

Project 2 Notebook

Introduction (40 points)

- 10 points for background on the protein/gene/species of interest and where the data is sourced from

Introduction

- 10 points for specific, measurable, and clear scientific question

Scientific Question

When examining dog breeds (*canis lupus familiaris*), will breeds of a similar size (e.g. Cocker Spaniel, English Cocker Spaniel) have more related genes and SNP's surrounding longevity than breeds of a different size (e.g. Doberman Pinscher, Miniature Pinscher)?

Note: I only selected 6 genes most closely associated with life span (HMGA2 , IGF1 (done) , IGSF1 (too big), IRS4 (too big), LCORL (done), and SMAD2 (too big)). There are more genes involved in this, but these are the most significant.

Note: Size will be determined by the AKC. You can filter by all AKC recognized dog breeds by size. This is categorical data; if it is easier for me to work with numerical instead, I will instead use the ideal height and weight, as outlined in the official standard of each breed.

:when possible used regions by paper if regions too big used dimension of genes on ncbi

- 10 points for clear, specific, and measurable scientific hypothesis that is in the form of an if-then statement

Scientific Hypothesis

If you examine canine breeds, then breeds of a similar size (e.g. Cocker Spaniel, English Cocker Spaniel) they will have more related SNPs and or fragments of genes surrounding longevity than breeds of a different size (e.g. Doberman Pinscher, Miniature Pinscher).

- 10 points for description of what analyses were done and how the data was downloaded for the project
Analysis Performed: ### SNP's
- EDA: Scatterplots of SNP nucleotide vs size to check for visible trends before analysis
- Multiple Sequence Alignment (of SNP+border sequences), which was then visualized with `msaPrettyPrint()` _ Clustering of MSA results, which were visualized as Dendrograms ### Gene Fragments:
- Multiple Sequence Alignment (of SNP+border sequences), which was then visualized with `msaPrettyPrint()` _ Clustering of MSA results, which were visualized as Dendrograms ### Data Sourcing
- Dog breed information (sizing): American Kennel Club LINK _ Data downloads:
- SNP and gene positionings:
- 25 points for definition of each of the packages loaded
- 5 points for correctly loading all of the packages needed

```

#for reading in fasta files
library("BiocManager")
#for reading in excel files
library("readxl")
#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: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")

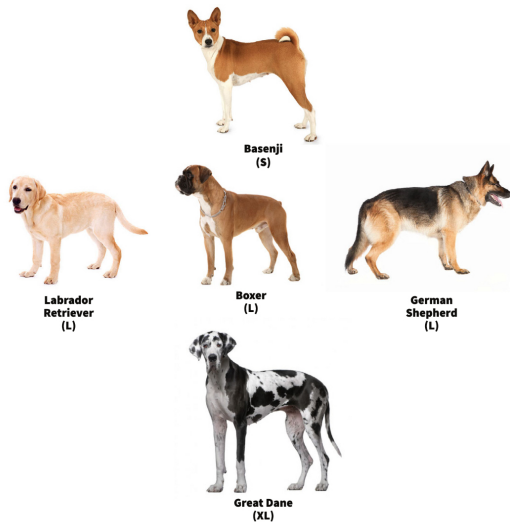
```

```
#for clustering of DNA seqs
library("DECIPHER")
```

```
## Loading required package: RSQLite
## Loading required package: parallel
```

```
#for cleaning up dendograms
library('dendextend')
```

```
##
## -----
## Welcome to dendextend version 1.15.2
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## You may ask questions at stackoverflow, use the r and dendextend tags:
##   https://stackoverflow.com/questions/tagged/dendextend
##
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))
## -----
##
## Attaching package: 'dendextend'
##
## The following object is masked from 'package:Biostrings':
##
##   nnodes
##
## The following object is masked from 'package:stats':
##
##   cutree
```



