ADEGENET E PARALLELSTRUCTURE: ANÁLISES XENÓMICOS DE ESTRUTURA POBOACIONAL EN ÁRNICA

Fernando Cabana[1], **Adrián Casanova** [1], **Manuel A. Rodríguez-Guitián**[2], **Andrés Blanco**[1], **Carlos Real**[3], **Rosa Romero**[2], **Carmen Bouza**[1], **Manuel Vera**[1]

[1] Departmento de Zooloxía, Xenética e Antropoloxía Física. Facultade de Veterinaria. Universidade de Santiago de Compostela. Campus Terra. 27002 Lugo, España
[2] Departmento de Produción Vexetal e Proxectos de Enxeñaría, Escola Politécnica Superior de Enxeñaría, Universidade de Santiago de Compostela. Campus Terra, 27002 Lugo, España
[3] Departmento de Bioloxía Funcional, Escola Politécnica Superior de Enxeñaría Universidade de Santiago de Compostela, Campus Terra, 27002 Lugo, España

Índice

INTRODUCIÓN

Arnica montana L.

- Planta asterácea de reprodución sexual e asexual.
- Distribuida ao longo de Europa, dende Noruega ata Montenegro, e dende Ucraína ata a Península Ibérica.
- Aínda que clasificada como "Least Concern" pola IUCN, atópase en listas vermellas de certos países pola disminución das poboacións.
- Colleita sen control e degradación dos hábitats.

- Propiedades antiinflamatorias.
- Lactonas sesquiterpénicas
  - Helenalinas
  - Dihidrohelenalinas
- Dous quimiotipos
  - + Potencia, + Alerxénico
  - - Potencia, - Alerxénico
    - Limitado á Península Ibérica e sur de Francia

SNP

Polimorfismo de nucleótido único.

METODOLOXÍA | PAQUETES DE R

...AAAGCTGGCATCGATTGGATTGCGAAGCACCCACTCGGCAGGGATTTGCATCTCAGTTTAAGGGCTAGTTACTGATGA...

AlfI encima de restrición (tesoira molecular)

Stacks

SNP 154_26 (CA-056)

T/C

Profundidade de cobertura = 7x

AGCACCCACTCGGCAGGGATTTGCATCTCAGTTTAA
AGCACCCACTCGGCAGGGATTTGCATCTCAGTTTAA
AGCACCCACTCGGCAGGGATTTGCATCTCAGTTTAA
AGCACCCACTCGGCAGGGATTTGCATCTCAGTTTAA
AGCACCCACTCGGCAGGGATTTGCATCTCAGTTTAA

T
5x

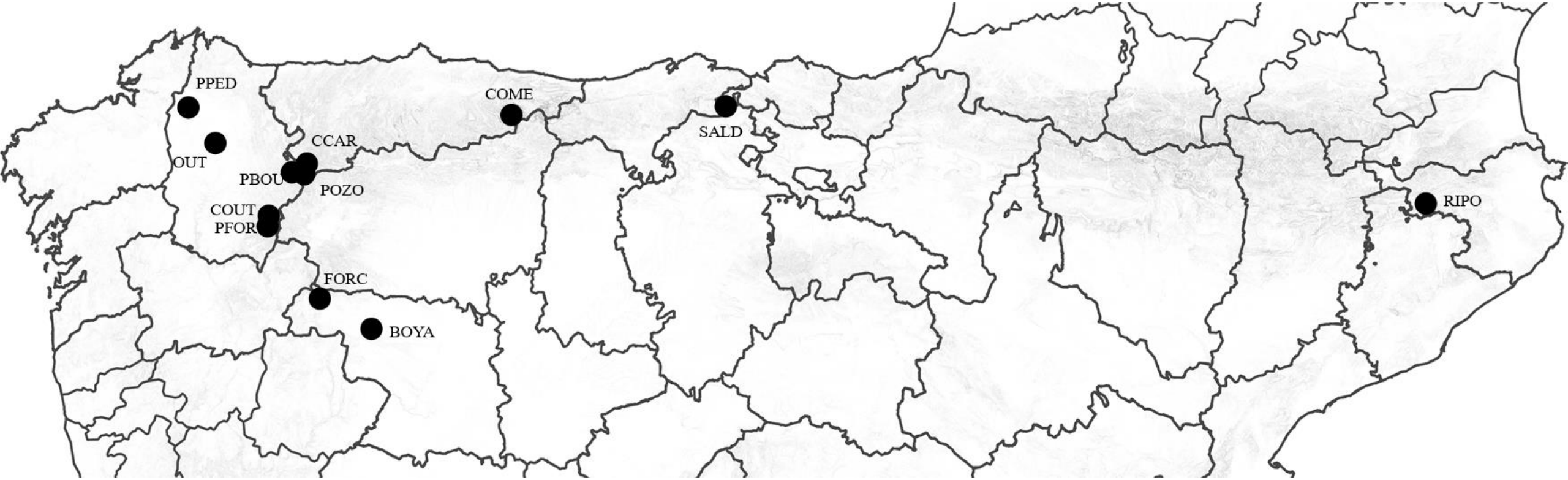
AGCACCCACTCGGCAGGGATTTGCACCTCAGTTTAA
AGCACCCACTCGGCAGGGATTTGCACCTCAGTTTAA

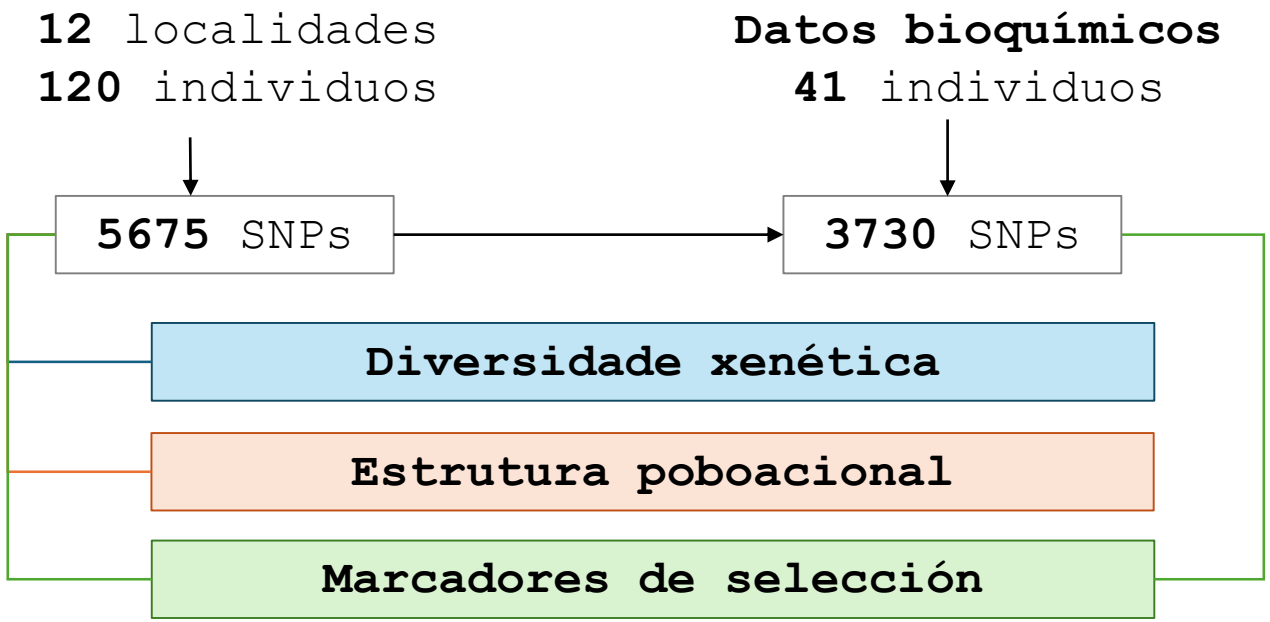
C
2x

- SNPs
- 12 localidades
- 120 individuos
- **5675 SNPs**

- Datas bioquímicos
- 6 localidades
- PPED-PBOU-CCAR POZO-COUT-PFOR
- 41 individuos
- **3730 SNPs**

Localidade	Código	N
Ponte Pedrido (Guitiriz, Lugo)	PPED	10
Outeiro de Rei (Lugo)	OUT	12
Ponte de Bous (Serra dos Ancares, Lugo)	PBOU	12
Catro Carballos (Serra dos Ancares, León)	CCAR	5
Marco do Pozo (Serra dos Ancares, Lugo)	POZO	4
Alto do Couto (Serra do Courel, Lugo)	COUT	10
Pico Formigueiros (Serra do Courel, Lugo)	PFOR	9
Covadonga (Asturias)	COME	11
Boya (Zamora)	BOYA	13
Forcadura (Zamora)	FORC	14
Salduero (Vizcaia)	SALD	13
Ripollés (Girona)	RIPO	7


