EASYstrata: a new workflow to infer evolutionary strata along sex chromosomes and other supergenes

Authors: Quentin Rougemont[1], Elise Lucotte[1], Loreleï Boyer[1], Alexandra Jalaber[1], Alodie Snirc[1], Tatiana Giraud[1], Ricardo C. Rodríguez de la Vega [1]

[1] Ecologie Systématique et Evolution, CNRS, Universite Paris-Saclay, AgroParisTech, 91198 Gif-sur-Yvette, France

Abstract

New reference genomes and transcriptomes are becoming increasingly available across the whole tree of life, opening new avenues to tackle exciting evolutionary questions. However, there are still challenges associated with annotating genomes, aligning sequences, inferring evolutionary processes and a lack of methodological standardisation. Here, we propose a new workflow designed for evolutionary analyses to overcome these challenges, facilitating the detection of recombination suppression and its consequences in terms of rearrangements and transposable element accumulation. To do so, we assemble multiple bioinformatic steps in a single, reproducible and easy-to-use workflow. We combine state-of-the-art tools to quickly detect transposable elements, annotate newly assembled genomes, infer gene orthology relationships, compute divergence between sequences, infer evolutionary strata (i.e. stepwise extension of recombination suppression) and their structural rearrangements, and visualise the results at different steps in the process. This workflow was called EASYstrata and was applied to reannotate a set of 42 published genomes from *Microbotryum* fungi. We show in case examples that we recover the same strata as previously described. While this tool was mostly developed with the goal to infer divergence between sex or mating-type chromosomes, it can be applied to any pair of haplotypes whose pattern of divergence might be of interest, e.g. supergenes in autosomes. This workflow will facilitate the study of the numerous non-model species for which newly sequenced phased diploid genomes are becoming available.

Contact: quentinrougemont@orange.fr

Introduction

The past ten years have seen an increasing number of available high-quality genomes across a wide range of organisms, including many non-model species (Formenti *et al.*, 2022). This opens up the possibility to better understand the evolution of genomes across the tree of life. For instance, it is now possible to study the patterns of recombination suppression along sex chromosomes (Hartmann *et al.*, 2021a; Yue *et al.*, 2023) and to empirically test the theoretical hypotheses developed to explain these patterns (Jay *et al.*, 2024). In particular, recombination suppression often progressively extends along sex chromosomes (Charlesworth *et al.*, 2014), and other supergenes (Charlesworth, 2016). Such stepwise extension of recombination suppression can be inferred in genomes by detecting patterns of evolutionary strata, *i.e.* discontinuous stair-like patterns of synonymous divergence between sex chromosomes or mating-type chromosomes (Lahn and Page, 1999; Bergero and Charlesworth, 2009; Branco *et al.*, 2017), Fig 1). Indeed, synonymous substitutions regularly accumulate with time between alleles on alternative sex

chromosomes from the onset of recombination suppression, making synonymous divergence a proxy for the time since recombination suppression. For detecting such stepwise extension of recombination suppression, synonymous divergence needs to be plotted against the ancestral gene order, because rearrangements following recombination suppression reshuffle gene order, therefore erasing footprints of evolutionary strata. The proxy for ancestral gene order is typically the sex chromosome from the homogametic sex (X or Z) in XY and ZW systems. In UV-like systems, in which neither of the two sex or mating-type chromosomes recombine, both accumulating rearrangements, an outgroup with recombining sex-related chromosomes can be used (Branco *et al.*, 2017; Yue *et al.*, 2023; Hough *et al.*, 2014; Umen and Coelho, 2019). The evolutionary and proximal hypotheses for explaining such stepwise extension of recombination suppression can be tested by investigating the degrees of rearrangements between sex chromosomes and the distribution of deleterious mutations (Jay *et al.*, 2024).

However, such a large amount of genomic data generates multiple challenges for analyses. While standardised procedures are now becoming available for genome assemblies (The Darwin Tree of Life Project Consortium, 2022), this is not yet the case for gene prediction. Indeed, analysing genomes annotated with different gene calling methods may introduce biases in downstream inference (Weisman *et al.*, 2022), for example if a method systematically predicts more genes than another method. In addition, the computation of synonymous divergence may be difficult given that the dedicated tools possess a steep learning curve (Álvarez-Carretero *et al.*, 2023). Identifying and objectively delimiting evolutionary strata is often not straightforward either, given the stochastic noise in synonymous divergence values. For all these reasons, inference of evolutionary strata in sex and mating-type chromosomes can be a considerable challenge, while being crucial to our understanding of the evolution of sex chromosomes and other supergenes.

Easy-to-use, reproducible and efficient workflows are still lacking in this research area, precluding routine analyses of evolutionary strata and limiting the reanalysis of existing data. There is an increasing need for such reproducible and standardised workflow to enable comparison across multiple organisms. Here, we provide a reproducible workflow for i) transposable element (TE) and gene calling, ii) gene filtration, quality assessment and subsequent inference of orthology relationships, iii) synteny pattern inferences based on both whole-genome and gene-based comparisons, and iv) synonymous divergence visualisation and evolutionary stratum inferences, as well as assessment of their association with chromosomal rearrangements. We called the workflow "EASYstrata" for Evolutionary analysis with Ancestral SYnteny for strata identification. EASYstrata is fully available at https://github.com/QuentinRougemont/EASYstrata, and can be deployed on any cluster. Below we describe its different modules, including options depending on the availability of ancestral gene order and of gene annotations. The pipeline has been used previously for studying the sex chromosome of *Silene* plant species (Moraga *et al.*, 2024), and we describe here its features and its application to a set of 42 previously published genomes in *Microbotryum* anther-smut fungi, focusing on two study cases among them. In this group of fungi, there were ancestrally two mating-type chromosomes, bearing each a mating-type locus (HD and PR loci, for homeodomain and pheromone receptor genes), and with two alleles at the PR mating-type locus (called a_1 and a_2). The species *M. lagerheimii* has been shown to be a good proxy with the ancestral gene arrangement of these two mating-type chromosomes (Branco *et al.*, 2017). We generated here a new reference genome for this species (Supp Methods and Results). Across the *Microbotryum* genus, there have been multiple events of mating-type chromosome rearrangements bringing PR and HD loci on the

same chromosome, and recombination suppression linking these two loci. Recombination suppression has then extended away from the mating-type loci in successive steps independently in several species (Branco *et al.*, 2018; Duhamel *et al.*, 2023). By using EASYstrata, we show in case examples that we recover the same strata as previously described. With this tool, we hope to stimulate research on the inference of evolutionary strata along sex chromosomes, mating-type chromosomes and other supergenes across a broad spectrum of organisms.

Methods

Input data and file configuration:

Three data sets are typically needed in the workflow (**Fig 1**): first, high-quality genome assemblies in fasta format. These can be in the form of two alternative mating-type chromosomes, a full diploid genome including a pair of sex chromosomes, such as X/Y or Z/W chromosomes, or any autosome bearing supergenes with two differentiated haplotypes. The two differentiated sex chromosomes or haplotypes are hereafter called haplotype1 and haplotype2. High-quality genome assemblies can be obtained from long-read sequencing technologies (e.g. Oxford Nanopore Technologies (ONT) or Pacbio HiFi). Second, the workflow needs a haplotype designated as a proxy for the ancestral gene order of the region of interest, in fasta format. This can be the X chromosome in XY systems (Lahn and Page, 1999) or the genome of a closely related species with recombining mating-type chromosomes when neither of the two haplotypes recombine and therefore both accumulate rearrangements, thus diverging from the ancestral gene order (Branco *et al.*, 2017). In the absence of a chromosome constituting a good proxy for the ancestral gene order, dedicated methods for ancestral gene order reconstruction may be used (e.g: Agora (Muffato *et al.*, 2023)). Even if no proxy for an ancestral state is available, at least one closely related species (outgroup) is recommended to help identify single-copy orthologs between haplotypes. Third, a set of protein sequences from as many closely related species as possible (in fasta format) can be provided, which will improve gene prediction in the genome annotation step with Braker (Gabriel *et al.*, 2024). If no such data are provided the pipeline will use the pre-partitionned OrthoDB clades provided at https://bioinf.uni-greifswald.de/bioinf/partitioned_odb12/ (Tegenfeldt *et al.*, 2024). In this case, a taxon name corresponding to the partition the species belong to should be provided (e.g. vertebrata, fungi, etc, detailed in the github). Optionally, RNAseq data (paired-end or single-end) can be used as input for the genome annotation step, in the form of a .txt file containing the list of RNAseq reads.

The workflow takes as input a config file including the path to the files described above as well as additional information and external data. An example of a config file is available in the github repository and summarised in **Table 1** below. Short names for the two haplotypes will be used as a basename to rename the fasta file of each haplotype genome, and the contig names within genomes, which will in turn serve as a basename ID in the genes name. Because the workflow relies on masking transposable elements prior to gene annotation using NCBI databases, a NCBI species name should be provided (NCBI species names can be accessed at https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cg). The evaluation of gene prediction completeness crucially relies on BUSCO (benchmarking universal single-copy orthologs (Simão *et al.*, 2015)), *i.e.,* evolutionarily-informed expectations of gene content for near-universal single-copy orthologs. A BUSCO lineage name is therefore also compulsory in the config file. In addition, the path to a .txt file containing the name of the

contigs or chromosomes of interest in the haplotype1 (or ancestral chromosome), corresponding to the target sex chromosome, mating-type chromosome, supergene or any focal region of interest should be provided in the config file. Below, we detail each step that will be performed to annotate a given pair of genomes and compute per-gene synonymous divergence (summarised in **Table 1** and **Fig 1**).

