KW Population Analysis

Caila Kucheravy

2024-01-07

Examine population structure using PCA. Script from E. de Greef, help from the PCAdapt vignette: https://bcm-uga.github.io/pcadapt/articles/pcadapt.html.

Prep the environment:

```
setwd("~/Dropbox/killer_whale_genomics/snps2/03-filtered_snps")

library(pcadapt)
library(ggplot2)
library(ggrepel)
library(patchwork)
library(dplyr)
```

```
##
## 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
```

```
library(RColorBrewer)
```

Load data and sample info:

```
snp_data <- read.pcadapt("kinship_removed/killerwhale2_snps.ID.filter1.miss.biallel.min100kb.autosomes.l

sample_info <- read.csv("kinship_removed/killerwhale_genomics_sample_info_round2_kinship_removed.csv", h
```

Remove duplicates (based on kinship file) and verify that snp file IDs are the in the same order as the metadata file:

```
sample_info <- sample_info %>%
  filter(!remove_kinship=="x")

snp_IDs <- read.table("kinship_removed/killerwhale2_snps.ID.filter1.miss.biallel.min100kb.autosomes.hwe
snp_IDs$vcf_ID <- sample_info$genome_sample_ID
snp_IDs$all_equal <- snp_IDs$V1==snp_IDs$vcf_ID #column should all say "TRUE"
snp_IDs$all_equal
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

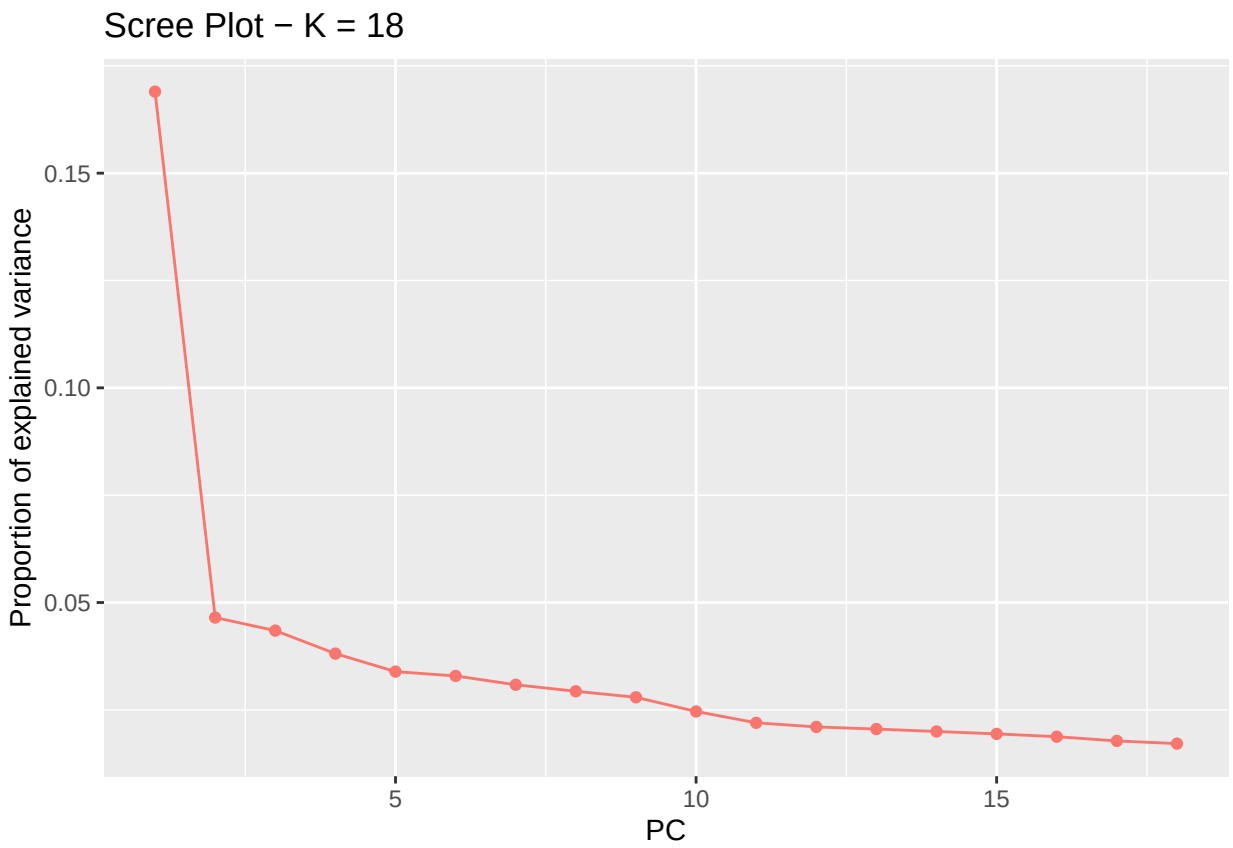
```
## [61] TRUE TRUE TRUE TRUE TRUE
```

Run pcadapt, setting k-value to the desired number of eigenvectors to be produced:

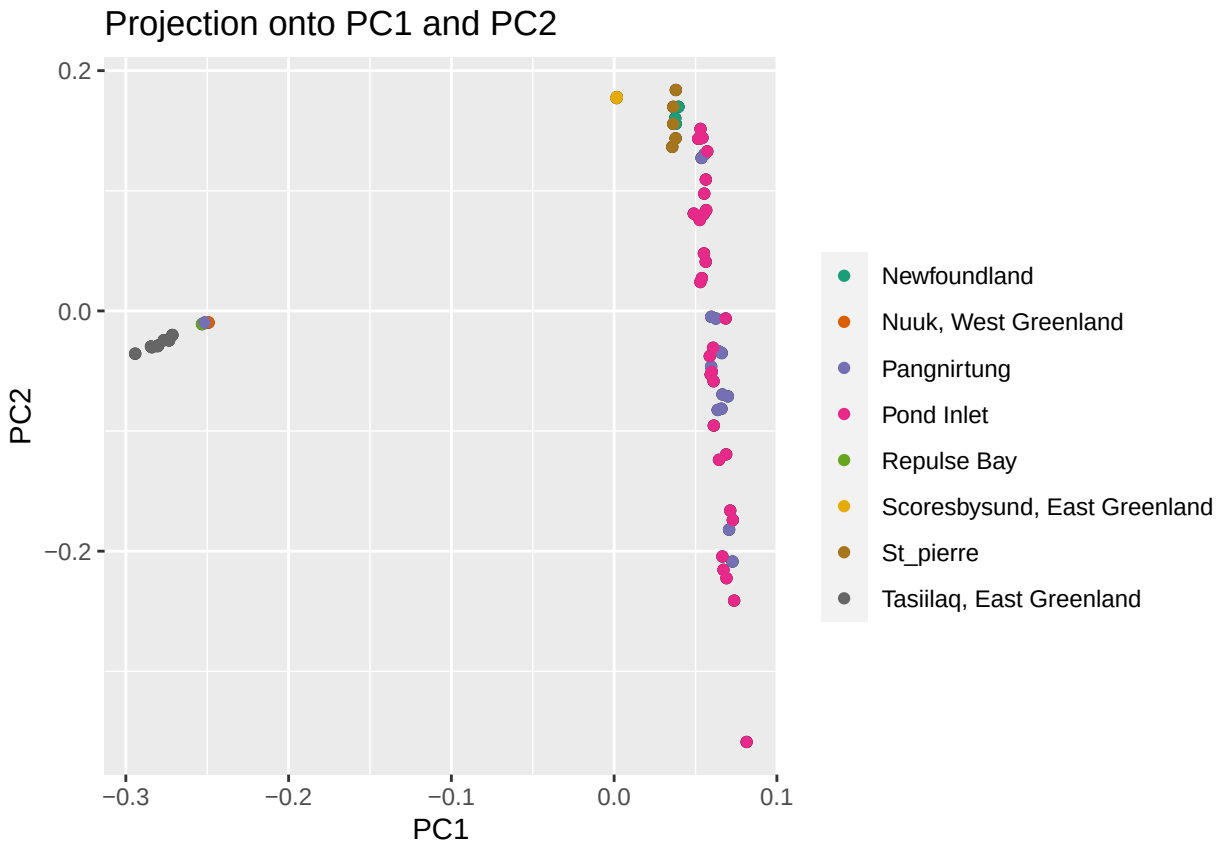
```
x <- pcadapt(input = snp_data, K = 18)
```

Plot screeplot and PCA:

```
# Screeplot:
plot(x, option = "screeplot")
```




```
# Quick PCA:
cols <- brewer.pal(8, "Dark2")
plot(x, option = "scores", pop = sample_info$location_name, col = cols, labels = sample_info$genome_samp
```



Examine PCA scores, loadings, and z-scores, and calculate proportion variance for first few eigenvectors:

```
# scores:
scores <- as.data.frame(x$scores)

# loadings:
loadings <- as.data.frame(x$loadings)

# z-scores:
z_scores <- as.data.frame(x$zscores)

# proportion variance
proportion <- as.data.frame(x$singular.values)
proportion$squared <- proportion$`x$singular.values`* proportion$`x$singular.values`
prop_var <- as.data.frame(proportion$squared)
PC1_proportion <- (round(prop_var[1,], digits=4))*100
PC2_proportion <- (round(prop_var[2,], digits=4))*100
PC3_proportion <- (round(prop_var[3,], digits=4))*100
PC4_proportion <- (round(prop_var[4,], digits=4))*100
```

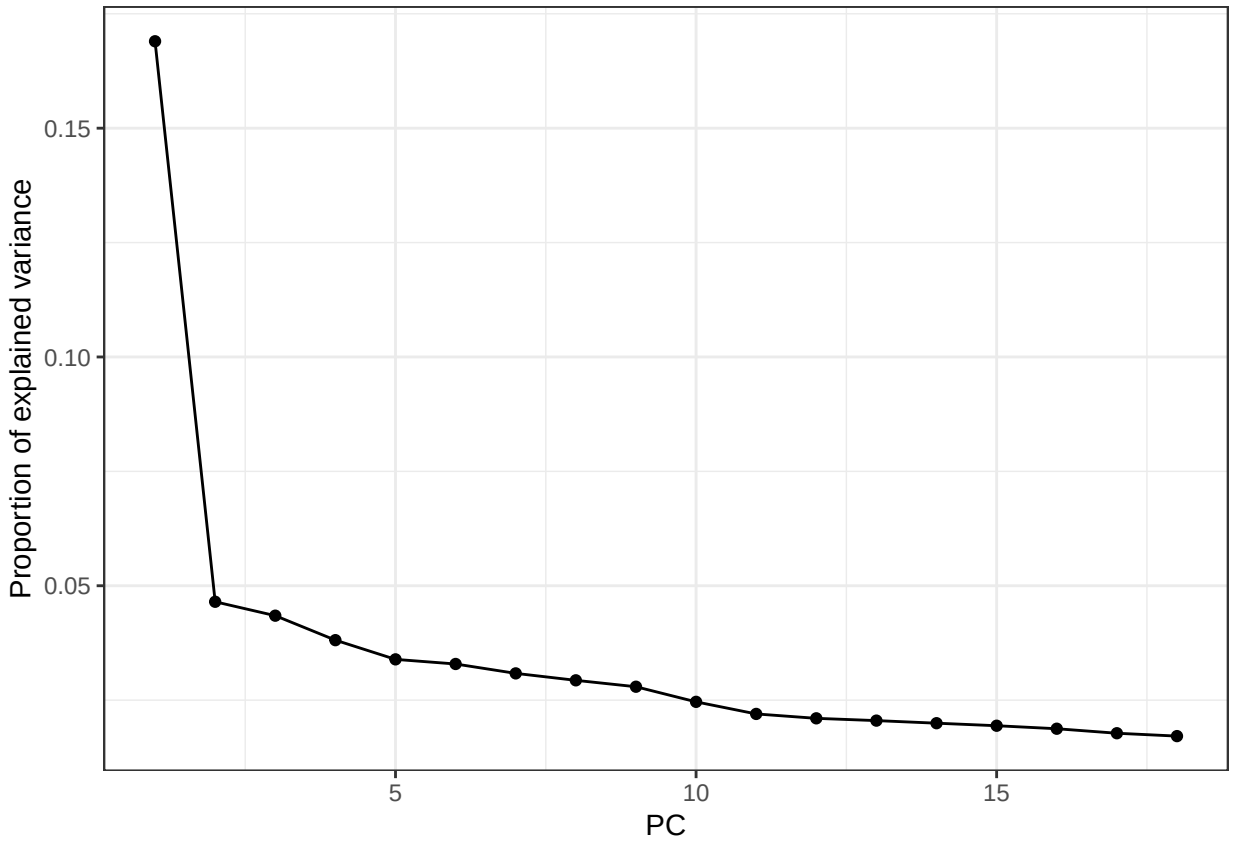
Make screeplot nicer:

```
prop_var$num <- 1:nrow(prop_var)

scree <- ggplot(data=prop_var, aes(x=num, y=prop_var$`proportion$squared`))+
  geom_point()+
  geom_line()+
  theme_bw()+
```

```
  ylab("Proportion of explained variance")+
  xlab("PC")

scree
```



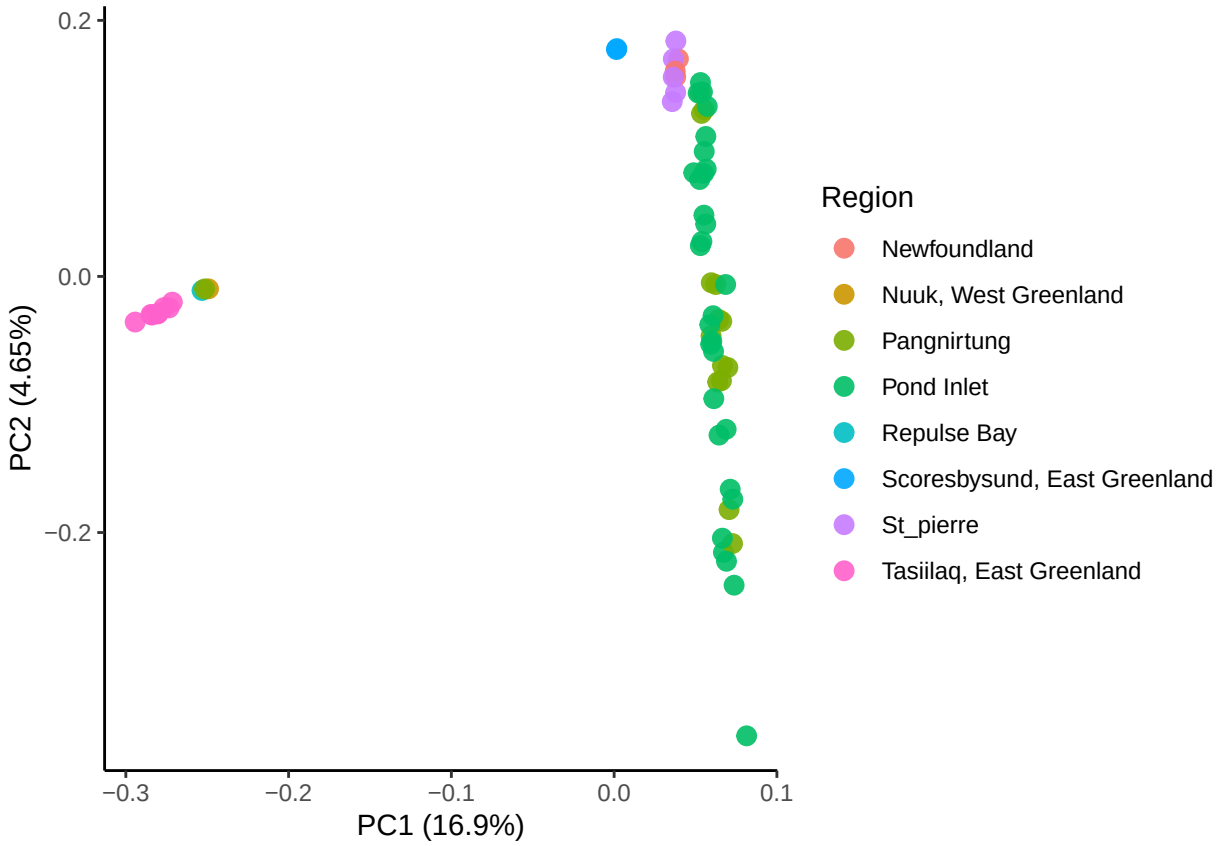

```
#ggsave("scree_plot_LDprunedr08_n57.png", width=6, height=4.5, dpi=300)
```

Make PCA nicer:

```
evec <- cbind(sample_info$genome_sample_ID, scores)
colnames(evec)[1] <- "sample"

pca <- ggplot(data=evec, aes(x=V1,y=V2))+
  #previously used point size 3, but increasing
  geom_point(aes(color=sample_info$location_name),size=3, alpha=0.9)+
  theme_classic()+
  xlab(paste("PC1 (", PC1_proportion, "%)", sep=""))+
  ylab(paste("PC2 (", PC2_proportion, "%)", sep=""))+
  #geom_text_repel(aes(label=sample_info$genome_sample_ID), size=2)+
  labs(color= "Region")

pca
```



We see the two populations as before. Let's take a closer look at the HA population to see if there is any substructure, etc.

Load data and sample info, remove duplicates and Greenland/Low Arctic whales, verify that snp file IDs match sample file:

```
snp_data2 <- read.pcadapt("kinship_LA_removed/killerwhale2_snps.ID.filter1.miss.biallel.min100kb.autoso

sample_info2 <- read.csv("kinship_LA_removed/killerwhale_genomics_sample_info_round2_kinship_removed.cs

sample_info2 <- sample_info2 %>%
  filter(!remove_kinship=="x") %>%
  filter(!remove_LA=="x")

snp_IDs2 <- read.table("kinship_LA_removed/killerwhale2_snps.ID.filter1.miss.biallel.min100kb.autosomes
snp_IDs2$vcf_ID <- sample_info2$genome_sample_ID
snp_IDs2$all_equal <- snp_IDs2$V1==snp_IDs2$vcf_ID #column should all say "TRUE"
snp_IDs2$all_equal
```

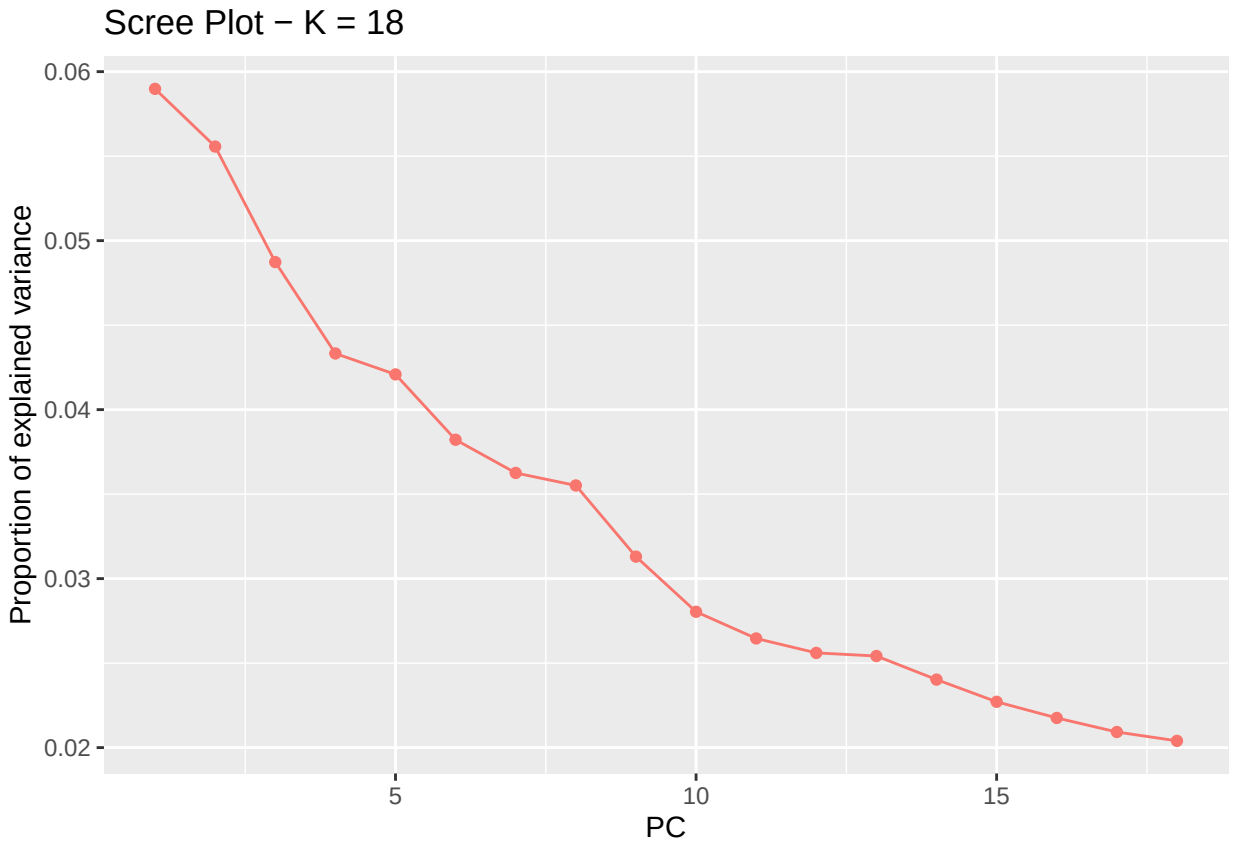
