Heatmaps generated from HMM peptide clustering

Tomas Bjorklund

Tue Nov 10 12:06:41 2020

This script clusters Polypeptide motifs using the Hammock hidden Markov model peptide clustering and generates Heatmaps for most functional motifs.

```
suppressPackageStartupMessages(library(knitr))
```

Loading samples

```
all.samples <- readRDS("data/allSamplesDataTable.RDS")
all.samples[, `:=`(Peptide, as.character(Peptide)), ]

setkey(all.samples, Group)
```

Generation of heatmaps for in vivo transported samples

```
select.samples <- all.samples[J(c("mRNA_30cpc_Str", "mRNA_3cpc_Str", "mRNA_30cpc_SN",
    "mRNA_3cpc_SN", "mRNA_30cpc_Th", "mRNA_3cpc_Th", "mRNA_30cpc_Ctx", "mRNA_3cpc_Ctx",
    "mRNA_30cpc_Str_4wks", "mRNA_3cpc_Str_4wks", "mRNA_30cpc_SN_4wks", "mRNA_3cpc_SN_4wks",
    "mRNA_30cpc_Th_4wks", "mRNA_3cpc_Th_4wks", "mRNA_30cpc_Ctx_4wks", "mRNA_3cpc_Ctx_4wks"))]

select.samples[, `:=`(BCcount, as.integer(mclapply(BC, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
select.samples[, `:=`(Animalcount, as.integer(mclapply(Animals, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
select.samples[, `:=`(Score, BCcount + Animalcount - 1), ]
select.samples.trsp <- unique(select.samples, by = c("Animals", "BC", "LUTnrs"))

fasta.names <- paste(1:nrow(select.samples.trsp), select.samples.trsp$Score,
    select.samples.trsp$Group, sep = "|")
write.fasta(as.list(select.samples.trsp$Peptide), fasta.names, "data/invivoSamplesPeptides.fasta",
    open = "w", nbchar = 60, as.string = TRUE)

# Generate Scoring table for Weblogo Weighting
select.samples.pepMerge <- select.samples.trsp[, sum(Score), by = c("Peptide")]
setnames(select.samples.pepMerge, "V1", "Score")
```

Executing Hammock Clustering

```
Sys.setenv(PATH = paste("/root/HMMER/binaries", Sys.getenv("PATH"), sep = ":"),
    HHLIB = "/home/rstudio/Hammock_v_1.1.1/hhsuite-2.0.16/lib/hh/")
unlink("/home/rstudio/data/HammockInVivo", recursive = TRUE, force = FALSE)
sys.out <- system(paste("java -jar /home/rstudio/Hammock_v_1.1.1/dist/Hammock.jar full -i /home/rstudio/data/
    detectCores(), sep = ""), intern = TRUE, ignore.stdout = TRUE)
# Alternative parameters --use_clinkage --alignment_threshold 23 --max_shift
# 13 --max_aln_length 37 --count_threshold 50 --max_inner_gaps 0
# --assign_thresholds 14.1,10.5,7.0
```

```r
hammock.log <- data.table(readLines("data/HammockInVivo/run.log"))

colnames(hammock.log) <- c("Hammock log file")
knitr::kable(hammock.log, longtable = T)
```

Hammock log file
2020-11-10 12:07:17.463:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-10 12:07:17.606: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/invivoSamplesPeptides.fasta -d /home/rstudio/data/HammockInVivo –max_shift 7 -c 250 –alignment_threshold 26 –assign_thresholds 50,40,30 -t 48
Complete list of input/output parameters:
-i, –input /home/rstudio/data/invivoSamplesPeptides.fasta
-d, –output_directory /home/rstudio/data/HammockInVivo
-t, –thread 48
-l, –labels null
Complete list of clinkage clustering parameters:
-f, –file_format fasta
-m, –matrix /home/rstudio/Hammock_v_1.1.1/matrices/blosum62.txt
-g, –alignment_threshold (–greedy_threshold)26
-x, –max_shift 7
-p, –gap_penalty 0
-C, –cache_size_limit 1
2020-11-10 12:07:17.606: Loading input sequences...
2020-11-10 12:07:17.757: 15768 unique sequences loaded.
2020-11-10 12:07:17.772: 39001 total sequences loaded.
2020-11-10 12:07:17.772: 15768 unique sequences after non-specified labels filtered out
2020-11-10 12:07:17.789: 39001 total sequences after non-specified labels fileterd out
2020-11-10 12:07:17.793: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-10 12:07:17.794: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-10 12:07:17.830: Generating input statistics...
2020-11-10 12:07:17.909: Initial greedy clusters limit not set. Setting automatically to: 394
2020-11-10 12:07:17.911: Greedy clustering...
2020-11-10 12:07:29.846: Ready. Clustering time: 11935
2020-11-10 12:07:29.847: Resulting clusers: 12939
2020-11-10 12:07:29.848: Building MSAs...
2020-11-10 12:07:30.280: Ready. Total time: 12369
2020-11-10 12:07:30.281: Saving results to output files...
2020-11-10 12:07:30.906: Greedy clustering results in: /home/rstudio/data/HammockInVivo/initial_clusters.tsv
2020-11-10 12:07:30.906: and: /home/rstudio/data/HammockInVivo/initial_clusters_sequences.tsv
2020-11-10 12:07:30.906: and: /home/rstudio/data/HammockInVivo/initial_clusters_sequences_original_order.tsv
2020-11-10 12:07:30.906:
Loading clusters...
2020-11-10 12:07:31.004: Maximal alignment length not set. Setting automatically to: 31
2020-11-10 12:07:31.013: Minimal number of match states not set. Setting automatically to: 5
2020-11-10 12:07:31.220: Overlap threshold not set. Setting automatically to:
2020-11-10 12:07:31.230: 10.83,6.19,0.0,
2020-11-10 12:07:31.230: Merge threshold not set. Setting automatically based on average sequence length to:

2020-11-10 12:07:31.239: 15.47,13.92,12.38,

Complete list of HMM-based clustering parameters:
-a, –part_threshold null
-s, –size_threshold null
-c, –count_threshold 250
-n, –assign_thresholds 50.0,40.0,30.0,
-v, –overlap_thresholds 10.83,6.19,0.0,
-r, –merge_thresholds 15.47,13.92,12.38,
-e, –relative_thresholds false
-b, –absolute_thresholds true
-h, –min_conserved_positions 5
-y, –max_gap_proportion 0.05
-k, –min_ic 1.2
-j, –max_aln_length 31
-u, –max_inner_gaps 0
-q, –extension_increase_length false

2020-11-10 12:07:31.336:
Clustering in 3 rounds...
2020-11-10 12:07:31.338:
2020-11-10 12:07:31.338: Round 1:

2020-11-10 12:07:31.338: 250 clusters remaining
2020-11-10 12:07:31.339: Building hmms and searching database...
2020-11-10 12:07:33.593: Extending clusters...
2020-11-10 12:07:33.631: 0 sequences to be inserted into clusters
2020-11-10 12:07:33.632: 0 clusters to be extended
2020-11-10 12:07:33.632: 0 sequences rejected
2020-11-10 12:07:33.637: 102 cluster pairs to check and merge.
2020-11-10 12:07:33.637: Merging clusters from 33 groups...
2020-11-10 12:07:33.668: Buiding hhs...
2020-11-10 12:07:33.739: HH clustering...
2020-11-10 12:07:35.528:
2020-11-10 12:07:35.529: Round 2:

2020-11-10 12:07:35.529: 242 clusters remaining
2020-11-10 12:07:35.529: Building hmms and searching database...
2020-11-10 12:07:37.252: Extending clusters...
2020-11-10 12:07:37.262: 0 sequences to be inserted into clusters
2020-11-10 12:07:37.262: 0 clusters to be extended
2020-11-10 12:07:37.262: 0 sequences rejected
2020-11-10 12:07:37.274: 2152 cluster pairs to check and merge.
2020-11-10 12:07:37.274: Merging clusters from 1 groups...
2020-11-10 12:07:37.301: Buiding hhs...
2020-11-10 12:07:37.369: HH clustering...
2020-11-10 12:07:44.828:
2020-11-10 12:07:44.828: Round 3:

2020-11-10 12:07:44.828: 229 clusters remaining
2020-11-10 12:07:44.828: Building hmms and searching database...
2020-11-10 12:07:46.573: Extending clusters...
2020-11-10 12:07:46.582: 7 sequences to be inserted into clusters
2020-11-10 12:07:46.583: 5 clusters to be extended
2020-11-10 12:07:46.594: 0 sequences rejected
2020-11-10 12:07:46.595: Overlap threshold is 0. Running full cluster merging.

Hammock log file

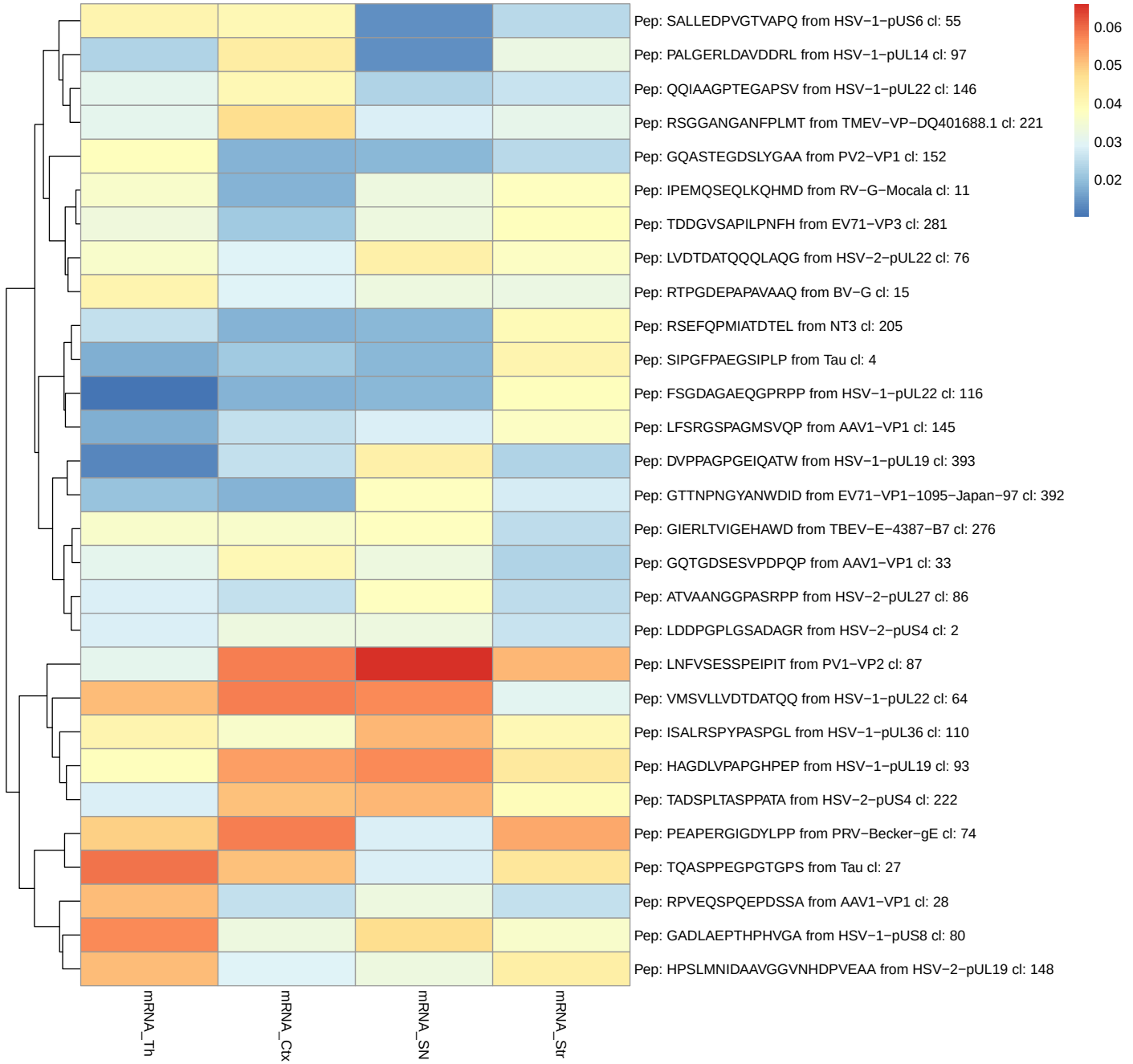
```
2020-11-10 12:07:46.625: Buiding hhs...
2020-11-10 12:07:46.637: HH clustering...
2020-11-10 12:07:59.910:
Ready. Clustering time : 28574
2020-11-10 12:07:59.910: Resulting clusers: 203
2020-11-10 12:07:59.911: Containing 2428 unique sequences and 10180 total sequences.
2020-11-10 12:07:59.920: Unique sequences not assigned: 13340, total sequences not assigned: 28821
2020-11-10 12:07:59.920: Saving results to outupt files...
2020-11-10 12:08:00.106: Results in: /home/rstudio/data/HammockInVivo/final_clusters_sequences.tsv
2020-11-10 12:08:00.106: and: /home/rstudio/data/HammockInVivo/final_clusters.tsv
2020-11-10 12:08:00.107: and: /home/rstudio/data/HammockInVivo/final_clusters_sequences_original_order.tsv
2020-11-10 12:08:00.107:
Calculating KLD...
2020-11-10 12:08:00.418: Final system KLD over match state MSA positions: 18.828779438507368
2020-11-10 12:08:00.418: Final system KLD over all MSA positions: 29.927410491876895
2020-11-10 12:08:00.420: Program successfully ended.
```

