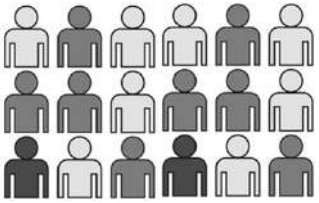
Microbial genomics lecture:
Use of next-generation (genome) sequencing in clinical microbiology

Rasmus Marvig, PhD
Bioinformatician - Head of Microbial Genomics, Department of Clinical Biochemistry, Rigshospitalet
Associate Professor, Department of Health Technology, Technical University of Denmark

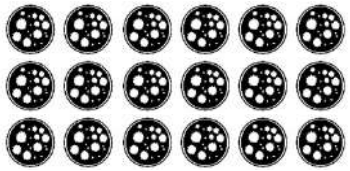
Conflicts of interest - secondary occupations:
Technical Assessor, DANAK - The Danish Accreditation Fund
Strategic Advisory Board Member, Freya Biosciences

Bacterial WGS powers diagnostics and research at Rigshospitalet (54,599 genomes sequenced per 2025-10-31)

Patients

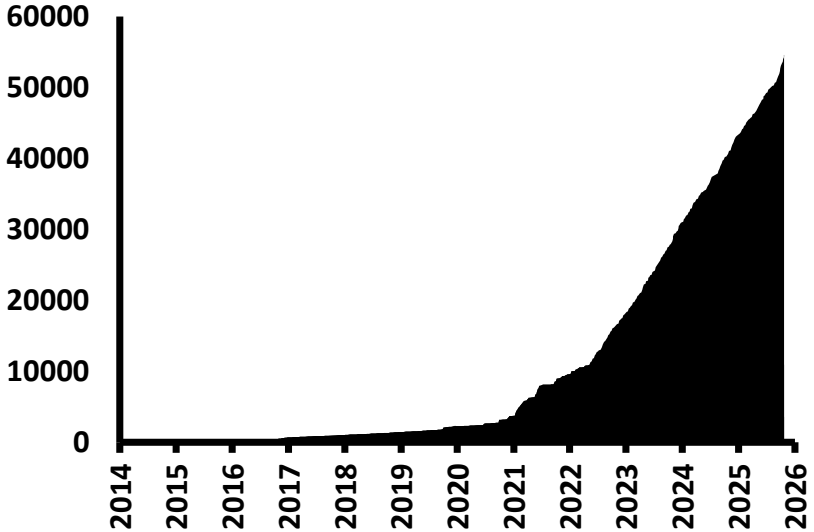


Bacterial isolates

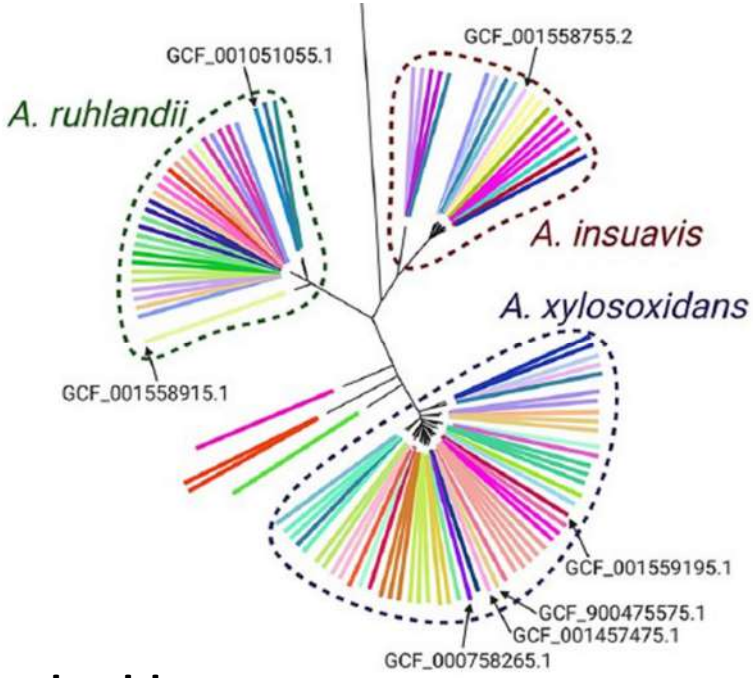


Genomic epidemiology

Early implementation of large-scale genomics.
Below: Accumulated number genomes since year 2015



Genomics superior to define bacterial species. Example below: Genetically distinct species hidden to standard test



Research shows positive impact of genomics on healthcare

One method to rule them all: Correct diagnostic typing with a single method (see *Gabrielaite et al., JCM, 2021*)

Earlier detection of persistent infections (see *Bartell et al., ERJ, 2021*)

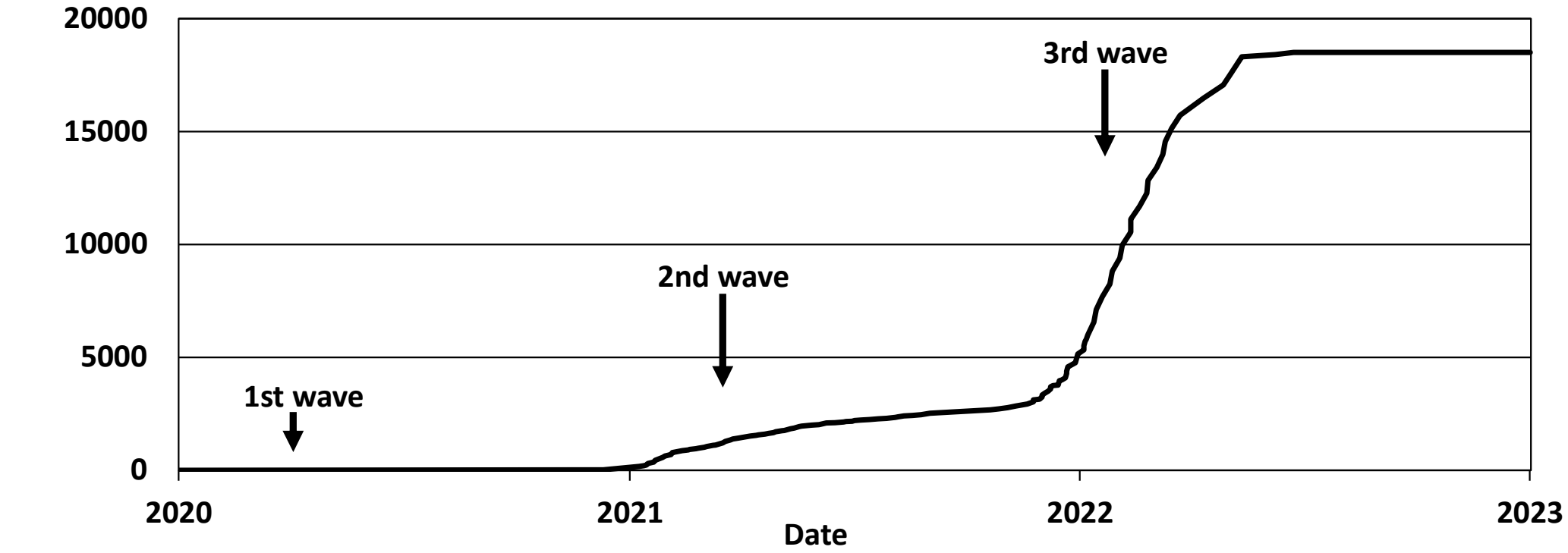
Detection of within-hospital transmission without need for prior suspicion (see *Eklöf et al., CID, 2022*)

Genomic detection of resistance not captured by standard tests (see *Misiakou et al., Microbial Genomics, 2023*)

Genomic epidemiology can generate annual savings of €1.25 million at Rigshositalet (see *Hertz et al., medRxiv, 2025*)

18,498 SARS-CoV-2 genomes sequenced at Rigshospitalet during pandemic waves in years 2020-2022

Jørgensen et al., Journal of Virological Methods, 2023: *WGS validation of a Sanger method*
Michaelsen et al., Genome Medicine, 2022: *Alpha B.1.1.7 transmission on national level*
Fonager et al., Eurosurveillance, 2022: *Omicron BA.2 transmission on national level*
Lyngse et al., Nature Communications, 2021: *Alpha B.1.1.7 household transmissibility*
Andersen et al., Danish Medical Journal, 2021: *B.1.1.29 transmission in elderly care home*



3,785 HIV-1 genomes sequenced in research project funded NIH (USA)

Rigshospitalet ALARM-INFO SØG MENU

Forside / Nyheder / Intranetnyheder

4000 prøver sendes fra USA til Rigshospitalet for specialanalyse

Med et tilbud om en hidtil uset præcision i typebestemmelse af hiv-virus har et amerikansk ledet internationalt forskningsprojekt valgt Enhed for Genomisk Medicin som samarbejdspartner til at analysere godt 4.000 prøver fra hiv-smittede patienter.

