README document for manuscript “Batch effect correction for heterogeneous DNA methylation data via Bayesian hierarchical beta mixtures”

1. **Data**
   1. Abstract

The simulated data in Simulation Case I, II and III are artificially generated by assigning parameters for the BUSbeta model and then sampling from the model. In the three cases, the simulated data all consist of 660 samples from 3 batches. The 660 samples have DNA methylation beta values for 1000 CpG sites and are assigned to 5 subgroups. The only difference in parameter specifications among these three cases is the true subgroup proportions which characterizes subgroup structures in each batch.

The Kabuki syndrome data consist of two data sources available in GEO with accession numbers GSE218186 and GSE97362, serving as batch one and batch two, respectively. Batch one comprises 7 Kabuki syndrome patients and 55 healthy controls, and batch two includes 19 Kabuki syndrome patients and 125 healthy controls. The healthy individuals and the patients are labeled as 2 subgroups.

The Down syndrome data consist of two data sources available in GEO with accession numbers GSE107211 and GSE174555, serving as batch one and batch two, respectively. Batch one contains 10 Down syndrome patients and 5 healthy controls, and batch two has 17 Down syndrome patients and 17 healthy controls. The healthy individuals and the patients are labeled as 2 subgroups.

* 1. Availability

The two data sources of the Kabuki syndrome data are publicly available for download via the website. GSE218186 is provided by Hildonen M et al. (2023) at

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE218186> and

<https://ftp.ncbi.nlm.nih.gov/geo/series/GSE218nnn/GSE218186/matrix>.

No registration is required. The txt file extracted from “GSE218186\_RAW.tar” is the DNA methylation data matrix at each CpG site for each sample. The txt file extracted from “GSE218186\_series\_matrix.txt.gz” contains the detailed information of all samples including the subgroup labels.

GSE97362 is provided by Butcher DT et al. (2017) at

<https://ftp.ncbi.nlm.nih.gov/geo/series/GSE97nnn/GSE97362/matrix>.

No registration is required. The txt file extracted from “GSE97362\_series\_matrix.txt.gz” contains both the DNA methylation data matrix at each CpG site for each sample and the detailed information of all samples including the subgroup labels.

The two data sources of the Down syndrome data are publicly available for download via the website. GSE107211 is provided by Henneman P et al. (2018) at

<https://ftp.ncbi.nlm.nih.gov/geo/series/GSE107nnn/GSE107211/matrix>.

No registration is required. The txt file extracted from “GSE107211\_series\_matrix.txt.gz” contains both the DNA methylation data matrix at each CpG site for each sample and the detailed information of all samples including the subgroup labels.

GSE174555 is provided by Naumova OY et al. (2021) at

<https://ftp.ncbi.nlm.nih.gov/geo/series/GSE174nnn/GSE174555/matrix>.

No registration is required. The txt file extracted from “GSE174555\_series\_matrix.txt.gz” contains both the DNA methylation data matrix at each CpG site for each sample and the detailed information of all samples including the subgroup labels.

* 1. Description

**The Kabuki syndrome data**

Citation:

Hildonen M, Ferilli M, Hjortshøj TD, Dunø M et al. DNA methylation signature classification of rare disorders using publicly available methylation data. Clin Genet 2023 Jun;103(6):688-692. PMID: 36705342.

Butcher DT, Cytrynbaum C, Turinsky AL, Siu MT et al. CHARGE and Kabuki Syndromes: Gene-Specific DNA Methylation Signatures Identify Epigenetic Mechanisms Linking These Clinically Overlapping Conditions. Am J Hum Genet 2017 May 4;100(5):773-788. PMID: 28475860.

In the data preprocessing procedure, we collect the subgroup information for each sample and select 148 informative CpG sites, which capture distinct DNA methylation patterns across subgroups, according to Aref-Eshghi et al. (2020).

