Pangenomic analysis of thermophiles

Iouliana Parisi, Alexios Loukas & Ilias Kappas

Department of Genetics, Development & Molecular Biology, School of Biology, Aristotle University of Thessaloniki, 541 24, Greece
iouliana@bio.auth.gr, aloukas@bio.auth.gr, ikappas@bio.auth.gr

INTRODUCTION

There has been growing interest in pangenomic studies, driven by the advancement of high-throughput next-generation sequencing and computational technologies. The analyses of pangenomes offer valuable insights into the genomic diversity and evolutionary dynamics within taxa. Thermophiles thrive at temperatures between 50 and 122˚C and are typically found in thermal vents and hot springs, where other life forms cannot survive. They are represented by most major prokaryotic lineages covering immense taxonomic, functional, physiological, and ecological diversity [1]. In this project, we analyze the pangenome of the largest, up to date, collection of thermophiles from Archaea and Bacteria by selecting **200 complete genomic sequences of thermophiles** from the NCBI Genomes Database. The dataset was not limited to species level but included genomes of strains and isolates from multiple thermophilic species found in Thermobase [2]. For the analysis, micropan [3] was employed, a platform for pangenome construction, visualization, and exploration across large prokaryotic datasets.

RESULTS

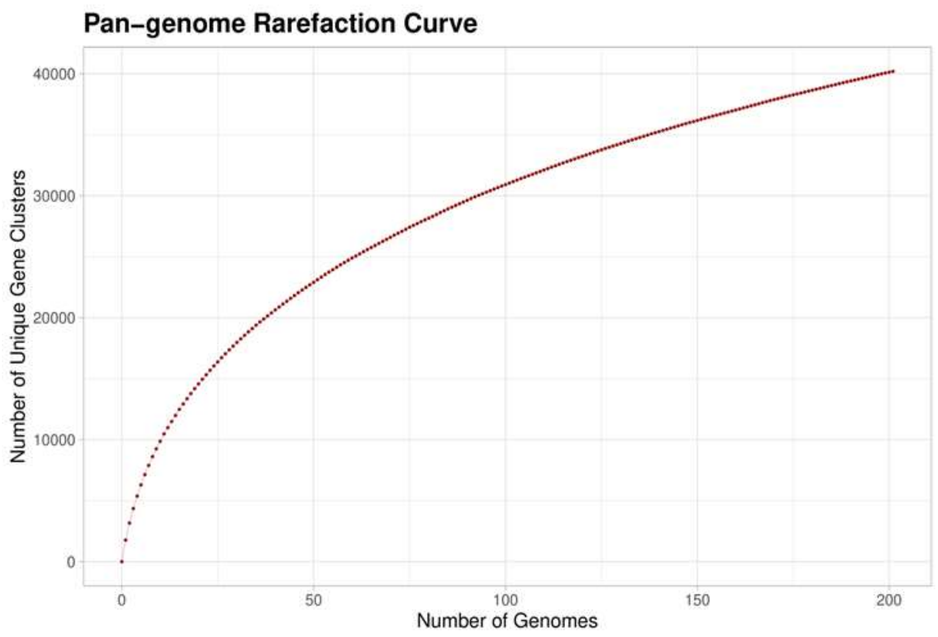

Fig. 1. Rarefaction curve of the open thermophilic pangenome.

The pangenome analysis uncovered **40200 gene clusters** throughout the 200 genomes. The estimate for the total number of clusters existing in this group was 57462, a rather accurate assessment using such a large dataset. The application of Heaps' law on the data gave an **alpha value of 0.474** indicating an open pangenome **(Fig. 1)**.

Using a binomial mixture model estimate, gene clusters were best grouped into 8 categories based on their occurrence frequency across genomes. The pangenome visualization **(Fig. 2)** shows a **significantly small core genome**, while over **75% of it consists of rare, unique clusters** (pink). Similarly, micropan simulations indicate that the average thermophilic genome is mostly made up of low detection probability clusters, with only a small portion belonging to the core genome (blue). Comparatively, an open pangenome with reported conserved core [4] was constructed using 5 thermophilic *Thermus* species genomes from the original dataset.

Additionally, based on genome distances computed from micropan, a dendrogram was constructed. To enhance its accuracy, weights were appointed to shell (core) genes. The dendrogram **(Fig. 3)** shows taxonomic consistency, with a **clear separation between Bacteria and Archaea, and genome clustering by phylum**.

