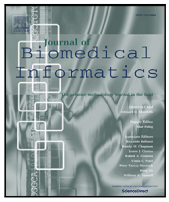
Journal of Biomedical Informatics 148 (2023) 104552

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



PanDelos-frags: A methodology for discovering pangenomic content of incomplete microbial assemblies

Vincenzo Bonnici [a,*], Claudia Mengoni [b], Manuel Mangoni [c,d], Giuditta Franco [b], Rosalba Giugno [b]

[a] Department of Mathematical, Physical and Computer Sciences, University of Parma, Parco Area delle Scienze 53/a (Campus), Parma, 43124, PR, Italy
[b] Department of Computer Science, University of Verona, Strada le Grazie, 15, Verona, 37134, VR, Italy
[c] Fondazione IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo (FG), 71013, Italy
[d] Department of Experimental Medicine, Sapienza University of Rome, Rome (RM), Italy

ARTICLE INFO

Keywords:
Pangenome
Gene families
Sequence homology
Fragmented genomes
Computational approach

ABSTRACT

Pangenomics was originally defined as the problem of comparing the composition of genes into gene families within a set of bacterial isolates belonging to the same species. The problem requires the calculation of sequence homology among such genes. When combined with metagenomics, namely for human microbiome composition analysis, gene-oriented pangenome detection becomes a promising method to decipher ecosystem functions and population-level evolution.

Established computational tools are able to investigate the genetic content of isolates for which a complete genomic sequence is available. However, there is a plethora of incomplete genomes that are available on public resources, which only a few tools may analyze. Incomplete means that the process for reconstructing their genomic sequence is not complete, and only fragments of their sequence are currently available. However, the information contained in these fragments may play an essential role in the analyses.

Here, we present PanDelos-frags, a computational tool which exploits and extends previous results in analyzing complete genomes. It provides a new methodology for inferring missing genetic information and thus for managing incomplete genomes. PanDelos-frags outperforms state-of-the-art approaches in reconstructing gene families in synthetic benchmarks and in a real use case of metagenomics.

PanDelos-frags is publicly available at https://github.com/InfOmics/PanDelos-frags.

1. Introduction

In 2005, Tettelin and colleagues introduced the term pangenome for describing the compositional set of genes in a genome dataset of given species [1]. In particular, genes were divided into three main categories: (i) *core genes*, shared by all genomes of the species and usually involved in essential cellular processes; (ii) *accessory or dispensable genes*, present in some of the strains; (iii) *singletons*, restricted to a single genome. The study of bacterial pangenomes has many applications in the clinical field: e.g. it allows us to both analyze the pathogenic potential and identify specific sequences to predict antigenic epitopes in order to design vaccines [2–4].

In recent years, the number of pangenomic studies has significantly increased [5], due to the reduced cost of genome sequencing and the development of improved pangenome analysis tools. Computational tools able to deal with complete genomes have shown good performance in analyzing both real and synthetic pangenomes, especially those combining alignment-free sequence similarity measures with machine-learning techniques. A traditional tool is Roary [6], which combines alignment BLAST (Basic Local Alignment Search Tool) scores with a pre-processing phase performed by means of CD-HIT (Cluster Database at High Identity with Tolerance) [7] and MCL (Markov Cluster Algorithm) clustering algorithms [8]. In this field, PanDelos [9] has shown to be on average the tool that better recognizes the homology relationships among genes belonging to different bacterial genomes [10].

Unfortunately, the availability of complete genomes, and thus gene sets, is not always guaranteed. A study regarding draft-quality genomes revealed that fragmentation compromises more than 80% of predicted open reading frames and that increased fragmentation correlated with a decreased genome assembly quality, by producing false functional gene annotation [11]. New technologies, such as long-read sequencing, may improve the quality of draft genomes because they help in solving assembly in the presence of repeat elements of the genome, and to identify intragenomic heterogeneities, for example, different copies of 16S rRNA genes [12]. However, pangenomic studies aim at capturing

* Corresponding author.
E-mail address: vincenzo.bonnici@unipr.it (V. Bonnici).

https://doi.org/10.1016/j.jbi.2023.104552
Received 28 February 2023; Received in revised form 6 September 2023; Accepted 19 November 2023

the complete genetic content of a genome. Unfortunately, discordant results are obtained by comparing current technologies [13]. As a result, technological advantages are still not enough to decipher the complete genomic information [14]. This is especially true when data coming from metagenomics experiments has to be analyzed, where the sequences of multiple organisms all at once are captured with increasing complexity in the reconstruction process making it even more cumbersome to reconstruct individual complete genomes [15]. The frequency of specific sub-populations influences the formation of chimeric contigs or produces assemblies with a greater frequency of inversions and insertion/deletions [16]. In any case, there is a plethora of draft-level genomes available in public resources made via short-read low-coverage approaches, some of which are non-cultivable or for which re-sequencing is an expensive operation, and may contain crucial information [17]. Overall, the partial information contained in the fragment of draft-level genomes could have a potentially crucial role in pangenomic analysis. For example, the possibility of managing incomplete information may turn out to be useful in fast and cheap responses to bacterial epidemics [18]. Metagenomics has been widely applied to characterize human-associated bacteria [19–21], allowing to overcome difficulties in cultivating and isolating some bacterial species. Pangenomic content discovery in metagenomics is a key step for understanding the genetic composition of these bacteria and thus their phenotype characterization and cultivar diversity [22]. In this type of study, it is frequent that incomplete genomes are formed by assembling sequencing data. Thus, methodologies able to better recognize fragmented genetic information are key instruments for their analysis.

When applied to incomplete genomes, tools such as Roary and PanDelos suffer the lack of ability to reconstruct the missing information that resides in uncovered portions of the genomes. For this reason, there is a continuous need for genomic data analysis consistent pipelines, and computational tools composing them. Some tools have already been developed in order to overcome the issue of dealing with incomplete genomes: for example GenAPI [23], Pan4Draft [24], and Panaroo [25]. GenAPI computes gene families by performing an initial cluster via CD-HIT-EST [7]. Then a representative of each cluster is chosen to be the gene with the longest sequences. GenAPI directly compares genetic sequences in their incomplete form. As a result, it can only be applied to analyze isolates belonging to the same species, and it may produce unexpected results, especially in less related genomes. Pan4Draft receives as input contigs produced by a preliminary assemble step, and the raw sequencing reads. Such reads are used for trying to close the gap between the contigs via de novo assembly of unmapped reads by means of Spades [26]. Pan4Draft performs an online query for identifying the reference known gene whose sequence is the most similar to the incomplete one. Thus, gene clusters are formed according to such a mapping after a consensus phase. The main issue in such an approach is that genes belonging to the same incomplete genome are mapped to known genes of different genomes. Panaroo merges the fragments of all the input genome into a unified graph-based pangenome. In this way, it is able to correct for intrinsic errors in the assembly and annotation of the fragments producing a better annotation/phenotype calling of the genes. The pangenomic graph is filtered by removing genes whose neighborhood context is not consistent among the input genomes, but such a behavior can potentially remove genes with low presence, such as singletons, if all the involved genomes do not have a certain level of similarity. This aggressive removal approach is mitigated by allowing several parameters to be set by the user. The clustering performed via CD-HIT and supported by context annotation coherence results in a more error-prone grouping of the genes. These approaches introduce specific procedures for dealing with fragmented information, but they ignore previous results regarding the comparison of classical methodologies. In fact, once the missing information is retrieved, the key task of a pangenomic tool is to compute sequence homology and to correctly cluster genes into families, and none of these tools uses the clustering approaches that, in recent years, have been shown to perform better.

In this work, we introduce *PanDelos-frags*, an extension of PanDelos that allows fragmented genomes to be analyzed by a suitable reconstruction of the missing information: the reference genome is selected in such a way as to have the highest number of common nucleotides in the mapping. *PanDelos-frags* extends PanDelos in two ways: (*i*) it adds a pre-processing phase for retrieving missing genomic information that is not covered by input fragments, and (*ii*) it applies a new sequence similarity measure to take into account the percentage of the genetic sequence that has been inferred (see Table 1 for a statement fo significance). In particular, the pre-processing phase aims at selecting a reference genome from a collection of reference sequences and produces a reconstructed version for the input draft genomes by arranging fragments to the reference. Reference-base genome rearrangement is an established technique for reconstructing genomes [27], here extended by an *ad hoc* scoring for reference selection, specific management of clipped parts and arrangement of unmapped fragments.

Compared to the other tools for fragmented genomes [23–25], *PanDelos-frags* is potentially more powerful when the sequencing process does not completely cover the entire genome. In fact, in contrast to other approaches, it tries to reconstruct the missing regions between the assembled fragments. These regions may contain the starting of genes, that will not be captured by the other tools. The selection of a single reference genome, one for each of the input fragmented genomes, allows *PanDelos-frags* to not mix information coming from multiple references. This approach is also helpful in deciding the correct rearrangement of fragments, even when the sequencing covers 100% of the sequenced genome. Single reference-based approach is preferable to multiple-reference reconstruction since a small portion of the bacterial genome is estimated to come from horizontal transferring [28]. This means that, under the assumption that the incomplete genome is close to an already sequenced one, the portion of the sequence to be reconstructed is with a high probability out of horizontally transferred regions. Thus a reconstruction that takes into account the known genome that is the most similar to the incomplete one is the most consistent approach. Lastly, error correction is not embedded in the procedure, but correcting errors at this level discards singletons, or less diffuse genes that can be potential targets for pangenomic applications.

We assessed the performance of *PanDelos-frags* on synthetic bacterial populations generated by simulating evolution and fragmentation with PANPROVA [29]. Statistical evaluation over such synthetic benchmarks shows that *PanDelos-frags* outperforms existing approaches, for both complete and incomplete genomes, by better capturing the set of homology relationships among the retrieved genes. Furthermore, an application to real data coming from a previous study in metagenomics [15] shows that *PanDelos-frags* enables the discovery of the presence of gene families in a wider range of metagenomic assemblies, with respect to Roary, that was originally used or the analysis, and to the other concurrent tools. The resultant families show to have a functional coherence, that is, the genes included by *PanDelos-frags* have a biological function similar to that of the genes composing the family.

The paper is structured in the following sections: Section 2 presents a formalization of the problem of retrieving pangenomic content by computing sequence homology for genetic sequences, Section 3 describes the proposed approach, Section 4 reports the results obtained by computational experiments on both synthetically generated bacterial populations and real metagenomic data, and Section 5 outlines some conclusions.

2. Background and preliminaries

Basic notions

Let $\Gamma = \{A, C, G, T\}$ be the quaternary genomic alphabet, and let Γ^* be the set of all strings, of any length, over Γ, and let Γ^+ to be

Table 1
Statement of significance.

Summary	Description
Problem	A plethora of incomplete microbial genomes is available on public resources but only a few tools can extract pangenomic content from them.
What is already known	When applied to incomplete genomes, classical tools suffer the lack of ability to reconstruct the missing genomic portions. Specialized solutions exist but they do not or only partially reconstruct the missing information, thus they potentially miss the recognition of some genes. Moreover, such approaches tend to mix information coming from multiple known organisms producing unrealistic reconstructions.
What this paper adds	The proposed approach, PanDelos-frags, introduces a novel technique for detecting pangenome content among incomplete genomes. It extends an existing approach, PanDelos, which performs relatively better than other approaches on complete sequences. It adds up a procedure for reconstructing the incomplete genome by inferring missing parts from one single reference sequence. In this way, it detects genes that are partially uncovered in a more realistic way than existing tools. as a result, it retrieves a more feasible set of genetic families in the context of isolated genomes as well as in metagenomic experiments.

Γ^* minus the empty string. Moreover, we denote with Γ^k the set of all possible strings of length k over Γ. We abstract a genome to be a formal string $G \in \Gamma^*$ having length $m = |G|$. We use the same length notation for the cardinality of a collection $\mathcal{G}$ of genomes, for example, $|\mathcal{G}| = n$ if $\mathcal{G} = \{G_1, \ldots, G_n\}$ as traditionally assumed in set theory, combinatorics of words, and formal language theory [30,31].

The W-C complementarity C of DNA strings is the natural extension to strings of the bijection over Γ, whose square is the identity function, associating A with T and C with G. As an example, $C(\mathsf{ATCG}) = \mathsf{TAGC}$. The reverse of a string $w \in \Gamma^*$ is the conventional function R which inverts the reading sense of the string, then $R(a_1 a_2 \ldots a_k) = a_k \ldots a_2 a_1$. The commutative composition of the two functions C and R is often called Mir (as a mirror function, whose square is the identity function of $\Gamma^\star$), and used to compute the reverse complement of DNA double strings [32], where the chemical orientation (corresponding to the reading sense) of the upper and lower filaments are opposite. For example, $Mir(\mathsf{ATCG}) = \mathsf{CGAT}$.

Given two strings, $v = (a_1 a_2 \ldots a_k)$ and $u = (b_1 b_2 \ldots b_t)$, the symbol $\cdot$ denotes the string concatenation operator, such that $v \cdot u = (a_1 a_2 \ldots a_k b_1 b_2 \ldots b_t)$.

