TECHNICAL UNIVERSITY OF MOLDOVA

DOCTORAL SCHOOL

As manuscript
UDC: 004.9:519.21:575

MUNTEANU VIOREL

PHYLOGENY BASED CONTINUOUS-TIME MARKOV MODELS FOR GENE DYNAMICS IN MICROBIAL PANGENOMES

122.03 Modeling, mathematical methods, software products

Summary of the doctoral thesis in informatics

Scientific Supervisor:

BostanVIOREL, doctor habilitate, university professor

CHIȘINĂU, 2026

The Ph.D. thesis has been elaborated within Department of **"Software Engineering and Automatics"**, **Faculty of Computers, Informatics and Microelectronics at Technical University of Moldova**, **Doctoral School of the Technical University of Moldova**

Author: MUNTEANU VIOREL
Scientific supervisor:
BOSTAN Viorel — corresponding member of the Academy of Sciences of Moldova, habilitated doctor in technical sciences, university professor, Technical University of Moldova, Moldova

Members of the advisory committee:

Chairman: GUȚULEAC Emilian, habilitated doctor in technical sciences, university professor, Technical University of Moldova, Moldova
Member: BOSTAN Viorel, corresponding member of the Academy of Sciences of Moldova, habilitated doctor in technical sciences, university professor, Technical University of Moldova, Moldova
Referent: MANGUL Serghei, doctor in computer science, university professor, University of Southern California, CA, US
Referent: LOBIUC Andrei, doctor in biology, university professor, "Ștefan cel Mare" University of Suceava, Romania
Referent: CEPOI Liliana, corresponding member of the Academy of Sciences of Moldova, habilitated doctor in biology, university professor, Technical University of Moldova, Moldova

The public defense of the doctoral thesis will take place on 30 March 2026 at 09:00 before the Public Doctoral Commission of the Doctoral School of the Technical University of Moldova (established by the decision of the Scientific Council of TUM of 27 January 2026, Protocol No. 1), at 9/7 Studenților Street, Building 3, Aula 3-3 "Amdaris", MD-2068, Chișinău, Republic of Moldova.

The doctoral thesis and the abstract can be consulted at the library of the Technical University of Moldova and on the ANACEC website (www.anacec.md).

Author:
VIOREL Munteanu

Scientific supervisor:
BOSTAN Viorel, habilitated doctor in technical sciences, university professor

RESEARCH CONCEPT GUIDELINES

The actuality of the research. The evolutionary dynamics of microbial genome content, driven by mutation, selection, recombination, horizontal gene transfer, and other gene duplication and loss events, underlie the ecological versatility and adaptive potential of microbial populations. [1–4]. These processes generate extensive genomic fluidity, promoting functional diversification and enabling rapid responses to environmental pressures in natural, engineered, and host-associated ecosystems. The pangenome has emerged as a powerful conceptual framework for capturing this genomic variability, yet existing approaches remain largely isolate-centric, limiting their applicability in complex environments and among uncultured microbial populations [5–7]. As a result, there is a growing need for bioinformatics methods capable of reconstructing and interpreting microbial pangenomes directly from metagenomic data, independently of predefined reference genomes [8]. Meeting this need is increasingly urgent, given that environmental and host-associated microbiomes are now recognized as major reservoirs of antimicrobial resistance, metabolic innovation, and pathogenic potential [9]. Quantifying *in situ* gene-content dynamics and deploying urban genomic surveillance systems that monitor Antimicrobial Resistance Genes (ARGs) and virulence determinants across city infrastructures have therefore become essential [6, 10]. In anthropogenically structured ecosystems, including urban wastewater, soils, and air, quantifying the rates and mechanisms of gene flux (horizontal gene transfer (HGT), recombination, gene gain and loss) is critical for tracing the emergence of adaptative traits and forecasting microbial responses to selective pressures [6, 11, 12]. Comparative genomics models have become essential for elucidating these processes, providing a rigorous mathematical framework for representing the evolution of sequences and discrete traits using continuous-time Markov chains (CTMCs) [13, 14]. These models enable the identification of evolutionary events along phylogenetic branches which are fundamental for characterizing the mechanisms that shape the evolutionary dynamics of characters across lineages. Extending and adapting these comparative frameworks to genome- and pangenome-scale analyses at gene-level resolution is increasingly necessary, as such approaches enable the capture of the full spectrum of genomic diversity, including uncultured and low-abundance microbial lineages. By integrating phylogenetic context with structural and functional genomic variation, these models can support a deeper exploration of the evolutionary forces driving adaptive trait emergence and the rapid reorganization of genomic repertoires within microbial communities.

The importance of addressed issue. The study of microbial genome evolution stands at the intersection of bioinformatics, mathematical modeling, metagenomic data analysis, and comparative

genomics [5–7]. Rapid advances in sequencing technologies have dramatically increased the volume and resolution of genomic data, expanding the scope of computational genomics, an emerging discipline concerned with understanding gene-content evolution across microbial populations [9, 15]. Integrating phylogenetic comparative methods with metagenomic analysis is crucial for resolving lineage-specific patterns of gene gain and loss, especially in environments where isolate genomes are unavailable. Such integrative approaches contribute to a deeper understanding of genome dynamics in natural communities and support applications in public-health microbiology, antimicrobial resistance (AMR) surveillance, and real-time evolutionary monitoring in complex ecosystems. Urban microbiomes present a particularly challenging context due to their pronounced genome plasticity, thus influencing microbial adaptation, transmission, and the dissemination of antimicrobial resistance within densely populated and anthropogenically shaped environments [4, 9, 10, 15]. Capturing these evolutionary processes is central to understanding how microbial functions emerge, persist, and spread in cities. Despite significant progress in the field, most computational tools of genome evolution still rely on isolate-based and reference-dependent frameworks [12, 16–18]. Addressing these challenges requires phylogeny-based pangenomic and metapangenomic bioinformatics software capable of reconstructing gene repertoires and modeling gain–loss dynamics directly from metagenomic data. Ancestral state reconstruction methods based on CTMCs, originally developed for trait evolution, can be adapted to model gene-content evolution, enabling the recovery of fine-scale evolutionary trajectories across environmental lineages [9, 14, 19, 20]. Therefore, there is a growing need for new methods and bioinformatics software capable of integrating tree-based evolutionary models with large-scale genomic datasets. Such software is essential for detecting lineage-specific adaptation, quantifying gene-turnover rates, and distinguishing conserved from variable components of the microbial pangenome by coupling metagenomic data with phylogenetic comparative and genomic methods, thereby addressing a major methodological gap in metagenomic surveillance [4, 6, 9, 10, 12, 15, 21–23]. Likewise, the ability to assess selective pressure on conserved genes and functions has become essential for identifying lineage-specific phyletic patterns of preferential gene retention, loss or acquisition that deviate from neutral expectations across distinct ecological contexts [24, 25].

The purpose and objectives of the research. The doctoral thesis aims to: (i) develop and validate a modular, scalable, and reproducible bioinformatics software for meta-pangenome reconstruction and analysis from metagenomic data; (ii) integrate probabilistic modeling and phylogeny-based inference to quantify gene-content variability and gene-turnover in pangenomes; and (iii) to develop and implement a statistical method to classify genes in pangenomes according

to selective pressure along taxonomic lineages.

Proposed research objectives:

- Develop and implement a reproducible, modular bioinformatics software that converts labeled genomic sequences into meta-pangenome datasets by unifying standardized annotation and orthogroup inference with construction of presence–absence gene matrices, quantitative openness metrics and core and accessory delineation, as well as recombination-free phylogenies from core alignments, yielding interoperable outputs for downstream modeling.
- Develop and implement a phylogeny-based CTMC probabilistic software that quantifies gene-content dynamics across phylogenies from metagenomic presence–absence data by inferring lineage-specific gain and loss processes and generating branch- and lineage-level summaries suitable for comparative analyses and genomic surveillance.
- Build a phylogeny-based CTMC-approach software that classifies genes by selective regime, by distinguishing symmetric from asymmetric gain–loss dynamics and quantifies the directionality of the gain–loss process (the tilt toward acquisition versus deletion), using gain/loss rate-contrast indices to capture the evolutionary tendency of gene-content change.
- Rigorously validate the end-to-end developed software by benchmarking its performance against gold-standard datasets, applying it to urban metagenomic data to reconstruct meta-pangenomes, and finally performing CTMC-based phylogenetic inference to estimate gene gain and loss counts and classify genes according to selective regime.

Scientific research methodology. In this thesis, we develop bioinformatics software for reconstructing meta-pangenomes from metagenomic data and for inferring lineage-specific gene gain and loss, as well as selection direction on a core-genome phylogeny, and demonstrate its application as a *proof-of-concept* using empirical *Klebsiella* genomes datasets. Three curated datasets were analyzed, an urban MAG sequences collection (for *Klebsiella* genus), and two isolate collections (*Klebsiella* genus collection of sequences and *K. pneumoniae* single species dataset), enabling comparisons across different quality data types and taxonomic scales (genus versus species). For each dataset, we constructed a meta-pangenome by predicting and annotating coding sequences, clustering orthologous groups and compiling a genome-by-orthogroup presence-absence matrix, and finally the core genome alignment was used to infer phylogeny.

The evolution of gene content was modeled on the phylogenetic tree using a two-state CTMC (gene presence/absence). State transitions along phylogenetic branches are described by a rate matrix $Q = \begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix}$, and the transition probabilities along a branch of length t are given by

$P(t) = e^{Qt}$ [14, 26, 27]. Gene gain (λ) and loss (μ) parameters were estimated independently for each orthologous gene group by maximum likelihood (ML) using the Felsenstein pruning algorithm [13], and parameter optimization was performed with a constrained quasi-Newton optimizer (L-BFGS-B) as implemented in the *stats* package in R [28].

PGGL (Pangenome Gene Gain and Loss) and PGGS (Pangenome Gene Selection) algorithms, based on CTMC-methods, downstream statistics, and visualizations were implemented in R programming language [28]; upstream assembly, annotation, orthogrouping, alignment, and phylogeny steps were executed with established command-line bioinformatics software tools [12]. The meta-pangenome reconstruction analyses were ran in versioned, containerized environments on local and HPC systems to ensure reproducibility and scalability [29].

The scientific novelty of the research results. This thesis introduces a new software framework that enable to advance microbial evolutionary genomics by introducing a phylogeny-based meta-pangenome approach that reconstructs gene repertoires directly from metagenomic assemblies and infers gene-content evolution without relying on complete isolate genomes or fixed species boundaries. The approach is tailored for the realities of environmental complex microbiomes, including fragmented assemblies, strain mixtures, HGT and recombination.

Methodologically, this work contributes two integrated components by introducing and implementing PGGL (Pangenome Gene Gain and Loss) software, a maximum-likelihood continuous-time Markov chain (ML-CTMC) framework applied to a fixed inferred core-genome phylogeny that models each orthologous group as a binary character (absent/present). The software returns gene-wise estimates of acquisition and deletion rates, marginal ancestral state probabilities, and expected branch-specific event counts computed via Felsenstein's pruning algorithm, thereby enabling lineage-level quantification of gene turnover across the pangenome. Building on that, PGGS (Pangenome Gene Selection) software introduces a phylogeny-based test for directional asymmetry in gene turnover. For each gene, an equal-rates (ER) model, which constrains the gain and loss rates to be the same, is contrasted with an all-rates-different (ARD) model, which allows the gain and loss rates to differ. Model support evaluated using Akaike's Information Criterion (AIC), partitions genes into gain-biased, loss-biased, or symmetric classes, offering an interpretable signal of selection pressure acting on gene presence–absence.

