# AlphaBeta

*Y.Shahryary, Rashmi Hazarika, Frank Johannes*

*2019-08-16*

## Contents

# 1 Introduction

**AlphaBeta** is a computational method for estimating epimutation rates and spectra from high-throughput DNA methylation data in plants.

The method has been specifically designed to:

**1.** Analyze 'germline' epimutations in the context of multi-generational mutation accumulation lines (MA-lines).

**2.** Analyze 'somatic' epimutations in the context of plant development and aging.

Heritable changes in cytosine methylation can arise stochastically in plant genomes independently of DNA sequence alterations. These so-called 'spontaneous epimutations' appear to be a byproduct of imperfect DNA methylation maintenance during mitotic and meitotic cell divisions.

Accurate estimates of the rate and spectrum of these stochastic events are necessary to be able to quantify how epimutational processes shape methylome diversity in the context of plant evolution, development and aging.

Here we describe AlphaBeta, a computational method for estimating epimutation rates and spectra from pedigree-based high-throughput DNA methylation data in plants.

The method requires that the topology of the pedigree is known, which is typically the case in the construction of mutation accumulation lines (MA-lines) in sexually or clonally reproducing plant species.

However, the method also works for inferring somatic epimutations in long-lived perrenials, such as trees, using leaf methylomes and coring data as input. In this case, AlphaBeta treats the tree branching structure as an intra-organismal phylogeny of somatic lineages that carry information about the epimutational history of each branch.

# 2 Preparing Files

*NOTE: In this tutorial we are reading methylome files generated using Bioconductor package Methimpute:*

You can find more information here: Methimpute package

## 2.1 List of files

List of filenames containing generations and lineage.

*User must define "generations.fn" file. The structure of "generations.fn" should be same as in our example*

```
# Sample 'generations.fn' file
generation.fn <- system.file("extdata", "generations.fn", package = "AlphaBeta")
file <- fread(generation.fn)
head(file)
```

```
                filename generation lineage
  1:   data/G3_26_r1.txt          3      26
  2:   data/G3_87_r1.txt          3      87
  3:   data/G3_87_r2.txt          3      87
  4: data/G31_109_r1.txt         31     109
```

## 2.2 Generate divergence matrix

Estimating epimutation rates from high-throughput DNA methylation data. Generation of divergence matrix and calculation of methylation levels.

```
dMatrix(genTable = generation.fn, cytosine = "CG", posteriorMaxFilter = 0.99)
```

```
# Sample output from dMatrix function
head(fread(system.file("extdata/dm", "AB-dMatrix-CG-0.99.csv", package = "AlphaBeta")))
```

```
     pair-1     pair-2     D-value
1: G3_26_r1   G3_87_r1 0.005516667
2: G3_26_r1   G3_87_r2 0.005856857
3: G3_26_r1 G31_109_r1 0.011792749
4: G3_26_r1 G31_109_r2 0.009345341
5: G3_26_r1 G31_119_r1 0.010905316
6: G3_26_r1 G31_119_r2 0.011464732
```

## 2.3 Generate methylation proportions

```
rc.meth.lvl(genTable = generation.fn, cytosine = "CG", posteriorMaxFilter = 0.99)
```

```
# Sample output from proportions function
head(fread(system.file("extdata/dm", "AB-methprop-CG-0.99.csv", package = "AlphaBeta")))
```

```
   Sample_name context rc.meth.lvls
1:    G3_26_r1      CG    0.2542201
2:    G3_87_r1      CG    0.2522355
3:    G3_87_r2      CG    0.2524761
4:  G31_109_r1      CG    0.2482041
5:  G31_109_r2      CG    0.2654014
6:  G31_119_r1      CG    0.2623544
```

## 2.4 Information about Sample file.

This file containing information on generation times and pedigree lineages.

