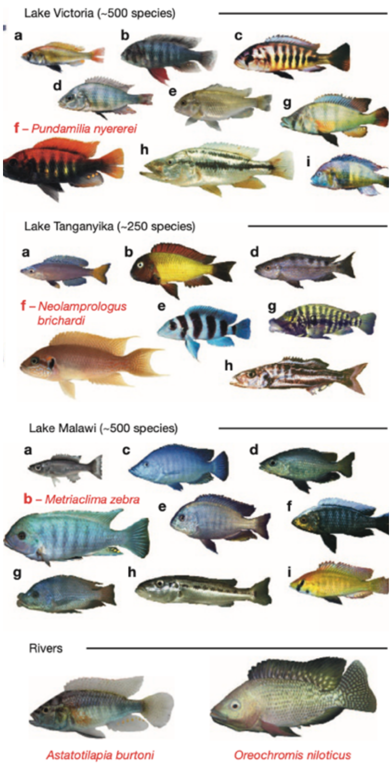
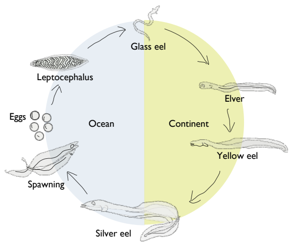
Using molecular approaches to understand the drivers of population divergence and speciation.

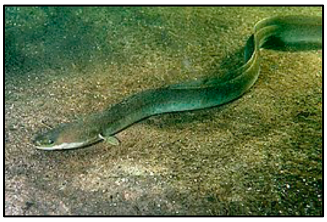
Cichlids



Brawand et al., 2014

European eel





Hylid frogs

Prof. Ian R. Sanders, Dr. Luca Fumagalli, Prof. Nicolas Salamin

Teaching assistants: Dr. Soon-Jae Lee, Dr. Angélica Pulido, Dr. Anna Hewett, Dr. Jaime Gonzalez, Dr Ricardo Arraiano, Marion Nyamari, Kenneth Kim.

Table of Contents

Introduction

Objective of this course

The main objective of this course is to give you an insight into how researchers use molecular techniques to address ecological and evolutionary questions. Although this course will contain a lot of technical information, **you must always keep the biological question in mind**! If you remember only one thing from this course, remember that (although we hope you remember more!).

As well as a clear question, **a scientist requires a powerful dataset on which to base solid conclusions**. In this practical, you will generate and analyse some data on your own, as well as some datasets we give to you. You will learn how to assess a dataset for its suitability to the task and recognise weaknesses inherent in some approaches. In the process, you will be introduced to several methods, both practical and analytical, which are used daily by ecologists and evolutionary biologists.

The "theme" of the course

Over this 8-day course, you will work on 3 projects. Each is an example of a natural biological system containing multiple populations / species. Although they come from different corners of the tree of life, **there is a common theme between these systems - they are all subject to ecological and evolutionary forces which promote or hinder population divergence and speciation**. The first (tree frogs) is an example of allopatric divergence (but with a twist), the second (cichlid fish) is a classic example of ecological divergence in sympatry, and the third (European eels) is in interesting case which we leave to you to categorise. You should already have heard about these modes of divergence in your fundamental evolutionary biology courses, but they are also briefly summarised below. Note however, that several other mechanisms of divergence and speciation exist. If only we had time to cover them all!

Allopatric divergence results from reproductive isolation caused by the separation of previously connected populations of the same species in physical space. For example, it is thought that approximately 2.7-3.5 million years ago (MYA) the Isthmus of Panama closed. In doing so, it broke the connection between the Pacific and the Caribbean seas, meaning many previously connected species were isolated to an extent where gene flow was impossible. The resulting divergence has been studied in many species in this region, including sea urchins, fish and isopods. Note that there are many geographic mechanisms which can lead to isolation, not just the formation of land-bridges.

Source: https://en.wikipedia.org/wiki/Allopatric_speciation

Ecological divergence refers to cases of population divergence and speciation where reproductive isolation is not caused by an isolation of species in physical space - i.e. the species remain sympatric throughout the process of divergence speciation. Instead, reproductive isolation is caused by differences which arise in their ecology, allowing them to move into different ecological niches. A good example can be found in the skinks of North America. Richmond & Reeder (2002) found evidence for repeated evolution of increased body size which has allowed some species to occupy higher altitudes and withstand harsh weather conditions.

Advice on preparing your report

The marking for this practical course will be based predominantly on your written report However, for those of you who take part in the wet lab section, a part of your final mark will also be based on your performance in the lab. During this wet lab portion of the course, make sure to keep strict notes of what you did and any auxiliary observations you made throughout, so you can include them in your final report when you write it. Below are some specific guidelines for writing the report but always remember to keep the work you report in the context of the biological question of the project.

Your report should start with a **general introduction** presenting the key concepts and main goals of this practical.

For those who participated in the wet lab session, your report should include a **Molecular Method** section, including methods followed and the results obtained during the first week of the class. When writing this section, keep the following points in mind: Are the lab protocols complete? It is ok to copy the protocol from the manual (as long as you cite it somewhere), but remember to make your observations/discussions distinct from this (for example, can you conclude if two genetic markers gave the same information about the species identity or relatedness to other species? If not, what can be the possible reason?). Sometimes (or usually), molecular experiments can be unsuccessful and fail to give enough results to be interpreted. Make an effort to understand the results and think about the potential cause and solution. Clearly label samples, reagents, volumes, concentrations etc and properly track them through the report. Note the date, time. Be exhaustive in the information you record. Note important observations, including failures during lab work if there were any.
For those not doing the wet lab, you do not need to include this section in your report, but it is advisable to read through that section of the protocol to give you an idea of the process preceding the data analyses.

The report should be organised according to the **three projects**, and the sub-sections should correspond to those found in a traditional scientific paper, i.e. Introduction, Methods, Results, Discussion and General Conclusion for each project. See below a short explanation on what is expected of each sub-section:

Introduction: Here you should set the scene of the project. Describe the reasons for and significance of this study, introduce the biological system, outline the main questions or objectives.

Methods: Here you should document *in detail* the approaches you used. Explain the logic of the analyses, i.e. why are you comparing observed and expected heterozygosity, Fst, etc. Note all software names and versions. In the first class, there will be considerable emphasis put on why it is so important to clearly document what you did.

Results: Here you should thoroughly describe the results of your analyses. Include tables and figures to support the text description of your results. E.g. A table of your samples with relevant information when available, like location, calculated population genetics statistics like He, Ho. Display all results of your analyses and make sure tables and figures are well labelled (e.g. title, legend, x-y axes labelled) and easy to understand.

Discussion: Here you discuss the implications and significance of your results in the light of the biological questions you outlined in the introduction. Also discuss any relevant technical details, E.g. the differences between phylogenetic trees obtained with nuclear, mitochondrial and RADseq data, bootstrap values, suitability of these markers for our purpose (i.e. be critical of your data / study design). ***Important:*** *Always look in the literature (especially literature mentioned in this manual) and include citations – how does your work fit with what is already known in phylogenetics, population structure in these species etc.*

General Conclusion: In the end of your report, add a **general conclusion**, where you put in perspective the three projects and describe the usefulness of molecular methods to understand the ecology and evolution of species.

Your report should not exceed 25 pages in total, with 5 pages maximum per project. Use normal size (i.e 11 or 12) and classical font (i.e. Times new Roman or Arial), as well as regular margin sizes. You should send your report before the **27/10/2025 midnight**, to angelica.pulido@unil.ch and soon-jae.lee@unil.ch

How to use this manual

In the sections below we will guide you through the wet-lab procedures as with any usual lab protocol of each project. For the analysis sections, although we will provide detailed instructions for some parts, we will not provide you the R code for the population genetic analyses. The reason for this is that, in the real world of molecular ecology and evolution, there is rarely someone to tell you exactly what to do. Thus, an integral part of a successful analysis involves looking around for the best approaches, software, parameter values etc. for your particular needs. So, in the sections that follow, we will pose questions, suggest potential functions or approaches that you might use, but it is up to you to find and interpret tutorials and R documentation. Of course you will also have assistants to ask during the TP.