Conceptually, the novelty is to bring comparative-phylogenetic logic to gene-content or function traits at meta-pangenome scale, delivering robust, lineage estimates of turnover and directional bias from empirical metagenomic assembled genomes (MAGs) and isolates alone. In practice, the software is modular and reproducible, integrates orthologous genes, curated ARGs and virulence factors layers within a unified phylogenetic context, and is designed for environmental

genomic surveillance, where cross-dataset and cross-location comparability is essential.

The scientific problem solved. This thesis addresses four fundamental limitations in microbial evolutionary genomics: (1) the absence of validated methods for reconstructing microbial pangenomes directly from metagenomic data, which is resolved through the development of a robust software framework for meta-pangenome reconstruction and analysis; (2) the distortion of phylogenetic signal caused by recombination and horizontal gene flow in microbial genomes, addressed by inferring phylogenies depleted of recombination signal from sequence alignment data; (3) the lack of validated bioinformatics methods for quantifying gene gain–loss dynamics at the pangenome level, overcome by developing and implementing software for estimating gene gain–loss counts and rates across phylogenies; and (4) the absence of methodological frameworks for estimating gene-level selection signals in pangenomes, addressed by constructing R-based software for classifying genes according to their inferred selective regime from pangenomic data. These challenges are addressed through the development of a phylogeny-based meta-pangenome software framework that reconstructs gene repertoires directly from metagenomic assemblies and quantifies the evolutionary dynamics of genome content. Specifically, gene presence–absence matrices derived from metagenomic data are analyzed on a recombination-free core-genome phylogeny, where two-state continuous-time Markov models are fitted independently to each orthologous group to estimate lineage-specific rates of gene gain (λ) and loss (μ). Directional selection on gene presence–absence is assessed by contrasting an equal-rates (ER) model, in which gain and loss are symmetric, with an all-rates-different (ARD) model that allows asymmetric turnover; support for the ARD model indicates preferential acquisition or deletion, whereas the ER model is consistent with symmetric or neutral dynamics. Together, these components yield robust, comparable estimates of pangenome structure and dynamics from metagenomic data, enabling lineage phyletic patterns evolutionary inference and cross-dataset comparison in environmental microbial surveillance.

Theoretical significance of the research. This thesis advances comparative genomic theory for microbial pangenomes by developing bioinformatics software that enables gene-content evolution to be estimated directly from metagenomic next-generation sequencing data. We introduce a discrete-gene likelihood models to orthogroup presence–absence data on recombination-free phylogenies and introduce a framework for testing selection pressure on gene content. Formally, the theoretical contributions of this work are as follows:

- Extends metagenomic next generation sequencing (NGS) analysis to pangenomics by reconstructing meta-pangenomes from MAG-derived gene presence–absence matrices and analyzing them on recombination-free core phylogenies, enabling lineage evolutionary

comparisons.

- Models orthogroup presence/absence as a binary continuous-time Markov process on a fixed core phylogeny, enabling gene-wise estimates of acquisition (gain) and deletion (loss) from presence–absence matrices.
- Uses likelihood-based ancestral reconstruction to obtain marginal ancestral state probabilities and branch-specific expected counts of gene gains and losses.
- Detects directional gene-content evolution by explicitly comparing symmetric and asymmetric gain–loss models, where support for unequal gain and loss rates indicates preferential acquisition or deletion along lineages, thereby enabling gene classification according to selective regime.

The applicative value of the thesis. In practical terms, this thesis delivers a bioinformatics software toolkit for metagenomic surveillance and monitoring, translating evolutionary modeling into outputs that are directly usable by public-health, environmental, and research teams. The software toolkit is designed to function on routine metagenomic NGS data and to produce standardized, interpretable summaries that support decision-making in analyzing complex microbial ecosystems. Its applicative value is reflected in the following operational capabilities and deliverables:

- The bioinformatics software enables the reconstruction of meta-pangenomes directly from metagenomic assemblies, allowing gene repertoire structure and variability to be characterized in environments where isolate genomes are unavailable or incomplete, such as wastewater, air, and built environments.
- By estimating gene gain and loss along phylogenetic lineages, the bioinformatics software toolkit provides quantitative measures of genome turnover that enable the identification of rapidly evolving lineages, assessment of adaptive potential, and prioritization of targets for detailed investigation or intervention.
- Bioinformatics software outputs in the form of quantitative summaries, such as gain and loss rates, directional turnover bias, and branch-specific event counts, are projected onto a phylogeny, yielding report-ready, lineage-based visualizations suitable for early-warning systems, hotspot detection, and longitudinal monitoring across sampling campaigns.
- The software supports the joint analysis of curated antimicrobial resistance, virulence, and mobile genetic element annotations within the same evolutionary context, enabling coordinated surveillance of traits with direct relevance to public health and environmental risk assessment.

- Standardized inference logic and reproducible software workflows allow results to be compared across sites, time points, and projects, facilitating coordinated surveillance efforts at institutional, regional, or national scales.
- Although motivated by urban genomic surveillance, the framework is transferable to other domains, including clinical microbiology, agriculture, aquaculture, marine systems, and natural ecosystems, without methodological redesign, supporting evolution-based analysis wherever metagenomic data are available.
- By identifying lineages and gene families exhibiting unusual gain–loss dynamics or directional bias, the software framework provides a principled basis for generating testable hypotheses that can be followed up by targeted sequencing, functional assays, or epidemiological investigation.

Scientific theses submitted for defense:

- Species-resolved meta-pangenomes can be reconstructed directly from quality-controlled environmental MAGs, with optional co-analysis of isolates, yielding gene-family presence–absence matrices, openness statistics, and recombination-free core-genome phylogenies suitable for downstream inference.
- Gene/orthogroup presence–absence derived from these meta-pangenomes supports phylogeny-based continuous-time Markov models that estimate branch-specific gain (λ) and loss (μ) and provide lineage-resolved rate indicators.
- Directional selection on gene content can be tested by contrasting symmetric (equal-rates) versus asymmetric (all-rates-different) CTMC parameterizations at the gene/orthogroup level, producing interpretable statistics that quantify bias toward acquisition or deletion.
- Application to complex environmental datasets yields an integrated set of selection-pressure indicators (λ, μ, selection indices, prioritized gene sequences) that is reproducible under a fixed analysis pipeline and directly consumable by One Health comparative and early-warning workflows.

Implementation of scientific results. This work was translated from methodological into practice through collaborations with National Agency for Public Health (ANSP), the Institute of Microbiology and Biotechnology from Technical University of Moldova, and the Ștefan cel Mare University of Suceava (Romania).

Approval of scientific results. The core results of the doctoral thesis were presented and discussed at the meetings and seminars of the Department of Software Engineering and Automatics, Faculty of Computers, Informatics and Microelectronics, Technical University of Moldova (2022-

2025). They were reported, discussed, positively evaluated at nine international and national scientific conferences, including, International Conference on Nanotechnologies and Biomedical Engineering (Chișinău, 2025); International Conference BioGENext: Next Generation Therapy Conference (Kyiv, 2024); International conference on Electronics, Communications and Computing (Chișinău, 2022, 2024); Technical and Scientific Conference for Undergraduate, Master's and Doctoral Students (TUM, Chișinău 2023).

Publications on the topic of the thesis. The main results of the thesis were published in 16 scientific papers, including 7 articles in ISI- and SCOPUS-indexed international journals, among them publications in *Nature Reviews Methods Primers* (IF = 56), *Nature Water* (IF = 24.24), and *Genome Biology* (IF = 9.4), as well as 9 papers presented and published in the proceedings of national and international conferences (the full list of publications is provided at the end of the thesis and in the Ph.D. summary). The total number of publications is 49 scientific papers, including 11 ISI and SCOPUS. The author has an h-index of 8 (SCI Hirsch index), and the total number of international citations exceeds 290.

The volume and structure of the thesis. The thesis comprises 116 pages and includes an introduction, 4 chapters, conclusions and recommendations, a bibliography from 348 sources, 5 annexes, 47 figures, and 11 tables.

Keywords: bioinformatics, biostatistics, mathematical modelling, continuous-time Markov model, metagenomics, pangenome, computational biology, comparative genomics.

THESIS CONTENT

The ***Introduction*** justifies the relevance and timeliness of the topic, presents a critical review of current research and technologies, states the thesis aim and objectives, and articulates the scientific novelty and the main theses advanced for defense. It also documents the validation of the results in peer-reviewed publications and lists the conferences where the core findings were presented.

In ***Chapter 1***, we describe the urban microbiome context, articulate the limitations of isolate-only analyses, and motivate meta-pangenomes as a necessary complement for recovering accessory diversity that drives adaptation and public health risk. We define core versus accessory structure, pangenome openness, and the rationale for phylogeny-based inference, thereby establishing the conceptual framework for the subsequent development of the software.

In ***Chapter 2***, we describe the data sources and present the development of a bioinformatics software framework that performs gene calling, ortholog clustering, functional annotation,

recombination-controlled core-genome phylogeny (Figure 1B), and pangenome statistical analyses for both MAGs and isolate genomes (Figure 1A).

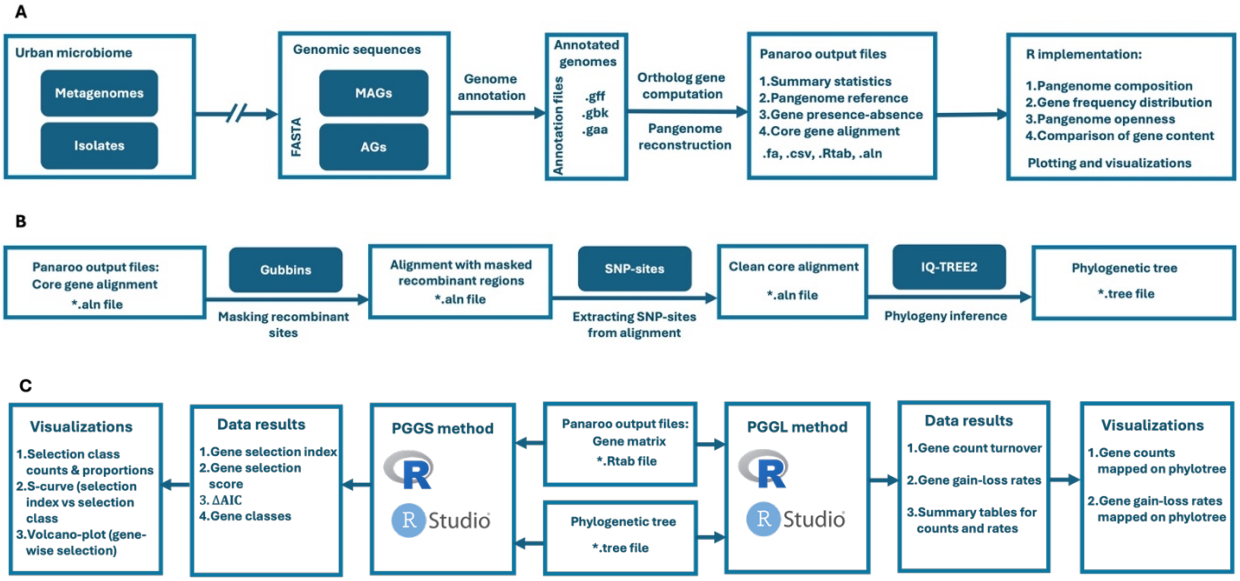

Fig.1. Design of the bioinformatics software framework for meta-pangenome reconstruction and phylogeny-based inference of gene-content evolution. (A) Genome annotation and meta-pangenome reconstruction; (B) Core-genome phylogeny inference; (C) Gene turnover quantification (PGGL method) and gene classification based on selection indices and calibrated effect scores (PGGS method).

We introduce and implement, as software, two modeling methods: (1) PGGL, which maps gene gains and losses onto the phylogeny and summarizes branch rates and genome burdens; and (2) PGGS, which quantifies selection from asymmetry between gain and loss rates, estimates calibrated effect sizes, and classifies genes into directional selection categories (Figure 1C).

