Presentation for Bioinformatics course in programme for Master's degree in Personalized Medicine

Microbial genomics in personalized medicine

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

Conflicts of interest - secondary occupations:
Technical Assessor, DANAK - The Danish Accreditation Fund

Fundamental diagnostic questions in clinical microbiology

Is there something ?

What is it ?

What can it do ?

In everyday life in clinical microbiology,
Microbial genomics are great in answering one of above questions

Fundamental diagnostic questions in clinical microbiology

➡ **Is there something ?**

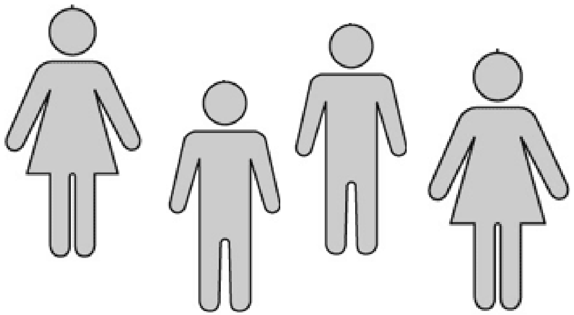
What is it ?

What can it do ?

In everyday life in clinical microbiology,
Microbial genomics are great in answering one of above questions

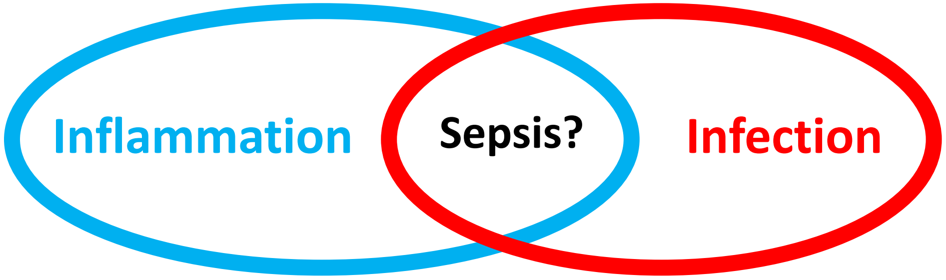
Difficult to determine if systemic inflammation is caused by infection (often no microbial pathogen is detected with conventional techniques)

Patient cohorte

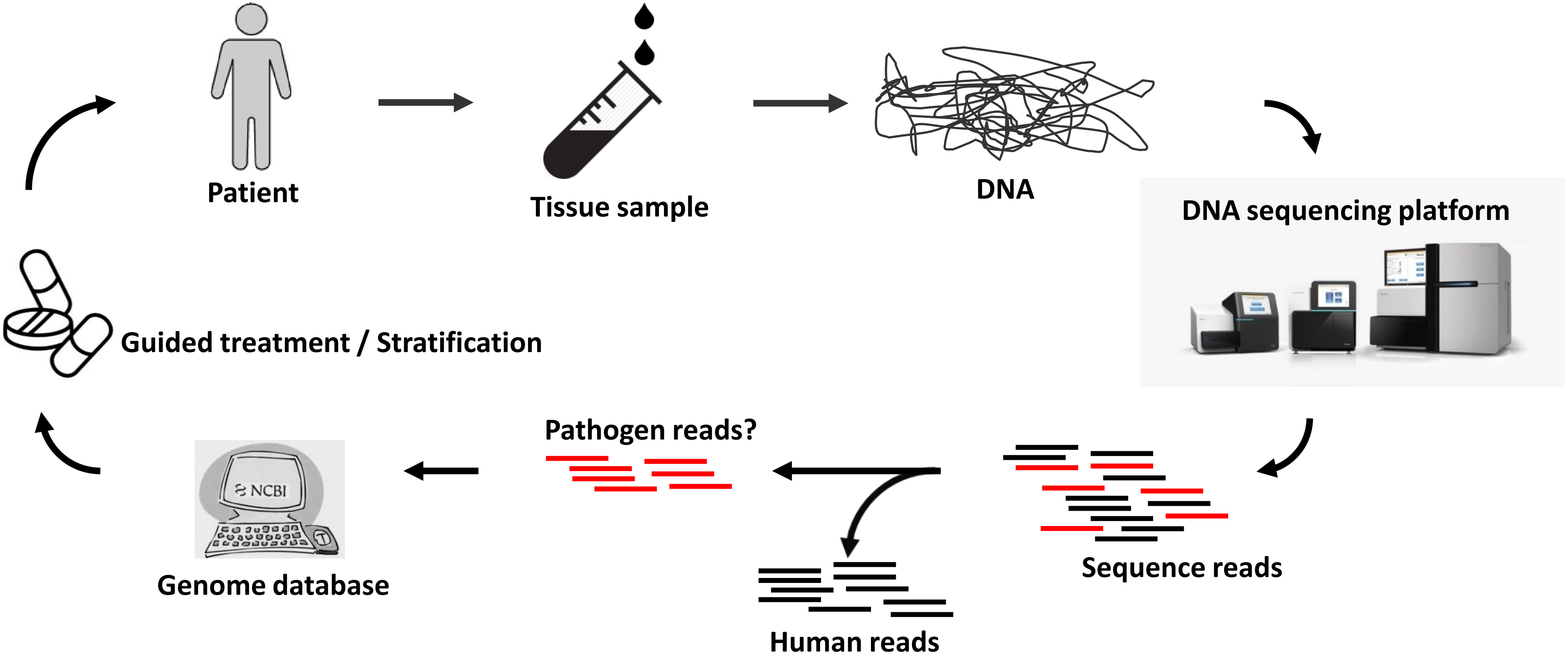


Immunocompromised patients with inflammation

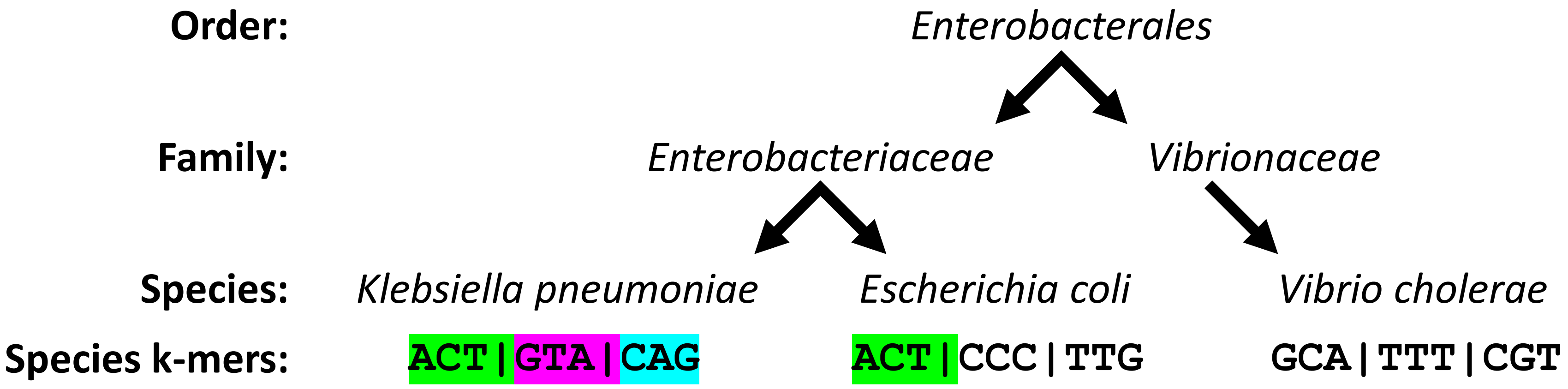
Definitions



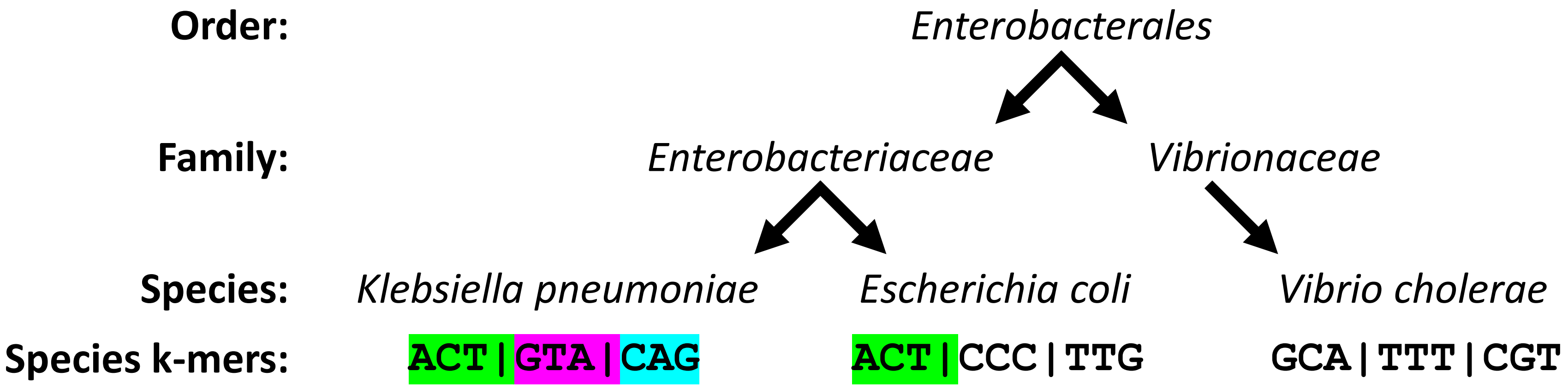
Strategy for detection of bacterial pathogen in host tissue:
Deep sequencing of DNA from host tissue that may contain pathogen



Principle: Assignment of species labels to short sequences (ACTGTACAG) by comparison to species-wide database with k-mers



Principle: Assignment of species labels to short sequences (ACTGTACAG) by comparison to species-wide database with k-mers