We have included questions throughout the manual - look for text like this. These questions are designed partially to help guide you through your analyses and also the report writing, so pay attention to these questions and try to incorporate as many answers as you can into your report. **Note** that it is not so important that you get these questions correct in your report, rather you will be marked on your ability to apply logic, based on the findings of your analyses.

Schedule

Schedule

Time	Monday Sept 29	Tuesday Sept 30	Wednesday Oct 1st	Thursday Oct 2nd	Friday Oct 3rd
08:00 - 08:45	DNA extraction	PCR	Sequencing prep	Lecture	
09:00 - 09:45		Gel preparation		IS	
10:15 - 11:00	Lecture	Lecture	Lecture	Lecture	
11:15 - 12:00	IS	IS	IS	NS	
12:15 - 13:00				DEE Seminar	
13:15 - 14:00	TP Introduction	Electrophoresis		Lecture	
14:15 - 15:00	DNA extraction			LF	Lecture
15:15 - 16:00	Quantification	Lecture		Introduction &	NS
16:15 - 17:00	Dilution	IS		Installation	
17:15 - 18:00		PCR purification			

Time	Monday Oct 6th	Tuesday Oct 7th	Wednesday Oct 8th	Thursday Oct 9th	Friday Oct 10th
08:00 - 08:45	Project 1	last comments	last comments	Project 4.	
09:00 - 09:45	Phylogeny- Sanger	project 2	project 3.	&	
10:15 - 11:00	Lecture	& start Project 3	& start Project 4	Final questions	
11:15 - 12:00	LF				
12:15 - 13:00					
13:15 - 14:00				DEE Seminar	
14:15 - 15:00	Project 2	Project 3.			
15:15 - 16:00	Cryptic speciation in	Genomic analyses of	Project 4. The	Personal Research	
16:15 - 17:00	hylid frogs.	divergence between	curious case of the	Work	
17:15 - 18:00		Lake Malawi cichlids	European Eel		

Wet-Lab Experiment	POL 203 and 205
Lecture	POL 334
Computer Analyses	POL 204.2

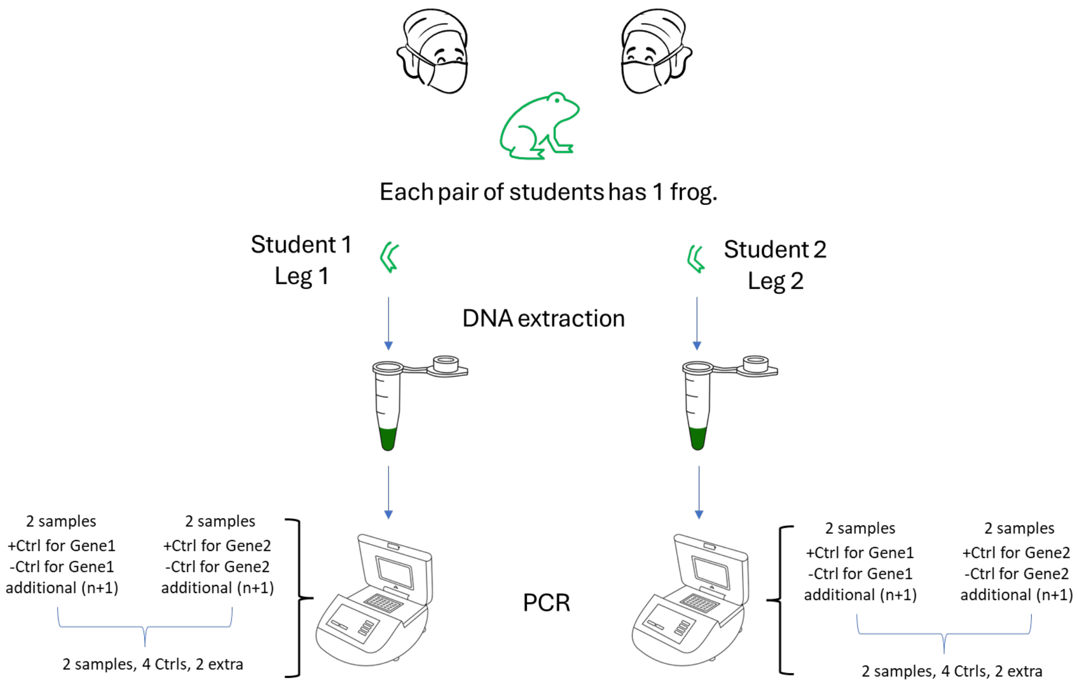
Wet-lab protocol for Sanger sequencing-based identification of unknown froglet

Ideally for the wet-lab stage of this course you should be working as a **TEAM (not individual)**, so please find a position at a workbench with partners.

Extraction of DNA from froglets.

Before any genetic work can be done on an organism, we must first extract DNA from a sample. This can be a tissue from organism, water from lake, soil etc. However, when working with wild animals, especially those of conservation concern, an important consideration is how to take a sample in a way that will cause as little damage / stress to the organism as possible. For frogs, the most non-invasive way of sampling is by using a buccal cell swab on the inside of their mouth. Unfortunately, these swabs can only be extracted from once, and so there are no swab samples left for this study. To give you experience of the extraction procedure, we have, therefore, provided you with unknown froglets from a lab-reared clutch. The parents of this clutch were reintroduced to the site where they were collected and so the effect on the natural population was minimal. We will quantify your extractions, and then we will provide you with aliquots of DNA from the original study which we will use for the PCR and sequencing.

Note: work at the molecular level is highly sensitive to contamination. It is thus necessary to wear gloves at all steps during lab work. Always sanitize your gloves with ethanol if you touch non-sterile material (for example, your glasses, pocket, cell phone, etc.).



1. You will be given 1 froglet in ethanol. These froglets were in long term storage in 70% ethanol at -20°C. However, ethanol can interfere with DNA extractions, so wash them first in purified water in a petri dish.

For each froglet, remove it from the water and place it on the tissue paper provided. Roll it around to remove excess water and then transfer it to the plastic petri dish. With tweezers and a scalpel, cut off one of the back feet of the froglet, and place it in a 1.5 ml Eppendorf tube. Place the rest of the frog in a separate 1.5 ml tube. **Be sure to assign a unique label to each of your samples, note it in your lab book and write it clearly on the top and side of both tubes. Permanent markers can be removed by ethanol, please be cautious.**

We have colour coded all the reagents for DNA extraction and added added the corresponding colours in the manual, so you will know which ones to use for which step.

Add 180 µL of Buffer ATL and 20 µL of proteinase K to the extraction tube (the one containing only the foot). Usually, you should keep the proteinase K on ice as enzymes can be very sensitive to temperature, however, the proteinase K we are using is stable at room temperature.

2. Mix thoroughly by vortexing and incubate at 56°C in the thermomixer for 2–4 hours until the tissue is fully lysed.

Note: during centrifugation steps, be careful to equilibrate the weight of the tubes in the centrifuge. Ask an assistant if you are unsure how to do this.

3. Collect your sample from the thermomixer and vortex for 15 sec. Centrifuge for a few seconds to collect the evaporated liquid from the caps. Add 200 µL Buffer AL to the sample, and mix thoroughly by vortexing and incubate at 56°C for 10 min. Add 200 µL ethanol 100% and mix again thoroughly by careful pipetting. **Can you see the change in the tube?**

4. Pipette the mixture from step 3 into the DNeasy Mini spin column (students should label with the sample ID) placed in a 2 ml collection tube (provided). Centrifuge at 8000 rpm for 1 min. Discard flow-through and collection tube.

5. Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 µL Buffer AW1, and centrifuge for 1 min at 8000 rpm. Discard flow-through and collection tube.

6. Place the DNeasy Mini spin column in a new 2 ml collection tube and add 500 µL Buffer AW2 and centrifuge for 3 min at 14,000 rpm. Discard-flow through and collection tube. (It is important to dry the membrane of the DNeasy Mini spin column, since residual ethanol may interfere with subsequent reactions. Following the centrifugation step, remove the DNeasy Mini spin column carefully so that the column does not come into contact with the flow-through in the collection tube).

7. Place the DNeasy Mini spin column in a clean 1.5 ml microcentrifuge tube (previously labeled with the sample ID) and carefully pipette 100 µL Buffer AE directly onto the DNeasy membrane (do not touch the membrane with the pipette tip, as this may break the membrane). Incubate at room temperature for 1 min, and centrifuge for 1 min at 8000 rpm to elute.

8. Take your sample to the nanodrop machine, where an assistant will help you check both quality and quantity of your extraction. Note down this concentration, 260/280 ratio and 260/230 ratio. You can then place your sample in a freezer box.

Q. What does the 260/280 ratio or 260/230 ratio show, respectively? How can you conclude if your DNA extraction was successful or not?

PCR amplification of genetic markers

Let us say we want to compare the certain genetic markers from different populations, first we need to collect the specific sequences from each of the samples. We have extracted entire DNA molecules existing in the organism, but we need to purify and amplify the sequences of our interest. How can we do this? The polymerase chain reaction (PCR) is a well-established technique to amplify specific DNA sequences in vitro. The polymerase chain reaction is based on the principle of DNA replication. In a PCR cycle, heat (94-96°C) is used to produce single stranded DNA (denaturation step). After this initial step, the temperature is reduced to ca. 50-65°C such that primers can anneal (annealing step). The annealing temperature is primer-specific. In a third step (elongation step; 68-72°C), the enzyme polymerase replicates the template in 5'-3' direction starting from each of the two primers. Thus, the stretch of DNA that is flanked by the primers is the target of amplification. PCR could be wisely used only after the finding of polymerases derived from those of heat-tolerant microorganisms such as *Thermus aquaticus*, so that the enzymes do not degrade during PCR cycles. Since during PCR also newly synthesized strands act as templates in following cycles, the amplification is exponential. Therefore, only relatively few cycles (25-40) are necessary to yield enough product for further steps such as DNA sequencing.

Today we will prepare the PCR reactions to amplify regions of the *recombination activating gene 1* and *rhodopsin* gene of unidentified frog samples.

Note: work at the molecular level is highly sensitive to contamination. It is thus necessary to wear gloves at all steps during lab work. Always sanitize your gloves with ethanol if you touched non-sterile material (for example, your glasses, pocket, cell phone, etc.).

Recombination activating gene I (Rag1) (~550bp)

1. Prepare a 1.5ml tube for the master-mix, label it (e.g. MMrag1).

2. Prepare 0.2 ml PCR tubes labelling them with "*rag1*" + the DNA sample name (sample, +ctrl, - ctrl).

3. Remove all the necessary reagents from the freezer and **place them on ice**. However, remember that enzymes are temperature-sensitive, so be sure to keep the PCR master mix on ice at all times.

4. You will be provided with a positive control (0.5 mL tube with a green + sign). This is DNA that we have tested before, and which gave a good result with both PCRs.

5. Prepare one master-mix for your PCR reactions using the following table (following the order from top to bottom). Do not forget to account for the volumes needed for both a positive control (which will be provided) and a negative control (replace DNA with water). The final volume of the individual reactions will be 25 µL.

   Q. What are the functions of the positive and negative controls here?

Reagent	**Volume (µL) for 1 sample**	**Volume for (N samples +2 controls) + 1**
H_20	8	
Master mix (2x)	12.5	
rag1-F 10 µM	1.25	
rag1-R 10 µM	1.25	
Total volume	23	

   Q. Why do we make extra 1 reaction amount of the mastermix here?

6. Distribute **23 µL master-mix** per well in the prepared PCR tubes.

7. Vortex the DNA extractions shortly and centrifuge them briefly to remove drops from the cap.

8. **Add 2 µL DNA** per individual to the corresponding tubes. Close the tubes well. Vortex your PCR reactions, and spin them down. Place your tubes on ice right beside the PCR machine.

9. The machine has been prepared for the following cycling conditions:

Temp (℃)	**Time**	**Cycles**
98	30’’	
98	10’’	34 x
51	30’’	
72	1’	
4	∞	

Rhodopsin (Size ~350 bp)

1. Follow the same procedure as used for *rag1* except use the following volumes for the preparation of the master mix.

Reagent	Volume (µL) for 1 sample	Volume for (N samples +2 controls) + 1
H_20	8.5	
Master mix (2X)	12.5	
Rho-F 10 µM	1	
Rho-R 10 µM	1	
	23	

2. This time, distribute **23 µL master-mix** per well in the prepared PCR tubes.

3. Vortex the DNA extractions shortly and centrifuge them briefly to remove drops from the cap.

4. **Add 2 µL DNA** per individual to the corresponding tubes. Close the tubes well. Vortex your PCR reactions, and spin them down. Place your tubes on ice right beside the PCR machine.

5. The *Rhod* PCR machine will be programmed as follows

Temp (°C)	Time	Cycles
98	30’’	
98	10’’	34 x
51	30’’	
72	1’	
4	∞	

Agarose gel electrophoresis & visualization

We will now check that amplification has been successful using gel electrophoresis, another classical method in molecular biology used on a daily basis. DNA molecules are negatively charged and therefore migrate towards the positive pole in an electric field. In gel electrophoresis, we suspend DNA in agarose gel and apply an electrical current. The agarose provides resistance to the movement of the DNA which is proportional to the size of the molecule, i.e. smaller molecules move faster through the agarose matrix than larger ones. As such, gel electrophoresis can be used to separate DNA strands by size, and to estimate their size through comparison to fragments of known lengths (size standard, or DNA ladder). We visualise DNA fragments in a gel using dyes which bind to the DNA molecules. When exposed to ultraviolet light these dyes will fluoresce. DNA fragments will then become visible as distinct bands in the gel.

Note: There is a chemical in this experiment that can bind the DNA fragments to visualise it under UV. This means your DNA in your skin is not an exception. It is thus necessary to wear gloves at all steps during lab work.

Preparation of the agarose gel

1. Prepare 30 ml of a 1.5% agarose gel solution by dissolving 0.45g agarose in 30 ml 1x TBE buffer. Heat the agarose gel solution in the microwave until the agarose melts and the solution gets liquid and clear. Do not overheat the agarose gel solution as it may evaporate.

2. Let the agarose cool slightly and add 3 µL of GelRed. Mix well.

3. Prepare a gel tray with the respective comb and pour the agarose into the gel tray. Cover with an aluminium foil to protect the GelRed from the light.

4. Before each sample can be loaded into the well in the gel, it has to be **mixed with loading dye** so that it falls to the bottom of the well. Take a piece of parafilm and deposit on it the appropriate number of separate drops of 2 µL loading dye each. Take 8 µL of PCR product per reaction and mix with an individual loading dye drop on the piece of parafilm. Make sure you remember which sample is where. **It is important to not make any bubbles during the mixing. The bubble will pop during the loading and make the contamination during the electrophoresis.**

5. Carefully remove the comb and check if there is any damage on the gel. Place the tray with the solid agarose gel into a gel chamber filled with a running buffer (1x TBE).

