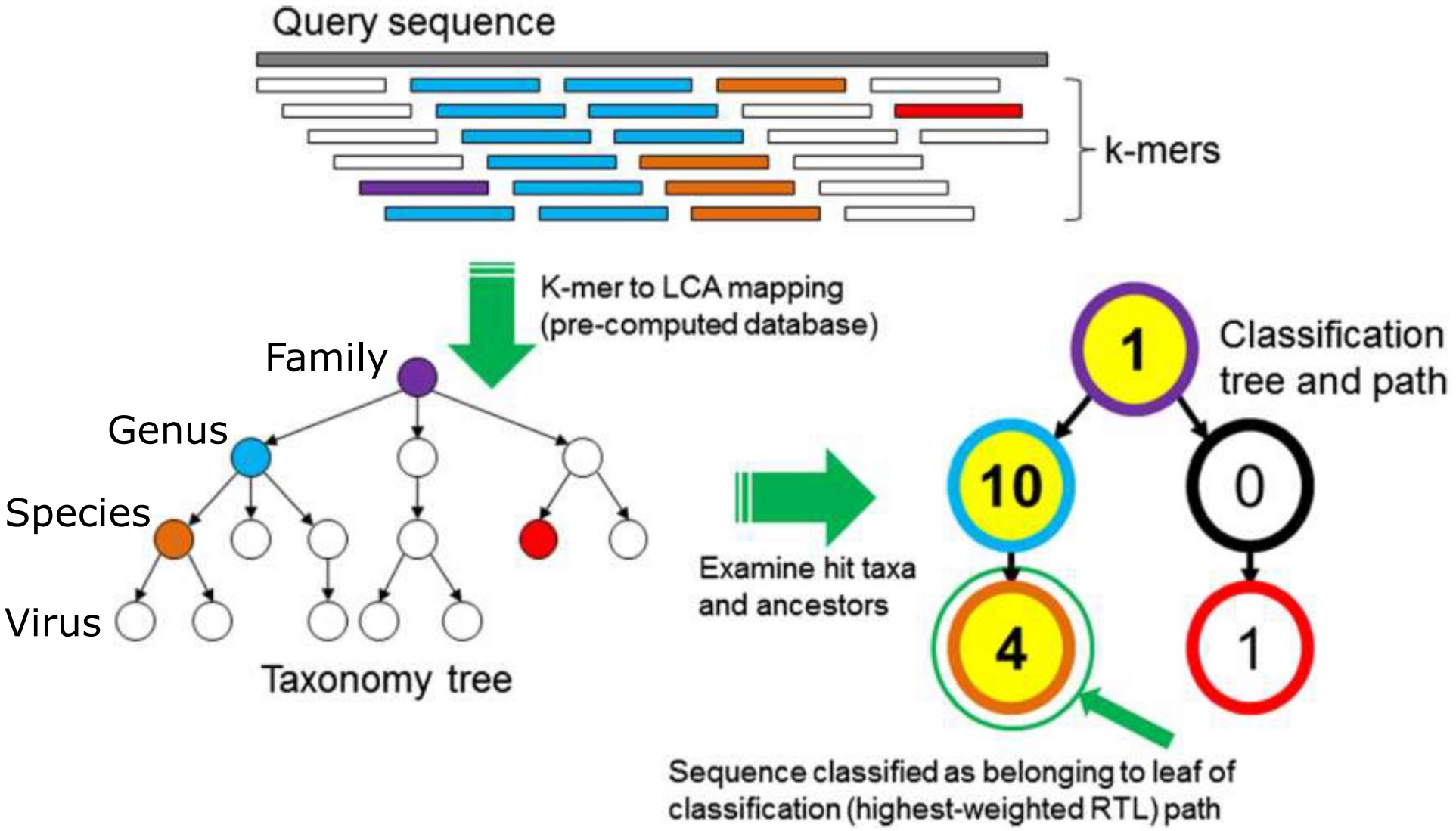


(default, K = 31)

Lowest common ancestor database

K-mer	LCA
ACC … GT	g__Escherichia
ACG … GT	s__E. coli
AGT … AA	p__Proteobacteria
…	
TGA … TT	d__Bacteria

