

An Introduction to the *REMP* Package

Yinan Zheng

May 5, 2024

Contents

1	Introduction	2
2	Installation	2
3	REMP: Repetitive Element Methylation Prediction	2
3.1	Groom methylation data	2
3.2	Prepare annotation data	4
3.3	Run prediction	4
3.4	Plot prediction	10
4	Extract RE-CpG methylation profiled by Illumina BeadChip array	11

1 Introduction

REMP predicts DNA methylation of locus-specific repetitive elements (RE) by learning surrounding genetic and epigenetic information. *REMP* provides genomewide single-base resolution of DNA methylation on RE that is difficult to measure directly using array-based or sequencing-based platforms, which enables epigenome-wide association study (EWAS) and differentially methylated region (DMR) analysis on RE. *REMP* also provides handy tool to extract methylation data of CpGs that are located within RE sequences.

REMP supports both Illumina methylation BeadChip array platforms (450k and EPIC) and sequencing platforms (e.g. TruSeq Methyl Capture EPIC). Both genome build hg19 and hg38 are supported.

2 Installation

Install *REMP* (release version):

```
> if (!requireNamespace("BiocManager", quietly=TRUE))
+   install.packages("BiocManager")
> BiocManager::install("REMP")
```

To install devel version:

```
> library(devtools)
> install_github("YinanZheng/REMP")
```

Load *REMP* into the workspace:

```
> library(REMP)
```

3 REMP: Repetitive Element Methylation Prediction

Currently *REMP* supports Human (hg19/hg38) Alu, LINE-1 (L1), and Long Terminal Repeat (LTR) (including endogenous retroviruses, ERV) repetitive element (RE) methylation prediction using Illumina 450k/EPIC array or sequencing platform.

3.1 Groom methylation data

Appropriate data preprocessing including quality control and normalization of methylation data are recommended before running *REMP*. Many packages are available to carry out these data preprocessing steps, for example, *minfi*, *wateRmelon*, and *methyumi*.

REMP is trying to minimize the requirement of the methylation data format. Users can maintain the methylation data in *RatioSet* or *GenomicRatioSet* object offered by *minfi*, *data.table*, *data.frame*, *DataFrame*, or *matrix*. Users can input either beta value or M-value. There are only two basic requirements of the methylation array data (450k/EPIC):

1. Each row should represent CpG probe and each column should represent sample.
2. The row names should indicate Illumina probe ID (i.e. cg00000029).

However, there are some other common data issues that may prevent *REMP* from running correctly. For example, if the methylation data are in beta value and contain zero methylation values, logit transformation (to create M-value) will create negative infinite value; or the methylation data contain NA, Inf, or NaN data. To tackle these potential issues, *REMP* includes a handy function `groomMethy` which can help detect and fix these issues. We highly recommend to take advantage of this function:

```

> # Get GM12878 methylation data (450k array)
> GM12878_450k <- getGM12878('450k')
> GM12878_450k <- grooMethy(GM12878_450k)
> GM12878_450k

class: RatioSet
dim: 482421 1
metadata(0):
assays(2): Beta M
rownames(482421): cg00000029 cg00000108 ... cg27666046
cg27666123
rowData names(0):
colnames(1): GM12878
colData names(0):
Annotation
  array: IlluminaHumanMethylation450k
  annotation: ilmn12.hg19
Preprocessing
  Method: NA
  minfi version: NA
  Manifest version: NA

```

For zero beta values, `grooMethy` will replace them with smallest non-zero beta value. For one beta values, `grooMethy` will replace them with largest non-one beta value. For NA/NaN/Inf values, `grooMethy` will treat them as missing values and then apply KNN-imputation to complete the dataset. If the imputed value is out of the original range (which is possible when `imputebyrow = FALSE`), mean value will be used instead. Warning: imputed values for multimodal distributed CpGs (across samples) may not be correct. Please check package *ENmix* to identify the CpGs with multimodal distribution.

For sequencing data, the users only need to prepare a methylation data matrix (row = CpGs, column = samples). The corresponding CpG location information (either in hg19 or hg38) should be prepared in a separate *GRanges* object and provide it to the `Seq.GR` argument in `grooMethy`. For an example of `Seq.GR`, please run:

```

> library(IlluminaHumanMethylation450kanno.ilmn12.hg19)
> getLocations(IlluminaHumanMethylation450kanno.ilmn12.hg19)

```

GRanges object with 485512 ranges and 0 metadata columns:

	seqnames	ranges	strand
	<Rle>	<IRanges>	<Rle>
cg00050873	chrY	9363356	*
cg00212031	chrY	21239348	*
cg00213748	chrY	8148233	*
cg00214611	chrY	15815688	*
cg00455876	chrY	9385539	*
...
ch.22.909671F	chr22	46114168	*
ch.22.46830341F	chr22	48451677	*
ch.22.1008279F	chr22	48731367	*
ch.22.47579720R	chr22	49193714	*
ch.22.48274842R	chr22	49888838	*

seqinfo: 24 sequences from hg19 genome; no seqlengths

Note that the row names of the CpGs in `Seq.GR` can be `NULL`.