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

```
#notebook functions
```

```

#align fasta from file_name with names from name file (visualization purposes) after alignment displays
mult_alignments<-function(file_name,fasta_names,name,big_aln=FALSE,dna_set=TRUE){

  #read in fasta for all dogs
  if(dna_set=="TRUE"){
    string_set<-readDNAStrngSet(file=file_name,use.names=FALSE)
  }
  else{
    string_set<-readAAStrngSet(file=file_name,use.names=FALSE)
  }

  #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,order="input")
  #if seq cant be display with msa pretty print, return
  if(big_aln==TRUE){
    return(alignment)
  }

  #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",askForOverwrite=FALSE,
    showNumbering="none",paperWidth=6,paperHeight=3)
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 dendrogram based on fasta files, names of items clustered in fasta_names
create_dendrogram<-function(fasta_path, fasta_names, fig_title){
  #grab DNA info from coallated file
  dna <- string_set<-readDNAStrngSet(file=fasta_path,use.names=FALSE)
  #get sequence names
  names(dna)=read.table(fasta_names, header = FALSE, sep = "\n")["V1"]
  #create distance matrix for clustering
  d1 <- DistanceMatrix(dna, type="dist")
  #form dendrogram
  dendrogram<-IdClusters(d1, method="complete", cutoff=0.05, showPlot=FALSE,type="dendrogram")
  #fix names being cut-off
  nodePar <- list(lab.cex = 0.6, pch = c(NA, 19),

```

```

        cex = 0.7, col = "black")
#plot results
plot(as.dendrogram(dendogram), ylab = "Height", nodePar = nodePar, main = fig_title)
  return(as.dendrogram(dendogram))
}

```

LCORL Analysis

```

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

## use default substitution matrix
print(alignment_name)

## [1] "LCORL.pdf"

```

logo

Species	1	2	3	4	5	6	7	8	9	10
boxer ref (L)	T	G	C	T	G	T	C	A	A	G
boxer noref (L)	G	A	A	G	A	A	A	A	A	A
german sheperd (L)	A	T	C	A	T	A	G	A	G	T
great dane (XL)	C	C	A	T	T	C	C	G	C	C
golden retriever (L)	A	C	A	A	T	T	C	G	T	T
basenji(S)	T	G	C	T	C	C	C	T	G	G
consensus	G	T	G	C	T	A	G	A		

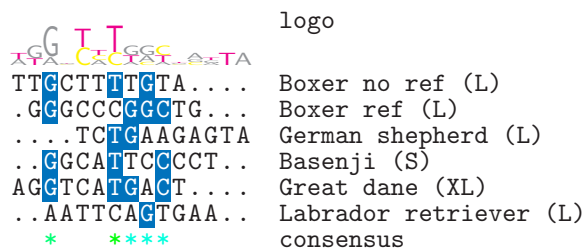
☐ non-conserved
☒ ≥ 50% conserved

```

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

## use default substitution matrix

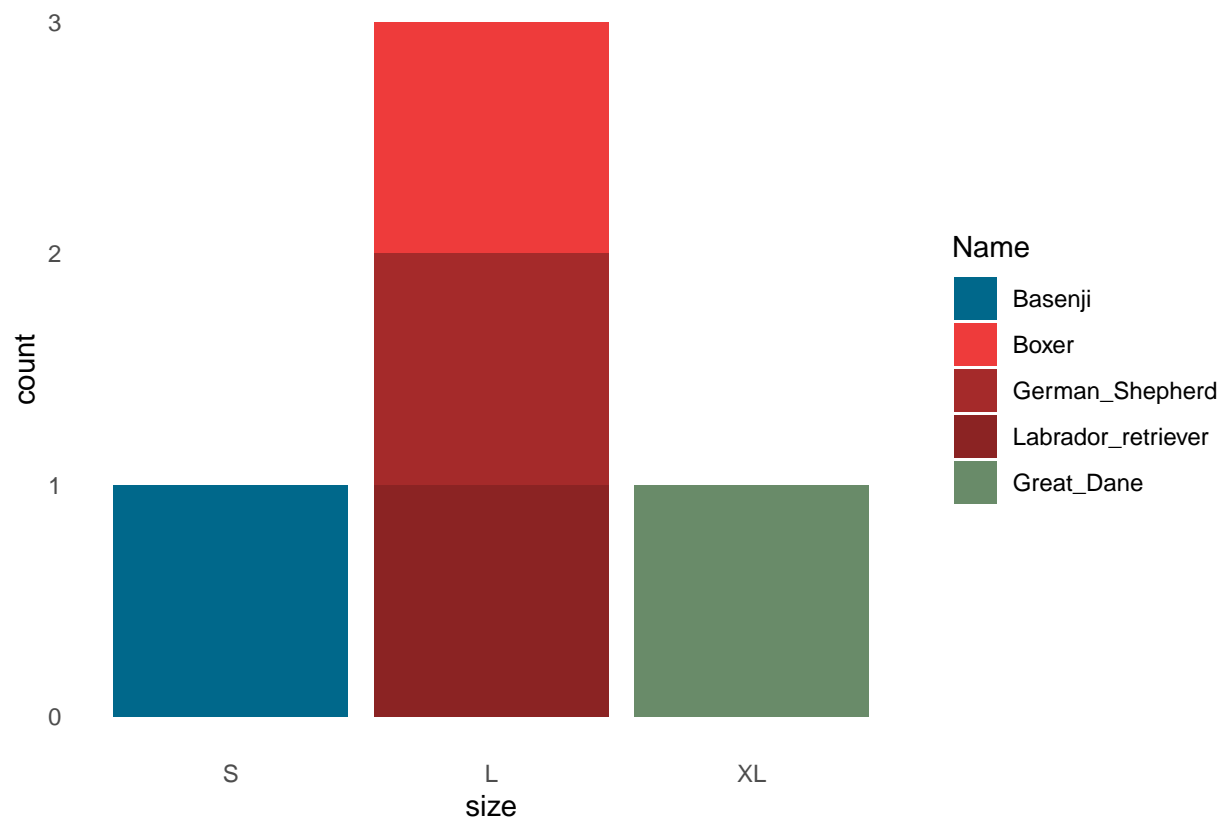
```



