Metagenomics 101

Session 10: Genome-based taxonomy & gene/genome collections

Anna Heintz-Buschart

May 2022

University of Amsterdam
Life Sciences

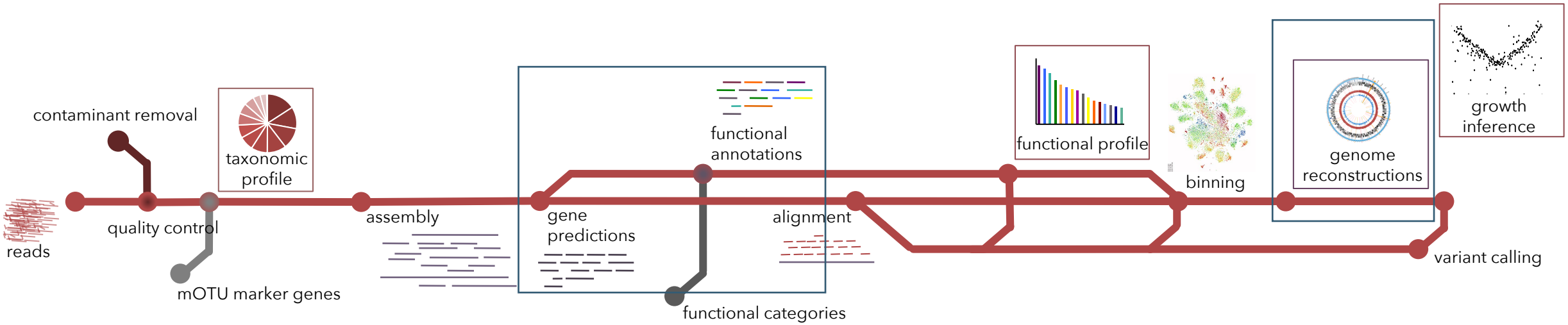
Today

- Part I:
