Splising Sites Selection

Nadzeya Boyeva

2024-11-05

Sites from EEJ gene related matrices

```
fd <- '/home/nadzeya/praktika'
fd116 <- '/media/user5/new/boyeva'

readTxt <- function(path) {
  read.table(file=path,
             sep="\t",
             header=TRUE,
             quote="\"",
             as.is=TRUE)
}
```

EEJ gene related matrices contain information about exon-exon junctions which is inferred from RNA-Seq reads mapping gaps. Rows of every matrix of this kind are individual EEJs, and columns are samples. To create a set of EEJs, where splicing actually occurs in Kasumi-1 cells, we need to combine all the EEJ counts from all the samples:

To keep all potentially active EEJs, select low filtering threshold: an EEJ should have at least 10 reads supporting it or CPM >= -1 (0.5 per million reads) at least in one sample.

Repeat filtration with higher CPM threshold:

. . . and with higher supporting samples number threshold:

Logos

To explore logos of splicing sites in selected EEJ regions, we import reference genome and extract sequences of splicing sites from it.

```
ref_path = file.path(fd, "Homo_sapiens.GRCh38.dna_sm.toplevel.fa")
ref_idx_path = file.path(fd, "Homo_sapiens.GRCh38.dna_sm.toplevel.fa.fai")
file <- FaFile(ref_path, index=ref_idx_path)
fasta <- open(file)
```

