SubRead Annotations

Manfred Schmid

Contents

03 July, 2020; 14:02

Setup

```
knitr::opts_chunk$set(fig.width=12, fig.height=8,
                      fig.path=paste0('Figures_subReadAnnotations/'),
                      dev='pdf',
                      echo=TRUE, warning=FALSE, message=FALSE,
                      error=TRUE)
```

```
suppressWarnings(library('tidyverse'))
suppressWarnings(library('magrittr'))
suppressWarnings(library('knitr'))
```

Exons

load table from subRead

```
(exon_anno_tbl <- rtracklayer::import('/Volumes/GenomeDK/ms_tools/subread-2.0.0-Linux-x86_64/annotation
```

```
## # A tibble: 261,752 x 8
##    seqnames start   end width strand GeneID    ExonID score
##    <fct>    <int> <int> <int> <fct>  <chr>     <chr>  <dbl>
##  1 chr1     11874 12227   354 +      100287102 1          0
##  2 chr1     12613 12721   109 +      100287102 2          0
##  3 chr1     13221 14409  1189 +      100287102 3          0
##  4 chr1     14362 14829   468 -      653635    1          0
##  5 chr1     14970 15038    69 -      653635    2          0
##  6 chr1     15796 15947   152 -      653635    3          0
##  7 chr1     16607 16765   159 -      653635    4          0
##  8 chr1     16858 17055   198 -      653635    5          0
##  9 chr1     17233 17368   136 -      653635    6          0
## 10 chr1     17606 17742   137 -      653635    7          0
## # ... with 261,742 more rows
```

add exon nr and counts

```
(exon_cnts <- exon_anno_tbl %>%
  mutate(GeneID = as.character(GeneID)) %>%
  group_by(GeneID) %>%
  summarize(exon_cnt = n(),
            tr_exons_width = sum(end-start)))
```

```
## # A tibble: 28,395 x 3
##    GeneID    exon_cnt tr_exons_width
##    <chr>        <int>          <int>
##  1 1                8           1758
##  2 10               3           1415
##  3 100             13           1851
##  4 1000            20           4874
##  5 10000           23           8411
##  6 100008587        1            155
##  7 100008588        1           1868
##  8 100008589        1           5069
##  9 100009601        2             71
## 10 100009602        2             71
## # ... with 28,385 more rows
```

```
(exon_anno_tbl %<>%
  group_by(GeneID) %>%
  mutate(exon_nr = as.integer(ifelse(strand == '+',
                                     rank(start, ties.method = 'first'),
                                     rank(-start, ties.method = 'first'))),
         exon_width = end-start) %>%
  dplyr::select(GeneID, ExonID, exon_nr, width) %>%
  left_join(., exon_cnts) %>%
  ungroup %>%
  mutate(class = case_when(.$exon_cnt == 1 ~ 'monoexonic',
                           .$exon_nr == 1 ~ 'multiexonic first exon',
                           .$exon_nr == .$exon_cnt ~ 'multiexonic last exon',
                           .$exon_nr < .$exon_cnt  ~ 'multiexonic internal')))
```

```
## # A tibble: 261,752 x 7
##    GeneID    ExonID exon_nr width exon_cnt tr_exons_width class
##    <chr>     <chr>    <int> <int>    <int>          <int> <chr>
##  1 100287102 1            1   354        3           1649 multiexonic fir~
##  2 100287102 2            2   109        3           1649 multiexonic int~
##  3 100287102 3            3  1189        3           1649 multiexonic las~
##  4 653635    1           11   468       11           1758 multiexonic las~
##  5 653635    2           10    69       11           1758 multiexonic int~
##  6 653635    3            9   152       11           1758 multiexonic int~
##  7 653635    4            8   159       11           1758 multiexonic int~
##  8 653635    5            7   198       11           1758 multiexonic int~
##  9 653635    6            6   136       11           1758 multiexonic int~
## 10 653635    7            5   137       11           1758 multiexonic int~
## # ... with 261,742 more rows
```