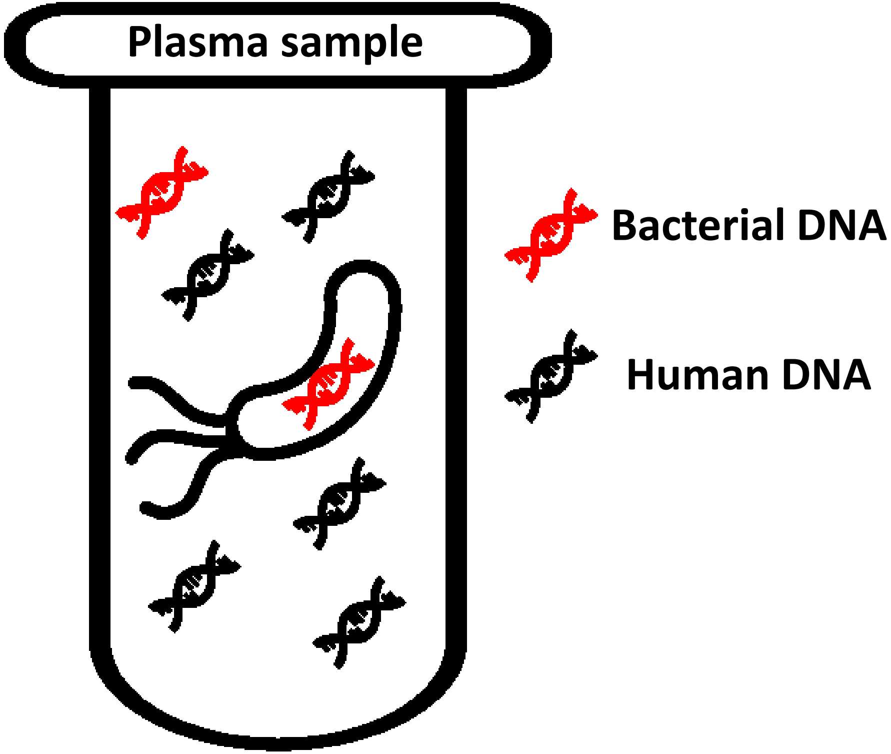


Query k-mer 1: ACT, unique to *Enterobacteriaceae*

Query k-mer 2 and 3: GTA , CAG, unique to *K. pneumoniae*

Sequence is labelled as *Klebsiella pneumoniae*

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

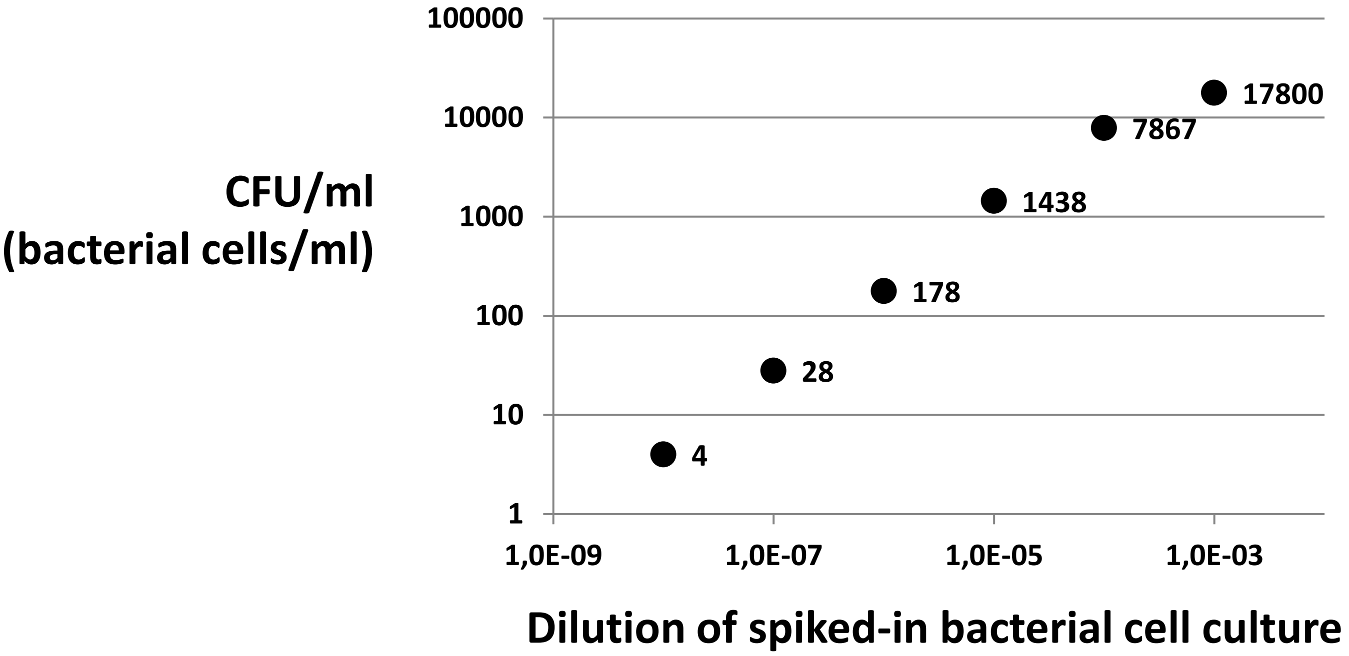
DNA amounts in plasma:

Plasma contains 10-30 ng/ml human DNA

10 microbial cells/ml equals 0.00005 ng

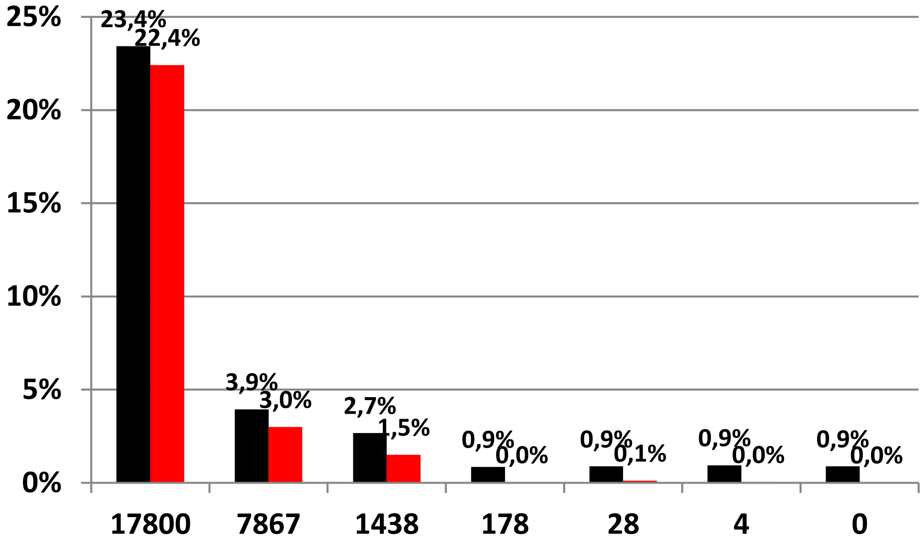
Bacterial to human DNA ratio is 1:200,000

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

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

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]

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

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.

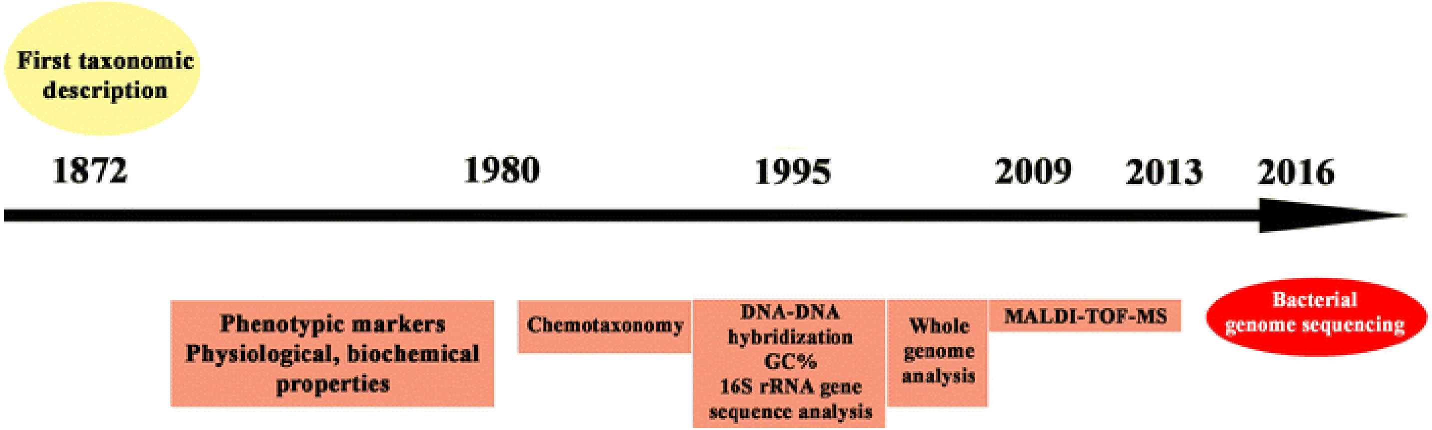
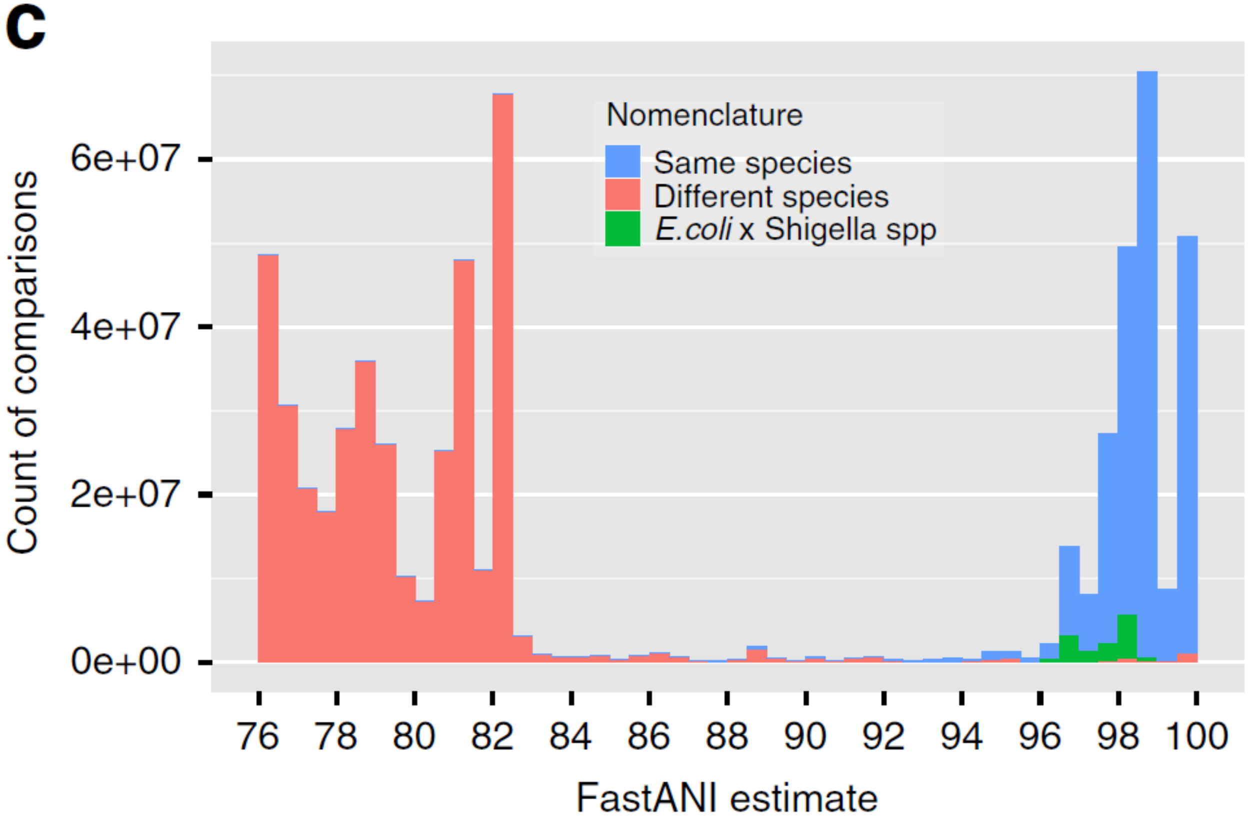


