Using Galaxy EU and R to Analyze RAD-seq Data

Audrey Jones
BIOL.5602: Bioinformatic Tools in Sequence Analysis
December 7, 2025

Main Databases, Tool, and Datasets Used:
SRR034310
https://usegalaxy.eu/
R

RADseq Galaxy EU Video Tutorial

Purpose:

The purpose of this manual is to provide a comprehensive guide for approaching RAD-seq (restriction site-associated DNA sequencing) data stored in public databases to draw population-level conclusions using Galaxy EU and R. It provides a step-by-step workflow on downloading and filtering RAD-seq data using Galaxy EU. Following this, are instructions on how to code to conduct further downstream analysis and produce box plots and principal component analysis (PCA) of this data using R programming.

Background:

Restriction site-associated DNA sequencing data, also known as RAD-seq data, is a high-throughput reduced-representation sequencing technique used to identify genetic variants for drawing population-level conclusions (1). It is used in applications such as tracking evolution, identifying adaptive traits, comparing populations, and aiding conservation efforts. To generate RAD-seq data DNA from multiple individuals of the same species is digested with a specific restriction enzyme. Due to the specificity of these enzymes, the regions of the fragmented DNA produced are homologous for all individuals (2). These fragments are sequenced using methods like Illumina. The reduced complexity achieved by restriction enzymes allows better accuracy, coverage, and ease of alignment. These advantages allow RAD-seq data to be utilized for capturing different variants between ecotypes which are distinct populations of a species that are genetically distinct due to their specific environment (3). RAD-seq data however does not always show the full scope of genetic variation present as only SNPs near restriction enzyme recognition sites are sequenced. This makes it hard to compare information across different studies, as the use of restriction enzymes results in reduced and missing data that may differ significantly based on the enzyme used and sensitivity of the species to the enzyme. Despite this, RAD-seq still serves as a powerful tool to develop population-wide genomic insights.

Bioinformatic tools are needed to process RAD-seq reads to ensure the quality and reliability of the data, and then to identify genetic variation statistics at the population level. The reads generated from multiple individuals can then be aligned, and variants such as single nucleotide polymorphisms (SNPs) are called. This allows genetic diversity to be both identified and quantified. Further downstream analysis can then be done allowing this data to be visualized and evolutionary patterns in terms of genetic variation to be better understood. The goal of this manual is to translate the raw RAD-seq data from the dataset into meaningful information on Galaxy EU, and then to produce a PCA and two boxplots of the data on R programming. The steps used within Galaxy EU to work with the dataset are outlined in the attached video tutorial.

Tool/Dataset:

The dataset used for this manual is RAD-seq data (SRR034310) collected by The University of Oregon. This data was originally used to identify genetic diversity in the form of SNPs between freshwater and marine three-spined stickleback populations (4). It contains 45,000 sequenced SNPs from 100 three-spined sticklebacks from two marine and three freshwater populations. This makes it particularly useful for this manual as the resulting RAD-seq analysis may show true genomic differences. To produce the RAD-seq data, the restriction enzyme SbfI was used to fragment the three-spined stickle backs genome into reproducible fragments (5). Illumina sequencing was then completed to sequence the resulting fragments. Data from 16 of these three-spined sticklebacks, as denoted by the barcode file and population map from Galaxy EU Training Network, were used for further filtering an analysis (6).

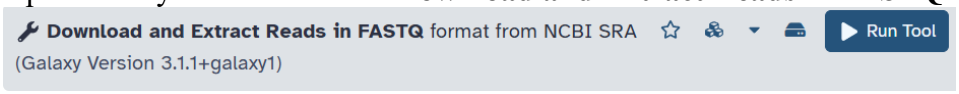
To derive usable information from the RAD-seq dataset in this manual, Galaxy EU was utilized (5). Galaxy EU is a useful web-based bioinformatics platform. It provides access to thousands of different programs allowing reproducible workflows to be produced through one interface. Within Galaxy EU, the dataset was uploaded via the *Download and Extract Reads in FASTQ* tool. *Stacks2: process RAD tags* program was then used to demultiplex the reads (remove barcode sequence) and filter out low-quality reads. *Falco* was then applied to check the sequence quality. Next, the *Stacks2: de novo map* tool grouped similar reads from each sample to call variants and identify SNPs. Following this, the *Stacks2: populations* tool was used to perform population-level analyses of the resulting variants.

Further downstream analysis of the data generated in Galaxy EU was performed using R programming. The use of R enables the data produced in Galaxy EU to be directly used in various visualizations and statistical tests, allowing additional biological conclusions to be drawn. This includes a principal component analysis (PCA), and multiple boxplots. The PCA displays the relatedness of SNPs between the three-spined stickleback samples. The first boxplot shows the nucleotide diversity in terms of pi across ecotypes, allowing comparison of levels of genetic diversity per locus. Lastly, the second boxplot produced compares the heterozygosity of genotypes across ecotypes to observe differences in genetic diversity between the freshwater and marine three-spined stickleback.

