Using branchpointer for annotation of intronic human splicing branchpoints

Beth Signal

October 29, 2019

Contents

1 Introduction

The spliceosome mediates the formation of an intron lariat though inter-action between the 5' splice site and branchpoint (Will and Luhrmann, 2011). A subsequent reaction at the 3' SS then removes the intron lariat, producing a spliced RNA product. Mapping of branchpoints generally requires sequencing of the intron lariat following cDNA synthesis (Gao et al., 2008; Taggart et al., 2012). However, intron lariats are rapidly de-branched and degraded, and much less abundant than the spliced RNA, resulting in the poor recovery of elements using sequencing. Most recently, Mercer et al. (2015) employed a targeted sequencing approach

to identify 59,359 branchpoints in 17.4% of annotated human gene introns. Whilst this constituted the largest annotation to date, the identification of branchpoints was restricted to highly-expressed genes with sufficient sequence coverage.

To address this limitation, and expand branchpoint annotations across the human genome, we have developed a machine-learning based model of branchpoints trained with this empirical annotation (Signal et al., 2016). This model requires only genomic sequence and exon annotations, and exhibits no discernible bias to gene type or expression, and can be applied using the R package, branchpointer. Aberrant splicing is known to lead to many human diseases (Singh and Cooper, 2012), however prediction of intronic variant effects have been typically limited to splice site alterations (McLaren et al., 2016; Wang et al., 2010). Therefore, in addition to annotation of branchpoints, branchpointer allows users to assess the effects of intronic mutations on branchpoint architecture.

Gao,K. et al. (2008) Human branch point consensus sequence is yUnAy. Nucleic Acids Res., 36, 2257–67.

McLaren,W. et al. (2016) The Ensembl Variant Effect Predictor. Genome Biol., 17, 122.

Mercer,T.R. et al. (2015) Genome-wide discovery of human splicing branchpoints. Genome Res., 25, 290–303.

Signal,B. et al. (2016) Machine-learning annotation of human splicing branchpoints. BioRxiv. doi: 10.1101/094003.

Singh,R.K. and Cooper,T.A. (2012) Pre-mRNA splicing in disease and therapeutics. Trends Mol. Med., 18, 472–482.

Taggart,A.J. et al. (2012) Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nat. Struct. Mol. Biol., 19, 719–21.

Wang,K. et al. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res., 38, e164.

Will,C.L. and Luhrmann,R. (2011) Spliceosome structure and function. Cold Spring Harb. Perspect. Biol., 3, a003707.

2 Preparation

2.1 Download genome annotations

Branchpointer requires a genome annotation GTF file for branchpoint annotation. We will be using the GENCODE annotation (http://www.gencodegenes.org/releases/current.html) as an example, although others and custom annotations can be used.

Create or move to a working directory where these files can be stored. Note that GTFs can be large files (over 1GB) when uncompressed.

```
wget ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_26/gencode.v26.annotation.gtf.gz

gunzip gencode.v26.annotation.gtf.gz
```

branchpointer requires either a BSGenome object, or a genome .fa file for sequence retrieval. The genome must correspond to the gtf used – i.e. gencodev26 uses GRCh38.

load a *BSGenome*:

```
library(BSgenome.Hsapiens.UCSC.hg38)
g <- BSgenome.Hsapiens.UCSC.hg38::BSgenome.Hsapiens.UCSC.hg38
```

or download .fa:

```
wget ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_26/GRCh38.primary_assembly.genome.fa.gz
GRCh38.primary_assembly.genome.fa.gz
```

2.2 Read in exon annotations

Start by loading branchpointer.

```
library(branchpointer)
```

gtfToExons will generate an exon annotation GRanges object from a gtf. To load in the gtf downloaded from the preparation section:

```
exons <- gtfToExons("gencode.v26.annotation.gtf")
```

We will load in a small gtf from the package data for the following examples.

```
smallExons <- system.file("extdata","gencode.v26.annotation.small.gtf",
                          package = "branchpointer")
exons <- gtfToExons(smallExons)
```

