Specificity, rarity and core subsets with HCore

M. Humberto Reyes-Valdés

June 14, 2019

CABANA workshop

Genomic analysis of crop diversity using R

As a part of a diversity analysis in a crop collection, we can calculate the rarity of accessions, based on marker information. This can serve to pick and preserve rare crop materials. The basis to calculate accession rarity is allele specificity. The theory behind such definitions can be used for core subset selection. A core subset is a fraction of a collection, aimed to capture most of its genetic diversity.

The theory was developed by Reyes-Valdés et al. (2018): An informational view of accession rarity and allele specificity in germplasm banks for management and conservation in PLOS ONE.

The implementation HCoreA3.R, makes easy the numerical process to estimate parameters and construct the core subset.

To show the process, we will use the small sample we picked from the genotyped wheat collection.

Besides rarity, HCore calculates the average Kullback-Leibler divergence between the accessions and the pooled frequencies. In the same paper it is mathematically demonstrated that the average Kullback-Leibler divergence is the same as the average rarity in a set of accessions.

Load data

```
setwd("~/cursos/cabanaIrapuato/lectures")
dat<-read.csv("tables/sample.csv",head=T)
dat[1:6,1:6]
```

```
##   marker allele SEEDDIV2819 SEEDDIV2891 SEEDDIV2899 SEEDDIV2907
## 1      3      1           0         1.0         0.5           1
## 2      3      2           1         0.0         0.5           0
## 3      6      1           1         0.5         0.0           0
## 4      6      2           0         0.5         1.0           1
## 5     19      1          NA         1.0         0.0           1
## 6     19      2          NA         0.0         1.0           0
```

Load HCore

```
source("HCoreA3.R")
```

Run HCore and assign the object to a variable

```
x<-s.biallelic(dat)
names(x)
```

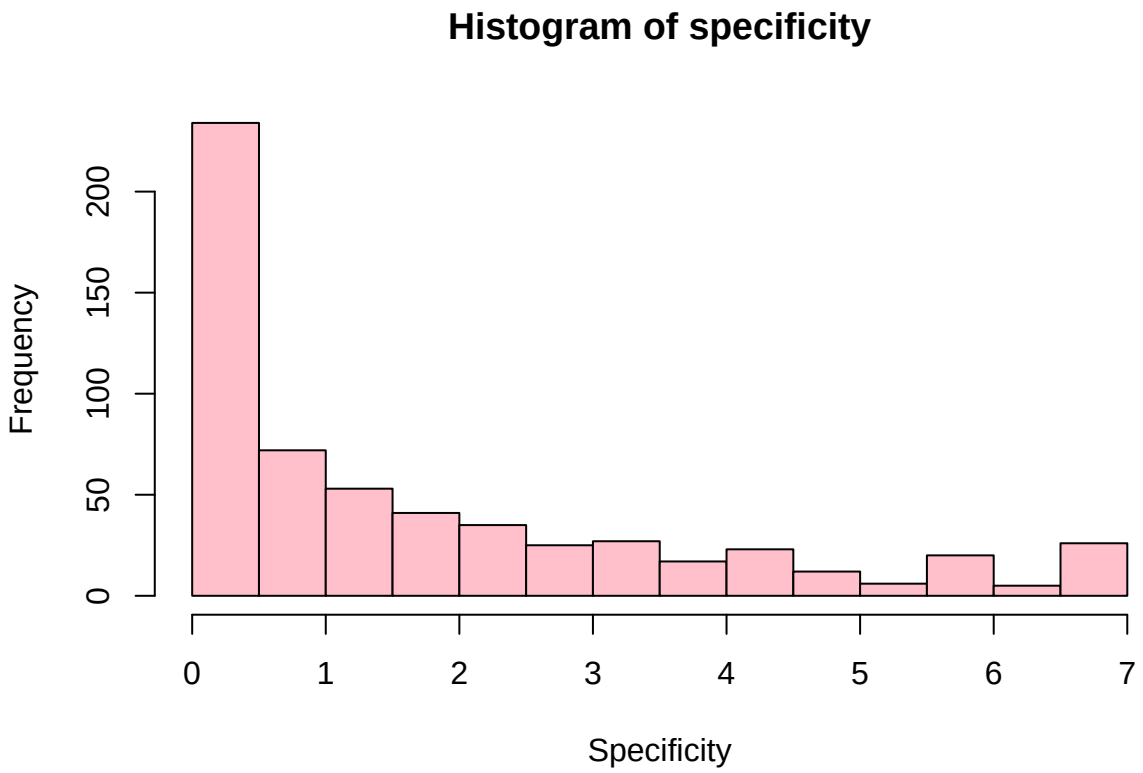
```
## [1] "specificity" "rareness"    "divergence"  "table"
```

Allele specificity

```
head(x$specificity)
```

```
## [1] 1.1098237 0.8227853 0.9088120 0.9923970 0.8811655 1.0350183
```

```
hist(x$specificity, main="Histogram of specificity",xlab="Specificity",col="pink")
```



We can see that the distribution is skewed towards alleles with low specificity.

