Peelii Larval Structure - Sites and Years

Alan Couch

Fri Jul 22 1:51:49 PM 2016

## Loading required package: knitr

## Loading project configuration

## Autoloading helper functions

## Running helper script: funRel.R

## Autoloading cache

## Autoloading data

## Loading data set: DARTallIdentifiers

## Loading data set: qslAllLarvaInfo

## Loading data set: qslAllLarvaInfo30MAY16

## Loading data set: qslMetaLarvAndAdultsUnion

## Loading data set: qslMPeeliiForRelated

## Loading data set: qslMPeeliiForRelated20JUN16

## Loading data set: Report.DMac15.1861

## Munging data

## Running preprocessing script: 01MungeGeneticsDatacontaminationfix.R

## Loading required package: 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

## Running preprocessing script: 02MungeLarvaSnpsByYears.R

## Running preprocessing script: 03MungeLarvaSnpsByYearCombos.R

a<-suppressPackageStartupMessages({  
library(plyr)  
library(dplyr)  
library(ggplot2)  
library(dart)  
library(tsne)  
})  
  
ptm <- proc.time()

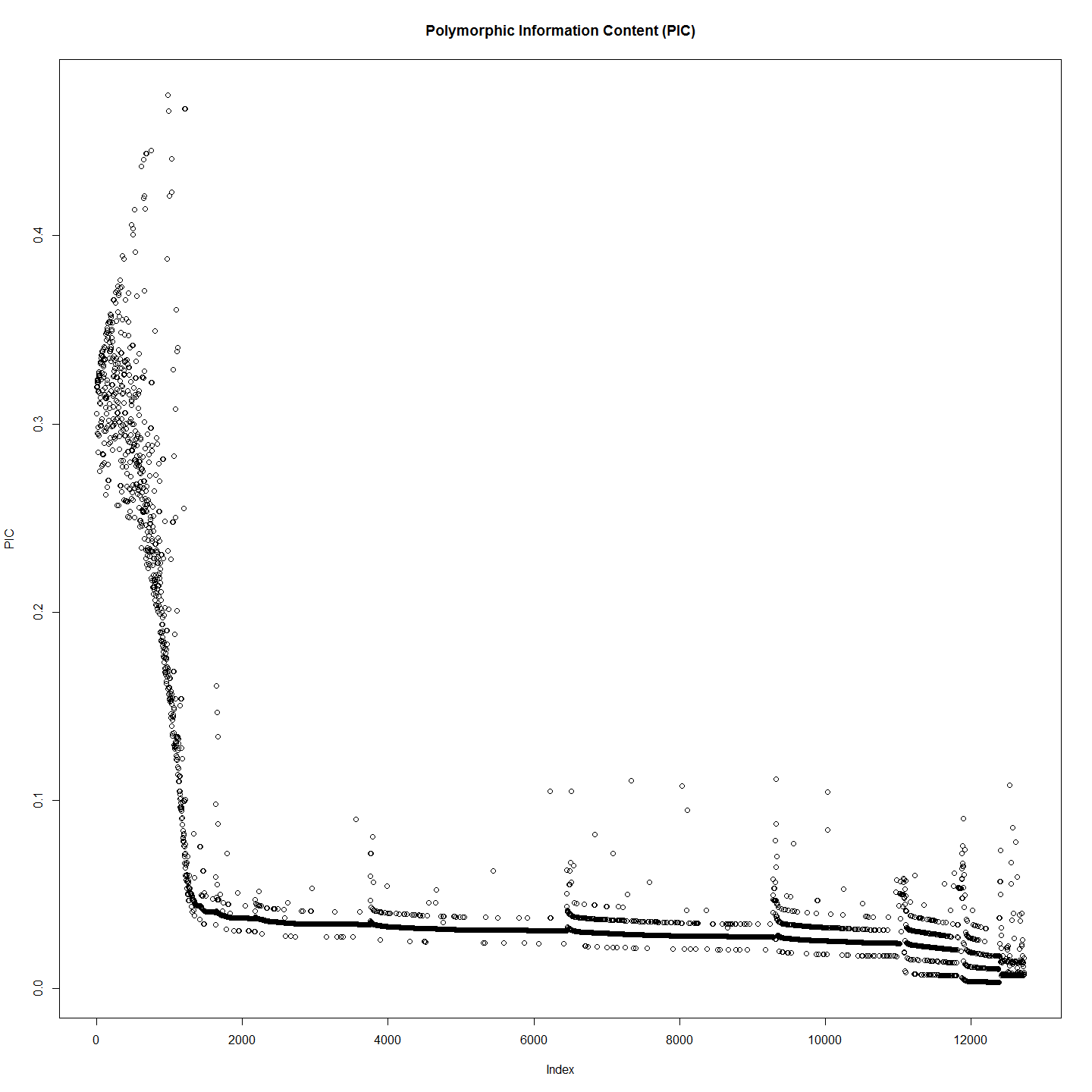
# Hardy Weinberg and Genotypic Disequilibrium

# Population Structure

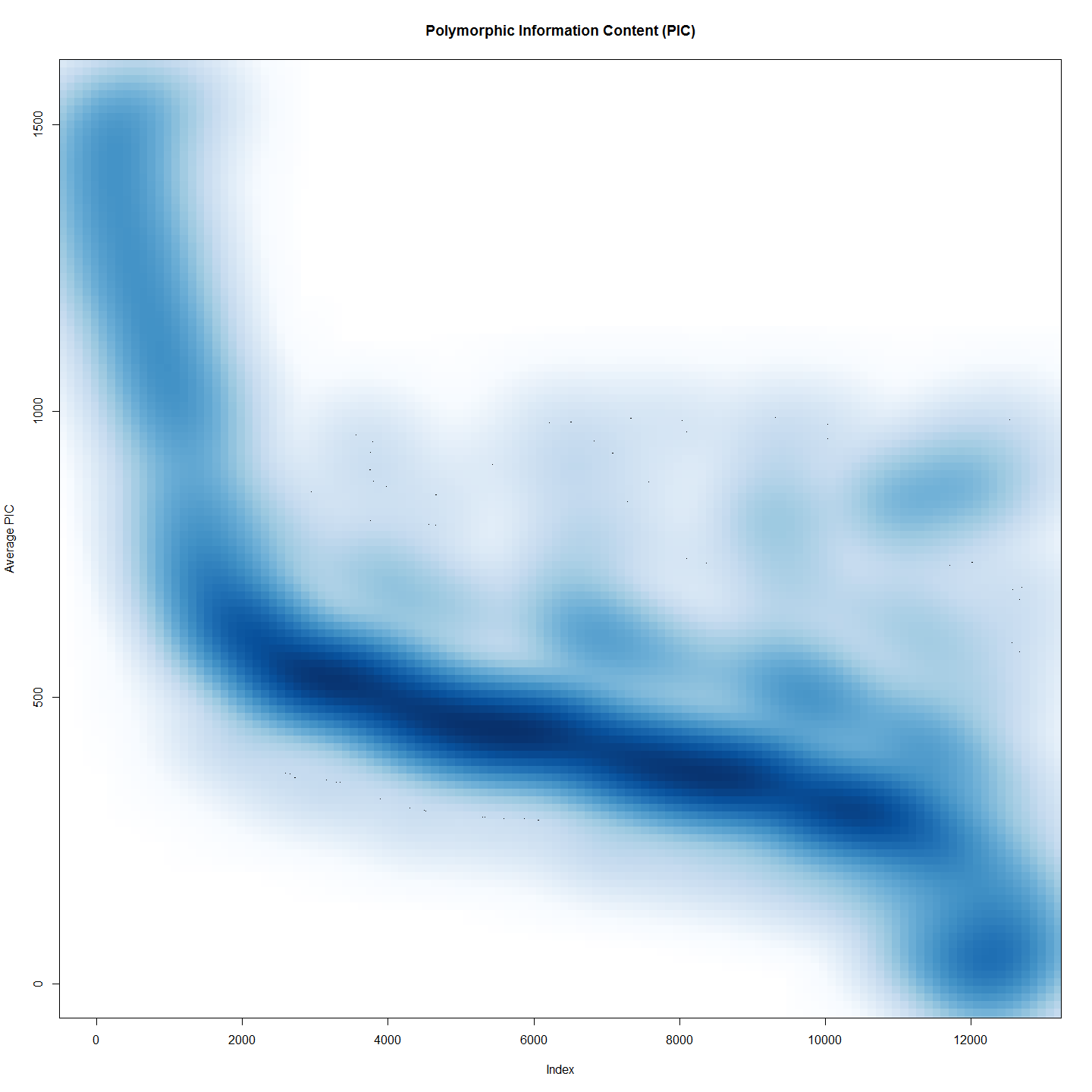
See: runPCAandTSNEYearsandSites.Rmd

# Polymorphic Information Content

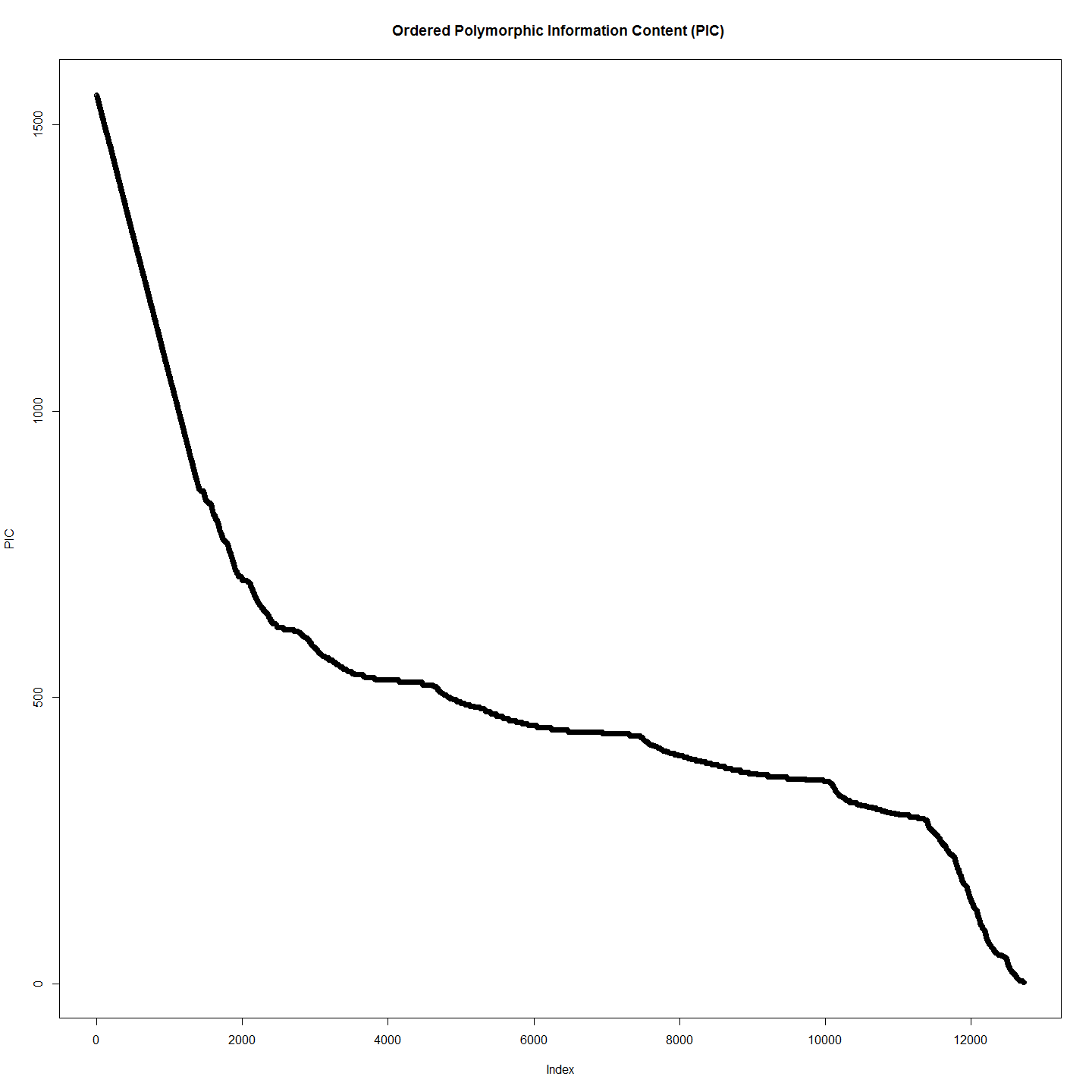
#A script to plot all M. peelii snps average PIC (PIC of reference and PIC of alternate state allele) against index number of the data frame. This shows that using the first X number of snps is using the alleles with the highest PIC.  
  
  
df<-allPeeliSnps; main<-"All Peelii snps"  
df<-df[-c(1:4),]  
  
pic<-df[2:nrow(df),14] #check pic in order  
plot(as.character(pic),main="Polymorphic Information Content (PIC)", ylab="PIC")# to show PIC is in order (ish) ideally cut off would be 1300 or so.



