ShiNyP: User Guide

true

Oct 2024

Contents

Chapter 1

Quickstart

Step 1: Pre-install Required Package

```
install.packages("BiocManager")
BiocManager::install(version = "3.19")
BiocManager::install("qvalue")
```

Step 2: Install the *ShiNyP* Package from GitHub

```
install.packages("remotes")
remotes::install_github("TeddYenn/ShiNyP", force = TRUE)
```

Step 3: Start the *ShiNyP* Platform

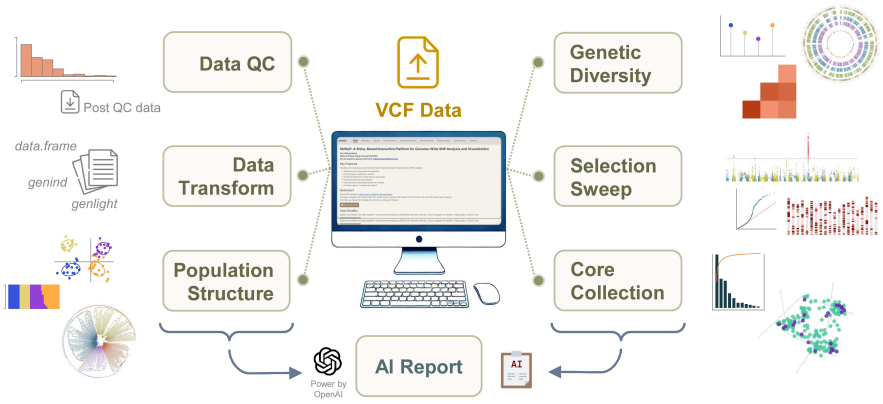
```
library(ShiNyP)
ShiNyP::run_ShiNyP()
```

Step 4: Run *ShiNyP* Analysis

Input your SNP data in VCF format, or feel free to use our **Demo Data**.

> **Note:** If you run in RStudio, you can click the **Open in Browser** button.

This is the user guide page for *ShiNyP*, live at https://teddyenn.github.io/ShiNyP_guide.io/.



ShiNyP is also accessible online at **https://teddyhuang.shinyapps.io/ShiNyP/**. But, please note that due to limited memory usage on this platform, we **DO NOT RECOMMEND** using it to analyze large SNP dataset. The online version is intended solely as a demo website for demonstration purposes. For real data analysis, please consider downloading the platform from GitHub repository **https://github.com/TeddYenn/ShiNyP** and running it locally on the R environment.

- Oct 2024: Initial release v1.0.0 of ***ShiNyP*** on GitHub.

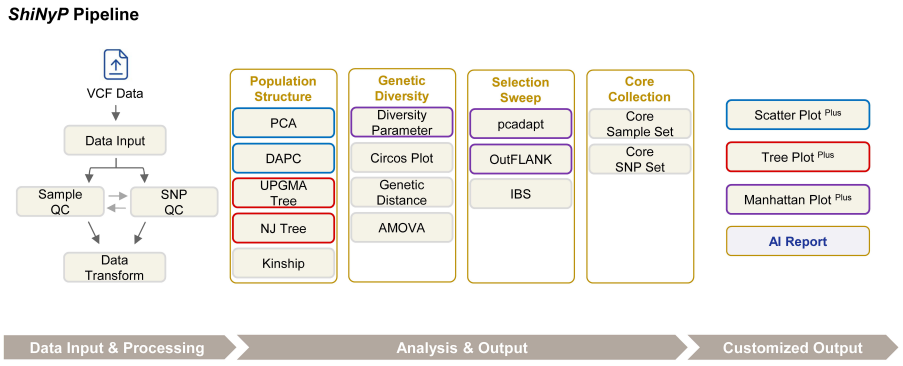
Chapter 2

ShiNyP

Input data: Genome-wide biallelic SNP in Variant Call Format (VCF) file format.

Analysis: Data QC, population genetics analysis, core collection...

Output: Publication-ready figures, tables, analyzed data objects, and AI-driven report.



Key Features

> Statistical and computational exploration
> Customizable visualization options
> Publication-ready figures and tables
> Analyzed data objects
> Auto-generate customized preliminary results
> AI-driven report - powered by OpenAI

Publication

Huang et al. (upcoming 2024) *ShiNyP*: A Shiny-Based Interactive Platform for Genome-Wide SNP Analysis and Visualization **https://www.example.com**

Support

If you encounter any issues or have suggestions for new features, please submit a report through our feedback form: **https://forms.gle/GPCggSo5czyNLfoB7 (Google Form)**

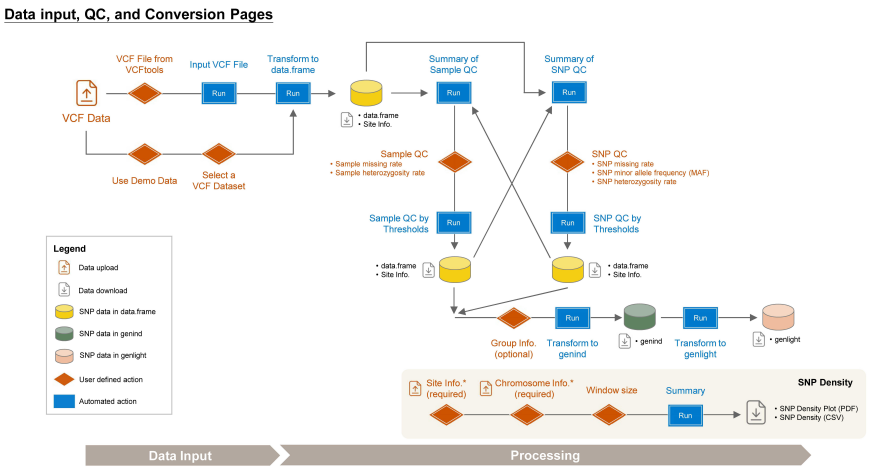
Websites

- **Journal Article**: https://www.example.com
- **GitHub Repository**: https://github.com/TeddYenn/ShiNyP
- **User Manual**: https://teddyenn.github.io/ShiNyP_guide.io/
- **Detailed Description of Statistical Methods**: https://www.example.com
- **Demo Datasets**: https://reurl.cc/QEx5lZ (Google Drive)
- **Demo Platform**: https://teddyhuang.shinyapps.io/ShiNyP/
- **Feedback Form**: https://forms.gle/GPCggSo5czyNLfoB7

Chapter 3

Data Input

This section contains two subpages: **VCF** and **data.frame/genind/genlight**, allowing you to upload various data types for analysis.



3.1 VCF

Required Dataset (one of the following):

- VCF file from PLINK
- VCF or gzipped VCF (vcf.gz) file from VCFtools
- VCF file in RDS format from ***ShiNyP***

The VCF file should contain chromosome and position information in the first two columns (`#CHROM` and `POS`), along with sample names and their genotypic information. For some whole genome sequencing (WGS) data, where SNP marker ID information is missing, ***ShiNyP*** will auto-generate the SNP ID names as #CHROM:POS, such as 2:12500, indicating chromosome 2, position 12500.

Step 1: Input your VCF File

1. Browse and upload one VCF file.
2. If your VCF file is from VCFtools, please tick the 'VCF file from VCFtools' checkbox.
3. After the progress bar shows 'Upload complete', click the **Input VCF File** button.

Or use our Demo Data

1. Click the **Use Demo Data** button and select one species. Detailed descriptions of the demo datasets are available at https://reurl.cc/QEx5lZ (Google Drive).

   Note: By default, the genotypic information for 5 samples and 10 SNPs will be displayed on the interactive table.

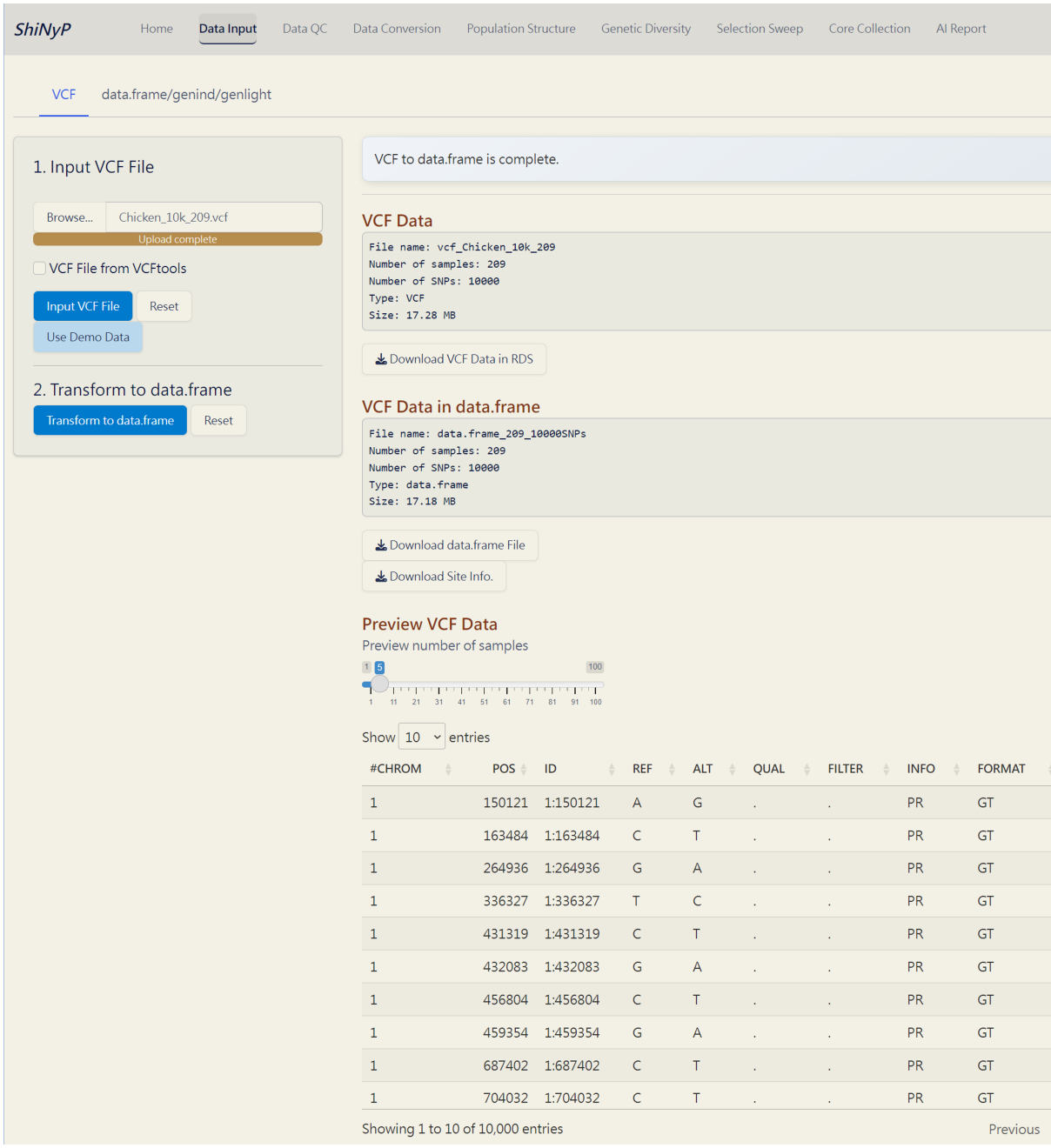
Step 2: Transform to data.frame

1. If you have already input a VCF file on ***ShiNyP***, click the **Transform to data.frame** button.
2. Download the `data.frame` file (in RDS format) and Site Info (in RDS format) so that you will not have to input the VCF file again; instead, you can upload the `data.frame` file.

Outputs:

- **VCF Data (RDS)**: VCF data stored in RDS format, which can be open and read in R environment.
- **data.frame (RDS)**: `data.frame` file. It's necessary for downstream analyses, *please download and save it!*
- **Site Info. (RDS)**: SNP site information file. It's necessary for downstream analyses, *please download and save it!*

Note: If your data is large (more than 1GB), it may take some time to process. Please be patient. The ***ShiNyP*** platform processes one task at a time (e.g., you must wait for the input process to finish before you can reset the data).



VCF Data Input!

