

Vignette for R package `asmn`

Anna Decker¹ and Paul Yousefi²

¹University of California, Berkeley, Division of Biostatistics

²University of California, Berkeley, Department of Environmental Health Sciences

October 13, 2014

1 Introduction

The `asmn` package performs the all-sample mean normalization procedure for Illumina BeadArray 450k methylation data. This package does not contain a complete pipeline for normalizing raw data, but the functions do take data in the `MethyLumiSet` format for integration with existing pipelines for analysis of methylation data. The functions can also take raw experimental and control data as well as feature information from `BeadStudio`, which can be read in as `data.frames`.

The `asmn` package is loaded by

```
> library(asmn)
```

To access the help files, type `help(package = asmn)` in the R console.

The example data come from the `TCGAMethylation450k` package, and are loaded as a `MethyLumiSet` object. The procedure to load the data is from a vignette for the `methylumi` package.

```
> library("methylumi")
> library("TCGAMethylation450k")
> idatPath <- system.file('extdata/idat', package='TCGAMethylation450k')
> mset450k <- methylumIDAT(getBarcodes(path=idatPath), idatPath=idatPath)
> sampleNames(mset450k) <- paste0('TCGA', seq_along(sampleNames(mset450k)))
> show(mset450k)
```

Object Information:

`MethyLumiSet` (storageMode: lockedEnvironment)

assayData: 485577 features, 10 samples

 element names: betas, methylated, methylated.OOB, pvals, unmethylated, unmethylated.OOB

protocolData: none

phenoData

 sampleNames: TCGA1 TCGA2 ... TCGA10 (10 total)

 varLabels: barcode

 varMetadata: labelDescription

featureData

 featureNames: cg00000029 cg00000108 ... rs9839873 (485577 total)

 fvarLabels: Probe_ID DESIGN COLOR_CHANNEL

 fvarMetadata: labelDescription

```

experimentData: use 'experimentData(object)'
Annotation: IlluminaHumanMethylation450k
Major Operation History:
      submitted          finished
1 2014-10-13 20:43:33 2014-10-13 20:44:37
2 2014-10-13 20:44:38 2014-10-13 20:44:40

                                     command
1 methylumIDAT(barcodes = getBarcodes(path = idatPath), idatPath = idatPath)
2                                     Subset of 485577 features.

```

2 Normalization factors

The normalization factors are calculated using the control data for each subject. The default settings of the `norm_factors()` function uses the mean of all control samples to create the normalization factors. The output from this function is a list of length 2 (one for each color channel), each containing a vector of normalization factors equal in length to the number of subjects.

One of either `controldata` or `methylumidata` must be supplied, but supplying both will produce an error, since `controldata` is the raw control data whereas `methylumidata` is a `MethyLumiSet` object, which may contain control and experimental data. The `subjects` argument is optional. Specifying a range of names or indices of subjects will calculate the normalization factors using only the control data for those subjects as opposed to using all samples. Finally, the `type` argument must be one of either "raw" or "methylumi," indicating the type of data being supplied.

```

> normfactors <- norm_factors(controldata=NULL,
+                             subjects=NULL,
+                             methylumidata=mset450k,
+                             type="methylumi")
> str(normfactors)

List of 2
 $ Red  : Named num [1:10] 0.886 1.022 1.034 1.033 1.402 ...
  ..- attr(*, "names")= chr [1:10] "TCGA1" "TCGA2" "TCGA3" "TCGA4" ...
 $ Green: Named num [1:10] 0.817 1.013 1.061 1.073 1.306 ...
  ..- attr(*, "names")= chr [1:10] "TCGA1" "TCGA2" "TCGA3" "TCGA4" ...

> normfactors

$Red
   TCGA1   TCGA2   TCGA3   TCGA4   TCGA5   TCGA6   TCGA7   TCGA8
0.8862207 1.0219058 1.0338912 1.0329013 1.4023652 0.9312108 0.7071845 0.6779672
   TCGA9   TCGA10
0.9065031 1.3998501

$Green
   TCGA1   TCGA2   TCGA3   TCGA4   TCGA5   TCGA6   TCGA7   TCGA8
0.8174033 1.0131220 1.0605838 1.0727148 1.3064618 1.0209169 0.6855948 0.7061097
   TCGA9   TCGA10
0.9453489 1.3717441

```

3 Normalization

The normalization factors can then be used in the `normalize()` function. For data of type `MethyLumiSet`, this function return the data object with the normalized data in the `betas()` slot.