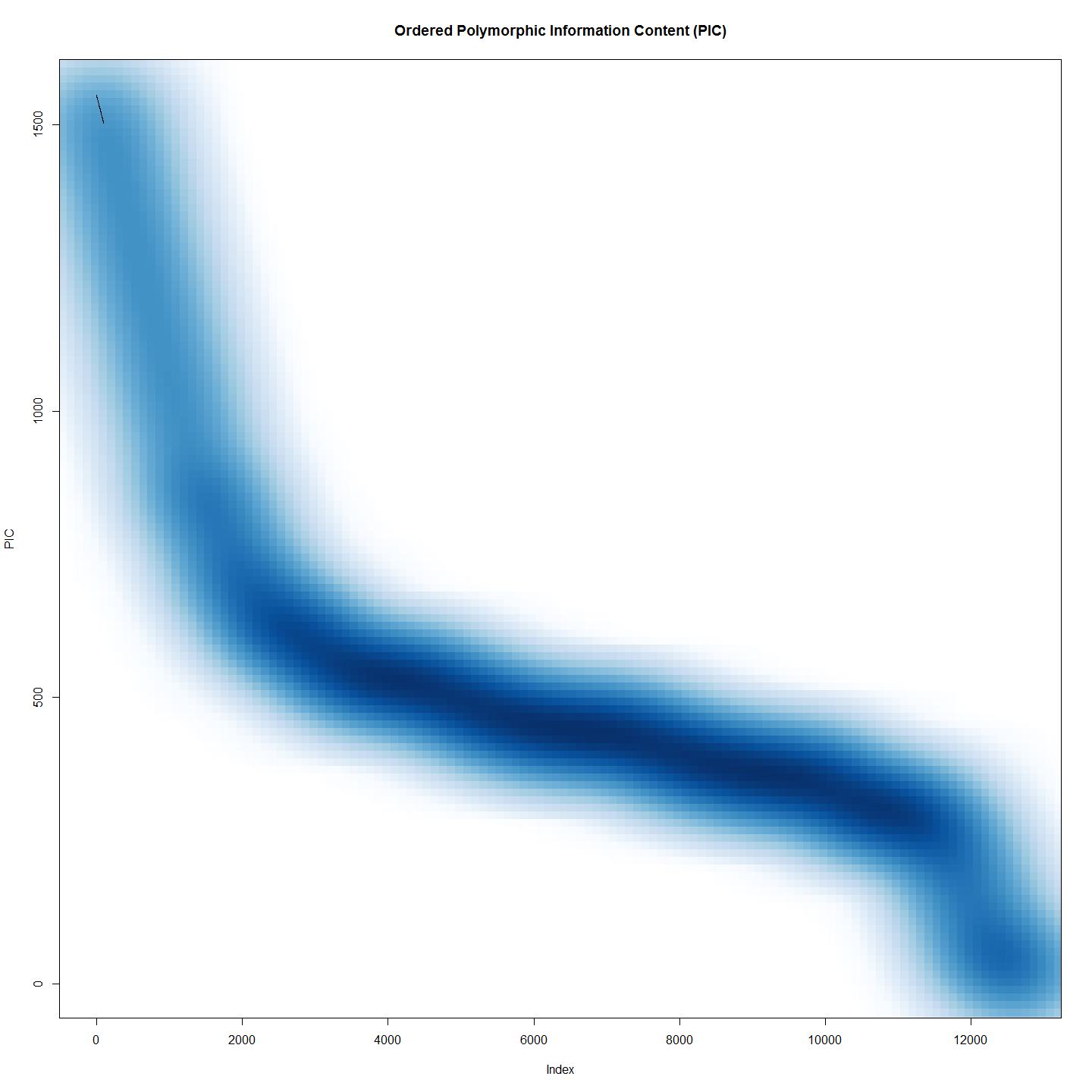
smoothScatter(pic,main="Polymorphic Information Content (PIC)", ylab="Average PIC")



# Place in strict order of PIC and have alook  
pico<-data.frame(pic)  
rownames(pico) <- NULL  
colnames(pico)[1]<-"PIC"  
  
pico <- data.frame(pico[order(as.numeric(pico$PIC),decreasing = TRUE),])  
  
pico[,2]<-pico[,1] #Order columns to facilitate plotting.  
pico[,1]<-row.names(pico) #Replace name with index number  
colnames(pico)[1]<-"Index"  
colnames(pico)[2]<-"PIC"  
  
plot(pico, main="Ordered Polymorphic Information Content (PIC) ")



smoothScatter(pico, main="Ordered Polymorphic Information Content (PIC)")



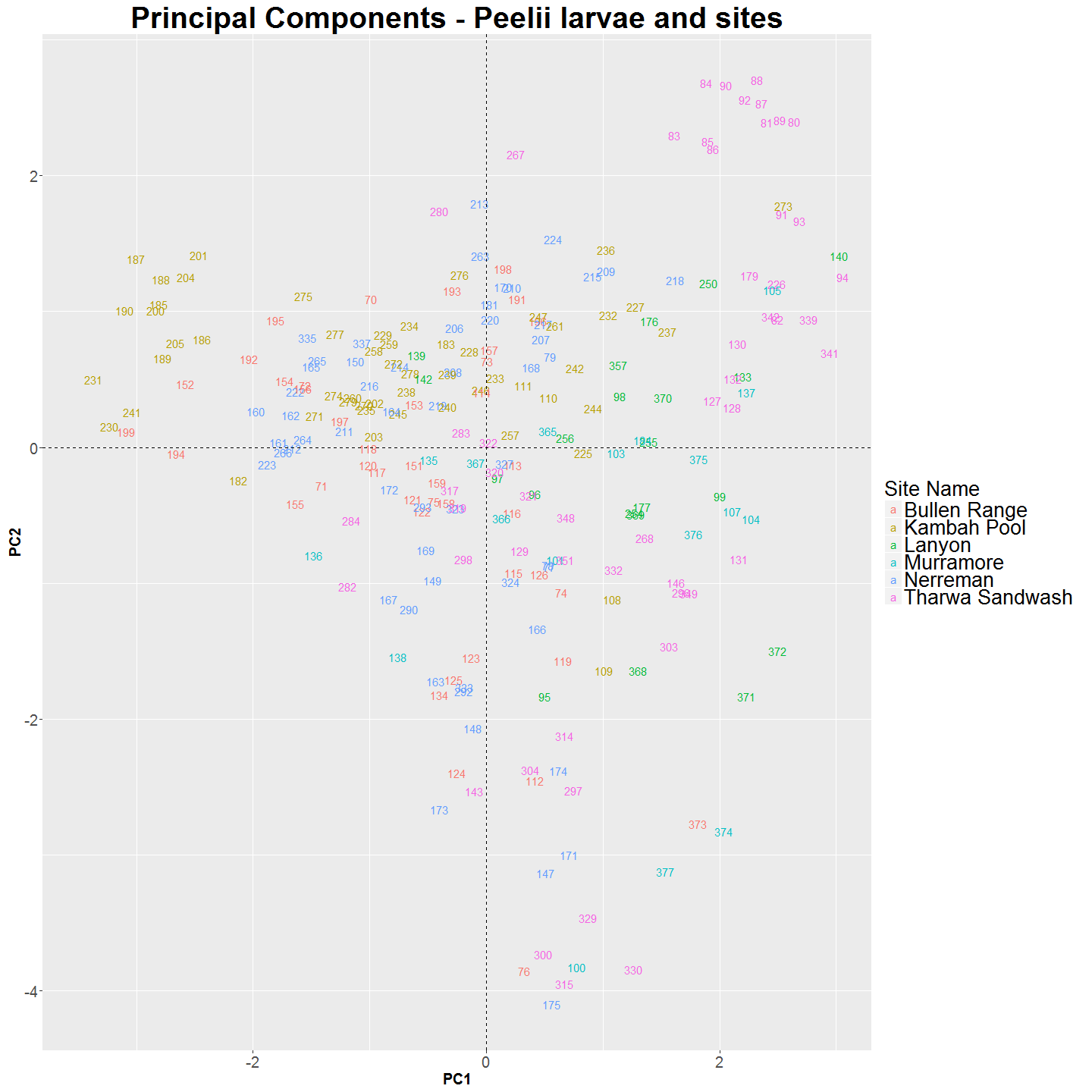
rm(pic)  
rm(pico)

## Genome Structure in Larvae by Years and Sites

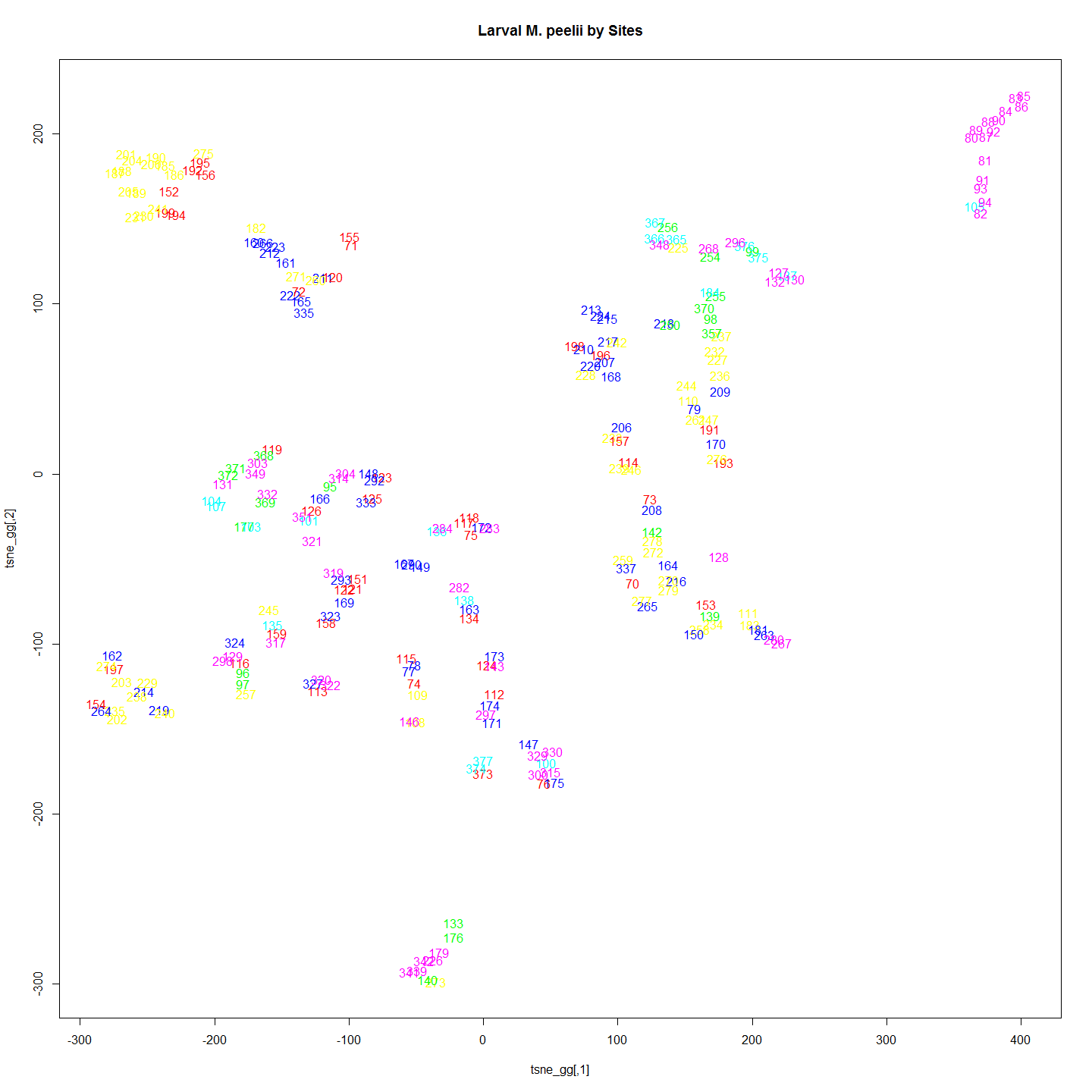
#First we do PCA and tSNE using peelii larvae and site  
#Vanilla PCA  
require(adegenet)  
require(ggplot2)  
  
system.time(pca1 <- glPca(gl[,], parallel=F, nf=3))

