Supplementary Experimental Data File

Selcen Ari
Alper Yilmaz

22 09 2019

Arrangement of CLASH dataset (Helwak et al. 2013)

CLASH dataset was retrieved from PubMed

```
clashelwak <- read.table("mmc1.txt", comment.char = "#",
    header = TRUE, skip = 1)

# hg19
```

Query of Human Genome 19.

```
# HG19
listEnsemblArchives()
listMarts(host = "http://grch37.ensembl.org")
ensemblgrch37 = useMart(host = "http://grch37.ensembl.org",
    biomart = "ENSEMBL_MART_ENSEMBL", dataset = "hsapiens_gene_ensembl")
hg19 <- getBM(attributes = c("ensembl_transcript_id",
    "ensembl_gene_id", "chromosome_name", "start_position",
    "end_position", "hgnc_symbol", "entrezgene_id",
    "strand"), mart = ensemblgrch37)
```

Adding miRNA and gene information

```
clashelwak <- clashelwak %>% separate(microRNA_name,
    c("Barcode", "Database", "mirna_name", "type"),
    sep = "_") %>% separate(mRNA_name, c("Ensembl_Gene_Id",
    "Ensembl_Transcript_Id", "Hugo_Symbol", "mRNA_Type"),
    sep = "_")
```

MiRNA releases are obtained from miRBase. In this step, release 21 (in Human genome 38) is utilised.

```
mirbasehg38 <- read.table("mirbasehg38.txt", comment.char = "#") %>%
    filter(V3 != "miRNA_primary_transcript") %>% separate(V9,
    c("ID", "Alias", "Name", "Precusor"), sep = ";") %>%
    mutate(ID = substr(ID, 4, length(ID)), Alias = substr(Alias,
        7, length(Alias)), Name = substr(Name, 6, length(Name)),
        Precusor = substr(Precusor, 14, length(Precusor))) %>%
    dplyr::select(chr = V1, start = V4, end = V5, strand = V7,
        ID, Alias, Name, Precusor)
```

CLASH dataset is published in miRBase erlease 15 and Human Genome 19 version.

```
clashelwakfinal <- read_tsv("mirna_mature.txt", col_names = FALSE) %>%
    filter(startsWith(X2, "hsa")) %>% dplyr::select(mirna_ID = X2,
    mirbase_ID = X3) %>% inner_join(mirbasehg38 %>%
    dplyr::select(ID, Name), by = c(mirbase_ID = "ID")) %>%
    dplyr::select(mirbase_ID, Name) %>% distinct() %>%
    inner_join(clashelwak, by = c(mirbase_ID = "Barcode")) %>%
    dplyr::select(Name, miRNA_seq, Ensembl_Gene_Id,
        Ensembl_Transcript_Id, Hugo_Symbol, mRNA_seq_extended,
        chimeras_decompressed, seed_type, seed_basepairs,
        folding_class, seq_ID, folding_energy, X5.UTR,
        CDS, X3.UTR) %>% inner_join(hg19, by = c(Ensembl_Gene_Id = "ensembl_gene_id",
    Ensembl_Transcript_Id = "ensembl_transcript_id",
    Hugo_Symbol = "hgnc_symbol")) %>% mutate(region1 = ifelse(X5.UTR ==
    "1", "5UTR", " "), region2 = ifelse(X3.UTR == "1",
    "3UTR", " "), region3 = ifelse(CDS == "1", "CDS",
    " ")) %>% unite(region, c(region1, region2, region3),
    sep = "||") %>% dplyr::select(chromosome_name,
    start_position, end_position, strand, Hugo_Symbol,
    Ensembl_Gene_Id, Ensembl_Transcript_Id, mRNA_seq_extended,
    Name, miRNA_seq, seq_ID, seed_type, seed_basepairs,
    folding_class, folding_energy, region) %>% as_tibble()
```

