PREDICTING AND ANALYSING ANTIMICROBIAL RESISTANCE IN PAN-GENOME

Abstract-

Bacterial genomes are dynamic and are shaped by a complex synthesis of several mechanisms of evolution as a result of horizontal transfer, gene loss, and duplication. The overall number of unique genes detected in a collection of related isolates in the pan-genome is frequently many times bigger than the genomes of individual isolates. We have developed a method to quickly find orthologous gene clusters across the entire genome. This workflow is connected to an effective but simple web-based visualization for interactive pan-genome exploration. The visualization is made up of interconnected components that enable rapid filtering, searching, and analysis of the evolutionary history of genes. PanX presents an alignment, a phylogenetic tree, maps mutations inside a gene cluster to the branches of the tree, and infers gene gain and loss on the core-genome phylogeny for each gene cluster. To determine the phylogenetic tree diversity and potential evolutionary trends of aquatic bacterial pathogen strains, as well as to clarify the pathogenesis mechanisms and estimate patterns of pathogen transmission across epidemiological scales, pan-genome genomic-derived aquatic pathogens are required. Using a web server or providing panX locally as a browser-based application, one can view customized pan-genomes.

Keywords: Pan-genome, Diamond, MCL, Pathogens, Phylogenetic tree, PanX, Genome phylogeny, Porphyromonas gingivalis

1. Introduction

Here, we propose panX, a web-based environment for visualizing and analyzing microbial pan-genome data that is predicated on a process for automatic pan-genome identification. The workflow first divides the genomes of many annotated genomes into genes, and then it clusters the genes into orthologous groups. It uses these clusters to identify the core genome, identify SNPs in the core genome to build a strain-level phylogeny, create multiple alignments of sequences in gene clusters, create trees for individual genes, and map the pattern of gene presence/absence onto the core genome tree.

Bacteria have the ability to laterally transfer genetic material from the environment in addition to vertically passing down their genome to offspring. Numerous methods, such as active absorption, mobile genetic elements, and viral gene transfer, are used to spread genes across bacteria. Gene duplication and loss occur regularly in addition to gene gain. The phylogenetic analysis of bacterial genomes is complicated by the presence of both vertical and horizontal transmission, which produces patterns of genetic diversity that are challenging to interpret.

Classifying genes into the core or accessory genome is a typical strategy used when analyzing datasets of bacterial genomes. All strains in a collection of isolates share core genes, and some strains share accessory genes, but not all strains share unique genes, and only one strain has unique genes.

2. Literature Review

The **pan-genome**, which is often several times larger than the core genome, is the collection of all genes found in a set of strains (for example, strains from one species). The core genome is frequently used to determine how closely related the genomes in the sample are to one another and to broadly represent the species tree, however substantial horizontal transfer has also been shown to occur in the core genome such that a tree inferred from the

diversity of the core genome does not always correspond to the phylogeny.

Various software applications try to figure out or minimize the impact of recombination on the phylogeny at the species level. Gene gain from the pan-genome can facilitate the acquisition of new metabolic processes by offering a library of functional genes. The development of drug-resistant mutations or habitat adaptability. It is now possible to find correlations between factors like habitats, symptoms, clinical manifestations, and the presence or lack of specific genes thanks to the exponentially growing number of sequenced bacterial genomes.

Pan-genomes have been created using a variety of software applications and pipelines, each with a different set of heuristics for comparing strains and creating clusters.

a. What is PanX?

PanX is a set of tools for thorough examination, dynamic exploration, and interactive visualization of bacterial pan-genomes. The DIAMOND, MCL, and phylogeny-aware post-processing components make up the analytic workflow. The visualization application consists of a number of related parts (statistical charts, gene cluster tables, alignment, comparative phylogenies, and metadata). Phylogenetic features like diversity and summary statistics like annotation can be used to quickly search and filter gene groups. The species tree can be used to map metadata and patterns of gene presence and absence. Such mapping makes it easier to find genes linked to traits like virulence, antibiotic resistance, or epidemiological factors like host age.

