NESCENT ACADEMY

*Hands-on instruction for your evolutionary toolbox*

# Phylogenetics - Orthology, phylogenetic experimental design and phylogeny reconstruction

## Lesser Tenrec (*Echinops telfairi*)

**Goals:**

**1. Use phylogenetic experimental design theory to select optimal taxa to sequence the genome of, in order to infer the place of your non-model organism within a "phylogenomic" phylogeny.**

**2. Use phylogenetic experimental design theory on phylogenomic data to select optimal genes to sequence for inference of a phylogeny.**

## Steps:

**1 - Identify single-copy orthologous genes for phylogenetics**

**2 - Select optimal taxa and build a phylogenetic tree with Phyml**

**3 - Construct a chronogram with PATHd8 (or r8s).**

**4 - Identify highly informative genes with PhyDesign**

**5 - Test the performance of your selected genes.**

**0. Software needed:**

**- Phyml**: software that estimates maximum likelihood phylogenies from alignments of nucleotide or amino acid sequences.

Download: http://code.google.com/p/phyml/

Online version: http://www.atgc-montpellier.fr/phyml/

**- PATHd8:** program for estimating divergence times on a phylogenetic tree

http://www2.math.su.se/PATHd8/

**- or r8s**: another program for estimating divergence times on a phylogenetic tree

http://loco.biosci.arizona.edu/r8s/

**- PhyDesign:** webbapp to estimate phylogenetic informativeness profiles

http://phydesign.townsend.yale.edu/

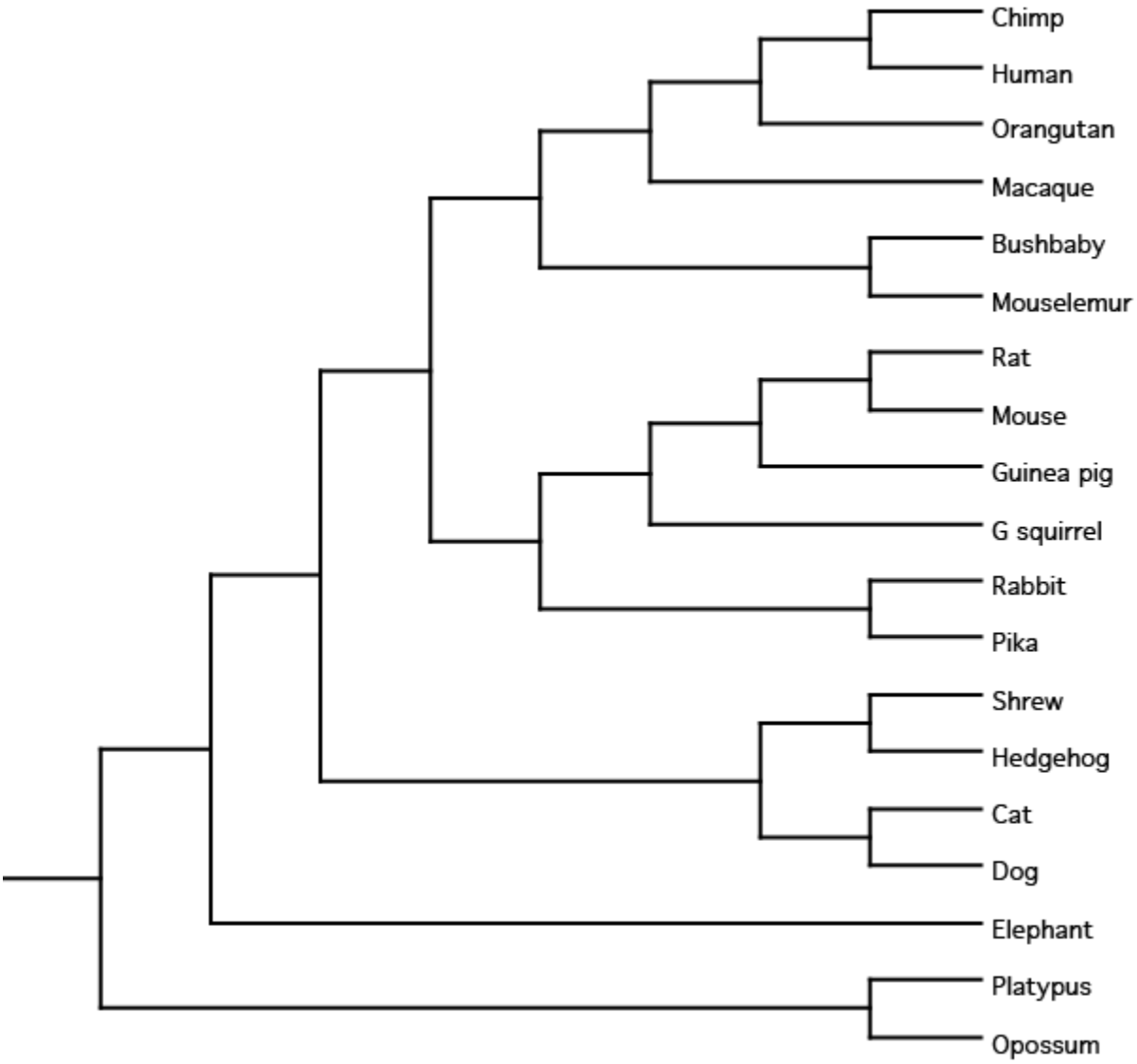
**- ReadSeq:** online alignment conversion tool

