





Towards a better understanding of the bacterial pan-genome

BIOOPEN 2021 – POST-CONFERENCE ARTICLE

DAWID GMITER , SYLWIA NAWROT, ILONA PACAK, KATARZYNA ZEGADŁO
WIEŚLAW KACA 

Jan Kochanowski University of Kielce, Faculty of Natural Science, Institute of Biology, Division of Microbiology and Parasitology, Uniwersytecka 7, 25-460 Kielce, Poland
E-mail: dawid.gmiter@ujk.edu.pl

ABSTRACT

The bacterial pan-genome is a relatively new concept that refers to the number of genes observed in a given set of bacterial genome sequences, either at the intra- or inter-species level. Determining the pan-genome of a given species of bacteria using a large number of strains allows one to compare multiple genes and to determine evolutionary links between isolates. This information can help to determine population structure, diversity in terms of prevalence in a given environment and pathogenicity of microorganisms. Within this review, we explain the most important issues related to pan-genome studies. We also include a brief description of some selected bacterial pan-genomes. Finally, we propose an easy-to-perform workflow to study bacterial pan-genomes that will facilitate non-experts in a pan-genome-based investigation.

KEYWORDS: pan-genome, bacterial pan-genome, genome comparison, Roary workflow

Introduction

A bacterial pan-genome can be defined as the total number of genes observed in a certain group of microorganisms. The pan-genome of individual bacterial species is most often analysed, but some studies focus on broader groups of microorganisms, for example, a genus. The term pan-genome was proposed in 2005 by Tettelin and co-workers (Guimarães *et al.* 2015, Mira *et al.* 2010). Next, Rouli *et al.* (2015) clearly defined a pan-genome, or supragenome, as ‘the entire gene repertoire of the study group’. Pan-genome research has become pos-

sible due to the development of next-generation sequencing (NGS) technology, which has allowed the sequencing of bacterial genomes (Guimarães *et al.* 2015, Rouli *et al.* 2015).

On the other hand, the idea that the genome of individual strains may differ significantly within a species was born in the 1980s, when using *Escherichia coli* and the technique of electrophoresis in a variable pulse field revealed that the size of the genome of strains of this species was between 4.5 and 5.5 mega base pairs

(Mbp). A relationship was also observed between the genome size and strain differentiation using the multilocus enzyme electrophoresis (MLEE) method (Mira *et al.* 2010). This method allows bacteria to be differentiated based on the relative migration rate pattern of a large group of intracellular enzymes. The different patterns result from mutations in the genes coding for these enzymes, which change the amino acid sequence of the proteins (Caierão *et al.* 2016).

Types of genes in the pan-genome

The pan-genome is a pool of genes that may occur with different frequency among the studied group of microorganisms (Costa *et al.* 2020; Guimarães *et al.* 2015). Based on the frequency of their occurrence, genes are assigned to one of three groups. The genes found in the genomes of all the microorganisms analysed are called core genes. The genes found in only some of the genomes studied are termed accessory genes. The third group of genes that make up the pan-genome are unique genes, the presence of which is found only in single genomes (see Figure 1). Depending on the scope of the analysis, unique genes may be strain specific (in the pan-genome analysis of one species) or species

specific (when the analysis is carried out at the inter-species level) (Costa *et al.* 2020; Guimarães *et al.* 2015; Mira *et al.* 2010; Rouli *et al.* 2015).

The genes of the aforementioned subtype found during the pan-genome analysis play different roles in the development of microorganisms. It is believed that the core is made of genes responsible for the basic functions of the bacterial cells, including housekeeping, cell division (replication) and homeostasis. Meanwhile, accessory and unique genes play a supporting role in relation to core genes. These genes are related to the growth environment of a bacterial species as well as the virulence of pathogenic bacteria. These genes are acquired through horizontal gene transfer, a phenomenon that can confer an adaptive advantage, and their presence can be a factor that facilitates development relative to strains lacking them (Costa *et al.* 2020; Mira *et al.* 2010; Rouli *et al.* 2015).

Types of pan-genomes

After determining the number of genes in the pan-genome and assigning them to individual subtypes, the next step to characterise it is to determine the ratio of the number of core genes to others. In

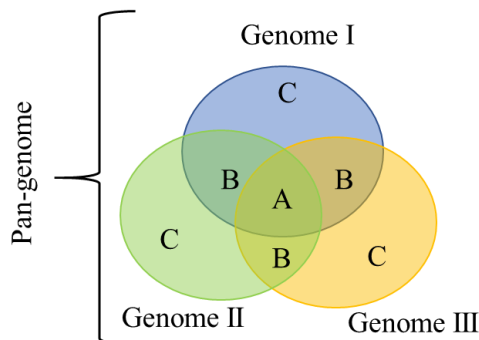


Figure 1. A Venn diagram that represents the three types of genes present in the pan-genome: A – core genes present in all analysed genomes; B – accessory genes, present only in some of the genomes studied; C – unique genes, characterising individual genomes.

addition, the number of new unique genes is observed when adding more genomes to the analysed pool. Based on the results of the second analysis, pan-genomes are divided into open and closed. An open pan-genome refers to when another genome added to the analysed pool increases the number of unique genes. Conversely, when adding more genomes does not increase the pool of unique genes, the pan-genome is termed closed (Rouli *et al.* 2015).