PPED
COME
SALD
CCAR
OUT
PBOU
POZO
COUT
PFOR
RIPO
FORC
BOYA


12 localidades
120 individuos
Datos bioquímicos
41 individuos
5675 SNPs
3730 SNPs
Diversidade xenética
Estrutura poboacional
Marcadores de selección

METODOLOXÍA | **PAQUETES DE R**

- DiveRsity
- PopGenReport
- genepop
- StAMPP
- pcadapt

- adegenet
- ParallelStructure

- ParallelStructure
  - STRUCTURE
  - StructureSelector
  - CLUMPAK

- adegenet
  - Inferir a estrutura xenómica en dúas etapas
    1. Definición do número de unidades poboacionais máis probable
       Criterio de Información Bayesiana (BIC)
    2. Análises DAPC
       - Análises Discriminantes de Compoñentes Principais
       - Mostrar as diferenzas entre grupos minimizando a variación dentro dos grupos
       - Análises de compoñentes principais (PCA)

       +
       - Análises discriminantes (DA)

Exploratory Analysis of Genetic and Genomic Data

Documentation for package ‘adegenet’ version 2.1.10

- DESCRIPTION file.

Help Pages

A B C D E F G H I K L M N O P R S T U V W X misc

-- A --

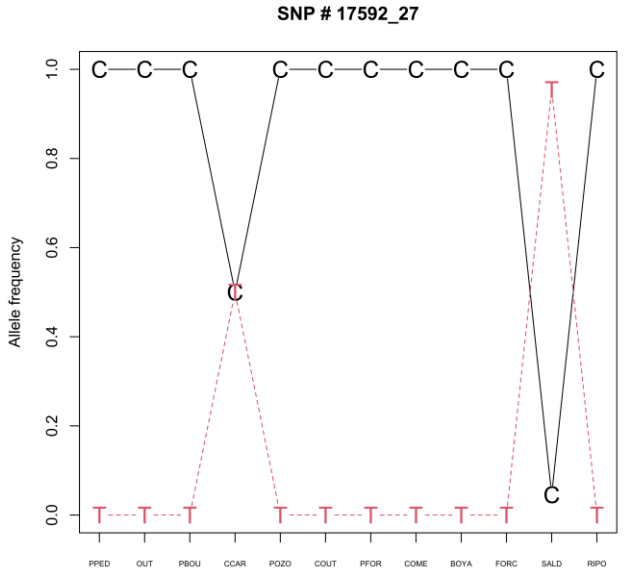
a.score	Compute and optimize a-score for Discriminant Analysis of Principal Components (DAPC)
addStrata	Access and manipulate the population strata for genind or genlight objects.
addStrata-method	Access and manipulate the population strata for genind or genlight objects.
addStrata<-	Access and manipulate the population strata for genind or genlight objects.
addStrata<--method	Access and manipulate the population strata for genind or genlight objects.
adegenet	The adegenet package
adegenet.package	The adegenet package
adegenetIssues	Functions to access online resources for adegenet
adegenetServer	Web servers for adegenet
adegenetTutorial	Functions to access online resources for adegenet

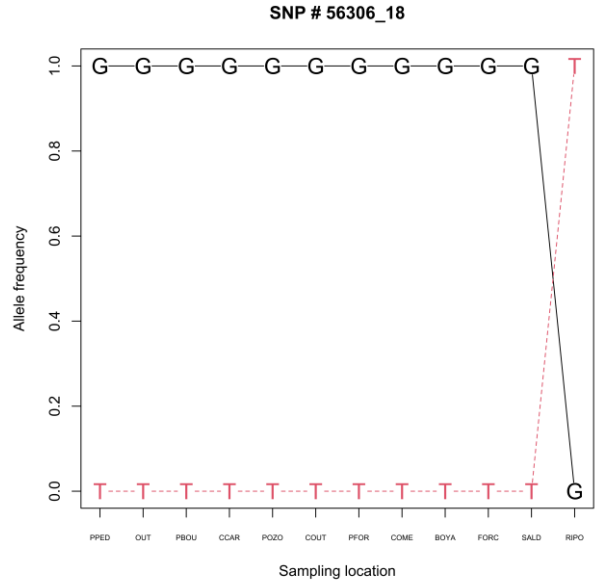
```r
library(adegenet)

sessionInfo()

# [1] adegenet_2.1.10 ade4_1.7-22

# Ler GENEPOP (input)
genind_Am_N120_5675<-read.genepop(file = "5675_SNPs_AmontanaN120_FINAL.gen",ncode = 3L,quiet = FALSE)

#Identificar clústers
Find.cluster <- find.clusters(genind_Am_N120_5675, max.n.clust=20, n.iter = 1000000, n.start = 100)
Find.cluster$Kstat
Find.cluster$grp

#DAPC con agrupamiento a priori
dapc1 <- dapc(genind_Am_N120_5675, genind_Am_N120_5675$pop)
#DAPC con agrupamiento segundo Find clusters
dapc1B <- dapc(genind_Am_N120_5675, Find.cluster$grp)
```

