compare_genomes: a comparative genomics workflow to streamline the analysis of evolutionary divergence across genomes

Jefferson Paril, Tannaz Zare, and Alexandre Fournier-Level*
School of BioSciences, University of Melbourne, Parkville, VIC, Australia
Corresponding authors: alexandre.fournier@unimelb.edu.au

Abstract

Summary

The dawn of cost-effective genome assembly is enabling deep comparative genomics to address fundamental evolutionary questions by comparing the genomes of multiple species. However, comparative genomics analyses often deploy multiple, often purpose-built frameworks, limiting their transferability and replicability. Here, we developed compare_genomes, a transferable and extensible comparative genomics workflow package which streamlines the identification of orthologous families within and across genomes and tests for the presence of several mechanisms of evolution (gene family expansion or contraction and substitution rates within protein-coding sequences).

Availability and Implementation

The workflow is available for Linux, written as a Nextflow workflow which calls established genomics and phylogenetics tools to streamline the analysis and visualisation of genome divergence. This workflow is freely available at https://github.com/jeffersonfparil/compare_genomes, distributed under the GNU General Public License version 3 (GPLv3).

Contact

Corresponding author: Jeff Paril jeff.paril@unimelb.edu.au. Queries and issues regarding the implementation can be submitted on the issue page of the github repository: https://github.com/jeffersonfparil/compare_genomes/issues.

Supplementary information

Synonymous to non-synonymous (Ka/Ks) nucleotide substitution ratio plots for the example data set are found in the github repository.

Introduction

The genomes of related organisms represent records of evolutionary histories along the tree of life. We can infer the drivers of their evolution by analysing the signatures of genome evolution during polyploidisation events, gene duplication or loss, or selection of adaptive mutations between species. This comparative genomics framework scrutinises alternative evolutionary histories across different species within and between clades which will offer insights into how the spatio-temporal dynamics and interactions between the biosphere and the environment prefer one group of biological solutions to fitness over others.

The increasing availability of affordable high-throughput DNA sequencing has allowed the assembly of the genomes of multiple species beyond the early set of model species towards species of specific biological relevance. This enabled the use of deep comparative genomics to answer fundamental evolutionary questions using multiple species within and across clades. However, the pipelines for these analyses are often study-specific and rarely described with enough details to be fully reproducible. This is a major impediment to the generalisation or meta-analysis of the results and hinders the transfer of these workflows. Hence, the bioinformatics community would benefit from a unified but open-source and portable comparative genomics workflow.

Web-based comparative genomics workflows exist (eg. PLAZA; Van Bel et al, 2022). However, the centralised design limits potential extension and usage is inherently limited to the computational resources provided, making it unsuitable to high throughput or high bandwidth usage. Various portable frameworks have been developed to run comparative genomics pipelines in a reproducible way (e.g. Snakemake by Mölder et al, 2021 and Nextflow by Di Tommaso et al, 2017). These tools work synergistically with package and environment management systems such as Docker (https://www.docker.com/) or Conda (https://docs.conda.io/en/latest/). These were used to generate genome assemblies and annotations, sequence alignments, variant callings, and transcriptomic data analyses. However, there is a noticeable lack of fully transparent and easily transferable comparative genomics analysis workflows. In this application note, we address this gap with compare_genomes, a comparative genomics workflow built under the Nextflow framework with packages managed by Conda.

Features and implementation

The compare_genomes workflow consists of nine analysis steps under the default setup (Figure 1: left panel). These steps are:

1. Download of the user-defined genome datasets (i.e. genome sequences, annotations, coding DNA sequences or CDS, protein sequences, protein-coding gene models, corresponding gene ontology terms, and protein sequences of specific genes of interest).
2. Identification of orthogroups using OrthoFinder (Emms and Kelly 2019) and within genome gene family for each orthogroup using HMMER3 (Mistry et al, 2013) and Panther HMMs (protein-coding gene family models; Thomas et al, 2022).
3. Inference of phylogenomic trees for each orthogroup with IQ-TREE 2 (Minh et al, 2020) using CDS alignments generated by MACSE (Ranwez et al, 2011) and the most likely nucleotide substitution model inferred by ModelFinder (Kalyaanamoorthy et al, 2017).
4. Inference of the rate of molecular evolution based on transversion rates among four-fold degenerate sites (4DTv) in single-copy genes between each pair of species using custom Julia scripts.
5. Assessment of whole-genome duplication events using 4DTv computed from multi-copy gene families.
6. Test of significant gene family expansion or contraction across genomes using CAFE (version 5; De Bie et al, 2006).
7. Gene ontology (GO) term enrichment analysis for significantly expanded gene families using the Panther GO API (Mi et al, 2019).
8. Visualisation of a summary of the results (i.e. Figure 1: right panel for a sample output).
9. Analysis of user-defined gene(s) of interest, i.e. gene family expansion/contraction analyses with CAFE, and estimation of non-synonymous to synonymous nucleotide substitution rates (Ka/Ks) using KaKs_Calculator 2.0 (Wang et al, 2010) and custom R script.

Compare_genomes was implemented using Nextflow because of the ease of integrating new and existing Linux-based bioinformatics pipelines. Analysis steps can be easily added or modified, for example adding a GO term enrichment analysis for the significantly contracted families, or substitute MACSE for another multiple sequence alignment tool.