3. marts 2016, kl 11:30

March 2016

Blanquart et al., Virus Evolution, 2025: *GWAS*
Cozzi-Lepri et al., HIV Medicine, 2024: *Drug resistance*
Gabrielaite et al., PLOS Computational Biology, 2023: ***GWAS***
Bennedbæk et al., Virus Evolution, 2021: ***Transmission***
Gabrielaite et al., Journal of Infectious Diseases, 2021: ***GWAS***
Baxter et al., HIV Medicine, 2021: *Drug resistance*

December 2024

There is a growing concern over infections as a threat to human health and welfare [1,2].

[1] Lancet. 2020. doi: 10.1016/S0140-6736(20)30925-9.

[2] Lancet. 2022. doi: 10.1016/S0140-6736(21)02724-0.

The Novo Nordisk Foundation [3], the Bill & Melinda Gates Foundation [4], and the The Wellcome Trust [5] all call to reduce the burden of infectious diseases.

[3] Novo Nordisk Foundation. Decreasing the burden and threat of infectious diseases. [cited 10 Sep 2024]. Available: https://novonordiskfonden.dk/en/what-we-support/health/decreasing-the-burden-and-threat-of-infectious-diseases/

[4] Bill & Melinda Gates Foundation. Program strategies. [cited 10 Sep 2024]. Available: https://www.gatesfoundation.org/our-work#program_strategies

[5] Wellcome Trust. Infectious Disease. [cited 10 Sep 2024]. Available: https://wellcome.org/what-we-do/infectious-disease

From https://novonordiskfonden.dk/nyheder

You will use 5 different tools to analyse microbial genome sequences: Tools are commonly used for microbial diagnostics and research

(11,807 citations) Kraken

(6,552 citations) GTDB-Tk

(unpublished) mlst

(1,866 citations) Parsnp

(unpublished) Abricate

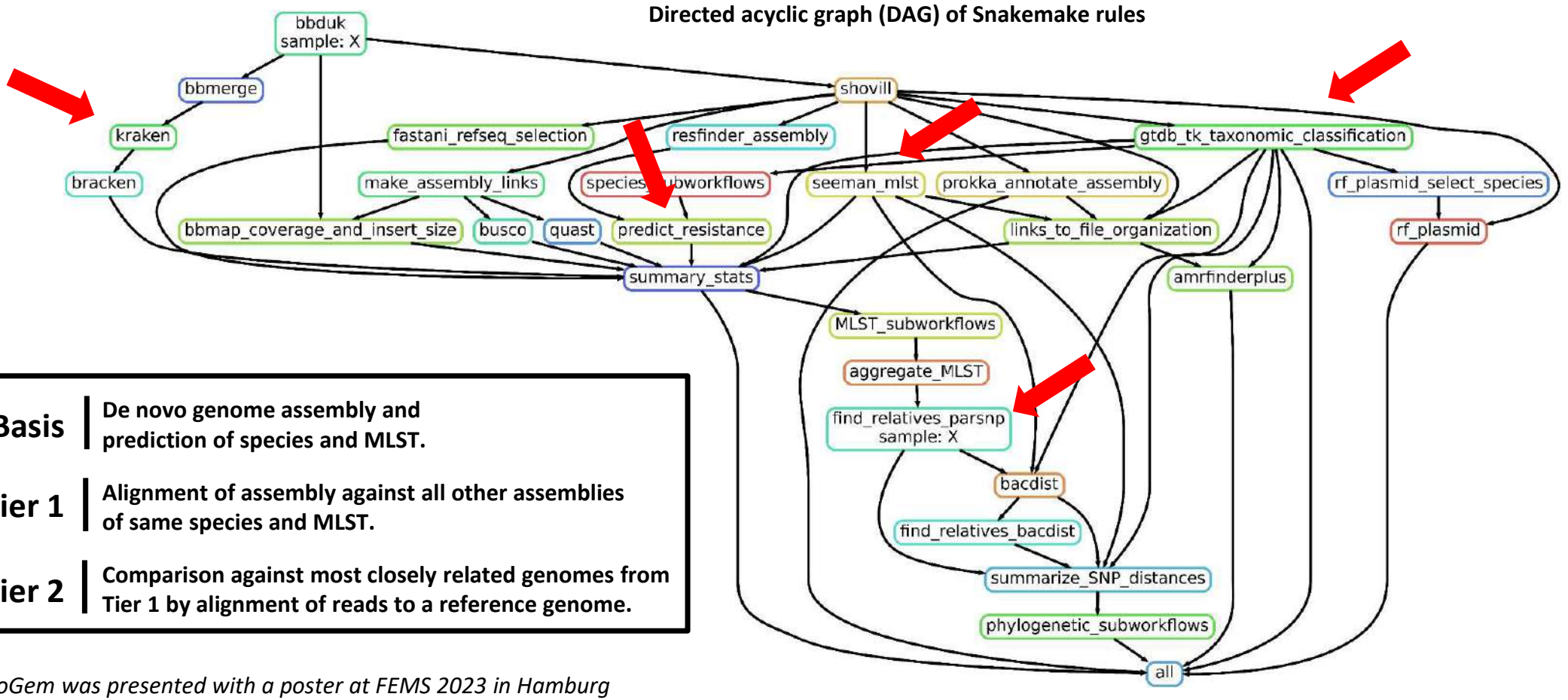
Microbial genomics exercise

Contents [hide]

https://teaching.healthtech.dtu.dk/22126/index.php/Microbial_genomics_exercise

Microgem pipeline: Tiered comparative genomic analysis of bacterial isolates

Available from https://repo-rh.ngc.dk/[...]/microgem



Basis	**De novo genome assembly and prediction of species and MLST.**
Tier 1	**Alignment of assembly against all other assemblies of same species and MLST.**
Tier 2	**Comparison against most closely related genomes from Tier 1 by alignment of reads to a reference genome.**

MicroGem was presented with a poster at FEMS 2023 in Hamburg

Fundamental diagnostic questions in clinical microbiology

Is there something ?

What is it ?

What can it do ?

In everyday life in clinical microbiology,
whole genome sequencing does great in answering one of above

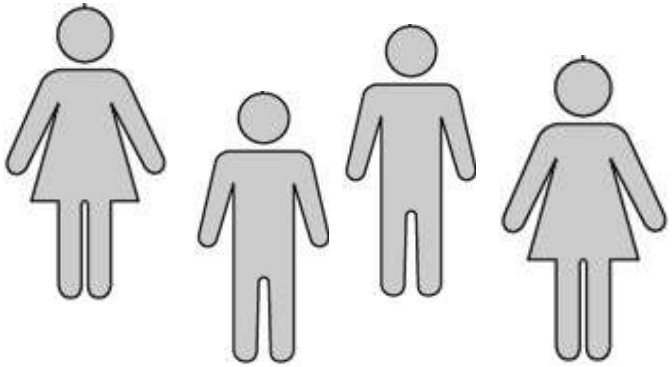
Fundamental diagnostic questions in clinical microbiology

- Is there something ?
- What is it ?
- What can it do ?

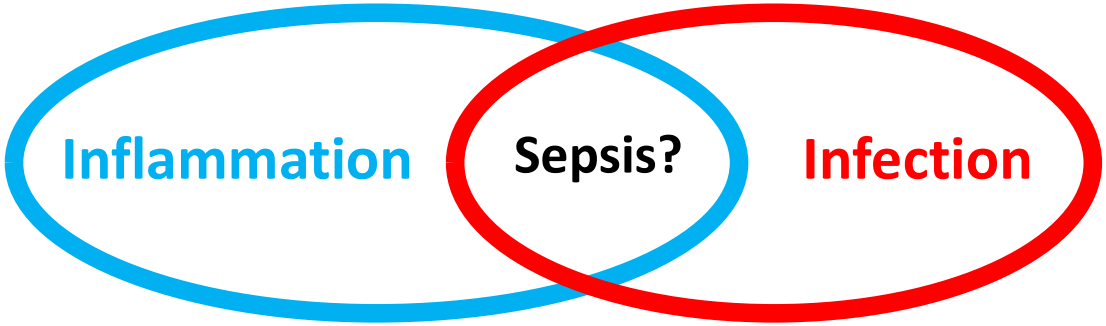
In everyday life in clinical microbiology,
whole genome sequencing does great in answering one of above

Difficult to determine if systemic inflammation is caused by infection (often no microbial pathogen is detected with conventional techniques)

