Heatmaps generated from HMM peptide clustering

Tomas Bjorklund

Mon Nov 9 21:56:38 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_3000cpc_Organoid_MD101", "mRNA_30cpc_Organoid_MD114",
    "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"))]

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[Animals == "mRNA_3000cpc_Organoid_MD101", `:=`(BCcount, as.integer(BCcount/5)),
    ]  # Removes unclear 3000cpc reads
select.samples <- select.samples[BCcount >= 1]
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/invivoSamplesPeptidesOrganoids.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/HammockInVivoOrganoids", 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/HammockInVivoOrganoids/run.log"))

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

Hammock log file
2020-11-09 21:57:17.520:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-09 21:57:17.692: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/invivoSamplesPeptidesOrganoids.fasta -d /home/rstudio/data/HammockInVivoOrganoids –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/invivoSamplesPeptidesOrganoids.fasta
-d, –output_directory /home/rstudio/data/HammockInVivoOrganoids
-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
2020-11-09 21:57:17.693: Loading input sequences...
2020-11-09 21:57:17.844: 16607 unique sequences loaded.
2020-11-09 21:57:17.860: 41142 total sequences loaded.
2020-11-09 21:57:17.861: 16607 unique sequences after non-specified labels filtered out
2020-11-09 21:57:17.878: 41142 total sequences after non-specified labels fileterd out
2020-11-09 21:57:17.883: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-09 21:57:17.883: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-09 21:57:17.919: Generating input statistics...
2020-11-09 21:57:17.985: Initial greedy clusters limit not set. Setting automatically to: 415
2020-11-09 21:57:17.987: Greedy clustering...
2020-11-09 21:57:31.228: Ready. Clustering time: 13241
2020-11-09 21:57:31.229: Resulting clusers: 13546
2020-11-09 21:57:31.229: Building MSAs...
2020-11-09 21:57:31.673: Ready. Total time: 13686
2020-11-09 21:57:31.674: Saving results to output files...
2020-11-09 21:57:32.204: Greedy clustering results in: /home/rstudio/data/HammockInVivoOrganoids/initial_clusters.tsv
2020-11-09 21:57:32.205: and: /home/rstudio/data/HammockInVivoOrganoids/initial_clusters_sequences.tsv
2020-11-09 21:57:32.205: and: /home/rstudio/data/HammockInVivoOrganoids/initial_clusters_sequences_original_order.tsv
2020-11-09 21:57:32.205:

Hammock log file

Loading clusters...
2020-11-09 21:57:32.308: Maximal alignment length not set. Setting automatically to: 31
2020-11-09 21:57:32.315: Minimal number of match states not set. Setting automatically to: 5
2020-11-09 21:57:32.459: Overlap threshold not set. Setting automatically to:
2020-11-09 21:57:32.469: 10.83,6.19,0.0,
2020-11-09 21:57:32.470: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-09 21:57:32.476: 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

2020-11-09 21:57:32.575:
Clustering in 3 rounds...
2020-11-09 21:57:32.577:
2020-11-09 21:57:32.578: Round 1:

2020-11-09 21:57:32.578: 250 clusters remaining
2020-11-09 21:57:32.578: Building hmms and searching database...
2020-11-09 21:57:34.856: Extending clusters...
2020-11-09 21:57:34.908: 0 sequences to be inserted into clusters
2020-11-09 21:57:34.908: 0 clusters to be extended
2020-11-09 21:57:34.909: 0 sequences rejected
2020-11-09 21:57:34.915: 104 cluster pairs to check and merge.
2020-11-09 21:57:34.915: Merging clusters from 32 groups...
2020-11-09 21:57:34.951: Buiding hhs...
2020-11-09 21:57:35.026: HH clustering...
2020-11-09 21:57:36.437:
2020-11-09 21:57:36.437: Round 2:

2020-11-09 21:57:36.437: 245 clusters remaining
2020-11-09 21:57:36.437: Building hmms and searching database...
2020-11-09 21:57:38.239: Extending clusters...
2020-11-09 21:57:38.251: 0 sequences to be inserted into clusters
2020-11-09 21:57:38.251: 0 clusters to be extended
2020-11-09 21:57:38.252: 0 sequences rejected
2020-11-09 21:57:38.265: 2353 cluster pairs to check and merge.
2020-11-09 21:57:38.266: Merging clusters from 1 groups...
2020-11-09 21:57:38.295: Buiding hhs...
2020-11-09 21:57:38.367: HH clustering...
2020-11-09 21:57:45.181:
2020-11-09 21:57:45.182: Round 3:

