Project 2 Notebook

Introduction

Scientific Question

hfi hihiao

```
#for reading in fasta files
library("BiocManager")
#for reading in excel files
library("readxl")
#forgot
library("seqinr")
#for multiple sequence alignment
library("msa")
```

```
## Loading required package: Biostrings
```

```
## Loading required package: BiocGenerics
```

```
##
## Attaching package: 'BiocGenerics'
```

```
## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs
```

```
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
## Loading required package: S4Vectors
```

```
## Loading required package: stats4
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname
```

```
## Loading required package: IRanges
```

```
## Loading required package: XVector
```

```
## Loading required package: GenomeInfoDb
```

```
##
## Attaching package: 'Biostrings'

## The following object is masked from 'package:seqinr':
##
##     translate

## The following object is masked from 'package:base':
##
##     strsplit

##
## Attaching package: 'msa'

## The following object is masked from 'package:BiocManager':
##
##     version
```

```
#for msa pretty print
library("tinytex")
#visualization of results
library("ggplot2")
#for clustering of DNA seqs
library("DECIPHER")
```

```
## Loading required package: RSQLite

## Loading required package: parallel
```

```
knitr::include_graphics("dog/Basenji.jpg")
```

Figure 1: Basenji (S)

```
knitr::include_graphics("dog/Boxer.jpg")
```

```
knitr::include_graphics("dog/greatdane.jpg")
```

```
knitr::include_graphics("dog/labret.jpg")
```

```
knitr::include_graphics("dog/german_sheperd.jpg")
```

```
#global variable
alignment_name<<-""
#notebook functions

#align fasta from file_name with names from name file (visualization purposes)
#after alignment displays msaprettyprint results for human readable data
mult_alingments<-function(file_name,fasta_names,name){
  #read in fasta for all dogs
  string_set<-readDNAStringSet(file=file_name,use.names=FALSE)
```

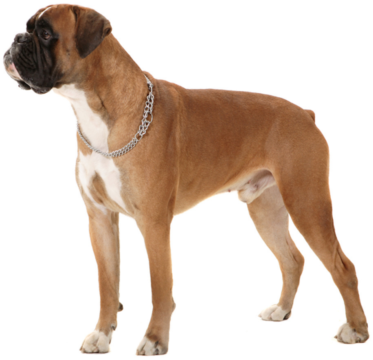


Figure 2: Boxer (L)

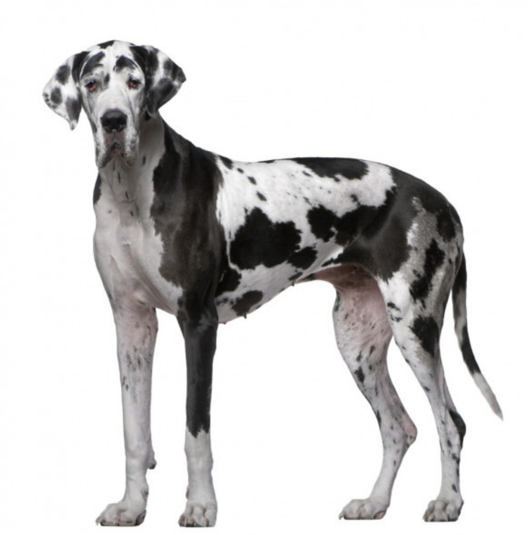


Figure 3: Great Dane (XL)

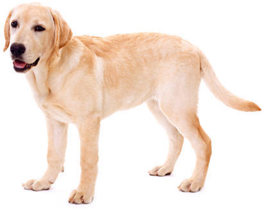


Figure 4: Golden Retriever (L)

