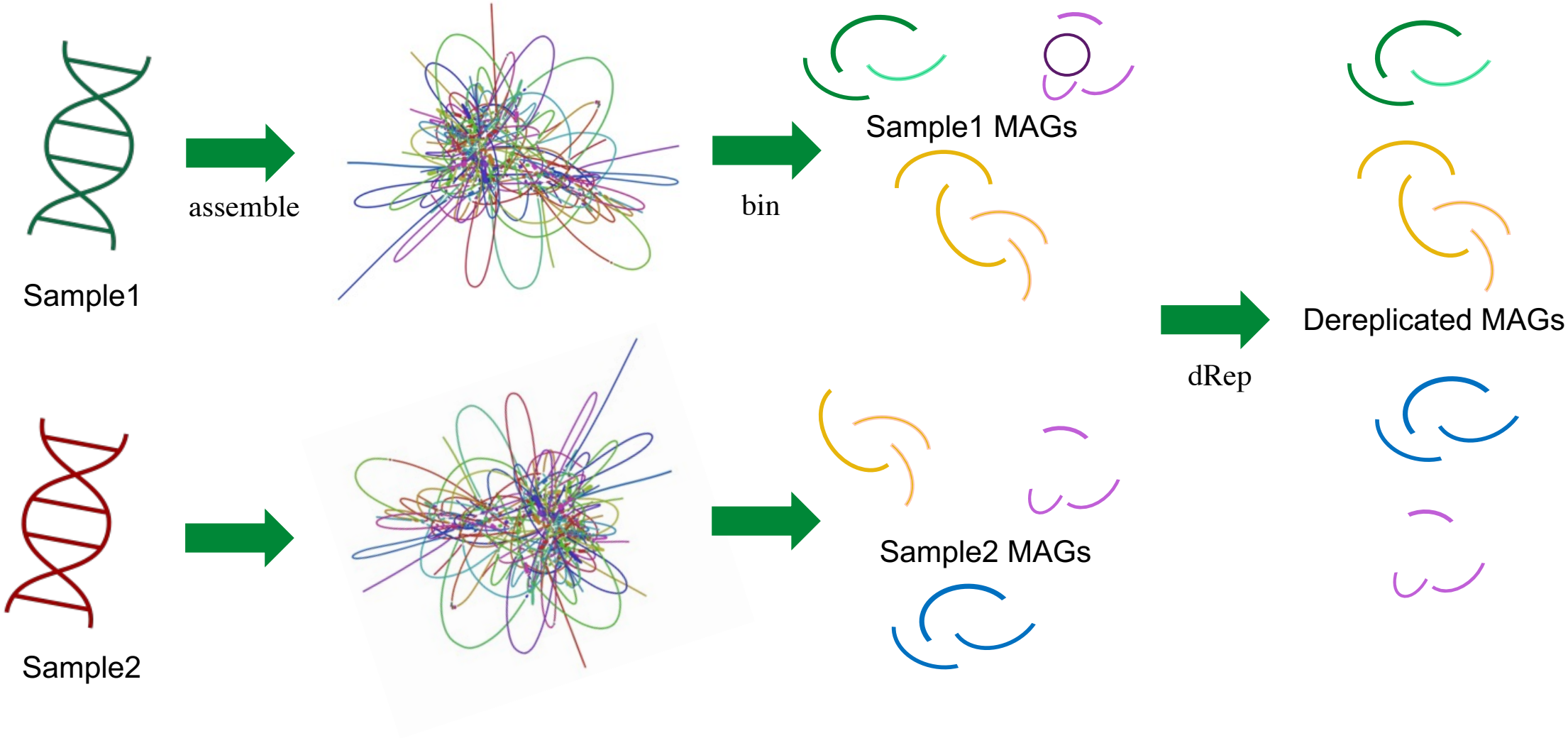
Judit Szarvas, PhD

Taxonomic classification of genomes

Learning objectives

- Define what “taxonomic annotation” is
- List and explain the main steps of the general workflow
- Explain how GTDB’s genome-based taxonomy was created
- Describe what GTDB contains
- Use GTDB-tk for taxonomic annotation
- Interpret the output of GTDB-tk classify_wf

Recap of process of getting MAGs



We have MAGs, what now...?