A simple way to determine whether the pan-genome is closed or open is to construct rarefaction curves. This tool is normally used by ecologists to determine graphically when further sampling would not increase the number of newly identified species. Using similar approaches, genes are counted as additional genomic sequences are added to the analysis. The results are presented in a graph of the total number of genes in a pan-genome versus the number of analysed genomes. If the curve reaches a plateau, the pan-genome is termed closed. Open pan-genomes are characterised by the fact that with each genome added, the number of genes increases at a constant rate (see Figure 2) (Mira *et al.* 2010).

Another tool is to apply Heap's law to determine the openness of an analysed pan-genome. Heap's law describes the

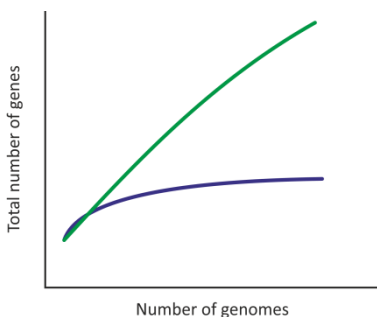


Figure 2. Rarefaction curves for open (green) and closed (blue) pan-genomes.

number of distinct words in a document as a function of the document length. It is represented by the formula $n = k \times N^{-\alpha}$.

In a pan-genome studies:

- n is the expected number of genes for a given number of genomes,
- N is the number of genomes and
- k and α are free parameters that are determined empirically.

According to Heap's law, when α is > 1 , the pan-genome is considered to be closed, and when α is < 1 , the pan-genome is considered to be open (Guimarães *et al.* 2015).

Examples of pan-genomes of selected bacteria

Within this paper we describe briefly some pan-genome studies of selected gram-negative and gram-positive bacteria. The presented information shows that utilisation of the pan-genome approach in microbiology could extend our understanding of molecular aspects of bacterial diversity, evolution and pathogenesis.

E. coli is an important urinary track pathogen; strain ST131 is becoming a serious problem due to its multi-drug resistance. Decano and Downing (2019) studied a cohort of 4,071 genomes of the ST131 strain to investigate the genetic diversity of the group based on the core and accessory genes. Their analysis indicated that the average number of genes in the pan-genome of ST131 increased as more genomes were added, indicating the open nature of the pan-genome of the entire collection. The authors found 26,479 genes, of which 3,712 genes present in all isolates formed the core. The tested strains formed three clades: A, B and C; these classifications were based on phylogenetic analyses of the ST131 *fimH* gene. Clade C was the largest group. The pan-genome was also tested as an independent sets of genomes. In all sets, the pan-genome was described as open. This comparison revealed inter-

clade but not intra-clade accessory genome divergence, which might result from ecological specialisation of the strains (Decano and Downing 2019).

In another study, the authors employed inter-species pan-genome analysis to compare the pan-genomes of *E. coli* and *Shigella* spp. to *Salmonella enterica*. This analysis indicated that *Shigella* should not be considered an independent genus – based on pan-genome diversity – because, after examining its genome, it turned out that it did not contain any specific genes not present in *E. coli*. This would mean that there are no barriers in the gene pool between the species. At the same time, *E. coli* and *S. enterica* maintained stable, species-restricted gene pools, despite intensive horizontal gene transfer between the species. Importantly, pan-genome analysis allowed the researchers to complement the current classification of the studied species, providing a new perspective to the understanding of bacterial evolution. Consequently, it can allow researchers to understand the interactions between strains in the environment, to track the evolution of individual lines, to predict the probability of certain diseases after infection with a given microorganism and to improve the treatment process and diagnostic tools (Gordienko *et al.* 2013).

Pseudomonas aeruginosa is the third most frequent opportunistic pathogen found in hospitals. The bacterium is resistant to most classes of antibiotics and causes major infections in immunocompromised patients, including individuals with cystic fibrosis. Understanding evolutionary processes and molecular mechanisms of *P. aeruginosa* on the pan-genome scale could help to explain the ineffectiveness of the designed vaccines against this pathogen and to understand the mechanisms by which *P. aeruginosa* strains avoid the human immune system.

Recently, Freschi *et al.* (2018) used 1,311 strains to update the pan-genome of this opportunistic pathogen. This approach allowed them to define the structure of the population and to determine the number of primary genes and to assign functions to them. Based on their data, the *P. aeruginosa* pan-genome comprises 54,272 genes: 665 core genes, 26,420 flexible genes and 27,187 unique genes. Overall, 33.1% of pan-genomic genes have not been assigned a function, and core genes account for only 1% of the total pan-genome. These findings demonstrated that determining pan-genomes or updating existing ones with larger data sets provides a better understanding of the population structure and evolution of microbes (Freschi *et al.* 2019). In contrast to the above-mentioned study, a previous analysis of 181 *P. aeruginosa* genomes showed that the pan-genome comprises 2,503 core genes (15%), 9,108 additional genes (54%) and 5,209 unique genes and is closed (Mosquera-Rendón *et al.* 2016).

