Supplementary Materials to “PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies.”

Ludovic Mallet[1], Tristan Bitard-Feildel[2], Franck Cerutti[1], and Hélène Chiapello[1]

[1]INRA UR875, Unité Mathématiques et Informatique Appliquées de Toulouse (MIAT), Auzeville, 31326 Castanet-Tolosan, France
[2]CNRS UMR7590, Sorbonne Universités, Université Pierre et Marie Curie – Paris 6 – MNHN – IRD – IUC, Paris, France.

January 12, 2017

Contents

1 Workflow

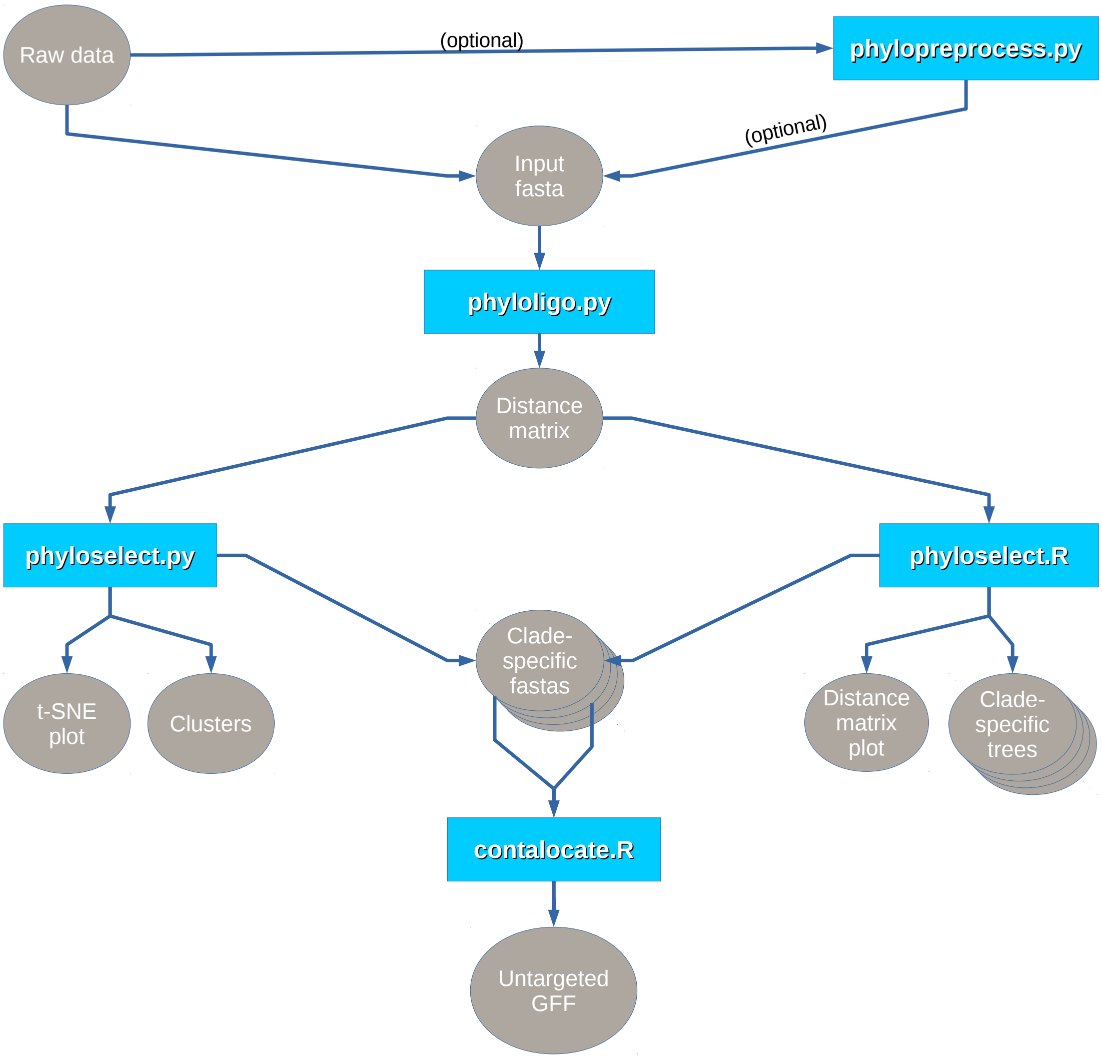


Figure 1: Workflow of PhylOligo.

2 Installation

PhyloOligo software needs python 3.4 or newer and several R and python packages.

2.1 Quick Install

Basic dependencies

If python or R are not installed on your system, call your distribution's package manager:

```
sudo apt-get install python3-dev python3-setuptools r-base git emboss samtools
#or
yum install python3-dev python3-setuptools r-base git emboss samtools
```

Clone/download the git repository

```
git clone https://github.com/itsmeludo/PhylOligo.git
```

or download it from https://github.com/itsmeludo/PhylOligo

Install python scripts and dependencies

If you have administrator rights or if you are working in a python virtual environment:

```
git clone https://github.com/itsmeludo/PhylOligo.git
cd PhylOligo
pip3 install .
```

You can also install it locally using:

```
git clone https://github.com/itsmeludo/PhylOligo.git
cd PhylOligo
pip3 install . --user
```

Or to install it locally in a folder of your choice:

```
pip3 install . --prefix /my/local/folder
```

If locally installed, be sure to add the local directory with executable in your executable path. On linux:

```
export PATH=$HOME/.local/bin:$PATH

phyloligo.py -h
```

2.2 Alternative install tricks

If the easy install procedure fails on your system, there are several options to install the dependencies.

Python requirements

If you want to install the dependencies separately use:

```
cd PhylOligo
pip3 install -r requirements.txt
```

Install R scripts and dependencies

In R, as root or user

```
R
install.packages(c("ape","getopt","gplots"))
```

Rights and paths

Link the programs into a directory listed in your $PATH

```
#cd PhylOligo

export PATH=‘pwd‘/src/:$PATH
chmod +x src/{*.py,*.R}
```

List of Dependencies:

- Python 3.x
  - BioPython biopython.org
  - sklearn http://scikit-learn.org/stable/install.html
  - Numpy numpy.org
  - matplotlib http://matplotlib.org
  - hdbscan https://pypi.python.org/pypi/hdbscan
  - Cython http://cython.org
  - h5py http://www.h5py.org
- R 3.x

  - ape http://ape-package.ird.fr
  - gplots https://cran.r-project.org/web/packages/gplots/index.html
  - getopt https://cran.r-project.org/web/packages/getopt/getopt.pdf

- EMBOSS http://emboss.sourceforge.net/download
- Samtools http://www.htslib.org/
- X11 onlyrequiredtorunphyloselect.R

3 Software and options

3.1 phyloligo.py

Generate the all-by-all contig distance matrix

- Load and index the genome assembly sequences.
- Compute the kmer/spaced-pattern composition profile of each sequence in the assembly.
- Compute a pairwise distance matrix for all sequences.

```
phyloligo.py -d JSD -i genome.fasta -o genome.JSD.mat -u 64
```

Parameters:

```
-h, --help            show this help message and exit
-i GENOME, --assembly GENOME
                      multifasta of the genome assembly
-k PATTERN, --lgMot PATTERN
                      word lenght / kmer length / k [default:4]. This option
                      is an alias for --pattern (see -p). If the type of the
                      parameter is an integer, it will be interpreted as the
                      lenght of the kmer to use. If the type of the
                      parameter is a string, it will be interpreted as a
                      spaced-pattern.
-s {both,plus,minus}, --strand {both,plus,minus}
                      strand used to compute microcomposition.
                      [default:both]
-d {Eucl,JSD}, --distance {Eucl,JSD}
                      how to compute distance between two signatures : Eucl
                      : Euclidean[default:Eucl], JSD : Jensen-Shannon
                      divergence
--freq-chunk-size FREQCHUNKSIZE
                      the size of the chunk to use in scoop to compute
                      frequencies
--dist-chunk-size DISTCHUNKSIZE
                      the size of the chunk to use in scoop to compute
                      distances
--method {scoop,joblib}
                      don't use scoop to compute distances use joblib
--large {None,memmap,h5py}
                      used in combination with joblib for large dataset
-c THREADS_MAX, --cpu THREADS_MAX
                      how many threads to use for windows microcomposition
                      computation[default:4]
-o OUT_FILE, --out OUT_FILE
                      output file[default:phyloligo.out]
-w WORKDIR, --workdir WORKDIR
                      working directory
-p PATTERN, --pattern PATTERN
                      spaced-word pattern string, only containing 1s and 0s,
                      i.e. '100101001', default='1111'. See -k / --lgMot.
```

