Optimal Sampling

Aims

1. Determination of optimal sampling for accurate estimates of genetic diversity
2. Determination of optimal number of loci for accurate estimates of genetic diversity

Description

- One script for He, Ho, Ar, Allele detection (calculation & pdf output).
- One script for FST, Jost's D, Cavalli-Sforza chord distance (calculation & pdf output).
- Two more scripts containing code to produce compact pdf outputs, displaying statistics only for sample sizes of 25, 30, 50, 75 individuals.

Before running the scripts

- The scripts' input file is a GenAlEx formatted excel sheet just like the one SFI, ASP and AUTH share for the common LGM data set. Because loci rank is determined based on the pooled empirical data set, a new excel sheet has to be created containing all the loci intended to be analyzed. If a locus has not been genotyped for all populations, it should be added as missing data. Note that the only statistic that is calculated from this pooled data set is loci rank. All other statistics are calculated on per population basis.

16		2551	9	250	200	400	251	200	400	250	200	400						
				DE_A	DE_R	DE_S	SL_A	SL_R	SL_S	GR_A	GR_R	GR_Seed						
Sample	Population	sfc0036	FS1_15	csolf31	csolf19	EEU75	DUKCT	EJV8T	EMILY	ERHBI	DZ447	DE567	sfc1143	csolfg05	csolfg06	csolfg29	concat14	
DE_I_FSY_A_007	DE_Adult	97 97	108 110	112 112	170 174	93 93	85 87	143 143	142 148	161 161	187 189	211 226	0 0	0 0	0 0	0 0	0 0	
DE_I_FSY_A_008	DE_Adult	97 103	118 118	108 116	172 174	93 107	87 87	143 143	146 148	159 161	189 189	226 226	0 0	0 0	0 0	0 0	0 0	
DE_I_FSY_A_010	DE_Adult	97 91	110 118	116 116	164 174	97 101	85 93	143 151	142 142	163 167	189 189	217 226	0 0	0 0	0 0	0 0	0 0	
DE_I_FSY_A_011	DE_Adult	97 103	118 118	122 126	166 170	107 107	85 85	143 143	146 150	161 161	187 191	211 211	0 0	0 0	0 0	0 0	0 0	
DE_I_FSY_A_015	DE_Adult	97 101	108 120	112 114	172 174	91 93	85 93	143 143	144 150	161 161	191 193	211 226	0 0	0 0	0 0	0 0	0 0	

Sheet tabs: Abies | Fagus | **Fagus-extra-markers** ⊕

- Note that population names are hardcoded and should be **exactly** the following in the GenAlEx input file: **SL_Adult, SL_Regen, SL_Seed, DE_Adult, DE_Regen, DE_Seed, GR_Adult, GR_Regen, GR_Seed**

- Note that the number of individuals in each population is hardcoded as following:
  - "Abies_DE_Adult", "Abies_GR_Adult", "Fagus_DE_Adult", "Fagus_GR_Adult": **250** trees
  - "Abies_SL_Adult": **249** trees
  - "Fagus_SL_Adult": **251** trees
  - "Abies_DE_Regen", "Abies_GR_Regen", "Abies_SL_Regen","Fagus_DE_Regen", "Fagus_GR_Regen", "Fagus_SL_Regen": **200** trees
  - "Abies_DE_Seed","Abies_SL_Seed","Fagus_DE_Seed", "Fagus_GR_Seed", "Fagus_SL_Seed": **400** trees
  - "Abies_GR_Seed": **382** trees

Please notify for any discrepancy.

Using the script

- The scripts has been tested to work on R 3.5.1 . The following guide assumes you are using the Rstudio application.

- The scripts has to be run individually for each population.

- To run the commands, place the cursor on line and press Ctrl + Enter
  - open rstudio by double clicking on

- place cursor on first library line and press Ctrl + Enter
- load excel input file: set the name of excel file and name of the sheet where the genotypes reside and press Ctrl + Enter
- set species name
- set population name
- set replication number

- Placing the cursor on a line, excecutes the code in this line

- Warnings (include images)

```
Console   Terminal ×
~/Optimal_Sampling_simulation_data/ ⇔
> # Load dataset
> # A GenAlEx formatted excel sheet is the required input
> obj <- read.genalexcel(
+     "LGM_DE_SI_GR_final.xlsx",     # name of excel file
+     sheet = "Abies",               # name of sheet where the genotypes reside
+     genclone = F)
warning message:
In df2genind(gena2, sep = "/", ind.names = ind.vec, pop = pop.vec,  :
  entirely non-type individual(s) deleted
>
```

```
Console   Terminal ×
~/Optimal_Sampling_simulation_
Joining, by = "samp_siz
Joining, by = "samp_siz
Joining, by = "samp_siz
Joining, by = "samp_siz
Joining, by = "samp_siz
Joining, by = "samp_siz
   user  system elapsed
  50.61    0.09   52.28
There were 50 or more w
>
```

```
Console   Terminal ×
~/Optimal_Sampling_simulation_data/ ⇔
  @tab does not contain integers; as of adegenet_2.0-0, numeric values are no longer used
47: In validityMethod(object) :
  @tab does not contain integers; as of adegenet_2.0-0, numeric values are no longer used
48: In validityMethod(object) :
  @tab does not contain integers; as of adegenet_2.0-0, numeric values are no longer used
49: In validityMethod(object) :
  @tab does not contain integers; as of adegenet_2.0-0, numeric values are no longer used
50: In validityMethod(object) :
  @tab does not contain integers; as of adegenet_2.0-0, numeric values are no longer used
>
```