Kraken: classification tree



Australian Centre for Ecogenomics

Wood and Salzberg, Genome Biology, 2014

Kraken analyses

- **A kraken analysis is highly dependent on the database used**
  - If the database does not contain the virus it will miss it
  - If the virus in your sample has diverged significantly away from the database virus sequence it will not classify all reads
  - If the virus only includes viruses – non-virus may be misclassified (prob low level)
  - If the virus contains ‘all’ virus sequences – there is a lot of garbage on GenBank
    - Retroviruses in both virus and host, EVEs, general noise
  - Kraken databases routinely ‘masked’ – low complexity removed

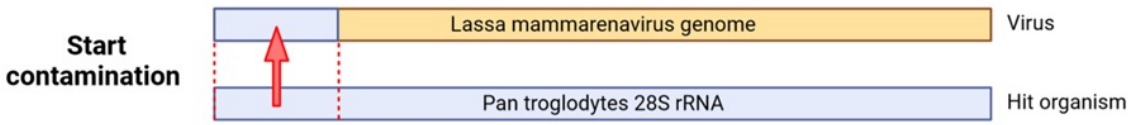


Figure 1. *E*xample of Lassa virus genome contaminated with chimpanzee ribosomal sequence.

Figure from Ailsa Orr, MSc Bioinformatics thesis, 2023

Kraken database

Collection	Contains	Date	Archive size (GB)	Index size (GB)	HTTPS URL	Inspect
Viral	Refeq viral	6/5/2023	0.5	0.6	.tar.gz	.txt
MinusB	Refeq archaea, viral, plasmid, human[1], UniVec_Core	6/5/2023	6.5	9.4	.tar.gz	.txt
Standard	Refeq archaea, bacteria, viral, plasmid, human[1], UniVec_Core	6/5/2023	51	67	.tar.gz	.txt
Standard-8	Standard with DB capped at 8 GB	6/5/2023	5.5	7.5	.tar.gz	.txt
Standard-16	Standard with DB capped at 16 GB	6/5/2023	11	15	.tar.gz	.txt
PlusPF	Standard plus Refeq protozoa & fungi	6/5/2023	55	71	.tar.gz	.txt
PlusPF-8	PlusPF with DB capped at 8 GB	6/5/2023	5.5	7.5	.tar.gz	.txt
PlusPF-16	PlusPF with DB capped at 16 GB	6/5/2023	11	15	.tar.gz	.txt
PlusPFP	Standard plus Refeq protozoa, fungi & plant	6/5/2023	108	148	.tar.gz	.txt
PlusPFP-8	PlusPFP with DB capped at 8 GB	6/5/2023	5.1	7.5	.tar.gz	.txt
PlusPFP-16	PlusPFP with DB capped at 16 GB	6/5/2023	10	15	.tar.gz	.txt
nt Database	Very large collection, inclusive of GenBank, RefSeq, TPA and PDB	5/2/2023	360	480	.tar.gz	.txt
EuPathDB46[2]	Eukaryotic pathogen genomes with contaminants removed	4/18/2023	8.4	11	.tar.gz	.txt

RAM requirement

RefSeq can miss divergent genotypes!

Just viruses can give false positives

Recomended

Subsampled

UniVec-Core vector sequences from bacteria, phage, yeast, and synthetic construct

NT – everything 480GB of RAM!!

Kmer metagenomics

- **Quick and dirty** (false positives and false negatives)
  - Very good for spotting possible contamination
  - Cross-validating with BLAST based metagenomics
  - Can tell you about the viruses (possibly) present in your sample

- But **does not give you genome sequences** like consensus calling and de novo
  - Classifies (assigns) reads to certain taxons (species, genus, family etc)

- Nucleotide – may miss divergent viruses in your sample

- Exact kmer match – suspectable to high sequence error (nanopore)
  - Kraken uses a kmer size of 31 – Q10 (1 in 10 error) would give ~3 errors

Kraken outputs

- **Output file**
  - Large (not really human readable)
  - One line per read – tax IDs and kmer counts
- **Report file**
  - Tab delimited summary of number of reads assign to different taxon levels
- **Krona file**
  - Need to use KronaTools to convert
  - Krona plot of the results - visual summary

