motifStack

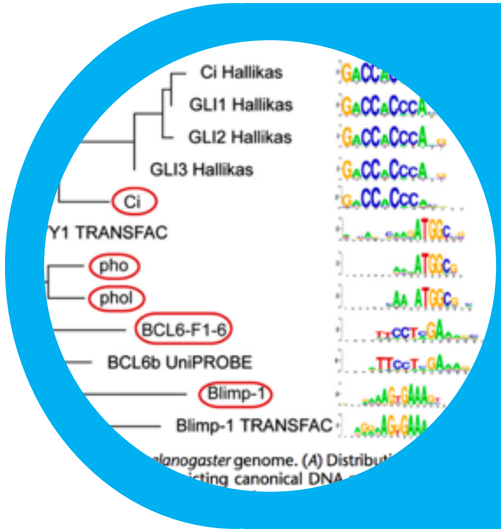
A Tool to Visualize Sequence Logo Alignments

Jianhong Ou
Julie Zhu

Install the workshop pkg

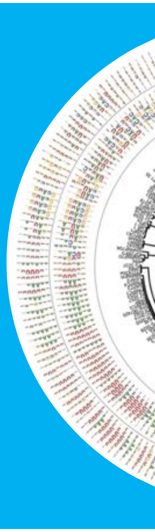
```
## set the working directory,
## replace "~/Downloads/workshop2020" by your path
wd <- "~/Downloads/workshop2020"
dir.create(wd)
setwd(wd)
library(BiocManager)
install("jianhong/workshop2020", build_vignettes = TRUE)
vignette(”motifStack", package="workshop2020")
```

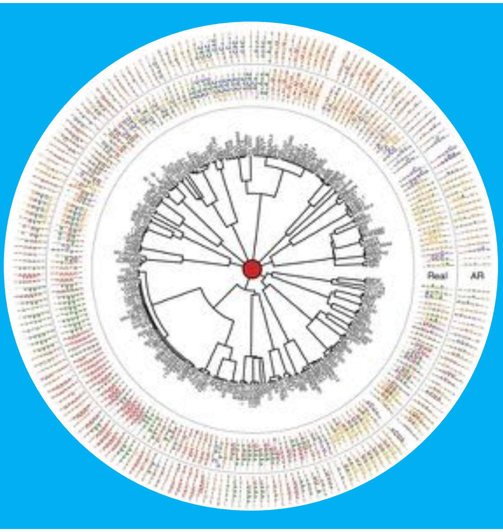
https://github.com/jianhong/workshop2020
https://bioconductor.org/packages/motifStack
https://www.nature.com/articles/nmeth.4555
Slides:
https://github.com/jianhong/workshop2020/blob/master/inst/extdata/motifStack_workshop2020.pdf



2012 motifStack package initialed

Enuameh MS, et al. 2013. doi: 10.1101/gr.151472.112





2015 compare two algorithms for 130 RBPs

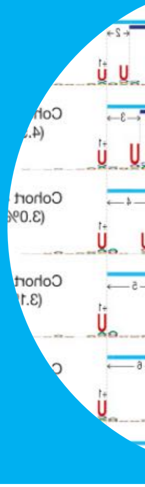
Pelossof R, et al. 2015. doi: 10.1038/nbt.3343

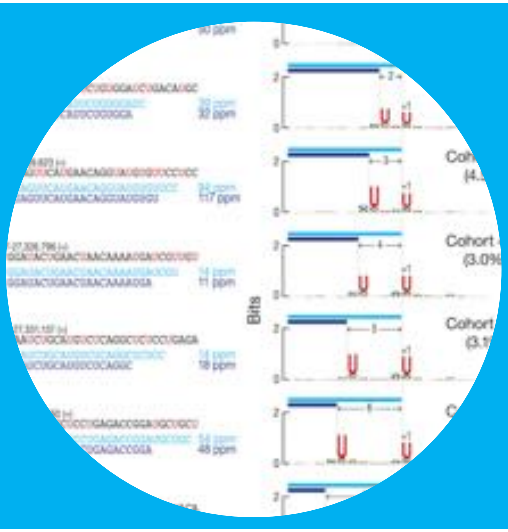
2016 plot Amino Acid (AA) logo

Kevorkian et.al., 2016. doi: 10.1016/j.biochi.2015.07.023

2016 impressed people by its beauty

O'Malley RC, et al. 2016. doi: 10.1016/j.cell.2016.04.038

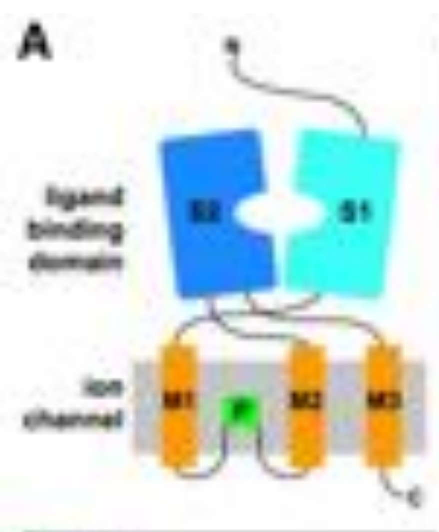


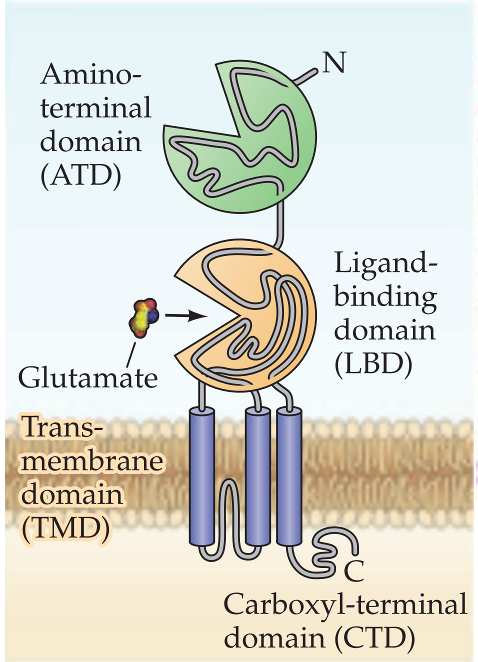


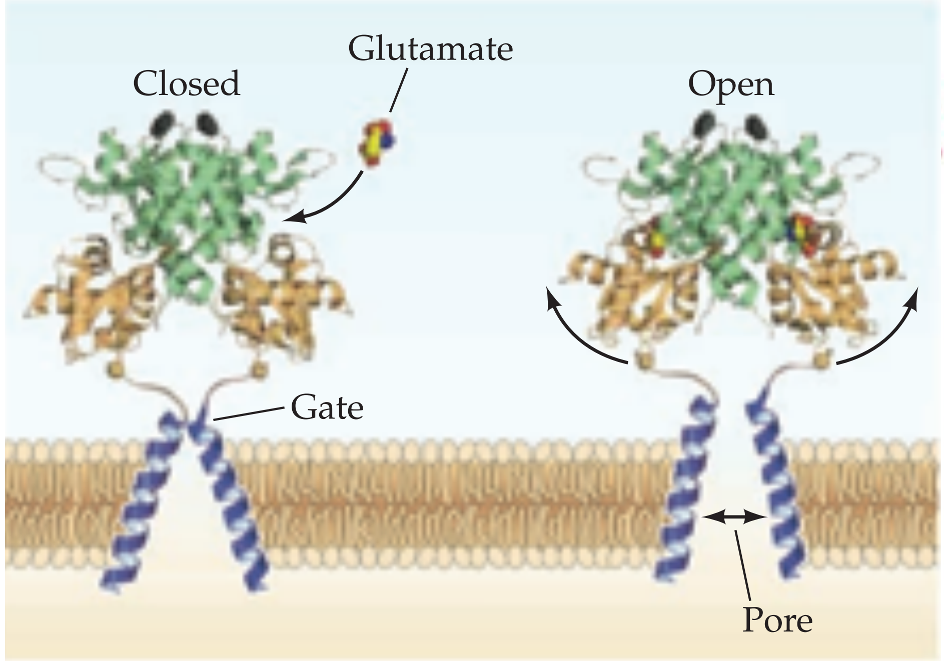
2018 plot piRNA sequence logo

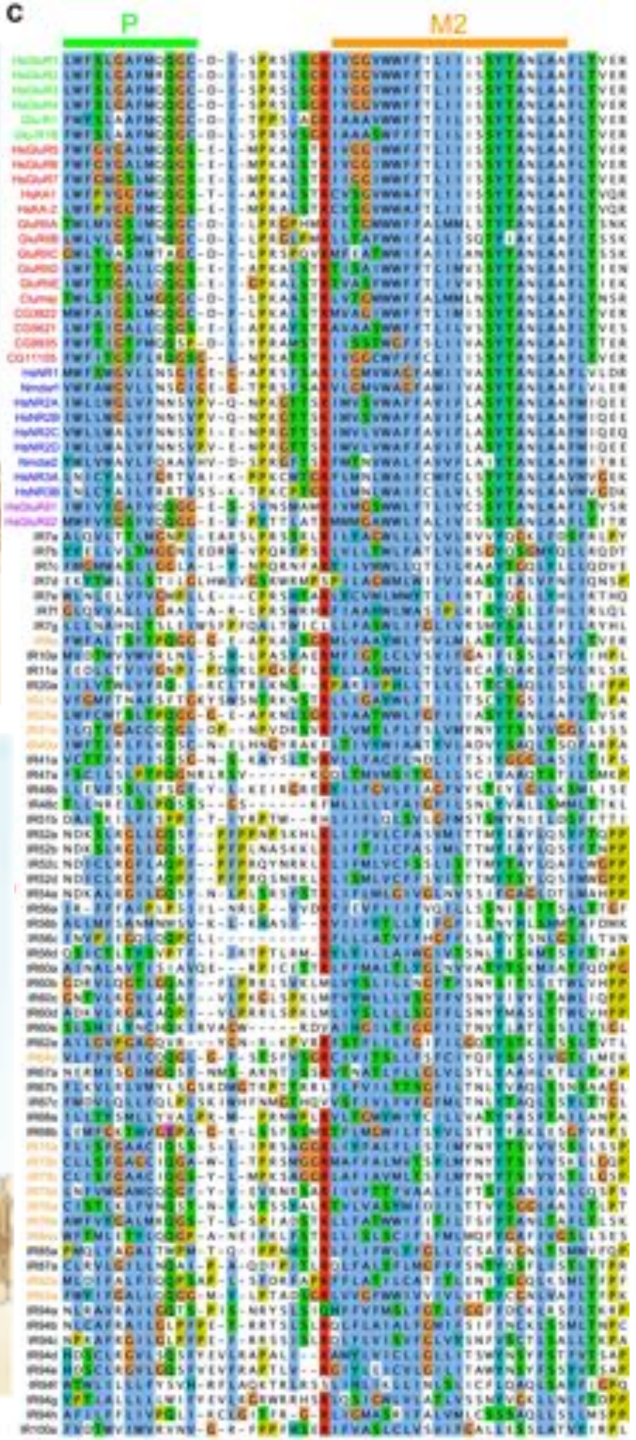
Gainetdinov et.al., 2018. doi: 10.1016/j.molcel.2018.08.007
Nishida et.al., 2018. doi: 10.1038/nature25788

?