Table 1 : Summary of the main compulsory and optional parameters to provide in the config file. If a single genome is provided, only the annotation step is performed. The orthoDBspecies value should be chosen in the following list : “Metazoa" "Vertebrata" "Viridiplantae" "Arthropoda" "Eukaryota" "Fungi" "Alveolata" and will be used for annotation if no additional data are available. With two genomes, the whole pipeline can be implemented. Alternatively, gtf/gff files may already be available, in which case only the latter stages of the workflow are executed. If no RNAseq data is available or if a bam file of aligned data is already available, the workflow will also be run accordingly. If a genome is available for the ancestral state, gene order is computed based on the order in this genome. Without a proxy of the ancestral genome, the order is based on the order in haplotype1. More details available on the project github page.

parameter	description
compulsory parameters	
genome1	full path to haplotype1 genome assembly (fasta, compressed or not)
haplotype1	a short name to be used in basename of chromosome contigs (eg: haplo1)
optional parameters	
genome2	full path to haplotype2 genome assembly (fasta, compressed or not)
haplotype2	a short name to be used in basename of chromosome contigs (eg: haplo2)
ancestral_genome	full path to ancestral genome (fasta, compressed or not) or to an outgroup
ancestral_gff	full path to ancestral gff or to an outgroup
Annotation (optional) parameters	
annotate	A string (YES/NO); if NO no annotation will be performed
fungus	A string (YES/NO); if NO the fungus option is not provided to Braker
orthoDBspecies	A string stating the lineage to be used for orthoDB protein download (see legend below),
RelatedProt	Full path to a set of external protein data (fasta format)
gtf1	name of gtf1; only used if annotate = NO
gtf2	name of gtf2; only used if annotate = NO
rnaseq	A string (YES/NO); if YES rnaseq data will be used
bam	An optional path to a bam file if rnaseq data are already aligned to the reference genome
RNAseqlist	Full path to txt file listing rnaseq data (PE or SE)
Transposable element TE (optional) parameters	
annotateTE	A string YES/NO if YES = annotate TE, if NO skip (assume TE already masked)
TE database	Full path to a custom TE database (if available)
ncbi_species	species name to be used for ncbi based de-novo TE discovery
BUSCO compulsory parameters	
busco_lineage	BUSCO species name for your target species (can be obtained using Busco –list-datasets)
chromosome/scaffold compulsory info	
scaffold	full path to a table listing the target genome to used as ref, the region of interest, and whether chromosomes should be reversed (“R”) or not (“N”)

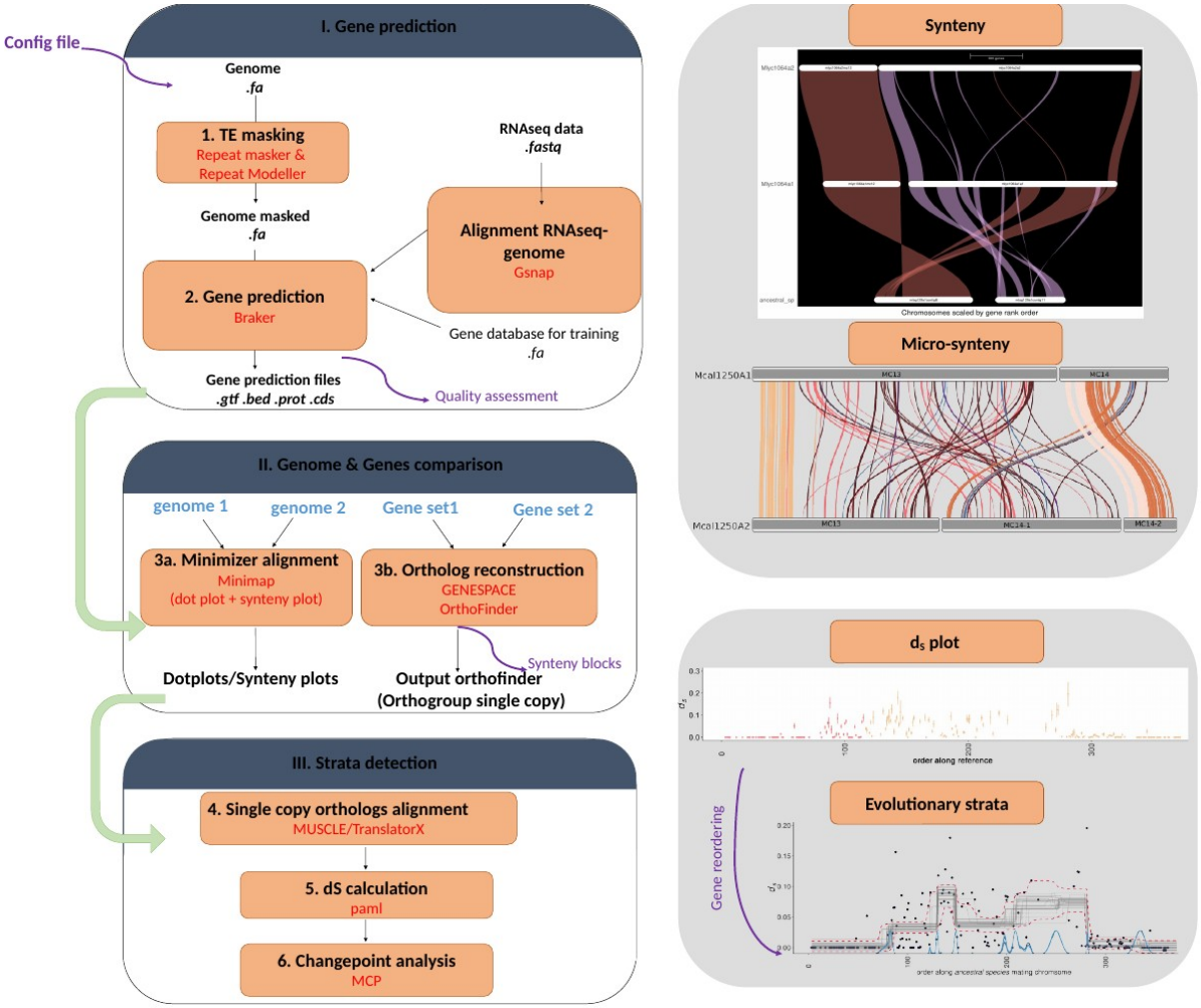

Figure 1 : Overview of the workflow used for haplotype alignment and visualisation, transposable element (TE) detection, gene prediction, synteny analysis, per-gene synonymous divergence (d_S) and changepoint analysis**.** A table of all third party tools is provided in **Table S1**. Details of how to install it are provided in the github INSTALL.md file.

Transposable element detection

De-novo detection of transposable elements (TEs) is performed for each haplotype separately using RepeatModeler 2.0.2 (Flynn *et al.*, 2020). Next, RepeatMasker (Smit *et al.*, 2013) is used to improve TE detection and masking using TE databases. It is then possible to use one or several of the following: *(i)* the *de-novo* families of TE detected by RepeatModeler, and optionally *(ii)* in-house TE libraries, *(iii)* RepBase database of TEs. The annotations labelled as “unknown” by RepeatModeler (i.e. without any known annotation) can be removed from the TE set to limit the rates of false positive TE identification that would correspond to genuine genes, or left as is, which may potentially induce an increased rate of false negative genes. Finally, if a soft-masked genome is already available, and no TE detection is needed, the option annotateTE = “NO” can be set directly in the config file, and this step will be skipped.

Genome annotation and filtering

The resulting soft masked genome is then fed to BRAKER 3 (Gabriel *et al.*, 2024), a method that combines GeneMark ETP (Brůna *et al.*, 2024) and AUGUSTUS (Stanke *et al.*, 2006).

While our pipeline has been optimised with BRAKER v2.1.6, we obtained similar results with BRAKER 3 and chose to focus on the latest because *i*) it is the currently maintained version, and *ii*) it does not require the licence associated with GeneMark. We propose two use cases for Braker: i) with a set of protein sequences from closely related species only, or ii) with a set of protein sequences from closely related species and RNAseq data for the focal species.

