PhyloChromoMap: Phylogenomic chromosome mapping

Contents

1. About:

PhyloChromoMap is a tool for depicting the phylogenetic history of every gene in the chromosomes of an organism. It requires a set of phylogenetic trees, loci of all genes that would be mapped and size of the chromosomes.

PhyloChromoMap was written by Mario Cerón-Romero with support of the laboratory of Laura Katz at Smith College. Questions, suggestions and bug reports can be sent to mceronromero@umass.edu.

2. Download:

PhyloChromoMap can be downloaded from
https://github.com/Katzlab/PhyloChromoMap_py

3. Dependencies:

PhyloChromoMap requires:

- python (https://www.python.org/)
- R and Rscript. To see if you have R correctly installed, type in the command line:

  ```
  >> which r
  ```

  The command should retrieve the path in which you have R. If you don't have R, download it from https://www.r-project.org/.

4. Quick start:

An example database (the genome of the microsporidian *Encephalitozoon hellem*) is included with the PhyloChromoMap distribution. Unzip the file PhyloChromoMap_py.zip and go to the folder PhyloChromoMap_py/source/. Then, execute on the command line:

```
>> python phylochromomap.py
```

The following sections walk through the various inputs and outputs associated with the produced map

5. Input:

a. Files:

   i. Folder with set of phylogenetic trees
   ii. Chromosome size file
   iii. Mapping file
   iv. Parameters file

b. Format

   i. Phylogenetic trees: Trees should be in newick format. The name must contain a unique code (e.g., OG5_126569_outrax.treerenamed, OG5_126572_outrax.treerenamed), henceforth called “tree code”.

PhyloChromoMap was designed for representing very deep phylogenetic history. It relies on detecting three levels in each phylogenic tree: a “major” clade, a “minor” clade and species. In our system, there are 8 major clades: Bacteria, Archaea, 5 eukaryotic “major” clades (i.e., Opisthokonta, Amoebozoa, Excavata, Archaeplastida, SAR) and a group of eukaryotic “orphans” or lineages with undefined placement in the eukaryotic tree of life (e.g., Chryptophytes, Haptophytes, labeled EE).

The user should define the major and minor clades. In our system, Glaucophytes (gl), Green algae (gr) and Red algae (rh) are minor clades of Archaeplastida. You can find our recommended minor clades per major clade at the end of this document.

Finally, the name of every tip in the phylogenetic trees should have the structure MC_mc_sp_seqID, where MC: Major clade, mc: minor clade, sp: species (a 4 letters code such as Hsap: *Homo sapiens*) and SeqID: sequence identifier.

ii. Chromosome size file: This is a file containing the size of the chromosomes in base pairs (bp). The format of the file is csv with two columns: Chromosome and size (e.g., chrI,161584). The file should not have any header.

iii. Mapping file: This file contains the mapping information for every gene. The format of the file is csv with 5 columns: Chromosome, starting position of locus, ending position of locus, protein identifier (e.g., XP_003886649) and tree code. When there is not a tree for a gene, the field should contain “no_tree” instead of a tree code. This file determines what is depicted in the map. The user could import only data of a subset of genes of interest (i.e., If the user is only interested in a particular region of the chromosomes or a strand of the DNA).

iv. Parameters file: The file “parametersFile.txt” should be filled based on the names of the sequences and input files.

6. Output:

a. Files:

i. Map matrix
ii. Map graph

b. Format:

i. Map matrix: This matrix is created for drawing the map using the R module “image”. The first column represents an interval (default: each 1000 bp). The second column says in which interval there is a ‘young’ gene. The next 8 columns are the proportion of minor clades per major clade. Each of these eight columns represents a major clade and the order depends on the major clade of the taxon of interest.

