# motifStack guide

Jianhong Ou, Lihua Julie Zhu

January 16, 2014

## Contents



## 1 Introduction

A sequence logo, based on information theory, has been widely used as a graphical representation of sequence conservation (aka motif) in multiple amino acid or nucleic acid sequences. Sequence motif represents conserved characteristics such as DNA binding sites, where transcription factors bind, and catalytic sites in enzymes. Although many tools, such as seqlogo[1], have been developed to create sequence motif and to represent it as individual sequence logo, software tools for depicting the relationship among multiple sequence motifs are still lacking. We developed a flexible and powerful open-source R/Bioconductor package, motifStack, for visualization of the alignment of multiple sequence motifs.

# 2 Prepare environment

You will need ghostscript: the full path to the executable can be set by the environment variable R_GSCMD. If this is unset, a GhostScript executable will be searched by name on your path. For example, on a Unix, linux or Mac "gs" is used for searching, and on Windows the setting of the environment variable GSC is used, otherwise commands "gswi64c.exe" then "gswin32c.exe" are tried.

Example on Windows: assume that the gswin32c.exe is installed at C:\Program Files\gs\gs9.06\bin, then open R and try:

```
> Sys.setenv(R_GSCMD="\"C:\\Program Files\\gs\\gs9.06\\bin\\gswin32c.exe\"")
```

# 3 Examples of using motifStack

## 3.1 plot a DNA sequence logo with different fonts and colors

Users can select different fonts and colors to draw the sequence logo (Figure 1).

```
> library(motifStack)
> pcm <- read.table(file.path(find.package("motifStack"),
+                             "extdata", "bin_SOLEXA.pcm"))
> pcm <- pcm[,3:ncol(pcm)]
> rownames(pcm) <- c("A","C","G","T")
> motif <- new("pcm", mat=as.matrix(pcm), name="bin_SOLEXA")
> ##pfm object
> #motif <- pcm2pfm(pcm)
> #motif <- new("pfm", mat=motif, name="bin_SOLEXA")
> opar<-par(mfrow=c(3,1))
> plot(motif)
> #try a different font
> plot(motif, font="mono,Courier")
> #try a different font and a different color group
> motif@color <- colorset(colorScheme='basepairing')
> plot(motif,font="Times")
> par(opar)
```

## 3.2 plot an amino acid sequence logo

Given that motifStack allows to use any letters as symbols, it can also be used to draw amino acid sequence logos (Figure 2).

```
> library(motifStack)
> protein<-read.table(file.path(find.package("motifStack"),"extdata","cap.txt"))
> protein<-t(protein[,1:20])
> motif<-pcm2pfm(protein)
> motif<-new("pfm", mat=motif, name="CAP",
```

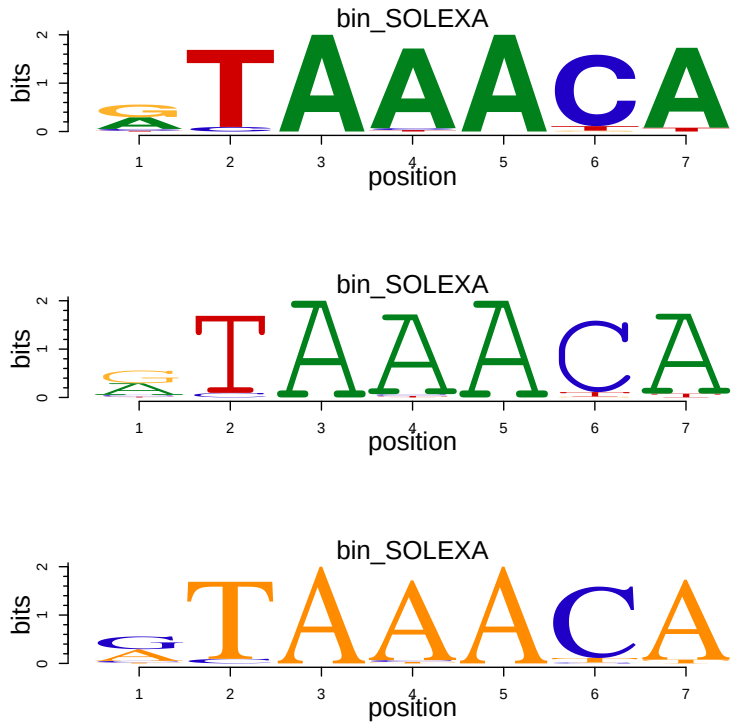


Figure 1: **DNA sequence logo.** Plot a DNA sequence logo with different fonts and colors.

```
+             color=colorset(alphabet="AA",colorScheme="chemistry"))
> plot(motif)
```