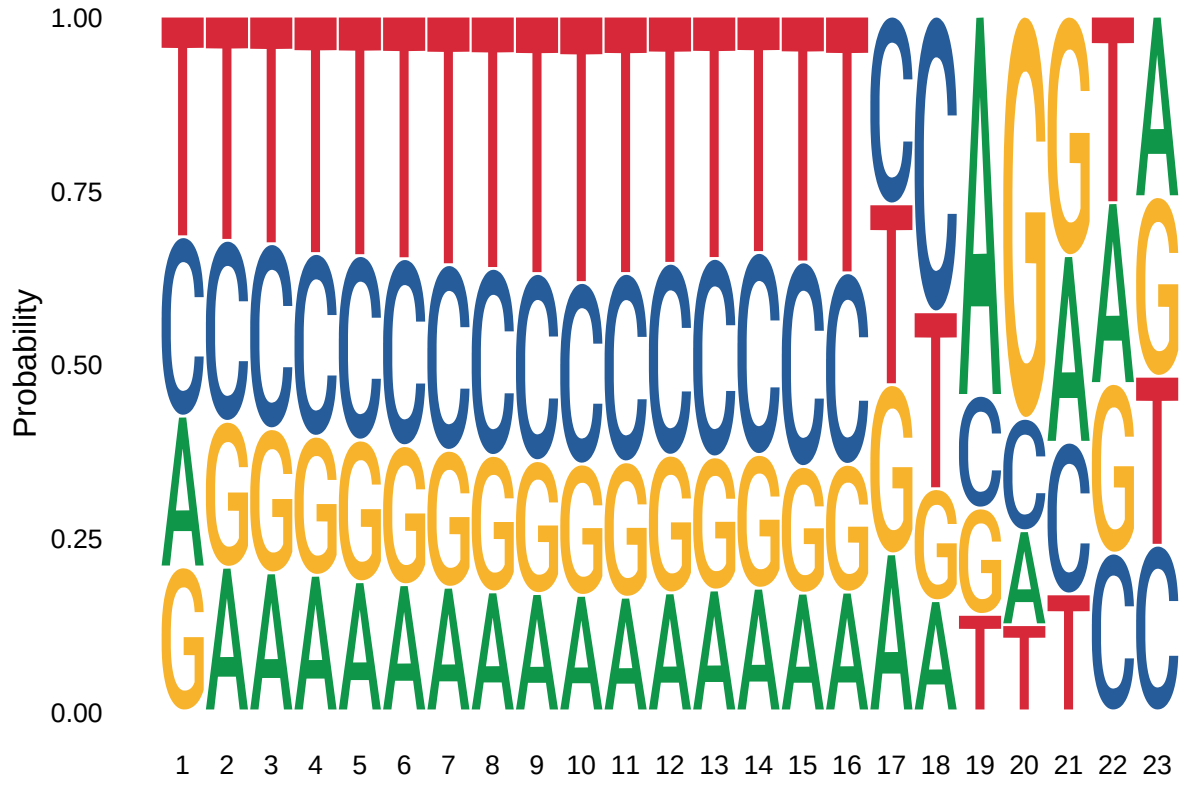
We can see that GT dinucleotide in 5' SS is not that conservative with lower filtering thresholds.

```
input <- "EEJ_raw10_cpm-1_s1.txt"

res <- get_SS_from_EEJ(read_from_file=TRUE, file.path(fd, input))
res$fiveSSs@seqnames <- gsub("chr", "", res$fiveSSs@seqnames)
x1 <- getSeq(x=fasta, res[[1]])
ggseqlogo(data.frame(x1), seq_type='dna', method='prob')
```

```
## Warning: The ‘<scale>‘ argument of ‘guides()‘ cannot be ‘FALSE‘. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the ggseqlogo package.