  - how to classify a MAG
  - genome-based phylogeny

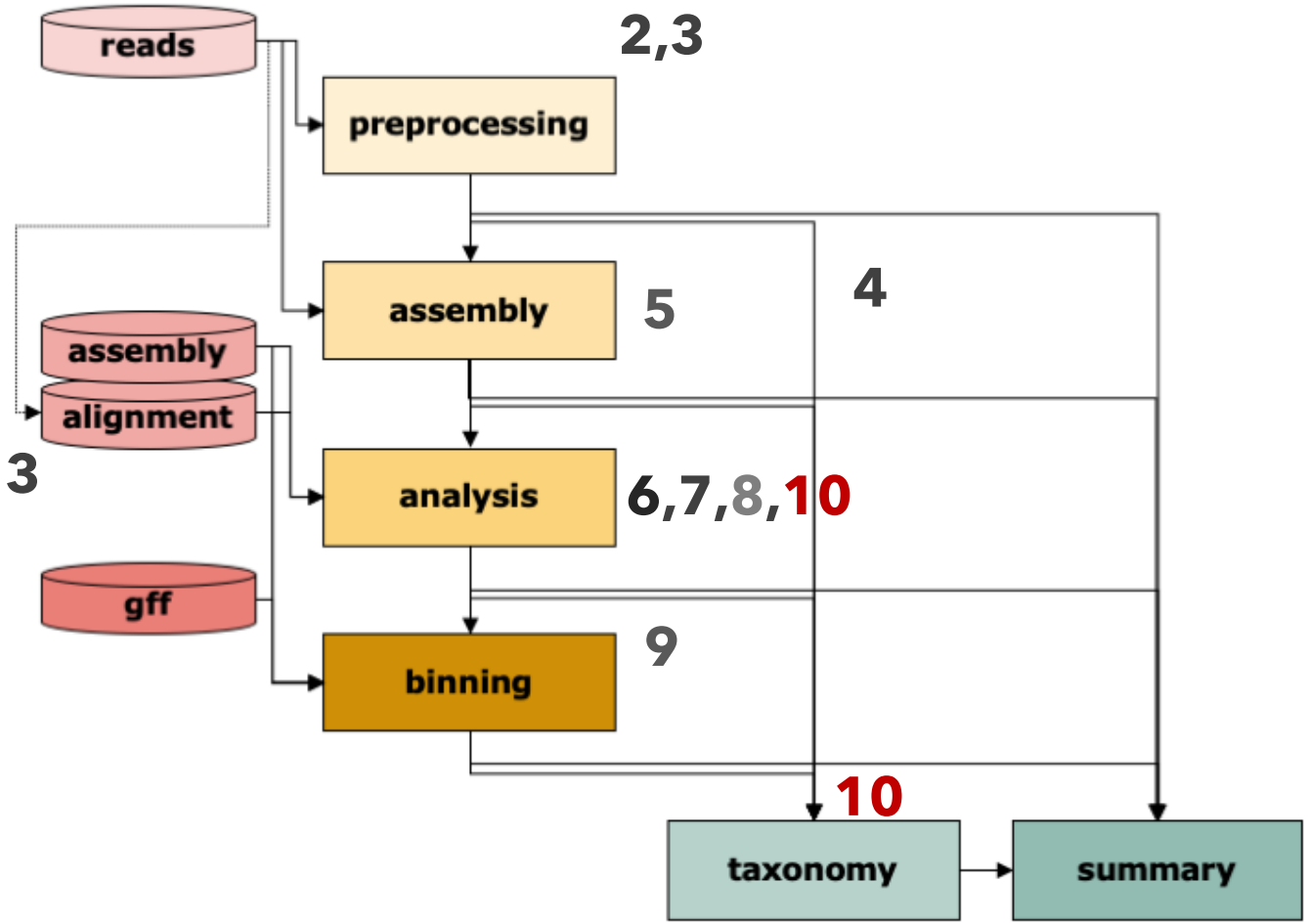
- Part II:
  - Working with genome/MAG collections
  - MAG collections' sloppy sister: gene catalogues