Patient cohorte

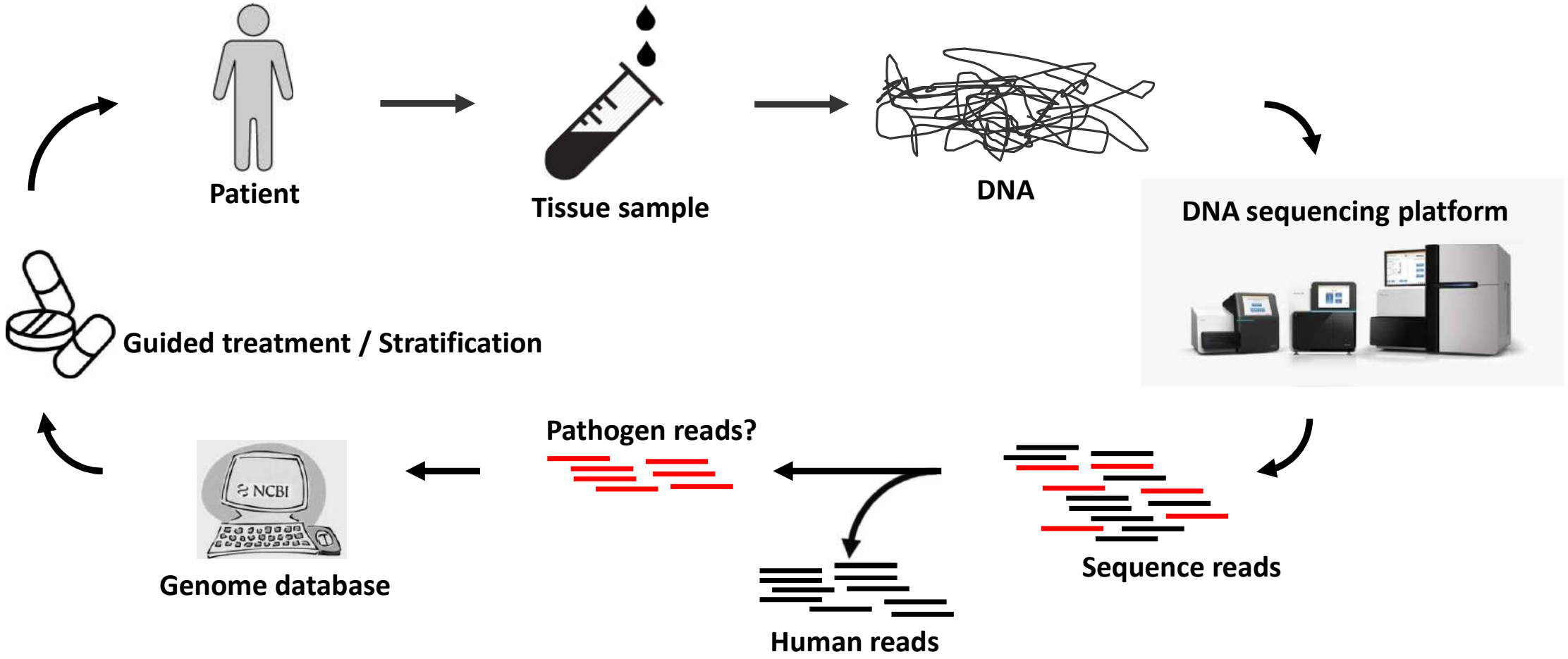


Definitions

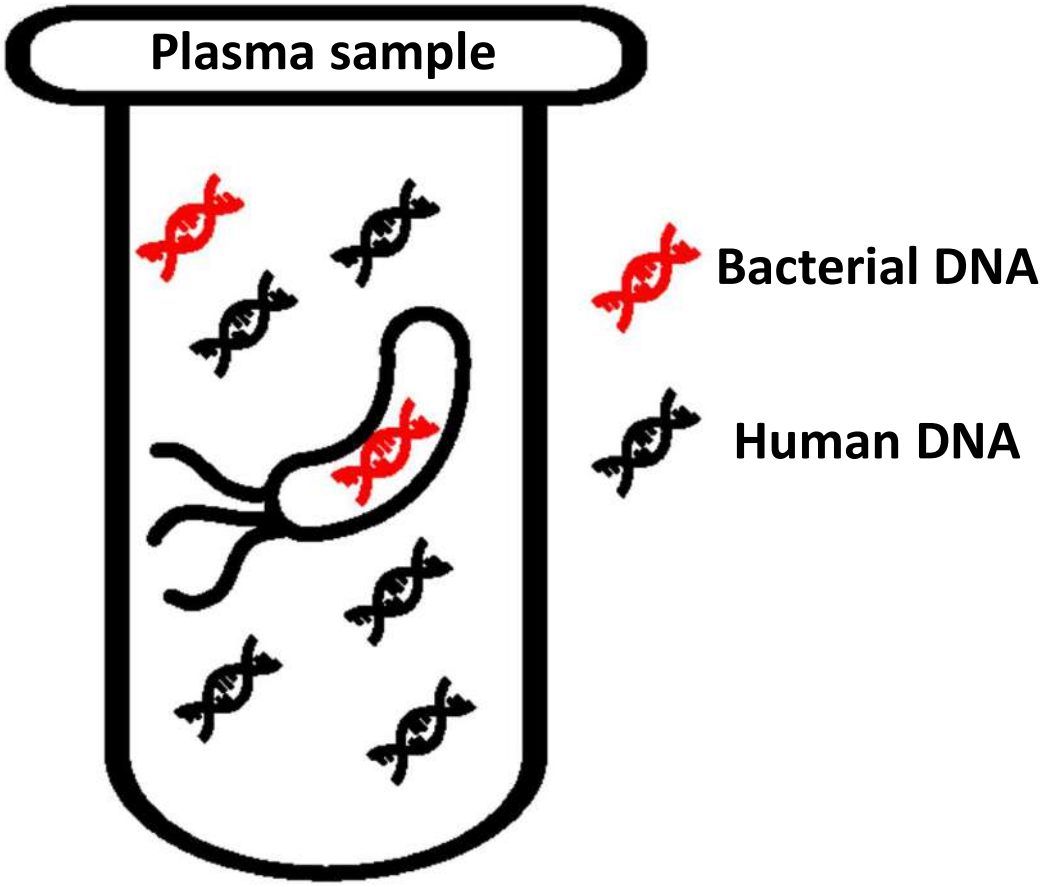


Immunocompromised patients with inflammation

Strategy for detection of bacterial pathogen in host tissue: Deep sequencing of DNA from host tissue that may contain pathogen



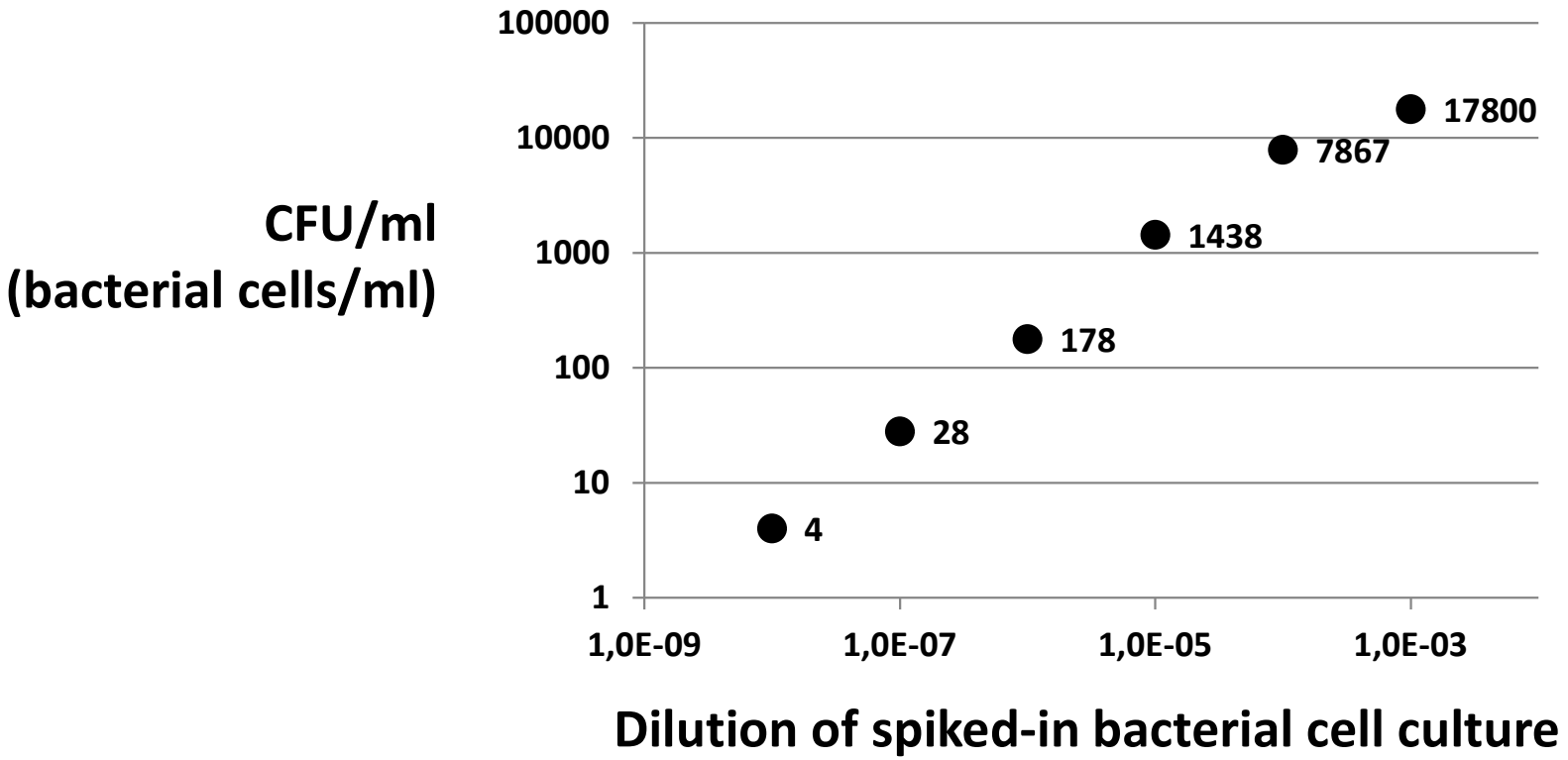
Challenge: There is a high abundance of human host DNA relative to microbial pathogen DNA in many samples types including plasma



