EBAii Assemblage & Annotation

Part 2: construction and analysis of procaryotic genomic dataset

H. Chiapello & V. Loux

Helene.chiapello@inrae.fr
https://orcid.org/0000-0001-5102-0632

Valentin.loux@inrae.fr
https://orcid.org/0000-0002-8268-915X

2. Construction and analysis of procaryotic genomic dataset

Outline

> 2.1 Downloading a dataset of public genomes

> 2.2 Analyzing the genome dataset

> 2.3 Comparing and dereplicating the dataset

Many slides from the "***Bioinformatique par la pratique***" migale training cycle "Comparison of microbial genomes" module

https://migale.inrae.fr/trainings

And thanks to Guillaume Gautreau for his help

Hélène Chiapello
Training

Valentin Loux
Technical coordinator

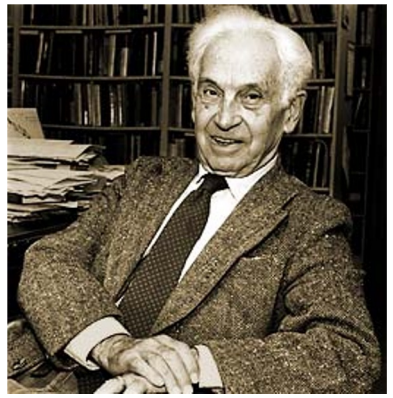
2.2 Analyzing a genome dataset

Why?

- Frequent problems in genome analysis and comparison
  - Heterogenous quality of sequencing and assembly
  - Presence of huge number or public genomes OR absence of any close genomes of the same species in public databases
  - Difficulties regarding microbial taxonomy (classification) and nomenclature (naming of genus, species and strain naming) for many non-model organisms
- Outline
  - 2.2.1 Introduction
  - 2.2.2 Dataset diversity analysis
  - 2.2.3 Dataset quality analysis

2.2.1 Introduction

- What is a species?



Ernst Mayr (1942) :

“Species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups”

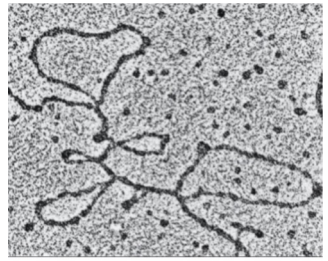
⇨ ***Not relevant for bacteria***

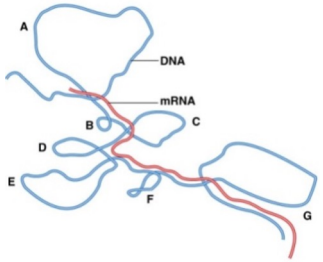
What is a bacterial species?

No consensual definition for procaryotes

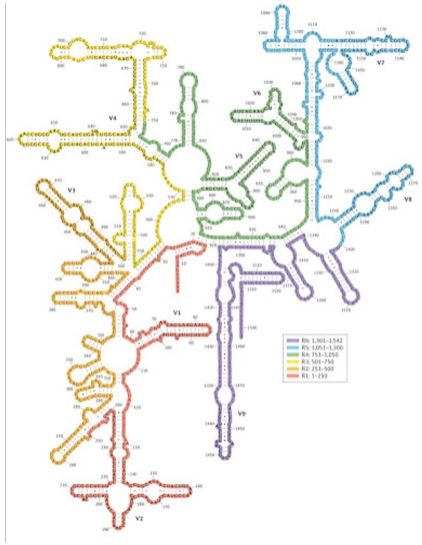
- No universal criteria
- Several approaches used to classify bacterial
  - Phenotypes and morphological criteria
  - DNA-DNA hybridization
- Universal markers
  - 16S rRNA
  - MLST (Multi Locus Sequence Typing)
- Genomic-based taxonomy are now becoming a gold-standard