Figure 5: German Shepherd (L)

```
  #read in seq names as list
  table=read.table(fasta_names, header = FALSE, sep = "\n")[["V1"]]
  #update names for pretty print
  names(string_set)<-table
  #align unnamed seqs
  alignment<-msa(string_set)
  #update global variable so multiple pretty print runs dont overrun eachother
  alignment_name<<-gsub(" ", "", paste(name,".pdf"), fixed = TRUE)
  #return pretty alignment, does not show up on my console
  msaPrettyPrint(alignment, file=alignment_name,output="pdf", showNames="right",showLogo="top",askForOve
  return(alignment_name)
}
#have figure with white background, no gridline and only axis ticks, no lines
tune_figure<-function(fig,addons){
  return(fig+theme_minimal()+theme(
    plot.background = element_blank(),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    panel.border = element_blank()))
}
#create dendogram based on fasta files, names of items clustered in fasta_names, fig_title is for fig
create_dendogram<-function(fasta_path, fasta_names, fig_title){
  dna <- string_set<-readDNAStringSet(file=fasta_path,use.names=FALSE)
  names(dna)=read.table(fasta_names, header = FALSE, sep = "\n")[["V1"]]
  d1 <- DistanceMatrix(dna, type="dist")
  dendogram<-IdClusters(d1, method="complete", cutoff=0.05, showPlot=FALSE,
                        type="dendrogram")
  nodePar <- list(lab.cex = 0.6, pch = c(NA, 19),
                cex = 0.7, col = "black")
  plot(as.dendrogram(dendogram), ylab = "Height", nodePar =
         nodePar,main=fig_title)
}
```

LCORL Analysis

```
#LCORL CALL
alignment<-mult_alingments("fasta/LCORL_file.txt","fasta/names.txt","LCORL")
```

```
## use default substitution matrix
```

logo

```
..CCATTCCGCCA..  great dane (XL)
.ACAATTTCGTT...  golden retriever (L)
ATTCATAGAGT....  german sheperd (L)
...TGCTGTGCAAG.  boxer ref (L)
...TGCTCCCTGGG.  basenji(S)
....GAAGAAAAAAA  boxer noref (L)
     ****** *    consensus
```

```
X  non-conserved
X  ≥50% conserved
```

```
#IGF1 CALL
alignment<-mult_alingments("fasta/igf1.fasta","fasta/igf1_names.txt","igf1")
```

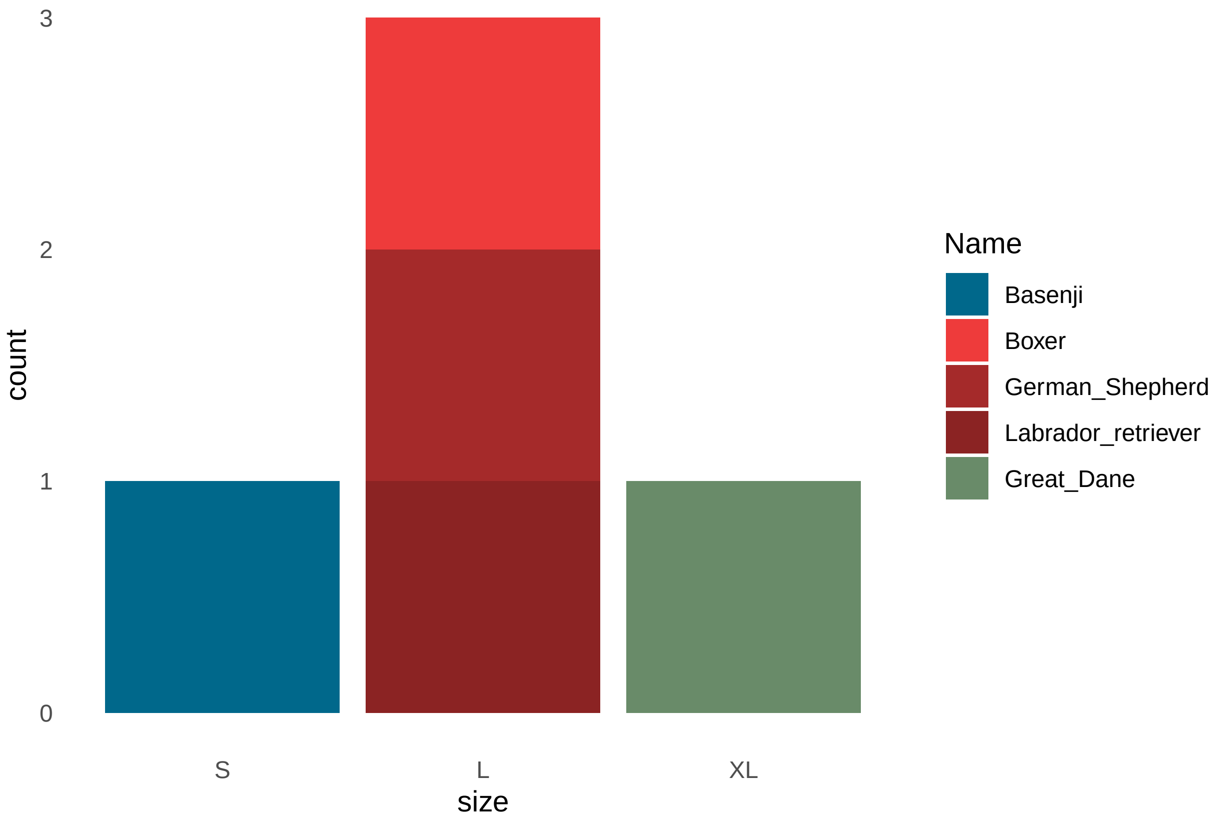
```
## use default substitution matrix
```