**The Down syndrome data**

Citation:

Henneman P, Bouman A, Mul A, Knegt L et al. Widespread domain-like perturbations of DNA methylation in whole blood of Down syndrome neonates. PLoS One 2018;13(3):e0194938. PMID: 29601581.

Naumova OY, Lipschutz R, Rychkov SY, Zhukova OV et al. DNA Methylation Alterations in Blood Cells of Toddlers with Down Syndrome. Genes (Basel) 2021 Jul 23;12(8). PMID: 34440289.

In the data preprocessing procedure, we collect the subgroup information for each sample and select 124 informative CpG sites, which capture distinct DNA methylation patterns across subgroups, according to Aref-Eshghi et al. (2020).

1. **Code**
   1. Abstract

All of the data preprocessing and analysis in this paper were completed using R. The code is provided to conduct generating simulated data or preprocessing on the raw data, implement BUSbeta via Markov chain Monte Carlo method, and compare against the competing methods BUS, the batch effect correction methods ComBat, BEclear, and the clustering methods Kmeans, GMM and MetaSparseKmeans.

* 1. Description

All of the R scripts are available as the supplementary code.

R license information: GPL (>= 2).

For R and R packages, we use R version 4.4.3 (2025-02-28, “Trophy Case”). Please note that for Windows system users, the version of Rtools needs to be compatible with the version of R. The used R packages are: (System\_preparation.R)

* aricode, version="1.0.3"
* BiocManager, version="1.30.25"
* circlize, version="0.4.16"
* ComplexHeatmap, version="2.25.1"
* ggbreak, version="0.1.4"
* ggplot2, version="3.5.1"
* invgamma, version="1.1"
* matrixStats, version="1.5.0"
* remotes, version="2.5.0"
* tidyr, version="1.3.1"
* truncnorm, version="1.0.9"
* umap, version="0.2.10.0"
* vroom, version="1.6.5"

For BUS,

* R >= 3.5.0
* BUScorrect, version 1.24.0

For ComBat,

* R >= 3.2
* Mgcv
* Genefilter
* BiocParallel
* sva, version 3.54.0

For BEclear,

* R >= 3.2
* BiocParallel (>= 1.14.2)
* BEclear, version 2.22.0

For GMM,

* R >= 3.0
* mclust, version 6.1.1

For MetaSparseKmeans,

* R >= 3.0
* MetaSparseKmeans, version 0.0.3

A MacBook Air was used for the real application analyses in this paper. The details of the computer are:

* Operating system: MacOS Sequoia 15.2
* CPU: Apple M3
* RAM: 24GB

The computing platform was used for the real application analyses in this paper. The details of the computing platform are:

* Operating system: CentOS 7.8.2003
* CPU: Intel Gold 5218 (16 cores, 32 threads) 2.3GHz
* RAM: 192GB
  1. Packages installation

**BUS**

BUS is a batch effects correction and subgroups clustering method based on a Bayesian hierarchical framework to model the genomics data heterogeneity within a batch via a Gaussian mixture (Luo and Wei, 2019). The following code of installation is provided in the System\_preparation.R file.

if (!require("BUScorrect", quietly = TRUE))

BiocManager::install("BUScorrect")

**ComBat**

ComBat is a batch effects correction method based on an empirical Bayes framework to estimate parameters and make the correction procedure robust to noises (Johnson et al., 2007). The following code of installation is provided in the System\_preparation.R file.

if (!require("sva", quietly = TRUE))

BiocManager::install("sva")

**BEclear**

BEclear is a batch effects correction method based on latent factor models and specifically designed for DNA methylation data (Akulenko et al., 2016). The following code of installation is provided in the System\_preparation.R file.

if (!require("BEclear", quietly = TRUE))

BiocManager::install("BEclear")

**GMM**

The Gaussian Mixture Model (GMM) is a model-based clustering approach that assumes data are generated from a combination of Gaussian distributions with unknown parameters. The following code of installation is provided in the System\_preparation.R file.