##   Please report the issue at <https://github.com/omarwagih/ggseqlogo/issues>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

We see the same situation for 3' SS dinucleotide – normally it should be more conservative.

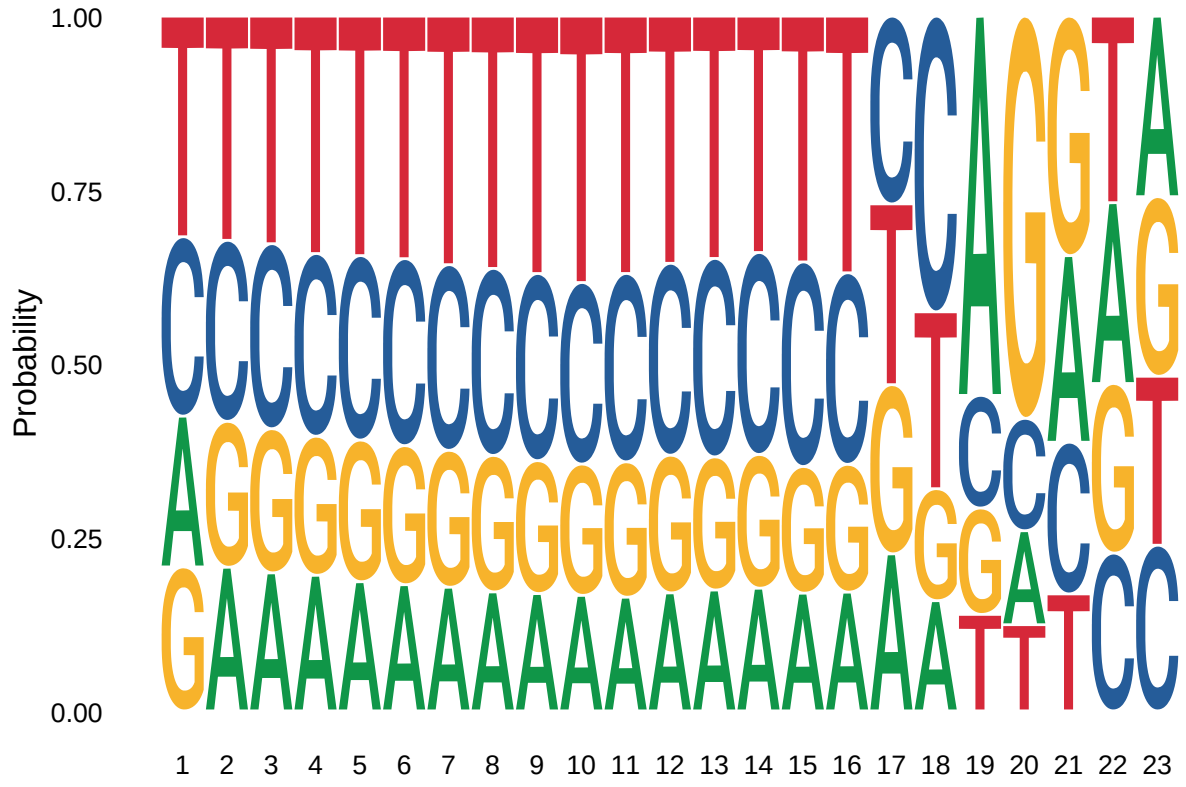
```
res$threeSSs@seqnames <- gsub("chr", "", res$threeSSs@seqnames)
x2 <- getSeq(x=fasta, res[[2]])
ggseqlogo(data.frame(x2), seq_type = 'dna', method='prob')
```



CPM threshold higher:

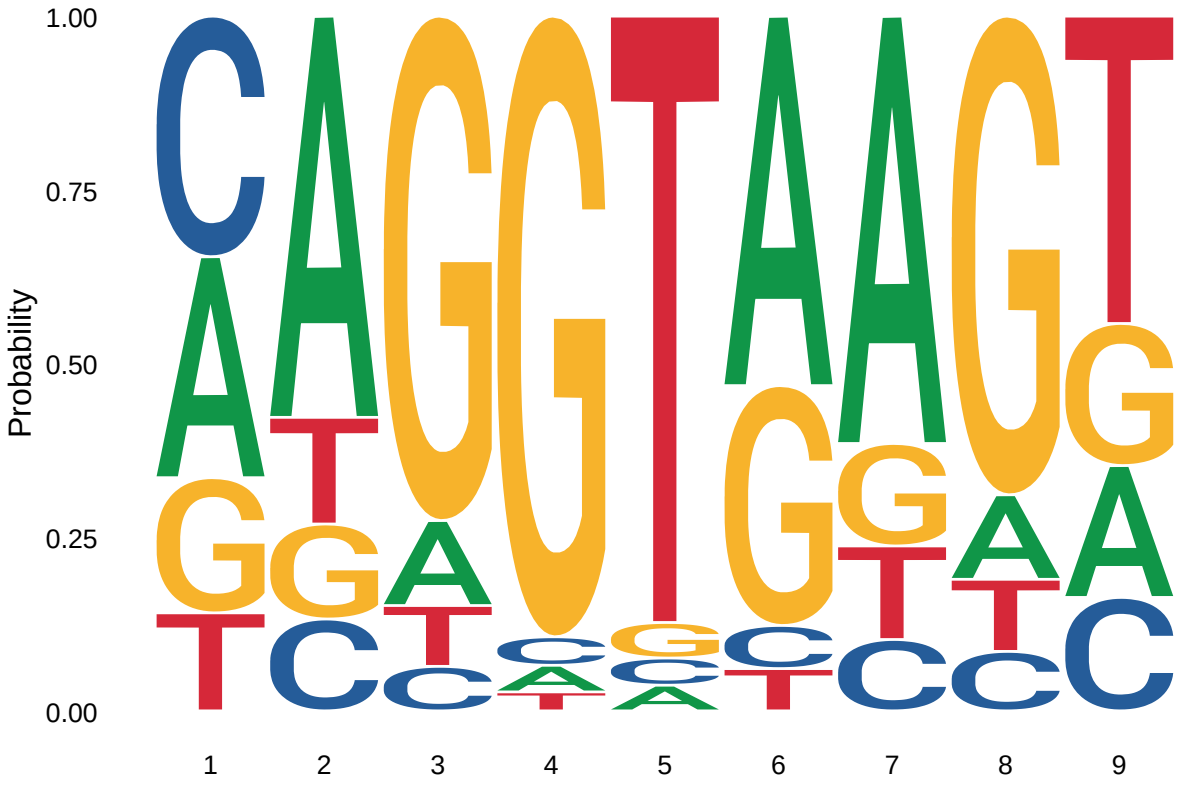
```
input <- "EEJ_raw10_cpm0_s1.txt"