In case *i*, BRAKER is run for five successive (independent) rounds, in order to take into account (limited) stochasticity, and the round providing the highest BUSCO score is kept for downstream processing. In use case *ii*, the same steps as case i are applied to the protein sequences (BRAKER and selection of the best round). Additionally, raw RNAseq reads (paired end or single end) are trimmed with Trimmomatic (Bolger *et al.*, 2014) and mapped to the reference genome using Gsnap (Wu *et al.*, 2016), with the resulting alignment being further sorted and filtered with Samtools (Li *et al.*, 2009). A table of read count and mapping quality is produced at this step, together with plots of read depth along each chromosome. Then, BRAKER runs GeneMark-ETP and Augustus on mapped RNAseq data. The best protein sequence round, as evaluated with BUSCO (see below), and the RNAseq run are combined using TSEBRA (Gabriel *et al.*, 2021). While BRAKER 3 enables the use of both RNAseq and external data along with TSEBRA in a single command call, we found that calling BRAKER with each data type separately and combining with TSBERA afterwards provided higher BUSCO scores (97% vs >98% on average); we therefore chose to stick with this approach. Moreover this approach enables the users to more easily customize the TSEBRA config file. At this step, a report is automatically generated using Braker utilities, enabling the assessment of the number of complete and partial genes, the number of single and multi exon genes, the number of introns per gene and the number of genes fully or partially supported by external evidence (i.e. RNAseq and databases). Various histograms are also automatically produced, which may be exploited to further filter genes with aberrant profiles.

The resulting gtf file, which contains both Augustus- and GeneMark-based gene predictions is then modified with a custom perl script (obtained here: https://github.com/Gaius-Augustus/BRAKER/issues/457). The file is parsed in order to insert the haplotype name in each gene ID. This step facilitates the analyses of genes in a phylogenetic context, especially when multiple species are to be studied.

This gtf is then used to extract the CDS and protein sequences from the genome of each haplotype using gffread (Pertea and Pertea, 2020). Finally, data are filtered to keep a single transcript per gene, the longest one, of prime importance to accurately infer orthology. This constitutes the final gff, CDS and protein sequence files for which BUSCO scores are eventually computed. The quality of the gene prediction is further assessed automatically through blasts against the Reviewed Swiss-Prot database available from uniprot. This enables the user to assess the proportion of predicted genes with at least one match against uniprot. Optionally, InterProScan will be run, combining several databases, such as pfam, panther, NCBIfam, cath-gene3D, in order to obtain prediction of the gene functions, family classification and domain predictions. The resulting hits, which are of interest in themselves for gene function inference, are also used to compute the proportion of genes with a hit against a database among all genes.

Identifying rearrangements

Minimizer-based whole genome alignment

Alignments of the two haplotype sequences against each other and against the ancestral genome proxy are performed with Minimap2 (Li, 2021) due to its rapidity and ease of use. Dotplots of primary alignments are automatically generated using PafR (Winter, 2020) and, if a list of sex chromosome regions or supergenes of interest are provided in the config file, synteny plots are automatically drawn, to facilitate visual inspection of rearrangements.

Synteny analysis based on orthologous gene sets

Synteny is investigated based on the gene positions in the different haplotypes of orthologous genes or alleles. To that aim, the pipeline first infers gene orthology relationships using OrthoFinder (Emms & Kelly, 2019, 2015) from the annotated gene set obtained from Braker. Instead of only extracting orthology information between single-copy genes, we took advantage of the recently developed GeneSpace pipeline (Lovell et al. 2022), which combines Diamond v2.0.8 (Buchfink & al. 2021), OrthoFinder v2.5.4 and MCScanX (Wang et al., 2012), to integrate gene order and orthology relationships and detect gene collinearity (i.e. local synteny) between genomes. GeneSpace is run with default parameters and automatically produces dotplots among all pairs of genomes, as well as riparian plots based on gene order and gene position. If a set of target scaffolds (e.g., sex chromosomes or supergene haplotypes) are provided, specific subplots are also constructed. The results can easily be further processed *a posteriori* by the user, for instance by exploiting pangenome gene sets and customising riparian and dotplots (e.g. Moraga et al. 2024).

The resulting single-copy orthologs extracted from OrthoFinder output files are further reformatted to serve as input to the Rideogram package (Hao *et al.*, 2020), enabling the visualisation of gene order from the target pairs of scaffolds (sex chromosomes or supergene haplotypes). A Circos plot is also produced using circlize (Gu et al. 2014) from the resulting single-copy orthologs (*i.e.*, one-to-one gametologs) highlighting the regions of interest. Optionally, bed files of the genes and transposable elements can be provided and will be plotted as inner tracks.

Calculating synonymous divergence between haplotypes

The set of single-copy orthologs identified by OrthoFinder is used as input on which the per-gene synonymous divergence (d_S) and non-synonymous divergence (d_N) are computed between chromosomes, as well as their standard errors (SE), using the yn00 program implemented in PAML (Yang, 2007). To that aim, the CDS (coding sequences, fasta format) for each single-copy orthologous pair are aligned using MUSCLE (Edgar, 2004), implemented in TranslatorX (Abascal *et al.*, 2010). Aligning with Macse (Ranwez *et al.*, 2011) yielded very similar results. All the processes involved are encompassed in a simple bash script. The per-gene synonymous divergence (d_S) values are plotted against the ancestral gene order with the goal to detect evolutionary strata. In cases where no proxy for the ancestral genome is available, the gene-rank order position along the haplotype 1 contigs is used as a proxy to plot the per-gene synonymous divergence (d_S) values.

Visualising synonymous divergence (d_S) along genomes and inferring evolutionary strata

The resulting per-gene d_S values are plotted along the contigs of the genome representing the ancestral gene arrangement, using the genomic coordinates in base pairs (from the ancestral-like haplotype) and also using the gene-rank order. The plots are performed in R with the packages ggplot2 (Wickham, 2016), dplyr (Wickham *et al.*, 2014), ggbreak (Xu *et al.*, 2021), tidyr (Wickham *et al.*, 2024), magrritr (Bache and Wickham, 2022), Cowplot (Wilke, 2023) and Patchwork (Pedersen, 2024). Colours from the Wesanderson (Karthik and Hadley, 2018) and Viridis (Garnier *et al.*, 2021) packages are used for graphical display.

The last step is to objectively infer the existence of evolutionary strata and delimit them, *i.e.* to detect discontinuous changes in d_S mean values along the ancestral gene order. This is performed using the Bayesian analysis implemented in the MCP R package (Lindeløv, 2024). If recombination suppression is a stepwise process, it is expected to result in different mean d_S values (i.e. strata of d_S, **Fig 2A**) in the different fragments that have successively stopped recombining (Charlesworth, 2010). Each stratum can be approximated by a linear model. In principle, different linear models for the distinct strata should fit better than a single model for the whole dataset, and changepoint location should correspond roughly to the boundaries of these strata (**Fig 2B**). Therefore, we implemented in the pipeline a series of models testing different events of recombination suppression, having generated evolutionary strata. Importantly, one-to-one gametologs need to be placed on their relative order including all genes (i.e. also not one-to-one gametologs), as the relative spacing between genes can influence the analysis. We implemented models with up to eight changepoints, which corresponds to six evolutionary strata if pseudo-autosomal (recombining) regions are present on each side of the non-recombining region. The Markov chain Monte Carlo (MCMC) function is run for 100,000 iterations across five replicates and 15,000 burn-in periods, and the results are automatically plotted for each model, including the intervals around the estimated changepoint (using qfit()). The most likely model is statistically chosen using leave-one-out cross-validation as implemented in the loo() function from the loo R package (Vehtari *et al.*, 2015). The estimated log-predictive density (ELPD) is computed for each model and models are compared using the loo_compare() function. Optionally, bayesian hypothesis testing can be performed to compare the significance of mean d_S differences between all pairs of inferred strata, as implemented in the hypothesis() function. BF > 5 is typically considered as good evidence of strong differences in mean d_S values. In addition, violin plots implemented in the ggstatsplots package (Patil, 2021) of d_S values in the different strata are plotted to further visualise the extent and significance of their differences. The coordinates of the strata are then used to visualise a posteriori, the distribution of d_S values along the genome coloured by stratum and to automatically redraw the ideogram also coloured by stratum. For numbers of changepoints from 3 to 8, all the results will be automatically exported in a first pass analysis that returns i) a pdf file displaying the location of the changepoint, ii) a pdf file displaying the convergence of the chain, iii) a pdf file with violin plots of the d_S values for each stratum, iv) a pdf file with colored plot of the d_S values along the ancestral order v) a pdf file with colored plot of the d_S values along physical order, as well as vi) various text files displaying model weights, changepoint location and Bayes Factor testing the support for the different strata.

Running the workflow at different stage of the process

Given the many different use cases of our workflow, we provide different options to run this pipeline (numbered from 1 to 7). While option 1 is used to perform a whole analyse, many users might only be interested in TE and gene prediction, available through option number 6, this can be useful if a single genome is to be annotated without any further analyses being required. Option number 2 will perform the TE, gene prediction, as well as GeneSpace and Synteny analyses. This can be useful, for instance, to help identify sex chromosomes when no such information is available a priori. Option 3 is used to perform GeneSpace and synteny analyses as well as computing synonymous and non-synonymous divergence and infer evolutionary strata (plotting d_S model comparison, Circos plots and Ideograms), which is useful if annotated genomes (gff and fasta files) are already available. Similarly, option 4 allows to compute d_S and infer evolutionary strata in cases where the synteny analysis was already performed. Option 5 is used to perform GeneSpace and synteny analyses only. Finally option 7 will be the option of choice to perform model comparisons and draw associated graphs including ideograms of a posterior strata. This can be highly useful given that option 1 will perform a first-pass model comparison along the 'default' ancestral gene order, which may not allow inferring strata if the 'default' gene order is not the ancestral gene order. Several user crafted modifications of the contig order may be required, or even deleting some contig fragments, for example if part of the ancestral sex or mating-type chromosome became autosomal in the studied species; typically, inferring strata will require visual inspection and biological interpretation from the GeneSpace riparian plots outputs, Circos plots, ideograms without priors and distribution of d_S values. These decision steps cannot be automated as they require interpretation.

