LDA.R

rob_c

Fri Apr 28 13:16:42 2017

```r
# SUMMARY:
## Al-Yotm distinct from Dead Sea and Kufranjah valley sites
## Dead Sea and Kufranjah are different ON AVERAGE, but a lot of overlap

# Simplified/clean theme for plotting
theme_simple <- function (base_size = 12, base_family = "") {
  theme_classic(base_size = base_size, base_family = base_family) %+replace%
    theme(
      axis.text = element_text(colour = "black"),
      axis.title.x = element_text(size=18),
      axis.text.x = element_text(size=12),
      axis.title.y = element_text(size=18,angle=90),
      axis.text.y = element_text(size=12),
      axis.ticks = element_blank(),
      panel.background = element_rect(fill="white"),
      panel.border = element_blank(),
      plot.title=element_text(face="bold", size=24),
      legend.position="none"
    )
}



# NOTES:
## Plotting PC results using ggplot2
## https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

library(ggplot2)
library(ggfortify)
```

```r
## Warning: package 'ggfortify' was built under R version 3.3.3
```

```r
library(MASS)

## Import Data
MorphData<-read.csv("MuhaidatEtAl_RawData.csv",header=T)
str(MorphData)
```

```
## 'data.frame':    57 obs. of  15 variables:
##  $ Loc      : Factor w/ 3 levels "AlYotm","DeadSea",..: 3 3 3 3 3 3 3 3 3
3 ...