Converting CLASH data to human genome 38 build.

```
lift19 <- clashelwakfinal %>% dplyr::select(1, 2, 3) %>%
    unite(start_end, c("start_position", "end_position"),
        sep = "-") %>% mutate(Chromosome = paste0("chr",
    chromosome_name, "")) %>% unite(chromosome_name,
    c("Chromosome", "start_end"), sep = ":")

write_tsv(lift19, "lift19.txt")

# Lift over process is made via UCSC liftover tool.
# (https://genome.ucsc.edu/cgi-bin/hgLiftOver)

lift19_del <- read_tsv("deleted_lift19.txt")
colnames(lift19_del)[1] <- "chromosome_loc"

lift19_del <- lift19_del %>% dplyr::filter(startsWith(chromosome_loc,
    "chr")) %>% separate(chromosome_loc, c("Chr", "End"),
    "-", remove = TRUE) %>% separate(Chr, c("Chr",
    "Start"), ":", remove = TRUE)

lift19_del$Start <- as.numeric(lift19_del$Start)
lift19_del$End <- as.numeric(lift19_del$End)

clashelwakfinal <- clashelwakfinal %>% mutate(Chromosome = paste0("chr",
    chromosome_name, "")) %>% dplyr::anti_join(lift19_del,
    by = c(Chromosome = "Chr", start_position = "Start",
        end_position = "End"))
```

```
hg38clash <- read.delim("hg38clashcomp.txt", header = FALSE)

clashelwakfinal <- clashelwakfinal %>% bind_cols(hg38clash)

colnames(clashelwakfinal)[18] <- "HG38build_loc"

clashelwakfinal <- clashelwakfinal %>% dplyr::mutate(Genom_build = rep("hg19"))
str(clashelwakfinal)


# Arrangement in dataset

clashelwakfinal <- clashelwakfinal %>% dplyr::select(cluster = seq_ID,
    chromosome = Chromosome, start_position, end_position,
    strand, hgnc_symbol = Hugo_Symbol, Ensembl_Gene_Id,
    Ensembl_Transcript_Id, target_seq = mRNA_seq_extended,
    miRNA = Name, miR_seq = miRNA_seq, seed_type, seed_type2 = seed_basepairs,
    seed_type3 = folding_class, Energy = folding_energy,
    HG38build_loc, Genom_build, region)

clashelwakfinal$cluster <- as.character(clashelwakfinal$cluster)
clashelwakfinal$strand <- as.character(clashelwakfinal$strand)
clashelwakfinal$target_seq <- as.character(clashelwakfinal$target_seq)
clashelwakfinal$miR_seq <- as.character(clashelwakfinal$miR_seq)
clashelwakfinal$seed_type <- as.character(clashelwakfinal$seed_type)
clashelwakfinal$HG38build_loc <- as.character(clashelwakfinal$HG38build_loc)
clashelwakfinal$seed_type2 <- as.numeric(clashelwakfinal$seed_type2)
clashelwakfinal$seed_type3 <- as.character(clashelwakfinal$seed_type3)
```

Interpreting the CLASH seed structures in dataset

```
clashelwakfinal <- clashelwakfinal %>% mutate(seed_type = ifelse(seed_type ==
    "noncanonical_seed" & seed_type2 > 4 & seed_type3 ==
    "I", paste0(seed_type2, "-mer"), seed_type), seed_type = ifelse(seed_type ==
    "noncanonical_seed" & seed_type2 > 4 & seed_type3 ==
    "II", paste0(seed_type2, "-mer_noncanonical"),
    seed_type), seed_type = ifelse(seed_type == "noncanonical_seed" &
    seed_type2 > 4 & seed_type3 == "III", paste0(seed_type2,
    "-mer_noncanonical"), seed_type), seed_type = ifelse(seed_type ==
    "noncanonical_seed" & seed_type2 > 4 & seed_type3 ==
    "IV", paste0(seed_type2, "-mer_noncanonical"),
    seed_type), seed_type = ifelse(startsWith(seed_type,
    "no"), "none", seed_type)) %>% dplyr::select(-seed_type2,
    -seed_type3)
```

Arrangement of CLEAR-CLiP Dataset (Moore et al. 2015)

CLASH dataset was retrieved from Nature web page

```
clearclip <- read_xlsx("CLEAR-CLIP.xlsx")

# Clearclip hg18
```