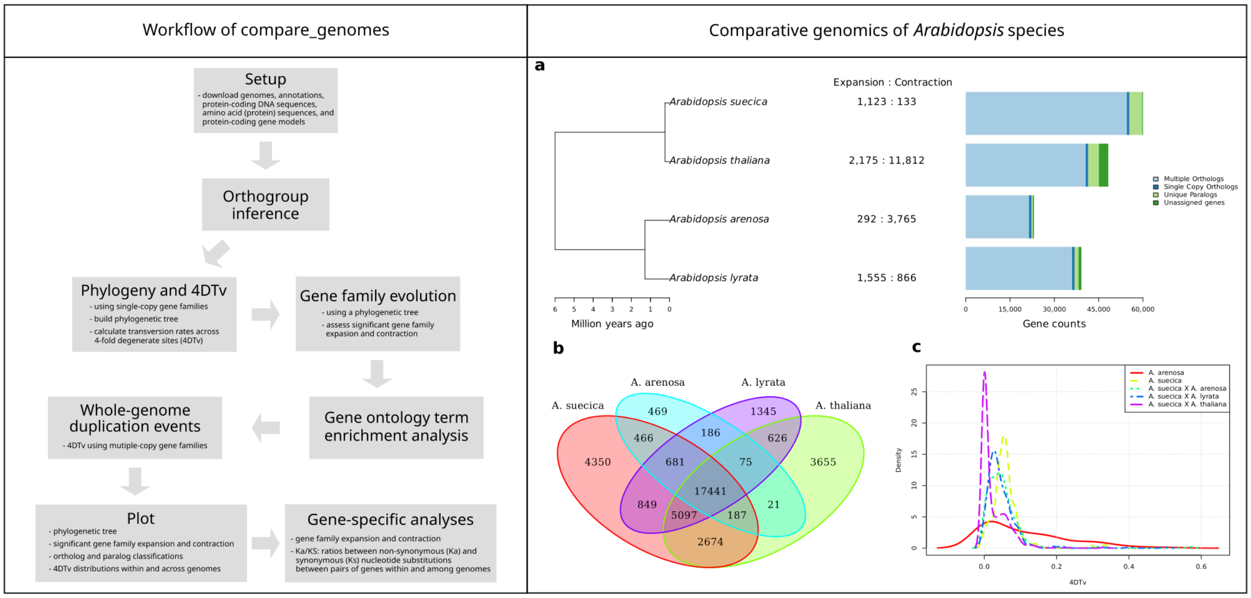

Workflow of compare_genomes
Setup
- download genomes, annotations, protein-coding DNA sequences, amino acid (protein) sequences, and protein-coding gene models
Orthogroup inference
Phylogeny and 4DTv
- using single-copy gene families
- build phylogenetic tree
- calculate transversion rates across 4-fold degenerate sites (4DTv)
Gene family evolution
- using a phylogenetic tree
- assess significant gene family expasion and contraction
Gene ontology term enrichment analysis
Whole-genome duplication events
- 4DTv using mutiple-copy gene families
Plot
- phylogenetic tree
- significant gene family expansion and contraction
- ortholog and paralog classifications
- 4DTv distributions within and across genomes
Gene-specific analyses
- gene family expansion and contraction
- Ka/KS: ratios between non-synonymous (Ka) and synonymous (Ks) nucleotide substitutions between pairs of genes within and among genomes
Comparative genomics of *Arabidopsis* species
a
Expansion : Contraction
Arabidopsis suecica
1,123 : 133
Arabidopsis thaliana
2,175 : 11,812
Arabidopsis arenosa
292 : 3,765
Arabidopsis lyrata
1,555 : 866
6 5 4 3 2 1 0
Million years ago
0 15,000 30,000 45,000 60,000
Gene counts
Multiple Orthologs
Single Copy Orthologs
Unique Paralogs
Unassigned genes
b
A. suecica
A. arenosa
A. lyrata
A. thaliana
469
1345
466
186
626
4350
681
75
3655
17441
849
5097
187
21
2674
c
A. arenosa
A. suecica
A. suecica X A. arenosa
A. suecica X A. lyrata
A. suecica X A. thaliana
Density
0 5 10 15 20 25
0.0 0.2 0.4 0.6
4DTv

Figure 1: **Left panel**: The steps performed by the compare_genomes comparative genomics workflow. **Right panel**: A sample output plot generated by the compare_genomes workflow using four *Arabidopsis* species.

Usage and example

A detailed user manual describing how to install, set up, and run the workflow is presented in the README page of the compare_genomes project repository: https://github.com/jeffersonfparil/compare_genomes. We included a tutorial and dataset analysing four *Pseudomonas spp.* species. The compare_genomes workflow was initially conceived to compare the newly released, high-quality reference genome of annual ryegrass (*Lolium rigidum*; an important weed in winter cropping) to related grass species, including the pasture crop perennial ryegrass (*Lolium perenne*). This analysis showed significant expansion of herbicide resistance-related gene families, including detoxification genes in the noxious weed annual ryegrass (Paril et al, 2022).

To illustrate the performance of our pipeline, we compared the genomes of four well characterised *Arabidopsis* species to test if it is able to recapture the evolutionary patterns expected in these species. The summary of the output of this example is shown in Figure 1 right panel. It shows that the phylogenetic relationship between species was accurately recapitulated as expected from the results of Novikova and colleagues in 2016. This summary figure also shows the differences in patterns of gene family contraction and expansion. *Arabidopsis suecica*, an allopolyploid hybrid of *A. thaliana* and *A. arenosa* (Novikova et al, 2017; Burns et al, 2021), experienced gene family expansion. A similar expansion was observed in *A. lyrata*, an outbreeder which diverged from *A. thaliana* around 5 million years ago (Schmickl et al, 2010). Similarities in gene family composition between species are presented as a Venn diagram. Finally, the presence of whole-genome duplication events was assessed using 4DTv density plot. This analysis recaptured the recent polyploidisation event of the *A. arenosa* sub-genome within the *A. suecica* allopolyploid genome detected by Novikova and colleagues in 2017. The Ka/Ks ratio analyses presented in the supplementary information show evidence of selection across several 15-bp windows in the GSTU13 (glutathione transferase - tau class; size of the window can be modified).

Conclusion

We developed compare_genomes, a transferable and extendible comparative genomics workflow built using the Nextflow framework and Conda package management system. It provides a wieldy pipeline to test for non-random evolutionary patterns which can be mapped to evolutionary processes to help identify the molecular basis of specific features or remarkable biological properties of the species analysed. Additionally, it provides a template which other comparative genomics pipelines can be built or patterned upon for improved reproducibility.

Acknowledgements

We thank the Department of Agriculture Fisheries and Forestry of Australia together with the Grains Research and Development Corporation (Grant ID: 4-FY9JQPE), Australian Research Data Commons (Grant ID: DP727), Commonwealth Scientific and Industrial Research Organisation for the funding, and the University of Melbourne for hosting this study.

