3UTR analysis for NGCs

Huayu

May 2, 2018

Human genome annotation data (Ensembl)

```
source("https://bioconductor.org/biocLite.R")

biocLite(suppressUpdates = TRUE)


#if (!require("ensembldb")) biocLite("ensembldb")
#if (!require("EnsDb.Hsapiens.v75")) biocLite("EnsDb.Hsapiens.v75")
#if (!require("Gviz")) biocLite("Gviz")
#if (!require("GenomicFeatures")) biocLite("GenomicFeatures")

library(ensembldb)
library(EnsDb.Hsapiens.v75)
library(Gviz)
library(GenomicFeatures)
```

```
ensdb <- EnsDb.Hsapiens.v75
ensdb
```

```
## EnsDb for Ensembl:
## |Backend: SQLite
## |Db type: EnsDb
## |Type of Gene ID: Ensembl Gene ID
## |Supporting package: ensembldb
## |Db created by: ensembldb package from Bioconductor
## |script_version: 0.2.3
## |Creation time: Tue Nov 15 23:35:19 2016
## |ensembl_version: 75
## |ensembl_host: localhost
## |Organism: homo_sapiens
## |genome_build: GRCh37
## |DBSCHEMAVERSION: 1.0
## | No. of genes: 64102.
## | No. of transcripts: 215647.
## |Protein data available.
```

```
columns(ensdb)
```

```
##  [1] "ENTREZID"            "EXONID"              "EXONIDX"
##  [4] "EXONSEQEND"          "EXONSEQSTART"        "GENEBIOTYPE"
##  [7] "GENEID"              "GENENAME"            "GENESEQEND"
## [10] "GENESEQSTART"        "INTERPROACCESSION"   "ISCIRCULAR"
## [13] "PROTDOMEND"          "PROTDOMSTART"        "PROTEINDOMAINID"
## [16] "PROTEINDOMAINSOURCE" "PROTEINID"           "PROTEINSEQUENCE"
## [19] "SEQCOORDSYSTEM"      "SEQLENGTH"           "SEQNAME"
## [22] "SEQSTRAND"           "SYMBOL"              "TXBIOTYPE"
```

```
## [25] "TXCDSSEQEND"         "TXCDSSEQSTART"       "TXID"
## [28] "TXNAME"              "TXSEQEND"            "TXSEQSTART"
## [31] "UNIPROTDB"           "UNIPROTID"           "UNIPROTMAPPINGTYPE"
```

Get 3'UTR annotations

```
hg3UTR <- threeUTRsByTranscript(ensdb)
hgGene <- genes(ensdb)
head(hg3UTR)
```

```
## GRangesList object of length 6:
## $ENST00000000233
## GRanges object with 1 range and 2 metadata columns:
##       seqnames                ranges strand |         exon_id exon_rank
##          <Rle>             <IRanges>  <Rle> |     <character> <integer>
##   [1]        7 [127231354, 127231759]      + | ENSE00000882271         6
##
## $ENST00000000412
## GRanges object with 1 range and 2 metadata columns:
##       seqnames             ranges strand |         exon_id exon_rank
##   [1]       12 [9092961, 9094413]      - | ENSE00002254457         7
##
## $ENST00000000442
## GRanges object with 1 range and 2 metadata columns:
##       seqnames               ranges strand |         exon_id exon_rank
##   [1]       11 [64083439, 64084210]      + | ENSE00001271942         7
##
## ...
## <3 more elements>
## -------
## seqinfo: 223 sequences from GRCh37 genome
```

```
head(hgGene)
```

```
## GRanges object with 6 ranges and 6 metadata columns:
##                    seqnames         ranges strand |         gene_id
##                       <Rle>      <IRanges>  <Rle> |     <character>
##   ENSG00000223972        1 [11869, 14412]      + | ENSG00000223972
##   ENSG00000227232        1 [14363, 29806]      - | ENSG00000227232
##   ENSG00000243485        1 [29554, 31109]      + | ENSG00000243485
##   ENSG00000237613        1 [34554, 36081]      - | ENSG00000237613
##   ENSG00000268020        1 [52473, 54936]      + | ENSG00000268020
##   ENSG00000240361        1 [62948, 63887]      + | ENSG00000240361
##                     gene_name                                entrezid
##                   <character>                             <character>
##   ENSG00000223972     DDX11L1                      100287596;100287102
##   ENSG00000227232      WASH7P                         653635;100287171
##   ENSG00000243485  MIR1302-10 100422831;100302278;100422834;100422919
##   ENSG00000237613     FAM138A                     641702;654835;645520
##   ENSG00000268020      OR4G4P
##   ENSG00000240361     OR4G11P
##                   gene_biotype seq_coord_system      symbol
##                    <character>      <character> <character>
```

```
##   ENSG00000223972   pseudogene       chromosome     DDX11L1
##   ENSG00000227232   pseudogene       chromosome      WASH7P
##   ENSG00000243485      lincRNA       chromosome  MIR1302-10
##   ENSG00000237613      lincRNA       chromosome     FAM138A
##   ENSG00000268020   pseudogene       chromosome      OR4G4P
##   ENSG00000240361   pseudogene       chromosome     OR4G11P
##   -------
##   seqinfo: 273 sequences from GRCh37 genome
```

