OPEN ACCESS Freely available online

ISSN: 2153-0602
Journal of
Data Mining in
Genomics & Proteomics

Opinion Article

Pan-Genome Analysis over Metagenome Assembled Genomes

James Harry*

Department of Biochemistry and Biotechnology, College of Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

DESCRIPTION

In 2005, the term pan-genome was used to describe the full gene set of all strains within a species. A pan-genes genome's are divided into three groups: core, auxiliary, and unique. Core genes must be shared by all strains within the species, while accessory genes are found in a subset of strains and unique genes are found only in one strain. In practice, a softer threshold can be used to determine key genes. Pan-genome analysis has been used in comparative genomics research since 2005, not only in prokaryotes but also in plants, fungi, mammals, and humans. Pan-genome analysis is useful in researching genomic diversity and phylogeny in bacteria, as well as disease outbreaks, virulence-associated genes, and antibiotic resistance. PGAP, GET HOMOLOGUES, ITEP, Roary, Anvi'o, BPGA, and PanX are only a few of the computational tools developed for bacterial pan-genome research. In recent years, surveys and comparisons of these technologies have been published. According to a recent study, insufficient and inconsistent gene annotations may result in underestimation of core genome size and overestimation of pan-genome size. This is especially significant since that draught isolate genomes and Metagenome-Assembled Genomes (MAGs) are being used in pan-genome analyses. Filtering, assembling, binning, and taxonomy assignment are used to construct approximate representations of individual genomes from metagenome shotgun sequencing reads. When the first large-scale MAG study was published in 2015, the word "MAG" first appeared in the literature. The Minimum Information about a Metagenome-Assembled Genome (MIMAG), a metagenomics community standard for distributing MAGs with obligatory metrics, was published by the Genomic Standards Consortium (GSC) in 2017. Hundreds of thousands of MAGs have been reassembled from a variety of settings, including the ocean, soil, freshwater, human gut, activated sludge, and animal gut. These MAGs are very important for improving metabolic capacity forecasts, discovering completely new taxa, and expanding the tree of life. However, a recent study found that 95% complete MAGs only captured 77% of population core genes and 50% of variable genes, and that MAG quality was generally lower than expected. Another study examined problems about gaps, assembly, and binning mistakes, all of which would greatly limit the use of MAGs. Even high-quality MAGs (according to MIMAG, >90% completeness and 5% contamination) may have assembly mistakes and chimaeras. Numerous research using MAGs in pan-genome investigation of human micro biomes have been published in the last four years. Combining MAGs with isolate genomes or using simply MAGs for pan-genome analysis has obviously become commonplace. However, it has never been determined to what extent the limitations (fragmentation, incompleteness, and contamination) of MAGs influence the accuracy of pan-genome conclusions. We expected that incorporating MAGs in pan-genome study results will result in biases and mistakes. We put this theory to the test by comparing entire genome pan-genome analysis results with MAGs simulated from complete genomes by adding fragmentation, incompleteness, and contamination. We have made recommendations based on our findings on how to reduce the accuracy loss caused by the use of MAGs.

In the last five years, large-scale metagenome assembly and binning to construct Metagenome Assembled Genomes (MAGs) has become conceivable. Millions of MAGs have been created as a result, and they are increasingly being used in pan-genomics workflows. However, due to mis-assembly and mis-binning, pan genome analysis of MAGs may suffer from the known difficulties with MAGs: fragmentation, incompleteness, and contamination. By comparing pan-genome analysis results of full bacterial genomes and simulated MAGs, we conducted a critical review of adding MAGs in pan-genome analysis.

We discovered that incompleteness resulted in a greater loss of core genes than fragmentation. Contamination had a minor impact on core genome size, but had a significant impact on accessory genome size. When using several pan-genome analytic technologies and a combination of MAGs and entire genomes, the core gene loss remained. Importantly, decreasing the core gene threshold and utilizing gene prediction algorithms that account fragmented genes partially eased core gene loss, albeit to a lesser extent when incompleteness was greater than 5%. The

Correspondence to: James Harry, Department of Biochemistry and Biotechnology, College of Science, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, E-mail: harryjam@hotmail.com

Received: 02-May-2022, Manuscript No. JDMGP-22-17227; **Editor assigned:** 04-May-2022, PreQC No. JDMGP-22-17227 (PQ); **Reviewed:** 18-May-2022, QC No JDMGP-22-17227; **Revised:** 25-May-2022, Manuscript No. JDMGP-22-17227 (R); **Published:** 02-Jun-2022. DOI: 10.4172/ 2153-0602.22.13.254.

Citation: Harry J (2022) Pan-Genome Analysis over Metagenome Assembled Genomes. J Data Mining Genomics Proteomics. 13: 254.

loss of core genes also resulted in faulty pan-genome functional predictions and phylogenetic trees.

To mitigate the accuracy loss, we conclude that decreasing the core gene threshold and predicting genes in metagenome mode are required in pan-genome analysis of MAGs. Future research will require improved MAG quality control and the development of new pan-genome analysis methods specifically developed for MAGs.