Kraken:
ultrafast metagenomic sequence classification using exact alignments

Derrick E. Wood and Steven L. Salzberg

Bioinformatics journal club

October 8, 2014

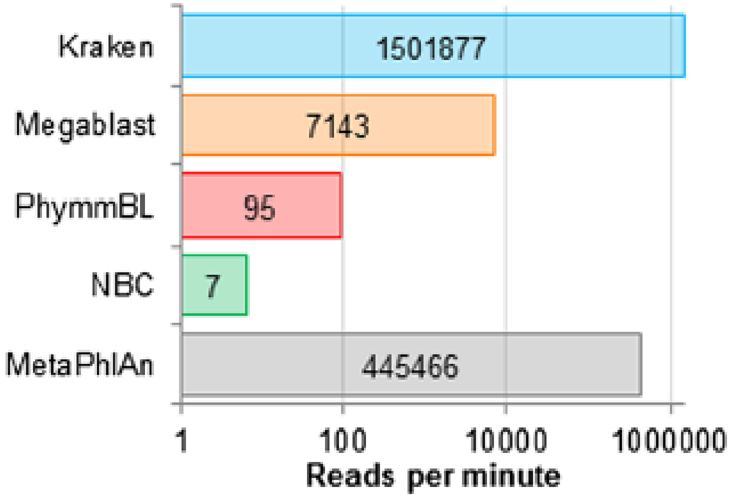
Märt Roosaare

Need for speed

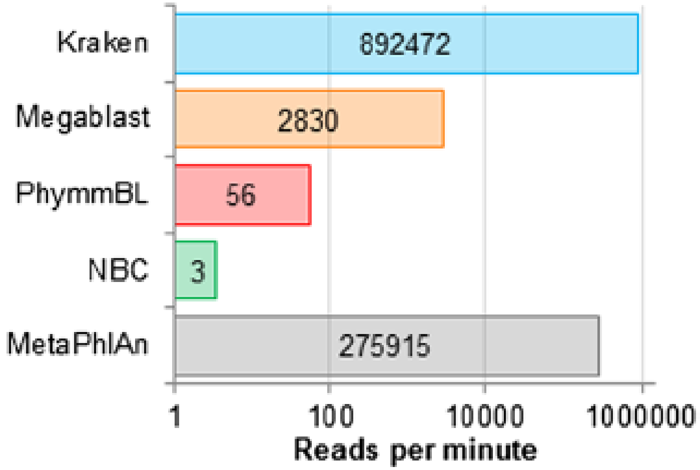
- Metagenomic data - huge amount of reads
  - Among them lots of novel sequences
- Otherwise good approaches (BLAST) - very slow (turnover times 24+ hours) - TOO SLOW for clinical diagnostics, not cost effective for other uses (metagenomics)
- Abundance estimation - reduce search space by using only selected marker genes from each organism - able to classify only a small set of reads

Need for speed

- Some programs might have higher accuracy than BLAST, but are even slower…
  - PhymmBL - uses Markov models, NBC - Naive Bayesian Classifier
- Others (abundance estimators) have speed, but not enough resolution (MetaPhlAn)



HiSeq (92bp)



MiSeq (156bp)

Kraken

- K-mer based, default k = 31
- Exact matching of k-mers (database vs reads)
- Hierarchical database: k-mer associated with lowest common ancestor (LCA - highest taxon that contains this k-mer)

Some drawbacks:

- Can only classify sequences that have k-mers in the database (low error tolerance)
- No confidence scores
- Not as sensitive as some other methods