3.2 data.frame/genind/genlight

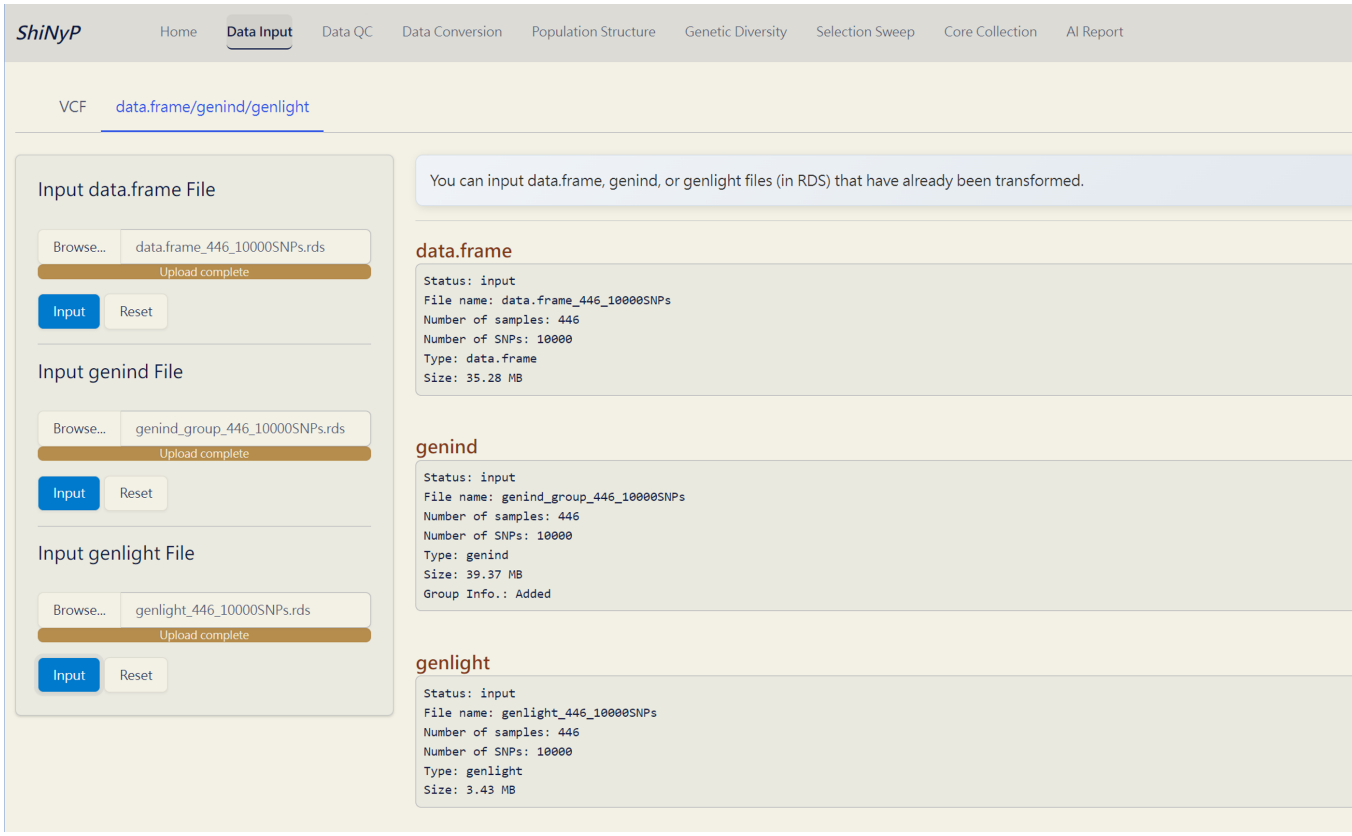
Required Dataset:

- **`data.frame`** in RDS file format
- **`genind`** in RDS file format
- **`genlight`** in RDS file format

`data.frame` file can be downloaded from the subpages VCF, Sample QC, and SNP QC.
`genind` and **`genlight`** files can be downloaded from the Data Conversion page.

One Step:

1. Browse and click the **Input** button to upload your **`data.frame`**, **`genind`**, and **`genlight`** files.



`data.frame`/`genind`/`genlight` Data Input!

Chapter 4

Data QC

This section contains three subpages: **Sample QC**, **SNP QC**, and **SNP Density**, allowing you to assess the quality of samples and SNPs in `data.frame`, as well as visualize SNP density across the genome.

4.1 Sample QC

Required Dataset (one of the following):

- `data.frame` file from the Data Input page
- SNP post-QC `data.frame` file from the subpage Data QC/SNP QC

Step 1: Get Summary First, obtain the sample summary statistics (missing rate and heterozygosity rate) by clicking both **Summary** buttons and you will see the results.

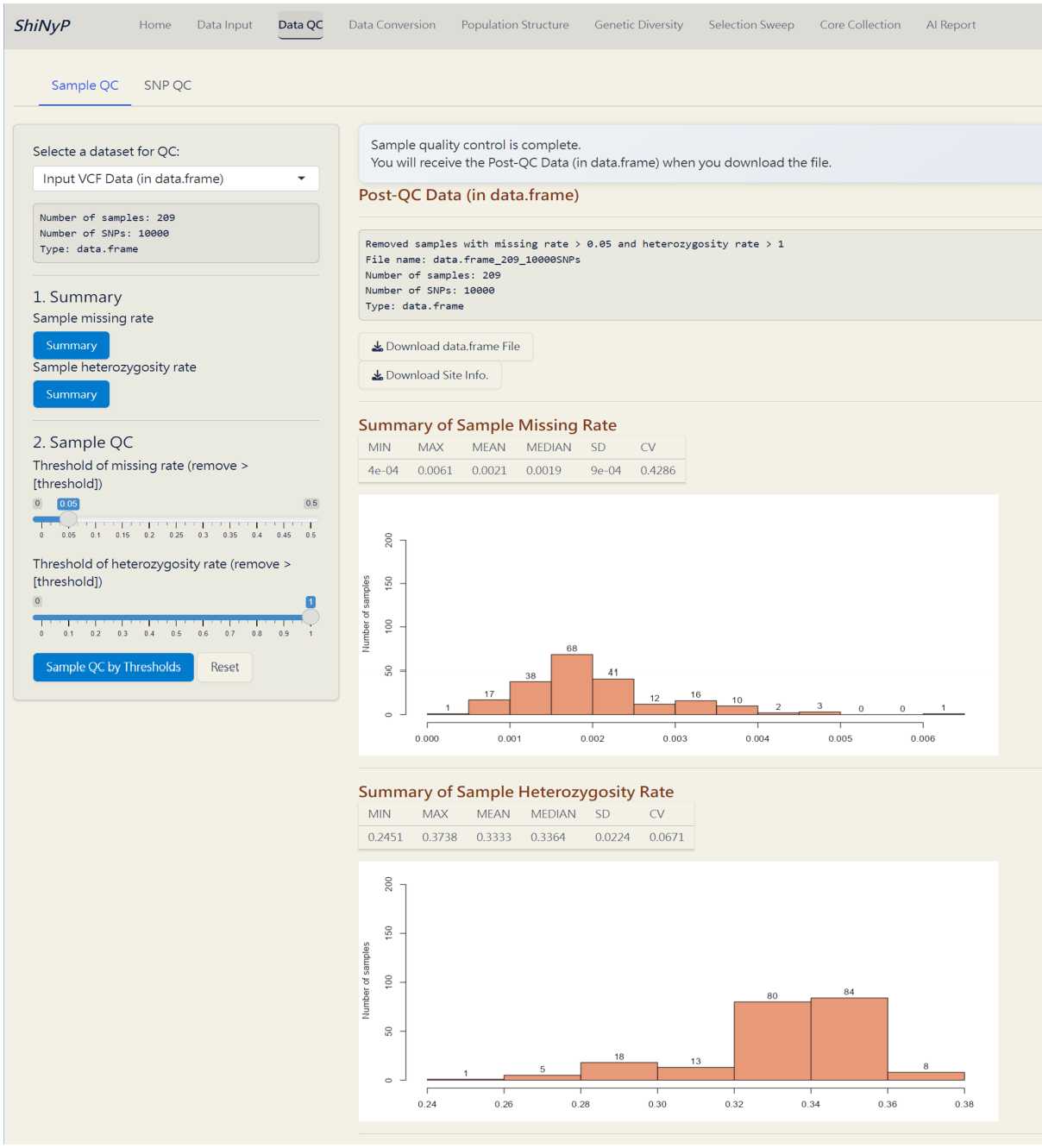
Step 2: Sample QC Adjust the thresholds and click the **Sample QC by Thresholds** button. This will generate the Post-QC `data.frame` file.

> **Note:** If you prefer not to perform sample QC by sample missing rate or heterozygosity rate, please set the threshold to 0.

Outputs:

- **data.frame (RDS)**: Updated `data.frame` file. It's necessary for downstream analyses, *please download and save it!*

- **Site Info. (RDS)**: Updated SNP site information file. It's necessary for downstream analyses, *please download and save it!*



Sample QC Complete!

4.2 SNP QC

Required Dataset (one of the following):

- `data.frame` file from the Data Input page
- Sample post-QC `data.frame` file from the subpage Data QC/Sample QC

Step 1: Get Summary First, obtain the SNP summary statistics [missing rate, minor allele frequency (MAF), heterozygosity rate, and Hardy-Weinberg equilibrium (HWE)] by clicking all **Summary** buttons and you will see the results.

Step 2: Sample QC Adjust the thresholds and click the **SNP QC by Thresholds** button. This will generate the Post-QC `data.frame` file.

> **Note:** If you prefer not to perform QC based on SNP missing rate or heterozygosity rate, set the missing rate threshold to 1, the MAF to 0, and the heterozygosity rate to 0 and 1. Additionally, leave the 'Do SNP QC by HWE' checkbox unticked to skip QC based on SNP HWE.

Outputs:

- **data.frame (RDS)**: Updated `data.frame` file. It's necessary for downstream analyses, *please download and save it!*
- **Site Info. (RDS)**: Updated SNP site information file. It's necessary for downstream analyses, *please download and save it!*


ShiNyP
Home
Data Input
Data QC
Data Conversion
Population Structure
Genetic Diversity
Selection Sweep
Core Collection
AI Report
Sample QC
SNP QC
Selecte a dataset for QC:
Sample Post-QC Data
Post-QC Data (Updated)
Number of samples: 209
Number of SNPs: 9683
Type: data.frame
1. Summary
SNP missing rate
Summary
SNP minor allele frequency (MAF)
Summary
SNP heterozygosity rate
Summary
SNP p-value for Hardy-Weinberg equilibrium (HWE)
Summary
2. SNP QC
Threshold of missing rate (remove > [threshold])
Threshold of MAF (remove < [threshold])
Threshold of heterozygosity rate (remove < [threshold])
Threshold of heterozygosity rate (remove > [threshold])
Threshold of -log(p) for HWE (remove > [threshold])
Do SNP QC by HWE
SNP QC by Thresholds
Reset
SNP quality control is complete.
You will receive the Post-QC Data (in data.frame) when you download the file.
Post-QC Data (in data.frame)
Removed SNPs with missing rate > 0.05, MAF < 0.05, heterozygosity rate < 0, heterozygosity rate > 1, and HWE -log10(p-value) > 6
File name: data.frame_209_9683SNPs
Number of samples: 209
Number of SNPs: 9683
Type: data.frame
Download data.frame File
Download Site Info.
Summary of SNP Missing Rate
MIN MAX MEAN MEDIAN SD CV
0.0000 0.1962 0.0021 0.0000 0.0131 6.3135
Number of SNPs
Summary of SNP Minor Allele Frequency (MAF)
MIN MAX MEAN MEDIAN SD CV
0.1000 0.5000 0.2713 0.2560 0.1190 0.4387
Summary of SNP Heterozygosity Rate
MIN MAX MEAN MEDIAN SD CV
0.0000 0.5000 0.3303 0.3397 0.0988 0.2991
Summary of SNP p-value for Hardy-Weinberg equilibrium (HWE)
MIN MAX MEAN MEDIAN SD CV
2.270073e-47 0.9997 0.2863 0.1752 0.2967 1.0362

SNP QC Complete!

4.3 SNP Density

Required Dataset (one of the following):

- **Site Info. (RDS)** of the current `data.frame`, downloadable from Data Input or Data QC pages.
- **Chromosome Info. (CSV)**: Reference genome information of the current study.