res <- get_SS_from_EEJ(read_from_file=TRUE, file.path(fd, input))
res$fiveSSs@seqnames <- gsub("chr", "", res$fiveSSs@seqnames)
x1 <- getSeq(x=fasta, res[[1]])
ggseqlogo(data.frame(x1), seq_type='dna', method='prob')
```


```
res$threeSSs@seqnames <- gsub("chr", "", res$threeSSs@seqnames)
x2 <- getSeq(x=fasta, res[[2]])
ggseqlogo(data.frame(x2), seq_type = 'dna', method='prob')
```

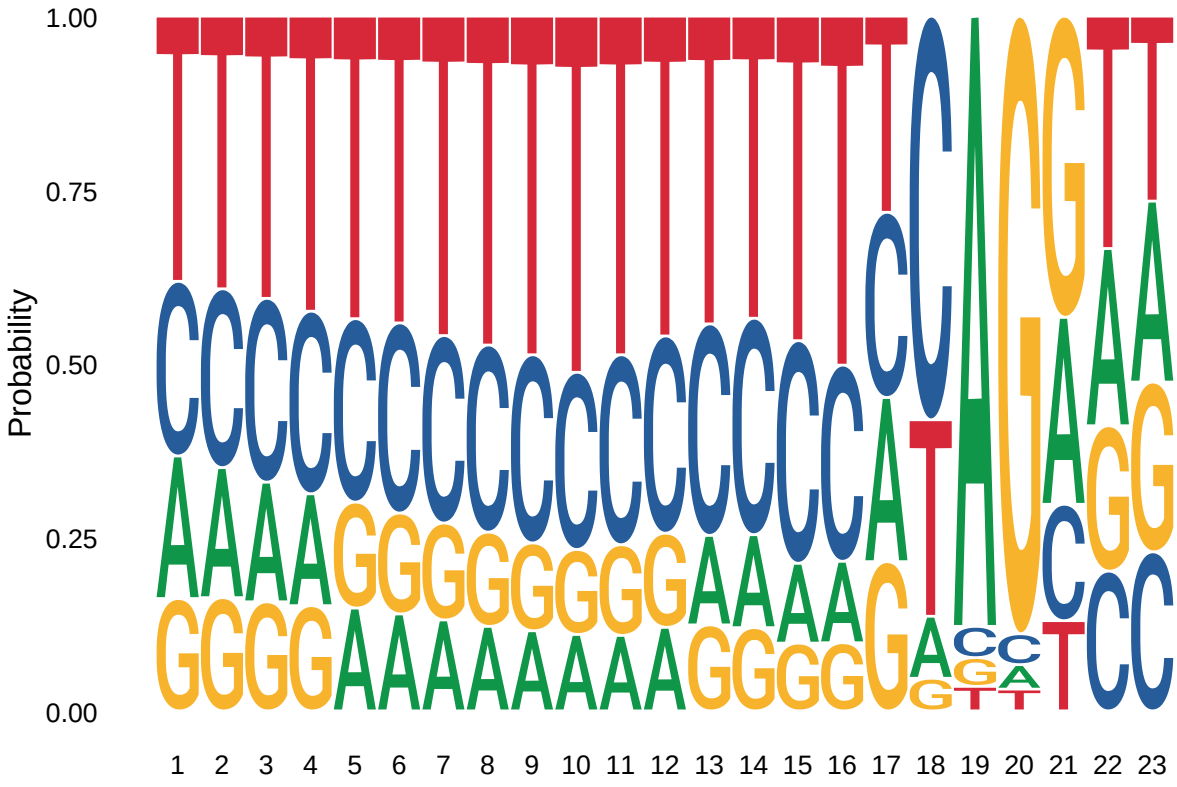


But if we take higher thresholds, dinucleotides are more conservative here.

```
input <- "EEJ_raw10_cpm0_s3.txt"

res <- get_SS_from_EEJ(read_from_file=TRUE, file.path(fd, input))
res$fiveSSs@seqnames <- gsub("chr", "", res$fiveSSs@seqnames)
x1 <- getSeq(x=fasta, res[[1]])
ggseqlogo(data.frame(x1), method='prob')
```




```
res$threeSSs@seqnames <- gsub("chr", "", res$threeSSs@seqnames)
x2 <- getSeq(x=fasta, res[[2]])
ggseqlogo(data.frame(x2), seq_type = 'dna', method='prob')
```



Intersection of Splice Sites from Different Sources

Load EEJ data:

```
eej <- readTxt(file.path(fd, "EEJ_raw10_cpm0_s3.txt"))
eej$seqnames <- gsub("chr", "", eej$seqnames)
ss_coords_eej <- get_SS_from_EEJ(read_from_file=FALSE, df=eej)
```

Load UCSC data (ucsc data):

```
bed_path = file.path(fd, "ucsc.bed")
ucsc <- import(con = bed_path, format = "BED")
ucsc@seqnames <- gsub("chr", "", ucsc@seqnames)
ucsc <- ucsc[nchar(as.character(ucsc@seqnames)) < 3, ]
ucsc@seqnames <- droplevels(ucsc@seqnames)
ucsc_df <- as.data.frame(ucsc)
ucsc_df <- ucsc_df[!duplicated(ucsc_df[c("seqnames", "start", "end", "strand")]),]
ss_coords_ucsc <- get_SS_from_ucsc(read_from_file = FALSE, df=ucsc_df)
ss_coords_ucsc
```