## 3.3 plot sequence logo stack

motifStack is designed to show multiple motifs in same canvas. To show the sequence logo stack, the distance of motifs need to be calculated first for example by using MotIV[2]::motifDistances, which implemented STAMP[3]. After alignment, users can use plotMotifLogoStack function to draw sequence logos stack (Figure 3) or use plotMotifLogoStackWithTree function to show the distance tree with the sequence logos stack (Figure 4) or use plotMotifStackWithRadialPhylog function to plot sequence logo stack in radial style (Figure 5) in the same canvas. There is a shortcut function named as motifStack. Use stack layout to call plotMotifLogoStack, treeview layout to call plotMotifLogoStackWithTree and radialPhylog to call plotMotifStackWithRadialPhylog.

```
> library(motifStack)
> #####Input#####
> pcms<-readPCM(file.path(find.package("motifStack"), "extdata"),"pcm$")
> motifs<-lapply(pcms,pcm2pfm)
> ## plot stacks
> motifStack(motifs, layout="stack", ncex=1.0)
```

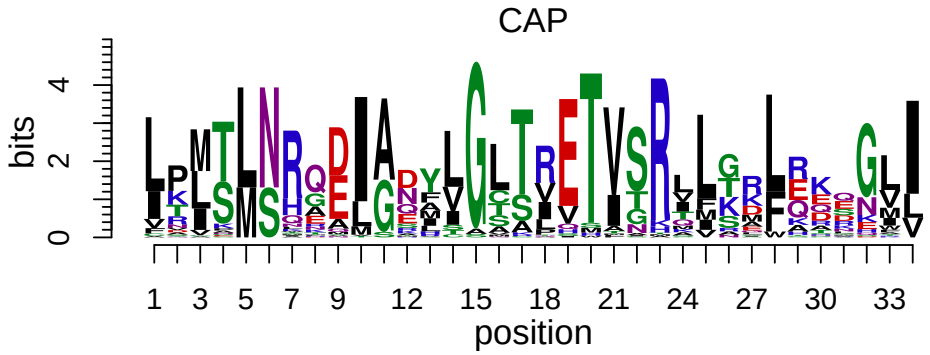


Figure 2: **Amino acid sequence logo.** Plot an sequence logo with any symbols as you want such as amino acid sequence logo

```
> ## plot stacks with hierarchical tree
> motifStack(motifs, layout="tree")
```

```
> ## When the number of motifs is too much to be shown in a vertical stack,
> ## motifStack can draw them in a radial style.
> ## random sample from MotifDb
> library("MotifDb")
> matrix.fly <- query(MotifDb, "Dmelanogaster")
> motifs2 <- as.list(matrix.fly)
> ## use data from FlyFactorSurvey
> motifs2 <- motifs2[grepl("Dmelanogaster\\-FlyFactorSurvey\\-",
+                          names(motifs2))]
> ## format the names
> names(motifs2) <- gsub("Dmelanogaster_FlyFactorSurvey_", "",
+                        gsub("_FBgn\\d+$", "",
+                             gsub("[^a-zA-Z0-9]","_",
+                                  gsub("(_\\d+)+$", "", names(motifs2)))))
> motifs2 <- motifs2[unique(names(motifs2))]
> pfms <- sample(motifs2, 50)
> ## creat a list of object of pfm
> motifs2 <- lapply(names(pfms),
+                   function(.ele, pfms){new("pfm",mat=pfms[[.ele]], name=.ele)}
+                   ,pfms)
> ## trim the motifs
> motifs2 <- lapply(motifs2, trimMotif, t=0.4)
> ## setting colors
```