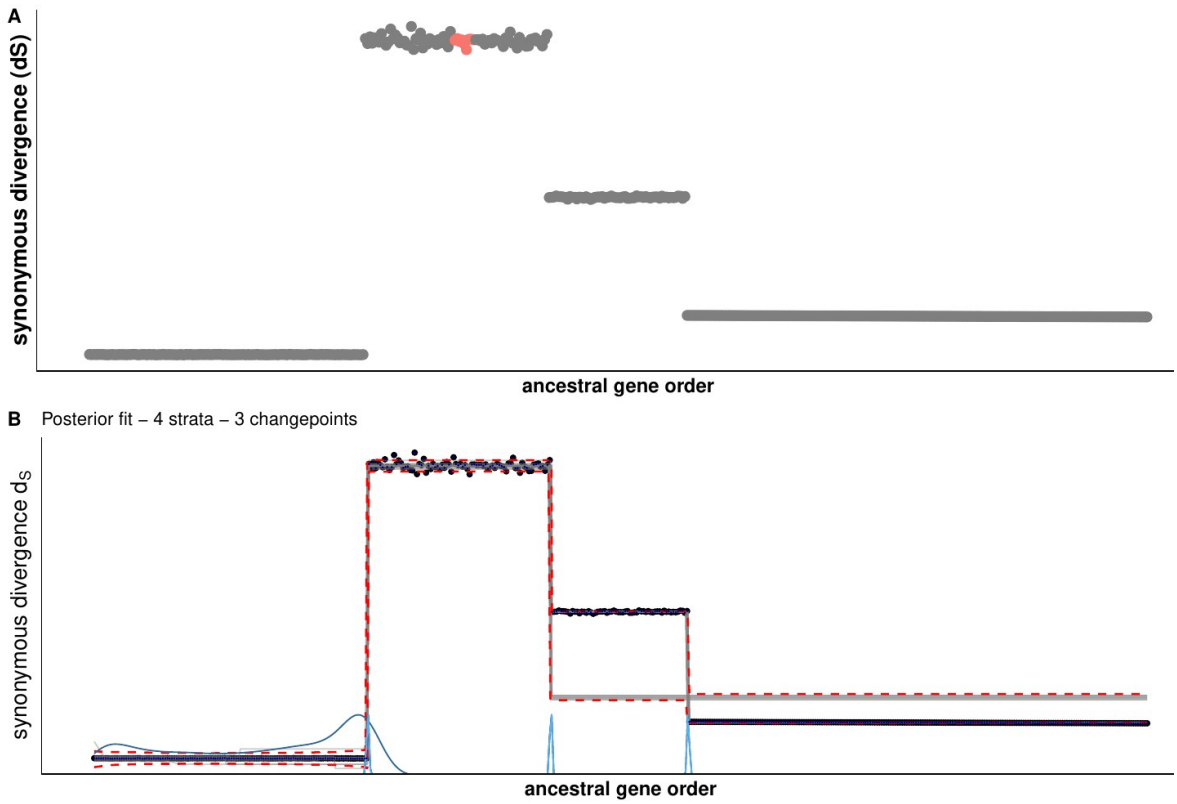

Figure 2: A) Expected pattern of d_S values of synonymous substitutions (d_S) between alleles in the two sex chromosomes or haplotypes under a scenario with two successive events of recombination suppression, the first one leading to the formation of the stratum with the highest mean d_S value and

containing the sex-determining locus (red dot) and the second one leading to a stratum with an intermediate level of synonymous divergence. **B)** Change point location on these simulated data inferred with the MCP (multiple change point) R package in a Bayesian framework. The blue curves at the bottom of the x-axis are the posterior distributions of the change point locations. The dashed red lines are the 2.5% and 97.% quantiles of the fitted (expected) values. The gray lines are 25 samples from the posterior values.

Results: application to a set of genomic data

We propose a new workflow that builds upon existing and up-to-date tools to detect evolutionary strata. Our approach includes three steps: 1) predicting genes and transposable elements, 2) comparing genes and genomes between haplotypes and 3) inferring evolutionary strata through changepoint analyses of evolutionary divergence (d_S). Each module of the workflow can be run separately in order to provide a high level of flexibility for different use cases: i) with all data (2 haplotypes, 1 ancestral sequence, and RNAseq data, so that all analyses will be performed); ii) with or without RNAseq (all analyses will be performed); iii) with only the two haplotypes and no ancestral-like sequence (only gene prediction - d_s/gene comparisons and changepoint analyses will be performed, option 1); iv) with the two haplotypes and an already annotated ancestral-like sequence (only gene prediction - d_s/gene comparisons and changepoint analyses will be performed, option 1); v) with already annotated haplotype and ancestral-like sequence (only the d_s computations, synteny and changepoint analyses will be performed, option 3); vi) with only a single haplotype (only gene annotation will be performed, option 6).

This workflow is mainly intended toward people unfamiliar with bioinformatics, in order to facilitate each of these steps. Below, we briefly show a detailed application of these different steps to a set of 42 published genomes of *Microbotryum* fungi. For the annotation part with different datasets, we focused on *Microbotryum lychnidis-dioicae*, one of the few species for which RNAseq data were available. For the second part (strata inference), we present results for *M. violaceum caroliniana,* a species with well assembled genomes and evolutionary strata of different ages, as well as additional examples from *Microbotryum* species in supplementary results.

Genome annotation

First, we performed on these case studies a classic transposable element (TE) detection that builds upon repeatmodeler and repeatmasker tools. Optionally, an external database of existing TEs for the studied organisms can be provided. Next, we performed genome annotation with BRAKER V3 with either *i*) RNAseq data only, *ii*) custom external evidence, *iii*) online external evidence from orthoDB for the target species or *iv*) all options together. Results for the genome of the mating type a_1 are presented in **Table 2**, (**Table S4** for the mating type a_2). We implemented a filtration step to systematically keep the longest transcript, as many alternative transcripts are produced by BRAKER by default (**Table 2, Table S4**). Mapping rates are computed on the fly after running GSNAP and plots are automatically drawn (**Fig S1-S2,** mean read depth was 160 in a_1 and 166 in a_2 genomes, respectively). Based on the longest transcript, the BUSCO results from the gene prediction tools all produced similar results. These results were in close agreement with the number of BUSCO genes predicted directly on the genome of the a_1 mating type (*i.e.,* C: 98.4% [S: 97.7%, D: 0.7%], F: 0.2%, M: 1.4%) (see **Table S3** for the genomes description) and were slightly better than previously published annotation scores (*M. lychnidis-dioicae* a_1: C: 96.3%

[S: 95.7%, D: 0.6%], F: 1.4%, M: 2.3%; *M. lychnidis-dioicae* a_2: C: 96.2% [S: 95.3%, D: 0.9%], F: 1.5%, M: 2.3%, n: 1764). The least accurate prediction scores were obtained when using only our custom database (built from previous annotations) (**Table 2**).

Additional evaluation is automatically implemented in our workflow by running blast against the uniprot database, enabling the users to further assess the proportion of genes with a significant blast and the length of the hits (**Table 2**). In addition, further annotation information can be obtained if the option to run InterProScan is set to "yes" in the config file (**Table 2**). While BUSCO enables an accurate assessment of the completeness of the gene prediction, it is noteworthy that tools such as BRAKER may fail to predict fast-evolving genes, even when they are well represented in a database. This is in particular the case of the pheromone receptor (PR) gene, which is of major interest in our focal species, controlling pre-mating fusion, and displaying million-year old trans-specific polymorphism (Devier *et al.*, 2009). In the absence of RNAseq data, this important gene was in general not predicted when applying our pipeline across hundreds of genomes. Another issue pertains to the best way of combining RNAseq and external database (ordthoDB12 and custom DB) in the final annotation with BRAKER. While we used TSEBRA following BRAKER's recommendation, we found that the weight given to each evidence impacted very slightly the final BUSCO score. For instance, in our studied case in **Table 2**, we found that the combination of RNAseq with the two other databases (ordthoDB12 and custom DB) actually resulted in a very small decrease in BUSCO accuracy, with six genes being missed in the combined annotation, as compared to results when using RNAseq only. Similarly, only four new genes were classified as fragmented compared to annotation using RNAseq alone. To circumvent this issue, we implemented the parsing of the list of missing genes and added them to the final prediction (which modestly increased the rate of false gene duplications as well: C: 98.9% [S: 96.2%,D: 2.7%], F: 0.5%, M: 0.6%). Yet, we recommend the user to carefully monitor the BUSCO score at each step and decide whether such genes should be added or not. In the same vein, we noted that the longest transcript among multiple alternative transcripts may not be a BUSCO gene, as the BUSCO gene sometimes corresponded to a transcript of shorter size. Therefore, we also implemented some minor corrections to retrieve those transcripts initially predicted but dropped during the size-selection step. Results obtained after reannotation of all 42 genomes from our lab are provided in **Table S5**.