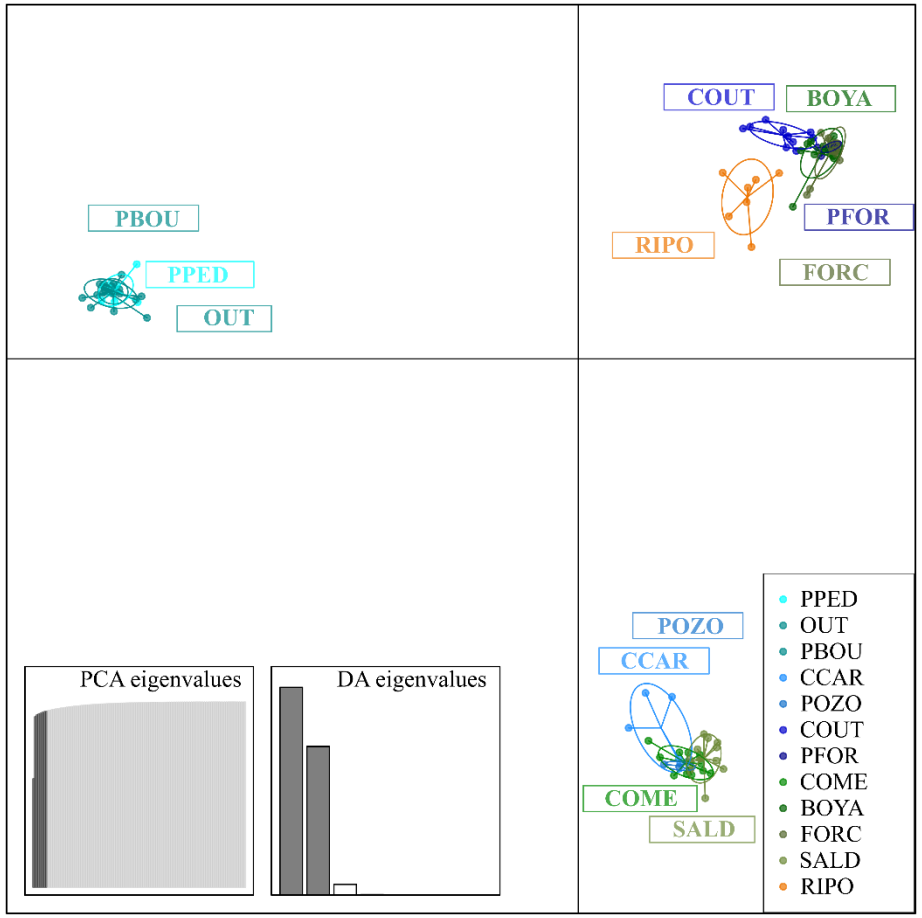
```r
#Selección de componentes principais

# 1A.PCaxes = kinfer - 1 PC criterion (Thia 2022)
genind_Am_N120_5675_DAPC_1a<-dapc(genind_Am_N120_5675, genind_Am_N120_5675$pop,n.pca = 6,var.contrib = TRUE, var.loadings=TRUE)
# 1B.PCaxes = kinfer - 1 PC criterion (Thia 2022)
genind_Am_N120_5675_DAPC_1b<-dapc(genind_Am_N120_5675, Find.cluster$grp,n.pca = 6,var.contrib = TRUE, var.loadings=TRUE)

#2. Criterio de varianza proprocional (e.g., >50%)
# 50% -> n.pca = 2
genind_Am_N120_5675_DAPC_2_50<-dapc(genind_Am_N120_5675, genind_Am_N120_5675$pop,n.pca = 2,var.contrib = TRUE, var.loadings=TRUE)


# 3. Validación cruzada
#   Función xvalDapc: Procedimiento para encontrar el punto perfecto entre retener demasiados PCs o demasiados pocos
xval <- xvalDapc(mat_Am_N120_5675, mat_Am_N120_5675_pop, n.pca.max = 300, training.set = 0.8, result = "groupMean",
                 center = TRUE, scale = FALSE, n.pca = NULL, n.rep = 100, xval.plot = TRUE, parallel = "multicore", ncpus = 12L)


#4. a-score
a_score <- optim.a.score(dapc,n.sim=1000) # I would recommend use n.sim = 1000
genind_Am_N120_5675_DAPC_a_score<-dapc(genind_Am_N120_5675, genind_Am_N120_5675$pop,n.pca = 12,var.contrib = TRUE, var.loadings=TRUE)
```

```
# ##############################################
# # Discriminant Analysis of Principal Components #
# ##############################################
# class: dapc
# $call: dapc.genind(x = genind_Am_N120_5675, pop = Find.cluster$grp,
#                    n.pca = 6, var.contrib = TRUE, var.loadings = TRUE)
#
# $n.pca: 6 first PCs of PCA used
# $n.da: 5 discriminant functions saved
# $var (proportion of conserved variance): 0.706
#
# $eig (eigenvalues): 1710 989.6 843.3 781.3 335.7 ...
#
# vector    length content
# 1 $eig      6      eigenvalues
# 2 $grp      120    prior group assignment
# 3 $prior    7      prior group probabilities
# 4 $assign   120    posterior group assignment
# 5 $pca.cent 11350  centring vector of PCA
# 6 $pca.norm 11350  scaling vector of PCA
# 7 $pca.eig  119    eigenvalues of PCA
#
# data.frame    nrow  ncol content
# 1 $tab          120   6    retained PCs of PCA
# 2 $means        7     6    group means
# 3 $loadings     6     5    loadings of variables
# 4 $ind.coord    120   5    coordinates of individuals (principal components)
# 5 $grp.coord    7     5    coordinates of groups
# 6 $posterior    120   7    posterior membership probabilities
# 7 $pca.loadings 11350 6    PCA loadings of original variables
# 8 $var.contr    11350 5    contribution of original variables
# 9 $var.load     11350 5    loadings of original variables
```

```
# Percentage of variation explained for each DA eigenvalue

genind_Am_N120_5675_DAPC_1b$eig[1]/sum(genind_Am_N120_5675_DAPC_1b$eig)
# 0.3598447
genind_Am_N120_5675_DAPC_1b$eig[2]/sum(genind_Am_N120_5675_DAPC_1b$eig)
# 0.2082802
genind_Am_N120_5675_DAPC_1b$eig[3]/sum(genind_Am_N120_5675_DAPC_1b$eig)
# 0.1774933
genind_Am_N120_5675_DAPC_1b$eig[4]/sum(genind_Am_N120_5675_DAPC_1b$eig)
#  0.1644386
genind_Am_N120_5675_DAPC_1b$eig[5]/sum(genind_Am_N120_5675_DAPC_1b$eig)
#  0.07065939
```

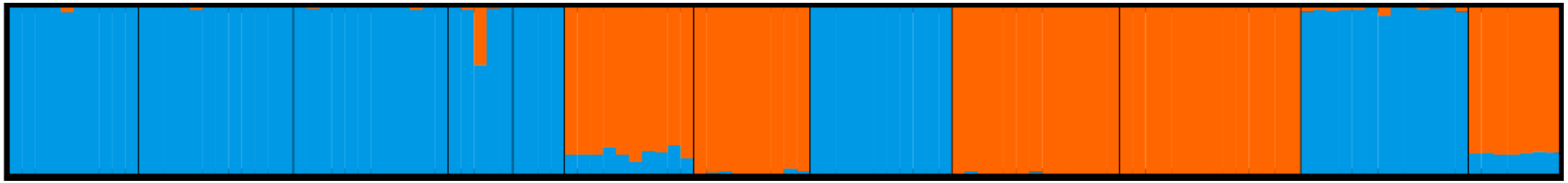





PBOU
PPED
OUT
COUT
BOYA
PFOR
RIPO
FORC
PCA eigenvalues
DA eigenvalues
POZO
CCAR
COME
SALD
PPED
OUT
PBOU
CCAR
POZO
COUT
PFOR
COME
BOYA
FORC
SALD
RIPO

- STRUCTURE
  - Análise bayesiana
  - Agrupa poboacións en clústeres
  - Representa cada individuo como barra vertical
  - Cores representando parte do xenoma
  - Asuncións de xenética de poboacións
    - Equilibrio de Hardy-Weinberg

K=2



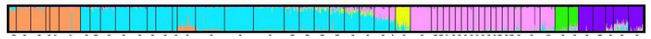
Structure Software

Pritchard Lab, Stanford University

Home | Research | Publications | Software | Data | Contact | Lab Members | Join Us

The program *structure* is a free software package for using multi-locus genotype data to investigate population structure. Its uses include inferring the presence of distinct populations, assigning individuals to populations, studying hybrid zones, identifying migrants and admixed individuals, and estimating population allele frequencies in situations where many individuals are migrants or admixed. It can be applied to most of the commonly-used genetic markers, including SNPS, microsatellites, RFLPs and AFLPs.

In 2016 John Novembre wrote a short historical perspective of Structure.



Download Structure 2.3.4.

fastSTRUCTURE for large SNP datasets is out now! Links to the preprint and software (beta release) by Anil, Matthew and Jonathan.

What to cite: The basic algorithm was described by Pritchard, Stephens & Donnelly (2000). Extensions to the method were published by Falush, Stephens and Pritchard (2003), and (2007) and Hubisz, Falush, Stephens and Pritchard (2009).

OPEN ACCESS Freely available online — PLOS ONE

ParallelStructure: A R Package to Distribute Parallel Runs of the Population Genetics Program STRUCTURE on Multi-Core Computers

Francois Besnier[1*], **Kevin A. Glover**[1,2]

1 Department of Population Genetics, Institute of Marine Research, Bergen, Norway, **2** Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Bergen, Bergen, Norway

```
#!/bin/bash
# this file is 5675_Amontana_N120.sh

#SBATCH -t 30:00:00
#SBATCH -c 50
#SBATCH --mem-per-cpu 1G
#SBATCH --mail-type=all
#SBATCH --mail-user=fernando.cabana@rai.usc.es

Rscript 5675_Amontana_N120.R
```