Figure from: Abdallah et al., Antonie Van Leeuwenhoek, 2017: Changes in bacterial taxonomic tools over the years

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)

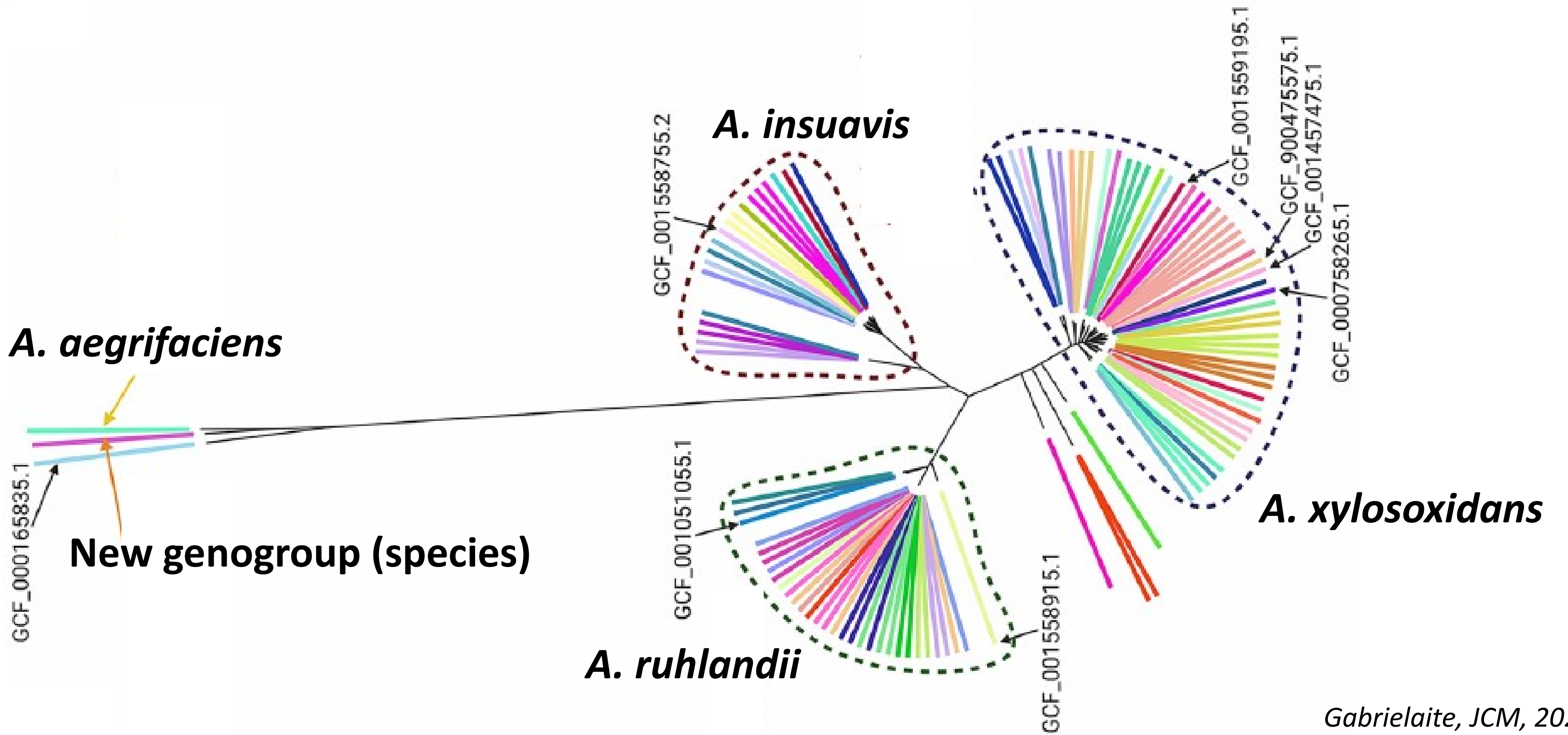
Nature Biotechnology, 2018
3,507 citations (Google Scholar)

Principle: Genome-based definition of species

Species	Genome sequence	Sequence similarity
Query genome	ACTGTACAGT	
Klebsiella pneumoniae	ACTGTACAGT	100% identity
Escherichia coli	ACTCCCTTGC	40% identity
Vibrio cholerae	GCATTTCGTA	30% identity

Conclusion: Query genome is *Klebsiella pneumoniae*

101 Achromobacter clinical isolates identified as *Achromobacter xylosoxidans* based on mass spectrometry or biochemical testing