6. Carefully load the samples in the well, including 4 µL of DNA ladder in **one of the wells** in the gel.

7. Close the chamber and run the gel for 35 min at 100 V.

8. When the gel has finished running, switch off the power and carefully remove the agarose gel and visualise it under UV light. Take a photo and label it. What is the size of your PCR fragments? (See Fig. 2 for the reference sizes of the ladder).

Q. What can you conclude from the gel, was your PCR successful and, if so, how do you know that you have amplified the correct region?

PCR purification before cycle sequencing reaction

To know the sequence differences of target genes among populations we need to sequence the amplified molecules from PCR. Before we can use the products from the PCR in the cycle sequencing reaction, we have to purify them from the remaining dNTPs, primers and salts which can inhibit the sequencing reaction. One way to do this is to run the PCR products on a gel, cut the band to be sequenced, and purify the gel. This method is usually applied when the PCR is not specific enough to exclusively amplify the target gene and instead amplifies other regions as well. In cases where the amplification is specific enough, such as the PCR we performed here, standard commercial kits are applied directly to the PCR product. Here we will use the Monarch PCR purification kit.

Note: work at the molecular level is highly sensitive to contamination. It is thus necessary to wear gloves at all steps during lab work. Always sanitise your gloves with ethanol if you touched non-sterile material (for example, your glasses, pocket, cell phone, etc.).

For each successful PCR product of *rag1* and *rhod* (positive controls and negative controls DO NOT need to be purified), label a purification kit column with the sample ID and gene name.

1. Dilute sample with DNA Cleanup Binding Buffer in 5:1 proportion of binding buffer to PCR product. You would have 40 µL of PCR product left after electrophoresis, mix this with 200 µL of DNA Cleanup Binding Buffer. Mix well by pipetting up and down or flicking the tube. DO NOT VORTEX.

2. Insert the column into collection tube and load sample onto the column. Spin for 1 minute, then discard flow-through.

3. Re-insert column into collection tube. Add 200 µL DNA Wash Buffer and spin for 1 minute. Discarding flow-through.

4. Repeat wash (Step 3).

5. Transfer column to a clean 1.5 ml microfuge tube. Use care to ensure that the tip of the column has not come into contact with the flow-through. If in doubt, re-spin for 1 minute to ensure traces of salt and ethanol are not carried over to next step.

6. Add 30 µL nuclease free water to the center of the matrix. Wait for 1 minute, then spin for 1 minute to elute DNA.

Q. What can you conclude from the nanodrop? How did / could you increase the confidence of your concentration estimates?

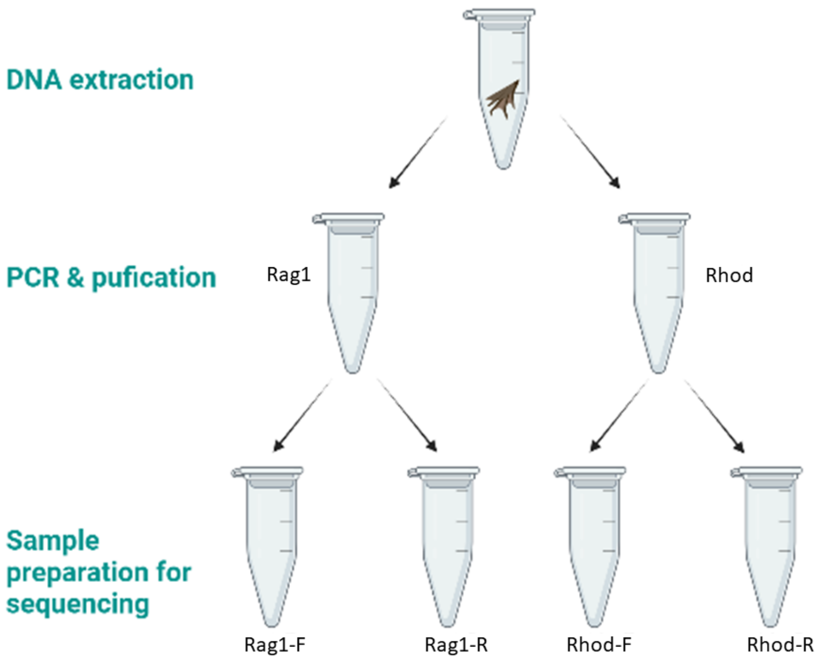
Preparation of samples for Sanger sequencing.

The cycle sequencing reaction is based on the principle of PCR. Yet, in addition to the dNTPs, fluorescently labelled ddNTPs are added. These nucleotides lack the 3'-OH group that is essential to prolong the DNA strand. As soon as a ddNTP has been incorporated, the synthesis of the fragment is terminated. With this method, strands with different sizes (from primer length plus 1 bp to the entire length of the sequence) are produced, each ending with a fluorescently labeled ddNTP. To allow the determination of the sequence, **only one strand can be sequenced at a time** (i.e., only one primer can be used at a time). After the cycle reaction, all fragments of different sizes are separated by capillary electrophoresis. The identity of the ending ddNTP of each fragment (i.e. either A, T, C or G) is determined thanks to the fluorescence of each, and the sequence of the whole PCR product is reconstructed from the successive bases that are detected by capillary electrophoresis. Cycle sequencing and capillary electrophoresis are performed by commercial labs where we will send our samples.

Note: work at the molecular level is highly sensitive to contamination. It is thus necessary to wear gloves at all steps during lab work. Always sanitise your gloves with ethanol if you touched non-sterile material (for example, your glasses, pocket, cell phone, etc.).

Preparation of samples

1. Prepare 0.2 ml tubes by labelling them with sample IDs. Each PCR product will be sequenced from both sides in separate reactions, i.e. from the forward primer side, and the reverse primer side. Therefore, each PCR product has to be prepared twice so that it can be mixed with both the F and R primers. We will send your samples to the commercial lab for sequencing specifying the individual sample names that you label the tubes with. It is therefore **very important** that you label the tubes with the number of the sample, the name of the gene, i.e. *rag1* or *Rag1*, and indicate the primer added, either forward (F) or reverse (R)). So for example, if one of your samples was number 14, label: "14*rag1*-F".



2. Vortex the PCR purifications and centrifuge them briefly to remove drops from the cap.

3. The amount of DNA added to the sequencing reaction depends on the length of the PCR product. The volume of PCR product needed is 12 µL. The required concentration of DNA in this volume is

   ng/µL of DNA in reaction = 1.5 x (size PCR product in bp/100)

   Ideally for about 900 bp fragments, a concentration of >17ng/ µL is optimal. If you have this concentration, or higher, you can dilute if required and make up the volume of 12 µL. If the concentration of your purified PCR product is less than that, add 12 µL of the concentration you have, and we can check if we still get some sequence data.

4. Prepare each purified PCR by diluting it with ddH2O to achieve the correct concentration and volume. If the volume of PCR product is too small to pipette accurately (e.g. less than 0.5µl), dilute your PCR product first.

5. Add the corresponding F or R primer to each tube: add **3 µL** Primer at **10 µM** to reach a final concentration of **2 µM**. The total volume of the sample is 15 µL.

6. Vortex your samples briefly, and spin them down. Leave them on the rack prepared by your assistant.

Computer-lab: Data analyses

Downloading the data for this practical

The first thing you must do is to **download all the data for this course** onto your computer. To do that go here: https://github.com/Angelica-Pulido/MMEE-2025/ and click on **Clone or Download >> Download ZIP.** Save the folder to your desktop and un-compress it if needed. You should then have all the course materials in the directory:
"**C:/Users/yourusername/Desktop/MMEE_2025/**" for windows.
"**/Users/yourusername/Desktop/MMEE_2025/**" for mac or linux.

