yCrypticRNAs

Nicole Uwimana, Benjamin Haibe-Kains, François Robert

2016-02-08

Contents

1. Introduction

The yCrypticRNAs package implements methods for cryptic transcription detection from RNA-seq data. Using a probabilistic method based on a cumulative distribution function, the package allows the identification of cryptic transcripts genome-wide. Along with the probabilistic method, the 5'/3' ratio (Cheung et al., 2008) and the 3' enrichment (DeGennaro et al., 2013) methods are implemented, giving the user all the tools for cryptic transcription detection. While previous methods only identify genes within which cryptic transcription occurs, this package -using a bootstrap approach- allow for the prediction of the position of the cryptic transcription start sites.

1.1 Requirements

yCrypticRNAs works only on Unix/linux platforms.

1.2 Installation

Install package

```
> install.packages("yCrypticRNAs")
```

Load the package in your curent workspace:

```
> library("yCrypticRNAs")
```

2. Cryptic scores

2.1 Datasets (coverageDataSet)

yCrypticRNAs was concieve for high-throughput RNA-Seq data. To be able to calculate cryptic scores a coverageDataSet is needed. The coverageDataSet is inherit form a data.table that contains the coverage values for each sample. For more information on the coverage files: `?coverageDataSet`.

```
> #create A coverageDataSet
> # Suppose an RNA-seq experiment that was done in duplicates in wild-type cells and mutant cells.
> # Data were sequenced in strand-specific manner and we wish to use fragments for paired-end reads.
> types <- c("wt", "wt", "mut", "mut")
> bamfiles <- system.file("extdata", paste0(types, "_rep", 1:2,".bam"),
+                         package = "yCrypticRNAs")
> scaling_factors <- c(0.069872847, 0.081113079, 0.088520251, 0.069911116)
>
> data("annotations")
> data <- coverageDataSet(bamfiles = bamfiles, annotations = annotations,
+                         types = types, sf = scaling_factors,
+                         paired_end = TRUE, as_fragments = TRUE)
```

2.2 Datasets (geneCoverage)

geneCoverage dataset contains coverage data for one specific gene without the intronic regions.

```
> # load introns annotations
> data("introns")
>
> # create geneCoverage dataset for FLO8 gene
> flo8 <- gene_coverage(coverageDataSet = data, name = "YER109C", introns = introns)
```

2.3 Calculate the cryptic score for a specific gene.

yCrypticRNAs offers three methods for the calculation of cryptic scores for each gene. The methods are: `zscore_score`, `ratio_score` (Cheung et al., 2008), and `enrichment_score` (DeGennaro et al., 2013). Each function takes an object of type geneCoverage as input.

```
> # using the 3'/5' ratio method
> ratio_score(geneCoverage = flo8)

Cryptic score for YER109C using the 3'/5' ratio method.

cryptic score =  6.42073610683135  [ 6.46 (mut_rep1/wt_rep1) 5.83 (mut_rep2/wt_rep1) 7.07 (mut_rep1/wt_
controls = 1.0948 (wt_rep1/wt_rep2) 0.9134 (wt_rep2/wt_rep1) 1.1089 (mut_rep1/mut_rep2) 0.9018 (mut_rep

> # using the 3' enrichemnt method
> enrichment_score(geneCoverage = flo8)

Cryptic score for YER109C using the  3' enrichment method.

cryptic score =  1.42034676034532
controls = 0.9538 (wt_rep1/wt_rep2) 0.9899 (mut_rep1/mut_rep2)
```

```
> # using the probabilistic method
> zscore_score(geneCoverage = flo8)

Cryptic score for YER109C using the probabilistic method.

cryptic score =  29.5132686915603
controls = -5.8741 (wt_rep2/wt_rep1) -4.3446 (mut_rep2/mut_rep1)
```

2.4 Calculate the cryptic scores for all genes.

The package offers a method for rapid calculation of cryptic score for all genes. For this, you must provide an object of type "CoverageDataSet" and specify the method you want to use.

