Building a gene annotation object

Jean-Philippe Fortin, Luke Hoberecht

Introduction

In this tutorial, we describe the process for making and using
rich gene annotation objects to be used throughout the crisprVerse ecosystem. Such objects enable users to retrieve coordinates of transcripts, exons, etc. Those objects are also used by several functions in the crisprDesign package to add gene annotations to both gRNA on-targets and off-targets. This is what the `txObject` argument in many of the functions expect.

We will also describe the process for constructing and using a transcription start site (TSS) annotation object (`tssObject` argument in many of the functions).

Installation

See the Installation tutorial to learn how to install the packages `crisprDesign` and `crisprDesignData` required in this tutorial.

Getting started

The packages can be loaded into an R session in the usual way:

```r
library(crisprDesign)
library(crisprDesignData)
```

Building gene annotation objects

In the crisprVerse, we represent gene annotations using `GRangesList` object, and this can be easily constructed using the commonly-used Bioconductor objects `TxDb` (see the GenomicFeatures package to learn more about `TxDb` objects). We will now show several ways of constructing such objects.

Building a GRangesList from Ensembl

We construct a gene annotation object for the human genome using the Ensembl release 104 (hg38). This can be done using the function `getTxDb` in `crisprDesign`:

```r
txdb <- getTxDb(organism="Homo sapiens", release=104)
```

This may take several minutes, and note that this requires an internet connection. In case it times out, one can increase the timeout option using the following:

```r
options(timeout = max(10000000, getOption("timeout")))
```

Once obtained, we can convert the object into a `GRangesList` using the function `TxDb2GRangesList` from `crisprDesign`:

```
grList <- TxDb2GRangesList(txdb)
```

We will specify that the genome is hg38:

```
GenomeInfoDb::genome(grList) <- "hg38"
```

And that's it! The `grList` object contains all of the information about the Ensembl release 104 gene model, and is ready to be used in the crisprVerse. Let's take a quick look at our gene annotation object:

```
names(grList)
```

```
## [1] "transcripts" "exons"       "cds"         "fiveUTRs"    "threeUTRs"  
## [6] "introns"     "tss"
```

```
grList$transcripts
```

```
## GRanges object with 111751 ranges and 14 metadata columns:
##     seqnames      ranges strand |           tx_id         gene_id
##        <Rle>   <IRanges>  <Rle> |     <character>     <character>
##            1 11869-14409      + | ENST00000456328 ENSG00000223972
##            1 12010-13670      + | ENST00000450305 ENSG00000223972
##            1 29554-31097      + | ENST00000473358 ENSG00000243485
##            1 30267-31109      + | ENST00000469289 ENSG00000243485
##            1 30366-30503      + | ENST00000607096 ENSG00000284332
##   .      ...         ...    ... .             ...             ...
##           MT   5826-5891      - | ENST00000387409 ENSG00000210144
##           MT   7446-7514      - | ENST00000387416 ENSG00000210151
##           MT 14149-14673      - | ENST00000361681 ENSG00000198695
##           MT 14674-14742      - | ENST00000387459 ENSG00000210194
##           MT 15956-16023      - | ENST00000387461 ENSG00000210196
##           protein_id                tx_type gene_symbol     exon_id exon_rank
##          <character>            <character> <character> <character> <integer>
##                 <NA>   processed_transcript     DDX11L1        <NA>      <NA>
##                 <NA> transcribed_unproces..     DDX11L1        <NA>      <NA>
##                 <NA>                 lncRNA MIR1302-2HG        <NA>      <NA>
##                 <NA>                 lncRNA MIR1302-2HG        <NA>      <NA>
##                 <NA>                  miRNA   MIR1302-2        <NA>      <NA>
##   .              ...                    ...         ...         ...       ...
##                 <NA>                Mt_tRNA       MT-TY        <NA>      <NA>
##                 <NA>                Mt_tRNA      MT-TS1        <NA>      <NA>
##      ENSP00000354665         protein_coding      MT-ND6        <NA>      <NA>
##                 <NA>                Mt_tRNA       MT-TE        <NA>      <NA>
##                 <NA>                Mt_tRNA       MT-TP        <NA>      <NA>
##     cds_start   cds_end  tx_start    tx_end   cds_len exon_start   exon_end
##     <integer> <integer> <integer> <integer> <integer>  <integer> <integer>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##   .       ...       ...       ...       ...       ...        ...       ...
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##          <NA>      <NA>      <NA>      <NA>      <NA>       <NA>      <NA>
##   -------
##   seqinfo: 25 sequences (1 circular) from hg38 genome
```