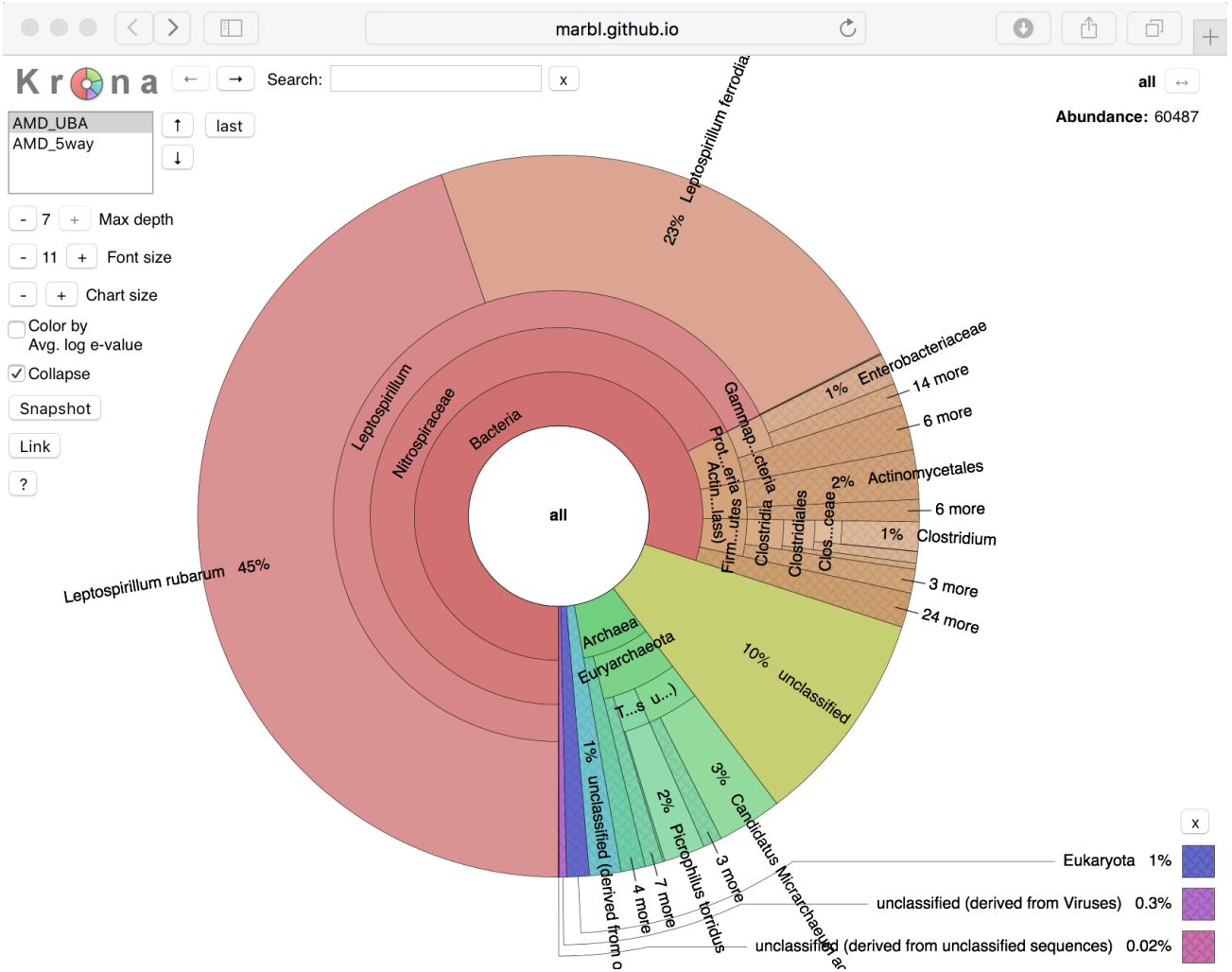
Example kraken report

- Text file summary of all the kraken hits:
  - **Percentage of reads** covered by the clade rooted at this taxon
  - **Number of reads** covered by the clade **rooted** at this taxon
  - **Number of reads** assigned **directly** to this taxon
  - A **rank code**, indicating (U)nclassified, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies. All other ranks are simply '-'.
  - **NCBI taxonomy ID**
  - Indented **scientific name**

- **EXAMPLE**

Krona plot

- Krona is an interactive visualization tool for exploring the composition of metagenomes within a Web browser.
- Krona uses multilevel pie charts to visualize both the most abundant organisms and their most specific classifications.
- **EXAMPLE**



Kraken alternatives

- **Kraken** – the first version of kraken (no longer supported)
- **KrakenUniq** – counts kmers, discounts duplicates, looking at how much of a database sequence (viral genome) is covered by kmers – uses Kraken1 databases (large)
- **Kraken2 –** newer version of Kraken, does incorporate elements of KrakenUniq
  - Minimum-hit-groups: how many groups of kmers within a read are needed to classify – e.g. stops a single kmer classifying a read
- **Bracken** takes the classified read counts and estimates the abundance of each taxon in the sample. **Pavian** can be used to explore and visualize this sample to spot the difference.