Get exon infos of all NGCs

```
load("NGCCors.Rdata")
ngcExonDb <- select(x = ensdb, keys = NGCCors$SYMBOL, columns = c("ENTREZID","GENEID", "TXID", "EXONID"
head(ngcExonDb)
```

```
##   ENTREZID          GENEID            TXID          EXONID EXONIDX SYMBOL
## 1     9423 ENSG00000065320 ENST00000173229 ENSE00001284677       1   NTN1
## 2     9423 ENSG00000065320 ENST00000173229 ENSE00001284696       2   NTN1
## 3     9423 ENSG00000065320 ENST00000173229 ENSE00001104530       3   NTN1
## 4     9423 ENSG00000065320 ENST00000173229 ENSE00001104535       4   NTN1
## 5     9423 ENSG00000065320 ENST00000173229 ENSE00001104539       5   NTN1
## 6     9423 ENSG00000065320 ENST00000173229 ENSE00000423960       6   NTN1
```

Extract from all 3'UTRs, ones for NGCs

```
ngc3UTR <- hg3UTR[names(hg3UTR) %in% unique(ngcExonDb$TXID)]
ngc3UTRUl <- unlist(ngc3UTR)
head(ngc3UTR)
```

```
## GRangesList object of length 6:
## $ENST00000002829
## GRanges object with 1 range and 2 metadata columns:
##       seqnames              ranges strand |         exon_id exon_rank
##          <Rle>           <IRanges>  <Rle> |     <character> <integer>
##   [1]        3 [50225549, 50226508]      + | ENSE00001911603        19
##
## $ENST00000166244
## GRanges object with 1 range and 2 metadata columns:
##       seqnames              ranges strand |         exon_id exon_rank
##   [1]        1 [22928235, 22930087]      + | ENSE00001156943        17
##
## $ENST00000173229
## GRanges object with 1 range and 2 metadata columns:
##       seqnames             ranges strand |         exon_id exon_rank
##   [1]       17 [9143286, 9147317]      + | ENSE00001126297         7
##
## ...
## <3 more elements>
## -------
## seqinfo: 223 sequences from GRCh37 genome
```

```
head(ngc3UTRUl)
```

```
## GRanges object with 6 ranges and 2 metadata columns:
##                    seqnames                ranges strand |         exon_id
##                       <Rle>             <IRanges>  <Rle> |     <character>
##   ENST00000002829        3 [ 50225549,  50226508]      + | ENSE00001911603
##   ENST00000166244        1 [ 22928235,  22930087]      + | ENSE00001156943
##   ENST00000173229       17 [  9143286,   9147317]      + | ENSE00001126297
##   ENST00000195173        3 [122629687, 122629829]      - | ENSE00003515311
##   ENST00000195173        3 [122628043, 122629148]      - | ENSE00002319897
##   ENST00000204961        X [ 68060498,  68061990]      + | ENSE00001041114
##                    exon_rank
##                    <integer>
##   ENST00000002829        19
##   ENST00000166244        17
##   ENST00000173229         7
##   ENST00000195173        22
##   ENST00000195173        23
##   ENST00000204961         5
##   -------
##   seqinfo: 223 sequences from GRCh37 genome
```

Match genenames to 3'UTR exons in NGC

```
head(values(ngc3UTRUl)$exon_id)
```

```
## [1] "ENSE00001911603" "ENSE00001156943" "ENSE00001126297" "ENSE00003515311"
## [5] "ENSE00002319897" "ENSE00001041114"
```

```
ngcSYMEXO <- unique(ngcExonDb[,c("SYMBOL", "EXONID")])

head(ngcSYMEXO)
```

```
##   SYMBOL          EXONID
## 1   NTN1 ENSE00001284677
## 2   NTN1 ENSE00001284696
## 3   NTN1 ENSE00001104530
## 4   NTN1 ENSE00001104535
## 5   NTN1 ENSE00001104539
## 6   NTN1 ENSE00000423960
```

```
ngc3UTRUl$SYMBOL <- ngcSYMEXO$SYMBOL[match(values(ngc3UTRUl)$exon_id, ngcSYMEXO$EXONID)]

head(ngc3UTRUl)
```

```
## GRanges object with 6 ranges and 3 metadata columns:
##                    seqnames                ranges strand |         exon_id
##                       <Rle>             <IRanges>  <Rle> |     <character>
##   ENST00000002829        3 [ 50225549,  50226508]      + | ENSE00001911603
##   ENST00000166244        1 [ 22928235,  22930087]      + | ENSE00001156943
##   ENST00000173229       17 [  9143286,   9147317]      + | ENSE00001126297
##   ENST00000195173        3 [122629687, 122629829]      - | ENSE00003515311
##   ENST00000195173        3 [122628043, 122629148]      - | ENSE00002319897
```

```
##   ENST00000204961        X [ 68060498,  68061990]      + | ENSE00001041114
##                   exon_rank      SYMBOL
##                   <integer> <character>
##   ENST00000002829        19      SEMA3F
##   ENST00000166244        17       EPHA8
##   ENST00000173229         7        NTN1
##   ENST00000195173        22      SEMA5B
##   ENST00000195173        23      SEMA5B
##   ENST00000204961         5       EFNB1
##   -------
##   seqinfo: 223 sequences from GRCh37 genome
```