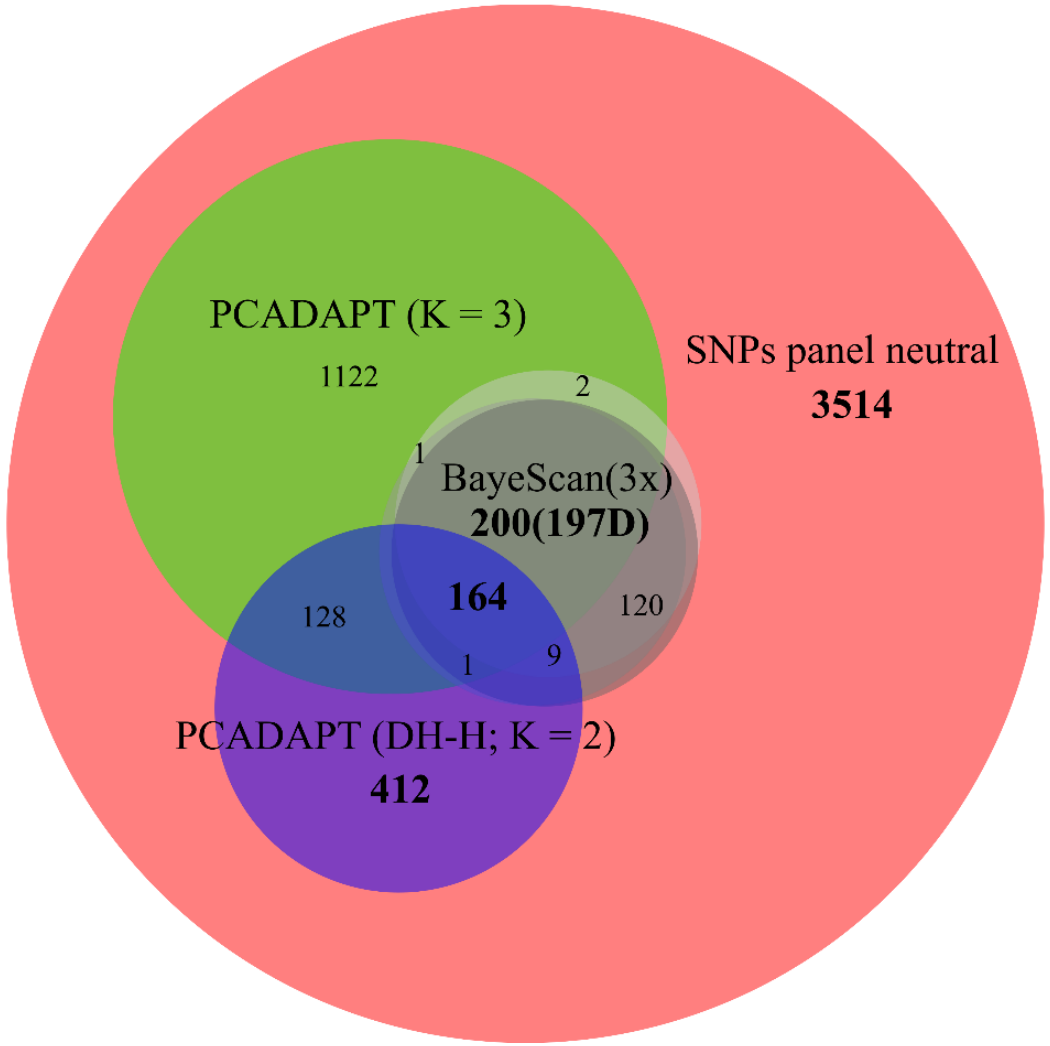
30th Anniversary CESGA GALICIA SUPERCOMPUTING CENTER

```
setwd("/mnt/lustre/scratch/nlsas/home/uvi/cursos/curso402/TFM/POPGEN/ARNICA_MONTANA/STRUCTURE_1/")

library("ParallelStructure")

parallel_structure(structure_path="/home/uvi/cursos/curso402/R/x86_64-pc-linux-gnu-library/console/",
                   joblist="20240416_5675_Amontana_N120_1_job", n_cpu=10, infile="./20240416_5675_Amontana_N120_1.stru",
                   numinds=120, numloci=5675, printqhat=1, plot_output=0,onerowperind=0,locprior=0,noadmix = 1,popflag = 0,
                   usepopinfo = 0,freqscorr = 0,markernames=1)
```

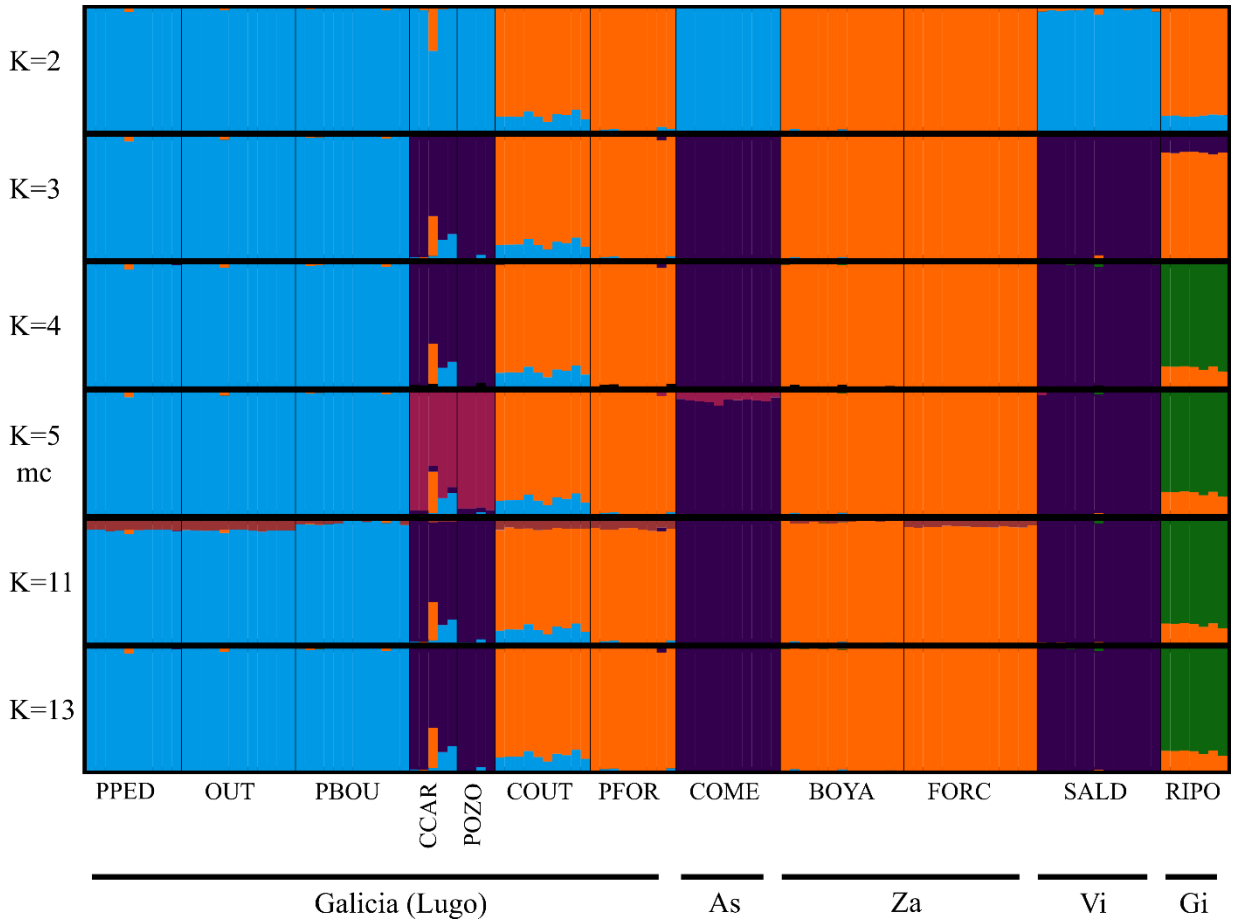
RESULTADOS

Resultado de análises de *outliers* (SNPs = 5,675)

