TECHNICAL UNIVERSITY OF MOLDOVA

DOCTORAL SCHOOL

Presented 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

Doctoral thesis in computer science

Scientific Supervisor:
Bostan VIOREL, doctor habilitate, university professor

Chişinău, 2026

TECHNICAL UNIVERSITY OF MOLDOVA

DOCTORAL SCHOOL

Presented 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

Doctoral thesis in computer science

Author: Munteanu VIOREL ______________________

Scientific Supervisor:
Bostan VIOREL, dr. hab., univ. prof. V. Bostan ______________________

Members of the advisory committee:

1. Siminiuc RODICA, dr. hab., univ. conf., TUM ______________________
2. Ciorbă DUMITRU, dr., univ. conf., TUM ______________________
3. Mangul SERGHEI, dr., univ. prof., Sage Bionetworks, US ______________________

Chișinău, 2026

UNIVERSITATEA TEHNICĂ A MOLDOVEI

ȘCOALA DOCTORALĂ

Cu titlul de manuscris
CZU: 004.9:519.21:575

MUNTEANU VIOREL

MODELE MARKOVIENE ÎN TIMP CONTINUU BAZATE PE FILOGENIE PENTRU DINAMICA GENELOR ÎN PANGENOMURILE MICROBIENE

122.03 Modelare, metode matematice, produse program

Teză de doctorat în informatică

Conducător științific:
Bostan VIOREL, doctor habilitat, profesor universitar

Chișinău, 2026

UNIVERSITATEA TEHNICĂ A MOLDOVEI

ȘCOALA DOCTORALĂ

Cu titlul de manuscris
CZU: 004.9:519.21:575

MUNTEANU VIOREL

MODELE MARKOVIENE ÎN TIMP CONTINUU BAZATE PE FILOGENIE PENTRU DINAMICA GENELOR ÎN PANGENOMURILE MICROBIENE

122.03 Modelare, metode matematice, produse program

Teză de doctorat în informatică

Autor: Munteanu VIOREL ______________________

Conducător științific:
Bostan VIOREL, dr. hab., prof. univ. V. Bostan ______________________

Membrii comisiei de îndrumare:

1. Siminiuc RODICA, dr. hab., conf. univ., UTM ______________________
2. Ciorbă DUMITRU, dr., conf. univ., TUM ______________________
3. Mangul SERGHEI, dr., prof. univ., Sage Bionetworks, SUA______________________

Chișinău, 2026

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.

CONTENTS

LIST OF FIGURES

Figure 1.6. The pan-genome concept. The pangenome includes the core genome shared by all strains and the accessory genome, which varies across strains. Closed pangenomes have low variability; open pangenomes show extensive gene diversity due to ecological and evolutionary pressures.

Figure 1.7. Workflow for meta-pangenome reconstruction from multiple metagenomic samples. Metagenomic reads from different samples are independently assembled and taxonomically deconvoluted to generate species-level bins. For each species, predicted genes from assigned assemblies are pooled across samples to generate a non-redundant gene set. These gene sets are clustered into homologous groups to form species-specific meta-pangenomes. Accumulation curves are then generated to assess meta-pangenome openness based on the number of gene clusters discovered as more samples are included.

Figure 1.8. Comparative features of open and closed microbial pangenomes. Gene presence-absence matrices for an open (A) and a closed (B) pangenome. Grey indicates gene presence, white indicates absence, solid lines denote gene frequency across genomes. (C) Gene accumulation (solid lines) and core gene depletion (dashed lines) curves for open (blue) and closed (orange) pangenome. (D) Gene frequency distribution show a U-shaped pattern for the open pangenome, with a higher proportion of rare and core genes, while the closed pangenome is enriched for conserved genes. (E) Genome fluidity estimates highlight greater gene content variability in the open pangenome relative to the closed counterpart.

Figure 1.9. Two-state continuous-time Markov chain (CTMC) models used in gene gain-loss inference. (A) One-parameter model, both gene gain and gene loss transitions occur at the same rate q. (B) Two-parameter model, gain and loss transitions occur at independent rates q_g for gene gain and q_l for gene loss.

Figure 2.1. Reproducible meta-pangenome workflow and downstream inferences. (A) Genome annotation and meta-pangenome reconstruction; (B) Phylogeny inference; (C) Gene turnover counts (PGGL method) and gene classification according selection index and scores (PGGS method).

Figure 3.1. Comparative annotation metrics of *Klebsiella* sp. genomes across isolate-derived (p=PKG) and metagenome-assembled (mp=MPKG) datasets. Each panel show one metrics: (top-left) the number of predicted CDS; (top-right) total genome size (in base pairs); (bottom-left) number of tRNA genes; (bottom-right) number of contigs per assembly.

Figure 3.2. *Klebsiella pneumoniae* pangenome annotation based on isolate-derived genomes (PKP dataset).

values indicate loss bias — and the y-axis showing the ΔAIC between equal-rates (ER) and all-rates-different (ARD) models, reflecting statistical support for asymmetric rates.

Figure 4.22. Ancestral state reconstructions on the simulated phylogeny obtained using Fitch parsimony (A), Bayesian stochastic character mapping (B), and maximum likelihood (C), with internal nodes annotated as matches, mismatches, or ambiguous relative to the known simulated states.

LIST OF TABLES

LIST OF ACRONYMS

AIC	Akaike Information Criterion
AMR	Antimicrobial Resistance
ANI	Average Nucleotide Identity
ARD	All-Rates-Different (model)
ASR	Ancestral State Reconstruction
CDS	Coding DNA Sequence
cgMLST	Core Genome Multi-Locus Sequence Typing
CTMC	Continuous-Time Markov Chain
DNA	Deoxyribonucleic Acid
EC	Enzyme Commission
ER	Equal-Rates (model)
FDR	False Discovery Rate
GO	Gene Ontology
HGT	Horizontal Gene Transfer
ICE	Integrative and Conjugative Elements
I/O	Input/Output
INDEL	Insertion-Deletion
KEGG	Kyoto Encyclopedia of Genes and Genomes
KpSC	*Klebsiella pneumoniae* Species Complex
LR	Likelihood Ratio
LRT	Likelihood-Ratio Test
MAG(s)	Metagenome-Assembled Genome(s)
MLST	Multi-Locus Sequence Typing
NCBI	National Center for Biotechnology Information
ORF	Open Reading Frame
OUT	Operational Taxonomic Unit
PGGL	Pangenome Gene Gain-Loss
PGGS	Pangenome Gene Selection
QC	Quality Control
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
ROI	Region of Interest
ST	Sequence Type
WGS	Whole-Genome Sequencing

INTRODUCTION

Timeliness and importance of the problem addressed. The evolutionary dynamics of microbial genomes, driven by mutation, selection, recombination, horizontal gene transfer (HGT) and gene gain-loss events, underpin the ecological versatility, adaptive potential, and public health relevance of microbial populations [5–8]. In natural and engineered environments, such as urban microbiomes, these dynamics are driven not only by mutation and selection, but also by rapid gene turnover, mediated through horizontal gene transfer, gene loss, and recombination [5, 8, 9]. While the pangenome has emerged as a powerful conceptual framework to capture this genomic fluidity, existing approaches remain largely isolate-centric, limiting their utility in complex, uncultured communities [10–12].

This thesis addresses a pressing need for methods capable of reconstructing and interpreting microbial pangenomes directly from metagenomic data, without reliance on predefined genomes or species boundaries [13]. By leveraging phylogenetic models of gene gain and loss and integrating these with high-resolution metagenomic assemblies and presence–absence matrices, this work enables lineage-aware inference of pangenome structure and evolutionary dynamics across diverse microbial taxa [13, 14].

Such approaches are timely and essential as environmental and host-associated microbiomes are now recognized as reservoirs of antimicrobial resistance, metabolic innovation, and pathogenic potential; quantifying in situ gene-content dynamics and deploying urban genomic surveillance that tracks antimicrobial resistance genes (ARGs,) virulence genes, and mobile elements across city infrastructures have become imperative [9, 11]. In anthropogenically structured ecosystems, including urban wastewater, soils, and air, quantifying the rates and mechanisms of gene flux (HGT, recombination, gain and loss) is critical for tracing the emergence of adaptative traits and forecasting microbial responses to selective pressures [1, 11, 15].

By formalizing phylogeny-aware comparative models for metagenomic data, this work provides a rigorous framework for microbial evolutionary inference that couples evolutionary theory to real-world community complexity, while supplying an operational analytics layer for urban genomic surveillance. It enables more predictive models of genome evolution, increases the resolution of surveillance, and sharpens the dissection of mechanisms underlying microbial adaptation in the metagenomic era.

Research field. This thesis is positioned at the intersection of bioinformatics, mathematical modeling, analysis of metagenomic data and comparative genomics [10–12]. It contributes to the emerging discipline of computational pangenomics by developing theoretical and algorithmic frameworks to infer gene content evolution in microbial and viral populations, particularly within urban ecosystems [16, 17]. The research integrates phylogenetic comparative methods with metagenomic data analysis to resolve lineage-specific patterns of gene gain, loss, and recombination, even in the absence of isolate genomes. It advances the broader understanding of genome dynamics in natural communities and has direct implications for public health microbiology, antimicrobial resistance (AMR) surveillance, and real-time evolutionary inference in complex environments.

Situation in the field and the research problem. Urban microbiomes are characterized by pronounced genome plasticity driven by frequent gene gain, gene loss, and homologous recombination, which together shape microbial adaptation, transmission, and antimicrobial resistance in densely populated, anthropogenically influenced ecosystems [8, 9, 16, 17]. Capturing these evolutionary processes is central to understanding how microbial functions emerge, persist, and spread in cities.

Most computational studies of genome evolution still rely on reference-based, isolate-derived frameworks. Such models underperform on real-world metagenomes, where assemblies are fragmented, strain mixtures are common, coverage is uneven, and genome boundaries are uncertain [1, 18–20]. Treating genomes as static entities obscures lineage-specific gene-content dynamics, and the lack of explicit phylogenetic integration hampers separation of vertical inheritance from convergent or horizontally transferred events [21–23].

This thesis addresses these limitations by introducing phylogeny-aware pangenomic frameworks that reconstruct gene repertoires and model gain–loss dynamics directly from metagenomic data. Using ancestral state reconstruction, continuous-time Markov chains (CTMC), and branch-specific event inference, the approach resolves fine-grained trajectories of gene content across environmental lineages, even without isolate genomes, while accommodating the high recombination rates, rapid turnover, and ecological selection typical of urban samples [14, 17, 22, 23].

By coupling tree-based models with presence–absence matrices across thousands of gene families from meta-pangenomes, the work delivers a computational toolkit for detecting lineage-specific adaptation, quantifying gene-turnover rates (λ/μ), and delineating conserved versus variable components of the microbial pangenome. This fills a key methodological gap in evolutionary metagenomics and strengthens urban genomic surveillance, enabling in situ analysis of genome dynamics with direct applications to public-health microbiology, antimicrobial resistance (AMR) monitoring, and microbial ecology [1, 8, 9, 11, 16, 17, 24–26]. We additionally assess selective pressure on gene presence/absence by testing rate asymmetry of gain versus loss on phylogenies within a continuous-time framework and classifying genes accordingly [27, 28]. This allows lineage-specific detection of preferential retention or acquisition beyond neutral expectations while remaining orthogonal to nucleotide-level selection signals methods.

The aim of the study. To develop and validate a modular, scalable, and reproducible bioinformatics software for meta-pangenome reconstruction and analysis from urban metagenomic data, integrating probabilistic modeling and phylogeny-based inference to quantify gene-content variability and gene-dynamics in pangenomes, and to develop a statistical classification method for identifying selective pressure on genes along taxonomic lineages. The bioinformatics framework should reconstruct pangenomes directly from metagenome-assembled genomes (MAGs) and estimates branch-wise gene gain and loss rates, classifying genes by inferred selection pressure along evolutionary lineages using continuous-time Markov models with likelihood-based ancestral reconstruction.

Objectives of the research:

- Develop and implement an end-to-end, reproducible, modular bioinformatics software that converts labeled genomic sequences into harmonized 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 evolution from metagenomic presence–absence data by inferring lineage-specific gain and loss dynamics and producing branch- and lineage-level summaries suitable for comparative analyses and urban 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 rate-contrast indices to capture the evolutionary tendency of gene-content change.
- Rigorously validate the end-to-end, reproducible software developed by applying them to urban microbiome datasets to reconstruct pangenomes, benchmark against gold-standard datasets and perform CTMC phylogeny-based inference of branch-specific gene gain and loss and classify genes by selective regime.

Research hypothesis. We hypothesize that applying continuous-time Markov chain models with likelihood-based ancestral state reconstruction to gene-family presence-absence data reconstructed from MAG-based meta-pangenomes will recover lineage-specific rates of gene gain (λ) and loss (μ) and a directionality in gene turnover, even when isolate genomes or complete assemblies are unavailable, building on established phylogenetic likelihood theory for discrete characters and its extension to gene-content evolution [29–31]. We further hypothesize that these estimates are biologically meaningful and robust in metagenomic settings characterized by recombination and horizontal transfer, aligning with the documented fluidity and selection shaping prokaryotic pangenomes [11, 32–34]. Given the availability of large, high-quality MAG catalogues and city-scale metagenomic surveys that enable species-level inference from metagenomes [35–37], we expect the same CTMC and ancestral state reconstruction framework, applied to urban meta-pangenomes, to reveal reproducible lineage-specific shifts in genome expansion, reduction, and turnover that covary with ecological pressures and public-health features such as AMR and virulence burdens, despite typical artifacts like fragmentation and strain mixing. Accordingly, we advance the following scientific theses for defense in this dissertation:

1. Species-resolved meta-pangenomes can be reconstructed directly from quality-controlled urban MAGs, with optional co-analysis of isolates, yielding gene-family presence–absence matrices, openness statistics, and recombination-aware core-genome phylogenies suitable for downstream inference.
2. Gene/orthogroup presence–absence derived from these meta-pangenomes supports phylogeny-aware continuous-time Markov models that estimate branch-specific gain (λ) and loss (μ) and provide lineage-resolved rate indicators for urban surveillance.
3. Directional selection on gene content can be tested by contrasting symmetric (ER) versus asymmetric (ARD) CTMC parameterizations at the gene/orthogroup level, producing interpretable statistics that quantify bias toward acquisition or deletion.
4. Integrated annotation over the meta-pangenome enables systematic prioritization of sequences of concern, including AMR genes, virulence loci, and prophage/viral segments, by combining prevalence, phylogenetic context, and mobile-element co-occurrence into ranked, lineage-resolved watchlists for urban genomic surveillance.
5. Application to urban compartments yields an integrated indicator set (λ, μ, selection indices, prioritized sequences) that is reproducible under a fixed analysis stack and directly consumable by One Health comparative and early-warning workflows.

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: (1) an urban MAG meta-pangenome (MPKG dataset; n = 64 high-quality MAGs), and two isolate collections (PKG dataset; n = 35 genomes, and PKP; n = 99 genomes), enabling comparisons across 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.

Gene-content evolution was modeled on the fixed core phylogeny using a two-state continuous-time Markov chain fitted per orthologous groups, likelihoods were computed with Felenstein's pruning algorithm [31], and the gain (λ) and loss (μ) parameters were estimated by maximizing likelihoods with a bound-constrained quasi-Newton optimizer (L-BFGF-B) implemented in stats R package [38].

To place the evolutionary results in a public-health frame, we integrated gene-content inferences with curated AMR and virulence resources and with prophage/viral calls, projecting all signals onto a single recombination-aware phylogeny to obtain lineage-level summaries of AMR burden, virulence potential, and virome integration. PGGL and PGGS methods, downstream statistics, and visualizations were implemented in R [38]; upstream assembly, annotation, orthogrouping, alignment, and phylogeny steps were executed with established command-line bioinformatics software, as detailed in the Methods chapter. All analyses ran in versioned, containerized environments on local and HPC systems to ensure reproducibility and scalability [39].

Scientific novelty and originality. This thesis advances microbial evolutionary genomics by introducing a phylogeny-aware meta-pangenome framework 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 urban microbiomes including fragmented assemblies, strain mixtures, and pervasive horizontal gene transfer and recombination, where classic isolate-centric comparative genomics is brittle.

Methodologically, the work contributes two integrated components. PGGL method is a maximum-likelihood, continuous-time Markov framework applied to a fixed core-genome phylogeny that treats each orthologous group as a two-state character (absent/present). It returns gene-wise estimates of the rate of acquisition (gain) and rate of deletion (loss), marginal ancestral states, and expected branch-specific events, computed via Felenstein's pruning algorithm; this enables lineage-aware quantification of gene turnover across large microbial cohorts. Building on that, PGGS method introduces a phylogeny-aware 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 (via likelihood criteria) classifies genes as gain-biased, loss-biased, or consistent with symmetry, providing an interpretable signal of selection acting on gene presence/absence that complements codon-level analyses.

Conceptually, the novelty is to bring comparative-phylogenetic logic to gene-content traits at meta-pangenome scale, delivering robust, lineage-aware estimates of turnover and directional bias from empirical MAGs and isolates alone. Practically, the framework is modular and reproducible, integrates curated AMR/virulence/virome layers on the same tree, and is purpose-built for urban genomic surveillance, where cross-dataset and cross-location comparability is essential.

The important scientific problem solved in the thesis. A key unmet need in microbial evolutionary genomics is to infer gene gain–loss dynamics and quantify genome-content variability directly from metagenomic cohorts, where assemblies are fragmented, strains co-occur, species boundaries are fuzzy, and horizontal gene transfer and recombination are pervasive. Classic pangenome and comparative-genomics pipelines were built for complete isolate genomes and stable taxonomies that under metagenomic conditions they lose power and interpretability, obscuring lineage-specific trajectories of genome expansion and reduction. The urgency is amplified in urban ecosystems, now profiled at scale, which exhibit high turnover of mobile genes yet are sampled predominantly by metagenomics [35, 36].

Two technical barriers make this a bona-fide scientific problem. First, recombination and gene flow distort tree-like signal, complicating phylogeny construction [33, 40, 41]. Second, existing gene gain–loss methods were largely designed for complete, well-delimited genomes and are not validated for presence–absence matrices derived from MAGs, as a result, the field lacks a standard, lineage-aware approach to estimate gain and loss from metagenomes or to test for directional bias in turnover at gene level [21, 27, 42]. Together, these gaps prevent rigorous, phylogeny-aware quantification of genome dynamics precisely where surveillance needs are greatest—wastewater, air, and built environments.

This thesis closes the gap by introducing a phylogeny-based meta-pangenome analysis bioinformatics software that reconstructs gene repertoires directly from metagenomic assemblies and, on a recombination-aware core phylogeny, fits two-state continuous-time Markov models to each orthogroup to estimate lineage-specific gene gain (λ) and loss (μ). To test selection on gene presence/absence, we contrast an equal-rates (ER) model, gain equals loss, with an all-rates-different (ARD) model that allows them to differ; support for ARD indicates directional turnover (preferential acquisition or deletion), whereas ER is consistent with symmetric/neutral turnover. This yields robust, comparable estimates from metagenomic data alone, enabling cross-dataset, lineage-aware evolutionary inference in urban microbiomes.

Theoretical significance of the research. This thesis advances comparative theory for microbial pangenomes by making gene-content evolution estimable directly from metagenomic cohorts. We adapt discrete-trait likelihood models to orthogroup presence/absence on recombination-aware phylogenies and formalize a test for selection on gene content, bridging isolate-centric models and the realities of fragmented, strain-mixed communities. Formally, the theoretical contributions of this work are as follows, each extending likelihood-based comparative models to gene-content data on recombination-based phylogenies:

1. 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.
2. 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.
3. Uses likelihood-based ancestral reconstruction to obtain marginal ancestral state probabilities and branch-specific expected counts of gene gains and losses.
4. 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 work. In practical terms, this framework operationalizes phylogeny-aware metagenomics for surveillance and monitoring. Its outputs and workflows are structured for direct use by public-health, environmental, and research teams. In applied contexts, this framework translates into the following operational capabilities and deliverables for surveillance and monitoring:

- 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.

Main scientific results. This work advances phylogeny-aware meta-pangenomics for urban metagenomes and demonstrates the approach on empirical *Klebsiella* datasets (genus *Klebsiella* and *K. pneumoniae*). The main results are:

1. 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.
2. 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.
3. 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.
4. 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). The computational framework (meta-pangenome reconstruction, lineage-aware gain/loss and selection pressure) was packaged as reproducible, containerized workflows and deployed on institutional server for routine analyses of urban microbiome (wastewater, air, and build-surface swabs). Additionally, building directly on the results and methods of this thesis, the Technical University of Moldova's Bioinformatics Laboratory secured a complex bilateral grant, namely "***UPGRADE: Genomic surveillance of urban pathogens for environmental and public health protection: a One Health approach***" to scale and operationalize the framework across partner institutions.

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 and Automata, Faculty of Computers, Informatics and Microelectronics, Technical University of Moldova (2022-2025) and the Department's Scientific Seminar (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 findings were disseminated in high-impact journals, including *Frontiers in Genetics*, *PeerJ Computer Science*, *Nature Reviews Methods Primers*, *Nature Water*, *Cell Genomics*, and *Genome biology* reflecting both the methodological advances and their relevance to computational biology and microbial surveillance.

Thesis structure. The thesis comprises 116 pages and includes an introduction, 4 chapters, conclusions and recommendations, a bibliography from 348 sources, 5 annexes, 46 figures, and 8 tables.

Summary of the sections of the thesis: In the ***Introduction*** we justify the relevance and timeliness of the topic, present a critical review of current research and technology, state the thesis aim and objectives, and articulate the scientific novelty and main theses advanced for defense. We also document the robustness and validation of the results and list the conferences where the core findings were presented.

In ***Chapter 1*** we describe the urban microbiome and virome context, articulate the limitations of isolate-only analyses, and motivate meta-pangenomes as the necessary complement for recovering accessory diversity that drives adaptation and risk. We define core versus accessory structure, pangenome openness, and the rationale for phylogeny-aware inference, establishing the conceptual scaffold for the pipeline and models that follow.

Later, in ***Chapter 2***, we develop a unified pipeline that standardizes gene calling, ortholog clustering, functional annotation, core-genome phylogeny with recombination control, and pangenome statistics for both MAGs and isolates. We introduce two modeling methods: (1) PGGL, which maps gains and losses on the phylogeny and summarizes branch-standardized rates and genome burdens; and (2) PGGS, which quantifies selection from the asymmetry between gain and loss rates and assigns calibrated effect sizes and directional classes.

In ***Chapter 3*** we apply the framework to reconstruct meta- and isolate-based pangenomes and show that input type shapes apparent architecture while core signals are conserved. MAGs broaden ecological coverage and expose low-prevalence, habitat-linked gene pools; isolates deliver higher contiguity, clearer species boundaries, and higher per-locus fidelity. Core–accessory structure is consistently recovered, openness follows phylogenetic and ecological scope rather than taxon labels, and structure-aware matrices with PCA reveal lineage-linked accessory islands that provide a stable reference for functional comparison.

In ***Chapter 4*** we estimate where gene-content turnover occurs and how it is biased. PGGL localizes gains and losses to specific branches and quantifies per-branch and per-genome burdens, while PGGS tests symmetry versus directionality and converts λ–μ asymmetry into a quantitative readout of selection at gene level. Together these components produce an interpretable atlas of counts, rates, and selection regimes that generalizes across datasets.

Each chapter ends with conclusions that synthesize the research and a summary of the main results. The final ***Conclusions and Recommendations*** chapter present the principal outcomes, published in peer-reviewed journals, and demonstrate the theoretical and practical value of the work on meta-pangenome reconstruction, phylogeny-aware gene turnover modeling, and integrated AMR, virulence across environmental and clinical contexts.

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

1. URBAN MICROBIOMES TO PANGENOMES CONCEPTS AND METHODS

1.1. Urban microbiome and virome

Historically, urban environments have housed only a fraction of the global population, with the majority residing in rural areas or small villages. However, trend has dramatically shifted in the last decades, with 55% of the world's population now living in urban areas, a figure project to rise nearly 70% by 2050 [43, 44]. Rapid urbanization has transformed the human-microbe interface, shaping unique microbial ecosystems within densely populated areas.

Since the advent of germ theory and John Snow's pioneering work on cholera, it has been evident that the dynamics of human-microbial interactions differ significantly between urban and natural settings [45, 46]. In urban environments, factors such as high population density, extensive human mobility, waste accumulation, infrastructure complexity, and build environments contribute to the diversity and spread of microorganisms, including both commensals and pathogenic species. The urban microbiome, composed of bacteria, viruses and fungi, is influenced by environmental factors such as air pollutions, wastewater systems, and urban wildlife, making it a crucial component of public health surveillance [16, 17, 47]. Similarly, the urban virome is the collection of viruses within an urban environment, also playing a particularly important role in shaping microbial communities and human health. Urban virome encompass bacteriophages that regulate bacterial populations, as well human and zoonotic viruses with potential implications for disease emergence [26, 48, 49]. Factors such as international travel, climate change, and antibiotic resistance further influence the urban microbial landscape, underscoring the need for comprehensive metagenomic surveillance to monitor emerging pathogens and antimicrobial resistance genes [16, 26, 50, 51].

Microbial communities in the built environment are increasingly recognized as potential reservoir of pathogens and contributors to human disease [52]. Urbanization has also been linked to rising allergy prevalence, likely driven by reduced microbial diversity and environment shifts that alter immune system development [53]. Despite growing evidence that cities shape human health, the mechanisms remain complex and poorly understood.

Our understanding of urban microbial dynamics beyond epidemic events is still nascent. Advances in metagenomics and environmental sequencing have revealed diverse and dynamic microbial ecosystems, yet their interactions with human populations and infrastructure remain largely unexplored [54]. As urbanization accelerates, deciphering these complex microbial networks will be critical for mitigating health risks and designing resilient urban environments.

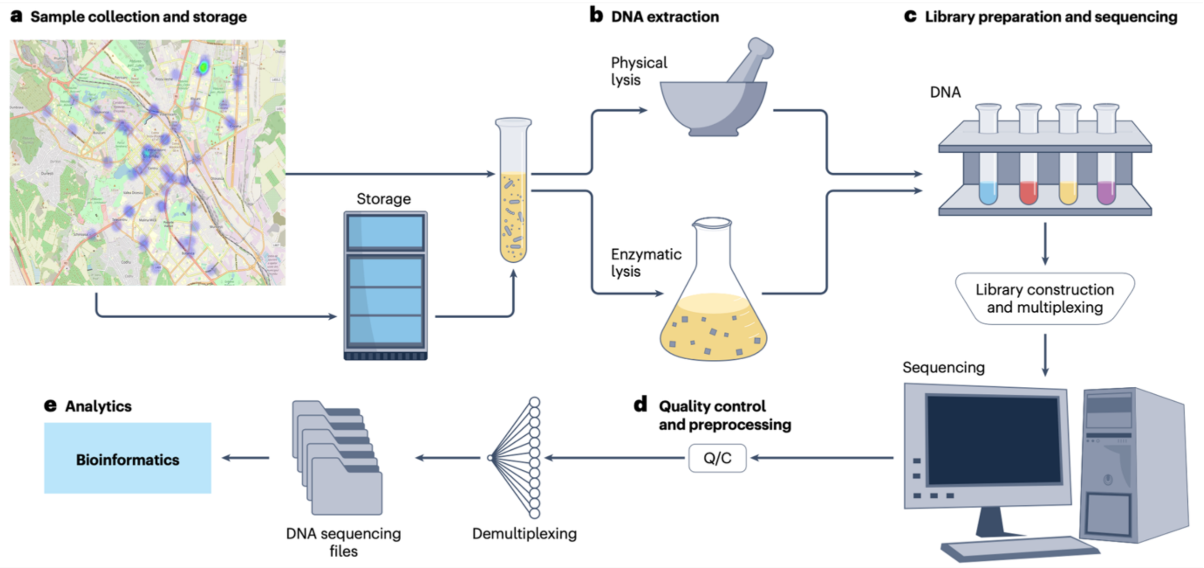

a Sample collection and storage
b DNA extraction
c Library preparation and sequencing
Physical lysis
DNA
Storage
Enzymatic lysis
Library construction and multiplexing
Sequencing
e Analytics
Bioinformatics
d Quality control and preprocessing
Q/C
DNA sequencing files
Demultiplexing

Figure 1.1. Workflow for metagenomic sampling, sequencing and analysis (adapted from [1]). (a) Environmental samples—such as air, surface swabs, or water—are collected from urban locations, georeferenced, and stored appropriately. (b) DNA is extracted from each sample using a combination of physical and enzymatic lysis methods. (c) Extracted DNA is used to prepare sequencing libraries, which are multiplexed using sample-specific barcodes for parallel sequencing. (d) Raw sequencing reads undergo quality control and preprocessing, including adapter trimming and removal of low-quality sequences. (e) Reads are demultiplexed by barcode and subjected to bioinformatics workflows for downstream metagenomic analysis, including taxonomic, functional, and statistical profiling.

Advances in next-generation sequencing (NGS) and metagenomics have transformed the study of urban microbiomes, enabling rapid, global profiling of microbial communities and their interactions with hosts (Figure 1.1). These technologies provide unprecedented resolution for characterizing microbial dynamics in cities, informing both clinical and public health strategies. NGS allows for culture-independent, high-throughput sampling and data generation, enabling simultaneous taxonomic and functional annotation – crucial for monitoring the emergence and spread of pathogens and antimicrobial resistance (AMR) [55, 56].

Integrating metagenomic surveillance with spatial and temporal analyses offers unprecedented opportunities to track the emergence, evolution and dissemination of AMR in urban environments. High-resolution molecular mapping can identify AMR hotspots, monitor the movement of resistance genes across microbial communities, and quantify the impact of environmental and public health interventions [16, 57]. Such approaches are crucial for assessing the role of horizontal gene transfer in AMR propagation and detecting reservoirs of resistance that may be overlooked in conventional surveillance strategies.

Urban microbiomes are dynamic ecosystems shaped by complex interactions between infrastructure, human activities, and environmental factors. Wastewater, transportation hubs, and healthcare facilities serve as critical nodes for microbial exchange, enabling the transmission of AMR determinants between commensal and pathogenic bacteria [58, 59]. Climate change and increasing global connectivity further exacerbate the spread of AMR, highlighting the need for a coordinated, global surveillance framework [16, 60].

A system-level approach – combining metagenomics, epidemiology, and advance bioinformatics methods, including machine learning (ML) based methods, will be essential to disentangle the complexity of AMR transmission and predict resistance trends [24]. However, significant gaps remain in understanding how urbanization shapes microbial adaption, necessitating interdisciplinary collaborations across genomics, public health, environmental science and computational biology. Deciphering these interactions will be pivotal in mitigating AMR and pathogens spread and designing resilient urban environments.

1.2. Analysis of metagenomic data

Metagenomic data analysis involves a structured series of computational steps that transform raw sequencing reads into interpretable information about microbial community composition and function. The initial stages focus on quality control and preprocessing, including read trimming, removal of sequencing artifacts, and assessment of coverage, assembly completeness, and contamination. These steps ensure data integrity and reduce technical noise in downstream analyses. Subsequent stages include metagenomic assembly, taxonomic classification, genome binning, and functional annotation. These processes are computationally intensive and typically require access to high-performance computing infrastructure. Advances in bioinformatics tools and pipelines have improved the scalability and accuracy of metagenomic analysis, but challenges such as low-abundance genome recovery, strain resolution, and contamination remain, particularly in complex or low-biomass samples.

1.2.1. Sequencing platforms

Sequencing technologies form the methodological foundation for metagenomic data analysis, enabling the recovery of microbial and viral genetic material from complex environments. Among these, Illumina sequencing platforms (Figure 1.2) are widely utilized for metagenomic studies due to their high-throughput, high-accuracy generation of short reads, typically ranging from 50 to 300 base pairs. The high degree of parallelization across millions of DNA fragments allows for deep sequencing coverage and comprehensive community profiling [1]. These platforms are particularly suitable for variant detection, differential gene expression analysis, and resistome surveillance in metagenomics, where cost-efficiency, data consistency, and analytical reproducibility are paramount [1]. Furthermore, the extensive ecosystems of optimized library preparation protocols, bioinformatics pipelines, and quality control standards reinforces the reliability of short-read platforms for large-scale environmental and clinical metagenomics [61].

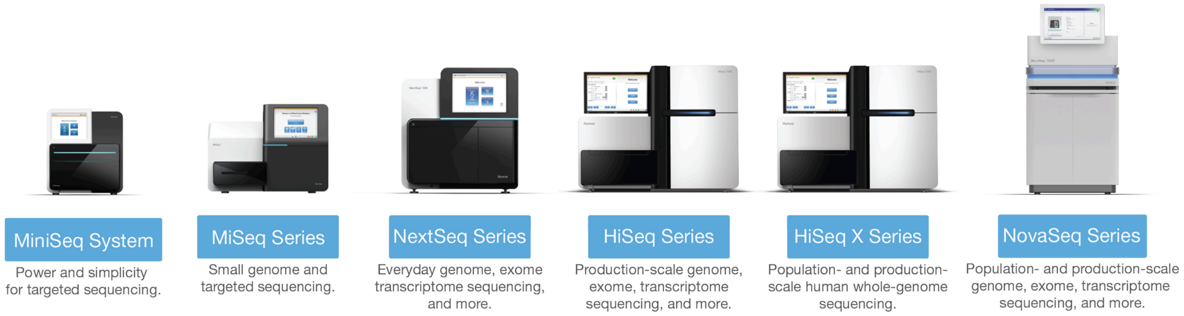

MiniSeq System
Power and simplicity for targeted sequencing.
MiSeq Series
Small genome and targeted sequencing.
NextSeq Series
Everyday genome, exome transcriptome sequencing, and more.
HiSeq Series
Production-scale genome, exome, transcriptome sequencing, and more.
HiSeq X Series
Population- and production-scale human whole-genome sequencing.
NovaSeq Series
Population- and production-scale genome, exome, transcriptome sequencing, and more.

Figure 2. Illumina short read sequencing systems covering from small benchtop sequencers to production-scale sequencing systems [2].

Long-read sequencing technologies, such as those developed by Oxford Nanopore Technologies (ONT), and Pacific Biosciences (PacBio) are PCR-free methods (Figure 1.3) and offer complementary advantages by producing reads spanning kilobases to megabases length. These platforms facilitate the resolution of structurally complex genomic regions, including those containing long repeats, transposons, and other mobile genetic elements. In urban metagenomics, long-read approaches have enabled the accurate reconstruction of bacterial and viral genomes, the mapping of antimicrobial resistance genes within their genomic context, and the assembly of complete plasmids or chromosomal segments [62, 63].

ONT-based nanopore sequencing is particularly suited for real-time analysis in-field deployment. Devices such as the MinION or PromethION (Figure 1.3A) offer scalable throughput and are increasingly used in pathogen surveillance, including rapid profiling of urban wastewater and outbreak samples [64–66]. PacBio's Sequell II systems (Figure 1.3B) employ circular consensus sequencing to produce high-fidelity (HiFi) reads, offering read accuracies comparable to Illumina but with longer read lengths, thus enabling strain-level resolution and complete genome closure [67].

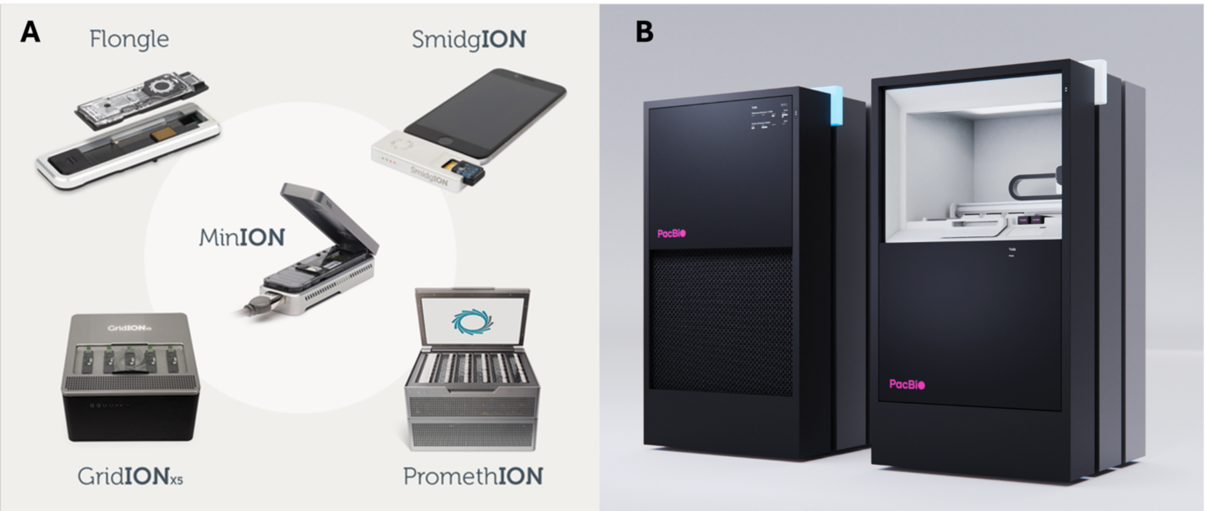


Figure 1.3. Representative long-read sequencing platforms used in metagenomic and microbial genomics. (A) Oxford Nanopore Technologies (ONT) sequencing devices [3]. (B) Pacific Bioscience (PacBio) Sequell II and and Sequel IIe platforms offering high-fidelity (HiFi) long-read sequencing through circular consensus sequencing [4].

Despite these strengths, long-read platforms have traditionally exhibit higher error rates than short-read technologies. While ONT reads previously showed per-base error rates in the range of 4-10%, recent updates, such as the R10.4.1 flow cell and Q20+ chemistry, have significantly improved consensus accuracy [68–70]. PacBio's HiFi sequencing updates using PacBio Sequel II system now routinely reach error rates below 1% [67]. However, both platforms still involve higher per-gigabase sequencing costs, longer run times, and more complex library preparation procedures, limiting their scalability for population-level studies [62].

To address these limitations, hybrid sequencing strategies have become increasingly common [67]. These combine the accuracy and affordability of short-read sequencing with the long-range continuity of long-read data. Hybrid assemblies improve contiguity, resolve repetitive regions, and enable high-confidence reconstruction of metagenome-assembled genomes (MAGs), resistance gene clusters, and mobile elements that are otherwise fragmented or misassembled using short reads alone [67, 71].

Selecting an appropriate sequencing platform requires careful consideration of sample type, target resolution (e.g., species-level versus strain-level), cost constraints, and downstream analytical goals. In urban microbial ecology, where diversity, mobility, and infrastructure-driven selection pressures intersect, multi-platform sequencing strategies offer a powerful means to capture both the genomic breadth and structural depth of microbial and viral communities.

1.2.2. Quality control of metagenomic data

Quality control (QC) is a critical initial step in the analysis of metagenomic data, ensuring that technical artifacts and low-quality sequences do not compromise downstream analyses. Prior to QC, a demultiplexing step is required to distinguish individual samples based on index sequences incorporated during library preparation. Tools such as Flexbar [72], Ultraplex [73], and others are commonly used for this purpose, supporting both barcode assignment and adapter trimming in a single step.

Whole-genome shotgun sequencing of metagenomic samples introduces various biases, including uneven coverage and random fragmentation, which can complicate assembly and interpretation. To mitigate these effects, raw reads must be assessed and filtered based on quality metrics. A fundamental measure of read quality is the Phred score, which estimates the probability of base-calling errors [74, 75]. Reads with Phred scores below 30 typically contain sequencing artifacta, such as PCR duplicates, base miscalls, and adapter contamination and should be excluded. Commonly used QC tools include FastQC [76], PRINSEQ [77], Trimmomatic [78], and BBTools [79] and others [80], which provide comprehensive diagnostics and support batch trimming and filtering workflows. These tools enable the removal of low-quality bases, sequencing adapters, and overrepresented k-mers, the latter of which can indicate contamination, primer bias, or repeats. QC reports often include metrics such as GC content, read length distribution, sequence duplication levels, and the frequency of ambiguous bases. Reads that fall below user-defined length thresholds after trimming are typically discarded.

For long-read sequencing technologies such as Oxford Nanopore and PacBio, QC requires platform-specific strategies, including tools like NanoFilt allow filtering based on read length, GC content, and mean read quality [81], while LongQC [82] performs more nuanced assessments by identifying anomalous or "nonsense" reads, often associated with poor signal output or pore-level errors. Because long-read platforms often sequence native DNA, they may also retain methylation signatures and other modifications, further motivating platform-aware filtering strategies.

An essential component of QC is host read decontamination, particularly for samples derived from human, animal, or plant hosts. Tools such as DeconSeq [83] and KneadData [84] can automatically align reads to host reference genomes and remove them from downstream analysis. This step reduces noise and computational burden while preventing spurious alignments during taxonomic classification or functional annotation. Overall, rigorous quality control enhances the accuracy and reproducibility of metagenomic studies, supporting reliable inferences about microbial community composition, function, and dynamics.

1.2.3. Metagenomic assembly

The assembly of sequencing reads into contiguous sequences represents a central step in metagenomic analysis, offering a more complete and interpretable view of microbial community structure than direct read-based approaches. Assembling reads into contigs or metagenome-assembled genomes (MAGs) enhances taxonomic resolution and enables more accurate functional annotation, particularly in complex microbial ecosystems. The process typically begins with short-read or long-read assemblers that reconstruct contiguous fragments from millions of overlapping sequences. Most short-read assemblers, such as MEGAHIT [85] and metaSPAdes [86], employ de Bruijn graph algorithms [87], which represent reads as networks of overlapping k-mers. While computationally efficient, these methods can struggle with uneven coverage, low-abundance genomes, and repetitive genomic elements common in environmental samples.

To address these limitations, long-read sequencing technologies have facilitated the development of specialized metagenomic assemblers such as metaFlye [88], which adjusts k-mer frequency estimation to account for heterogeneous coverage, and hifiasm-meta [89], which modifies read selection criteria to improve assembly from low-abundance species using high-fidelity long reads. Tools like metaMDBG [90] introduce minimizer-based strategies that reduce memory usage and improve scalability, particularly for high-quality long-read datasets. Hybrid assembly approaches, which combine short-read accuracy with long-read contiguity, have proven especially powerful in resolving complex genomes from mixed communities. Assemblers such as hybridSPAdes [91], Opera-MS [92], and Unicycler [93] consistently outperform single-platform approaches in terms of accuracy, completeness, and strain resolution. Recent methods such as haplotype-resolved hierarchical clustering-based hybrid assembly (HCBHA) go a step further by phasing reads into haplotypes prior to assembly, enabling near-complete genome reconstruction from metagenomic samples with high strain-level diversity [94].

Nevertheless, strain-level resolution in metagenomes remains a computationally challenging task. Unlike diploid genomes, microbial communities may harbor dozens of closely related strains, often differing by only a few single-nucleotide polymorphisms. Resolving these variants requires strategies borrowed from viral quasispecies analysis and human haplotyping, although the number of co-occurring genotypes is typically unknown and often much larger. This problem, formally categorized as NP-hard [95, 96], is tackled by heuristic algorithms that may fail to capture rare strains or produce artificial recombinant sequences. An alternative strategy involves identifying strains based on variation in conserved genes or gene families, bypassing the need for complete genome reconstruction [97].

Once contigs are assembled, they must be assigned to their likely source organisms through a process known as binning. This step is essential for recovering individual MAGs and relies on features not used during assembly, such as nucleotide composition, coverage patterns across multiple samples, and in some cases, paired-end or long-range linkage information. Binning tools such as MetaBAT [98], CONCOCT [99], and VAMB [100] implement a variety of unsupervised learning techniques, from modified k-medoid clustering to Gaussian mixture modeling and deep variational autoencoders, to group contigs into bins that represent draft genomes. Some methods also incorporate mate-pair data to improve accuracy [101, 102]. The quality of bins is assessed in terms of completeness and contamination, with high-quality MAGs serving as proxies for uncultivated genomes. These reconstructed genomes enable phylogenetic placement, metabolic reconstruction, and ecological inference for previously uncharacterized taxa, expanding our understanding of microbial life in natural and human-impacted environments.

1.2.4. Taxonomic classification and profiling

Taxonomic classification and profiling are foundational to metagenomic data analysis, enabling identification and quantification of microorganisms within complex microbial communities based on sequence similarity to known reference genomes. Unlike metagenomic assembly approaches, which aim to reconstruct novel genomes de novo from DNA fragments, classification and profiling depends on existing sequence information to determine the identity and relative abundance of taxa in a sample [103]. These analyses are critical not only for characterizing community composition but also for detecting shifts in microbial populations associated with environmental perturbations, disease states, or selective pressures such as antimicrobial exposure.

Two primary strategies are used for taxonomic inference from metagenomic data. Taxonomic classification assigns individual sequencing reads or contigs to specific taxa, typically through a process known as taxonomic binning, and then aggregates these assignments to estimate abundance profiles. In contrast, taxonomic profiling uses the overall sequence composition to estimate the relative abundance of taxa directly. While both approaches rely on comparisons to reference databases, they differ in granularity, computational cost, and susceptibility to database bias.

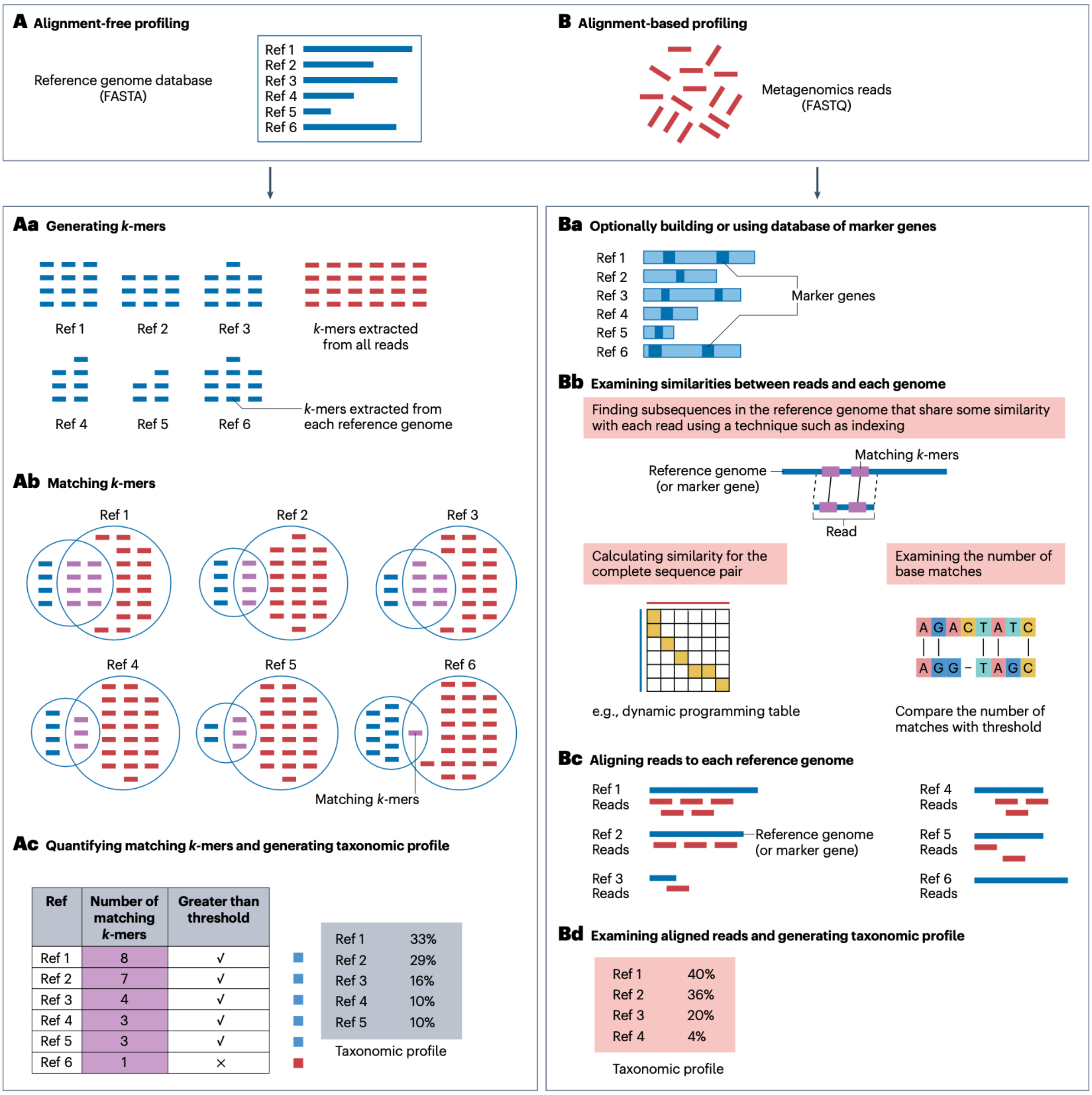
Alignment-based methods remain a cornerstone of taxonomic classification due to their high precision and interpretability (Figure 1.4B). These strategies use alignment tools to map sequencing reads to full reference genomes or curated gene sets, often applying the lowest common ancestor (LCA) algorithm to resolve ambiguous alignments [103]. This is particularly important when dealing with short or low-quality reads that align to multiple closely related genomes. The

MEGAN tool exemplifies this approach by using BLAST [104] or DIAMOND [105] alignments followed by LCA-based binning to classify reads, and it has been adapted for long-read data in MEGAN-LR, which enhances performance on contigs and long reads by incorporating additional alignment context [106].

To reduce computational overhead, marker gene-based methods have been developed, focusing on conserved, single-copy genes that are taxonomically informative. Tools such as MetaPhyler [107] and PhyloSift [108] rely on sets of such genes. MetaPhyler uses 31 protein-coding markers spanning major taxonomic ranks, while PhyloSift extends this to 37 primary gene families plus four auxiliary marker sets covering rRNA, mitochondria, eukaryotes, and viruses, amounting to around 800 gene families. The reduced reference size in these approaches not only enhances computational efficiency but also improves specificity by limiting the influence of redundant or ambiguous genomic regions.

One of the most widely adopted tools in this space is MetaPhlAn [109], which uses a curated database of clade-specific marker genes selected from over two million candidate sequences for their intra-clade conservation and inter-clade uniqueness. MetaPhlAn4 introduces new capabilities, including support for user-defined custom databases derived from metagenome-assembled genomes (MAGs), thus extending its utility to poorly characterized environments and increasing detection sensitivity for novel or rare taxa [109]. Marker-based strategies, such as those implemented in mOTUs, have also enabled the generation of taxonomic units based on universally conserved single-copy genes, facilitating strain-level differentiation even in the absence of full reference genomes. More than 7,700 mOTUs were initially generated using this approach, and over 20,000 additional reference mOTUs were recently integrated based on marker genes derived from over 150,000 MAGs [110, 111]. However, while marker gene methods tend to exhibit higher specificity (i.e., fewer false positives), they can miss organisms not well represented in the marker set and often exhibit reduced sensitivity (i.e., more false negatives) [110].

In parallel, alignment-free methods have emerged as computationally scalable alternatives, particularly valuable in high-throughput applications (Figure 1.4A). These methods decompose reads into short subsequences called *k*-mers and compare them to k-mer databases constructed from reference genomes.


A Alignment-free profiling
Reference genome database (FASTA)
Ref 1
Ref 2
Ref 3
Ref 4
Ref 5
Ref 6
B Alignment-based profiling
Metagenomics reads (FASTQ)
Aa Generating *k*-mers
Ref 1
Ref 2
Ref 3
k-mers extracted from all reads
Ref 4
Ref 5
Ref 6
k-mers extracted from each reference genome
Ab Matching *k*-mers
Ref 1
Ref 2
Ref 3
Ref 4
Ref 5
Ref 6
Matching *k*-mers
Ac Quantifying matching *k*-mers and generating taxonomic profile
Ba Optionally building or using database of marker genes
Ref 1
Ref 2
Ref 3
Ref 4
Ref 5
Ref 6
Marker genes
Bb Examining similarities between reads and each genome
Finding subsequences in the reference genome that share some similarity with each read using a technique such as indexing
Matching *k*-mers
Reference genome (or marker gene)
Read
Calculating similarity for the complete sequence pair
Examining the number of base matches
A G A C T A T C
A G G – T A G C
e.g., dynamic programming table
Compare the number of matches with threshold
Bc Aligning reads to each reference genome
Ref 1 Reads
Ref 2 Reads
Reference genome (or marker gene)
Ref 3 Reads
Ref 4 Reads
Ref 5 Reads
Ref 6 Reads
Bd Examining aligned reads and generating taxonomic profile
Taxonomic profile
Taxonomic profile

Ref	Number of matching *k*-mers	Greater than threshold	
Ref 1	8	√	🟦
Ref 2	7	√	🟦
Ref 3	4	√	🟦
Ref 4	3	√	🟦
Ref 5	3	√	🟦
Ref 6	1	×	🟥

Ref 1	33%
Ref 2	29%
Ref 3	16%
Ref 4	10%
Ref 5	10%

Ref 1	40%
Ref 2	36%
Ref 3	20%
Ref 4	4%

Figure 1.4. Overview of taxonomic profiling strategies in metagenomics (adapted from [1]). (A) Alignment-free profiling reduces computational demand by decomposing both metagenomic reads and reference genomes into *k*-mers (Aa), identifying shared *k*-mers (Ab), and quantifying these matches to generate a taxonomic profile (Ac). (B) Alignment-based methods provide higher sensitivity by aligning reads to reference genomes or marker gene sets. This involves constructing or using a predefined marker database (Ba), identifying read-reference similarities through indexing, dynamic programming, or *k*-mer matching (Bb), aligning reads to references (Bc), and quantifying the alignments to infer taxonomic composition (Bd).

Kraken [112], for example, assigns taxonomic labels using exact *k*-mer matches and a majority-vote scheme, which, while fast, may struggle with precision at lower taxonomic ranks due to shared *k*-mers among closely related taxa. Bracken [113] refines Kraken's output using a Bayesian re-estimation of read distributions across the taxonomic tree, resulting in improved abundance estimation and accuracy.

To enhance computational performance, sketching algorithms such as MinHash compress *k*-mer datasets into smaller, information-rich signatures that retain similarity relationships while drastically reducing memory and runtime requirements [114, 115]. More recent strategies employ discriminative subsets of *k*-mers to optimize classification accuracy and avoid the pitfalls of redundant sequence similarity [116].

In parallel, alignment-free methods have emerged as computationally scalable alternatives, particularly valuable in high-throughput applications. These methods decompose reads into short subsequences called *k*-mers and compare them to k-mer databases constructed from reference genomes.

Kraken [112], for example, assigns taxonomic labels using exact *k*-mer matches and a majority-vote scheme, which, while fast, may struggle with precision at lower taxonomic ranks due to shared *k*-mers among closely related taxa. Bracken [113] refines Kraken's output using a Bayesian re-estimation of read distributions across the taxonomic tree, resulting in improved abundance estimation and accuracy. To enhance computational performance, sketching algorithms such as MinHash compress *k*-mer datasets into smaller, information-rich signatures that retain similarity relationships while drastically reducing memory and runtime requirements [114, 115]. More recent strategies employ discriminative subsets of *k*-mers to optimize classification accuracy and avoid the pitfalls of redundant sequence similarity [116].

Hybrid methods aim to combine the strengths of alignment-based precision with the efficiency of alignment-free classification. Metalign exemplifies this hybrid paradigm by first applying MinHash-based filtering to identify the most likely reference genomes and then performing targeted alignments to assign reads, achieving both speed and high taxonomic accuracy [117, 118]. Such approaches are particularly valuable in resource-constrained or time-sensitive environments where comprehensive classification is required without compromising computational efficiency.

The nature of the sequencing technology used—short-read versus long-read—also influences taxonomic classification outcomes. Short-read platforms such as Illumina generate highly accurate but fragmented sequences, which may not capture enough taxonomic signal for precise classification, particularly at low abundance. In contrast, long-read technologies such as Oxford Nanopore Technologies (ONT) and PacBio HiFi provide extended read lengths that allow for greater specificity and resolution. A benchmarking study evaluating 11 taxonomic classifiers, including five designed for long-read data, demonstrated that long-read classifiers consistently outperformed short-read tools in terms of precision and recall, particularly for low-abundance taxa and at finer taxonomic levels [97]. Tools such as BugSeq [119], MEGAN-LR [106], DIAMOND [105], and sourmash [114] showed robust performance on PacBio HiFi data, detecting species present at just 0.1% relative abundance with high precision. Moreover, long-read data improved classification even without the need for stringent post-classification filtering, which is often necessary for short-read datasets. Nevertheless, long-read classification is still affected by factors such as read length distribution and error rates. For example, datasets dominated by short long-reads (<2 kb) were associated with reduced classification accuracy and biased abundance estimates [62, 97].

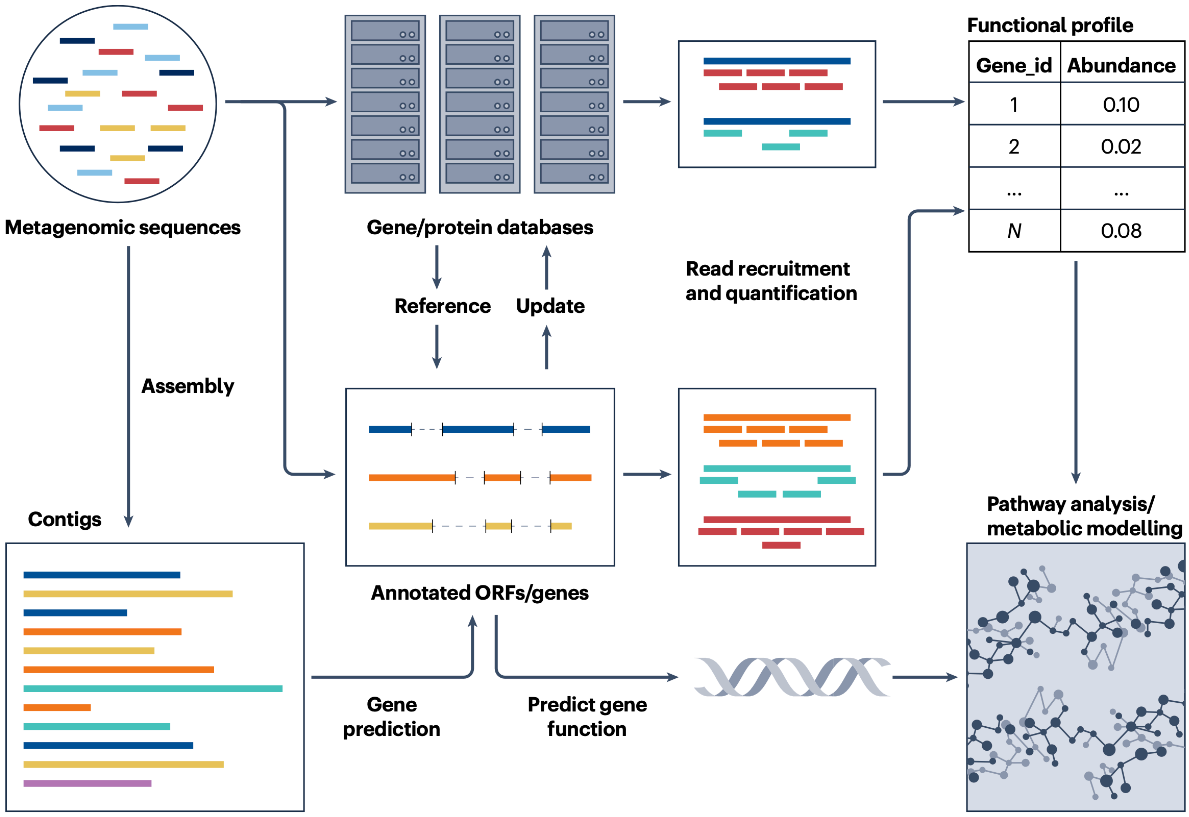
Beyond static profiling, real-time taxonomic classification has become feasible with streaming long-read platforms such as ONT. The ability to analyze reads as they are generated allows for near-instantaneous detection of microbial taxa, bypassing the need for genome assembly or binning prior to classification. This capability is particularly valuable in clinical and public health settings, where timely identification of pathogens can guide treatment decisions and outbreak response [120, 121]. In such contexts, long-read classifiers demonstrate not only high precision but also practical advantages in terms of responsiveness and deployment flexibility.

Together, the growing suite of taxonomic classification and profiling tools offers researchers a diverse set of strategies for interrogating the structure and dynamics of microbial communities. Selection of the appropriate approach depends on the goals of the study, the characteristics of the sample, the available reference data, and the computational resources at hand. When used in concert with complementary methods such as functional profiling and strain-level analysis, taxonomic classifiers form a critical component of modern metagenomics workflows.