Generation of Weblogo visualization

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockInVivo/final_clusters.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))
id.order <- as.list(ham.clusters$cluster_id)
ham.clusters.all <- data.table(read.table("/home/rstudio/data/HammockInVivo/final_clusters_sequences.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))
ham.clusters.all[, `:=`(alignment, gsub("\\-", "\\_", alignment))]
setkey(select.samples, Peptide)
setkey(select.samples.trsp, Peptide)

unlink("/home/rstudio/data/WEBlogosInVivo", recursive = TRUE, force = FALSE)
dir.create(file.path("/home/rstudio/data/", "WEBlogosInVivo"), showWarnings = FALSE)
dir.create(file.path("/home/rstudio/data/HammockInVivo/", "alignments_final_Scored"),
    showWarnings = FALSE)

setkey(ham.clusters.all, cluster_id)
setkey(ham.clusters, cluster_id)
setkey(select.samples.pepMerge, Peptide)


opts_chunk$set(out.width = "100%", fig.align = "center")
generateWeblogo <- function(in.name) {
    # in.name <- ham.clusters$cluster_id[12] in.name <- 67777
    this.fa <- read.fasta(file = paste("/home/rstudio/data/HammockInVivo/alignments_final/",
        in.name, ".aln", sep = ""))
    allSeqs <- unlist(getSequence(this.fa, as.string = TRUE))
    allSeqs <- data.table(unlist(lapply(allSeqs, function(x) gsub("([-])", "",
        toupper(x)))))
    allSeqs.out <- select.samples.pepMerge[J(allSeqs)]
    allSeqs.out$Annot <- data.table(getName(this.fa))
    allSeqs.out[, `:=`(Annot, paste(Annot, "_", Score, sep = ""))]
    allSeqs.out$Alignment <- data.table(toupper(unlist(getSequence(this.fa,
        as.string = TRUE))))

    allSeqs.out <- allSeqs.out[rep(1:.N, Score)][, `:=`(Indx, 1:.N), by = Peptide]
    allSeqs.out[, `:=`(Annot, paste(Annot, "_", Indx, sep = ""))]
```

```
    write.fasta(as.list(allSeqs.out$Alignment), allSeqs.out$Annot, nbchar = 60,
        paste("/home/rstudio/data/HammockInVivo/alignments_final_Scored/", in.name,
            ".aln", sep = ""), open = "w")


    this.main <- ham.clusters[J(in.name)]
    main.gene <- select.samples.trsp[J(this.main$main_sequence)]$GeneName[1]
    this.title <- paste("## Peptide", this.main$main_sequence, "from", main.gene,
        "with cluster number", in.name, sep = " ")

    tmp <- system(paste("weblogo --format PDF --sequence-type protein --size large --errorbars NO --resolutio
        this.title, "' < /home/rstudio/data/HammockInVivo/alignments_final_Scored/",
        in.name, ".aln > /home/rstudio/data/WEBlogosInVivo/", in.name, ".pdf",
        sep = ""), intern = TRUE, ignore.stdout = FALSE)
}

invisible(mclapply(id.order, generateWeblogo, mc.cores = detectCores()))


ham.clusters.merged <- ham.clusters

ham.clusters.merged[, `:=`(mRNA_Str, mRNA_30cpc_Str + mRNA_3cpc_Str + mRNA_30cpc_Str_4wks +
    mRNA_3cpc_Str_4wks)]
ham.clusters.merged[, `:=`(mRNA_SN, mRNA_30cpc_SN + mRNA_3cpc_SN + mRNA_30cpc_SN_4wks +
    mRNA_3cpc_SN_4wks)]
ham.clusters.merged[, `:=`(mRNA_Th, mRNA_30cpc_Th + mRNA_3cpc_Th + mRNA_30cpc_Th_4wks +
    mRNA_3cpc_Th_4wks)]
ham.clusters.merged[, `:=`(mRNA_Ctx, mRNA_30cpc_Ctx + mRNA_3cpc_Ctx + mRNA_30cpc_Ctx_4wks +
    mRNA_3cpc_Ctx_4wks)]
ham.clusters.merged[, `:=`(c("mRNA_30cpc_Str", "mRNA_30cpc_SN", "mRNA_30cpc_Th",
    "mRNA_30cpc_Ctx", "mRNA_3cpc_Str", "mRNA_3cpc_SN", "mRNA_3cpc_Th", "mRNA_3cpc_Ctx",
    "mRNA_30cpc_Str_4wks", "mRNA_30cpc_SN_4wks", "mRNA_30cpc_Th_4wks", "mRNA_30cpc_Ctx_4wks",
    "mRNA_3cpc_Str_4wks", "mRNA_3cpc_SN_4wks", "mRNA_3cpc_Th_4wks", "mRNA_3cpc_Ctx_4wks"),
    NULL)]


ham.clusters.merged.melt <- melt(ham.clusters.merged, id = c("cluster_id", "main_sequence",
    "sum"))
setkeyv(ham.clusters.merged.melt, "variable")
ham.clusters.topTen <- setorder(setDT(ham.clusters.merged.melt), -value)[, head(.SD,
    14), keyby = variable]
# ham.clusters.topTen <- ham.clusters.merged.melt[, head(.SD, 15),
# by=variable]
ham.clusters.select <- ham.clusters.merged.melt[ham.clusters.merged.melt$cluster_id %in%
    unique(ham.clusters.topTen$cluster_id)]

ham.clusters.select[, `:=`(geneName, lapply(main_sequence, function(x) select.samples.trsp[J(x)]$GeneName[1]))
ham.clusters.select[, `:=`(listName, paste("Pep:", main_sequence, "from", geneName,
    "cl:", cluster_id, sep = " "))]

select.samples.out <- acast(ham.clusters.select, listName ~ variable, value.var = "value")  #Utilizes reshape
select.samples.out[is.na(select.samples.out)] <- 0
select.samples.out <- select.samples.out[, c(3, 4, 2, 1)]
select.samples.out.scaled <- scale(select.samples.out, center = FALSE, scale = colSums(select.samples.out))
# select.samples.out.scaled <-
# select.samples.out.scaled[order(round(select.samples.out.scaled[,1],digits
# = 2),round(select.samples.out.scaled[,2],digits =
# 2),round(select.samples.out.scaled[,3],digits =
```

```
# 2),round(select.samples.out.scaled[,4],digits = 2),decreasing=TRUE),]
pheatmap(select.samples.out.scaled, cluster_rows = TRUE, show_rownames = TRUE,
    cluster_cols = FALSE)
```



Generation of heatmaps for in vitro samples

```
select.samples <- all.samples[J(c("mRNA_3cpc_HEK293T", "mRNA_30cpc_HEK293T",
    "mRNA_3cpc_pNeuron", "mRNA_30cpc_pNeuron"))]

select.samples[, `:=`(BCcount, as.integer(mclapply(BC, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
select.samples[, `:=`(Animalcount, as.integer(mclapply(Animals, function(x) length(table(strsplit(paste(t(x),
```

```
    collapse = ","), ","))), mc.cores = detectCores()))]
select.samples[, `:=`(Score, BCcount + Animalcount - 1), ]
select.samples.trsp <- unique(select.samples, by = c("Animals", "BC", "LUTnrs"))

fasta.names <- paste(1:nrow(select.samples.trsp), select.samples.trsp$Score,
    select.samples.trsp$Group, sep = "|")
write.fasta(as.list(select.samples.trsp$Peptide), fasta.names, "data/invitroSamplesPeptides.fasta",
    open = "w", nbchar = 60, as.string = TRUE)

# Generate Scoring table for Weblogo Weighting
select.samples.pepMerge <- select.samples.trsp[, sum(Score), by = c("Peptide")]
setnames(select.samples.pepMerge, "V1", "Score")
```

Executing Hammock Clustering

