ISH - R package for Intra-Sample Heterogeneity Scores

Michael Scherer

April 9, 2018

Contents

1 Introduction

This vignette desribes the functionalities included in the ISH R package. We describe the Intra-Sample Heterogeneity Scores FDRP, qFDRP, PDR, MHL, Epipolymorphism and Entropy. The package is able to compute each of those scores from bisulfite sequencing data. Input should be a *bam* file that contains reads that have been aligned to a reference genome. While PDR, qFDRP and FDRP are independent of the employed mapping tool, Epipolymorphism and Entropy require the reads to be aligned with bismark. In addition to the aligned reads, the user needs to specify the sites for which the scores should be computed in either of two ways: `GRanges`http://bioconductor.org/packages/release/bioc/html/GenomicRanges.html or `RnBSet` (`https://rnbeads.org/`). Here, we only discuss how to use the package. A detailed description of each of the scores can be found in the corresponding publications.

2 Installation

The package is available from GitHub and can be installed by the following command, given that the devtools package is installed:

```
> devtools::install_github("schmic05/ISH_package")
> library(ISH)
```

You can test if the installed version is functioning by employing one of the examples in the package:

```
> qfdrp <- ish.run.example()

2018-04-09 15:49:55     1.0  STATUS STARTED ISH score example
2018-04-09 15:49:56     1.0  STATUS     STARTED qFDRP caluclation
2018-04-09 15:50:37     1.0  STATUS     COMPLETED qFDRP caluclation
2018-04-09 15:50:37     1.0  STATUS COMPLETED ISH score example
```

3 Computing ISH scores

3.1 FDRP, qFDRP and PDR

FDRP, qFDRP and PDR do not require any additional tools or scripts and can be computed directly from your *bam* file. In this case, the `score` argument of the `compute.score` function needs to be one of "fdrp", "qfdrp" or "pdr". You need to specify the CpG sites for which the scores should be computed in either of two forms: `GRanges` or `RnBSet`.

1. `GRanges`: This object should contain the positions of the CpGs for which analysis is to be conducted. The `GRanges` object should contain a single entry for each CpG and only have length 1 for each of the entries. Then you can either run `compute.score.GRanges` directly or call the generic function `compute.score`.

   ```
   > example.bam <- system.file(file.path("extData","small_example.bam"),package="ISH")
   > example.GRanges <- GRanges(Rle(rep("chr2",10)),IRanges(start = c(2298361,2298554,2298
   +                                                                  2298915,2298921),end
   +
   > pdr <- compute.score(bam.file=example.bam,example.GRanges,score="pdr")

   2018-04-09 15:50:37     1.0  STATUS STARTED PDR calculation
   2018-04-09 15:50:48     1.0  STATUS COMPLETED PDR calculation
   ```

   This returns a *data.frame*, with the CpG positions (chromosome, start, end) in the first columns and the corresponding score in the last column.

   ```
   > dim(pdr)

   [1] 10  4

   > head(pdr)

     chromosome   start     end       PDR
   1       chr2 2298361 2298362       NaN
   2       chr2 2298554 2298555       NaN
   3       chr2 2298732 2298733 0.2413793
   4       chr2 2298743 2298744 0.2372881
   5       chr2 2298787 2298788 0.2337662
   6       chr2 2298792 2298793 0.2337662
   ```

2. **RnBSet:** In addition to `GRanges` objects, the ISH package supports `RnBSet` objects as input. Here, the annotation is inferred from the object's annotation with the addition of only selecting those sites that have a coverage higher than *coverage.threhold* in the `RnBSet`

object, given coverage information is present. For more details on how to set options for analysis, see subsection 4.1.

```
> example.rnb.set <- system.file(file.path("extData","small_rnbSet.zip"),package="ISH")
> example.rnb.set <- load.rnb.set(example.rnb.set)
> set.option(coverage.threshold = 10)
> fdrp <- rnb.calculate.fdrp(example.rnb.set,example.bam)
> to.plot <- data.frame(qFDRP=qfdrp$qFDRP,FDRP=fdrp$FDRP)
> to.plot <- melt(to.plot)
> plot <- ggplot(to.plot,aes(x=value,y=..count..,fill=variable))+geom_histogram()+facet
> plot
```

3.2 MHL

In contrast to the scores above, MHL requires a working version of `perl` installed on your machine. For Linux, this should in general be `/usr/bin/perl`, which is per default set in this package. In case you are using MacOS (why we do not support Windows is argued in subsection 4.2), you first need to specify the option *perl.path*. Furthermore, a working version of `samtools` is required by the programs that compute MHL.

```
> set.option(perl.path = "/usr/bin/perl")
> set.option(samtools.path = "/usr/bin/")
> mhl <- compute.score.rnb(bam.file = example.bam, rnb.set = example.rnb.set, score="mhl")
```

3.3 Epipolymorphism and Entropy

Epipolymorphism and Entropy calculations depend on the methclone software (https://code.google.com/archive/p/methclone/) to compute epiallele counts and then uses R functions to compute the final scores. This package comes with an executable version of methclone and has been tested for several Debian versions. If you have trouble with the methclone version, please contact the author. In contrast to the scores discussed above, Epipolymorphism and Entropy do not require an annotation object (either `GRanges` or `RnBSet`), since methclone operates as a black box and produces scores at positions directly inferred from the *bam* file.

```
> epipoly <- compute.score(example.bam,score="epipolymorphism")
> entropy <- compute.score(example.bam,score="entropy")
> to.plot <- data.frame(Epipolymorphism=epipoly$Epipolymorphism,Entropy=entropy$Entropy)
> to.plot <- melt(to.plot)
> plot <- ggplot(to.plot,aes(x=value,y=..density..,color=variable))+geom_density()+theme_bw(
> plot
```

4 Advanced Configuration

4.1 Option settings

The ISH package provides a bunch of options to set, which influence how the data is handled. This includes setting coverage thresholds on the annotation, distances between individual CpGs, or quality thresholds on reads to be considered in the calculation. For a detailed description of each of the options, see the R documentation.

```
> ?set.option
```

4.2 Windows troubleshooting

Using this package on a Windows OS, one can only compute qFDRP, FDRP and PDR, since they don't rely on external tools. In contrast to that, MHL depends on both `perl` and `samtools`, and since `samtools` is not easily installable on a Windows machine, we exclude this computation in case of a Windows. Epipolymorphism and Entropy depend on the methclone software, which is not supported for Windows and we thus also exclude this.