1.2.5. Functional analysis of metagenomic data

Functional analysis in metagenomics aims to elucidate the biochemical capabilities and ecological roles encoded in the genomes of complex microbial communities. In the context of urban metagenomes, spanning wastewater, public transit surfaces, hospital effluents, and air microbiomes, this type of analysis provides critical insights into how microbial consortia adapt to anthropogenic pressures and influence environmental and public health. Functional analysis reveals the metabolic potential of urban microbial communities, including their involvement in nutrient cycling, xenobiotic degradation, antimicrobial resistance, and host-associated processes [122, 123]. Rather than focusing solely on which organisms are present, functional analysis characterizes what these organisms can do, thus serving as a more stable and ecologically informative layer of metagenomic interpretation [124, 125].

The conceptual foundations of functional metagenomics date back to early studies that cloned environmental DNA into *Escherichia coli* hosts and screened for specific enzymatic activities without sequencing [126]. These pioneering efforts demonstrated that complex environments harbor extensive, previously uncharacterized biochemical diversity [127–129]. The shift from experimental to computational methods was driven by high-throughput sequencing and the development of large functional gene databases, allowing for genome-wide annotation and pathway-level reconstruction directly from metagenomic reads or contigs.



Functional profile

Gene_id	Abundance
1	0.10
2	0.02
...	...
N	0.08

Figure 1.5. Overview of functional profiling in metagenomic analysis (adapted from [1]). Sequencing reads can be directly mapped to known genes in reference databases to estimate their abundance (top), while *ab initio* methods identify open reading frames (ORFs) for functional assignment (middle). Predicted ORFs are annotated using homology or domain-based approaches, enabling quantification and profiling. Both strategies contribute to the construction of functional profiles, which can be further explored through pathway analysis and metabolic modeling (right).

Unlike taxonomic markers, which vary across individuals and cohorts, microbial functional profiles tend to be more conserved, offering a more reproducible and ecologically relevant perspective on community structure [124]. This is particularly important in urban systems, where spatial and temporal heterogeneity in community composition may obscure biological signals unless functional redundancy and metabolic potential are considered.

Computational functional analysis of metagenomes involves two interrelated steps: functional annotation and functional profiling. Functional annotation assigns putative functions to sequences, often based on homology or domain similarity, while functional profiling quantifies the abundance and distribution of these functions across samples [173–179]. Together, they enable the reconstruction of microbial metabolic capabilities and community-level ecological strategies in built environments.

Annotation typically begins with the identification of open reading frames (ORFs) in assembled contigs or even directly from unassembled reads (Figure 1.5). Gene prediction in metagenomic data is inherently challenging due to short contigs, sequencing errors, and the unknown taxonomic origin of sequences. Tools such as MetaGeneAnnotator [130] and Prodigal [131] have been widely adopted for prokaryotic gene prediction, using features like ribosomal binding site motifs, GC content, codon usage, and hexamer statistics. Deep learning-based models such as CNN-MGP [132] and Meta-MFDL [133] improve sensitivity and specificity in fragmented data by learning high-dimensional representations of sequence patterns.

Predicted ORFs are then subjected to functional annotation using a variety of strategies (Figure 1.5). Homology-based tools such as eggNOG-mapper [134] and KOALA/BlastKOALA [135] align sequences against curated databases including COG [136], KEGG Orthology [137], and Pfam [138], facilitating the assignment of enzymatic functions, pathway membership, and broader functional categories. These tools are particularly useful in urban microbiome studies where many sequences derive from taxa with no cultured representatives; orthology-based mapping leverages evolutionary relationships to extend annotation coverage [139].

Annotation based on protein structure and domain similarity is especially valuable for highly divergent or previously unknown genes. Profile Hidden Markov Models (HMMs), used in tools such as InterProScan [140] and KOfamKOALA [141], match sequences to precomputed domain profiles with adaptive thresholds (Figure 1.5). These methods reduce computational overhead while preserving accuracy and are well suited for annotating metagenomes from highly variable urban environments.

Despite advances in annotation tools, a substantial portion of metagenomic ORFs remains unclassified. This “functional dark matter” highlights the limitations of current reference databases, which are biased toward well-studied, culturable organisms [142]. Approaches like FunGeCo use gene neighborhood context to infer functions from co-located and co-regulated genes, while MG-RAST’s subsystem-based annotation strategy [143] organizes gene families into coherent metabolic modules, improving biological interpretability. Databases such as AGNOSTOS-DB [144] aggregate protein clusters with unknown functions, facilitating the mapping and classification of novel functional elements in urban samples.

Advanced embedding models such as ProtTrans [145] and ProSE [146] apply transfer learning to protein sequences, capturing semantic and structural features that enable function prediction even for highly novel genes. These deep representation learning tools are increasingly integrated into annotation pipelines to improve the discovery of functional novelty in urban metagenomes, where previously unseen proteins are frequently encountered.

Following annotation, functional profiling quantifies the abundance of genes or gene products across samples (Figure 1.5). Profiling tools differ in whether they operate on raw reads, predicted proteins, or gene families, and whether they rely on alignment-based or alignment-free algorithms. MG-RAST [147], for instance, uses BLAST-based alignments to compare predicted proteins against the M5nr database [148], while DIAMOND [105] accelerates this process by using seed-and-extend heuristics and double indexing, making it suitable for large-scale urban datasets. eggNOG-mapper, GhostKOALA, and BlastKOALA integrate taxonomic and functional assignments, enabling joint interpretation of ecological and phylogenetic patterns.

Profile HMM-based profilers such as KOfamKOALA[141] further increase throughput and accuracy by reducing reference database complexity using adaptive score thresholds. Tools like InterProScan [140] annotate protein domains or motifs, allowing for proteomic-level functional inference even when full gene annotation is incomplete. These domain-level annotations are particularly informative in urban samples where mobile genetic elements and horizontally transferred genes may encode truncated or rearranged proteins.

Pathway-based functional analysis extends beyond gene-by-gene annotation to reconstruct metabolic pathways and functional modules (Figure 1.5). Tools such as HUMAnN [149] and gutSMASH [150] map functional annotations to metabolic pathways in curated databases such as MetaCyc [151], KEGG [137], and KBase [152]. These analyses allow for the inference of metabolic interactions across species and within communities, including nutrient fluxes, xenobiotic degradation, and biosynthetic capabilities. Pathway-based functional mapping is particularly relevant in urban contexts where selective pressures such as antibiotics, detergents, and industrial pollutants shape microbial metabolic architecture.

The complexity and interconnectedness of microbial functions in urban environments often require integrative approaches. For example, DrugBank [153] can be used to infer potential microbial contributions to drug metabolism or resistance in urban sewage metagenomes. Tools like MetaErg [154] incorporate secondary features such as signal peptides, transmembrane domains, and subcellular localization to enhance annotation confidence and ecological inference. Despite extensive tool development, no single pipeline offers complete coverage or accuracy. Benchmarking studies indicate that ab initio gene finders often outperform homology-based methods in novel environments but are sensitive to model assumptions and taxonomic biases [155]. Integrative frameworks combining statistical models, HMMs, and deep learning approaches appear to offer the best performance across diverse environments, including the multifaceted landscape of the urban microbiome.

Ultimately, functional analysis of metagenomic data in urban settings provides a lens through which to examine the metabolic resilience, adaptability, and potential risk of microbial communities in human-impacted environments. Whether tracking antimicrobial resistance, pollutant degradation, or host-interactive pathways, functional metagenomics represents a cornerstone for mechanistically understanding the urban microbiome's role in shaping both environmental quality and public health outcomes.

1.3. Metagenome meets pangenome analysis

The integration of metagenomics and pangenomics, termed meta-pangenomics, represents a powerful paradigm for studying microbial communities beyond taxonomic profiles, capturing evolutionary and ecological patterns encoded in genome content variation. While traditional pangenomic analysis has focused on collections of isolate genomes to characterize core and accessory gene pools across microbial species, metagenomics enables the recovery of genomic fragments and metagenome-assembled genomes (MAGs) directly from environmental samples, bypassing the need for cultivation. Their convergence enables high-resolution dissection of gene content dynamics within and across microbial populations in situ.

This integrative approach is particularly relevant for microbial communities inhabiting urban and artificial environments, where selective pressures imposed by infrastructure, pollution, human activity, and antimicrobial exposure drive rapid adaptation. Urban metagenomes—derived from wastewater, transit systems, hospital air, built surfaces, and other engineered ecosystems—reveal microbial assemblages characterized by high turnover, frequent horizontal gene transfer, and complex gene exchange networks. In such settings, traditional single-isolate genome approaches are insufficient to capture the breadth of genetic variability and functional potential.

Meta-pangenomic analysis enables the extraction of gene presence–absence matrices from metagenomic datasets and links them to phylogenetic or ecological frameworks. This allows for the quantification of genome plasticity, identification of lineage-specific adaptations, and inference of gene gain/loss events across community members. By examining how core and accessory genes are distributed within urban microbiomes, this approach provides insight into microbial responses to artificial surfaces, anthropogenic chemicals, and spatial connectivity across infrastructure.

A central challenge in applying pangenomic concepts to metagenomic data lies in accounting for incomplete genomes, strain heterogeneity, and the fragmented nature of assemblies. Methodological innovations, such as phylogenetic gene turnover modeling, probabilistic estimation of gene presence, and quantification of pangenome openness are required to address these limitations.

1.3.1. Urban metagenome deconvolution and MAG reconstruction

A major barrier in microbial ecology has long been the inability to culture the vast majority of environmental microorganisms under laboratory conditions [156]. This limitation arises from the complex and often unknown physiological requirements of many microbial taxa, which may depend on highly specific nutrient regimes, microaerobic or anoxic conditions, or even syntrophic interactions with other microbial partners [157, 158]. As a result, conventional cultivation-based approaches provide only a narrow window into the diversity and function of natural microbial communities, particularly in dynamic and anthropogenically impacted environments such as urban settings.

The advent of metagenomic sequencing has fundamentally transformed this landscape, enabling cultivation-independent access to the collective genetic material extracted from environmental samples [159–161]. In urban environments, ranging from subway surfaces and wastewater to air microbiomes, metagenomics facilitates *in situ* analysis of microbial communities, capturing both abundant and rare taxa, and shedding light on the ecological strategies that underpin microbial survival in built systems [15–17].

Metagenomic sequencing produces complex datasets composed of short reads from hundreds or thousands of taxa. These reads are typically assembled into longer contiguous sequences (contigs), although assembly is often hindered by the presence of conserved genomic regions across related taxa. Advances in metagenome-specific assembly algorithms have improved the resolution and completeness of metagenomic assemblies, even in high-diversity or low-biomass samples. Tools such as IDBA-UD [162], MetaVelvet [163], SOAPdenovo2 [164], ABYSS [165], Khmer [166, 167], Ray-meta [168], MEGAHIT [85], and metaSPAdes [86] are among the widely adopted tools for assembling complex environmental datasets.

Following assembly, metagenome deconvolution involves grouping contigs into genome-level bins based on intrinsic sequence features such as GC content, tetranucleotide frequency, coverage depth, and differential abundance patterns across samples. This process, referred to as binning, has enabled the recovery of metagenome-assembled genomes (MAGs), which represent the composite genome of an uncultivated species or a set of closely related strains [20]. MAGs provide genome-resolved insights into the functional potential, metabolic capabilities, and ecological roles of microbes in natural and artificial ecosystems.

Binning algorithms have undergone significant refinement and have been reviewed extensively [169, 170]. Modern approaches combine multiple features and often include machine learning classifiers to enhance bin accuracy. Nonetheless, the completeness and purity of MAGs depend on sequencing depth, assembly quality, and contamination from host or cohabiting microbial DNA [171–173]. Particularly in urban samples, where complex microbial consortia interact with human-associated and environmental DNA, careful curation of MAGs is essential.

The expansion of public MAG repositories has illuminated a remarkable spectrum of microbial genomic diversity, but the proliferation of incomplete or mis-binned genomes has raised concerns regarding data reliability. Studies have shown that false positives, chimeric assemblies, and redundancy are common among public MAG collections, especially when minimal quality filtering is applied [174]. Consequently, rigorous quality control and validation steps are essential before MAGs are used for downstream ecological or evolutionary analyses.

A suite of tools has been developed for MAG evaluation, including MetaQUAST [175], CheckM [176], MAGpy [177], Anvi'o [178], AMBER [179], and DAS Tool[180]. These tools assess genome completeness, contamination, redundancy, and other quality metrics. Additional strategies such as re-assembly after read recruitment, refinement using coverage profiles, and manual curation are increasingly applied to improve MAG fidelity and ensure that only high-quality genomes are retained for analysis.

Despite their limitations, MAGs have become a cornerstone for genome-centric metagenomics, enabling systems-level interrogation of microbial community structure, population genetics, and functional traits. In urban microbiomes, genome-resolved analyses provide a high-resolution view of microbial adaptation to anthropogenic stressors such as heavy metals, biocides, antibiotics, and nutrient imbalances. High-quality MAGs serve as the foundation for downstream comparative genomics, phylogenetic modeling, metabolic reconstruction, and pangenome analysis—making them indispensable for elucidating microbial dynamics in modern urban ecosystems.

1.3.2. Strain-level resolution and intraspecies diversity in urban metagenomes

Microbial communities in urban environments, such as wastewater systems, built surfaces, and air microbiomes, often consist of complex mixtures of coexisting strains within a single species. This intraspecies diversity is a critical, yet often under-resolved, layer of microbial community structure that can influence ecological dynamics, metabolic capacity, and pathogen evolution [1, 17]. Strain-level heterogeneity has been linked to clinically relevant traits, including host adaptation, antimicrobial resistance, and variation in immune modulation, making it particularly relevant in densely populated and infrastructure-mediated environments [181, 182].

Early genome-resolved studies using gel microdroplet cultivation revealed substantial strain-level variation within the human oral and fecal microbiomes, uncovering near-complete genomes from coexisting subpopulations [183]. Subsequent analyses demonstrated that dominant skin and gut bacterial species exhibit considerable strain-level heterogeneity, often shaped by fine-scale environmental gradients such as pH, moisture, and host-derived compounds [184, 185]. In urban contexts, such micro-scale variation is further modulated by infrastructure type, surface materials, cleaning agents, and human activity patterns—creating distinct ecological niches that may favor specific genotypes.

From a bioinformatics perspective, the presence of multiple coexisting strains presents a significant challenge to metagenome-assembled genome (MAG) reconstruction. Conventional binning algorithms are generally designed to group contigs into species-level bins and lack the resolution to distinguish between closely related strains. This often results in composite bins that aggregate genomic fragments from multiple genotypes, obscuring true strain structure and complicating downstream evolutionary and functional inference[186].

Strain heterogeneity also confounds short-read assemblies by disrupting synteny and collapsing homologous regions, particularly in conserved genes and repeat-rich loci. This leads to fragmented or misassembled genomes and limits the ability to infer genome-wide linkage between variants. Although some strain-level signals can be removed during assembly, residual variation often persists in MAGs, masking strain-specific adaptations and functional signatures.

To address this, several computational frameworks have emerged that aim to resolve strains directly from metagenomic short-read data. Notable tools include StrainPhlAn [187], which uses clade-specific marker genes to track strain variants; ConStrains [188], which infers strain-resolved population structure from SNP patterns; MetaSNV [189], which detects strain variation using single nucleotide variants; and DESMAN [190], which leverages haplotype reconstruction in a probabilistic framework to deconvolve strain-level genotypes. While promising, these tools often rely on high coverage and consistent reference databases, which may be limiting in complex or low-biomass urban samples.

The term "strain" remains inconsistently defined across studies, often used interchangeably with subspecies, clonal type, or genotype. In metagenomic analyses, where cultivation is absent and phenotype data are unavailable, strain identification is typically operational, inferred from genomic similarity or gene content profiles.

Emerging long-read sequencing platforms such as PacBio SMRT and Oxford Nanopore Technologies, as well as chromosome conformation capture techniques like Hi-C, offer the potential to resolve strain structure more robustly. These methods extend read lengths and capture physical linkage across distant genomic regions, improving both assembly contiguity and binning accuracy [191, 192]. While not yet widely implemented in urban metagenomic studies due to cost, technical complexity, and throughput constraints, these technologies represent promising directions for high-resolution microbial ecology.

In urban microbiome research, resolving strain-level variation is not merely a technical refinement—it is essential for understanding microbial adaptation to anthropogenic stressors, tracing the movement of resistance and virulence genes, and developing accurate ecological and epidemiological models. As binning algorithms, assembly techniques, and sequencing technologies continue to evolve, strain-resolved metagenomics will become increasingly integral to genome-centric studies of urban microbial ecosystems.

1.3.3. Meta-pangenome reconstruction from urban metagenomes

The pangenome concept has played a transformative role in microbial genomics, offering a framework to describe and quantify the total genomic repertoire of a species as a combination of shared (core) and variable (accessory or unique) genes (Figure 1.6) [10, 12, 193]. Traditionally, pangenomes are reconstructed using isolate genomes obtained through culture-based methods. While powerful, this approach is inherently limited in scope, as it excludes uncultivable taxa, environmentally rare lineages, and transient or conditionally expressed genes—features that are especially relevant in complex, dynamic environments like urban microbiomes.

Urban environments, characterized by dense human populations, engineered infrastructure, pollutant gradients, and fluctuating microclimatic conditions, exert unique selective pressures on microbial communities [1]. These include exposure to heavy metals, antibiotics, surfactants, and highly variable nutrient conditions, all of which influence microbial adaptation and genome plasticity. In such contexts, the concept of a species' pangenome must be extended to account for both the ecological specificity and the diversity of gene pools that arise from environmental structuring and anthropogenic stress.

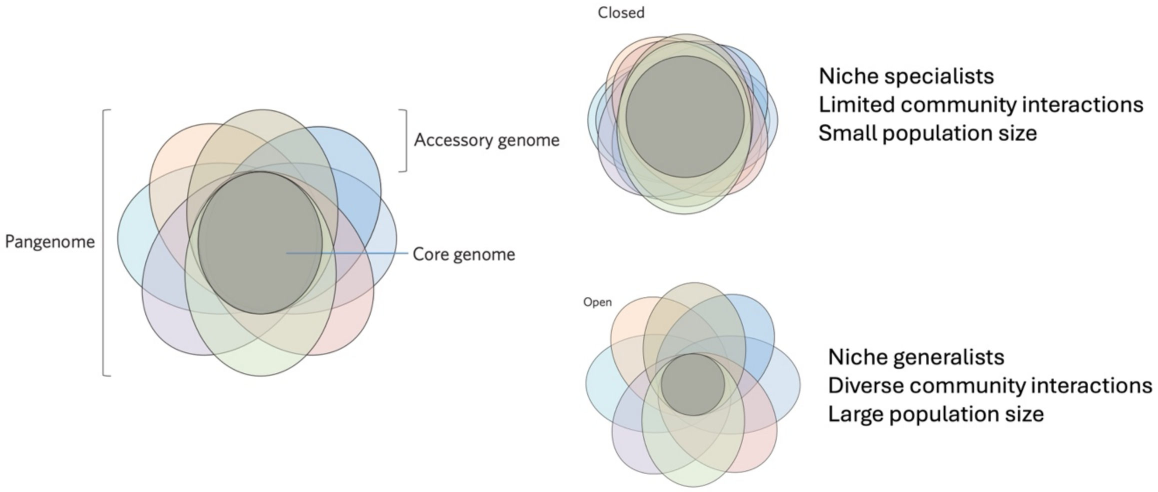


Figure 1.6. The pan-genome concept including the core genome shared by all strains and the accessory genome, which varies across strains. Closed pangenomes have low variability; open pangenomes show extensive gene diversity due to ecological and evolutionary pressures.

To address these complexities, the meta-pangenome concept was introduced , defined as the complete collection of genes observed for a microbial species across all metagenomic and genomic samples from a specific environment [10, 12, 193]. This includes genes from isolate genomes, MAGs, and unbinned contigs, making it a more inclusive and ecologically grounded representation of a species' functional potential in situ. In computational terms, the meta-pangenome reflects the total gene space that can be accessed by a species in a particular environment—capturing both conserved elements and dynamic gene acquisitions shaped by selective gradients and horizontal gene transfer.

The workflow for meta-pangenome reconstruction begins with the assembly of metagenomic reads from multiple samples collected across time points, locations, or environmental compartments (Figure 1.7). Each metagenome is assembled into contigs, from which MAGs are reconstructed using binning algorithms based on coverage, GC content, tetranucleotide frequencies, and differential abundance. Taxonomic assignment of bins is performed to identify contigs associated with the same microbial species across different samples.

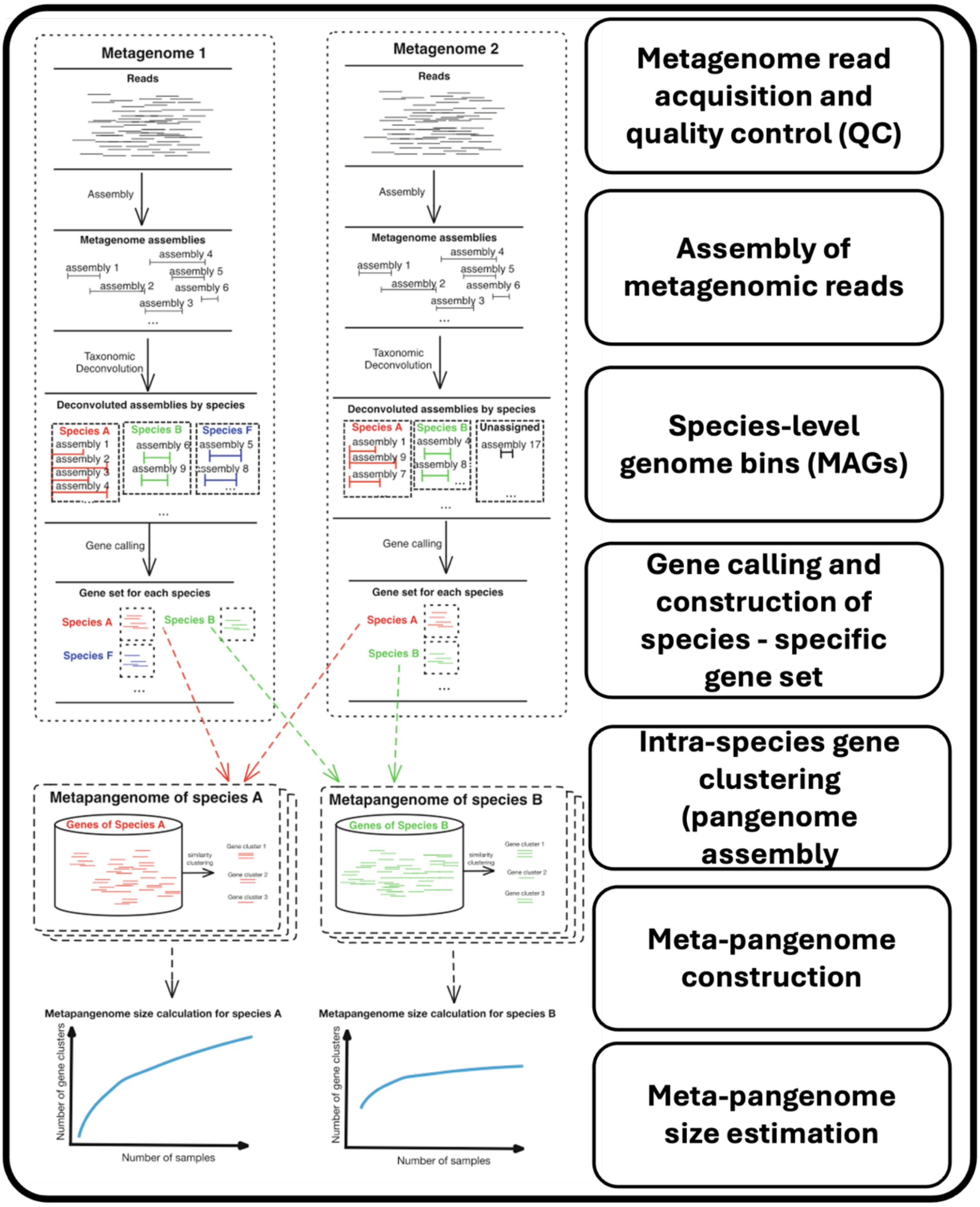


Figure 1.7. Workflow for meta-pangenome reconstruction from multiple metagenomic samples. Metagenomic reads from different samples are independently assembled and taxonomically deconvoluted to generate species-level bins. For each species, predicted genes from assigned assemblies are pooled across samples to generate a non-redundant gene set. These gene sets are clustered into homologous groups to form species-specific meta-pangenomes. Accumulation curves are then generated to assess meta-pangenome openness based on the number of gene clusters discovered as more samples are included.

Subsequently, species-specific contigs and MAGs are subjected to gene prediction using tools such as Prodigal [131], MetaGeneAnnotator [130], or deep learning-based predictors for fragmented contigs [133]. Following quality control (e.g. using CheckM [176], MetaQUASR [175], or DAS Tool [180]), coding sequences (CDSs) are clustered into homologous gene families using sequence similarity thresholds, using tools such as Roary [194], CD-HIT [195], yielding non-redundant gene catalogs per species. These gene cluster are then annotated using reference databases such as eggNOG [196], KEGG [137] or Pfam [138], and stratified into: (1) core genes, consistently detected across nearly all genomes; (2) accessory genes, found only in a subset of metagenomes; and (3) unique genes, present in only one or very few samples (Figure 1.6).

This stratification mirrors the classic pangenome model, but contextualizes it within the spatial, temporal, and environmental variability of urban microbial ecosystems. Core meta-pangenome genes likely reflect essential functions for survival or competitive fitness across diverse urban niches, whereas accessory and unique genes may be linked to local adaptation, resistance, niche partitioning, or recent horizontal gene transfer events.

A key meta-pangenome analysis is the gene cluster accumulation curve, which plots the number of non-redundant gene clusters discovered as a function of the number of samples analyzed [193]. This allows us to assess whether a species' meta-pangenome is closed, where new samples contribute few or no new genes, or open, indicating continual gene discovery and high genomic plasticity. Open meta-pangenomes are commonly associated with generalist species or those experiencing strong selective heterogeneity, such as those in wastewater, industrial sites, or public transit systems. Closed meta-pangenomes are more often observed in specialists confined to stable, restricted niches.

1.4. Pangenome openness-closeness

Pangenome arise from the continual flux of gene content, shaped by the processes of gene gain and loss, with horizontal gene transfer (HGT) serving as a predominant mechanism of gene acquisition in prokaryotes [5, 7, 8, 197]. Once introduced, these genes are subject to drift and natural selection, giving rise to the characteristic architecture of microbial pangenomes (Figure 1.8). One of the most widely observed patterns is the increase in accessory gene content and the decline in core gene content as additional genomes of the same species are sequenced [10, 12, 193]. This pattern reflects the accumulation of rare and niche-specific genes, while genes that are universally conserved become proportionally less dominant.

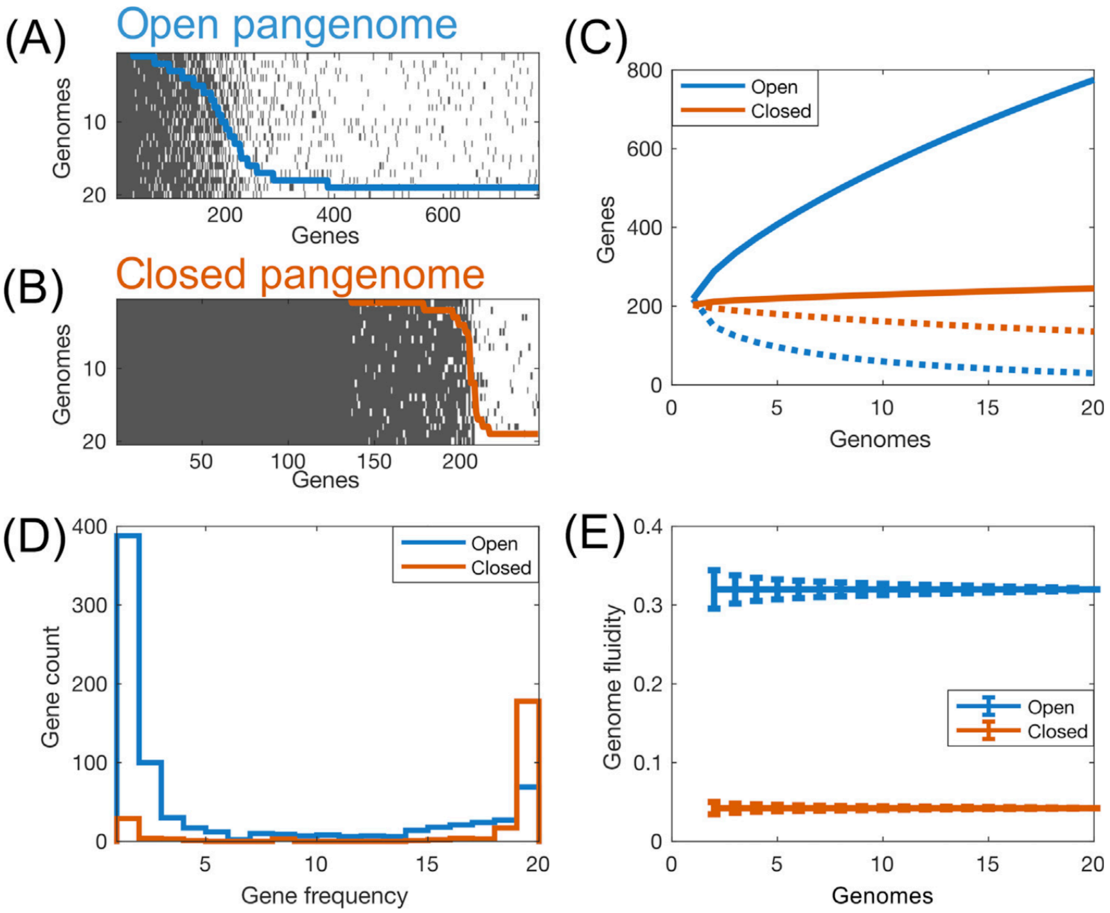


Figure 1.8. Comparative features of open and closed microbial pangenomes. Gene presence-absence matrices for an open (A) and a closed (B) pangenome. Grey indicates gene presence, white indicates absence, solid lines denote gene frequency across genomes. (C) Gene accumulation (solid lines) and core gene depletion (dashed lines) curves for open (blue) and closed (orange) pangenome. (D) Gene frequency distribution show a U-shaped pattern for the open pangenome, with a higher proportion of rare and core genes, while the closed pangenome is enriched for conserved genes. (E) Genome fluidity estimates highlight greater gene content variability in the open pangenome relative to the closed counterpart.

A second hallmark is the emergence of a U-shaped gene frequency distribution (Figure 1.8D), wherein genes tend to be either ubiquitous (core) or rare (unique), with relatively few genes shared by an intermediate number of genomes, reflecting a bimodal structure of microbial pangenomes: one set of genes supporting essential cellular functions, and another set shaped by environmental selection, lateral gene flow, and episodic adaptation [198, 199]. Importantly, the specific shape and parameters of these patterns vary markedly across species. For instance, the rate at which new genes are discovered may plateau rapidly in some taxa, whereas in others it continues to rise even after hundreds of genomes have been sampled [198, 200]. Likewise, the proportion of core genes may remain relatively stable or diminish sharply depending on population structure, genome plasticity, and ecological niche breadth. To capture this variation, pangenomes have been broadly categorized as either open or closed [10, 11, 201].

An open pangenome is characterized by continued gene discovery with each new genome added, suggesting extensive gene flow, high genomic diversity, and environmental heterogeneity. A closed pangenome, by contrast, implies a limited gene pool and relatively stable genome content. Although this binary classification is intuitive, it oversimplifies the underlying dynamics. To provide a more quantitative and scalable measure of pangenome variability, metrics such as genome fluidity have been introduced (Figure 1.8E) [198]. Genome fluidity estimates gene content dissimilarity between genome pairs and enables comparative analyses across taxa or environments, making it particularly useful in studies of diverse and high-turnover ecosystems such as urban microbiomes.

1.5. Pangenome gene gain-loss dynamics

Estimating gene gain and loss across microbial genomes requires integration of phylogenetic information, gene content data, and models of character state evolution. This framework is essential for analyzing genome dynamics and understanding the evolutionary processes that shape pangenome structure. A rooted phylogenetic tree provides the topological and temporal context for comparative analysis [202]. Internal nodes represent ancestral genomes, while branch lengths reflect evolutionary divergence, typically in substitutions per site or time units. The tree serves as the scaffold for modeling gene transitions and reconstructing evolutionary trajectories. Gene presence-absence data are encoded as a binary matrix, where each element indicates whether a gene family is detected in a genome. This matrix represents observed states at the leaves of the tree and is the input for ancestral reconstruction.

Ancestral state reconstruction (ASR) is used to infer gene presence or absence at internal nodes. From these reconstructions, one can derive either (i) counts of gain and loss events by comparing states along branches [27, 203], or (ii) gain and loss rates by integrating transition models with branch lengths [14, 204]. These approaches differ in granularity, notably event counts depend on discrete changes in inferred states, whereas rate estimation assumes a stochastic process and yields branch-specific or lineage-specific rate parameters.

The reconstruction of ancestral gene content presents several challenges, owing to limitations in the available data and model assumptions [205–208]. Despite extensive reviews [206, 207, 209] and empirical comparisons of methods [210–213], the continued misinterpretation of phylogenies indicates that fundamental aspects of ASR require reiteration. Given a reliable inferred phylogeny [202], gene presence and absence states are mapped onto the tips of a rooted tree and optimized to internal nodes using parsimony, maximum likelihood, or Bayesian inference (Table 1.1) [208, 214]. These methods enable inference of gene gain and loss either as discrete event counts or as transition rates when branch lengths are available and modeled explicitly. Each ASR method is assumption-based, regardless of claims to the contrary, however they differ substantially in the processes they consider. Consequently, different methods applied to the same data may yield divergent results [211–213, 215].

Table 1.1. Summary of ancestral state reconstruction (ASR) methods for gene gain-loss inference.

Type	Proprieties	Disadvantages	Ref.
Parsimony	Minimize discrete-state changes; uses step matrices; ignores branch lengths.	No probabilities; underestimates frequent changes.	[206, 216]
Maximum Likelihood	Estimates asymmetric gain/loss rates using CTMCs and branch lengths.	Sensitive to model assumptions; fails if model is misspecified.	[209, 210, 217, 218]
Bayesian Inference	Use CTMCs with priors; estimates posteriors; integrates model/tree uncertainty.	High computational cost; sensitive to priors and model fit.	[219]

1.5.1. Parsimony based gene gain-loss inference

Maximum parsimony is one of the earliest and most widely applied approaches for ancestral state reconstruction, including gene presence–absence data in phylogenetic tree. It seeks the scenario that minimizes the number of state changes, gene gains (0→1) and losses (1→0), required to explain the observed distribution across the tips of a rooted phylogenetic tree [220]. In the context of gene content evolution, this translates into identifying the minimal set of gain-loss events needed to reconstruct the presence-absence states at internal nodes.

The most commonly used algorithm is Fitch parsimony [221], which proceeds in two passes, firstly a postorder traversal to determine sets of possible states at internal nodes based on tip data, followed by a preorder traversal that assigns specific states by minimizing transitions between parent and child nodes. A change is counted when the sets between a node and its child do not intersect. Parsimony can be extended by applying weighted step matrices, such as Wagner [222] or Dollo [223, 224] models, to differentially penalize gains and losses. Dollo parsimony, in particular, enforces a single gain per gene and allows for multiple losses, aligning with models where horizontal transfer or innovation is rare but deletion is frequent [223–225].

Despite its simplicity and computational efficiency, parsimony assumes that all state changes are equally probable and independent of evolutionary time. As such, it does not incorporate branch lengths and underestimates transitions when gene turnover is rapid or heterogeneously distributed across the tree [205, 226]. This limitation leads to systematic biases, particularly on long branches where multiple events are likely to have occurred. Additionally, without a probabilistic framework, parsimony offers no estimates of uncertainty, likelihood, or statistical support for alternative reconstructions [205, 206, 226]. While still used in exploratory settings or as initialization for more complex models [227], parsimony is generally considered suboptimal for modeling gene gain and loss where event rates vary or branch-specific dynamics are of interest. These cases are better addressed by maximum likelihood or Bayesian inference methods that incorporate time-scaled models of character evolution [228].

Parsimony approaches, such as Fitch, Wagner, and Dollo parsimony, aim to infer the minimum number of gain and loss events required to explain the observed distribution of gene presence and absence across a phylogeny. Several tools implement these methods (Table 1.2), applying general-purpose algorithms for ancestral state reconstruction. However, parsimony frameworks are inherently limited in their ability to accommodate rate heterogeneity, account for branch length information, or quantify uncertainty in ancestral assignments.

1.5.2. Maximum Likelihood for gene gain-loss inference

Maximum Likelihood (ML) methods treat ancestral states as parameters and infer their values by maximizing the probability of observing the gene presence-absence data at the tree tips, given a specified evolutionary model and phylogeny [229, 230]. These approaches were initially developed for nucleotide and protein sequences evolution but are equally applicable to discrete binary traits such as gene content evolution as a continuous-time Markov process [14, 31].

Under this model, gene states transition between presence (1) and absence (0) along branches of known lengths, with probabilities defined by a rate matrix Q. For a branch of length t, the transition probability $P(i \rightarrow j, t)$ is computed from the matrix exponential $P(t) = \exp{(Qt)}$. The likelihood of the entire tree is obtained by summing of er all possible internal state assignments, consistent with Felsenstein's pruning algorithm [31]. For an internal node x with descendants y and z, the subtree likelihood is computed as:

$$L_x = \sum_{S_x \in \Omega} P(S_x) \sum_{S_y \in \Omega} P(S_y | S_x, t_{xy}) L_y \sum_{S_z \in \Omega} P(S_z | S_x, t_{xz}) L_z \qquad (1.1)$$

Where we are computing the likelihood of the subtree rooted at node x with direct descendants y and z, S_i denotes the character state for the i-th node, t_{ij} is the branch length (evolutionary time) between nodes i and j, and Ω is the set of all possible character states (for example gene presence or absence). Thus, the objective of ASR is to find the assignment to S_x for all x internal nodes that maximizes the likelihood of the observed data for a given tree.

Table 1.2. Computational tools for ancestral state reconstruction (ASR) and gain-loss (GGL) even estimation, categorized by inference framework.

Method	Name	Notes	Ref.
Maximum Likelihood	**Diversitree**	Contains several classical and contemporary comparative phylogenetic methods, including methods for analyzing trait evolution and estimating speciation/extinction rates.	[236]
	EREM	Estimates the parameters of the model and reconstruct ancestral states (presence and absence in internal nodes) and events (gain and loss events along branches).	[237]
	HyPhy	ASR using an efficient maximum likelihood joint reconstruction method.	[238]
	ape (R)	Implements ML ancestral-state reconstruction for discrete traits, also supports continuous traits	[239]
	BGB	Likelihood framework for ancestral-range (discrete-state), returns node posteriors and event summaries.	[240]
	DEC	Dispersal-Extinction-Cladogenesis model for discrete-state evolution, ML estimation of dispersal/extinction rates and ASR on a fixed tree.	[241]
Bayesian	**BayesTraits**	ASR with Bayesian framework to account for uncertainty in phylogenetic tree and the model of evolution.	[242]
	BayArea	Bayesian MCMC inference of ancestral discrete states – yields full posterior distributions over node states and parameters.	[243]
	BBM (implemented in RASP)	MCMC ancestral-state reconstruction for binary/multistate characters.	[244]
Parsimony	**Count**	Performs ASR and infers family- and lineage-specific characteristics along the evolutionary tree.	[21]
	ANGES	Reconstructs ancestral genome maps.	[245]
	DIVA	Parsimony-based ancestral-state (range) reconstruction minimizing dispersal/extinction costs.	[246]

ML reconstructions can be carried in two ways. In marginal likelihood reconstruction, the most likely state is assigned to each node independently, based on its local likelihood. In joint likelihood reconstruction, the entire set of internal nodes states is optimized simultaneously to maximize the global likelihood. Marginal methods are computationally efficient but may yield suboptimal results due to local maxima, while joint methods are more robust but computationally demanding [231].

ML methods offer improved accuracy over parsimony when gene turnover rates vary across lineages or traits [232, 233]. However, they assume rate constancy over time (i.e., no heterotachy), which can lead to biases when evolutionary rates shift due to ecological or genomic context. Furthermore, ML does not provide uncertainty estimates unless combined with bootstrapping or Bayesian approaches. Distinguishing between rate heterogeneity across characters and rate variation over time remains a challenge [234]. Additionally, ML approach requires model specification and is therefore sensitive to model misspecification [205, 206, 235]. When the likelihood surface is multimodal or highly non-convex, ML may yield unstable point estimates, in which case Bayesian approaches are preferable. Building on the general ML framework, several computational tools have been developed to infer gene gain and loss dynamics from ASR using continuous-time Markov chain (CTMC) models (Table 1.2). These tools implement the core principles of ML estimation while providing model flexibility, support for rate heterogeneity, scalability to large phylogenies.

1.5.3. Bayesian inference for gene gain-loss estimation

Bayesian inference provides a probabilistic framework for ASR by estimating the posterior probability distribution of gene presence or absence at internal nodes, conditional on the observed data and a model of gene content evolution. Unlike ML, which yields a single point estimate, Bayesian methods characterize the uncertainty in ancestral reconstructions and evolutionary parameters by integrating over their joint posterior distributions. Formally, this approach applies Bayes' theorem:

$$P(S|D,\theta) = \frac{P(D|S,\theta)P(S|\theta)}{P(D|\theta)} \propto P(D|S,\theta)P(S|\theta)P(\theta) \qquad (1.2)$$

where S denotes the ancestral states, D is the observed data (e.g. gene presence-absence matrix), and θ encompasses both the phylogenetic tree and the evolutionary model parameters. The likelihood term $P(D|S,\theta)$ is typically computed using Felsenstein's pruning algorithm [31, 226], while the priors $P(S|\theta)$ and $P(\theta)$ reflect assumptions about ancestral states and model complexity, respectively.

Two main variants of Bayesian inference are commonly used. The empirical Bayes approach conditions on fixed estimates of the phylogeny and model parameters, typically derived from maximum likelihood, and computes posterior probabilities of ASR under these assumptions. This method, implemented in software such PAML [247], provides a tractable solution for large datasets by avoiding integration over the space of trees and models, but it assumes no uncertainty in the estimated phylogeny or rates [247].

In contrast, the hierarchical (full) Bayesian approach infers the joint posterior distribution over ancestral states, model parameters, and tree topology. This is typically achieved using Markov chain Monte Carlo (MCMC) sampling, with the Metropolis-Hastings algorithm exploring high-dimensional posterior space [248]. While this method provides a more complete quantification of uncertainty, averaging over plausible trees and models, it is computationally intensive and often restricted to datasets of modest size due to slow convergence in large tree space [249].

Bayesian methods have been used to model gene gain and loss processes explicitly by defining priors over gain-loss rates and integrating over uncertain reconstruction to obtain posterior expectations of event counts. This probabilistic treatment is particularly advantageous when modeling space or noisy gene presence-absence data, or when testing hypotheses about the timing and frequency of gene acquisitions across linages.

Despite their strengths, hierarchical Bayesian methods remain computationally demanding, and their practical advantages over empirical Bayes approaches are debated [250]. Nonetheless, the ability to directly quantify confidence in inferred gene histories makes Bayesian inference a powerful tool for reconstructing genome evolution under uncertainty.

1.5.4. Models for gene gain-loss inference

Many models have been developed to estimate ASR of discrete and continuous characters from extant descendants [211, 251]. These models assume that trait evolution can be modeled as a stochastic process along a phylogeny. For gene gain-loss analysis, the traits of interest are binary (presence or absence of a gene), and their evolution is typically modeled using a continuous-time Markov chain (CTMC).

For discrete case, each gene can occupy one of k possible states, $k = 2$ for presence (1) and absence (0). The CTMC assumes that each state transition to any other with a defined instantaneous rate. The process is memoryless, and transitions are determined by a set of rates $q = \{q_{ij}: 1 \leq i, j \leq k, i, i \neq j\}$, where ach transition follows an exponential waiting time distribution. Once a lineage reaches a given state, transition “clocks” are initiated for each possible destination state, and the next state is selected based on which expires first. These transition rates are typically estimated using maximum likelihood (ML) methods or, alternatively, using Bayesian inference [205].

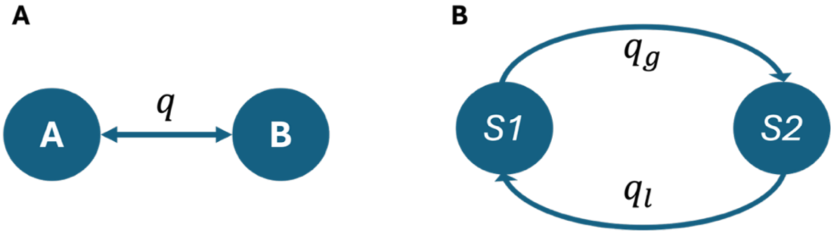


Figure 1.9. Two-state continuous-time Markov chain (CTMC) models used in gene gain-loss inference. (A) One-parameter model, both gene gain and gene loss transitions occur at the same rate q. (B) Two-parameter model, gain and loss transitions occur at independent rates q_g for gene gain and q_l for gene loss.

To estimate the ancestral state of a node v under an ML framework, one first determines the maximum likelihood estimate $\hat{q}$, for the rate matrix. Then, for each possible state at node v, the likelihood is computed given $\hat{q}$, and the state maximizing the conditional likelihood is selected. Alternatively, a Bayesian framework allows integrating over uncertainty by specifying priors over both rates and states and computing the posterior distribution over possible ancestral states.

Because models with many states can results in larger parameter spaces (up to $k(k-1)$ parameters), simplified models are often preferred. For example, the Markov k-state one-parameter model assumes a common transition rate q for all allowed transitions (Figure 1.9A). Certain transitions can be excluded by fixing their rates to zero. The asymmetrical Markov k-state two-parameter model assumes ordered states and restricts transitions to adjacent states, with separated rates q_g for increase and q_l for decrease (e.g., gain and loss) (Figure 1.9B). These simplifications help reduce overfitting, improve numerical stability and facilitate biological interpretability, especially in large-scale pangenomic datasets.

The binary state speciation and extinction (BiSSE) model [251] represents a notable extension of these ideas, jointly estimates binary character evolution and diversification dynamics. While originally applied to macroevolutionary traits, its structure is conceptually relevant for modeling gene family dynamics, where duplications and deletions may be viewed analogous to lineage birth and death events.

1.6. Conclusions to chapter 1

This chapter demonstrates that advances in genome-resolved metagenomics have exposed fundamental limitations in how microbial diversity and function are currently analyzed, particularly in complex environments and *One Health* settings. The following conclusions synthesize these insights and motivate the methodological developments introduced in this thesis:

1. Public-health-relevant microbial variation is primarily determined by gene content and genomic organization rather than by taxonomic identity alone. Approaches based on species profiling therefore fail to resolve the evolutionary and functional context of resistance, virulence, and metabolic traits. Comparative genomic methods operating at the pangenome level are required to capture this variation and to enable function-centric interpretation across related genomes.
2. Metagenomic datasets are inherently fragmented and heterogeneous, combining isolate genomes, MAGs, and unbinned contigs of variable completeness. Existing bioinformatic workflows inadequately account for this structure, leading to biased gene catalogs and unstable comparisons. Robust meta-pangenome inference therefore requires dedicated bioinformatic methods and software capable of integrating heterogeneous genomic evidence while explicitly accounting for uncertainty.
3. Gene gain and loss are central processes shaping microbial adaptation and diversification, yet current pangenome tools lack the capacity to model these processes explicitly within a phylogenetic framework. The absence of scalable software for pangenome-level gene gain–loss analysis prevents the inference of phyletic patterns and lineage-specific dynamics. New phylogeny-aware models and implementations are therefore needed to quantify gene turnover in an evolutionary context.
4. The interconnected nature of environmental, clinical, and agricultural microbiomes demands analytical frameworks that operate consistently across domains. Without genome-resolved, comparative methods, the *One Health* paradigm remains largely conceptual. The development of meta-pangenome-based approaches enables unified analysis of gene flow, functional emergence, and selective pressures, supporting applications in medicine, public health, agriculture, and biotechnology.

These conclusions motivate the thesis’ methodological program: building reproducible, hybrid-aware pipelines for species-level meta-pangenomes; applying phylogeny-informed turnover inference that remains robust to metagenomic uncertainty; and delivering standardized products that translate genome-resolved metagenomics into decision-relevant indicators for surveillance.

2. METHODS FOR MICROBIOME AND VIROME DATA ANALYSIS

To solve the problem of quantifying gene-content variability and its evolutionary drivers in heterogenous microbiome datasets, we developed a modular, reproducible framework for meta-pangenome analysis. The challenge is twofold: (i) assemblies derived from metagenomes (MAGs) are fragmented and uneven, whereas isolate genomes are high quality but culturally biased; and (ii) bacterial recombination and rapid tur nover in accessory genes can obscure vertical signal and inflate false inferences. Our framework addresses these issues by standardizing inputs, enforcing quality thresholds, masking recombination before tree building, and applying explicit probabilistic models of gene gain and loss on a fixed phylogeny (Figure 2.1).

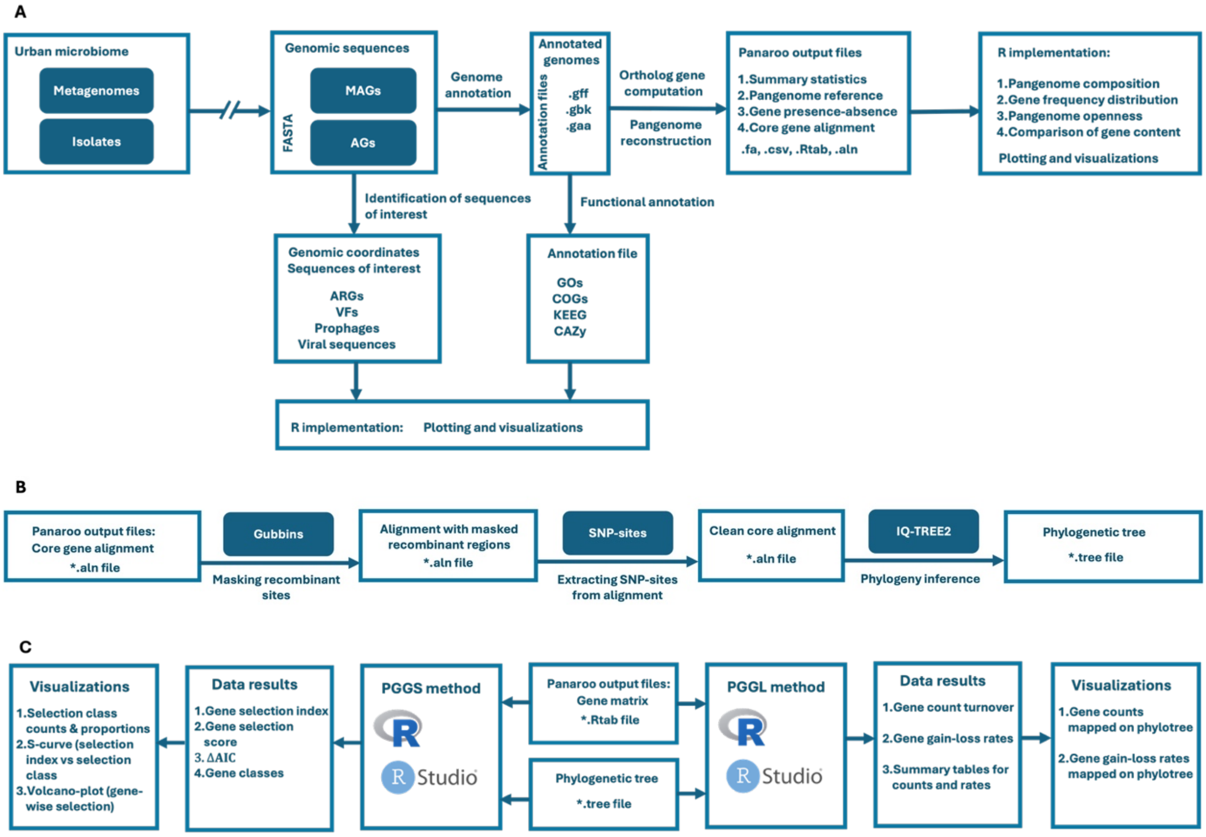


Figure 2.1. Reproducible meta-pangenome workflow and downstream inferences. (A) Genome annotation and meta-pangenome reconstruction; (B) Phylogeny inference; (C) Gene turnover counts (PGGL method) and gene classification according selection index and scores (PGGS method).

Starting from taxonomically curated genomes (MAGs and isolates), we annotate coding sequences and assign functions, cluster orthologs to build a gene presence–absence matrix, and from it derive a recombination-filtered core alignment to infer a maximum-likelihood phylogeny (Figure 2.1A-B). On this scaffold we deploy two complementary inference methods: the PGGL

(Pangenome Gene Gain Loss) method that estimates per-gene gain (λ) and loss (μ) rates under a two-state CTMC using Felsenstein’s pruning, delivering node posteriors and branch-level transition calls that are aggregated along root-to-tip paths to yield per-genome summaries, and the PGGS (Pangenome Gene Selection) then tests for asymmetric turnover by contrasting equal-rates versus all-rates-different models, and quantifies directionality with rate-based indices (log(λ/μ) and a normalized selection score) (Figure 2.1C). In parallel, a uniform virome screen reports fragmentation-robust prophage metrics, including length-weighted viral signal, total viral bp, segment counts, and hallmark-gene density. All components are scripted for parallel execution on HPC with harmonized I/O for downstream statistics and visualization.

2.1.Datasets and software

2.1.1. Empirical datasets

We validated the framework on three *Klebsiella* cohorts spanning environments and taxonomic scales: (i) MPKG (n=64 MAGs from urban environments); (ii) PKG (n≈35 public isolate genomes spanning the genus); and (iii) PKP (n=99 *K. pneumoniae* isolates from the Republic of Moldova). Together, these datasets enable controlled comparisons of MAGs versus isolates, genus-versus species-level processes, and environmental versus clinical contexts within a single, phylogeny-aware analytical protocol.

MPKG dataset

We built the MPKG (Meta-Pangenome for *Klebsiella* genus) dataset from urban-labeled metagenomes extracted from a corpus of 1,023 SRA experiments spanning four peer-reviewed projects: (i) New York City MTA subway samples Metagenome (BioProject: PRJNA271013) [252], which produced the first city-scale mass-transit metagenomes; (ii) metagenomics based spatiotemporal study of Chicago river microbiome (BioProject: PRJNA336577) [253]; (iii) urban waterway sediments in Singapore (BioProject: PRJNA267173) [254]; and (iv) urban metagenomes emphasizing antimicrobial resistance, also in Singapore (BioProoject: PRJNA400857) [255], and available as metagenomic assembled genomes (MAGs) in bioproject “Urban MAG antimicrobial resistance diversity” (BioProject: PRJNA850115) [256]. Of the 3,838 assemblies available in this study, we retained 64 high-quality *Klebsiella* genus MAGs spanning built-environment niches, including subway surfaces, river sediments and wastewater (Table 2.1, Table A1.1). The MPKG dataset comprises 64 urban *Klebsiella* MAGs, dominated by *K. pneumoniae* (n=28), and *K. michiganensis* (n=15), with additional representation from *K. oxytoca* (n=8), *K. variicola* (n=7), *K. aerogenes* (n=3), *K. africana* (n=2), and *K. huaxiensis* (n=1), (Table A5.1).

Table 2.1. Source datasets for the MPKG data, including raw data and curated MAGs

BioProject	Study	Environment/city	Source
Raw data available as FASTQ files			
PRJNA271013	NYC subway metagenomes	Mass-transit surfaces, New York City, US	[47]
PRJNA336577	Chicago river microbiome	Urban river/waterway, Chicago, US	[253]
PRJNA267173	Urban waterway sediments	Tropical urban waterways, Singapore	[254]
PRJNA400857	Antimicrobial resistance of urban water samples	Urban freshwater and wastewater, Singapore	[255]
Curated urban data available as metagenomic assembled genomes (MAGs)			
PRJNA850115	Urban MAG antimicrobial resistance diversity		[253]

PKG dataset

We compiled the PKG dataset from NCBI Assembly (GenBank/RefSeq) [257], restricting to isolate-derived genomes and prioritizing complete genomes. We selected a non-redundant set of five genomes per species, yielding n = 35 genomes across seven *Klebsiella* species: *K. aerogenes, K. africana, K. huaxiensis, K. michiganensis, K. oxytoca, K. penumoniae,* and *K. variicola*. This sampling is intended for genus-level benchmarking and cross-datasets comparisons. Overall, 29/35 (83%) assemblies are complete, the remainder are high-contiguity scaffolds/contigs retained to ensure species representation (Table A1.2).

PKP dataset

We obtained the PKP dataset directly from the National Agency for Public Health (ANSP, Republic of Moldova) as assembled isolate genomes (FASTA, AGs). The set includes 99 clinical *K. pneumoniae* genome assembled sequences collected during 2020-2023 across Moldova's hospitals and ambulatory clinics, with specimen's primary from blood and urine and several cerebrospinal fluid (CSF) cases. These assembled genomes and their metadata were used as input to our unified pangenome and phylogeny-aware gain-loss analyses, full isolate-level identifiers are listed in Table A1.3.

2.1.2. Bioinformatics software

The computational analyses performed in this study utilized a range of open-source software tools for genome annotation, pangenome reconstruction, phylogenetic inference, and gene gain-loss modelling. All tools are freely available, and their sources and intended purposes in the workflow are summarized in Table 2.2.

Table 2.2. Summary of software tools used throughout the study.

Tool/Package	Purpose	Reference
Prokka	Prokaryotic genome annotation	[258]
eggNOG-mapper v.2	Functional annotation	[134]
Panaroo	Pangenome construction and gene presence-absence matrix generation	[259]
IQ-TREE 2	Maximum likelihood phylogenetic tree construction	[260]
Gubbins	Ancestral sequence reconstruction, phylogeny construction and recombination identification	[261]
R (base)	Environment for analysis	[262]
ape (R package)	Phylogenetic tree manipulation and evolutionary analysis	[263]
phytools (R)	Comparative methods and visualization of phylogenetic data	[264]
expm (R)	Matix exponentiation for transition probabilities	[265]
ggplot2 (R)	Data visualization using a layered grammar of graphics	[266]
micropan (R)	Tools for pangenome analysis	[267]
ggtree (R)	Visualization and annotation of phylogenetic trees	[268]
Pheatmap (R)	Heatmap generation (visualizations)	[269]
Phangorn (R)	Phylogenetic analysis	[270]

2.2.Pangenome reconstruction

We implemented a standardized framework for bacterial meta-pangenome analysis, comprising genome annotations, orthologous gene clustering, and core genome alignment. Annotated coding sequences were clustered into orthologous groups to delineate the core and accessory gene content across genomes. This enabled the construction of a gene presence-absence matrix and identification of shared and variable genomic regions. The resulting data formed the basis for subsequent analyses, including phylogenetic reconstruction and characteristic of pangenome structure.

2.2.1. Sequence annotation

All genomes assemblies were annotated with Prokka [258], a widely used prokariotic genome annotation pipeline for speed and NCBI-compliant outputs. Prokka predict coding sequences using Prodigal [131] and includes standard annotation of tRNAs, rRNAs, and other genomic features. For each genome, it produces annotation files in multiple formats (`*.gff`, `.*gbk`, `*.faa`), which were used in downstream pangenome and functional analyses (Listing A2.1). Functional annotation of predicted proteins was performed using eggNOG-mapper v2 [134], which assigns orthologs based on the eggNOG v6 database [196]. This includes a high-coverage annotation of Gene Ontology (GO) terms, KEGG orthology and pathways, Clusters of Orthologous Groups (COG) functional categories, and CAZyme familes. In our analyses, the HMM-based search mode of eggNOG-mapper was enabled to increase annotation sensitivity. This mode uses precomputed Hidden Markov Models (HMMs) to improve ortholog detection, particularly for divergent or poorly characterized proteins. The `*.faa` files generated by Prokka were used as input, and resulting annotations were parsed to support functional enrichment and pangenomic profiling (Listing A2.2).

2.2.2. *Ortholog gene computation and clustering*

Identification of orthologous genes across multiple genomes is a fundamental step in comparative genomics and pangenome analysis. In this study, we used Panaroo [259] for ortholog clustering and pangenome construction (Listing A2.3). Panaroo operates on annotated genome files (`*.gff` from Prokka) and uses CD-HIT [195] for initial sequence similarity searches, followed by graph-based refinement incorporating synteny to correct for assembly and annotation errors [259]. Orthologous genes are clustered using a high sequence identity threshold (98%), and clusters are further merged when supported by conserved genomic context. Panaroo classifies genes into core (present in ≥99% of genomes), soft-core (95-99%), shell (15-95%), and cloud (<15%) categories, reflecting their distribution across the genome set. This classification provides insights into both conserved and variable genomic context. The output includes a gene presence-absence matrix (`*.Rtab`) and a set of reference sequences, as well as a core gene alignment suitable for phylogenetic inference. These files serve as inputs for subsequent analyses, including gain-loss modeling, frequency distribution assessment, and lineage-specific comparisons.

