Imputation benchmark of β-value and M-value from DNA methylation data under different missing data mechanisms.

CURRENT STATUS: UNDER REVISION

BMC Bioinformatics BMC Series

Pietro Di Lena
University of Bologna

pietro.dilena@unibo.it*Corresponding Author*
ORCiD: https://orcid.org/0000-0002-1838-8918

Claudia Sala
University of Bologna

Andrea Prodi
Smart Cities Living Lab, ISOF CNR

Christine Nardini
Istituto per le Applicazioni del Calcolo Mauro Picone, CNR

DOI:
10.21203/rs.2.20718/v1

SUBJECT AREAS
Bioinformatics

KEYWORDS
Imputation, DNA methylation, M-value, β-value, Missing data mechanisms, MCAR, MAR, MNAR

Abstract

Background: High-throughput technologies enable the cost-effective collection and analysis of DNA methylation data throughout the human genome. This naturally entails missing values management that can complicate the analysis of the data. Several general and specific imputation methods are suitable for DNA methylation data. However, there are no detailed studies of their performances under different missing data mechanisms -(completely) at random or not- and different representations of DNA methylation levels (β and *M*-value).

Results: We make an extensive analysis of the imputation performances of seven imputation methods on simulated missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) methylation data. We further consider imputation performances on the β- and M-value popular representations of methylation levels. Overall, β -values enable better imputation performances than M-values. Imputation accuracy is lower for mid-range β -values, while it is generally more accurate for values at the extremes of the β -value range. The MAR values distribution is on the average more dense in the mid-range in comparison to the expected β -value distribution. As a consequence, MAR values are on average harder to impute.

Conclusions: The results of the analysis provide guidelines for the most suitable imputation approaches for DNA methylation data under different representations of DNA methylation levels and different missing data mechanisms.