3 Branchpoint annotations in intronic regions

3.1 Read query and calculate location attributes

Query regions must contain a branchpoint window - that is the region located at -18 to -44 from the 3' splice site. Each region given will be treated as only one query, and associated with the closest 3' exon. To cover multiple 3'exons, please provide branchpointer with separate region queries. For known regions, queries can be supplied as a table with the following formatting:

```
queryIntronFile <- system.file("extdata","intron_example.txt",
                               package = "branchpointer")
queryIntronTable <- read.delim(queryIntronFile)
head(queryIntronTable)
```

```
##             id chromosome    start      end strand
## 1 BRCA1_intron      chr17 43045820 43045846      -
## 2 BRCA2_intron      chr13 32346783 32346809      +
```

Query files can be read into branchpointer using readQueryFile(), and location information retrieved for these sequences:

```
queryIntron <- readQueryFile(queryIntronFile,
                             queryType = "region",
```

```
                              exons = exons)
head(queryIntron)

## GRanges object with 2 ranges and 6 metadata columns:
##       seqnames            ranges strand |           id to_3prime to_5prime
##          <Rle>         <IRanges>  <Rle> |  <character> <numeric> <numeric>
##   [1]    chr17 43045820-43045846      - | BRCA1_intron        18      1823
##   [2]    chr13 32346783-32346809      + | BRCA2_intron        18      2156
##       same_gene       exon_3prime       exon_5prime
##       <logical>       <character>       <character>
##   [1]      TRUE ENSE00001814242.1 ENSE00003687053.1
##   [2]      TRUE ENSE00000939171.1 ENSE00003753873.1
##   -------
##   seqinfo: 2 sequences from an unspecified genome; no seqlengths
```

Alternatively, to generate branchpoint window region queries by from the gtf annotation, the exon object can be used:

Note that when searching for genes, transcripts, or exons, the ids used must be in the same format as in the annotation file (i.e. ENSG00000XXXXXX.X, ENST00000XXXXXX.X, ENSE00000XXXXXX.X). If you are unsure of an id, aliases can typically be found through ensembl (ensembl.org), or through a biomaRt query.

```
queryIntronFromGTF <- makeBranchpointWindowForExons("ENSE00000939171.1",
                                                    idType = "exon_id",
                                                    exons = exons)
head(queryIntronFromGTF)

## GRanges object with 1 range and 12 metadata columns:
##       seqnames            ranges strand |            gene_id      gene_type
##          <Rle>         <IRanges>  <Rle> |        <character>    <character>
##   [1]    chr13 32346783-32346809      + | ENSG00000139618.14 protein_coding
##           transcript_id transcript_type           exon_id exon_number
##             <character>     <character>       <character> <character>
##   [1] ENST00000380152.7  protein_coding ENSE00000939171.1          13
##       to_3prime to_5prime same_gene       exon_3prime       exon_5prime
##       <numeric> <integer> <logical>       <character>       <character>
##   [1]        18      2156      TRUE ENSE00000939171.1 ENSE00000939169.1
##                       id
##             <character>
##   [1] ENSE00000939171.1
##   -------
##   seqinfo: 4 sequences from an unspecified genome; no seqlengths

# for multiple ids:
queryIntronFromGTF <- makeBranchpointWindowForExons(c("ENSE00000939171.1",
                                                      "ENSE00001814242.1"),
                                                    idType = "exon_id",
                                                    exons = exons)
head(queryIntronFromGTF)

## GRanges object with 2 ranges and 12 metadata columns:
##       seqnames            ranges strand |            gene_id      gene_type
##          <Rle>         <IRanges>  <Rle> |        <character>    <character>
```

```
##   [1]    chr13 32346783-32346809      + | ENSG00000139618.14 protein_coding
##   [2]    chr17 43045820-43045846      - | ENSG00000012048.20 protein_coding
##          transcript_id transcript_type           exon_id exon_number
##            <character>     <character>       <character> <character>
##   [1] ENST00000380152.7  protein_coding ENSE00000939171.1          13
##   [2] ENST00000357654.7  protein_coding ENSE00001814242.1          23
##       to_3prime to_5prime same_gene       exon_3prime       exon_5prime
##       <numeric> <integer> <logical>       <character>       <character>
##   [1]        18      2156      TRUE ENSE00000939171.1 ENSE00000939169.1
##   [2]        18      1823      TRUE ENSE00001814242.1 ENSE00003687053.1
##                       id
##             <character>
##   [1] ENSE00000939171.1
##   [2] ENSE00001814242.1
##   -------
##   seqinfo: 4 sequences from an unspecified genome; no seqlengths
```

