Natural Family-Free Genomic Distance

Diego P. Rubert
Universidade Federal de Mato Grosso do Sul
Fábio V. Martinez
Universidade Federal de Mato Grosso do Sul
Marília Braga (mbraga@cebitec.uni-bielefeld.de)
Bielefeld University: Universitat Bielefeld https://orcid.org/0000-0002-4092-2646

Research Article

Keywords: Comparative genomics, Genome rearrangement, DCJ-indel distance

Posted Date: February 18th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-198423/v1

Abstract

Background: A classical problem in comparative genomics is to compute the rearrangement distance, that is the minimum number of large-scale rearrangements required to transform a given genome into another given genome.

The traditional approaches in this area are family-based, i.e., require the classification of DNA fragments of both genomes into families. Furthermore, the most elementary family-based models, which are able to compute distances in polynomial time, restrict the families to occur at most once in each genome. In contrast, the distance computation in models that allow multifamilies (i.e., families with multiple occurrences) is NP-hard. Very recently, Bohnenk amper et

al. (J. Comput. Biol., 2020) proposed an ILP formulation for computing the genomic distance of genomes with multifamilies, allowing structural rearrangements, represented by the generic double cut and join (DCJ) operation, and content-modifying insertions and deletions of DNA segments. This ILP is very efficient, but must maximize a matching of the genes in each multifamily, in order to prevent the free lunch artifact that would otherwise let empty or almost

empty matchings give smaller distances.

Results: In this paper, we adopt the alternative family-free setting that, instead of family classification, simply uses the pairwise similarities between DNA fragments of both genomes to compute their rearrangement distance. We adapted the ILP mentioned above and developed a model in which pairwise similarities are used to assign weights to both matched and unmatched genes, so that an optimal solution does not necessarily maximize the matching. Our model

then results in a natural family-free genomic distance, that takes into consideration all given genes, without prior classification into families, and has a search space composed of matchings of any size. In spite of its bigger search

space, our ILP seems to be boosted by a reduction of the number of co-optimal solutions due to the weights. Indeed, it converged faster than the original one by Bohnenk amper et al. for instances with the same number of multiple

connections. We can handle not only bacterial genomes, but also fungi and insects, or sets of chromosomes of mammals and plants. In a comparison study of six fruit fly genomes, we obtained accurate results.

Full Text

This preprint is available for download as a PDF.

Figures

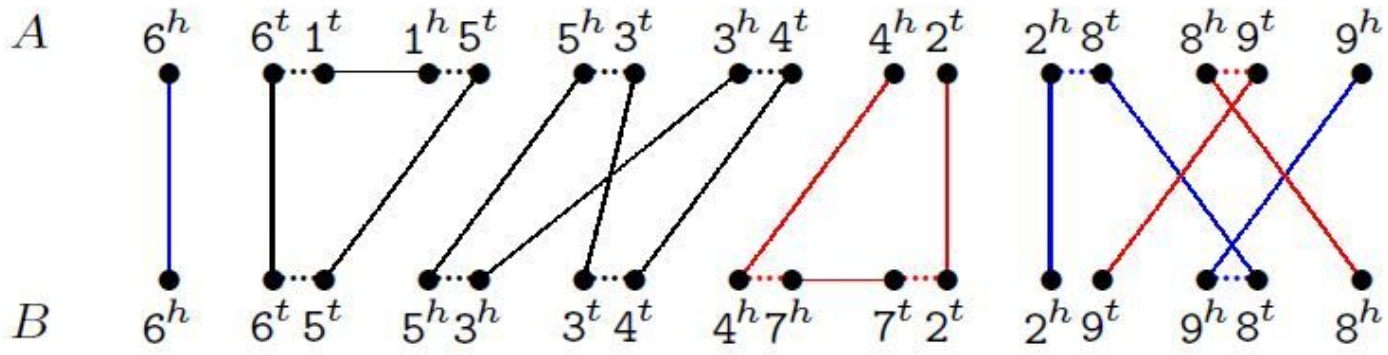

Figure 1

Relational diagram of two singular genomes. For genomes A = f[61534]; [289]g and B = f[653472]; [98]g, the relational diagram contains two cycles, two AB-paths (represented in blue), one AA-path and one BB-path (both represented in red). Short dotted horizontal edges are adjacency edges, long horizontal edges are indel edges, top-down edges are extremity edges.