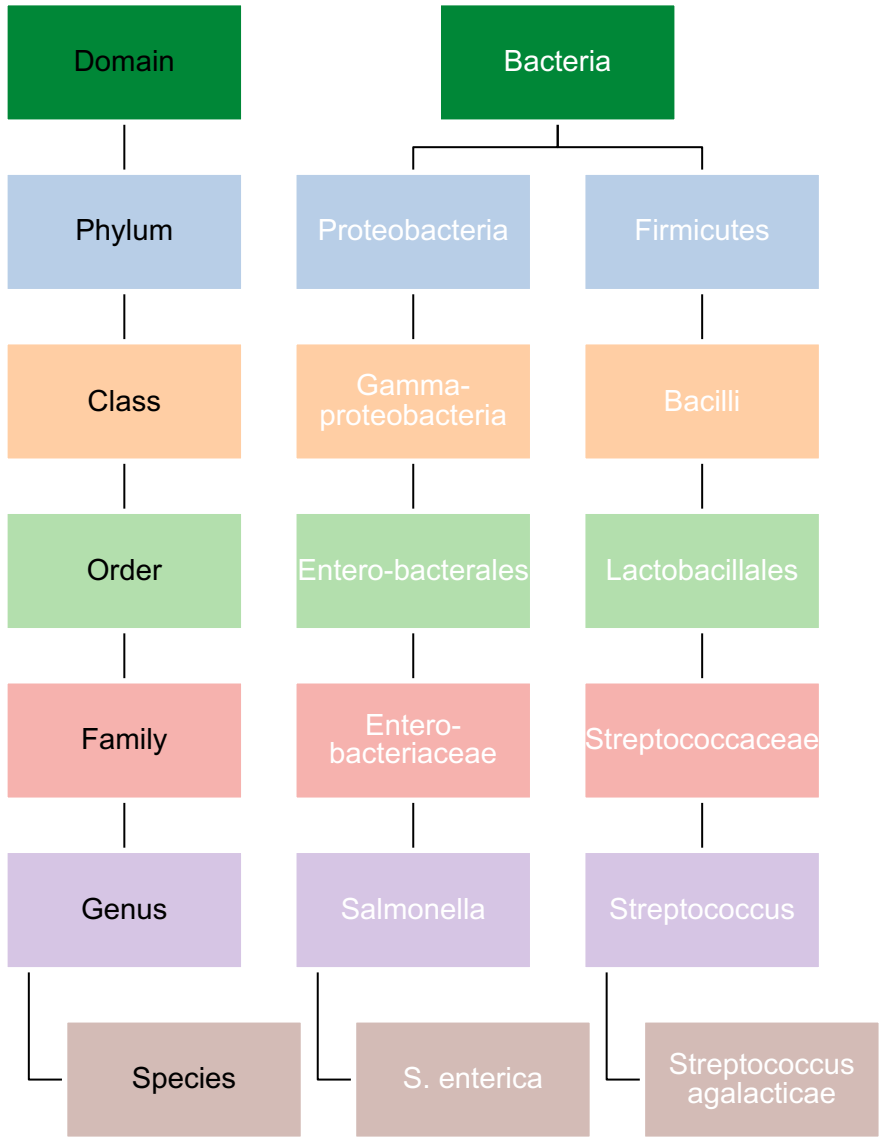
- Annotation of sequences:
  - Open reading frames / coding regions
  - Gene function, metabolic pathway reconstruction
  - Mobile genetic elements

- Use our quality dereplicated MAGs as references for mapping, to estimate abundances
- Taxonomic classification

Taxonomic classification

Predict the taxonomic label of the genome

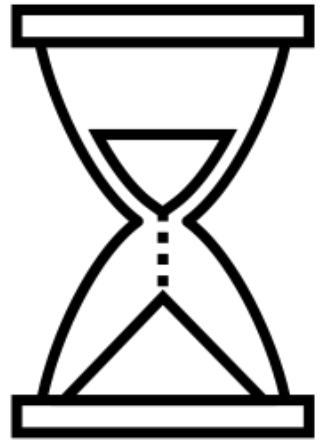
Which tool(s) already introduced in this class could be also used for this purpose?



Exercise

- Run GTDB-Tk classify_wf on dereplicated MQ & HQ MAGs: https://learn.inside.dtu.dk/d2l/le/content/126041/viewContent/526170/View

- 10 mins to submit your jobscripts

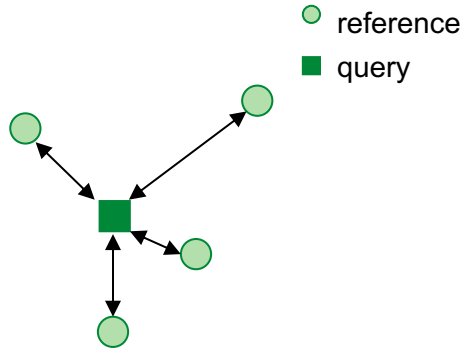


General workflow – Define reference set

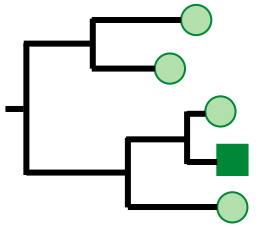
- Take a reference set of labelled sequences
  - should be appropriate for the investigated ecosystem or host
    i.e. don't map soil samples to sequences from marine bacteria, or cow rumen samples to human gut microbiome catalog
  - if individual genes or proteins, they should be present in MAGs
    - rRNA sequences are often missing from MAGs

General workflow – Distance estimation

- Estimate the evolutionary divergence between the query and the references
  - full genomes:
    - estimate average nucleotide identities (ANIs) (hash-based comparison, fx. mash)