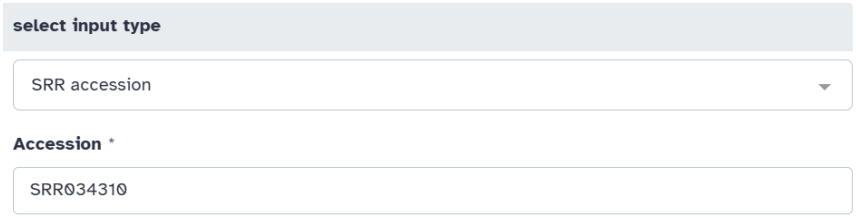
Manual:

Load data into Galaxy EU via the SRR accession number found in NCBI

1. Locate the desired dataset in NCBI and copy the SRR accession number (SRR034310).
2. Open Galaxy EU and load the **Download and Extract Reads in FASTQ** tool.



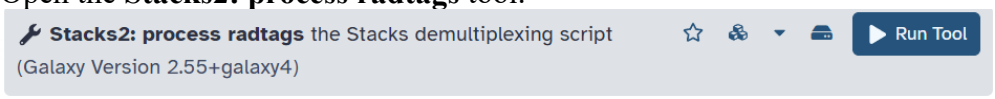
3. Under select input type, choose “SRR accession” and paste the SRR accession number into the accession box.



4. Run the tool to import data.

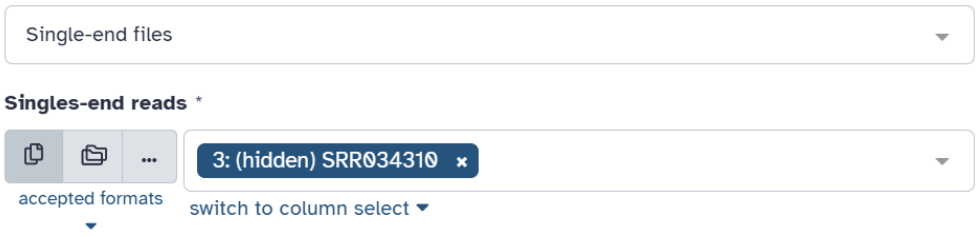
Demultiplex reads in Galaxy EU

1. Open the **Stacks2: process radtags** tool.

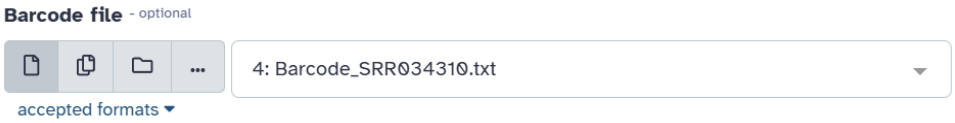


2. Set tool parameters to read the single-end data produced (labeled under the accession number previously input). This is because single-end reads were used in the original

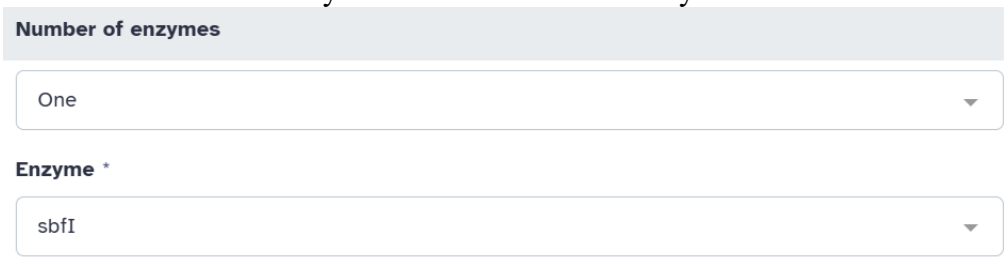
RAD-seq dataset.



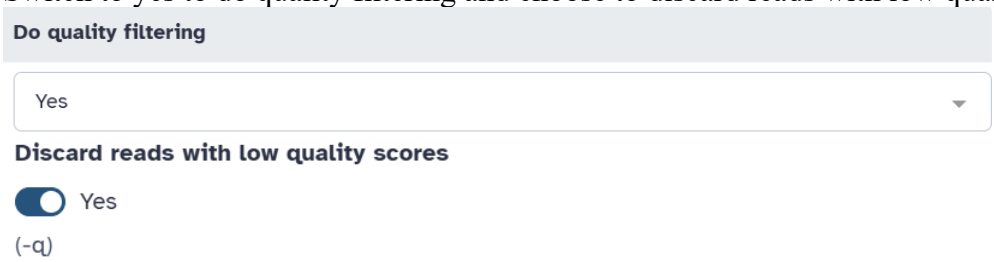
3. Insert a barcode file by selecting upload dataset and uploading the below txt file from the Galaxy EU Training Network (6). This will select data from 16 specific three-spined sticklebacks based on the barcodes provided.
    a. SRR034310 Barcode File.txt



4. Under the number of enzymes select one and the enzyme SbfI.



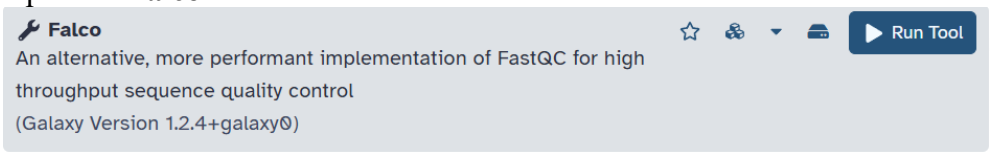
5. Switch to yes to do quality filtering and choose to discard reads with low quality scores.