Building a tssObject

Building a TSS annotation object requires only one additional step after constructing the `GRangesList` object described above. This can be obtained using the function `getTssObjectFromTxObject` in `crisprDesign`:

```
tssObject <- getTssObjectFromTxObject(grList)
tssObject
```

```
## GRanges object with 52547 ranges and 5 metadata columns:
##          seqnames    ranges strand |           tx_id         gene_id
##             <Rle> <IRanges>  <Rle> |     <character>     <character>
##    11402        1     65419      + | ENST00000641515 ENSG00000186092
##    11442        1    923923      + | ENST00000616016 ENSG00000187634
##    11444        1    925731      + | ENST00000342066 ENSG00000187634
##    11445        1    960584      + | ENST00000338591 ENSG00000187961
##    11446        1    960639      + | ENST00000622660 ENSG00000187961
##      ...      ...       ...    ... .             ...             ...
##   123058        Y  24047689      - | ENST00000382407 ENSG00000172352
##   123073        Y  24813186      - | ENST00000382365 ENSG00000187191
##   123074        Y  24813186      - | ENST00000315357 ENSG00000187191
##   123075        Y  24813186      - | ENST00000446723 ENSG00000187191
##   123080        Y  25052074      - | ENST00000382287 ENSG00000185894
##          gene_symbol        promoter                     ID
##          <character>     <character>            <character>
##    11402       OR4F5 ENST00000641515 ENSG00000186092_ENST..
##    11442      SAMD11 ENST00000616016 ENSG00000187634_ENST..
##    11444      SAMD11 ENST00000342066 ENSG00000187634_ENST..
##    11445      KLHL17 ENST00000338591 ENSG00000187961_ENST..
##    11446        <NA> ENST00000622660 ENSG00000187961_ENST..
##      ...         ...             ...                    ...
##   123058       CDY1B ENST00000382407 ENSG00000172352_ENST..
##   123073        DAZ3 ENST00000382365 ENSG00000187191_ENST..
##   123074        DAZ3 ENST00000315357 ENSG00000187191_ENST..
##   123075        DAZ3 ENST00000446723 ENSG00000187191_ENST..
##   123080       BPY2C ENST00000382287 ENSG00000185894_ENST..
##   -------
##   seqinfo: 25 sequences (1 circular) from hg38 genome
```

Using gene annotation objects

The gene (or TSS) annotation objects described above are often necessary for the full characterization of CRISPR gRNAs as they as inputs for several of the `crisprDesign` functions, including `queryTxObject`, `queryTssObject`, `addGeneAnnotation`, `addTssAnnotation`, and `addSpacerAlignments`.

For convenience, we provide in the crisprDesignData package precomputed gene annotation for human and mouse:

Object name	Object class	Version	Description
`txdb_human`	`GRangesList`	Release 104	Ensembl gene model for human (hg38/GRCh38)
`txdb_mouse`	`GRangesList`	Release 102	Ensembl gene model for mouse (mm10/GRCm38)

Object name	Object class	Version	Description
`tss_human`	`GRanges`	Release 104	Ensembl-based TSS coordinates for human (hg38/GRCh38)
`tss_mouse`	`GRanges`	Release 102	Ensembl-based TSS coordinates for human (mm10/GRCm38)

Building a gene annotation object from a GFF file

If you have a General Feature Format (GFF) file from which you want to construct the gene annotation object, you can pass this to the `file` argument of the `crisprDesign` function `getTxDb`; this will create the `TxDb` object using the `GenomicFeatures` function `makeTxDbFromGFF`.