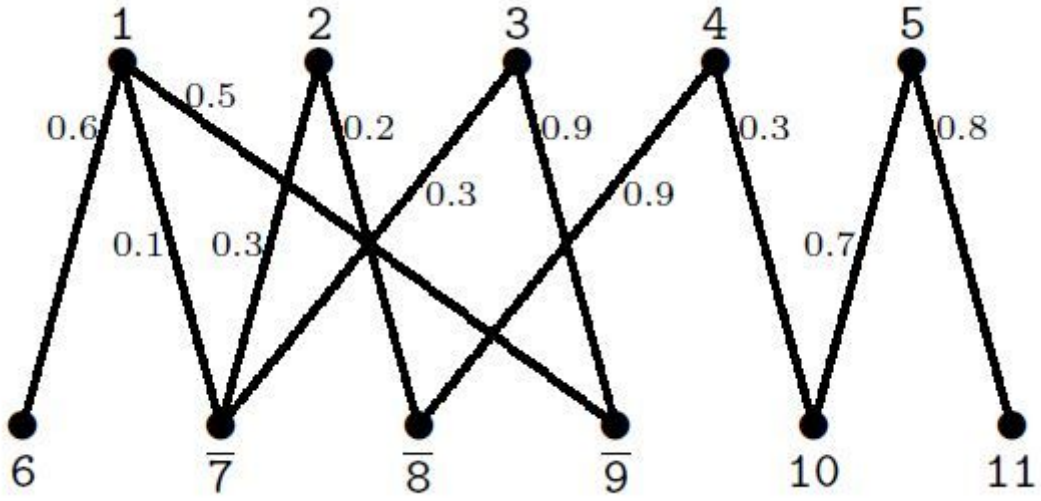

Figure 2

please see the manuscript file for the full caption

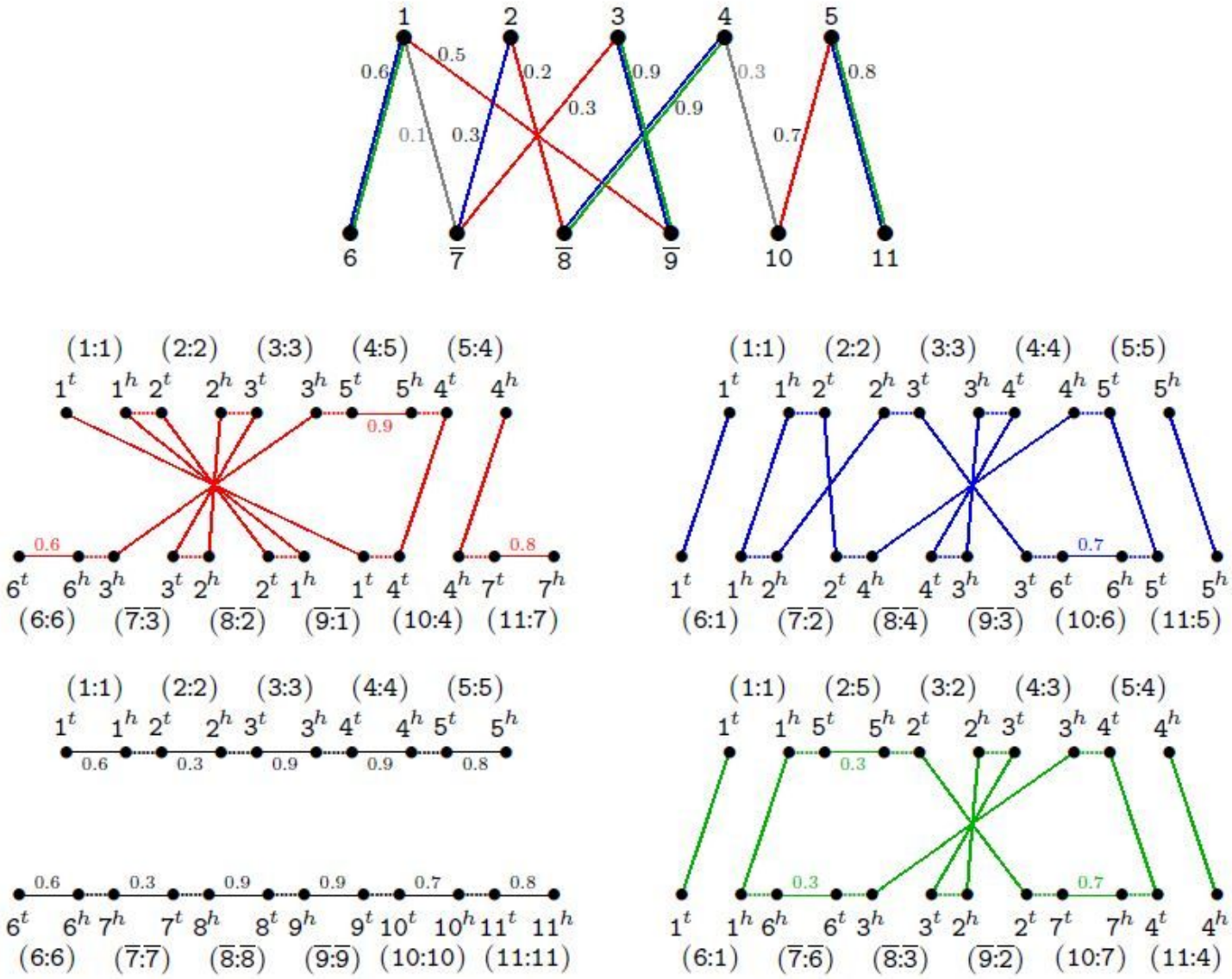

Figure 3

please see the manuscript file for the full caption

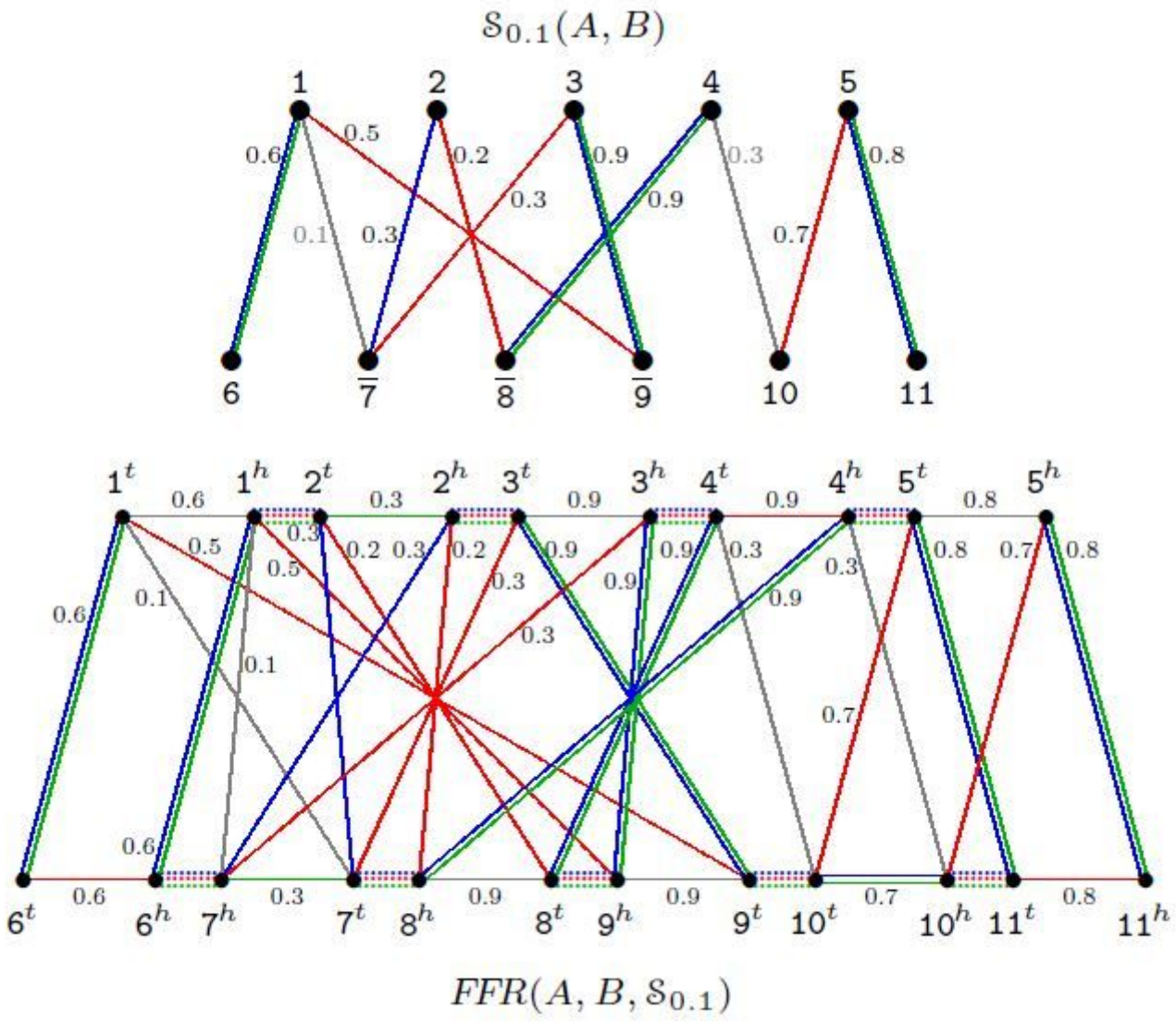

Figure 4

please see the manuscript file for the full caption

