MedCrave
Step into the World of Research

MOJ Proteomics & Bioinformatics

Short Communication

Open Access

CrossMark

Core *Pseudomonas* genome from 10 *Pseudomonas* species

Abstract

Core genome of a set of organisms represents the set of homologous genes shared between the set of organisms with many applications. The *Pseudomonas* genus is highly diverse with both plant and animal pathogens. Hence, the core genome of *Pseudomonas* genus can be useful. Current studies presented contradictory results with the core genome of *Pseudomonas* genus marginally larger than that of *Pseudomonas aeruginosa*. In this study, we attempt to identify a core *Pseudomonas* genome from 10 publicly available annotated genomes by intersecting homologous coding sequences using BLAST. Our results suggest a 218-gene core genome, which is 3.46% of the coding sequences of *P. aeruginosa*. 136 of 218 genes were mapped to official gene symbols and were enriched in 8 clusters in Gene Ontology biological processes related to central metabolism.

Volume 9 Issue 3 - 2020

Xue Ting Tan,[1,2] Avettra Ramesh,[1,2] Victor CC Wang,[1,2] Nur Jannah Kamarudin,[1,2] Shermaine SM Chew,[1,2] Madhurya V Murthy,[1,2] Nikita V Yablochkin,[1,2] Karthiga Mathivanan,[1,2] Maurice HT Ling[1,2,3]

[1]Department of Applied Sciences, Northumbria University, United Kingdom
[2]School of Life Sciences, Management Development Institute of Singapore, Singapore
[3]HOHY PTE LTD, Singapore

Correspondence: Maurice HT Ling, School of Life Sciences, Management Development Institute of Singapore, 501 Stirling Road, Singapore 148951, Republic of Singapore, Singapore, Email mauricling@acm.org

Received: June 27, 2020 | **Published:** July 17, 2020

Introduction

The core genome for a set of related genomes represents a set of orthologous genes within a set of related genomes,[1] which may be from different strains of a species[2] or different species of a genus.[3] Hence, core genome represents the intersection of the set of genomes under study. Therefore, phylogenetically related genomes tend to share more genes and likely to have a larger core genome.[4] This is different from pan-genome, which is the entire set of all genes from the genomes under study.[5] There are many applications of core genomes. For example, the core genome is crucial to observe genomic distance within a species, which can then be used for disease surveillance and outbreak monitoring.[6,7] It can also be used to study speciation events[8] and the evolutionary history of an organism.[9]

The *Pseudomonas* genus is one of the most diverse bacterial genera[10] inhabiting a wide variety of environments,[11] including pathogens of both plants and animals.[12] For example, Batrich et al.,[13] found a variety of *Pseudomonas* species demonstrating antibiotics resistance and metal tolerance near Lake Michigan. Hence, it is useful to elucidate the core genome of *Pseudomonas* genus for further applications. A study by Hesse et al.,[14] examined 166 *Pseudomonas* type strains to deduce a core genome of 794 genes while Freschi et al.,[15] focused on identifying *Pseudomonas aeruginosa*core genome and used 1,311 *P. aeruginosa* genomes sequences to obtain a 665-gene *P. aeruginosa* core genome.However, there is a contradiction–shouldthe core genome of *P. aeruginosa* is 665 genes,[15] it is not likely for the core genome of *Pseudomonas* genus to be only 794 genes.[14] This may be due to low stringency criteria in identifying orthologs used by Hesse et al.,[14] which is 30% identity at 50% coverage; as compared to Freschi et al.,[15] which is 50% identity at 85% coverage. This suggests that the core genome of *Pseudomonas* genus warrants further study.

Here, we attempt to identify a core *Pseudomonas* genome from 10 publicly available annotated genomes. Our results suggest a 218-gene core genome, which is 3.46% of the coding sequences of *P. aeruginosa*.

Materials and methods

Genome data set: The genome of 10 Pseudomonas species; namely, (i) *Pseudomonas aeruginosa* (Accession CP045002.1; P1), (ii) *Pseudomonas mandelii* (Accession NZ_CP005960.1; P2), (iii) *Pseudomonas balearica* (Accession CP045858.1; P3), (iv) *Pseudomonas chlororaphis* (Accession NZ_CP027716.1; P4), (v) *Pseudomonas fluorescens* (Accession NZ_CP048607.1; P5), (vi) *Pseudomonas fulva* (Accession NZ_CP023048.1; P6), (vii) *Pseudomonas orientalis* (Accession NZ_CP018049.1; P7), (viii) *Pseudomonas psychrophila* (Accession NZ_CP049044.1; P8), (ix) *Pseudomonas putida* (Accession NZ_CP026115.2; P9), and (x) *Pseudomonas synxantha* (Accession NZ_CP027754.1; P10); were obtained from NCBI.

