Wageningen University and Research (WUR), December 12, 2025

Assessing pangenome structure influence on gene co-occurrence estimation approaches in bacterial pangenomes

T. Sardjoe [1]

[1] *Bioinformatics Group, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands*

Supervisors: Dr. A. Kupczok [1], Prof. Dr. Ir. D. de Ridder [1]

Keywords: *Pangenome structure, Gene co-occurrence, Bacterial pangenomes, Genomic fluidity, Pangenomics*

Abstract

Understanding how pangenome structure affects the detection of gene co-occurrence is essential for interpreting functional relationships between bacterial genes. This study evaluates the influence of openness and genomic fluidity, together called pangenome structure, on three widely used co-occurrence estimation approaches: Coinfinder, Goldfinder, and PanForest. Using pangenomic data from 114 bacterial species, pangenomes were classified into Closed, Moderate, and Open structural categories based on calculated openness and fluidity metrics. Representative species from each class were analysed using a real pangenomic dataset and two pseudo-simulated datasets designed to provide controlled co-occurrence ground truths. These simulations introduced perfect and near-perfect duplicated gene patterns across the D-value distribution to assess tool performance. Across real pangenomes, Goldfinder identified the largest number of gene pairs, of which many involved genes with low lineage independence. Agreement between tools was limited, particularly at the gene-pair level. In the pseudo-simulated datasets, Goldfinder achieved the highest recall, Coinfinder the highest precision, and PanForest performed poorly overall. All methods showed reduced performance in open pangenomes, reflecting the increased noise associated with large, variable accessory genomes. D-value analyses further revealed that Coinfinder and PanForest preferentially detect high D-value genes, whereas Goldfinder uniquely recovers low D-value genes. These findings reveal that pangenome structure strongly shapes the reliability of gene co-occurrence detection, with closed pangenomes yielding more reliable signals. We also exposed knowledge gaps, including the lack of standardised validation frameworks, the limitations of binary openness classifications, and the uncertain generalisability of current machine-learning approaches. This highlights the need for more rigorous benchmarking, continuous openness metrics, and more powerful models to capture complex gene-gene dependencies.

1. Introduction

At the turn of the century, Perna et al. (2001) [1] were the first to publish a comparative study of single-species bacterial genomes, showing a disparity in gene content between two strains of *Escherichia coli*. Soon thereafter Welch et al. (2002) [2] performed a three-way comparison on the two aforementioned *E. coli* strains and one other pathogenic *E. coli* genome, showing a less than 40% overlap of protein coding sequences between the three strains. At the time, this was an unexpected outcome. However, the coining of the pangenome concept by Tettelin et al. (2005) [3] has since provided an explanation for the observed results and facilitated a paradigm shift in the field of (micro)biology [4, 5, 6].

The *pangenome concept* poses that a single reference genome is inadequate for capturing the genetic diversity of a species, particularly in bacteria, where related isolates can have dramatically differing gene contents. It states that the collective gene set across all isolates of a strain is larger than the genes set found in any single isolate[3, 7]. In pangenomes, genes are subdivided into three categories: the *core genome*, the *accessory genome* and *isolate-specific genes*. Genes present across all isolates of a strain constitute the core genome and are most likely functionally essential, particularly in metabolism and ribosomal activity [8]. Genes present in more than one but not all isolates comprise the accessory genome, the composition of which varies based on several factors, such as lifestyle [9]. Genes that are only found in a single isolate are called isolate-specific genes (also known as strain-specific or unique genes). An important property of pangenomes is *openness*, the degree to which accessory genes are found by increasing the number of sequenced genomes. For *open pangenomes*, the number of accessory genes will increase with every genome added to the pangenome, whereas *closed pangenomes* have a finite number of genomes that can be added before the number of accessory genes no longer increases [7]. Another property of pangenomes is *genomic fluidity*, the measure of dissimilarity between individual genomes of a species at the gene level. A high fluidity means a high turnover of genes and thus a larger accessory genome, whereas low fluidity suggests a more stable gene content and a larger core genome [10]. It has been shown that *horizontal gene transfer* (HGT) is the main driver of gene gain in (open) pangenomes and that the balance of gene gain and loss is what drives pangenome formation [11, 5].

Pangenome analyses are particularly relevant in cases such as investigating multi-drug resistant (MDR) pathogens, as it allows for the elucidation of drivers of the genotypic diversity that shapes the emergence and spread of antibiotic resistance [12, 13]. The emergence of MDR bacterial pathogens has become an increasing threat to the antimicrobial strategies we employ [14]. Given the complex, often strain-specific interactions genes in pangenomes can have with antibiotics [7], it is crucial to increase our understanding of how such genes are shared and conserved across bacterial strains. Genes often fulfil biologically important functions, such as host specificity or antibiotic resistance, together with other genes [15]. When genes consistently appear together in genomes of a species, this is known as *gene co-occurrence*. Gene co-occurrence both drives and is a consequence of the functional relation between genes [15]. Genes can co-occur due to being *co-localised* or due to *co-selection*. Genes that co-localise may be inherited together due to their proximity to one another on a chromosome, not necessarily due to shared function. Co-selected genes are retained together because they confer functional advantages that are selected for simultaneously in both genes, occurring when the presence of one gene affects the selective pressure of another. This is how bacterial strains can spread multiple resistance traits [16], and it has been suggested that selection for certain gene combinations affects the *structure of pangenomes* [17]. Pangenome structure is typically achieved through rarefaction analysis, however, recently a new quantitative measure has been proposed [18]. To discover whether co-occurrence of genes is due to co-selection or simply because they are physically linked (i.e., co-localised) on the genome, several tools have been developed. Phylogeny-aware approaches, such as Coinfinder [19] and Goldfinder [20] allow for the inference of co-occurring genes pairs using phylogenetic context to control or account for shared ancestry. In contrast, machine learning approaches such as PanForest [21] employ random forest models that predict accessory gene presence/absence solely based on presence/absence patterns of all other accessory genes within a genome.

A comparative study on the (identical) pangenome of *S. pneumoniae* between Coinfinder and Goldfinder has shown wildly differing numbers of associated gene pairs [20]. Furthermore, comparative pangenome studies focus on related strains and hardly ever include gene co-occurrence analyses. When gene co-occurrence results are published, only a single pangenome is reconstructed and results are not compared to pangenomes of varying structures (openness/fluidity) [19, 20, 21]. Given that closed pangenomes have a more conserved, stable accessory gene set and open pangenomes have a large and unstable accessory gene set [7], we expect that the gene co-occurrence estimation approaches inherently are more effective in closed pangenomes. This could potentially skew results when comparing closed to open pangenome gene co-occurrences, leading to the following research question: "What is the effect of pangenome structure on gene co-occurrence estimation approaches?". We hypothesise that the effect of pangenome structure is quantifiable and that closed pangenomes will yield more reliable co-occurrence detection, while open pangenomes are clouded by noise in their vast accessory genomes. A comparative study between phylogeny-aware and a machine learning gene co-occurrence approach, including an emphasis on the effect of pangenome structure on the outcome of these methods, has so far not been published. In this study Coinfinder, Goldfinder, and PanForest gene co-occurrence estimation approaches are evaluated on pseudo-simulated and real bacterial pangenomes classified into pangenome structure categories in an attempt to discover what the effect of pangenome structure is on gene co-occurrence estimation approaches.

2. Materials & Methods

Here we present an overview of our approach to quantifying what the effect of pangenome structure is on gene co-occurrence estimation approaches. A dataset was created that consists of bacterial pangenome gene presence/absence (GPA) matrices and phylogenetic strain trees. The pangenome structure characteristics were computed from the GPA matrices. Subsequently, pangenome structure classes were defined to measure their effect on gene co-occurrence estimation results. The data was then subjected to three gene co-occurrence estimation approaches (Coinfinder, Goldfinder and PanForest) and their results were assessed across the different pangenome structure classes. These steps are described in further detail below.

2.1. Data collection

Pangenomic data was obtained from Anja Barth's Master thesis [22]. This dataset consisted of `panX` [23] output of 403 bacterial species genomes collected from the National Center of Biotechnology Information (NCBI). PanX reconstructs pangenomes by clustering the genes in given genomes into orthologous groups, from which the core genome is identified. Using SNPs within the core genome, panX builds a strain-level phylogeny. For each gene cluster, multiple sequence alignments are generated to construct individual gene trees. PanX maps the presence/absence of gene clusters onto the strain-level phylogeny, derived from the core genome. `PhyloThin` [24] was used to correct for oversampling bias, which happens when certain groups or variations are overrepresented in the genomic data and can lead to inaccurate or non-representative findings. Subsequently, supplementary data from Dewar et al. (2024) [25] was collected. This dataset contained lifestyle traits (host-associated or free-living, host reliance, host location, motility and effect on host), as well as genome characteristics (genome size and effective population size) for 126 bacterial species. This dataset also included pangenome fluidity estimates.

2.2. Data preparation

The intersection of the bacterial species in the two datasets left 114 species. For each species a binary GPA matrix, a phylogenetic strain tree, lifestyle traits and other metadata were now available in one centralized dataset (see appendix I). The R package micropan (`v2.1`) [26] was used to calculate the pangenome openness (Heaps' law approach, as suggested by Tettelin et al. (2024) [27]), using 500 permutations. The genomic fluidity was recalculated for each pangenome because the results from Dewar et al. (2024) were not corrected for oversampling, a change in fluidity is to be expected when correcting for oversampling.

2.3. Defining pangenome structure classes

To elucidate the effect of pangenome structure on gene co-occurrence estimation approaches, first classes of said "structure" were established. Pangenomes were classified according to their fluidity and openness using k-means clustering. Clusters sizes of k = 3,4, and 5 were evaluated. To maximize the separation between the categories the k with the highest silhouette score [28] was chosen (k = 3), using the scikit-learn `sklearn.metrics v1.7.1` Python library. The three classes that were defined were as follows: closed low-fluidity pangenomes (Closed), moderately open medium fluidity pangenomes (Moderate) and open high fluidity pangenomes (Open). Each pangenome was then assigned to a cluster based on its proximity to a cluster centroid in the

Table 1. Corresponding class, number of taxa, number of genes, genomic fluidity and openness (Heaps' alpha) of the selected pangenomes

Class	Species	# Taxa	# Genes	Fluidity	Openness
Super-closed	*Yersinia pestis*	62	4,310	0.070	2.000
Closed	*Bordetella holmesii*	64	3,105	0.009	1.792
Closed	*Brucella melitensis*	100	3,127	0.021	1.673
Closed	*Corynebacterium pseudotuberculosis*	121	2,302	0.068	1.758
Closed	*Neisseria gonorrhoeae*	122	2,609	0.100	1.345
Closed	*Mannheimia haemolytica*	67	3,622	0.173	1.396
Moderate	*Bordetella pertussis*	385	3,732	0.014	0.747
Moderate	*Helicobacter pylori*	335	3,069	0.135	0.844
Moderate	*Pseudomonas aeruginosa*	289	31,247	0.167	0.536
Moderate	*Campylobacter jejuni*	180	4,130	0.190	0.827
Moderate	*Streptococcus pneumoniae*	177	4,592	0.192	0.831
Open	*Stenotrophomonas maltophilia*	85	15,976	0.259	0.567
Open	*Escherichia coli*	420	39,372	0.278	0.523
Open	*Acinetobacter buamannii*	168	19,579	0.279	0.578
Open	*Enterococcus faecium*	216	9,986	0.283	0.765
Open	*Bacillus cereus*	123	25,810	0.284	0.461

two-dimensional feature space (Figure S1). Five pangenomes were manually selected per class (see Table 1) based on several considerations: their status as model organism (*E. coli, P. aeruginosa*), the ability to compare to previous studies (*S. pneumoniae*), and the number of sequenced genomes in the pangenome. The full selected species table is available in appendix II.