Burkholderia cepacia is a gram-negative bacterium that does not produce spores and cannot ferment glucose. It is common in humid environments (e.g. around plant roots) and is a common cause of opportunistic nosocomial infections, with cystic fibrosis patients being the most vulnerable to infection (Mahenthiralingam *et al.* 2008). Recombination and positive selection are two fundamental evolutionary forces that can be studied by performing comparative genomic analyses. *B. cepacia* species are very difficult to distinguish genotypically and phenotypically due to their high level of recombination, which is strongly supported by about 5.8% of the basic orthologous genes, while 1.1% of these genes support positive selection (Zhou *et al.* 2020). This problem can be solved by using combined methods that ensure proper recognition of species, even those

that are closely related to each other. It is suggested to combine the core-gene-based phylogenetic study with the analysis of digital DNA-DNA hybridization and Average Nucleotide Identity (dDDH/ANI) clusters and the formation of species trees (Jin *et al.* 2020).

The analysis of the genomes of bacteria isolated from cystic fibrosis patients is often a source of valuable information about changes in genomes under the influence of the host's immune system and the therapies used. Phylogenetic analysis of 2,148 orthologous gene clusters from *Burkholderia cenocepacia* isolates collected from 16 cystic fibrosis patients confirmed compliance with patient-specific clades and allowed the observation of pathogen transmission among patients (there was evidence of shared clonal lines), as well as frequent repeated loss of genes and the entire chromosome III (Lee *et al.* 2017). Based on the above-mentioned studies, the analysis of the *Burkholderia* spp. genome has contributed to a more in-depth understanding of the phylogenetic tree of these microorganisms, and thus to the development of more effective treatment methods and improved diagnosis of infections.

Staphylococcus aureus often causes hospital- and community-acquired infections that, due to the presence of methicillin-resistant strains, are very difficult to treat and can lead to sepsis and death (Guo *et al.* 2020). Strain antibiotic resistance is one of the major problems of modern medicine; this phenomenon may be better understood by examining the evolutionary pathway and origin of resistance genes in common bacterial pathogens. Indeed, an analysis comparing 152 fully sequenced *S. aureus* strains with 7,529 reference genomes of other bacteria found that 55% of known resistance genes for this bacterium belong to its accessory genome and 27%

of them were located in Staphylococcal Cassette Chromosome *mec* (SCC*mec*), and in most cases they were acquired laterally from other species (John *et al.* 2019). *S. aureus* co-exists on the skin, throat and nose alongside *Staphylococcus epidermidis*. Approximately half of their genomes are shared, and homologous recombination between the species is rare. However, they contain a significant proportion of interspecific mobile elements, which are genes responsible for metal detoxification, methicillin resistance (SCC*mec* island) and are associated with the pathogenicity island (SaPin1) (Méric *et al.* 2015). Genome sequencing allows for the analysis of the structural and evolutionary changes of microorganisms over the years. The analysis of *S. aureus* USA300, which represents a line of methicillin-resistant *S. aureus* strains in the United States, revealed that pan-genome evolved from 2004 to 2010 (Jamroz *et al.* 2016).

Staphylococcus lugdunensis has unique properties among *Staphylococcus* and occurs on the human skin. This coagulase-negative microorganism can produce various virulence factors and has the ability to cause severe infections, especially in hospital conditions. Interestingly, it is easily treated with antimicrobials, a feature that is quite unheard of for this type of bacteria. Phylogenetic studies of the *S. lugdunensis* genome have shown its high conservation in terms of antibiotic sensitivity and extremely rare methicillin resistance even in hospital conditions, which distinguishes it from all other staphylococci. To investigate the *S. lugdunensis* genome, researchers used 16 different strains from Europe, Asia and North America, isolated between 1988 and 2015. They found *S. lugdunensis* has a very closed pan-genome with a fairly limited number of new genes. This is an infrequent feature for *Staphylococcus* spp., which

have an open pan-genome (Argemi *et al.* 2018).

Bacillus cereus sensu lato is a diverse group of bacteria containing many species found in different environments and exhibiting a variety of phenotypes. Many species of this genus have medical or agricultural significance. The pan-genome of *B. cereus s.l.* consists of approximately 60,000 genes, 598 of which are core genes. The accessory pan-genome consists of 32,324 genes, of which 27,067 are unique. Gene analyses indicate the presence of open pan-genome for *B. cereus s.l.* (Bazinet 2017).

Pan-genome study workflow

With the increased number of studies focused on bacterial pan-genomes, new *in silico* tools have been developed. Most of them are command-line-based software programs that allow testing the pan-genome-based diversity. The majority of the software programs have been reviewed by Guimarães *et al.* (2015). The major disadvantages of the software programs is that they require computing skills, which might be problematic for users who cannot code. Therefore, within this chapter we propose a simple workflow that allows performing solid pan-genomic and phylogenetic investigation of bacterial species of interest (Figure 3). This workflow should be used as a guide for beginners. It reviews tools and briefly describe their usage, but we advise that one should expand their computational skills.