http://www.ebi.ac.uk/cgi-bin/readseq.cgi

- **Geneious**: you can use this software to visualize, reroot, and build with PhyML pluggin phylogenetic trees. You can also use it to convert alignments into different formats. We will work with NEXUS and phylip formats.

http://www.geneious.com/

Previous published studies based on gene data in the literature give you this tree




• To discover the phylogenetic placement of the Tenrec, you have been granted money to sequence the Tenrec genome, plus 9 more!

• We'll provide you with the single copy orthologous genes among all sequenced genomes: build a phylogenetic tree with your 9 selected species plus the Tenrec.

## 1. Selecting single copy orthologous genes:

Single-copy genes are useful markers for phylogenetic inference.

Some vocabulary (https://www.msu.edu/~jhjacksn/Reports/similarity.htm)
Genes can be....
- Homologs: Homologs have common origins but may or may not have common activity.
- Analogs: Analogs have common activity but not common origin
- Paralogs: Paralogs are homologs produced by gene duplication. They diverged after duplication event.
- Orthologs: Orthologs are homologs produced by speciation. They diverged after speciation event.
- Xenologs: Xenologs are homologs resulting from horizontal gene transfer.

We are not going to do an activity for this step of the workshop. Please, check the wiki for a presentation on the topic. Large-scale orthology assignment would need its own workshop. There is no fully automated (and easy) way to assign orthology among multiple genomes.
Here, you have three software packages that one can use to assign orthology:

- OrthoInspector: http://lbgi.igbmc.fr/orthoinspector/
- BrunchClust: http://www.bioinformatics.org/branchclust/index.html
- OrthoMCL: http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=orthomcl#Software

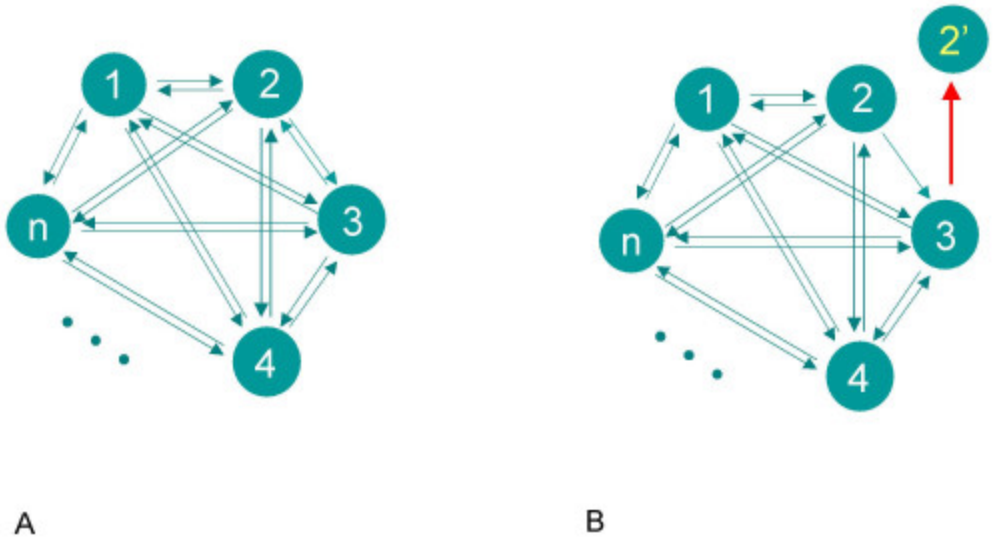
There are also databases which offer tools to assign orthology but most of them are limited to pairwise comparisons and to a limited number of species.
Ex:
- Cluster of Orthologous Groups (COG): http://www.ncbi.nlm.nih.gov/COG/
- Eukaryotic Gene Orthologues (EGO): http://compbio.dfci.harvard.edu/tgi/ego/
- Inparanoid eukaryotic orthologs database: http://inparanoid.sbc.su.se/
- ENSEMBL: http://www.ensembl.org/info/docs/compara/homology_method.html

Some of the methods for determining orthology include the following: Reciprocal Blast Hit (a common step for most of the methods), conservation of gene order in corresponding syntenic blocks and phylogenomics (tree-based) methods, among others.