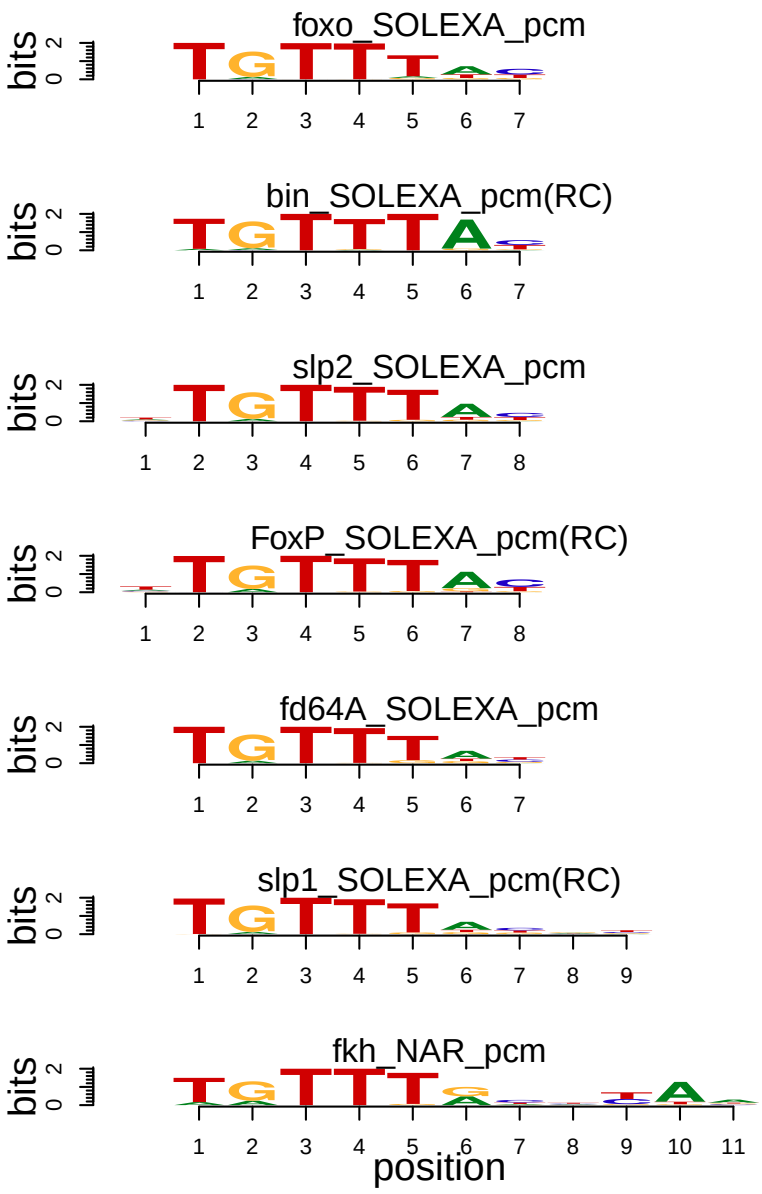


Figure 3: **Sequence logo stack.** Plot motifs with sequence logo stack style.

```
> library(RColorBrewer)
> color <- brewer.pal(12, "Set3")
> ## plot logo stack with radial style
> motifStack(motifs2, layout="radialPhylog",
+            circle=0.3, cleaves = 0.2,
+            clabel.leaves = 0.5,
+            col.bg=rep(color, each=5), col.bg.alpha=0.3,
+            col.leaves=rep(color, each=5),
+            col.inner.label.circle=rep(color, each=5),
```