Query of Human Genome 18

```
# HG18
listEnsemblArchives()
listMarts(host = "may2009.archive.ensembl.org")
ensembl54 = useMart(host = "may2009.archive.ensembl.org",
    biomart = "ENSEMBL_MART_ENSEMBL", dataset = "hsapiens_gene_ensembl")

hg18 <- getBM(attributes = c("ensembl_transcript_id",
    "ensembl_gene_id", "chromosome_name", "start_position",
    "end_position", "hgnc_symbol", "entrezgene", "strand"),
    mart = ensembl54)
```

Adding Genome Information to dataset

```
clearclipfinal <- hg18 %>% inner_join(clearclip, by = c(entrezgene = "gene.id",
    hgnc_symbol = "gene.symbol")) %>% distinct()
```

Converting human genome build

```
lift18 <- clearclipfinal %>% unite(start_end, c("start_position",
    "end_position"), sep = "-") %>% unite(location,
    c("chr", "start_end"), sep = ":") %>% dplyr::select(location)

write_tsv(lift18, "lift18.txt")

deleted_lift18 <- read_tsv("deleted_lift18.txt")

colnames(deleted_lift18)[1] <- "Chromosome_loc"

deleted_lift18 <- deleted_lift18 %>% dplyr::filter(startsWith(Chromosome_loc,
    "chr")) %>% separate(Chromosome_loc, c("Chr", "End"),
    "-", remove = TRUE) %>% separate(Chr, c("Chr",
    "Start"), ":", remove = TRUE)

deleted_lift18$Start <- as.numeric(deleted_lift18$Start)
deleted_lift18$End <- as.numeric(deleted_lift18$End)

clearclipfinal <- clearclipfinal %>% dplyr::anti_join(deleted_lift18,
    by = c(chr = "Chr", start_position = "Start", end_position = "End"))

hg38clearclip <- read.delim("hg38clearclip.txt", header = FALSE)
```

```
clearclipfinal <- clearclipfinal %>% bind_cols(hg38clearclip)

colnames(clearclipfinal)[28] <- "HG38build_loc"

clearclipfinal <- clearclipfinal %>% dplyr::mutate(Genom_build = rep("hg18"))

clearclipfinal <- clearclipfinal %>% dplyr::select(cluster = cluster.ID,
    chromosome = chr, start_position, end_position,
    strand = strand.y, hgnc_symbol, Ensembl_Gene_Id = ensembl_gene_id,
    Ensembl_Transcript_Id = ensembl_transcript_id,
    target_seq = target.map, miRNA, miR_seq = miR.map,
    seed_type = "seed match", Energy = MFE, HG38build_loc,
    Genom_build, region)
```

Seed type manipulation

```
clipdata_seed <- data_frame(seed_type = c("5mer_1",
    "5mer_2", "5mer_3", "6mer", "6mer.indel", "6mer.mm",
    "6mer_off.mm", "6merA1", "6merA1.indel", "6merA1.mm",
    "7merA1", "7merA1.indel", "7merA1.mm", "7merm8",
    "7merm8.indel", "7merm8,mm", "8mer", "8mer.indel",
    "8mer.mm", "NA"), seed_type_com = c("5-mer", "5-mer_noncanonical",
    "5-mer_noncanonical", "6-mer", "6-mer_noncanonical",
    "6-mer_noncanonical", "6-mer_noncanonical", "6-merA1",
    "6-merA1_noncanonical", "6-merA1_noncanonical",
    "7-merA1", "7-merA1_noncanonical", "7-merA1_noncanonical",
    "7-mer-8m", "7-mer-8m_noncanonical", "7-mer-8m_noncanonical",
    "8-mer", "8-mer_noncanonical", "8-mer_noncanonical",
    "none"))

clearclipfinal <- clearclipfinal %>% inner_join(clipdata_seed,
    by = "seed_type") %>% dplyr::select(1:11, seed_type = seed_type_com,
    Energy, HG38build_loc, Genom_build, region)


clearclipfinal$HG38build_loc <- as.character(clearclipfinal$HG38build_loc)
```

Integration of two experimental dataset

```
experimentalmirnagene <- bind_rows(clashelwakfinal,
    clearclipfinal) %>% distinct()
```