## user system elapsed   
## 55.76 0.03 56.14

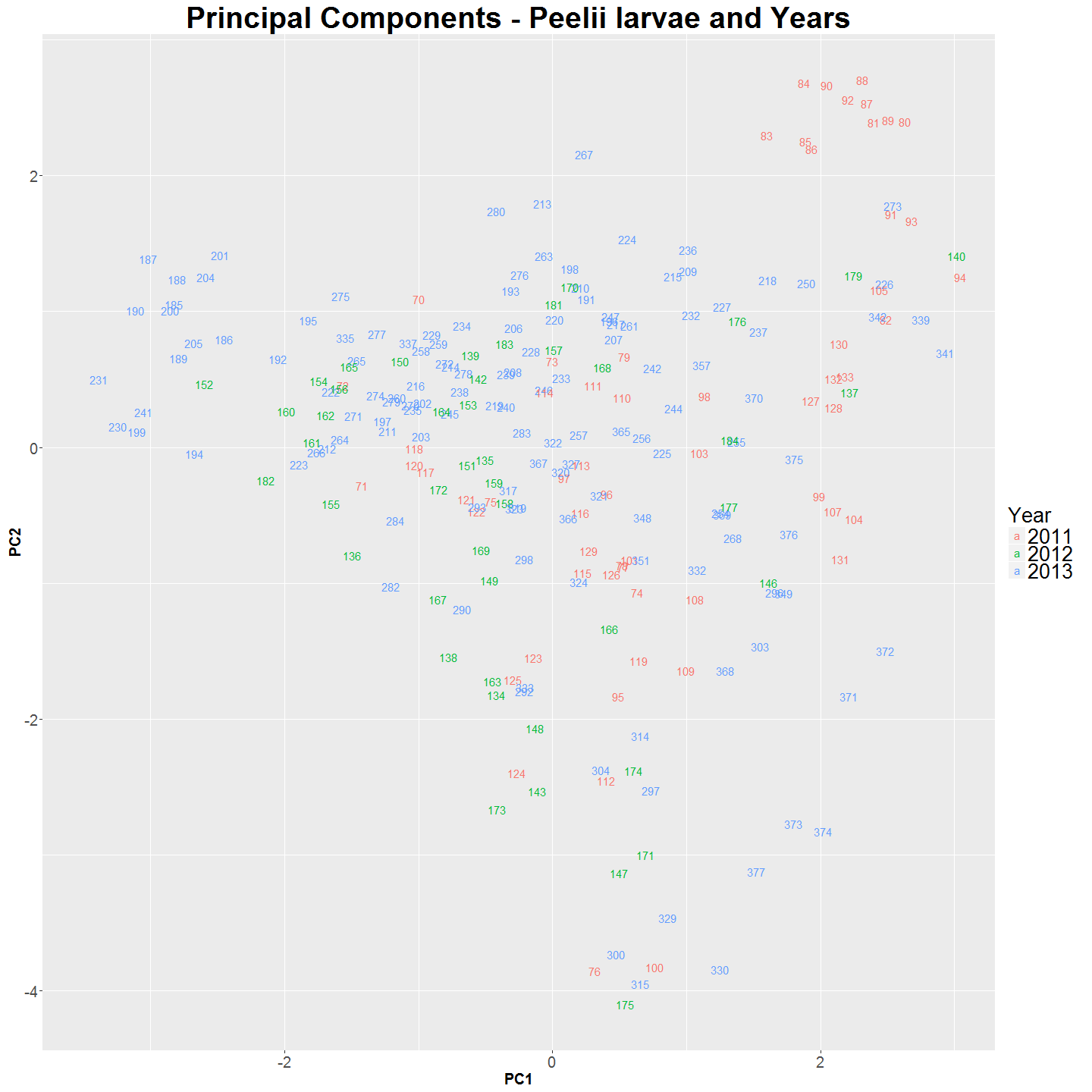
gg <- data.frame(pca1$scores,cluster=pop(gl))  
gg$labels <- rownames(gg)  
  
ggplot(gg, aes(x=PC1, y=PC2))+  
 geom\_text(aes(label=labels, color=factor(cluster)),hjust=-0.2, size=4) +  
 geom\_hline(yintercept=0,linetype=2) +  
 geom\_vline(xintercept=0,linetype=2) +  
 scale\_color\_discrete(name="Site Name") +  
 ggtitle("Principal Components - Peelii larvae and sites")+  
 theme(plot.title = element\_text(lineheight=.8, face="bold", size=29))+  
 theme(axis.text=element\_text(size=15), axis.title=element\_text(size=15,face="bold"),legend.text = element\_text(size = 20),legend.title = element\_text(size = 20))



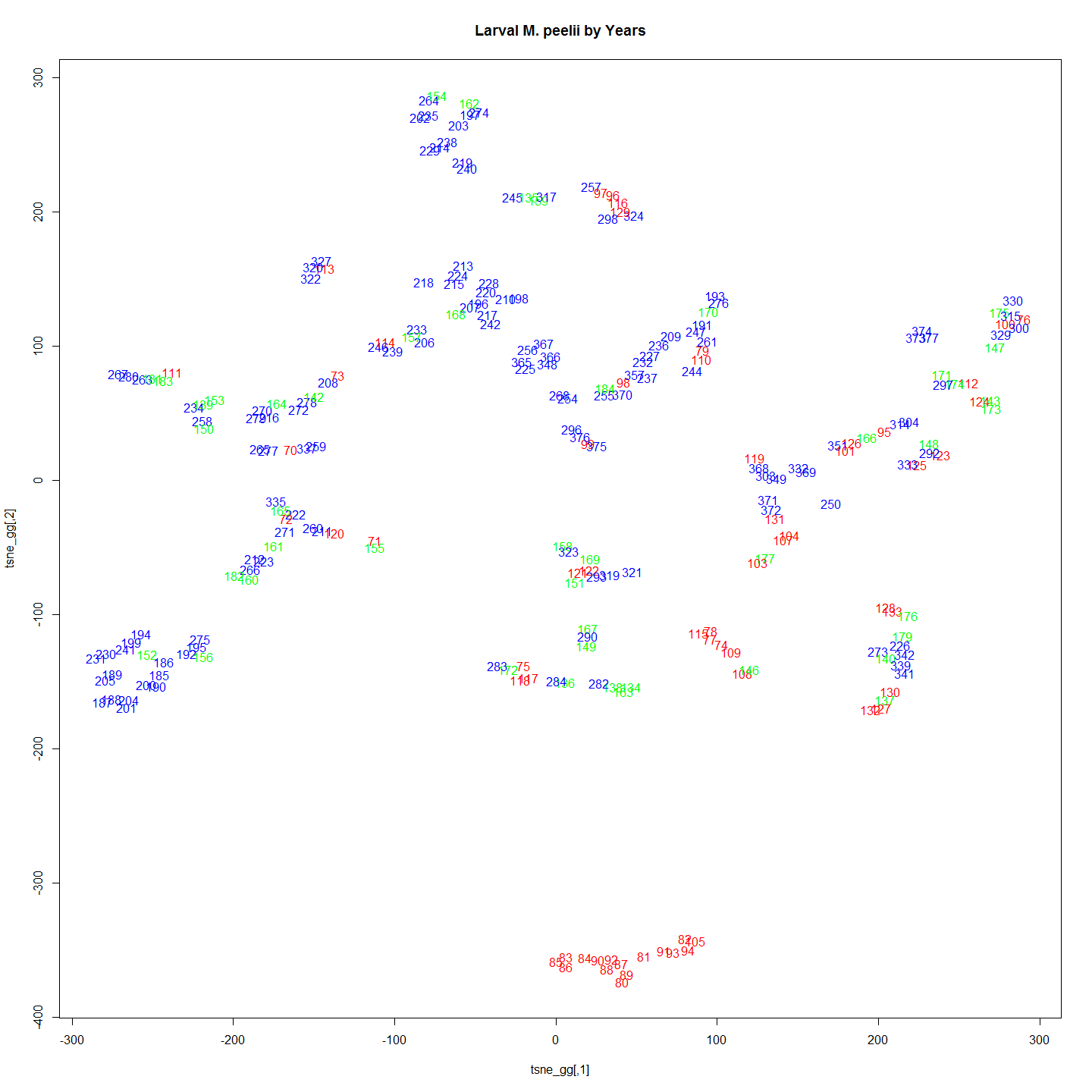
#tsne 2d  
require(tsne)  
colors<- rainbow(length(unique(gg$cluster)))  
names(colors) <- unique(gg$cluster)  
#ecb <-function(x,y){ plot(x,t='n',main="Larval M. peelii by Sites"); text(x,labels= row.names(gg),col=colors[gg$cluster])} #to make series  
#tsne\_gg <- tsne(gg[,1:3], max\_iter=10000,k=2, epoch\_callback = ecb, perplexity=5,epoch=1000)  
tsne\_gg <- tsne(gg[,1:3], max\_iter=10000,k=2, perplexity=5)  
plot(tsne\_gg,t='n',main="Larval M. peelii by Sites"); text(tsne\_gg,labels= row.names(gg),col=colors[gg$cluster])



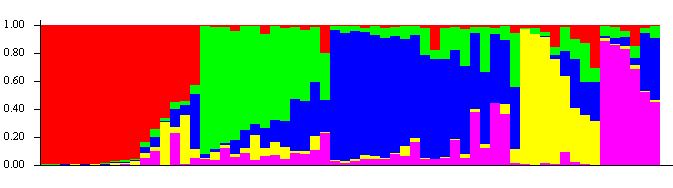
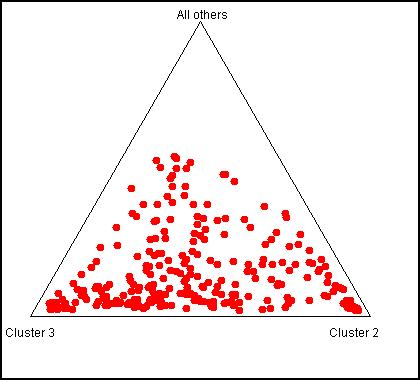
#Second we do PCA and tSNE with the same larvae coloured by year   
gg<- data.frame(pca1$scores,cluster=as.character(gl@other$covariates))  
gg$labels <- rownames(gg)  
ggplot(gg, aes(x=PC1, y=PC2))+  
 geom\_text(aes(label=labels, color=factor(cluster)),hjust=-0.2, size=4) +  
 # geom\_hline(yintercept=0,linetype=2) +  
 # geom\_vline(xintercept=0,linetype=2) +  
 scale\_color\_discrete(name="Year") +  
 ggtitle("Principal Components - Peelii larvae and Years")+  
 theme(plot.title = element\_text(lineheight=.8, face="bold", size=29))+  
 theme(axis.text=element\_text(size=15), axis.title=element\_text(size=15,face="bold"),legend.text = element\_text(size = 20),legend.title = element\_text(size = 20))

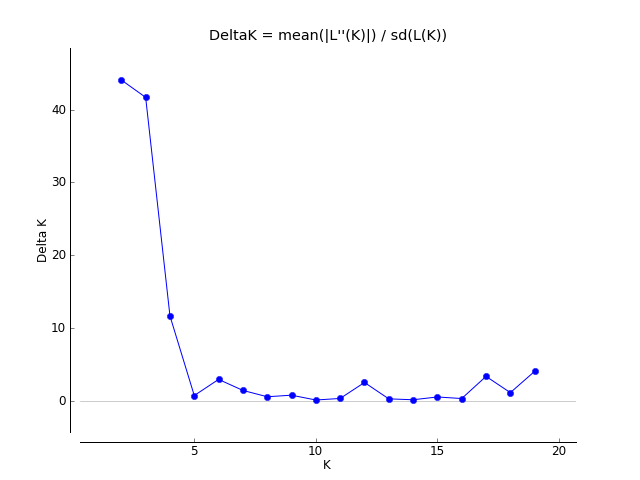


#tsne 2d  
require(tsne)  
colors<- rainbow(length(unique(gg$cluster)))  
names(colors) <- unique(gg$cluster)  
#ecb <-function(x,y){ plot(x,t='n',main="Larval M. peelii by Years"); text(x,labels= row.names(gg),col=colors[gg$cluster])} #to make series   
#tsne\_gg <- tsne(gg[,1:3], max\_iter=10000, epoch\_callback = ecb, perplexity=5,epoch=1000)  
tsne\_gg <- tsne(gg[,1:3], max\_iter=10000, perplexity=5)  
plot(tsne\_gg,t='n',main="Larval M. peelii by Years"); text(tsne\_gg,labels= row.names(gg),col=colors[gg$cluster])

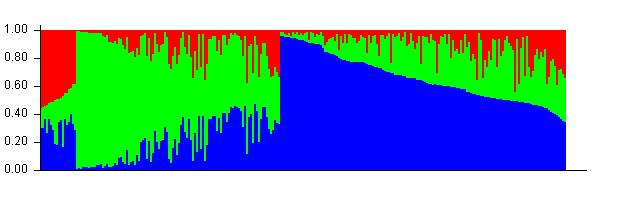
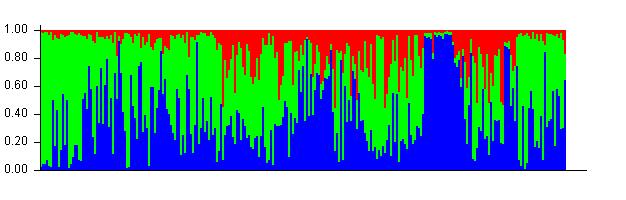