Table 2 : Evaluation of genome annotation results in the case *Microbotryum lychnidis-dioicae* 1064. BUSCO (v5.7.1) scores are provided along with a number of basic statistics for the genome of one mating type (*M. lychnidis-dioicae 1064* a_1). The results for the genome of the second mating type (*M. lychnidis-dioicae 1064* a_2) are provided in **Table S4**. Only the longest transcript was kept for each gene for the BUSCO evaluation. Results are presented for data either i) with only RNAseq data, ii) only the OrthoDB11 dataset for fungi, or iii) a custom database (CustomDB) made from previous gene annotation work in the lab. Column "all" refers to the combination of the three previous data; %uniprot match and %interproscan match refer to the number of Braker genes with at least one match against either uniprot database or the whole databases used in InterProScan. Finally, the number of single-copy orthologs between the ancestral-like mating-type chromosome of *M. lagerheimii* and the a_1/a_2 mating-type chromosomes of *M. lychnidis-dioicae* are provided.

	RNAseq only	OrthoDB11	CustomDB	All
Number of genes	9850	9778	9363	10568
Number of transcripts	11945	13629	12670	14091
Mean CDS length (bp)	1,626.15	1,633.53	1,704.16	1,527.57
Number of introns per gene	4.71	4.3	4.45	4.54
Complete genes	99.81%	99.65	99.77	NA
BUSCO	C:98.9% [S:98.3%,D:0.6%] F:0.2%,M:0.9%	C:98.6% [S:98.0%,D:0.6%] F:0.3%,M:1.1%	C:97.4% [S:96.8%,D:06%] F:0.3%,M:1.2%	C:98.3% [S:97.4%,D:0.6%] F:0.4%,M:1.6%
Complete BUSCOs (C)	1744	1739	1723	1734
Complete and single-copy BUSCOs (S)	1734	1728	1712	1718
Complete and duplicated BUSCOs (D)	10	11	11	10
Fragmented BUSCOs (F)	3	5	10	7
Missing BUSCOs (M)	17	20	31	23
Total BUSCO groups searched	1764	1764	1764	1764
Uniprot match (%)	62.49	64.49	54.84	59.84
InterproScan match (%)	90.18	92.39	93.57	89.13
Single-copy orthologs present on both mating-type chromosomes	620	631	623	634

Genome and Gene Comparisons

Following the gene prediction steps, different operations will be launched if requested. Once the genes are correctly filtered, they are fed to GeneSpace which internally runs Orthofinder, Blast and MCScanX to produce a rich output, suitable both for whole genome comparison, as well as to focus on the non-recombining region of interest (ex: **Fig S3, Fig 3A**). The output itself can be further processed by the user. Next, we used the output of the Phylogenetic Hierarchical Orthogroups (N0.tsv) to infer single-copy orthologs whose numbers observed for the different annotation approaches are provided in the last row of Table 2; highly similar numbers were inferred with all alternatives approaches. These single-copy orthologs are further used for two purposes: *i*) construct an ideogram to display microsynteny (gene-by-gene links between all single-copy orthologs) considering either both mating-type chromosomes, or one mating-type chromosome and an ancestral-like genome (**Fig 5**); and *ii*) construct a circos plot based on the same data (**Fig 3B**) and highlight genes of interest, if any. The purpose of these ideogram and circos plots are two fold: (i) helping understanding the extent of rearrangements between the two mating-type or sex chromosomes, and relative to the ancestral sequence if such sequence is available, (ii) draw

hypotheses regarding the number of evolutionary strata. Moreover, the backbone code that we provide will provide both uncolored ideograms and circos plots as well as colored ideograms based on the distribution of strata inferred from the MCP analysis (**Fig S4**). As a complement to the gene synteny analysis, broad patterns of synteny along sex chromosomes are plotted automatically using the pafr plot_synteny function following a minimap alignment of the genome between each other (**Fig S5**) and dotplots are generated on the fly (**Fig S6**). While this is not a central analysis for the identification of evolutionary strata, it is a straightforward approach to identify inversions, as well as possible assembly errors.

Inference of evolutionary strata

Next, orthofinder single-copy orthologs are reformatted and corresponding CDS are passed as input to translatorX for alignment. The orthologs are used to compute d_S+/-SE and d_N values, as shown in **Fig 4A**. The use of TranslatorX further produces information-rich html tables for each alignment, that can be easily parsed for debugging or to focus on particular genes of interest. For the changepoint analysis (**Fig 5A-C**), the gene order is crucial to inference. When this information is not known *a priori*, users can run the pipeline 'as is' using default values, identify reversed contigs, and rerun the last plotting steps, which typically takes a few minutes (**Fig 4B**).

We applied the pipeline to the *M. v. caroliniana* dataset. After a first pass we identified the contigs forming the mating-type chromosomes, and their homology with those of *M. lagerheimii* **(Fig 3)** and plotted automatically the d_S values (**Fig 4A**). The inference of the rearrangements having formed the *M. v. caroliniana* mating-type chromosomes allowed to decide the order of genes to be plotted on the X axis for the d_S plot (**Fig 4B**) and to perform the changepoint analysis. The model comparison from the changepoint analysis indicates that the best models was those with 7 and 5 changepoints based on the leave-one-out weight (weight = 0.577 and 0.339, **Table S7**) approach, corresponding to 5 and 3 evolutionary strata along with the 2 PARs (pseudo-autosomal regions) on each side (**Table S8** for example of the MCP output). Analysis of the different changepoint locations supported previous inferences of evolutionary strata (Branco et al. 2018, **Fig 5 and Fig 6, Fig S7-S10**). While the leave-one-out cross validation procedure provides insight into which models perform best, the shape of the posterior fitted values (**Fig 5 A-E**) also need to be carefully inspected. Similarly, the posterior plot and mixing and convergence of the MCMC are central checks that need to be performed, as for any Bayesian analysis (**Fig S11**).