(should be provide manually)

```
# Sample file
head(fread(system.file("extdata/dm", "sampleInfo.csv", package = "AlphaBeta")))
```

```
       Sample Generation Lineage
1:   G3_26_r1          3      26
2:   G3_87_r1          3      87
3:   G3_87_r2          3      87
4: G31_109_r1         31     109
5: G31_109_r2         31     109
6: G31_119_r1         31     119
```

## 2.5 File containing lineage branch points.

(should be provide manually)

```
# Sample file
head(fread(system.file("extdata/dm", "branchPoints.csv", package = "AlphaBeta")))
```

```
   BP Generation Lineage
1:  1          0    none
2:  2          2      87
3:  3         30     109
4:  4         30     119
5:  5         30      29
6:  6         30      39
```

# 3 Germline epimutations

Models ABneutral, ABselectMM and ABselectUU can be used to estimate the rate of spontaneous epimutations from pedigree-based high-throughput DNA methylation data. The models are generally designed for pedigree data arising from selfing diploid species.

## 3.1 Calculate divergence times

Divergence time (delta t) is calculated as follows: delta t = t1 + t2 - 2*t0, where t1 is the time of sample 1 (in generations), t2 is the time of sample 2 (in generations) and t0 is the time (in generations) of the most recent common founder of samples 1 and 2.

To calculate divergence times of the pedigree should be provided in the form of 4 files as shown below.

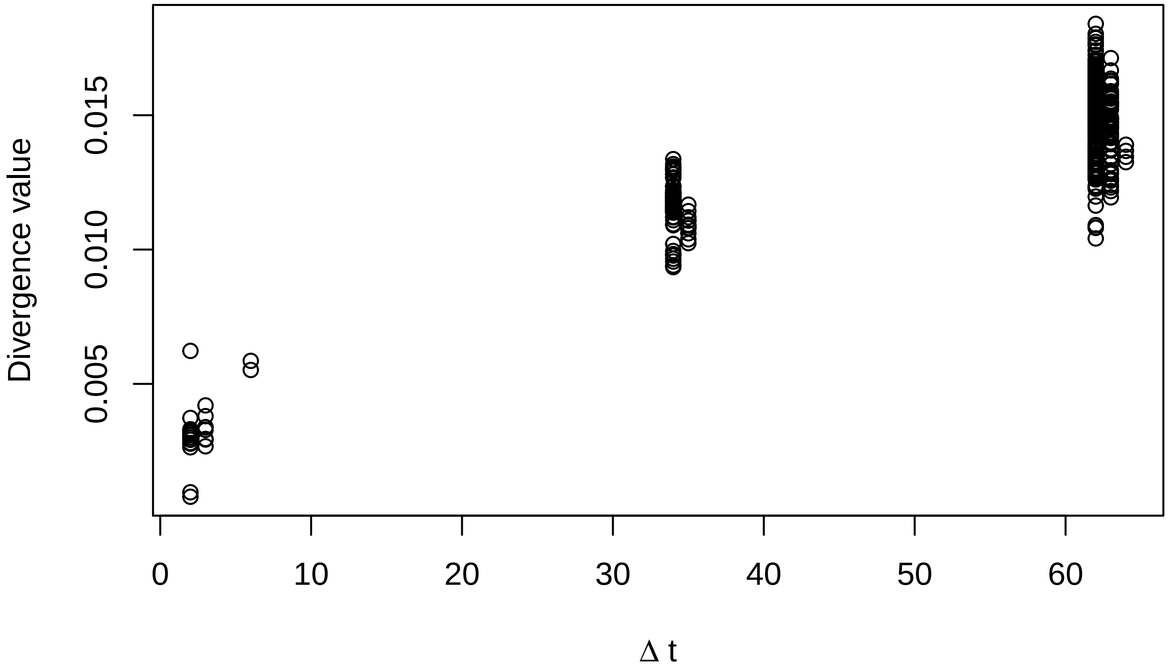
```
props.name <- read.table(system.file("extdata/dm", "AB-methprop-CG-0.99.csv", package = "AlphaBeta"),
    sep = "\t", header = TRUE)
sample.info <- read.table(system.file("extdata/dm", "sampleInfo.csv", package = "AlphaBeta"),
    sep = "\t", header = TRUE)
branch.points <- read.table(system.file("extdata/dm", "branchPoints.csv", package = "AlphaBeta"),
    sep = "\t", header = TRUE)
dmatrix <- read.table(system.file("extdata/dm", "AB-dMatrix-CG-0.99.csv", package = "AlphaBeta"),
    sep = "\t", header = TRUE)
context <- "CG"
```

calculate divergence times of the pedigree:

```
pedigree <- convertDMATRIX(sample.info = sample.info, branch.points = branch.points,
    dmatrix = dmatrix, design = "sibling")
head(pedigree)
```

```
     time0 time1 time2     D.value
[1,]     0     3     3 0.005516667
[2,]     0     3     3 0.005856857
[3,]     0     3    31 0.011792749
[4,]     0     3    31 0.009345341
[5,]     0     3    31 0.010905316
[6,]     0     3    31 0.011464732
```

