Methylation status calling with METHimpute

Aaron Taudt *

*aaron.taudt@gmail.com

April 9, 2018

Contents

1 Introduction

Methimpute implements a powerful HMM-based binomial test for methylation status calling. Besides improved accuracy over the classical binomial test, the HMM allows imputation of the methylation status of **all cytosines** in the genome. It achieves this by borrowing information from neighboring covered cytosines. The confidence in the methylation status call is reported as well. The HMM can also be used to impute the methylation status for binned data instead of individual cytosines. Furthermore, *methimpute* also outputs context-specific conversion rates, which might be used to optimize the experimental procedure.

For the exact workings of *methimpute* we refer the interested reader to our publication at https://doi.org/10.1101/190223.

2 Methylation status calling on individual cytosines

The following examples explain the necessary steps for methylation status calling (and imputation). To keep the calculation time short, it uses only the first 200.000 bp of the Arabidopsis genome. The example consists of three steps: 1) Data import, 2) estimating the distance correlation and 3) methylation status calling. At the end of this example you will see that positions without counts are assigned a methylation status, but the confidence (column "posteriorMax") is generally quite low for those cytosines, whereas it is high for well-covered cytosines ($>=0.99$).

2.1 Separate-context model

The separate-context model runs a separate HMM for each context. This assumes that only within-context correlations are important, and between-context correlations do not need to be considered.

```
library(methimpute)

# ===== Step 1: Importing the data ===== #

# We load an example file in BSSeeker format that comes with the package
file <- system.file("extdata","arabidopsis_bsseeker.txt.gz", package="methimpute")
bsseeker.data <- importBSSeeker(file)
print(bsseeker.data)

## GRanges object with 110927 ranges and 2 metadata columns:
##              seqnames            ranges strand |  context   counts
##                 <Rle>         <IRanges>  <Rle> | <factor> <matrix>
##      [1]         chr1          [34, 34]      - |      CHG      0:4
##      [2]         chr1          [80, 80]      - |      CHH      2:9
##      [3]         chr1          [84, 84]      + |      CHH      1:1
##      [4]         chr1          [85, 85]      + |      CHH      1:1
##      [5]         chr1          [86, 86]      + |      CHH      1:1
##      ...          ...               ...    ... .      ...      ...
## [110923]         chr1 [533552, 533552]      - |       CG      2:2
## [110924]         chr1 [533554, 533554]      - |       CG      2:2
```

```
##   [110925]     chr1 [533595, 533595]     + |     CHG     0:1
##   [110926]     chr1 [533601, 533601]     + |     CHG     0:2
##   [110927]     chr1 [533614, 533614]     + |      CG     0:2
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths

# Because most methylation extractor programs report only covered cytosines,
# we need to inflate the data to inlcude all cytosines (including non-covered sites)
fasta.file <- system.file("extdata","arabidopsis_sequence.fa.gz", package="methimpute")
cytosine.positions <- extractCytosinesFromFASTA(fasta.file,
                                                contexts = c('CG','CHG','CHH'))
methylome <- inflateMethylome(bsseeker.data, cytosine.positions)
print(methylome)