Search for ACUAA in 3'UTRs of NGCs

Get sequence of 3'UTRs

```
Dna <- getGenomeFaFile(ensdb)
```

```
## snapshotDate(): 2017-04-25
```

```
## require("Rsamtools")
```

```
## loading from cache '//vf-d2-home/d2home$/hzhang/MyDocs/AppData/.AnnotationHub/10878'
##     '//vf-d2-home/d2home$/hzhang/MyDocs/AppData/.AnnotationHub/14664'
```

```
ngc3UTRSeq <- getSeq(Dna, ngc3UTRUl)
head(ngc3UTRSeq)
```

```
##   A DNAStringSet instance of length 6
##     width seq                                               names
## [1]   960 GGCCAGCTGCCTGTGCCTGCC...ACACTGGCTCTGGGACTAGA 3
## [2]  1853 TGTACAGCCAGCAGGGCCCAG...ATAAATTCTGCCTCATCTTT 1
## [3]  4032 CGCCGAGGCAGCGGGCGGGCG...AAATAAACTCTTGTACACTA 17
## [4]   143 CCCTAGCAGTGTACCTGTCTT...ACACACCCATGGAATTCAAG 3
## [5]  1106 ACCCTGAACAAGAATAACTTG...ATTAAAGATGATATCCAGTC 3
## [6]  1493 GTGCCCGGCACGGCCTCAGGC...TGAGAACTAAAAAAAAAAAA X
```

```
ngc3UTRUl$Count_ACTAA <- vcountPattern(pattern = "ACTAA", subject = ngc3UTRSeq)
ngc3UTR_ACTAA <- vmatchPattern(pattern = "ACTAA", subject = ngc3UTRSeq)

ngc3UTR_ACTAA@NAMES <- values(ngc3UTRUl)$SYMBOL

QKIMotif <- DNAStringSet(
                      x = paste("ACTAAY",
                      sapply(X = 1:20, FUN = function(x, string) paste(rep(string, times = x), collapse
                      "TAAY",
                      sep = ""
                      ))

ngc3UTRUl$Count_QKIMotif <- apply(
                      X = sapply(X = QKIMotif, FUN = vcountPattern, subject = ngc3UTRSeq),
                      MARGIN = 1,
                      FUN = sum)
```

```
head(ngc3UTRUl)
```

```
## GRanges object with 6 ranges and 5 metadata columns:
##                    seqnames                ranges strand |         exon_id
##                       <Rle>             <IRanges>  <Rle> |     <character>
##   ENST00000002829        3 [ 50225549,  50226508]      + | ENSE00001911603
##   ENST00000166244        1 [ 22928235,  22930087]      + | ENSE00001156943
##   ENST00000173229       17 [  9143286,   9147317]      + | ENSE00001126297
##   ENST00000195173        3 [122629687, 122629829]      - | ENSE00003515311
##   ENST00000195173        3 [122628043, 122629148]      - | ENSE00002319897
##   ENST00000204961        X [ 68060498,  68061990]      + | ENSE00001041114
##                    exon_rank      SYMBOL Count_ACTAA Count_QKIMotif
##                    <integer> <character>   <integer>      <integer>
##   ENST00000002829        19      SEMA3F           0              0
##   ENST00000166244        17       EPHA8           1              0
##   ENST00000173229         7        NTN1           2              0
##   ENST00000195173        22      SEMA5B           0              0
##   ENST00000195173        23      SEMA5B           0              0
##   ENST00000204961         5       EFNB1           1              0
##   -------
##   seqinfo: 223 sequences from GRCh37 genome
```

```
head(ngc3UTR_ACTAA)
```

```
## MIndex object of length 6
## $SEMA3F
## IRanges object with 0 ranges and 0 metadata columns:
##        start       end     width
##    <integer> <integer> <integer>
##
## $EPHA8
## IRanges object with 1 range and 0 metadata columns:
##           start       end     width
##       <integer> <integer> <integer>
##   [1]      1040      1044         5
##
## $NTN1
## IRanges object with 2 ranges and 0 metadata columns:
##           start       end     width
##       <integer> <integer> <integer>
##   [1]      3839      3843         5
##   [2]      3994      3998         5
##
## ...
## <3 more elements>
```