2.2.3. *Pangenome openness-closeness*

To evaluate whether the analysed pangenome is open or closed, we applied a rarefaction-based modelling using the `micropan` R package [267]. This involved generating gene accumulation curves from 1,000 random permutations of genome sampling order and fitting a power-low mode to observed gene discovery rate:

$$\log G(n) = \log k + \gamma \log n, \qquad (2.1)$$

where $G_{(n)}$ represents the cumulative number of distinct orthologous gene families observed after sampling n genomes, k is a constant (the intercept), and γ is the scalling exponent that quantifies the rate of pangenome expansion [193]. Based on this model, the openess-closeness coefficient was computed as $\alpha = 1 - \gamma$. Confidence intervals for α were derived from the standard error of the fitted slope parameter. This framework allows statistical interpretation of pangenome dynamics, where a positive γ indicates an open pangenome, and a near-zero or negative value suggests a closed structure.

2.3.Gene gain and loss model

We developed and implemented a maximum likelihood framework for inferring gene gain and loss from binary presence-absence profiles across a phylogeny. For each orthologous group, the method aligns the gene vector to the tip labels of a rooted, bifurcating tree and computes the likelihood of the observed data under a continuous-time Markov chain model (CTMC) [271] using Felsenstein's pruning algorithm [31]. Transition rates are estimated by minimizing the deviance via bounded quasi-Newton optimization [272] and standard errors are derived from the inverse Hessian of the deviance [273], approximated numerically to assess parameter uncertainty. Ancestral states are reconstructed by combining post-order likelihoods with pre-order upward messages, yielding marginal posterior probabilities for each internal node [14]. Branch-specific gain and loss events are inferred by computing the difference in posterior presence probabilities between child and parent nodes. Branches are classified as gains or losses, enabling confidence-aware event detection. When analyzing multiple genes, branch-level events are aggregated along root-to-tip paths to produce genome-wide summaries of gain and loss dynamics, facilitating comparative analyses across lineages.

2.3.1. Phylogenetic tree based on core genome alignment

To reconstruct the evolutionary relationships among genomes, we performed phylogenetic inference based on core genome alignment obtained from Panaroo [259]. This alignment comprises nucleotide sequences of single copy orthologs shared by the majority of genomes in the dataset (typically $\geq 95\%$), providing a robust scaffold for phylogenetic analysis.

Homologous recombination, which is widespread in bacterial genomes, can obscure vertical inheritance signals and mislead phylogenetic inference. To account for this, we used

Gubbins [261] to identify and mask recombinant regions in the alignment (Listing A2.4). Gubbins iteratively detects recombination by modeling elevated densities of base substitutions along the alignment using a phylogenetic framework. The output is a recombination-filtered alignment, retaining only vertically inherited single nucleotide polymorphisms (SNPs) for tree construction.

The filtered core alignment was used to construct a phylogenetic tree with IQ-TREE 2 [260], which implements an efficient maximum likelihood (ML) framework (Listing A2.4). IQ-TREE automatically detects the best-fit nucleotide substitution model via ModelFinder [274] and supports ultrafast bootstrap approximation (UFBoot) for statistical branch support. We applied 1,000 UFBoot replicates to assess clade reliability and inferred the final tree using the optimized ML topology under the selected model. The resulting phylogeny (Newick format) was midpoint-rooted using the `phangorn` package in R [270] and visualized with `ggtree` [268].

2.3.2. Probabilistic inference of gene gain and loss events using a continuous-time Markov chain (CTMC) model

To identify evolutionary transition in gene content across a phylogeny, we developed a method PGGL (Pangenome Gene Gain-Loss), implemented in R [38], that integrates ancestral state reconstruction with probabilistic gain-loss detection procedure. The inference is grounded in a continuous-time Markov chain (CTMC) model for binary traits [14], specifically modeling the presence or absence of each gene across the nodes of a rooted phylogenetic tree.

Phylogenetic representation

Let phylogenetic tree be defined as $T = (V, E, l)$, where $V = \{1,2, \dots ,2n - 1\}$ is a set of all nodes, $E \subset V \times V$ is the set of directed edges $(u \rightarrow v)$, representing parent-to-child relationships, $l: E \rightarrow \mathbb{R}_{>0}$ assigns a positive branch length l_{uv} to each edge, interpreted as expected evolutionary time, $V_{tip} \subset V$ denotes the set of n tip nodes (observed taxa), and $V_{int} = V \backslash V_{tip}$ denotes the internal (ancestral) nodes.

Binary-state CTMC model of gene content evolution

Each gene is modeled independently as a time-homogeneous CTMC on the tree T [271], with binary state space $S = \{0,1\}$, where 0 denotes gene absence and 1 denotes presence. The infinitesimal generator (rate matrix) of the process is:

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \tag{2.2}$$

with $\lambda = q_{01} > 0$ denoting gain rate, and $\mu = q_{10} > 0$ denoting loss rate. The stationary distribution $\pi = (\pi_0, \pi_1)$ is:

$$\pi_0 = \frac{\mu}{\lambda+\mu}, \pi_1 = \frac{\lambda}{\lambda+\mu}, \qquad (2.3)$$

and $\gamma = \lambda + \mu$ is the total transition rate. The stationary variant was not explicitly implemented because it introduces an additional constraint on the root distribution. Instead, the analysis was conducted under the general formulation of the model, which does not assume a priori evolutionary equilibrium and allows the parameters to be estimated directly and without constraint from the data.

Transition probabilities and the Kolmogorov forward equation

Given a branch of length t, the transition probability matrix is derived as $P(t) = e^{Qt}$ [14, 275, 276], yielding:

$$P(t) = e^{Qt} = \begin{bmatrix} \pi_0 + \pi_1 e^{-(\lambda+\mu)t} & \pi_1(1 - e^{-(\lambda+\mu)t}) \\ \pi_0(1 - e^{-(\lambda+\mu)t}) & \pi_1 + \pi_0 e^{-(\lambda+\mu)t} \end{bmatrix}. \qquad (2.4)$$

These expressions are the solution to the Kolmogorov forward equation:

$$\frac{d}{dt} P(t) = QP(t), \qquad (2.5)$$

with initial condition $P(0) = I$, the identity matrix [271, 277].

The assumption of exponentially distributed waiting times between gene gain and loss events is grounded in the classical convergence theorem established in 1963 [278]. This theorem rigorously demonstrates that the superposition of a large number of independent stochastic processes, each characterized by infinitesimal event intensity, converges in distribution to a Poisson process. In the biological context, this corresponds to the accumulation of numerous rare and independent gene turnover events, such as insertions, deletions, or horizontal gene transfers, acting along evolutionary lineages. In this asymptotic regime, event occurrences become statistically independent, and the inter-event times follow an exponential distribution, the only continuous distribution satisfying the memoryless property. Consequently, the limiting process conforms to the axioms of a continuous-time Markov chain (CTMC). This provides the formal theoretical foundation for representing gene content evolution as a CTMC with transition probability matrix $P(t) = e^{Qt}$ where Q is the infinitesimal generator whose off-diagonal elements denote the instantaneous rates of gene gain (λ) and loss (μ).

Likelihood estimation via Felsenstein's pruning algorithm

To compute the likelihood of the observed gene presence-absence pattern $x \in \{0,1\}^n$, I empolyed the pruning algorithm introduced by Felsenstein [226]. The algorithm recursively computes the conditional likelihood at each internal node, given a hypothesized state. The likelihood at a node v, denoted $L_v(s)$, represents the probability of obsering all data below the node v, assuming v is in state $s \in \{0,1\}$.

The recursion begins at the tip nodes, where the observed states are known. For a tip node a tip node $t \in V_{tip}$, the conditional likelihood is defined as:

$$L_t(s) = \begin{cases} 1, & \text{if } s = x_t \\ 0, & \text{otherwise} \end{cases} \qquad (2.6)$$

where $x_t \in \{0,1\}$ is the observed gene presence (1) or absence (0) in taxon t. This initialization reflects a deterministic assignment, the likelihood is 1 if the assumed state matches the observed data, and 0 otherwise.

For an internal node $v \in V_{int}$ with child nodes j and k, we compute the conditional likelihood $L_v(s)$ for each state $s \in \{0,1\}$ by summing over all possible state assignment at the child node. Formally, the recursion is given by:

$$L_v(s) = \sum_{s_{j\in\{0,1\}}} P_{s\to s_j}(l_{vj}) L_j(s_j) \cdot \sum_{s_k\in\{0,1\}} P_{s\to s_k}(l_{vk}) L_k(s_k) (2.7)$$

where l_{vj} and l_{vk} are the branch lengths from node v to children j and k, $P_{s\to s_i}(t)$ is the CTMC transition probability from state s to s_i over time t (l_{vj} and l_{vk}), and $L_j(s_j)$ and $L_k(s_k)$ are the likelihoods at child nodes (j and k).

Denoting $t_{vj} = l_{vj}$ branch length from v to j, and $t_{vk} = l_{vk}$ branch length from v to k, we use the Kolmogorov-derived transition probabilities,

$$P_{00}(t) = \pi_0 + \pi_1 e^{-(\lambda+\mu)t} \qquad (2.8)$$

$$
\begin{aligned}
P_{01}(t) &= \pi_1(1 - e^{-(\lambda+\mu)t}) \\
P_{10}(t) &= \pi_0(1 - e^{-(\lambda+\mu)t}) \\
P_{11}(t) &= \pi_1 + \pi_0 e^{-(\lambda+\mu)t}
\end{aligned}
$$

to expand the two cases ($X_v = 0$, and $X_v = 1$):

$$
\begin{aligned}
L_v(0) &= [P_{00}(l_{vj}) \cdot L_j(0) + P_{01}(l_{vj}) \cdot L_j(1)] \cdot [P_{00}(l_{vk}) \cdot L_k(0) + P_{01}(l_{vk}) \cdot L_k(1) \\
&= [(\pi_0 + \pi_1 e^{-\gamma l_{vj}}) \cdot L_j(0) + \pi_1(1 - e^{-\gamma l_{vj}}) \cdot L_j(1)] \cdot \\
&\quad [(\pi_0 + \pi_1 e^{-\gamma l_{vk}}) \cdot L_k(0) + \pi_1(1 - e^{-\gamma l_{vk}}) \cdot L_k(1)],
\end{aligned}
\tag{2.9}
$$

and

$$
\begin{aligned}
L_v(1) &= [P_{10}(l_{vj}) \cdot L_j(0) + P_{11}(l_{vj}) \cdot L_j(1)] \cdot [P_{10}(l_{vk}) \cdot L_k(0) + P_{11}(l_{vk}) \cdot L_k(1) \\
&= [\pi_0(1 - e^{-\gamma l_{vj}}) \cdot L_j(0) + (\pi_1 + \pi_0 e^{-\gamma l_{vj}}) \cdot L_j(1)] \cdot \\
&\quad [\pi_0(1 - e^{-\gamma l_{vk}}) \cdot L_k(0) + (\pi_1 + \pi_0 e^{-\gamma l_{vk}}) \cdot L_k(1)].
\end{aligned}
\tag{2.10}
$$

Once the upward recursion reaches the root node ρ, we obtain $L_\rho(0)$ and $L_\rho(1)$, the conditional likelihoods assuming the root is in each possible state. Since the true ancestral state at the root is unknown, we marginalize over both possibilities using the stationary probabilities:

$$
\mathcal{L}(\lambda, \mu | x) = \pi_0 \cdot L_\rho(0) + \pi_1 \cdot L_\rho(1), \tag{2.11}
$$

where x refers to the observed data at the tips of the tree (the vector of gene presence (1) or absence (0)) across all extant taxa. This final likelihood represents the probability of the observed gene presence-absence data under the specified CTMC model and parameter values and is used as the objective function in maximum likelihood (ML) estimation, $f(\alpha, \beta) = -\ell(e^{\alpha},\ e^{\beta})$, and serves as a basis for statistical model comparisons and ancestral reconstructions. Parameters (λ, μ) were estimated by maximizing the pruning-based log-likelihood $\ell(\lambda, \mu) = \log\{\pi_0 L_\rho(0) + \pi_1 L_\rho(1)\}$ [31], equivalently minimizing the deviance $D(\lambda, \mu) = -2\ell(\lambda, \mu)$. To ensure positive rates, we optimize on the log scale $\alpha = \log(\lambda)$ and $\beta = \log(\mu)$ and minimize $D(\alpha, \beta)$ with bound-constrained L-BGFS-B [272]. After convergence, estimates are back transformed to the natural scale ($\hat{\lambda} = \exp \hat{\alpha}$, $\hat{\mu} = \exp \hat{\beta}$).

Uncertainty was quantified from the observed information, defined as the inverse Hessian of the negative log-likelihood at the MLE. Let $\theta = (\alpha, \beta)^T$ with $\alpha = \log \lambda$ and $\beta = \log \mu$. The observed information matrix is $I(\hat{\theta}) = \nabla^2[-\ell(\alpha, \beta]|_{\hat{\theta}}$, so $Var(\hat{\theta}) = I(\hat{\theta})^{-1}$. Back transforming to rates $\hat{\lambda} = e^{\hat{\alpha}}$ and $\hat{\mu} = e^{\hat{\beta}}$ the delta methods give $Var(\hat{\lambda}) \approx \hat{\lambda}^2 Var(\hat{\alpha})$, $Var(\hat{\mu}) \approx \hat{\mu}^2 Var(\hat{\beta})$, and $Cov(\hat{\lambda}, \hat{\mu}) \approx \hat{\lambda}\hat{\mu} Cov(\hat{\alpha}, \hat{\beta})$; hence $SE(\hat{\lambda}) \approx \hat{\lambda} SE(\hat{\alpha})$ and $SE(\hat{\mu}) \approx \hat{\mu} SE(\hat{\beta})$ [267]. We report these uncertainties to quantify the precision of $\hat{\lambda}$ and $\hat{\mu}$, enable confidence intervals and hypothesis tests (e.g. $\lambda = \mu$ versus $\lambda \neq \mu$), and to gauge the reliability of gene-level inferences rather than relying on point estimates alone.

Posterior probabilities at internal nodes

Following likelihood maximization of (λ, μ), we compute node marginal posterior probabilities of each state at each internal node using Bayes' rule [277], by combining the upward conditional likelihoods form pruning (L_v) with downward (parent-to-child) message (F_v). For node v and state $s \in \{0,1\}$,

$$\Pr(X_v = 1 | x_{tips}) = \frac{\pi_1 \cdot L_v(1) \cdot F_v(1)}{\pi_0 \cdot L_v(0) \cdot F_v(0) + \pi_1 \cdot L_v(1) \cdot F_v(1)}, \qquad (2.12)$$

$$\Pr(X_v = 0 | x_{tips}) = \frac{\pi_0 \cdot L_v(0) \cdot F_v(0)}{\pi_0 \cdot L_v(0) \cdot F_v(0) + \pi_1 \cdot L_v(1) \cdot F_v(1)},$$

or in generalized form:

$$\Pr(X_v = s | x_{tips}) = \frac{\pi_s \cdot L_v(s) \cdot F_v(s)}{\sum_{s' \in S} \pi_{s'} \cdot L_v(s') \cdot F_v(s')}. \qquad (2.13)$$

At the root, $F_{root}(s) = 1; \pi$ is uniform or stationary prior. This results in a matrix $\mathcal{A} \in [0,1]^{m \times 2}$ where m is the number of internal nodes and each row gives the marginal posterior over states $\{0,1\}$.

Pangenome gain-loss rates inference

To infer transition on branches, I evaluated the posterior probability state 1 at both parent and child nodes of each edge $e = (u \to v)$. We define the change in presence probability as:

$$\Delta_e^{(g)} = \mathbb{P}(X_v = 1) - \mathbb{P}(X_u = 1). \tag{2.14}$$

The $\Delta_e^{(g)}$ represents the change in the posterior probability of the gene being in state "1" form the parent to the child node of edge e. We classify the events as: (i) gain if $\Delta_e^{(g)} > \delta$, (ii) loss if $\Delta_e^{(g)} < -\delta$, (iii) no changes if $|\Delta_e^{(g)}| \leq \delta$ and (iv) unknown if either posterior is undefined. We use a threshold for $\delta \in [0,1]$, $\delta = 0.05$, to determine whether a change has occurred along a branch for a particular gene g on edge e. So, if $-\delta \leq \Delta_e^{(g)}$, i.e. the change is very small (within the threshold), then we classify it as no change, or when $\Delta_e^{(g)} \in (-\delta, \delta)$, the change is not significant enough to be called a gain or loss, and it is labeled as no change. This acts as a buffer zone that filters out noise or uncertainty in the reconstruction process.

To summarize gene gain and loss events at the level of individual genomes (tips), we trace the root-to-tip path for each genome in the tree and aggregate all gain-loss events that occurred on that path. Given a set of per-gene branch events $\left\{\Delta_e^{(g)}\right\}_{g=1}^{G}$, we compute per-genome counts:

$$Gain_j = \sum_{g=1}^{G} \sum_{e \in \mathcal{P}_j} \mathbb{I}\left(\Delta_e^{(g)} > \delta\right), \tag{2.15}$$

$$Loss_j = \sum_{g=1}^{G} \sum_{e \in \mathcal{P}_j} \mathbb{I}\left(\Delta_e^{(g)} < -\delta\right), \tag{2.16}$$

where $\mathcal{P}_j$ is the set of branches on the path from root to tip j, and $\mathbb{I}(\cdot)$ is the indicator function. This yields a matrix summarizing the total number of inferred gene gains and losses per genome. Similarly, per-node summaries (internal or ancestral) are computed by aggregating all events leading into each node. These summaries are useful for identifying clade-specific expansions, contractions, or evolutionary bottlenecks.

2.3.3. Gene-wise estimation of selection pressure in pangenomes

Having established a probabilistic model for gene gain and loss using a continuous-time Markov chain (CTMC) model (Section 2.3.2), we next sought to quantify the degree to which evolutionary transitions in gene content are biased toward gain or loss. This is conceptually linked to directional selection acting on gene retention or deletion. For this purpose, we implemented a comparative likelihood framework that tests for asymmetry in transition rates and defines quantitative indices of evolutionary bias for each gene.

This procedure, implemented in the PGGS (Pangenome Gene Selection) classification method, that evaluates each gene g independently, using its observed binary presence-absence pattern $x^{(g)} \in \{0,1\}^n$ across the tips of the phylogenetic tree T. The analysis builds on the maximum-likelihood framework established in the previous section, reusing the CTMC likelihood function computed via Felsenstein's pruning algorithm [31, 271].

To implement this framework, we compare two nested models of gene evolution: (1) the first model (ER, Equal-Rates) assumes symmetric transition: the gain and loss rates are equal ($\lambda = \mu = q$), and (2) the second model (ARD, All-Rates-Different) allows distinct gain and loss rates [14, 271]. Formally, the infinitesimal generator matrices are:

$$Q_{ER} = \begin{bmatrix} -q & q \\ q & -q \end{bmatrix}, Q_{ARD} = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \qquad (2.17)$$

For each gene g, the corresponding CTMC likelihood $\mathcal{L}(\theta | x^{(g)}, T)$ is computed under both models. This yields MLE results, including $\hat{q}^{(g)}$ under ER model, and $(\hat{\lambda}^{(g)}, \hat{\mu}^{(g)})$ under the ARD model.

Next to determine whether the asymmetric model (ARD) better explains the observed data, we employed the Akaike Information Criterion (AIC) as an alternative model selection strategy. AIC evaluates each model's goodness-of-fit while penalizing model complexity, and is defined as:

$$\text{AIC} = 2k - 2\log\hat{L} \qquad (2.18)$$

where *k* is the number of free parameters, and $\hat{L}$ is the maximum likelihood under the model [279]. For the ER model (with one free parameter, *q*), AIC is computed as:

$$AIC_{ER} = -2\log L_{ER} + 2 \qquad (2.19)$$

while for the ARD model (with two parameters, λ and μ), it is

$$AIC_{ARD} = -2\log L_{ARD} + 4. \qquad (2.20)$$

A lower AIC indicates a more parsimonious model that balances explanatory power with model simplicity. For each gene, the two AIC values are compared, and the model with the lower AIC is considered better supported. When the difference in AIC ($\Delta AIC = AIC_{ER} - AIC_{ARD}$) exceed 2, the

ARD model is interpreted as significantly superior, implying asymmetric turnover dynamics. In contrast, small ΔAIC values (e.g., | $\Delta AIC < 2$|) suggest comparable fit, favoring ER model under the principle of parsimony [280].

Following parameter estimation under the ARD model, we define the Selection Index log-ratio (SI) metric to quantify the asymmetry in gene turnover:

$$SI = log\left(\frac{\hat{\lambda}^{(g)}}{\hat{\mu}^{(g)}}\right). \tag{2.21}$$

This index reflects the fold-change between gain and loss rates on a log scale. Positive values indicate that gains are more frequent than losses, while negative values suggest loss-dominated evolution.

Additionally, we define the Selection Score (SS) metric based on normalized difference of estimates:

$$SS_g = \frac{\hat{\lambda}^{(g)} - \hat{\mu}^{(g)}}{\hat{\lambda}^{(g)} + \hat{\mu}^{(g)}} \in (-1,1). \tag{2.22}$$

This bounded, symmetric score facilitates direct comparison across genes with very different turnover dynamics. A value of 0 indicates balanced gain-loss rates, while values approaching +1 or -1 indicate directional gain or loss pressure, respectively.

To support interpretably and downstream comparative analysis, we discretize the continuous SS score into qualitative selection classes (SC). This categorization reflects increasing levels of gain- or loss-bias:

$$SC_g = \begin{cases} strong\ loss, & if\ SS_g \in [-1.0, -0.8] \\ moderate\ loss, & if\ SS_g \in [-0.8, -0.5] \\ neutral, & if\ SS_g \in [-0.5, 0.5] \\ moderate\ gain, & if\ SS_g \in [0.5, 0.8] \\ strong\ gain, & if\ SS_g \in [0.8, 1.0] \end{cases} \tag{2.23}$$

The classification scheme provides a biologically meaningful interpretation of gene turnover asymmetry. It allows for stratification of genes into functional categories (e.g., core, accessory, adaptative), facilitates visualization of evolutionary patterns across species or environments.

2.4. Simulation of phylogenetic trees with known ancestral states

A rooted phylogenetic tree with 101 and 1001 terminal taxa was simulated under a Yule (pure-birth) process using `rtree` function form the `ape` package [281]. The simulated trees were checked and, if necessary, modified to ensure that they are rooted, strictly bifurcating, and contained only positive branch lengths. Missing or non-positive branch lengths were replaced with a small positive constant (1×10^{-8}) to ensure numerical stability in downstream analyses.

Binary orthologous gene evolution was modeled as a two-state continuous-time Markov process with states {0, 1}, parametrized by an asymmetric rate matrix allowing distinct gain (0→1) and loss (1→0) rates. The root state was fixed to state 0, thereby defining a known ancestral condition at the base of the tree.

Complete evolutionary histories of the orthologous gene, including ancestral states at all internal nodes and state transitions along branches, were generated using stochastic character mapping under the specified Markov model [282]. Stochastic simulation and subsequent extraction of true tip node states were performed using the `phytools` package, yielding ground-truth ancestral states by construction. Branch lengths were interpreted as evolutionary time units under the continuous-time model. All simulations were performed with a fixed random seed to ensure reproducibility.

2.5. Ancestral state reconstruction and benchmarking

To assess the accuracy of the developed PGGL maximum likelihood ancestral state reconstruction method, we designed a benchmarking analysis based on simulated phylogenetic data with known ground-truth ancestral states. Ancestral states for a binary orthologous gene (states {0, 1}) were reconstructed from observed tip states using three approaches: (1) Fitch parsimony [283], maximum likelihood (ML; PGGL), and Bayesian stochastic character mapping [219], and the inferred states were directly compared to the true simulated ancestral states.

Parsimony-based ancestral state reconstruction was performed using the Fitch algorithm for unordered binary characters [283]. Tip states were propagated toward the root using a postorder traversal of the tree, combining descendant state sets by intersection when possible or by union otherwise. Internal nodes were classified as unambiguously assigned to state 0 or 1, or as ambiguous when both states were equally parsimonious. This approach provided a non-parametric baseline for comparison.

Bayesian ancestral state reconstruction was performed using stochastic character mapping, implemented with the `make.simmap` function from the `phytools` R package [219, 284]. Character histories were sampled conditional on the observed tip states under a continuous-time Markov model, using either the true transition rate matrix from simulation or rates estimated from the data under an ARD Mk model [284, 285]. Posterior state probabilities at internal nodes were obtained by summarizing across multiple stochastic maps, and the modal posterior state was used as the Bayesian point estimate.

For all three approaches, reconstructed ancestral states were aligned to internal node identifiers and evaluated by direct comparison with the known simulated ancestral states. ML and Bayesian reconstructions were classified as correct or incorrect (matches and mismatches), while parsimony-inferred ambiguous nodes were tracked separately (matches, mismatches and ambiguous).

2.6. Conclusions to chapter 2

This chapter produces a set of bioinformatic software components, algorithms, and methodological outputs that enable quantitative, phylogeny-based meta-pangenome analysis from heterogeneous metagenomic data. These contributions can be summarized as follows:

1. A unified meta-pangenome analysis software framework was developed and implemented that integrates isolate genomes and metagenome-assembled genomes through standardized curation, annotation, and ortholog clustering, resolving inconsistencies arising from heterogeneous genome quality and enabling reproducible pangenome reconstruction and comparison across mixed genomic inputs.
2. A recombination-filtered maximum-likelihood species phylogeny is inferred and used as a fixed evolutionary reference for gene-content analyses, restoring explicit evolutionary context and enabling lineage-resolved comparison of gene presence–absence variation across genomes.
3. We developed a stochastic gene gain–loss inference algorithm (PGGL) that infers gene-content turnover as a continuous-time Markov process on the phylogeny, enabling estimation of gain and loss rates and counts and reconstruction of branch-, lineage-, and clade-specific gene-content changes as quantitative measures of genome plasticity.
4. We developed a rate-based gene selection inference method (PGGS) that contrasts symmetric and asymmetric gain–loss models to detect directional biases in gene turnover, enabling lineage- and clade-specific identification of genes exhibiting non-neutral evolutionary dynamics.

5. We integrated all developed methods into a single scalable and reproducible software framework, enabling controlled and repeatable meta-pangenome analyses across datasets, species, and ecological contexts.

Collectively, the methods introduced in this chapter transform meta-pangenomes from descriptive gene catalogs into quantitative evolutionary objects, enabling systematic inference of gene-content dynamics from complex metagenomic data and providing a methodological foundation for the applications presented in subsequent chapters.

3. RESULTS FOR META-PANGENOME RECOSTRUCTION AND ANALYSIS

The pangenome encompasses the full complement of gene families identified across all genomes within a defined taxonomic group [286]. It integrates the conserved core genome, which comprises genes shared by all members of the group, with the accessory genome, consisting of genes present in only a subset of genomes [11]. In this chapter we reconstruct and compare *Klebsiella* pangenomes across three complementary cohorts, urban metagenome-assembled genomes (MPKG), multi-species isolate genomes (PKG), and a species-focused set of *K. pneumoniae* isolates (PKP), using a single, standardized pipeline for annotation, ortholog clustering, and gene presence–absence profiling. We quantify pangenome architecture (core, shell, cloud) and gene-frequency spectra, assess openness via rarefaction (Heap's law) [193, 287], and interrogate structure in the accessory genome using dimensionality reduction and hierarchical clustering. Across datasets we find a clear gradient in pangenome organization: the MPKG meta-pangenome exhibits an expanded cloud component and broad accessory diversity, the PKG isolate set shows balanced core–accessory proportions, and the PKP clinical cohort is enriched for core genes with reduced accessory dispersion. Gene accumulation analyses indicate that meta-pangenomes remain strongly open, whereas the species-level PKP set approaches functional saturation. Projection of gene presence–absence matrices separate genomes by species (and by sequence type within *K. pneumoniae*), revealing lineage-specific accessory modules consistent with ecological filtering and recent horizontal acquisition. Together, these results establish a quantitative baseline for *Klebsiella* pangenome diversity across environmental and clinical contexts, and provide the comparative framework used in subsequent sections for functional interpretation and downstream evolutionary analyses.

3.1.Sequence structural annotation of *Klebsiella* sp. MAGs and isolates

To evaluate the genomic characteristics of the *Klebsiella* genus across different sequencing contexts, we performed comparative annotation analyses on the three datasets: (1) the *Klebsiella* genus meta-pangenome (MPKG dataset) derived from metagenome-assembled genomes (MAGs) recovered from urban environments, (2) the *Klebsiella* genus pangenome (PKG dataset) and *K. pneumoniae* isolates (PKP dataset). Across *Klebsiella* species, genomes from the PKG and PKP isolates dataset exhibited consistently higher numbers of predicted coding sequences than those from MPKG MAGs, consistent with known short-read metagenomic assembly limitations and with community standards for MAG quality assessment, especially under-representation of repeat-rich and multicopy loci [18, 19] (Figure 3.1, top-left panel, Figure 3.2). For instance, *K. michiganensis* had an average of 6219.6 CDS in PKG, compared to 4998.3 CDS in MPKG (Table

3.1). Similarly, *K. oxytoca* displayed 5987.0 in PKG versus 5304.4 CDS in MPKG (Table 3.1). Intra-species variability in CDS counts, captured by the coefficient of variation (CV), was substantially higher in MAGs. For example, *K. michiganensis* in MPKG had a CV-CDS of 0.146, compared to 0.068 in PKG 0.068 indicating greater heterogeneity in gene prediction among MAGs, likely due to variable assembly completeness (Table 3.1).

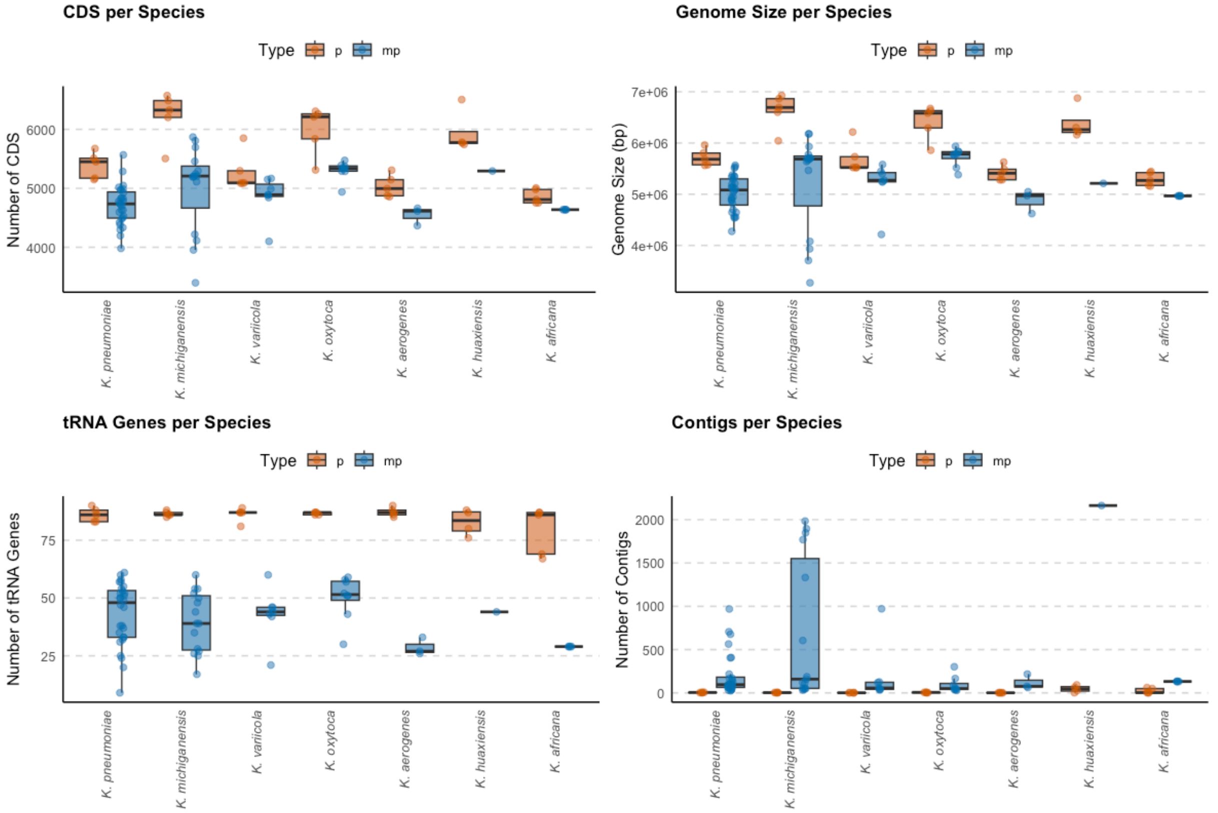


Figure 3.1. Comparative annotation metrics of *Klebsiella* sp. genomes across isolate-derived (p=PKG) and metagenome-assembled (mp=MPKG) datasets. Each panel show one metrics: (top-left) the number of predicted CDS; (top-right) total genome size (in base pairs); (bottom-left) number of tRNA genes; (bottom-right) number of contigs per assembly.

MAGs were notably deficient in annotated tRNA genes (Figures 3.1-3.2), a well-recognized consequence of fragmented assemblies that collapse or break rRNA/tRNA loci in short-read MAGs [18, 288]. While isolate genomes consistently contained ~80-87 tRNAs across species (e.g., *K. pneumoniae* 86.0, *K. aerogenes* 87.2), the same species in MPKG had markedly reduced tRNA counts (e.g. *K. pneumoniae* 43.1, *K. aerogenes* 28.7), with higher variability (CV-tRNA up to 0.329) (Table 3.1). These differences are illustrated in Figure 11 (bottom-left), which shows widespread underrepresentation and dispersion in tRNA counts among MAGs. This trend reflects the inherent difficulty in assembling and annotating tRNA loci in fragmented, short-read metagenomic data (Table 3.1).

Table 3.1. Summary of genome annotation metrics for *Klebsiella* sp. from isolate assembled genomes (PKG dataset) (n = 34).

Species	n	CDS (mean)	CV CDS	tRNA (mean)	CV tRNA	Contigs (mean)	Genome size (bp)
K. aerogenes	5	5036.40	0.037	87.20	0.022	1.6	5,417,138
K. africana	5	4861.00	0.025	79.20	0.129	23.4	5,292,441
K. huaxiensis	4	5950.25	0.062	82.75	0.069	46.5	6,389,098
K. michiganensis	5	6219.60	0.068	86.40	0.013	3.2	6,625,047
K. oxytoca	5	5987.00	0.070	86.60	0.006	4.2	6,407,511
K. pneumoniae	5	5391.40	0.041	86.00	0.036	4.2	5,716,564
K. variicola	5	5283.20	0.062	86.20	0.035	1.8	5,701,695

CDs, tRNA, and contigs represent mean values per genome; CVs measure within-species variability.

Contiguity metrics further differentiate the two types of datasets (Figure 3.1, bottom-right), the average number of contigs and dispersion being significantly higher in MAGs, in line with CAMI benchmarks and routine quality assessments for MAGs [176, 288] (Table 3.1). For example, *K. huaxiensis* in MPKG displayed a mean of 2163 contigs, whereas in PKG the value dropped to 46.5. In most PKG genomes, contig counts were consistently below 5 (e.g., *K. variicola* 1.8, *K. aerogenes* 1.6) (Table 3.2). This pronounced contrast is depicted in Figure 3.1 (bottom-right), where MAGs show wide-ranging and elevated contig numbers, indicative of their fragmented assembly states.

Table 3.2. Summary statistics of *Klebsiella pneumoniae* genomes in the PKP (Pangenome - *Klebsiella pneumoniae*) dataset (n = 99).

Feature	Mean	SD	CV
Contigs	137	81.8	0.598
Genome Size (bp)	5,634,535	138,597	0.025
CDS	5,335	159	0.030
tRNA genes	80.1	1.89	0.024

As expected, genome sizes in PKG and PKP were generally larger and more consistent, reflecting the higher completeness and lower contamination typical of isolate genomes compared with MAGs [19, 176] (Figure 3.1, top-right, Figure 3.2). For example, *K. michiganensis* in PKG had an average genome size of 6.63 Mbp, while the corresponding MPKG estimate was 5.24 Mbp (Table 3.1-3.2). A similar pattern was observed in *K. oxytoca* (PKG – 6.41 Mbp; MPKG – 5.73 Mbp), consistent with the underrepresentation of accessory and non-core regions in MAGs. These genome size discrepancies are visually apparent in Figure 3.1 (top-right), where PKG genomes cluster around larger, less variable sizes, while MAGs exhibit reduced and more dispersed sizes. While both datasets include representatives from *K. pneumoniae*, *K. michiganensis*, *K. variicola*, *K. oxytoca*, and *K. aerogenes*, their inter-species differences were more pronounced in the PKG datset, particularly for genome size and CDS content (Tables 3.1-3.2). Moreover, intra-species variability (CV) was consistently higher in MPKG, reflecting the influence of fragmented assemblies, uneven sequencing depth, and inconsistent binning quality.

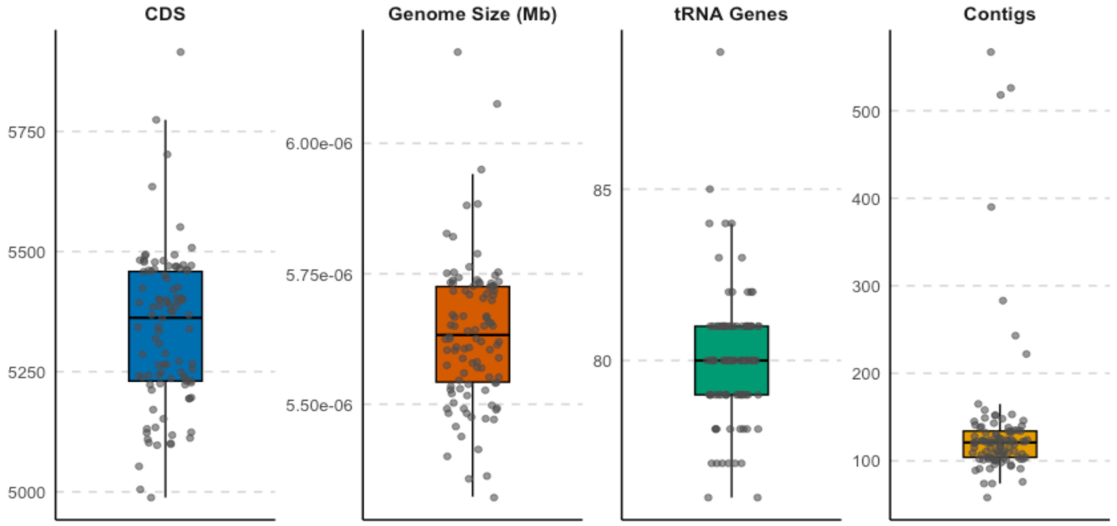


Figure 3.2. ***Klebsiella pneumoniae*** **pangenome annotation based on isolate-derived genomes (PKP dataset).**

Next, we annotated 99 *K. pneumoniae* isolate genomes to derive a high-resolution pangenome profile (PKP data). As shown in Figure 3.2, key features such as the number of coding sequences (CDS), genome size, tRNA genes and contig counts reveal relatively constrained within-species variability. The mean number of CDS per genome in the PKP dataset was 5,335±159, with a coefficient of variation (CV) of 0.030, comparable to the MPKG value of 4,714±334.6 (CV=0.071) (Table 3.1-3.2). Likewise, the genome size in PKP genomes averaged 5.63 Mbp, with minimal variation (CV=0.025), whereas the MPKG *K. penumoniae* genomes averaged 5.02 Mbp, but with more pronounced variability and the presence of low-quality assemblies (as suggested by the higher number of contigs, 204 vs. 137, respectively). The number of tRNA genes was also highly conserved across PKP genomes (80.1±1.9, CV=0.024) compared to the MKPG data (43.1±13.6, CV=0.315), likely reflecting the effects of metagenome-assembled genome fragmentation and incomplete recovery in environmental datasets. Contigs counts, a proxy for assembly fragmentation, were significantly higher and more dispersed in the MPKG (*K. pneumoniae*, mean = 2014, CV = 0.61) than in the PKP dataset (mean = 137, CV = 0.60), although variation was substantial in both.

Table 3.3. Comparative summary of unique functional annotations across Klebsiella species in meta-pangenome (MPKG), isolate pangenome (PKG), and single-species (K. pneumoniae) isolate (PKP) datasets.

	CAZy	**COG**	**EC**	**GO**	**KO**	**KP**	**PFAM**
MPKG dataset							
K. aerogenes	38	20	1141	4961	2993	450	2756
K. africana	34	20	1126	4840	2886	456	2563
K. huaxiensis	33	20	1020	4525	2576	450	2444
K. michiganensis	46	21	1272	5290	3438	468	3359
K. oxytoca	44	20	1257	5050	3242	452	3097
K. pneumoniae	36	21	1257	5149	3415	474	3375
K. variicola	38	20	5075	5075	3180	450	3020
PKG dataset							
K. aerogenes	38	21	1154	5010	3120	452	3009
K. africana	37	20	1192	5014	3172	460	2917
K. huaxiensis	37	21	1262	5090	3353	476	3208
K. michiganensis	44	20	1262	5168	3412	478	3315
K. oxytoca	42	20	1262	5107	3287	452	3118
K. pneumoniae	36	20	1211	5057	3255	452	3104
K. variicola	34	20	1195	5028	3206	452	3043
PKP dataset							
K. pneumoniae	37	22	1271	5263	3472	468	3411

CAZy = Carbohudrate- Active enZYmes; COG = Clusters of Orthologous Groups; EC = Enzyme Commission Number; GO = Gene Ontology; KO = KEGG Orthology; KP = KEGG Pathway; PFAM = Protein Families

These results underscore the robustness of the isolate-based pangenome (PKP) in resolving fine-scale genomic features and reinforce that MAG-based annotations, while informative, are subject to biases introduced by incomplete assemblies and environmental contamination. The integration of high-quality isolate genomes remains essential for pangenome structure interpretation and core/accessory gene modeling.

3.2.Functional annotation of *Klebsiella* sp. MAGs and isolates

To investigate the functional landscape of the *Klebsiella* genus across urban and clinical contexts, we performed comprehensive annotation of both metagenome-assembled genomes (MAGs) and isolate genomes using a curated suite of functional databases: CAZy (carbohydrate-active enzymes) [289], COG (orthologous groups) [290], EC (enzyme classification) [291], GO (gene ontology) [292], KEGG Orthology (KO) and KEGG Pathway (KP) [293], and PFAM (protein families) [138], functional transfers being assigned with eggNOG-mapper v2 [134].

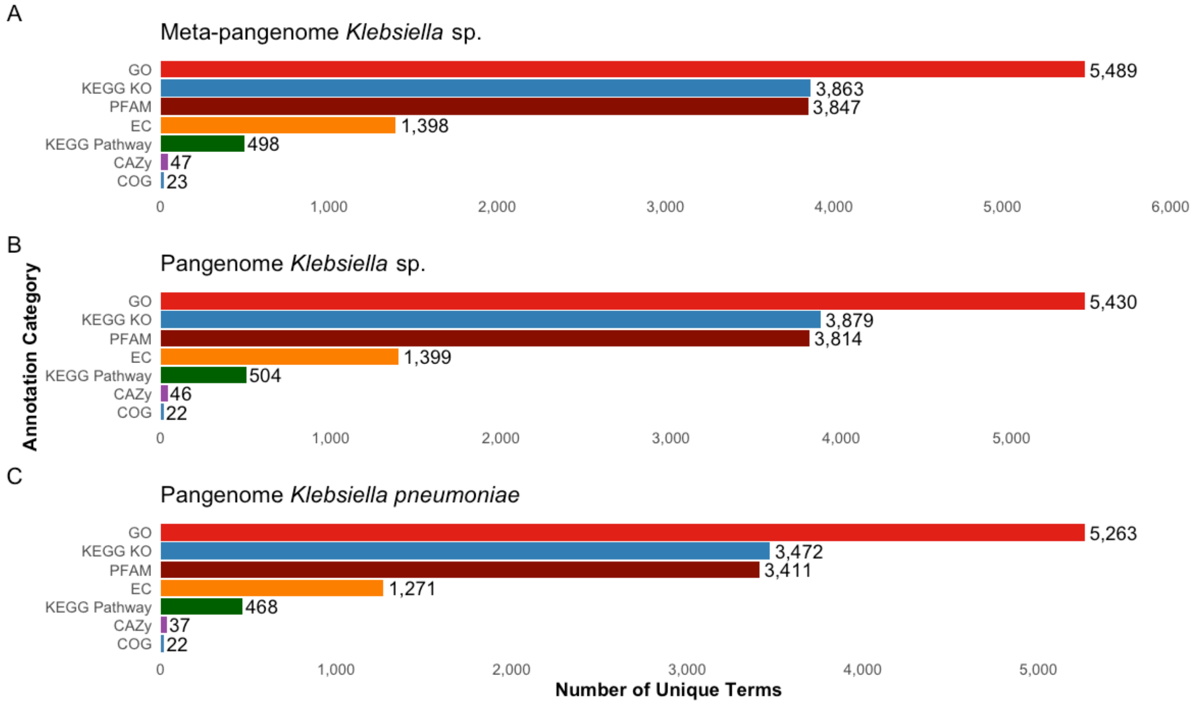


Figure 3.3. Functional annotation summary across *Klebsiella* datasets. (A) Meta-pangenome (MPKG dataset) captured from MAGs. (B) Isolate pangenome (PKG dataset). (C) *K. pneumoniae* (PKP dataset) from isolate samples.

Functional profiles were compared across three datasets: (1) the meta-pangenome (MPKG dataset), comprising high-quality MAGs from urban environment; (2) the isolate-only pangenome (PKG dataset); and (3) a focused set of *K. pneumoniae* isolate genomes (PKP dataset). The MPKG dataset, integrating only environmental MAGs (urban samples), revealed substantial inter-species variation in functional repertoire. Among the seven species, *K. michiganensis* and *K. pneumoniae* consistently exhibited the largest number of unique functional terms, particularly in EC (1,272 and 1,257, respectively), GO (5,290 and 5,149), and PFAM domains (3,359 and 3,375). This functional richness is indicative of metabolic versatility and potentially broader environmental adaptability.

Notably, *K. variicola* showed elevated KP values (5,075), suggesting extensive pathway reconstruction or annotation saturation within MAGs (Figure 3.3A; Table 3.3).

In contrast, the PKG dataset, composed exclusively of isolate genomes with standardized sequencing and annotation workflows, demonstrated a more conserved functional profile across species. The variance in total functional terms was narrower, with GO and PFAM annotations distributed more evenly. For instance, *K. oxytoca*, *K. huaxiensis*, and K. *michiganensis* all possessed comparable PFAM domain counts (~3,100–3,300), suggesting a functional core that is less influenced by environmental context (Figure 3.3B; Table 3.3). The lower dispersion in CAZy and KP annotations across PKG species further supports this inference. These findings underscore the methodological constraint of isolate-only analyses, which may underrepresent accessory or niche-specific functions present in complex microbial communities.

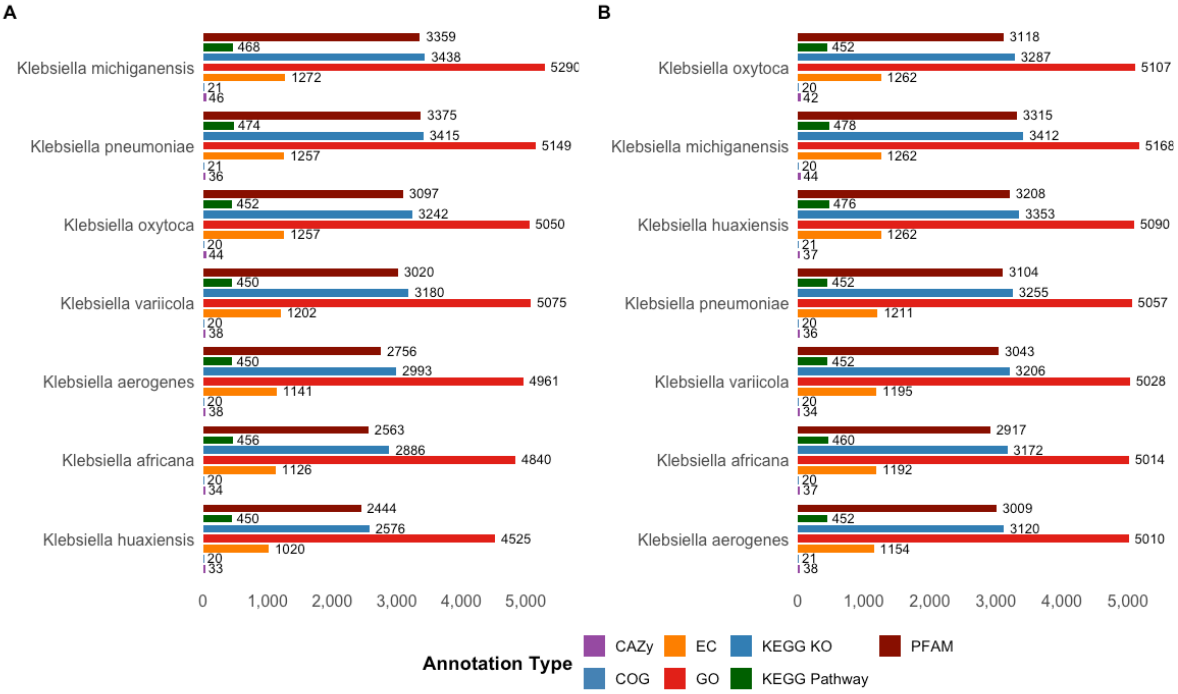


Figure 3.4. Species-level functional annotation profiles across Klebsiella genomes. (A) MPKG dataset composed exclusively of MAGs reveals substantial variation in functional term counts across species, particularly in GO and PFAM annotations. (B) PKG dataset based on isolate genomes shows greater annotation consistency, with overall higher counts for most annotation types.

A targeted comparison within *K. pneumoniae* (PKP) revealed that isolate genomes maintain functional repertoires on par with or exceeding those in MAGs and multi-species pangenomes. Specifically, PKP genomes encoded 1,271 EC entries, 5,263 GO terms, and 3,411

PFAM domains (Table 3.3, Figure 3.3C), reflecting the species' broad metabolic potential and highlighting the robustness of isolate-based functional profiling when deep coverage is available. However, the elevated KO (3,472) and GO annotation counts suggest that even within a single species, functional diversity may be underappreciated in isolate-based pangenomes when environmental signals are excluded.

Cross-dataset visualization of functional annotation distributions reveals systematic differences in term recovery between metagenomic and isolate-based approaches (Figure 3.4). Contrary to initial expectations, the isolate-derived pangenome (PKG) consistently outperforms the meta-pangenome (MPKG) in the number of unique GO terms and PFAM domains across most *Klebsiella* species. This is particularly evident for *K. aerogenes*, *K. huaxiensis*, and *K. africana*, where isolate genomes recover a broader range of protein families and ontology terms than their MAG counterparts. These trends suggest that despite the ecological breadth and composite nature of MAGs, isolate genomes retain higher annotation fidelity, likely due to improved assembly quality and completeness.

The MPKG dataset nonetheless remains valuable for uncovering functional diversity in under-sampled or uncultured lineages, such as *K. michiganensis* and *K. variicola*, where annotation patterns differ more substantially from their PKG counterparts. Moreover, in some cases, KP and EC term counts are comparable or even higher in MAGs, underscoring their utility in metabolic reconstruction. The species-level comparison of *K. pneumoniae* (Figure 3.3A-C) further reinforces the benefit of data integration: the combination of MAGs and isolate genomes produces a more complete and ecologically representative functional map. Together, these results advocate for a hybrid pangenomic strategy that leverages the complementary strengths of isolate genomes precision and MAGs breadth to maximize functional resolution across the meta-pangenomes and pangenomes.

3.3.Ortholog gene computation and clustering

Orthology was inferred for protein-coding genes across all genomes (in MPKG, PKG and PKP datasets) from Prokka-annotated GFF3 files [258] and computed using Panaroo [259]. This approach clusters orthologous coding sequences (CDSs) using CD-HIT [195] at 98% identity and integrates synteny correction and sequence alignments of core gene families, which were subsequently used for downstream comparative analysis. We computed the core, shell and cloud components of each pangenome based on gene prevalence thresholds: core (≥ 99%), shell (15 – 99%), and cloud (< 15%), consistent with the categorization scheme proposed by [259] (Figure 3.5). Notably, the MPKG meta-pangenome exhibited a substantial expansion of cloud genes, comprising 53.8% of its total gene content (Figure 3.5A). This elevated proportion of low-frequency genes is indicative of the high genomic plasticity, gene gain-loss dynamics, and functional heterogeneity characteristics of metagenome-assembled genomes (MAGs) originating from urban environment. The predominance of cloud genes reflects the fragmented and heterogenous nature of microbial populations captured via metagenomic assembly, including rare and habitat-specific functions. In contrast, the PKP pangenome, composed solely of clinical *K. pneumoniae* isolates, revealed a markedly expanded core genome, accounting for 40.8% of the total genes (Figure 3.5C). This enrichment of universally conserved genes is consistent with a tighter phylogenetic clustering and shared clinical niche, where selective pressures likely favor maintenance of a stable functional core.

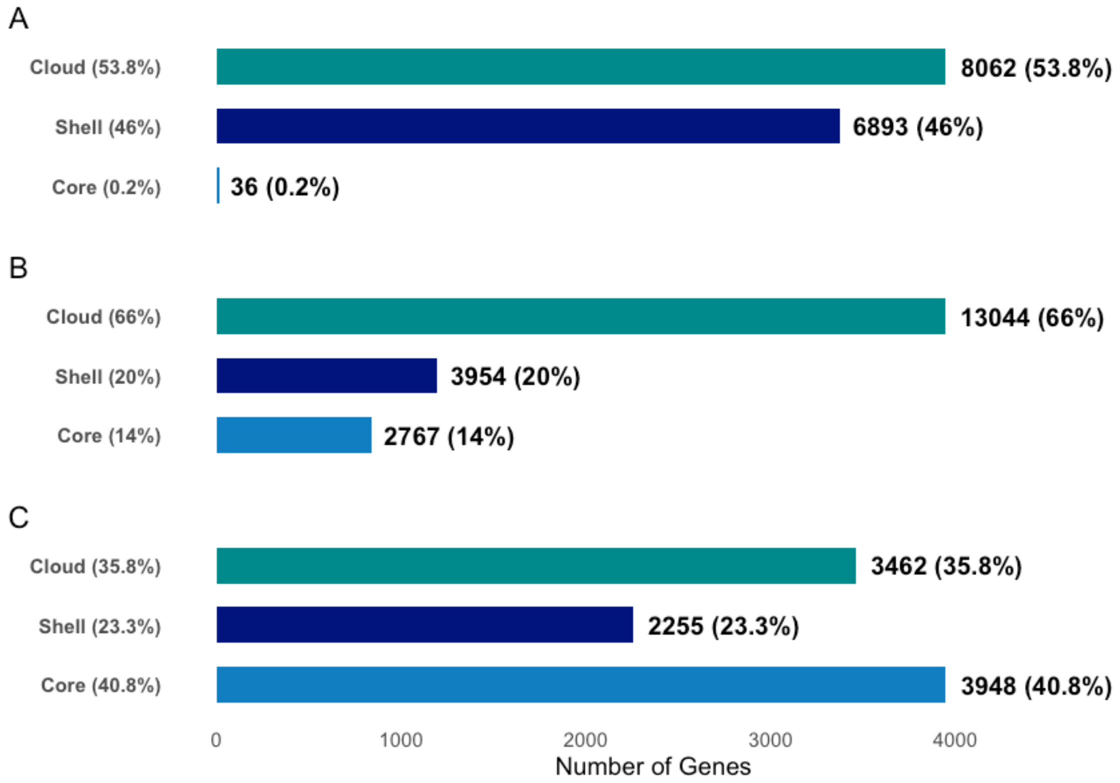


Figure 3.5. Summary of gene composition across datasets (core, shell and cloud) in (A) MPKG, (B) PKG, and (C) PKP datasets

The intermediate PKG dataset, which aggregates isolates genomes from multiple *Klebsiella* species, presented a balanced distribution, with a moderately sized core (29.7%), shell (31.6%), and cloud (38.7%) compartment, reflecting taxonomic heterogeneity but more uniform sampling compared to MAGs (Figure 3.5B).

To further evaluate the distribution of gene frequencies across genomes, we plotted histograms of gene family occurrence (Figure 3.6). All data sets exhibited the characteristic U-shaped distribution typical of prokaryotic pangenomes, where most genes are either universally conserved (core) or rare (cloud), and relatively few genes exists at intermediate frequencies [259, 294]. However, the degree of skewness varied, MPKG and PKG datasets presented a flatter curve with longer tails in low-frequency region (Figure 3.6A-B), indicating elevated accessory genome diversity in mixed-species datasets, compared to the narrower peak and expanded high-frequency tail observed in PKP dataset (Figure 3.6C).

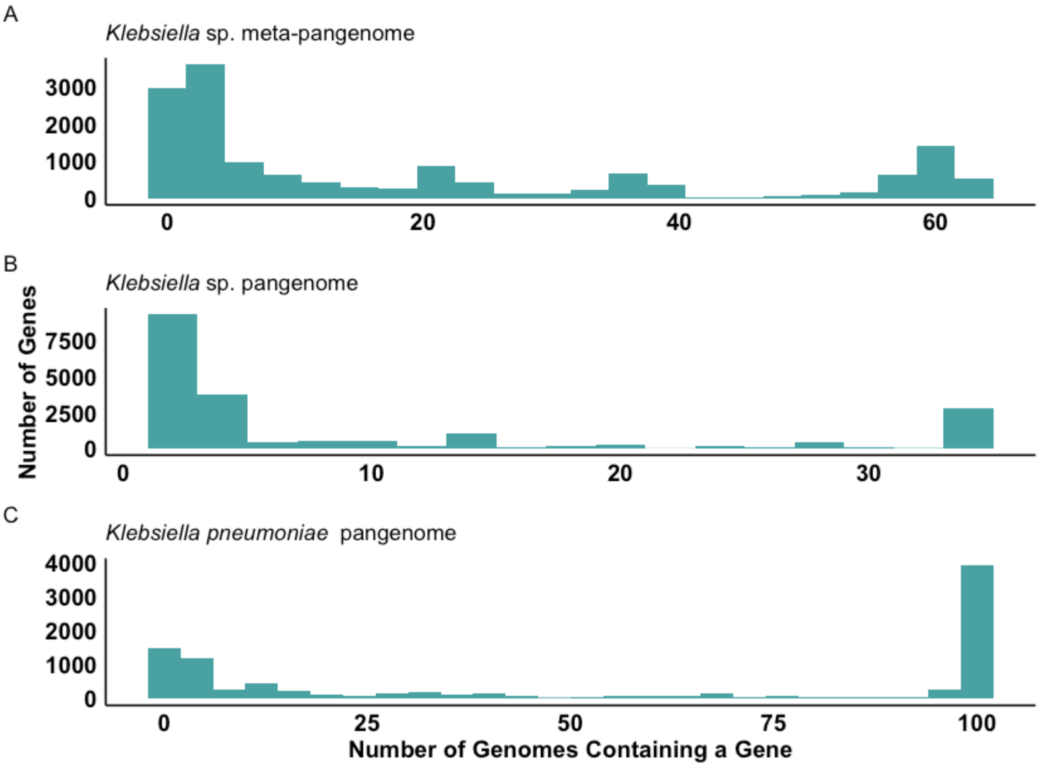


Figure 3.6. Gene frequency distribution across the number of genomes in (A) MPKG, (B) PKG, and (C) PKP datasets

These patterns delineate a scale-dependent pangenome architecture: the meta-pangenome (MPKG) is enriched for low-frequency "cloud" genes, the single-species clinical set (PKP) shows the converse with an expanded core and contracted cloud, and the multi-species isolate set (PKG) is intermediate. This gradient reflects the evolutionary heterogeneity and ecological plasticity sampled by each dataset and is consistent with the canonical U-shaped gene-frequency distribution of bacterial pangenomes

3.4.Gene discovery dynamics and meta-pangenome openness

To investigate the openness/closeness of *Klebsiella* pangenomes, we constructed rarefaction curves and modeled gene discovery rates across genome permutations using a power-low fit (Heap's law) as proposed in Tettelin et al [286]. The rarefaction analysis estimates the cumulative number of non-redundant orthologous gene families discovered as new genomes are sequentially added to the dataset, thereby allowing inference on whether the pangenome remains expandable (open) or saturated (closed) [286, 295].