## GRanges object with 199978 ranges and 2 metadata columns:
##          seqnames            ranges strand | context   counts
##             <Rle>         <IRanges>  <Rle> | <factor> <matrix>
##      [1]     chr1            [1, 1]     + |     CHH      0:0
##      [2]     chr1            [2, 2]     + |     CHH      0:0
##      [3]     chr1            [3, 3]     + |     CHH      0:0
##      [4]     chr1            [8, 8]     + |     CHH      0:0
##      [5]     chr1            [9, 9]     + |     CHH      0:0
##      ...      ...               ...    ... .     ...      ...
##  [199974]    chr1 [533554, 533554]     - |      CG      2:2
##  [199975]    chr1 [533557, 533557]     + |     CHH      0:0
##  [199976]    chr1 [533560, 533560]     + |      CG      0:0
##  [199977]    chr1 [533561, 533561]     - |      CG      0:0
##  [199978]    chr1 [533565, 533565]     - |     CHH      0:0
##   -------
##   seqinfo: 1 sequence from an unspecified genome

# ===== Step 2: Obtain correlation parameters ===== #

# The correlation of methylation levels between neighboring cytosines is an important
# parameter for the methylation status calling, so we need to get it first. Keep in mind
# that we only use the first 200.000 bp here, that's why the plot looks a bit meagre.
distcor <- distanceCorrelation(methylome, separate.contexts = TRUE)
fit <- estimateTransDist(distcor)
print(fit)

## $transDist
##        CG-CG      CHG-CHG      CHH-CHH
## 1.866779e+02 1.942438e+09 2.374003e+01
##
## $plot

## Warning:  Removed 23 rows containing missing values (geom_path).
```
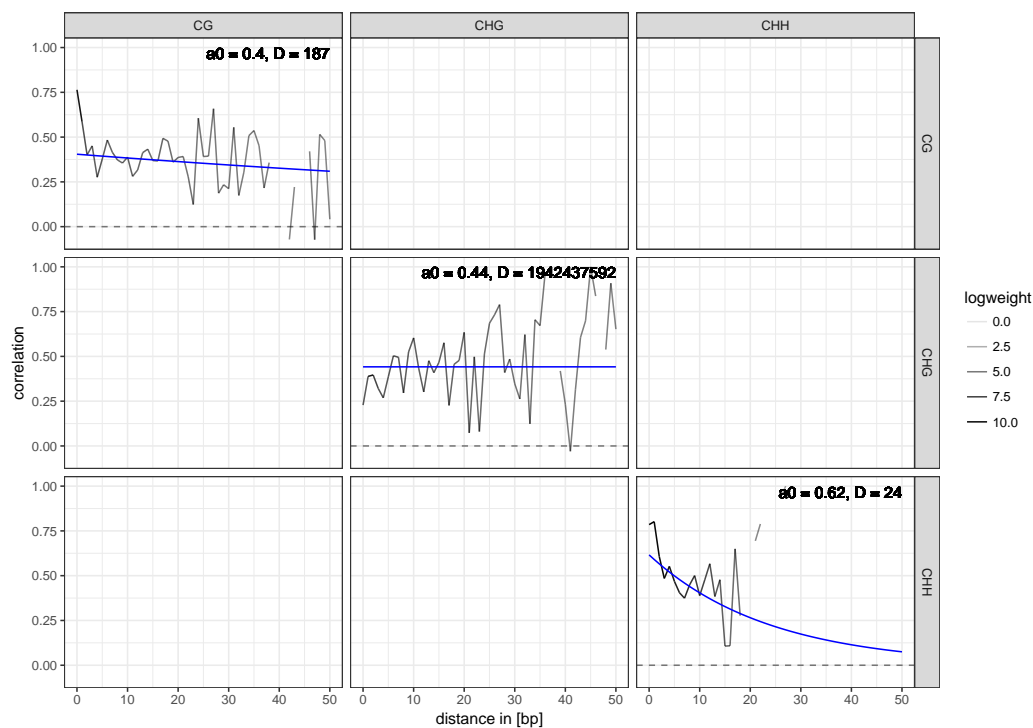
Methylation status calling with METHimpute


```
# ===== Step 3: Methylation status calling (and imputation) ===== #

model <- callMethylationSeparate(data = methylome, transDist = fit$transDist, verbosity = 0)
# The confidence in the methylation status call is given in the column "posteriorMax".
# For further analysis one could split the results into high-confidence
# (posteriorMax >= 0.98) and low-confidence calls (posteriorMax < 0.98) for instance.
print(model)

## GRanges object with 199978 ranges and 9 metadata columns:
##           seqnames           ranges strand |  context  counts  distance
##              <Rle>        <IRanges>  <Rle> | <factor> <matrix> <numeric>
##     [1]       chr1           [1, 1]      + |      CHH      0:0       Inf
##     [2]       chr1           [2, 2]      + |      CHH      0:0         0
##     [3]       chr1           [3, 3]      + |      CHH      0:0         0
##     [4]       chr1           [8, 8]      + |      CHH      0:0         4
##     [5]       chr1           [9, 9]      + |      CHH      0:0         0
##     ...        ...              ...    ... .      ...      ...       ...
## [199974]      chr1 [533554, 533554]      - |       CG      2:2         0
## [199975]      chr1 [533557, 533557]      + |      CHH      0:0         8
## [199976]      chr1 [533560, 533560]      + |       CG      0:0         5
## [199977]      chr1 [533561, 533561]      - |       CG      0:0         0
## [199978]      chr1 [533565, 533565]      - |      CHH      0:0         7
##         transitionContext posteriorMax posteriorMeth posteriorUnmeth
##                  <factor>    <numeric>     <numeric>       <numeric>
##     [1]              <NA>    0.5004034     0.5004034       0.4995966
##     [2]           CHH-CHH    0.6301245     0.3698755       0.6301245
##     [3]           CHH-CHH    0.7265858     0.2734142       0.7265858
##     [4]           CHH-CHH    0.7516365     0.2483635       0.7516365
##     [5]           CHH-CHH    0.8163712     0.1836288       0.8163712
##     ...               ...          ...           ...             ...
```