Given a genome G, a *genomic region* is either a substring of G, that is, the string $G[i, j]$ from some position i to some position j of G, with $i \leq j$, or its mirror string. Defined over genomic regions, we have the binary function *strand*, which informs about the strand on which the region resides: "+" stays for the upper filament and "-" for the lower filament, and functions $left$ and $right$, which return the left and right coordinates of the region (respectively i and j), by referring to the $5' - 3'$ filament of the genome, even if the region is located on the other filament.

Both a *gene* g and a *fragment* f are genomic regions of the given genome G. We call $\widetilde{G}$ the set of fragments extracted from G and $\widehat{G}$ the collection of all its genes. A well-known concept in the context of alignment-free methods is the genomic dictionary of k-mers, $D_k(G)$, which are elements of Γ^k (i.e., strings of length k) appearing as substrings of a given genome G [33,34]. More formally:

$$D_k(G) = \{v \in \Gamma^k \mid \exists i, 1 \leq i \leq |G| - k + 1 : G[i, i + k - 1] = v\}.$$

Sequence similarity

From the literature, there are several (e.g., alignment-based) distances between sequences, such as the Hamming distance or Damerau–Levenshtein (also known as edit) distance, which may be defined also over two strings of possibly different length. Once we have established how to measure the distance $dist(w_1, w_2)$ between two given strings w_1 and w_2, we may set a non-negative threshold d and consider w_1 and w_2 *similar* when $dist(w_1, w_2) \leq d$. We formally denote such similarity by $w_1 \simeq_d w_2$, a symmetric binary relation which is reduced to the exact matching (i.e., the two strings are equal) in the case of $d = 0$.

An *optimal occurrence* of the word w is defined as the couple (i, j) corresponding to a genomic region (it may be helpful here to remind that for each couple of indexes, we have two possible genomic regions, those within the two filaments) $G[i, j]$ at a minimal distance from w, in particular, the couple with the minimum i is taken among all couples having this property. Genomic regions corresponding to such an optimal occurrence are said to be optimally similar to w:

$$G[i, j] \simeq_d^\star w.$$

A sort of optimal approximate coverage may be now defined, by mapping a set of words over a genome (each word occurring once) by a non-exact matching. First, we define the *word coverage* of a word $w \in \Gamma^*$ within a genome G as (the set of positions engaged by the optimal occurrence of w):

$$cov(w, G) = \{p : w \simeq_d^\star G[i, j], 1 \leq i \leq p \leq j \leq |G|\}.$$

More in general, for a set of strings $W = \{w_1, w_2, \ldots, w_n\}$, the coverage of W over a genome G is

$$cov(W, G) = \bigcup_{w \in W} cov(w, G).$$

Multiple (even overlapping) occurrences of one string w within a given genome G are represented in the following collection:

$$occ(w, G) = \{(i, j) : G[i, j] \simeq_0 w, 1 \leq i \leq j \leq |G|\}.$$

If we restrict ourselves to k-mers dictionaries, for a prefixed value k, we define the generalized Jaccard distance between two strings $w_1, w_2 \in \Gamma^*$:

$$J_k(w_1, w_2) = \frac{\sum_{l \in \{D_k(w_1) \cup D_k(w_2)\}} min(|occ(l, w_1)|, |occ(l, w_2)|)}{\sum_{l \in \{D_k(w_1) \cup D_k(w_2)\}} max(|occ(l, w_1)|, |occ(l, w_2)|)}.$$

Recent studies have shown a good performance of such a measure for retrieving homology among genetic sequences belonging to a set of genomes [35].

Gene families

Genes are transmitted in a vertical way from an ancestor to its descendants, by means of reproduction. Alternatively, genes may be transmitted in a horizontal way from one living organism to another without a direct relationship [36]. Horizontal transmission is common between bacteria. Indeed, a relatively large portion of bacterial genomes is composed of horizontally transferred genes [37].

Roughly speaking, two genes are homologous if they "implement" a common biological function, and often this property corresponds to a structural similarity in their sequence. In this study, we consider two genes homologous if their sequences are *similar* (under some threshold d), in the sense defined in the above section, and such a similarity, combined with statistical evaluation, represents a proof of homology [38].

In particular, two genes respectively located within two different genomes are considered homologous if one is considered the product of transmission of the other, from one genome to the other. The transmitted gene may be an exact copy of the ancestor gene, or it may present some alterations. The two genes are also said to belong to the same gene family. More in general, a gene family is composed of a set of genes for which a direct or indirect relation of transmission exists. On the other hand, genes belonging to different gene families may have a relatively high sequence similarity. Thus, sequence similarity itself is not sufficient to ensure homology. Moreover, the higher the phylogenetic distance between two genomes is, the more permissive

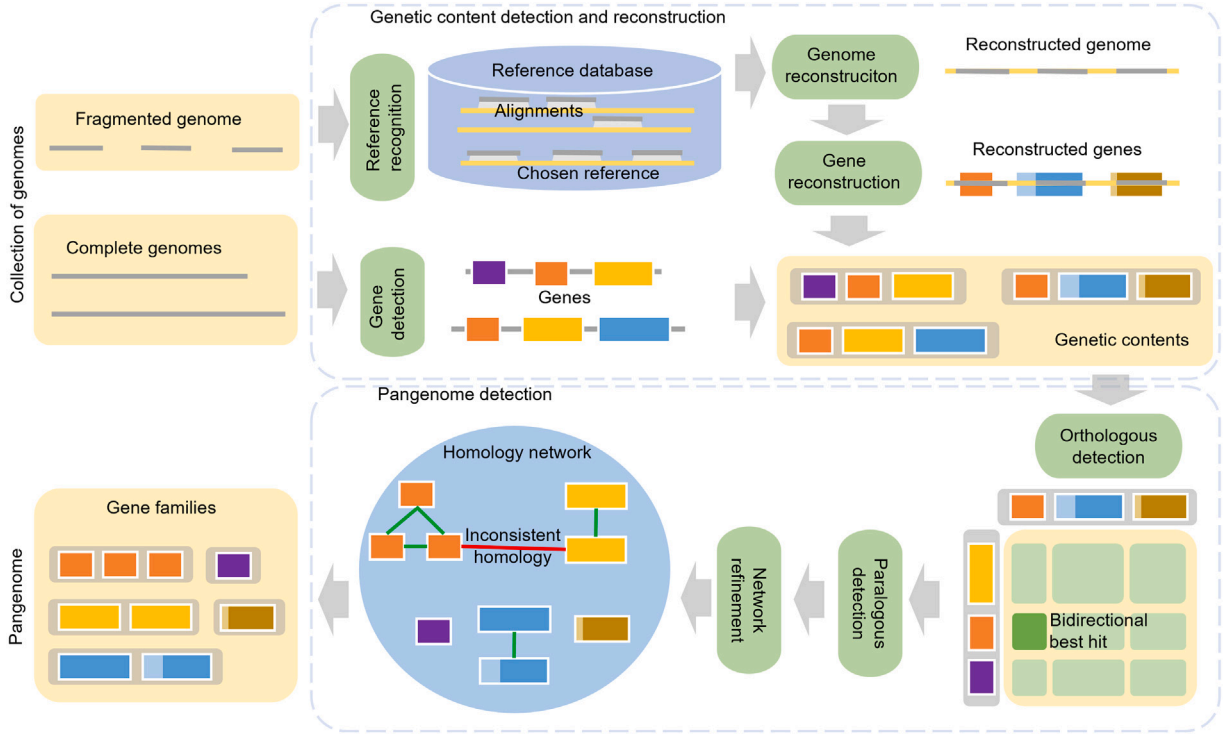


Fig. 1. Workflow of *PanDelos-frags*. The tool takes as input a set of complete and/or fragmented genomes and returns as output the set of reconstructed gene families, namely the pangenome of the set of input genomes. The workflow is mainly divided into two steps: (*i*) reconstruction and detection of the genetic content of the input genomes; (*ii*) detection of the pangenomic content which results in the set of recognized gene families.

the parameter d should be. In fact, a copy of a gene g which has been acquired by means of multiple transmission is led to show a higher level of alterations with respect to g than to its direct ancestor.

In pangenomic analysis, usually, we are not aware of the phylogenetic relationships among the genomes of interest. Therefore, here we represent the homology relationship between two genes g and h as $g \doteq h$, which implies $g \simeq_d h$, for some value d, but the vice-versa is not true. In this way, we are able to model homologous relations such as orthologs (derived from speciation events), paralogs (derived from a gene duplication event, internally to one genome), and xenologs (derived from horizontal transfer or lineage fusion).

Given a set of genomes $\mathcal{G} = \{G_1, \ldots, G_n\}$, a gene family F is composed of genes from the genomes in $\mathcal{G}$ being pairwise homologous. That is, inside a gene family each gene corresponds to a different homologous one. Since a gene belongs to only one gene family and to only one genome, we define string functions $genome(g)$ and $family(g)$ respectively as the genome and the family to which the gene g belongs, respectively. We define the *diffusivity* of a gene family F as the number of genomes $\mathcal{G}$ to which the genes in F belong to. More formally:

$$\delta(F) = |\bigcup_{g \in F} genome(g)|.$$

Given a set of genomes $\mathcal{G} = \{G_1, G_2, \ldots, G_m\}$, the pangenome $P(\mathcal{G})$ is a set of gene families $\{F_1, F_2, \ldots, F_n\}$ such that $\bigcup_{F_i \in P(\mathcal{G})} F_i = \bigcup_{G_j \in \mathcal{G}} G_j$.

3. Methods

Fig. 1 shows the workflow of the proposed approach. *PanDelos-frags* takes as input a collection of input genomes. Such genomes may be complete, namely, the entire genomic sequence is known, or fragmented. Fragmented genomes come in the form of a set of genomic sequences, called fragments, that compose a specific genome. The association between a fragment and its genome is a required input.

For what concerns complete genomes, a gene detection procedure is run in order to recognize the coordinates of the genes and to extract their sequence, while fragmented genomes follow a different flow.

In the case of fragmented genomes, since portions of the genes of a fragmented genome may reside in regions that are not covered by the fragments of such a genome, a specialized reconstruction procedure is applied. The aim is to reconstruct the alleged complete sequence of the genome, and thus the sequence of the genes contained in it. The procedure starts by selecting, from a collection of complete reference genomes, the one that better aligns with the set of fragments of a given fragmented genome. Once a reference genome is chosen, the genomic sequence of the genome is reconstructed. Then, a specialized gene detection procedure is applied in order to recognize genes within such a reconstructed genome. For each gene, we keep track of the percentage of its sequence that has been inferred from the reference genome.

At this point, a set of genes composing the genetic content of each input genome is available and can be used as input for a revised approach for detecting the pangenomic content of a set of complete genomes. For this task, we employed a software called PanDelos [9], where we needed to adjust the sequence similarity measure in order to take into account the information regarding the portion of each reconstructed gene that does not overlap the input fragments. According to PanDelos methodology, an orthologous detection phase is performed by searching for bidirectional best hits in a two-by-two genome comparison. Then, a paralogous detection phase exploits the information regarding sequence similarity between orthologous for selecting the minimum level of sequence similarity between paralogous genes. The retrieved homology (orthologous and paralogous) relationships are integrated into a unified homology network which is refined in order to discard unfeasible homologies. As a result, the connected components of such a refined network identify the gene families that compose the requested pangenomic information.

In what follows, Section 3.1 gives details on the pre-processing phase of *PanDelos-frags*, which is aimed at both reconstructing the genomic sequences of fragmented genomes and detecting the genetic content of all input genomes, while Section 3.2 describes the pangenomic detection phase, which performs homology detection.

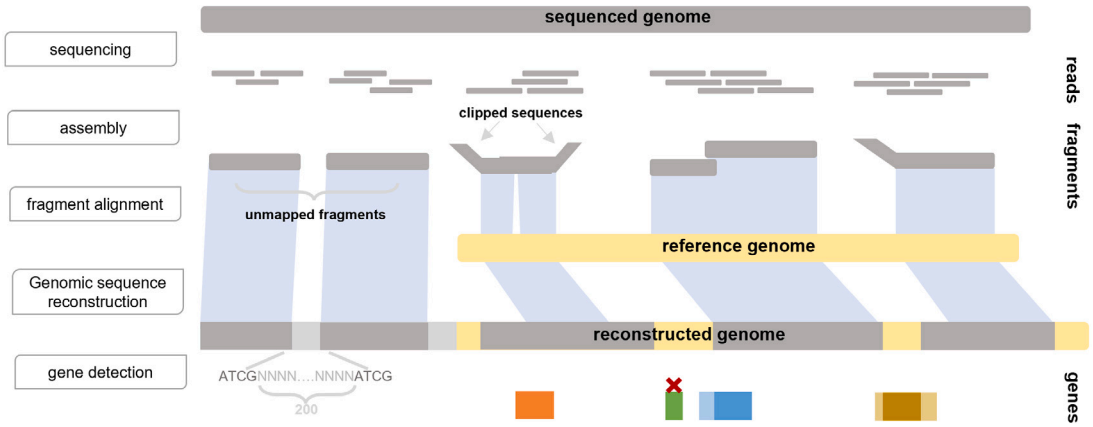


Fig. 2. Pre-processing phase of *PanDelos-frags* for reconstructing genetic content of a fragmented genome by inferring the missing genomic portions from a selected reference genome.