2020-11-09 21:57:45.182: 234 clusters remaining

Hammock log file

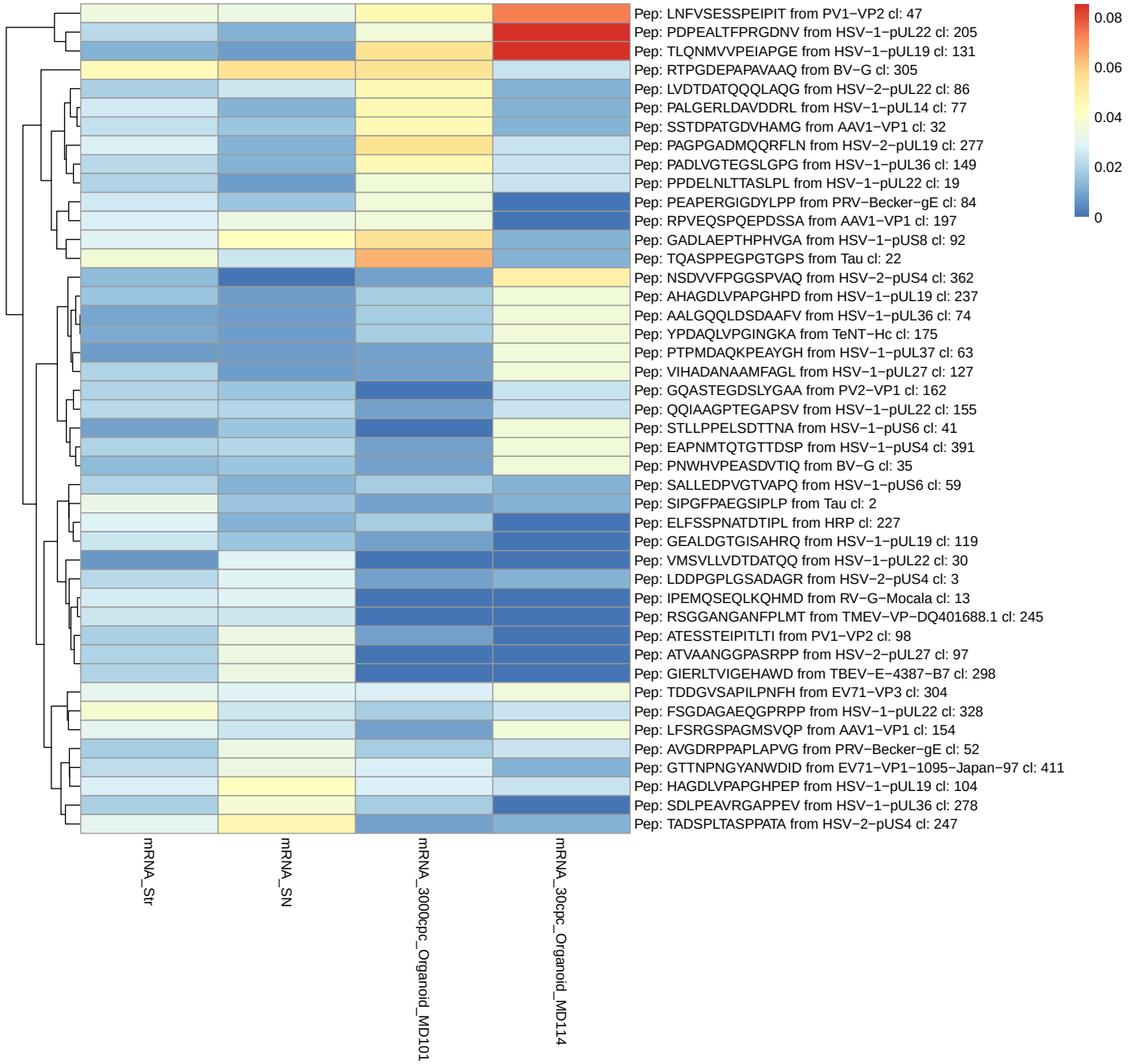
```
2020-11-09 21:57:45.182: Building hmms and searching database...
2020-11-09 21:57:46.921: Extending clusters...
2020-11-09 21:57:46.955: 9 sequences to be inserted into clusters
2020-11-09 21:57:46.955: 7 clusters to be extended
2020-11-09 21:57:46.963: 2 sequences rejected
2020-11-09 21:57:46.965: Overlap threshold is 0. Running full cluster merging.
2020-11-09 21:57:46.992: Buiding hhs...
2020-11-09 21:57:47.006: HH clustering...
2020-11-09 21:58:00.822:
Ready. Clustering time : 28246
2020-11-09 21:58:00.822: Resulting clusers: 207
2020-11-09 21:58:00.823: Containing 2523 unique sequences and 10550 total sequences.
2020-11-09 21:58:00.833: Unique sequences not assigned: 14084, total sequences not assigned: 30592
2020-11-09 21:58:00.833: Saving results to outupt files...
2020-11-09 21:58:01.068: Results in: /home/rstudio/data/HammockInVivoOrganoids/final_clusters_sequences.tsv
2020-11-09 21:58:01.068: and: /home/rstudio/data/HammockInVivoOrganoids/final_clusters.tsv
2020-11-09 21:58:01.069: and:
/home/rstudio/data/HammockInVivoOrganoids/final_clusters_sequences_original_order.tsv
2020-11-09 21:58:01.069:
Calculating KLD...
2020-11-09 21:58:01.449: Final system KLD over match state MSA positions: 19.153047632455923
2020-11-09 21:58:01.449: Final system KLD over all MSA positions: 30.417697944217444
2020-11-09 21:58:01.449: Program successfully ended.
```

Generation of Weblogo visualization

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockInVivoOrganoids/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/HammockInVivoOrganoids/final_clusters_sequences
    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/HammockInVivoOrganoids/", "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/HammockInVivoOrganoids/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/HammockInVivoOrganoids/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/HammockInVivoOrganoids/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 = " "))]
```

```r
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_3000cpc_Organoid_MD101", "mRNA_30cpc_Organoid_MD114",
    "mRNA_3cpc_HEK293T", "mRNA_30cpc_HEK293T", "mRNA_3cpc_pNeuron", "mRNA_30cpc_pNeuron"))]  # 'mRNA_3000cpc_

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[Animals == "mRNA_3000cpc_Organoid_MD101", `:=`(BCcount, as.integer(BCcount/5)),
    ]  # Removes unclear 3000cpc reads
select.samples <- select.samples[BCcount >= 1]
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/invitroSamplesPeptidesOrganoids.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/HammockInVitroOrganoids", 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/HammockInVitroOrganoids/run.log"))

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

Hammock log file
2020-11-09 21:58:21.589:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-09 21:58:21.722: Program started in mode “full”.
Command-line arguments:
full -i /home/rstudio/data/invitroSamplesPeptidesOrganoids.fasta -d /home/rstudio/data/HammockInVitroOrganoids –max_shift 7 -c 50 -t 48
Complete list of input/output parameters:
-i, –input /home/rstudio/data/invitroSamplesPeptidesOrganoids.fasta
-d, –output_directory /home/rstudio/data/HammockInVitroOrganoids
-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

