Making sense of RAD-seq data: from filtering to population structure and genotype-environment association analyses

Pavel Škaloud

FACULTY OF SCIENCE
Charles University

Algal speciation & evolution lab

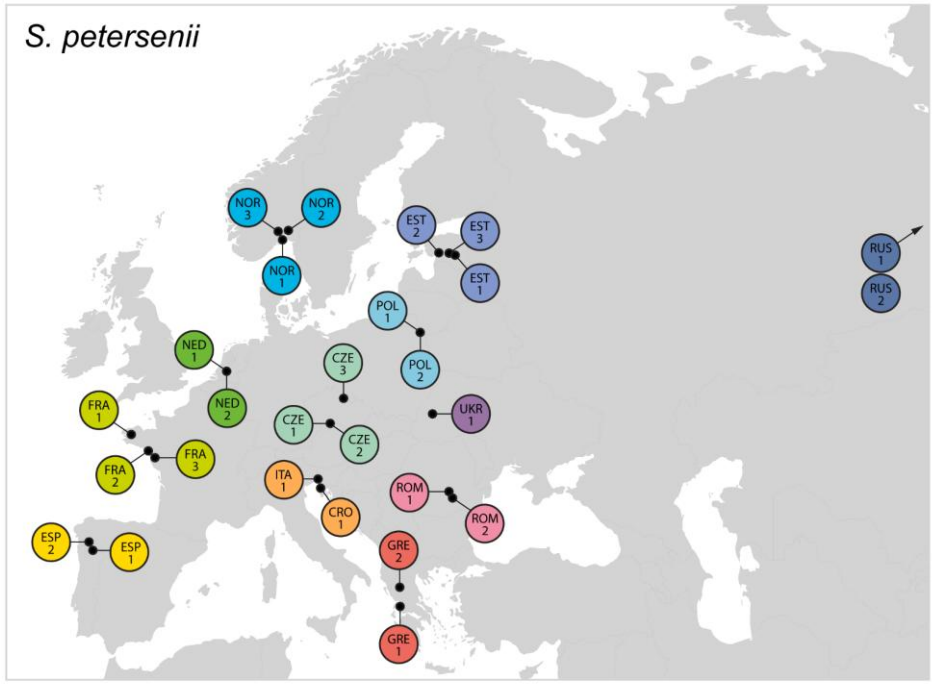
Škaloud, P., **Jadrná, I.**, Dvořák, P., **Škvorová, Z.**, **Pusztai, M.**, **Čertnerová, D.**, **Bestová, H.** & Rengefors, K. (2024)
Rapid diversification of a free-living protist is driven by adaptation to climate and habitat.
Current Biology 34: 92-105.

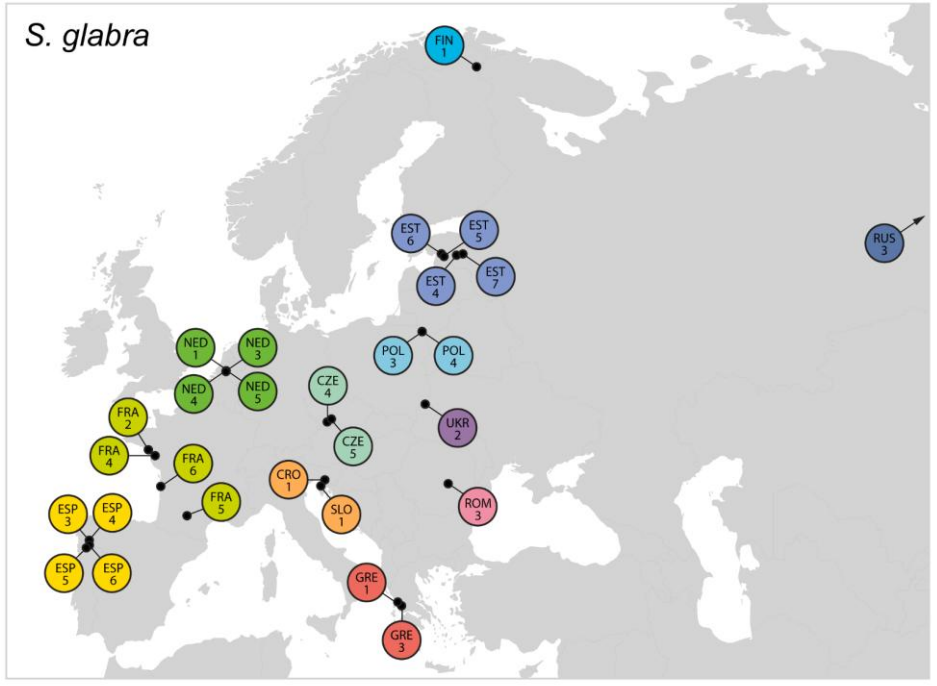
Jadrná, I., **Škvorová, Z.** & **Škaloud, P.** (2026)
Beyond protist generalizations: different evolutionary trajectories in closely related microbial species.
ISME Communications (accepted).

To download the papers:

To assess the Mendeley Data:

To download this presentation:





To download the paper:

To assess the Mendeley Data:

To download this presentation:

RAD-seq data

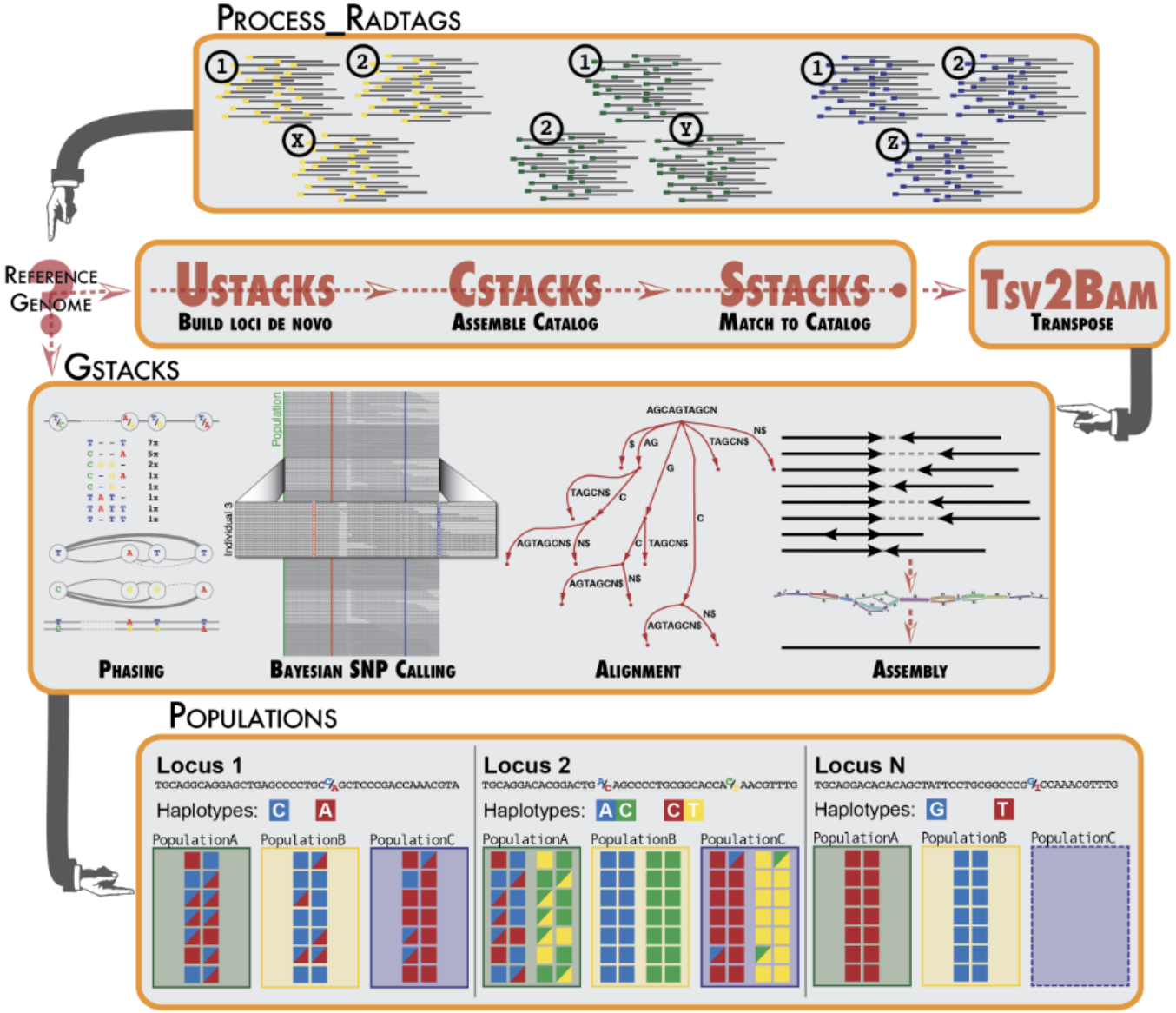
Working with RAD-seq data is much easier when you have access to **high-performance computing**, where you can submit the jobs. I typically run my pipelines on the Czech national HPC system (Metacentrum), but any comparable cluster works just as well.