This is a manual step for inspecting the divergence data and removing outlier samples (if any):



Read in the proportions data:

```
outliers <- "none"
dmatrix <- dmatrix[which(dmatrix[, 1] != outliers), ]
dmatrix <- dmatrix[which(dmatrix[, 2] != outliers), ]
pedigree <- pedigree[c(as.numeric(rownames(dmatrix))), ]

props <- props.name[which(as.character(props.name[, 2]) == context), ]
props <- props.name[which(!is.element(props.name[, 1], outliers) == TRUE), ]
```

Calculate initial proportions of unmethylated cytosines after removal of outliers:

```
p0uu_in <- 1 - mean(as.numeric(as.character(props[, 3])))
p0uu_in
```

```
  [1] 0.7435074
```

## 3.2 Run Models

### 3.2.1 Run Model with no selection (ABneutral)

This model assumes that heritable gains and losses in cytosine methylation are selectively neutral.

```
# output directory
output.data.dir <- paste0(getwd(), "/")

output <- ABneutral(pedigree.data = pedigree, p0uu = p0uu_in, eqp = p0uu_in, eqp.weight = 1,
    Nstarts = 2, out.dir = output.data.dir, out.name = "CG_global_estimates_ABneutral")
```

```
  Progress: 0.5

  Progress: 1
```

*NOTE: it is recommended to use at least 50 Nstarts to achieve best solutions*

Showing summary output of only output:

```
summary(output)
```

```
                    Length Class      Mode
  estimates           20   data.frame list
  estimates.flagged   20   data.frame list
  pedigree          2457   -none-     numeric
  settings             2   data.frame list
  model                1   -none-     character
  for.fit.plot       315   -none-     numeric
```

```
head(output$pedigree)
```

```
       time0 time1 time2     div.obs delta.t    div.pred      residual
  [1,]     0     3     3 0.005516667       6 0.006941634 -1.424967e-03
  [2,]     0     3     3 0.005856857       6 0.006941634 -1.084777e-03
  [3,]     0     3    31 0.011792749      34 0.010996737  7.960122e-04
  [4,]     0     3    31 0.009345341      34 0.010996737 -1.651396e-03
  [5,]     0     3    31 0.010905316      34 0.010996737 -9.142122e-05
  [6,]     0     3    31 0.011464732      34 0.010996737  4.679954e-04
```

### 3.2.2 Run model with selection against spontaneous gain of methylation (ABselectMM)

This model assumes that heritable losses of cytosine methylation are under negative selection. The selection parameter is estimated.

```
output <- ABselectMM(pedigree.data = pedigree, p0uu = p0uu_in, eqp = p0uu_in, eqp.weight = 1,
    Nstarts = 2, out.dir = output.data.dir, out.name = "CG_global_estimates_ABselectMM")
```

```
## Progress: 0.5

## Progress: 1
```

```r
summary(output)
```

```
##                   Length Class      Mode     
## estimates           22   data.frame list     
## estimates.flagged   22   data.frame list     
## pedigree          2457   -none-     numeric  
## settings             2   data.frame list     
## model                1   -none-     character
## for.fit.plot       315   -none-     numeric  
```

### 3.2.3 Run model with selection against spontaneous loss of methylation (ABselectUU)

This model assumes that heritable gains of cytosine methylation are under negative selection. The selection parameter is estimated.

```r
output <- ABselectUU(pedigree.data = pedigree, p0uu = p0uu_in, eqp = p0uu_in, eqp.weight = 1,
    Nstarts = 2, out.dir = output.data.dir, out.name = "CG_global_estimates_ABselectUU")
```

```
## Progress: 0.5

## Progress: 1
```

```r
summary(output)
```

```
##                   Length Class      Mode     
## estimates           22   data.frame list     
## estimates.flagged   22   data.frame list     
## pedigree          2457   -none-     numeric  
## settings             2   data.frame list     
## model                1   -none-     character
## for.fit.plot       315   -none-     numeric  
```