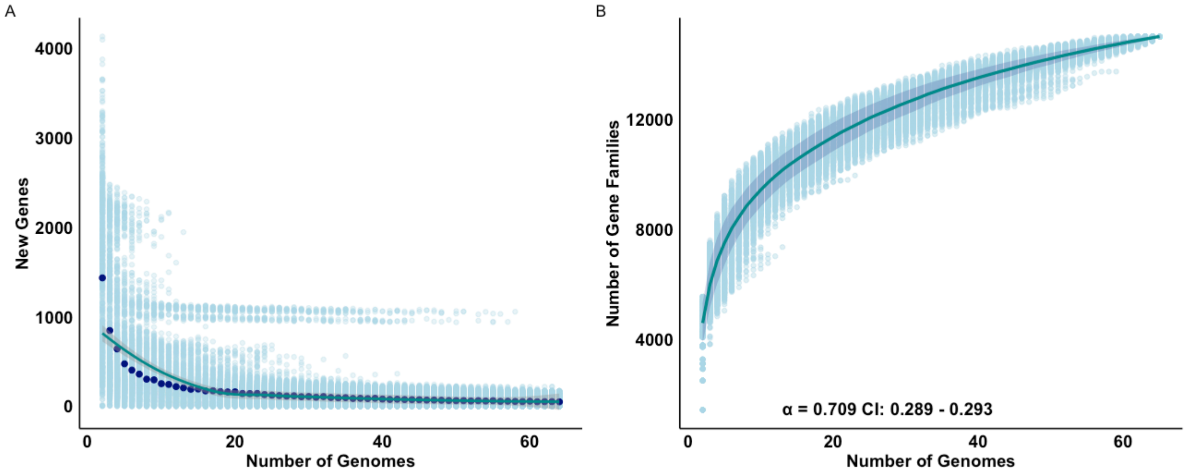


Figure 3.7. Gene discovery dynamics and meta-pangenome openness in the MPKG dataset. (A) Discovery rate of new gene families per genome, showing diminishing returns with increasing sampling. (B) Cumulative number of unique gene families fitted to Heap's law model.

To quantitatively assess these dynamics, two complementary estimations were conducted for each reconstructed pangenome. First, we evaluated the number of newly discovered gene families per genome along the sampling order, thereby quantifying the incremental contribution of each genome to the pangenome. Second, we computed the cumulative number of unique gene families across genomes and fitted the growth curve using Heap's low, formalized as $n = k \cdot N^{\gamma}$, where n is the total number of observed gene families for N genomes, and $\gamma > 0$ is indicative of an open pangenome, and when $\gamma < 0$, the pangenome is considered closed [286]. The rarefaction curve was generate using `rarefaction()` function of the micropan package [267] with 1000 permutations.

The MPKG meta-pangenome dataset, derived from 64 MAGs sourced from urban environment sources, displayed an intermediate, clearly open pangenome (Figure 3.7B). The gene discovery curve (Figure 3.7A) reveals a steep initial slope, with each genome contributing a high number of novel genes even at 40 additions, with saturation being reached after 50 additions (Figure 3.7A). The rarefaction curve surpassed 13,000 gene families, and the fitted Heap's exponent yielded $\gamma = 0.291$ ($\alpha = 0.709$, 95% CI: $0.289 - 0.293$), a strong signal of highly open pangenome. This reflects the vast gene content heterogeneity, ecological versatility, and potential for horizontal gene transfer among urban *Klebsiella* strains.

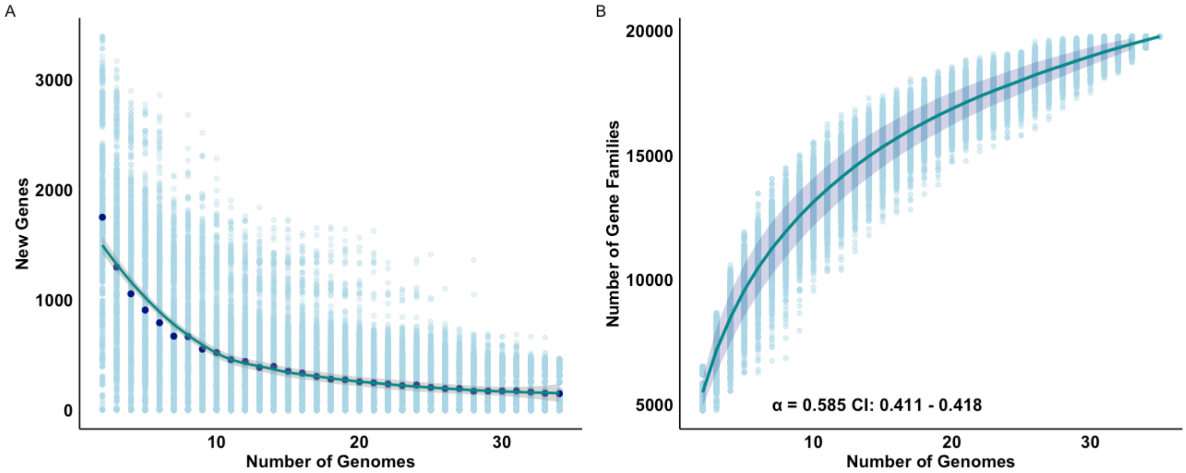


Figure 3.8. Gene discovery dynamics and meta-pangenome openness in the PKG dataset. (A) Discovery rate of new gene families per genome, showing diminishing returns with increasing sampling. (B) Cumulative number of unique gene families fitted to Heap's law model.

The PKG dataset (34 genomes from various *Klebsiella* species) presented an intermediate strong open pangenome. While the gene discovery rate showed a downward trend, it remained above ~50 genes/genome, with a cumulative gene family count exceeding 19,000 (Figure 3.8A). The fitted α value was 0.585 (95% CI:0.411-0.418), consistent with a strong open pangenome (Figure 3.8B). The slowdown in novel gene acquisition suggests a near-saturation of core and shell components, yet the presence of rare cloud genes continues to contribute to diversity. This pattern aligns with expectations for genus-level reconstructions spanning multiple species and ecological niches.

In the species-specific pangenome of 99 clinical *K. pneumoniae* isolates, gene discovery reached near saturation, falling below 10 after ~50 genomes and approaching ~5 in the final bins (Figure 3.9A). The rarefaction curve count plateaued near 9,500 gene families (Figure 3.9B), and the fitted exponent remained positive, γ =0.139, α=0.861, 95% CI: 0.138-0.141), indicating a formally open but functionally a closing pangenome. This pattern is consistent with reduced ecological diversity and high genomic conservation within clinical strains.

These results establish a gradient of pangenome openness across ecological and taxonomic contexts. The MPKG dataset reflects the highest degree of openness and gene influx, driven by environmental heterogeneity. The PKG dataset occupies an intermediate position, with saturation beginning to emerge despite continued novelty. In contrast, the PKP dataset reveals a decelerating rate of discovery, with most functional diversity already captured, a trend typical for narrow, clonal clinical populations [11, 194]. Despite all α values exceeding zero and thus meeting the criterion for openness, the empirical gene discovery curves provide additional resolution. Specifically, the rate of new gene acquisition per genome clearly declines with increasing sample size and is an essential complement to cumulative rarefaction modeling.

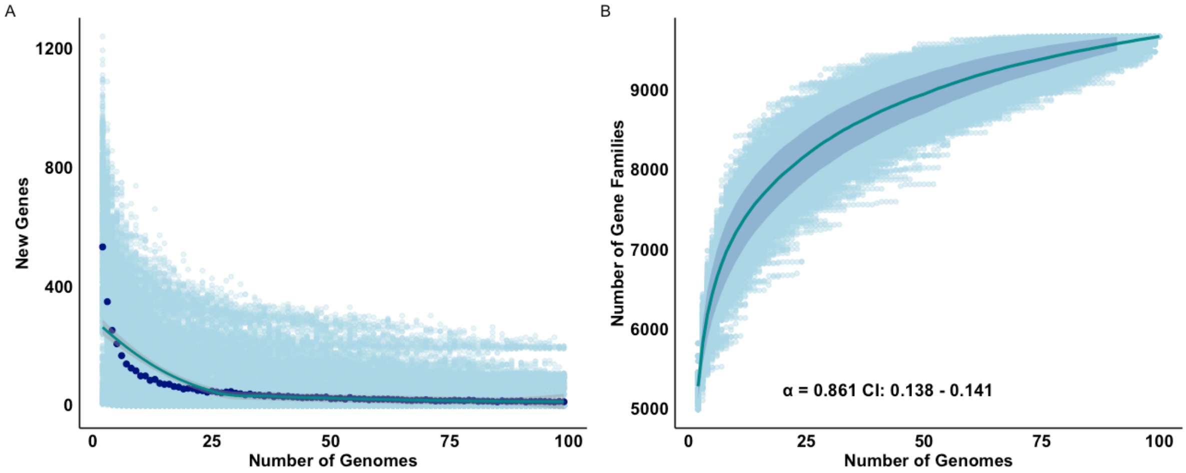


Figure 3.9. Gene discovery dynamics and meta-pangenome openness in the PKP dataset. (A) Discovery rate of new gene families per genome, showing diminishing returns with increasing sampling. (B) Cumulative number of unique gene families fitted to Heap's law model.

3.5. Pangenome structure and gene content clustering

To investigate the internal diversity and structure of the reconstructed meta-pangenomes, we conducted a comparative analysis of gene presence-absence patterns across the three datasets. Two complementary strategies were employed, dimensionality reduction using principal component analysis (PCA) [296, 297], and hierarchical clustering of binary gene presence-absence matrices [298, 299]. Together, these methods elucidate the underlying genomic architecture, inter- and intra-species variability, and patterns of accessory gene distribution. Across datasets, PCA of the gene presence-absence matrices revealed strong separation between genomes based on species or sequence type (ST) (Figures 3.10, 3.13-3.14). In the MPKG dataset, derived from urban MAGs, the PCA projection of the accessory gene presence-absence matrix reveals a high degree of inter-species separation across the first two principal components, which jointly capture the dominant presence of genomic variation (Figure 3.10).

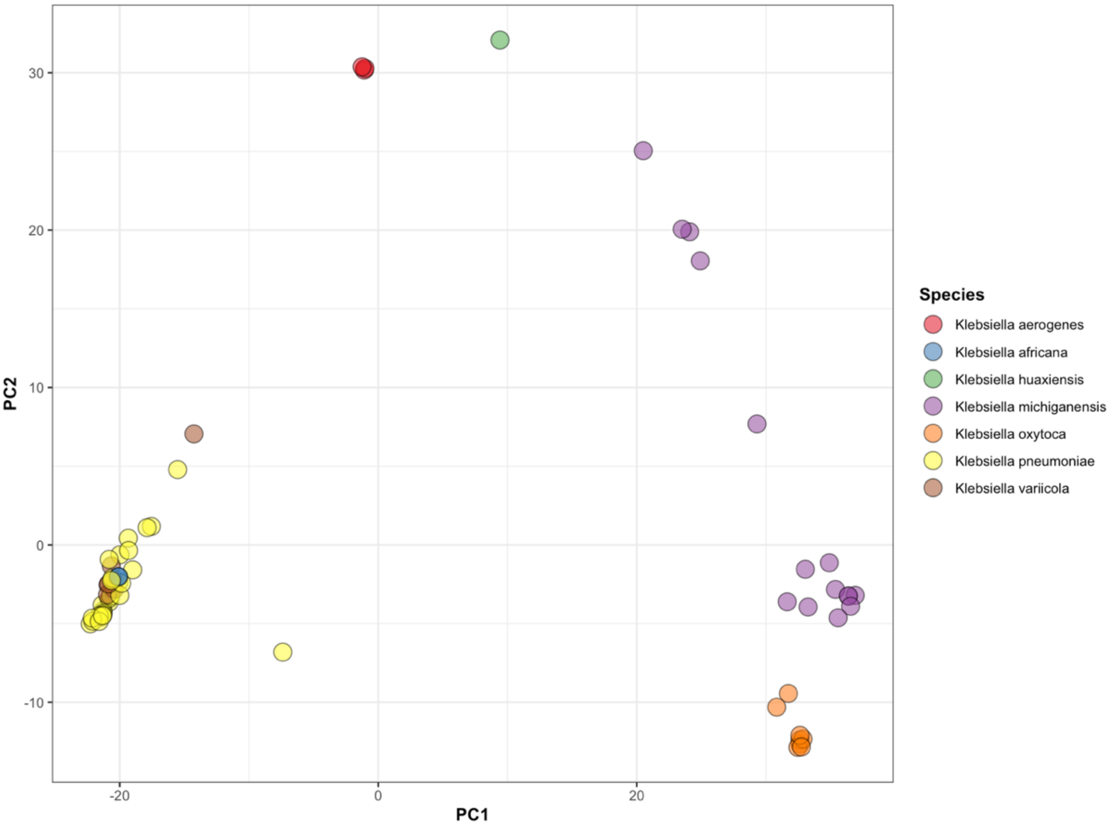


Figure 3.10. Principal Component Analysis (PCA) of MPKG meta-pangenome (*Klebsiella* spp.) gene presence-absence matrix, colored by species.

Clusters corresponding to distinct *Klebsiella* species, including *K. michiganensis*, *K. huaxiensis*, *K. oxytoca*, and *K. aerogenes* are clearly distinguishable, with each species occupying a unique region in the PCA space (Figure 3.10). Notably, *K. pneumoniae* MAGs appear tightly clustered near the origin, reflecting high intra-species homogeneity in gene content, while the environmental species (urban landscapes) form dispersed and distant clusters, underscoring broader accessory genome repertoires and potential ecological adaption [300]. The separation of *K. variicola*, *K. africana*, and *K. michiganensis* in orthogonal directions further supports the hypothesis that urban *Klebsiella* lineages exhibit species-specific accessory gene architectures, likely shaped by environmental filtering and horizontal gene acquisition [11, 300–302].

The gene presence-absence heatmap of hierarchical clustering for the MPKG meta-pangenome (Figure 3.11) shows a broad, dense block of universally conserved orthologous groups spanning nearly all MAGs (core genome), overlaid by patchy, clade-restricted islands that define the accessory gene space. Several accessory bocks are restricted to *K. michiganensis* and to members of the *K. oxytoca* (Figure 3.11), consistent with the high plasmid and integrative and conjugative elements (ICE) burden and ecological breadth reported for the lineages [303–306].

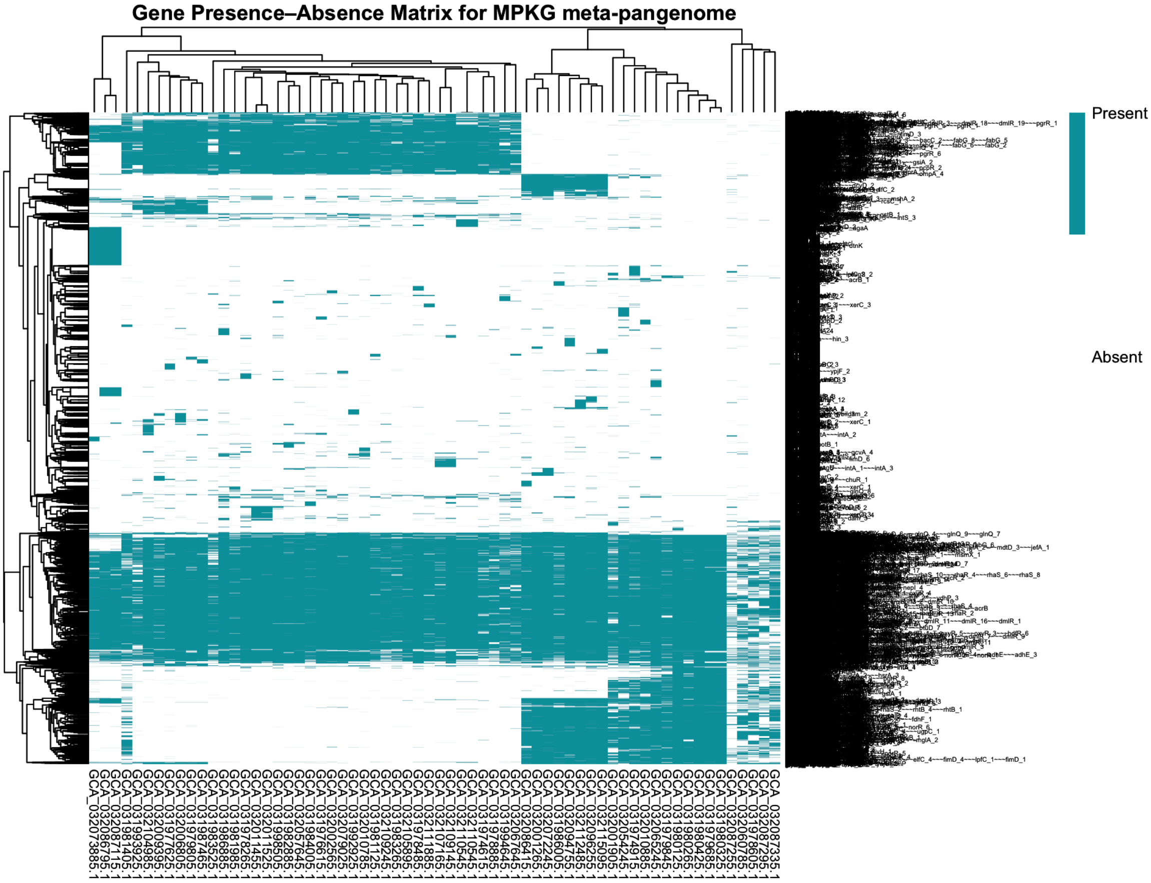


Figure 3.11. Presence-absence heatmap based on hierarchical clustering of genomes and orthologous groups of the MPKG meta-pangenome (*Klebsiella* sp.).

Distinct islands are also apparent among *K. variicola* MAGs, in line with its environmental and plant-associated niche and corresponding accessory functions [301]. By contrast, *K. pneumoniae* MAGs cluster tightly with fewer large accessory islands, reflecting the more homogenous core observed in KpSC clinical lineages (Figure 3.11) [300, 307]. We also note smaller, lineage-delimited blocks for *K. aerogenes*, *K. huaxiensis*, and *K. africana*, suggesting species-specific gene modules that likely track habitat filtering and horizontal gene acquisition (Figure 3.11). The numerous MAG-specific islands across taxa are consistent with mobile genetic elements and prophage cargo-a pervasive driver of gene-content heterogeneity in *Enterobacterales* [308, 309]. Hierarchical clustering of both genomes and genes recapitulates species-level structure and mirrors the PCA separation, as expected when pangenome construction corrects annotation and assembly artefacts and leverages gene-context information [259]. A small number of columns with unusually sparse signal corresponds to MAGs with small assembly sizes and very low N50 values (e.g., N50 < 10kb and total size < 4Mb), a pattern typical of incompletely recovered bins rather than true biological absence [18, 19, 176, 288]. These observations reinforce that core-accessory structure is recoverable from metagenomic data, while interpretation of rare or patchy genes should be conditioned on MAG quality metrics.

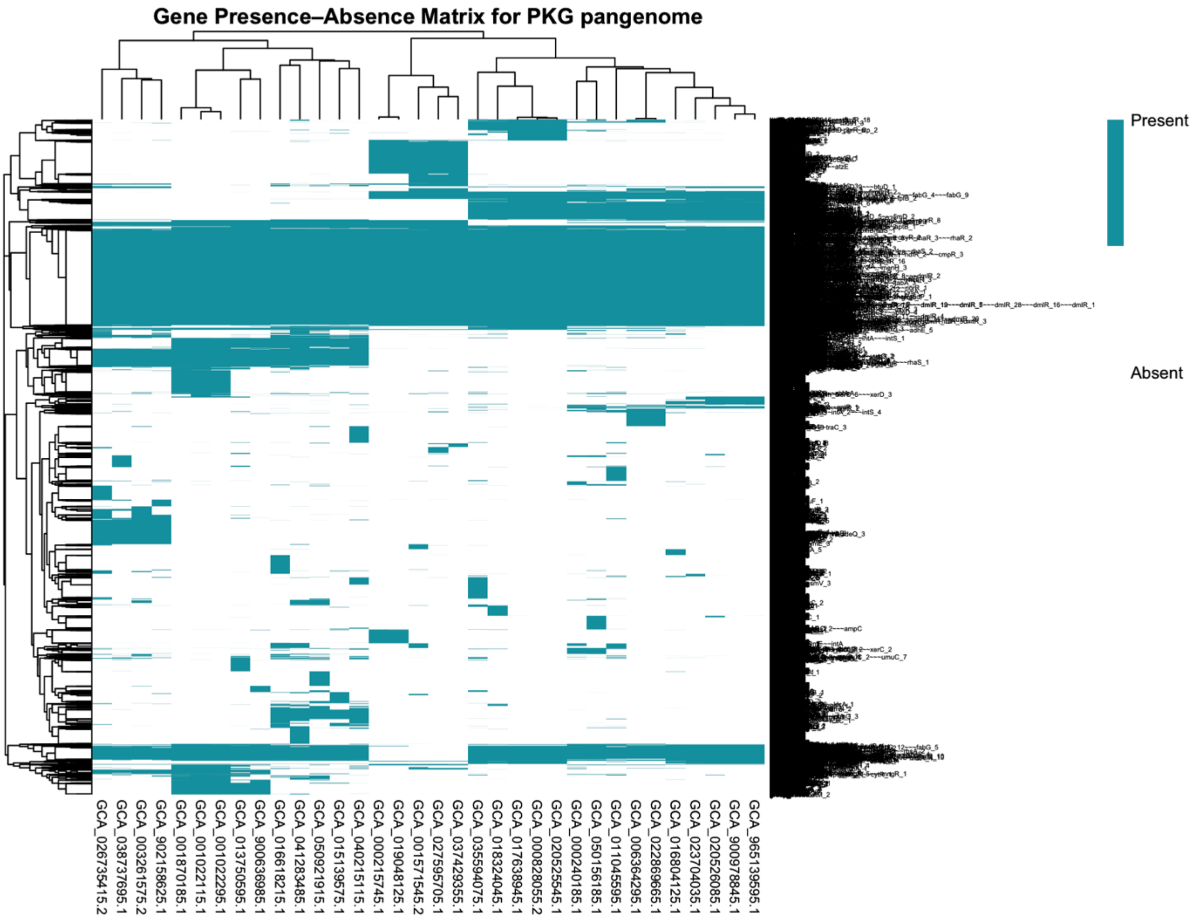


Figure 3.12. Presence-absence heatmap based on hierarchical clustering of genomes and orthologous groups of the PKG pangenome (*Klebsiella* spp.).

The PKG pangenome, derived from RefSeq/GenBank [310, 311] isolate assemblies, exhibits a sharper, contiguous core block than the MAG-based MPKG analysis, consistent with higher completeness and uniform annotation in isolates. Hierarchical clustering resolves clades that match established taxonomy for members of the *Klebsiella pneumoniae* species complex (KpSC), including *K. pneumoniae*, *K. variicola*, and *K. africana* form a coherent cluster with a large species-core and several clade-restricted accessory modules (Figure 3.12) [312, 313]. The *Klebsiella oxytoca* complex appears as a second block, with *K. oxytoca* and *K. michiganensis* grouping together and *K. huaxiensis* adjacent but distinct; *K. aerogenes* falls outside these complexes as an outgroup with a clearly differentiated accessory repertoire [314]. Within each species block, compact, high-frequency cores are flanked by patchy islands that likely reflect lineage-specific gene pools shaped by plasmids, prophages, and integrative conjugative elements rather than assembly artefacts, an interpretation supported by the near-complete status and consistent genome sizes of these isolates. These patterns align with known population structure and gene-flow dynamics in *Klebsiella* (species-complex clustering; species-specific accessory content) and the prominent role of mobile elements in *Enterobacterales* pangenomes [312–314]. As expected for high-quality isolate genomes, columns with extensive absences are rare; where present, they most plausibly reflect genuine clade-specific deletions or missing accessory modules rather than incomplete recovery. Methodologically, the clean separation of species in the heatmap and dendrogram is consistent with pangenome reconstructions that correct spurious merges and leverage gene-neighborhood information [259].

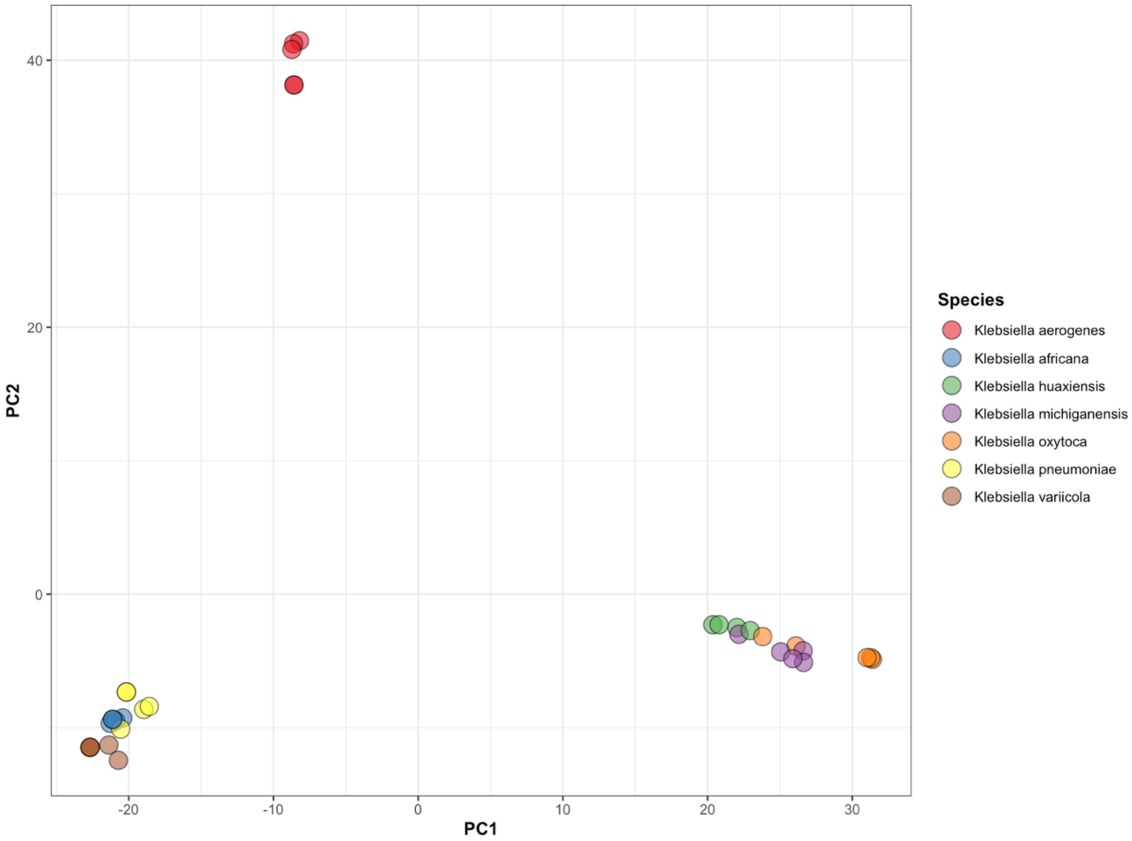


Figure 3.13. Principal Component Analysis (PCA) of PKG pangenome (*Klebsiella* sp.) gene presence-absence matrix, colored by species.

Relative to the MPKG meta-pangenome, the PKG matrix shows (i) a more compact and continuous core (fewer “false absences”), (ii) fewer sparsely populated columns, and (iii) accessory islands that align tightly with taxonomic clusters rather than with assembly quality—patterns attributable to the higher completeness and lower fragmentation of isolate assemblies versus MAGs. In MPKG, by contrast, some sparse columns track MAGs with small assembly size/N50, and patchy accessory signal can be inflated by fragmentation; these effects are largely absent in PKG, strengthening inference on species-specific accessory repertoires.

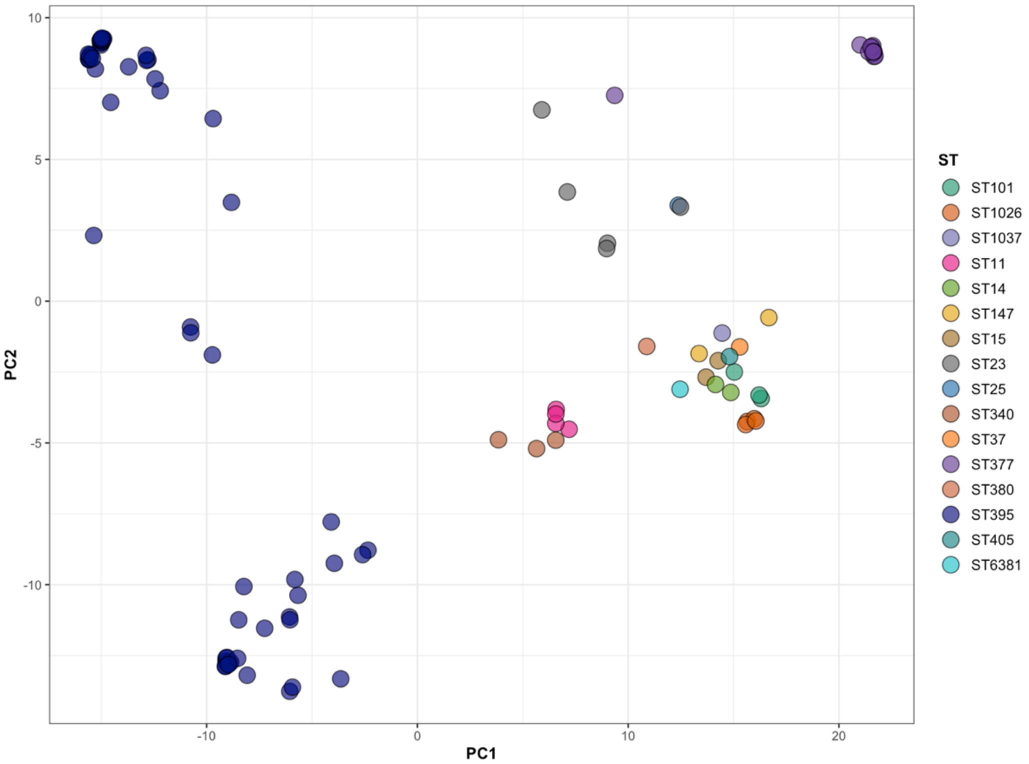


Figure 3.14. Principal Component Analysis (PCA) of PKP pangenome (*Klebsiella pneumoniae*) gene presence-absence matrix, colored by sequence type (ST).

Principal component analysis of the binary gene presence–absence matrix resolves the PKG genomes into species-specific clouds (Figure 3.13). PC1 and PC2 capture the dominant axes of gene-content variation and cleanly partition the *Klebsiella pneumoniae* species complex (KpSC), the *K. oxytoca* complex (*K. oxytoca*, *K. michiganensis*), and the outlying *K. aerogenes* and *K. huaxiensis* clade, consistent with established taxonomy and recent population-genomic surveys. Compact clusters (e.g., *K. variicola*, *K. pneumoniae*) indicate highly shared accessory repertoires, whereas the broader spread observed for *K. michiganensis* suggests greater within-species genomic plasticity. The concordance between gene-content ordination and species boundaries mirrors structure seen in core-genome phylogenies and genotyping frameworks for the KpSC, and is expected given the high completeness of these isolate assemblies [11, 312–314].

As well, we quantified within-species structure for *Klebsiella pneumoniae* using hierarchical clustering and PCA of the gene presence-absence matrix from the PKP dataset (clinical high-quality isolate assemblies). The first two PCs segregate genomes by multilocus sequence type (ST), with discrete clusters for major epidemic or clinically important lineages, including ST147, ST15, ST23, and ST395, consistent with established KpSC population structure [312, 313] (Figure 3.14). The ST395 cluster is especially compact, indicating a highly conserved accessory repertoire compatible with recent clonal expansion or streamlining. By contrast, ST23 and ST15 exhibit greater dispersion across PC space, suggesting more heterogeneous accessory gene pools shaped by episodic horizontal acquisition and loss (Figure 3.14).

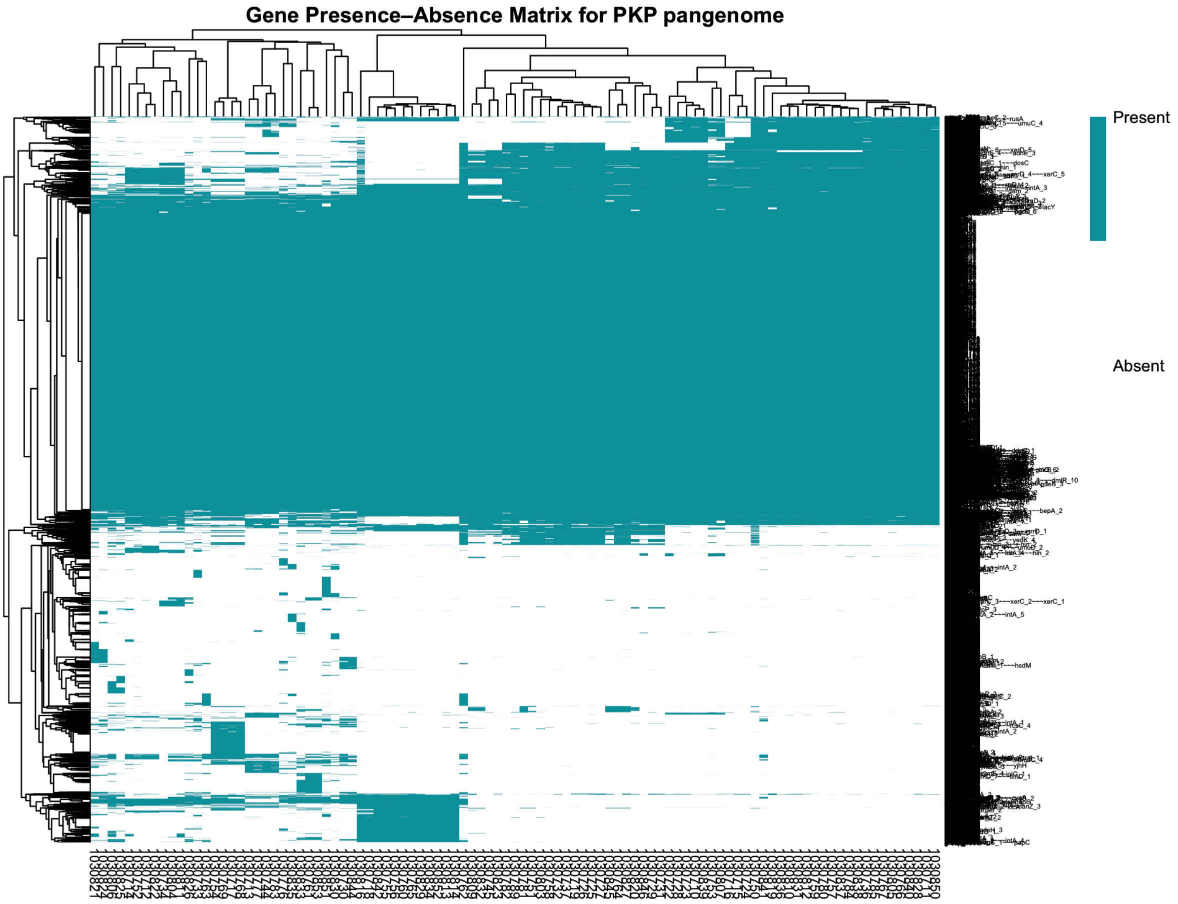


Figure 3.15. Presence-absence heatmap based on hierarchical clustering of genomes and orthologous groups of the PKP pangenome (*Klebsiella pneumoniae*).

The hierarchical clustering corroborates the PCA, presenting a dense, universally present core flanked by ST-restricted accessory modules (Figure 3.15). Blocks enriched in ST147/ST15 likely reflect AMR-associated plasmids and integrative conjugative elements, whereas ST23 blocks are consistent with virulence-associated loci (e.g., siderophore and capsule modules) frequently reported for hypervirulent *K. pneumoniae* [312, 313]. The predominance of large, coherent accessory islands and the rarity of columns with widespread absence argues for genuine lineage-specific content rather than assembly artefact, which is expected given the near-complete status and standardized annotation of these isolates [18].

The PKP, a single species isolate dataset, resolves the finest structure and provides high resolution of sequence type specific accessory modules. These modules are clearly delineated, and the core block is compact and nearly universal. PKG (multi-species isolates) retains strong species cores with clade-restricted islands, whereas MPKG (heterogeneous MAGs) shows the greatest patchiness, reflecting fragmentation and mixed ancestry. The progressive sharpening of signal with increasing assembly completeness and phylogenetic focus matches expectations from pangenome theory and *Klebsiella* population genomics, which predict clearer accessory architecture and fewer artefactual absences under these conditions.

3.6. Conclusions to chapter 3

This chapter provides empirical validation of the meta-pangenome framework developed in chapters 2 by applying it to real isolate and metagenome-assembled genome datasets of *Klebsiella* spp. across multiple ecological and different taxonomic scales. The main validated methodological outcomes are summarized as follows:

1. The end-to-end meta-pangenome software workflow is validated on empirical data, demonstrating that standardized gene prediction, functional annotation, ortholog clustering, and gene-frequency summarization can be applied reproducibly to isolate genomes and MAGs within a single analytical protocol, enabling controlled cross-dataset comparisons.
2. Genome-level structural summaries validate input harmonization and quality propagation, as the framework consistently recovers known differences between isolates and MAGs in coding sequence counts, tRNA completeness, genome size, and contiguity, indicating that downstream analyses reflect underlying data quality rather than pipeline artefacts.
3. Functional annotation and aggregation steps are validated across multiple databases (GO, KEGG, PFAM, EC, COG, CAZy), recovering narrow and stable functional profiles for high-quality isolate cohorts and broader, ecologically enriched repertoires for MAG-based

meta-pangenomes, consistent with expectations from assembly completeness and environmental sampling.

4. Ortholog clustering and gene-frequency stratification are validated by recovery of canonical pangenome structure, including core–shell–cloud architecture and characteristic gene-frequency distributions, across datasets differing in phylogenetic scope and sampling depth.
5. Rarefaction and pangenome openness modelling validate gene discovery dynamics, correctly distinguishing strongly open meta-pangenomes from partially saturated genus-level and near-saturated single-species cohorts, confirming that the framework captures both mathematical and practical aspects of pangenome openness.
6. Presence–absence matrices and visualization steps are validated through coherent structural patterns, including clade-restricted accessory islands, species-level partitioning in multi-species datasets, and sequence-type-specific modules in clinical *K. pneumoniae* datasets, demonstrating sensitivity to biologically meaningful organization while remaining robust to MAG fragmentation.
7. Multivariate analyses validate the consistency of gene-content representations, as dimensionality-reduction methods reproduce expected relationships between genomes, separating MAGs by environmental heterogeneity, isolates by species boundaries, and clinical strains by clonal lineage.

Collectively, these results validate the methodological pipeline integrated in a bioinformatics software as a reliable foundation for evolutionary inference, confirming that its outputs are structurally coherent, biologically interpretable, and suitable for downstream modelling of gene gain–loss histories and directional turnover developed in subsequent chapters.

4. GENE DYNAMICS PATTERNS IN META-PANGENOMES

To quantify evolutionary turnover and directional pressures in microbial gene content from urban meta-pangenomes, we developed, implemented, and applied two complementary, phylogeny-aware methods to the MPKG (urban MAGs), PKG, and PKP *Klebsiella* datasets. The PGGL (Pangenome Gene Gain–Loss) method converts gene presence–absence into per-branch, per-node, and per-genome counts of gains and losses by coupling continuous-time Markov modeling with ancestral state reconstruction on the species tree. The PGGS (Pangenome Gene Selection) methods then provided gene-wise evidence of directional bias for gain versus loss, via likelihood-based comparison of symmetric and asymmetric turnover models, summarized with an interpretable effect size and ΔAIC thresholds.

Applied together, the framework resolved a heterogeneous, lineage-structured flux landscape, a defining feature of open bacterial pangenomes embedded in large, environmentally structured gene pools [11, 193, 315]. PGGL localized where turnover clustered (innovation bursts on internal branches; pruning on terminal branches), while PGGS quantified how individual genes deviated from balanced dynamics, recovering near-neutral behavior for core housekeeping modules and pronounced biases for accessory functions linked to mobility, defense, and niche-tuned metabolism [8, 315–318]. Together, PGGL and PGGS methods delivered a compact, comparative atlas of rates (events along paths) and directionality or bias (gain versus loss) that underpins the result sections that follow.

4.1.Probabilistic inference and visualization of gene gain and loss events

To localize where gene content changed along meta-pangenome phylogenies, we developed and applied PGGL (Pangenome Gene Gain-Loss) (Algorithm 2; Algorithm A3.1), implemented in R [38], a phylogeny-based inference that treats each orthologous group as a binary trait evolving on the species tree. For every gene, PGGL performs maximum-likelihood ancestral state reconstruction under a two-state continuous Markov model using Felsenstein's pruning algorithm (Algorithm 1) [31, 239, 271]. This yields node-wise posterior probabilities of presence-absence across the tree. The PGGL then evaluates each parent-child edge and calls an event when posterior support for presence changes beyond a small threshold (default δ =0.05), labelling increases as gains and decreases as losses.

The PGGL returns a branch-level atlas of events and corresponding summaries that are directly interpretable for comparative genomics. First, edgewise calls identify innovation and attrition hotspots on the tree, exposing clusters of gain-rich branches that precede radiating clades and loss-enriched terminal edges consistent with post-acquisition streamlining.

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

```
Inputs:
    • Rooted, strictly bifurcating tree $T$ with branch length.
    • Tip vector for gene $g$: $x_g \in \{0,1,?\}^n$ aligned to the tips of $T$.
    • Target state $s^* \in \{0,1\}$ (default1).
    • Event threshold $\delta > 0$ (default 0.05).
    • Model: ER ($\lambda = \mu$) or ARD ($\lambda \neq \mu$); root prior: uniform [1/2, 1/2] or stationary.
Output:
Rows for every edge $e = (u \rightarrow v)$: g, u, v, parent label, child label, gain, loss, $\lambda_g$, $\mu_g$.
Procedure:
    1. Initialize tip likelihoods (upward message).
       For each tip t:
        • if $x_g(t) = 0$: $L_t = [1,0]$
        • if $x_g(t) = 1$: $L_t = [0,1]$
        • if missing: $L_t = [1,1]$
    2. Define the 2-state CTMC
        • Rates: $Q(\lambda, \mu) = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix}$
        • Transition on branch of length $t$: $P(t) = \exp\{Qt\}$.
    3. Fit ($\lambda_g$ $\mu_g$) by maximum likelihood (ML).
        • Post-order recursion (for a given $\lambda, \mu$):
          For each internal node $a$ with children $c_1, c_2$ and branch lengths $t_1, t_2$:
          $v_1 = P(t_1)L_{c_1}, v_2 = P(t_2)L_{c_2}$ ,$L_a = v_1 \odot v_2$.
          Apply scaling $L_a \leftarrow L_a / \sum L_a$ and accumulate log-scales to avoid underflow.
        • Root prior and log-likelihood:
          $\pi = \log(\pi^T L_{root})$ + (sum of log-scales).
        • Optimize $\lambda, \mu \geq 0$ (ER: $\lambda = \mu$) to maximize $\log L$ (bound-constrained L-BGFS-B).
        • Record the MLEs as $\hat{\lambda}_g, \hat{\mu}_g$.
    4. Compute node posteriors for the target state (upward + downward).
        • Recompute all $L_u$ using $\hat{Q}$.
        • Root marginal: $M_{root} \propto \pi \odot L_{root}$; normalize to sum 1.
        • Downward pass: for each edge $a \rightarrow c$ with length $t$:
          $M_c \propto (M_a^T P(t))^T \odot L_c$; normalize.
          Tips with observed states remain at posterior 1/0; missing tips resolve via this pass.
        • Store $p_u := M_u[s^*]$ for every node $u$.
    5. Call edge events by posterior change.
        • For each edge $e = (u \rightarrow v)$, $\Delta_e = p_v - p_u$:
        • gain = 1 if $\Delta_e > \delta$ else 0
        • loss = 1 if $\Delta_e < \delta$ else 0
    6. Append output rows.
       For every edge $e$ write: (g, u, v, parent label, child label, gain, loss, $\hat{\lambda}_g, \hat{\mu}_g$) to R.
Notes:
    • Complexity per gene: one likelihood evaluation is $O(|E|)$; optimization uses a few dozen
      evaluations in practice.
    • Using stationary root prior often stabilizes calls on deep trees; uniform prior is
      conservative when presence/absence is balances.
```

Second, per-genome root-to-tip aggregates quantify the total burden of changes experienced by each genome, producing heavy-tailed distributions in which a minority of genomes carry disproportionate flux, candidates for recent ecological transitions or elevated interaction with mobile gene pools. Third, per-node tallies highlight clade-specific episodes of expansion or contraction, facilitating hypothesis-driven inspection of lineages that repeatedly acquire or purge accessory modules.

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

```
Inputs:
    • Rooted, strictly bifurcating tree $T$ with branch lengths.
    • Presence-absence matrix $P \in \{0,1,?\}^{n \times G}$ (tips × rows), rows aligned to tips of $T$.
    • Target state $s^* \in \{0,1\}$ (default 1); threshold $\delta > 0$ (default 0.05).
    • Model ER ($\lambda = \mu$) or ARD ($\lambda \neq \mu$); root prior (uniform or stationary).
Outputs:
    • Long table R: for every gene $g$ and edge $e = (u \to v)$: g, u, v, parent label, child label,
      gain, loss, $\hat{\lambda}_g, \hat{\mu}_g$.
    • Summaries: per-tip, per-node, or per-edge aggregates derived from R.
Procedure:
    1. Ensure $P$ rows match the tip order of $T$.
    2. Initialize $R \leftarrow \emptyset$.
    3. For each gene $g = 1, \dots, G$:
        • Extract $x_g = P[, g]$. If invariant (all 0 or all 1, ignoring "?"), skip.
        • Run **Algorithm 1 (PGGL-Gene)** on $(T, x_g, s^*, \delta, model, prior)$ to obtain $R_g$ (all
          edge rows for gene $g$ with $\hat{\lambda}_g, \hat{\mu}_g$).
        • Append $R \leftarrow R \cup R_g$.
    4. Return $R$.
    5. (optional post-processing) From $R$, compute per-genome (root-to-tip) or per-node event
       counts/rates as needed.
Notes:
    • Time complexity $O(G|E|)$; trivially parallelizable over genes.
    • Use the same $\delta$ and root prior across genes for comparability.
    • Missing tips ("?") are handled inside Algorithm 1; no special casing needed here.
```

Where appropriate, event counts are normalized by path length to allow like-for-like rate comparisons across uneven branch durations. Sensitivity analyses varying δ yield stable qualitative patterns, indicating that major features are not threshold artifacts.

4.1.1. Summary of gene gain-loss counts and rates in meta-pangenomes

Bacterial pangenomes are not static inventories but dynamic gene pools that expand and contract as lineages diversify, exchange DNA, and adapt to new niches [10, 11, 193]. Quantifying how often genes are gained or lost, and where on the phylogeny these transitions concentrate, provides a mechanistic read-out of genome plasticity, ecological opportunity, and constraint [319, 320]. In a meta-pangenome, the familiar core–shell–cloud structure emerges from the balance between innovation and attrition. Innovation comprises acquisition via mobile elements (plasmids, ICEs, prophage), duplication, and occasional (neo)functionalization; attrition reflects deletion, pseudogenization, and purifying selection that prunes costly or destabilizing cargo [11, 318, 320]. Crucially, gains and losses are not mirror processes: horizontal transfer and phage drive punctuated bursts of acquisition, whereas streamlining and stability selection often produce gradual loss [319, 321]. These opposing forces act on a phylogenetic scaffold whose branch lengths, effective population sizes, recombination regimes, and ecological exposures shape the tempo of turnover [312, 322]. Interpreting the resulting patterns therefore requires both robust aggregations, to stabilize noisy presence–absence calls, and explicit mapping to the tree, so that events are attributed to lineages rather than samples. Here, we analyze event counts and continuous-time gain–loss rates across three complementary *Klebsiella* cohorts: (i) MPKG, a meta-pangenome built from high-quality urban MAGs; (ii) PKG, cultured isolates spanning multiple species; and (iii) PKP, a clinical collection enriched for *Klebsiella pneumoniae* sequence types sampled in the Republic of Moldova. We proceed in three steps designed for comparability across datasets, first, aggregating gain–loss counts at genome resolution, next, localizing events to specific branches of the tree, and finally, estimating branch-length–normalized rates under a CTMC framework, all within a uniform inference pipeline. We implement this design by reporting totals, branch-localized events, and rate estimates per unit evolutionary time, enabling a clean separation of magnitude from tempo and allowing direct cross-cohort contrasts.

4.1.2. Pangenome gain-loss events patterns quantification

The distributions of gene gain and loss events (Figures 4.1–4.3, Tables 4.1–4.2) demonstrate pronounced heterogeneity across genomes, both in terms of absolute counts and in the relative balance between gains and losses [10, 11, 193, 319]. In the PKP dataset (Figure 4.1), which includes only *Klebsiella pneumoniae* isolate genomes, several lineages exhibit very high cumulative turnover, with more than 1,000–1,300 events per genome, predominantly attributable to gene losses. In contrast, other genomes within the same dataset show substantially lower turnover, in some cases fewer than 200–300 events (Figure 4.1). This contast is in accord with recent publication characterizing *K. penumoniae* and microbial evolution genomes [312, 318, 323–326].

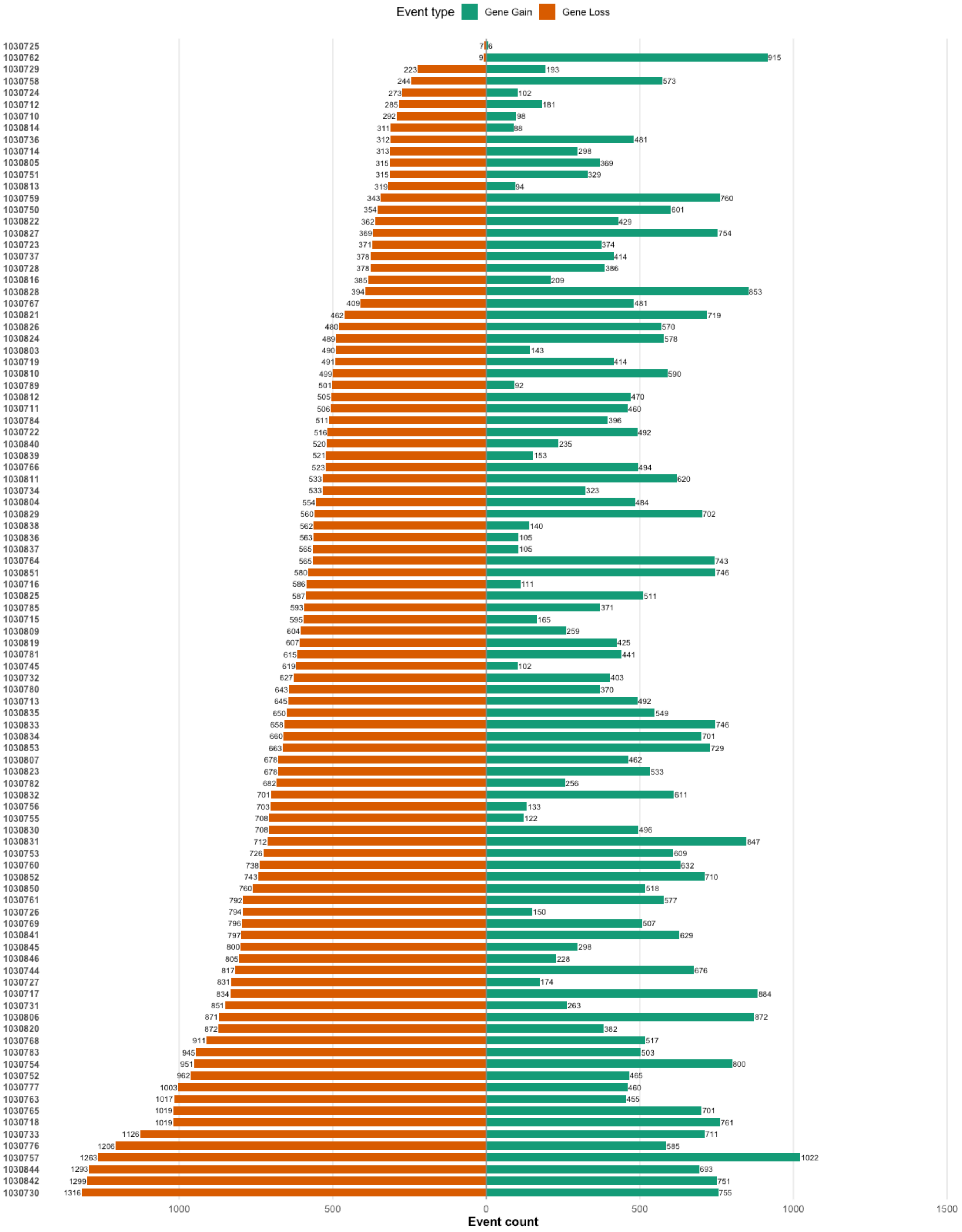


Figure 4.1. Gene gain–loss event counts per isolate genome in the PKP pangenome.

Across all three datasets, the balance of gene turnover was consistently skewed toward losses when event counts were normalized by the number of genes evaluated per genome (Tables 4.1–4.2). In the MPKG dataset, the aggregated normalized contribution of gains was 3.58, compared with 8.05 for losses, corresponding to relative proportions of 30.8% gains versus 69.2% losses. The PKG dataset showed the strongest loss bias, with normalized gains contributing only 0.88 against 4.24 losses (17.2% versus 82.8%) (Table 4.1).

Table 4.1. Normalized gene gain and loss contribution across MPKG, PKG, and PKP datasets

Dataset	Nr. genomes	Gains	Losses	% Gains	% Losses
MPKG	64	3.581	8.049	0.31	0.69
PKG	35	0.878	4.236	0.17	0.83
PKP	99	8.02	10.85	0.43	0.57

By contrast, the PKP dataset was more balanced, with normalized gains (8.03) and losses (10.86) corresponding to proportions of 42.5% and 57.5%, respectively (Table 4.1).

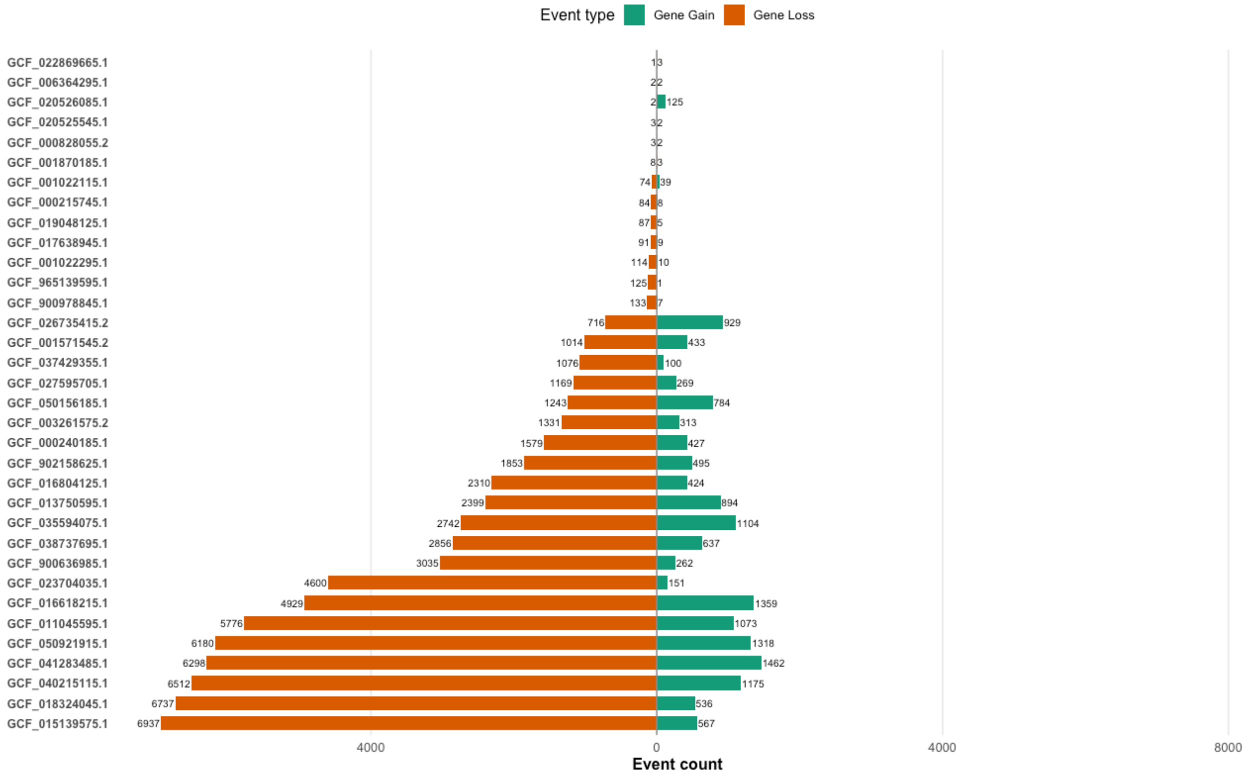


Figure 4.2. Gene gain-loss event counts per isolate genome in the PKG pangenome.

These normalized values reflect the aggregate contribution of each event type across genomes, after adjusting for the number of genes analyzed per dataset. While all datasets reveal a predominance of loss events, the extent of this bias varies, being strongest in PKG and weakest in

PKP, a hierarchy consistent with sustained deletional pressure and episodic influx of mobile cargo in clinical settings [312, 318, 323–326].

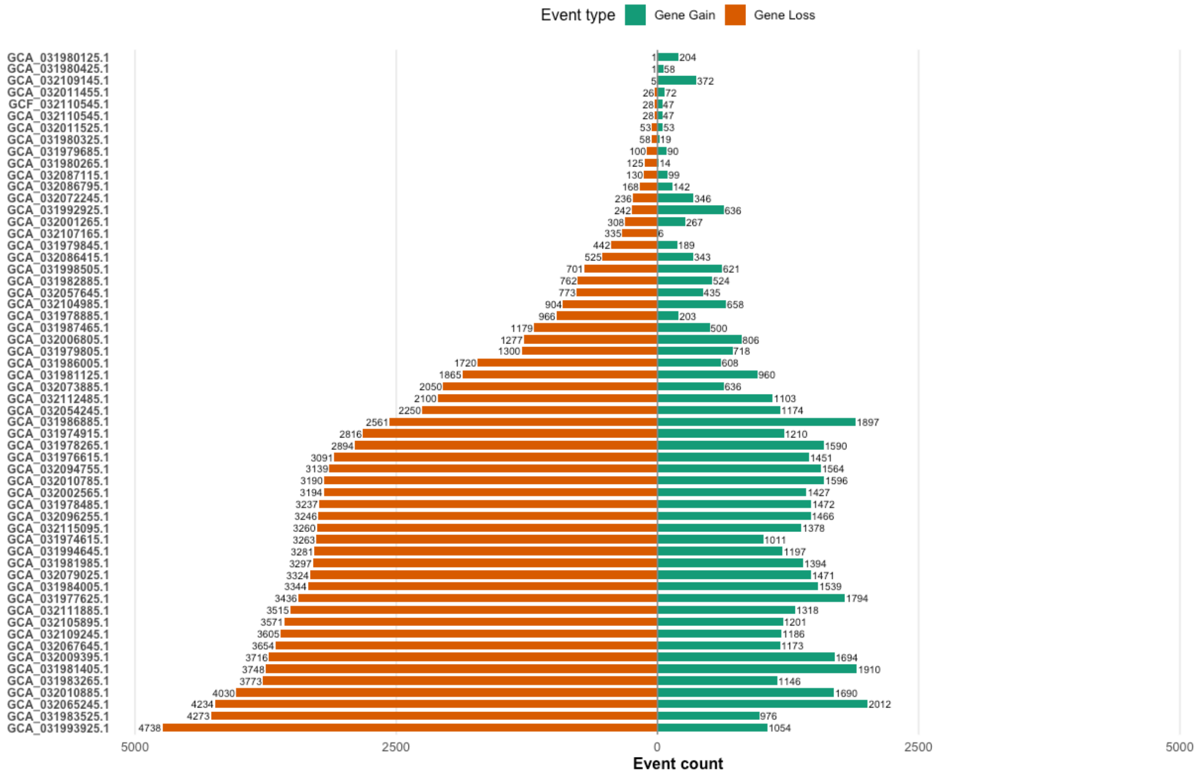


Figure 4.3. Gene gain-loss event counts per MAG in the MPKG meta-pangenome.