```
## $fiveSSs
## GRanges object with 252413 ranges and 2 metadata columns:
##          seqnames              ranges strand |                 name     score
```

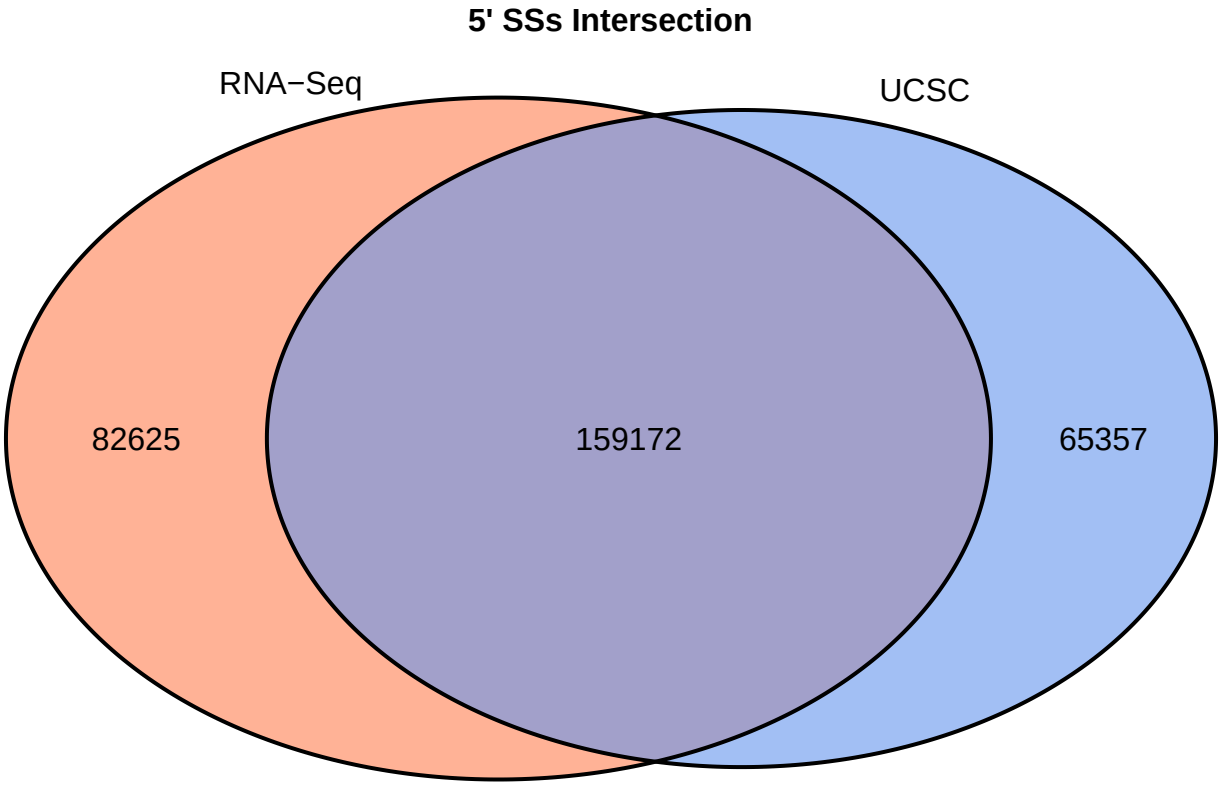
```
##             <Rle>           <IRanges>  <Rle> |            <character> <numeric>
##        1        1 201283902-201283910      + | NM_000299_intron_0_0..         0
##        2        1 201294043-201294051      + | NM_000299_intron_1_0..         0
##        3        1 201313558-201313566      + | NM_000299_intron_2_0..         0
##        4        1 201316695-201316703      + | NM_000299_intron_3_0..         0
##        5        1 201317777-201317785      + | NM_000299_intron_4_0..         0
##      ...      ...                 ...    ... .                      ...       ...
##   732804       22   50784070-50784078      + | NR_026982_intron_0_0..         0
##   732875       22   50738310-50738318      + | NM_001097_intron_0_0..         0
##   732876       22   50739472-50739480      + | NM_001097_intron_1_0..         0
##   732877       22   50739975-50739983      + | NM_001097_intron_2_0..         0
##   732878       22   50744204-50744212      + | NM_001097_intron_3_0..         0
##   -------
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
##
## $threeSSs
## GRanges object with 252413 ranges and 2 metadata columns:
##          seqnames              ranges strand |                   name     score
##             <Rle>           <IRanges>  <Rle> |            <character> <numeric>
##        1        1 201293922-201293944      + | NM_000299_intron_0_0..         0
##        2        1 201313146-201313168      + | NM_000299_intron_1_0..         0
##        3        1 201316533-201316555      + | NM_000299_intron_2_0..         0
##        4        1 201317552-201317574      + | NM_000299_intron_3_0..         0
##        5        1 201318598-201318620      + | NM_000299_intron_4_0..         0
##      ...      ...                 ...    ... .                      ...       ...
##   732804       22   50785153-50785175      + | NR_026982_intron_0_0..         0
##   732875       22   50739251-50739273      + | NM_001097_intron_0_0..         0
##   732876       22   50739674-50739696      + | NM_001097_intron_1_0..         0
##   732877       22   50744041-50744063      + | NM_001097_intron_2_0..         0
##   732878       22   50744633-50744655      + | NM_001097_intron_3_0..         0
##   -------
##   seqinfo: 24 sequences from an unspecified genome; no seqlengths
```