cesnet metacentrum

- ABOUT METACENTRUM
- CURRENT AFFAIRS
- DOCUMENTATION AND SERVICES
- GETTING AN ACCOUNT
- MY ACCOUNT
- CURRENT STATE
- USER SUPPORT
- AMBASADORS
- SEMINARS
- PORTAL MAP

```
#!/bin/bash
#PBS -N RadSeq
#PBS -l select=1:ncpus=32:ompthreads=32:mem=64gb:scratch_local=200gb
#PBS -l walltime=6:00:00
#PBS -m ae

# define a DATADIR variable: directory where the input files are taken from and where output will be copied to
DATADIR=/auto/brno2/home/skaloud/ngsdata/RADseq/knihovna17

# append a line to a file "jobs_info.txt" containing the ID of the job, the hostname of node it is run on and the path to a scratch directory
echo "$PBS_JOBID is running on node `hostname -f` in a scratch directory $SCRATCHDIR" >> $DATADIR/jobs_info.txt

#loads the module
module add stacks

# test if scratch directory is set
# if scratch directory is not set, issue error message and exit
test -n "$SCRATCHDIR" || { echo >&2 "Variable SCRATCHDIR is not set!"; exit 1; }

# copy input files
# if the copy operation fails, issue error message and exit
cp $DATADIR/data/barcodes.txt $SCRATCHDIR || { echo >&2 "Error while copying input file(s)!"; exit 2; }
cp $DATADIR/data/K17-Skaloud_S23_L002_R1_001.fastq.gz $SCRATCHDIR || { echo >&2 "Error while copying input file(s)!"; exit 2; }
cp $DATADIR/data/K17-Skaloud_S23_L002_R2_001.fastq.gz $SCRATCHDIR || { echo >&2 "Error while copying input file(s)!"; exit 2; }

# move into scratch directory
cd $SCRATCHDIR
mkdir stacks

# run stacks
# if the calculation ends with an error, issue error message an exit
process_radtags -1 K17-Skaloud_S23_L002_R1_001.fastq.gz -2 K17-Skaloud_S23_L002_R2_001.fastq.gz -o stacks -b barcodes.txt -e sbfI -c -q -r --inline_null --
threads 32 || { echo >&2 "Calculation ended up erroneously (with a code $?) !!"; exit 3; }

# move the output to user's DATADIR or exit in case of failure
cp -R stacks/ $DATADIR/metacentrum/ || { echo >&2 "Result file(s) copying failed (with a code $?) !!"; exit 4; }

# clean the SCRATCH directory
clean_scratch
```

Stacks – analysis of raw RAD-seq data



https://catchenlab.life.illinois.edu/stacks/

Stacks – analysis of raw RAD-seq data

Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Methods in Ecology and Evolution 2017, 8, 1360–1373 doi: 10.1111/2041-210X.12775

Lost in parameter space: a road map for STACKS

Josephine R. Paris[1], Jamie R. Stevens[1] and Julian M. Catchen*[,2]

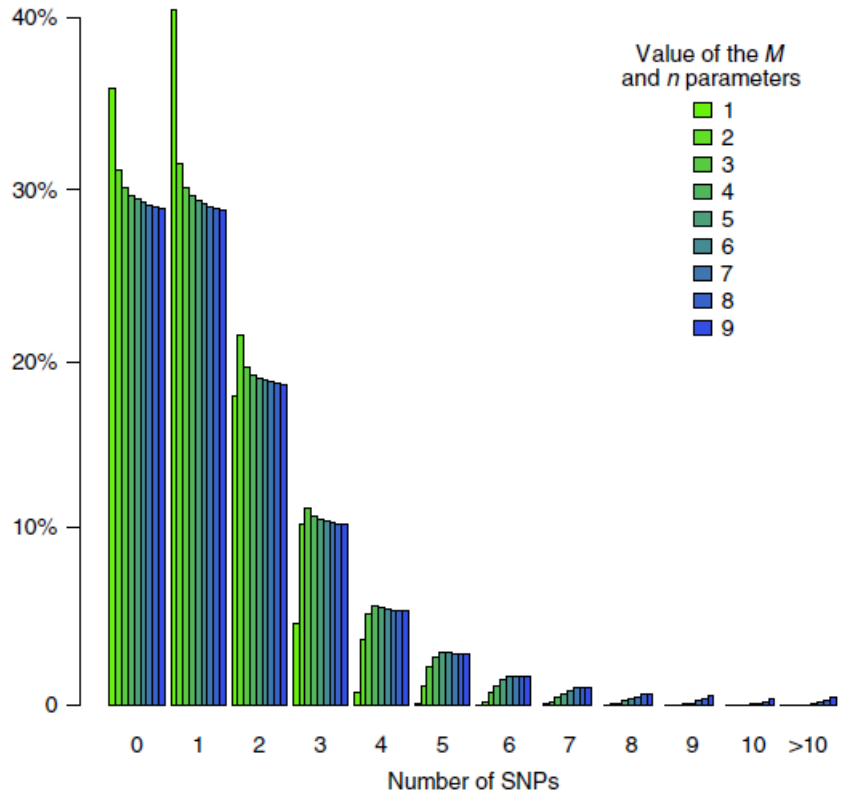
[1]*Biosciences, College of Life and Environmental Sciences, University of Exeter, Exeter, UK; and* [2]*Department of Animal Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA*

PROTOCOL

Deriving genotypes from RAD-seq short-read data using Stacks

Nicolas C Rochette & Julian M Catchen

Department of Animal Biology, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA. Correspondence should be addressed to J.M.C. (jcatchen@illinois.edu).



Parameter	Default value	STACKS component	Description
m	3	`ustacks`	Minimum number of raw reads required to form a stack (a putative allele)
M	2	`ustacks`	Number of mismatches allowed between stacks (putative alleles) to merge them into a putative locus
n	1	`cstacks`	Number of mismatches allowed between stacks (putative loci) during construction of the catalog

Stacks – analysis of raw RAD-seq data

Barcode sequences

Restriction enzyme

Position of barcodes

process_radtags -1 read1.fastq.gz -2 read2.fastq.gz -o stacks -b barcodes.txt -e sbfI -c -q -r --inline_null --threads 32

61906032 total sequences
1865322 barcode not found drops (3.0%)
211988 low quality read drops (0.3%)
10715922 RAD cutsite not found drops (17.3%)
49112800 retained reads (79.3%)

Stacks parameters

Population assignment of samples

denovo_map.pl -M 4 -n 4 -m 3 -T 24 --popmap popmap_file -o outputdir/ --samples . --paired