References

Burns, Robin, Terezie Mandáková, Joanna Gunis, Luz Mayela Soto-Jiménez, Chang Liu, Martin A. Lysak, Polina Yu Novikova, and Magnus Nordborg. “Gradual Evolution of Allopolyploidy in Arabidopsis Suecica.” Nature Ecology & Evolution 5, no. 10 (October 2021): 1367–81. https://doi.org/10.1038/s41559-021-01525-w.

“Conda — Conda Documentation.” Accessed January 5, 2023. https://docs.conda.io/en/latest/.

De Bie, Tijl, Nello Cristianini, Jeffery P. Demuth, and Matthew W. Hahn. “CAFE: A Computational Tool for the Study of Gene Family Evolution.” Bioinformatics 22, no. 10 (May 15, 2006): 1269–71. https://doi.org/10.1093/bioinformatics/btl097.

Di Tommaso, Paolo, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. “Nextflow Enables Reproducible Computational Workflows.” Nature Biotechnology 35, no. 4 (April 2017): 316–19. https://doi.org/10.1038/nbt.3820.

“Docker: Accelerated, Containerized Application Development.” Accessed January 5, 2023. https://www.docker.com/.

Emms, David M., and Steven Kelly. “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics.” Genome Biology 20, no. 1 (November 14, 2019): 238. https://doi.org/10.1186/s13059-019-1832-y.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. “ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates.” Nature Methods 14, no. 6 (June 2017): 587–89. https://doi.org/10.1038/nmeth.4285.

Mi, Huaiyu, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. "PANTHER Version 14: More Genomes, a New PANTHER GO-Slim and Improvements in Enrichment Analysis Tools." Nucleic Acids Research 47, no. D1 (January 8, 2019): D419–26. https://doi.org/10.1093/nar/gky1038.

Minh, Bui Quang, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era." Molecular Biology and Evolution 37, no. 5 (May 1, 2020): 1530–34. https://doi.org/10.1093/molbev/msaa015.

Mistry, Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. "Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions." Nucleic Acids Research 41, no. 12 (July 2013): e121. https://doi.org/10.1093/nar/gkt263.

Mölder, Felix, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa Sochat, Jan Forster, et al. "Sustainable Data Analysis with Snakemake." F1000Research, January 18, 2021. https://doi.org/10.12688/f1000research.29032.1.

Novikova, Polina Yu, Nora Hohmann, Viktoria Nizhynska, Takashi Tsuchimatsu, Jamshaid Ali, Graham Muir, Alessia Guggisberg, et al. "Sequencing of the Genus Arabidopsis Identifies a Complex History of Nonbifurcating Speciation and Abundant Trans-Specific Polymorphism." Nature Genetics 48, no. 9 (September 2016): 1077–82. https://doi.org/10.1038/ng.3617.

Novikova, Polina Yu., Takashi Tsuchimatsu, Samson Simon, Viktoria Nizhynska, Viktor Voronin, Robin Burns, Olga M. Fedorenko, et al. "Genome Sequencing Reveals the Origin of the Allotetraploid Arabidopsis Suecica." Molecular Biology and Evolution 34, no. 4 (April 1, 2017): 957–68. https://doi.org/10.1093/molbev/msw299.

Paril, Jefferson, Gunjan Pandey, Emma M. Barnett, Rahul V. Rane, Leon Court, Thomas Walsh, and Alexandre Fournier-Level. "Rounding up the Annual Ryegrass Genome: High-Quality Reference Genome of Lolium Rigidum." Frontiers in Genetics 13 (2022). https://www.frontiersin.org/articles/10.3389/fgene.2022.1012694.

Ranwez, Vincent, Sébastien Harispe, Frédéric Delsuc, and Emmanuel J. P. Douzery. "MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons." PLOS ONE 6, no. 9 (September 16, 2011): e22594. https://doi.org/10.1371/journal.pone.0022594.

