Supplementary Experimental Data File

Selcen Arı

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_positi
```

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.

```
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
  dplyr::select(chr= V1, start = V4, end = V5, strand = V7, ID, Alias, Name, Precusor)->mirbasehg38
```

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

```
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
  inner_join(hg19, by = c("Ensembl_Gene_Id"= "ensembl_gene_id", "Ensembl_Transcript_Id"="ensembl_transc
  mutate(region1 = ifelse(X5.UTR == "1", "5UTR", " "), region2= ifelse(X3.UTR == "1", "3UTR", " "),regi
  unite(region, c(region1, region2, region3), sep = "||")%>%
  dplyr::select(chromosome_name, start_position, end_position,strand, Hugo_Symbol, Ensembl_Gene_Id, Ens
  as_tibble() -> clashelwakfinal
```

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%>% dplyr::select(cluster=seq_ID, chromosome = Chromosome, start_position, end_position,

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

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_positi
```

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%>%
  dplyr::select(cluster=cluster.ID, chromosome = chr, start_position, end_position, strand = strand.y, l
```

Seed type manipulation

```
data_frame(seed_type= c("5mer_1","5mer_2", "5mer_3", "6mer", "6mer.indel", "6mer.mm", "6mer_off.mm", "6m
           seed_type_com= c("5-mer", "5-mer_noncanonical", "5-mer_noncanonical", "6-mer", "6-mer_noncano

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


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

Integration of two experimental dataset

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

Adding Coefficients of Interaction factors

```
experimentalmirnagene <- experimentalmirnagene%>%
  mutate(region2  = str_replace_all(region, "NA", ""), region3 = str_replace_all(region2, "\\|", ""), re
  dplyr::select(-region2, -region3)%>%
  mutate(region_effect = as.double(ifelse(region %in% c("3UTRCDS", "CDS3UTR", "5UTR3UTR", "CDS5UTR3UTR"
```

```
seed_type_effect <- data_frame( seed_type = c("5-mer", "5-mer_noncanonical", "6-mer", "6-mer_noncanonica
                                seed_type_effect= c(0.05, 0.04, 0.07, 0.05, 0.07, 0.05, 0.23, 0.19, 0.1
                                )

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

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

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

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.