Folia Biologica et Oecologica 18: 72–84 (2024)

Acta Universitatis Lodziensis

The characterization of genome sequences diversity of *Pseudomonas aeruginosa* strains from international reference panel using wide range of *in silico* techniques

LEON PETRUŃKO*, KLAUDIA MUSIAŁ, DAWID GMITER

Jan Kochanowski University of Kielce, Faculty of Natural Sciences, Institute of Biology, Department of Microbiology, Uniwersytecka 7, 25-406 Kielce, Poland
E-mail: leonpetrunko@gmail.com

ABSTRACT

Pseudomonas aeruginosa is an important pathogen in patients suffering from Cystic Fibrosis as well as acute opportunistic infections in people without it. For the reason of *P. aeruginosa* having a broad range of habitats it's diversity and adaptability lead it to be a very diverse species. Previous attempts of classifying *P. aeruginosa* strains based on their biochemical and genetic characteristics were made. In presented studied we performed additional characteristic of P. aeruginosa panel strains genomes using wide range of in silico approaches, including Single Nucleotide Polymorphisms (SNPs) - based phylogeny, as well as pan-genome and Intergenic Regions (IGRs) investigation. We shed light on strains diversity, expanding our knowledge about the strains assembled in this international panel. The results of our study may become the basis for further research aimed at fully understanding the pathogenesis of *P. aeruginosa*.

KEYWORDS: *Pseudomonas aeruginosa*, reference panel, pan-genome, intergenic regions (IGRs)

Introduction

Pseudomonas aeruginosa is a Gram-negative bacterium which causes many opportunistic infections, including wound, urinary tract, and respiratory tract. This pathogen is especially important in patients with Cystic Fibrosis (CF) with over half of patients suffering from CF chronically infected by adulthood. Given the incredible diversity of its habitats (soil, still water, plants), which leads to huge biochemical and genetical diversity, it is to no surprise that classifying *P. aeruginosa* strains had been difficult. A large part of the research of *P. aeruginosa* was based on PAO1 strain which is considered a laboratory strain and may have diversified greatly throughout its existence. Considering the diversity of bacterial strains within the species, this may have led to conclusions not relevant to clinical scenario of *P. aeruginosa* infections. To felicitate the research, the

DOI: https://doi.org/10.18778/1730-2366.18.11

international reference panel of *P. aeruginosa* strains was developed. The strains were previously well characterized based on their biochemical, phenotypic, and genotypic characteristics (Cullen *et al*., 2015; De Soyza *et al*., 2013; Freschi *et al*., 2018).

However, additional genomic studies will allow better understanding of the diversity of mentioned strains. Therefore, the purpose on the presented work was the phylogenetic analysis of panel *P. aeruginosa* strains using wide range of *in silico* methods, including SNPs and pan-genome based phylogeny. For wider insight, our study focused as well on diversity of non-coding regions present in strains genomes.

Materials and methods

Genome sequences

All 40 *P. aeruginosa* raw sequences were obtained from the National Center for Biotechnology Information in a FASTA format. Detailed information about the sequences could be found in the work by (Freschi *et al*., 2018).

SNPs-based phylogeny

The REALPHY webserver (Bertels *et al*., 2014) was used to perform SNPs-based phylogenetic analysis. The default options were used and the genome of *P. aeruginosa* strain PAO1 was used as a reference sequence. The loci containing elevated densities of base substitutions, which are marked as recombinations, were identified from REALPHY generated alignment file using Gubbins v3.3.0 (Croucher *et al*., 2015) and visualized using Phandango v1.3.1 (Hadfield *et al*., 2018). The obtained maximum likelihood (ML) phylogenetic tree was midpoint-rooted and visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

Functional annotation with Prokka

The Prokka, (Galaxy Version 1.14.6+galaxy1), (Seemann, 2014) available online on the Galaxy web Server (https://usegalaxy.org/) was used with the default options to annotate the sequences, resulting in a GFF3-formated files.

Pan-genome and IGRs investigation

Initially *P. aeruginosa* pan-genome was analysed using Roary with the following setups: -e -n (to perform alignments using MAFFT), -i 90 (90% sequence identity cut-off) (Page *et al*., 2015). Next the -s option (to not split paralogs into separate clusters) was used in second analysis. The output files from the second analysis were subjected for the IGRs analysis using Piggy with the default parameters (Thorpe *et al*., 2018). Data visualization (genes and IGR presence/absence) was performed using Phandango. The phylogenetic trees based on core genes and core IGRs alignment were created using SeaView (Gouy *et al*., 2010) using ML method (substitution model GTR, branch support: aLTR – SH-like)). To reduce the memory and run time, the alignment files were pre filter using snp_sites v2.5.1 (Page *et al*., 2016). The visualization of trees was done using FigTree. The R package phytools was used for tree comparison (Revell, 2012).