Kraken - database creation

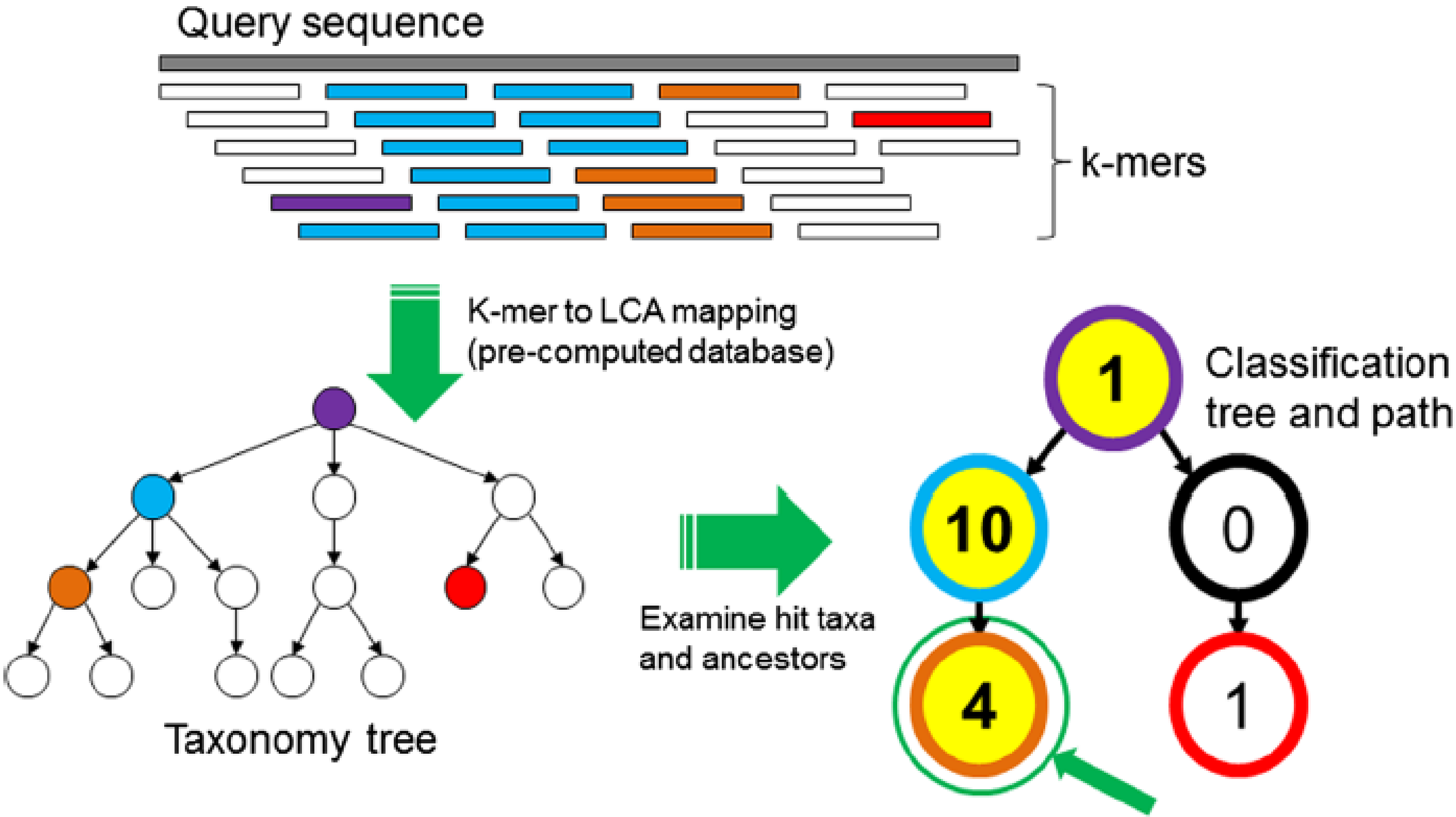
1. Choose library of genomes (NCBI RefSeq)

2. Split the library into k-mers (Jellyfish)

3. Process all the sequences to obtain taxon information (NCBI taxonomy database)
    - By default, all k-mers are given taxon identifiers of the sequence they are from
    - If a k-mer already has its taxon ID set when processed, Kraken finds respective LCA

Database size: 70 GB, must fit into RAM

Kraken - classification process



Query sequence

k-mers

K-mer to LCA mapping (pre-computed database)