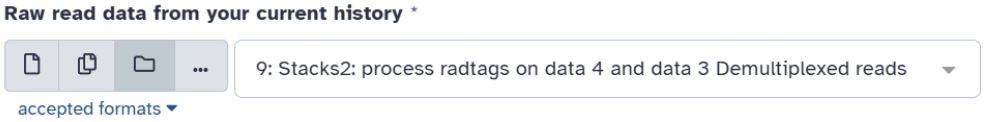
6. Run the tool to separate the data into reads from individual samples and filter out low quality reads.

Run Falco in Galaxy EU

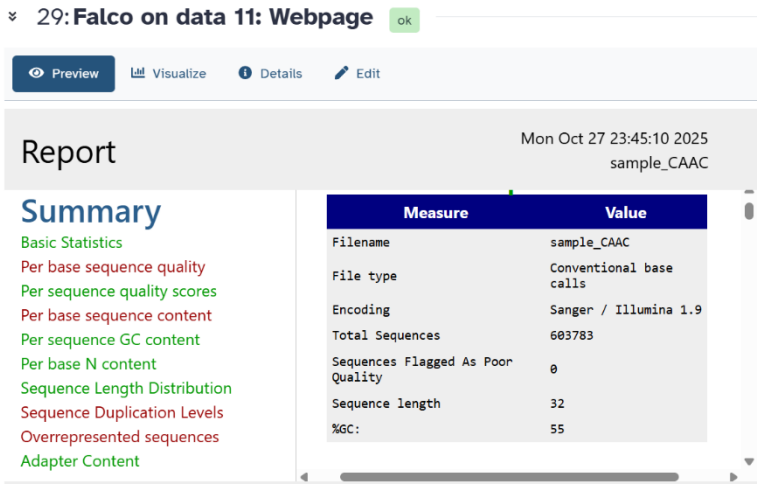
1. Open the **Falco** tool.



2. Under *raw read data* from your current history, select the outputted demultiplexed reads from Stacks2: process radtags.

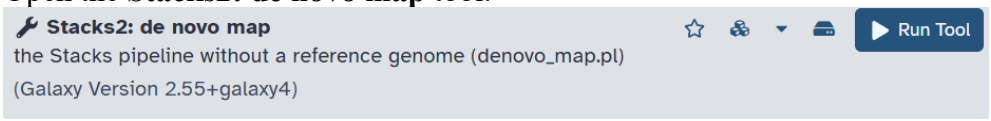


3. Run the tool to complete quality control and ensure sequence. This tool is used rather than FastQC as it is faster and optimized to be used on RAD-seq data within the Stacks2 pipeline.
4. Analyze the expected outcome, as seen below, to confirm that low quality sequences have been removed.

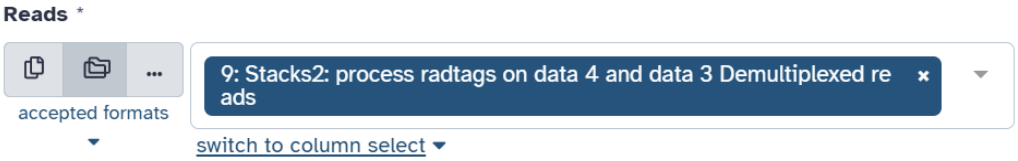


Run the Stacks2: de novo map tool in Galaxy EU

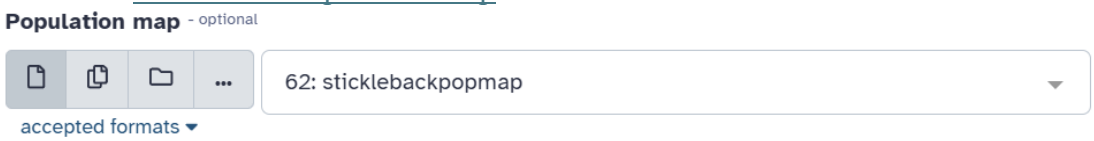
1. Open the **Stacks2: de novo map** tool.



2. Under *short read data from individuals* select the outputted demultiplexed reads from Stacks2: process radtags for the reads category.



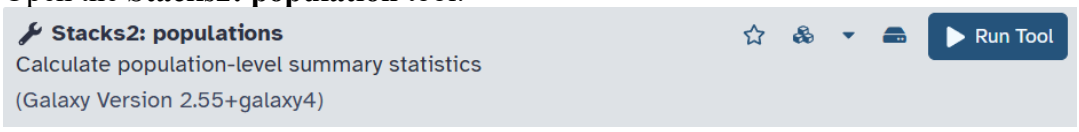
3. Insert a population map by selecting upload dataset and uploading the file below from the Galaxy EU Training Network (6). This will allow the 16 three-spined sticklebacks selected via the previous barcode file to be used in further analysis.
    a. Stickleback Population Map