- The end

Metagenomics (+ other omics) pipeline



Metagenomics (+ other omics) pipeline



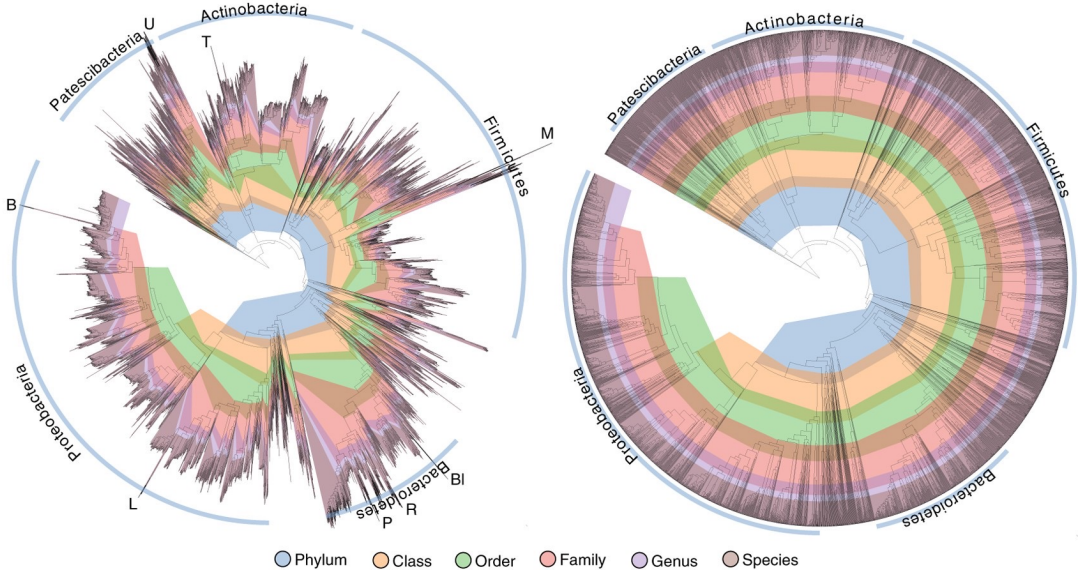
Genome-based phylogeny

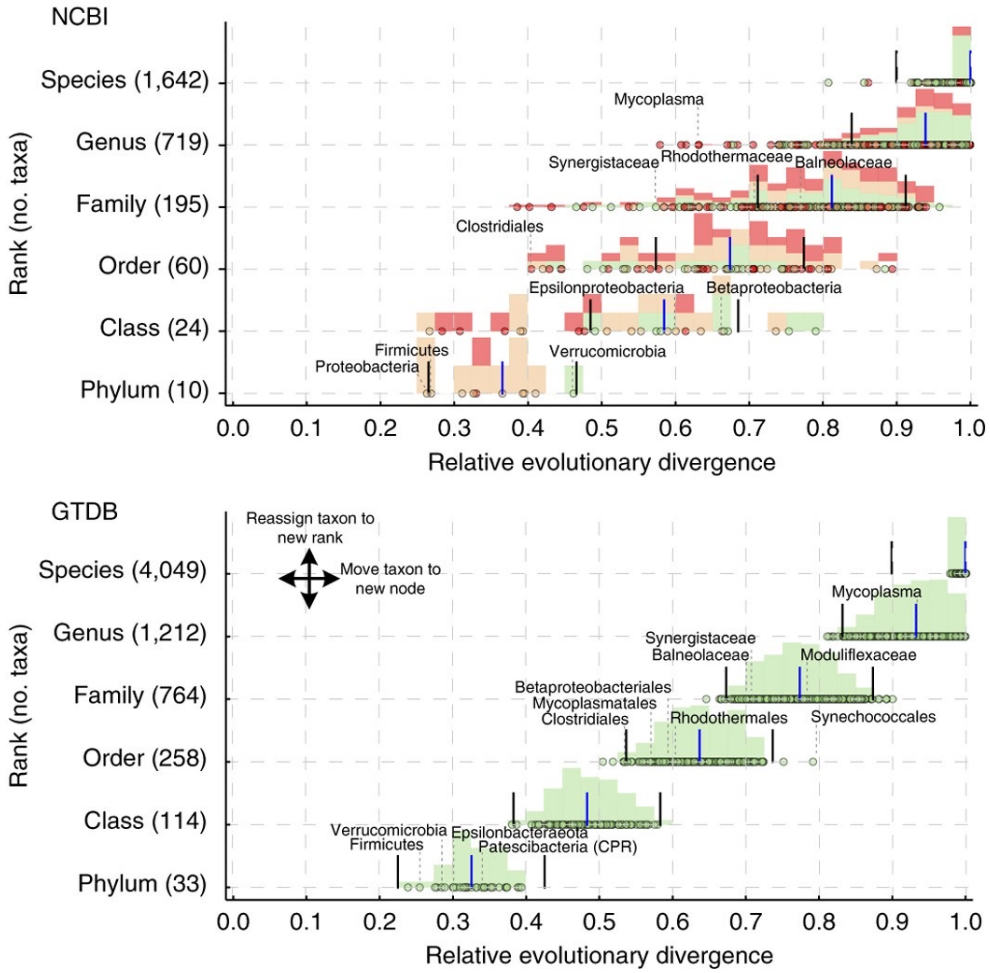
- reminder: GTDB

GTDB

Genome-based phylogeny

- reminder: GTDB





GTDB

Parks et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36, 996–1004 (2018)

GTDBtk: GTDB for your own MAGs

- call genes on the MAG (using **prodigal** - if you haven't done it yet)
- search for 120 bacterial marker genes and 53 archaea marker genes (using **HMMER**) – align "winner" set to HMM
- concatenate aligned sequences and trim (5,000 aa)
- find maximum-likelihood placement in tree (pplacer)
- classify