% ADN-ADN hybridation	>70%





% rRNA 16S identity	>98,7%

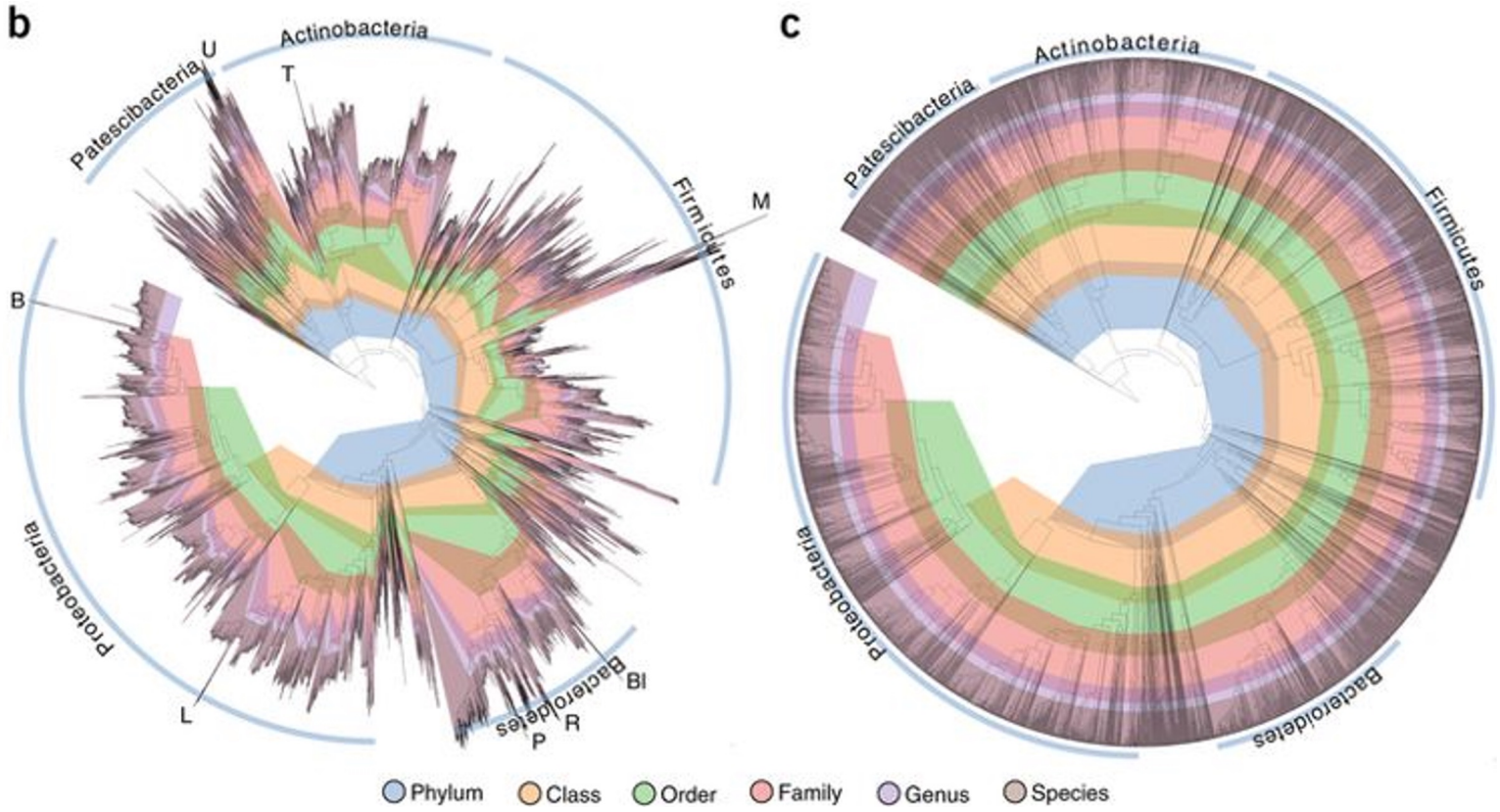


Example: the Genome-based taxonomy for prokaryotic genomes

- Objective: a standardized microbial taxonomy based on genome phylogeny
- Taxonomy inferred from concatenated single copy marker proteins