3.2 phyloselect.R

Regroup contigs by compositional similarity on a tree and explore branching

- Load the distance matrix produced by PhylOligo.
- Optionally create a hierarchically sorted distance matrix.
- Build a cladogram from the distance matrix.
- Interactively ask the user to explore the cladogram and select clads that might correspond to untargeted sequences based on the interpretation of the topology.
- Export clad-specific fasta files:
  - To inspect their potential origin for example with blast or GOHTAM (Ménigaud *et al.*, 2012)
  - To use as learning material in ContaLocate

```
phyloselect.R -d -m -c 0.95 -s 4000 -t BIONJ -f c -w 20  -i genome.JSD.mat -a
    genome.fasta -o genome_conta
```

Parameters:

```
  -i|--matrix
                  All-by-all contig distance matrix, tab separated (required)
  -a|--assembly
                  Multifasta file of the contigs (required)
  -f|--tree_draw_method
                  Tree building type. [phylogram, cladogram,
                  fan, unrooted, radial] by default cladogram.
  -t|--tree_building_method
                  Tree drawing type [NJ, UPGMA, BIONJ, wardD,
                  wardD2, Hsingle, Hcomplete, WPGMA, WPGMC, UPGMC] by default NJ.
  -m|--matrix_heatmap
                  Should a matrix heatmap should be produced
  -c|--distance_clip_percentile
                  Threshold to exclude very distant contigs based on the distance
                  distribution. Use if the tree is squashed by repeats or
                  degenerated/uninformative contigs [0.97]
  -s|--contig_min_size
                  Min length in bp of contigs to use in the matrix and tree.
                  Use if the tree is squashed by repeats or
                  degenerated/uninformative contigs [4000]
  -d|--dump_R_session
                  Should the R environment be saved for later exploration?
                  The filename will be generated from the outfile parameter
                  or its default value
  -g|--max_perc
                  Max edge assembly length percentage displayed (%)
  -l|--min_perc
                  Min edge assembly length percentage displayed (%)
  -k|--keep_perc
                  Ratio of out-of-range percentages to display (%)
  -o|--outfile
                  Outfile name, default:phyloligo.out
  -b|--branchlength
                  Display branch length
  -w|--branchwidth
                  Branch width factor [40]
  -v|--verbose
                  Says what the program do.
  -h|--help
                  This help.
```

note: PhyloSelect uses the library Ape and its interactive clade selection function on a tree plot with the mouse. X11 is therefore required. If the program has to run on a server -typically for memory reasons- please use the -X option of ssh to allow X11 forwarding.

3.3 phyloselect.py

Regroup contigs by compositional similarity: hierarchical DBSCAN or K-medoids clustering and multidimensional scaling display with t-SNE.

- Load the distance matrix produced by PhylOligo.
- Cluster the sequences
- Export cluster-specific fasta files:
  - To inspect their potential origin for example with blast or GOHTAM (Ménigaud *et al.*, 2012)
  - To use as learning material in ContaLocate

```
phyloselect.py -i genome.JSD.mat -t -m hdbscan --noX -o genome_conta
```

Parameters:

```
-h, --help               show this help message and exit
-i DISTMAT               The input matrix file
-t                       Perform tsne for visualization and pre-clustering
-p PERPLEXITY            Change the perplexity value
-m {hdbscan,kmedoids}
                         Method to use to compute cluster on transformed
                         distance matrix
--minclustersize MIN_CLUSTER_SIZE
                         Set the minimal cluster size of an HDBSCAN cluster
--minsamples MIN_SAMPLES
                         Set the minimal sample size of an HDBSCAN cluster
-k NBK                   Number of cluster
-f FASTAFILE             Path of the original fasta file used for the
                         computation of the distance matrix
--interactive            Allow the user to run the script in an interactive
                         mode and change clustering parameter on the fly
                         (require -t)
--large {memmap,h5py}
                         Used in combination with joblib for large dataset
--noX                    Instead of showing pictures, store them in png
-o OUTPUTDIR
```

3.4 contalocate.R

Extract DNA segments with homogeneous oligonucleotide composition from a genome assembly. Once you have explored your assembly's oligonucleotide composition, identified and selected -potentially partial- untargeted material, use ContaLocate to target species-specific DNA according to a double parametrical threshold.

- Learn a compositional profile for the host and the untargeted organism, previously identified with phyloligo.py / phyloselect.py,R.
- Scan the assembly for regions similar in composition to the two aforementioned profiles.
- Compute one threshold value for each scan based on the distribution of the metric.
- Locate the untargeted regions according to the 2 thresholds, distant from the host and close the the untargeted profile.
- Generate a GFF3 map of the untargeted region positions in the genome.

If both the host and untargeted learning material are avaialble:

```
contalocate.R -i genome.fasta -r genome_host.fa -c genome_conta_1.fa
```

The training set for the host genome can be omitted if the amount of untargeted sequences is negligible/very small. In this case, the profile of the host will be trained on the whole genome, including the untargeted sequences which might create a bias proportional to the relative amount of untargeted material.

```
contalocate.R -i genome.fasta -c genome_conta_1.fa
```

The set up of the thresholds can be manually enforced. The user will interactively prompted to set the thresholds given the distribution of windows divergence.

```
contalocate.R -i genome.fasta -c genome_conta_1.fa -m
```

Parameters:

```
  -i|--genome
                          Multifasta of the genome assembly (required)
  -r|--host_learn
                          Host training set (optional)
  -c|--conta_learn
                          Contaminant training set (optional) if missing and
                          sliding window parameters are given, the sliding
                          windows composition will be compared to the whole
                          genome composition to contrast potential HGTs
                          (prokaryotes and simple eukaryotes only)
  -t|--win_step
                          Step of the sliding windows analysis to locate the
                          contaminant (optional) default: 500bp or 100bp
  -w|--win_size
                          Length of the sliding window to locate the
                          contaminant (optional) default: 5000bp
  -W|--outputdir
                          path to outputdir directory
  -d|--dist
                          Divergence metric used to compare profiles: (KL), JSD or Eucl
  -m|--manual_threshold
                          You will be asked to manually set the thresholds
  -h|--help
                          This help
```

3.5 phylopreprocess.py

Preprocess the original assembly/raw reads in order to filter out entries, reduce computational time and increase signal. Filter short sequences or highly conserved repeats.

- Reads an assembly or long sequencing reads multi-fasta file
- Output filtered dataset

```
phylopreprocess.py [-h] -i INPUTFASTA [-p PERCENTILE] [-m MIN_READSIZE] [-s
    SAMPLING] [-r] [-o OUTPUTFASTA]
```

Parameters:

```
-h, --help          show this help message and exit
-i INPUTFASTA
-p PERCENTILE       remove read of size not in Xth percentile
-m MIN_READSIZE     remove reads shorter than the provided minimal size
-s SAMPLING         percentage of read to sample
-r                  the order of the reads are randomized
-o OUTPUTFASTA
```

4 Pipeline examples

4.1 Workstation

```
assembly=/path/to/assembly.fa
cpus=64
name="organism"
pattern=4
distance="JSD"
work_dir=`pwd`



phyloligo.py -c 24 -o ${name}_${distance}_k${pattern}.mat -i $assembly -k
    $pattern -d ${distance}

phyloselect.R -i ${name}_${distance}_k${pattern}.mat -a $assembly -d -w 20 -c
    0.90 -s 4000 -m -f c -t BIONJ -o PhyloSelect_${name}



*TODO* add contalocalte
```

4.2 SGE grid - SMP

```
#!/bin/bash

assembly=/path/to/assembly.fa
cpus=64
name="organism"
pattern="1111"
distance="JSD"
work_dir=`pwd`


#$ -S /bin/bash
#$ -cwd
#$ -V
#$ -pe parallel_smp $cpu
#$ -l mem=1G
#$ -l h_vmem=1G
#$ -N PhylOligo_grid_test_$name



echo "phyloligo.py -c \$NSLOTS -o ${name}_${distance}_k${pattern}.mat -i
    $assembly --pattern $pattern -d ${distance} --method joblib --large h5py" |
    qsub -N PhylOligo_${name}_${distance}_k${pattern} -l mem=12G -l h_vmem=64G


echo "phyloselect.py -i ${name}_${distance}_k${pattern}.mat -t -m hdbscan --
    large h5py --noX -o $work_dir"| qsub -N PhyloSelect_${name} -l mem=10G -l
    h_vmem=30G -hold_jid PhylOligo_${name}_${distance}_k${pattern}
```

