PangenePro: an automated pipeline for rapid identification and classification of gene family members

Kinza Fatima*[1] and Muhammad Tahir Ul Qamar*†[2]

[1]University of California [Riverside] – United States
[2]Government College University of Faisalabad – Pakistan

Abstract

Genome data has accumulated rapidly in recent years due to the influx of next-generation sequencing technologies, causing the availability of large datasets, including plant genomes in public databases. Plants are constantly exposed to various environmental stresses, prompting the evolution of several mechanisms and gene families to confer tolerance and resistance against these stresses. In the past few years, numerous gene families have been studied and published using the datasets available in public databases. The genome-wide Identification (GWI) studies aid in identifying and characterizing the gene members, providing insights into their structural and functional diversity, as well as their expression patterns under stress conditions. These in-silico analyses provide an initial framework for further functional elucidation and breeding research. However, these GWI studies are often single-reference-based, insufficient to capture the genetic diversity of multiple individuals within a species. Additionally, the manual GWI of gene family members is time-consuming and requires significant effort with a likelihood of errors and biases. The objective of this study is to develop a pipeline that automates the identification of gene family members from multiple plant genomes and classifies these members into pan-gene sets. The first module of this pipeline identifies the members of the gene family of interest from multiple genomes through alignment, filtering, and domain profiling. The next module categorizes the identified members from multiple genomes at the pangene level into Core, Accessory, and Unique genes. Finally, these pangenes sets are visually represented through summary tables, bar plots, upset plots, and Venn diagrams. To validate this pipeline, an assessment was conducted on both simpler (Arabidopsis, 2n with genome size in Mbp) and complex (Arachis, both 2n and 4n with genome size in Gbs) genomes. This pipeline enabled the automated, quicker, and efficient identification of gene family members from multiple related genomes at the pan-gene level.

Keywords: PangenomeWide, Automated pipeline, Pangene sets, Bioinformatics, Gene families, plants

*Speaker
†Corresponding author: tahirulqamar@gcuf.edu.pk