2.4. Fritz and Purvis' D-statistic

The Fritz and Purvis' D-statistic [29], also referred to as *lineage independence*, measures phylogenetic signal in binary traits, such as gene presence/absence. It quantifies whether the distribution of a gene along a phylogeny is more concentrated than expected under random association or more dispersed than expected under a Brownian motion model. D-values $\approx$ 1 indicate a random distribution, while D-values ≈ 0 indicate strong phylogenetic concentration. Negative D-values show very strongly concentrated genes, i.e., genes that are likely to co-occur due to shared ancestry rather than functional association. D-values higher than 1 suggest overdispersion and co-occurrences may reflect ecological adaptations or functional linkage. The D-statistic can thus be used to evaluate whether detected associations primarily reflect phylogenetic inheritance or instead to ecological adaptations or functional linkages, beyond shared ancestry.

2.5. Gene co-occurrence estimation

The three created datasets were subjected to the three different gene co-occurrence estimation approaches. The latest versions at the time of writing were used, namely `Coinfinder` (Release build @ Dec 15 2024 17:04:59) [19], `Goldfinder` (downloaded on 14/10/2025) [30], and `PanForest` (downloaded on 14/10/2025) [21].

Coinfinder evaluates associations (and dissociations) from the observed GPA matrix, testing for associations among genes as they occur in present genomes. It reports significant patterns of gene co-occurrence (or avoidance) by lineage-independent statistical association, indirectly accounting for phylogeny by requiring genes to exceed a D-value cutoff that is to be determined by the user based on their data (it should be noted that for Coinfinder there is no standard approach for selecting this D-value cutoff). The cutoff removes low D-value genes (i.e., genes most likely inherited by vertical descent) from the analysis. For the remaining genes, each pair is subjected to a binomial significance test to evaluate the null-hypothesis that their observed frequency of co-occurrence does not differ from random expectations. Multiple testing correction is then applied (Bonferroni [31] by default), retaining only the significant gene pairs. Non-significant results remove only the tested pair (X,Y), while other pairs involving X or Y are evaluated independently. These significant gene pairs are saved in files and serve as the output, as well as a file that contains computed D-values for all genes that constitute a gene pair. From the significant associations, Coinfinder also generates a gene co-occurrence network, in which genes are the nodes and significant associations are the edges. This network can be used to visualise and highlight gene clusters of genes that tend to associate or dissociate.

Coinfinder was run using the `-a`, `-n`, and `-E` flags to look for genes that tend to associate/co-occur, without multiple testing correction, whilst outputting all results, regardless of significance, respectively. Following the successful processing of all pangenomes and to bring the results in line with the other tools, Benjamini-Hochberg (BH) false discovery rate (FDR) correction [32] was applied. A Python script was written to apply BH multiple testing correction to the associating genes and gene pairs. Only genes and gene pairs below the threshold of $\alpha = 0.05$ (BH FDR corrected) were kept for subsequent analyses.

Goldfinder reconstructs ancestral gene histories and tests for co-gains and co-losses along the phylogeny, thereby defining associations in terms of repeated evolutionary events rather than contemporary co-occurrence. To establish null expectations, Goldfinder simulates random gene histories under different root state assumptions (present at root, absent at root, or unknown root state). These distributions record the frequency of Fitch scores, which represent the minimum number of gains and losses required to explain a gene's presence/absence pattern; low scores indicate few gene gain/loss events, high scores reflect many such events. By explicitly simulating gene histories on the strain tree, Goldfinder provides a phylogenetically grounded null distribution of expected co-gain and co-loss frequencies, removing the need for a lineage-independence cutoff such as a D-value cutoff. Observed gene pairs are evaluated against the this null model and an empirical p-value (BH FDR corrected) is determined for each gene pair, based on the proportion of simulated pair scores exceeding the observed gene pair score. Statistically significant associations (or dissociations) are reported. The output includes a filtered set of gene-gene associations, optionally outputting a gene network, which can be visualised and further explored.

Goldfinder was run using the `-c both` flag to look for both associating and dissociating genes (required to generate gene network Cytoscape output) and `-g 50000`, simulating 50,000 genes for the estimation of the null distribution. All other parameters were left as their default.

PanForest is fundamentally different from Coinfinder and Goldfinder in its treatment of gene associations. PanForest frames co-occurrence as a predictive relationship and uses random forest machine learning on gene presence-absence data. It attempts to predict the presence or absence of a gene given all other genes in the pangenome. Phylogeny is partially accounted for using D-value filtering, where the input GPA is filtered to only retain genes with a D-value greater than zero such that non-informative features that associate with low D-value genes are excluded from the random forest models. The output of the PanForest pipeline is an asymmetric n by n Gini importance score matrix (also known as Mean Decrease Gini) [33], where n is the number of genes/gene families in the pangenome after D-value filtering. Gini importance quantifies how well a predictor gene helps distinguish genomes with or without the target gene, averaged over all trees in the random forest. This score serves as a quantitative measure of how strongly the predictor gene influences the prediction of the presence/absence state. Because this matrix is asymmetric, co-occurring gene pairs are reported in both directions. For a fair comparison to the other two methods, the matrix is symmetrised by averaging the upper and lower triangular. A gene pair is then only reported once as `gene_a - gene_b` and not as `gene_b - gene_a` as well. PanForest also outputs precision, recall and F1-score performance metrics [34] based on the presence or absence of a gene in each genome in the test set that the random forest model, generated on the training set, predicted.

PanForest has a three-step process to its workflow. First, the genes in the input GPA matrix with a D-value < 0 were removed. Beavan et al. (2024) use this approach, but do not provide the necessary code, so a custom Python script was developed. Then, a Python script provided by PanForest that processes the GPA matrix, was run. This script removes genes present in more than 99% or less than 1% of genomes and additionally collapses genomes with identical patterns of gene presence and absence into the same vector, called a family group. This step is necessary here as random forest models need informative features and balanced, non-redundant samples to work effectively. Removing the core genes and ultra-rare genes and collapsing identical genomes into family groups ensures that only the most informative non-redundant features remain. The third step is to run the random forest analysis itself. The flags that were used were `-n 1000` and `-d 16`, meaning 1000 decision trees, with a maximum depth of 16 nodes per tree, as per Beavan et al. (2024) [21]. Furthermore, the `-pres 1` and `-abs 1` flags were used to set the minimum percentage of genomes featuring a gene and the minimum percentage of genomes missing a gene to 1%, respectively.

Note that a gene pair is defined as two genes that have passed all filtering/cutoffs specific to each tool. The D-value of a gene pair is the minimum D-value of either gene in the association.

2.6. D-value cutoff

As described previously, no standard approach for determining a D-value exists for Coinfinder. Therefore the D-values Coinfinder calculated for every gene were used to create a D-value distribution of all genes for all pangenomes, which would aid in determining the D-value cutoff for Coinfinder. To determine this D-value cutoff, two approaches were tried.

First, an elbow method was used to determine the optimal cutoff point, as suggested in supplementary figure 3 of Whelan et al. 2021 [17]. Using an elbow cutoff provides a principled approach that can easily be replicated and finds the balance between excluding vertically inherited genes and retaining as many true co-occurring genes as possible. An R script was written that implemented an "elbow finder" function that takes candidate cutoff values and the number of associated pair counts, computes the differences between successive cutoff values to find the index of the steepest drop, draws a reference line between the steepest drop and the final cutoff point in the series and finally, finds the cutoff point with the largest perpendicular distance to the reference line. This cutoff point represent the elbow method determined cutoff value. This proved to work well for pangenomes above a certain number of genes (> 100), producing cutoff values near where the curve begins to flatten. However, if fewer genes were present, the elbow was calculated on too few data points and resulted in spurious cutoffs.

Therefore, a second method was a simplified, replicable approach that takes the sorted D-values of all genes and takes the third quartile (top 25%) as the cutoff point, meaning all genes below this cutoff (75%) are discarded. To verify that this cutoff is suitable, plots of gene pairs retained as a function of D-value cutoff were generated. For most pangenomes it was found that the third quartile cutoff aligned well with the distribution and the previously defined elbow-cutoff (Figure S2). It also produced more robust results on pangenomes with fewer genes. The simplified third quartile approach, being less computationally intensive and more replicable, was thus used to generate the results in this work. Because the D-value cutoff is pangenome-specific, a separate D-value was calculated independently for every pangenome. This cutoff was imposed on the associating gene pairs reported by Coinfinder only in the corresponding pangenome.

2.7. Controlling the ground truth

Subsequently, two datasets were created that would act as a controlled test sets ("pseudo-simulations"). These datasets provide a ground truth to evaluate whether the tools can detect co-occurring genes across the D-value distribution. The term "pseudo-simulation" is used because the datasets are derived from the real pangenomes but artificially modified by duplicating and shuffling certain genes, rather than being generated entirely in silico. For each previously selected pangenome (Table 1), the D-value distribution of all genes was examined and the 5th to 95th percentile of the distribution was taken to exclude extreme outliers. Then, 10% of these filtered genes were selected with evenly spaced indices between them from the minimum D-value (5th percentile) to maximum D-value (95th percentile). The gene IDs of this selection were matched against the gene IDs of the corresponding GPA matrix. Any gene matches had their entry duplicated in the GPA matrix and suffixed with `_dup`. These duplicated genes were then randomly shuffled into the GPA matrix.

As described previously, PanForest collapses identical gene presence/absence patterns into family groups. This would collapse perfectly duplicated genes into family groups as well, preventing a gene to ever be associated with its duplicate. So, in addition to perfect co-occurrence patterns, near-perfect ones were introduced by repeating the process above for a clean GPA matrix. In addition to duplicating the gene pattern, one genome per duplicate had a single presence or absence flipped to make a near-perfect co-occurrence pattern, preventing the collapse of genes with their duplicates. These datasets are referred to as the perfect co-occurrence pseudo-simulated dataset and the one-flip co-occurrence pseudo-simulated dataset from here on.

2.8. Co-occurrence result analyses

The outputs, gene pairs found to be associated more often than expected by chance, were collected and compared between the three tools, aggregated by pangenome structure category. UpSet plots were created using the ggplot2 (`3.5.2`) [35] extension package ggupset (`v0.4.1`) [36] to visualise the (dis)agreement in unique gene counts and gene pairs between the tools. As gene pairs can be found in multiple species in a single pangenome structure category, only one instance of a gene pair was kept to prevent species prevalence from influencing the results (i.e., a `gene_a` - `gene_b` association can only be reported once for all species in a pangenome structure category. Analysing the data revealed only a handful of these pairs). UpSet plots were chosen specifically because the sizes of intersections can vary wildly between the different tools and a more traditional Venn diagram would thus prove difficult to read and interpret. A global Chi-squared test for equality of proportions was performed to assess whether the proportion of reported genes or gene pairs relative to the total genes or gene pairs differed across the categories. When significant, post-hoc one-sample proportion tests were conducted to assess which categories differed from the overall mean proportion, with p-values adjusted (BH FDR). Furthermore, pairwise Fisher's exact tests were performed to assess whether the gene/gene pair proportions between the categories differed significantly from one another, again with BH FDR multiple testing correction adjustments to the p-values.

