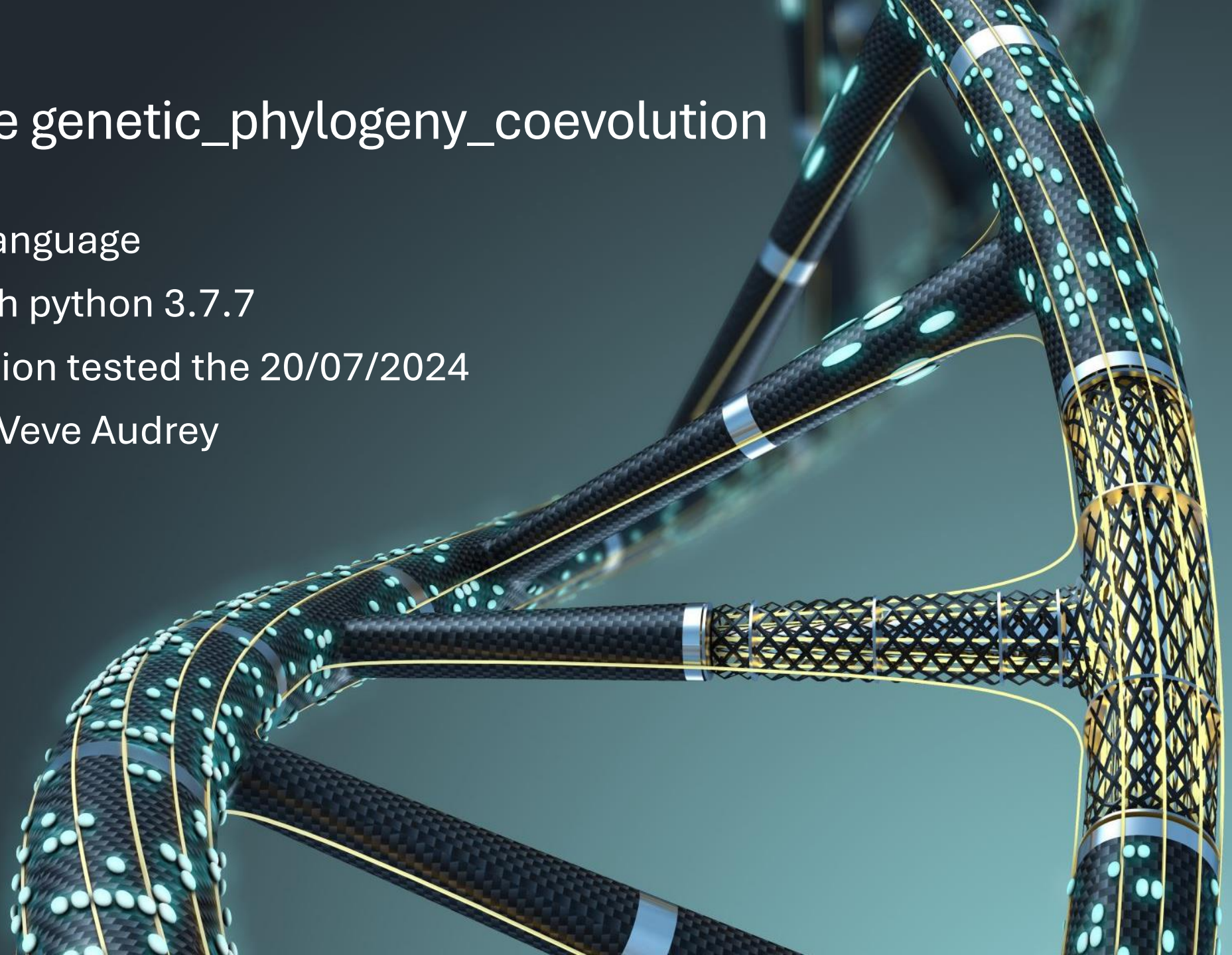
Pipeline genetic_phylogeny_coevolution

Python language

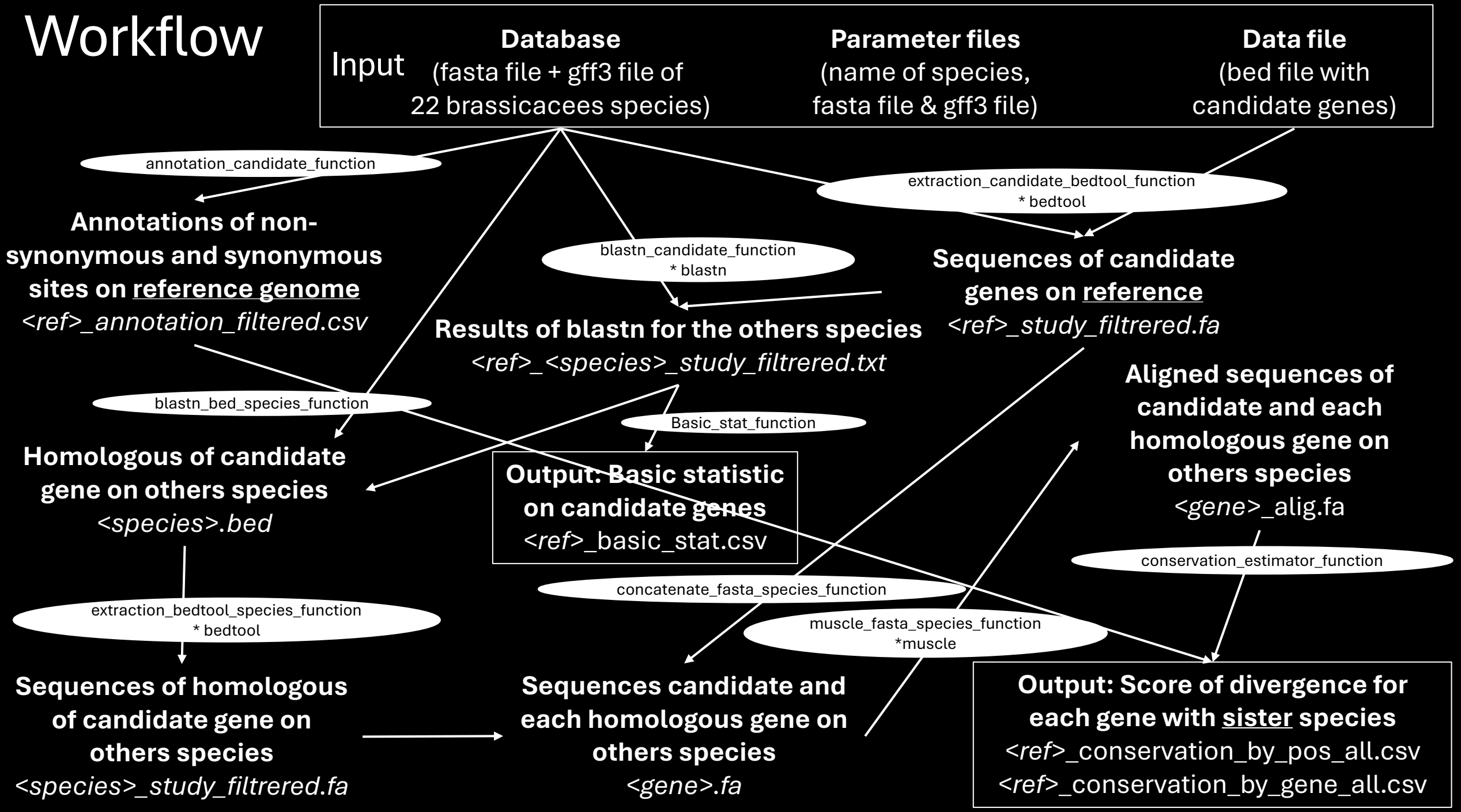
Build with python 3.7.7

Last version tested the 20/07/2024

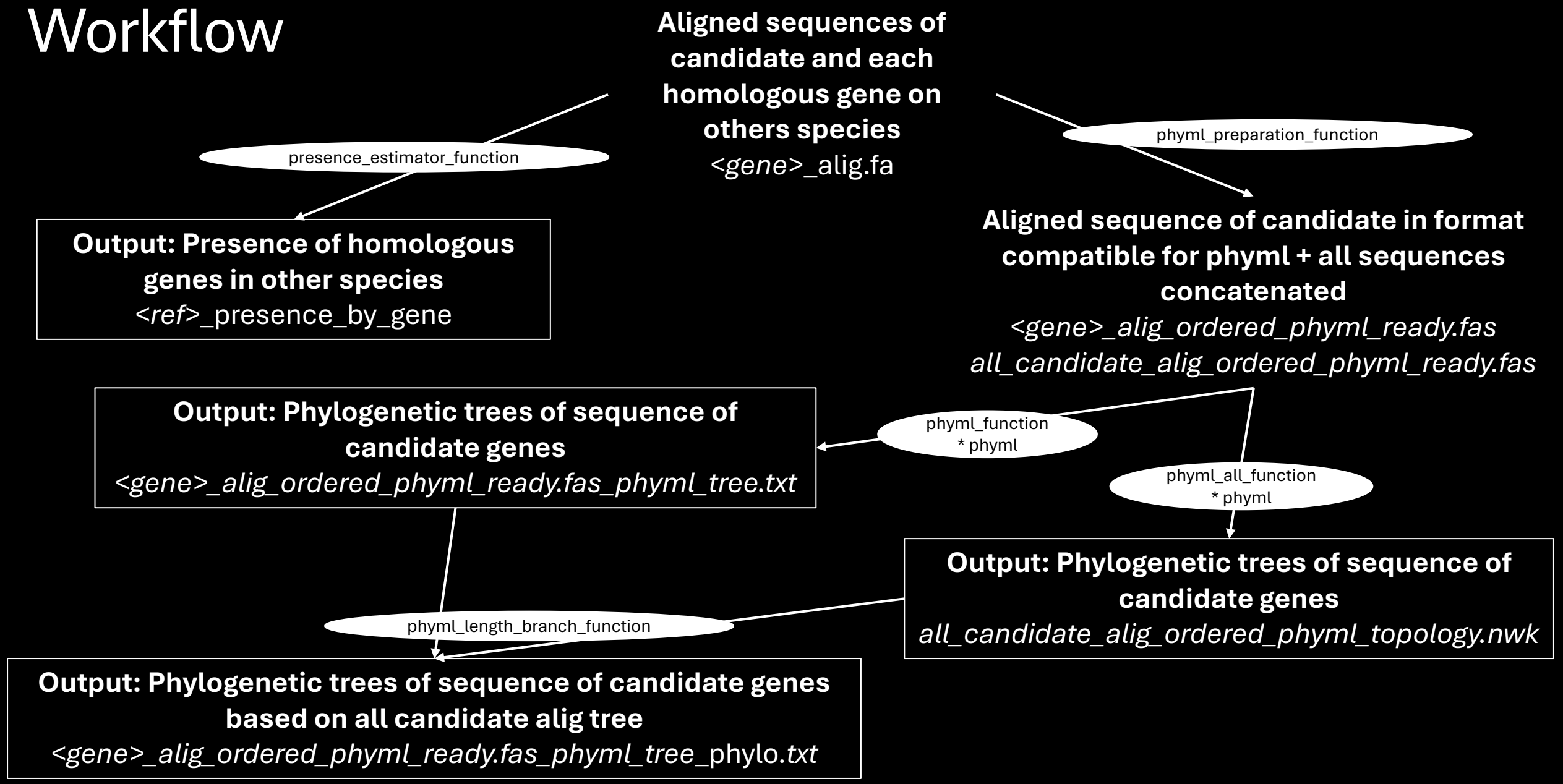
By Dr Le Veve Audrey



Workflow


Input
Database
(fasta file + gff3 file of
22 brassicacees species)
Parameter files
(name of species,
fasta file & gff3 file)
Data file
(bed file with
candidate genes)
annotation_candidate_function
Annotations of non-synonymous and synonymous sites on reference genome
<ref>_annotation_filtered.csv
extraction_candidate_bedtool_function
* bedtool
Sequences of candidate genes on reference
<ref>_study_filtrered.fa
blastn_candidate_function
* blastn
Results of blastn for the others species
<ref>_<species>_study_filtrered.txt
blastn_bed_species_function
Basic_stat_function
Homologous of candidate gene on others species
<species>.bed
Output: Basic statistic on candidate genes
<ref>_basic_stat.csv
Aligned sequences of candidate and each homologous gene on others species
<gene>_alig.fa
conservation_estimator_function
concatenate_fasta_species_function
extraction_bedtool_species_function
* bedtool
muscle_fasta_species_function
*muscle
Sequences of homologous of candidate gene on others species
<species>_study_filtrered.fa
Sequences candidate and each homologous gene on others species
<gene>.fa
Output: Score of divergence for each gene with sister species
<ref>_conservation_by_pos_all.csv
<ref>_conservation_by_gene_all.csv