ENTREZ to ENSEMBL

```
library("AnnotationDbi")
library("org.Hs.eg.db")
columns(org.Hs.eg.db)
```

```
##  [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"
##  [5] "ENSEMBLTRANS" "ENTREZID"     "ENZYME"       "EVIDENCE"
##  [9] "EVIDENCEALL"  "GENENAME"     "GO"           "GOALL"
## [13] "IPI"          "MAP"          "OMIM"         "ONTOLOGY"
## [17] "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"
## [21] "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
## [25] "UNIGENE"      "UNIPROT"
```

```
entrezid_map <- select(org.Hs.eg.db,
        key=unique(exon_anno_tbl$GeneID), columns=c('ENSEMBL', 'SYMBOL', "REFSEQ"),
        keytype="ENTREZID")

head(entrezid_map)
```

```
##    ENTREZID         ENSEMBL    SYMBOL       REFSEQ
## 1 100287102 ENSG00000223972   DDX11L1    NR_046018
## 2    653635 ENSG00000227232    WASH7P    NR_024540
## 3 102466751 ENSG00000278267 MIR6859-1    NR_106918
## 4 100302278 ENSG00000284332 MIR1302-2    NR_036051
## 5    645520 ENSG00000237613   FAM138A    NR_026818
## 6     79501 ENSG00000186092     OR4F5 NM_001005484
```

Gencode gene info

```
(gencode_tbl <- rtracklayer::import('/Volumes/GenomeDK/annotations/hg38/Gencode_v28/gencode.v28.genes.g
   dplyr::mutate(ENSEMBL = sub('\\..*', '', gene_id)) %>%
   dplyr::select(ENSEMBL, gene_name, gene_type))
```

```
## # A tibble: 58,381 x 3
##    ENSEMBL         gene_name    gene_type
##    <chr>           <chr>        <chr>
##  1 ENSG00000223972 DDX11L1      transcribed_unprocessed_pseudogene
##  2 ENSG00000227232 WASH7P       unprocessed_pseudogene
##  3 ENSG00000278267 MIR6859-1    miRNA
##  4 ENSG00000243485 RP11-34P13.3 lincRNA
##  5 ENSG00000284332 MIR1302-2    miRNA
##  6 ENSG00000237613 FAM138A      lincRNA
##  7 ENSG00000268020 OR4G4P       unprocessed_pseudogene
##  8 ENSG00000240361 OR4G11P      transcribed_unprocessed_pseudogene
##  9 ENSG00000186092 OR4F5        protein_coding
## 10 ENSG00000238009 RP11-34P13.7 lincRNA
## # ... with 58,371 more rows
```

```
sort(table(gencode_tbl$gene_type),decreasing = T)
```

```
##
##                      protein_coding               processed_pseudogene
##                               19901                              10219
##                            lincRNA                          antisense
##                               7490                               5501
##             unprocessed_pseudogene                           misc_RNA
##                               2664                               2213
##                              snRNA                              miRNA
##                               1900                               1881
##                                TEC                             snoRNA
##                               1067                                943
##                     sense_intronic transcribed_unprocessed_pseudogene
##                                899                                855
##               processed_transcript                               rRNA
##                                556                                544
##   transcribed_processed_pseudogene                    IG_V_pseudogene
##                                472                                188
##                  sense_overlapping                          IG_V_gene
##                                183                                144
##     transcribed_unitary_pseudogene                          TR_V_gene
##                                123                                106
##                 unitary_pseudogene                          TR_J_gene
##                                 95                                 79
##                             scaRNA      bidirectional_promoter_lncRNA
##                                 49                                 47
##             polymorphic_pseudogene                          IG_D_gene
##                                 38                                 37
##                    TR_V_pseudogene           3prime_overlapping_ncRNA
##                                 33                                 32
##                            Mt_tRNA                          IG_J_gene
##                                 22                                 18
##                         pseudogene                          IG_C_gene
##                                 18                                 14
##                    IG_C_pseudogene                           ribozyme
##                                  9                                  8
##                          TR_C_gene                               sRNA
##                                  6                                  5
##                          TR_D_gene                    TR_J_pseudogene
##                                  4                                  4
##                    IG_J_pseudogene                         non_coding
##                                  3                                  3
##                            Mt_rRNA    translated_processed_pseudogene
##                                  2                                  2
##                      IG_pseudogene                       macro_lncRNA
##                                  1                                  1
##                              scRNA                           vaultRNA
##                                  1                                  1
```