Both the perfect and the one-flip co-occurrence pseudo-simulated datasets were subsequently used to evaluate the accuracy of the gene co-occurrence methods. UpSet plots were generated to visualise the relative proportions of duplicated genes found by each tool, across the pangenome structure categories. To evaluate if the performance metrics (precision, recall, and F1-score) differed across the pangenome structure categories, linear mixed-effect models (lme4 `v1.1.37` [37]) were used with category as a fixed effect and method as a random effect. Type III ANOVA tests (Satterthwaite method [38]) from lmerTest `v3.1.3` [39] were used to assess the significance of pangenome structure category effect on the mixed models. For metrics where category effects were found to be significant, post-hoc pairwise comparisons between the categories were made using estimated marginal means (EMMeans `v2.0.0`, [40]), for which the p-values were adjusted using BH FDR. Finally, to investigate the homogeneity of these metrics over D-value, scatter plots were generated to visualise the distribution of the duplicated genes found by each tool across the pangenomes structure categories as a function of D-value, attempting to elucidate the effect of lineage dependency on the performance of the tools in the different pangenomes structure categories.

3. Results

Three separate datasets were used in subsequent analyses. First, a real pangenome dataset containing only empirical pangenomic data. Second, a perfect co-occurrence pseudo-simaulated dataset, created by duplicating 10% of the genes across the 5th to 95th percentile of the D-value distribution. Third, a one-flip co-occurrence pseudo-simulated dataset, generated the same way but with one random presence presence/absence entry flipped per duplicated gene. The pseudo-simulated datasets serve as a benchmark for testing whether duplicated genes can be correctly recovered.

Cutoffs recap

Goldfinder has no need for a lineage-independence cutoff such as a D-value cutoff as it explicitly simulates gene histories on the strain tree, providing a null distribution to which observed gene pairs are evaluated against. Coinfinder relies on a D-value cutoff that must be determined by the user from the data. PanForest has two fixed cutoffs; a D-value filter of greater than zero for genes in the GPA matrix that serves as the input for the random forest analyses and a fixed 0.01 Gini importance score cutoff in the output importance matrix.

Not all three tools produced significantly associated gene pairs for each pangenome, Goldfinder reported significantly association gene pairs for all 16 pangenomes, Coinfinder for 15 pangenomes (not for *Bordetella holmesii*, from the Closed category), and PanForest for 13 pangenomes (not for *Corynebacterium pseudotuberculosis*, from the Closed category, *Campylobacter jejuni*, from the Moderate category, and *Acinetobacter baumannii*, from the Open category). All results reported in this chapter are based on the pangenomes with significantly associated gene pairs.

3.1. Comparative analysis of gene pairs and unique genes across pangenome structure categories

In order to understand the influence of pangenome structure on gene co-occurrence estimation approaches, first the total counts of associating gene pairs and their constituting genes were collected and aggregated by pangenome structure category (Table 2).

Table 2. Total associating gene pair and gene counts aggregated by pangenome structure category in the real pangenome dataset.

Category	Method	Total gene pairs	Total genes
Closed	Coinfinder	3,079	1,403
Closed	Goldfinder	32,151	914
Closed	PanForest	3,182	401
Moderate	Coinfinder	13,832	2,979
Moderate	Goldfinder	120,869	4,300
Moderate	PanForest	2,573	630
Open	Coinfinder	62,272	6,392
Open	Goldfinder	326,018	11,979
Open	PanForest	4,163	865

UpSet plots visualise the co-occurring gene pairs reported by each tool, aggregated by pangenome structure category (Figure 1). Goldfinder consistently reports the largest number of tool-specific gene pairs across all pangenome structure categories, most of

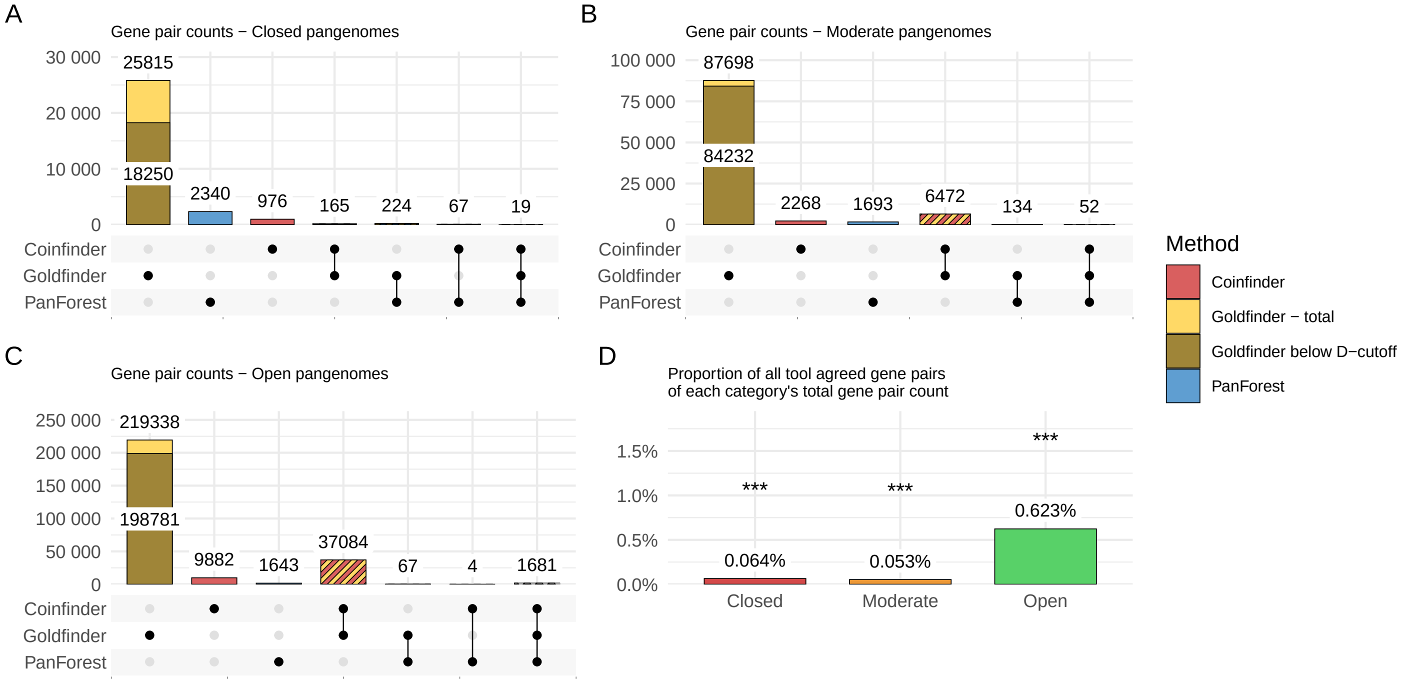

Figure 1. Numbers of gene pairs detected as co-occurring in real pangenomes. A,B,C, Aggregated UpSet plot of co-occurring gene pair counts reported by Coinfinder, Goldfinder, and PanForest for the pangenomes in the (A) Closed, (B) Moderate and (C) Open pangenome structure category respectively. The y-axis indicates gene pair counts and the text label shows the precise gene pair count for a tool/intersection. The combination matrix below shows the intersections between the tools. **D**, Relative proportion of gene pairs reported by all three methods in a pangenome structure category against the total gene pair count of that category. The significance of one-sample proportion tests is indicated above each bar (*** - $p < 0.001$).

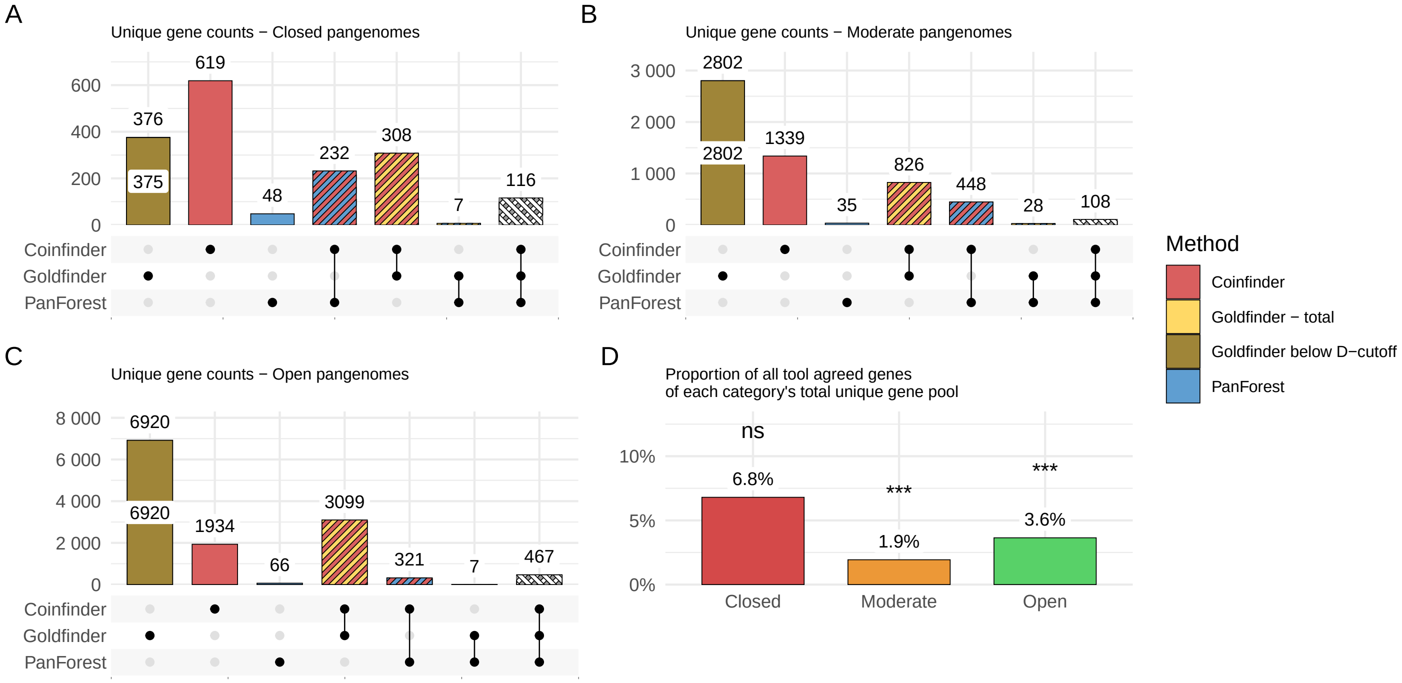

Figure 2. Number of unique genes extracted from the co-occurring gene pairs in real pangenomes. A,B,C, Aggregated UpSet plot of unique gene counts reported by Coinfinder, Goldfinder, and PanForest for the pangenomes in the (A) Closed, (B) Moderate and (C) Open pangenome structure category, respectively. The y-axis indicates unique gene counts and the text label shows the precise unique gene count for a tool/intersection. The combination matrix below the bars shows the intersection of unique genes between the tools. **D**, Relative proportion of unique genes reported by all three methods in a pangenome structure category against the total unique gene count of that category. The significance of one-sample proportion tests is indicated above each bar (ns - not significant, *** - $p < 0.001$).

which fall below Coinfinder's *D*-value cutoff. Agreement between tools is generally low, except for the Coinfinder-Goldfinder intersection. This intersection, apart from the Closed category, reports the highest number of gene-pairs other after Goldfinder-specific pairs. Only a small fraction of gene pairs are shared across all three tools: 19 in the Closed category, 52 in the Moderate category, and 1681 in the Open category. Overall, the results highlight that tool-specific outputs (especially Goldfinder) are more prevalent, and consensus across all tools is rare.

The intersection of agreement between the three tools is visualised as the relative proportion of the total gene pairs present in their respective pangenome structure categories (Figure 1D). The relative proportion of three-tool agreement is very low in the Closed category (0.064%) and the Moderate category (0.053%), but higher in the Open category (0.623%). The asterisks above the bars show that the one-sample proportion test is significant and that all the categories differed from the overall mean proportion. Pairwise Fisher's exact tests show that the Closed and Moderate category three-tool agreements do not differ significantly from one another, but the Closed - Open and Moderate - Open categories do (Table 3). Overall, while the absolute number of reported gene pairs increases from the Closed to the Open category, the fraction agreed upon by all three tools remains small, with only a modest rise in the Open category.

