# MSc phylogenetics practical

In this practical we will carry out some simple analyses that use phylogenies (evolutionary relationships inferred from genetic sequences) to answer some interesting biological questions. You’ll be using Linux because 90% of bioinformatics servers also do, and running it on the Raspberry Pi because we think they’re cool :-)

Throughout the practical, commands you should type into the terminal window will be rendered in this font, e.g.:

MacBook-Pro:pi\_biolinux\_cluster joe$ echo 'hello, world!'

hello, world!

Some of the datasets in this practical have been simulated so that no two workstations will give identical answers to every question. If you have questions during the practical, ask your instructors. They are:

Joe Parker

Jim Clarkson

Alex Papadopulos

Note: the datasets and other files needed for this practical can be found at <https://github.com/lonelyjoeparker/MSc-phylogenetics-lab>. If you need a hint, a cheatsheet is at ‘~/Desktop/phylogenetics/commands.sh’

## Part 0 – Getting started (10-15mins)

*# Yes, most things in computing are numbered from 0… n-1, not 1… n. Get used to it!*

You should already have read [“How to read a phylogenetic tree”](http://epidemic.bio.ed.ac.uk/how_to_read_a_phylogeny) by Andrew Rambaut, and “[Basic Shell Commands](http://www.hongkiat.com/blog/basic-shell-commands-for-bloggers/)” (which you can run interactively on [Tutorialspoint](http://www.tutorialspoint.com/execute_bash_online.php)). Let’s get started:

1. Your desktop should have a folder visible called ‘phylogenetics’. This is where your data will live, but we’ll be using a window that connects directly with the operating system (a ‘terminal’) which uses the bash language. Open a terminal now, by holding the Ctrl-Alt-T keys together.
2. You should now have a terminal window open. The black screen you can see looks like it dates from the 1970s because it does! In fact, pretty much all computers, deep down, date from the 1970s, because not much has changed at that level since then.
3. The line you can see with a blinking cursor at the end is called a *prompt.* It is split into three parts, telling us:

**User@user-pi:~$** (username @ computer name: directory$)

1. “~” is a funny name for a directory, which you’ve probably not seen before. We can use the ‘pwd’ command (‘**p**rint **w**orking **d**irectory’) to find out where we really are. Type that now, followed by a return (you always type return to enter your input).
2. We want to move to the directory where the data is, that ‘phylogenetics’ folder you can see on the desktop. To change directories in Linux, Unix, or Mac terminals, we use the ‘cd’ command, followed by the name of the directory we want to go to. These names must be specified relative to the current directory, so if we’re currently in a directory called /home/trouble/ahead and we want to move into /home/trouble/ahead/for/starters , we need to move using the command:

* cd for/starters not
* cd starters

as the computer can’t find the ‘starters’ directory, without first being told to look in the ‘for’ directory.

1. To go up a directory we can use the special file ‘..’ e.g. cd ‘..’ will take us up to the parent directory. Type these commands, looking at the output

* cd Desktop/phylogenetics
* pwd
* cd ..
* pwd
* cd Desktop/phylogenetics/part\_1

… we’re now ready for the practical.

## Part 1 – Aligning sequences, building trees, and looking for error (30mins)

*# Unfunny pithy remark here, spot the pattern?*

We are going to construct a phylogeny of some matK genes taken from specimens in the gardens. They were collected from orchids and sequenced. First we’ll look at the sequences. They are in the ‘orchids.short.initial.fa’ file. You can use the command ‘less’ to look at the file (press ‘q’ to quit, or spacebar to scroll through the sequences):

**User@user-pi:~$** less orchids.short.initial.fa

That is a bit cumbersome. We’ll use a graphical user interface called CONTEXT to look at it. Open the context browser with this command:

**User@user-pi:~$** java -jar ../software/CONTEXT-PhylogenomicDatasetBrowser-v0.8.1.jar

Click ‘add a file’ to open the orchids.short.initial.fa file. You can see that the sequences are of various lengths. We are going to align them using a program called muscle. First open a new terminal by pressing ‘control-alt-T’ again (leave CONTEXT running in the old terminal), then navigate back to the part\_1 directory, and type:

**User@user-pi:~$** muscle -in orchids.short.initial.fa -out orchids.short.out -maxiters 2

Once muscle has finished open the output (‘orchids.short.out’) in CONTEXT.

**QUESTION 1:** Compare the two files. What has changed since you ‘aligned’ the sequences?

Now we will build a phylogeny from the aligned sequences. First we have to convert the file format from .fasta to a format RAxML uses:

**User@user-pi:~$**java -jar ../software/PrepareFilesForPaml.jar orchids.short.out

Now we can build the tree using RAxML:

**User@user-pi:~$** ../software/raxmlHPC -m GTRCAT -n orchids.short.initial -p 1 -s orchids.short.out.stops.removed.phy

Finally, we can look at the phylogeny we’ve generated (RAxML\_bestTree.orchids.short.initial.tre) in less again:

**User@user-pi:~$** less RAxML\_bestTree.orchids.short.initial.tre

It is in a format that takes a bit of skill to look at, so we’ll use another graphical program, FigTree, to view it more easily. Open FigTree and choose ‘File > Open’:

User@user-pi:~$ java -jar ../software/figtree.jar

You can see the phylogenetic tree that we have *inferred* – that is, our best guess at the evolutionary history connecting the orchid species. But one sequence looks odd…