Taxonomy tree

Examine hit taxa and ancestors

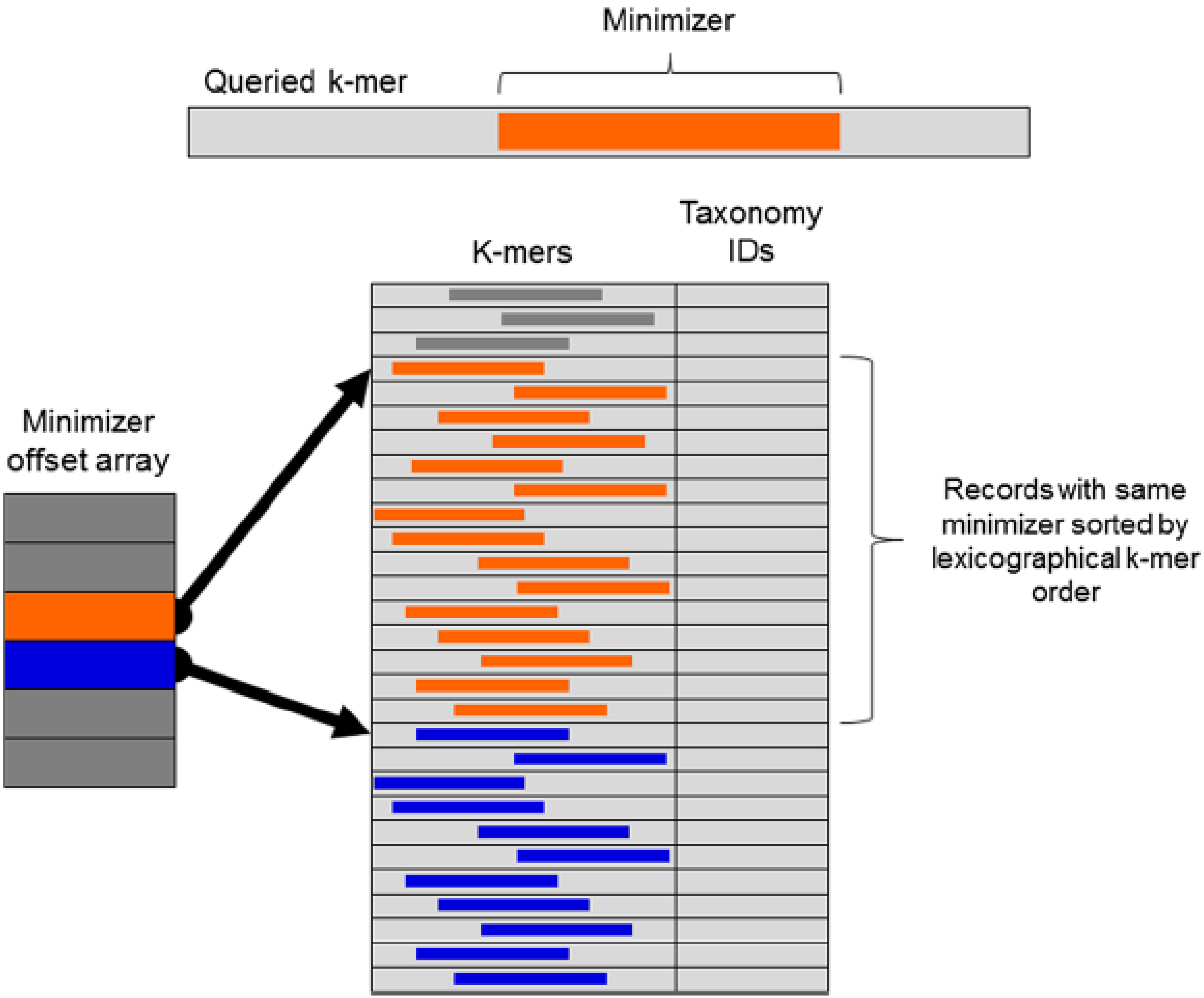
Classification tree and path

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

Kraken – further speed improvements

- Main idea – adjacent k-mers are often queried one after the other and they share substantial amount of sequence

- Using smaller substrings („minimizers“) of a k-mer to group them together and reduce search space
- Using same search range for next query (if query fails => compute minimizer, if it is the same, k-mer is not in database)