- **Centrifuge –** very similar to first Kraken but lower memory needed (but now there is Kraken2)
- **Kaiju -** reads are translated into amino acid sequences
- **VirFinder, DisCVR, Genome Detective**

Nanopore Data

Microb Genom. 2022; 8(10): mgen000886.
Published online 2022 Oct 21. doi: 10.1099/mgen.0.000886
PMCID: PMC9676057
PMID: 36269282

Benchmarking taxonomic classifiers with Illumina and Nanopore sequence data for clinical metagenomic diagnostic applications

Kumeren N. Govender[1,*] and David W. Eyre[1,2]

Simulating both with and without sequencing error for both the Illumina and Oxford Nanopore platforms, we evaluated commonly used classification tools including Kraken2, Bracken and Centrifuge, utilizing mini (8 GB) and standard (30-50 GB) databases.

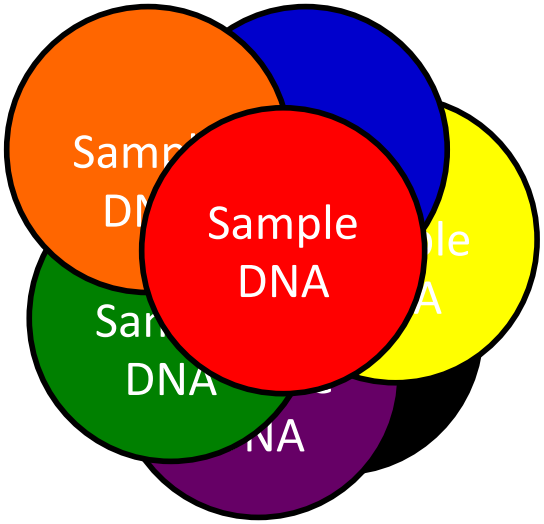
By filtering out shorter Nanopore reads we found performance similar or superior to Illumina sequencing, despite higher sequencing error rates. Misclassification was more common when the misclassified species had a higher average nucleotide identity to the true species.

Practical – Part 1

- **Human sample [instructions]**
- **Vampire bat sample**
- **VIZIONS sample**

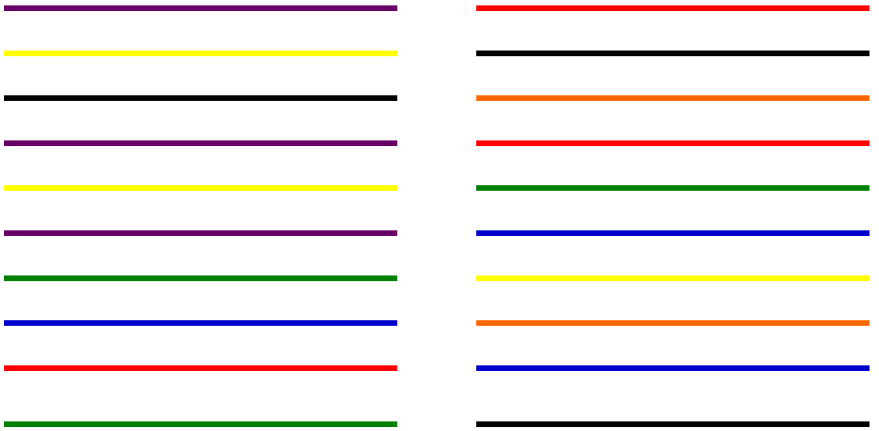
- Run kraken – report file
- Create Krona plot

