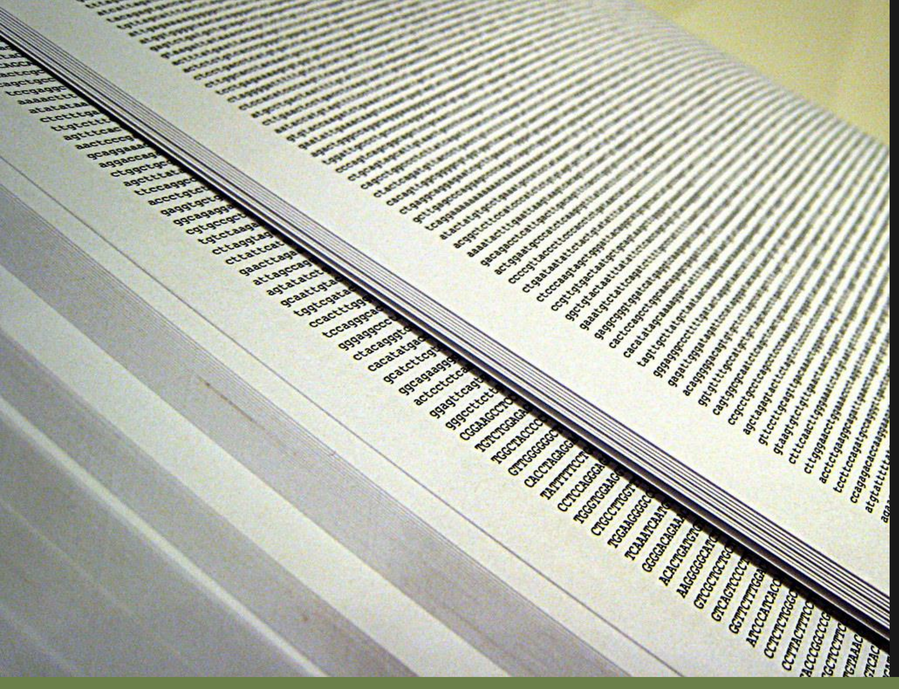
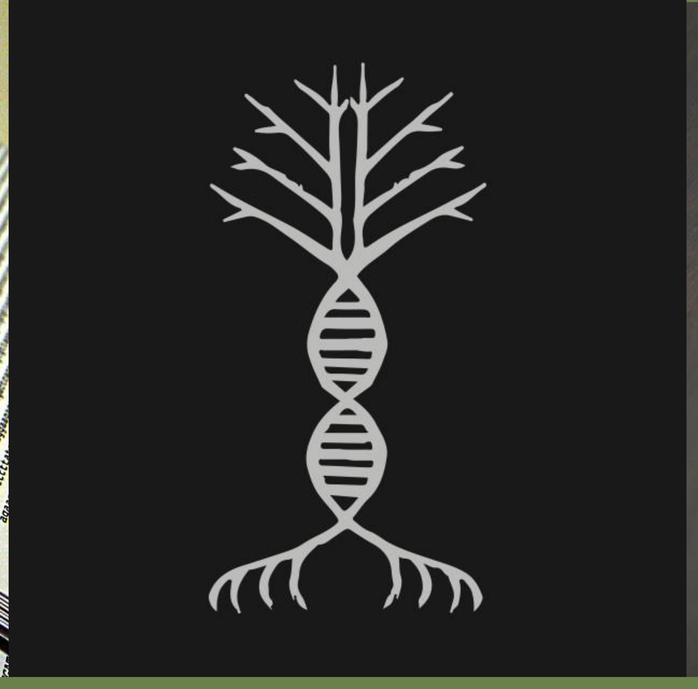




An automated bioinformatics pipeline: from importing SNP data to estimating population genetic parameters

Anna-Maria Farsakoglou[1], Ivan Scotti[2], Bruno Fady[2], Filippos A. Aravanopoulos[1]

[1] Laboratory of Forest Genetics and Tree Breeding, School of Forestry and Natural Environment, Aristotle University of Thessaloniki, Thessaloniki, Greece, aravanop@auth.gr

