SNVs in Splicing SItes

Nadzeya Boyeva

2024-11-09

Search of SNVs in Splicing Sites

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

Read VCF files from first, second and third nascent RNA samples, and all samples merged into one file on alignment step. Retrieve common SNVs for all four files (nas_common).

```
nas_vcf1_path <- file.path(fd, "nas1_snps_f.vcf.gz")
nas_vcf2_path <- file.path(fd, "nas2_snps_f.vcf.gz")
nas_vcf3_path <- file.path(fd, "nas3_2_snps_f.vcf.gz")
nas_merged_vcf_path <- file.path(fd, "nas_merged_snps_f.vcf.gz")

nas1 <- load_vcf(nas_vcf1_path)
```

```
## Warning in .bcfHeaderAsSimpleList(header): duplicate keys in header will be
## forced to unique rownames
```

```
nas2 <- load_vcf(nas_vcf2_path)
```

```
## Warning in .bcfHeaderAsSimpleList(header): duplicate keys in header will be
## forced to unique rownames
```

```
nas3 <- load_vcf(nas_vcf3_path)
```

```
## Warning in .bcfHeaderAsSimpleList(header): duplicate keys in header will be
## forced to unique rownames
```

```
nas_merged <- load_vcf(nas_merged_vcf_path)
```

```
## Warning in .bcfHeaderAsSimpleList(header): duplicate keys in header will be
## forced to unique rownames
```

```
nas_common <- nas_merged[nas_merged@ranges@NAMES %in%
                           Reduce(intersect, list(nas1@ranges@NAMES,
                                                  nas2@ranges@NAMES,
                                                  nas3@ranges@NAMES,
                                                  nas_merged@ranges@NAMES))]
```

Load filtered EEJ dataset. DC is for double-checked: each EEJ has read count >=10 and log10CPM >=-1 at least in one experiment. Unify chromosome notation. Selects metadata EEJ columns. Extract SSs coordinates from EEJ coordinates.

```
ss_df <- readTxt(file.path(fd, "EEJ_filtered_DC.txt"))
ss_df$seqnames <- gsub("chr", "", ss_df$seqnames)
ss_df_short <- ss_df[,1:8] # EEJ metadata
ss_coords <- get_SS_from_EEJ(read_from_file=FALSE, df=ss_df_short)
```

Find SNVs overlapping with SSs.

```
dc_nas_common <- find_overlaps_jointSS(ss_coords, nas_common, ss_df_short, source="EEJ")
head(dc_nas_common)
```

```
##                            eej_id         gene_id seqnames   start     end
## 4754    chr1:1402256-1402462_str- ENSG00000242485        1 1402256 1402462
## 14590   chr1:1825499-1873958_str- ENSG00000078369        1 1825499 1873958
## 14934   chr1:1839238-1873958_str- ENSG00000078369        1 1839238 1873958
## 27209   chr1:3496018-3496605_str- ENSG00000162591        1 3496018 3496605
## 27211   chr1:3496063-3496605_str- ENSG00000162591        1 3496063 3496605
## 1634318 chr1:6098929-6098969_str+ ENSG00000069424        1 6098929 6098969
##         width strand intron_length ss ss_start  ss_end        snp_id snp_pos
## 4754      207      -             0  5  1402456 1402464 1:1402457_A/G 1402457
## 14590   48460      -         48458  5  1873952 1873960 1:1873952_G/A 1873952
## 14934   34721      -         34719  5  1873952 1873960 1:1873952_G/A 1873952
## 27209     588      -           586  5  3496599 3496607 1:3496604_C/T 3496604
## 27211     543      -           541  5  3496599 3496607 1:3496604_C/T 3496604
## 1634318    41      +             0  5  6098927 6098935 1:6098935_G/A 6098935
##         snp_ref snp_alt snp_pos_in_ss
## 4754          A       G             1
## 14590         G       A             0
## 14934         G       A             0
## 27209         C       T             5
## 27211         C       T             5
## 1634318       G       A             8
```