  ***Click here**: Download an example of **Chromosome Info.(CSV)**.*

  Example: The **Chromosome Info.** of rice (Data source: https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_034140825.1/)

Chr	Start	End
Chr01	0	43929697
Chr02	0	36447916
Chr03	0	37399924
Chr04	0	36078568
Chr05	0	30400764
Chr06	0	32122276
Chr07	0	29936421
Chr08	0	28605474
Chr09	0	27474823
Chr10	0	23931887
Chr11	0	31111469
Chr12	0	28271460

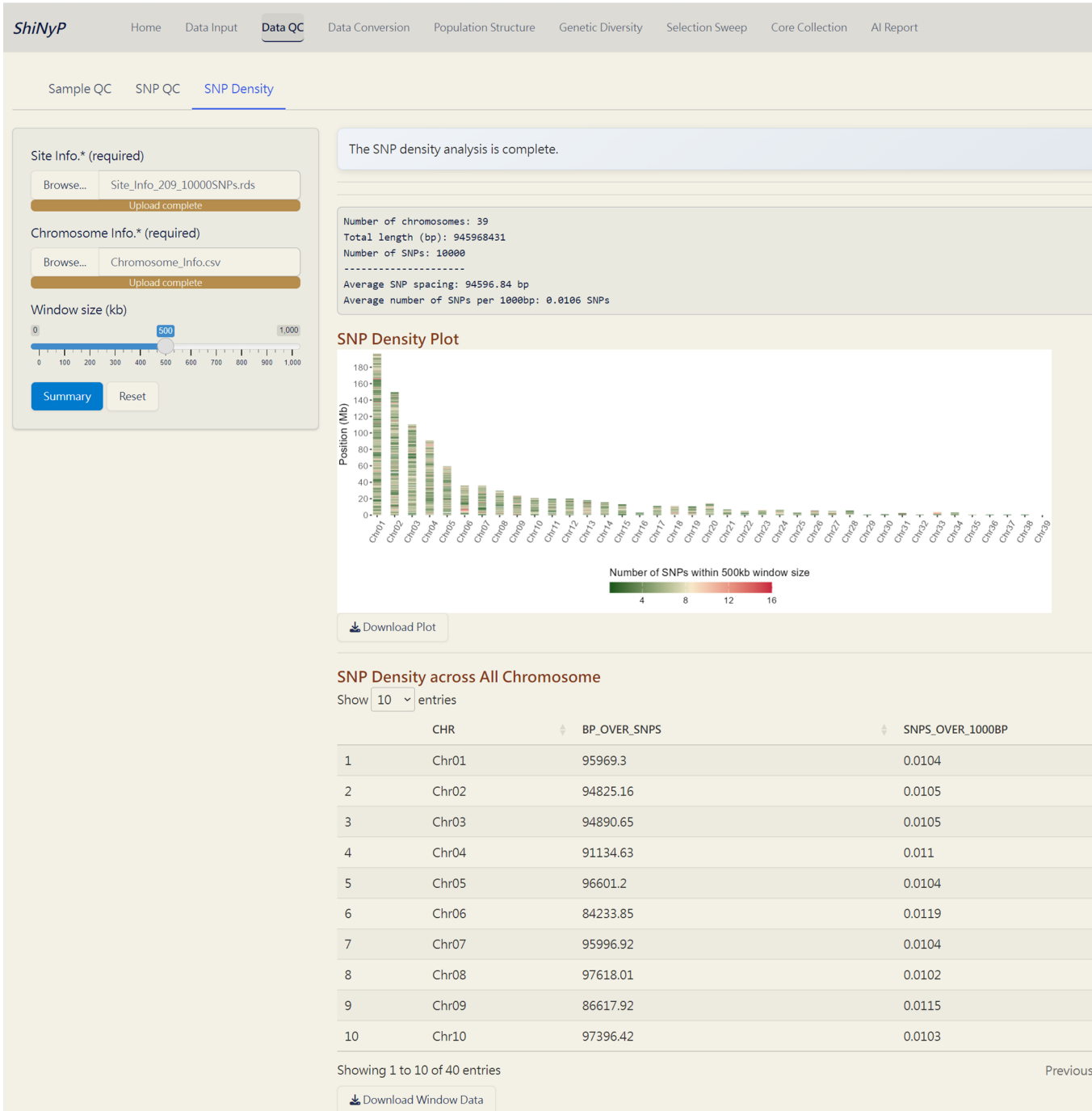
Steps:

1. Upload **Site Info. (RDS)** and **Chromosome Info. (CSV)**.
2. Choose a window size in kilobases (kb).
3. Click the **Summary** button. This will calculate the density of SNPs across the genome.

Outputs:

- **SNP Density Plot (PDF):** An ideogram visualizing SNP density across the genome within a defined window size. A gradient color palette is used to represent varying SNP densities: *green* for lower densities, *yellow* for medium densities, and *red* for higher densities, with *grey* indicating regions with zero SNP.

- **SNP Density (CSV):** A table detailing SNP density across each chromosome. *bp_over_SNPs*: The total base pairs (bp) per SNP in each window, representing the average spacing between SNPs. *SNPs_over_1000bp*: The number of SNPs per 1,000 base pairs, providing a normalized measure of SNP density across the genome.



SNP Density Complete!

Chapter 5

Data Transform

This section allow you to convert your SNP data in `data.frame` into `genind` and `genlight`.

Required Dataset (one of the following):

- Input VCF Data (`data.frame` file) from the Data Input page.
- Post-QC Data (`data.frame` file) from the Data QC page.

Step 1: Transform `data.frame` to `genind` Click the **Transform to genind** button. This will generate the `genind` file.

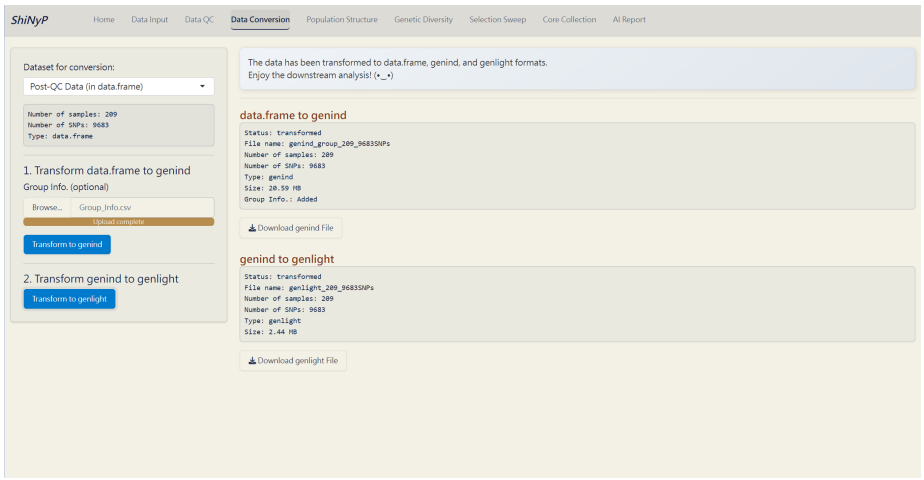
> **Note:** After obtaining the clustering results from the Population Structure/DAPC subpage, you can add **Group Info.** to the `genind` file by inputting the '**DAPC_Group_Info.csv**'. This step is necessary for analyses like 'Genetic Distance' and 'AMOVA'.

Step 2: Transform `genind` to `genlight` Click the **Transform to genlight** button. This will generate the `genlight` file.

Outputs:

- **genind (RDS)**: `genind` file. It's necessary for downstream analyses, *please download and save it!*
- **genlight (RDS)**: `genlight` file. It's necessary for downstream analyses, *please download and save it!*

Note: Please download and save your **`data.frame`**, **`genind`**, and **`genlight`** files after transformation. This will save you from having to input the large VCF file again next time.

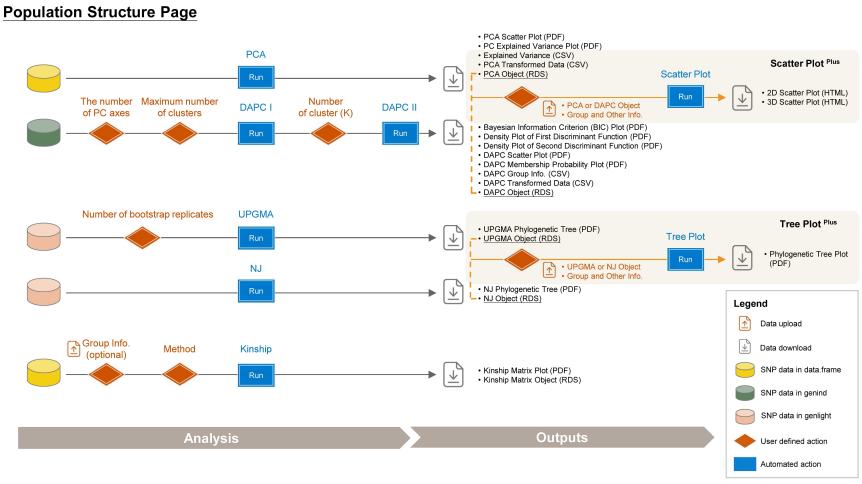


Data Transformation Complete!

Chapter 6

Population Structure

This section contains seven subpages: **PCA**, **DAPC**, **UPGMA Tree**, **NJ Tree**, **Kinship**, **Scatter Plot**$^{\textbf{Plus}}$, and **Tree Plot**$^{\textbf{Plus}}$ allowing you to conduct various population structure analyses and customize your plot.



6.1 PCA (Principal Component Analysis)

A widely used method to uncover underlying population structure by reducing the dimensionality of genetic data.

Required Dataset:

- `data.frame`

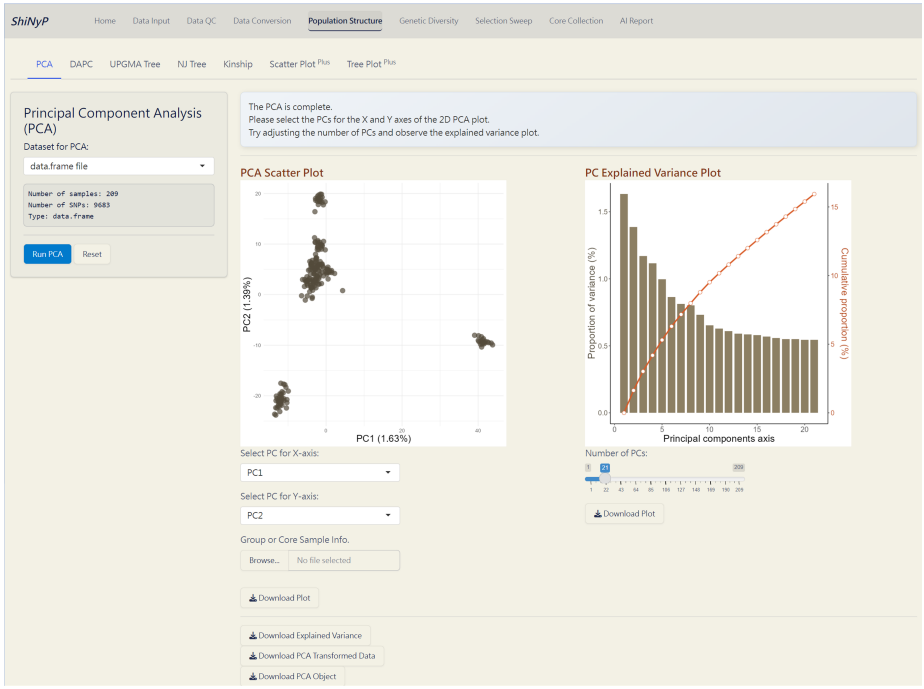
One Step:

1. Click the **Run PCA** button to generate PCA plots and the following downloadable files.

   Note: You can upload the **Group Info.** (from Population Structure/DAPC) or **Core Sample Info.** (from Core Collection/Core Sample Set) to classify individuals and color them in the PCA Scatter Plot.

Outputs:

- **PCA Scatter Plot (PDF)**: A scatter plot showing the distribution of samples based on principal components, with each dot representing an individual.
- **PC Explained Variance Plot (PDF)**: Visualizes the variance explained by each principal component.
- **Explained Variance (CSV)**: Contains the explained variance of each principal component.
- **PCA Transformed Data (CSV)**: Dataset transformed into principal components, with samples as rows and principal components as columns.
- **PCA Object (RDS)**: Contains all PCA results for future use and reproducibility, and can be used as input data in the Population Structure/Scatter Plot$^{\text{Plus}}$ subpage.



PCA Complete!

6.2 DAPC (Discriminant Analysis of Principal Components)

A multivariate method for identifying and visualizing genetic clusters by combining PCA and Linear Discriminant Analysis (LDA) [Jombart et al., 2010]. For more information, visit https://adegenet.r-forge.r-project.org/files/tutorial-dapc.pdf.

Required Dataset:

- `genind`

Step 1: Cluster Identification

1. Click the **Run DAPC I** button to determine the optimal number of clusters (the lowest BIC value indicates the optimal number of clusters).

Note: The default number of PC axes for cluster identification is set to retain PCs that capture up to 80% of the total variance. You can refer the "PC Explained Variance Plot" in the Population Structure/PCA subpage.