2020-11-09 21:58:21.723: Loading input sequences...
2020-11-09 21:58:21.759: 2349 unique sequences loaded.
2020-11-09 21:58:21.762: 2659 total sequences loaded.
2020-11-09 21:58:21.762: 2349 unique sequences after non-specified labels filtered out
2020-11-09 21:58:21.767: 2659 total sequences after non-specified labels fileterd out
2020-11-09 21:58:21.768: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-09 21:58:21.768: Up to 10 000 unique sequences. Using clinkage clustering. Use –use_greedy to force greedy clustering
2020-11-09 21:58:21.774: Generating input statistics...
2020-11-09 21:58:21.782: Clinkage clustering threshold not set. Setting automatically to: 26
2020-11-09 21:58:21.784: Clinkage clustering...
2020-11-09 21:58:34.841: Ready. Clustering time: 13057
2020-11-09 21:58:34.843: Resulting clusers: 1177
2020-11-09 21:58:34.843: Building MSAs...
2020-11-09 21:58:35.240: Ready. Total time: 13456
2020-11-09 21:58:35.241: Saving results to output files...
2020-11-09 21:58:35.487: Clinkage clustering results in:
/home/rstudio/data/HammockInVitroOrganoids/initial_clusters.tsv
2020-11-09 21:58:35.487: and: /home/rstudio/data/HammockInVitroOrganoids/initial_clusters_sequences.tsv
2020-11-09 21:58:35.487: and:
/home/rstudio/data/HammockInVitroOrganoids/initial_clusters_sequences_original_order.tsv
2020-11-09 21:58:35.488:
Loading clusters...
2020-11-09 21:58:35.510: Maximal alignment length not set. Setting automatically to: 30
2020-11-09 21:58:35.511: Minimal number of match states not set. Setting automatically to: 5
2020-11-09 21:58:35.548: Assign threshold sequence not set. Setting automatically to:
2020-11-09 21:58:35.552: 14.37,11.35,8.32,
2020-11-09 21:58:35.552: Overlap threshold not set. Setting automatically to:
2020-11-09 21:58:35.553: 10.59,6.05,0.0,
2020-11-09 21:58:35.553: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-09 21:58:35.554: 15.13,13.61,12.1,

Complete list of HMM-based clustering parameters:
-a, –part_threshold null
-s, –size_threshold null
-c, –count_threshold 50
-n, –assign_thresholds 14.37,11.35,8.32,
-v, –overlap_thresholds 10.59,6.05,0.0,
-r, –merge_thresholds 15.13,13.61,12.1,
-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

2020-11-09 21:58:35.584:
Clustering in 3 rounds...
2020-11-09 21:58:35.586:
2020-11-09 21:58:35.586: Round 1:

2020-11-09 21:58:35.586: 50 clusters remaining
2020-11-09 21:58:35.586: Building hmms and searching database...
2020-11-09 21:58:36.088: Extending clusters...
2020-11-09 21:58:36.091: 19 sequences to be inserted into clusters
2020-11-09 21:58:36.091: 14 clusters to be extended
2020-11-09 21:58:36.105: 13 sequences rejected
2020-11-09 21:58:36.106: 0 cluster pairs to check and merge.
2020-11-09 21:58:36.106: Merging clusters from 0 groups...
2020-11-09 21:58:36.114: Buiding hhs...
2020-11-09 21:58:36.115: HH clustering...
2020-11-09 21:58:36.123:
2020-11-09 21:58:36.123: Round 2:

2020-11-09 21:58:36.123: 50 clusters remaining
2020-11-09 21:58:36.124: Building hmms and searching database...
2020-11-09 21:58:36.568: Extending clusters...
2020-11-09 21:58:36.570: 24 sequences to be inserted into clusters
2020-11-09 21:58:36.570: 17 clusters to be extended
2020-11-09 21:58:36.580: 18 sequences rejected
2020-11-09 21:58:36.581: 26 cluster pairs to check and merge.
2020-11-09 21:58:36.581: Merging clusters from 5 groups...
2020-11-09 21:58:36.589: Buiding hhs...
2020-11-09 21:58:36.609: HH clustering...
2020-11-09 21:58:37.218:
2020-11-09 21:58:37.219: Round 3:

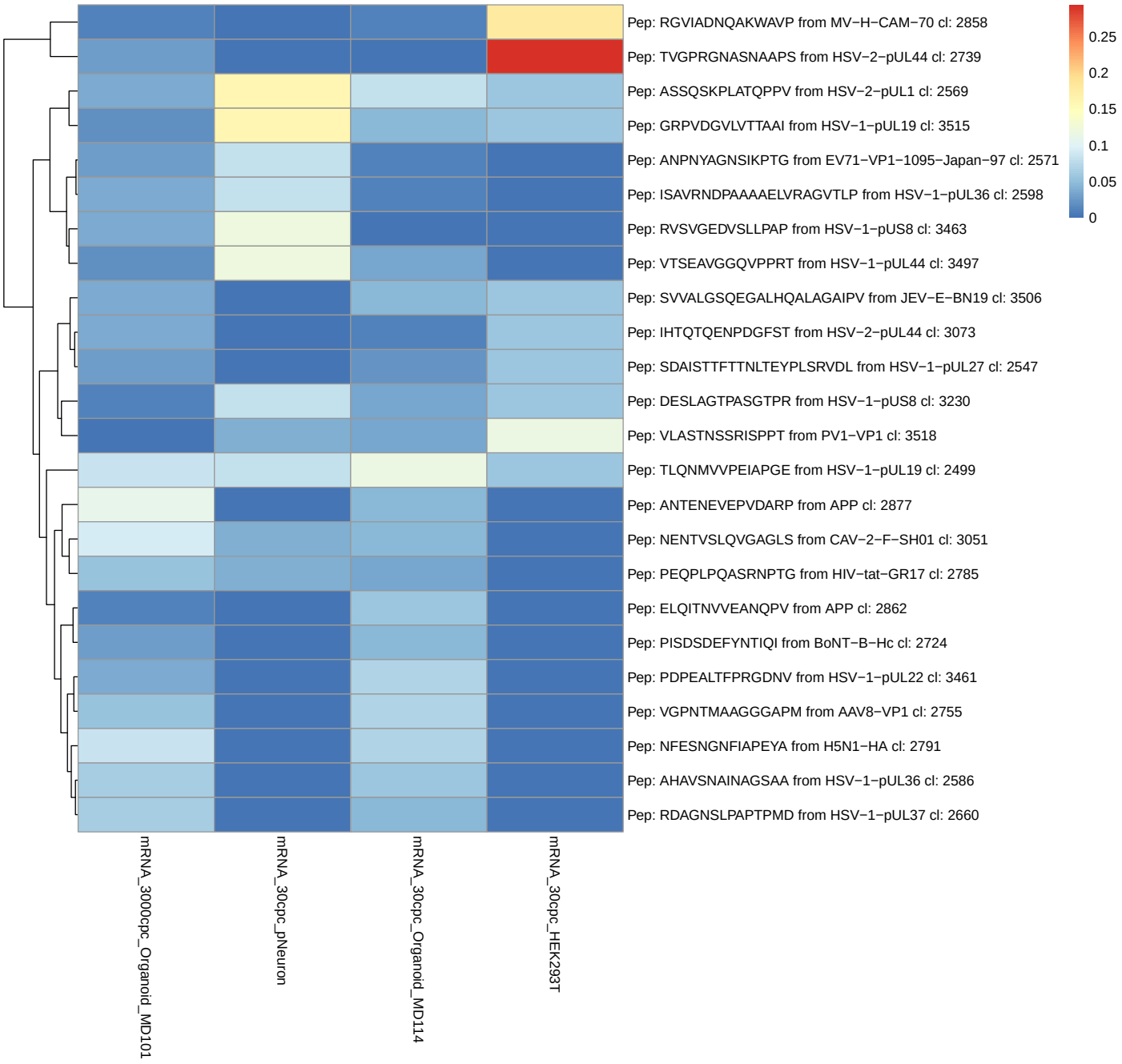
2020-11-09 21:58:37.219: 49 clusters remaining
2020-11-09 21:58:37.219: Building hmms and searching database...
2020-11-09 21:58:37.634: Extending clusters...
2020-11-09 21:58:37.637: 72 sequences to be inserted into clusters
2020-11-09 21:58:37.638: 32 clusters to be extended
2020-11-09 21:58:37.655: 55 sequences rejected
2020-11-09 21:58:37.656: Overlap threshold is 0. Running full cluster merging.
2020-11-09 21:58:37.662: Buiding hhs...
2020-11-09 21:58:37.682: HH clustering...
2020-11-09 21:58:39.389:
Ready. Clustering time : 3805
2020-11-09 21:58:39.389: Resulting clusers: 46
2020-11-09 21:58:39.389: Containing 269 unique sequences and 392 total sequences.
2020-11-09 21:58:39.391: Unique sequences not assigned: 2080, total sequences not assigned: 2267
2020-11-09 21:58:39.391: Saving results to outupt files...
2020-11-09 21:58:39.426: Results in: /home/rstudio/data/HammockInVitroOrganoids/final_clusters_sequences.tsv
2020-11-09 21:58:39.426: and: /home/rstudio/data/HammockInVitroOrganoids/final_clusters.tsv
2020-11-09 21:58:39.427: and:
/home/rstudio/data/HammockInVitroOrganoids/final_clusters_sequences_original_order.tsv
2020-11-09 21:58:39.427:
Calculating KLD...
2020-11-09 21:58:39.488: Final system KLD over match state MSA positions: 12.44161313602254
2020-11-09 21:58:39.488: Final system KLD over all MSA positions: 17.31084346721962