Gabrielaite, JCM, 2021

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

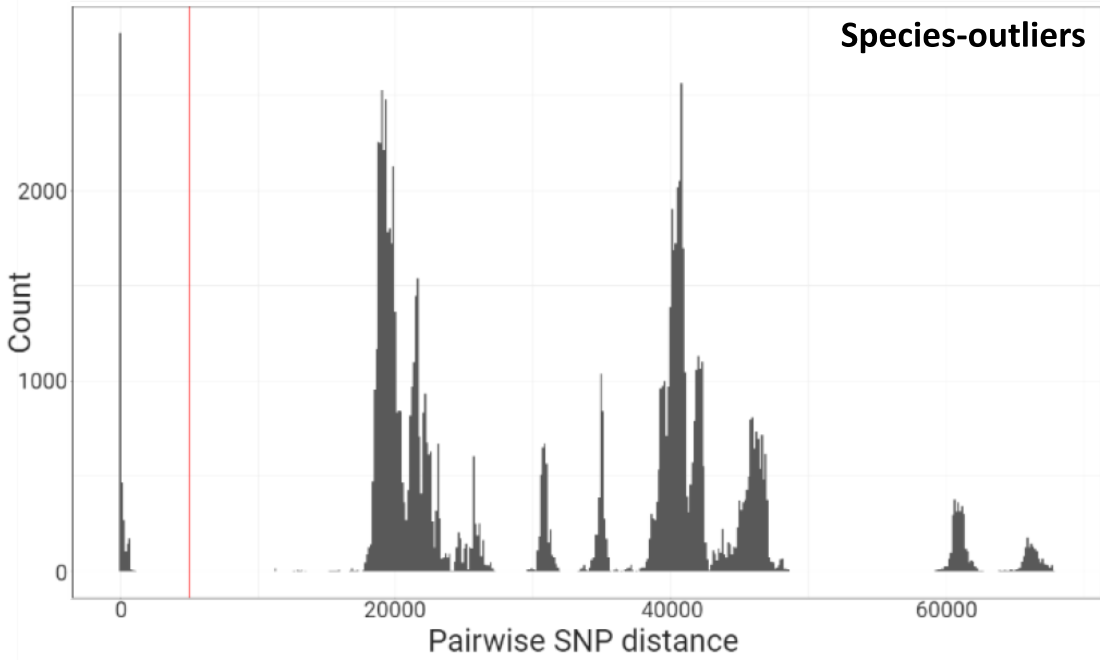
Sub-species typing

Same strain

Different strains

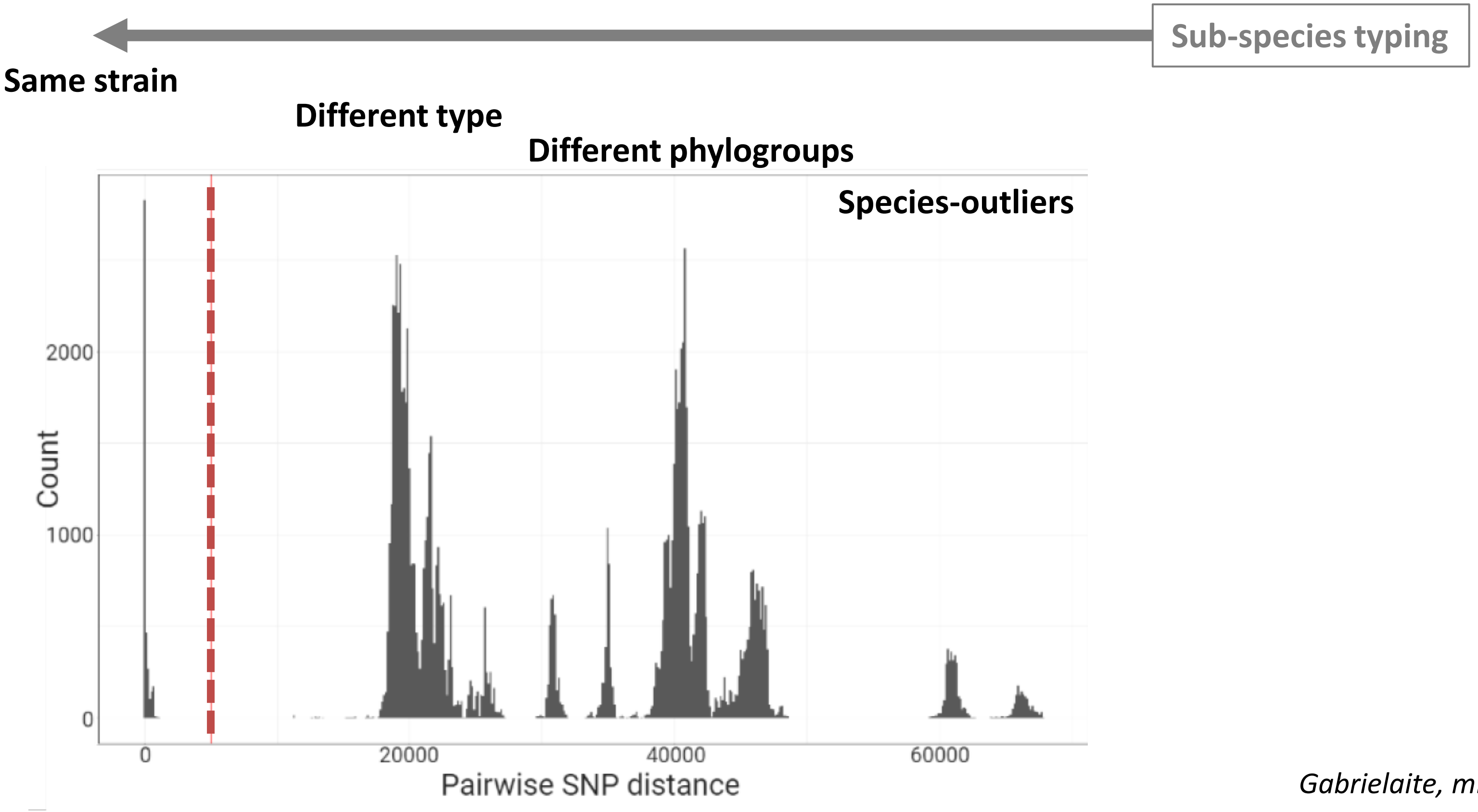
Different phylogroups

Species-outliers



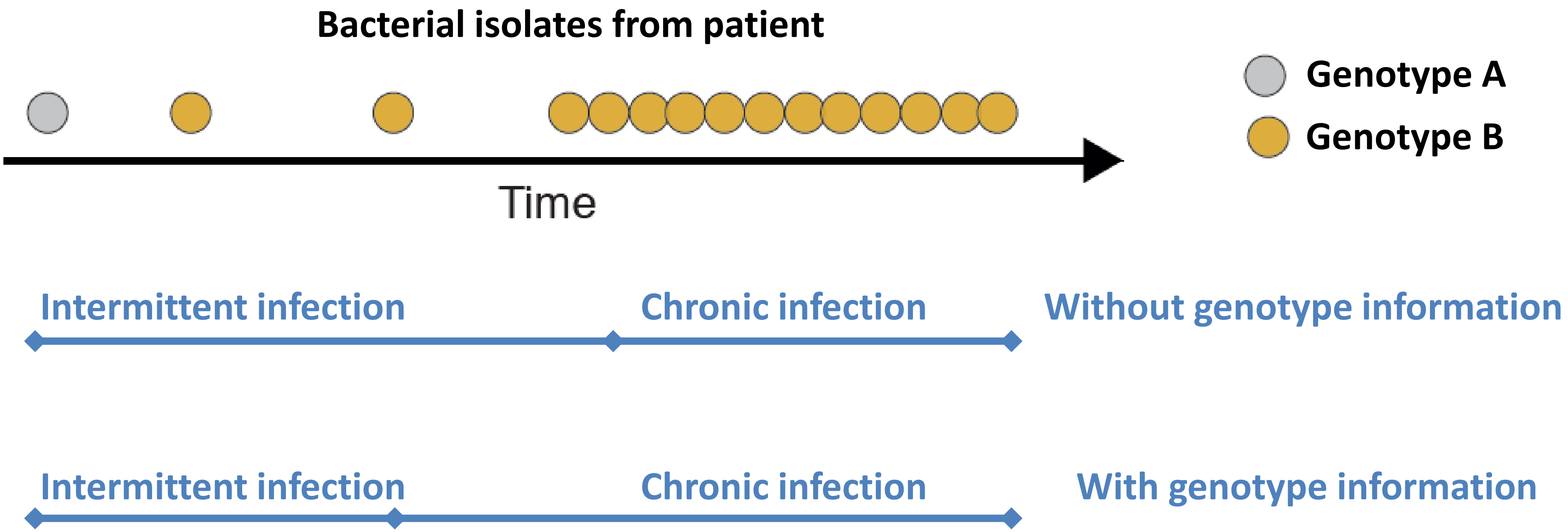
Gabrielaite, mBio, 2020

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



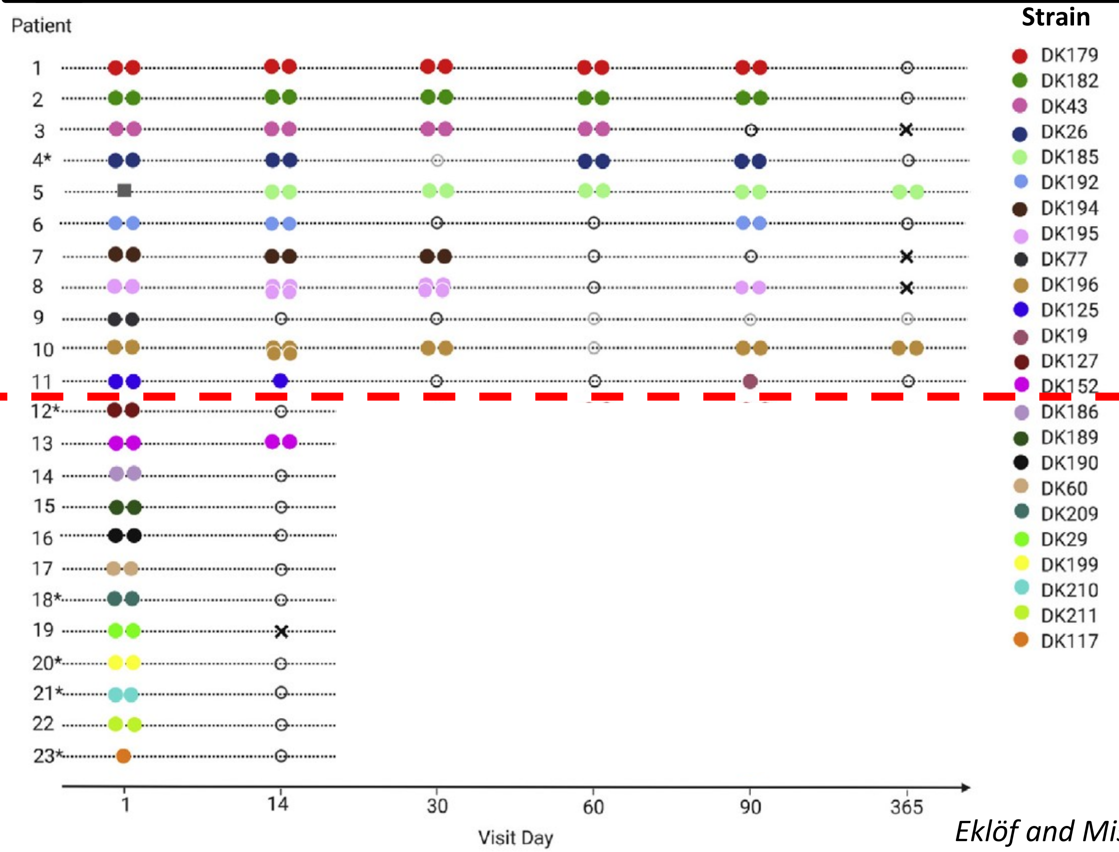
Gabrielaite, mBio, 2020

Genotyping offers increased resolution into strain dynamics and helps to determine if patient is chronically infected



Marvig, Nat Genet, 2015; Bartell, ERJ, 2021

Recurrence of *P. aeruginosa* in COPD patients is caused by persistence of the same strain – negative culture is false proxy for eradication