4.3 SGE grid - Multi node

```
#!/bin/bash

assembly=/path/to/assembly.fa
cpus=64
name="organism"
pattern="1111"
distance="JSD"
work_dir=`pwd`


#$ -S /bin/bash
#$ -cwd
#$ -V
#$ -pe parallel_smp $cpu
#$ -l mem=1G
#$ -l h_vmem=1G
#$ -N PhylOligo_grid_test_$name
```

```
#SSH connexion between nodes must be allowed for scoop to work properly

echo "phyloligo.py -c \$NSLOTS -o ${name}_${distance}_k${pattern}.mat -i
    $assembly --pattern $pattern -d ${distance} --method scoop --freq-chunk-size
     3000 --dist-chunk-size 500" | qsub -N PhylOligo_${name}_${distance}_k${
    pattern} -l mem=12G -l h_vmem=64G


echo "phyloselect.py -i ${name}_${distance}_k${pattern}.mat -t -m hdbscan --noX
    -o $work_dir"| qsub -N PhyloSelect_${name} -l mem=10G -l h_vmem=30G -
    hold_jid PhylOligo_${name}_${distance}_k${pattern}
```

5 Examples

5.1 *Magnaporthe oryzae*

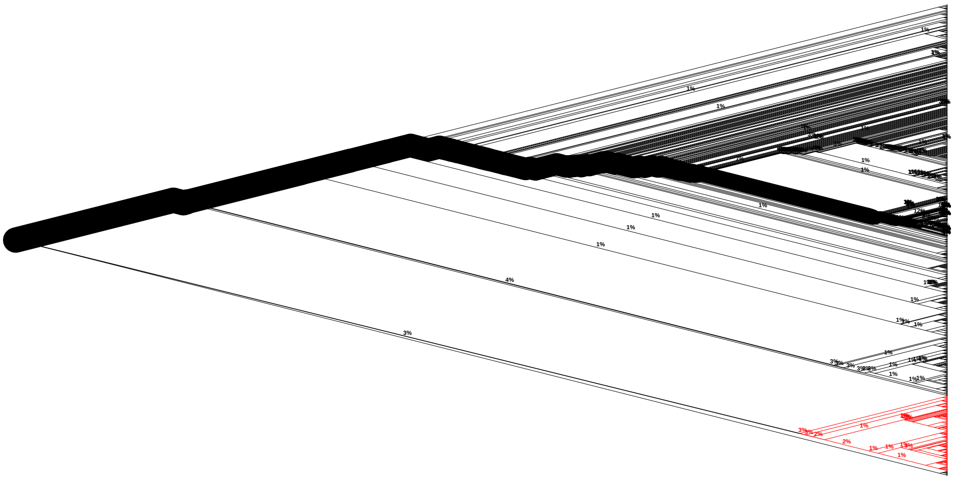


Figure 2: Interactive exploration of the *Magnaporthe oryzae* TH12 assembly (Chiapello *et al.*, 2015). This slection will be called “Clade A”, the user suspect this is an untargeted set of sequences.

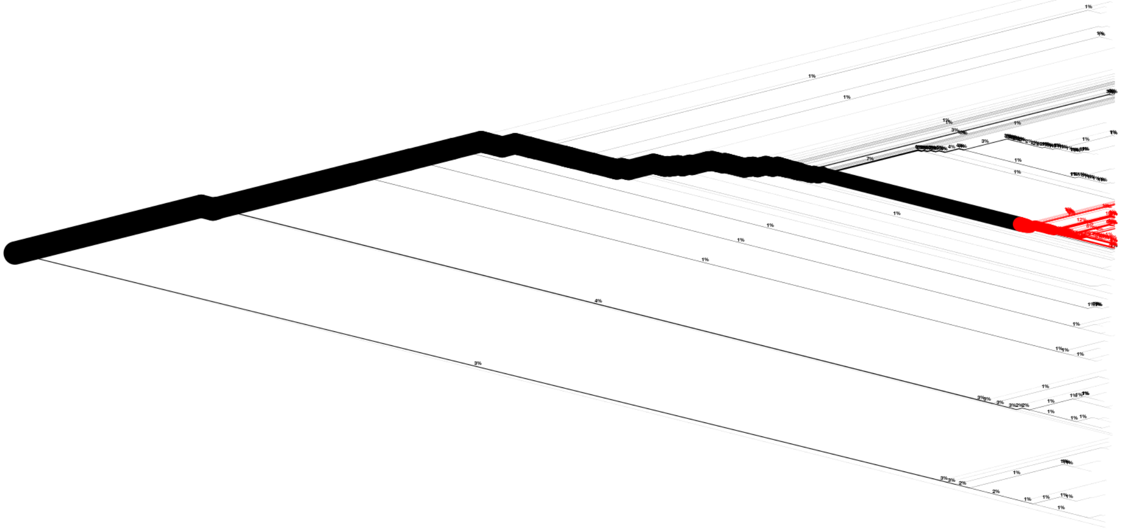


Figure 3: Interactive exploration of the *Magnaporthe oryzae* TH12 assembly (Chiapello *et al.*, 2015). This slection will be called “Clade B”, the user suspect this correspond to the host/targeted sequences .

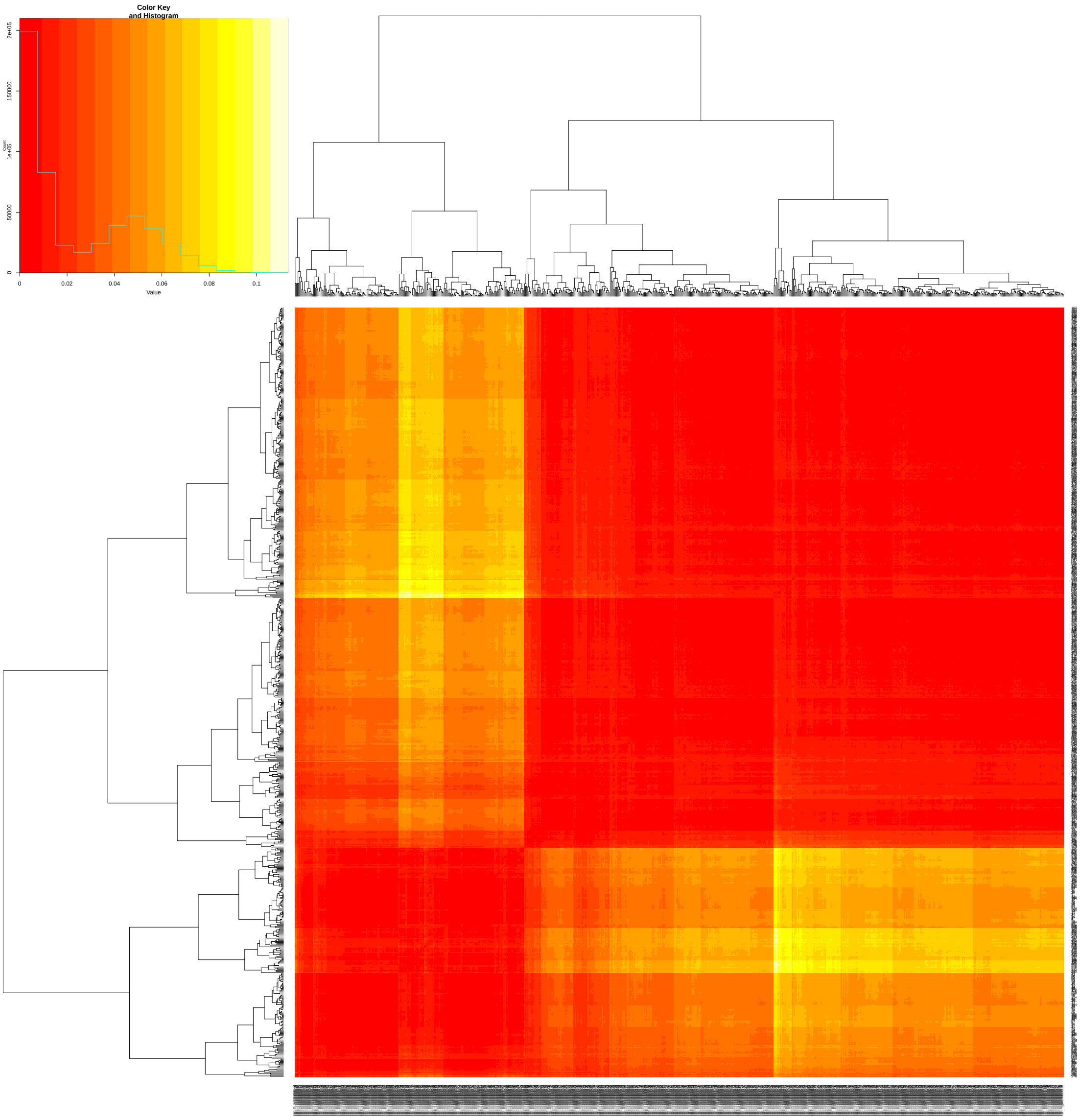
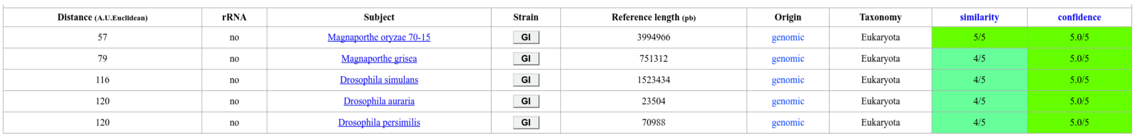
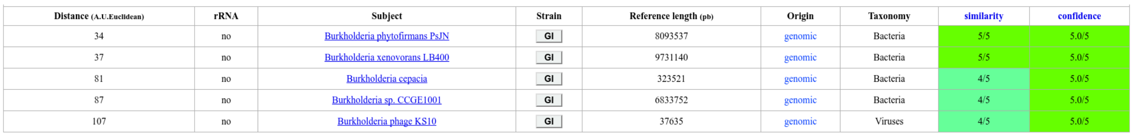


