UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FARMACIA



TESIS DOCTORAL

Development of Omics and Bioinformatics Tools for the Analysis of Bacterial Populations of Public Health Interest

———————

Desarrollo de herramientas ómicas y bioinformáticas para el análisis de poblaciones bacterianas de interés en salud pública

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Miguel Díez Fernández De Bobadilla

DIRECTORES

María Teresa Coque González
Val Fernández Lanza

© Miguel Díez Fernández De Bobadilla, 2024

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE FARMACIA



TESIS DOCTORAL

Development of Omics and Bioinformatics Tools for the Analysis of Bacterial Populations of Public Health Interest

---

Desarrollo de herramientas ómicas y bioinformáticas para el análisis de poblaciones bacterianas de interés en salud pública

MIGUEL DÍEZ FERNÁNDEZ DE BOBADILLA

DIRECTORES

MARIA TERESA COQUE GONZÁLEZ Y VAL FERNÁNDEZ LANZA

PROGRAMA DE DOCTORADO
MICROBIOLOGÍA Y PARASITOLOGÍA

And you, my father, there on the sad height,

Curse, bless, me now with your fierce tears, I pray.

Do not go gentle into that good night.

Rage, rage against the dying of the light.

—Dylan Thomas (1914–1953), Do not go gentle into that good night"

An important scientific innovation rarely makes its way by gradually winning over and converting its opponents: it rarely happens that Saul becomes Paul. What does happen is that its opponents gradually die out, and that the growing generation is familiarized with the ideas from the beginning: another instance of the fact that **the future lies with the youth.**

—Max Planck (1858 - 1947), Scientific autobiography, 1950, p. 33, 97

The result, therefore, of our present enquiry is, that we find no vestige of a beginning, no prospect of an end.

—James Hutton (1727-1797), Theory of the Earth, 1795

I'm tired of Marvin asking me, "What's Going On?

March to the streets 'cause I'm willing and I'm able

Categorize me, I defy every label

And while you're selling dope, we're gonna keep selling hope

We rising up now, you gotta deal, you gotta cope

Will you be electric sheep?

Electric ladies, will you sleep?

Or will you preach?

—Janelle Monae (1985-), Q.U.E.E.N. The electric lady, 2013

Quiero dar las gracias a todas las personas que me han acompañado en la creación y escritura de este trabajo:

Gracias a los impuestos de los trabajadores que han permitido la financiación de este trabajo.

Gracias a los voluntarios que han dado sus muestras de manera altruista para hacer avanzar el conocimiento.

Gracias a los trabajadores del hospital que sin su colaboración hubiera sido imposible.

Gracias a todas las instituciones de enseñanza pública que han hecho mi mundo más grande.

Gracias a mis compañeros que han hecho que todo sea más divertido.

Gracias a mi familia y amigos por el cariño, el apoyo y la preocupación.

Gracias a Teresa y Val porque hemos llegado hasta aquí y, sin embargo, os tengo aún más aprecio y consideración que el primer día.

Gracias a Cristina y Alba por haber sido parte indispensable de esta tesis.

También me gustaría aprovechar este espacio para reflexionar sobre nuestra responsabilidad como científicos con la sociedad y el planeta. Ninguna de nuestras investigaciones es neutra ni apolítica. Todos nuestros actos tienen un impacto y pueden ayudar a perpetuar situaciones de opresión o a crear otras nuevas, como bien atestiguan las vidas que han sido daños colaterales en nombre del Desarrollo. El progreso o el avance del conocimiento no debe justificar ningún tipo de maltrato o injusticia. La ciencia ha sido y sigue siendo cooptada por el sistema económico, social y político para legitimarse y perpetuarse. Sin embargo, la ciencia también sigue siendo una de las herramientas para liberarnos de esta violencia. Espero que el dolor nunca nos sea ajeno, que nos incite a salir de las torres de marfil y que nos haga luchar por ayudar a construir un sistema mejor.

eppur si muove

ABSTRACT

Precision Public Health represents an evolving approach to disease treatment and prevention that considers individual variability in genes, environment, and lifestyle. However, existing limitations in tools for analyzing microbial communities, subhierarchical units (such as clones, extrachromosomal elements, and genes), and complex data have impeded the validation of hypotheses and hindered the implementation of Precision Public Health.

This thesis focuses on developing omics tools for assessing the abundance and diversity of microbial populations. Concurrently, bioinformatics tools have been validated for analyzing extensive and diverse databases, encompassing omics data, various datasets, and clinical information. The developed bioinformatic tools enable the analysis of bacterial genomes, categorizing their diversity at the species and intraspecies levels. These tools contribute to estimating diversity within species, validating existing nomenclature, and enhancing the understanding of pathogen adaptation.

Additionally, bioinformatic tools have been devised to analyze accessory genomes, shedding light on their impact on species definition and facilitating the categorization of mobile genetic elements such as plasmids. The development of bioinformatic tools for analyzing metagenomes, specifically resistomes, aims to enhance the analysis of gene populations conferring resistance to antimicrobials, heavy metals, and biocides, including relaxases as indicators of conjugative extrachromosomal elements (mobilome).

Furthermore, omic tools have been crafted to harmonize multiple databases, encompassing public databases, microbiological data (omics data), and clinical data (clinical and epidemiological data). The ultimate goal is to establish portable and user-friendly pipelines for analyzing metadata from diverse sources.

RESUMEN

La Salud Pública de Precisión representa un enfoque en constante evolución para el tratamiento y la prevención de enfermedades que tiene en cuenta la variabilidad individual en genes, ambiente y estilo de vida. Sin embargo, las limitaciones existentes en las herramientas para analizar comunidades microbianas, unidades subjerárquicas (como clones, elementos extracromosómicos y genes) y datos complejos han dificultado la validación de hipótesis y obstaculizado la implementación de la Salud Pública de Precisión.

En esta tesis, el enfoque se centra en el desarrollo de herramientas ómicas para evaluar la abundancia y diversidad de las poblaciones microbianas. Simultáneamente, se han validado herramientas de bioinformática para analizar bases de datos extensas y diversas, que incluyen datos ómicos, diversos conjuntos de datos e información clínica. Las herramientas de bioinformática desarrolladas permiten el análisis de genomas bacterianos, categorizando su diversidad a niveles de especie e intraspecíficos. Estas herramientas contribuyen a estimar la diversidad dentro de las especies, validar la nomenclatura existente y mejorar la comprensión de la adaptación de patógenos.

Además, se han ideado herramientas de bioinformática para el análisis de genomas accesorios, arrojando luz sobre su impacto en la definición de especie y facilitando la categorización de elementos genéticos móviles como plásmidos. El desarrollo de herramientas de bioinformática para analizar metagenomas, específicamente resistomas, tiene como objetivo mejorar el análisis de poblaciones génicas que confieren resistencia a antimicrobianos, metales pesados y biocidas, incluidas las relaxasas como indicadores de elementos extracromosómicos conjugativos (mobiloma).

Adicionalmente, se han creado herramientas ómicas para armonizar múltiples bases de datos, que incluyen bases de datos públicas, datos microbiológicos (datos ómicos) y datos clínicos (datos clínicos y epidemiológicos). El objetivo final es establecer *pipelines* fáciles de usar para analizar metadatos de diversas fuentes.

ABBREVIATIONS

UMI: Unified Microbiome Initiative

UN: United Nations

OTU: objective taxonomic units

TCP: Targeted capture approaches

WGS: whole-genome sequencing

AMR: Antimicrobial Resistance

OUS: Operational Units of Surveillance

AMU: Antimicrobial Use

ARGs: Antimicrobial resistance genes

AMC: Antimicrobial consumption

ANI: Average nucleotide identity

K-NNN: k-nearest neighbor network

MRGs: Genes encoding resistance to metals

BRGs: Genes encoding resistance to biocides

PCA: Principal Component Analysis

UMAP: Uniform Manifold Approximation and Projection

T-sne: t-distributed Stochastic Neighbor Embedding

GLM: general linear model

IA: Infectious agent

INDEX

1. INTRODUCTION 21
1.1 A Brief Historical Perspective of Microbial Categorization 21
1.2 The advent of a new field 22
1.2.1. The origins of computer science 22
1.2.2. Computation applied to biology 23
1.3 Omics Tools 24
1.3.1. Metagenomics as innovation driver 24
1.3.2. Weaknesses and Limitations of the Metagenomic Toolkit 24
1.4 Measuring the microbiome 26
1.4.1. Units of analysis and indicators 27
1.4.2. Validity, reliability, comparability 29
2. HYPOTHESIS AND OBJECTIVES 35
3. MATERIALS AND METHODS 41
3.1 Cohorts and samples analyzed 41
3.2 Genomics 43
3.2.1. DNA extraction and preparation 43
3.2.2. Genomic analysis 44
3.2.3. Genomic statistics 45
3.3 Metagenomics 46
3.3.1. DNA extraction and preparation 46
3.3.2. Sequence capture platform 46
3.3.3. Targeted Metagenomics Workflow 46
3.3.4. Metagenomic analysis 47
3.3.5. Metagenomic Statistics 48
3.4 Bioinformatic software development 49

3.4.1. PATO: Pangenome Analysis Toolkit....49
3.4.2. BACTAX-ID....50
4. RESULTS....55
4.1. Development of bioinformatic tools for in-depth metagenome analysis....55
4.1.1. In-depth metagenomic analysis of intestinal microbial communities from Amerindians to address the "baseline human resistomes." 56
4.1.2. In-depth metagenomic analysis of "the hospital resistome" to address the selection and transmission of antimicrobial resistance....66
4.2. Development of bioinformatic tools for in-depth genome analysis....70
4.2.1. PATO (Pangenome Analysis Toolkit)....70
4.2.2 BACTAX-ID....123
4.3. Application of Genomic Tools to the Analysis of Bacterial Species and Plasmids135
4.3.1. Enterococcus....136
5. DISCUSSION....143
5.1 Metagenomic tools to analyze resistomes....143
5.2 Bioinformatic analysis for in-depth genomic analysis....145
5.3 Bioinformatic analysis for in-depth species analysis. The BacTaxID....147
5.4 Analysis of bacterial species....148
6. CONCLUSIONS....153
7. REFERENCES....159
8. ANNEXES....169
Annex 1. Analysis pipeline....169
Annex 2. Statistics data frame....169
Annex 3. Poster and oral presentations submitted to national or international conferences....169

1. INTRODUCTION

1. INTRODUCTION

1.1 A Brief Historical Perspective of Microbial Categorization

Microbial taxonomy and phylogeny were initiated in the 19th century, but their development was slow due to the limitations of tools and the absence of universal criteria to establish biological diversity. It was not until well into the 20th century that these fields progressed significantly. In 1923, David Hendricks Bergey published the first treatise on microbial taxonomy at the request of the American Society of Bacteriologists, which he then chaired (1). The objective of Bergey's Manual was to incorporate progressive knowledge in the field and facilitate work in this discipline. based on the biochemical and physiological characteristics of microorganisms.

In 1977, Carl Woese, using genetic sequences of DNA encoding RNA (16S rRNA in prokaryotes and 18S rRNA in eukaryotes), published a new "tree of life," establishing the diversity of living organisms into three domains and six kingdoms. It subdivided the kingdom Monera into the kingdoms Eubacteria and another, described initially in this work, called Archaea (2). This classification sparked intense debate with public opposition and disagreement from significant microbiologists and evolutionists such as Ernst Mayr (3).

In the same year, Frederick Sanger applied the sequencing method, allowing access to the genetic sequence of microorganisms, for which he received the Nobel Prize in Chemistry in 1980 (4). The spectacular development of sequencing technology, bioinformatics, and statistical tools has given rise to Genomics as an interdisciplinary branch of Molecular Biology, expanding since its inception.

In 1998, microbiology began to be approached from the community perspective ("from one microorganism to many") when Jo Hendelsman coined the term "metagenomics" to define the genomic analysis of a population or community of similar but not identical microorganisms, adding the statistical concept of meta-analysis (5).

Further technical advances in genomics, such as single-cell omics (sc-DNA, sc-RNAseq) or 3D genomics, which combines 3C (chromosome conformation capture-based technologies) and next-generation sequencing, NGS (6) alongside (machine learning or artificial intelligence) are

enabling the transition from a descriptive science to a quantitative one. This scenario will allow us to understand the composition, variation, and interactions between microbial communities and, thus, design strategies to improve and design applications in different fields, preserve their diversity, and ensure ecosystem service maintenance.

Current research and political agendas underscore the strategic value of research in Microbiology and have led to the “Unified Microbiome Initiative”, a global initiative for collaboration between the scientific community and public sectors to study microbial processes and develop new technologies that can accelerate fundamental discoveries about the functions of the microbiome, the capabilities of microbial ecosystems on Earth, and the translation of these results into practical applications (7). Quantitative biotechnology is considered an essential part of what has been called "the Fourth Industrial Revolution" and one of the critical elements to achieving the sustainable development goals set by the United Nations in 2015 (http://www.un.org/sustainabledevelopment/sustainable-development-goal).

1.2 The advent of a new field

1.2.1. The origins of computer science

Since ancient times, the need to automate calculations has been evident among various civilizations, such as the Mesopotamians, ancient Chinese dynasties, and the Greeks (8). Devices like the abacus and advancements in mathematics emerged to cope with the complexities of growing societies, often aimed at managing tax payments efficiently.

However, it wasn't until the 17th century that Gottfried Wilhelm Leibniz introduced binary arithmetic (9). He proposed that every number could be represented using a binary system of 0s and 1s, envisioning it as a potential universal language for computation and communication.

In the 19th century, Charles Babbage designed the "Analytical Engine," which was considered the first precursor to modern computers (10). This groundbreaking invention was intended to handle diverse calculations and execute programs stored on punch cards. Despite Babbage's inability to construct it due to technological and financial challenges, his collaborator Ada Lovelace recognized its transformative potential. Lovelace even devised an algorithm for calculating Bernoulli numbers, providing a visionary glimpse into the future of computer science (11).

The 1970s marked a pivotal moment with the advent of microprocessors. This innovation enabled the development of personal computers, democratizing access to computing and unleashing its potential across various scientific disciplines (12). This era witnessed a profound democratization of technology, laying the foundation for its widespread integration into every aspect of modern science.

1.2.2. Computation applied to biology.

The field of bioinformatics, though relatively recent, began to take shape in the 1970s when Ben Hesper and David B. Searls coined the term "bioinformatics," defining it as the study of informatic processes in biotic systems (13). This conceptualization laid the foundation for a discipline that integrates computational and biological sciences to unravel the complexities of living systems.

Margaret Dayhoff emerged as a pioneer in bioinformatics with her groundbreaking work on the "Atlas of Protein Sequence and Structure." Published during the field's early days, this catalog systematically organized protein sequences and structures, exemplifying efforts to develop tools for organizing, analyzing, and interpreting biological information (14).

The 1980s and 1990s witnessed a significant evolution in bioinformatics, paralleling advancements in DNA sequencing techniques such as Sanger sequencing (15). These breakthroughs enabled the rapid and accurate discovery of new sequences, driving the development of algorithms and programs to analyze vast amounts of genomic data. Notably, Altschul et al.'s creation of the Basic Local Alignment Search Tool (BLAST) in 1990 revolutionized sequence comparison, offering a fast method to assess sequence similarity (16).

The emergence of newer sequencing techniques has propelled bioinformatics into every corner of the biological sciences. The vast amount of data generated and archived in public databases has spurred the development of innovative methods to manage and analyze large datasets (17). This accessibility to extensive biological information has democratized research, opening doors for scientists to explore new frontiers and make groundbreaking discoveries in diverse fields of biology.

1.3 Omics Tools

1.3.1. Metagenomics as innovation driver

The term "metagenomics" broadly refers to the genomic analysis of a population or community of microorganisms. Coined in 1998 by Jo Hendeslman, this term encompasses the analysis of a collection of biologically similar but not identical entities, incorporating the statistical concept of meta-analysis (18). In a world deeply intertwined with the Microbiosphere, unraveling microbial communities' intricacies in natural environments is a significant challenge in 21st-century science. This pursuit offers innovative solutions to practical challenges across various fields of exceptional socioeconomic importance, such as medicine, agriculture, food science, nutrition, sustainability, energy, biodiversity exploration, forensics, ecology, bioremediation, and numerous industrial applications.

1.3.2. Weaknesses and Limitations of the Metagenomic Toolkit

Metagenomic methodologies are increasingly employed for the analysis of diverse microbial ecosystems, including the exploration of antibiotic resistance gene pools, commonly referred to as “resistomes” (19–23). Metagenomics aims to comprehensively represent microbial content in a sample, encompassing high and low-abundance microorganisms. However, these sequence-based approaches face limitations in detecting sequences near or below their detection thresholds (sensitivity). The short length of sequences also precludes the identification of allelic gene variants. This is relevant when discriminating variants of relevance in Public Health (i.e., gene alleles with different susceptibility to antimicrobials or different adaptive phenotypes) (specificity) (24,25). Both sensitivity and specificity are crucial attributes for effectively prioritizing the risks associated with acquiring and transmitting antibiotic-resistance genes in metagenomes (26,27). The challenges impeding the full utilization of metagenomics potential are outlined in Table 1.

Table 1. Critical limitations of current high-throughput metagenomic technologies. QC: quality control; OTU: objective taxonomic units

Process Step	Limitations and Difficulties
Sample preparation and QC	DNA/RNA quality important for all approaches
	Amplification of difficult quantitation and detection of minority populations (low DNA concentration in samples)
	Difficulty in multiplex samples in the same array
	Normalization between samples
Sequencing strategy	Results skewed by dominant populations
	Coverage/breadth (no. of different genes detected)
	Sampling depth (no. of sequence or OTUs per gene)
	Quantitation
Data Processing	Raw data processing
	Taxonomic resolution
	Phylogeny
	Signal threshold
	Requires a priori knowledge
	Analysis of a diversity (referred to a single gene)
	Data comparison between samples

Most metagenomic approaches fail in detecting and characterizing taxa with low relative abundance, which include “knowns”, “known unknowns” (bacteria under the culturable threshold), or “unknown unknowns” (unculturable bacteria) microbial units (28). Despite their low densities, these taxa may hold crucial ecological or pathogenic roles, serving as genetic and

functional diversity reservoirs with strategic industrial, economic, or social significance. For samples containing sequences at low concentrations, such as minority or rare populations in complex samples, capture-based or targeted sequencing metagenomics strategies became essential.

Targeted capture approaches (TCP) have emerged as efficient and cost-effective alternatives, offering high-throughput solutions to overcome the limitations of whole-genome sequencing (WGS) and multiplex PCR. In-solution targeted capture platforms, representing the next generation, provide significant technical advancements compared to array-based platforms and other genome partitioning methods. These platforms have become standard not only for well-defined industrial objectives, such as in the fermentation industry, but also hold tremendous potential for advancing environmental and ecological studies, particularly in biodiversity research. TCP offers scalability, cost-effectiveness, and enhanced data quality, including reduced variance in target coverage, more accurate SNP calling, higher reproducibility, and longer assembled contigs. They are particularly advantageous over other metagenomic tools when analyzing large sample sizes or when preselecting data/traits for analysis, such as addressing detailed population genetic questions or identifying rare variants associated with phenotypes (29). With their capability to isolate specific sequences from a complex blend of DNA from various organisms, TCP demonstrates remarkable versatility, applicable not only to metagenomic studies but also to health, animal, industrial, or environmental surveys. Moreover, they find utility in research involving host-parasite or pathogen interactions (personalized medicine, personalized nutrition), as well as in fields working with potentially ancient/damaged DNA (forensics, legal, biodefense, conservation, environment).

TCP has found extensive success in human medicine for detecting genetic variants associated with diseases, providing a competitive edge in human diagnostics for several prominent companies in the US (Affymetrix, IDT, NimBleGene, Roche Diagnostics), or the European Union (Haloplex).

1.4 Measuring the microbiome

Identifying the relative contributions of the host, the microbial threat, and their environments to the acquisition and the transmission of infectious agent (IA) or antimicrobial resistance (AMR) units requires a careful selection of the *indicators* and the sample (*unit of analysis*).

The available information on IAs and/or AMR came from relatively few sources using similar *indicators* and methodologies However, a reflection on the *validity* and *reliability* of the AMR information to be applied in various One Health sectors (human, animal, environment), demographic, and economic contexts (local, regional, global, and inter-sectorial levels) has been neglected to date.

1.4.1. Units of analysis and indicators

This section defines the ***unit of analysis*** and ***indicators*** used in AMR and discusses the factors that affect their *validity* and *reliability* (Chart 1).

The unit of analysis. Microbial infectious diseases and AMR disproportionally impact populations and settings, and the "who" or "what" that might occur needs analysis. The *units of analysis* are human groups (such as single individuals, social, professional, or health-risk groups), healthcare settings (tertiary, secondary, primary, or single wards), and economic sectors (One Health, economic growth sectors). Age, sex, body health condition, genetics, and the *exposome* influence host susceptibility, host immunity, and social behavior. The classical high-risk populations prone to acquire IAs or AMRR units are individuals with severe underlying diseases, immunocompromised conditions (children and elderly), living in poor socioeconomic structural environments, or significantly exposed to factors that increase the risk of selecting or spreading AMR (e.g., invasive procedures).

The *indicators* refer to the percentage of the microbial threat at particular sites. They include some IAs (sentinel microorganisms) and AMR units (clones, plasmids, genes).

Sentinel microbial species. Only a few bacterial species acquire and successfully spread adaptive genes of medical relevance, such as antibiotic resistance genes (ARGs). They constitute the "*watch lists*" or "*priority lists*" *of microbial threats* of the WHO and the Centers for Disease Control and Prevention (CDC). These lists include a few traditional IAs (*Neisseria gonorrhoeae*, *M. tuberculosis*, *Streptococcus group A*, *Streptococcus pneumoniae*), commensal opportunistic pathogens (Enterobacterales, enterococci, staphylococci), or environmental organisms (*Pseudomonas aeruginosa* and *Acinetobacter baumannii*). The Division of Healthcare Quality Promotion at the CDC reported that non-Legionella water-related opportunistic pathogens cause more than one-fifth of HAIs and hospital outbreaks. They are Enterobacterales already included in the above-mentioned

watch lists and other increasingly reported opportunistic pathogens (*Elizabethkingia*, *Achromobacter*, and *Burkholderia*) (30).

The heterogeneity within a species has been neglected in the surveillance and monitoring of IAs. Again, a limited number of subpopulations within each of the species mentioned, called high-risk clonal complexes (HiRCCs) (31) or "pandemic clones" (32), cause HAIs or are involved in clinical outbreaks. HiRCCs are increasingly detected in healthy humans, which suggests they are part of the commensal flora and eventually spread to close contacts when the population size increases and the shedding is high (33). Genomic typing to track HiRCCs of sentinel species is progressively implemented at local (hospitals, farms), national, and international surveillance programs.(28,34,35).

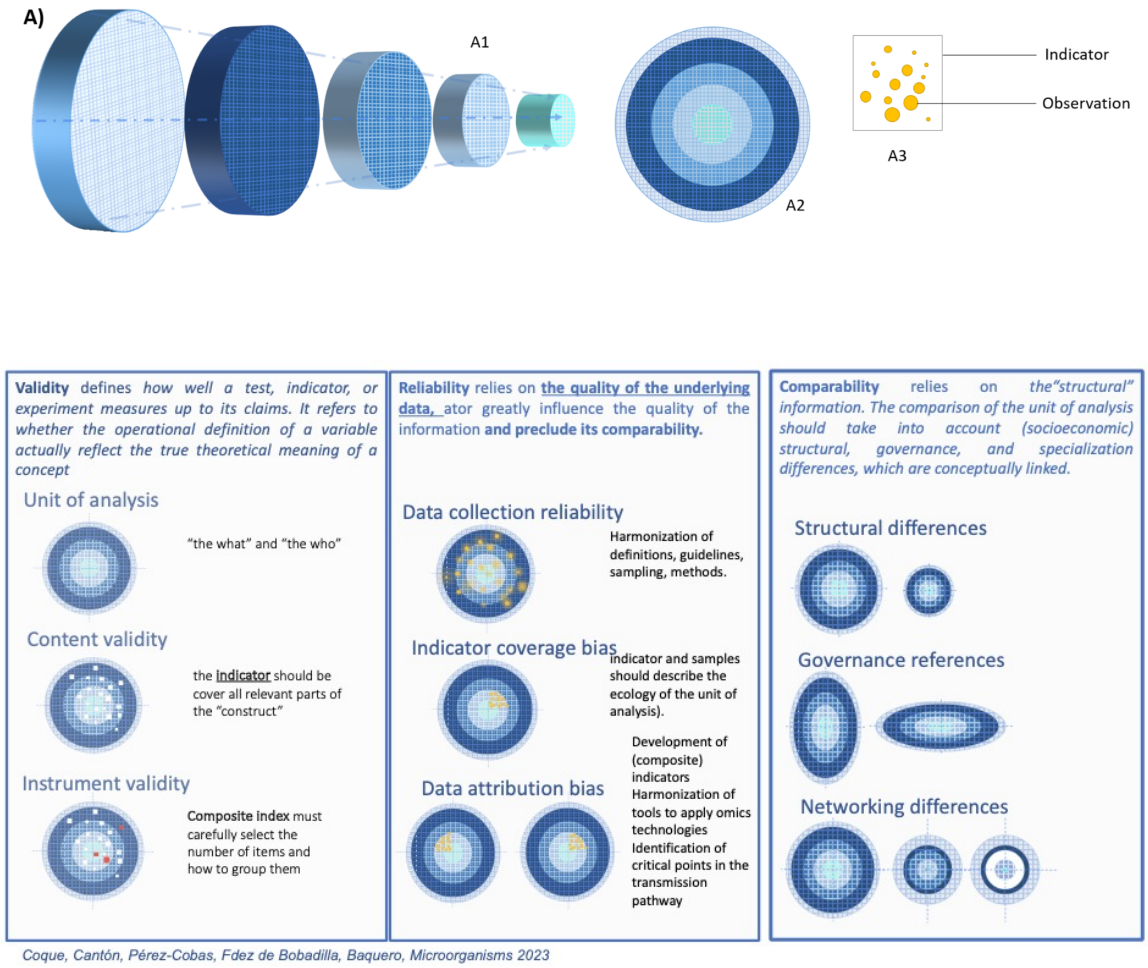
Genes, plasmids, and mobile elements. Adaptive genes (i.e., ARGs) can spread across bacterial species, genera, or microbial communities from humans, animals, and environmental reservoirs through transposable elements, often embedded in plasmids. Again, only a few members of mobile genetic elements (plasmid, transposons, integrons) are responsible for the spread of ARGs or other adaptive traits of medical relevance. However, its diversification (e.g., F plasmids in Enterobacterales) and promiscuity often result in mosaic elements that difficult their identification and, thus, the dynamics of ARGs and plasmids in microbiomes beyond the analysis of every single entity.

Microbiomes and resistomes. Metagenomics-enabled surveillance methods are still in their infancy, although they may improve the detection and characterization of microbial communities associated with IAs and AMR units (36). The analysis of the intestinal and respiratory microbiomes and the resistomes may also allow clarifying fundamental questions, such as the host's susceptibility to infection, therapeutic regiments, fecal transplantation, long-term AMR colonization, or health response to antimicrobial disturbances (28). The microbiome of the hospital-built environment is also an area of increasing interest (37). Knowledge in this area is still scarce, as well as the metagenomic tools to analyze microbiomes and resistomes (38,39).

In microbial epidemiology and Public Health, the proportion of hospital-acquired infections (HAIs) cases is a marker for significant emergent threats. Selectors such as antimicrobial use (AMU) and/or consumption (AMC) have also been used as indicators (28).

1.4.2. Validity, reliability, comparability

In public health microbiology, different indicators are used alone or in combination. The pitfalls of multi-indicator analysis can be grouped into the categories of scope, data quality, and comparability (Chart 1). Although it is widely used in the social sciences, an evaluation of multi-indicators' suitability in decision-making has been neglected in the AMR field.



Footnote: Panel (A) Unit of analysis 1, indicators, and observations. Circles represent the different healthcare levels targeted by surveillance systems (primary, secondary, and tertiary). Microorganisms 2023, 11, 1050 (13 of 32). A1. The nested healthcare networks. From the left to the right: public health and sanitation services, primary health sector, secondary health sectors, tertiary health sectors, and intensive care units (or any high-risk hospital areas). These sectors comprise multiple patches between them and those of other levels. There exist feedback loops between them. A2. A multilayered concentric circle exemplifies challenges in AMR measurement in panel (B). A3. A white square represents a simple indicator. Each yellow circle indicates a set of observations. The size of the circle is proportional to the number of observations. Glow yellow circles reflect an unrepresentative sample [175– 182]. 1. The unit of analysis = what and whom to measure (individuals and populations) within each level; 2 The indicator ="operational units of surveillance", OUS) (Sentinel organisms, genes, GEs. Resistomes, and microbiomes). AMU/AMC; 3 The composite indicators (Drug resistance indexes and Metagenomic indexes). Diagrams inspired in [182].

Chart 1. Challenges of AMR measurements. The validity, reliability, and comparability of the information.

Resistance (AMR) is a natural and ancient phenomenon that resembles the phylogeny of bacterial populations (21,40). This suggests that emerging selective pressures will favor pre-existing resistance determinants that would have been circulating within the microbial pangenomes for millennia.

The resistome concept, the ensemble of antibiotic resistance genes (ARGs) present in microbiomes, is a framework for understanding antibiotic resistance's origin, emergence, and evolution (20,22). The diversity and abundance of ARGs in particular hosts or environmental microbiomes are influenced by the historical interactions with microbial antibiotic producers (21), interactions with transferable ARGs present in microbiomes, and the more recent anthropogenic-induced selective pressures, which included exposition to disparate antimicrobial agents (antibiotics, biocides, metals, and other pharmaceuticals and residues from late industrialism and lifestyle antimicrobial practices and habits) (21,41,42). The resistomes of humans, animals, and environmental sources have been increasingly analyzed in recent years (41,43,43). Several studies also evaluated the effect of antibiotics and other factors in the gut resistomes of Western human populations. However, the resistomes of “pristine” or “minimally impacted environments” by anthropogenic activities are still exceptional (42), and baseline data of the environmental microbiotas before perturbations are still scarce. This biased knowledge precludes distinguishing “background” (intrinsic levels of ARGs in each environment) from “contamination” (levels resulting from human activity) as well as quantifying and monitoring changes after specific interventions or activities.

2. HYPOTHESIS AND OBJECTIVES

2. HYPOTHESIS AND OBJECTIVES

Microbiology Public Health demands accurately identifying bacterial species and subhierarchical elements of epidemiological relevant (clones, plasmids, genes) to track, control, and prevent communicable diseases and other microbial health threats such as antimicrobial resistance ." (44,45). However, the limitations in omic and bioinformatic tools for accurately analyzing the abundance and the diversity of microbial communities and accessory genomes and any of their sub-hierarchical structures (clone and extrachromosomal elements such as plasmids, transposons, and genes) have hindered the validation of hypotheses using theoretical models and, consequently, the implementation of Precision Public Health approaches. *The specificity and sensitivity of available genomic and metagenomic tools preclude the accurate analysis of species, accessory genomes, and gene populations*.

The **MAIN OBJECTIVE** of this thesis is the ***development of omics tools*** for analyzing the abundance and diversity of microbial populations and developing and validating ***bioinformatics tools*** for analyzing large and/or multiple databases (omics data, different databases, and/or clinical data).

The **SPECIFIC AIMS** are

1. Develop and validate ***bioinformatic tools to analyze bacterial genomes*** and categorize their diversity *at species and intraspecies levels*. These tools will be applied to estimate the diversity within species, validate available nomenclature, and understand the adaptation of specific pathogens.

2. Develop and validate ***bioinformatic tools to analyze accessory genomes.*** These tools will be applied to estimate the impact of the accessory genomes in species definition and to categorize mobile genetic elements such as plasmids.

3. Develop and validate ***bioinformatic tools to analyze metagenomes (resistomes).*** This objective aims to improve the analysis of targeted metagenomic tools (ResCap, developed by our group (29)) to analyze gene populations conferring resistance to antimicrobials such as antibiotics (resistomes), heavy metals (metalomes), or

biocides and also relaxases as a proxy of conjugative extrachromosomal elements (mobilome).

4. Development and application of ***a multi-omic integration pipeline***, including public databases, microbiological data (multi-omics data), and clinical data (clinical and epidemiological data). The goal is to establish portable and easy-to-use pipelines to analyze metadata in combination with omics data from different sources.

3. MATERIALS AND METHODS

3. MATERIALS AND METHODS

3.1 Cohorts and samples analyzed

The studies described in Chapter 4.1 analyzed several cohorts and their associated metadata.

Cohorts and samples in chapter 4.1.

The study population comprises 95 Amerindians living in an isolated, ethnically homogeneous, traditional community in the Southern part of French Guiana (Wayampi indigenous people). The Wayampi community consists of three villages arbitrarily named village TS (41.2%), village Z (18.1%), and village YP (23.6%), with restricted access to residents and located 100 km South of the closest village. Wayampi lived in large huts (13.9 + 8.6 inhabitants per hut, ranging from 4 to 38 individuals) with no restrooms or hygienic facilities. The diet consisted of local food from traditional crops, fishing, or hunting. They did not raise farm animals except a few free-running chickens. The only source of antibiotics is the health post in village Z, where drugs are delivered free of charge upon prescription by the resident nurses. When necessary, the villagers are hospitalized in Cayenne, the capital of French Guiana. Only visitors who benefit from permits rarely delivered by the regional authority are admitted to the village. Despite the restricted access to antibiotics and health care institutions, Wayampii people are heavily exposed to mercury due to gold mining activities (46). Demographic (age, sex, marital status, the number of children), lifestyle (including travel frequency, duration, destinations), environmental (size and location of households, contact with animals, daily activities, baby-sitting of children under five years old), medical history (current pregnancy, chronic disease, previous hospitalizations, and surgery, and oral and parenteral antibiotic intake during the year before sampling) data were collected from each individual in a standardized manner using data recorded at the health post and from interviews of local informers. Volunteers were included if they were healthy at physical examination and signed an informed consent form. The Regional Ethics Committee approved the study (Comité de Protection des Personnes SudOuest et Outre Mer III, 2010-A00682-37).

Another cohort comprising 32 European hospitalized individuals, 29 from a French hospital and 6 from Dutch hospitals recruited in the frame of the EvoTAR European project, was used for comparison. The samples of these individuals were taken before exposition to antibiotics (47).

Cohorts and samples in chapter 4.3.

The chapter focused on comparative genomics of different bacterial species. It used three different samples: fecal and clinical isolates of hospitalized patients and environmental isolates.

Fecal and environmental isolates were collected during a 34-month prospective study performed in the largest ICU ward of a single hospital between 2019 and 2021. We included 2417 samples recovered from sinks (surface, drain, and p-trap, n=1126) and room surfaces (bed rails, ventilator touchscreens, and pots, n=1291). Sink surfaces were sampled with polyurethane sponges placed in a sterile bag impregnated with 10 mL HiCap Neutralizing Broth (EZ-10HC-PUR, EZ Reach™ Sponges, World Bioproducts, Bioing sro, Czech Republic). Sink drainage and p-trap samples were collected with standard sterile swabs and aspiration probes. The samples were plated onto BD CHROMagar™ Orientation Medium, BD CHROMagar ESBL-Biplate, and mSuperCARBA (Becton Dickinson, Franklin Lakes, USA). The plates were then incubated for 48 h at 37 ºC. Colonies of different morphotypes (size, color, and shape) were subcultured onto BD CHROMagar™ Orientation Medium, incubated for 24 h at 37 ºC, and further identified with MALDI-TOF MS (MALDI Biotyper, Bruker, Billerica, MA, USA). Stocks of sub-cultured bacterial isolates identified as *Serratia* species by MALDI-TOF-MS (reliability score >2) were stored at −80 °C in 1500 µL Luria Bertani broth + 15% glycerol for further analysis.

A sample of clinical isolates from the collections in the Microbiology department causing individual episodes of bloodstream infections or involved in hospital outbreaks were also included for comparison

3.2 Genomics

3.2.1. DNA extraction and preparation

In some instances, bacterial isolates were sequenced "de novo" by Illumina and further closed by long-read sequencing (Oxford Nanopore, Oxford, UK).

Short-read sequencing (Illumina). Following manufacturer recommendations, DNA extraction was performed with the Chemagic DNA Bacterial External Lysis Kit (PerkinElmer, USA). DNA quality and concentration were measured in a NanoDrop 2000 Spectrophotometer (Thermo Scientific, Waltham, MA, USA) and Qubit® 2.0 Fluorometer (Life Technologies, Waltham, Massachusetts, USA). DNA fragmentation of 4 µg DNA was performed in a 46-µL elution volume in Covaris G-tubes by centrifuging at 4200–5000 rpm (Eppendorf 5424 centrifuge) for 90 s to achieve fragment sizes of ~20 kb, and fragment length was assessed with the TapeStation 2200 (Agilent, Waldbronn, Germany). Nextera DNA libraries were constructed using the Nextera XT library preparation kit and the Nextera XT v2 index kit (Illumina, San Diego, CA, USA). The library was sequenced on a HiSeq4000 (Illumina) using the reagent kit v2 (Oxford Genomics Center, Oxford, UK) to generate 250-bp paired-end reads.