```
Sys.setenv(PATH = paste("/root/HMMER/binaries", Sys.getenv("PATH"), sep = ":"),
    HHLIB = "/home/rstudio/Hammock_v_1.1.1/hhsuite-2.0.16/lib/hh/")
unlink("/home/rstudio/data/HammockInVitro", recursive = TRUE, force = FALSE)
sys.out <- system(paste("java -jar /home/rstudio/Hammock_v_1.1.1/dist/Hammock.jar full -i /home/rstudio/data/
    detectCores(), sep = ""), intern = TRUE, ignore.stdout = TRUE)
# Alternative parameters --use_clinkage --alignment_threshold 23 --max_shift
# 13 --max_aln_length 37 --count_threshold 50 --max_inner_gaps 0
# --assign_thresholds 14.1,10.5,7.0
hammock.log <- data.table(readLines("data/HammockInVitro/run.log"))

colnames(hammock.log) <- c("Hammock log file")
knitr::kable(hammock.log, longtable = T)
```

Hammock log file
2020-11-10 12:08:13.970:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-10 12:08:14.107: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/invitroSamplesPeptides.fasta -d /home/rstudio/data/HammockInVitro –max_shift 7 -c 50 -t 48
Complete list of input/output parameters:
-i, –input /home/rstudio/data/invitroSamplesPeptides.fasta
-d, –output_directory /home/rstudio/data/HammockInVitro
-t, –thread 48
-l, –labels null
Complete list of clinkage clustering parameters:
-f, –file_format fasta
-m, –matrix /home/rstudio/Hammock_v_1.1.1/matrices/blosum62.txt
-g, –alignment_threshold (–greedy_threshold)null
-x, –max_shift 7
-p, –gap_penalty 0
-C, –cache_size_limit 1

Hammock log file

2020-11-10 12:08:14.108: Loading input sequences...
2020-11-10 12:08:14.124: 462 unique sequences loaded.
2020-11-10 12:08:14.126: 518 total sequences loaded.
2020-11-10 12:08:14.126: 462 unique sequences after non-specified labels filtered out
2020-11-10 12:08:14.129: 518 total sequences after non-specified labels fileterd out
2020-11-10 12:08:14.129: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-10 12:08:14.129: Up to 10 000 unique sequences. Using clinkage clustering. Use –use_greedy to force greedy clustering
2020-11-10 12:08:14.133: Generating input statistics...
2020-11-10 12:08:14.135: Clinkage clustering threshold not set. Setting automatically to: 26
2020-11-10 12:08:14.137: Clinkage clustering...
2020-11-10 12:08:14.748: Ready. Clustering time: 611
2020-11-10 12:08:14.749: Resulting clusers: 348
2020-11-10 12:08:14.749: Building MSAs...
2020-11-10 12:08:14.886: Ready. Total time: 749
2020-11-10 12:08:14.887: Saving results to output files...
2020-11-10 12:08:14.941: Clinkage clustering results in: /home/rstudio/data/HammockInVitro/initial_clusters.tsv
2020-11-10 12:08:14.941: and: /home/rstudio/data/HammockInVitro/initial_clusters_sequences.tsv
2020-11-10 12:08:14.941: and: /home/rstudio/data/HammockInVitro/initial_clusters_sequences_original_order.tsv
2020-11-10 12:08:14.941:
Loading clusters...
2020-11-10 12:08:14.954: Maximal alignment length not set. Setting automatically to: 30
2020-11-10 12:08:14.955: Minimal number of match states not set. Setting automatically to: 5
2020-11-10 12:08:14.982: Assign threshold sequence not set. Setting automatically to:
2020-11-10 12:08:14.986: 14.35,11.33,8.31,
2020-11-10 12:08:14.986: Overlap threshold not set. Setting automatically to:
2020-11-10 12:08:14.986: 10.58,6.04,0.0,
2020-11-10 12:08:14.987: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-10 12:08:14.987: 15.11,13.6,12.09,

Complete list of HMM-based clustering parameters:
-a, –part_threshold null
-s, –size_threshold null
-c, –count_threshold 50
-n, –assign_thresholds 14.35,11.33,8.31,
-v, –overlap_thresholds 10.58,6.04,0.0,
-r, –merge_thresholds 15.11,13.6,12.09,
-e, –relative_thresholds false
-b, –absolute_thresholds true
-h, –min_conserved_positions 5
-y, –max_gap_proportion 0.05
-k, –min_ic 1.2
-j, –max_aln_length 30
-u, –max_inner_gaps 0
-q, –extension_increase_length false

2020-11-10 12:08:15.012:
Clustering in 3 rounds...
2020-11-10 12:08:15.014:
2020-11-10 12:08:15.014: Round 1:

2020-11-10 12:08:15.014: 50 clusters remaining
2020-11-10 12:08:15.014: Building hmms and searching database...
2020-11-10 12:08:15.446: Extending clusters...
2020-11-10 12:08:15.447: 0 sequences to be inserted into clusters
2020-11-10 12:08:15.448: 0 clusters to be extended

Hammock log file

2020-11-10 12:08:15.449: 0 sequences rejected
2020-11-10 12:08:15.449: 0 cluster pairs to check and merge.
2020-11-10 12:08:15.450: Merging clusters from 0 groups...
2020-11-10 12:08:15.461: Buiding hhs...
2020-11-10 12:08:15.462: HH clustering...
2020-11-10 12:08:15.471:
2020-11-10 12:08:15.471: Round 2:

2020-11-10 12:08:15.471: 50 clusters remaining
2020-11-10 12:08:15.471: Building hmms and searching database...
2020-11-10 12:08:15.835: Extending clusters...
2020-11-10 12:08:15.836: 3 sequences to be inserted into clusters
2020-11-10 12:08:15.836: 3 clusters to be extended
2020-11-10 12:08:15.843: 3 sequences rejected
2020-11-10 12:08:15.844: 9 cluster pairs to check and merge.
2020-11-10 12:08:15.844: Merging clusters from 5 groups...
2020-11-10 12:08:15.852: Buiding hhs...
2020-11-10 12:08:15.871: HH clustering...
2020-11-10 12:08:16.151:
2020-11-10 12:08:16.151: Round 3:

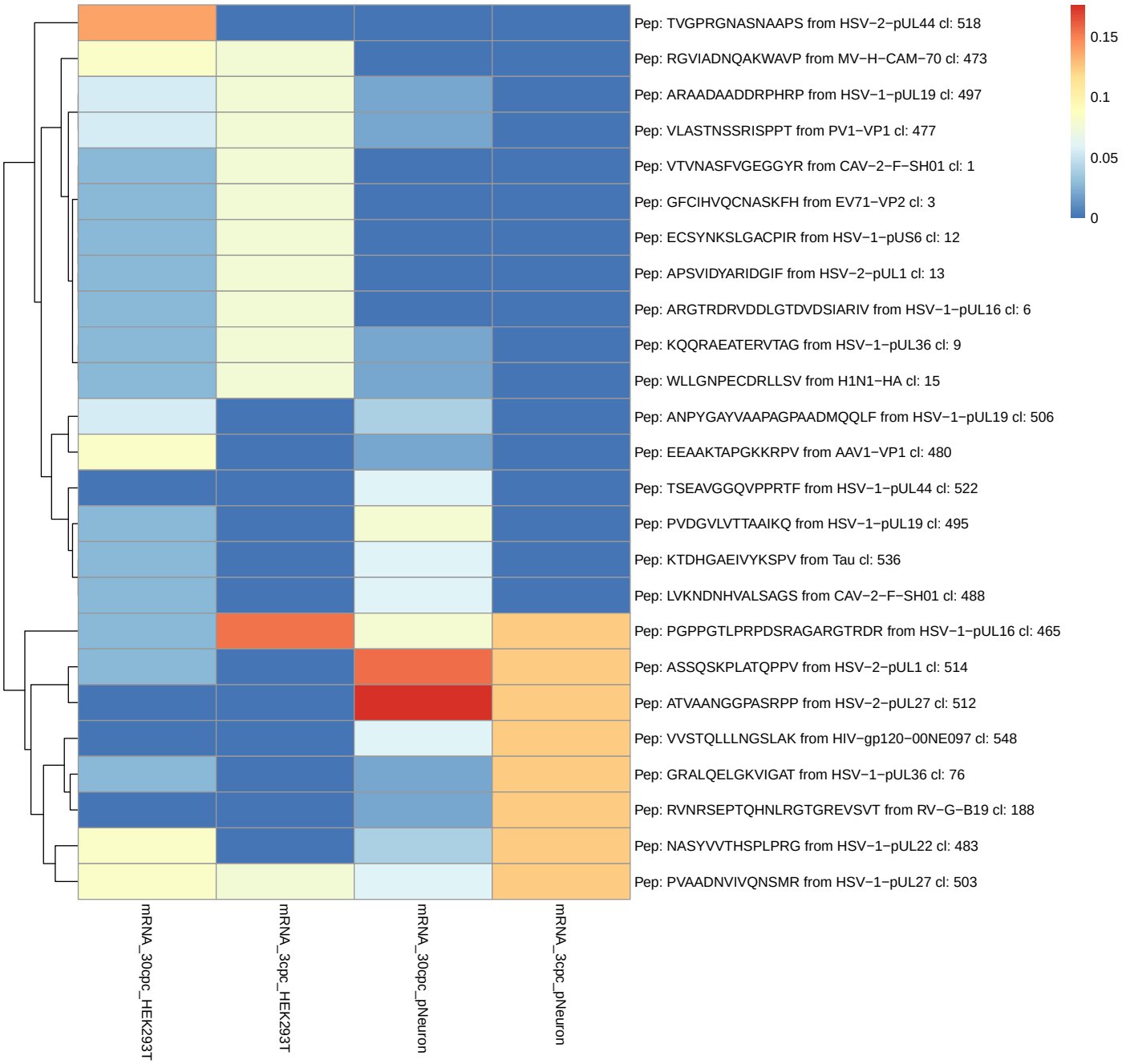
2020-11-10 12:08:16.151: 50 clusters remaining
2020-11-10 12:08:16.152: Building hmms and searching database...
2020-11-10 12:08:16.522: Extending clusters...
2020-11-10 12:08:16.523: 19 sequences to be inserted into clusters
2020-11-10 12:08:16.524: 15 clusters to be extended
2020-11-10 12:08:16.533: 11 sequences rejected
2020-11-10 12:08:16.534: Overlap threshold is 0. Running full cluster merging.
2020-11-10 12:08:16.540: Buiding hhs...
2020-11-10 12:08:16.564: HH clustering...
2020-11-10 12:08:19.465:
Ready. Clustering time : 4453
2020-11-10 12:08:19.466: Resulting clusers: 44
2020-11-10 12:08:19.466: Containing 108 unique sequences and 161 total sequences.
2020-11-10 12:08:19.467: Unique sequences not assigned: 354, total sequences not assigned: 357
2020-11-10 12:08:19.467: Saving results to outupt files...
2020-11-10 12:08:19.485: Results in: /home/rstudio/data/HammockInVitro/final_clusters_sequences.tsv
2020-11-10 12:08:19.485: and: /home/rstudio/data/HammockInVitro/final_clusters.tsv
2020-11-10 12:08:19.485: and: /home/rstudio/data/HammockInVitro/final_clusters_sequences_original_order.tsv
2020-11-10 12:08:19.485:
Calculating KLD...
2020-11-10 12:08:19.486: 13 clusters omitted from KLD calculation because each of them only contains a single unique sequence.
2020-11-10 12:08:19.524: Final system KLD over match state MSA positions: 7.501060786231158
2020-11-10 12:08:19.524: Final system KLD over all MSA positions: 7.227012879783852
2020-11-10 12:08:19.524: Program successfully ended.

Generation of Weblogo visualization

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockInVitro/final_clusters.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))
id.order <- as.list(ham.clusters$cluster_id)
ham.clusters.all <- data.table(read.table("/home/rstudio/data/HammockInVitro/final_clusters_sequences.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))
```

```
ham.clusters.all[, `:=`(alignment, gsub("\\-", "\\_", alignment))]
setkey(select.samples, Peptide)
setkey(select.samples.trsp, Peptide)

unlink("/home/rstudio/data/WEBlogosInVitro", recursive = TRUE, force = FALSE)
dir.create(file.path("/home/rstudio/data/", "WEBlogosInVitro"), showWarnings = FALSE)
dir.create(file.path("/home/rstudio/data/HammockInVitro/", "alignments_final_Scored"),
    showWarnings = FALSE)

setkey(ham.clusters.all, cluster_id)
setkey(ham.clusters, cluster_id)
setkey(select.samples.pepMerge, Peptide)


opts_chunk$set(out.width = "100%", fig.align = "center")
generateWeblogo <- function(in.name) {
    # in.name <- ham.clusters$cluster_id[12] in.name <- 6777
    this.fa <- read.fasta(file = paste("/home/rstudio/data/HammockInVitro/alignments_final/",
        in.name, ".aln", sep = ""))
    allSeqs <- unlist(getSequence(this.fa, as.string = TRUE))
    allSeqs <- data.table(unlist(lapply(allSeqs, function(x) gsub("([-])", "",
        toupper(x)))))
    allSeqs.out <- select.samples.pepMerge[J(allSeqs)]
    allSeqs.out$Annot <- data.table(getName(this.fa))
    allSeqs.out[, `:=`(Annot, paste(Annot, "_", Score, sep = ""))]
    allSeqs.out$Alignment <- data.table(toupper(unlist(getSequence(this.fa,
        as.string = TRUE))))

    allSeqs.out <- allSeqs.out[rep(1:.N, Score)][, `:=`(Indx, 1:.N), by = Peptide]
    allSeqs.out[, `:=`(Annot, paste(Annot, "_", Indx, sep = ""))]

    write.fasta(as.list(allSeqs.out$Alignment), allSeqs.out$Annot, nbchar = 60,
        paste("/home/rstudio/data/HammockInVitro/alignments_final_Scored/",
            in.name, ".aln", sep = ""), open = "w")


    this.main <- ham.clusters[J(in.name)]
    main.gene <- select.samples.trsp[J(this.main$main_sequence)]$GeneName[1]
    this.title <- paste("## Peptide", this.main$main_sequence, "from", main.gene,
        "with cluster number", in.name, sep = " ")

    tmp <- system(paste("weblogo --format PDF --sequence-type protein --size large --errorbars NO --resolutio
        this.title, "' < /home/rstudio/data/HammockInVitro/alignments_final_Scored/",
        in.name, ".aln > /home/rstudio/data/WEBlogosInVitro/", in.name, ".pdf",
        sep = ""), intern = TRUE, ignore.stdout = FALSE)
}

invisible(mclapply(id.order, generateWeblogo, mc.cores = detectCores()))


ham.clusters.merged <- ham.clusters

# ham.clusters.merged[,mRNA_Str := mRNA_30cpc_Str + mRNA_3cpc_Str +
# mRNA_30cpc_Str_4wks + mRNA_3cpc_Str_4wks] ham.clusters.merged[,mRNA_SN :=
# mRNA_30cpc_SN + mRNA_3cpc_SN + mRNA_30cpc_SN_4wks + mRNA_3cpc_SN_4wks]
# ham.clusters.merged[,mRNA_Th := mRNA_30cpc_Th + mRNA_3cpc_Th +
# mRNA_30cpc_Th_4wks + mRNA_3cpc_Th_4wks] ham.clusters.merged[,mRNA_Ctx :=
# mRNA_30cpc_Ctx + mRNA_3cpc_Ctx + mRNA_30cpc_Ctx_4wks + mRNA_3cpc_Ctx_4wks]
```

```
# ham.clusters.merged[,c('mRNA_30cpc_Str',
# 'mRNA_30cpc_SN','mRNA_30cpc_Th','mRNA_30cpc_Ctx','mRNA_3cpc_Str',
# 'mRNA_3cpc_SN','mRNA_3cpc_Th','mRNA_3cpc_Ctx','mRNA_30cpc_Str_4wks',
# 'mRNA_30cpc_SN_4wks','mRNA_30cpc_Th_4wks','mRNA_30cpc_Ctx_4wks','mRNA_3cpc_Str_4wks','mRNA_3cpc_SN_4wks','m
# := NULL]

library(reshape)
ham.clusters.merged.melt <- melt(ham.clusters.merged, id = c("cluster_id", "main_sequence",
    "sum"))
setkeyv(ham.clusters.merged.melt, "variable")
ham.clusters.topTen <- setorder(setDT(ham.clusters.merged.melt), -value)[, head(.SD,
    8), keyby = variable]
# ham.clusters.topTen <- ham.clusters.merged.melt[, head(.SD, 15),
# by=variable]
ham.clusters.select <- ham.clusters.merged.melt[ham.clusters.merged.melt$cluster_id %in%
    unique(ham.clusters.topTen$cluster_id)]

ham.clusters.select[, `:=`(geneName, lapply(main_sequence, function(x) select.samples.trsp[J(x)]$GeneName[1]))
ham.clusters.select[, `:=`(listName, paste("Pep:", main_sequence, "from", geneName,
    "cl:", cluster_id, sep = " "))]

select.samples.out <- acast(ham.clusters.select, listName ~ variable, value.var = "value")  #Utilizes reshape
select.samples.out[is.na(select.samples.out)] <- 0
select.samples.out <- select.samples.out[, c(2, 3, 1, 4)]
select.samples.out.scaled <- scale(select.samples.out, center = FALSE, scale = colSums(select.samples.out))
# select.samples.out.scaled <-
# select.samples.out.scaled[order(round(select.samples.out.scaled[,1],digits
# = 2),round(select.samples.out.scaled[,2],digits =
# 2),round(select.samples.out.scaled[,3],digits =
# 2),round(select.samples.out.scaled[,4],digits = 2),decreasing=TRUE),]
pheatmap(select.samples.out.scaled, cluster_rows = TRUE, show_rownames = TRUE,
    cluster_cols = FALSE)
```




```
select.samples <- all.samples[J(c("DNA_pscAAVlib", "DNA_pscAAVlib_Prep2", "DNA_AAVlib_DNAse_3cpc",
    "DNA_AAVlib_DNAse_30cpc"))]

select.samples[, `:=`(BCcount, as.integer(mclapply(BC, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
select.samples[, `:=`(Score, BCcount)]
select.samples.trsp <- unique(select.samples, by = c("Animals", "BC", "LUTnrs"))
```