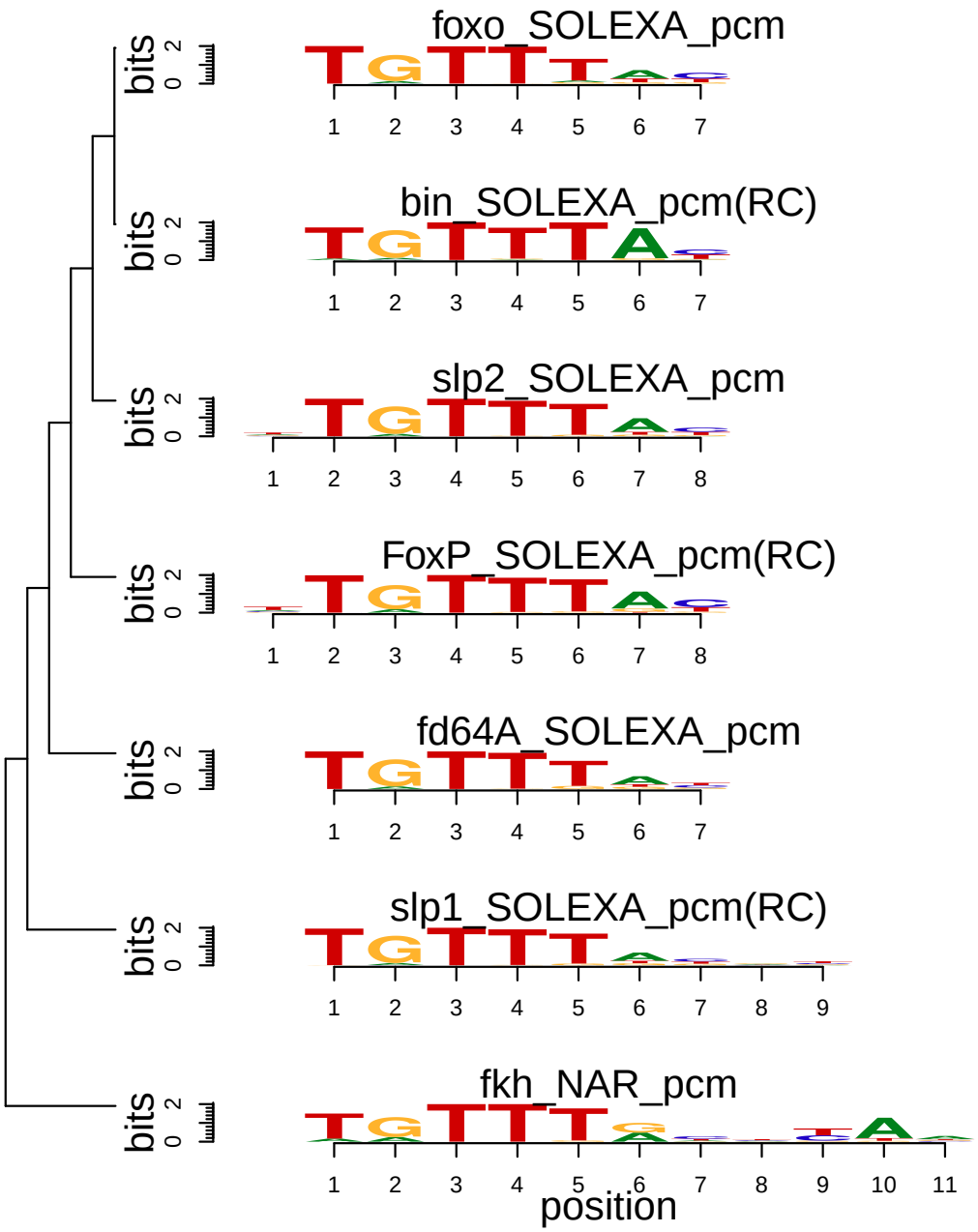


Figure 4: **Treeview layout logo stack.** Sequence logo stack with hierarchical cluster tree.

```
+             inner.label.circle.width=0.05,
+             col.outer.label.circle=rep(color, each=5),
+             outer.label.circle.width=0.02,
+             circle.motif=1.2,
+             angle=350)
```