Chapter 3 presents the research results obtained using the reproducible bioinformatics software developed for reconstructing and analyzing meta-pangenomes from both isolate assemblies and metagenome-assembled genomes (MAGs). The workflow integrates gene prediction and functional annotation, ortholog clustering, gene-frequency summarization, rarefaction modeling, and pangenome structure visualization within a unified analytical protocol applied consistently to MAGs and isolates. Harmonized inputs and outputs enable direct, controlled comparisons across ecological settings and taxonomic scales.

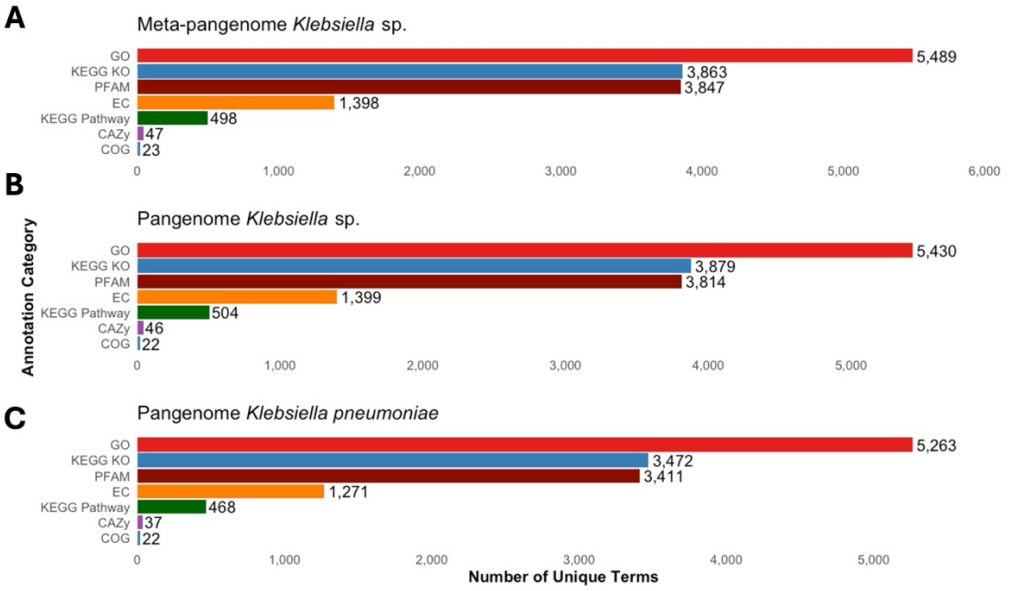

Fig.2. Functional annotation summary across *Klebsiella* datasets. (A) Meta-pangenome (MPKG dataset) captured from MAGs. (B) Isolate pangenome (PKG dataset). (C) *K. pneumoniae* (PKP dataset) from isolate samples.

The structural annotation results reveal systematic differences between data sources. Isolate genomes consistently carry higher and less variable coding sequence counts, near complete tRNA inventories, larger and more stable genome sizes, and far fewer contigs than MAGs. These patterns reflect known limitations of short-read metagenomics, including fragmentation, uneven coverage, binning errors, and community quality criteria for MAGs [16, 17, 30, 31]. Comparative analyses of gene content recover consistent core–accessory structure and frequency patterns across datasets, with isolate assemblies displaying tighter internal coherence, whereas MAGs exhibit greater dispersion reflecting heterogeneous sampling depth and variable genome completeness [16, 17, 30, 31].

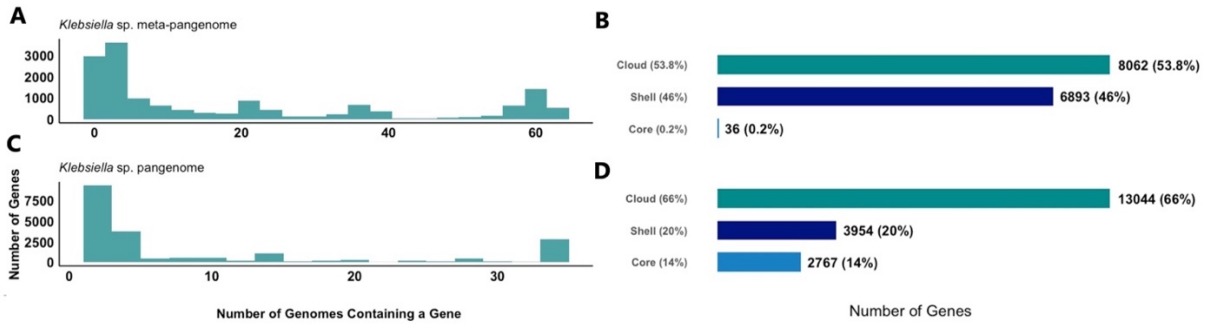

Fig.3. Comparative gene frequency distribution (A, C) and gene composition (core, shell and cloud) in meta-pangnome (A, B) and pangenome (C, D).

Functional profiling against curated resources (GO, KEGG, PFAM, EC, COG, CAZy) captures conserved and niche-specific functions (Figure 2), with high-quality isolate genomes showing narrow dispersion of ontology terms and protein domains consistent with uniform annotation and near-complete assemblies (Figure 2B–C) [32, 33].

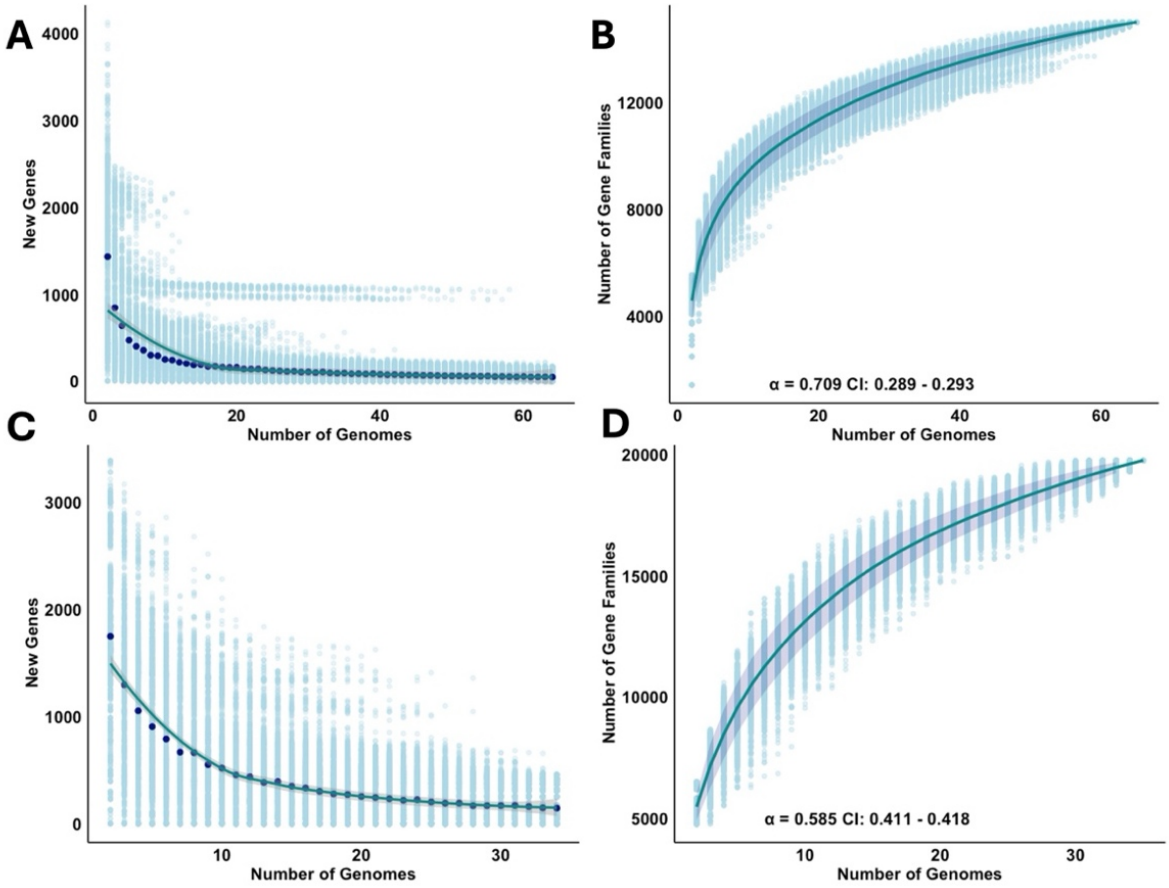

Fig.4. Gene discovery dynamics (A, C) and cumulative number of unique gene families fitted to Heap's law model (B, D) in meta-pangenome (A, B) and pangenome (C, D).

MAG-derived meta-pangenomes recover broader functional repertoires, particularly in enzymatic classes and pathway reconstructions, reflecting both ecological breadth and the ability of metagenomics to sample uncultured diversity present in urban environments (Figure 2A) [16, 17]. A focused *K. pneumoniae* isolate cohort achieves the highest annotation fidelity and the lowest within-species variance, while preserving extensive coverage of enzyme classes, ontology terms, and conserved domains. Collectively, these results support a hybrid approach, where MAGs expand functional breadth by accessing unsampled lineages, whereas isolates raise confidence in per-gene calls and minimize artefacts introduced by fragmentation.

Ortholog clustering and gene-frequency stratification recover the expected core–shell–cloud architecture of bacterial pangenomes in all datasets (Figure 3) [34, 35]. The MAG-derived meta-

pangenome has the smallest core and the largest shell, indicative of many low-frequency genes typical of mixed and environmentally diverse data (Figure 3A-B). The single-species *K. pneumoniae* set has the largest core and cloud, in line with tighter phylogenetic divergence and more uniform clinical sampling, with the multi-species isolate data occupying an intermediate position (Figure 3C-D). Gene-frequency histograms show the characteristic asymmetric U-shape distribution, with most gene families either widespread or rare and relatively few at intermediate prevalence, as described in classical pangenome theory (Figure 3B, D)) [7, 34, 35].

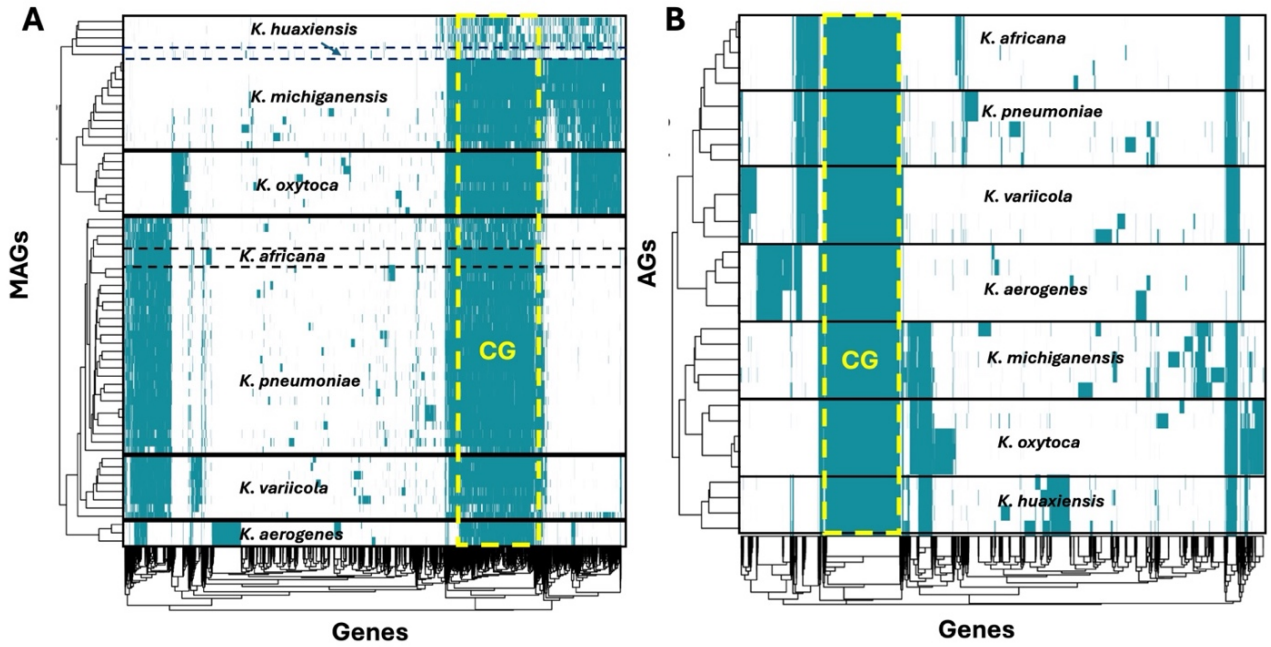

Fig.5. Comparative hierarchical clustering of orthologous genes and genomes in meta-pangenome (A) and pangenome (B). Note: CG-Core Genome, MAGs-Metagenomic Assembled Genomes, AGs-Assembled Genomes.