☒ non-conserved
☒ $\geq 50\%$ 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", "Great_Dane"))
p<-ggplot(data = snps, aes(size))+scale_x_discrete(limits = c("S", "L", "XL"))+geom_bar(aes(fill = Name))
tune_figure(p,add_ons)
  
```



```

abbrev_x <- c("A", "C", "G", "T")
print(length(abbrev_x))
  
```

```
## [1] 4
```

```
print(length(seq(0,4,by=1)))
```

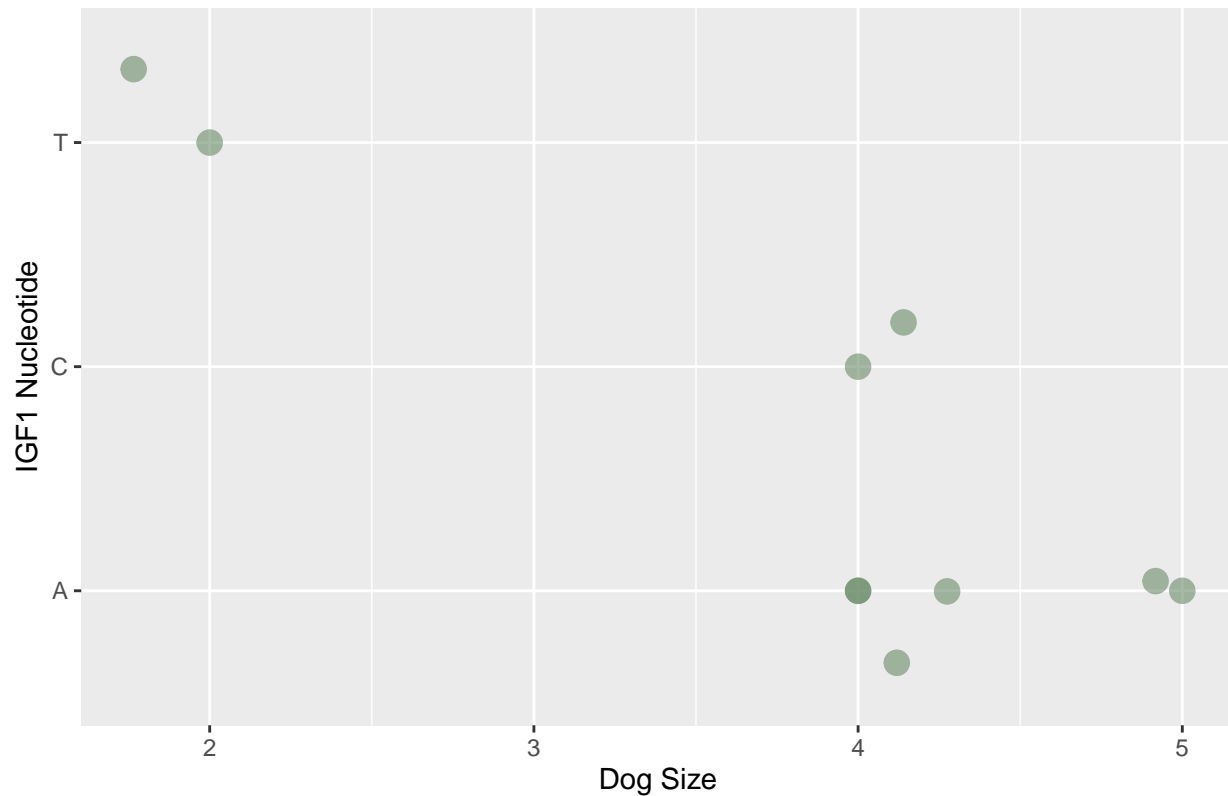
```
## [1] 5
```

