Sequence analysis

Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data

Kemal Akman1,*, Thomas Haaf2, Silvia Gravina3, Jan Vijg3 and Achim Tresch1

1Tresch Group, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany, 2AG Haaf, Institute of Human Genetics, Julius Maximilians University, 97070 Wuerzburg, Germany and 3Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

1 Received on December 26, 2013; revised on February 16, 2014; accepted on March 4, 2014

2 USAGE AND APPLICATION

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite (Frommer et al., 1992). The method is based on bisulfite conversion of unmethylated cytosines into uracil and has become a standard in DNA methylation profiling. Its advantage is accuracy, as the degree of methylation at each cytosine can be quantified with great precision (Fraga and Esteller, 2002). More recently, bisulfite sequencing has been applied in a genome-wide manner, which requires advanced computational analysis for determining DNA methylation patterns and changes therein. Thus far, such analysis has been limited to the comparison of individual CpG sites between samples. However, because of the bisulfite conversion of cytosines into uracils, and eventually thymines, sequence complexity is much reduced, with no account for incomplete conversion and/or sequencing errors. Bisulfite sequencing also often suffers from low coverage. Here, we present BS-Seq Epimutation Analysis Toolkit (BEAT), a novel tool for analyzing bisulfite-converted DNA sequences. To overcome the aforementioned limitations in the estimation of methylation rates, BEAT aggregates data from consecutive cytosines into regions by using a Bayesian binomial-beta mixture model. The model is derived, described in detail and evaluated in our Supplementary Material. For each region, it calculates a posterior methylation probability distribution that can be used for the comparison of DNA methylation between samples. Anticipating technological progress in the DNA methylation field, BEAT includes an error model adapted to single-cell BS-Seq data.

1 INTRODUCTION

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite (Frommer et al., 1992). The method is based on bisulfite conversion of unmethylated cytosines into uracil and has become a standard in DNA methylation profiling. Its advantage is accuracy, as the degree of methylation at each cytosine can be quantified with great precision (Fraga and Esteller, 2002). More recently, bisulfite sequencing has been applied in a genome-wide manner, which requires advanced computational analysis for determining DNA methylation patterns and changes therein. Thus far, such analysis has been limited to the comparison of individual CpG sites between samples. However, because of the bisulfite conversion of cytosines into uracils, and eventually thymines, sequence complexity is much reduced, with no account for incomplete conversion and/or sequencing errors. Bisulfite sequencing also often suffers from low coverage. Here, we present BS-Seq Epimutation Analysis Toolkit (BEAT), a novel tool for analyzing bisulfite-converted DNA sequences. To overcome the aforementioned limitations in the estimation of methylation rates, BEAT aggregates data from consecutive cytosines into regions by using a Bayesian binomial-beta mixture model. The model is derived, described in detail and evaluated in our Supplementary Material. For each region, it calculates a posterior methylation probability distribution that can be used for the comparison of DNA methylation between samples. Anticipating technological progress in the DNA methylation field, BEAT includes an error model adapted to single-cell BS-Seq data.

1 INTRODUCTION

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite (Frommer et al., 1992). The method is based on bisulfite conversion of unmethylated cytosines into uracil and has become a standard in DNA methylation profiling. Its advantage is accuracy, as the degree of methylation at each cytosine can be quantified with great precision (Fraga and Esteller, 2002). More recently, bisulfite sequencing has been applied in a genome-wide manner, which requires advanced computational analysis for determining DNA methylation patterns and changes therein. Thus far, such analysis has been limited to the comparison of individual CpG sites between samples. However, because of the bisulfite conversion of cytosines into uracils, and eventually thymines, sequence complexity is much reduced, with no account for incomplete conversion and/or sequencing errors. Bisulfite sequencing also often suffers from low coverage. Here, we present BS-Seq Epimutation Analysis Toolkit (BEAT), a novel tool for analyzing bisulfite-converted DNA sequences. To overcome the aforementioned limitations in the estimation of methylation rates, BEAT aggregates data from consecutive cytosines into regions by using a Bayesian binomial-beta mixture model. The model is derived, described in detail and evaluated in our Supplementary Material. For each region, it calculates a posterior methylation probability distribution that can be used for the comparison of DNA methylation between samples. Anticipating technological progress in the DNA methylation field, BEAT includes an error model adapted to single-cell BS-Seq data.

1 INTRODUCTION

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite (Frommer et al., 1992). The method is based on bisulfite conversion of unmethylated cytosines into uracil and has become a standard in DNA methylation profiling. Its advantage is accuracy, as the degree of methylation at each cytosine can be quantified with great precision (Fraga and Esteller, 2002). More recently, bisulfite sequencing has been applied in a genome-wide manner, which requires advanced computational analysis for determining DNA methylation patterns and changes therein. Thus far, such analysis has been limited to the comparison of individual CpG sites between samples. However, because of the bisulfite conversion of cytosines into uracils, and eventually thymines, sequence complexity is much reduced, with no account for incomplete conversion and/or sequencing errors. Bisulfite sequencing also often suffers from low coverage. Here, we present BS-Seq Epimutation Analysis Toolkit (BEAT), a novel tool for analyzing bisulfite-converted DNA sequences. To overcome the aforementioned limitations in the estimation of methylation rates, BEAT aggregates data from consecutive cytosines into regions by using a Bayesian binomial-beta mixture model. The model is derived, described in detail and evaluated in our Supplementary Material. For each region, it calculates a posterior methylation probability distribution that can be used for the comparison of DNA methylation between samples. Anticipating technological progress in the DNA methylation field, BEAT includes an error model adapted to single-cell BS-Seq data.