Workflow


Aligned sequences of candidate and each homologous gene on others species
<gene>_alig.fa
presence_estimator_function
phyml_preparation_function
Output: Presence of homologous genes in other species
<ref>_presence_by_gene
Aligned sequence of candidate in format compatible for phyml + all sequences concatenated
<gene>_alig_ordered_phyml_ready.fas
all_candidate_alig_ordered_phyml_ready.fas
phyml_function
* phyml
phyml_all_function
* phyml
Output: Phylogenetic trees of sequence of candidate genes
<gene>_alig_ordered_phyml_ready.fas_phyml_tree.txt
Output: Phylogenetic trees of sequence of candidate genes
all_candidate_alig_ordered_phyml_topology.nwk
phyml_length_branch_function
Output: Phylogenetic trees of sequence of candidate genes based on all candidate alig tree
<gene>_alig_ordered_phyml_ready.fas_phyml_tree_phylo.txt

Input files

- Database and parameter file

the database file contains all the reference genomes in fasta format, their indexations in fai format and the gff3 files corresponding mentioned in Parameter_file.txt

The Parameter_file.txt contains one header line, and the following lines are compound of name of the species, the name of the fasta file corresponded to the reference genome and the gff, separated by tabulation

Parameter_file.txt

Fichier Modifier Affichage

```
#Species        fasta Files in Database gff files in Database
A.halleri       Ahalleri_765_v2.03.fa   Ahalleri_765_v2.1.0.gene_exons.gff3
A.lyrata        Alyrata_384_v1.fa       Alyrata_384_v2.1.gene_exons.gff3
A.thaliana      Athaliana_167_TAIR10.fa Athaliana_167_gene_exons.gff3
B.oleracea      Boleraceacapitata_446_v1.0.fa   Boleraceacapitata_446_v1.0.gene_exons.gff3
B.stricta       Bstricta_278_v1.fa      Bstricta_278_v1.2.gene_exons.gff3
B.rapa  Brapassp_trilocularisR500_795_v2.0.fa   Brapassp_trilocularisR500_795_v2.1.gene_exons.gff3
C.amplexicaulis Camplexicaulis_470_v1.fa        Camplexicaulis_470_v1.1.gene_exons.gff3
D.strictus      Dstrictus_582_v2.0.fa   Dstrictus_582_v2.1.gene_exons.gff3
D.sophioides    Dsophioides_482_v1.fa   Dsophioides_482_v1.1.gene_exons.gff3
C.violacea      Cviolacea_585_v2.0.fa   Cviolacea_585_v2.1.gene_exons.gff3
C.rubella       Crubella_474_v1.fa      Crubella_474_v1.1.gene_exons.gff3
C.grandiflora   Cgrandiflora_266_v1.fa  Cgrandiflora_266_v1.1.gene_exons.gff3
T.arvense       Tarvense_479_v1.fa      Tarvense_479_v1.1.gene_exons.gff3
R.islandica     Rislandica_473_v1.fa    Rislandica_473_v1.1.gene_exons.gff3
M.maritima      Mmaritima_477_v1.fa     Mmaritima_477_v1.1.gene_exons.gff3
L.sativum       Lsativum_478_v1.fa      Lsativum_478_v1.1.gene_exons.gff3
L.annua Lannua_476_v1.fa        Lannua_476_v1.1.gene_exons.gff3
I.amara Iamara_485_v1.fa        Iamara_485_v1.1.gene_exons.gff3
E.vesicaria     Evesicaria_487_v1.fa    Evesicaria_487_v1.1.gene_exons.gff3
E.syriacum      Esyriacum_483_v1.fa     Esyriacum_483_v1.1.gene_exons.gff3
E.salsugineum   Esalsugineum_173_v1.fa  Esalsugineum_173_v1.0.gene_exons.gff3
S.parvula       Sparvula_574_v2.0.fa    Sparvula_574_v2.2.gene_exons.gff3
```

Input files

- <u>Data and bedfile file</u>

the data file contains the bedfile, named candidate.bed, corresponding to the coordonate of candidate genes in reference genome (*A. lyrata* by default). The bedfile must contain four columns, with the chromosome, the start, the endof the candidate gene and the type of gene in last column (ex: tissu, control vs condition, ...)