Single SNP per locus

Output formats

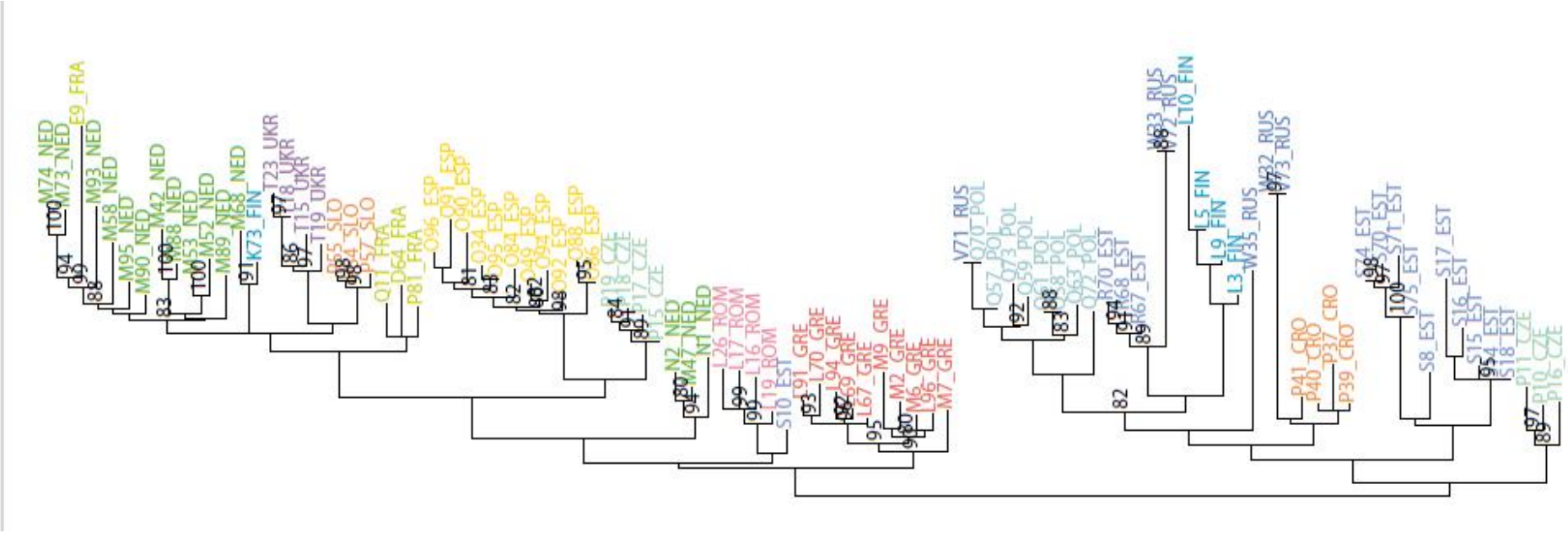
populations -P . -t 24 -M popmap_file -O outputdir --write-single-snp --phylip --structure --vcf

Initial check of the data structure – IQ tree

Download and clean the PHYLIP output file by correcting the spacing between the two numbers in the first line (only one space) and deleting the trailing line that contains “#”.

Phylip output file

iqtree -s populations.fixed.phylip -m GTR+ASC -bb 1000 -redo

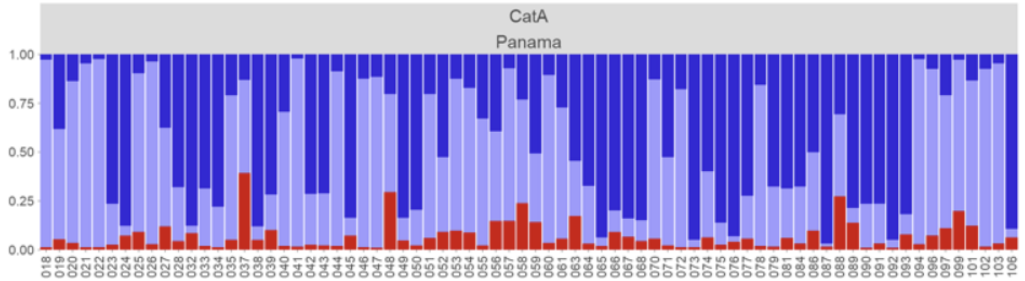


Initial check of the data structure – STRUCTURE

- **Uploading the** `mainparams` **and** `extraparams` **files** into the main directory
- **Preprocessing all input files using a single script:**
  - **Set the number of individuals** by assigning the appropriate value to `INDS` (e.g., `INDS=36`).
  - **Extract the** `NUMLOCI` **value** and insert it into the copied `mainparams` file.
  - **Remove the first comment line** from `populations.structure`.

```
INDS=36
NUM=$(head -2 populations.structure | tail -1 | tr '\t' '\n' | wc -l)
NUM=$((NUM - 2))
echo $NUM
cp ../mainparams .
cp ../extraparams .
sed -i "s/^#define NUMLOCI [0-9]\+/#define NUMLOCI ${NUM}/" mainparams
sed -i "s/^#define NUMINDS [0-9]\+/#define NUMINDS ${INDS}/" mainparams
sed -i '1{/^#/d}' populations.structure
```

- **Run STRUCTURE**
- **Visualization by pophelper.R**



pophelper

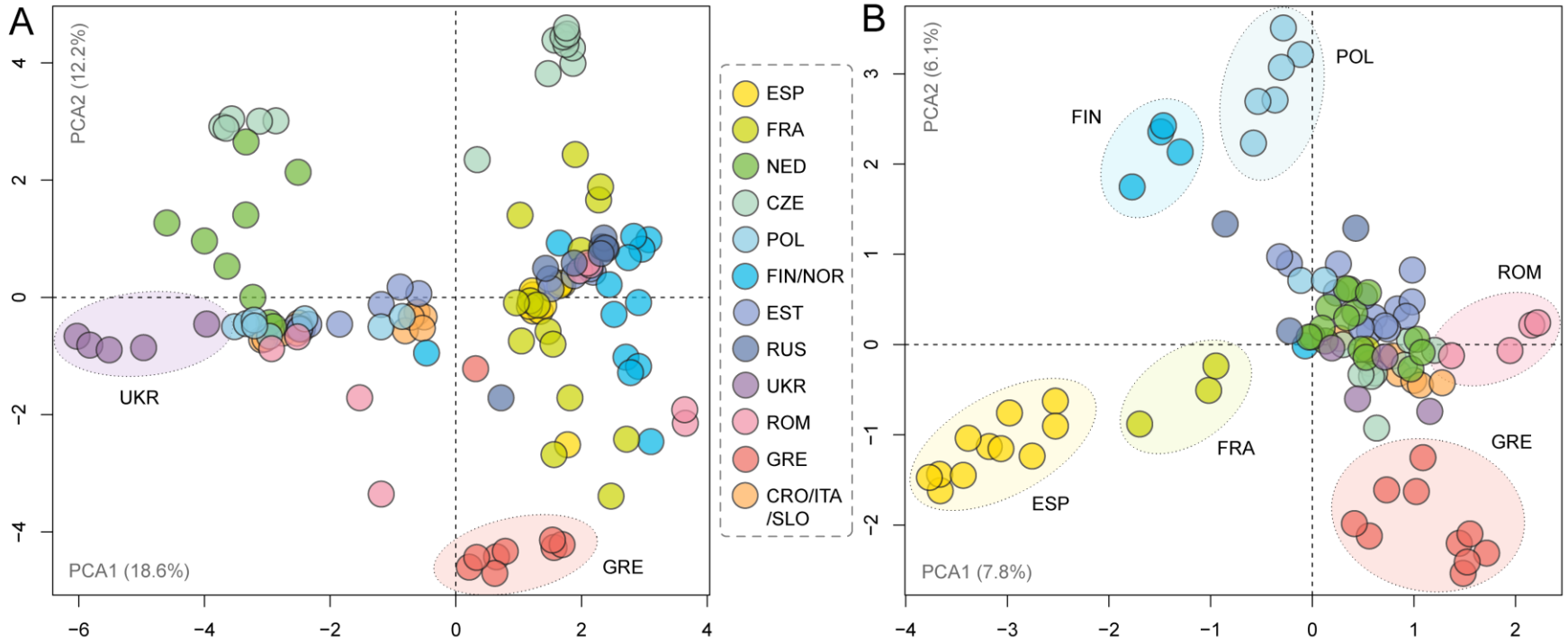
pophelper is an R package and web app to analyse and visualise population structure. **pophelper** curently supports output run files generated from population analysis programs such as STRUCTURE, TESS, TESS3, BAPS and numeric delimited formats such as ADMIXTURE or fastSTRUCTURE. The **pophelper** package can be used to read run files to R, tabulate runs, summarise runs, estimate *K* using the Evanno method, align clusters within and across K and generate barplot figures.