Long-read sequencing (Minion). DNA was extracted with the MagnaPure 96 System (Roche, Basilea, Switzerland) and quantified with a Qubit® 2.0 Fluorometer. We used a minimum concentration of 20 ng/µL for the library preparation and the Rapid Barcoding kit 96 (Oxford Nanopore Technologies, Oxford, UK), following the manufacturer's instructions. The libraries were loaded onto flow cell versions FLO-MIN106 R9.4 MinION (Oxford Nanopore Technologies, Oxford, UK) and sequenced for 72 h. Base calling was performed in real-time with Guppy integrated into MinKNOW (Oxford Nanopore). Adapter sequences were removed with qcat (Oxford Nanopore), and fastq reads were filtered and quality measured with NanoFilt (>10,000 bases) and Filtlong v0.2.039 (500Mbp best reads) before the assembly. Hybrid assemblies were created with Unicycler v0.4.7 (53) and checked by visualization with Bandage v0.8.1.

3.2.2. Genomic analysis

Sequenced genomes were first annotated with Bakta 1.7.0 (48). Average nucleotide identity (ANI) was used to obtain species designations for each isolate. To compute ANI, we used alignments of all shared genes (using orthogroup clusters) for each pair of isolates. We compared reference assemblies obtained from the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/taxonomy) to obtain species designations. Subsequently, the core and accessory genomes were defined using the Pangenome Analysis Toolkit (PATO), facilitating population structure, annotation adaptive features, and creation of gene networks (49). Specifically, the PATO functions "core_plots" were employed to calculate the pangenome size, core, and accessory genome, "classifier" to identify the closest reference species in genome databases, and the "tree_dist" command to generate dendrograms. Clades were defined based on pairwise ANI scores using the unsupervised clustering MClust (50).

Core genome. Core genes were defined as those present in >90% of the isolates (blastP identity cutoff 90%) and analyzed through a pseudo-multisequence alignment with FastreeMP (50) midpoint root to build a maximum likelihood phylogenetic tree. The tree was displayed and annotated using ggtree version 2.2.4 in the R package. We employed iTOL (51) to edit and generate metadata linked to the tree.

Accessory genome. It was defined as the genes detected in less than 80% of the genomes. It was determined by employing PATO with default parameters (80% identity and 80% coverage) (49). The accessory genes enriched in any subclade were further analyzed using the PATO function "accnet_enrichment_analysis," which performs a multi-hyperparametric test to find overrepresented genes in a cluster compared with the whole population (i.e., the genome data set). The accessory network was illustrated with Gephi for rearranging and uploading the layout in R to plot the network.

Plasmid analysis. Plasmids were reconstructed from FASTA files, typed using the MOB-recon and MOB-typer modules of the MOB-suite software (52), and annotated with Bakta (48). A distance matrix using gene-by-gene presence-absences using the "accent" function of PATO with a default similarity parameter of 70% was created using Jaccard similarity. Then, we generated a k-nearest neighbor network (K-NNN) to allow reciprocal connections with the thresholds of 10 neighbors and 0.5 Jaccard distance. This implies that any plasmid

is linked to its ten most similar plasmids provided they share more than 0.5 Jaccard similarity. Plasmids were clustered from the K-NNN using mclust (50).

The plasmid network was built with Cytoscape, imported to R with the tidygraph R package, and set with the Louvain cluster algorithm (igraph R package). Information about plasmid features (predicted mobility, replicase, and relaxase IDs) was added to the network. Plasmid distribution over phylogenetic trees was built with the ggtree R software. All data manipulation and visualization were performed with the tidyverse R metapackage.

The categorization of replicases and genes encoding resistance to antibiotics, biocides, and heavy metals was performed by submitting hybrid assemblies of plasmid sequences to Plasmidfinder (53), Resfinder (54) and BacMet databases (55). Gview (https://server.gview.ca) was employed to generate the circular map and clinker (56) for the linear visualization and gene cluster comparison.

3.2.3. Genomic statistics

The Mantel test implemented in R was employed to compare the structure of the core and accessory trees by determining the correlation between the distance matrices. A nonpaired Student's t-test (stats R package) was applied for the significance test of the number of accessory genes per genome. We used a statistical test to analyze the shared accessory genes within each clade/subclade, modeling the distribution of shared gene pairs between genomes in each group. This analysis was further corrected according to the similarity between genomes. To evaluate the association of the metadata variables (origin, collection place, and year) with the core and accessory tree, we employed the "envfit" function in the vegan R package. This function calculates the multiple regression of environmental variables with the ordination axes and estimates the significance with a permutation test. The p-values were adjusted based on the Bonferroni correction. The goodness-of-fit statistic is the squared correlation coefficient (r2), which measures the correlation of the variable with the ordination. We assessed the significance of the SNPs/Mb within each clade/subclade using a Wilcoxon test for each group comparison. The p-values were subsequently adjusted using the Bonferroni correction method.

3.3 Metagenomics

3.3.1. DNA extraction and preparation

DNA was extracted from 250 mg of fecal samples using the Powerfecal Qiagenby (QIAGEN, UK) following manufacturer instructions. DNA quality and concentration were measured in a NanoDrop 2000 Spectrophotometer (Thermo Scientific, Waltham, MA, USA) and Qubit® 2.0 Fluorometer (Life Technologies, Waltham, Massachusetts, USA).

3.3.2. Sequence capture platform

The resistomes and mobilomes were characterized by targeted metagenomics using the ResCap-targeted sequence panel "ResCap" (29). ResCap comprises well-known and hypothetical genes encoding resistance to antimicrobials (antibiotics, heavy metals, and biocides) and genes coding for relaxases, enzymes involved in DNA mobilization and transfer of conjugative elements that typically carry ARGs (plasmids, conjugative transposons). It is based on the custom SeqCap EZ library format (NimbleGen, Madison, USA). Specifically, it comprises 78,600 non-redundant genes (81,117 redundant genes) with a target space of 88.13 Mb. Probes targeting the antibiotic resistome include 47,806 putative ARGs and 7963 functionally characterized genes, defined as "canonical ARGs." Probes targeting the metal and biocide resistome contain 30,794 putative resistance genes and 704 canonical genes encoding resistance to metals (MRGs) and biocides (BRGs) (55). The platform also provides probes for 2517 relaxase genes of the ConjDB database (57) for plasmid identification and classification.

3.3.3. Targeted Metagenomics Workflow

The ResCap workflow consists of (i) whole-metagenome shotgun library construction, (ii) hybridization and capture, and (iii) captured DNA sequencing. All the steps were performed according to NimbleGen standard protocols for Illumina platforms, as previously described (29).

The raw sequencing data was analyzed using the sequence aligner bowtie2 to capture the on-target sequences (58) from the ResCap database. Gene identification from the capture extraction was accomplished using KMA (59), a mapping method designed to map raw reads directly against redundant databases. KMA is robust and precise, capable of distinguishing between gene alleles in a sample. To improve and update the annotations, we built a custom-

made library based on the latest version of the databases used in the ResCap platform. This library was compiled from Resfinder (54), the experimental and predicted genes from BacMet (55), and the relaxases from ConjDB (57,60). Furthermore, we developed a pipeline to convert the gene data from BacMet initially presented in the protein version to nucleotide sequences.

Subsequently, we filtered out captured genes with a p-value higher than 0.05 and an identity lower than 90% for the ARG database (ARG), relaxases (REL), and the biocides and metals from our custom database (BAC). The identity threshold for the predicted biocides and metals (BAC) database was set at 100% due to the extensive size of the database. Finally, we conducted a rarefaction analysis to normalize the results. The rarefaction was performed using the "rrarefy" function from the package vegan implemented in R.

3.3.4. Metagenomic analysis

We first analyzed the alpha diversity of the data through the Shannon index, Chao 1 richness estimator, and Simpson index implemented in the "phyloseq" package. The normality of the diversity estimated values was evaluated with the Shapiro test (rstats). Then, the metric's distribution was compared between groups using the Wilcoxon test for non-normal distributions and the t-student for normal distributed data.

We employed different dimension reduction algorithms, including Principal Component Analysis (PCA) from the stats package, Uniform Manifold Approximation and Projection (UMAP, umap package), and T-sne (t-distributed Stochastic Neighbor Embedding) using the Rtsne. These algorithms were utilized to identify composition patterns between groups and sample clusters. To determine if there was a statistically significant difference in composition between Amerindians and European individuals, a permutational multivariate analysis of variance (PERMANOVA) ("adonis2" function, vegan) was performed.

We compared the number and diversity of genes between populations with the Mann-Whitney test, and we used the package "ggstatsplot" to visualize the data distribution for both groups. To calculate the sampling depth, we performed a pseudo-rarefaction by counting the number of newly added genes by each new individual for each cohort.

We calculated the core genes for each population, defined by the genes that appear in all individuals of each cohort and both cohorts. Then, a pseudo-core was calculated with the genes that appear in at least 80 % of the population.

3.3.5. Metagenomic Statistics

We performed a chi-square test for every resistance gene in the platform to its percentage of apparition. Next, genes with a p-value below 0.05 were selected, and the appearance percentage for the gene in the two populations (European and Amerindians) was calculated. We used a GLM (general linear model) to analyze the differential gene abundance between groups, and the results were visualized in a volcano plot and heat maps from the package pheatmap.

The ADONIS test examined whether metadata variables explain gene composition differences. Gene composition comparisons between groups were performed using the Linear Discriminant Analysis (LDA) effect size (LEfSe) (61). This method identifies genes with statistically significant differences in abundance between groups of samples. All analyses were performed using R software (version 3.0.2; R Foundation for Statistical Computing, Vienna, Austria).

3.4 Bioinformatic software development

3.4.1. PATO: Pangenome Analysis Toolkit

PATO was designed as an R package for several compelling reasons. Firstly, R stands out as a premier environment for robust large-scale data analysis. Secondly, the expansive repertoire of packages within R, complemented by the Bioconductor repository, covers a diverse spectrum, including genomics, population analysis, data visualization, and statistical analysis. Thirdly, the versatility of R, being free, multiplatform, and capable of parallel computing, aligns seamlessly with PATO's objectives.

3.4.1.1. Programming Language and third-party software

PATO incorporates both internal R functions and external software, enhancing its flexibility and expanding its capabilities to ensure a comprehensive and practical toolkit for users. The foundational operations of PATO are reliant on MMSeqs2 (62), MASH (63), and Minimap2 (64). PATO incorporates the executable files of these software applications to circumvent potential incompatibility issues, ensuring compatibility with Unix and MacOS environments. Multithreading properties have been meticulously implemented within PATO functions to optimize computational efficiency.

Beyond these primary executables, PATO encompasses additional auxiliary programs beyond these primary executables, including bedtools (65), k8 (https://github.com/attractivechaos/k8), and a customized script suite. This comprehensive integration of diverse computational tools enhances PATO's functionality and versatility, enabling a robust and efficient computational framework.

3.4.1.2. External databases

PATO leverages external databases to annotate pathogenicity and antibiotic resistance sequences. Specifically, it integrates information from the Virulence Factor DataBase (66) for pathogenicity-related annotations and the ResFinder database (54) for antibiotic resistance-related annotations. Additionally, PATO incorporates a bespoke database comprising reference and representative bacterial genomes sourced from RefSeq. Notably, all entries in this database were last updated in October 2020, ensuring a comprehensive and contemporary foundation for sequence annotation within the PATO framework.

3.4.1.3. Pipeline adaptability

PATO is architecturally constructed with a suite of S3 objects and functions designed to enhance robustness and interoperability across its functionalities. This meticulous design ensures a seamless flow of data, where the output of one function seamlessly serves as input for another, fostering comprehensive compatibility.

In certain instances, PATO employs temporary files and folders within specific functions, enabling the judicious reuse of generated data when necessary. This resourceful approach optimizes computational efficiency and contributes to the overall resilience of PATO's computational framework.

3.4.1.4. Software comparison

We conducted a comprehensive comparative analysis of PATO against prominent software tools in the field, including PanX, PanACoTA, Panaroo, Roary, PIRATE, and PEPPAN (67–72).

3.4.2. BACTAX-ID

3.4.2.1 Genome retrieval

In January 2021, 130,995 genomic sequences were sourced from the National Center for Biotechnology Information (NCBI). The dataset is exclusively comprised of complete genomes belonging to genera with an abundance surpassing 1000 genomes. Taxonomic assignments were made based on the authoritative taxonomy provided by NCBI, which served as the primary reference for sequence identification. Our analysis encompassed 44 genera, and the sequences were acquired through the File Transfer Protocol (FTP) for comprehensive retrieval.

3.4.2.2 Genomic Distance

We employed the MASH software (63) to establish a distance metric between pairs of genomes, forming the foundation of our identification system. MASH computes the similarity between two genomes, yielding a percentage of identity as a quantifiable measure of their genomic relatedness.

3.4.2.3 Code

The algorithm was implemented in Python 3, with Pandas as the primary library for effectively managing extensive databases. SQL, specifically PostgreSQL, was employed to establish and maintain the relational database.

3.4.2.4 Algorithm

Our system leverages a hierarchical genome organization scheme, mirroring the nested structure of a matryoshka doll. This hierarchical framework encompasses seven distinct levels of depth, facilitating efficient genome classification.

Furthermore, our algorithm incorporates the concept of pseudo-cliques from graph theory (73). Within graph theory, a clique represents a complete subgraph where every pair of vertices within the subset are connected. Our algorithm adopts the concept of a pseudo-clique, characterized by a high degree of connectivity between most, but not necessarily all, vertices. The algorithm is detailed in the results section.

4. RESULTS

4. RESULTS

4.1. Development of bioinformatic tools for in-depth metagenome analysis

During the work of this doctoral thesis, a targeted sequence capture (TCP) previously developed by the Ecoevobiome group to analyze the pool of ARGs was improved and applied to screen different metagenomes (29). The platform combines targeted metagenomics using last-generation in-solution capture platforms with bioinformatics methods to establish a standardized framework that allows quantitative and qualitative analyses of resistomes. It offers significantly enhanced sensitivity and specificity concerning the metagenomic methods that have been developed. It allows the analysis of genes related to the selection and transfer of antimicrobial resistance (biocides, heavy metals, and the extrachromosomal elements in which they are located). The improved pipeline and user-friendly user interface are described in **ANNEX 1**.

The platform was applied to the analysis of different populations, namely, i) pristine communities (Amerindian cohort), ii) animal cohorts selected by antibiotics, and iii) contaminated environments (hospital wastewaters). They will perform during the development of the collaborative grants STARCs (JPIAMR-AC16/0039) and MicroCarbaflux (CIBERINFEC22PE02) under the perspective of OneHealth. This chapter will comprehensively describe the first example in section 4.1.1. Some figures will be added in sections 4.1.2 and comparatively discussed in SECTION 5.

These works have been presented at international conferences as **oral communications (see Annex 3)**.

4.1.1. In-depth metagenomic analysis of intestinal microbial communities from Amerindians to address the “baseline human resistomes.”

4.1.1.1. Motivation

Natural microbiomes and resistomes in people living ancestral lifestyles, such as indigenous communities with sporadic exposure to antibiotics and health systems, became essential to understand how antibiotics could act on pristine microbiomes (74). Differences between the composition of microbiomes from westernized and non-industrialized human populations have been documented (74–77). The gut microbiota of Papua New Guinea populations exhibited greater bacterial diversity and composition and lower inter-individual variation than the gut microbiome of American individuals born in the U.S. (74,78). In addition, the high diversity of ARGs conferring resistance to synthetic antibiotics indicates that AMR is a default mechanism in pristine environments such as soils and microbiomes of indigenous populations (20,74,79,80). Nonetheless, the low sensitivity and specificity of most metagenomic approaches make it difficult to characterize resistomes accurately (29).

In this chapter, through targeted metagenomics (29), we deeply analyzed the resistomes of Wayampis, an Amerindian human population in a remote geographical community with different ecological networks, lifestyles, and exceptional exposure to antibiotics and heavy metals in comparison to Western people. Specifically, we aim to: (i) compare the resistome, metalome, and mobilome diversity and composition between Amerindian and Western populations; (ii) understand the influence of some demographic factors in shaping the resistome, metalome, and mobilome profile of the Amerindians; and (iii) improve and update the bioinformatic pipeline to analyze resistome datasets.

4.1.1.2. Alfa- and Beta-diversity in Amerindians and European cohorts according to Gene Categories

We explored the diversity of genes encoding resistance to antibiotics (ARGs), metals (MRGs), and biocides (BCRs), and genes encoding relaxases (REL) as a proxy of plasmids in the fecal samples of Amerindian and European populations. **Figure 1A** shows the diversity of these gene groups using Shannon, Chao 1, Evenness Simpson, and Dominance Simpson indexes. To statistically compare both groups, we checked if they followed normal distributions. We found that only MRG for dominance Simpson was normally distributed

(Shapiro-Wilk normality test) (**Annex 2**). Based on the t-student test, there were no significant differences between the groups.

The alpha diversity for all gene categories ranged from 0 to 5 using the Shannon index, with similar median values for Amerindians and Europeans for ARGs (2.8 vs. 2.7), BCRs (2.9 vs. 2.7), MRGs (3.4 vs 3), and REL (3.4 vs 3.5) (**Figure 1A**). Based on the Chao 1 richness estimator, the fecal samples of Amerindians and Europeans differ in the content of MRGs (110 vs. 72) and BCRs (71 vs. 54) but not in ARGs (57 for both cohorts) or REL (92 vs 96). Nonetheless, differences in the distribution of ARGs and BCRs were detected according to the Dominance and/or Evennes Simpson tests (**Annex 2**).

To further identify the composition of the different gene categories, we performed a Uniform Manifold Approximation and Projection (UMAP) coupled with the ADONIS test (**Figure 1B**). The UMAP grouped Amerindians and Europeans by ARGs and REL categories but not by MRGs or BRGs categories (ADONIS; p-values<0.001) (**Annex 2**).

All these results showed that the diversity and richness metrics of Amerindians' gut resistome, metalome, and mobilome are similar to that of Europeans. However, they differ in the gene composition of all gene categories, the gene evenness being significantly different for ARGs and BRGs.

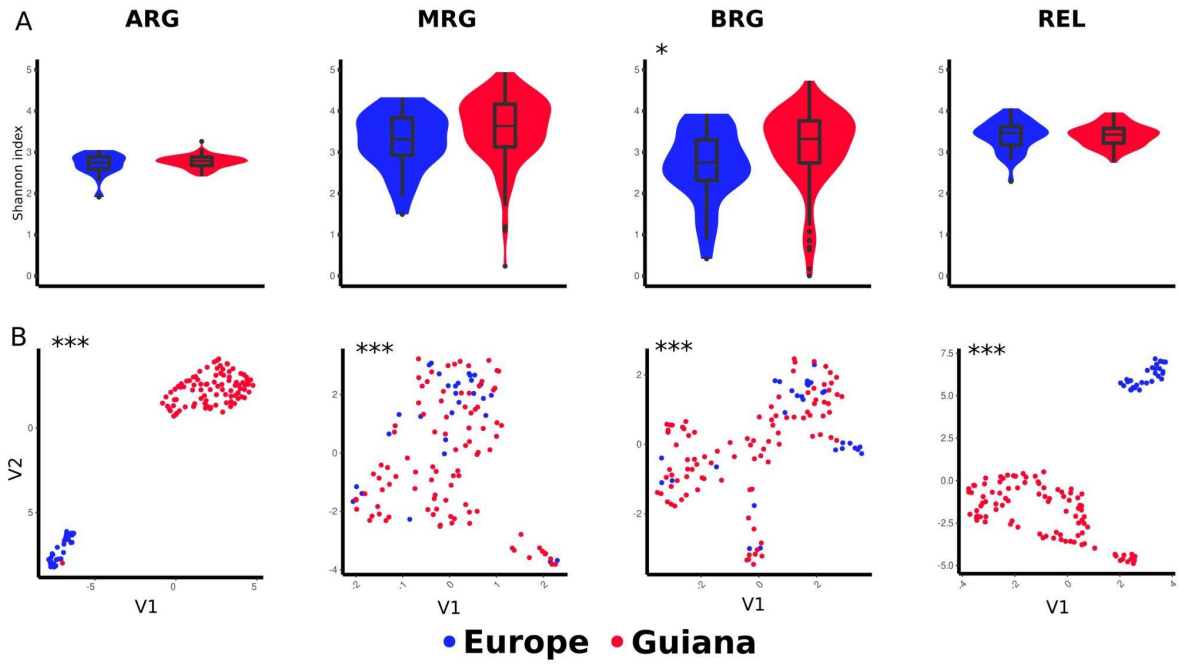


Figure 1. Alpha and beta diversity comparison among Amerindian and European populations. ARG=antibiotic resistance gene; MRGs=Metal resistance gene; BRG=biocide resistance gene; REL=relaxase. **(A)** Alpha diversity is based on the Shannon Index. Box-plot combined with Violin-plot for the Shannon Index estimations of ARGS, MRGs, BRGs, and REL categories. * for statistically relevant differences (BRG) **(B)** Uniform Manifold Approximation and Projection (UMAP) plots to compare resistance gene profiles for ARGS, MRGs, BRGs, and REL categories. The ADONIS test p-values are at the panel's top (***; p-values < 0.001).

4.1.1.3. Diversity and distribution of ARGs, BAC, and REL in Amerindians and European cohorts

After estimating the overall community heterogeneity (alfa- and beta-diversity), we explored the diversity of each gene category within and between cohorts. The overall number of ARGs and REL in the Wayampi cohort was similar to that of Europeans (259 vs 264 and 331 vs 312, respectively), while the number of BAC (MRGs+BRGs) greatly differs between the two cohorts (2039 vs. 3670) (**Figure 2A**). These genes were unequally distributed, some being exclusively found in Wayampis or Europeans [ARGs (103 vs 109, representing 40% of such resistomes), REL (91 vs 72, representing 23-27% of this gene category), and BAC (2390 vs 759, representing 66 and 37%)] (**Figure 2B**).

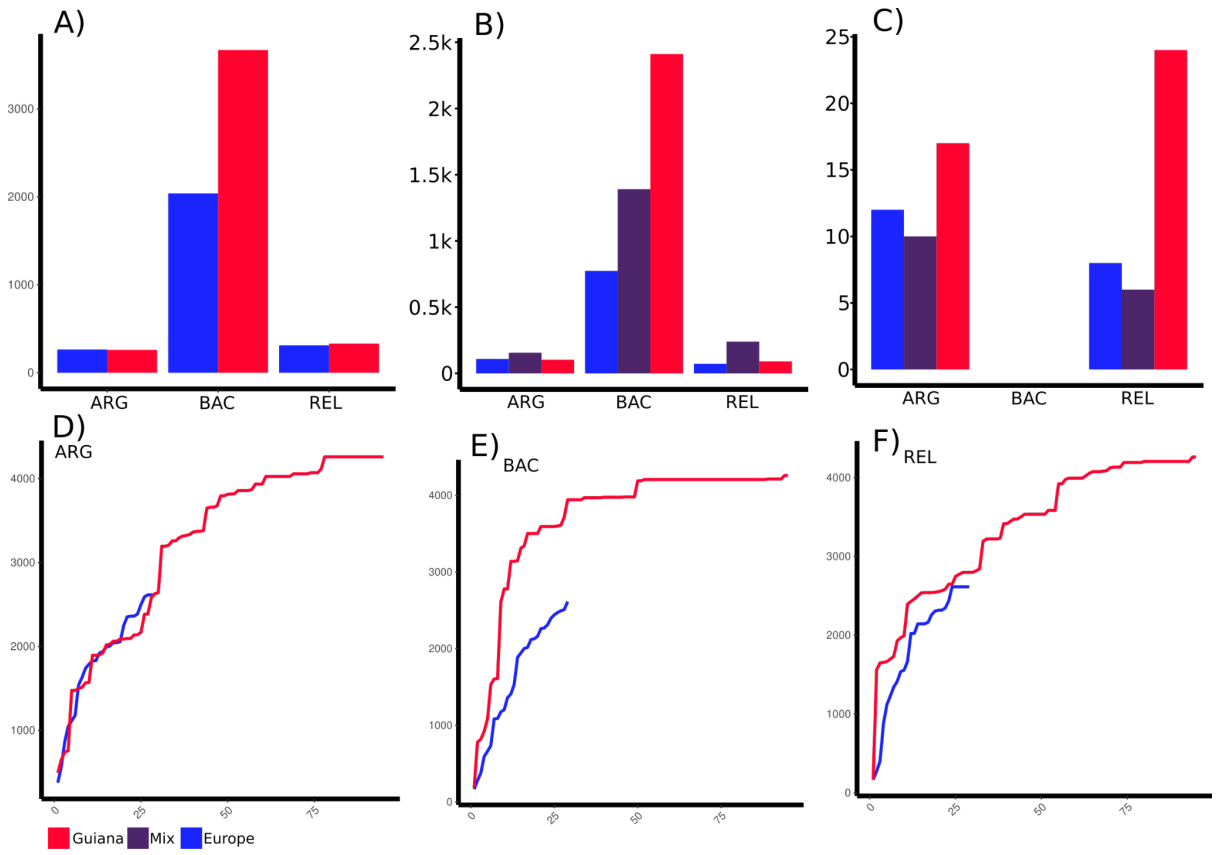


Figure 2. Distribution and diversity of genes for Amerindian and European populations. A) The number of genes appearing at least once in each population and combined per gene category. B) Core resistance genes per gene category, cohort, and combined. C) Number of unique genes for each population in each category. D) Rarefaction curves of the gene diversity for Amerindians and Europeans in ARG. E) Rarefaction curves of the gene diversity for Amerindians and Europeans in BAC. F) Rarefaction curves of the gene diversity for Amerindians and Europeans in REL. ARG: antibiotic resistance gene; MRGs: metal resistance gene; BRG: biocide resistance gene; REL: relaxase. The "y-axis" refers to the number of genes. For D), E), and F), the "x-axis" refers to the number of samples taken into account.

The cumulative frequency distribution of unique genes allowed us to classify them as "core genes" (present in all individuals of a given cohort)(**Figure 2C**), "high prevalence genes" (present in mostly all the people in the study), and "low prevalence genes" (present in a few individuals) (**Figure 3**) according to the early categorization of pangenomes by Wolf and Koonin (81) . Most of the genes were medium prevalence genes. Interestingly, the same distribution was found for each cohort and each gene type (**Figure 2C**) (**Annex 2**)).

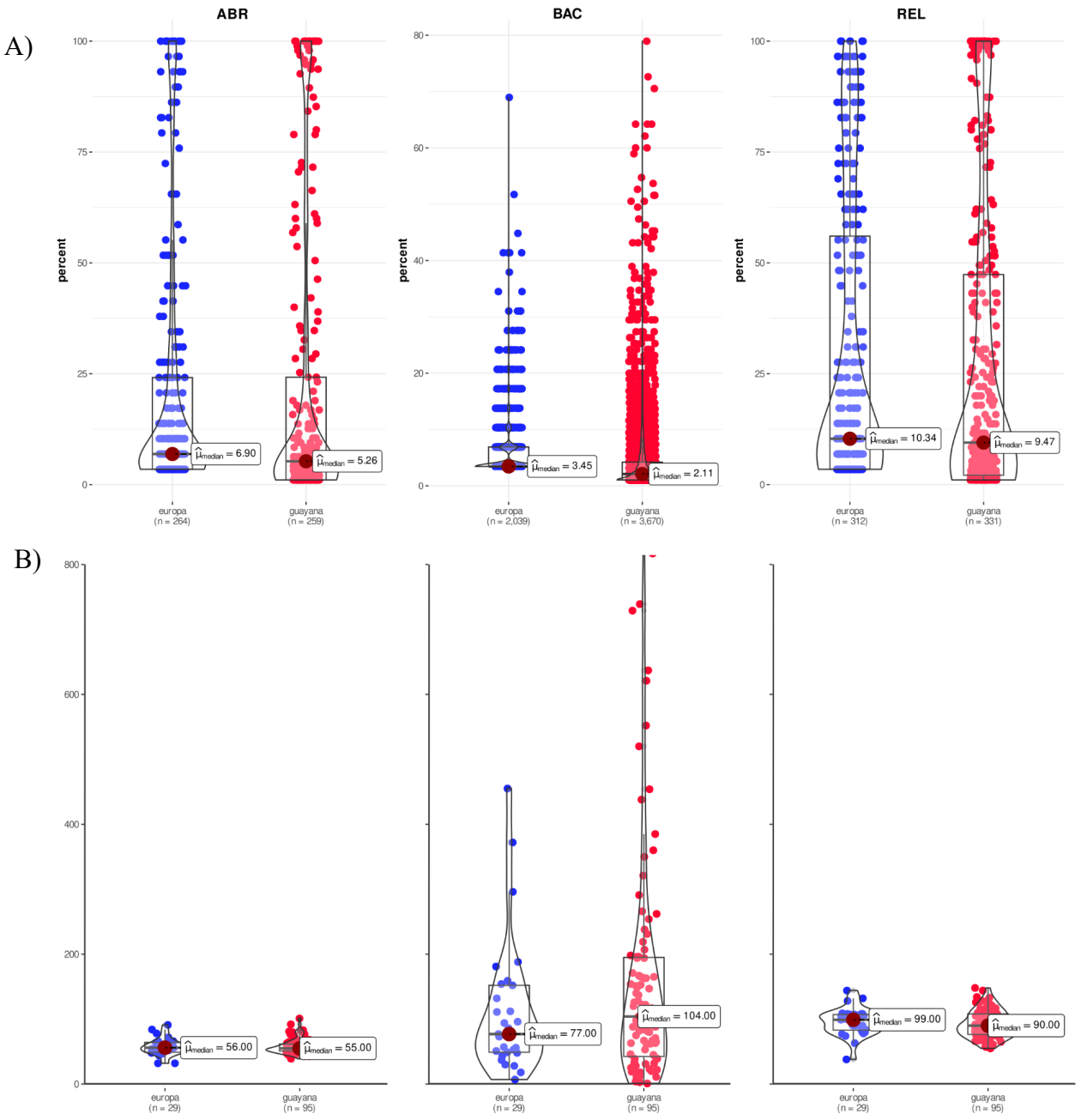


Figure 3. Distribution and diversity of genes for Amerindian and European populations. A) Frequency and mean of each gene in each population. B) The number of genes per person in each population and their mean. ARG: antibiotic resistance genes; MRGs: Metal resistance genes; BRG: biocide resistance genes; REL: relaxase. The "y-axis" refers to the quantity of genes per person.

Rarefaction analysis per group of individuals and database showed that the number of samples in the study was enough to yield correct estimates for the resistome diversity (i.e. most rarefaction curves reached a plateau for the cohort sample sizes (**Figure 2 D, E, F**) (82).

4.1.1.4. Categorization of ARGs, MRGs, BRGs, and REL categories

Figure 4 shows the relative abundance of ARGs, BRGs, MRGs, and REL for each population. Within the category of ARGs, genes coding for resistance to tetracyclines predominate in both groups, those coding for resistance to beta-lactams are only overrepresented in Amerindians, and those associated with resistance to macrolides and aminoglycosides are only overrepresented in Europeans (**Figure 4A**). For relaxases, MOBP1-type genes predominate in both cohorts, although they were more abundant among Europeans, while MOBV was overrepresented in Wayampis samples (**Figure 4D**). Conversely, the distribution of MRGs is similar between the two cohorts, and the predominant genes are associated with resistance to copper and arsenic, followed by iron, silver, zinc, and nickel (**Figure 4B**). The distribution of BRGs is also similar in both cohorts, although Amerindians are enriched in quaternary acids and Europeans in genes associated with resistance to xanthenes (**Figure 4C**). A more detailed analysis of the richness (number of distinct genes) and evenness (abundance of each one) within each gene category indicates the presence of "core genes", "highly prevalent genes," and "sporadic genes," which would be linked to the phylogenetic structure of the bacterial communities and/or the mobilome of the different taxa. In the following lines, the pools of ARGs, MRGs, and BRGs were analyzed in more detail due to the relevance of some variants in biomedicine.

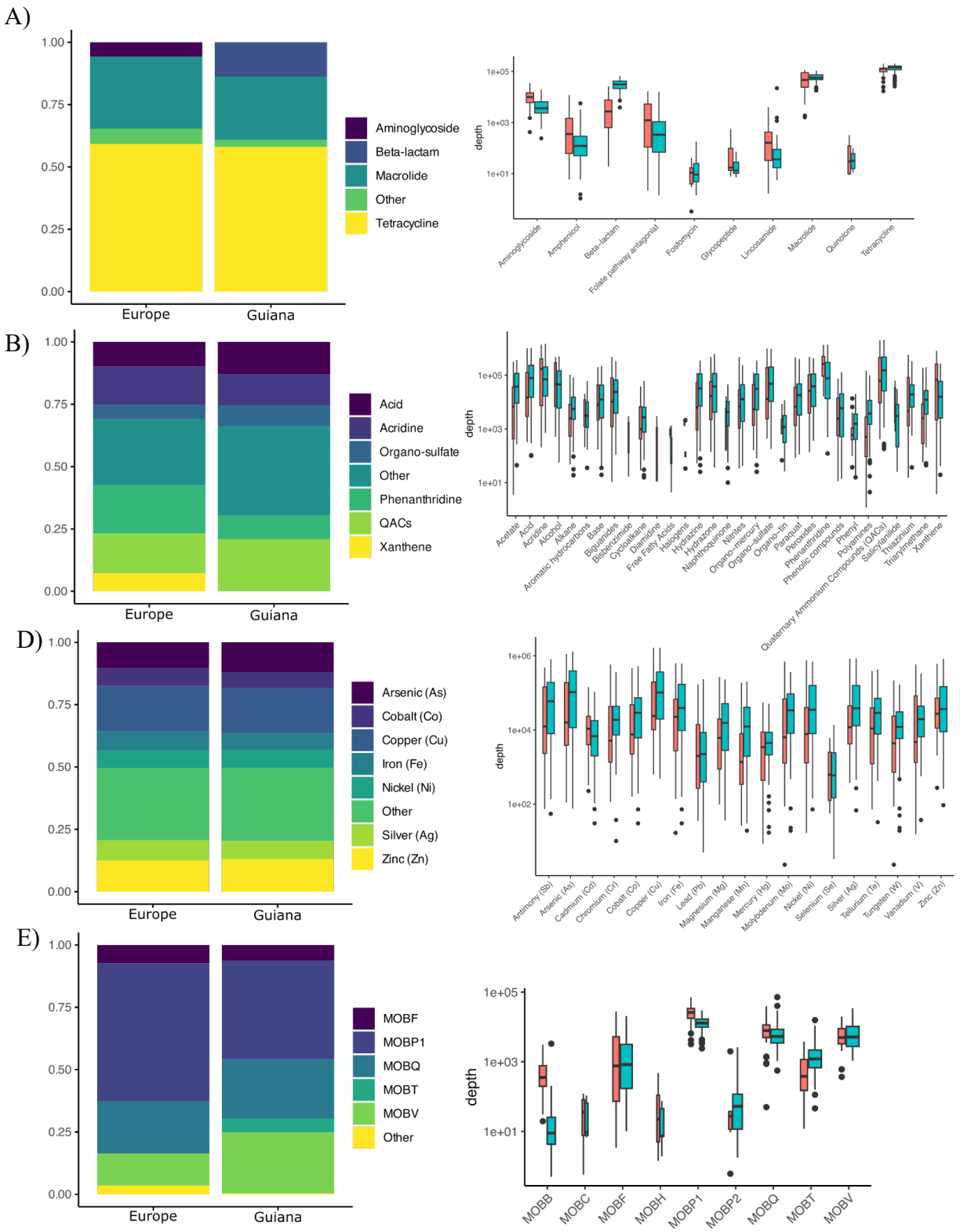


Figure 4. Gene diversity and depth. (A) Percentage of ARG genes in the population and the depth of lectures for each category in logarithm, **(B)** Percentage of BRG genes in the population and the depth of lectures for each category in logarithm, **(C)** Percentage of Metal genes in the population and the depth of lectures for each category in logarithm **(D)** Percentage of REL genes in the population and the depth of lectures for each category in logarithm, (red is key for Europe and blue for Amerindians).

The resistome. The ARGs' pool of Wayampi and Europeans comprises a similar number of genes (259 and 264, respectively), and more than half were detected in both groups (156 ARGs). They were categorized as genes encoding resistance to beta-lactams, tetracyclines, macrolides, aminoglycosides, phenicols, quinolones, glycopeptides, and fosfomycin. Among them, a set of "core genes" for the resistomes of both populations include some associated with resistance to tetracyclines [*tet*(32), *tet*(O), *tet*(Q), *tet*(W), *tet*(X)], beta-lactamase (*cfxA6*), aminoglycosides (*ant6´-I, aph3-III*), or macrolides (*ermF* and *ermB*). They are well-recognized ARGs, although they seem to be markers of anaerobic species and/or Firmicutes or specific mobile genetic elements. For example, CTnDOT comprises *tet(Q)* and *ermF* and is present in species of the *Bacteroides* genus and other anaerobes such as *Prevotella* or *Capnacitophaga*. Both *aph3-III* and *ant6´-I,* are common in a plethora of transposable elements of Lactobacillales (streptococci, enterococci, and staphylococci) as well as other commensals within the Firmicutes phylum. Besides these core genes, we also detected chromosomally located ARGs associated with the presence of different species, such as beta-lactamases of class A (bla_{LEN} and bla_{OKP} of *Klebsiella* spp); class B (bla_{Cfi}), class C (bla_{MIR} and bla_{ACT} of Enterobacter), class D (bla_{OXA} of Campylobacter and intestinal spirochaetes), aminoglycosides (*aac-6-Ii* of *Enterococcus faecium* or *Enterococcus casseliflavus*), macrolides (*msrC* of *Enterococcus faecalis*), or phenicol (*catS* of streptococci).

Regarding well-known ARGs conferring clinical resistance to first-line antibiotics, we detected ARGs usually located on emblematic mobile genetic elements such as class 1 integrons (*dfrA1, sul1, aadA1*), transposable units in Firmicutes [(erm(B), *aph(3)-III*)] or enterobacterales [(*aph(3´´)-Ib aph(6´)-Id,* also known as *strA/strB*; and *oqxA/oqxB*]; or widely resistance genes (*sul2*). It is interesting to note the redundancy of some ARGs (represented by different alleles of the same gene at the individual or population level) and their distribution between both groups.