3.2 Prepare annotation data

To run *REMP* for RE methylation prediction, users first need to prepare some annotation datasets. The function `initREMP` is designed to do the job.

Suppose users will predict Alu methylation using Illumina 450k array data:

```
> data(Alu.hg19.demo)
> remparcel <- initREMP(arrayType = "450k",
+                       REtype = "Alu",
+                       annotation.source = "AH",
+                       genome = "hg19",
+                       RE = Alu.hg19.demo,
+                       ncore = 1)
> remparcel
```

```
REMPParcel object
RE type: Alu
Genome build: hg19
Illumina platform: 450k
Valid (max) Alu-CpG flanking window size: 1200
Number of RE: 500
Number of Alu-CpG: 5138
```

For demonstration, we only use 500 selected Alu sequence dataset which comes along with the package (`Alu.hg19.demo`). We specify `RE = Alu.hg19.demo`, so that the annotation dataset will be generated for the 500 selected Alu sequences. Most of the time, specifying `RE` is not necessary, as the function will fetch the complete RE sequence dataset from package *AnnotationHub* using `fetchRMSK`. Users can also use this argument `RE` to provide customized RE dataset.

`annotation.source` allows the users to switch the source of the annotation databases, including the RefSeq Gene annotation database and RepeatMasker annotation database. If `annotation.source = "AH"`, the database will be obtained from the AnnotationHub package. If `annotation.source = "UCSC"`, the database will be downloaded from the UCSC website <http://hgdownload.cse.ucsc.edu/goldenpath>. The corresponding build ("`hg19`" or "`hg38`") can be specified in the argument `genome`. Most of the time "`hg19`" is used for array data. But if "`hg38`" is specified, the function will liftover the CpG probe location information to "`hg38`" and obtain annotation databases in "`hg38`".

If `arrayType = "Sequencing"`, users should provide the genomic location information of the CpGs in a *GRanges* object to `Seq.GR`. Note that the genome build of `Seq.GR` provided should match the genome build specified in `genome`.

All data are stored in the *REMPParcel* object:

```
> saveParcel(remparcel)
```

It is recommended to specify a working directory using argument `work.dir` in `initREMP` so that the annotation data generated can be re-used. Without specifying working directory, the annotation dataset will be created under the temporal directory `tempdir()` by default. Users can also turn on the `export` argument in `initREMP` to save the data automatically.

3.3 Run prediction

Once the annotation data are ready, users can pass the annotation data parcel to `remp` for prediction:

```
> remp.res <- remp(GM12878_450k,
+                  REtype = 'Alu',
```

```
+         parcel = remparcel,
+         ncore = 1,
+         seed = 777)
```

If `parcel` is missing, `remp` will then try to search the *REMP*Parcel data file in the directory indicated by `work.dir`. If `work.dir` is also missing, `remp` will try to search the *REMP*Parcel data file in the temporal directory `tempdir()`.

By default, `remp` uses Random Forest (`method = 'rf'`) model (package *ranger* for fast implementation) for prediction. Random Forest model is recommended because it offers more accurate prediction results and it automatically enables Quantile Regression Forest (Nicolai Meinshausen, 2006) for prediction reliability evaluation. `remp` constructs predictors to carry out the prediction. For Random Forest model, the tuning parameter `param = 6` (i.e. `mtry` in *ranger* or *randomForest*) indicates how many predictors will be randomly selected for building the individual trees. The performance of random forest model is often relatively insensitive to the choice of `mtry`. Therefore, auto-tune will be turned off using random forest and `mtry` will be set to one third of the total number of predictors. It is recommended to specify a seed for reproducible prediction results.

Besides random forest, `remp` provides other machine learning engines for users to explore, including Extreme Gradient Boosting, SVM with linear kernel, and SVM with radial kernel).