Reproducibility

```
sessionInfo()
```

```
## R version 4.2.1 (2022-06-23)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.7
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/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] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] BSgenome.Hsapiens.UCSC.hg38_1.4.4 BSgenome_1.65.2
##  [3] rtracklayer_1.57.0                Biostrings_2.65.2
##  [5] XVector_0.37.0                    GenomicRanges_1.49.1
##  [7] GenomeInfoDb_1.33.5               IRanges_2.31.2
##  [9] S4Vectors_0.35.1                  crisprDesignData_0.99.17
## [11] crisprDesign_0.99.133             crisprScore_1.1.14
## [13] crisprScoreData_1.1.3             ExperimentHub_2.5.0
## [15] AnnotationHub_3.5.0               BiocFileCache_2.5.0
## [17] dbplyr_2.2.1                      BiocGenerics_0.43.1
## [19] crisprBowtie_1.1.1                crisprBase_1.1.5
## [21] crisprVerse_0.99.8                rmarkdown_2.15.2
##
## loaded via a namespace (and not attached):
##  [1] rjson_0.2.21                  ellipsis_0.3.2
##  [3] Rbowtie_1.37.0                bit64_4.0.5
##  [5] lubridate_1.8.0               interactiveDisplayBase_1.35.0
##  [7] AnnotationDbi_1.59.1          fansi_1.0.3
##  [9] xml2_1.3.3                    codetools_0.2-18
## [11] cachem_1.0.6                  knitr_1.40
## [13] jsonlite_1.8.0                Rsamtools_2.13.4
## [15] png_0.1-7                     shiny_1.7.2
## [17] BiocManager_1.30.18           readr_2.1.2
## [19] compiler_4.2.1                httr_1.4.4
## [21] basilisk_1.9.2                assertthat_0.2.1
## [23] Matrix_1.4-1                  fastmap_1.1.0
## [25] cli_3.3.0                     later_1.3.0
## [27] htmltools_0.5.3               prettyunits_1.1.1
## [29] tools_4.2.1                   glue_1.6.2
## [31] GenomeInfoDbData_1.2.8        dplyr_1.0.9
## [33] rappdirs_0.3.3                tinytex_0.41
## [35] Rcpp_1.0.9                    Biobase_2.57.1
## [37] vctrs_0.4.1                   crisprBwa_1.1.3
## [39] xfun_0.32                     stringr_1.4.1
## [41] mime_0.12                     lifecycle_1.0.1
## [43] restfulr_0.0.15               XML_3.99-0.10
## [45] zlibbioc_1.43.0               basilisk.utils_1.9.1
## [47] vroom_1.5.7                   VariantAnnotation_1.43.3
## [49] hms_1.1.2                     promises_1.2.0.1
## [51] MatrixGenerics_1.9.1          parallel_4.2.1
## [53] SummarizedExperiment_1.27.1   RMariaDB_1.2.2
## [55] yaml_2.3.5                    curl_4.3.2
## [57] memoise_2.0.1                 reticulate_1.25
## [59] biomaRt_2.53.2                stringi_1.7.8
## [61] RSQLite_2.2.16                BiocVersion_3.16.0
## [63] highr_0.9                     BiocIO_1.7.1
## [65] randomForest_4.7-1.1          GenomicFeatures_1.49.6
## [67] filelock_1.0.2                BiocParallel_1.31.12
## [69] rlang_1.0.4                   pkgconfig_2.0.3
## [71] matrixStats_0.62.0            bitops_1.0-7
## [73] evaluate_0.16                 lattice_0.20-45
## [75] purrr_0.3.4                   GenomicAlignments_1.33.1
## [77] bit_4.0.4                     tidyselect_1.1.2
## [79] magrittr_2.0.3                R6_2.5.1
## [81] generics_0.1.3                DelayedArray_0.23.1
## [83] DBI_1.1.3                     pillar_1.8.1
## [85] KEGGREST_1.37.3               RCurl_1.98-1.8
## [87] tibble_3.1.8                  dir.expiry_1.5.0
## [89] crayon_1.5.1                  utf8_1.2.2
## [91] tzdb_0.3.0                    progress_1.2.2
## [93] grid_4.2.1                    blob_1.2.3
## [95] digest_0.6.29                 xtable_1.8-4
## [97] httpuv_1.6.5                  Rbwa_1.1.0
```