N
Amino-terminal domain (ATD)
Ligand-binding domain (LBD)
Glutamate
Trans-membrane domain (TMD)
C
Carboxyl-terminal domain (CTD)

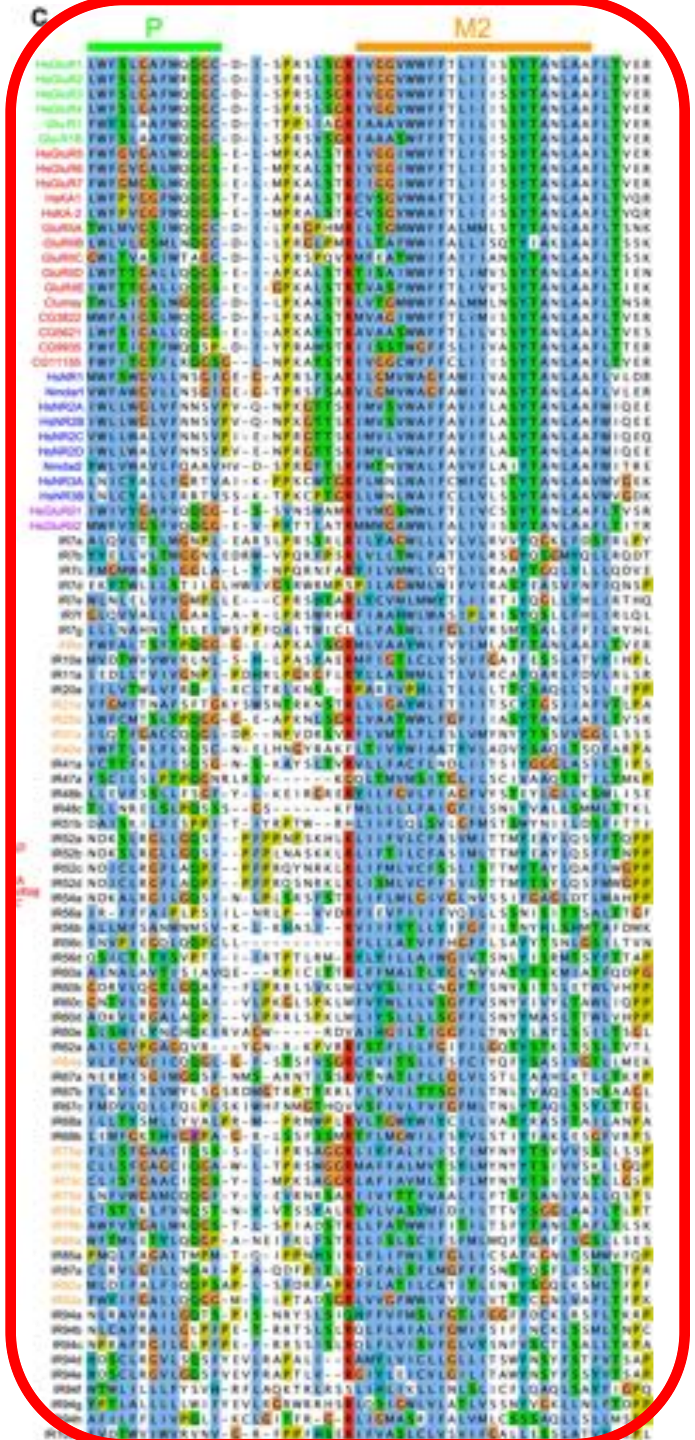




Sequence Motif
Stereotypical element

Purves et.al., 2017. Neuroscience Sixth Edition

Benton et.al., 2009. doi: 10.1016/j.cell.2008.12.001



PCM

	[,1]	[,2]	[,3]	[,4]
A	7	1	0	8
C	4	1	2	8
D	1	9	5	0
E	1	3	2	0
F	15	3	25	3
G	4	0	2	4
H	2	0	0	1
I	12	4	10	4
K	0	1	5	0
L	16	19	16	12
M	5	3	1	6
N	9	5	2	2
P	1	2	0	3
Q	1	0	4	0
R	0	1	5	0
S	1	2	4	15
T	3	1	4	12
V	5	2	1	15
W	3	35	1	0
Y	3	1	3	0

PFM

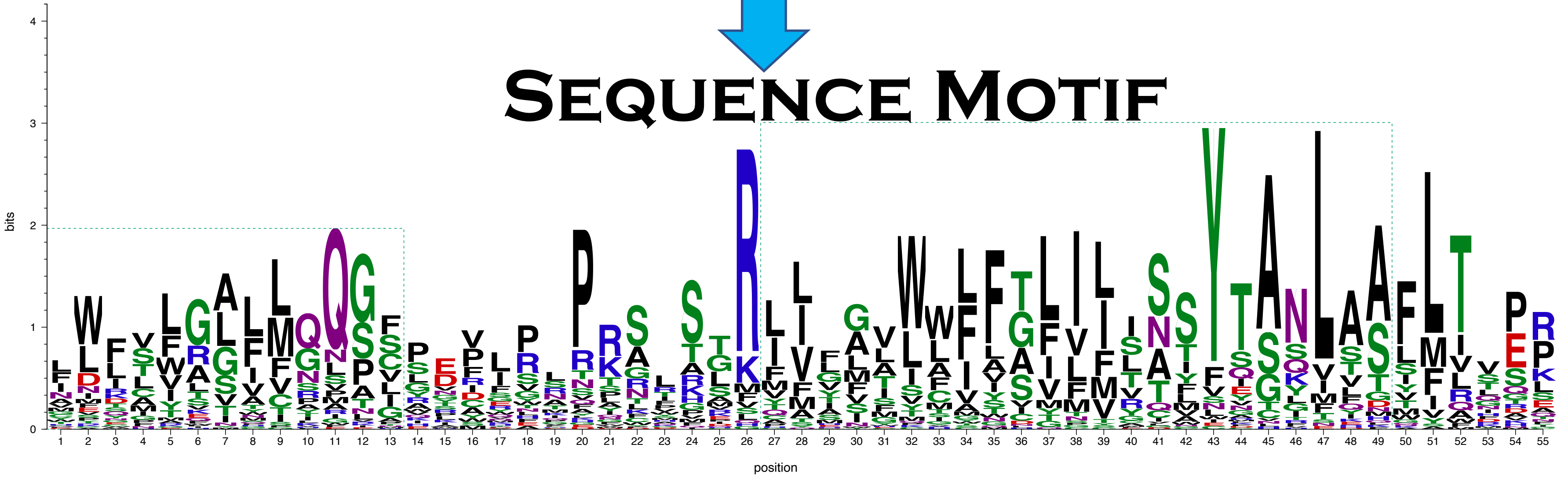
	[,1]	[,2]	[,3]	[,4]
A	0.075	0.011	0.000	0.086
C	0.043	0.011	0.022	0.086
D	0.011	0.097	0.054	0.000
E	0.011	0.032	0.022	0.000
F	0.161	0.032	0.272	0.032
G	0.043	0.000	0.022	0.043
H	0.022	0.000	0.000	0.011
I	0.129	0.043	0.109	0.043
K	0.000	0.011	0.054	0.000
L	0.172	0.204	0.174	0.129
M	0.054	0.032	0.011	0.065
N	0.097	0.054	0.022	0.022
P	0.011	0.022	0.000	0.032
Q	0.011	0.000	0.043	0.000
R	0.000	0.011	0.054	0.000
S	0.011	0.022	0.043	0.161
T	0.032	0.011	0.043	0.129
V	0.054	0.022	0.011	0.161
W	0.032	0.376	0.011	0.000
Y	0.032	0.011	0.033	0.000

PWM

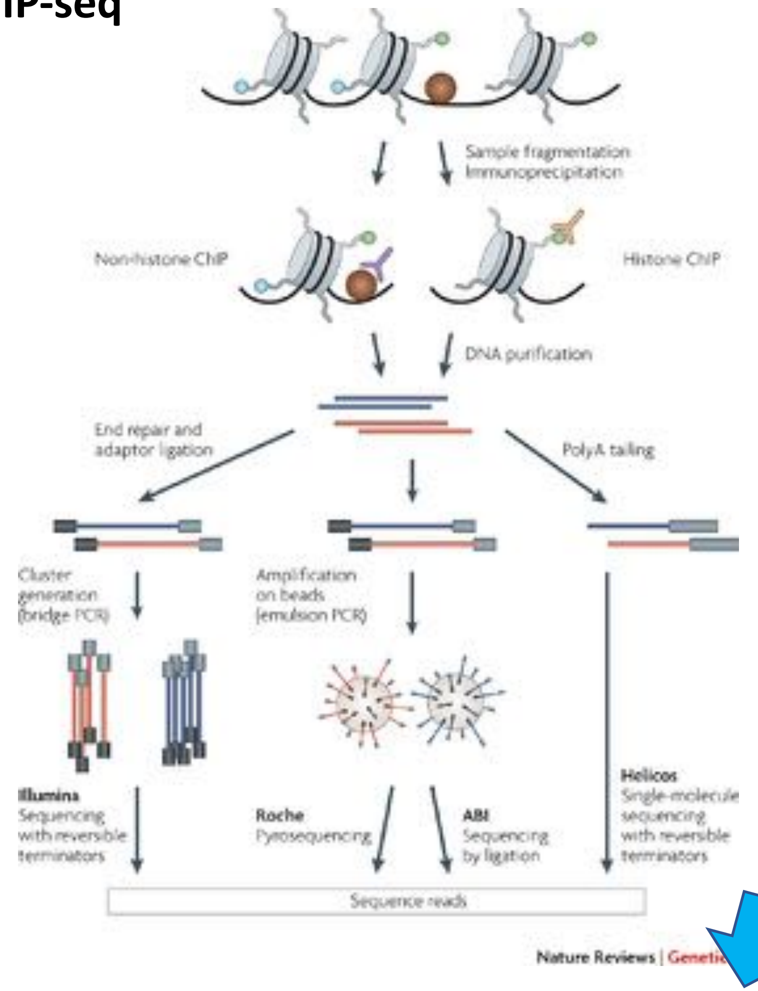
	[,1]	[,2]	[,3]	[,4]
A	0.590	-2.217	-15.678	0.783
C	-0.217	-2.217	-1.202	0.783
D	-2.217	0.953	0.120	-15.678
E	-2.217	-0.632	-1.202	-15.678
F	1.690	-0.632	2.442	-0.632
G	-0.217	-15.678	-1.202	-0.217
H	-1.217	-15.678	-15.678	-2.217
I	1.368	-0.217	1.120	-0.217
K	-15.678	-2.217	0.120	-15.678
L	1.783	2.031	1.798	1.368
M	0.105	-0.632	-2.202	0.368
N	0.953	0.105	-1.202	-1.217
P	-2.217	-1.217	-15.678	-0.632
Q	-2.217	-15.678	-0.202	-15.678
R	-15.678	-2.217	0.120	-15.678
S	-2.217	-1.217	-0.202	1.690
T	-0.632	-2.217	-0.202	1.368
V	0.105	-1.217	-2.202	1.690
W	-0.632	2.912	-2.202	-15.678
Y	-0.632	-2.217	-0.617	-15.678

$$PWM(b,i) = \log \frac{PFM(b,i)}{P(b)}$$

Sequence Motif

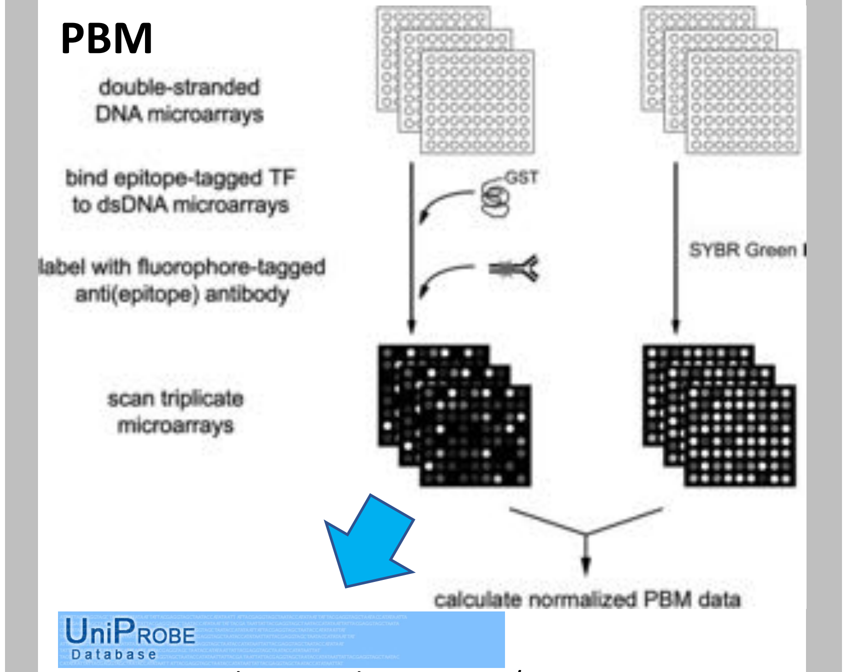


ChIP-seq


Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads
Nature Reviews | Genetics

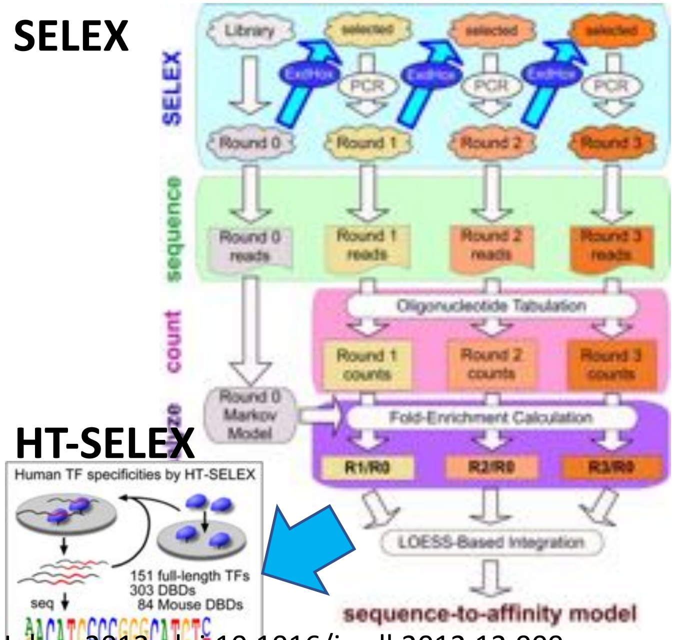
Park. 2009. doi: 10.1038/nrg2641

PBM


double-stranded DNA microarrays
bind epitope-tagged TF to dsDNA microarrays
label with fluorophore-tagged anti(epitope) antibody
scan triplicate microarrays
GST
SYBR Green I
calculate normalized PBM data
UniPROBE
Database