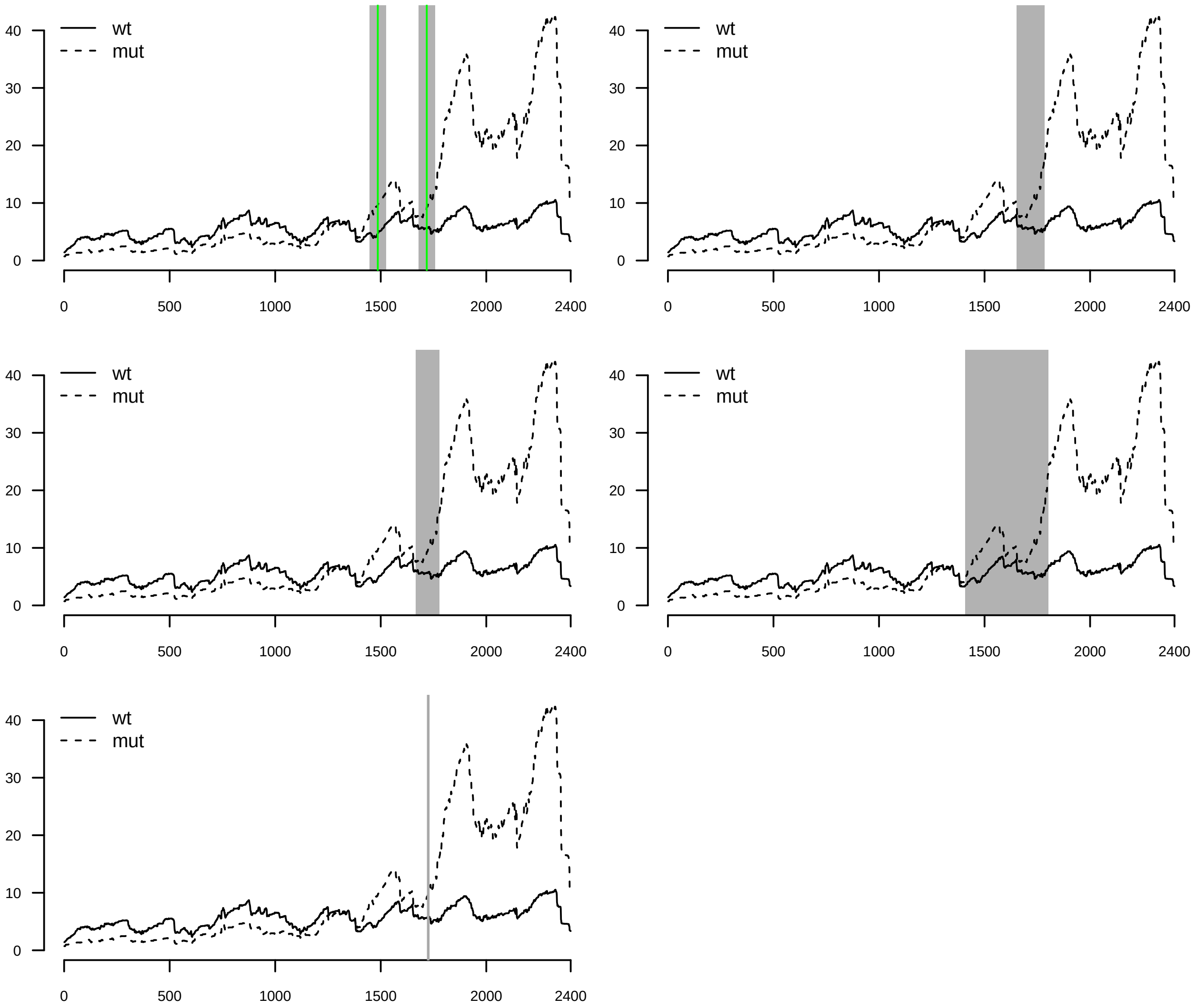
```
> genome_wide_scores(coverageDataSet = data, method = "ratio",
+                    outfile = "/tmp/ratio_scores.txt", introns = introns)
> genome_wide_scores(coverageDataSet = data, method = "enrichment",
+                    outfile = "/tmp/enrichment_scores.txt", introns = introns)
> genome_wide_scores(coverageDataSet = data, method = "probabilistic",
+                    outfile = "/tmp/zscores.txt", introns = introns)
```

3. Cryptic transcription start sites (cTSS)

The package implements a probabilistic method able to estimate the positions of cryptic transcription start sites (cryptic zones). We have defined five ways of identifying cryptic zones (for more information see ?initiation_sites).

3.1 Calculate cTSS

```
> #calculating the cTSS for FLO8 gene
> flo8_cTSS <- initiation_sites(
+   name = "YER109C", bamfiles = bamfiles,
+   types = types, annotations = annotations,
+   introns = introns, sf = scaling_factors, replicates = 100
+ )
>
> #visualize results
> par(mfrow = c(3,2))
> plot(flo8, cTSS = flo8_cTSS, method = "methodC_gaussian")
> plot(flo8, cTSS = flo8_cTSS, method = "methodA")
> plot(flo8, cTSS = flo8_cTSS, method = "methodB")
> plot(flo8, cTSS = flo8_cTSS, method = "methodC")
> plot(flo8, cTSS = flo8_cTSS, method = "methodD")
```



References

Cheung V, et al. (2008) Chromatin- and transcription-related factors repress tran-scription from within coding regions throughout the Saccharomyces cerevisiae genome. PLoS Biol. Nov 11;6(11):e277.

DeGennaro CM, et al. (2013) Spt6 regulates intragenic and antisense transcription, nucleosome positioning, and histone modifications genome-wide in fission yeast. Mol Cell Biol. Dec;33(24):4779-92.