ORTHOSKIM: *in silico* gene capture in genomic and transcriptomic libraries

ORTHOSKIM is a pipeline providing different tools to capture targeted genes in genomic or transcriptomic libraries and to produce phylogenetic matrices for these genes.

This software was developed under the PhyloAlps project.

> **Applications:** ORTHOSKIM can be run on genomes skims libraries to capture chloroplast (CDS+rRNA+trnL-UAA), mitochondrial (CDS+rRNA) and ribosomal (rRNA+spacers) genes thanks to specific modes of assembly and capture. In addition, we provided additional mode to capture nuclear genes and BUSCO markers in transcriptomic or target sequences capture libraries.

citation:

- Pouchon et al. *in prep.* ORTHOSKIM: in silico gene capture in genomic and transcriptomic libraries for phylogenomic and barcoding applications.
- Inger Greve Alsos, Sebastien Lavergne, Marie Kristine Føreid Merkel, Marti Boleda, Youri Lammers, Adriana Alberti, Charles Pouchon, France Denoeud, Iva Pitelkova, Mihai Puşcaş, Cristina Roquet, Bogdan-Iuliu Hurdu, Wilfried Thuiller, Niklaus E. Zimmermann, Peter M. Hollingsworth, Eric Coissac, The Treasure Vault Can be Opened: Large-Scale Genome Skimming Works Well Using Herbarium and Silica Gel Dried Material, Plants, 10.3390/plants9040432, 9, 4, (432), (2020).

Table of contents

1. Installation

ORTHOSKIM is tested on Unix environment and requires:

- Exonerate
- SPAdes
- QUAST (optional)
- Diamond
- Blast
- MAFFT
- trimAl
- Needs Awk, Python

Some python libraries are also required, and can be installed via conda *install*:

- ete3==3.0.0b35
- joblib==0.16.0
- numpy==1.19.1
- Bio==0.3.0

ORTHOSKIM is installed from the source code:

```
wget https://github.com/cpouchon/OrthoSkim/archive/master.zip
unzip master.zip
cd OrthoSkim-master/
```

2. Input files

ORTHOSKIM required a sample file, a config file, and references sequences for targeted regions.

2.1. Configuration file

The following section describes the config file required for ORTHOSKIM. This tells ORTHOSKIM where to find files and all relevant informations. Users have to modify the *config_orthoskim.txt* file provided before running the pipeline. Default values are set for filtering and assembly steps.

```
nano config_orthoskim.txt
```

```
# ORTHOSKIM (v.1.0) config file
# Global parameters ------------------------------------------------------------------------------------------------
TOOLS=~/OrthoSkim-master/tools.sh                                                ## [1] path to file with tools ali
RES=~/run_orthoskim                                                              ## [2] output directory for all OR
EVALUE=0.00001                                                                   ## [3] evalue threshold for mappin
THREADS=15                                                                       ## [4] Number of threads which wil
VERBOSE=0                                                                        ## [5] Set verbose to TRUE (1) or
PLANT_MODEL=yes                                                                  ## [6] plants analyzed (yes/no)
```

```
# preprocessing the data ---------------------------------------------------------------------------------------------
LIST_FILES=~/OrthoSkim-master/ressources/listSamples.tab                    ## [7] Samples table. Specific for

# [assembly] mode ----------------------------------------------------------------------------------------------------
MEMORY=30                                                                   ## [8] Max memory which will be us
KMER=55                                                                     ## [9] Kmer size used in assembly,

# [filtering] mode: Filtering for contaminants in assemblies
SIMILARITY_CONTA_THSLD=65                                                   ## [10] Similarity threshold (%) u
MAPPING_CONTA_LENGTH=50                                                     ## [11] Minimal value of mapping.
TAXONOMIC_PHYLUM_EXPECTED=Embryophyta                                       ## [12] Taxonomic Phylum expected

# [database] mode: sequences of reference ----------------------------------------------------------------------------
MITO_ANNOTATIONS=~/OrthoSkim-master/data/mitochondrion_viridiplantae.gb     ## [13] file with mitochondrial an
NRDNA_ANNOTATIONS=~/OrthoSkim-master/data/nucrdna_viridiplantae.gb          ## [14] file with nucrdna annotati
CHLORO_ANNOTATIONS=~/OrthoSkim-master/data/chloroplast_viridiplantae.gb     ## [15] file with chloroplast anno
MITO_DB_FMT=genbank                                                         ## [16] database format: genbank,e
NRDNA_DB_FMT=genbank                                                        ## [17] database format: genbank,e
CHLORO_DB_FMT=genbank                                                       ## [18] database format: genbank,e
MITO_SIZE_MIN=200000                                                        ## [19] minimal size of mitochondr
MITO_SIZE_MAX=1000000                                                       ## [20] maximal size of mitochondr
NRDNA_SIZE_MIN=2000                                                         ## [21] minimal size of nuclear ri
NRDNA_SIZE_MAX=9000                                                         ## [22] maximal size of nuclear ri
CHLORO_SIZE_MIN=140000                                                      ## [23] minimal size of chloroplas
CHLORO_SIZE_MAX=200000                                                      ## [24] maximal size of chloroplas
SEEDS_THRESHOLD=0.8                                                         ## [25] minimal percent of seed co

# [capture] mode: extraction steps from mapping assemblies into a reference -------------------------------------------
MINLENGTH=90                                                                ## [26] minimal length of alignmen
REFPCT=0.4                                                                  ## [27] minimal covered exonic par
COVERAGE=3                                                                  ## [28] Minimal contigs coverage (
MINCONTLENGTH=500                                                           ## [29] Minimal contigs length all
EXO_SCORE=50                                                                ## [30] minimal score of mapping i

#---------  [busco] target ----------------------------------------------------------------------------------------------
BUSCO_REF=~/OrthoSkim-master/data/BUSCO_viridiplantae.fa                    ## [31] multi-fasta of BUSCO seque
BUSCO_TYPE=exon                                                             ## [32] type of sequence captured:

#---------  [nuclear] target --------------------------------------------------------------------------------------------
NUC_NT_REF=~/OrthoSkim-master/data/nucleusNT_unaligned.fa                   ## [33] multi-fasta of nuclear gen
NUC_AA_REF=~/OrthoSkim-master/data/nucleusAA_unaligned.fa                   ## [34] multi-fasta of nuclear gen
NUC_TYPE=exon                                                               ## [35] type of sequence captured:

#---------  [mitochondrion] target --------------------------------------------------------------------------------------
SEEDS_MITO_CDS=~/OrthoSkim-master/ressources/mitoCDS.seeds                  ## [36] CDS mitochondrial seeds of
SEEDS_MITO_rRNA=~/OrthoSkim-master/ressources/mitorRNA.seeds                ## [37] rRNA mitochondrial seeds o
MITO_REF_CDS=~/OrthoSkim-master/data/mit_CDS_unaligned.fa                   ## [38] multi-fasta file/name of m
MITO_REF_rRNA=~/OrthoSkim-master/data/mit_rRNA_unaligned.fa                 ## [39] multi-fasta file/name of m
MITO_TYPE=exon                                                              ## [40] Type of structure extracte

#--------- [chloroplast] target ----------------------------------------------------------------------------------------
SEEDS_CHLORO_CDS=~/OrthoSkim-master/ressources/chloroCDS.seeds              ## [41] chloroplast CDS seeds of r
SEEDS_CHLORO_rRNA=~/OrthoSkim-master/ressources/chlororRNA.seeds            ## [42] chloroplast rRNA seeds of
SEEDS_CHLORO_tRNA=~/OrthoSkim-master/ressources/chlorotRNA.seeds            ## [43] chloroplast tRNA seeds of
CHLORO_REF_CDS=~/OrthoSkim-master/data/chloro_CDS_unaligned.fa              ## [44] multi-fasta file/name of c
CHLORO_REF_rRNA=~/OrthoSkim-master/data/chloro_rRNA_unaligned.fa            ## [45] multi-fasta file/name of c
CHLORO_REF_tRNA=~/OrthoSkim-master/data/chloro_tRNA_unaligned.fa            ## [46] multi-fasta file/name of c
CHLORO_TYPE=exon                                                            ## [47] Type of sequence captured:

#--------- [nucrdna] target --------------------------------------------------------------------------------------------
NRDNA_REF=~/OrthoSkim-master/data/nucrdna_rRNA_unaligned.fa                 ## [48] multi-fasta file/name of r
SEEDS_NRDNA=~/OrthoSkim-master/ressources/nucrdna.seeds                     ## [49] ribosomal rRNA seeds of re
NRDNA_TYPE=exon                                                             ## [50] Type of sequence captured:

# [alignment] mode ---------------------------------------------------------------------------------------------------
SELECTION=on                                                                ## [50] Option to perform a select
TAXALIST=~/OrthoSkim-master/ressources/selTaxa_Primulaceae.tab              ## [51] list of taxa to select if
TRIMMING=on                                                                 ## [52] Option to trim alignments
MISSING_RATIO=1.0                                                           ## [53] maximal missing data thres
GENES_TO_CONCAT=~/OrthoSkim-master/ressources/listGenes_To_Concat.tab       ## [54] list of genes to include i

# [checking] mode ----------------------------------------------------------------------------------------------------
BARCODES=( matK rbcL )                                                      ## [55] list of barcodes used for
```

```
BARCODES_TYPE=chloroplast_CDS                                                    ## [56] Subdirectory in $RES/Extra
DB_LOCAL=off                                                                     ## [57] Option to perform a blast
BLAST_NT_DB=~/path_to_ntdb/nt                                                    ## [58] location of local NCBI nt
BLAST_NT_ACCESSION_TAXID=/bettik/pouchon/blastDB/nucl_gb.accession2taxid          ## [59] list of corresponding betw
TAXALIST=~/OrthoSkim-master/ressources/selTaxa_Primulaceae.tab                   ## [60] list of taxa for which tax
FAMILIES_LOCAL=off                                                               ## [61] option to include families
CORRESPONDING_FAMILIES=ecofind_out.tab                                           ## [62] table with query taxid and

# only for phyloskims users ----------------------------------------------------------------------------------------
CHLORO_GENES=~/OrthoSkim-master/ressources/listGenes.chloro                      ## [63] list of chloroplast genes
MITO_GENES=~/OrthoSkim-master/ressources/listGenes.mito                          ## [64] list of mitochondrial gene
NRDNA_GENES=~/OrthoSkim-master/ressources/listGenes.rdna                         ## [65] list of rdna nuclear genes
```