Clustering DNAse resistant virions

```
select.samples <- all.samples[J(c("DNA_AAVlib_DNAse_3cpc", "DNA_AAVlib_DNAse_30cpc"))]

select.samples[, `:=`(BCcount, as.integer(mclapply(BC, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
```

```
select.samples[, `:=`(Score, BCcount)]
select.samples.trsp <- unique(select.samples, by = c("Animals", "BC", "LUTnrs"))

fasta.names <- paste(1:nrow(select.samples.trsp), select.samples.trsp$Score,
    select.samples.trsp$Group, sep = "|")
write.fasta(as.list(select.samples.trsp$Peptide), fasta.names, "data/DNAsePeptides.fasta",
    open = "w", nbchar = 60, as.string = TRUE)

# Generate Scoring table for Weblogo Weighting
select.samples.pepMerge <- select.samples.trsp[, sum(Score), by = c("Peptide")]
setnames(select.samples.pepMerge, "V1", "Score")
```

Executing Hammock Clustering

```
Sys.setenv(PATH = paste("/root/HMMER/binaries", Sys.getenv("PATH"), sep = ":"),
    HHLIB = "/home/rstudio/Hammock_v_1.1.1/hhsuite-2.0.16/lib/hh/")
unlink("/home/rstudio/data/HammockDNAse", recursive = TRUE, force = FALSE)
sys.out <- system(paste("java -jar /home/rstudio/Hammock_v_1.1.1/dist/Hammock.jar full -i /home/rstudio/data/
    detectCores(), sep = ""), intern = TRUE, ignore.stdout = TRUE)
# Alternative parameters --use_clinkage --alignment_threshold 23
# --alignment_threshold 26 --assign_thresholds 50,40,30
hammock.log <- data.table(readLines("data/HammockDNAse/run.log"))

colnames(hammock.log) <- c("Hammock log file")
knitr::kable(hammock.log, longtable = T)
```

Hammock log file
2020-11-10 12:08:58.796:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-10 12:08:58.933: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/DNAsePeptides.fasta -d /home/rstudio/data/HammockDNAse –max_shift 7 -c 2000 -t 48
Complete list of input/output parameters:
-i, –input /home/rstudio/data/DNAsePeptides.fasta
-d, –output_directory /home/rstudio/data/HammockDNAse
-t, –thread 48
-l, –labels null
Complete list of clinkage clustering parameters:
-f, –file_format fasta
-m, –matrix /home/rstudio/Hammock_v_1.1.1/matrices/blosum62.txt
-g, –alignment_threshold (–greedy_threshold)null
-x, –max_shift 7
-p, –gap_penalty 0
-C, –cache_size_limit 1
2020-11-10 12:08:58.934: Loading input sequences...
2020-11-10 12:08:59.206: 49840 unique sequences loaded.
2020-11-10 12:08:59.235: 203706 total sequences loaded.

Hammock log file

2020-11-10 12:08:59.236: 49840 unique sequences after non-specified labels filtered out
2020-11-10 12:08:59.276: 203706 total sequences after non-specified labels fileterd out
2020-11-10 12:08:59.289: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-10 12:08:59.289: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-10 12:08:59.354: Generating input statistics...
2020-11-10 12:08:59.479: Greedy clustering threshold not set. Setting automatically to: 27
2020-11-10 12:08:59.479: Initial greedy clusters limit not set. Setting automatically to: 1246
2020-11-10 12:08:59.481: Greedy clustering...
2020-11-10 12:10:13.784: Ready. Clustering time: 74303
2020-11-10 12:10:13.785: Resulting clusers: 35413
2020-11-10 12:10:13.785: Building MSAs...
2020-11-10 12:10:15.098: Ready. Total time: 75617
2020-11-10 12:10:15.098: Saving results to output files...
2020-11-10 12:10:16.325: Greedy clustering results in: /home/rstudio/data/HammockDNAse/initial_clusters.tsv
2020-11-10 12:10:16.326: and: /home/rstudio/data/HammockDNAse/initial_clusters_sequences.tsv
2020-11-10 12:10:16.326: and: /home/rstudio/data/HammockDNAse/initial_clusters_sequences_original_order.tsv
2020-11-10 12:10:16.326:
Loading clusters...
2020-11-10 12:10:16.544: Maximal alignment length not set. Setting automatically to: 32
2020-11-10 12:10:16.555: Minimal number of match states not set. Setting automatically to: 5
2020-11-10 12:10:17.031: Assign threshold sequence not set. Setting automatically to:
2020-11-10 12:10:17.036: 15.14,11.95,8.77,
2020-11-10 12:10:17.036: Overlap threshold not set. Setting automatically to:
2020-11-10 12:10:17.039: 11.16,6.38,0.0,
2020-11-10 12:10:17.039: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-10 12:10:17.042: 15.94,14.34,12.75,
2020-11-10 12:10:17.213: 6 clusters rejected because of match states and information content constraints.