Rarity and divergence

```
head(x$table)
```

```
##           pop  rareness divergence
## 1 SEEDDIV2819 0.4921676  0.5131915
## 2 SEEDDIV2891 0.4245511  0.4331133
## 3 SEEDDIV2899 0.5208033  0.5351608
## 4 SEEDDIV2907 0.5271215  0.5374895
## 5 SEEDDIV2827 0.6043301  0.6352046
## 6 SEEDDIV2835 0.4609223  0.4806979
```

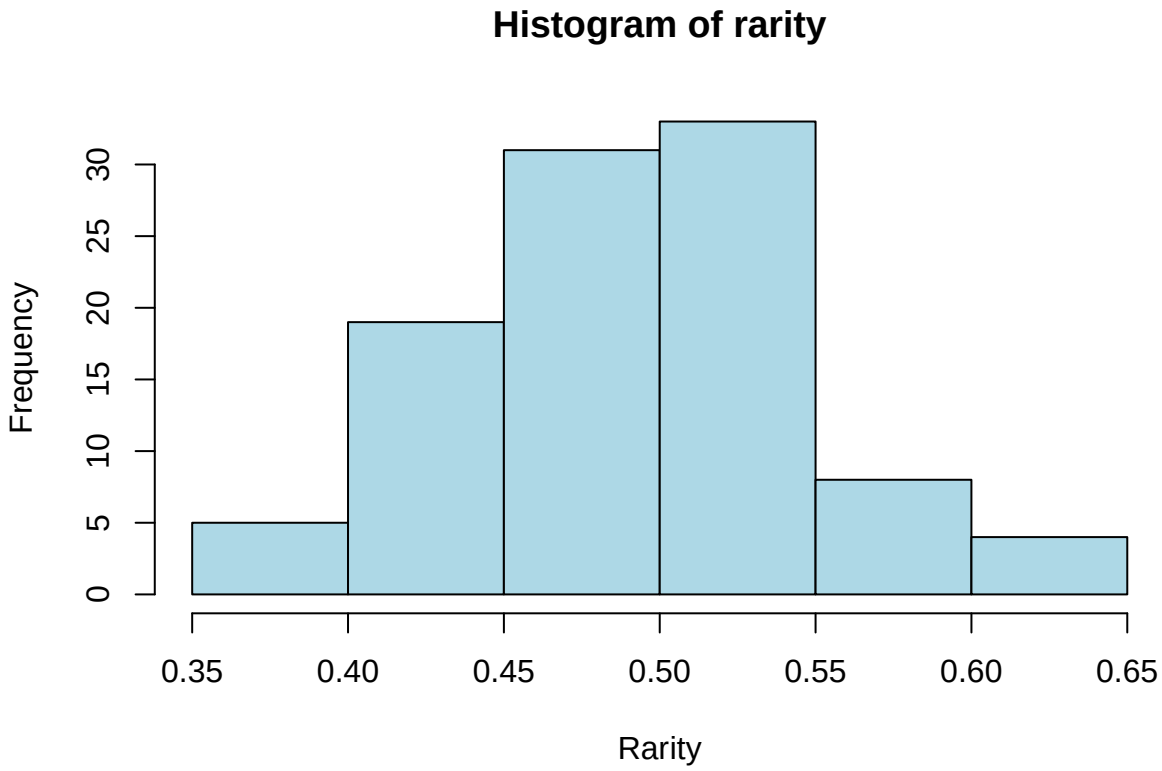
```
library(dplyr) #A package to handle tables
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
hist(x$rareness,main="Histogram of rarity",xlab="Rarity",col="light blue")
```




```
#The 10 rarest accessions
head(arrange(x$table,desc(rareness)),10)
```

```
##            pop  rareness divergence
## 1  SEEDDIV3002 0.6075625  0.6310273
## 2  SEEDDIV2827 0.6043301  0.6352046
## 3  SEEDDIV2828 0.6034884  0.6249794
## 4  SEEDDIV2820 0.6021141  0.6149766
## 5  SEEDDIV2853 0.5931663  0.6118790
## 6  SEEDDIV2922 0.5887970  0.6126360
## 7  SEEDDIV2852 0.5885105  0.5232233
## 8  SEEDDIV2830 0.5861748  0.6169671
## 9  SEEDDIV2861 0.5765573  0.5997430
## 10 SEEDDIV2900 0.5658469  0.5231352
```