Table 3. Pairwise Fisher's exact test at the gene pair level - difference in the unique gene pair proportions reported by all three methods across the pangenome structure categories.

Category 1	Category 2	*p*-value	*p*-adj	Significance
Closed	Moderate	0.482	0.482	ns
Closed	Open	1.94×10^{-51}	2.91×10^{-51}	***
Moderate	Open	2.49×10^{-157}	7.47×10^{-157}	***

Subsequently, we examined the unique genes involved in the co-occurring gene pairs, defined as the set of distinct genes that appear in at least one reported pair. Because some genes occur in multiple pairs, the set of unique genes is smaller than the total number of genes across all pairs. These genes were aggregated by pangenome structure category and tool, and duplicate genes were removed. The remaining counts were again visualised in UpSet plots (Figure 2). Coinfinder reports the largest number of tool-specific unique genes in the Closed category, while Goldfinder reports the largest number in the Moderate and Open categories, with all but one of its unique genes falling below Coinfinder's *D*-value cutoff. PanForest consistently contributes the smallest set. The Coinfinder-Goldfinder intersection is largest across categories, whereas the Goldfinder-PanForest intersection is minimal. Only a moderate proportion of unique genes are shared by all three tools: 116 in the Closed category, 108 in the Moderate category, and 467 in the Open category.

The agreement between all three tools is visualised as the relative proportion of the total unique genes present in their respective pangenome structure categories (Figure 2D). The relative proportion is highest in the Closed category (6.8%), followed by the Open category (3.6%) and lowest in the Moderate category (1.9%). The asterisks above the bars show that the one-sample proportion test is significant for the Moderate and Open categories, showing that their proportions differed from the overall mean proportion, however, the Closed category did not. Pairwise Fisher's exact tests show that all unique gene proportions differ significantly between the pangenome structure categories (Table 4). Overall, tool-specific genes (apart from PanForest) are generally the largest sets, except for the Open category, where the Goldfinder-Coinfinder intersection outnumbers the Coinfinder-specific counts. However, consensus at the unique gene level is stronger than at the gene-pair level. The relative proportion of unique genes in the Closed category is largest, followed by Open, and lastly by the Moderate pangenome structure category.

Table 4. Pairwise Fisher's exact test at the individual gene level - difference in the unique gene pair proportions reported by all three methods across the pangenome structure categories.

Category 1	Category 2	*p*-value	*p*-adj	Significance
Closed	Moderate	1.17×10^{-20}	3.52×10^{-20}	***
Closed	Open	7.93×10^{-09}	7.93×10^{-09}	***
Moderate	Open	2.13×10^{-10}	3.19×10^{-10}	***

3.2. Evaluation of tool performance across pangenome structure categories in pseudo-simulated data

To evaluate tool performance, we first examined the perfect co-occurrence pseudo-simulated dataset (Figure 3), focusing on the duplicated gene pairs (e.g., `gene_a` and `gene_a_dup`). PanForest is not included here, as its design collapses identical GPA patterns into family groups, preventing assessment of duplicated genes. Total duplicate gene pair recovery rates range from ∼21% in the Closed category to ∼46% in the Moderate category and ∼60% in the Open category. Across all pangenome structure categories, Goldfinder consistently recovered the largest number of duplicated gene pairs. Coinfinder recovered fewer pairs overall, with some overlap and occasional unique recoveries. The Open category shows the strongest intersection between Goldfinder and Coinfinder.

Identifying a proportion of the duplicated gene pairs represents only the true positives. However, for each correctly reported pair, a considerable number of false positives may also be present. Therefore, precision, recall, and their harmonic mean F1-score were calculated (Figure 3D). Coinfinder consistently achieves the highest precision, while Goldfinder showed higher recall. As a result, Goldfinder performs best in the Closed category, whereas Coinfinder performs best in the Moderate category. In the Open category, both tools lose nearly all precision, leading to very low F1-scores overall. Across categories, predictive performance was most strong in the Closed category, followed by the Moderate category and then the Open category. The ANOVA tests to assess whether the performance metrics differed across pangenome structure categories were significant for recall ($p = 0.0013$) and F1-score ($p = 0.0275$) but not for precision ($p = 0.1984$). The post-hoc pairwise comparisons for recall reveal that the only categories that differ significantly from one another are the Closed and Open category (p-adj $= 0.0056$). The post hoc comparisons for F1 were not found to be significant. The full ANOVA and post hoc EMMeans test results can be found in Table S1 and Table S2 respectively.

The one-flip pseudo-simulated dataset differs from the perfect co-occurrence dataset in that duplicated genes do not share identical GPA patterns, allowing PanForest to correctly report duplicated gene pairs (Figure 4). The total recovery of duplicated gene pairs remains generally the same for all pangenome structure categories. Across all pangenome structure categories, Goldfinder consistently recovers the largest proportion of duplicated pairs,

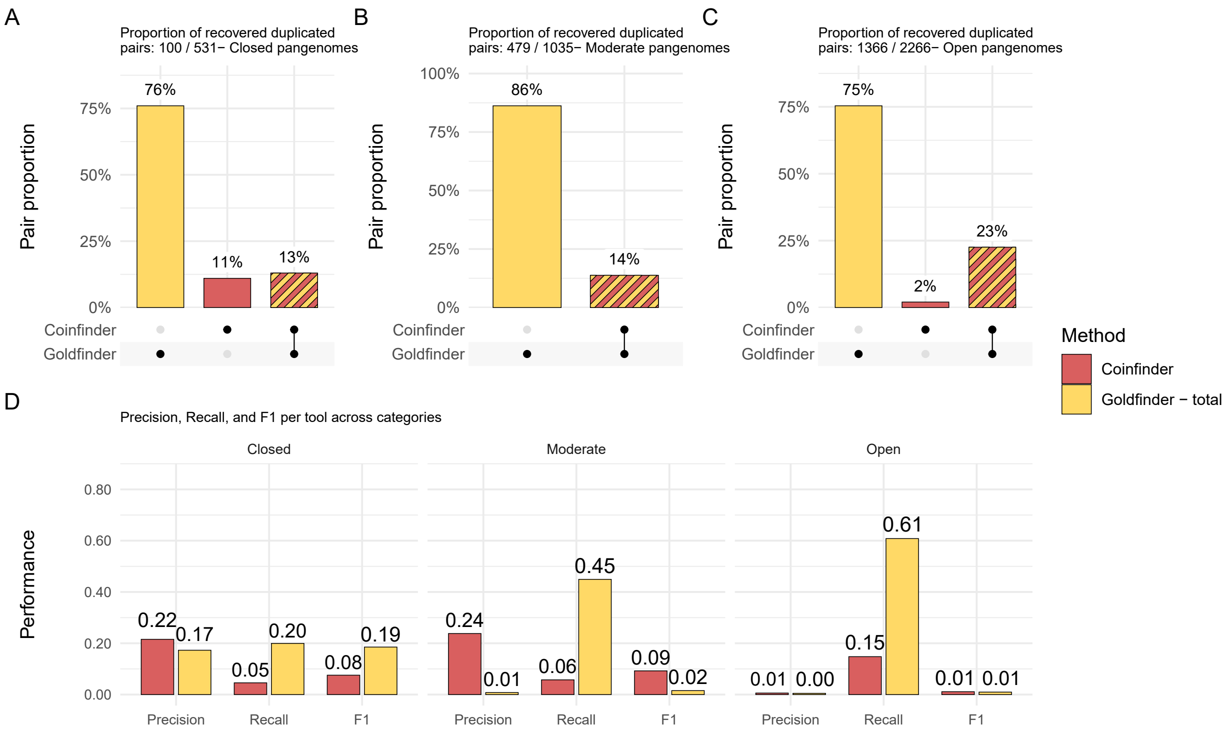

Figure 3. Pseudo-simulated perfect co-occurrence - relative proportion of duplicated gene pairs recovered. A,B,C, UpSet plots of the proportion of correctly reported duplicated gene pairs (original gene with its duplicate in a pair) in the perfect co-occurrence pseudo-simulation dataset. The absolute count of duplicated gene pairs recovered/total in a category is shown at the top of each panel. **D**, Precision, recall and F1-score metrics for each tool per pangenome structure category. Each column corresponds to one of these metrics, a row to the pangenome structure category, and the bars to each one of the specific tools, indicated by colour.

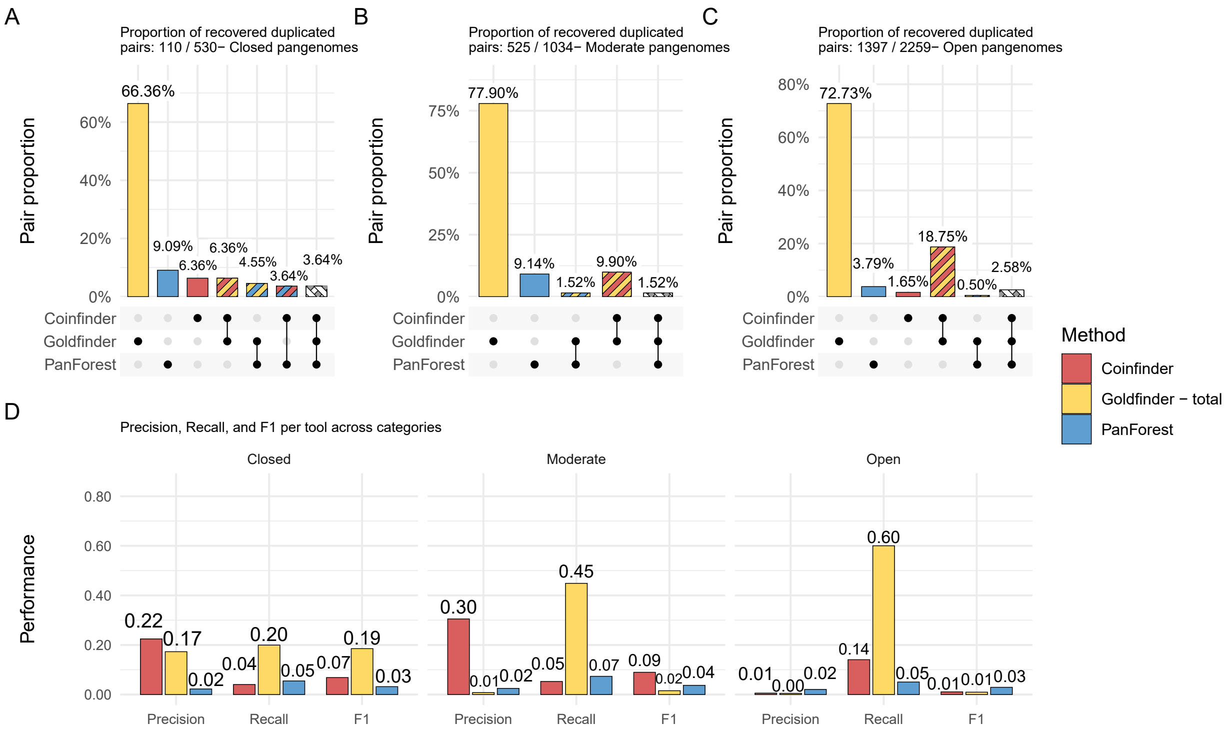

Figure 4. Pseudo-simulated one-flip co-occurrence - relative proportion of duplicated gene pairs recovered. A,B,C, UpSet plots of the proportion of correctly reported duplicated gene pairs (original gene with its duplicate in a pair) in the one-flip co-occurrence pseudo-simulation dataset. The absolute count of duplicated gene pairs recovered/total in a category is shown at the top of each panel. **D**, Precision, recall and F1-score metrics for each tool per pangenome structure category. Each column corresponds to one of these metrics, a row to the pangenome structure category, and the bars to each one of the specific tools, indicated by colour.

followed by PanForest and then Coinfinder. Coinfinder does not recover pairs that are not reported by the other tools in the Moderate category. Intersections show that Coinfinder-Goldfinder overlap is the largest, while no gene pairs are shared between Coinfinder and PanForest in the Moderate and Open categories. The Coinfinder-Goldfinder intersection in the Open category is by far the largest. Three-tool agreement is consistently low, ranging from 1.52% to 3.64%.