Gene discovery dynamics were quantified through rarefaction and power-law modelling (Figure 4) [7, 36, 37]. The isolate-based pangenome remains strongly open across sampling permutations, with sustained discovery of novel gene families (Figure 4C-D). The multi-species MAG-based meta-pangenome remains open but grows more slowly (Figure 4A-B).

Presence–absence heatmaps reveal dense, near-universal cores layered with accessory islands (Figure 5). In the MAG meta-pangenome, accessory modules are patchy and clade-restricted, with islands. Columns with unusually sparse signal correspond to MAGs with small assembly sizes and very low N50, a pattern typical of incompletely recovered bins rather than true biological absence (Figure 5A) [16, 17, 30, 31]. In the multi-species isolate data, species partition

cleanly and recapitulate known taxonomy, the KpSC complex (*K. pneumoniae*, *K. variicola*, *K. africana*) groups together and the *K. oxytoca* complex (*K. oxytoca*, *K. michiganensis*) forms a distinct block, additionally *K. huaxiensis* sits adjacent but remains separate, and *K. aerogenes* falls outside these complexes with a differentiated accessory repertoire (Figure 5B). Principal component analysis (PCA) on binary gene content shows broad dispersion for MAGs, tracking environmental and compositional heterogeneity, while isolate-based species form compact, well-separated clusters that align with taxonomy. Within *K. pneumoniae*, major sequence types occupy distinct regions, revealing lineage-specific accessory repertoires that agree with population-genomic analyses of the KpSC [38, 39].

The primary contribution of this chapter lies in development and validation of pangenome reconstruction software workflow. Applied on three datasets showed that: (1) the MAG-based meta-pangenome maximizes ecological breadth and accessory diversity at the cost of higher structural variance; (2) the multi-species isolate cohort offers genus-level resolution with clear species structure and reduced noise; and (3) the single-species clinical cohort resolves clonal-lineage architecture with near-universal cores and discrete, ST-specific modules. Sampling strategy, assembly quality, and phylogenetic scope therefore jointly shape apparent pangenome size, its openness, and the visibility of accessory pathways. These results provide a robust foundation for the evolutionary modelling developed later in the thesis, including probabilistic inference of gain–loss histories and tests of directional turnover.

In ***Chapter 4*** we present two complementary methods implemented as bioinformatics tools. The PGGL (Pangenome Gene Gain–Loss) method, implemented as R software, treats each orthologous group as a two-state character on a fixed species tree and infers where gains and losses occurred while estimating the rate of gene turnover as the expected number of transitions per unit branch length, using CTMC-ML and pruning to obtain node posteriors, edge-level events, and specific branch-length rate estimates (Algorithm 1 and 2).

The PGGS (Pangenome Gene Selection) method, implemented as R software, then evaluates whether individual genes follow direction-free turnover or directional dynamics, summarizing magnitude and sign via ratio-based effect sizes, with model support adjudicated by Akaike's Information Criterion (Algorithm 1 and 3) [40, 41]. Applied to ecologically broad urban metagenomic data, and to cultured/clinical isolate collection of genomes, this method resolves the lineage-structured pattern of gene acquisition and deletion and classify genes into neutral versus gain- or loss-biased regimes consistent with their acquisitional or deletional pressures. Normalized gene gain (λ) and loss (μ) rate distributions reveal a consistent loss-dominated regime across

species, with μ exceeding λ in both isolate-based pangenomes and meta-pangenomes, in line with long-recognized deletional pressure in bacteria (Figure 6) [6, 42].

Algorithm 1. Per-gene gain-loss calling and rate estimation (PGGL-Gene).

Input: A rooted, strictly bifurcating phylogeny with branch lengths; a binary presence–absence vector for gene $\boldsymbol{g}$ aligned to the tree tips (with missing values allowed); a gain–loss model (ER or ARD); a root prior; and an event-calling threshold $\boldsymbol{\delta}$.
Output: For each branch of the tree, inferred gene gain and loss events together with ML estimates of the gain ($\hat{\boldsymbol{\lambda}}$) and loss ($\hat{\boldsymbol{\mu}}$) rates for gene $\boldsymbol{g}$.

Procedure:

1. **Model specification:** Model the evolution of gene presence–absence as a two-state CTMC process on the phylogeny, with transitions corresponding to gene acquisition ($0 \rightarrow 1$, rate $\boldsymbol{\lambda}$) and gene deletion ($1 \rightarrow 0$, rate $\boldsymbol{\mu}$).
2. **Likelihood computation and rate estimation:** Encode observed tip states as state likelihood vectors and compute the likelihood of the data using Felsenstein's pruning algorithm. Estimate gene-specific gain and loss rates ($\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\mu}}$) by ML under either an equal-rates (ER) or all-rates-different (ARD) parameterization.
3. **Ancestral state reconstruction:** Using the fitted rates, compute marginal posterior probabilities of gene presence at all internal nodes via standard upward–downward message passing on the tree.
4. **Even calling on branches:** For each branch, compare posterior probabilities of gene presence between parent and child nodes. A gene gain or loss event is recorded when the posterior change exceeds the predefined threshold $\boldsymbol{\delta}$.
5. **Reporting:** Output, for each branch, the inferred gain and loss events together with the corresponding gene-level rate estimates.

When applied to meta-pangenome data, the PGGL software displays broader, heavy-tailed rate distributions, indicating heterogeneous turnover in which most lineages evolve slowly while a minority exhibit elevated loss rates together with localized, lineage-specific increases in gain rates, consistent with episodic ecological opportunity and variable access to mobile gene pools (Figure 6B). In contrast, isolate datasets show tighter distributions and the strongest loss bias, reflecting higher assembly contiguity and more complete recovery of deletional tracts (Figure 6A) [17, 31].

Algorithm 2. Gene Gain-Loss (PGGL) pangenome event calling.

Input: A rooted, strictly bifurcating phylogeny with branch lengths; a binary presence–absence matrix $\boldsymbol{P}$ (genomes × genes) aligned to the tree tips; a gain–loss model (ER or ARD); a root prior; and an event-calling threshold $\boldsymbol{\delta}$.
Ouput: For each gene and each branch, inferred gene gain and loss events together with gene-specific maximum-likelihood estimates of gain and loss rates ($\hat{\boldsymbol{\lambda}}_{\boldsymbol{g}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{g}}$)**.** Additionally, genome-level summaries of gene gain and loss (e.g. root-to-tip counts and rates) obtained by aggregating events across genes along each genome's phylogenetic path.

Procedure:

1. **Matrix–tree alignment:** Ensure that rows of $\boldsymbol{P}$ (genomes) correspond to the tip order of the phylogeny.
2. **Per-gene inference:** Treat each column of $\boldsymbol{P}$ as a gene-specific presence-absence vector $\boldsymbol{x_g}$**.** Excluding invariant genes, apply **Algorithm 1 (PGGL-Gene)** to infer branch level gain and loss events and estimate ($\hat{\boldsymbol{\lambda}}_{\boldsymbol{g}}, \hat{\boldsymbol{\mu}}_{\boldsymbol{g}}$)**.**
3. **Pangenome aggregation:** Combine all gene-level branch annotations into a single pangenome-wide event table.
4. **Genome-level summarization:** For each genome (tip), aggregate inferred events across all genes along its root-to-tip path to compute genome-specific gain and loss counts and rates.
5. **Reporting:** Return: gene-, branch-, and genome-level summaries of pangenome dynamics

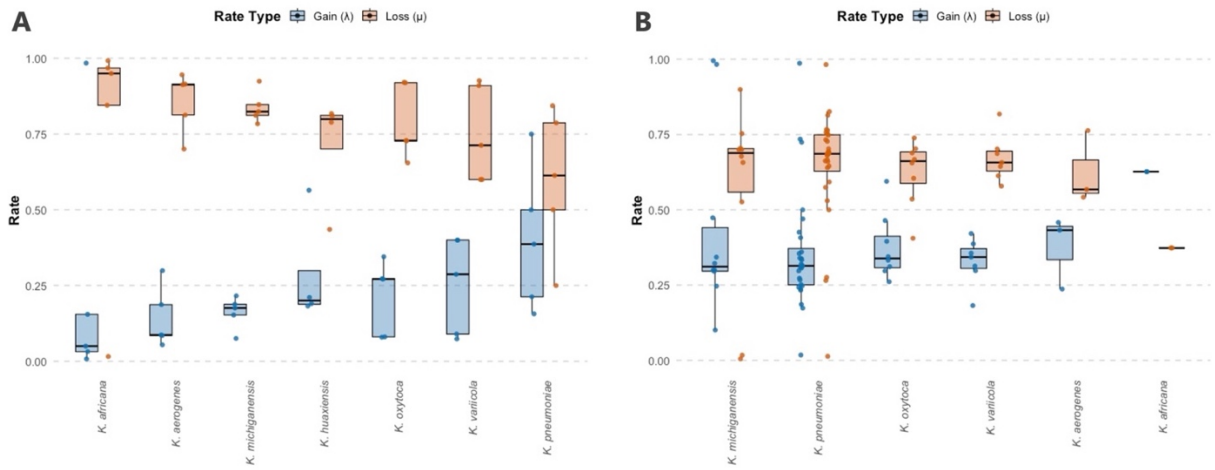

Fig.6. Normalized summary distribution of gene gain and loss rates per species in (A) pangeome and (B) meta-pagenome

Species-level summaries of absolute gene gain and loss events emphasize the predominance of losses over gains across cohorts, with substantial between-species variation in total event burden (Figure 7). While losses account for the majority of events in both meta-pangenome and pangenome datasets, repeated gains are concentrated in a subset of lineages, consistent with acquisition in accessory gene pools linked to mobility, defense, niche-tuned metabolism, and antimicrobial resistance, in agreement with prior functional analyses of horizontally transferred modules (Figure 7) [43–45].