Core subset selection

```
#Number of accessions
dim(dat)[2]-2
```

```
## [1] 100
```

```
#Obtain a minicore of 20 accessions
core<-h.coreA(dat,20)
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
```

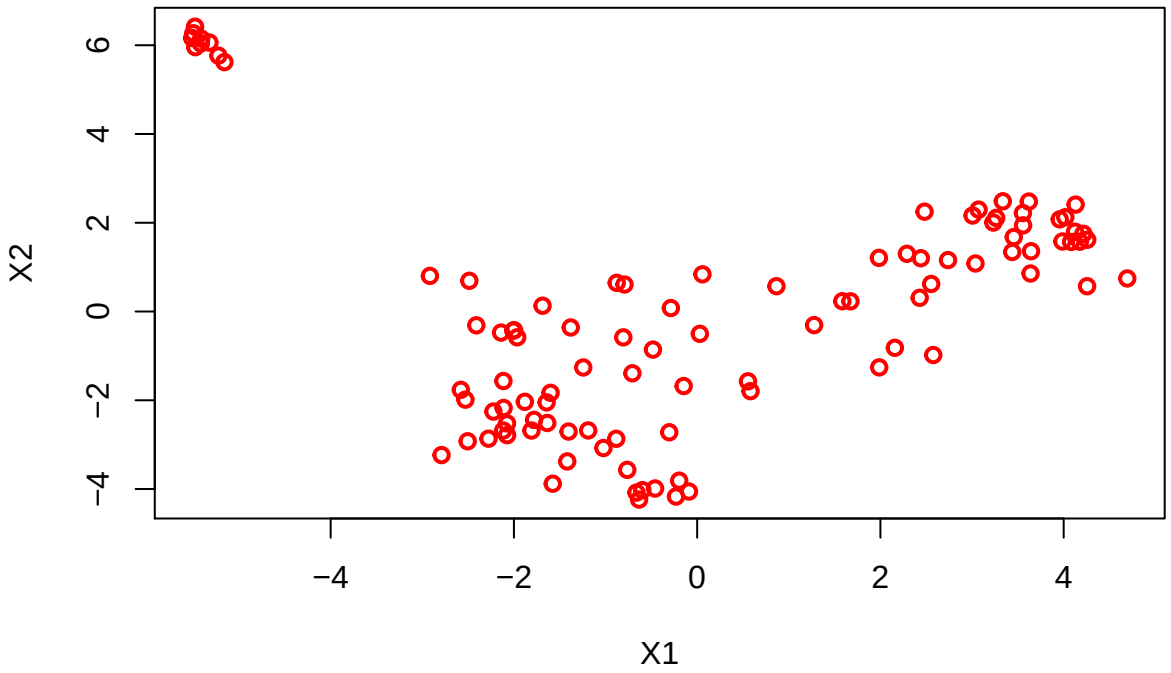
```
core
```

```
##  [1] "SEEDDIV2827" "SEEDDIV2846" "SEEDDIV2861" "SEEDDIV2946" "SEEDDIV2953"
##  [6] "SEEDDIV2850" "SEEDDIV2820" "SEEDDIV2876" "SEEDDIV2824" "SEEDDIV2829"
## [11] "SEEDDIV2830" "SEEDDIV2922" "SEEDDIV2899" "SEEDDIV2907" "SEEDDIV2896"
## [16] "SEEDDIV2993" "SEEDDIV2882" "SEEDDIV3002" "SEEDDIV2884" "SEEDDIV2913"
```

