Core collections - Computer lab

Plant Genetic Resources (3502-470)

Prof. Karl Schmid, University of Hohenheim

15 May 2022

Background

The pipe operator %>% allows to create analysis pipelines. Look up the help function to learn more about it.

Core collections are small subsets of larger *ex situ* collections of accessions that aim to represent the diversity present in the complete collection.

The purpose of this computer lab is to construct a core collection using molecular markers and to compare different approaches for the construction of a core collection.

We use a real data set of 1,311 maize landraces (and wild ancestors) genotyped with 983 SNP markers (Heerwaarden et al. 2011) that will be our *collection* (we use a reduced marker set compared to the original dataset). Using the SNP data, we construct a core collection using the 'H' method. Accessions are grouped by subspecies and geographic origin (as in van Heerwaarden et al., see figure above). We want to construct a core collection of approx. $n = 100$ accessions.

Construction of the collection with the H method

The 'H' method measures genetic diversity, h_i, within each group i. Then, a weight

$$w_i = \frac{h_i}{1 - h_i}$$

is assigned to each group

From each group,

$$n \times \frac{w_i}{\sum_i w_i}$$

accessions (relative to the weight of the group) are picked at random to form the core collection. These numbers of accessions are non-integer and will be rounded up to the next integer. In the following, a core collection $n = 100$ accessions will be constructed.

Genetic diversity is measured with Nei's gene diversity, which is the mean expected heterozygosity across SNPs under the assumption of a Hardy-Weinberg-Equilibrium. The functions for calculating diversity are included in the pegas package.

```
library(tidyverse) # for dealing with data frames and plotting
library(pegas)
library(adegenet)
```

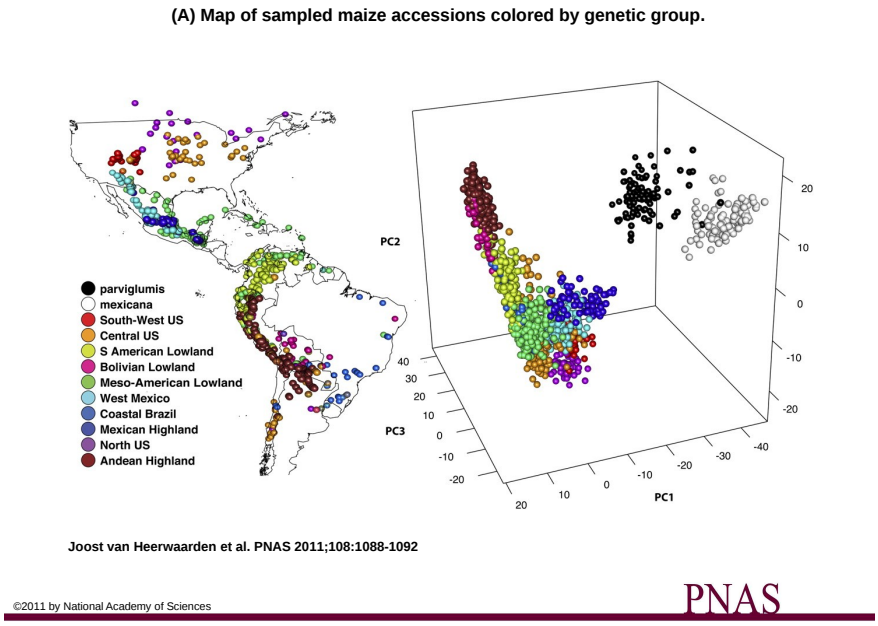

Figure 1: Groups of maize and teosinte accessions. Source: (Heerwaarden et al. 2011)

To fully understand what the functions are doing in the following, it is always a good idea to check the help function in the console, e.g., `?df2genind`

Read data into a data frame and transform it into a `genind` object.

```
df <- read.table("~/pgr/data/maize_SNPdata.txt", sep = "\t", header = TRUE, row.names = 1)
snpdata <- df2genind(df[, -1], pop = df[, 1], sep = "/")
```

Next, we investigate the grouping of the accessions. How many subpopulations are there and how large are they?

```
pop_ind <- table(snpdata@pop)
names(pop_ind)   # names of subpopulations
length(pop_ind)  # numbers of accessions in subpopulations
pops <- tibble("population" = names(pop_ind), "accessions" = pop_ind) # construct a population summary
pops
```

`tibble` is a nice function for constructing data frames. The name of columns can be easily defined as shown in the example.