For a detailed demonstration and walkthrough, refer the online vignette. For information about changes in the latest version, check version history.

Initial check of the data structure – PCA

- **In R**
- **Converting the vcf output file to genlight**
- **Using the function glPcaFast** (https://botany.natur.cuni.cz/hodnocenidat/Lesson_05_tutorial.pdf) to run the PCA

```
vcf <- read.vcfR("path to vcf file")
aa.genlight <- vcfR2genlight(vcf)
pca.2 <- glPcaFast(aa.genlight, nf=300)
```



Parameter testing and locus filtering in Stacks

Testing parameter combinations in the RAD-seq pipeline – *denovo_map.pl*

- Using various parameters (*m*, *M*, *n*) across a defined range (e.g., m 3-5, M+n 1-6)
- Selection of the optimal combination based on the highest number of polymorphic loci present in ≥80% of samples.

Testing the M parameter

```
denovo_map.pl -M 1 -n 1 -m 3 -T 24 --popmap popmap_file -o denovo_m3M1/ --samples . –paired
denovo_map.pl -M 2 -n 2 -m 3 -T 24 --popmap popmap_file -o denovo_m3M2/ --samples . --paired
denovo_map.pl -M 3 -n 3 -m 3 -T 24 --popmap popmap_file -o denovo_m3M3/ --samples . --paired
```

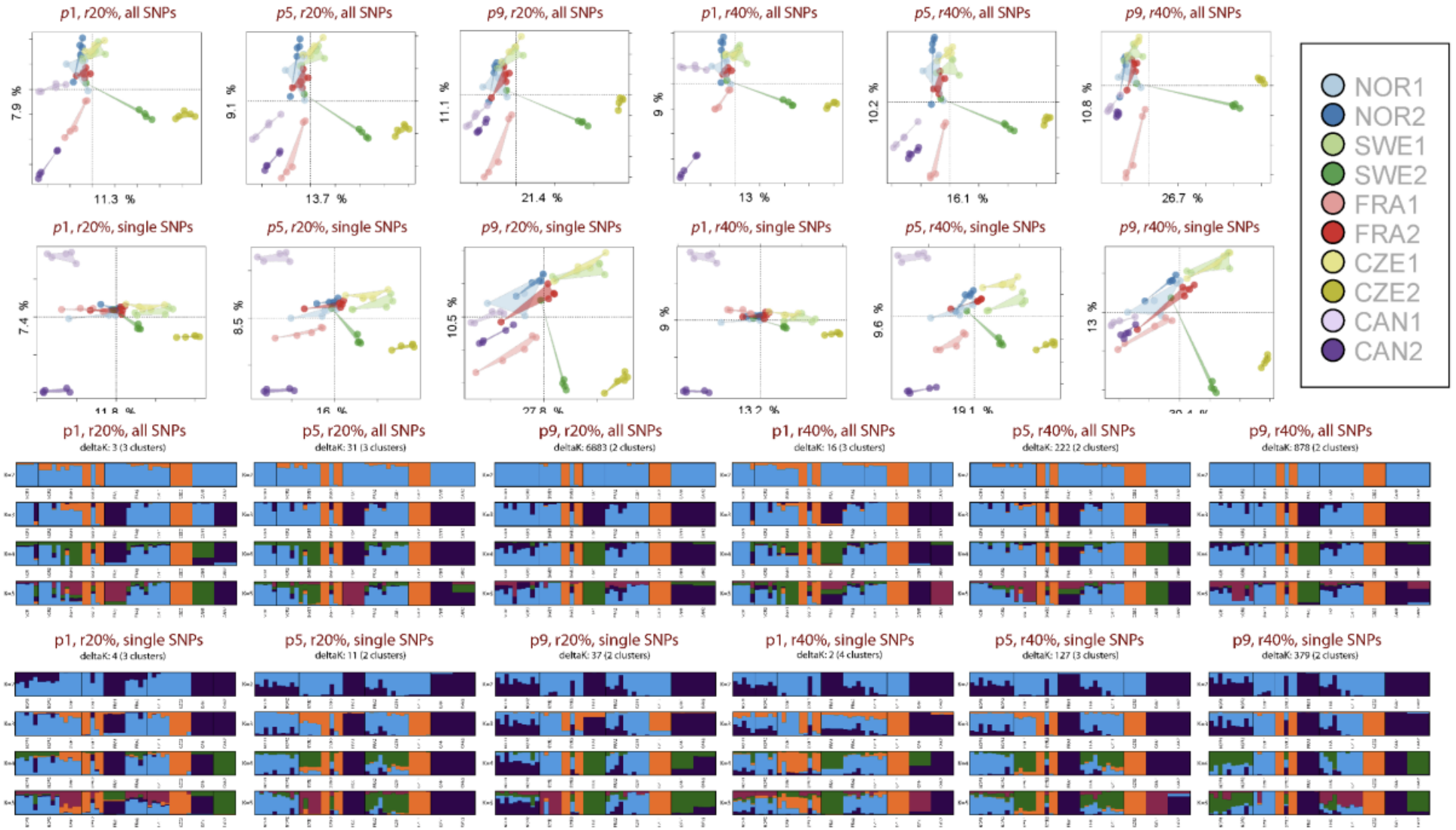
Testing locus-filtering thresholds - *Populations*

- `--min-populations` (*p*): minimum number of populations in which a locus must occur.
- `--min-samples-per-pop` (*r*): minimum percentage of individuals required within each population.
- `min-samples-overall` (*R*): minimum percentage of individuals across all populations required to retain a locus. (e.g., thresholds: **3%, 5%, 10%, 20%**, etc.)
- Selection of the optimal locus filtering thresholds by inspecting Structure/IQ tree/PCA results

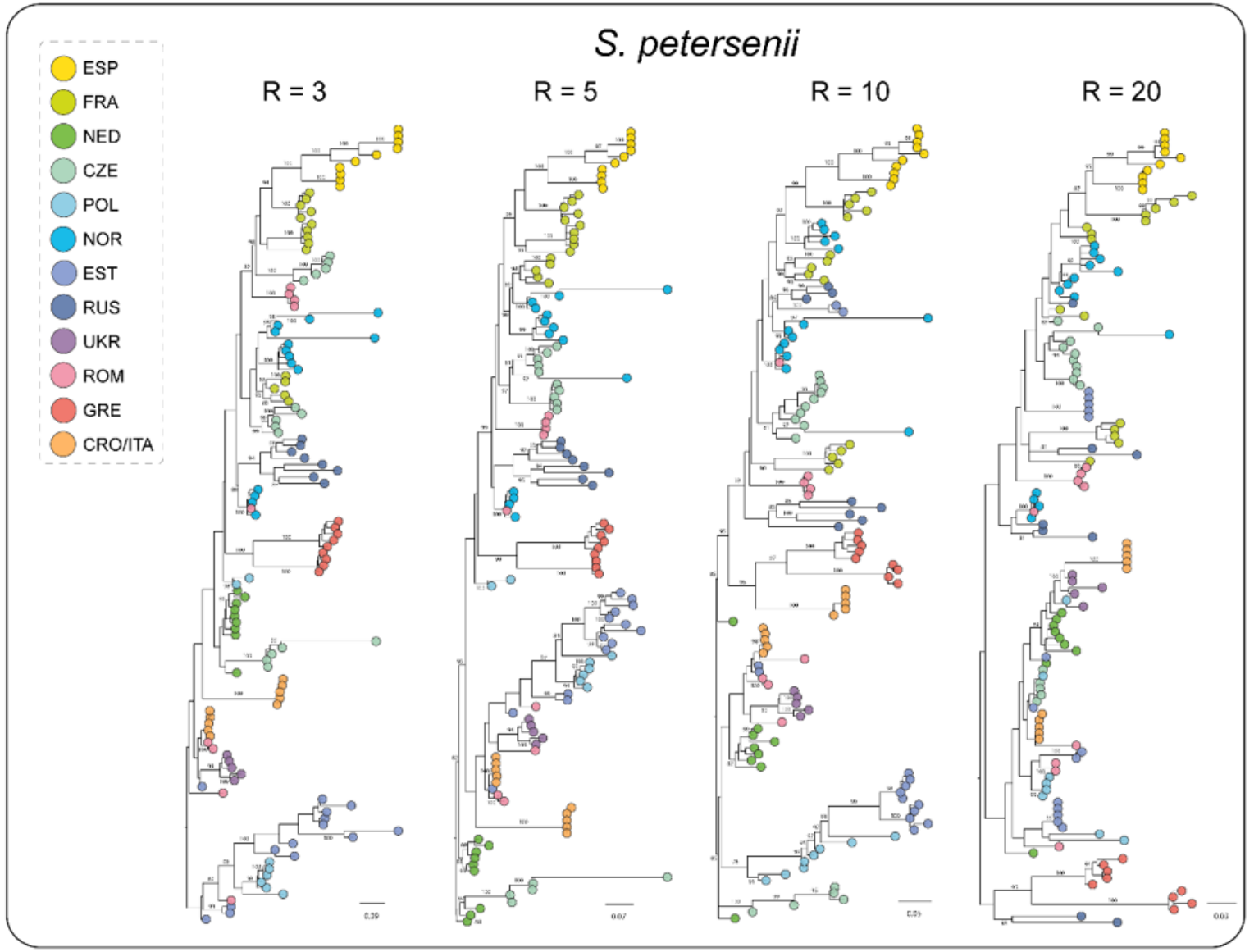
Testing the R filtering option

```
populations -P . -t 24 -M popmap_file -O populations_R05 -R 5 --write-single-snp --phylip –structure
populations -P . -t 24 -M popmap_file -O populations_R10 -R 10 --write-single-snp --phylip –structure
populations -P . -t 24 -M popmap_file -O populations_R20 -R 20 --write-single-snp --phylip --structure
```