MC of interest	Order of MC in the counts
Op	Op, Am, Ex, EE, Pl, Sr, Za, Ba
Am	Am, Op, Ex, EE, Pl, Sr, Za, Ba
Ex	Ex, EE, Pl, Sr, Am, Op, Za, Ba
EE	EE, Pl, Sr, Ex, Am, Op, Za, Ba
Pl	Pl, EE, Sr, Ex, Am, Op, Za, Ba
Sr	Sr, Pl, EE, Ex, Am, Op, Za, Ba

The next set of 9 columns represents the same information for the second chromosome and so on. The order of the chromosomes is the same than in the chromosome size file.

ii. Map graph: As in the matrix, the order of the chromosomes is the same than in the chromosome size file. The first chromosome is at the bottom and the last chromosome is at the top. For each chromosome the first row (totally black) represents the chromosome itself. The second row (from bottom to top) represents the presence or absence of 'young' genes. The remaining rows are heatmaps representing the proportion of minor clades per major clade (the same order than in the matrix) that contain the gene.

The heatmaps have 4 color tones:

1. 0 – 25 % minor clades (no color)
2. 25% – 50 % minor clades (lightest color)
3. 50% – 75 % minor clades
4. 75% – 100 % minor clades (Darkest color)

7. **Optional taxa structure:**

Our advice to users is to adapt our naming system without changing its structure. For example, if you are studying *Drosophila melanogaster* and your trees only consider Arthropods, you can use our major clade codes as:

MC code	MCs in Arthropoda
Ba	Nothing
Za	Unclassified Arthropoda
Sr	Arachnida
Pl	Merostomata
EE	Pycnogonida
Ex	Unclassified Mandibulata
Am	Myriapoda
Op	Pancrustacea

Here we see that there are 7 major clades for Arthropoda and the default number is 8. So, in order to keep the same structure, the user can keep 'Ba' and specify any number of minor clades for Ba in the parameters file (greater than 0 in order to avoid a division error).

8. **Centromeres:**

   a. Map centromeres: The user can map the centromeres in each chromosome using the script “mapCentromeres.py”. The input file is the chromosome size file with two extra columns, the starting and ending positions of the centromere. A line of the chromosome file would look like:

   ```
   chrI,230218,151465,151582
   ```

   The output would be a new map and a new matrix with the information of the centromere per chromosome. In the map, the centromere will be colored in red.

   b. Example: An example database (the genome of *Saccharomyces cerevisiae*) is included with the PhyloChromoMap distribution for testing the centromere mapping method. Follow the next steps to run a quiz analysis:

      i. Set the path to the folder “S_cerevisiae_CentroANDhypo” in the field “path to files:” of the parameters file.

      ii. Run phylochromomap.py:

      ```
      >> python phylochromomap.py
      ```

      iii. Run mapCentromeres.py:

      ```
      >> python mapCentromeres.py
      ```

9. **Hypotheses of conservation:**

   a. Map hypotheses: The use can also explore hypotheses of conservation such as “the gene should be present in all major clades”. The script map_hypotheses.py maps all genes that meet that criterion. This script allows to map a variable number of hypotheses (less than 10 is recommended).

   b. Input files:

      i. Files and format: This analysis requires the outputs of phylochomomap.py and a file listing the hypotheses of conservation (hypotheses.csv). The first column in hypotheses.csv contains the hypotheses. The second column contains the colors for the genes that meet the criterion for each hypothesis. Columns from the third to the last specify the minimal number (as frequency) of minor clades per major clade for considering a major clade as present. The default value for each major clade is 0.25 and the order of the numbers depend on the major clade of studied lineage (see section 6). Then, a line in hypotheses.csv should look like:

      ```
      Sr;Pl;Op;EE;Am;Ex;Ba;Za,red,0.25,0.25,0.25,0.25,0.25,0.25,0.25,0.25
      ```

ii. Hypothesis notation:

- Clades to be evaluated as present or absent should be separated with “;” (e.g., Am;Ex)
- Absence should be specified with “*” (e.g., *Sr)
- For a clade that is either present or absent use “?” (e.g., ?Op)
- The presence of a number of major clades from a group would be specified using the signs “[]”, “|”, “+” and “-”. For instance, [Sr|Pl|Op]+2 means that at least 2 of the major clades should be present. In contrast, [Sr|Pl|Op]-2 means that at most 2 of the major clades should be present. Finally, [Sr|Pl|Op]2 means that the 3 major clades should be present.