Results and Discussion

P. aeruginosa is a pathogenic bacteria know from its phenotypic and genotypic diversity. To felicitate studies, the panel of reference strains was developed – collected strains represent isolates obtained from different sources, mostly from cystic fibrosis patients, from all over the word (De Soyza *et al*., 2013). So far, the detailed phenotypic characteristic of strains was performed (Cullen *et al*., 2015), together with genomic studies (Freschi *et al*., 2018). However, many aspects of the genomics of panel strains

are yet to be studied. Therefore, within this work, we performed phylogenetic and pan-genome characteristic of mentioned strains. Our work also focused on evaluation of the diversity of non-coding regions within the genome, known as intergenic regions (IGRs).

In first step SNPs-based phylogenetic analysis was performed using online tool REALPHY based on core genomic content. The obtained ML and minpoint-rooted tree are presented on Figure 1. Results revealed clear separation of strains into two major phylogenetic groups: first (green) includes eight strains, meanwhile remaining (n = 32) form wider group (yellow). Within the groups, we observed the presence of closely related strains. These results correspond to the previous one (Freschi *et al*., 2018), were panel strains also grouped in similar way based on core genome single nucleotide variant (SNV), expect for the strain Mi162, that in our analysis fall into wider group.

Next, the Gubbins software was used to investigate putative events of recombination based on alignment file provided by REALPHY. The Gubbins uses spatial scanning statistics to identify loci containing elevated densities of base substitutions suggestive of horizontal sequence transfer while concurrently constructing a maximum likelihood

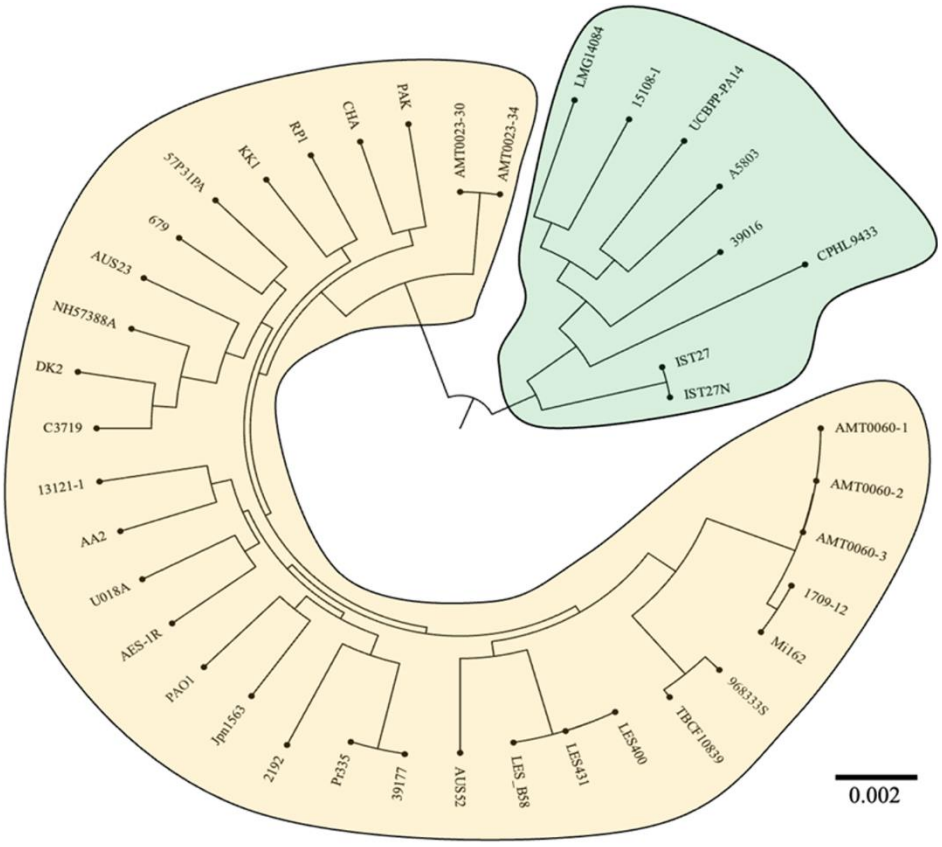

Figure 1. SNPs-based phylogenetic tree of panel *P. aeruginosa* strains genome sequences obtained using REALPHY webserver. Tree was visualized and midpoint-rooted using FigTree.

phylogeny based on the putative point mutations outside these regions of high sequence diversity (Croucher *et al*., 2015) (Fig. 2.).

