WORKSHOP CONSERVATION GENOMICS 2025

Study system:

Podarcis raffonei is an insular lizard species endemic to the Aeolian Islands (Sicily). The species currently persists on only **three islets** and **one peninsula**.

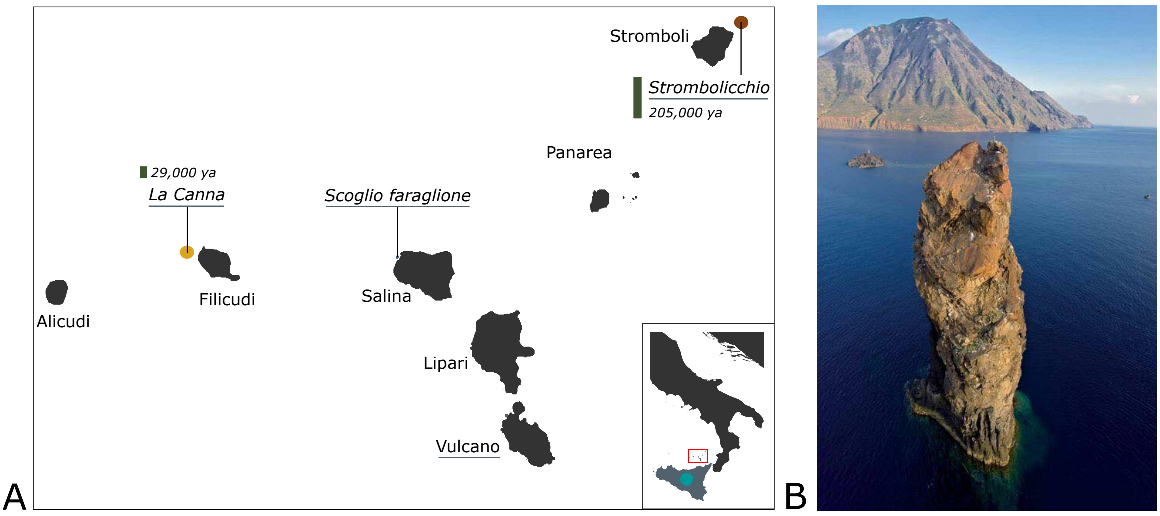
For this workshop, genomic data are available for:

- **La Canna (LC):** 10 individuals, estimated *Ne* ≈ 50
- **Strombolicchio (ST):** 11 individuals, estimated *Ne* ≈ 100–200

Individuals from other species have also been sequenced, including 11 *Podarcis waglerianus* (WG), the siter species of *P. raffonei*, found in Sicily with a large population size.

Raw reads are available on NCBI, and the analyses have been conducted in *Gabrielli et al., In review*.

Photography of an individual of *Podarcis raffonei*, on La Canna stack (credit: Daniele Salvi)



(A) Distribution range of the Aeolian wall lizard and the Sicilian wall lizard (bottom right, different scale). The red square in the inset indicates the position of the Aeolian archipelago. The Aeolian wall lizard is found in four islands and islets (underlined in grey), including our study sites of La Canna and Strombolicchio. The geologically estimated age of La Canna and Strombolicchio is indicated in italic. (B) Picture of the La Canna stack (credit: Piera Rapisarda), a volcanic neck 70m high and 1.6km distant from the island of Filicudi in the background.

Connect to the genobioinfo cluster:
ssh user@genobioinfo.toulouse.inrae.fr
Enter your password

Copy the folder with all data and scripts to your own folder
You can work directly from your space /home/user/work
cd /home/user/work
cp -r /work/project/latitudinal_mysteries/maeva/workshop2025 ./

Part 1 – Genetic diversity and PCA

In this section, we will perform preliminary analyses to estimate genetic diversity within each population and describe population structure.

Tasks

1. What is the **number of SNPs and individuals** of the vcf file ?
2. What is the **number of SNPs and individuals** in each of the **three groups** (LC, ST, waglerianus)?
3. **Compute heterozygote counts** in each **individual** using vcftools https://vcftools.sourceforge.net/ (script: 1a_genetic_diversity.sh).
4. **Run a PCA** on the SNP dataset using the provided PLINK script (script: 1b_PCA.sh).
5. Transfer output files to your computer, and **plot the PCA** in R (script plot_pca_plink.R)
6. **Interpretation:**
    - What does the PCA indicate about population structure?
    - Do LC and ST individuals cluster separately?
    - Are there signs of substructure or outliers?

Part 2 – Runs of homozygosity

We will estimate **runs of homozygosity (ROH)** to evaluate levels of inbreeding. We use **ROHan**: https://github.com/grenaud/ROHan
Because ROHan can be computationally intensive, each participant will run it on **one individual** from each of the three populations. You will be assigned your individual during the session.

Tasks

1. Run ROHan for your assigned individual (script: 2a_ROHan.sh)
2. Examine the **.summary** file and calculate the **percentage of the genome in ROH** for your individual (script: 2b_stats_ROH.sh).
3. Review the **PDF output files**:
    - **.het.pdf**: What patterns do you observe regarding heterozygosity across the genome?
    - **.hmm.pdf**: What does the hidden Markov model inference reveal about ROH segments?
4. Report the **mean ROH length** for your individual

5. Plot the proportion of the genome in ROH>5Mb; 2-5Mb; 1-2Mb and 500km-1Mb (script: plot_ROH.R)
6. **Interpretation:**
   - What do your results suggest about the **inbreeding level** in your assigned population?
   - How do proportions of different ROH classes compare among individuals or populations?