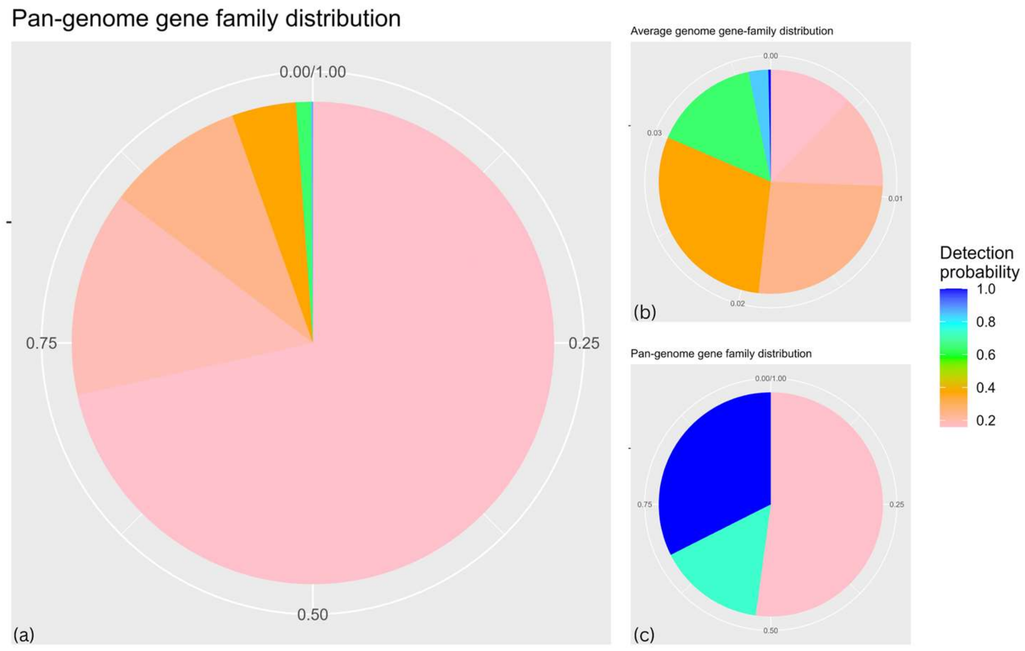

Fig. 2. Distribution of gene clusters (gene families) in the thermophilic pangenome based on their detection frequencies in all genomes (a) and in the average simulated genome (b) in contrast to a typical open pangenome visualization (c) of the genus *Thermus*.

DISCUSSION

The observed widely open pangenome indicates **high genomic diversity in thermophiles**, potentially reflecting adaptation to varied and extreme environmental conditions. A considerably small core genome suggests that **only a few genes have been conserved during prokaryotic diversification**. Housekeeping and thermo-adaptive functions are expected within the core and extended core, though these may vary across genomes.