The metalome. BCRs and MRGs were the most abundant category (n=2411). (**Figure 4B)** reflects an apparent similar distribution of MRGs categories between the two groups of individuals. Many genes are housekeeping genes responsible for the intake of essential metals and, thus, present in most people as they were associated with commensal taxa (**Annex 2**). Within the overrepresented categories related to the metals zinc, arsenic, copper, silver, and iron, we found relevant genes involved in the detoxification of these compounds. They included pCoABCDRS and *tcrB* (copper), *silACFPRS* (silver), *ars* (arsenic). Those

linked with copper comprised a high percentage of the population, while the others were located in less than 10% of the population.

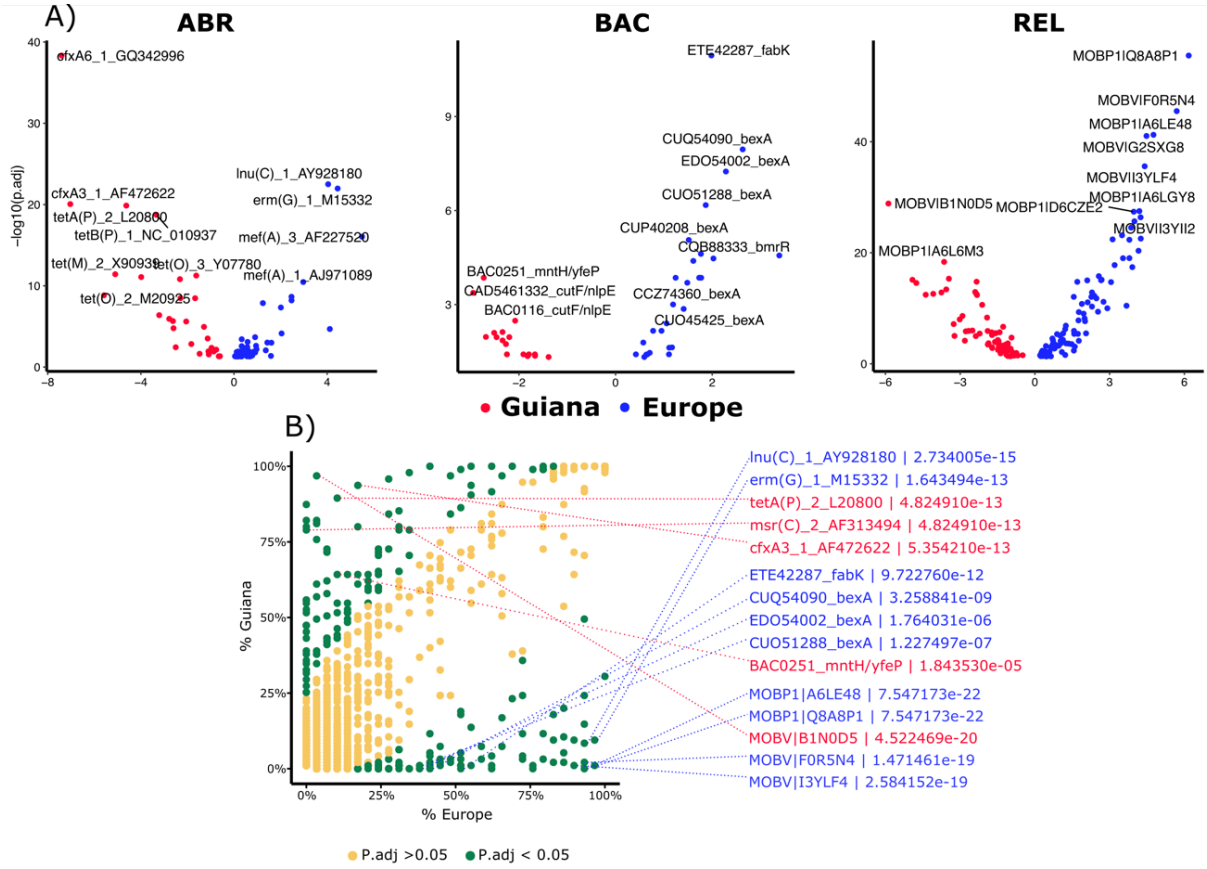


Figure 5. Most discriminative genes Amerindian and European populations (A) Most discriminative genes to each group per category (ABRs, BRGsC, RELs) based on the Generalized linear model (GLM). **(B)** Most discriminative genes and their adjusted p-values for each cohort using the Fisher test. It shows the genes significantly overrepresented in each population. Among Wayampis, ARGs include species-specific markers of enterococci [*aac(6´)-Ii* and *msr(C)*], Enterobacter (*bla*MIR), or Clostridium [*tetA/B(P)*].

The biocide resistance gene pool. In contrast with that observed for MRGs, differences were observed for BRGs, the categories of genes associated with the susceptibility of quaternary ammonium salts and xanthenes being those showing more remarkable differences (common genes target these compounds). Many genes correspond to the intake regulation of disparate compounds but are present in less than 10% of the population. Some acquired genes associated with resistance and/or tolerance to known antiseptics and co-selected with ARGs were detected. They were *qacE* (located on widespread class 1 integron) or *sugE* (often linked to blaCMY genes in plasmids in epidemiological papers).

Most discriminative genes among the cohorts. Utilizing a Generalized Linear Model (GLM) to analyze ARG genes within the Guiana population, we identified genes from the cfx and tet families as particularly noteworthy. These genes exhibited significant characteristics, corroborated by a Fisher test, further underscoring their prominence within the population. (**Figure 5)**

4.1.1.5. Variables to explain resistome composition in Amerindians.

The analysis of the Guiana population resistome shows no single cluster under UMAP nor PCA dimension reduction. The metadata analysis of the Guiana population was based on the ADONIS test to identify those variables significantly impacting the resistome composition (**Figure 6**).

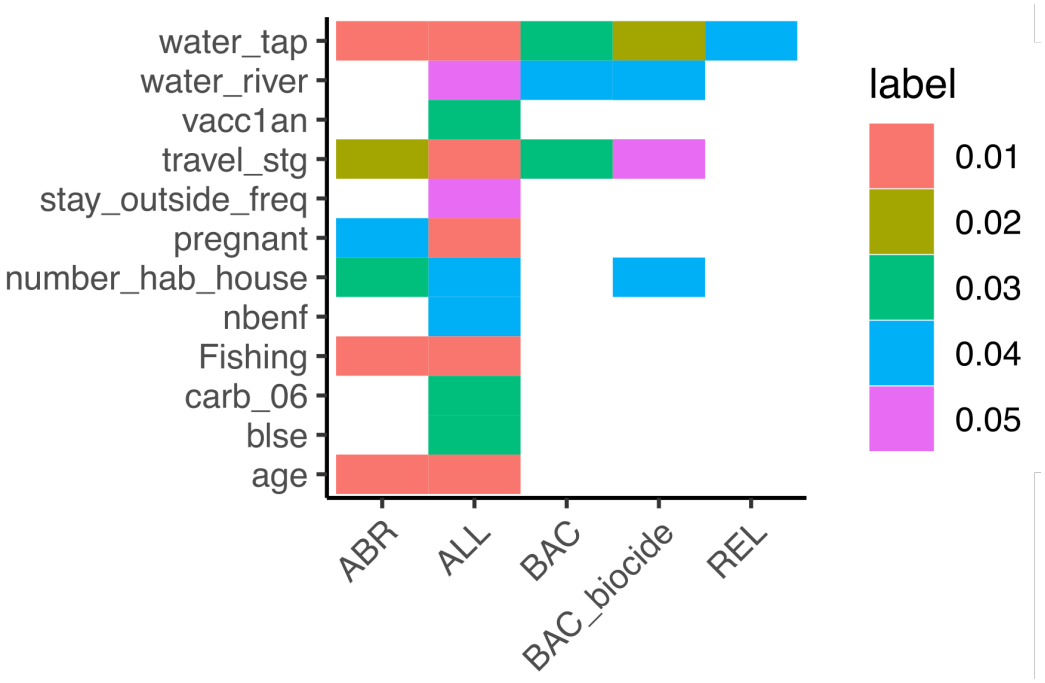


Figure 6. Metadata analysis of the Guiana population based on the ADONIS test. Significant variables explaining differences in resistome composition and their associated p-values are shown per category.

According to ADONIS, only five variables were statistically significant for different gene categories from a total of 88 examined. While the variable "drinking water from a tap, was significative for all categories (ARGs, BRGs, MRGs, REL), ARGs were associated with "fishing" and "age" MRGs and BRGs with "traveling" and "water river" and BRGs with the "number of inhabitants per house". When all the gene categories are combined, we added from all the previously described "vaccinated less than a year ago," "staying outside the

village frequently", "the number of children", "the house they inhabit", and "having taken beta-lactams".

4.1.1.6. Novelty and Significance

This study is original and novel as it constitutes the first application of in-depth targeted metagenomics of high specificity and sensitivity to analyze resistome in host populations.

The deep analysis of resistomes in host and non-host microbiomes is of high value in different academic and industrial sectors as AMR is a well-recognized "Global Health challenge" that threats the sustainability of all ecosystem services and demands the active participation of different economic sectors (pharmaceutical, agro-food, health) (UNEP, 2023, UNEP 2022). First, it constitutes one of the few works that provide a detailed catalog of the baseline resistome, essential to interpreting results in different ecosystems (26,83). Second, it confirms the relationship between resistome and phylogeny (21) and the need to further explore the non-acquired resistome and their impact on the susceptibility to treatments (Ruppe 2019).

4.1.2. In-depth metagenomic analysis of "the hospital resistome" to address the selection and transmission of antimicrobial resistance

4.1.2.1. Motivation

The impact of antibiotic therapy and hospital interventions on the patients´ microbiome has also been analyzed in specific patient cohorts with specific chronic diseases. However, we still lack a general view of the dynamics of common opportunistic pathogens associated with antimicrobial resistance in the compartments of the hospital ecosystem and, more specifically, high-risk hospital wards such as intensive care units (ICUs).

The main goal of **chapter 4.1.2**. is to understand the selection and transmission of ARGs in high-risk hospital areas. For that, we performed a longitudinal and prospective population-wide analysis of the microbiome and resistome of a large ICU ward of a 1200-bed tertiary hospital for a 12-month using pooled samples, multi-omics, metadata (patient records, clinical microbiology information, hospital ward metrics, microbiomics).

This work has been presented in national and international conferences as posters or oral communication (**Annex 3**). Below are those in which the candidate appears as first author.

- M.D. Fernandez-de-Bobadilla. Dynamics of a "hospital ward resistome" using pooled fecal patient samples, advanced targeted metagenomics, and metadata analysis. FLAMR. 2022. Oral Presentation.

- M.D. Fernández-De-Bobadilla, V. F. Lanza, F. Baquero, T. M. Coque . Selection and transmisión patterns of antimicrobial resistance in a large Intensive Care Unit using a population-wide surveillance approach. IMMEM 2022. Poster.

- M.D. Fernández-De-Bobadilla, V. F. Lanza, F. Baquero, T. M. Coque Meta-análisis de resistomas: Análisis de la resistencia a antibióticos bajo una perspectiva One Health. SEIMC 2022, Granada.Oral Presentation

- M.D. Fernández-De-Bobadilla. Patrones de selección y transmisión de clones y plásmidos de enterococos en una UCI. SEIMC 2023. Santiago de Compostela. Oral Presentation (manuscript in preparation)

4.1.2.2. The Ward Resistome

Figure 7 shows the catalog of "canonical" ARGs identified in "pooled" patient fecal samples throughout the 12 months of study. The recovery of ARGs encoding resistance to first-line antibiotics for the treatment of Gram-negative infections includes those conferring resistance to carbapenems (*bla*$_{GES}$, *bla*$_{KPC,}$ *bla*$_{OXA}$) and/or third-generation cephalosporins (*bla*$_{CTX\text{-}M\text{-}15}$). While the genes encoding carbapenemases were coincidental with the emergence of hospital outbreaks, those coding for extended-spectrum beta-lactamases were detected during the whole period. Other genes associated with first-line antibiotics used for the treatment of infections caused by Gram-positive pathogens, such as beta-lactams (*mec*) or glycopeptides (*van A, B, D, G*) in staphylococci and enterococci (**Figure 8**), respectively, were detected during the whole period of study.

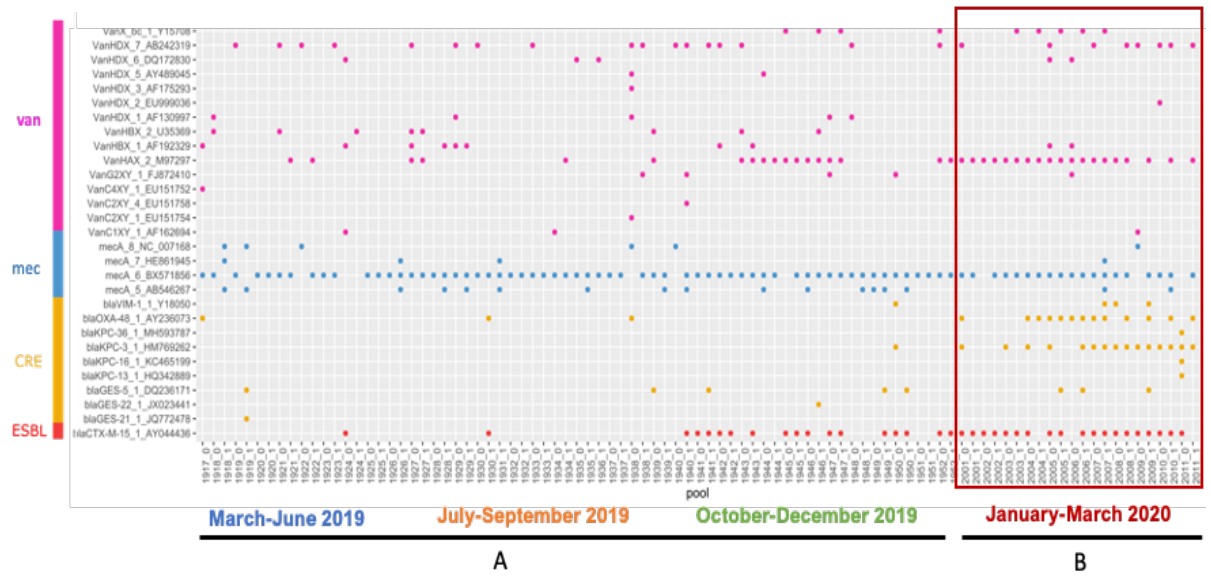


Figure 7. High-priority ARGs in pooled fecal samples from ICU patients

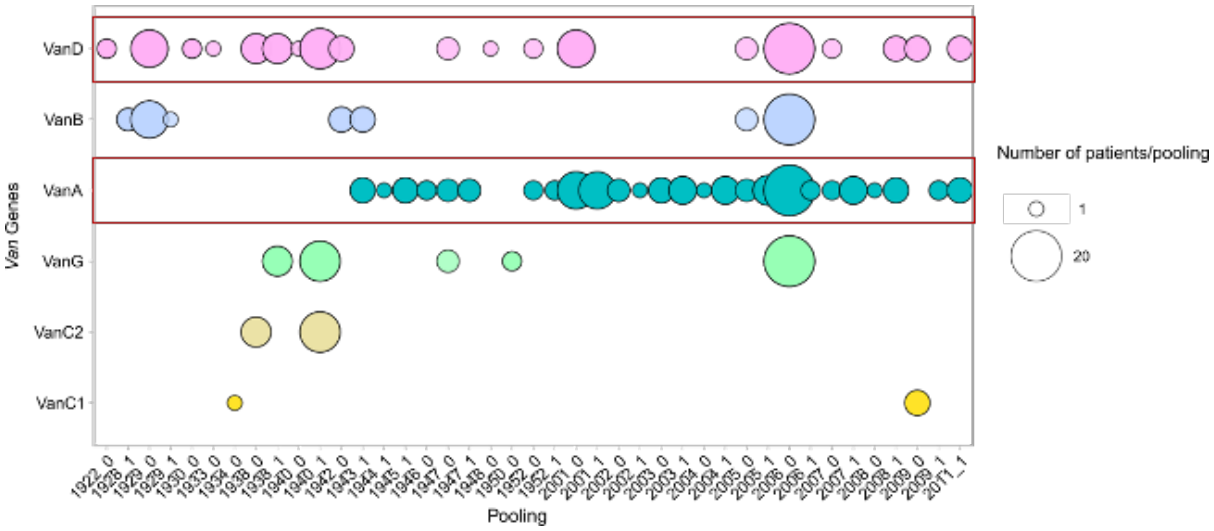


Figure 8. The Enterococcal resistome. The temporal distribution of *van* genes was determined by the ResCap enrichment platform. The "x-axis" represents the time pooled samples were collected, and the "y-axis", the type of the *van* alleles detected. The size of the circles is proportional to the reads.

4.1.2.3. Novelty and significance

This study is original and novel as it constitutes the first application of in-depht targeted metagenomics to "population surveillance," considering the ICU ward as a single entity to evaluate the impact of interventions in the ward ecosystem.

ICU stay promotes colonization pressure and thus the risk of selection, transmission, and persistence of ARGs. The distribution of "*van*" genes suggests “hidden” clonal transmission of VRE (*vanA*) and undetected hidden intestinal reservoirs (*vanD, vanG, vanB*).

4.2. Development of bioinformatic tools for in-depth genome analysis

4.2.1. PATO (Pangenome Analysis Toolkit)

PATO exhibits versatility in processing a diverse array of sequence file formats, encompassing whole genome nucleotide files (fna), nucleotide features or CDS files (ffn), and proteome files (faa). Moreover, PATO incorporates a specialized function for loading GFF files, which is a prerequisite for integrating the FASTA sequence at the end of the file. Notably, this format aligns with the output structure of PROKKA (Seemann, 2014), a widely utilized tool for bacterial annotation.

PATO has logically divided its features into four sections to make it easier for users to access its functions: Quality Control, Main Functions, Analysis, and Visualization. This modular organization facilitates a streamlined and intuitive navigation of PATO's functionalities.

Table 2. Table of the PATO functions

Input -> I Output -> O	Functions	Objects								
		files	mmseq	gf _list	core_genome	core_snp_genome	mash	accnet	accnet_enr	igraph
QC	classif er	I		I						
	extract_non_redundant						I/O	I/O		
	non_redundant_hier						I	I		
	non_redundant_pangenomes	I								
	non_redundant				I		I	I		
	outliers						I	I		
	remove_outliers						I/O	I/O		
Main Functions	load_gf _list	I		O						
	accnet		I					O		
	mash	I		I			O			
	mmseqs	I		I			O			
	core_genome		I		O					
	core_snp_genome	I		I		O				
	pagenomes_from_mmseqs		I					O		
	pangenomes_from_f les		I					O		
Visualization	core_plots		I							
	similarity_network						I	I		
	similarity_tree						I	I		
	snps_map					I				
	umap_plot						I	I		
	heatmap_of_annotation									
	network_of_annotation									O
	plot_knnn_network									I
Analysis	accnet_enrichment_analysis							I	O	
	accnet_with_padj							O	I	
	annotate		I							
	clustering						I	I		I
	coincidents							I		
	core_snps_matrix				I	I				
	export_accnet_aln							I		
	export_core_to_fasta				I	I				
	export_to_gephi							I		I
	knnn						I	I		O
	sharedness		I					I		
	singles							I		
	snps_pairwaise	I		I						
	twins							I		
	plamidome			I/O						

4.2.1.1. Functions and classes

PATO incorporates diverse wrapper functions to integrate external software into its framework seamlessly. This includes MMSeqs2, Mash, and Minimap2 binaries, along with supplementary tools like bedtools, k8, and a suite of Perl and JavaScript scripts.

Functionally, PATO is conceived as a comprehensive toolkit. Its modular design enables most functions to utilize outputs from preceding functions, constructing an integrated workflow tailored to analyze extensive genomic datasets. Being an R package, PATO facilitates interoperability with other packages such as ape, vegan, igraph, and pheatmap. Special attention has been devoted to ensuring compatibility with tidyverse packages, leveraging the %>% pipe command. Furthermore, PATO is designed to seamlessly interface with ggplot2 and its extensions, including ggtree, fostering a versatile and adaptable analytical environment.

4.2.1.2. Classes

PATO employs specific data classes (S3 objects) to establish a resilient framework, ensuring compatibility across its functions. The majority of these classes are lists of other kinds of objects, such as data frames, character vectors, matrices, etc.

PATO encompasses eight distinctive object classes:

- mash object
- mmseq object
- accnet object
- nr_list object
- gff_list object
- core_genome object
- core_snps_genome object
- accnet_enr object

Additionally, PATO interfaces with external objects such as igraph or ggplot, expanding its versatility.

The incorporation of these object classes serves the purpose of facilitating the exchange of structured data between functions. These objects are designed for inspection, enabling users to explore their contents, and specific data elements can be leveraged with external functions

or packages, fostering enhanced collaboration and interoperability within the PATO environment.

Mash Class

A Mash class within PATO is a list comprising three fundamental elements: table, matrix, and path.

1. Matrix: This component is a square and symmetric matrix delineating the distances among genomes. The matrix is structured with the input file names serving as both row names and column names.

2. Table: The table is a “data.table” or “data.frame” encapsulating all distances in a list format (long format). Its columns include "Source," "Target," and "Dist," providing a comprehensive representation of pairwise genomic distances.

3. Path: The path element denotes the full file path to the temporary folder. This temporary repository is designed for potential reuse in subsequent runs, optimizing computational efficiency by minimizing processing times.

MMSeq Class

A mmseq object in PATO is a list incorporating three essential components: *table, annot*, and *path*.

1. Table: The first element, table, is a data.table or data.frame featuring four columns denoted as c("Prot_genome", "Prot_Prot", "Genome_genome", "Genome_Prot"). This output, generated by MMSeqs2, elucidates the clustering of input genes or proteins. The first column refers to the genome containing the representative gene or protein of the cluster. The second column designates the representative protein of the cluster, essentially serving as the cluster name. The third column specifies the genome encompassing the gene or protein from the fourth column.

2. Annot: The second element, “annot”, is a data.frame or data.table providing the original annotations for all representative genes or proteins within each cluster. This component comprises two columns: Prot_Prot (mirroring the second column of the table element) and the gene or protein annotation.

3. Path: The third element, path, represents the full file path to the temporary folder. This designated temporary folder can be recycled in subsequent runs, contributing to computational efficiency by abbreviating processing times.

Accnet Class

An accnet object within PATO is a structured list comprising three integral elements: list, matrix, and annot.

1. List: The first element, list, takes the form of a data.frame or data.table featuring columns denoted as c("Source", "Target", "degree"). This component captures the associations among proteins or genes (Target) and genomes (Source), along with the degree of the corresponding protein or gene within the overall network. The third column may vary in specific instances, such as when the accnet object originates from the accnet_with_padj function or pangenomes. It might encompass information such as the protein or gene's frequency in the pangenome (for pangenomes) or the p-value indicating the association between the genome and the protein or gene (for accnet_with_padj).

2. Matrix: The second element, matrix, is a presence/absence matrix critical for understanding genomic content. Notably, this matrix incorporates a "Source" column. In cases where the matrix is extracted from PATO, users are advised to convert this column into row names (e.g., my_matrix <- my_accnet$matrix %>% column_to_rownames("Source")).

3. Annot: The third element, annot, is a data.table or data.frame that encapsulates the original annotation of the representative gene or protein within the given cluster. This element is a crucial reference for interpreting the functional characteristics associated with the genomic entities.

nr_list Class

The nr_list class represents the outcome of non-redundant functions, encapsulating clustering results. Comprising three elements—Source, centrality, and cluster—it is an S3 object that can be converted into a data.frame with three columns using the as.data.frame function. The Source element contains genome accessions or names, centrality represents the genome's centrality based on vertex degrees within its cluster, and cluster indicates the membership of the Source genome in the cluster.

gff_list Class

A gff_list is an object storing results from the load_gff_list() output. It includes paths to GFF, FNA, FFN, and FAA files. Utilizable as input for classifier, mash, mmseq, core_snp_genome, and snps_pairwise functions, a gff_list object comprises two elements: path, indicating the pathname to the temporary directory, and files, containing subdirectories (FAA, FNA, FFN, and GFF) storing corresponding FASTA files and GFF. This temporary folder can be recycled in subsequent runs to optimize computational efficiency.

core_genome and core_snps_genome

Objects generated by core_ functions, these contain alignment information (core or core_snps) and can be exported to FASTA format or used in other functions. The core_genome class has two elements: core_genome (a data.frame with columns "Genome" and "Seq," containing the core-genome alignment) and path (indicating the temporary directory). On the other hand, core_snps_genome comprises alignment, bed (a BED table with conserved regions of the alignment, featuring coordinates in the reference base), path (pathname to the temporary directory), and reference (indicating the chosen reference genome).

accnet_enr Class

An object resulting from accnet_enrichment_analysis(), this data.frame can be exported as a gephi file, including adjusted p-values as edge weights. It encompasses columns such as Target (Protein), Source (Genome), Cluster (Cluster of genome), and various informative measures, including frequencies, odds ratios, p-values, and annotations about protein functionality.

4.2.1.3. Main functions

The principal functions within PATO serve as the foundation for subsequent analysis, visualization, and quality control operations. Primarily acting as wrappers for external software, these functions initiate the execution of the respective external tools, subsequently processing, transforming, and constructing corresponding R objects with the derived results.

PATO relies on MMSeqs2 as a pivotal component for clustering family proteins and/or genes, a critical step in constructing both the core-genome and the accessory genome. Recognized for its superior speed and accuracy, MMSeqs2 stands out as one of the most

efficient gene and protein clustering methods. Additionally, integrating MASH facilitates rapid estimation of sequence similarity among the datasets.

Furthermore, PATO leverages Minimap2 to generate whole-genome alignments and construct the core SNP genome, which entails identifying all shared polymorphic positions across the set of genomes. Complementing Minimap2, PATO incorporates the paftools.js software application for variant calling based on the results from Minimap2. This process adheres to the pipeline recommended by Minimap2 developers, ensuring robust variant identification within the genomic dataset.

load_gff_list()

PATO exhibits versatility in managing diverse sequence files, encompassing whole genome, CDS, or amino acid files, albeit not universally across all functions. A unifying approach to consolidating this information involves utilizing GFF files. Crucially, these GFF files MUST conclude with the sequence in FASTA format, a convention adhered to by tools such as PROKKA. The load_gff_list() function adeptly parses these GFF files, establishing distinct directories to house various file types—FNA (whole genome sequence), FFN (CDS sequence), and FAA (amino acid sequence)—in addition to the original GFF files. Users retain the flexibility to specify the desired file type for use in most main functions.

Notably, the objects generated by load_gff_list() possess reusability across different R sessions, facilitating seamless integration and continuity of analyses.

mash()

The mash() function serves as a wrapper for the Mash tool.

Mash, an acronym for "Fast genome and metagenome distance estimation using MinHash," is a methodology designed to swiftly estimate genomic and metagenomic distances by applying MinHash techniques. MASH accommodates both nucleotide and amino acid fasta files, employing them to derive an estimation of similarity distance.

mmseqs()

The mmseqs() function operates as a wrapper for the MMSeqs2 software.

MMseqs2, which stands for "Many-against-Many sequence searching," is a robust software suite designed for the exploration and clustering of extensive protein and nucleotide sequence datasets. The algorithm begins by renaming all files, generating a unique header

for each gene or protein. Subsequently, the algorithm executes sequence clustering using the mmseqs linclust pipeline. The resultant clustering outcomes are then imported into R and parsed to construct the mmseq object.

mmseqs() exhibits versatility in its input, accommodating either a list of files or a gff_list object. However, in the latter case, the function exclusively supports feature files, namely genes (FFN) or proteins (FAA).

accnet()

Accnet, a representation of the "accessory genome" utilizing a bipartite network, is a pivotal component in PATO's analytical framework (refer to Lanza, V. F., Baquero, F., de la Cruz, F., & Coque, T. M. AcCNET [Accessory Genome Constellation Network]: Comparative genomics software for accessory genome analysis using bipartite networks. Bioinformatics, 33, 283–285 [2017]). The accnet() function constructs a graph featuring two distinct node types: genomes and proteins/genes. A connection is established between a genome and a protein/gene if the genome harbors the corresponding protein—effectively delineating protein families or clusters.

The accnet object, a primary data type in PATO, facilitates various representations of the accessory genome, including dendrograms, networks, and plots. PATO incorporates dedicated functions to analyze the accessory genome comprehensively.

In the accnet() function, the accnet object is generated, with user-defined parameters influencing the maximum frequency criterion for a protein/gene to be considered part of the accessory genome. Users can also decide whether to include single protein/genes—those present in only one sample.

Critical to the accnet function is the mmseqs object, which serves as the input. This reliance on protein families' definition underscores the importance of homologous cluster determination in accurately characterizing the accessory genome.

pangenomes_from_files()

The pangenomes_from_files() function performs clustering with a specified maximum distance among samples, generating a comprehensive set of pangenomes. Genome frequency can further refine these pangenomes, allowing users to filter based on the number of genomes within each. The function constructs a new accnet object, establishing

associations between pangenomes and the respective genes/proteins. Additionally, users can filter included proteins based on their presence frequency in the dataset.

Specifically designed to handle substantial datasets, exceeding 5,000 genomes, this function operates as a tailored pipeline for accelerated accnet object generation.

pangenomes_from_mmseqs()

The pangenomes_from_mmseqs() function creates genome clusters (pan-genomes) using a predefined list, which can originate externally (e.g., from mlst or BAPS) or internally (e.g., clustering function). Similar to the pangenomes_from_files() function, it allows for pangenome refinement based on genome frequency. The resulting accnet object connects the pangenome and associated genes/proteins. Users can apply protein filtering based on their presence frequency. However, it is advisable to refrain from using this function for datasets larger than 5,000 genomes due to its dependence on an mmseq object, which might surpass computational capabilities.

core_genome()

The core_genome() function identifies and generates a hard core-genome alignment, comprising genes in 100% of genomes without repetitions. Utilizing the output of mmseqs(), this function defines paralogous genes of the core genome based on parameters set in mmseqs(). The function employs a rapid pseudo-multiple sequence alignment (MSA) method, leveraging mmseqs2's result2msa for protein sequences and a blast-based approach for DNA sequences (requiring blast installation in the PATH). Although less accurate than classical MSAs (clustal, mafft, or muscle), this approach is notably faster. Given that many phylogenetic inference tools focus on variant columns without insertions or deletions, the impact on final phylogenetic trees is minimal. Impressively, core_genome() can construct a core-genome alignment for thousands of genomes in minutes, as demonstrated in the Benchmarking section.

core_snp_genome()

The core_snp_genome() function is designed to identify the core SNP genome. This method employs minimap2 to align all genomes against a reference genome. In cases where a reference genome is not explicitly specified, core_snp_genome() randomly selects one. Following the alignment process, conserved regions across all genomes are identified.

Subsequently, variant calling is executed using paftools, an integral tool bundled with minimap2.

In the final step, the identified SNPs are refined by filtering them through common regions, yielding the ultimate core SNP genome. This comprehensive process ensures the identification of shared polymorphic positions across the entire dataset, providing a refined representation of the core SNP genome.

4.2.1.4. Analysis

PATO incorporates a comprehensive suite of functions dedicated to creating and analyzing pangenomes. The underlying concept involves treating each pangenome as an individual sample, enabling holistic analyses of the entire dataset. Depending on dataset size, PATO deploys two distinct pipelines—one tailored for datasets smaller than 5000 genomes and another optimized for larger datasets exceeding 5000 genomes. The functionality encompasses the construction of genome groups, conversion into pangenomes, and utilization of user-defined or pre-existing pangenomes.

Upon obtaining pangenomes, PATO facilitates the generation of an accnet object, allowing for analysis with standard tools implemented for accnet objects. Remarkably, PATO enables the streamlined analysis of tens of thousands of genomes within a single study (refer to Supplementary 'Pangenome Analysis').

Additionally, PATO integrates the latest version of AccNET (Lanza et al., 2017), along with requisite functions for the comprehensive analysis of accessory genome networks.

Within the realm of analysis-specific functions, PATO includes tools for statistical analyses of the accessory genome and gene associations (coincidences, twins, singles), clustering functions for both whole genomes and accessory genomes, plasmid sequence extraction, SNP distribution analysis, and feature annotation for antibiotic resistance genes and virulence factors.

Furthermore, PATO features a pre-existing pipeline for the enrichment analysis of genes within accessory genomes (Fernández-de-Bobadilla et al., 2020). This functionality discerns over-represented genes within a specific group compared to the entire population—a valuable insight into gene enrichment patterns.

PATO's repertoire extends to constructing SNP matrices, utilizing the mlplasmids package (Arredondo-Alonso et al., 2018) for extracting the plasmidome (installation required) and offering functions to export data for network analysis, visualization, and phylogenetic reconstruction to external software applications. The multifaceted capabilities of PATO empower researchers to conduct thorough analyses and gain nuanced insights from large-scale genomic datasets.

accnet_enrichment_analysis()

The accnet_enrichment_analysis() function performs a statistical assessment, explicitly employing the Hypergeometric test (phyper), on genes/proteins within a specified cluster. Leveraging cluster definitions from the cluster table, this test evaluates the statistical significance of gene/protein frequency within the cluster compared to its frequency in the overall population. This robust analytical approach provides valuable insights into the enrichment patterns of genes/proteins within defined clusters.

accnet_with_padj()

The accnet_with_padj() function is dedicated to constructing and enhancing networks, utilizing adjusted p-values as edge weights. The resultant network is exportable to Gephi (export_to_gephi) for accurate visualization. This function contributes to a more nuanced understanding of relationships between nodes, enhancing the interpretability of the generated networks.

annotate()

The annotate() function serves the purpose of annotating virulence factors and antibiotic resistance genes within genomes (files). Utilizing the mmseqs2 software and databases such as the Virulence Factor Database and ResFinder, the annotation process can either re-use previous computational steps from mmseqs or establish a new index database from the provided files. Opting for re-use substantially reduces computational time. Noteworthy is the use of mmseqs2's search algorithm, which selectively returns high-identity matches, ensuring a focused and accurate annotation process.

clustering()

The clustering() function in PATO serves the purpose of clustering genomes based on mash data, accnet data, or igraph data. It accommodates objects generated by the accnet, mash,

and knnn functions. The specific clustering methodology varies depending on the type of input object.

For accnet objects:

- mclust: This method employs Gaussian Finite Mixture Models for clustering, which can be combined with dimensionality reduction (d_reduction). The Mclust package is utilized, and the function is designed to identify the optimal cluster number.

- upgma: Hierarchical Clustering is performed using the UPGMA algorithm, requiring the user to provide the desired number of clusters (n_cluster).

- ward.D2: This method executes Hierarchical Clustering using the Ward algorithm, with the user specifying the number of clusters (n_cluster).

- hdbscan: Density-based spatial clustering of applications with noise is implemented using the DBSCAN package. This method automatically determines the optimal number of clusters.

For mash objects:

- mclust: Similar to the accnet objects, this method employs Gaussian Finite Mixture Models and can be combined with dimensionality reduction (d_reduction). The Mclust package is utilized, and the function is tailored to identify the optimal cluster number.

- upgma: Similar to the accnet objects, this method performs Hierarchical Clustering using the UPGMA algorithm, requiring the user to provide the desired number of clusters (n_cluster).

- ward.D2: Similar to the accnet objects, this method executes Hierarchical Clustering using the Ward algorithm, with the user specifying the number of clusters (n_cluster).

- hdbscan: Similar to the accnet objects, this method performs Density-based spatial clustering of applications with noise using the DBSCAN package, determining the optimal number of clusters.

For igraph objects:

- greedy: This method identifies community structure via greedy optimization of modularity.

- louvain: This method implements community structure identification using the multi-level modularity optimization algorithm.

- walktrap: Community structure is determined via short random walks.

Clustering of igraph objects relies on the network building process (see knnn function), and the number of clusters may vary with different settings of the k-nn network. Network-based methods are computationally faster than distance-based methods, offering efficiency in handling large datasets.

coincidents()

The coincidents() function in PATO identifies groups of coincident genes/proteins within an accnet object. In contrast to accnet_enrichment_analysis(), coincidents() don't necessitate the pre-definition of genome/pangenome clusters. The function employs multi-dimensional scaling through UMAP and subsequently identifies clusters using HDBSCAN.

core_snps_matrix()

This function computes the number of SNPs among samples within the core genome and produces a SNP matrix. Results can be presented in a normalized form by the alignment length (SNP per Megabase) or as the raw count of SNPs. Additionally, it offers the option to exclude columns containing one or more insertions/deletions, specifically in the case of core_genome objects.

export_accnet_aln()

export_accnet_aln() writes the presence/absence matrix into a FASTA alignment file (1 = A, 0 = C), compatible with external software.

export_core_to_fasta()

This function exports the alignments of the core-genome and core_snp_genome to a multi-alignment FASTA file.

export_to_gephi()

export_to_gephi() facilitates the export of a network (accnet or igraph) to the universal graphml format. Larger networks, particularly accnet networks, may exceed visualization capabilities within the R environment. It is recommended to use external software such as Gephi or Cytoscape. Gephi, in particular, can efficiently handle larger networks, and the

ForceAtlas2 layout is adept at arranging accnet networks with up to 1,000 genomes. If the accnet object originates from the accnet_with_padj() function, the graph files will incorporate edges-weight proportional to the p-value associating the protein with the genome.

knnn()

knnn() constructs a network with the k best neighbors for each genome or pangenome. The network can be generated with or without repetitions among the best neighbors. The 'repeats' option determines whether the best k neighbors include bidirectional links. If G_i and G_j are genomes, and G_j is one of the best k-neighbors of G_i, the repeats option, if TRUE, removes G_i from the k-neighbor list and replaces it with the next-best neighbor. If FALSE, G_i remains in the list.

plasmidome()

The plasmidome() function in PATO leverages the mlplasmids package to identify contigs associated with plasmids. It's important to note that mlplasmids currently supports processing genomes of *Enterococcus faecium*, *Escherichia coli,* or *Klebsiella pneumoniae*. PATO does not include the mlplasmids package by default; therefore, manual installation on the user's computer is required. The input genomes must be in gff_list format (refer to the load_gff_list() function). Plasmidome identifies contigs attributed to plasmids and generates a new GFF structure containing comprehensive features information, including FAA, FNA, FFN, and GFF.

sharedness()

The sharedness() function within PATO constructs a matrix detailing the count of shared genes/proteins across multiple samples.

singles()

The singles() function identifies genes/proteins exclusively present in individual genomes or clusters of genomes.

snps_pairwise()

snps_pairwise() calculates the total number of Single Nucleotide Polymorphisms (SNPs) between all pairs of sequences. Unlike core_snps_matrix, this function computes the raw count of SNPs for each sequence pair, mitigating biases related to core genome size and/or

reference. However, due to its pairwise nature, this function can be computationally intensive for large datasets.

twins()

Twins nodes in PATO refer to genes/proteins sharing identical edges, indicating they are present in the same set of genomes or samples within an accnet object.