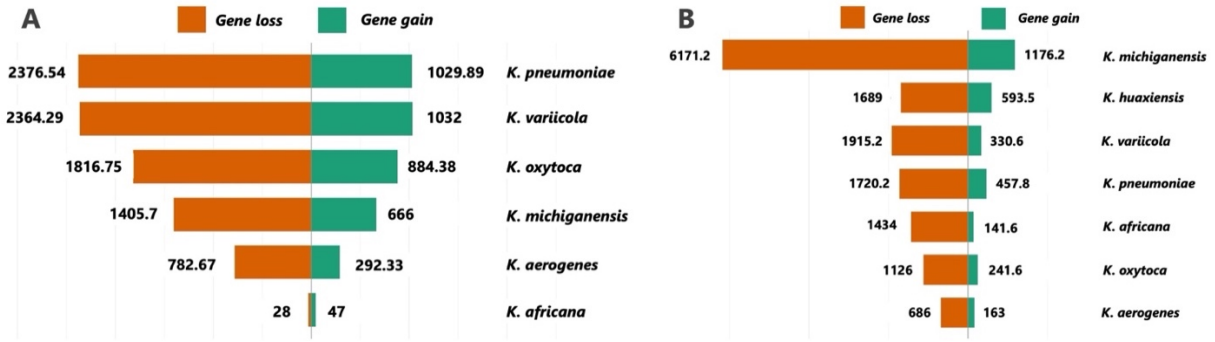

Fig.7. Species level gene gain and loss in (A) meta-pagenome and (B) pagenome.

The PGGS software operates downstream of PGGL by using the per-gene maximum-likelihood rates to adjudicate between symmetric and directional turnover. For each gene, PGGS compares an equal-rates regime (ER, $\lambda = \mu$), consistent with direction-free turnover under nearly neutral or fluctuating selection, with an all-rates-different regime (ARD, $\lambda \neq \mu$) that captures directional processes attributable to long-term selection or process asymmetries such as deletional bias and uneven horizontal-gene-transfer supply (Algorithm 1 and 3) [6, 42, 45–47]. Model support

is evaluated with Akaike's Information Criterion (AIC score) [40, 41]. Direction and magnitude are summarized by the selection index and the selection score metrics, which are ratio-based and thus portable across datasets with different branch-length calibrations. Class labels (NS/WS/MS/SS) provide a discrete, analysis-ready stratification for ranking.

Algorithm 3. PGGS (Pangenome Gene Selection) algorithm.

Input: A rooted, strictly bifurcating phylogeny with branch lengths; a binary presence–absence matrix ***P*** (genomes × genes) aligned to the tree tips; a frequency window defining informative genes; a gain–loss model (ER and ARD); a root prior; and an event threshold.
Output: For each informative gene, model-specific gain and loss rate estimates under ER and ARD, model-comparison statistics (log-likelihoods, **AIC**, **ΔAIC**), a gene-selection score, and an assigned selection class.

Procedure:

1. **Gene filtering:** Restrict analysis to informative genes by excluding invariant or extremely rare/common genes based on their presence frequency across genomes.
2. **Model fitting:** For each retained gene, apply **Algorithm 1 (PGGL-Gene)** twice, once under the ER model and once under the ARD model, to estimate gain and loss rates and compute corresponding likelihoods.
3. **Model comparison:** Quantify rate asymmetry by comparing ER and ARD fits using information criteria (**AIC and ΔAIC**).
4. **Selection scoring:** Derive a gene-level selection index from the relative magnitude and asymmetry of inferred gain and loss rates.
5. **Classification:** Assign each gene to a discrete selection class (neutral, weakly selected, moderately selected, strongly selected) based on its selection score and model support.
6. **Reporting:** Return a gene-wise table summarizing rate estimates, model-comparison statistics, selection scores, and selection classes.

Applied to metagenomic and isolate datasets, the PGGS software recovers directional asymmetry of gene turnover in environmentally diverse, genus-wide meta-pangenome data, characterized by a predominance of loss-biased genes and a smaller but non-trivial fraction of gain-biased genes (Figure 8A). This directional signal is attenuated in genus-wide isolate pangenomes and approaches effective symmetry in species-restricted isolate panels, consistent with the expectation that increased ecological breadth and heterogeneity in gene supply amplify directional gene-content turnover (Figure 8B) [36, 48, 49]. The classifications are stable to the choice of root prior (uniform versus stationary) and to modest perturbations of branch lengths and optimizer settings, with ΔAIC orderings for strongly asymmetric genes reproducible across runs. Together, the PGGL and PGGS tools developed in this thesis deliver a compact, interpretable atlas of pangenome gene-content dynamics that integrates where changes occur (event maps and rates) with how individual genes deviate from symmetry (model-supported direction and effect size).

Unlike frequency-only heuristics such as core/accessory genome partitions, which quantify variability but cannot separate localized gains from losses, this approach is explicitly phylogeny-based and links turnover to the evolutionary scaffolding on which it accrued. The resulting outputs are directly usable for figure generation and downstream functional enrichment analyses and are

generalizable across data types, thereby providing a consistent and reproducible basis for cross-cohort inference in meta-pangenome studies.

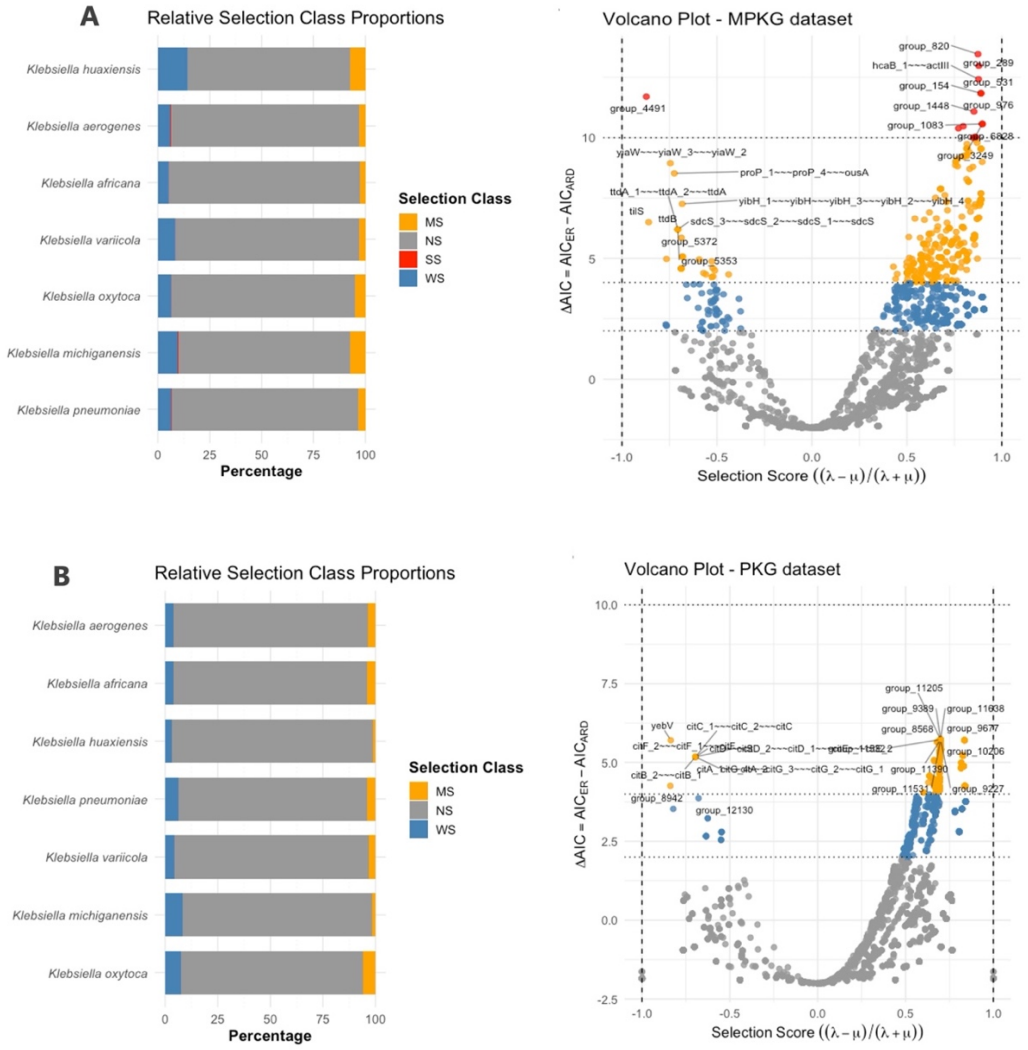

Fig.8. Selection class counts and volcano plots for selection score for genes in (A) meta-pangenome and (B) pangenome.

To evaluate the reconstruction accuracy of the developed PGGL and PGGS software, we performed a controlled benchmarking analysis on simulated phylogenies with fully specified evolutionary histories. Trees were generated under a Yule (pure-birth) diversification process [50], and binary gene presence–absence evolution was simulated along branches using a continuous-time Markov chain (CTMC) with asymmetric gain and loss rates. This design allows direct comparison between inferred and true ancestral states. Fitch parsimony [51], Bayesian stochastic character mapping (SCM) [52], and the maximum likelihood (ML) framework implemented in this thesis as the inferential core of PGGL and PGGS were compared.

Table 1. Comparative performance metrics for ancestral state reconstruction methods

Method	Accuracy	Precision	Recall	Specificity	Ambiguous nodes
Maximum Likelihood	0.963	0.966	0.973	0.945	0
Bayesian SCM	0.963	0.964	0.975	0.943	0
Fitch Parsimony	0.944	0.964	0.944	0.943	392

The ML implementation achieves complete node coverage and an effective accuracy of approximately 96%, closely matching Bayesian SCM while maintaining balanced error rates (Table 1). In contrast, Fitch parsimony produces a substantial proportion of ambiguous internal nodes, leading to markedly reduced effective accuracy once unresolved states are incorporated into the evaluation (Table 1). These results demonstrate that the ML framework underlying PGGL and PGGS provides accurate and computationally efficient ancestral state inference, comparable to Bayesian SCM but substantially more scalable for large phylogenies.

Each chapter of the thesis ends with conclusions and a summary of the main results obtained. **The final conclusions and recommendations** summarize the main results published in peer-reviewed journals and justify the theoretical and practical value of the developed methods for metagenomic data analysis.

GENERAL CONCLUSIONS AND RECOMMENDATIONS

This thesis demonstrates that metagenomics and pangenomics can be fused into a single analytical framework that resolves how gene content is organized and changes in complex environments. The key advance is to treat gene presence-absence as an evolutionary trait on a fixed phylogeny, allowing event localization, rate estimation, and directionality tests that are portable across assembly types and taxonomic scopes. Accordingly, the main conclusions are:

1. A robust and fully reproducible meta-pangenome reconstruction and analysis bioinformatics software framework was developed to integrate isolate genomes and heterogeneous metagenomic data into a unified gene-content representation, allowing comparative and evolutionary analysis of genomic diversity based on gene presence–absence and functional annotation across mixed-quality datasets.
2. A method for inferring a recombination-filtered maximum-likelihood species phylogeny was developed and integrated into the analysis framework, providing the explicit evolutionary structure required for probabilistic modeling of gene presence–absence evolution and for

likelihood-based inference of gene gain and loss processes across heterogeneous genome collections.

3. A phylogeny-based maximum-likelihood gene gain–loss inference algorithm (PGGL; Pangenome Gene Gain Loss) was developed and implemented as R software to model the evolution of inferred gene presence–absence states of orthologous groups as a continuous-time Markov process on a species phylogeny, enabling quantitative estimation of branch-, lineage-, and clade-specific gene gain and loss rates and event counts.
4. A rate-based gene selection inference and corresponding classification algorithm, with an R software implementation (PGGS; Pangenome Gene Selection), was developed for pangenome analyses, based on the explicit comparison of gene gain and loss rates under symmetric and asymmetric evolutionary models to identify statistically supported gene- and lineage-specific gain- or loss-biased regimes.
5. The meta-pangenome reconstruction framework and the PGGL and PGGS inference algorithms were empirically validated on isolate and metagenome-assembled genome datasets across multiple taxonomic scales, yielding consistent gene presence–absence representations, stable gene gain–loss inference, and gene classification from species-level to higher phylogenetic resolutions.
6. The developed method based on maximum-likelihood inference under a continuous-time Markov framework was benchmarked at the level of ancestral gene-state reconstruction against Bayesian stochastic mapping and Fitch parsimony, showing reliable reconstruction accuracy with reduced ambiguity and improved computational efficiency in genome-scale analyses.