Minimizer

Queried k-mer

K-mers

Taxonomy IDs

Minimizer offset array

Records with same minimizer sorted by lexicographical k-mer order

Kraken - other variants

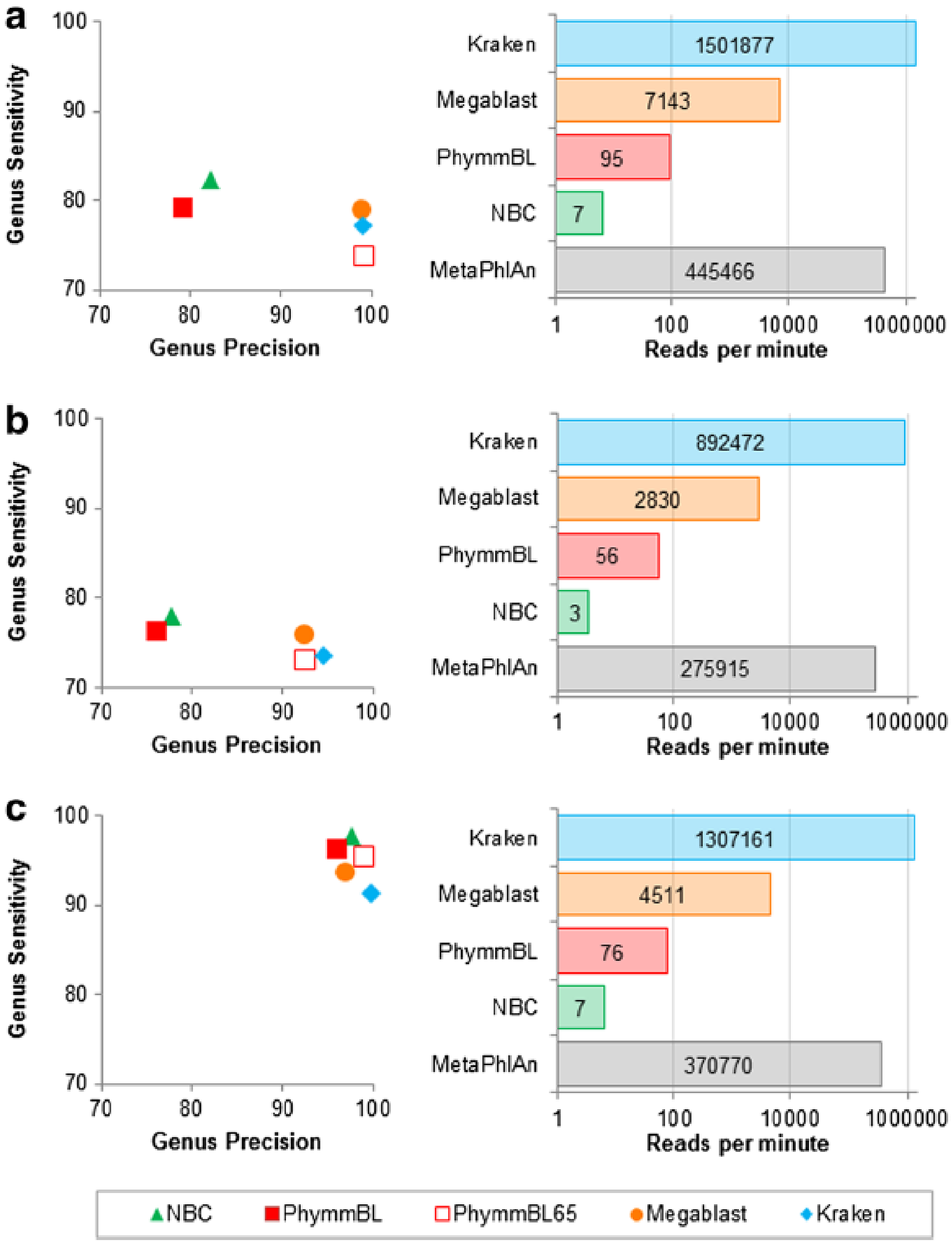
- MiniKraken - reduced database size (4GB), uses every 19th k-mer in database (good for desktop computers and personal users)
- Kraken-Q and MiniKraken-Q - first k-mer found in database is used for final classification
- Kraken-GB - uses GenBank’s draft data as well, larger database (8500 vs 2200 genomes)

Kraken - performance test data

- Simulated datasets:
  - HiSeq (92bp) and MiSeq (156bp) - reads from bacterial WGS projects (GAGE-B or NCBI Sequence Read Archive) - 10 genomes each
  - simBA-5 (5x more errors) - RefSeq genomes (607 genera)

