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 19, 2017

Contents

1 Workflow **1**

2 Installation **2**
- 2.1 Quick Install . . . 2
- 2.2 Alternative install tricks . . . 3

3 Software manual and options **4**
- 3.1 phylopreprocess.py . . . 4
- 3.2 phyloligo.py . . . 4
- 3.3 phyloselect.R . . . 5
- 3.4 phyloselect.py . . . 6
- 3.5 contalocate.R . . . 7

4 Pipeline examples **8**
- 4.1 Workstation . . . 8
- 4.2 SGE grid - SMP . . . 8
- 4.3 SGE grid - Multi node . . . 8
- 4.4 SGE grid - Very large dataset . . . 9

5 Examples **9**
- 5.1 *Magnaporthe oryzae* . . . 9
- 5.2 *Hypsibius dujardini* . . . 14

6 Bibliography **14**

1 Workflow

PhylOligo is a package of tools to analyse the heterogeneity of oligonucleotide composition of genomic assembly fragments to explore and locate sequences from potential untargeted organisms. The package contains several programs arranged in a workflow.

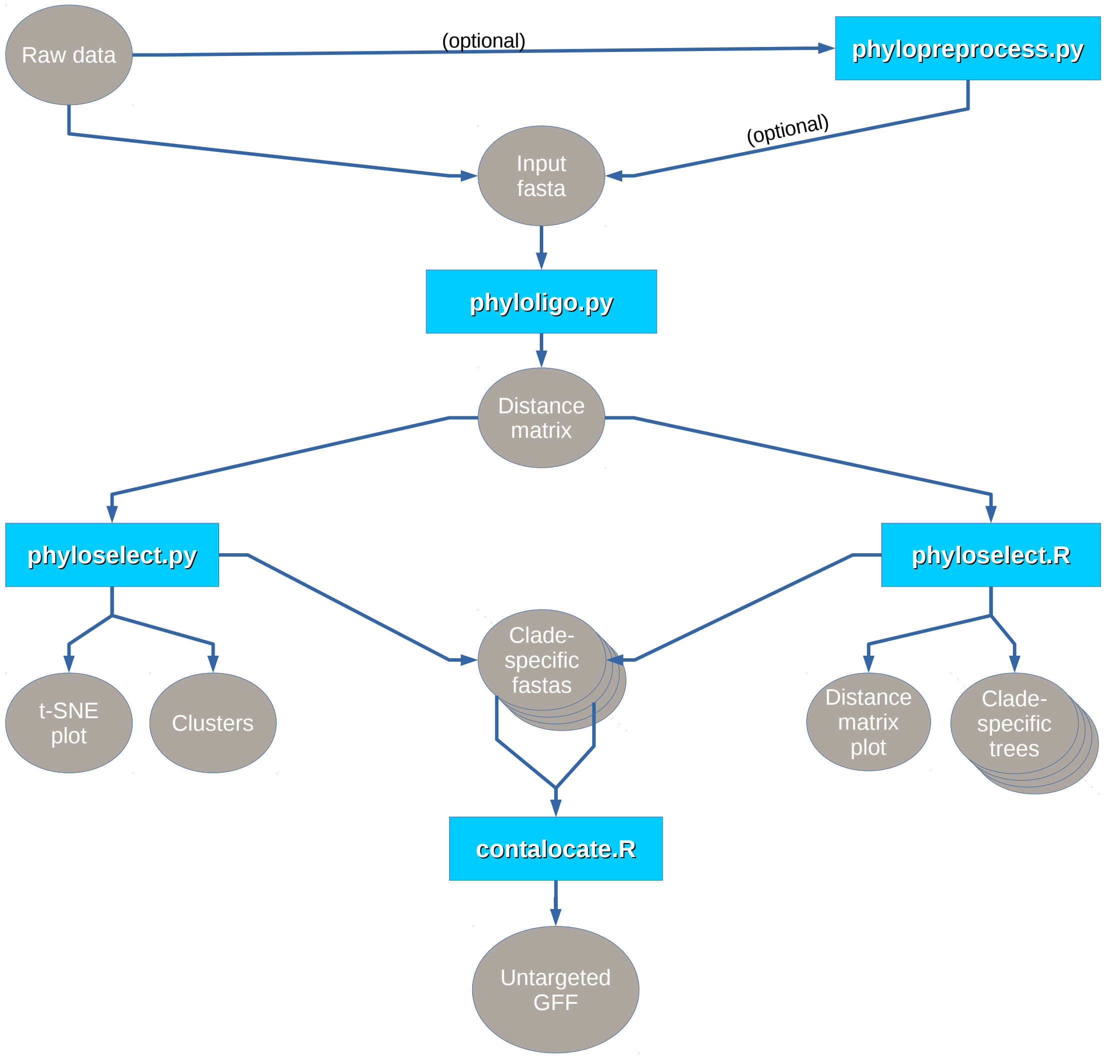

Figure 1: Workflow of PhylOligo. Blue frames: programs and scripts. Grey blobs: data files and output files.

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 manual and options

3.1 phylopreprocess.py

Pre-process the original contigs/scaffolds/long reads in order to filter out entries, reduce computational time and increase signal. Filter short sequences or highly conserved repeats. Sub-sampling can be used in order to perform quick tests or to reduce the size of a dataset to allow for its computation given the computational resources available. Note that this step is optional and that `phyloselect.R` also contains sequence filters in order to test out different values without having to recompute the frequencies and the distance matrix with `phyloligo.py`. Sequences shorter than 1kb should be considered as poorly informative or representative of their species compositional profile. In order to grant a more refined selection of materials to establish an accurate compositional profile prototype and the detection of potential untargeted sequences, sequences below about 5Kb could be filtered if it can be hypothesized that a possible contaminant would not have shorter sequences or be completely filtered out.

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

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

Parameters:

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

3.2 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.
```

3.3 phyloselect.R

Regroup contigs by compositional similarity on a tree and explore the topology.

- 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.

3.4 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
```

3.5 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 genome 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.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
```

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 $cpus -o ${name}_${distance}_k${pattern}.mat -i $assembly -k
    $pattern -d ${distance} --method joblib

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



# filenames depends on the selection made by the user.
contalocate.R -i genome.fasta -r PhyloSelect_${name}_1.fa -c PhyloSelect_${name}
    _2.fa
```

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}
```

4.4 SGE grid - Very large dataset

```
#!/bin/bash

assembly=/path/to/assembly.fa
cpus=240
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 "phylopreprocess.py -i $assembly -m 4000 -o ${assembly}_filtered_m4000.fa"
    | qsub -N PhylOligo_${name}_${distance}_k${pattern} -l mem=12G -l h_vmem=64G

echo "phyloligo.py -c \$NSLOTS -o ${name}_${distance}_k${pattern}.mat -i ${
    assembly}_filtered_m4000.fa --pattern $pattern -d ${distance} --method
    joblib --large h5py" | qsub -N PhylOligo_${name}_${distance}_k${pattern} -l
    mem=48G -l h_vmem=100G