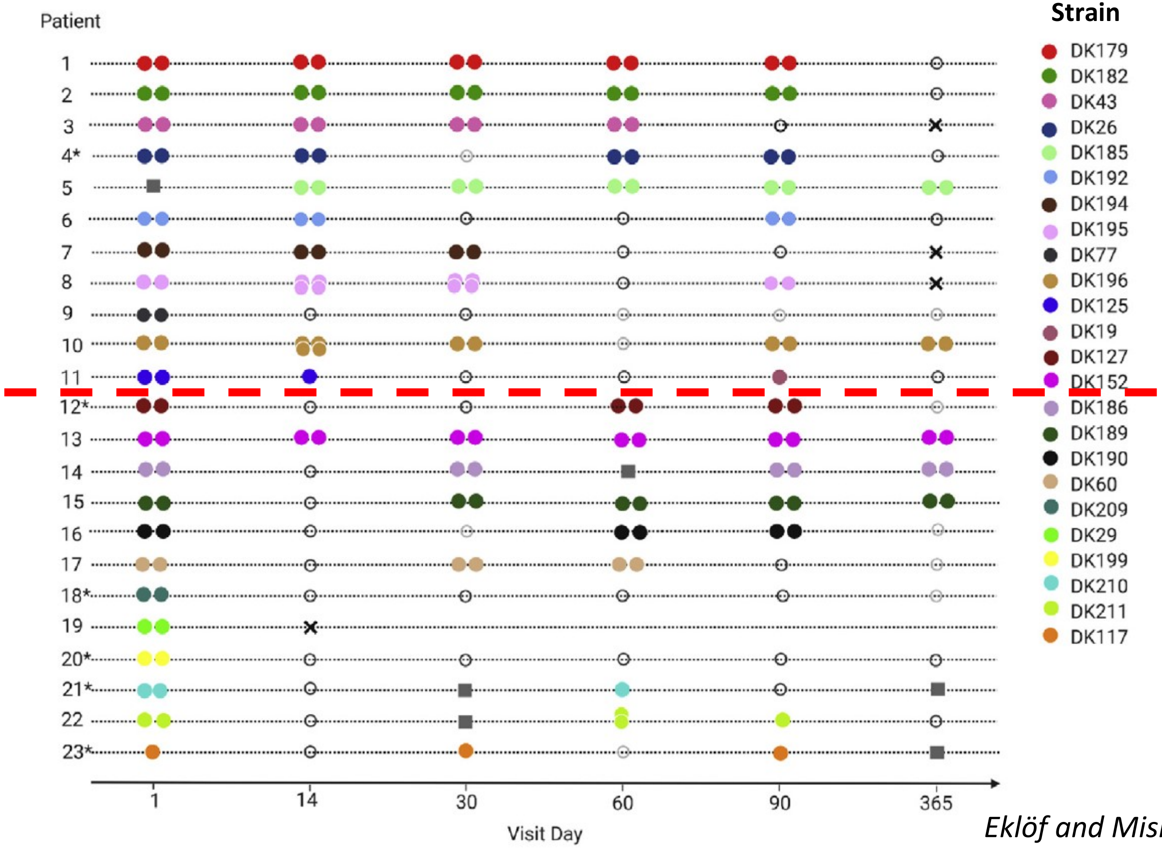


RCT control group:
No antipseudomonal treatment

RCT treatment group:
Antipseudomonal antibiotic treatment (intravenous b-lactam and oral ciprofloxacin) between day 1 and 14

Eklöf and Misiakou et al., Clinical Microbiology and Infection, 2022

Recurrence of *P. aeruginosa* in COPD patients is caused by persistence of the same strain – negative culture is false proxy for eradication

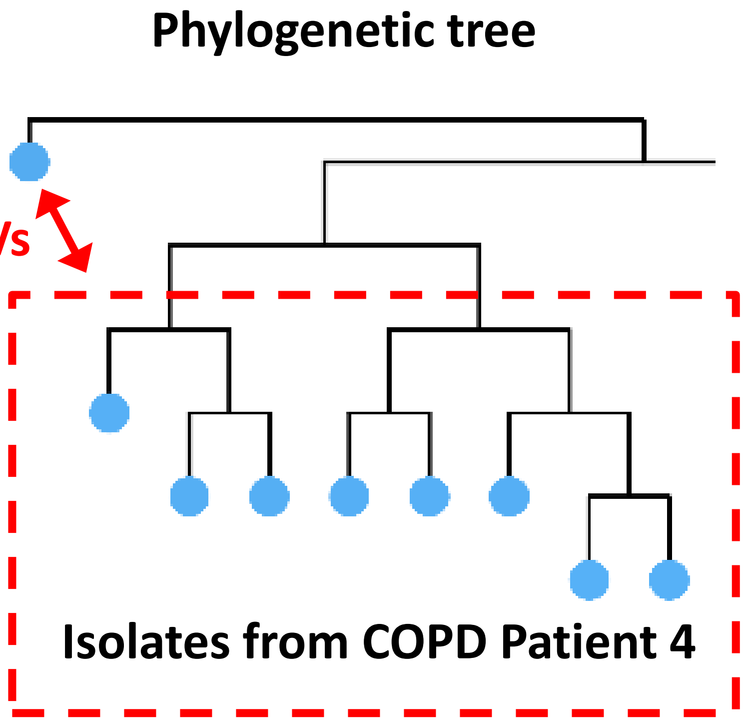


RCT control group:
No antipseudomonal treatment

RCT treatment group:
Antipseudomonal antibiotic treatment (intravenous b-lactam and oral ciprofloxacin) between day 1 and 14

Eklöf and Misiakou et al., Clinical Microbiology and Infection, 2022

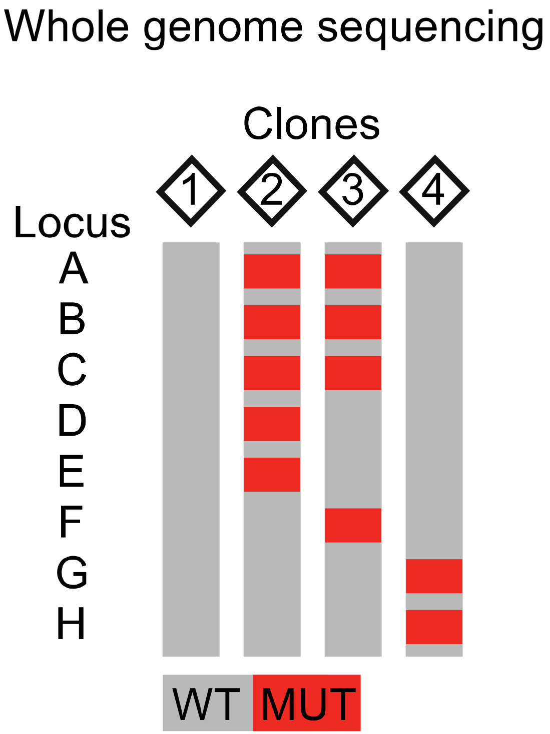
Incidental finding: Close genetic relationship between isolates of COPD Patient 4 and CF Patient P2805 suggests within-hospital transmission

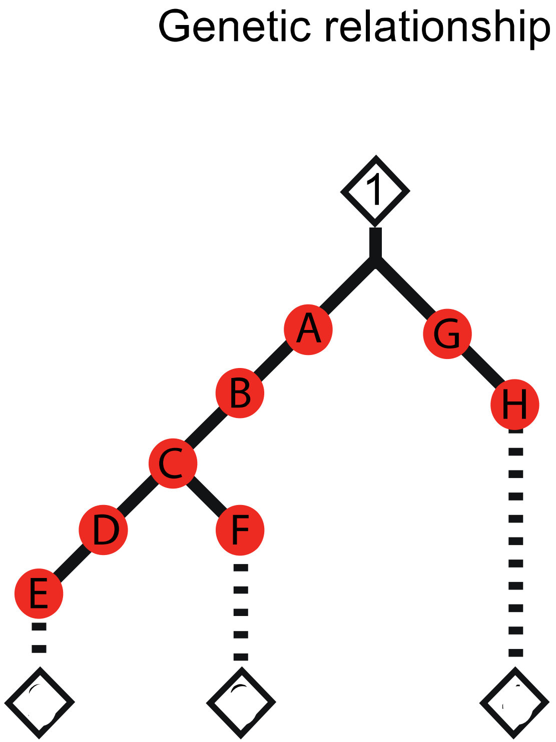


Both patients attended the same department at Rigshospitalet around the time of detection

Eklöf and Misiakou et al., Clinical Microbiology and Infection, 2022

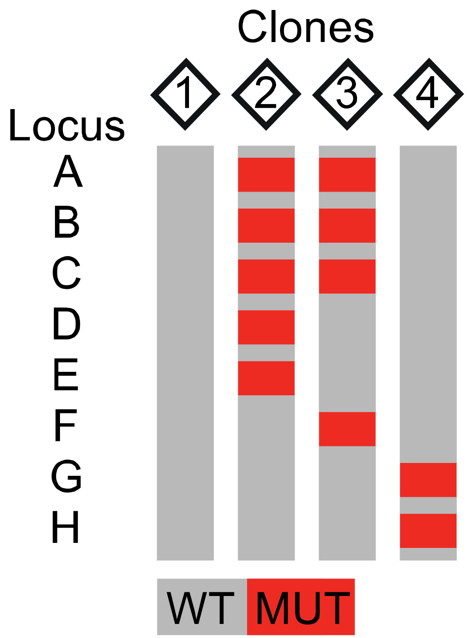
Phylogenomics: Inference of genetic relationship based on genomes