logo

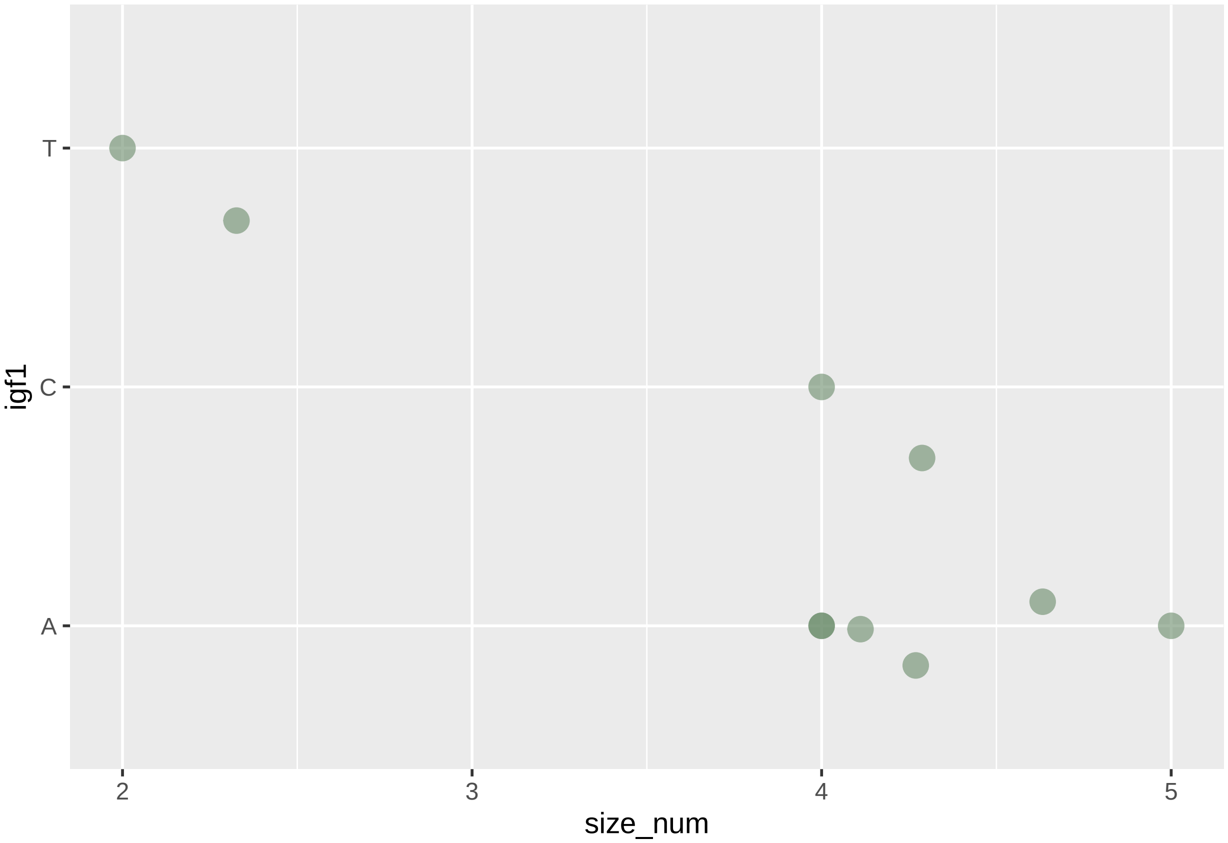
```
.GGGCCCGGCTG...  Boxer ref (L)
AGGTCATGACT....  Great dane (XL)
..GGCATTCCCCT..  Basenji (S)
....TCTGAAGAGTA  German shepherd (L)
..AATTCAGTGAA..  Labrador retriever (L)
TTGCTTTTGTA....  Boxer no ref (L)
  *    ****      consensus
```

```
X  non-conserved
X  ≥50% conserved
```