Is fold change correlated with number of ACUAAs in 3'UTR

```
load("QKINGCSummary.Rdata")

ngc3UTRUl$logFC <- SummaryList$logFC$humanMPData_Array[match(ngc3UTRUl$SYMBOL, SummaryList$logFC$SYMBOL
```

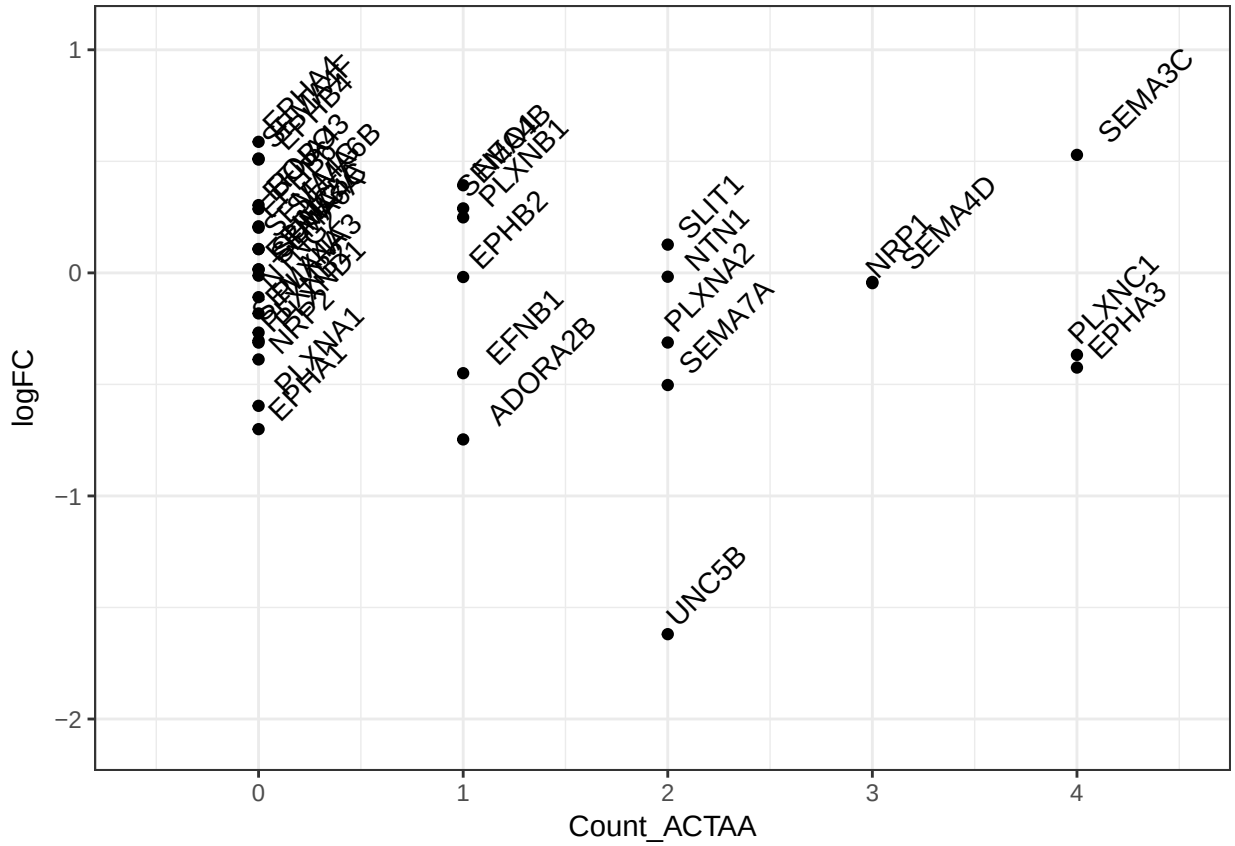
```
ngc3UTRUlDf <- as.data.frame(ngc3UTRUl[!duplicated(names(ngc3UTRUl))])

ngc3UTRUlDf <- ngc3UTRUlDf[ngc3UTRUlDf$SYMBOL != "QKI",]

ngc3UTRUlDf <- ngc3UTRUlDf[!is.na(ngc3UTRUlDf$logFC),]
ngc3UTRUlDf <- ngc3UTRUlDf[!duplicated(ngc3UTRUlDf$SYMBOL),]



library(ggplot2)
ggplot(data = ngc3UTRUlDf, aes(x = Count_ACTAA, y = logFC)) +
  geom_point() +
  geom_text(aes(label = SYMBOL), angle = 45, hjust = -0.2, vjust = 0.5, position = position_jitter(widt
  scale_x_continuous(expand = c(0.05, 0.5)) +
  scale_y_continuous(expand = c(0.05, 0.5)) +
  theme_bw()
```




```
ggsave(filename = "ACTAAvsFC.svg", device = "svg", width = 5, height = 5)
```

```
## Warning: package 'gdtools' was built under R version 3.4.3
```

