rnaSeqMap: RNASeq analyses using xmap database back-end

Anna Lesniewska, Michal Okoniewski

October 13, 2015

Vignette for v.2.11.1 - no Xmap database needed

Contents

1 Recent changes and updates

8.09.2012 - the vignette contains mainly the material covered by the ECCB 2012 tutorial chunk 28.06.2012 - switched to GAlignments in RS class and used them to get coverage in ND - corrected all the rle into Rle 04.10.2011 - added data modification generators : generatorAddSquare(), generatorAdd(),generatorMultiply(), generatorTrunc(), generatorPeak(), generatorSynth()

local coverage normalizations: standarizationNormalize(), densityNormalize(), min_maxNormalize()

local coverage difference measures: ks_test(), diff_area(), diff_derivative_area(),qq_plot(), qq_derivative_plot(), pp_plot(), pp_derivative_plot(), hump_diff1(), hump_diff2()

14.05.2011 - added parseGff3()

2 Introduction

`rnaSeqMap` is a "middleware" library for RNAseq secondary analyses. It constitutes an API for such operations as:

- access to any the reads of the experiment in possibly fastest time, according to any chromosome coordinates
- accessing sets of reads according to genomic annotation in Ensembl
- calculation of coverage and number of reads and transformations of those values
- creating input for significance analysis algorithms - from edgeR and DESeq
- precisely finding significant and consistent regions of expression
- splicing analyses
- visualizations of genes and expression regions

The library is independent from the sequencing technology and reads mappina software. It needs either reads described as genome coordinates in the extended xmap database, or can alternatively read data as big as they can fit in the operational memory. The use with modified xmap database is recommended, as it overcomes memory limitations - thus the library can be efficiently run on not very powerful machines.

The internal features of `rnaSeqMap` distinctive for this piece of software are:

- sequencing reads and annotations in one common database - extended XMAP [**?**]
- algorithm for finding irreducible regions of genomic expression - according to Aumann and Lindell [**?**]
- nucleotide-level splicing analysis
- connectors for further gene- and region-level expression processing to DESeq [**?**] and edgeR [**?**]
- the routines for coverage, splicing index and region mining algorithm have been implemented in C for speed

3 Using the SeqReadsand NucleotideDistr objects

The reads are provided into the objects built according to genome coordinates from BAM files described in the "covdesc" file.

```
>         rs <- newSeqReads(ch,st, en, str);
>         rs <- getBamData(rs,idx.both, cvd=cvd)
>         nd <- getCoverageFromRS(rs, idx.both)
```

4 Processing schema to get the coverage measures using the "camel wrapper"

To get the coverage difference measures described in [2] Encode the experimental design in the sample description/covdesc file The comparison may be done between any two group of samples (1+1, n+n, n+m) Get samples indices, eg:

```
>  idxT <- which(samples$condition=="T")
>  idxC <- which(samples$condition=="C")
```

Prepare the table of genome coordinates to query Encode them as GenomicRanges object, eg:

```
>  regions.gR <- rnaSeqMap:::.fiveCol2GRanges(tmp)
```

Run the wrapper for all camel comparisons

```
>  regionsCamelMeasures <- gRanges2CamelMeasures(regions.gR,samples,idxT,idxC,sums=su
```

Run detection filtering by the density of coverage, eg:

```
> idx <- which(regionsCamelMeasures[,"covDensC1"]>10 | regionsCamelMeasures[,"covDens
> regionsCamelMeasures <- regionsCamelMeasures[idx, ]
```

Order the regions by a selected measure:

```
> o <- order(regionsCamelMeasures[,"QQ.mm"], decreasing=T)
> regionsCamelMeasures <- regionsCamelMeasures [o, ]
```

5 Using Aumann-Lindell two-sliding-window algorithm to find expressed genomic regions

The regions will be found as new object containing mindiff (second parameter value) for the nucleotides for which there are irreducible regions of coverage with given mindiff and minimal length - minsup. For the details of the algorithm see [1, 3]

```
> nd.AL <- findRegionsAsND(nd, 15, minsup=5)
```

References

[1] Leśniewska, A., Okoniewski, M. J. (2011). rnaSeqMap: a Bioconductor package for RNA sequencing data exploration. BMC bioinformatics, 12, 200. doi:10.1186/1471-2105-12-200

[2] Okoniewski, M. J., Leśniewska, A., Szabelska, A., Zyprych-Walczak, J., Ryan, M., Wachtel, M., Morzy, T., et al. (2011). Preferred analysis methods for single genomic regions in RNA sequencing revealed by processing the shape of coverage. Nucleic acids research. doi:10.1093/nar/gkr1249

[3] Aumann, Y., Lindell, Y. (2003). A Statistical Theory for Quantitative Association Rules. J. Intell. Inf. Syst., 20(3), 255–283.