`remp` will return a *REMP*set object, which inherits Bioconductor's *RangedSummarizedExperiment* class:

```
> remp.res

class: REMProduct
dim: 4952 1
metadata(8): REannotation RECpG ... GeneStats Seed
assays(3): rempB rempM rempQC
rownames: NULL
rowData names(1): RE.Index
colnames(1): GM12878
colData names(1): mtry

> # Display more detailed information
> details(remp.res)

RE type: Alu
Genome build: hg19
Methylation profiling platform: 450k
Flanking window size: 1000
Prediction model: Random Forest
QC model: Quantile Regression Forest
Seed: 777
Covered 4952 CpG sites in 500 Alu

Number of Alu-CpGs by chromosome:
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8
 429  195  257  197  142  382  202  153

chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16
  93   172   192   355    52   122   201   280

chr17 chr18 chr19 chr20 chr21 chr22
 361    65   803   107    56   136

Training information:
 500 profiled Alu are used for model training.
```

490 Alu-CpGs that have at least 2 neighboring profiled CpGs are used for model training.

Coverage information:

The data cover 500 Alu (4952 Alu-CpG).

Gene coverage by Alu (out of total # of RefSeq genes):

530 (2.13%) total genes;

460 (2.4%) protein-coding genes;

103 (1.43%) non-coding RNA genes.

Distribution of methylation value (beta value):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.01414972	0.43123992	0.63885476	0.56753362	0.75513368	0.90365646

Distribution of reliability score (lower score = higher reliability):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.6177377	1.3028736	1.6043964	1.6758817	2.0213579	5.7489818

Prediction results can be obtained by accessors:

```
> # Predicted RE-CpG methylation value (Beta value)
> rempB(remp.res)
```

DataFrame with 4952 rows and 1 column

```
      GM12878
<numeric>
1      0.653224
2      0.651030
3      0.663961
4      0.677392
5      0.673879
...      ...
4948  0.796753
4949  0.796552
4950  0.799788
4951  0.799387
4952  0.806483
```

```
> # Predicted RE-CpG methylation value (M value)
> rempM(remp.res)
```

DataFrame with 4952 rows and 1 column

```
      GM12878
<numeric>
1      0.913576
2      0.899622
3      0.982469
4      1.070209
5      1.047085
...      ...
4948  1.97090
4949  1.96911
4950  1.99809
4951  1.99448
4952  2.05918
```

```
> # Genomic location information of the predicted RE-CpG
> # Function inherit from class 'RangedSummarizedExperiment'
> rowRanges(remp.res)
```

GRanges object with 4952 ranges and 1 metadata column:

	seqnames	ranges	strand	RE.Index
	<Rle>	<IRanges>	<Rle>	<Rle>
[1]	chr1	1149956-1149957	+	Alu_0000214
[2]	chr1	1149966-1149967	+	Alu_0000214
[3]	chr1	1149999-1150000	+	Alu_0000214
[4]	chr1	1150049-1150050	+	Alu_0000214
[5]	chr1	1150057-1150058	+	Alu_0000214
...
[4948]	chr22	50914839-50914840	-	Alu_1118131
[4949]	chr22	50914853-50914854	-	Alu_1118131
[4950]	chr22	50914860-50914861	-	Alu_1118131
[4951]	chr22	50914869-50914870	-	Alu_1118131
[4952]	chr22	50914907-50914908	-	Alu_1118131

seqinfo: 24 sequences from an unspecified genome; no seqlengths

```
> # Standard error-scaled permutation importance of predictors
> rempImp(remp.res)
```

DataFrame with 18 rows and 1 column

	GM12878
	<numeric>
RE.swScore	5.47630
RE.Length	1.79475
RE.CpG.density	6.50236
RE.InTSS	3.58959
RE.In5UTR	2.78372
...	...
Methy.mean.mov1	19.11597
Methy.mean.mov2	16.94418
Methy.mean.mov3	10.62029
Methy.mean.mov4	9.09739
Methy.std	6.82098

```
> # Retrive seed number used for the reesults
> metadata(remp.res)$Seed
```

```
[1] 777
```

Trim off less reliable predicted results:

```
> # Any predicted CpG values with quality score less than
> # threshold (default = 1.7) will be replaced with NA.
> # CpGs contain more than missingRate * 100% (default = 20%)
> # missing rate across samples will be discarded.
> remp.res <- rempTrim(remp.res, threshold = 1.7, missingRate = 0.2)
> details(remp.res)
```

RE type: Alu
Genome build: hg19

Methylation profiling platform: 450k
 Flanking window size: 1000
 Prediction model: Random Forest - trimmed (1.7)
 QC model: Quantile Regression Forest
 Seed: 777
 Covered 2812 CpG sites in 397 Alu

Number of Alu-CpGs by chromosome:

chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
293	138	149	130	89	132	108	91

chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16
42	84	114	229	32	60	92	167

chr17	chr18	chr19	chr20	chr21	chr22
194	41	474	73	8	72

Coverage information:

The data cover 397 Alu (2812 Alu-CpG).