Hammock log file
2020-11-09 21:58:39.488: Program successfully ended.

Generation of Weblogo visualization

```
ham.clusters <- data.table(read.table("/home/rstudio/data/HammockInVitroOrganoids/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/HammockInVitroOrganoids/final_clusters_sequence
    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/HammockInVitroOrganoids/", "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/HammockInVitroOrganoids/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/HammockInVitroOrganoids/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/HammockInVitroOrganoids/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
2020-11-09 21:59:19.724:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-09 21:59:19.856: 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
2020-11-09 21:59:19.857: Loading input sequences...
2020-11-09 21:59:20.110: 49840 unique sequences loaded.
2020-11-09 21:59:20.139: 203706 total sequences loaded.

Hammock log file

2020-11-09 21:59:20.139: 49840 unique sequences after non-specified labels filtered out
2020-11-09 21:59:20.181: 203706 total sequences after non-specified labels fileterd out
2020-11-09 21:59:20.193: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-09 21:59:20.193: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-09 21:59:20.259: Generating input statistics...
2020-11-09 21:59:20.379: Greedy clustering threshold not set. Setting automatically to: 27
2020-11-09 21:59:20.379: Initial greedy clusters limit not set. Setting automatically to: 1246
2020-11-09 21:59:20.381: Greedy clustering...
2020-11-09 22:00:29.907: Ready. Clustering time: 69526
2020-11-09 22:00:29.908: Resulting clusers: 35413
2020-11-09 22:00:29.909: Building MSAs...
2020-11-09 22:00:31.169: Ready. Total time: 70788
2020-11-09 22:00:31.169: Saving results to output files...
2020-11-09 22:00:32.411: Greedy clustering results in: /home/rstudio/data/HammockDNAse/initial_clusters.tsv
2020-11-09 22:00:32.411: and: /home/rstudio/data/HammockDNAse/initial_clusters_sequences.tsv
2020-11-09 22:00:32.411: and: /home/rstudio/data/HammockDNAse/initial_clusters_sequences_original_order.tsv
2020-11-09 22:00:32.411:
Loading clusters...
2020-11-09 22:00:32.611: Maximal alignment length not set. Setting automatically to: 32
2020-11-09 22:00:32.624: Minimal number of match states not set. Setting automatically to: 5
2020-11-09 22:00:33.022: Assign threshold sequence not set. Setting automatically to:
2020-11-09 22:00:33.030: 15.14,11.95,8.77,
2020-11-09 22:00:33.030: Overlap threshold not set. Setting automatically to:
2020-11-09 22:00:33.034: 11.16,6.38,0.0,
2020-11-09 22:00:33.035: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-09 22:00:33.039: 15.94,14.34,12.75,
2020-11-09 22:00:33.214: 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