Berger et.al., 2006. doi: 10.1385/1-59745-097-9:245

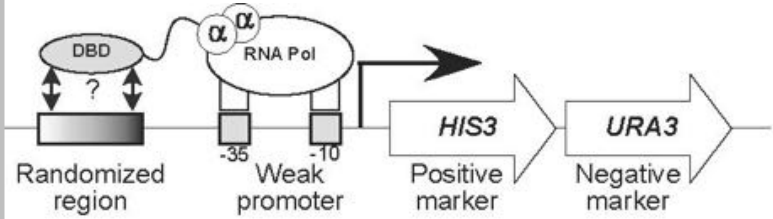
SELEX

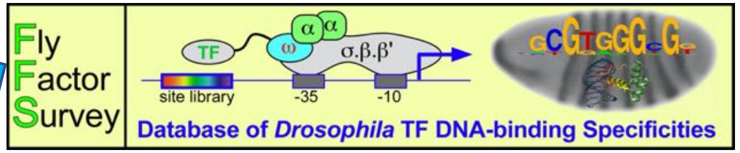


HT-SELEX

Jolma 2013. doi:10.1016/j.cell.2012.12.009
Slattery et.al., 2011. doi: 10.1016/j.cell.2011.10.053

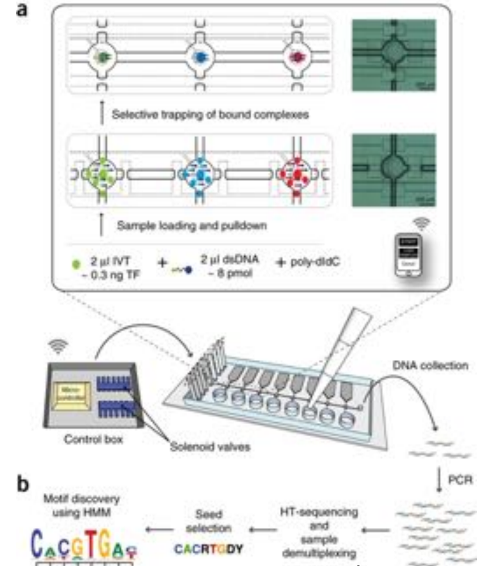
B1H





Meng et.al., 2005. doi: 10.1038/nbt1120

SMiLE-seq


a
Selective trapping of bound complexes
Sample loading and pulldown
2 µl IVT ~ 0.3 ng TF
2 µl dsDNA ~ 8 pmol
+ poly-dIdC
Control box
Solenoid valves
DNA collection
PCR
b
Motif discovery using HMM
Seed selection
HT-sequencing and sample demultiplexing
CACRTGDY

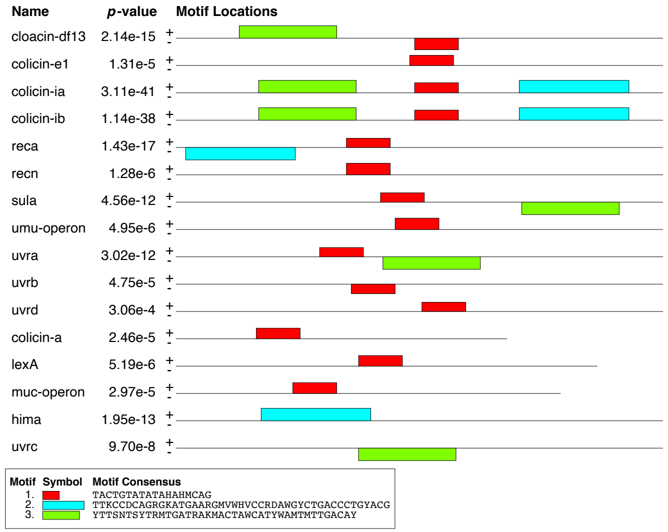
Isakova et.al., 2017. doi: 10.1038/nmeth.4143

J20 JASPAR2020

HOCOMOCO

CIS-BP

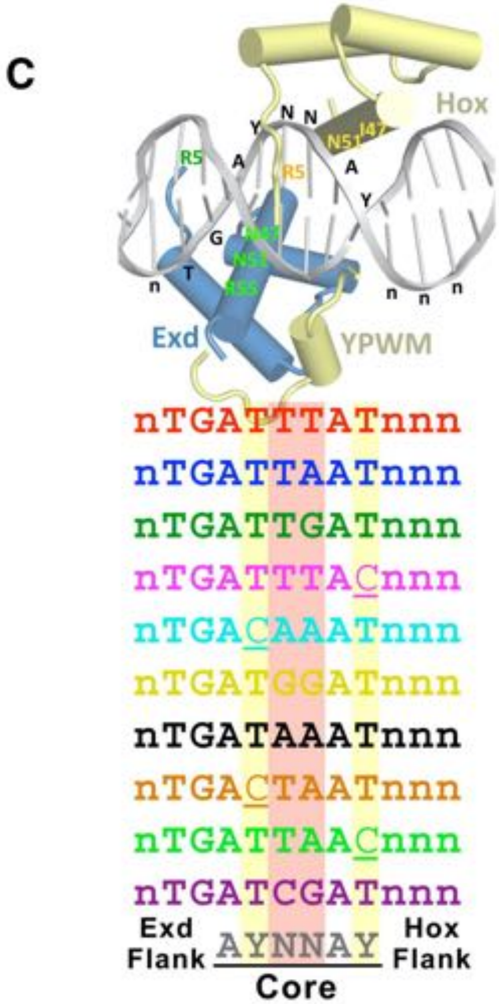
MotifDb



Name	p-value	Motif Locations
cloacin-df13	2.14e-15	
colicin-e1	1.31e-5	
colicin-ia	3.11e-41	
colicin-ib	1.14e-38	
reca	1.43e-17	
recn	1.28e-6	
sula	4.56e-12	
umu-operon	4.95e-6	
uvra	3.02e-12	
uvrb	4.75e-5	
uvrd	3.06e-4	
colicin-a	2.46e-5	
lexA	5.19e-6	
muc-operon	2.97e-5	
hima	1.95e-13	
uvrc	9.70e-8	

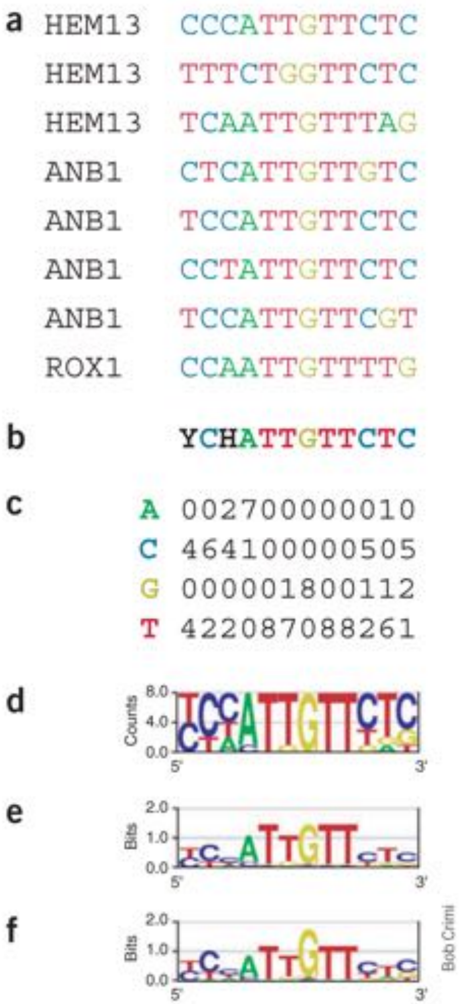
Motif	Symbol	Motif Consensus
1.	🟥	TACTGTATATAHAHMCAG
2.	🟦	TTKCCDCAGRGKATGAARGMVWHVCCRDAWGYCTGACCCTGYACG
3.	🟩	YTTSNTSYTRMTGATRAKMACTAWCATYWAMTMTTGACAY

Bailey et.al., 2015. doi: 10.1093/nar/gkv416



nTGATTTATnnn
nTGATTAATnnn
nTGATTGATnnn
nTGATTTACnnn
nTGACAAATnnn
nTGATGGATnnn
nTGATAAATnnn
nTGACTAATnnn
nTGATTAACnnn
nTGATCGATnnn

Slattery et.al., 2011. doi: 10.1016/j.cell.2011.10.053



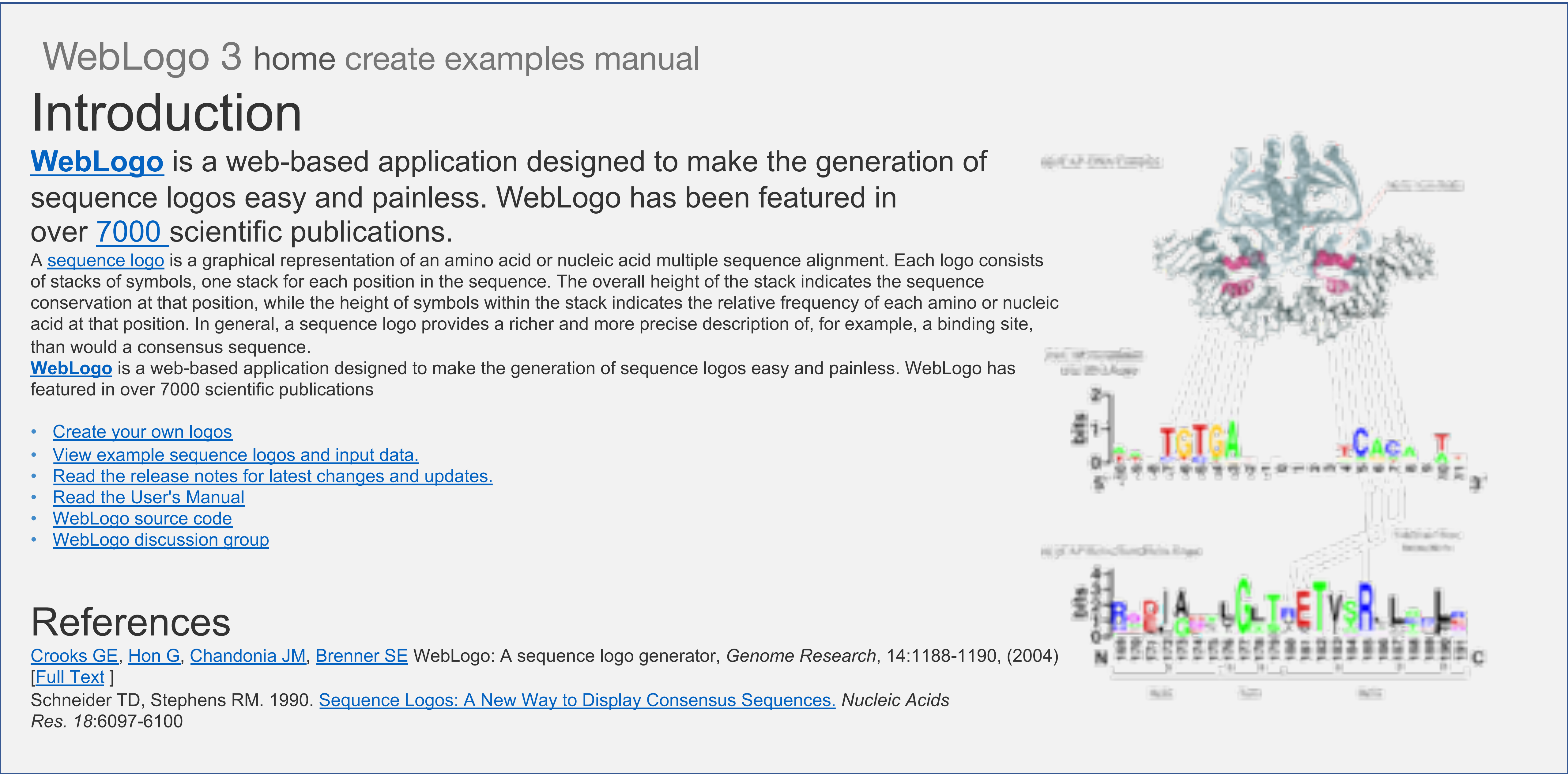
```
a HEM13 CCCATTGTTCTC
  HEM13 TTTCTGGTTCTC
  HEM13 TCAATTGTTTAG
  ANB1  CTCATTGTTGTC
  ANB1  TCCATTGTTCTC
  ANB1  CCTATTGTTCTC
  ANB1  TCCATTGTTCGT
  ROX1  CCAATTGTTTTG

b       YCHATTGTTCTC

c     A 002700000010
      C 464100000505
      G 000001800112
      T 422087088261
```

D’haeseleer. 2006. doi: 10.1038/nbt0406-423

Plot single sequence logo



WebLogo 3 home create examples manual

Introduction

WebLogo is a web-based application designed to make the generation of sequence logos easy and painless. WebLogo has been featured in over 7000 scientific publications.