Methylation status calling with METHimpute

```
##   [199974]          CG-CG    0.9999755    0.9999755    2.451116e-05
##   [199975]        CHH-CHH    0.7742228    0.2257772    7.742228e-01
##   [199976]          CG-CG    0.9057913    0.9057913    9.420875e-02
##   [199977]          CG-CG    0.8304996    0.8304996    1.695004e-01
##   [199978]        CHH-CHH    0.7480418    0.2519582    7.480418e-01
##                   status rc.meth.lvl
##                 <factor>   <numeric>
##       [1]    Methylated  0.14214505
##       [2]  Unmethylated  0.10509472
##       [3]  Unmethylated  0.07771417
##       [4]  Unmethylated  0.07060353
##       [5]  Unmethylated  0.05222859
##       ...           ...         ...
##   [199974]    Methylated  0.85151435
##   [199975]  Unmethylated  0.06419243
##   [199976]    Methylated  0.77221618
##   [199977]    Methylated  0.70882454
##   [199978]  Unmethylated  0.07162390
##   -------
##   seqinfo: 1 sequence from an unspecified genome

# Bisulfite conversion rates can be obtained with
1 - model$params$emissionParams$Unmethylated

##          prob
## CG  0.9904123
## CHG 0.9907659
## CHH 0.9998944
```
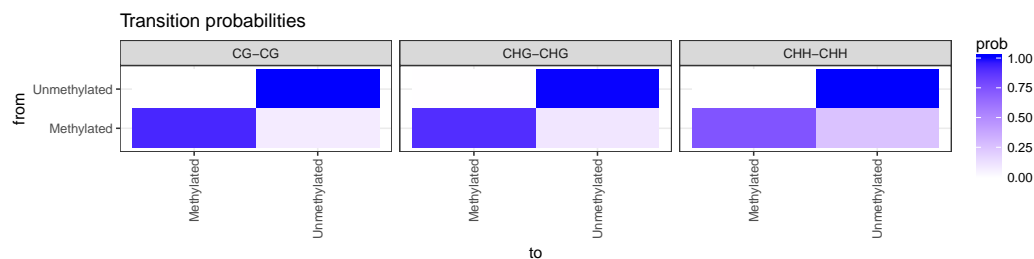
You can also check several properties of the fitted Hidden Markov Model, such as convergence or transition probabilities, and check how well the fitted distributions describe the data.
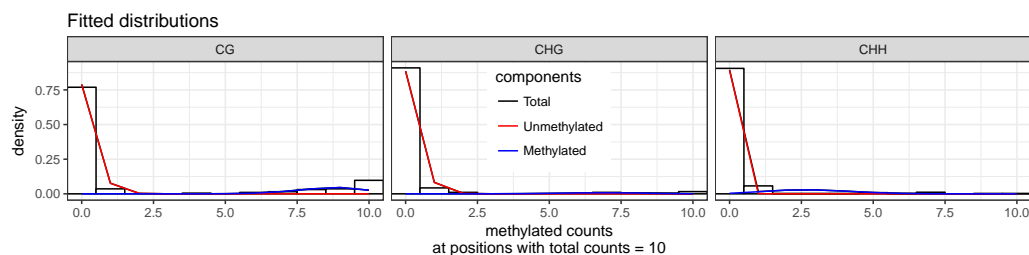
```
plotConvergence(model)
```
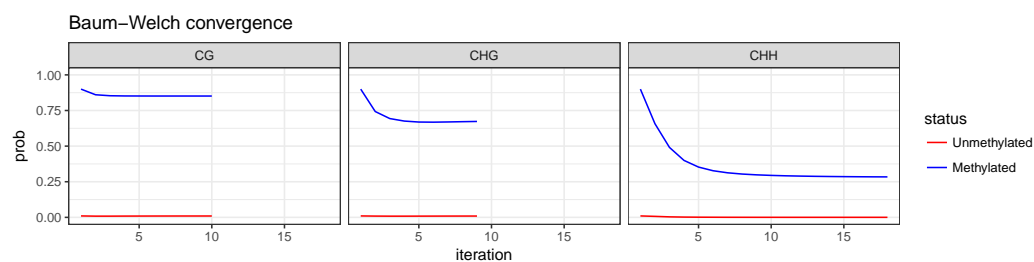


```
plotTransitionProbs(model)
```

```
plotHistogram(model, total.counts = 10)
```