Metagenomics methods



DNA extraction

Sequencing quality control



De novo assembly

BLAST

Species A

Species B

Species C

Species D

Species E

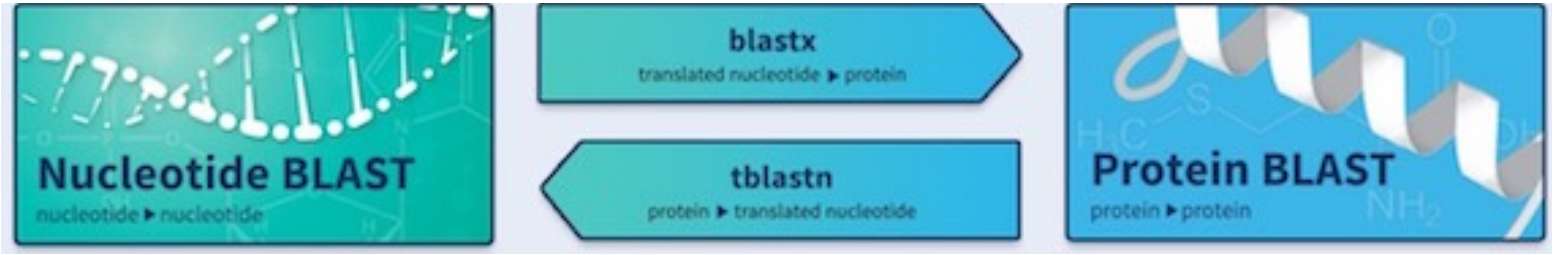
Species F

Species G

De novo contigs

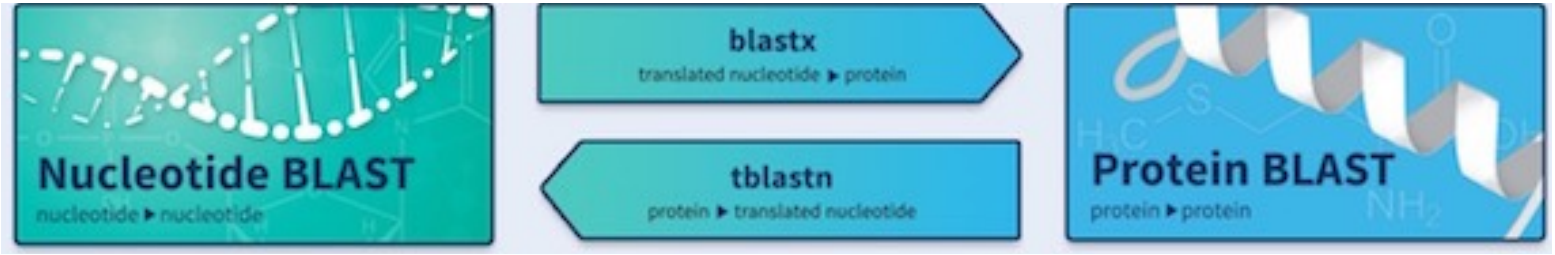
- De novo assembly gives us contigs
  - Some big
  - Some small
- What are they?
- Are they viruses?
- Which viruses?
- **BLAST – find closest hits, what species is it?**

BLAST types



- **BLAST** – **B**asic **L**ocal **A**lignment **S**earch **T**ool
- **BLASTn** – nucleotide Vs nucleotide
- **BLASTp** – protein Vs protein
- **BLASTx** – nucleotide Vs protein
  - 6 frame (3 fwd + 3 rev) and BLAST the resultant protein sequence against protein DB

BLAST types



- With metagenomics you are often sequencing samples from hosts or geographical locations not previously sequenced – possibly contain novel viruses
- Some viruses are highly divergent – with much closer similarity at the amino acid (protein) level compared to the nucleotide level
- **BLASTx – 6 frame translate the contigs, compare to protein database**

BLAST databases

CVR
Medical Research Council
University of Glasgow
Centre for Virus Research

- **BLAST** – **B**asic **L**ocal **A**lignment **S**earch **T**ool
- **NT DB** – all GenBank **nucleotide** sequences – duplicate sequences merged (non-redundant)
- **NR DB** – all GenBank **protein** sequences – duplicate sequences merged (non-redundant)
- **RefSeq DB** – curated representative genome/transcript sequences for each organism
  - Nucleotide
  - Protein
- **VM – for speed and disk space – we will be using a BLAST DB that just contains protein from Viral RefSeqs (one for each viral species) – NOT RECOMMENDED**

DIAMOND

CVR
Medical Research Council
University of Glasgow
Centre for Virus Research

- DIAMOND is a sequence aligner for protein and translated DNA searches, designed for high performance analysis of big sequence data.