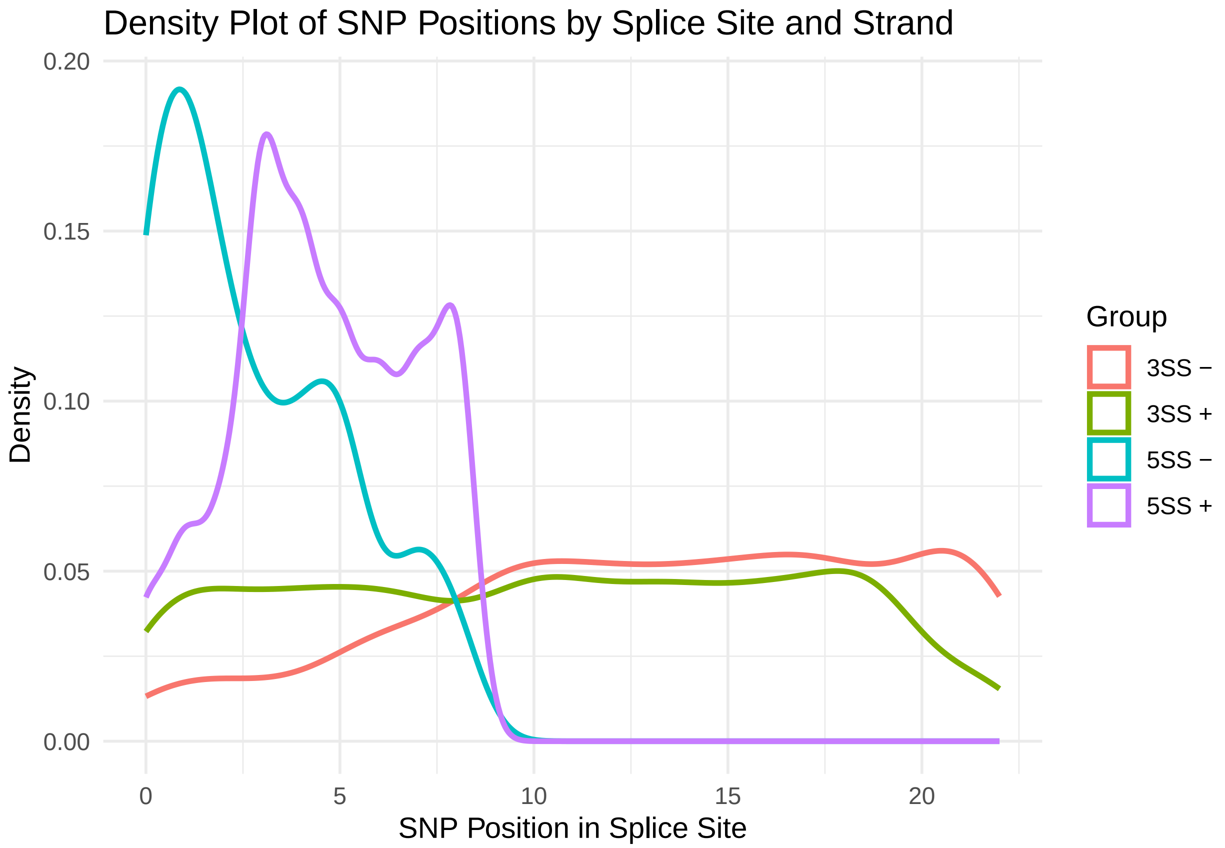
Create separate dataframes for 5' and 3' SSs on + and - strands.

```
#dc_nas_common <- readTxt(file.path(fd, "dc_eej_nas_common.txt"))
dc_nas_common_5ss_plus <- dc_nas_common[(dc_nas_common$ss == 5 & dc_nas_common$strand == '+'),]
dc_nas_common_5ss_minus <- dc_nas_common[(dc_nas_common$ss == 5 & dc_nas_common$strand == '-'),]
dc_nas_common_3ss_plus <- dc_nas_common[(dc_nas_common$ss == 3 & dc_nas_common$strand == '+'),]
dc_nas_common_3ss_minus <- dc_nas_common[(dc_nas_common$ss == 3 & dc_nas_common$strand == '-'),]
```