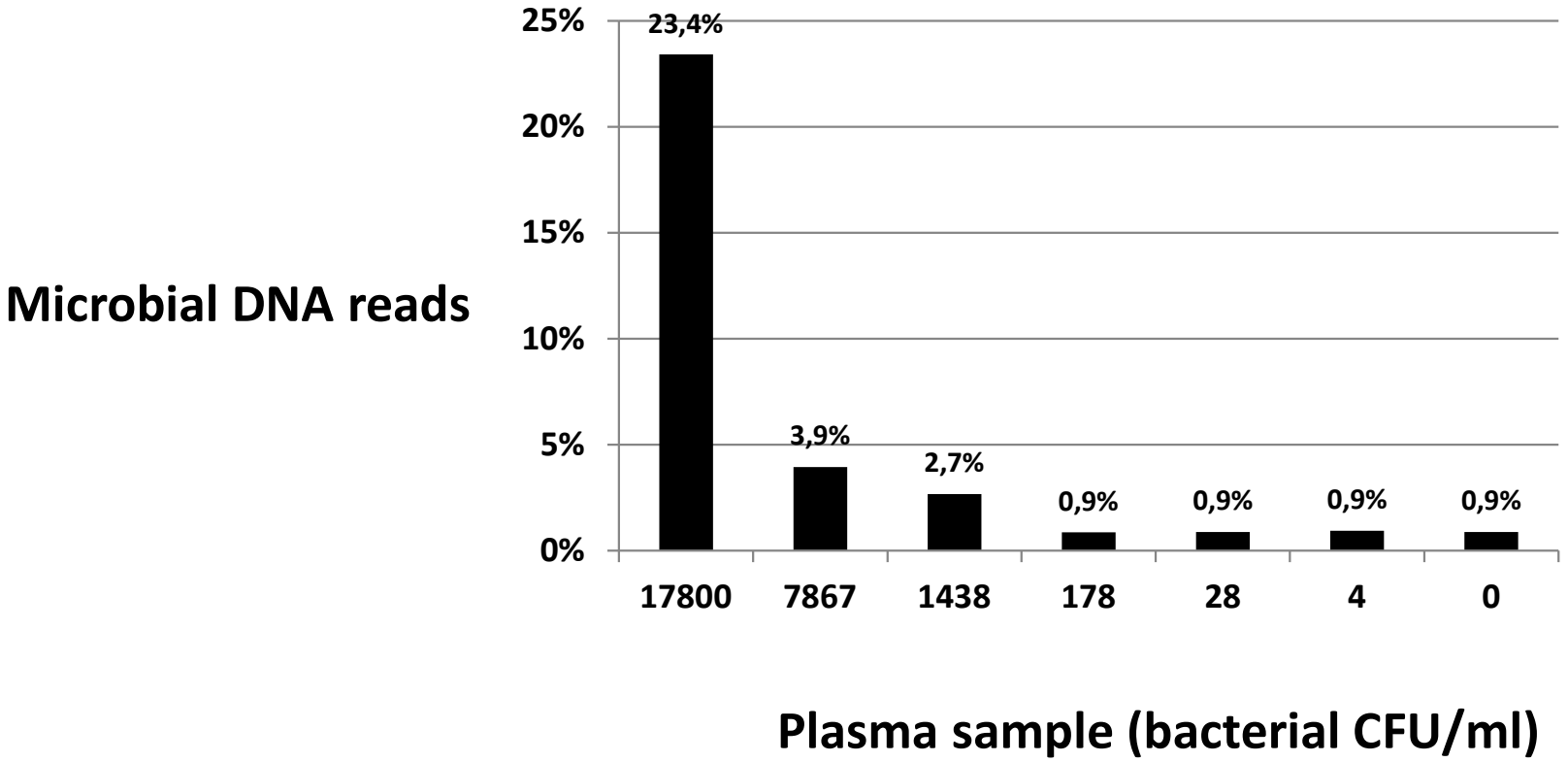
Sizes of genomes are very different:
Human genome is 3,000,000,000 nt
Microbial genome is 5,000,000 nt

DNA amounts in plasma:
Plasma contains 10-30 ng/ml human DNA
10 microbial cells/ml equals 0.00005 ng
Bacterial to human DNA ratio is 1:200,000

Plasma sample from healthy donor with spike-in of different amounts of bacterial cells (*Escherichia coli*)

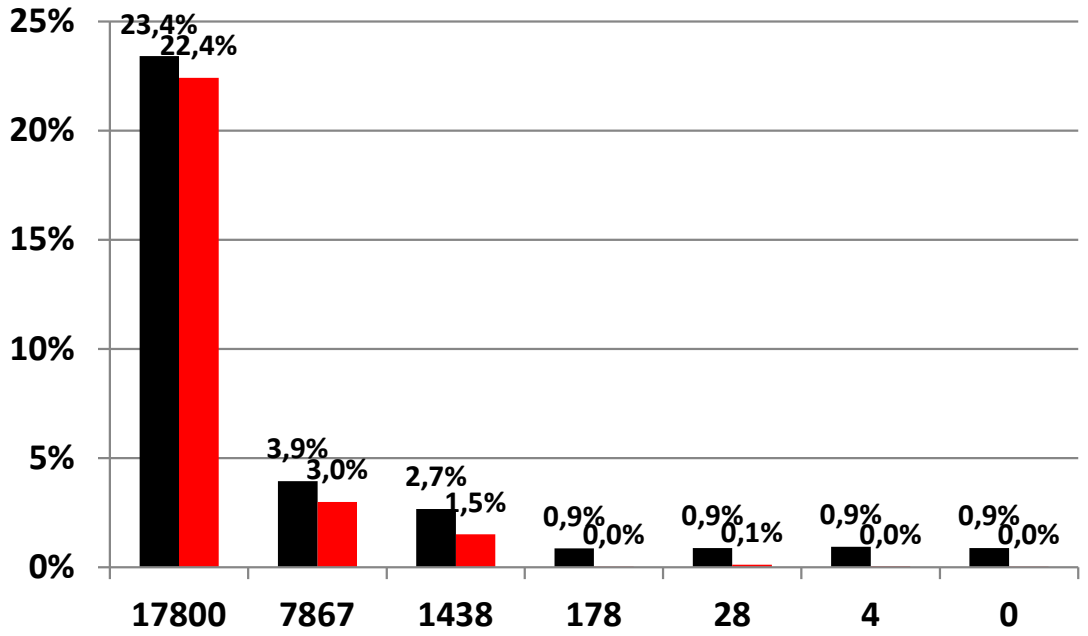


Fraction of total DNA reads that are classified as microbial DNA reads



Fraction of total DNA reads that are classified as microbial DNA reads

Microbial DNA reads
(red is E. coli reads)



Plasma sample (bacterial CFU/ml)

Another challenge of deep sequencing of metagenomic DNA libraries: Example - Contamination likely explains 'food genes in blood' claim