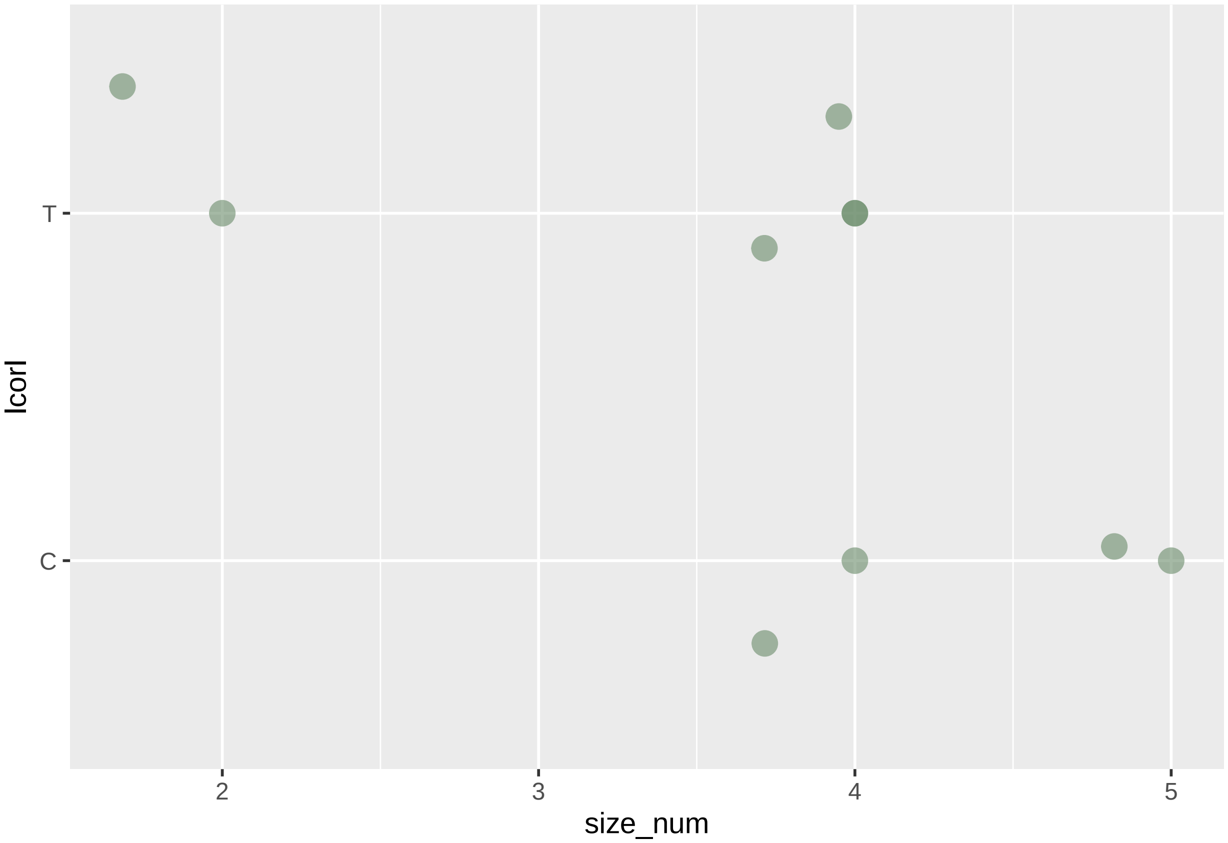
```
#visualize size breakdown of dogs
snps<-read_excel("dog snps.xlsx")
#fix ordering of legend
snps$Name <- factor(snps$Name, levels = c("Basenji", "Boxer", "German_Shepherd","Labrador_retriever","G
p<-ggplot(data = snps, aes(size))+scale_x_discrete(limits = c("S","L","XL"))+geom_bar(aes(fill = Name))
tune_figure(p,add_ons)
```



```
#visualize IGF1 SNP by size
p<-ggplot(data = snps, mapping = aes(y=igf1,x=size_num))+geom_point(size=4,alpha=0.6,color="darkseagreen
p+geom_jitter(size=4,alpha=0.6,color="darkseagreen4")
```