Plot SNVs distribution in SSs.

```r
# Combine the data into one dataframe with group labels
dc_combined <- rbind(
  data.frame(snp_pos_in_ss = dc_nas_common_5ss_plus$snp_pos_in_ss, group = "5SS +"),
  data.frame(snp_pos_in_ss = dc_nas_common_5ss_minus$snp_pos_in_ss, group = "5SS -"),
  data.frame(snp_pos_in_ss = dc_nas_common_3ss_plus$snp_pos_in_ss, group = "3SS +"),
  data.frame(snp_pos_in_ss = dc_nas_common_3ss_minus$snp_pos_in_ss, group = "3SS -")
)

# Create the ggplot density plot
ggplot(dc_combined, aes(x = snp_pos_in_ss, color = group)) +
  geom_density(linewidth = 1) +  # Add density lines
  labs(
    title = "Density Plot of SNP Positions by Splice Site and Strand",
    x = "SNP Position in Splice Site",
    y = "Density",
    color = "Group"
  ) +
  theme_minimal()
```

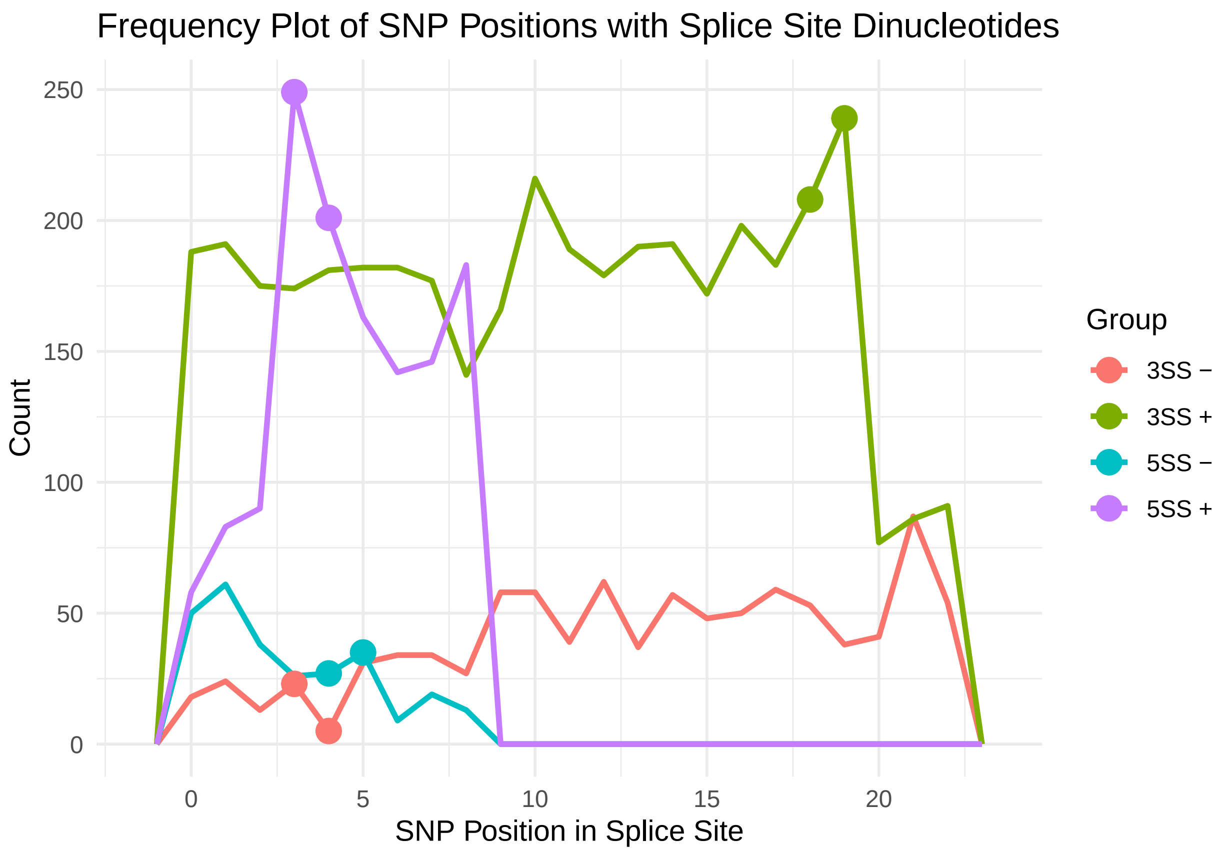


Create count plot of SNVs in each location in SSs. Circles on each line represent theoretical locations of SSs' dinucleotides.

```r
counts <- dc_combined %>%
  group_by(group, snp_pos_in_ss) %>%
  summarise(count = n(), .groups = "drop")
```

```r
dinucleotide_counts <- counts %>%
  filter(
    (group == "5SS +" & (snp_pos_in_ss == 3 | snp_pos_in_ss == 4)) |  # GU for 5SS +
    (group == "5SS -" & (snp_pos_in_ss == 4 | snp_pos_in_ss == 5)) |  # GU for 5SS -
    (group == "3SS +" & (snp_pos_in_ss == 18 | snp_pos_in_ss == 19)) | # AG for 3SS +
    (group == "3SS -" & (snp_pos_in_ss == 3 | snp_pos_in_ss == 4))    # AG for 3SS -
  )

ggplot(dc_combined, aes(x = snp_pos_in_ss, color = group)) +
  geom_freqpoly(binwidth = 1, linewidth = 1) +  # Main frequency lines
  geom_point(
    data = dinucleotide_counts,
    aes(x = snp_pos_in_ss, y = count, color = group),
    size = 4, shape = 19
  ) +
  labs(
    title = "Frequency Plot of SNP Positions with Splice Site Dinucleotides",
    x = "SNP Position in Splice Site",
    y = "Count",
    color = "Group"
  ) +