4.2.1.5. Quality control

The Quality Control module of PATO encompasses essential functions designed to systematically filter outliers and identify species within genomic datasets. Instances of sequences that deviate from expected patterns, whether originating from public databases or our proprietary datasets, are effectively identified and classified by PATO. PATO employs a tailored reference database containing comprehensive reference and representative bacterial genomes from RefSeq to facilitate this process.

Redundancies in datasets pose another common challenge, particularly with databases exhibiting disproportionate representations of certain phylogenetic branches, such as isolates from outbreaks. PATO addresses this issue by implementing functions dedicated to constructing a non-redundant representation of genomes, ensuring a more balanced and representative dataset. This module enhances the overall quality and reliability of subsequent analyses within the PATO framework.

classifier()

The classifier() function in PATO processes a list of genome files, either nucleotide or protein sequences, to determine the most similar species for each file. Leveraging the NCBI RefSeq database's complete collection of reference and representative genomes, the classifier() employs Mash to identify the best-matching reference genome for each input file.

extract_non_redundant()

The extract_non_redundant() function creates a non-redundant representation of the input object (accnet or mash). It is recommended that mash objects be used to eliminate redundancies and subsequently perform mmseq() before accnet(). Changes in the dataset may impact gene/protein clustering and core-genome definition.

non_redundant()

The non_redundant() function generates a non-redundant subset of samples from an accnet or mash object, producing an nr_list object. This subset can be defined based on a specific distance (mash or Jaccard distance, depending on the input), a specified number of elements, or a fraction of the entire sample set. The function performs an iterative search, and the exact number of returned elements may slightly differ from the specified input due to the threshold parameter.

non_redundant_hier()

The non_redundant_hier() function creates a non-redundant subset of samples from an accnet or mash object, returning an nr_list object. Similar to non_redundant(), it allows customization based on distance, the number of elements, or a fraction of the sample set. This function, however, performs a hierarchical search, consuming less memory and proving more efficient for large datasets (>2000).

non_redundant_pangenomes()

The non_redundant_pangenomes() function eliminates redundant sequences from a list of nucleotide or protein sequences. Specifically designed for vast datasets (>10,000), this function accepts a distance threshold for removing similar sequences.

outliers()

The outliers() function identifies outliers in the dataset by applying a threshold distance (mash or Jaccard distance, depending on the input) or through standard deviation multiples. The function also generates a Manhattan plot for visualizing the distribution structure of the input object's distances.

remove_outliers()

The remove_outliers() function eliminates outliers or other specified elements from an accnet or mash object.

4.2.1.6. Visualization

The utilization of R in PATO is motivated by the extensive repertoire of packages available for robust data visualization. PATO incorporates a suite of functions dedicated to visualizing analysis outcomes. Standard visualizations cover core, accessory, and pangenome distributions, beta diversity employing dimensional reduction through UMAP (Uniform

Manifold Approximation and Projection) as proposed by Becht et al. (2018), and similarity trees, including MASH trees and accessory trees.

A distinctive contribution lies in our innovative representation of population structure. We introduce the K-nearest neighbor network visualization by leveraging a combination of k-nearest neighbor (k-NN) and spanning tree concepts. This function meticulously selects the optimal K neighbors for each genome, constructing an informative network. Notably, our pioneering approach incorporates the option to avoid reciprocal connections, ensuring that each genome's set of nearest neighbors remains unconnected among themselves. This novel feature enhances the discernment of relationships within the population structure, offering a nuanced perspective on genetic associations.

core_plots()

The core_plots() function, designed for mmseq objects, facilitates the visualization of core, accessory, and pan-genome data at specified points with a defined number of replicates. Typically, core-genome elements are expected to be present in all dataset genomes. However, some genes/proteins may be absent inherent to random selection and sequencing errors. To address this, core_plots() employs bootstrapping to generate a distribution at each point, enhancing the robustness of the analysis.

similarity_network()

This function constructs a similarity network, linking genomes with distances below a user-defined threshold. The suitability of the specified distance is contingent upon the data type; for instance, MASH values may range between 0-0.05 for intra-species connections, while accnet distances might assume higher values. Distinguishing it from knnn(), each genome in similarity_network() connects to all genomes within the stipulated threshold distance.

similarity_tree()

similarity_tree() generates a similarity tree (pseudo-phylogenetic) based on samples, utilizing either whole-genome data (mash) or accessory genome information (accnet).

snps_map()

The snps_map() function produces a plot akin to Manhattan plots, illustrating the cumulative number of SNPs per contig-position pair.

umap_plot()

umap_plot() employs UMAP, a dimension reduction technique, to create a 2D plot representing the population structure of the dataset (accnet or mash). The implementation relies on the UMAP function from the uwot package.

heatmap_of_annotation()

heatmap_of_annotation() generates a heatmap based on the results obtained from the annotate() function. Users can apply filters based on identity and/or E-value. If installed, the function uses pheatmap instead of the heatmap function.

network_of_annotation()

network_of_annotation() constructs an igraph object using the results from the annotate() function. It offers flexibility in filtering results by identity and/or E-value.

plot_knnn_network()

plot_knnn_network() leverages the threejs R package to visualize the k-nnn network, providing options to select layout algorithms and visualize various types of clustering. The function returns a list containing the network, layout, and colors for further use in other analyses.

4.2.1.7. Case study

Quality Control and DataSet manipulation

The provided example involves the analysis of 2,941 high-quality *E. coli* and *Shigella* genomes obtained from the NCBI Assembly database. A copy of the dataset can be accessed at https://doi.org/10.6084/m9.figshare.13049654. The dataset primarily comprises protein multi-fasta files (*.faa).

PATO demonstrates versatility in handling nucleotide and amino acid multi-fasta files and GFF files containing FASTA sequences, such as those generated by PROKKA. While PATO can utilize whole-genome sequence files, certain functions restrict their use with this data type.

Species Classification and Outliers Detection

PATO incorporates a classification function designed to discern the most likely species for each input file. This methodology relies on determining the minimum Mash distance of each

file in comparison to a curated reference database. The reference database encompasses all reference and representative genomes sourced from RefSeq. Input files can consist of either nucleotide or amino acid sequences in FASTA format.

Table 3. Species classification data frame.

organism_name	Number of genomes
Citrobacter gillenii	1
Escherichia coli O157:H7 str. Sakai	323
Escherichia coli str. K-12 substr. MG1655	2520
Escherichia marmotae	4
Shigella boydii	18
Shigella dysenteriae	7
Shigella flexneri 2a str. 301	45
Shigella sonnei	40

When downloading genomes from public databases, encountering outliers—genomes deviating from the selected species—can be expected. PATO addresses this challenge by implementing a dedicated function to identify and, if necessary, remove outliers. In this context, a predefined threshold of 0.06 (corresponding to approximately 94% Average Nucleotide Identity, ANI) is employed to classify a genome as an outlier (as this is the fastest way to cluster species). This ensures rigorous filtering of genomes that significantly differ from the intended species.

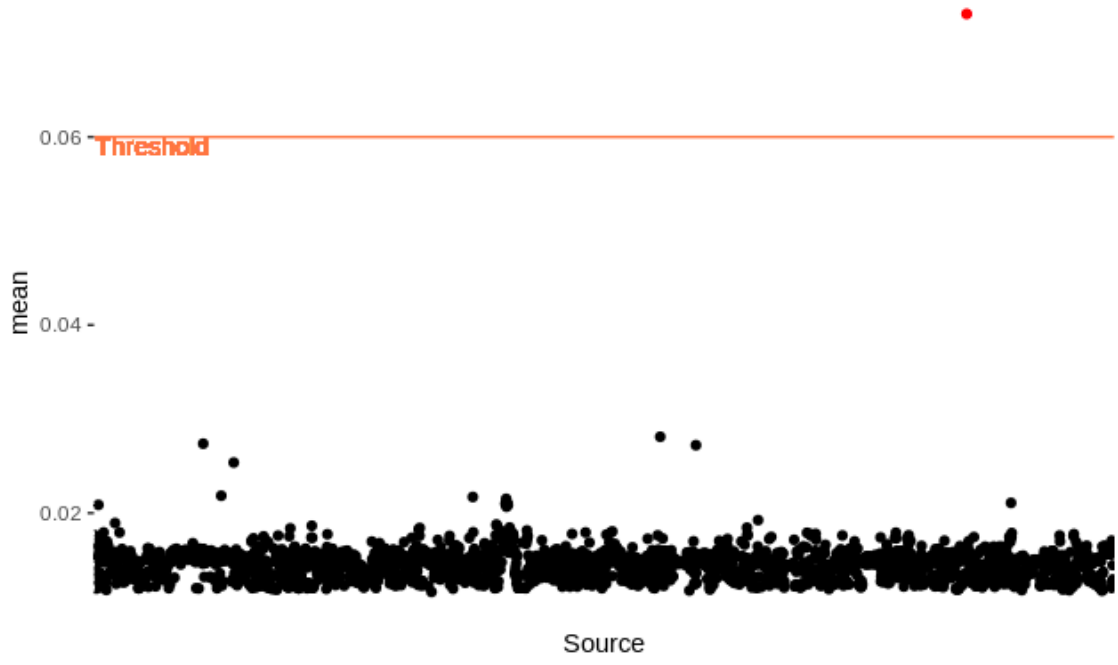


Figure 9. Outliers plot

Redundancies

The prevalence of redundancies, often induced by oversampling of specific strains such as those associated with outbreaks or overrepresented bacterial lineages, poses a significant challenge in microbial diversity analysis within databases. PATO incorporates a functionality designed to curate a non-redundant subset to address this issue. Users can tailor the selection criteria based on a fixed distance (similarity), a percentage of samples, or a specific number of genomes. Using an iterative approach, the function considers a tolerance threshold, recognizing that it may not always be possible to find an exact solution.

Additionally, PATO introduces the function non_redundant_hier(), which employs a hierarchical clustering approach to handle large datasets, particularly those exceeding 1000 genomes, efficiently.

This function adopts a "divide and conquer" strategy, where the dataset is split into distinct subsets that are clustered independently before the results are combined again. This iterative process continues until the desired number of non-redundant elements is attained. The hierarchical methodology enhances computational efficiency, which is particularly well-suited when dealing with large datasets for which traditional methods resulted cumbersome.

Population Structure / Pangenome Description

In this example, we continue with the high-quality E. coli and Shigella spp dataset (you can download a copy from https://doi.org/10.6084/m9.figshare.13049654) after the QC described above.

Core genome analysis

Within PATO, an array of tools is incorporated to scrutinize the core genome comprehensively. This encompasses exploring the pangenome's structural components, and delineating the core, accessory, and pangenome elements. Additionally, functionalities for crafting a core-genome alignment or a core-snp-genome are provided.

For a detailed understanding, an examination of the dataset's pangenome composition is detailed below :

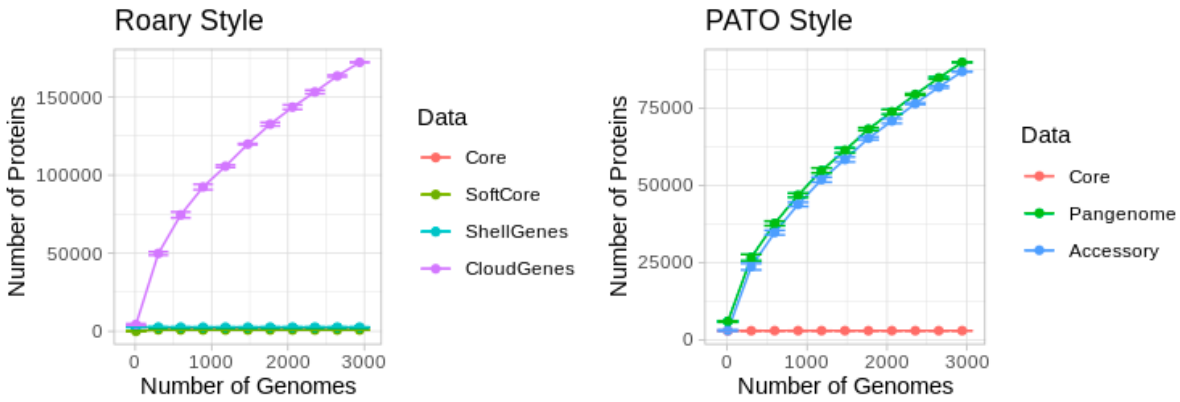


Figure 10. Plots of the pangenome composition

PATO facilitates the construction of a core-genome alignment suitable for integration with external software tools, including IQTree (http://www.iqtree.org/), RAxML-NG (http://www.exelixis-lab.org/software.html), FastTree (http://www.microbesonline.org/fasttree/, and various other phylogenetic inference applications. The computation of the core genome relies on a mmseq object. Thus, it is worth noting that the definition of the core-genome i can vary depending on these parameters.

For this exercise, we will continue with the example of *E. coli*. We generate a core-genome alignment comprising approximately 800 *E. coli* genomes derived from the non-redundant dataset. We selected the best sample from each cluster, which was determined by the

maximum centrality value. In cases where multiple samples share the same centrality value within a cluster, we select the first one. Subsequently, we initiate the core genome analysis for this curated list of samples.

The resulting files are exported in Multi-alignment FASTA format, which is compatible with a broad array of phylogenetic tools. We opt to utilize FastTree for subsequent phylogenetic inference.

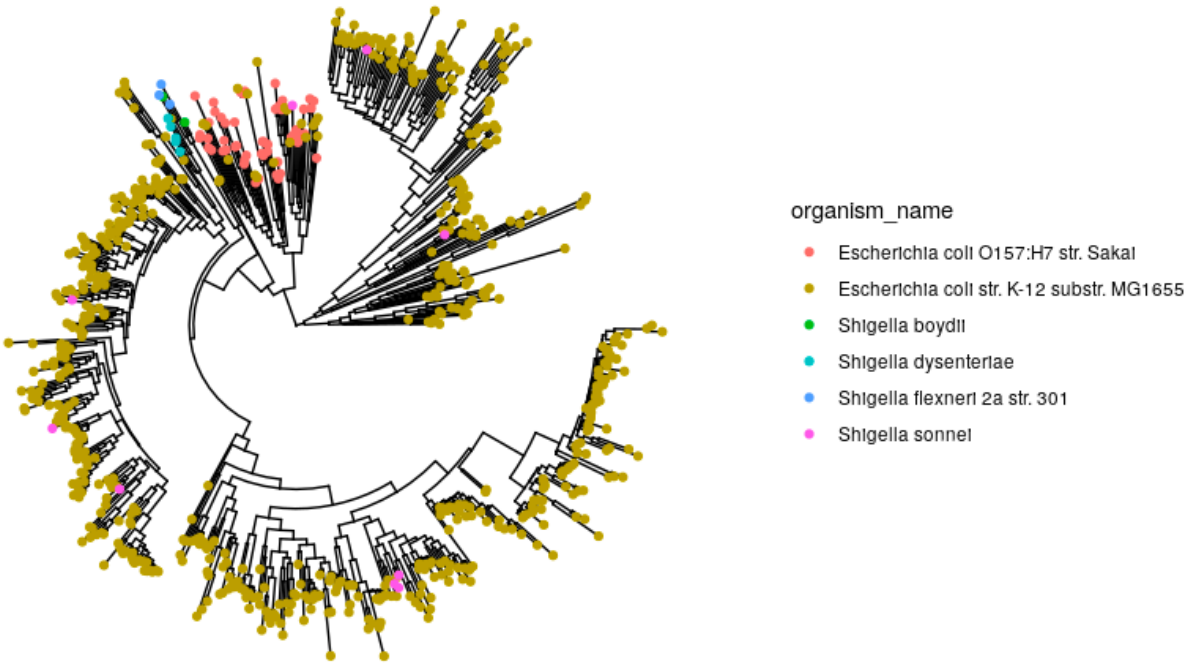


Figure 11. ML Core-genome phylogenetic tree

In the conclusive phase of the analysis, a SNPs matrix can be derived from the core genome alignment. This matrix provides a comprehensive overview of the total number of SNPs (or aa changes in the context of protein alignments) shared among the samples. The matrix allows for flexible normalization either by the total length of the genome measured in megabases or as a raw count of genetic variants.

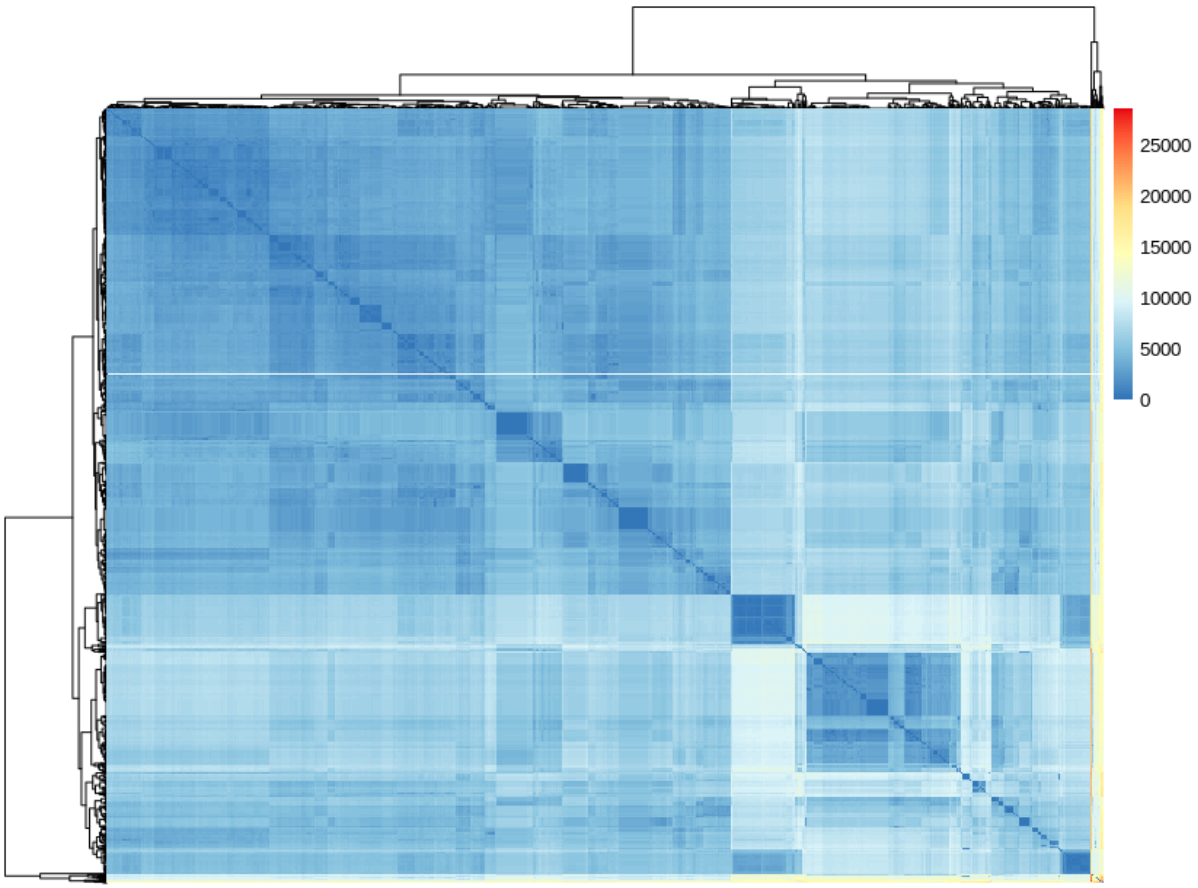


Figure 12. HeatMap of the core-snp-genome SNP matrix

Accessory Genome Analysis

We delineate the accessory genome by considering proteins present in no more than 80% of the genomes using the specialized object “accnet”, along with functions for its analysis and visualization, which include the intricate relationships between genes/proteins and genomes. Visualizing an accnet object unveils its bipartite network structure.

A compelling facet of accessory genome analysis is the identification of over-represented genes/proteins within a set of genomes. The accnet_enrichment_analysis() function scrutinizes the over-representation of genes/proteins in a cluster in contrast to the entire population, offering insights into accessory genome dynamics. Clusters can be externally defined based on categorical metadata (e.g., Sequence Type, Source, Serotype) or internally generated through clustering processes. It is crucial to account for redundancies that may skew analysis outcomes. Biases may emerge if certain strains/genomes dominate the dataset, leading to an overestimation of significant genes/proteins due to uneven dataset diversity.

To mitigate this, we recommend utilizing a non-redundant set of samples. In this example, we select ed 800 non-redundant genomes.

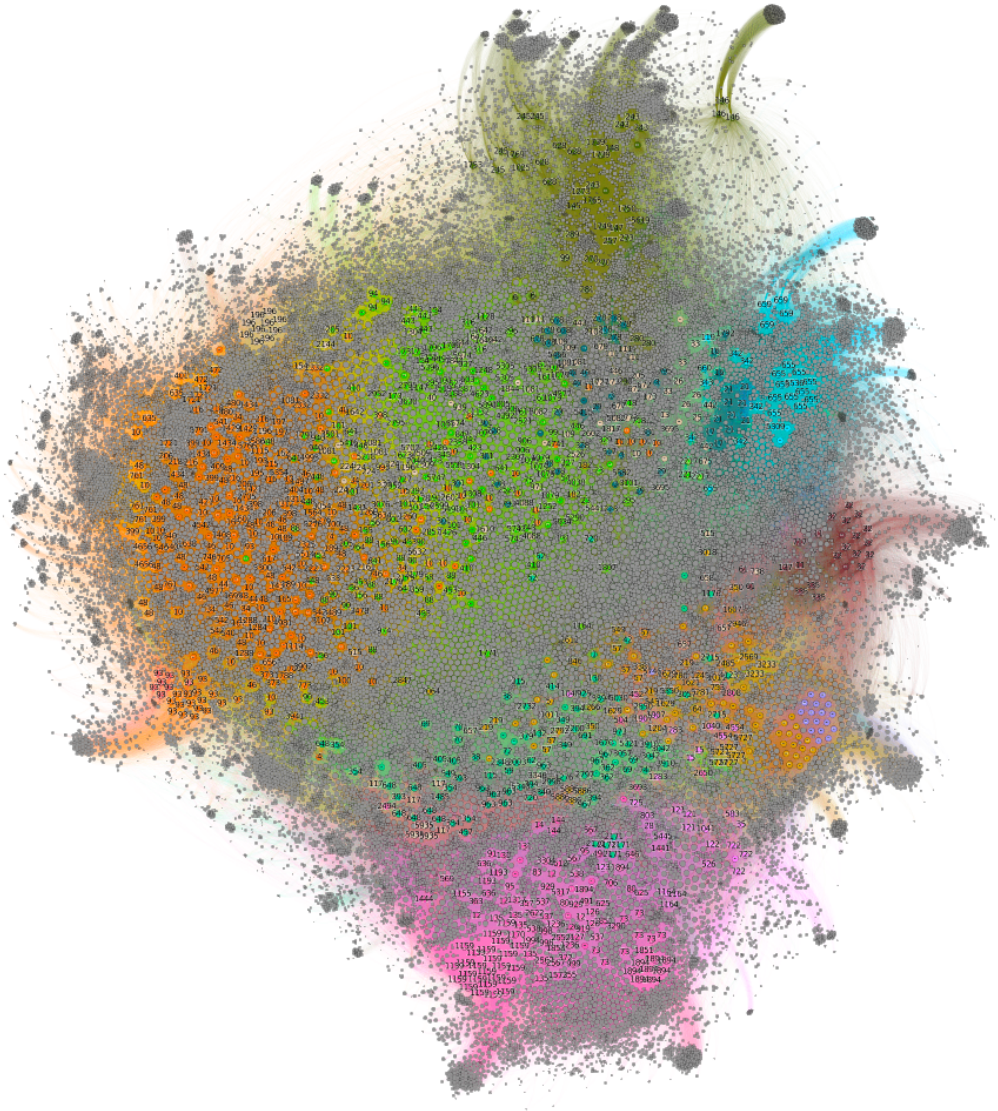


Figure 13. Non-redundant Accnet network, each color represents a strain.

New network with the adjusted p-values as edge-weight after performing an enrichment analysis.

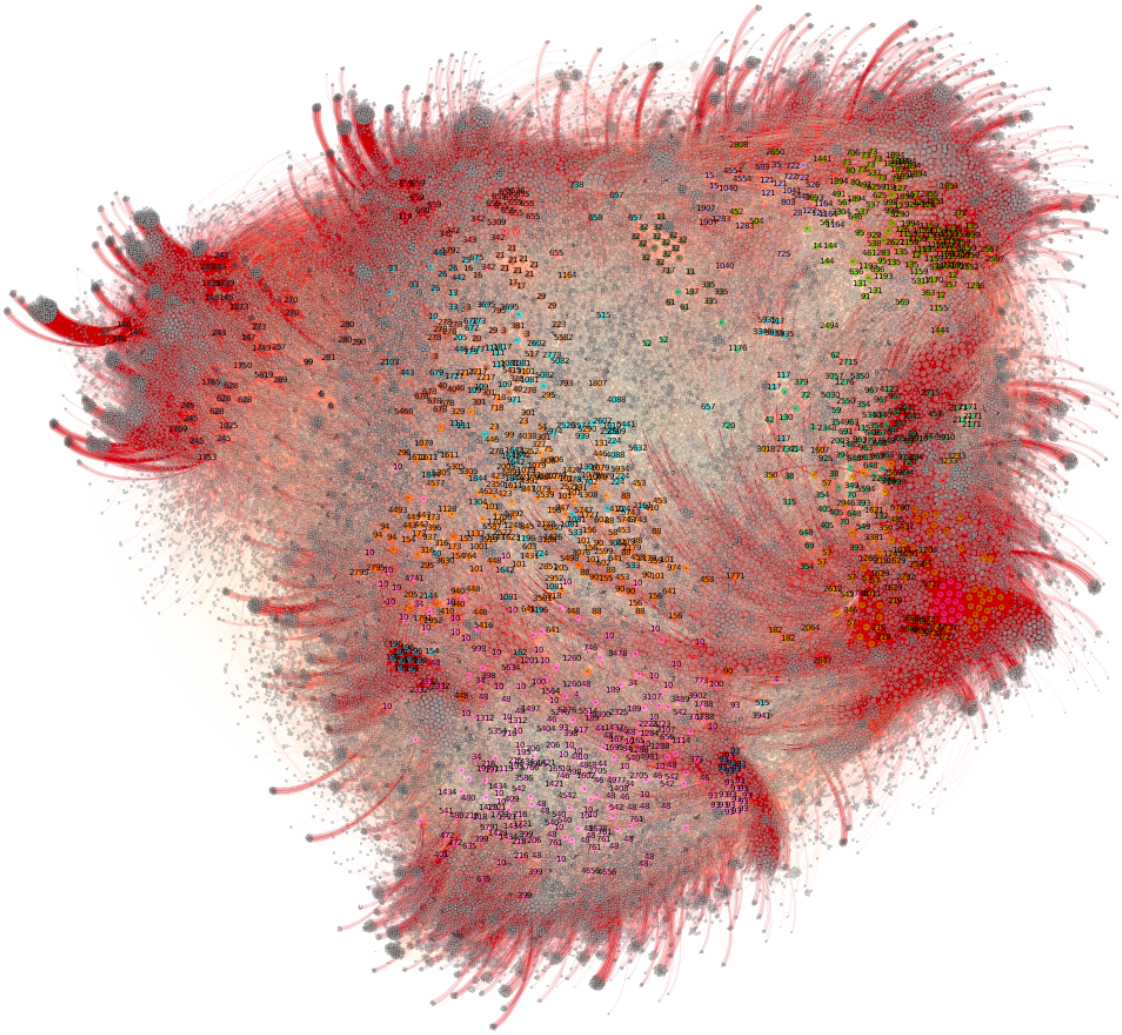


Figure 14. Accnet network with enrichment analysis added. Edge colors are proportional to the p-value (adjusted).

PATO incorporates specific functions designed to unravel the distribution patterns of genes/proteins, namely singles(), twins(), and coincidents().

singles(): Identifies genes/proteins exclusive to a single sample (genome, pangenome, etc.).

twins(): Discerns genes/proteins sharing identical connections, indicating common presence in the same genomes.

coincidents(): Identifies genes/proteins with analogous connections, implying a propensity to co-occur.

Distinguishing between twins() and coincidents() involves several key factors. coincidents() offers higher flexibility, rendering it less sensitive to outliers, but it may exhibit slower speeds. On the other hand, twins() excels in terms of speed, accuracy, and sensitivity to noise or outliers, as any missed connection automatically excludes a protein from the twin group. Notably, the swifter processing of twins() occasionally results in the formation of substantial clusters resembling pseudo-cores.

Population Structure

PATO provides robust tools for dissecting population structures. This involves scrutinizing the entire genome's population structure, grounded in MASH data, or the accessory genome structure using AccNET data.

Dataset visualization is facilitated through dendrograms, essentially tree-like representations. PATO offers the capability to visualize both MASH and AccNET data as dendrograms. In the case of AccNET data, PATO initially computes a distance matrix utilizing the presence/absence matrix coupled with the Jaccard distance metrics. The similarity_tree() function encompasses various methodologies for dendrogram construction. Phylogenetic approaches, such as Neighbor-Joining and FastME (Minimum Evolution), as well as hierarchical clustering methods like complete linkage, UPGMA (Unweighted Pair Group Method with Arithmetic Mean), Ward's minimum variance, and WPGMC (Weighted Pair Group Method with Centroids), are incorporated.

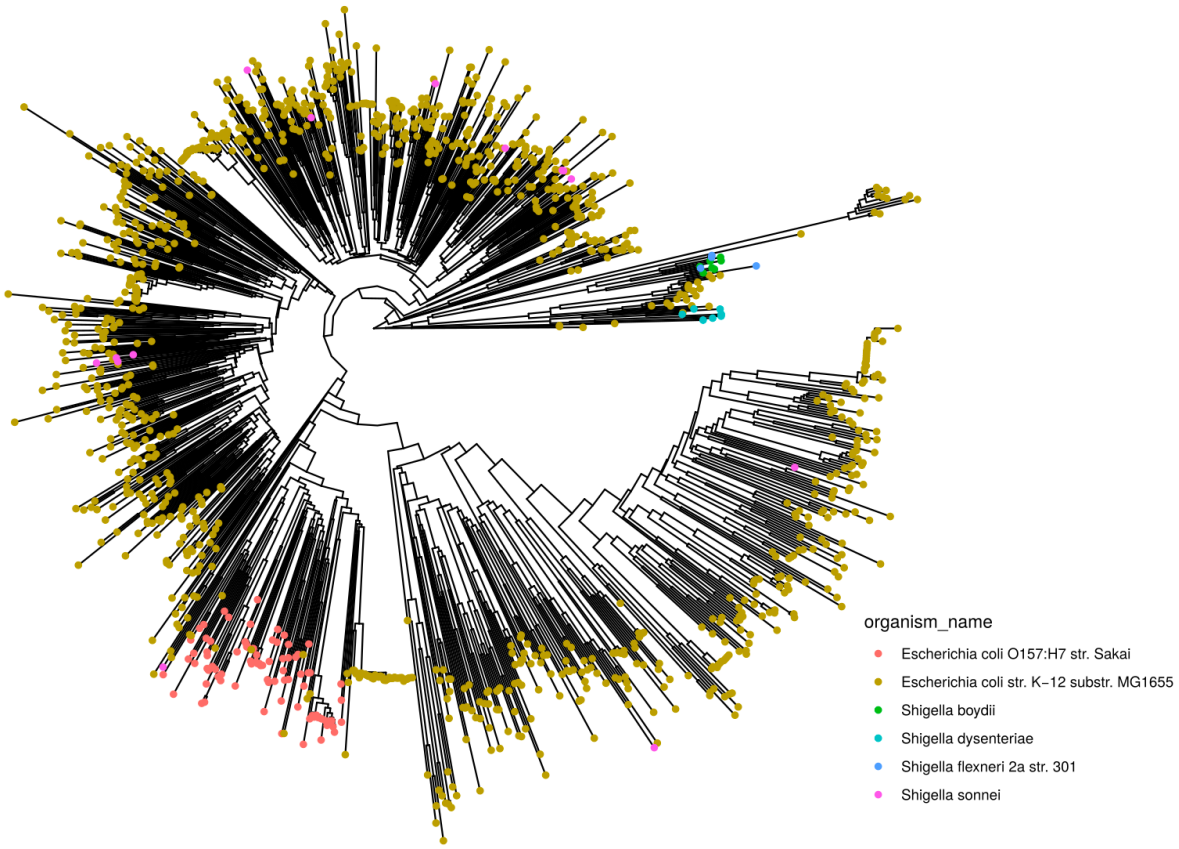


Figure 15: Neighbor-Joining tree of MASH distances

Certain Maximum Likelihood inference tree software accepts binary data (0-1) alignments as input. Consequently, using this binary representation, we can leverage accessory data (accnet) to deduce a tree, albeit non-phylogenetic. This parallels the functionality of similarity_tree(); however, rather than relying on distance metrics, it is grounded in Maximum Likelihood principles. The export_accnet_aln() function can be employed to export this alignment for further analysis.

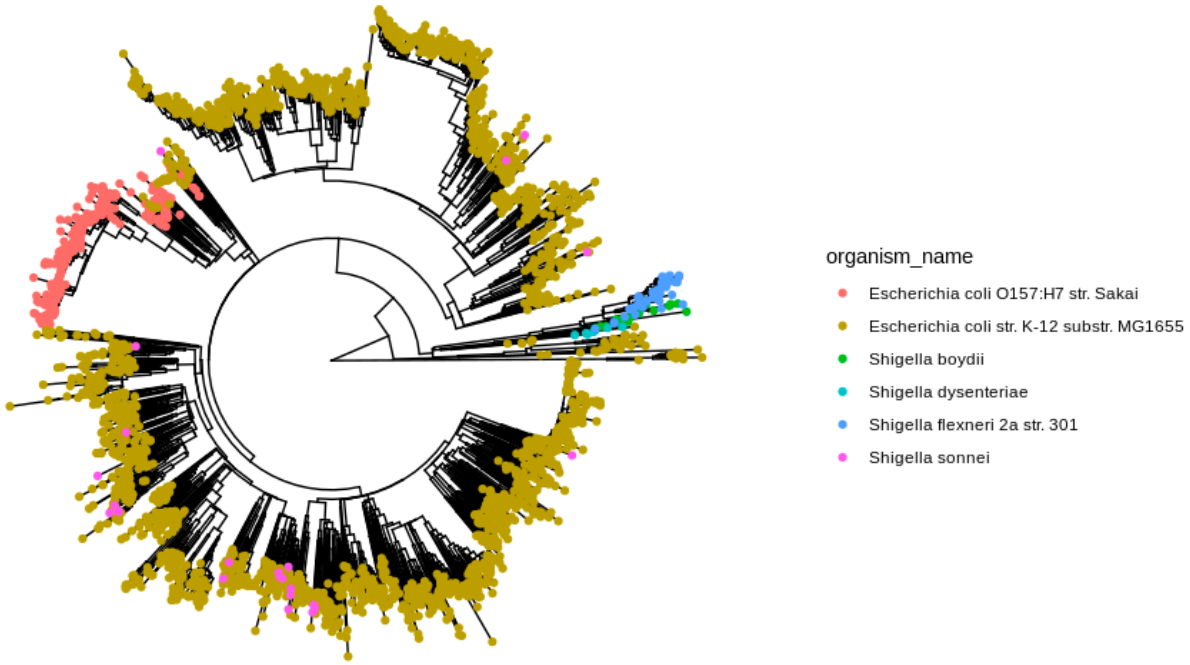


Figure 16: ML tree of accessory genome

Binary alignments generated from accessory data can be substantial in size. Accessory genomes often contain numerous spurious genes/proteins that contribute little information to the alignment. To mitigate this, export_accnet_aln() incorporates a min_freq parameter, allowing users to filter genes/proteins based on their frequency. Implementing this option significantly reduces alignment length, enhancing computational efficiency.

PATO offers a suite of functions for data visualization. While tree-related functions have been discussed, the platform also features methods for visualizing genome relationships as networks. The K-Nearest Neighbor Network (knnn()) provides a network plot depicting data relationships. The resulting igraph object can be visualized using various packages. PATO introduces its own function, plot_knnn_network(), utilizing threejs and its igraph for network layout and plotting. Given the potential size of the network, it is highly recommended to use this function or external software like Gephi (https://gephi.org/). The export_to_gephi() function facilitates exporting these networks to Gephi for further analysis.