3.1.1 Using bedtools with a genome .fa file

During the prediction step, if a BSGenome object is not specified, sequences covering each site +/- 250 nt can be retrieved using bedtools. The absolute location of the bedtools binary must be provided for calls from within R. To find the location of your installed bedtools binary, using the command line type:

```
which bedtools
```

If chromosome names in the .fa genome file do not match those in the query (i.e chr1 in query, 1 in .fa), the argument rm_chr should be set to FALSE.

3.2 Predict branchpoint probabilities

Branchpoint probability scores can now be evaluated using the branchpointer model. This will generate a new GRanges object with a row for each site (of 27) in branchpoint window regions. If a SNP query type is provided (See next section), this will also perform an in silico mutation of the sequence.

We recommend use of the cut-off probability 0.52 to distinguish branchpoints and non-branchpoint sites. U2 binding energy can be used as a measurement of branchpoint strength when the probability score is above the cut-off.

All features required for the model to predict branchpoint probability are contained within the output object, along with the score and U2 binding energy.

```
branchpointPredictionsIntron <- predictBranchpoints(queryIntron,
                                                    queryType = "region",
                                                    BSgenome = g)
head(branchpointPredictionsIntron)
```

```
## GRanges object with 6 ranges and 31 metadata columns:
##       seqnames            ranges strand |           id to_3prime to_5prime
##          <Rle>         <IRanges>  <Rle> |  <character> <numeric> <numeric>
```