2020-11-09 22:00:33.480:
Clustering in 3 rounds...
2020-11-09 22:00:33.482:
2020-11-09 22:00:33.482: Round 1:

2020-11-09 22:00:33.482: 1994 clusters remaining
2020-11-09 22:00:33.483: Building hmms and searching database...
2020-11-09 22:00:56.006: Extending clusters...
2020-11-09 22:00:56.212: 13792 sequences to be inserted into clusters
2020-11-09 22:00:56.224: 1553 clusters to be extended
2020-11-09 22:01:01.999: 10675 sequences rejected

Hammock log file

2020-11-09 22:01:02.089: 5089 cluster pairs to check and merge.
2020-11-09 22:01:02.090: Merging clusters from 98 groups...
2020-11-09 22:01:02.347: Buiding hhs...
2020-11-09 22:01:03.196: HH clustering...
2020-11-09 22:04:10.096:
2020-11-09 22:04:10.097: Round 2:

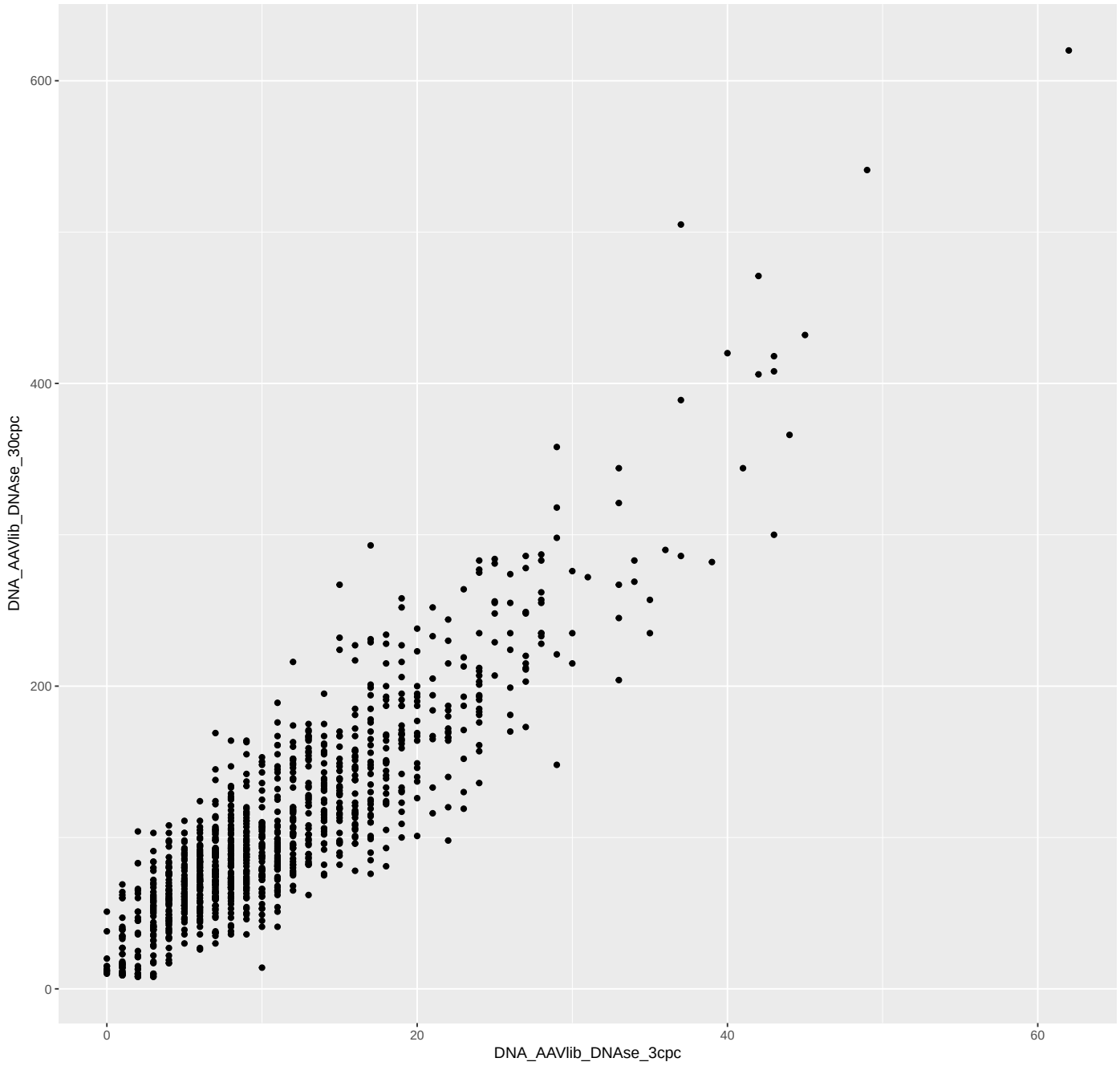
2020-11-09 22:04:10.097: 1530 clusters remaining
2020-11-09 22:04:10.097: Building hmms and searching database...
2020-11-09 22:04:24.554: Extending clusters...
2020-11-09 22:04:24.674: 11386 sequences to be inserted into clusters
2020-11-09 22:04:24.679: 1184 clusters to be extended
2020-11-09 22:04:37.092: 9109 sequences rejected
2020-11-09 22:04:37.358: 51972 cluster pairs to check and merge.
2020-11-09 22:04:37.359: Merging clusters from 1 groups...
2020-11-09 22:04:37.527: Buiding hhs...
2020-11-09 22:04:38.633: HH clustering...
2020-11-09 22:06:03.044:
2020-11-09 22:06:03.044: Round 3:

2020-11-09 22:06:03.045: 1374 clusters remaining
2020-11-09 22:06:03.045: Building hmms and searching database...
2020-11-09 22:06:16.699: Extending clusters...
2020-11-09 22:06:16.793: 14554 sequences to be inserted into clusters
2020-11-09 22:06:16.798: 1204 clusters to be extended
2020-11-09 22:06:26.047: 10703 sequences rejected
2020-11-09 22:06:26.052: Overlap threshold is 0. Running full cluster merging.
2020-11-09 22:06:26.196: Buiding hhs...
2020-11-09 22:06:27.065: HH clustering...
2020-11-09 22:08:29.364:
Ready. Clustering time : 475884
2020-11-09 22:08:29.365: Resulting clusers: 1127
2020-11-09 22:08:29.366: Containing 25589 unique sequences and 134333 total sequences.
2020-11-09 22:08:29.388: Unique sequences not assigned: 24251, total sequences not assigned: 69373
2020-11-09 22:08:29.388: Saving results to outupt files...
2020-11-09 22:08:30.297: Results in: /home/rstudio/data/HammockDNAse/final_clusters_sequences.tsv
2020-11-09 22:08:30.297: and: /home/rstudio/data/HammockDNAse/final_clusters.tsv
2020-11-09 22:08:30.297: and: /home/rstudio/data/HammockDNAse/final_clusters_sequences_original_order.tsv
2020-11-09 22:08:30.297:
Calculating KLD...
2020-11-09 22:08:30.299: 21 clusters omitted from KLD calculation because each of them only contains a single unique sequence.
2020-11-09 22:08:32.826: Final system KLD over match state MSA positions: 20.23129789753592
2020-11-09 22:08:32.826: Final system KLD over all MSA positions: 36.0292448781723