For the purposes of this workflow, we recommend installing the latest version of Ubuntu, which is a Linux program based on Debian recommended for beginners (Möller *et al.* 2010). Ubuntu can be installed as a dual boot with Windows or on a Windows 10 machine by downloading the Ubuntu application (Lloyd 2018).

STEP 1: pan-genome

This workflow is based on the use of Roary, a standalone pipeline allowing the calculation of a pan-genome. The installation is relatively easy but there are some requirements (Page *et al.* 2015; Sitto and Battistuzzi 2020). For beginners, we recommend to install Conda first and to work with Roary as a Bioconda package. To install conda, run the following command in the Linux Terminal (Grüning *et al.* 2018):

```
curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
sh Miniconda3-latest-Linux-x86_64.sh
```

Then one need to set up channels:

```
conda config --add channels r
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
```

And install Roary:

```
conda install roary
```

Roary requires the annotated assemblies in the GFF3 format and there are a few steps required to generate such files. We recommend that the GFF3 files

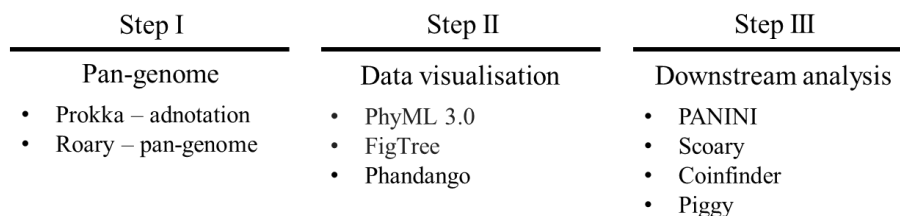


Figure 3. Scheme for a pan-genome analysis workflow.

be generated by Prokka (Seemann 2014, Page *et al.* 2015). The use of complete and finished genome sequences will give the best annotation results, but it is expected that the typical input will be a set of scaffold sequences. Prokka can be easily installed and used locally. Prokka is also available online within the Galaxy server (<https://usegalaxy.org/>); this approach might facilitate analysis in the case of beginners.

After installation of Roary, the user must optimise just a few parameters, but some additional options might be considered for more in-depth analysis. The simple command `roary *.gff` will run Roary with default parameters. The basic options to optimise the program are presented in Figure 4.

The major advantage of Roary is it is relatively simple to use and a pan-genome of even thousands of samples can be analysed on a standard desktop PC. In addition to the basic performance, Roary offers the `query_pan_genome` scripts, which perform set operations on

the pan-genome to see the gene differences between groups of isolates (Sitto and Battistuzzi 2020).

Roary generates a set of output files, of which the most important and useful are (Sitto and Battistuzzi 2020):

- `summary_statistics.txt` – a text file that summarises the number of genes founded in the studied data, where genes are grouped into core, soft-core, shell and cloud based on the frequency within the studied genomes;
- `gene_presence_absence.csv`;
- `gene_presence_absence.Rtab`;
- `accessory_binary_genes.fa.newick` – a maximum likelihood tree generated based on the gene presence absence; and
- `core_gene_alignment.aln` – a file that contains the alignment of all core genes.

These files could be used for data visualisation as well as for downstream analysis with additional software.

```
Usage: roary [options] *.gff

Options: -p INT      number of threads [1]
         -o STR      clusters output filename [clustered_proteins]
         -f STR      output directory [.]
         -e          create a multiFASTA alignment of core genes using PRANK
         -n          fast core gene alignment with MAFFT, use with -e
         -i          minimum percentage identity for blastp [95]
         -cd FLOAT  percentage of isolates a gene must be in to be core [99]
         -qc        generate QC report with Kraken
         -k STR      path to Kraken database for QC, use with -qc
         -a          check dependancies and print versions
         -b STR      blastp executable [blastp]
         -c STR      mcl executable [mcl]
         -d STR      mcxdeblast executable [mcxdeblast]
         -g INT      maximum number of clusters [50000]
         -m STR      makeblastdb executable [makeblastdb]
         -r          create R plots, requires R and ggplot2
         -s          dont split paralogs
         -t INT      translation table [11]
         -ap        allow paralogs in core alignment
         -z          dont delete intermediate files
         -v          verbose output to STDOUT
         -w          print version and exit
         -y          add gene inference information to spreadsheet, doesnt work with -e
         -iv STR    Change the MCL inflation value [1.5]
         -h          this help message

Example: Quickly generate a core gene alignment using 8 threads
roary -e --mafft -p 8 *.gff
```

Figure 4. List of Roary options available with the command `roary -h`.

STEP 2: data visualisation

Data obtained from Roary analysis could be easily visualised using local and online applications. The `core_gene_alignment.aln` file could be used to generate the phylogenetic tree based on core gene single nucleotide polymorphisms (SNPs). This tree can be prepared by using the online version of PhyML 3.0 (<http://www.atgc-montpellier.fr/phyml/>) (Guindon *et al.* 2010), but it needs to be converted to the PHYLIP format. Trees based on `accessory_binary_genes.fa`, `newick` and core genes can be visualised and modified using FigTree (<http://tree.bio.ed.ac.uk/software/figtree/>) (Rambaut 2013). FigTree is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures. It can be used on a desktop PC running Mac, Windows or Linux OS.