You can use this path with the `setwd()` function when working in R so you can easily find all of the files you need.

We will perform most of our genetics analyses in the computing language R, using the R studio interface. R is one of the most common scripting languages used for statistical analyses of data, especially in population genetics. Hopefully you are already a little familiar with R, but don't worry if you find it challenging, everyone does at first - that's what the assistants are there for. Just ask us and we'll help you.

While this course isn't designed specifically to improve your competence in R, a crucial skill in population genetic analyses is the ability to find the most suitable tool to analyse your data, to understand how that tool works and to properly use it to achieve your goals. With this in mind, we will not give you step by step instructions for this part of the course. Rather we will guide you towards the R packages and functions that you will need for each task, and it will be mostly up to you to write the code required. But you are not alone, you can look at the help documentation for any function in R by running it with a '?' in front of it. E.g. `?abline`

Useful packages for population genetic analyses

There are many packages in R which allow various population genetics tests. For the analyses of the data in this course, we suggest you use these packages, but you are of course free to find more if you wish.

APE and **PHANGORN:**

Multiple packages may be used to perform phylogenetic analyses in R. Two of them are APE and PHANGORN, used to manipulate DNA sequences, compute distances between these sequences and infer phylogenetic trees.

Some commands and functions from APE and PHANGORN that you might find useful are listed below but see https://www.rdocumentation.org/packages/ape and https://www.rdocumentation.org/packages/phangorn. Use also help pages of each function for more detailed information:

`read.dna()`: reads sequences from a file (may be a fasta file for example).

`phyDat()`: convert several DNA formats (as the one produced by *read.dna*) into the phyDat format.
`dist.ml()`: compute pairwise distances for an object of class *phyDat*.
`NJ()`: performs the neighbor-joining tree estimation.
`root()`: re-roots a phylogenetic tree with respect to the specified outgroup or at the node specified.

HIERFSTAT:

HIERFSTAT is a package developed for R to study population structure with the help of genetic markers. Wright's F-statistics are commonly used to determine population structure and differentiation. HIERFSTAT allows you to estimate F-statistics and many other basic statistics important for the analyses of population-level genetic data.

Some commands and functions from HIERFSTAT that you might find useful are listed below but see here https://www.rdocumentation.org/packages/hierfstat and help pages of each function for more detailed information:

`read.fstat.data()`: import data formatted as a .dat file (the .dat file is the format used by the software FSTAT). The data will be read in as a dataframe.
`basic.stats()`: calculate basic statistics such as expected heterozygosity, observed heterozygosity, F_{IS}, etc for each population and each locus.
`pairwise.neifst()`: to calculate population pairwise Fst
`indpca()`: to carry out a PCA on the centred matrix of individual allele frequencies

ADEGENET:

ADEGENET and HIERFSTAT are built using many of the same libraries and dependencies. There is, therefore, a lot of overlap between them and they complement each other very well. For example, ADEGENET also uses the Genind object type and contains functions for calculating heterozygosity, allele frequencies, F_{ST} etc. However, ADEGENET also implements methods that are very useful for spatial genetics, for example examining Isolation By Distance (IBD), and PCA.

Some functions that you will find useful are below. There is also a great tutorial for ADEGENET, you can find it here: http://adegenet.r-forge.r-project.org/files/tutorial-basics.pdf. Take a look specifically at the PCA and IBD sections which should prove very helpful.

`read.fstat()`: "Genind" object - a specific object class for this package. Take a look at the structure of the object, you can access slots within it using the `$` accessor, E.g. `mygenindobj$Pop` will show you the information in the Pop slot, which you can edit if you wish.
`genind2genpop()`: transform a genind object to a genpop object, necessary for some functions
`dist.genpop()`: Create a Euclidean distance matrix from genetic data
`dist()`: Create a Euclidean distance matrix from spatial coordinates
`mantel.randtest()`: Perform a mantel test for isolation by distance by comparing a matrix of genetic distances (such as pairwise F_{ST}) and pairwise geographic distances.

LEA:

LEA is an R package dedicated to landscape genomics and ecological association tests. LEA can run analyses of population structure and genome scans for local adaptation. It includes statistical methods for estimating ancestry coefficients from large genotypic matrices and evaluating the number of ancestral populations (snmf, pca).
Some commands and functions that you might find useful are listed below but see https://www.rdocumentation.org/packages/LEA, and help pages of each function for more detailed information:

`snmf()`: estimates admixture coefficients using sparse Non-Negative Matrix Factorization algorithms, and provides STRUCTURE-like outputs.
`cross.entropy()`: Return the cross-entropy criterion for runs of snmf with K ancestral populations.

Plotting in R

One of the most powerful aspects in R is its ability to produce scientific plots quickly from the data and results that you are analysing. This is particularly useful in the data exploration phase of a project, but also for making the final plots for reports, theses and papers. Here are some of the most commonly used plotting functions:

`barplot()`: Used to plot 1 dimensional data, e.g. heights of all students in the class.
`plot()`: Used to plot 2 dimensional data. E.g. height vs age. This function may also be used to plot phylogenetic trees.
`heatmap()`: Used to produce a graphical representation of a 2D matrix (e.g. pairwise F_{ST}). Take a look at the *Rowv*, *Colv* and *scale* parameters.
`abline()`: Adds a line to a plot (e.g. an average `abline(mean(x))`)
`text()`: Adds text to a plot (e.g. useful for adding population names to scatter plots etc.)
`barchart()`: This function displays a bar plot/bar chart representation of the Q-matrix computed from an snmf run.

There are many more plot types and functions, so feel free to explore the possibilities!

Project 1. Phylogenetic identification of unknown froglet using Sanger sequencing

Using Sanger sequencing on the samples you processed in the wet-lab

Sanger sequence analysis

Before we can analyse the sequences we produced in the web-lab part, we have to quality check them and correct/remove errors. To do this, we need to align the forward and reverse sequences for each individual and check them for congruence (note that this step controls for sequence reading errors, not those errors occurring during PCR). Once this is done, we can build a consensus sequence for each individual. These steps were done for you by the TAs, before this practical. You can find the sequences in the github repository on: **MMEE-2025/1.Frogs_Sanger/Data/Phylo/.**

Now you will have to align all sequences such that homologous positions among individuals are identified. We will perform this step online, with the freely available program Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). Finally, using R, we will reconstruct the phylogenetic relationships among the samples.

To reconstruct the phylogenetic relationships between our frog samples, we are going to compute a phylogenetic tree using the simple, distance-based, Neighbour-Joining (NJ) method. To root the tree, we will also need one or more outgroups which we know are more distantly related to the other frogs in our tree. We will supply you with some sequences from *Hyla* for this purpose. Before estimating the NJ-tree, we have to integrate the outgroup sequences in the sequence alignment.

Q. Why should we include an outgroup when constructing a phylogenetic tree?

Sequences Alignment

1. Open the my_file.fasta file containing your sequences for *RAG1* and *RHOD* with any text editor software.
   The file may be found in the folder : **./MolGen-2025/1.Frogs_Sanger/Data/Phylo/** (in case your file is empty, you can use **rag1_Sanger_reference.fasta and rhod_Sanger_reference.fasta**).