To calculate average heterozygosity of each group with command `Hs` from the `adegenet` package, the data needs to be converted into the `genpop` format (check with `?genpop` for more information).

```
snpgenpop <- genind2genpop(snpdata)
h <- Hs(snpgenpop) # calculates the average heterozygosity (= Nei's gene diversity) of each subpopulati
h
```

Next, calculate weights for each population based on heterozygosity

```
weights<- h/(1-h)
weights
```

rrr

Calculate weighted number of accessions to select from each subpopulation

```
coresize <- 100  # size of core collection
acc_num <- coresize * weights/ sum(weights) # calculates the number of accessions
                                            # of each subpopulation
acc_num <- ceiling(acc_num)  # round numbers to integers
coresize <- sum(acc_num)     # sum of accessions for core collection
acc_num
```

We have now calculated how many accessions should be sampled randomly from each subgroup and we will proceed to sampling these accessions.

```
x <- vector("list", length(pop_ind)) # list for subpopulation genind objects

# separate subpopulations for sampling
subpops <- seppop(snpdata)

# sample according heterozygosity weight, drop is to account for alleles not being sampled
for (i in 1:length(pop_ind)) {
  x[[i]] <- subpops[[i]][sample(pop_ind[[i]], acc_num[i]), drop = TRUE]
  }

# combine genotypes into a single dataset (the core collection)
corecoll <- repool(x)
```

To see whether our structured approach to construct a core collection brings any advantage compared to a random sampling, we make a random core collection of comparable size.

```
randcoll <- snpdata[sample(1:1311, coresize), drop = TRUE]
```

Differences between the original and the core collection

We now investigate how the core collection differs from the original collection based on the following parameters:

- Mean expected heterozygosity (i.e., Nei's gene diversity)
- Number of polymorphic SNPs (i.e., how much diversity)
- Mean allele frequency
- Distribution of minor allele frequencies over SNPs

Mean expected heterozygosity

First use the summary function for the `genind` object to calculate the population-wide heterozygosity per locus.

```r
# Use the summary for genind object to generate population-wide heterozygosity per locus
h_orig <- mean(summary(snpdata)$Hexp)
h_cc <- mean(summary(corecoll)$Hexp)
h_rc <- mean(summary(randcoll)$Hexp)
```

The differences should be visualised to allow to answer the question whether the differences are real, or not.

par(mfrow=c(1,2))

barplot(c(h_orig, h_cc, h_rc), names.arg = c("All", "H strategy", "Random"),

las = 2, main = "Mean exp. heterozygosity")

Count number of polymorphic SNPs in each collection

poly_orig <- length(which(isPoly(snpdata, thres = 0))) poly_cc <- length(which(isPoly(corecoll, thres = 0))) poly_rc <- length(which(isPoly(randcoll, thres = 0)))

barplot(c(poly_orig, poly_cc, poly_rc), names.arg = c("All", "H strategy", "Random"), las = 2, main = "Number of polymorphic SNPs") par(mfrow = c(1,1))

```
### Exercises:

* Interprete your results. Can the heterozygosity of a core collection be higher than of the complete c
* Test whether core collections with $n = 50$ have a smaller diversity than with $n = 100$.

## Allele frequencies in core vs. complete collection

To further see what has changed from original to core collection, we look at a histogram of the minor a
```

```r
origloci <- genind2loci(snpdata)
coreloci <- genind2loci(corecoll)
randloci <- genind2loci(randcoll)

# we define a function to extract the minar allele frequency
maf_f <- function(x){
min(x$allele[names(x$allele)!="NA"])/sum(x$allele[names(x$allele)!="NA"])
}

# plot as density distribution
plot(density(sapply(summary(origloci),maf_f), from = 0, to = 1),
     main = "",xlab = "Allele frequencies")
points(density(sapply(summary(coreloci),maf_f), from = 0, to = 1), col = "red", type = "l")
points(density(sapply(summary(randloci),maf_f), from = 0, to = 1), col = "blue", type = "l")
legend("topright",legend=c("All","H strategy","Random"),col=c("black","red","blue"),lty=1,bty="n")
```

Discussion questions

1. Is the shape of the allele frequencies (site frequency spectrum) expected from a neutral model?

2. Are allele frequency distributions of the core collections significantly different from the complete sample? What does a comparison reveal about the population structure of the maize samples?

OPTIONAL

- Construct a third core collection in which the number of accessions is selected from each subpopulation corresponds to the percentage of accessions in the complete set of accessions.
- Is the new collection expected to have a higher or lower diversity than the above collections? Calculate the diversity levels and compare with the other methods.

Other programs and R packages to build core collections

- Core hunter (available at www.corehunter.org (Beukelaer et al. 2012)) screens many, but not all, possible subsets of a collection of a given size to find the subset that maximises either genetic diversity or allelic richness (or a composite measure thereof). It does not sample from each group, but samples directly from the complete collection without clustering.
- R package ccChooser: Implements several possibilities to build a core collection based on phenotypic data. It uses clustering and then selecting accessions from each cluster by different strategies.

References

Beukelaer, Herman De, Petr Smýkal, Guy F. Davenport, and Veerle Fack. 2012. "Core Hunter II: Fast Core Subset Selection Based on Multiple Genetic Diversity Measures Using Mixed Replica Search." *BMC Bioinformatics* 13: 312. https://doi.org/10.1186/1471-2105-13-312.

Heerwaarden, Joost van, John Doebley, William H. Briggs, Jeffrey C. Glaubitz, Major M. Goodman, Jose de Jesus Sanchez Gonzalez, and Jeffrey Ross-Ibarra. 2011. "Genetic Signals of Origin, Spread, and Introgression in a Large Sample of Maize Landraces." *Proceedings of the National Academy of Sciences* 108 (3): 1088–92. https://doi.org/10.1073/pnas.1013011108.