Schmickl, Roswitha, Marte H. Jørgensen, Anne K. Brysting, and Marcus A. Koch. "The Evolutionary History of the Arabidopsis Lyrata Complex: A Hybrid in the Amphi-Beringian Area Closes a Large Distribution Gap and Builds up a Genetic Barrier." BMC Evolutionary Biology 10, no. 1 (April 8, 2010): 98. https://doi.org/10.1186/1471-2148-10-98.

Thomas, Paul D., Dustin Ebert, Anushya Muruganujan, Tremayne Mushayahama, Laurent-Philippe Albou, and Huaiyu Mi. "PANTHER: Making Genome-Scale Phylogenetics Accessible to All." Protein Science 31, no. 1 (2022): 8–22. https://doi.org/10.1002/pro.4218.

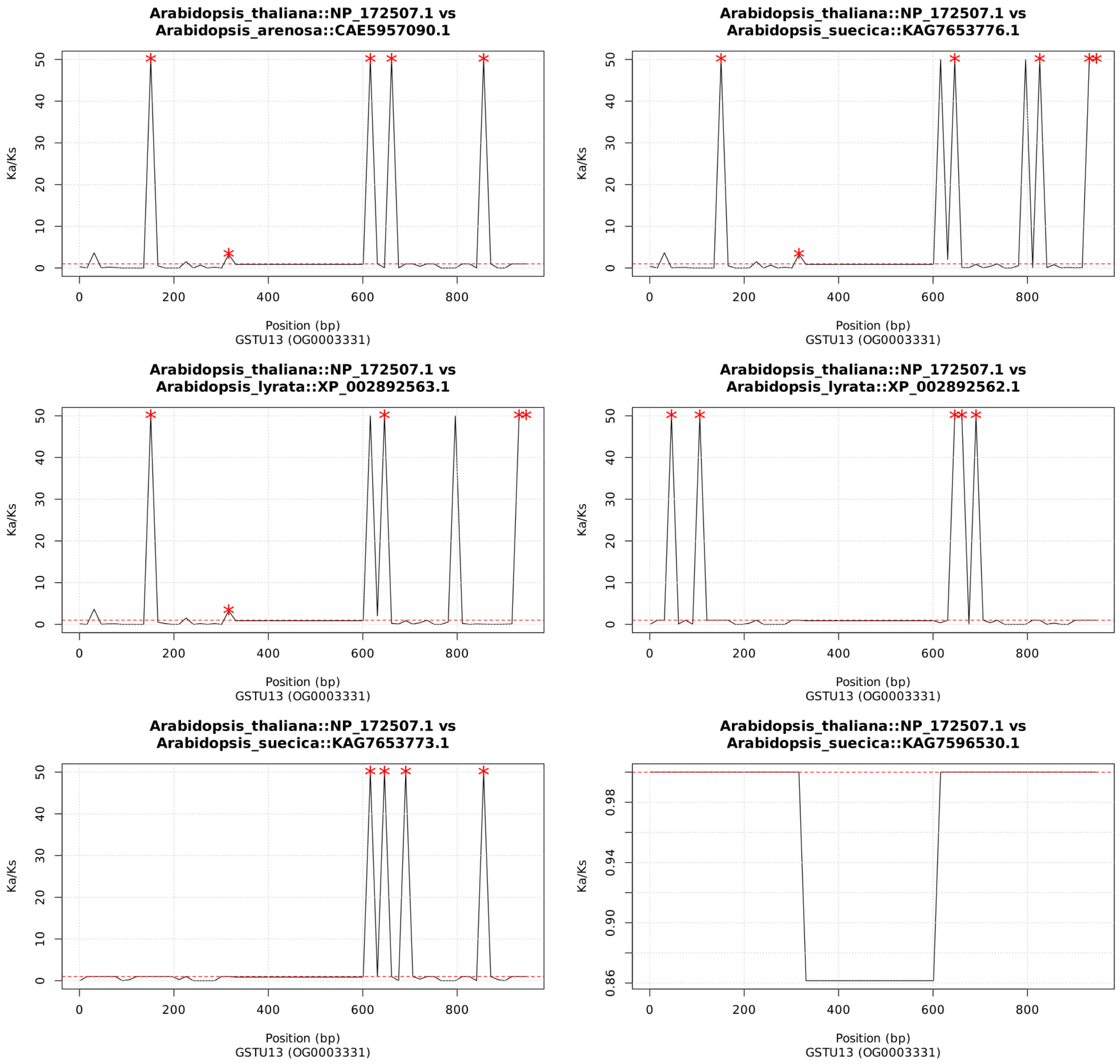
Van Bel, Michiel, Francesca Silvestri, Eric M Weitz, Lukasz Kreft, Alexander Botzki, Frederik Coppens, and Klaas Vandepoele. "PLAZA 5.0: Extending the Scope and Power of