```
ucsc_df$eej_id <- paste0("chr", as.character(ucsc_df$seqnames), ":",
                         as.character(ucsc_df$start - 1), "-",
                         as.character(ucsc_df$end + 1), "_str",
                         ucsc_df$strand)
```

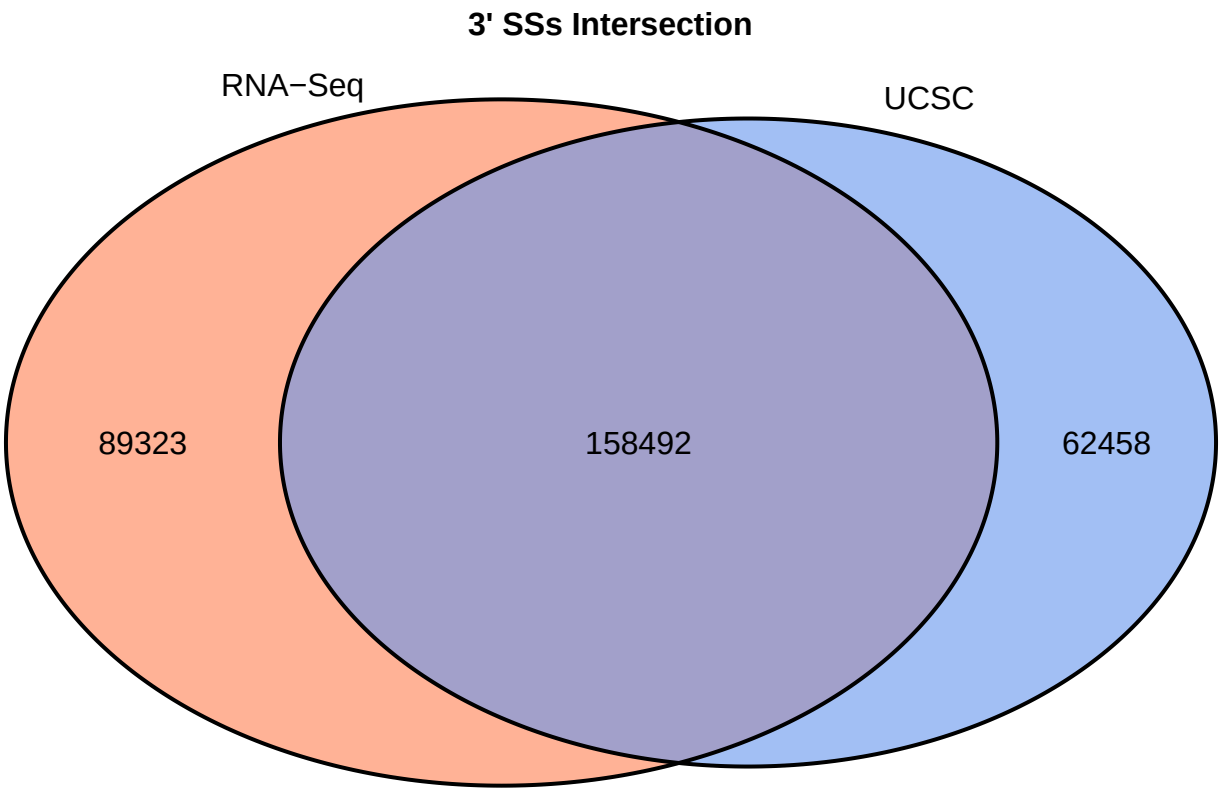
```
eej_5ss <- paste0("chr", as.character(ss_coords_eej$fiveSSs@seqnames), ":",
                  as.character(start(ss_coords_eej$fiveSSs@ranges)), "-",
                  as.character(end(ss_coords_eej$fiveSSs@ranges)),
                  "_str", ss_coords_eej$fiveSSs@strand)