echo "phyloselect.py -i ${name}_${distance}_k${pattern}.mat -t -m hdbscan --
    large h5py --noX -o $work_dir"| qsub -N PhyloSelect_${name} -l mem=800G -l
    h_vmem=3000G -hold_jid PhylOligo_${name}_${distance}_k${pattern}
```

5 Examples

5.1 *Magnaporthe oryzae*

This example shows how a bacterial genome was identified in the assembly of the phytopathgenic fungus *Magnaporthe oryzae*. Nine isolates were sequenced (Chiapello *et al.*, 2015) of which four exhibited an unexpectedly larger genome size compared to other genomes of the same species.

Investigations with PhylOligo and comparison of the different isolates (Figure 2, Figure 6) revealed the presence of contigs with distinct oligonucleotide composition as seen on Figure 3. These regions were used to learn a prototype of this composition, determine divergence thresholds (see Figure 7 and Figure 8) and the whole genome was scanned with ContaLocate. A whole bacterial genome was identified, as well as several chimeric scaffolds, *i.e.* containing DNA from 2 organisms. With GOHTAM (Ménigaud *et al.*, 2012), the bacterial genome was identified to be unsequenced at the time and compositionally close to *Burkholderia phytofirmans* and *Burkholderia xenovorans*.

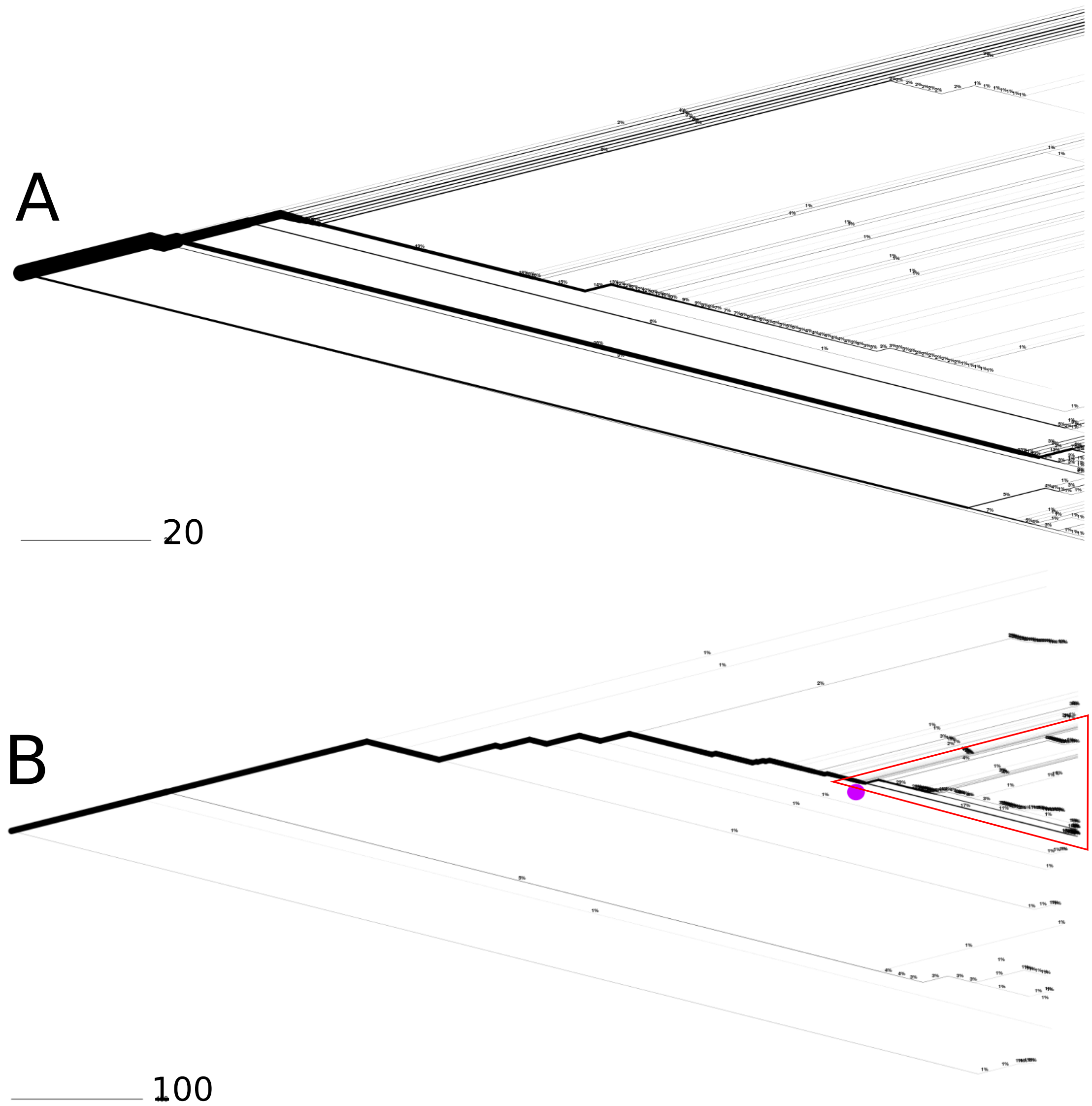

Figure 2: Exploration of 2 isolates of *Magnaporthe oryzae* (Chiapello *et al.*, 2015). **A**: Exploration of *Magnaporthe oryzae* TH16. This isolate was found to contain no untargeted material. **B**: Exploration of *Magnaporthe oryzae* TH12. The topology and width pattern of the subtree identified in the red triangle is very similar to the whole tree in Figure 2 A. The user suspects that this conserved pattern accounts for the targeted organism, and that the extra clades might represent untargeted sequences, as the clade banches very early on the cladogram and represent a small amount of sequences in the assembly.