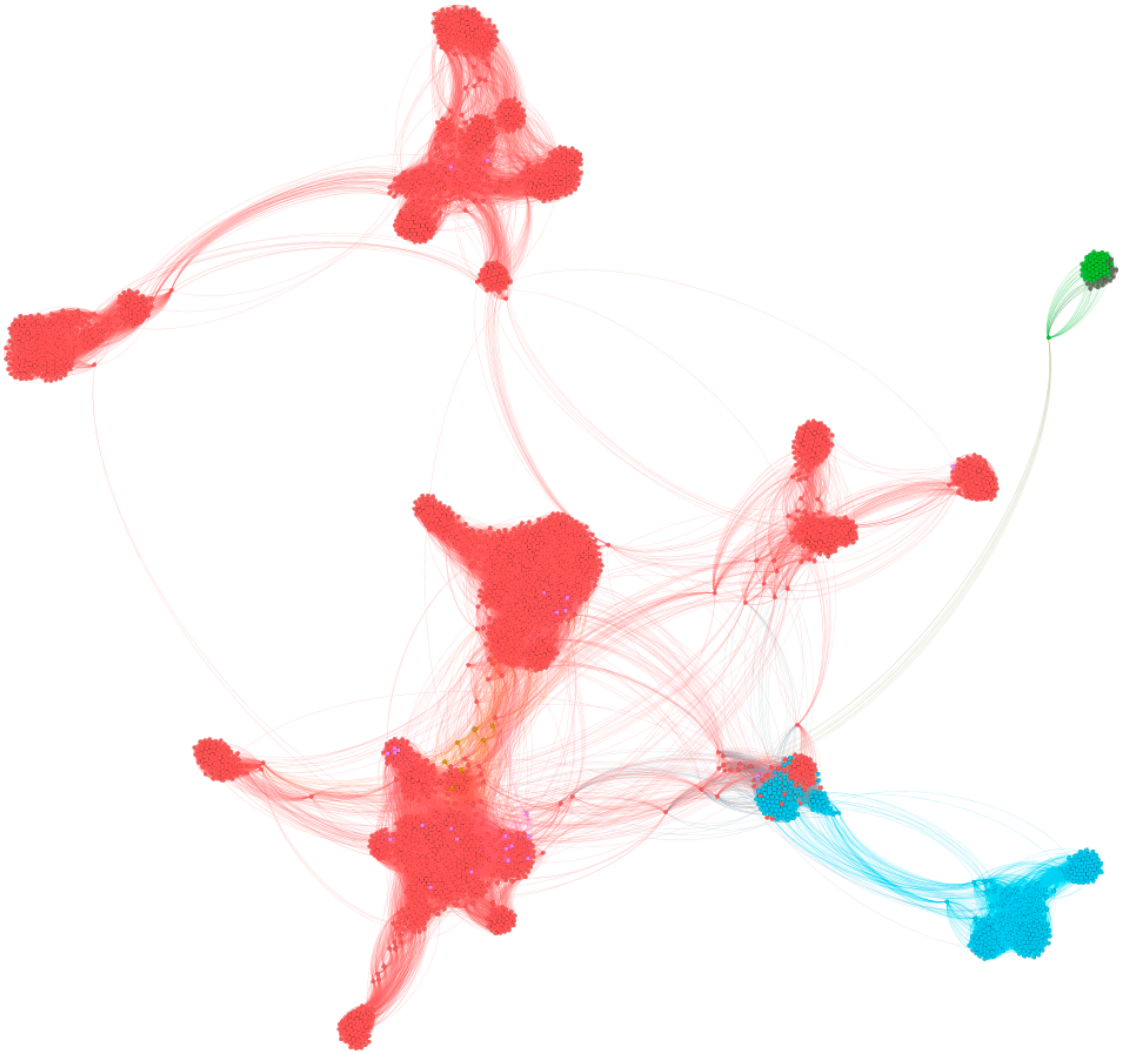


Figure 17. Gephi visualization of the K-NNN with 50 neighbors.

PATO incorporates diverse clustering methods tailored to different types of data objects, aiming to streamline the clustering process through the versatile clustering() function. Specifically designed for mash and accnet objects, the following clustering techniques are available:

1. mclust: Utilizing Gaussian Finite Mixture Models, this method, compatible with dimensionality reduction, is implemented via the Mclust package to determine the optimal number of clusters.

2. upgma: This method performs Hierarchical Clustering using the UPGMA algorithm, necessitating user-provided cluster numbers for effective execution.

3. Ward.D2: To Employ Hierarchical Clustering through the Ward algorithm, users must specify the desired number of clusters.

4. hdbscan: Executing Density-based Spatial Clustering of Applications with Noise using the DBSCAN package, this method determines the optimal cluster number.

All of the aforementioned methods support multidimensional scaling (MDS). PATO employs the UMAP algorithm for MDS, a technique akin to other machine learning MDS approaches such as t-SNE or Isomap. Distinctly categorized into two types, preserving distance structure within the data and favoring local distance preservation over global distance, UMAP, t-SNE, and Isomap fall into the latter category. This type of MDS synergizes effectively with clustering algorithms, which inherently strive to identify local structures.

Additionally, the clustering() function accommodates knnn networks as input, integrating network clustering algorithms, including:

- Greedy: Identifies community structure through greedy optimization of modularity.

- Louvain: Implements a multi-level modularity optimization algorithm for discovering community structure.

- walk trap: Establishes community structure via short random walks.

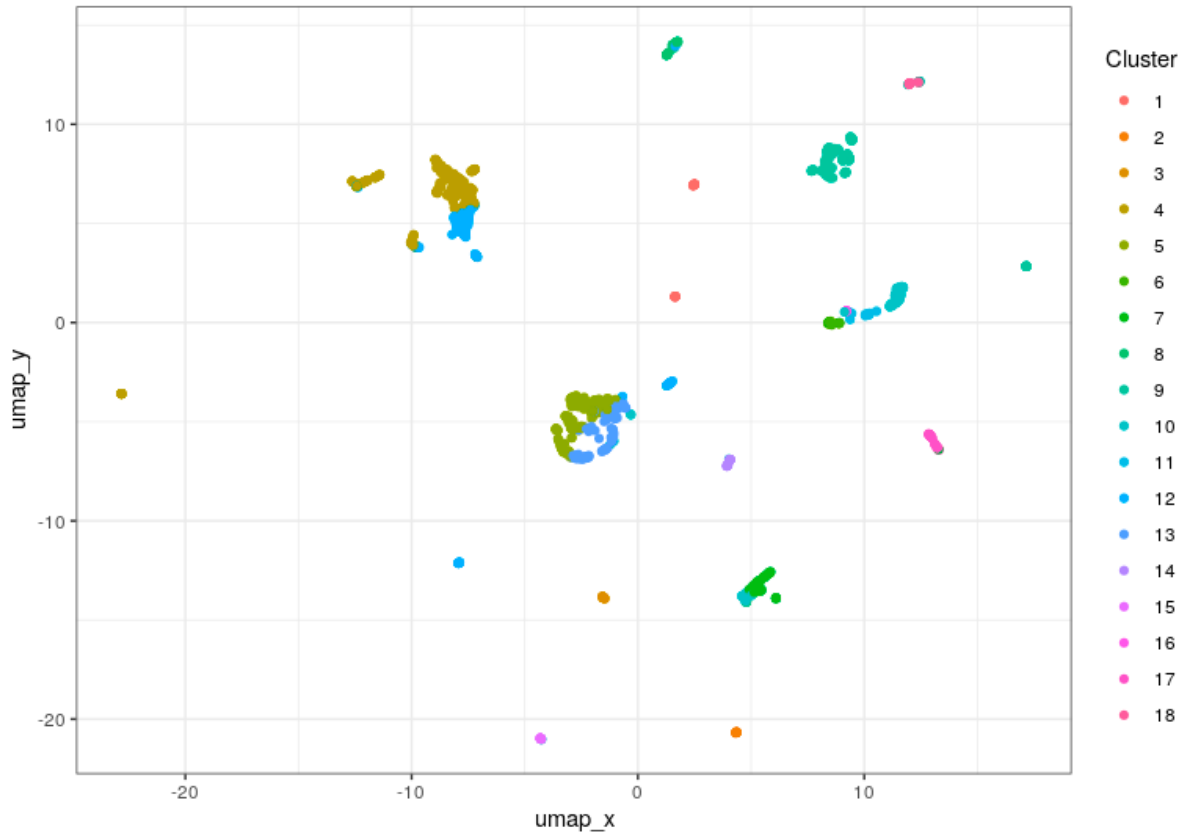


Figure 18. UMAP visualization with cluster definition

We can also use the plot_knnn_network() to visualize the network clustering results.

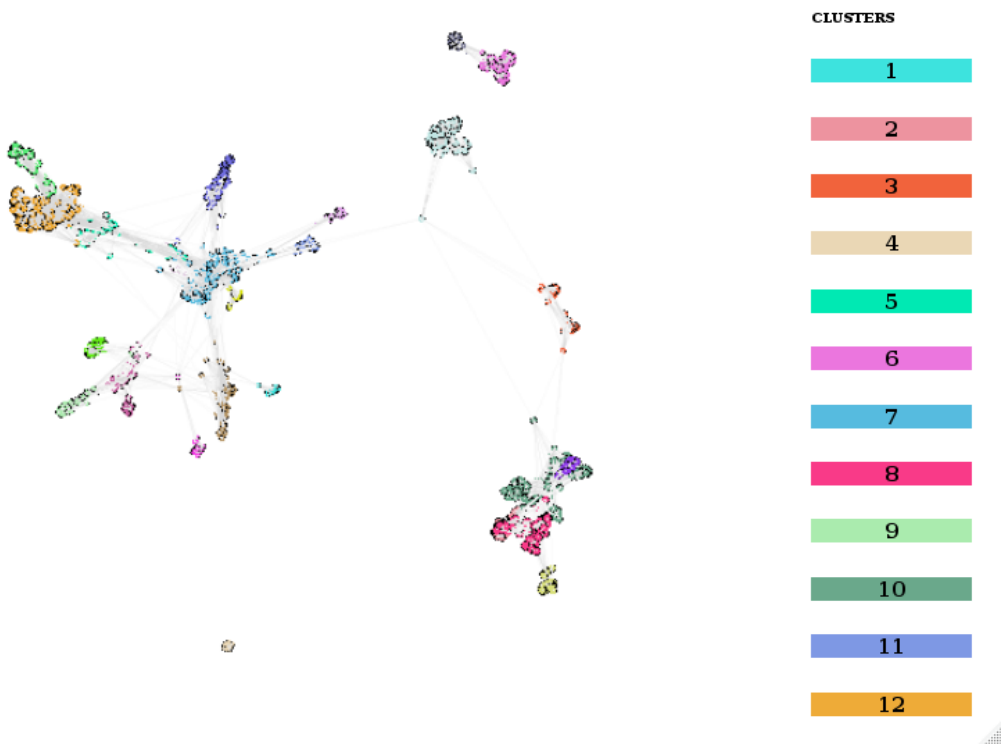


Figure 19. K-NNN with clustering results, each color represents a cluster from the network.

Annotation

PATO encompasses a sophisticated gene annotation function, namely annotate(), designed to annotate Antibiotic Resistance Genes (ARGs) and Virulence Factors (VFs). Antibiotic resistance predictions leverage MMSeqs2 over the ResFinder database (https://cge.cbs.dtu.dk/services/ResFinder/) (doi: 10.1093/jac/dks261). For Virulence Factors, VFDB (http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi) (doi: 10.1093/nar/gky1080) is employed, presenting two gene sets: VF_A (core dataset) and VF_B (full dataset). VF_A exclusively includes experimentally verified VFs, while VF_B encompasses both VF_A genes and those predicted as VFs.

The outcome of the annotate() function is a table featuring all positive hits for each gene/protein in the dataset files. The annotation process is versatile, allowing reuse of mmseqs() results for enhanced efficiency. Users can target all genes/proteins or specifically focus on accessory genes/proteins during the query. While the results are comprehensive, they are presented in a raw format, necessitating user-driven curation. We recommend leveraging tidyverse tools for efficient data refinement.

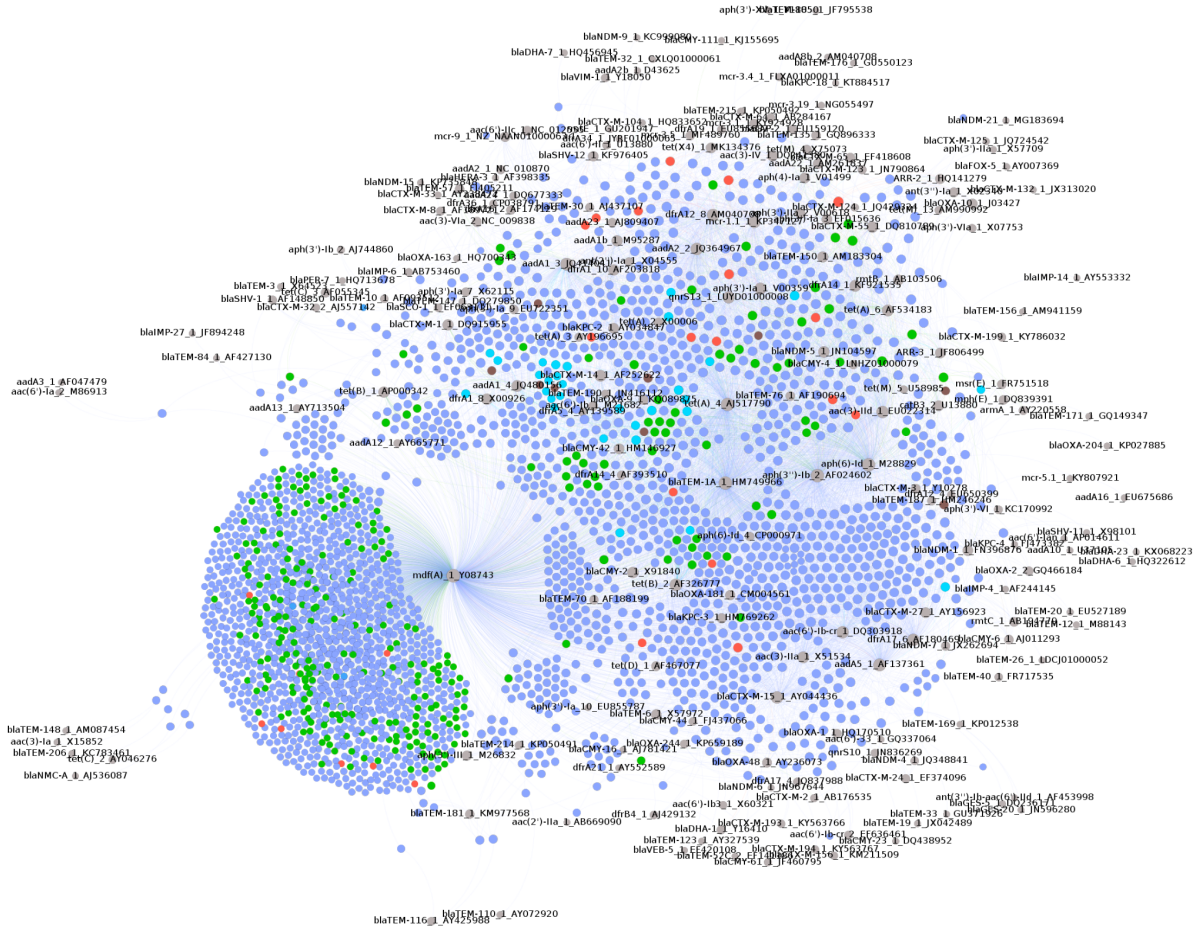


Figure 20. Gephi visualization of the network of genes encoding antibiotic resistance (ARGs)

Outbreak and Transmission Analysis of bacterial genes and strains *In microbial genomics, a primary objective is to discern outbreaks or transmission events across diverse sources. Essentially, the aim is to identify identical (or highly similar) strains among different subjects or sources, often involving strains differing by only a few Single Nucleotide Polymorphisms (SNPs).*

For this illustration, we leverage a dataset from [https://msphere.asm.org/content/5/1/e00704-19]. (downloadable at https://doi.org/10.6084/m9.figshare.13482435.v1).

In their study, the authors conducted whole-genome comparative analyses on 60 *E. coli* isolates sourced from soils and various fecal sources (cattle, chickens, and humans) in rural Bangladeshi households (Montealegre 2020). PATO provides diverse methodologies to identify Outbreak/Transmission events. A conventional approach involves calculating the core-genome, constructing a phylogenetic tree, and determining the average number of SNP

differences (or alternatively, Average Nucleotide Identity - ANI) between strains. Four alternatives are available:

1. MASH Similarity

2. Core-genome + SNP Matrix (Roary-like)

3. Core_snp_genome + SNP Matrix (Snippy-like)

4. SNPs Pairwise (most resource-intensive but highly accurate)

The samples are designed with the name of the household number and the source (H: Human, C: Cattle, CH: Chicken, S: Soil). For example, the strain *E. coli* HH15CH is the householder 15 and the source is chicken feces.

MASH Similarity Analysis

A rapid and straightforward method for identifying high similarities among genomes involves employing MASH distance calculations. It is important to note that MASH utilizes the entire genome to estimate genomic distance, thereby encompassing accessory elements such as plasmids, transposons, and phages in the distance calculation.

Typically, the number of Single Nucleotide Polymorphisms (SNPs) serves as a metric for clonal relatedness. In the subsequent analysis, we will compare the MASH approach with SNP-based methodologies.

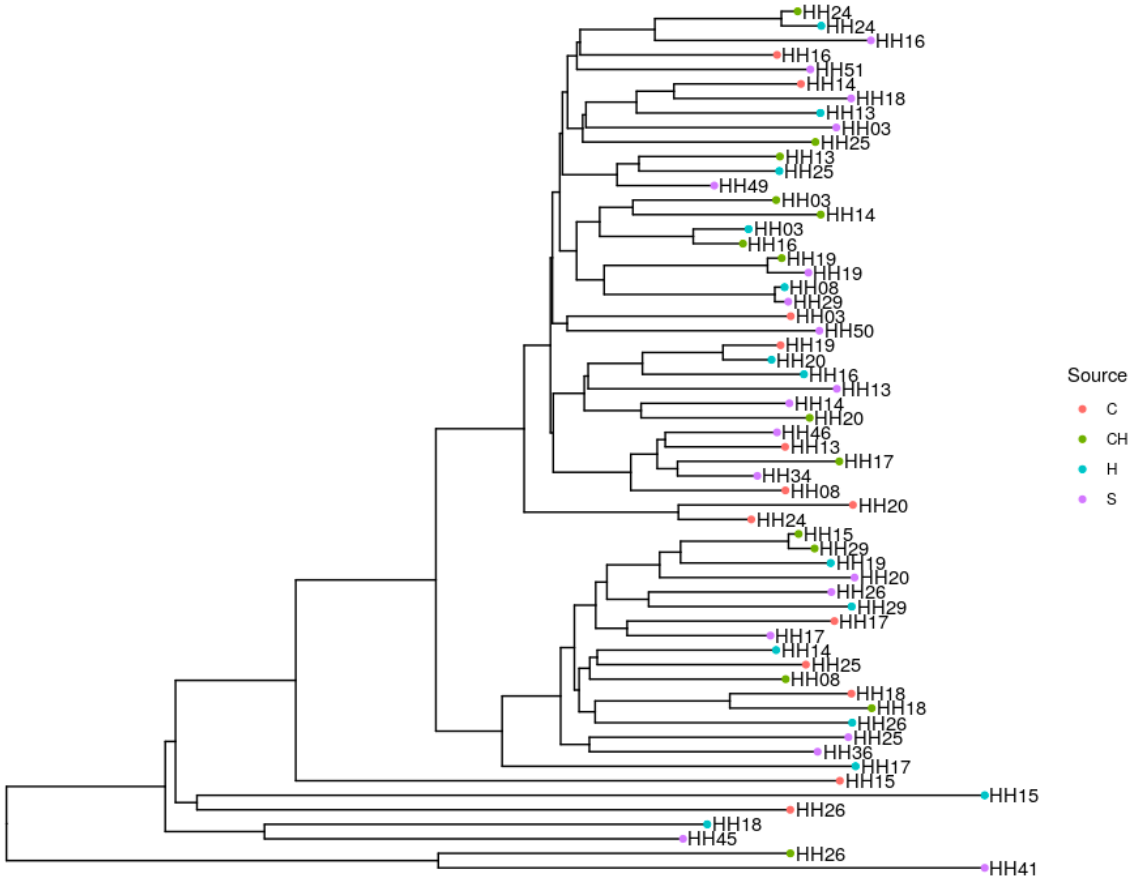


Figure 21. Mash similarity tree

Core-Genome and SNP Matrix Analysis (Roary-like)

In this instance, we will employ the PATO pipeline to calculate the core-genome and determine the number of SNPs among the sampled genomes.

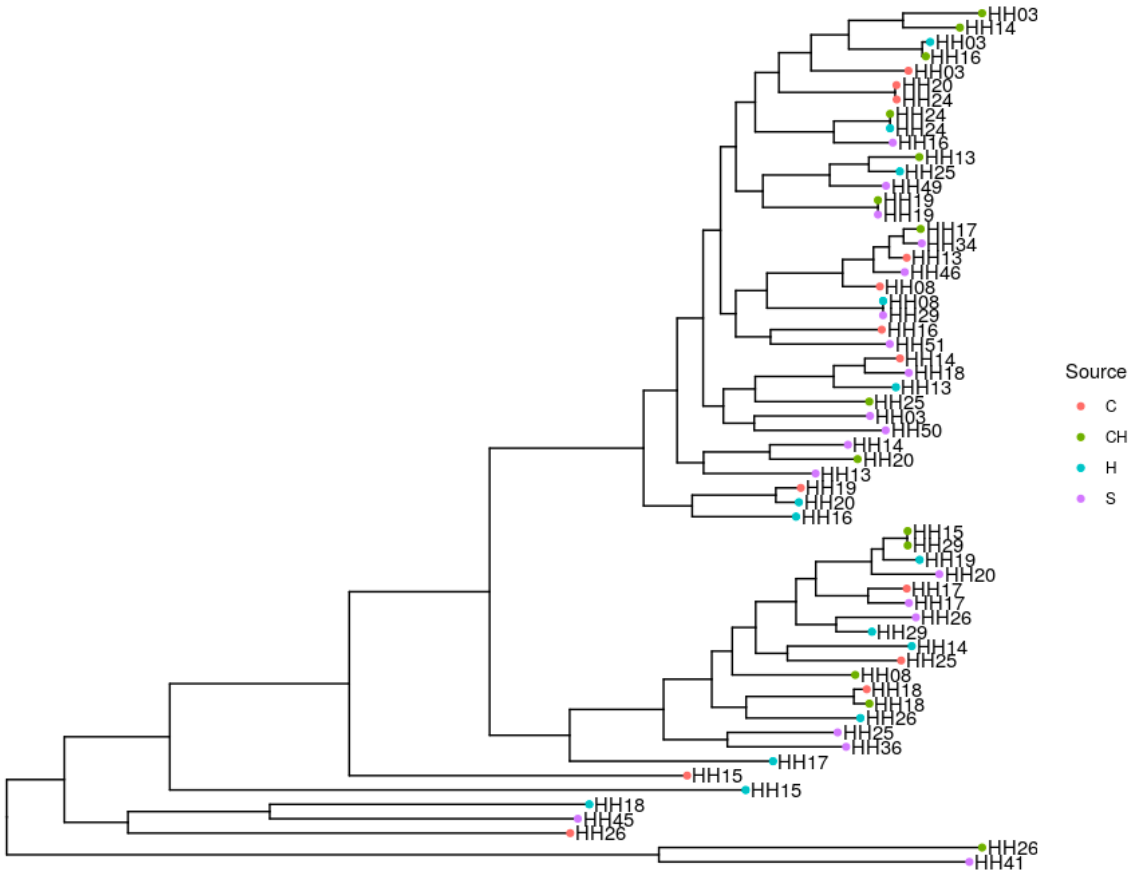


Figure 22. Core-genome + snp_matrix (roary-like) tree, each color dot corresponds to a source.

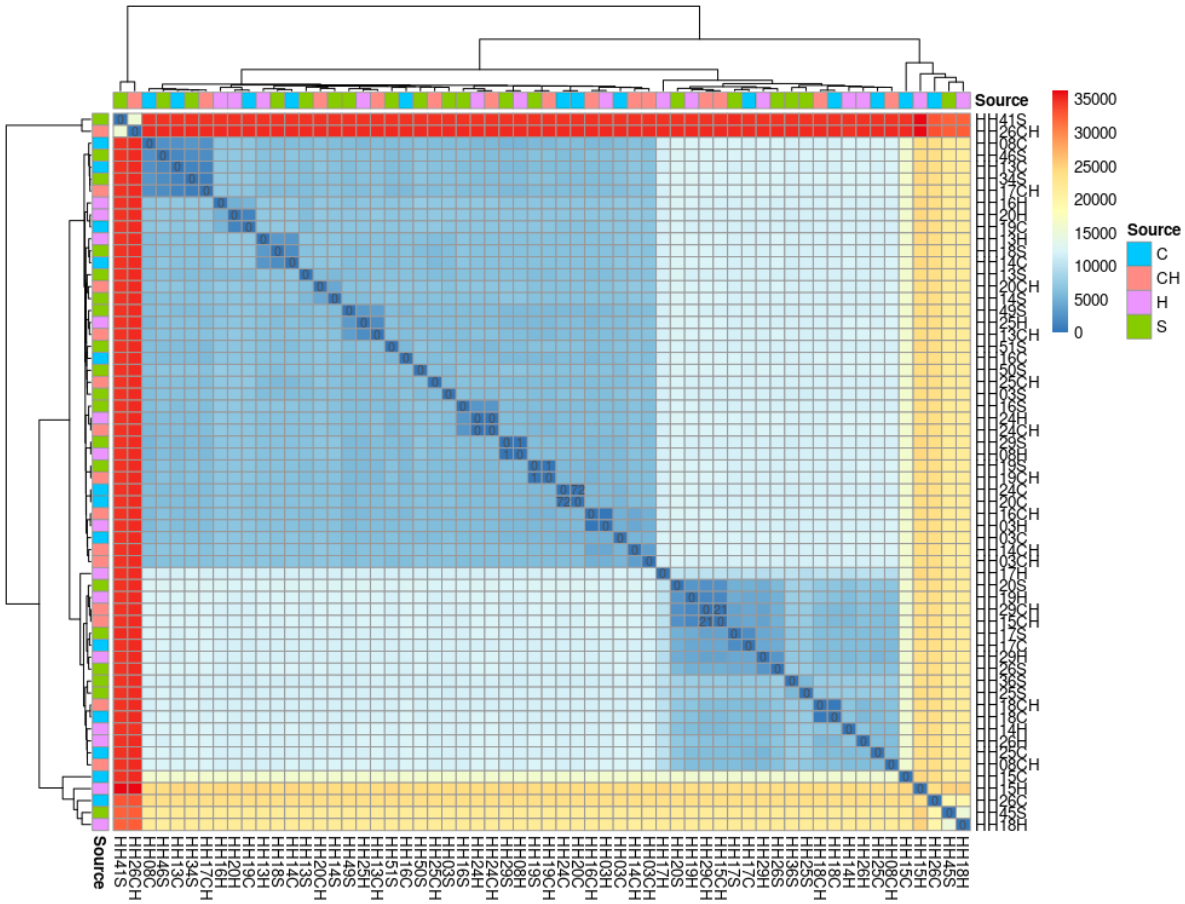


Figure 23. Core-genome + snp_matrix (roary-like) heatmap. The source of each genome is represented in both axis.

Our investigation reveals three potential transmission events: HH24 from Human to Chicken (or vice versa), HH29 Soil to HH08 Human (or vice versa), and Human HH19 to Chicken (or vice versa), each exhibiting less than 4 SNPs per Megabase (SNPs/Mb). PATO performs this computation by normalizing the number of SNPs against the alignment length in Megabases, specifically within the core genome.

Core SNP Genome Analysis with SNP Matrix (Snippy-like)

An alternative approach for identifying potential transmission events involves utilizing the core_snp_genome() function. In this scenario, all genomes are aligned to a reference genome. The function identifies common regions across all samples and extracts Single Nucleotide Polymorphisms (SNPs) within those regions. This methodology closely resembles the Snippy pipeline (https://github.com/tseemann/snippy).

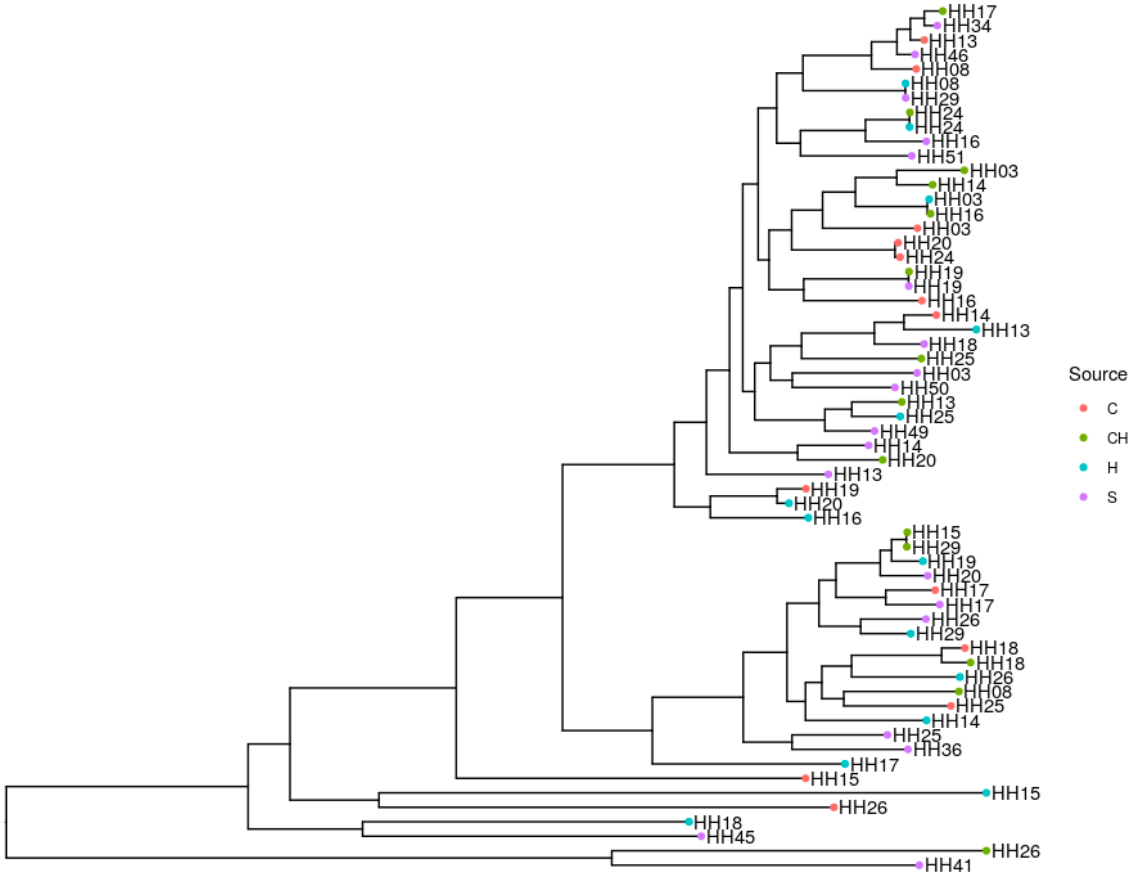


Figure 24. Core SNP Genome Analysis with SNP Matrix (Snippy-like) tree.

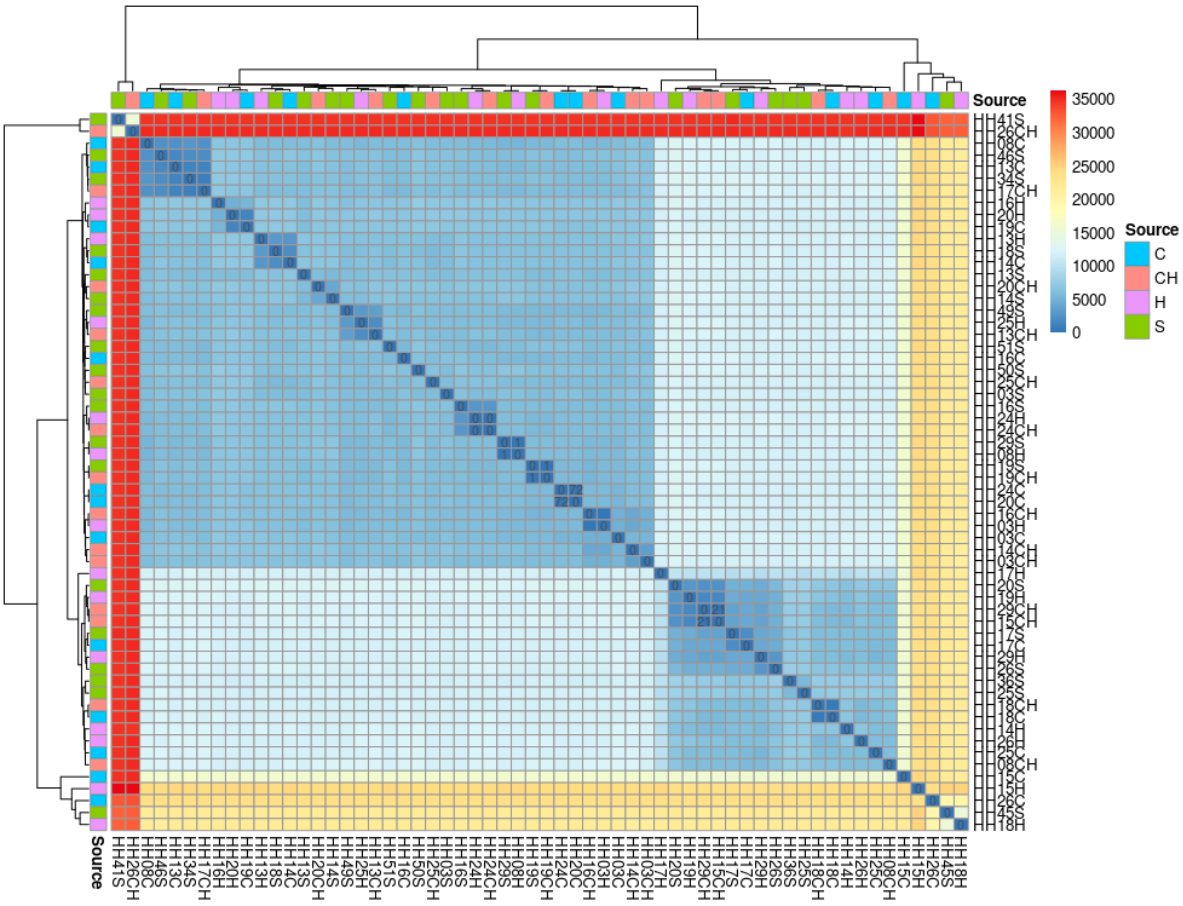


Figure 25. Core SNP Genome Analysis with SNP Matrix (Snippy-like) heatmap.

In adopting this strategy, it is crucial to acknowledge that the total count of Single Nucleotide Polymorphisms (SNPs) may escalate. However, it is noteworthy that the pairs remain consistent, exemplified by HH24H<->HH24CH, HH29S<->HH08H, and HH19S<->HH19CH.

SNPs-Pairwise Analysis

The last two methodologies share a common limitation: their reliance on the dataset structure. In both instances, the count of Single Nucleotide Polymorphisms (SNPs) is determined within the core-genome, a metric contingent upon the dataset's composition. If the dataset harbors an outlier or exhibits heterogeneous diversity, the core-genome might contract, leading to an underestimation of the SNPs count, even when normalized by the alignment length. To address this challenge, PATO introduces the snps_pairwise() function, facilitating a pairwise computation of SNP numbers. However, this approach incurs computational intensity, as it involves aligning all sequences against each other. This stands in contrast to the core_snp_genome() function, which performs alignments in O(n)

complexity. The latter represents the actual number of SNPs per-megabase, excluding insertions and deletions.