2. *Optional*: Checking the chromatogram of your sequence
   a. connect to **https://www.gear-genomics.com/teal/**
   b. browse *.ab1 file and launch analysis

3. Building the consensus sequence for your sample:
   a. connect to **https://www.bioinformatics.org/sms/rev_comp.html** and get the complementary sequence of **your reverse read**.
   b. Merge the reverse complemented read with the forward read:
      i. Connect to **https://www.ebi.ac.uk/Tools/msa/clustalo/** (or **https://www.genome.jp/tools-bin/clustalw**)

ii. Specify that your input sequences are DNA sequences.
iii. select Pearson/FASTA format as output.
iv. submit your alignment.
v. Copy and paste the **overlapping part of the two reads** in a *new.fasta* file (keep in mind that a *.*fasta* file is composed by a header line, starting with a ">" and the sequence itself)

4. We have provided you with some representative sequences from other *Hylid* species. These sequences will firstly act as outgroups to your data, and secondly allow you to compare the divergence between our samples and the divergence between other, already-established species pairs. Import the outgroup sequences which can be found here:
./MolGen-2025/1.Frogs_Sanger/Data/Phylo/Rag1_Reference_sequences.fasta for rag1.
./MolGen-2025/1.Frogs_Sanger/Data/Phylo/Rhod_Reference_sequences.fasta for rhod.
Add these outgroup sequences to the *new.fasta* file with our own sequence.

5. Run a Clustal Omega alignment online and export that alignment as *fasta* file, by:
   a. Connecting to https://www.ebi.ac.uk/Tools/msa/clustalo/ (or https://www.genome.jp/tools-bin/clustalw)
   b. Specify that your input sequences are DNA sequences.
   c. load your file with all your sequences and the outgroup sequences.
   d. select Pearson/FASTA format as output.
   e. submit your alignment.
   f. Copy and paste the results in a *new2.fasta* file.

Simple phylogeny in R

Using R and the packages and tools specified above, load the aligned sequences, compute the distance between these sequences and draw a Neighbour-Joining tree.
You can use *Hyla japonica* as an outgroup to your tree.

Q. What can you say about your sample? Does your sample cluster with other species?
Q. What does this result means regarding your sample?
Q. What complementary analysis can you think of to better characterize you sample?

Project 2. Reconstructing the phylogenetic history of Hylid frogs using RAD sequencing

Phylogenetics of the genus *Hyla* using RAD-seq data.

Overview

In conservation biology, a fundamental piece of information required for effective management strategies is the distribution of a species and the population structure within that range. Molecular methods have, in the past few decades, proved invaluable for this task, and as technologies have advanced, we have been able to describe subtle distinctions between different lineages and species which were previously not possible.

In this project, we will work on a system of tree frogs from southern Switzerland and Italy. It was traditionally thought that these frogs were a single species, called *Hyla intermedia*. However, some phylogenetic evidence then arose (Stöck et al., 2008) that identified two distinct lineages within the range, which were proposed to be two different species. In a recent paper, researchers from the DEE took a closer look at this system using both traditional and modern genetic approaches (Dufresnes et al., 2018). We will recreate some of this work and try to decide, based on the results, if these really are two separate species and if conservation efforts should be made separately for both of them.

To do so, we will give you genomic data from the same samples, collected using Restriction Site Associated DNA sequencing (RADseq - see next pages for information). We will run some general population genetics analyses and compare this to the sequence phylogeny we made in the previous section to gain more information and helps us determine the status of the frogs in this region.

Q. What is RADseq, what are its advantages and why might we use RADseq here in addition to the gene sequences?

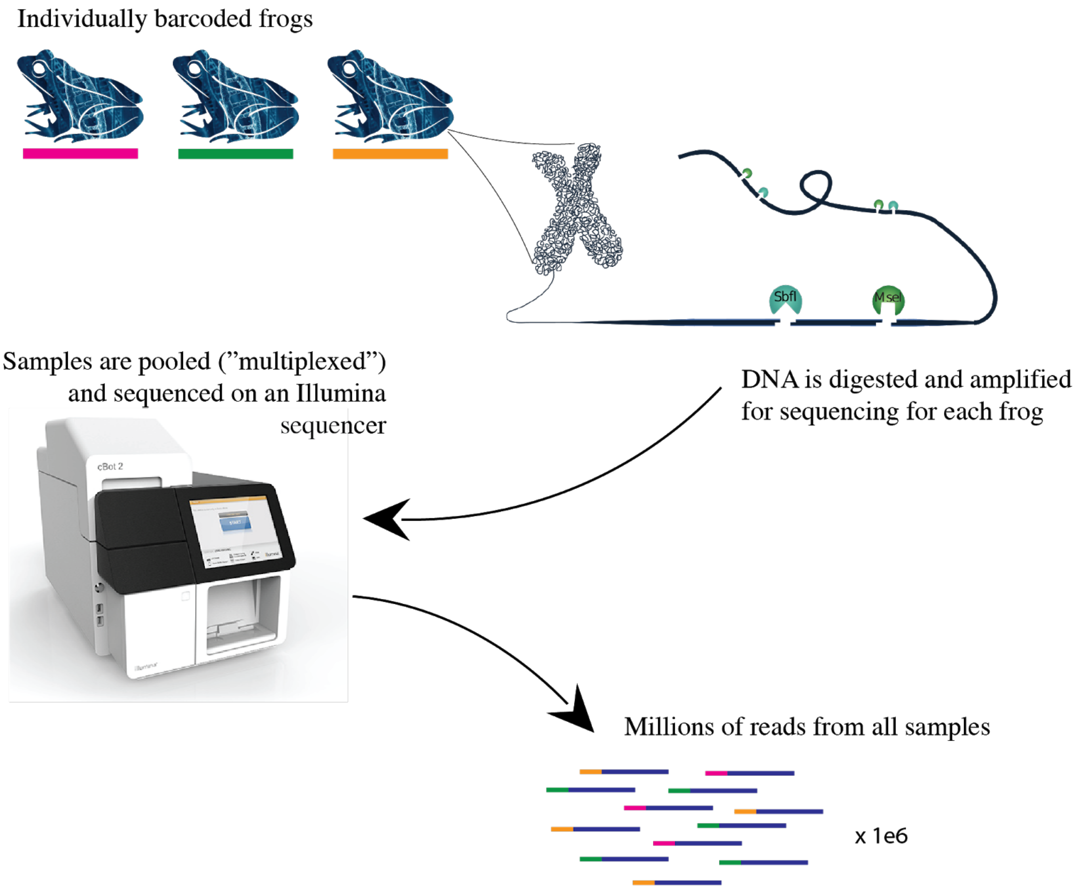
Summary of raw RADseq data processing.

The initial steps in the RAD marker analysis involve a bioinformatics pipeline which can take several days to complete. As such you will not be asked to do these steps, however you should be aware of how they work. Below, and in the graphical summary on the next page, we outline the major steps in producing RADseq data.

The RADseq library is made in the lab by first digesting (cutting) the genome of each sample with a specific combination of enzymes. As enzymes cut in specific places in the genome, they will generally **cut in the same place across all samples**. We then **incorporate a sample-specific barcode** to the resulting fragments. All fragments from all samples are then pooled together and sequenced using a high throughput sequencing technology like Illumina HiSeq.

The data which is produced by the sequencer contains a mix of all the pooled (multiplexed) samples, thus the first step of the data processing is to "demultiplex" the data, which means to split the sequences into subsets corresponding to the individual they come from. This step was performed using the program Stacks (http://creskolab.uoregon.edu/stacks/). Stacks contain several steps required for raw data processing, we will not describe them all here, but we encourage you to look at the pipeline yourself and familiarise with them.

One of the major outputs of the Stacks pipeline is a "VCF" file, which stands for "Variant Call Format". This file contains all variants (e.g. Single Nucleotide Polymorphisms, SNPs) for all samples in the dataset. Once in VCF format, the polymorphism data can be reformatted into one of several commonly used data formats. Here we will give you the data in "FSTAT" or GENEPOP formats, which are simply named after the software whose developers originally invented them. These formats are easy to load into the R packages we will be using for most of our analyses.

