Data_cleanup

Grace Saville

11/02/2022

Preamble

```
library(plyr)
getwd()
setwd("C:/Users/airhe/OneDrive/Documents/Masters/Project 3/kiore-project")
```

Loading the data

```
data <- read.delim("./data/Genotyping-007.010-01_SNP_Raw_data.tsv")
dim(data) #478 rows (specimens), 333 columns (SNP loci)
data[1,1:17] # SNP data in columns 17 to 333
```

```
class(data[5,17]) # character
count(data$island.1) #  how many samples from each island there are
count(data[,206]) # how many of each base combination there are in a SNP column
```

Specimens per Island before data clean-up

Island	freq
(blank)	2
Aotea (Great Barrier I)	10
Borneo	25
Doubtful Sound	1
Great Mercury Island	1
Halmahera	25
Hatutaa	21
Honuea	21
Kaikura Island	20
Kamaka	21
Kayangel	21
Late Island	21
Luzon	1
Mainland	3
Malenge	25
Mohotani	14
Motukawanui	21

Island	freq
New Britain	26
New Guinea	25
Normanby Island	25
Rakiura (Stewart Isl)	21
Reiono	21
Rimatuu (Tetiaroa)	21
Slipper Island	21
Sulawesi	25
Tahanea	20
Wake Island	20

To do:

- cleanup: remove invariant/monomorphic columns
- cleanup: remove SNPs/columns with few samples (which cutoff? 60%)
- cleanup: remove rows with >53% missing
- cleanup: remove SNPs not in HW equilibrium
- cleanup: remove samples that are weird in the structure analysis
- re-run NeighborNet and Mantel test
- aggregate samples at island level, calculate island-to-island Fst and/or Nei distance
- calculate heterozygosity, should decrease with distance, possibly be shaped by island size
- PCA on the SNPs