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

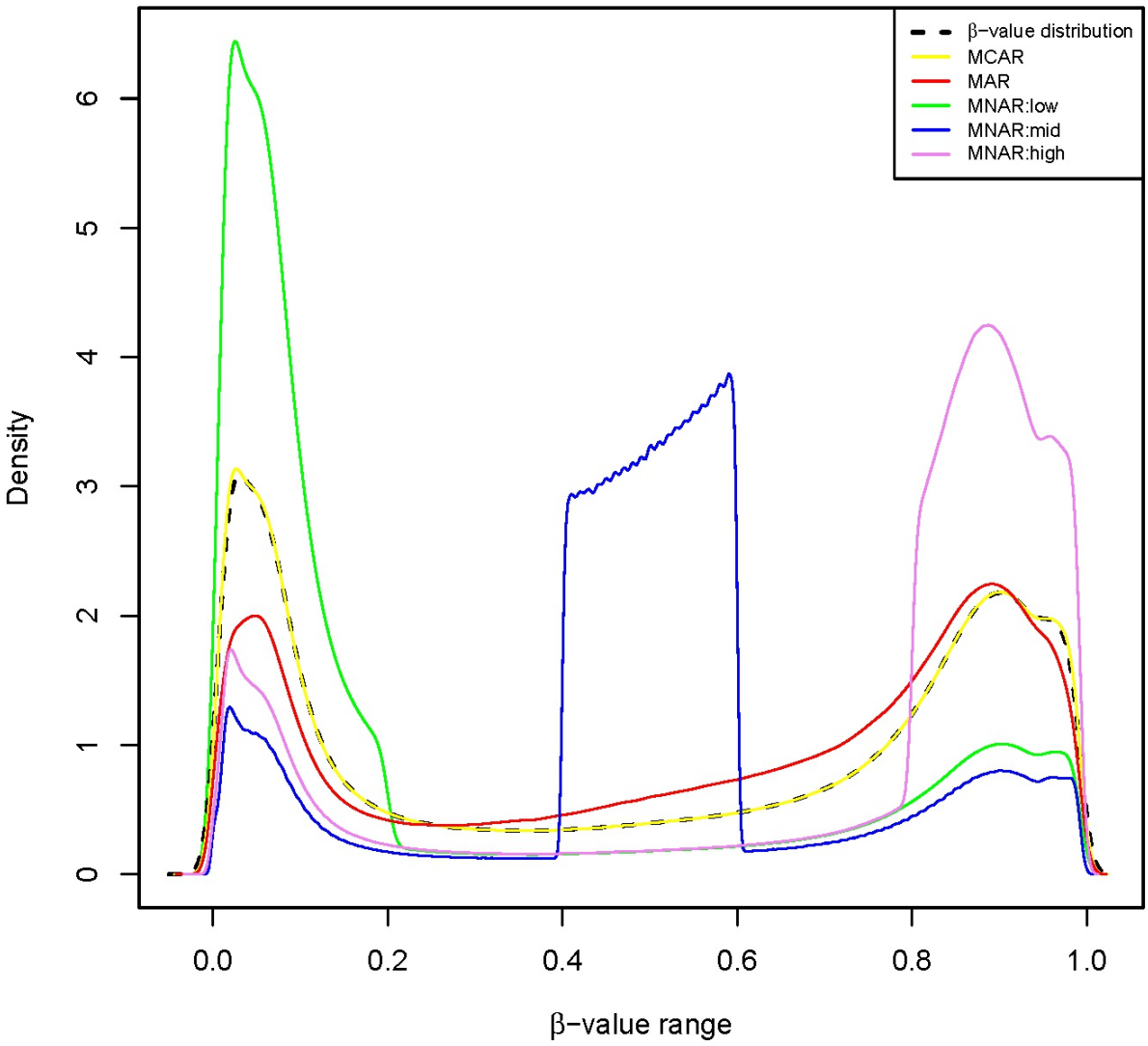


Figure 1

β-value distributions of different missingness mechanisms. Comparison of the β-value distribution against the distribution of simulated MCAR, MAR and MNAR missing values.