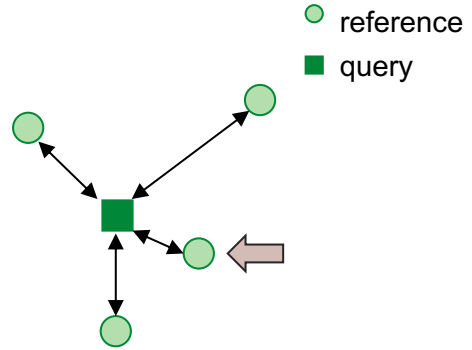


  - single copy marker genes or proteins:
    - phylogenetic placement

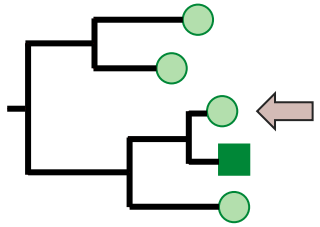


General workflow – Classification

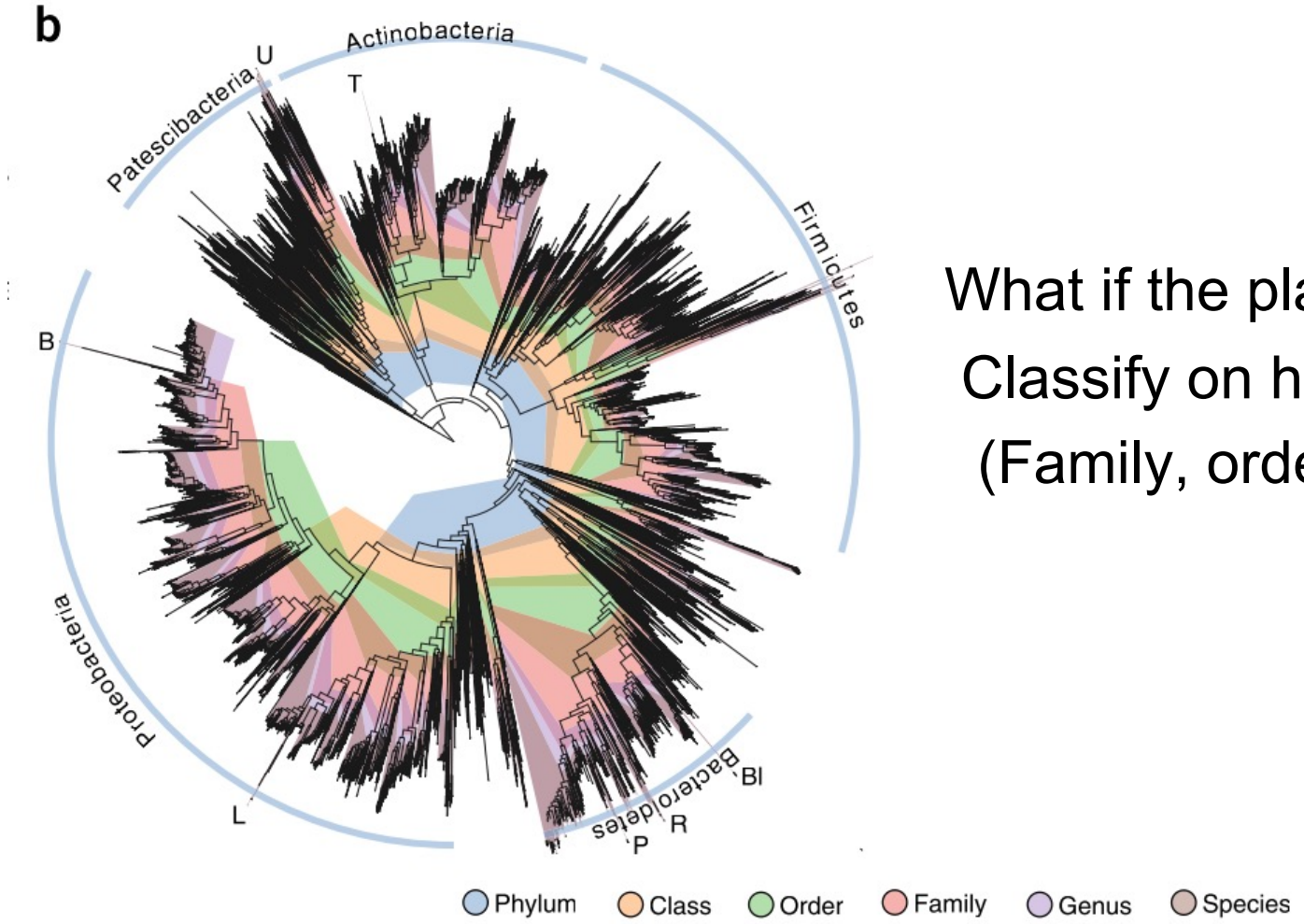
- Classify query based on pre-determined threshold(s)
  - closest reference, if below species ANI threshold