##  $ BLen     : num  4.8 5 5.4 5.6 5.5 5.2 5.9 5.7 4.9 5.9 ...
##  $ BWdth    : num  0.4 0.5 0.4 0.4 0.4 0.5 0.4 0.5 0.4 0.5 ...
##  $ BVeins   : int  3 5 NA 3 3 3 3 5 3 5 ...
##  $ BSpines  : int  2 2 2 2 2 2 3 2 2 2 ...
##  $ SpLen    : num  1 1 1 1.2 1.1 0.9 0.9 1.3 1 1.5 ...
##  $ NodeLen  : num  1.3 1.6 1.5 1.6 1.3 1.8 1.8 1.6 1.6 2 ...
##  $ LLen     : num  7 7 9 8.5 7.5 10 8 8 9.1 10 ...
##  $ LWdth    : num  1 1 1.1 0.9 0.7 1.2 1 1.2 0.8 1.2 ...
##  $ LTeeth   : int  4 5 4 4 5 3 4 NA 5 4 ...
##  $ ASFilLen : num  1.3 1.3 1.2 1.3 1.2 1.3 1.3 1.2 1.3 1.3 ...
##  $ ASAnthLen: num  0.5 0.5 0.5 0.5 0.4 0.5 0.5 0.5 0.6 0.5 ...
##  $ ASApenLen: num  0.5 0.5 0.6 0.5 0.5 0.5 0.6 0.5 0.6 0.5 ...
##  $ PSFilLen : num  1.2 NA 1.1 1.2 1.1 1.1 1.1 1.4 1.2 1.2 ...
##  $ PSAnthLen: num  0.5 0.5 0.5 0.5 0.5 NA 0.6 0.5 0.6 0.4 ...
```

```r
## Add midvalue for missing data
for(Row in 1:nrow(MorphData)){
  for(Col in 2:ncol(MorphData)){
    if(is.na(MorphData[Row,Col])){
      MorphData[Row,Col]<-mean(MorphData[MorphData$Loc==MorphData$Loc[Row],Co
l],na.rm=T)
    }
  }
}


## Inspect pairwise scatterplots
pairs(MorphData,col=rgb(0,0,0,0.3),pch=16)
```
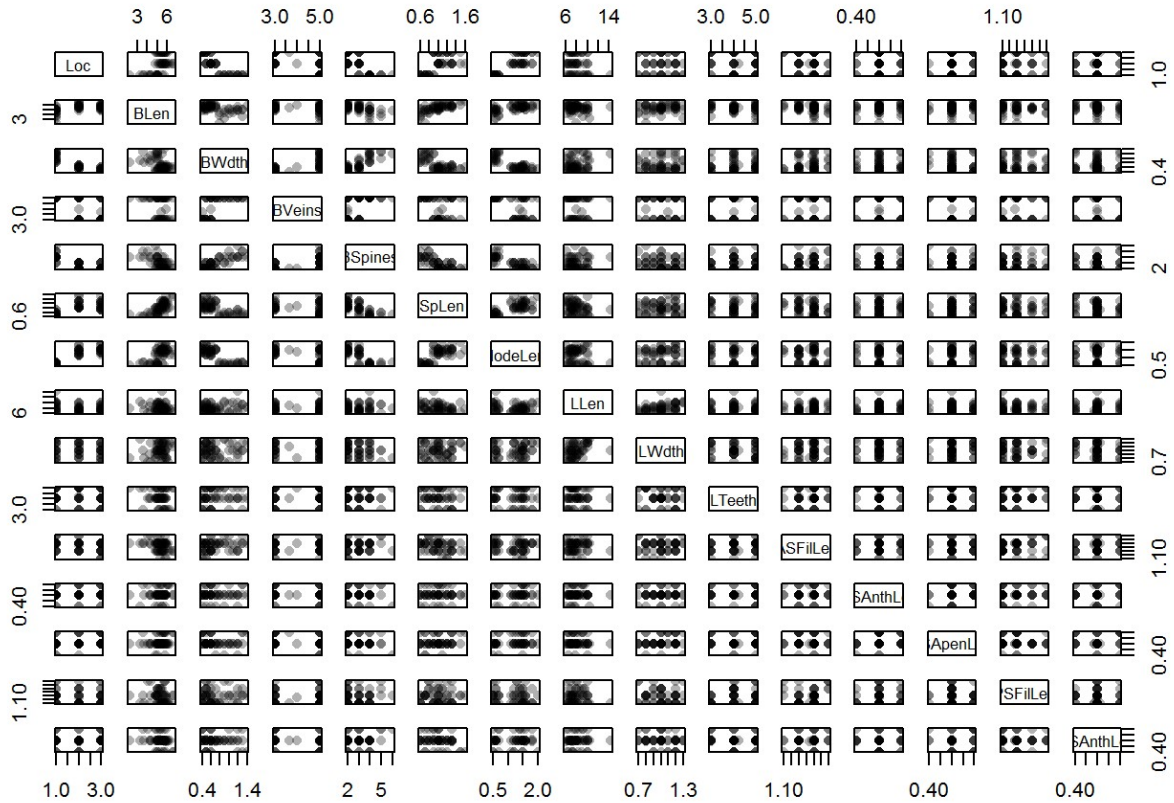
```
## Scale data to mean and sd prior to analysis
scale<-function(x){
  return((x-mean(x,na.rm=T))/sd(x,na.rm=T))
}
MorphScaled<-data.frame(sapply(MorphData[,2:ncol(MorphData)],scale))