Fitted distributions

2.2   Interacting-context model

The interacting-context model runs a single HMM for all contexts. This takes into account the within-context and between-context correlations and should be more accurate than the separate-context model if sufficient data is available. However, we have observed that in low coverage settings too much information from well covered contexts is diffusing into the low covered contexts (e.g. CHH and CHG will look like CG with very low coverage). In this case, please use the separate-context model in section 2.1.

```
library(methimpute)

# ===== Step 1: Importing the data ===== #

# We load an example file in BSSeeker format that comes with the package
file <- system.file("extdata","arabidopsis_bsseeker.txt.gz", package="methimpute")
bsseeker.data <- importBSSeeker(file)
print(bsseeker.data)

## GRanges object with 110927 ranges and 2 metadata columns:
##            seqnames            ranges strand |  context    counts
##               <Rle>         <IRanges>  <Rle> | <factor> <matrix>
##       [1]      chr1          [34, 34]      - |      CHG       0:4
##       [2]      chr1          [80, 80]      - |      CHH       2:9
##       [3]      chr1          [84, 84]      + |      CHH       1:1
##       [4]      chr1          [85, 85]      + |      CHH       1:1
##       [5]      chr1          [86, 86]      + |      CHH       1:1
##       ...       ...               ...    ... .      ...       ...
##  [110923]      chr1 [533552, 533552]      - |       CG       2:2
##  [110924]      chr1 [533554, 533554]      - |       CG       2:2
##  [110925]      chr1 [533595, 533595]      + |      CHG       0:1
##  [110926]      chr1 [533601, 533601]      + |      CHG       0:2
##  [110927]      chr1 [533614, 533614]      + |       CG       0:2
##  -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths

# Because most methylation extractor programs report only covered cytosines,
# we need to inflate the data to inlcude all cytosines (including non-covered sites)
fasta.file <- system.file("extdata","arabidopsis_sequence.fa.gz", package="methimpute")
cytosine.positions <- extractCytosinesFromFASTA(fasta.file,
                                                contexts = c('CG','CHG','CHH'))
methylome <- inflateMethylome(bsseeker.data, cytosine.positions)
```

```
print(methylome)

## GRanges object with 199978 ranges and 2 metadata columns:
##            seqnames           ranges strand |   context    counts
##               <Rle>        <IRanges>  <Rle> | <factor> <matrix>
##       [1]      chr1           [1, 1]      + |       CHH      0:0
##       [2]      chr1           [2, 2]      + |       CHH      0:0
##       [3]      chr1           [3, 3]      + |       CHH      0:0
##       [4]      chr1           [8, 8]      + |       CHH      0:0
##       [5]      chr1           [9, 9]      + |       CHH      0:0
##       ...       ...              ...    ... .       ...      ...
##  [199974]      chr1 [533554, 533554]      - |        CG      2:2
##  [199975]      chr1 [533557, 533557]      + |       CHH      0:0
##  [199976]      chr1 [533560, 533560]      + |        CG      0:0
##  [199977]      chr1 [533561, 533561]      - |        CG      0:0
##  [199978]      chr1 [533565, 533565]      - |       CHH      0:0
##   -------
##   seqinfo: 1 sequence from an unspecified genome

# ===== Step 2: Obtain correlation parameters ===== #

# The correlation of methylation levels between neighboring cytosines is an important
# parameter for the methylation status calling, so we need to get it first. Keep in mind
# that we only use the first 200.000 bp here, that's why the plot looks a bit meagre.
distcor <- distanceCorrelation(methylome)
fit <- estimateTransDist(distcor)
print(fit)

## $transDist
##        CG-CG       CG-CHG       CG-CHH      CHG-CHG      CHG-CHH      CHH-CHH
## 2.670669e+08 3.764952e+06 8.663764e+01 4.124081e+00 6.909617e+07 4.409969e+01
##
## $plot

## Warning:  Removed 24 rows containing missing values (geom_path).
```
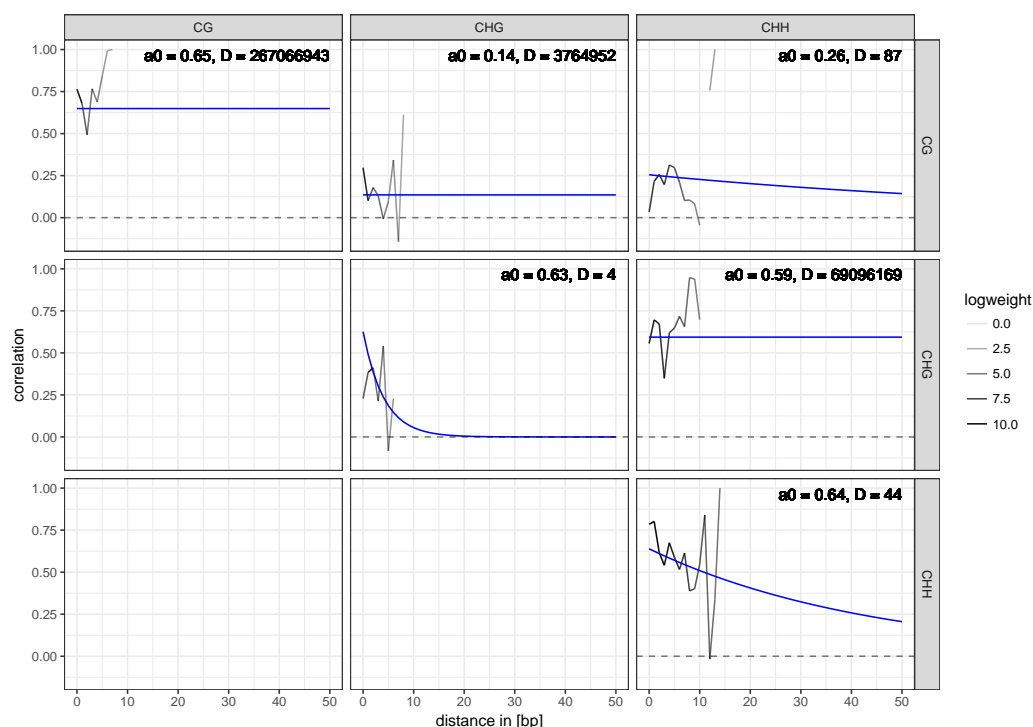
Methylation status calling with METHimpute


```
# ===== Step 3: Methylation status calling (and imputation) ===== #

model <- callMethylation(data = methylome, transDist = fit$transDist)

##  Iteration            log(P)          dlog(P)    Time in sec
##          0              -inf                -              0
##          1     -40631.538451              inf              0
##          2     -26304.340662     14327.197788              0
##          3     -24210.680437      2093.660225              0
##          4     -23635.136896       575.543541              1
##          5     -23374.140217       260.996679              1
##          6     -23224.427468       149.712749              1
##          7     -23134.096595        90.330873              1
##          8     -23079.925956        54.170639              1
##          9     -23047.280535        32.645421              1
##         10     -23027.416404        19.864131              1
##         11     -23015.215509        12.200895              1
##         12     -23007.665767         7.549743              1
##         13     -23002.971021         4.694745              1
##         14     -23000.044357         2.926664              1
##         15     -22998.218617         1.825740              1
##         16     -22997.079103         1.139514              1
##         17     -22996.365832         0.713271              2
## HMM: Convergence reached!

# The confidence in the methylation status call is given in the column "posteriorMax".
# For further analysis one could split the results into high-confidence
# (posteriorMax >= 0.98) and low-confidence calls (posteriorMax < 0.98) for instance.
print(model)

## GRanges object with 199978 ranges and 9 metadata columns:
```

