Species determination – what's in my sample?

What is my sample?

- When might you not know what your sample is?

- One species
  - You have a malaria sample but don’t know which species
  - Misidentification or no identification from culture/MALDI-TOF

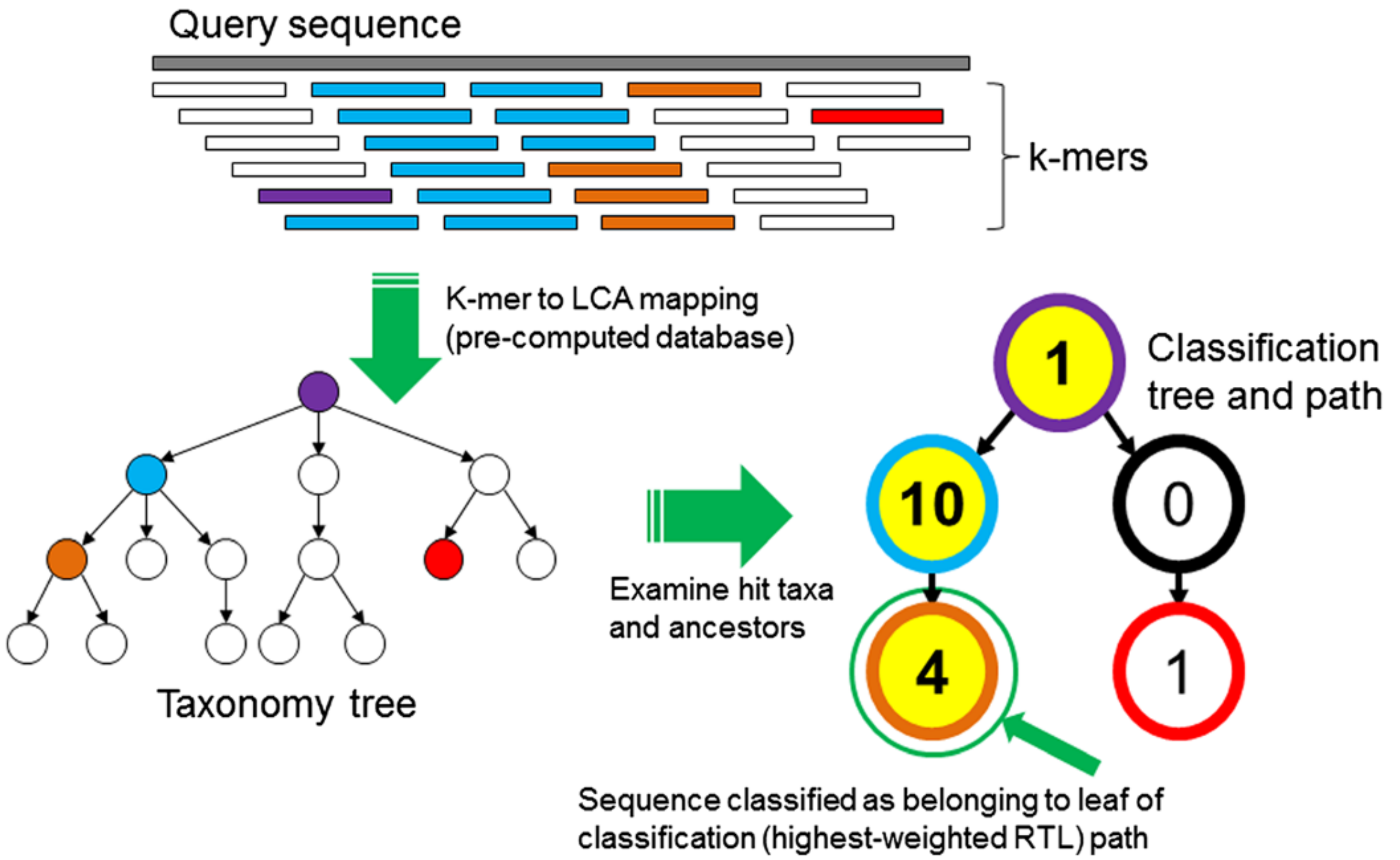
- Metagenomic samples

- Contamination

Taxonomic Classifiers

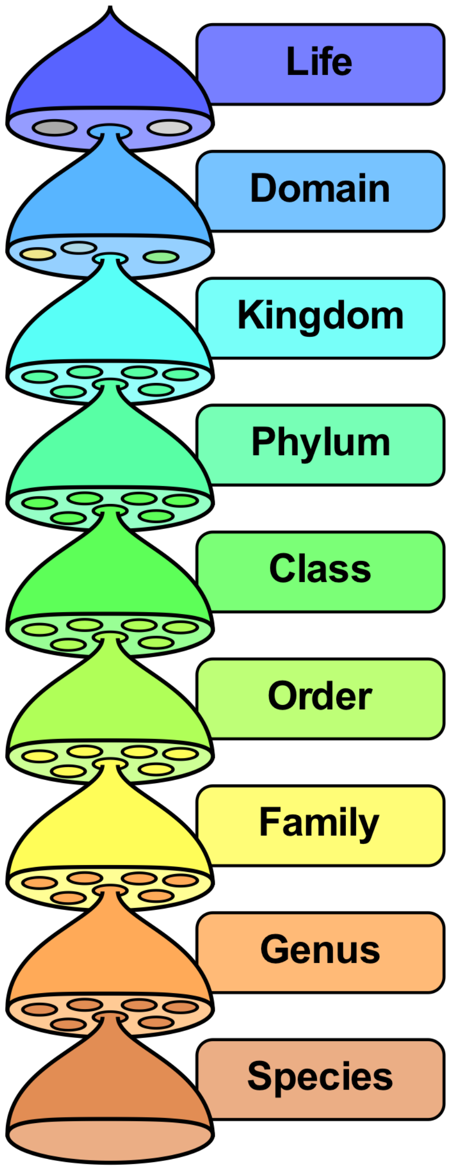
- Compare sequence reads against a database and determine the species
- BLAST works for a single sequence, too slow for a whole run
- Classifiers use database indexing and k-mer searching
- Similar accuracy to BLAST but much much faster

Kraken taxonomic classifier



Query sequence
k-mers
K-mer to LCA mapping (pre-computed database)
Taxonomy tree
Examine hit taxa and ancestors
Classification tree and path
1
10
0
4
1
Sequence classified as belonging to leaf of classification (highest-weighted RTL) path