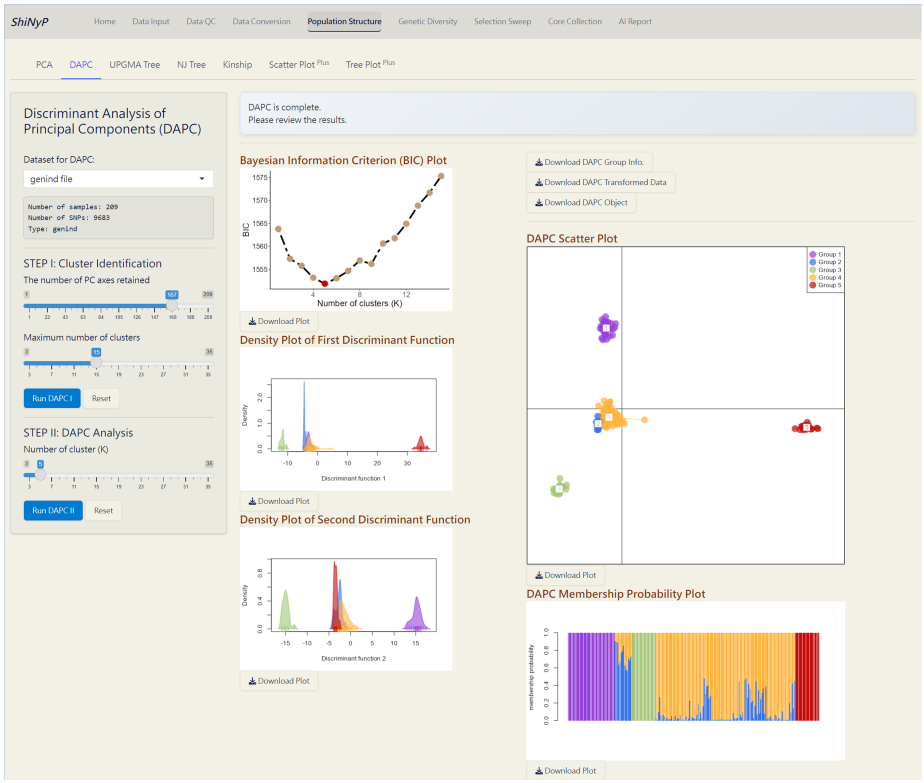
Step 2: DAPC Analysis

1. Choose the number of cluster (K) based on the "Bayesian Information Criterion (BIC) Plot".
2. Click the **Run DAPC II** button to generate DAPC plots and the following downloadable files.

Note: You can download the "DAPC Object" and upload it on Population Structure/Scatter Plot^{Plus} subpage to customize your 2D and 3D scatter plots.

Outputs:

- **Bayesian Information Criterion (BIC) Plot (PDF)**: Visual representation of the BIC for model selection.
- **Density Plot of First & Second Discriminant Function (PDF)**: Displays the density of the first and second discriminant functions, with each row bar representing an individual.
- **DAPC Scatter Plot (PDF)**: A scatter plot showing the distribution of samples based on discriminant functions (x-axis: first discriminant function; y-axis: second discriminant function), with each dot representing an individual.
- **DAPC Membership Probability Plot (PDF)**: Visualizes membership probabilities of individuals in different groups, with each row bar representing an individual.
- **DAPC Group Info. (CSV)**: Contains the group assignments for each individual based on DAPC. This file used in various subpages.
- **DAPC Transformed Data (CSV)**: Dataset transformed into discriminant functions with samples as rows and discriminant functions as columns.
- **DAPC Object (RDS)**: Contains all results from the DAPC analysis for future reproducibility. It can be used as input data in the Population Structure/Scatter Plot^{Plus} and Core Collection/Core SNP Set subpages.



DAPC Complete!

6.3 UPGMA (Unweighted Pair Group Method with Arithmetic mean) Tree

A classic approaach for constructing rooted trees based on genetic distance data. UPGMA tree is generated by *poppr* and *ggtree* packages [Yu et al., 2016, Kamvar et al., 2014].

Required Dataset:

- `genlight`

Steps:

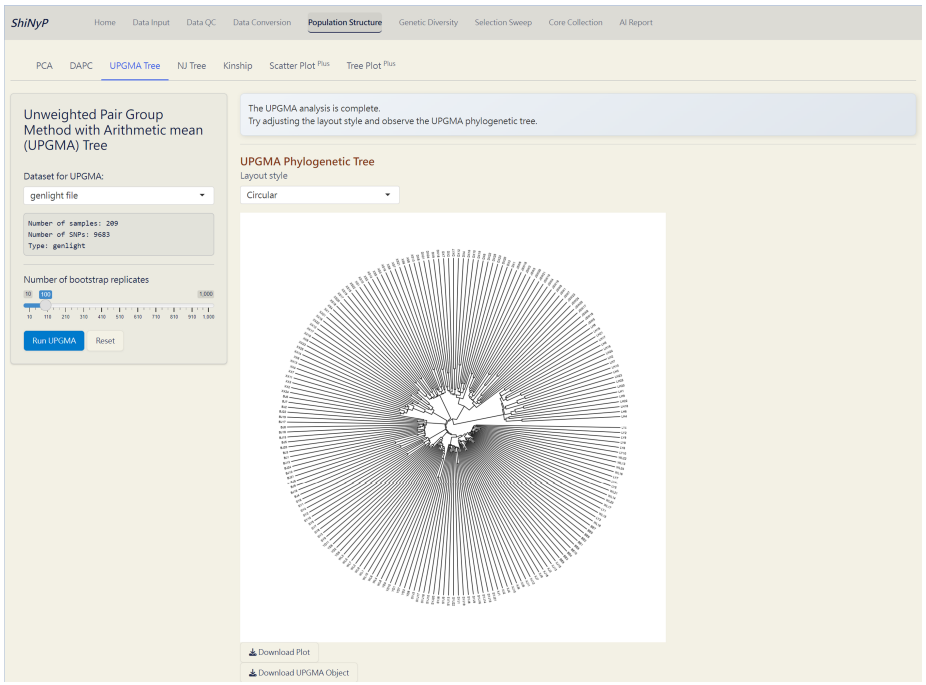
1. Choose the number of bootstrap replicates, which will be used for assessing the confidence of the branching structure.

2. Click the **Run UPGMA** button to generate tree plot.

   Note: You can download the "UPGMA Object" and upload it on Population Structure/Tree Plot^{Plus} subpage to customize your phylogenetic tree.

Outputs:

- **UPGMA Phylogenetic Tree (PDF)**: A UPGMA rooted tree with a user-defined layout style.
- **UPGMA Object (RDS)**: Contains all information of the UPGMA tree, and can be used as input data in the Population Structure/Tree Plot^{Plus} subpage.



UPGMA Tree Complete!

6.4 NJ (Neighbor-Joining) Tree

A method for building unrooted trees using genetic distance data. NJ tree is generated by *ape* and *ggtree* packages [Paradis and Schliep, 2018, Yu et al., 2016].

Required Dataset:

- `genlight`

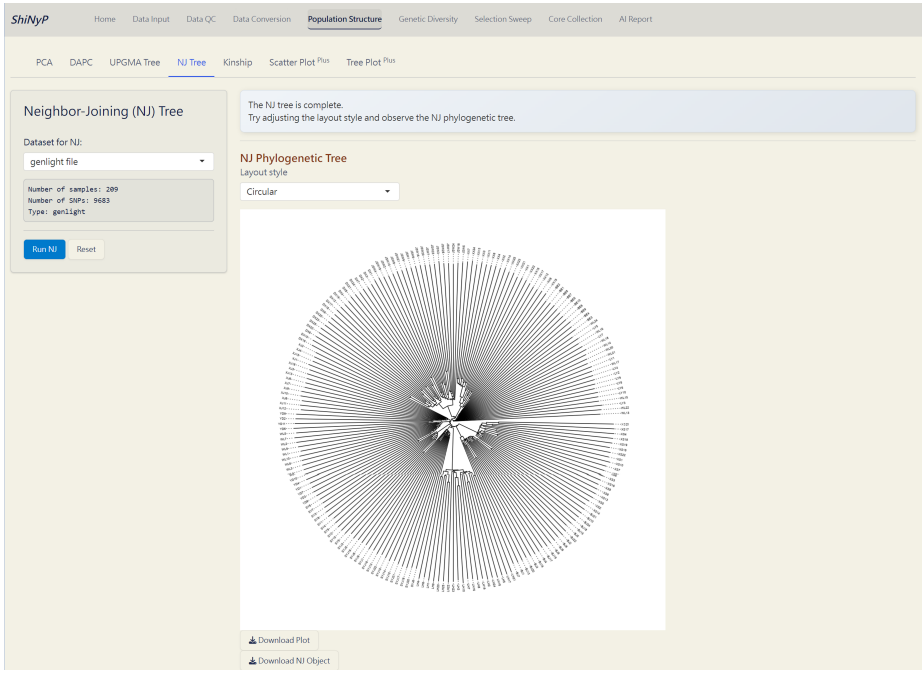
One Step:

1. Click the **Run NJ** button to generate tree plot.

   Note: You can download the "NJ Object" and upload it on Population Structure/Tree Plot^{Plus} subpage to customize your phylogenetic tree.

Outputs:

- **NJ Phylogenetic Tree (PDF)**: A NJ unrooted tree with a user-defined layout style.
- **NJ Object (RDS)**: Contains all information of the NJ tree, and can be used as input data in the Population Structure/Tree Plot^{Plus} subpage.



NJ Tree Complete!

6.5 Kinship Analysis

A statistical method for assessing genetic relationships and relatedness among individuals based on shared alleles [Kang et al., 2010]. Kinship matrix is generated by *statgenGWAS* package.For more information, visit https://rdrr.io/cran/statgenGWAS/man/kinship.html.

Required Dataset:

- `data.frame`

Steps:

1. Upload **Group Info.** from Population Structure/DAPC (optional). If uploaded, the order of samples will follow the group assignment; otherwise, it will follow the order of the original VCF data.

2. Choose a method to run kinship analysis.

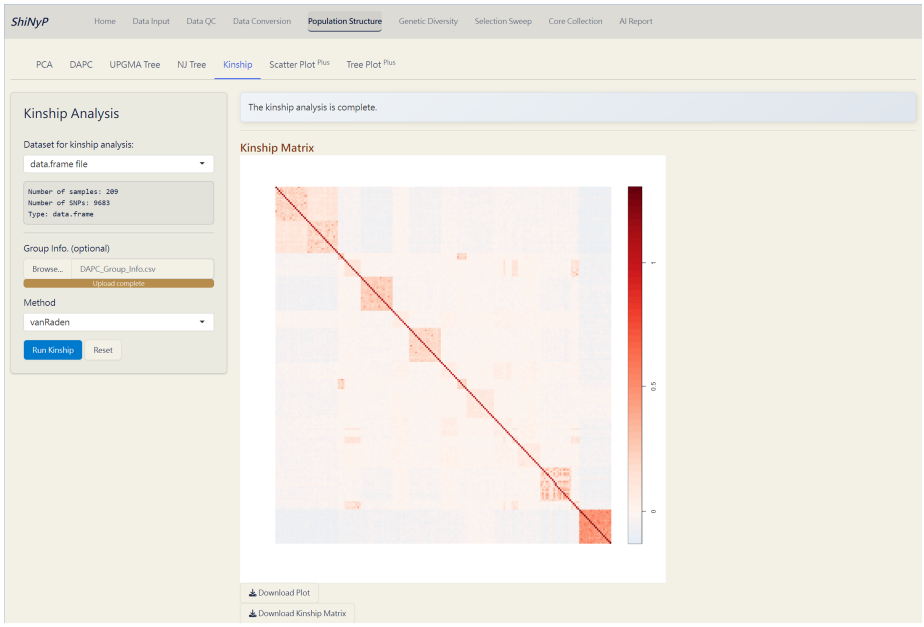
3. Click the **Run Kinship** button to generate the kinship matrix.

Outputs:

- **Kinship Matrix Plot (PDF)**: A visual representation of the kinship matrix.

- **Kinship Matrix (RDS)**: Contains the kinship matrix data.

  Note: This kinship matrix can be directly used as input for *GAPIT* package in genome-wide association studies (GWAS), helping to control for confounding effects.



Kinship Analysis Complete!