```
#visualize IGF1 SNP by size
```

```
p<-ggplot(data = snps, mapping = aes(y=igf1,x=size_num))+geom_point(size=4,alpha=0.6,color="darkseagreen4")
```

```
p+geom_jitter(size=4,alpha=0.6,color="darkseagreen4")
```

IGF1 SNP Distribution by Dog Size

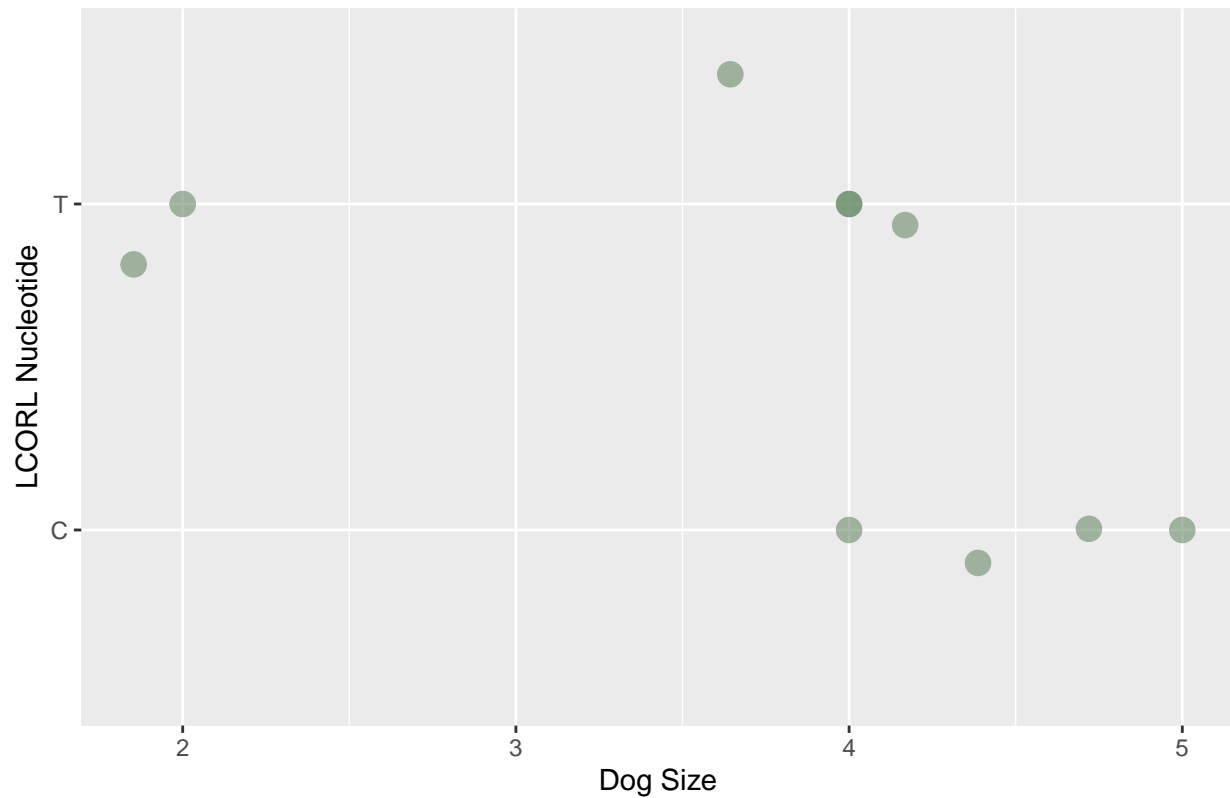