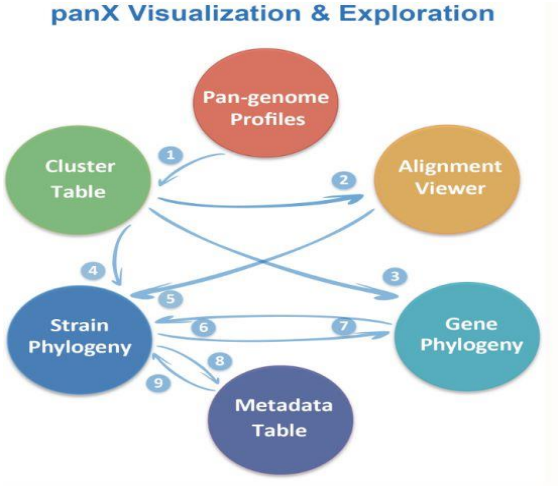

Fig.1 PanX visualization and exploration

b. Pan-genome analysis web application

Here is the browser-based visualization tool to allow users to explore the pan-genome that was created by the pipeline mentioned above. The application's design resembles a sizable dashboard, on which numerous pan-genome features can be simultaneously queried.

3. Methodology and materials used

c. Data collection:

- Go to NCBI and search Porphyromonas gingivalis.
- Download the Porphyromonas gingivalis fasta file format.
- Download Porphyromonas gingivalis strains about 5.
- Porphyromonas gingivalis

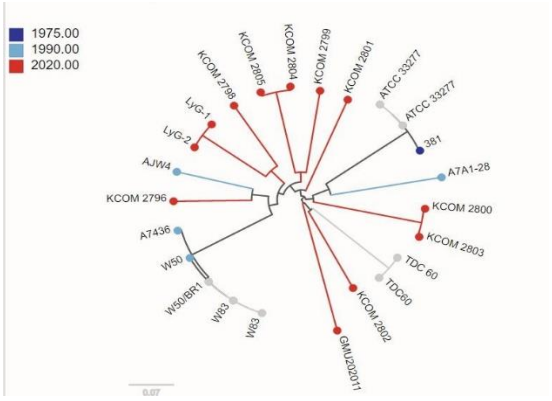

Fig.2 Data Collection

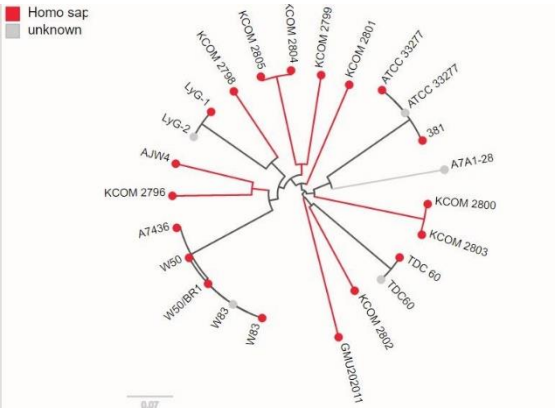

Fig.3 Host

Fig.4 Consensus

d. Orthologous gene clusters identification

PanX's computational pipeline's first phases (shown in Figure 5. Panax analysis) are mostly comparable to those of previous pan-genome construction tools.

This algorithm identifies groups of homologous genes by similarity search using **DIAMOND** and clustering by **MCL**. It then analyses the structure of the trees to construct phylogenies of these gene groupings and divides them into roughly orthologous clusters. It thus combines the speed of graph methods to identify groups of similar sequences with tree-based methods applied to individual clusters to accurately split homologous sequences into orthologous groups. The visualisation is made up of interconnected components that enable quick filtering, searching, and examination of the evolutionary history of genes.

PanX shows an alignment, phylogenetic tree, maps mutations within each gene cluster to the branches of the tree, and infers gene gain and loss for each gene cluster.