6.6 Scatter Plot Plus

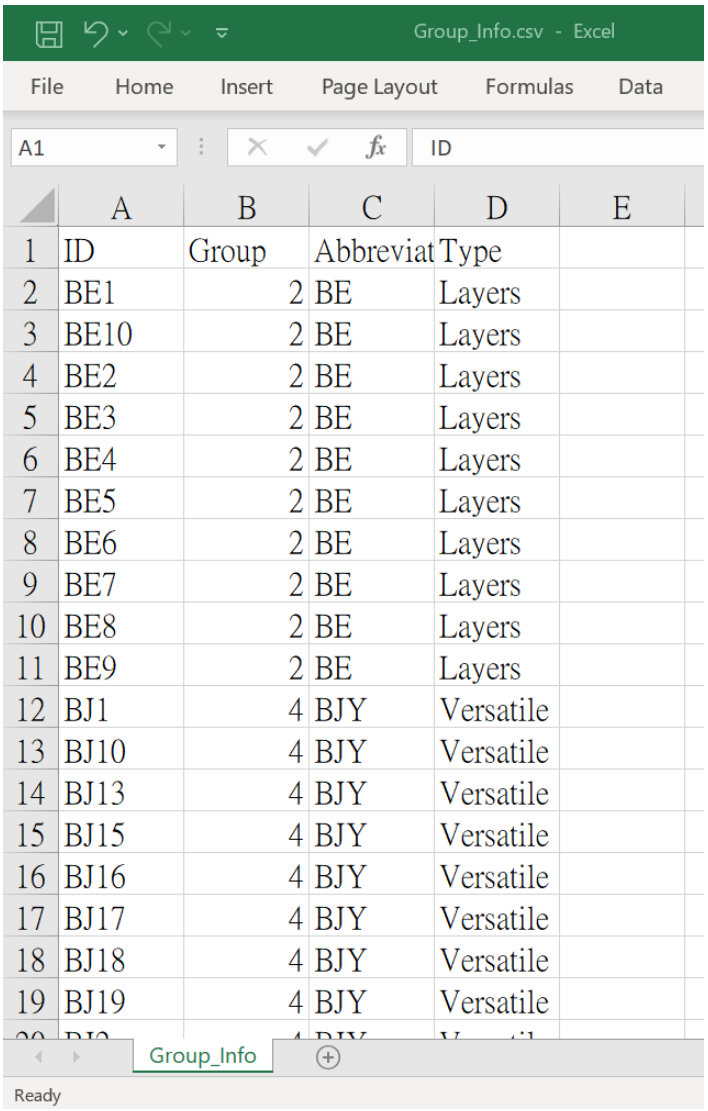
Customize your scatter plot based on the results from Population Structure/PCA or Population Structure/DAPC.

Required Files:

- **PCA Object** (PCA_prcomp_Object.rds file) or **DAPC Object** (DAPC_dapc_Object.rds file)
- **Group and Other Info.** (modifiable from DAPC_Group_Info.csv)

> **Note**: You can add more information about samples by adding new variables to the Group Info. file. Ensure that the sample order remains unchanged.

Example of Group Info. file (CSV).



Group_Info.csv - Excel

File Home Insert Page Layout Formulas Data

A1 ID

	A	B	C	D	E
1	ID	Group	Abbreviat	Type	
2	BE1	2	BE	Layers	
3	BE10	2	BE	Layers	
4	BE2	2	BE	Layers	
5	BE3	2	BE	Layers	
6	BE4	2	BE	Layers	
7	BE5	2	BE	Layers	
8	BE6	2	BE	Layers	
9	BE7	2	BE	Layers	
10	BE8	2	BE	Layers	
11	BE9	2	BE	Layers	
12	BJ1	4	BJY	Versatile	
13	BJ10	4	BJY	Versatile	
14	BJ13	4	BJY	Versatile	
15	BJ15	4	BJY	Versatile	
16	BJ16	4	BJY	Versatile	
17	BJ17	4	BJY	Versatile	
18	BJ18	4	BJY	Versatile	
19	BJ19	4	BJY	Versatile	

Group_Info

Ready

Steps:

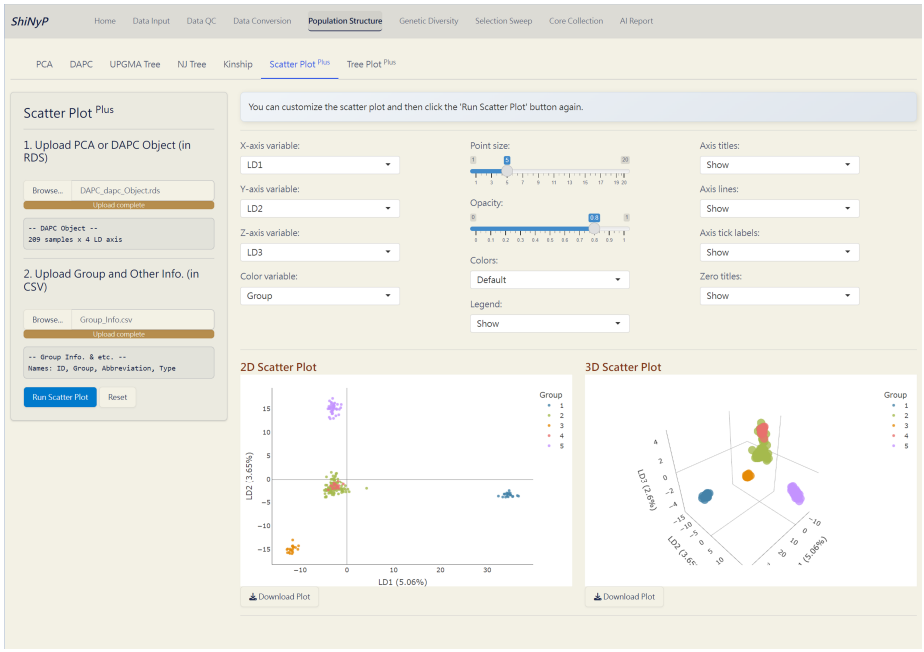
1. Upload **PCA or DAPC Object (RDS)**
2. Upload **Group and Other Info. (CSV)**
3. Click the **Run Scatter Plot** button to generate the 2D and 3D interactive scatter plots.

4. Customize the scatter plot and click the **Run Scatter Plot** button again.

   Note: The scatter plots are downloaded as HTML files and can be opened with browsers like Chrome or Edge.

Outputs:

- **2D Scatter Plot (HTML)**: Two-dimensional interactive scatter plot with user-defined attributes.
- **3D Scatter Plot (HTML)**: Three-dimensional interactive scatter plot with user-defined attributes.



Scatter Plot Plus Complete!

6.7 Tree Plot Plus

Customize your phylogenetic tree plot based on the results from Population Structure/UPGMA or Population Structure/NJ.

Required Files:

- **UPGMA Object** (UPGMA_phylo_Object.rds file) or **NJ Object** (NJ_phylo_Object.rds file)
- **Group and Other Info.** (modifiable from DAPC_Group_Info.csv)

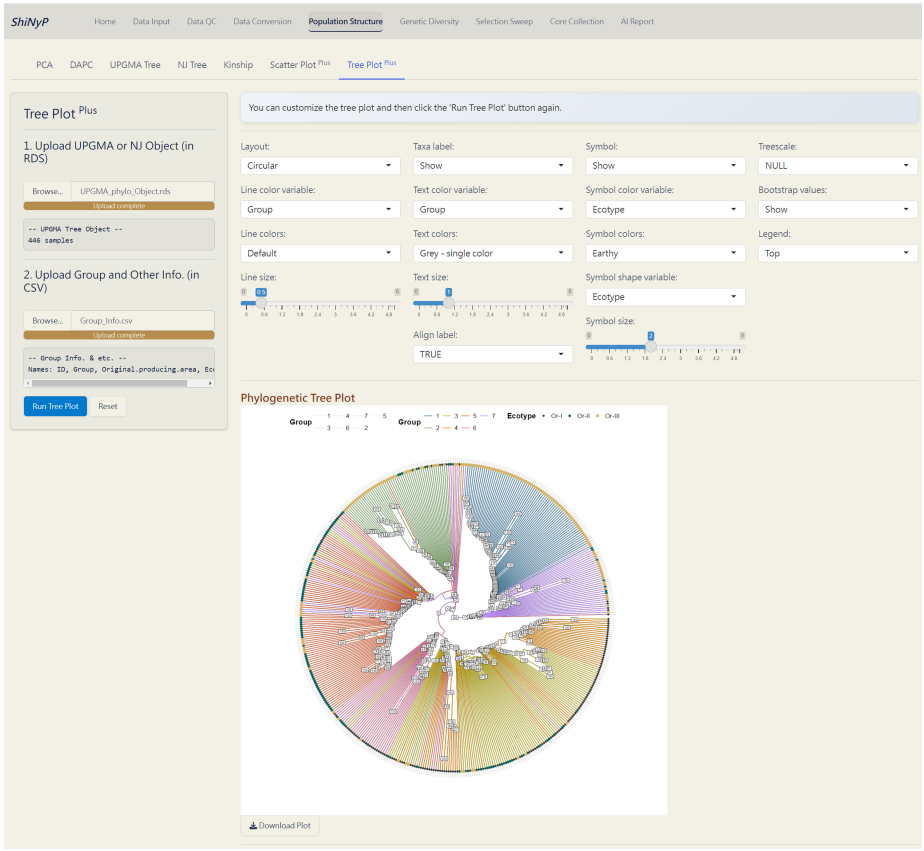
Note: You can add more information about samples by adding new variables to the Group Info. file. Ensure that the sample order remains unchanged.

Steps:

1. Upload **UPGMA or NJ Object (RDS)**
2. Upload **Group and Other Info. (CSV)**
3. Click the **Run Tree Plot** button to generate the tree plot.
4. Customize the tree plot and click the **Run Tree Plot** button again.

Outputs:

- **Phylogenetic Tree Plot (PDF)**: A phylogenetic tree plot with user-defined layout style and attributes.

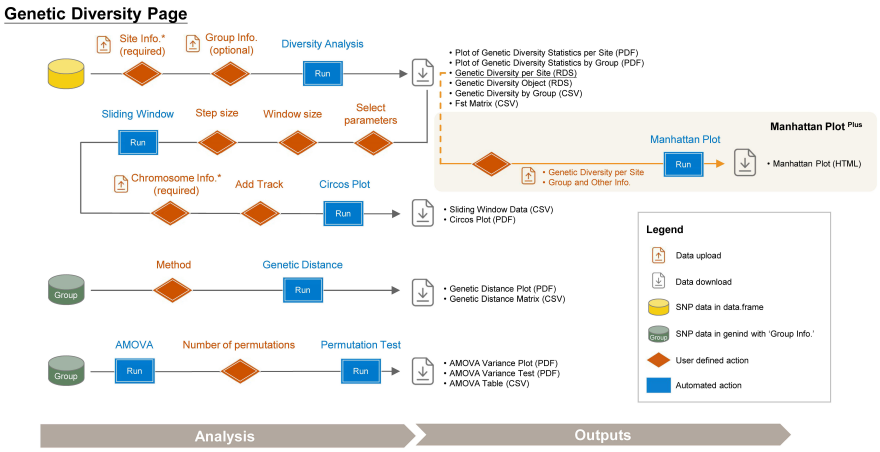


Tree Plot Plus Complete!

Chapter 7

Genetic Diversity

This section contains four subpages: **Diversity Parameter**, **Circos Plot**, **Genetic Distance**, and **AMOVA**, allowing you to conduct various population diversity and differentiation analyses.



7.1 Diversity Parameter

Calculate key diversity parameters for each SNP site. This approach is performed using the function from *snpReady* package [Granato et al., 2018].

Required Datasets:

- `data.frame`

- **Site Info. (RDS)** of the current `data.frame`, downloadable from Data Input or Data QC pages.

Steps:

1. Upload **Site Info.** (required).

2. Upload **Group Info.** from DAPC (optional). If uploaded, population-based parameters will be calculated.

3. Click the **Run Diversity Analysis** button to generate genetic diversity and the following downloadable files.

Outputs:

- **Plot of Genetic Diversity Statistics per Site (PDF)**: A genome-wide scatter plot visualizing the user-selected parameter.