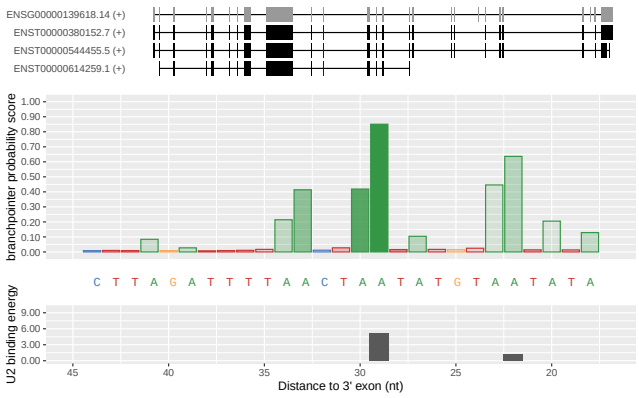
```
##   [1]    chr17 43045820-43045846      - | BRCA1_intron        18      1823
##   [2]    chr13 32346783-32346809      + | BRCA2_intron        18      2156
##   [3]    chr17 43045820-43045846      - | BRCA1_intron        18      1823
##   [4]    chr13 32346783-32346809      + | BRCA2_intron        18      2156
##   [5]    chr17 43045820-43045846      - | BRCA1_intron        18      1823
##   [6]    chr13 32346783-32346809      + | BRCA2_intron        18      2156
##       same_gene       exon_3prime       exon_5prime
##       <logical>       <character>       <character>
##   [1]      TRUE ENSE00001814242.1 ENSE00003687053.1
##   [2]      TRUE ENSE00000939171.1 ENSE00003753873.1
##   [3]      TRUE ENSE00001814242.1 ENSE00003687053.1
##   [4]      TRUE ENSE00000939171.1 ENSE00003753873.1
##   [5]      TRUE ENSE00001814242.1 ENSE00003687053.1
##   [6]      TRUE ENSE00000939171.1 ENSE00003753873.1
##                                                  seq      status
##                                          <character> <character>
##   [1] TCCAGGAGAATGAATTGACACTAATCTCTGCTTGTGTTCTCTGTCTCCAG         REF
##   [2] TATTCTCTTAGATTTTAACTAATATGTAATATAAAATAATTGTTTCCTAG         REF
##   [3] TCCAGGAGAATGAATTGACACTAATCTCTGCTTGTGTTCTCTGTCTCCAG         REF
##   [4] TATTCTCTTAGATTTTAACTAATATGTAATATAAAATAATTGTTTCCTAG         REF
##   [5] TCCAGGAGAATGAATTGACACTAATCTCTGCTTGTGTTCTCTGTCTCCAG         REF
##   [6] TATTCTCTTAGATTTTAACTAATATGTAATATAAAATAATTGTTTCCTAG         REF
##       to_3prime_point to_5prime_point test_site seq_pos0 seq_pos1 seq_pos2
##             <integer>       <numeric> <numeric> <factor> <factor> <factor>
##   [1]              44            1797  43045846        A        G        A
##   [2]              44            2130  32346783        C        T        T
##   [3]              43            1798  43045845        G        A        A
##   [4]              43            2131  32346784        T        T        A
##   [5]              42            1799  43045844        A        A        T
##   [6]              42            2132  32346785        T        A        G
##       seq_pos3 seq_pos4 seq_pos5 seq_neg1 seq_neg2 seq_neg3 seq_neg4
##       <factor> <factor> <factor> <factor> <factor> <factor> <factor>
##   [1]        A        T        G        G        G        A        C
##   [2]        A        G        A        T        C        T        T
##   [3]        T        G        A        A        G        G        A
##   [4]        G        A        T        C        T        C        T
##   [5]        G        A        A        G        A        G        G
##   [6]        A        T        T        T        C        T        C
##       seq_neg5 canon_hit1 canon_hit2 canon_hit3 canon_hit4 canon_hit5
##       <factor>  <numeric>  <numeric>  <numeric>  <numeric>  <numeric>
##   [1]        C         42         53         62         81         83
##   [2]        A          3         42         54         61         87
##   [3]        C         41         52         61         80         82
##   [4]        T          2         41         53         60         86
##   [5]        A         40         51         60         79         81
##   [6]        T          1         40         52         59         85
##       ppt_start ppt_run_length    branchpoint_prob U2_binding_energy
##       <numeric>      <numeric>           <numeric>         <numeric>
##   [1]        19             24  0.0316661369143931               0.5
##   [2]        16              7 0.00860352554505515               0.5
##   [3]        18             24   0.008682352791488               0.1
```

```
##   [4]        15              7 0.00991772863948898              1.2
##   [5]        17             24  0.0297617254720141              1.4
##   [6]        14              7 0.00904756761173039              1.2
##   -------
##   seqinfo: 2 sequences from an unspecified genome; no seqlengths
```

The window scores can be plotted using plotBranchpointWindow(), with optional plots for gene and isoform structure. The main panel displays the probability scores of each site within the branchpoint window. The opacity of the bars is representative of relative U2 binding energy (darker = stronger), and the lower panel shows U2 binding energy for all sites above the provided probability cutoff.

BRCA2 intron (ENSE00000939169.1 - ENSE00000939171.1):

```
plotBranchpointWindow(queryIntron$id[2],
                      branchpointPredictionsIntron,
                      probabilityCutoff = 0.52,
                      plotMutated = FALSE,
                      plotStructure = TRUE,
                      exons = exons)
```



4 Effects of SNPs on branchpoint annotations

In addition to locating branchpoints in intronic windows, branchpointer can be used to evaluate the local effects of SNPs on branchpoints. The general workflow is the same as for annotation of intronic windows, however queryType="SNP" must be used.

4.1 Read query and calculate location attributes

Query SNPs should be located nearby a branchpoint window to have any potential effects on branchpoint architecture SNP queries can be supplied as a table formatted as follows:

```
querySNPFile <- system.file("extdata","SNP_example.txt",
                            package = "branchpointer")
querySNPTable <- read.delim(querySNPFile)
head(querySNPTable)
```

```
##           id chromosome chrom_start strand ref_allele alt_allele
## 1 rs786205083       chr2    71590178      +          A          G
## 2 rs587776767      chr11     2165787      -          A          T
```

When reading in exceptionally large numbers of SNPs, it is recommended to set filter = TRUE. This adds a pre-filtering step which removes any SNPs not located in an intron or 50nt from a 3' exon.