A similar but more extreme pattern is observed in the PKG dataset (Figure 4.2). Here, certain genomes accumulate >6,000 gene loss events, far exceeding the median across the dataset. The PKP dataset (Figure 4.1; Tables 4.1–4.2) presents a more balanced distribution, with most genomes showing between 500 and 1,200 events, indicating persistent but not overwhelming acquisition superimposed on background pruning [312, 323, 324].

In the MAG-derived dataset, most species show ~30–33% gains versus ~67–70% losses per genome. *Klebsiella pneumoniae* and *K. variicola* carry the largest tip burdens (≈1,029–1,032 gains and ≈2,364–2,376 losses per genome), indicating intense accessory churn at shallow phylogenetic depth. *K. oxytoca* is intermediate (≈884/1,816), whereas *K. aerogenes* shows a lower density (≈292/782). An apparent gain bias in *K. africana* (≈47/28 per genome) reflects small sample size (n=2) and should be interpreted cautiously. MAGs also display smaller annotated gene complements than isolates (e.g., *K. pneumoniae* mean 4,568 genes), with larger dispersion for some taxa (e.g., *K. variicola* SD ≈476 genes), consistent with environmental heterogeneity and partial recovery of accessory contigs (Table 4.2) [288, 327]. These results indicate that in urban compartments, high-turnover species (*K. pneumoniae, K. variicola*) dominate the event landscape, yet the direction of turnover remains loss-skewed (Figure 4.4; Table 4.2).

Table 4.2. Summary of events counts in MPKG, PKG and PKP datasets

Features	***Kpn***	***Kmi***	***Kox***	***Kva***	***Kae***	***Kaf***	***Khu***
			MPKG dataset				
Nr. genomes	28	10	8	7	3	2	NA
Mean genes	4568	5207	5211	4752	4248	4545	NA
SD tip genes	323	159	171	476	137	0	NA
Terminal gain events	28837	6660	7075	7224	877	94	NA
Terminal loss events	66543	14057	14534	16550	2348	56	NA
Gain per genome (mean)	1029	666	884	1032	292	47	NA
Loss per genome (mean)	2376	1405	1816	2364	782	28	NA
			PKG dataset				
Nr. genomes	5	5	5	5	5	5	4
Mean genes	5306	6174	5969	5270	5018	4824	5856
SD tip genes	223	435	416	328	191	78	360
Terminal gain events	2289	5881	1208	1653	815	708	2374
Terminal loss events	8601	30856	5630	9576	3430	7170	6756
Gain per genome (mean)	458	1176	242	331	163	142	593
Loss per genome (mean)	1720	6171	1126	1915	686	1434	1689
			PKP dataset				
Nr. genomes	99	NA	NA	NA	NA	NA	NA
Mean genes	5284	NA	NA	NA	NA	NA	NA
SD tip genes	156	NA	NA	NA	NA	NA	NA
Terminal gain events	45885	NA	NA	NA	NA	NA	NA
Terminal loss events	62069	NA	NA	NA	NA	NA	NA
Gain per genome (mean)	463	NA	NA	NA	NA	NA	NA
Loss per genome (mean)	627	NA	NA	NA	NA	NA	NA

Isolate genomes also accentuate the loss predominance, notably most species fall between ~15–26% gains and ~74–85% losses. *K. michiganensis* is an outlier with the highest per-genome burden (≈1,176 gains and ≈6,171 losses), followed by *K. huaxiensis* (≈593/1,689) and K. variicola (≈331/1,915). Even species with smaller genomes, such as *K. aerogenes* (≈163/686), remain strongly loss-skewed (Figure 4.2; Table 4.2). Compared with MPKG, isolates recover larger gene sets (e.g., *K. pneumoniae* mean 5,306 genes) and reveal broader between-strain variability (e.g., *K. oxytoca* SD ≈416 genes), consistent with improved capture of plasmids and genomic islands in high-quality assemblies (Table 4.2) [5–7,8–10]. Beyond the dominance of *K. michiganensis*, the spread in total per-genome burden is steep, ranging from ≈7,347 events in *K. michiganensis* to ≈849 in *K. aerogenes*, an ~8.6-fold difference that underscores strong lineage heterogeneity (Figure 4.5; Table 4.2). Mid-ranked species cluster near ~2.2–2.3×10^3 total events per genome, yet their directionality diverges, notably *K. huaxiensis* carries the highest relative gain share (~26%), *K. pneumoniae* is intermediate (~21%), and *K. variicola* is lower (~15%), indicating distinct balances between acquisition and pruning despite similar overall burdens (Figure 4.5; Table 4.2). No species exhibits a net-gain regime in the isolate cohort, losses exceed gains across the board, reinforcing a cohort-wide reductive bias that is consistent with long-recognized deletional pressures in bacteria [318, 325, 326]. The rank order by total burden proceeds from *K. michiganensis* to *K. huaxiensis*, *K. variicola*, *K. pneumoniae*, *K. africana*, *K. oxytoca*, and *K. aerogenes* (Figure 4.5; Table 4.2). Estimates for *K. huaxiensis* (n=4) and *K. africana* (n=5) remain sample-size limited and should be interpreted cautiously.

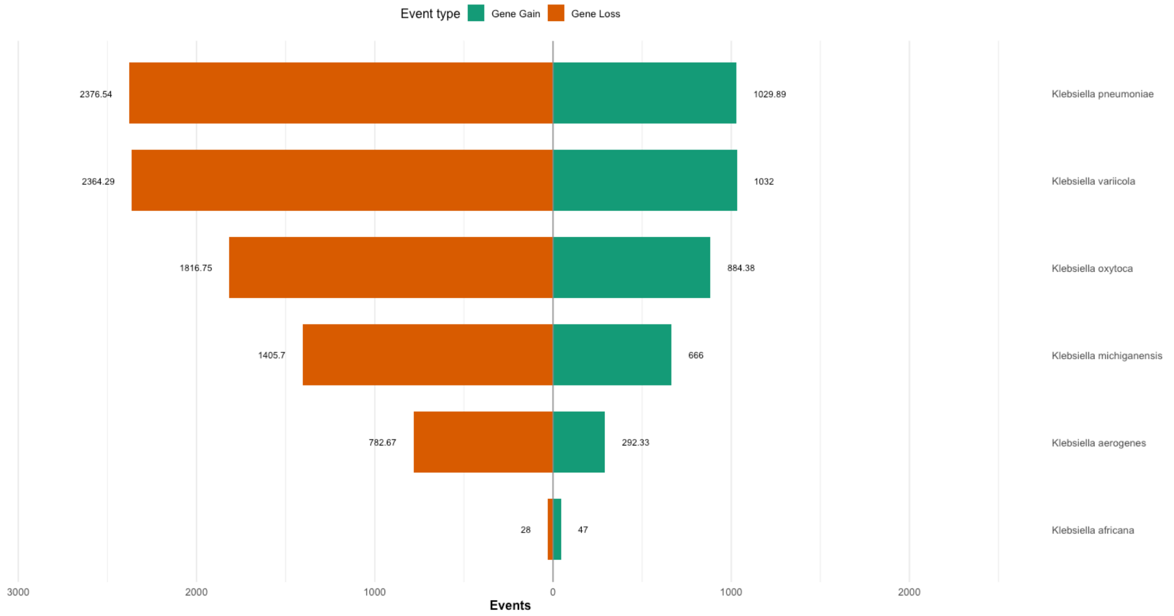


Figure 4.4. Species-level gene gain and loss count in the MPKG meta-pangenome

Within the large *K. pneumoniae* clinical set (n=99), turnover approaches near-balance relative to PKG at ≈463 gains versus ≈627 losses per genome (~43/57%). Despite a net loss bias, the absolute gain load is substantial, consistent with persistent acquisition of mobile cargo in clinical contexts [312, 323, 324]. The within-dataset dispersion is modest (mean 5,284 genes; SD ≈156), reflecting the narrower taxonomic scope (Figure 4.1; Table 4.2).

Considered together, the MPKG, PKG, and PKP datasets analyses resolve three reproducible patterns that frame our interpretation of *Klebsiella* gene turnover. First, terminal losses exceed gains in every cohort, and the effect size is cohort-dependent: (1) MAGs show a moderate skew (≈30–33% gains, 67–70% losses), (2) isolates show the strongest skew (≈9–26% gains, 74–91% losses), and (3) clinical *K. pneumoniae* is intermediate (≈43% gains, 57% losses) (Tables 4.1–4.2; Figures 4.4–4.5). Second, species rank by total per-genome burden aligns with accessory-space expectations, with *K. pneumoniae* and *K. variicola* leading in MAGs and *K. michiganensis* dominating in isolates, where it exhibits >7× more losses than gains and an ~8-fold spread over *K. aerogenes* (Table 4.2; Figure 4.5). Finally, data type shapes observables, notably, fragmented MAG assemblies under-recover plasmids and genomic islands, which can inflate apparent terminal losses in some taxa, whereas high-quality isolate assemblies reveal genuine, lineage-specific reductive dynamics, including pronounced pruning in *K. michiganensis* and *K. variicola* (Tables 4.1–4.2; Figures 4.4–4.5) [318, 325–328]. Signals for rare taxa remain uncertain owing to limited sample size (e.g., *K. africana* in MPKG and *K. huaxiensis* in PKG).

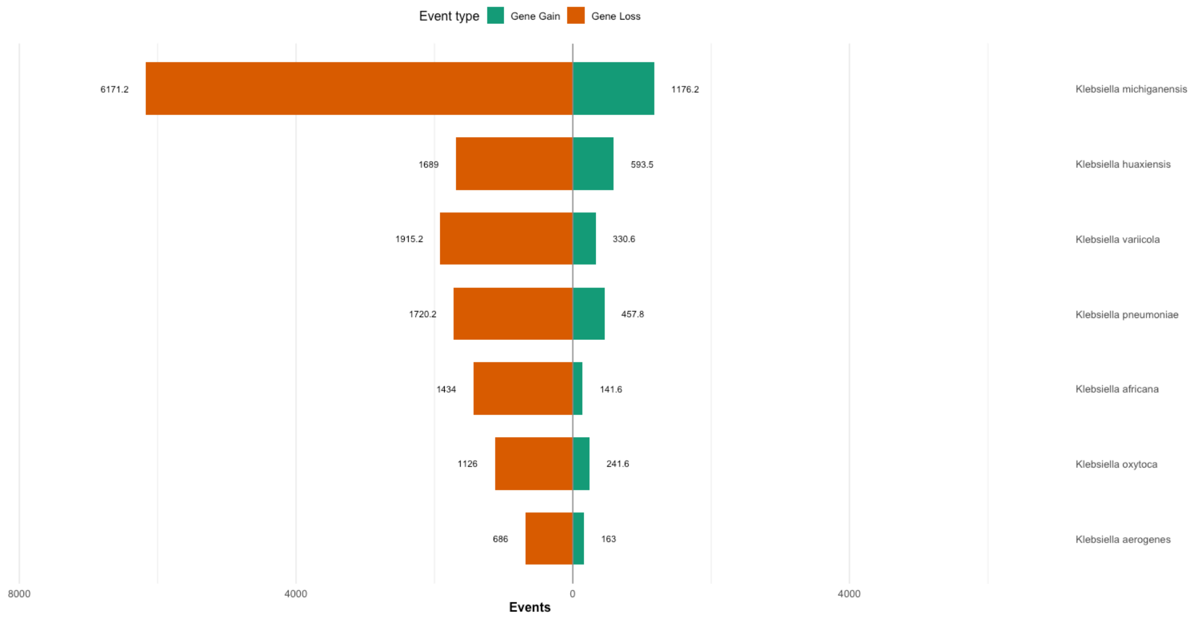


Figure 4.5. Species-level gene gain and loss in the PKG pangenome.

4.1.3. *Pangenome gain-loss events mapped on phylogenetic trees*

Event totals capture magnitude but confound evolutionary time, because identical numbers of transitions can accumulate on branches of very different length. To enable fair comparison across the tree, we estimated genome-specific rates of gain (λ) and loss (μ) under an all-rates-different continuous-time Markov model and standardized by branch length, yielding the expected number of acquisitions or deletions per unit evolutionary time [219, 235]. These rate estimates are interpreted in conjunction with the mapped counts.

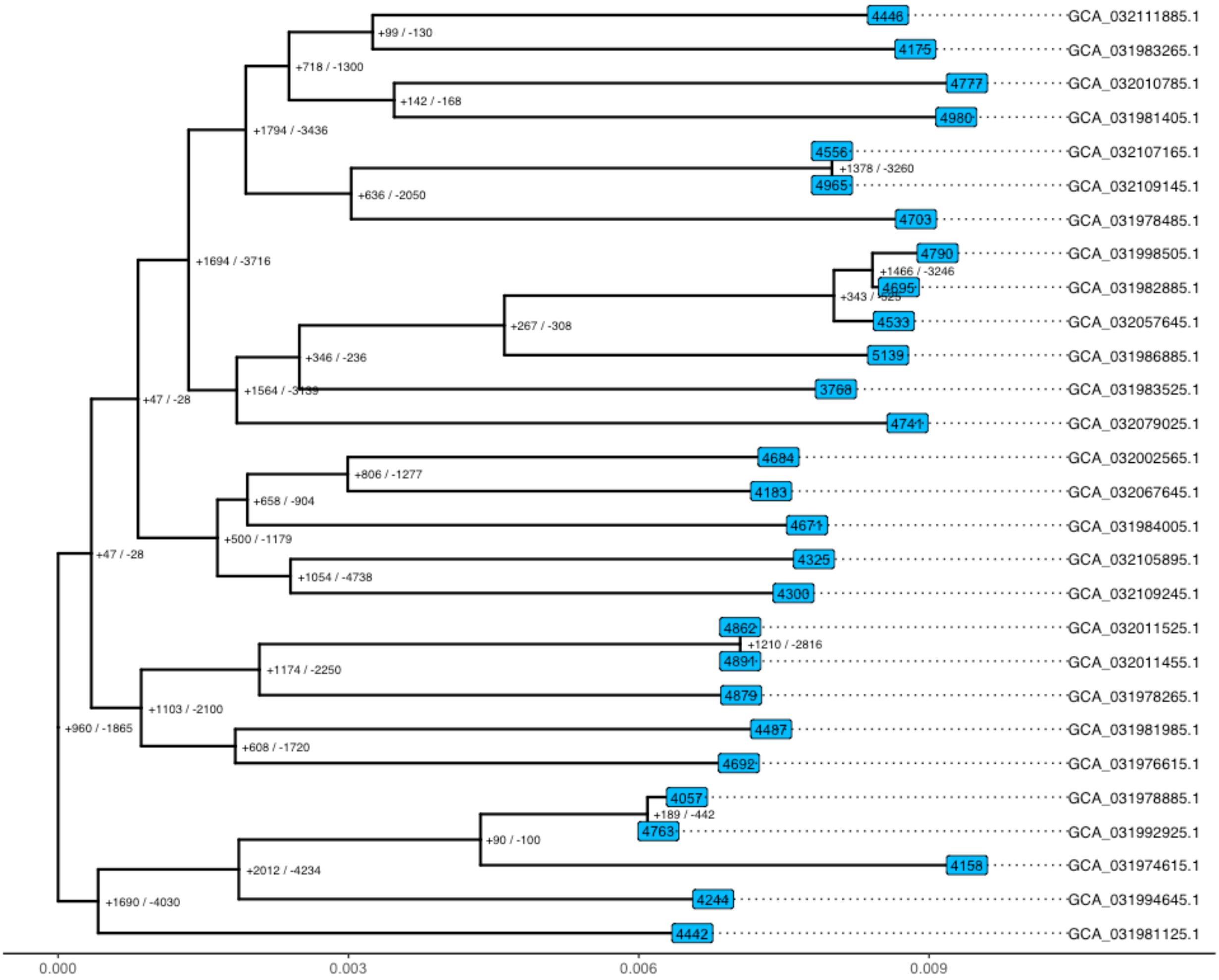


Figure 4.6. Gene gain-loss counts in *K. pneumoniae* phylogenetic tree pruned from MPKG dataset phylogeny

In the MPKG whole-tree projection, terminal branches are widely loss-skewed, whereas gains are concentrated on a minority of short, shallow branches and on several interior branches that subtend multi-tip clusters of Klebsiella pneumoniae (Figure A4.1). Within those clusters, most daughter terminals retain a loss excess even when their immediate ancestor carries elevated gain counts, and pairs of sister terminals with comparable branch lengths frequently differ in both the sign and magnitude of their terminal balances. The *K. pneumoniae* pruning rendered at higher resolution shows the same architecture, multiple interior branches with elevated gains followed by terminals in which losses outnumber gains by several-fold, interspersed with a smaller set of terminals where gains approach parity (Figure 4.6). The combination of (i) interior branches with concentrated gains, (ii) terminal branches with pronounced loss excess, and (iii) large between-sister differences despite similar branch lengths demonstrates that raw totals are a poor proxy for the pace of change on individual branches and motivates the use of branch-length–standardized λ and μ for cross-lineage comparison (Figure A4.1; Figure 4.6) [219, 235]. The local enrichment of gains on shallow branches is compatible with recent influx of mobile elements, whereas the prevailing terminal loss excess accords with well-documented deletional pressures in bacterial chromosomes [318, 325, 326].

In the isolate cohort spanning multiple species (PKG), tree-mapped counts show a pervasive excess of losses on both internal and terminal branches (Figure A4.2). Deep internal branches that delimit species complexes contain long tracts of inferred deletions, and most tips remain strongly loss-skewed despite high assembly completeness, indicating a biological signal rather than an assembly artefact (Figure A4.2). Within this full tree, clusters assigned to *K. michiganensis* carry the heaviest terminal burdens, with uniformly large loss counts across adjacent tips, consistent with the species-level per-genome summaries.

The *K. pneumoniae* subtree pruned from the same phylogeny displays the same architecture at higher resolution, occasional interior branches with elevated gains are followed by terminals in which losses outnumber gains in nearly all genomes (Figure 4.7). No tip-set within the *K. pneumoniae* pruning exhibits a persistent gain-dominated regime. Together, the PKG trees support sustained reductive dynamics in isolates, most pronounced in *K. michiganensis* on the full tree and evident across *K. pneumoniae*, with only infrequent, localized gain episodes superimposed on a background of deletional pressure (Figure A4.2; Figure 4.7), in line with prior reports [312, 318, 325]. Because PKP comprises clinical *Klebsiella pneumoniae* isolates, hospital selection (antibiotics, indwelling devices, host-associated niches) and dense plasmid and prophage traffic are expected to elevate terminal acquisitions despite background genome reduction [312, 323, 324]. Within the clinical *K. pneumoniae* cohort, the whole-tree projection remains loss-skewed but shows a higher frequency of tip-proximal gains than the isolate panel (Figure A4.3). Gains tend to occur on short terminal branches and are interspersed with loss-dominated segments, whereas deep internal branches rarely accumulate large, shared acquisitions. The ST395 subtree, which represents the largest sequence type in the dataset, displays repeated shallow gain events superimposed on persistent terminal losses across closely related isolates (Figure 4.8). This combination, frequent small gains on tips with continued net attrition, is consistent with ongoing exchange of mobile elements in hospital settings coupled with genome-wide deletional pressure, and it explains why the cohort-level balance in PKP is closer to parity than in PKG [312, 319, 323]. The within-ST dispersion in terminal balances further indicates that closely related clinical lineages can differ appreciably in recent accessory-genome editing, a pattern widely noted for Klebsiella pneumoniae populations undergoing plasmid and prophage traffic [319, 325, 326].

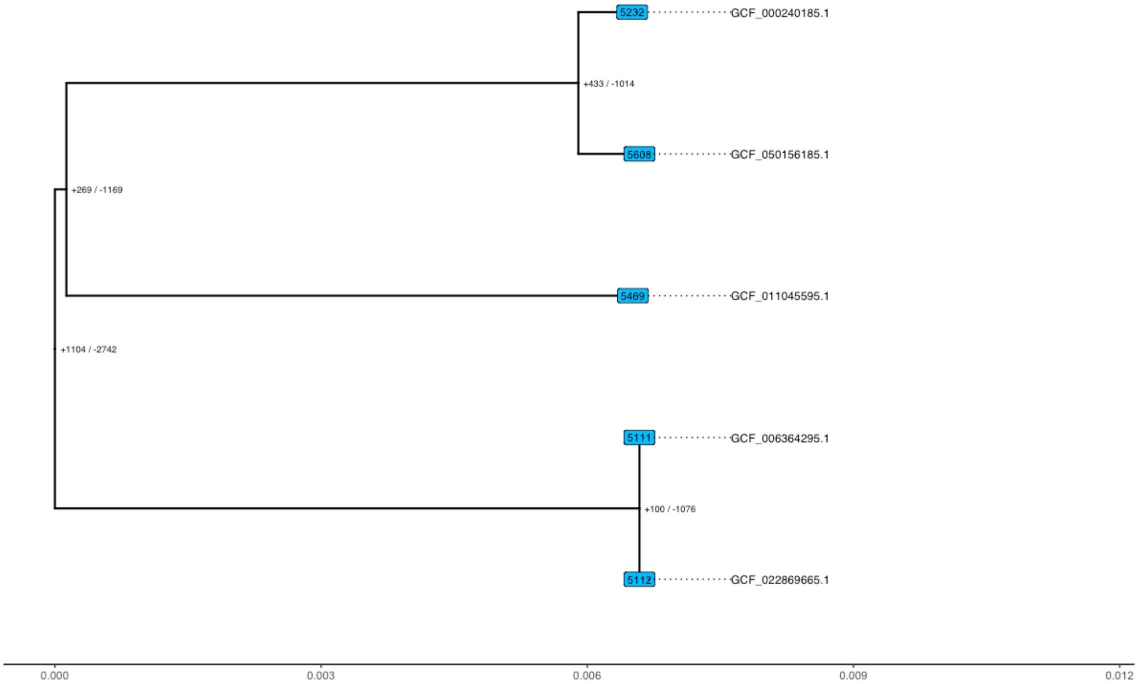


Figure 4.7. Gene gain-loss counts in *K. pneumoniae* phylogenetic tree pruned from PKG dataset phylogeny.

Read across the three trees, three observations are consistent: (i) loss predominance is universal, with a dataset-dependent effect size that is strongest in high-contiguity isolate trees (PKG), (ii) intermediate in clinical *K. pneumoniae* (PKP), and (iii) more moderate in MAGs (MPKG), where fragmented assemblies under-recover short, mobile-element–rich contigs and can inflate inferred terminal losses [19, 327, 328]. The comparatively higher gain fraction in PKP versus PKG is consistent with healthcare-associated plasmid exchange and AMR-gene mobilization repeatedly documented for *K. pneumoniae* [312, 323].

Internal-branch gains occur but are limited in scope and are followed by heterogeneous terminal outcomes among sister tips. Because tip-wise raw counts correlate weakly with branch length, comparisons among closely related isolates are made using branch-length–standardized estimates of gain (λ) and loss (μ), which express expected transitions per unit evolutionary time [219].

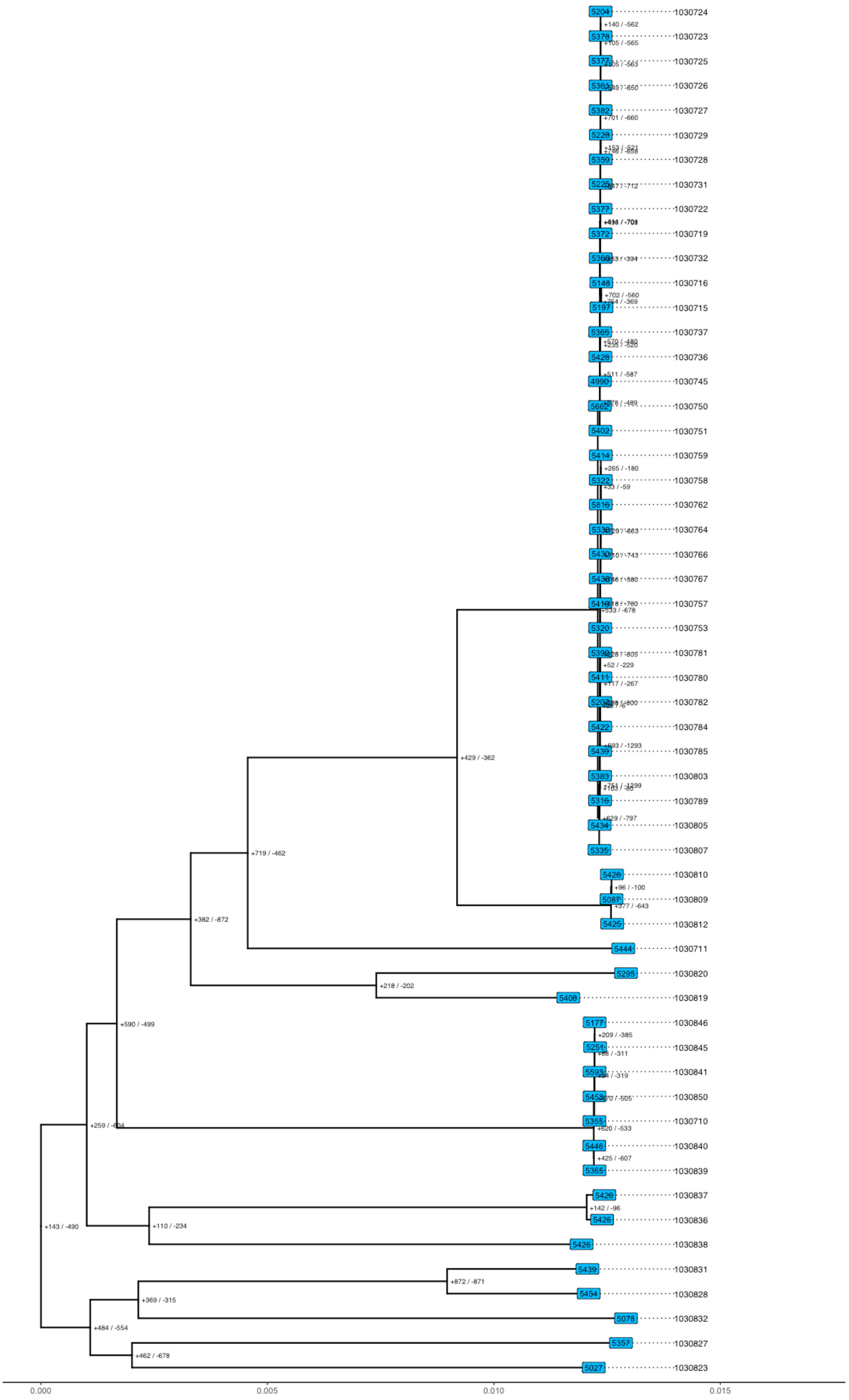


Figure 4.8. Gene gain-loss counts in *K. pneumoniae* ST395 sequence type phylogenetic tree pruned from PKP dataset phylogeny.

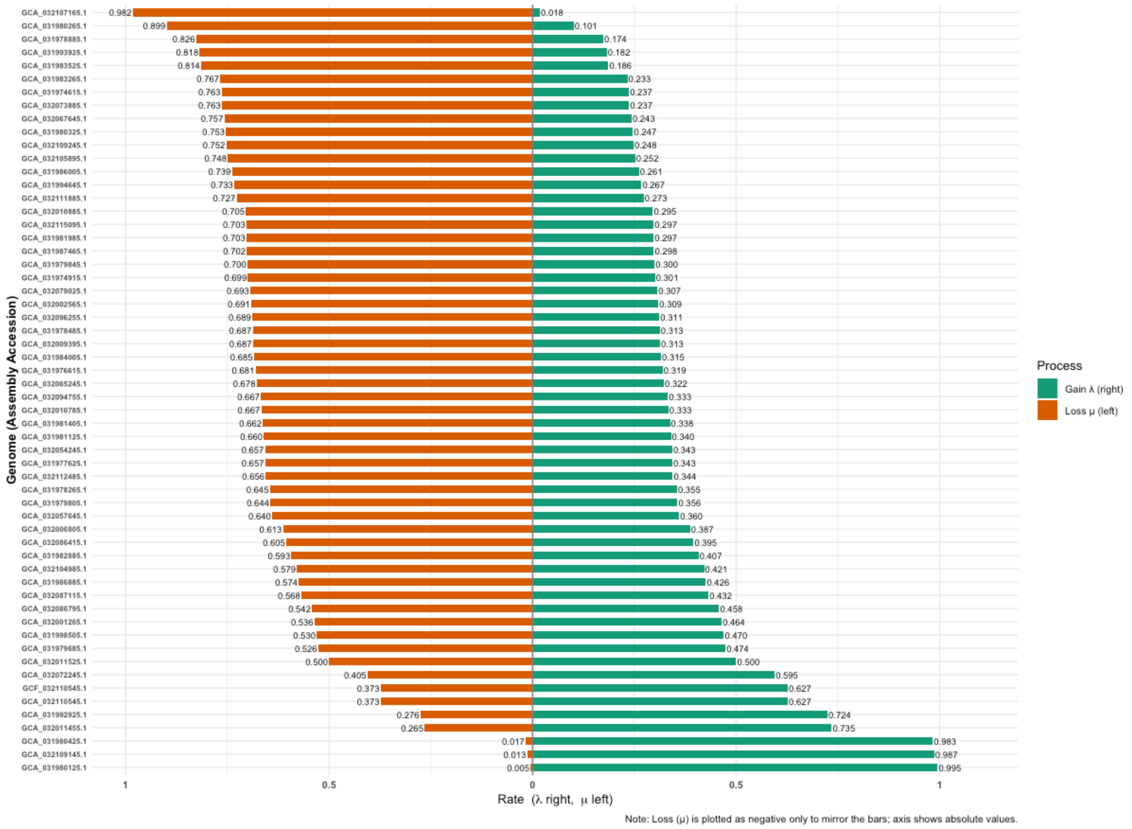
4.1.4. Pangenome gain-loss rate quantification

Event counts measure the magnitude of genome editing but confound it with evolutionary time, since identical totals can accumulate slowly along long branches or rapidly along short ones. To separate these dimensions, we estimated genome-specific gain (λ) and loss (μ) under an all-rates-different continuous-time Markov framework and normalized by branch length, yielding expected acquisitions and deletions per unit evolutionary time [219, 235]. Read together with counts, these branch-length–standardized rates help distinguish changes concentrated on internal branches from ongoing turnover at the tips.

Table 4.3. Summary gene gain-loss rates by species in MPKG, PKG.

Species	Nr. genomes	Gain rate (mean)	Loss rate (mean)
		MPKG dataset	
K. variicola	7	0.329	0.671
K. pneumoniae	28	0.356	0.644
K. oxytoca	8	0.375	0.625
K. aerogenes	3	0.376	0.624
K. michiganensis	10	0.436	0.564
K. africana	2	0.627	0.373
		PKG dataset	
K. aerogenes	5	0.143	0.857
K. michiganensis	5	0.162	0.838
K. oxytoca	5	0.210	0.790
K. africana	5	0.246	0.754
K.variicola	5	0.250	0.750
K. huaxiensis	4	0.287	0.713
K. penumoniae	5	0.401	0.599

Across datasets, branch-length–standardized estimates of gain (λ) and loss (μ) resolve a consistent yet dataset-specific picture of accessory-genome turnover. In the urban MAG panel (MPKG), per-unit-time gain–loss rates span a broad range across genomes, but the median pattern is a moderate, consistent bias toward losses (for most species λ≈0.33–0.44 and μ≈0.56–0.67), with substantial among-genome heterogeneity (Figure 4.9). Most species center around λ≈0.33–0.44 with the complementary μ≈0.56–0.67 (Table 4.3). Within this range, both *K. pneumoniae* (λ≈0.356, μ≈0.644) and *K. variicola* (λ≈0.329, μ≈0.671) show a reproducible deletional bias, the latter is slightly more loss-skewed, implying stronger net attrition of accessory loci [312]. By contrast, *K. michiganensis* exhibits the highest relative acquisition tempo among well-sampled MAG taxa (λ≈0.436, μ≈0.564), consistent with frequent uptake of mobile cargo in urban compartments where gene flow is intense [329, 330].



Note: Loss (μ) is plotted as negative only to mirror the bars; axis shows absolute values.

Figure 4.9. Gene gain-loss rates per genome in MPKG meta-pangenome.

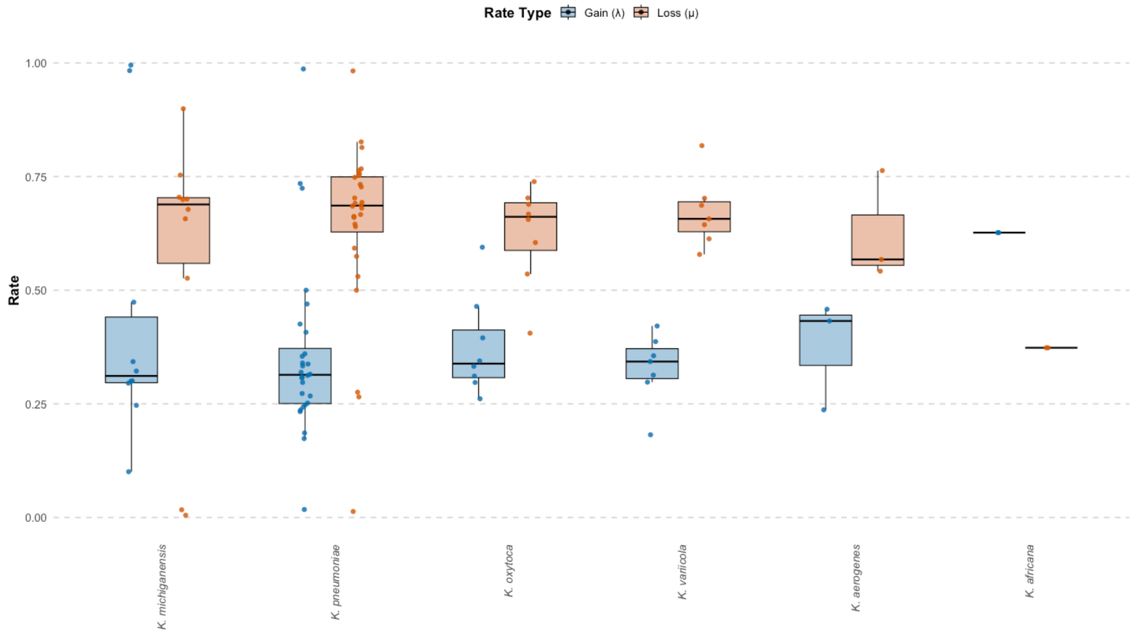
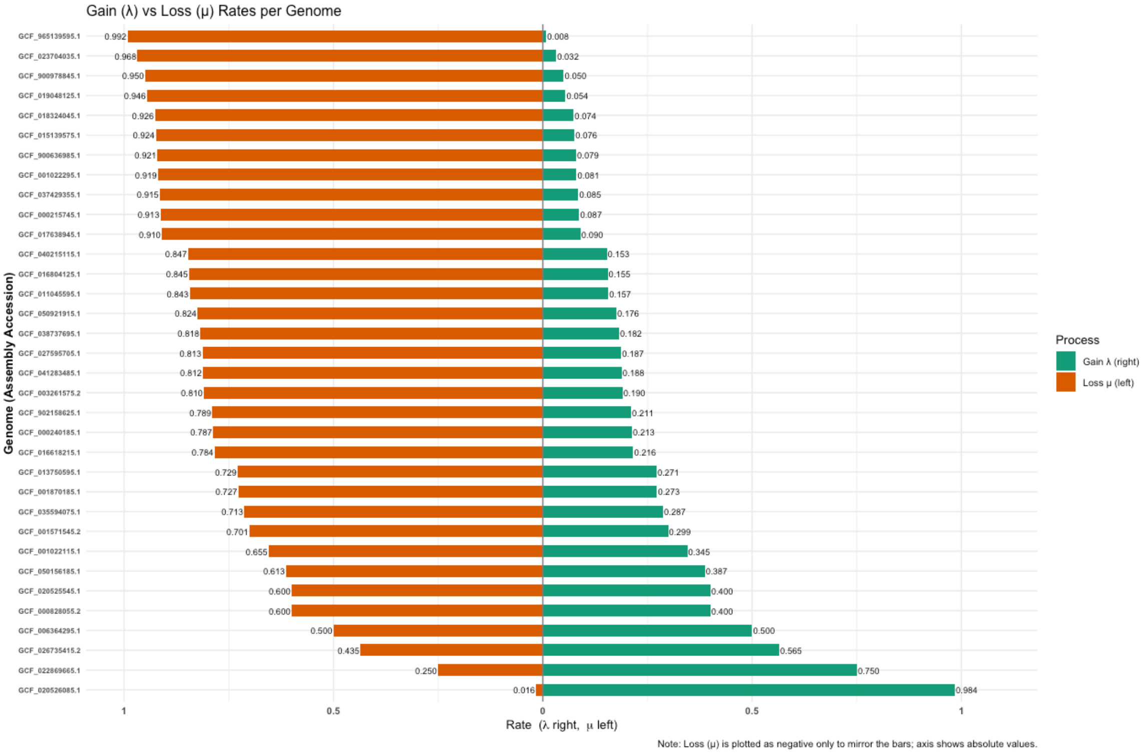


Figure 4.10. Distribution of gene gain and loss rates per species in MPKG meta-pangenome.

An apparent gain-biased mean for *K. africana* (λ≈0.627, μ≈0.373) is observed across only two MAGs, given the small n and known under-recovery of short, mobile-element–rich contigs in MAGs, we report the point estimates without drawing species-wide conclusions (Figures 4.9–4.10; Table 4.3). Genome-wise rate bars in MPKG are heavy-tailed, notably a minority of tips approach λ≥0.45 while sister tips with comparable branch lengths remain strongly loss-skewed, indicating that among-genome differences in λ and μ are driven by lineage biology rather than evolutionary depth alone (Figure 4.9), a pattern corroborated by species-level distributions (Figure 4.10). The modest overall skew in MPKG is also compatible with partial under-recovery of short accessory contigs in fragmented MAGs [288, 327].



Note: Loss (μ) is plotted as negative only to mirror the bars; axis shows absolute values.

Figure 4.11. Gene gain-loss rates per genome in PKG pangenome.

In the cultured-isolate panel (PKG), the loss bias is strongest. Across species, λ contracts to ≈0.14–0.29 while μ rises to ≈0.71–0.86 (Table 4.3), yielding narrowly distributed, uniformly loss-skewed genome profiles (Figure 4.11). *K. aerogenes* (λ≈0.143, μ≈0.857) and *K. michiganensis* (λ≈0.162, μ≈0.838) exemplify high deletional pressure, whereas *K. huaxiensis* lies toward the gain-richer end of the spectrum (λ≈0.287, μ≈0.713). Across species, the rate distributions are well separated, notably *K. michiganensis* shows a pronounced loss excess with wide among-genome dispersion, whereas *K. variicola* and *K. oxytoca* cluster tightly at high μ and low λ (Figure 4.12).The compressed λ and consistently high μ in PKG are expected when high-contiguity assemblies reveal genuine deletions rather than missing contigs and align with long-recognized deletional bias and reductive evolution in bacteria [318, 325].

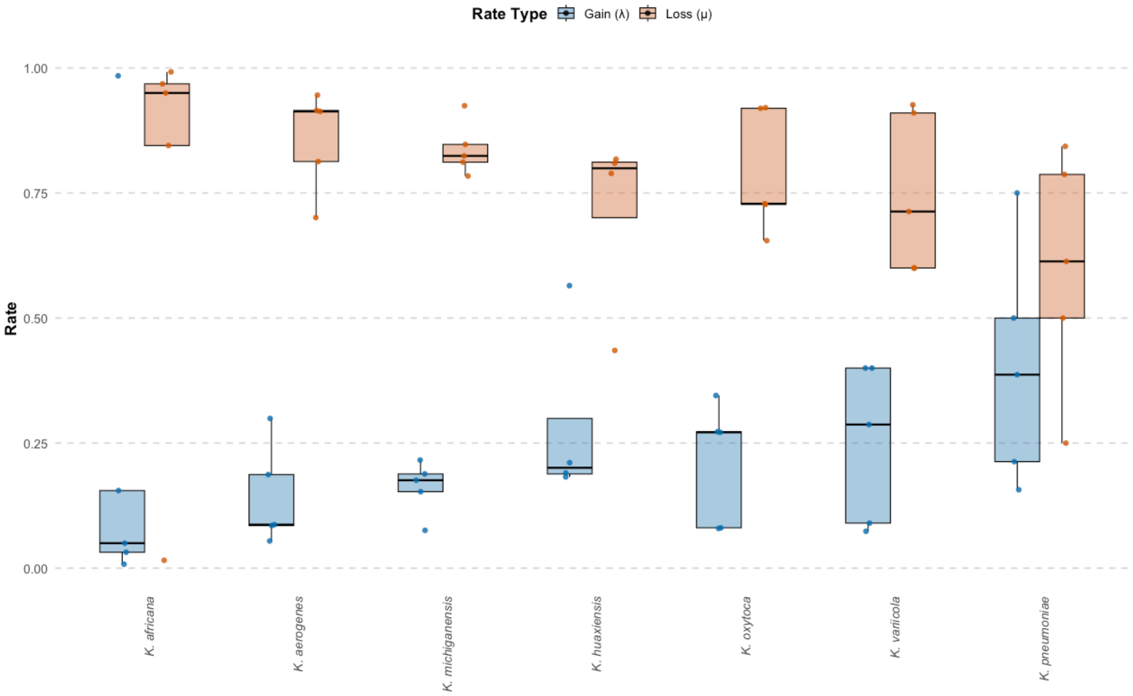


Figure 4.12. Distribution of gene gain and loss rates per species in PKG pangenome.

Within the clinical *K. pneumoniae* dataset (PKP), genomes and sequence types converge on an intermediate bias between MPKG and PKG. Genome-level estimates (Figure 4.13) and ST-resolved summaries (Figure 4.14; Table A5.4) show that the dominant ST395 (n=56) centers at λ≈0.410, μ≈0.590, with active turnover but net loss, whereas several clinically important or widely distributed STs display substantially higher acquisition tempo, notably ST15 (λ≈0.575, μ≈0.425), ST405 (λ≈0.563, μ≈0.437), ST1037 (λ≈0.543, μ≈0.457), ST101 (λ≈0.492, μ≈0.508), and elevated λ also in ST11 and ST147 (Table A5.4). The within-ST dispersion is large, especially in ST395, indicating that even closely related hospital lineages differ in near-term exchange and purging of accessory cargo. This structure is characteristic of episodic plasmid and prophage influx superimposed on background deletional dynamics in healthcare settings, where antibiotic exposure, devices, and dense patient-to-patient transmission elevate opportunities for horizontal transfer [312, 313, 323].

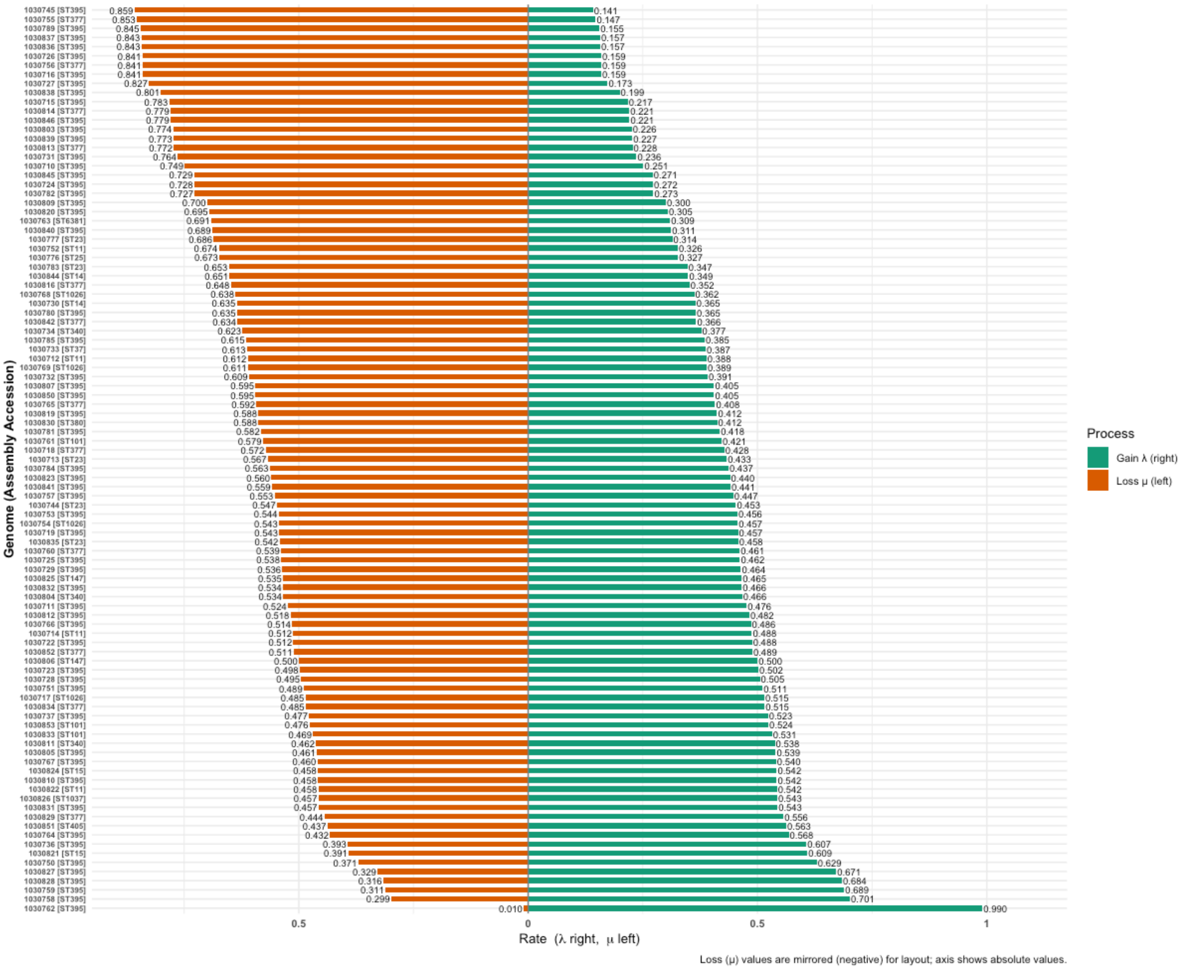


Figure 4.13. Gene gain-loss rates per genome in PKP pangenome (clinical *K. pneumoniae* isolates).

Across datasets, the estimated gain (λ) and loss (μ) rates show a reproducible pattern. Losses dominate in every cohort, and the magnitude of this bias depends on data type, λ is typically ~0.33–0.44 in MAGs, ~0.14–0.29 in isolates, and ~0.36–0.58 in clinical *K. pneumoniae*, with sequence-type. Lineage context shapes these signals, internal branches often carrying clusters of gains that establish clade-specific accessory backbones, whereas closely related tips diverging in time, with heavy-tailed λ in MPKG, compressed low λ with uniformly high μ in PKG, and ST-stratified λ in PKP. Differences among cohorts reflect both measurement and biology. Fragmented MAGs can under-recover short mobile-element contigs, which depresses apparent λ or inflates μ, yet the persistent upper tails of λ in MPKG indicate genuine rapid gene influx in specific environmental lineages [328, 331]. In contrast, complete isolate assemblies reveal long tracts of deletion and the strongest loss bias (Figures 4.11–4.13), consistent with well-documented deletional pressure and genome streamlining in bacteria [32, 318, 325].

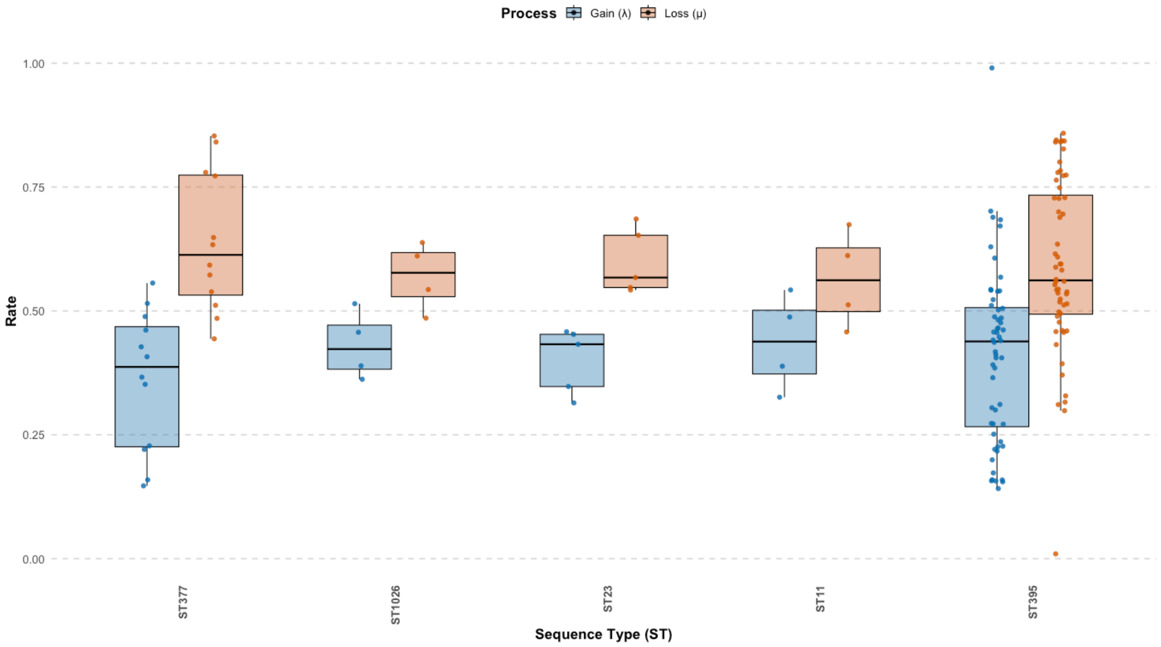


Figure 4.14. Gene gain loss rate distribution by ST type in PKP pangenome (STs with ≥ 4 genomes).

Two caveats qualify interpretation, namely that estimates for small groups such as *K. africana* in MPKG (n=2) are dominated by sampling uncertainty and are reported without species-level generalization, and that because λ and μ are computed per unit branch length, comparisons presume an internally consistent molecular clock within each cohort, with violations expected to widen dispersion without altering the cross-cohort ordering [332–334]. In summary, the results indicate discontinuous, lineage-structured gene turnover with a net excess of losses, environment-specific and sequence-type–specific acquisition regimes generate the upper tails in λ and plausibly explain the higher gain fraction in the clinical PKP cohort relative to the multi-species PKG panel, consistent with documented plasmid and prophage exchange and AMR-gene mobilization in *K. pneumoniae* [312, 323, 335, 336].

4.1.5. Gain-loss rates mapping onto meta-pangenome phylogenies

Projecting branch-length–standardized gain (λ) and loss (μ) onto the phylogenies resolves where turnover concentrates along lineages rather than across samples. In the MAG panel, the full MPKG tree shows a broadly loss-skewed background punctuated by compact clusters of elevated λ on short, tip-proximal branches; deeper internal paths remain moderate for both processes, consistent with founder acquisitions followed by heterogeneous terminal outcomes rather than uniformly accelerated change across clades (Figure A4.4).

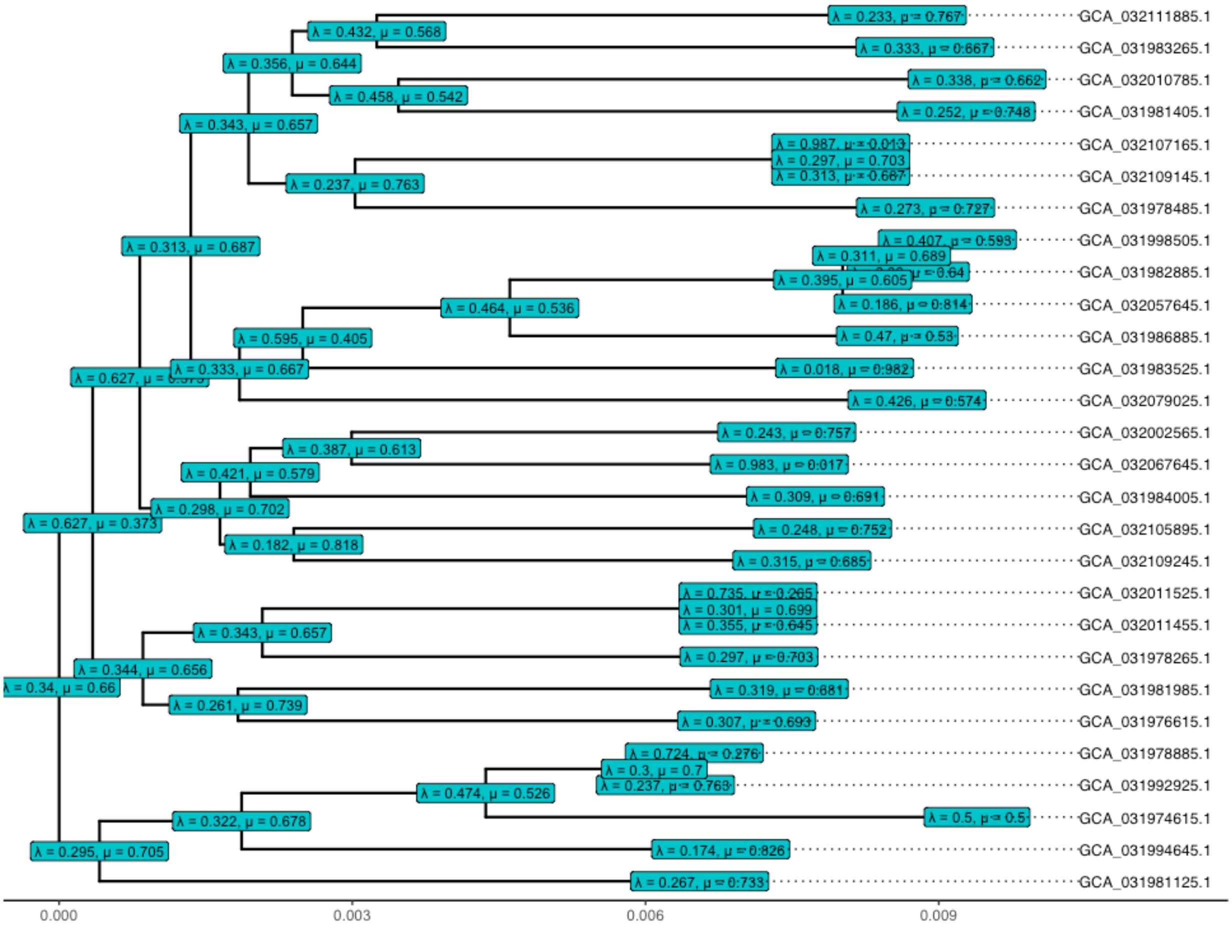


Figure 4.15. Gene gain-loss rates for *K. pneumoniae* mapped on phylogenetic tree (pruned from MPKG dataset).

Pruning to *Klebsiella pneumoniae* sharpens this picture, with multiple subclades displaying independent, shallow λ spikes embedded within a μ-dominated scaffold, closely related tips with comparable branch length frequently diverge in rate balance, which argues for lineage-specific access to mobile DNA and differential removal of recently acquired genes rather than time-alone effects (Figure 4.14).

Isolate trees (PKG dataset) accentuate reductive dynamics, notably long segments of adjacent branches carry consistently high μ with only sparse, low-amplitude λ excursions, and within the *K. pneumoniae* pruning the short tips rarely overturn the loss background, clades assigned to *K. michiganensis* show the most coherent runs of elevated μ across neighboring tips, mirroring the strong isolate-level loss bias reported from the rate summaries (Figure A4.5; Figure 4.16). The continuity of these high-μ tracts across adjacent lineages supports sustained gene removal within these backgrounds and is not readily explained by assembly gaps, which would be expected to affect contiguous taxa idiosyncratically rather than in blocks.

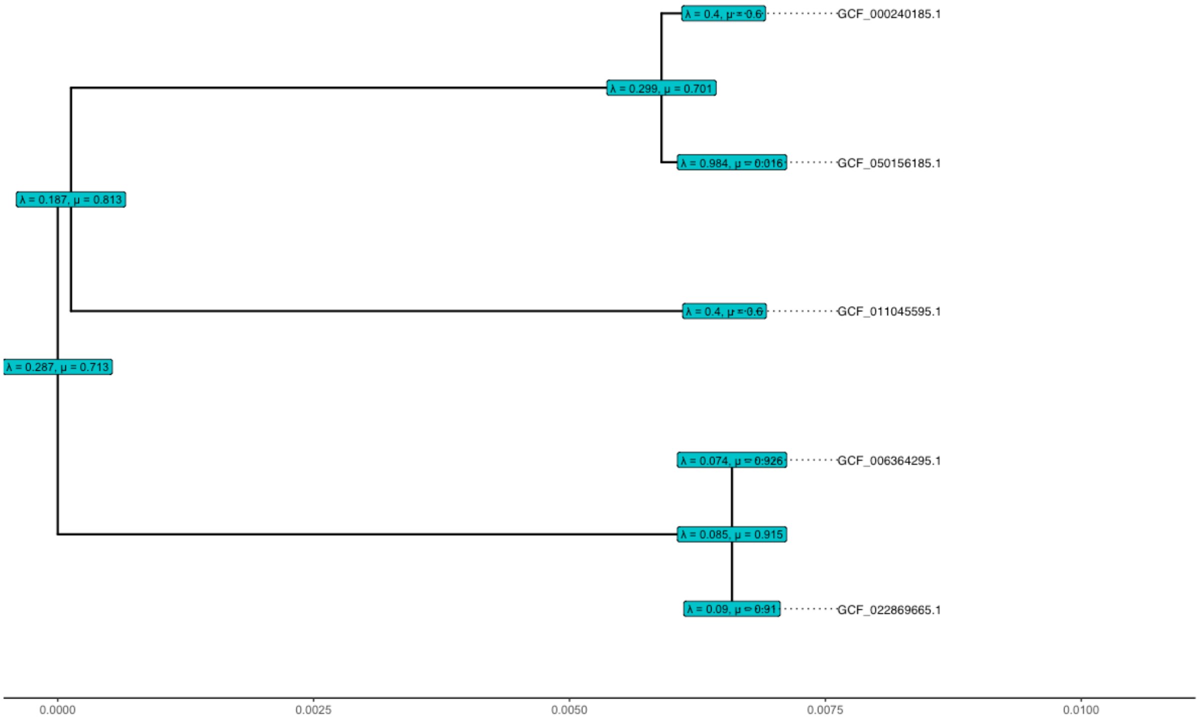


Figure 4.16. Gene gain-loss rates for *K. pneumoniae* mapped on phylogenetic tree (pruned from PKG dataset).

Clinical *K. pneumoniae* occupies an intermediate regime. The whole PKP tree preserves a loss-skewed backbone yet includes numerous tip-proximal elevations in λ, a pattern compatible with ongoing plasmid and prophage exchange in hospital environments (Figure A4.6).

The dominant sequence type, ST395, is internally stratified: several sublineages approach or exceed roughly one-half of events per unit branch length as gains on short branches, while adjacent tips remain loss-dominated, indicating repeated but uneven accessory influx superimposed on background removal (Figure 4.17). This within-ST dispersion aligns with the cohort-level rate distributions that are closer to balance than in PKG, and with the well-documented, episodic movement of resistance and cargo plasmids through *K. pneumoniae* populations under clinical selection [323, 337, 338].

Loss events predominate across all datasets, although the magnitude of the bias differs by data type, being strongest in high-contiguity isolate panels (PKG), intermediate in the clinical *K. pneumoniae* collection (PKP), and weakest in MAGs once tip-proximal acquisition bursts are considered. Founder-like gains on internal branches establish clade-level accessory backbones, after which closely related tips follow divergent trajectories, with acquisition-rich pockets confined to specific subclades and short terminal segments rather than distributed uniformly across the tree (Figures A4.4–A4.6; Figures 4.15-4.17).