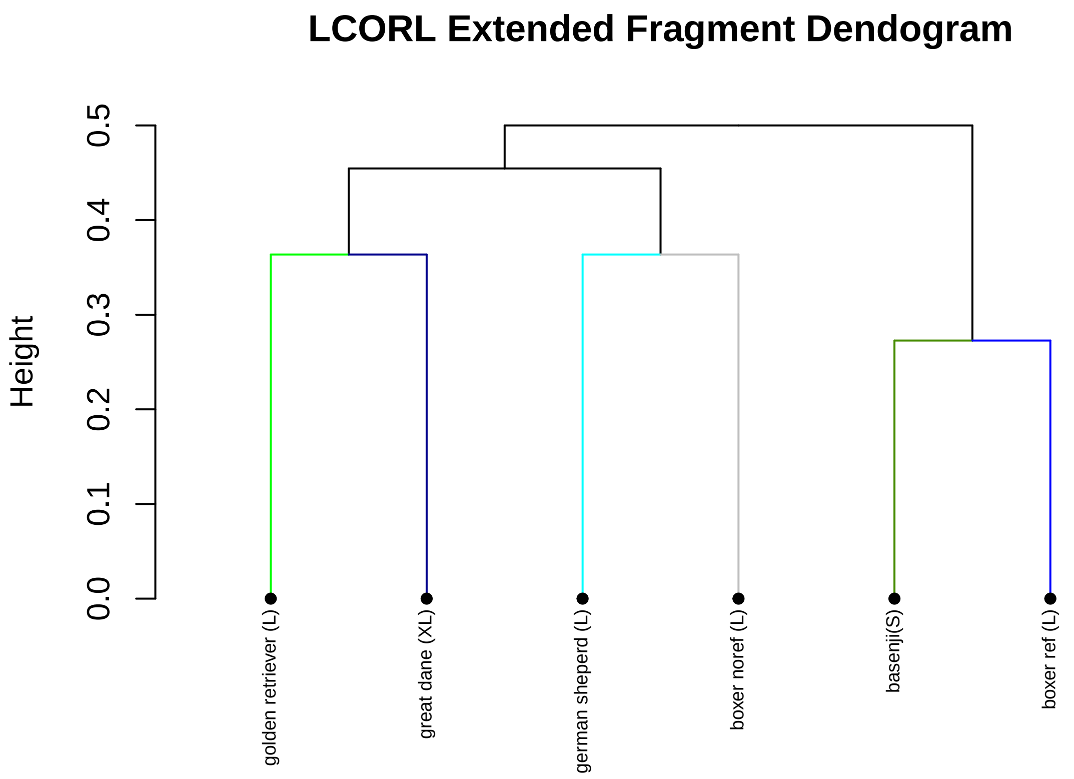


```
#visualize LCORL SNP by size
p<-ggplot(data = snps, mapping = aes(y=lcorl,x=size_num))+geom_point(size=4,alpha=0.6,color="darkseagree
p+geom_jitter(size=4,alpha=0.6,color="darkseagreen4")
```



```
#Cluster LCORL extended fragment
create_dendogram("fasta/LCORL_file.txt", "fasta/names.txt", "LCORL Extended Fragment Dendogram")
```