candidate.bed

Fichier Modifier Affichage

#Chrom	start	end	Genes	Cluster
scaffold_103	4936	7534	AL103U10020	Root
scaffold_144	302	881	AL144U10010	Root
scaffold_1	146301	149871	AL1G10410	Root
scaffold_1	377030	380594	AL1G11110	Root
scaffold_1	422251	424145	AL1G11220	Root
scaffold_1	454231	455675	AL1G11320	Root
scaffold_1	501270	503644	AL1G11450	Root
scaffold_1	536111	538652	AL1G11530	Root
scaffold_1	596585	598002	AL1G11710	Root
scaffold_1	628244	630982	AL1G11800	Root
scaffold_1	650407	652260	AL1G11840	Root
scaffold_1	777219	779320	AL1G12270	Root
scaffold_1	798556	799184	AL1G12340	Root
scaffold_1	806192	807344	AL1G12370	Root
scaffold_1	810324	811634	AL1G12390	Root
scaffold_1	841019	844200	AL1G12460	Root
scaffold_1	851658	854434	AL1G12490	Root
scaffold_1	935514	939755	AL1G12690	Root
scaffold_1	1084899	1086115	AL1G13190	Root
scaffold_1	1160919	1162019	AL1G13370	Root
scaffold_1	1162035	1162884	AL1G13380	Root
scaffold_1	1170647	1173194	AL1G13400	Root
scaffold_1	1189773	1190464	AL1G13430	Root
scaffold_1	1195901	1196794	AL1G13450	Root
scaffold_1	1272552	1274010	AL1G13650	Root
scaffold_1	1380224	1383500	AL1G13880	Root
scaffold_1	1404736	1406227	AL1G13930	Root
scaffold_1	1411744	1413197	AL1G13960	Root
scaffold_1	1472424	1474717	AL1G14140	Root
scaffold_1	1503442	1505172	AL1G14220	Root
scaffold_1	1618698	1619435	AL1G14490	Root
scaffold_1	1754630	1757166	AL1G14850	Root

Programs required

The pipeline requires python and some other programs:bedtools 2.26, phyml 3.3, muscle 3.8 and blast 2.10. During pipeline step, these different programs are called according to our personal servor. Thus, the names to call these different could be different on your servor. You can change the name with the different options, respectively:

-bedt or --bedtools <name>

-fi or –phyml <name>

-m or –muscle <name>

-bns or --blastn <name>

Steps on workflow

Each step define in the workflow could be skeep using the name of the step (see workflow section) and the parameter « F ». For example, to skeep the step that give the basic statistic, use :

--basic_stat_function or --bsf F

Be careful, if one step is skeeped, all following dependant steps could generate an error if the input files required are not available.

Others parameter

By default, the pipeline works with *A. lyrata* as reference genome and it gff3 file, *A. halleri* as sister species, the species in file Parameter_file.txt, a evalue of blastn of 1e-5 and a minimum coverage of the candidate homologous gene found by blastn and the query of 70%. To change these paramètres, uses respectively:

-P or --parameter_file <Parameter_file.txt>

-R or --ref <A.lyrata>

-g or --gff <Alyrata_384_v2.1.gene_exons.gff3>

-S or --sister <A.halleri>

-e or --eval_blastn <1e-5>

-c or --coverage <70>

Output files

The pipeline could generate different phylogenetic tree in nwk format:

- The phylogenetic trees of all genes (*<gene>_alig_ordered_phyml_ready.fas_phyml_tree.txt*).
- The phylogenetic tree without lenght of branches based on the concatenation of all control genes found in 19 species or more (*all_candidate_alig_ordered_phyml_topology.nwk*). This tree could be used as reference topology for other gene and to estimate length of branches.
- The phylogenetic tree of all genes with lenght of branches following the reference topology (*<gene>_alig_ordered_phyml_ready.fas_phyml_tree*_phylo.*txt*)

Output files

The pipeline could generate a summary file in csv format sep « ; » that give, for each species in Parameter_file.txt, excepted the reference species, if homologous genes were found or not (<*ref*>_presence_by_gene)

Output files

The pipeline could generate a summary file in csv format sep « ; » that give, for each genes, the conservation rate for all positions, the non-synonymous and synonymous position of the reference genome, compared to sister species and all species by pair (<*ref*>_conservation_by_pos_all.csv).

Moreover, the pipeline could generate a summary file in csv format sep « ; » that give, for each genes, the mean conservation rate for all genes on all sites, on non-synonymous and synonymous sites of the reference genome, compared to sister species and the mean obtained compared to all species by pair (<*ref*>_conservation_by_gene_all.csv).