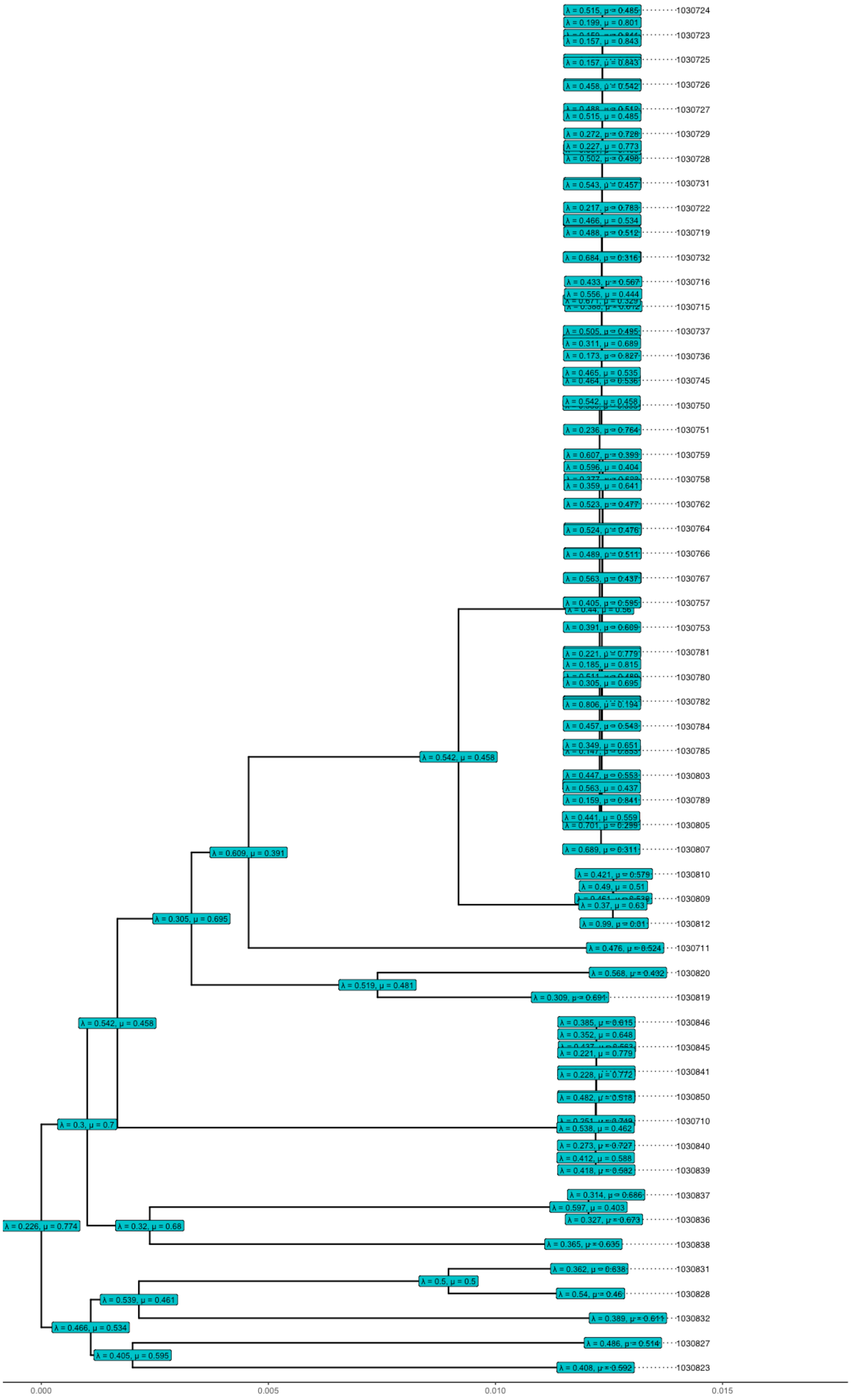


Figure 4.17. Gene gain-loss rates for *K. pneumoniae* ST395 sequence type mapped on phylogenetic tree (pruned from PKP dataset).

Differences between datasets reflect both biology and measurement, relating that fragmented environmental assemblies can under-recover short, mobile-element–rich contigs and thereby depress apparent gain rates or inflate loss rates, yet acquisition-rich tails persist in MPKG, indicating genuine rapid influx in a subset of environmental lineages, by contrast, complete isolate genomes reveal contiguous tracts of gene loss and the strongest deletional skew, consistent with long-recognized deletional bias in bacteria [319, 325] and with known limitations of MAG recovery for accessory elements [174].

Estimates for very small groups, such as *K. africana* in MPKG (n=2), are dominated by sampling uncertainty and are reported without species-level generalization; comparisons of λ and μ among tips further assume a consistent molecular-clock (branch-length) calibration within each cohort [332, 339].

4.2.Gene-wise selection patterns pressure estimation

To investigate the evolutionary pressure shaping the structure of bacterial meta-pangenomes or pangenomes, we developed the PGGS (Pangenome Gene Selection), an R-implemented classifier built on the `phytools` package (Algorithm 3; Algorithm A3.2) [264]. PGGS operates downstream of PGGL-Gene (Algorithm 1) by using the per-gene maximum-likelihood gain and loss rates ($\hat{\lambda}_g, \hat{\mu}_g$) estimated on a rooted, branch-length-calibrated phylogeny, it performs model-selection-based classification of turnover bias for every orthologous group. For each gene, PGGS compare a symmetric history (ER, $\lambda = \mu$) with an asymmetric history (ARD, $\lambda \neq \mu$), evaluates support with Akaike's Information Criterion (AIC) [279, 280], and report the best-supported regime together with direction (gain- versus loss-biased) and magnitude, summarized by the selection index $\log(\hat{\lambda}_g/\hat{\mu}_g)$ and selection score $(\hat{\lambda}_g - \hat{\mu}_g)$/$(\hat{\lambda}_g + \hat{\mu}_g)$. Likelihoods are computed on fixed tree with standards two-state CTMC pruning [14, 31, 271] (Algorithm 1). The software returns a ranked, tidy table with estimates, including ΔAIC, selection index and selection score, suitable for figure generation and functional enrichment. Biologically, ER denotes no directional bias in long-term turnover, presence and absence being exchangeable after accounting for shared ancestry and branch lengths, reflecting nearly neutral or fluctuating selection [340–342]. Equal-rate model is therefore consistent with: (i) effectively neutral or fluctuating selection that averages to zero across lineages and environments; (ii) a balance between gene supply (HGT) and deletion processes; or (iii) dynamics dominated by phylogenetic inertia rather than consistent gains or consistent losses. Importantly, ER does not mean "no evolution", with $\lambda = \mu$ small, a gene can appear nearly core because both gains and losses are rare, and with $\lambda = \mu$ large, the same can be highly labile yet show no preferred direction. Data patterns that support ER include gains and losses dispersed across the tree without monotonic trends, posterior probabilities at internal nodes that do not consistently drift upward or downward, and AIC indicating that allowing $\lambda \neq \mu$ provides negligible improvements ($\Delta AIC \leq 2$). In stationary ER, the long-run equilibrium frequency is 0.5, but on finite trees the realized frequency also reflects the root prior and total tree length [14, 31, 264, 271].

Algorithm 3. PGGS (Pangenome Gene Selection) algorithm for estimating selection pressure and bias in pangenome genes.

```
Inputs:
  • Rooted, strictly bifurcating tree $T$ with branch lengths; tip order $L$.
  • Binary matrix $X \in \{0,1,?\}^{n\times G}$ (rows = tips of $T$).
  • Frequency window for informative genes: $f_{min}, f_{max}$ (defaults 0.05, 0.95).
  • Root prior: uniform [1/2,1/2] or stationary.
  • (For PGGL-Gene calls) targe state $s^*$ (default 1), event threshold $\delta$ (default = 0.05)
Outputs: Table R with one row per gene $g$: $\hat{\lambda}_g, \hat{\mu}_g$, $sel\_index_g$ $(selection\ index)$, $sel\_score_g$ $(selection\ score)$, $\ell_{ER}$, $\ell_{ARD}$, $AIC_{ER}$, $AIC_{ARD}$, $\Delta AIC$, $sel\_class$ $(selection\ class)$
Procedure:
  1. Align rows to tips. Reorder $X$ so its rows match $L$.
  2. Filter genes. Keep $g$ with $var(X_{\cdot g}) > 0$ and $f_{min} < \frac{1}{n}\sum_{i=1}^{n} X_{ig} < f_{max}$.
  3. Initialize $R \leftarrow \emptyset$.
  4. For each gene $g$ in the filtered set:
       • Set $y \leftarrow X_{\cdot g}$.
       • ER fit via Algorithm 1 (PGGL-Gene):
         call PGGL-Gene ($T, y, s^*, \delta = 0$, model = ER, prior) → get single rate $\hat{r}_g$ and log-likelihood $\ell_{ER}$.
         ARD fit via Algorithm 1 (PGGL-Gene):
         call PGGL-Gene ($T, y, s^*, \delta = 0$, model = ARD, prior) → get single rate $(\hat{\lambda}_g, \hat{\mu}_g)$ and log-likelihood $\ell_{ARD}$.
       • Information criteria:
         $AIC_{ER} = -2\ell_{ER} + 2$, $AIC_{ARD} = -2\ell_{ARD} + 4$, $\Delta AIC = AIC_{ER} - AIC_{ARD}$.
       • Selection summaries:
         $sel\_index_g = \log\left(\frac{\hat{\lambda}_g+\varepsilon}{\hat{\mu}_g+\varepsilon}\right)$, $sel\_score_g = \frac{\hat{\lambda}_g-\hat{\mu}_g}{\hat{\lambda}_g+\hat{\mu}_{g+\varepsilon}}$.
       • Class label (by $\Delta AIC$):
         NS if ≤ 2; WS if 2 <≤ 4; MS if 4 <≤ 10; SS if > 10
       • Append row for $g$ to $R$.
  5. Return R.
```

The assymetric rates model ($\lambda \neq \mu$) captures a directional bias in turnover, attributable to long-term selection, with two distinct biological interpretations, specifically, the cases when $\mu > \lambda$ (loss-biased) and $\lambda > \mu$ (gain-biased). In loss biased cases ($\mu > \lambda$), repeated, lineage-independent losses outweigh gains. This pattern is expected when deletional mechanisms and/or purifying selection against maintenance cost dominate—classical "streamlining", so dispensable or conditionally useful genes are pruned more often than they are reacquired [11, 325]. Typical data signatures include deep or intermediate ancestral presence with nested, clade-specific absences, posterior presence that declines along internal backbones, negative selection index $\log(\hat{\lambda}/\hat{\mu}) < 0$, and substantial AIC support for ARD (ΔAIC>2). Complementary, in gain-biased cases ($\lambda > \mu$) independent acquisitions outpace losses. This arises when a gene is repeatedly supplied (e.g., by mobile elements) and confers advantages often enough to be retained, yielding patchy yet convergent presence across distant lineages. Empirically, gain-biased genes are enriched for horizontally transferred, niche-adaptive modules (AMR determinants, efflux/secretion systems, carbohydrate-use islands, prophage/ICE cargo), and event maps show "seeding" on internal branches followed by tip-proximal retention [325, 343, 344]. The selection index gain-biased cases are positive, and ARD is AIC-favored.

As a tool for meta-pangenome gene classification, PGGS emphasizes phylogeny-aware inference and interpretable, portable outputs. Frequency-only heuristics (core/shell/rare partitions) quantify variability but cannot distinguish recent localized gains from widespread losses or deep ancestral asymmetries [193, 318, 345]. PGGS evaluates fit on the tree, producing labels that reflect where and how turnover accumulated, providing an actionable classification, including neutral/symmetric or gain-/loss-biased with weak, moderate, or strong support, together with effect sizes suitable for ranking and enrichment [279, 280]. Loss-biased genes flagged by PGGS trace clade-defining attrition on event maps, consistent with deletional bias and streamlining [11, 325, 326], whereas gain-biased genes concentrate on internal "seeding" branches with tip-proximal retention, consistent with horizontally acquired, selectively maintained loci [193, 343, 344, 346].

For practice, PGGS includes a pragmatic frequency window to remove near-invariant cores and ultra-rare singletons that drive boundary estimates, improving identifiability without altering qualitative conclusions. Classifications remain robust under either root prior (uniform or stationary) and under modest changes to branch lengths and optimizer settings, with ΔAIC orderings for strongly asymmetric genes conserved across runs [279, 280]. Computational cost grows linearly with the number of genes and edges, enabling parallel execution at cohort scale. In applied analyses, PGGS consistently recovers the expected predominance of loss-bias (genome streamlining) and highlights focused pockets of gain-bias enriched for mobile, surveillance-relevant functions, thereby converting binary gene histories on a fixed phylogeny into a reproducible, interpretable map of directional selection on gene content that complements event mapping and improves upon frequency-only summaries for prioritizing genes of biological and public-health interest.

4.2.1. *Assessment of selection pressure on pangenomes in Klebsiella genus*

We next asked how strongly directional forces shape gene-content turnover across Klebsiella lineages and whether such forces differ between metagenome-derived and isolate-derived cohorts.

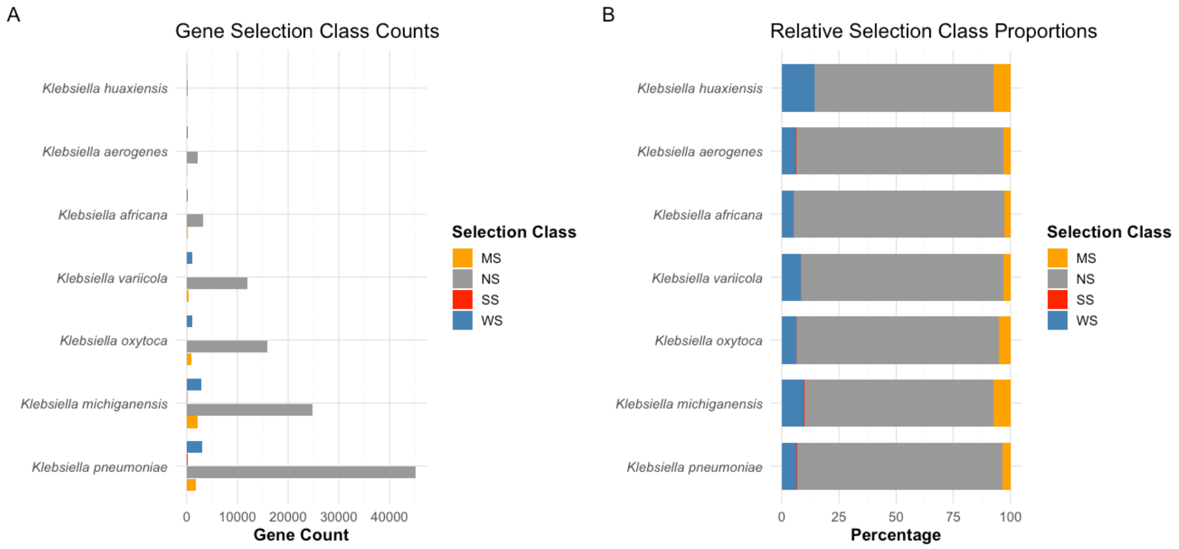


Figure 4.18. Selection class counts and proportions across *Klebsiella* species in the MPKG dataset.

Using the PGGS pipeline on three complementary datasets: (i) MPKG (urban, MAG-based, genus-level), (ii) PKG (isolate-based, genus-level), and (iii) PKP (isolate-based, *K. pneumoniae*), we classified every orthologous group by the best-supported CTMC model (ER versus ARD), and summarized direction and magnitude with the selection index $\log(\hat{\lambda}_g/\hat{\mu}_g)$ and selection score $(\hat{\lambda}_g - \hat{\mu}_g)/(\hat{\lambda}_g + \hat{\mu}_g)$ computed on the fixed phylogeny. In this framework, ER denotes direction-free turnover (nearly neutral or fluctuating selection), whereas ARD denotes directional turnover attributable to long-term selection and/or process asymmetries such as deletional bias or uneven HGT supply.

Across species in MPKG, directional turnover is widespread, a substantial minority of genes falling into the non-neutral classes (WS/MS/SS corresponding to weak, moderate and strong selection signals) in every species, with WS the most common non-neutral label and MS and SS forming smaller but consistent tails; both gain- and loss-biased members are present, and loss-biased classes outnumber gain-biased classes (Figure 4.18A-B). The selection-index–selection-score relationship covers nearly the full theoretical range, with both positive and negative tails well populated, indicating the presence of strong gain- and loss-biased gene subsets (Figure

4.20A). The volcano plot shows a U-shaped relationship between ΔAIC and the selection score; model support for asymmetry increases with the absolute selection score, and genes with strongly positive scores (gain-biased) or strongly negative scores (loss-biased) exhibit the highest ΔAIC values (Figure 4.21A).

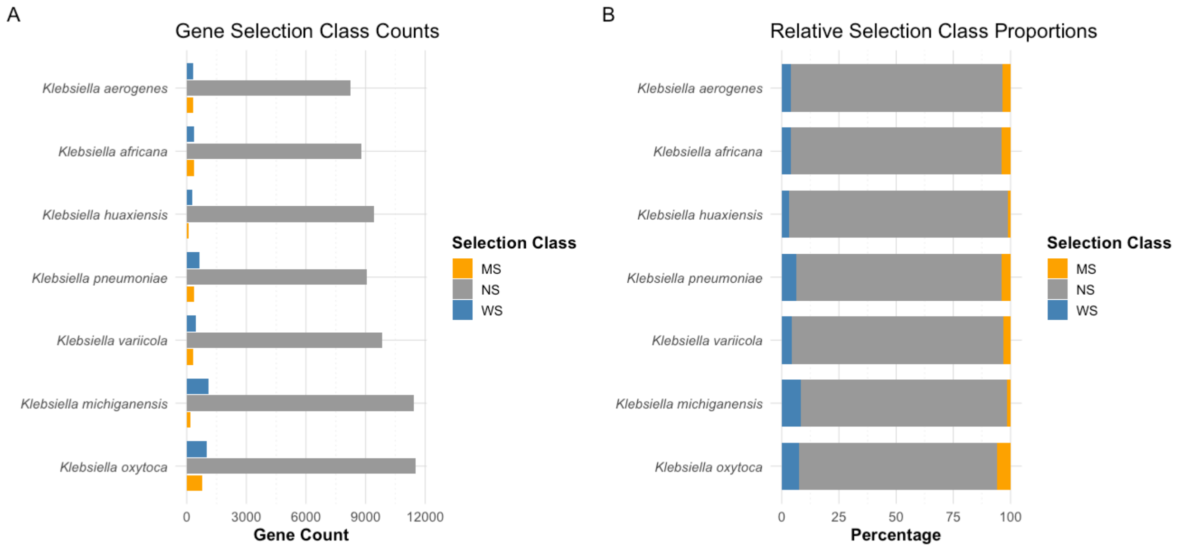


Figure 4.19. Selection class counts and proportions across *Klebsiella* species in the PKG dataset.

Together these patterns indicate that in the ecologically diverse, urban MAG set, directional turnover is common, repeated losses dominate overall (consistent with streamlining), but repeated gains occur in a non-trivial subset, features expected when mobile elements supply niche-adaptive functions that are selectively retained [11, 193, 343, 346].

In the isolate-based PKG pangenome, support for directional gain–loss dynamics is substantially weaker than in the urban MAG-based MPKG cohort. Most genes are NS (neutral selection under ER) with only modest WS/MS (weak and moderate selection) tails across species (Figure 4.19A-B). The S-curve is compressed toward the origin (Figure 4.20B), and the volcano plot concentrates near ΔAIC≈0 with few extreme points (Figure 4.21B), indicating that symmetric turnover explains the majority of presence–absence patterns in this more homogeneous sampling frame [318, 319].

Under current sampling and phylogenetic resolution, classifications are exclusively neutral, selection scores and ΔAIC values collapse at the origin (not shown), implying that symmetric gain–loss dynamics suffice for the *K. pneumoniae* pangenome in this cohort.

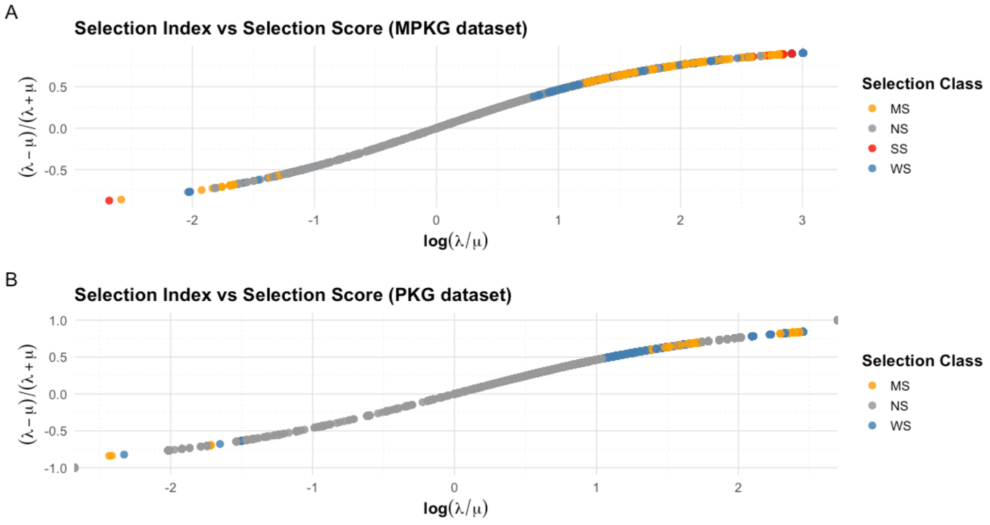


Figure 4.20. S-curve showing selection index versus selection score for the MPKG and PKG datasets.

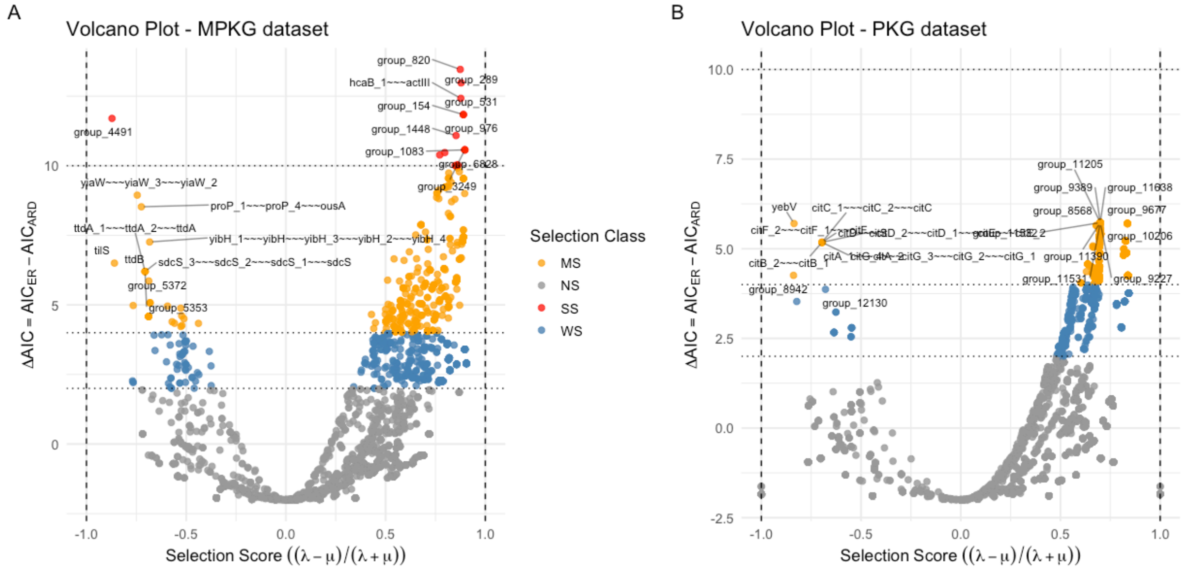


Figure 4.21. Volcano plots of selection score ((λ – μ) / (λ + μ)) versus model support (ΔAIC: ER versus ARD CTMC models) for the MPKG and PKG datasets. Panel A shows results for the MPKG dataset, and Panel B for the PKG dataset. Each point represents an orthologous group, with the x-axis showing the normalized selection score — positive values indicate gain bias, negative values indicate loss bias — and the y-axis showing the ΔAIC between equal-rates (ER) and all-rates-different (ARD) models, reflecting statistical support for asymmetric rates.

The aggregate picture matches modern views of pangenome evolution. A predominance of loss-bias, most evident in MPKG, aligns with deletional pressure and higher streamlining of accessory content [11, 325], whereas focused pockets of gain-bias are consistent with repeated acquisition and retention of mobile, surveillance-relevant modules including AMR determinants, secretion/efflux systems, carbohydrate-use islands, prophage/ICE cargo), where HGT supply is high and selection favors maintenance [193, 343, 346].

By evaluating fit on the tree, PGGS yields phylogeny-aware labels that localize where turnover accumulated and viewed alongside PGGL event maps, reveal clade-backbone attrition for loss-biased genes and internal "seeding" branches with tip-proximal retention for gain-biased genes. These figure-level signals—broad non-neutral tails in MPKG meta-pangenome (Figures 4.18, 4.20A, 4.21A) versus compressed, near-neutral distributions in PKG and PKP pangenomes (Figures 4.19, 4.20B, 4.21B), results that support that ecological heterogeneity and mobile-element supply amplify directional turnover at the genus level, while clonal, species-level isolate cohorts appear closer to symmetric dynamics.

4.3. Benchmarking ancestral state reconstruction on simulated trees

The performance of ancestral state reconstruction (ASR) methods was evaluated using simulated phylogenies with fully known evolutionary histories. Phylogenetic trees were simulated under a Yule (pure-birth) diversification process [347], after which gene presence–absence evolution was simulated along branches using a continuous-time Markov chain (CTMC) with asymmetric gain and loss rates. This design provides an objective benchmark for ASR accuracy, which is generally not achievable with empirical data. Two commonly used approaches were compared: Fitch parsimony [221], Bayesian stochastic character mapping (SCM) [219], and maximum likelihood (ML), developed in this thesis.

The benchmarking results indicate that both maximum likelihood and Bayesian stochastic character mapping achieve consistently high reconstruction accuracy across the phylogeny, whereas Fitch parsimony is constrained by a substantial proportion of unresolved internal nodes (Figure 4.22). Errors under the probabilistic methods are rare and show no systematic association with specific tree regions or depths, suggesting stable performance throughout the evolutionary history. Given its comparable accuracy but markedly lower computational cost, maximum likelihood emerges as the most practical primary reconstruction approach (Figure 4.22). In contrast, Fitch parsimony [221] produces a large number of ambiguous ancestral assignments (Figure 4.22). These ambiguities occur whenever multiple equally parsimonious reconstructions exist and are particularly common in regions characterized by frequent state transitions. Although parsimony correctly infers many resolvable nodes, the high proportion of ambiguous states substantially limits its utility for downstream analyses requiring complete ancestral reconstructions.

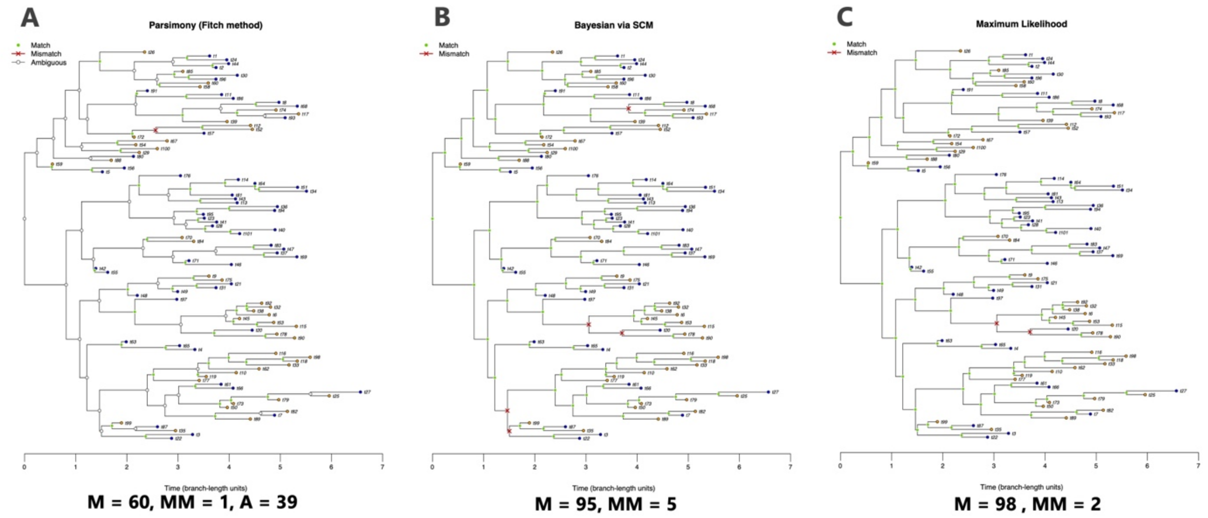


Figure 4.22. Ancestral state reconstructions on the simulated phylogeny obtained using Fitch parsimony (A), Bayesian stochastic character mapping (B), and maximum likelihood (C), with internal nodes annotated as matches, mismatches, or ambiguous relative to the known simulated states.

Quantitative reconstruction outcomes are summarized in Table 4.4, which presents the confusion matrix for each method alongside node coverage. Fitch parsimony reconstructs only 60.8% of internal nodes unambiguously, leaving 392 nodes unresolved. When evaluated solely on resolvable nodes, parsimony attains a raw accuracy of 94.4%. However, once ambiguous nodes are penalized by normalizing correct assignments to the total number of internal nodes, the effective accuracy declines sharply to 57.4%. This result highlights a fundamental limitation of parsimony-based ASR on large phylogenies, although reliable when informative, it frequently fails to provide comprehensive reconstructions.

To facilitate a standardized comparison across methods, the confusion-matrix outcomes were used to derive coverage-adjusted performance metrics, summarized in Table 4.5. These metrics provide a more comprehensive assessment of reconstruction quality by integrating both classification accuracy and the proportion of nodes assigned a definite ancestral state. Both probabilistic approaches achieve complete node coverage, assigning a state to every internal node. Maximum likelihood reconstruction, implemented via marginal likelihood estimation under a discrete-state CTMC [235], attains an effective accuracy of 96.3%. Bayesian SCM, which samples character histories conditional on the observed data and rate matrix [219, 348], achieves an identical effective accuracy of 96.3%. Confusion-matrix–derived metrics further reveal nearly indistinguishable error profiles, with ML producing 21 false positives and 16 false negatives, and SCM producing 22 false positives and 15 false negatives. Balanced accuracy exceeds 0.959 for both methods, indicating robust performance across character states and confirming that reconstruction accuracy is not biased toward either state. Despite exhibiting a high conditional accuracy of 94.4%, Fitch parsimony performs substantially worse once unresolved nodes are incorporated into the evaluation framework. Its effective accuracy drops to 57.4%, reflecting the large fraction of ambiguous reconstructions. This contrast underscores the importance of accounting for reconstruction completeness when benchmarking ASR methods, as conditional accuracy alone may substantially overestimate practical performance.

Table 4.4. Confusion matrix and node coverage for ancestral state reconstruction methods on a simulated phylogeny with 1000 terminal taxa.

Method	TP	TN	FP	FN	Coverage
Maximum Likelihood	596	367	21	16	1.000
Bayesian SCM	697	366	22	15	1.000
Fitch Parsimony	357	217	13	21	0.608

Despite their similar reconstruction accuracy, ML and Bayesian SCM differ substantially in computational efficiency and practical applicability. ML-based ASR relies on deterministic likelihood calculations implemented via dynamic programming and converges rapidly even for large phylogenies. In contrast, Bayesian SCM requires repeated Monte Carlo sampling of full character histories, resulting in substantially higher computational demands and longer runtimes, particularly for large trees or repeated analyses. In the present benchmark, ML reconstruction completed orders of magnitude faster than SCM while yielding indistinguishable point-estimate accuracy.

Table 4.5. 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

These results demonstrate that probabilistic ASR methods strongly outperform parsimony once ambiguity is appropriately accounted for, and that maximum likelihood reconstruction represents the most efficient and practical approach under correct model specification. ML achieves accuracy equivalent to Bayesian SCM while offering superior computational scalability, making it particularly well suited for large phylogenies and high-throughput evolutionary analyses. Bayesian SCM remains valuable as a complementary framework for uncertainty assessment and methodological validation but is not strictly required for accurate ancestral state inference at scale.

4.4.Conclusions to chapter 4

In this chapter, we introduce phylogeny-based methods for quantitative inference of gene-content evolution in microbial pangenomes and evaluates their performance on empirical datasets. The main conclusions of this chapter are as follows:

1. The PGGL (Pangenome Gene Gain–Loss) method was developed and implemented as software to infer phyletic patterns of gene-content evolution by modeling gene gain and loss as a continuous-time Markov process on a fixed species phylogeny, enabling branch-, lineage-, and clade-specific estimation of gain and loss counts and rates as quantitative measures of gene turnover.
2. The PGGS (Pangenome Gene Selection) method was developed and impelemented as software to infer directional biases in gene turnover by contrasting symmetric and asymmetric gain–loss rate models using information-theoretic model selection, enabling gene- and lineage-specific identification and quantification of gain- or loss-biased evolutionary regimes.
3. Maximum likelihood reconstruction under a continuous-time Markov chain (CTMC) achieves accuracy comparable to Bayesian stochastic mapping while substantially reducing the ambiguity associated with Fitch parsimony and, by avoiding computationally intensive sampling, enables reliable and scalable inference of ancestral gene presence–absence states across large phylogenies and genome-scale datasets.
4. Application of PGGL and PGGS to empirical isolate and metagenomic datasets validates both methods, demonstrating recovery of coherent phyletic patterns, predominant loss bias consistent with deletional pressure, and localized gain events associated with mobile and niche-adaptive gene modules, while remaining robust to differences in assembly quality, ecological breadth, and sampling design.

GENERAL CONCLUSIONS AND RECOMMENDATIONS

This thesis demonstrates that metagenomics and pangenomics can be fused into a single, lineage-aware 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 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 specifically for pangenome analyses 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 algorithm and R software implementation (PGGS; Pangenome Gene Gain–Loss Selection) was developed for pangenome analyses, based on 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 and stable gene gain–loss inference 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 pangenome 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. Result reporting should include explicit identifiers for genes, genomes, and branches, along with minimal metadata on data provenance and analytical parameters, to ensure full traceability and reproducibility in downstream studies.
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. 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.
2. Sequencing Platforms | Illumina NGS platforms. 2025. https://www.illumina.com/systems/sequencing-platforms.html. Accessed May 6.
3. Oxford Nanopore flow cells and sequencing devices. 2025. *Oxford Nanopore Technologies*. https://nanoporetech.com/products/sequence. Accessed May 6.
4. Sequencing systems. 2025. *PacBio*. https://www.pacb.com/sequencing-systems/. Accessed May 6.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
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. 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.
16. 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.

17. 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.
18. 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.
19. 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.
20. 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.
21. Csűös, Miklós. 2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26: 1910–1912. https://doi.org/10.1093/bioinformatics/btq315.
22. 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.
23. 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.
24. 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.
25. 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.
26. 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.
27. 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.
28. 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.
29. Kannan, Lavanya, Hua Li, Boris Rubinstein, and Arcady Mushegian. 2013. Models of gene gain and gene loss for probabilistic reconstruction of gene content in the last universal common ancestor of life. *Biology Direct* 8: 32. https://doi.org/10.1186/1745-6150-8-32.
30. Pagel, Mark. 1994. Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings: Biological Sciences* 255. Royal Society: 37–45.

31. 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.
32. Domingo-Sananes, Maria Rosa, and James O. McInerney. 2021. Mechanisms That Shape Microbial Pangenomes. *Trends in Microbiology* 29: 493–503. https://doi.org/10.1016/j.tim.2020.12.004.
33. Didelot, Xavier, and Martin C. J. Maiden. 2010. Impact of recombination on bacterial evolution. *Trends in Microbiology* 18: 315–322. https://doi.org/10.1016/j.tim.2010.04.002.
34. Brito, I. L., S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, et al. 2016. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* 535. Nature Publishing Group: 435–439. https://doi.org/10.1038/nature18927.
35. Ryon, Krista A., Braden T. Tierney, Alina Frolova, Andre Kahles, Christelle Desnues, Christos Ouzounis, Cynthia Gibas, et al. 2022. A history of the MetaSUB consortium: Tracking urban microbes around the globe. *iScience* 25: 104993. https://doi.org/10.1016/j.isci.2022.104993.
36. Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2021. A genomic catalog of Earth’s microbiomes. *Nature Biotechnology* 39. Nature Publishing Group: 499–509. https://doi.org/10.1038/s41587-020-0718-6.
37. Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* 176: 649-662.e20. https://doi.org/10.1016/j.cell.2019.01.001.
38. 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.
39. 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.
40. Galtier, Nicolas, and Vincent Daubin. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363. Royal Society: 4023–4029. https://doi.org/10.1098/rstb.2008.0144.
41. Posada, David, and Keith A. Crandall. 2002. The Effect of Recombination on the Accuracy of Phylogeny Estimation. *Journal of Molecular Evolution* 54: 396–402. https://doi.org/10.1007/s00239-001-0034-9.
42. Uyeda, Josef C, Rosana Zenil-Ferguson, and Matthew W Pennell. 2018. Rethinking phylogenetic comparative methods. *Systematic Biology* 67: 1091–1109. https://doi.org/10.1093/sysbio/syy031.
43. Affairs, United Nations Department of Economic and Social. 2019. *World Urbanization Prospects: The 2018 Revision*. United Nations. https://doi.org/10.18356/b9e995fe-en.
44. Ritchie, Hannah, Veronika Samborska, and Max Roser. 2024. Urbanization. *Our World in Data*.
45. Neiderud, Carl-Johan. 2015. How urbanization affects the epidemiology of emerging infectious diseases. *Infection Ecology & Epidemiology* 5. Taylor & Francis: 27060. https://doi.org/10.3402/iee.v5.27060.
46. Tulchinsky, Theodore H. 2018. Chapter 5 - John Snow, Cholera, the Broad Street Pump; Waterborne Diseases Then and Now. In *Case Studies in Public Health*, ed. Theodore H. Tulchinsky, 77–99. Academic Press. https://doi.org/10.1016/B978-0-12-804571-8.00017-2.

47. Afshinnekoo, Ebrahim, Cem Meydan, Shanin Chowdhury, Dyala Jaroudi, Collin Boyer, Nick Bernstein, Julia M. Maritz, et al. 2015. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Systems* 1. Elsevier: 72–87. https://doi.org/10.1016/j.cels.2015.01.001.

48. Wylie, Kristine M., George M. Weinstock, and Gregory A. Storch. 2012. Emerging view of the human virome. *Translational Research* 160. Elsevier: 283–290. https://doi.org/10.1016/j.trsl.2012.03.006.

49. Gao, Zihao, Jun Wu, Alexander G. Lucaci, Jian Ouyang, Lan Wang, Krista Ryon, Eran Elhaik, et al. 2024. Diversity and Distinctive Traits of the Global RNA Virome in Urban Environments. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network. https://doi.org/10.2139/ssrn.4871972.

50. Crits-Christoph, Alexander, Rose S. Kantor, Matthew R. Olm, Oscar N. Whitney, Basem Al-Shayeb, Yue Clare Lou, Avi Flamholz, et al. 2021. Genome Sequencing of Sewage Detects Regionally Prevalent SARS-CoV-2 Variants. *mBio* 12. American Society for Microbiology: e02703-20. https://doi.org/10.1128/mBio.02703-20.

51. Miller, Ruth R., Vincent Montoya, Jennifer L. Gardy, David M. Patrick, and Patrick Tang. 2013. Metagenomics for pathogen detection in public health. *Genome Medicine* 5: 81. https://doi.org/10.1186/gm485.

52. Cooley, J. D., W. C. Wong, C. A. Jumper, and D. C. Straus. 1998. Correlation between the prevalence of certain fungi and sick building syndrome. *Occupational and Environmental Medicine* 55. BMJ Publishing Group Ltd: 579–584. https://doi.org/10.1136/oem.55.9.579.

53. Nicolaou, N., N. Siddique, and A. Custovic. 2005. Allergic disease in urban and rural populations: increasing prevalence with increasing urbanization. *Allergy* 60: 1357–1360. https://doi.org/10.1111/j.1398-9995.2005.00961.x.

54. Gilbert, Jack A., and Brent Stephens. 2018. Microbiology of the built environment. *Nature Reviews Microbiology* 16. Nature Publishing Group: 661–670. https://doi.org/10.1038/s41579-018-0065-5.

55. Afshinnekoo, Ebrahim, Chandrima Bhattacharya, Ana Burguete-García, Eduardo Castro-Nallar, Youping Deng, Christelle Desnues, Emmanuel Dias-Neto, et al. 2021. COVID-19 drug practices risk antimicrobial resistance evolution. *The Lancet Microbe* 2: e135–e136. https://doi.org/10.1016/S2666-5247(21)00039-2.

56. Fresia, Pablo, Verónica Antelo, Cecilia Salazar, Matías Giménez, Bruno D'Alessandro, Ebrahim Afshinnekoo, Christopher Mason, Gastón H. Gonnet, and Gregorio Iraola. 2019. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. *Microbiome* 7: 35. https://doi.org/10.1186/s40168-019-0648-z.

57. Van Goethem, Marc W., Rian Pierneef, Oliver K. I. Bezuidt, Yves Van De Peer, Don A. Cowan, and Thulani P. Makhalanyane. 2018. A reservoir of 'historical' antibiotic resistance genes in remote pristine Antarctic soils. *Microbiome* 6: 40. https://doi.org/10.1186/s40168-018-0424-5.

58. Gatica, Joao, and Eddie Cytryn. 2013. Impact of treated wastewater irrigation on antibiotic resistance in the soil microbiome. *Environmental Science and Pollution Research* 20: 3529–3538. https://doi.org/10.1007/s11356-013-1505-4.

59. Lax, Simon, Daniel P. Smith, Jarrad Hampton-Marcell, Sarah M. Owens, Kim M. Handley, Nicole M. Scott, Sean M. Gibbons, et al. 2014. Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science* 345. American Association for the Advancement of Science: 1048–1052. https://doi.org/10.1126/science.1254529.

60. Berendonk, Thomas U., Célia M. Manaia, Christophe Merlin, Despo Fatta-Kassinos, Eddie Cytryn, Fiona Walsh, Helmut Bürgmann, et al. 2015. Tackling antibiotic resistance: the environmental framework. *Nature Reviews Microbiology* 13. Nature Publishing Group: 310–317. https://doi.org/10.1038/nrmicro3439.

61. Barba, Marina, Henryk Czosnek, and Ahmed Hadidi. 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. *Viruses* 6. Multidisciplinary Digital Publishing Institute: 106–136. https://doi.org/10.3390/v6010106.
62. Gehrig, Jeanette L., Daniel M. Portik, Mark D. Driscoll, Eric Jackson, Shreyasee Chakraborty, Dawn Gratalo, Meredith Ashby, and Ricardo Valladares. 2022. Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microbial Genomics* 8. Microbiology Society,: 000794. https://doi.org/10.1099/mgen.0.000794.
63. Amarasinghe, Shanika L., Shian Su, Xueyi Dong, Luke Zappia, Matthew E. Ritchie, and Quentin Gouil. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* 21: 30. https://doi.org/10.1186/s13059-020-1935-5.
64. Kadam, Pratibha Prakash, Tejal Mestry, Nerges Mistry, and Kayzad Soli Nilgiriwala. 2025. Wastewater-based genomic surveillance of SARS-CoV-2 in vulnerable communities in Mumbai. *The Indian Journal of Medical Research* 160. Scientific Scholar: 570–577. https://doi.org/10.25259/ijmr_299_24.
65. Wastewater sequencing — an early warning system for infectious disease outbreaks. 2022. *Oxford Nanopore Technologies*. November 10.
66. Xia, Yu, An-Dong Li, Yu Deng, Xiao-Tao Jiang, Li-Guan Li, and Tong Zhang. 2017. MinION Nanopore Sequencing Enables Correlation between Resistome Phenotype and Genotype of Coliform Bacteria in Municipal Sewage. *Frontiers in Microbiology* 8. Frontiers. https://doi.org/10.3389/fmicb.2017.02105.
67. Eisenhofer, Raphael, Joseph Nesme, Luisa Santos-Bay, Adam Koziol, Søren Johannes Sørensen, Antton Alberdi, and Ostaizka Aizpurua. 2024. A comparison of short-read, HiFi long-read, and hybrid strategies for genome-resolved metagenomics. *Microbiology Spectrum* 12. American Society for Microbiology: e03590-23. https://doi.org/10.1128/spectrum.03590-23.
68. Sanderson, Nicholas D., Katie M.V. Hopkins, Matthew Colpus, Melody Parker, Samuel Lipworth, Derrick Crook, and Nicole Stoesser. 2024. Evaluation of the accuracy of bacterial genome reconstruction with Oxford Nanopore R10.4.1 long-read-only sequencing. *Microbial Genomics* 10. Microbiology Society,: 001246. https://doi.org/10.1099/mgen.0.001246.
69. Bogaerts, Bert, An Van den Bossche, Bavo Verhaegen, Laurence Delbrassinne, Wesley Mattheus, Stéphanie Nouws, Maxime Godfroid, et al. 2024. Closing the gap: Oxford Nanopore Technologies R10 sequencing allows comparable results to Illumina sequencing for SNP-based outbreak investigation of bacterial pathogens. *Journal of Clinical Microbiology* 62. American Society for Microbiology: e01576-23. https://doi.org/10.1128/jcm.01576-23.
70. Hoffmann, Maria, Jay Hee Jang, Sandra M. Tallent, and Narjol Gonzalez-Escalona. 2024. Single Laboratory Evaluation of the Q20+ Nanopore Sequencing Kit for Bacterial Outbreak Investigations. *International Journal of Molecular Sciences* 25. Multidisciplinary Digital Publishing Institute: 11877. https://doi.org/10.3390/ijms252211877.
71. Gounot, Jean-Sebastien, Minghao Chia, Denis Bertrand, Woei-Yuh Saw, Aarthi Ravikrishnan, Adrian Low, Yichen Ding, et al. 2022. Genome-centric analysis of short and long read metagenomes reveals uncharacterized microbiome diversity in Southeast Asians. *Nature Communications* 13. Nature Publishing Group: 6044. https://doi.org/10.1038/s41467-022-33782-z.
72. Roehr, Johannes T, Christoph Dieterich, and Knut Reinert. 2017. Flexbar 3.0 – SIMD and multicore parallelization. *Bioinformatics* 33: 2941–2942. https://doi.org/10.1093/bioinformatics/btx330.
73. Ultraplex: A rapid, flexible, all-in-one ... | Wellcome Open Research. 2025. https://wellcomeopenresearch.org/articles/6-141/v1. Accessed May 6.

74. Ewing, Brent, LaDeana Hillier, Michael C. Wendl, and Phil Green. 1998. Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment. *Genome Research* 8. Cold Spring Harbor Lab: 175–185. https://doi.org/10.1101/gr.8.3.175.
75. Ewing, Brent, and Phil Green. 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Research* 8. Cold Spring Harbor Lab: 186–194. https://doi.org/10.1101/gr.8.3.186.
76. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. 2023. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed October 9.
77. PRINSEQ. 2013. *SourceForge*. November 10.
78. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
79. BBMap download | SourceForge.net. 2023. https://sourceforge.net/projects/bbmap/. Accessed September 20.
80. Yang, Chao, Debajyoti Chowdhury, Zhenmiao Zhang, William K. Cheung, Aiping Lu, Zhaoxiang Bian, and Lu Zhang. 2021. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal* 19: 6301–6314. https://doi.org/10.1016/j.csbj.2021.11.028.
81. De Coster, Wouter, Svenn D'Hert, Darrin T Schultz, Marc Cruts, and Christine Van Broeckhoven. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 34: 2666–2669. https://doi.org/10.1093/bioinformatics/bty149.
82. Fukasawa, Yoshinori, Luca Ermini, Hai Wang, Karen Carty, and Min-Sin Cheung. 2020. LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 Genes|Genomes|Genetics* 10: 1193–1196. https://doi.org/10.1534/g3.119.400864.
83. Schmieder, Robert, and Robert Edwards. 2011. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLOS ONE* 6. Public Library of Science: e17288. https://doi.org/10.1371/journal.pone.0017288.
84. biobakery/kneaddata. 2025. Python. bioBakery.
85. Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31: 1674–1676. https://doi.org/10.1093/bioinformatics/btv033.
86. Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* 27. Cold Spring Harbor Lab: 824–834. https://doi.org/10.1101/gr.213959.116.
87. Bruijn, de, N.G. 1946. A combinatorial problem. *Proceedings of the Section of Sciences of the Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam* 49: 758–764.
88. Kolmogorov, Mikhail, Derek M. Bickhart, Bahar Behsaz, Alexey Gurevich, Mikhail Rayko, Sung Bong Shin, Kristen Kuhn, et al. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods* 17. Nature Publishing Group: 1103–1110. https://doi.org/10.1038/s41592-020-00971-x.
89. Feng, Xiaowen, Haoyu Cheng, Daniel Portik, and Heng Li. 2022. Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods* 19. Nature Publishing Group: 671–674. https://doi.org/10.1038/s41592-022-01478-3.
90. Benoit, Gaëtan, Sébastien Raguideau, Robert James, Adam M. Phillippy, Rayan Chikhi, and Christopher Quince. 2024. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nature Biotechnology* 42. Nature Publishing Group: 1378–1383. https://doi.org/10.1038/s41587-023-01983-6.

91. Antipov, Dmitry, Anton Korobeynikov, Jeffrey S. McLean, and Pavel A. Pevzner. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32: 1009–1015. https://doi.org/10.1093/bioinformatics/btv688.

92. Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nature Biotechnology* 37. Nature Publishing Group: 937–944. https://doi.org/10.1038/s41587-019-0191-2.

93. Wick, Ryan R., Louise M. Judd, Claire L. Gorrie, and Kathryn E. Holt. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology* 13. Public Library of Science: e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

94. Liu, Lei, Yulin Wang, Yu Yang, Depeng Wang, Suk Hang Cheng, Chunmiao Zheng, and Tong Zhang. 2021. Charting the complexity of the activated sludge microbiome through a hybrid sequencing strategy. *Microbiome* 9: 205. https://doi.org/10.1186/s40168-021-01155-1.

95. Cilibrasi, Rudi, Leo van Iersel, Steven Kelk, and John Tromp. 2005. On the Complexity of Several Haplotyping Problems. In *Algorithms in Bioinformatics*, ed. Rita Casadio and Gene Myers, 128–139. Berlin, Heidelberg: Springer. https://doi.org/10.1007/11557067_11.

96. Nicholls, Samuel M, Wayne Aubrey, Kurt De Grave, Leander Schietgat, Christopher J Creevey, and Amanda Clare. 2021. On the complexity of haplotyping a microbial community. *Bioinformatics* 37: 1360–1366. https://doi.org/10.1093/bioinformatics/btaa977.

97. Portik, Daniel M., C. Titus Brown, and N. Tessa Pierce-Ward. 2022. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics* 23: 541. https://doi.org/10.1186/s12859-022-05103-0.

98. Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3. PeerJ Inc.: e1165. https://doi.org/10.7717/peerj.1165.

99. Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods* 11. Nature Publishing Group: 1144–1146. https://doi.org/10.1038/nmeth.3103.

100. Nissen, Jakob Nybo, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, et al. 2021. Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* 39. Nature Publishing Group: 555–560. https://doi.org/10.1038/s41587-020-00777-4.

101. Muralidharan, Harihara Subrahmaniam, Nidhi Shah, Jacquelyn S. Meisel, and Mihai Pop. 2021. Binnacle: Using Scaffolds to Improve the Contiguity and Quality of Metagenomic Bins. *Frontiers in Microbiology* 12. Frontiers. https://doi.org/10.3389/fmicb.2021.638561.

102. Wu, Yu-Wei, Yung-Hsu Tang, Susannah G. Tringe, Blake A. Simmons, and Steven W. Singer. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2: 26. https://doi.org/10.1186/2049-2618-2-26.

103. Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Research* 17. Cold Spring Harbor Lab: 377–386. https://doi.org/10.1101/gr.5969107.

104. Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

105. Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* 12. Nature Publishing Group: 59–60. https://doi.org/10.1038/nmeth.3176.

106. Huson, Daniel H., Benjamin Albrecht, Caner Bağcı, Irina Bessarab, Anna Górska, Dino Jolic, and Rohan B. H. Williams. 2018. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct* 13: 6. https://doi.org/10.1186/s13062-018-0208-7.

107. Liu, Bo, Theodore Gibbons, Mohammad Ghodsi, Todd Treangen, and Mihai Pop. 2011. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* 12: S4. https://doi.org/10.1186/1471-2164-12-S2-S4.

108. Darling, Aaron E., Guillaume Jospin, Eric Lowe, Frederick A. Matsen Iv, Holly M. Bik, and Jonathan A. Eisen. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2. PeerJ Inc.: e243. https://doi.org/10.7717/peerj.243.

109. Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*. Nature Publishing Group: 1–12. https://doi.org/10.1038/s41587-023-01688-w.

110. Martínez-Porchas, Marcel, Enrique Villalpando-Canchola, and Francisco Vargas-Albores. 2016. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. *Heliyon* 2. Elsevier. https://doi.org/10.1016/j.heliyon.2016.e00170.

111. Alser, Mohammed, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, et al. 2021. Technology dictates algorithms: recent developments in read alignment. *Genome Biology* 22: 249. https://doi.org/10.1186/s13059-021-02443-7.

112. Wood, Derrick E., and Steven L. Salzberg. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15: R46. https://doi.org/10.1186/gb-2014-15-3-r46.

113. Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science* 3. PeerJ Inc.: e104. https://doi.org/10.7717/peerj-cs.104.

114. Brown, C. Titus, and Luiz Irber. 2016. sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software* 1: 27. https://doi.org/10.21105/joss.00027.

115. Ondov, Brian D., Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. 2019. Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biology* 20: 232. https://doi.org/10.1186/s13059-019-1841-x.

116. Koslicki, David, Stephen White, Chunyu Ma, and Alexei Novikov. 2024. YACHT: an ANI-based statistical test to detect microbial presence/absence in a metagenomic sample. *Bioinformatics* 40: btae047. https://doi.org/10.1093/bioinformatics/btae047.

117. LaPierre, Nathan, Mohammed Alser, Eleazar Eskin, David Koslicki, and Serghei Mangul. 2020. Metalign: efficient alignment-based metagenomic profiling via containment min hash. *Genome Biology* 21: 242. https://doi.org/10.1186/s13059-020-02159-0.

118. Liu, Shaopeng, and David Koslicki. 2022. CMash: fast, multi-resolution estimation of k-mer-based Jaccard and containment indices. *Bioinformatics* 38: i28–i35. https://doi.org/10.1093/bioinformatics/btac237.

119. Fan, Jeremy, Steven Huang, and Samuel D. Chorlton. 2021. BugSeq: a highly accurate cloud platform for long-read metagenomic analyses. *BMC Bioinformatics* 22: 160. https://doi.org/10.1186/s12859-021-04089-5.

120. Tedersoo, Leho, Mads Albertsen, Sten Anslan, and Benjamin Callahan. 2021. Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. *Applied and Environmental Microbiology* 87. American Society for Microbiology: e00626-21. https://doi.org/10.1128/AEM.00626-21.
121. Chen, Liang, Na Zhao, Jiabao Cao, Xiaolin Liu, Jiayue Xu, Yue Ma, Ying Yu, et al. 2022. Short- and long-read metagenomics expand individualized structural variations in gut microbiomes. *Nature Communications* 13. Nature Publishing Group: 3175. https://doi.org/10.1038/s41467-022-30857-9.
122. Riesenfeld, Christian S., Patrick D. Schloss, and Jo Handelsman. 2004. Metagenomics: Genomic Analysis of Microbial Communities. *Annual Review of Genetics* 38. Annual Reviews: 525–552. https://doi.org/10.1146/annurev.genet.38.072902.091216.
123. Sass, Peter, ed. 2023. *Antibiotics: Methods and Protocols*. Vol. 2601. Methods in Molecular Biology. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2855-3.
124. Proctor, Lita M., Heather H. Creasy, Jennifer M. Fettweis, Jason Lloyd-Price, Anup Mahurkar, Wenyu Zhou, Gregory A. Buck, et al. 2019. The Integrative Human Microbiome Project. *Nature* 569. Nature Publishing Group: 641–648. https://doi.org/10.1038/s41586-019-1238-8.
125. Dabdoub, Shareef M., Sukirth M. Ganesan, and Purnima S. Kumar. 2016. Comparative metagenomics reveals taxonomically idiosyncratic yet functionally congruent communities in periodontitis. *Scientific Reports* 6. Nature Publishing Group: 38993. https://doi.org/10.1038/srep38993.
126. Healy, F. G., R. M. Ray, H. C. Aldrich, A. C. Wilkie, L. O. Ingram, and K. T. Shanmugam. 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. *Applied Microbiology and Biotechnology* 43: 667–674. https://doi.org/10.1007/BF00164771.
127. Williamson, Lynn L., Bradley R. Borlee, Patrick D. Schloss, Changhui Guan, Heather K. Allen, and Jo Handelsman. 2005. Intracellular Screen To Identify Metagenomic Clones That Induce or Inhibit a Quorum-Sensing Biosensor. *Applied and Environmental Microbiology* 71. American Society for Microbiology: 6335–6344. https://doi.org/10.1128/AEM.71.10.6335-6344.2005.
128. Rondon, Michelle R., Paul R. August, Alan D. Bettermann, Sean F. Brady, Trudy H. Grossman, Mark R. Liles, Kara A. Loiacono, et al. 2000. Cloning the Soil Metagenome: a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms. *Applied and Environmental Microbiology* 66. American Society for Microbiology: 2541–2547. https://doi.org/10.1128/AEM.66.6.2541-2547.2000.
129. Riesenfeld, Christian S., Robert M. Goodman, and Jo Handelsman. 2004. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environmental Microbiology* 6: 981–989. https://doi.org/10.1111/j.1462-2920.2004.00664.x.
130. Noguchi, Hideki, Takeaki Taniguchi, and Takehiko Itoh. 2008. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Research* 15: 387–396. https://doi.org/10.1093/dnares/dsn027.
131. Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11: 119. https://doi.org/10.1186/1471-2105-11-119.
132. Al-Ajlan, Amani, and Achraf El Allali. 2019. CNN-MGP: Convolutional Neural Networks for Metagenomics Gene Prediction. *Interdisciplinary Sciences: Computational Life Sciences* 11: 628–635. https://doi.org/10.1007/s12539-018-0313-4.

133. Zhang, Shao-Wu, Xiang-Yang Jin, and Teng Zhang. 2017. Gene Prediction in Metagenomic Fragments with Deep Learning. *BioMed Research International* 2017: 4740354. https://doi.org/10.1155/2017/4740354.

134. Cantalapiedra, Carlos P, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* 38: 5825–5829. https://doi.org/10.1093/molbev/msab293.

135. Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. 2016. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology* 428. Computation Resources for Molecular Biology: 726–731. https://doi.org/10.1016/j.jmb.2015.11.006.

136. Galperin, Michael Y, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* 49: D274–D281. https://doi.org/10.1093/nar/gkaa1018.

137. Kanehisa, Minoru, Miho Furumichi, Yoko Sato, Yuriko Matsuura, and Mari Ishiguro-Watanabe. 2025. KEGG: biological systems database as a model of the real world. *Nucleic Acids Research* 53: D672–D677. https://doi.org/10.1093/nar/gkae909.

138. Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research* 49: D412–D419. https://doi.org/10.1093/nar/gkaa913.

139. Anand, Swadha, Bhusan K Kuntal, Anwesha Mohapatra, Vineet Bhatt, and Sharmila S Mande. 2020. FunGeCo: a web-based tool for estimation of functional potential of bacterial genomes and microbiomes using gene context information. *Bioinformatics* 36: 2575–2577. https://doi.org/10.1093/bioinformatics/btz957.

140. Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30: 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

141. Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36: 2251–2252. https://doi.org/10.1093/bioinformatics/btz859.

142. Maranga, Mary, Pawel Szczerbiak, Valentyn Bezshapkin, Vladimir Gligorijevic, Chris Chandler, Richard Bonneau, Ramnik J. Xavier, Tommi Vatanen, and Tomasz Kosciolek. 2023. Comprehensive Functional Annotation of Metagenomes and Microbial Genomes Using a Deep Learning-Based Method. *mSystems* 8. American Society for Microbiology: e01178-22. https://doi.org/10.1128/msystems.01178-22.

143. Meyer, F., D. Paarmann, M. D'Souza, R. Olson, EM Glass, M. Kubal, T. Paczian, et al. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386. https://doi.org/10.1186/1471-2105-9-386.

144. Vanni, Chiara, Matthew S Schechter, Silvia G Acinas, Albert Barberán, Pier Luigi Buttigieg, Emilio O Casamayor, Tom O Delmont, et al. 2022. Unifying the known and unknown microbial coding sequence space. Edited by C Titus Brown, Gisela Storz, C Titus Brown, and Byron Smith. *eLife* 11. eLife Sciences Publications, Ltd: e67667. https://doi.org/10.7554/eLife.67667.

145. Elnaggar, Ahmed, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, et al. 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44: 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381.