Claim:

OPEN ACCESS Freely available online

PLOS ONE

Complete Genes May Pass from Food to Human Blood

Sándor Spisák[1,2]*, Norbert Solymosi[3,4], Péter Ittzés[3], András Bodor[3], Dániel Kondor[3], Gábor Vattay[3], Barbara K. Barták[5], Ferenc Sipos[5], Orsolya Galamb[5], Zsolt Tulassay[1,5], Zoltán Szállási[2],

Disproof:

OPEN ACCESS Freely available online

PLOS ONE

Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data

Richard W. Lusk*

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, Michigan, United States of America

Analysis of DNA sequencing results from low microbial biomass samples remains to be challenging

nature microbiology

Consensus Statement

https://doi.org/10.1038/s41564-025-02035-2

Guidelines for preventing and reporting contamination in low-biomass microbiome studies

Received: 9 December 2021

Accepted: 15 May 2025

Published online: 20 June 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Numerous important environments harbour low levels of microbial biomass, including certain human tissues, the atmosphere, plant seeds, treated drinking water, hyper-arid soils and the deep subsurface, with some environments lacking resident microbes altogether. These low microbial biomass environments pose unique challenges for standard DNA-based sequencing approaches, as the inevitability of contamination from external sources becomes a critical concern when working near the limits of detection. Likewise, lower-biomass samples can be disproportionately impacted by cross-contamination and practices suitable for handling higher-biomass samples may produce misleading results when applied to lower microbial biomass samples. This Consensus Statement outlines strategies to reduce contamination and cross-contamination, focusing on marker gene and metagenomic analyses. We also provide minimal standards for reporting contamination information and removal workflows. Considerations must be made at every study stage, from sample collection and handling through data analysis and reporting to reduce and identify contaminants. We urge researchers to adopt these recommendations when designing, implementing and reporting microbiome studies, especially those conducted in low-biomass systems.

Diagnostic metagenome sequencing is in the future expected to be part of the clinical toolbox for detection of infectious microbial pathogens

AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Spectrum

Check for updates

| Clinical Microbiology | Research Article

Application of rapid Nanopore metagenomic cell-free DNA sequencing to diagnose bloodstream infections: a prospective observational study

Morten Eneberg Nielsen,[1] Kirstine Kobberøe Søgaard,[2,3] Søren Michael Karst,[1] Anne Lund Krarup,[3,4] Mads Albertsen,[1] Hans Linde Nielsen[2,3]

Nielsen et al., Microbiology Spectrum, 2025

AAU based research that has led to start up (SeeQ Diagnostics)

In Task01 you will use Kraken to analysis data from Nielsen *et al.*

Other commercial solutions:
https://noscendo.com
https://kariusdx.com

Fundamental diagnostic questions in clinical microbiology

- Is there something ?
- What is it ?
- What can it do ?

Taxonomy: A system for classifying and organizing organisms into a hierarchical structure of groups (taxa) based on shared characteristics

Why do we need a taxonomy ?

- **To categorize organisms so we can easily communicate.**
- **Taxonomy uses hierarchical categorization to organize the diversity of life.**

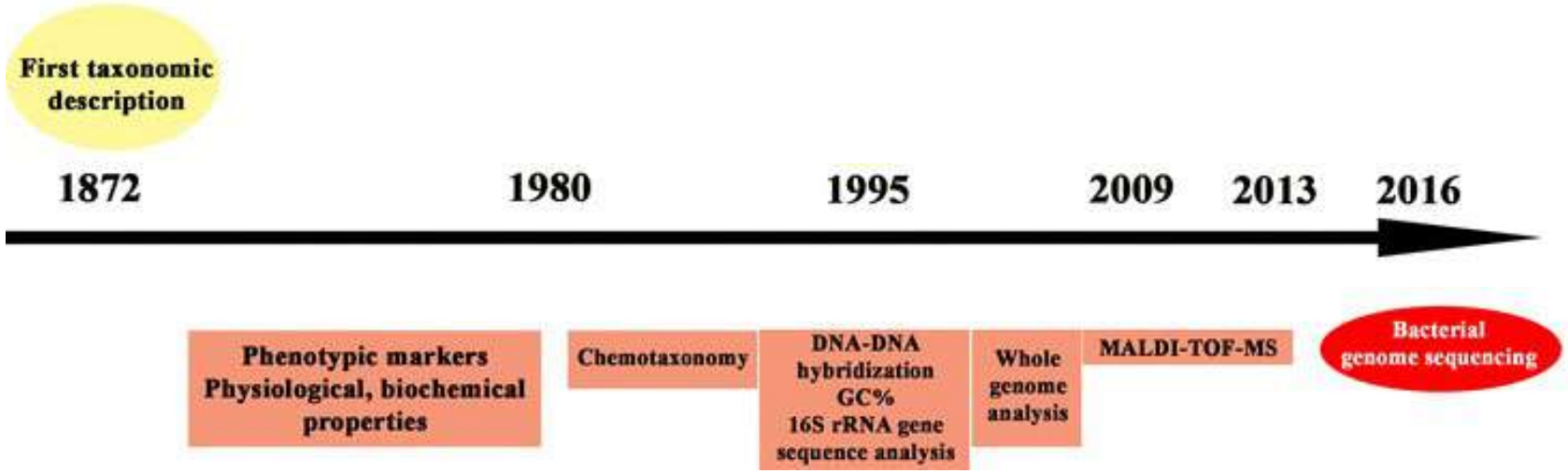
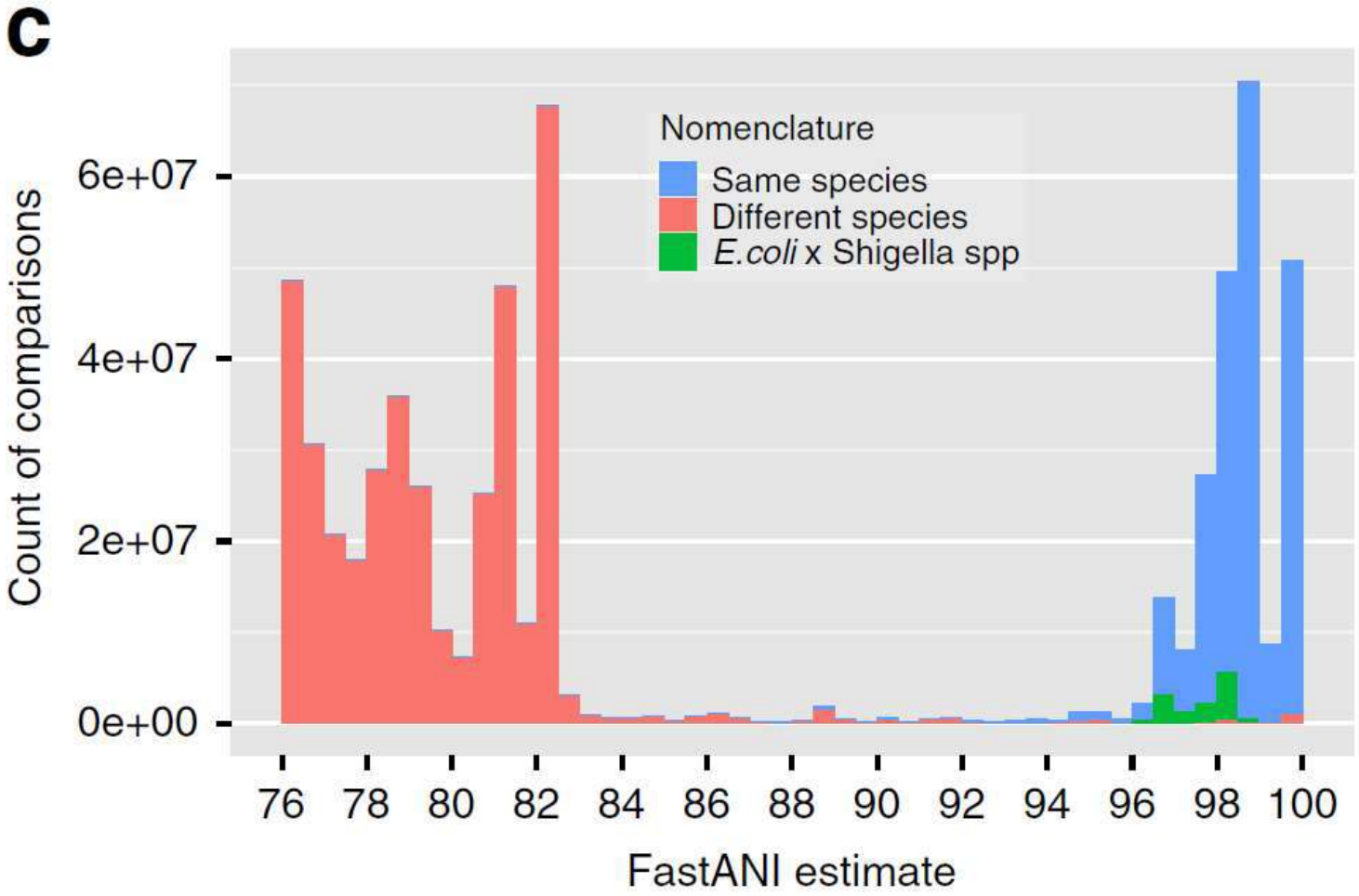