To translate these findings into durable practice, metagenomic pangenomics should be operated as a tiered, FAIR, and quality-aware system that couples upstream assemblies to downstream evolutionary inference and public-health reporting. The items below prioritize actions that raise fidelity, portability, and decision value:

1. Analyses should be conducted concurrently across environmental, isolate, and clinical data layers, using uniform analytical thresholds and a shared data model, so that inferred signals are directly comparable across distinct contexts and over longitudinal series.
2. For high-information-value datasets, particularly metagenome-assembled genomes containing repetitive regions, the use of long-read or hybrid assemblies is preferable. Complete recovery of rRNA and tRNA operons, as well as repeat-rich mobile regions, reduces artifactual gene absences caused by assembly fragmentation, enables more precise delineation of accessory islands, and improves localization of gain–loss events and prophage boundaries.

3. Event-mapping and gene-flux directionality tools should be implemented as R packages, tested across diverse datasets, with stable interfaces, example datasets, reproducible documentation, and dedicated plotting functions.
4. Systematic assessment of inferential robustness is recommended via sensitivity analyses to root choice, branch-length scaling, and optimization settings, particularly for genes exhibiting strong directional asymmetry. Such evaluation strengthens the biological interpretation of selection metrics and reduces the risk of conclusions driven by technical parameterization.
5. Integrating gain–loss inferences with standardized functional annotations (e.g., COG, KEGG, PFAM) is essential for biological interpretation of observed patterns and for conducting comparable functional enrichment analyses across cohorts and ecological contexts.

Bibliography

1. Koonin, Eugene V., and Yuri I. Wolf. 2008. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* 36: 6688–6719. https://doi.org/10.1093/nar/gkn668.
2. Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405. Nature Publishing Group: 299–304. https://doi.org/10.1038/35012500.
3. Vos, Michiel, Matthijn C. Hesselman, Tim A. te Beek, Mark W. J. van Passel, and Adam Eyre-Walker. 2015. Rates of Lateral Gene Transfer in Prokaryotes: High but Why? *Trends in Microbiology* 23: 598–605. https://doi.org/10.1016/j.tim.2015.07.006.
4. Treangen, Todd J., and Eduardo P. C. Rocha. 2011. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLOS Genetics* 7. Public Library of Science: e1001284. https://doi.org/10.1371/journal.pgen.1001284.
5. Medini, Duccio, Claudio Donati, Hervé Tettelin, Vega Masignani, and Rino Rappuoli. 2005. The microbial pan-genome. *Current Opinion in Genetics & Development* 15. Genomes and Evolution: 589–594. https://doi.org/10.1016/j.gde.2005.09.006.
6. McInerney, James O., Alan McNally, and Mary J. O'Connell. 2017. Why prokaryotes have pangenomes. *Nature Microbiology* 2. Nature Publishing Group: 1–5. https://doi.org/10.1038/nmicrobiol.2017.40.
7. Vernikos, George, Duccio Medini, David R Riley, and Hervé Tettelin. 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology* 23. Host–Microbe Interactions: Bacteria • Genomics: 148–154. https://doi.org/10.1016/j.mib.2014.11.016.
8. Ma, Bing, Michael France, and Jacques Ravel. 2020. Meta-Pangenome: At the Crossroad of Pangenomics and Metagenomics. In *The Pangenome: Diversity, Dynamics and Evolution of Genomes*, ed. Hervé Tettelin and Duccio Medini, 205–218. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_9.
9. Danko, David, Daniela Bezdan, Evan E. Afshin, Sofia Ahsanuddin, Chandrima Bhattacharya, Daniel J. Butler, Kern Rei Chng, et al. 2021. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* 184: 3376-3393.e17. https://doi.org/10.1016/j.cell.2021.05.002.
10. Boucher, Yan, Christophe J. Douady, R. Thane Papke, David A. Walsh, Mary Ellen R. Boudreau, Camilla L. Nesbø, Rebecca J. Case, and W. Ford Doolittle. 2003. Lateral Gene Transfer and the Origins of Prokaryotic Groups. *Annual Review of Genetics* 37. Annual Reviews: 283–328. https://doi.org/10.1146/annurev.genet.37.050503.084247.
11. Zolfo, Moreno, Francesco Asnicar, Paolo Manghi, Edoardo Pasolli, Adrian Tett, and Nicola Segata. 2018. Profiling microbial strains in urban environments using metagenomic sequencing data. *Biology Direct* 13: 9. https://doi.org/10.1186/s13062-018-0211-z.
12. Liu, Shaopeng, Judith S. Rodriguez, Viorel Munteanu, Cynthia Ronkowski, Nitesh Kumar Sharma, Mohammed Alser, Francesco Andreace, et al. 2025. Analysis of metagenomic data. *Nature Reviews Methods Primers* 5. Nature Publishing Group: 1–28. https://doi.org/10.1038/s43586-024-00376-6.
13. Felsenstein, Joseph. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17: 368–376. https://doi.org/10.1007/BF01734359.
14. Pagel, Mark. 1999. The Maximum Likelihood Approach to Reconstructing Ancestral Character States of Discrete Characters on Phylogenies. *Systematic Biology* 48. [Oxford University Press, Society of Systematic Biologists]: 612–622.
15. Hendriksen, Rene S., Patrick Munk, Patrick Njage, Bram van Bunnik, Luke McNally, Oksana Lukjancenko, Timo Röder, et al. 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature Communications* 10. Nature Publishing Group: 1124. https://doi.org/10.1038/s41467-019-08853-3.
16. Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology* 35. Nature Publishing Group: 833–844. https://doi.org/10.1038/nbt.3935.
17. Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. Minimum information about a single amplified genome

(MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* 35. Nature Publishing Group: 725–731. https://doi.org/10.1038/nbt.3893.
18. Albertsen, Mads, Philip Hugenholtz, Adam Skarshewski, Kåre L. Nielsen, Gene W. Tyson, and Per H. Nielsen. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology* 31. Nature Publishing Group: 533–538. https://doi.org/10.1038/nbt.2579.
19. Ishikawa, Sohta A, Anna Zhukova, Wataru Iwasaki, and Olivier Gascuel. 2019. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. *Molecular Biology and Evolution* 36: 2069–2085. https://doi.org/10.1093/molbev/msz131.
20. Boussau, Bastien, and Vincent Daubin. 2010. Genomes as documents of evolutionary history. *Trends in Ecology & Evolution* 25: 224–232. https://doi.org/10.1016/j.tree.2009.09.007.
21. Aßmann, Eva, Timo Greiner, Hugues Richard, Matthew Wade, Shelesh Agrawal, Fabian Amman, Sindy Böttcher, et al. 2025. Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. *Nature Water*. https://doi.org/10.1038/s44221-025-00444-5.
22. Gordeev, Victor, Martin Hölzer, Daniel Desirò, Iryna V. Goraichuk, Sergey Knyazev, Helena Solo-Gabriele, Pavel Skums, et al. 2025. Leveraging wastewater sequencing to strengthen global public health surveillance. *BMC Global and Public Health* 3: 23. https://doi.org/10.1186/s44263-025-00138-w.
23. Munteanu, Viorel, Michael Saldana, Nitesh Kumar Sharma, Wenhao O. Ouyang, Eva Aßmann, Victor Gordeev, Nadiia Kasianchuk, et al. 2023. SARS-CoV-2 Wastewater Genomic Surveillance: Approaches, Challenges, and Opportunities. *arXiv.org*. September 23.
24. Cohen, Ofir, Haim Ashkenazy, Frida Belinky, Dorothée Huchon, and Tal Pupko. 2010. GLOOME: gain loss mapping engine. *Bioinformatics* 26: 2914–2915. https://doi.org/10.1093/bioinformatics/btq549.
25. Nei, M, and T Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* 3: 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410.
26. Arnold O. Allen. 1990. Probability, Statistics, and Queueing Theory. *ScienceDirect*.
27. Grigelionis, B. 1963. On the Convergence of Sums of Random Step Processes to a Poisson Process. *Theory of Probability & Its Applications* 8. Society for Industrial and Applied Mathematics: 177–182. https://doi.org/10.1137/1108017.
28. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2024. https://www.r-project.org/. Accessed August 2.
29. Sharma, Nitesh Kumar, Ram Ayyala, Dhrithi Deshpande, Yesha Patel, Viorel Munteanu, Dumitru Ciorba, Viorel Bostan, et al. 2024. Analytical code sharing practices in biomedical research. *PeerJ Computer Science* 10. PeerJ Inc.: e2066. https://doi.org/10.7717/peerj-cs.2066.
30. Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nature Methods* 14. Nature Publishing Group: 1063–1071. https://doi.org/10.1038/nmeth.4458.
31. Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* 25. Cold Spring Harbor Lab: 1043–1055. https://doi.org/10.1101/gr.186072.114.
32. O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 44: D733–D745. https://doi.org/10.1093/nar/gkv1189.
33. Tatusova, Tatiana, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky, and James Ostell. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research* 44: 6614–6624. https://doi.org/10.1093/nar/gkw569.
34. Rocha, Jaqueline, Isabel Henriques, Margarita Gomila, and Célia M. Manaia. 2022. Common and distinctive genomic features of Klebsiella pneumoniae thriving in the natural environment or in clinical

settings. *Scientific Reports* 12. Nature Publishing Group: 10441. https://doi.org/10.1038/s41598-022-14547-6.
35. Cooper, Helena B., Ben Vezina, Jane Hawkey, Virginie Passet, Sebastián López-Fernández, Jonathan M. Monk, Sylvain Brisse, Kathryn E. Holt, and Kelly L. Wyres. 2024. A validated pangenome-scale metabolic model for the Klebsiella pneumoniae species complex. *Microbial Genomics* 10. Microbiology Society,: 001206. https://doi.org/10.1099/mgen.0.001206.
36. Tettelin, Hervé, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, et al. 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome." *Proceedings of the National Academy of Sciences* 102. Proceedings of the National Academy of Sciences: 13950–13955. https://doi.org/10.1073/pnas.0506758102.
37. Snipen, Lars, and Kristian Hovde Liland. 2015. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics* 16: 79. https://doi.org/10.1186/s12859-015-0517-0.
38. Wyres, Kelly L., Margaret M. C. Lam, and Kathryn E. Holt. 2020. Population genomics of Klebsiella pneumoniae. *Nature Reviews Microbiology* 18. Nature Publishing Group: 344–359. https://doi.org/10.1038/s41579-019-0315-1.
39. Lam, Margaret M. C., Ryan R. Wick, Stephen C. Watts, Louise T. Cerdeira, Kelly L. Wyres, and Kathryn E. Holt. 2021. A genomic surveillance framework and genotyping tool for Klebsiella pneumoniae and its related species complex. *Nature Communications* 12. Nature Publishing Group: 4188. https://doi.org/10.1038/s41467-021-24448-3.
40. Akaike, Hirotugu. 1998. A New Look at the Statistical Model Identification. In *Selected Papers of Hirotugu Akaike*, ed. Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa, 215–222. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-1694-0_16.
41. Burnham, Kenneth P., and David R. Anderson, ed. 2004. *Model Selection and Multimodel Inference*. New York, NY: Springer. https://doi.org/10.1007/b97636.
42. Mira, Alex, Howard Ochman, and Nancy A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* 17. Elsevier: 589–596. https://doi.org/10.1016/S0168-9525(01)02447-7.
43. Ochman, Howard, Jeffrey G. Lawrence, and Eduardo A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405. Nature Publishing Group: 299–304. https://doi.org/10.1038/35012500.
44. Frost, Laura S., Raphael Leplae, Anne O. Summers, and Ariane Toussaint. 2005. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology* 3. Nature Publishing Group: 722–732. https://doi.org/10.1038/nrmicro1235.
45. Soucy, Shannon M., Jinling Huang, and Johann Peter Gogarten. 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* 16. Nature Publishing Group: 472–482. https://doi.org/10.1038/nrg3962.
46. Ohta, Tomoko. 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature* 246. Nature Publishing Group: 96–98. https://doi.org/10.1038/246096a0.
47. Bell, Graham. 2010. Fluctuating selection: the perpetual renewal of adaptation in variable environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365. Royal Society: 87–97. https://doi.org/10.1098/rstb.2009.0150.
48. Touchon, Marie, Claire Hoede, Olivier Tenaillon, Valérie Barbe, Simon Baeriswyl, Philippe Bidet, Edouard Bingen, et al. 2009. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics* 5. Public Library of Science: e1000344. https://doi.org/10.1371/journal.pgen.1000344.
49. Rocha, Eduardo P C. 2018. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. *Molecular Biology and Evolution* 35: 1338–1347. https://doi.org/10.1093/molbev/msy078.
50. Felsenstein, Joseph, and Joseph Felsenstein. 2003. *Inferring Phylogenies*. Oxford, New York: Oxford University Press.
51. Fitch, Walter M. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* 20. [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.]: 406–416. https://doi.org/10.2307/2412116.