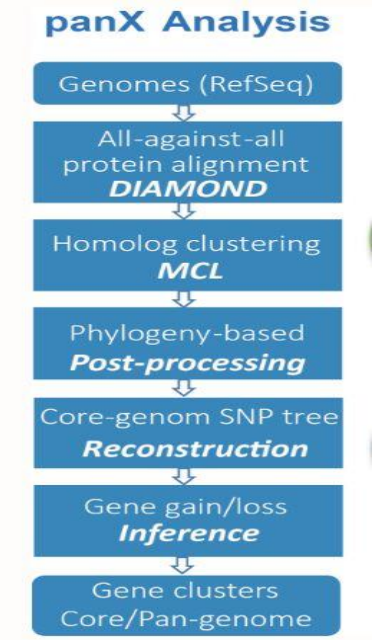

Fig.5 Panx Analysis

e.Identification of Homologous sequences

Annotated genome sequences in GenBank format are used as input by panX. It uses DIAMOND to conduct an all-against-all similarity search to find homologous proteins. If performed on a computing cluster, panX uses 64 CPUs by default, and the DIAMOND similarity search can be multi-threaded. It creates a file with all gene pairings with significant hits and the accompanying bitscore from the diamond output. Utilizing bitscore in place of e-value prevents underflow issues and combines homologous region length and similarity. **The Markov Clustering Algorithm** (MCL) uses the table of similarity scores as input to generate groups of ostensibly orthologous genes.

Ribosomal RNAs (rRNA) and other non-protein coding genes must be handled separately because DIAMOND aligns proteins. From annotated genomes in GenBank format, PanX extracts annotated

rRNAs and uses blastn to compare sequences to one another. Following that, MCL clusters the output of blastn in a manner similar to how DIAMOND compares proteins. Other sequence similarity search techniques like blastx or blastn can totally replace the DIAMOND similarity search if desired.

f. Divide and Conquer method for large datasets

First clustering small batches of genomes and then combining several batches, it is possible to eliminate the majority of these redundant comparisons. In particular, we perform the DIAMOND and MCL phases on subsets of 50 genomes (big enough to make use of DIAMOND's double indexing method, yet small enough such that the all-against-all comparison is still not expensive), and derive gene clusters of this "sub-pan-genome". The sample sequences of all gene clusters are then combined to form a "pseudo genome," which includes the representative sequences of each gene cluster "representing the batch as a whole. Then, using the DIAMOND + MCL procedures, the pseudo genomes representing the various batches are once more grouped.

Sequences represented by the pseudo genomes are eventually combined to create whole clusters. For really large pan-genomes, this "divide-and-conquer" approach can be used repeatedly.

g. Splitting into orthologous clusters

According to our observations, it is best to aggressively cluster proteins and separate clusters containing paralogous sequences in a post-processing phase. In a phylogenetic tree, the groups of paralogs are frequently easily discernible. After aligning the protein sequences with MAFFT, PanX reconstructs trees from the sequences in each cluster. By introducing a gap of length three for each gap in the amino acid alignment, the protein alignment is then utilized to create a codon alignment of the corresponding nucleotide sequences. Using FastTree, panX then reconstructs a tree from nucleotide sequence alignment. FastTree's runtime scales roughly as n3/2 with the number of sequences.

h. Generating gene orthologous groups by splitting into clusters

Groups of distantly related genes, such as those resulting from an ancient duplication, are connected by long branches and can be easily identified in a gene tree—at least for pan-genomes of low or moderate diversity—because branch lengths reflect evolutionary distances among genes within one cluster. At branches whose length exceeds an adaptive threshold, PanX divides trees into subtrees.

The core genes' genetic diversity will be of the same order of magnitude as the mutational distance from the collection of genomes' most recent common ancestor (MRCA). Branches that are significantly longer will likely correlate to duplications that occurred before the MRCA. Therefore, it is necessary to cut the cluster along these branches.

i. Splitting of closely identical paralogs

Recent gene duplication events will be missed if branches are split longer than bc. PanX determines a paralogy score for each branch in each tree in order to more accurately identify paralogous groups. The number of strains represented on both sides of a branch in the phylogeny determines the paralogy score of that branch. All branches can be simultaneously determined for this score in linear time using two tree traversals.

j. Fragmented clusters are merged

Only a few genes are incorrectly clustered, either because no homology was initially found or because the MCL clustering

process failed. Such unclustered sequences appear as several singleton clusters that are of the same size. PanX determines the average length of sequences in each cluster and looks for peaks in the distribution of gene cluster length to find those sequences. This empirical gene length distribution exhibits unclustered genes as spikes, which panX can detect by looking for peaks in the distribution in comparison to a smoothed background distribution.