Figure from: Abdallah et al., Antonie Van Leeuwenhoek, 2017: Changes in bacterial taxonomic tools over the years

High throughput ANI (average nucleotide identity) analysis of 90K prokaryotic genomes reveals clear species boundaries



⇒ **Birth of the FastANI tool**

Nature Communications, 2018
5,075 citation (Google Scholar)

Genome Taxonomy Database has defined 136,646 bacterial species (based on analysis of 732,475 bacterial genomes)

Nature Biotechnology, 2018
3,507 citations (Google Scholar)

Result example:
Use of FastANI to compare genome against reference genome database

List of best matches when query genome is compared against selection of Refseq reference genomes:

Genome match	% ANI	Mapped_fragments	Total_Fragments
Pseudomonas_aeruginosa_GCF_001045685.1_ASM104568v1_genomic.fna.gz	99	1973	2228
Pseudomonas_delhiensis_GCF_900099945.1_IMG-taxon_2671180033_annotated_assembly_genomic.fna.gz	85	1340	2228
Pseudomonas_humi_GCF_001748265.1_ASM174826v1_genomic.fna.gz	85	1386	2228
Pseudomonas_citronellolis_GCF_001654435.1_ASM165443v1_genomic.fna.gz	85	1368	2228
Pseudomonas_jinjuensis_GCF_900103845.1_IMG-taxon_2663762768_annotated_assembly_genomic.fna.gz	84	1156	2228
Pseudomonas_knackmussii_GCF_000689415.1_PKB13_genomic.fna.gz	84	1195	2228
Pseudomonas_panipatensis_GCF_900099785.1_IMG-taxon_2663762770_annotated_assembly_genomic.fna.gz	84	1161	2228
Pseudomonas_nitroreducens_GCF_002091755.1_ASM209175v1_genomic.fna.gz	84	1237	2228
Pseudomonas_denitrificans_GCF_008807415.1_ASM880741v1_genomic.fna.gz	84	1304	2228
Pseudomonas_resinovorans_GCF_000412695.1_ASM41269v1_genomic.fna.gz	82	1128	2228

Conclusion:

Species of query genome is Pseudomonas aeruginosa as % ANI is >95% and >50% of fragments mapped

Use of GTDB for species classification not feasible until memory friendly update to GTDB toolkit

Bioinformatics, 38(23), 2022, 5315–5316
https://doi.org/10.1093/bioinformatics/btac672
Advance Access Publication Date: 11 October 2022
Applications Note

OXFORD

Genome analysis

GTDB-Tk v2: memory friendly classification with the genome taxonomy database

Pierre-Alain Chaumeil [1,2,*], **Aaron J. Mussig** [1], **Philip Hugenholtz** [1] **and Donovan H. Parks** [1,*]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, QLD 4072, Australia and [2]Research Computing Center, The University of Queensland, St Lucia, QLD 4072, Australia

*To whom correspondence should be addressed.

Associate Editor: Karsten Borgwardt

Received on July 10, 2022; revised on September 23, 2022; editorial decision on October 3, 2022; accepted on October 7, 2022

Abstract

Summary: The Genome Taxonomy Database (GTDB) and associated taxonomic classification toolkit (GTDB-Tk) have been widely adopted by the microbiology community. However, the growing size of the GTDB bacterial reference tree has resulted in GTDB-Tk requiring substantial amounts of memory (~320 GB) which limits its adoption and ease of use. Here, we present an update to GTDB-Tk that uses a divide-and-conquer approach where user genomes are initially placed into a bacterial reference tree with family-level representatives followed by placement into an appropriate class-level subtree comprising species representatives. This substantially reduces the memory requirements of GTDB-Tk while having minimal impact on classification.

Availability and implementation: GTDB-Tk is implemented in Python and licenced under the GNU General Public Licence v3.0. Source code and documentation are available at: https://github.com/ecogenomics/gtdbtk.

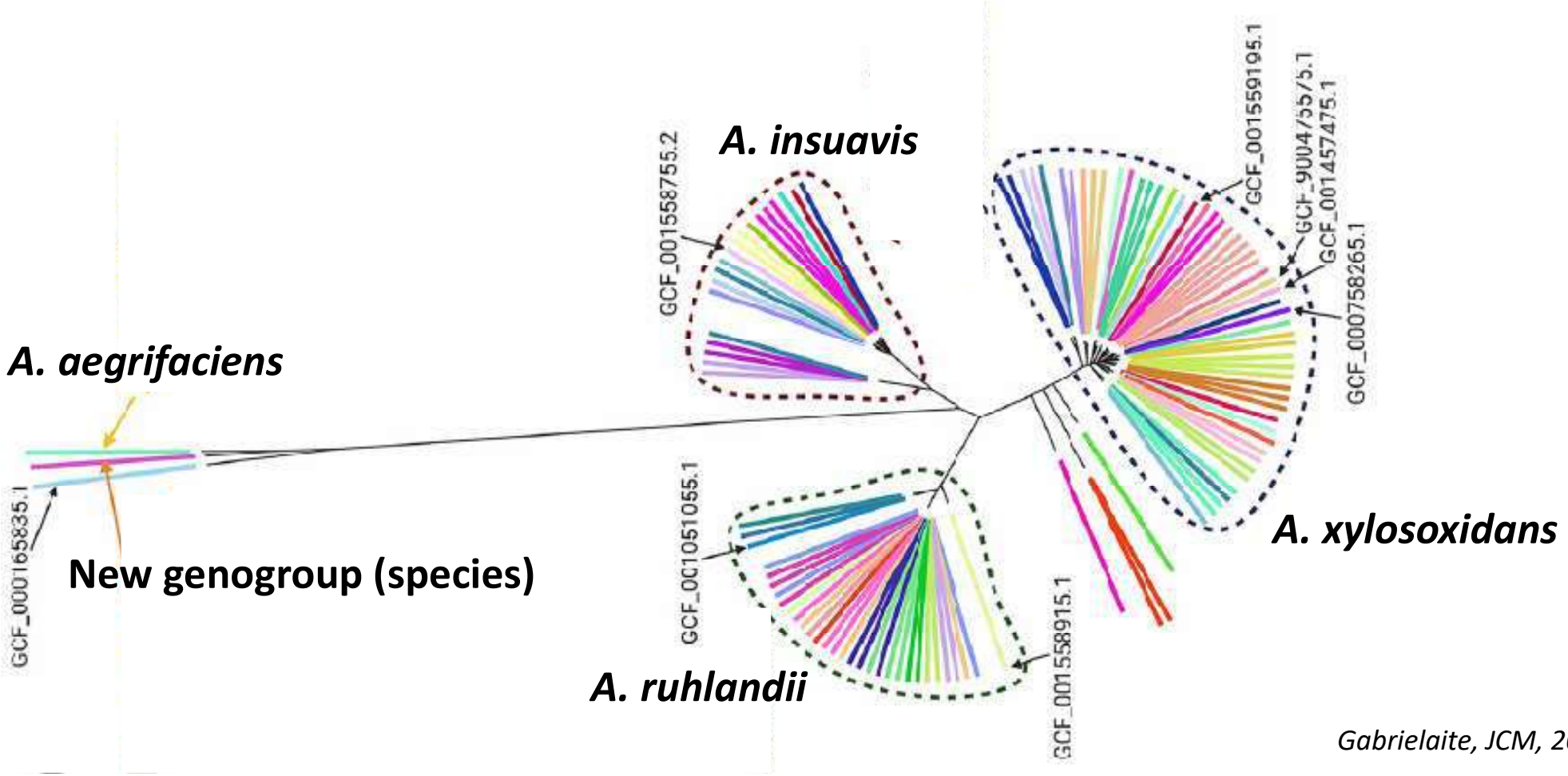
Contact: p.chaumeil@uq.edu.au or donovan.parks@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

GTDB-Tk now uses skani
(it was using fastANI until v2.3.2)