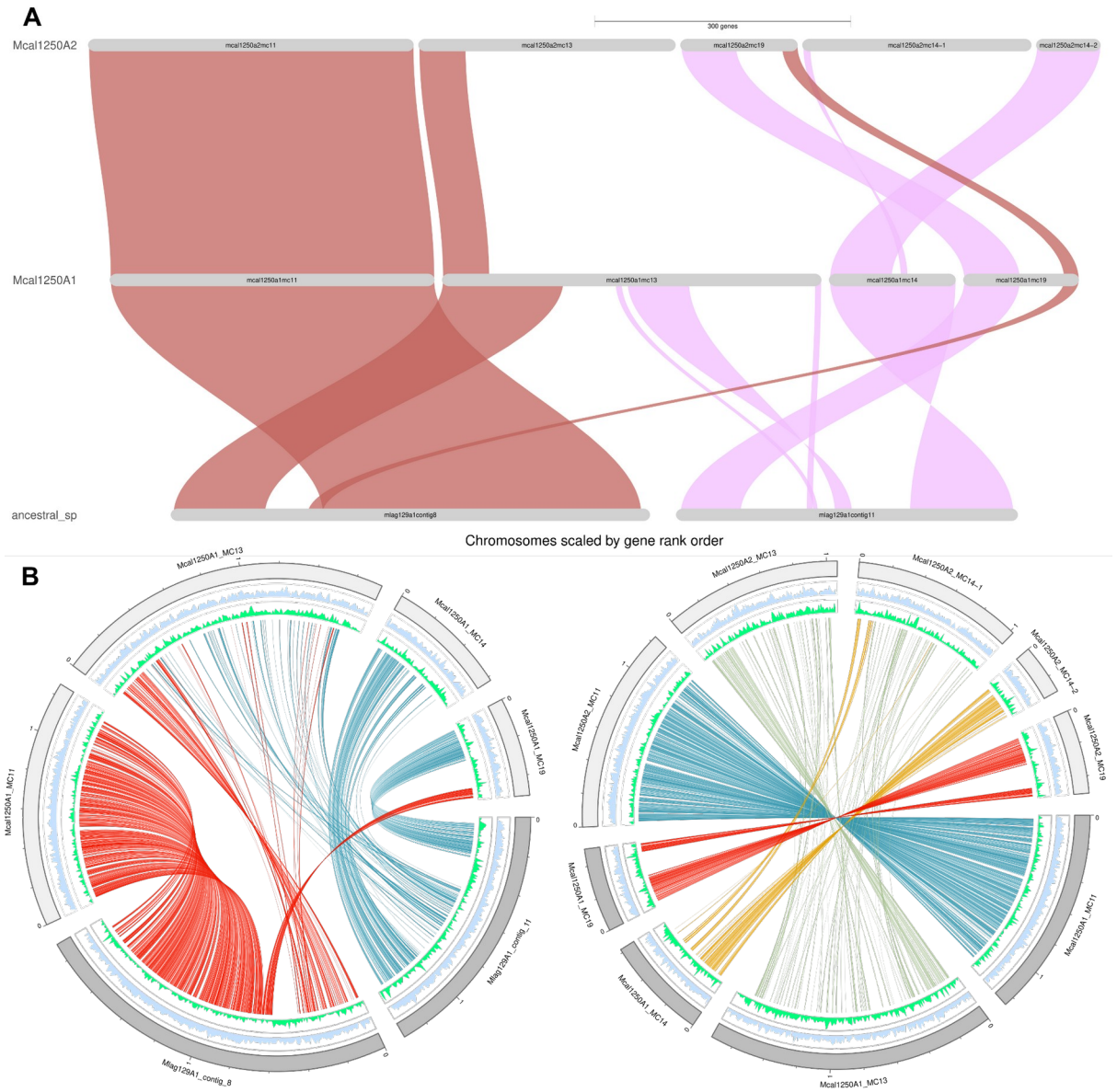

Figure 3 : GeneSpace (A) and Circos plots (B) showing synteny and orthology relationship between mating-type chromosomes of *M. lagerheimii* and *M. v. caroliniana 1250* a_1 and between the two mating-type chromosomes of *M. v. caroliniana 1250* (a_1 and a_2). Mlag129A1_contig 8 corresponds to the HD chromosome and Malg129A1contig11 to the PR chromosome. A) Plots showing major synteny blocks as inferred by combining orthofinder, blasts and MCScan results. Synteny groups must include at least five consecutive genes. The large synteny observed between contig8 and the two corresponding contigs in *M. v. caroliniana* a_1 and a_2 indicates that a large part of the ancestral HD mating-type chromosome is now an autosome. **B)** Circos plots of the *M. v. caroliniana 1250* a_1 mating-type chromosome compared to the *M. lagerheimii* HD and PR mating-type chromosomes (left) and *M. v. caroliniana 1250* a_1 versus a_2 mating-type chromosomes (right). Red link = single-copy orthologous genes between PR contigs. Blue links = single-copy orthologs genes between HD contigs. Inner circle displays gene density in light blue and TE density in spring green respectively.

While a first analysis is implemented with default settings in the “automated” approach of our workflow, the R environment will be entirely exported and we strongly recommend the users to further explore the results manually and test if the evolutionary strata inferred fit with their working hypothesis and get support from other analyses, such as the identified

rearrangements from the synteny analysis, or comparisons with closely related species that may display, for example, different patterns for young evolutionary strata. Here for instance, priors were used regarding the location of the changepoints along the gene order based on the visual change in mean d_S values. The changepoint analysis supported the existence of the previously reported strata in Branco et al., (2017). There was no significant difference in the mean d_S between the youngest evolutionary stratum supporting previous work (light blue in Branco et al. 2017) and the PAR (p>0.5, **Fig 5,** see also **Fig S7-S9** including all detailed statistical tests; BayesFactor from the MCP analysis also returned only modest evidence for the difference in d_S values (BF = 2.99, see table **S9** for all bayes factor).

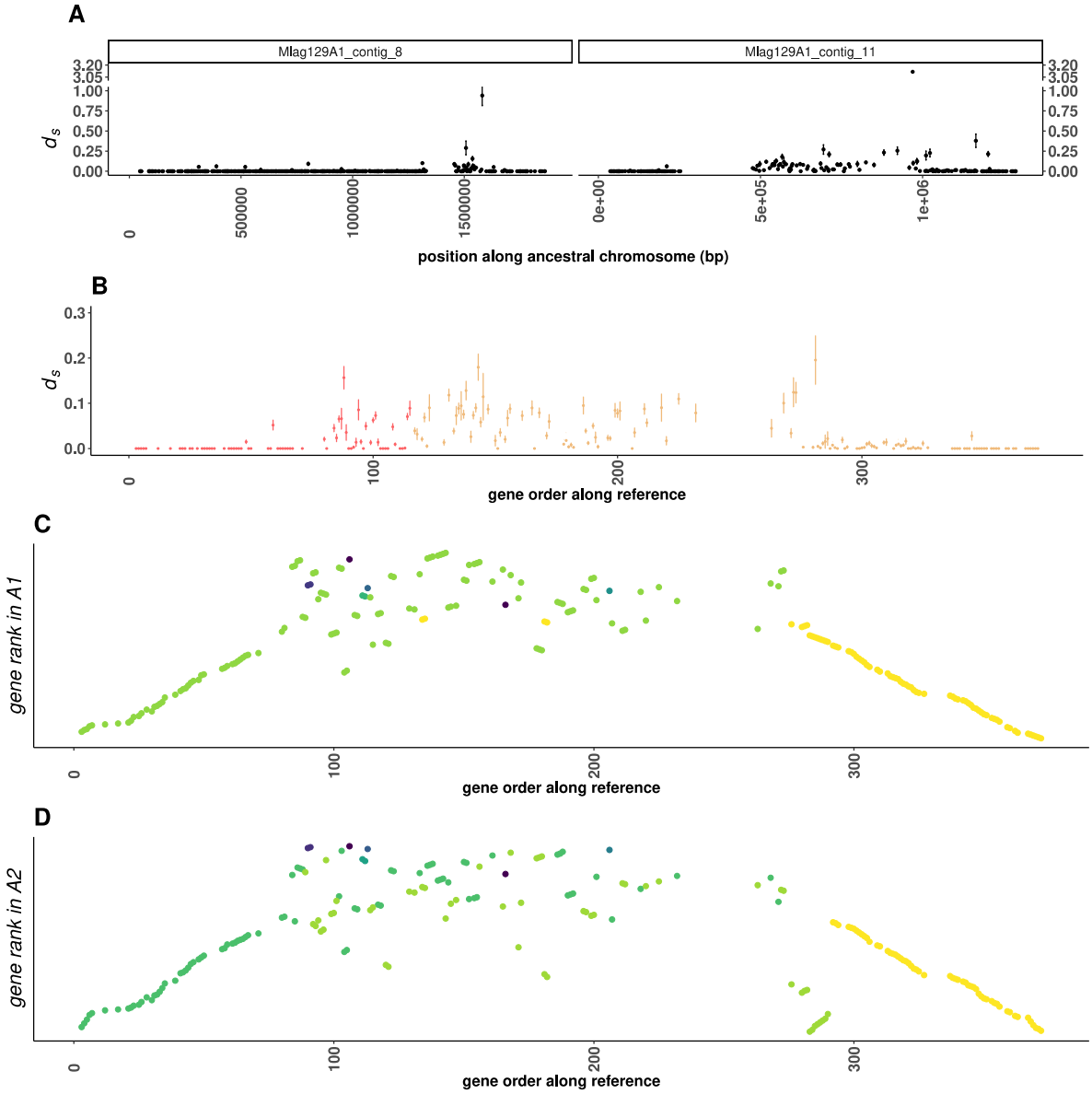

Figure 4: Observed d_S values and rearrangements in the *Microbotryum v. caroliniana* case. A) d_S values between the *M. v. caroliniana* a_1 and a_2 mating-type chromosomes along the mating-type chromosomes of *M. lagerheimii*, with the contig 11 corresponding to the PR chromosome and the contig 8 corresponding to the HD chromosome. B) Distribution of the d_S values considering the ancestral gene order instead of the physical position along the *M. lagerheimii* mating-type chromosomes. Contigs are reversed to match the order of the chromosomal rearrangement and fusion in *M. v. caroliniana* (i.e. contig 8 (HD) in red and contig 11 (PR) in orange). The large region in contig 8 that became an autosome upstream of the centromere in panel A (>500 kbp) has been

removed. The region downstream of the centromere in contig 11 (<460 kbp) is also removed. C) and D) Current rank of the genes in the a_1 and a_2 mating-type chromosomes along the ancestral gene order, showing the rearrangements.

