RnBeads – Annotations

Yassen Assenov, Fabian Müller, Pavlo Lutsik, Michael Scherer

July 4, 2024

1 Introduction

`RnBeads` is accompanied by four annotation packages: `RnBeads.hg19`, `RnBeads.mm10`, `RnBeads.mm9`, `RnBeads.rn5`. They contain genomic locus and Infinium probe annotation tables, as well as definition tables for tiling regions, genes, promoters and CpG islands. The `RnBeads` package contains routines for handling and updating these annotations. The vignette named **RnBeads – Comprehensive DNA Methylation Analysis** also contains examples that use functions presented here.

The data packages described above are hosted on the `RnBeads` website[1]. Note that at least one of the data packages should be installed prior to installing `RnBeads`.

2 Data Extraction and Processing

2.1 Sites

Currently, every data package examines cytosines in the context of CpG and contains an annnotation table of all CpGs in the respective genome. CpG density and GC content are also computed for the neighborhood of length 100 base pairs centered on each locus. The total number of dinucleotides annotated in HG19 is 28,217,009 represented both on the forward and reverse DNA strands.

2.2 HumanMethylation450 Probes

The Infinium HumanMethylation450 BeadChip interrogates over 470,000 loci in the human genome. Control and methylation probe annotation is obtained from online sources, validated for consistency and enriched with locations for the SNP-associated loci obtained from dbSNP. In addition, the annotation includes information about CpG density and GC content of the sequence neighborhood around every probe target as described in the section above.

2.3 dbSNP

Single-Nucleotide Polymorphisms (SNPs) in the human genome are downloaded from the respective FTP directory of dbSNP[2]. `RnBeads` uses the provided VCF files; for mouse and rat these are one file per chromosome. The following table shows the references in dbSNP used for the

[1] rnbeads.org
[2] www.ncbi.nlm.nih.gov/snp/organisms/

supported assemblies.

Package	Reference
RnBeads.hg19	GRCh37.p10
RnBeads.mm10	GCF_000001635.21
RnBeads.rn5	GCF_000001895.4

The data package `RnBeads.mm9` does not include SNP-related information. Only SNPs that meet all of the following criteria are considered:

1. **assembly** is `GRCh37.p5,reference`;
2. **chrom** is the expected chromosome;
3. **allele origin** is not `somatic`;
4. (for HG19) major allele frequency is at most 0.95.

These SNP positions are used to enrich the annotations for CpG dinucleoties and Infinium probes.

2.4 Regions

Every data package defines the following sets of regions for the dedicated assembly:

Tiling regions Tiling regions with a window size of 5 kilobases are defined over the whole genome.

Genes and promoters Ensembl[3] gene definitions are downloaded using the `biomaRt` package. A promoter is defined as the region spanning 1,500 bases upstream and 500 bases downstream of the transcription start site of the corresponding gene.

GpG islands The CpG island track is downloaded from the dedicated FTP directory of the UCSC Genome Browser[4].

CpG density and GC content are computed for all region types listed above.

3 Annotations

3.1 Genome Assemblies

Every annotation table and other data structure is specific to a genome assembly. The list of supported assemblies can be obtained by the calling the function `rnb.get.assemblies`, as shown in the example below.

```
> rnb.get.assemblies()

[1] "hg19" "hg38" "mm9"
```

[3] www.ensembl.org/
[4] genome.ucsc.edu/

No.	Column Name	Description
1	ID	Probe identifier.
2	Target	Probe category.
3	Color	Probe color.
4	Description	Probe description, abbreviated.
5	AVG	...
6	Evaluate Green	If probe is used in the evaluation of the green channel.
7	Evaluate Red	If probe is used in the evaluation of the red channel.
8	Expected Intensity	Expected intensity of the probe.
9	Sample-dependent	Flag indicating if the intensity of the probe is sample-dependent.
10	Index	Index of the probe in its category.

Table 1: Columns in the HumanMethylation450 control probe annotation table.

3.2 Built-in Annotation Tables

Control probe information for the Illumina 450K array is stored (and can be extracted) as a `data.frame` in which every row corresponds to a control probe. Every other annotation table is represented by a dedicated `GRangesList` object, that is, a list of consistent `GRanges` objects, one per chromosome. Every instance of `GRanges` defines the genomic positions of the corresponding sites, probes or regions. Identifiers, if present, can be obtained using the `names` method. Strand information is also included when applicable. Any additional annotation is stored as metadata. Please refer to the documentations of the `GenomicRanges` package and the `GRanges` class for more details.