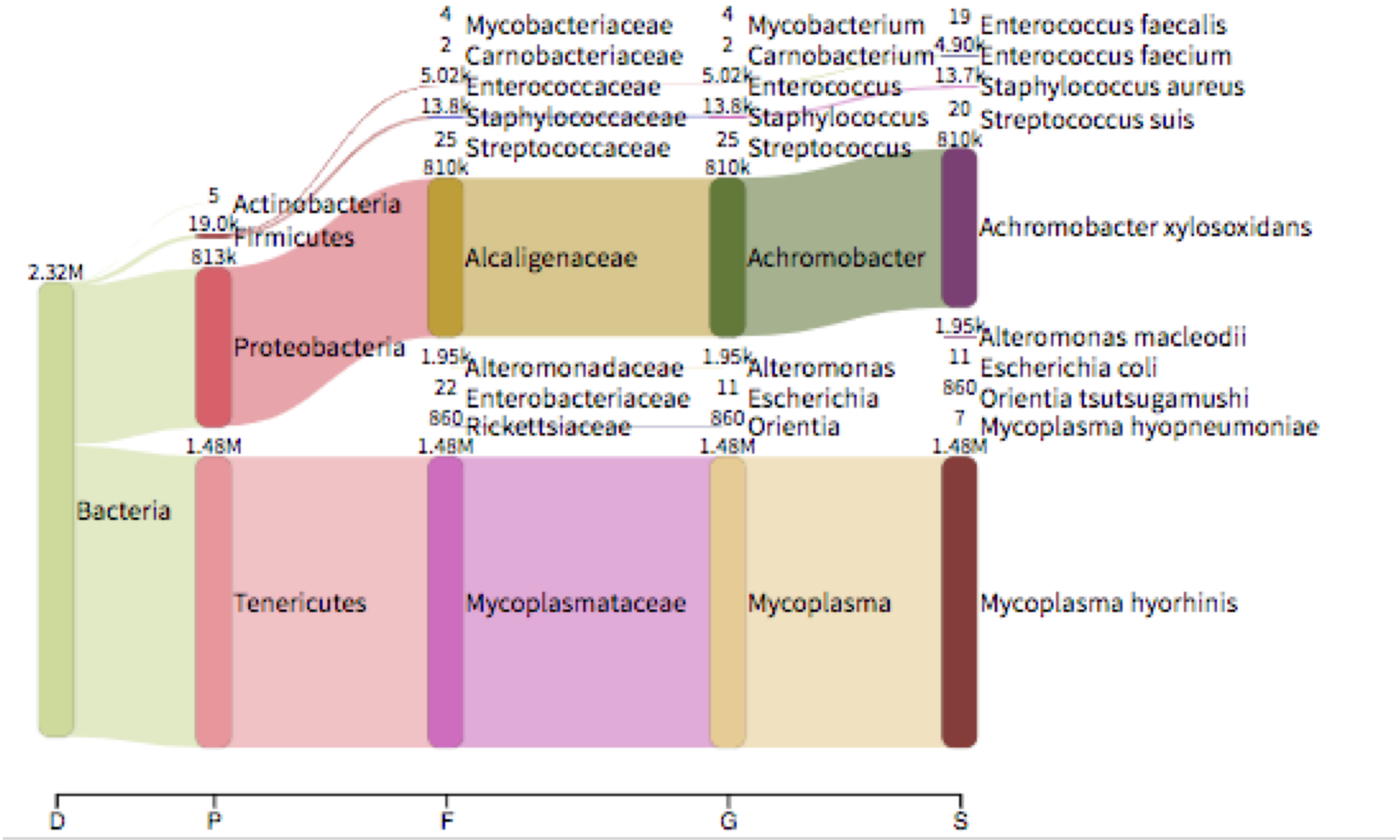
Wood and Salzberg Genome Biology 2014 **15**:R46



Life	
Domain	Bacteria
Kingdom	
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacteriales
Family	Enterobacteriaceae
Genus	Escherichia
Species	Escherichia coli

```
0.24   8553    8553    U       0       unclassified
99.76  3553969 0       -       1       root
99.76  3553969 217     -       131567    cellular organisms
65.03  2316784 542     D       2           Bacteria
41.67  1484567 0       P       544448        Tenericutes
41.67  1484567 0       C       31969           Mollicutes
41.67  1484566 0       O       2085              Mycoplasmatales
41.67  1484566 0       F       2092                Mycoplasmataceae
41.67  1484566 822     G       2093                  Mycoplasma
41.65  1483728 1434758 S       2100                    Mycoplasma hyorhinis
 0.52  18488   18488   -       936139                    Mycoplasma hyorhinis MCLD
 0.33  11708   11708   -       1118964                   Mycoplasma hyorhinis SK76
 0.25  9015    9015    -       872331                    Mycoplasma hyorhinis HUB-1
 0.20  6995    6995    -       1129369                   Mycoplasma hyorhinis GDL-1
 0.08  2764    2764    -       634997                    Mycoplasma hyorhinis DBS 1050
22.81  812626  157     P       1224          Proteobacteria
22.73  809640  0       C       28216           Betaproteobacteria
22.73  809640  0       O       80840             Burkholderiales
22.73  809640  0       F       506                 Alcaligenaceae
22.73  809640  0       G       222                   Achromobacter
22.73  809640  0       S       85698                   Achromobacter xylosoxidans
22.73  809640  809640  -       1216976                   Achromobacter xylosoxidans ATCC 27061
```

Visualisation



Breitwieser FP, Salzberg SL. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. bioRxiv 2016: 084715.

Pavian demonstration

Pitfalls of classification

- What is in your database?
  - Standard databases are bacterial and viral
  - More species, more sequences, bigger databases
  - How correct is your database? Draft genomes have contaminants
- Confidence of classification
  - What if reads are not in the database?
  - Do you look at genus, species, or strain level?
  - How confident is each match? Tradeoff sensitivity vs specificity
  - Kraken confidence threshold moves up the tree until confidence is met