2020-11-09 22:08:32.826: 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
2020-11-09 22:08:56.339:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-09 22:08:56.472: 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
2020-11-09 22:08:56.473: Loading input sequences...
2020-11-09 22:08:56.891: 60179 unique sequences loaded.
2020-11-09 22:08:56.925: 2906509 total sequences loaded.
2020-11-09 22:08:56.925: 60179 unique sequences after non-specified labels filtered out
2020-11-09 22:08:56.977: 2906509 total sequences after non-specified labels fileterd out

Hammock log file

2020-11-09 22:08:56.992: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-09 22:08:56.992: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-09 22:08:57.072: Generating input statistics...
2020-11-09 22:08:57.222: Greedy clustering threshold not set. Setting automatically to: 27
2020-11-09 22:08:57.222: Initial greedy clusters limit not set. Setting automatically to: 1504
2020-11-09 22:08:57.224: Greedy clustering...
2020-11-09 22:10:45.826: Ready. Clustering time: 108602
2020-11-09 22:10:45.827: Resulting clusers: 40062
2020-11-09 22:10:45.827: Building MSAs...
2020-11-09 22:10:47.326: Ready. Total time: 110102
2020-11-09 22:10:47.327: Saving results to output files...
2020-11-09 22:10:48.970: Greedy clustering results in: /home/rstudio/data/HammockLibDNAse/initial_clusters.tsv
2020-11-09 22:10:48.970: and: /home/rstudio/data/HammockLibDNAse/initial_clusters_sequences.tsv
2020-11-09 22:10:48.971: and:
/home/rstudio/data/HammockLibDNAse/initial_clusters_sequences_original_order.tsv
2020-11-09 22:10:48.971:
Loading clusters...
2020-11-09 22:10:49.276: Maximal alignment length not set. Setting automatically to: 32
2020-11-09 22:10:49.289: Minimal number of match states not set. Setting automatically to: 5
2020-11-09 22:10:49.864: Assign threshold sequence not set. Setting automatically to:
2020-11-09 22:10:49.868: 15.3,12.08,8.86,
2020-11-09 22:10:49.869: Overlap threshold not set. Setting automatically to:
2020-11-09 22:10:49.873: 11.28,6.44,0.0,
2020-11-09 22:10:49.873: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-09 22:10:49.878: 16.11,14.5,12.89,
2020-11-09 22:10:50.109: 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

2020-11-09 22:10:50.373:
Clustering in 3 rounds...
2020-11-09 22:10:50.376:
2020-11-09 22:10:50.376: Round 1:

2020-11-09 22:10:50.376: 1997 clusters remaining
2020-11-09 22:10:50.376: Building hmms and searching database...
2020-11-09 22:11:13.750: Extending clusters...
2020-11-09 22:11:13.946: 15752 sequences to be inserted into clusters
2020-11-09 22:11:13.958: 1560 clusters to be extended
2020-11-09 22:11:20.379: 10906 sequences rejected
2020-11-09 22:11:20.464: 4665 cluster pairs to check and merge.

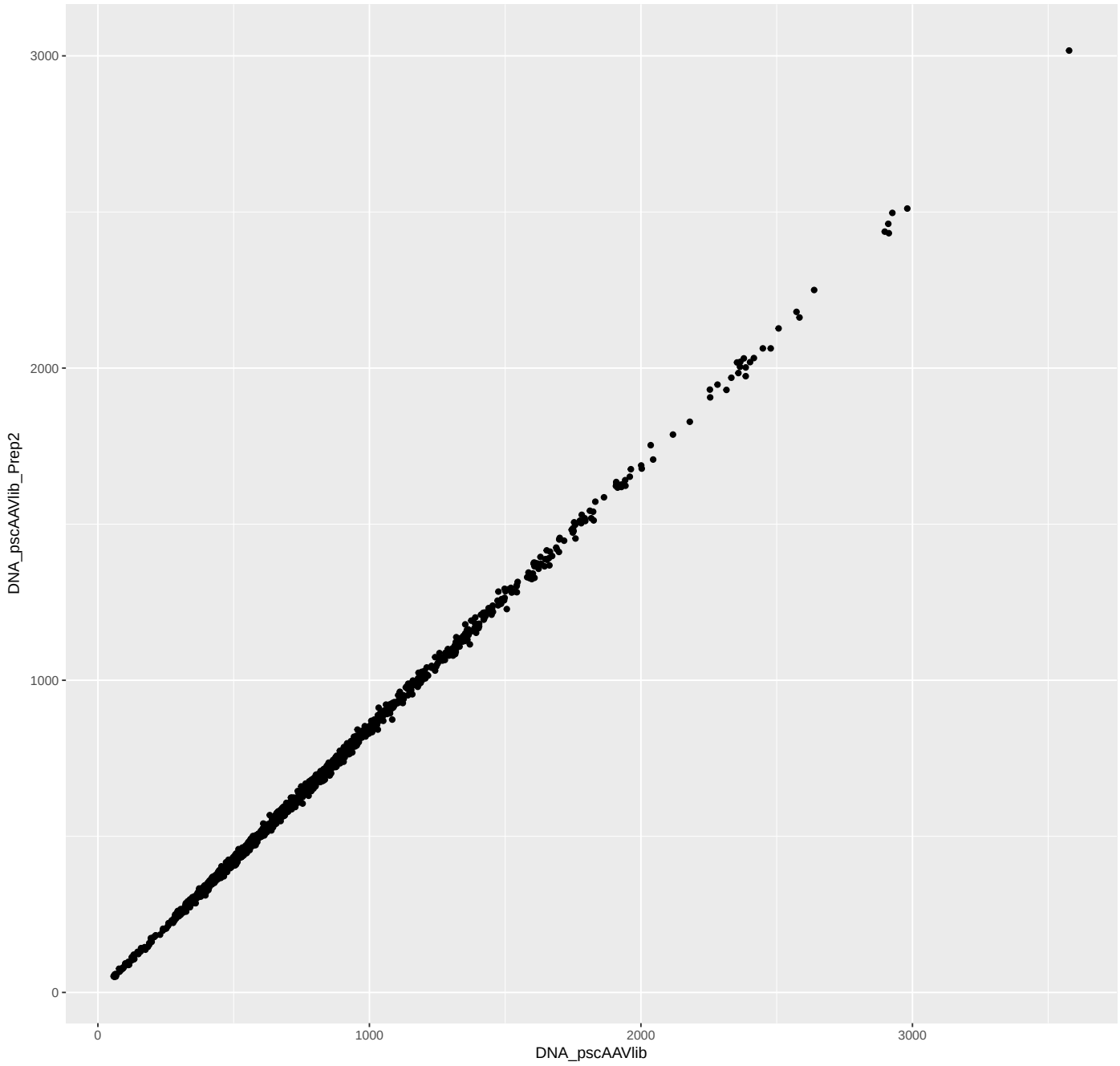
Hammock log file