Parameter testing and locus filtering in Stacks



Parameter testing and locus filtering in Stacks



Final STRUCTURE analyses

Removing loci with low abundance (optional)

- Export selected loci to `loc.txt` in R.
- Convert VCF → STRUCTURE using **PGDSpider** (edit SPID file: add locus IDs under *Only input following regions*).
- Save SPID file and run conversion.
- Open output in Excel, split by space, save as tab-delimited file.
- Re-save the file on Linux using an R script.

```
vcf <- read.vcfR("file.vcf")
dp <- extract.gt(vcf, element = "DP", as.numeric=TRUE)
myMiss <- apply(dp, MARGIN = 1, function(x){ sum( is.na(x) ) } )
myMiss <- myMiss / ncol(dp)vcf2 <- vcf[myMiss < 0.5, ]
write.table(vcf2@fix[,1], file = "loci.txt")
```

ParallelStructure

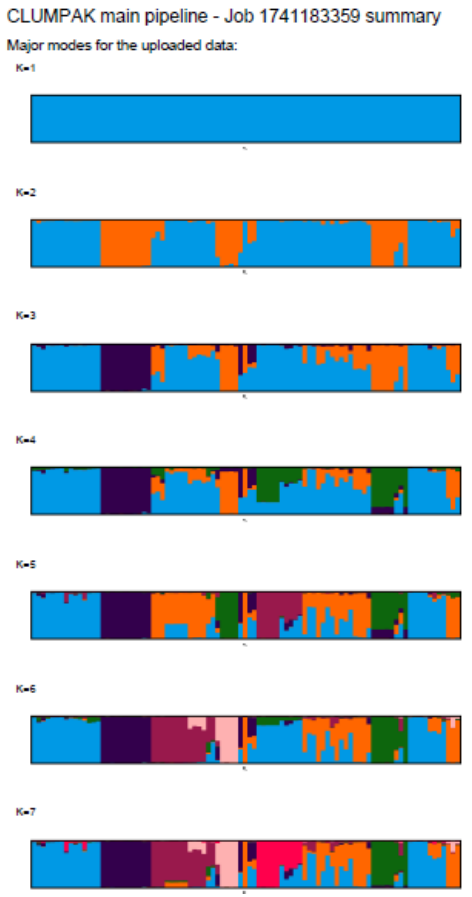
- Parallel runs on the HPC cluster: https://github.com/V-Z/structure-multi-pbspro
- For each **K = 1–10**, run **50 independent STRUCTURE replicates** (total = 500 runs).
- Burn-in = **50,000**, MCMC after burn-in = **100,000**.

CLUMPAK

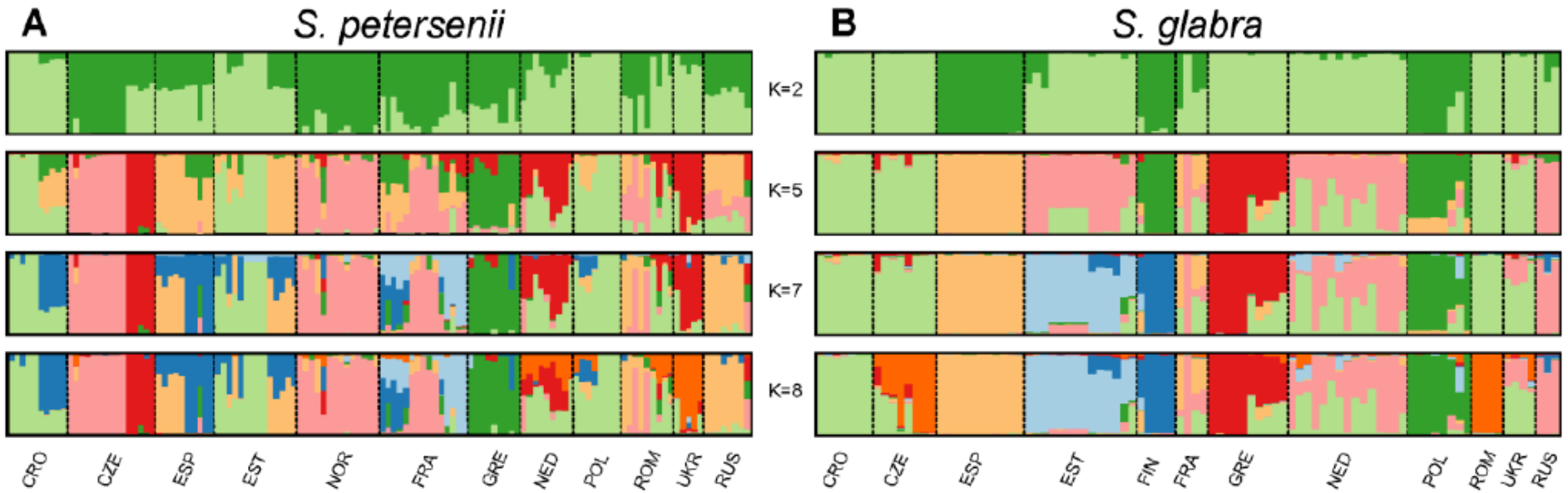
- Analyse ParallelStructure outputs.
- Main Pipeline: identify major/minor clustering modes and generate plots.
- BestK: compute ΔK.