Complete list of HMM-based clustering parameters:
-a, –part_threshold null
-s, –size_threshold null
-c, –count_threshold 2000
-n, –assign_thresholds 15.14,11.95,8.77,
-v, –overlap_thresholds 11.16,6.38,0.0,
-r, –merge_thresholds 15.94,14.34,12.75,
-e, –relative_thresholds false
-b, –absolute_thresholds true
-h, –min_conserved_positions 5
-y, –max_gap_proportion 0.05
-k, –min_ic 1.2
-j, –max_aln_length 32
-u, –max_inner_gaps 0
-q, –extension_increase_length false

2020-11-10 12:10:17.478:
Clustering in 3 rounds...
2020-11-10 12:10:17.480:
2020-11-10 12:10:17.481: Round 1:

2020-11-10 12:10:17.481: 1994 clusters remaining
2020-11-10 12:10:17.481: Building hmms and searching database...
2020-11-10 12:10:40.669: Extending clusters...
2020-11-10 12:10:40.894: 13792 sequences to be inserted into clusters
2020-11-10 12:10:40.907: 1553 clusters to be extended
2020-11-10 12:10:46.107: 10675 sequences rejected

Hammock log file

2020-11-10 12:10:46.187: 5089 cluster pairs to check and merge.
2020-11-10 12:10:46.187: Merging clusters from 98 groups...
2020-11-10 12:10:46.416: Buiding hhs...
2020-11-10 12:10:47.264: HH clustering...
2020-11-10 12:14:21.355:
2020-11-10 12:14:21.355: Round 2:

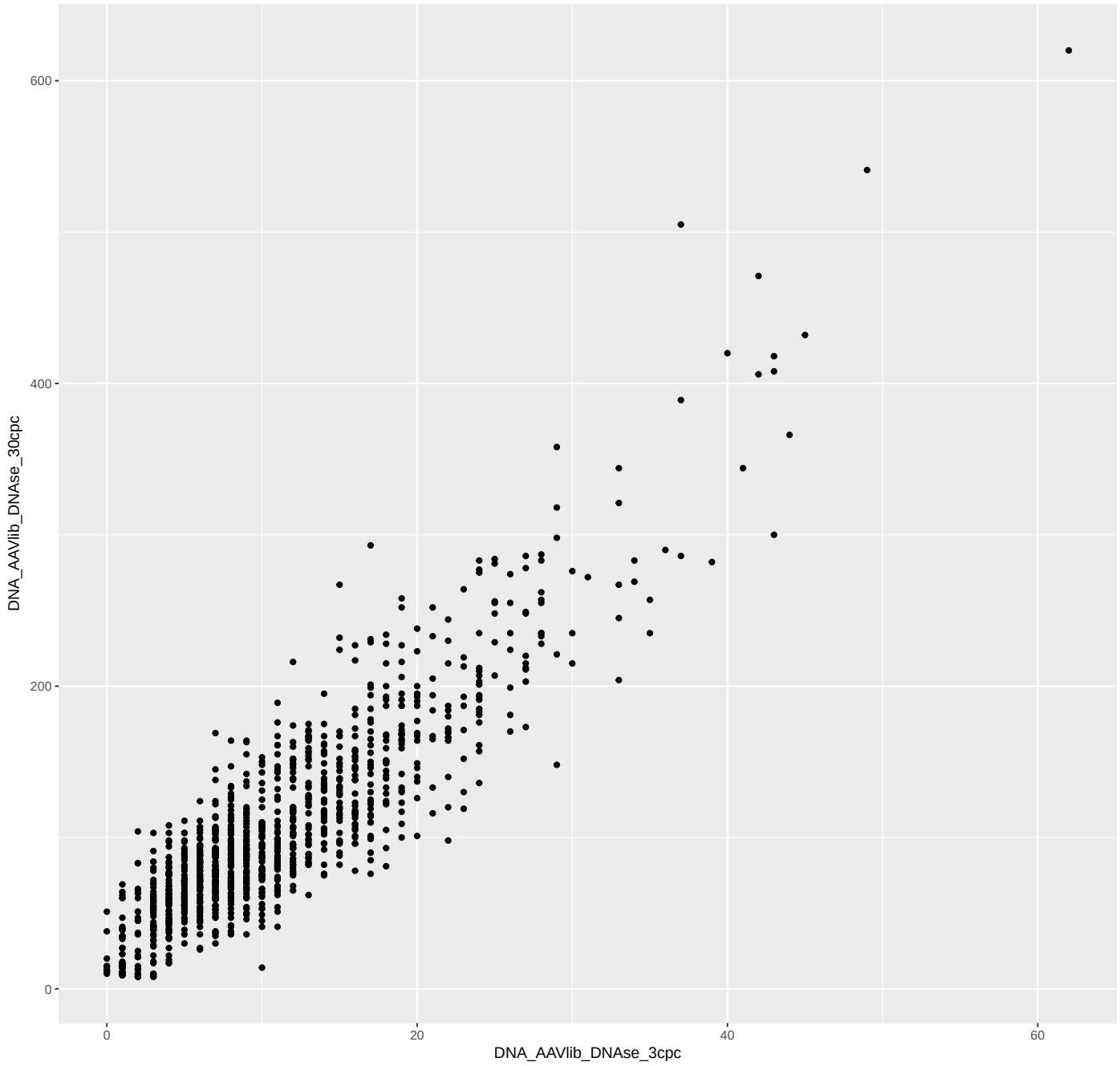
2020-11-10 12:14:21.356: 1530 clusters remaining
2020-11-10 12:14:21.356: Building hmms and searching database...
2020-11-10 12:14:35.489: Extending clusters...
2020-11-10 12:14:35.618: 11386 sequences to be inserted into clusters
2020-11-10 12:14:35.623: 1184 clusters to be extended
2020-11-10 12:14:42.098: 9109 sequences rejected
2020-11-10 12:14:42.466: 51972 cluster pairs to check and merge.
2020-11-10 12:14:42.467: Merging clusters from 1 groups...
2020-11-10 12:14:42.657: Buiding hhs...
2020-11-10 12:14:43.193: HH clustering...
2020-11-10 12:16:20.799:
2020-11-10 12:16:20.800: Round 3:

2020-11-10 12:16:20.800: 1374 clusters remaining
2020-11-10 12:16:20.800: Building hmms and searching database...
2020-11-10 12:16:34.704: Extending clusters...
2020-11-10 12:16:34.791: 14554 sequences to be inserted into clusters
2020-11-10 12:16:34.796: 1204 clusters to be extended
2020-11-10 12:16:47.474: 10703 sequences rejected
2020-11-10 12:16:47.483: Overlap threshold is 0. Running full cluster merging.
2020-11-10 12:16:47.640: Buiding hhs...
2020-11-10 12:16:48.836: HH clustering...
2020-11-10 12:19:10.315:
Ready. Clustering time : 532835
2020-11-10 12:19:10.315: Resulting clusers: 1127
2020-11-10 12:19:10.316: Containing 25589 unique sequences and 134333 total sequences.
2020-11-10 12:19:10.343: Unique sequences not assigned: 24251, total sequences not assigned: 69373
2020-11-10 12:19:10.343: Saving results to outupt files...
2020-11-10 12:19:11.188: Results in: /home/rstudio/data/HammockDNAse/final_clusters_sequences.tsv
2020-11-10 12:19:11.188: and: /home/rstudio/data/HammockDNAse/final_clusters.tsv
2020-11-10 12:19:11.188: and: /home/rstudio/data/HammockDNAse/final_clusters_sequences_original_order.tsv
2020-11-10 12:19:11.188:
Calculating KLD...
2020-11-10 12:19:11.190: 21 clusters omitted from KLD calculation because each of them only contains a single unique sequence.
2020-11-10 12:19:13.442: Final system KLD over match state MSA positions: 20.23129789753592
2020-11-10 12:19:13.443: Final system KLD over all MSA positions: 36.0292448781723
2020-11-10 12:19:13.443: Program successfully ended.

Generation of Scatter plot generation

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockDNAse/final_clusters.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))



pred.points <- ggplot(data = ham.clusters, aes(x = DNA_AAVlib_DNAse_3cpc, y = DNA_AAVlib_DNAse_30cpc)) +
```

```
    labs(x = "DNA_AAVlib_DNAse_3cpc", y = "DNA_AAVlib_DNAse_30cpc") + geom_point()
print(pred.points)
```



Clustering DNAse resistant virions with library

```
select.samples <- all.samples[J(c("DNA_pscAAVlib", "DNA_pscAAVlib_Prep2"))]

select.samples[, `:=`(BCcount, as.integer(mclapply(BC, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
select.samples[, `:=`(Score, BCcount)]
select.samples.trsp <- unique(select.samples, by = c("Animals", "BC", "LUTnrs"))
```

```
fasta.names <- paste(1:nrow(select.samples.trsp), select.samples.trsp$Score,
    select.samples.trsp$Group, sep = "|")
write.fasta(as.list(select.samples.trsp$Peptide), fasta.names, "data/LibDNAsePeptides.fasta",
    open = "w", nbchar = 60, as.string = TRUE)

# Generate Scoring table for Weblogo Weighting
select.samples.pepMerge <- select.samples.trsp[, sum(Score), by = c("Peptide")]
setnames(select.samples.pepMerge, "V1", "Score")
```

Executing Hammock Clustering

```
Sys.setenv(PATH = paste("/root/HMMER/binaries", Sys.getenv("PATH"), sep = ":"),
    HHLIB = "/home/rstudio/Hammock_v_1.1.1/hhsuite-2.0.16/lib/hh/")
unlink("/home/rstudio/data/HammockLibDNAse", recursive = TRUE, force = FALSE)
sys.out <- system(paste("java -jar /home/rstudio/Hammock_v_1.1.1/dist/Hammock.jar full -i /home/rstudio/data/
    detectCores(), sep = ""), intern = TRUE, ignore.stdout = TRUE)
# Alternative parameters --use_clinkage --alignment_threshold 23
# --alignment_threshold 26 --assign_thresholds 50,40,30
hammock.log <- data.table(readLines("data/HammockLibDNAse/run.log"))

