# Phylogenomic analysis of Australasian marsupials

## Background

Marsupials form one of the three major groups of mammals (along with monotremes and placentals) and are native to Australasia and the Americas. Their name comes from their most distinctive anatomical feature, the marsupium, which is an abdominal pouch that carries and protects the young offspring.

There are about 334 extant marsupial species, grouped into 22 families. Australasia is home to 16 marsupial families, comprising more than two-thirds of all known species. There is convincing evidence that the Australasian families form a natural or 'monophyletic' group, known as Eomarsupialia. In other words, all Australasian marsupials are thought to be descended from a single common ancestor, to the exclusion of all other marsupials. Fossil and genetic evidence suggests that this ancestor lived about 60–70 million years ago.

Yellow-footed rock-wallaby (*Petrogale xanthopus*). Photo by Simon Ho.

In the past, evolutionary relationships were reconstructed on the basis of anatomical and skeletal characteristics, but genetic data are now widely used for this purpose. Major advances in DNA-sequencing technology have allowed researchers to assemble very large genetic data sets for analysing evolutionary relationships. Nevertheless, there are still some parts of the marsupial phylogeny that have remained particularly difficult to resolve.

The marsupial moles (genus *Notoryctes*) are a particularly enigmatic group. The genus is placed in its own family, Notoryctidae, and contains two species: the northern and southern marsupial moles. These animals live underground and only rarely appear on the surface. Because of their fossorial (burrowing) lifestyle, marsupial moles have evolved an unusual and highly specialised morphology that has made them difficult to place in the marsupial phylogeny.

Southern marsupial mole (*Notoryctes typhlops*). Illustration by Richard Lydekker.

There is also uncertainty about the relationships among the different groups of Australasian possums. Most researchers agree that these marsupials can be classified into two superfamilies: Phalangeroidea (brushtail possums, cuscuses, and pygmy possums) and Petauroidea (gliders and ringtail possums). The two possum superfamilies are believed to have a close relationship with Macropodiformes (bettongs, kangaroos, potoroos, wallabies, and allies), but the exact evolutionary relationships among these three marsupial groups have proven to be difficult to resolve with any confidence.

Recent studies have produced large amounts of genetic data from marsupials, providing unprecedented opportunities for reconstructing the phylogeny of this group. Such genome-scale data sets offer a rich source of information about evolutionary history, but they also present substantial challenges for analysis. In terms of phylogenetic analyses, the chief difficulty is that different genes can support different sets of evolutionary relationships. This is because recombination breaks the links between genes, allowing them to follow separate incongruent evolutionary histories. For this reason, the phylogenetic trees inferred from individual genes ("gene trees") might differ from each other and from the actual relationships among the species of interest (the "species tree"). This is sometimes referred to as discordance between gene trees and the species tree.

In this exercise you will analyse a sample of genes (taken from a much larger data set) to examine signals of evolutionary relationships. In Part 1, you will estimate and compare the phylogenies inferred using DNA sequences from two individual genes (gene trees). In Part 2, you will analyse the results from a collection of genes in order to extract their collective phylogenetic signal (the species tree) and to evaluate the power of genomic data to resolve evolutionary relationships.

## Data and Software

In the *Data* folder, you should have the following four data files for this exercise:

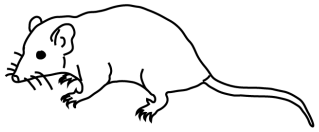
- ***gene_ddo.fasta*** contains the aligned partial DNA sequences (546 bp) of the *DDO* gene from 45 marsupials. This gene encodes the enzyme D-aspartate oxidase. The sequences are in Fasta format, which is a simple, widely used data format. Each DNA sequence has a label (indicated by a '>' symbol) on the preceding line.
- ***genetree_ddo.pdf*** shows the gene tree estimated from the *DDO* gene.
- ***marsupials.500genes.tre*** contains 500 gene trees, each estimated from a single gene. The trees are in Newick format, which is a simple format that depicts an evolutionary tree using commas and parentheses. The tree also contains information about the lengths of its branches, and support values for the nodes in the tree.

You will need three computer programs for this exercise:

- **MEGA** (https://www.megasoftware.net/), version 7 or above. This is a user-friendly program that can perform a range of population genetic and phylogenetic analysis.
- **ASTRAL** (https://github.com/smirarab/ASTRAL) can infer a species tree given a set of gene trees. This program has been included in the *Data* folder. It consists of the Java JAR file *astral.5.6.3.jar* and everything in the folder *lib*.
- **FigTree** (https://github.com/rambaut/figtree) is a popular program for viewing phylogenetic trees.