A sequence logo is a graphical representation of an amino acid or nucleic acid multiple sequence alignment. Each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence.

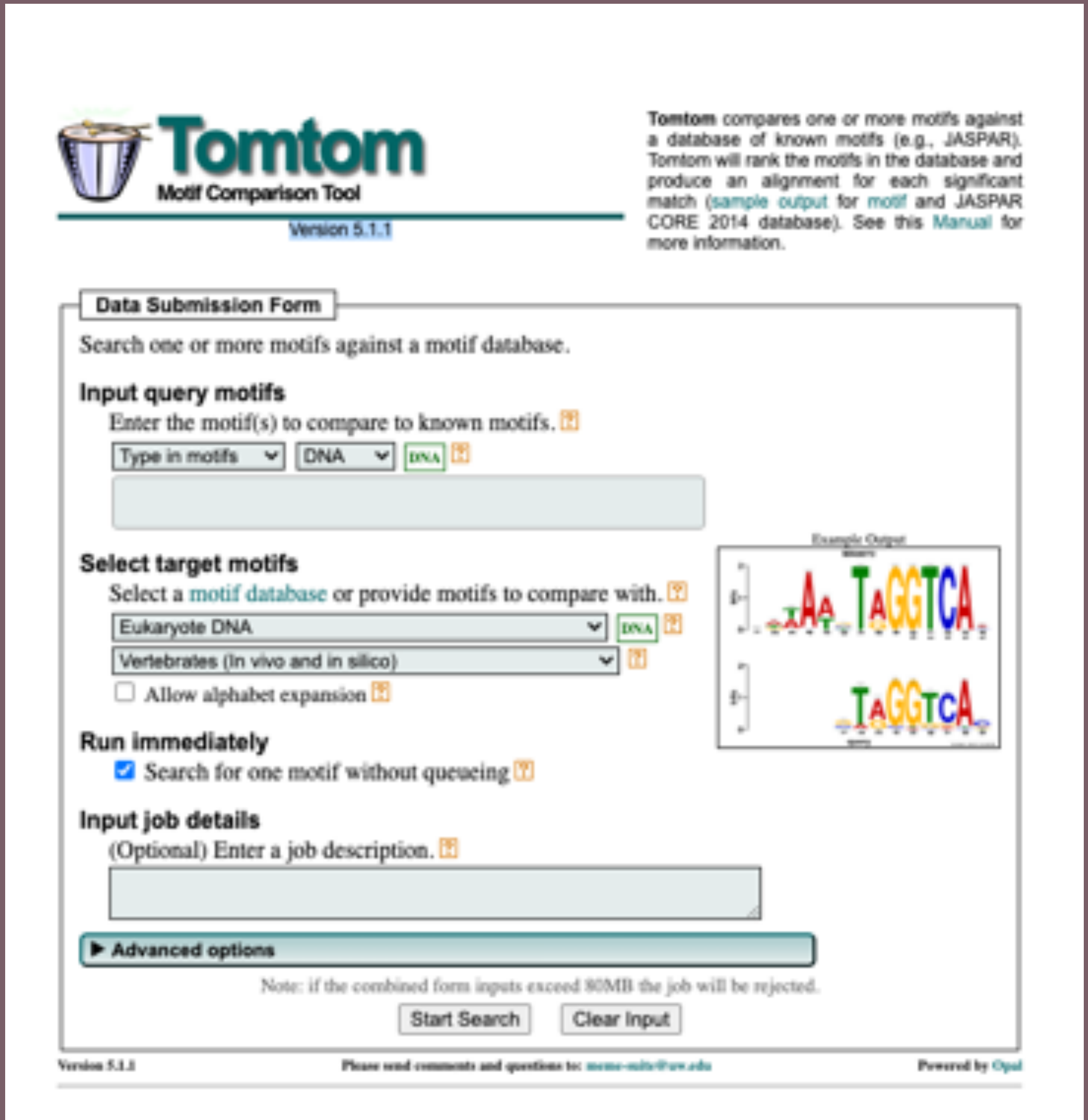
WebLogo is a web-based application designed to make the generation of sequence logos easy and painless. WebLogo has featured in over 7000 scientific publications

- Create your own logos
- View example sequence logos and input data.
- Read the release notes for latest changes and updates.
- Read the User's Manual
- WebLogo source code
- WebLogo discussion group

References

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004) [Full Text]

Schneider TD, Stephens RM. 1990. Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res. 18*:6097-6100



Tomtom compares one or more motifs against a database of known motifs (e.g., JASPAR). Tomtom will rank the motifs in the database and produce an alignment for each significant match (sample output for motif and JASPAR CORE 2014 database). See this Manual for more information.

Data Submission Form

Search one or more motifs against a motif database.

Input query motifs

Enter the motif(s) to compare to known motifs.

Type in motifs | DNA | DNA

Select target motifs

Select a motif database or provide motifs to compare with.

Eukaryote DNA | DNA

Vertebrates (In vivo and in silico)

□ Allow alphabet expansion

Run immediately

☑ Search for one motif without queueing

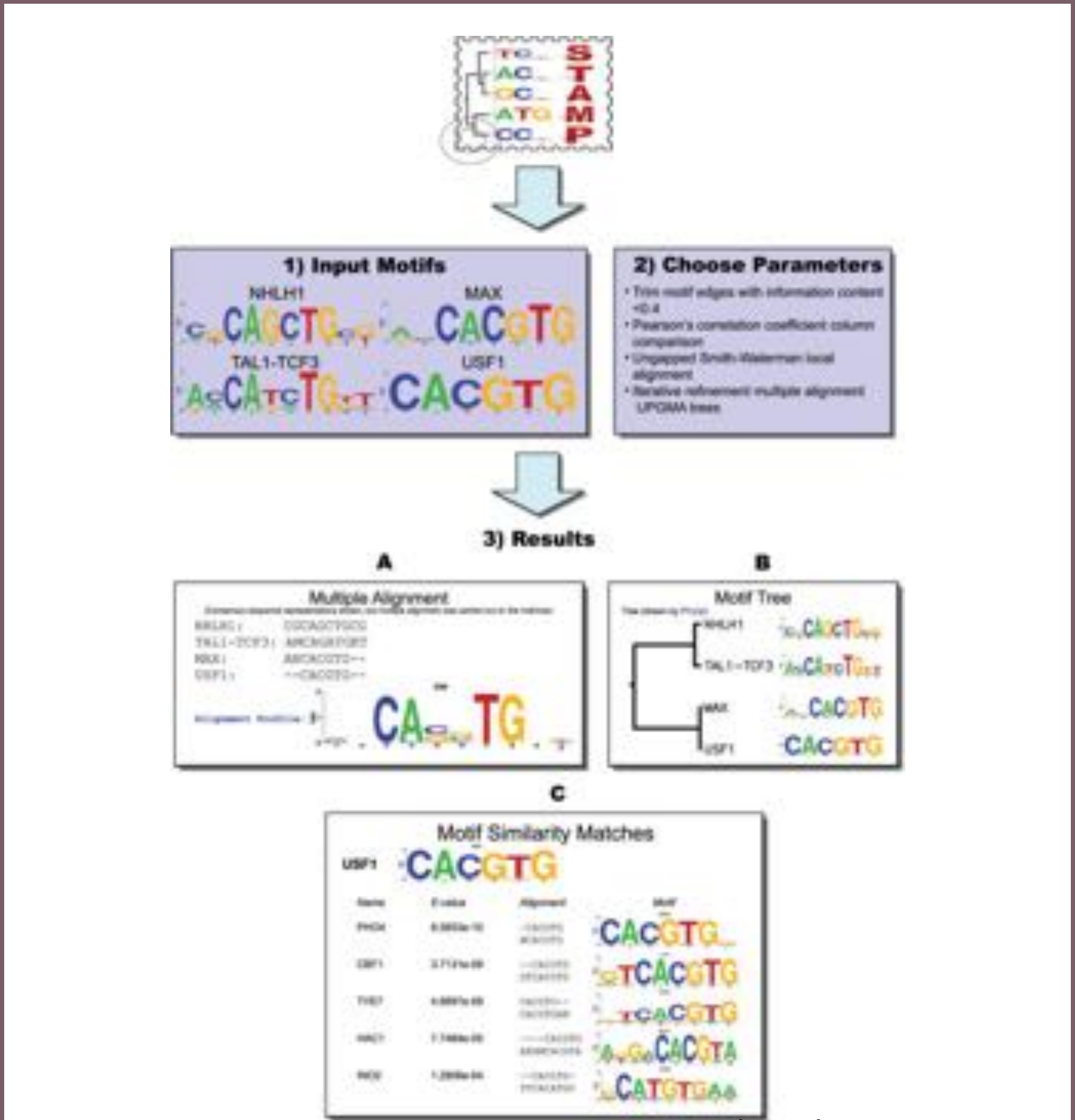
Input job details

(Optional) Enter a job description.

Note: if the combined form inputs exceed 80MB the job will be rejected.

Version 5.1.1 Please send comments and questions to: meme-suite@uw.edu Powered by Opal

Gupta et.al., 2007. doi: 10.1186/gb-2007-8-2-r24



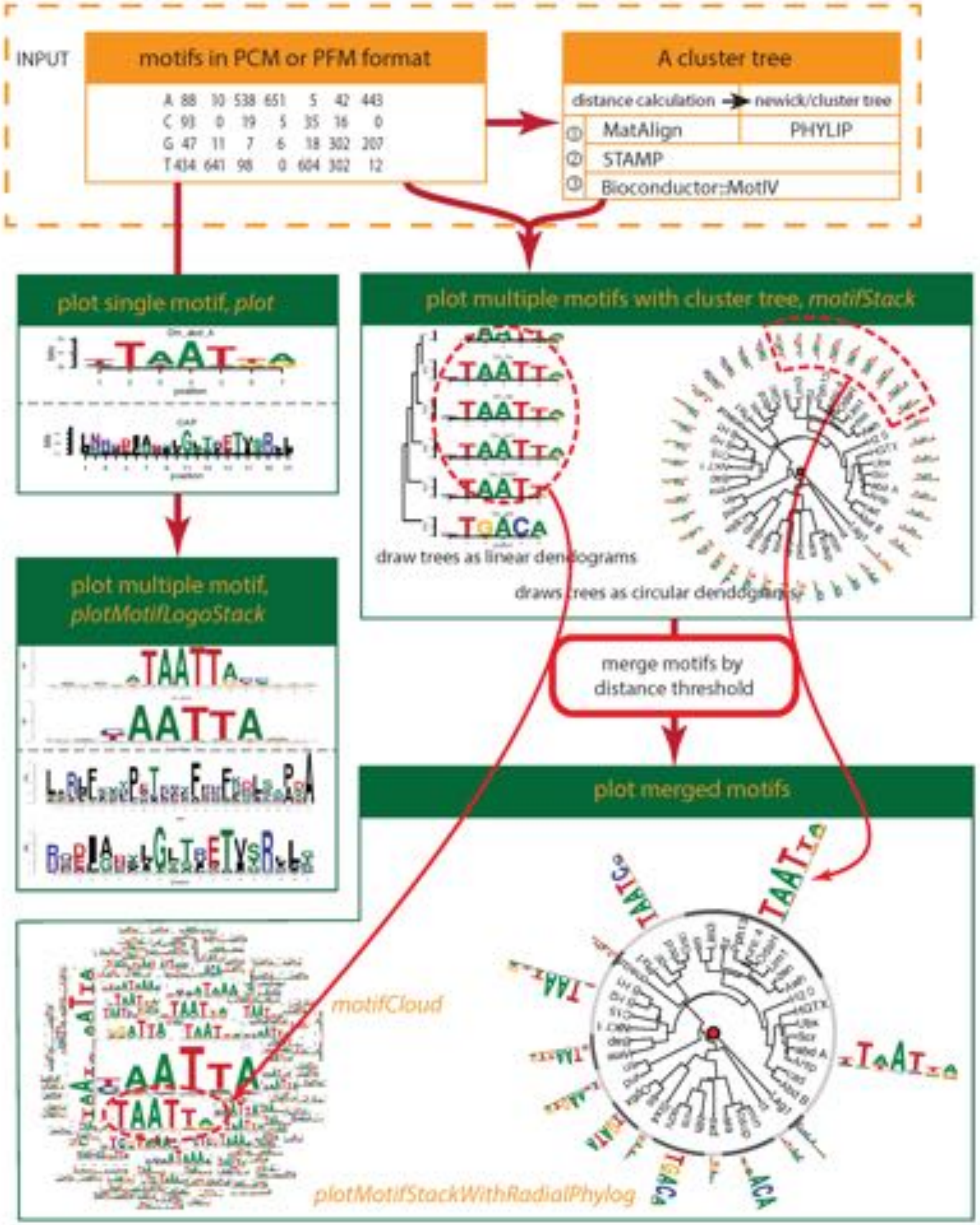
STAMP
1) Input Motifs
NHLH1
MAX
TAL1-TCF3
USF1
2) Choose Parameters
• Trim motif edges with information content <0.4
• Pearson's correlation coefficient column comparison
• Ungapped Smith-Waterman local alignment
• Iterative refinement multiple alignment UPGMA trees
3) Results
A
Multiple Alignment
B
Motif Tree
NHLH1
TAL1-TCF3
MAX
USF1
C
Motif Similarity Matches
USF1
Name
E-value
Alignment
Motif

