nearBynding Vignette

Veronica Busa

2022-11-01

Contents

Introduction

nearBynding is a package designed to discern annotated RNA structures proximal to protein binding. nearBynding allows users to annotate RNA structure contexts via CapR or input their own annotations in BED/bedGraph format. It accomodates protein binding information from CLIP-seq experiments as either aligned CLIP-seq reads or peak-called intervals. This vignette will walk you through:

```
* The external software necessary to support this pipeline
* Creating a concatenated transcriptome
* Extracting and folding RNA from the transcriptome via CapR
* Mapping protein-binding and RNA structure information onto a transcriptome
* Running StereoGene to identify RNA structure proximal to protein binding
* Visualizing binding results
* Determining the distance between binding contexts
```

Before running any of these examples, it is highly recommended that the user establishes a new empty directory and uses `setwd()` to make certain that all outputs are deposited there. Some of the functions below create multiple output files.

Installation

```
if(!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("nearBynding")
```

External Software Dependencies

Add all dependency directories to your PATH after installation.

bedtools

bedtools is available for installation here.

Installation instructions will vary by operating system.

CapR

Download the zip file from the github repository, unzip the file, and move it to a directory where you want to permanently store the function.

In the command line, access the folder where the unzipped file is stored.

```
cd CapR-master
make
./CapR
```

If installation is successful, the final line will output

```
Error: The number of argument is invalid.
```

StereoGene

Download the zip file from the github repository, unzip the file, and move it to a directory where you want to permanently store the function.

In the command line, access the folder where the unzipped file is stored.

```
cd stereogene-master
cd src
make
./stereogene -h
```

If installation is successful, the final line will output a menu of argument options.

1. Concatenate the Transcriptome

Although nearBynding is designed to support whole-genome analyses, we will exclusively be evaluating protein-coding genes of chromosomes 4 and 5 through this vignette.

First, a list of transcripts must be identified for analysis. A recommended criterium for selection is that the transcripts be expressed in the cell type used for CLIP-seq experiments. For this vignette, 50 random transcripts have been selected, and the 3'UTR structure of each transcript will be used for analysis, though any region of a transcript such as 5'UTR or CDS could be assessed instead.

This step creates a chain file that will be used to map the selected regions of transcripts end-to-end, excluding the intergenic regions and undesired transcripts that comprise the majority of the genome.

```
# load transcript list
load(system.file("extdata/transcript_list.Rda", package="nearBynding"))
# get GTF file
gtf<-system.file("extdata/Homo_sapiens.GRCh38.chr4&5.gtf",
                    package="nearBynding")

GenomeMappingToChainFile(genome_gtf = gtf,
                         out_chain_name = "test.chain",
                         RNA_fragment = "three_prime_utr",
                         transcript_list = transcript_list,
                         alignment = "hg38")
```

A file containing the sizes of each concatenated chromosome in the chain file will be required for downstream analysis.

```
getChainChrSize(chain = "test.chain",
                out_chr = "chr4and5_3UTR.size")
```

2. Fold RNA via CapR

In order to fold the 3'UTRs, the sequences must first be extracted. This is achieved with the following code:

```
ExtractTranscriptomeSequence(transcript_list = transcript_list,
                    ref_genome = "Homo_sapiens.GRCh38.dna.primary_assembly.fa",
                    genome_gtf = gtf,
                    RNA_fragment = "three_prime_utr",
                    exome_prefix = "chr4and5_3UTR")
```

The reference genome can be found through Ensembl, but for users who do not want to download that 3.2GB file for the sake of this vignette, the FASTA output of the above code is available via:

```
chr4and5_3UTR.fa <- system.file("extdata/chr4and5_3UTR.fa",
                                package="nearBynding")
```

These sequences can then be submitted to CapR for folding.

```
runCapR(in_file = chr4and5_3UTR.fa)
```

Warning: This step can take hours or even days depending on how many transcripts are submitted, how long the RNA fragments are, and the maximum distance between base-paired nucleotides submitted to the CapR algorithm.

3. Map to Transcriptome

The nearBynding pipeline can accomodate either a BAM file of aligned CLIP-seq reads or a BED file of peak intervals. BAM files must be sorted and converted to a bedGraph file first, whereas BED files can be read-in directly.