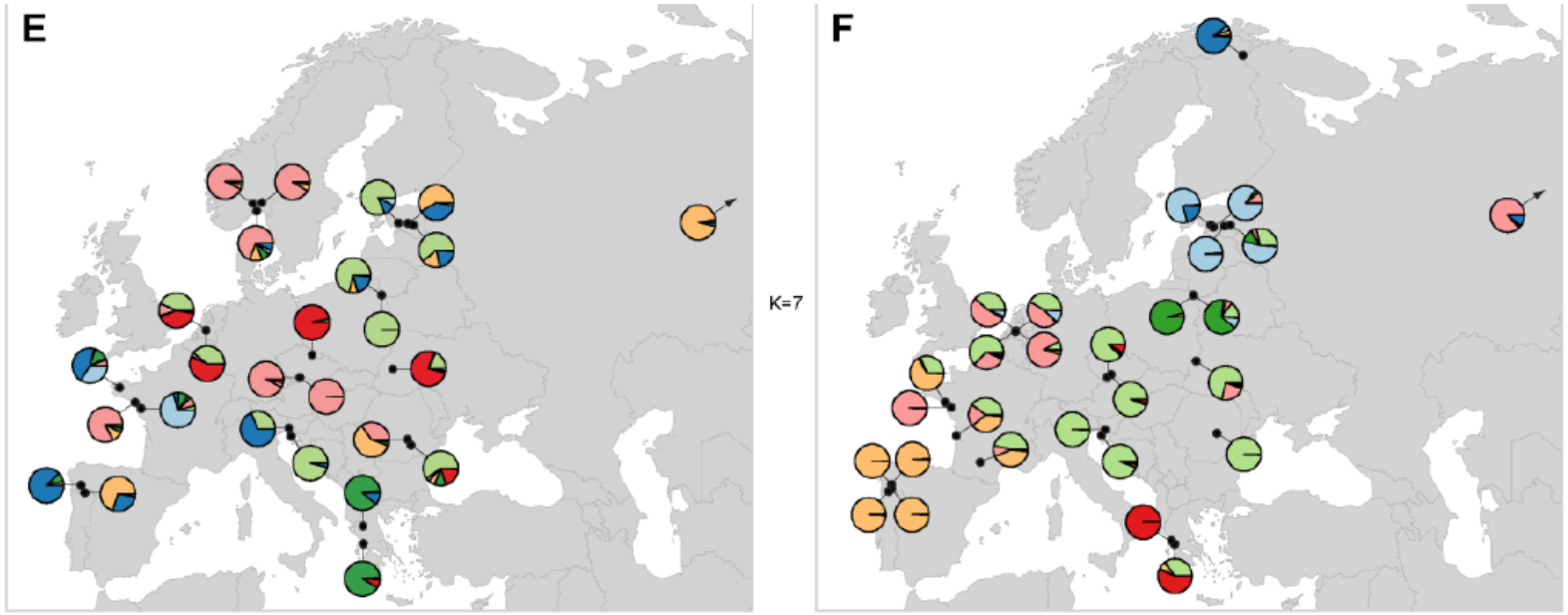
Mapping STRUCTURE results

- Visualize inferred genetic clusters on a map of Europe using the R script.



Final STRUCTURE analyses





Genotype-environment association analyses

Outlier loci detection to detect loci which are likely to be affected by selection

- Vcf file generated by Stacks using as the input file.
- Using a custom script to identify and extract outlier loci *(https://github.com/dvorikus/Synura-RADs)*.
- Outlier loci are detected based on **q-values**.

Obtaining climatic or chemical soil data:

- Worldclim
- CHELSA
- Soilgrid

WorldClim

Historical climate data

This is WorldClim version 2.1 climate data for 1970-2000. This version was releas
2020.

There are monthly climate data for minimum, mean, and maximum temperature,
solar radiation, wind speed, water vapor pressure, and for total precipitation. The
"bioclimatic" variables.

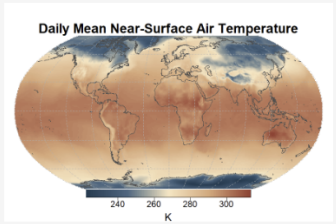
The data is available at the four spatial resolutions, between 30 seconds (~1 km2
minutes (~340 km2). Each download is a "zip" file containing 12 GeoTiff (.tif) fi
each month of the year (January is 1; December is 12).

CHELSA

Home › Datasets › CHELSA-climatologies

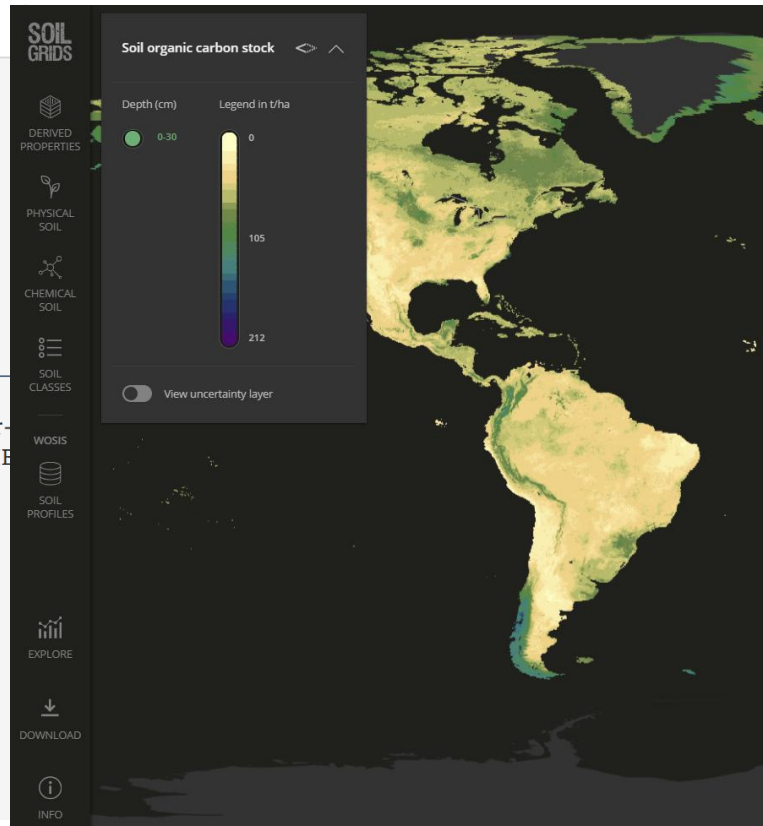
CHELSA-climatologies

Overview | Variables | Related Datasets



CHELSA-climatologies is a global, kilometer-
scale climate dataset generated with the CHE
downscaling model. It consists of long term
climatological means.

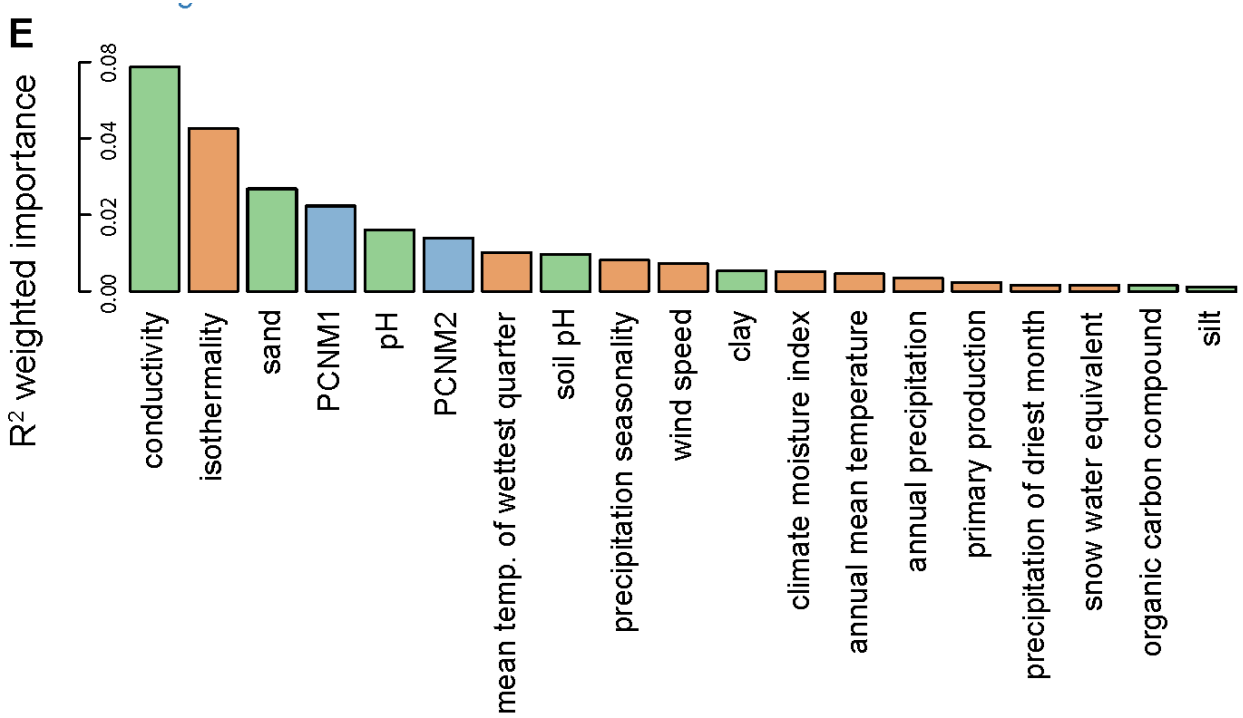
Start Time	1981-2010
End Time	2071-2100
Status	active
Version	2.1