### 3.2.4 Run model that considers no accumulation of epimutations (ABnull)

This is the null model of no accumulation.

```r
output <- ABnull(pedigree.data = pedigree, out.dir = output.data.dir, out.name = "CG_global_estimates_ABnull")

summary(output)
```

```
##                   Length Class  Mode     
## estimates            1   -none- numeric  
## estimates.flagged    0   -none- NULL     
## pedigree          2457   -none- numeric  
## settings             0   -none- NULL     
## model                1   -none- character
## for.fit.plot      1755   -none- numeric  
```

## 3.3 Comparison of different models and selection of best model

### 3.3.1 Testing ABneutral vs. ABnull

```r
file1 <- system.file("extdata/models/", "CG_global_estimates_ABneutral.Rdata", package = "AlphaBeta")
file2 <- system.file("extdata/models/", "CG_global_estimates_ABnull.Rdata", package = "AlphaBeta")

out <- FtestRSS(pedigree.select = file1, pedigree.null = file2)
```

```
out$Ftest
```

```
        RSS_F         RSS_R          df_F          df_R        Fvalue        pvalue
 7.084342e-04  4.124786e-03  3.460000e+02  3.500000e+02  4.171374e+02 6.260446e-131
```

### 3.3.2 Testing ABselectMM vs.ABneutral

```
file1 <- system.file("extdata/models/", "CG_global_estimates_ABselectMM.Rdata", package = "AlphaBeta")
file2 <- system.file("extdata/models/", "CG_global_estimates_ABnull.Rdata", package = "AlphaBeta")

out <- FtestRSS(pedigree.select = file1, pedigree.null = file2)

out$Ftest
```

```
        RSS_F         RSS_R          df_F          df_R        Fvalue        pvalue
 6.507729e-04  4.124786e-03  3.460000e+02  3.500000e+02  4.617618e+02 2.662626e-137
```

### 3.3.3 Testing ABselectUU vs.ABneutral

```
file1 <- system.file("extdata/models/", "CG_global_estimates_ABselectUU.Rdata", package = "AlphaBeta")
file2 <- system.file("extdata/models/", "CG_global_estimates_ABnull.Rdata", package = "AlphaBeta")

out <- FtestRSS(pedigree.select = file1, pedigree.null = file2)

out$Ftest
```

```
        RSS_F         RSS_R          df_F          df_R        Fvalue        pvalue
 6.509786e-04  4.124786e-03  3.460000e+02  3.500000e+02  4.615886e+02 2.812040e-137
```

## 3.4 Bootstrap analysis with the best model

i.e ABneutral in our case

```
inputModel <- system.file("extdata/models/", "CG_global_estimates_ABneutral.Rdata",
    package = "AlphaBeta")

# Bootstrapping models CG
output.data.dir <- paste0(getwd(), "/")

Boutput <- BOOTmodel(pedigree.data = inputModel, Nboot = 2, out.dir = output.data.dir,
    out.name = "Boot_CG_global_estimates_ABneutral")
```

```
Bootstrap interation: 0.5
```

```
Bootstrap interation: 1
```

```
summary(Boutput)
```

```
                Length Class      Mode     
standard.errors 24     -none-     numeric  
boot.base       20     data.frame list     
settings         2     data.frame list     
N.boots          1     -none-     numeric  
N.good.boots     1     -none-     numeric  
boot.results    19     data.frame list     
model            1     -none-     character
```

```
Boutput$standard.errors
```

```
                     SE        2.5%        97.5%
alpha      6.268189e-06 0.0001069090 0.0001153303
beta       1.823272e-05 0.0003104196 0.0003349153
beta/alpha 2.816994e-04 2.9035865373 2.9039630093
weight     3.589985e-03 0.0183813955 0.0232045509
intercept  6.837474e-05 0.0023625633 0.0024544250
PrMMinf    3.710272e-05 0.2558070694 0.2558569170
PrUMinf    3.723571e-05 0.0006355098 0.0006855361
PrUUinf    1.329971e-07 0.7435073946 0.7435075733
```