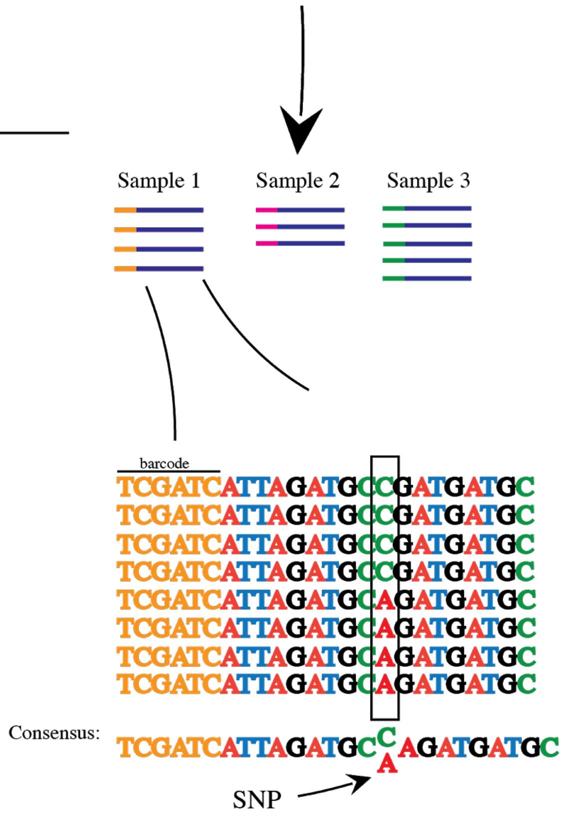


Stacks

Raw sequencing reads are quality checked and "demultiplexed".

There are several other steps here which are required for identifying homology between loci of different individuals

SNP Calling: During Stacks, we also identify variant sites, they are positions where there is a heterozygous position in an individual. This process is done for all samples and the variants are then summarised for all loci and all samples in a .VCF file which stands for "Variant Call Format".



Population genetics of Hylid frogs in R

With the R tools mentioned above, and others you might find, explore and analyse the frog SNP dataset, produced with the RADseq approach. The data set can be found here:
./MolGen-2025/1_Frogs_RADseq/Data/RADpopgen/hyla_FSTAT.dat

We encourage you to explore the data thoroughly:

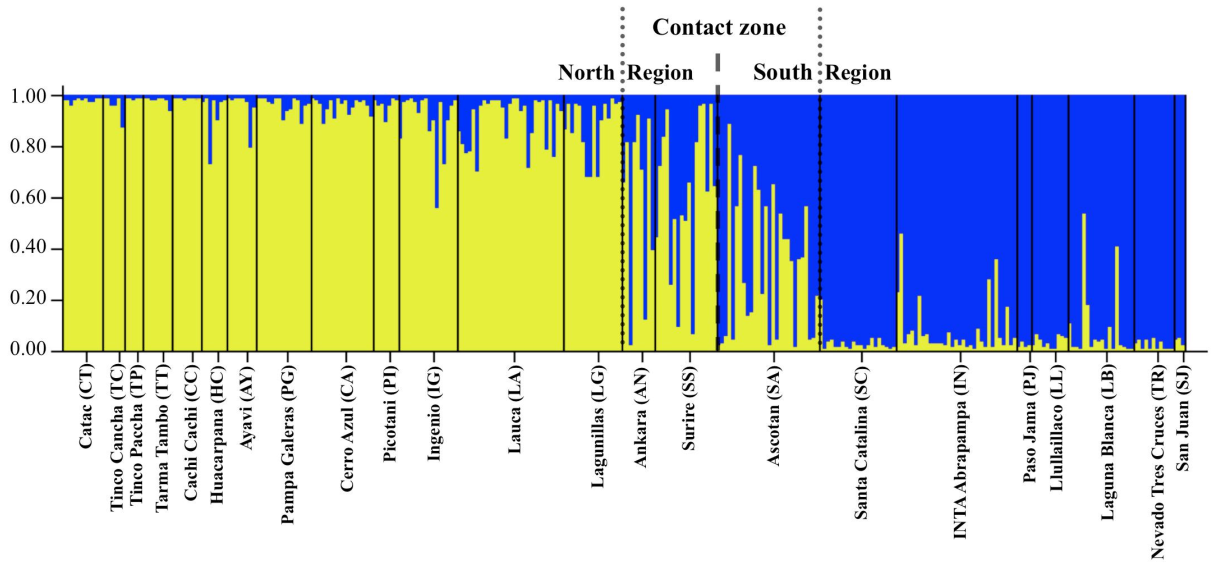
- Look at within population statistics, like observed and expected heterozygosity, inbreeding (F_{IS}) and expectations of Hardy Weinberg.
- Look at between population statistics - pairwise F_{ST} between populations, patterns of isolation by distance (i.e. a correlation between physical distance and genetic distance between populations (see also Mantel test)).
- Look at population clustering (i.e. are some populations more closely related to each other than to any others?) using a PCA.

Remember to keep the biological question in mind all the way through - are these frogs one species, or two?

Q. What do these calculations tell you about the divergence between populations?
Q. Are there some populations which are more divergent than others?
Q. Do the populations fall into discrete “groups” which might be considered as different species?
Q. What was the most likely evolutionary history of these species? Has the hybrid zone always been there? How did reproductive isolation occur?

Population structure analyses of frogs in R

We often talk about organisms sampled from nature as coming from "populations", however, this is a reasonably arbitrary classification. What is a population - is it a single pond full of frogs? Is it a network of ponds that are close and always connected to each other? Is it a network of ponds which sometimes connect to each other during high rainfall?
You can see that drawing a classification like this could be quite meaningless. However, one way to accurately determine a meaningful population genetics unit of organisms is to use an approach called genetic clustering. This is a method which groups individuals based on how closely related they are to each other. The advantage of genetic clustering is that it is not subject to inconsistent human impressions of what constitutes a cohesive population, but simply identifies groups ("clusters") of individuals that are more closely related to each other than they are to any other sample in the dataset. STRUCTURE is the most widely known program for this task (http://pritch.bsd.uchicago.edu/structure.html, Pritchard et al, 2000) but many other tools have been developed since. The idea is to assign individuals to K clusters based on their genotypes and estimate population allele frequencies. The program incorporates population genetics models, most importantly Hardy-Weinberg equilibrium, into its clustering and so has many advantages over more general multivariate clustering approaches like PCA.



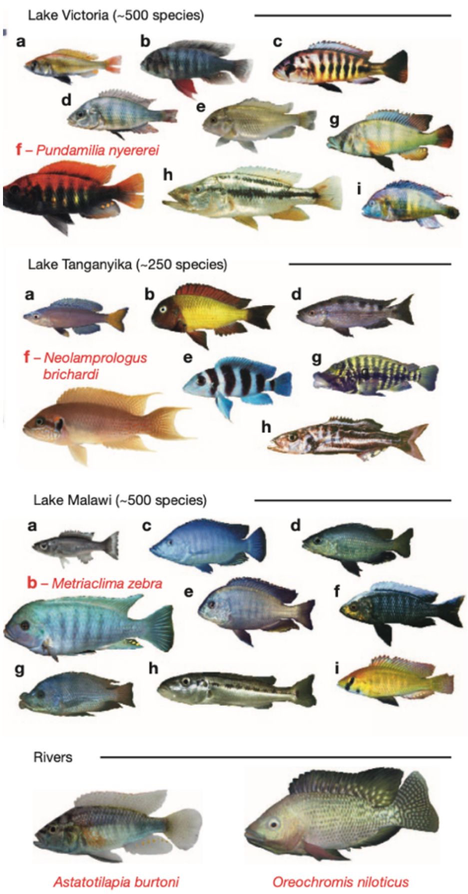
As an example of the use of STRUCTURE (or similar software), this figure above shows individuals of two subspecies of vicuña (*Vicugna vicugna mensalis* and *V. v. vicugna*) which were genotyped at a set of 15 microsatellites (González et al., 2019), and assigned to K=2 clusters. Each vertical bar depicts an individual and is coloured to denote the probability that each individual belongs to cluster 1 (yellow) or cluster 2 (blue). Individuals in this study are classified as either *V. v. mensalis* or *V. v. vicugna*, but you can see that the genetic data tells a more complicated story, with no convenient separation between these two subspecies and lots of hybridisation occurring in the contact zone between them.