Gene coverage by Alu (out of total # of RefSeq genes):

419 (1.68%) total genes;
 362 (1.89%) protein-coding genes;
 85 (1.18%) non-coding RNA genes.

Distribution of methylation value (beta value):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.03748213	0.61906555	0.72691418	0.66424329	0.79475308	0.90365646

Distribution of reliability score (lower score = higher reliability):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.6177377	1.1818386	1.3420702	1.3347675	1.5086062	1.6998495

(Optional) Aggregate the predicted methylation of CpGs in RE by averaging them to obtain the RE-specific methylation level:

```
> remp.res <- rempAggregate(remp.res, NCpG = 2)
> details(remp.res)
```

RE type: Alu (aggregated by mean: min # of CpGs: 2)

Genome build: hg19

Methylation profiling platform: 450k

Flanking window size: 1000

Prediction model: Random Forest - trimmed (1.7)

QC model: Quantile Regression Forest

Seed: 777

Covered 336 Alu (aggregated by mean: min # of CpGs: 2)

Number of Alu (aggregated by mean: min # of CpGs: 2) by chromosome:

chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
33	17	19	13	8	21	13	10

chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16
9	10	14	24	6	6	12	19


```
chr17 chr18 chr19 chr20 chr21 chr22
    23     4    58     8     1     8
```

Coverage information:

The data cover 336 Alu (aggregated by mean: min # of CpGs: 2)

Gene coverage by Alu (aggregated by mean: min # of CpGs: 2) (out of total # of RefSeq genes)

353 (1.42%) total genes;

302 (1.58%) protein-coding genes;

72 (1%) non-coding RNA genes.

Distribution of methylation value (beta value):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.04913212	0.59104405	0.71030962	0.64305168	0.77820466	0.85147583

Distribution of reliability score (lower score = higher reliability):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.8480023	1.2536260	1.3774244	1.3712054	1.5017999	1.6976878

Aggregating CpGs in the same RE for RE-level methylation data is beneficial because 1) it greatly reduces the data dimension for downstream analysis and 2) it may produce more robust RE methylation estimation. Note that by default, RE with 2 or more predicted CpG sites will be aggregated. Therefore, the downside of doing this is the reduced coverage of RE. The assumption of doing this is the CpG methylation level within each RE are similar.

To add genomic regions annotation of the predicted REs:

```
> # By default gene symbol annotation will be added
> remp.res <- decodeAnnot(remp.res)
> rempAnnot(remp.res)
```

GRanges object with 336 ranges and 12 metadata columns:

	seqnames	ranges	strand	swScore	repName
	<Rle>	<IRanges>	<Rle>	<integer>	<character>
[1]	chr1	1149947-1150242	+	2216	AluSx1
[2]	chr1	1885941-1886230	+	1962	AluJb
[3]	chr1	14032197-14032500	+	2768	AluSg
[4]	chr1	21273562-21273862	+	1867	AluJb
[5]	chr1	22378622-22378926	+	2142	AluSx1
...
[332]	chr22	44964234-44964539	+	2284	AluSq2
[333]	chr22	50311037-50311333	+	2251	AluSx1
[334]	chr22	50493205-50493515	+	2229	AluSx
[335]	chr22	42078691-42078952	-	1881	AluSx1
[336]	chr22	50914608-50914917	-	2526	AluY

	repClass	repFamily	Index	InNM.symbol	InNR.symbol
	<character>	<character>	<Rle>	<character>	<character>
[1]	SINE	Alu	Alu_0000214	TNFRSF4	<NA>
[2]	SINE	Alu	Alu_0000634	CFAP74	<NA>
[3]	SINE	Alu	Alu_0004565	PRDM2	<NA>
[4]	SINE	Alu	Alu_0006920	EIF4G3	<NA>
[5]	SINE	Alu	Alu_0007535	CDC42	<NA>
...
[332]	SINE	Alu	Alu_1104534	<NA>	LINC00207
[333]	SINE	Alu	Alu_1105703	CRELD2 ALG12	CRELD2
[334]	SINE	Alu	Alu_1105737	<NA>	<NA>
[335]	SINE	Alu	Alu_1115677	SNU13	<NA>