BAM file input

```
bam <- system.file("extdata/chr4and5.bam", package="nearBynding")
sorted_bam<-sortBam(bam, "chr4and5_sorted")

CleanBAMtoBG(in_bam = sorted_bam)
```

Map Protein Binding and RNA Structure to the Transcriptome

BED or bedGraph files can then be mapped onto the concatenated transcriptome using the chain file created by `GenomeMappingToChainFile()`. This way, only the protein binding from transcriptomic regions of interest will be considered in the final protein binding analysis.

```
liftOverToExomicBG(input = "chr4and5_sorted.bedGraph",
                   chain = "test.chain",
                   chrom_size = "chr4and5_3UTR.size",
                   output_bg = "chr4and5_liftOver.bedGraph")
```

For BED file inputs, use the additional argument `format = "BED"`.

The RNA structure information from the CapR output also needs to mapped onto the concatenated transcriptome. There are six different binding contexts calculated by CapR – *stem*, *hairpin*, *multibranch*, *exterior*, *internal*, and *bulge*. `processCapRout()` parses the CapR output, converts it into six separate bedGraph files, and then performs `liftOverToExomic()` for all the files.

For this sake of this vignette, the CapR outfile is available:

```
processCapRout(CapR_outfile = system.file("extdata/chr4and5_3UTR.out",
                                          package="nearBynding"),
               chain = "test.chain",
               output_prefix = "chr4and5_3UTR",
               chrom_size = "chr4and5_3UTR.size",
               genome_gtf = gtf,
               RNA_fragment = "three_prime_utr")
```

It is possible for users to input their own RNA structure information rather than using CapR. The information should be in BED file format and can be input into `liftOverToExomicBG()` to map the RNA structure data to the same transcriptome as the protein binding data.

4. Calculate Cross-correlation via StereoGene

This is the process that directly answers the question, "What does RNA structure look like around where the protein is binding?" StereoGene is used to calculate the cross-correlation between RNA structure and protein binding in order to visualize the RNA structure landscape surrounding protein binding.

If CapR is used to determine RNA structure, `runStereogeneOnCapR()` initiates StereoGene for a given protein against all CapR-generated RNA structure contexts.

For the sake of this vignette, use `outfiles()` to pull the StereoGene output files to your local directory if you do not want to run StereoGene.

```
runStereogeneOnCapR(protein_file = "chr4and5_liftOver.bedGraph",
                    chrom_size = "chr4and5_3UTR.size",
                    name_config = "chr4and5_3UTR.cfg",
                    input_prefix = "chr4and5_3UTR")
```

If external RNA structure data is being tested, `runStereogene()` initiates analysis for a given protein and a single RNA structure context.

Note: The input track file order matters! The correct order is:

1) RNA structure
2) protein binding

Otherwise, data visualization will be inverted and all downstream analysis will be backwards.

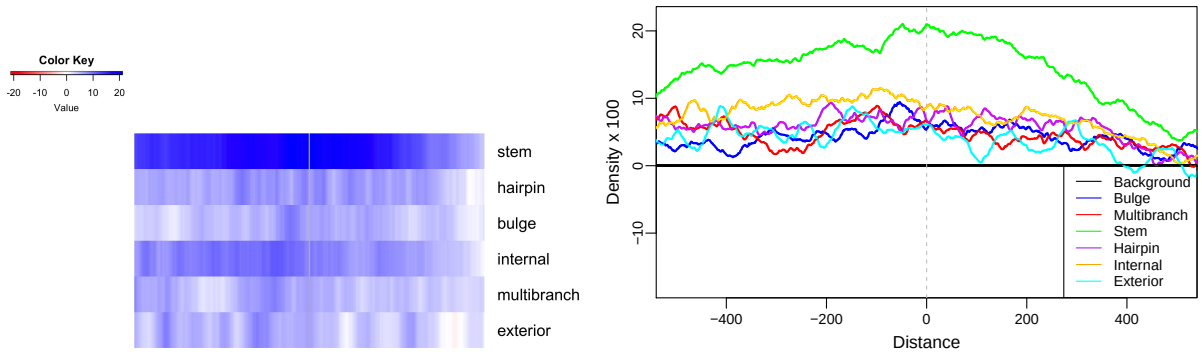
```
runStereogene(track_files = c("chr4and5_3UTR_stem_liftOver.bedGraph",
                              "chr4and5_liftOver.bedGraph"),
              name_config = "chr4and5_3UTR.cfg")
```

5. Visualize Results

The cross-correlation output of StereoGene can be visualized as either a heatmap or a line plot. Examples of both are below.