## AMERICAS

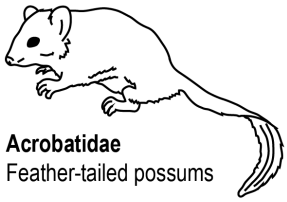
**Didelphidae**
Pouched opossums



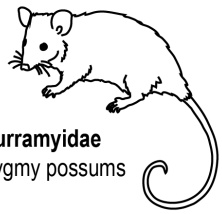
**Marmosidae**
Mouse opossums

**Microbiotheriidae**
Monito del monte

## AUSTRALASIA



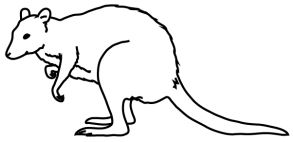
**Acrobatidae**
Feather-tailed possums



**Burramyidae**
Pygmy possums

**Dasyuridae**
Quolls, antechinus, dunnarts, Tasmanian devil, and allies

**Hypsiprymnodontidae**
Musky rat-kangaroo

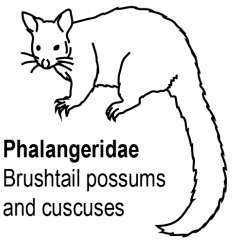


**Macropodidae**
Wallabies and kangaroos

**Myrmecobiidae**
Numbat

**Notoryctidae**
Marsupial moles

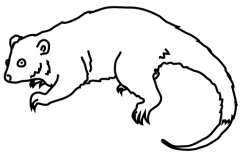
**Peramelidae**
Bandicoots



**Phalangeridae**
Brushtail possums and cuscuses

**Phascolarctidae**
Koala

**Potoroidae**
Bettongs, potoroos, and rat-kangaroos



**Pseudocheiridae**
Ring-tailed possums

**Tarsipedidae**
Honey possum

**Thylacomyidae**
Bilby

**Vombatidae**
Wombats

Three American and 15 Australasian marsupial families analysed in this exercise.

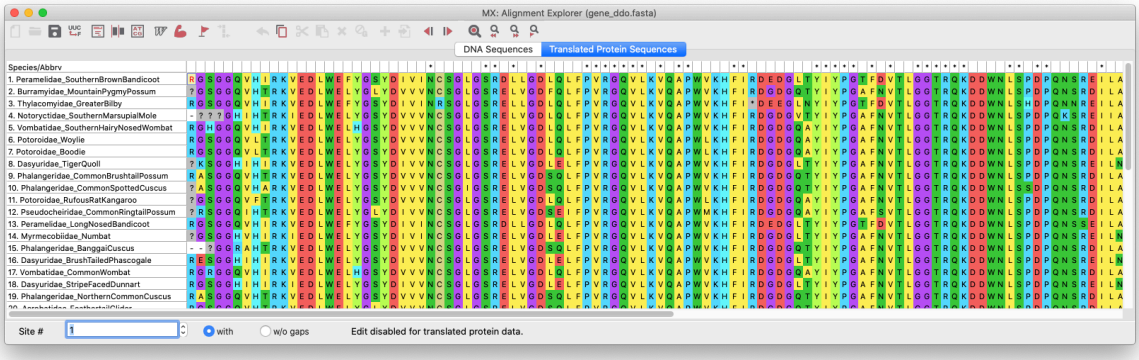
## Part 1: Phylogenetic analysis of a single gene

In this exercise you will use the free phylogenetic software MEGA. This software is able to implement a wide range of genetic analyses, including several different methods for phylogenetic analysis. We will be using the GUI version for convenience, but a command-line version is also available.

a) Open the program MEGA.

b) Open the file *gene_ddo.fasta* in the *Data* folder, either via **File** -> **Open A File/Session** or by dragging and dropping the data file into the MEGA window.

c) When prompted, select **Analyze** and confirm that you are working with nucleotide sequences from a protein-coding gene that uses the standard genetic code.




MEGA window showing the translated sequences of the *DDO* gene.

We will infer the evolutionary relationships among marsupials using a phylogenetic method based on maximum likelihood. This is a statistical approach that was first applied to phylogenetic analysis in the 1970s and formalised in the early 1980s. In maximum-likelihood analysis, we search for the phylogenetic tree that produces the highest likelihood score.