2.2. Dependencies

The path to all dependencies which are required in ORTHOSKIM must be supplied in the *tools.sh* file, using following command:

```
nano tools.sh
```

```
#!/bin/bash

SPADES=/Users/pouchonc/PhyloAlps/OrthoSkim/TOOLS/SPAdes-3.13.0-Darwin/bin/spades.py
DIAMOND=/Users/pouchonc/miniconda2/bin/diamond
EXONERATE=/usr/local/bin/exonerate
QUAST=/Users/pouchonc/miniconda2/bin/quast.py
BLASTDB=/Users/pouchonc/miniconda2/bin/makeblastdb
BLASTN=/Users/pouchonc/miniconda2/bin/blastn
MAFFT=/path/to/mafft
TRIMAL=/path/to/trimal
```

2.3. Sample file

A sample file must be supplied in the **$LIST_FILES** tab file (line 7 in *config_orthoskim.txt*).
This tab must contain for each sample the following columns :

- the sample name following *Genus_species_taxid_sampleid_otherids*
- the file-path to forward reads
- the file-path reverse reads

```
head ~/OrthoSkim/ressources/listSamples.tab

Veronica_crassifolia_996476_CAR009639_BGN_NFI   /Users/pouchonc/PhyloAlps/CDS/Veronica_crassifolia:996476/BGN_NFIOSW_4_
Androsace_helvetica_199610_CLA000520_BGN_ETA    /Users/pouchonc/PhyloAlps/CDS/Androsace_helvetica:199610/BGN_ETAOSW_2_1
```

2.4. References files (database)

ORTHOSKIM uses a multi-taxa references bank to capture targeted genes into assemblies for all the different targets (see *3. Pipeline description* below part).

This bank of references is created in ORTHOSKIM pipeline for the *nucrdna*, *chloroplast* and *mitochondrion* targets directly from genomic annotations collected by users in a single file for each comportament (genbank or embl format required, file-path set in config file at lines 13-15). These annotations can be collected directly from the NCBI for example. To achieve this, seeds are required for each type of gene (CDS, rRNA + tRNA for chloroplast) to identify each gene with a standard name (header) as following "*>genename_taxid_Genus_species_other-arguments*" (e.g. *cox1_3702_Arabidopsis_thaliana* for cox1 gene). Location of seeds is given in lines 36-37, 41-43 and 49 of the config file.