To better explore the impact of putative recombination events on *P. aeruginosa* panel strains phylogeny, the trees obtained using REALPHY and Gubbins were compared using R package

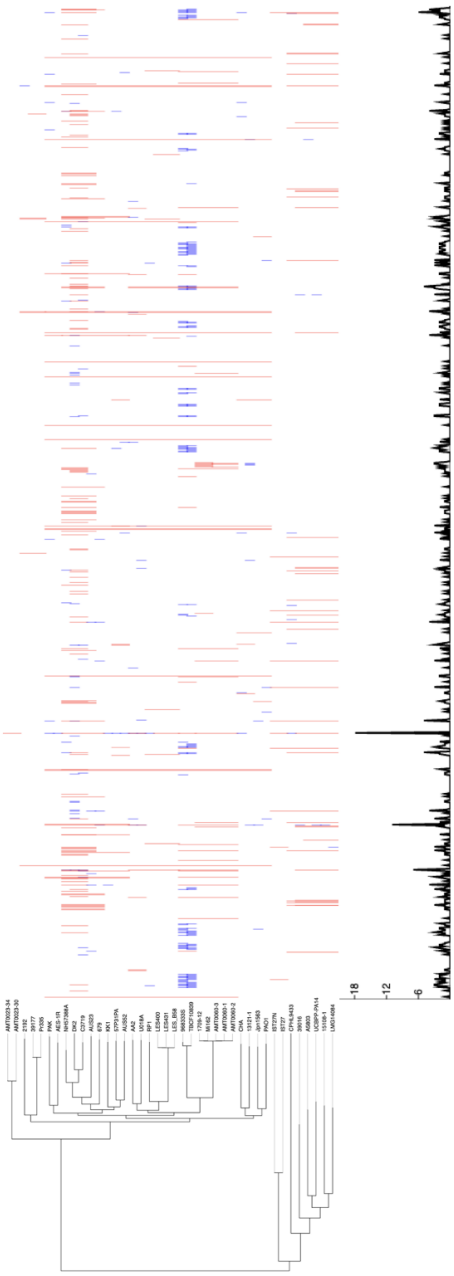

Figure 2. Phylogenetic tree after exclusion of putative recombination regions and matrix showing the recombination pattern within the *P. aeruginosa* genomes. Each column of the matrix refers to a base in the analysed sequence; each row represents an isolate in the phylogeny. Red blocks indicate predicted recombination regions occurring in the internal branch, therefore common to many isolates – common origin. Blue blocks represent recombinations occurring at terminal branches that are unique to individual isolates.

phytools (Fig. 3). Results shown similar clustering of strains into two major groups, consistent in the same strains. Strains from green group cluster identically, however exclusion of regions identified by Gubbins impacted the pattern within orange group. This suggests that horizontal gene transfer occurs at higher level among these strains.

Further, we focused on panel strains pan-genome investigation. It is well recognized that genes within bacterial pan-genomes are categorized as core, additional and unique genes when present in all, most of or only one of the studied genomes, respectively (Gmiter *et al*., 2021). For pan-genome investigation we used Roary, a frequently used software, that allows downstream analysis based on the generated files (Gmiter *et al*., 2021). The Roary software performs pan-genome analysis based on GFF3-formatted input files, containing the annotated genomes sequences (Page *et al*., 2015; Sitto and Battistuzzi, 2020). In step one, software counts identified genes and categorizes them, based on their frequency of occurrence between studied

