Kraken2

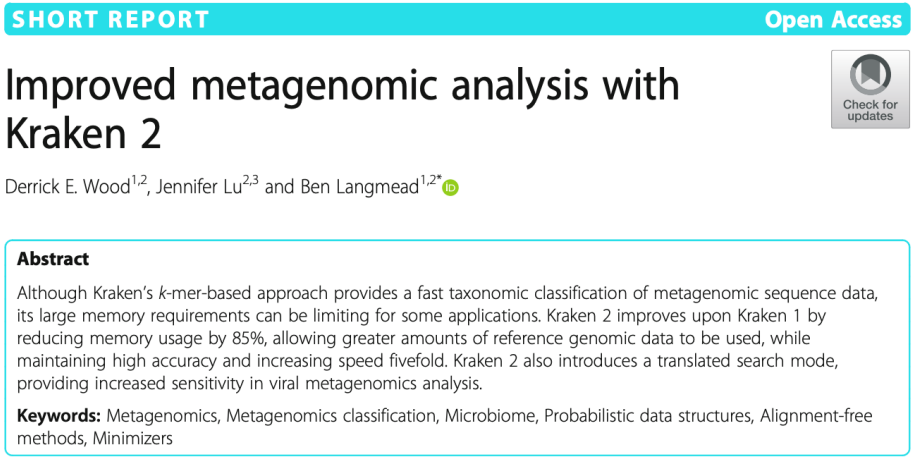
Quick introduction

Dr. Natalia Zajac

31.03.2023

Kraken2

- Building Kraken database
- Classification of reads
- Visualisation of results



SHORT REPORT Open Access

Improved metagenomic analysis with Kraken 2

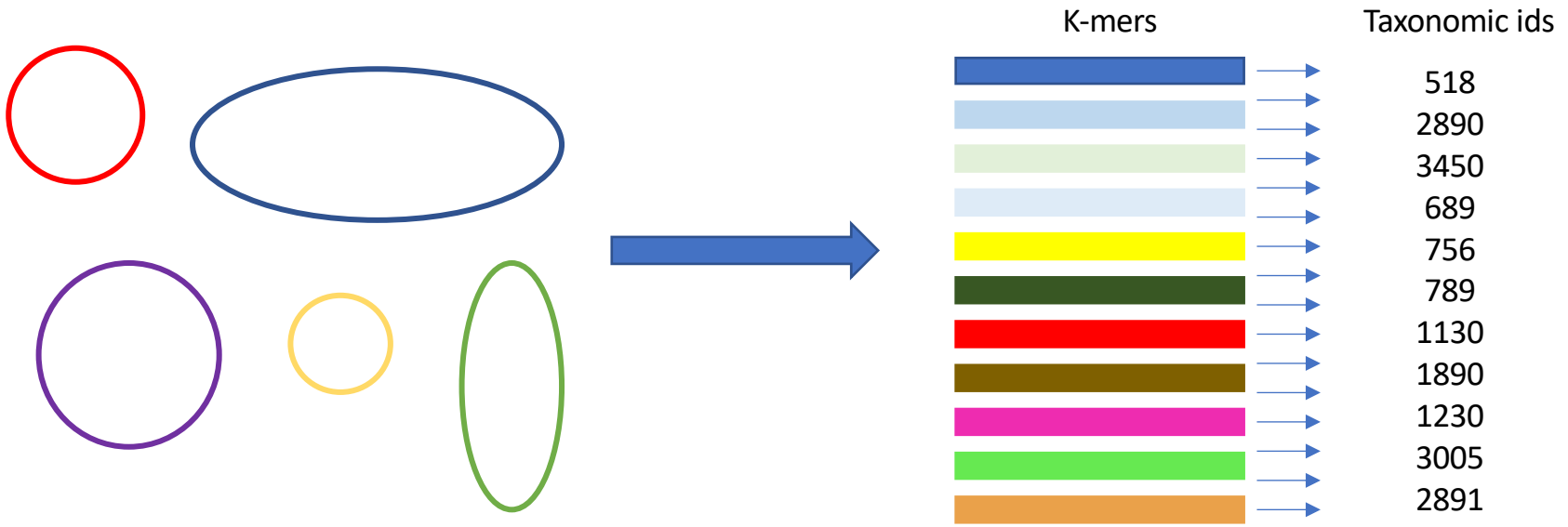
Derrick E. Wood[1,2], Jennifer Lu[2,3] and Ben Langmead[1,2*]

Abstract

Although Kraken's *k*-mer-based approach provides a fast taxonomic classification of metagenomic sequence data, its large memory requirements can be limiting for some applications. Kraken 2 improves upon Kraken 1 by reducing memory usage by 85%, allowing greater amounts of reference genomic data to be used, while maintaining high accuracy and increasing speed fivefold. Kraken 2 also introduces a translated search mode, providing increased sensitivity in viral metagenomics analysis.