Part 3 – Genetic load using annotation-based tools

We will now estimate genetic load using an annotation-based approach with **SnpEff**: https://pcingola.github.io/SnpEff/snpeff/introduction/
To use SnpEff, you need a **well-annotated reference genome**. The software uses this annotation to predict the effect of SNPs in a VCF file based on their impact on the encoded amino acids. SnpEff adds an **ANN** field to each variant in the VCF, classifying mutations into four broad categories of predicted impact:

- **LOW** – include synonymous mutations
- **MODERATE** – include non-synonymous (missense) mutations
- **MODIFIER** – different effects (see site)
- **HIGH** – include variants introducing or removing a STOP codon, likely to impair protein structure and function

In this section, we will focus on:

- **Missense mutations** (MODERATE), which are generally considered *slightly deleterious*
- **Nonsense mutations** (HIGH), typically *strongly deleterious*
- **Synonymous mutations** (LOW), which can serve as a **neutral control** for differences in genetic diversity among populations due to demographic history

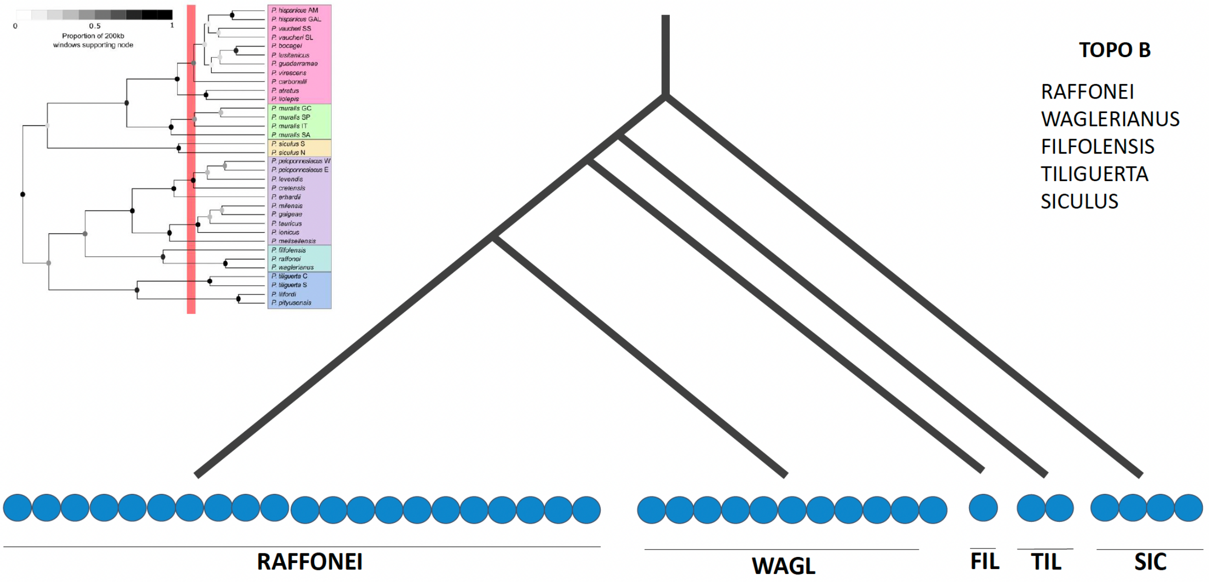
Tasks

1. **Run SnpEff** to annotate your VCF (script: 3a_SnpEff.sh). The initial step, building the SnpEff genome database for your reference, has already been completed.
2. **Generate genotype counts** for each individual (script: 3b_count_genotypes.sh).
3. **Interpretation**
   - What do your results suggest about patterns of genetic load in the three groups of interest?

Part 4 – Genetic load using annotation-based tools, and with outgroups

The goal of this section is to demonstrate the importance of **outgroups** for **polarizing allele,** that is, identifying which allele is **ancestral** and which is **derived**. This allows us to count **homozygous derived genotypes** that are predicted to be deleterious.

Below is the phylogeny of *Podarcis* species relevant to our study.



Tasks

1. **Identify suitable outgroups**
   - Which species in this phylogeny can be used as outgroups?

We will then use a Python script to assign an **ancestral state** to each site. Only sites where the ancestral allele can be inferred will be retained, this requires both outgroups to be **homozygous** (some missing data is tolerated to avoid losing too many sites). The script assigns different tags depending on the allele frequency in the focal populations (details in the associated presentation).

2. **Generate the ancestral-state tag** file for the VCF (script: 4a_command_pol_miss.sh).
3. **Generate genotype counts**, this time using the ancestral state information from the tag file (script: 4b_count_genotypes.sh).
4. **Interpretation**
   - Compare these polarized results with those obtained without polarization.
   - What do your results now suggest about genetic load in the three groups of interest?