Determining core genome by intersecting genomes: The core genome of *Pseudomonas* was determined as the intersection of the 10 *Pseudomonas* genomes. Operationally, the intersection of 2 genomes; such as, *P. aeruginosa* (P1) and *P. mandelii* (P2); was determined by constructing a BLAST database out of the coding sequences of *P. aeruginosa* and the coding sequences of *P. mandelii* were used as query in BLASTN[16] version 2.10.0. The expectation value (E-value) in BLAST is defined as per-search expected false positive rate[17] and was set to less than 1E-9,[18] which had been used in pan-genomics[19] and homology.[20] Only the top match was taken for each of the query sequences. The result represented the core genome of *P. aeruginosa* and *P. mandelii* (denoted as P1P2). Subsequently, the coding sequences of *P. balearica* (P3) was used to construct a BLAST database for sequence comparison with P1P2 under the same E-value threshold.

The result represented the core genome of *P. aeruginosa*, *P. mandelii* and *P. balearica* (denoted as P1P2P3). This process was repeated until all 10 *Pseudomonas* genomes were intersected, which represented the core genome and was denoted as P1P2P3P4P5P6P7P8P9P10.

Determining functions of core genome: The functional properties of the core genome were determined by gene set enrichment analysis[21–23] for biological processes using PANTHER[24,25] on the official gene symbols.

Results and discussion

The number of coding sequence (CDS) ranges from to 4274 in *P. balearica* to 6305 in *P. aeruginosa* (Table 1). Using genome intersection, a 218-gene core genome was identified, which amounts to 3.46% of *P. aeruginosa* genome (Table 2). A study on 23 *Corallococcus* genomes[26] suggest that the size of pan-genome[5] can be estimated to be $8127N^{0.5481}$ genes, where N is the number of genomes. Using this estimation,[26] the size of pan-genome of the 10 *Pseudomonas* species is estimated to be 28,750 CDS or genes. Inglin et al.,[27] examined 98 complete genomes of the genus *Lactobacillus* and found the core and pan-genome to be 266 genes and 20,800 genes, respectively. This amounts to 1.28% of the pan-genome being the core genome. We evaluate the use of this core genome to pan-genome ratio in this case. Using this ratio, where the size of core genome is 1.28% of pan-genome, on our estimated 28,750-gene *Pseudomonas* pan-genome, we will expect a core genome of 368 genes, which 68% more than that identified in this study. The difference may be due to the higher stringency on the E-value threshold used in this study (E-value<1E-9), which is commonly used as threshold for pan-genomics[19] and homology[20] studies, as compared to Inglin et al.,[27] whom uses E-value of less than 1E-5. This suggests that the estimation of the size of pan-genome[26] from number of genomes and the estimation of the size of core genome from the size of pan-genome by ratio[27] may be a useful heuristic (Table 1&2).

Table 1 Number of Coding Sequences (CDS) in each organism

Label	Organism	Accession number	Number of CDS
P1	*P. aeruginosa*	CP045002.1	6305
P2	*P. mandelii*	NZ_CP005960.1	6139
P3	*P. balearica*	CP045858.1	4274
P4	*P. chlororaphis*	NZ_CP027716.1	5886
P5	*P. fluorescens*	NZ_CP048607.1	5914
P6	*P. fulva*	NZ_CP023048.1	4541
P7	*P. orientalis*	NZ_CP018049.1	5248
P8	*P. psychrophila*	NZ_CP049044.1	4737
P9	*P. putida*	NZ_CP026115.2	5561
P10	*P. synxantha*	NZ_CP027754.1	6135

Table 2 Progressive reduction of number of CDS

CDS Set	Number of CDS	Percentage
P1	6305	100.00%
P2	6139	97.37%
P1P2	1320	20.94%
P1P2P3	1294	20.52%
P1P2P3P4	796	12.62%
P1P2P3P4P5	575	9.12%
P1P2P3P4P5P6	402	6.38%
P1P2P3P4P5P6P7	344	5.46%
P1P2P3P4P5P6P7P8	237	3.76%
P1P2P3P4P5P6P7P8P9	230	3.65%
P1P2P3P4P5P6P7P8P9P10	218	3.46%

