Practical 2.2: Phylogenomic Analysis of Australasian Marsupials

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 monophyletic group, known as Eomarsupialia. Fossil and genetic evidence suggests that their most recent common ancestor lived about 60–70 million years ago. Nevertheless, there are still some parts of the marsupial phylogeny that have remained 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.

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

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 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 Section A, you will estimate and compare the phylogenies inferred using DNA sequences from two individual genes (gene trees). In Section B, you will analyse the results from a collection of genes in order to extract their collective phylogenetic signal (the species tree).

Data and Software

You will need the following four data files for this exercise:

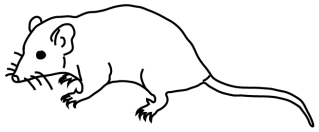
- **gene_ddo.fasta** contains the aligned nucleotide sequences of part of the *DDO* gene from 45 marsupials. This gene encodes the enzyme D-aspartate oxidase.
- **gene_ubox.fasta** contains the aligned nucleotide sequences of the *UBOX* gene, which encodes RING finger protein 37, from the same 45 taxa.
- **marsupials.100genes.tre** contains 100 gene trees, each estimated from a single gene. The trees are in Newick format and include node support values.
- **marsupials.500genes.tre** contains 500 gene trees, each estimated from a single gene.

This exercise will require three computer programs:

- ***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.
- ***FigTree*** (https://github.com/rambaut/figtree) is a popular program for viewing phylogenetic trees.

AMERICAS

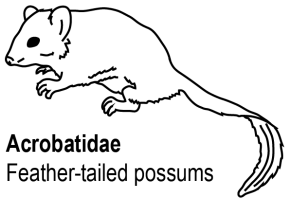
Didelphidae
Pouched opossums



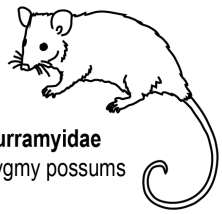
Marmosidae
Mouse opossums

Microbiotheriidae
Monito del monte

AUSTRALASIA



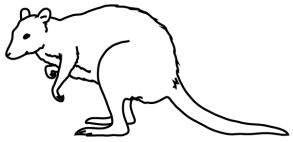
Acrobatidae
Feather-tailed possums



Burramyidae
Pygmy possums

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

Hypsiprymnodontidae
Musky rat-kangaroo

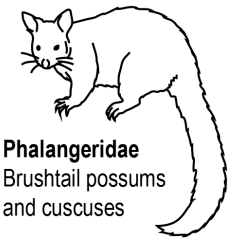


Macropodidae
Wallabies and kangaroos

Myrmecobiidae
Numbat

Notoryctidae
Marsupial moles

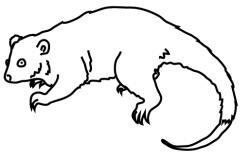
Peramelidae
Bandicoots



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.

Section A: Phylogenetic analysis of individual genes

- Load the file **gene_ddo.fasta** in *MEGA* by dragging and dropping the data file into the *MEGA* window.
- When prompted, select "Analyze". Note that the data set consists of nucleotide sequences of a protein-coding gene (using the Standard genetic code).
- To infer the gene tree using maximum likelihood, 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.

Q. *When estimating the evolutionary relationships among highly divergent organisms (such as the relationships among marsupial families), why might we choose to focus on protein-coding genes rather than on non-coding regions of the genome?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

- 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.

Q. *What is the sister group to the marsupial mole (family Notoryctidae)?*

. . . . . .

Q. *What is the sister group to the Petauroidea (families Pseudocheiridae, Acrobatidae, and Tarsipedidae), which includes the gliders and ringtail possums?*

. . . . . .

. . . . . .

. . . . . .

- Save the tree by going to the "Image" menu and saving it in your preferred image format, in case you need to look at the tree again later.

  Note that you have analysed nucleotide 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.

  Now you should look at nucleotide sequences from a second gene to see whether they supports the same set of relationships as the gene that you have just analysed. Use *MEGA* to open the file **gene_ubox.fasta**, which contains the sequences of a second gene from the same 45 marsupials. In *MEGA*, repeat all of the steps above in order to estimate the gene tree from this data set.

Q. *In terms of the relationships that you described on the previous page, are there any differences between the two gene trees?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *What are some potential reasons for the incongruence between the two gene trees?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Section B: Phylogenetic analysis of multiple gene trees

When analysing a large collection of genes, for example in a phylogenomic analysis, we encounter gene-tree incongruence on a large scale. In Section A of this practical exercise, you saw that gene trees can conflict with each other. But even though gene trees might be discordant with each other and with the species tree, all of the gene trees are 'embedded' in the same species tree. Therefore, if we have enough gene trees, we can still reconstruct the underlying species tree.

In this section 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.100genes.tre* contains 100 gene trees for the same 45 marsupials that were analysed in Section A.

- Open the file **marsupials.100genes.tre** in *FigTree*.
- 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.

- Put the files **marsupials.100genes.tre** and **marsupials.500genes.tre** into the folder that contains *ASTRAL*.
- Open a command prompt (Windows) or a Terminal window (Mac) and change directory to the folder containing ASTRAL. You can do this by typing `cd` and then dragging and dropping the folder into the command-line window.
- Once you are in the folder that contains *ASTRAL* (astral.5.7.4.jar), you can run *ASTRAL* using the following command:
  `java -jar astral.5.7.4.jar -i marsupials.100genes.tre -o speciestree.100genes.tre`
- Use FigTree to open the output file from *ASTRAL*, **speciestree.100genes.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.

Q. *What is the sister group to the Petauroidea (families Pseudocheiridae, Acrobatidae, and Tarsipedidae), which includes the gliders and ringtail possums? Is this relationship strongly supported?*

. . . . . .

. . . . . .

. . . . . .

- Now go back to the command prompt and analyse the larger set of 500 gene trees by typing the following command:
  java -jar astral.5.6.3.jar -i marsupials.500genes.tre -o speciestree.500genes.tre
- 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.

Q. *Now what is the sister group to the Petauroidea (families Pseudocheiridae, Acrobatidae, and Tarsipedidae), which includes the gliders and ringtail possums? Is this relationship strongly supported?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Q. *What is the sister group of the marsupial mole (family Notoryctidae)? Is this relationship strongly supported?*

. . . . . .

. . . . . .

. . . . . .

. . . . . .

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 100 gene trees begins to show some support for some unusual relationships among the possum groups, but a different set of relationships are supported when more genes are added to the data set. Even with a large collection of 500 gene trees, 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) by Duchêne et al. (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 of 500 gene trees. A follow-up study using a genus-level sample of Australasian marsupials is now being led by Mezzalina Vankan at the University of Sydney, as part of the Oz Mammals Genomics initiative.