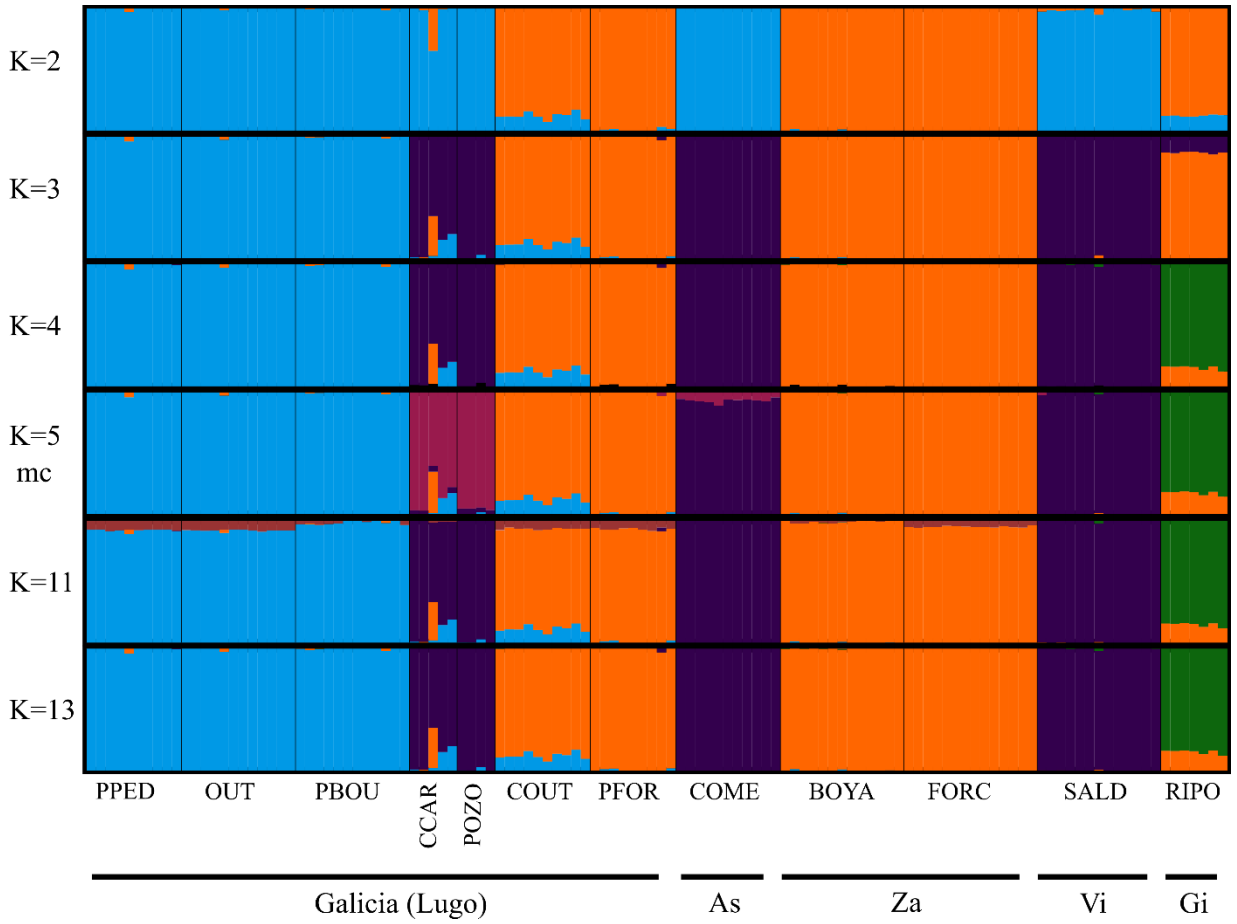
Paneis diverxente, bioquímico e neutral



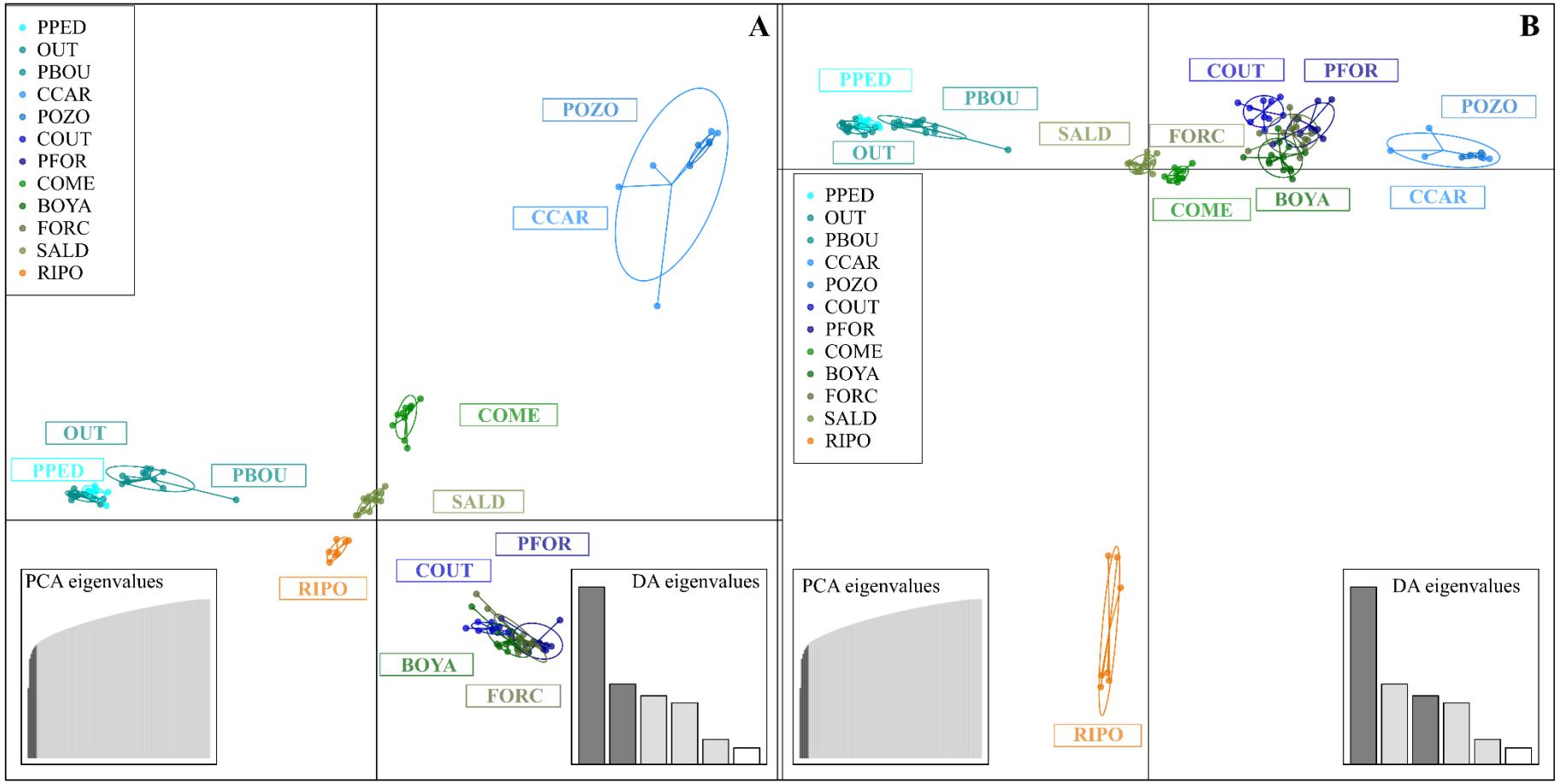
STRUCTURE 5,675 SNPs



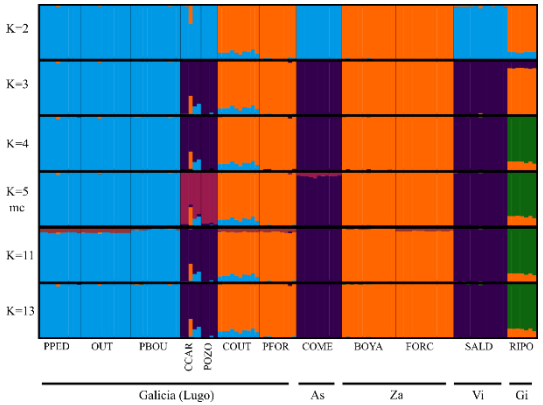
STRUCTURE 5,675 SNPs

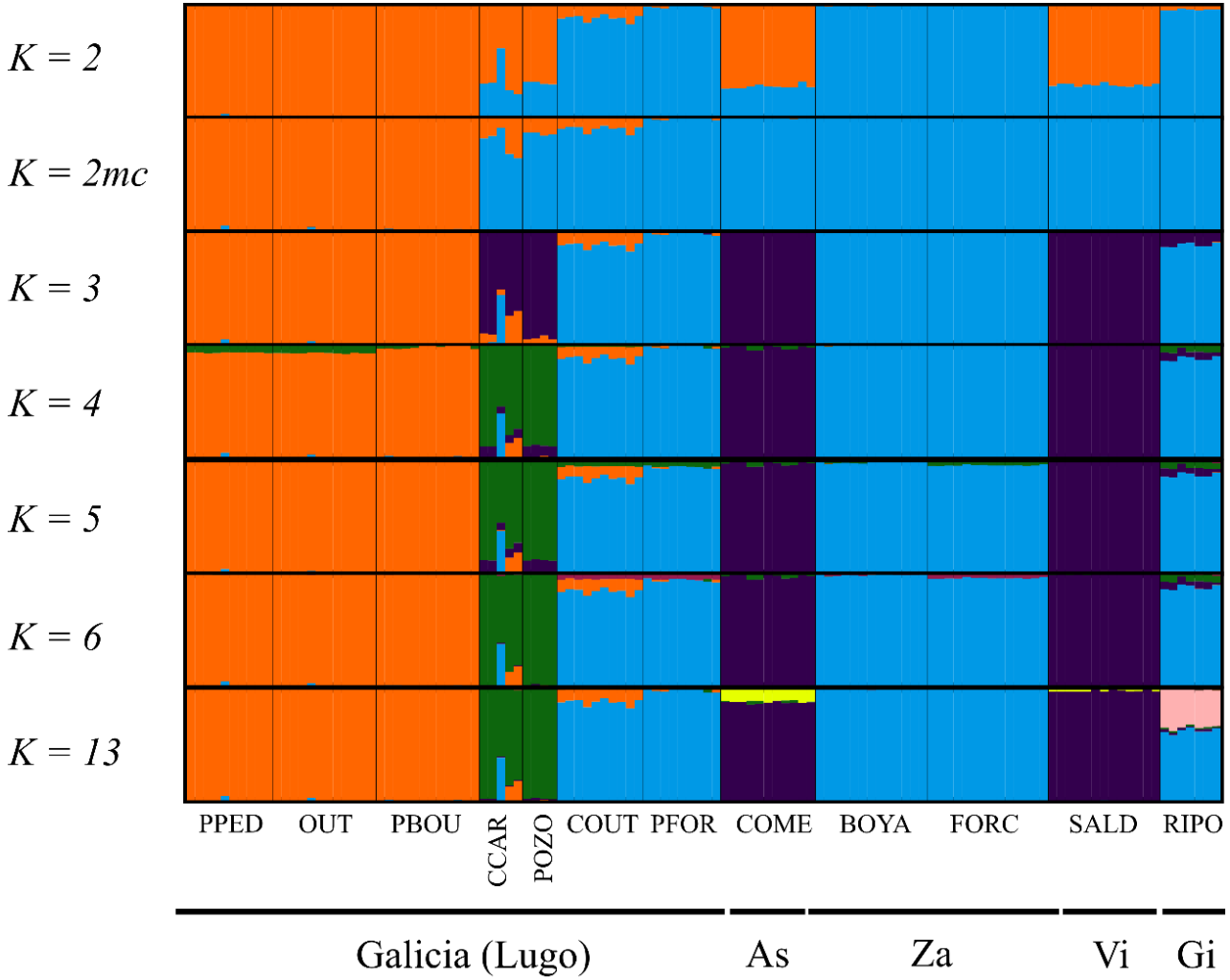


DAPC 5,675 SNPs

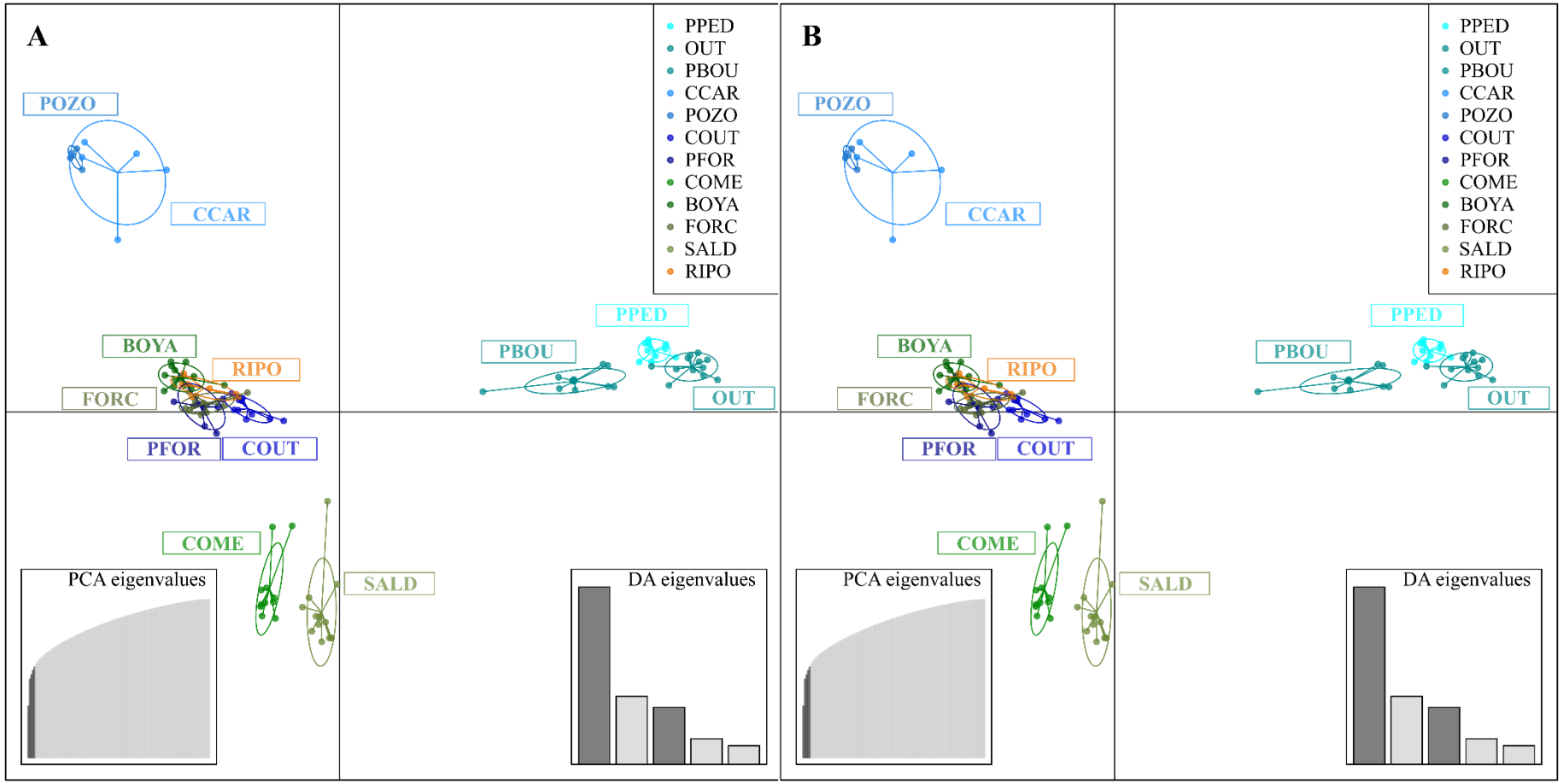


STRUCTURE 3,514 SNPs

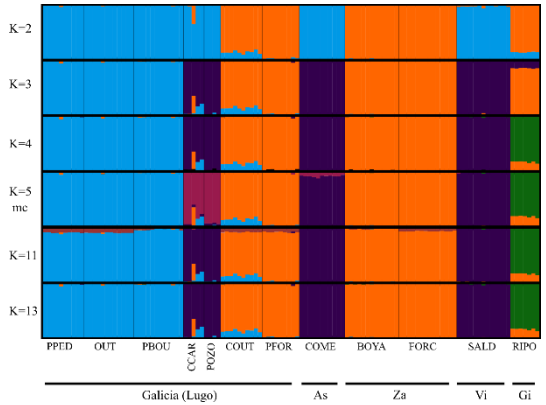


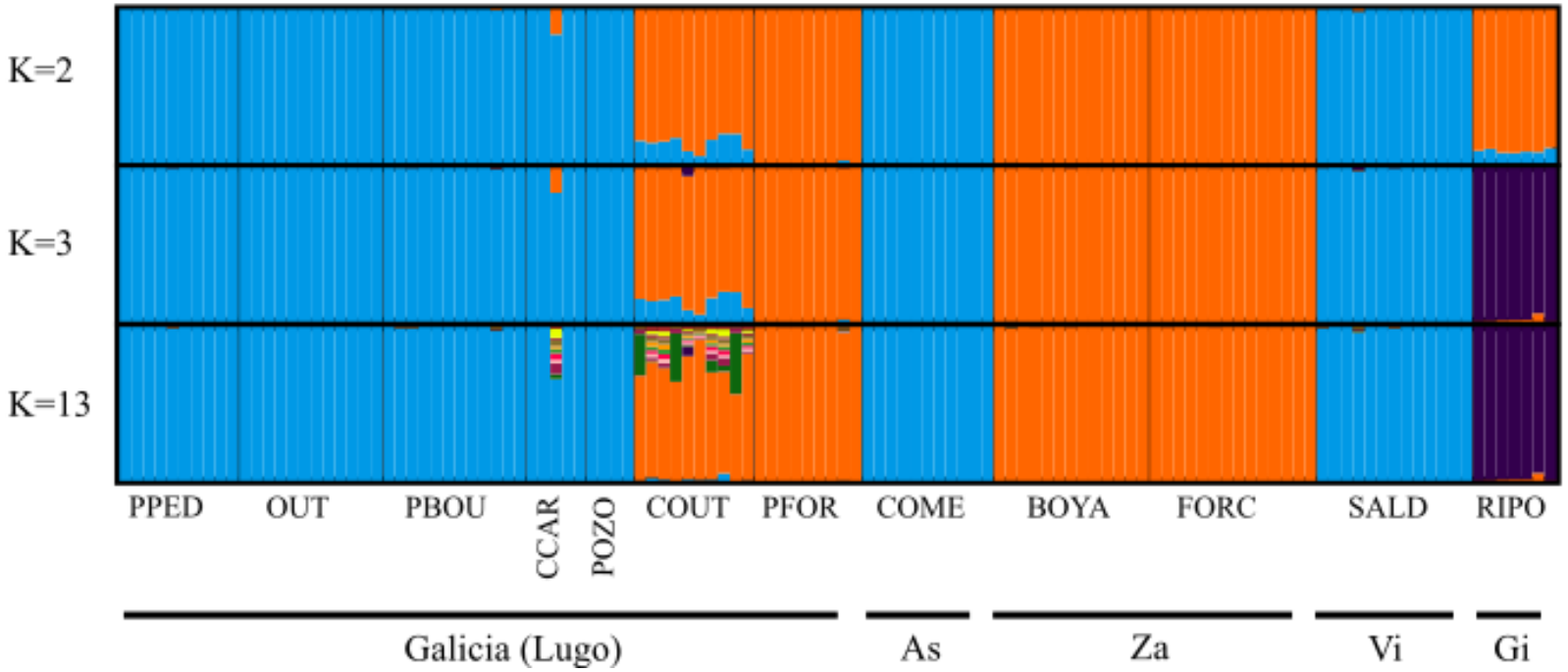