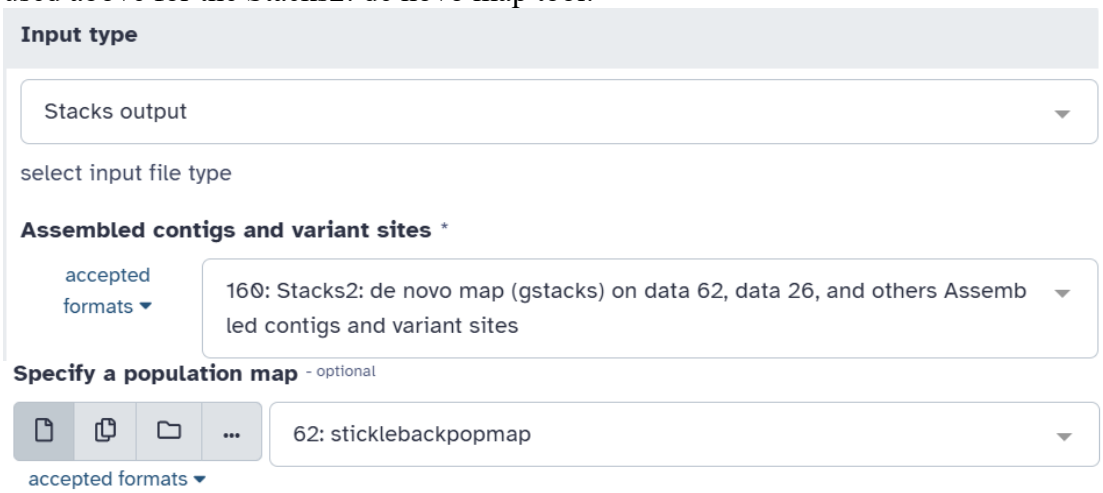
4. Run the tool to map the reads without the use of a reference genome.

Run the Stacks2: populations tool in Galaxy EU

1. Open the **Stacks2: population** tool.

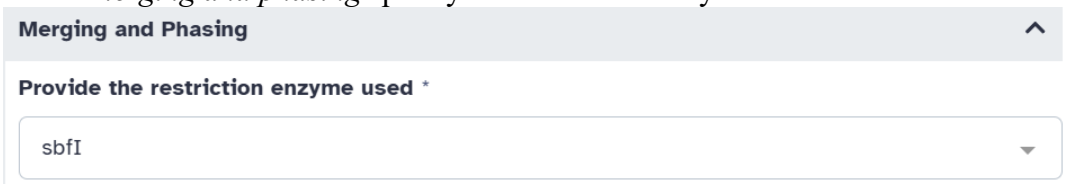


2. Under *input type* select the stacks output from the Stacks2: de novo map tool containing the gstacks data. The gstacks data is used as it contains the finalized genotype data generated by the above Galaxy EU workflow. Then, upload the same population map used above for the Stacks2: de novo map tool.

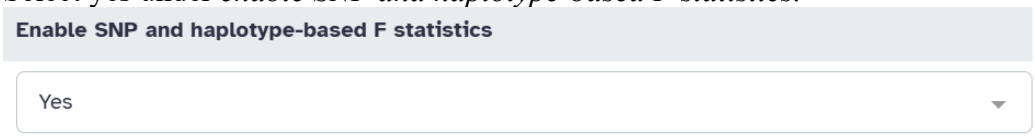


3. Under *data filtering options* change the minimum percentage of individuals in a population required to process a locus for that population to 0.75. This allows some biological variation to be present across indivduals.

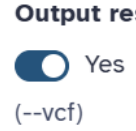
4. Under *merging and phasing* specify the restriction enzyme used as SbfI.



5. Select yes under *enable SNP and haplotype-based F statistics.*



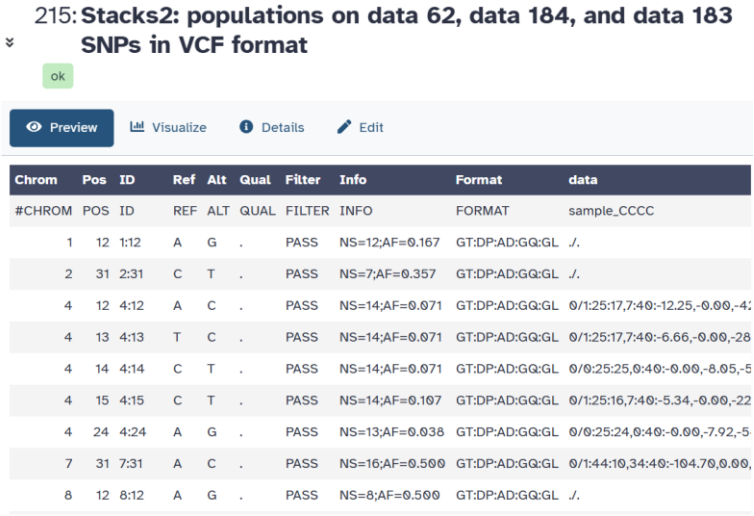
6. Under *output options*, choose to output the data in variant call format (VCF).