# 4 Somatic epimutations

Models ABneutralSOMA, ABselectMMSOMA and ABselectUUSOMA can be used to estimate the rate of spontaneous epimutations from pedigree-based high-throughput DNA methylation data. The models are generally designed for pedigree data arising from clonally or asexually propagated diploid species. The models can also be applied to long-lived perrenials, such as trees, using leaf methylomes and coring data as input. In this case, the tree branching structure is treated as an intra-organismal pedigree (or phylogeny) of somatic lineages.

## 4.1 Loading data and generation of pedigree

```r
props.name <- read.table(system.file("extdata/soma/", "AB-methprop-CG-0.99.csv",
    package = "AlphaBeta"), sep = "\t", header = TRUE, stringsAsFactors = FALSE)
sample.info <- read.table(system.file("extdata/soma/", "sampleInfo.csv", package = "AlphaBeta"),
    sep = "\t", header = TRUE, stringsAsFactors = FALSE)
dmatrix <- read.table(system.file("extdata/soma/", "AB-dMatrix-CG-0.99.csv", package = "AlphaBeta"),
    sep = "\t", header = TRUE, stringsAsFactors = FALSE)
```

Samples:

```r
head(props.name)
```

```
  sample_name context Unmethylated
1        13_1      CG    0.5473480
2        13_2      CG    0.5473662
3        13_3      CG    0.5476725
4        13_5      CG    0.5474322
5        14_2      CG    0.5475183
6        14_3      CG    0.5475606
```

```r
head(sample.info)
```

```
  sample_name Branch_date Branchpoint_date Stem
1        13_1          29               31   13
2        13_2          41               44   13
3        13_3          70               72   13
4        13_5          80              113   13
5        14_2          35               41   14
6        14_3          41               47   14
```

```r
head(dmatrix)
```

```
  Pair_1 Pair_2     D.value
1   13_5   14_5 0.003796614
2   13_5   14_3 0.003974756
3   13_2   14_5 0.003995156
4   13_5   14_4 0.004040671
5   13_1   14_5 0.004046553
6   13_3   14_5 0.004048672
```

## 4.2 Generate pedigree from the input files

*NOTE: Here in our example "makePHYLO" function calculates divergence times of a tree branching structure with total age of 330 years derived from coring-based measurements.*

```
# 'tall' is total age of tree
pedigree.out <- makePHYLO(tall = 330, pedigree = dmatrix, sample.info = sample.info)
pedigree.out <- pedigree.out[[1]]
head(pedigree.out)
```

```
       time0 time1 time2     D.value
  [1,]     0   297   287 0.003796614
  [2,]     0   297   324 0.003974756
  [3,]     0   327   287 0.003995156
  [4,]     0   297   287 0.004040671
  [5,]     0   328   287 0.004046553
  [6,]     0   328   287 0.004048672
```

## 4.3 Calculate the proportion of unmethylated cytosines

```
p0uu_in <- mean(props[, 3])
p0uu_in
```

```
  [1] 0.2564926
```

## 4.4 Run Models

### 4.4.1 Run Model with no selection (ABneutralSOMA)

This model assumes that somatically heritable gains and losses in cytosine methylation are selectively neutral.

```
outneutral <- ABneutralSOMA(pedigree.data = pedigree.out, p0uu = p0uu_in, eqp = p0uu_in,
    eqp.weight = 0.001, Nstarts = 2, out.dir = output.data.dir, out.name = "ABneutralSOMA_CG_estimates")
```

```
  Progress: 0.5

  Progress: 1
```

```
summary(outneutral)
```

```
                    Length Class      Mode
  estimates           20   data.frame list
  estimates.flagged   20   data.frame list
  pedigree           196   -none-     numeric
  settings             2   data.frame list
  model                1   -none-     character
  for.fit.plot      3275   -none-     numeric
```

```
head(outneutral$pedigree)
```

```
       time0 time1 time2     div.obs delta.t div.pred residual
  [1,]     0   297   287 0.003796614     584       NA       NA
  [2,]     0   297   324 0.003974756     621       NA       NA
  [3,]     0   327   287 0.003995156     614       NA       NA
  [4,]     0   297   287 0.004040671     584       NA       NA
  [5,]     0   328   287 0.004046553     615       NA       NA
  [6,]     0   328   287 0.004048672     615       NA       NA
```