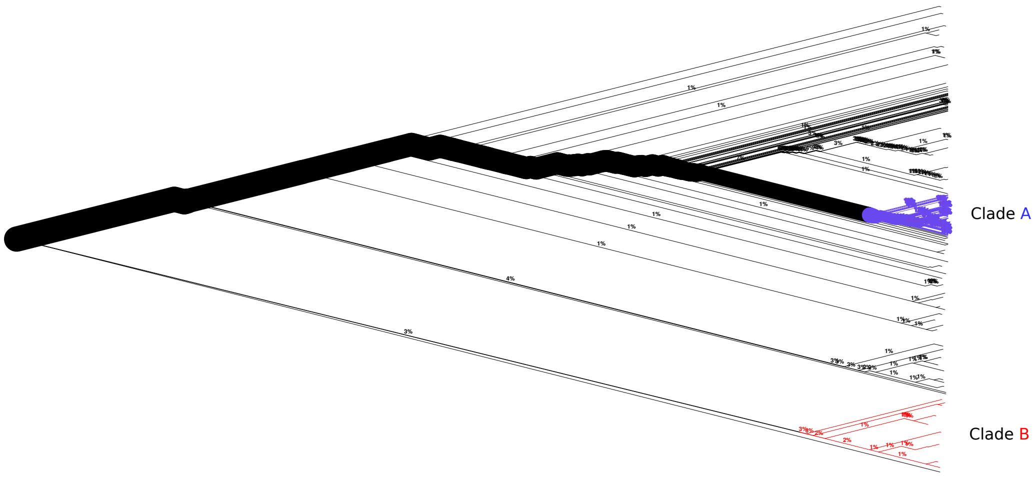

Figure 3: Exploration of the *Magnaporthe oryzae* TH12 assembly (Chiapello *et al.*, 2015). The width of the branches is set proportional to the cumulative size of contigs in the sub tree. The thicker path on the tree indicates a set of contigs with homgeneous oligonucleotide composition cumulating the majority of the assembled sequences The selection in blue will be called "Clade A", the user suspect this correspond to the host/targeted sequences. The selection in red will be called "Clade B", the user suspect this is an untargeted set of sequences, as the clade banches very early on the cladogram and represent a small amount of sequences in the assembly.

Distance (A.U.Euclidean)	rRNA	Subject	Strain	Reference length (pb)	Origin	Taxonomy	similarity	confidence
20	no	Magnaporthe oryzae 70-15	GI	3994966	genomic	Eukaryota	5/5	5.0/5