Methylation status calling with METHimpute

```
##           seqnames             ranges strand | context   counts distance
##              <Rle>          <IRanges>  <Rle> | <factor> <matrix> <numeric>
##      [1]       chr1             [1, 1]      + |      CHH      0:0       Inf
##      [2]       chr1             [2, 2]      + |      CHH      0:0         0
##      [3]       chr1             [3, 3]      + |      CHH      0:0         0
##      [4]       chr1             [8, 8]      + |      CHH      0:0         4
##      [5]       chr1             [9, 9]      + |      CHH      0:0         0
##      ...        ...                ...    ... .      ...      ...       ...
## [199974]       chr1 [533554, 533554]      - |       CG      2:2         0
## [199975]       chr1 [533557, 533557]      + |      CHH      0:0         2
## [199976]       chr1 [533560, 533560]      + |       CG      0:0         2
## [199977]       chr1 [533561, 533561]      - |       CG      0:0         0
## [199978]       chr1 [533565, 533565]      - |      CHH      0:0         3
##          transitionContext posteriorMax posteriorMeth posteriorUnmeth
##                   <factor>    <numeric>     <numeric>       <numeric>
##      [1]              <NA>    0.6606448     0.3393552       0.6606448
##      [2]           CHH-CHH    0.7068376     0.2931624       0.7068376
##      [3]           CHH-CHH    0.7471539     0.2528461       0.7471539
##      [4]           CHH-CHH    0.7624977     0.2375023       0.7624977
##      [5]           CHH-CHH    0.7963517     0.2036483       0.7963517
##      ...               ...          ...           ...             ...
## [199974]            CG-CG    0.9999923     0.9999923    7.668704e-06
## [199975]           CG-CHH    0.5029823     0.4970177    5.029823e-01
## [199976]           CHH-CG    0.5294277     0.4705723    5.294277e-01
## [199977]            CG-CG    0.5595388     0.4404612    5.595388e-01
## [199978]           CG-CHH    0.7689333     0.2310667    7.689333e-01
##               status rc.meth.lvl
##             <factor>   <numeric>
##      [1] Unmethylated  0.11618208
##      [2] Unmethylated  0.10047755
##      [3] Unmethylated  0.08677090
##      [4] Unmethylated  0.08155434
##      [5] Unmethylated  0.07004477
##      ...          ...         ...
## [199974]   Methylated   0.8238826
## [199975] Unmethylated   0.1697838
## [199976] Unmethylated   0.3906849
## [199977] Unmethylated   0.3660465
## [199978] Unmethylated   0.0793664
## -------
##   seqinfo: 1 sequence from an unspecified genome

# Bisulfite conversion rates can be obtained with
1 - model$params$emissionParams$Unmethylated

##          prob
## CG  0.9943607
## CHG 0.9979215
## CHH 0.9991911
```