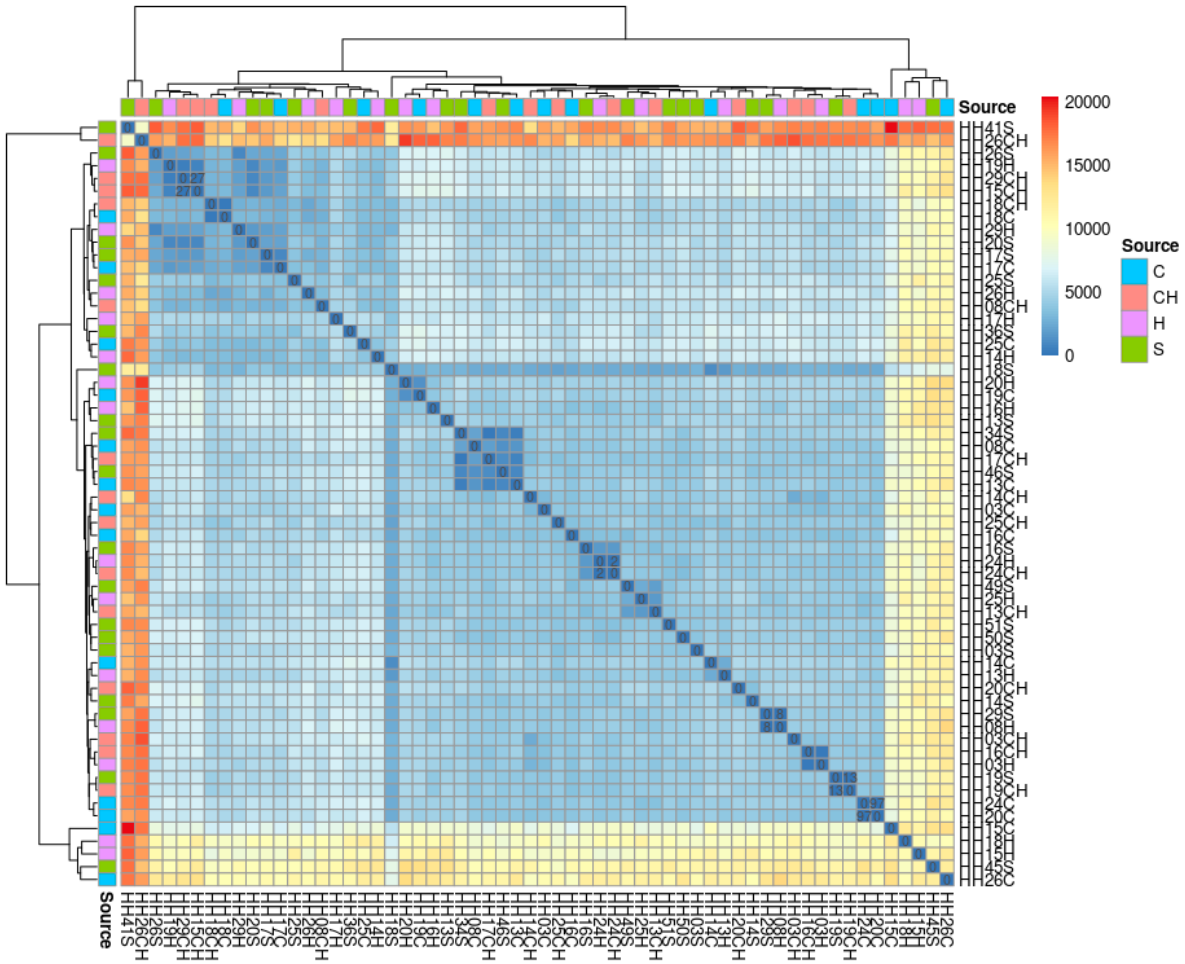


Figure 26. HeatMap of the pairwaise SNP Matrix

Upon analysis, the results indicate that the transmission event between HH24H and HH24CH is recent, whereas the transmission between HH29<->HH08 and HH19 appears more distant.

It is noteworthy that these findings diverge from the MASH results, likely attributed to the exclusion of "dispensable" or "accessory" elements (e.g., plasmids, phages, ICEs) in approaches solely based on SNPs differences.

Pangenome analysis

PATO is adeptly designed to unravel the intricate relationships within pangenomes across diverse species. This specialized functionality aims to dissect clusters of genomes, treating each pangenome as an individual entity. In this instance, we meticulously scrutinize a colossal dataset comprising 49,591 genomes from the bacterial phylum Firmicutes, meticulously curated from the NCBI database.

For datasets of this magnitude, it is judicious to employ a dedicated pangenome analysis pipeline. This specialized pipeline orchestrates the clustering of genomes into distinct pangenomes, delineating entities such as species clusters, intra-species phylogroups, or sequence types (ST). The diversity within each cluster is contingent upon the specified distance parameter, ensuring the creation of homogeneous clusters with similar phylogenetic distances. Subsequently, the pipeline generates an accnet object, enabling the utilization of a common accnet object for the array of functions available within PATO. Additionally, the pipeline incorporates configurable parameters for determining the minimum number of genomes required to constitute a pangenome and the minimum frequency threshold for the inclusion of a protein/gene family within a pangenome.

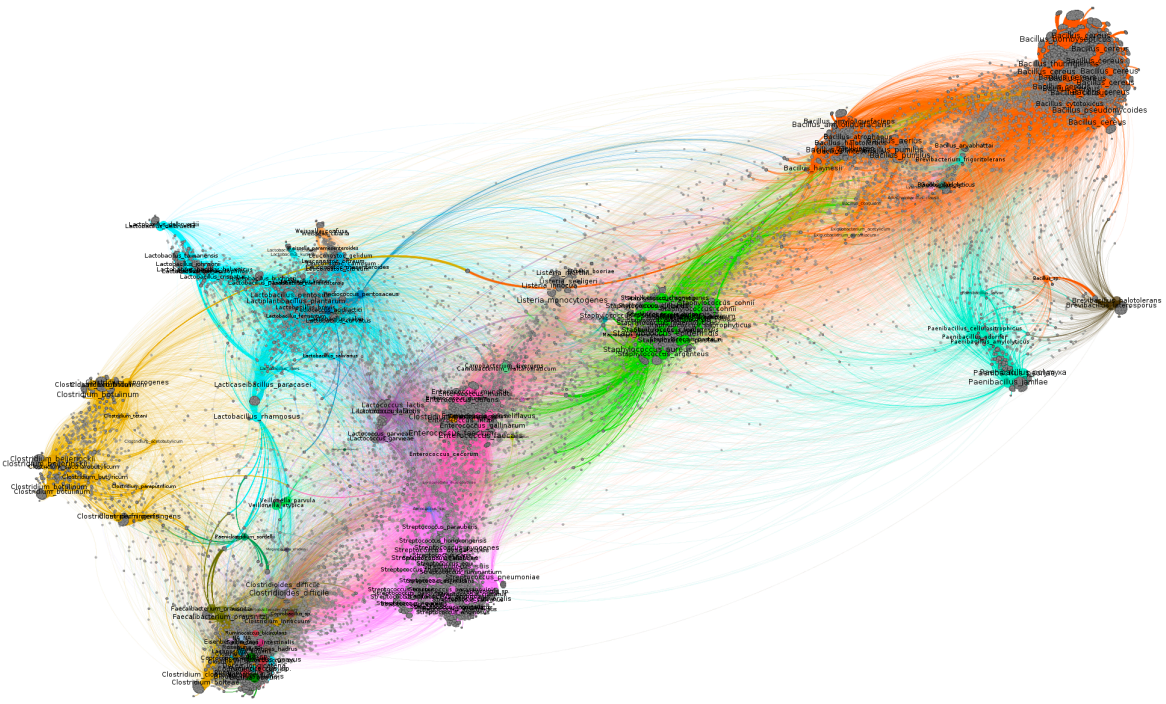


Figure 27. Pangenomes network of firmicutes: Gephi visualization

An alternative avenue for comprehending and visualizing the outcomes lies in the intricate landscape of sharedness among pangenomes. PATO introduces the sharedness() function, a pivotal tool for rendering this nuanced interplay visible. Facilitating the interpretation of results derived from the sharedness() function, the renowned pheatmap package becomes indispensable. However, there is a prerequisite: the input value for pheatmap must be a matrix. Consequently, a preliminary transformation of our result into a matrix format becomes imperative.

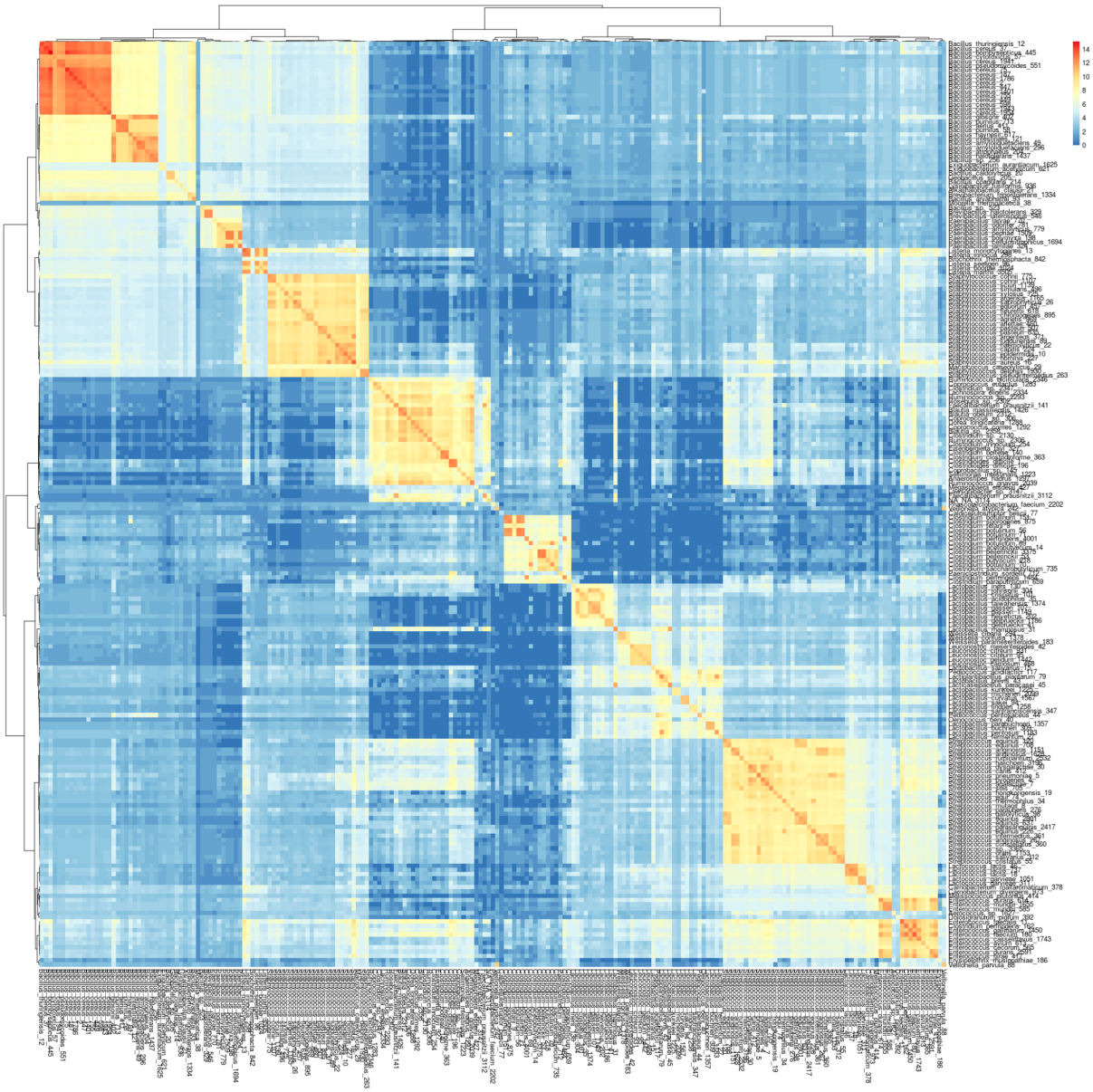


Figure 28. Sharedness heatmap

4.2.1.8. Benchmarking

Performance

To assess and benchmark PATO against prominent pangenome analysis software, we conducted a rigorous evaluation using distinct sets of *E. coli* genomes, varying in quantity (20, 50, 100, 200, 400 genomes). The comparative analysis involved PanACoTA v1.3.1, Panaroo v1.2.7, Peppan v1.0.5, PanX v1.5.1, PIRATE v1.0.4, and Roary v3.12.0, with PATO included in the lineup. All software, including PATO, was executed with default parameters. The evaluation was performed on a robust workstation featuring 24 cores (Dual Xenon) and 48 GB of RAM.

For PATO_roary, a set of functions emulating Roary's internal pipeline was meticulously crafted. This encompassed GFF file loading, orthologous clustering, accessory genome delineation, core-genome definition, core-genome representation, and accessory genome clustering. Similarly, Pato_panacota replicated the PanACoTA pipeline for pangenome creation, involving GFF file loading, orthologous clustering, and accessory genome definition.

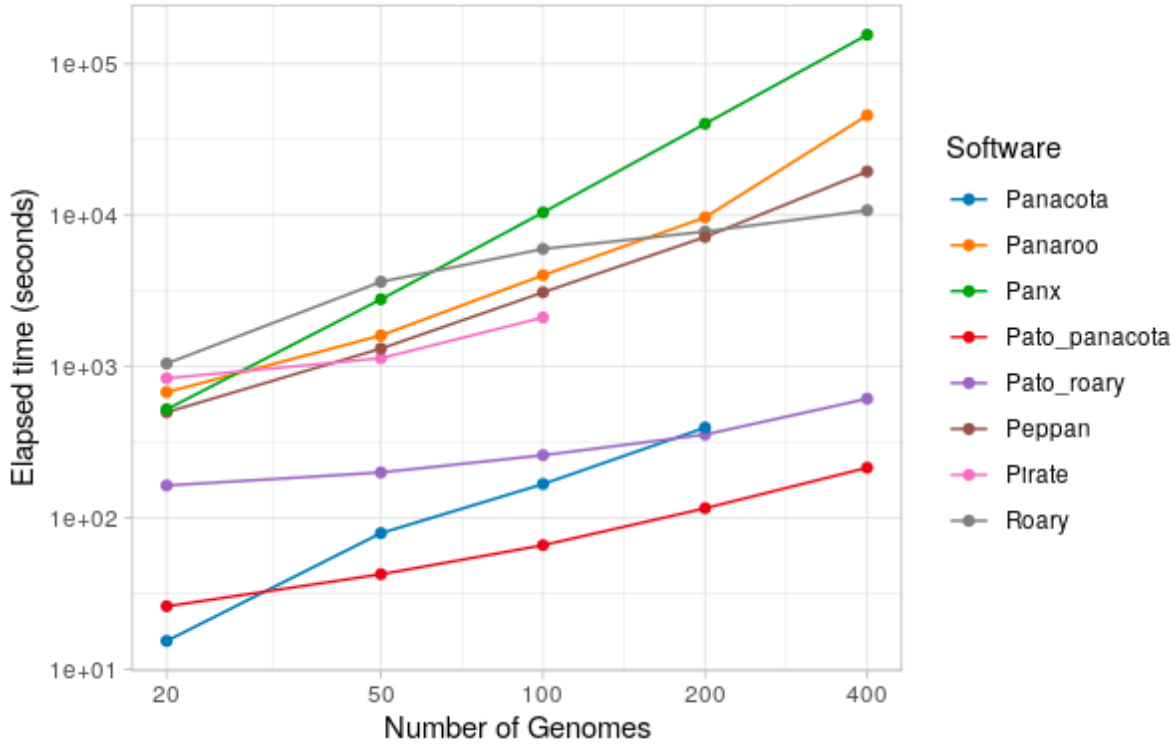


Figure 29. PATO Benchmarking among pangenome software

PATO stands out as a remarkably swift and efficient software solution in comparison to other tools. Notably, both PanX and PanACoTA encountered significant challenges, reaching the upper limit of available RAM during execution—PanX with 200 genomes and PanACoTA with 400 genomes—resulting in the need to halt their processes.

A distinctive feature of PATO and PanACoTA is the absence of paralogue splitting processing, in contrast to other software in the evaluation. This characteristic contributes significantly to the accelerated performance of PATO and PanACoTA, distinguishing them as faster alternatives. PATO, specifically designed to handle large-scale analyses involving hundreds or even thousands of genomes, intentionally forgoes paralogue splitting processes. This strategic decision was driven by the recognition that such processes are computationally

intensive and often yield results that may not justify the substantial investment of computational time. The efficiency gains achieved by foregoing these time-consuming procedures position PATO as an expedient solution for high-throughput genomic analyses.

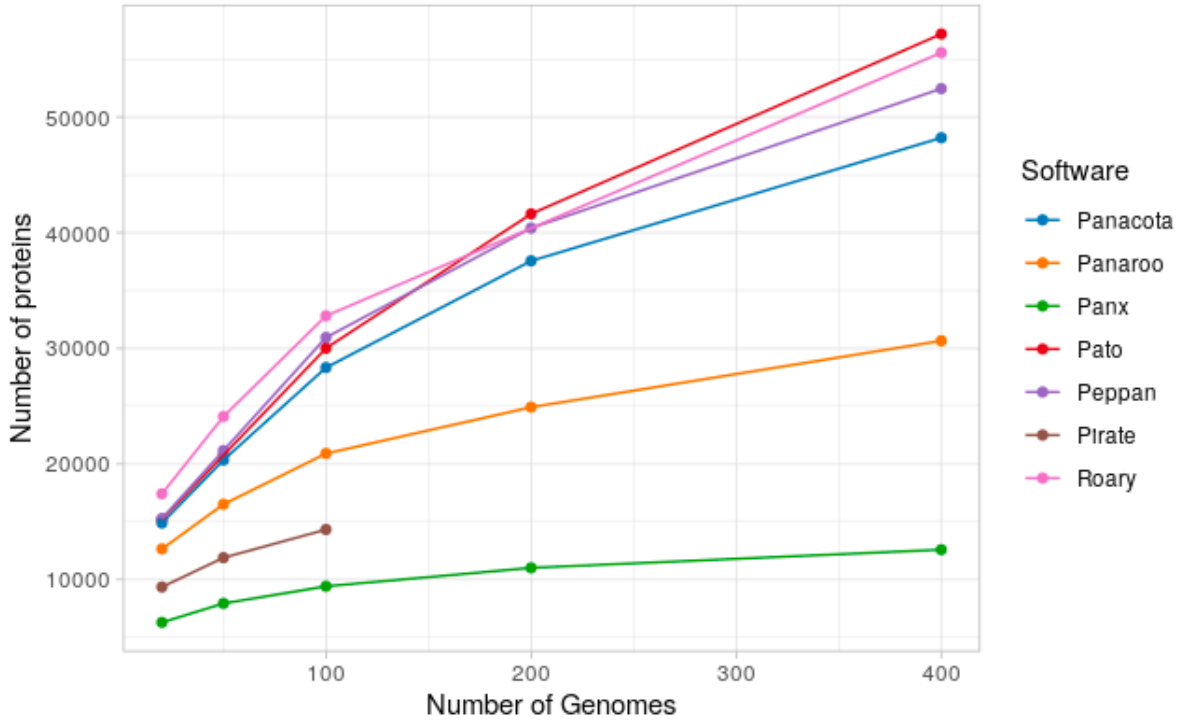


Figure 30. PATO Benchmarking in Pangenome size

PATO consistently yields results comparable to PanACoTA, Roary, and PEPPAN in terms of pangenome size while showcasing a clear advantage in performance. In contrast, more intricate pangenome analysis tools such as Panaroo or PanX tend to produce noticeably smaller pangenomes. Although PATO's outcomes align with state-of-the-art methods, users have the flexibility to import results from more exhaustive software, like Panaroo, and leverage an array of tools designed for in-depth analysis and visualization.

While the paralogue processing employed by other tools aims for a highly accurate core-genome, PATO adopts a more conservative approach. It excludes any gene from the core-genome that possesses a paralog in any genome. While this may lead to slightly smaller core-genomes, it ensures the authenticity of genes within the core-genome and minimizes the impact of horizontal transfers.

A significant source of bias in pangenome analyses is the presence of single genes—those unique to a single genome and absent in others. Such discrepancies often arise from challenges in both assembly and ORF prediction. To address this, PATO's accessory genome

analysis tools provide users with the option to filter out genes with frequencies of appearance below a user-defined threshold, thereby mitigating the impact of erroneous single-gene identifications.

Phylogenetics

PATO offers two distinct pipelines for constructing phylogenetic trees. The first follows the classical method of gene-to-gene alignment, identifying core-genome genes, aligning groups of orthologs, and concatenating the alignments. This approach mirrors the methodology of most pangenome analysis programs such as Roary, PIRATE, and PanX. The second pipeline involves aligning all genomes against a reference genome, extracting mutations (SNPs) in regions common to all genomes, known as the core SNP genome. This methodology aligns with Snippy's approach.

We conducted a performance comparison between PATO and Snippy using the montealegre et al. dataset. Microbenchmarking was employed, aiming to reproduce Snippy's pipeline steps within PATO using distinct functions.

The benchmarking results, measured in seconds, revealed a substantial performance advantage for PATO's equivalent pipeline:

Table 4. Time benchmarking between PATO and Snippy. Unit: seconds

expr	min	lq	mean	uq	max	neval
pato_like_snippy_time	149.9739	149.9739	149.9739	149.9739	149.9739	1
snippy_time	3425.9932	3425.9932	3425.9932	3425.9932	3425.9932	1

PATO demonstrated nearly 23 times faster execution than Snippy for a dataset of 60 genomes on a Dual Intel Xeon workstation with 24 effective cores (12 real cores) and 48 GB RAM.

Comparing the core alignments, Snippy generated a 4,680,136 bp alignment, while PATO produced a slightly smaller 4,392,982 bp alignment. However, the number of identified SNPs differed, with Snippy detecting 260,909 and PATO identifying 155,300. The variation may stem from the core_snps_genome() function, which currently does not account for multi-variant positions (changes of two or more nucleotides in a row) in PATO. This issue is acknowledged and will be addressed in future PATO versions.

It's important to note that PATO and Snippy, while sharing a similar approach, employ distinct pipelines. Snippy simulates NGS data, aligns it against a reference using BWA, and performs variant calling with Freebayes. In contrast, PATO primarily utilizes Minimap2 tools for both genome alignment and variant calling. Minimap2 is explicitly designed for genome alignment and variant calling on its results, contributing to the observed differences in outcomes between Snippy and PATO.

A subsequent comparison involved phylogenetic trees generated by the same software (Fasttree) using four different alignments from various pipelines (PATO core_genome, PATO core_snp_genome, Roary, and Snippy).

In our comparative analysis of phylogenetic trees, we employed the TreeDist package, which implements the Robinson-Foulds method to measure the similarities between trees.

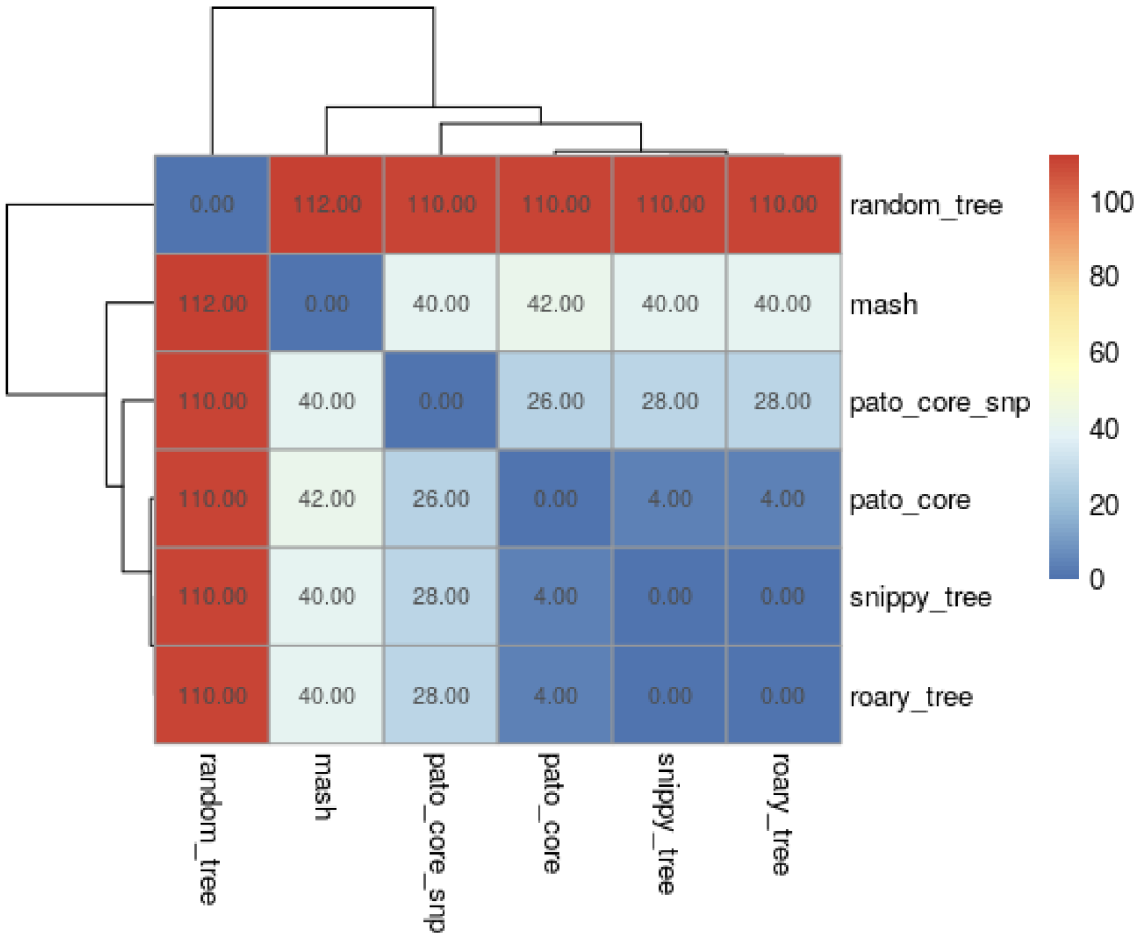


Figure 31. HeatMap of the Robinson Foulds distances

The phylogenetic trees generated by Roary and Snippy exhibited a high degree of similarity. Meanwhile, the tree produced by PATO using the roary-like approach closely resembled the Roary and Snippy trees, demonstrating a high level of consistency. On the other hand, the tree derived from PATO's snippy-like approach, while accurate, did not precisely mirror the Roary and Snippy trees. Despite these differences, the alignments produced by all methods, including MASH, exhibited considerable similarity.

To provide a baseline for unrelated tree structures, we incorporated a randomly generated tree into our analysis. The inclusion of this unrelated tree allowed us to gauge the distinctiveness of the trees obtained through the different methodologies.

Population structure

PATO integrates a suite of tools designed to unravel the population structure within a set of genomes. These tools primarily constitute the implementation of various clustering algorithms applied to the genomic data processed by PATO.

To validate our outcomes, we subjected the identical dataset to analysis using alternative software, including PopPPunk, Snapclust, and FastBaps. Additionally, we incorporated the classification based on phylogroups established by the Clermont scheme.

For seamless integration of the population structure database of PopPunk, we utilized version v1.2.2 (latest version 2.4.0). PATO employed diverse methods, namely MClust, MClust with Dimensional Reduction, HDBscan, Ward.D2, and Louvain (utilizing k-nnn).

To facilitate result comparison, we depicted the population structure based on MASH distance using the umap_plot() function and color-coded the strains based on their respective clusters. For result uniformity, we organized the cluster numbers based on cluster size, where the cluster labeled as number one represents the largest cluster. This approach ensures a coherent representation of population structures for comprehensive evaluation.

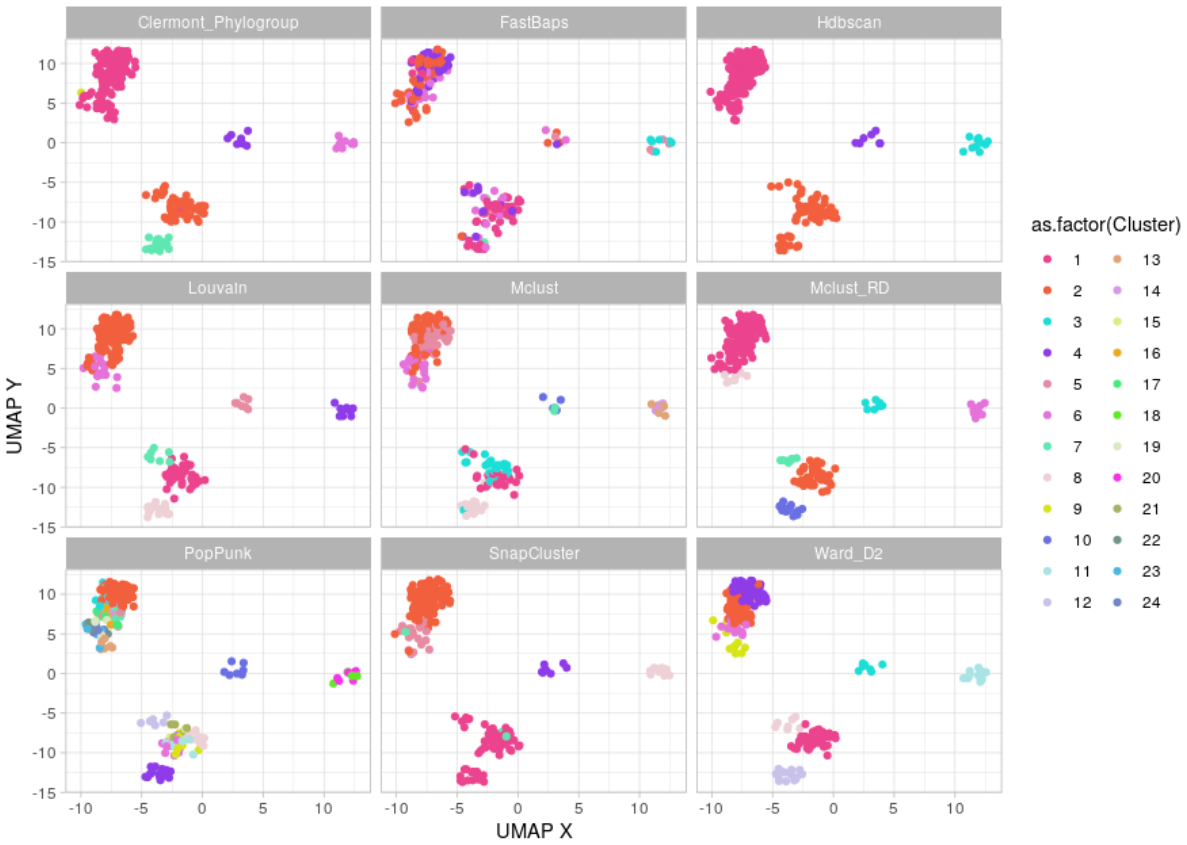


Figure 32. Clustering comparison. UMAP plot of mash data

The clustering outcomes can be systematically assessed using the adjustedRandIndex() function incorporated in the MClust module. This approach enables the quantification of dissimilarities among various clustering and population structure methods. The findings

indicate a notable concordance between the results obtained through PATO's HDbscan, Louvain, and Mclust + DR methods, along with SnapClust. Strikingly, these outcomes closely align with the phylogenetic groups proposed by the Clermont scheme, showcasing the reliability and accuracy of these methods in capturing the underlying population structure within the genomic dataset.

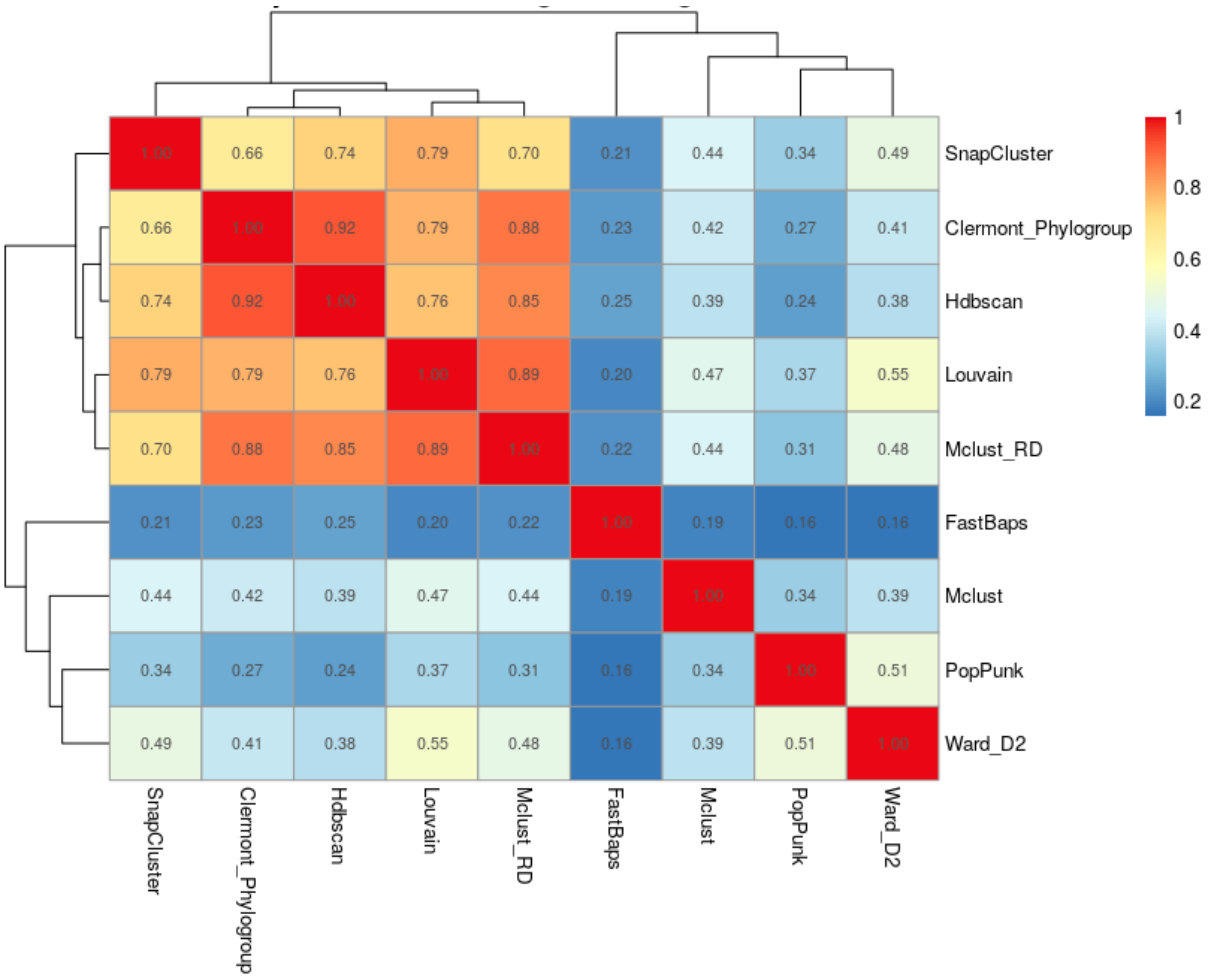


Figure 33. AdjustRandIndex in clustering methods

Gene Association Test

Exploring gene associations within the PATO framework involves distinguishing between strict associations and associative associations. Strict associations refer to genes that appear in precisely the same gene sets, such as the identification of twins. Associative associations involve genes clustering based on their presence or absence across genomes. This

methodology bears similarity to Coinfinder (doi: https://doi.org/10.1099/mgen.0.000338), albeit with distinctive features.

Coinfinder operates by constructing an association network through statistical association tests between genes. Statistical corrections for multiple tests are applied, and the associations are further adjusted based on the phylogenetic relationships among genomes. However, the subsequent analysis involves examining the network components, specifically the complete sub-networks.

In contrast, PATO utilizes protein clustering reliant on gene distance determined by their presence-absence patterns in genomes. A 2D projection is computed using UMAP, followed by gene/protein clustering employing the HDbscan algorithm. Unlike Coinfinder, PATO does not correct gene associations based on genome phylogeny, potentially leading to associations influenced by genome similarity. This influence can be mitigated by opting for non-redundant datasets. Users are advised to consider these methodological nuances and potential limitations.

For a comprehensive comparison, we conducted an analysis of 900 non-redundant genomes, enhancing Coinfinder results through multiple clustering iterations of the association network.

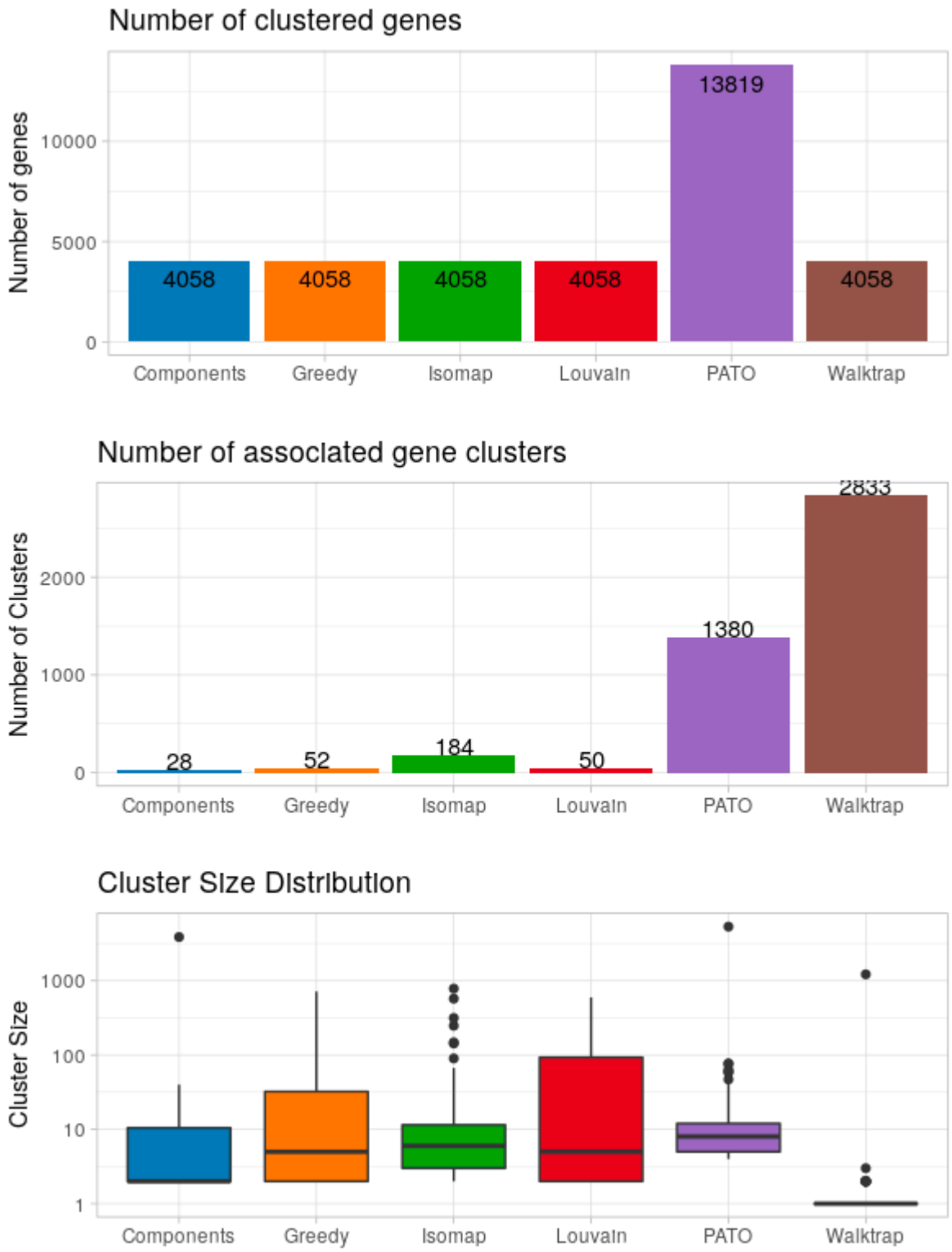


Figure 34. Cluster statistics comparison

PATO identifies 13,819 genes, whereas Coinfinder reveals positive associations in 4,058 genes. Initially, Coinfinder identifies 28 distinct groups, while PATO uncovers 1,380 groups, each containing at least 4 members and represented in at least 5 genomes. Despite

the increased number of groups, the average size remains comparable. Employing network clustering methods refines Coinfinder's original results and enhances the group count. Advanced methods such as Louvain or Fast Greedy can even double the number of clusters. It's crucial to note that PATO's outcomes are not adjusted for the phylogeny of genomes, introducing potential biases due to the over-representation of closely related genome groups.