33	no	Magnaporthe grisea	GI	751312	genomic	Eukaryota	5/5	5.0/5
90	no	Beauveria bassiana	GI	301164	genomic	Eukaryota	4/5	5.0/5
96	no	Hypocrea jecorina	GI	687282	genomic	Eukaryota	4/5	5.0/5
104	no	Isaria farinosa	GI	27141	genomic	Eukaryota	4/5	5.0/5

Figure 4: Identification of clade A (see Figure 3) with GOHTAM (Ménigaud *et al.*, 2012).

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 5: Identification of clade B (see Figure 3) with GOHTAM (Ménigaud *et al.*, 2012).

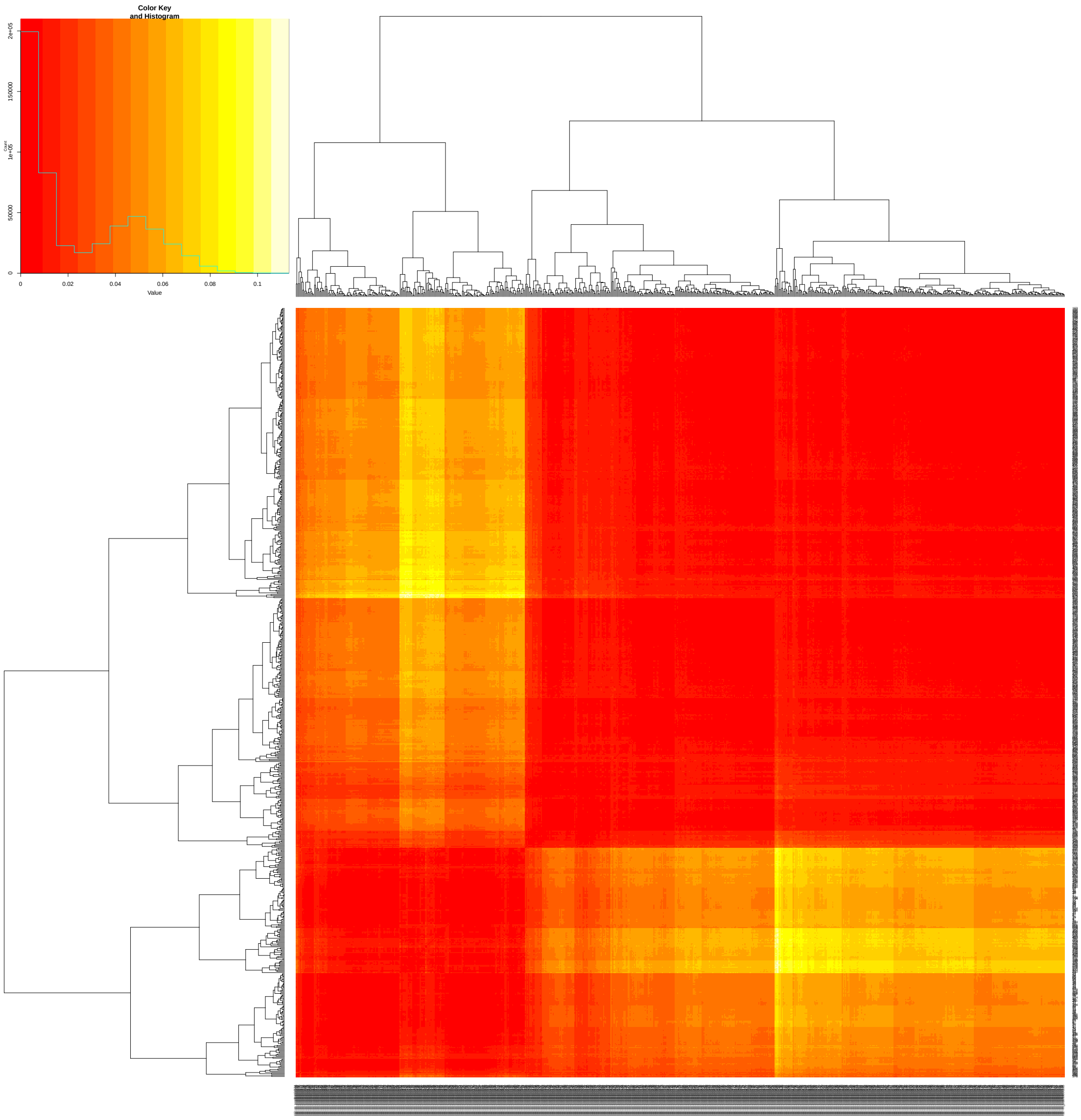

Figure 6: Sorted distance matrix of contigs of *Magnaporthe oryzae* TH12 assembly (Chiapello *et al.*, 2015). The following parameters in phyloselect.R were used: -d -w 20 -c 0.97 -m

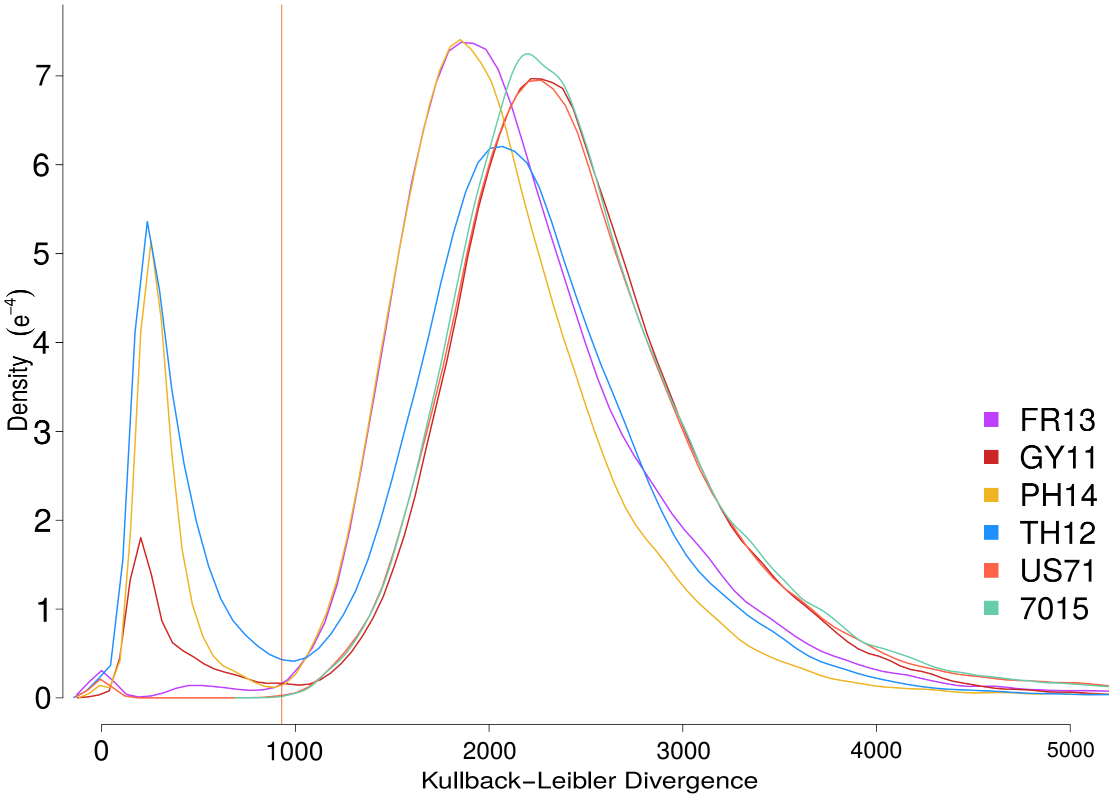

Figure 7: Distribution of distances between the composition profile of clade A (see Figure 3) and the scanning windows over the whole assembly. The untargeted threshold is the vertical red line.