3   Methylation status calling on binned data

The following examples explain the necessary steps for methylation status calling (and imputation) on binned data, such as commonly used 100bp bins. To keep the calculation time short, it uses only the first 200.000 bp of the Arabidopsis genome. The example consists of four steps: 1) Data import, 2) binning and 3) methylation status calling.

```
library(methimpute)

# ===== Step 1: Importing the data ===== #

# We load an example file in BSSeeker format that comes with the package
file <- system.file("extdata","arabidopsis_bsseeker.txt.gz", package="methimpute")
bsseeker.data <- importBSSeeker(file)
print(bsseeker.data)

## GRanges object with 110927 ranges and 2 metadata columns:
##            seqnames            ranges strand |  context   counts
##               <Rle>         <IRanges>  <Rle> | <factor> <matrix>
##        [1]     chr1          [34, 34]      - |      CHG      0:4
##        [2]     chr1          [80, 80]      - |      CHH      2:9
##        [3]     chr1          [84, 84]      + |      CHH      1:1
##        [4]     chr1          [85, 85]      + |      CHH      1:1
##        [5]     chr1          [86, 86]      + |      CHH      1:1
##        ...      ...               ...    ... .      ...      ...
##   [110923]     chr1 [533552, 533552]      - |       CG      2:2
##   [110924]     chr1 [533554, 533554]      - |       CG      2:2
##   [110925]     chr1 [533595, 533595]      + |      CHG      0:1
##   [110926]     chr1 [533601, 533601]      + |      CHG      0:2
##   [110927]     chr1 [533614, 533614]      + |       CG      0:2
##   -------
##   seqinfo: 1 sequence from an unspecified genome; no seqlengths

# Because most methylation extractor programs report only covered cytosines,
# we need to inflate the data to inlcude all cytosines (including non-covered sites)
fasta.file <- system.file("extdata","arabidopsis_sequence.fa.gz", package="methimpute")
cytosine.positions <- extractCytosinesFromFASTA(fasta.file,
                                              contexts = c('CG','CHG','CHH'))
methylome <- inflateMethylome(bsseeker.data, cytosine.positions)
print(methylome)

## GRanges object with 199978 ranges and 2 metadata columns:
##            seqnames            ranges strand |  context   counts
##               <Rle>         <IRanges>  <Rle> | <factor> <matrix>
##        [1]     chr1            [1, 1]      + |      CHH      0:0
##        [2]     chr1            [2, 2]      + |      CHH      0:0
##        [3]     chr1            [3, 3]      + |      CHH      0:0
##        [4]     chr1            [8, 8]      + |      CHH      0:0
##        [5]     chr1            [9, 9]      + |      CHH      0:0
##        ...      ...               ...    ... .      ...      ...
##   [199974]     chr1 [533554, 533554]      - |       CG      2:2
##   [199975]     chr1 [533557, 533557]      + |      CHH      0:0
##   [199976]     chr1 [533560, 533560]      + |       CG      0:0
##   [199977]     chr1 [533561, 533561]      - |       CG      0:0
##   [199978]     chr1 [533565, 533565]      - |      CHH      0:0
```

```
##   -------
##   seqinfo: 1 sequence from an unspecified genome

# ===== Step 2: Binning into 100bp bins ===== #
binnedMethylome <- binMethylome(methylome, binsize = 100, contexts = c('total','CG'))
print(binnedMethylome$CG)

## GRanges object with 5335 ranges and 3 metadata columns:
##        seqnames            ranges strand | cytosines  context   counts
##           <Rle>         <IRanges>  <Rle> | <integer> <factor> <matrix>
##    [1]     chr1       [  1,  100]      * |         0       CG     7:19
##    [2]     chr1       [101,  200]      * |         6       CG    41:62
##    [3]     chr1       [201,  300]      * |         0       CG     3:58
##    [4]     chr1       [301,  400]      * |         2       CG     4:43
##    [5]     chr1       [401,  500]      * |         1       CG     0:19
##    ...      ...               ...    ... .       ...      ...      ...
## [5331]     chr1 [533001, 533100]      * |         0       CG     0:55
## [5332]     chr1 [533101, 533200]      * |        14       CG    3:171
## [5333]     chr1 [533201, 533300]      * |        10       CG     0:16
## [5334]     chr1 [533301, 533400]      * |         2       CG     0:35
## [5335]     chr1 [533401, 533500]      * |         8       CG     1:44
##   -------
##   seqinfo: 1 sequence from an unspecified genome

# ===== Step 3: Methylation status calling (and imputation) ===== #

binnedModel <- callMethylation(data = binnedMethylome$CG)

## Iteration            log(P)            dlog(P)   Time in sec
##       0               -inf                  -             0
##       1       -27231.884280              inf             0
##       2       -16567.652577    10664.231703             0
##       3       -14617.459829     1950.192748             0
##       4       -13264.059065     1353.400764             0
##       5       -12272.683299      991.375766             0
##       6       -11836.442261      436.241038             0
##       7       -11645.945339      190.496922             0
##       8       -11556.833035       89.112304             0
##       9       -11524.535678       32.297356             0
##      10       -11514.865975        9.669703             0
##      11       -11512.066277        2.799698             0
##      12       -11511.303950        0.762327             0
## HMM: Convergence reached!

print(binnedModel)

## GRanges object with 5335 ranges and 10 metadata columns:
##        seqnames            ranges strand | cytosines  context   counts
##           <Rle>         <IRanges>  <Rle> | <integer> <factor> <matrix>
##    [1]     chr1       [  1,  100]      * |         0       CG     7:19
##    [2]     chr1       [101,  200]      * |         6       CG    41:62
##    [3]     chr1       [201,  300]      * |         0       CG     3:58
##    [4]     chr1       [301,  400]      * |         2       CG     4:43
##    [5]     chr1       [401,  500]      * |         1       CG     0:19
##    ...      ...               ...    ... .       ...      ...      ...
## [5331]     chr1 [533001, 533100]      * |         0       CG     0:55
## [5332]     chr1 [533101, 533200]      * |        14       CG    3:171
## [5333]     chr1 [533201, 533300]      * |        10       CG     0:16
```

```
##    [5334]      chr1 [533301, 533400]      * |        2        CG      0:35
##    [5335]      chr1 [533401, 533500]      * |        8        CG      1:44
##          distance transitionContext posteriorMax posteriorMeth posteriorUnmeth
##         <numeric>         <factor>    <numeric>     <numeric>       <numeric>
##    [1]       Inf            <NA>    1.0000000    1.00000000   3.278506e-109
##    [2]         0           CG-CG    1.0000000    1.00000000    6.634096e-70
##    [3]         0           CG-CG    0.6422146    0.35778544    6.422146e-01
##    [4]         0           CG-CG    0.9742299    0.97422989    2.577011e-02
##    [5]         0           CG-CG    0.9649472    0.03505281    9.649472e-01
##    ...       ...             ...          ...           ...             ...
##    [5331]      0           CG-CG    1.0000000  4.023375e-08     1.0000000
##    [5332]      0           CG-CG    1.0000000  1.362313e-15     1.0000000
##    [5333]      0           CG-CG    0.9990175  9.824911e-04     0.9990175
##    [5334]      0           CG-CG    0.9999928  7.221815e-06     0.9999928
##    [5335]      0           CG-CG    0.9999085  9.153180e-05     0.9999085
##              status rc.meth.lvl
##            <factor>   <numeric>
##    [1]   Methylated  0.23143841
##    [2]   Methylated  0.23143841
##    [3] Unmethylated  0.08546719
##    [4]   Methylated  0.22558103
##    [5] Unmethylated  0.01211214
##    ...          ...         ...
## [5331] Unmethylated 0.004144878
## [5332] Unmethylated 0.004144868
## [5333] Unmethylated 0.004368182
## [5334] Unmethylated 0.004146510
## [5335] Unmethylated 0.004165673
##  -------
##  seqinfo: 1 sequence from an unspecified genome
```