Figure 4: Sorted distance matrix of contigs of *Magnaporthe oryzae* TH12 assembly (Chiapello *et al.*, 2015). The following parameters in phyloselect.R were used: -d -w 20 -c 0.97 -m



Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
57	no	Magnaporthe oryzae 70-15	GI	3994966	genomic	Eukaryota	5/5	5.0/5
79	no	Magnaporthe grisea	GI	751312	genomic	Eukaryota	4/5	5.0/5
116	no	Drosophila simulans	GI	1523434	genomic	Eukaryota	4/5	5.0/5
120	no	Drosophila auraria	GI	23504	genomic	Eukaryota	4/5	5.0/5
120	no	Drosophila persimilis	GI	70988	genomic	Eukaryota	4/5	5.0/5

Figure 5: Identification of clade A (see Figure 2) with GOHTAM.



Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
34	no	Burkholderia phytofirmans PsJN	GI	8093537	genomic	Bacteria	5/5	5.0/5
37	no	Burkholderia xenovorans LB400	GI	9731140	genomic	Bacteria	5/5	5.0/5
81	no	Burkholderia cepacia	GI	323521	genomic	Bacteria	4/5	5.0/5
87	no	Burkholderia sp. CCGE1001	GI	6833752	genomic	Bacteria	4/5	5.0/5
107	no	Burkholderia phage KS10	GI	37635	genomic	Viruses	4/5	5.0/5

Figure 6: Identification of clade B (see Figure 3) with GOHTAM.

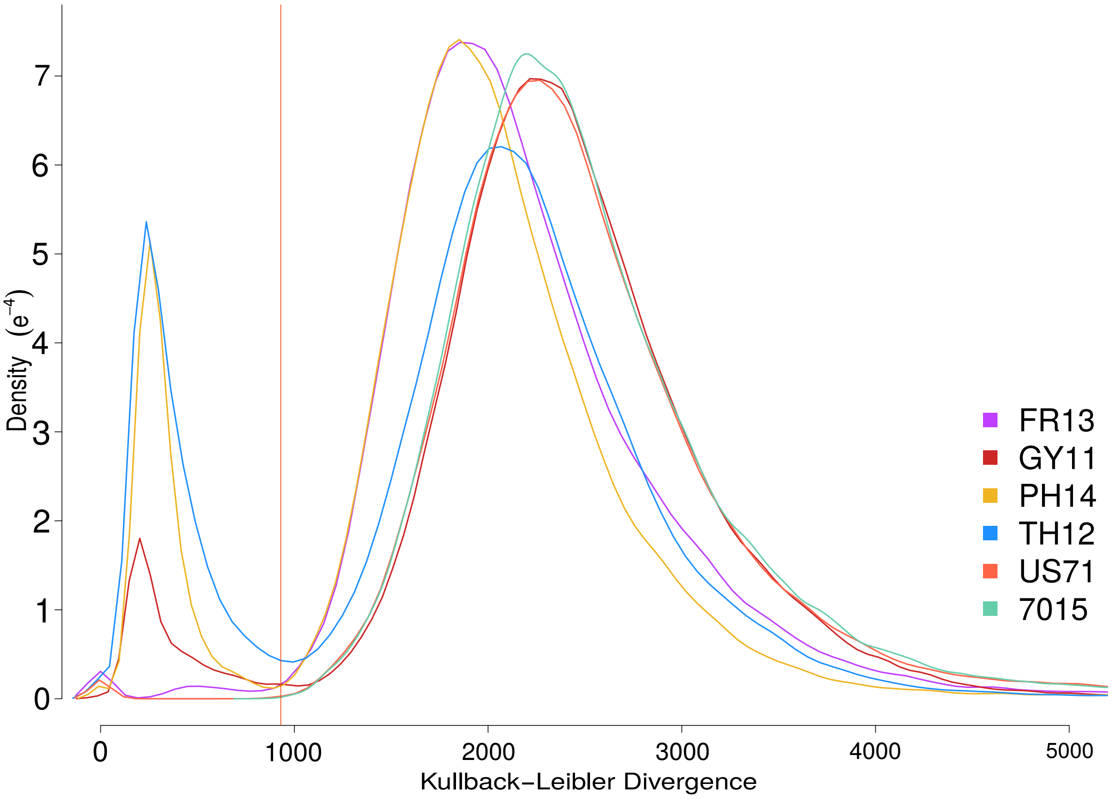


Figure 7: Distribution of distances between the composition profile of clade A (see Figure 2) and the scanning windows over the whole assembly. The untargeted threshold is the vertical red line.

