GONE Practical

Marianne Dehasque

June 18, 2025

Introduction

GONE

For this practical, we will use `GONE`, a software that infers recent effective population size from genetic data based on linkage disequilibrium.

The original `GONE` paper can be found here. However, today we will use `GONE2`, which hasn't been officially published yet, but is faster and more user friendly. Whereas the original GONE software depends on multiple scripts and a separate INPUT_PARAMETERS_FILE, GONE2 can be run directly from the command line with different parameters.

The `GONE2` github repository with installation and usage instructions can be found here. Note that for the practical today all necessary softwares and tools have already been installed.

Dataset

We will use the killer whale dataset from Kardos et al, 2023 (*Nature Ecology and Evolution*). In this paper, Kardos and colleagues study the population dynamics of an endangered North Pacific killer whale population.

For this practical, the main author kindly provided a pre-filtered dataset consisting of genotype information from 151 individuals from five different killer whale populations.

Data types

In this practical, we will use two types of data: PLINK PED/MAP files, and PLINK binary BED files.

Given their compact size, PLINK binary files are most commonly used to store and filter data. Since `GONE2` requires PED/MAP files as input, a workflow will typically convert between these two data types. This can be done using the following commands:

```
#Convert PLINK binary files to PED/MAP file
plink --bfile <filename> --recode --out <filename>

#Convert PLINK PED/MAP to PLINK binary files
plink --file <filename> --make-bed --out <filename>
```

Below more detailed information on both data types is given. This is additional reference information. It is not needed to run the practical, so you can skip it now, but it can be useful if you want to poke into the data files.

PLINK PED/MAP files are used to store genotype data in a text-based format. They consist of two main files:

- `.ped` file: Contains genotype and sample information.
- `.map` file: Contains variant information.

The `.ped` file is a tab-delimited text file where each row corresponds to an individual sample. It contains both sample metadata and genotype data. The columns are structured as follows:

- `Family ID`: Identifier for the family (can be set to 0 if not applicable).
- `Individual ID`: Unique identifier for the individual.
- `Paternal ID`: Identifier for the father (0 if not available).
- `Maternal ID`: Identifier for the mother (0 if not available).
- `Sex`: Sex of the individual (1 = male, 2 = female, 0 = unknown).
- `Phenotype`: Phenotype information (1 = unaffected, 2 = affected, -9 = missing).
- `Genotype Data`: Genotypes for each variant, represented as pairs of alleles (e.g., A A, A B, B B). Missing genotypes are denoted as 0 0.

The `.map` file is a tab-delimited text file where each row corresponds to a variant. It contains the following columns:

- `Chromosome`: Chromosome number or ID.
- `Variant ID`: Unique identifier for the variant (e.g., rsID).
- `Genetic Distance`: Genetic distance (can be set to 0 if not available).
- `Base-pair Position`: Physical position of the variant on the chromosome.

PLINK binary files are used to store genotype data in a compact and efficient binary format. They typically consist of three main files:

- `.bed` file: Contains the binary genotype data.
- `.bim` file: Contains variant information.
- `.fam` file: Contains individual sample information.

The `.bed` file stores the genotype data in a compact binary format. It does not have a human-readable structure but is designed for efficient storage and access by PLINK and other compatible tools.

Genotype Encoding:

- 00: Homozygous for the reference allele (AA)
- 01: Missing genotype
- 10: Heterozygous (AB)
- 11: Homozygous for the alternate allele (BB)

The `.bim` file is a text file that contains variant information. Each row corresponds to a variant, and it has the following columns:

- `chrom`: Chromosome number or ID
- `variant ID`: Unique identifier for the variant (e.g., rsID)
- `genetic distance`: Genetic distance (can be set to 0 if not available)
- `base-pair position`: Physical position of the variant on the chromosome
- `allele 1`: Reference allele (usually coded as the minor allele)
- `allele 2`: Alternate allele

The `.fam` file is a text file that contains information about each individual sample in the dataset. Each row corresponds to an individual and has the following columns:

- `Family ID`: Identifier for the family (can be set to 0 if not applicable)
- `Individual ID`: Unique identifier for the individual
- `Paternal ID`: Identifier for the father (0 if not available)
- `Maternal ID`: Identifier for the mother (0 if not available)
- `Sex`: Sex of the individual (1 = male, 2 = female, 0 = unknown)
- `Phenotype`: Phenotype information (1 = unaffected, 2 = affected, -9 = missing)

Exploring the data

Open this link in your browser to open the GONE training environment for this practical: [link]

You should be presented with a page where you can create a new GitHub Codespace. For this practical, the default environment with 2 cores should suffice.

In the training environment, you will see following folders:

- DATA: contains the data for this practical
- SCRIPTS: contains (plotting) scripts you can use
- GONE2: contains the GONE2 installation