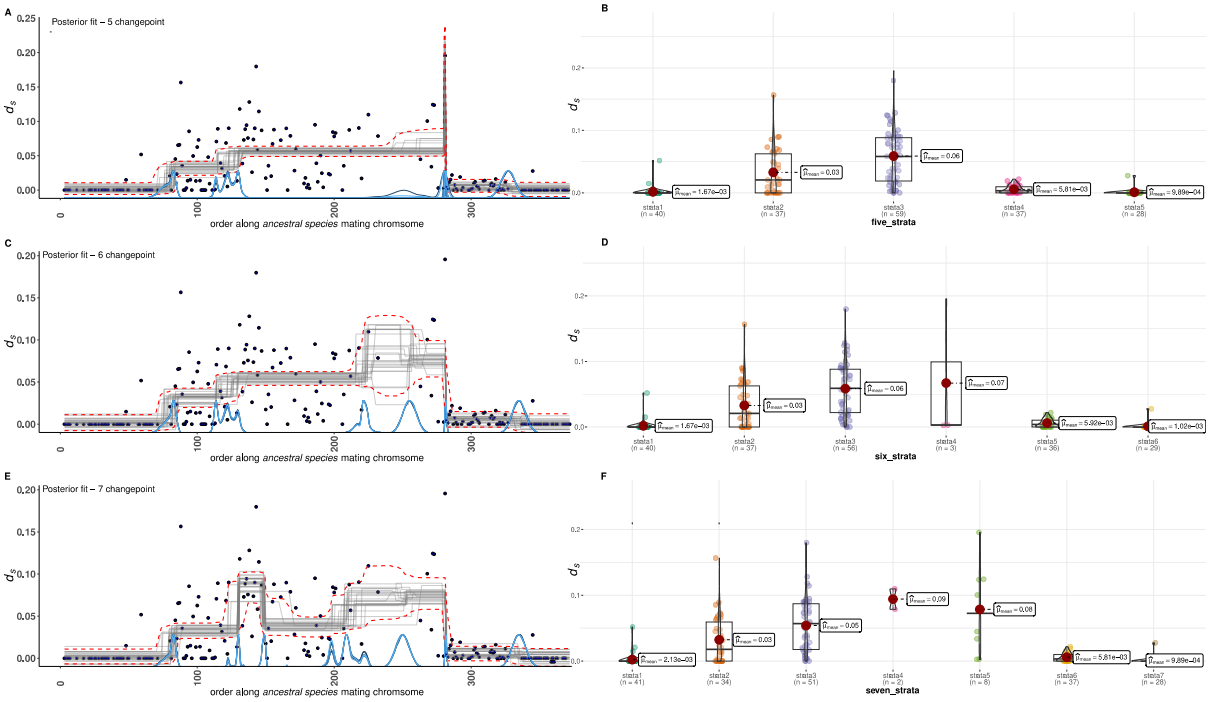

Figure 5: Changepoint inference (A-C-E) and corresponding violin plots (B-D-F) displaying the distribution of d_S values and significance of mean differences between evolutionary strata in *Microbotryum violaceum caroliniana*. Each changepoint panel displays the distribution of raw data (i.e., d_S values as black dots) along with 25 draws from the joint posterior distribution (grey lines) and 95% highest density interval (red lines). Posterior distributions of the changepoints are shown in blue with one line for each chain. Posterior fits are displayed for models with 5, 6 and 7 changepoints, corresponding to 3, 4 and 5 evolutionary strata, respectively. The analysis was run with a prior regarding the location of changepoint, based on observed discrete changes in the distribution of mean d_S values. The same figures for a wider range of changepoints with and without priors are provided in **Fig S7** and **Fig S12,** respectively. The right panel displays the corresponding violin plots, showing the observed distribution of d_S values for each evolutionary stratum, as colored points. Each red dot displays the mean of d_s values along with its inferred value.

Finally, we provide additional results and examples of evolutionary strata inference from the species *M. v. tatarinowii* 1400 in **Figs S13 to S19** as well as the species *Microbotryum lychnidis-dioicae* which was annotated with RNAseq above (**Figs S20 to S26**).

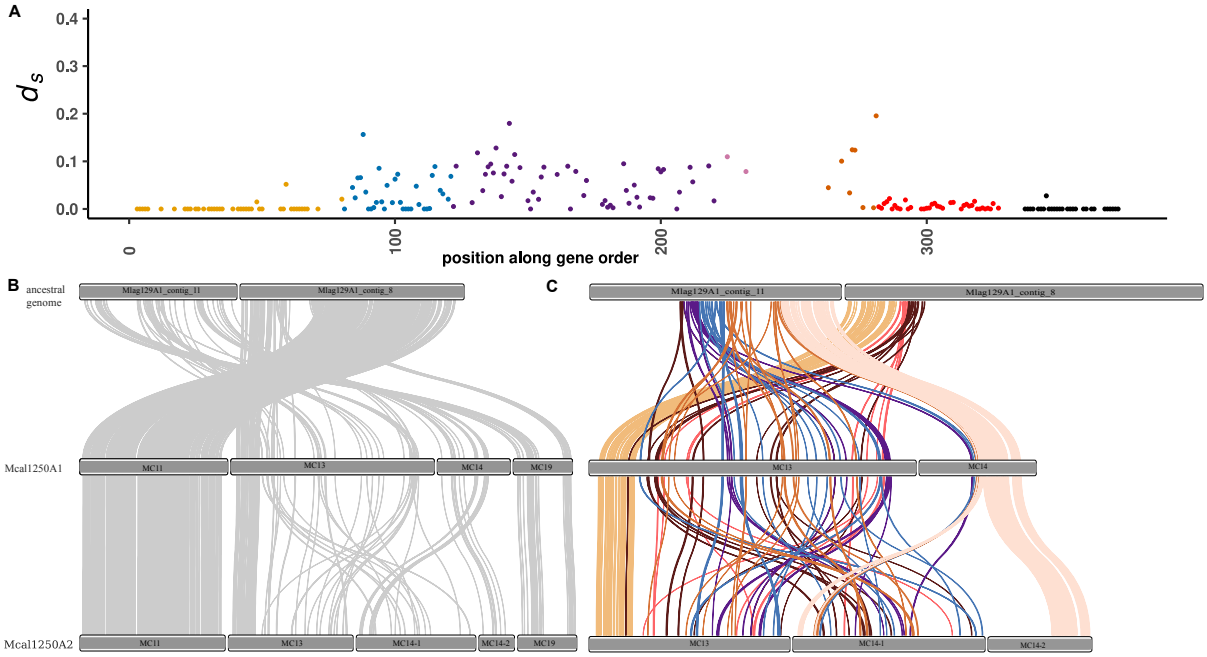

Figure 6: Evolution of mating-type chromosomes in *Microbotryum violaceum caroliniana.* A) Localisation of the evolutionary strata in *M. v. caroliniana* and their d_S values along the ancestral gene order, using as proxy *M. lagerheimii.* Each point represents the value for a gene (See **Fig S4** for the same plots with different numbers of evolutionary strata). Panels B and C display ideograms plot showing links between single-copy ortholog genes of *M. v. caroliniana* a_1 or a_2 and the genome used as a proxy for the ancestral gene order (*M. lagerheimii*). **B)** The most simple version with grey links displays all links, including those on the current autosomal region in *M. v. caroliniana.* **C)** Single-copy orthologs (n = 201) used for the changepoint analysis colored according to their inferred evolutionary strata in the MCP analysis.

Discussion

The increasing number of available genome assemblies opens up new avenues for evolutionary research but can be at the same time a significant challenge, given the numerous softwares, dependencies and sometimes complex tools that might be needed to correctly annotate and analyse large-scale genomic data. In addition, the inference of evolutionary strata has often relied on simple visual inspection of d_S plots without any statistical tests or other lines of evidence, that can be presence /absence of strata in closely related species (Branco et al. 2017). Here, we provide a straightforward workflow that aims to simplify the various tasks required to infer evolutionary strata along non-recombining regions in genomes, for example in sex chromosomes. This workflow is flexible and should scale well with genome assemblies of any size. We provide here examples of its application with the small genomes of *Microbotryum* fungi (30 to 45 Mb), and it was also useful for identifying evolutionary strata in the giant Y chromosome of the plant *Silene latifolia* (2.9 Gb genome with a ~ 350 and ~500 Mb long X and Y chromosomes respectively) (Moraga *et al.*, 2024). It is also possible to use it to simply perform gene prediction, in which case only the TE discovery and Braker steps will be run. Similarly, if gene models and genome assemblies are already available, it is possible to skip the gene prediction step, so that only the assembly mapping, orthology, synteny and/or d_S analyses will be performed, providing a high level of flexibility to the end users.

The quality of the analyses will depend on the availability of an accurate proxy for the ancestral gene order, especially for mating-type chromosomes and other supergenes for which both haplotypes are non-recombining and therefore potentially rearranged. In the absence of a proxy for an ancestral arrangement, we recommend to use the order along the X or Z chromosome in the case of sex chromosomes, and an outgroup is still recommended to help identify single-copy orthologs. Another critical factor is the quality of the genome assembly. Users should aim at providing the most accurate and contiguous assembly, as assessed by the BUSCO completeness at the genome level, QV statistics, N50 and other standard measures (see e.g. (Rougemont, 2024). In particular, orientation errors and other kinds of misassemblies that mimic inversions may result in false interpretations of degeneration signals associated with the loss of recombination. By default, the pipeline performs minimap alignments and constructs dot plots on the fly among the genomes to be studied, which provides an easy way to spot remaining assembly errors. While minimap is not designed to investigate full synteny, other minimizer-based approaches, such as nt-synt (Coombe *et al.*, 2024), could be used to complement the implemented analysis.

In addition, the accuracy of gene annotation may be affected by the availability of RNAseq data. This can be an issue if the studied group is poorly represented in existing databases, resulting in few external protein sequence datasets being available. At the end of the genome annotation steps, BUSCO scores are automatically computed. Users should aim at scores well above 90% of complete gene sets to confidently trust the rest of the analysis. By default, we implemented a checkpoint, automatically stopping the analysis if BUSCO scores are below 85%. In addition, some proteins might have a highly dynamic or particular evolution, especially considering the rate of degeneration of sex chromosomes (Charlesworth, 2021), and may thus be hard to predict; this is particularly the case of the PR gene in the *Microbotryum* case, but is likely to be a more general issue. An easy approach to circumvent this problem, that we successfully implemented, is to transfer the annotation, using Miniprot (Li, 2023), from a phylogenetically close species in which the protein was identified. We included an example script to do so in our workflow.