Comparative and Functional Genomics in Plants." Nucleic Acids Research 50, no. D1 (January 7, 2022): D1468–74. https://doi.org/10.1093/nar/gkab1024.

Wang, Da-Peng, Hao-Lei Wan, Song Zhang, and Jun Yu. "γ-MYN: A New Algorithm for Estimating Ka and Ks with Consideration of Variable Substitution Rates." Biology Direct 4, no. 1 (June 16, 2009): 20. https://doi.org/10.1186/1745-6150-4-20.

Wang, Dapeng, Yubin Zhang, Zhang Zhang, Jiang Zhu, and Jun Yu. "KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies." Genomics, Proteomics & Bioinformatics 8, no. 1 (March 1, 2010): 77–80. https://doi.org/10.1016/S1672-0229(10)60008-3.

Supplementary Information



Arabidopsis_thaliana::NP_172507.1 vs Arabidopsis_arenosa::CAE5957090.1

Arabidopsis_thaliana::NP_172507.1 vs Arabidopsis_suecica::KAG7653776.1

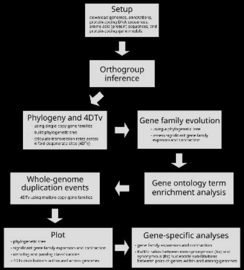
Arabidopsis_thaliana::NP_172507.1 vs Arabidopsis_lyrata::XP_002892563.1

Arabidopsis_thaliana::NP_172507.1 vs Arabidopsis_lyrata::XP_002892562.1

Arabidopsis_thaliana::NP_172507.1 vs Arabidopsis_suecica::KAG7653773.1

Arabidopsis_thaliana::NP_172507.1 vs Arabidopsis_suecica::KAG7596530.1

Supplementary Figure 1: Ka/Ks (ratio of the number of nonsynonymous substitutions per non-synonymous site to the number of synonymous substitutions per synonymous site per unit time) across non-overlapping 15-bp sliding windows in GSTU13 (tau class glutathione transferase) gene between four *Arabidopsis* species. Red asterisks refer to windows with significant deviations from neutral expectations at α=0.1%.



Setup

download genomes, annotations, protein-coding DNA sequences, amino acid (protein) sequences, and protein-coding gene models

Orthogroup inference

Phylogeny and 4DTv

- using single copy gene families
- build phylogenetic tree
- calculate transversion rates across 4-fold degenerate sites (4DTv)

Gene family evolution

- using a phylogenetic tree
- assess significant gene family expansion and contraction

Gene ontology term enrichment analysis

Whole-genome duplication events

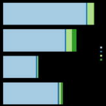
- 4DTv using multiple copy gene families

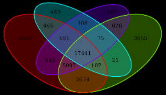
Plot

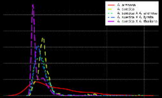
- phylogenetic tree
- significant gene family expansion and contraction
- ortholog and paralog classifications
- 4DTv distributions within and across genomes

Gene-specific analyses

- gene family expansion and contraction
- Ka/Ks ratios between non-synonymous (Ka) and synonymous (Ks) nucleotide substitutions between pairs of genes within and among genomes




469
466
186
626
681
75
3650
17441
869
5097
187
21
2674


A. arenosa
A. suecica
A. suecica X A. arenosa
A. suecica X A. lyrata
A. suecica X A. thaliana