Parks et al. 2018, 2021

https://gtdb.ecogenomic.org/



2.2.2 Evaluating genome diversity in a dataset

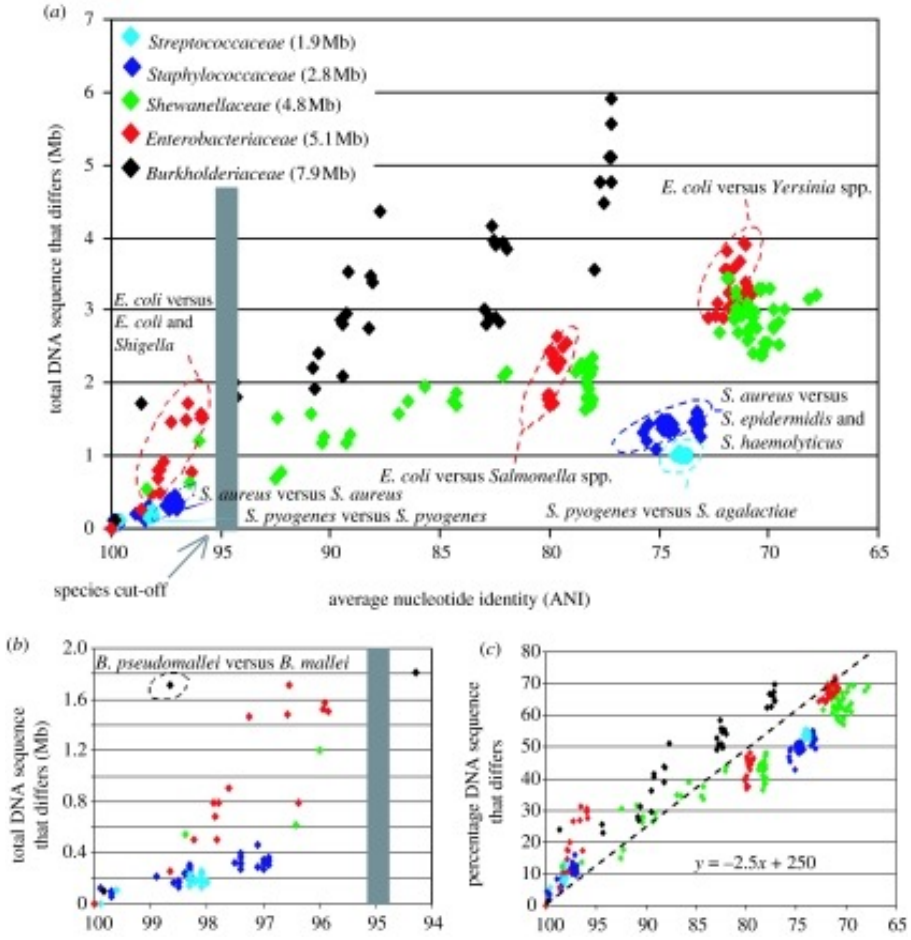
- Why ?
  - Identify outlier genomes
  - Identify groups of (very) similar genomes and de-replicate datasets
  - Estimate genome similarity in a dataset and design an adapted comparative strategy

How ?

- Alignment based approaches (ANI)
- k-mer based approaches (MASH)

Average Nucleotide Identity (ANI)

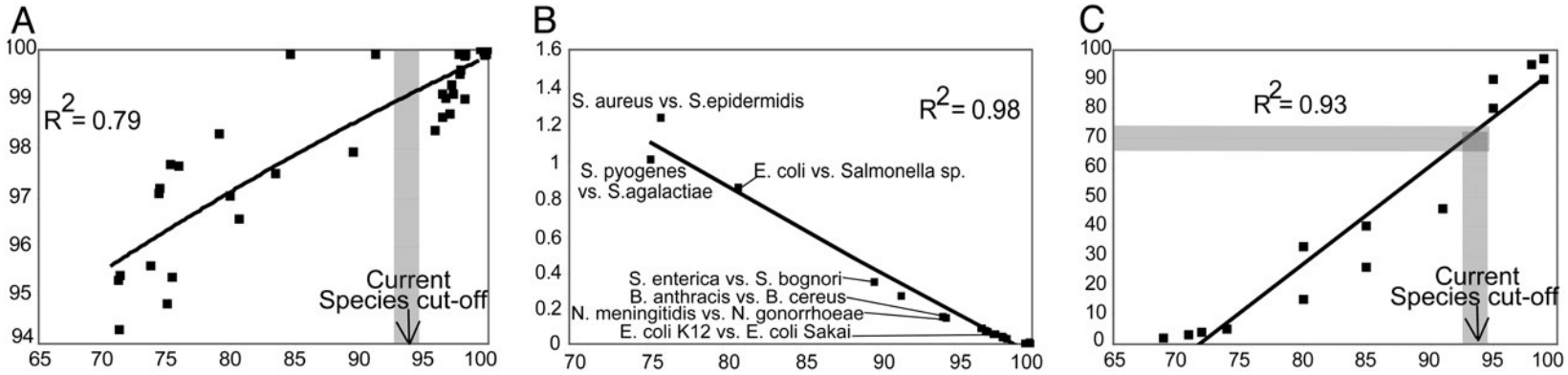
- Meet the need for a robust measure of genomic relatedness and a systematic and scalable species assignation technique
- Mean identity percent of aligned regions of a pair of genomes
- Rely on pairwise alignments from
  - aligned core genes
  - genomic alignments
- Can easily be used to build phylogenetics tree using distance methods
- Is implemented in several bioinformatics tools: ANIn (nucmer based, Richter 2009) gANI (coding regions, Varghese 2015),…