Overlapping with microRNA binding site

Convert QkiMirNgc dataframe to Granges objects

```
QkiMirNgc <- read.delim(file = "QkiMirCombinedData.txt", stringsAsFactors = FALSE)

makeStartEndfromGenoCoord <- function(genome_coordinates)
  {
    mat1 <- do.call(rbind, strsplit(x = genome_coordinates, split = "\\[|:|\\]"))
    mat2 <- do.call(rbind, strsplit(x = mat1[,4], split = "\\-"))
    mat <- cbind(mat1, mat2)
    mat <- as.data.frame(mat[,c(3, 6, 7, 5)])
    colnames(mat) <- c("seqnames", "start", "end", "strand")
    mat
  }


FromGenoCoordtoGranges <- function(df, coord.var)
{
  grangeMat <- makeStartEndfromGenoCoord(df[,coord.var])
  df <- cbind(df,grangeMat)
  gr <- makeGRangesFromDataFrame(df, seqnames.field = "seqnames", start.field = "start", end.field = "en
  values(gr) <- subset(df, select = -c(seqnames, start, end, strand))
  gr
}

QkiMirNgcGr <- FromGenoCoordtoGranges(df = QkiMirNgc, coord.var = "genome_coordinates")

head(QkiMirNgcGr)
```

```
## GRanges object with 6 ranges and 25 metadata columns:
##       seqnames                ranges strand |      SYMBOL  mirna_name
##          <Rle>             <IRanges>  <Rle> | <character> <character>
##   [1]        2 [222290690, 222290710]      - |       EPHA4  hsa-miR-93
##   [2]        2 [222290690, 222290710]      - |       EPHA4  hsa-miR-93
##   [3]        2 [222290690, 222290710]      - |       EPHA4  hsa-miR-93
##   [4]        2 [222290671, 222290692]      - |       EPHA4  hsa-miR-93
##   [5]        2 [222290671, 222290692]      - |       EPHA4  hsa-miR-93
##   [6]        2 [ 74901371,  74901394]      + |      SEMA4F  hsa-miR-93
##       X.mirbase_acc   gene_id transcript_id ext_transcript_id
##         <character> <integer>   <character>       <character>
##   [1]  MIMAT0000093      2043    uc010zln.1          AK290306
##   [2]  MIMAT0000093      2043    uc002vmr.2          BC026327
##   [3]  MIMAT0000093      2043    uc002vmq.2         NM_004438
##   [4]  MIMAT0000093      2043    uc010zln.1          AK290306
##   [5]  MIMAT0000093      2043    uc002vmr.2          BC026327
##   [6]  MIMAT0000093     10505    uc010ysb.1          AK304358
##                 mirna_alignment               alignment
##                     <character>             <character>
##   [1]  gaUGGACGUGCUUGUCGUGAAAc    |||| ||:  |: |||||||
##   [2]  gaUGGACGUGCUUGUCGUGAAAc    |||| ||:  |: |||||||
##   [3]  gaUGGACGUGCUUGUCGUGAAAc    |||| ||:  |: |||||||
##   [4]  gaUGGACGUGCUUGUCGUGAAAc    | :|| | |||| |||||||
```

```
##   [5]  gaUGGACGUGCUUGUCGUGAAAc    | :|| | |||| |||||||
##   [6] gauggacGUGCU-UGUCGUGAAAc       ||: | ||||||||||
##                gene_alignment mirna_start mirna_end gene_start  gene_end
##                   <character>   <integer> <integer>  <integer> <integer>
##   [1]  uuACCU-CAUCCAU-GCACUUUa           2        22         38        58
##   [2]  uuACCU-CAUCCAU-GCACUUUa           2        22        178       198
##   [3]  uuACCU-CAUCCAU-GCACUUUa           2        22         38        58
##   [4]  uuAAUUG-AAGAACUGCACUUUu           2        22         56        77
##   [5]  uuAAUUG-AAGAACUGCACUUUu           2        22        196       217
##   [6] uaaaacaCAUAAUACAGCACUUUa           2        17        761       784
##                  genome_coordinates conservation align_score  seed_cat
##                           <character>    <numeric>   <integer> <integer>
##   [1] [hg19:2:222290690-222290710:-]       0.8304         153         7
##   [2] [hg19:2:222290690-222290710:-]       0.8304         153         7
##   [3] [hg19:2:222290690-222290710:-]       0.8304         153         7
##   [4] [hg19:2:222290671-222290692:-]       0.8094         163         7
##   [5] [hg19:2:222290671-222290692:-]       0.8094         163         7
##   [6]   [hg19:2:74901371-74901394:+]       0.5128         159         7
##         energy mirsvr_score             Type   miR.Sib   miR.Pat
##      <numeric>    <numeric>      <character> <numeric> <numeric>
##   [1]    -16.90      -1.3338 miR down/Gene up      18.1    17.575
##   [2]    -16.90      -1.3304 miR down/Gene up      18.1    17.575
##   [3]    -16.90      -1.2058 miR down/Gene up      18.1    17.575
##   [4]    -17.12      -1.2034 miR down/Gene up      18.1    17.575
##   [5]    -17.12      -1.1881 miR down/Gene up      18.1    17.575
##   [6]    -16.59      -1.1009 miR down/Gene up      18.1    17.575
##      Gene.Sib_MONO Gene.Pat_MONO     Remarks
##          <numeric>     <numeric> <character>
##   [1]     0.6614030      1.248838    Monocyte
##   [2]     0.6614030      1.248838    Monocyte
##   [3]     0.6614030      1.248838    Monocyte
##   [4]     0.6614030      1.248838    Monocyte
##   [5]     0.6614030      1.248838    Monocyte
##   [6]     0.4955519      1.003386    Monocyte
##   -------
##   seqinfo: 9 sequences from an unspecified genome; no seqlengths
```