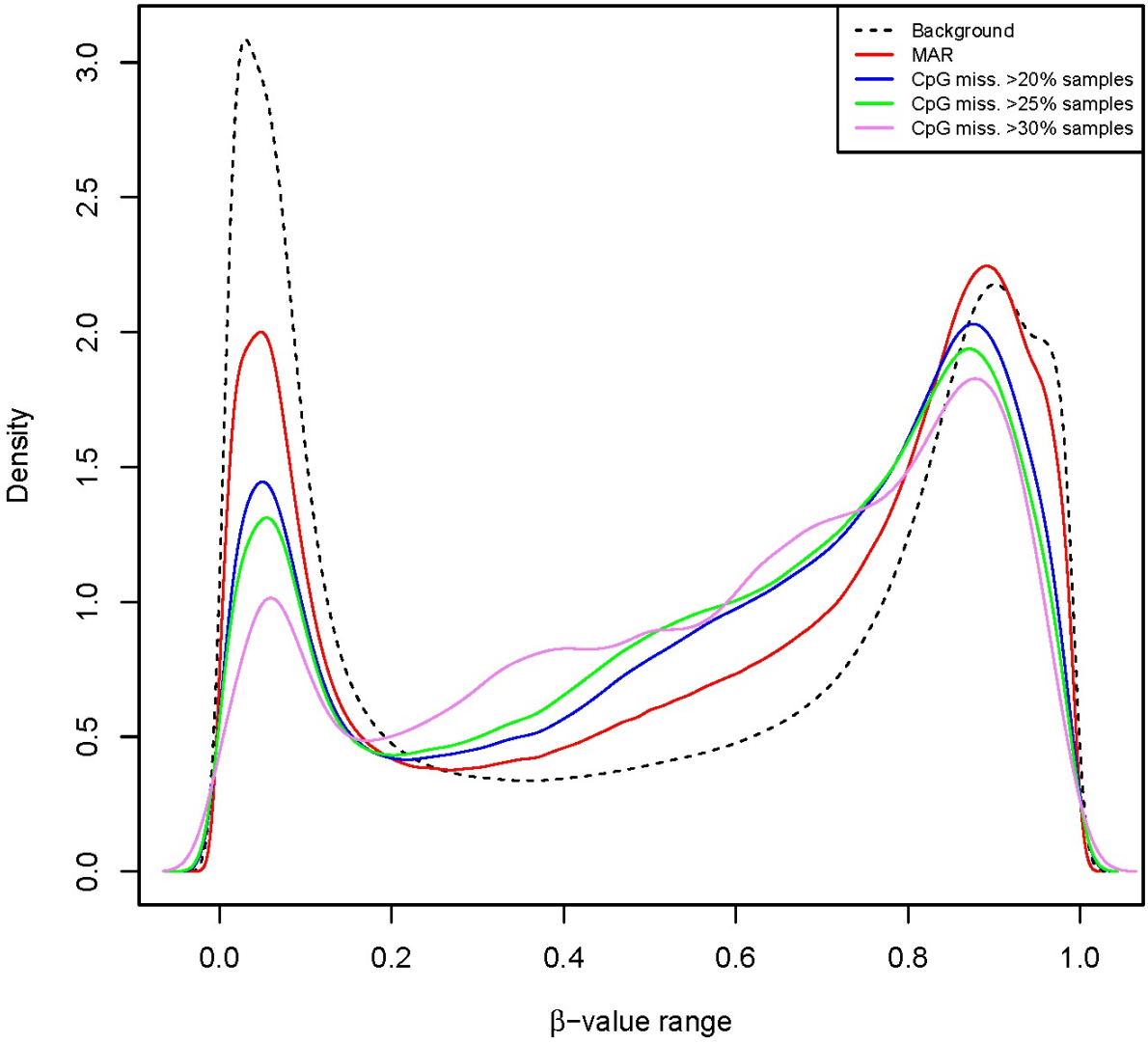
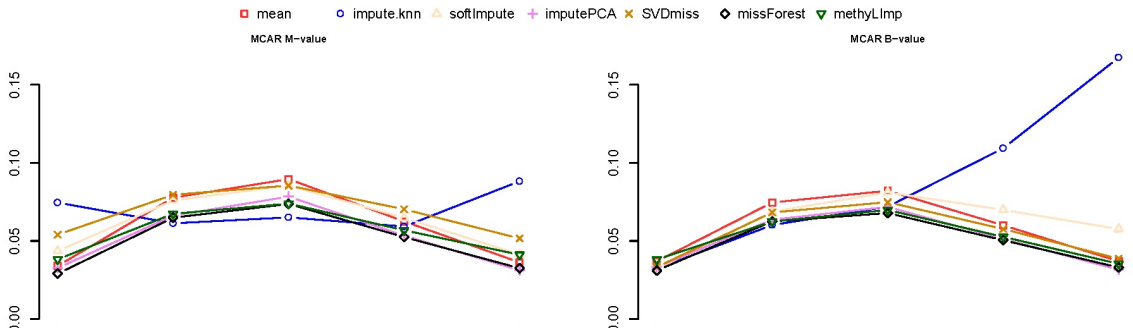


Figure 2

β-value distributions of CpGs with frequently missing values. Comparison of the β-value distribution against the β-value distribution of CpGs with missing values on > 20%;> 25%;> 30% samples. MAR simulated distribution included.