3.1. Genetic content detection and reconstruction

It is supposed that a genome has been sequenced in such a way that the produced reads are assembled into fragments. Such fragments are the input of the pre-processing procedure of *PanDelos-frags*.

Given a collection of genomes, the aim is to recognize the genes from each input genome and to pass such information to the successive phase. If a genome is given in a fragmented format, then it is processed in order to reconstruct the complete sequence of the genome. The reconstruction procedure infers both the missing regions of the genome and the order of the fragments, by retrieving the necessary information for a selected reference genome. To this purpose, a collection $\mathcal{R}$ of relevant reference genomes is selected. Such a collection must contain only complete genomes. This means that only one reference genome is used for the reconstruction of the potentially missing portions of the sequenced genome. In this way, we avoid as much as possible the mixing of multiple genomic sequences which will diverge from a plausible reconstructed sequence.

Given a fragmented genome $\widetilde{G}$ (as a set of fragments extracted from the original G), the reference genome R_G which is the most similar to G is retrieved from $\mathcal{R}$. To recognize R_G, every fragment of $\widetilde{G}$ is aligned to $\mathcal{R}$. The alignment is performed by constructing a BLAST database of $\mathcal{R}$ [39]. Default parameters are used to find all the feasible alignments of a fragment within genomes of $\mathcal{R}$. The reference genome which maximizes the coverage of the alignments with fragments is selected:

$$R_G = \max_{R \in \mathcal{R}} |cov(\widetilde{G}, R)|$$

Once a reference genome is selected, the actual reconstruction procedure is applied. Fig. 2 shows the key aspects of the procedure. Fragments of $\widetilde{G}$ are aligned to R_G by means of the BWA algorithm with default parameters [40]. For each fragment $f \in \widetilde{G}$, we search for the optimal occurrence of f in R_G for some threshold parameter d.

Fragments for which an occurrence has not been found are concatenated at the beginning of the $5'-3'$ strand of the reconstructed genome. Since the actual order of such fragments is unknown, the fragments are lined up according to the increasing order of their length. A fixed number of N symbols (specifically 200) is inserted between each pair of unmapped fragments, and between the last unmapped fragment and the being of the portion of the reconstructed genome that contains mapped fragments. These N symbols are necessary in order to avoid the recognition of genes straddling two unmapped fragments.

For what concerns aligned fragments, we recall that an alignment produces gaps and/or it may clip leading or trailing parts of the fragment. In general, a cut-and-paste approach is applied to replace regions of the reference genome with the mapping fragments, by ignoring gaps and clippings of the alignment. More precisely, given a fragment f and its mapping coordinates $(i, j) = opt_d(f, R)$ on the reference genome R, the resultant genomic string is given by $R[1, i-1] \cdot f \cdot R[j+1, |R|]$. An exception arises when the alignments of multiple fragments overlap (on the same reference genome R). In this case, we proceed pairwise as in the following. Given two fragments, f and f', and their alignment coordinates (i, j) and (i', j'), respectively, such that $i \leq i' \leq j \leq j'$, the resultant genome is given by $R[1, i-1] \cdot f \cdot f'[j+1, j'] \cdot R[j'+1, |R|]$. Namely, we resolve the overlap by keeping the portion of the left fragment that overlaps and by discarding the overlapping part of the right fragment. Such a procedure is iteratively applied from left to right to sequences of overlapping fragments.

Once a reconstructed genome is obtained, a gene recognition tool searches for genes within it. By default, we use PRODIGAL (PROkaryotic DYnamic programming Gene-finding ALgorithm) for this phase [41], which is a well-established algorithm for automated gene prediction in microbial organisms. From the output, we discard the genes that entirely reside in regions that are not covered by given fragments and keep genes that entirely or partially reside in regions obtained by mapping fragments within the reference genome. Intuitively, if a gene recognized by PRODIGAL is not at least partially covered by an input fragment (that is, there is no overlapping of a significant length), we do have not enough information to assume that such a gene is present in the sequenced genome, and we discard it.

For what concerns the complete genomes, gene recognition is performed on the input genome without running the reconstruction procedure, then any recognized gene is kept for the successive phase.

3.2. Pangenome detection

PanDelos-frags inherits from PanDelos [9] the procedure for pangenome detection. PanDelos takes as an input a set of genomes. This implies that the coordinates of genes are already known and no gene detection phase is run by the methodology. Sequence similarity is computed by means of the generalized Jaccard's similarity distance reported in Section 2.

The reconstruction procedure may produce genes that just partially reside on the fragments of input genomes. Thus, a portion of such genes is inferred from the reference genome, and we cannot ensure that such a portion is contained exactly in the sequenced genome. For this reason, we need to apply a modified Jaccard's similarity distance which takes into account our confidence with the reconstructed gene sequence. Such confidence is encoded by a scaling factor that evaluates the percentage of the sequence that has been inferred from the reference genome.

Given a gene g, we denote with $|inferred(g)|$ the number of nucleotides of g inferred from the reference genome. Formally, let $\widetilde{G} =$

$\{f_1, f_2, \ldots, f_n\}$ be a fragmented genome and G' the reconstructed sequences of it, and let $G'[i,j] = g$ for some i and j, then $inferred(g) = \{p : i \leq p \leq j\} \setminus cov(\widetilde{G}, G')$.

Given two genes, g and g', the modified Jaccard's similarity distance (for some parameter k) is defined as:

$$J'_k(g,g') = J_k(g,g') \cdot (1 - \frac{|inferred(g)| + |inferred(g')|}{|g| + |g'|}).$$

The value of k is chosen according to the total length of all genes. Such a choice is made in accordance with theoretical results regarding the comparison between real genomes and random strings [42]. Specifically:

$$k = \lceil log_{|\Gamma|} \sum_{g \in \widehat{G}, G \in \mathcal{G}} |g| \rceil.$$

The extraction of pairs of genes candidate to be homologous is computed by taking into account genomes in a pairwise way. For each pair of different genomes G_1 and G_2 the tool evaluates the homology between each pair of genes $(g_1, g_2) \in \widehat{G_1} \times \widehat{G_2}$ (while $\times$ being the Cartesian product of the two sets of genes). Jaccard-based similarities are computed in order to discover bidirectional hits. Namely, for each gene in one of the two genomes, we search for the most similar genes within the other genome. Formally, for one gene of one genome $g_1 \in \widehat{G_1}$ the set of best hits with respect to another genome G_2 is defined as

$$BH(g_1, G_2) = \{g^\star \in \widehat{G_2} : J'_k(g_1, g^\star) \geq J'_k(g_1, g), g \in \widehat{G_2}\}.$$

The set of bidirectional best hits between two genomes is naturally given by

$$BBH(G_1, G_2) = \{(g_1, g_2) \in \widehat{G_1} \times \widehat{G_2} : g_1 \in BH(g_2, G_1) \wedge g_2 \in BH(g_1, G_2)\}.$$

The $BBH(G_1, G_2)$ is filtered by applying a threshold k ensuring that the involved sequences share a minimum amount of k-mers. In particular, $(g_1, g_2) \in BBH(G_1, G_2)$ must satisfies the following conditions:

$$\frac{\sum_{w \in D_k(g_1) \cap D_k(g_2)} |occ(w, g_1)|}{|g_1| - k + 1} \geq \frac{2}{k}$$

and

$$\frac{\sum_{w \in D_k(g_1) \cap D_k(g_2)} |occ(w, g_2)|}{|g_2| - k + 1} \geq \frac{2}{k}.$$

The resultant set is referred to as $\widehat{BBH}(G_1, G_2)$. Once $\widehat{BBH}(G_1, G_2)$ is retrieved, a specialized procedure detects paralogous genes. For each genome G, a similarity threshold p for considering two genes paralogous is computed, such that:

$$p(G) = \min_{g \in \widehat{G}, g' \in \widehat{G'}, G \neq G'} J'_k(g, g').$$

The set of paralogous relationships between genes of a genome G is defined as:

$$PAR(G) = \{(g_1, g_2) \in \widehat{G} \times \widehat{G} : J'_k(g_1, g_2) \geq p(G)\}.$$

Finally, we build up a homology network, as an undirected graph (V, E) in which vertices are all the genes within the genomes of $\mathcal{G}$, namely $V = \bigcup_{g \in \widehat{G}, G \in \mathcal{G}} g$, and edges represent homology relationships between such genes, that is $E = \left\{\bigcup_{G \in \mathcal{G}} PAR(G)\right\} \bigcup \left\{\bigcup_{G_1, G_2 \in \mathcal{G} : G_1 \neq G_2} BBH(G_1, G_2)\right\}$.

A connected component is said to be inconsistent if it contains two distinct genes, g_1, g_2, within one same genome G, such that $(g_1, g_2) \notin PAR(G)$. Otherwise, the component is said to be consistent. Inspired by the Girvan–Newman algorithm for community detection [43], we scan for inconsistent components and iteratively remove edges from inconsistent components until they become consistent. At each iteration, the algorithm selects the edge within an inconsistent component with the highest edge betweenness, defined as the number of the shortest paths crossing through an edge in a graph [43]. After the removal of such inconsistencies, each connected component of the graph is given as a specific gene family output.

In summary, with respect to PanDelos, the proposed approach introduces the genetic content detection and reconstruction phase which is necessary for managing fragmented genomes. Moreover, it modifies the previous pangenome detection strategy. It introduces in the sequence similarity measure a factor that takes into account the percentage of genetic sequence that has been reconstructed (details are given in what follows). Moreover, since portions of genetic sequences are inferred from reference genomes, the proposed approach discards a filtering strategy presented in PanDelos which forces two sequences to share a given amount of k-mer content (see [9]). In the same fashion as PanDelos, the pangenome detection strategy is parameter-free. However, the tool has inner parameters due to the application of BLAST for reference recognition, the mapping fragments to the reference via BWA, the use of PRODIGAL for gene detection, and the length (200) of the N islands that are inserted between unmapped fragments.

4. Experiments and discussion

Our experimental setup is here reported, along with a discussion on the results of the experiments designed to prove the effectiveness of *PanDelos-frags* in detecting pangenomic content, among both complete and incomplete genomes.

In Section 4.1, we present a performance analysis of the proposed approach, with respect to the competing methodologies, by means of artificially generated benchmarks. We created three synthetic bacterial populations and simulated fragmentation at varying levels. The use of such synthetic data allows us to systematically evaluate the performance of the tools in varying experimental settings, such as the percentage of the genome that is sequenced, as for usual reverse engineering analysis.

In Section 4.2, we show the application of the proposed methodology to a previously published study in the field of metagenomics [15]. Thank to pangenomic analyses, Pasolli et al. discovered thousands of microbial genomes from yet-to-be-named species. We perform pangenomic analysis on three sets of metagenomic assembled genomes belonging to different species and explore the obtained results in comparison to competing pangenomic tools.

4.1. Synthetic data

In order to assess in a systematic way the performance of *PanDelos-frags*, we performed an evaluation on synthetic data as well. The advantage of this type of data is that input genomes are generated by means of a procedure such that we always know which is the expected output (for a pangenomic discovery content methodology). In this way, we can compare the expected output with the output obtained by running a tool, and calculate performance statistics.

4.1.1. Experimental setup

Synthetic benchmarks were generated by means of PANPROVA [29], a tool that generates a population of synthetic genomes by simulating an evolution starting from a single root genome. Genomic sequence alterations and variations in the gene set of a genome are taken into account for simulating the evolution. At each step of the simulation, a genome that currently composes the synthetic population is selected for being reproduced. The genomic sequence of the ancestor genome is modified by deleting or duplicating genes that are present in it. Subsequently, new genes are added to the sequence in random positions of the genome, in order to simulate horizontal gene transfer. Such new genes are extracted from a pool of genes built before the evolution process. Each time a gene is extracted from the pool, it is also removed from it. Then, insertions, deletions and alterations are applied to its genomics sequence in order to obtain the final sequence of the new genome. Single-nucleotide variations are applied to intergenic portions of the genome, while single-codon variations are applied to genetic regions in order to avoid frame-shifting effects. Each process of gene set modification or genomic sequence alteration is driven by some

Table 2
Total number of genes contained in the synthetic benchmarks, produced by means of PANPROVA, and retrieved by the compared tools. (all) reports the complete set of genes recognized by a tool. (PP) identifies the genes retrieved by the tool that have a correspondence with PANPROVA genes.

Species	*P. aeruginosa*			*M. genitalium*			*E. coli*		
Sequenced perc.	50	80	100	50	80	100	50	80	100
PANPROVA	29,445	46,482	57,346	3270	4606	4,903	27,437	43,361	53,534
PanDelos-frags (all)	37,310	59,468	74,430	6280	9597	11,170	34,698	55,123	69,100
PanDelos-frags (PP)	26,460	41,994	52,327	1697	2461	2,674	23,414	36,692	45,614
Roary (all)	37,041	59,154	74,214	5275	8644	10,811	34,358	54,728	68,798
Roary (PP)	25,721	41,304	52,029	1099	1921	2,440	22,565	35,937	45,281
GenAPI (all)	37,428	59,830	75,046	5448	8962	11,195	35,022	55,713	70,077
GenAPI (PP)	25,721	41,304	52,029	1099	1921	2,440	22,565	35,937	45,281
Panaroo (all)	34,580	57,396	73,317	2953	6823	9,424	29,758	52,833	67,168
Panaroo (PP)	25,721	41,304	52,029	1099	1921	2,440	22,565	35,937	45,281

fixed parameters that define the level of alteration that is reached at each evolutionary step.