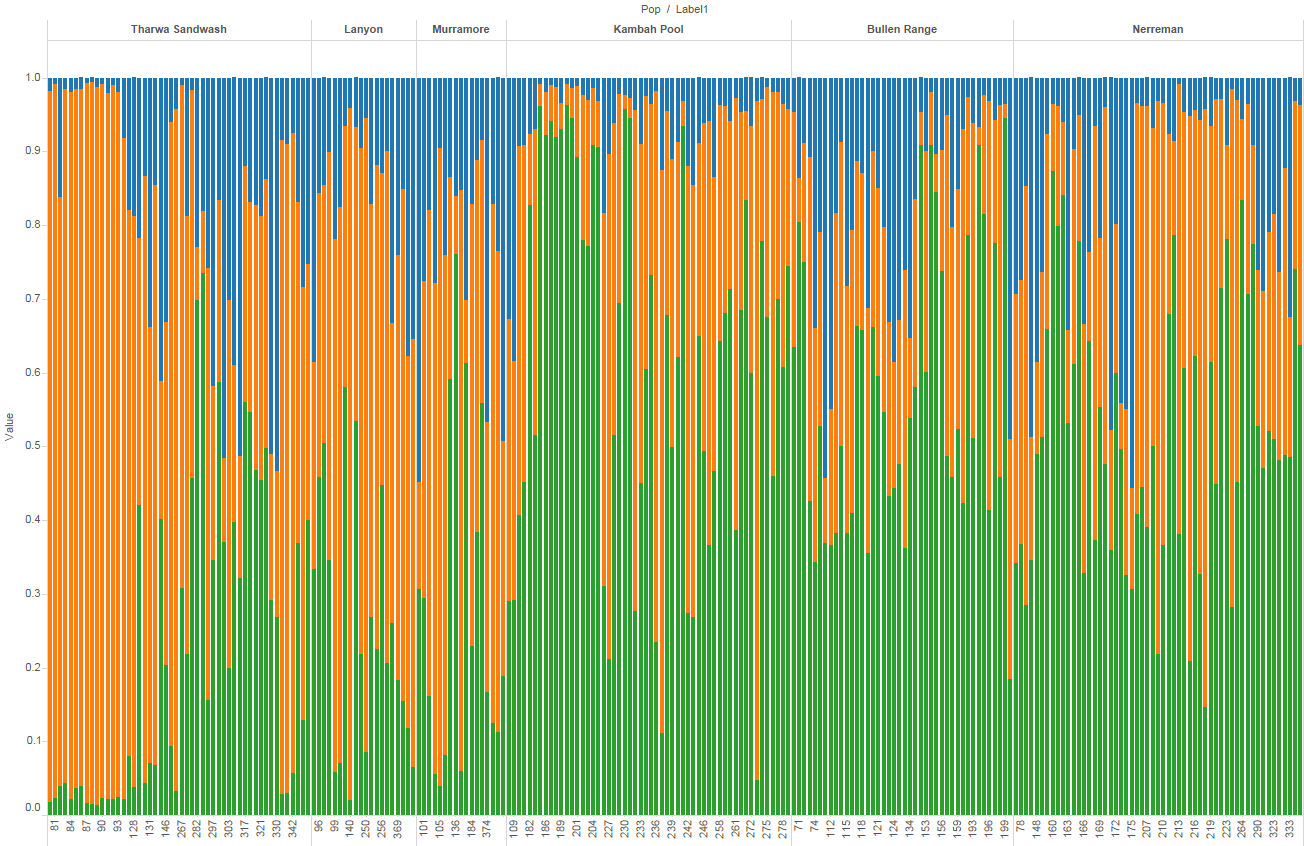
# ####################################Incase we need 3d plots  
# require(scatterplot3d)  
# tsne\_gg <- tsne(gg[,1:3], max\_iter=1000,k=3, perplexity=5)  
# scatterplot3d(tsne\_gg[,1],tsne\_gg[,2],tsne\_gg[,3], pch=16, highlight.3d=TRUE, main="3D Scatterplot")  
#   
# require(rgl)  
# #plot3d(tsne\_gg[,1],tsne\_gg[,2],tsne\_gg[,3], col="red", size=3)  
# plot3d(tsne\_gg[,1],tsne\_gg[,2],tsne\_gg[,3], col=colors[gg$cluster], size=3)

A little structure is evident if we look at structure based on collection site.  or as a triangle plot 

Following the evanno method we settled on 3 groups. 

so with three groups.

The structure is a associated with collection site. 

When the structure is examined based on the nest site rather than the collection site which better accounts for putative barriers and unidirectional dispersal. then:

after nests chapter is done

#Looking for DAPC, fixed differences, and Correspondence Analysis between populations.  
  
require(reshape2)  
require(dart)  
require(pegas)  
all.dart <- read.dart("otherData/larvalPeeliSnps.csv", topskip = 5)

## Added the following covmetrics:  
## CloneID SNP SnpPosition CallRate OneRatioSnp FreqHomRef FreqHomSnp FreqHets AvgCountRef AvgCountSnp RepAvg .  
## Number of rows per Clone. Should be only 2s:  
## 2  
## Recognised: 243 individuals and 6364 SNPs in otherData/larvalPeeliSnps.csv

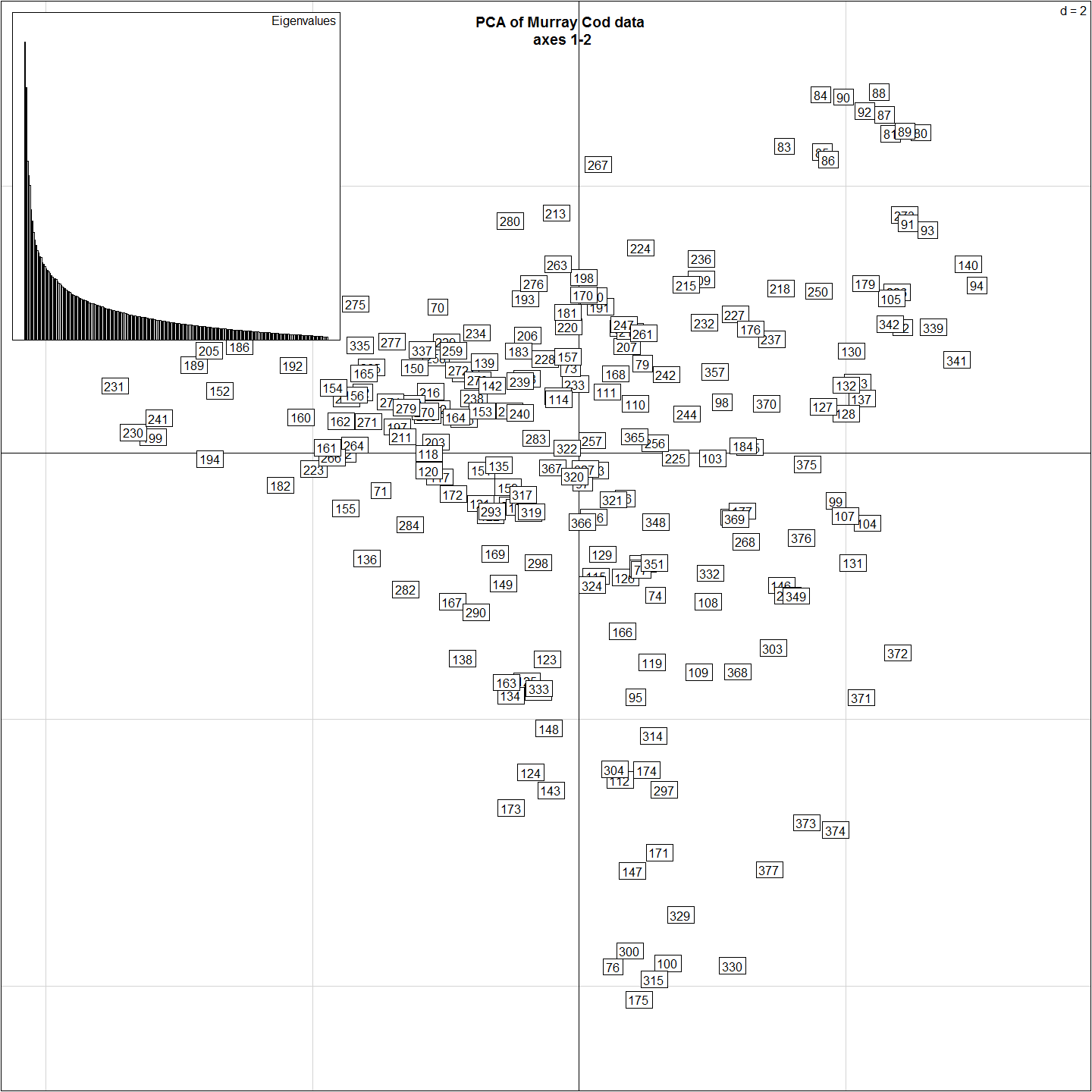
gl.dart <- dart2genlight(all.dart, covfilename = "otherData/qslLarvalPeeliiMetaForPCA.csv") # this glObject is then suitable for sharing via GitHub or such for open data as required by PeerJ etc. Need better file but.

## Start conversion....  
## Please note conversion of bigger data sets will take some time!  
## Once finished, we recommend to save the object using save(object, file="object.rdata")  
  
## Try to add covariate file: otherData/qslLarvalPeeliiMetaForPCA.csv .  
## Ids of covariate file does not match the number of ids in the genetic file. Maybe this is fine if a subset matches.  
## Ids of covariate file (at least a subset of) are matching!  
## Found 243 matching ids out of 251 ids provided in the covariate file.  
## Added pop factor.  
## Added latlon data.  
## Added YearOnly to the other$covariates slot.  
## Added estimatedAge to the other$covariates slot.  
## Added Day.of.Year to the other$covariates slot.  
## Added hatchedDoY to the other$covariates slot.  
## Added Delta13C to the other$covariates slot.  
## Added Delta15N to the other$covariates slot.  
## Added CNRatio to the other$covariates slot.  
## Added Distance.to.Angle.Crossing..m. to the other$covariates slot.  
## Added raceCladeName to the other$covariates slot.  
## Added cladeGoTName to the other$covariates slot.  
## Added mumPulldown to the other$covariates slot.  
## Added Fathers to the other$covariates slot.  
## Added inbredCoeff to the other$covariates slot.

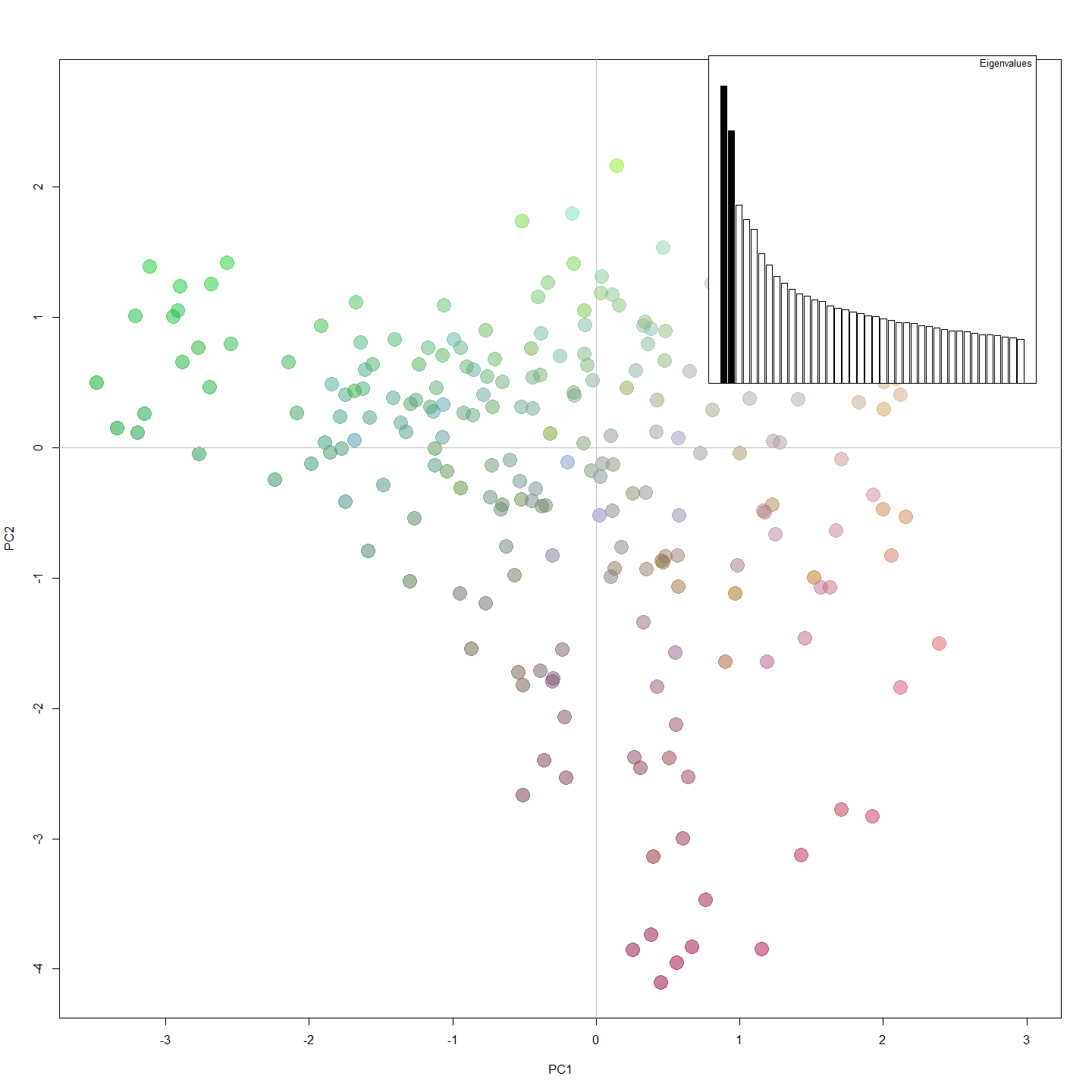
# cant work gl.fixed.diff(gl.dart, t=0)  
  