You can also check several properties of the fitted Hidden Markov Model, such as convergence or transition probabilities, and check how well the fitted distributions describe the data. This last point is important because the binomial distributions that the HMM uses were originally meant to describe individual cytosines and not bins. However, we have observed that they still capture the bimodal distributions of methylation levels in binned data quite well. Note that the histogram for our example looks quite sparse due to the very low number of bins that were used.
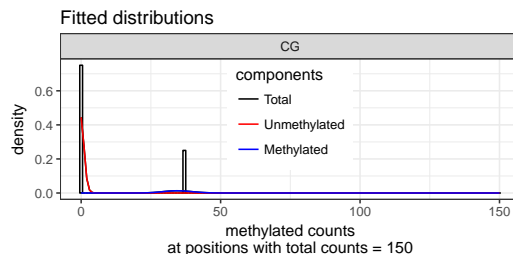
```
plotConvergence(binnedModel)
```
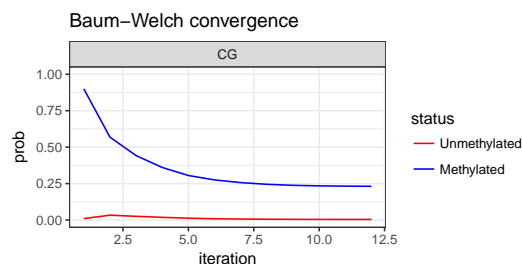


```
plotTransitionProbs(binnedModel)
```

Methylation status calling with METHimpute

Transition probabilities
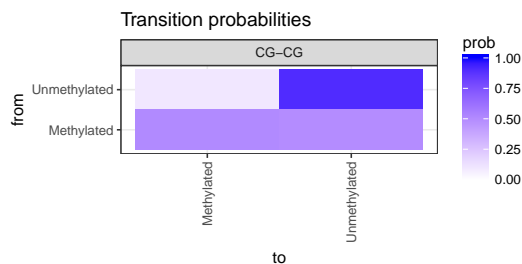
```
plotHistogram(binnedModel, total.counts = 150)
```


Fitted distributions

4 Description of columns in the output

- **context** The sequence context of the cytosine.
- **counts** Counts for methylated and total number of reads at each position.
- **distance** The distance in base-pairs from the previous to the current cytosine.
- **transitionContext** Transition context in the form "previous-current".
- **posteriorMax** Maximum posterior value of the methylation status call, can be interpreted as the confidence in the call.
- **posteriorMeth** Posterior value of the "methylated" component.
- **posteriorUnmeth** Posterior value of the "unmethylated" component.
- **status** Methylation status.
- **rc.meth.lvl** Recalibrated methylation level, calculated from the posteriors and the fitted parameters (see ?methimputeBinomialHMM for details).