colnames(hammock.log) <- c("Hammock log file")
knitr::kable(hammock.log, longtable = T)
```

Hammock log file
2020-11-10 12:19:42.220:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-10 12:19:42.358: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/LibDNAsePeptides.fasta -d /home/rstudio/data/HammockLibDNAse –max_shift 7 -c 2000 -t 48
Complete list of input/output parameters:
-i, –input /home/rstudio/data/LibDNAsePeptides.fasta
-d, –output_directory /home/rstudio/data/HammockLibDNAse
-t, –thread 48
-l, –labels null
Complete list of clinkage clustering parameters:
-f, –file_format fasta
-m, –matrix /home/rstudio/Hammock_v_1.1.1/matrices/blosum62.txt
-g, –alignment_threshold (–greedy_threshold)null
-x, –max_shift 7
-p, –gap_penalty 0
-C, –cache_size_limit 1
2020-11-10 12:19:42.359: Loading input sequences...
2020-11-10 12:19:42.771: 60179 unique sequences loaded.
2020-11-10 12:19:42.804: 2906509 total sequences loaded.
2020-11-10 12:19:42.805: 60179 unique sequences after non-specified labels filtered out
2020-11-10 12:19:42.857: 2906509 total sequences after non-specified labels fileterd out

Hammock log file

2020-11-10 12:19:42.872: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-10 12:19:42.873: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-10 12:19:42.955: Generating input statistics...
2020-11-10 12:19:43.117: Greedy clustering threshold not set. Setting automatically to: 27
2020-11-10 12:19:43.117: Initial greedy clusters limit not set. Setting automatically to: 1504
2020-11-10 12:19:43.119: Greedy clustering...
2020-11-10 12:21:17.302: Ready. Clustering time: 94183
2020-11-10 12:21:17.303: Resulting clusers: 40062
2020-11-10 12:21:17.303: Building MSAs...
2020-11-10 12:21:18.919: Ready. Total time: 95800
2020-11-10 12:21:18.920: Saving results to output files...
2020-11-10 12:21:20.311: Greedy clustering results in: /home/rstudio/data/HammockLibDNAse/initial_clusters.tsv
2020-11-10 12:21:20.311: and: /home/rstudio/data/HammockLibDNAse/initial_clusters_sequences.tsv
2020-11-10 12:21:20.312: and:
/home/rstudio/data/HammockLibDNAse/initial_clusters_sequences_original_order.tsv
2020-11-10 12:21:20.312:
Loading clusters...
2020-11-10 12:21:20.549: Maximal alignment length not set. Setting automatically to: 32
2020-11-10 12:21:20.560: Minimal number of match states not set. Setting automatically to: 5
2020-11-10 12:21:20.973: Assign threshold sequence not set. Setting automatically to:
2020-11-10 12:21:20.977: 15.3,12.08,8.86,
2020-11-10 12:21:20.978: Overlap threshold not set. Setting automatically to:
2020-11-10 12:21:20.982: 11.28,6.44,0.0,
2020-11-10 12:21:20.982: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-10 12:21:20.986: 16.11,14.5,12.89,
2020-11-10 12:21:21.179: 3 clusters rejected because of match states and information content constraints.

Complete list of HMM-based clustering parameters:
-a, –part_threshold null
-s, –size_threshold null
-c, –count_threshold 2000
-n, –assign_thresholds 15.3,12.08,8.86,
-v, –overlap_thresholds 11.28,6.44,0.0,
-r, –merge_thresholds 16.11,14.5,12.89,
-e, –relative_thresholds false
-b, –absolute_thresholds true
-h, –min_conserved_positions 5
-y, –max_gap_proportion 0.05
-k, –min_ic 1.2
-j, –max_aln_length 32
-u, –max_inner_gaps 0
-q, –extension_increase_length false

2020-11-10 12:21:21.432:
Clustering in 3 rounds...
2020-11-10 12:21:21.435:
2020-11-10 12:21:21.435: Round 1:

2020-11-10 12:21:21.435: 1997 clusters remaining
2020-11-10 12:21:21.435: Building hmms and searching database...
2020-11-10 12:21:45.015: Extending clusters...
2020-11-10 12:21:45.227: 15752 sequences to be inserted into clusters
2020-11-10 12:21:45.240: 1560 clusters to be extended
2020-11-10 12:21:51.159: 10906 sequences rejected
2020-11-10 12:21:51.229: 4665 cluster pairs to check and merge.

Hammock log file

2020-11-10 12:21:51.229: Merging clusters from 84 groups...
2020-11-10 12:21:51.442: Buiding hhs...
2020-11-10 12:21:52.254: HH clustering...
2020-11-10 12:24:04.540:
2020-11-10 12:24:04.540: Round 2:

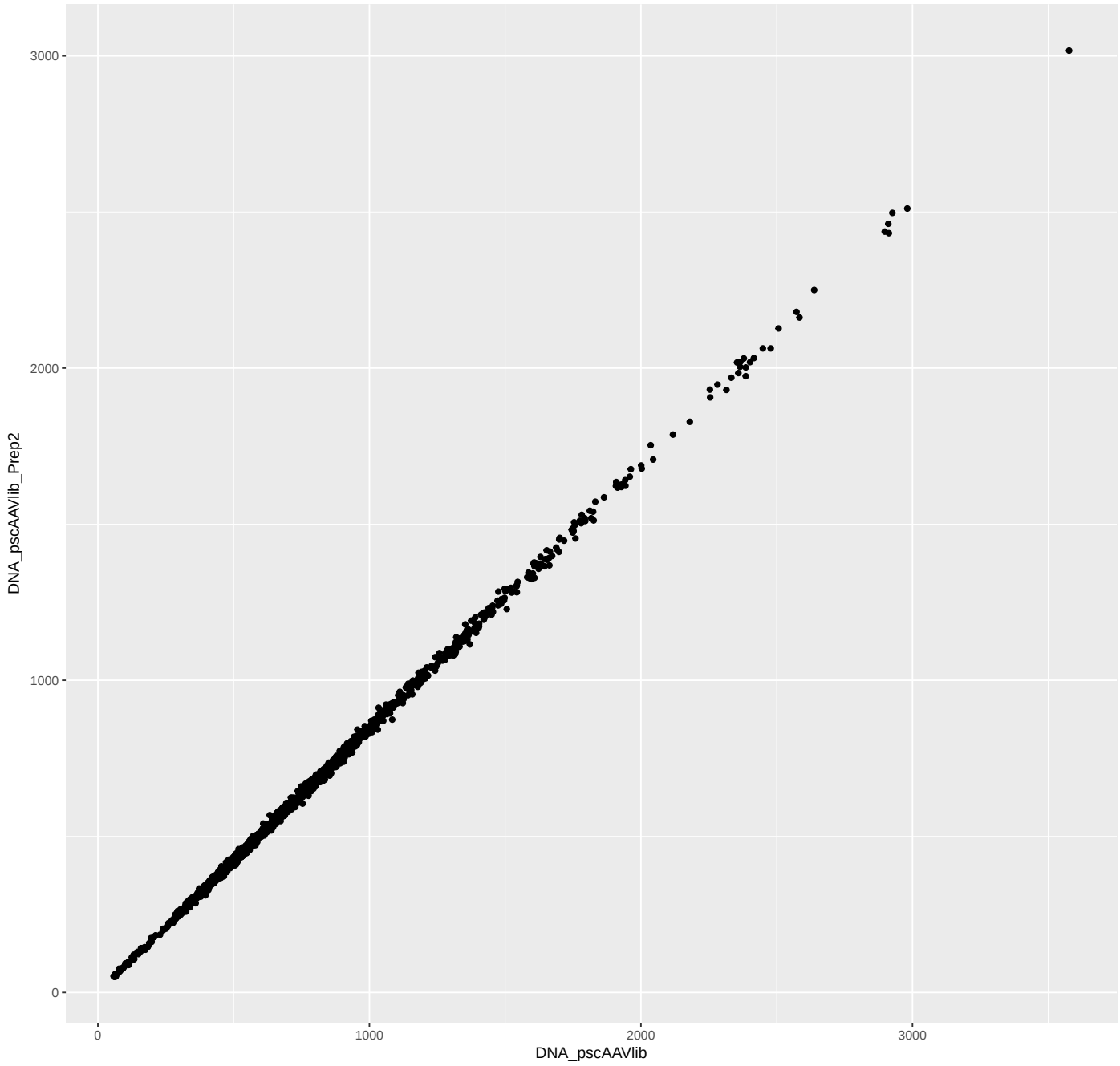
2020-11-10 12:24:04.540: 1739 clusters remaining
2020-11-10 12:24:04.541: Building hmms and searching database...
2020-11-10 12:24:21.926: Extending clusters...
2020-11-10 12:24:22.092: 12758 sequences to be inserted into clusters
2020-11-10 12:24:22.098: 1313 clusters to be extended
2020-11-10 12:24:31.689: 10118 sequences rejected
2020-11-10 12:24:32.073: 67740 cluster pairs to check and merge.
2020-11-10 12:24:32.073: Merging clusters from 1 groups...
2020-11-10 12:24:32.276: Buiding hhs...
2020-11-10 12:24:32.750: HH clustering...
2020-11-10 12:26:25.018:
2020-11-10 12:26:25.019: Round 3:

2020-11-10 12:26:25.019: 1558 clusters remaining
2020-11-10 12:26:25.019: Building hmms and searching database...
2020-11-10 12:26:42.198: Extending clusters...
2020-11-10 12:26:42.303: 16600 sequences to be inserted into clusters
2020-11-10 12:26:42.309: 1331 clusters to be extended
2020-11-10 12:26:50.802: 11821 sequences rejected
2020-11-10 12:26:50.811: Overlap threshold is 0. Running full cluster merging.
2020-11-10 12:26:50.982: Buiding hhs...
2020-11-10 12:26:51.484: HH clustering...
2020-11-10 12:29:01.024:
Ready. Clustering time : 459590
2020-11-10 12:29:01.024: Resulting clusers: 1340
2020-11-10 12:29:01.025: Containing 34315 unique sequences and 1936731 total sequences.
2020-11-10 12:29:01.043: Unique sequences not assigned: 25864, total sequences not assigned: 969778
2020-11-10 12:29:01.043: Saving results to outupt files...
2020-11-10 12:29:02.189: Results in: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences.tsv
2020-11-10 12:29:02.190: and: /home/rstudio/data/HammockLibDNAse/final_clusters.tsv
2020-11-10 12:29:02.190: and: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences_original_order.tsv
2020-11-10 12:29:02.190:
Calculating KLD...
2020-11-10 12:29:02.192: 21 clusters omitted from KLD calculation because each of them only contains a single unique sequence.
2020-11-10 12:29:05.241: Final system KLD over match state MSA positions: 21.284280102889714
2020-11-10 12:29:05.241: Final system KLD over all MSA positions: 39.68035578266766
2020-11-10 12:29:05.242: Program successfully ended.

Generation of Scatter plot generation

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockLibDNAse/final_clusters.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))



pred.points <- ggplot(data = ham.clusters, aes(x = DNA_pscAAVlib, y = DNA_pscAAVlib_Prep2)) +
    labs(x = "DNA_pscAAVlib", y = "DNA_pscAAVlib_Prep2") + geom_point()
```

```
print(pred.points)
```



Clustering DNAse resistant virions with library

```
select.samples <- all.samples[J(c("DNA_pscAAVlib_Prep2", "DNA_AAVlib_DNAse_3cpc",
    "DNA_AAVlib_DNAse_30cpc"))]

select.samples[, `:=`(BCcount, as.integer(mclapply(BC, function(x) length(table(strsplit(paste(t(x),
    collapse = ","), ","))), mc.cores = detectCores())))]
select.samples[, `:=`(Score, BCcount)]
select.samples.trsp <- unique(select.samples, by = c("Animals", "BC", "LUTnrs"))
```

```
fasta.names <- paste(1:nrow(select.samples.trsp), select.samples.trsp$Score,
    select.samples.trsp$Group, sep = "|")
write.fasta(as.list(select.samples.trsp$Peptide), fasta.names, "data/LibDNAsePeptides.fasta",
    open = "w", nbchar = 60, as.string = TRUE)

# Generate Scoring table for Weblogo Weighting
select.samples.pepMerge <- select.samples.trsp[, sum(Score), by = c("Peptide")]
setnames(select.samples.pepMerge, "V1", "Score")
```