```
2020-11-09 22:11:20.464: Merging clusters from 84 groups...
2020-11-09 22:11:20.692: Buiding hhs...
2020-11-09 22:11:22.081: HH clustering...
2020-11-09 22:13:15.739:
2020-11-09 22:13:15.740: Round 2:

2020-11-09 22:13:15.740: 1739 clusters remaining
2020-11-09 22:13:15.740: Building hmms and searching database...
2020-11-09 22:13:33.023: Extending clusters...
2020-11-09 22:13:33.185: 12758 sequences to be inserted into clusters
2020-11-09 22:13:33.190: 1313 clusters to be extended
2020-11-09 22:13:40.893: 10118 sequences rejected
2020-11-09 22:13:41.348: 67740 cluster pairs to check and merge.
2020-11-09 22:13:41.349: Merging clusters from 1 groups...
2020-11-09 22:13:41.539: Buiding hhs...
2020-11-09 22:13:42.031: HH clustering...
2020-11-09 22:15:19.555:
2020-11-09 22:15:19.555: Round 3:

2020-11-09 22:15:19.555: 1558 clusters remaining
2020-11-09 22:15:19.556: Building hmms and searching database...
2020-11-09 22:15:36.660: Extending clusters...
2020-11-09 22:15:36.761: 16600 sequences to be inserted into clusters
2020-11-09 22:15:36.767: 1331 clusters to be extended
2020-11-09 22:15:48.405: 11821 sequences rejected
2020-11-09 22:15:48.414: Overlap threshold is 0. Running full cluster merging.
2020-11-09 22:15:48.581: Buiding hhs...
2020-11-09 22:15:49.246: HH clustering...
2020-11-09 22:17:41.044:
Ready. Clustering time : 410671
2020-11-09 22:17:41.044: Resulting clusers: 1340
2020-11-09 22:17:41.045: Containing 34315 unique sequences and 1936731 total sequences.
2020-11-09 22:17:41.065: Unique sequences not assigned: 25864, total sequences not assigned: 969778
2020-11-09 22:17:41.065: Saving results to outupt files...
2020-11-09 22:17:42.213: Results in: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences.tsv
2020-11-09 22:17:42.213: and: /home/rstudio/data/HammockLibDNAse/final_clusters.tsv
2020-11-09 22:17:42.213: and: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences_original_order.tsv
2020-11-09 22:17:42.214:
Calculating KLD...
2020-11-09 22:17:42.215: 21 clusters omitted from KLD calculation because each of them only contains a single unique
sequence.
2020-11-09 22:17:45.393: Final system KLD over match state MSA positions: 21.284280102889714
2020-11-09 22:17:45.393: Final system KLD over all MSA positions: 39.68035578266766
2020-11-09 22:17:45.393: 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
2020-11-09 22:18:07.895:
Hammock version 1.1.1 Run with –help for a brief description of command line parameters.
2020-11-09 22:18:08.028: 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
2020-11-09 22:18:08.029: Loading input sequences...
2020-11-09 22:18:08.446: 60086 unique sequences loaded.
2020-11-09 22:18:08.481: 1535104 total sequences loaded.
2020-11-09 22:18:08.481: 60086 unique sequences after non-specified labels filtered out
2020-11-09 22:18:08.536: 1535104 total sequences after non-specified labels fileterd out

Hammock log file

2020-11-09 22:18:08.551: Shortest sequence: 14 AA. Longest sequence: 22 AA.
2020-11-09 22:18:08.551: More than 10 000 unique sequences. Using greedy clustering. Use –use_clinkage to force clinkage clustering
2020-11-09 22:18:08.635: Generating input statistics...
2020-11-09 22:18:08.802: Greedy clustering threshold not set. Setting automatically to: 27
2020-11-09 22:18:08.802: Initial greedy clusters limit not set. Setting automatically to: 1502
2020-11-09 22:18:08.804: Greedy clustering...
2020-11-09 22:19:57.794: Ready. Clustering time: 108990
2020-11-09 22:19:57.795: Resulting clusers: 39583
2020-11-09 22:19:57.795: Building MSAs...
2020-11-09 22:19:59.273: Ready. Total time: 110469
2020-11-09 22:19:59.274: Saving results to output files...
2020-11-09 22:20:00.828: Greedy clustering results in: /home/rstudio/data/HammockLibDNAse/initial_clusters.tsv
2020-11-09 22:20:00.829: and: /home/rstudio/data/HammockLibDNAse/initial_clusters_sequences.tsv
2020-11-09 22:20:00.829: and:
/home/rstudio/data/HammockLibDNAse/initial_clusters_sequences_original_order.tsv
2020-11-09 22:20:00.829:
Loading clusters...
2020-11-09 22:20:01.070: Maximal alignment length not set. Setting automatically to: 32
2020-11-09 22:20:01.081: Minimal number of match states not set. Setting automatically to: 5
2020-11-09 22:20:01.485: Assign threshold sequence not set. Setting automatically to:
2020-11-09 22:20:01.490: 15.3,12.08,8.86,
2020-11-09 22:20:01.490: Overlap threshold not set. Setting automatically to:
2020-11-09 22:20:01.494: 11.27,6.44,0.0,
2020-11-09 22:20:01.495: Merge threshold not set. Setting automatically based on average sequence length to:
2020-11-09 22:20:01.499: 16.1,14.49,12.88,
2020-11-09 22:20:01.707: 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

2020-11-09 22:20:01.977:
Clustering in 3 rounds...
2020-11-09 22:20:01.979:
2020-11-09 22:20:01.980: Round 1:

2020-11-09 22:20:01.980: 1994 clusters remaining
2020-11-09 22:20:01.980: Building hmms and searching database...
2020-11-09 22:20:25.155: Extending clusters...
2020-11-09 22:20:25.383: 16358 sequences to be inserted into clusters
2020-11-09 22:20:25.396: 1568 clusters to be extended
2020-11-09 22:20:31.946: 11310 sequences rejected
2020-11-09 22:20:32.036: 4703 cluster pairs to check and merge.