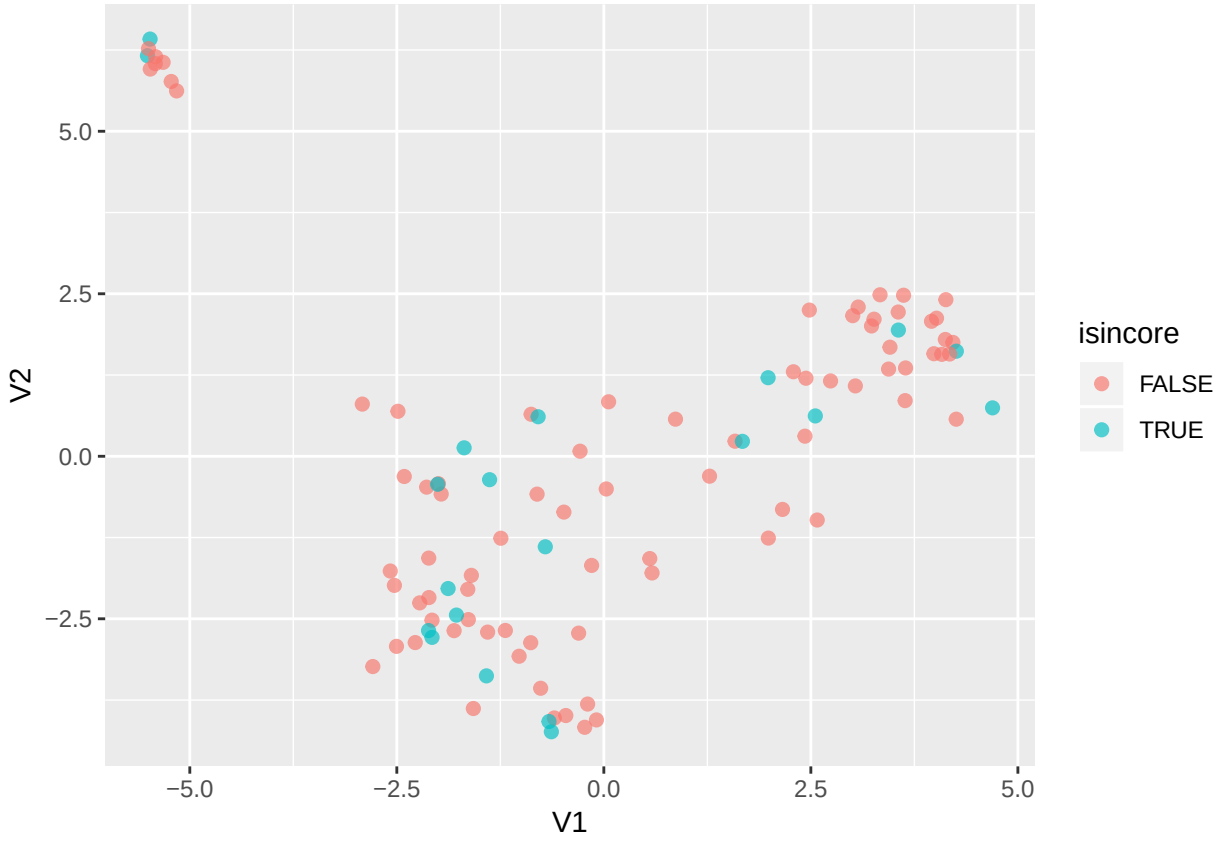
```
#Visualize in MDS
#Multidimensional scaling
#We don't need imputation
for.pc<-dat[-c(1,2)]
for.pc<-t(for.pc)
d<-dist(for.pc)
fit<-cmdscale(d,eig=T,k=2)
plot(fit$points[,1],fit$points[,2],col="red",lwd=2, xlab="X1",ylab="X2")
#Colorize
x<-as.data.frame(fit$points)
head(x)
```

```
##                      V1         V2
## SEEDDIV2819 -2.1145987 -1.5661560
## SEEDDIV2891  0.5547715 -1.5734630
## SEEDDIV2899 -5.5113300  6.1622303
## SEEDDIV2907 -1.8806391 -2.0344407
## SEEDDIV2827 -0.7068930 -1.3922921
## SEEDDIV2835 -2.1401077 -0.4747442
```

```
isincore<-is.element(row.names(x),core) #Core indicator vector
x<-mutate(x, isincore=isincore)
library(ggplot2)
```




```
ggplot(x, aes(V1,V2,color=isincore))+geom_point(alpha=2/3,size=2)
```



$$\left(\hat{\mathrm{O}}_{\triangledown}\hat{\mathrm{O}}\right)$$
$$\left(\left[\begin{array}{c}\mathrm{vvv}\\ \mathrm{vvv}\end{array}\right]\right)$$
$$\lambda\frown\lambda$$
Humberto Reyes