Executing Hammock Clustering

```
Sys.setenv(PATH = paste("/root/HMMER/binaries", Sys.getenv("PATH"), sep = ":"),
    HHLIB = "/home/rstudio/Hammock_v_1.1.1/hhsuite-2.0.16/lib/hh/")
unlink("/home/rstudio/data/HammockLibDNAse", recursive = TRUE, force = FALSE)
sys.out <- system(paste("java -jar /home/rstudio/Hammock_v_1.1.1/dist/Hammock.jar full -i /home/rstudio/data/
    detectCores(), sep = ""), intern = TRUE, ignore.stdout = TRUE)
# Alternative parameters --use_clinkage --alignment_threshold 23
# --alignment_threshold 26 --assign_thresholds 50,40,30
hammock.log <- data.table(readLines("data/HammockLibDNAse/run.log"))

colnames(hammock.log) <- c("Hammock log file")
knitr::kable(hammock.log, longtable = T)
```

Hammock log file
2020-11-10 12:29:27.806:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-10 12:29:27.942: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/LibDNAsePeptides.fasta -d /home/rstudio/data/HammockLibDNAse –max_shift 7 -c 2000 -t 48
Complete list of input/output parameters:
-i, –input /home/rstudio/data/LibDNAsePeptides.fasta
-d, –output_directory /home/rstudio/data/HammockLibDNAse
-t, –thread 48
-l, –labels null
Complete list of clinkage clustering parameters:
-f, –file_format fasta
-m, –matrix /home/rstudio/Hammock_v_1.1.1/matrices/blosum62.txt
-g, –alignment_threshold (–greedy_threshold)null
-x, –max_shift 7
-p, –gap_penalty 0
-C, –cache_size_limit 1
2020-11-10 12:29:27.943: Loading input sequences...
2020-11-10 12:29:28.360: 60086 unique sequences loaded.
2020-11-10 12:29:28.395: 1535104 total sequences loaded.
2020-11-10 12:29:28.395: 60086 unique sequences after non-specified labels filtered out
2020-11-10 12:29:28.451: 1535104 total sequences after non-specified labels fileterd out

Hammock log file

2020-11-10 12:29:28.466: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-10 12:29:28.466: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-10 12:29:28.551: Generating input statistics...
2020-11-10 12:29:28.724: Greedy clustering threshold not set. Setting automatically to: 27
2020-11-10 12:29:28.724: Initial greedy clusters limit not set. Setting automatically to: 1502
2020-11-10 12:29:28.726: Greedy clustering...
2020-11-10 12:30:56.819: Ready. Clustering time: 88092
2020-11-10 12:30:56.819: Resulting clusers: 39583
2020-11-10 12:30:56.820: Building MSAs...
2020-11-10 12:30:58.433: Ready. Total time: 89707
2020-11-10 12:30:58.433: Saving results to output files...
2020-11-10 12:30:59.928: Greedy clustering results in: /home/rstudio/data/HammockLibDNAse/initial_clusters.tsv
2020-11-10 12:30:59.929: and: /home/rstudio/data/HammockLibDNAse/initial_clusters_sequences.tsv
2020-11-10 12:30:59.929: and:
/home/rstudio/data/HammockLibDNAse/initial_clusters_sequences_original_order.tsv
2020-11-10 12:30:59.929:
Loading clusters...
2020-11-10 12:31:00.169: Maximal alignment length not set. Setting automatically to: 32
2020-11-10 12:31:00.180: Minimal number of match states not set. Setting automatically to: 5
2020-11-10 12:31:00.580: Assign threshold sequence not set. Setting automatically to:
2020-11-10 12:31:00.585: 15.3,12.08,8.86,
2020-11-10 12:31:00.585: Overlap threshold not set. Setting automatically to:
2020-11-10 12:31:00.592: 11.27,6.44,0.0,
2020-11-10 12:31:00.592: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-10 12:31:00.599: 16.1,14.49,12.88,
2020-11-10 12:31:00.821: 6 clusters rejected because of match states and information content constraints.

Complete list of HMM-based clustering parameters:
-a, –part_threshold null
-s, –size_threshold null
-c, –count_threshold 2000
-n, –assign_thresholds 15.3,12.08,8.86,
-v, –overlap_thresholds 11.27,6.44,0.0,
-r, –merge_thresholds 16.1,14.49,12.88,
-e, –relative_thresholds false
-b, –absolute_thresholds true
-h, –min_conserved_positions 5
-y, –max_gap_proportion 0.05
-k, –min_ic 1.2
-j, –max_aln_length 32
-u, –max_inner_gaps 0
-q, –extension_increase_length false

2020-11-10 12:31:01.082:
Clustering in 3 rounds...
2020-11-10 12:31:01.084:
2020-11-10 12:31:01.084: Round 1:

2020-11-10 12:31:01.084: 1994 clusters remaining
2020-11-10 12:31:01.085: Building hmms and searching database...
2020-11-10 12:31:24.508: Extending clusters...
2020-11-10 12:31:24.742: 16358 sequences to be inserted into clusters
2020-11-10 12:31:24.755: 1568 clusters to be extended
2020-11-10 12:31:30.958: 11310 sequences rejected
2020-11-10 12:31:31.038: 4703 cluster pairs to check and merge.

Hammock log file

2020-11-10 12:31:31.038: Merging clusters from 84 groups...
2020-11-10 12:31:31.246: Buiding hhs...
2020-11-10 12:31:32.412: HH clustering...
2020-11-10 12:33:45.088:
2020-11-10 12:33:45.088: Round 2:

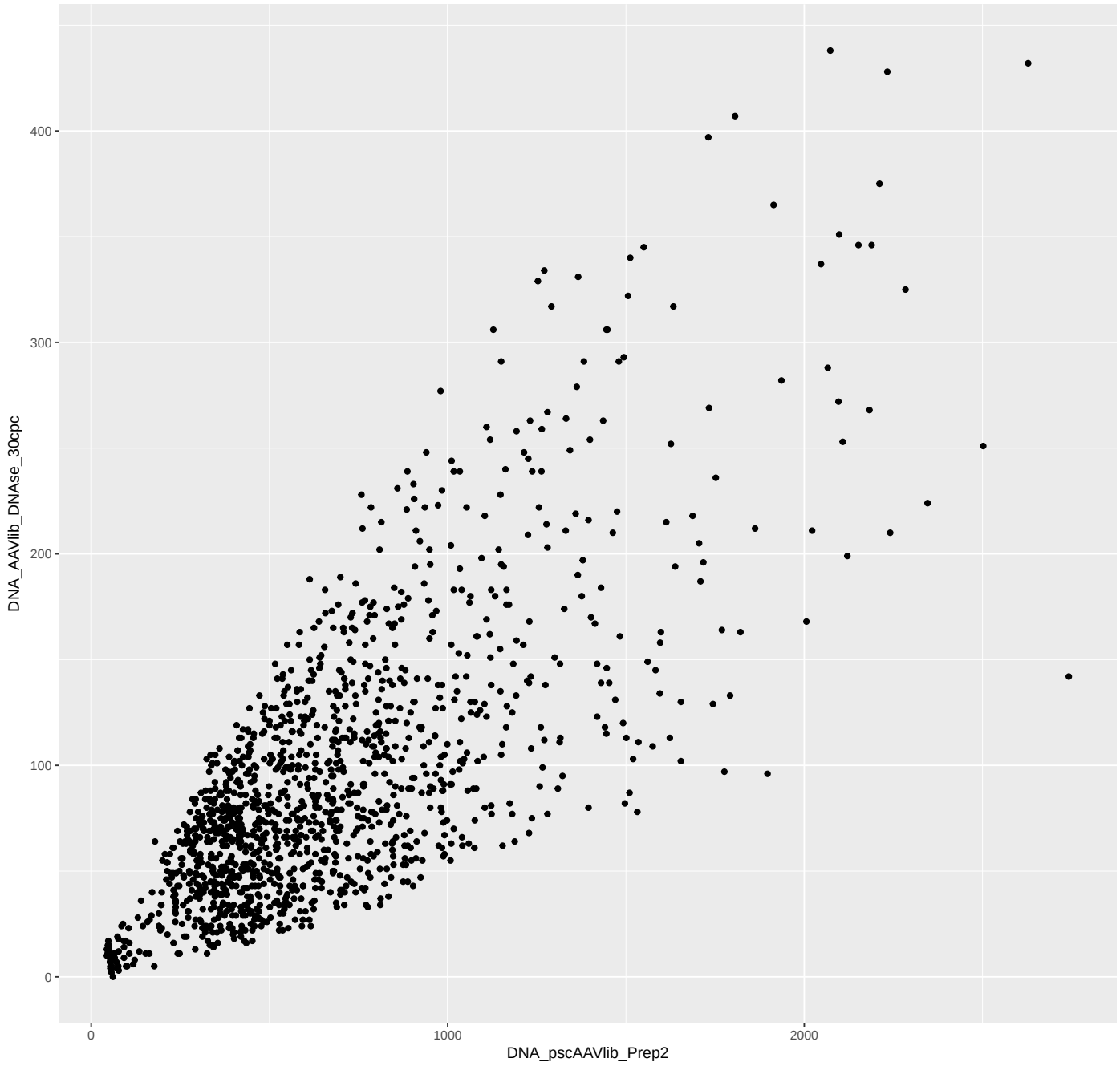
2020-11-10 12:33:45.089: 1735 clusters remaining
2020-11-10 12:33:45.089: Building hmms and searching database...
2020-11-10 12:34:02.317: Extending clusters...
2020-11-10 12:34:02.485: 12559 sequences to be inserted into clusters
2020-11-10 12:34:02.490: 1300 clusters to be extended
2020-11-10 12:34:07.647: 10059 sequences rejected
2020-11-10 12:34:07.981: 72128 cluster pairs to check and merge.
2020-11-10 12:34:07.981: Merging clusters from 1 groups...
2020-11-10 12:34:08.159: Buiding hhs...
2020-11-10 12:34:09.919: HH clustering...
2020-11-10 12:35:53.899:
2020-11-10 12:35:53.899: Round 3:

2020-11-10 12:35:53.899: 1571 clusters remaining
2020-11-10 12:35:53.900: Building hmms and searching database...
2020-11-10 12:36:10.743: Extending clusters...
2020-11-10 12:36:10.842: 16562 sequences to be inserted into clusters
2020-11-10 12:36:10.847: 1360 clusters to be extended
2020-11-10 12:36:23.354: 11873 sequences rejected
2020-11-10 12:36:23.359: Overlap threshold is 0. Running full cluster merging.
2020-11-10 12:36:23.522: Buiding hhs...
2020-11-10 12:36:23.905: HH clustering...
2020-11-10 12:38:37.796:
Ready. Clustering time : 456714
2020-11-10 12:38:37.796: Resulting clusers: 1345
2020-11-10 12:38:37.797: Containing 34643 unique sequences and 1030396 total sequences.
2020-11-10 12:38:37.816: Unique sequences not assigned: 25443, total sequences not assigned: 504708
2020-11-10 12:38:37.816: Saving results to outupt files...
2020-11-10 12:38:39.000: Results in: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences.tsv
2020-11-10 12:38:39.000: and: /home/rstudio/data/HammockLibDNAse/final_clusters.tsv
2020-11-10 12:38:39.001: and: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences_original_order.tsv
2020-11-10 12:38:39.001:
Calculating KLD...
2020-11-10 12:38:39.002: 31 clusters omitted from KLD calculation because each of them only contains a single unique sequence.
2020-11-10 12:38:42.144: Final system KLD over match state MSA positions: 21.314423213245473
2020-11-10 12:38:42.145: Final system KLD over all MSA positions: 39.98132228244945
2020-11-10 12:38:42.145: Program successfully ended.