**Try Reciprocal BLAST:**
Let's assume were interested in the TRPM5 protein from mouse. This gene is involved in signaling of sweet and amino acid taste perception in the tongue.
- Go to the NCBI BLAST page and select Genomes -> Human
- Select blastp tab and enter the following GI number: 12383054
- Select Database: Build Protein and click ”BLAST.”

Look at the list of e-values. It looks like there are several significant hits. Scroll down to see the details of each hit. How does the top hit compare to the rest of them?
Now take the GI number of putative ortholog and BLAST it against the human protein build. (Repeat steps 1-4 with the mouse genome) Do you get a reciprocal best hit? In other words, do you return to your original query, the mouse TRPM5 protein?

Now repeat your reciprocal BLAST search with the mouse V2R2 protein sequence, with GI: 9910304 . How do you explain the results?




**The reciprocal best BLAST hit method**. Circles represent genes from n different taxa, arrows signify best BLAST hit relationship; (A) – case of strict reciprocity, (B) – failing of reciprocity in the presence of paralogs. Poptsova and Gogarten *BMC Bioinformatics* 2007 **8**:120

## 2. Building phylogenetic tree with PhyML with single-copy

## orthologous genes:

Online version of PhyML
http://www.atgc-montpellier.fr/phyml/

Or using Geneious when you install the plugin for PhyML.

It can be run by menus just typing:

```
$ phyml
```

or command-line:

```
$ phyml -i <YOUR_FILE> -b 0 -m GTR -f m -v e -c 4 -a e -s NNI -o tlr
```

- Compare new tree with neighbour. What is the sister species of the Tenrec?

Let's imagine that you realize you've all been sequencing related genomes. You get together with all your collaborators and you get all 20 genomes and their single-copy orthologs. Build a phylogenetic tree with the full set of species.
We will use this new phylogenetic tree to build a chronogram.

## 3. Constructing a chronogram with PATHd8:

You spend all your money sequencing genomes but you still have work to do.
Now you have only money for PCR and sanger sequencing for 1 gene for the many other taxa within this clade.
Which gene to select? Write down your choice. Now, we can use PhyDesign for selecting genes.

To use PhyDesign you need:

- an alignment with your genes of interest (YOU HAVE THAT)
- an ultrametric tree/chronogram

### 3.1. Preparing input file for PATHd8:

Input is given as a phylogenetic tree with branch lengths. An arbitrary number of age constraints can be specified, either as fixage, minage or maxage. The output contains the chronogram obtained from the PATHd8 analysis as well as a list of estimated node ages, their mean path lengths, and their estimated substitution rates. The output also contains results from the MPL analysis described in Britton et al. (2002), which can be of interest when there are no fossil datings or if the clock hypothesis is of interest (a clock test is performed at each node).

Example of an input file:

```
Sequence length = 1823;
((((Rat:0.007148,Human:0.001808):0.024345,Platypus:0.016588):0.01292,(
Ostrich:0.018119,Alligator:0.006232):0.004708):0.028037,Frog:0);
mrca: Rat, Ostrich, minage=260;
mrca: Human, Platypus, fixage=125;
mrca: Alligator, Ostrich, minage=150;
name of mrca: Platypus, Rat, name=crown_mammals;
name of mrca: Human, Rat, name=crown_placentals;
name of mrca: Ostrich, Alligator, name=crown_Archosaurs;
# my own notes, not executed by PATHd8
# fossil Archosaurus minage=260
# fossil Archaeopteryx minage=150
# fossil Eomaia minage=125
```

- Take the phylogenetic tree with all species and root the tree with the outgroup (marsupials). To build a proper chronogram, you would remove the outgroup from the tree once rooted because we don't know where to root the outgroups. However, it is not a critical step for this exercise.
- Paste the rooted tree in the PATHd8 input file (you have sample files in the workshop folder).
- You also need to describe calibration points in the PATHd8 input file with the commands MRCA (Most Recent Common Ancestor). Follow the PATHd8 sample file as an example.

3.2. Running PATHd8 with the file created:

```
$ PATHd8 <infile> <outfile>
```

PATHd8 will output a chronogram. Copy the tree (only the tree) and save it as text file.

## 4. Selecting genes for phylogenetic analyses with PhyDesign:

**http://phydesign.townsend.yale.edu/**

Input your alignment in NEXUS format and your chronogram in newick format. In the NEXUS file you have the coordinates of all genes.

Please, visit the instructions and FAQs online:
http://phydesign.townsend.yale.edu/instructions.html
http://phydesign.townsend.yale.edu/faq.html

Choose 1 gene. In reality you would sequence the selected gene in some novel taxa, but for sake of convenience for this exercise, we're going to reconstruct the tree for the same taxa, but just with one gene. Also, for the sake of this exercise, please choose the gene you think will perform best, and the gene you think it will perform worst. Infer the phylogenies and compare them. Individual files for each in FASTA format are provided in the workshop folder.