Like before, precision, recall, and their harmonic mean F1-score were calculated (Figure 4D). Coinfinder and Goldfinder show similar performance patterns to the perfect co-occurrence pseudo-simulated dataset, while PanForest consistently has low precision and recall, and shows little variation between pangenome structure categories. As a results, PanForest's F1-scores are low but marginally higher than Goldfinder and Coinfinder. The ANOVA tests are only significant for recall ($p = 0.0028$). The post-hoc pairwise comparisons of recall were found to be significant only between the Closed and Open category (p-adj $= 0.0059$). The full ANOVA and post hoc EMMeans test results can be found in Table S3 and Table S4 respectively. Overall, Goldfinder performs best in the Closed category, Coinfinder in the Moderate category, while all three tools performs poorly in the Open category, with PanForest only performing marginally better.

3.3. Influence of D-value on gene recovery across pangenome structure categories in pseudo-simulated data

To investigate the influence of D-value on the performance of the tools across the pangenome structure categories, the D-value distribution of the duplicated genes was examined. From the D-value distribution of the pseudo-simulated perfect co-occurrence dataset (Figure 5), a few things become evident, namely that Goldfinder is able to find the duplicated genes for low D-value, below the D-value cutoff that is imposed by Coinfinder. Coinfinder, on the other hand, is better at finding genes with high D-values, which is reflected by the upwards section at the end of each distribution . The D-value cutoff also slightly shifts upwards going from the Closed to the Open pangenome structure category.

This was repeated for the one-flip co-occurrence dataset (Figure 6). Here, a similar pattern emerges, where Goldfinder is able to find low D-value genes and struggles with high D-value genes. Coinfinder shows a similar, if not more extreme pattern as well, able to find high D-value genes in all three of the pangenome structure categories. PanForest is unable to find duplicated genes below a D-value around 0 to 0.5. However, like Coinfinder, it is able to detect genes in higher D-value ranges than Goldfinder but this effect seems to diminish the more open a pangenome gets.

4. Discussion

In this work, the effect of pangenome structure on three gene co-occurrence estimation approaches (Coinfinder, Goldfinder and PanForest) was assessed through a combination of a real pangenome dataset and pseudo-simulated datasets. The analyses showed that relative proportions of co-occurring gene pairs and unique genes, as well as performance metrics, differ across the pangenome structure categories. The results also showed that the tools are able to recover genes in differing D-value ranges. In the following discussion, we highlight the strengths and weaknesses of our methods, make a critical comparison with existing literature, and implications for future research directions and recommendations.

4.1. Strengths of this study

Biological relevance of co-occurrence patterns

The co-occurring gene pairs reported by each tool, and intersections, reveal an association between pangenome structure and gene pair counts. The Open pangenome structure category shows the highest proportion of co-occurring gene pairs that are agreed upon by all tools, which seems contradictory to biological expectations of pangenomes and our hypothesis that the gene co-occurrence estimation approaches inherently are more effective in closed pangenomes. However, collapsing gene pairs into their unique genes produced the opposite pattern: closed pangenomes had the largest proportion of associated genes, while open pangenomes had the smallest proportion. These differences between proportions proved to be statistically significant, demonstrating that pangenome structure has an influence on the proportion of gene co-occurrence: closed pangenomes report a higher proportion of co-occurrence than open pangenomes.

Tool performance across pangenome structure categories

Evaluation of tool performance on the pseudo-simulated datasets showed that tools differed in their ability to recover duplicated genes across the different pangenome structure categories. Goldfinder consistently achieved the highest recall, whereas Coinfinder consistently achieved the highest precision. For Goldfinder, only the Closed category achieved relatively high values for both metrics; in the Moderate and Open categories recall remained high but precision dropped sharply, leading to many false positives. Coinfinder showed the opposite pattern: precision remained high in the Closed and Moderate categories, but recall was very low, indicating that many duplicated gene pairs were missed. In the Open category, Coinfinder's recall increased, yet precision collapsed to near zero, again leading to many false positives. These opposing tendencies demonstrate that high recall or high precision alone provide an incomplete picture of tool performance.

The combined influence of both metrics is best captured by the F1-score, which reveals the overall decline in performance as pangenomes become more open. Goldfinder performed best in the Closed category, Coinfinder in the Moderate category and PanForest (marginally) in the Open category, although all tools performed poorly in the latter category. This is likely due to the more variable GPA patterns, making it difficult for the tools to discern duplicates from the large and more noisy accessory gene pool. Again, supporting our hypothesis that closed pangenomes provide a genomic context in which co-selected gene pairs are more easily detected. Statistical tests confirmed that recall differences across categories were significant, particularly between Closed and Open pangenomes. Precision and F1-score differences were not statistically significant, but both metrics show a clear descriptive decline from Closed to Open categories. This pattern suggests that while recall was the only metric with statistically supported variation, precision and F1-score still contribute interpretive value by illustrating the overall loss of performance as pangenomes become more open. In summary, these findings show that tool performance varied systematically across the pangenome structure categories and that performance is lost with pangenomes openness.

Influence of D-value on gene recovery

We revealed that tools differed in their ability to recover duplicated genes across D-value ranges in the pseudo-simulated datasets.

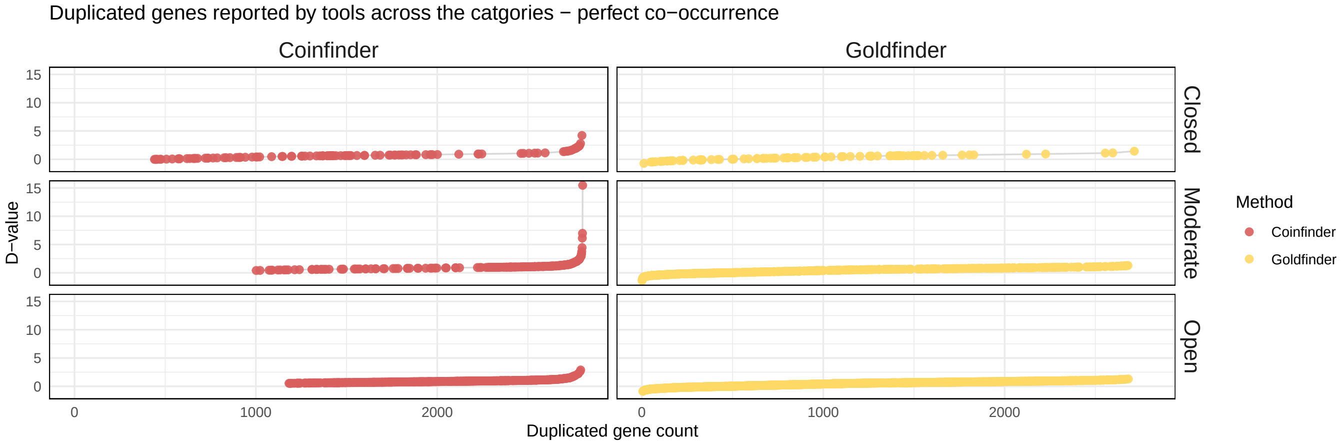

Figure 5. Pseudo-simulated perfect co-occurrence dataset - correctly reported duplicated genes *D*-value distribution. The duplicated genes reported by each method in the Closed, Moderate, and Open pangenome structure categories respectively (top to bottom). The *x*-axis shows the index of each duplicated gene. the *y*-axis shows the *D*-value of a gene. A circle represents a correctly reported duplicated gene. The colour of a circle corresponds to the tool that reported the gene.

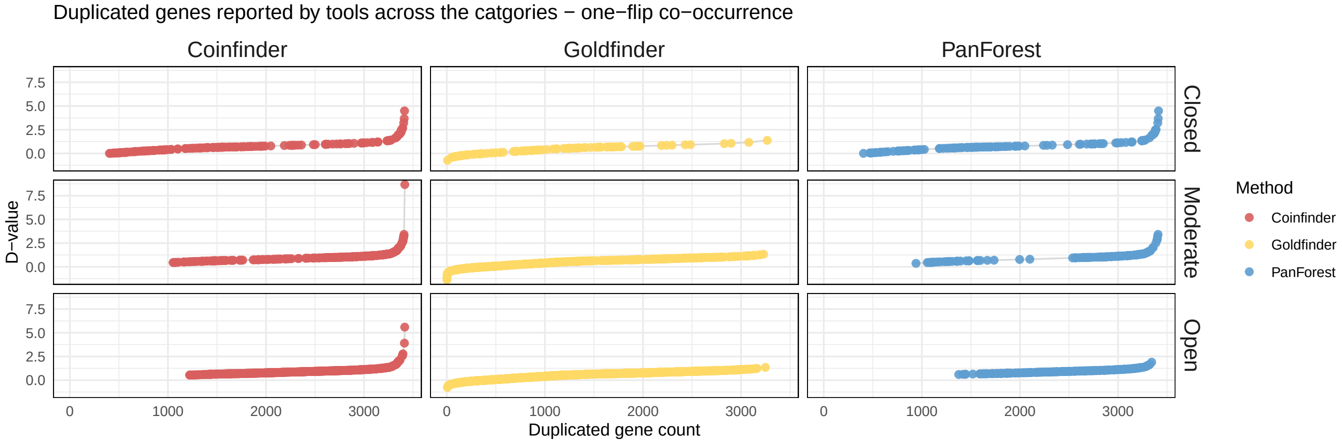

Figure 6. Pseudo-simulated one-flip dataset - correctly reported duplicated genes *D*-value distribution. The duplicated genes reported by each method in the Closed, Moderate, and Open pangenome structure categories respectively (top to bottom). The *x*-axis shows the index of each duplicated gene. the *y*-axis shows the *D*-value of a gene. A circle represents a correctly reported duplicated gene. The colour of a circle corresponds to the tool that reported the gene.

Coinfinder and PanForest were much better at recovering high *D*-value genes. This reflects their design: Coinfinder's binomial tests assume lineage independence, and high *D*-value genes provide greater statistical power and more valid associations. Similarly, PanForest excludes genes with a *D*-value of zero or lower from the GPA matrices used as input for its random forest analyses, meaning that high *D*-value genes serve as stronger predictors. PanForest also shows that its ability to recover high *D*-value genes diminishes the more open a pangenome structure becomes. In contrast, Goldfinder was the only tool to report low *D*-value genes. This difference arises because both Coinfinder and PanForest impose *D*-value cutoffs at different points in their pipelines, whereas Goldfinder does not. Despite being implemented differently, these cutoffs have similar effects in restricting recovery to high *D*-value genes. These findings show that differences in reported co-occurrence are directly influenced by how each method accounts for phylogeny.

Together, our research recovers biologically consistent co-occurrence patterns, provides insight into tool performance trade-offs across pangenome structures, and highlights how methodological assumptions about phylogeny shape gene recovery. These findings future pangenome studies and reveal implications for interpreting co-occurrence results.