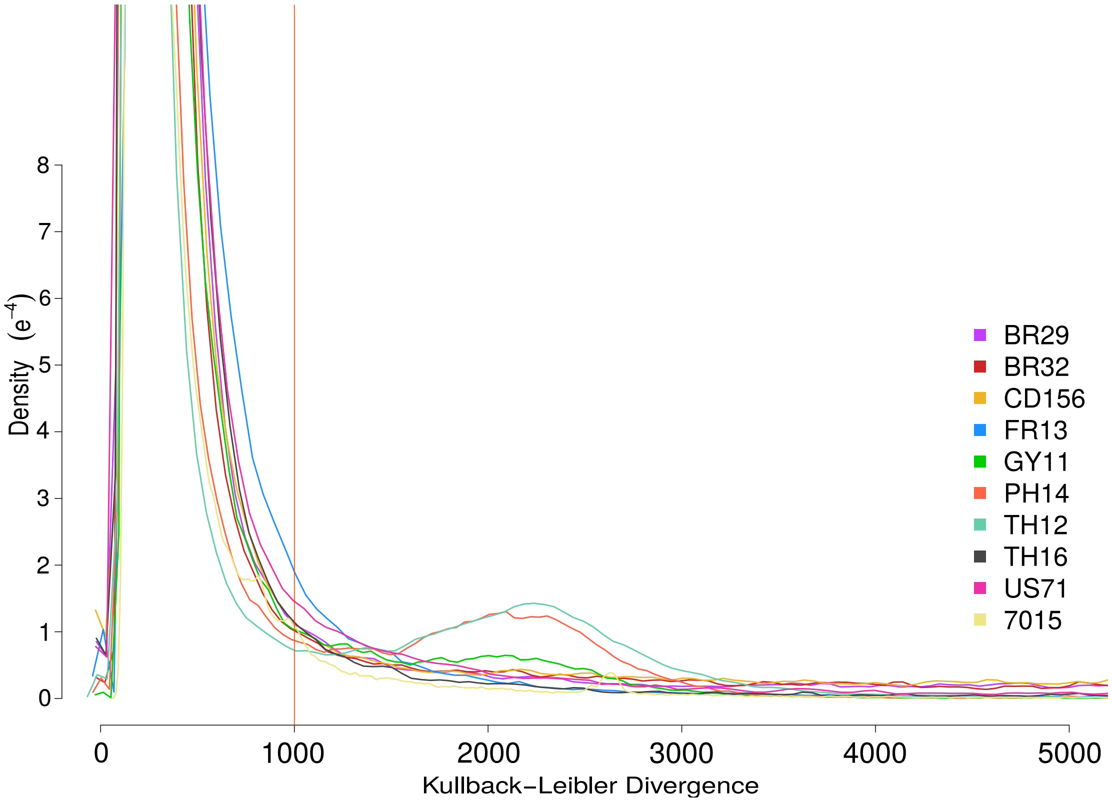


Figure 8: Distribution of distances between the composition profile of clade B (see Figure 3) and the scanning windows over the whole assembly. The host threshold is the vertical red line.

5.2 *Aeschynomene evenia*

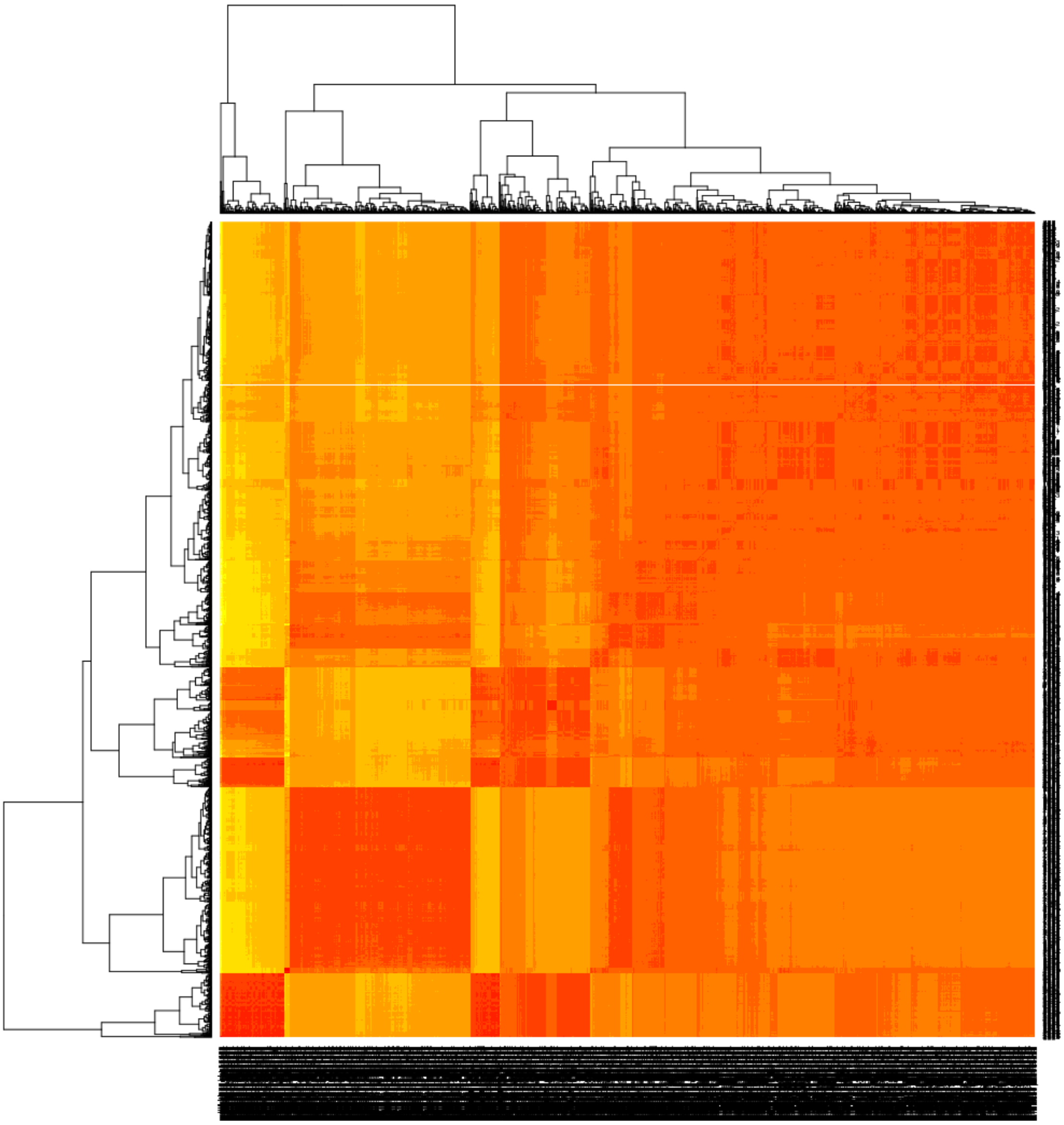


Figure 9: Sorted distance matrix of contigs of *Aeschynomene evenia* (unpublished genome).

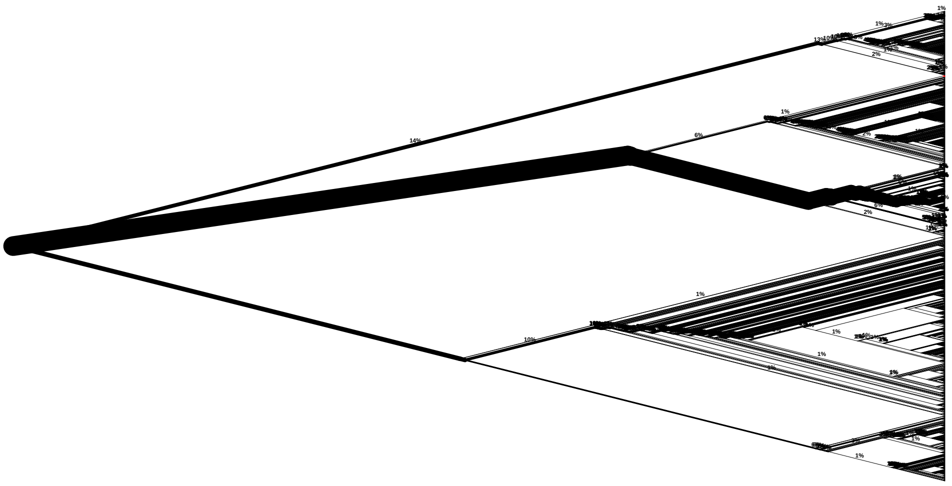


Figure 10: Interactive exploration of the *Aeschynomene evenia*.

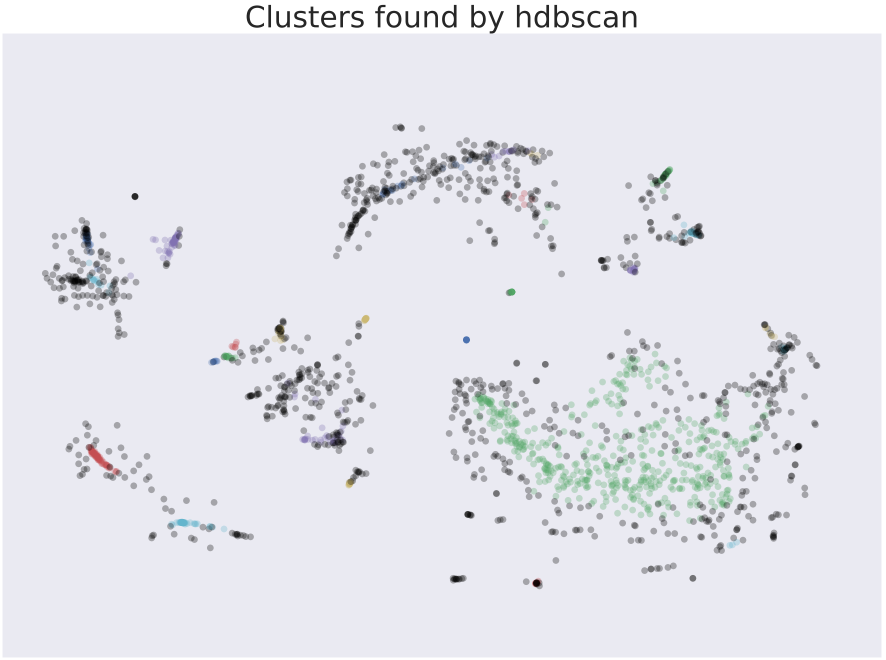


Figure 11: Automated clustering and dimensional reduction of the *Aeschynomene evenia* assembly. Phyloselect.py with default parameters.