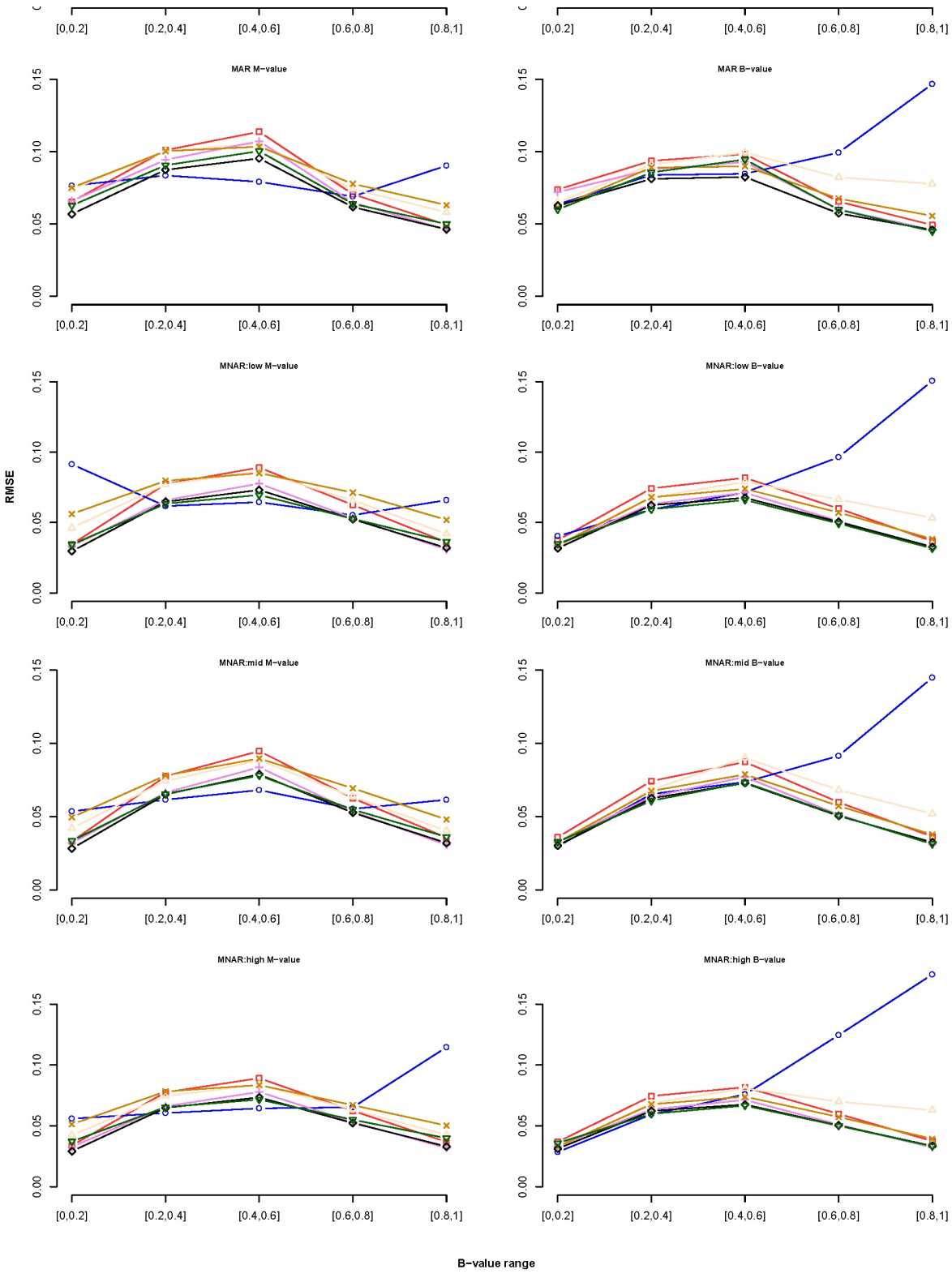


Figure 3

Healthy datasets. Average RMSE with respect to β-value range. Average RMSE for M-value and β-value imputation with respect to different - value ranges and with respect to the MCAR, MAR, MNAR (low, mid, high) missing data mechanisms.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

AdditionalFile2.pdf
AdditionalFile1.pdf