The `gene_presence_absence.csv` file together with any phylogenetic tree can be display using the Phandango website. One can simply drag and drop the Roary results into the web browser and then interactively play around with the data. Phandango can also be used to visualise the metadata for the samples, but it needs to be collected within a single file (csv format file). The sample IDs must be the same for all types of data used, or Phandango will not compile the data accurately (Hadfield *et al.* 2018).

The R script implemented in Roary allows one to generate two additional graphs: rarefaction curves, to conclude whether the analysed pan-genome is open or closed, and a plot of the number of new genes as a function of the number of studied sequences (Page *et al.* 2015; Sitto and Battistuzzi 2020).

STEP 3: downstream analysis

We propose the use of four additional software programs that will expand the pan-genome results obtained with Roary.

These programs are the PANINI, Scoary, Coinfinder and Piggy pipelines.

PANINI

When the number of samples in a pan-genome analysis exceeds 100, one can use PANINI (Pangenome Neighbour Identification for Bacterial Populations). PANINI is a web-based tool that allows a user to identify the neighbours for each isolate in a data set. The tool is integrated with the Microreact platform for rapid online visualisation and exploration of pan-genomes, together with relevant epidemiological, geographical, temporal and other metadata (Abudahab *et al.* 2019).

The tool requires three types of input data: the `gene_presence_absence.Rtab` file (Roary output), any phylogenetic tree and a file containing the metadata of interest. After introducing the `gene_presence_absence.Rtab` file (drag and drop), the network file (DOT format) can be sent to Microreact when the other files are uploaded. The result will be displayed online, and Microreact allows a user to browse and to select data of interest (Abudahab *et al.* 2019).

The file containing metadata should be prepared in the csv format. It could include any relevant data. The geographical distribution of isolates could be displayed against a world map; this information must be specified by latitude and longitude columns. The file requires an ID column and each ID must be the same as in the other input files and must be unique. The metadata will be displayed against a phylogenetic tree comprised of coloured dots (default mode). The colour can be defined by user by the additional column `maned: data_name_colour`. The colour must be defined by hex triplet number (Abudahab *et al.* 2019). The use of PANINI is intuitive and well described by the tool's authors, and there is also a video walkthrough.

Scoary

Roary output data could be used for a pan-genome wide association study. For this purpose, we recommend Scoary (Brynildsrud *et al.* 2016). Scoary is designed to take the `gene_presence_absence.csv` file from Roary and a traits file created by the user. It calculates the associations between all genes in the accessory genome and the traits. A traits file is a binary table (csv format) in which individual isolates are described with 1 or 0, sequentially with or without a trait. For example, when analysing the relationship between genetic diversity and the source of isolation, if the strain was isolated from a particular source of interest, the value is 1, while the remaining isolates receive the value 0 (Brynildsrud *et al.* 2016). The easiest way to install Scoary is with the pip package manager:

```
pip install scoary
```

The use of Scoary requires the following basic command:

```
scoary -g <gene_presence_absence.csv> -t <traits.csv>
```

The user can also modify the parameters by additional flags, as shown at Figure 5.

Scoary outputs a single csv file per trait in the traits file. The file contains a list of genes with additional statistical characteristics. The output data need to be filtered manually. Candidate genes

can be determined to be significantly related to a trait if they have achieved a ‘naïve’ p value < 0.05, a Benjamini-Hochberg corrected p value < 0.05 and an empirical p value < 0.05. A particular gene is considered to be positively related to a trait when the odds ratio is > 1 and to be negatively related when the odds ratio is < 1 (Espadinha *et al.* 2019, Touchon *et al.* 2020).

Coinfinder

Coinfinder allows one to assess the occurrence of interactions between genes in the pan-genome. The software tests for the occurrence of gene association and dissociation events among the accessory genes. The algorithm on which the program is based assumes the rejection of core genes and strongly unique genes to increase the precision of the analysis. The application uses the output (the `gene_presence_absence.csv` table) generated by Roars (Whelan *et al.* 2020). We recommend installing Coinfinder with Conda:

```
conda install -c defaults -c bioconda -c conda-forge coinfinder
```

Coinfinder requires gene information (the `gene_presence_absence.csv` table) and a phylogeny as input. The phylogeny should be Newick formatted; we recommend using the core SNP-based phylogeny from the Roary output (Whelan *et al.* 2020). To run Coinfinder with default parameters, use the following line:

```
usage: scoary [-h] [-t TRAITS] [-g GENES] [-n NEWICKTREE] [-s START_COL]
             [--delimiter DELIMITER] [--restrict_to RESTRICT_TO] [--outdir OUTDIR] [-u]
             [-p P_VALUE_CUTOFF [P_VALUE_CUTOFF ...]]
             [-c [{I,B,BH,PW,EPW,P} [{I,B,BH,PW,EPW,P} ...]]] [-m MAX_HITS]
             [--include_input_columns GRABCOLS] [-w] [--no-time] [-e PERMUTE]
             [--no_pairwise] [--collapse] [--threads THREADS] [--test]
             [--citation] [--version]

Scoary version 1.6.16 - Screen pan-genome for trait-associated variants
```

