GALEON - A bioinformatic tool for gene cluster identification and analysis in chromosome-level assemblies

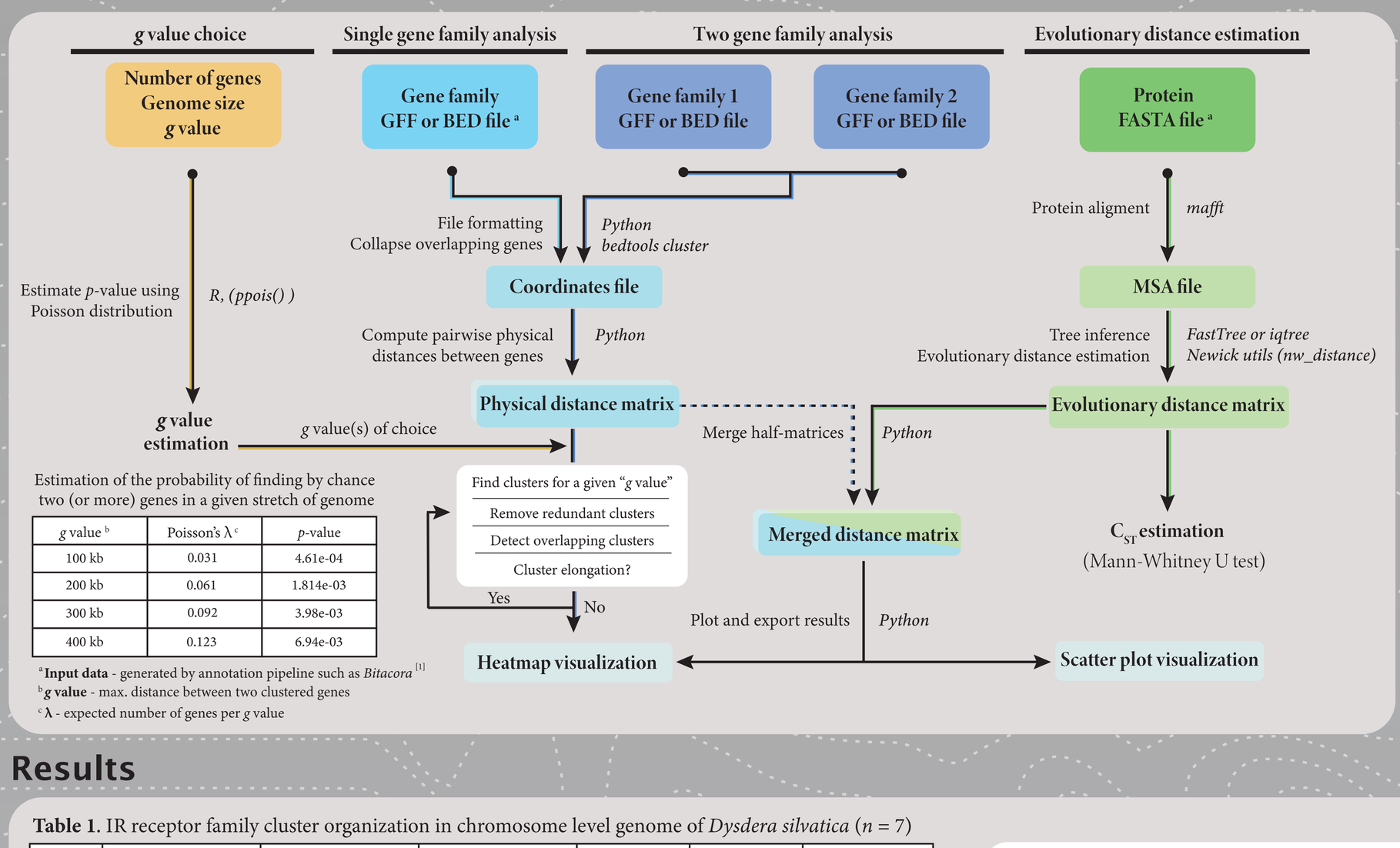
Vadim A. Pisarenco [1,2], Joel Vizueta [3] and Julio Rozas [1,2]

(1) Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain
(2) Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
(3) Villum Centre for Biodiversity Genomics, Section for Ecology and Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark

Abstract

Gene clusters, groups of genes encoding functionally-related proteins, are commonly found in eukaryotic genomes. One of the most abundant types are gene families, those encoding a set of homologous genes commonly originated through gene duplication often via unequal crossing-over. As a result, they are found arranged in tandem in the genome. Despite the increasing number of whole genome information for many species, the comprehensive evolutionary analysis of gene family members remains largely unexplored. Two primary, not exclusive, challenges hinder this exploration: the analysis of large gene family sizes, and those of very recent (young) family members. These issues stem from limitations in current assemblies (including many of those based on long reads) that can not accurately assemble extensive stretches of repetitive DNA regions (such as those of recently originated copies). Current sequencing techniques (based on long reads and chromatin contacts) offer a promising avenue for addressing these challenges. The discovery and visualization of gene clusters is usually case-specific, involving arbitrary criteria and custom bioinformatic pipelines. To overcome such limitations, we present GALEON, a user-friendly bioinformatic tool to identify, analyze and visualize physically clustered gene family genes in chromosome-level genomes. The software uses simple input file formats with gene coordinates (BED or GFF3), and protein sequence data. Specifically, the gene cluster analysis is assessed by analyzing the distribution of pairwise physical distances between the gene set's members and the average genome gene density. GALEON also allows the study (and comparison) of two gene families at once and, if the protein sequence data is provided, can also explore the relationship between physical and evolutionary distances. Overall, GALEON represents a novel tool for the study of clustered genes to gain valuable insights into the origin, evolution and function of gene families, while also providing an utility to evaluate the local quality of assembled genomic regions.

Methods



Estimation of the probability of finding by chance two (or more) genes in a given stretch of genome

g value [b]	Poisson's λ [c]	*p*-value
100 kb	0.031	4.61e-04
200 kb	0.061	1.814e-03
300 kb	0.092	3.98e-03
400 kb	0.123	6.94e-03

[a] **Input data** - generated by annotation pipeline such as *Bitacora* [1]
[b] ***g* value** - max. distance between two clustered genes
[c] **λ** - expected number of genes per *g* value

C_{ST} estimation

C_{ST} - measures the proportion of the evolutionary distance that is attributable to unclustered genes

$$C_{ST} = \frac{D_T - D_C}{D_T}$$

D_T - average pairwise evolutionary distances (amino acid replacements per site) between family genes

$$D_T = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}$$

- n - number of gene family members in the surveyed chromosome/scaffold (or the genome)
- d_{ij} - evolutionary distance between sequences i and j

D_C - average pairwise evolutionary distances between copies from within a cluster, averaged across all clusters of the same scaffold (or the genome)

$$D_C = \frac{1}{m} \sum_{k=1}^{m} D_{Ck}$$

- m - number of clusters (in a scaffold or the genome)

D_{Ck} - average pairwise evolutionary distances within the cluster k

$$D_{Ck} = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}$$

- n - number of gene copies in cluster k
- d_{ij} - evolutionary distance between sequences i and j

Results

Table 1. IR receptor family cluster organization in chromosome level genome of *Dysdera silvatica* ($n = 7$)

Family	Chromosome ID	Num. of clusters	Clustered genes	Total genes	C_{ST}	MU *p*-value
IR	ChrX	0	0	10	NA	NA
IR	Chr1	7	19	32	0.757	***
IR	Chr2	5	11	28	0.844	***
IR	Chr3	2	5	16	0.982	*
IR	Chr4	1	2	9	0.418	***
IR	Chr5	5	21	32	0.462	***
IR	Chr6	1	2	11	0.616	*

NA, not applicable; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$. Data from [2].

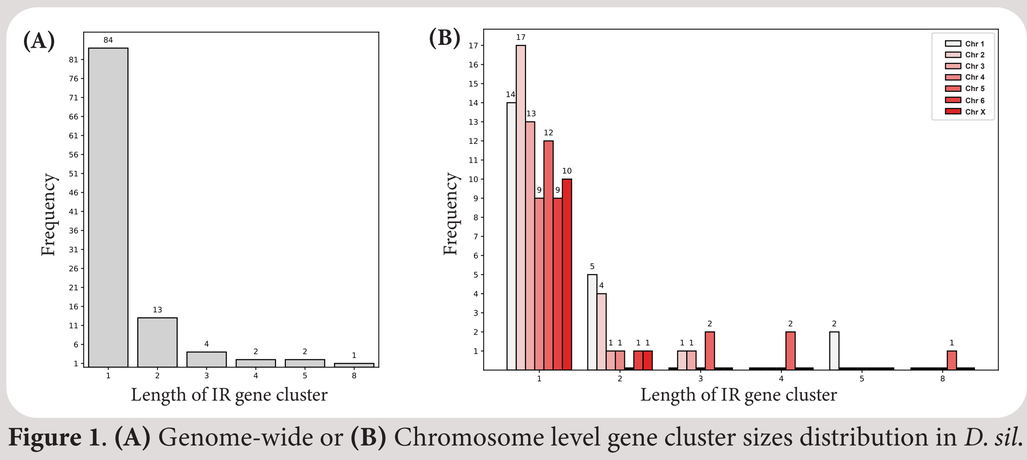


Figure 1. (A) Genome-wide or **(B)** Chromosome level gene cluster sizes distribution in *D. sil.*