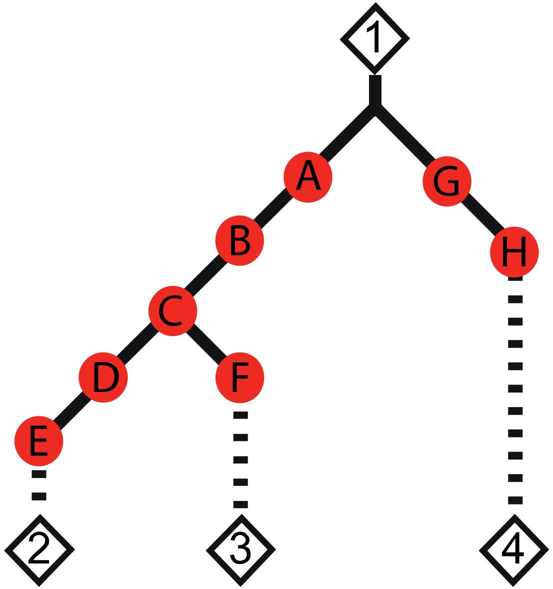


Phylogenomics: Inference of genetic relationship based on genomes

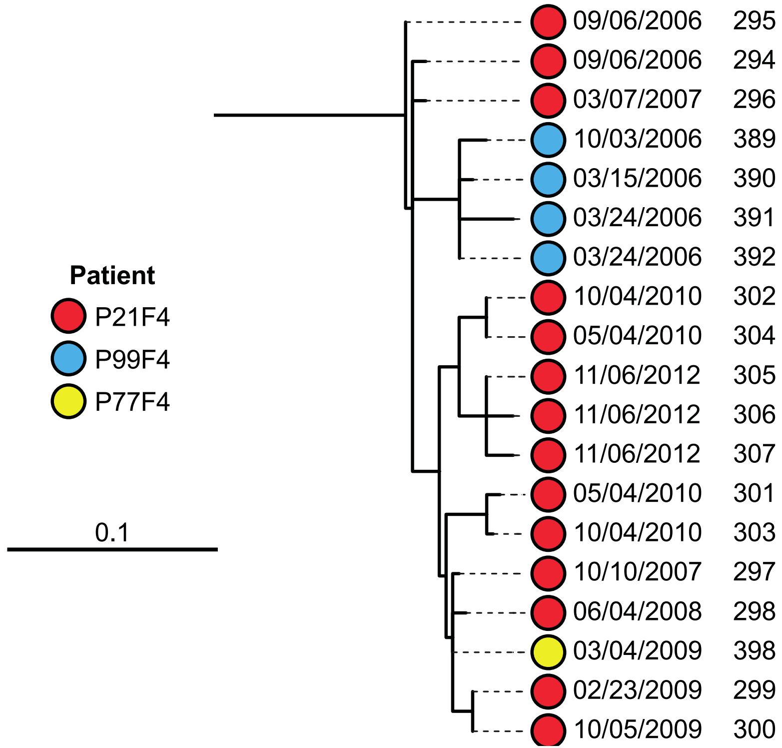
Whole genome sequencing



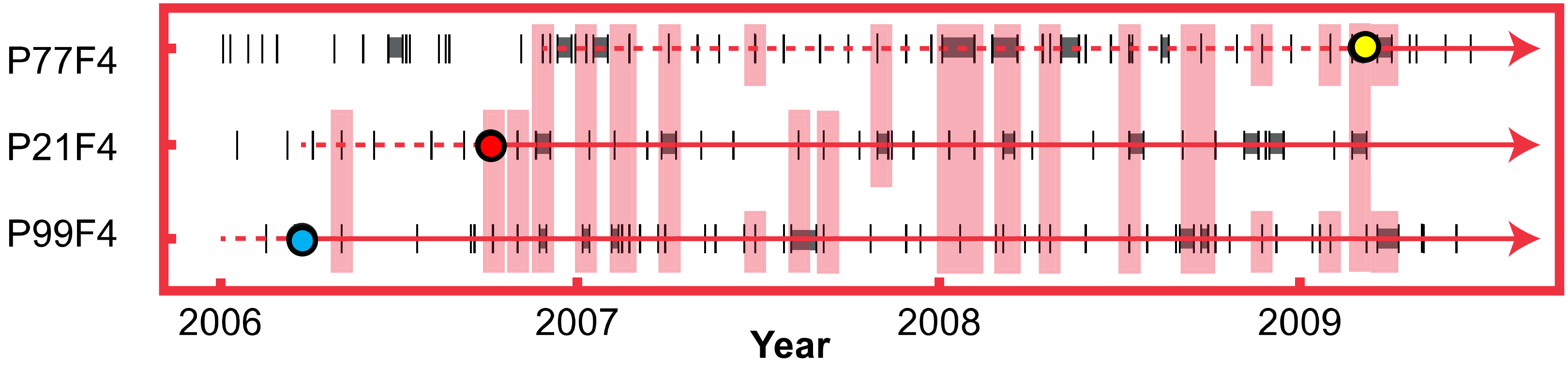
Genetic relationship



Structure phylogenetic tree can be used to infer direction of transmission: Example with <13 SNPs between patients



Temporal overlaps in the patient's hospital visits suggest that transmission took place at the hospital



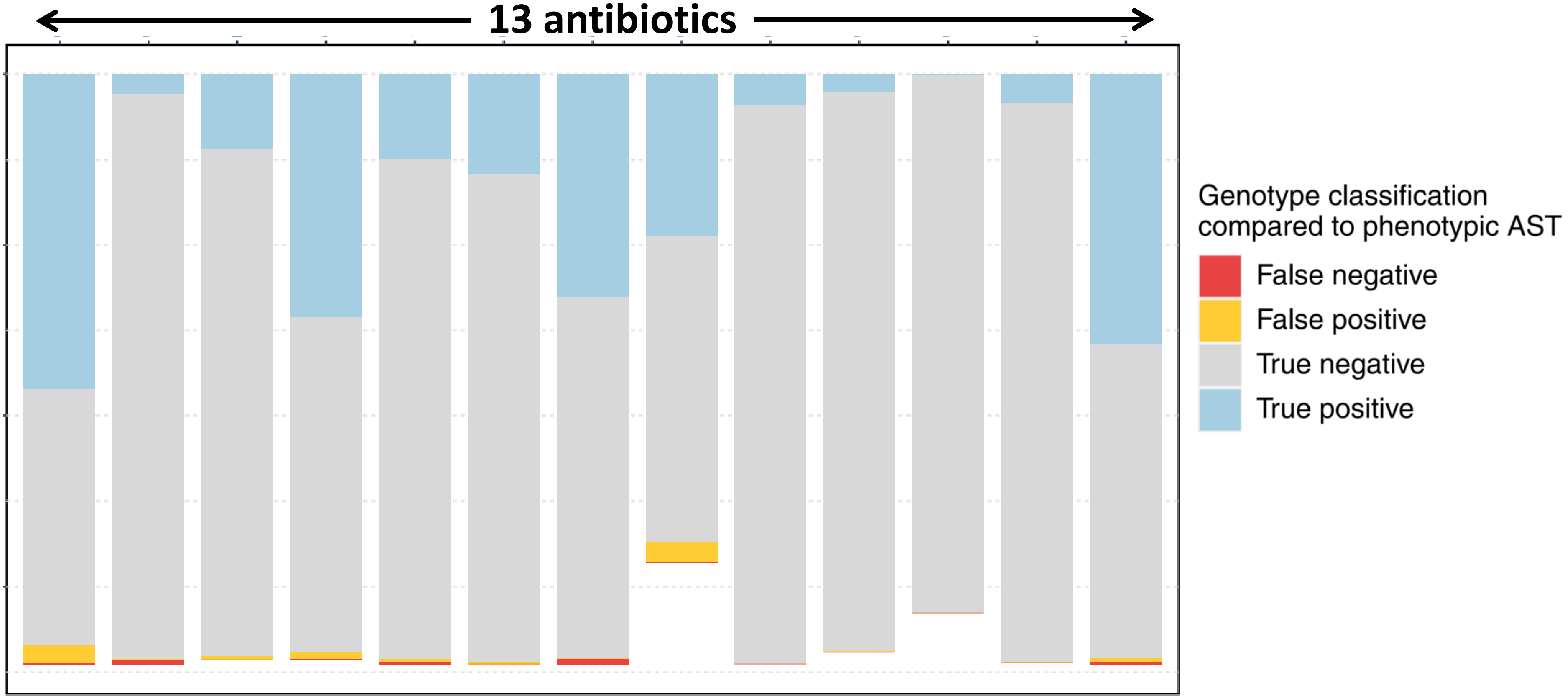
Fundamental diagnostic questions in clinical microbiology

Is there something ?

What is it ?

What can it do ?

98.9% accuracy of genotypic prediction of AST phenotype for 864 *Salmonella* spp. isolates (Sherry, Nature Communications, 2023)



Antibiotic determinants observed in a collection of genomes including only one *P. aeruginosa* genome per person (2,089 genomes)

AMRFinderPlus mutations

Observations (count)	Gene	Mutations	Drug class for resistance
<=1	Not shown		
2	*mexZ*	G195E,Q134STOP	BETA-LACTAM
5	*ampR*	D135G,D135N	BETA-LACTAM
5	*ampD*	H157Y,Q155STOP,Q44STOP,H77Y	BETA-LACTAM
6	*phoQ*	V260G,LG364del	COLISTIN
22	*fusA1*	A555E,E100G,R371C,T456A,Y552C,P618L, R680C,T671A,P554L	AMINOGLYCOSIDE
30	*parC*	S87W,S87L	QUINOLONE
31	*pmrB*	A54V,H340R,A248T	COLISTIN
33	*gyrB*	E468D,S466Y,S466F	QUINOLONE
41	*ftsI*	A454V,A539T,F507L,Q458R,V471G,F533L, N427S,H394R,V523M,G63S,P527S,R504C	BETA-LACTAM
46	*gyrA*	Q106L,D87G,T83A,D87Y,D87N	QUINOLONE
49	*oprD*	L11P,S278P, 13 different STOP mutations	BETA-LACTAM

Mutations *gyrA*(T83I), *oprD*(V359L), *parE*(A473V), and *pmrB*(V15I) were ignored in mutation analysis as they were observed in 48 to 983 of the genomes:.
blaOXA- and *blaPDC*-genes were ignored in gene presence analysis.