Mahony et.al., 2007. doi: 10.1093/nar/gkm272

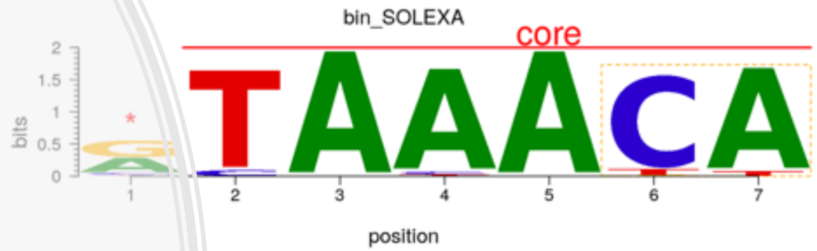
From single motif to multiple motifs

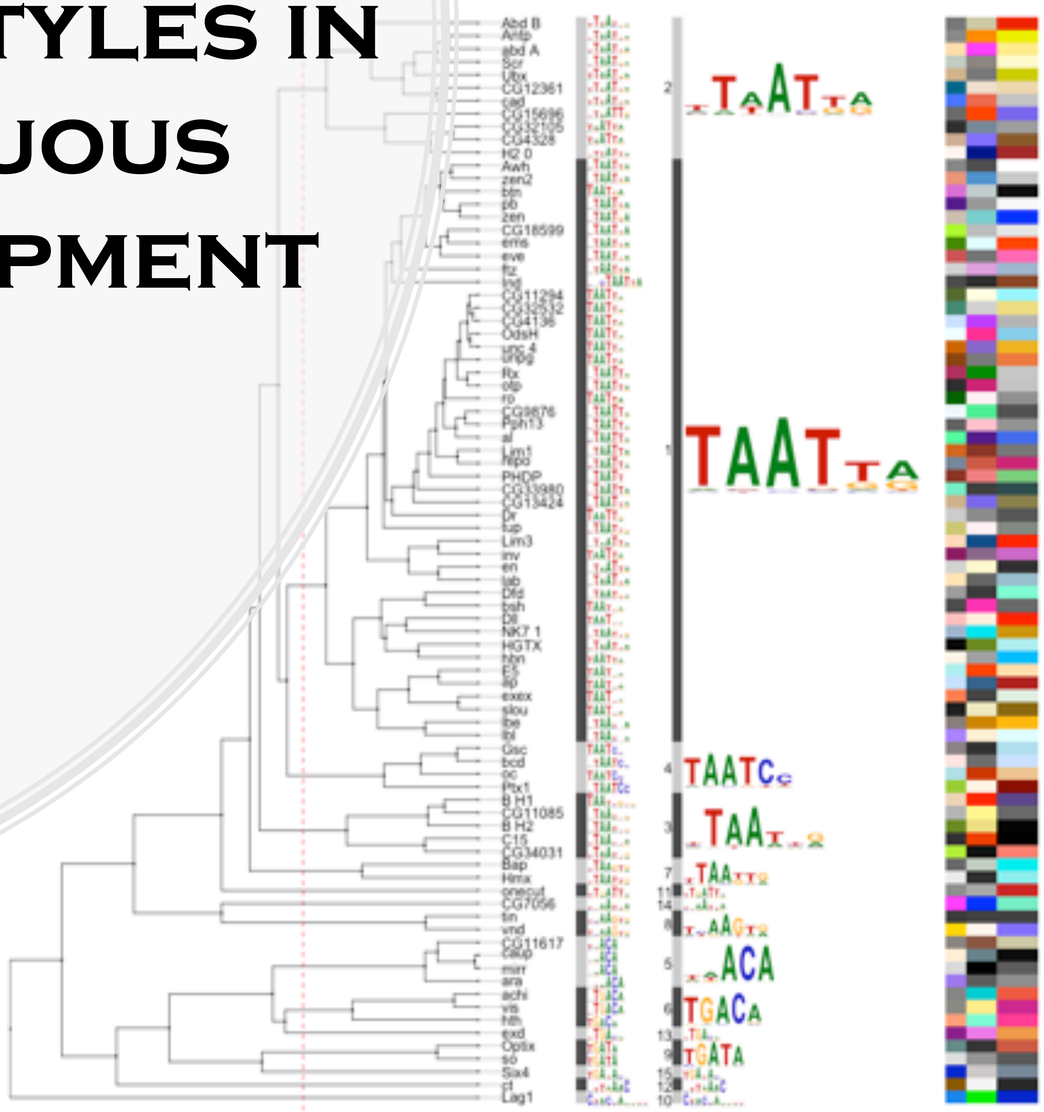
- ❈ Plot aligned motifs
- ❈ Powerful tool to visualize bunch of sequence logos
- ❈ Highlight grouped motifs by their signatures
- ❈ Multiple style and technique to show and label motifs

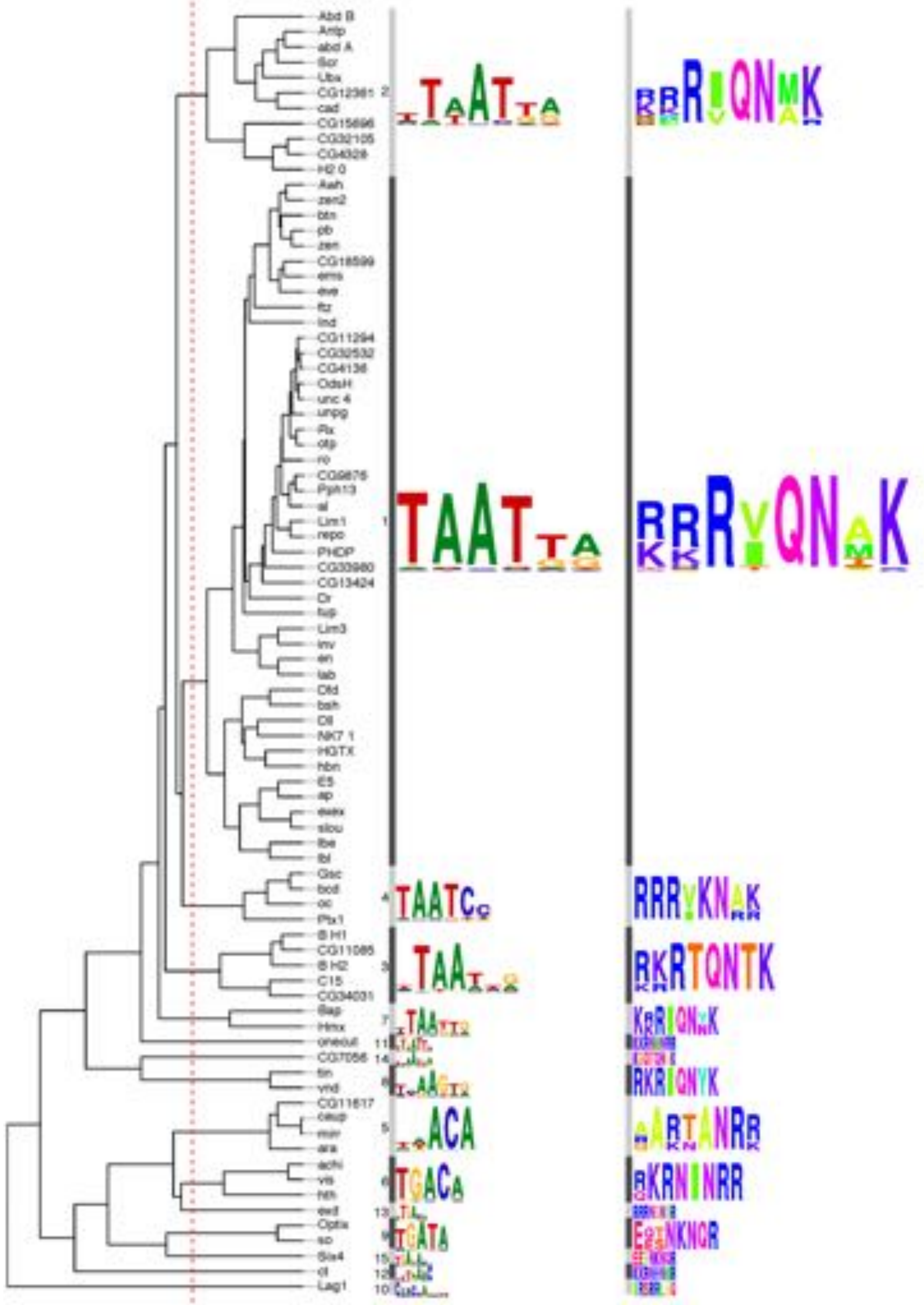
Ou J, Wolfe SA, Brodsky MH, Zhu LJ (2018) doi: 10.1038/nmeth.4555


INPUT
motifs in PCM or PFM format
A 88 10 538 651 5 42 443
C 93 0 19 5 35 16 0
G 47 11 7 6 18 302 207
T 434 641 98 0 604 302 12
A cluster tree
distance calculation → newick/cluster tree
① MatAlign
PHYLIP
② STAMP
③ Bioconductor::MotIV
plot single motif, plot
plot multiple motifs with cluster tree, motifStack
draw trees as linear dendrograms
draws trees as circular dendrograms
plot multiple motif, plotMotifLogoStack
merge motifs by distance threshold
plot merged motifs
motifCloud
plotMotifStackWithRadialPhylog

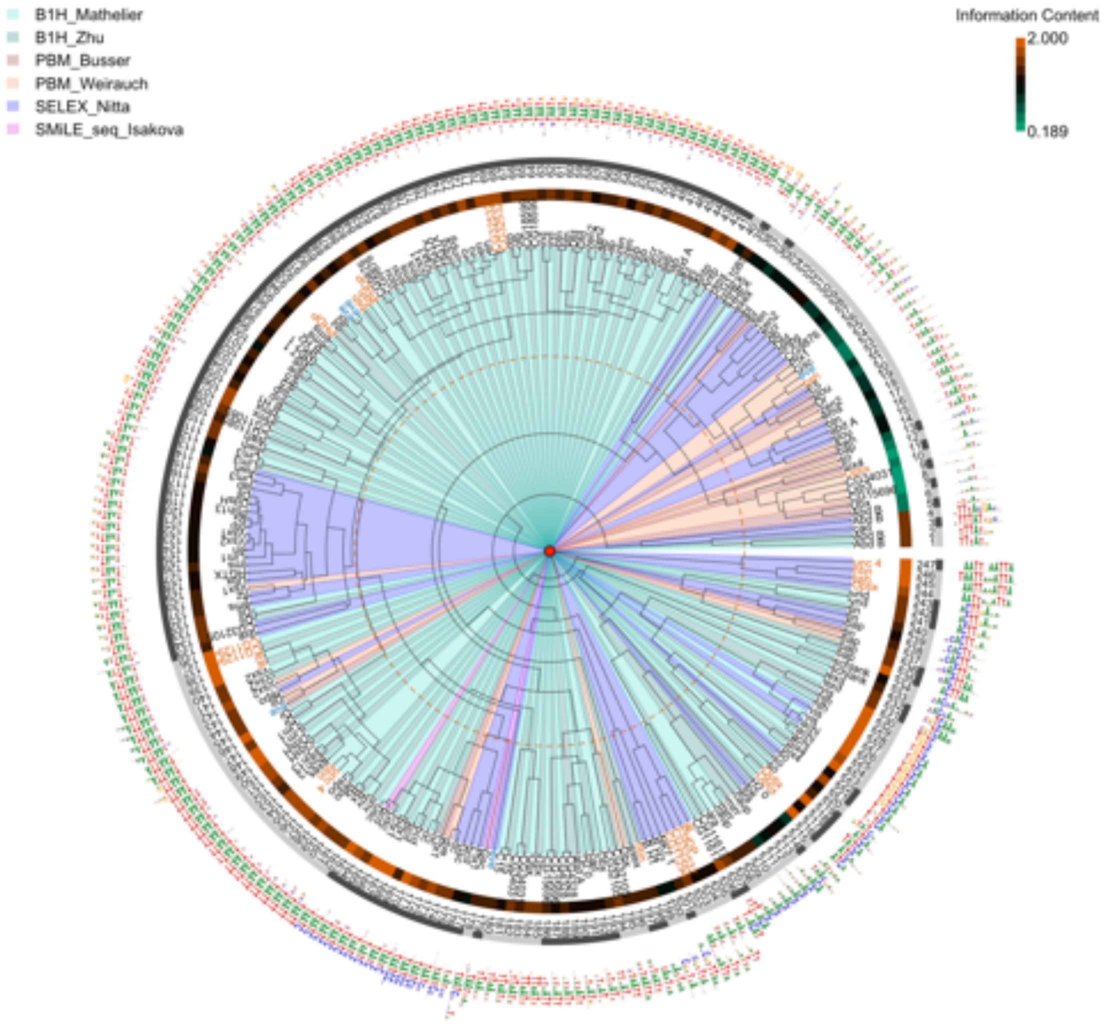
More and more styles in continuous development