4.2. Weaknesses of this study

Assumptions about tool agreement

The assumption is made that the agreement between the three tools represents the most reliable source of information. While reasonable, given a lack of a ground truth, may not have been the best strategy. The poor performance of PanForest may have hindered the discovery of gene-gene associations that were found between Goldfinder and Coinfinder. Alternatively, tool outputs could be validated against independent biological datasets, such as co-localization, functional interaction networks, or co-expression data, which would provide an external benchmark to validate the co-occurrence estimation agreement.

Constraints of pseudo-simulated validation

The validation strategy relied on artificially inserted duplicated genes, which do not represent real biological processes. Goldfinder reconstructs ancestral gene histories and artificial duplicates are not part of the natural evolutionary history. From Goldfinder's perspective these genes look like they have been "gained" in a way that is not congruent with the phylogeny, negatively affecting the null model. PanForest is at a disadvantage in the validation of its results, as perfect co-occurrence is collapsed and the near-perfect co-occurrence duplicated genes are seen as redundant, overlapping information, contributing little additional predictive power. Combined with the fact that PanForest cutoffs are solely based on the paper by Beavan et al. (2024) [21], in which they explicitly set their parameters and thresholds to be as strict as possible to limit themselves to the most dispersed genes in the ∼39,000 gene pangenome of *E. coli*. This makes the experimental setup of pseudo-simulation an inherently poor validation of PanForest's results. These methodological implications limit the generalisability of pseudo-simulation-based validation. Alternatively, fully simulated pangenomes calibrated to match the Fitch score distribution and core/accessory structure of real pangenomes could embed biologically plausible co-occurrence mechanisms. This would provide a controlled ground truth while retaining realistic lineage dispersal, enabling fair benchmarking across pangenome structure categories.

Category selection bias

The Moderate pangenome structure category did not uniformly produce results that were the middle ground between the Open and Closed category (e.g., gene pair and unique gene proportions). The clustering of the pangenomes may explain inconsistencies in reported co-occurrence between the categories, with the selected Open pangenomes bordering the Moderate cluster (Figure S1). Instead of making a selection of pangenomes based on model organisms, comparability to previous studies, and the number of sequenced genomes in the pangenome, a more rigorous experimental approach would have been to pick pangenomes based on cluster separation.

Restricted *D*-value range in pseudo-simulations

Duplicated genes were inserted only within the 5th to 95th percentile of the *D*-value distribution. Excluding these extreme *D*-value genes, which can be biologically interesting, may have masked differences in tool ability ability to detect these hyper-conserved or extremely dispersed genes. Alternatively, these extreme D-values could have been included at first to assess if striking difference appear between the tools, instead of excluding them from the beginning.

Need for external validation

Our research did not validate reported co-occurring genes against independent biological evidence. This limits confidence in the biological relevance of tool outputs. Future studies would greatly benefit from robust validation datasets and benchmarking standards that allow co-occurrence tools to be evaluated against known biological relationships.

Together, these weaknesses show that our research results are constrained by assumptions about tool agreement, the limitations of pseudo-simulation, category selection bias, restricted *D*-value ranges, and the absence of external validation. While these issues temper confidence in the results, they also highlight clear directions for methodological improvement, such as grounding tool outputs in independent biological evidence, developing more realistic simulation strategies, refining dataset selection, and unifying validation approaches. Addressing these areas would strengthen the robustness and generalisability of future pangenome studies.

4.3. Literature comparison

As far as we are aware, a direct comparison between multiple gene co-occurrence approaches, with an emphasis on pangenome structure, has not been published. Direct comparison with prior single-species studies is not feasible. The *S. pneumoniae* pangenome analysed by Coinfinder and Goldfinder, and *E. coli* pangenome analysed by PanForest contain far more genomes than our datasets. Furthermore, our results are aggregated by structure category rather than reported per species. Instead, our results are confronted with broader literature on pangenome structure, validation strategies, and tool-specific performance claims.

Pangenome structure

As discussed in the introduction, the classical definition of the accessory genome of open pangenomes is described as ever expanding as the number of genomes added to the pangenome increase [3]. A major correspondence is found with our work, showing that as pangenomes become more open, not only are more genes and gene pairs found to be associating, we also reveal that performance of the tools declines. This supports our hypothesis that closed pangenomes yield more reliable co-occurrence detection, while open pangenomes are clouded by noise in their vast accessory genomes. Another major correspondence is that the Moderate pangenomes structure category does not behave as an intermediate class. This is at least partly caused by our class definition but may also have a more deep-found origin, as shown by Munteanu et al. (2025) [18]. They describe that genomic fluidity and openness tend to lose resolution in transitional genome states (i.e., our definition of Moderate pangenomes). They suggest that binary classification of pangenomes into open or closed obscures the multidimensional nature of gene content variation. The solution they propose is a new method to quantitively assess the distribution of gene presence across genomes in a pangenome. Stepping away from the binary (or in our case tertiary) classification of pangenomes into a continuous assessment of pangenome structure could provide a more unbiased evaluation.

Gene co-occurrence validation

Our study did not include a validation of the gene co-occurrence results other than that of the pseudo-simulated datasets. This is a minor conflict with literature as the gene co-occurrence studies themselves partly validate their results. The study of Coinfinder [19] provides validation on a few manually selected genes by inspecting genes expected to be co-occurring as part of a multi-protein enzyme, compared to the same genes in a related species from a sister taxon. They successfully annotated six out of eleven genes using KEGG [41] and found that Coinfinder identified consistent co-occurrence relationships between the six genes. Goldfinder [20] does not provide a validation of their co-occurrence results, however, their results are based on the same dataset as Coinfinder. PanForest [21] performed an extensive validation of their results. They used EGGNOG mapper [42], Gene Ontology (GO) enrichment [43], and KEGG pathways analysis [44] on their most predictable genes. They found no enriched GO terms and only two KEGG pathway terms. They also found a 68.7% of these gene pairs to be separated by a distance of ten or fewer genes. This demonstrates that gene co-occurrence validation is generally in a poor state and more effort is required to establish a ground truth baseline upon which these methods can be evaluated fairly.

Tool-specific claims

Coinfinder claims to be an accurate and efficient tool for the identification of coincident gene relationships within pangenomes. A minor correspondence is found in the claim that Coinfinder is accurate, as the tool generally produced the highest precision in our work. A minor conflict is the claim of efficiency, as Coinfinder does not allow for BH FDR correction of p-values as a parameter in its script. This required us to generate results regardless of significance to be able to apply BH FDR correction manually, which heavily increased compute times, disputing their claim of efficiency in our specific case.

Goldfinder claims to be an efficient method that takes the phylogeny of strains into account to be able to distinguish co-selection from pairs that co-occur due to genome-wide linkage. A minor conflict is found in our study as Goldfinder was unable to recover duplicated genes with high D-values (i.e., the genes most likely under co-selection) and was strong in recovering low D-value genes (i.e., the genes that co-occur due to genome-wide linkage). Furthermore, the authors say that Goldfinder infers fewer gene pairs under co-selection but more total genes involved in co-selection than Coinfinder. The second half of that statement was a minor correspondence with our study, as Goldfinder indeed found more total genes than Coinfinder, however, Goldfinder reported many more gene pairs than Coinfinder, across all pangenome structure categories. This is most likely due to the influence of the number of genomes in the pangenome. The authors of Coinfinder demonstrate that with a reduction of total genomes in their dataset, the number of observed gene-gene associations decreased substantially. The Goldfinder study used the same 534 genome pangenome of *S. pneumoniae* to compare against Coinfinder, where our total genome counts were between 62 to 420, with an average of 182. This shows that while Coinfinder's reported number of associations is heavily dependent on the number of genomes in the pangenome, Goldfinder is not influenced nearly as strongly, which is not discussed in the Goldfinder paper.

PanForest claims that the random forest approach has two advantages over the phylogeny-aware approaches. First, the models can be tested on an independent test set that is not used to generate the model. The authors say that they have demonstrated that the gene presence or absence is predictable based only on the other genes in the genome. This is a minor conflict with our results, as we have shown PanForest to perform poorly, regardless of pangenome structure category. This disputes their claim that meaningful predictive power was gained solely based the presence or absence of genes. Second, random forests are able to evaluate much more complex relationships, as well as the pairwise assessment of the phylogeny-aware methods. This claim was not verified in our study, as we removed directionality from the Gini importance score matrix by averaging the upper and lower triangular. However, this is the primary strength of this approach and the authors of PanForest therefore propose that other machine learning algorithms, such as neural networks, may be able to improve the predictions. This could lead to new understanding of pangenome evolution.

Our contribution to the existing literature is the novelty of our study, with no prior study having directly compared multiple gene co-occurrence estimation approaches while explicitly considering pangenome structure. We have found support for our hypothesis that closed pangenome structures yield more reliable co-occurrence detection and added to the discussion of stepping away from the binary classification of pangenome structure. Furthermore, we have shown the limitations of existing validation strategies and call for improved benchmarking standards. Finally, we assessed the claims of the individual gene co-occurrence estimation approaches, which adds the nuance of context-dependency to their efficiency and accuracy claims, as well as highlighting the challenges that machine learning approaches face.

5. Conclusion

In this study, we set out to find out: "What is the effect of pangenome structure on gene co-occurrence estimation approaches?". Our hypothesis was that the effect of pangenome structure is quantifiable and that closed pangenomes will yield more reliable co-occurrence detection, while open pangenomes are clouded by noise in their vast accessory genomes. Our results show that the effect of pangenome structure is quantifiable, where closed pangenomes have a higher proportion of co-occurrence compared to open pangenomes. We have shown that the performance of gene co-occurrence estimation approaches is influenced by pangenome structure, with diminishing performance the more open a pangenome becomes.

These findings show that pangenome structure has a direct effect on the co-occurrence signal quality. Closed pangenomes, having a smaller and more stable accessory genome, provide a genomic context in which co-occurring gene pairs are more consistently preserved and therefore easier to identify. On the other hand, the large accessory genomes under extensive gene turnover in open pangenomes create a genomic context that is highly variable and noisy, which makes the detection of co-occurrence signals more difficult. Altogether, this means that methods that perform well in closed pangenomes do not necessarily generalise to more open pangenomes.

Despite these insights, important knowledge gaps remain. The current gene co-occurrence estimation approaches lack a standardised validation approach, making claims of accuracy or fair comparisons between methods difficult. The binary classification of pangenome structure into open and closed may limit our ability to capture transitional and intermediate pangenomes accurately. Furthermore, the effectiveness and generalisability of machine learning methods remains unclear. These knowledge gaps outline the need for more rigorous benchmarking and development of gene co-occurrence estimation approaches.

Future research should focus on developing robust validation datasets and benchmarking standards that allow co-occurrence tools to be evaluated against known biological relationships. Adopting continuous metrics of pangenome openness, as proposed by Munteanu et al. (2025) [18], would enable more precise assessments of how pangenome structure influences gene co-occurrence. Even though machine learning approaches may not yet perform at the level of phylogeny-aware approaches, further research into more advanced models, including neural networks, may help capture complex gene–gene dependencies that current tools overlook. Taken together, these efforts will deepen our understanding of the complex gene-gene interactions that shape pangenome evolution in bacteria.

Recommendations

Based on the conclusion, a few concrete recommendations are given for future projects.

Develop standardised validation frameworks for gene co-occurrence estimation approaches

As current gene co-occurrence estimation approaches lack a standardised validation approach, comparisons between them are unreliable. This project would involve designing and implementing a benchmarking framework that allows for the fair comparison of methods. A gold-standard dataset could be collected across multiple bacterial species, specifically focusing on species with documented intricate protein complexes, in which genes are expected to co-occur. Metrics can be defined or adopted to unify assessment of gene co-occurrence in this validation framework and existing tools can validated fairly to identify their strengths and weaknesses. The outcome of this project would be a reproducible validation pipeline that could become the community standard for assessing gene co-occurrence estimation performance.