5.3 *Ganoderma lucidum*

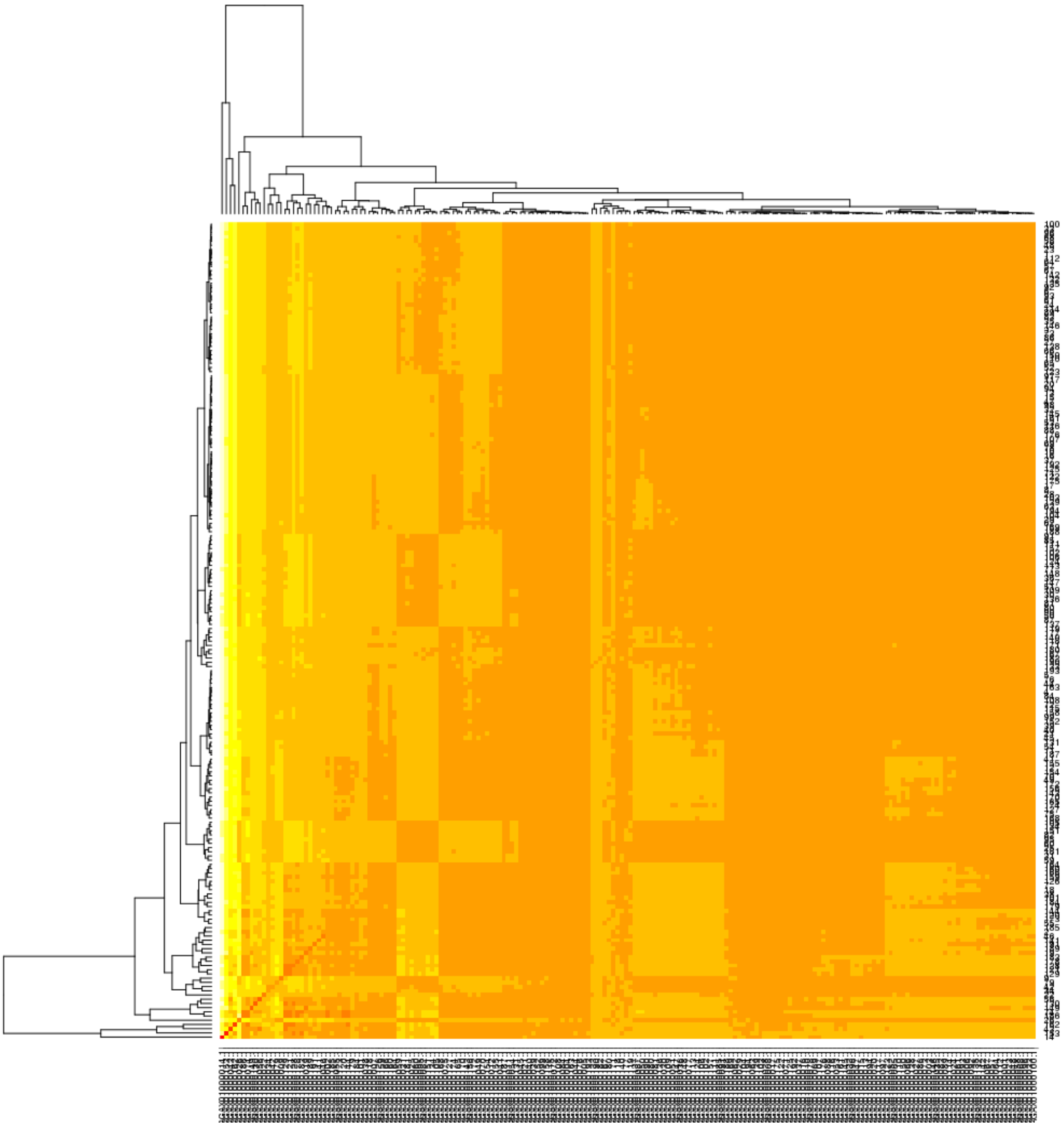


Figure 12: Sorted distance matrix of contigs of *Ganoderma lucidum* (Chen *et al.*, 2012).

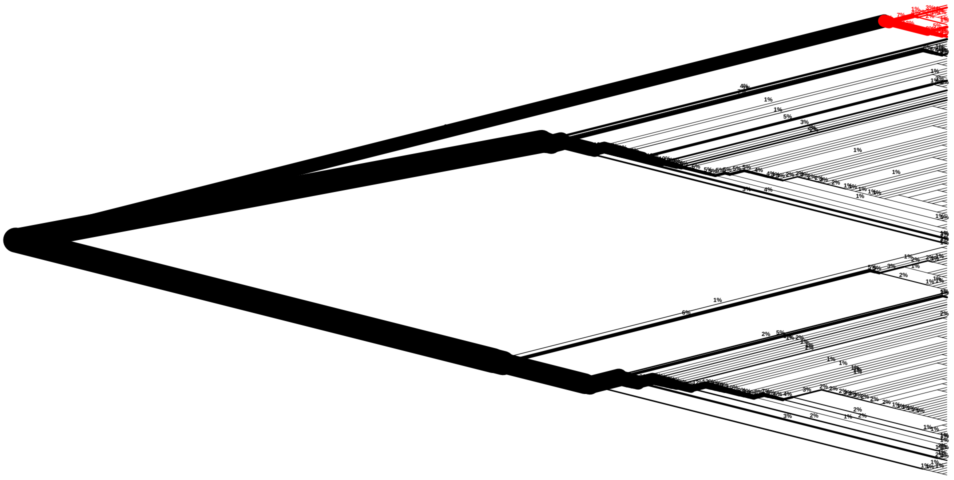
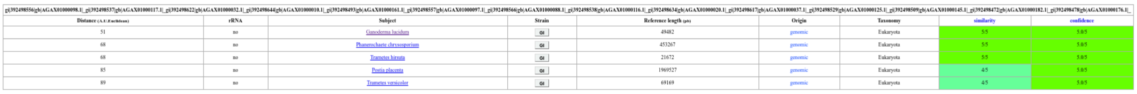


Figure 13: Interactive exploration of the *Ganoderma lucidum* assembly (Chen *et al.*, 2012).



gi|392498556|gb|AGAX01000098.1|_gi|392498537|gb|AGAX01000117.1|_gi|392498622|gb|AGAX01000032.1|_gi|392498644|gb|AGAX01000010.1|_gi|392498493|gb|AGAX01000161.1|_gi|392498557|gb|AGAX01000097.1|_gi|392498566|gb|AGAX01000088.1|_gi|392498538|gb|AGAX01000116.1|_gi|392498634|gb|AGAX01000020.1|_gi|392498617|gb|AGAX01000037.1|_gi|392498529|gb|AGAX01000125.1|_gi|392498509|gb|AGAX01000145.1|_gi|392498472|gb|AGAX01000182.1|_gi|392498478|gb|AGAX01000176.1|_

Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
51	no	Ganoderma lucidum	GI	49482	genomic	Eukaryota	5/5	5.0/5
68	no	Phanerochaete chrysosporium	GI	453267	genomic	Eukaryota	5/5	5.0/5
68	no	Trametes hirsuta	GI	21672	genomic	Eukaryota	5/5	5.0/5
85	no	Postia placenta	GI	1969527	genomic	Eukaryota	4/5	5.0/5
89	no	Trametes versicolor	GI	69169	genomic	Eukaryota	4/5	5.0/5

Figure 14: Identification of the red clade from Figure 13 with GOHTAM.