Genetic diversity within five important bacterial groups. Konstantidinis et al. 2006. The bacterial species definition in the genomic era
DOI: 10.1098/rstb.2006.1920

Average Nucleotide Identity (ANI)

- ANI strongly correlates (R = 0.79 for logarithmic correlation) with the 16S rRNA gene sequence identity and can resolve areas where the 16S rRNA gene is inadequate (intra-species level)
- The average rate of synonymous substitutions shows a tight correspondence to ANI, suggesting that ANI may also be a useful descriptor of the evolutionary distance
- ANI shows a strong linear correlation to DNA–DNA reassociation values, and the 70% DNA–DNA reassociation standard corresponds to ≈93–94% ANI i.e. strains that show >94% ANI should belong to the same species

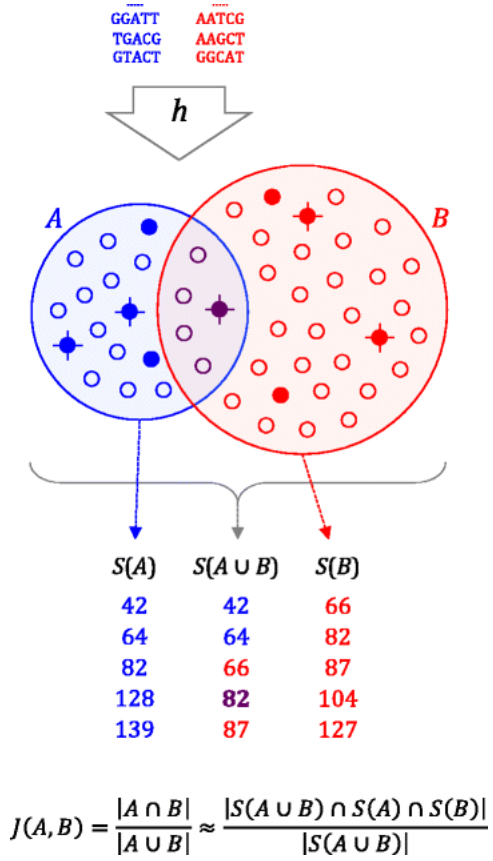


ANI vs 16 rRNA %id | ANI vs average syn. Mutation rate | ANI vs DNA-DNA reassociation value

Konstantidinis et al. 2005. Genomic insights that advance the species definition for prokaryotes https://doi.org/10.1073/pnas.0409727102