#PCA  
pca1 <- glPca(gl.dart, nf = 3 ,parallel = FALSE)  
pca1

## === PCA of genlight object ===  
## Class: list of type glPca  
## Call ($call):glPca(x = gl.dart, nf = 3, parallel = FALSE)  
##   
## Eigenvalues ($eig):  
## 2.077 1.762 1.245 1.144 1.078 0.908 ...  
##   
## Principal components ($scores):  
## matrix with 243 rows (individuals) and 3 columns (axes)   
##   
## Principal axes ($loadings):  
## matrix with 6364 rows (SNPs) and 3 columns (axes)

scatter(pca1, posi="topleft")  
title("PCA of Murray Cod data\n axes 1-2")



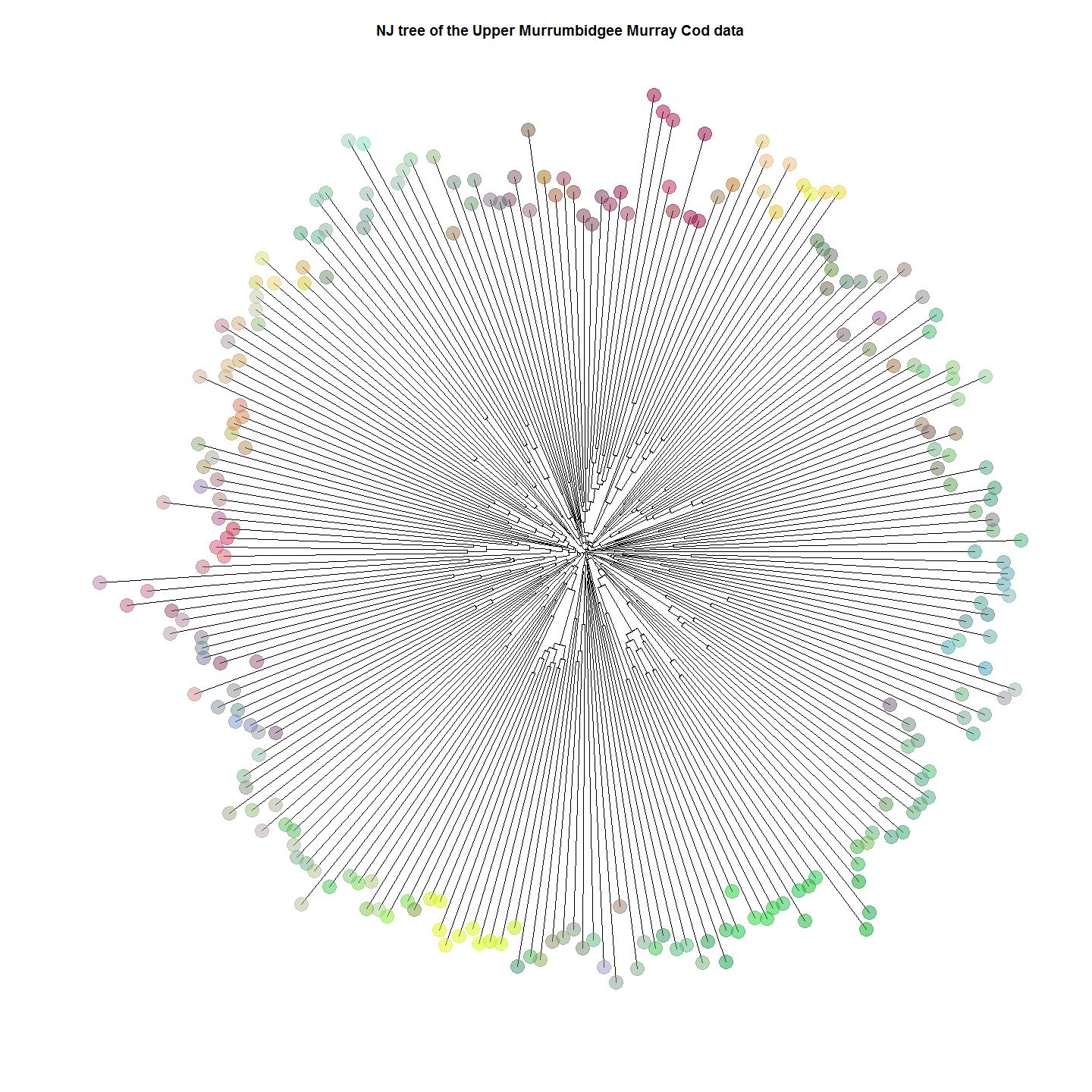
myCol <- colorplot(pca1$scores,pca1$scores, transp=TRUE, cex=4)  
abline(h=0,v=0, col="grey")  
add.scatter.eig(pca1$eig[1:40],2,1,2, posi="topright", inset=.05, ratio=.3)



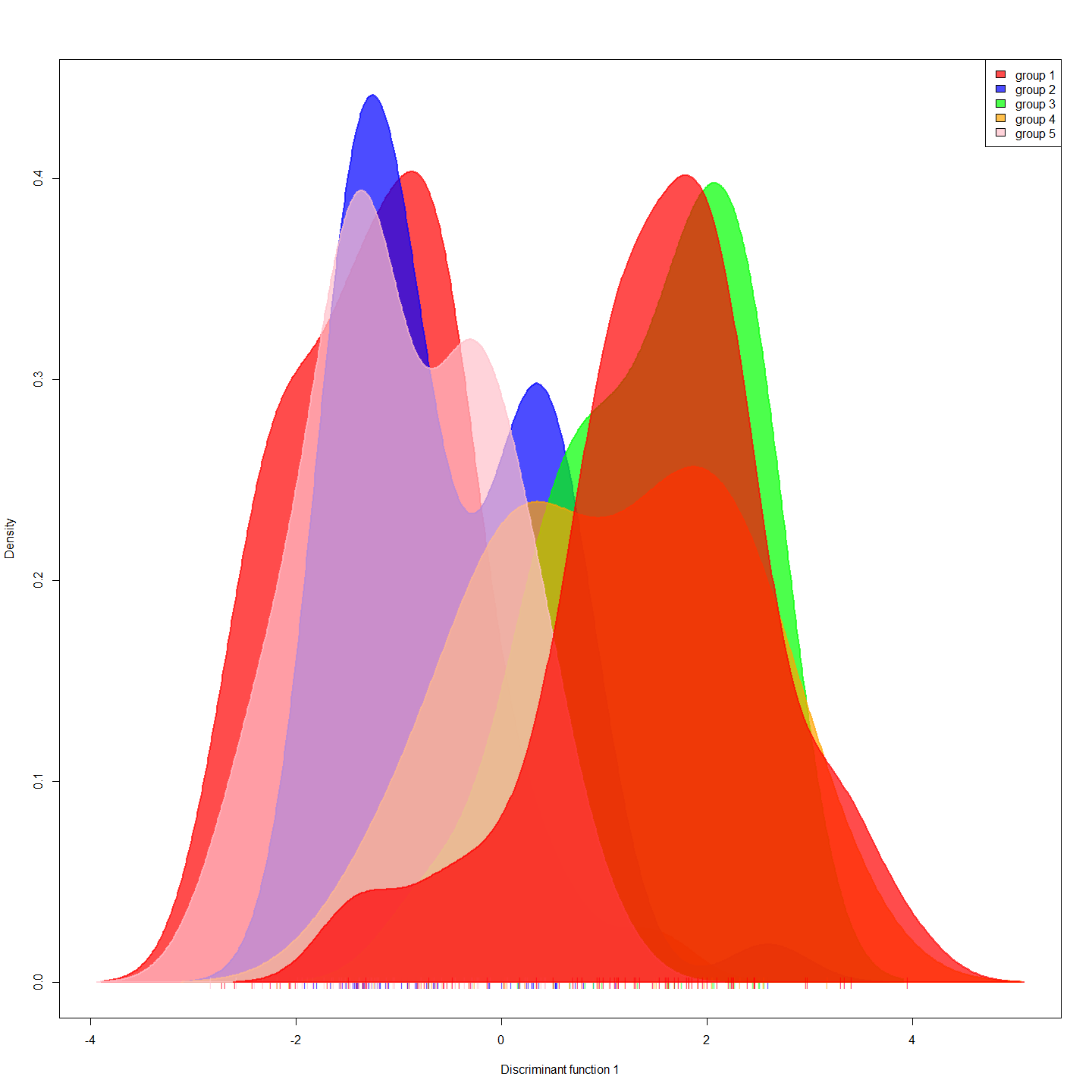
require(ape)  
tre <- nj(dist(as.matrix(gl)))  
tre

##   
## Phylogenetic tree with 243 tips and 241 internal nodes.  
##   
## Tip labels:  
## 185, 186, 187, 188, 189, 190, ...  
##   
## Unrooted; includes branch lengths.

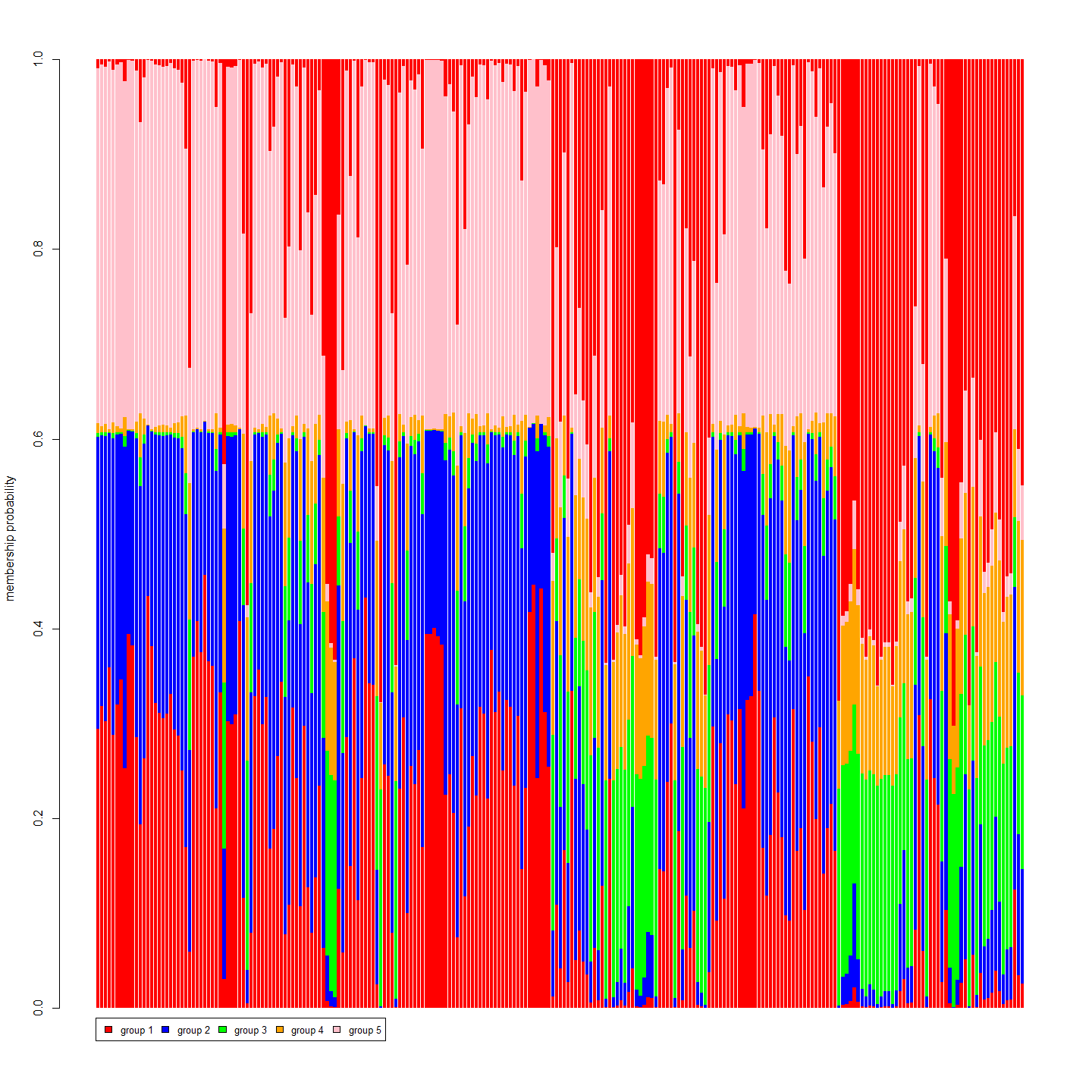
plot(tre, typ="fan", show.tip=FALSE)  
tiplabels(pch=20, col=myCol, cex=4)  
title("NJ tree of the Upper Murrumbidgee Murray Cod data")



#Discriminant Component Analysis   
dapc2 <- dapc(gl.dart, n.pca=10, n.da=1, parallel=FALSE)  
scatter(dapc2,scree.da=FALSE, bg="white", posi.pca="topright", legend=TRUE, txt.leg=paste("group", 1:5), col=c("red","blue","green","orange", "pink"))