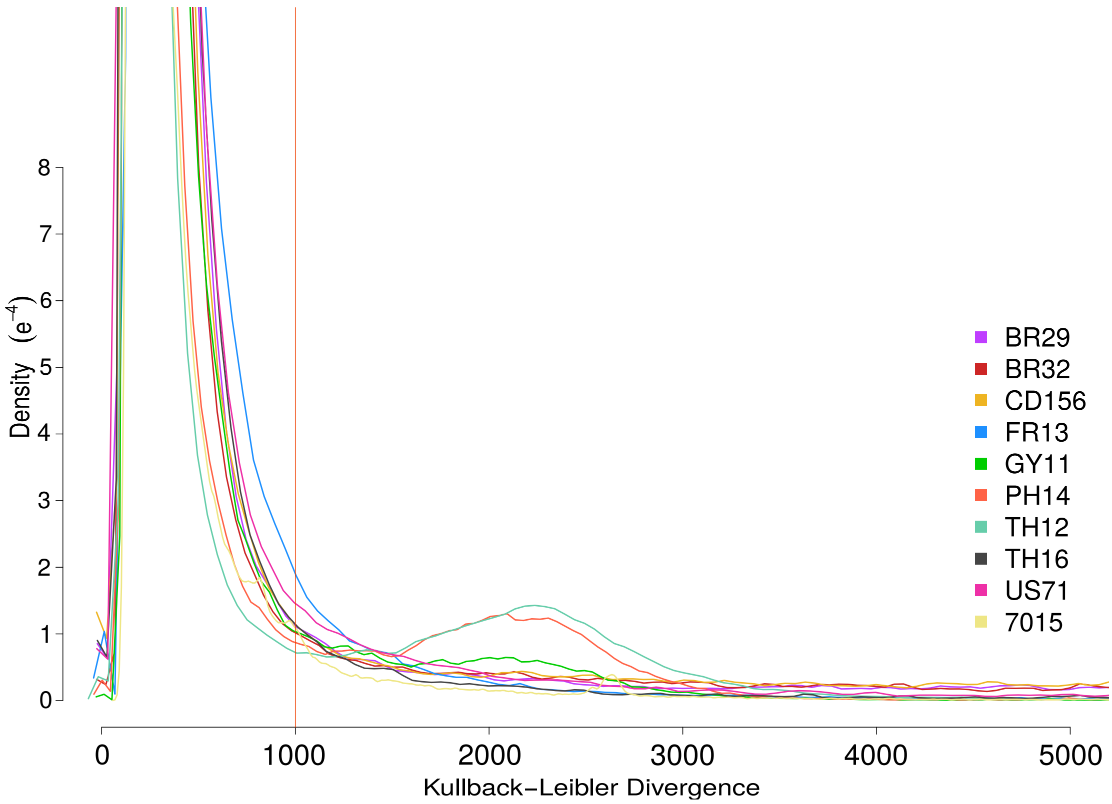

Figure 8: Distribution of distances between the composition profile of clade B (see **??**) and the scanning windows over the whole assembly. The host threshold is the vertical red line.

5.2 *Hypsibius dujardini*

The recent sequencing of the tardigrade (Boothby *et al.*, 2015) yielded a controversy about the composition of its genome sequence. Running PhylOligo on the genome assembly reveal the presence of sets of contigs with an homogeneous oligonucleotide composition grouping in diverging clades, which is in agreement with the previously proposed multiple contamination of the sample (Delmont and Eren, 2016; Koutsovoulos *et al.*, 2016). Unlike the example of *Magnaporthe oryzae*, the cladogram displays many branching clades each containing a substantial fraction of the assembled data.

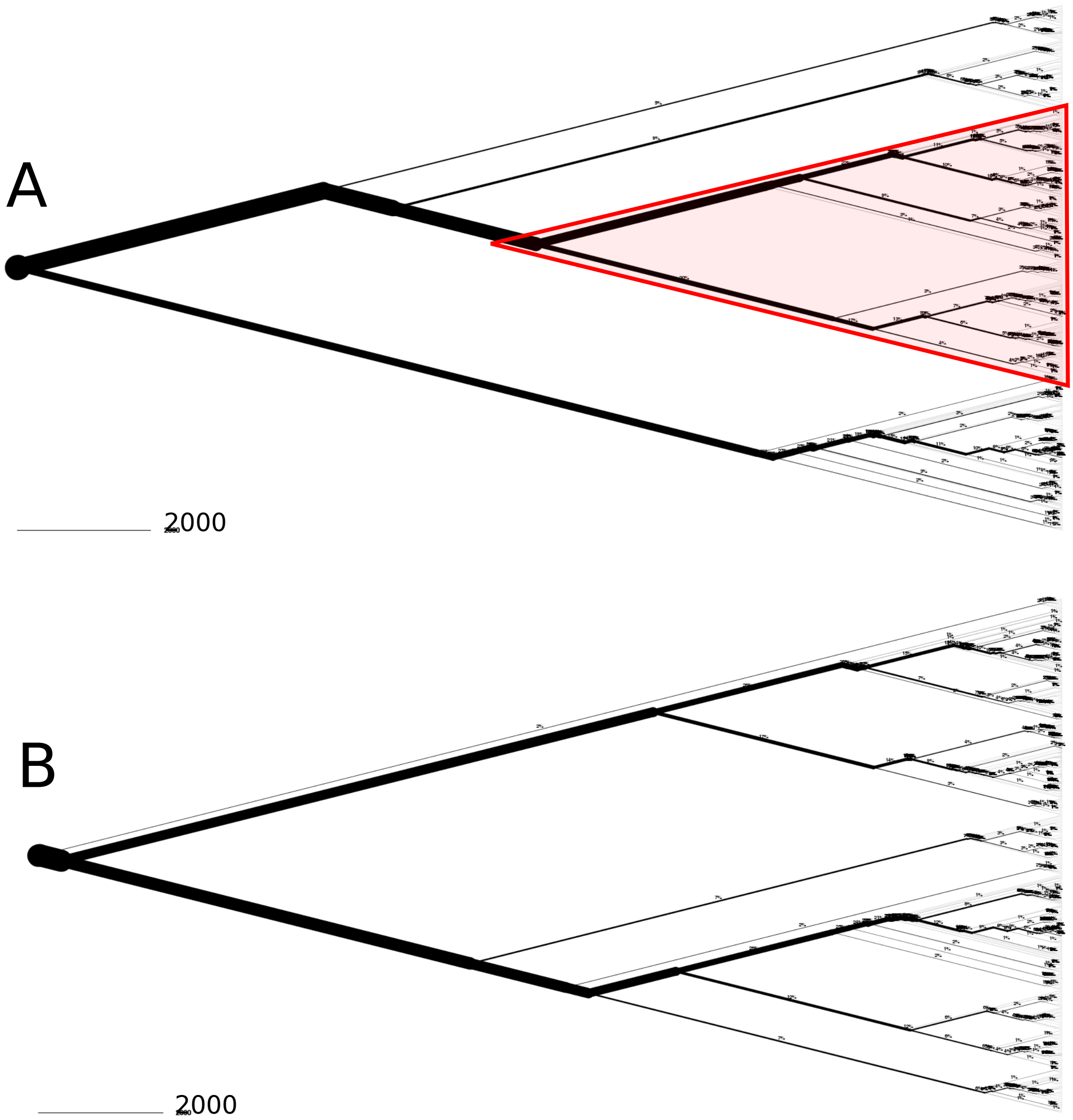

Figure 9: **A**: Exploration of the *Hypsibius dujardini* original assembly (Boothby *et al.*, 2015). **B**: Exploration of the *Hypsibius dujardini* curated assembly (Delmont and Eren, 2016). A very similar tree topology and branch width pattern can be identified in the original assembly (Figure 9 A) in the red triangle.

6 Bibliography

Boothby, T. C., Tenlen, J. R., Smith, F. W., Wang, J. R., Patanella, K. A., Osborne Nishimura, E., Tintori, S. C., Li, Q., Jones, C. D., Yandell, M., Messina, D. N., Glasscock, J., and Goldstein, B. (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences*, **112**(52), 15976–15981.

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.

Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A. A., and Blaxter, M. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade hypsibius dujardini. *Proceedings of the National Academy of Sciences*, **113**(18), 5053–5058.

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.