MASH: fast (meta)genome distance estimation using MinHash

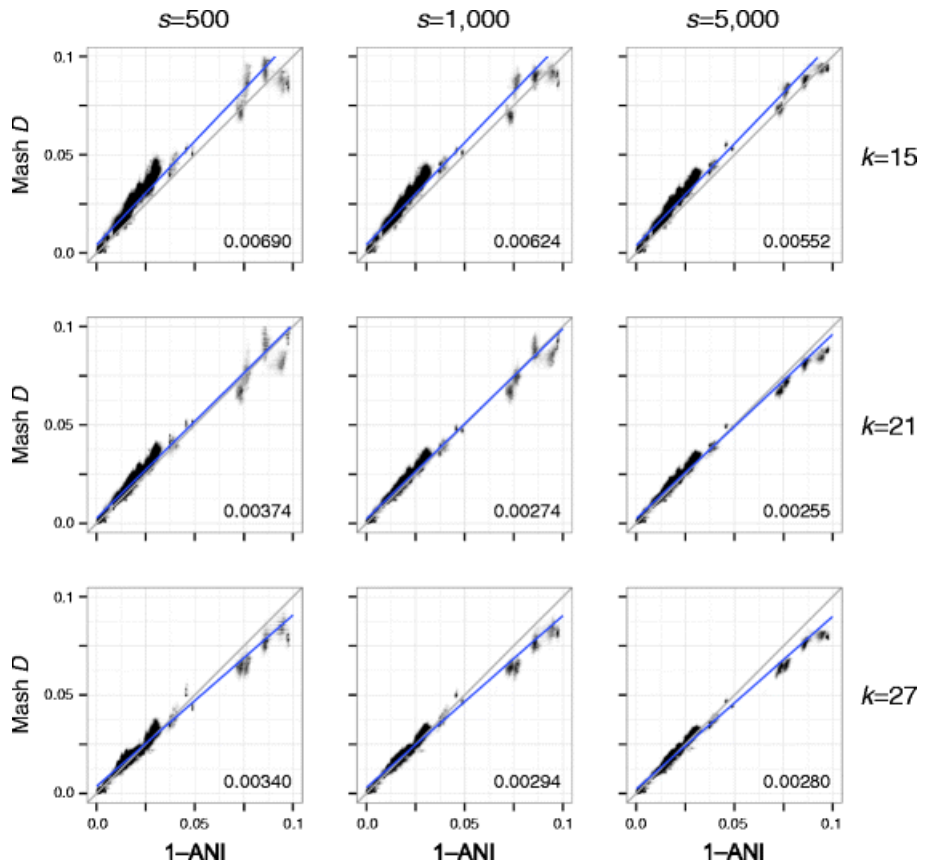
- Mash allows to compute a pairwise mutation distance without alignment using k-mer counts
- Mash provides two basic functions for sequence comparisons:
  - **sketch**: converts a sequence or collection of sequences into a MinHash sketch
  - **dist**: compares two sketches and returns an estimate of the Jaccard index (i.e. the fraction of shared k- mers), a P value, and the Mash distance



Ondov, B.D., Treangen, T.J., Melsted, P. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

MASH distances correlate well with ANI

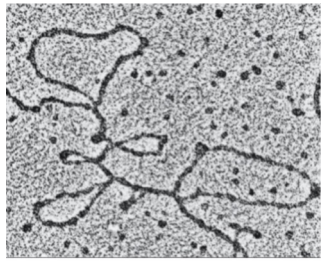
- Dataset: 500 complete E. coli genomes
  - Each plot column shows a different sketch size
  - Each plot row a different k-mer size k.
  - Gray lines: model relationship D = 1–ANI
- Increasing the sketch size improves the accuracy of the MASH distance, especially for more divergent sequences.
- Limit on how well the MASH distance can approximate ANI, especially for more divergent genomes (e.g. ANI considers only the core genome)



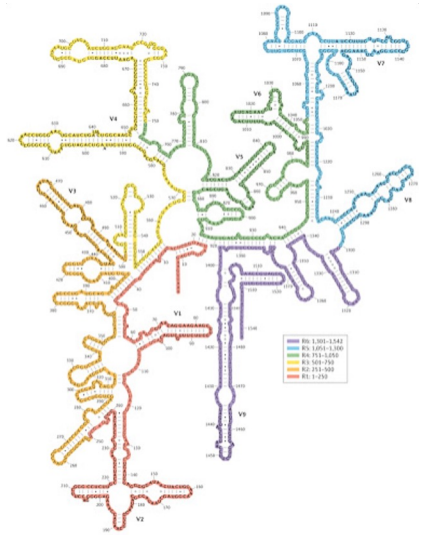
Ondov, B.D., Treangen, T.J., Melsted, P. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17, 132 (2016). https://doi.org/10.1186/s13059-016-0997-x