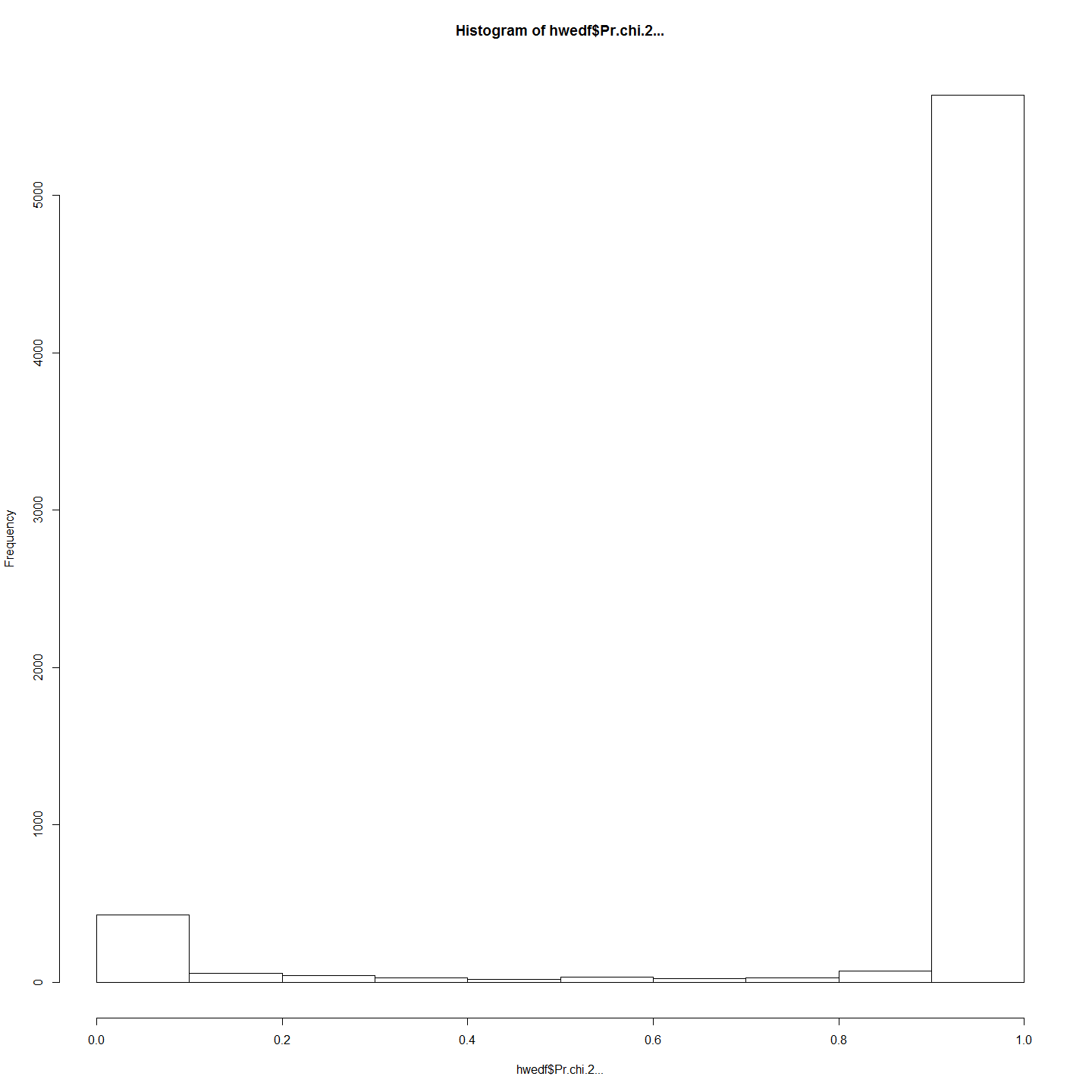
compoplot(dapc2, col=c("red","blue","green","orange", "pink"),lab="", txt.leg=paste("group", 1:5), ncol=5)



#Analysis on genInd object (individuals)  
#Hardy Weinberg  
source("genlight2genInd.R")  
gI<-genlight2genind(gl.dart)

## Start conversion....  
## Please note conversion of bigger data sets will take some time!  
## Once finished, we recommend to save the object using >save(object, file="object.rdata")  
##   
## Matrix converted.. Prepare genind object...  
## Finished! Took 77 seconds.

hwe<-hw.test(gI, B=0)  
hwedf<-data.frame(hwe)  
hist(hwedf$Pr.chi.2...)



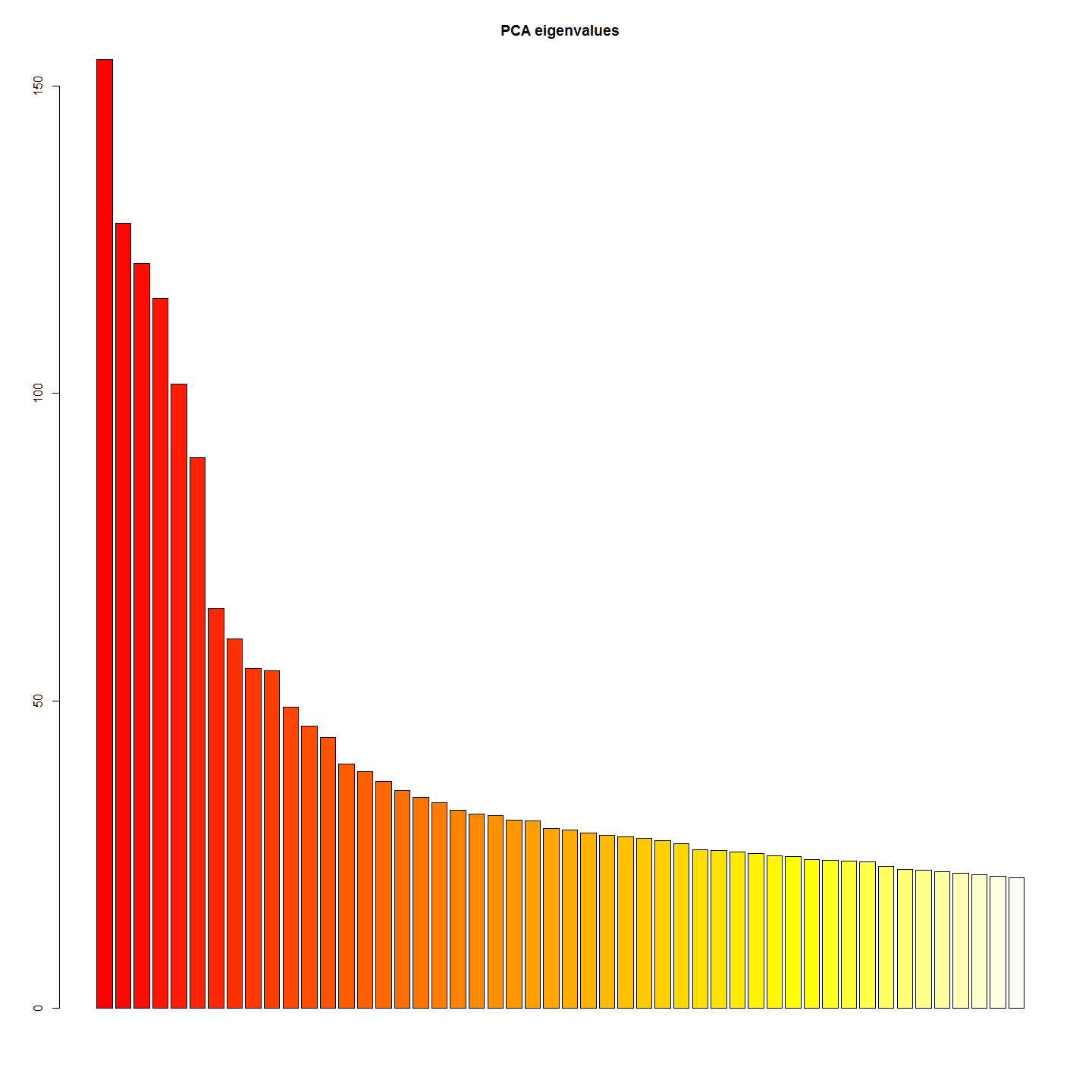
#There are   
sum(is.na(gI$tab)) #missing values in the data

## [1] 30511

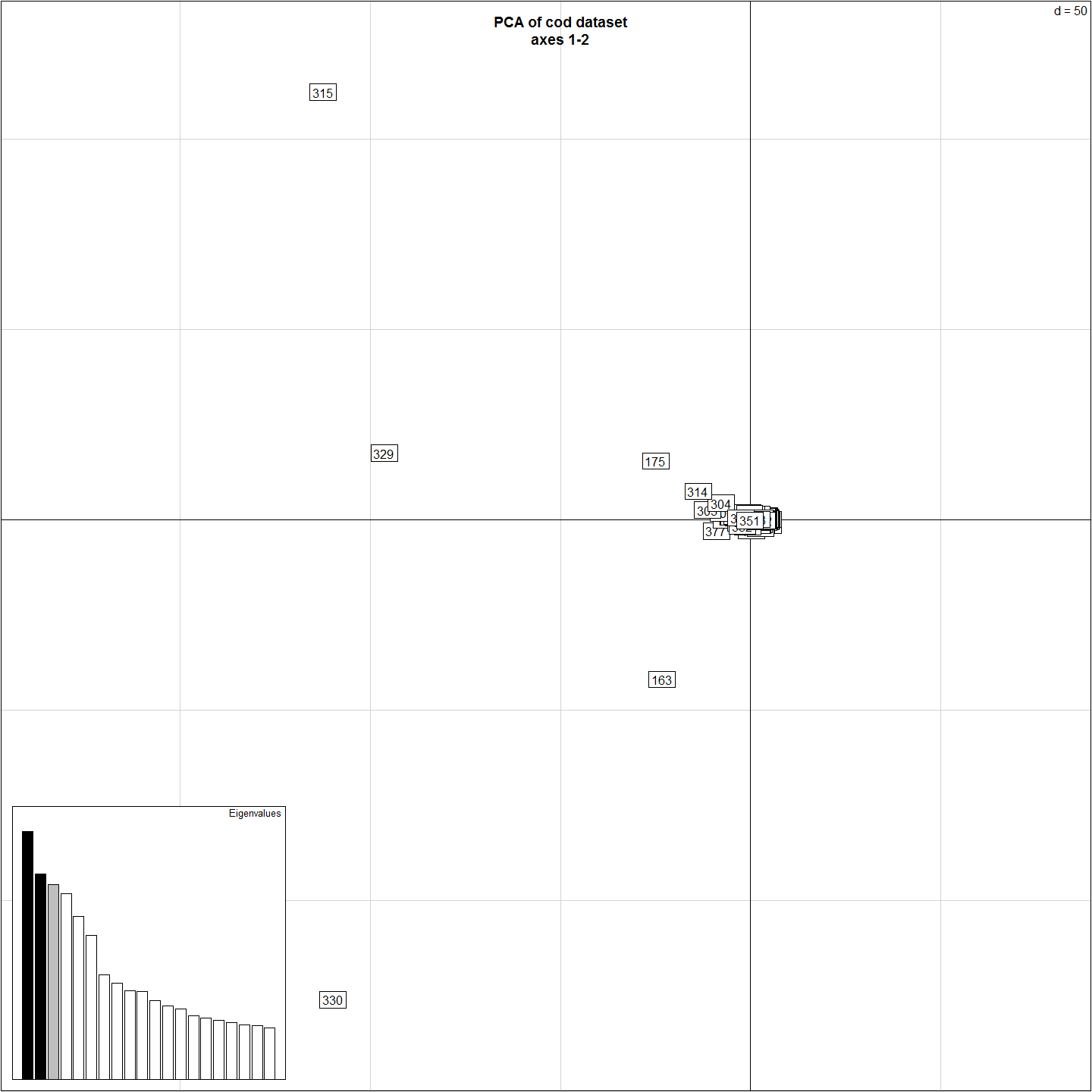
forPCOA<-scaleGen(gI, NA.method="mean") #scale the data  
forPCOA[1:5,1:5]

## 6661549-60-C/T.A 6661549-60-C/T.B 6650588-15-C/T.A 6650588-15-C/T.B 6661794-21-A/G.A  
## 185 2.006758 -2.006758 1.499011 -1.499011 0.1184337  
## 186 2.006758 -2.006758 1.499011 -1.499011 1.7172889  
## 187 0.289708 -0.289708 1.499011 -1.499011 0.1184337  
## 188 0.289708 -0.289708 1.499011 -1.499011 1.7172889  
## 189 0.289708 -0.289708 1.499011 -1.499011 0.1184337