In an R session, every annotation object can be accessed using the function `rnb.get.annotation`. The code snippet below shows how the control probe annotations can be extracted. Table 1 briefly describes the characteristics of these probes.

```
> control.annotation <- rnb.get.annotation("controls450")
> head(control.annotation)
```

In a similar manner, the command below extracts the annotation of HumanMethylation450 probes. The metadata structure for the methylation and SNP-associated probes is presented in Table 2.

```
> probe.annotation <- rnb.get.annotation("probes450")
> probe.annotation
```

Every data package integrates various region types with the CpG sites (and Infinium probes for HG19) presented in the previous section. Obtaining a region annotation information is very similar to accessing site or probe annotation. The code snippet below extracts the available gene definitions. Table 3 provides information about the metadata of these regions.

```
> gene.annotation <- rnb.get.annotation("genes")
> attr(gene.annotation, "version")
> gene.annotation
```

No.	Column Name	Description
1	Design	Probe design type.
2	Color	Color channel.
3	Context	Probe context.
4	Random	Flag indicating if the probeś location was randomly chosen.
5	HumanMethylation27	Flag indicating if the probe is also covered by HumanMethylation27k assay.
6	Mismatches A	Number of base mismatches between the provided and expected probe sequence.
7	Mismatches B	Number of base mismatches between the provided and expected probe sequence.
8	CGI Relation	Relation to a CpG island.
9	CpG	Number of CpG dinucleotides in the sequence neighborhood of the target.
10	GC	Percentage of C and G bases in the sequence neighborhood of the target.
11	SNPs 3	Number of SNPs that overlap with the last 3 bases of a probe's target sequence.
12	SNPs 5	Number of SNPs that overlap with the last 5 bases of a probe's target sequence.
13	SNPs Full	Number of SNPs that overlap with the probe's sequence.

Table 2: Metadata accompanying the HumanMethylation450 probe definitions.

No.	Column Name	Description
1	symbol	Gene symbols associated with this region.
2	entrezID	Entrez gene identifiers associated with this region.
3	CpG	Number of CpG dinucleotides in the region.
4	GC	Total number of C and G bases in the region.

Table 3: Metadata accompanying the Ensembl gene and promoter definitions.

The function `rnb.region.types` returns all supported region definitions for a given genome assembly. These currently include `"cpgislands"`, `"genes"`, `"promoters"`, `"tiling"`.

The `GRangesList` objects are well optimized for storage and fast access of genomic elements. However, exporting such objects to tables in human-readable form is not trivial. The function `rnb.annotation2data.frame` provides an easy way to convert a probe or region annotation to a single table. The code snippet below exemplifies this function by saving the gene promoter definitions to a comma-separated value (CSV) file named `promoters.csv`.

```
> promoters <- rnb.annotation2data.frame(rnb.get.annotation("promoters"))
> write.csv(promoters, file = "promoters.csv", row.names = FALSE)
```

3.3 Custom Annotations

Custom region annotations can be included using the function `rnb.set.annotation`. The information is either loaded from a BED file, or it must be presented as a `data.frame` containing at least the following three columns - **chromosome**, **start** and **end**. Note that the included genomic regions remain available in the current R session only.

The vignette named **RnBeads – Comprehensive DNA Methylation Analysis** contains a detailed example for including custom annotations to an analysis.

4 Data Sources

The data packages accompanying `RnBeads` obtained data for the annotation tables from the following sources:

Other Bioconductor packages Methylation and control probe annotation is obtained from the package FDb.InfiniumMethylation.hg19. In addition, `RnBeads` uses the packages:

BSgenome.Hsapiens.UCSC.hg19

BSgenome.Mmusculus.UCSC.mm9

BSgenome.Mmusculus.UCSC.mm10

BSgenome.Rnorvegicus.UCSC.rn5

to validate the probe annotations and to enrich all annotation tables with sequence-based properties.

CRAN packages Additional control probe information is obtained from the package HumMeth27QCReport.

dbSNP Information on SNPs in the human genome is downloaded as BED files from the FTP site of dbSNP.

Gene Expression Omnibus (GEO) Methylation and control probe annotation is obtained from GEO record GPL13534 and validated for consistency with the other sources.

Biomart Ensembl gene definitions (Ensembl versions 73 (for hg19, mm10 and rn5), and 67 (for mm9) are downloaded using the biomaRt package.

UCSC Genome Browser The CpG island tracks are downloaded from the FTP site of the UCSC Genome Browser.