Test results

Kraken vs other metods



A - HiSeq | B - MiSeq | C - SimBA5

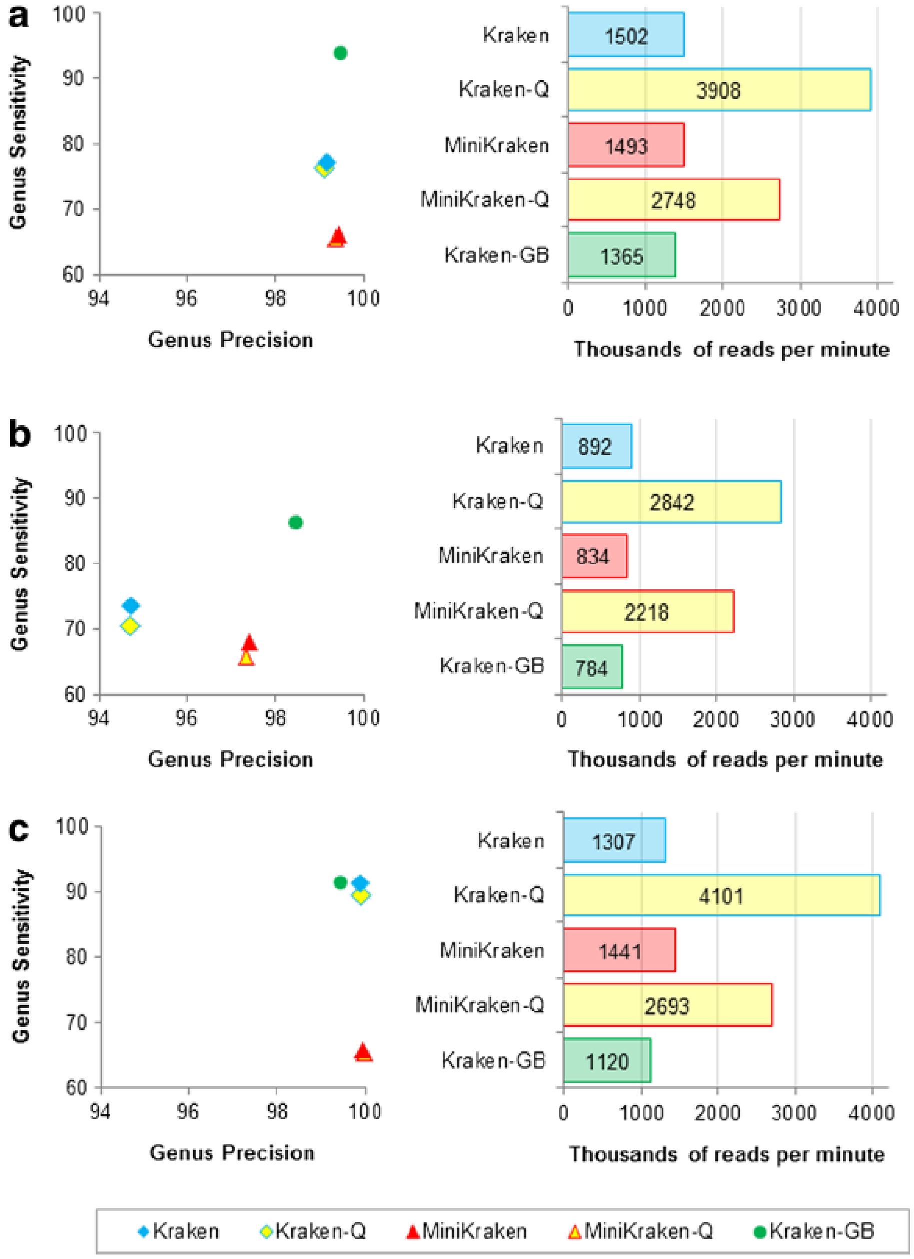
Speed - NBC and PhymmBL practically unusable, Megablast for small datasets

Sensitivity - Kraken leaves reads unclassified if there is insufficient evidence. Also, exact matching does not tolerate errors (compared to Megablast)

SimBA5 metagenome - despite of errors, sensitivity and precision still highest compared to other datasets - simulation not comparable to true WGS?