combine subread exons with gencode those

```
(exon_anno_tbl %<>%
  left_join(., dplyr::rename(entrezid_map, GeneID=ENTREZID)) %>%
   left_join(., gencode_tbl))
```

```
## # A tibble: 4,801,919 x 12
##    GeneID  ExonID exon_nr width exon_cnt tr_exons_width class     ENSEMBL
##    <chr>   <chr>    <int> <int>    <int>          <int> <chr>     <chr>
##  1 100287~ 1            1   354        3           1649 multiexo~ ENSG000~
##  2 100287~ 2            2   109        3           1649 multiexo~ ENSG000~
##  3 100287~ 3            3  1189        3           1649 multiexo~ ENSG000~
##  4 653635  1           11   468       11           1758 multiexo~ ENSG000~
##  5 653635  2           10    69       11           1758 multiexo~ ENSG000~
##  6 653635  3            9   152       11           1758 multiexo~ ENSG000~
##  7 653635  4            8   159       11           1758 multiexo~ ENSG000~
##  8 653635  5            7   198       11           1758 multiexo~ ENSG000~
##  9 653635  6            6   136       11           1758 multiexo~ ENSG000~
## 10 653635  7            5   137       11           1758 multiexo~ ENSG000~
## # ... with 4,801,909 more rows, and 4 more variables: SYMBOL <chr>,
## #   REFSEQ <chr>, gene_name <chr>, gene_type <chr>
```

```
sort(table(exon_anno_tbl$gene_type),decreasing = T)
```

```
##
##                     protein_coding                            lincRNA
##                            4355224                              20056
##                          antisense transcribed_unprocessed_pseudogene
##                               8754                               6492
##               processed_transcript                              miRNA
##                               4628                               1859
##     transcribed_unitary_pseudogene             polymorphic_pseudogene
##                               1361                               1202
##             unprocessed_pseudogene   transcribed_processed_pseudogene
##                               1010                                551
##                     sense_intronic                             snoRNA
##                                383                                372
##      bidirectional_promoter_lncRNA                  sense_overlapping
##                                236                                224
##                          TR_C_gene               processed_pseudogene
##                                128                                 77
##                              snRNA                                TEC
##                                 33                                 29
##                               rRNA                             scaRNA
##                                 18                                 17
##                           misc_RNA           3prime_overlapping_ncRNA
##                                 11                                  3
##                         non_coding                           ribozyme
##                                  3                                  1
##                              scRNA                 unitary_pseudogene
##                                  1                                  1
```

```
save(exon_anno_tbl, file='../data/subRead_exon_annotations.RData')
```