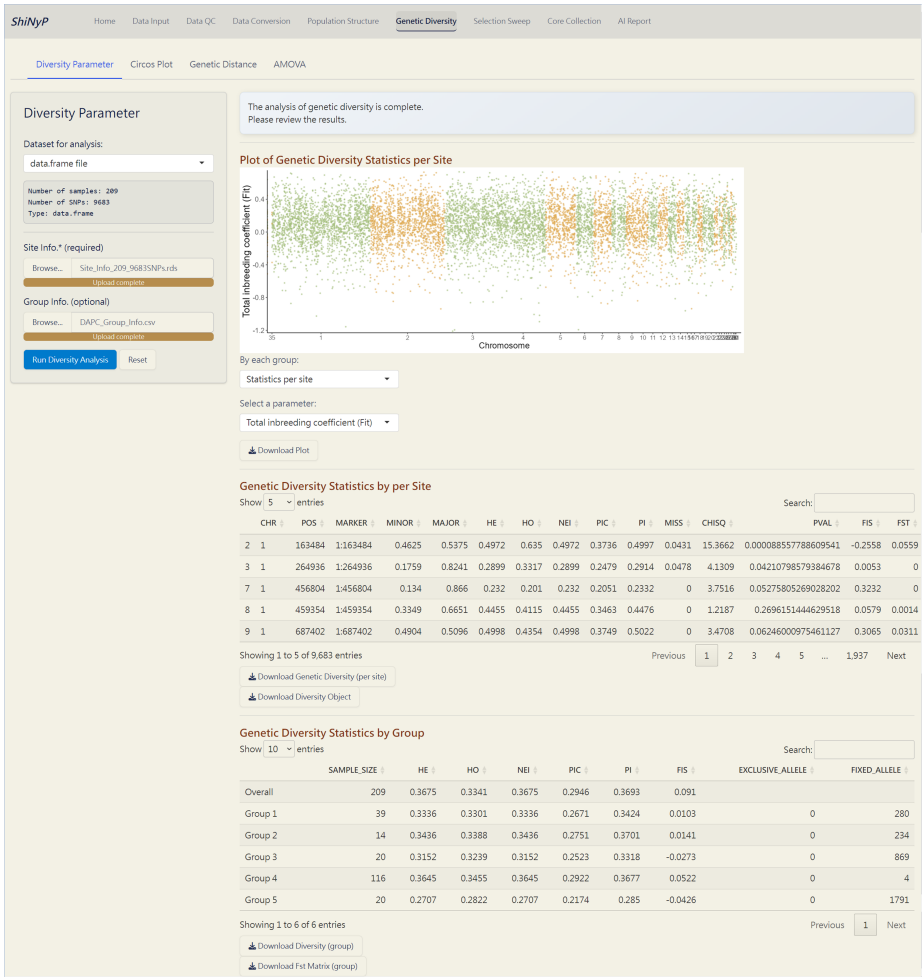
- **Plot of Genetic Diversity Statistics by Group (PDF)**: A lollipop plot visualizing the user-selected parameter.

- **Genetic Diversity per Site (RDS)**: Contains site information and diversity statistics, can be used as input data in the Selection Sweep/Manhattan Plot$^{\text{Plus}}$.

- **Genetic Diversity Object (RDS)**: Contains all genetic diversity results for future use and reproducibility.

- **Genetic Diversity by Group (CSV)**: A table showing genetic diversity based on defined group assignments.

- **Fst Matrix (CSV)**: A table showing pairwise Fst based on defined group assignments.



Diversity Analysis Complete!

7.2 Circos Plot

Genome-wide diversity is visualized using Circos plots generated with the *circlize* package [Gu et al., 2014] based on results of diversity parameters in a sliding window format.

Required Dataset:

- Auto-import the results from the Genetic Diversity/Diversity Parameter subpage.

- **Chromosome Info. (CSV)**: Reference genome information of the current study. For more details about this file, refer to **Section** 4.3 **(SNP Density)**.

Step 1: Sliding Window

1. Select parameters to generate sliding window data.

2. Choose window size (kb) and step size (kp).

3. Click the **Run Sliding Window** button to generate sliding window data for circos plot.

Step 2: Circos Plot

1. Upload **Chromosome Info. (CSV)**.

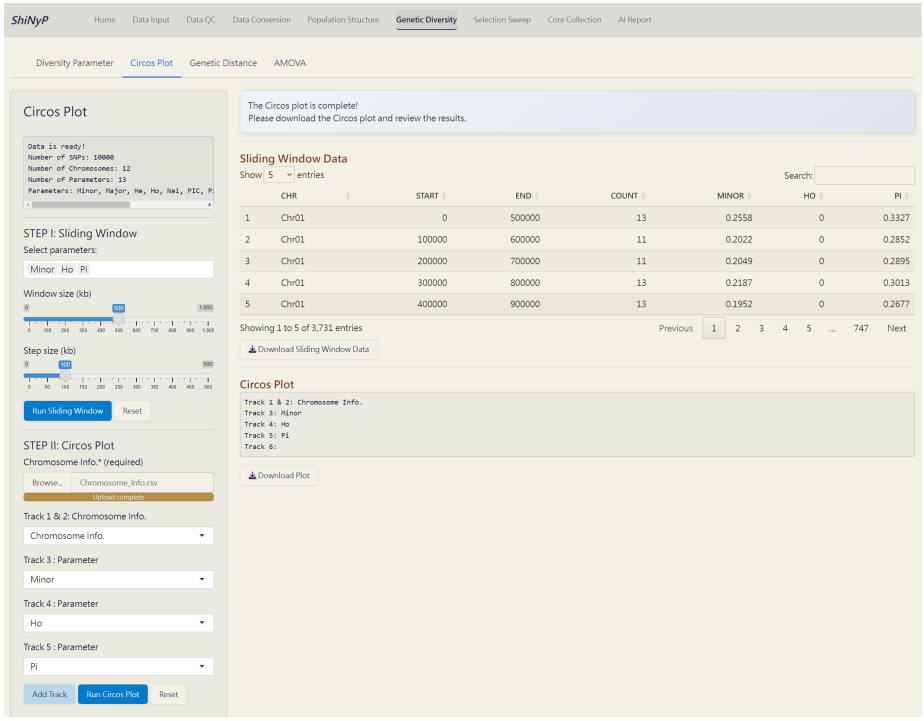
2. Select a parameter for each track, and add tracks if necessary (up to a maximum of 6).

3. Click the **Run Circos Plot** button to generate the circos plot.

Outputs:

- **Sliding Window Data (CSV)**: A sliding window dataset based on user-selected parameters.

- **Circos Plot (PDF)**: A circos plot visualizing the user-selected parameters, with the top 1% of each parameter colored in red.



Circos Plot Complete!

7.3 Genetic Distance

Pairwise genetic distance between populations is computed using *hierfstat* package. For more information, visit https://rdrr.io/cran/hierfstat/man/genet.dist.html.

Required Dataset:

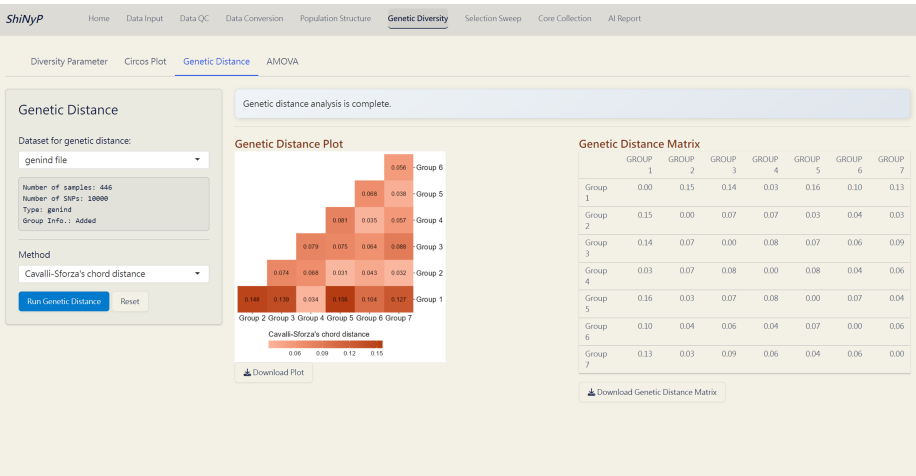
- **`genind`** with 'Group Info.', downloadable from Data Conversion page after you have both the **`data.frame`** and Group Info.

Steps:

1. Select a method.
2. Click the **Run Genetic Distance** button to generate the pairwise genetic distance.

Outputs:

- **Genetic Distance Plot (PDF)**: A plot of the pairwise genetic distance matrix based on the user-selected method.
- **Genetic Distance Matrix (CSV)**: A pairwise genetic distance matrix based on the user-selected method.



Genetic Distance Complete!

7.4 AMOVA (Analysis of MOlecular VAriance)

A method for assessing genetic variations and relationships within and between populations [Excoffier et al., 1992].This approach is performed using the function from *hierfstat* and *poppr* packages [Kamvar et al., 2014, GOUDET, 2004].

Required Dataset:

- **`genind`** with 'Group Info.', downloadable from Data Conversion page after you have both the **`data.frame`** and Group Info.

Step 1: Run AMOVA

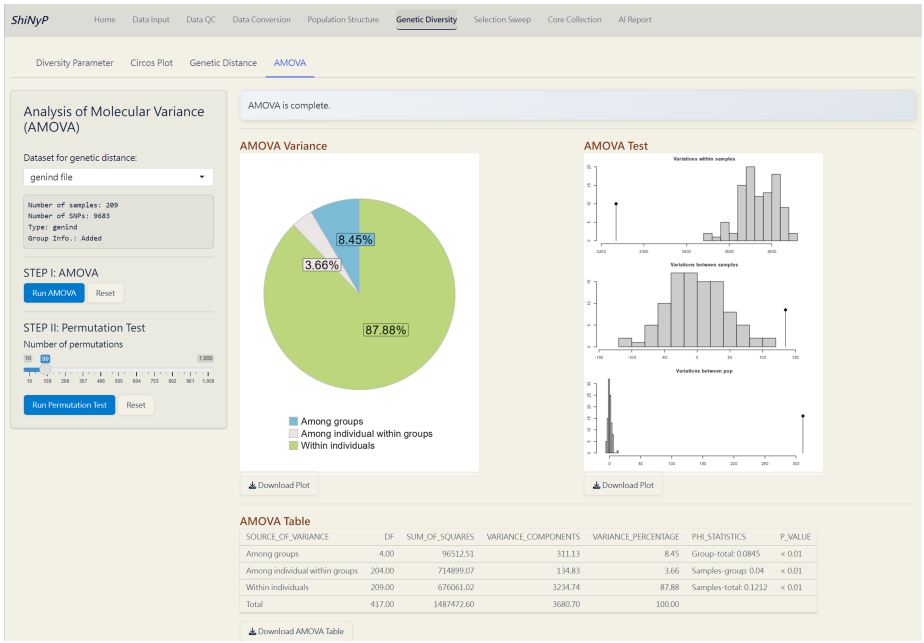
1. Click the **Run AMOVA** button to partition genetic variation among and within populations.

Step 2: Run Permutation Test

1. Choose the number of randomizations for the permutation test to detect the significance of three hierarchical levels. We recommend using 9, 99 (default), 199, 499, 799, or 999 permutations for more classical p-values.
2. Click the **Run Permutation Test** button to perform the statistical test.

Outputs:

- **AMOVA Variance Plot (PDF)**: A pie chart showing the explained genetic variance of population strata among defined groups.
- **AMOVA Variance Test (PDF)**: A plot showing the significance test of population strata among defined groups. The histograms depict randomized strata distributions, with the black line representing genetic variance components.
- **AMOVA Table (CSV)**: A table with detailed AMOVA results.

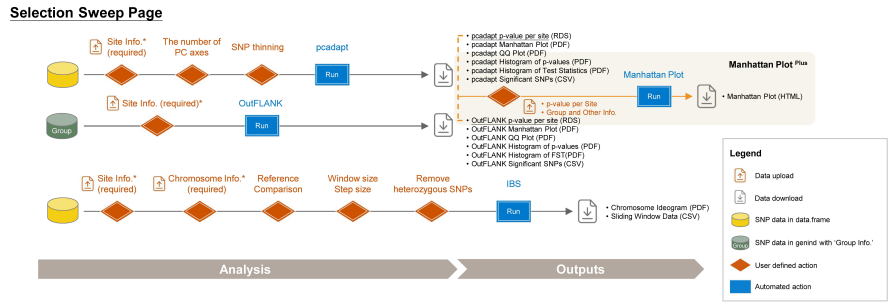


AMOVA Complete!

Chapter 8

Selection Sweep

This section contains four subpages: **pcadapt**, **OutFLANK**, **IBS**, and **Manhattan Plot**$^{\textbf{Plus}}$, allowing you to detect selection signatures in different scenario and customize your plot.



8.1 pcadapt

A PCA-based approach identifies selective outliers relative to population structure [Luu et al., 2016].

Required Datasets:

- `data.frame`
- **Site Info. (RDS)** of the current `data.frame`, downloadable from Data Input or Data QC pages.

Steps:

1. Upload **Site Info.** (required).

2. Click SNP Thinning button (optional) and choose window size (number of SNPs) and r^2 threshold. For more information, visit https://bcm-uga.github.io/pcadapt/articles/pcadapt.html.

3. Click the **Run pcadapt** button to perform genome scan for selection.

Outputs:

- **pcadapt p-value per site (RDS)**: A dataset containing p-values and adjusted p-values for each site.