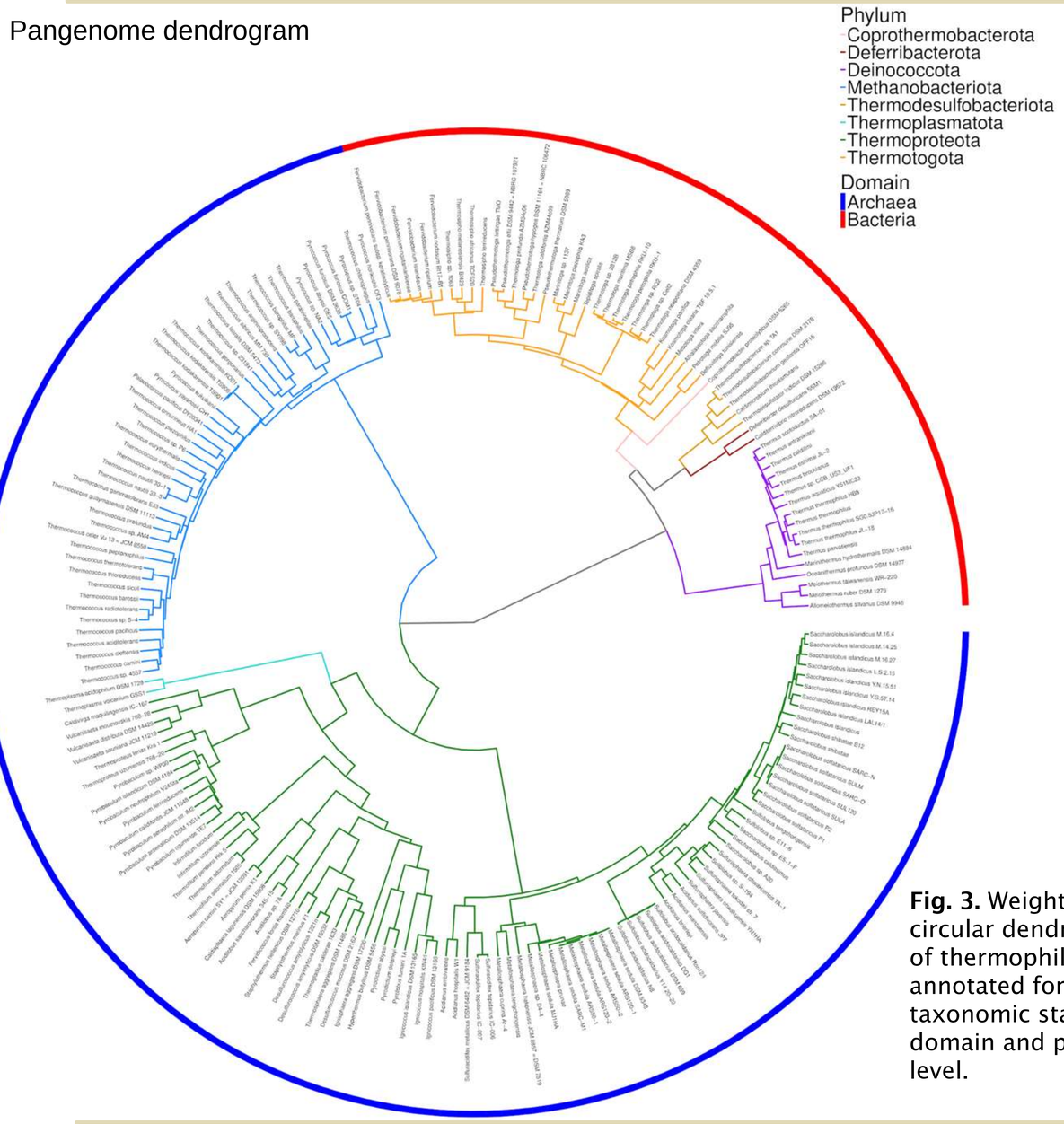

Fig. 3. Weighted circular dendrogram of thermophiles, annotated for their taxonomic status on domain and phylum level.

On the other hand, the large number of unique clusters depicts the vast genomic diversity of thermophiles, which is expected given the taxonomic breadth of the dataset spanning both Bacteria and Archaea. **Prokaryotes demonstrate tremendous variation in gene content even within strains** [5]. The study of extremophiles at the upper thermal boundary of life is of enormous interest considering the evolutionary implications of gene transfer and genetic exchange, their biotechnological applications, and their implication in theories on the origin of life on Earth or other planetary bodies [6].

REFERENCES

[1] Zhou Y, et al. Diversity of thermophilic prokaryotes. Front Microbiol. 2022;13:984632.
[2] DiGiacomo J, et al. ThermoBase: a database of the phylogeny and physiology of thermophilic and hyperthermophilic organisms. PLoS One. 2022;17(5):e0268253.
[3] Snipen L, Liland KH. micropan: an R-package for microbial pan-genomics. BMC Bioinformatics. 2015;16:79.
[4] Vishnivetskaya TA, et al. Complete genome analysis of *Thermus parvatiensis* and comparative genomics of *Thermus* spp. provide insights into genetic variability and evolution of natural competence as strategic survival attributes. Front Microbiol. 2017;8:1410.
[5] Tettelin H, Medini D. The pangenome: Diversity, dynamics and evolution of genomes. Cham: Springer; 2020.
[6] de la Haba RR, et al. Extremophiles: Microbial genomics and taxogenomics. Front Microbiol. 2022;13:984632.

Acknowledgements

This work was supported by "Aristotelis" AUTh HPC cluster.