With the above R tools, and others in these packages, run sNMF for K=2 for your samples. Use the results to plot a bar-chart of individuals. The dataset can be found here:

./MolGen-2025/2.Frogs_RADseq/Data/RADstructure/Hyla.struct.geno

Q. What do the sNMF results tell you about the distribution of diversity between these populations?
Q. What is "K " and how can you use this to help infer whether these frogs fall into two species or not?
Q. Do you think these two lineages should be considered separate species and why?
Whatever your conclusion is, remember to logically back up your decision!

Project 3. Genomic analyses of divergence between Lake Malawi cichlids



Brawand et al., 2014

Cichlids are an extremely diverse group which make up over 10% of all Teleost (ray-finned) fish. The most famous are those found in the lakes of Africa, which have become textbook models for the study of speciation due to the extremely rapid radiation of thousands of species in only 1-2 Million years (Kocher, 2004). They are also extremely important models for the study of **ecological speciation**, as most of the divergence in this group has arisen while they were still in sympatry.

In this project we will analyse some data from Hahn et al. (2017) who studied the divergence between four species of cichlid. Not only did they do this at a population scale, but also on the genomic scale. That is, they identified the regions of the genome that diverge the most between the species, as a way of generating hypotheses about the mechanism of speciation.

We have provided two files. The first contains RADseq data similar to that analysed in Project 1. The second dataset is perhaps the most interesting and will give you an insight to the bleeding edge of population genomics - i.e. how do **different parts of the genome differ in their evolutionary history**.

This dataset contains F_{ST} that has been calculated **per locus**. i.e. for each locus, a calculated F_{ST} represents the difference in allele frequency between populations at that locus alone. Loci which are near (and so tightly linked) to loci under selection will show specific patterns of F_{ST}. This information will give clues to the type and direction of selection. For example, regions of the genome containing many loci with elevated F_{ST} (relative to the genome average) are likely under diversifying selection. Plotting these values along the genome is therefore a powerful way to identify candidate regions/genes driving evolutionary processes such as speciation.

Population genetics of lake Malawi cichlids

For the first part of this project, you will analyse and explore a SNP dataset from several cichlid species (find the data here: **./MMEE-2025/3.Cichlids/Data/cichlid_hierfstat.dat.**) Use your newly found expertise in R to:

- Look at within population statistics, like observed and expected heterozygosity, inbreeding (F_{IS}) and expectations of Hardy Weinberg. Plot the mean observed and expected heterozygosity for the 4 populations against each other.
- Look at population pairwise F_{ST}. Try to use the basic R plotting function heatmap() with the matrix you obtain using HIERFSTAT.
- Look at population clustering using a PCA.

Q. Do you observe any difference in observed and expected heterozygosity for the 4 populations?
Q. Considering that these 4 species live together in the same lake, how can you interpret the populations pairwise F_{ST} values you obtained? Which can be the biological explanation and evolutionary mechanism that drives this pattern?
Q. Are there some populations which are more diverged than others? If so, can you hypothesise as to why?
Q. Do the individuals of the same populations fall into discrete "clusters"?

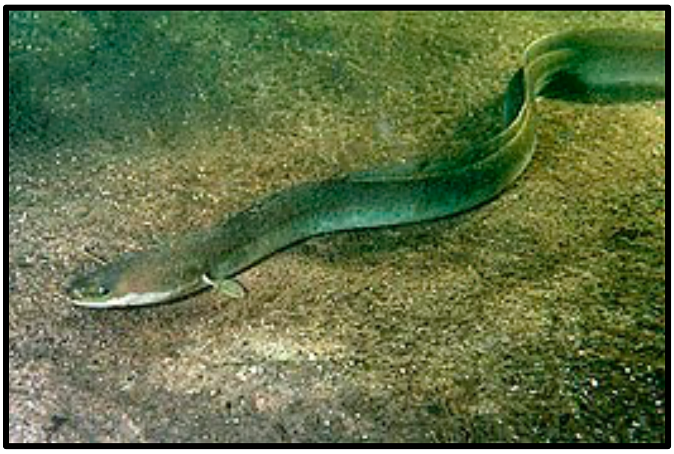
Detecting candidate loci under selection

In the second part of this project, you will identify candidate loci under selection by using a dataset of pairwise locus F_{ST} values obtained using 2 out of the 4 populations. That can be found here: **./MMEE-2025/3.Cichlids/Data/Di_1-Di_2_global_mod.tsv**

- Look at the content of the file and try to understand the different information it provides.
- Generate a "Manhattan" plot using this data (i.e. plot F_{ST} against the location in the genome)
- Use a statistical test to detect possible outliers. HINT: you can do it visually, using a normal distribution or use the R-package called `outliers` (https://cran.r-project.org/web/packages/outliers/outliers.pdf) or `EnvStats` (https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/rosnerTest)

Q. Can you visually identify the presence and number of candidate loci under selection? If so, in which chromosomes are these loci located?
Q. How many of these loci are statistically identified as actual outliers?
Q. How would you use these loci to better understand the evolutionary mechanism underlying the observed genetic differentiation?

Project 4. The curious case of the European Eel



The final dataset we will give you is one from the European Eel, *Anguilla anguilla*. By this point you should be well acquainted with the techniques needed to explore a genetic dataset. Thus, we will not spoil the fun by telling you the story behind this system.

Examine population diversity and especially differentiation between populations (i.e. F_{ST}). Compare the levels of divergence and the geographic distances with those of the previous two datasets and **try to come up with an explanation for what you see**.

For clues, you can read the paper from which this data was taken (Pujolar et al., 2014). This final task represents the final step of any scientific project - once you have found an interesting result, you must go looking through the literature to see how it compares to what others have found.

Bibliography

Dufresnes, C., Mazepa, G., Rodrigues, N., Brelsford, A., Litvinchuk, S. N., Sermier, R., et al. (2018). Genomic Evidence for Cryptic Speciation in Tree Frogs From the Apennine Peninsula, With Description of Hyla perrini sp. nov. *Frontiers in Ecology and Evolution* 6.

González, B. A., Vásquez, J. P., Gómez-Uchida, D., Cortés, J., Rivera, R., Aravena, N., et al. (2019). Phylogeography and Population Genetics of Vicugna vicugna: Evolution in the Arid Andean High Plateau. *Front. Genet.* 10, 445.

Hahn, C., Genner, M. J., Turner, G. F., and Joyce, D. A. (2017). The genomic basis of cichlid fish adaptation within the deepwater "twilight zone" of Lake Malawi. *Evol Lett* 1, 184–198.

Kocher, T. D. (2004). Adaptive evolution and explosive speciation: the cichlid fish model. *Nat. Rev. Genet.* 5, 288–298.

Richmond, J. Q., and Reeder, T. W. (2002). Evidence for parallel ecological speciation in scincid lizards of the Eumeces skiltonianus species group (Squamata: Scincidae). *Evolution* 56, 1498–1513.

Stöck, M., Dubey, S., Klütsch, C., Litvinchuk, S. N., Scheidt, U., and Perrin, N. (2008). Mitochondrial and nuclear phylogeny of circum-Mediterranean tree frogs from the Hyla arborea group. *Mol. Phylogenet. Evol.* 49, 1019–1024.