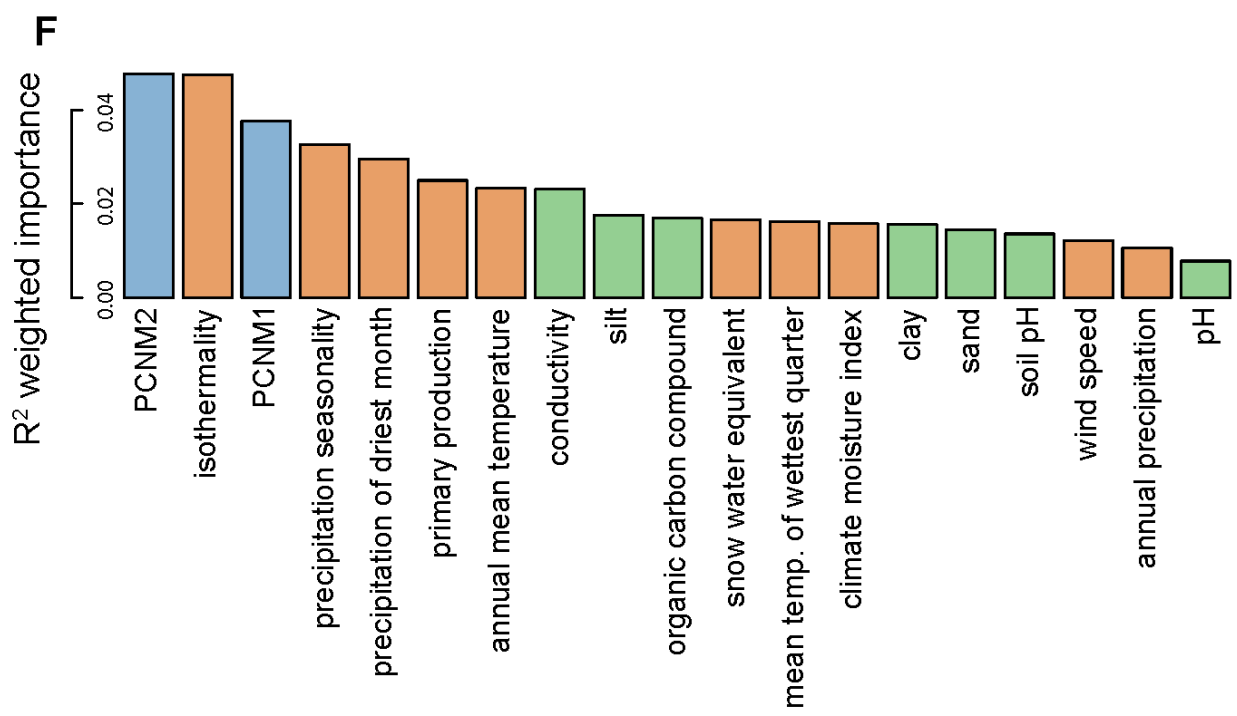


Genotype-environment association analyses

Gradient forest analysis to assess the relative effects of variables on the genetic diversity

- Run in R using *gradientForest* and *extendedForest* packages.
- Accounting for hierarchical sampling by setting *locality* as a random factor and performing 300 runs
- Model settings: 500 trees, correlation threshold 0.5, maxLevel = 3.

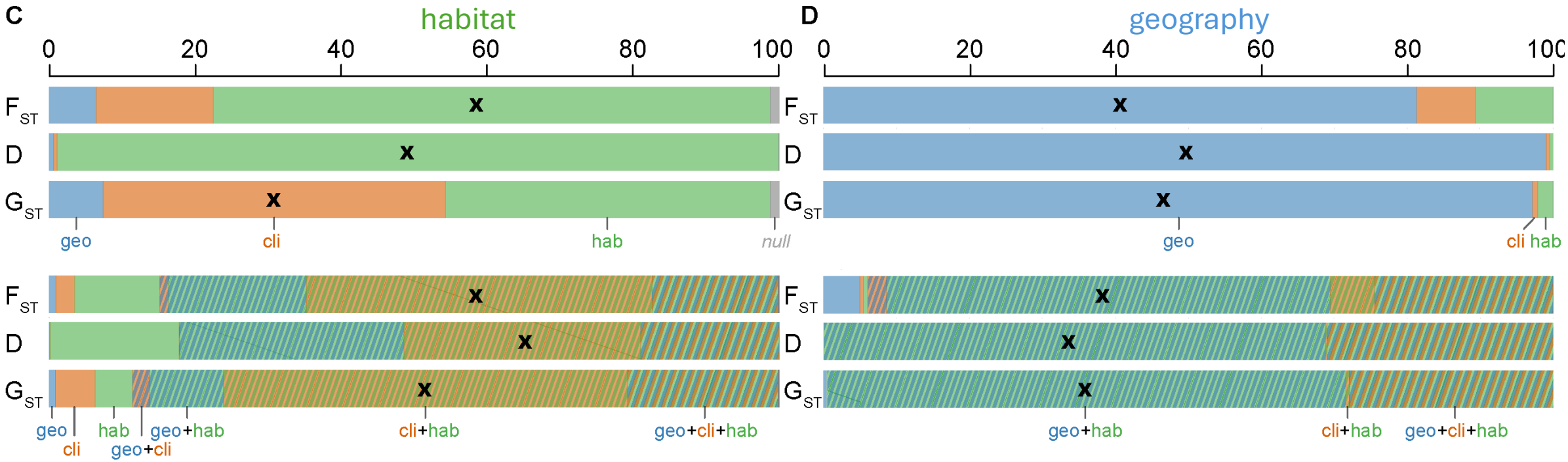




Genotype-environment association analyses

Generalized linear mixed modelling (GLMM) to assess the relative effects of variables on the genetic diversity