Test Results

Kraken’s different versions



A - HiSeq | B - MiSeq | C - SimBA5

MiniKraken’s smaller database leads to lower sensitivity, but not lower precision

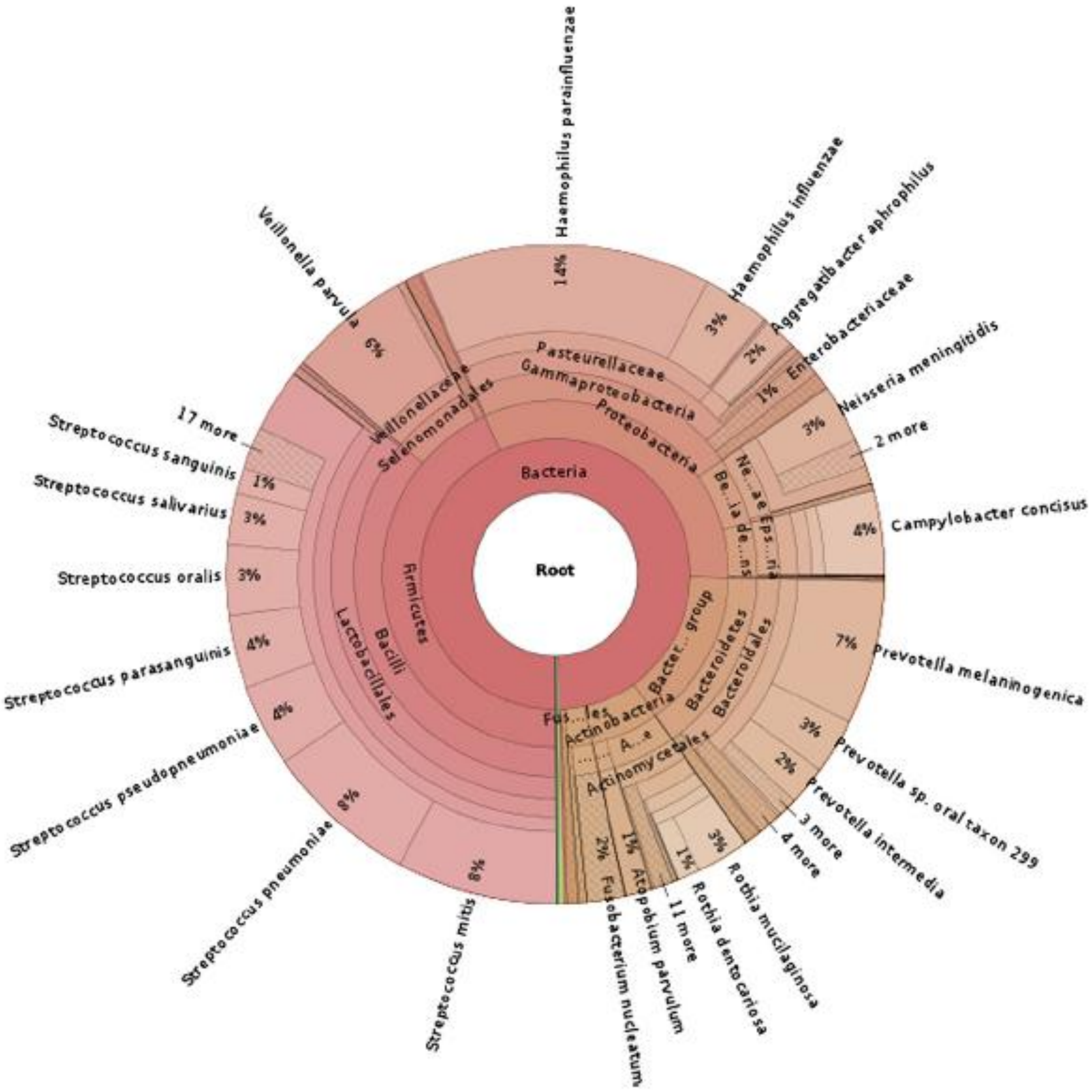
Large database (Kraken-GB) gives more sensitivity (effect lowers with more diversity), but can lower precision (contaminated data, hard to remove)

Classification by FIRST k-mer only (-Q) does not affect results much, but gives a large, 2-3 fold speed increase. Larger database allows more speed (Kraken-Q)

Kraken - conclusions

- Bigger database does not affect speed much
- Bigger database rises sensitivity in case of lower diversity, but can lower precision (contaminated, uncontrolled data) - out of 10 species in HiSeq/MiSeq, at least 1 was not in RefSeq database
- Use smaller, curated database and do not take ALL k-mers from a read into account when classifying?