Simulate pangenomes using Fitch score distributions

The binary classification of pangenome structure obscures intermediate pangenomes. This project is centred around developing and implementing a pangenome simulation script using Fitch score distributions (the number of gene gain/loss events along phylogenies). The goal of this project is to create simulations that can be used to generate biologically accurate synthetic pangenomes with varying degrees of openness, with a controlled ground truth for co-occurrence. By tuning the Fitch score distribution, transitional states between open and closed pangenomes can be simulated and explored. The output of this project would be a pangenome simulation pipeline that can be used as a first step in a possible transition away from the binary classification of pangenome structure.

Explore neural networks to discover complex gene-gene dependencies

The main strength of the machine learning approach is that it may be able to capture dependencies that are overlooked by current phylogeny-aware approaches. This project focuses on the development of deep learning architectures (e.g., graph neural networks) that are able to distil complex interactions between all genes in a pangenome, to find predictive patterns in the data. The main objectives of this project would be to represent pangenomes as graphs of presence/absence, train models to predict co-selected pairs using embeddings, and to evaluate the interpretability and biological relevance of the learned features. The output of this project would be a novel approach to uncover complex gene-gene interactions in pangenomes.

References

[1] Nicole T. Perna, Guy Plunkett, Valerie Burland, Bob Mau, Jeremy D. Glasner, Debra J. Rose, George F. Mayhew, Peter S. Evans, Jason Gregor, Heather A. Kirkpatrick, György Pósfai, Jeremiah Hackett, Sara Klink, Adam Boutin, Ying Shao, Leslie Miller, Erik J. Grotbeck, N. Wayne Davis, Alex Lim, Eileen T. Dimalanta, Konstantinos D. Potamousis, Jennifer Apodaca, Thomas S. Anantharaman, Jieyi Lin, Galex Yen, David C. Schwartz, Rodney A. Welch, and Frederick R. Blattner. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. *Nature*, 409 (6819):529–533, January 2001. ISSN 1476-4687. . URL https://www.nature.com/articles/35054089. Publisher: Nature Publishing Group.

[2] R. A. Welch, V. Burland, G. Plunkett, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S.-R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. *Proceedings of the National Academy of Sciences*, 99(26): 17020–17024, December 2002. . URL https://www.pnas.org/doi/full/10.1073/pnas.252529799. Publisher: Proceedings of the National Academy of Sciences.

[3] Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. Scott Durkin, Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D.

Peterson, Christopher R. Hauser, Jaideep P. Sundaram, William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, Mary J. Rosovitz, Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, Michelle L. Gwinn, Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha Watkins, Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig E. Rubens, Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael R. Wessels, Rino Rappuoli, and Claire M. Fraser. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America*, 102(39): 13950–13955, September 2005. ISSN 0027-8424. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1216834/.

[4] Rafael Della Coletta, Yinjie Qiu, Shujun Ou, Matthew B. Hufford, and Candice N. Hirsch. How the pan-genome is changing crop genomics and improvement. *Genome Biology*, 22(1):3, January 2021. ISSN 1474-760X. . URL https://doi.org/10.1186/s13059-020-02224-8.

[5] Nikita Chordia Golchha, Anand Nighojkar, and Sadhana Nighojkar. Bacterial Pangenome: A Review on the Current Strategies, Tools and Applications. *Medinformatics*, June 2024. ISSN 3029-1321. . URL https://ojs.bonviewpress.com/index.php/MEDIN/article/view/2496.

[6] Alex Mira. The bacterial pan-genome: a new paradigm in microbiology. *International Microbiology*, (13):45–57, 2010. ISSN 1618-1905. . URL https://doi.org/10.2436/20.1501.01.110.

[7] Hervé Tettelin and Duccio Medini, editors. *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-38281-0. . URL https://link.springer.com/10.1007/978-3-030-38281-0.

[8] Jason C. Hyun, Jonathan M. Monk, and Bernhard O. Palsson. Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity. *BMC Genomics*, 23(1):7, January 2022. ISSN 1471-2164. . URL https://doi.org/10.1186/s12864-021-08223-8.

[9] L. Rouli, V. Merhej, P.-E. Fournier, and D. Raoult. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7:72–85, September 2015. ISSN 2052-2975. .

[10] Andrey O Kislyuk, Bart Haegeman, Nicholas H Bergman, and Joshua S Weitz. Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics*, 12:32, January 2011. ISSN 1471-2164. . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3030549/.

[11] Michael A. Brockhurst, Ellie Harrison, James P. J. Hall, Thomas Richards, Alan McNally, and Craig MacLean. The Ecology and Evolution of Pangenomes. *Current Biology*, 29 (20):R1094–R1103, October 2019. ISSN 0960-9822. . URL https://www.sciencedirect.com/science/article/pii/S0960982219310280.

[12] Junhui Qiu, Yulan Shi, Fei Zhao, Yi Xu, Hui Xu, Yan Dai, and Yi Cao. The Pan-Genomic Analysis of Corynebacterium striatum Revealed its Genetic Characteristics as an Emerging Multidrug-Resistant Pathogen. *Evolutionary Bioinformatics*, 19:11769343231191481, January 2023. ISSN 1176-9343. . URL https://doi.org/10.1177/11769343231191481. Publisher: SAGE Publications Ltd STM.

[13] Pieter De Maayer, Teigra Green, Sara Jordan, Theo H. M. Smits, and Teresa A. Coutinho. Pan-genome analysis of the Enterobacter hormaechei complex highlights its genomic flexibility and pertinence as a multidrug resistant pathogen. *BMC Genomics*, 26(1):408, April 2025. ISSN 1471-2164. . URL https://doi.org/10.1186/s12864-025-11590-1.

[14] Yeji Kim, Changdai Gu, Hyun Uk Kim, and Sang Yup Lee. Current status of pan-genome analysis for pathogenic bacteria. *Current Opinion in Biotechnology*, 63:54–62, June 2020. ISSN 0958-1669. . URL https://www.sciencedirect.com/science/article/pii/S0958166919301387.

[15] Carolin M. Kobras, Andrew K. Fenton, and Samuel K. Sheppard. Next-generation microbiology: from comparative genomics to gene function. *Genome Biology*, 22(1):123, April 2021. ISSN 1474-760X. . URL https://doi.org/10.1186/s13059-021-02344-9.

[16] Andrea Di Cesare, Ester Eckert, and Gianluca Corno. Co-selection of antibiotic and heavy metal resistance in freshwater bacteria. *Journal of Limnology*, 75(s2), April 2016. ISSN 1723-8633. . URL https://www.jlimnol.it/jlimnol/article/view/jlimnol.2016.1198.

[17] Fiona J Whelan, Rebecca J Hall, and James O McInerney. Evidence for Selection in the Abundant Accessory Gene Content of a Prokaryote Pangenome. *Molecular Biology and Evolution*, 38(9):3697–3708, September 2021. ISSN 1537-1719. . URL https://doi.org/10.1093/molbev/msab139.

[18] Viorel Munteanu, Alexei Leahu, Dumitru Ciorbă, Eugeniu Catlabuga, Nicolae Drabcinski, Damian Dubciuc, Victor Iapăscurtă, and Viorel Bostan. The Pangenome Variability Index: A Quantitative Measure for Assessing Gene Content Diversity in Microbial Genomes. In Victor Sontea, Ion Tiginyanu, and Serghei Railean, editors, *7th International Conference on Nanotechnologies and Biomedical Engineering*, pages 253–261, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-032-06497-4. .

[19] Fiona Jane Whelan, Martin Rusilowicz, and James Oscar McInerney. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microbial Genomics*, 6(3): e000338, 2020. ISSN 2057-5858. . URL https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000338. Publisher: Microbiology Society,.

[20] Athina Gavriilidou, Emilian Paulitz, Christian Resl, Nadine Ziemert, Anne Kupczok, and Franz Baumdicker. Goldfinder: Unraveling Networks of Gene Co-occurrence and Avoidance in Bacterial Pangenomes, May 2024. URL https://www.biorxiv.org/content/10.1101/2024.04.29.591652v1. Pages: 2024.04.29.591652 Section: New Results.

[21] Alan J. S. Beavan, Maria Rosa Domingo-Sananes, and James O. McInerney. Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. *Proceedings of the National Academy of Sciences*, 121(1): e2304934120, January 2024. . URL https://www.pnas.org/doi/10.1073/pnas.2304934120. Publisher: Proceedings of the National Academy of Sciences.

[22] Anja Barth. A population genetic view of bacterial adaptability using persistent pangenome size estimates. Master's thesis, Eberhard Karls Universität Tübingen, September 2024.

[23] Wei Ding, Franz Baumdicker, and Richard A Neher. panX: pan-genome analysis and exploration. *Nucleic Acids Research*, 46(1):e5, January 2018. ISSN 0305-1048. . URL https://doi.org/10.1093/nar/gkx977.

[24] Hannah Götsch and Franz Baumdicker. Unpublished manuscript. URL https://www.baumdickergroup.de/index.php/software/29-phylothin.

[25] Anna E. Dewar, Chunhui Hao, Laurence J. Belcher, Melanie Ghoul, and Stuart A. West. Bacterial lifestyle shapes pangenomes. *Proceedings of the National Academy of Sciences*, 121(21):e2320170121, May 2024. . URL https://www.pnas.org/doi/10.1073/pnas.2320170121. Publisher: Proceedings of the National Academy of Sciences.

[26] Lars Snipen and Kristian Hovde Liland. micropan: an R-package for microbial pan-genomics. *BMC Bioinformatics*, 16(1):79, March 2015. ISSN 1471-2105. . URL https://doi.org/10.1186/s12859-015-0517-0.

[27] Hervé Tettelin, David Riley, Ciro Cattuto, and Duccio Medini. Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*, 11(5):472–477, October 2008. ISSN 1369-5274. . URL https://www.sciencedirect.com/science/article/pii/S1369527408001239.

[28] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987. ISSN 0377-0427. . URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[29] Susanne A. Fritz and Andy Purvis. Selectivity in Mammalian Extinction Risk and Threat Types: a New Measure of Phylogenetic Signal Strength in Binary Traits. *Conservation Biology*, 24(4):1042–1051, 2010. ISSN 1523-1739. . URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1523-1739.2010.01455.x. _eprint: https://conbio.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1523-1739.2010.01455.x.

[30] Franz Baumdicker and Anne Kupczok. Tackling the Pangenome Dilemma Requires the Concerted Analysis of Multiple Population Genetic Processes | Genome Biology and Evolution | Oxford Academic. *Genome Biology and Evolution*, 15(5), May 2023. . URL https://academic.oup.com/gbe/article/15/5/evad067/7137407.

[31] Olive Jean Dunn. Multiple Comparisons among Means. *Journal of the American Statistical Association*, 56(293):52–64, March 1961. ISSN 0162-1459. . URL https://doi.org/10.1080/01621459.1961.10482090.

[32] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, January 1995. ISSN 0035-9246. . URL https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

[33] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in Forests of randomized trees. volume 26, December 2013.

[34] Cyril Goutte and Eric Gaussier. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval*, pages 345–359, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31865-1. .

[35] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

[36] Constantin Ahlmann-Eltze. *ggupset: Combination Matrix Axis for 'ggplot2' to Create 'UpSet' Plots*. 2025. URL https://github.com/const-ae/ggupset.

[37] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. .

[38] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics*, 2(6):110–114, December 1946. ISSN 0006-341X. . URL https://doi.org/10.2307%2F3002019. Place: England.

[39] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26, 2017. .

