AlphaBeta

Y.Shahryary, Rashmi Hazarika, Frank Johannes

2019-11-07

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)
```

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")))
```

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")))
```

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")))
```

2.5 File containing lineage branch points.

(should be provide manually)

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

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)
```

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
```

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")
```

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

Showing summary output of only output:

```
summary(output)
```

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

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")

summary(output)
```

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.

```
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")

summary(output)
```

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

This is the null model of no accumulation.

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

summary(output)
```

3.3 Comparison of different models and selection of best model

3.3.1 Testing ABneutral vs. ABnull

```
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
```

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
```

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
```

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")

summary(Boutput)
```

```
Boutput$standard.errors
```

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

```
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:

```
head(props.name)
head(sample.info)
head(dmatrix)
```

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)
```

4.3 Calculate the proportion of unmethylated cytosines

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

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")
summary(outneutral)
```

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

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")

summary(outselectMM)
```

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")
summary(outselectUU)
```

5 R session info

```
sessionInfo()
```