k. Core genome SNP tree

To create a core-genome SNP matrix, PanX extracts all variable locations from the nucleotide alignments of all single-copy core genes. Using FastTree and RaxML, this SNP matrix is used to create a core genome phylogenetic tree, which is then further improved, using a method similar to that employed in nextflu. This core genome tree may not accurately depict the history of each gene in the core genomes due to homologous recombination, and as only variable sites were employed, branch lengths do not correspond to sequence similarity. The distribution of phenotypes and the evolution of the mobile genome may both be studied using this core genome SNP phylogeny as a scaffold. It is nevertheless a fair approximation of the relationships between the many strains.

l. Algorithm of SNP CALLING

Step 1: Load the Porphyromonas gingivalis fasta file and read the entire sequence as seq_a.

Step 2: Then load one of the strains as a fasta file and read the entire sequence as seq_b.

Step 3: Convert the sequences in tuples and then compare them by using zip.

Step 4: With the help of for loop find the mutations.

4. Result

```
missense C - G
silent T - A
silent T - A
silent T - A
silent T - A
silent T - A
missense C - G
missense G - C
silent T - A
silent T - A
silent T - A
missense G - C
silent T - A
silent T - A
silent T - A
missense G - C
silent T - A
silent T - A
missense C - G
missense C - G
missense G - C
silent T - A
missense G - C
silent T - A
missense G - C
missense C - G
silent T - A
silent T - A
silent T - A
missense G - C
missense G - C
missense G - C
missense C - G
silent T - A
silent T - A
silent T - A
silent T - A

nonsense:
0
missense:
84
silent:
100
{'Count of nonsense': 0, 'Count of missense': 84, 'Count of silent': 100}
```

Fig.6 SNP calling output

m. Cluster post-processing stage

In the cluster post-processing stage, phylogenetic trees for a gene cluster have already been inferred. Using these trees and the joint maximum likelihood method as implemented in tree time, PanX uses these trees to infer the ancestral sequences of internal nodes. This ancestral reconstruction is used to map the branches of the tree with likely mutations.

n. Algorithm of Phylogenetic Trees

Step 1: First find the alignment between the Porphyromonas gingivalis and 5 strains.

Step 2: Convert it into .phy or .phylip(all the genes of different strains and Porphyromonas gingivalis of the same length).

Step 3: Calculate the distance matrix and distance tree.

Step 4: From the distance tree construct the phylogenetic tree using **UPGMA** algorithm.

Input: Alignment of Porphyromonas gingivalis and 5 strains.

Output:

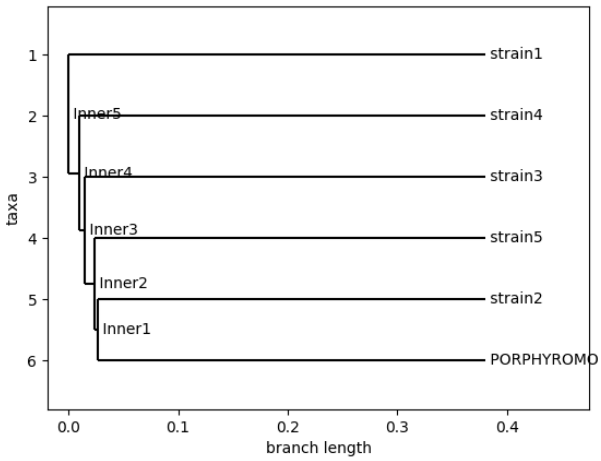

Fig.7 Phylogenetic Trees output

o. Gene gain or loss

Then, using a similar ancestral inference approach, we determine whether each gene cluster is present or absent on internal nodes of the core genome SNP tree. Based on this ancestral reconstruction, certain gain and loss events are connected to specific branches. In order to enhance the possibility that the observed presence/absence pattern of genes will occur, the gain and loss rates are tuned. The ideal gain and loss rates are always greater than each other, however, the ratio varies amongst species, with a median ratio of 22.