MorphScaled$Loc<-as.factor(MorphData$Loc)

# Linear discriminant function analysis
(BlephLDA<-lda(Loc ~ ., data=MorphScaled))
```

```
## Call:
## lda(Loc ~ ., data = MorphScaled)
##
## Prior probabilities of groups:
##     AlYotm    DeadSea Kufranjah
## 0.3333333 0.3333333 0.3333333
##
## Group means:
##                BLen       BWdth      BVeins     BSpines       SpLen
## AlYotm    -0.9082828   1.2385737   0.8048484   1.1492857  -1.0936903
## DeadSea    0.5943910  -0.3825007  -0.2299567  -0.4701623   0.6714428
## Kufranjah  0.3138919  -0.8560730  -0.5748917  -0.6791234   0.4222475
##               NodeLen        LLen        LWdth       LTeeth     ASFilLen
## AlYotm    -1.2934771   0.34813637  -9.493046e-17   0.11281514   0.3438730
## DeadSea    0.5197006  -0.24816478   1.543217e-01   0.05707119   0.1114529
## Kufranjah  0.7737765  -0.09997159  -1.543217e-01  -0.16988633  -0.4553259
##              ASAnthLen    ASApenLen     PSFilLen    PSAnthLen
## AlYotm    -0.06452092  -0.24348747   0.09745486   0.003838539
## DeadSea    0.03226046  -0.03478392  -0.07796389  -0.203442582
## Kufranjah  0.03226046   0.27827139  -0.01949097   0.199604042
##
## Coefficients of linear discriminants:
##                    LD1          LD2
## BLen        0.56800326  -0.198396091
## BWdth      -1.64395526  -0.486772238
## BVeins     -0.25508231   0.482271419
## BSpines    -0.03403730  -0.309409345
## SpLen      -0.07400690  -1.593549477
## NodeLen     2.30851067   0.700925267
## LLen       -0.10456205   0.349062218
## LWdth      -0.18501256  -0.458459037
## LTeeth     -0.13656424  -0.343481761
## ASFilLen   -0.37778904  -0.799127697
## ASAnthLen  -0.08513061  -0.178142530
## ASApenLen   0.18021581   0.596656287
## PSFilLen    0.22693391   0.418529527
## PSAnthLen   0.01028254   0.004758835
##
## Proportion of trace:
##    LD1    LD2
## 0.9732 0.0268
```

```
## Extract scaling vectors
scalvec<-data.frame(BlephLDA$scaling)

## Extract predictions
BlephLDAval <- data.frame(predict(BlephLDA)$x)
ldahist(data = BlephLDAval[,1], g=MorphScaled$Loc)
```


group AlYotm

group DeadSea

group Kufranjah
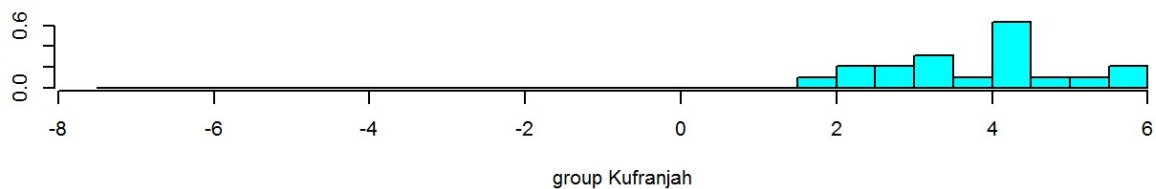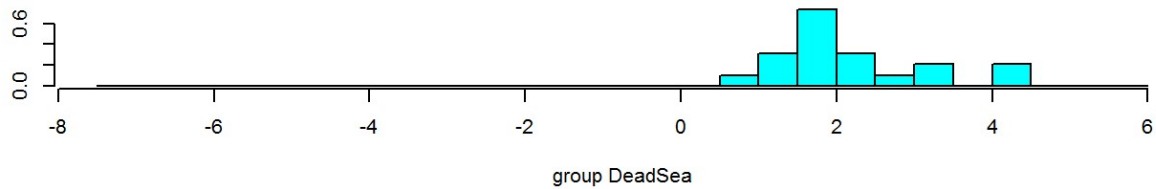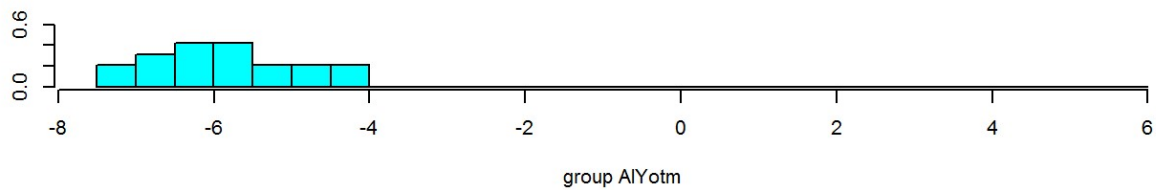
```
BlephLDAval$Loc<-MorphScaled$Loc

## Plot results
p<-ggplot(data=BlephLDAval,aes(x=LD1,y=LD2,group=Loc))+
  stat_ellipse(geom="polygon",aes(colour=Loc),fill=NA,size=1.2,alpha=0.3)+
  stat_ellipse(geom="polygon",aes(fill=Loc,colour=Loc),size=1.2,alpha=0.3)+
  geom_point(aes(shape=Loc,fill=Loc,colour=Loc),size=I(4),alpha=I(0.8))+
  xlab("LD Axis 1")+ylab("LD Axis 2")+theme_simple()

print(p)
```
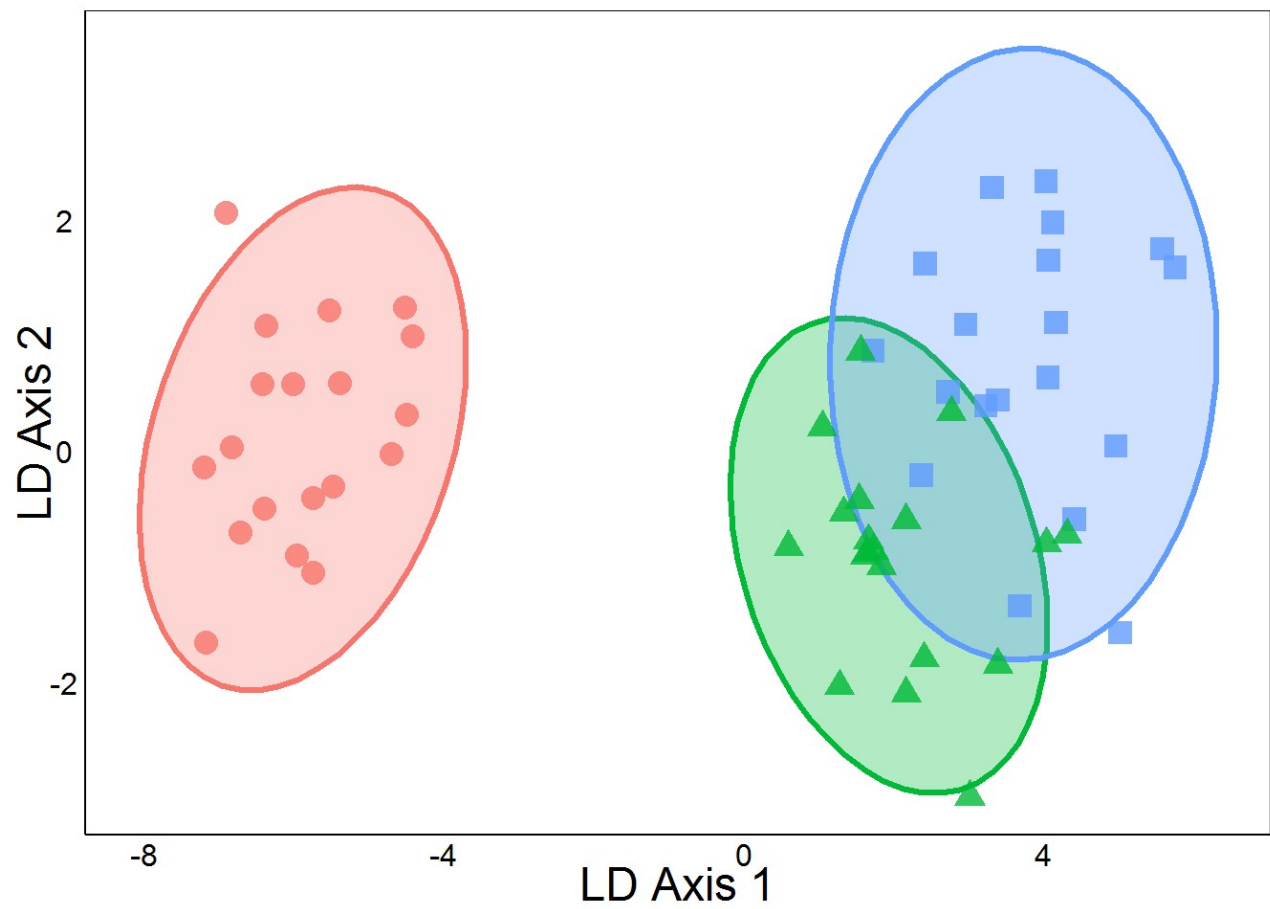
```
pdf("LDAplot.pdf",width=6,height=6)
  print(p)
dev.off()
```

```
## png
##   2
```