  theme_minimal()
```



Add reference and alternative sequences of SSs.

```r
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)
```

```r
dc_nas_common <- add_refseqs(fasta, dc_nas_common, source="EEJ")
dc_nas_common <- add_altseqs(dc_nas_common, source="EEJ")
write.table(dc_nas_common, file=file.path(fd, "dc_eej_nas_common.txt"), sep='\t')
head(dc_nas_common)
```

```
##                            eej_id         gene_id seqnames   start     end
## 4754    chr1:1402256-1402462_str- ENSG00000242485        1 1402256 1402462
## 14590   chr1:1825499-1873958_str- ENSG00000078369        1 1825499 1873958
## 14934   chr1:1839238-1873958_str- ENSG00000078369        1 1839238 1873958
## 27209   chr1:3496018-3496605_str- ENSG00000162591        1 3496018 3496605
## 27211   chr1:3496063-3496605_str- ENSG00000162591        1 3496063 3496605
## 1634318 chr1:6098929-6098969_str+ ENSG00000069424        1 6098929 6098969
##         width strand intron_length ss ss_start  ss_end        snp_id snp_pos
## 4754      207      -             0  5  1402456 1402464 1:1402457_A/G 1402457
## 14590   48460      -         48458  5  1873952 1873960 1:1873952_G/A 1873952
## 14934   34721      -         34719  5  1873952 1873960 1:1873952_G/A 1873952
## 27209     588      -           586  5  3496599 3496607 1:3496604_C/T 3496604
## 27211     543      -           541  5  3496599 3496607 1:3496604_C/T 3496604
## 1634318    41      +             0  5  6098927 6098935 1:6098935_G/A 6098935
##         snp_ref snp_alt snp_pos_in_ss    refseq    altseq
## 4754          A       G             1 AAGCACCTG AGGCACCTG
## 14590         G       A             0 GCATACCTG ACATACCTG
## 14934         G       A             0 GCATACCTG ACATACCTG
## 27209         C       T             5 CCCTACCCT CCCTATCCT
## 27211         C       T             5 CCCTACCCT CCCTATCCT
## 1634318       G       A             8 AAGGCCAGG AAGGCCAGA
```