As part of the analysis you will also run "bootstrap" replicates to estimate the statistical support for the relationships in the tree. This method involves resampling from the sequence alignment and then re-estimating the tree a number of times. If a particular grouping appears in all of these replicates of the tree, then the node in the tree is said to have 100% bootstrap support.

d) In the main MEGA window, click on the icon for **Phylogeny**, and select **Construct/Test Maximum Likelihood Tree**. Check that you have the following settings for the analysis:
Test of Phylogeny -> Bootstrap method
No. of Bootstrap Replications -> 100
Substitutions Type -> Nucleotide
Model/Method -> Hasegawa-Kishino-Yano model
Rates among Sites -> Gamma Distributed (G)
No of Discrete Gamma Categories -> 4

The rest of the settings can be left at the default choices. Click on **OK** to start the analysis, which will run for a few minutes.

e) The estimate of the phylogeny will appear in a new window. Check that the root of the tree is placed between the American opossum families (Didelphidae and Marmosidae) and the rest of the taxa. If this is not the case, select the branch leading to these two taxa and select **Subtree** -> **Root**.

Have a look at the tree and examine the relationships among the marsupial families. If you are not familiar with marsupials, you might find the figure on page 3 helpful. In the tree, the numbers next to the nodes represent percentage support values from the bootstrap analysis, with values >80% indicating strong support.

In this exercise we are particularly interested in the phylogenetic placement of the marsupial mole (family Notoryctidae), as well as the relationships among the Phalangeroidea (brushtail possums, cuscuses, and pygmy possums), Petauroidea (gliders and ringtail possums), and Macropodiformes (bettongs, kangaroos, potoroos, wallabies, and allies).

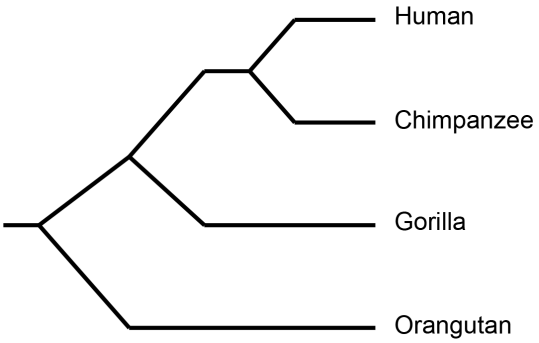
f) You can now close MEGA. If you want to look at the gene tree again, you can see one that was prepared earlier in the file *genetree_ddo.pdf*.

Note that you have analysed DNA sequences from a single gene: the inferred phylogeny is known as a gene tree. This tree might not represent the actual relationships among the species that are being studied, because individual gene trees can differ from each other and from the species tree.

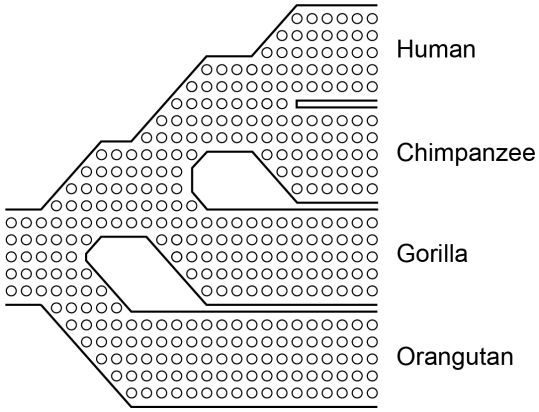
To gain a more reliable estimate of the relationships among marsupial families, we need to increase the size of the data set. In the next part of this exercise you will analyse a much larger collection of gene trees using a method that can combine their phylogenetic signals to produce a single estimate of the species tree.

## Part 2: Phylogenetic analysis of multiple gene trees

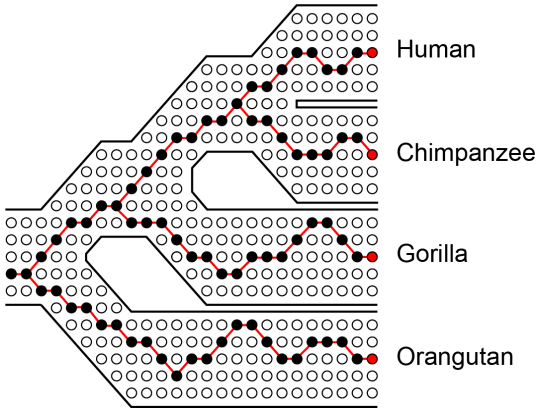
When we think of the evolutionary relationships among species, we typically envisage phylogenetic tree similar to the one on the right. This is what we consider to be the *species tree* and is usually what we wish to estimate when we perform phylogenetic analyses of DNA sequence data.