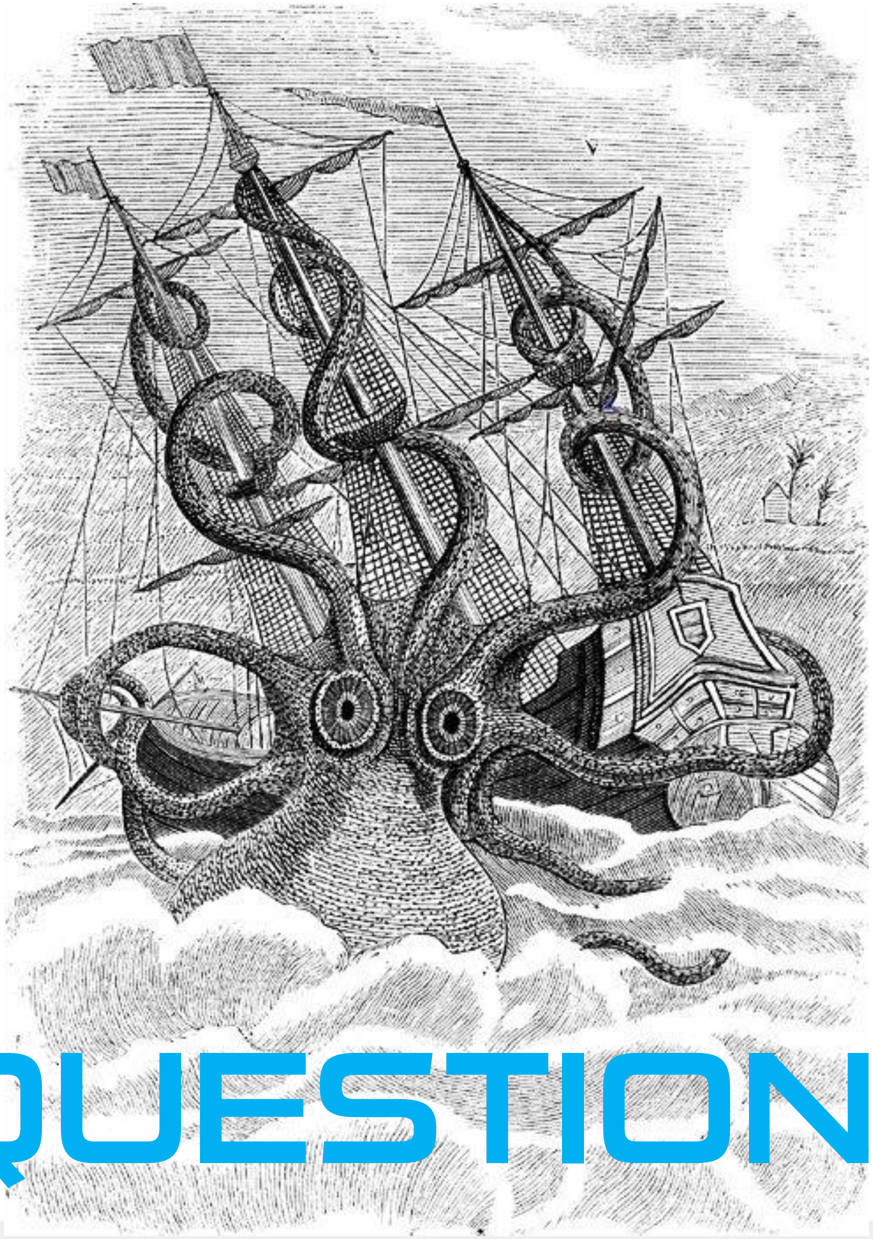
Kraken - human microbiome data



Human Microbiome Project data - 3 saliva samples

Almost 70% of reads not classified - novel sequences not present in databases (only 11% showed homology to sequences in databases, according to BLAST)

Quick and efficient - no assembling of reads or other operations required to get an overview - organisms and their abundance



QUESTIONS?