The `normfactors` argument is the output from `norm_factors()`. Either one of `rawdata` or `methylumidata` must be supplied, depending on the format of the data set to be normalized (similar to the creation of the normalization factors). The `type` argument must be either "raw" or "methylumi" indicating the type of data to be normalized. If `type = "raw"`, then the `featuredata` argument must be supplied, containing information on the assay type and color channel for each probe. This information is stored in the `MethyLumiSet` data in the `featureData` slot.

```
> featureData(mset450k)
```

```
An object of class 'AnnotatedDataFrame'
```

```
featureNames: cg00000029 cg00000108 ... rs9839873 (485577 total)
```

```
varLabels: Probe_ID DESIGN COLOR_CHANNEL
```

```
varMetadata: labelDescription
```

```
> str(fData(mset450k))
```

```
'data.frame':      485577 obs. of  3 variables:
```

```
$ Probe_ID      : chr  "cg00000029" "cg00000108" "cg00000109" "cg00000165" ...
```

```
$ DESIGN        : chr  "II" "II" "II" "II" ...
```

```
$ COLOR_CHANNEL: chr  "Both" "Both" "Both" "Both" ...
```

```
> normdata <- normalize_asmn(normfactors = normfactors,
```

```
+      rawdata=NULL,
```

```
+      featuredata=NULL,
```

```
+      methylumidata=mset450k,
```

```
+      type="methylumi")
```

```
> show(normdata)
```

```
Object Information:
```

```
MethyLumiSet (storageMode: lockedEnvironment)
```

```
assayData: 485577 features, 10 samples
```

```
element names: betas, methylated, methylated.OOB, pvals, unmethylated, unmethylated.OOB
```

```
protocolData: none
```

```
phenoData
```

```
sampleNames: TCGA1 TCGA2 ... TCGA10 (10 total)
```

```
varLabels: barcode
```

```
varMetadata: labelDescription
```

```
featureData
```

```
featureNames: cg00000029 cg00000108 ... rs9839873 (485577 total)
```

```
fvarLabels: Probe_ID DESIGN COLOR_CHANNEL
```

```
fvarMetadata: labelDescription
```

```
experimentData: use 'experimentData(object)'
```

```
Annotation: IlluminaHumanMethylation450k
```

```
Major Operation History:
```

```
submitted finished
```

```
1 2014-10-13 20:43:33 2014-10-13 20:44:37
```

2 2014-10-13 20:44:38 2014-10-13 20:44:40

```
command
1 methylumIDAT(barcodes = getBarcodes(path = idatPath), idatPath = idatPath)
2 Subset of 485577 features.
```

For raw BeadStudio data, this function returns a `data.frame` of the beta values ordered by CpG site identifier. The methylated and unmethylated sites must be identified by "SignalA" and "SignalB" headers according to the BeadStudio documentation.

4 Other data types

Coercing the resulting `MethyLumiSet` object to a different data type after normalization can be achieved using the `as()` function. See the vignette for `methylumi` for other examples.

For example, coercing the new data set from above into a `MethyLumiM` object:

```
> normdataM <- as(normdata, 'MethyLumiM')
> show(normdataM)
```

```
MethyLumiM (storageMode: lockedEnvironment)
assayData: 485577 features, 10 samples
  element names: detection, exprs, methylated, unmethylated
protocolData: none
phenoData
  sampleNames: TCGA1 TCGA2 ... TCGA10 (10 total)
  varLabels: barcode
  varMetadata: labelDescription
featureData
  featureNames: cg00000029 cg00000108 ... rs9839873 (485577 total)
  fvarLabels: Probe_ID DESIGN COLOR_CHANNEL
  fvarMetadata: labelDescription
experimentData: use 'experimentData(object)'
Annotation: IlluminaHumanMethylation450k
```