146. Bepler, Tristan, and Bonnie Berger. 2021. Learning the protein language: Evolution, structure, and function. *Cell Systems* 12: 654-669.e3. https://doi.org/10.1016/j.cels.2021.05.017.
147. Keegan, Kevin P., Elizabeth M. Glass, and Folker Meyer. 2016. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In *Microbial Environmental Genomics (MEG)*, ed. Francis Martin and Stephane Uroz, 207–233. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-3369-3_13.
148. Wilke, Andreas, Travis Harrison, Jared Wilkening, Dawn Field, Elizabeth M. Glass, Nikos Kyrpides, Konstantinos Mavrommatis, and Folker Meyer. 2012. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 13: 141. https://doi.org/10.1186/1471-2105-13-141.
149. Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, et al. 2012. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. *PLOS Computational Biology* 8. Public Library of Science: e1002358. https://doi.org/10.1371/journal.pcbi.1002358.
150. Pascal Andreu, Victòria, Hannah E. Augustijn, Lianmin Chen, Alexandra Zhernakova, Jingyuan Fu, Michael A. Fischbach, Dylan Dodd, and Marnix H. Medema. 2023. gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota. *Nature Biotechnology* 41. Nature Publishing Group: 1416–1423. https://doi.org/10.1038/s41587-023-01675-1.
151. Caspi, Ron, Richard Billington, Luciana Ferrer, Hartmut Foerster, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, et al. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research* 44: D471–D480. https://doi.org/10.1093/nar/gkv1164.
152. Arkin, Adam P., Robert W. Cottingham, Christopher S. Henry, Nomi L. Harris, Rick L. Stevens, Sergei Maslov, Paramvir Dehal, et al. 2018. KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology* 36. Nature Publishing Group: 566–569. https://doi.org/10.1038/nbt.4163.
153. Wishart, David S., Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. 2006. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research* 34: D668–D672. https://doi.org/10.1093/nar/gkj067.
154. Dong, Xiaoli, and Marc Strous. 2019. An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs. *Frontiers in Genetics* 10. Frontiers. https://doi.org/10.3389/fgene.2019.00999.
155. Dimonaco, Nicholas J, Wayne Aubrey, Kim Kenobi, Amanda Clare, and Christopher J Creevey. 2022. No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study. *Bioinformatics* 38: 1198–1207. https://doi.org/10.1093/bioinformatics/btab827.
156. Bakken, Lars R. 1985. Separation and Purification of Bacteria from Soil. *Applied and Environmental Microbiology* 49: 1482–1487.
157. Amann, R I, W Ludwig, and K H Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews* 59. American Society for Microbiology: 143–169. https://doi.org/10.1128/mr.59.1.143-169.1995.
158. Eckburg, Paul B., Elisabeth M. Bik, Charles N. Bernstein, Elizabeth Purdom, Les Dethlefsen, Michael Sargent, Steven R. Gill, Karen E. Nelson, and David A. Relman. 2005. Diversity of the Human Intestinal Microbial Flora. *Science* 308. American Association for the Advancement of Science: 1635–1638. https://doi.org/10.1126/science.1110591.

159. Handelsman, Jo. 2004. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* 68. American Society for Microbiology: 669–685. https://doi.org/10.1128/mmbr.68.4.669-685.2004.

160. Iverson, Vaughn, Robert M. Morris, Christian D. Frazar, Chris T. Berthiaume, Rhonda L. Morales, and E. Virginia Armbrust. 2012. Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. *Science* 335. American Association for the Advancement of Science: 587–590. https://doi.org/10.1126/science.1212665.

161. Mackelprang, Rachel, Mark P. Waldrop, Kristen M. DeAngelis, Maude M. David, Krystle L. Chavarria, Steven J. Blazewicz, Edward M. Rubin, and Janet K. Jansson. 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 480. Nature Publishing Group: 368–371. https://doi.org/10.1038/nature10576.

162. Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420–1428. https://doi.org/10.1093/bioinformatics/bts174.

163. Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* 40: e155. https://doi.org/10.1093/nar/gks678.

164. Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 18. https://doi.org/10.1186/2047-217X-1-18.

165. Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J. M. Jones, and İnanç Birol. 2009. ABySS: A parallel assembler for short read sequence data. *Genome Research* 19. Cold Spring Harbor Lab: 1117–1123. https://doi.org/10.1101/gr.089532.108.

166. Pell, Jason, Arend Hintze, Rosangela Canino-Koning, Adina Howe, James M. Tiedje, and C. Titus Brown. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences* 109. Proceedings of the National Academy of Sciences: 13272–13277. https://doi.org/10.1073/pnas.1121464109.

167. Howe, Adina Chuang, Jason Pell, Rosangela Canino-Koning, Rachel Mackelprang, Susannah Tringe, Janet Jansson, James M. Tiedje, and C. Titus Brown. 2012. Illumina Sequencing Artifacts Revealed by Connectivity Analysis of Metagenomic Datasets. arXiv. https://doi.org/10.48550/arXiv.1212.0159.

168. Boisvert, Sébastien, Frédéric Raymond, Élénie Godzaridis, François Laviolette, and Jacques Corbeil. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology* 13: R122. https://doi.org/10.1186/gb-2012-13-12-r122.

169. Sangwan, Naseer, Fangfang Xia, and Jack A. Gilbert. 2016. Recovering complete and draft population genomes from metagenome datasets. *Microbiome* 4: 8. https://doi.org/10.1186/s40168-016-0154-5.

170. Breitwieser, Florian P, Jennifer Lu, and Steven L Salzberg. 2019. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics* 20: 1125–1136. https://doi.org/10.1093/bib/bbx120.

171. Zaheer, Rahat, Noelle Noyes, Rodrigo Ortega Polo, Shaun R. Cook, Eric Marinier, Gary Van Domselaar, Keith E. Belk, Paul S. Morley, and Tim A. McAllister. 2018. Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports* 8. Nature Publishing Group: 5890. https://doi.org/10.1038/s41598-018-24280-8.

172. Pereira-Marques, Joana, Anne Hout, Rui M. Ferreira, Michiel Weber, Ines Pinto-Ribeiro, Leen-Jan van Doorn, Cornelis Willem Knetsch, and Ceu Figueiredo. 2019. Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in Microbiology* 10. Frontiers. https://doi.org/10.3389/fmicb.2019.01277.

173. Bowers, Robert M., Alicia Clum, Hope Tice, Joanne Lim, Kanwar Singh, Doina Ciobanu, Chew Yee Ngan, Jan-Fang Cheng, Susannah G. Tringe, and Tanja Woyke. 2015. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 16: 856. https://doi.org/10.1186/s12864-015-2063-6.

174. Shaiber, Alon, and A. Murat Eren. 2019. Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. *mBio* 10. American Society for Microbiology: 10.1128/mbio.00725-19. https://doi.org/10.1128/mbio.00725-19.

175. Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32: 1088–1090. https://doi.org/10.1093/bioinformatics/btv697.

176. 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.

177. Stewart, Robert D, Marc D Auffret, Timothy J Snelling, Rainer Roehe, and Mick Watson. 2019. MAGpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (MAGs). *Bioinformatics* 35: 2150–2152. https://doi.org/10.1093/bioinformatics/bty905.

178. Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3. PeerJ Inc.: e1319. https://doi.org/10.7717/peerj.1319.

179. Meyer, Fernando, Peter Hofmann, Peter Belmann, Ruben Garrido-Oter, Adrian Fritz, Alexander Sczyrba, and Alice C McHardy. 2018. AMBER: Assessment of Metagenome BinnERs. *GigaScience* 7: giy069. https://doi.org/10.1093/gigascience/giy069.

180. Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* 3. Nature Publishing Group: 836–843. https://doi.org/10.1038/s41564-018-0171-1.

181. Zhou, Zhemin, Nabil-Fareed Alikhan, Khaled Mohamed, Yulei Fan, the Agama Study Group, Mark Achtman, Derek Brown, et al. 2020. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. *Genome Research* 30. Cold Spring Harbor Lab: 138–152. https://doi.org/10.1101/gr.251678.119.

182. Napit, Rajindra, Anupama Gurung, Ajit Poudel, Ashok Chaudhary, Prajwol Manandhar, Ajay Narayan Sharma, Samita Raut, et al. 2025. Metagenomic analysis of human, animal, and environmental samples identifies potential emerging pathogens, profiles antibiotic resistance genes, and reveals horizontal gene transfer dynamics. *Scientific Reports* 15. Nature Publishing Group: 12156. https://doi.org/10.1038/s41598-025-90777-8.

183. Fitzsimons, Michael S., Mark Novotny, Chien-Chi Lo, Armand E. K. Dichosa, Joyclyn L. Yee-Greenbaum, Jeremy P. Snook, Wei Gu, et al. 2013. Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome Research* 23. Cold Spring Harbor Lab: 878–888. https://doi.org/10.1101/gr.142208.112.

184. Oh, Julia, Allyson L. Byrd, Clay Deming, Sean Conlan, Heidi H. Kong, and Julia A. Segre. 2014. Biogeography and individuality shape function in the human skin metagenome. *Nature* 514. Nature Publishing Group: 59–64. https://doi.org/10.1038/nature13786.

185. Greenblum, Sharon, Rogan Carr, and Elhanan Borenstein. 2015. Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. *Cell* 160. Elsevier: 583–594. https://doi.org/10.1016/j.cell.2014.12.038.

186. Huson, Daniel H., Suparna Mitra, Hans-Joachim Ruscheweyh, Nico Weber, and Stephan C. Schuster. 2011. Integrative analysis of environmental sequences using MEGAN4. *Genome Research* 21. Cold Spring Harbor Lab: 1552–1560. https://doi.org/10.1101/gr.120618.111.

187. Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research* 27. Cold Spring Harbor Lab: 626–638. https://doi.org/10.1101/gr.216242.116.

188. Luo, Chengwei, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J. Xavier, and Dirk Gevers. 2015. ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology* 33. Nature Publishing Group: 1045–1052. https://doi.org/10.1038/nbt.3319.

189. Costea, Paul Igor, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork. 2017. metaSNV: A tool for metagenomic strain level analysis. *PLOS ONE* 12. Public Library of Science: e0182392. https://doi.org/10.1371/journal.pone.0182392.

190. Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biology* 18: 181. https://doi.org/10.1186/s13059-017-1309-9.

191. Frank, J. A., Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope. 2016. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports* 6. Nature Publishing Group: 25373. https://doi.org/10.1038/srep25373.

192. Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* 58. 3D Chromatin Architecture: 268–276. https://doi.org/10.1016/j.ymeth.2012.05.001.

193. 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.

194. Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31: 3691–3693. https://doi.org/10.1093/bioinformatics/btv421.

195. Li, Weizhong, and Adam Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.

196. Hernández-Plaza, Ana, Damian Szklarczyk, Jorge Botas, Carlos P Cantalapiedra, Joaquín Giner-Lamia, Daniel R Mende, Rebecca Kirsch, et al. 2023. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Research* 51: D389–D394. https://doi.org/10.1093/nar/gkac1022.

197. Puigbò, Pere, Alexander E. Lobkovsky, David M. Kristensen, Yuri I. Wolf, and Eugene V. Koonin. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology* 12: 66. https://doi.org/10.1186/s12915-014-0066-4.

198. Kislyuk, Andrey O., Bart Haegeman, Nicholas H. Bergman, and Joshua S. Weitz. 2011. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* 12: 32. https://doi.org/10.1186/1471-2164-12-32.

199. 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.

200. Maistrenko, Oleksandr M, Daniel R Mende, Mechthild Luetge, Falk Hildebrand, Thomas S B Schmidt, Simone S Li, João F Matias Rodrigues, et al. 2020. Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME Journal* 14: 1247–1259. https://doi.org/10.1038/s41396-020-0600-z.

201. Brockhurst, Michael A., Ellie Harrison, James P. J. Hall, Thomas Richards, Alan McNally, and Craig MacLean. 2019. The Ecology and Evolution of Pangenomes. *Current Biology* 29: R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012.

202. Holder, Mark, and Paul O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* 4. Nature Publishing Group: 275–284. https://doi.org/10.1038/nrg1044.

203. Cohen, Ofir, and Tal Pupko. 2010. Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution* 27: 703–713. https://doi.org/10.1093/molbev/msp240.

204. Beaulieu, Jeremy M., Brian C. O'Meara, and Michael J. Donoghue. 2013. Identifying Hidden Rate Changes in the Evolution of a Binary Morphological Character: The Evolution of Plant Habit in Campanulid Angiosperms. *Systematic Biology* 62: 725–737. https://doi.org/10.1093/sysbio/syt034.

205. Schluter, Dolph, Trevor Price, Arne Ø. Mooers, and Donald Ludwig. 1997. Likelihood of Ancestor States in Adaptive Radiation. *Evolution* 51: 1699–1711. https://doi.org/10.1111/j.1558-5646.1997.tb05095.x.

206. Cunningham, Clifford W., Kevin E. Omland, and Todd H. Oakley. 1998. Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology & Evolution* 13: 361–366. https://doi.org/10.1016/S0169-5347(98)01382-2.

207. Omland, Kevin E. 1999. The Assumptions and Challenges of Ancestral State Reconstructions. *Systematic Biology* 48. [Oxford University Press, Society of Systematic Biologists]: 604–611.

208. Crisp, Michael D., and Lyn G. Cook. 2005. Do early branching lineages signify ancestral traits? *Trends in Ecology & Evolution* 20: 122–128. https://doi.org/10.1016/j.tree.2004.11.010.

209. Pagel, Mark. 1999. Inferring the historical patterns of biological evolution. *Nature* 401. Nature Publishing Group: 877–884. https://doi.org/10.1038/44766.

210. Nepokroeff, Molly, Kenneth J. Sytsma, Warren L. Wagner, and Elizabeth A. Zimmer. 2003. Reconstructing Ancestral Patterns of Colonization and Dispersal in the Hawaiian Understory Tree Genus Psychotria (Rubiaceae):A Comparison of Parsimony and Likelihood Approaches. *Systematic Biology* 52: 820–838. https://doi.org/10.1093/sysbio/52.6.820.

211. Webster, Andrea J., and Andy Purvis. 2002. Testing the accuracy of methods for reconstructing ancestral states of continuous characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 269. Royal Society: 143–149. https://doi.org/10.1098/rspb.2001.1873.

212. Oakley, Todd H., and Clifford W. Cunningham. 2000. Independent Contrasts Succeed Where Ancestor Reconstruction Fails in a Known Bacteriophage Phylogeny. *Evolution* 54. Oxford University Press: 397–405.

213. Manos, Paul S., and Alice M. Stanford. 2001. The Historical Biogeography of Fagaceae: Tracking the Tertiary History of Temperate and Subtropical Forests of the Northern Hemisphere. *International Journal of Plant Sciences* 162. The University of Chicago Press: S77–S93. https://doi.org/10.1086/323280.

214. Joy, Jeffrey B., Richard H. Liang, Rosemary M. McCloskey, T. Nguyen, and Art F. Y. Poon. 2016. Ancestral Reconstruction. *PLOS Computational Biology* 12. Public Library of Science: e1004763. https://doi.org/10.1371/journal.pcbi.1004763.
215. Gàlvez-Morante, Alex, Laurent Guéguen, Paschalis Natsidis, Maximilian J Telford, and Daniel J Richter. 2024. Dollo Parsimony Overestimates Ancestral Gene Content Reconstructions. *Genome Biology and Evolution* 16: evae062. https://doi.org/10.1093/gbe/evae062.
216. Friedman, William E., and Sandra K. Floyd. 2001. PERSPECTIVE: THE ORIGIN OF FLOWERING PLANTS AND THEIR REPRODUCTIVE BIOLOGY–A TALE OF TWO PHYLOGENIES. *Evolution* 55: 217–231. https://doi.org/10.1111/j.0014-3820.2001.tb01288.x.
217. Losos, Jonathan B. 1999. Uncertainty in the reconstruction of ancestral character states and limitations on the use of phylogenetic comparative methods. *Animal Behaviour* 58: 1319–1324. https://doi.org/10.1006/anbe.1999.1261.
218. Rodríguez, F., J. L. Oliver, A. Marín, and J. R. Medina. 1990. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology* 142: 485–501. https://doi.org/10.1016/S0022-5193(05)80104-3.
219. 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.
220. Swofford, David L., and Wayne P. Maddison. 1987. Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences* 87: 199–229. https://doi.org/10.1016/0025-5564(87)90074-5.
221. 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.
222. Kluge, Arnold G., and James S. Farris. 1969. Quantitative Phyletics and the Evolution of Anurans. *Systematic Zoology* 18. [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.]: 1–32. https://doi.org/10.2307/2412407.
223. Farris, James S. 1977. Phylogenetic Analysis Under Dollo's Law. *Systematic Zoology* 26. [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.]: 77–88. https://doi.org/10.2307/2412867.
224. Le Quesne, Walter J. 1974. The Uniquely Evolved Character Concept and its Cladistic Application. *Systematic Zoology* 23. [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.]: 513–517. https://doi.org/10.2307/2412469.
225. Rogozin, Igor B., Yuri I. Wolf, Vladimir N. Babenko, and Eugene V. Koonin. 2006. Dollo parsimony and the reconstruction of genome evolution. In *Parsimony, Phylogeny, and Genomics*, ed. Victor A. Albert, 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199297306.003.0011.
226. Felsenstein, Joseph. 1973. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology* 22. [Oxford University Press, Society of Systematic Biologists, Taylor & Francis, Ltd.]: 240–249. https://doi.org/10.2307/2412304.
227. Stamatakis, Alexandros. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690. https://doi.org/10.1093/bioinformatics/btl446.
228. Li, Guoliang, Mike Steel, and Louxin Zhang. 2008. More Taxa Are Not Necessarily Better for the Reconstruction of Ancestral Character States. *Systematic Biology* 57: 647–653. https://doi.org/10.1080/10635150802203898.

229. Yang, Z, S Kumar, and M Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141: 1641–1650. https://doi.org/10.1093/genetics/141.4.1641.

230. Koshi, Jeffrey M., and Richard A. Goldstein. 1996. Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution* 42: 313–320. https://doi.org/10.1007/BF02198858.

231. Pupko, Tal, Itsik Pe, Ron Shamir, and Dan Graur. 2000. A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution* 17: 890–896. https://doi.org/10.1093/oxfordjournals.molbev.a026369.

232. Eyre-Walker, Adam. 1998. Problems with Parsimony in Sequences of Biased Base Composition. *Journal of Molecular Evolution* 47: 686–690. https://doi.org/10.1007/PL00006427.

233. Pupko, Tal, Itsik Pe'er, Masami Hasegawa, Dan Graur, and Nir Friedman. 2002. A branch-and-bound algorithm for the inference of ancestral amino-acid sequences when the replacement rate varies among sites: Application to the evolution of five gene families. *Bioinformatics* 18: 1116–1123. https://doi.org/10.1093/bioinformatics/18.8.1116.

234. Gruenheit, Nicole, Peter J. Lockhart, Mike Steel, and William Martin. 2008. Difficulties in Testing for Covarion-Like Properties of Sequences under the Confounding Influence of Changing Proportions of Variable Sites. *Molecular Biology and Evolution* 25: 1512–1520. https://doi.org/10.1093/molbev/msn098.

235. Pagel, Mark. 1997. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255. Royal Society: 37–45. https://doi.org/10.1098/rspb.1994.0006.

236. FitzJohn, Richard G. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods in Ecology and Evolution* 3: 1084–1092. https://doi.org/10.1111/j.2041-210X.2012.00234.x.

237. Carmel, Liran, Yuri I. Wolf, Igor B. Rogozin, and Eugene V. Koonin. 2010. EREM: Parameter Estimation and Ancestral Reconstruction by Expectation-Maximization Algorithm for a Probabilistic Model of Genomic Binary Characters Evolution. *Advances in Bioinformatics* 2010: 167408. https://doi.org/10.1155/2010/167408.

238. Pond, Sergei L. Kosakovsky, Simon D. W. Frost, and Spencer V. Muse. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676–679. https://doi.org/10.1093/bioinformatics/bti079.

239. Paradis, Emmanuel, and Klaus Schliep. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526–528. https://doi.org/10.1093/bioinformatics/bty633.

240. Matzke, Nicholas J. 2018. nmatzke/BioGeoBEARS: BioGeoBEARS: BioGeography with Bayesian (and likelihood) Evolutionary Analysis with R Scripts. Zenodo. https://doi.org/10.5281/zenodo.1478250.

241. Ree, Richard H, and Stephen A Smith. 2008. Maximum Likelihood Inference of Geographic Range Evolution by Dispersal, Local Extinction, and Cladogenesis. *Systematic Biology* 57: 4–14. https://doi.org/10.1080/10635150701883881.

242. Meade, Andrew, and Mark Pagel. 2022. Ancestral State Reconstruction Using BayesTraits. In *Environmental Microbial Evolution: Methods and Protocols*, ed. Haiwei Luo, 255–266. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-2691-7_12.

243. Landis, Michael J., Nicholas J. Matzke, Brian R. Moore, and John P. Huelsenbeck. 2013. Bayesian Analysis of Biogeography when the Number of Areas is Large. *Systematic Biology* 62: 789–804. https://doi.org/10.1093/sysbio/syt040.

244. Yu, Yan, A. J. Harris, Christopher Blair, and Xingjin He. 2015. RASP (Reconstruct Ancestral State in Phylogenies): A tool for historical biogeography. *Molecular Phylogenetics and Evolution* 87: 46–49. https://doi.org/10.1016/j.ympev.2015.03.008.
245. Jones, Bradley R., Ashok Rajaraman, Eric Tannier, and Cedric Chauve. 2012. ANGES: reconstructing ANcestral GEnomeS maps. *Bioinformatics* 28: 2388–2390. https://doi.org/10.1093/bioinformatics/bts457.
246. Ronquist, Fredrik. 1997. Dispersal-Vicariance Analysis: A New Approach to the Quantification of Historical Biogeography. *Systematic Biology* 46: 195–203. https://doi.org/10.1093/sysbio/46.1.195.
247. Yang, Ziheng. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24: 1586–1591. https://doi.org/10.1093/molbev/msm088.
248. Huelsenbeck, John P., and Jonathan P. Bollback. 2001. Empirical and Hierarchical Bayesian Estimation of Ancestral States. *Systematic Biology* 50. [Oxford University Press, Society of Systematic Biologists]: 351–366.
249. Lutzoni, François, Mark Pagel, and Valérie Reeb. 2001. Major fungal lineages are derived from lichen symbiotic ancestors. *Nature* 411. Nature Publishing Group: 937–940. https://doi.org/10.1038/35082053.
250. Hanson-Smith, Victor, Bryan Kolaczkowski, and Joseph W. Thornton. 2010. Robustness of Ancestral Sequence Reconstruction to Phylogenetic Uncertainty. *Molecular Biology and Evolution* 27: 1988–1999. https://doi.org/10.1093/molbev/msq081.
251. Maddison, Wayne P., Peter E. Midford, and Sarah P. Otto. 2007. Estimating a Binary Character's Effect on Speciation and Extinction. *Systematic Biology* 56: 701–710. https://doi.org/10.1080/10635150701607033.
252. Afshinnekoo, Ebrahim, Cem Meydan, Shanin Chowdhury, Dyala Jaroudi, Collin Boyer, Nick Bernstein, Julia M. Maritz, et al. 2015. Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Systems* 1. Elsevier: 72–87. https://doi.org/10.1016/j.cels.2015.01.001.
253. Magnúsdóttir, Stefanía, Joao Pedro Saraiva, Alexander Bartholomäus, Majid Soheili, Rodolfo Brizola Toscan, Junya Zhang, Ulisses Nunes da Rocha, and CLUE-TERRA consortium. 2023. Metagenome-assembled genomes indicate that antimicrobial resistance genes are highly prevalent among urban bacteria and multidrug and glycopeptide resistances are ubiquitous in most taxa. *Frontiers in Microbiology* 14. Frontiers. https://doi.org/10.3389/fmicb.2023.1037845.
254. Saxena, Gourvendu, Suparna Mitra, Ezequiel M. Marzinelli, Chao Xie, Toh Jun Wei, Peter D. Steinberg, Rohan B. H. Williams, Staffan Kjelleberg, Federico M. Lauro, and Sanjay Swarup. 2018. Metagenomics Reveals the Influence of Land Use and Rain on the Benthic Microbial Communities in a Tropical Urban Waterway. *mSystems* 3. American Society for Microbiology: 10.1128/msystems.00136-17. https://doi.org/10.1128/msystems.00136-17.
255. Noman, Sohail M., Muhammad Shafiq, Shabana Bibi, Bharti Mittal, Yumeng Yuan, Mi Zeng, Xin Li, Oluwaseyi Abraham Olawale, Xiaoyang Jiao, and Muhammad Irshad. 2023. Exploring antibiotic resistance genes, mobile gene elements, and virulence gene factors in an urban freshwater samples using metagenomic analysis. *Environmental Science and Pollution Research* 30: 2977–2990. https://doi.org/10.1007/s11356-022-22197-4.
256. Magnúsdóttir, Stefanía, Joao Pedro Saraiva, Alexander Bartholomäus, Majid Soheili, Rodolfo Brizola Toscan, Junya Zhang, Ulisses Nunes da Rocha, and CLUE-TERRA consortium. 2023. Metagenome-assembled genomes indicate that antimicrobial resistance genes are highly prevalent among urban bacteria and multidrug and glycopeptide resistances are ubiquitous in most taxa. *Frontiers in Microbiology* 14. Frontiers. https://doi.org/10.3389/fmicb.2023.1037845.

257. Sayers, Eric W, Jeffrey Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Ryan Connor, Michael Feldgarden, et al. 2025. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research* 53: D20–D29. https://doi.org/10.1093/nar/gkae979.

258. Seemann, Torsten. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30: 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

259. Tonkin-Hill, Gerry, Neil MacAlasdair, Christopher Ruis, Aaron Weimann, Gal Horesh, John A. Lees, Rebecca A. Gladstone, et al. 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* 21: 180. https://doi.org/10.1186/s13059-020-02090-4.

260. Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37: 1530–1534. https://doi.org/10.1093/molbev/msaa015.

261. Croucher, Nicholas J., Andrew J. Page, Thomas R. Connor, Aidan J. Delaney, Jacqueline A. Keane, Stephen D. Bentley, Julian Parkhill, and Simon R. Harris. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* 43: e15. https://doi.org/10.1093/nar/gku1196.

262. R: The R Project for Statistical Computing. 2025. https://www.r-project.org/. Accessed April 21.

263. Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20: 289–290. https://doi.org/10.1093/bioinformatics/btg412.

264. Revell, Liam J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

265. Maechler, Martin, Christophe Dutang, Vincent Goulet, Douglas Bates (cosmetic clean up, in svn r42), David Firth (expm(method= “PadeO” and “TaylorO”)), Marina Shapira (expm(method= “PadeO” and “TaylorO”)), Michael Stadelmann (“Higham08*” methods, and see ?expm.Higham08...). 2024. expm: Matrix Exponential, Log, “etc” (version 1.0-0).

266. Wickham, Hadley. 2016. *ggplot2*. Use R! Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4.

267. 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.

268. Yu, Guangchuang, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8: 28–36. https://doi.org/10.1111/2041-210X.12628.

269. Kolde, Raivo. 2019. pheatmap: Pretty Heatmaps (version 1.0.12).

270. Schliep, Klaus Peter. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27: 592–593. https://doi.org/10.1093/bioinformatics/btq706.

271. Lewis, Paul O. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology* 50: 913–925. https://doi.org/10.1080/106351501753462876.

272. Byrd, Richard H., Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* 16. Society for Industrial and Applied Mathematics: 1190–1208. https://doi.org/10.1137/0916069.

273. Pawitan, Yudi. 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press. https://doi.org/10.1093/oso/9780198507659.001.0001.

274. Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14. Nature Publishing Group: 587–589. https://doi.org/10.1038/nmeth.4285.

275. Arnold O. Allen. 1990. Probability, Statistics, and Queueing Theory. *ScienceDirect*.

276. Harvey, Paul H, and Mark D Pagel. 1991. *The Comparative Method in Evolutionary Biology*. Oxford University Press. https://doi.org/10.1093/oso/9780198546412.001.0001.

277. Yang, Ziheng. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199602605.001.0001.

278. 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.

279. 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.

280. Burnham, Kenneth P., and David R. Anderson, ed. 2004. *Model Selection and Multimodel Inference*. New York, NY: Springer. https://doi.org/10.1007/b97636.

281. Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20: 289–290. https://doi.org/10.1093/bioinformatics/btg412.

282. 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.

283. Fitch, Walter M. 1971. Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Biology* 20: 406–416. https://doi.org/10.1093/sysbio/20.4.406.

284. Revell, Liam J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 12. PeerJ Inc.: e16505. https://doi.org/10.7717/peerj.16505.

285. Revell, Liam J., and Luke J. Harmon. 2022. *Phylogenetic Comparative Methods in R*. Princeton University Press.

286. Tettelin, Hervé, David Riley, Ciro Cattuto, and Duccio Medini. 2008. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology* 11. Antimicrobials/Genomics: 472–477. https://doi.org/10.1016/j.mib.2008.09.006.

287. Heaps, H. S. 1978. *Information Retrieval: Computational and Theoretical Aspects*. USA: Academic Press, Inc.

288. 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.

289. Drula, Elodie, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, and Nicolas Terrapon. 2022. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research* 50: D571–D577. https://doi.org/10.1093/nar/gkab1045.

290. Galperin, Michael Y, Yuri I Wolf, Kira S Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V Koonin. 2021. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research* 49: D274–D281. https://doi.org/10.1093/nar/gkaa1018.

291. McDonald, Andrew G., and Keith F. Tipton. 2023. Enzyme nomenclature and classification: the state of the art. *The FEBS Journal* 290: 2214–2231. https://doi.org/10.1111/febs.16274.

292. Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. Gene Ontology: tool for the unification of

biology. *Nature Genetics* 25. Nature Publishing Group: 25–29. https://doi.org/10.1038/75556.

293. Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 27: 29–34. https://doi.org/10.1093/nar/27.1.29.
294. Colquhoun, Rachel M., Michael B. Hall, Leandro Lima, Leah W. Roberts, Kerri M. Malone, Martin Hunt, Brice Letcher, et al. 2021. Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biology* 22: 267. https://doi.org/10.1186/s13059-021-02473-1.
295. Park, Sang-Cheol, Kihyun Lee, Yeong Ouk Kim, Sungho Won, and Jongsik Chun. 2019. Large-Scale Genomics Reveals the Genetic Characteristics of Seven Species and Importance of Phylogenetic Distance for Estimating Pan-Genome Size. *Frontiers in Microbiology* 10. Frontiers. https://doi.org/10.3389/fmicb.2019.00834.
296. Greenacre, Michael, Patrick J. F. Groenen, Trevor Hastie, Alfonso Iodice D'Enza, Angelos Markos, and Elena Tuzhilina. 2022. Principal component analysis. *Nature Reviews Methods Primers* 2. Nature Publishing Group: 1–21. https://doi.org/10.1038/s43586-022-00184-w.
297. *Principal Component Analysis*. 2002. Springer Series in Statistics. New York: Springer-Verlag. https://doi.org/10.1007/b98835.
298. Murtagh, Fionn, and Pierre Legendre. 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification* 31: 274–295. https://doi.org/10.1007/s00357-014-9161-z.
299. Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, and David Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95. Proceedings of the National Academy of Sciences: 14863–14868. https://doi.org/10.1073/pnas.95.25.14863.
300. . Wyres, Kelly L, and Kathryn E Holt. 2018. *Klebsiella pneumoniae* as a key trafficker of drug resistance genes from environmental to clinically important bacteria. *Current Opinion in Microbiology* 45. Antimicrobials * Microbial Systems Biology: 131–139. https://doi.org/10.1016/j.mib.2018.04.004.
301. Rosenblueth, Mónica, Lucía Martínez, Jesús Silva, and Esperanza Martínez-Romero. 2004. *Klebsiella variicola*, A Novel Species with Clinical and Plant-Associated Isolates. *Systematic and Applied Microbiology* 27: 27–35. https://doi.org/10.1078/0723-2020-00261.
302. Brisse, Sylvain, Virginie Passet, and Patrick A. D. Grimont. 2014. Description of Klebsiellaquasipneumoniae sp. nov., isolated from human infections, with two subspecies, Klebsiellaquasipneumoniae subsp. quasipneumoniae subsp. nov. and Klebsiellaquasipneumoniae subsp. similipneumoniae subsp. nov., and demonstration that Klebsiella singaporensis is a junior heterotypic synonym of Klebsiella variicola. *International Journal of Systematic and Evolutionary Microbiology* 64. Microbiology Society,: 3146–3152. https://doi.org/10.1099/ijs.0.062737-0.
303. Long, Haiyan, Ya Hu, Yu Feng, and Zhiyong Zong. 2022. Genome Analysis of Klebsiella oxytoca Complex for Antimicrobial Resistance and Virulence Genes. *Antimicrobial Agents and Chemotherapy* 66. American Society for Microbiology: e02183-21. https://doi.org/10.1128/aac.02183-21.
304. Moradigaravand, Danesh, Veronique Martin, Sharon J. Peacock, and Julian Parkhill. 2017. Population Structure of Multidrug-Resistant Klebsiella oxytoca within Hospitals across the United Kingdom and Ireland Identifies Sharing of Virulence and Resistance Genes with K. pneumoniae. *Genome Biology and Evolution* 9: 574–584. https://doi.org/10.1093/gbe/evx019.
305. Simoni, Serena, Francesca Leoni, Laura Veschetti, Giovanni Malerba, Maria Carelli, Maria M. Lleò, Andrea Brenciani, et al. 2022. The Emerging Nosocomial Pathogen Klebsiella

michiganensis: Genetic Analysis of a KPC-3 Producing Strain Isolated from Venus Clam. *Microbiology Spectrum* 11. American Society for Microbiology: e04235-22. https://doi.org/10.1128/spectrum.04235-22.

306. Prah, Isaac, Yoko Nukui, Shoji Yamaoka, and Ryoichi Saito. 2022. Emergence of a High-Risk Klebsiella michiganensis Clone Disseminating Carbapenemase Genes. *Frontiers in Microbiology* 13. Frontiers. https://doi.org/10.3389/fmicb.2022.880248.
307. Holt, Kathryn E., Heiman Wertheim, Ruth N. Zadoks, Stephen Baker, Chris A. Whitehouse, David Dance, Adam Jenney, et al. 2015. Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proceedings of the National Academy of Sciences* 112. Proceedings of the National Academy of Sciences: E3574–E3581. https://doi.org/10.1073/pnas.1501049112.
308. Brüssow, Harald, Carlos Canchaya, and Wolf-Dietrich Hardt. 2004. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews* 68. American Society for Microbiology: 560–602. https://doi.org/10.1128/mmbr.68.3.560-602.2004.
309. Bobay, Louis-Marie, Marie Touchon, and Eduardo P. C. Rocha. 2013. Manipulating or Superseding Host Recombination Functions: A Dilemma That Shapes Phage Evolvability. *PLOS Genetics* 9. Public Library of Science: e1003825. https://doi.org/10.1371/journal.pgen.1003825.
310. Haft, Daniel H, Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Kathleen O'Neill, Wenjun Li, et al. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Research* 46: D851–D860. https://doi.org/10.1093/nar/gkx1068.
311. Sayers, Eric W, Mark Cavanaugh, Karen Clark, Kim D Pruitt, Stephen T Sherry, Linda Yankie, and Ilene Karsch-Mizrachi. 2024. GenBank 2024 Update. *Nucleic Acids Research* 52: D134–D137. https://doi.org/10.1093/nar/gkad903.
312. 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.
313. 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.
314. Yang, Jing, Haiyan Long, Ya Hu, Yu Feng, Alan McNally, and Zhiyong Zong. 2021. Klebsiella oxytoca Complex: Update on Taxonomy, Antimicrobial Resistance, and Virulence. *Clinical Microbiology Reviews* 35. American Society for Microbiology: e00006-21. https://doi.org/10.1128/CMR.00006-21.
315. 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.
316. Polz, Martin F., Eric J. Alm, and William P. Hanage. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics* 29. Elsevier: 170–175. https://doi.org/10.1016/j.tig.2012.12.006.
317. Shapiro, B. Jesse. 2017. The population genetics of pangenomes. *Nature Microbiology* 2. Nature Publishing Group: 1574–1574. https://doi.org/10.1038/s41564-017-0066-6.
318. Rocha, Eduardo P. C. 2008. The Organization of the Bacterial Genome. *Annual Review of Genetics* 42. Annual Reviews: 211–233. https://doi.org/10.1146/annurev.genet.42.110807.091653.
319. 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.

320. 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.
321. Smith, J M, N H Smith, M O'Rourke, and B G Spratt. 1993. How clonal are bacteria? *Proceedings of the National Academy of Sciences* 90. Proceedings of the National Academy of Sciences: 4384–4388. https://doi.org/10.1073/pnas.90.10.4384.
322. Vos, Michiel, and Xavier Didelot. 2009. A comparison of homologous recombination rates in bacteria and archaea. *The ISME Journal* 3: 199–208. https://doi.org/10.1038/ismej.2008.93.
323. Partridge, Sally R., Stephen M. Kwong, Neville Firth, and Slade O. Jensen. 2018. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clinical Microbiology Reviews* 31. American Society for Microbiology: 10.1128/cmr.00088-17. https://doi.org/10.1128/cmr.00088-17.
324. Schmutzer, Michael, and Timothy Giles Barraclough. 2019. The role of recombination, niche-specific gene pools and flexible genomes in the ecological speciation of bacteria. *Ecology and Evolution* 9: 4544–4556. https://doi.org/10.1002/ece3.5052.
325. 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.
326. Kuo, Chih-Horng, and Howard Ochman. 2010. The Extinction Dynamics of Bacterial Pseudogenes. *PLOS Genetics* 6. Public Library of Science: e1001050. https://doi.org/10.1371/journal.pgen.1001050.
327. Chen, Lin-Xing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. Accurate and complete genomes from metagenomes. *Genome Research* 30. Cold Spring Harbor Lab: 315–333. https://doi.org/10.1101/gr.258640.119.
328. Meyer, Fernando, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, et al. 2022. Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nature Methods* 19. Nature Publishing Group: 429–440. https://doi.org/10.1038/s41592-022-01431-4.
329. Xu, Panpan, Di Zhang, Wanqing Zhuo, Lin Zhou, Yue Du, Peipei Zhang, Lijuan Ma, and Yajuan Wang. 2024. Characterization of a Highly Virulent <em>Klebsiella michiganensis</em> Strain Isolated from a Preterm Infant with Sepsis. *Infection and Drug Resistance* 17. Dove Press: 4973–4983. https://doi.org/10.2147/IDR.S481750.
330. Sun, Yong, Qingqing Cai, Tianyu Li, Jingbo Chen, and Yuan Fang. 2025. Genome assembly of Klebsiella michiganensis based on metagenomic next-generation sequencing reveals its genomic characteristics in population genetics and molecular epidemiology. *Frontiers in Microbiology* 16. Frontiers. https://doi.org/10.3389/fmicb.2025.1546594.
331. Shaiber, Alon, and A. Murat Eren. 2019. Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories. *mBio* 10. American Society for Microbiology: 10.1128/mbio.00725-19. https://doi.org/10.1128/mbio.00725-19.
332. Ho, Simon Y. W., and Sebastián Duchêne. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Molecular Ecology* 23: 5947–5965. https://doi.org/10.1111/mec.12953.
333. Bromham, Lindell, and David Penny. 2003. The modern molecular clock. *Nature Reviews Genetics* 4. Nature Publishing Group: 216–224. https://doi.org/10.1038/nrg1020.
334. Drummond, Alexei J., Simon Y. W. Ho, Matthew J. Phillips, and Andrew Rambaut. 2006. Relaxed Phylogenetics and Dating with Confidence. *PLOS Biology* 4. Public Library of Science: e88. https://doi.org/10.1371/journal.pbio.0040088.

335. Martin, Rebekah M., and Michael A. Bachman. 2018. Colonization, Infection, and the Accessory Genome of Klebsiella pneumoniae. *Frontiers in Cellular and Infection Microbiology* 8. Frontiers. https://doi.org/10.3389/fcimb.2018.00004.

336. Lam, Margaret M. C., Ryan R. Wick, Kelly L. Wyres, Claire L. Gorrie, Louise M. Judd, Adam W. J. Jenney, Sylvain Brisse, and Kathryn E. Holt. 2018. Genetic diversity, mobilisation and spread of the yersiniabactin-encoding mobile element ICEKp in Klebsiella pneumoniae populations. *Microbial Genomics* 4. Microbiology Society,: e000196. https://doi.org/10.1099/mgen.0.000196.

337. Mathers, Amy J., Gisele Peirano, and Johann D. D. Pitout. 2015. The Role of Epidemic Resistance Plasmids and International High-Risk Clones in the Spread of Multidrug-Resistant Enterobacteriaceae. *Clinical Microbiology Reviews* 28. American Society for Microbiology: 565–591. https://doi.org/10.1128/cmr.00116-14.

338. Lerminiaux, Nicole A., and Andrew D.S. Cameron. 2019. Horizontal transfer of antibiotic resistance genes in clinical environments. *Canadian Journal of Microbiology* 65. NRC Research Press: 34–44. https://doi.org/10.1139/cjm-2018-0275.

339. Drummond, Alexei J., and Remco R. Bouckaert. 2015. *Bayesian Evolutionary Analysis with BEAST*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139095112.

340. 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.

341. Ohta, Tomoko. 1992. The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology, Evolution, and Systematics* 23. Annual Reviews: 263–286. https://doi.org/10.1146/annurev.es.23.110192.001403.

342. Kimura, Motoo. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511623486.

343. 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.

344. 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.

345. Cordero, Otto X., and Martin F. Polz. 2014. Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology* 12. Nature Publishing Group: 263–273. https://doi.org/10.1038/nrmicro3218.

346. 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.

347. Felsenstein, Joseph, and Joseph Felsenstein. 2003. *Inferring Phylogenies*. Oxford, New York: Oxford University Press.

348. Nielsen, Rasmus. 2002. Mapping Mutations on Phylogenies. *Systematic Biology* 51: 729–739. https://doi.org/10.1080/10635150290102393.

ANNEX 1. Source datasets and metadata for MPKG, PKG and PKP datasets

Table A1.1. MPKG dataset metadata

Organism Scientific Name	Taxonomy id	Assembly Name	Assembly Accession	Level	Contig N50	Size	BioProject	BioSample
Klebsiella pneumoniae	573	ASM3197461v1	GCA_031974615.1	Scaffold	8924	4536574	PRJNA850115	SAMN29156382
Klebsiella michiganensis	1134687	ASM3197491v1	GCA_031974915.1	Scaffold	46608	5637797	PRJNA850115	SAMN29156368
Klebsiella pneumoniae	573	ASM3197661v1	GCA_031976615.1	Scaffold	95709	5095361	PRJNA850115	SAMN29156457
Klebsiella variicola	244366	ASM3197762v1	GCA_031977625.1	Scaffold	143594	5397350	PRJNA850115	SAMN29159067
Klebsiella pneumoniae	573	ASM3197826v1	GCA_031978265.1	Scaffold	94162	5340135	PRJNA850115	SAMN29159037
Klebsiella pneumoniae	573	ASM3197848v1	GCA_031978485.1	Scaffold	129128	5132067	PRJNA850115	SAMN29159025
Klebsiella michiganensis	1134687	ASM3197860v1	GCA_031978605.1	Scaffold	1989	3934098	PRJNA850115	SAMN29159021
Klebsiella pneumoniae	573	ASM3197888v1	GCA_031978885.1	Scaffold	10475	4645042	PRJNA850115	SAMN29159009
Klebsiella michiganensis	1134687	ASM3197968v1	GCA_031979685.1	Scaffold	123103	5701375	PRJNA850115	SAMN29158982
Klebsiella variicola	244366	ASM3197980v1	GCA_031979805.1	Scaffold	134212	5243289	PRJNA850115	SAMN29158979
Klebsiella michiganensis	1134687	ASM3197984v1	GCA_031979845.1	Scaffold	11302	5464758	PRJNA850115	SAMN29158975
Klebsiella michiganensis	1134687	ASM3198012v1	GCA_031980125.1	Scaffold	143358	5929369	PRJNA850115	SAMN29158961
Klebsiella michiganensis	1134687	ASM3198026v1	GCA_031980265.1	Scaffold	65644	5700745	PRJNA850115	SAMN29158955
Klebsiella michiganensis	1134687	ASM3198032v1	GCA_031980325.1	Scaffold	132575	5686946	PRJNA850115	SAMN29158950
Klebsiella michiganensis	1134687	ASM3198042v1	GCA_031980425.1	Scaffold	140658	5776060	PRJNA850115	SAMN29158947
Klebsiella pneumoniae	573	ASM3198112v1	GCA_031981125.1	Scaffold	60934	4818337	PRJNA850115	SAMN29159252
Klebsiella pneumoniae	573	ASM3198140v1	GCA_031981405.1	Scaffold	113467	5511502	PRJNA850115	SAMN29159237
Klebsiella pneumoniae	573	ASM3198198v1	GCA_031981985.1	Scaffold	16914	4879073	PRJNA850115	SAMN29159208
Klebsiella pneumoniae	573	ASM3198288v1	GCA_031982885.1	Scaffold	109697	5030109	PRJNA850115	SAMN29159166
Klebsiella pneumoniae	573	ASM3198326v1	GCA_031983265.1	Scaffold	29240	4565805	PRJNA850115	SAMN29159148
Klebsiella pneumoniae	573	ASM3198352v1	GCA_031983525.1	Scaffold	35376	4275052	PRJNA850115	SAMN29159136
Klebsiella pneumoniae	573	ASM3198400v1	GCA_031984005.1	Scaffold	33098	5135774	PRJNA850115	SAMN29159111

Klebsiella oxytoca	571	ASM3198600v1	GCA_031986005.1	Scaffold	28907	5381023	PRJNA850115	SAMN29160655
Klebsiella pneumoniae	573	ASM3198688v1	GCA_031986885.1	Scaffold	99630	5565665	PRJNA850115	SAMN29160633
Klebsiella variicola	244366	ASM3198746v1	GCA_031987465.1	Scaffold	150064	5272748	PRJNA850115	SAMN29160620
Klebsiella pneumoniae	573	ASM3199292v1	GCA_031992925.1	Scaffold	172752	5138273	PRJNA850115	SAMN29159268
Klebsiella variicola	244366	ASM3199392v1	GCA_031993925.1	Scaffold	4502	4214911	PRJNA850115	SAMN29160832
Klebsiella pneumoniae	573	ASM3199464v1	GCA_031994645.1	Scaffold	8884	5533584	PRJNA850115	SAMN29160816
Klebsiella pneumoniae	573	ASM3199850v1	GCA_031998505.1	Scaffold	114717	5332399	PRJNA850115	SAMN29160718
Klebsiella oxytoca	571	ASM3200126v1	GCA_032001265.1	Scaffold	202468	5755155	PRJNA850115	SAMN29160900
Klebsiella michiganensis	1134687	ASM3200190v1	GCA_032001905.1	Scaffold	4644	5661280	PRJNA850115	SAMN29160886
Klebsiella pneumoniae	573	ASM3200256v1	GCA_032002565.1	Scaffold	177418	5068932	PRJNA850115	SAMN29160869
Klebsiella variicola	244366	ASM3200680v1	GCA_032006805.1	Scaffold	182862	5579730	PRJNA850115	SAMN29159946
Klebsiella variicola	244366	ASM3200939v1	GCA_032009395.1	Scaffold	67096	5240718	PRJNA850115	SAMN29159879
Klebsiella pneumoniae	573	ASM3201078v1	GCA_032010785.1	Scaffold	156149	5198552	PRJNA850115	SAMN29159843
Klebsiella michiganensis	1134687	ASM3201088v1	GCA_032010885.1	Scaffold	75241	5712396	PRJNA850115	SAMN29159841
Klebsiella pneumoniae	573	ASM3201145v1	GCA_032011455.1	Scaffold	124252	5357102	PRJNA850115	SAMN29159824
Klebsiella pneumoniae	573	ASM3201152v1	GCA_032011525.1	Scaffold	95670	5288370	PRJNA850115	SAMN29159818
Klebsiella michiganensis	1134687	ASM3205424v1	GCA_032054245.1	Scaffold	47792	6178226	PRJNA850115	SAMN29161927
Klebsiella pneumoniae	573	ASM3205764v1	GCA_032057645.1	Scaffold	70901	4925819	PRJNA850115	SAMN29162232
Klebsiella michiganensis	1134687	ASM3206078v1	GCA_032060785.1	Scaffold	2060	3707121	PRJNA850115	SAMN29161482
Klebsiella michiganensis	1134687	ASM3206524v1	GCA_032065245.1	Scaffold	176732	6178048	PRJNA850115	SAMN29161626
Klebsiella pneumoniae	573	ASM3206764v1	GCA_032067645.1	Scaffold	4823	4551093	PRJNA850115	SAMN29161552
Klebsiella oxytoca	571	ASM3207224v1	GCA_032072245.1	Scaffold	159845	5780502	PRJNA850115	SAMN29161022
Klebsiella aerogenes	548	ASM3207388v1	GCA_032073885.1	Scaffold	26370	4622068	PRJNA850115	SAMN29160963
Klebsiella pneumoniae	573	ASM3207902v1	GCA_032079025.1	Scaffold	116783	5211853	PRJNA850115	SAMN29161167
Klebsiella oxytoca	571	ASM3208641v1	GCA_032086415.1	Scaffold	14214	5516882	PRJNA850115	SAMN29163403
Klebsiella aerogenes	548	ASM3208679v1	GCA_032086795.1	Scaffold	100691	4971102	PRJNA850115	SAMN29164030
Klebsiella aerogenes	548	ASM3208711v1	GCA_032087115.1	Scaffold	112433	5048697	PRJNA850115	SAMN29164014

Klebsiella huaxiensis	2153354	ASM3208725v1	GCA_032087255.1	Scaffold	2790	5212165	PRJNA850115	SAMN29164007
Klebsiella michiganensis	1134687	ASM3208729v1	GCA_032087295.1	Scaffold	1894	3271432	PRJNA850115	SAMN29164004
Klebsiella michiganensis	1134687	ASM3208733v1	GCA_032087335.1	Scaffold	2323	4078816	PRJNA850115	SAMN29164003
Klebsiella oxytoca	571	ASM3209475v1	GCA_032094755.1	Scaffold	227978	5936634	PRJNA850115	SAMN29163503
Klebsiella oxytoca	571	ASM3209625v1	GCA_032096255.1	Scaffold	81670	5833154	PRJNA850115	SAMN29163464
Klebsiella variicola	244366	ASM3210498v1	GCA_032104985.1	Scaffold	45181	5447014	PRJNA850115	SAMN29162995
Klebsiella pneumoniae	573	ASM3210589v1	GCA_032105895.1	Scaffold	39861	4699530	PRJNA850115	SAMN29163251
Klebsiella pneumoniae	573	ASM3210716v1	GCA_032107165.1	Scaffold	73725	4944969	PRJNA850115	SAMN29163214
Klebsiella pneumoniae	573	ASM3210914v1	GCA_032109145.1	Scaffold	247680	5375292	PRJNA850115	SAMN29163146
Klebsiella pneumoniae	573	ASM3210924v1	GCA_032109245.1	Scaffold	13386	4646431	PRJNA850115	SAMN29163141
Klebsiella africana	2489010	ASM3211054v1	GCA_032110545.1	Scaffold	47932	4967355	PRJNA850115	SAMN29163107
Klebsiella pneumoniae	573	ASM3211188v1	GCA_032111885.1	Scaffold	79541	4888767	PRJNA850115	SAMN29163385
Klebsiella oxytoca	571	ASM3211248v1	GCA_032112485.1	Scaffold	221850	5828758	PRJNA850115	SAMN29163368
Klebsiella oxytoca	571	ASM3211509v1	GCA_032115095.1	Scaffold	290550	5788139	PRJNA850115	SAMN29163299
Klebsiella africana	2489010	ASM3211054v1	GCF_032110545.1	Scaffold	47932	4967355	PRJNA850115	SAMN29163107

Table A1.2. PKG dataset metadata

Organism Scientific Name	Taxonomy id	Assembly Name	Assembly Accession	Level	Contig N50	Size	BioProject	BioSample
Klebsiella aerogenes	1028307	ASM21574v1	GCF_000215745.1	Complete	5280350	5280350	PRJNA66537	SAMN02603581
Klebsiella aerogenes	548	ASM157154v2	GCF_001571545.2	Complete	5452368	5626434	PRJNA279469	SAMN03448049
Klebsiella aerogenes	548	ASM1904812v1	GCF_019048125.1	Complete	5281764	5281764	PRJNA231221	SAMN16357584
Klebsiella aerogenes	548	ASM2759570v1	GCF_027595705.1	Complete	5443727	5487446	PRJNA667445	SAMN16387652
Klebsiella aerogenes	548	ASM3742935v2	GCF_037429355.1	Complete	5409695	5409695	PRJNA812595	SAMN40557126
Klebsiella africana	2489010	ASM1680412v1	GCF_016804125.1	Complete	5291121	5422247	PRJNA646592	SAMN15547534
Klebsiella africana	2489010	ASM2052608v1	GCF_020526085.1	Complete	5243981	5443802	PRJNA768294	SAMN22024779
Klebsiella africana	2489010	ASM2370403v1	GCF_023704035.1	Complete	5268423	5268423	PRJNA646592	SAMN15547540
Klebsiella africana	2489010	200023.1	GCF_900978845.1	Contig	195731	5156720	PRJEB29143	SAMEA4969318
Klebsiella africana	2489010	CIP111653T	GCF_965139595.1	Scaffold	158804	5171015	PRJEB85433	SAMEA117660618
Klebsiella huaxiensis	2153354	ASM326157v2	GCF_003261575.2	Complete	6183608	6300829	PRJNA353728	SAMN08861555
Klebsiella huaxiensis	2153354	ASM2673541v2	GCF_026735415.2	Contig	6314697	6875149	PRJNA907249	SAMN31964237
Klebsiella huaxiensis	2153354	ASM3873769v1	GCF_038737695.1	Contig	284828	6220856	PRJNA316969	SAMN40982453
Klebsiella huaxiensis	2153354	SB6422	GCF_902158605.1	Scaffold	318410	6208089	PRJEB15325	SAMEA5610086
Klebsiella huaxiensis	2153354	SB6421	GCF_902158625.1	Scaffold	200809	6159557	PRJEB15325	SAMEA5610085
Klebsiella michiganensis	1134687	ASM1513957v1	GCF_015139575.1	Complete	5935402	6041841	PRJDB9036	SAMD00196009
Klebsiella michiganensis	1134687	ASM1661821v1	GCF_016618215.1	Complete	6330740	6865058	PRJNA664790	SAMN16233476
Klebsiella michiganensis	1134687	ASM4021511v1	GCF_040215115.1	Complete	6491865	6692448	PRJNA1120221	SAMN41684531
Klebsiella michiganensis	1134687	ASM4128348v1	GCF_041283485.1	Complete	6301770	6925468	PRJNA812595	SAMN32093411
Klebsiella michiganensis	1134687	ASM5092191v1	GCF_050921915.1	Complete	6545023	6600419	PRJNA907198	SAMN31956146
Klebsiella oxytoca	571	ASM102211v1	GCF_001022115.1	Complete	6229565	6673117	PRJNA246471	SAMN03733750
Klebsiella oxytoca	571	ASM102229v1	GCF_001022295.1	Complete	6217725	6630537	PRJNA246471	SAMN03733663

Klebsiella oxytoca	571	ASM187018v1	GCF_001870185.1	Complete	6155924	6582387	PRJNA246471	SAMN03733631
Klebsiella oxytoca	571	ASM1375059v1	GCF_013750595.1	Complete	6091081	6293551	PRJNA605147	SAMN15148597
Klebsiella oxytoca	571	45889_C01	GCF_900636985.1	Complete	5857964	5857964	PRJEB6403	SAMEA3923594
Klebsiella pneumoniae	1125630	ASM24018v2	GCF_000240185.1	Complete	5333942	5682322	PRJNA78789	SAMN02602959
Klebsiella pneumoniae	573	ASM636429v1	GCF_006364295.1	Complete	5303035	5573867	PRJNA231221	SAMN11056490
Klebsiella pneumoniae	573	ASM1104559v1	GCF_011045595.1	Complete	5803733	5803733	PRJNA559783	SAMN12559024
Klebsiella pneumoniae	573	ASM2286966v1	GCF_022869665.1	Complete	5303036	5563484	PRJNA605254	SAMN14078806
Klebsiella pneumoniae	573	ASM5015618v1	GCF_050156185.1	Complete	5785925	5959413	PRJNA1043403	SAMN38339031
Klebsiella variicola	244366	ASM82805v2	GCF_000828055.2	Complete	5521203	5521203	PRJNA272370	SAMN01174581
Klebsiella variicola	244366	ASM1763894v1	GCF_017638945.1	Complete	5519584	5519584	PRJNA716670	SAMN18446056
Klebsiella variicola	244366	ASM1832404v1	GCF_018324045.1	Complete	5564085	5732600	PRJDB11476	SAMD00294603
Klebsiella variicola	2590157	ASM2052554v1	GCF_020525545.1	Complete	5521194	5521194	PRJNA768294	SAMN22024815
Klebsiella variicola	244366	ASM3559407v1	GCF_035594075.1	Complete	5727068	6213895	PRJNA1050876	SAMN38756473

Table A1.3. PKP dataset metadata

No.	Isolate ID	ENA accession number	Year of isolation	Biosubstrate	Isolate origin	Hospital department
1.	1030710	ERS24931633	2020	blood	Hospital	Intensive care unit
2.	1030711	ERS24931634	2020	blood	Hospital	Intensive care unit
3.	1030712	ERS24931635	2020	blood	Hospital	Intensive care unit
4.	1030713	ERS24931636	2020	CSF	Hospital	Intensive care unit
5.	1030714	ERS24931637	2020	CSF	Hospital	Intensive care unit
6.	1030715	ERS24931638	2020	blood	Hospital	Intensive care unit
7.	1030716	ERS24931639	2020	blood	Hospital	Intensive care unit
8.	1030717	ERS24931640	2020	blood	Hospital	Intensive care unit
9.	1030718	ERS24931641	2020	blood	Hospital	Intensive care unit
10.	1030719	ERS24931642	2020	blood	Hospital	Other*
11.	1030776	ERS24931699	2020	urine	Ambulatory	Consultative
12.	1030777	ERS24931700	2020	urine	Ambulatory	Consultative
13.	1030722	ERS24931645	2021	Blood	Hospital	Other*
14.	1030723	ERS24931646	2021	Blood	Hospital	Other*
15.	1030724	ERS24931647	2021	Blood	Hospital	Intensive care unit
16.	1030725	ERS24931648	2021	Blood	Hospital	Other*
17.	1030726	ERS24931649	2021	Blood	Hospital	Other*
18.	1030727	ERS24931650	2021	Blood	Hospital	Other*
19	1030728	ERS24931651	2021	Blood	Hospital	Intensive care unit
	1030729	ERS24931652	2021	Blood	Hospital	Intensive care unit
21	1030730	ERS24931653	2021	Blood	Hospital	Paediatry
22	1030731	ERS24931654	2021	Blood	Hospital	Intensive care unit
23.	1030732	ERS24931655	2021	Blood	Hospital	Internal medicine
24.	1030733	ERS24931656	2021	Blood	Hospital	Paediatry
25.	1030734	ERS24931657	2021	Blood	Hospital	Other*
26.	1030736	ERS24931659	2021	Blood	Hospital	Other*
27.	1030737	ERS24931660	2021	Blood	Hospital	Intensive care unit
28.	1030741	ERS24931664	2022	blood	Hospital	Intensive care unit
29.	1030744	ERS24931667	2022	blood	Hospital	Surgery
30.	1030745	ERS24931668	2022	blood	Hospital	Intensive care unit
31.	1030750	ERS24931673	2022	blood	Hospital	Intensive care unit
32.	1030751	ERS24931674	2022	blood	Hospital	Intensive care unit
33.	1030752	ERS24931675	2022	blood	Hospital	Other*
34.	1030753	ERS24931676	2022	blood	Hospital	Intensive care unit
35.	1030754	ERS24931677	2022	CSF	Hospital	Paediatry
36.	1030755	ERS24931678	2022	blood	Hospital	Intensive care unit
37.	1030756	ERS24931679	2022	blood	Hospital	Intensive care unit
38.	1030757	ERS24931680	2022	blood	Hospital	Intensive care unit
39.	1030758	ERS24931681	2022	blood	Hospital	Internal medicine
40.	1030759	ERS24931682	2022	blood	Hospital	Intensive care unit
41.	1030760	ERS24931683	2022	blood	Hospital	Intensive care unit
42.	1030762	ERS24931685	2022	blood	Hospital	Internal medicine
43.	1030763	ERS24931686	2022	blood	Hospital	Intensive care unit
44.	1030764	ERS24931687	2022	blood	Hospital	Intensive care unit
45.	1030765	ERS24931688	2022	blood	Hospital	Paediatric intensive care unit
46.	1030766	ERS24931689	2022	blood	Hospital	Intensive care unit
47.	1030767	ERS24931690	2022	blood	Hospital	Intensive care unit
48.	1030768	ERS24931691	2022	blood	Hospital	Paediatric intensive care unit
49.	1030769	ERS24931692	2022	blood	Hospital	Paediatric intensive care unit
50.	1030780	ERS24931703	2023	Urine	Hospital	Intensive care unit
51.	1030781	ERS24931704	2023	Urine	Hospital	Urology