Repeat the same steps for VCF resulting from merged alignment.

```r
dc_nas_merged <- find_overlaps_jointSS(ss_coords, nas_merged, ss_df_short, source="EEJ")
dc_nas_merged <- add_refseqs(fasta, dc_nas_merged, source="EEJ")
dc_nas_merged <- add_altseqs(dc_nas_merged, source="EEJ")
write.table(dc_nas_merged, file=file.path(fd, "dc_eej_nas_merged.txt"), sep='\t')
head(dc_nas_merged)
```

```
##                            eej_id         gene_id seqnames  start    end width
## 37281     chr1:904943-905120_str+ ENSG00000272438        1 904943 905120   178
## 468604    chr1:904944-905116_str+ ENSG00000272438        1 904944 905116   173
## 798716    chr1:953288-953782_str- ENSG00000188976        1 953288 953782   495
## 1059109   chr1:953470-953782_str- ENSG00000188976        1 953470 953782   313
## 798716.1  chr1:953288-953782_str- ENSG00000188976        1 953288 953782   495
## 1059109.1 chr1:953470-953782_str- ENSG00000188976        1 953470 953782   313
##           strand intron_length ss ss_start ss_end       snp_id snp_pos snp_ref
## 37281          +           176  5   904941 904949 1:904947_G/A  904947       G
## 468604         +           171  5   904942 904950 1:904947_G/A  904947       G
## 798716         -             0  5   953776 953784 1:953778_G/C  953778       G
## 1059109        -             0  5   953776 953784 1:953778_G/C  953778       G
## 798716.1       -             0  5   953776 953784 1:953779_A/C  953779       A
## 1059109.1      -             0  5   953776 953784 1:953779_A/C  953779       A
##           snp_alt snp_pos_in_ss    refseq    altseq
## 37281           A             6 GACTCCGCC GACTCCACC
## 468604          A             5 ACTCCGCCG ACTCCACCG
## 798716          C             2 ACGAACCTT ACCAACCTT
## 1059109         C             2 ACGAACCTT ACCAACCTT
## 798716.1        C             3 ACGAACCTT ACGCACCTT
## 1059109.1       C             3 ACGAACCTT ACGCACCTT
```