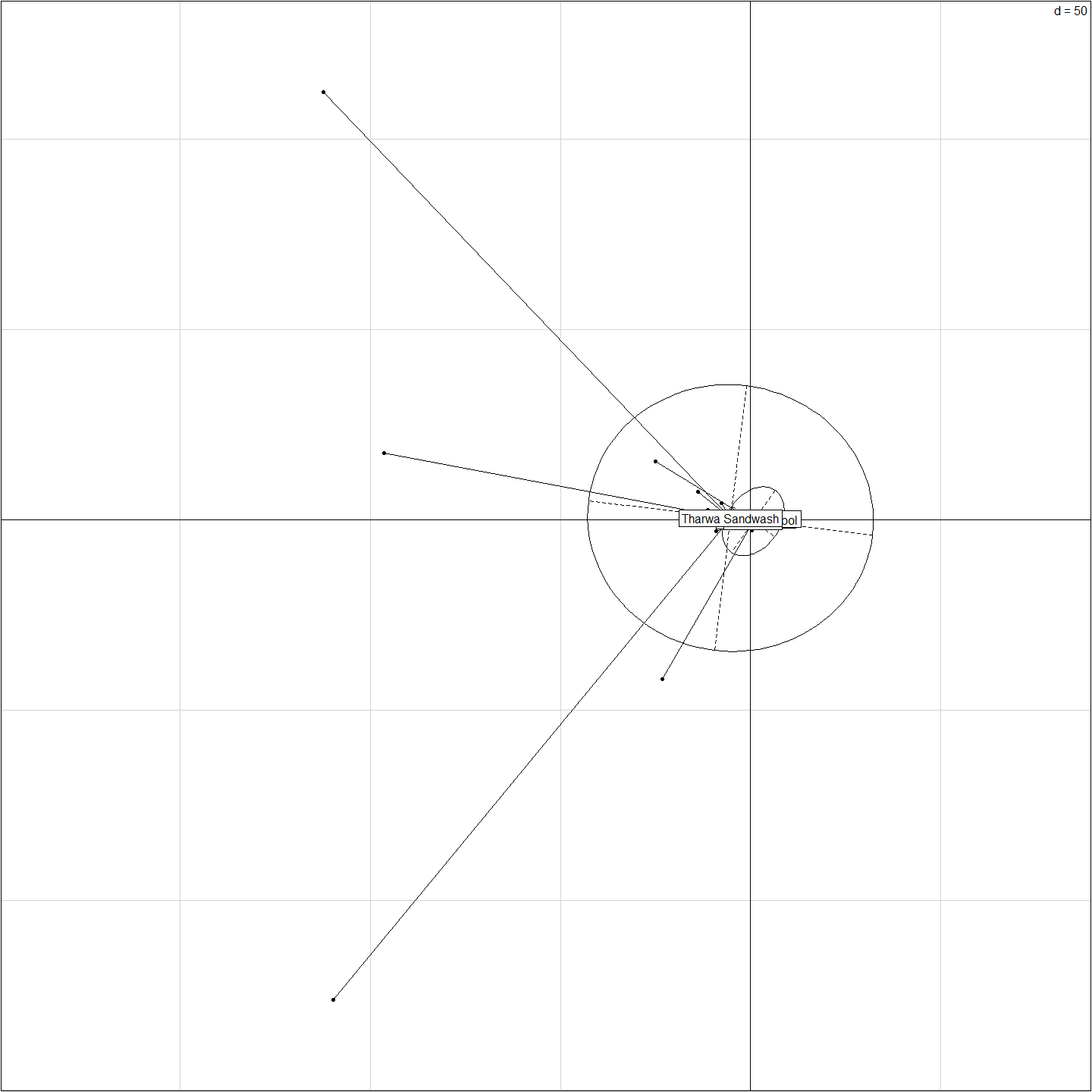
pca1 <- dudi.pca(forPCOA,cent=FALSE,scale=FALSE,scannf=FALSE,nf=3)  
barplot(pca1$eig[1:50],main="PCA eigenvalues", col=heat.colors(50))



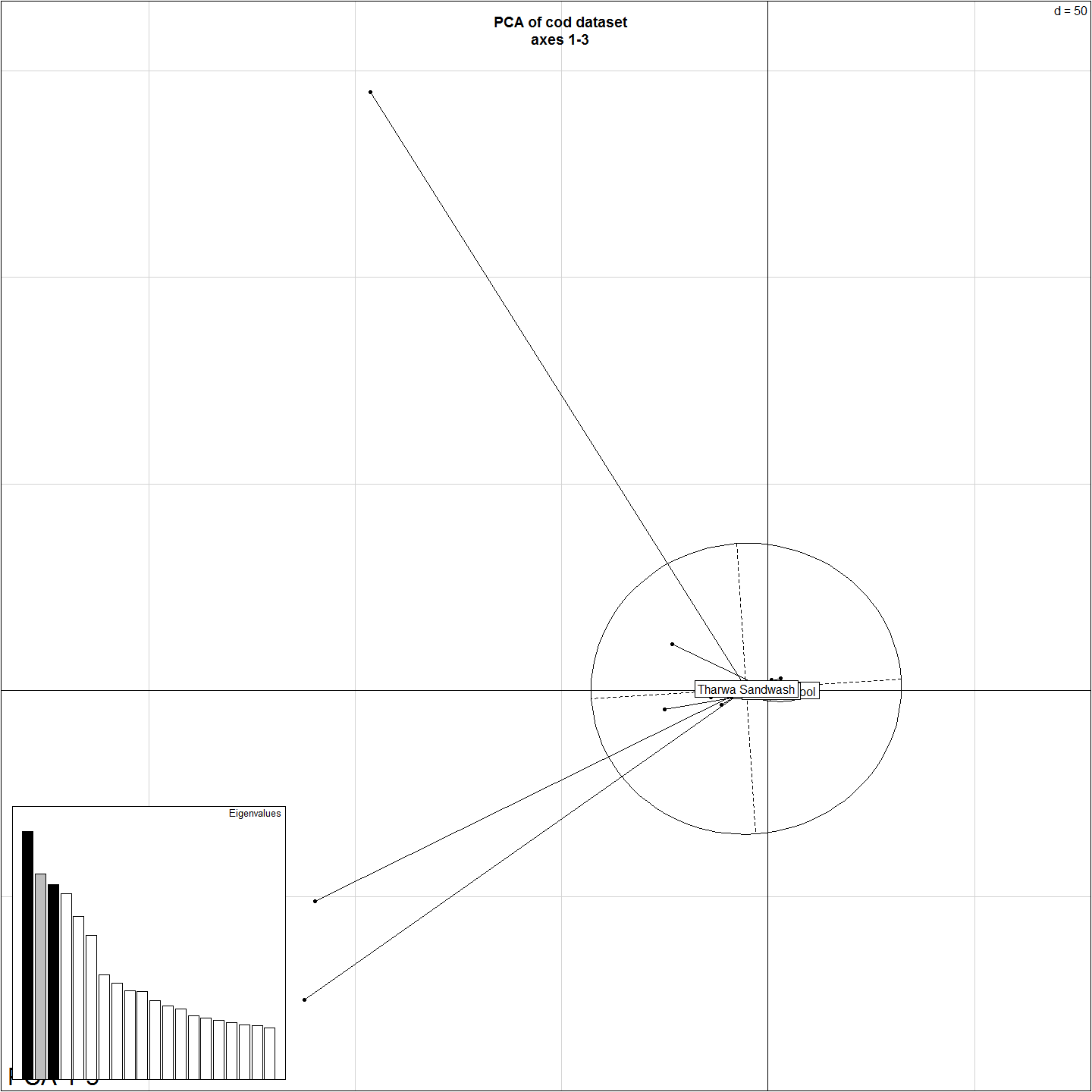
s.label(pca1$li)  
title("PCA of cod dataset\naxes 1-2")  
add.scatter.eig(pca1$eig[1:20], 3,1,2)



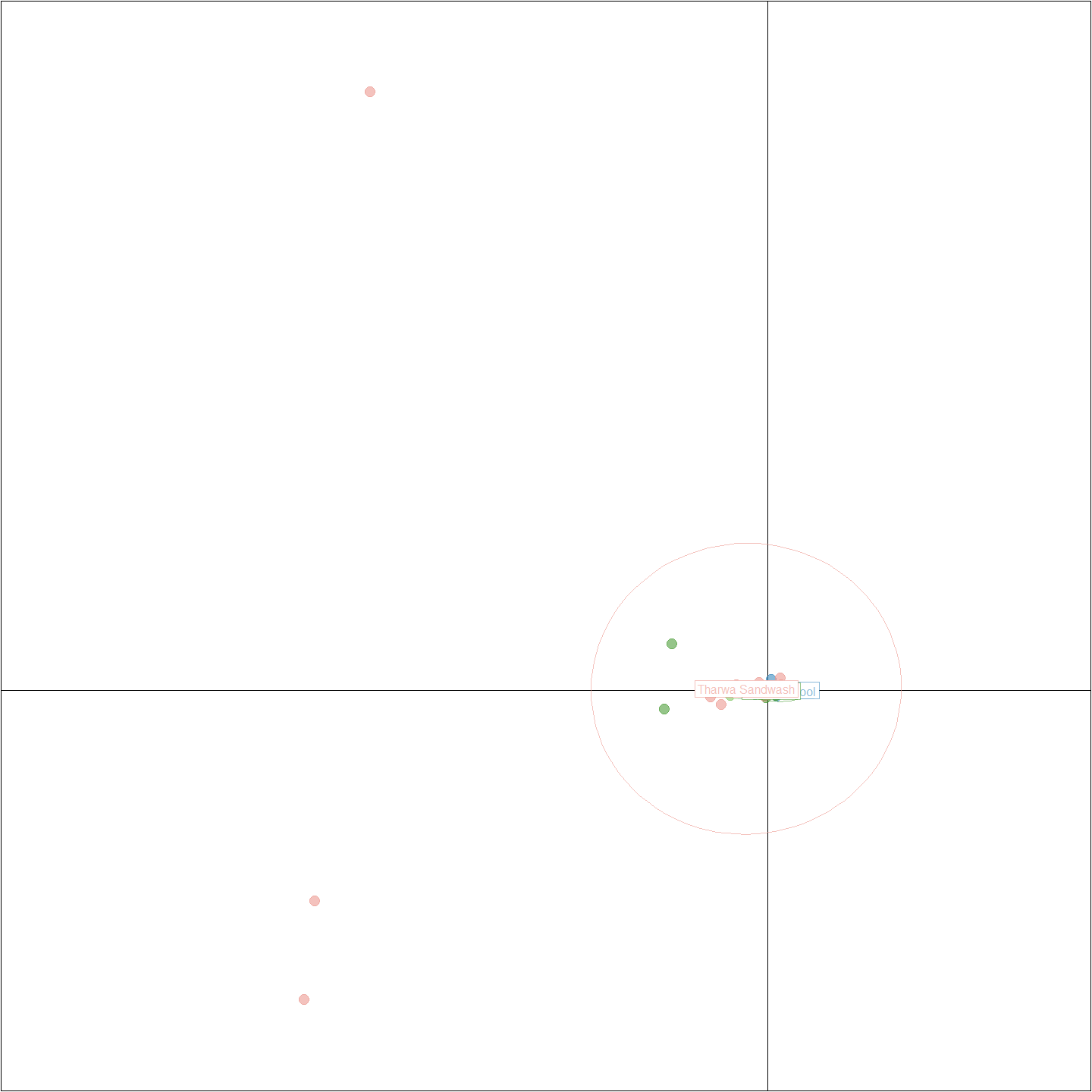
s.class(pca1$li, pop(gI))



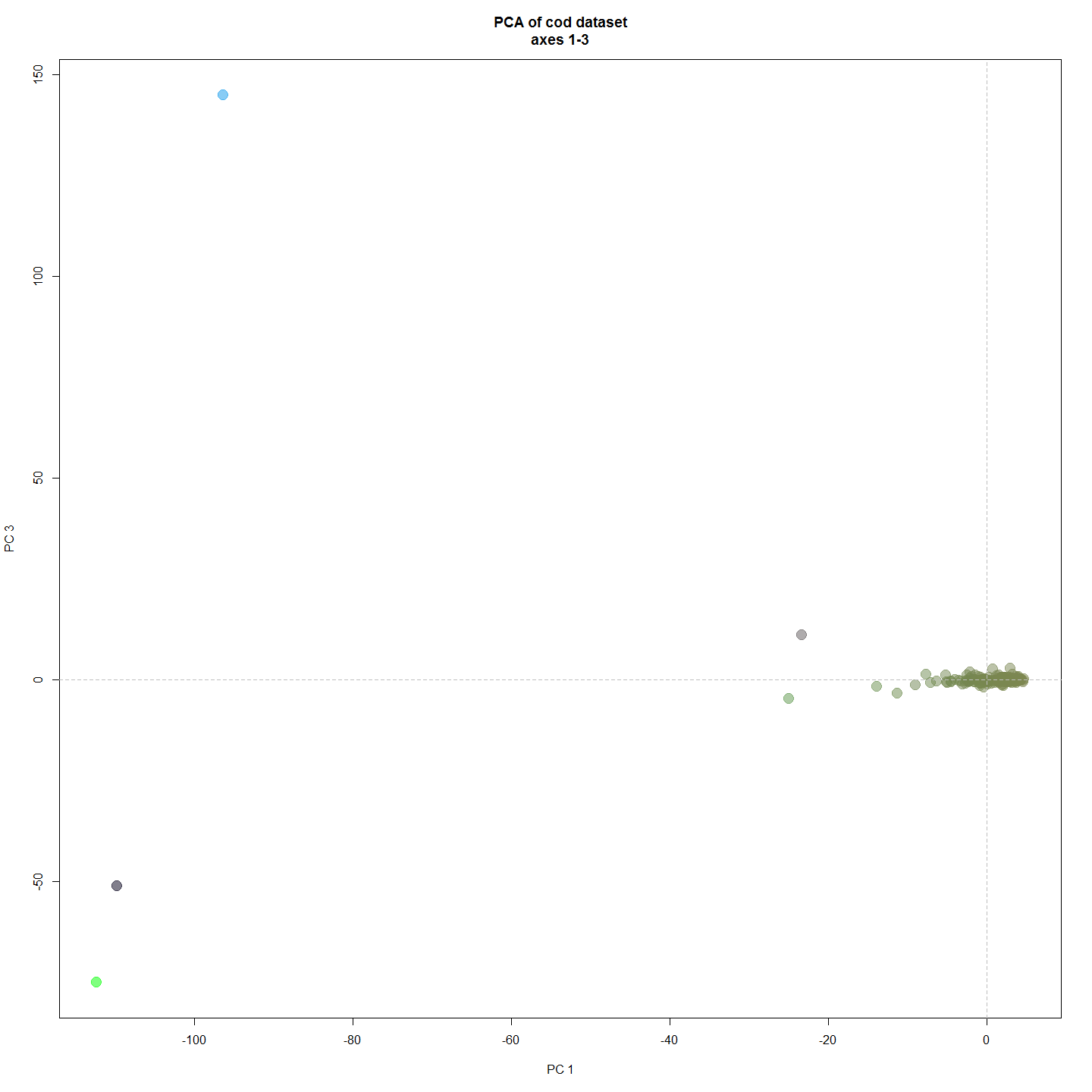
s.class(pca1$li,pop(gI),xax=1,yax=3,sub="PCA 1-3",csub=2)  
title("PCA of cod dataset\naxes 1-3")  
add.scatter.eig(pca1$eig[1:20],nf=3,xax=1,yax=3)