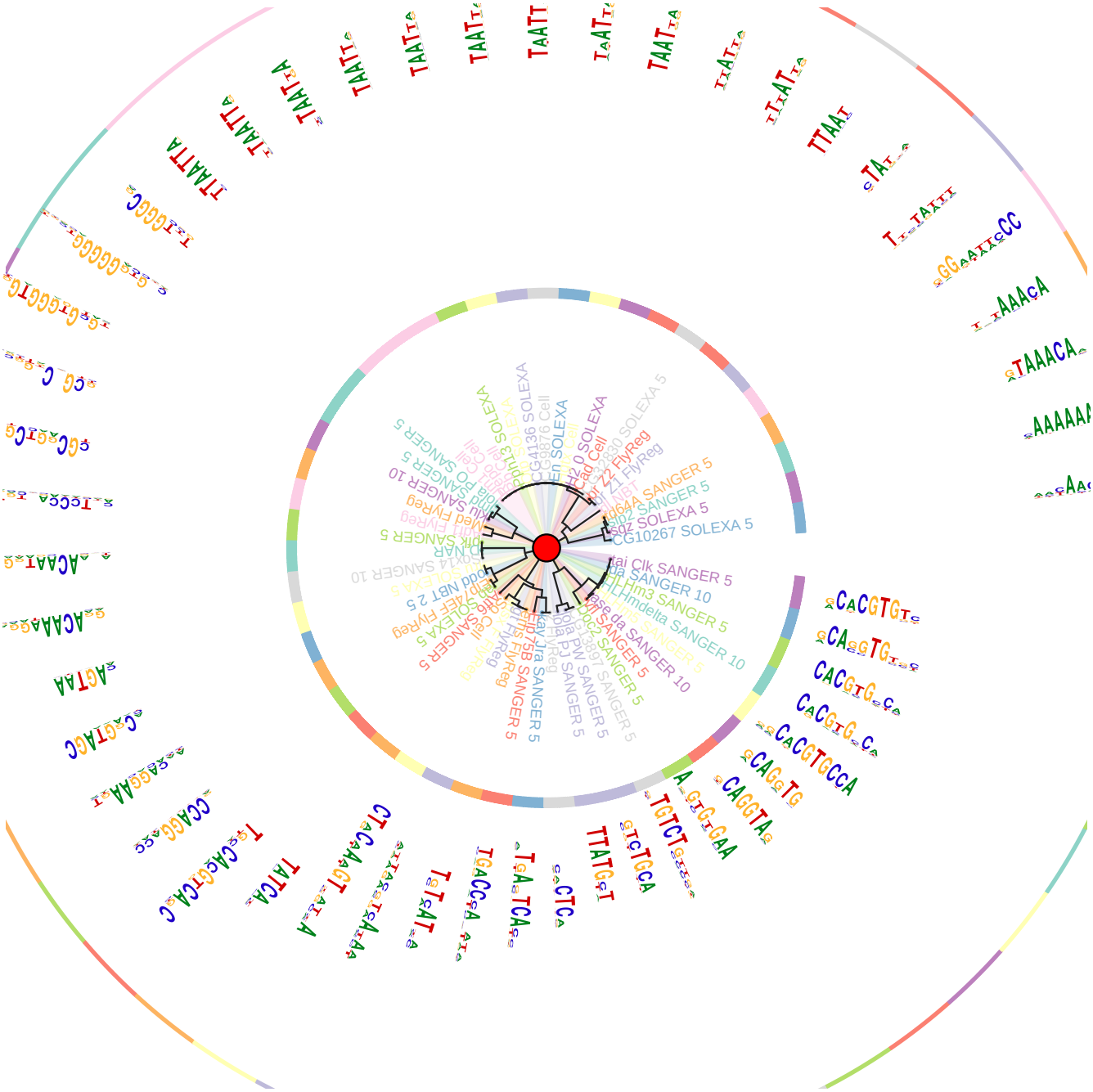

Figure 5: **Sequence logo stack in radial style** Plot motifs in a radial style when the number of motifs is too much to be shown in a vertical stack.