What will we make today



INSTALL motifStack PACKAGE

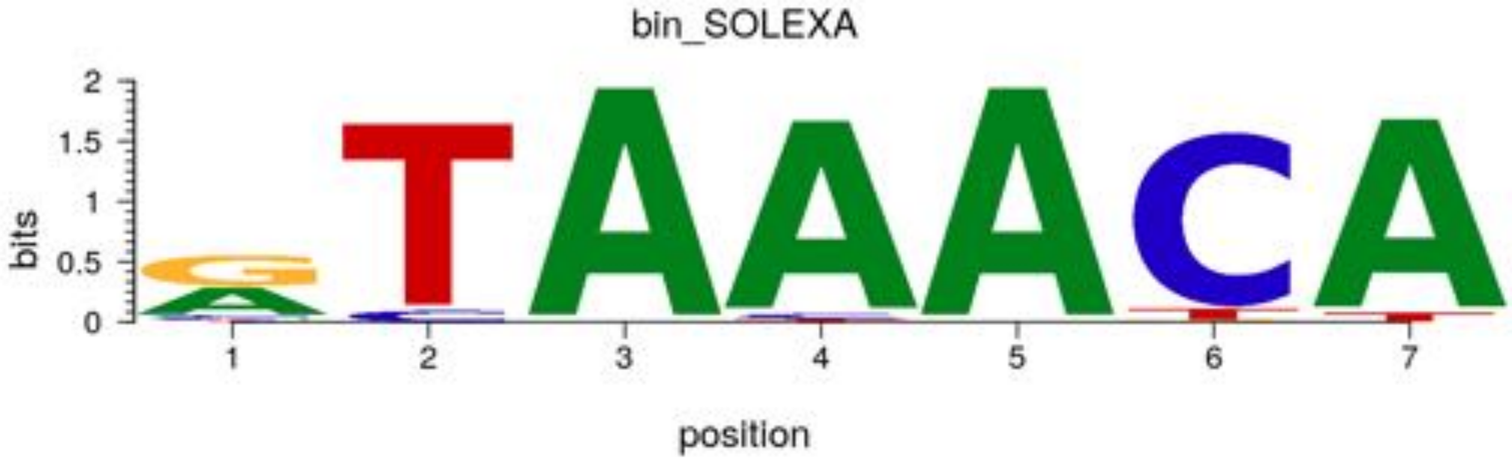
```
if(packageVersion("motifStack")<"1.33.3"){
  BiocManager::install("jianhong/motifStack", build_vignettes=TRUE)
}
```

- Starting from version 1.33.2, *motifStack* does not require cario or ghostscript anymore. It will use cario if cario (>=1.6) is install or use ghostscript if gs command is available. Otherwise, *motifStack* will use embed font to plot the sequence logo.

- MatAlign algorithm was included in *motifStack* since 1.33.2.

PLOT A DNA SEQUENCE LOGO

```
library(motifStack)
pcm <- importMatrix(system.file(" extdata ", "bin_SOLEXA.pcm” , package = " motifStack "),
                                        format = "pcm", to = "pcm")
plot(pcm)
```



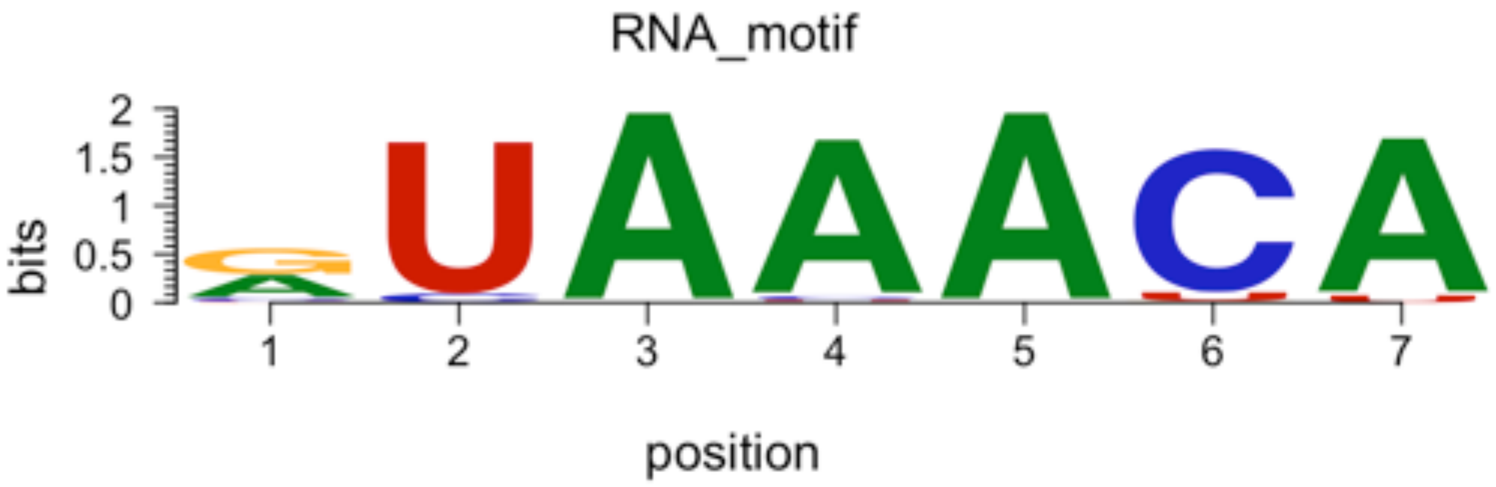
$$PWM(b,i) = \log \frac{PFM(b,i)}{P(b)}$$

$$IC = \sum_b PFM_{b,i} \log \frac{PFM(b,i)}{P(b)} = \sum_b PFM_{b,i} PWM(b,i)$$

Strormo 2000. doi: 10.1093/bioinformatics/16.1.16

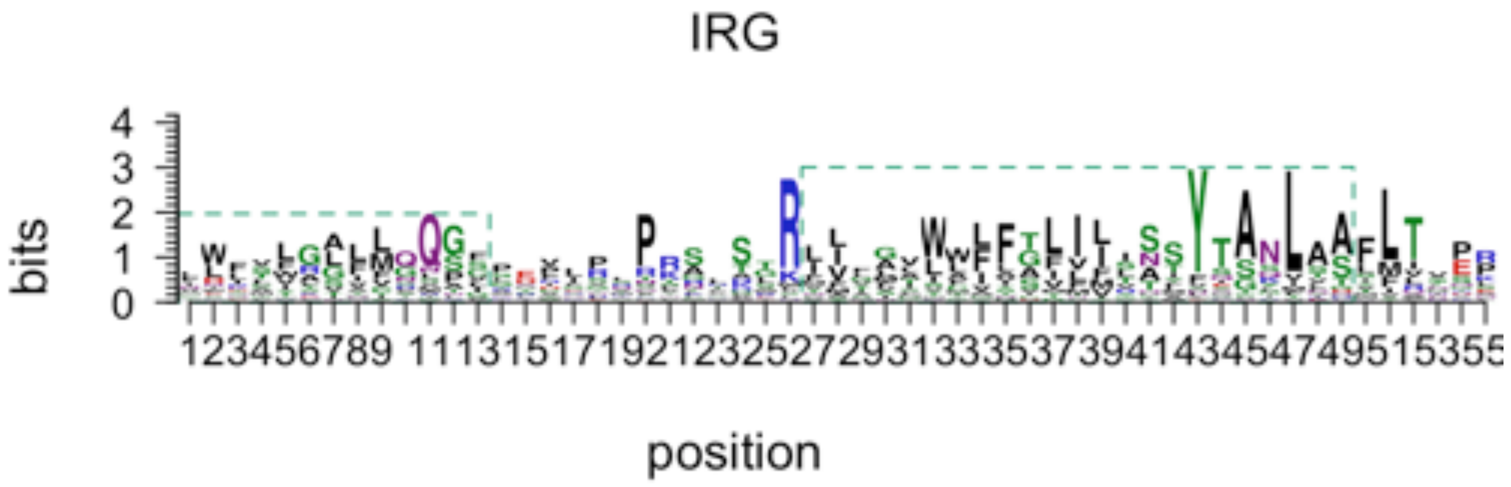
PLOT AN RNA SEQUENCE LOGO

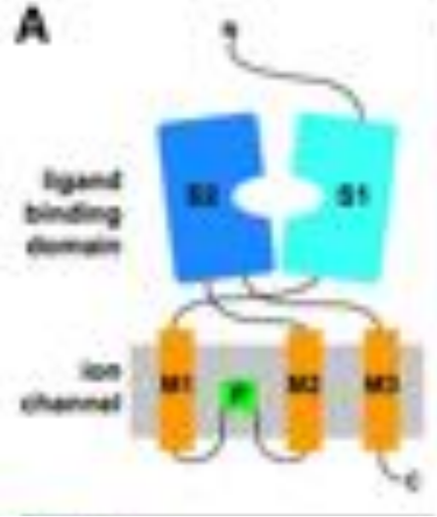
```
library(motifStack)
pcm <- read.table(file.path(find.package("motifStack"), "extdata", "bin_SOLEXA.pcm"))
pcm <- pcm[,3:ncol(pcm)]
rownames(pcm) <- c("A","C","G","U")
motif <- new("pcm", mat=as.matrix(pcm), name="bin_SOLEXA")
plot(motif)
```



PLOT AN AMINO ACID (AA) SEQUENCE LOGO

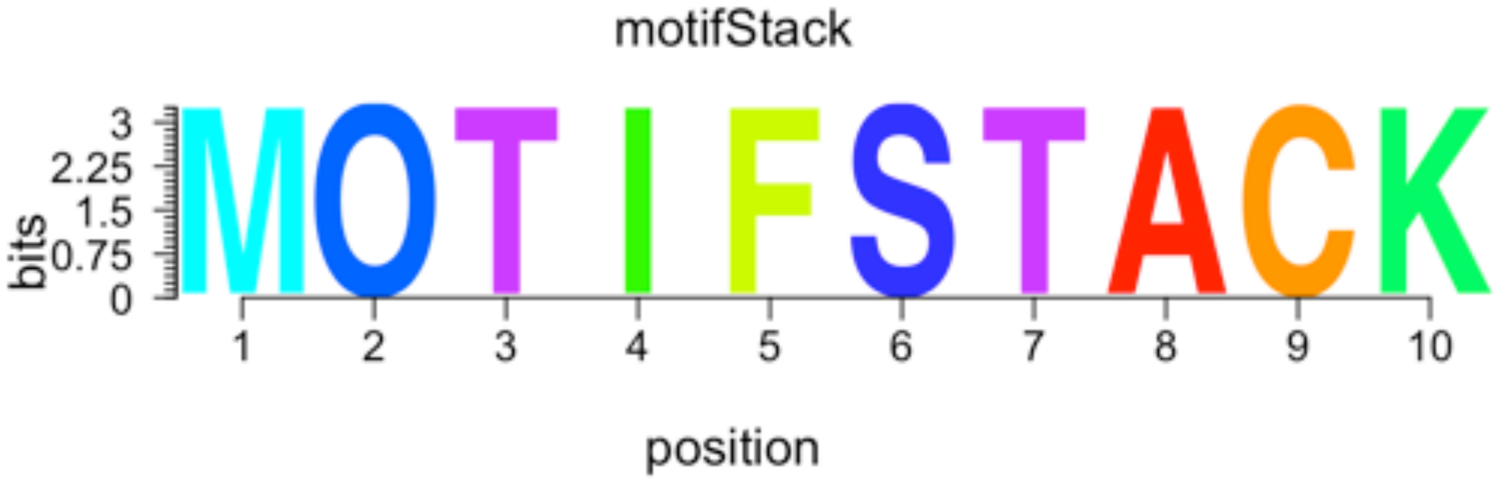
```
library(Biostrings)
protein<-read.table(system.file("extdata", "motifStack", "irg.txt",
                                package = "workshop2020"))
protein<-t(protein[,2:21])
rownames(protein) <- sort(AA_STANDARD)
protein_motif<-new("pcm", mat=protein, name="IRG",
                     color=colorset(alphabet="AA",colorScheme="chemistry"),
                     alphabet = "AA",
                     markers=list(new("marker", type="rect", start=c(1,27), stop=c(13,49),
                                          gp=gpar(col="#009E73", fill=NA, lty=2))))
plot(protein_motif)
```