```
## Output table to csv
BlephLDAscales<-data.frame(round(BlephLDA$scaling,3))
BlephLDAscales$desc<-c("Bract Length","Bract Width","Veins per bract","Lateral
spines","Longest spine length","Internode length",
                       "Leaf length","Leaf width","Teeth per leaf","Filament le
ngth","Anther length","Appendage length",
                       "Filament length","Anther length")

write.csv(BlephLDAscales,"LDAscales.csv",row.names=T)

## Test significance of LD Axes
anova(lm(BlephLDAval$LD1~MorphScaled$Loc))
```

```
## Analysis of Variance Table
##
## Response: BlephLDAval$LD1
##                Df Sum Sq Mean Sq F value    Pr(>F)
## MorphScaled$Loc  2 1018.7  509.36  509.36 < 2.2e-16 ***
## Residuals       54   54.0    1.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(BlephLDAval$LD2~MorphScaled$Loc))
```

```
## Analysis of Variance Table
##
## Response: BlephLDAval$LD2
##                Df Sum Sq Mean Sq F value    Pr(>F)
## MorphScaled$Loc  2 28.096  14.048  14.048 1.224e-05 ***
## Residuals       54 54.000   1.000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Plot trait vectors
BlephLDAscales$Loc<-NA
s<-2 # scale for vector
k<-0.2# Keep this proportion of largest vectors

keep<-abs(BlephLDAscales$LD1)>=sort(abs(BlephLDAscales$LD1),decreasing=T)[floor(nrow(BlephLDAscales)*k)] |