Figure 5. List of Scoary options available with the command `scoary -h`.

```

./cofinder [OPTIONS]
File input- specify either:
  -l or --input          The path to the gene_presence_absence.csv output from Roary
                        -or-
                        The path of the Alpha-to-Beta file with (alpha)(TAB)(beta)
                        set if -i is in the gene_presence_absence.csv format from Roary
  -I or --inputroary    Phylogeny of Betas in Newick format (required)
  -p or --phylogeny
Max mode (mandatory for coincidence analysis):
  -a or --associate     Overlap; identify groups that tend to associate/co-occur (default).
  -d or --dissociate   Separation; identify groups that tend to dissociate/avoid.
Significance- specify:
  -L or --level        Specify the significance level cutoff (default: 0.05)
Significance correction- specify:
  -n or --bonferroni   Bonferroni correction multiple correction (recommended & default)
  -n or --nocorrection No correction, use value as-is
  -c or --fraction     (Connectivity analysis only) Use fraction rather than p-value
Alternative hypothesis- specify:
  -g or --greater      Greater (recommended & default)
  -l or --less         Less
  -t or --twotailed   Two-tailed
Miscellaneous:
  -x or --num_cores   The number of cores to use (default: 2)
  -v or --verbose     Verbose output.
  -r or --filter      Permit filtering of saturated and low-abundance data.
  -U or --upfilthreshold Upper filter threshold for high-abundance data filtering (default: 1.0 i.e. any alpha in >=100% of betas.
  -F or --filthreshold Threshold for low-abundance data filtering (default: 0.05 i.e. any alpha in <=5% of betas.
  -q or --query       Query a specific gene.
  -T or --test        Runs the test cases and exits.
  -E or --all         Outputs all results, regardless of significance.
Output:
  -o or --output      The prefix of all output files (default: coincident).

```

Figure 6. List of Coinfinder options available with the command *cofinder -h*.

```

cofinder -i <gene information> [-I]
|-p <phylogeny> -o <output prefix>
[--associate|--dissociate]

```

One might also change some options (see Figure 6).

Coinfinder produces a number of output files, with the default prefix of `coincident_`, which have been well described by (Whelan *et al.* 2020). The tool identifies pairs of associating/dissociating genes that are clustered in components or sets of genes that are related to each other. In addition, the results obtained with Coinfinder can be visualised using the Gephi graphics program. The produced charts should be interpreted as follows. Individual genes are represented by individual points on the plot (nodes). The lines connecting these points (edges) indicate the presence of interactions between the genes of the studied genomes. The groups of genes for which the presence of statistically significant correlations were found are depicted with different colours; these groups are called components. Genes within a given component show an association or dissociation. Occasional relationships between genes from different components can also be observed (Whelan *et al.* 2020).

Piggy

The above-mentioned tools allow for a detailed description of a pan-genome. All are focused on genes: their distribution, interaction and importance. To gain better insight into the phylogeny of a particular species, one can use the Piggy pipeline. Piggy works similarly to Roary, except it is focused on the intergenomic regions (termed IGRs) rather than genes. Piggy also detects and specifies highly divergent ('switched') intergenic regions (IGRs) upstream of genes. Similarly to Roary, IGRs can be described as core, present in all samples, accessory or unique. Based on a core IGR alignment file, a user can create the phylogenetic tree. Therefore, the use of this tool allows a user to understand not only the gene-based phylogeny, but also phylogeny based on non-coding regions. This information could be useful to better understand the evolution of tested strains. On the other hand, this tool provides insight into regions of genome sequences that might play a regulatory role (Thorpe *et al.* 2018).

Piggy can be easily obtained from github (<https://github.com/>), by a simple command line:

```
Piggy - version 1.5
--in_dir|-i <STR> input folder [default - current folder]
--out_dir|-o <STR> output folder [default - current folder/piggy_out]
--roary_dir|-r <STR> folder where roary output is stored [required]
--threads|-t <INT> threads [default - 1]
--nuc_id|-n <INT> min percentage nucleotide identity [default - 90]
--len_id|-l <INT> min percentage length identity [default - 90]
--edges|-e <STR> keep IGRs at the edge of contigs [default - off]
--size|-s <STR> size of IGRs to extract [i-j] [default 30-1000]
--method|-m <STR> method for detecting switched IGRs [g - gene_pair, u - upstream] [default - g]
--R_plots|-R make R plots (requires R, Rscript, ggplot2, reshape2) [default - off]
--fast|-f fast mode (doesn't align IGRs or detect switched regions) [default - off]
--help|-h help
--version|-v version
```

Figure 7. List of Piggy options available with the command *piggy -h*.

```
git clone https://github.com/harry-
thorpe/piggy.git
```

After cloning the Piggy repository, its directory should be either added to user \$PATH or one can run Piggy by specifying its location in the terminal. For Piggy to work, Roary must be run first. The output folder produced by Roary is required as an input to Piggy. We recommend running Roary with the -s flag to keep paralogs together, so secondary Roary analysis can be performed (Thorpe *et al.* 2018).