[40] Russell V. Lenth and Julia Piaskowski. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. 2025. URL https://rvlenth.github.io/emmeans/.

[41] Minoru Kanehisa and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. ISSN 0305-1048. . URL https://doi.org/10.1093/nar/28.1.27.

[42] Jaime Huerta-Cepas, Kristoffer Forslund, Luis Pedro Coelho, Damian Szklarczyk, Lars Juhl Jensen, Christian von Mering, and Peer Bork. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, 34(8):2115–2122, August 2017. ISSN 1537-1719. .

[43] The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, December 2020. ISSN 0305-1048. . URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7779012/.

[44] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44 (D1):D457–462, January 2016. ISSN 1362-4962. .

6. Acknowledgements

Franz Baumdicker for providing the data that was foundational for this work.

Appendices

I. Full species intersected dataset

The full species intersected dataset is available from GitHub.

II. Selected species intersected dataset

The selected species intersected dataset is available from GitHub.

III. AI statement

Generative AI was used to assist in debugging self-written scripts. Generative AI was NOT used for writing or experimental design choices.

Data availability

All raw data and scripts are accessible from GitHub.

7. Supplementary data

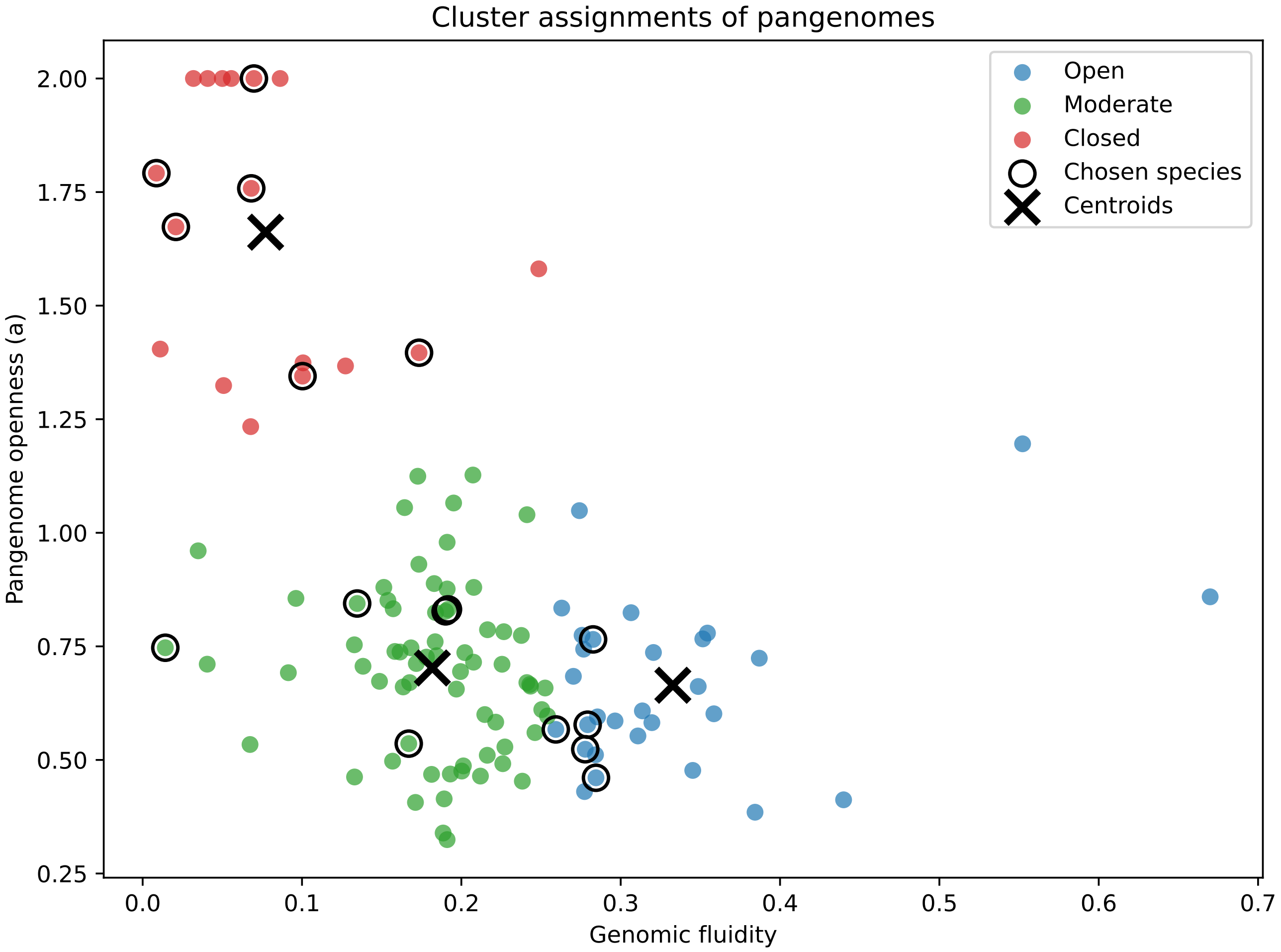

Figure S1. *k*-means clustering result with optimal silhouette score at k=3 of the 114 pangenomes. The *y*-axis shows the Heaps' law alpha value (openness) and the *x*-axis the genomic fluidity. Each circle represents a pangenome in the dataset. The pangenomes are assigned to a cluster based on their proximity to a cluster centroid, shown as a black cross. Red circles correspond to the Closed category assignment, green circles to the Moderate category assignment, blue circles to the Open category assignment. A black outline around a circle indicates a pangenome that is featured in the selected dataset

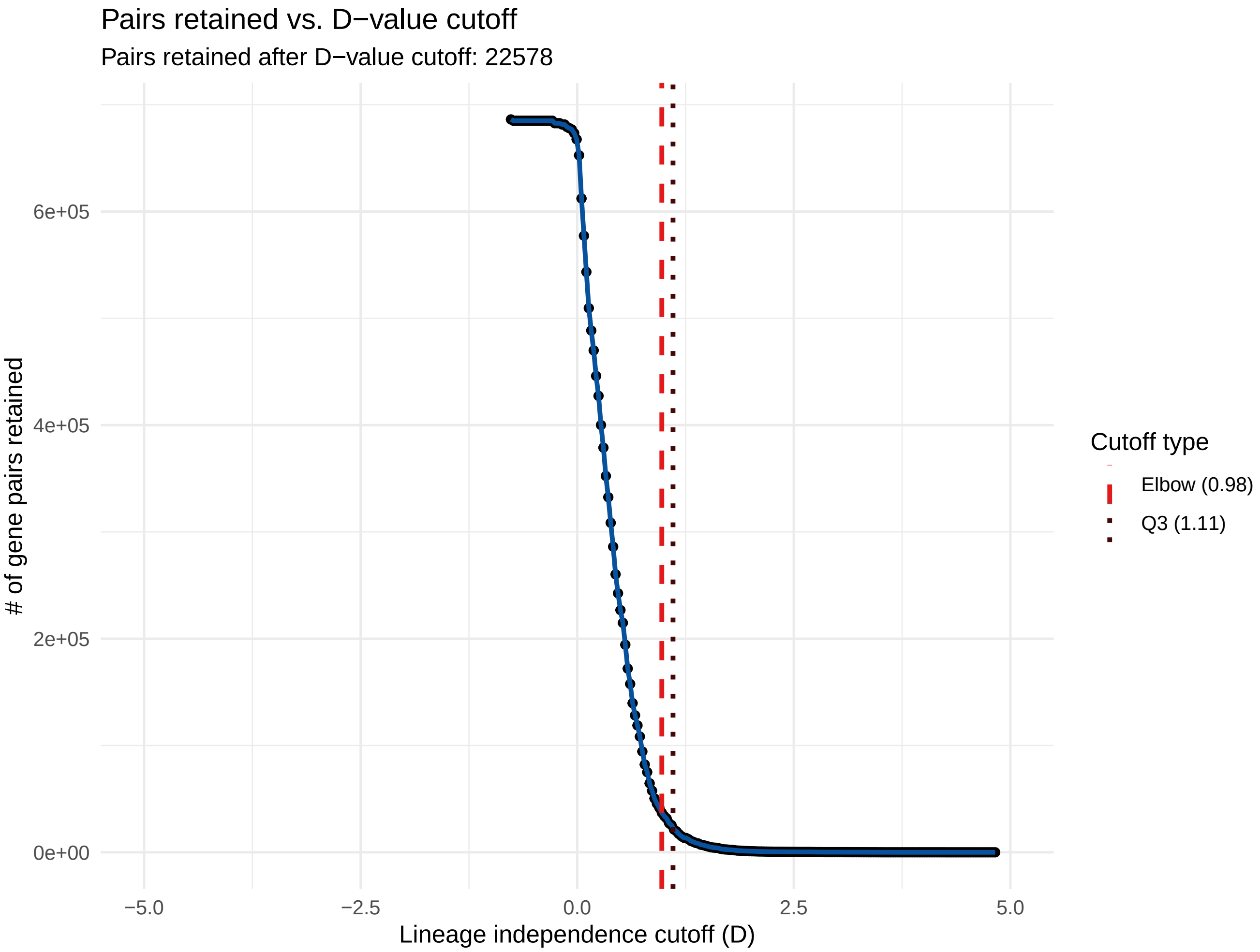

Figure S2. Pairs retained over *D*-value plot for the gene pairs in the pangenome of *Serratia marcescens*. The red dashed line shows the *D*-cutoff point as determined by the elbow detection method. The black dotted line shows the *D*-cutoff point as obtained by taking the third quartile in the *D*-value distribution as the cutoff point. The pairs retained in the subtitle of the figure are from the Q3 cutoff line.

Table S1. Perfect co-occurrence pseudo-simulation dataset - type III ANOVA results, testing the effect of pangenome structure (Closed, Moderate, Open) on performance metrics (Precision, Recall, F1-score).

Metric	DF	*F*-value	*p*-value	Significance
Precision	27	1.7181	0.1984	ns
Recall	27	8.4859	0.001382	**
F1-score	27	4.1135	0.02757	*

Table S2. Perfect co-occurrence pseudo-simulation dataset - post hoc pairwise EMMeans contrasts comparing significant ANOVA effects across pangenome categories for each performance metric.

Metric	Categories	Estimate	SE	*p*-value	*p*-adj	Significance
Recall	Closed - Moderate	-0.138	0.064	0.096	0.192	ns
Recall	Closed - Open	-0.263	0.064	9.28×10^{-4}	5.57×10^{-3}	**
Recall	Moderate - Open	-0.125	0.065	0.156	0.216	ns
F1-score	Closed - Moderate	0.012	0.007	0.18	0.216	ns
F1-score	Closed - Open	0.019	0.007	0.0238	0.0713	ns
F1-score	Moderate - Open	0.007	0.007	0.606	0.606	ns

Table S3. One-flip co-occurrence pseudo-simulation dataset - type III ANOVA results, testing the effect of pangenome structure (Closed, Moderate, Open) on performance metrics (Precision, Recall, F1-score).

Metric	DF	*F*-value	*p*-value	Significance
Precision	39	1.4727	0.2418	ns
Recall	39	6.8593	0.002802	**
F1-score	39	2.0146	0.147	ns

Table S4. One-flip co-occurrence pseudo-simulation dataset - post hoc pairwise EMMeans contrasts comparing significant ANOVA effects across pangenome categories for each performance metric.

Metric	Categories	Estimate	SE	*p*-value	*p*-adj	Significance
Recall	Closed - Moderate	-0.101	0.049	0.11	0.165	ns
Recall	Closed - Open	-0.180	0.049	0.00197	0.00592	**
Recall	Moderate - Open	-0.079	0.050	0.272	0.272	ns