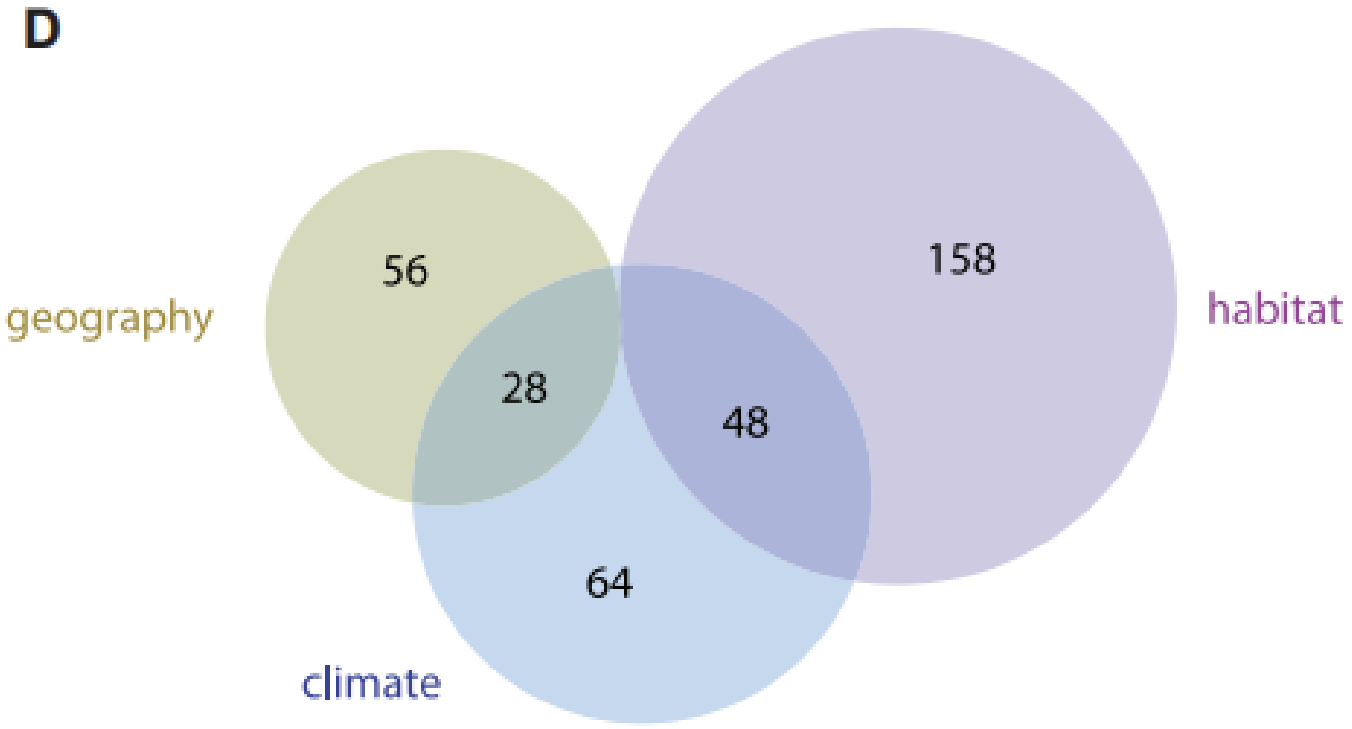
- Run in R using MCMCglmm
- Response variables: pairwise F_{ST}, G_{ST}, D_{Jost} distances (obtained from Stacks and FinePop.
- Eight models tested: null model, three single-predictor models, three two-predictor combinations, full model with all predictors.
- Model comparison via DIC, ΔDIC, and DIC weights.



Genotype-environment association analyses

Bayesian hierarchical modelling (BHM) to identify loci associated with the analyzed factors

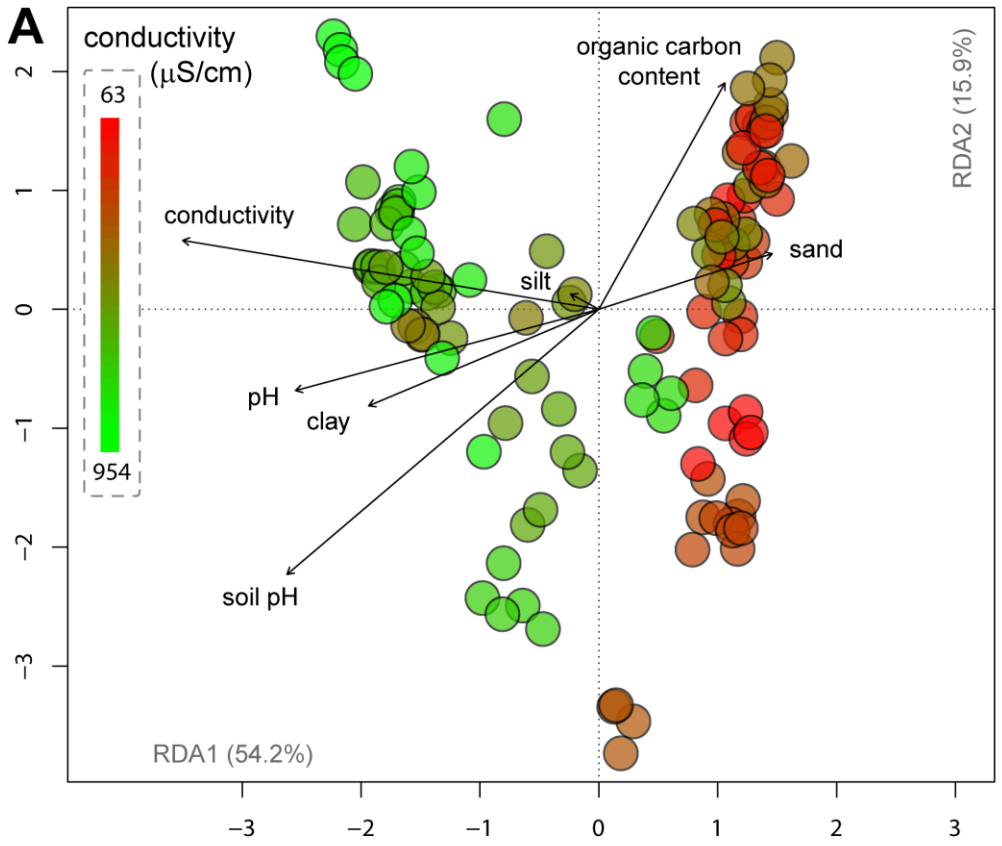
- Run in BayPass v2.1 under the standard covariate model (default settings).
- Input: population allele counts, converted from VCF using geste2baypass.py (RAD_Tools)

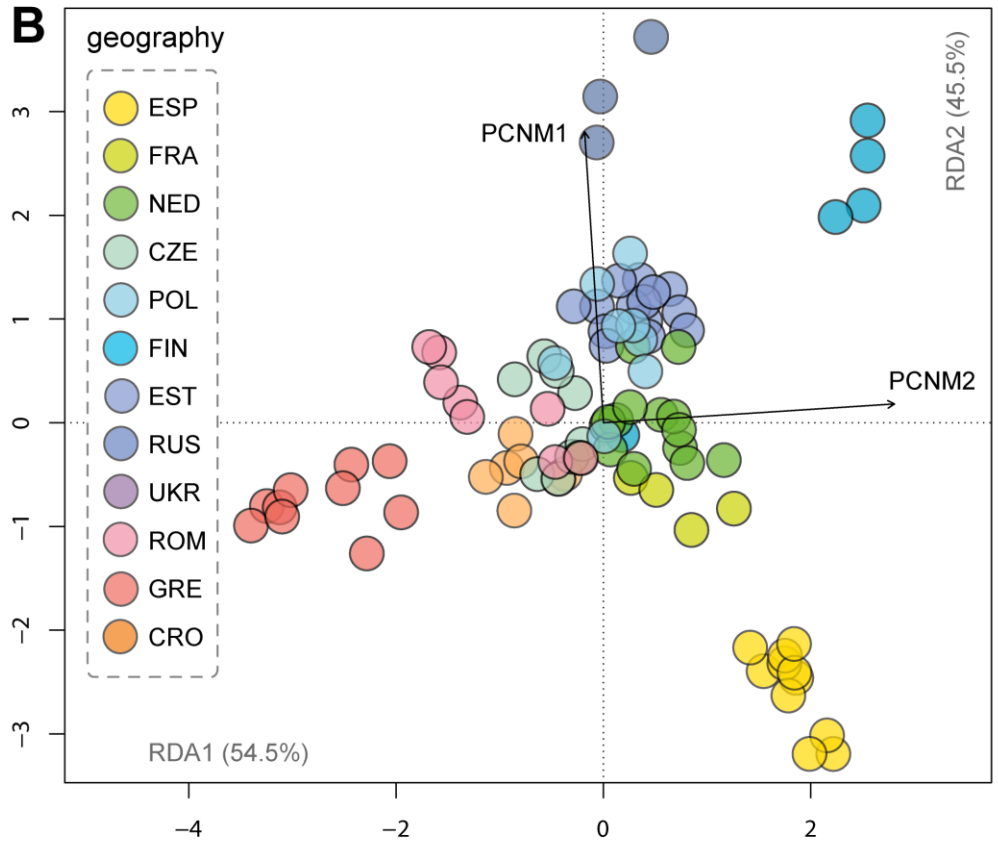


Genotype-environment association analyses

Redundancy Analysis (RDA) to detect the associations of environmental variables with genetic diversity

- Performed in R using the package vegan
- Missing genotypes replaced by the most common genotype per SNP.
- Genepop file loaded, converted to a data frame, missing values imputed via a simple mode-function.
- RDA model: allele matrix ~ environmental predictors
- Significance tested using anova.cca with 999 permutations (p = 0.001).





Thank you for your attention!