  - in the same clade of phylogenetic tree



General concept – Classification



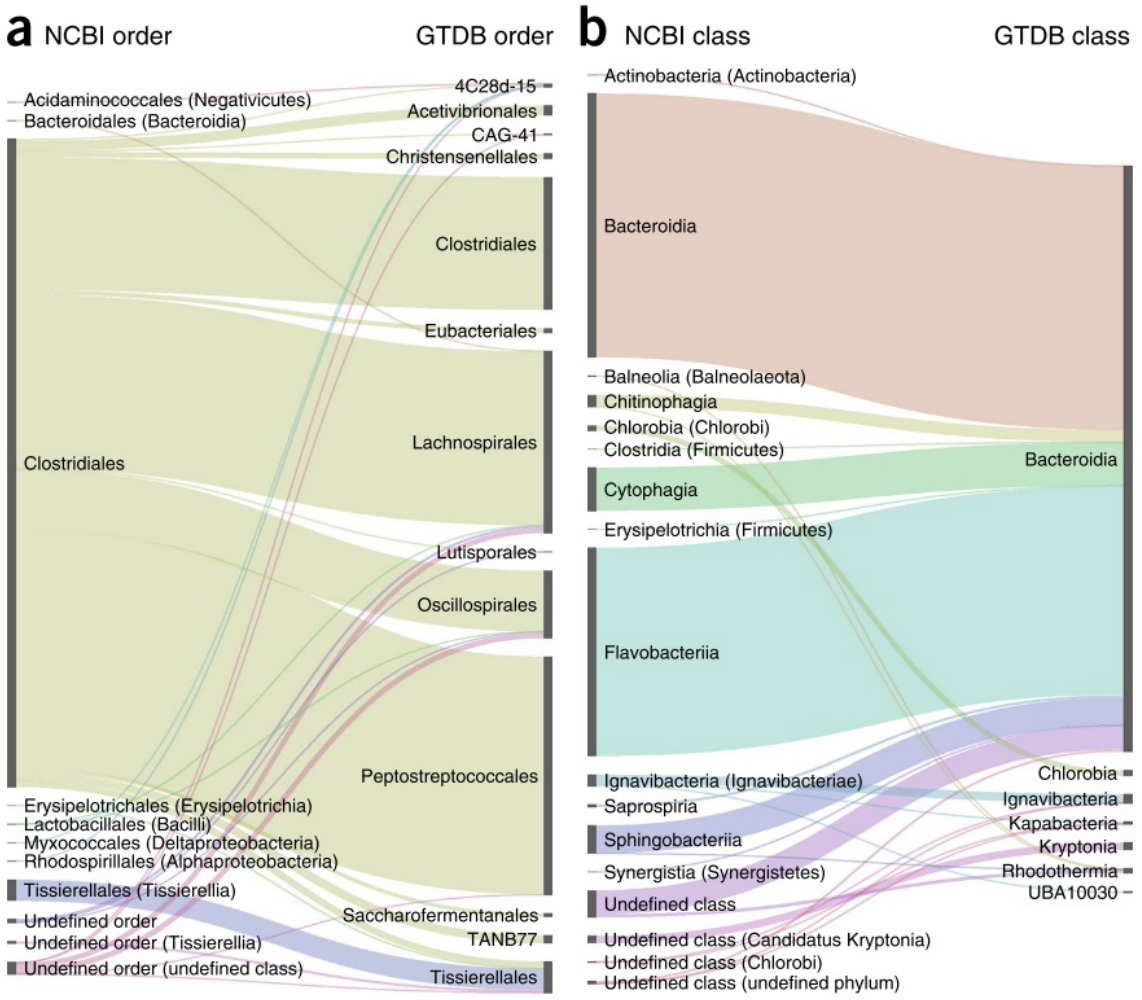
What if the placement uncertain?

Classify on higher taxonomical level (Family, order, etc)

Parks et al, *Nat. Biotechnol.* 2018
doi:10.1038/nbt.4229

Genome Taxonomy DataBase (GTDB)

- https://gtdb.ecogenomic.org/
- General purpose reference database for bacteria and archea
- Genome-based taxonomy:
  - Re-names and normalizes taxonomic ranks
  - Boundaries are described



Parks et al, *Nat. Biotechnol.* 2018
doi:10.1038/nbt.4229

Genome-based taxonomy

- Genomes and MAGs downloaded from NCBI RefSeq/GenBank
- Universal, single-copy proteins
  - present in ≥90%, single-copy in ≥95% of genomes
  - 120 bacterial, 122 (currently 53) archeal
- Aligned, MSA trimmed of too diverse or too sparse columns
- Bootstrapped phylogenetic trees inferred

Parks et al, *Nat. Biotechnol.* 2020
doi:10.1038/s41587-020-0501-8.

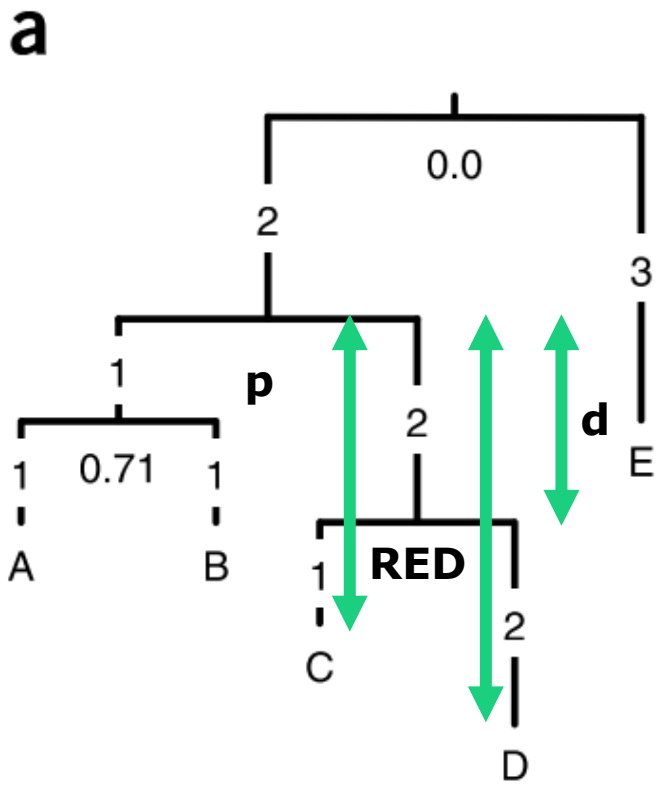
Genome-based taxonomy

- Labels from NCBI Taxonomy or 16S rRNA based novel names
- Manual curation to “split” polyphyletic groups:
  - Group with type material kept original name
  - New groups received alphabetic suffixes, ie. Bacillus_A, Bacillus_B

Parks et al, *Nat. Biotechnol.* 2020
doi:10.1038/s41587-020-0501-8.