Here, we thus provide a simple way to assess changes in mean d_s values along non-recombining regions in order to identify evolutionary strata. Our workflow uses functions implemented in the MCP package, which is one of the most comprehensive for this purpose to our knowledge. Its Bayesian basis allows an assessment of the credible intervals around mean changepoints, as well as an assessment of the support for different changepoint sites and different models. However, the results of the MCP analyses, especially the support for a given model, should be critically weighted by the user to make sure that the inferred strata make biological sense and that they are consistent with other types of evidence. Ideally, this analysis should be complemented by analyses of i) structural rearrangements (e.g. inversions, deletions, transpositions and duplications), as these likely affect recombination and accumulate with time in non-recombining regions (Jay et al 2024), ii) the presence/absence of footprints of recombination suppression in closely related species, if available, as young evolutionary strata may be specific to some species (Branco *et al.*, 2017, 2018), and iii) the segregation of alleles and trans-specific polymorphism to infer the age of recombination suppression relative to speciation events (Hartmann *et al.*, 2021b; Branco *et al.*, 2018, 2017; Duhamel *et al.*, 2022). As these analyses heavily rely on the availability of outgroup data, they are not covered by our workflow, but can be conducted using phylogenetic approaches. Trans-specific polymorphism can be examined by inferring genealogical trees for the set of genes belonging to each inferred stratum. In case of evolutionary strata having evolved between different speciation events, the expectation is

that gene trees in the different strata display different topologies and different branch lengths.

Our workflow combines a curated set of qualitative and up-to-date softwares. However, as bioinformatics is a fast-paced field, it is possible to modify our workflow to take advantage of newly developed tools. For instance, we obtained quick and high-quality gene prediction for *Microbotryum* fungi using the newly developed Helixer pipeline (Stiehler *et al.*, 2021). This could be easily combined with Braker to provide a high-quality set of genes (Moraga *et al.*, 2023b). We expect that the flexibility of our workflow will facilitate the systematic discovery of evolutionary strata across the tree of life to help understand patterns and evolutionary processes.

Data Availability

The code and data are available on GitHub: https://github.com/QuentinRougemont/EASYstrata/.

All dataset used to illustrate the pipeline were previously published. The new genome assembly for *M. lagerheimii* 129.010A1 is available on github.

Acknowledgments

We would like to thank Jean-Philippe Vernadet for help with the server maintenance as well as Michel Bartoli for strain collection of the new *M. lagerheimii* 129.01 strain.

Funding

This work was supported by the European Research Council (ERC) EvolSexChrom (832352) grant to TG.

References

Abascal,F. *et al.* (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res*, **38**, W7-13.

Álvarez-Carretero,S. *et al.* (2023) Beginner's Guide on the Use of PAML to Detect Positive Selection. *Molecular Biology and Evolution*, **40**, msad041.

Bache,S.M. and Wickham,H. (2022) magrittr: A Forward-Pipe Operator for R. *https://magrittr.tidyverse.org, https://github.com/tidyverse/magrittr.*

Bergero,R. and Charlesworth,D. (2009) The evolution of restricted recombination in sex chromosomes. *Trends Ecol Evol*, **24**, 94–102.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Branco,S. *et al.* (2017) Evolutionary strata on young mating-type chromosomes despite the lack of sexual antagonism. *Proceedings of the National Academy of Sciences*, **114**, 7067–7072.

Branco,S. *et al.* (2018) Multiple convergent supergene evolution events in mating-type chromosomes. *Nat Commun*, **9**, 2000.

Brůna,T. *et al.* (2024) GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.*, **34**, 757–768.

Charlesworth,B. *et al.* (2014) The evolutionary dynamics of sexually antagonistic mutations in pseudoautosomal regions of sex chromosomes. *Evolution*, **68**, 1339–1350.

Charlesworth,D. (2016) The status of supergenes in the 21st century: recombination suppression in Batesian mimicry and sex chromosomes and other complex adaptations. *Evolutionary Applications*, **9**, 74–90.

Chen,S. (2023) Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta*, **2**, e107.

Coombe,L. *et al.* (2024) Multi-genome synteny detection using minimizer graph mappings. 2024.02.07.579356.
De Coster,W. and Rademakers,R. (2023) NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics*, **39**, btad311.
Devier,B. *et al.* (2009) Ancient trans-specific polymorphism at pheromone receptor genes in basidiomycetes. *Genetics*, **181**, 209–223.
Duhamel,M. *et al.* (2023) Dynamics of transposable element accumulation in the non-recombining regions of mating-type chromosomes in anther-smut fungi. *Nat Commun*, **14**, 5692.
Duhamel,M. *et al.* (2022) Onset and stepwise extensions of recombination suppression are common in mating-type chromosomes of Microbotryum anther-smut fungi. *Journal of Evolutionary Biology*, **35**, 1619–1634.
Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.
Flynn,J.M. *et al.* (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, **117**, 9451–9457.
Formenti,G. *et al.* (2022) The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution*, **37**, 197–202.
Gabriel,L. *et al.* (2024) BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv*, 2023.06.10.544449.
Gabriel,L. *et al.* (2021) TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics*, **22**, 566.
Garnier,S. *et al.* (2021) sjmgarnier/viridis: viridis 0.6.0 (pre-CRAN release).
Hao,Z. *et al.* (2020) RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci*, **6**, e251.
Hartmann,F.E. *et al.* (2021a) Recombination suppression and evolutionary strata around mating-type loci in fungi: documenting patterns and understanding evolutionary and mechanistic causes. *New Phytologist*, **229**, 2470–2491.
Hartmann,F.E. *et al.* (2021b) Recombination suppression and evolutionary strata around mating-type loci in fungi: documenting patterns and understanding evolutionary and mechanistic causes. *New Phytologist*, **229**, 2470–2491.
Hough,J. *et al.* (2014) Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 7713–7718.
Jay,P. *et al.* (2024) Why do sex chromosomes progressively lose recombination? *Trends in Genetics*, **40**, 564–579.
Karthik,R. and Hadley,W. (2018) wesanderson: A Wes Anderson Palette Generator_. R package version 0.3.6.
Kolmogorov,M. *et al.* (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*, **37**, 540–546.
Lahn,B.T. and Page,D.C. (1999) Four Evolutionary Strata on the Human X Chromosome. *Science*, **286**, 964–967.
Li,H. (2021) New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, **37**, 4572–4574.
Li,H. (2023) Protein-to-genome alignment with miniprot. *Bioinformatics*, **39**, btad014.
Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
Li,K. *et al.* (2023) Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat Commun*, **14**, 6556.
Lindeløv,J.K. (2024) mcp: An R Package for Regression With Multiple Change Points.
Moraga,C. *et al.* (2024) The Silene latifolia genome and its giant Y chromosome. 2023.09.21.558754.
Muffato,M. *et al.* (2023) Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. *Nat Ecol Evol*, **7**, 355–366.

Patil,I. (2021) Visualizations with statistical details: The ‘ggstatsplot’ approach. *Journal of Open Source Software*, **6**, 3167.
Pedersen,T.L. (2024) patchwork: The Composer of Plots.
Pertea,G. and Pertea,M. (2020) GFF Utilities: GffRead and GffCompare.
Ranwez,V. *et al.* (2011) MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. *PLOS ONE*, **6**, e22594.
Rougemont,Q. (2024) Genome Assembly Workflow, QuentinRougemont/genome_assembly: v1.0-beta (v1.0-beta). *Zenodo*.
Simão,F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
Smit *et al.* (2013) RepeatMasker Open-4.0.
Stanke,M. *et al.* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*, **34**, W435-439.
Stiehler,F. *et al.* (2021) Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning. *Bioinformatics*, **36**, 5291–5298.
Tegenfeldt,F. *et al.* (2024) OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Research*, gkae987.
The Darwin Tree of Life Project Consortium (2022) Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences*, **119**, e2115642118.
Umen,J. and Coelho,S. (2019) Algal Sex Determination and the Evolution of Anisogamy. *Annu Rev Microbiol*, **73**, 267–291.
Vasimuddin,Md. *et al.* (2019) Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In, *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*., pp. 314–324.
Vehtari,A. *et al.* (2015) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv.org*.
Walker,B.J. *et al.* (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, **9**, e112963.
Weisman,C.M. *et al.* (2022) Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Curr Biol*, **32**, 2632-2639.e2.
Wickham,H. *et al.* (2014) dplyr: A Grammar of Data Manipulation. R package version 1.1.4,. *https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.*
Wickham,H. (2016) ggplot2 Springer International Publishing, Cham.
Wickham,H. *et al.* (2024) tidyr: Tidy Messy Data. R package version 1.3.1.
Wilke,C.,O. (2023) cowplot: Streamlined Plot Theme and Plot Annotations for ‘ggplot2’. R package version 1.1.2.
Winter,D. (2020) Pafr R Package Scholar & Citation Profiles. *R Package Scholars and Citations*.
Wu,T.D. *et al.* (2016) GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. *Methods Mol Biol*, **1418**, 283–334.
Xu,S. *et al.* (2021) Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. *Front. Genet.*, **12**.
Yue,J. *et al.* (2023) The origin and evolution of sex chromosomes, revealed by sequencing of the Silene latifolia female genome. *Curr Biol*, **33**, 2504-2514.e3.