RRBSdata: An RRBS data set with simulated DMRs.

Katja Hebestreit, Hans-Ulrich Klein

October 10, 2014

Contents

1 Introduction

This package contains RRBS data with simulated differentially methylated regions (DMR's). In particular, it comprises 12 samples that are divided into 'cancer' and 'control' samples. We simulated 10,000 DMRs with different lengths and differences. This data set was used for [1] and can be used for any method evaluation to find DMRs.

2 Simulation

Instead of simulating bisulfite sequencing data, we used a real data set in which we incorporated DMRs. This procedure has the advantage that technical and biological characteristics of bisulfite sequencing data are present in the simulation data, for example, the correlation of DNA methylation of nearby CpG sites. And, especially, biological and technical DNA methylation variation across CpG sites and across samples is preserved. To obtain a

dataset with known DMRs, we used 12 human control samples from a previously published RRBS data set [2]. We simulated DMRs of different lengths and intensities by altering the number of methylated reads of some of the CpG sites in half of the samples, which could be considered as the cancer samples. We downloaded CpG island positions from UCSC database and filtered out all islands that were not covered in any of the samples (27 718 remaining islands). Only islands with not less than 10 covered CpG sites were considered to receive a DMR (24,698 islands). Within these 24,698 CpG islands we incorporated 10 000 DMRs with methylation differences of 10, 20, 30, or 40%. The DMRs spanned 10, 20, 30, or 40% of the CpG sites of the respective islands. We made sure that we gained the same amount of DMRs for each of the 16 combinations of methylation difference and percentage of modified CpG sites, that is, we gained 1 250 DMRs per combination.

We incorporated the DMRs as follows: 1.) We devided the 12 control samples into 6 "cancer" samples and 6 "control" samples. 2.) Within each CpG island we only considered regions to receive a DMR if each of its CpG sites was covered in at least half of all samples. Those regions are referred to as "covered island regions". 3.) For each CpG site within covered island regions we determined its minimum and maximum smoothed methylation level across all samples. 4.) For each CpG island we determined the maximum percentage of neighbored CpG sites that can be altered by determining the percentage of CpG sites within its biggest covered island region on all CpG sites within the island. 5.) Each of the 10,000 DMRs was sampled into a covered island region of a CpG island that was appropriate to harbor the DMR, in terms of minimum and maximum DNA methylation and CpG percentage of its largest covered island region on all CpG sites within the island. Not more than one DMR was incorporated per CpG island. 6.) Whenever it was possible to increase the DNA methylation by the amount of the difference of the respective DMR (that is, the resulting methylation level is below 1), it was increased in the cancer samples. Otherwise, the DNA methylation was decreased by the amount of difference. 7.) To increase the DNA methylation in a cancer sample within a region by a certain amount, the number of methylated reads of the respective CpG sites was increased. For instance, a CpG site of coverage 12 with 3 methylated reads (and a relative methylation level of 0.25) within a region that should get a methylation difference of 0.3, received 4 additional methylated reads (because: 0.3x12 = 4).

The scripts to simulate the data are available under `RRBSdata/R`.

3 Data objects

There are three data objects: `rrbs`, `islands` and `diffMethCpGs`.

The `rrbs` object is a *BSraw-class* object from the *BiSeq* package. It comprises the RRBS data:

```
> library(BiSeq)

> data(rrbs)
> rrbs

class: BSraw
dim: 10502 10
metadata(0):
assays(2): totalReads methReads
rownames(10502): 1456 1457 ... 4970981 4970982
rowData names(0):
colnames(10): APL1 APL2 ... APL11624 APL5894
colData names(1): group

> head(colData(rrbs))

DataFrame with 6 rows and 1 column
            group
         <factor>
APL1       APL
APL2       APL
APL3       APL
APL7       APL
APL8       APL
APL10961   control
```

The `islands` object is a *GRanges-class* object comprising all CpG islands that were considered to contain DMRs. Please see `?islands` for information regarding the columns.

The `diff.meth.cpgs` object is a *GRanges-class* object comprising all differentially methylated CpG sites:

```
> data(diffMethCpGs)
```

References

[1] Hans-Ulrich Klein and Katja Hebestreit. Global test and biseq are the methods of choice for testing genomic regions for differential methylation in bisulfite sequencing data. *In preparation.*

[2] Till Schoofs, Christian Rohde, Katja Hebestreit, Hans-Ulrich Klein, Stefanie Göllner, Isabell Schulze, Mads Lerdrup, Nikolaj Dietrich, Shuchi Agrawal-Singh, Anika Witten, Monika Stoll, Eva Lengfelder, Wolf-Karsten Hofmann, Peter Schlenke, Thomas Büchner, Klaus Hansen, Wolfgang E Berdel, Frank Rosenbauer, Martin Dugas, and Carsten Müller-Tidow. Dna methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood*, Nov 2012. URL: `http://dx.doi.org/10.1182/blood-2012-08-448860`, `doi:10.1182/blood-2012-08-448860`.