1 INTRODUCTION

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite (Frommer et al., 1992). The method is based on bisulfite conversion of unmethylated cytosines into uracil and has become a standard in DNA methylation profiling. Its advantage is accuracy, as the degree of methylation at each cytosine can be quantified with great precision (Fraga and Esteller, 2002). More recently, bisulfite sequencing has been applied in a genome-wide manner, which requires advanced computational analysis for determining DNA methylation patterns and changes therein. Thus far, such analysis has been limited to the comparison of individual CpG sites between samples. However, because of the bisulfite conversion of cytosines into uracils, and eventually thymines, sequence complexity is much reduced, with no account for incomplete conversion and/or sequencing errors. Bisulfite sequencing also often suffers from low coverage. Here, we present BS-Seq Epimutation Analysis Toolkit (BEAT), a novel tool for analyzing bisulfite-converted DNA sequences. To overcome the aforementioned limitations in the estimation of methylation rates, BEAT aggregates data from consecutive cytosines into regions by using a Bayesian binomial-beta mixture model. The model is derived, described in detail and evaluated in our Supplementary Material. For each region, it calculates a posterior methylation probability distribution that can be used for the comparison of DNA methylation between samples. Anticipating technological progress in the DNA methylation field, BEAT includes an error model adapted to single-cell BS-Seq data.

1 INTRODUCTION

Bisulfite sequencing (BS-Seq) is a sequence-based method to accurately detect DNA methylation at specific loci, which involves treating DNA with sodium bisulfite (Frommer et al., 1992). The method is based on bisulfite conversion of unmethylated cytosines into uracil and has become a standard in DNA methylation profiling. Its advantage is accuracy, as the degree of methylation at each cytosine can be quantified with great precision (Fraga and Esteller, 2002). More recently, bisulfite sequencing has been applied in a genome-wide manner, which requires advanced computational analysis for determining DNA methylation patterns and changes therein. Thus far, such analysis has been limited to the comparison of individual CpG sites between samples. However, because of the bisulfite conversion of cytosines into uracils, and eventually thymines, sequence complexity is much reduced, with no account for incomplete conversion and/or sequencing errors. Bisulfite sequencing also often suffers from low coverage. Here, we present BS-Seq Epimutation Analysis Toolkit (BEAT), a novel tool for analyzing bisulfite-converted DNA sequences. To overcome the aforementioned limitations in the estimation of methylation rates, BEAT aggregates data from consecutive cytosines into regions by using a Bayesian binomial-beta mixture model. The model is derived, described in detail and evaluated in our Supplementary Material. For each region, it calculates a posterior methylation probability distribution that can be used for the comparison of DNA methylation between samples. Anticipating technological progress in the DNA methylation field, BEAT includes an error model adapted to single-cell BS-Seq data.
the output of BEAT. The R-objects required for this analysis are included in the BEAT package.

The most important step is the setting of the user-defined parameters.

```r
# Load sample data
data(BEAT)
# Initialize working path
localpath <- system.file('extdata', package = 'BEAT')
# Set sample names and prefix of data files
sampNames <- c('reference', 'sample')
# Set reference vs. non-ref status per sample
is.reference <- c(TRUE, FALSE)
# Set BS-conversion rate per sample
pplus <- c(0.2, 0.5)
convrates <- 1-pplus
# Create parameter object
params <- makeParams(localpath, sampNames, convrates, is.reference, pminus = 0.2,
regionSize = 10000, minCounts = 5)
# Pool CG positions into genomic regions
positions_to_regions(params)
# Model methylation levels and status
generate_results(params)
# Call epimutations
epiCalls <- epimutation_calls(params)
```

3 CONCLUSION

The BEAT package delivers methods for the estimation of methylation levels, methylation status and for calling epimutation events in a two-sample comparison. To our knowledge, it is the first tool providing a rigid statistical model for handling BS-Seq samples. It has an in-built correction for conversion errors and is therefore tailored to the analysis of BS-Seq samples with possibly different BS-conversion rates.

**Funding:** NIH grants (AG17242, AG034421, AG032117 and SENS Foundation to J.V. and S.G.). Jeff Schell professorship for Computational Biology grant (SFB680) from the Deutsche Forschungsgemeinschaft (to A.T.). BMBF e:Bio grant (to K.A.).

**Conflict of Interest:** none declared

### REFERENCES


---

**Fig. 1.** (a) Methylation estimates and epimutation calls on a DNA segment. For all regions with sufficient read coverage, the black curves show the methylation estimates for a single cell sample (top), a reference sample (bottom) and their difference (middle). Regions with methylating epimutations are marked in red, while regions with demethylating epimutations are marked in blue. Samples used for our analysis in this article were obtained from neuronal cells of young mice (data unpublished). (b) Scatterplot of methylation estimates of a multi-cell reference sample (x-axis) versus those of a sample (y-axis) for all common regions with sufficient coverage. Each dot represents a single region that is covered by both samples. Red dots indicate methylating epimutations in the sample, while blue dots indicate demethylating epimutations in the sample. Four dots representing exemplary regions with epimutations at the corresponding boundary value ranges for demethylating and methylating epimutations have been annotated with their values of methylated (k) and total (n) counts. Note that there exists no boundary line separating the red and the blue region because our Bayesian model assigns different methylation estimates to tuples (k1/n1), (k2, n2) with equal empirical methylation level k1/n1 = k2/n2