Find the exact possion of ACUAAs

```
makeExactPosition <- function(irMatch, grInfo)
{
  if (length(irMatch) != 0)
  {
    if (as.character(strand(grInfo)) == "-")
    {
      irMatch <- IRanges(start = width(grInfo) - end(irMatch) + 1,
                         end =  width(grInfo) - start(irMatch) + 1,
                         width = width(irMatch))
    }
    n <- length(irMatch)
    gr <- GRanges(
      seqnames = Rle(rep(as.character(seqnames(grInfo)), n)),
```

```
      ranges = shift(irMatch, start(grInfo) - 1),
      strand = Rle(rep(as.character(strand(grInfo)), n))
    )
    values(gr) <- as.data.frame(rep(grInfo, n), row.names = 1:n)
    names(values(gr))[1:5] <- paste("UTR_", names(values(gr))[1:5], sep = "")
    return(gr)
  }
  return(NA)
}

ngc3UTR_ACTAA_ectPos <- list()

for (i in 1:534)
  ngc3UTR_ACTAA_ectPos[[i]] <- makeExactPosition(irMatch = ngc3UTR_ACTAA[[i]], grInfo = ngc3UTRU1[i,])

head(ngc3UTR_ACTAA_ectPos)
```

```
## [[1]]
## [1] NA
##
## [[2]]
## GRanges object with 1 range and 11 metadata columns:
##       seqnames              ranges strand | UTR_seqnames UTR_start
##          <Rle>           <IRanges>  <Rle> |     <factor> <integer>
##   [1]        1 [22929274, 22929278]      + |            1  22928235
##         UTR_end UTR_width UTR_strand         exon_id exon_rank      SYMBOL
##       <integer> <integer>   <factor>     <character> <integer> <character>
##   [1]  22930087      1853          + ENSE00001156943        17       EPHA8
##       Count_ACTAA Count_QKIMotif     logFC
##         <integer>      <integer> <numeric>
##   [1]          1              0      <NA>
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
##
## [[3]]
## GRanges object with 2 ranges and 11 metadata columns:
##       seqnames             ranges strand | UTR_seqnames UTR_start
##          <Rle>          <IRanges>  <Rle> |     <factor> <integer>
##   [1]       17 [9147124, 9147128]      + |           17   9143286
##   [2]       17 [9147279, 9147283]      + |           17   9143286
##         UTR_end UTR_width UTR_strand         exon_id exon_rank      SYMBOL
##       <integer> <integer>   <factor>     <character> <integer> <character>
##   [1]   9147317      4032          + ENSE00001126297         7        NTN1
##   [2]   9147317      4032          + ENSE00001126297         7        NTN1
##       Count_ACTAA Count_QKIMotif       logFC
##         <integer>      <integer>   <numeric>
##   [1]          2              0 -0.01728598
##   [2]          2              0 -0.01728598
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
##
## [[4]]
## [1] NA
##
```

```
## [[5]]
## [1] NA
##
## [[6]]
## GRanges object with 1 range and 11 metadata columns:
##       seqnames              ranges strand | UTR_seqnames UTR_start
##          <Rle>           <IRanges>  <Rle> |     <factor> <integer>
##   [1]        X [68061976, 68061980]      + |            X  68060498
##         UTR_end UTR_width UTR_strand         exon_id exon_rank      SYMBOL
##       <integer> <integer>   <factor>     <character> <integer> <character>
##   [1]  68061990      1493          + ENSE00001041114         5       EFNB1
##       Count_ACTAA Count_QKIMotif      logFC
##         <integer>      <integer>  <numeric>
##   [1]           1              0 -0.4497018
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

Is the positions overlapping with the postion of miRNA binding?

Flatten the ACTAA postions in a GRange