Keywords: Metagenomics, Metagenomics classification, Microbiome, Probabilistic data structures, Alignment-free methods, Minimizers

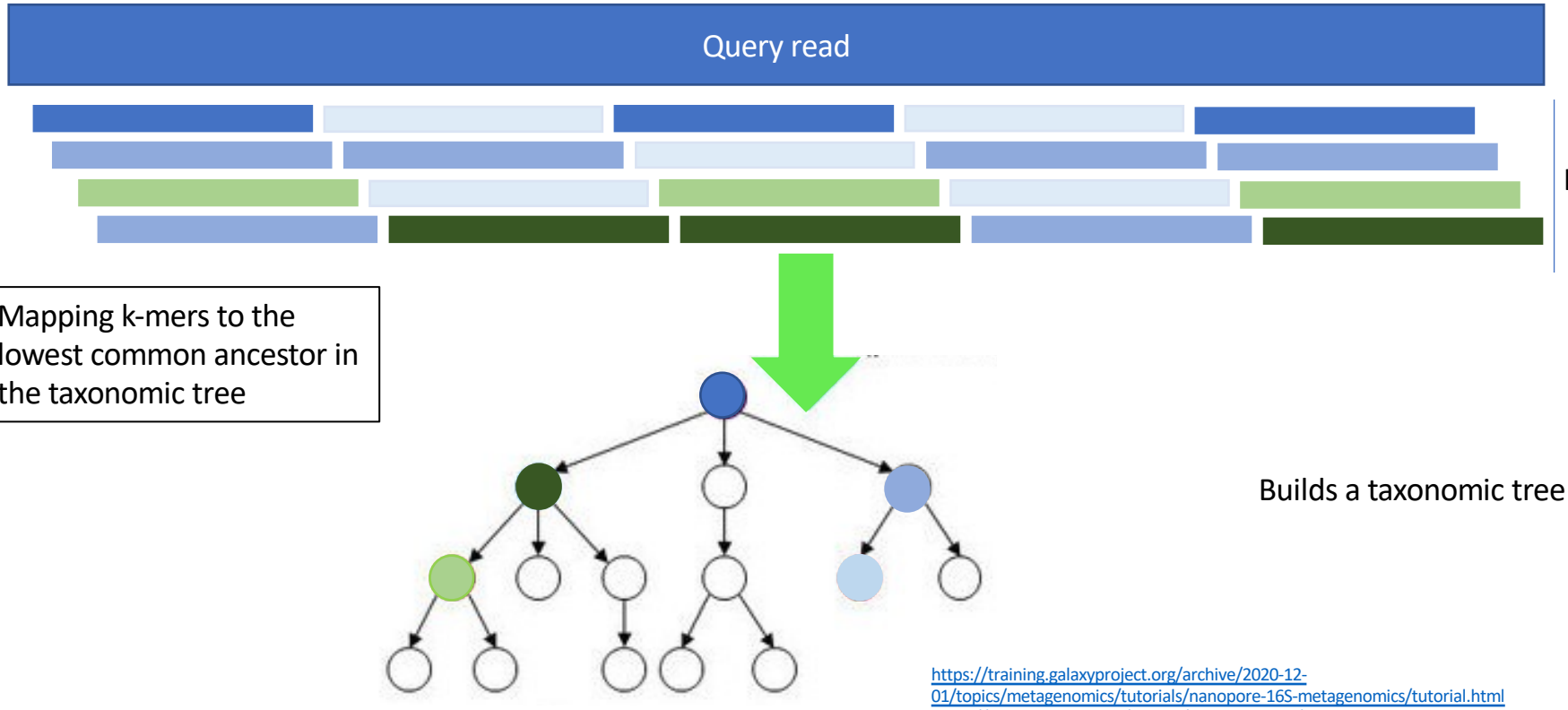
Kraken database



Kraken database

- The complete Max database is 75Gb large and contains Archea, Bacteria and Viral genomes
- It can also have a version with Eukaryotes which is > 100Gb
- The Mini version is only 8Gb and contains a subset
- You can also build a custom database for things you are interested in and when you add something to the database it needs to be in the fasta format with an **NCBI accession number** or a **taxonomic ID**