ORTHOSKIM creates a reference multi-fasta file for the coding regions (CDS) with amino acid sequences, and nucleotidic sequences for the non-coding regions (*i.e.* rRNA + tRNA only for *chloroplast* target). Location of these output files are set in the *config_orthoskim.txt* file at lines 38-39, 44-46 and 48.

> **NOTE:** As a selection on assemblies is done (see *3.3.1.b.* section), users have to collect all three mitochondrion, chloroplast and nucrdna genomes before to run ORTHOSKIM if plant models are analyzed (l.6), or both mitochondrion and nucrdna genomes for other models. All seeds are also required for corresponding regions. Moreover, as a taxonomic selection is done according to the query taxon, we recommend to include as many divergent taxa as possible in the annotations.

Here, an example of output CDS bank from mitonchondrial annotations (using the mode `-m database` and the target `-t mitochondrion`).

```
head ~/OrthoSkim/data/mit_CDS_unaligned.fa

>cox2_103999_Codonopsis_lanceolata
MRELEKKNTHDFILPAPADAAEPWQLGFQDGATPIMQGIIDLHHDIFFFLIMILVLVLWILVRALWLFSSKRNPIPQRIVHGTTIEILRTIFPSIILMFIAIPSFALLYSMDEVVVDPA
>cox2_104537_Roya_obtusa
MILKSLFQVVYCDAAEPWQLGFQDAATPMMQGIIDLHHDIMFFITIIITFVLWMLVRVLWHFHYKKNPIPQRFVHGTTIEIIWTIIPSIILMFIAIPSFALLYSMDEVVDPAITIKAIG
>cox2_111617_Ulva_fasciata
MKNFSFSYCILITLFNISVISSCDAPLSATSAMLDRFGFQEPASPLMEGLIALHSDIWAIMLFVAGFVLYMMCAILYNFSASSSEISYKVHHHSLIEIVWTTIPALILCVIAIPSFTLL
>cox1_112509_Hordeum_vulgare_subsp._vulgare
MTNLVRWLFSTNHKDIGTLYFIFGAIAGVMGTCFSVLIRMELARPGDQILGGNHQLYNVLITAHAFLMIFFMVMPAMIGGFGNWFVPILIGAPDMAFPRLNNISFWLLPPSLLLLLSSA
>nad1_119543_Anomodon_attenuatus
MRLYIIGILAKILGIIIPLLLGVAFLVLAERKIMASMQRRKGPNVVGLFGLLQPLADGLKLMIKEPILPSSANLFIFLMAPVMTFMLSLVAWAVIPFDYGMVLSDLNVGILYLFAISSL
```

Concerning the *nucleus_aa* and *nucleus_nt*, users have to provide the multi-fasta files of genes, and set their location in the config file to the corresponding sections (lines 33-34 of the config file). The gene name restrictions have to be respected.
For the *busco* target, the multi-fasta file must contain the BUSCO dataset of ancestral sequences in amino acid sequences, called *ancestral_variants* in datasets. The location of this database is given in line 31 of the config file).

Here, an overview of the busco sequences needed:

```
head ~/OrthoSkim/data/BUSCO_viridiplantae.fa

>10018_0
IASVVSEIGLGSEPAFKVPEYDFRSPVDKLQKATGIPKAVFPVLGGLAVGLIALAYPEVLYWGFENVDILLESRPKGLSADLLLQLVAVKIVATSLCRASGLVGGYYAPSLFIGAATGM
>10018_1
VASVVSEIGLGSEPAFKVPEYDFRSAVDSLKKTLGLPKAVLPALGGLIVGLIALAYPEVLYWGFENVDILLESRPRGLSAELLLQLVAVKVVATSLCRASGLVGGYYAPSLFIGAATGM
...
```

By default, ORTHOSKIM is supplied with sequences for plants containing the BUSCO plant set (viridiplantaeae_odb10), 353 UCE designed for angiosperms (Johnson et al., 2018) and a subset of annotations for chloroplast, mitochondrion and nucrdna genomes (in *data/* directory). More annotations can be downloaded as shown in the *4.2 ORTHOSKIM tutorials* section.
Users can easily adapted the files for other models by respecting the recommendations (see documentation).

3. Pipeline description

The gene capture is driven on genomic or transcriptomic global assemblies. This allowed to capture from a single assembly run different targeted genes (*e.g.* chloroplast, mitochondrial and ribosomal genes) thanks to alignments of contigs into gene database.

ORTHOSKIM pipeline uses different mode to compute the databases, capture targeted regions, align them between taxa, or to check assemblies.

> **Note**: A *mode_done.log* file is created containing samples that were correctly processed, whereas unprocessed samples were added into *mode_error.log* file. This file could be used to remove processed samples from the initial sample file if

> the script has to be rerun. Command lines are also print if users want to rerun specific commands on samples.

3.1. Database (optional)

ORTHOSKIM provides a mode to create gene database for the mitochondrial, chloroplast and ribosomal regions with `-m database` mode along with `-t mitochondrion, chloroplast, nucrdna` targets. To do this, genomic annotations of these compartments has to be collected across taxa in a single file for each regions and set into the config file.

ORTHOSKIM will then extract all notified CDS, rRNA and tRNA genes and align them into given seeds thanks to *exonerate* to keep a standard gene name. Output files (l. 38-39, 44-46 and 48) are created containing a bank of genes, all well identified. Only genes given for the seeds will be included.

> **NOTE**: Users have to collect all three genomes and corresponding seeds to run ORTHOSKIM (or two for non plant model). If users want to capture nuclear or busco markers, this step is skipped. In such case, users have to collected genes of reference for these markers into the *config_orthoskim.txt* file, by following instructions for the sequence header.

3.2. Global assemblies and cleaning

3.2.1. genomic/transcriptomic assembly

Global assemblies are performed for each taxon of the taxa file (l.7) by using SPAdes and have to be run using the `-m assembly` mode and `-t spades` or `-t rnaspades` target (according to the type of library). SPAdes will be run by using the assembly options (**$THREADS**,**$MEMORY**,**$KMER**) specified in the config file (l. 4, 8-9).

ORTHOSKIM will then output a *samplename/* subdirectory into the **${RES}/Assembly/SPADES/** or **${RES}/Assembly/RNASPADES/** given per sample included in the taxa file.

After SPAdes runs, ORTHOSKIM has to preprocess SPAdes scaffolding contigs by renaming the file according to the same sample name provided in the taxa file and ordering them into **${RES}/Assembly/Samples/unfiltered/** directory.
This is made under `-m reformate` mode and `-t spades` or `-t rnaspades` targets according to the version used.