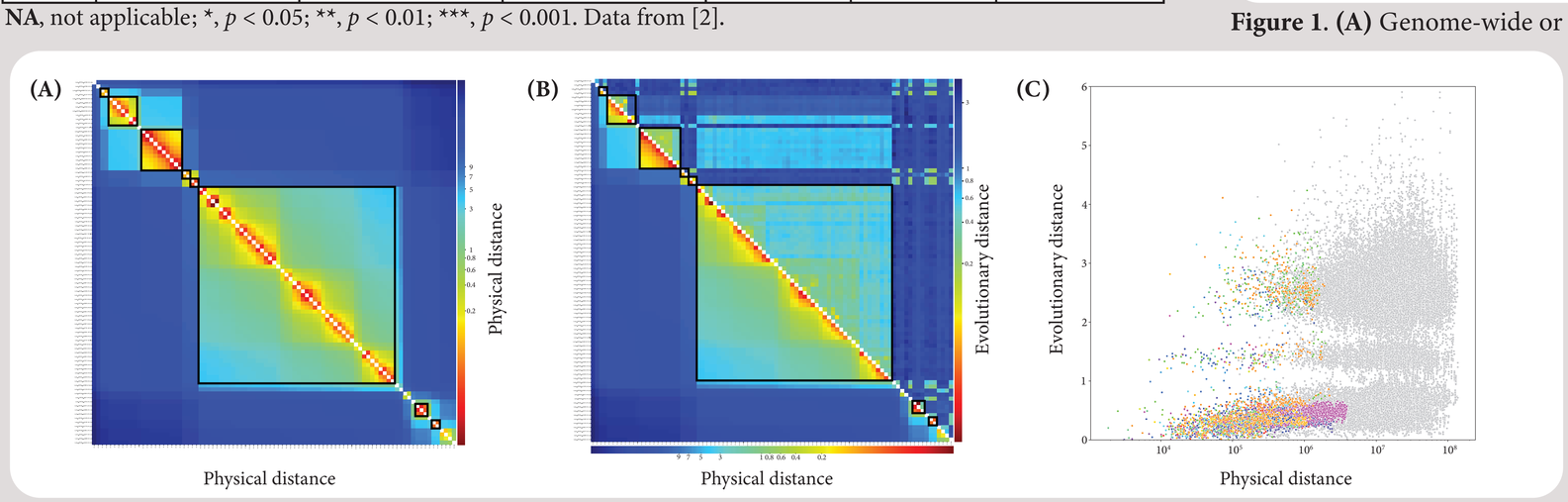


Figure 2. GPCR gene family analysis in chromosome level genome of *Lautoconus ventricosus* ($n = 35$). **(A)** Physical distances (in 100 kb units) plotted as a heatmap in Scaffold_11. **(B)** Physical distances (upper matrix) vs Evolutionary distances (lower matrix) in amino acid substitutions per site in Scaffold_11. **(C)** Genome-wide scatter plot of physical distances in bp vs Evolutionary distances in amino acid substitutions per site. Clustered genes' data is shown in color; Not-clustered genes' data is shown in grey. Data from [3].

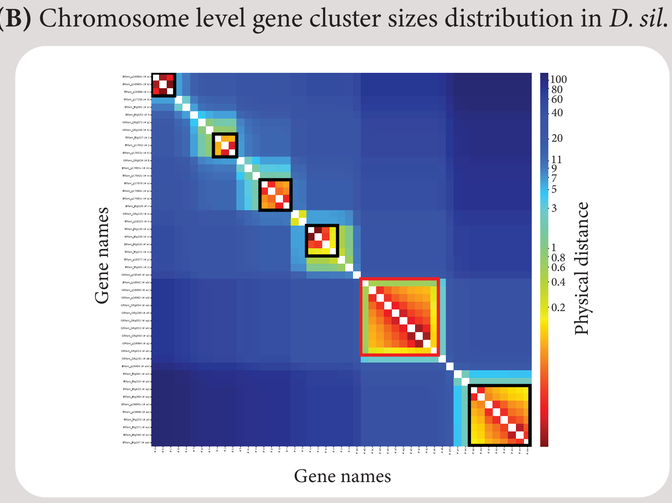


Figure 3. Two gene families analysis in *D. sil.* Physical distances (in 100 kb units) between IR and GR gene families. Clusters are formed by one family gene members (black square) or two families genes (red square). Data from [2].

References

[1] Vizueta *et al.* (2020). *Molecular Ecology Resources*, 20, 1445–1452.
[2] Escuer *et al.* (2022). *Molecular Ecology Resources*, 22(1), 375-390.
[3] Rondón *et al.* (2023). *Molecular Phylogenetics and Evolution* (under review).

Acknowledgements

All the members of *Evolutionary Genomics & Bioinformatics* research group. Funded by Ministerio de Ciencia, Innovación y Universidades of Spain (PID2019-103947GB-C21 and PID2022-138477NB-C22) and Generalitat de Catalunya (2021SGR00279)