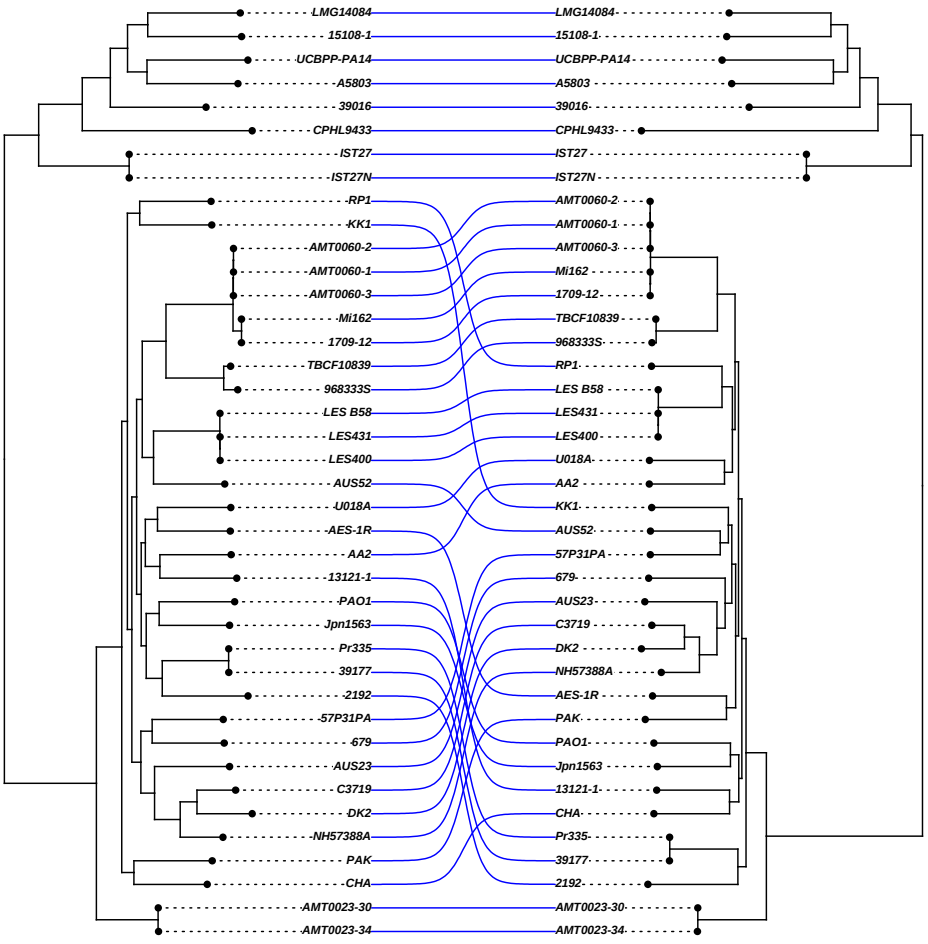

Figure 3. Comparison of phylogenetic trees of *P. aeruginosa* panel strains based on SNPs obtained using REALPHY (left) and Gubbins (right).

strains, into four groups, defined by the software designers; for example, if a gene could be found in 99 to 100% of genomes it is called the core gene (Page *et al*., 2015; Sitto and Battistuzzi, 2020). The Table 1 present the count of genes within categories identified by the Roary, used with the default options and after including the -s parameter (option needed for further analysis). Generally, approximately 5,000 genes can be found in core and soft-core gene categories. Meanwhile, Shell and cloud gene categories, which correspond to the additional and unique genes, accounts greater number of genes. Despite the used parameters of the analysis, the resulted reretraction curves (Fig. 4) indicates that the pan-genome of studied set of *P. aeruginosa* strain is open. It is indicated by the fact that Total genes curve on Fig 4. does not reach plateau (Gmiter *et al*., 2021). Corresponding results were obtained previously (Freschi *et al*., 2019; Mosquera-Rendón *et al*., 2016), indicating similar level of panel strains diversity in comparison to the other isolates.

Roary also creates a gene presence/absence table and then provides phylogenetic tree based on it (so-called pan-genome based phylogeny). The tree together with the matrix is presented on Figure 5. Sequences clustered into two major pan-genome groups, indicated with red and purple.

Table. *Pseudomonas aeruginosa* pan-genome count.

Gene type	Frequency between strains	Default	With -s parameter
Core genes	(99% <= strains <= 100%)	4313	4502
Soft core genes	(95% <= strains < 99%)	884	764
Shell genes	(15% <= strains < 95%)	1680	1382
Cloud genes	(0% <= strains < 15%)	6469	5098
Total genes	(0% <= strains <= 100%)	13346	11746

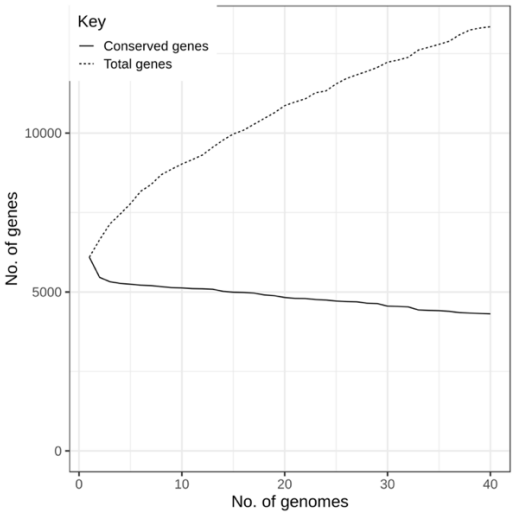

Figure 4. Gene accumulation curves of the *P. aeruginosa* panel strains pan-genome (dashed) and core genes (solid).