For CapR-derived RNA structure, all six contexts can be viewed simultaneously.

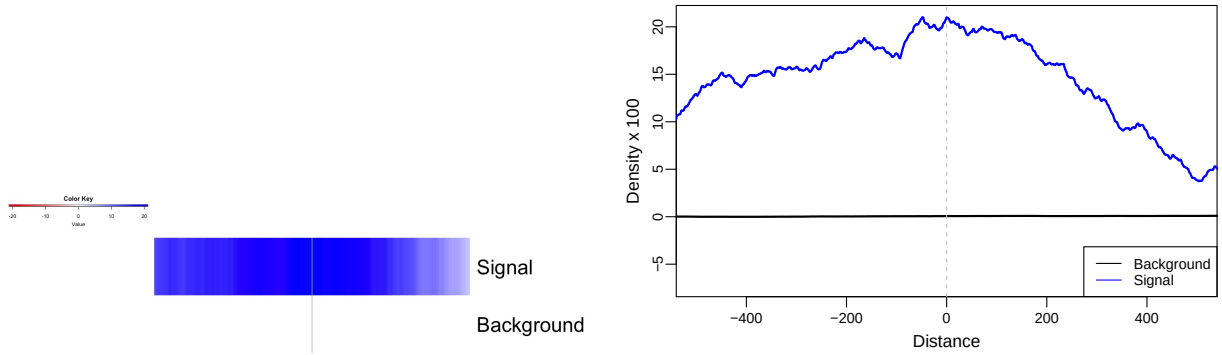
```
visualizeCapRStereogene(CapR_prefix = "chr4and5_3UTR",
                        protein_file = "chr4and5_liftOver",
                        heatmap = TRUE,
                        out_file = "all_contexts_heatmap",
                        x_lim = c(-500, 500))
visualizeCapRStereogene(CapR_prefix = "chr4and5_3UTR",
                        protein_file = "chr4and5_liftOver",
                        x_lim = c(-500, 500),
                        out_file = "all_contexts_line",
                        y_lim = c(-18, 22))
```



Warning: This step may take up to an hour for a full transcriptome.

Alternatively, a single context can be viewed at a time.

```
visualizeStereogene(context_file = "chr4and5_3UTR_stem_liftOver",
                    protein_file = "chr4and5_liftOver",
                    out_file = "stem_line",
                    x_lim = c(-500, 500))
visualizeStereogene(context_file = "chr4and5_3UTR_stem_liftOver",
                    protein_file = "chr4and5_liftOver",
                    heatmap = TRUE,
                    out_file = "stem_heatmap",
                    x_lim = c(-500, 500))
```



Although this specific, limited example does not provide a particularly visually stimulating image, larger data sets (which provide many more StereoGene windows) result in narrower peaks and less noise.

6. Calculate Distance

In order to determine the similarity of two binding contexts, we can calculate the Wasserstein distance between curves. A small value suggests two binding contexts are very similar, whereas larger values suggest substantial differences.

For example, calculate the distance between the stem and hairpin contexts visualized above.

```
bindingContextDistance(RNA_context = "chr4and5_3UTR_stem_liftOver",
                       protein_file = "chr4and5_liftOver",
                       RNA_context_2 = "chr4and5_3UTR_hairpin_liftOver")
#> [1] 11.75237
```

Now compare it to the distance between internal and hairpin contexts.

```
bindingContextDistance(RNA_context = "chr4and5_3UTR_internal_liftOver",
                       protein_file = "chr4and5_liftOver",
                       RNA_context_2 = "chr4and5_3UTR_hairpin_liftOver")
#> [1] 1.788653
```

We can see that the stem context is less similar to the hairpin context than the internal context, and this is reflected in the calculated distances.