Adding Coefficients of Interaction factors

Numeric valuesof interaction factors are shown at Table S2-3 in Supplementary Tables file.

```
experimentalmirnagene <- experimentalmirnagene %>%
    mutate(region2 = str_replace_all(region, "NA",
        ""), region3 = str_replace_all(region2, "\\|",
        ""), region = str_replace_all(region3, c(`3'UTR` = "3UTR",
        `5'UTR` = "5UTR"))) %>% dplyr::select(-region2,
    -region3) %>% mutate(region_effect = as.double(ifelse(region %in%
    c("3UTRCDS", "CDS3UTR", "5UTR3UTR", "CDS5UTR3UTR",
        "CDS3UTRintron"), "0.93", ifelse(region %in%
    c("CDS", "CDSintron"), "0.42", ifelse(region %in%
    c("3UTR", "3UTRintron"), "0.84", ifelse(region %in%
    c("5UTR", "5UTRintron"), "0.01", ifelse(region %in%
    c("5UTRCDS", "CDS5UTR"), "0.42", ifelse(region %in%
    c("intron", ""), "0.01", ifelse(region %in% c("exon_unclassified",
    ""), "0.2", NA)))))))))
```

```
seed_type_effect <- data_frame(seed_type = c("5-mer",
    "5-mer_noncanonical", "6-mer", "6-mer_noncanonical",
    "6-merA1", "6-merA1_noncanonical", "7-mer", "7-mer_noncanonical",
    "7-merA1", "7-merA1_noncanonical", "7-mer-8m",
    "7-mer-8m_noncanonical", "8-mer", "8-mer_noncanonical",
    "9-mer", "9-mer_noncanonical", "none"), seed_type_effect = c(0.05,
    0.04, 0.07, 0.05, 0.07, 0.05, 0.23, 0.19, 0.19,
    0.16, 0.25, 0.21, 0.43, 0.35, 0.43, 0.35, 0.01))

experimentalmirnagene <- experimentalmirnagene %>%
    inner_join(seed_type_effect, by = "seed_type")
```

```
## Saving dataset
saveRDS(experimentalmirnagene, "data/experimentalmirnagene.RDS")
```

```
experimentalmirnagene <- readRDS("data/experimentalmirnagene.RDS")
experimentalmirnagene
```

```
## # A tibble: 45,340 x 18
##    cluster chromosome start_position end_position strand hgnc_symbol
##    <chr>   <chr>              <int>        <int> <chr>  <chr>
##  1 0727A-~ chr5           162864575    162873157 1      CCNG1
##  2 L1HS-1~ chr14           95552565     95624347 -1     DICER1
##  3 L2HS-8~ chr6           109307640    109416022 -1     SESN1
##  4 L2HS-1~ chr5            36876861     37066515 1      NIPBL
##  5 L2-407~ chr4           106603784    106817143 -1     INTS12
##  6 L1HS-7~ chr5           130977407    131132710 -1     FNIP1
##  7 L1HS-4~ chr11          134123389    134135749 1      ACAD8
##  8 0727A-~ chr15           59397277     59417244 1      CCNB2
##  9 L2HS-1~ chr19           37001597     37019562 -1     ZNF260
## 10 L2HS-9~ chr11           64889252     64902004 -1     SYVN1
## # ... with 45,330 more rows, and 12 more variables: Ensembl_Gene_Id <chr>,
## #   Ensembl_Transcript_Id <chr>, target_seq <chr>, miRNA <chr>,
## #   miR_seq <chr>, seed_type <chr>, Energy <dbl>, HG38build_loc <chr>,
## #   Genom_build <chr>, region <chr>, region_effect <dbl>,
## #   seed_type_effect <dbl>
```

The context of dataset is shown at Table S5 in Supplementary Tables.

REFERENCES

Helwak, Aleksandra, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. 2013. "Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding." *Cell* 153 (3): 654–65. https://doi.org/10.1016/j.cell.2013.03.043.

Moore, Michael J., Troels K. H. Scheel, Joseph M. Luna, Christopher Y. Park, John J. Fak, Eiko Nishiuchi, Charles M. Rice, and Robert B. Darnell. 2015. "miRNA-Target Chimeras Reveal miRNA 3'-End Pairing as a Major Determinant of Argonaute Target Specificity." *Nature Communications* 6 (November): 8864. https://doi.org/10.1038/ncomms9864.