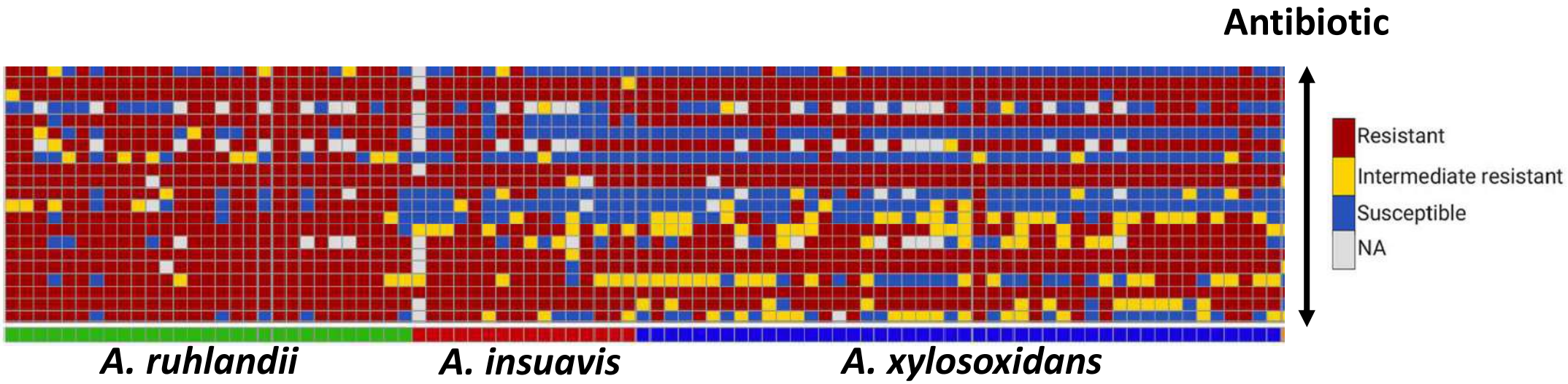
In Task02 you will use GTDB-Tk v2.6.1 to classify bacterial genomes

101 Achromobacter clinical isolates identified as *Achromobacter xylosoxidans* based on MALDI-TOF or API N20 typing



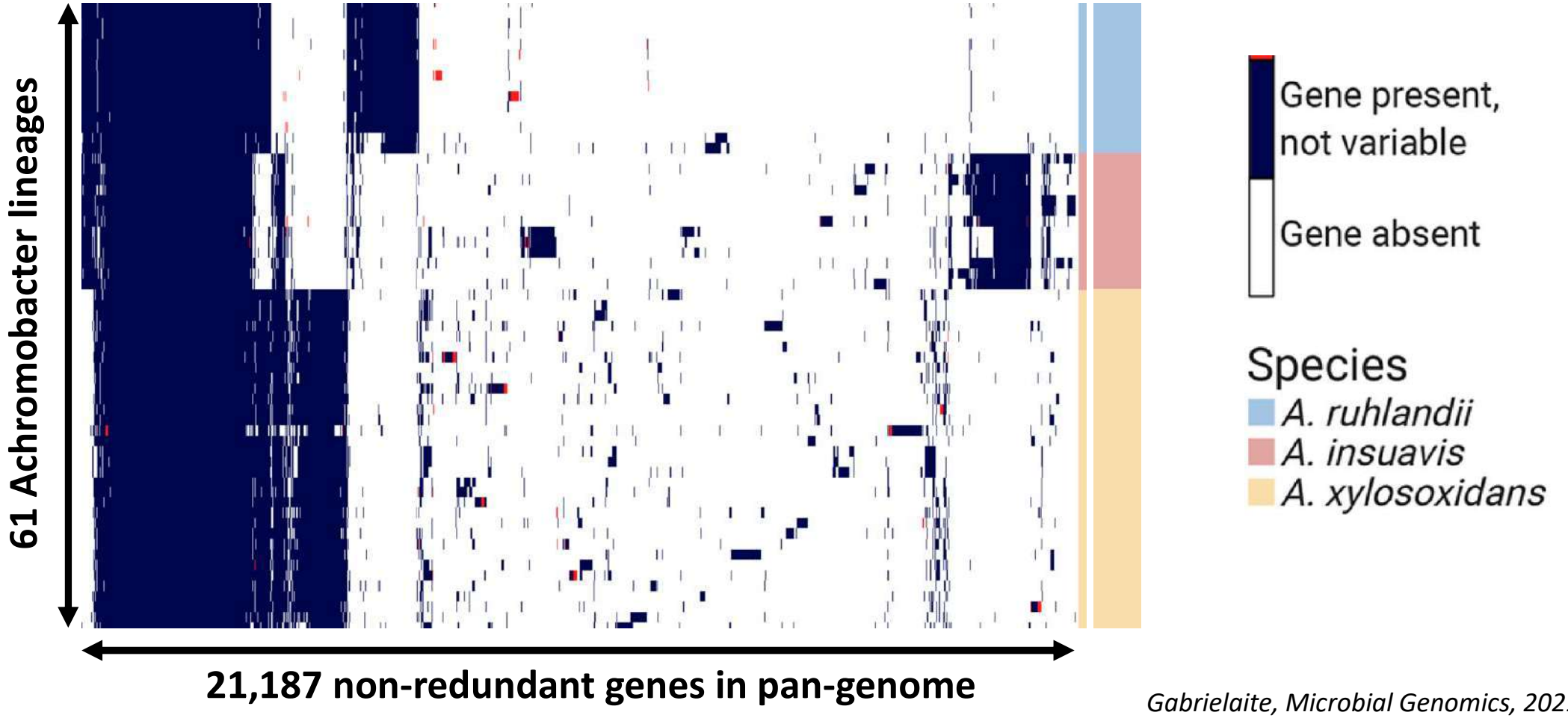
Gabrielaite, JCM, 2021

Correct species typing helps to guide treatment: E.g. different species has different antibiotic susceptibility profiles



Gabrielaite, JCM, 2021

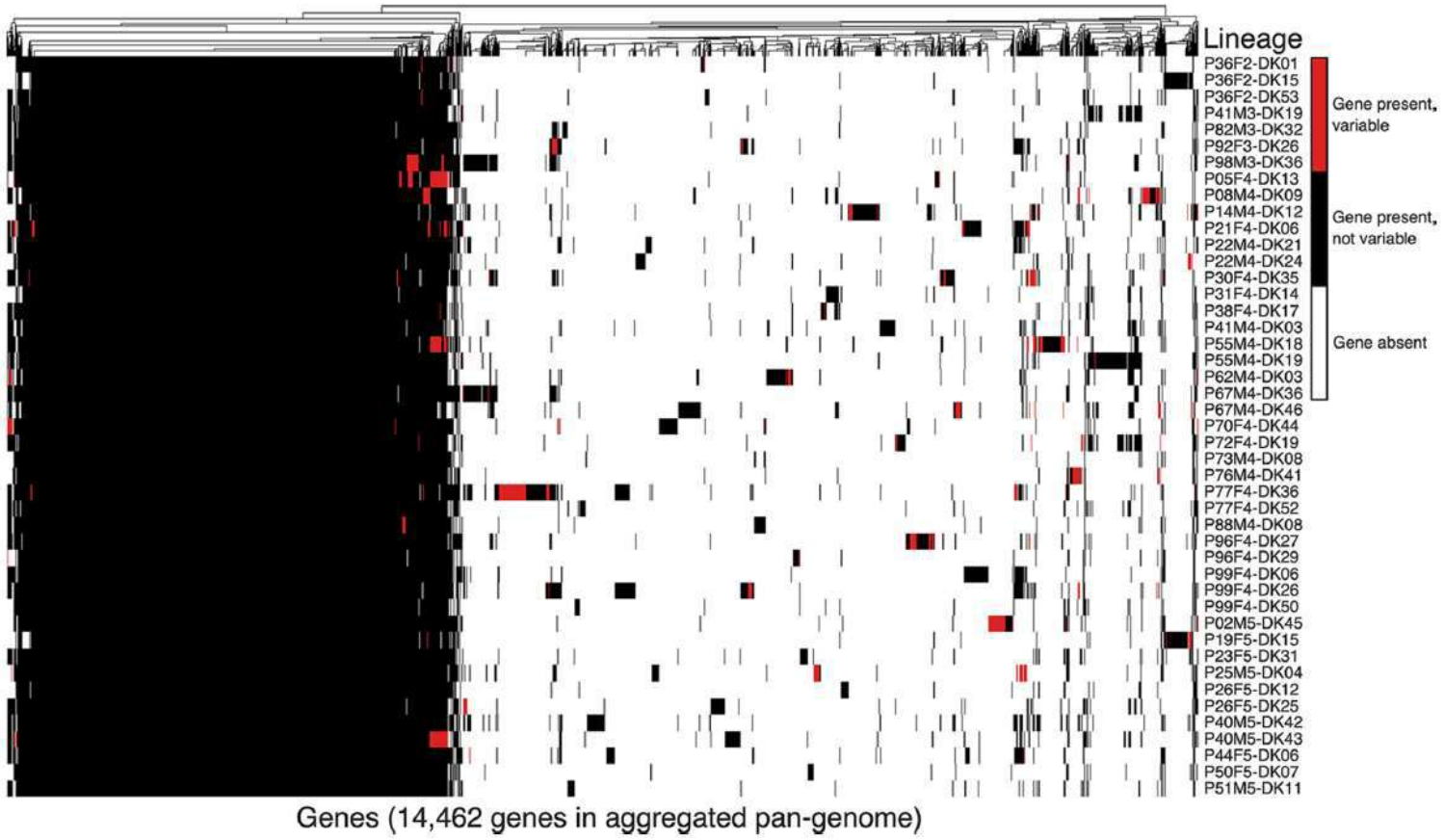
Genome content (gene content) varies within and between species