```
#visualize LCORL SNP by size
```

```
p<-ggplot(data = snps, mapping = aes(y=lcorl,x=size_num))+geom_point(size=4,alpha=0.6,color="darkseagreen4")
```

```
p+geom_jitter(size=4,alpha=0.6,color="darkseagreen4")
```

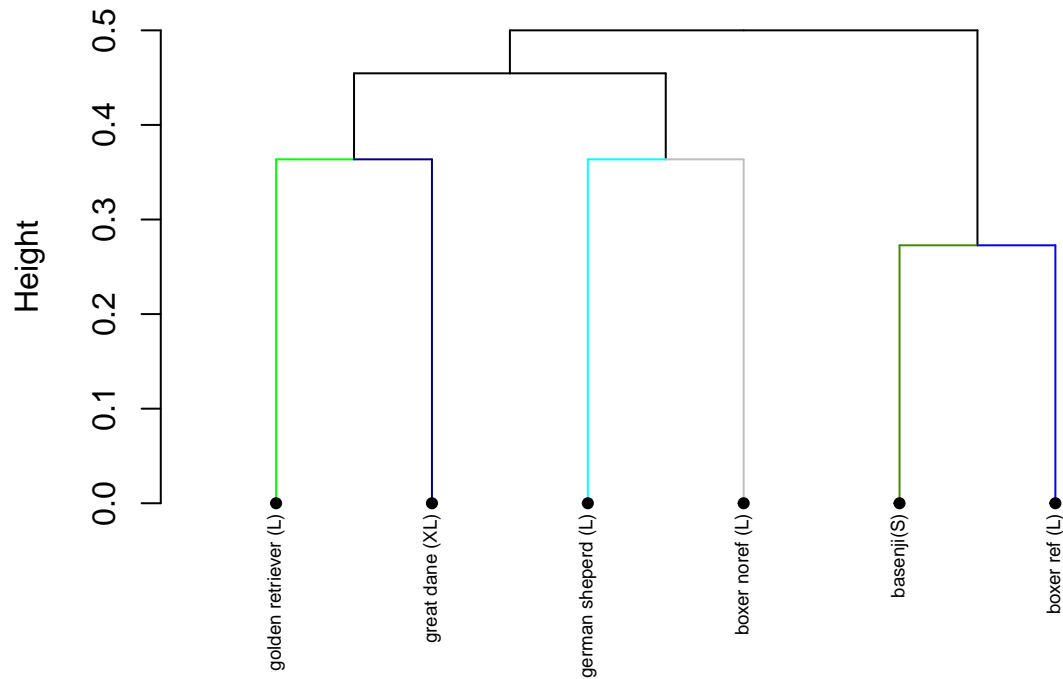
LCORL SNP Distribution by Dog Size