Citation: Tan XT, Ramesh A, Wang VCC, et al. Core *Pseudomonas* genome from 10 *Pseudomonas* species. *MOJ Proteomics Bioinform.* 2020;9(3):68–71. DOI: 10.15406/mojpb.2020.09.00282

Of the 218-genes core genome identified, 136 (62.4%) genes were mapped to official gene symbols for gene set enrichment analysis.[21–23] Our results show an enrichment in eight biological process ontological terms; namely, (i) Guanosine-containing compound metabolic process (GO:1901068), (ii) glutamine family amino acid metabolic process (GO:0009064), (iii) purine nucleotide metabolic process (GO:0006163), (iv) purine-containing compound biosynthetic process (GO:0072522), (v) tRNA aminoacylation for protein translation (GO:0006418), (vi) small molecule biosynthetic process (GO:0044283), (vii) response to nutrient levels (GO:0031667), and (viii) aerobic respiration (GO:0009060).

The first five enriched terms (GO:1901068, GO:0009064, GO:0006163, GO:0072522, and GO:0006418) represent central metabolic processes for growth, which is similar to the core genome of *Comamonas*.[28] Small molecule biosynthetic process (GO:0044283) are often related to response to nutrient levels (GO:0031667), which are also found in the core genome of *Acidithiobacillus*.[29] Aerobic respiration is expected as *Pseudomonas* are generally aerobic.[30,31] Hence, the biological processes of *Pseudomonas* core genome identified in this study are supported by current studies in other bacterial genus.

In conclusion, this study identified a 218-gene core genome of *Pseudomonas*, which is linked to central metabolic processes and nutrient metabolism.

Data availability

The data files for this study can be downloaded at https://bit.ly/CorePseudomonasGenome, which is a zip file containing four folders; namely, (i) FASTA Files contain the 10 *Pseudomonas* genomes, (ii) BLAST Files contain the results from BLASTN, (iii) Intersection Files contain the progressive genomic intersections after BLAST where P1P2P3P4P5P6P7P8P9P10.fasta is the core genome of the 10 *Pseudomonas* species, and (iv) Core Genome contains the description and GSEA results of the core genome.

Acknowledgments

None.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Funding

None.