52. Huelsenbeck, John P., Rasmus Nielsen, and Jonathan P. Bollback. 2003. Stochastic Mapping of Morphological Characters. *Systematic Biology* 52: 131–158. https://doi.org/10.1080/10635150390192780.

LIST OF THE AUTHOR'S PUBLICATIONS ON THE SUBJECT OF THE Ph.D.THESIS

Articles in international journals listed ISI and SCOPUS:

1) LIU, S., RODRIGUEZ, JS., **MUNTEANU, V.**, RONKOWSKI, C., SHARMA, NK., ALSER, M., ANDREACE, F., BLEKHMAN, R., BŁASZCZYK, D., CHIKHI, R., CRANDALL, KA., LIBERA, KD., FRANCIS, D., FROLOVA, A., GANCZ, AS., HUNTLEY, NE., JAISWAL, P., KOSCIOLEK, T., ŁABAJ, PP., ŁABAJ, W., LUAN, T., MASON, C., MOUSTAFA, M., MURALIDHARAN, HS., MUTLU, O., GHIASI, NM., RAHNAVARD, A., SUN, F., TIAN, S., TIERNEY, BT., SYOC, EV., VICEDOMINI, R., ZACKULAR JP., ZELIKOVSKY, A., ZELIŃSKA, K., GANDA, E., DAVERNPORT, ER., POP, M., KOSLICKI, D., MANGUL, S., Analysis of metagenomic data. In: *Nature Reviews Methods Primers*, 2025, vol. 5, pp. 5. ISSN 2662-8449 (Impact Factor 56.0)
2) AßMANN, E., GREINER, T., RICHARD, H., WADE, R., AGRAWAL, S., AMMAN, F., BÖTTCHER, S., LACKNER, S., LANDTHALER, M., MANGUL, S., **MUNTEANU, V.**, PSOMOPOULUS, F., SMITH, M., TROFIMOVA, M., ULLRICH, A., VON KLEIST, M., WYLER, E., HÖLZER, M., IRRGANG, G. Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. In: *Nature Water*, 2025, vol. 3, pp. 753-763. ISSN 2731-6084. (Impact Factor 24.1)
3) HUANG, YN., JAISWAL, PV., RAJES, A., YADAV, A., YU, D., LIU, F., SCHEG, G., SHIH, E., BOLDIREV, G., NAKASHIDZE, I., SARKAR, A., MEHTA, JH, WANG, K., PATEL, KK., MIRZA, MAB., HAPANI, KC., PENG, Q., AYYALA, R., GUO, R., KAPUR, S., RAMESH, T., CIORBĂ, D., **MUNTEANU, V**., BOSTAN, V., DIMIAN, M., ABEDALTHAGAFI, MS., MANGUL, S. The systematic assessment of completeness of public metadata accompanying omics studies in the Gene Expression Omnibus data repository. In: *Genome Biology*, 2025, vol. 26, pp. 274, ISSN 1474-760X. (Impact Factor 9.4)
4) HUANG, Y., **MUNTEANU, V.**, LOVE, MI., RONKOWSKI, CF., DESHAPANDE, D., WONG-BERINGER, A., CORBETT-DETIG, R., DIMIAN, M., MOORE, JH., GARMIRE, LX., REDDY, TBK., BUTTE, AJ., ROBINSON, MD., ESKIN, E., ABEDALTHAGAFI, M., S., MANGUL, S. Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies. In: *Cell Genomics*, 2025, vol. 5/5, pp. 100845. ISSN 2666-979X. (Impact Factor 9.0)
5) SHARMA, NK., AYYALA, R., DESHPANDE, D., PATEL, Y., **MUNTEANU, V.**, CIORBĂ, D., BOSTAN, V., FICUSTEAN, A., VAHED, M., SAKAR, A., GUO, R., MOOR, A., DARCI-MAHER, N., NOGOY, N., ABEDALTHAGAFI, M., MANGUL, S. Analytical code sharing practices in biomedical research. In: *Peer J Computer Science*, 2024, vol. 10, pp. e20666, ISSN 2376-5992. (Impact Factor 3.8)
6) DESHPANDE, D., CHHUNGANI, K., CHANG Y., KARLSBERG, A., LOEFFLER, C., ZHANG, J., MUSZYNSKA, A., **MUNTEANU, V.**, YANG, H., ROTMAN, J., TAO, L., BALLIU, B., TSENG, E., ESKING E., ZHAO, F., MOHAMMADI, P., ŁABAJ, PP., MANGUL, S. RNA-Seq data science: From raw data to effective interpretation. In: *Frontiers in Genetics*. 2023, vol. 14, pp. 997383. ISSN 1664-8021. (Impact Factor 2.8)
7) GORDEEV, V., HÖLZER, M., DESIRÒ, D., GORAICHUK, IV., KNYAZEV, S., SOLO-GABRIELE, H., SKUMS, P., KARTHIKEYAN, S., EVANS, A., AGRAWAL, S., LUCACI, AG., MASON, CE., SU, JM., GIBAS, C., NARAJAN, N., PERES DA SILVA, R., DRABCINSKI, N., **MUNTEANU, V.**, ZHAN, L., RUBIN, J., WU, NC., TRISTER, A., CIORBĂ, D., BOSTAN, V., LOBIUC, A., COVASA, M., OPHOFF, RA., ZELIKOVSKY, A., DIMIAN, M., MANGUL, S. Leveraging wastewater sequencing to strengthen global public health surveillance. In: *BMC Global and Public Health*. 2025, vol. 3, pp. 23. ISSN 2731-913X.

Articles in the works of scientific manifestations included in the Web of Science and SCOPUS databases:

1) **MUNTEANU, V.**, LEAHU, A., CIORBĂ, D., CATLABUGA, E., DRABCINSKI, N., DUBCIUC, D., IAPĂSCURTĂ, V., BOSTAN, V. The Pangenome Variability Index: A Quantitative Measure for

Assessing Gene Content Diversity in Microbial Genomes. In: *International Conference on Nanotechnologies and Biomedical Engineering,* October 7-10, 2025, Chisinau, Moldova, vol. 2, pp. 1-9, ISBN 978-3-030-31865-9

2) BOLDIREV, G., SHARMA, NK., **MUNTEANU, V.**, BHAVATHARINI, A., KOSLICKI, D., ZELIKOVSKY, A., MANGUL, S. Assessing microbial genome representation across various reference databases: A coprehensive evalutaion. BioGENext: Next Generation Therapy Conference. September 17-20, 2024, Kyiv, Ukraine. In: *Biopolymers and Cell*. 2024, vol. 40, pp. 169-244, ISSN 1993-6842
3) SHARMA, NK., CHHUGANI, K., **MUNTEANU, V.**, SKUMS, P., ZELIKOVSKY, A., MANGUL, S. Realistic assortment of novel metagenomic benchmarks with diverse biological and technological characteristics. BioGENext: Next Generation Therapy Conference. September 17-20, 2024, Kyiv, Ukraine. In: *Biopolymers and Cell*. 2024, vol. 40, pp. 169-244, ISSN 1993-6842

Articles in the works of scientific events included in other databases accepted by ANACEC:

1) **MUNTEANU, V.**, DRABCINSKI, N., CIORBĂ, D., BOSTAN, V. Entropy-based Kullback-Leibler Taxonomic Classification of Biological Sequences. In: *Electronics, Communications and Computing*, October 17-18, 2024, Chisinau, Moldova, pp. 143-144, ISBN 978-9975-64-480-8
2) CATLABUGA, E., DRABCINSKI, N., **MUNTEANU, V.**, SUDACEVSCHI, V. Rare Events Detection and Forecasting in Dynamic Systems. In: *Electronics, Communications and Computing*, October 17-18, 2024, Chisinau, Moldova, pp. 167-168, ISBN 978-9975-64-480-8
3) **MUNTEANU, V.**, DRABCINSKI, N., CIORBĂ, D., MANGUL, S., BOSTAN., V. The reusability of public omics data across 5 million research publications. In: *Electronics, Communications and Computing*, October 17-18, 2024, Chisinau, Moldova, pp. 182-183, ISBN 978-9975-64-480-8
4) **MUNTEANU, V.**, CIORBĂ, D., POPIC, V., MANGUL, S. Developing bioinformatics capacity in Moldova. In: *Electronics, Communications and Computing*, October 20-21, 2022, Chisinau, Moldova, pp. 22-23, ISBN 978-9975-45-898-6

Articles in the works of scientific events included in the Register of materials published on the basis of organized scientific events in the Republic of Moldova:

1) POPOVA, D., **MUNTEANU, V.** Large Language Models in Academia: a case study at the Technical University of Moldova. In: *Technical and Scientific Conference for Unergraduate, Master's and Doctoral Students.* Technical University of Moldova, March 27-29, Chisinau, Moldova, vol I, pp. 324-331, ISBN 978-9975-64-458-7
2) BAS, A., **MUNTEANU, V.** Comprehensive assessment of sequence read archive metadata completeness. In: *Technical and Scientific Conference for Unergraduate, Master's and Doctoral Students.* March 27-29, vol II, pp. 1040-1045, ISBN 978 9975-64-460-0

ADNOTARE

la teza cu titlul "**Modele markoviene în timp continuu bazate pe filogenie pentru dinamica genelor în pangenomurile microbiene**", înaintată de competitorul **MUNTEANU Viorel**, pentru conferirea gradului științific de doctor în informatică, la specialitatea **122.3** "**Modelare, metode matematice, produse program**".

Structura tezei: teza a fost realizată în cadrul Universității Tehnice a Moldovei (UTM), Departamentul Inginerie Software și Automatică, Facultatea Calculatoare, Informatică și Microelectronică. Este scrisă în limba engleză și constă din introducere, 4 capitole, concluzii generale și recomandări, bibliografie din 348 de titluri, 116 text de bază, 47 figuri și 11 tabele. Rezultatele obținute au fost publicate în 16 lucrări științifice, inclusiv: 10 articole recenzate și în reviste cotate ISI și SCOPUS (dintre care 7 cu Factor de Impact); 6 articole în reviste din Registrul Național al revistelor de profil; 3 lucrări prezentate, recenzate și publicate la conferințe naționale și internaționale.

Cuvinte-cheie: bioinformatică, biostatistică, modelare matematică, metagenomică, pangenom, microbiom urban, genomică comparativă.