These are some examples of hypotheses:

- **Sr;Pl;Op;EE;Am;Ex;Ba;Za (Gene from LUCA)**: All clades present.
- **Sr;Pl;Op;EE;Am;Ex;*Ba;*Za (Gene from LECA)**: All clades except Ba and Za are present.
- **[Sr|Pl|Op|EE|Am|Ex]+5;[Ba|Za]2 (Gene from LUCA, relaxed)** : From the group composed by Sr,Pl,Op,EE,Am and Ex at least 5 are present. Moreover, both Ba and Za are also present.
- **[Sr|Pl|EE]+2;[Op|Am|Ex]0;Ba;*Za (Photosynthetic gene)** : From the group composed by Sr, Pl and EE at least 2 are present. Op, Am and Ex are absent. Ba is present and Za is absent.

c. Output files: There will be two output files after running this analysis: a new map and a new matrix with the information after testing all the hypotheses in each gene. In the map, if a gene meet the criterion for an hypothesis, the gene will be colored with the corresponding color (set in the file hypothesis.csv).

d. Example: The database in the folder S_cerevisiae_CentroANDhypo also allows to test the hypotheses of conservation mapping method. Follow the next steps to run a quiz analysis:

i. Set the path to the folder “S_cerevisiae_CentroANDhypo” in the field “path to files:” of the parameters file.

ii. Run phylochromomap.py:

```
>> python phylochromomap.py
```

iii. Run map_hypotheses.py:

```
>> python map_hypotheses.py
```

10. Appendix. Our system of major and minor clades

Code (MC_mc)	Major clade (MC)	Minor clade (mc)
Am_ar	Amoebozoa	Archamoebae
Am_di	Amoebozoa	Discosea
Am_my	Amoebozoa	Mycetozoa
Am_hi	Amoebozoa	Himatismenida
Am_is	Amoebozoa	incertaesedis
Am_th	Amoebozoa	Thecamoebida
Am_tu	Amoebozoa	Tubulinea
Am_va	Amoebozoa	Vannellidae
EE_ap	Orphans (Enything else)	Apusozoa
EE_br	Orphans (Enything else)	Breviatea
EE_cr	Orphans (Enything else)	Cryptophyta
EE_ha	Orphans (Enything else)	Haptophyceae
EE_is	Orphans (Enything else)	incertaesedis
EE_ka	Orphans (Enything else)	Katablepharidophyta
Ex_eu	Excavata	Euglenozoa
Ex_fo	Excavata	Fornicata
Ex_he	Excavata	Heterolobosea
Ex_is	Excavata	incertae sedis
Ex_ja	Excavata	Jakobida
Ex_ma	Excavata	Malawimonadidae
Ex_ox	Excavata	Oxymonadida
Ex_pa	Excavata	Parabasalia
Op_ch	Opisthokonta	Choanoflagellida
Op_fu	Opisthokonta	Fungi
Op_ic	Opisthokonta	Ichthyosporea
Op_is	Opisthokonta	incertae sedis
Op_me	Opisthokonta	Metazoa
Op_nu	Opisthokonta	Nucleariidae and Fonticula group
Pl_gl	Plantae	Glaucophytes
Pl_gr	Plantae	Green algae
Pl_rh	Plantae	Red algae
Sr_ap	SAR	Apicomplexa
Sr_ch	SAR	Chromerida
Sr_ci	SAR	Ciliates
Sr_di	SAR	Dinoflagellates
Sr_is	SAR	incertae sedis
Sr_pe	SAR	Perkinsea
Sr_rh	SAR	Rhizaria
Sr_st	SAR	Stramenopiles