For this study, we generated 3 synthetic bacterial populations starting from 3 root genomes belonging to different species, that are *Escherichia coli*, *Mycoplasma genitalium* and *Pseudomonas aeruginosa*. For the population generated starting for the *M .genitalium* and *P. aeruginosa* genome, we built a pool of HGT (Horizontal Gene Transfer) genes from 6 other species. Instead, the pool used for the *E. coli* populations was built by extracting genes from other *Escherichia* species. Populations having a total of 100 genomes were generated. Subsequently, 10 genomes were selected from such populations. The selection was made with two different approaches. The first approach extracts the genomes that are the most phylogenetic related to the root, and for this reason, we refer to these benchmarks as *root* benchmarks. The second approach randomly selects from the population a given number of genomes that have not been reproduced during evolution. In the phylogenetic tree of the population, such genomes are the leaves, and for this reason, the benchmarks are called *leaves* benchmarks. Then, for each selected group of genomes, we simulated the fragmentation by randomly extracting fragments of variable length from the genomes. The fragmentation was simulated for different levels of coverage of the resultant fragment set over the genome from which they were extracted. The percentages were varied from 50% to 100% by steps of 10%.

4.1.2. Genes and gene families

We evaluated the ability of *PanDelos-frags* to reconstruct the pangenomic content of the genes that were partially or completely kept during the fragmentation by PANPROVA. In order to do this, we first mapped the genes reconstructed by *PanDelos-frags* against the original set of gene sequences, which are the genes of the synthetic population before fragmentation. For this purpose, we used BLASTN, with parameter *-qcov* 50, to make sure that at least more than half of the gene was being mapped to the original gene. Then, we repeated the approach in the opposite direction, using the original genes as queries and the reconstructed genes as databases. We considered as mapped only the genes that had a bidirectional best hit with their mapping gene.

Table 2 reports the total number of genes that are originally contained in the benchmarks produced with PANPROVA and that are retrieved by the compared tools. The annotation of the input genomes used as root in PANPROVA is retrieved from public resources.[1]. The annotation is very accurate and partially manually curated. In fact, if we run PRODIGAL on them then a large number of genes is recognized. We decided to keep curated annotations rather than automatic predictions produced by PROKKA in order to make the comparison fair to the tool that uses error correction procedures. The number of genes that result from the synthetic evolution and that survived, completely or partially, the fragmentation is reported as PANPROVA in Table 2. Roary, GenAPI and Panaroo take as input annotated fragments, which we produced by applying PROKKA to the fragments, and then they run specific procedures for evaluating the annotation. *PanDelos-frags* takes as input unannotated fragments that will be rearranged in the reconstructed genome. The annotation is then retrieved by running PRODIGAL on the genomic sequence. The rows of the table labeled as *all* report the complete pangenomic content discovered by each tool. On the contrary, the rows labeled as *PP* inform about the number of genes for which we found a correspondence with the output of PANPROVA. *PanDelos-frags* is the tool that on average generates the highest number of genes, but it is also the tool that has always the highest correspondence with PANPROVA. Roary, GenAPI and Panaroo generate a variable number of genes because they apply different strategies, but they have the same corresponding genes. This behavior is due to the fact that none of these tools reconstruct the mission regions, and the annotation only relies on the fragment content, which is the same for all of them.

4.1.3. Homology relationships

Let $\mathcal{G}$ be the set of genomes of a synthetic population. Let $P'(\mathcal{G})$ be the pangenome extracted by a computational tool. The aim of a systematic evaluation is to compare $P(\mathcal{G})$ with $P'(\mathcal{G})$. There are two key factors that can lead a tool to produce a pangenome $P'(\mathcal{G})$ which is different from the real pangenome $P(\mathcal{G})$. The first factor is due to the unavailability of portions of the genomic sequences which may cause the loss of some genes. This situation arises in fragmented genomes when the set of fragments does not totally cover the original genome. The second factor is implicitly due to the methodology of a tool. In particular, a tool may be led to merge or divide gene families in discordance with effective homology relationships. Thus, we are interested in evaluating the difference between $P(\mathcal{G})$ and $P'(\mathcal{G})$ in terms of homology relationships. To this purpose, we compare the homology networks $W(V, E)$ and $W'(V, E')$ built up from the two pangenomes. We notice that the set of nodes V is the same since the two pangenomes regard the same genomes, and thus the same set of genes. We are to compare the respective sets of edges: $E = \{(g, g') \in V \times V \mid \exists F \in P(\mathcal{G}) : g \in F \wedge g' \in F\}$ and $E' = \{(g, g') \in V \times V \mid \exists F \in P'(\mathcal{G}) : g \in F \wedge g' \in F\}$.

With this aim, we calculated the following four sets:

- $TP = \{E \cup E'\}$, that is the set of homology relationships that are in common between the two pangenomes. This set is also referred to as the *true positive* homologies.
- $FP = \{E' \setminus E\}$, that is the set of homologies reported by the computational tool that do not belong to the true set of homologies. This set is also referred to as *false positive* homologies.
- $FN = \{E \setminus E'\}$, that is the set of homology relationships that have been missed by the computational tool. This set is also referred to as *false negative* homologies.
- $TN = \{V \times V \setminus E \cup E'\}$, that is the set of homologies relations that are not in both E and E'. This set is also referred to as the *true negative* homologies.

On top of these four sets, well-known statistics are computed.

[1] https://www.ncbi.nlm.nih.gov/

Table 3
F1-score of *PanDelos-frags* over the synthetic benchmarks obtained by varying the genome used as root and the percentage of genomic sequence that is virtually sequenced.

Species	Type	Percentage of sequenced sequence					
		50	60	70	80	90	100
M. gen.	leaf	0.98	0.99	0.99	0.99	0.99	0.99
M. gen.	root	0.99	0.99	0.99	0.99	0.99	0.99
E. coli	leaf	0.95	0.95	0.96	0.96	0.97	0.97
E. coli	root	0.98	0.98	0.97	0.98	0.98	0.98
P. aer.	leaf	0.97	0.97	0.97	0.97	0.97	0.97
P. aer.	root	0.99	0.99	0.99	0.99	0.99	0.99

Table 4
False discovery rate of *PanDelos-frags* over the synthetic benchmarks obtained by varying the genome used as root, and the percentage of genomic sequence that is virtually sequenced.

Species	Type	Percentage of sequenced sequence					
		50	60	70	80	90	100
M. gen.	leaf	0.008	0.003	0.002	0.001	0.001	0.001
M. gen.	root	0.002	0.014	0.002	0.001	0.001	<0.001
E. coli	leaf	0.044	0.035	0.027	0.021	0.015	0.011
E. coli	root	0.013	0.012	0.011	0.011	0.008	0.006
P. aer.	leaf	0.006	0.006	0.005	0.004	0.003	0.002
P. aer.	root	0.002	0.002	0.001	0.001	0.001	<0.001

It has to be noticed that this type of homology network is very sparse. In fact, families compose cliques within the network but no edges are among these cliques. Thus, the effective set of edges is very small compared to the possible one $V \times V$. Such missing relationships highly relate to TN, and for this reason, measures based on TN could be not so much informative. Thus, we avoid the calculation of measures such as accuracy and true negative rate, which are based on TN. On the other hand, we computed F1-score and false discovery rate (FDR) which are described in what follows.

The F1-score is a combination of the precision and recall by their harmonic mean. The precision is defined as $\frac{|TP|}{|TP|+|FP|}$, and it informs about the portion of homologies relations output by the tool that are known (real) homologies. The recall is defined as $\frac{|TP|}{|TP|+|FN|}$, and it informs about the portion of true positive relationships that have been output by the tool. The F1-score is defined as $2 \cdot \frac{precision \cdot recall}{precision+recall} = \frac{2 \cdot |TP|}{2 \cdot |TP|+|FP|+|FN|}$. The FDR measure informs about the portion of relations output by the tool that are not true relations: it is defined as $\frac{|FP|}{|FP|+|TP|}$.

We evaluated the homology relationships only between the genes that were partially or completely kept during the fragmentation by PANPROVA and that we were able to match by means of the procedure described in the previous section.

Table 3 reports F1-scores of *PanDelos-frags* over the synthetic benchmarks. As expected, the approach has better performance on *root* benchmarks, because the involved genomes are very similar to each other. However, it is interesting to notice that such performance is not affected by the percentage of the genome that is virtually sequenced. *PanDelos-frags* is able to capture the greatest portion of homology relations between genes that are not completely discarded by fragmentation. For what concerns *leaf* benchmarks, *PanDelos-frags* still shows good performance. Low sequencing percentage affects the ability of the proposed approach to retrieve the correct set of homology relationships, however, an average F1-score of 0.96 is obtained. Table 4 shows FDR values. These results follow the trend of F1-scores, with a maximum FDR value of 0.044 obtained for the *leaf* benchmark built from the *E. coli* genome and by simulating a 50% of sequencing covering.

4.1.4. Diffusivity and core genes

We also compared the pangenomes detected by the different tools by running the same analyses on the three *leaves* benchmarks because these benchmarks are the ones that most reflect a realistic situation.

Figs. 3, 4, and 5 show the diffusivity retrieved by each tool in analyzing respectively the *M. genitalium*, *E. coli* and *P. aeruginosa* benchmarks by simulating 50%, 80% and 100% of sequencing coverage. The pairwise comparison of the retrieved diffusivity between *PanDelos-frags* and the other tools is reported in Appendix Figs. A.1, A.2, and A.3.of The charts show the actual diffusivity distribution of the population as it is obtained by the generation via PANPROVA and after the simulation of the sequencing phase. Thus, genes that were discarded by the sequencing process are not included in computing the diffusivity of gene families. Such a diffusivity distribution is to be considered the golden truth of the analysis. Namely, the tool that retrieves a diffusivity distribution which is the most similar to the PANPROVA distribution is intended to be the best approach. In addition, the charts report the F1-score of each tool on the legend of the figure.

PanDelos-frags is generally the approach that recognizes a number of core families, or families with a diffusivity similar to the core ones, which better approximated the actual distribution of PANPROVA, with a few exceptions. As expected, GenAPI and Panaroo are the tools that show the most similar performance to *PanDelos-frags*. The reason is that Roary can only detect genes that are entirely included in the input fragments, and no reconstruction phases are performed by the tool. As a result, Roary has an average F1-score of 0.42. GenAPI and Panaroo show F1-score between 0.73 and 0.96, which are in any case lower than the *PanDelos-frags*'s values. The F1-scores of the homologies of each experiment are reported in the legends of Figs. 3, 4, and 5.

We compute an index to evaluate the ability of a tool to retrieve the actual number of core gene families. Let c be the actual number of core families, and let n be the number of core families identified by a tool, the index is defined as $abs(n-c)/c \cdot 100$, where abs means the absolute value. We call this index the absolute percentage difference. Intuitively, a lower index indicates a smaller difference in the number of core genes identified by the tool compared to the actual number of core gene families. We calculate for each tool the mean value of the index over the six different fragmentation levels. In *M. genitalium* the absolute percentage differences for the four tools are *PanDelos-frags* 26.8%, Roary 94.2%, GenAPI 87.5%, and Panaroo 83.5%. In *E. coli* the percentages for the four tools are *PanDelos-frags* 12.2%, Roary 98.5%, GenAPI 32.6%, and Panaroo 36.8%. In *P. aeruginosa* the percentages for the four tools are *PanDelos-frags* 6.5%, Roary 96.9%, GenAPI 41.6%, and Panaroo 20.8%. In all three cases, *PanDelos-frags* is the tool that most closely resembles the true size of core gene families.