Genome-based taxonomy

- Rank normalization: based on relative evolutionary divergence (RED) of internal nodes of the tree

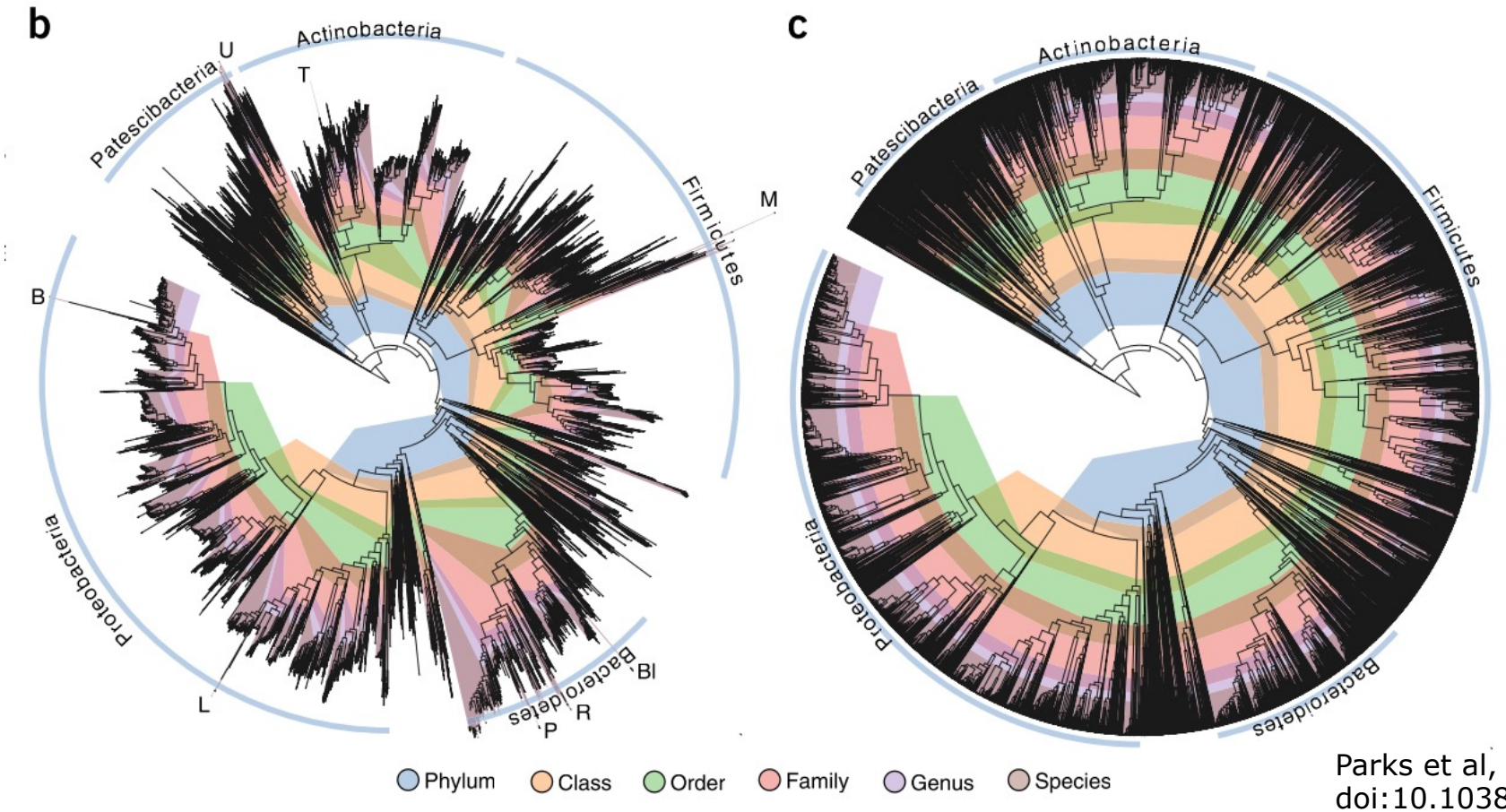


$$RED = p + (d/u) \times (1 - p)$$

Parks et al, *Nat. Biotechnol.* 2018
doi:10.1038/nbt.4229

Genome-based taxonomy

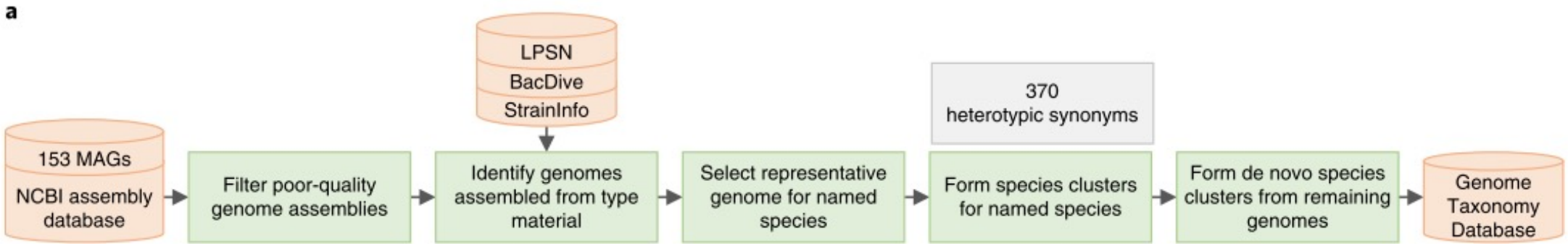
- Rank normalization: based on relative evolutionary divergence (RED) of internal nodes of the tree