## 3.4 plot a sequence logo cloud

We can also plot a sequence logo cloud for DNA sequence logo (Figure 6).

```
> ## assign groups for motifs
> groups <- rep(paste("group",1:5,sep=""), each=10)
```

```
> names(groups) <- names(pfms)
> ## assign group colors
> group.col <- brewer.pal(5, "Set3")
> names(group.col)<-paste("group",1:5,sep="")
> ## use MotIV to calculate the distances of motifs
> jaspar.scores <- MotIV::readDBScores(file.path(find.package("MotIV"),
+                                                "extdata",
+                                                "jaspar2010_PCC_SWU.scores"))
> d <- MotIV::motifDistances(pfms)
> hc <- MotIV::motifHclust(d)
> ## convert the hclust to phylog object
> phylog <- hclust2phylog(hc)
> ## reorder the pfms by the order of hclust
> leaves <- names(phylog$leaves)
> pfms <- pfms[leaves]
> ## create a list of pfm objects
> pfms <- lapply(names(pfms), function(.ele, pfms){
+                                  new("pfm",mat=pfms[[.ele]], name=.ele)}
+                ,pfms)
> ## extract the motif signatures
> motifSig <- motifSignature(pfms, phylog, groupDistance=0.01, min.freq=1)
> ## draw the motifs with a tag-cloud style.
> motifCloud(motifSig, scale=c(6, .5),
+            layout="rectangles",
+            group.col=group.col,
+            groups=groups,
+            draw.legend=T)
```

## 3.5 plot grouped sequence logo

To plot grouped sequence logo, except do motifCloud, we can also plot it with radialPhylog style (Figure 7).

```
> ## get the signatures from object of motifSignature
> sig <- signatures(motifSig)
> ## extract the motif names for each signature.
> ## the motif names are separated by ";"
> ## and then set the inner-circle color for each signature
> pfmNames <- lapply(sig, function(.ele){unlist(strsplit(.ele@name, ";"))})
> pfmNames <- mapply(function(.ele, .name){cbind(.ele, .name)},
+       pfmNames, paste("gps", 1:length(pfmNames), sep=""))
> pfmNames <- do.call(rbind, pfmNames)
> pfmNames <- pfmNames[match(names(phylog$leaves), pfmNames[,1]),]
> bd.color <- c("gray80","gray30")
```


Figure 6: **Sequence logo cloud with rectangle packing layout** Like tag-cloud, the sequence logo size is determined by the number of motifs of the signature. The group sources of the motifs for each signature are shown as a pie graph in topleft corner.

```
> in.color <- rle(pfmNames[,2])
> in.color$values <- bd.color[rep(1:2,
+                     length(in.color$lengths))[1:length(in.color$lengths)]]
> pfmNames <- cbind(pfmNames, color=inverse.rle(in.color))
> ## plot the logo stack with radial style.
> plotMotifStackWithRadialPhylog(phylog=phylog, pfms=sig,
+                                circle=0.4, cleaves = 0.3,
+                                clabel.leaves = 0.5,
+                                col.bg=rep(color, each=5), col.bg.alpha=0.3,
+                                col.leaves=rep(color, each=5),
+                                col.inner.label.circle=pfmNames[,3],
```

```
+                                   inner.label.circle.width=0.03,
+                                   angle=350, circle.motif=1.2,
+                                   motifScale="logarithmic")
```

# 4 References

## References

[1] seqLogo: Sequence logos for DNA sequence alignments. R package version 1.22.0.

[2] MotIV: Motif Identification and Validation. Eloi Mercier and Raphael Gottardo (2010). R package version 1.10.0.

[3] STAMP: a web tool for exploring DNA-binding motif similarities. Mahony S, Benos PV, Nucleic Acids Res. 2007, 35(Web Server issue): W253-W258.

# 5 Session Info

```
> toLatex(sessionInfo())
```

- R version 3.0.2 Patched (2013-12-18 r64484), i386-w64-mingw32
- Locale: LC_COLLATE=C, LC_CTYPE=English_United States.1252, LC_MONETARY=English_United States.1252, LC_NUMERIC=C, LC_TIME=English_United States.1252
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, utils
- Other packages: BiocGenerics 0.8.0, Biostrings 2.30.1, IRanges 1.20.6, MotIV 1.18.0, MotifDb 1.4.0, RColorBrewer 1.0-5, XML 3.98-1.1, XVector 0.2.0, ade4 1.6-2, grImport 0.9-0, motifStack 1.6.5
- Loaded via a namespace (and not attached): BSgenome 1.30.0, BiocStyle 1.0.0, GenomicRanges 1.14.4, RCurl 1.95-4.1, Rsamtools 1.14.2, bitops 1.0-6, lattice 0.20-24, rGADEM 2.10.0, rtracklayer 1.22.1, seqLogo 1.28.0, stats4 3.0.2, tools 3.0.2, zlibbioc 1.8.0