- Much faster BLASTx (100x – 10,000x)

- Needs a protein database (can convert the BLAST ones)

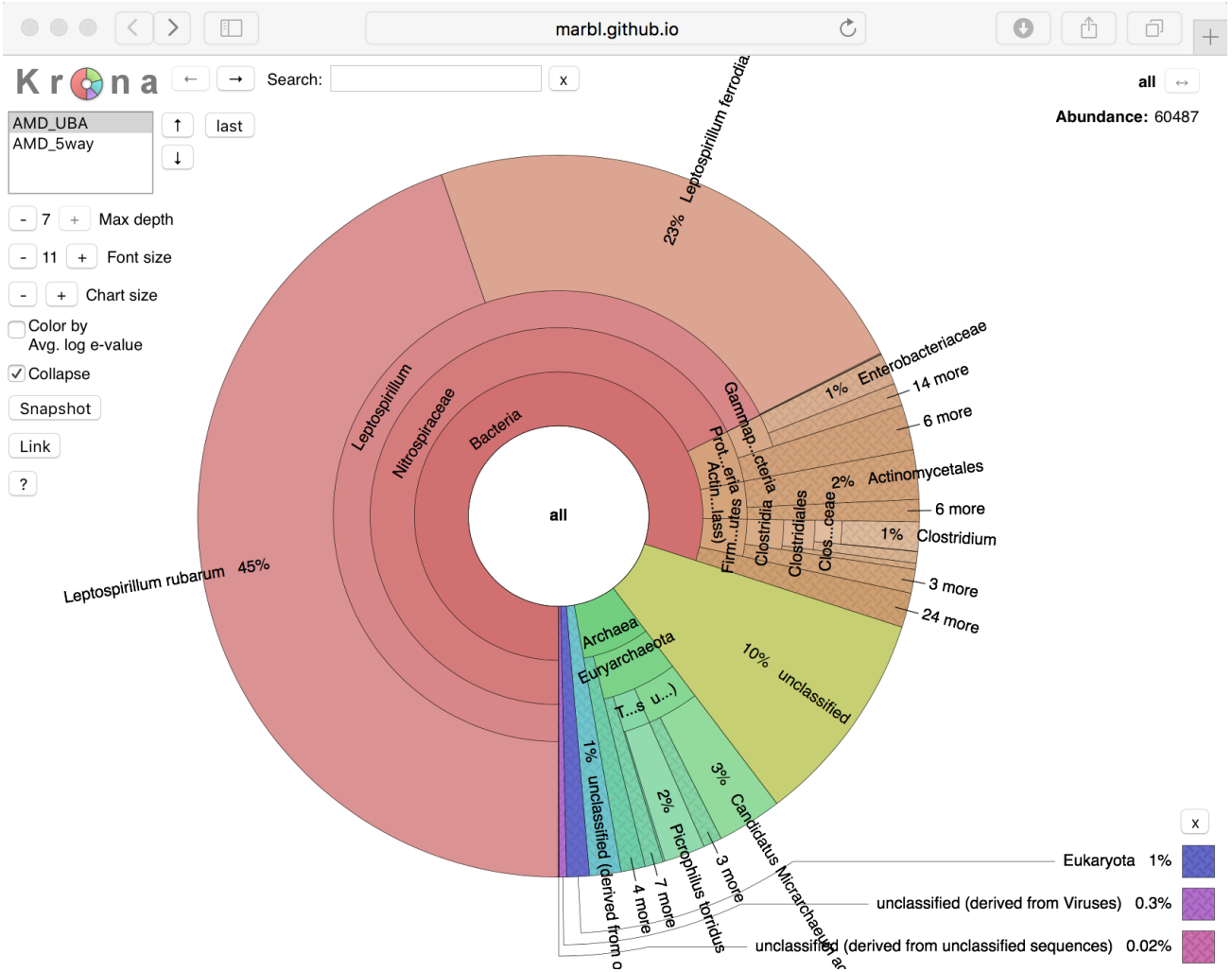
- The bigger the database the slower the speed, the more sequences the slower the speed:

- NR – slow hours
- RefSeq – medium up to 2 hours
- Viruses – fast – minutes [not recommended]

```
diamond blastx --threads 6 --outfmt 6 --db ~/Diamond/viral-refseq-prot --out
contigs_diamond.txt -q contigs.fasta
```

Krona plot

```
ktImportBLAST contigs_diamond.txt -o contigs_diamond.html -tax ~/Diamond/
```



Other options

CVR
Medical Research Council
University of Glasgow
Centre for Virus Research

- MEGAN – Meta Genome ANalyzer - works well with DIAMOND (https://software-ab.cs.uni-tuebingen.de/download/public/tutorial-aug2021/welcome.html)

Practical – Part 2

- **Human sample [instructions]**
- **Vampire bat sample**
- **VIZIONS sample**

- Run spades (it will take a while)
- Pre-run contigs
- Diamond
- Krona