```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [31] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [46] TRUE TRUE TRUE TRUE TRUE TRUE TRUE
```

Run pcadapt:

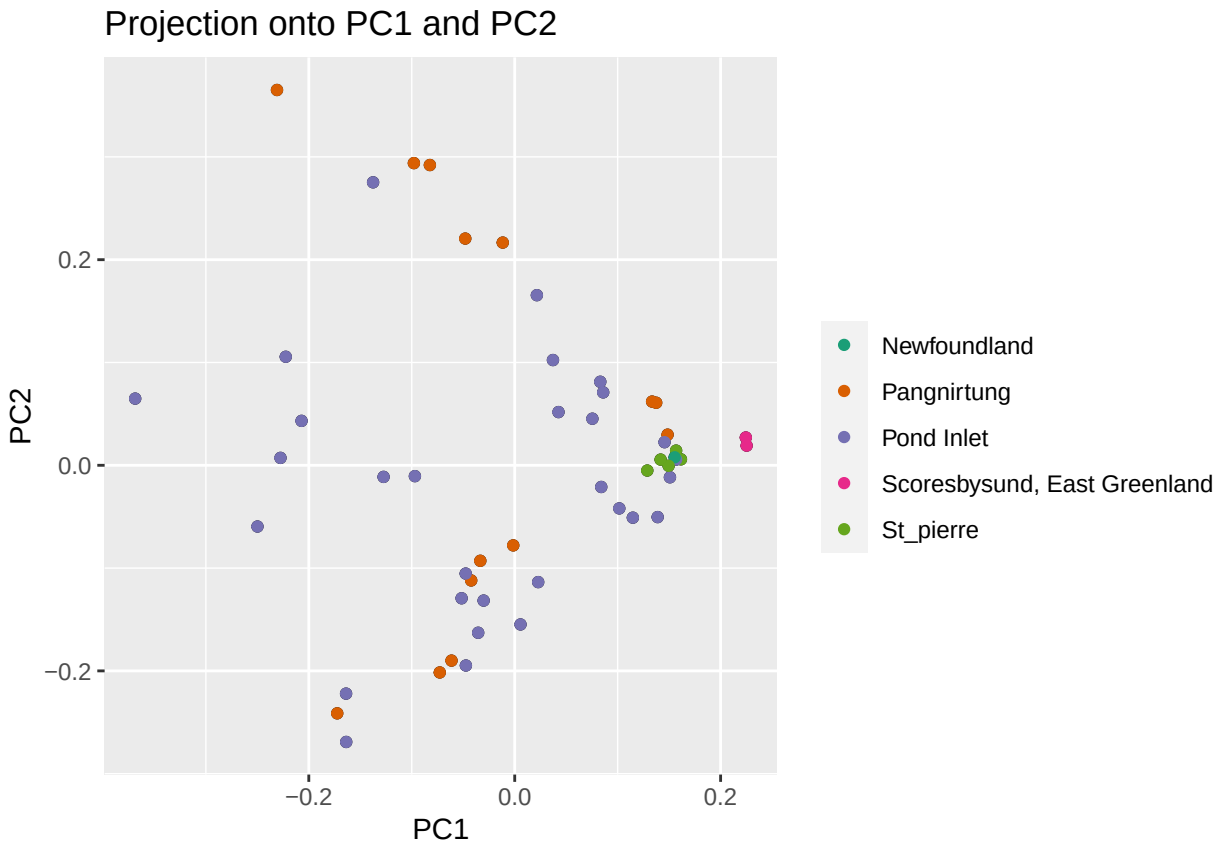
```
x2 <- pcadapt(input = snp_data2, K = 18)
```

Plot screeplot and PCA:

```r
# Screeplot:
plot(x2, option = "screeplot")
```




```r
# Quick PCA:
plot(x2, option = "scores", pop = sample_info2$location_name, col = cols, labels = sample_info2$genome_
```


Projection onto PC1 and PC2
PC2
0.2
0.0
−0.2
−0.2
0.0
0.2
PC1
Newfoundland
Pangnirtung
Pond Inlet
Scoresbysund, East Greenland
St_pierre