Each SNP will be associated with the closest 3' exon. If SNPs are distal from branchpoint windows, the max_dist argument will remove any greater than the specified distance. Filtering prior to exon associations can therefore speed up processing in instances where it is unknown if the majority of SNPs fall nearby branchpoint windows.

```
querySNP <- readQueryFile(querySNPFile,
                          queryType = "SNP",
                          exons = exons,
                          filter = TRUE)
head(querySNP)
```

```
## GRanges object with 2 ranges and 8 metadata columns:
##       seqnames    ranges strand |              id  ref_allele  alt_allele
##          <Rle> <IRanges>  <Rle> |     <character> <character> <character>
##   [1]     chr2  71590178      + | rs786205083_pos           A           G
##   [2]    chr11   2165787      - | rs587776767_neg           A           T
##       to_3prime to_5prime same_gene       exon_3prime       exon_5prime
##       <numeric> <numeric> <logical>       <character>       <character>
##   [1]        33       492      TRUE ENSE00003642866.1 ENSE00003663865.1
##   [2]        24        66      TRUE ENSE00003550033.1 ENSE00001878270.1
##   -------
##   seqinfo: 2 sequences from an unspecified genome; no seqlengths
```

Queries can be provided as stranded or unstranded. In the case of unstranded queries, any value except "+" or "-" will cause branchpointer to run on both strands.

Alternatively, appropriate attributes can be pulled from biomart when a list of refsnp ids is provided:

```
library(biomaRt)
mart <- useMart("ENSEMBL_MART_SNP", dataset="hsapiens_snp",host="www.ensembl.org")
querySNP <- makeBranchpointWindowForSNP(c("rs587776767","rs786205083"),
                                        mart.snp = mart,
                                        exons = exons,
                                        filter  = FALSE)
head(querySNP)
```

```
## GRanges object with 2 ranges and 8 metadata columns:
##       seqnames    ranges strand |              id  ref_allele  alt_allele
##          <Rle> <IRanges>  <Rle> |     <character> <character> <character>
##   [1]     chr2  71590178      + | rs786205083_pos           A           G
##   [2]    chr11   2165787      - | rs587776767_neg           A           T
##       to_3prime to_5prime same_gene       exon_3prime       exon_5prime
##       <numeric> <numeric> <logical>       <character>       <character>
##   [1]        33       492      TRUE ENSE00003642866.1 ENSE00003663865.1
##   [2]        24        66      TRUE ENSE00003550033.1 ENSE00001878270.1
```

```
##   -------
##   seqinfo: 2 sequences from an unspecified genome; no seqlengths
```

By default, all SNPs retrieved will be unstranded, and hence further processing will be done on both strands

4.2 Predict branchpoint probabilities

Using a .fa and bedtools:

```
branchpointPredictionsSNP <- predictBranchpoints(querySNP,
                                        queryType = "SNP",
                                        genome = "GRCh38.primary_assembly.genome.fa",
                                        bedtoolsLocation="/Apps/bedtools2/bin/bedtools")
```