4.1.5. Phylogenetic inference

Lastly, we aimed to evaluate the impact of retrieving a set of core genes that does not reflect the actual core genes of a species. Their presence in all genomes under investigation allows us to study the evolutionary relationship between genomes. In fact, core genes are often employed to build phylogenetic trees of species [44]. We compute the true phylogeny of each of the three synthetic species by building a binary, unrooted tree from the single copy core genes from the original complete genomes of PANPROVA. Core genes are aligned using MAFFT (Multiple Alignment using Fast Fourier Transform) [45] as multiple sequence aligner and a maximum likelihood tree is built on the concatenated alignments using FastTree [46] with the popular substitution model for nucleotides GTR+CAT [47]. The same process is applied to the core genes reconstructed and detected by the four tools at the percentages of sequenced sequences of 50%, 80% and 100%. To evaluate the distance between phylogenies retrieved by the tools and the actual phylogeny we use the normalized Robinson–Foulds (nRF) distance metric [48] computed using the computational framework of ETE3 toolkit [49]. Formally, given two phylogenetic trees named T_1 and T_2 the Robinson–Foulds (RF) metric is defined as $dist(T_1, T_2) = i(T_1) + i(T_2) - 2v(T_1, T_2)$, where $i(T_1)$ denotes the number of internal edges, while $v(T_1, T_2)$ indicates the number of internal splits shared by the two trees. The nRF distance is derived by dividing RF by the maximal possible distance $i(T_1) + i(T_2)$. Intuitively, the maximum distance is

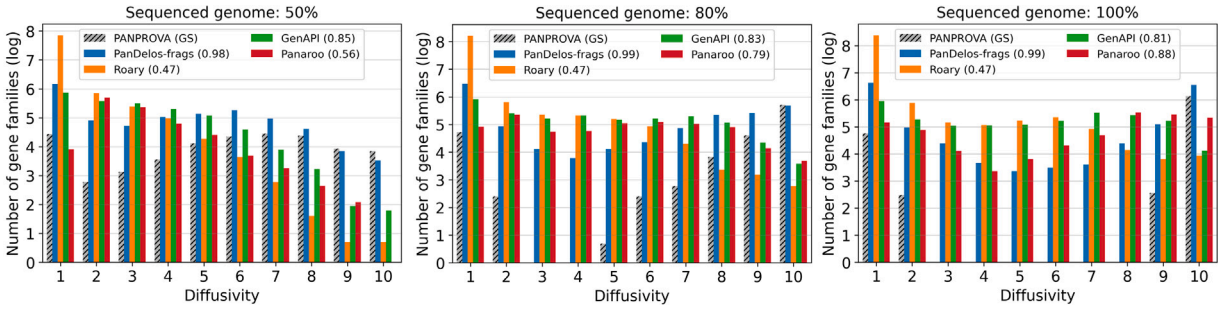


Fig. 3. Diffusivity distributions and F1-scores of the compared tools for the *M. genitalium* synthetic dataset, by varying the percentage of genomic sequence that is virtually sequenced.

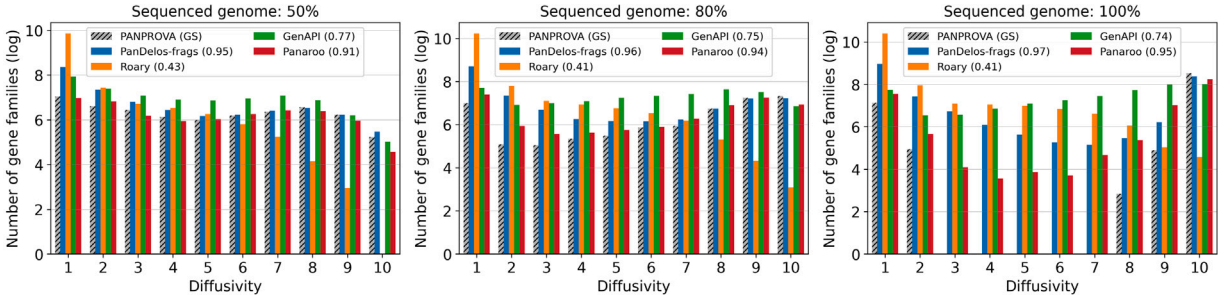


Fig. 4. Diffusivity distributions and F1-scores of the compared tools for the *E. coli* synthetic benchmark, by varying the percentage of genomic sequence that is virtually sequenced.

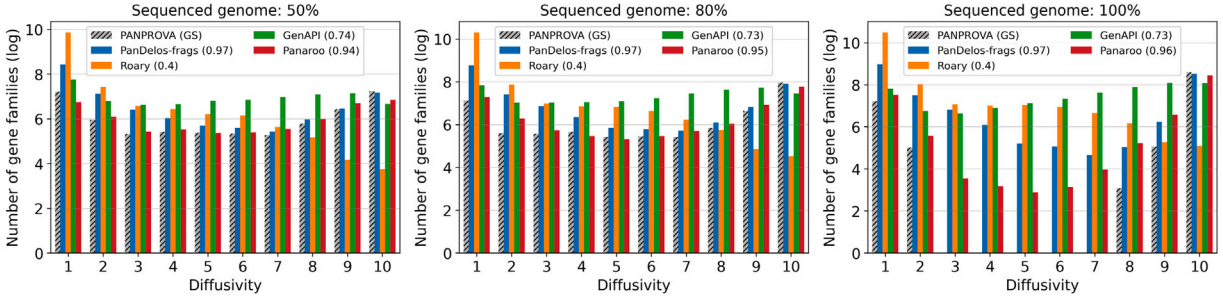


Fig. 5. Diffusivity distributions and F1-scores of the compared tools for the *P. aeruginosa* synthetic benchmark, by varying the percentage of genomic sequence that is virtually sequenced.

given by an nRF of 1, when none of the splits corresponds between two trees, while the minimum value is 0 for topologically identical trees. Table 5 shows the nRF of the experiments. When the percentage of the sequenced genome is 100% the tools perform mostly the same but with the decreasing of the sequenced genome we can appreciate some differences. As expected, Roary is the tool which most suffers from the fragmentation of the analyzed genomes, while *PanDelos-frags* is the tool which resembles more closely the true phylogeny in almost all cases.

Table 5
Normalized Robinson–Foulds distance between golden-truth phylogenetic trees generated by PANPROVA and trees reconstructed by means of the gene families recognized by the computational tools on varying the percentage of sequenced sequence for the genomes composing the given populations.

Species	Type	Sequenced percentage	*PanDelos-frags*	Roary	GenAPI	Panaroo
M. gen.	leaf	50%	0.43	0.83	0.71	1
M. gen.	leaf	80%	0.57	0.71	0.43	0.57
M. gen.	leaf	100%	0.57	0.57	0.57	0.57
E. coli	leaf	50%	0.71	0.86	0.86	0.86
E. coli	leaf	80%	0.43	0.71	0.71	0.57
E. coli	leaf	100%	0.57	0.57	0.57	0.57
P. aer.	leaf	50%	0.57	0.71	0.57	0.57
P. aer.	leaf	80%	0.57	0.57	0.57	0.57
P. aer.	leaf	100%	0.57	0.71	0.57	0.57

4.2. Metagenome

The characterization of human-associated bacteria is hindered by the difficulty in isolating and cultivating certain species that are prevalent in samples [50]. The advent of metagenome shotgun sequencing made it possible to perform culture-independent analyses, and reconstruct nearly complete genomes without the need for reference genomes that are referred to as metagenomic assembled genomes (MAGs) [51]. MAGs are obtained from metagenome shotgun sequencing by first collecting a sample of DNA from an environmental source, which can also be the human body. This DNA is fragmented into smaller pieces and sequenced using a shotgun sequencing method. The resulting short reads are then assembled into longer contiguous sequences (contigs) using specialized assembly software. Contigs are binned into genomic groups based on their similarities, and annotated with functional information, to provide a comprehensive picture of the genomic information present in the environmental sample.

Metagenomic assembly allows us to reconstruct genomic information of microorganisms present in environmental samples, which in turn led to a great increase in the number of human-associated bacterial genomes obtained using this technique [19–21]. Although this represents a great resource for studying the relationship between bacterial

Table 6
Number of genes and gene families retrieved by the compared tools on the metagenomics benchmark.

	Genes			Gene families		
	A. defectiva	*B. nordii*	*P. aeruginosa*	*A. defectiva*	*B. nordii*	*P. aeruginosa*
PanDelos-frags	26,080	120,835	212,287	2963	19,623	11,375
Roary	12,965	40,684	50,363	3078	22,478	13,273
GenAPI	14,633	41,663	54,878	2659	19,777	9,179
Panaroo	12,824	36,320	50,413	2320	8,603	8,496

species and their gene families' evolution, recent studies have shown that even MAGs reflecting the highest quality standards, according to recent guidelines for MAGs quality (completeness >95%, contamination <5%) [52], are subject to biases when their pangenomes are investigated using tools for pangenomic analysis that were originally intended for complete genomes [53]. These include loss of core genes and overestimation of pangenome size [44]. Namely, the study by Pasolli et al. [15] focuses on a systematic reconstruction of human-associated bacterial genomes via a single-sample assembly of 9,428 metagenomic samples, from which 154,723 microbial genomes are reconstructed. The reconstructed genomes all respect quality control criteria and exceed the thresholds for medium quality MAGs (completeness >50%, contamination <5%).

In our study, we consider MAGs from three species belonging to different phyla that were reconstructed, annotated, and made available online [15,54]. All the genomes we selected were assigned a taxonomic annotation up to species level, as we require to include in the analysis one reference genome to be able to compare gene families in downstream analyses, although the presence of a reference is not strictly required for running the pangenomic analysis. The datasets of the metagenomic experiments are composed of 17 *Abiotrophia defectiva* MAGs, 29 *Bacteroides nordii* MAGs, and 39 *Pseudomonas aeruginosa* MAGs. The aim of this analysis is to compare the pangenomes of the three species obtained using *PanDelos-frags* with the pangenomes produced by three other tools for pangenomic analysis: Roary, which is one of the most popular tools for pangenome analysis (which however does not take into account the fragmentation of genes), GenAPI and Panaroo, two more recent tools that can deal with fragmented genomes. Pan4Draft has been built to perform pangenomics analysis on draft genomes as well, by creating an assembly of unmapped reads that are mapped back on the assembled genes. However, since it requires the assembled contigs as input, as well as the raw-reads files, we could not test the tool on MAGs.

Table 6 reports the number of genes and gene families retrieved by the compared approaches along the metagenomic benchmark. *PanDelos-frags* is the tool that retrieves the highest number of genes. The MAGs have completeness >50%, thus the strategy of reconstructing the missing portions of the genomes gives *PanDelos-frags* ad advantage in recognizing broken genes. The error-correction approach of Panaroo makes such a tool to be the one with the lowest number of identified genes and, consequently, gene families. The number of genes found by *PanDelos-frags* may seem an overestimation of the real amount of genes. However, the assembled reference genomes of *A defectiva*, *B. nordii* and *P. aeruginosa* reported in NCBI (National Center for Biotechnology Information)[2] have 1,897, 4,357 and 5,697 genes, respectively, that multiplied by the number of MAGs of each species is equal to 32,249, 126,353 and 222,183, respectively. This means that the count of *PanDelos-frags* is very close to the expected one. Moreover, such a count is made by taking into account genes that are totally or partially contained in input fragments, making it even more reasonable because genes in reference-only regions were excluded.

We have defined in Section 2 the concept of diffusivity, as the (cardinality of the) set of genomes to which the genes of a family belong to. We use this notion in Figs. 6a, 7a, and 8a to describe how the gene families extracted via pangenomic analysis are represented across multiple genomes, and to visualize the overall pangenome distribution. In all three species, the tools show a common trend in their distribution. The highest peaks of the distributions are found on the far-most left side of the distribution, where singletons are represented, and on the far-most right side, where core gene families are. Although the number of singletons is comparable across tools, we see that in all three experiments *PanDelos-frags* is able to retrieve a greater number of core gene families than all the other tools.

For this type of study, we are not aware of the actual diffusivity distribution for the involved genomes. Thus, we perform a comparison between the tools by defining a mapping between the gene families retrieved by one tool with the gene families retrieved by the other tools. The mapping is performed according to the genes that are present in the reference genomes which we have included for the analysis. Since Roary was used in the original study, we perform this comparison between *PanDelos-frags* and Roary, in order to show potential differences in the results of the same study.

For each gene g_r in the reference genome, we identify with $F^R(g_r)$ the gene family of g_r according to Roary, and with $F^P(g_r)$ the gene family of g_r according to *PanDelos-frags*. Thus, we built a bijective mapping of the families extracted by a tool to the families extracted by the other tool. Given that, a tool may assign to the same family two (or more) different reference genes, we duplicate such families to obtain the bijection. Let δ be the diffusivity, which will have different values according to Roary and according to *PanDelos-frags*. The goal is to compare $\delta(F^R(g_r))$ and $\delta(F^P(g_r))$ for each reference gene g_r. For the comparison, we employ a matrix M such that, given a set of m input genomes containing a reference genome G_r, the matrix has m row and m columns. The rows identify the diffusivity assigned by Roary, and the columns identify the diffusivity assigned by *PanDelos-frags*. Thus, $M[i,j]$ is given by $|\{g_r \in G_r : \delta(F^P(g_r)) = i\} \bigcup \{g_r \in G_r : \delta(F^R(g_r)) = j\}|$.

Namely, $M[i,j]$ reports the number of reference genes for which *PanDelos-frags* has assigned diffusivity i to the corresponding family, and Roary diffusivity j. The secondary diagonal of M identifies the reference genes, and thus the families, for which *PanDelos-frags* and Roary are in accordance with their diffusivity. Cells above the secondary diagonal identify families for which Roary was able to retrieve a bigger (with higher diffusivity value) family of *PanDelos-frags*. While cells below the secondary diagonal identify families for which *PanDelos-frags* assigns a bigger diffusivity. In Figs. 6b, 7b, and 8b we plot the matrix of diffusivity computed on the three metagenomics datasets. To understand how the matrix should be interpreted, we focus for example on Fig. 6b. The number in the upper right corner cell indicates that there exist 93 core gene families (diffusivity equal to the number of genomes in the experiment, e.g. 18) identified both by *PanDelos-frags* and Roary. However, *PanDelos-frags* identifies in total 204 core gene families, meaning that, apart from the 93 core gene families that correspond in both tools, all the other gene families that *PanDelos-frags* identified as core gene families were assigned a lower diffusivity by Roary. More specifically, the distribution of their lower diffusivity can be appreciated in all the cells in column 18, distributed over the different rows.