[336]	SINE	Alu	Alu_1118131	SBF1	<NA>
	InTSS.symbol	In5UTR.symbol	InCDS.symbol	InExon.symbol	
	<character>	<character>	<character>	<character>	
[1]	TNFRSF4	<NA>	<NA>	<NA>	
[2]	<NA>	<NA>	<NA>	CFAP74	
[3]	<NA>	PRDM2	<NA>	<NA>	
[4]	<NA>	<NA>	EIF4G3	<NA>	
[5]	CDC42	<NA>	<NA>	<NA>	
...	
[332]	LINC00207	<NA>	<NA>	<NA>	
[333]	CRELD2	ALG12	<NA>	<NA>	
[334]	<NA>	<NA>	<NA>	<NA>	
[335]	<NA>	SNU13	SNU13	<NA>	
[336]	SBF1	<NA>	<NA>	<NA>	
	In3UTR.symbol				
	<character>				
[1]	<NA>				
[2]	CFAP74				
[3]	<NA>				
[4]	<NA>				
[5]	<NA>				
...	...				
[332]	<NA>				
[333]	<NA>				
[334]	<NA>				
[335]	<NA>				
[336]	<NA>				

seqinfo: 24 sequences from hg19 genome

Seven genomic region indicators will be added to the annotation data in the input *REMP* object:

- InNM: in protein-coding genes (overlap with refSeq gene's "NM" transcripts + 2000 bp upstream of the transcription start site (TSS))
- InNR: in noncoding RNA genes (overlap with refSeq gene's "NR" transcripts + 2000 bp upstream of the TSS)
- InTSS: in flanking region of 2000 bp upstream of the TSS. Default upstream limit is 2000 bp, which can be modified globally using `remp_options`
- In5UTR: in 5' untranslated regions (UTRs)
- InCDS: in coding DNA sequence regions
- InExon: in exon regions
- In3UTR: in 3'UTRs

Note that intron region and intergenic region information can be derived from the above genomic region indicators: if "InNM" and/or "InNR" is not missing but "InTSS", "In5UTR", "InExon", and "In3UTR" are missing, then the RE is strictly located within intron region; if all indicators are missing, then the RE is strictly located in intergenic region.

3.4 Plot prediction

Make a density plot of the predicted methylation (beta values):

```
> remplot(remp.res, main = "Alu methylation (GM12878)", col = "blue")
```

4 Extract RE-CpG methylation profiled by Illumina BeadChip array

REMP offers a handy tool to extract methylation data of CpGs that are located in RE. Similar as *remp*, users can choose the source of annotation database (AH: AnnotationHub or UCSC: UCSC website) and genome build (hg19 or hg38).

```
> # Use Alu.hg19.demo for demonstration
> remp.res <- remprofile(GM12878_450k,
+                        REtype = "Alu",
+                        annotation.source = "AH",
+                        genome = "hg19",
+                        RE = Alu.hg19.demo)
> details(remp.res)
```

```
RE type: Alu
Genome build: hg19
Methylation profiling platform: 450k
Flanking window size: N/A
Prediction model: Profiled
QC model: N/A
Covered 602 CpG sites in 500 Alu
```

Number of Alu-CpGs by chromosome:

```
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8
  58   22   33   23   17   59   26   18
```

```
chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16
  12   27   20   42    8   11   20   30
```

```
chr17 chr18 chr19 chr20 chr21 chr22
  39    9   97   13    5   13
```

Coverage information:

```
The data cover 500 Alu (602 Alu-CpG).
Gene coverage by Alu (out of total # of RefSeq genes):
  530 (2.13%) total genes;
  460 (2.4%) protein-coding genes;
  103 (1.43%) non-coding RNA genes.
```

Distribution of methylation value (beta value):

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.001000 0.344500 0.649500 0.567397 0.814000 0.959000
```

```
> # All accessors and utilites for REMProduct are applicable
> remp.res <- rempAggregate(remp.res)
> details(remp.res)
```

```
RE type: Alu (aggregated by mean: min # of CpGs: 2)
Genome build: hg19
Methylation profiling platform: 450k
Flanking window size: N/A
```

Prediction model: Profiled

QC model: N/A

Covered 86 Alu (aggregated by mean: min # of CpGs: 2)

Number of Alu (aggregated by mean: min # of CpGs: 2) by chromosome:

chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8
13	2	6	2	1	14	3	2

chr10	chr11	chr12	chr14	chr15	chr16	chr17	chr18
2	3	9	1	2	4	4	3

chr19	chr20	chr21	chr22
9	3	1	2

Coverage information:

The data cover 86 Alu (aggregated by mean: min # of CpGs: 2)

Gene coverage by Alu (aggregated by mean: min # of CpGs: 2) (out of total # of RefSeq genes)

98 (0.39%) total genes;

83 (0.43%) protein-coding genes;

24 (0.33%) non-coding RNA genes.

Distribution of methylation value (beta value):

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0408311	0.1983081	0.6649476	0.5265950	0.7829270	0.9288642