Transcripts table

```
(transcripts <- rtracklayer::import('/Volumes/GenomeDK/ms_tools/subread-2.0.0-Linux-x86_64/annotation/hg
```

```
## GRanges object with 28421 ranges and 5 metadata columns:
##                  seqnames          ranges strand |        name     score
##                     <Rle>       <IRanges>  <Rle> | <character> <numeric>
##       [1]           chr1     11874-14409      + |   100287102         0
##       [2]           chr1     14362-29370      - |      653635         0
##       [3]           chr1     17369-17436      - |   102466751         0
##       [4]           chr1     30366-30503      + |   100302278         0
##       [5]           chr1     34611-36081      - |      645520         0
##       ...            ...             ...    ... .         ...       ...
##   [28417] NW_009646208.1     21277-23071      + |   102723722         0
##   [28418] NW_009646208.1     25775-30157      - |   101929829         0
##   [28419] NW_011332699.1     64666-77517      - |   105379672         0
##   [28420] NW_011332699.1    76210-170143      + |       85316         0
##   [28421] NW_011332701.1 2741969-2798262      - |   102725021         0
##               itemRgb           thick                          blocks
##           <character>       <IRanges>                   <IRangesList>
##       [1]     #FF0000     11874-14409         1-354,740-848,1348-2536
##       [2]     #FF0000     14362-29370      1-468,609-677,1435-1586,...
##       [3]     #FF0000     17369-17436                            1-68
##       [4]     #FF0000     30366-30503                           1-138
##       [5]     #FF0000     34611-36081         1-564,667-871,1111-1471
##       ...         ...             ...                             ...
##   [28417]     #FF0000     21277-23071      1-250,493-576,966-1008,...
##   [28418]     #FF0000     25775-30157      1-254,353-494,949-1136,...
##   [28419]     #FF0000     64666-77517     1-220,453-564,3270-3342,...
##   [28420]     #FF0000    76210-170143 1-203,1673-1862,56463-56625,...
##   [28421]     #FF0000 2741969-2798262 1-731,1522-1659,11728-11842,...
##   -------
##   seqinfo: 55 sequences from an unspecified genome; no seqlengths
```

```
(tr_anno_tbl <- exon_anno_tbl %>%
  distinct(GeneID, exon_cnt, tr_exons_width, ENSEMBL, SYMBOL, REFSEQ, gene_name, gene_type))
```

```
## # A tibble: 287,698 x 8
##    GeneID  exon_cnt tr_exons_width ENSEMBL    SYMBOL   REFSEQ   gene_name
##    <chr>      <int>          <int> <chr>      <chr>    <chr>    <chr>
##  1 100287~        3           1649 ENSG00000~ DDX11L1  NR_0460~ DDX11L1
##  2 653635        11           1758 ENSG00000~ WASH7P   NR_0245~ WASH7P
##  3 102466~        1             67 ENSG00000~ MIR6859~ NR_1069~ MIR6859-1
##  4 100302~        1            137 ENSG00000~ MIR1302~ NR_0360~ MIR1302-2
##  5 645520         3           1127 ENSG00000~ FAM138A  NR_0268~ FAM138A
##  6 79501          1            917 ENSG00000~ OR4F5    NM_0010~ OR4F5
##  7 79501          1            917 ENSG00000~ OR4F5    NP_0010~ OR4F5
##  8 729737         3           5471 <NA>       LOC7297~ NR_0399~ <NA>
##  9 102725~        4           1169 <NA>       LOC1027~ NR_1483~ <NA>
## 10 102723~       11           2109 ENSG00000~ WASH9P   XR_0017~ RP11-34P1~
## # ... with 287,688 more rows, and 1 more variable: gene_type <chr>
```

–> no good fit with RefSeq !?

```
(tr_anno_tbl <- exon_anno_tbl %>%
  distinct(GeneID, exon_cnt, tr_exons_width, ENSEMBL, SYMBOL, gene_name, gene_type))
```

```
## # A tibble: 31,756 x 7
##    GeneID    exon_cnt tr_exons_width ENSEMBL  SYMBOL  gene_name gene_type
##    <chr>        <int>          <int> <chr>    <chr>   <chr>     <chr>
##  1 100287102        3           1649 ENSG000~ DDX11L1 DDX11L1   transcrib~
##  2 653635          11           1758 ENSG000~ WASH7P  WASH7P    unprocess~
##  3 102466751        1             67 ENSG000~ MIR685~ MIR6859-1 miRNA
##   4 100302278         1             137 ENSG000~ MIR130~ MIR1302-2 miRNA
##   5 645520            3            1127 ENSG000~ FAM138A FAM138A   lincRNA
##   6 79501             1             917 ENSG000~ OR4F5   OR4F5     protein_c~
##   7 729737            3            5471 <NA>     LOC729~ <NA>      <NA>
##   8 102725121         4            1169 <NA>     LOC102~ <NA>      <NA>
##   9 102723897        11            2109 ENSG000~ WASH9P  RP11-34P~ unprocess~
##  10 102465909         1              67 ENSG000~ MIR685~ MIR6859-2 miRNA
## # ... with 31,746 more rows
```