Using a BSgenome:

```
#for query SNPs
branchpointPredictionsSNP <- predictBranchpoints(querySNP,
                                        queryType = "SNP",
                                        BSgenome = g)
head(branchpointPredictionsSNP)
```

```
## GRanges object with 6 ranges and 33 metadata columns:
##       seqnames    ranges strand |              id  ref_allele  alt_allele
##          <Rle> <IRanges>  <Rle> |     <character> <character> <character>
##   [1]     chr2  71590178      + | rs786205083_pos           A           G
##   [2]    chr11   2165787      - | rs587776767_neg           A           T
##   [3]     chr2  71590178      + | rs786205083_pos           A           G
##   [4]    chr11   2165787      - | rs587776767_neg           A           T
##   [5]     chr2  71590178      + | rs786205083_pos           A           G
##   [6]    chr11   2165787      - | rs587776767_neg           A           T
##       to_3prime to_5prime same_gene       exon_3prime       exon_5prime
##       <numeric> <numeric> <logical>       <character>       <character>
##   [1]        33       492      TRUE ENSE00003642866.1 ENSE00003663865.1
##   [2]        24        66      TRUE ENSE00003550033.1 ENSE00001878270.1
##   [3]        33       492      TRUE ENSE00003642866.1 ENSE00003663865.1
##   [4]        24        66      TRUE ENSE00003550033.1 ENSE00001878270.1
##   [5]        33       492      TRUE ENSE00003642866.1 ENSE00003663865.1
##   [6]        24        66      TRUE ENSE00003550033.1 ENSE00001878270.1
##                                                    seq      status
##                                              <character> <character>
##   [1] AACCACTCCAGCCACTCACTCTGGCACCTCTGTTTTTTCCCTTGGTGAAG         REF
##   [2] CCGGTGGGCGGCAGCTGTCTCTGGGCTGATGCTGCCCGGCTTCCCCGCAG         REF
##   [3] AACCACTCCAGCCACTCACTCTGGCACCTCTGTTTTTTCCCTTGGTGAAG         REF
##   [4] CCGGTGGGCGGCAGCTGTCTCTGGGCTGATGCTGCCCGGCTTCCCCGCAG         REF
##   [5] AACCACTCCAGCCACTCACTCTGGCACCTCTGTTTTTTCCCTTGGTGAAG         REF
##   [6] CCGGTGGGCGGCAGCTGTCTCTGGGCTGATGCTGCCCGGCTTCCCCGCAG         REF
##       to_3prime_point to_5prime_point test_site seq_pos0 seq_pos1 seq_pos2
##             <integer>       <numeric> <numeric> <factor> <factor> <factor>
##   [1]              44             481  71590167        T        C        C
```

```
##   [2]             44             46   2165807        G        G        C
##   [3]             43            482  71590168        C        C        A
##   [4]             43             47   2165806        G        C        G
##   [5]             42            483  71590169        C        A        G
##   [6]             42             48   2165805        C        G        G
##       seq_pos3 seq_pos4 seq_pos5 seq_neg1 seq_neg2 seq_neg3 seq_neg4
##       <factor> <factor> <factor> <factor> <factor> <factor> <factor>
##   [1]        A        G        C        C        A        C        C
##   [2]        G        G        C        G        T        G        G
##   [3]        G        C        C        T        C        A        C
##   [4]        G        C        A        G        G        T        G
##   [5]        C        C        A        C        T        C        A
##   [6]        C        A        G        G        G        G        T
##       seq_neg5 canon_hit1 canon_hit2 canon_hit3 canon_hit4 canon_hit5
##       <factor>  <numeric>  <numeric>  <numeric>  <numeric>  <numeric>
##   [1]        A          3         42         80        107        124
##   [2]        C          6         42         63         78         81
##   [3]        C          2         41         79        106        123
##   [4]        G          5         41         62         77         80
##   [5]        C          1         40         78        105        122
##   [6]        G          4         40         61         76         79
##       ppt_start ppt_run_length    branchpoint_prob U2_binding_energy
##       <numeric>      <numeric>           <numeric>         <numeric>
##   [1]        19             19 0.00779341576259746               1.4
##   [2]        18              7 0.00761276989591145                 0
##   [3]        18             19  0.0109008696263172               1.7
##   [4]        18              7  0.0081540780039032               0.2
##   [5]        17             19  0.0101902457784256                 2
##   [6]        18              7 0.00868429623612561               0.8
##   -------
##   seqinfo: 2 sequences from an unspecified genome; no seqlengths
```

```
#to summarise effects:
querySNPSummary <- predictionsToSummary(querySNP,branchpointPredictionsSNP)
head(querySNPSummary)
```