Parks et al, *Nat. Biotechnol.* 2018
doi:10.1038/nbt.4229

GTDB construction

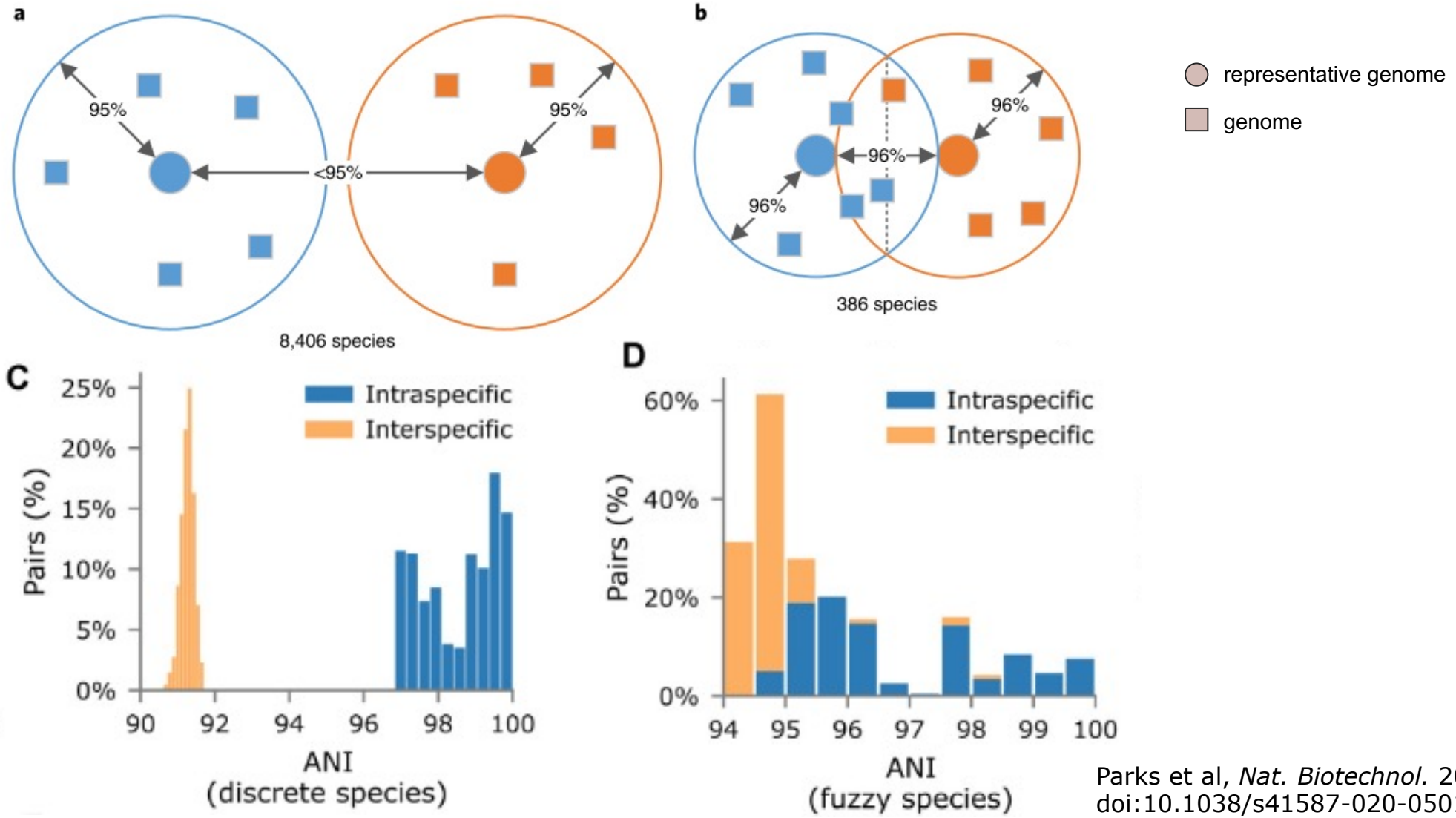
Selecting species representative genomes, centering around type strains



ANI and alignment fraction (AF) based clustering

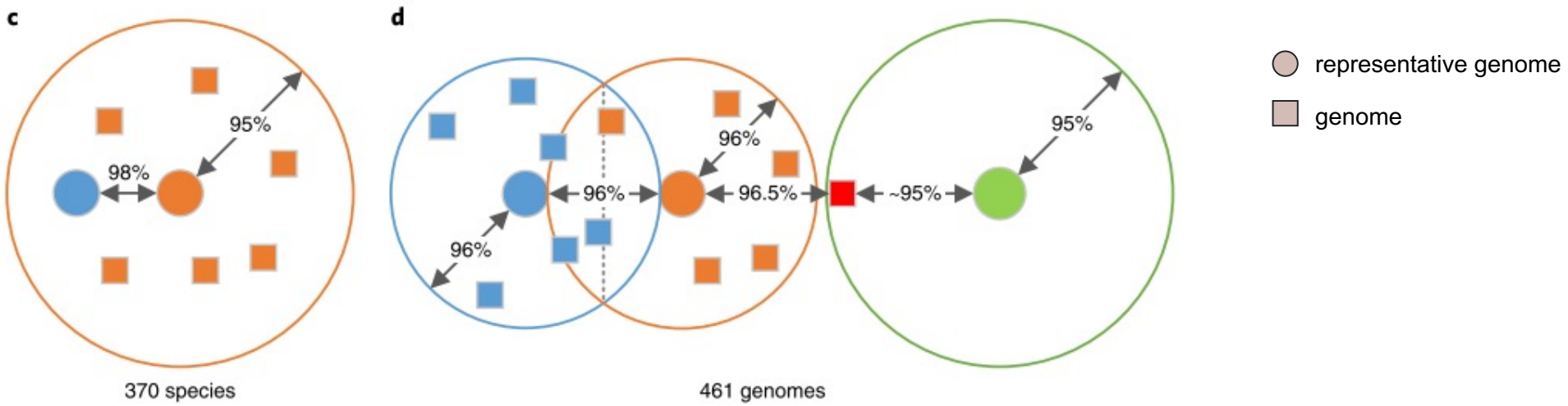
Parks et al, *Nat. Biotechnol.* 2020
doi:10.1038/s41587-020-0501-8