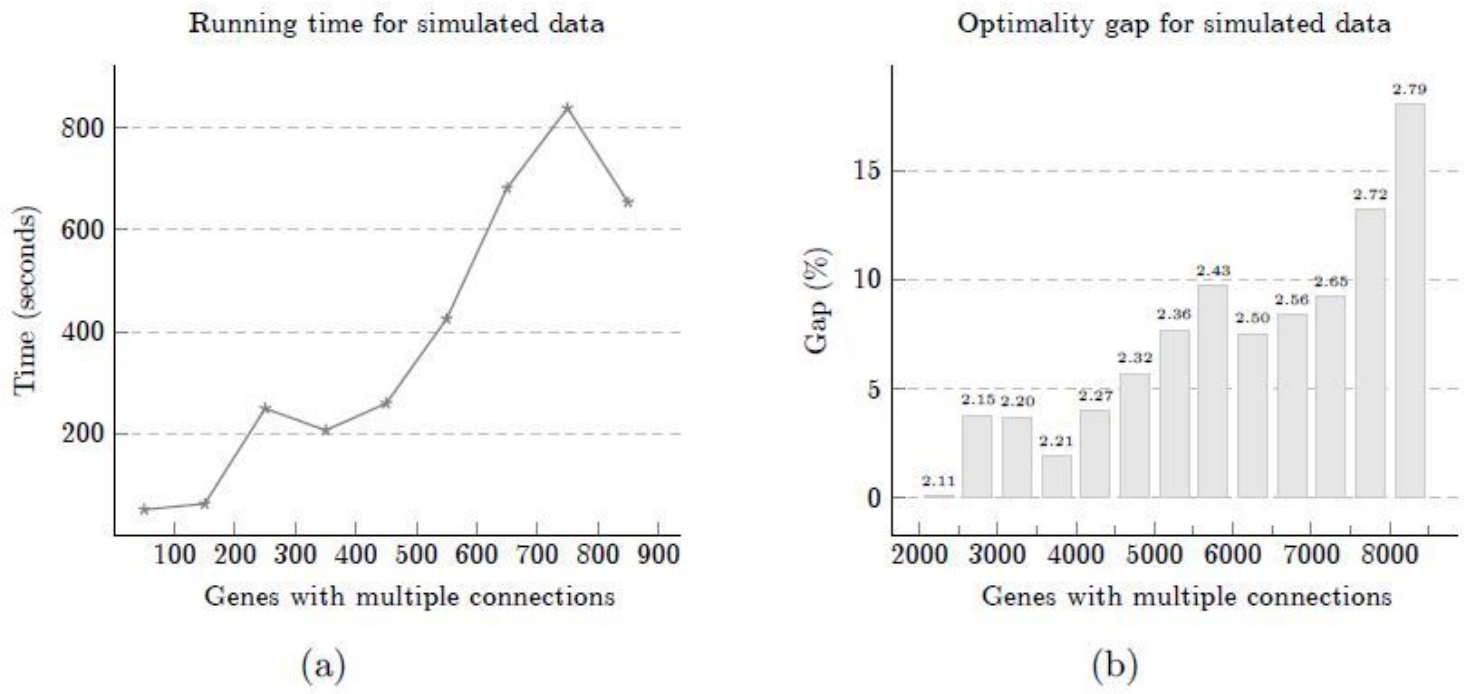

Figure 5

Performance of the ILP computing the family-free DCJ-indel distance of simulated genomes. The experiment results are displayed in two parts and in both of them instances are grouped by the number of genes with multiple connections (i.e. vertices with degree > 1 in δ0:1): (a) shows the average running time for instances grouped in intervals of 100 and up to 900, and (b) shows the average optimality gap and the average number of connections for groups of instances that did not nish within the time limit of 1 hour (in intervals of 500).

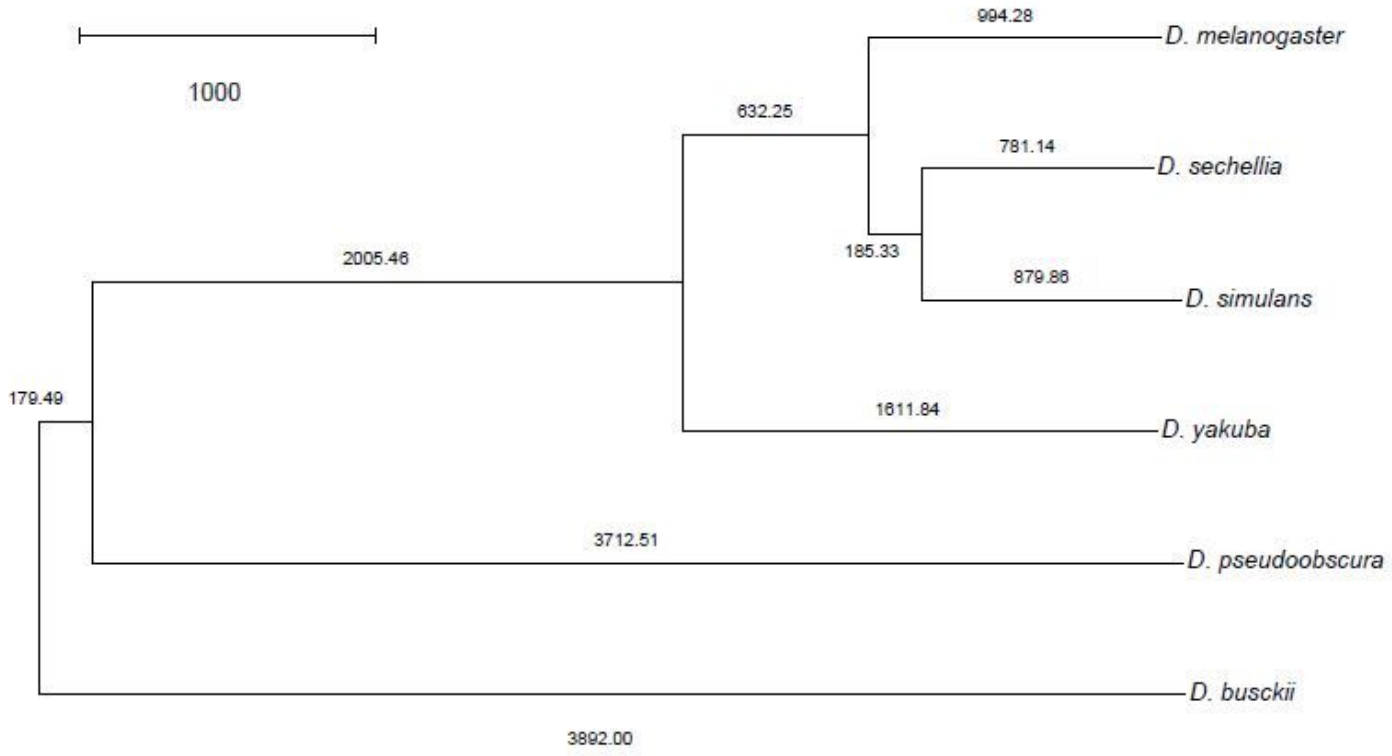

Figure 6

Phylogenetic tree computed based on the distances given by the family-free approach. This tree was computed by the Neighbor-Joining method [27, 28] based on distance matrices of pairwise comparisons of complete Drosophila genomes calculated by Diff.