DAPC 3,514 SNPs

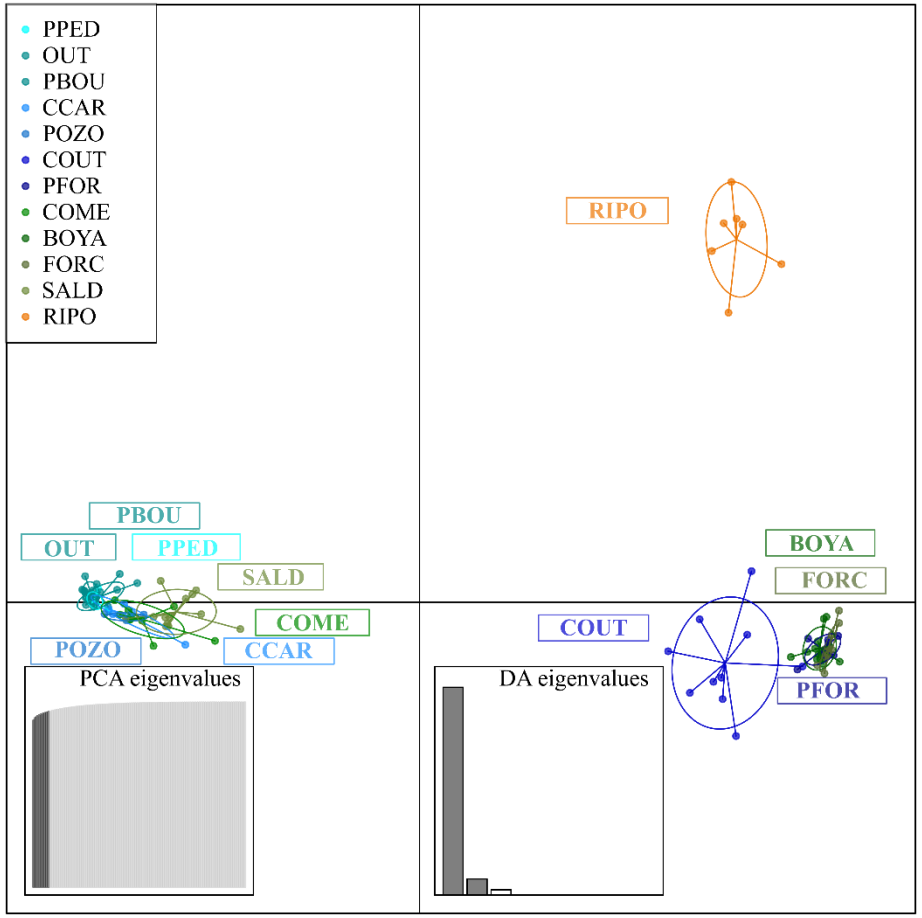


STRUCTURE 197 SNPs diverxentes

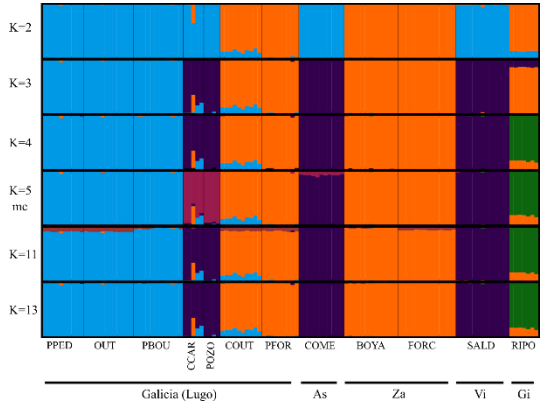




DAPC 197 SNPs diverxentes



STRUCTURE 412 SNPs

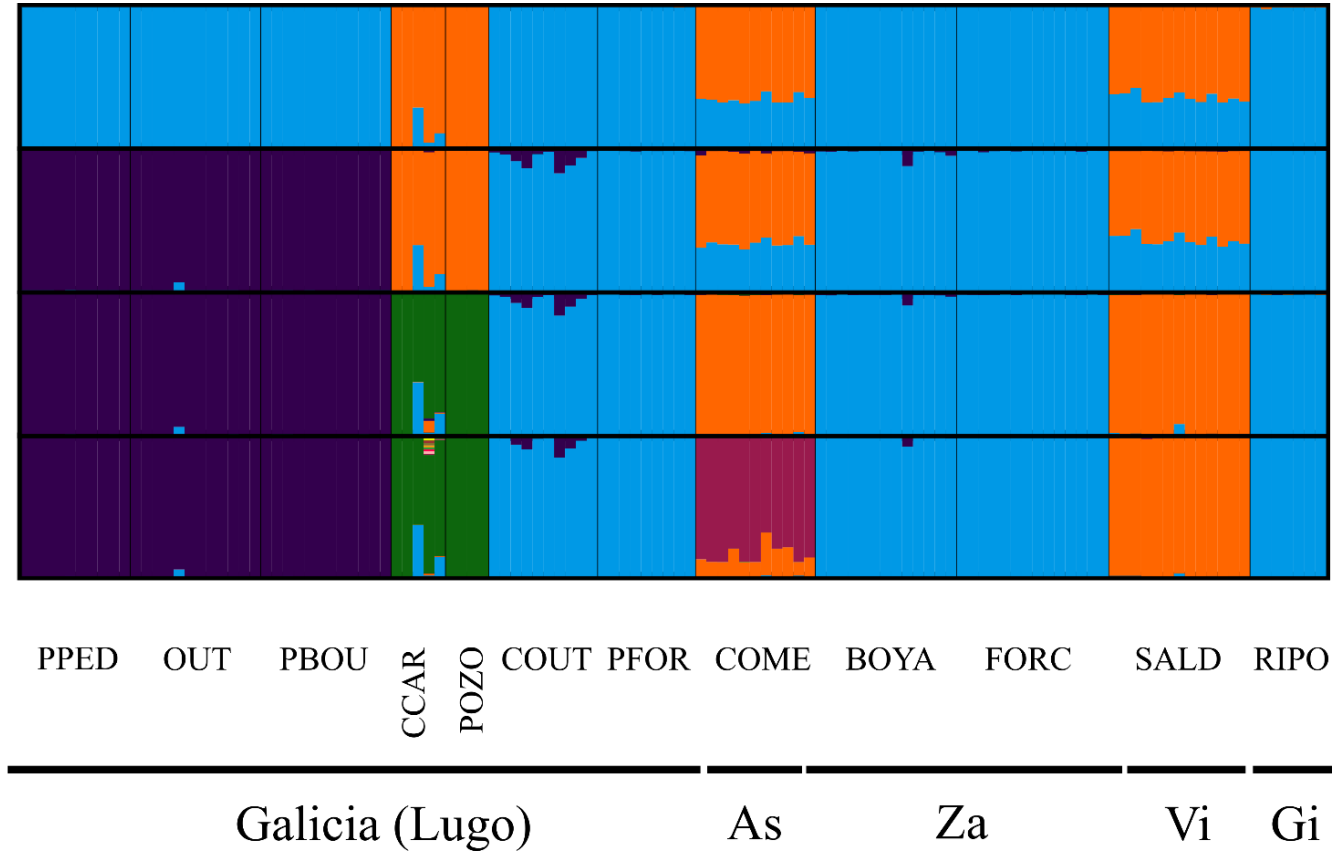


$K = 2$

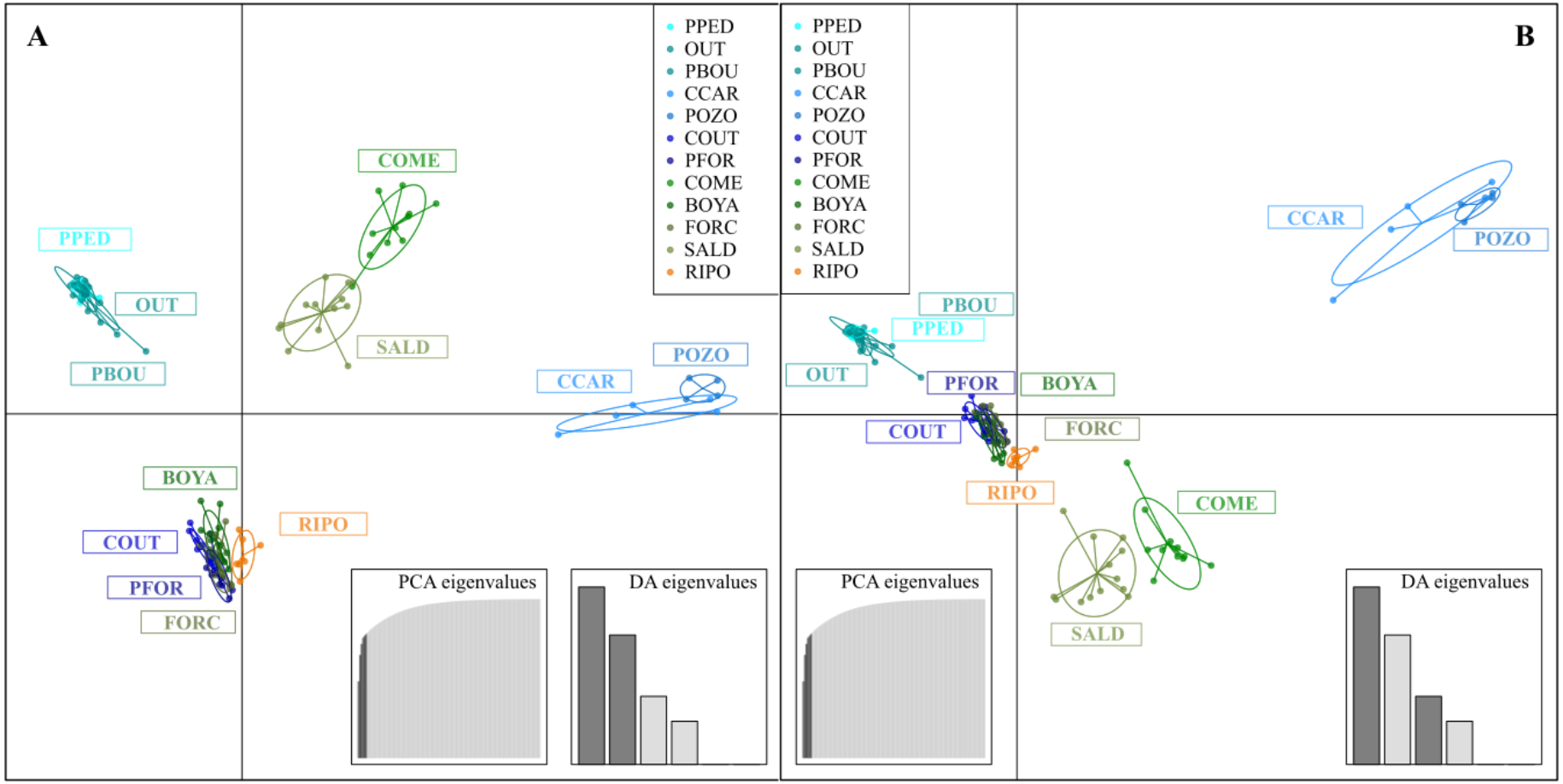
$K = 3$

$K = 4$

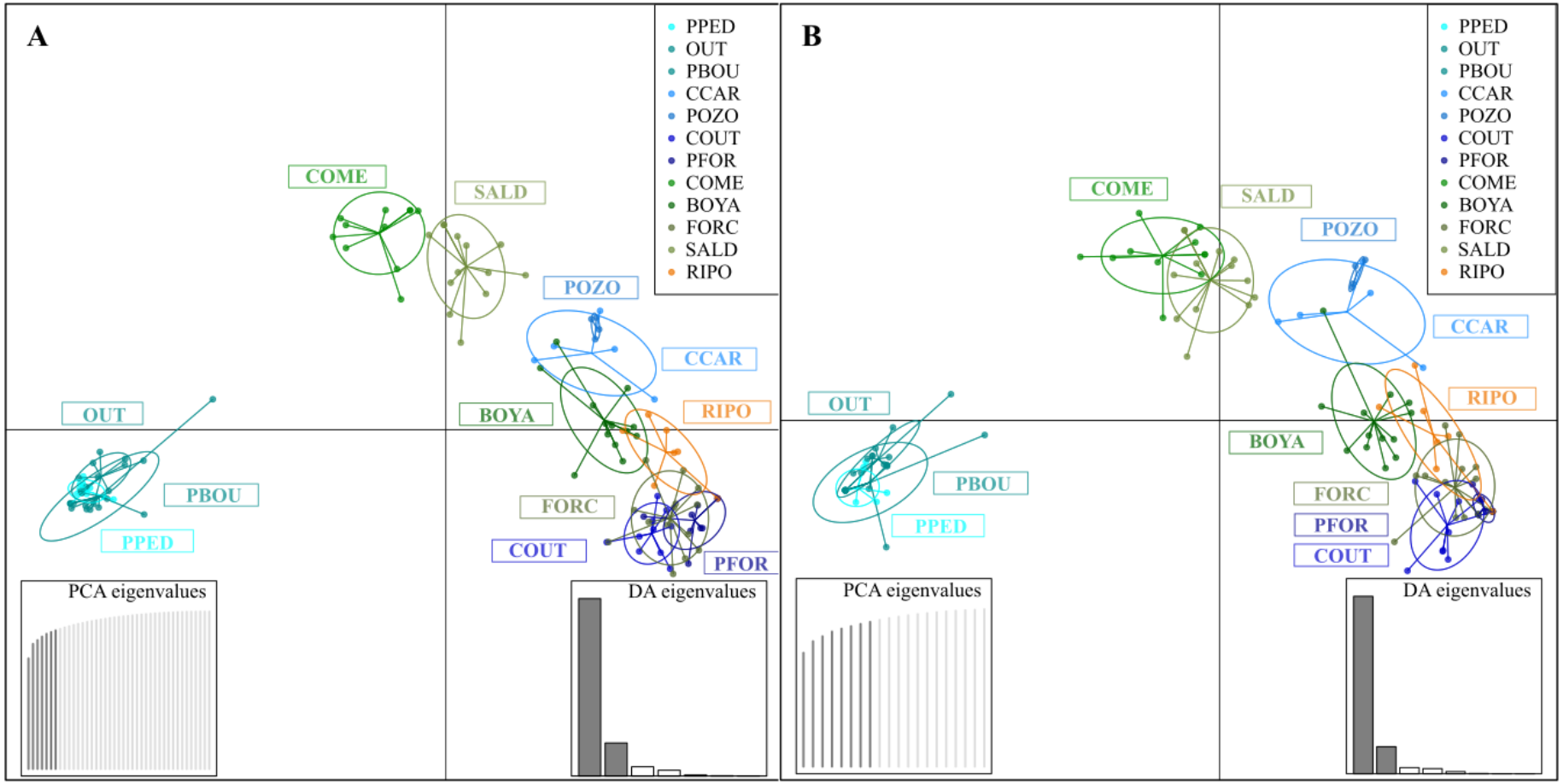
$K = 13$



DAPC 412 SNPs



DAPC 41-20 SNPs



CONCLUSIÓNS

- A través das análises de estrutura poboacional, podemos definir ata cinco posibles unidades de xestión (MUs, Management units).
- Atopáronse *outliers* candidatos a estar baixo selección, potencialmente implicados en adaptacións locais, e *outliers* potencialmente implicados na diferenciación de quimiotipos.
- Os resultados obtidos son consistentes con previos estudos xenéticos da especie.

XI XORNADA DE USUARIOS DE R EN GALICIA

24 de outubro de 2024

tade de Matemáticas (Campus Vida), Santiago de Compo

XI XORNADA DE USUARIOS DE EN GALICIA
24 de OUTUBRO de 2024
SANTIAGO DE COMPOSTELA
r-users.gal

CESGA
GALICIA SUPERCOMPUTING CENTER

DEPUTACIÓN DE LUGO

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

CAMPUS TERRA

XUNTA DE GALICIA | CONSELLERÍA DE EDUCACIÓN, CIENCIA, UNIVERSIDADES E FORMACIÓN PROFESIONAL