3.2.2. assemblies cleaning

The capture of genes will be run only on cleaned assemblies after running `-m cleaning` mode. This step identifies contigs which are not expected in the assembly dataset and removes them.

To do this, all contigs are blast against rRNA databases SILVA and RFAM supplied in sortmerna (v.4.2.0), composed of the 5S, 5.8S, 16S, 23S, 18S and 28S genes for bacteria, archaea and eukarya. Moreover, contigs are also blasted against to own DBFAM database including a subset of chloroplast, mitochondria and nucrdna genomes for eukarya.
The best hits are identified for each contigs, and only contigs mapping to the expected taxonomy are kept according to the taxonomy corresponding file provided (*~/OrthoSkim-master/ressources/rRNA_database_taxonomy.txt*). The expected taxonomy is set by the user at the line 12 (**$TAXONOMIC_PHYLUM_EXPECTED**).

> **NOTE:** Please check the taxonomy provided in the ~/OrthoSkim-master/ressources/rRNA_database_taxonomy.txt file to set a correct phylum (*e.g.* "Embryophyta", "Eumetazoa","Arthropoda","Annelida" etc). We also recommend to keep low values for parameters of **$SIMILARITY_CONTA_THSLD** and **$MAPPING_CONTA_LENGTH** (l. 10-11) as a taxonomic comparison is done between entries in the database.

3.3. Gene capture

The capture of targeted genomic regions is made using the `-m capture` mode according to three steps:

3.3.1. Selection

3.3.1.a. gene selection

For all targets (with the exception of BUSCO), ORTHOSKIM will first select the closest reference for each gene and for each taxa from the given database of references.

To achieve this, the selection is made according to the NCBI taxonomy thanks to the taxid number given in the sample name. If the taxid does not exist in the NCBI taxonomy, ORTHOSKIM will use seeds as references for the chloroplast, mitochondrion and nucrdna targets, or the longest sequences for other targets.

For the BUSCO, no selection is made into the sequences as ancestral variants sequences (already aligned) are used for the reference.

After this, if CDS are targeted, a diamond database is created for each amino acid sequences provided in the retained sequences (with *diamond makedb*). Otherwise, a blast database (*makeblastdb* program) is formatted.

3.3.1.b. contig selection

Cleaned contigs are selected to reduce the computational time of the following alignments and to correctly identify the right genomic origin of the targeted genes.

To achieve this, for the mitochondrion, chloroplast and nucrdna targets, we identified the contigs by mapping them with blast directly on five closest genomes from the provided annotations for each taxa for all three genomes in plant models (or both mitochondrion and nucrdna genomes for others). For example, if a contig align more on chloroplast than on mitochondrion or nucrdna, it will be identified as chloroplast. Only genomes with a minimal/maximal size given in **$[MITO,CHLORO,NRDNA]_SIZE** arguments will be considered (lines 19-24 of the config file).

For example, as near to 35% of the ancestral plastid genomes has been estimated to be transferred and conserved in to mitochondrial genomes (Park et al., 2020), this step allows to avoid capturing a mitochondrial copy of a targeted chloroplast gene leading to taxonomic mis-positioning, and *vice versa*. It allows also to attribute the right RNA gene copy to its original cellular compartment.

For the other targets, the selection is performed by mapping the contigs directly on the selected genes by using diamond or blast if the sequences are proteic or nucleotidic. A threshold on the kmer coverage (**$COVERAGE**), the contig length (**$MINCONTLENGTH**) and the minimal evalue (**$EVALUE**) is set by users to exclude all contigs below these values for the following step.

3.3.2. Exon/intron gene prediction

Alignments are conducted on the selected contigs and the selected genes from exonerate by incorporating all the appropriate gaps and frameshifts, and by modelling introns. The *protein2genome* mode is used when CDS are targeted or the *genome2genome* mode for other targets. A *gff* output table is created in **${RES}/Mapping/[nucleus,mitochondrion,chloroplast]/** folder for each sample. Only sequences with a mapping score above the **$EXO_SCORE** value are kept (l. 30 of the config file).
By default we set this score at 50. We recommend to not set too high values (if the gene length is short) as a selection in alignment scores is next performed. Otherwise short genes could be skipped.

> **Note:** Concerning plant models, we performed a second control during the gene alignment to ensure the right origin of organelle. To achieve this, for example, during the chloroplast capture, we align the mitochondrial seeds on selected chloroplast contigs to check if a contig position best align on selected genes than on seeds. This allows to verify if chimeric organelle contig were assembled on the conserved regions and thus wrongly pass the selection of contigs. Seeds of both mitochondrion and chloroplast have to be done by users even if only chloroplast genes will be captured.

3.3.3. gene extraction

Extraction of selected genes is conducted from the gff table by identifying the best alignment for each covered regions of each gene. Type of gene structure extracted (i.e. exon, intron or all) is choosen by the users in the config file. This step is conducted

into multiple processors using the specified in the the *config_orthoskim.txt* file (l. 4).
For the nucrdna target, ITS1 and ITS2 barcodes are extracted from the intronic regions of rRNA probes designed during the database step.

Output gene files are created in the **${RES}/Extraction/[mitochondrion,chloroplast]_[CDS,tRNA,rRNA]/** or **${RES}/Extraction/[nucleus_aa,nucleus_nt,nucrdna,busco,uce]/** as following:

```
ls -l ~/RES/Extraction/busco/

-rw-r--r--  1 pouchonc  staff  1758  5 jui 11:11 10104.fa
-rw-r--r--  1 pouchonc  staff  1964  5 jui 11:11 10521.fa
-rw-r--r--  1 pouchonc  staff  5071  5 jui 11:11 10785.fa
-rw-r--r--  1 pouchonc  staff  1400  5 jui 11:11 11487.fa
-rw-r--r--  1 pouchonc  staff  2040  5 jui 11:11 11505.fa
-rw-r--r--  1 pouchonc  staff  1778  5 jui 11:11 1504.fa
```

> **Note:** Once genes were captured, users can use the *checking* mode (-m) on some genes to check the family rank found for these genes for each queried taxa. A blast is done on NCBI database and a comparison is made according to the given taxid. Please see required parameters on the config file.
> Users have to download and unzip the corresponding file between accesions and taxids as following:

```
wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid//nucl_gb.accession2taxid.gz
```

A subdirectory is created **${RES}/Errors/** with a *ValidationSamples.out* file. This file indicates for each taxa and for each gene if the checking is TRUE/FALSE/NA, as following:

```
Abies_alba_45372_PHA000002_RSZ_RSZAXPI000687-79    TRUE    TRUE
Abies_balsamea_90345_TROM_V_43901_CDM_AOZ    TRUE    TRUE
Abies_sibirica_97169_TROM_V_97238_CDM_AVE    TRUE    TRUE
```

> If users want to combine chloroplast_tRNA (e.g. trnL-UAA) and CDS genes (e.g. matK and rbcL), a new directory must be created in the **${RES}/Extraction/** subdirectory with gene files inside; users have next to set the name of this directory in the config file (l. 56).