To run Piggy, a user must specify the direction to files containing annotated assemblies in the GFF3 format, Roary output files and the output direction. The list of options are shown in Figure 7. Piggy produces several output files:

- cluster_intergenic_alignment_files,
- switched_region_alignment_files
and
- IGR_presence_absence.csv.

The IGR_presence_absence.csv file can be visualised with Phandango, as can the gene_presence_absence.csv file generated by Roary. Piggy also generates two additional graphs, one presenting gene and IGR accumulation curves, and a histogram showing the frequency of genes and IGR regions identified within the tested set of samples (Thorpe *et al.* 2018).

Conclusions

The bacterial pan-genome is a relatively new concept for microbial genomics, but in recent years the number of studies focused on its investigation has increased rapidly. This approach allows one to better understand the diversity and evolution of bacteria, both at the inter- and intra-species levels. The results might help to improve taxonomy and ultimately lead to the development of more specific and sensitive methods of bacteria identification. On the other hand, knowing the relationship of the pan-genomic diversity of bacteria isolated from different sources (and/or time points) can be used as a basis to design new therapeutics. Hence, it is worth developing the area of pan-genome research.

Within this paper, we have reviewed some in silico tools that could be used by beginners who have an interest in pan-genome investigation. Each of them covers important features of pan-genome studies, namely core genes and pan-genome-based phylogeny, pan-genome-level diversity of strains, pan-genome wide association study and, in contrast to gene-focused studies, the diversity of non-coding sequences. When used in combination, the reviewed tools allow solid pan-genome investigation of bacterial species of interest.

Acknowledgements

D. Gmitter received a PhD scholarship from the National Science Centre, Poland Grant No. 2019/32/T/NZ1/00515.

References

- Abudahab, K., Prada, J.M., Yang, Z., Bentley, S.D., Croucher, N.J., Corander, J., Aanensen, D.M. 2019. PANINI: pangenome neighbour identification for bacterial populations. *Microbial Genomics*, 5(4): e000220.
- Argemi, X., Matelska, D., Ginalski, K., Riegel, P., Hansmann, Y., Bloom, J., Pestel-Caron, M., Dahyot, S., Lebeurre, J., Prévost, G. 2018. Comparative genomic analysis of *Staphylococcus lugdunensis* shows a closed pan-genome and multiple barriers to horizontal gene transfer. *BMC Genomics*, 19(1): 1–16.
- Bazinnet, A.L. 2017. Pan-genome and phylogeny of *Bacillus cereus sensu lato*. *BMC Evolutionary Biology*, 17(1): 1–16.
- Brynildsrud, O., Bohlin, J., Scheffer, L., Eldholm, V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, 17(1): 1–9.
- Caierão, J., Paiva, J.A.C.D., Sampaio, J.L.M., da Silva, M.G., Santos, D.R. de S., Coelho, F.S., Fonseca, L. de S., Duarte, R.S., Armstrong, D.T., Regua-Mangia, A.H. 2016. Multilocus enzyme electrophoresis analysis of rapidly-growing mycobacteria: An alternative tool for identification and typing. *International Journal of Infectious Diseases*, 42: 11–16.
- Costa, S.S., Guimarães, L.C., Silva, A., Soares, S.C., Baraúna, R.A. 2020. First steps in the analysis of prokaryotic pan-genomes. *Bioinformatics and Biology Insights*, 14: 1–9.
- Decano, A.G., Downing, T. 2019. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Scientific Reports*, 9(1): 1–13.
- Espadinha, D., Sobral, R.G., Mendes, C.I., Méric, G., Sheppard, S.K., Carriço, J.A., Lencastre, H. de, Miragaia, M. 2019. Distinct phenotypic and genomic signatures underlie contrasting pathogenic potential of *Staphylococcus epidermidis* clonal lineages. *Frontiers in Microbiology*, 10: 1971.
- Freschi, L., Vincent, A.T., Jeukens, J., Emond-Rheault, J.G., Kukavica-Ibrulj, I., Dupont, M.J., Charette, S.J., Boyle, B., Levesque, R.C. 2019. The *Pseudomonas aeruginosa* Pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biology and Evolution*, 11(1): 109–120.
- Gordienko, E.N., Kazanov, M.D., Gelfand, M.S. 2013. Evolution of pan-genomes of *Escherichia coli*, *Shigella* spp., and *Salmonella enterica*. *Journal of Bacteriology*, 195(12): 2786–2792.
- Grüning, B., Dale, R., Sjödin, A., Chapman, B.A., Rowe, J., Tomkins-Tinch, C.H., Köster, J., The Bioconda Team. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, 15: 475–476.
- Guimarães, L.C., Benevides De Jesus, L., Vinícius, M., Viana, C., Silva, A., Thiago, R., Ramos, J., De, S., Soares, C., Azevedo, V. 2015. Inside the pan-genome – methods and software overview. *Current Genomics*, 16: 245–252.
- Guindon, S., Dufayard, J., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3): 307–321.
- Guo, Y., Song, G., Sun, M., Wang, J., Wang, Y. 2020. Prevalence and therapies of antibiotic-resistance in *Staphylococcus aureus*. *Frontiers in Cellular and Infection Microbiology*, 10: 107.
- Hadfield, J., Croucher, N.J., Goater, R.J., Abudahab, K., Aanensen, D.M., Harris, S.R. 2018. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, 34(2): 292–293.
- Jamrozny, D.M., Harris, S.R., Mohamed, N., Peacock, S.J., Tan, C.Y., Parkhill, J., Anderson, A.S., Holden, M.T.G. 2016. Pangenomic perspective on the evolution of the *Staphylococcus aureus* USA300 epidemic. *Microbial Genomics*, 2(5): e000058.
- Jin, Y., Zhou, J., Zhou, J., Hu, M., Zhang, Q., Kong, N., Ren, H., Liang, L., Yue, J. 2020. Genome-based classification of *Burkholderia cepacia* complex provides new insight into its taxonomic status. *Biology Direct*, 15(1): 1–14.
- John, J., George, S., Nori, S.R.C., Nelson-Sathi, S., Pisani, D. 2019. Phylogenomic analysis reveals the evolutionary route of resistant genes in *Staphylococcus aureus*. *Genome Biology and Evolution*, 11(10): 2917–2926.
- Lee, A.H.Y., Flibotte, S., Sinha, S., Paiero, A., Ehrlich, R.L., Balashov, S., Ehrlich, G.D., Zlosnik, J.E.A., Mell, J.C., Nislow, C. 2017. Phenotypic diversity and genotypic flexibility of *Burkholderia cenocepacia* during long-term chronic infection of cystic fibrosis lungs. *Genome Research*, 27(4): 650–662.
- Lloyd, J.P.B. 2018. Ubuntu on Windows for computational biology. protocols.io. Available from: <https://www.protocols.io/view/ubuntu-on>