–> much more reasonable list!

```
filter(tr_anno_tbl, gene_type == 'protein_coding')
```

```
## # A tibble: 19,463 x 7
##    GeneID exon_cnt tr_exons_width ENSEMBL     SYMBOL gene_name gene_type
##    <chr>     <int>          <int> <chr>       <chr>  <chr>     <chr>
##  1 79501         1            917 ENSG000001~ OR4F5  OR4F5     protein_co~
##  2 729759        1            938 ENSG000002~ OR4F29 OR4F29    protein_co~
##  3 81399         1            938 ENSG000002~ OR4F16 OR4F16    protein_co~
##  4 81399         1            938 ENSG000002~ OR4F16 OR4F29    protein_co~
##  5 81399         1            938 ENSG000002~ OR4F16 OR4F3     protein_co~
##  6 148398       14           2540 ENSG000001~ SAMD11 SAMD11    protein_co~
##  7 26155        19           2781 ENSG000001~ NOC2L  NOC2L     protein_co~
##  8 339451       12           3438 ENSG000001~ KLHL17 KLHL17    protein_co~
##  9 84069        15           2545 ENSG000001~ PLEKH~ PLEKHN1   protein_co~
## 10 84808         4           3900 ENSG000001~ PERM1  PERM1     protein_co~
## # ... with 19,453 more rows
```

```
save(tr_anno_tbl, file='../data/subRead_tr_annotations.RData')
```

sessionInfo

```
sessionInfo()
```

```
## R version 3.5.0 (2018-04-23)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS  10.14.6
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] org.Hs.eg.db_3.7.0   AnnotationDbi_1.44.0 IRanges_2.14.10
##  [4] S4Vectors_0.18.3     Biobase_2.40.0       BiocGenerics_0.26.0
##  [7] bindrcpp_0.2.2       knitr_1.20           magrittr_1.5
## [10] forcats_0.3.0        stringr_1.3.1        dplyr_0.7.5
## [13] purrr_0.2.5          readr_1.1.1          tidyr_0.8.1
## [16] tibble_1.4.2         ggplot2_3.1.0        tidyverse_1.2.1
##
## loaded via a namespace (and not attached):
##  [1] httr_1.3.1                  bit64_0.9-7
##  [3] jsonlite_1.5                modelr_0.1.2
##  [5] assertthat_0.2.0            blob_1.1.1
##  [7] GenomeInfoDbData_1.1.0      cellranger_1.1.0
##  [9] Rsamtools_1.32.0            yaml_2.1.19
## [11] RSQLite_2.1.1               pillar_1.2.3
## [13] backports_1.1.2             lattice_0.20-35
## [15] glue_1.2.0                  digest_0.6.15
## [17] GenomicRanges_1.32.3        XVector_0.20.0
## [19] rvest_0.3.2                 colorspace_1.3-2
## [21] htmltools_0.3.6             Matrix_1.2-14
## [23] plyr_1.8.4                  psych_1.8.4
## [25] XML_3.98-1.11               pkgconfig_2.0.1
## [27] broom_0.4.4                 haven_1.1.1
## [29] zlibbioc_1.26.0             scales_0.5.0
## [31] BiocParallel_1.14.1         withr_2.1.2
## [33] SummarizedExperiment_1.10.1 lazyeval_0.2.1
## [35] cli_1.0.0                   mnormt_1.5-5
## [37] crayon_1.3.4                readxl_1.1.0
## [39] memoise_1.1.0               evaluate_0.10.1
## [41] nlme_3.1-137                xml2_1.2.0
## [43] foreign_0.8-70              tools_3.5.0
## [45] hms_0.4.2                   matrixStats_0.53.1
## [47] munsell_0.5.0               DelayedArray_0.6.0
## [49] Biostrings_2.48.0           compiler_3.5.0
## [51] GenomeInfoDb_1.16.0         rlang_0.2.1
## [53] grid_3.5.0                  RCurl_1.95-4.10
## [55] rstudioapi_0.7              bitops_1.0-6
## [57] rmarkdown_1.10              gtable_0.2.0
## [59] DBI_1.0.0                   reshape2_1.4.3
## [61] R6_2.2.2                    GenomicAlignments_1.16.0
## [63] lubridate_1.7.4             rtracklayer_1.40.3
## [65] bit_1.1-14                  utf8_1.1.4
## [67] bindr_0.1.1                 rprojroot_1.3-2
## [69] stringi_1.2.3               Rcpp_0.12.17
## [71] tidyselect_0.2.4
```