AMRFinderPlus genes

Observations (count)	Gene	Drug class for resistance
<=4	Not shown	
5	*aadA7*	AMINOGLYCOSIDE
5	*blaNDM-1*	BETA-LACTAM
5	*dfrA1*	TRIMETHOPRIM
5	*rmtB4*	AMINOGLYCOSIDE
6	*aac(6')-Ib3*	AMINOGLYCOSIDE
6	*aadA11*	AMINOGLYCOSIDE
6	*ant(2'')-Ia*	AMINOGLYCOSIDE
6	*blaIMP-1*	BETA-LACTAM
6	*tbtB*	EFFLUX
6	*tbtM*	EFFLUX
8	*aadA6*	AMINOGLYCOSIDE
10	*floR2*	PHENICOL
10	*qnrVC1*	QUINOLONE
10	*tet(G)*	TETRACYCLINE
21	*aph(6)-Id*	AMINOGLYCOSIDE
27	*aph(3'')-Ib*	AMINOGLYCOSIDE
32	*sul1*	SULFONAMIDE
>=1375	Not shown	

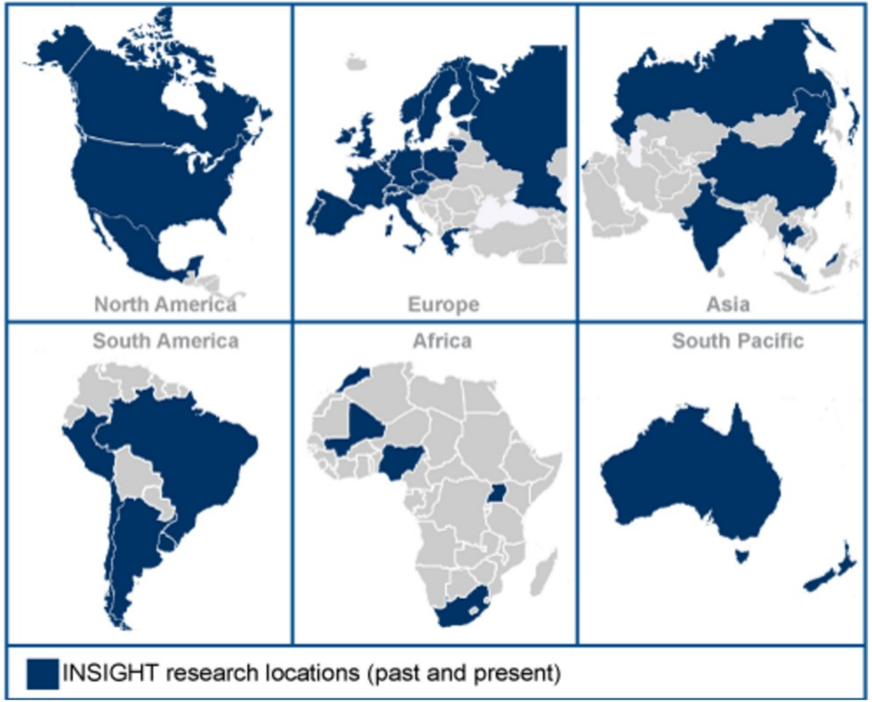
Positive predictive values (PPVs) for resistant phenotype in *in vitro* antibiotic susceptibility test

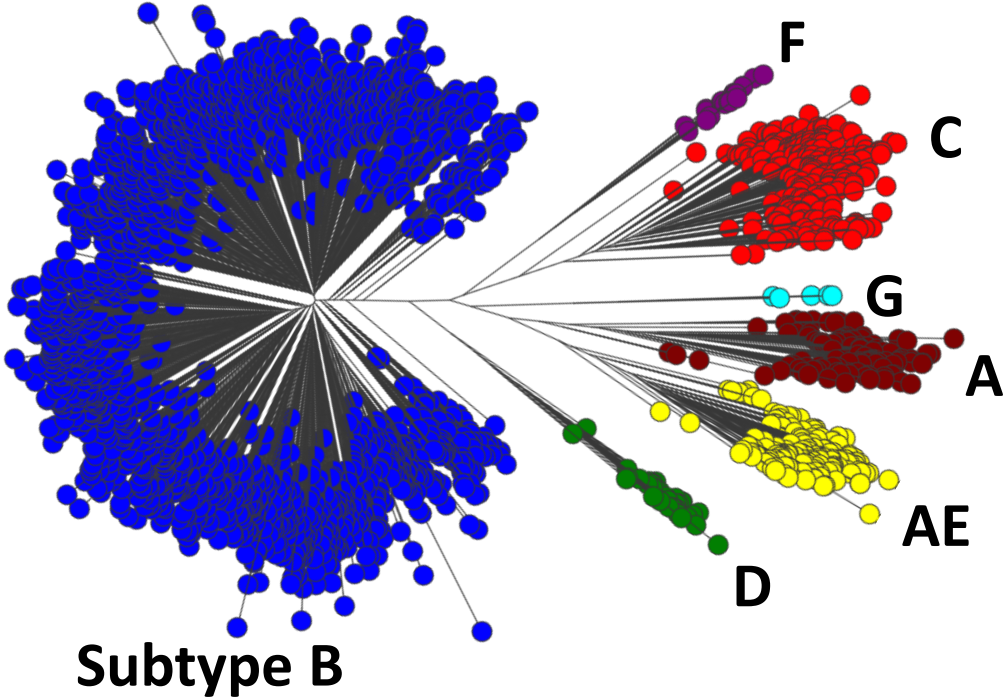
AMRFinderPlus mutations

Observations (count)	Gene	Mutations	Drug class for resistance	
<=1	Not shown			
2	*mexZ*	G195E,Q134STOP	BETA-LACTAM	
5	*ampR*	D135G,D135N	BETA-LACTAM	
5	*ampD*	H157Y,Q155STOP,Q44STOP,H77Y	BETA-LACTAM	
6	*phoQ*	V260G,LG364del	COLISTIN	**PPV 20% for Colistin resistant (40%)**
22	*fusA1*	A555E,E100G,R371C,T456A,Y552C,P618L, R680C,T671A,P554L	AMINOGLYCOSIDE	**PPV 100%/55% for Gentamicin/Tobramycin resistant (100%/64%)**
30	*parC*	S87W,S87L	QUINOLONE	**PPV 97% for Ciprofloxacin resistant (100%)**
31	*pmrB*	A54V,H340R,A248T	COLISTIN	**PPV 19% for Colistin resistant (43%)**
33	*gyrB*	E468D,S466Y,S466F	QUINOLONE	**PPV 33% for Ciprofloxacin resistant (64%)**
41	*ftsI*	A454V,A539T,F507L,Q458R,V471G,F533L, N427S,H394R,V523M,G63S,P527S,R504C	BETA-LACTAM	**PPV 76% for Ceftazidime resistant (83%)**
46	*gyrA*	Q106L,D87G,T83A,D87Y,D87N	QUINOLONE	**PPV 37% for Ciprofloxacin resistant (65%)**
49	*oprD*	L11P,S278P, 13 different STOP mutations	BETA-LACTAM	**PPV 56% for Meropenem resistant (75%)**

Parenthesis shows PPV for either resistant (R) or susceptible at increased exposure (I)

Phylogenetic tree of 2,501 HIV-1 genomes from START clinical trial
Samples have been collected in 2009-2013 from 35 countries





Bennedbæk, Virus Evolution, 2021

2,122 persons: Data available for association analysis using generalized linear models

Virus genetics

3,010 AA variants (observed min >=5% of virus genomes)

(adjustment for age, sex, and overall genetic structure in human and virus population)

Biomarkers

Viral load, CD4+, CD8+, hs CRP, IL-6, D-dimer

HIV GWAS studies:
Gabrielaite, Bennedbæk et al., J Infect Dis, 2021
Gabrielaite, PloS Comput Biol, 2023

No significant associations btw virus genetics and biomarkers …

2,122 persons: Data available for association analysis using generalized linear models

Virus genetics

3,010 AA variants (observed min >=5% of virus genomes)

(adjustment for age, sex, and overall genetic structure in human and virus population)

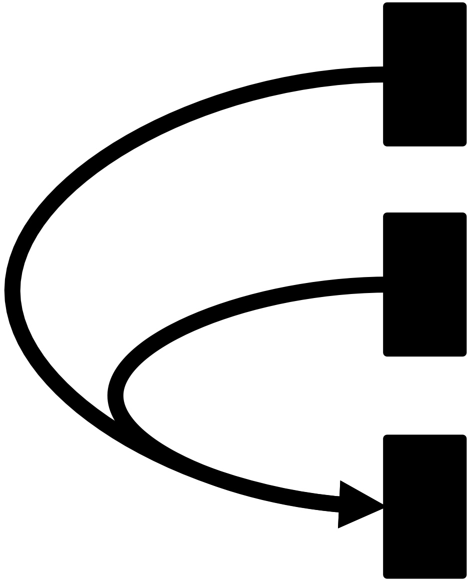
Biomarker

Viral load

Gabrielaite, Bennedbæk et al., J Infect Dis, 2021

Gabrielaite, PloS Comput Biol, 2023

2,122 persons: Data available for association analysis using generalized linear models



Virus genetics
3,010 AA variants (observed min >=5% of virus genomes)

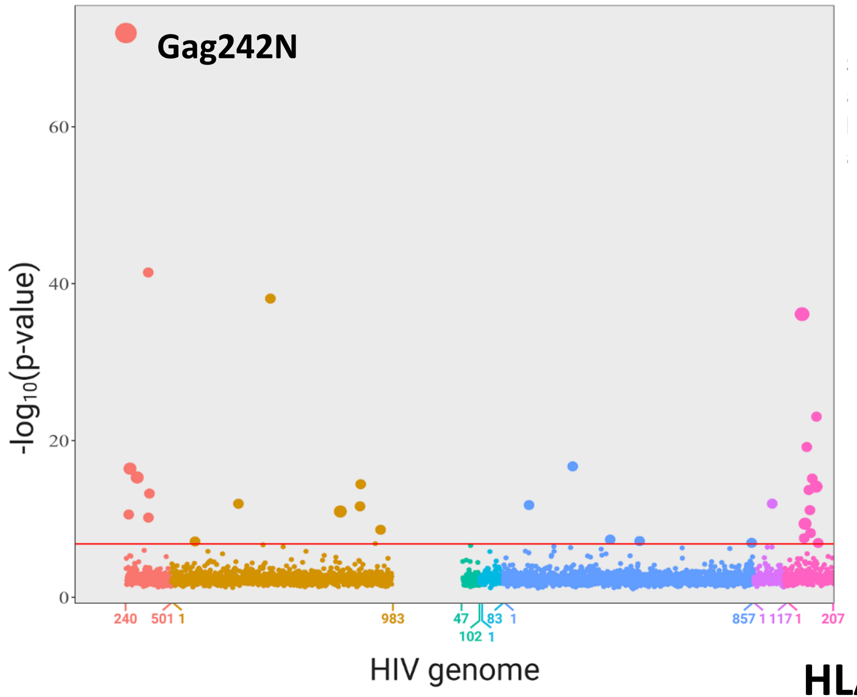
Adjustment for host immune type
388,501 SNPs → 130 HLA alleles