To quantify the dissimilarities between PATO and Coinfinder, we measured the similarity using the Adjusted Rand Index method. However, given the diverse nature of the results, this similarity index may not provide accurate assessments.

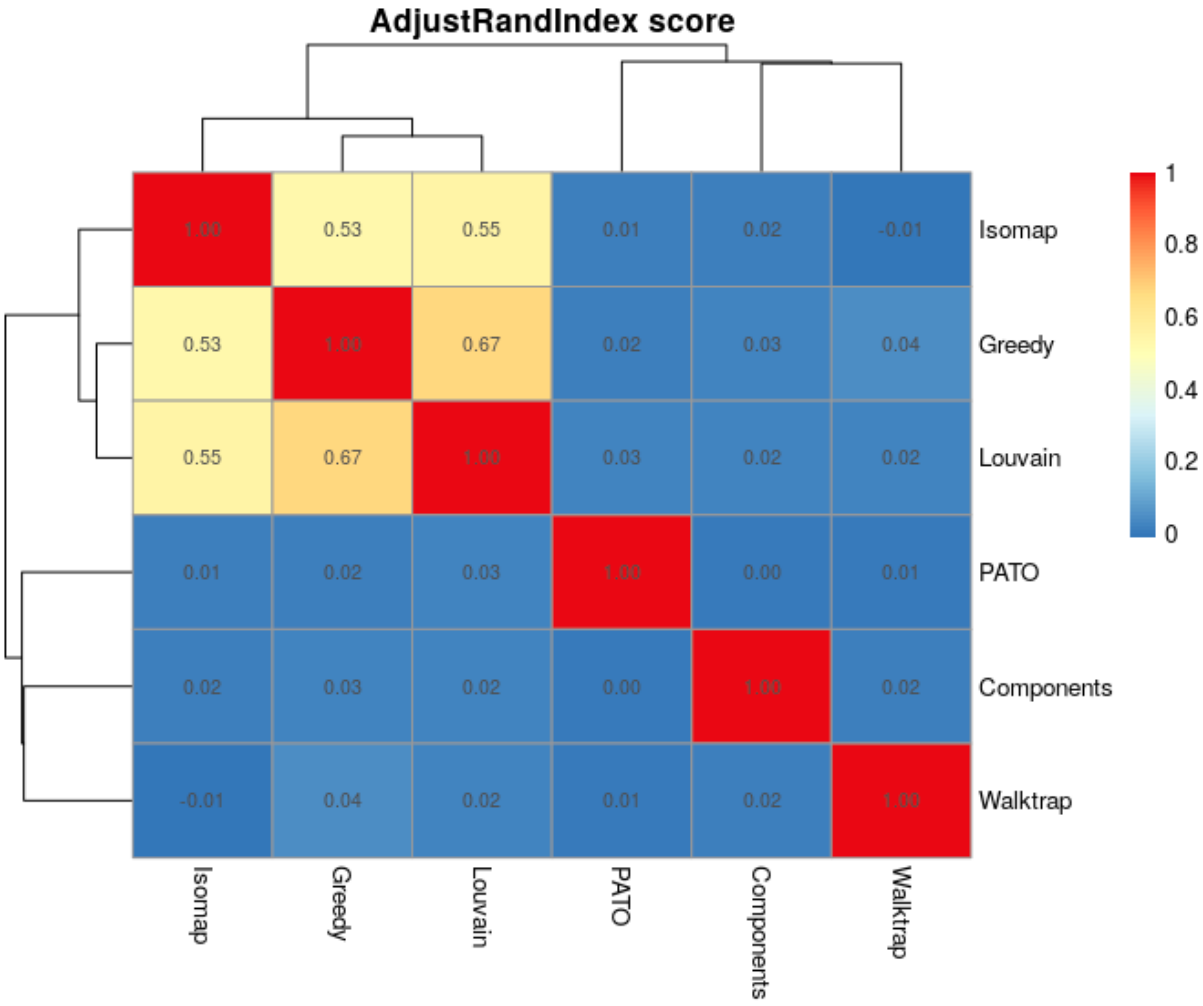


Figure 35. PATO and Coinfinder comparison

4.2.2 BACTAX-ID

4.2.2.1. Algorithm

We aimed to develop an algorithm capable of creating a robust, self-explained, universal hierarchical scheme. We base our clustering method on genome similarity i.e., genomic distance. Distance matrices, characterized by an infinite range of values between 0 and 100, operate within a non-Euclidean space, rendering traditional Euclidean representations inadequate. Specifically, distances between points such as A to B, B to C, and C to A cannot be accurately depicted in an Euclidean space. Our algorithm aimed to accommodate unlimited samples, supporting nested groups *ad infinitum*. We prioritized infinite scalability, ensuring that once sample classification began, it remained unchanged. The clusters formed by the algorithm were intended to be permanent entities, providing a stable organizational structure. To effectively handle continuous variations, we pursued a hierarchical arrangement. This systematic structuring allowed for comprehensive data organization, facilitating nuanced analysis. Importantly, our algorithm's versatility enabled its application to any sample possessing a pairwise distance matrix among its components. Our algorithm operationalized iteratively, analyzing samples one by one. This iterative process allowed for assigning labels to new samples and adding new layers of hierarchical labels to previously analyzed ones. Ultimately, our algorithm's iterative nature ensured adaptability and scalability across various datasets.

Clique and Pseudo-clique

We opted for a clique-based approach among genomes to establish a new cluster. Upon introducing a new genome into the database, it systematically identified the best matches among all existing genomes. A cluster would form when a minimum of five genomes (a configurable parameter) exhibited genomic identities surpassing a specified threshold (e.g., 95% identity) and demonstrated mutual similarity beyond that threshold, constituting a clique. Subsequently, a distinct group with a specific ID number was assigned. We incorporated the pseudo-clique concept to facilitate cluster expansion, providing enhanced flexibility. When a new genome entered the database, it assessed its best match; if the corresponding genome belonged to a clique, the newcomer compared itself to other cluster components. Admission to the cluster occurred if a specified percentage (e.g., 80%) of the cluster members displayed genomic identities above the threshold (95% identity). Otherwise, the genome awaited the entry of new genomes, allowing the potential formation

of its own cluster. This method afforded adaptability and facilitated the organic growth of clusters.

Hierarchical model

In this scenario, employing a single classification layer, akin to MLST, poses challenges, as it fails to elucidate the proper relationships between distinct clusters. Moreover, it may lead to complications, such as disparate cluster compositions contingent on the order of genome entry into the system. To address these issues, we devised a hierarchical multilayer system analogous to matryoshka dolls, where each level encapsulates the subsequent one. We established seven levels ranging from 94% to 99.95% identity to obtain a nuanced perspective of each genus. For each level, we applied the aforementioned classification method of clique and pseudo-clique. Consequently, this approach yielded a numerical code for each genome corresponding to its classification across each level.

Technical approximation

We designed an algorithm to assign a taxonomic identification marker to each genome, encapsulating information relative to other genomes. The algorithm initiates by determining the clustering of genomes into groups, utilizing genomic similarity data generated by the Mash software—a data frame presenting pairwise genomic similarities. In the algorithm's initial step, each new genome is comprehensively compared with those existing in the database. The database, housing relational information, is stored in SQL for efficient accessibility. Simultaneously, a sketches .msh file storing all genomes in the database is maintained. To optimize processing efficiency, a sequential, tiered analysis approach is implemented. This eliminates the need for exhaustive comparisons between individual genomes and the entire database. Instead, each genome is compared only with its corresponding subgroups within the database. This stratified analysis strategy significantly reduces computational demands while simultaneously facilitating the incorporation of additional genomes into the database, thereby promoting its scalability.

Nomenclature.

We introduce a novel nomenclature system designed to categorize genomes intricately through a hierarchical structure. This innovative scheme elucidates the hierarchical classification of genomes, similar to the organization of the Enzyme Code, where each level signifies a progressively specific definition within the enzyme reaction. In our system,

BacTaxID, each tier represents a distinct level of similarity. Commencing at the genus level, we systematically define increasingly specific groups, including species, phylogroups, clonal complexes, sequence types, and beyond. Notably, our system maintains uniformity across all genera, facilitating comprehensive comparisons within and across different genera. This framework promotes intra-genus and inter-genus analyses, enhancing the depth and breadth of genomic classification.

For example, in our classification, if we can give it a complete ID, one strain from Klebsiella means that the strain is part of a group at every level. Its ID would look like *Klebsiella pneumoniae* 1.1.3.4.1.1, for example. That means shows to which groups it is part of each level. So, inside the group pneumonia, the second level would be in group 1. Inside group 1, for the third level, it would be in group 1. For the fourth level, in group 3, et cetera, this way, it is easy to compare the degrees of similarity that every klebsiella has to each other, depending on how many numbers they share.

4.2.2.2. Visualization

To assess the representation of genomes within the databases, we computed the rate of new cluster creation for each level across various genera. Our analysis revealed that lower levels of genomic identity tend to plateau early. Pseudo-rarefaction plots were generated for each genus to visualize cluster growth with the number of genomes. This approach allows us to identify when curves become flatter, indicating a more comprehensive representation of genomes at that level. Conversely, higher-level curves exhibit greater verticality, signifying the emergence of new outbreaks within the species.

Furthermore, we calculated the percentage of genomes at each level successfully identified in a cluster. This metric serves as an indicator of the thoroughness of the description at each level.

The classification results are conveyed through three distinct forms of presentation.

Table displaying the percentage of identified genomes: The table shows the percentage of the studied genomes that have an ID for each level (from T1 to T7). For each increasing level, always less or the same number of genomes are identified. A high percentage shows that that species is well sampled for that level.

Rarefaction plot: The X-axis represents the number of genomes, while the Y-axis indicates the count of groups within a level (t). Each line grows as new groups are added. When the lines flatten, it suggests that the level is well-sampled in the libraries, as additional genomes do not contribute to new group formation. Besides, if a line falls short (does not reach the end of the plot), it suggests that new group additions have ceased for an extended period.

Sunburst plot: A nested arrangement of circles, with each circle symbolizing a level of identification and the extent of color, indicating the number of formed groups (ID). White spaces may be observed in the upper levels as not all genomes possess identification for those levels. The size of each segment corresponds to the percentage of that identification within the population.

This comprehensive analysis was conducted across 44 microbial genera. We have chosen nine bacterial genera of interest in Biomedicine to demonstrate the usefulness of the identification algorithm.

Staphylococcus

Figure 36 shows that 95% of the 12,026 available *Staphylococcus* genomes are identified up to level t5. However, the last two levels significantly decrease the percentage of identified genomes. Notably, the clusters in levels t5 and t6 exhibit substantial growth, whereas the earlier levels remain relatively stable. As illustrated in the sunburst plot, this genus appears relatively stable, which agrees with its clonal structure according to the works in population genetics of *S. aureus* (Planet, Trends Microbiology 2017).

Level	Percentage
t1	99.5676
t2	99.25162
t3	98.85249
t4	98.22884
t5	95.72593
t6	86.60402
t7	13.58723
Number of Genomes	12026

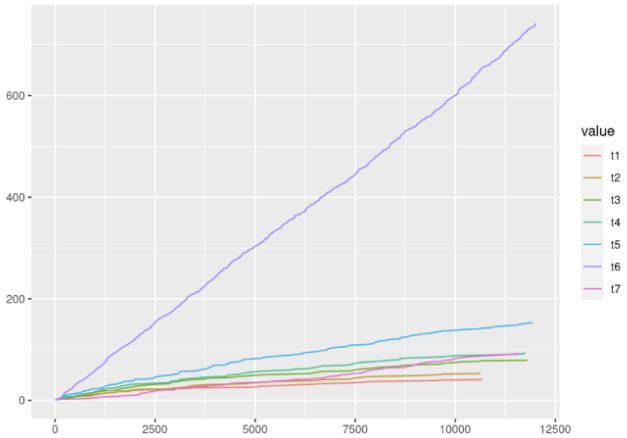




Figure 36. Table of genomes identified, rarefaction plot, and sunburst plot of *Staphylococcus*

Klebsiella

For the genus *Klebsiella*, cluster identification significantly decreases from t4; however, approximately 21% of the genomes obtain an ID at t7, indicating a noteworthy identification rate. The genus appears to exhibit stability in the lower levels but demonstrates high clonality in the upper ones.

Level	Percentage
t1	99.75183
t2	99.4168
t3	99.21827
t4	98.48616
t5	89.29148
t6	71.41085
t7	21.01998
Number of Genomes	8059

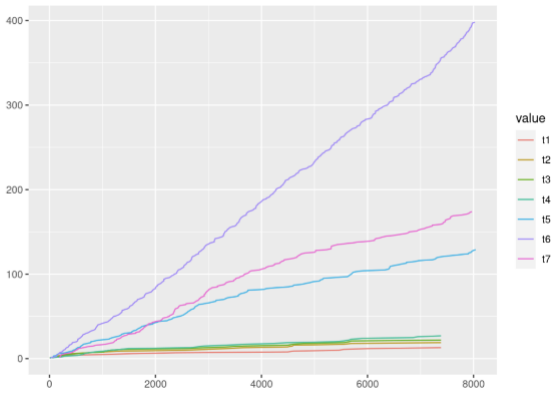




Figure 37. Table of genomes identified, rarefaction plot and sunburst plot of *Klebsiella*

Bacillus

This genus exhibits remarkable diversity across all levels, as shown in **Figure 38**. Notably, only 90% of the genomes receive an ID at level t1, which is low compared to other genera. All curves in Figure 38 display a steep incline, indicating dynamic evolutionary processes within this genus.

Level	Percentage
t1	90.69117
t2	86.97563
t3	83.0803
t4	78.10627
t5	61.40631
t6	35.95685
t7	14.163
Number of Genomes	5006

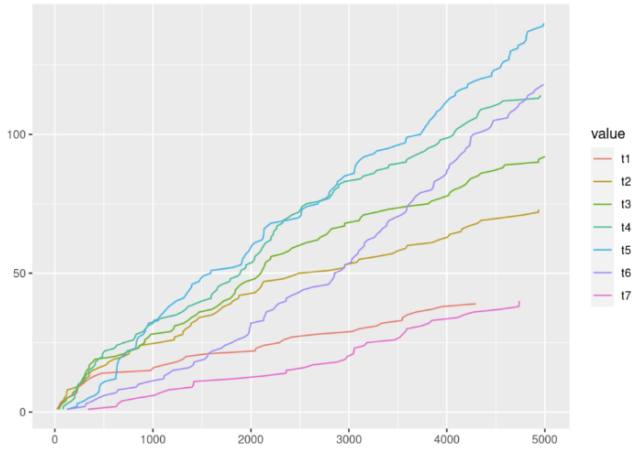




Figure 38. Table of genomes identified, rarefaction plot, and sunburst plot of *Bacillus.*

Escherichia and Shigella

Figure 39 shows the nearly coincidental t1 cluster for *Escherichia* and *Shigella,* which confirms they represent the same species. A gradual decline in identification occurs from t5 onward, with a low percentage of identification at the last level (t7). The curves for t6 and t5 exhibit steep slopes, indicating the ongoing discovery of new clusters in these levels, while t7 shows a less pronounced trend.

Level	Percentage
t1	99.93107
t2	99.87593
t3	99.26477
t4	97.48185
t5	92.202
t6	71.01829
t7	14.3553
Number of Genomes	21762

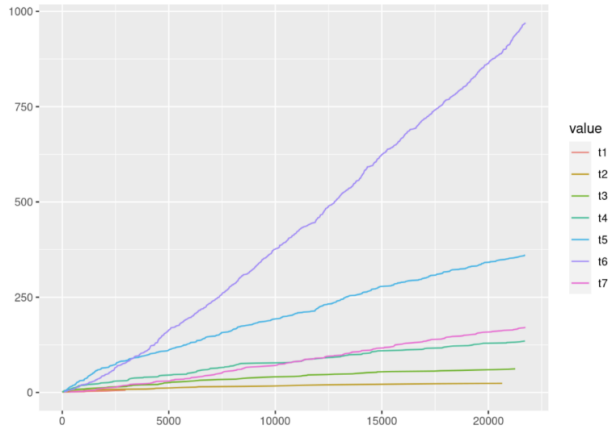




Figure 39. Table of genomes identified, rarefaction plot, and sunburst plot of *Escherichia* and *Shigella*.

Enterococcus

The notable divergence between *E. faecalis* and *E. faecium*, the predominant species of this genus, is evident. In this case, there is a substantial generation of clusters in t6, and t7 exhibits a remarkably low genome identification rate, at only 11%. The observed patterns indicate a divergence in clonalization dynamics between these two species.

Level	Percentage
t1	97.37526
t2	96.8815
t3	95.2183
t4	91.60603
t5	81.49688
t6	57.58836
t7	11.90229

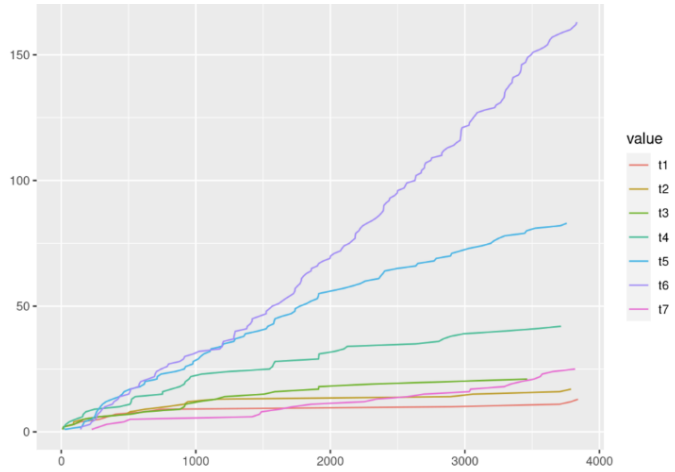




Figure 40. Table of genomes identified, rarefaction plot, and sunburst plot of *Enterococcus.*

Salmonella

This genus is notably well-represented in the repositories, with a substantial percentage of genomes possessing an ID, reaching 48% in t7. The distinctive high slope observed in the last level, t7, indicates a pattern of notable stability and gradual clonality.

Level	Percentage
t1	99.95296
t2	99.85889
t3	99.65898
t4	99.01223
t5	94.95532
t6	90.05174
t7	48.08325
Number of Genomes	8504

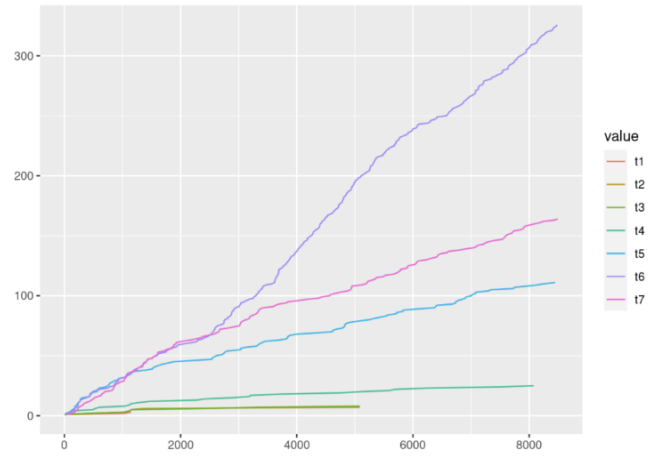




Figure 41. Table of genomes identified, rarefaction plot, and sunburst plot of *Salmonella.*

Enterobacter

In this genus, most levels display steep slopes, indicating the continual emergence of new groups across all levels. Although numerous genomes lack identification, this could be attributed to the relatively limited number of genomes available for analysis.

Level	Percentage
t1	96.91584
t2	95.3999
t3	91.68845
t4	77.62676
t5	57.03084
t6	40.14637
t7	8.991113
Number of Genomes	1913

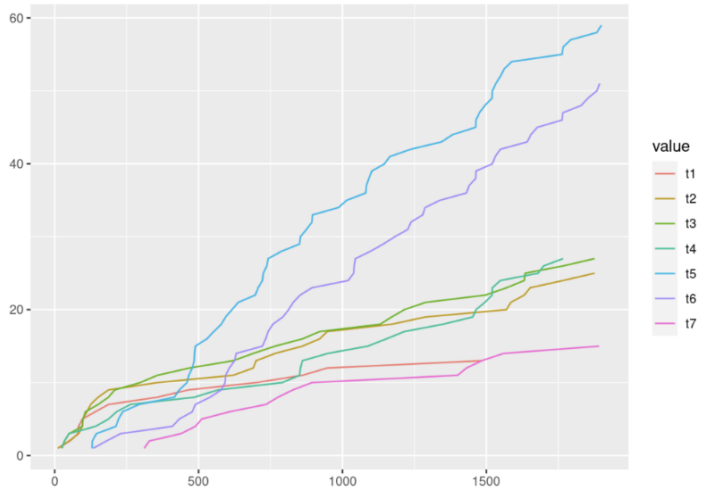


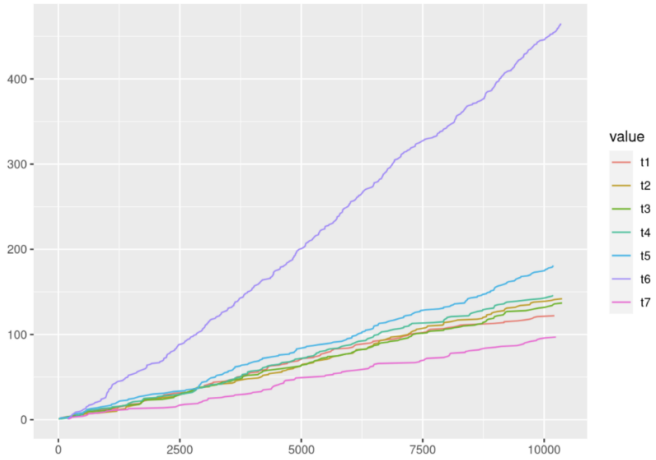


Figure 42. Table of genomes identified, rarefaction plot, and sunburst plot of *Enterobacter.*

Pseudomonas

In this case, all the slopes exhibit considerable elevation in Pseudomonas, indicating the emergence of new groups at each level, with t6 standing out for its exceptionally high slope. The percentage of genomes with an identification is relatively modest compared to other groups, reaching 92% in t1. However, the primary group (*Pseudomonas aeruginosa*) demonstrates stability up to level 6, while other groups exhibit notable clonality.

Level	Percentage
t1	92.47975
t2	88.61358
t3	84.99807
t4	81.14153
t5	73.49595
t6	58.54223
t7	16.61203
Number of Genomes	10372



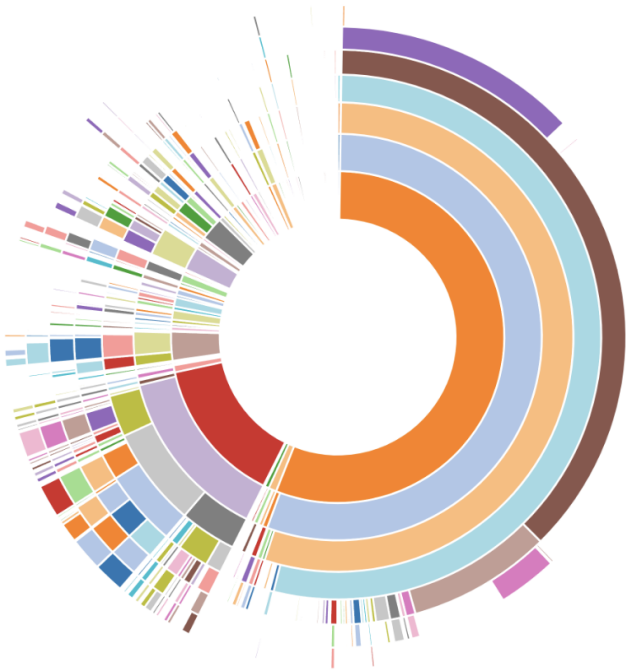


Figure 43. Table of genomes identified, rarefaction plot, and sunburst plot of *Pseudomonas.*

Listeria

This genus exhibits a unique pattern, with the lower levels (t1-t3) showing a substantial number of genomes belonging to an ID. At the same time, t7 has a notably low percentage, and t6 remains relatively high at 78%. This trend suggests a remarkably high degree of clonality in t6, evident in both the slope and the associated figure. The pronounced clonality in t6 makes identifying distinct groups in t7 challenging.

Level	Percentage
t1	99.32432
t2	99.01247
t3	97.60915
t4	95.55613
t5	89.21518
t6	78.69023
t7	3.456341
Number of Genomes	3848

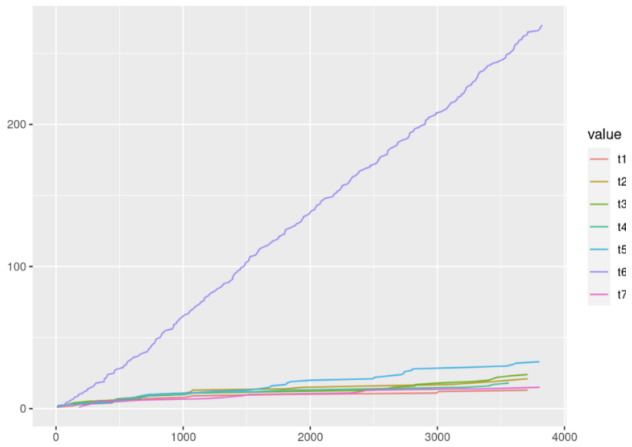




Figure 44. Table of genomes identified, rarefaction plot, and sunburst plot of *Listeria*.

4.3. Application of Genomic Tools to the Analysis of Bacterial Species and Plasmids

Genomics has been increasingly and widely applied to tracking infectious diseases by sequencing the genomes of bacteria, parasites, and viruses since the late 1990s. Genomic surveillance is the process of constantly monitoring pathogens and analyzing their genetic

similarities and differences, and it results strategic for pathogens with pandemic and epidemic potential. It is also widely applied to track the evolutionary processes of microbial populations *in vivo* or *in vitro*.

Surveillance Analysis of antibiotic resistance pathogens is focused on species included in the watch list of the WHO (WHO 2017) or the CDC (CDC 2019). However, genomic surveillance has traditionally used genomic tools that have poorly characterized the accessory genomes, including plasmids and transposons, despite ARGs being usually located on these mobile genetic elements.

In this chapter, we apply the tools developed in section 4.2. (AccNET and PATO) Plus, in combination with available genomic tools, we analyze core genomes to characterize the population structure of significant pathogens involved in clonal or polyclonal outbreaks in our institution. They include

- *Enterococcus faecalis*
- *Enterococcus faecium*

This work has been presented at various national and international conferences as posters or oral communications (Annex 3). Below are those in which the candidate appears as the first author.

- M. D. Fernández-De-Bobadilla. Patrones de selección y transmisión de clones y plásmidos de enterococos en una UCI. SEIMC 2023. Santiago de Compostela. Oral Presentation (manuscript in preparation)

4.3.1. Enterococcus

4.3.1.1. Motivation

Antibiotic use, colonization pressure, and environmental contamination influence their selection and transmission in ICUs. Most available knowledge came from outbreaks and vancomycin-resistant enterococci (VRE), which provide fractionated information. We prospectively analyzed the selection and transmission of enterococci in the largest ICU of a tertiary hospital for 12 months.

4.3.1.2. Genomic analysis

We identified 129 enterococci (86 *E. faecium*, Efm; *32 E. faecalis*, Efc; 11 others) from 179 patients (72% of colonized patients, most vancomycin susceptible VSE).

Figures 45 and 46 show that the clonal diversity of Efm (27 ST80, 18 ST117, and 6 STs in 1-3 isolates) and Efc (8 ST179 and 6 STs with fewer than five isolates) decreased over time, coincidental with different ICU outbreaks of ESBL/carbapenemase-producing enterobacteria and pseudomonas (see chapter 4.1.2).

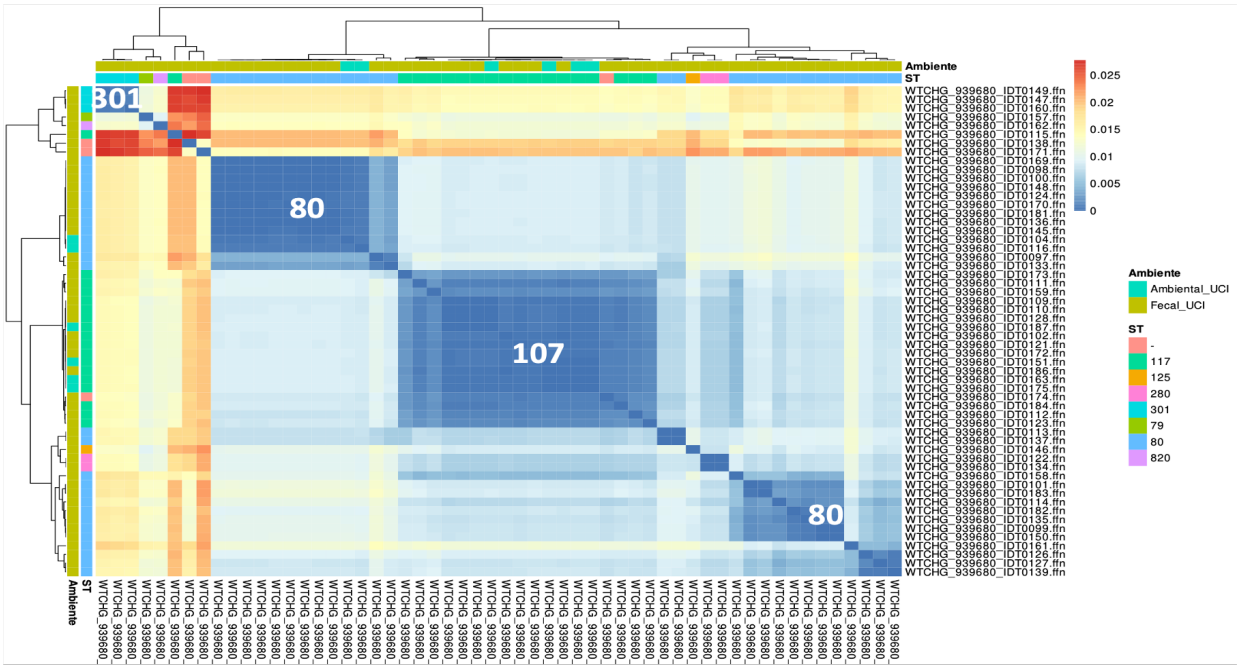


Figure 45. Genomic analysis of *E. faecium* ICU clones. Heatmap of pairwise genomic distances based on the ANI. Numbers represent the STs.

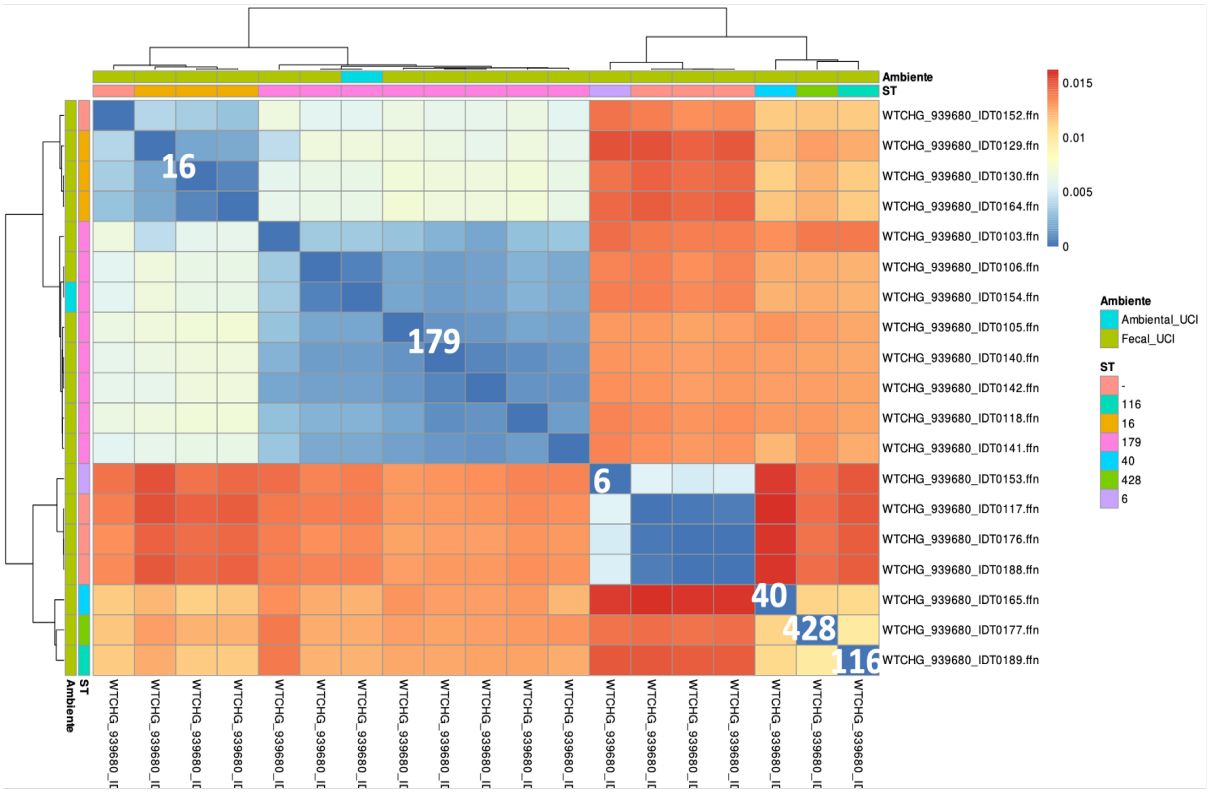


Figure 46. Genomic analysis of *E. faecalis* ICU clones. Heatmap of pairwise genomic distances based on the ANI. Numbers represent the STs.

4.3.1.3. Plasmid analysis

Figure 47 shows the plasmidome of enterococci, which comprises 495 plasmids clustered in 121 plasmid groups (PGs). Only a few of these carry resistance genes. Consistent patterns were observed per cluster.

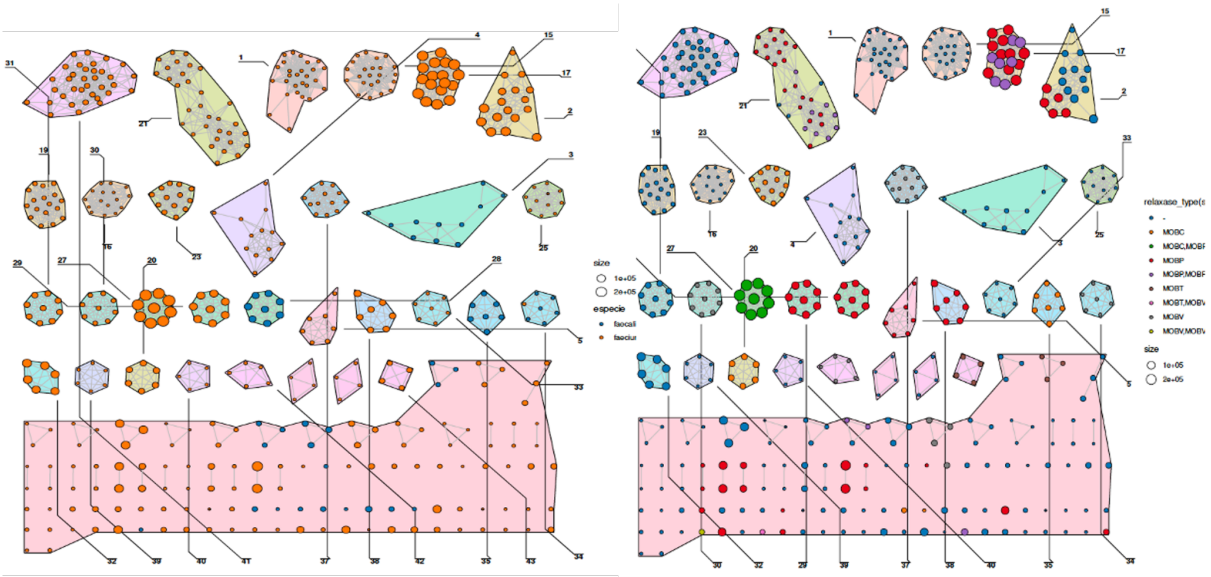


Figure 47. The plasmidome of enterococci. Right: Per relaxase; Left: per species.