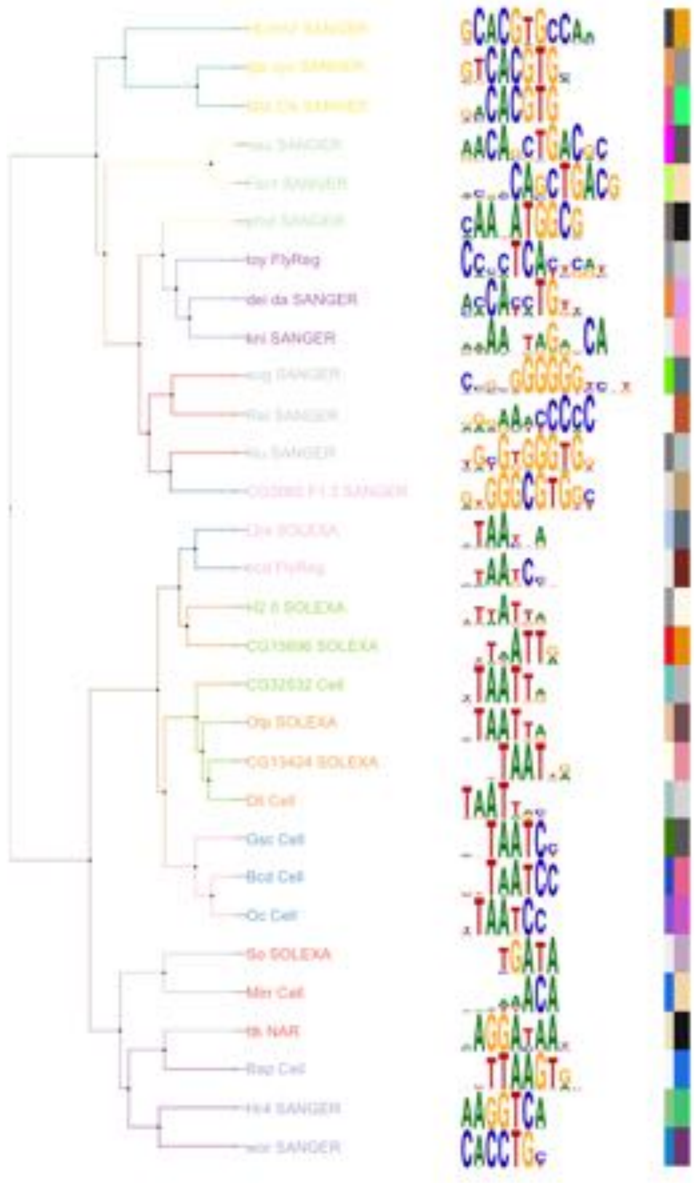
PLOT A CUSTOMIZED LOGO

```
m <- matrix(0, nrow = 10, ncol = 10,
            dimnames = list(strsplit("motifStack", "")[[1]],
                            strsplit("motifStack", "")[[1]]))
for(i in seq.int(10)) m[i, i] <- 1
ms <- new("pfm", mat=m, name="motifStack")
plot(ms)
```



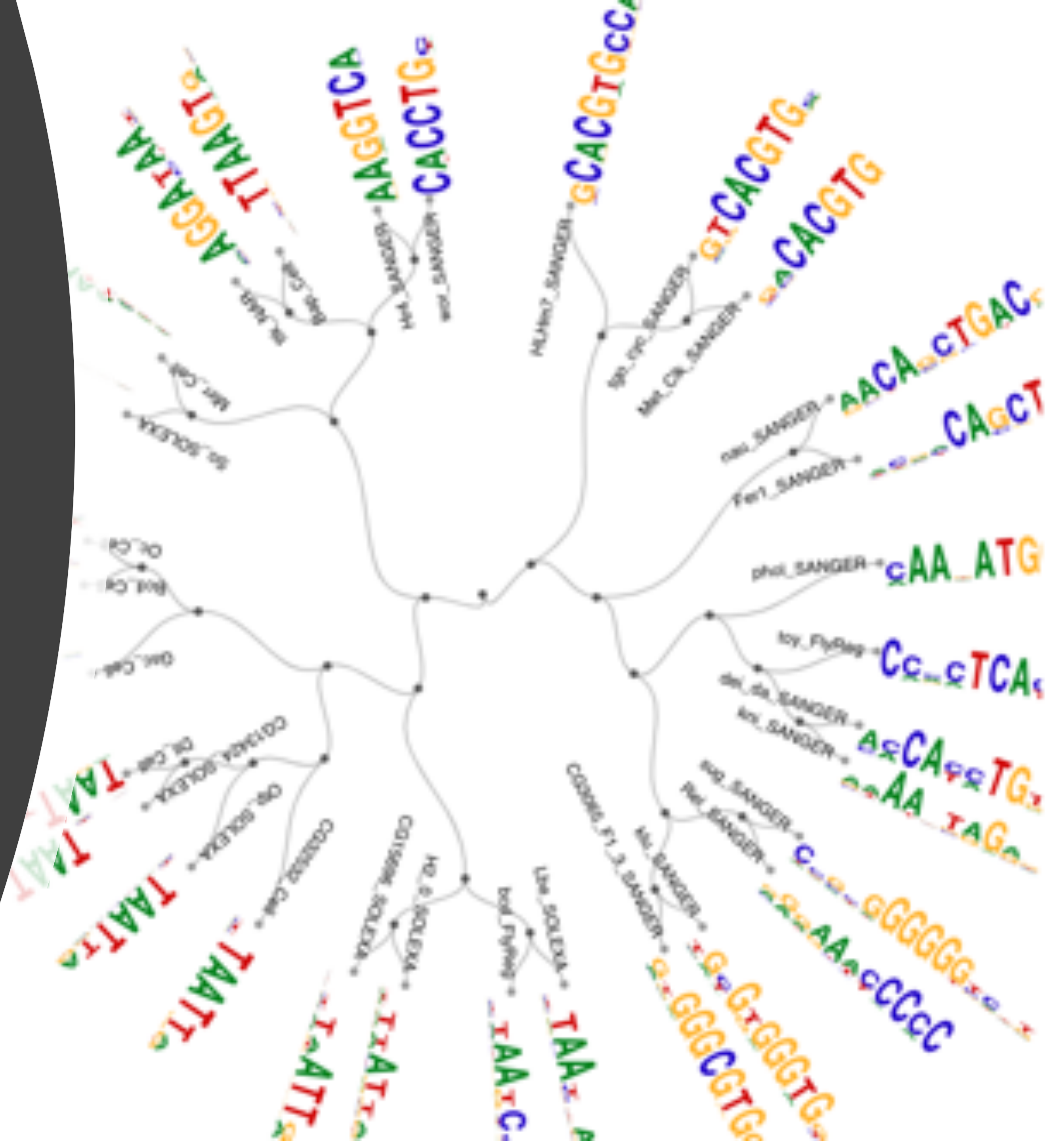
Plot multiple sequence logos

```
library(MotifDb); library(ade4); library(RColorBrewer)
matrix.fly <- MotifDb::query(MotifDb, "FlyFactorSurvey")
motifs2 <- as.list(matrix.fly)
## format the name
names(motifs2) <- gsub("(_[\\.0-9]+)*_FBgn\\d+$", "",
                        elementMetadata(matrix.fly)$providerName)
names(motifs2) <- gsub("[^a-zA-Z0-9]", "_", names(motifs2))
motifs2 <- motifs2[unique(names(motifs2))]
## subsample motifs
set.seed(1); pfms <- sample(motifs2, 30)
## cluster the motifs
hc <- clusterMotifs(pfms)
## convert the hclust to phylog object
phylog <- ade4::hclust2phylog(hc)
## reorder the pfms by the order of hclust
leaves <- names(phylog$leaves)
pfms <- pfms[leaves]
## create a list of pfm objects
pfms <- mapply(pfms, names(pfms), FUN=function(.pfm, .name){
                  new("pfm",mat=.pfm, name=.name)})
color <- brewer.pal(12, "Set3")
## plot the logo stack with pile style.
motifPiles(phylog=phylog, pfms=pfms,
           col.tree=rep(color, each=3), col.leaves=rep(rev(color), each=3),
           r.anno=c(0.02, 0.03), col.anno=list(sample(colors(), 30), sample(colors(), 30)))
```



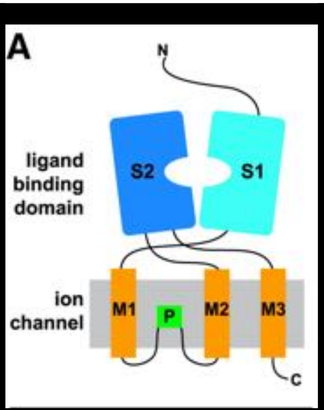
Plot interactive sequence logos

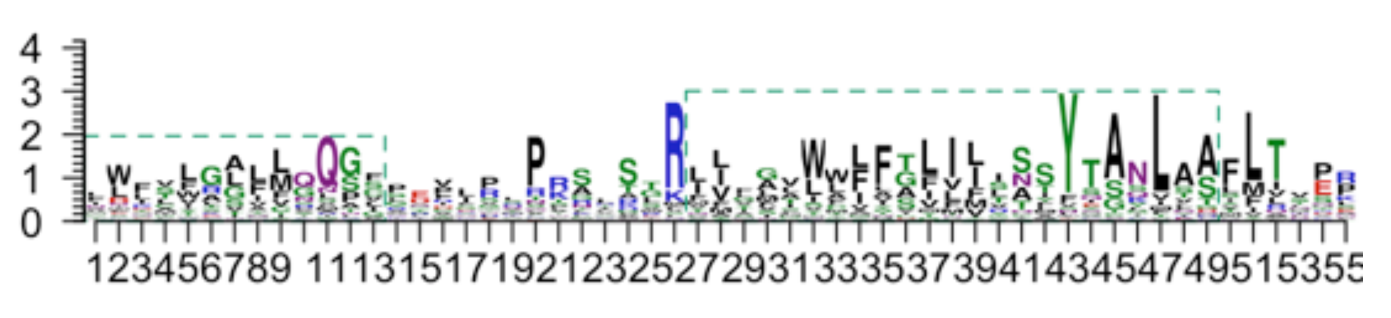
```
browseMotifs(pfms = pfms, phylog = phylog,
             layout="radialPhylog",
             yaxis = FALSE, xaxis = FALSE)
```