References

1. Barajas HR, Romero MF, Martínez-Sánchez S, et al. Global Genomic Similarity and Core Genome Sequence Diversity of the Streptococcus Genus as a Toolkit to Identify Closely Related Bacterial Species in Complex Environments. *Peer J*. 2019;6:e6233.
2. Goodall ECA, Robinson A, Johnston IG, et al. The Essential Genome of Escherichia coli K-12. *mBio*. 2018 20;9(1):e02096.
3. Alcaraz LD, Moreno-Hagelsieb G, Eguiarte LE, et al. Understanding the Evolutionary Relationships and Major Traits of Bacillus through Comparative Genomics. *BMC Genomics*. 2010;11:332.
4. Guimarães LC, Florczak-Wyspianska J, de Jesus LB, et al. Inside the Pan-Genome - Methods and Software Overview. *Curr Genomics*. 2015;16(4):245–252.
5. Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of Pan-Genome analyses. *Curr Opin Microbiol*. 2015;23:148–154.
6. Aggelen H van, Kolde R, Chamarthi H, et al. A Core Genome Approach that Enables Prospective and Dynamic Monitoring of Infectious Outbreaks. *Sci Rep*. 2019;9(1):7808.
7. Guglielmini J, Bourhy P, Schiettekatte O, et al. Genus-Wide Leptospira Core Genome Multilocus Sequence Typing for Strain Taxonomy and Global Surveillance. *PLoS Negl Trop Dis*. 2019;13(4):e0007374.
8. Segerman B. The Genetic Integrity of Bacterial Species: The Core Genome and The Accessory Genome, Two Different Stories. *Front Cell Infect Microbiol*. 2012;2.
9. Sarkar SF, Guttman DS. Evolution of the Core Genome of Pseudomonas syringae, A Highly Clonal, Endemic Plant Pathogen. *Appl Environ Microbiol*. 2004;70(4):1999–2012.
10. Jun S-R, Wassenaar TM, Nookaew I, et al. Diversity of Pseudomonas Genomes, Including Populus-Associated Isolates, as Revealed by Comparative Genome Analysis. Kivisaar M, editor. *Appl Environ Microbiol*. 2016;82(1):375–383.
11. Silby MW, Winstanley C, Godfrey SAC, et al. Pseudomonas Genomes: Diverse and Adaptable. *FEMS Microbiol Rev*. 2011;35(4):652–680.
12. Otero-Asman JR, Wettstadt S, Bernal P, et al. Diversity of Extracytoplasmic Function Sigma (σECF) Factor-Dependent Signaling in Pseudomonas. Mol Microbiol. 2019;112(2):356–373.
13. Batrich M, Maskeri L, Schubert R, et al. Pseudomonas Diversity Within Urban Freshwaters. *Front Microbiol*. 2019;10:195.
14. Hesse C, Schulz F, Bull CT, et al. Genome-Based Evolutionary History of Pseudomonas spp. *Environ Microbiol*. 2018;20(6):2142–2159.
15. Freschi L, Vincent AT, Jeukens J, et al. The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biol Evol*. 2019;11(1):109–120.
16. Altschul SF, Gish W, Miller W, et al. Basic Local Alignment Search Tool. *J Mol Biol*. 1990;215(3):403–410.
17. Pearson WR. Finding Protein and Nucleotide Similarities with FASTA. *Curr Protoc Bioinforma*. 2016;53.
18. Herman RA, Song P. Validation of Bioinformatic Approaches for Predicting Allergen Cross Reactivity. *Food Chem Toxicol*. 2019;132:110656.
19. Håfström T, Jansson DS, Segerman B. Complete Genome Sequence of Brachyspira intermedia Reveals Unique Genomic Features in Brachyspira Species and Phage-Mediated Horizontal Gene Transfer. *BMC Genomics*. 2011;12:395.
20. Cruz-Morales P, Orellana CA, Moutafis G, et al. Revisiting the Evolution and Taxonomy of Clostridia, a Phylogenomic Update. *Genome Biol Evol*. 2019;11(7):2035–2044.
21. Felten A, Vila Nova M, Durimel K, et al. First Gene-Ontology Enrichment Analysis Based on Bacterial Coregenome Variants: Insights into Adaptations of Salmonella Serovars to Mammalian- and Avian-Hosts. *BMC Microbiol*. 2017;17(1):222.
22. Hung J-H, Yang T-H, Hu Z, et al. Gene Set Enrichment Analysis: Performance Evaluation and Usage Guidelines. *Brief Bioinform*. 2012;13(3):281–291.
23. Subramanian A, Tamayo P, Mootha VK, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
24. Mi H, Muruganujan A, Casagrande JT, et al. Large-Scale Gene Function Analysis with the PANTHER Classification System. *Nat Protoc*. 2013;8(8):1551–1566.
25. Mi H, Muruganujan A, Ebert D, et al. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res*. 2019;47(D1):D419–D426.

26. Livingstone PG, Morphew RM, Whitworth DE. Genome Sequencing and Pan-Genome Analysis of 23 Corallococcus spp. Strains Reveal Unexpected Diversity, With Particular Plasticity of Predatory Gene Sets. *Front Microbiol*. 2018;9:3187.

27. Inglin RC, Meile L, Stevens MJA. Clustering of Pan- and Core-genome of Lactobacillus provides Novel Evolutionary Insights for Differentiation. *BMC Genomics*. 2018;19(1):284.

28. Wu Y, Zaiden N, Cao B. The Core- and Pan-Genomic Analyses of the Genus Comamonas: From Environmental Adaptation to Potential Virulence. *Front Microbiol*. 2018;9:3096.

29. Zhang X, Liu Z, Wei G, et al. In Silico Genome-Wide Analysis Reveals the Potential Links Between Core Genome of Acidithiobacillus thiooxidans and Its Autotrophic Lifestyle. *Front Microbiol*. 2018;9:1255.

30. Leppik RA, Park RJ, Smith MG. Aerobic Catabolism of Bile Acids. *Appl Environ Microbiol*. 1982;44(4):771–776.

31. Arai H. Regulation and Function of Versatile Aerobic and Anaerobic Respiratory Metabolism in Pseudomonas aeruginosa. *Front Microbiol*. 2011;2:103.

Citation: Tan XT, Ramesh A, Wang VCC, et al. Core *Pseudomonas* genome from 10 *Pseudomonas* species. *MOJ Proteomics Bioinform.* 2020;9(3):68–71. DOI: 10.15406/mojpb.2020.09.00282