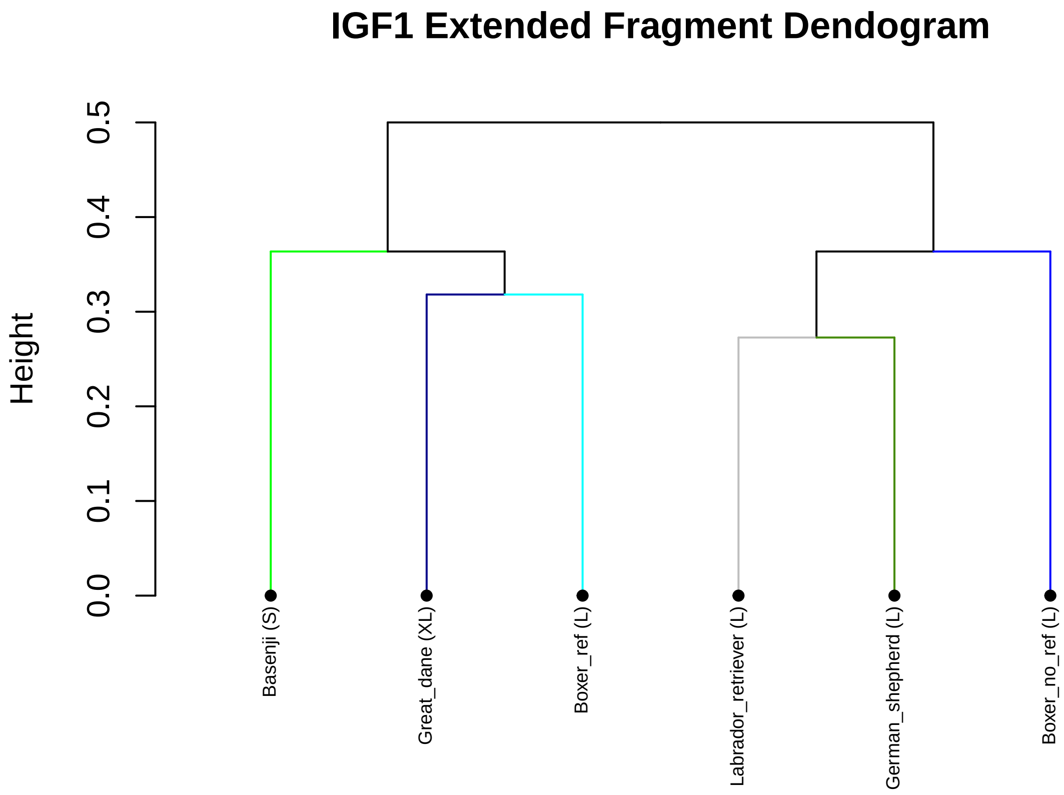
```
## ================================================================================
##
## Time difference of 0 secs
##
## ================================================================================
##
## Time difference of 0.01 secs
```



http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning#plot.dendrogram-function for look and non cut off stuff

```
#Cluster IGF1 extended fragment
create_dendogram("fasta/igf1.fasta", "fasta/igf1_names.txt", "IGF1 Extended Fragment Dendogram")
```

```
## ================================================================================
##
## Time difference of 0 secs
##
## ================================================================================
##
## Time difference of 0.01 secs
```


IGF1 Extended Fragment Dendogram
Height
0.0
0.1
0.2
0.3
0.4
0.5
Basenji (S)
Great_dane (XL)
Boxer_ref (L)
Labrador_retriever (L)
German_shepherd (L)
Boxer_no_ref (L)