Genome clustering scenarios 1



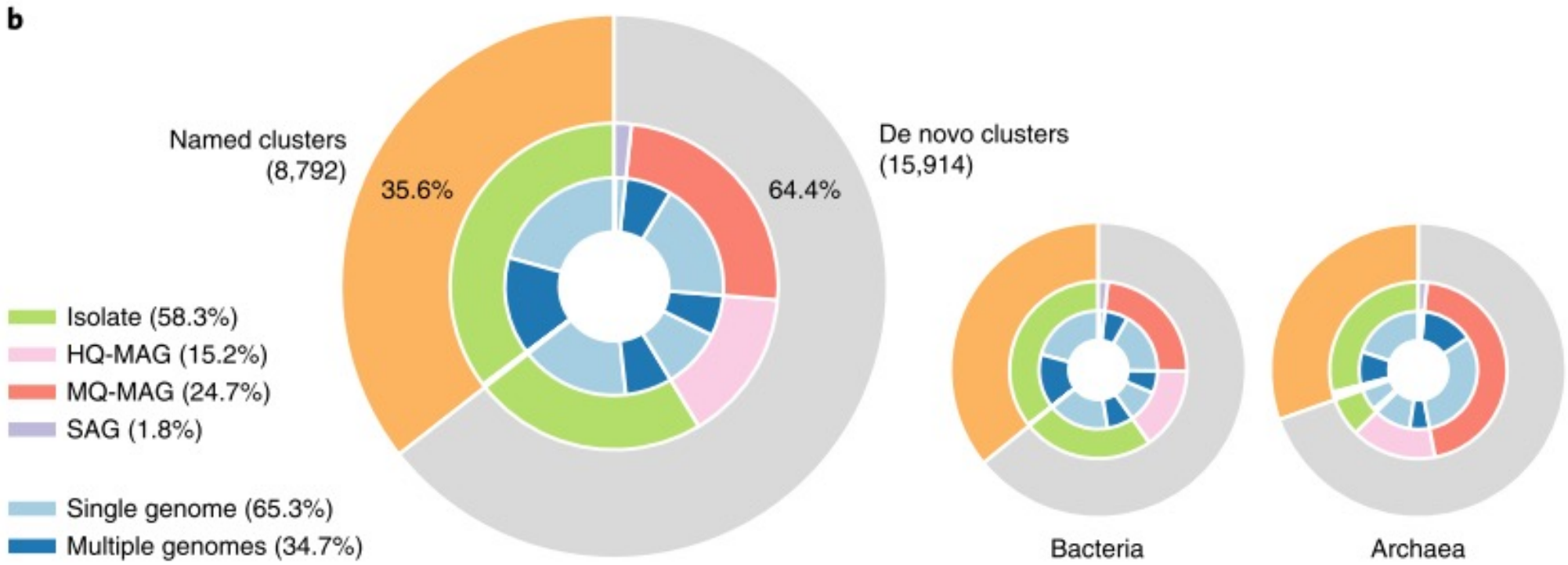
Parks et al, *Nat. Biotechnol.* 2020
doi:10.1038/s41587-020-0501-8

Genome clustering scenarios 2



Parks et al, *Nat. Biotechnol.* 2020
doi:10.1038/s41587-020-0501-8

Genome cluster characteristics



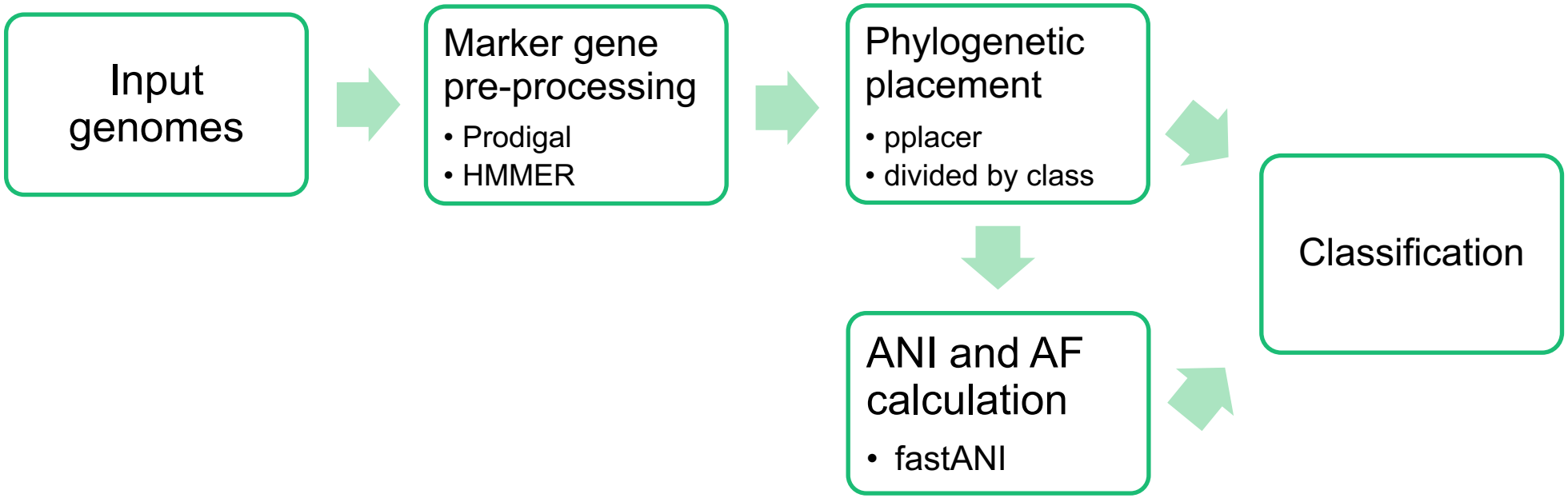
Parks et al, *Nat. Biotechnol.* 2020
doi:10.1038/s41587-020-0501-8