eej_3ss <- paste0("chr", as.character(ss_coords_eej$threeSSs@seqnames), ":",
                  as.character(start(ss_coords_eej$threeSSs@ranges)), "-",
                  as.character(end(ss_coords_eej$threeSSs@ranges)),
                  "_str", ss_coords_eej$threeSSs@strand)

ucsc_5ss <- paste0("chr", as.character(ss_coords_ucsc$fiveSSs@seqnames), ":",
                  as.character(start(ss_coords_ucsc$fiveSSs@ranges)), "-",
                  as.character(end(ss_coords_ucsc$fiveSSs@ranges)),
                  "_str", ss_coords_ucsc$fiveSSs@strand)

ucsc_3ss <- paste0("chr", as.character(ss_coords_ucsc$threeSSs@seqnames), ":",
```

```
                    as.character(start(ss_coords_ucsc$threeSSs@ranges)), "-",
                    as.character(end(ss_coords_ucsc$threeSSs@ranges)),
                    "_str", ss_coords_ucsc$threeSSs@strand)
```

```
venn.plot <- venn.diagram(
  x = list(
    eej = eej_5ss,
    ucsc = ucsc_5ss
  ),
  filename = NULL,
  fill = c("coral", "cornflowerblue"),
  alpha = 0.6,
  cat.cex = 1,
  cex = 1,
  cat.pos = c(-22, 20),
  category.names = c("RNA-Seq", "UCSC"),
  main = "5' SSs Intersection",
  main.cex = 1,
  main.pos = c(0.5, 0.98),
  main.fontfamily = "sans",
  main.fontface = "bold",
  cat.fontfamily = "sans",
  fontfamily = "sans",
)

grid.newpage()
grid.draw(venn.plot)
```




```
venn.plot <- venn.diagram(
  x = list(
    eej = eej_3ss,
    ucsc = ucsc_3ss
  ),
  filename = NULL,
  fill = c("coral", "cornflowerblue"),
  alpha = 0.6,
  cat.cex = 1,
  cex = 1,
  cat.pos = c(-22, 20),
  category.names = c("RNA-Seq", "UCSC"),
  main = "3' SSs Intersection",
  main.cex = 1,
  main.pos = c(0.5, 0.98),
  main.fontfamily = "sans",
  main.fontface = "bold",
  cat.fontfamily = "sans",
  fontfamily = "sans",
)

grid.newpage()
grid.draw(venn.plot)
```



Here we can see which percent of SSs from RNA-Seq data and from UCSC annotation (overall, unique to each source and common) are in genes which are normally expressed in Kasumi-1 cell line.

```
expressed_genes <- readTxt(file.path(fd, "RUNX1-RUNX1T1 project, list of expressed genes"))
```

```
ucsc_df$refseq_id <- sapply(strsplit(ucsc_df$name, "_"), function(x) paste(x[1], x[2], sep = "_"))

unique_to_eej <- eej[eej$eej_id %in% setdiff(eej$eej_id, ucsc_df$eej_id),]
unique_to_ucsc <- ucsc_df[ucsc_df$eej_id %in%setdiff(ucsc_df$eej_id, eej$eej_id),]
common <- eej[eej$eej_id %in% intersect(eej$eej_id, ucsc_df$eej_id),]
mean(unique_to_eej$gene_id %in% expressed_genes$gene_id)
```

```
## [1] 0.8920784
```

```
mean(unique_to_ucsc$refseq_id %in% expressed_genes$refseq_id)
```

```
## [1] 0.03647059
```

```
mean(common$gene_id %in% expressed_genes$gene_id)
```

```
## [1] 0.7956512
```

```
mean(eej$gene_id %in% expressed_genes$gene_id)
```

```
## [1] 0.8418429
```

```
mean(ucsc_df$refseq_id %in% expressed_genes$refseq_id)
```

```
## [1] 0.2086343
```

```
ucsc_unique <- ucsc_df[ucsc_df$eej_id %in%setdiff(ucsc_df$eej_id, eej$eej_id),]
write.table(ucsc_unique, file = file.path(fd, "ucsc_unique.txt"), sep = '\t')
```