**QUESTION 2:** Which sequence is the contaminant? Why that one?

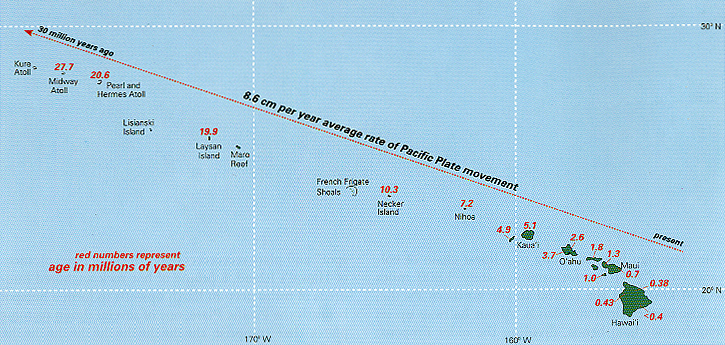
In ‘orchids.short.filtered.fa’, we’ve removed the contaminant. Re-align the sequences and repeat the phylogeny inference using the cleaned data set. **Hint: you’ll need to change the names from the previous commands.**

**QUESTION 3:** How might the contamination have arisen?

## Part 2a – Phylogeography of island chains (30mins)

*# Commenting code liberally is good. Comments in Linux typically start with a ‘#’*

* There are two datasets from two genera of flowering plants, both distributed on the hawaian islands. One is wind-dispered, the other is dispersed by mammals. We will investigate their districutopns and look for a phylogeographic pattern.
* The islands themselves have formed at different times from a chain of seamounts arising from two tectonic plates:



First you’ll need to navigate to the ‘part\_2’ directory in the command prompt window. There are two datasets, genus\_A\_master\_ungapped.fa and genus\_B\_master\_ungapped.fa. Each has one species from each island. We’ll build a tree for genus A first. We have to align the sequences, as before:

muscle -in genus\_A\_master\_ungapped.fa -out genus\_A\_aligned.out -maxiters 2

Then convert into phylip format:

java -jar PrepareFilesForPaml.jar genus\_A\_aligned.out

Then use RAxML to infer the tree:

../software/raxmlHPC -m GTRCAT -n genus\_A -p 222 -s genus\_A\_aligned.out.stops.removed.phy

Open the finished tree in FigTree again, and compare the phylogeny with the map.

**QUESTION 4:** Which island do you think was colonised last? Why?

Repeat the analysis for genus B. Compare this phylogeny with the one inferred for genus A.

**QUESTION 5:** What is the difference between the two genera? Which do you think is wind-dispersed, and which is dispersed by large mammals? Why?

## Part 2b – Rates of evolution of island chains (30mins)

*# Commenting code liberally is good. Comments in Linux typically start with a ‘#’*

Assuming the rate of evolution has been steady over time, we attempt to guess how old the clades in the tree are. The file ‘mc.paml\_MCMC\_slow.(time).MCC.tre’ contains a tree that has branch lengths in units of time (millions of years ago, MYA). Open it in FigTree.

**QUESTION 6:** Which dispersal event is the oldest?

Finally, we can use a programme called BEAST to infer how old clades within the tree are numerically. Your instructors will show you a trace file from BEAST. (mc.paml\_MCMC\_slow.log)

## Part 3 – Speciation in tropical rainforests (30mins)

*# Most code comments debase into wisecracks of dubious value soon enough…*

The shape of the tree itself also gives us useful information. Figs 1a and 1b show two common shapes of tree; 1a shows a lineage that has undergone a burst of rapid speciation and evolution. Fig 1b shows a lineage which is evolving more slowly. **[joe – no image]**

Navigate to the '~/Desktop/phylogenetics/part\_3' directory. The file ‘Trich\_ITS.NXS.short.fa’ contains a dataset from the ITS (internal transcribed spacer) domain of the ribosomal genome. Align the sequences and infer a tree with RAxML, then examine it in FigTree.

Half the samples have been collected from tropical rainforest. The other half have been collected from seasonal dry forest.

**QUESTION 7:** Which half of the phylogeny has evolved recently? Which half a while ago?

### Glossary of useful terms

* **Alignment** – data (usually DNA or amino-acid sequences) that has been optimally arranged so homologous positions can be compared. May include '-' characters known as 'gaps' to pad out shorter sequences where nucleotides/amino acids have been deleted during evolution (aka 'indels')
* **Bayesian inference** – phylogeny can be inferred using a likelihood function to produce posterior probabilities for a tree and then infer the most likely tree for a given dataset.
* **Branch** – a line that represents the path of evolutionary history within a phylogeny
* **Clade** – a group of taxa in a phylogeny that share character(s) from a single hypothetical ancestor
* **Likelihood** – is defined to be a quantity proportional to the probability of observing the data given the model. Usually expressed in log-likelihood units, abbreviated as 'lnL'
* **MCMC** – Markov chain Monte Carlo methods are algorithms. They can be used for sampling probability distributions
* **Node** – the position in a tree where branches separate. Each ‘node’ represents a real or hypothetical ancestor
* **Phylogeny** – evolutionary tree inferred from a dataset (often DNA sequences)
* **Posterior probability** - the probability of assigning observations to groups given the data
* **Tip** – the tips of a tree normally represent the extant (living) taxa that were sampled
* **Taxon** – the formal name of an organism or group of organisms