5. Pan-genome analysis

We created 120 artificial pan-genome simulations with 30 artificial genomes each to evaluate the accuracy of clustering algorithms. The simulation allows for horizontal transmission as well as gene loss and gain, and it evolves ancestors along coalescent trees. One sample gene from each of the 37 KEGG ortholog groups found in the Escherichia coli strain K-12 served as the starting point for the simulation in order to produce accurate ancestral sequences. This resulted in 2803 unique genes. We created pan-genome simulations using these as the ancestral sequences by performing the following steps: We created associated trees with various rates of horizontal transfer for each of the 2803 genes.

All 2803 genes evolve in accordance with the same clonal genealogy of the population, or one common species tree, if the gene transfer rate is zero. In contrast, if some genes are impacted by gene transfer, the individual gene trees may vary. However, because of their shared connection to clonal genealogy, the gene trees continue to be highly dependent on one another.

Tool used: pan-genome analysis http://pgaweb.vlcc.cn

Input: Porphyromonas gingivalis and strains.

Output: Based on the length of genome

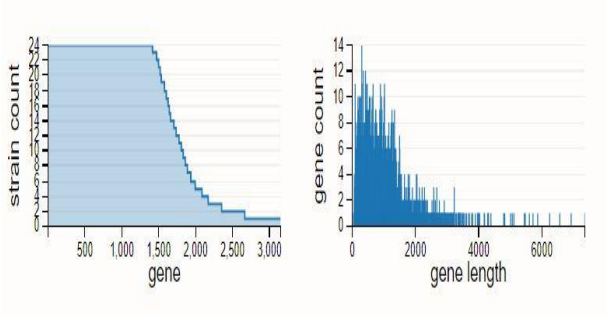

Fig.8 Pan-genome Analysis

6. Conclusion

Large collections of bacterial genomes, which are characterized by the evolution of specific genes as well as the gain and loss of genes, are intended to be explored using PanX. It is designed to prioritize combining breadth and depth; in addition to summary statistics and species trees, it also allows users to choose particularly interesting groupings of genes or do gene-specific searches. Individual mutations and gain/loss events can then be mapped to the gene tree and the core tree, respectively, allowing for detailed analysis of alignments and phylogenetic trees of genes.

The meta-data, such as resistance characteristics connected to the various strains, can then be compared to the evolutionary patterns of genes. It can help studies by combining meta-data with the molecular evolution of genes and genomes in one visualization or its derivatives might be used to track food-borne epidemics, follow the global spread of drug-resistant bacteria, or help with infection management with the rising availability and prompt publishing of such data from routine surveillance.

7. References

[1] Ding, Wei, Franz Baumdicker, and Richard A. Neher. "panX: pan-genome analysis and exploration." *Nucleic acids research* 46, no. 1 (2018): e5-e5.

[2] Kim, Y., Gu, C., Kim, H. U., & Lee, S. Y. (2020). Current status of pan-genome analysis for pathogenic bacteria. *Current opinion in biotechnology*, *63*, 54-62.

[3] Soucy S.M., Huang J., Gogarten J.P.. Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 2015; 16:472–482.

[4] Ochman H., Lawrence J.G., Groisman E.A.. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000; 405:299–304.

[5] Chewapreecha C., Marttinen P., Croucher N.J., Salter S.J., Harris S.R., Mather A.E., Hanage W.P., Goldblatt D., Nosten F.H., Turner C. et al.. Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014; 10:e1004547.

[6] Huson D.H., Bryant D.. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 2006; 23:254–267.

[7] Holden M. T.G., Hsu L.-Y., Kurt K., Weinert L.A., Mather A.E., Harris S.R., Strommenger B., Layer F., Witte W., de Lencastre H. et al.. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. Genome Res. 2013; 23:653–664.

[8] Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics. 2014; 30:2068–2069.

[9] Vernikos G., Medini D., Riley D.R., Tettelin H.. Ten years of pan-genome analyses. Curr. Opin. Microbiol. 2015; 23:148–154.

[10] Vernikos, G., Medini, D., Riley, D. R., & Tettelin, H. (2015). Ten years of pan-genome analyses. *Current opinion in microbiology*, *23*, 148-154.