In *A. defectiva* (Fig. 6b), out of 1,831 gene families, 931 (50.8%) have bigger diffusivity in *PanDelos-frags*, 890 (48.6%) have the same diffusivity, and 10 (0.5%) have bigger diffusivity in Roary. Respectively, in *B. nordii* (Fig. 7b), out of 4,335 matching gene families, the

[2] https://www.ncbi.nlm.nih.gov/datasets/taxonomy/46125/, https://www.ncbi.nlm.nih.gov/datasets/taxonomy/291645/ and https://www.ncbi.nlm.nih.gov/datasets/taxonomy/287/

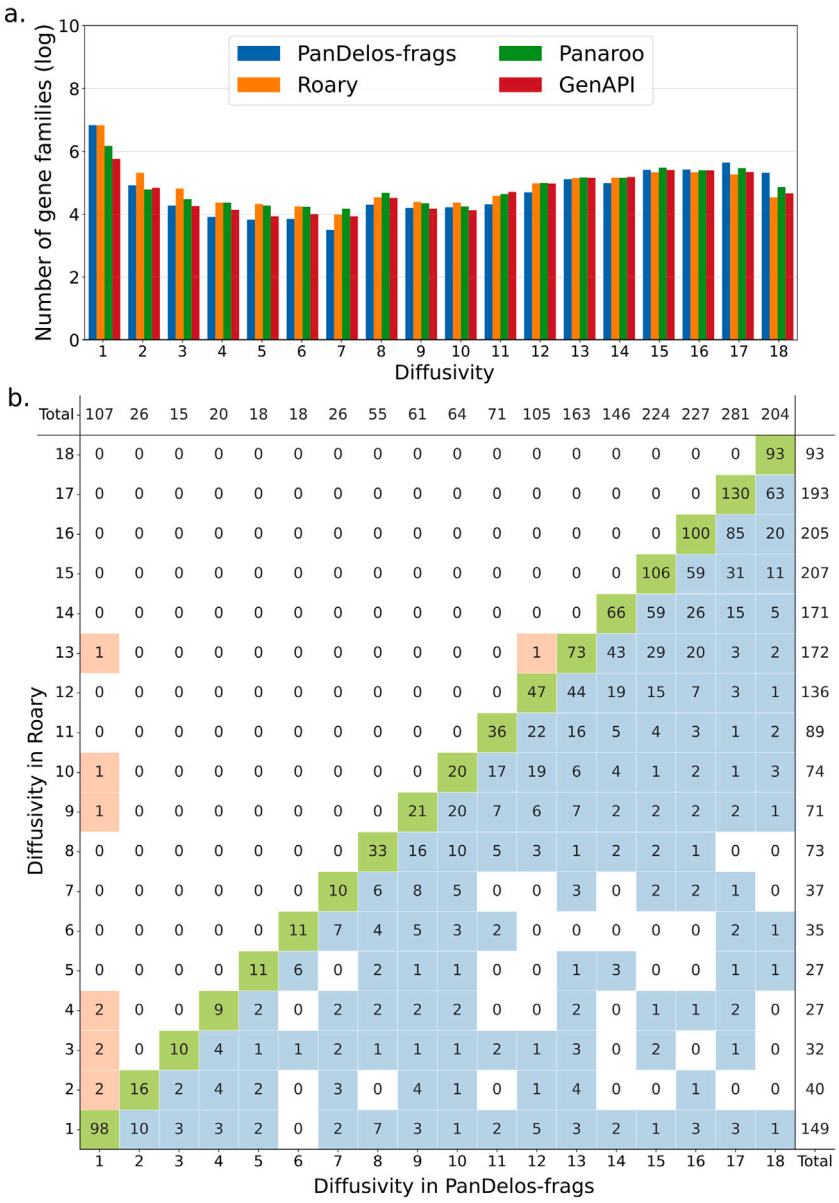


Fig. 6. Diffusivity in of gene families in *A. defectiva* metagenomic data. (a) Diffusivity distributions retrieved by the compared tools. (b) Diffusivity concordance between *PanDelos-frags* and Roary.

number of gene families in the three categories is 2,439 (56.3%), 1,852 (42.7%), and 44 (1%). In *P. aeruginosa* (Fig. 8b), out of 5,655 matching gene families, the numbers are 4,494 (79.5%), 1,110 (19.6%), and 51 (0.9%). In general, we associate a bigger diffusivity with a better performance of the tool. However, since in a metagenomic setting, we are not able to know which is the actual family size, we validated the relevance of the genes included by *PanDelos-frags* and excluded by Roary by means of functional annotation. Specifically, we investigated whether the genes that belonged to a family with a larger diffusivity in *PanDelos-frags* than in Roary were functionally coherent with the rest of the family they had been assigned to. We took all the families where *PanDelos-frags* had a greater diffusivity than Roary, which is for the three species, respectively 931, 2,439, and 4,494 gene families, and we mapped the genes to the set of annotated coding sequences of the reference genome of the investigated species. This was achieved using Diamond [55] with default parameters. We compared the function that was assigned to the *PanDelos-frags* specific genes, defined as all the genes in the family that were assigned to a certain family by *PanDelos-frags* and not by Roary, to those of the majority of the remaining genes in the family. We found that respectively in *A. defectiva*, *B. nordii*, and *P. aeruginosa* the function was coherent in 97%, 90.3%, and 94.5% of the families. On the opposite, when we look at the functions assigned to the genes in the families that have greater diffusivity in Roary than in *PanDelos-frags*, which are for the three species 10, 44 and 51, we report that the percentage of families where the genes identified only by Roary having a coherent function with the rest of the family is 70%, 86.4%, and 84.3%. These experiments suggest that *PanDelos-frags* is able to identify a greater number of core gene families, and overall families which are more diffused across the genomes, thus giving a complete description of the pangenomes of the analyzed species. By investigating the function of the genes that are assigned to a family only by *PanDelos-frags* we are able to see that in more than 90% of cases, the genes detected are functionally coherent with the rest of the family, suggesting that the genes are correctly assigned to the families.

5. Conclusions

Extraction of pangenomic content from metagenome samples is a key step in the investigation of human microbiome composition.

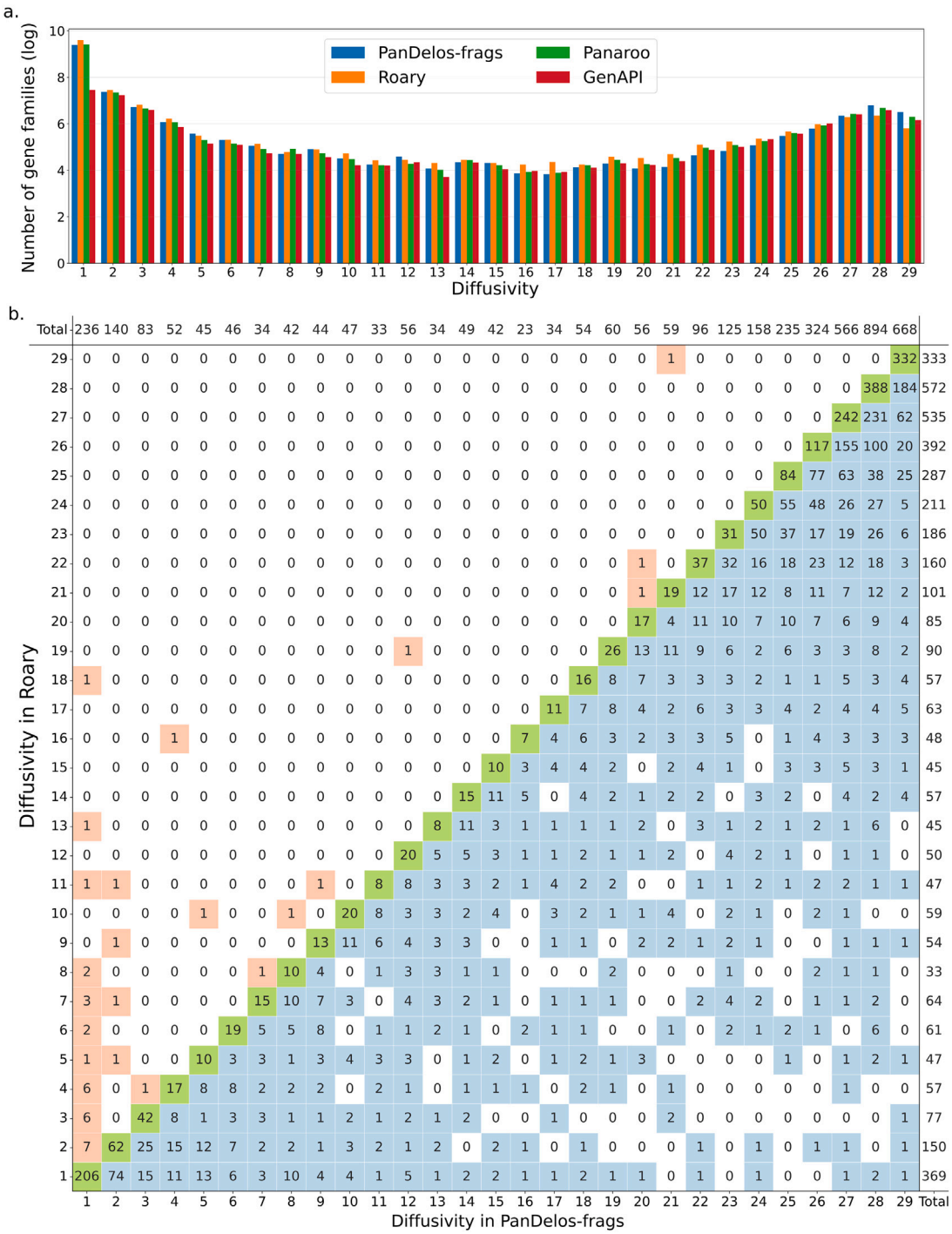


Fig. 7. Diffusivity in of gene families in *B. nordii* metagenomic data. (a) Diffusivity distributions retrieved by the compared tools. (b) Diffusivity concordance between *PanDelos-frags* and Roary.

Several tools exist for extracting the pangenomic content of a group of genomes for which the complete sequence is available. However, when analyzing data from metagenome experiments, usually it is not possible to assemble genomes at their complete stage, while fragments are retrieved. It is then crucial to develop specialized computational tools that are able to extract pangenomic content from such fragmented information.

Most of the existing approaches able to deal with fragmented genomes lack in exploiting previous results obtained for complete genomes in which alignment-free methodologies have shown the best performance. In fact, it is shown that approaches combining alignment-free sequence similarity with artificial intelligence techniques better solve the problem of grouping genes into gene families. Here, we presented *PanDelos-frags*, a computational tool that extends a state-of-the-art algorithm in order to work on fragmented genomes. It includes a specialized procedure for inferring the missing piece of information by reconstructing genomic sequences according to a genomic reference database. This allows the proposed approach to recognize genes that were partially corrupted by the fragmentation, because of low-coverage sequencing or arrangement ambiguity due to repeated portions of the genomes. However, this approach comes with the limitation of assuming that a closely-related completely-assembled reference genome is available. A suitable sequence similarity measure is also defined, to tackle the fact that a portion of the genetic sequences is inferred. As

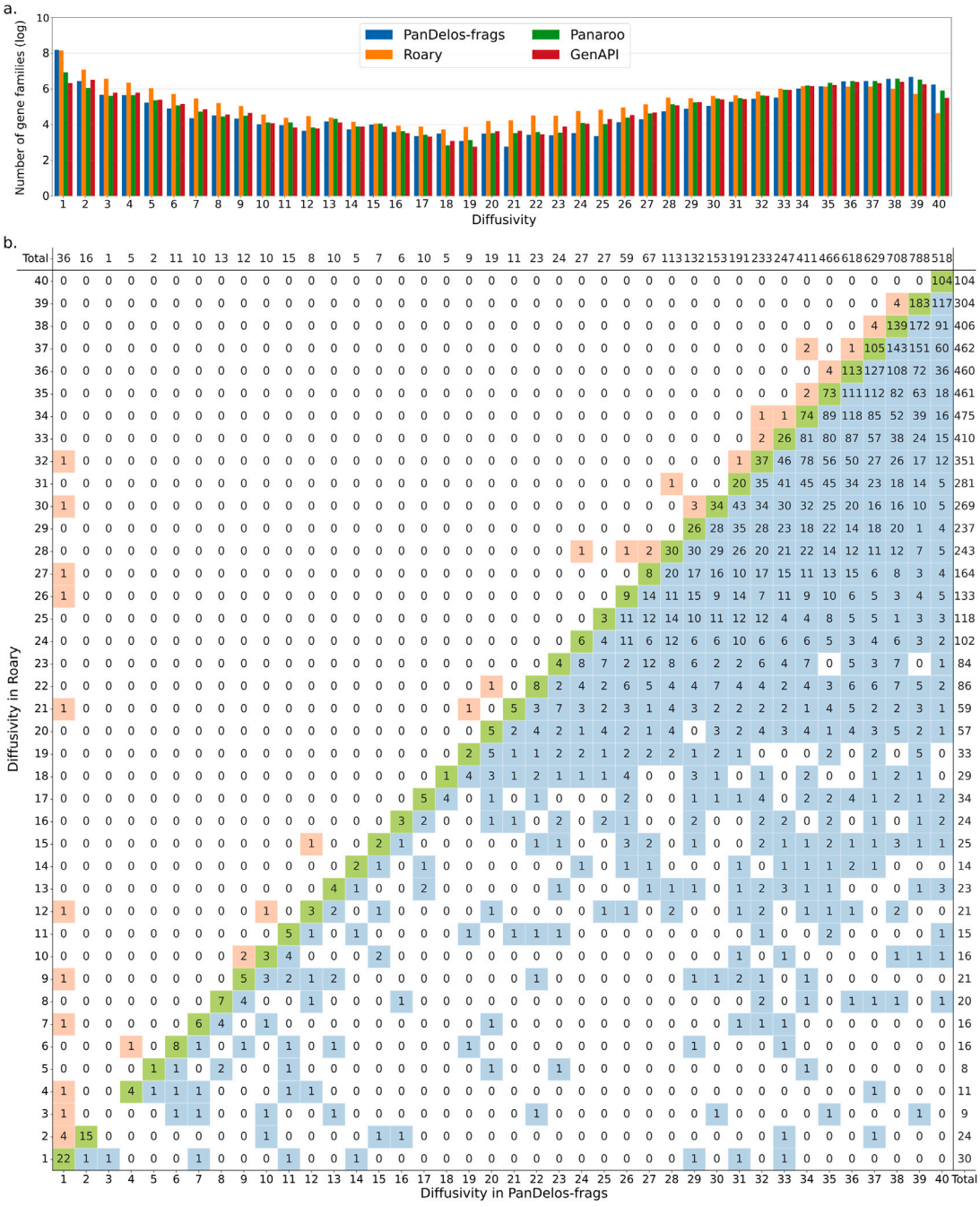


Fig. 8. Diffusivity in of gene families in *P. aeruginosa* metagenomic data. (a) Diffusivity distributions retrieved by the compared tools. (b) Diffusivity concordance between *PanDelos-frags* and Roary.