Questions? Comments? Please email Veronica Busa at vbusa1@jhmi.edu

```
sessionInfo()
#> R version 4.2.1 (2022-06-23)
#> Platform: x86_64-pc-linux-gnu (64-bit)
#> Running under: Ubuntu 20.04.5 LTS
#>
```

```
#> Matrix products: default
#> BLAS:   /home/biocbuild/bbs-3.16-bioc/R/lib/libRblas.so
#> LAPACK: /home/biocbuild/bbs-3.16-bioc/R/lib/libRlapack.so
#>
#> locale:
#>  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>  [3] LC_TIME=en_GB              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    stats     graphics  grDevices utils     datasets  methods
#> [8] base
#>
#> other attached packages:
#> [1] Rsamtools_2.14.0     Biostrings_2.66.0    XVector_0.38.0
#> [4] GenomicRanges_1.50.0 GenomeInfoDb_1.34.0  IRanges_2.32.0
#> [7] S4Vectors_0.36.0     BiocGenerics_0.44.0  nearBynding_1.8.0
#>
#> loaded via a namespace (and not attached):
#>  [1] MatrixGenerics_1.10.0
#>  [2] Biobase_2.58.0
#>  [3] httr_1.4.4
#>  [4] bit64_4.0.5
#>  [5] R.utils_2.12.1
#>  [6] gtools_3.9.3
#>  [7] assertthat_0.2.1
#>  [8] BiocFileCache_2.6.0
#>  [9] blob_1.2.3
#> [10] GenomeInfoDbData_1.2.9
#> [11] yaml_2.3.6
#> [12] progress_1.2.2
#> [13] pillar_1.8.1
#> [14] RSQLite_2.2.18
#> [15] lattice_0.20-45
#> [16] glue_1.6.2
#> [17] digest_0.6.30
#> [18] transport_0.13-0
#> [19] colorspace_2.0-3
#> [20] R.oo_1.25.0
#> [21] htmltools_0.5.3
#> [22] Matrix_1.5-1
#> [23] TxDb.Hsapiens.UCSC.hg38.knownGene_3.16.0
#> [24] XML_3.99-0.12
#> [25] pkgconfig_2.0.3
#> [26] biomaRt_2.54.0
#> [27] zlibbioc_1.44.0
#> [28] scales_1.2.1
#> [29] BiocParallel_1.32.0
#> [30] tibble_3.1.8
#> [31] KEGGREST_1.38.0
```

```
#> [32] ggplot2_3.3.6
#> [33] generics_0.1.3
#> [34] ellipsis_0.3.2
#> [35] cachem_1.0.6
#> [36] SummarizedExperiment_1.28.0
#> [37] GenomicFeatures_1.50.0
#> [38] cli_3.4.1
#> [39] magrittr_2.0.3
#> [40] crayon_1.5.2
#> [41] memoise_2.0.1
#> [42] evaluate_0.17
#> [43] R.methodsS3_1.8.2
#> [44] fansi_1.0.3
#> [45] gplots_3.1.3
#> [46] xml2_1.3.3
#> [47] data.table_1.14.4
#> [48] tools_4.2.1
#> [49] prettyunits_1.1.1
#> [50] hms_1.1.2
#> [51] BiocIO_1.8.0
#> [52] lifecycle_1.0.3
#> [53] matrixStats_0.62.0
#> [54] stringr_1.4.1
#> [55] plyranges_1.18.0
#> [56] munsell_0.5.0
#> [57] DelayedArray_0.24.0
#> [58] AnnotationDbi_1.60.0
#> [59] compiler_4.2.1
#> [60] caTools_1.18.2
#> [61] rlang_1.0.6
#> [62] grid_4.2.1
#> [63] RCurl_1.98-1.9
#> [64] rjson_0.2.21
#> [65] rappdirs_0.3.3
#> [66] bitops_1.0-7
#> [67] rmarkdown_2.17
#> [68] gtable_0.3.1
#> [69] restfulr_0.0.15
#> [70] codetools_0.2-18
#> [71] DBI_1.1.3
#> [72] curl_4.3.3
#> [73] R6_2.5.1
#> [74] GenomicAlignments_1.34.0
#> [75] knitr_1.40
#> [76] dplyr_1.0.10
#> [77] rtracklayer_1.58.0
#> [78] fastmap_1.1.0
#> [79] bit_4.0.4
#> [80] utf8_1.2.2
#> [81] filelock_1.0.2
#> [82] KernSmooth_2.23-20
#> [83] stringi_1.7.8
#> [84] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2
```

```
#> [85] parallel_4.2.1
#> [86] Rcpp_1.0.9
#> [87] vctrs_0.5.0
#> [88] png_0.1-7
#> [89] dbplyr_2.2.1
#> [90] tidyselect_1.2.0
#> [91] xfun_0.34
```