GTDB-tk

- Software for automatic taxonomic classification of genomes
- https://ecogenomics.github.io/GTDBTk
- GTDB based:
  - marker genes
  - species representative genomes
- Classify by:
  - phylogenetic placement and RED
  - ANI and AF

Chaumeil et al, *Bioinformatics.* 2020
doi:10.1093/bioinformatics/btz848

GTDB-tk Classify workflow

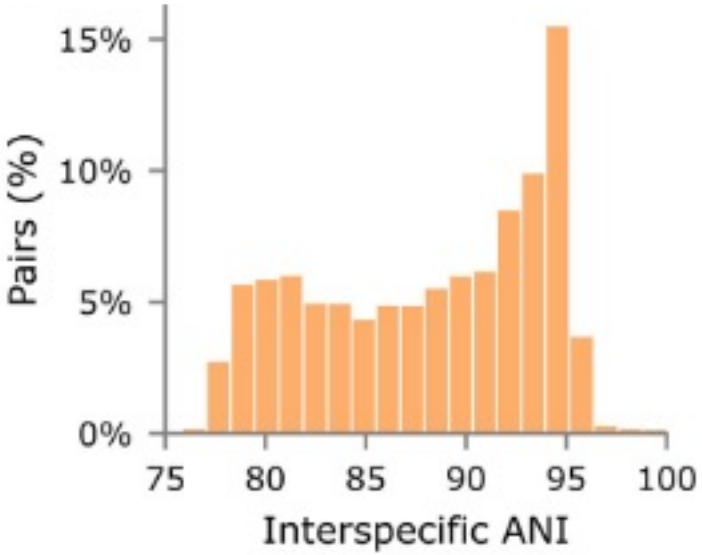


Species classification

- ANI trumps pplacer placement
- Generally, genomes of the same species have >95 ANI%
- GTDB-tk has ANI thresholds for each cluster

gtdbtk.bac120.summary.tsv

fastani_reference	fastani_reference_radius	fastani_taxonomy	fastani_ani	fastani_af
GCF_900102635.1	96.7402	d__Bacteria;p__Proteobacteria;c__Ga	98.52	0.89
GCF_003144325.1	96.4726	d__Bacteria;p__Proteobacteria;c__Ga	98.98	0.95
GCF_011045835.1	96.4726	d__Bacteria;p__Proteobacteria;c__Ga	99.49	0.93
GCF_003044425.1	95.8064	d__Bacteria;p__Proteobacteria;c__Ga	96.91	0.91
GCF_001883995.1	95.3658	d__Bacteria;p__Firmicutes;c__Bacilli;(	95.88	0.81
GCF_000742895.1	95.2518	d__Bacteria;p__Firmicutes;c__Bacilli;(	97.87	0.85
GCF_013267835.1	95.0019	d__Bacteria;p__Proteobacteria;c__Ga	96.24	0.89
GCF_000194945.1	95.0	d__Bacteria;p__Firmicutes;c__Bacilli;(	95.67	0.94



Closest genome in genus

Parks et al, *Nucleic Acids Research* 2022
doi:10.1093/nar/gkab776

Species classification

- Summary output notes method of classification:

`gtdbtk.bac120.summary.tsv`

classification	classification_method	note
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Enterobacteriaceae;g__Providencia;s__	taxonomic classification defined by topology and ANI	classification based on placement in class-level tree
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas_E;s__Pseudomonas_E chengduensis	taxonomic classification defined by topology and ANI	topological placement and ANI have congruent species assignments
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Neisseriaceae;g__Neisseria;s__Neisseria sicca_C	ANI	topological placement and ANI have incongruent species assignments

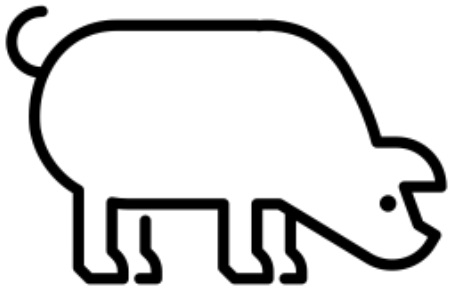
Manual curation of results is recommended!

- For novel organisms, the *de novo* workflow is needed, where a maximum likelihood tree is inferred on the protein alignment, giving more precise results.

Exercise

- Look at GTDB-Tk classify_wf results

https://learn.inside.dtu.dk/d2l/le/content/126041/viewContent/526170/View




DTU