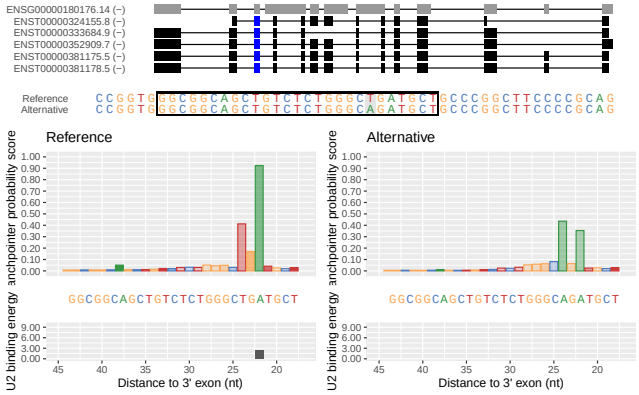
```
## GRanges object with 2 ranges and 18 metadata columns:
##       seqnames    ranges strand |              id  ref_allele  alt_allele
##          <Rle> <IRanges>  <Rle> |     <character> <character> <character>
##   [1]     chr2  71590178      + | rs786205083_pos           A           G
##   [2]    chr11   2165787      - | rs587776767_neg           A           T
##       to_3prime to_5prime same_gene       exon_3prime       exon_5prime
##       <numeric> <numeric> <logical>       <character>       <character>
##   [1]        33       492      TRUE ENSE00003642866.1 ENSE00003663865.1
##   [2]        24        66      TRUE ENSE00003550033.1 ENSE00001878270.1
##       BP_num_REF BP_num_ALT deleted_n created_n dist_to_BP_REF
##        <numeric>  <numeric> <numeric> <numeric>      <numeric>
##   [1]          2          1         1         0              0
##   [2]          1          0         1         0              2
##       dist_to_BP_ALT      max_prob_REF      max_prob_ALT max_U2_REF
##            <numeric>         <numeric>         <numeric>  <numeric>
##   [1]              8 0.922264790347396  0.56160979432899          2
```

```
##   [2]          <NA> 0.923687584223331 0.435892118572635        2.5
##       max_U2_ALT
##        <numeric>
##   [1]        0.5
##   [2]       <NA>
##   -------
##   seqinfo: 2 sequences from an unspecified genome; no seqlengths
```

The window scores in the reference and alternative sequences can be visualised using plotBranchpointWindow().

rs587776767 in TH intron

```
plotBranchpointWindow(querySNP$id[2],
                      branchpointPredictionsSNP,
                      probabilityCutoff = 0.52,
                      plotMutated = TRUE,
                      plotStructure = TRUE,
                      exons = exons)
```



5 Performance

Branchpointer is vectorised where possible to decrease run times.

When reading in SNP queries, a prefiltering step can be applied by setting filter=TRUE. This step adds less than 10 seconds to readQueryFile(), and can reduce run times when the locations of SNPs are unknown [Figure 1].

Figure 1: Time taken for readQueryFile() on SNP query files

predictBranchpoints() is the rate limiting step, taking 55 seconds per 100 region queries. This can be lowered using parallelisation by setting useParallel=TRUE and specifying the number of cores [Figure 2].

Figure 2: Time taken for predictBranchpoints() on region queries

5.1 Example run times

```
# Step times for annotating branchpoints in introns:
gtfToExons()
# user  system elapsed
# 41.385   3.848  47.096

# Set 1. 294 lincRNA introns on chr22:
makeBranchpointWindowForExons()
# user  system elapsed
# 0.196   0.024   0.226
predictBranchpoints()
# user  system elapsed
# 208.934   4.157 225.849

# Set 2. 3693 protein coding exons on chr22:
makeBranchpointWindowForExons()
# user  system elapsed
# 0.245   0.013   0.261
predictBranchpoints()
# user   system  elapsed
# 2332.519   38.266 2482.032

# Step times for annotating branchpoints with SNPs:
# 29899 GWAS SNPS
readQueryFile(filter = TRUE)
# user  system elapsed
# 5.997   1.608   7.773
readQueryFile(filter = FALSE)
# user  system elapsed
# 1.744   0.427   2.339

# 298 filtered SNPS
predictBranchpoints()
# user  system elapsed
# 172.495   2.485 181.876

predictionsToSummary()
# user  system elapsed
# 0.057   0.003   0.061
```