Hammock log file

2020-11-09 22:20:32.036: Merging clusters from 84 groups...
2020-11-09 22:20:32.257: Buiding hhs...
2020-11-09 22:20:33.650: HH clustering...
2020-11-09 22:22:29.889:
2020-11-09 22:22:29.890: Round 2:

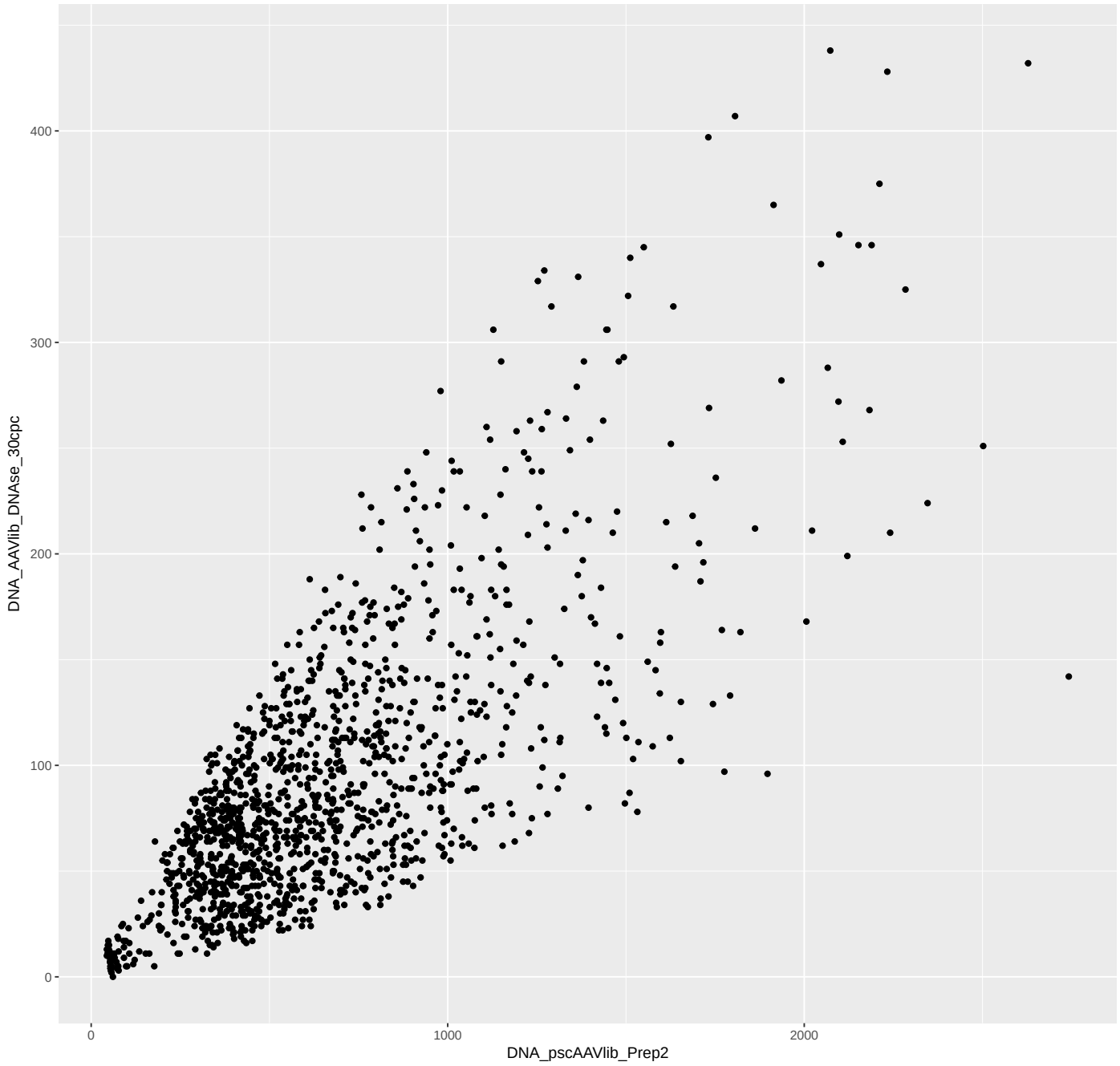
2020-11-09 22:22:29.890: 1735 clusters remaining
2020-11-09 22:22:29.890: Building hmms and searching database...
2020-11-09 22:22:46.849: Extending clusters...
2020-11-09 22:22:47.007: 12559 sequences to be inserted into clusters
2020-11-09 22:22:47.013: 1300 clusters to be extended
2020-11-09 22:22:57.721: 10059 sequences rejected
2020-11-09 22:22:58.136: 72128 cluster pairs to check and merge.
2020-11-09 22:22:58.136: Merging clusters from 1 groups...
2020-11-09 22:22:58.360: Buiding hhs...
2020-11-09 22:22:58.819: HH clustering...
2020-11-09 22:24:29.278:
2020-11-09 22:24:29.279: Round 3:

2020-11-09 22:24:29.280: 1571 clusters remaining
2020-11-09 22:24:29.280: Building hmms and searching database...
2020-11-09 22:24:46.379: Extending clusters...
2020-11-09 22:24:46.475: 16562 sequences to be inserted into clusters
2020-11-09 22:24:46.480: 1360 clusters to be extended
2020-11-09 22:24:57.290: 11873 sequences rejected
2020-11-09 22:24:57.295: Overlap threshold is 0. Running full cluster merging.
2020-11-09 22:24:57.459: Buiding hhs...
2020-11-09 22:24:57.854: HH clustering...
2020-11-09 22:26:54.059:
Ready. Clustering time : 412082
2020-11-09 22:26:54.060: Resulting clusers: 1345
2020-11-09 22:26:54.060: Containing 34643 unique sequences and 1030396 total sequences.
2020-11-09 22:26:54.094: Unique sequences not assigned: 25443, total sequences not assigned: 504708
2020-11-09 22:26:54.094: Saving results to outupt files...
2020-11-09 22:26:55.318: Results in: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences.tsv
2020-11-09 22:26:55.318: and: /home/rstudio/data/HammockLibDNAse/final_clusters.tsv
2020-11-09 22:26:55.319: and: /home/rstudio/data/HammockLibDNAse/final_clusters_sequences_original_order.tsv
2020-11-09 22:26:55.319:
Calculating KLD...
2020-11-09 22:26:55.320: 31 clusters omitted from KLD calculation because each of them only contains a single unique sequence.
2020-11-09 22:26:58.397: Final system KLD over match state MSA positions: 21.314423213245473
2020-11-09 22:26:58.397: Final system KLD over all MSA positions: 39.98132228244945
2020-11-09 22:26:58.398: 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 30.07174 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     2020-11-09
```

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
```