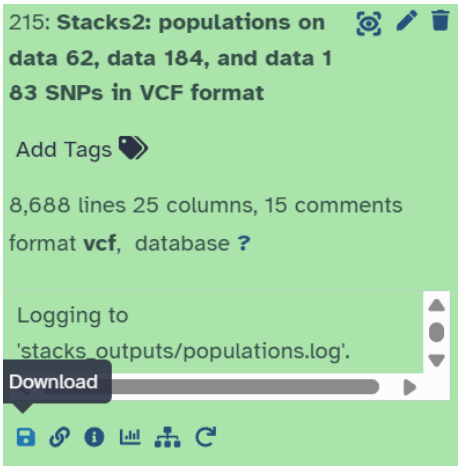


7. Run the tool to calculate population-level summary statistics.

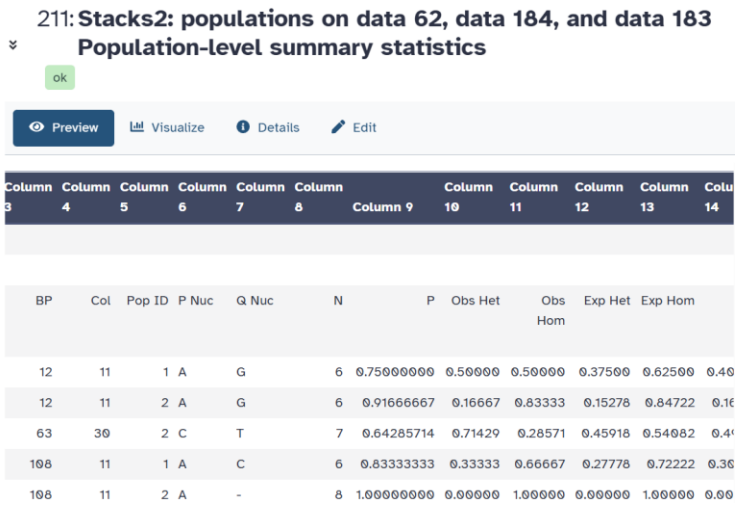
Export files from Galaxy EU

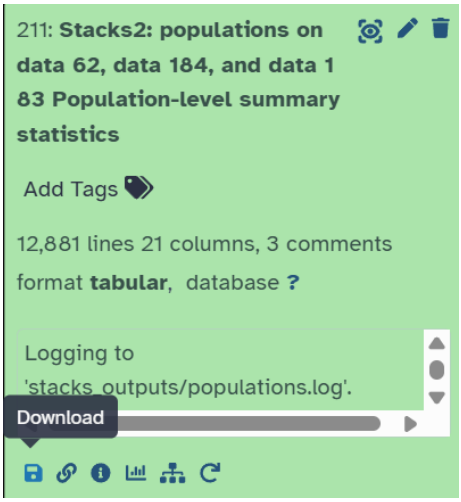
1. Find the dataset from Stacks2: populations that show the SNPs in VCF format. The VCF file contains the genetic variants identified in the SNPs for each of the 16 three-spined sticklebacks pulled out of the RAD-seq dataset. Click on this dataset and in the bottom left corner select the download option to download this file.





2. Find the dataset from Stacks2: populations that show the population-level summary statistics. Click on this dataset and in the bottom left corner select the download option to download this file.



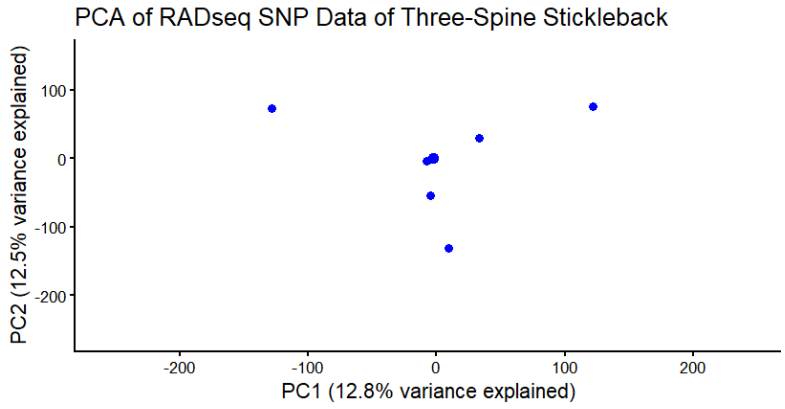


3. Move these two downloads into a new folder on your computer desktop called "RADseq." This will allow the file to be easily located for further processing on R programming.