  abs(BlephLDAscales$LD2)>=sort(abs(BlephLDAscales$LD2),decreasing=T)[floor(nrow(BlephLDAscales)*k)]
p+geom_segment(data=BlephLDAscales[keep,],aes(x=0,xend=LD1*s,y=0,yend=LD2*s),
               arrow = arrow(length = unit(0.5, "cm")),colour="grey",inherit_aes=FALSE)+
    geom_text(data=BlephLDAscales[keep,],aes(x=LD1*s,y=LD2*s,label=desc),size=5)#+coord_fixed()
```

```
## Warning: Ignoring unknown parameters: inherit_aes
```
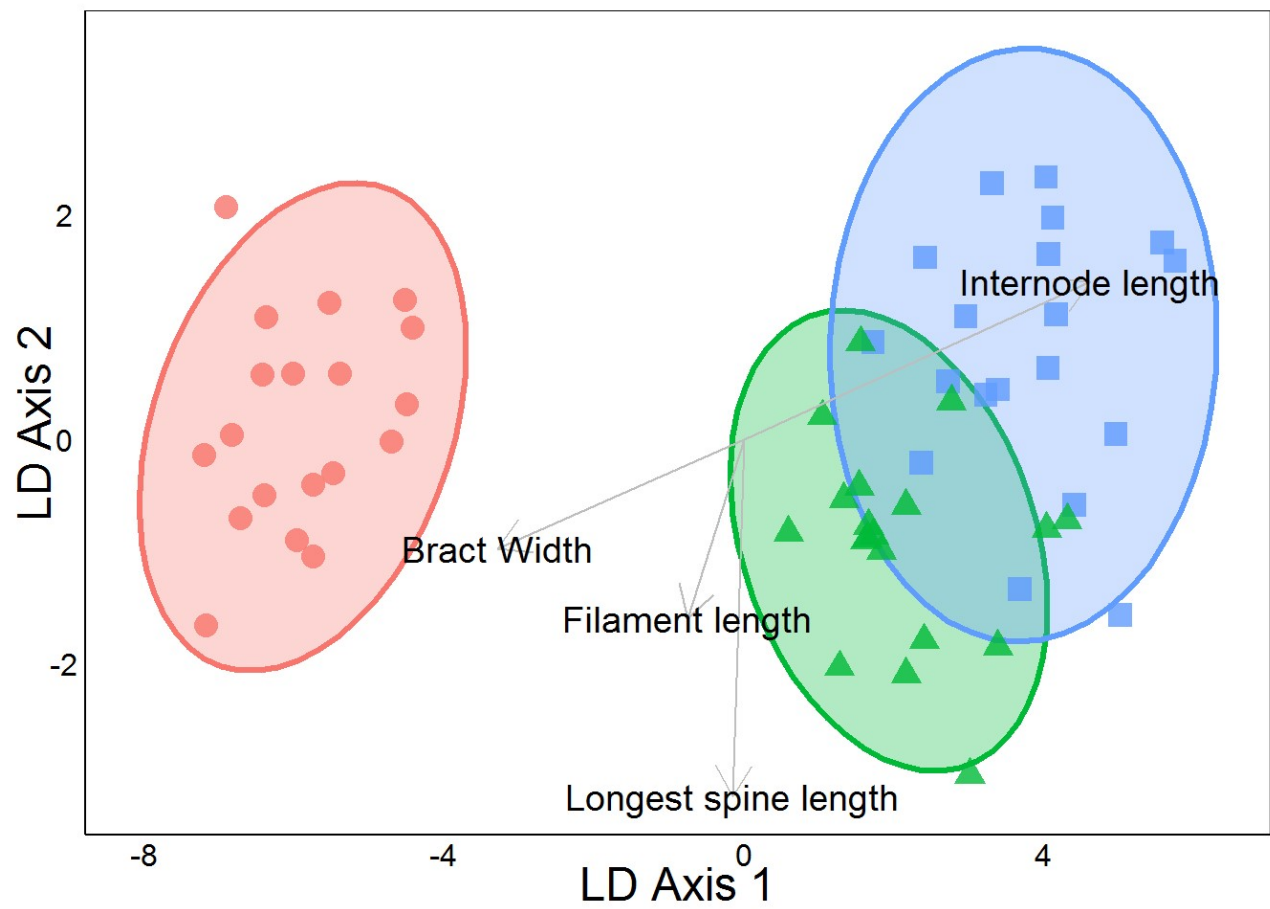
```
## Software version info
sessionInfo()
```

```
## R version 3.3.2 (2016-10-31)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 14393)
##
## locale:
## [1] LC_COLLATE=English_Canada.1252  LC_CTYPE=English_Canada.1252
## [3] LC_MONETARY=English_Canada.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Canada.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] MASS_7.3-45    ggfortify_0.4.1 ggplot2_2.2.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.9      knitr_1.15.1      magrittr_1.5     munsell_0.4.3
##  [5] colorspace_1.3-2 R6_2.2.0          stringr_1.1.0    plyr_1.8.4
##  [9] dplyr_0.5.0      tools_3.3.2       grid_3.3.2       gtable_0.2.0
## [13] DBI_0.6-1        htmltools_0.3.5   lazyeval_0.2.0   rprojroot_1.2
## [17] digest_0.6.12    assertthat_0.1    tibble_1.2       gridExtra_2.2.1
## [21] tidyr_0.6.1      evaluate_0.10     rmarkdown_1.3    labeling_0.3
## [25] stringi_1.1.2    scales_0.4.1      backports_1.0.5
```