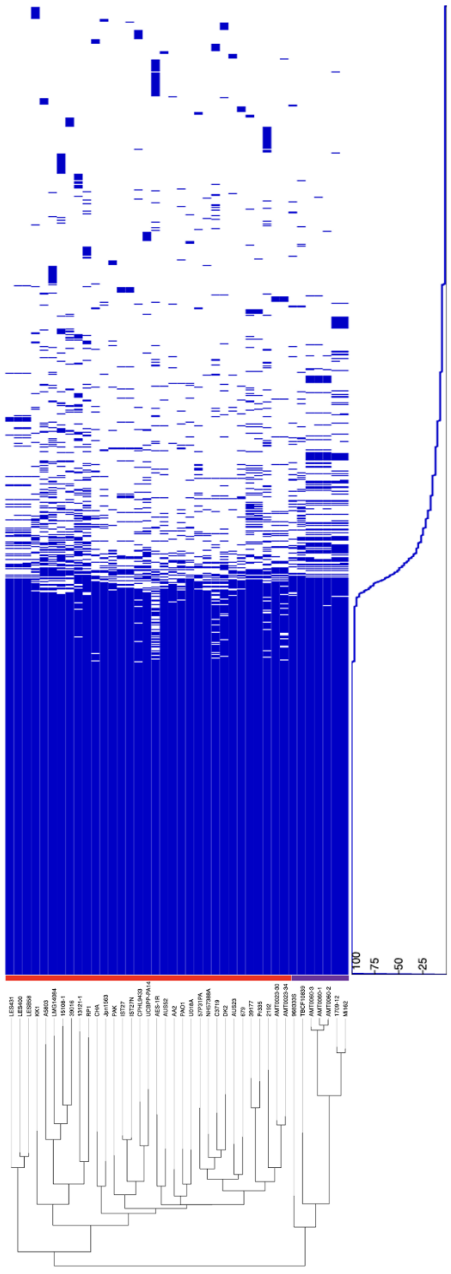

Figure 5. Pan-genome based phylogenetic tree of 40 genomes from *P. aeruginosa* reference panel together with the gene presence/absence matrix. Visualized using Phandango.

To better explore patterns of strains clustering based on SNPs and genomic content, trees from Fig. 1 and Fig. 5 were compared using R package phytools. The results are presented on Figure 6. It could be seen that both phylogenetic approaches provide different pattern of strains clustering. That indicates different evolution of strains genomes content in comparison to the genome sequences. A

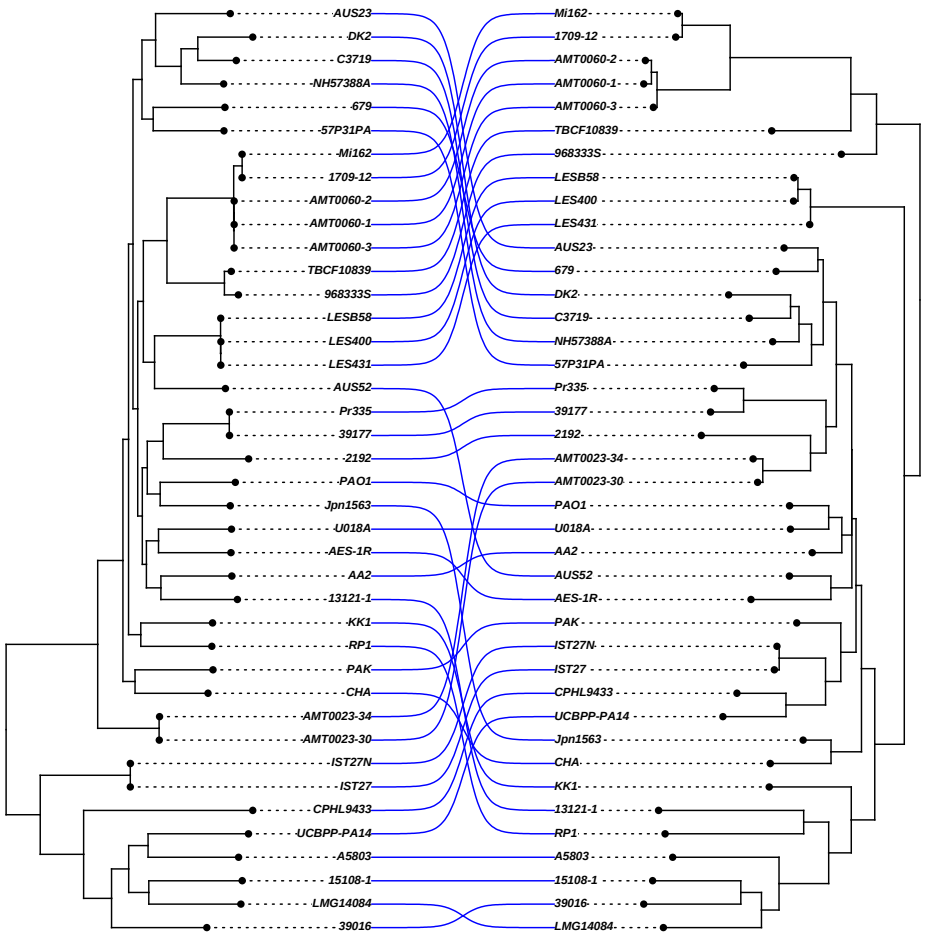

Figure 6. Phylogenetic trees of panel *P. aeruginosa* strains based on SNPs obtained using REALPHY (left) and pan-genome content (right).

similar variability in the way strains were grouped, depending on which one of the above phylogenetic methods was used, was observed in the case of strains belonging to the genus *Aeromonas* (Science *et al*., 2019).

Much of the research analyzing *P. aeruginosa* pan-genome focused on genes often overlooked non-coding intergenic regions (IGRs). IGRs sequence contains a significant amount of crucial genomic information, being biologically relevant elements such as promoter, terminators, and regulatory binding sites (Nielsen *et al*., 2023). For the linkage between genotypes and phenotypes of microorganisms to be effective, IGRs must be taken into consideration. IGRs may also undergo Horizontal Regulatory Transfer (HRT) (Matus-Garcia *et al*., 2012; Ragan and Beiko, 2009). It is thought that as much as 32% of core regulatory regions in *Escherichia coli* and 51% of overall core IGRs were acquired via HRT (Oren *et al*., 2014). Thus, in some cases IGRs may lead to more genetical diversity than genes themselves.