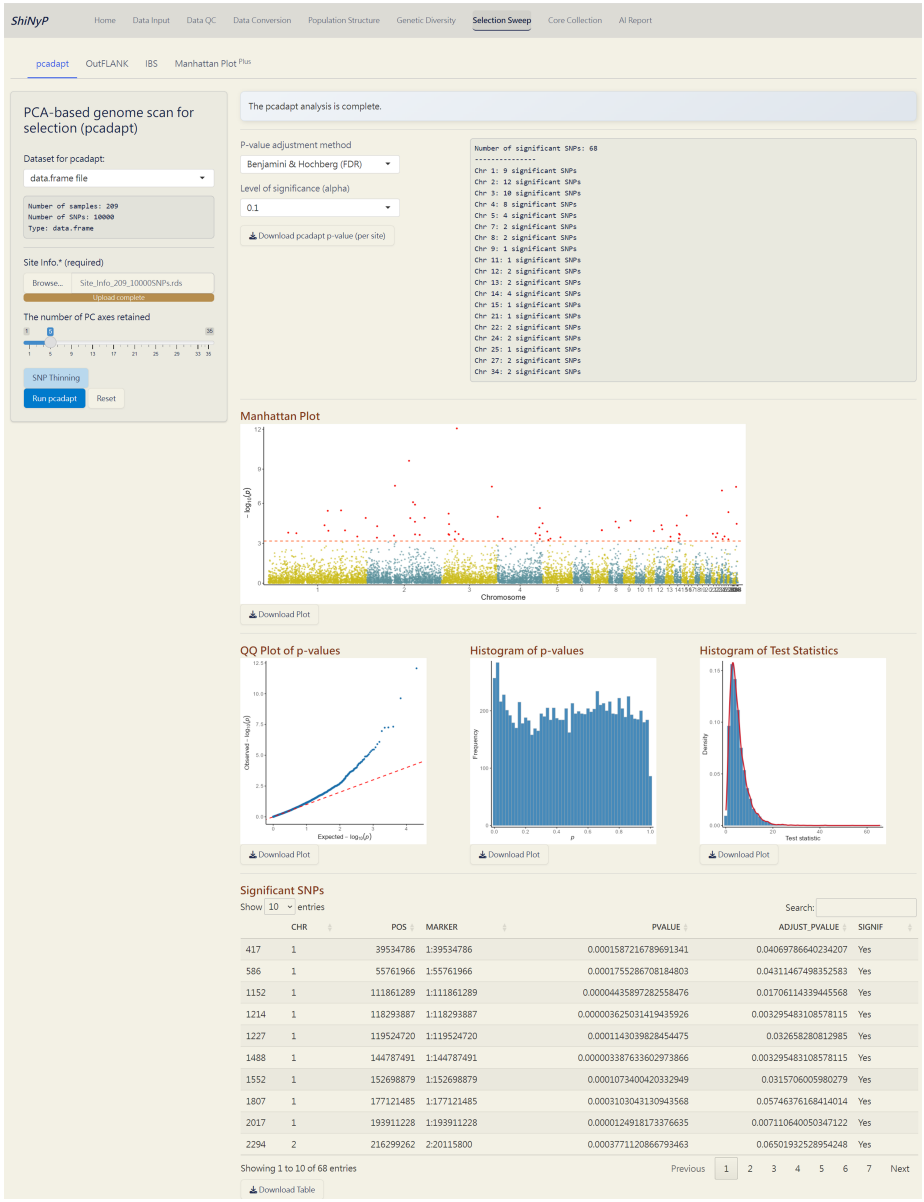
- **pcadapt Manhattan Plot (PDF)**: A Manhattan plot visualizing the p-values per site across the genome. Significant SNPs are highlighted in red.

- **pcadapt QQ Plot (PDF)**: A QQ plot comparing the distribution of observed p-values to the expected distribution under the null hypothesis.

- **pcadapt Histogram of p-values (PDF)**: A histogram showing the distribution of p-values across all sites.

- **pcadapt Histogram of Test Statistics (PDF)**: A histogram showing the distribution of test statistics across all sites.

- **pcadapt Significant SNPs (CSV)**: A table listing SNPs identified as significant by pcadapt, including their site info., p-values, and adjusted p-values.



The pcadapt Complete!

8.2 OutFLANK

A Fst-based approach detects selection signals by comparing genetic differentiation between defined group assignments [Whitlock and Lotterhos, 2015]. For

more information, visit https://rpubs.com/lotterhos/outflank.

Required Datasets:

- `genind` with 'Group Info.', downloadable from Data Conversion page after you have both the `data.frame` and Group Info.
- **Site Info. (RDS)** of the current `data.frame`, downloadable from Data Input or Data QC pages.

Steps:

1. Upload **Site Info.** (required).

2. Click the **Run OutFLANK** button to perform genome scan for selection.

Outputs:

- **OutFLANK p-value per site (RDS)**: A dataset containing p-values and adjusted p-values for each site.

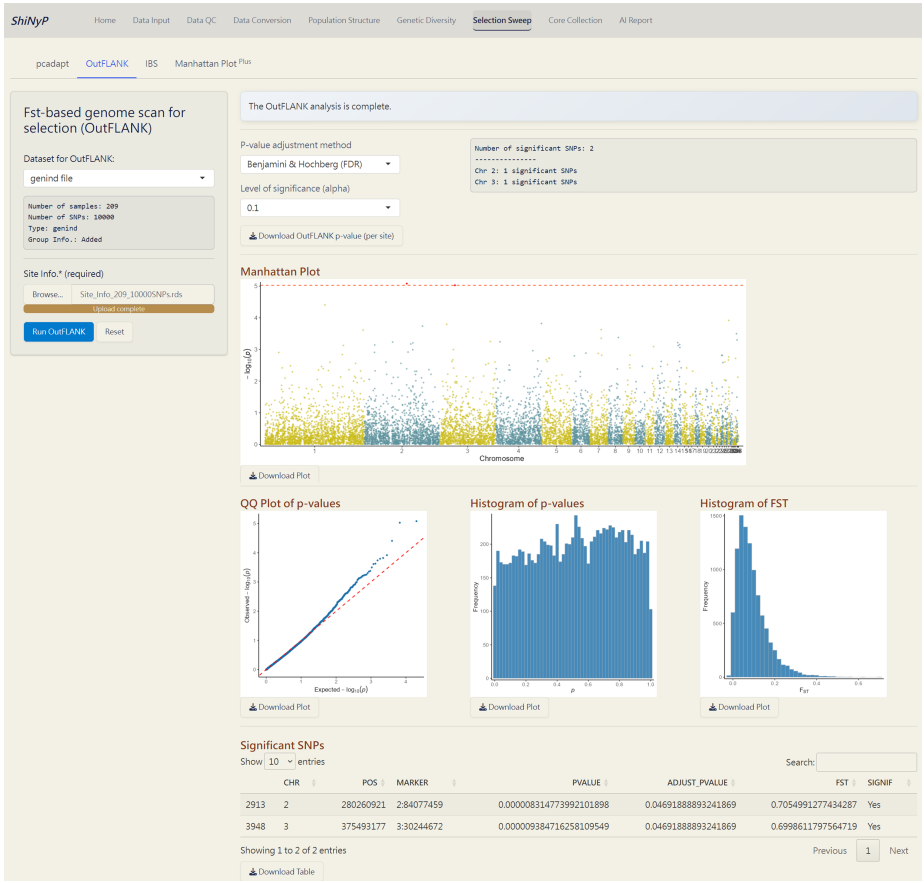
- **OutFLANK Manhattan Plot (PDF)**: A Manhattan plot visualizing the p-values per site across the genome. Significant SNPs are highlighted in red.

- **OutFLANK QQ Plot (PDF)**: A QQ plot comparing the distribution of observed p-values to the expected distribution under the null hypothesis.

- **OutFLANK Histogram of p-values (PDF)**: A histogram showing the distribution of p-values across all sites.

- **OutFLANK Histogram of Fst (PDF)**: A histogram showing the distribution of Fst values across all sites.

- **OutFLANK Significant SNPs (CSV)**: A table listing SNPs identified as significant by OutFLANK, including their site info., Fst values, and p-values.



The OutFLANK Complete!

8.3 IBS (Identity By State)

An approach to detect differences in genomic regions between pairs of individuals, useful for identifying pedigree relationships.

Required Datasets:

- `data.frame`
- **Site Info. (RDS)** of the current `data.frame`, downloadable from Data Input or Data QC pages.
- **Chromosome Info. (CSV)**: Reference genome information of the current study. For more details about this file, refer to **Section** 4.3 **(SNP Density)**.

Steps:

1. Upload **Site Info.** (required).

2. Upload **Chromosome Info. (CSV)** (required).

3. Choose the reference and comparison samples.

4. Select window size (kb) and step size (kp).

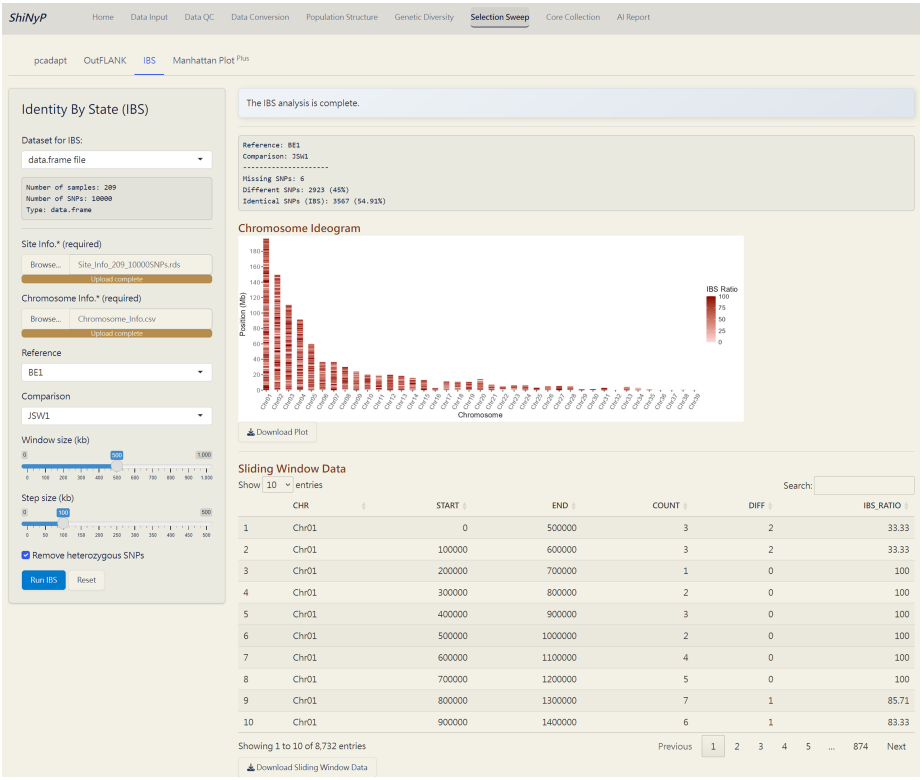
5. To remove heterozygous SNPs from the reference sample, click the Remove heterozygous SNPs checkbox (optional).

6. Click the **Run IBS** button to perform IBS analysis.

Outputs:

- **Chromosome Ideogram (PDF)**: An ideogram visualizing the IBS results, using a gradient palette to represent the differences across chromosomes.

- **Sliding Window Data (CSV)**: A sliding window dataset with IBS results, including SNP count, different SNPs, and the ratio of different SNPs per window.



The IBS Complete!

8.4 Manhattan Plot Plus

Customize your phylogenetic tree plot based on the results from Genetic Diversity/Diversity Parameter, Selection Sweep/pcadapt, or Selection Sweep/OutFLANK.

Required Files:

- **Genetic Diversity per Site** (Genetic_Diversity_per_Site.rds), **pcadapt p-value per Site** (pcadapt_p-value_per_site.rds), or **OutFLANK p-value per Site** (OutFLANK_p-value_per_site.rds).
- **Chromosome Info. (CSV)**: Reference genome information of the current study. For more details about this file, refer to **Section** 4.3 **(SNP Density)**.

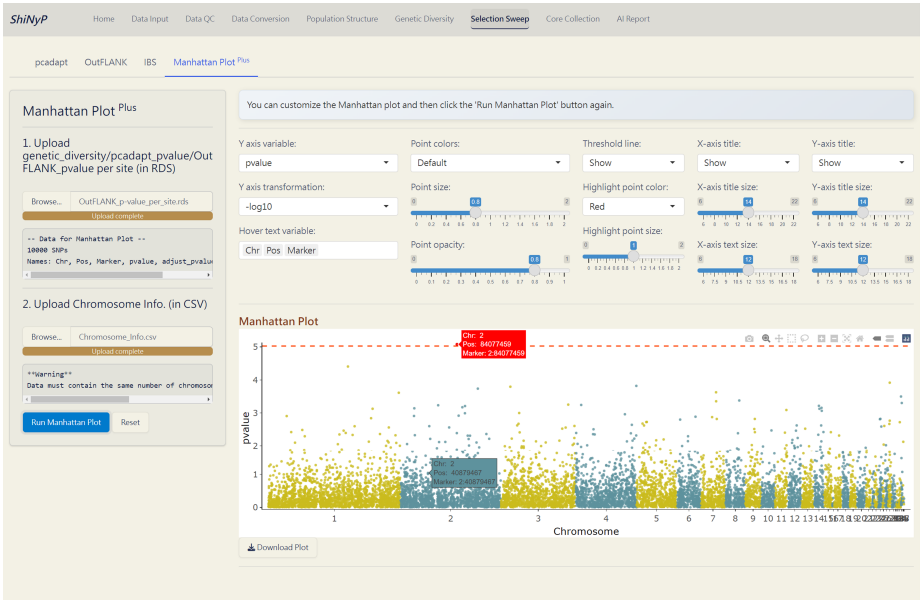
Steps:

1. Upload **genetic_diversity/pcadapt_pvalue/OutFLANK_pvalue per site (RDS)**.
2. Upload **Chromosome Info. (CSV)**.
3. Click the **Run Manhattan Plot** button to generate the Manhattan plot.
4. Customize the Manhattan plot and click the **Run Manhattan Plot** button again.