3.4. Summary statistics

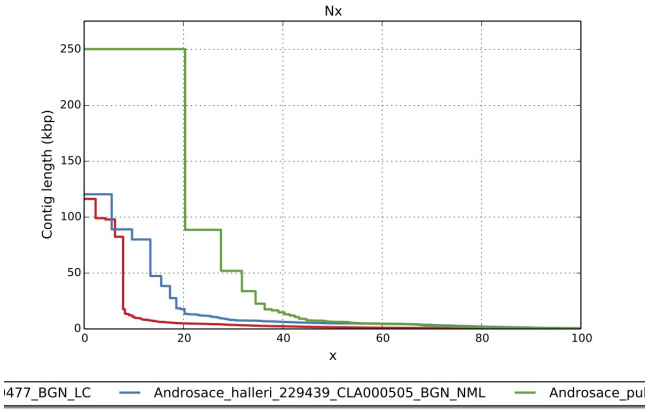
a. on assemblies

ORTHOSKIM allows to output summary statistic on contigs assemblies thanks to QUAST by specifying the `-m stat_assembly` mode.

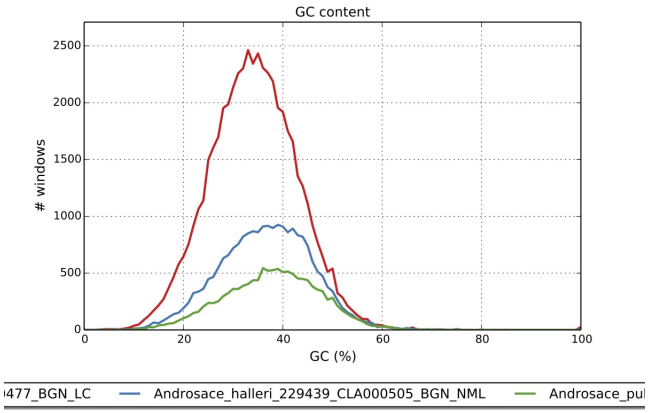
The output *transposed_report.txt* tab file will be in **${RES}/report_SPAdes_assemblies/** directories given indication on the assembly.

```
head ~/RES/report_chloro_assemblies/transposed_report.txt

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total l

Assembly                                          # contigs (>= 0 bp)  # contigs (>= 1000 bp)  # contigs (>= 5000 b
Veronica_crassifolia_996476.CAR009639.BGN_NFI.chloro  1                   1                      1
Androsace_helvetica_199610.CLA000520.BGN_ETA.chloro   1                   1                      1
Doronicum_columnae_118758.PHA003018.BGN_EEH.chloro    1                   1                      1
```

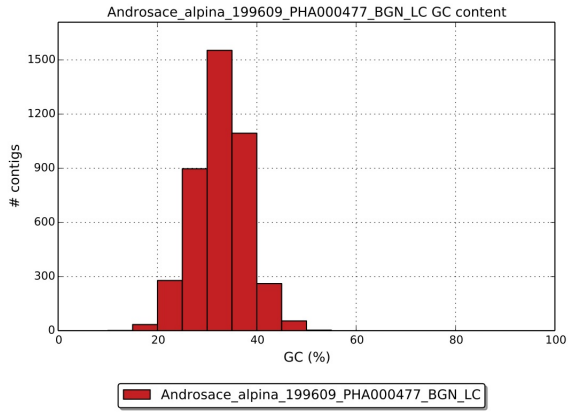
ORTHOSKIM will also output the *report.pdf* file generated with QUAST containing:



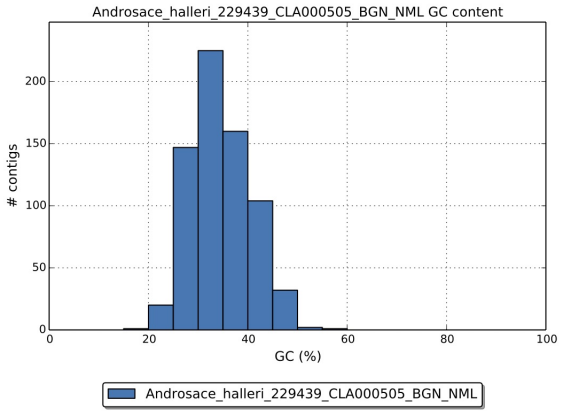
Nx values varying from 0 to 100%



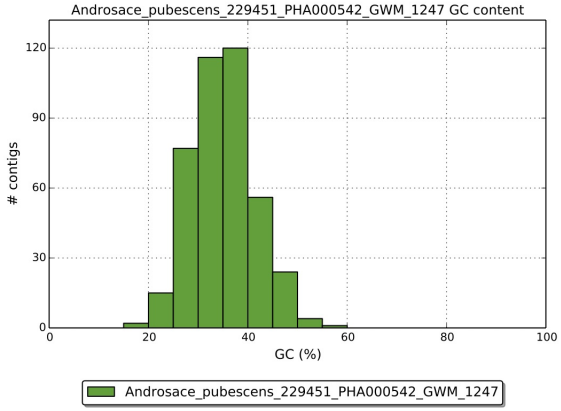
GC content in the contigs for all samples