- windows-for-computational-biology-sfuebnw (accessed 28.06.2021).
- Mahenthalingam, E., Baldwin, A., Dowson, C.G. 2008. Burkholderia cepacia complex bacteria: Opportunistic pathogens with important natural biology. *Journal of Applied Microbiology*, 104(6): 1539–1551.
- Méric, G., Miragaia, M., De Been, M., Yahara, K., Pascoe, B., Mageiros, L., Mikhail, J., Harris, L. G., Wilkinson, T.S., Rolo, J., Lambale, S., Bray, J.E., Jolley, K.A., Hanage, W.P., Bowden, R., Maiden, M.C.J., Mack, D., De Lencastre, H., Feil, E.J., Corander J., Sheppard, S.K. 2015. Ecological overlap and horizontal gene transfer in *Staphylococcus aureus* and *Staphylococcus epidermidis*. *Genome Biology and Evolution*, 7(5): 1313–1328.
- Mira, A., Martín-Cuadrado, A.B., D'Auria, G., Rodríguez-Valera, F. 2010. The bacterial pangenome: A new paradigm in microbiology. *International Microbiology*, 13(2): 45–57.
- Möller, S., Krabbenhöft, H.N., Tille, A., Paleino, D., Williams, A., Wolstencroft, K., Goble, C., Holland, R., Belhachemi, D., Plessy, C. 2010. Community-driven computational biology with Debian Linux. *BMC Bioinformatics*, 11(SUPPL. 12): S5.
- Mosquera-Rendón, J., Rada-Bravo, A.M., Cárdenas-Brito, S., Corredor, M., Restrepo-Pineda, E., Benítez-Páez, A. 2016. Pangenome-wide and molecular evolution analyses of the *Pseudomonas aeruginosa* species. *BMC Genomics*, 17(1): 1–15.
- Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J. 2015. Roary: rapid large-scale prokaryote pangenome analysis. *Bioinformatics*, 31(22): 3691–3693.
- Rambaut A. 2013. FigTree. Available from: <http://tree.bio.ed.ac.uk/software/figtree/> (accessed 28.06.2021).
- Rouli, L., Merhej, V., Fournier, P.E., Raoult, D. 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, 7: 72–85.
- Seemann, T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14): 2068–2069.
- Sitto, F., Battistuzzi, F.U. 2020. Estimating pangenomes with Roary. *Molecular Biology and Evolution*, 37(3): 933–939.
- Thorpe, H.A., Bayliss, S.C., Sheppard, S.K., Feil, E.J. 2018. Piggy: A rapid, large-scale pangenome analysis tool for intergenic regions in bacteria. *GigaScience*, 7(4): 1–11.
- Touchon, M., Perrin, A., De Sousa, J.A.M., Vangchhia, B., Burn, S., O'Brien, C.L., Denamur, E., Gordon, D., Rocha, E.P.C. 2020. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genetics*, 16(6): e1008866.
- Whelan, F.J., Rusilowicz, M., McInerney, J.O. 2020. Coinfinder: Detecting significant associations and dissociations in pangenomes. *Microbial Genomics*, 6(3): 1–7.
- Zhou, J., Ren, H., Hu, M., Zhou, J., Li, B., Kong, N., Zhang, Q., Jin, Y., Liang, L., Yue, J. 2020. Characterization of *Burkholderia cepacia* complex core genome and the underlying recombination and positive selection. *Frontiers in Genetics*, 11: 1–15.