Scopul lucrării: dezvoltarea unui software bioinformatic reproductibil și scalabil destinat reconstrucției meta-pangenomului din date metagenomice, estimării cantitative a conținutului genic și a ratelor de câștig și pierdere de gene pe arborii filogenetici la nivel de linie taxonomică și clasificării genelor în funcție de presiunea selectivă exercitată de-a lungul acestor linii.

Obiectivele cercetării: (1) dezvoltarea unui software bioinformatic pentru adnotarea genomică, clusterizarea genelor ortoloage, aliniere a secvențelor și inferență filogenetică pentru construcția meta-pangenomului din genomuri asamblate din date metagenomice; (2) definirea și implementarea unui nou model pentru inferența câștigului și pierderii de gene pe arborii filogenetici la nivel de pangenom; (3) dezvoltarea și implementarea unui model statistic pentru detectarea presiunii selecției la nivel de gene și genomuri în pangenomuri; (4) validarea pe seturi empirice, inclusiv metagenom urban și izolate, cu studiu de caz pe genul *Klebsiella* și specia *Klebsiella pneumoniae*.

Noutatea și originalitatea științifică: rezultatele obținute contribuie la soluționarea problemei lipsei unor instrumente computaționale pentru reconstrucția pangenomului și inferența evolutivă a conținutului genomic direct din date metagenomice, prin dezvoltarea unui software bioinformatic scalabil și reproductibil care integrează reconstrucția pangenomului, inferența câștigului și pierderii de gene prin modele Markov cu timp continuu și clasificarea genelor în funcție de presiunea selectivă la nivel de genă și genom, permițând operaționalizarea conceptului de meta-pangenom și analiza integrată a dinamicii fluxului genic și a presiunii selective în comunități microbiene complexe, depășind limitările abordărilor centrate exclusiv pe izolate genomice.

Probleme științifică și de cercetare soluționată: lucrarea introduce un software bioinformatic scalabil și reproductibil pentru reconstrucția pangenomului, care integrează modele Markov în timp continuu și reconstrucția stărilor ancestrale pentru analiza evoluției conținutului genomic direct din date metagenomice. Produsul software dezvoltat permite aplicarea conceptului de meta-pangenom ca extensie a pangenomului, oferind un instrument riguros pentru analiza dinamicii fluxului genic și a presiunii selective asupra genelor în medii complexe.

Semnificația teoretică și valoarea aplicativă a lucrării: lucrarea contribuie la extinderea pangenomului către nivelul de meta-pangenom, oferind o bază metodologică pentru analiza dinamicii fluxului genic și a presiunii selective asupra genelor în medii diverse, inclusiv mediul urban. Din perspectivă aplicativă, software-ul dezvoltat face posibilă utilizarea meta-pangenomului ca instrument de supraveghere genomică, cu relevanță pentru sănătatea publică, agricultură, biotehnologie și abordarea *One Health*, permițând reconstrucția repertoriului genetic din date metagenomice complexe și realizarea de comparații robuste între ecosisteme și clade taxonomice.

Implementarea rezultatelor științifice: Metodele sunt implementate ca set de instrumente bioinformatice cu acces deschis, modulare, scalabile și reproductibile. Acestea sunt utilizat în activități de instruire, implementare în colaborare cu Agenția Națională de Sănătate Publică (ANSP) și laboratoare partenere (Institutul de Microbiologie și Biotehnologie al UTM), prin aplicații pilot de supraveghere genomică urbană și integrarea protocoalelor în fluxuri de lucru operaționale.

ANNOTATION

to the thesis entitled "***Phylogeny based continuous-time markov models for gene dynamics in microbial pangenomes***", submitted by the candidate **MUNTEANU Viorel** for the award of the scientific degree of Doctor in Computer Sciences, in the specialty **122.3** "**Modeling, mathematical methods, software products**".

Thesis structure: the thesis was carried out at the Technical University of Moldova (TUM), Department of Software Engineering and Automatics, Faculty of Computers, Informatics and Microelectronics. It is written in English and consists of an introduction, 4 chapters, general conclusions and recommendations, a bibliography of 348 sources, 116 pages of main text, 47 figures and 11 tables. The results obtained were published in 16 scientific works, including: 10 per-reviewed articles in ISI- and SCOPUS-indexed journals (of which 7 with Impact Factor); 6 articles in journals from the National Register of specialized journals; 3 papers presented, peer-reviewed and published at national and international conferences.

Keywords: bioinformatics, biostatistics, mathematical modelling, metagenomics, pangenome, urban microbiome, comparative genomics.

Aim of the work: the development of a reproducible and scalable bioinformatics software intended for the reconstruction of the meta-pangenome from metagenomic data, the estimation of gene counts and gene gain and loss rates on phylogenetic trees at the taxonomic lineage level, and the classification of genes according to the selective pressure acting along these lineages.

Research objectives: (1) the development of a bioinformatics software for genomic annotation, orthologous gene clustering, sequence alignment, and phylogenetic inference for the construction of the meta-pangenome from genomes assembled from metagenomic data; (2) the definition and implementation of a novel model for inferring gene gain and loss on phylogenetic trees at pangenome level; (3) the development and implementation of a statistical model for detecting selective pressure at the gene and genome levels in pangenomes; (4) validation on empirical datasets, including urban metagenomes and isolates, with a case study focusing on the genus *Klebsiella* and the species *Klebsiella pneumoniae*.

Scientific novelty and originality: the obtained results contribute to addressing the lack of computational tools for pangenome reconstruction and evolutionary inference of genomic content directly from metagenomic data, through the development of a scalable and reproducible bioinformatics software that integrates pangenome reconstruction, inference of gene gain and loss using continuous-time Markov models, and gene classification according to selective pressure at the gene and genome levels, thereby enabling the operationalization of the meta-pangenome concept and the integrated analysis of gene flow dynamics and selective pressure in complex microbial communities, overcoming the limitations of approaches centered exclusively on genomic isolates.

Scientific and research problem solved: the thesis introduces a scalable and reproducible bioinformatics software for pangenome reconstruction, which integrates continuous-time Markov models and ancestral state reconstruction to analyze the evolution of genomic content directly from metagenomic data. The developed software product enables the application of the meta-pangenome concept as an extension of the pangenome, providing a rigorous tool for analyzing gene flow dynamics and selective pressure acting on genes in complex environments.

Theoretical significance and practical value of the work: the research contributes to extending the pangenome toward the meta-pangenome level, providing a methodological basis for analyzing gene flow dynamics and selective pressure acting on genes across diverse environments, including urban settings. From an applied perspective, the developed software enables the use of the meta-pangenome as a genomic surveillance tool, with relevance to public health, agriculture, biotechnology, and the One Health approach, allowing reconstruction of genetic repertoires from complex metagenomic data and robust comparisons across ecosystems and taxonomic clades.

Implementation of the scientific results: the methods are implemented as an open-access, modular, scalable, and reproducible set of bioinformatics tools. They are used in training activities and in implementations carried out in collaboration with the National Agency for Public Health (ANSP) and partner laboratories (the Institute of Microbiology and Biotechnology, from TUM), through pilot applications of urban genomic surveillance and the integration of protocols into operational workflows.

АННОТАЦИЯ

к диссертации на тему “**Модели Маркова с непрерывным временем, основанные на филогении, для исследования динамики генов в микробных пангеномах**”, представленной соискателем **МУНТЯНУ Виорелом** на соискание ученой степени доктора по специальности **122.3** “**Моделирование, математические методы, программные продукты**” в области компьютерных наук.

Структура диссертации. диссертация выполнена в Техническом Университете Молдовы, на Кафедре Программной Инженерии и Автоматики, Факультете Вычислительной Техники, Информатики и Микроэлектроники. Работа написана на английском языке и включает: введение, 4 глав, общие выводы и рекомендации, библиографию из 348 источников, 116 страниц основного текста, 47 рисунков и 11 таблиц. Полученные результаты опубликованы в 16 научных работах, в том числе: 10 рецензируемых статьях в журналах, индексируемых в ISI и SCOPUS (из них 7 с импакт-фактором); 6 статьях в журналах Национального реестра профильных изданий; 3 докладах, представленных, рецензированных и опубликованных на национальных и международных конференциях.

Ключевые слова: биоинформатика, биостатистика, математическое моделирование, метагеномика, филогенетическая инференция, пангеном, городской микробиом, сравнительная геномика.

Цель работы: разработка воспроизводимого и масштабируемого биоинформатического программного обеспечения, предназначенного для реконструкции, мета-пангенома из метагеномных данных, оценки численности генов и скоростей приобретения и потери генов на филогенетических деревьях на уровне таксономических линий, а также классификации генов в зависимости от действующего вдоль этих линий селективного давления.

Задачи исследования: (1) разработка биоинформатического программного обеспечения для геномной аннотации, кластеризации ортологичных генов, выравнивания последовательностей и филогенетической инференции с целью построения мета-пангенома на основе геномов, собранных из метагеномных данных; (2) разработка и реализация новой модели для инференции процессов приобретения и потери генов на филогенетических деревьях на уровне пангенома; (3) разработка и реализация статистической модели для выявления селективного давления на уровне генов и геномов в пангеномах; (4) валидация на эмпирических наборах данных, включая городские метагеномы и изоляты, с исследованием на примере рода *Klebsiella* и вида *Klebsiella pneumoniae*.

Научная новизна и оригинальность: полученные результаты решают проблему отсутствия вычислительных инструментов для реконструкции пангенома и эволюционной инференции геномного содержания непосредственно из метагеномных данных. В работе разработано масштабируемое и воспроизводимое биоинформатическое программное обеспечение, объединяющее реконструкцию пангенома, инференцию процессов приобретения и потери генов на основе марковских моделей с непрерывным временем, а также классификацию генов по уровню селективного давления, что позволяет проводить интегрированный анализ динамики генетического потока и селективного давления в сложных микробных сообществах.

Решаемая научная проблема: диссертация представляет воспроизводимое биоинформатическое программное обеспечение для реконструкции пангенома, которое интегрирует марковские модели с непрерывным временем и реконструкцию предковых состояний для анализа эволюции геномного содержания непосредственно из метагеномных данных. Разработанный программный продукт обеспечивает применение концепции мета-пангенома как расширения пангенома, предоставляя строгий инструмент для анализа динамики генетического потока и селективного давления.

Теоретическая значимость и практическая ценность: исследование способствует расширению пангенома до уровня мета-пангенома, обеспечивая методологическую основу для анализа динамики генетического потока и селективного давления, действующего на гены в разнообразных средах. Разработанное программное обеспечение позволяет использовать мета-пангеном в качестве инструмента геномного мониторинга, имеющего значение для здравоохранения, сельского хозяйства, биотехнологии и подхода *One Health*, обеспечивая реконструкцию генетических репертуаров из сложных метагеномных данных и проведение устойчивых сравнений между экосистемами и таксономическими кладами.

Внедрение научных результатов: методы реализованы как открытый, модульный, масштабируемый и воспроизводимый набор биоинформатических инструментов. Они используются в учебной деятельности и внедрениях, проводимых в сотрудничестве с Национальным Агентством Общественного Здоровья и партнёрскими лабораториями (Институт Микробиологии и Биотехнологии, ТУМ), через пилотные проекты и интеграцию протоколов в операционные рабочие процессы.

MUNTEANU VIOREL

PHYLOGENY BASED CONTINUOUS-TIME MARKOV MODELS FOR GENE DYNAMICS IN MICROBIAL PANGENOMES

122.03 Modeling, mathematical methods, software products

Summary of the doctoral thesis in computer science

Approved for printing: *25.02.2026*
Offset paper. RISO Typing
Print sheets 2,5

Paper size 60×84 1/16
Circulation 50 ex.
Order no.

TUM, MD 2004, mun. Chisinau, bd. Stefan cel Mare si Sfant, no. 168.
"TEHNICA-UTM" Publishing House
MD-2045, Chisinau mun., 9/9 Studentilor street