a result, *PanDelos-frags* shows better performance compared to existing tools in reconstructing gene families, thus the pangenomic content. Tests on real metagenomic data coming from previous experiments show that by means of *PanDelos-frags*: (*i*) a more complete set of genes is extracted from fragmented genomes; (*ii*) the presence of a gene family spans across a bigger set of input genomes; and (*iii*) the resultant enlarged gene families still show functional coherence. In addition, the performance of *PanDelos-frags* was systematically evaluated by means of synthetic benchmarks. The PANPROVA tool was used to create a set of synthetic bacterial populations, by simulating evolution with specific parameters of sequence variation and horizontal gene transfer. A statistical evaluation shows that *PanDelos-frags* better captures the set of homology relationships among genes, when compared to existing approaches, and enables a better phylogenetic analysis of gene families. These results are shown by varying the percentage of genomic sequence that has been virtually sequenced.

As a future development, we aim to reduce the limitation due to the fact that a reference genome similar to the sequenced one must be available. A possible solution could be the development and training of modern generative machine-learning models, able to deal with the risk of producing unrealistic sequences generated as a consensus of the genomes used to train the model, that in our case will be used to fill the gap between fragments. Moreover, we plan to introduce error correction procedures by enabling users to switch them on/off according to the suspected presence of a specific error type. Lastly, we plan the development of a user-friendly interface, especially focused on downstream analyses.

CRediT authorship contribution statement

Vincenzo Bonnici: Conceptualization, Methodology, Investigation, Formal analysis, Writing – reviewing & editing, Software, Validation, Visualization, Data retrieve-curation, Experiment, Original draft preparation, Supervision. **Claudia Mengoni:** Conceptualization, Methodology, Investigation, Formal analysis, Writing – reviewing & editing, Software, Validation, Visualization, Data retrieve-curation, Experiment, Original draft preparation. **Manuel Mangoni:** Software, Validation, Visualization, Data retrieve-curation, Experiment, Original draft preparation. **Giuditta Franco:** Conceptualization, Methodology, Investigation, Formal analysis, Writing – reviewing & editing, Supervision. **Rosalba Giugno:** Conceptualization, Methodology, Investigation, Formal analysis, Writing – reviewing & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Vincenzo Bonnici reports financial support was provided by University of Parma. Vincenzo Bonnici reports a relationship with University of Parma that includes: employment. This project has been founded by the University of Parma (Italy), project number MUR_DM737_B_MAFI_BONNICI. V. Bonnici is partially supported by IDnAM-GNCS, projects number CUP E55F22000270001 and CUP E53C22001930001 V. Bonnici and R. giugno are also supported by the CINI InfoLife laboratory.

Acknowledgments

This project has been partially founded by the University of Parma (Italy), project number MUR_DM737_B_MAFI_BONNICI. V. Bonnici is partially supported by INdAM-GNCS, project number CUP_E55F22000270001, and by the CINI InfoLife laboratory.

Appendix. Supplementary figures

See Figs. A.1–A.3.

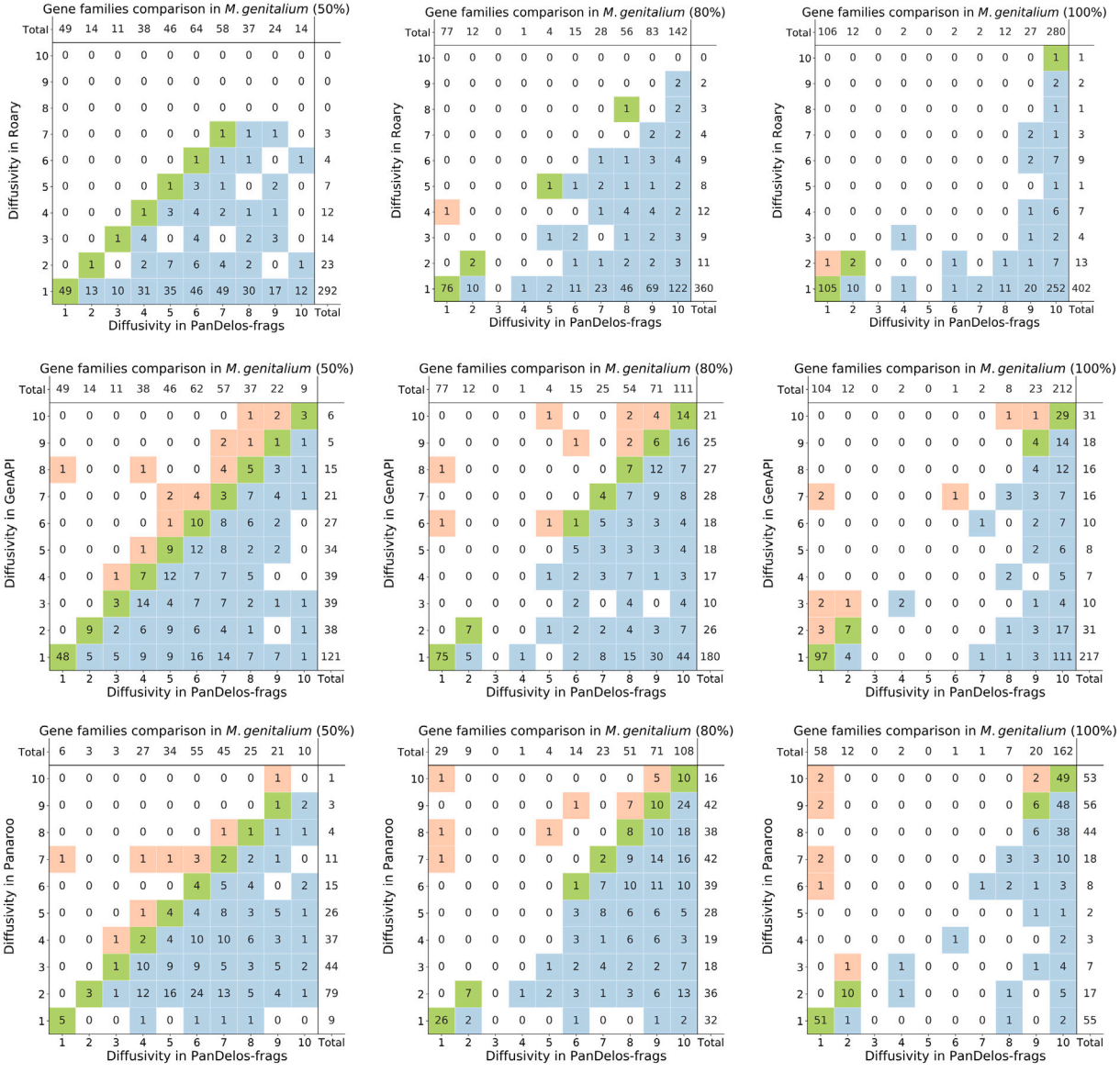


Fig. A.1. Diffusivity concordance between *PanDelos-frags* and other tools in *M. genitalium* synthetic data.

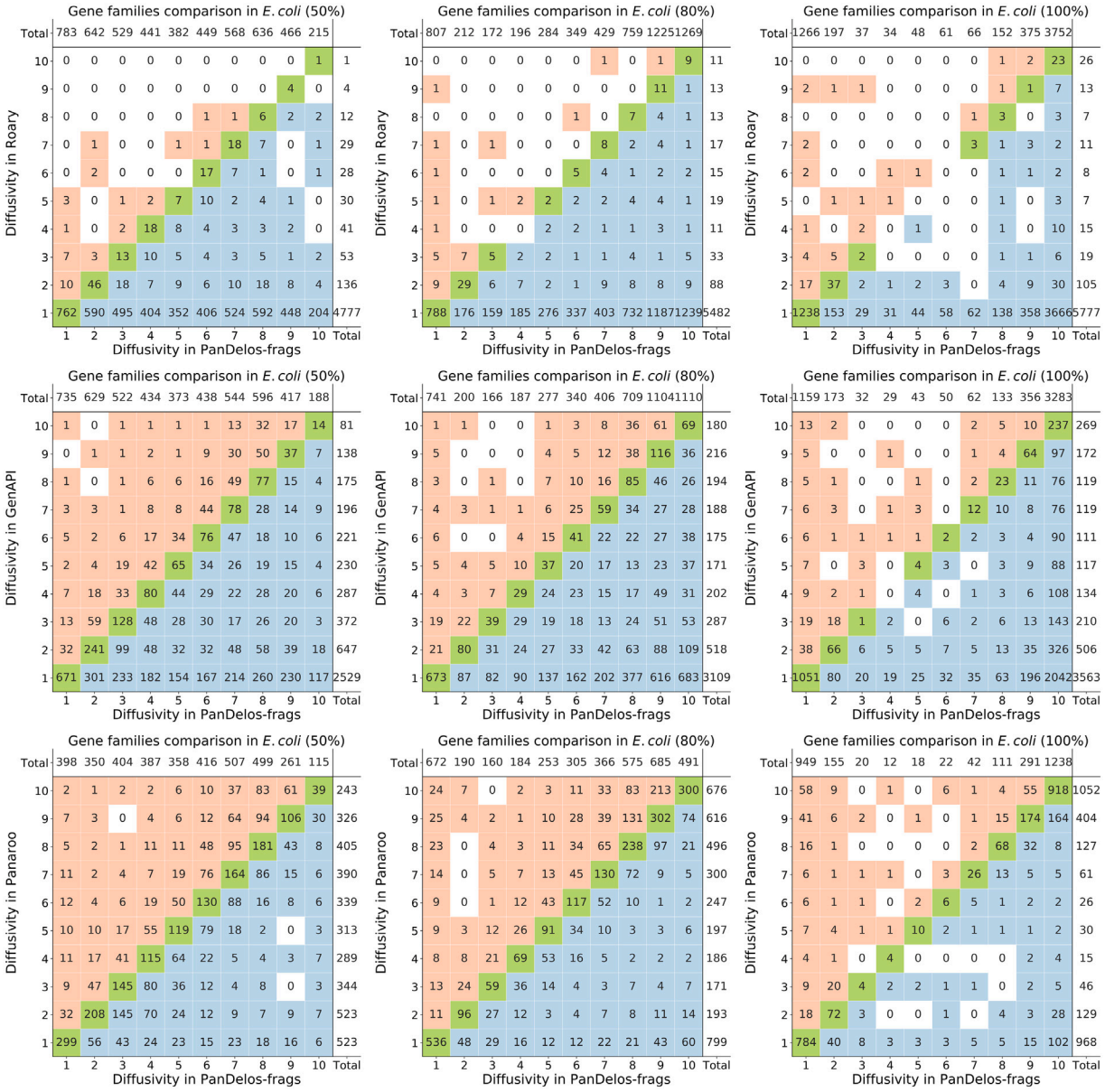


Fig. A.2. Diffusivity concordance between *PanDelos-frags* and other tools in *E. coli* synthetic data.