Biomarker
Viral load

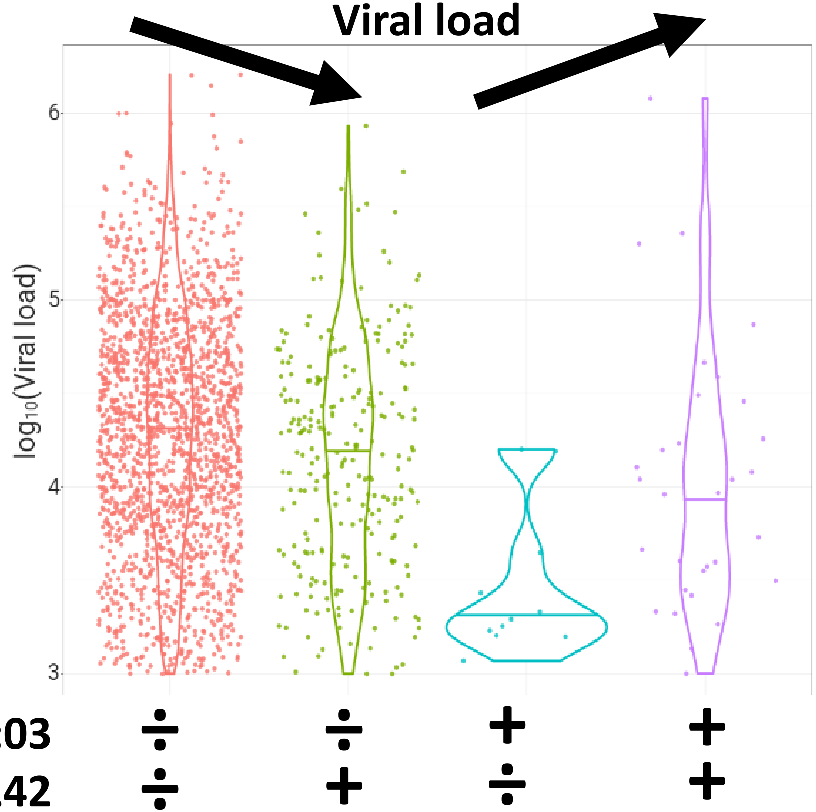
Gabrielaite, Bennedbæk et al., J Infect Dis, 2021
Gabrielaite, PloS Comput Biol, 2023

Interaction between HIV genetic variant and HLA allele (immune type) explains viral load variance

Virus genome to human HLA association



Viral load



Gabrielaite, Bennedbæk et al., J Infect Dis, 2021

Gabrielaite, PloS Comput Biol, 2023

Microbial genomics in a clinical setting can be used for

… for high resolution genetic typing

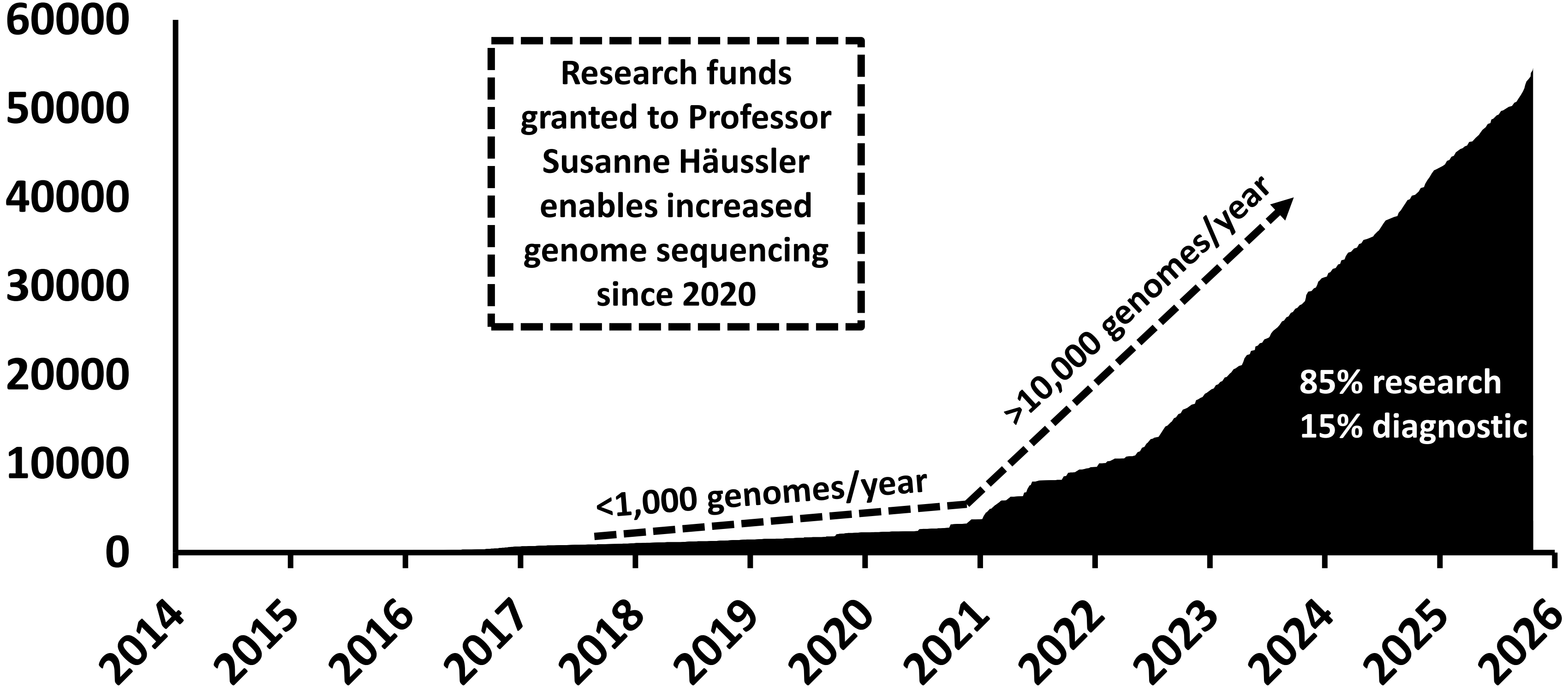
… detection of transmission

… identification of genetic predictive markers of clinical phenotypes

… a lot, but not all and still difficult to translate into clinical action

Hospital directorship: Are genomics cost-effective ?

Bacterial WGS at Rigshospitalet (accumulated count):
54,599 genomes sequenced per 2025-10-31



What is the potential if we sequence the genome of all bacterial isolates ?

ELSEVIER

Clinical Microbiology and Infection
Available online 7 October 2025
In Press, Journal Pre-proof ⓘ What's this?

Original Article

Estimating the potential economic and health impact of integrated genomic surveillance in a hospital setting

Frederik Boetius Hertz [1 2 ¶], Karen Leth Nielsen [1 ¶], Dmytro Strunin [1 ¶], Jelena Erdmann [4 5 ¶], Martin Lucas Jørgensen [3], Theiss Bendixen [3], Roshkan Srinathan [3], Rasmus L. Marvig [6], Steen Christian Rasmussen [1], Asger Nellemann Rasmussen [1], Christian Salgaard Jensen [1], Jenny Dahl Knudsen [1], Susanne Häussler [1 4 5]

Abstract

Objectives

Integrated genomic surveillance, combining whole genome sequencing (WGS) of bacterial isolates with patient movement data, promises improved detection and prevention of pathogen transmission. However, evidence on its cost-effectiveness and clinical utility remains limited, not least because the full extent of transmission in hospital settings is difficult to capture.

Methods

We conducted a 28-month observational study at Rigshospitalet, Copenhagen, collecting patient movement data and sequencing 18,438 bacterial isolates from 7,398 patients across diverse species, infection sites, and resistance profiles. We estimated the hypothetical benefits of implementing integrative WGS surveillance, assuming that continued transmission could be prevented when WGS information was acted upon immediately.

Results

We found that 1,975 of 7,398 of culture-positive hospitalized patients (26.7 %) harboured a pathogen genetically related to another patient´s isolate. 1359 of those (68.8 %) had an epidemiological link in the hospital, with *Enterococcus faecium* by far being the most prevalent involved in transmissions. WGS-informed prevention could hypothetically generate net savings of €1.35 million annually if transmission was stopped once a clonal isolate was detected in a second patient. Furthermore, this approach could potentially have prevented an estimated 1,284 hospital-acquired infections over the 28-month study period, including 94 cases characterized by the recovery of bloodstream isolates.

Conclusions

Our integrated genomic surveillance approach reveals previously unexplored transmission landscapes. We discovered that transmission is widespread, varies between species, and is not limited to resistant isolates. Our results emphasise the potential of integrated genomic surveillance, the identification of local transmission hotspots, potential greater savings by including susceptible isolates, and an incremental cost-effectiveness ratio classification by pathogen species.

PROMISING*: An investigator initiated randomized clinical trial to test if WGS all bacterial isolates improves infection control

******Pr**evention of nosoco**mi**al infections through prospective surveillance of bacterial pathogens by whole genome **s**equenc**ing** – a cluster randomised clinical trial*

Rigshospitalet (RH)

Hvidovre Hospital

Rigshospitalet (RH)	Arm	Hvidovre Hospital
RH Dept A RH Dept B	*Active arm:* *WGS of all isolates*	HH Dept A HH Dept B
RH Dept C RH Dept D	*Control arm:* *WGS upon indication*	HH Dept C HH Dept D

Investigator: Susanne Haussler
Trial contact: Melissa Hornbæk Øvre

Primary trial endpoint:
Does the intervention (WGS-informed infection control) reduce hospital-acquired infections ?

Please help us improve healthcare through research

Thanks for your attention