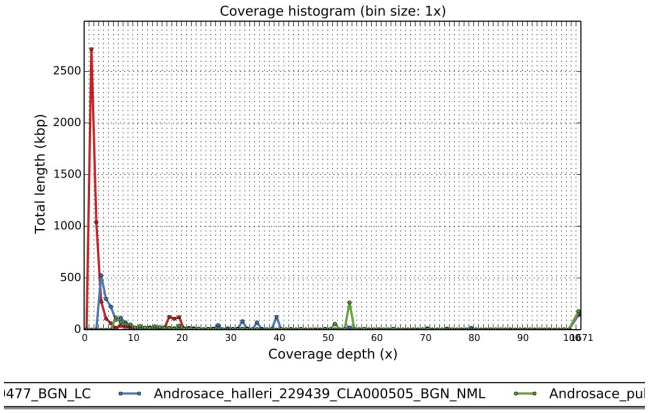
GC content in the contigs for sample1



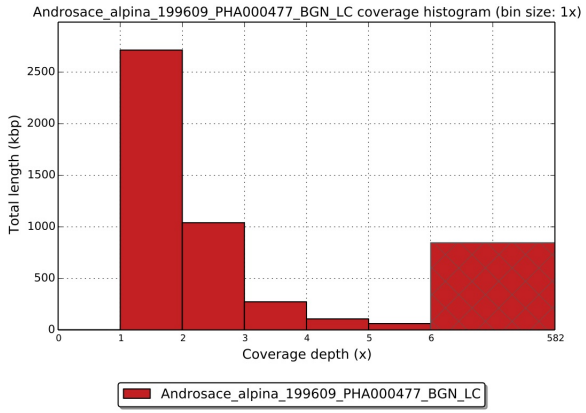
GC content in the contigs for sample2



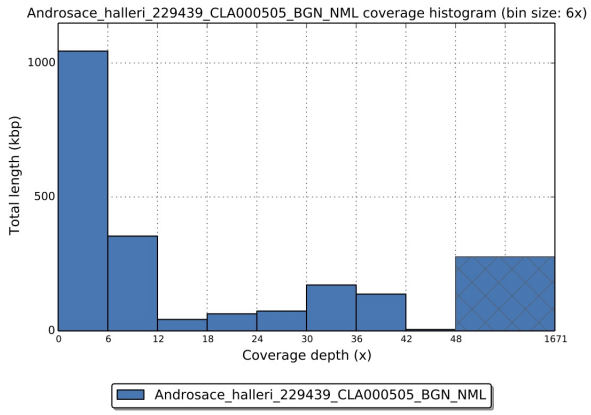
GC content in the contigs for sample3



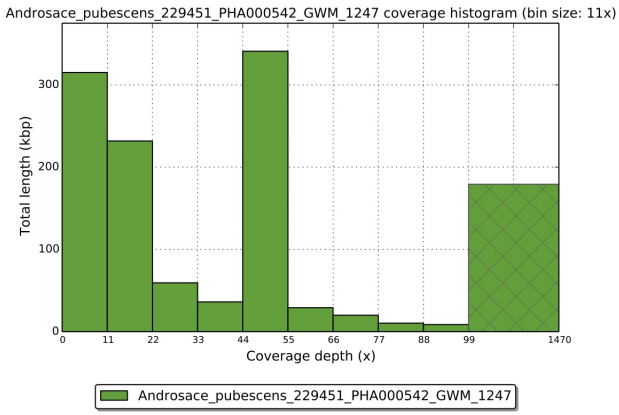
distribution of total contig lengths at different
read coverage (only for SPAdes mode)



coverage distribution for sample1



coverage distribution for sample2



coverage distribution for sample3

> **Note**: see QUAST manual for more details. Outputs for icarus genome visualizer were also kept in the directory to visualize assemblies.

b. on capture

ORTHOSKIM allows to get statistic from the gene capture by using the `-m stat_capture` mode for sequences for the different targets (multiple targets can be supplied). The pipeline output a *report.tab* into this path containing:

- the gene name (gene_name)
- the taxa coverage (taxa)
- the mean length (mean)
- the minimal length of sequence found (minlen)
- the maximal length of sequence found (maxlen)
- the standard deviation (std)
- the 25th percentil (pct25)
- the 50th percentil (pct50)
- the 75th percentil (pct75)

```
head ~/RES/chloroplast_CDS_report.log
gene    taxa  mean  min   max   std   pct25   pct50   pct75
rpoC2    7   3316  1831  4152  880   2743    3561    4093
rps19    7   280   273   309   11    276     276     276
ycf1     6  2026  378   5607  1769  820     1346    2462
rpoC1    7   1842  945   2121  413   1795    2058    2092
psbA     7  1059  1059  1059  0     1059    1059    1059
atpI     7  741   741   744   1     741     741     741
rpl2     7  763   483   828   115   792     801     825
ndhH     7  1179  1179  1179  0     1179    1179    1179
rbcL     7  1425  1425  1425  0     1425    1425    1425
```

> **Note**: The full summary statistics of gene capture, as shown in our paper, can be obtained by using the *FullStat.py* function provided in the src/ directory as following:

```
~/OrthoSkim-master/src/FullStat.py -pfind -p Extraction/chloroplast_CDS/ -t chloroplast_CDS_done.log > stat_cp.txt
```

with -p: path where genes are extracted and -t: list of taxa to compute statistics

3.5. Alignment of taxa

ORTHOSKIM provides a mode to align taxa for each captured genes by using the `-m alignment` mode. We use MAFFT to align each gene individually with the '--adjustdirectionaccurately' option. This alignment can be filtered if the option is chosen by users using trimAl with the heuristic 'automated1' method (*on/off* at line 52 of the config file).
In addition, users can choose which taxa will be aligned by stating if a selection is made on taxa (*on/off* at line 50 of the config file). In such case, a list of taxa to align has to be given (l. 51).

ORTHOSKIM will output the concatenated alignment of genes along with a partition file under a RAxML-style format suitable for phylogenetic inferences. For such needs, a list if gene that will processed has to be given (l. 54). A tab with information about gappy or missing data is also produced by sample.

```
-rw-r--r--    1 pouchonc  staff        1341  5 mai 10:41 concatenated.fa
-rw-r--r--    1 pouchonc  staff          21  5 mai 10:41 concatenated.info
-rw-r--r--    1 pouchonc  staff         101  5 mai 10:41 concatenated.missingdata
-rw-r--r--    1 pouchonc  staff          19  5 mai 10:41 concatenated.partitions
```

```
head ~/PATH/concatenated.fa
>Carex_elongata_240685_PHA001842_BGN_MAS
CTTACTATAAATTTCATTGTTGTCGATATTGACATGTAGAAT-GGACTCTCTCTTTATTCTCGTTTGATTTATCA-TCATTTTTTCAATCTAACAAACTCTAAAATGAATAAAATAAAT
```

```
>Dipsacus_fullonum_183561_TROM_V_159792_CDM_BFO
CTTACTAAAAATTTCATTGTTGCCGGTATTGACATGTAGAATGGGACTCTATCTTTATTCTCGTCCGATTAATCAGTTCTTCAAAAGATCTATCAGACTATGGAGT----------
```

```
head ~/PATH/concatenated.info
1    625 trnL-UAA    part1
```

```
head ~/PATH/concatenated.missingdata
Carex_elongata_240685_PHA001842_BGN_MAS    0.0096
Dipsacus_fullonum_183561_TROM_V_159792_CDM_BFO    0.1808
```

```
head ~/PATH/concatenated.partition
DNA, part1 = 1-625
```

4. Running ORTHOSKIM

ORTHOSKIM uses a command line interface (CLI) that can be accessed through a terminal. Please use the -help (-h) flag to see a description of the main arguments.

```
./orthoskim -h
```

ORTHOSKIM is called step by step. Recommendations about steps are given in the previous description (section 3). After edition of the *tools.sh* and *config_orthoskim.txt* files (with all required files and formats), ORTHOSKIM is called by using the different modes.