The analysis of IGRs in *P. aeruginosa* panel strains genomes was performed using Piggy, based on Roary files generated with -s option, to not split paralogs into separate clusters (Thorpe *et al.*, 2018). IGRs can be denoted as core and additional, depending on frequency. Figure 7 presents the matrix of core and additional IGRs among strains.

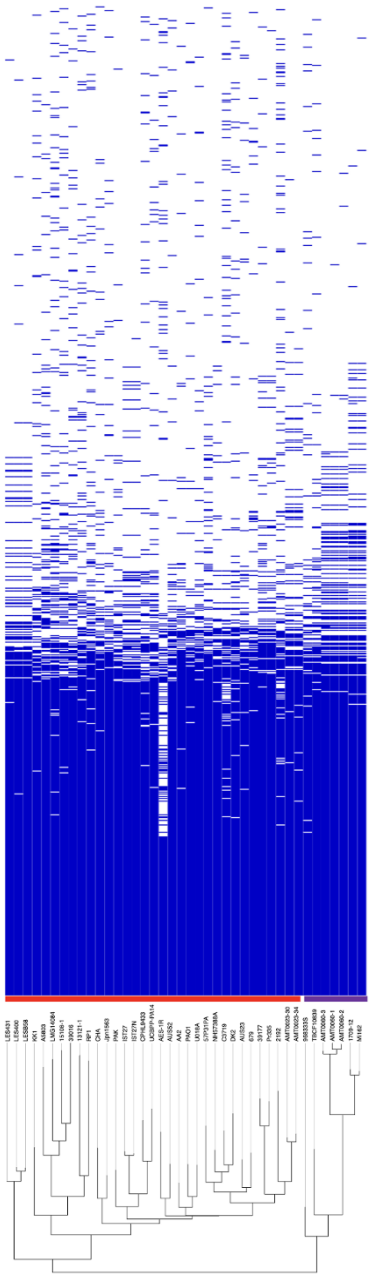

Figure 7. Distribution of IGRs among studied 40 genome sequences of panel *P. aeruginosa* strains. The pan-genome based phylogenetic tree was used for visualization using Phandango.

Piggy provides a core IGRs alignment file – similar to the way Roary provides the core genes alignment file. Both files can be used to perform phylogenetic analysis. Here we used SeaView to obtain ML phylogenetic trees. Figure 8 presents comparison of the mentioned trees based on core genes and core IGRs. Both approaches generated similar patterns of strains clustering into two major groups. However, the groups differ in terms of strains diversity within. Smaller group of eight strains (green group, Fig 1.) cluster almost identical, except of strain 15108-1, which shown more similarities to LMG14084 in IGRs based phylogeny, and strain 39016, for which phylogenetic distance to LMG14084 is higher in IGRs based phylogeny. Meanwhile, strains from second group shown higher level of core genes and core IGRs dissimilarities. These results suggest different evolutionary pressure on genomic regions in a particular panel strain.

Finally, Figure 9 presents the characteristic of genes and IGRs frequency and distribution in *P. aeruginosa* panel strains pan-genome. We