UCSC Intron Annotation

```
bed_path = file.path(fd, "introns.bed")
introns <- import(con = bed_path, format = "BED")
introns@seqnames <- gsub("chr", "", introns@seqnames)
introns <- introns[nchar(as.character(introns@seqnames)) < 3, ]
introns@seqnames <- droplevels(introns@seqnames)
introns_df <- as.data.frame(introns)
head(introns_df)
```

```
##   seqnames     start       end width strand
## 1        1 201283905 201293941 10037      +
## 2        1 201294046 201313165 19120      +
## 3        1 201313561 201316552  2992      +
## 4        1 201316698 201317571   874      +
## 5        1 201317780 201318617   838      +
## 6        1 201318796 201319815  1020      +
##                                    name score
## 1 NM_000299_intron_0_0_chr1_201283905_f     0
## 2 NM_000299_intron_1_0_chr1_201294046_f     0
## 3 NM_000299_intron_2_0_chr1_201313561_f     0
## 4 NM_000299_intron_3_0_chr1_201316698_f     0
## 5 NM_000299_intron_4_0_chr1_201317780_f     0
## 6 NM_000299_intron_5_0_chr1_201318796_f     0
```

```
ss_coords_introns <- get_SS_from_introns(read_from_file = FALSE, df=introns_df)
introns_nas_common <- find_overlaps_jointSS(ss_coords_introns, nas_common, introns_df, source="introns"
introns_nas_common <- add_refseqs(fasta, introns_nas_common, source="introns")
introns_nas_common <- add_altseqs(introns_nas_common, source="introns")

introns_nas_common$refseq_id <- sapply(strsplit(introns_nas_common$name, "_"), function(parts) {
  paste(parts[1], parts[2], sep = "_")
})

write.table(introns_nas_common, file=file.path(fd, "introns_nas_common.txt"), sep='\t')
head(introns_nas_common)
```

```
##       seqnames     start       end width strand
## 33662        1    922510    922671   162      -
## 9248         1  24902667  24907258  4592      -
## 9258         1  24902667  24907258  4592      -
## 9295         1  24902667  24907258  4592      -
## 47472        1  74733532  74736541  3010      +
## 3162         1 111183588 111184190   603      +
##                                          name score ss  ss_start    ss_end
## 33662      NR_168405_intron_3_0_chr1_922510_r     0  5    922666    922674
## 9248     NM_004350_intron_3_0_chr1_24902667_r     0  5  24907253  24907261
## 9258  NM_001320672_intron_5_0_chr1_24902667_r     0  5  24907253  24907261
## 9295  NM_001031680_intron_4_0_chr1_24902667_r     0  5  24907253  24907261
## 47472    NR_027962_intron_0_0_chr1_74733532_f     0  5  74733529  74733537
## 3162    NM_006090_intron_7_0_chr1_111183588_f     0  5 111183585 111183593
##                snp_id   snp_pos snp_ref snp_alt snp_pos_in_ss    refseq
## 33662    1:922671_C/T    922671       C       T             5 CAGGCAAGG
## 9248   1:24907257_A/G  24907257       A       G             4 AAGGTACGG
## 9258   1:24907257_A/G  24907257       A       G             4 AAGGTACGG
## 9295   1:24907257_A/G  24907257       A       G             4 AAGGTACGG
## 47472  1:74733530_G/A  74733530       G       A             1 GGGGTGGTC
## 3162  1:111183591_A/G 111183591       A       G             6 CTGGTAAGT
##          altseq    refseq_id
## 33662 CAGACAAGG    NR_168405
## 9248  AAGGCACGG    NM_004350
## 9258  AAGGCACGG NM_001320672
## 9295  AAGGCACGG NM_001031680
## 47472 GAGGTGGTC    NR_027962
## 3162  CTGGTAGGT    NM_006090
```