Bioinformatics, 36(6), 2020, 1925–1927
doi: 10.1093/bioinformatics/btz848
Advance Access Publication Date: 15 November 2019
Applications Note

OXFORD

GTDB

Genome analysis

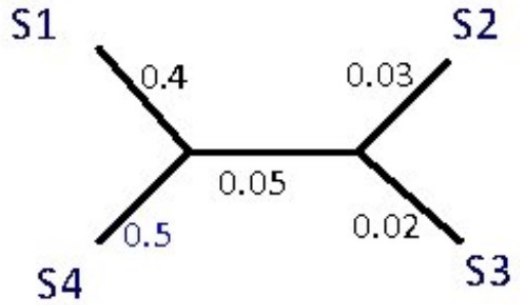
GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database

Pierre-Alain Chaumeil*, Aaron J. Mussig, Philip Hugenholtz and Donovan H. Parks*

pplacer & more classification

- pplacer: phylogenetic (GTDB) tree is given (T), placement of our MAG's concatenated markers (Q1) is searched

```
S1  = -AGGCTATCACCTGACCTCCA-AA
S2  = TAG-CTATCAC--GACCGC--GCA
S3  = TAG-CT-------GACCGC--GCT
S4  = TAC----TCAC--GACCGACAGCT
Q1  = -------T-A--AAAC--------
```



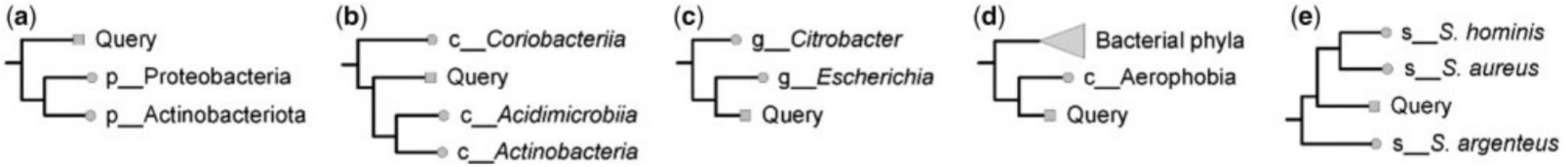
1. For every edge in T, let T_e be the tree created by adding Q1 to that edge. Compute the maximum likelihood (ML) score of the tree T_e for the extended alignment. (Use the ML scores to assign probabilities p(e) to all edges e!)
2. Return T_e that has the best ML score.

GTDB

https://slidetodoc.com/tipp-and-sepp-plus-pasta-tandy-warnow-department/
Matsen, Kodner, Armbrust (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11:538.