```
ngc3UTR_ACTAA_ectPosUl <- unlist(GRangesList(ngc3UTR_ACTAA_ectPos[!is.na(ngc3UTR_ACTAA_ectPos)]))

head(ngc3UTR_ACTAA_ectPosUl)
```

```
## GRanges object with 6 ranges and 11 metadata columns:
##       seqnames              ranges strand | UTR_seqnames UTR_start
##          <Rle>           <IRanges>  <Rle> |     <factor> <integer>
##   [1]        1 [22929274, 22929278]      + |            1  22928235
##   [2]       17 [ 9147124,  9147128]      + |           17   9143286
##   [3]       17 [ 9147279,  9147283]      + |           17   9143286
##   [4]        X [68061976, 68061980]      + |            X  68060498
##   [5]        3 [52468250, 52468254]      - |            3  52467069
##   [6]        3 [52467317, 52467321]      - |            3  52467069
##         UTR_end UTR_width UTR_strand         exon_id exon_rank      SYMBOL
##       <integer> <integer>   <factor>     <character> <integer> <character>
##   [1]  22930087      1853          + ENSE00001156943        17       EPHA8
##   [2]   9147317      4032          + ENSE00001126297         7        NTN1
##   [3]   9147317      4032          + ENSE00001126297         7        NTN1
##   [4]  68061990      1493          + ENSE00001041114         5       EFNB1
##   [5]  52469618      2550          - ENSE00000770774        16      SEMA3G
##   [6]  52469618      2550          - ENSE00000770774        16      SEMA3G
##       Count_ACTAA Count_QKIMotif       logFC
##         <integer>      <integer>   <numeric>
##   [1]           1              0        <NA>
##   [2]           2              0 -0.01728598
##   [3]           2              0 -0.01728598
##   [4]           1              0 -0.44970175
##   [5]           2              0        <NA>
##   [6]           2              0        <NA>
##   -------
##   seqinfo: 19 sequences from an unspecified genome; no seqlengths
```

```
head(ngc3UTR_ACTAA_ectPosUl + 20)
```

```
## GRanges object with 6 ranges and 11 metadata columns:
##       seqnames              ranges strand | UTR_seqnames UTR_start
##          <Rle>           <IRanges>  <Rle> |     <factor> <integer>
##   [1]        1 [22929254, 22929298]      + |            1  22928235
##   [2]       17 [ 9147104,  9147148]      + |           17   9143286
##   [3]       17 [ 9147259,  9147303]      + |           17   9143286
##   [4]        X [68061956, 68062000]      + |            X  68060498
##   [5]        3 [52468230, 52468274]      - |            3  52467069
##   [6]        3 [52467297, 52467341]      - |            3  52467069
##         UTR_end UTR_width UTR_strand         exon_id exon_rank      SYMBOL
##       <integer> <integer>   <factor>     <character> <integer> <character>
##   [1]  22930087      1853          + ENSE00001156943        17       EPHA8
##   [2]   9147317      4032          + ENSE00001126297         7        NTN1
##   [3]   9147317      4032          + ENSE00001126297         7        NTN1
##   [4]  68061990      1493          + ENSE00001041114         5       EFNB1
##   [5]  52469618      2550          - ENSE00000770774        16      SEMA3G
##   [6]  52469618      2550          - ENSE00000770774        16      SEMA3G
##       Count_ACTAA Count_QKIMotif       logFC
##         <integer>      <integer>   <numeric>
##   [1]           1              0        <NA>
##   [2]           2              0 -0.01728598
##   [3]           2              0 -0.01728598
##   [4]           1              0 -0.44970175
##   [5]           2              0        <NA>
##   [6]           2              0        <NA>
##   -------
##   seqinfo: 19 sequences from an unspecified genome; no seqlengths
```

For every NGC in the list, find the overlap of extended QRE region and miR binding range

```
QREMirOverlap <- list()
for (i in unique(QkiMirNgcGr$SYMBOL))
QREMirOverlap[[i]] <-  findOverlaps(QkiMirNgcGr[QkiMirNgcGr$SYMBOL == i], ngc3UTR_ACTAA_ectPosUl[ngc3UTR


sapply(QREMirOverlap, length)
```

```
##   EPHA4  SEMA4F   EPHA1 ADORA2B  SEMA3C   EPHA3   EPHB4   EFNB1   UNC5B
##       0       0       0       0       0       0       0       0       0
##  SEMA7A  SEMA4G   EPHB2  SEMA6A   SLIT1
##       4       0       0       0       0
```

```
QREMirOverlap[sapply(QREMirOverlap, length) > 0]
```

```
## $SEMA7A
## Hits object with 4 hits and 0 metadata columns:
##       queryHits subjectHits
##       <integer>   <integer>
##   [1]         1           2
##   [2]         1           4
```