We detail instructions here through the description of arguments and the tutorials below.

4.1. ORTHOSKIM arguments

-c (config file): config file edited by users. See instructions above.

-m (mode): different modes encoded in ORTHOSKIM.

> - **alignment:**: Give taxa alignments of selected genes. Each gene are aligned individually with MAFFT and then concatenated. Multiple targets (-t) can be set. A selection of taxa can be performed to decide to which taxa will be align. Alignments can also be trimmed or not.
> A concatenation and a partition file are generated.

> - **database:** compute the reference bank of gene database for the chloroplast, mitochondrion and nucrdna targets. Annotation needs to be collected in a single file in genbank/embl format. Seeds are required from one organism for each targeted genes with a standard gene name. CDS genes are given in proteic sequences and others in nucleotidic sequences.

> - **capture:** Capture of genes from targeted markers. A selection of the closest reference is made for each gene according to the taxonomy. If errors occurred during this step, OrthoSkim will use seeds as reference (exception for busco and uce targets). Users has to collected seeds for the targeted genes. Users choose to capture exonic, intronic or both regions.

> - **checking:** Checking of the family rank found for given gene with blast into the NCBI database and taxonomic comparison with taxid given for the queried taxa.

> - **cleaning:** Cleaning of contigs according blast mapping into RNA databases and DBFAM databases. An expected taxonomic level is required to consider as "good" contigs for which the best-hit corresponds to this level.

> - **assembly:** Perform global assembly using SPAdes assembler. Specificities for assembly are given in the config file (Kmer, memory, threads).

> - **reformate:** Extract and reformate the scaffold fasta file for each taxa. A Samples/ subdirectory is generated containing all taxa contig files.

> - **stat_assembly:** Compute summary statistics of assemblies using QUAST. Graphs and a table with information over the contigs number, the contigs size, the GC content, the N50 value are generated.

> - **stat_capture:** Compute summary statistics of extraction. A file (target_report.log) is generated including the taxa recovery, the mean size and the range size by gene. Multiple targets (-t) can be set.

-t (targets): targeted regions by the mode (-m) used.

> For *database* mode:
>
> - **chloroplast** (creation of chloroplast database containing CDS+rRNA+trnL-UAA genes)
> - **mitochondrion** (creation of mitochondrial database containing CDS+rRNA genes)
> - **nucrdna** (creation of ribosomal database containing rRNA genes and probes for spacer regions)

> For *alignment*, *capture* and *stat_capture* modes:
>
> - **busco** (BUSCO markers)
> - **chloroplast_CDS** (coding sequence of chloroplast)
> - **chloroplast_rRNA** (non coding chloroplast rRNA genes)
> - **chloroplast_tRNA** (only tRNA trnL-UAA gene)
> - **mitochondrion_CDS** (coding sequence of mitochondrion)
> - **mitochondrion_rRNA** (non coding mitochondrial rRNA genes)
> - **nucleus_aa** (nucleus genes in proteic sequences in databse)
> - **nucleus_nt** (nucleus genes in nucleotidic sequences in databse)

> For *assembly* and *reformate* modes:
>
> - **spades** (use of SPAdes software to compute genomic assemblies)
> - **rnaspades** (use RNA version of SPAdes software to compute transcriptomic assemblies)

4.2. ORTHOSKIM tutorials

In this section, we describe a tutorial to capture chloroplast, mitochondrial and ribosomal genes for our list of taxa.

4.2.1. databases

To begin, users have to install all dependencies, create a sample file, edit the *config_orthoskim.txt* and the *tools.sh* files and collect annotations files for the targeted compartments. By default, subsets of genomic annotations are given for Viridiplantae with ORTHOSKIM to quickly run the software.

Here, we show an example to collect these annotations from the NCBI for the chloroplast for plants.

```
wget -m -np -nd 'ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plastid/' -A.genomic.gbff.gz
gunzip *.genomic.gbff.gz
cat *.genomic.gbff >> plastid.genomic.gb
rm *.genomic.gbff
```

We supplied with ORTHOSKIM a function *AnnotFilter.py* to filter annotations according to taxonomy (e.g. viridiplantae). Here, we collected all annotations of viridiplantae.

```
~/OrthoSkim-master/AnnotFilter.py -i plastid.genomic.gb -f genbank -l viridiplantae -o ~/OrthoSkim-master/data/chloropl
Filtering annotations on taxonomy
1 level(s) of taxonomy set: viridiplantae
      parsing annotations [.....................................................] 100 %
4869 / 5201 annotations selected on taxonomy
```

> **NOTE:** the output (given with **-o**) has to be the same which is set in the config file (line 15: **CHLORO_ANNOTATIONS=~/OrthoSkim-master/data/chloroplast_plants.gb**). Morevover, multiple taxonomic levels can be given in -l with a coma separator (*e.g.* -l asteraceae,helianthae).