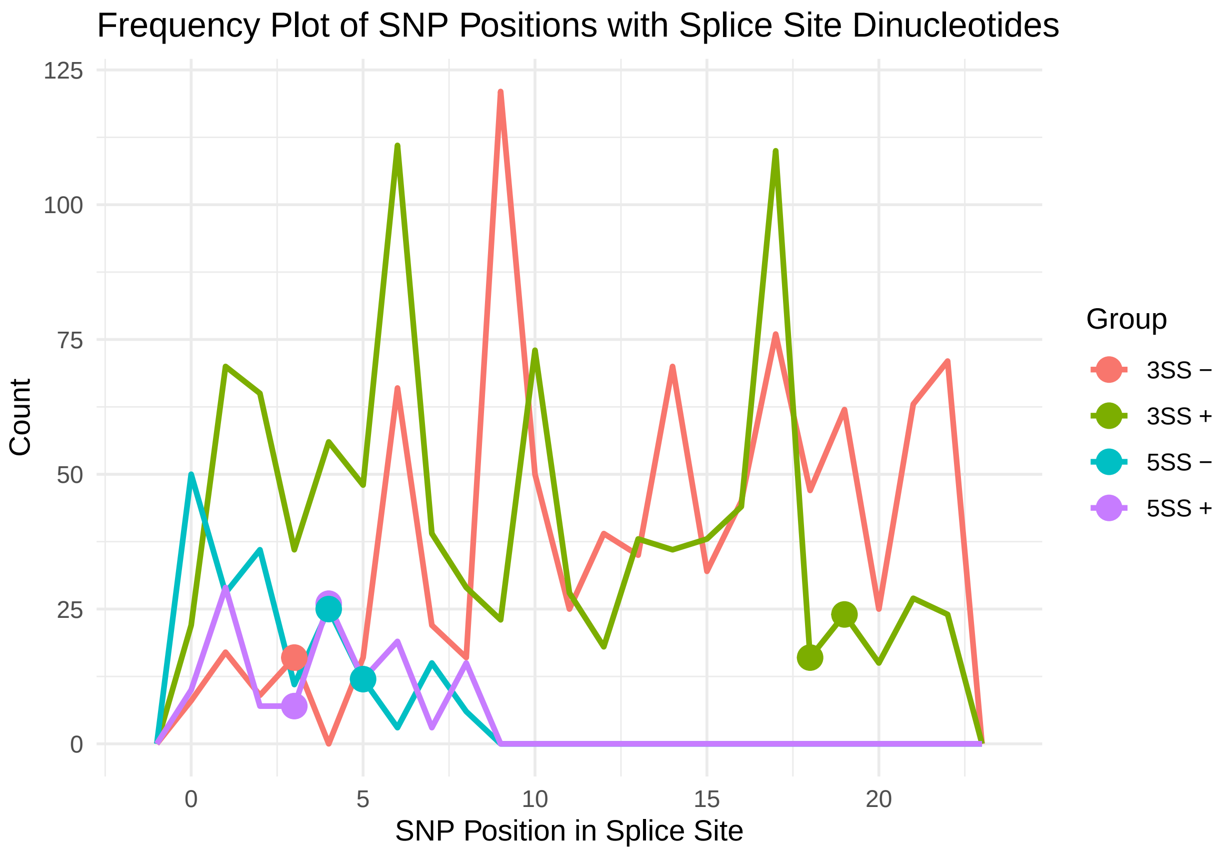
```
introns_nas_common_5ss_plus <- introns_nas_common[(introns_nas_common$ss == 5 & introns_nas_common$stran
introns_nas_common_5ss_minus <- introns_nas_common[(introns_nas_common$ss == 5 & introns_nas_common$stra
introns_nas_common_3ss_plus <- introns_nas_common[(introns_nas_common$ss == 3 & introns_nas_common$stran
introns_nas_common_3ss_minus <- introns_nas_common[(introns_nas_common$ss == 3 & introns_nas_common$stra

introns_combined <- rbind(
  data.frame(snp_pos_in_ss = introns_nas_common_5ss_plus$snp_pos_in_ss, group = "5SS +"),
  data.frame(snp_pos_in_ss = introns_nas_common_5ss_minus$snp_pos_in_ss, group = "5SS -"),
  data.frame(snp_pos_in_ss = introns_nas_common_3ss_plus$snp_pos_in_ss, group = "3SS +"),
  data.frame(snp_pos_in_ss = introns_nas_common_3ss_minus$snp_pos_in_ss, group = "3SS -")
)
```

```
introns_counts <- introns_combined %>%
  group_by(group, snp_pos_in_ss) %>%
  summarise(count = n(), .groups = "drop")

introns_dinucleotide_counts <- introns_counts %>%
  filter(
    (group == "5SS +" & (snp_pos_in_ss == 3 | snp_pos_in_ss == 4)) |  # GU for 5SS +
    (group == "5SS -" & (snp_pos_in_ss == 4 | snp_pos_in_ss == 5)) |  # GU for 5SS -
    (group == "3SS +" & (snp_pos_in_ss == 18 | snp_pos_in_ss == 19)) | # AG for 3SS +
    (group == "3SS -" & (snp_pos_in_ss == 3 | snp_pos_in_ss == 4))    # AG for 3SS -
  )

ggplot(introns_combined, aes(x = snp_pos_in_ss, color = group)) +
  geom_freqpoly(binwidth = 1, linewidth = 1) +  # Main frequency lines
  geom_point(
    data = introns_dinucleotide_counts,
    aes(x = snp_pos_in_ss, y = count, color = group),
    size = 4, shape = 19
  ) +
  labs(
    title = "Frequency Plot of SNP Positions with Splice Site Dinucleotides",
    x = "SNP Position in Splice Site",
    y = "Count",
    color = "Group"
  ) +
  theme_minimal()
```



Load list of expressed genes (leg).