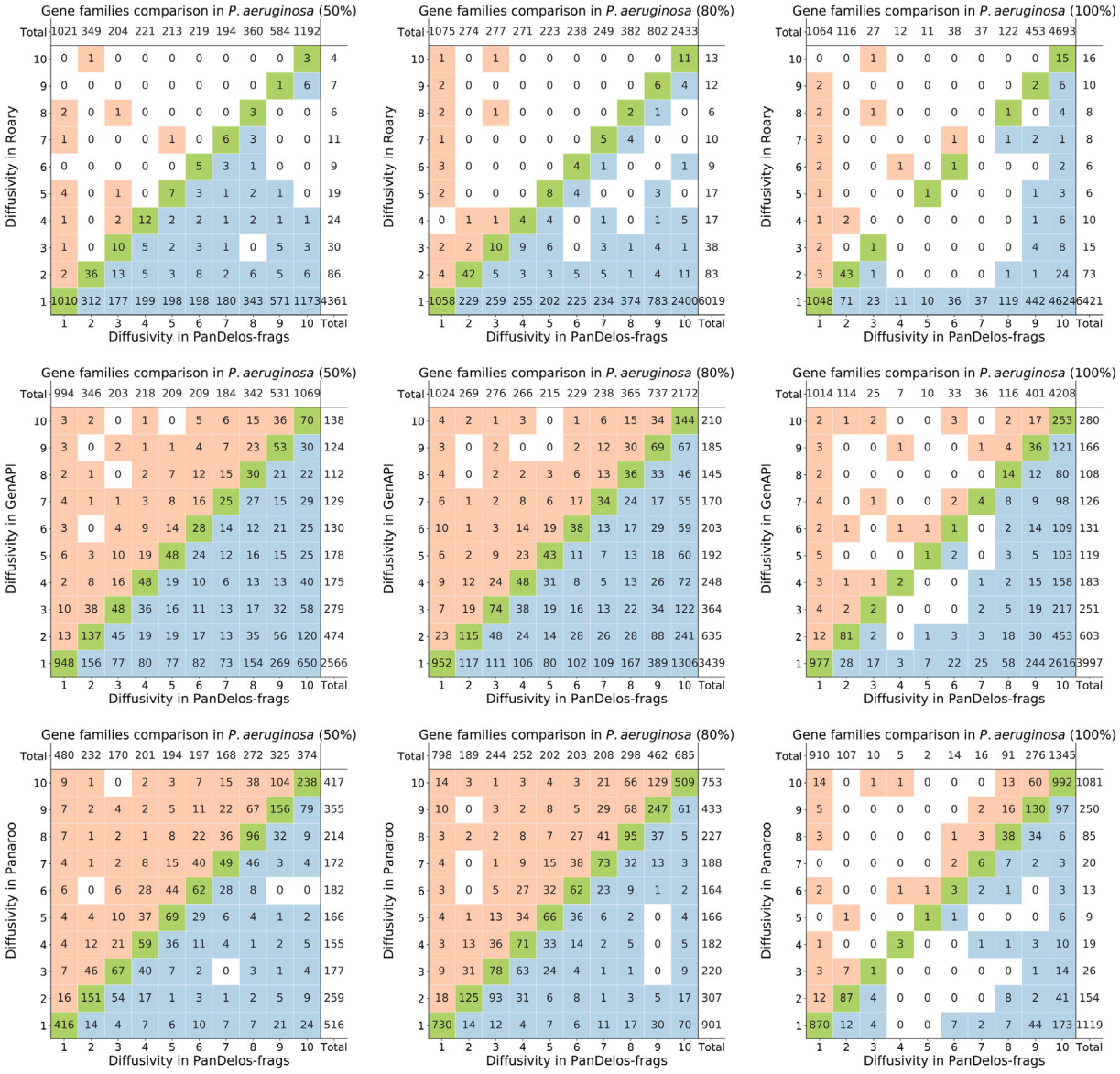


Fig. A.3. Diffusivity concordance between *PanDelos-frags* and other tools in *P. aeruginosa* synthetic data.

References

[1] H. Tettelin, V. Masignani, M.J. Cieslewicz, C. Donati, D. Medini, N.L. Ward, S.V. Angiuoli, J. Crabtree, A.L. Jones, A.S. Durkin, et al., Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome", Proc. Natl. Acad. Sci. 102 (39) (2005) 13950–13955.

[2] H. Anani, R. Zgheib, I. Hasni, D. Raoult, P.-E. Fournier, Interest of bacterial pangenome analyses in clinical microbiology, Microb. Pathog. 149 (2020) 104275.

[3] D. Serruto, L. Serino, V. Masignani, M. Pizza, Genome-based approaches to develop vaccines against bacterial pathogens, Vaccine 27 (25–26) (2009) 3245–3250.

[4] A. Muzzi, V. Masignani, R. Rappuoli, The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials, Drug Discov. Today 12 (11–12) (2007) 429–439.

[5] D. Medini, C. Donati, R. Rappuoli, H. Tettelin, The pangenome: a data-driven discovery in biology, Pangenome Diversity Dyn. Evol. Genomes (2020) 3–20.

[6] A.J. Page, C.A. Cummins, M. Hunt, V.K. Wong, S. Reuter, M.T. Holden, M. Fookes, D. Falush, J.A. Keane, J. Parkhill, Roary: rapid large-scale prokaryote pan genome analysis, Bioinformatics 31 (22) (2015) 3691–3693.

[7] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (23) (2012) 3150–3152.

[8] A.J. Enright, S. Van Dongen, C.A. Ouzounis, An efficient algorithm for large-scale detection of protein families, Nucleic Acids Res. 30 (7) (2002) 1575–1584.

[9] V. Bonnici, R. Giugno, V. Manca, PanDelos: a dictionary-based method for pan-genome content discovery, BMC Bioinform. 19 (15) (2018) 47–59.

[10] V. Bonnici, E. Maresi, R. Giugno, Challenges in gene-oriented approaches for pangenome content discovery, Brief. Bioinform. 22 (3) (2021) bbaa198.

[11] J.L. Klassen, C.R. Currie, Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation, BMC Genomics 13 (1) (2012) 1–11.

[12] H. Derakhshani, S.P. Bernier, V.A. Marko, M.G. Surette, Completion of draft bacterial genomes by long-read sequencing of synthetic genomic pools, BMC Genomics 21 (1) (2020) 1–11.

[13] P. Zhang, D. Jiang, Y. Wang, X. Yao, Y. Luo, Z. Yang, Comparison of de novo assembly strategies for bacterial genomes, Int. J. Mol. Sci. 22 (14) (2021) 7668.

[14] E. Altermann, H.E. Tegetmeyer, R.M. Chanyi, The evolution of bacterial genome assemblies—Where do we need to go next, Microbiome Res. Rep. 1 (2) (2022) 15.

[15] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, et al., Extensive unexplored human microbiome diversity resource extensive unexplored human microbiome diversity revealed by over 150 000 genomes from metagenomes spanning age, geography, and lifestyle, Cell 176 (2019) 649–662.

[16] C.L. Brown, I.M. Keenum, D. Dai, L. Zhang, P.J. Vikesland, A. Pruden, Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes, Sci. Rep. 11 (1) (2021) 3753.

[17] E.G. Barbosa, F.F. Aburjaile, R.T. Ramos, A.R. Carneiro, Y. Le Loir, J. Baumbach, A. Miyoshi, A. Silva, V. Azevedo, Value of a newly sequenced bacterial genome, World J. Biol. Chem. 5 (2) (2014) 161.

[18] L. Rouli, V. Merhej, P.-E. Fournier, D. Raoult, The bacterial pangenome as a new tool for analysing pathogenic bacteria, New Microbes New Infect. 7 (2015) 72–85.

[19] J. Qin, R. Li, J. Raes, M. Arumugam, K.S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, et al., A human gut microbial gene catalogue established by metagenomic sequencing, Nature 464 (7285) (2010) 59–65.

[20] T.H.M.P. Consortium, Structure, function and diversity of the healthy human microbiome, Nature 486 (7402) (2012) 207–214.

[21] C. Quince, A.W. Walker, J.T. Simpson, N.J. Loman, N. Segata, Shotgun metagenomics, from sampling to analysis, Nature Biotechnol. 35 (9) (2017) 833–844.

[22] D.R. Utter, G.G. Borisy, A.M. Eren, C.M. Cavanaugh, J.L. Mark Welch, Meta-pangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity, Genome Biol. 21 (1) (2020) 1–25.

[23] M. Gabrielaite, R.L. Marvig, GenAPI: a tool for gene absence-presence identification in fragmented bacterial genome sequences, BMC Bioinform. 21 (1) (2020) 1–8.

[24] A. Veras, F. Araujo, K. Pinheiro, L. Guimarães, V. Azevedo, S. Soares, A. da Costa da Silva, R. Ramos, Pan4Draft: a computational tool to improve the accuracy of pan-genomic analysis using draft genomes, Sci. Rep. 8 (1) (2018) 1–8.

[25] G. Tonkin-Hill, N. MacAlasdair, C. Ruis, A. Weimann, G. Horesh, J.A. Lees, R.A. Gladstone, S. Lo, C. Beaudoin, R.A. Floto, S.D. Frost, J. Corander, S.D. Bentley, J. Parkhill, Producing polished prokaryotic pangenomes with the panaroo pipeline, Genome Biol. 21 (1) (2020) 1–21.

[26] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477.

[27] G. Tamazian, P. Dobrynin, K. Krasheninnikova, A. Komissarov, K.-P. Koepfli, S.J. O'Brien, Chromosomer: a reference-based genome arrangement tool for producing draft chromosome sequences, Gigascience 5 (1) (2016) s13742–016.

[28] H. Ochman, J.G. Lawrence, E.A. Groisman, Lateral gene transfer and the nature of bacterial innovation, Nature 405 (6784) (2000) 299–304.

[29] V. Bonnici, R. Giugno, PANPROVA: pangenomic prokaryotic evolution of full assemblies, Bioinformatics 38 (9) (2022) 2631–2632.

[30] M. Lothaire, Applied Combinatorics on Words, Cambridge University Press, 2005.

[31] G. Rozenberg, A. Salomaa, Handbook of Formal Languages (Vol 1), Springer Nature, 1997.

[32] J. Percus, Mathematics of Genome Analysis, Cambridge University Press, 2007.

[33] A. Castellini, G. Franco, V. Manca, A dictionary based informational genome analysis, BMC Genomics 13 (1) (2012) 1–14.

[34] V. Bonnici, G. Franco, V. Manca, Spectral concepts in genome informational analysis, Theoret. Comput. Sci. 894 (2021) 23–30.

[35] V. Bonnici, A. Cracco, G. Franco, A k-mer based sequence similarity for pangenomic analyses, in: Machine Learning, Optimization, and Data Science: LOD 2021, Revised Selected Papers, 2022, pp. 31–44.

[36] J.P. Demuth, M.W. Hahn, The life and death of gene families, Bioessays 31 (1) (2009) 29–39.

[37] S.M. Soucy, J. Huang, J.P. Gogarten, Horizontal gene transfer: building the web of life, Nature Rev. Genet. 16 (8) (2015) 472–482.

[38] C. Webber, C.P. Ponting, Genes and homology, Curr. Biol. 14 (9) (2004) R332–R333.

[39] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications, BMC Bioinform. 10 (2009) 1–9.

[40] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows–Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760.

[41] D. Hyatt, G.-L. Chen, P.F. LoCascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinform. 11 (1) (2010) 1–11.

[42] V. Bonnici, V. Manca, Informational laws of genome structures, Sci. Rep. 6 (1) (2016) 1–10.

[43] M. Girvan, M.E. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. 99 (12) (2002) 7821–7826.

[44] T. Li, Y. Yin, Critical assessment of pan-genomic analysis of metagenome-assembled genomes, Brief. Bioinform. 23 (6) (2022) bbac413.

[45] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (4) (2013) 772–780.

[46] M.N. Price, P.S. Dehal, A.P. Arkin, FastTree: computing large minimum evolution trees with profiles instead of a distance matrix, Mol. Biol. Evol. 26 (7) (2009) 1641–1650.

[47] S. Tavaré, Some probabilistic and statistical problems in the analysis of DNA sequences, Lect. Math. Life Sci. (Am. Math. Soc.) 17 (1986) 57–86.

[48] D.F. Robinson, L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. 53 (1–2) (1981) 131–147.

[49] J. Huerta-Cepas, F. Serra, P. Bork, ETE 3: reconstruction, analysis, and visualization of phylogenomic data, Mol. Biol. Evol. 33 (6) (2016) 1635–1638.

[50] E.J. Stewart, Growing unculturable bacteria, J. Bacteriol. 194 (16) (2012) 4151–4160.

[51] Y. Zhou, M. Liu, J. Yang, Recovering metagenome-assembled genomes from shotgun metagenomic sequencing data: Methods, applications, challenges, and opportunities, Microbiol. Res. (2022) 127023.

[52] R.M. Bowers, N.C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. Reddy, F. Schulz, J. Jarett, A.R. Rivers, E.A. Eloe-Fadrosh, et al., Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea, Nature Biotechnol. 35 (8) (2017) 725–731.

[53] A. Meziti, L.M. Rodriguez-R, J.K. Hatt, A. Peña-Gonzalez, K. Levy, K.T. Konstantinidis, The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: Insights from comparing MAGs against isolate genomes derived from the same fecal sample, Appl. Environ. Microbiol. 87 (2021).

[54] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, et al., The SGB collection, 2019, URL https://opendata.lifebit.ai/table/SGB.

[55] B. Buchfink, C. Xie, D.H. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (1) (2015) 59–60.