Once all annotations are collected, we compute the database for the three targets. The seeds were collected on *Arabidopsis thaliana* by including CDS, rNA genes plus trnL-UAA.

```
./orthoskim -m database -t chloroplast -c config_orthoskim.txt
./orthoskim -m database -t mitochondrion -c config_orthoskim.txt
./orthoskim -m database -t nucrdna -c config_orthoskim.txt
```

> **NOTE:** We supplied with ORTHOSKIM a python function SortDB.py allowing to select a subset of lineages by family in gene or genome databases. It allows to reduce the computational time of capture steps by reducing the number of sequences and keeping a taxonomic diversity within the database. This function can be run directly on outputs as following:

```
SortDB.py -i chloroplast_CDS.fa -f fasta -l 3 -o selected_chloroplast_CDS.fa -m gene
SortDB.py -i chloroplast_ncbi.gb -f genbank -l 5 -o selected_chloroplast_CDS.embl -m genome
```

with -i input genes/genomes file; -l number of queried lineages by family; -f input file format (embl/ genbank/fasta); -o output name (format fasta for genes or embl for genomes); -m mode (gene/genome)

4.2.2. assemblies and filtering

We next perform global assemblies of our samples and reformate the outputs. After that, assemblies were cleaned by removing all potential contaminants.

```
./orthoskim -m assembly -t spades -c config_orthoskim.txt
./orthoskim -m reformate -t spades -c config_orthoskim.txt
./orthoskim -m cleaning -c config_orthoskim.txt
```

> **Note:** For the cleaning step, we set the expected phyllum at "Embryophyta" (l.12 of the config file).

If you want to get summary statistics of assemblies, users can run the following command:

```
./orthoskim -m stat_assembly -c config_orthoskim.txt
```

4.2.3. gene capture

The next step consists on capture all targeted genes into these assemblies. To do this, we run the `capture` mode with our different targets.

```
./orthoskim -m capture -t chloroplast_CDS -c config_orthoskim.txt
./orthoskim -m capture -t chloroplast_rRNA -c config_orthoskim.txt
./orthoskim -m capture -t chloroplast_tRNA -c config_orthoskim.txt
./orthoskim -m capture -t mitochondrion_CDS -c config_orthoskim.txt
./orthoskim -m capture -t mitochondrion_rRNA -c config_orthoskim.txt
./orthoskim -m capture -t nucrdna -c config_orthoskim.txt
```

> **Note**: in this example for the chloroplast tRNA, we change the CHLORO_TYPE (l.45) from "exon" to "intron", to capture the intron of the trnL-UAA.

Summary statistics about the capture can be obtained by using the following mode:

```
./orthoskim -m stat_capture -t chloroplast_CDS -t chloroplast_rRNA -t chloroplast_tRNA -t mitochondrion_CDS -t mitochon
```

> **NOTE:** Here, multiple targets (-t) are given in the command same line.

4.2.4. alignments

Finally, we compute a supermatrix by aligning captured genes (here on chloroplast data) that can be used for phylogenetic inferences.

```
./orthoskim -m alignment -t chloroplast_CDS -t chloroplast_rRNA -c config_orthoskim
```

> **NOTE:** all outputs are detailed in the previous section.

5. Additional modes for PhyloDB users

Additional modes were implemented for PhyloDB users (*i.e.* for PHA, PHN, PHC member project) to use ORTHOSKIM along with annotations performed under these projects with Org.Asm assembler. Users can easily use all modes supplied in ORTHOSKIM in complement.

5.1. Sample file

Sample file can be created directly from samples location into the GriCAD infrastructures on the */bettik/LECA/phyloskims/release/* folder. This tab is produced by the `-m phyloskim_indexing` mode. This allowed to screen each sample that will be used for the gene extraction from `-p path/to/seek/files/` . Unwanted samples must be removed from the list before processing other modes.

```
./orthoskim -m indexing -c config_orthoskim.txt -p /bettik/LECA/phyloskims/release/
```

5.2. List of genes files

The extraction of orthologous regions and the creation of databases from annotations are based on a given list of genes. This list is supplied in **$CHLORO_GENES**, **$MITO_GENES** and **$NRDNA_GENES** (lines [63-65] of the config file) and must contain:

- the type of gene (*e.g.* CDS,rRNA,tRNA,misc_RNA)
- the gene name

```
head ~/OrthoSkim/ressources/listGenes.chloro

tRNA    trnV
tRNA    trnA
tRNA    trnN
rRNA    rrn16S
rRNA    rrn23S
rRNA    rrn4.5S
rRNA    rrn5S
CDS     psbA
CDS     matK
CDS     rps16
CDS     psbK
```

By default, ORTHOSKIM provided a list for tRNA, rRNA and CDS genes in chloroplast (see **$CHLORO_GENES** and **$MITO_GENES**). For the ribosomal complex, the gene type correspond to rRNA (*i.e.* for rrn18S, 5.8S rRNA, rrn28S) and misc_RNA (*i.e.* ITS1 and ITS2) (see **$NRDNA_GENES**) as annotated in Org.Asm assembler.

5.3. phyloDB database of references

ORTHOSKIM provides a mode to create a database from the all annotations performed within the project by using the `-m phyloskim_database` mode. To do this, all genes found in these annotations files are extracted with the header restrictions. Output files are created according to the name and the path set in the config file (**$CHLORO_REF_CDS**, **$CHLORO_REF_rRNA**, **$CHLORO_REF_tRNA** and **$NRDNA_REF** at lines 42-44 and 46 of the config file).

> **Note:** For chloroplast annotations, only genes found in single and circular contig will be extracted.

5.4. phyloDB extraction from annotations

For each sample of the sample file, ORTHOSKIM will perform genes extraction directly from annotation with `-m phyloskim_extraction_targeted` mode, according to a list of genes for `-t [chloroplast, nucrdna]` targets.

Results are output in **RES/** directory by creating subdirectories for each compartment and gene type, including a multifasta file per gene. For example, for chloroplast CDS provided in **$CHLORO_GENES**, ORTHOSKIM will output **RES/chloroplast_CDS/** subdirectory with CDS gene files.

```
ls -l ~/RES/chloroplast_CDS/

-rw-r--r--  1 pouchonc  staff   4874 16 avr 10:40 accD.fa
-rw-r--r--  1 pouchonc  staff   4952 16 avr 10:40 atpA.fa
-rw-r--r--  1 pouchonc  staff   4853 16 avr 10:40 atpB.fa
-rw-r--r--  1 pouchonc  staff   1580 16 avr 10:40 atpE.fa
-rw-r--r--  1 pouchonc  staff   2057 16 avr 10:40 atpF.fa
```

6. Funding

The PhyloAlps data collection was largely funded from the European Research Council under the European Community's Seventh Framework Programme FP7/2007-2013 grant agreement 281422 (TEEMBIO), the Sixth European Framework Programme (GOCE-CT-2007-036866), the Swiss SNF (Grant 31003A_149508/1), the ANR DIVERSITALP Project (ANR-07-BDIV-014), ANR project Origin-Alps (ANR-16-CE93-0004), France Génomique (ANR-10-INBS-09-08) and the NextBarcode project (Institut Français de Bioinformatique).

7. Support

For questions and comments, please contact: contact@orthoskim.org