Back to procaryote taxonomy

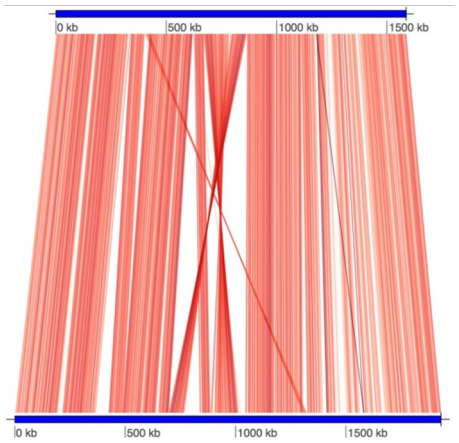
% ADN-ADN hybridation	>70%



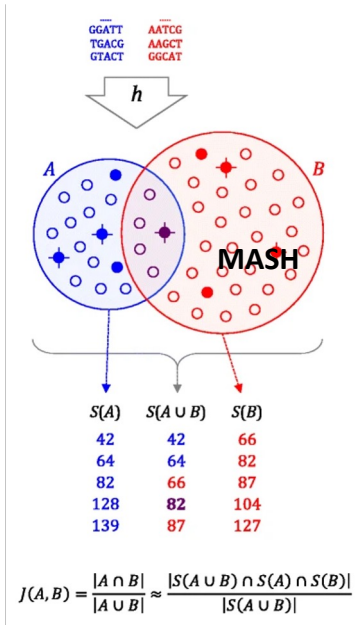
% rRNA 16S identity	>98,7%



% genome identity (ANI)	>94%



K-mer distance	<0,06



2.2.3 Quality analysis

- Already presented: evaluate Quality using the 3Cs

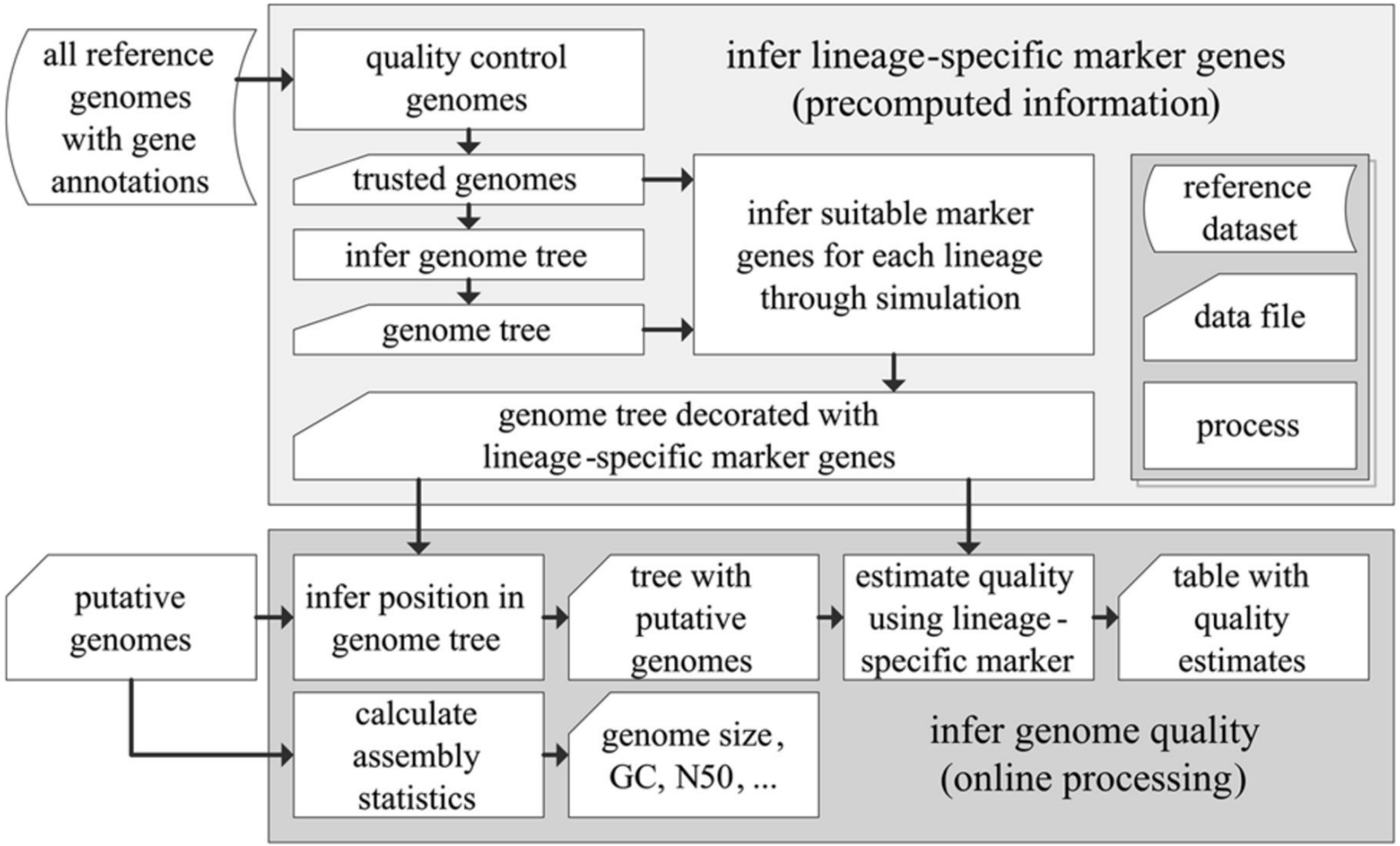
1. **Contiguity**. Produce the longest possible contigs.
2. **Correctness**. Assemble contigs with few/no errors.
3. **Completeness**. Cover the entire original sequence and minimize missing regions