Example scripts used to test times are found in inst/scripts, and were run on a 2.4GHz Macbook Pro with 8GB RAM

6 Session info

```
sessionInfo()
```

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.3 LTS
##
## Matrix products: default
## BLAS:   /home/biocbuild/bbs-3.10-bioc/R/lib/libRblas.so
## LAPACK: /home/biocbuild/bbs-3.10-bioc/R/lib/libRlapack.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] biomaRt_2.42.0                  branchpointer_1.12.0
##  [3] caret_6.0-84                    ggplot2_3.2.1
##  [5] lattice_0.20-38                 BSgenome.Hsapiens.UCSC.hg38_1.4.1
##  [7] BSgenome_1.54.0                 rtracklayer_1.46.0
##  [9] Biostrings_2.54.0               XVector_0.26.0
## [11] GenomicRanges_1.38.0            GenomeInfoDb_1.22.0
## [13] IRanges_2.20.0                  S4Vectors_0.24.0
## [15] BiocGenerics_0.32.0             knitr_1.25
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-141               bitops_1.0-6
##  [3] matrixStats_0.55.0         lubridate_1.7.4
##  [5] bit64_0.9-7                httr_1.4.1
##  [7] progress_1.2.2             tools_3.6.1
##  [9] backports_1.1.5            R6_2.4.0
## [11] rpart_4.1-15               DBI_1.0.0
## [13] lazyeval_0.2.2             colorspace_1.4-1
## [15] nnet_7.3-12                gbm_2.1.5
## [17] withr_2.1.2                gridExtra_2.3
## [19] prettyunits_1.0.2          tidyselect_0.2.5
## [21] curl_4.2                   bit_1.1-14
## [23] compiler_3.6.1             Biobase_2.46.0
## [25] DelayedArray_0.12.0        labeling_0.3
## [27] scales_1.0.0               askpass_1.1
## [29] rappdirs_0.3.1             stringr_1.4.0
## [31] digest_0.6.22              Rsamtools_2.2.0
```

```
## [33] rmarkdown_1.16              pkgconfig_2.0.3
## [35] htmltools_0.4.0             dbplyr_1.4.2
## [37] highr_0.8                   rlang_0.4.1
## [39] RSQLite_2.1.2               generics_0.0.2
## [41] BiocParallel_1.20.0         dplyr_0.8.3
## [43] ModelMetrics_1.2.2          RCurl_1.95-4.12
## [45] magrittr_1.5                GenomeInfoDbData_1.2.2
## [47] Matrix_1.2-17               Rcpp_1.0.2
## [49] munsell_0.5.0               stringi_1.4.3
## [51] yaml_2.2.0                  MASS_7.3-51.4
## [53] SummarizedExperiment_1.16.0 zlibbioc_1.32.0
## [55] BiocFileCache_1.10.0        plyr_1.8.4
## [57] recipes_0.1.7               grid_3.6.1
## [59] blob_1.2.0                  crayon_1.3.4
## [61] cowplot_1.0.0               splines_3.6.1
## [63] hms_0.5.1                   zeallot_0.1.0
## [65] pillar_1.4.2                reshape2_1.4.3
## [67] codetools_0.2-16            XML_3.98-1.20
## [69] glue_1.3.1                  evaluate_0.14
## [71] data.table_1.12.6           BiocManager_1.30.9
## [73] vctrs_0.2.0                 foreach_1.4.7
## [75] openssl_1.4.1               gtable_0.3.0
## [77] purrr_0.3.3                 kernlab_0.9-27
## [79] assertthat_0.2.1            xfun_0.10
## [81] gower_0.2.1                 prodlim_2018.04.18
## [83] class_7.3-15                survival_2.44-1.1
## [85] timeDate_3043.102           tibble_2.1.3
## [87] iterators_1.0.12            GenomicAlignments_1.22.0
## [89] AnnotationDbi_1.48.0        memoise_1.1.0
## [91] lava_1.6.6                  BiocStyle_2.14.0
## [93] ipred_0.9-9
```