4.3.1.4. Novelty and Significance

This study is original and novel as it represents a combined approach to exploring enterococcal diversity using distinct tools to analyze core and accessory genomes. These tools make it easier to trace specific strains.

5. DISCUSSION

5. DISCUSSION

Technical limitations to characterize the abundance and diversity of the microbial world still preclude our understanding of microbial ecosystems and keep challenging the microbial taxonomy and phylogeny of microbes of biomedical relevance in the 21st century. In addition, the lack of debate about the *validity, reliability*, and *comparability* of the (microbial) information generated in Life sciences thwart the utility of the results and, thus, the implementation of efficient interventions. Questions as the *unit of analysis* (samples analyzed) or *indicators* (including the metrics for expressing omic results) are unmet needs in most surveys.

In this thesis, we developed, optimized and/or validated bioinformatic tools to analyze genomic and metagenomic datasets from well-designed specific cohorts (*units of analysis*) and explore the *validity* and *comparability* of traditional *indicators* (sentinel species, HiRCCs, plasmids, and ARGs). The information obtained will be helpful in developing metrics and risk maps.

In the following paragraphs, we will discuss the main findings regarding the goals we initially pursued.

5.1 Metagenomic tools to analyze resistomes

Limitations of the metagenomic tools limit the understanding of microbial ecosystems, including resistomes. In addition, the lack of debate about the validity, reliability, and comparability of the information generated precludes the utility of the results of metagenomic surveys and, thus, the implementation of efficient interventions (24,25,28). Questions as the unit of analysis (samples analyzed) or indicators (including the metrics for expressing resistome results) are unmet needs in most surveys. In this thesis, besides using novel bioinformatic tools, we carefully select specific cohorts (units of analysis) and explore the validity of traditional indicators (some ARGs).

Resistome analysis constitutes the current framework to understand AMR's origin, emergence, and evolution[1]. However, the limited sensibility and specificity of available metagenomic platforms difficult to detect ARGs present at low concentrations in most

healthy microbiomes and the discrimination between gene allelic variants of medical relevance. In **chapter 4.1**, we first used a targeted metagenomics platform developed by our group (29) to analyze the resistomes.

Our analysis of the Amerindians resistome allowed us to analyze ARGs pools in pristine environments and thus establish baseline reference results. Monitoring microbiome changes demands having reference values for comparisons, which are missing in most available studies (84). In the case of AMR, there is a need to characterize the composition of microbiomes given that ARGs are metabolic genes that become resistant to different compounds throughout evolution due to multiple selection events (85). Clemente et al. (74) first demonstrated how the microbiome of the uncontacted Amerindians exhibited the highest diversity ever reported in any human group (26,27). However, this work uses applied standard metagenomics, and neither discriminates between ARG alleles nor provides information about other genes encoding resistance to non-antibiotic antimicrobials or MGEs whose selection could predate the emergence of some ARGs in human populations.

The disparities between Wayampii and Europeans´resistomes found in our study confirm the differences in the microbial composition of these populations. The alpha and beta diversity values from comparing Amerindian and European resistomes revealed differences in the richness and composition of ARGs, MRGs, BRGs, and relaxase genes. Analysis of the demographic factors related to lifestyle and environmental conditions that may exercise microbial selection or favor microbial transmission in different geographic areas seem to contribute to the similarities in richness (similar genes in all individuals tested) but distinctions in gene composition (various combinations of genes in different populations). These socio-geographic factors would explain the homogeneity of resistomes in Wayampiis in contrast with Westerns and the enrichment in some gene populations related to a diet near gold mines (i.e. *mer* genes).

One of the most interesting results of this work was the pool of ARGs found in most people in each group, which was mostly the same ("the core resistome"). This pool comprises ARGs widely recognized as "canonical ARGs" or acquired functional genes apparently resistant to different families of antibiotics (*tetM, ermB*,..). Silencing of these genes has widely been documented in works analyzing individual isolates (86), but recent works suggest that these genes are located downstream sequences that might control their expression and enable their presence with a low cost if they are not functional or they express at basal levels (87).

The emphasis on targeted metagenomics underscores its precision in unraveling resistome intricacies within the Wayampiis. This methodological precision enhances our understanding compared to broader metagenomic approaches. However, how can this precision be leveraged for future research demands the development of metrics that enable the measurement of the resistome or any set of microbiome genes. This issue is related to the *validity* of the information and, more importantly, the *instrument validity* (28). The creation of complex indexes could only be approached with accurate detection of genes and reference values for comparison. taking into account baseline data.

Finally, it should be mentioned that the categorization of gene types in our and other studies relies on the current databases, which are largely biased either by the source of information (most are fed with clinical isolates) or classification criteria (sequence homology instead of experimental validation of the genotype-phenotype correlation). This is currently a hot area of research necessary to understand what ecological dynamics govern the distribution and abundance of ARGs and other microbiome genes , and how might they influence the overall microbial ecosystem within and between human populations.

5.2 Bioinformatic analysis for in-depth genomic analysis

WGS is increasingly used to identify bacterial pathogens and characterize their transmission pathways (88). Next-generation sequencing (NGS), mainly short-reads technology such as Illumina, has been applied extensively to develop whole-genome SNP (wgSNP) or core genome Multi-Locus Sequence Typing (cgMLST) schemes assessing genetic relatedness between genomes of many bacterial species. However, short-read technologies preclude the resolution of large structural variations, gene duplications, or, more importantly, the identification of the accessory genome. In that regard, one of the main disadvantages of short sequencing lies in its inability to close plasmids, a key element in HGT (89). The last weakness reduces the coverage depth involving a lack of characterization regarding drug resistance, virulence, and transmission analysis. Different bioinformatic solutions have been developed to analyze plasmids using different criteria, such as COPLA (90), MOB-SUITE (52), or mlplasmids (91). However, different tools are needed for analyzing different AMR units (clones, plasmids, genes, et cetera).

The **chapter 4.2**. , describes PATO framework, a comprehensive suite of tools necessary for a comprehensive microbial genomics analysis I developed during the thesis. It mainly

focuses on quality control, dataset manipulation, population structure, and pangenome description and addresses some needs highlighted by the EU to analyze AMR (92,93).

First, PATO exhibits versatility in managing nucleotide and amino acid multi-fasta files, proving to be an invaluable resource for researchers dealing with various data types. Moreover, its dedicated function for outliers detection and removal, employing a predefined threshold, guarantees robustness in addressing anomalies within the dataset, thereby significantly enhancing the quality control process.

Secondly, he tool provides a non-repetitive function to enable users to curate a dataset using customizable selection criteria such as distance, percentage of samples, or a specific number of genomes, thus offering flexibility. This iterative approach enhances computational efficiency for large datasets and ensures streamlined data curation processes tailored to individual research needs.

Thirdly, PATO provides a comprehensive suite of tools for population structure and pangenome description. Its capabilities include robust core genome analysis tools for exploring structural components and generating core-genome alignments to facilitate phylogenetic inference. Additionally, PATO provides specialized functions for accessory genome analysis, including enrichment analysis and assessment of gene/protein distribution patterns through the dedicated object type (accnet), enabling a nuanced understanding of genomic diversity. Moreover, the integration of network analysis tools such as singles(), twins(), and coincidents() enhances the exploration of gene associations, allowing users to differentiate between exclusive, identical, and co-occurring genes, thereby providing valuable insights into genomic relationships and interactions. This approach can also be applied to plasmids to create networks of the plasmids present in a community, to create new cluster classifications, and to find correlations among clusters and genes.

In benchmarking against other prominent pangenome analysis software, PATO demonstrates remarkable efficiency, showcasing its ability to handle large-scale analyses swiftly and with computational prowess (94). Notably, PATO's accelerated performance is attributed to its streamlined approach, particularly in forgoing paralogue splitting processes, rendering it a favorable choice for high-throughput genomic analyses, where speed and efficiency are paramount. PATO's phylogenetic tree construction pipelines with Snippy (95) yield a notable performance advantage, underscoring its efficacy and reliability across

various methodologies and emphasizing the consistency of results. PATO introduces a distinct approach to gene association by employing protein clustering and UMAP projection. This, in turn, offers a novel perspective that differs from methodologies such as Coinfinder and enriches the understanding of associative associations between genes.

In conclusion, the PATO framework emerges as a robust and efficient toolkit for microbial genomics analysis, offering various functionalities. Its performance advantages, versatility, and unique approaches to gene association make it a valuable resource for researchers in the field. However, users are encouraged to be mindful of methodological nuances and potential biases, particularly in interpreting results related to population structure and gene associations. PATO's ongoing development and refinement promise continued advancements in microbial genomics research (available at https://github.com/irycisBioinfo/PATO).

5.3 Bioinformatic analysis for in-depth species analysis. The BacTaxID

Chapter 4.2 describes a bioinformatic tool to address a challenge in Public health, such as the emergence of novel epidemic threats. This can occur by selection and further spread of a given clone or by recombination between genomes, resulting in species of increased epidemicity (96). BaxTaxID introduces a novel perspective on intra-species classification, a critical aspect of epidemiology (28). Intra-species classification, commonly known as typing, is the primary tool for characterizing outbreaks and defining HiRCCSs. However, prevailing typing schemes such as MLST, cgMLST, or wgMLST rely on specific labels assigned to distinct sets of alleles, lacking inherent relationships. Consequently, researchers need help determining the similarity between different labels, making it difficult to discern, for example, whether ST11 is more similar to ST12 or ST1012.

In contrast, BacTaxID offers a hierarchical, self-interpretable, and uniform scheme that harmonizes all species schemes within a formalized framework. This hierarchical framework enables unbiased comparisons across diverse population structures of various species or statistical analyses. This is achieved by minimizing compositional bias, ensured by the homogeneity of groups at each level of the hierarchy. The findings presented in this thesis underscore substantial variability in the population structures of specific genera. Genera like

Enterobacter or *Bacillus* exhibit considerable diversity at initial levels (low ANIs), potentially indicative of recent speciation events. Conversely, genera such as *Klebsiella* or *Escherichia/Shigella* reveal more clonal structures, with variability manifesting predominantly at later levels (high ANIs). The results agree with the population structure of these species.

It is crucial to note that, despite BacTaxID's normalized framework, databases lack standardization, leading to a biased representation of clinically relevant strains. Addressing this discrepancy is paramount for ensuring the robustness and generalizability of the classification system.

5.4 Analysis of bacterial species

The **chapter 4.3** show the applications of omic tools to disentangle the dynamics of different *sentinel species* and AMR units such as ARGs and plasmids. The advent of NGS technologies is increasingly integrated into hospital surveillance, however, a notable gap exists in characterizing accessory genomes and mobile genetic elements.

During the performance of this thesis, I carried the analysis of several bacterial species of nosocomial opportunistic pathogens with different lifestyles; either commensal, such as *Escherichia coli* (97), or *Klebsiella*, or associated with environmental reservoirs, such as *Serratia* (98) and *Elizabethkingia* (Annex. 3). This chapter details the example of *Enterococcus faecalis* and *Enterococcus faecium,* which were prospectively analyzed, as well as the ARGs and MGEs encoding resistance to first-line antibiotics (i.e., glycopeptides) in the largest ICU of a tertiary hospital over 12 months. The selection of this *unit of analysis* provided the opportunity to explore the dynamics in the absence of clinical outbreaks.

The dynamic landscape reflects a decreasing clonal diversity over time, coinciding with outbreaks caused by carbapenem-resistant enterobacterales (VIM-1 and OXA-48 producing Enterobacterales and GES-producing *Pseudomonas aeruginosa*) pathogens (Annex. 3). This result showed an increase in the use of antibiotics and further selection of enterococci, which resulted in an increase in colonized patients and the transmission of different HiRCCs. What is surprising is the lack of documented outbreaks despite the transmission. Plasmid analysis highlights the complexity of the plasmidome of *enterococci*, with many plasmid

groups carrying different rep and rel sequences and plasmid occurrence different from what is commonly reported.

Applying ResCap to fecal poolings of the patients, we identified a diversity of van genes (v*anA, vanB, vanC, vanD, vanG*), some of which were not detected by culturing of fecal samples. Possibilities include an anaerobic reservoir in the microbiome for these genes.

The significance of this chapter lies in its novel combined approach, utilizing metagenomics and genomic tools to explore enterococcal diversity comprehensively. It represents the first application of in-depth targeted metagenomics to "population surveillance," treating the ICU ward as a single entity. Our findings emphasize the ICU's role in promoting colonization pressure and the associated risks of selection, transmission, and persistence of ARGs.

In summary, this section provides a multidimensional view, integrating genomics, metagenomics, and epidemiology to enhance our understanding of bacterial pathogen dynamics and AMR in healthcare settings.

6. CONCLUSIONS

6. CONCLUSIONS

The conclusions of this Doctoral Thesis are.

1. ResCap, a tailored sequence capture platform developed by our group, is a powerful approach to accurately characterize resistomes (pool of genes coding for resistance to antibiotics, heavy metals, and biocides) and metalomes using a targeted metagenomics approach. The improvements in the bioinformatic pipeline and the interface to analyze the resulting sequencing data developed in this thesis provide a necessary and friendly -user tool that allows fast, accurate, and intelligible analysis of resistomes in disparate microbial ecosystems.

2. The application of ResCap to the analysis of specific cohorts such as "pristine ecosystems" (Amerindian villages which have never been in contact with antimicrobials or other Western selective pressures for AMR) or "ward resistomes" (pooled samples of inpatients) allows not only to validate the methodology but also respond to different unmet epidemiological questions—first, the need to set "baseline resistomes" to interpret results accurately. Second, to understand the selection, transmission, and persistence of ARGs in different microbial ecosystems.

3. The Pangenome Analysis Toolkit (PATO) can perform the most demanding bioinformatic analyses in minutes with an accuracy comparable to state-of-the-art software but 20-30× times faster. PATO also integrates all the necessary functions for the complete analysis of the most common objectives in microbiology studies. Finally, PATO includes the tools required for visualizing the results, and it is possible to integrate them with other analytical packages available in R.

4. The application of different tools to analyze accessory genomes (included in PATO) enabled the accurate characterization of opportunistic pathogens and their extrachromosomal elements, such as plasmids. To date, most genomic analyses mainly rely on the core genome analysis.

5. The novel tool BACTAX-ID developed in this thesis is a universally applicable identification system that is scalable and neutral. It enables classification at a subspecies level with a high degree of specificity. The simplicity of its nomenclature

renders it well-suited for applications in microbial epidemiology, microbial ecology, and the biogeography of both known and novel species.

6. Combined approaches for both in-depth metagenomic analysis and genomics in high-risk hospital wards (e.g., intensive Care Units, ICUs) enhanced available knowledge about colonization pressure, transmission, and environmental contamination by “antimicrobial resistant priority pathogens” (e.g., vancomycin-resistant enterococci). The study on enterococci demonstrates the risk of transmission involving multiple clones, some of which may endure for prolonged durations, potentially linked with "super spreader" patients. The dispersion of "*van*" genes mirrors the clonal transmission of Vancomycin-Resistant Enterococci (VRE), precisely the VanA type, and underscores the presence of latent, concealed intestinal reservoirs harboring VanD, VanG, and VanB types within these clones.

7. REFERENCES

7. REFERENCES

1. Krieg NR, Holt JG, Bergey DH. Bergey's manual of systematic bacteriology. Baltimore: Williams & Wilkins; 1984.
2. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990 Jun;87(12):4576–9.
3. O?Malley MA. Ernst Mayr, the Tree of Life, and Philosophy of Biology. Biology and Philosophy. 2010;25(4):529–52.
4. Sanger Sequencing - an overview | ScienceDirect Topics [Internet]. [cited 2024 Apr 1]. Available from: https://www.sciencedirect.com/topics/medicine-and-dentistry/sanger-sequencing
5. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004 Dec;68(4):669–85.
6. Chen M, Liu X, Liu Q, Shi D, Li H. 3D genomics and its applications in precision medicine. Cellular & Molecular Biology Letters. 2023 Mar 6;28(1):19.
7. A unified initiative to harness Earth's microbiomes | Science [Internet]. [cited 2024 Apr 1]. Available from: https://www.science.org/doi/10.1126/science.aac8480
8. Boyer CB, Merzbach UC. A History of Mathematics. John Wiley & Sons; 2011. 695 p.
9. Strickland L, Lewis HR. Leibniz on Binary: The Invention of Computer Arithmetic. MIT Press; 2022. 243 p.
10. Charles Babbage's Difference Engines and the Science Museum | Science Museum [Internet]. [cited 2024 Apr 8]. Available from: https://www.sciencemuseum.org.uk/objects-and-stories/charles-babbages-difference-engines-and-science-museum
11. Stein D. Ada: a life and a legacy. 1. paperback ed. Cambridge, Mass.: MIT Press; 1987. 321 p. (MIT Press series in the history of computing).
12. Mazor S. The history of the microcomputer-invention and evolution. Proceedings of the IEEE. 1995 Dec;83(12):1601–8.
13. Hagen JB. The origins of bioinformatics. Nat Rev Genet. 2000 Dec;1(3):231–6.
14. Dayhoff MO, National Biomedical Research Foundation. Atlas of protein sequence and structure. Vol. 5, Supplement 3 ; [edited by] Margaret O. Dayhoff. Silver Spring [Md: National Biomedical Research Foundation; 1979.
15. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463–7.
16. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403–10.
17. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010 Jan;11(1):31–46.
18. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol. 1998 Oct;5(10):R245-249.
19. Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, Larsson DGJ. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Front Microbiol. 2014;5:648.
20. D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic

resistome. Science. 2006 Jan 20;311(5759):374–7.
21. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, et al. Bacterial phylogeny structures soil resistomes across habitats. Nature. 2014 May 29;509(7502):612–6.
22. Perry JA, Westman EL, Wright GD. The antibiotic resistome: what's new? Curr Opin Microbiol. 2014 Oct;21:45–50.
23. Wright GD. The antibiotic resistome. Expert Opin Drug Discov. 2010 Aug;5(8):779–88.
24. Lynch MDJ, Neufeld JD. Ecology and exploration of the rare biosphere. Nat Rev Microbiol. 2015 Apr;13(4):217–29.
25. Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L. High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. mBio. 2015 Jan 27;6(1):e02288-14.
26. Martínez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. Nat Rev Microbiol. 2015 Feb;13(2):116–23.
27. Martínez JL, Coque TM, Baquero F. Prioritizing risks of antibiotic resistance genes in all metagenomes. Nat Rev Microbiol. 2015 Jun;13(6):396.
28. Coque TM, Cantón R, Pérez-Cobas AE, Fernández-de-Bobadilla MD, Baquero F. Antimicrobial Resistance in the Global Health Network: Known Unknowns and Challenges for Efficient Responses in the 21st Century. Microorganisms. 2023 Apr;11(4):1050.
29. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted metagenomics. Microbiome. 2018 Jan 15;6(1):11.
30. Perkins KM, Reddy SC, Fagan R, Arduino MJ, Perz JF. Investigation of healthcare infection risks from water-related organisms: Summary of CDC consultations, 2014—2017. Infect Control Hosp Epidemiol. 2019 Jun;40(6):621–6.
31. Baquero F, Coque TM. Multilevel population genetics in antibiotic resistance. FEMS Microbiology Reviews. 2011 Sep 1;35(5):705–6.
32. Riley LW. Pandemic lineages of extraintestinal pathogenic Escherichia coli. Clin Microbiol Infect. 2014 May;20(5):380–90.
33. Faith JJ, Colombel JF, Gordon JI. Identifying strains that contribute to complex diseases through the study of microbial inheritance. Proc Natl Acad Sci U S A. 2015 Jan 20;112(3):633–40.
34. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using Genomics to Track Global Antimicrobial Resistance. Front Public Health. 2019;7:242.
35. European Centre for Disease Prevention and Control. ECDC strategic framework for the integration of molecular and genomic typing into European surveillance and multi-country outbreak investigations: 2019–2021. [Internet]. LU: Publications Office; 2019 [cited 2024 Apr 8]. Available from: https://data.europa.eu/doi/10.2900/805317
36. Ko KKK, Chng KR, Nagarajan N. Metagenomics-enabled microbial surveillance. Nat Microbiol. 2022 Apr;7(4):486–96.
37. National Academies of Sciences, Engineering, and Medicine, National Academy of Engineering, Division on Engineering and Physical Sciences, Health and Medicine Division, Division on Earth and Life Studies, Board on Infrastructure and the Constructed Environment, et al. Microbiomes of the Built Environment: A Research Agenda for Indoor Microbiology, Human Health, and Buildings [Internet]. Washington (DC): National Academies Press (US); 2017 [cited 2024 Apr 8]. (The National Academies Collection: Reports funded by National Institutes of Health). Available from: http://www.ncbi.nlm.nih.gov/books/NBK458827/

38. Santiago M, Eysenbach L, Allegretti J, Aroniadis O, Brandt LJ, Fischer M, et al. Microbiome predictors of dysbiosis and VRE decolonization in patients with recurrent C. difficile infections in a multi-center retrospective study. AIMS Microbiol. 2019 Jan 17;5(1):1–18.
39. Jeffery IB, Claesson MJ, O'Toole PW, Shanahan F. Categorization of the gut microbiota: enterotypes or gradients? Nat Rev Microbiol. 2012 Sep;10(9):591–2.
40. Benveniste R, Davies J. Aminoglycoside Antibiotic-Inactivating Enzymes in Actinomycetes Similar to Those Present in Clinical Isolates of Antibiotic-Resistant Bacteria. Proc Natl Acad Sci USA. 1973 Aug;70(8):2276–80.
41. Lee K, Kim DW, Lee DH, Kim YS, Bu JH, Cha JH, et al. Mobile resistome of human gut and pathogen drives anthropogenic bloom of antibiotic resistance. Microbiome. 2020 Jan 7;8(1):2.
42. Scott LC, Lee N, Aw TG. Antibiotic Resistance in Minimally Human-Impacted Environments. Int J Environ Res Public Health. 2020 Jun;17(11):3939.
43. Maestre-Carballa L, Navarro-López V, Martinez-Garcia M. A Resistome Roadmap: From the Human Body to Pristine Environments. Frontiers in Microbiology [Internet]. 2022 [cited 2023 Oct 19];13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.858831
44. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015 Feb 26;372(9):793–5.
45. David S. Thaler. Precision Public Health: What Is It? | Blogs | CDC [Internet]. 2018 [cited 2024 Apr 1]. Available from: https://blogs.cdc.gov/genomics/2018/05/15/precision-public-health-2/
46. Laperche V, Maury-Brachet R, Blanchard F, Dominique Y, Durrieu G, Massabuau JC, et al. Étude réalisée dans le cadre des projets de Service Public du BRGM - PDR05GUY02. 2007;
47. Ruppé E, Ghozlane A, Tap J, Pons N, Alvarez AS, Maziers N, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. Nat Microbiol. 2019 Jan;4(1):112–23.
48. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom. 2021 Nov;7(11):000685.
49. Fernández-de-Bobadilla MD, Talavera-Rodríguez A, Chacón L, Baquero F, Coque TM, Lanza VF. PATO: Pangenome Analysis Toolkit. Bioinformatics. 2021 Dec 7;37(23):4564–6.
50. Fraley C, Raftery A. Model-based Methods of Classification: Using the mclust Software in Chemometrics. Journal of Statistical Software. 2007 Jan 10;18:1–13.
51. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Research. 2021 Jul 2;49(W1):W293–6.
52. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom. 2018 Aug;4(8):e000206.
53. Carattoli A, Hasman H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). Methods Mol Biol. 2020;2075:285–94.
54. Florensa AF, Kaas RS, Clausen PTLC, Aytan-Aktug D, Aarestrup FM. ResFinder - an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. Microb Genom. 2022 Jan;8(1):000748.
55. Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DGJ. BacMet: antibacterial biocide and metal resistance genes database. Nucleic Acids Res. 2014 Jan

1;42(Database issue):D737–43.
56. Gilchrist CLM, Chooi YH. clinker & clustermap.js: automatic generation of gene cluster comparison figures. Bioinformatics. 2021 Aug 25;37(16):2473–5.
57. Guglielmini J, Néron B, Abby SS, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. Key components of the eight classes of type IV secretion systems involved in bacterial conjugation or protein secretion. Nucleic Acids Res. 2014 May;42(9):5715–27.
58. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Apr;9(4):357–9.
59. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. 2018 Aug 29;19(1):307.
60. Cury J, Abby SS, Doppelt-Azeroual O, Néron B, Rocha EPC. Identifying Conjugative Plasmids and Integrative Conjugative Elements with CONJscan. In: de la Cruz F, editor. Horizontal Gene Transfer: Methods and Protocols [Internet]. New York, NY: Springer US; 2020 [cited 2023 Nov 23]. p. 265–83. (Methods in Molecular Biology). Available from: https://doi.org/10.1007/978-1-4939-9877-7_19
61. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biology. 2011 Jun 24;12(6):R60.
62. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017 Nov;35(11):1026–8.
63. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology. 2016 Jun 20;17(1):132.
64. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018 Sep 15;34(18):3094–100.
65. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current Protocols in Bioinformatics. 2014;47(1):11.12.1-11.12.34.
66. Chen L, Zheng D, Liu B, Yang J, Jin Q. VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. Nucleic Acids Res. 2016 Jan 4;44(D1):D694-697.
67. Bayliss SC, Thorpe HA, Coyle NM, Sheppard SK, Feil EJ. PIRATE: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. Gigascience. 2019 Oct 1;8(10):giz119.
68. Zhou Z, Charlesworth J, Achtman M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. Genome Res. 2020 Nov;30(11):1667–79.
69. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol. 2020 Jul 22;21(1):180.
70. Perrin A, Rocha EPC. PanACoTA: a modular tool for massive microbial comparative genomics. NAR Genom Bioinform. 2021 Jan 12;3(1):lqaa106.
71. Ding W, Baumdicker F, Neher RA. panX: pan-genome analysis and exploration. Nucleic Acids Res. 2018 Jan 9;46(1):e5.
72. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.
73. Amstutz P. Théorie des graphes et ses applications deC. Berge. Ann Télécommun. 1961 Jan 1;16(1):61–61.
74. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, et al. The microbiome of uncontacted Amerindians. Science Advances. 2015 Apr 17;1(3):e1500183.
75. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al.

Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proceedings of the National Academy of Sciences. 2010 Aug 17;107(33):14691–6.
76. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun. 2014 Apr 15;5(1):3654.
77. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012 Jun;486(7402):222–7.
78. Gosalbes MJ, Vázquez-Castellanos JF, Angebault C, Woerther PL, Ruppé E, Ferrús ML, et al. Carriage of Enterobacteria Producing Extended-Spectrum β-Lactamases and Composition of the Gut Microbiota in an Amerindian Community. Antimicrob Agents Chemother. 2016 Jan;60(1):507–14.
79. Grenet K, Guillemot D, Jarlier V, Moreau B, Dubourdieu S, Ruimy R, et al. Antibacterial Resistance, Wayampis Amerindians, French Guyana. Emerg Infect Dis. 2004 Jun;10(6):1150–3.
80. Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. Nat Rev Microbiol. 2007 Mar;5(3):175–86.
81. Sela I, Wolf YI, Koonin EV. Assessment of assumptions underlying models of prokaryotic pangenome evolution. BMC Biol. 2021 Feb 10;19(1):27.
82. Colwell RK, Mao CX, Chang J. Interpolating, Extrapolating, and Comparing Incidence-Based Species Accumulation Curves. Ecology. 2004;85(10):2717–27.
83. Martínez JL, Coque TM, Baquero F. Prioritizing risks of antibiotic resistance genes in all metagenomes. Nat Rev Microbiol. 2015 Jun;13(6):396.
84. Huijbers PMC, Flach CF, Larsson DGJ. A conceptual framework for the environmental surveillance of antibiotics and antibiotic resistance. Environment International. 2019 Sep 1;130:104880.
85. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, et al. Country-specific antibiotic use practices impact the human gut resistome. Genome Res. 2013 Jul;23(7):1163–9.
86. Novais C, Freitas AR, Silveira E, Baquero F, Peixe L, Roberts AP, et al. A tet(S/M) hybrid from CTn6000 and CTn916 recombination. Microbiology (Reading). 2012 Nov;158(Pt 11):2710–1.
87. Hu J, Chen J, Ma L, Hou Q, Zhang Y, Kong X, et al. Characterizing core microbiota and regulatory functions of the pig gut microbiome. ISME J. 2024 Jan 12;18(1):wrad037.
88. Ferdinand AS, Kelaher M, Lane CR, da Silva AG, Sherry NL, Ballard SA, et al. An implementation science approach to evaluating pathogen whole genome sequencing in public health. Genome Medicine. 2021 Jul 28;13(1):121.
89. Arredondo-Alonso S, Willems RJ, van Schaik W, Schürch AC. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microbial Genomics. 2017;3(10):e000128.
90. Redondo-Salvo S, Bartomeus-Peñalver R, Vielva L, Tagg KA, Webb HE, Fernández-López R, et al. COPLA, a taxonomic classifier of plasmids. BMC Bioinformatics. 2021 Jul 31;22(1):390.
91. Arredondo-Alonso S, Rogers MRC, Braat JC, Verschuuren TD, Top J, Corander J, et al. mlplasmids: a user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. Microbial Genomics. 2018;4(11):e000224.
92. Angers-Loustau A, Petrillo M, Bengtsson-Palme J, Berendonk T, Blais B, Chan KG, et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation

sequencing technologies. F1000Res. 2018;7:ISCB Comm J-459.
93. Petrillo M, Fabbri M, Kagkli DM, Querci M, Van den Eede G, Alm E, et al. A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing. F1000Res. 2022 Mar 16;10:80.
94. Fernández-de-Bobadilla MD, Talavera-Rodríguez A, Chacón L, Baquero F, Coque TM, Lanza VF. PATO: Pangenome Analysis Toolkit. Bioinformatics. 2021 Dec 7;37(23):4564–6.
95. Seemann T. tseemann/snippy [Internet]. 2024 [cited 2024 Apr 8]. Available from: https://github.com/tseemann/snippy
96. Croucher NJ, Klugman KP. The Emergence of Bacterial "Hopeful Monsters." mBio. 2014 Jul 29;5(4):e01550-14.
97. Rodríguez I, Figueiredo AS, Sousa M, Aracil-Gisbert S, Fernández-de-Bobadilla MD, Lanza VF, et al. A 21-Year Survey of Escherichia coli from Bloodstream Infections (BSI) in a Tertiary Hospital Reveals How Community-Hospital Dynamics of B2 Phylogroup Clones Influence Local BSI Rates. Castanheira M, editor. mSphere. 2021 Dec 22;6(6):e00868-21.
98. Aracil-Gisbert S, Fernández-De-Bobadilla MD, Guerra-Pinto N, Serrano-Calleja S, Pérez-Cobas AE, Soriano C, et al. The ICU environment contributes to the endemicity of the "Serratia marcescens complex" in the hospital setting. mBio. 2024 Apr 2;0(0):e03054-23.

8. ANNEXES

8. ANNEXES

Annex 1. Analysis pipeline

https://github.com/migueldiezfdz

Annex 2. Statistics data frame

https://docs.google.com/spreadsheets/d/1LK2k4hF5wqRKjXUFVd8GPthZwNSnzusxR_gsbENeHE0/edit?usp=sharing

Annex 3. Poster and oral presentations submitted to national or international conferences

1. Title of the work: Patrones de selección y transmisión de clones y plásmidos de enterococos en una UCI Name of the conference: XXVI Congreso SEIMC City of event: Santiago de Compostela, Galicia, Spain Date of event: 01/06/2023 End date: 03/06/2023 Organising entity: Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica Type of entity: Associations and Groups D. Fernández-de-Bobadilla 1; F. Lanza; M. Coque. "Patrones de selección y transmisión de clones y plásmidos de enterococos en una UCI".
2. Title of the work: Bacterial Taxonomic Identifier (BACTAX-ID) a new method for bacterial classification Name of the conference: FUTURE LEADERS IN AMR RESEARCH WORKSHOP City of event: Madrid, Date of event: 17/03/2023 D. Fernández-de-Bobadilla.
3. Title of the work: Selection and transmisión patterns of antimicrobial resistance in a large Intensive Care Unit using a population-wide surveillance approach Name of the conference: 13th International Meeting on Microbial Epidemiological Markers (IMMEM XIII) Corresponding author: Yes City of event: Bath, United Kingdom Date of event: 14/09/2022 End date: 17/09/2022 Organising entity: ESCMID D. Fernandez-de-Bobadilla; Aracil-Gisbert; F. Lanza; Baquero; M. Coque.
4. Title of the work: Meta-análisis de resistomas: Análisis de la resistencia a antibióticos bajo una perspectiva ONE HEALTH Name of the conference: SEIMC 2022 Corresponding author: Yes City of event: Granada, Andalusia, Spain Date of

event: 02/06/2022 End date: 04/06/2022 D. Fernández-de-Bobadilla; F. Lanza; Baquero; M. Coque.

5. Title of the work: Dynamics of a “hospital ward resistome” using pooled fecal patient samples, advanced targeted metagenomic, and metadata analysis. Name of the conference: 2022 Meetings in Microbial Ecology & Evolution Corresponding author: Yes City of event: Plön, Germany Date of event: 16/05/2022 31564b2f8b42a23742fc9dbd79b1ae8c 9 End date: 20/05/2022 Organising entity: Max-Planck-Institute for Evolutionary Biology D. Fernandez-de-Bobadilla; Aracil-Gisbert; Pérez-Cobas 3; Guerra-Pinto; F. Lanza; Baquero; M. Coque.
6. Title of the work: Bacterial Taxonomic Identifier BACTAX-ID a new method for plasmid, viral and bacterial classification. Name of the conference: ECCMID 2021 City of event: Online, Date of event: 09/07/2021 End date: 12/07/2021 Organising entity: ESCMID Miguel D. Fernández-de-Bobadilla; Fernando Baquero; Teresa M. Coque; Val F. Lanza.
7. Title of the work: Ecology and phylogeny of the genus Elizabethkingia in the hospital environment Name of the conference: ECCMID 2021 City of event: Online, Date of event: 09/07/2021 End date: 12/07/2021 Organising entity: ESCMID Miguel D. FERNÁNDEZ DE BOBADILLA; Natalia GUERRA-PINTO; Sonia ARACIL-GISBERT; María Cruz SORIANO-CUESTA; Almudena ORTIZFERNÁNDEZ (; Sergio GALLEGO-ZARZOSA; Marina LÓPEZ-OLIVENCIA; Rafael Cantón; Raul de Pablo; Fernando Baquero; Val F. Lanza; Teresa M. Coque.
8. Title of the work: PATO: a suite of bioinformatics tools for analysing pangenomes Name of the conference: ECCMID 2021 City of event: Online, Date of event: 09/07/2021 End date: 12/07/2021 Organising entity: ESCMID Miguel D. FERNÁNDEZ-DE-BOBADILLA; Alba TALAVERARODRÍGUEZ; Lucía CHACÓN; Fernando BAQUERO; Teresa COQUE; Val F. LANZA.
9. Title of the work: PATO: Un paquete de herramientas bioinformáticas para analizar pangenomas Name of the conference: SEIMC 2021 City of event: Online, Date of event: 05/06/2021 End date: 11/06/2021 Organising entity: Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica Type of entity: Associations and Groups D. Fernandez-de-Bobadilla; Alba Talavera-Rodríguez; Lucía Chacón; Fernando Baquero; Teresa M. Coque; Val F. Lanza.
10. Title of the work: Analysis of the intestinal resistome of Amerindians from the French Guiana and individuals from European countries Name of the conference:

12th International Meeting on Microbial Epidemiological Markers (IMMEM XII) City of event: Dubrovnik, Croatia Date of event: 18/09/2019 End date: 21/09/2019 Organising entity: ESCMID Miguel D. Fernández-de-Bobadilla; Fernando Baquero; Jose Luis Martinez; Antoine Andremont; Etienne Ruppe; Val F. Lanza; Teresa M. Coque. 31564b2f8b42a23742fc9dbd79b1ae8c 10

11. Title of the work: Análisis del resistoma intestinal de individuos amerindios de la Guayana francesa e individuos europeos Name of the conference: SEIMC 2019 XXIII CONGRESO NACIONAL DE LA SOCIEDAD ESPAÑOLA DE ENFERMEDADES INFECCIOSAS Y MICROBIOLOGÍA CLÍNICA City of event: Madri, Community of Madrid, Spain Date of event: 23/05/2019 End date: 25/05/2019 Organising entity: Sociedad Española de Enfermedades Infecciosas y Microbiología Clínica Type of entity: Associations and Groups Miguel Díez Fernández de Bobadilla; Fernando Baquero; Jose Luis Martínez; Antoine Andremont; Etienne Ruppé; Val Fernandez Lanza; Teresa Coque.
12. Title of the work: Accnet2: A tool for accessory genome comparison and statistical analysis. Name of the conference: JBI-2018 XIV SYMPOSIUM ON BIOINFORMATICS City of event: Granada, Andalusia, Spain Date of event: 14/11/2018 End date: 16/11/2018 Organising entity: Spanish National Bioinformatics Institute (INB) Val Fernández Lanza; Miguel Díez Fernández de Bobadilla; Alba Talavera; Fernando Baquero; Teresa Coque.
13. Title of the work: One locus sequencing typing (OLST), a novel metagenomic approach to analyze intra-species abundance and diversity Name of the conference: MEEGID XIV INTERNATIONAL CONFERENCE ON MOLECULAR EPIDEMIOLOGY AND EVOLUTIONARY GENETICS OF INFECTIOUS DISEASES City of event: Sitges, Catalonia, Spain Date of event: 07/11/2018 End date: 09/11/2018 Organising entity: Elsevier Miguel Díez Fernández de Bobadilla; Sonia Aracil Gisbert; Fernando Baquero; Teresa Coque.