[2] INRAe, UR629, Ecologie des Forêts Méditerranéennes (URFM), Avignon, France

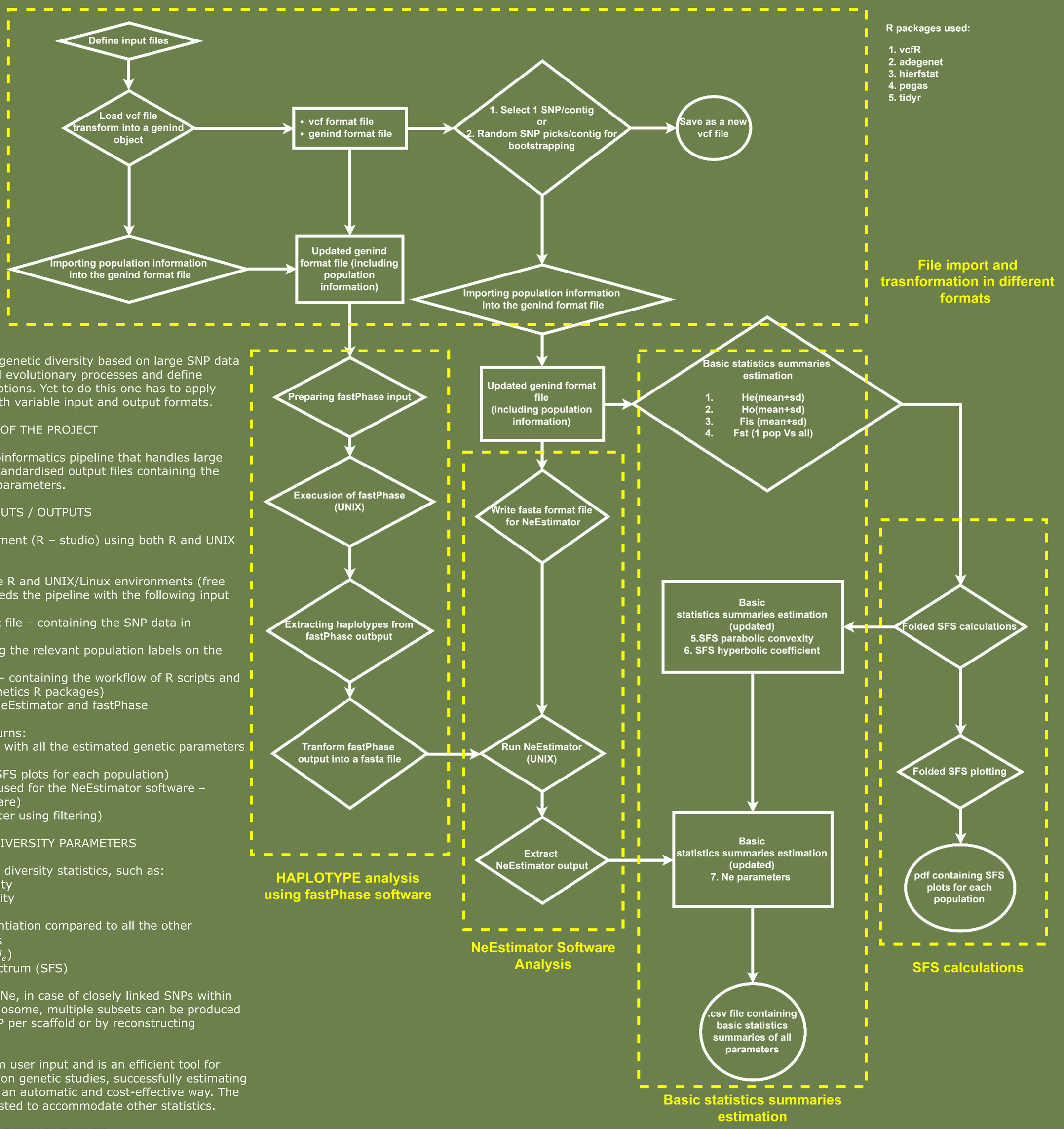
Software used:

1. **NeEstimator: http://www.molecularfisheries laboratory.com.au/neestimator-software**
2. **fastPhase: http://scheet.org/software.html**

R packages used:

1. **vcfR**
2. **adegenet**
3. **hierfstat**
4. **pegas**
5. **tidyr**

PIPELINE FLOWCHART



The description of patterns of genetic diversity based on large SNP data sets is the basis to understand evolutionary processes and define science-based management options. Yet to do this one has to apply multiple software packages with variable input and output formats.

AIM OF THE PROJECT

A user-friendly and flexible bioinformatics pipeline that handles large SNP data files and produces standardised output files containing the estimated population genetic parameters.

INPUTS / OUTPUTS

The pipeline runs in R environment (R – studio) using both R and UNIX commands.

We built our pipeline under the R and UNIX/Linux environments (free and open source). The user feeds the pipeline with the following input files:

- ✓ .vcf file (Variant Call Format file – containing the SNP data in compressed format .vcf.gz.)
- ✓ .txt file (text file - containing the relevant population labels on the first column)
- ✓ .Rmd file (R markdown file – containing the workflow of R scripts and commands from several genetics R packages)
- ✓ Other relevant files to run NeEstimator and fastPhase

The pipeline automatically returns:

- ✓ .csv file (containing a table with all the estimated genetic parameters for each population)
- ✓ .pdf file (containing folded SFS plots for each population)
- ✓ .dat file (a fasta file that is used for the NeEstimator software – potential use in other software)
- ✓ .vcf.gz file (a new vcf file after using filtering)

GENETIC DIVERSITY PARAMETERS

The pipeline computes genetic diversity statistics, such as:

1. Expected (H_e) heterozygosity
2. Observed (H_0) heterozygosity
3. Inbreeding coefficient (F_{IS})
4. $F_{ST(i)}$ as the genetic differentiation compared to all the other populations of each species
5. Effective Population Size (N_e)
6. Folded Site Frequency Spectrum (SFS)

For an unbiased estimation of Ne, in case of closely linked SNPs within the same scaffold/gene/chromosome, multiple subsets can be produced by selecting randomly one SNP per scaffold or by reconstructing haplotypes.

This pipeline requires minimum user input and is an efficient tool for analyzing SNP data in population genetic studies, successfully estimating standard diversity statistics in an automatic and cost-effective way. The open-source code can be adjusted to accommodate other statistics.

FUTURE IMPROVEMENTS

- ✓ Import different formats of genetic data (GenAlEx, Structure etc.)
- ✓ Estimate more genetic parameters (N_m, F_{ST} outliers etc.)
- ✓ Export data in other formats to be used in external software

ARISTOTLE UNIVERSITY OF THESSALONIKI
INRAE la science pour la vie, l'humain, la terre
GENTREE