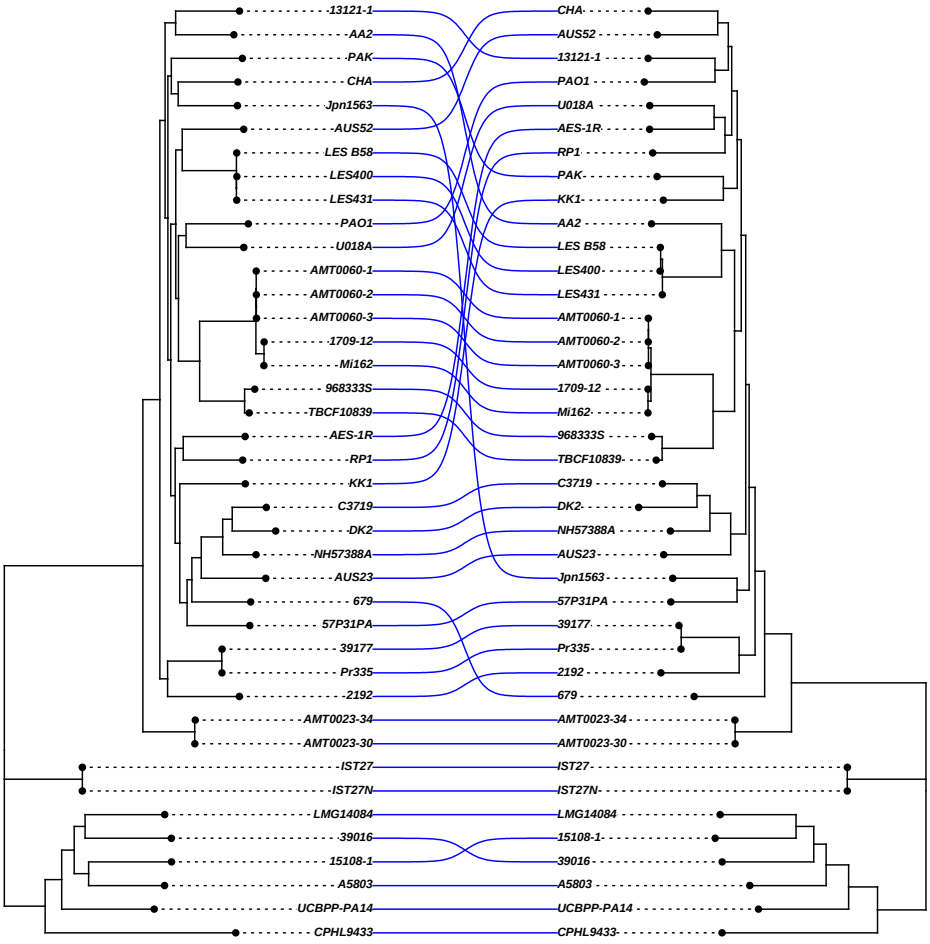

Figure 8. The phylogenetic trees of *P. aeruginosa* panel strains based on core gene sequences (left) and core IGRs sequences (right).

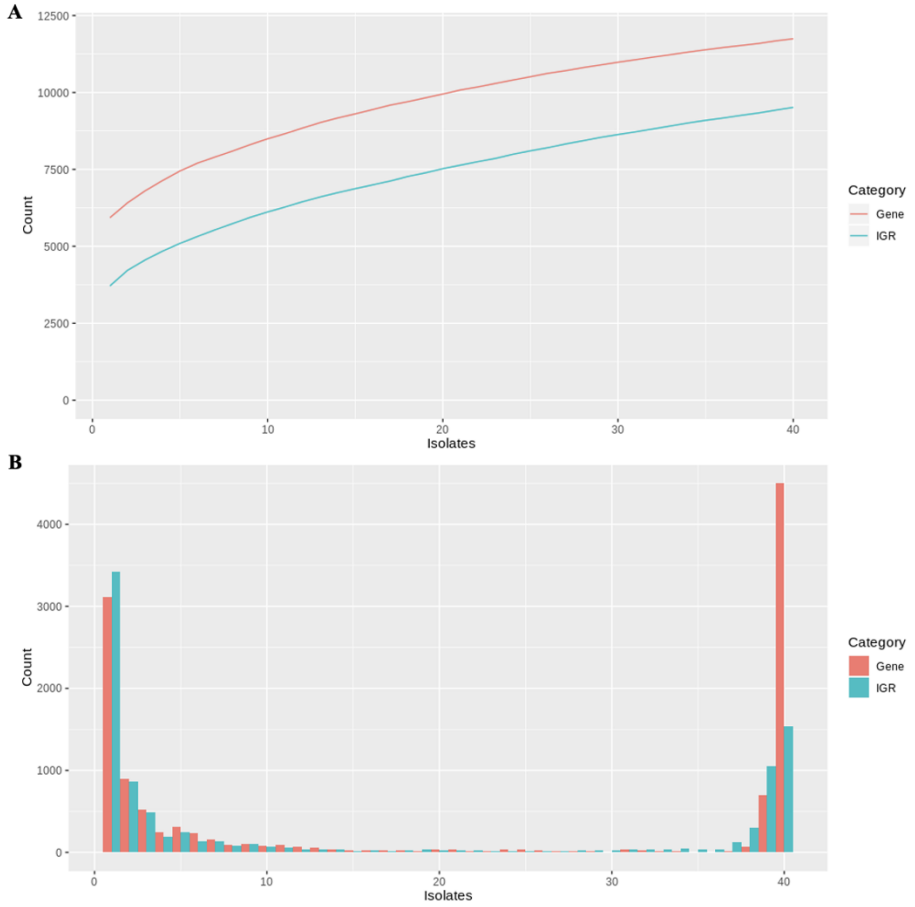

Figure 9. Properties of the panel *P. aeruginosa* strains pan-genome and its IGRs. (A) Number of unique IGRs (green) and genes (orange) as a function of the number of isolates included in the pangenome. (B) Distribution of unique IGRs (green) and genes (orange) across the pangenome, illustrated with a frequency histogram (number of IGRs/genes present in the given number of genomes).

observed that the number of unique IGRs increases with the number of analysed genomes, with a similar pattern in case of the number of unique genes (Fig. 9A). Moreover, the majority of IGRs was present in either most of the strains or only a few of them, which means that they were either very common or very rare, with only few exceptions (Fig. 9B). The obtained results remain in agreement with previous observations for other bacterial species (Nielsen *et al*., 2023; Thorpe *et al*., 2018).