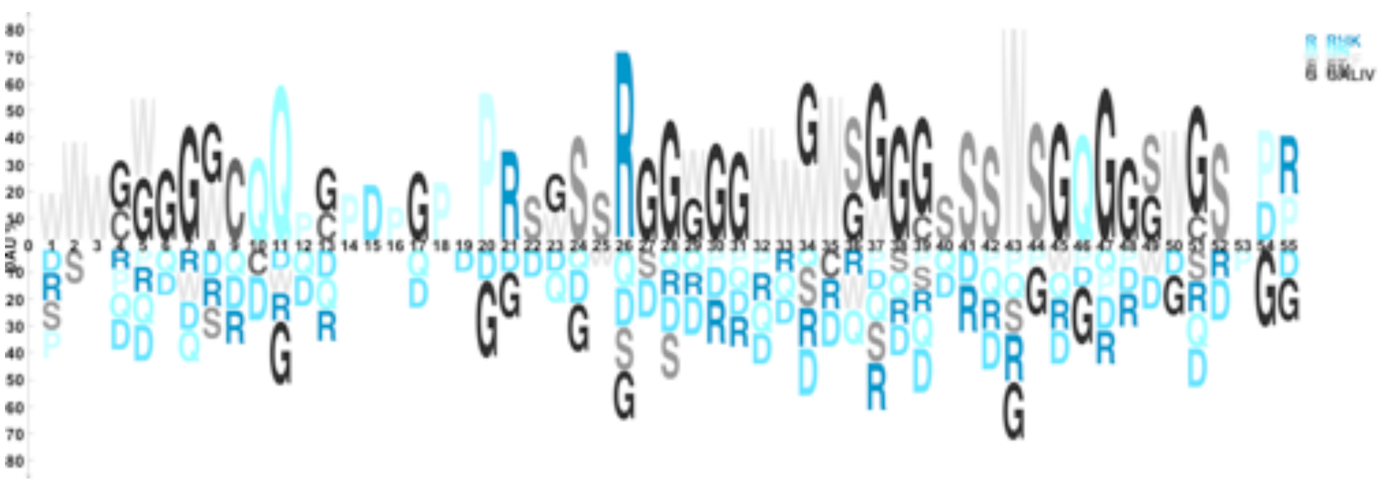


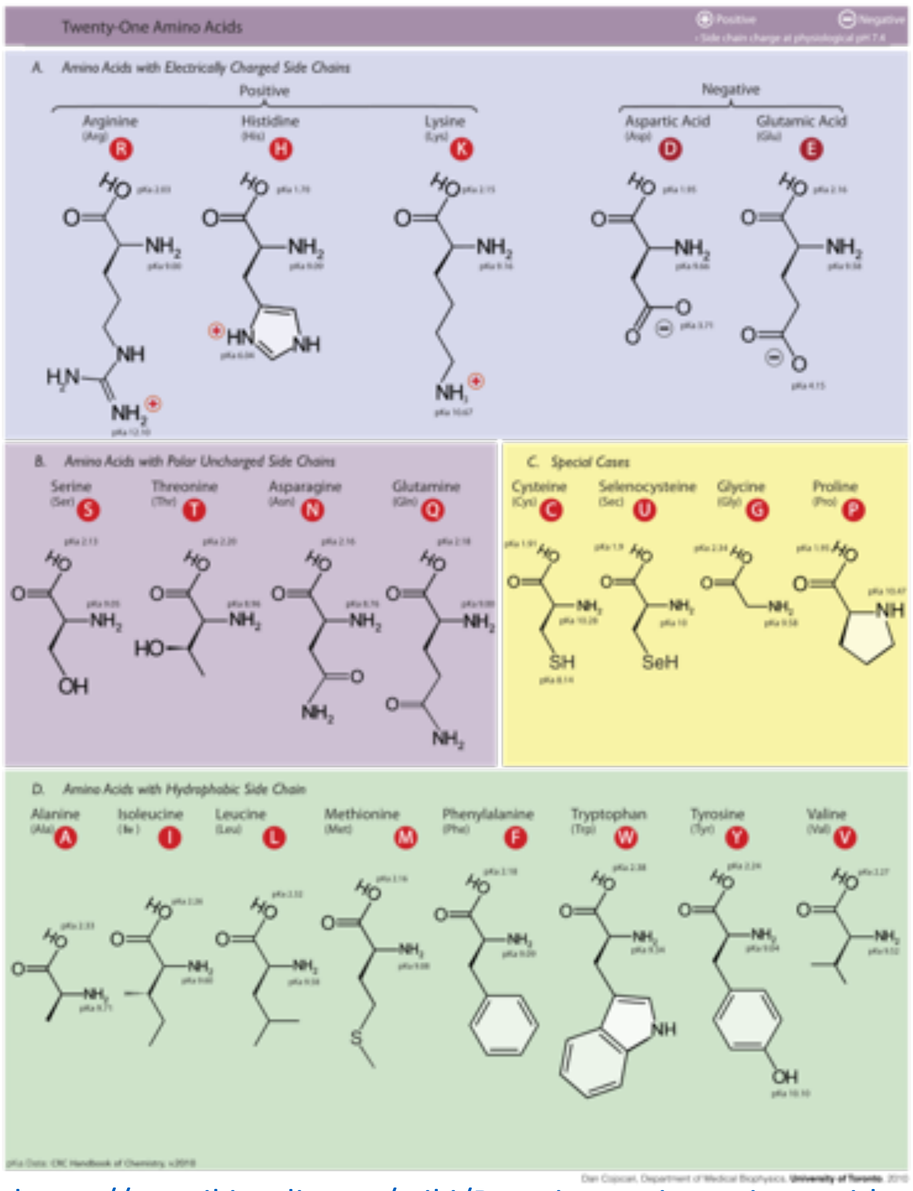
GOTO VIGNETTE

dagLogo



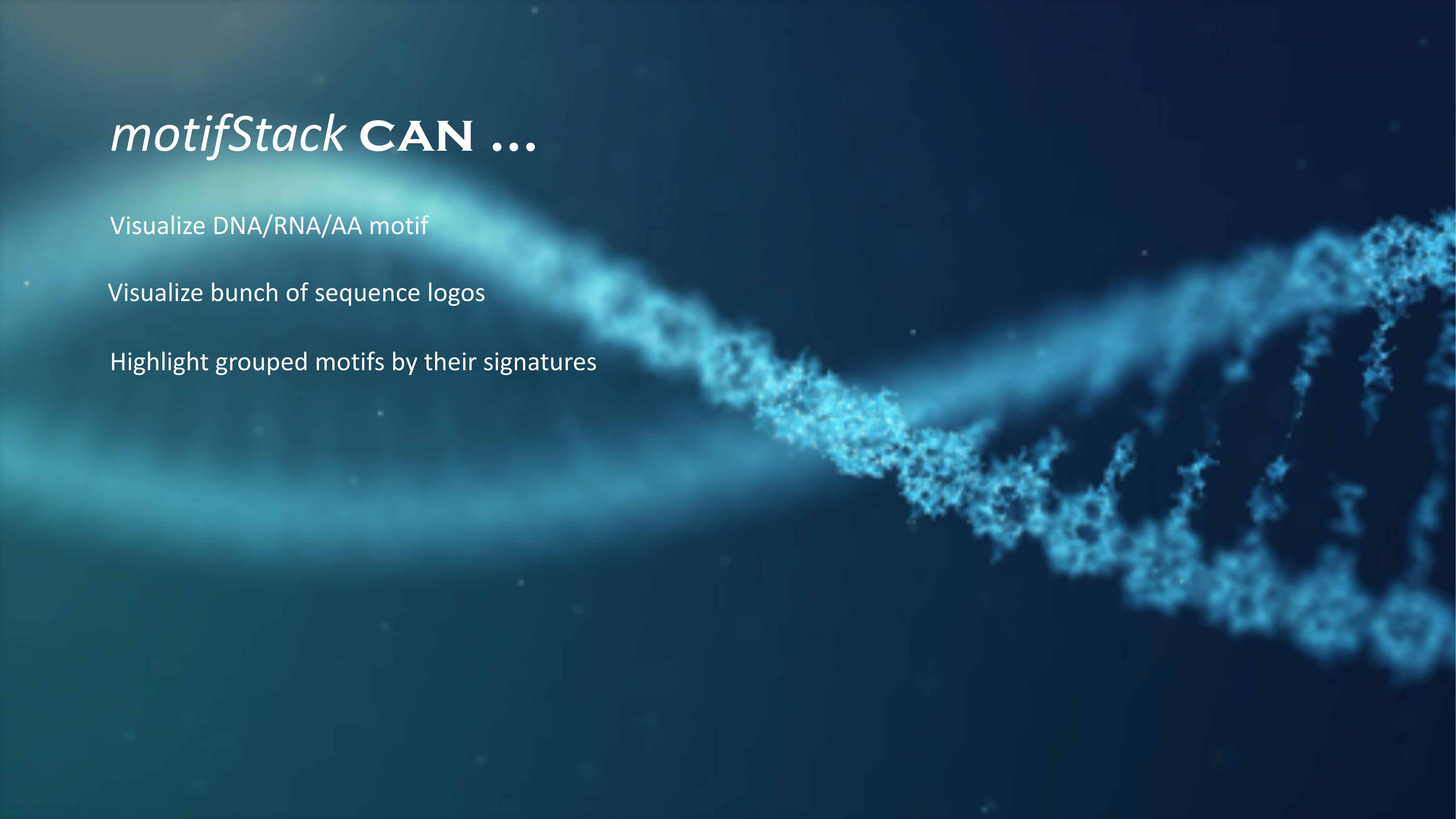






https://en.wikipedia.org/wiki/Proteinogenic_amino_acid

GOTO VIGNETTE

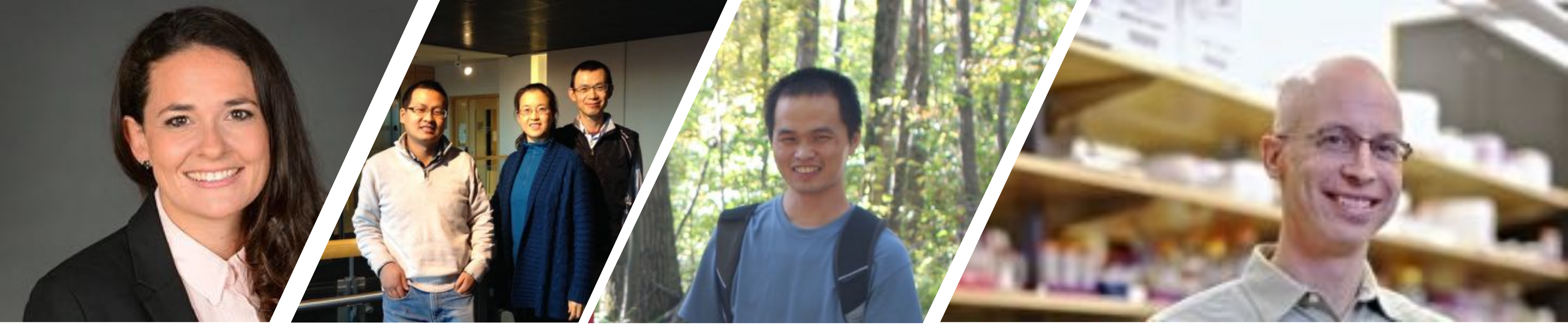


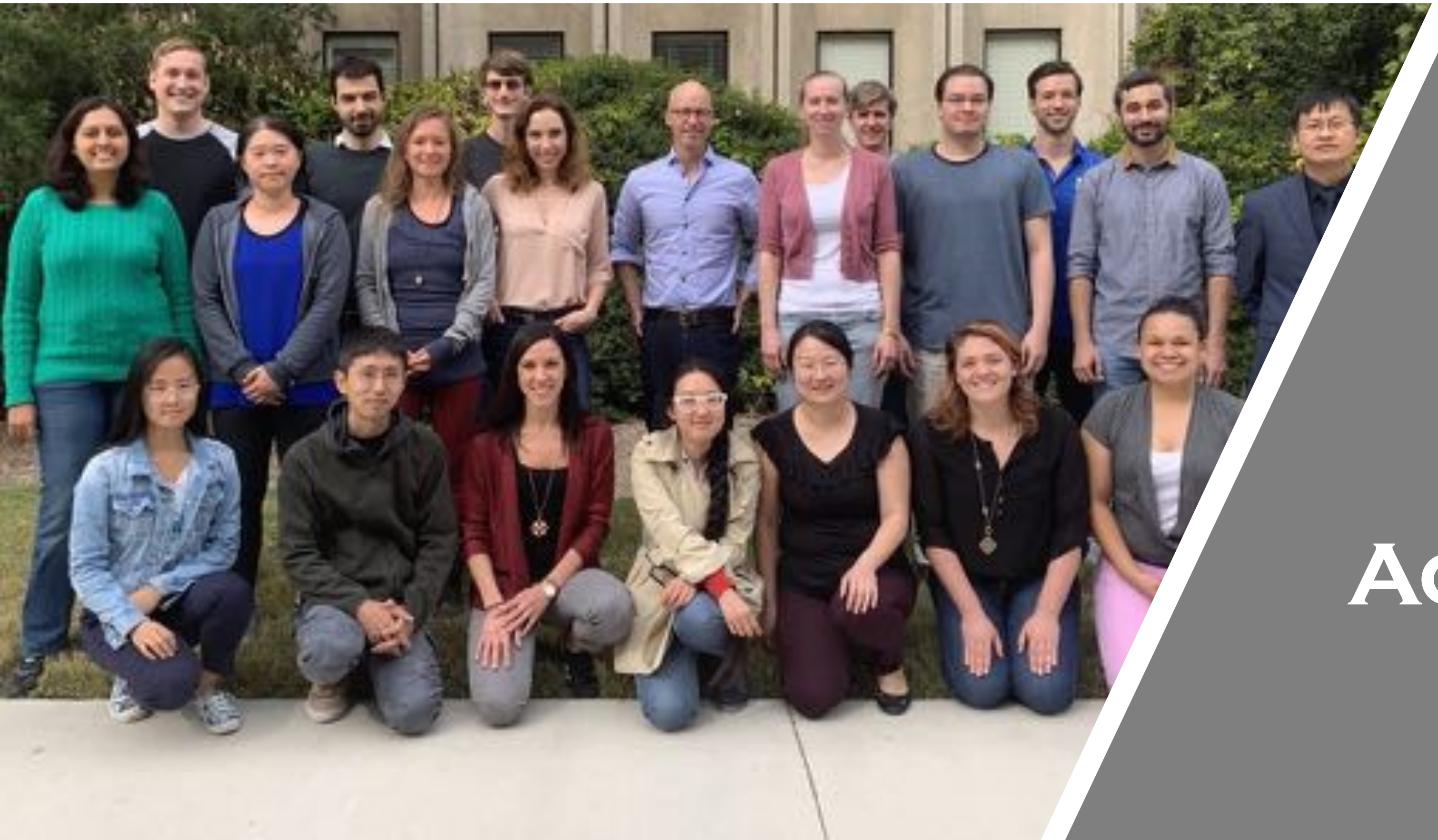
motifStack **CAN …**

Visualize DNA/RNA/AA motif

Visualize bunch of sequence logos

Highlight grouped motifs by their signatures





Acknowledgement