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

```
##           gene_id seqnames     start       end strand  width gene_symbol
## 1 ENSG00000000419    chr20  50934867  50958555      -  23689        DPM1
## 2 ENSG00000000457     chr1 169849631 169894267      -  44637       SCYL3
## 3 ENSG00000000460     chr1 169662007 169854080      + 192074    C1orf112
## 4 ENSG00000000938     chr1  27612064  27635277      -  23214         FGR
## 5 ENSG00000001036     chr6 143494811 143511690      -  16880       FUCA2
## 6 ENSG00000001084     chr6  53497341  53616970      - 119630        GCLC
##                                                     gene_name previous_symbol
## 1 dolichyl-phosphate mannosyltransferase subunit 1, catalytic
## 2                                    SCY1 like pseudokinase 3
## 3                         chromosome 1 open reading frame 112
## 4               FGR proto-oncogene, Src family tyrosine kinase            SRC2
## 5                                            alpha-L-fucosidase 2
## 6                  glutamate-cysteine ligase catalytic subunit     GLCLC, GLCL
##            synonyms uniprot_id refseq_id ncbi_gene_id    hgnc_id
## 1       MPDS, CDGIE     O60762 NM_003859         8813  HGNC:3005
## 2     PACE-1, PACE1     Q8IZE3 NM_181093        57147 HGNC:19285
## 3          FLJ10706     Q9NSG2 NM_018186        55732 HGNC:25565
## 4   c-fgr, p55c-fgr     P09769 NM_005248         2268  HGNC:3697
## 5 MGC1314, dJ20N2.5     Q9BTY2 NM_032020         2519  HGNC:4008
## 6               GCS     P48506                   2729  HGNC:4311
```

```
introns_nas_common_dinucl <- rbind(
  introns_nas_common[introns_nas_common$ss == "5" &
                       introns_nas_common$strand == "+" &
                       (introns_nas_common$snp_pos_in_ss %in% c("3", "4")),],
  introns_nas_common[introns_nas_common$ss == "5" &
                       introns_nas_common$strand == "-" &
                       (introns_nas_common$snp_pos_in_ss %in% c("4", "5")),],
  introns_nas_common[introns_nas_common$ss == "3" &
                       introns_nas_common$strand == "+" &
                       (introns_nas_common$snp_pos_in_ss %in% c("18", "19")),],
  introns_nas_common[introns_nas_common$ss == "3" &
                       introns_nas_common$strand == "-" &
                       (introns_nas_common$snp_pos_in_ss %in% c("3", "4")),])

leg_dinucl <- leg[match(introns_nas_common_dinucl$refseq_id, leg$refseq_id),]
introns_nas_common_dinucl$gene_symbol <- leg_dinucl$gene_symbol
introns_nas_common_dinucl$gene_name <- leg_dinucl$gene_name

write.table(introns_nas_common_dinucl, file=file.path(fd, "introns_nas_common_dinucleotides.txt"), sep=
```

Check Intersections With List of Expressed Genes

```
mean(is.na(dc_nas_common[!match(dc_nas_common$gene_id,
                                leg$gene_id),]))
```

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

```
mean(is.na(dc_nas_merged[!match(dc_nas_merged$gene_id,
                                leg$gene_id),]))
```

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

```
mean(is.na(introns_nas_common[!match(introns_nas_common$refseq_id, leg$refseq_id),]))
```

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