5 Plots and enrichment analysis

This package also offers plotting functions for a simple enrichment analysis. Let's say we are interested in the methylation level around genes and transposable elements. We would also like to see how the imputation works on cytosines with missing data compared to covered cytosines.

```
# Define categories to distinguish missing from covered cytosines
model$data$category <- factor('covered', levels=c('missing', 'covered'))
model$data$category[model$data$counts[,'total']>=1] <- 'covered'
model$data$category[model$data$counts[,'total']==0] <- 'missing'

# Note that the plots look a bit ugly because our toy data has only 200.000 datapoints.
data(arabidopsis_genes)
seqlengths(model$data) <- seqlengths(arabidopsis_genes)[seqlevels(model$data)] # this
```
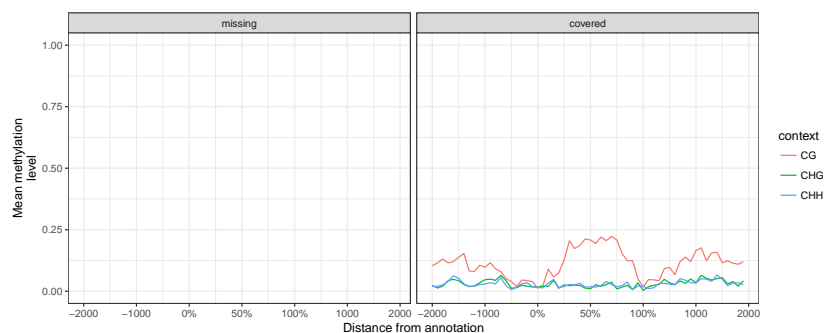
Methylation status calling with METHimpute

```
                                    # line should only be necessary for our toy example
plotEnrichment(model$data, annotation=arabidopsis_genes, range = 2000,
               category.column = 'category')
```
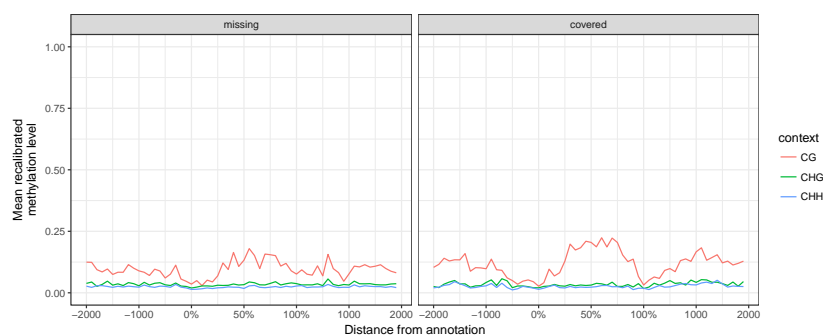
```
## $meth.lvl
```

```
## Warning:  Removed 180 rows containing missing values (geom_path).
```
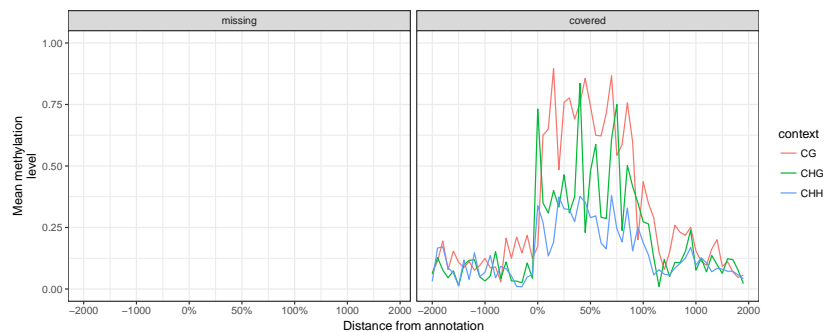


```
##
## $rc.meth.lvl
```



```
data(arabidopsis_TEs)
plotEnrichment(model$data, annotation=arabidopsis_TEs, range = 2000,
               category.column = 'category')
```
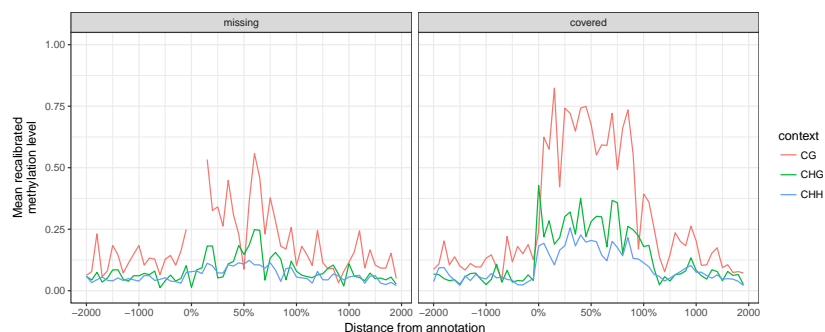
```
## $meth.lvl
```

```
## Warning:  Removed 180 rows containing missing values (geom_path).
```



```
##
## $rc.meth.lvl
```

6   Export results

You can export the results as TSV file with the following columns:

- chromosome, position, strand, context, counts.methylated, counts.total, posteriorMax, posteriorMeth, posteriorUnmeth, status, rc.meth.lvl

```
exportMethylome(model, filename = tempfile())
```

Please see section 4 for a description of the columns.

7   Session Info

```
toLatex(sessionInfo())
```

- R version 3.4.3 (2017-11-30), x86_64-apple-darwin15.6.0
- Locale: en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
- Running under: macOS High Sierra 10.13.3
- Matrix products: default
- BLAS: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRblas.0.dylib
- LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: BiocGenerics 0.24.0, devtools 1.13.4, GenomeInfoDb 1.14.0, GenomicRanges 1.30.3, ggplot2 2.2.1, IRanges 2.12.0, knitr 1.19, methimpute 1.1.1, S4Vectors 0.16.0
- Loaded via a namespace (and not attached): backports 1.1.2, BiocStyle 2.6.1, Biostrings 2.46.0, bitops 1.0-6, colorspace 1.3-2, compiler 3.4.3, digest 0.6.15, evaluate 0.10.1, GenomeInfoDbData 1.0.0, grid 3.4.3, gtable 0.2.0, highr 0.6, htmltools 0.3.6, labeling 0.3, lazyeval 0.2.1, magrittr 1.5, memoise 1.1.0, minpack.lm 1.2-1, munsell 0.4.3, pillar 1.1.0, plyr 1.8.4, Rcpp 0.12.16, RCurl 1.95-4.10, reshape2 1.4.3, rlang 0.1.6, rmarkdown 1.8, rprojroot 1.3-2, scales 0.5.0, stringi 1.1.6, stringr 1.2.0, tibble 1.4.2, tools 3.4.3, withr 2.1.1, XVector 0.18.0, yaml 2.1.16, zlibbioc 1.24.0