Kraken classification



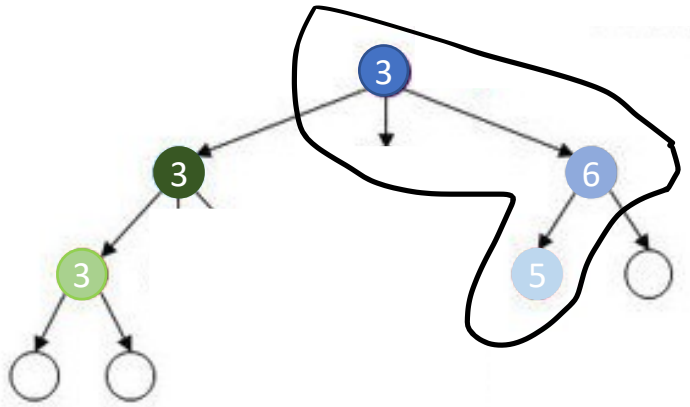
Mapping k-mers to the lowest common ancestor in the taxonomic tree

Builds a taxonomic tree

https://training.galaxyproject.org/archive/2020-12-01/topics/metagenomics/tutorials/nanopore-16S-metagenomics/tutorial.html
https://www.coursera.org/lecture/metagenomics/demo-of-metagenomic-classification-using-kraken-7GIn6

Kraken classification

- Determine highest weighted root-to-leaf path



Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

https://training.galaxyproject.org/archive/2020-12-01/topics/metagenomics/tutorials/nanopore-16S-metagenomics/tutorial.html
https://www.coursera.org/lecture/metagenomics/demo-of-metagenomic-classification-using-kraken-7GIn6

1.Percentage of reads covered by the clade rooted at this taxon
2.Number of reads covered by the clade rooted at this taxon
3.Number of reads assigned directly to this taxon
4.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 '-'.
5.NCBI taxonomy ID
6.indented scientific name

```
| 70.22   54418   54418   U       0       unclassified
  29.78   23073   0       R       1       root
  29.78   23073   0       D       10239     Viruses
  29.69   23010   0       D1      2731342     Monodnaviria
  29.20   22625   0       K       2732091       Sangervirae
  29.20   22625   0       P       2732412         Phixviricota
  29.20   22625   0       C       2732413           Malgrandaviricetes
  29.20   22625   0       O       2732414             Petitvirales
  29.20   22625   0       F       10841                 Microviridae
  29.20   22625   121     F1      1910950                 Bullavirinae
  29.02   22491   0       G       1910954                   Sinsheimervirus
  29.02   22491   22491   S       10847                       Escherichia virus phiX174
   0.01   9       2       G       1910952                   Gequatrovirus
   0.01   4       0       S       1986034                     Escherichia virus G4
   0.00   3       3       S1      489829                        Enterobacteria phage ID18
sensu lato
   0.00   1       1       S1      10843                         Escherichia phage G4
   0.00   3       0       S       1910969                     Escherichia virus Talmos
   0.00   3       3       S1      511969                        Escherichia phage ID2
Moscow/ID/2001
   0.01   4       1       G       1910951                   Alphatrevirus
   0.00   2       0       S       1945586                     Escherichia virus NC29
   0.00   2       2       S1      338110                        Escherichia phage NC29
   0.00   1       0       S       1945588                     Escherichia virus ID62
   0.00   1       1       S1      338107                        Escherichia phage ID62
   0.50   385     0       K       2732092       Shotokuvirae
   0.50   385     0       P       2732415         Cossaviricota
   0.50   385     0       C       2732421           Papovaviricetes
   0.50   385     0       O       2732532             Sepolyvirales
   0.50   385     4       F       151341                Polyomaviridae
```

Summary

- Chops all genomes into k-mers and links them to a taxonomic ID
- With your query you search for exacts hits in the database
- Searches for highest weighted root-to-leaf paths and assignes taxonomic IDs of the lowest node to read
- Tutorial: SUSHI: MetaAtlas_data – which species do you find?