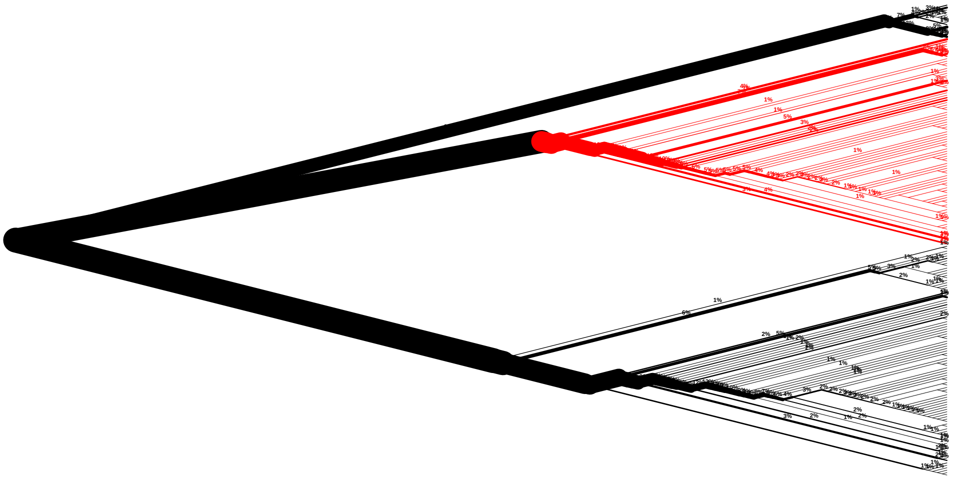
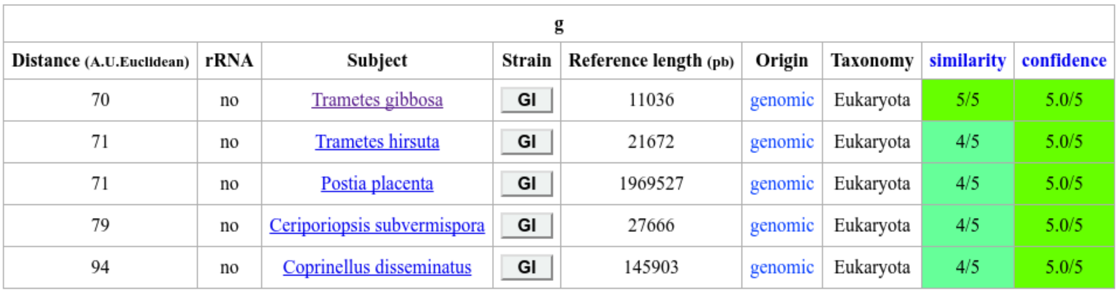


Figure 15: Interactive exploration of the *Ganoderma lucidum* assembly (Chen *et al.*, 2012).



g								
Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
70	no	Trametes gibbosa	GI	11036	genomic	Eukaryota	5/5	5.0/5
71	no	Trametes hirsuta	GI	21672	genomic	Eukaryota	4/5	5.0/5
71	no	Postia placenta	GI	1969527	genomic	Eukaryota	4/5	5.0/5
79	no	Ceriporiopsis subvermispora	GI	27666	genomic	Eukaryota	4/5	5.0/5
94	no	Coprinellus disseminatus	GI	145903	genomic	Eukaryota	4/5	5.0/5

Figure 16: Identification of the red clade from Figure 15 with GOHTAM.

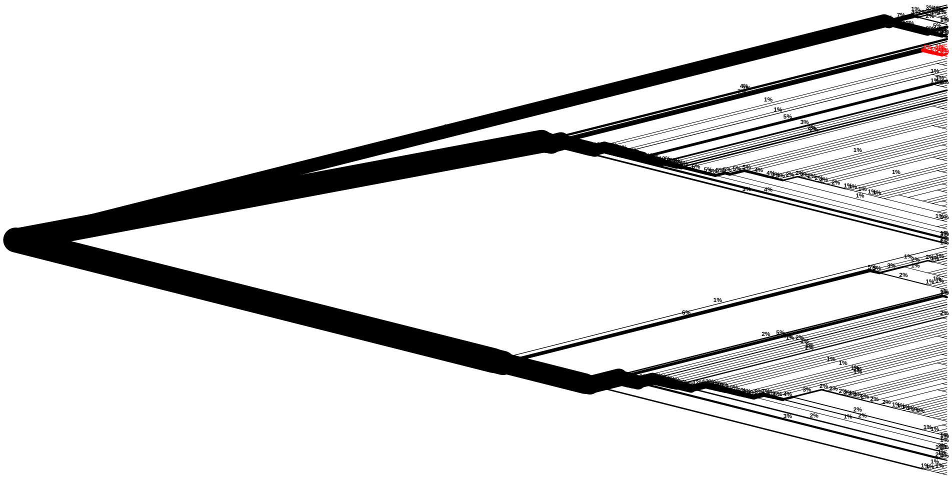
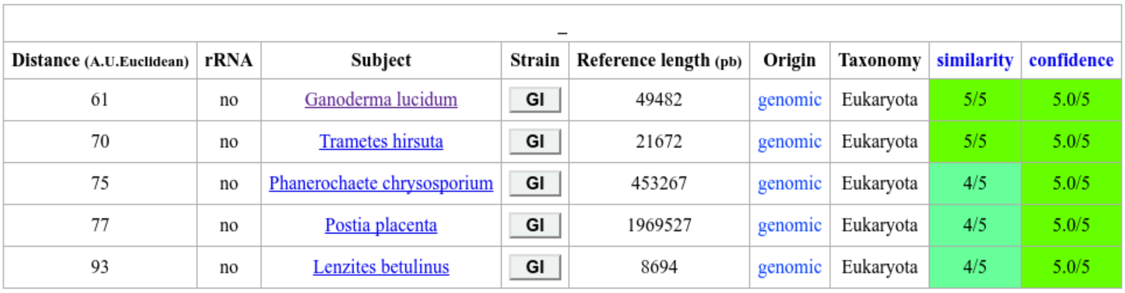


Figure 17: Interactive exploration of the *Ganoderma lucidum* assembly (Chen *et al.*, 2012).



_								
Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
61	no	Ganoderma lucidum	GI	49482	genomic	Eukaryota	5/5	5.0/5
70	no	Trametes hirsuta	GI	21672	genomic	Eukaryota	5/5	5.0/5
75	no	Phanerochaete chrysosporium	GI	453267	genomic	Eukaryota	4/5	5.0/5
77	no	Postia placenta	GI	1969527	genomic	Eukaryota	4/5	5.0/5
93	no	Lenzites betulinus	GI	8694	genomic	Eukaryota	4/5	5.0/5

Figure 18: Identification of the red clade from Figure 17 with GOHTAM.

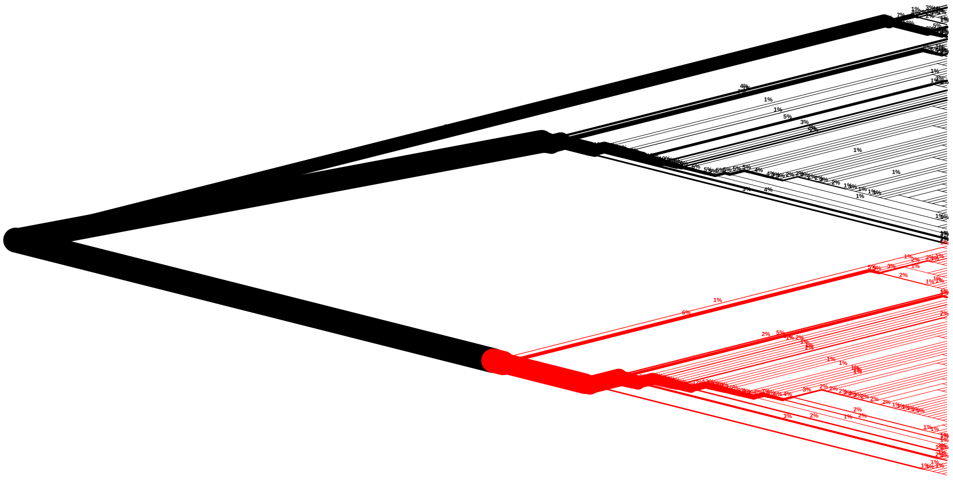
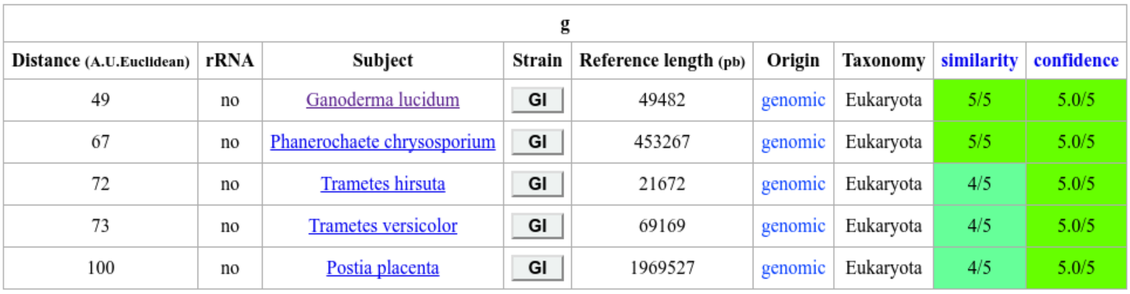