### 4.4.2 Run model with selection against spontaneous gain of methylation (ABselectMMSOMA)

This model assumes that somatically heritable losses of cytosine methylation are under negative selection. The selection parameter is estimated.

```
outselectMM <- ABselectMMSOMA(pedigree.data = pedigree.out, p0uu = p0uu_in, eqp = p0uu_in,
    eqp.weight = 0.001, Nstarts = 2, out.dir = output.data.dir, out.name = "ABselectMMSOMA_CG_estimates")
```

```
  Progress: 0.5

  Progress: 1
```

```
summary(outselectMM)
```

```
                    Length Class      Mode
  estimates           22   data.frame list
  estimates.flagged   22   data.frame list
  pedigree           196   -none-     numeric
  settings             2   data.frame list
  model                1   -none-     character
  for.fit.plot      3275   -none-     numeric
```

### 4.4.3 Run model with selection against spontaneous loss of methylation (ABselectUUSOMA)

This model assumes that somatically heritable gains of cytosine methylation are under negative selection. The selection parameter is estimated.

```
outselectUU <- ABselectUUSOMA(pedigree.data = pedigree.out, p0uu = p0uu_in, eqp = p0uu_in,
    eqp.weight = 0.001, Nstarts = 2, out.dir = output.data.dir, out.name = "ABselectUUSOMA_CG_estimates")
```

```
  Progress: 0.5

  Progress: 1
```

```
summary(outselectUU)
```

```
                    Length Class      Mode
  estimates           22   data.frame list
  estimates.flagged   22   data.frame list
  pedigree           196   -none-     numeric
  settings             2   data.frame list
  model                1   -none-     character
  for.fit.plot      3275   -none-     numeric
```

## 5 R session info

```
sessionInfo()
```

```
  R version 3.6.1 (2019-07-05)
  Platform: x86_64-pc-linux-gnu (64-bit)
  Running under: Ubuntu 18.04.3 LTS

  Matrix products: default
  BLAS:   /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
  LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.2.20.so

  locale:
   [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8        LC_COLLATE=C
   [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8    LC_PAPER=en_US.UTF-8       LC_NAME=C
   [9] LC_ADDRESS=C               LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] data.table_1.12.2 AlphaBeta_0.99.2

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.2          formatR_1.7         compiler_3.6.1      pillar_1.4.2        prettyunits_1.0.2
 [6] remotes_2.1.0       tools_3.6.1         testthat_2.2.1      digest_0.6.20       pkgbuild_1.0.4
[11] pkgload_1.0.2       evaluate_0.14       lattice_0.20-38     memoise_1.1.0       tibble_2.1.3
[16] pkgconfig_2.0.2     rlang_0.4.0         Matrix_1.2-17       cli_1.1.0           rstudioapi_0.10
[21] commonmark_1.7      yaml_2.2.0          parallel_3.6.1      expm_0.999-4        xfun_0.8
[26] knitr_1.24          withr_2.1.2         stringr_1.4.0       dplyr_0.8.3         roxygen2_6.1.1
[31] xml2_1.2.2          gtools_3.8.1        desc_1.2.0          fs_1.3.1            devtools_2.1.0
[36] grid_3.6.1          rprojroot_1.3-2     tidyselect_0.2.5    glue_1.3.1          R6_2.4.0
[41] processx_3.4.1      BiocParallel_1.19.2 rmarkdown_1.14      sessioninfo_1.1.1   callr_3.3.1
[46] purrr_0.3.2         magrittr_1.5        htmltools_0.3.6     backports_1.1.4     ps_1.3.0
[51] usethis_1.5.1       assertthat_0.2.1    numDeriv_2016.8-1.1 optimx_2018-7.10    stringi_1.4.3
[56] crayon_1.3.4
```