Gabrielaite, Microbial Genomics, 2021

A clinical collection of 446 *Pseudomonas aeruginosa* genomes share 4,760 genes (core-genome) among a pan-genome with 14,462 genes

Core genome of 4,760 genes



https://github.com/MigleSur/Pactyper

Gabrielaite et al., mBio, 2020

Genetic distances (single nucleotide variants) between genomes of the same bacterial species: Real data for 446 isolates of *P. aeruginosa*

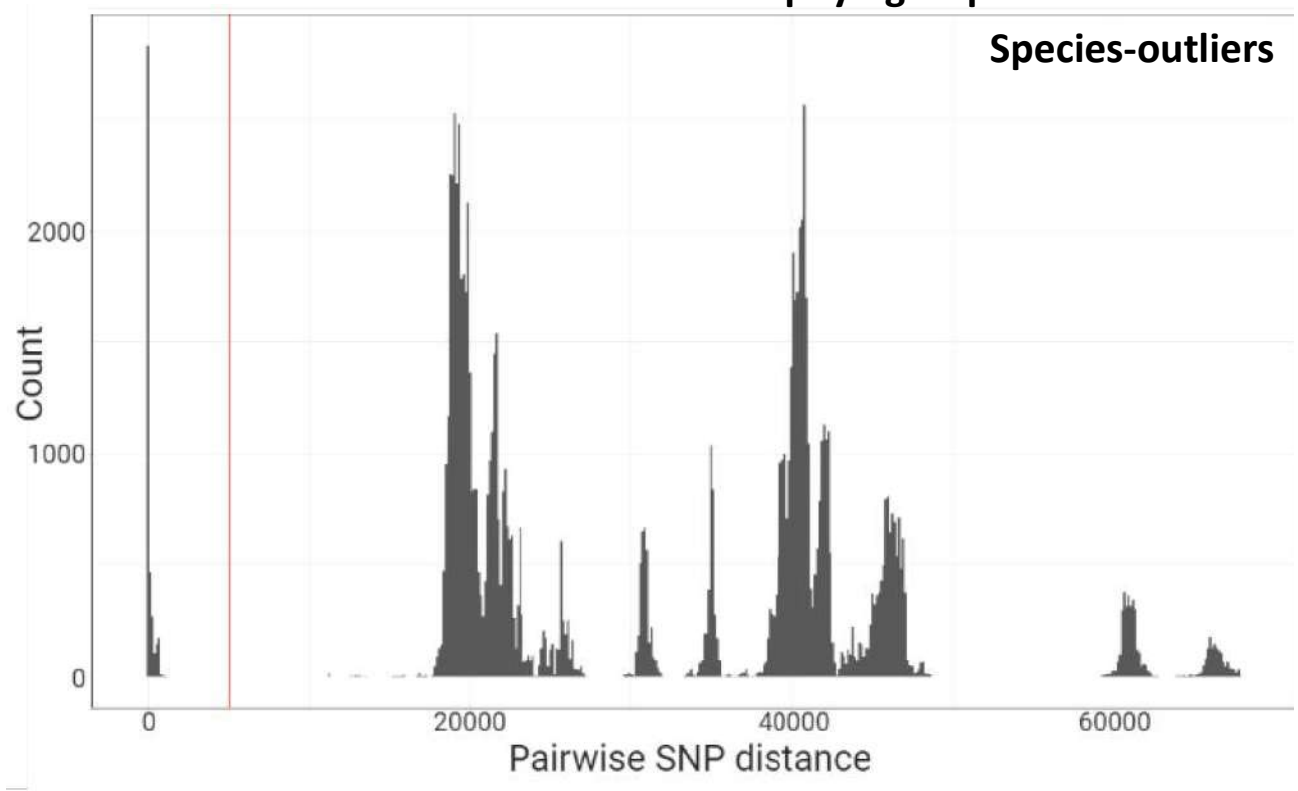
Sub-species typing

Same strain

Different strains

Different phylogroups

Species-outliers



Gabrielaite, mBio, 2020

Histogram of pairwise SNP distances between 446 P. aeruginosa isolates in the core genome: Red dotted line is chosen threshold for typing

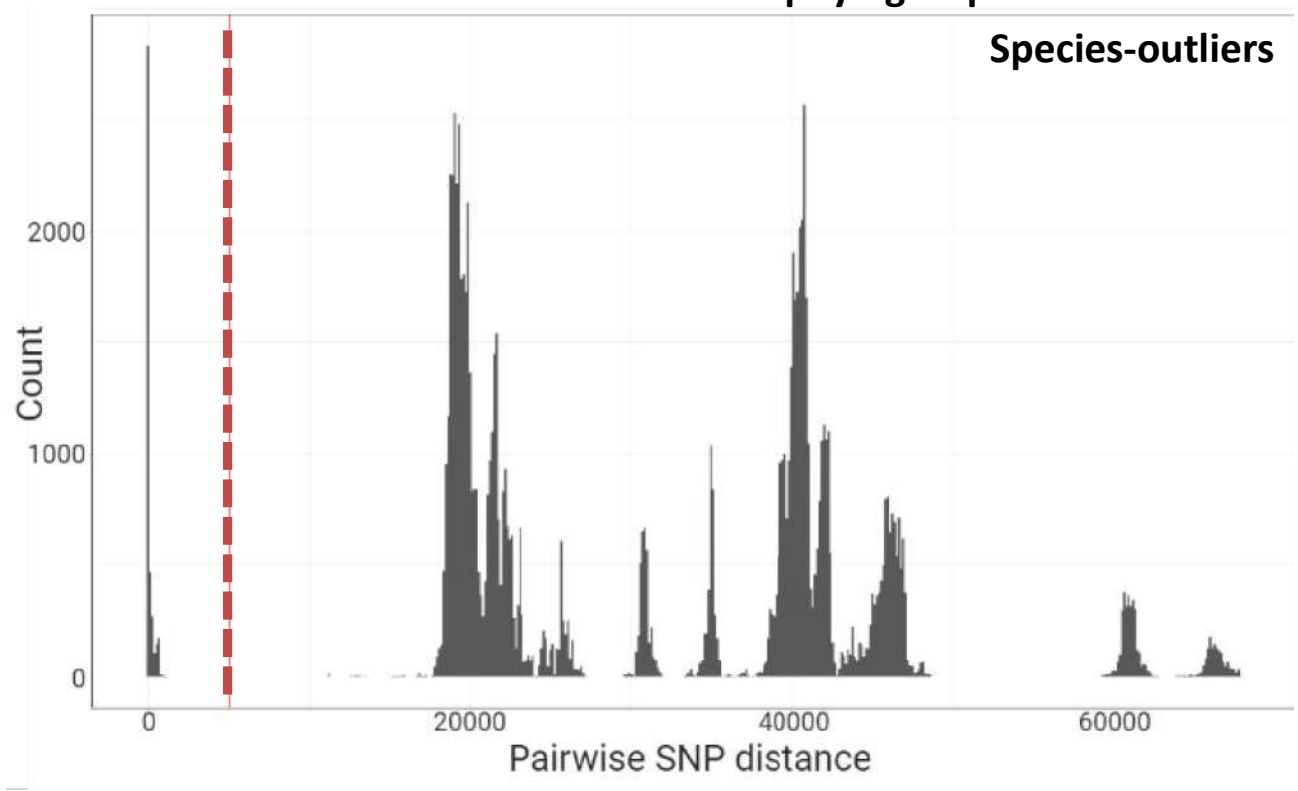
Sub-species typing

Same strain

Different type

Different phylogroups

Species-outliers



In Task03 you will use mlst for subspecies-level typing of bacterial genomes

Gabrielaite, mBio, 2020