if (!require("mclust", quietly = TRUE))

install.packages("mclust", version="6.1.1")

**MetaSparseKmeans**

MetaSparseKmeans is a clustering approach that extend the sparse K-means clustering approach to a meta-analytic setting for discovering sample subgroups in multiple batches (Huo et al., 2016). The following code of installation is provided in the System\_preparation.R file.

if (!require("MetaSparseKmeans", quietly = TRUE))

remotes::install\_github("Caleb-Huo/MetaSparseKmeans")

* 1. Instructions for Use

All simulated data generating, real data preprocessing and analysis can be reproduced. All figures including Figures 1 and 2 in the manuscript and Figures S1, S2 and S3, as well as the tables including Table 1 in the manuscript and Tables S6 and S7 in the supplementary materials can be reproduced.

Detailed workflow information is contained in the "README.docx" in "Real\_application" directory. One should firstly check and install the R packages by conducting the code files "System\_preparation.R."

The general steps in the simulation case I are:

1. Generate the simulated data.

2. Conduct BIC analysis to the simulated data.

3. Apply BUSbeta to the simulated data. There are 1,000 iterations in the MCMC with 500 burn-in steps. Total execution time is about 18 minutes on a MacBook Air with Apple M3 CPU and 24GB of RAM.

4. Implement the BUS, ComBat and BEclear.

5. Draw Figures 1(a)-(h) in the manuscript.

6. Implement Kmeans, GMM and MetaSparseKmeans and generate Table 1 in the manuscript.

The general steps in the simulation case II are:

1. Generate the simulated data.

2. Conduct BIC analysis to the simulated data.

3. Apply BUSbeta to the simulated data. There are 1,000 iterations in the MCMC with 500 burn-in steps. Total execution time is about 18 minutes on a MacBook Air with Apple M3 CPU and 24GB of RAM.

4. Implement the BUS, ComBat and BEclear.

5. Draw Figures S1(a)-(h) in the supplementary materials.

6. Implement Kmeans, GMM and MetaSparseKmeans and generate Table S6 in the supplementary materials.

The general steps in the simulation case III are:

1. Generate the simulated data.

2. Conduct BIC analysis to the simulated data.

3. Apply BUSbeta to the simulated data. There are 1,000 iterations in the MCMC with 500 burn-in steps. Total execution time is about 18 minutes on a MacBook Air with Apple M3 CPU and 24GB of RAM.

4. Implement the BUS, ComBat and BEclear.

5. Draw Figures S2(a)-(h) in the supplementary materials.

6. Implement Kmeans, GMM and MetaSparseKmeans and generate Table S7 in the supplementary materials.

The general steps in the first real application (Kabuki syndrome data) are:

1. Read raw data from the txt files.

2. Conduct data preprocessing.

3. Conduct BIC analysis to the preprocessed data.

4. Apply BUSbeta to the preprocessed data. There are 3,000 iterations in the MCMC with 1,500 burn-in steps. Total execution time is about 1.5 minutes on a MacBook Air with Apple M3 CPU and 24GB of RAM.

5. Implement the BUS, ComBat and BEclear.

6. Draw Figures 2(a)-(f) in the manuscript.

7. Implement Kmeans, GMM and MetaSparseKmeans.

The general steps in the second real application (Down syndrome data) are:

1. Read raw data from the txt files.

2. Conduct data preprocessing.

3. Conduct BIC analysis to the preprocessed data.

4. Apply BUSbeta to the preprocessed data. There are 3,000 iterations in the MCMC with 1,500 burn-in steps. Total execution time is about 0.5 minutes on a MacBook Air with Apple M3 CPU and 24GB of RAM.

5. Implement the BUS, ComBat and BEclear.

6. Draw Figures S3(a)-(f) in the supplementary materials.

7. Implement Kmeans, GMM and MetaSparseKmeans.