col <- funky(15)  
s.class(pca1$li, pop(gI),xax=1,yax=3, col=transp(col,.6), axesell=FALSE,  
 cstar=0, cpoint=3, grid=FALSE)



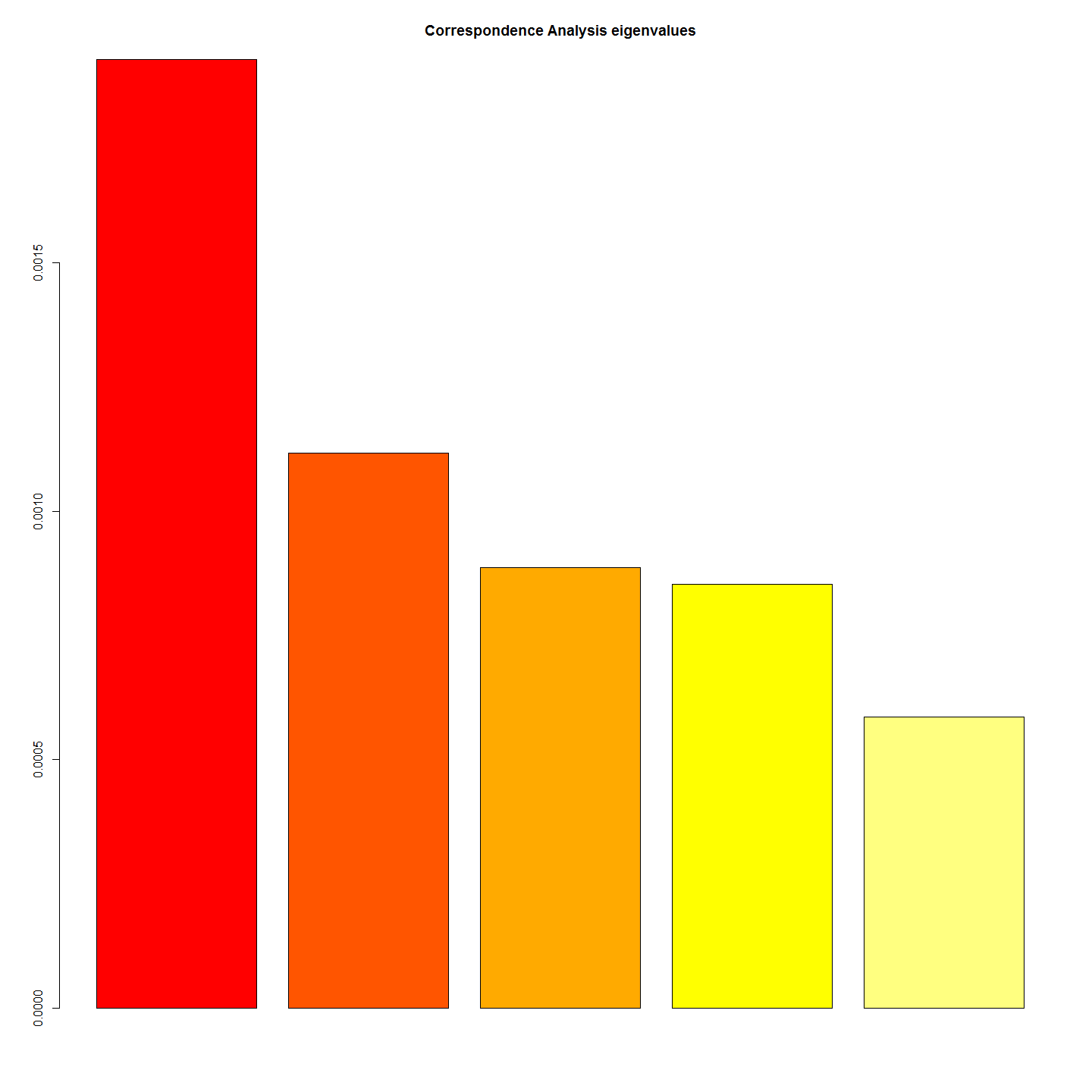
colorplot(pca1$li[c(1,3)], pca1$li, transp=TRUE, cex=3, xlab="PC 1", ylab="PC 3")  
title("PCA of cod dataset\naxes 1-3")  
abline(v=0,h=0,col="grey", lty=2)



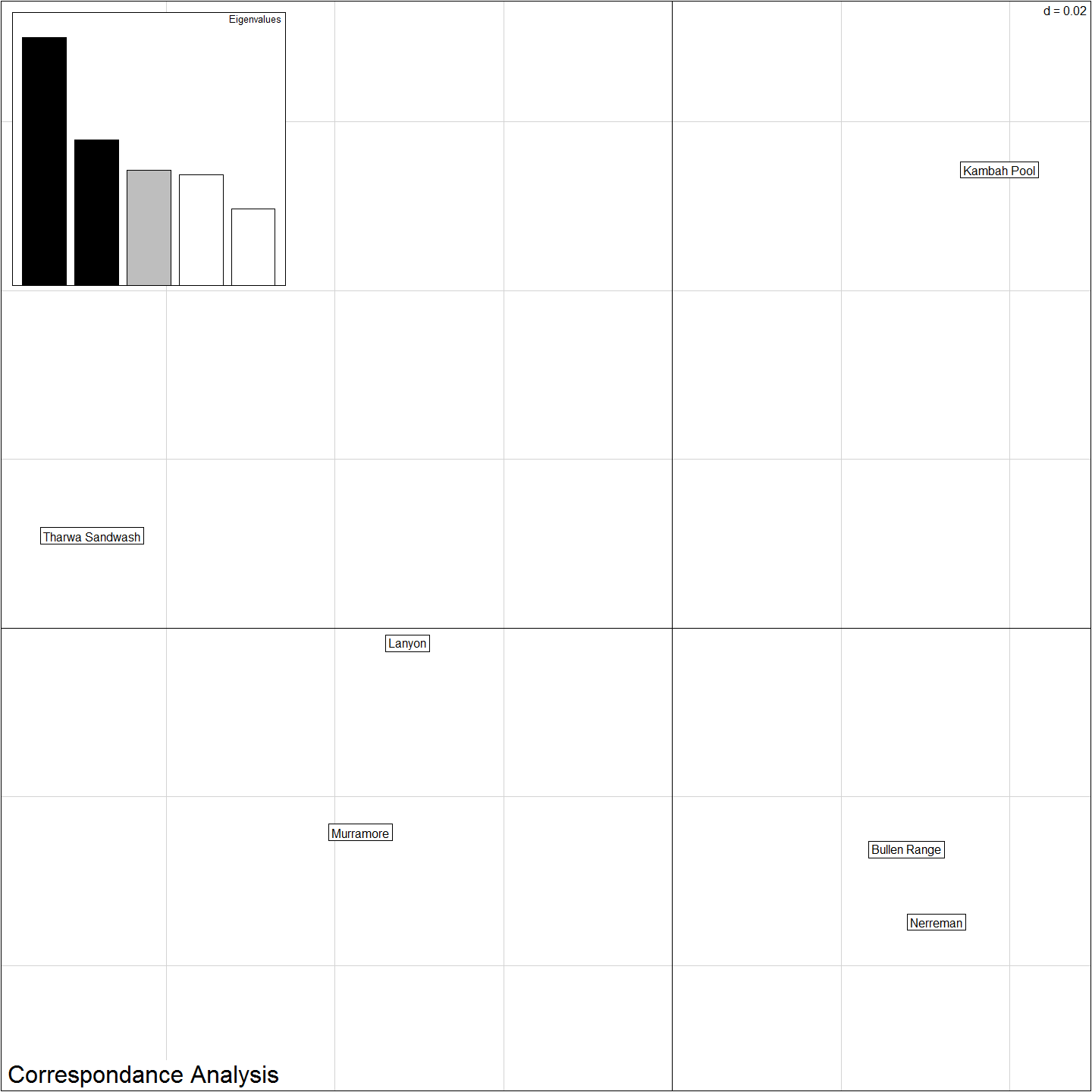
#Correspondance Analysis (genepop object)  
obj <- genind2genpop(gI)

##   
## Converting data from a genind to a genpop object...   
##   
## ...done.

ca1 <- dudi.coa(tab(obj),scannf=FALSE,nf=3)  
barplot(ca1$eig,main="Correspondence Analysis eigenvalues",  
 col=heat.colors(length(ca1$eig)))



s.label(ca1$li, sub="Correspondance Analysis",csub=2)  
add.scatter.eig(ca1$eig,nf=3,xax=1,yax=2,posi="topleft")



#Isolation By Distance  
# Dgen <- dist.genpop(obj,method=2)  
# Dgeo <- dist(gI$other$latlong) #need xys for sites in gI for this to work.  
# ibd <- mantel.randtest(Dgen,Dgeo)  
# ibd

all\_labels()

## [1] "Project\_Template\_and\_Knitr" "Set\_Global\_Options" "unnamed-chunk-1" "unnamed-chunk-2"   
## [5] "unnamed-chunk-3" "unnamed-chunk-4" "unnamed-chunk-5" "PCAtSNEStructure"   
## [9] "unnamed-chunk-6" "dapcCorrespondence" "Include\_Chunk\_Labels\_and\_Session Information" "averagePIC"   
## [13] "dapcCorroStructure"

proc.time()-ptm

## user system elapsed   
## 841.42 8.00 861.64

#Session Information  
sessionInfo()

## R version 3.3.0 (2016-05-03)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 7 x64 (build 7601) Service Pack 1  
##   
## locale:  
## [1] LC\_COLLATE=English\_Australia.1252 LC\_CTYPE=English\_Australia.1252 LC\_MONETARY=English\_Australia.1252 LC\_NUMERIC=C LC\_TIME=English\_Australia.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] pegas\_0.9 ape\_3.5 reshape2\_1.4.1 tsne\_0.1-2 dart\_0.3 adegenet\_2.0.1 ade4\_1.7-4 ggplot2\_2.1.0 plyr\_1.8.4   
## [10] dplyr\_0.4.3 ProjectTemplate\_0.6 knitr\_1.13   
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_0.12.5 spdep\_0.6-5 formatR\_1.4 LearnBayes\_2.15 tools\_3.3.0 boot\_1.3-18 digest\_0.6.9 evaluate\_0.9 gtable\_0.2.0 nlme\_3.1-128   
## [11] lattice\_0.20-33 Matrix\_1.2-6 igraph\_1.0.1 shiny\_0.13.2 DBI\_0.4-1 yaml\_2.1.13 parallel\_3.3.0 coda\_0.18-1 stringr\_1.0.0 gtools\_3.5.0   
## [21] grid\_3.3.0 R6\_2.1.2 rmarkdown\_0.9.6 sp\_1.2-3 gdata\_2.17.0 seqinr\_3.1-5 deldir\_0.1-12 magrittr\_1.5 gmodels\_2.16.2 splines\_3.3.0   
## [31] scales\_0.4.0 htmltools\_0.3.5 MASS\_7.3-45 assertthat\_0.1 mime\_0.4 colorspace\_1.2-6 xtable\_1.8-2 httpuv\_1.3.3 labeling\_0.3 KernSmooth\_2.23-15  
## [41] stringi\_1.1.1 lazyeval\_0.1.10 munsell\_0.4.3