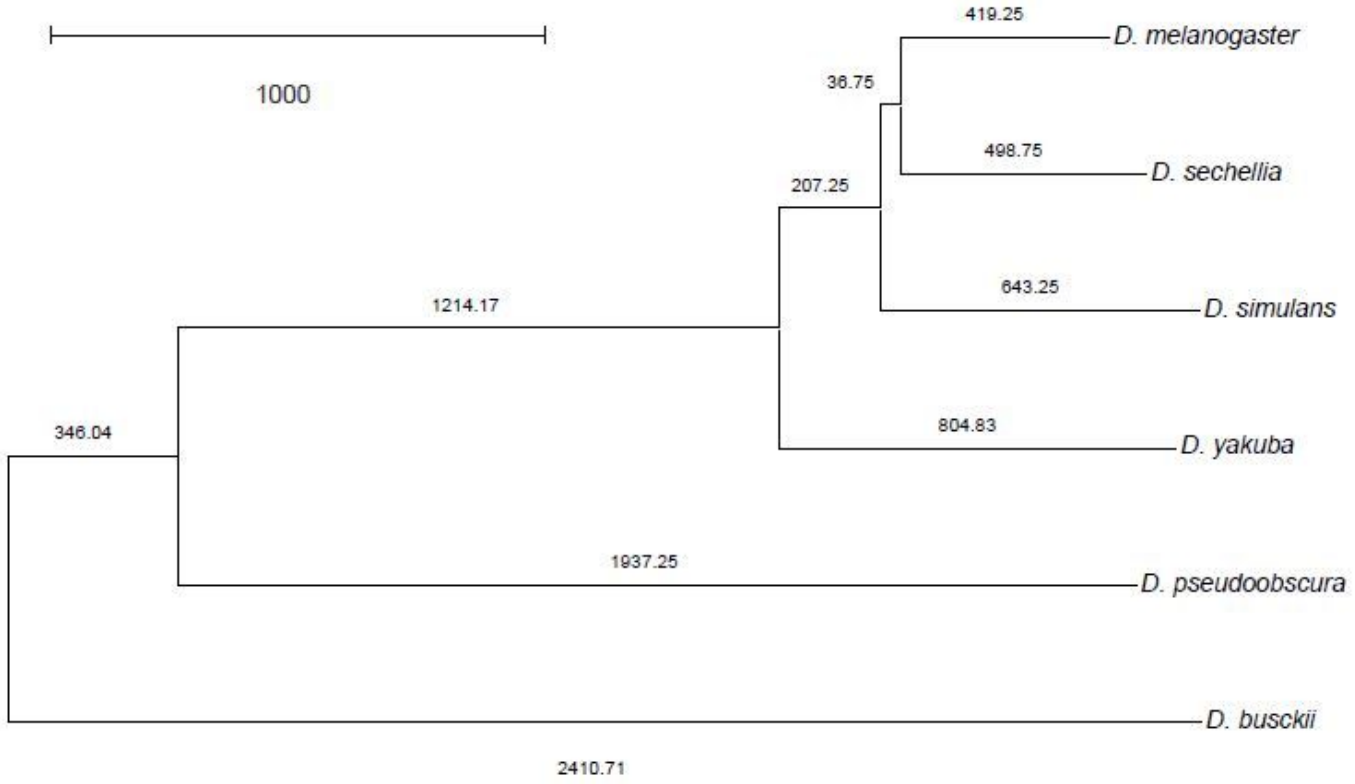

Figure 7

Phylogenetic tree computed based on the distances given by the family-based approach. This tree, reproduced from the results originally published in [16], was computed by the Neighbor-Joining method based on distance matrices of pairwise comparisons of complete Drosophila genomes calculated by Ding. The gene families here were generated in [16] by computing OMA orthologies [25] on the same genome assemblies used in the present study.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- additionalfile1.pdf
- additionalfile2.ods