Code for a PCA Plot in R using the SNPs VCF

The complete R code can be accessed here - RADseq PCA Plot.R

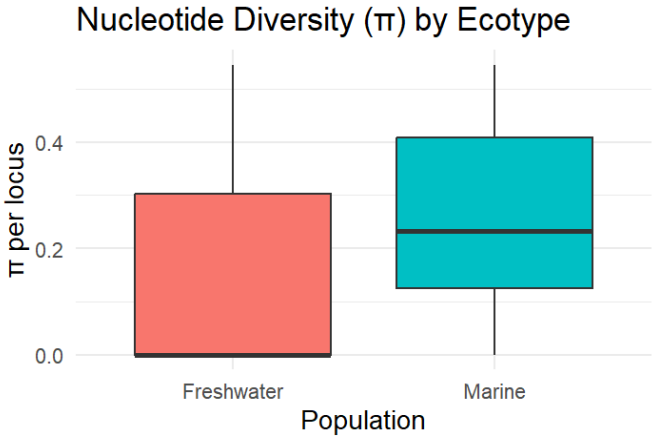
1. Install the packages "*vcfr*", "*adegenet*" and "*ggplot2*", then load their libraries.
```
if (!require("vcfR")) install.packages("vcfR")
library(vcfR)
if (!require("adegenet"))install.packages("adegenet")
library(adegenet)
if (!require("ggplot2")) install.packages("ggplot2")
library(ggplot2)
```
2. Set the working directory in R to the "RADseq" folder on your computer desktop using the function and locate the previously saved SNPs VCF file within this folder.
```
setwd("C:/Users/audre/OneDrive/Desktop/RADseq")
getwd()
list.files(pattern = "vcf")
```
3. Read the imported VCF and fill in missing values with NA. This is to clearly mark missing values so further anyalsis aren't altered.
```
snps <- vcfR::read.vcfR("GalaxyRADseqSNPvcf.vcf", convertNA  = TRUE)
```
4. Convert the VCF into a geneid product that the *adegenet* package can work with.
```
genind_obj <- vcfR2genind(snps)
```
5. Extract the genotypes present in the geneid to a usable numeric form and fill in missing genotype values with the mean genotype value.
```
snps_num <- tab(genind_obj, NA.method = "mean")
```
6. Remove SNPs with zero variance as they cannot contribute to the PCA.
```
snps_num <- snps_num[, apply(snps_num, 2, var) != 0]
```
7. Run the PCA. Set center equal to true, so the PCA focuses on variation around the mean.
```
pca_res <- prcomp(snps_num, center = TRUE, scale. = TRUE)
```
8. Convert the PCA into a data frame, allowing it to be plotted.
```
pca_df <- as.data.frame(pca_res$x)
```
9. Edit the x and y axes to display the percentage variance explained.
```
xlab_text <- paste0("PC1 (", round(summary(pca_res)$importance[2,1]*100, 1), "% variance explained)")
ylab_text <- paste0("PC2 (", round(summary(pca_res)$importance[2,2]*100, 1), "% variance explained)")
```
10. Create the PCA plot, edit titles, and theme accordingly.
```
ggplot(pca_df, aes(x = PC1, y = PC2)) +
geom_point(color = "blue", size = 2) +
theme_classic(base_size = 12) +
labs(title = "PCA of RADseq SNP Data of Three-Spine Stickleback", x = xlab_text, y = ylab_text) +
coord_cartesian(xlim = c(min(pca_df$PC1) * 2, max(pca_df$PC1) * 2),ylim = c(min(pca_df$PC2) * 2, max(pca_df$PC2) * 2))
```
11. Expected output. It shows a PCA plot representing RAD-seq data from marine and freshwater three-spined stickleback, with each dot representing a singular fish. It is seen that there is no clear clustering representing different genomic backgrounds of the marine and freshwater stickleback SNPs sequenced.



Code for Visuals in R using the population-level summary statistics
The complete R code can be accessed here - RADseq Boxplots.R

1. Install the packages "*ggplot2*" and "*dplyr*", then load their libraries.
```
if (!require("ggplot2")) install.packages("ggplot2")
library(ggplot2)
if (!require("dplyr")) install.packages("dplyr")
library(dplyr)
```
2. Set the working directory in R to the "RADseq" folder on your computer desktop using the function and locate the previously saved population-level summary statistics tabular file within this folder.
```
setwd("C:/Users/audre/OneDrive/Desktop/RADseq")
getwd()
stats_file <- "populationstats.tabular"
```
3. Convert the tabular file into a data frame, so R can use the data to formulate the boxplots. Then, note that the headers are not present, so the data is accurately represented in the data frame.
```
stats_df <- read.delim("populationstats.tabular", header = FALSE)
```
4. Name the columns of the data frame in accordance with the labels present in Galaxy EU.
```
colnames(stats_df) <- c("LocusID","Chr","BP","Col","PopID",+
"P_Nuc","Q_Nuc","N","P","Obs_Het","Obs_Hom","Exp_Het","Exp_Hom",+
"Pi","Pi_smooth","Pi_p","Fis","Fis_smooth","Fis_p","HWE_p","Private")
```
5. Reassign the labels for the PopID column to show freshwater and marine, rather than one and two.
```
stats_df$PopID <- factor(stats_df$PopID, levels = c(1,2),+
labels = c("Freshwater","Marine"))
```
6. Convert the data in the Pi, Fis, Obs_het, Exp_het, and Fis columns to be treated as numbers.
```
stats_df$Pi       <- as.numeric(stats_df$Pi)
stats_df$Fis      <- as.numeric(stats_df$Fis)
stats_df$Obs_Het   <- as.numeric(stats_df$Obs_Het)
stats_df$Exp_Het   <- as.numeric(stats_df$Exp_Het)
stats_df$Fis_p    <- as.numeric(stats_df$Fis_p)
```
7. Remove the data that does not have a PopID associated with it. This is because this data is not associated with 16 three-spined sticklebacks being used for analysis.
```
stats_df_edit <- stats_df %>% filter(!is.na(PopID))
```
8. Make a boxplot of pi vs. ecotype…
    a. Use the data frame produced above and set the x-axis to represent the PopID and the y-axis to represent the pi value. Then, set the boxplot to fill in color based on the PopID.
    ```
    ggplot(stats_df_edit, aes(x = PopID, y = Pi, fill = PopID)) +
    ```
    b. Create the boxplot.
    ```
    geom_boxplot() +
    ```
    c. Provide and label for the chart, x-axis, and y-axis.
    ```
    labs(title = "Nucleotide Diversity (Pi) by Ecotype",+
    x = "Population",y = "Pi per locus")
    ```

d. Expected output. It shows that marine three-spined sticklebacks have a greater average pi per locus compared to the freshwater populations.