Outputs:

- **Manhattan Plot (PDF)**: A Manhattan plot with user-defined layout style and attributes.

  Note: If generating a plot for p-values, make sure to use '-log10' transformation for the Y axis.

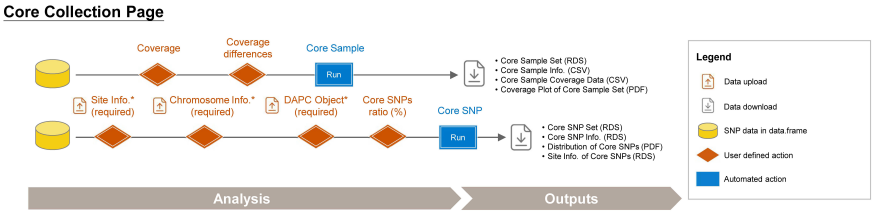


Manhattan Plot ^{Plus} Complete!

Chapter 9

Core Collection

This section contains two subpages: **Core Sample Set**, and **Core SNP Set**, allowing you to capture the key samples and SNPs.



9.1 Core Sample Set

Establish a core collection that represents the genetic variation of the entire population. This approach is modified function from GenoCore [Jeong et al., 2017].

Required Datasets:

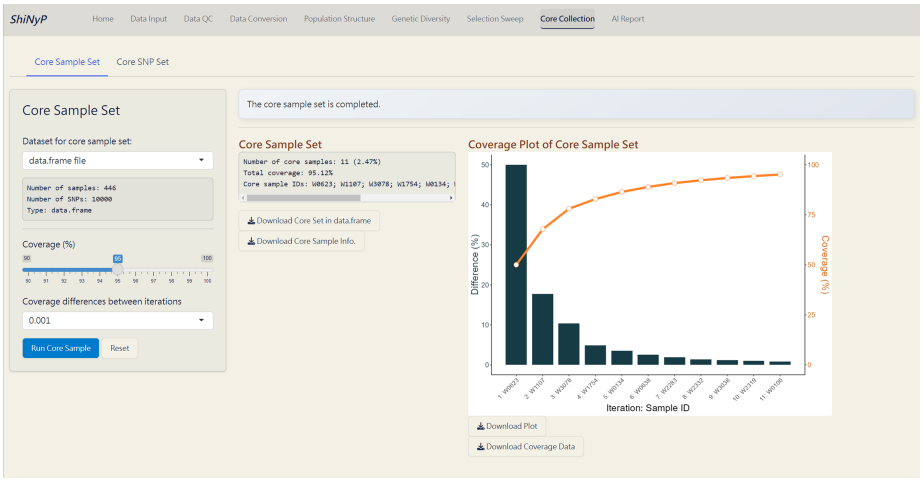
- `data.frame`

Steps:

1. Choose the minimum genetic coverage (%).
2. Choose the genetic coverage differences between iterations.
3. Click the **Run Core Sample** button to perform core collection.

Outputs:

- **Core Sample Coverage Data (CSV)**: A table listing the coverage (%) of each iteration and coverage differences between iterations.
- **Core Sample Set (RDS)**: A `data.frame` of core samples and their genotypic information.
- **Core Sample Info. (CSV)**: A table listing whether each sample is included in the core collection or not, and can be used as input data in the Population Structure/PCA subpage.
- **Coverage Plot of Core Sample Set (PDF)**: Visualizes the sample coverage by each iteration.



The Core Sample Set Complete!

9.2 Core SNP Set

Establish a core SNP collection that represents the genetic variation observed in the full dataset.

Required Datasets:

- `data.frame`
- **Site Info. (RDS)** of the current `data.frame`, downloadable from Data Input or Data QC pages

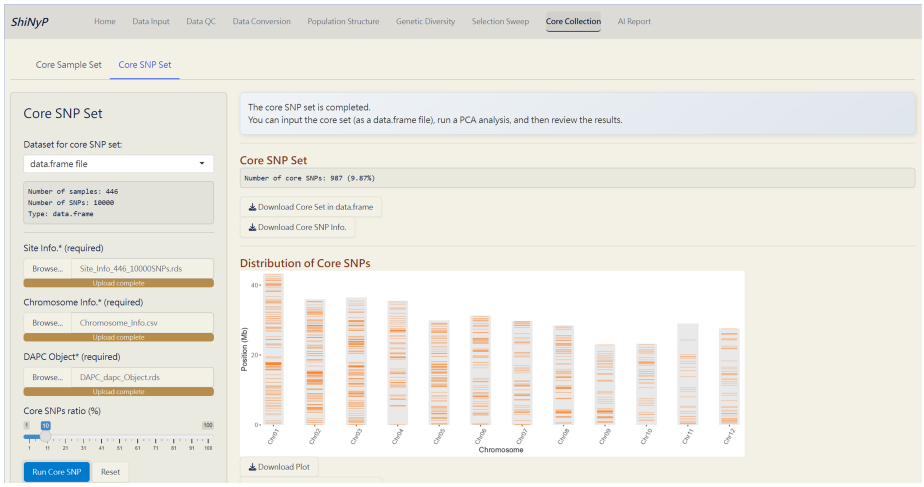
- **Chromosome Info. (CSV)**: Reference genome information of the current study. For more details about this file, refer to **Section** 4.3 **(SNP Density)**.
- **DAPC Object** (DAPC_dapc_Object.rds), downloadable from Population Structure/DAPC.

Steps:

1. Upload required datasts: **Site Info. (RDS)**, **Chromosome Info. (CSV)**, and **DAPC Object (RDS)**.
2. Choose the maximum core SNPs ratio (%).
3. Click the **Run Core SNP** button to perform core collection.

Outputs:

- **Core SNP Set (RDS)**: A `data.frame` of core SNPs and their genotypic information.
- **Core SNP Info. (RDS)**: A table listing whether each SNP is included in the core collection or not.
- **Distribution of Core SNPs (PDF)**: An ideogram labeling the core SNPs.
- **Site Info. of Core SNPs (RDS)**: Core SNPs site information file.



The Core SNP Set Complete!

Chapter 10

AI Report

This page allows you to generate your preliminary results from prior analysis, input your OpenAI API key, select an AI model, and get an AI-driven report. Powered by the *openai* package (https://github.com/irudnyts/openai).

Step 1: Preliminary Results

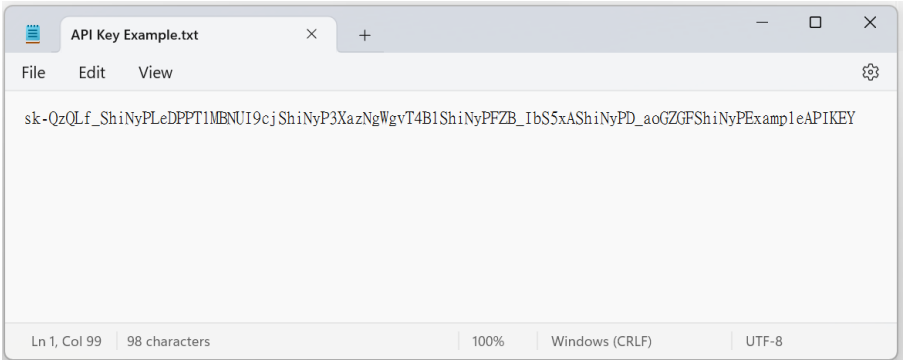
1. Enter the species name for the current study.
2. Click the **Auto-generate** button to obtain **Preliminary Results** from the ***ShiNyP*** workflow.

   Note: You can download the Preliminary Results as a `.txt` file, edit it as needed, and upload it again for 'AI-driven Report' use.

Step 2: AI-driven Report

1. Select an AI model. We recommend using **GPT-4o mini**, which offers the most cost-efficient performance. For more information, visit: https://platform.openai.com/docs/models.
2. Specify the task for your Preliminary Results to OpenAI.
3. Upload the `.txt` file containing your OpenAI API key (e.g., "sk-……").

   Example of API key file (TXT).



4. Click the **Get Report** button to obtain **AI-driven Report**.

How to get the OpenAI API key:

1. **Sign Up or Log In** to the OpenAI website: https://platform.openai.com/docs/overview.
2. **Check Your Usage** to track (free) credits and current consumption: https://platform.openai.com/usage.
3. If your (free) credits are **insufficient**, you can manage billing and payments by visiting: https://platform.openai.com/settings/organization/billing/overview.
4. **Generate a New API Key** by going to: https://platform.openai.com/api-keys.
5. **Copy and Paste** the generated key into a Notepad file and save it as a `.txt` file for ***ShiNyP*** use.

Upon signing up, OpenAI provides free credits valid for 3 months. After the free trial credits expire or are exhausted, you'll be billed based on your usage. Costs depend on the model and the number of tokens processed.

{Figure}

AI Report Complete!

Chapter 11

INDEX

Bibliography

Bibliography

L Excoffier, P E Smouse, and J M Quattro. Analysis of molecular variance inferred from metric distances among dna haplotypes: application to human mitochondrial dna restriction data. *Genetics*, 131(2):479–491, 06 1992. doi: 10.1093/genetics/131.2.479. URL http://dx.doi.org/10.1093/genetics/131.2.479.

JÉRÔME GOUDET. hierfstat, a package for r to compute and test hierarchical *f*-statistics. *Molecular Ecology Notes*, 5(1):184–186, 12 2004. doi: 10.1111/j.1471-8286.2004.00828.x. URL http://dx.doi.org/10.1111/j.1471-8286.2004.00828.x.

Italo S. C. Granato, Giovanni Galli, Evellyn Giselly de Oliveira Couto, Massaine Bandeira e Souza, Leandro Freitas Mendonça, and Roberto Fritsche-Neto. snpready: a tool to assist breeders in genomic analysis. *Molecular Breeding*, 38(8), 07 2018. doi: 10.1007/s11032-018-0844-8. URL http://dx.doi.org/10.1007/s11032-018-0844-8.

Zuguang Gu, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. *circlize*implements and enhances circular visualization in r. *Bioinformatics*, 30(19):2811–2812, 06 2014. doi: 10.1093/bioinformatics/btu393. URL http://dx.doi.org/10.1093/bioinformatics/btu393.

Seongmun Jeong, Jae-Yoon Kim, Soon-Chun Jeong, Sung-Taeg Kang, Jung-Kyung Moon, and Namshin Kim. Genocore: A simple and fast algorithm for core subset selection from large genotype datasets. *PLOS ONE*, 12(7):e0181420, 07 2017. doi: 10.1371/journal.pone.0181420. URL http://dx.doi.org/10.1371/journal.pone.0181420.

Thibaut Jombart, Sébastien Devillard, and François Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1):94, 2010. doi: 10.1186/1471-2156-11-94. URL http://dx.doi.org/10.1186/1471-2156-11-94.

Zhian N. Kamvar, Javier F. Tabima, and Niklaus J. Grünwald. *Poppr*: an r package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, 2:e281, 03 2014. doi: 10.7717/peerj.281. URL http://dx.doi.org/10.7717/peerj.281.

Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348–354, 03 2010. doi: 10.1038/ng.548. URL http://dx.doi.org/10.1038/ng.548.

Keurcien Luu, Eric Bazin, and Michael G. B. Blum. *pcadapt*: an r package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, 17(1):67–77, 09 2016. doi: 10.1111/1755-0998.12592. URL http://dx.doi.org/10.1111/1755-0998.12592.

Emmanuel Paradis and Klaus Schliep. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in r. *Bioinformatics*, 35(3):526–528, 07 2018. doi: 10.1093/bioinformatics/bty633. URL http://dx.doi.org/10.1093/bioinformatics/bty633.

Michael C. Whitlock and Katie E. Lotterhos. Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution off_{st}. *The American Naturalist*, 186(S1):S24–S36, 10 2015. doi: 10.1086/682949. URL http://dx.doi.org/10.1086/682949.

Guangchuang Yu, David K. Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36, 09 2016. doi: 10.1111/2041-210x.12628. URL http://dx.doi.org/10.1111/2041-210X.12628.