classification

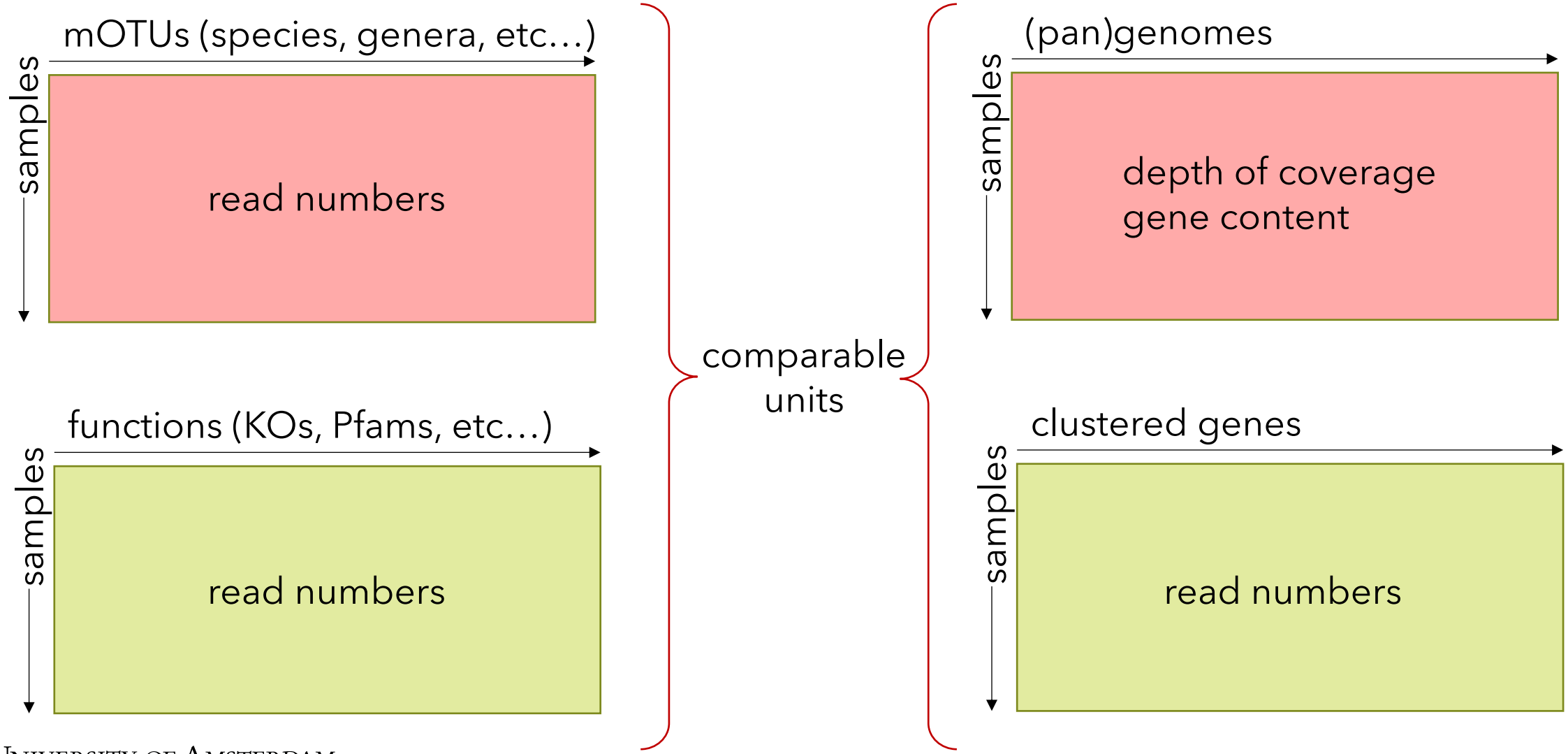
- based on pplacer output
- ambiguities in rank resolved using RED (relative evolutionary divergence)
- genomes that land within a genus are also compared to the species representatives by ANI (average nucleotide identity)