9. Make a boxplot of observed heterozygosity vs. ecotype…
   a. Use the data frame produced above and set the x-axis to represent the PopID and the y-axis to represent Obs_het. Then, set the boxplot to fill in color based on the PopID.

```
ggplot(stats_df_edit, aes(x = PopID, y = Obs_Het, fill = PopID)) +
```

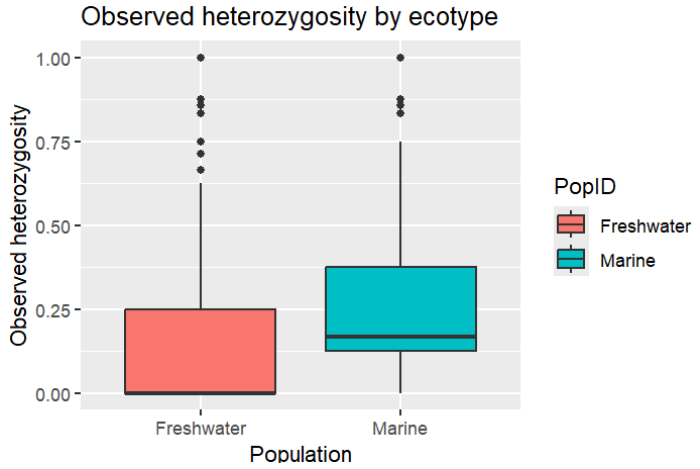
   b. Create the boxplot.

```
geom_boxplot() +
```

   c. Provide labels for the chart, x-axis, and y-axis.

```
labs(title="Observed heterozygosity by ecotype", +
x="Population", y="Observed heterozygosity")
```

   d. Expected output. It shows that marine three-spined sticklebacks have a greater observed heterozygosity than the freshwater populations.



Conclusion:

Using the bioinformatic tools Galaxy EU and R, RAD-seq data from freshwater and marine populations of three-spined stickleback can be evaluated for population-level genomic analysis. Galaxy EU can process Illumina reads and perform quality filtering, demultiplexing, mapping, and variant calling. R then directly utilizes this data to produce visuals representing the

genomic variation within this data. The PCA plot generated in R displays the relatedness of SNPs between the three-spined stickleback samples. This plot, as seen in Figure 1, showed that there is no clear clustering of the freshwater and marine three-spined stickleback population based on genetic similarities. This may be due to only 16 three-spined sticklebacks being compared, due to the small size of the barcode file and population map from the Galaxy EU Training Network (6). This is hypothesized, as the original study from the University of Oregon demonstrated significant differences in freshwater and marine three-spined stickleback populations when examining data from all 100 fish (4). Next, the boxplot of nucleotide diversity (pi) in comparison to ecotype shows the nucleotide diversity in terms of pi across ecotypes, allowing comparison of levels of genetic diversity per locus. As seen in Figure 2, this plot indicated that the marine three-spined stickleback population has a higher average pi value and thus has a higher amount of genetic diversity. Following this, a boxplot comparing the heterozygosity of genotypes across ecotypes to observe differences in genetic diversity between the freshwater and marine three-spined stickleback was completed. This plot, as seen in Figure 3, showed that the marine population has a higher median observed heterozygosity, signaling that they maintain greater genetic diversity.

Overall, analysis of RAD-seq data in Galaxy EU and R demonstrated that marine three-spined stickleback populations retain greater genetic diversity than freshwater populations. This may be because the marine three-spined stickleback is both ancestral and has a larger population size, allowing greater levels of genetic diversity to be maintained. While this manual focused on the three-spined stickleback, this workflow can be adapted to be used with other data sets. For example, it can be applied to examine the effects of selective pressure within ecotypes, trace evolutionary relationships, or look at genetic structure/variation in conservation studies.

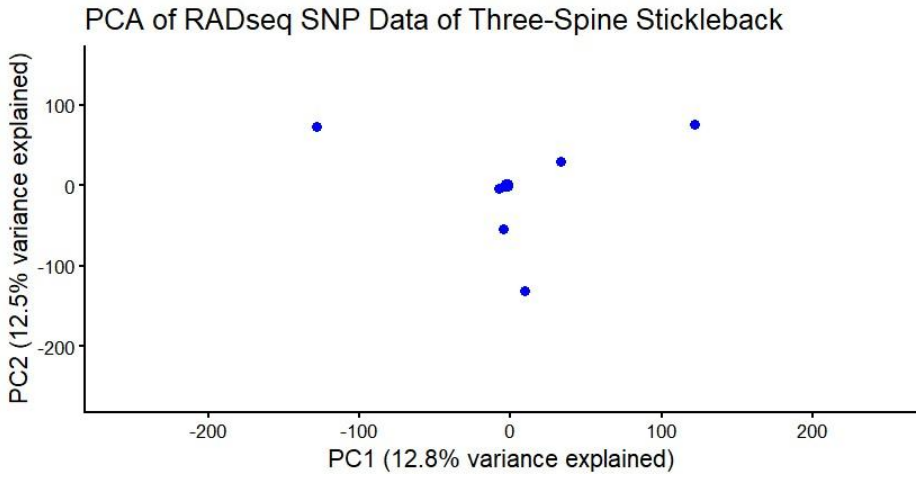


Figure 1: PCA of three-spine stickleback RADseq data. This PCA plot shows genetic variation of freshwater and marine three-spine stickleback based on RAD-seq data for SNPs. Each dot represents a singular fish.