52.	1030782	ERS24931705	2023	Urine	Hospital	Urology
53.	1030783	ERS24931706	2023	Urine	Hospital	Urology
54.	1030784	ERS24931707	2023	Urine	Hospital	Internal medicine
55.	1030785	ERS24931708	2023	Urine	Hospital	Intensive care unit
56.	1030789	ERS24931712	2023	Urine	Ambulatory	Consultative
57.	1030803	ERS24931726	2023	Urine	Hospital	Urology
58.	1030804	ERS24931727	2023	Urine	Hospital	Urology
59.	1030805	ERS24931728	2023	Urine	Hospital	Intensive care unit
60.	1030806	ERS24931729	2023	Urine	Hospital	Internal medicine
61.	1030807	ERS24931730	2023	Urine	Ambulatory	Consultative
62.	1030809	ERS24931732	2023	Blood	Hospital	Intensive care unit
63.	1030810	ERS24931733	2023	Blood	Hospital	Intensive care unit
64.	1030811	ERS24931734	2023	Blood	Hospital	Intensive care unit
65.	1030812	ERS24931735	2023	Blood	Hospital	Intensive care unit
66.	1030813	ERS24931736	2023	Blood	Hospital	Intensive care unit
67.	1030814	ERS24931737	2023	Blood	Hospital	Intensive care unit
68.	1030816	ERS24931739	2023	Blood	Hospital	Intensive care unit
69.	1030819	ERS24931742	2023	Urine	Hospital	Intensive care unit
70.	1030820	ERS24931743	2023	Urine	Hospital	Urology
71.	1030821	ERS24931744	2023	Urine	Hospital	Internal medicine
72.	1030822	ERS24931745	2023	Urine	Hospital	Intensive care unit
73.	1030823	ERS24931746	2023	Urine	Ambulatory	Consultative
74.	1030824	ERS24931747	2023	Urine	Hospital	Internal medicine
75.	1030825	ERS24931748	2023	Blood	Hospital	Intensive care unit
76.	1030826	ERS24931749	2023	Blood	Hospital	Internal medicine
77.	1030827	ERS24931750	2023	Blood	Hospital	Intensive care unit
78.	1030828	ERS24931751	2023	Blood	Hospital	Intensive care unit
79.	1030829	ERS24931752	2023	Blood	Hospital	Intensive care unit
80.	1030830	ERS24931753	2023	Blood	Hospital	Intensive care unit
81.	1030831	ERS24931754	2023	Blood	Hospital	Intensive care unit
82.	1030832	ERS24931755	2023	Urine	Hospital	Internal medicine
83.	1030833	ERS24931756	2023	Urine	Hospital	Urology
84.	1030834	ERS24931757	2023	Urine	Hospital	Internal medicine
85.	1030835	ERS24931758	2023	Blood	Hospital	Intensive care unit
86.	1030836	ERS24931759	2023	Blood	Hospital	Intensive care unit
87.	1030837	ERS24931760	2023	Blood	Hospital	Intensive care unit
88.	1030838	ERS24931761	2023	Blood	Hospital	Intensive care unit
89.	1030839	ERS24931762	2023	Blood	Hospital	Intensive care unit
90.	1030840	ERS24931763	2023	Blood	Hospital	Intensive care unit
91.	1030841	ERS24931764	2023	Blood	Hospital	Intensive care unit
92.	1030842	ERS24931765	2023	Blood	Hospital	Intensive care unit
93.	1030844	ERS24931767	2023	Blood	Hospital	Intensive care unit
94.	1030845	ERS24931768	2023	Blood	Hospital	Intensive care unit
95.	1030846	ERS24931769	2023	Blood	Hospital	Intensive care unit
96.	1030850	ERS24931773	2023	Urine	Hospital	Gynecology
97.	1030851	ERS24931774	2023	Urine	Hospital	Intensive care unit
98.	1030852	ERS24931775	2023	Urine	Hospital	Intensive care unit
99.	1030853	ERS24931776	2023	Blood	Hospital	Intensive care unit

ANNEX 2. Command-line protocols for meta-pangenome reconstruction and analysis

Listing A2.1. Batch Prokka annotation for *Klebsiella* genomes (recursive over .fna files)

```
#!/bin/bash
INPUT_DIR="/home/…/pangenome_project/klebsiella_pangenome/data_k
lebsiella"
OUTPUT_DIR="/home/…/pangenome_project/klebsiella_pangenome/klebs
iella_prokka_results"

mkdir -p "$OUTPUT_DIR"

# Recursively find all .fna files
find "$INPUT_DIR" -type f -name "*.fna" | while read i; do
  base=$(basename "$i" .fna)

  prokka --kingdom Bacteria --genus Klebsiella --prefix "$base" \
         --locustag "$base" --outdir "$OUTPUT_DIR/$base" \
         --cpus 0 --force "$i"

  if [ $? -eq 0 ]; then
    echo "Prokka annotation completed for: $base"
  else
    echo "Prokka failed for: $base"
  fi
done

echo "All Prokka annotations finished!"
```

Listing A2.2. Batch eggNOG-mapper function annotation for all .faa files

```
#!/bin/bash
# Define paths
DB_PATH="/home/…/pangenome_project/klebsiella_pangenome/eggnog-
mapper/data"
INPUT_FOLDER="/home/…/pangenome_project/klebsiella_pangenome/kle
bsiella_prokka_results"
OUTPUT_DIR="klebsiella_mapping_results"

# Set CPU usage to use all available CPUs
NUM_CPUS=0  # Use all CPUs
MP_METHOD="spawn"  # Recommended for multiprocessing compatibility

# Create the output directory if it doesn't exist
mkdir -p "$OUTPUT_DIR"

# Find all .faa files in subdirectories and process them
find "$INPUT_FOLDER" -type f -name "*.faa" | while read -r
faa_file; do
    echo "Processing: $faa_file"
    # Extract the base name of the file (without extension and
path)
    BASENAME=$(basename "$faa_file" .faa)

    # Define the output file name
    OUTPUT_FILE="${OUTPUT_DIR}/${BASENAME}_eggnog"

    # Run eggNOG-mapper (emapper.py)
    emapper.py  --data_dir  "$DB_PATH"  -i  "$faa_file"  --output
"$OUTPUT_FILE" --cpu "$NUM_CPUS" --mp_start_method "$MP_METHOD"

    echo "Finished processing: $faa_file"
done

echo "All FAA files have been mapped!"
```

Listing A2.3. Batch Panaroo pangenome build from Prokka GFFs

```
#!/bin/bash
# Force maximum system resources based on provided specs
THREADS=32          # Force using all 32 CPU cores
MEMORY_LIMIT=126    # Force using 126GB of RAM

# Define Prokka output and Panaroo output directories
PROKKA_OUTPUT_DIR="/home/…/pangenome_project/klebsiella_pangenom
e/klebsiella_prokka_results"
PANAROO_OUTPUT_DIR="/home/…/pangenome_project/klebsiella_pangeno
me/klebsiella_panaroo_results"

# Create Panaroo output directory if it doesn't exist
mkdir -p "$PANAROO_OUTPUT_DIR"

# Print resource usage
echo "   Running Panaroo with FORCED FULL RESOURCES:"
echo "   - CPUs: $THREADS threads (MAX)"
echo "   - Memory: ${MEMORY_LIMIT}GB (MAX)"
echo "   - Codon Table: $CODON_TABLE"
echo "   - Input: $PROKKA_OUTPUT_DIR"
echo "   - Output: $PANAROO_OUTPUT_DIR"

# Run Panaroo with forced resource allocation
panaroo -i "$PROKKA_OUTPUT_DIR"/*/*.gff -o "$PANAROO_OUTPUT_DIR"
\
        --clean-mode strict -a core --aligner mafft \
        --threads "$THREADS"

# Check if Panaroo ran successfully
if [ $? -eq 0 ]; then
  echo "Panaroo pangenome analysis completed using ALL available
resources!"
else
```

```
  echo "Panaroo failed!"
  exit 1
fi

echo "Panaroo analysis finished successfully!"
```

Listing A2.4. Phylogenetic tree based on core genome alignment

```
# bash shell
# use data from panaroo outputs
#!/bin/bash
# === Configuration ===
INPUT_DIR="/home/…/pangenome_project/klebsiella_pangenome/klebsi
ella_panaroo_results"
OUTPUT_DIR="/home/…/pangenome_project/klebsiella_pangenome/phylo
geny_results"
INPUT_ALN="${INPUT_DIR}/core_gene_alignment.aln"
OUTPUT_PREFIX="${OUTPUT_DIR}/gubbins"

# === Create output directory if it doesn't exist ===
mkdir -p "$OUTPUT_DIR"

# === Run Gubbins using all available cores ===
run_gubbins --prefix "$OUTPUT_PREFIX" \
               --threads "$(nproc)" \
               "$INPUT_ALN"

# bash shell

# extract SNPs from the alignment using spn-sites
$   snp-sites   -c   gubbins.filtered_polymorphic_sites.fasta   >
clean.core.aln
# reconstruct the phylogenetic tree with 1000 of bootstrap
$ iqtree2 -s clean.core.aln -B 1000 -nt AUTO --prefix tree_clean_ST
```

Listing A2.5. Identification of Antimicrobial-Resistance Genes (ARGs) with ABricate

```
#!/bin/bash
# === Detect script's own directory ===
SCRIPT_DIR="$(cd "$(dirname "${BASH_SOURCE[0]}")" && pwd)"

# === Set input directory with genome folders ===
BASE_DIR="/home/…/pangenome_project/klebsiella_pangenome/data_kl
ebsiella"

# === Abricate configuration ===
DB="resfinder"
MINID=90
MINCOV=90

echo "Output will be saved to: $SCRIPT_DIR"
```

```
# === Loop through each subfolder ===
for folder in "$BASE_DIR"/*; do
  if [ -d "$folder" ]; then
    # Find the first .fna file
    fna_file=$(find "$folder" -maxdepth 1 -name "*.fna" | head -n
1)

    if [ -z "$fna_file" ]; then
      echo "No .fna file found in $folder"
      continue
    fi

    # Get folder name to use as sample ID
    sample_name=$(basename "$folder")

    # Set output file in the script's directory
    out_file="$SCRIPT_DIR/${sample_name}_resfinder.tab"

    echo "Processing $sample_name"
    abricate --db "$DB" --minid "$MINID" --mincov "$MINCOV"
"$fna_file" > "$out_file"
    echo "Result saved to: $out_file"
  fi
done

echo "All outputs saved in: $SCRIPT_DIR"

# summarizing results
abricate --summary *_resfinder.tab > resfinder_summary.tab
```

Listing A2.6. Identification of Virulence Factors (VFs) with ABricate

```
#!/bin/bash
# ========== Configuration ==========
INPUT_DIR="/home/…/pangenome_project/klebsiella_pangenome/data_k
lebsiella"
DB="vfdb"
MINID=90
MINCOV=90
REPORT_NAME="result_virulence.tab"
# ===================================

# Get path where this script resides
SCRIPT_DIR="$(cd "$(dirname "${BASH_SOURCE[0]}")" && pwd)"

echo "Starting virulence gene screening with Abricate using
database: $DB"
echo "Minimum identity: $MINID%, Minimum coverage: $MINCOV%"
echo "Scanning genome folders in: $INPUT_DIR"
echo "Output directory: $SCRIPT_DIR"
```

```
for folder in "$INPUT_DIR"/*; do
  if [ -d "$folder" ]; then
    fna_file=$(find "$folder" -maxdepth 1 -type f -name "*.fna" |
head -n 1)

    if [ -z "$fna_file" ]; then
      echo "No .fna file found in: $folder"
      continue
    fi

    sample_name=$(basename "$folder")
    output_file="$SCRIPT_DIR/${sample_name}_$REPORT_NAME"

    echo "Screening $fna_file"
    abricate  --db  "$DB"  --minid  "$MINID"  --mincov  "$MINCOV"
"$fna_file" > "$output_file"
    echo "Output saved to: $output_file"
  fi
done

echo "Virulence factor screening complete."

# summarizing results
abricate --summary *_result_virulence.tab > virulence_summary.tab
```

Listing A2.7. Identification of viral sequences with VirSorter2

```
#!/bin/bash
# ==========================================
# VirSorter2 Multi-FASTA Batch Runner
# ==========================================
# Directory that contains subfolders with FASTA files
INPUT_ROOT="/home/…/pangenome_project/klebsiella_pangenome/data_
klebsiella"
OUTPUT_ROOT="/home/…/pangenome_project/klebsiella_pangenome/virs
orter_results"
# Use all CPU cores
THREADS=$(nproc)

# VirSorter2 parameters
MIN_SCORE=0.8
MIN_LENGTH=10000
INCLUDE_GROUPS="all"

# Create output directory if it doesn't exist
mkdir -p "$OUTPUT_ROOT"

# Recursively find all .fasta files in subdirectories
find "$INPUT_ROOT" -type f -name "*.fna" | while read fna; do
  # Get unique name using relative path, replacing "/" with "__"
  rel_path=$(realpath --relative-to="$INPUT_ROOT" "$fna")
  tag=${rel_path//\//__}
  tag=${tag%.fna}
  outdir="$OUTPUT_ROOT/${tag}_vir"
  echo "Processing: $fna"
  echo "Output:   $outdir"

  virsorter run \
    -w "$outdir" \
    -i "$fna" \
    --include-groups "$INCLUDE_GROUPS" \
    -j "$THREADS" \
    --min-score "$MIN_SCORE" \
    --min-length "$MIN_LENGTH"

  if [ $? -eq 0 ]; then
    echo "Completed: $tag"
  else
    echo "Failed: $tag"
  fi
done

echo "All VirSorter2 jobs finished."
```

ANNEX 3. PGGL and PGGS algorithms

Algorithm A3.1. Pangenome Gene Gain-Loss (PGGL) Algorithm

Input:

- rooted, strictly bifurcating tree T with tip order $L = (l_1, \dots, l_n)$, internal nodes I, branch lengths $\ell_e > 0$, and edge $E = \{(u \to v)\}$.
- Binary presence-absence matrix $P \in \{0,1\}^{n \times G}$ (rows align to L, columns are genes).
- Target state s*∈ $\{0,1\}$ whose gains/losses are detected (default s*=1)
- Threshold $\tau > 0$ for declaring events from probability changes (default $\tau = 0.05$)

Output:

- Long table R with one row per gene g and edge $(u \to v)$, containing parent/child Ids and labels, $gain_{g,e} \in \{0,1\}$, $loss_{g,e} \in \{0,1\}$, and fitted rates (λ_g, μ_g)

```
PGGL(T, P, model="ARD", threshold=τ, state=s*):
1. Align rows:
   Ensure rows of P match the tip order L of T.

2. Precompute:
   n ← number of tips;  m ← number of internal nodes
   all_nodes ← (labels for tips L) ‖ (string IDs for internal nodes)
   R ← empty table

3. For each gene g = 1..G:
   3.1.  x ← P[:, g]
        If x is invariant (all 0 or all 1, ignoring missing): continue

   3.2.  Inlined CTMC gain-loss fitter with node posteriors
        # States and parameters
        S = {0,1}; nl = 2
        Parameterization:
           if model=ER  : p=(p1),   q01=p1,  q10=p1,   k=1
           if model=ARD : p=(p1,p2),q01=p1,  q10=p2,   k=2

        BUILD_Q(p):
           Q ← [[-q01, q01],
                [ q10,-q10]]

        Ptrans(Q,t) = exp(Q·t)         # eigen or generic matrix exponential

        Dev(p):                        # -2 log-likelihood via pruning with scaling
           if any p_r < 0 or non-finite: return +∞
           Q ← BUILD_Q(p)
           Initialize conditional likelihoods L_u ∈ ℝ^2:
             for each tip u with observed x_u:
                L_u = [1,0] if x_u=0,  L_u=[0,1] if x_u=1
             for missing: L_u=[1,1]
           logS ← 0
           Postorder over internal node a with children c1,c2 and lengths t1,t2:
              v1 ← Ptrans(Q,t1)  · L_{c1}
              v2 ← Ptrans(Q,t2)  · L_{c2}
              v  ← v1 ⊙ v2
              s  ← Σ_s v[s]
              L_a ← v / s
              logS ← logS + log(s)
           Choose root prior π (stationary from Q or uniform [½,½])
           logL ← logS + log( Σ_s π_s · L_root[s] )
           return −2·logL

        # Optimize
        p0 ← vector of 0.1 (length k)
        pˆ ← argmin_p Dev(p)  subject to p ≥ 0
        if Dev(pˆ) is non-finite: continue  # skip gene on fitting failure
        # Rates for reporting
        if model=ARD: (λ, μ) ← (p^1, p^2)
        else         : (λ, μ) ← (p^, p^)

        # Internal-node posteriors for state s* (upward values already in L_u)
        # Downward pass to get node marginals (standard upward-downward):
        # (store upward L_u from final Dev pass)
        Q^ ← BUILD_Q(p^)
        For each edge a→c with length t (top-down order):
           P ← Ptrans(Q^, t)
           # parent marginal M_a ∝ (π at root or previously computed at a)
           # combine sibling info implicitly via upward L_sibling:
           # compute child marginal M_c ∝ (M_a · P) ⊙ (upward of c)
           Normalize M_c to sum 1
        After pass: obtain posterior π_u(s) for all internal nodes u.

   3.3. Build node probability vector for target state s*
        prob ∈ ℝ^{n+m}
        For internal nodes u: prob[u] ← π_u(s*)
        For tips t:           prob[t] ← 1 if x_t = s* else 0   (or 0.5 if missing)

   3.4. Edge-wise gain / loss calls by probability change
        For each edge e=(u→v) in E:
           Δ_e ← prob[v] − prob[u]
           gain_e ← 1 if Δ_e >  +τ else 0
           loss_e ← 1 if Δ_e <  −τ else 0
           Append row to results dataframe:
             (gene=g,
              parent=u, child=v,
              parent_label=all_nodes[u], child_label=all_nodes[v],
              gain=gain_e, loss=loss_e,
              lambda=λ, mu=μ)

    4. Return R table.
```

Algorithm 3.2. PGGS (Pangenome Gene Selection) method for estimating selection pressure in pangenome genes.

```
Input: rooted tree T with tip order L; binary matrix $X \in \{0,1\}^{n\times G}$
Output: table R with per-gene $\lambda$, $\mu$, selections index, selection score, AICs, ΔAIC,
selection class
ALGORITHM: PGGS(T, X, min_freq, max_freq):
1. Align rows to tips:
   Ensure rows of X match the tip order L of T.
2. Filter informative, mid-frequency genes:
   K ← { g ∈ {1..G} : variance(X[:,g]) > 0  ∧
                      min_freq < (1/n)·Σ_i X[i,g] < max_freq }.
3. Initialize empty result table R.
4. For each gene g ∈ K do:
   4.1  y ← X[:,g] ∈ {0,1}^n.
   4.2  Define common CTMC machinery (used twice, ER and ARD):
        (a) State space S={0,1}, nl=2.
        (b) Given parameter vector p, build Q(p):
              if model=ER:  p=(p1);  q01=q10=p1
              if model=ARD: p=(p1,p2); q01=p1, q10=p2
              Q ← [[-q01, q01],
                   [ q10,-q10]].
        (c) Transition matrix P(Q,t) ← exp(Q·t)     # via eigendecomposition or
generic matrix exponential.
        (d) Pruning deviance Dev(p) = −2·logLik(p):
              i.   Initialize conditional likelihoods L_u ∈ ℝ^2 at each node u:
                   if u is tip with observed y_u: L_u = [1,0] if y_u=0 else [0,1];
                   if missing: L_u = [1,1].
```

```
          ii.  Postorder over internal node a with children c1,c2 and branch lengths t1,t2:
               v1 ← P(Q(p), t1)  · L_{c1}
               v2 ← P(Q(p), t2)  · L_{c2}
               v  ← v1 ⊙ v2
               s  ← Σ_s v[s]
               L_a ← v / s
               accumulate logS ← logS + log(s)
          iii. Choose root prior π (stationary: solve πQ(p)=0, Σπ=1; or uniform [½,½]).
          iv.  logLik(p) ← logS + log( Σ_s π_s · L_root[s] ).
          v.   Return −2·logLik(p).

      (e) Constrained optimization and Hessian SEs:
          - Choose initial $p_0$ (all 0.1 for each free rate), bounds p ≥ 0.
          - pˆ ← argmin_p Dev(p)  subject to p ≥ 0  (generic bounded optimizer).
          - If Dev(pˆ) is non-finite: mark fit as failed.
          - logL ← −½·Dev(pˆ).
          - H ← numerical Hessian of Dev at pˆ.
          - If H invertible and well-conditioned:
               SE ← sqrt(diag(H^{-1}));
            else:
               SE ← (NaN,…,NaN).

   4.3  Fit ER model:
        - model ← ER;  k_ER ← 1
        - Run steps 4.2(b–e) → obtain (pˆ_ER, logER, SE_ER) or failure.

   4.4  Fit ARD model:
        - model ← ARD; k_ARD ← 2
        - Run steps 4.2(b–e) → obtain (pˆ_ARD, logARD, SE_ARD) or failure.

   4.5  If either fit failed:
        - Append NA row for gene g to R and continue to next g.

   4.6  Read ARD rates:
        - λ ← pˆ_ARD[1]  (q01)
        - μ ← pˆ_ARD[2]  (q10)

   4.7  Information criteria and selection summaries:
        - AIC_ER  ← −2·logER  + 2·k_ER
        - AIC_ARD ← −2·logARD + 2·k_ARD
        - ΔAIC    ← AIC_ER − AIC_ARD
        - sel_index ← log(λ/μ)
        - sel_score ← (λ − μ)/(λ + μ)

   4.8  Class label:
        - sel_class ←
            NS if ΔAIC ≤ 2;
            WS if 2 < ΔAIC ≤ 4;
            MS if 4 < ΔAIC ≤ 10;
            SS if ΔAIC > 10.
   4.9  Append row to R table:
        (gene=g, λ, μ, sel_index, sel_score,
         logLik_ER=logER, logLik_ARD=logARD,
         AIC_ER, AIC_ARD, ΔAIC, sel_class).
5. Return R table.
```

ANNEX 4. SUPPLEMENTARY FIGURES

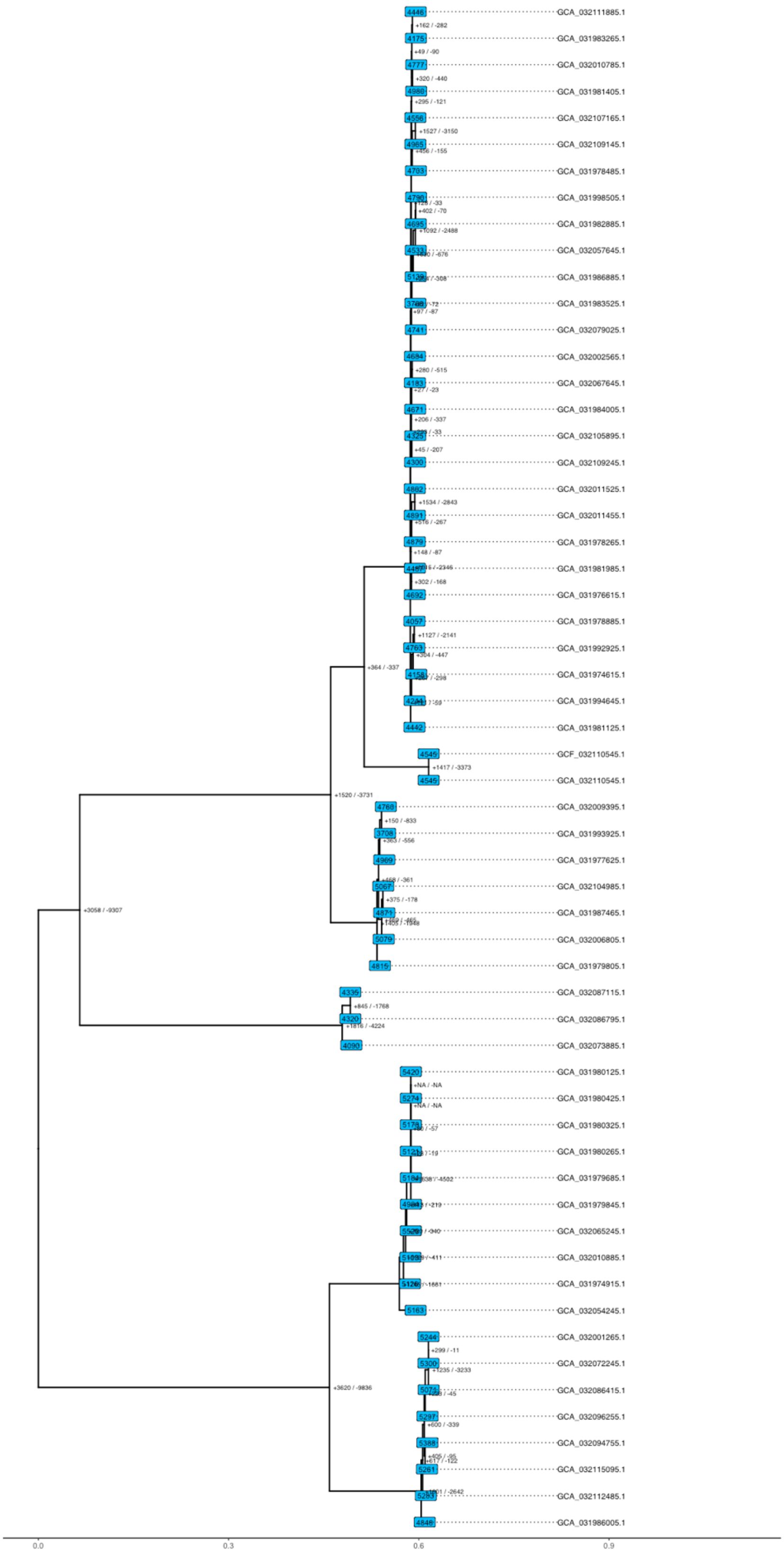


Figure A4.1. Gene gain-loss counts mapped on phylogenetic tree of MPKG dataset (meta-pangenome).

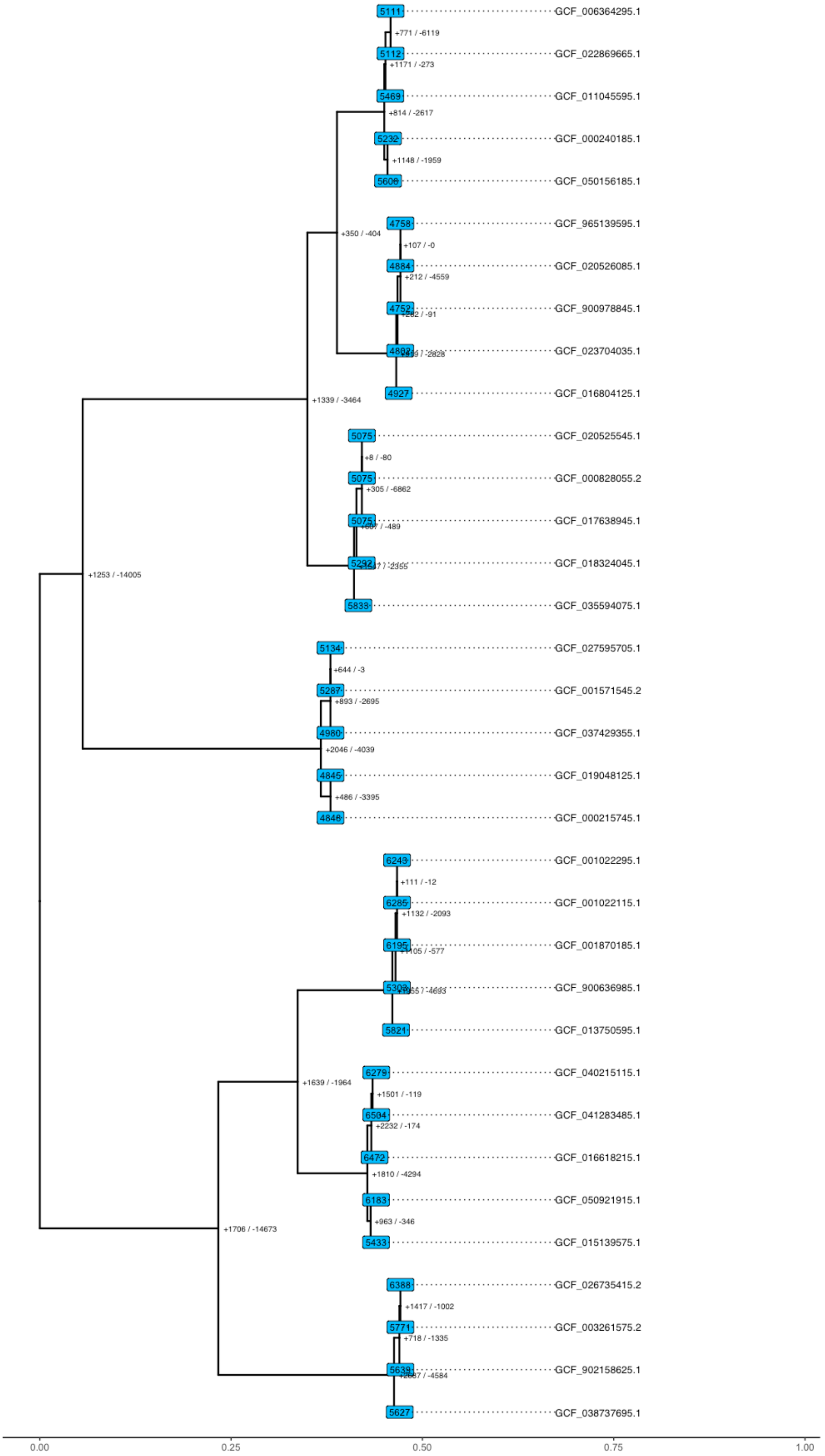


Figure A4.2. Gene gain-loss counts mapped on phylogenetic tree of PKG dataset (pangenome).

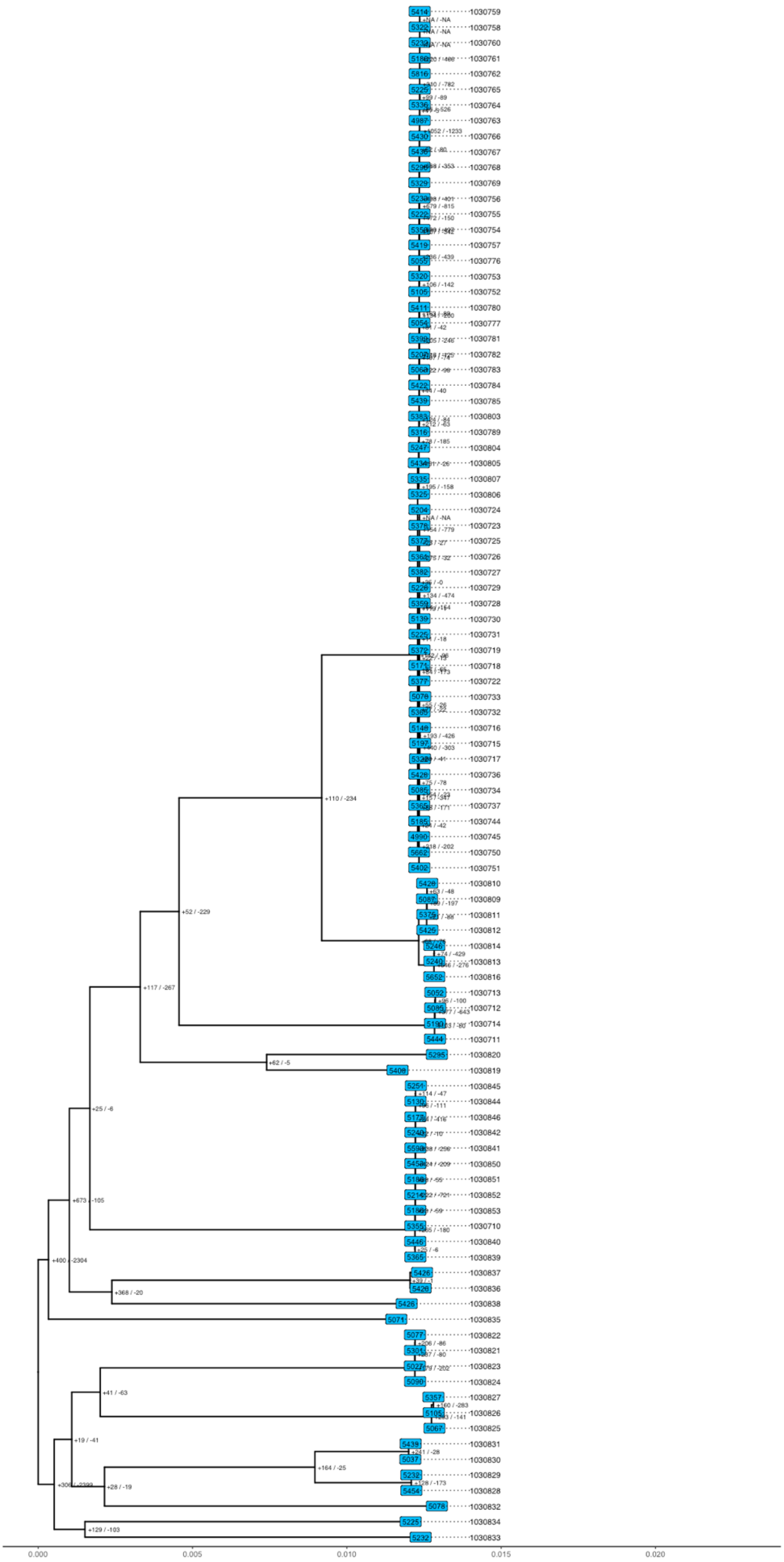


Figure A4.3. Gene gain-loss counts mapped on phylogenetic tree of PKP dataset (pangenome).

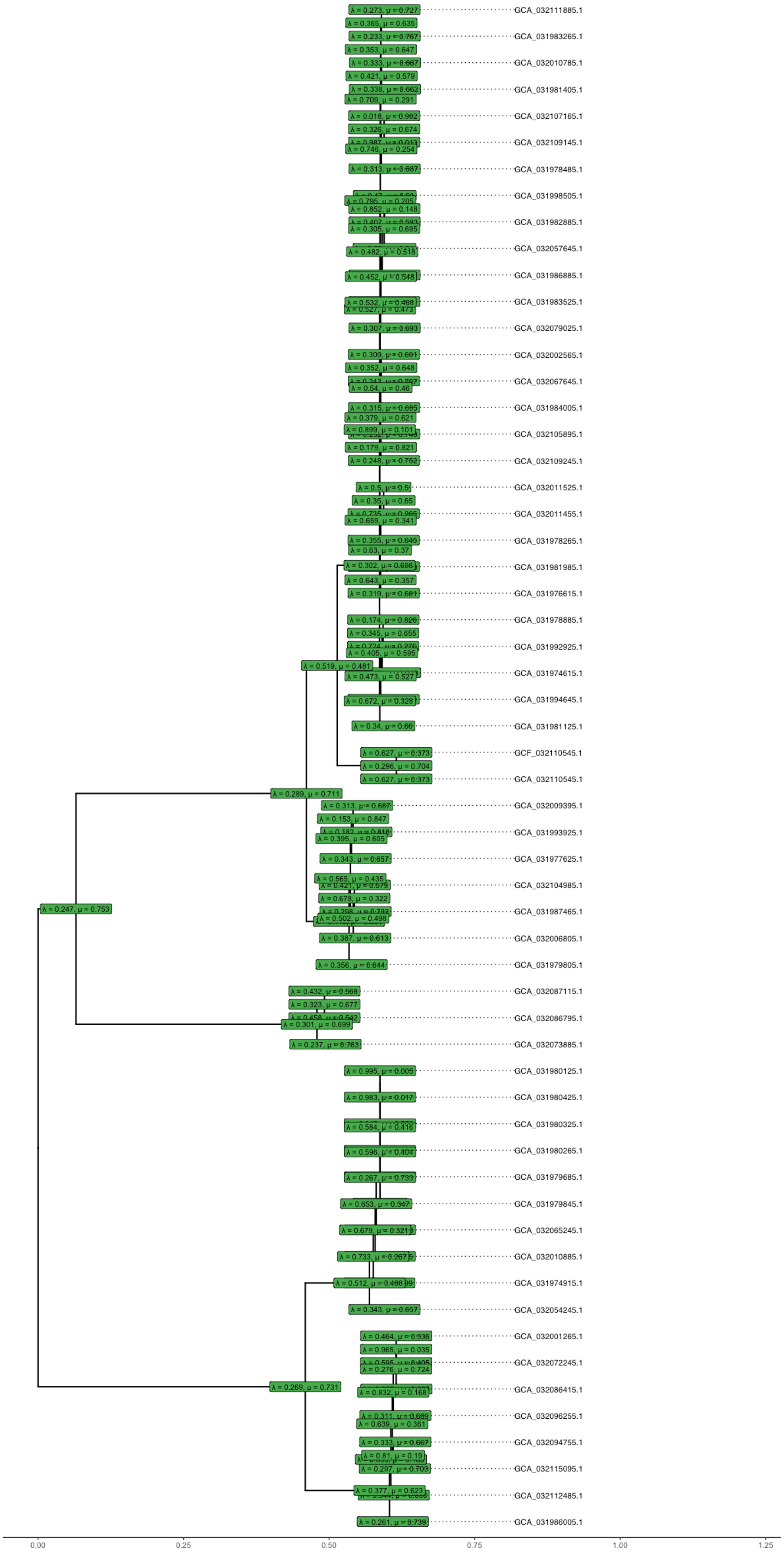


Figure A4.4. Gene gain-loss rates in MPKG dataset (meta-pangenome) mapped on phylogenetic tree.

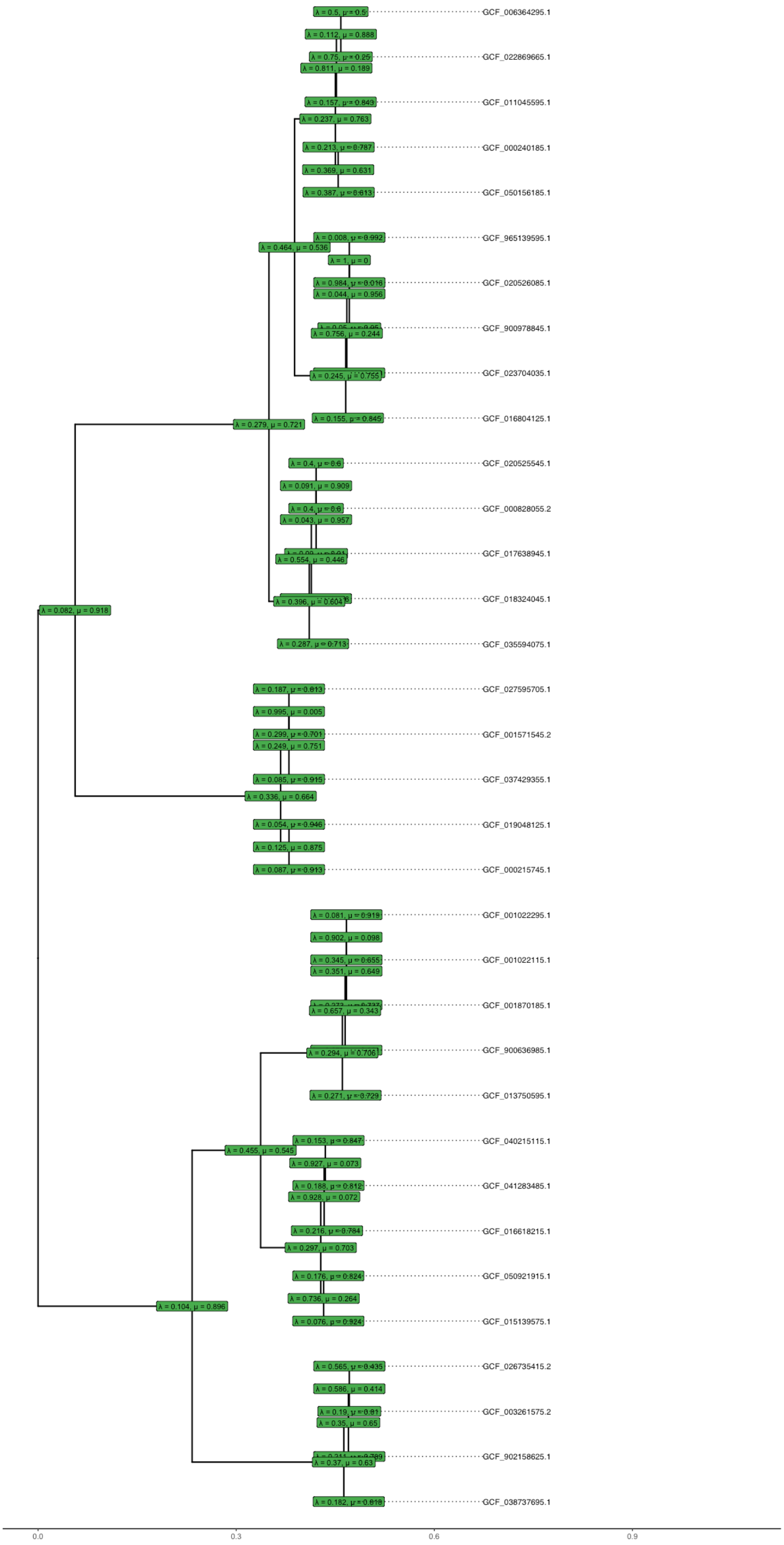


Figure A4.5. Gene gain-loss rates in PKG dataset (pangenome) mapped on phylogenetic tree.

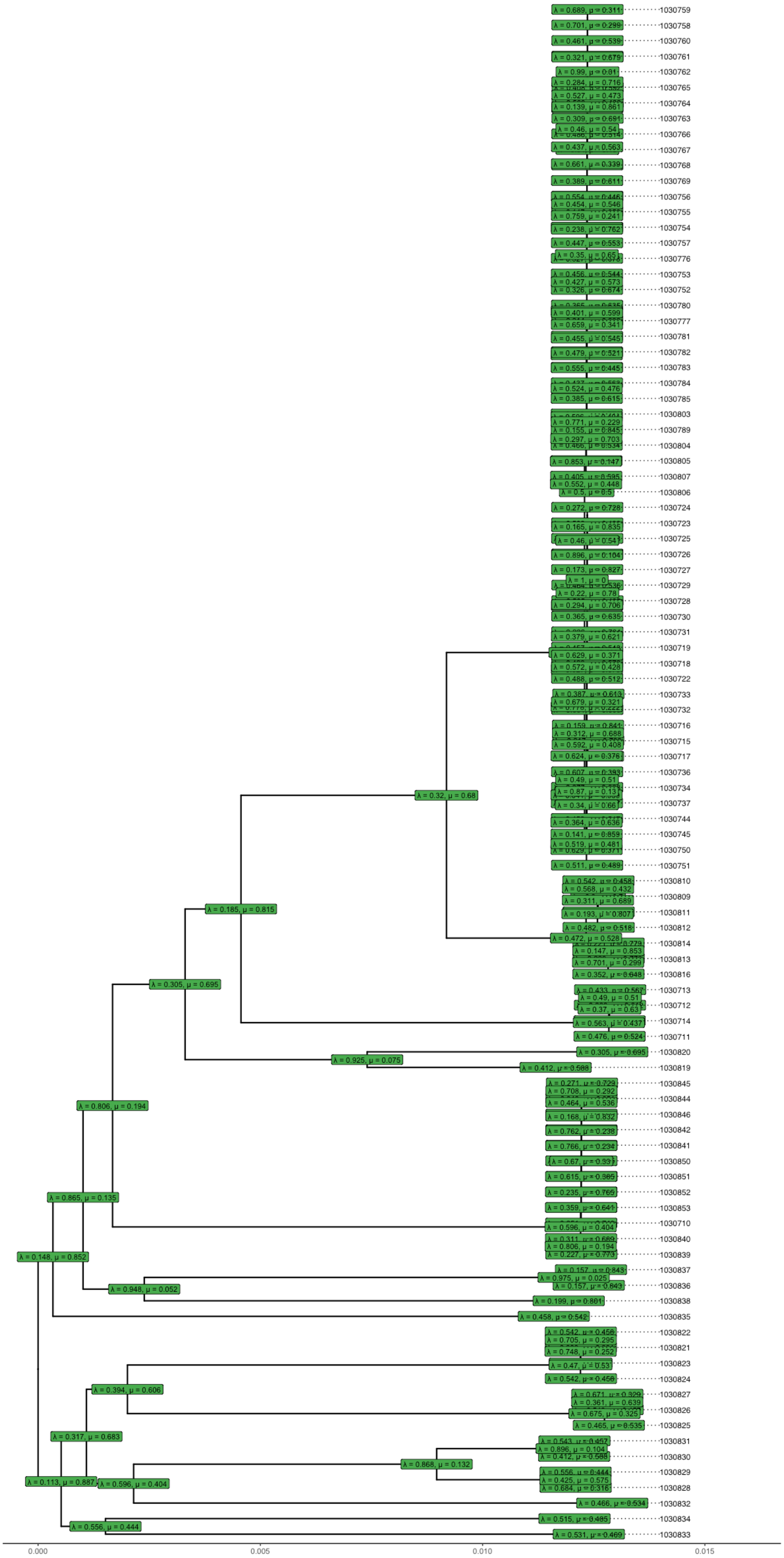


Figure A4.6. Gene gain-loss rates in PKP dataset (pangenome) mapped on phylogenetic tree

ACKNOWLEDGEMENTS

First and foremost, I am profoundly grateful to my advisors, Prof. **Viorel Bostan** and Prof. **Serghei Mangul**, for their masterful mentorship, unwavering support, and invaluable guidance throughout my PhD journey. Their expertise and encouragement have been instrumental in shaping both my research and my growth as a scientist.

I extend my heartfelt thanks to Prof. **Dumitru Ciorbă**, Dean of the Department of Computers, Informatics and Microelectronics, Technical University of Moldova (TUM), and Prof. **Rodica Siminiuc,** Head of TUM Doctoral and Postdoctoral School Studies, for their consistent support in advancing my academic aspirations. I am equally thankful to my friends and colleagues **Nicolae D.**, **Victor G.** and **Eugeniu C.** from the Department and the Laboratory of Bioinformatics at TUM, whose collaboration, encouragement, and access to resources have been instrumental in advancing my research and contributing to my personal growth.

Special gratitude goes to **Elena Postolache** and the late **Vasile Roman**, my dedicated schoolteachers, who planted and nurtured the seeds of curiosity and ambition in me during the early stages of my academic and career journey. Their belief in my potential has been a constant source of inspiration.

To my parents, **Valentina** and **Vasile**, and my sister, **Viorica**, I owe everything. Their unwavering support, constant encouragement, and sacrifices have been my foundation, giving me the strength to pursue my dreams and overcome challenges along the way.

Finally, I am eternally indebted to my family, **Dana** and **Raluca**. Their unconditional love, patience, and encouragement have been my steadfast anchor throughout this journey. They have nurtured my curiosity and wonder for the world, undoubtedly playing a vital role in my decision to pursue this path.

To all who have supported and believed in me, thank you from the bottom of my heart. This achievement is as much yours as it is mine.

Viorel Munteanu,
Chișinău, 2026

Disclaimer of liability

I, the undersigned, hereby declare on my own responsibility that the materials presented in the doctoral thesis are the results of my own research and scientific work. I acknowledge that, otherwise, I will bear the consequences in accordance with the legislation in force.

Viorel Munteanu

Signature

Date

CURRICULUM VITAE

VIOREL MUNTEANU

January 28, 2026

Bioinformatics Researcher, Team Lead
at Bioinformatics Laboratory from
Technical University of Moldova (TUM),
9/7 Studenţilor str.,
Study building Nr. 3, office 210,
MD-2045, Chișinău, Moldova
P: +373 603 387 71
E: viorel.munteanu@lt.utm.md
O: https://orcid.org/0000-0002-4133-5945

RESEARCH INTERESTS

Bioinformatics, Computational biology, Comparative & Evolutionary Genomics, Data Science, Machine Learning

EDUCATION & TRAININGS

Ph.D. in Computer Science, Technical University of Moldova (TUM), Moldova Advisor: Prof. Viorel Bostan (Technical University of Moldova, MD); Co-advisor: Prof. Serghei Mangul (Univeristy of Southern California, LA, US)	Sep 2022 – Feb 2026
Advanced training program in Bioinformatics and Computational Medicine University of Southern California, School of Pharmacy, Mangul Lab, Los Angeles, CA, US	Nov 2021 - Apr 2022
Research to service: Planning and running a bioinformatics core facility European Molecular Biology Organisation (EMBO), Heidelberg, Germany	Oct 2021
M.S. in Microbial Genomics, University of Lausanne (UNIL), Lausanne, Switzerland Advisor: Maria Péchy-Tarr (Swiss Government Excellence Scholarships)	Sep 2007 - May 2009
B.S. in Genetics, Moldova State University (MSU), Chisinau, Moldova	Sep 2003 - Jun 2007

SELECTED HONORS AND AWARDS

- **Diploma of Honor**, Ministry of Education and Research of Republic of Moldova for outstanding contributions to bioinformatics and genomics research, December 2025

- **Government Excellence Scholarship** (doctoral level), Ministry of Education and Research of the Republic of Moldova, 2024 – 2025

- **ASHG** abstract reviewers' choice award (top 10% of submitted abstracts), 2025

- **Swiss Government Excellence Scholarship** (Master degree), 2007-2009

PROFESSIONAL APPOINTMENTS

- **Bioinformatician/Research Data Analyst**, **University of Suceava**, Suceava, Romania — 2023 - present

- **Bioinformatics Researcher, Team lead at Bioinformatics Laboratory**, **Technical University of Moldova**; Chișinău, Moldova — 2020 – present

- **Junior Lecturer, Technical University of Moldova**; Chișinău, Moldova — 2020 – present

- **Bioinformatician/Research Data Analyst, University of Southern California (Mangul Lab)**; Los Angeles, CA, US — 2022 - 2025

- **Scientific Researcher, Oncological Institute of Republic of Moldova**; Chisinau, Moldova - Jan 2015 - Jun 2019

- **Scientific Researcher, Shemyakin-Ovechnikov Institute of Bioorganic Chemistry**, Russian Academy of Sciences; Moscow, Russia - Jan 2011 - Dec 2011

- **Scientific Researcher, Molecular Biology Center, University of Academy of Sciences of Moldova** ; Chisinau, Moldova - Sep 2009 - Dec 2014

CONTRIBUTION TO EDUCATION

Courses and workshops designed and taught at Technical University of Moldova:

- BPC, 4 units, *Basic Python Programming*, Technical University of Moldova
- SD, 6 units, *Data Science*, Technical University of Moldova
- BI, 4 units, *Bioinformatics and Computational Genomics*, Technical University of Moldova
- Member of *East European Bioinformatics and Computational Genomics School (EEBG)* Organizing Committee, https://sites.google.com/view/eebgschool2023/home, https://eebg2025.edu.pl/history/

CURRENT RESEARCH PROJECTS

- **Romania–Republic of Moldova Complex Bilateral Projects,** Technical University of Moldova – Ștefan cel Mare University of Suceava, 2025–2027. *Urban Pathogen Genomic Surveillance for Public and Environmental Health Protection: A One Health Framework (UPGRADE)*. **Funding agency:**

CCCDI – UEFISCDI, Project No. PN-IV-PCB-RO-MD-2024-0555, within PNCDI IV. **Total funding:** €120,000. **Role:** Bioinformatics Group Lead

- **Romania–Republic of Moldova Complex Bilateral Projects,** Technical University of Moldova – Floreasca Emergency Clinical Hospital, Bucharest, Romania), 2025–2027. *Collaborative Genomic Research on Genetic Variations for Cardiovascular Health in Eastern Europe (CardioGen).* **Funding agency:** CCCDI – UEFISCDI, Project No. PN-IV-PCB-RO-MD-2024-0303, within PNCDI IV. **Total funding:** €120,000. **Role:** Bioinformatics Group Lead

- **National Agency for Research and Development (Republic of Moldova),** Technical University of Moldova, 2024–2027**.** *Innovations in Biomedical Engineering: Advanced Technologies and Applications for Data Acquisition, Processing and Analysis (BIOTECH)***. Total funding:** €410,000. **Role:** Bioinformatics Group Lead

- **Romania's National Recovery and Resilience Plan (PNRR),** Ștefan cel Mare University of Suceava, 2023–2026**.** *Artificial Intelligence–Powered Personalized Health and Genomics Libraries for Analysis of Long-Term Effects in COVID-19 Patients (AI-PHGL-COVID)***. Total funding:** €4.0 million. **Role:** Bioinformatics Group Lead

- **Romania's National Recovery and Resilience Plan (PNRR),** Ștefan cel Mare University of Suceava, 2023–2026**.** *Metagenomics and Bioinformatics Tools for Wastewater-Based Genomic Surveillance of Viral Pathogens for Early Prediction of Public Health Risks (MetBio-WGSP)***. Total funding:** €3.0 million. **Role:** Bioinformatics Group Lead

CONTRIBUTION TO RESEARCH

Bibliometric indexes:

Citations >**290**; h-index **7**; i10-index **5**

Journal Articles:

1. Munteanu, V., Saldana, M.A., Dreifuss, D. *et al.* SARS-CoV-2 wastewater genomic surveillance: approaches, challenges, and opportunities. ***Genome Biol*** **27**, 1 (2026). https://doi.org/10.1186/s13059-025-03927-6
2. Liu, S., Rodriguez, J.S., Munteanu, V (joint first author), *et al.* Analysis of metagenomic data. ***Nat Rev Methods Primers*** **5**, 5 (2025). https://doi.org/10.1038/s43586-024-00376-6
3. Huang, Y. N., Munteanu, V. (joint first author), Love, M. I., Ronkowski, C. F., Deshpande, D., Wong-Beringer, A., *et al*. Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies. ***Cell Genomics***, *5*(5) (2025). 10.1016/j.xgen.2025.100845
4. Aßmann, E., Greiner, T., Richard, H., Munteanu V., *et al.* Augmentation of wastewater-based epidemiology with machine learning to support global health surveillance. ***Nat Water*** **3**, 753–763 (2025). https://doi.org/10.1038/s44221-025-00444-5
5. Gordeev, V., Hölzer, M., Desirò, D., Munteanu V., *et al.* Leveraging wastewater sequencing to strengthen global public health surveillance. ***BMC Glob. Public Health*** **3**, 23 (2025). https://doi.org/10.1186/s44263-025-00138-w

6. Huang, YN., Jaiswal, P.V., Rajes, A., Munteanu V., *et al.* The systematic assessment of completeness of public metadata accompanying omics studies in the Gene Expression Omnibus data repository. ***Genome Biol*** **26**, 274 (2025). https://doi.org/10.1186/s13059-025-03725-0
7. Sharma NK, Ayyala R, Deshpande D, Patel Y, Munteanu V, Ciorba D, Bostan V, Fiscutean A, Vahed M, Sarkar A, Guo R, Moore A, Darci-Maher N, Nogoy N, Abedalthagafi M, Mangul S. 2024. Analytical code sharing practices in biomedical research. ***PeerJ Computer Science*** 10:e2066 https://doi.org/10.7717/peerj-cs.2066
8. Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, Muszyńska A, Munteanu V, Yang H, Rotman J, Tao L, Balliu B, Tseng E, Eskin E, Zhao F, Mohammadi P, P. Łabaj P and Mangul S (2023) RNA-seq data science: From raw data to effective interpretation. ***Front. Genet.*** 14:997383. doi: 10.3389/fgene.2023.997383

Perr-reviewed conference and workhops articles:

1. Munteanu, V. *et al.* (2025). The Pangenome Variability Index: A Quantitative Measure for Assessing Gene Content Diversity in Microbial Genomes. In: Sontea, V., Tiginyanu, I., Railean, S. (eds) 7th International Conference on Nanotechnologies and Biomedical Engineering. ICNBME 2025. ***IFMBE Proceedings***, vol 135. Springer, Cham. https://doi.org/10.1007/978-3-032-06497-4_26
2. Ciubara, R., Odajiu, O., Munteanu, V., Arnaut, O., Belîi, A., Iapăscurtă, V. (2025). Multi-agent Decision Support for Sepsis: Balancing Precision and Hallucination Risks in Biomedical Engineering. In: Sontea, V., Tiginyanu, I., Railean, S. (eds) 7th International Conference on Nanotechnologies and Biomedical Engineering. ICNBME 2025. ***IFMBE Proceedings***, vol 135. Springer, Cham. https://doi.org/10.1007/978-3-032-06497-4_63
3. Iapăscurtă, V., Falenciuc, R., Munteanu, V., Arnaut, O. (2025). Advancing Biomedical Engineering: An Agent-Based Approach to Pulmonary Edema Simulation. In: Sontea, V., Tiginyanu, I., Railean, S. (eds) 7th International Conference on Nanotechnologies and Biomedical Engineering. ICNBME 2025. ***IFMBE Proceedings***, vol 135. Springer, Cham. https://doi.org/10.1007/978-3-032-06497-4_7
4. Iapăscurtă, V., Munteanu, V., Belîi, A. (2025). Exploring Maternal-Placental-Fetal Interactions: A Hybrid Modeling Approach for Biomedical Engineering. In: Sontea, V., Tiginyanu, I., Railean, S. (eds) 7th International Conference on Nanotechnologies and Biomedical Engineering. ICNBME 2025. ***IFMBE Proceedings***, vol 135. Springer, Cham. https://doi.org/10.1007/978-3-032-06497-4_30
5. Sharma N K, Chhugani K, Munteanu V,, Skums P, Zelikovsky A, Mangul S, Realistic assortment of novel metagenomics benchmarks with diverse biological and technological characteristics, ***Biopolymers & Cell*** 2024; 40(3):219-219. doi: 10.7124/bc.000AFF
6. Boldirev G, Sharma N K, Munteanu V, Bhavatharini A, Koslicki D, Zelikovsky A, Mangul S, Assessing microbial genome representation across various reference databases: A comprehensive evaluation ***Biopolymers & Cell*** 2024; 40(3):220-220. doi: 10.7124/bc.000AFD

Non-Perr-reviewed preprints:

1. Oleksyk, T. K., Yakymenko, D., Bożek, S., Munteanu, V., Pilch, W., Comarova, Z., ... & Mangul, S. (2026). Leveraging a hybrid cross-disciplinary training model to accelerate global bioinformatics capacity. ***bioRxiv***, 2026-01. doi: https://doi.org/10.64898/2026.01.21.700760
2. Oleksyk, T. K., Wolfsberger, W. W., Chhugani, K., Huang, Y. N., Pokrytiuk, V., Shchubelka, K., Munteanu, V., ... & Mangul, S. (2025). Challenges and Recommendations in Establishing National Human Diversity Genomic Projects. ***ArXiv,*** preprint arXiv:2510.19869. https://doi.org/10.48550/arXiv.2510.19869
3. Sharma, G., Munteanu, V., Ghiasi, N. M., Banerjee, J., Varma, S., Foschini, L., ... & Mangul, S. (2025). A decentralized future for the open-science databases. ***ArXiv***, arXiv-2509. https://doi.org/10.48550/arXiv.2509.19206

4. Sharma, S., İlgün, E., Okçu, T., Ostash, V., Bashynska, V., Alkan, C., Munteanu, V., ... & Mangul, S. (2025). Robust software development practices improve citations of RNA-seq tools. ***bioRxiv***, 2025-09. https://doi.org/10.1101/2025.09.05.674580
5. Sarwal, V., Munteanu, V., Suhodolschi, T., Ciorba, D., Eskin, E., Wang, W., & Mangul, S. (2023). Biollmbench: A comprehensive benchmarking of large language models in bioinformatics. ***bioRxiv***, 2023-12. https://doi.org/10.1101/2023.12.19.572483
6. Munteanu, V., Gordeev, V., Saldana, M., Aßmann, E., Su, J. M., Drabcinski, N., ... & Mangul, S. (2023). A rigorous benchmarking of methods for SARS-CoV-2 lineage abundance estimation in wastewater. ***ArXiv*** *preprint arXiv:2309.16994*. https://doi.org/10.48550/arXiv.2309.16994

MEMBERSHIPS

- International Society for Computational Biology (ISCB). Membership ID38741
- Romanian Society of Bioinformatics (RSBI)

SKILLS

Languages: R, Python, Unix shell scripts, Jupyter Notebook, Git, C++
Operating Systems: Unix/Linux/Ubuntu, Mac OS, Windows
Teaching Resources: Github Classroom, Google Classroom

LIST OF SCIENTIFIC AND METHODOLOGICAL-DIDACTIC PUBLICATIONS RELATED TO THE THESIS TOPIC

By Mr. **MUNTEANU Viorel**, full-time PhD student, specialization: 122.03 Modeling, mathematical methods, software products

LIST OF PUBLICATIONS RELATED TO THE THESIS TOPIC

Articles published in international journals indexed in 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 proceedings of scientific events 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 proceedings 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 proceedings of scientific events included in the Register of materials published based on scientific events organized 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