```
##   [3]         2           2
##   [4]         2           4
##   -------
##   queryLength: 2 / subjectLength: 5
```

Visulization of the overlapping

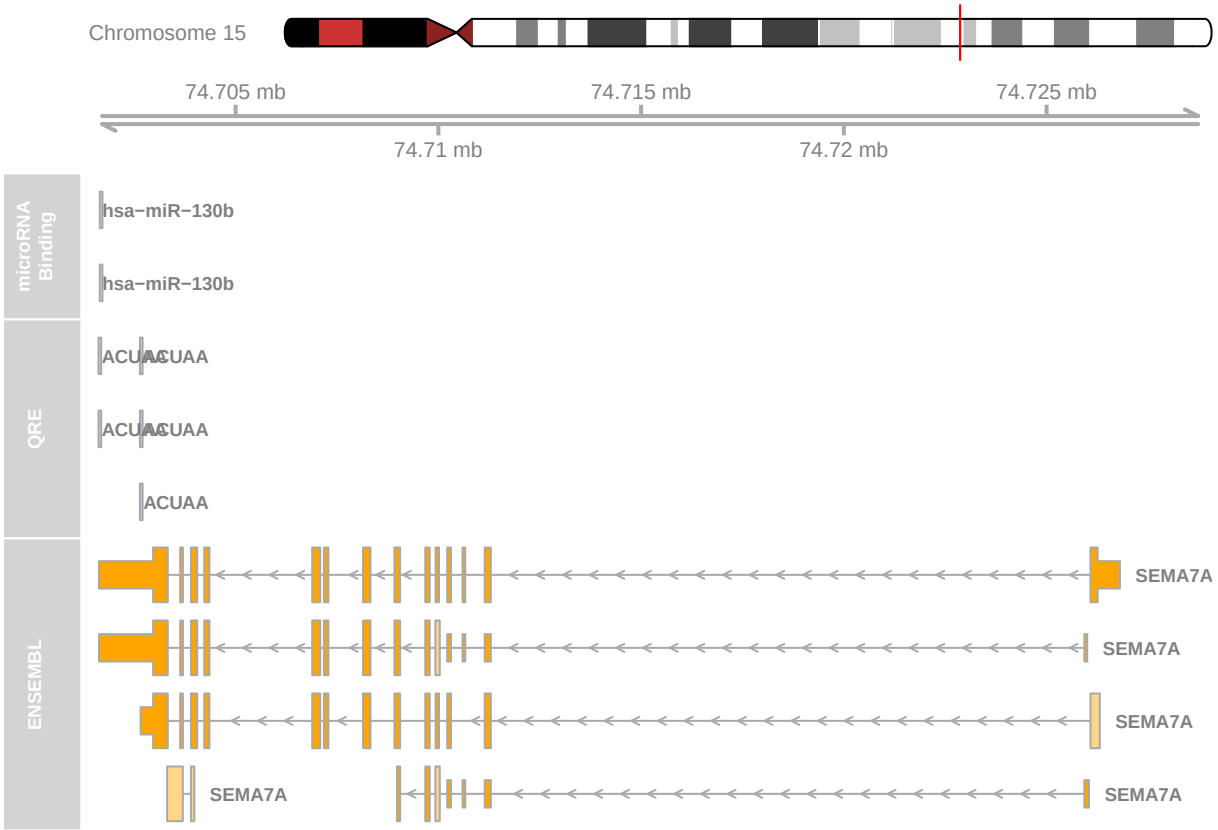
```
i <- "SEMA7A"

miRGr <- QkiMirNgcGr[QkiMirNgcGr$SYMBOL == i]
seqlevels(miRGr) <- paste("Chr", seqlevels(miRGr), sep = "")

QREGr <- ngc3UTR_ACTAA_ectPosUl[ngc3UTR_ACTAA_ectPosUl$SYMBOL== i]
seqlevels(QREGr) <- paste("Chr", seqlevels(QREGr), sep = "")

miRTrack <- AnnotationTrack(miRGr, name = "microRNA Binding", id = values(miRGr)$mirna_name)
QRETrack <- AnnotationTrack(QREGr, name = "QRE", id = "ACUAA")
txdb <- loadDb(
  system.file("extdata", "hg19_knownGene_sample.sqlite", package = "GenomicFeatures")
)
biomTrack <- BiomartGeneRegionTrack(genome = "hg19", symbol = "SEMA7A", transcriptAnnotation = "symbol"
iTrack <- IdeogramTrack(genome = "hg19", chromosome = "chr15")
gTrack <- GenomeAxisTrack(name = "chr15")
```

```
plotTracks(list(iTrack, gTrack, miRTrack, QRETrack, biomTrack), groupAnnotation = "id", just.group = "r
```



```
plotTracks(list(iTrack, gTrack, miRTrack, QRETrack, biomTrack), from = 74701610, to = 74702984, groupAn
```