Generation of Scatter plot generation

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockLibDNAse/final_clusters.tsv",
    header = TRUE, skip = 0, sep = "\t", stringsAsFactors = FALSE, fill = TRUE))



pred.points <- ggplot(data = ham.clusters, aes(x = DNA_pscAAVlib_Prep2, y = DNA_AAVlib_DNAse_30cpc)) +
    labs(x = "DNA_pscAAVlib_Prep2", y = "DNA_AAVlib_DNAse_30cpc") + geom_point()
```

```
print(pred.points)
```




```
print("Total analysis time:")
```

```
[1] "Total analysis time:"
```

```
print(Sys.time() - strt1)
```

```
Time difference of 31.74036 mins
```

```
devtools::session_info()
```

```
Session info -------------------------------------------------------------
```

```
 setting  value
 version  R version 3.4.2 (2017-09-28)
 system   x86_64, linux-gnu
```

```
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 tz       UTC
 date     2020-11-10
```

Packages ----------

```
 package               * version  date       source
 acepack                 1.4.1    2016-10-29 CRAN (R 3.4.2)
 ade4                    1.7-8    2017-08-09 CRAN (R 3.4.2)
 annotate                1.54.0   2017-11-29 Bioconductor
 AnnotationDbi           1.38.2   2017-11-29 Bioconductor
 AnnotationFilter        1.0.0    2017-11-29 Bioconductor
 AnnotationHub           2.8.3    2017-11-29 Bioconductor
 backports               1.1.1    2017-09-25 CRAN (R 3.4.2)
 base                  * 3.4.2    2017-10-06 local
 base64enc               0.1-3    2015-07-28 CRAN (R 3.4.2)
 Biobase               * 2.36.2   2017-11-29 Bioconductor
 BiocGenerics          * 0.22.1   2017-11-29 Bioconductor
 BiocInstaller           1.26.1   2017-10-10 Bioconductor
 BiocParallel            1.10.1   2017-11-29 Bioconductor
 biomaRt                 2.32.1   2017-11-29 Bioconductor
 Biostrings            * 2.44.2   2017-11-29 Bioconductor
 biovizBase              1.24.0   2017-11-29 Bioconductor
 bit                     1.1-12   2014-04-09 CRAN (R 3.4.2)
 bit64                   0.9-7    2017-05-08 CRAN (R 3.4.2)
 bitops                  1.0-6    2013-08-17 CRAN (R 3.4.2)
 blob                    1.1.0    2017-06-17 CRAN (R 3.4.2)
 BSgenome                1.44.2   2017-11-29 Bioconductor
 checkmate               1.8.4    2017-09-25 CRAN (R 3.4.2)
 cluster                 2.0.6    2017-03-16 CRAN (R 3.4.2)
 codetools               0.2-15   2016-10-05 CRAN (R 3.4.2)
 colorspace              1.3-2    2016-12-14 CRAN (R 3.4.2)
 compiler                3.4.2    2017-10-06 local
 curl                    2.8.1    2017-07-21 CRAN (R 3.4.2)
 data.table            * 1.10.4-2 2017-10-12 url
 datasets              * 3.4.2    2017-10-06 local
 DBI                     0.7      2017-06-18 CRAN (R 3.4.2)
 DelayedArray          * 0.2.7    2017-11-29 Bioconductor
 DESeq2                * 1.16.1   2017-11-29 Bioconductor
 devtools              * 1.13.3   2017-08-02 CRAN (R 3.4.2)
 dichromat               2.0-0    2013-01-24 CRAN (R 3.4.2)
 digest                  0.6.12   2017-01-27 CRAN (R 3.4.2)
 doParallel            * 1.0.11   2017-09-28 CRAN (R 3.4.2)
 ensembldb               2.0.4    2017-11-29 Bioconductor
 evaluate                0.10.1   2017-06-24 CRAN (R 3.4.2)
 foreach               * 1.4.3    2015-10-13 CRAN (R 3.4.2)
 foreign                 0.8-69   2017-06-21 CRAN (R 3.4.2)
 formatR               * 1.5      2017-04-25 CRAN (R 3.4.2)
 Formula                 1.2-2    2017-07-10 CRAN (R 3.4.2)
 futile.logger         * 1.4.3    2016-07-10 cran (@1.4.3)
 futile.options          1.0.0    2010-04-06 cran (@1.0.0)
 genefilter              1.58.1   2017-11-29 Bioconductor
 geneplotter             1.54.0   2017-11-29 Bioconductor
 GenomeInfoDb          * 1.12.3   2017-11-29 Bioconductor
 GenomeInfoDbData        0.99.0   2017-11-29 Bioconductor
 GenomicAlignments     * 1.12.2   2017-11-29 Bioconductor
 GenomicFeatures         1.28.5   2017-11-29 Bioconductor
 GenomicRanges         * 1.28.6   2017-11-29 Bioconductor
```

```
GGally                   1.3.2    2017-08-02 CRAN (R 3.4.2)
ggbio                  * 1.24.1   2017-11-29 Bioconductor
ggplot2                * 2.2.1    2016-12-30 CRAN (R 3.4.2)
graph                    1.54.0   2017-11-29 Bioconductor
graphics               * 3.4.2    2017-10-06 local
grDevices              * 3.4.2    2017-10-06 local
grid                   * 3.4.2    2017-10-06 local
gridExtra                2.3      2017-09-09 CRAN (R 3.4.2)
gtable                   0.2.0    2016-02-26 CRAN (R 3.4.2)
highr                    0.6      2016-05-09 CRAN (R 3.4.2)
Hmisc                    4.0-3    2017-05-02 CRAN (R 3.4.2)
hms                      0.3      2016-11-22 CRAN (R 3.4.2)
htmlTable                1.9      2017-01-26 CRAN (R 3.4.2)
htmltools                0.3.6    2017-04-28 CRAN (R 3.4.2)
htmlwidgets              0.9      2017-07-10 CRAN (R 3.4.2)
httpuv                   1.3.5    2017-07-04 CRAN (R 3.4.2)
httr                     1.3.1    2017-08-20 CRAN (R 3.4.2)
interactiveDisplayBase   1.14.0   2017-11-29 Bioconductor
IRanges                * 2.10.5   2017-11-29 Bioconductor
iterators              * 1.0.8    2015-10-13 CRAN (R 3.4.2)
kableExtra             * 0.5.2    2017-09-15 url
knitr                  * 1.17     2017-08-10 CRAN (R 3.4.2)
labeling                 0.3      2014-08-23 CRAN (R 3.4.2)
lambda.r                 1.2      2017-09-16 cran (@1.2)
lattice                  0.20-35  2017-03-25 CRAN (R 3.4.2)
latticeExtra             0.6-28   2016-02-09 CRAN (R 3.4.2)
lazyeval                 0.2.0    2016-06-12 CRAN (R 3.4.2)
locfit                   1.5-9.1  2013-04-20 CRAN (R 3.4.2)
magrittr                 1.5      2014-11-22 CRAN (R 3.4.2)
Matrix                   1.2-11   2017-08-21 url
matrixStats            * 0.52.2   2017-04-14 CRAN (R 3.4.2)
memoise                  1.1.0    2017-04-21 CRAN (R 3.4.2)
methods                * 3.4.2    2017-10-06 local
mime                     0.5      2016-07-07 CRAN (R 3.4.2)
munsell                  0.4.3    2016-02-13 CRAN (R 3.4.2)
nnet                     7.3-12   2016-02-02 CRAN (R 3.4.2)
OrganismDbi              1.18.1   2017-11-29 Bioconductor
parallel               * 3.4.2    2017-10-06 local
pheatmap               * 1.0.8    2015-12-11 CRAN (R 3.4.2)
plyr                   * 1.8.4    2016-06-08 CRAN (R 3.4.2)
ProtGenerics             1.8.0    2017-11-29 Bioconductor
R6                       2.2.2    2017-06-17 CRAN (R 3.4.2)
RBGL                     1.52.0   2017-11-29 Bioconductor
RColorBrewer             1.1-2    2014-12-07 CRAN (R 3.4.2)
Rcpp                     0.12.13  2017-09-28 url
RCurl                    1.95-4.8 2016-03-01 CRAN (R 3.4.2)
readr                    1.1.1    2017-05-16 CRAN (R 3.4.2)
reshape                * 0.8.7    2017-08-06 CRAN (R 3.4.2)
reshape2               * 1.4.2    2016-10-22 CRAN (R 3.4.2)
rlang                    0.1.2    2017-08-09 CRAN (R 3.4.2)
rmarkdown                1.6      2017-06-15 url
rpart                    4.1-11   2017-04-21 CRAN (R 3.4.2)
rprojroot                1.2      2017-01-16 CRAN (R 3.4.2)
Rsamtools              * 1.28.0   2017-11-29 Bioconductor
RSQLite                  2.0      2017-06-19 CRAN (R 3.4.2)
rtracklayer              1.36.6   2017-11-29 Bioconductor
rvest                    0.3.2    2016-06-17 CRAN (R 3.4.2)
S4Vectors              * 0.14.7   2017-11-29 Bioconductor
scales                   0.5.0    2017-08-24 CRAN (R 3.4.2)
```

```
seqinr                * 3.4-5    2017-08-01 CRAN (R 3.4.2)
shiny                   1.0.5    2017-08-23 CRAN (R 3.4.2)
splines                 3.4.2    2017-10-06 local
stats                 * 3.4.2    2017-10-06 local
stats4                * 3.4.2    2017-10-06 local
stringi                 1.1.5    2017-04-07 url
stringr                 1.2.0    2017-02-18 CRAN (R 3.4.2)
SummarizedExperiment  * 1.6.5    2017-11-29 Bioconductor
survival                2.41-3   2017-04-04 CRAN (R 3.4.2)
tibble                  1.3.4    2017-08-22 CRAN (R 3.4.2)
tools                   3.4.2    2017-10-06 local
utils                 * 3.4.2    2017-10-06 local
VariantAnnotation       1.22.3   2017-11-29 Bioconductor
VennDiagram           * 1.6.17   2016-04-18 url
withr                   2.0.0    2017-07-28 url
XML                     3.98-1.9 2017-06-19 CRAN (R 3.4.2)
xml2                    1.1.1    2017-01-24 CRAN (R 3.4.2)
xtable                  1.8-2    2016-02-05 CRAN (R 3.4.2)
XVector               * 0.16.0   2017-11-29 Bioconductor
yaml                    2.1.14   2016-11-12 CRAN (R 3.4.2)
zlibbioc                1.22.0   2017-11-29 Bioconductor
```