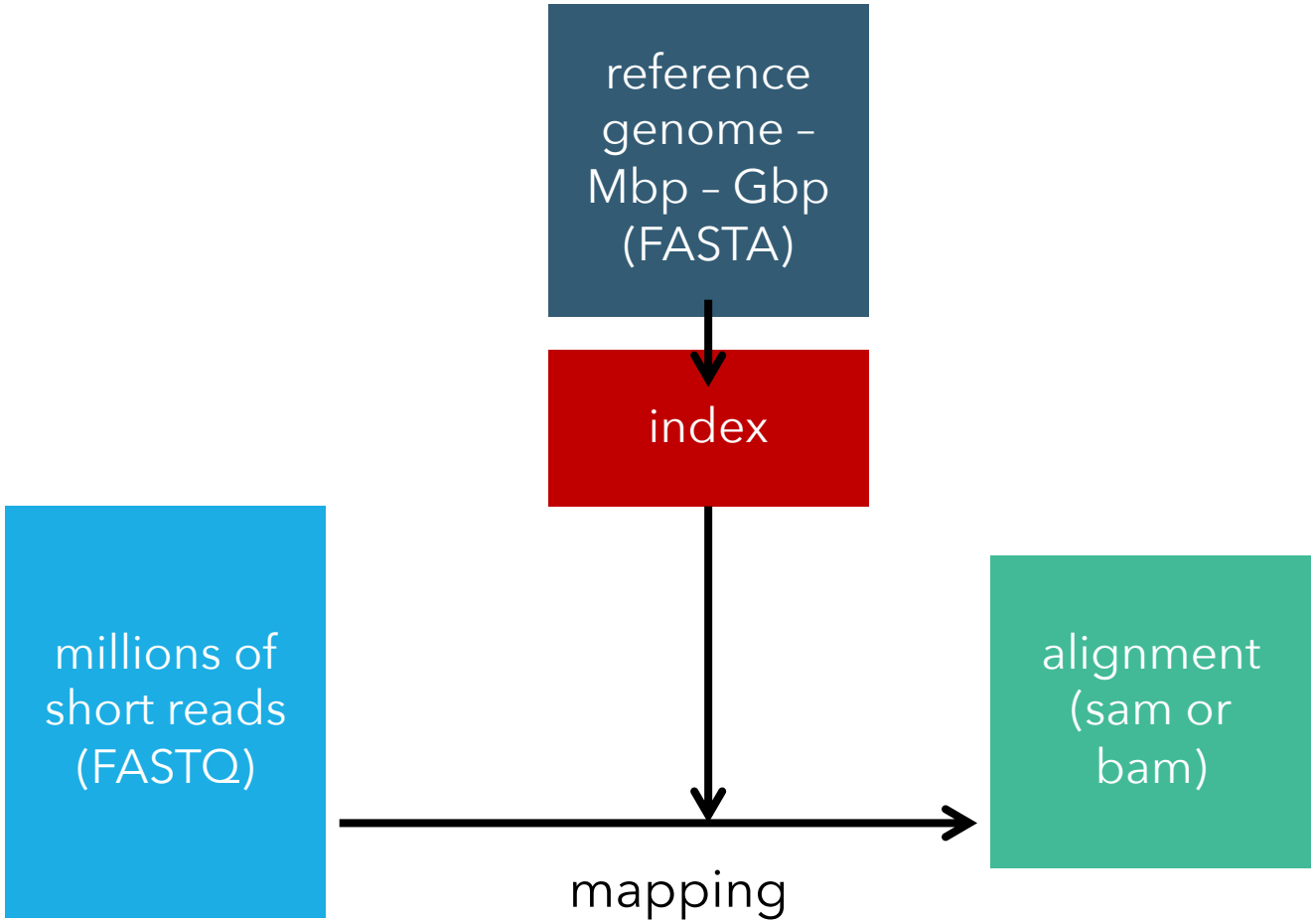
GTDB

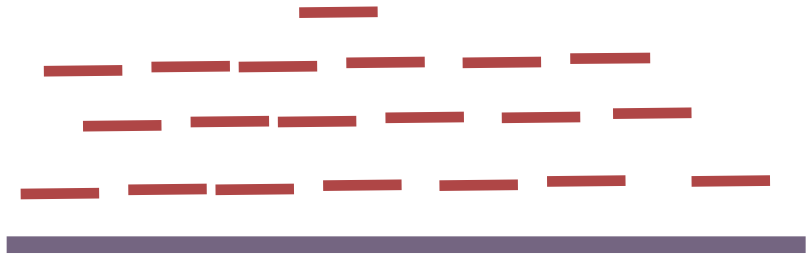
Chaumeil, Mussig, Hugenholtz, Parks (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36(6):1925–7.

Collections to compare samples



Mapping reads





Thanks for your attention!

a.u.s.heintzbuschart@uva.nl

SP C2.205

github.com/a-h-b

twitter.com/_a_h_b_

University of Amsterdam
Life Sciences