Figure 19: Interactive exploration of the *Ganoderma lucidum* assembly (Chen *et al.*, 2012).



g								
Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
49	no	Ganoderma lucidum	GI	49482	genomic	Eukaryota	5/5	5.0/5
67	no	Phanerochaete chrysosporium	GI	453267	genomic	Eukaryota	5/5	5.0/5
72	no	Trametes hirsuta	GI	21672	genomic	Eukaryota	4/5	5.0/5
73	no	Trametes versicolor	GI	69169	genomic	Eukaryota	4/5	5.0/5
100	no	Postia placenta	GI	1969527	genomic	Eukaryota	4/5	5.0/5

Figure 20: Identification of the red clade from Figure 19 with GOHTAM.

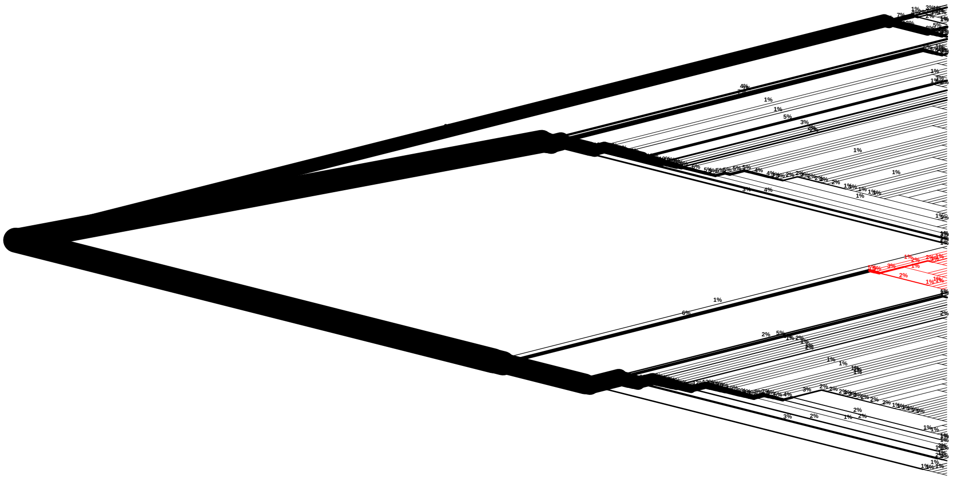
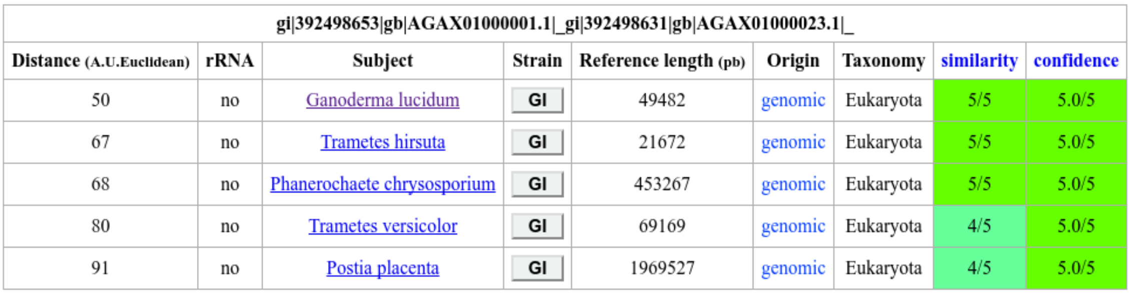


Figure 21: Interactive exploration of the *Ganoderma lucidum* assembly (Chen *et al.*, 2012).



gi|392498653|gb|AGAX01000001.1|_gi|392498631|gb|AGAX01000023.1|_								
Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
50	no	Ganoderma lucidum	GI	49482	genomic	Eukaryota	5/5	5.0/5
67	no	Trametes hirsuta	GI	21672	genomic	Eukaryota	5/5	5.0/5
68	no	Phanerochaete chrysosporium	GI	453267	genomic	Eukaryota	5/5	5.0/5
80	no	Trametes versicolor	GI	69169	genomic	Eukaryota	4/5	5.0/5
91	no	Postia placenta	GI	1969527	genomic	Eukaryota	4/5	5.0/5

Figure 22: Identification of the red clade from Figure 21 with GOHTAM.

5.4 *Hypsibius dujardini*

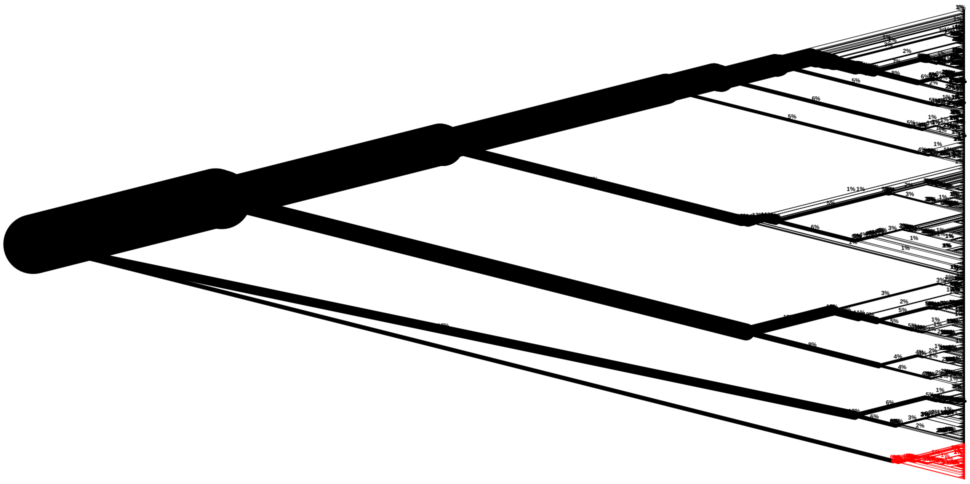


Figure 23: Interactive exploration of the *Hypsibius dujardini* assembly (Delmont and Eren, 2016).

6 Bibliography

Chen, S., Xu, J., Liu, C., Zhu, Y., Nelson, D. R., Zhou, S., Li, C., Wang, L., Guo, X., Sun, Y., Luo, H., Li, Y., Song, J., Henrissat, B., Levasseur, A., Qian, J., Li, J., Luo, X., Shi, L., He, L., Xiang, L., Xu, X., Niu, Y., Li, Q., Han, M. V., Yan, H., Zhang, J., Chen, H., Lv, A., Wang, Z., Liu, M., Schwartz, D. C., and Sun, C. (2012). Genome sequence of the model medicinal mushroom ganoderma lucidum. *Nature communications*, **3**, 913.

Chiapello, H., Mallet, L., Guérin, C., Aguileta, G., Amselem, J., Kroj, T., Ortega-Abboud, E., Lebrun, M.-H., Henrissat, B., Gendrault, A., Rodolphe, F., Tharreau, D., and Fournier, E. (2015). Deciphering genome content and evolutionary relationships of isolates from the fungus Magnaporthe oryzae attacking different host plants. *Genome Biology and Evolution*, **7**(10), 2896–2912.

Delmont, T. O. and Eren, A. M. (2016). Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, **4**, e1839.

Ménigaud, S., Mallet, L., Picord, G., Churlaud, C., Borrel, A., and Deschavanne, P. (2012). Gohtam: a website for 'genomic origin of horizontal transfers, alignment and metagenomics'. *Bioinformatics*, **28**(9), 1270–1271.