Conclusions

P. aeruginosa represents species that is known for its genetic and phenotypic diversity. To better explore mechanisms of its pathogenicity, the reference panel of word-wide isolates was developed and primarily characterized. Within presented study we performed additional characteristic of panel strains genomes using wide range of *in silico* approaches, including SNPs- based phylogeny, as well as pan-genome and IGRs investigation. We shed light on strains diversity,

concluding that strains form two major phylogenetic groups, when used methods based on comparison of DNA sequences. Further, these groups are consistent with the same strains. However, strains from these groups might be characterized by different level of diversity. We observed different pattern of potential recombination events, as well as dynamics of coding and non-coding regions evolution. On the other hand, studied strains shown distinct patter of clustering in pan-genome based phylogeny comparing to previous. Obtained results suggest evolutionary pressure from the environment of their isolation source. This diversity, in turn, most likely affect strains pathogenicity.

Acknowledgements

Study was supported by the Polish National Science Centre Grant 2019/32/T/NZ1/00515 (D.G.).

Reference

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P. B., Van Nimwegen, E. 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Molecular Biology and Evolution, 31(5).

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., Harris, S.R. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research, 43(3), e15.

Cullen, L., Weiser, R., Olszak, T., Maldonado, R.F., Slachmuylders, L., Brackman, G., Paunova-Krasteva, T.S., Zarnowiec, P., Czerwonka, G., Reilly, J., Drevinek, P., Kaca, W., Melter, O., De Soyza, A., Perry, A., Winstanley, C., Stoitsova, S.R., Lavigne, R., Mahenthiralingam, E., Sá-Correia, I., Coenye, T., Drulis-Kawa, Z., Augustyniak, D., Valvano, M.A., McClean, S. 2015. Phenotypic characterization of an international Pseudomonas aeruginosa reference panel: strains of cystic fibrosis (CF) origin show less in vivo virulence than non-CF strains. Microbiology, 161(10), 1961–1977.

De Soyza, A., Hall, A.J., Mahenthiralingam, E., Drevinek, P., Kaca, W., Drulis-Kawa, Z., Stoitsova, S.R., Toth, V., Coenye, T., Zlosnik, J.E.A., Burns, J. L., Sá-Correia, I., De Vos, D., Pirnay, J.P., Kidd, T.J., Reid, D., Manos, J., Klockgether, J., Wiehlmann, L., Tümmler, B., McClean, S., Winstanley, C. 2013. Developing an international Pseudomonas aeruginosa reference panel. MicrobiologyOpen, 2(6).

Freschi, L., Bertelli, C., Jeukens, J., Moore, M.P., Kukavica-Ibrulj, I., Emond-Rheault, J.G., Hamel, J., Fothergill, J.L., Tucker, N.P., McClean, S., Klockgether, J., De Soyza, A., Brinkman, F.S.L., Levesque, R.C., Winstanley, C. 2018. Genomic characterisation of an international Pseudomonas aeruginosa reference panel indicates that the two major groups draw upon distinct mobile gene pools. FEMS Microbiology Letters, 365(14).

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.

Gmiter, D., Nawrot, S., Pacak, I., Zegadło, K., Kaca, W. 2021. Towards a better understanding of the bacterial pan-genome. Acta Universitatis Lodziensis. Folia Biologica et Oecologica, 17, 84–96.

Gouy, M., Guindon, S., Gascuel, O. 2010. Sea view version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution, 27(2), 221–224.

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.

Matus-Garcia, M., Nijveen, H., Van Passel, M.W.J. 2012. Promoter propagation in prokaryotes. Nucleic Acids Res. 40, 10032–10040.

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.

Nielsen, F.D., Møller-Jensen, J., Jørgensen, M.G. 2023. Adding context to the pneumococcal core genes using bioinformatic analysis of the intergenic pangenome of Streptococcus pneumoniae. Frontiers in Bioinformatics, 3.

Oren, Y., Smith, M.B., Johns, N.I., Kaplan Zeevi, M., Biran, D., Ron, E.Z., Corander, J., Wang, H.W., Alm, E.J., Pupko, T. 2014. Transfer of noncoding DNA drives regulatory rewiring in Bacteria. Proc. Natl. Acad. Sci. U. S. A. 111, 16112–16117.

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 pan-genome analysis. Bioinformatics, 31(22), 3691–3693.

Page, A.J., Taylor, B., Delaney, A.J., Soares, J., Seemann, T., Keane, J.A., Harris, S.R. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics, 2(4), e000056.

Ragan, M.A., Beiko, R.G. 2009. Lateral genetic transfer: Open issues. Philos. Trans. R. Soc. B Biol. Sci. 364, 2241–2251.

Revell, L.J. 2012. Phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3(2), 217–223.

Science, E., Zhong, C., Han, M., Yang, P., Chen, C., Yu, H., Wang, L., Ning, K. 2019. Comprehensive analysis reveals the evolution and pathogenicity of *Aeromonas*, Viewed from Both Single Isolated Species and Microbial Communities. mSystems, 4(5), 1–19.

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 pan-genome analysis tool for intergenic regions in bacteria. GigaScience, 7(4), 1–11.