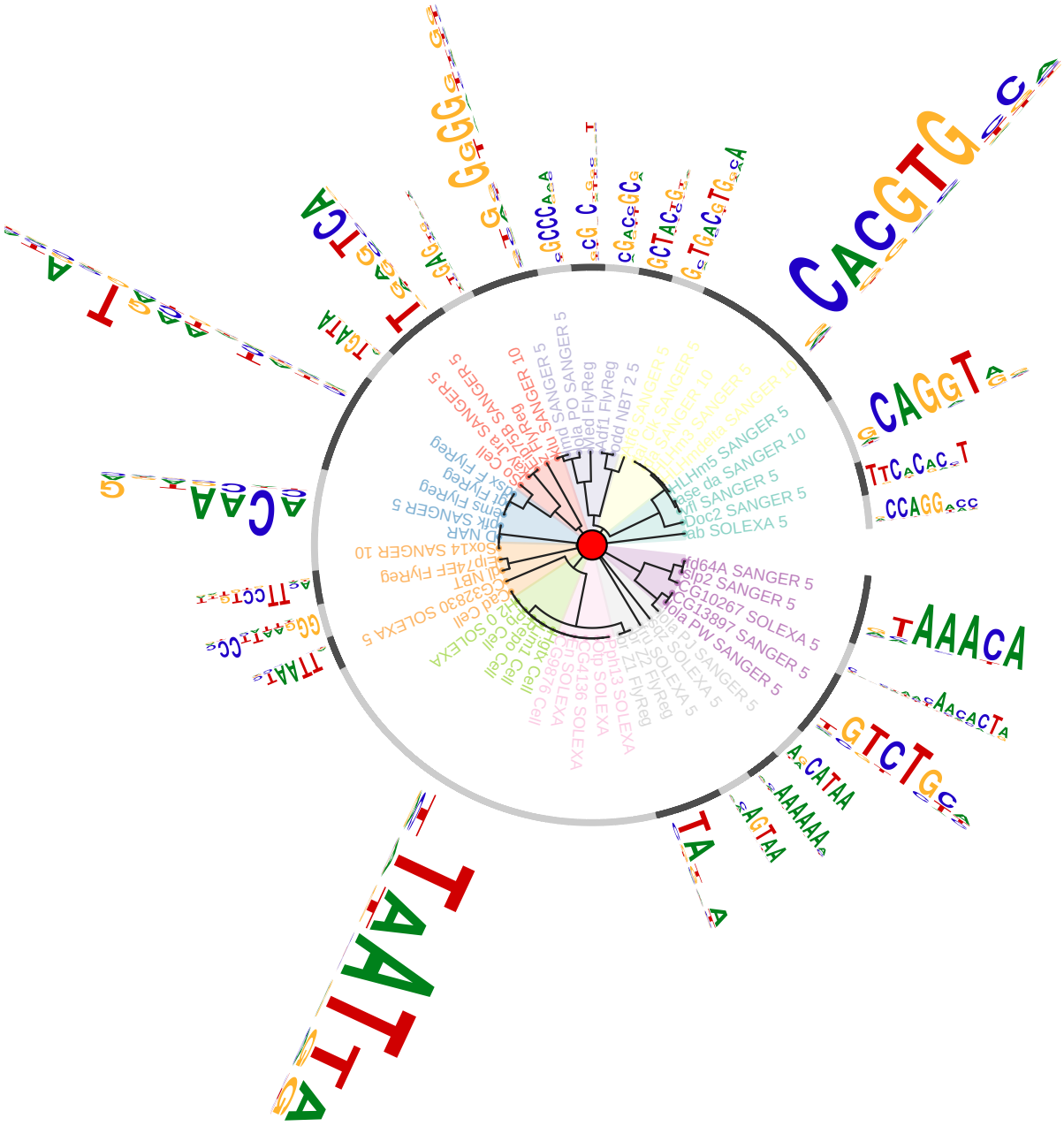

Figure 7: **Grouped sequence logo with radialPhylog style layout.** Like tag-cloud, the sequence logo size is determined by the number of motifs for the signature. The gray-black circle indicates the range of each signature.