```
#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
```


LCORL Extended Fragment Dendrogram



'dendrogram' with 2 branches and 6 members total, at height 0.5

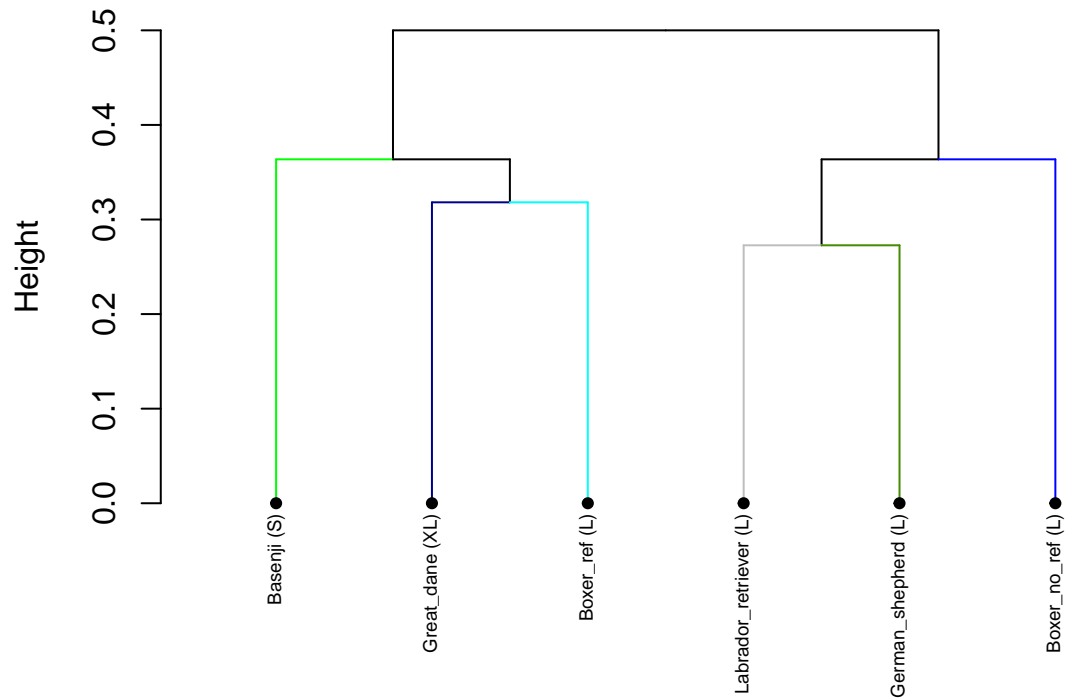
<http://www.sthda.com/english/wiki/beautiful-dendrogram-visualizations-in-r-5-must-known-methods-unsupervised-machine-learning#plot.dendrogram-function> for look and non cut off stuff

#Cluster IGF1 extended fragment

`create_dendrogram("fasta/igf1.fasta", "fasta/igf1_names.txt", "IGF1 Extended Fragment Dendrogram")`

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

IGF1 Extended Fragment Dendrogram



```
## 'dendrogram' with 2 branches and 6 members total, at height 0.5
```

```
string_set<-readAAStringSet(file="fasta/IGSF1.fasta",use.names=TRUE)
#read in seq names as list
table=read.table("fasta/IGSF1_names.txt", header = FALSE, sep = "\n")["V1"]
#update names for pretty print
names(string_set)<-table
mult<-msa(string_set)
```

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

```
igsf1<-mult_alingments("fasta/IGSF1.fasta","fasta/IGSF1_names.txt","IGSF1",TRUE,FALSE)
```

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

```
library(seqinr)
```

```
##
```

```
## Attaching package: 'seqinr'
```

```
## The following object is masked from 'package:Biostrings':
```

```
##
```

```
## translate
```

```
library(ape)
```

```
##
```

```
## Attaching package: 'ape'
```

```
## The following objects are masked from 'package:seqinr':
```

```
##
```

```
## as.alignment, consensus
```

```
## The following objects are masked from 'package:dendextend':
##
##   ladderize, rotate
## The following object is masked from 'package:Biostrings':
##
##   complement
igsf1_aln <- msaConvert(igsf1, type="seqinr::alignment")
d <- dist.alignment(igsf1_aln, "identity")
dendrogram<-IdClusters(d, method="complete", cutoff=0.05, showPlot=FALSE,type="dendrogram")

## =====
##
## Time difference of 0 secs
#fix names being cut-off
nodePar <- list(lab.cex = 0.6, pch = c(NA, 19),
               cex = 0.7, col = "black")
#plot results
plot(as.dendrogram(dendrogram), ylab = "Height", nodePar =nodePar,main="igsf1")
```