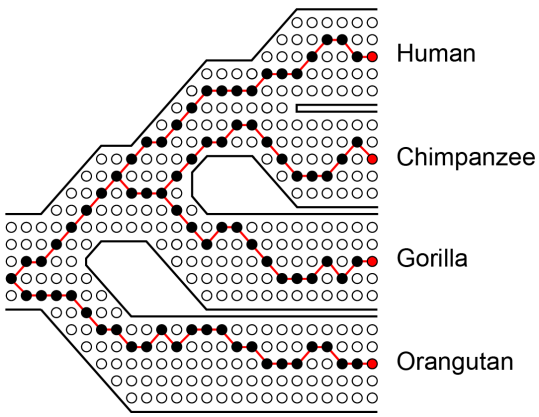

However, when we are tracing the histories of individual genes, it is more appropriate to consider the species tree as consisting of populations through time, as shown on the right. In this depiction, each circle represents an individual, and each column of circles represents the population at a particular point in time (i.e., one generation). The black lines represent the species boundaries.




We now start with our sampled individuals (red circles) and trace their ancestry back in time. Each sampled lineage chooses a parent at random from the previous generation. Individuals cannot cross species boundaries (black lines), which represent reproductive barriers. When two lineages choose the same parent, they 'coalesce'. This process continues until all of the lineages have coalesced into a single ancestral lineage.




By chance, each gene will trace out a slightly different history. Some genes will trace out histories that suggest relationships that differ from the species tree. In the example on the right, the gene tree shows a close relationship between the chimpanzee and gorilla, to the exclusion of human. Thus, the gene tree is discordant with the species tree. This is known as incomplete lineage sorting.



In this part of the exercise you will use the software ASTRAL to infer the species tree from a collection of gene trees. Estimating the gene trees from the individual genes can be time-consuming, but this step has already been done for you. The file *marsupials.500genes.tre* contains 500 gene trees. Each of these trees has been estimated from the nucleotide sequences of a single gene, sampled from the same 45 marsupials that were analysed in Part 1.

a) Run FigTree and open the file *marsupials.500genes.tre*.

Have a look at some of the trees in this file by using the left/right arrows at the top of the window. You might notice a lot of incongruence among the gene trees. For example, look out for the placement of the marsupial mole (family Notoryctidae) in the first 10 gene trees. In general, the relationships in the individual trees are poorly supported.

The software ASTRAL implements a very rapid method that can infer the overall species tree from a collection of gene trees. It does this by looking at all of the 'quartets' of taxa that are supported by each gene tree, then finding the species tree that offers the best agreement with these quartets. Although this approach sounds simple, the method has been shown to perform very well under a range of conditions and is widely used in phylogenomic studies.

b) Open a command prompt (Windows) or a Terminal window (Mac) and change directory to the folder containing all of the files that you are using for this exercise. You can do this by typing:
   cd
   and then dragging and dropping the folder into the command-line window.

c) Once you are in the folder that contains ASTRAL (astral.5.6.3.jar), you can run ASTRAL using the following command:
   java -jar astral.5.6.3.jar -i marsupials.500genes.tre -o speciestree.500genes.tre

d) Use FigTree to open the output file from ASTRAL,*speciestree.500genes.tre*. When prompted, type "probability" into the dialogue box. Select the branch leading to the American opossum families (Didelphidae and Marmosidae) and click on **Reroot**. Check the box next to **Node Labels**, then expand the options by clicking on the triangle to the left of the check-box. In the drop-down menu next to **Display**, select **probability**. This will show the support values for the nodes in the phylogenetic tree. Keep this display open.

In this exercise you have seen how individual gene trees can present a misleading depiction of the evolutionary relationships among species. Analysing a set of 500 gene trees shows support for some unusual relationships among the possum groups. But even with a data set of this size, some of the relationships among the marsupial families are not able to be determined with complete confidence.

A more comprehensive analysis of a larger data set (1550 genes), published by David Duchêne, Simon Ho, and colleagues in 2018, was able to resolve all of the evolutionary relationships among Australasian marsupial families. Analysis of the full data set confirms the relationships that you found in the ASTRAL analysis here.