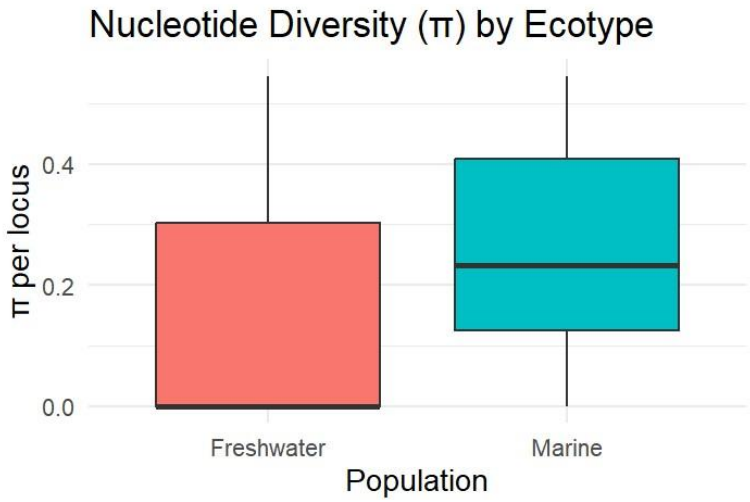


Figure 2: Nucleotide diversity (pi) in comparison to ecotype. A boxplot is seen showing nucleotide diversity in terms of pi between freshwater and marine three-spine stickleback. Marine three-spine stickleback are seen to have a greater pi value per locus.

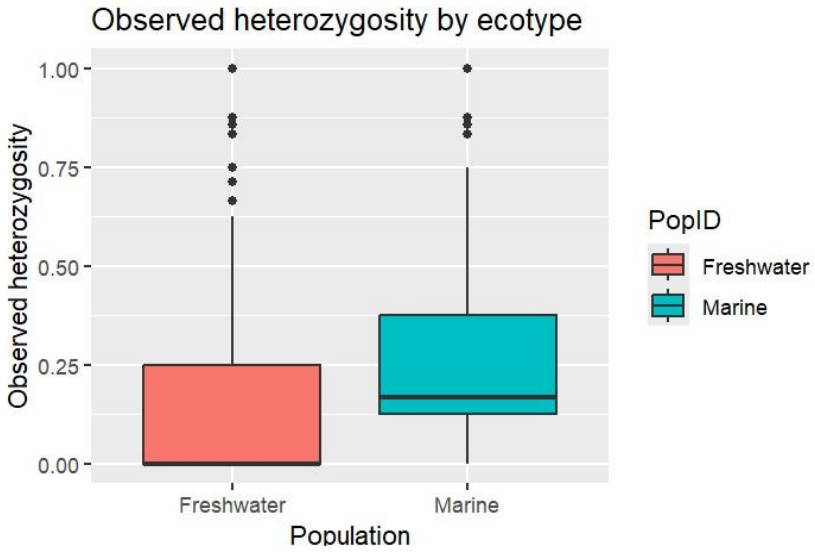


Figure 3: Observed heterozygosity in comparison to ecotype. A boxplot is seen showing observed heterozygosity between freshwater and marine three-spine stickleback. Marine three-spine stickleback are seen to have a higher median observed heterozygosity.

Sources

1. Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics*, *9*(5-6), 416–423. https://doi.org/10.1093/bfgp/elq031
2. Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome research*, *17*(2), 240–248. https://doi.org/10.1101/gr.5681207
3. John E. McCormack, Sarah M. Hird, Amanda J. Zellmer, Bryan C. Carstens, Robb T. Brumfield, (2013). Applications of next-generation sequencing to phylogeography and phylogenetics, Molecular Phylogenetics and Evolution, Volume 66, Issue 2, Pages 526-538, ISSN 1055-7903, https://doi.org/10.1016/j.ympev.2011.12.007.
4. Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS genetics*, *6*(2), e1000862. https://doi.org/10.1371/journal.pgen.1000862
5. *SRA Archive: NCBI*. (n.d.). https://www.ncbi.nlm.nih.gov/Traces/index.html?view=study&acc=SRP001747
6. Galaxy EU Training Network. (2025, June 3). *Ecology / RAD-Seq de-novo data analysis / Hands-on: RAD-Seq de-novo data analysis*. https://training.Galaxy EUproject.org/training-material/topics/ecology/tutorials/de-novo-rad-seq/tutorial.html
7. *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. (n.d.). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
8. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354
9. De Sena Brandine, G., & Smith, A. D. (2021). Falco: high-speed FastQC emulation for quality control of sequencing data. *F1000Research*, *8*, 1874. https://doi.org/10.12688/f1000research.21142.2
10. Knaus, B. J., & Grünwald, N. J. (2016). vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44–53. https://doi.org/10.1111/1755-0998.12549
11. Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129
12. *Create Elegant Data Visualizations Using the Grammar of Graphics*. (n.d.). https://ggplot2.tidyverse.org/
13. *A grammar of data manipulation*. (n.d.). https://dplyr.tidyverse.org/