- An additional key point for microbes: evaluate **Contamination**
  - From genomic fragments of divergent taxa
  - From genomic fragments of multiple strains (i.e. strain **heterogeneity**)

CheckM

- a set of tools for assessing the quality of genomes recovered from isolates, single cells, or metagenomes
- provides robust estimates of genome **completeness** and **contamination**
  - use collocated sets of genes that are ubiquitous and single-copy within a phylogenetic lineage
  - propose a fixed vocabulary for defining genome quality based on estimates of completeness and contamination
- Evaluate by simulations the accuracy of quality estimates

CheckM consists of a workflow for precomputing lineage-specific marker genes for each branch within a reference genome tree (top box) and an online workflow for inferring the quality of putative genomes (bottom box).



CheckM relies on several other tools and data

- *prodigal* to predict genes
- A reference genome tree based on 43 phylogenetically informative marker genes and 5656 trusted reference genomes
  - Marker genes are identified in assemblies using HMMER
  - The resulting genes are used to placed the genome into the tree using *pplacer*
- Lineage-specific marker sets determined for all nodes within the reference genome tree by identifying single-copy genes present in ≥97% of all descendant genomes.

CheckM report

Provides classic quality metrics and plots, including:

- Results of binning

  >Marker lineage, #genomes, #markers, #marker sets

- CheckM metrics

  > Completeness, Contamination, Strain heterogeneity

- Classical Quality metrics

> #ambiguous bases, #scaffolds, #contigs, N50 (scaffolds), N50 (contigs), Mean scaffold length (bp),Mean contig length (bp), Longest scaffold (bp),Longest contig (bp), GC, GC std (scaffolds > 1kbp)

CheckM report – binning part

Marker lineage: indicates the taxonomic rank of the lineage-specific marker set used to estimated genome completeness, contamination, and strain heterogeneity.

#genomes: number of reference genomes used to infer the lineage-specific marker set

#markers: number of marker genes within the inferred lineage-specific marker set

#marker sets: number of co-located marker sets within the inferred lineage-specific marker set

0-5+: number of times each marker gene is identified

CheckM report

- Completeness: estimated completeness of genome as determined from the presence/absence of marker genes and the expected colocalization of these genes
- Contamination: estimated contamination of genome as determined by the presence of multi-copy marker
- Strain heterogeneity: % determined from the number of multi-copy marker pairs which exceed a specified **amino acid identity threshold** (default = 90%).
  - High strain heterogeneity suggests the majority of reported contamination is from one or more closely related organisms (i.e. potentially the same species),
  - Low strain heterogeneity suggests the majority of contamination is from more phylogenetically diverse sources

CheckM: proposed genome quality classification scheme

- ***Finished genomes***: genomes assembled into a single contiguous sequence containing no gaps or ambiguities and where extensive efforts have been made to identify errors
- ***Noncontiguous finished:*** *g*enomes assembled into multiple sequences as a result of repetitive regions, but otherwise of a finished quality
- ***Draft genomes:*** all other genomes

Table 3. Controlled vocabulary of draft genome quality based on estimated genome completeness and contamination

Completeness	Classification	Contamination	Classification
≥90%	Near	≤5%	Low*
≥70% to 90%	Substantial	5% to ≤10%	Medium
≥50% to 70%	Moderate	10% to ≤15%	High
<50%	Partial	>15%	Very high

(*) Genomes estimated to have 0% contamination can be designated as having "no detectable contamination".

Donovan H. Parks et al. Genome Res. 2015;25:1043-1055

CheckM result interpretation limits

- CheckM is dedicated to eubacterial and archeal genomes
  - Eukaryotic or phage genomes will be reported as highly incomplete
  - The quality of plasmids must also be assessed independently of CheckM
- The novelty of a genome will also influence the accuracy of CheckM estimates
  - Estimates for bacterial and archaeal genomes from deep basal lineages with few reference genomes are generally based on domain-level marker sets
  - Quality estimates may be not reliable for genomes of novel lineages
  - Gene loss or duplication may be an issue

Conclusion : use CheckM as a tool to detect outliers and further investigate!

Questions ?