Missing value imputation in methylation data

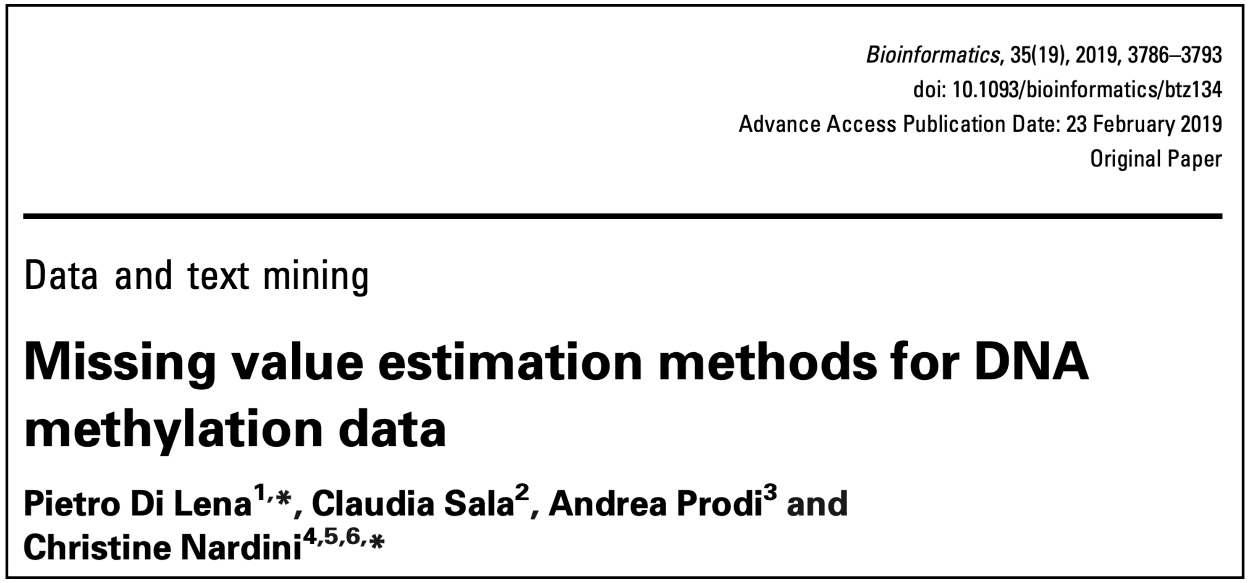
Oslo Bioinformatics Workshop Week 2025
December 10th

Anna Plaksienko, postdoc
at Oslo Centre of
Biostatistics and Epidemiology

Our plan

1. What is methylation data?
2. What kind of missing values does it have?
3. How many missing values are there usually?
4. Why impute missing values?
5. Overview of the methods for missing value imputation
6. Overview of the *methyLImp2* method
7. Try of the `methyLImp2` package

Content of this workshop is based on the following three papers

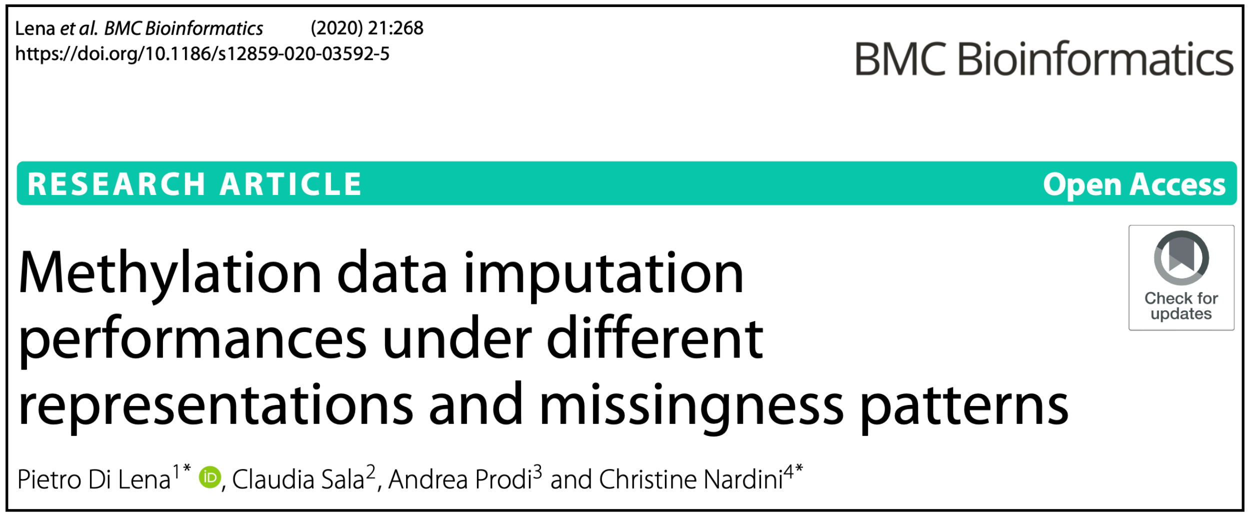


Bioinformatics, 35(19), 2019, 3786–3793
doi: 10.1093/bioinformatics/btz134
Advance Access Publication Date: 23 February 2019
Original Paper

Data and text mining

Missing value estimation methods for DNA methylation data

Pietro Di Lena[1,]*, Claudia Sala[2], Andrea Prodi[3] and Christine Nardini[4,5,6,]*



Lena *et al. BMC Bioinformatics* (2020) 21:268
https://doi.org/10.1186/s12859-020-03592-5

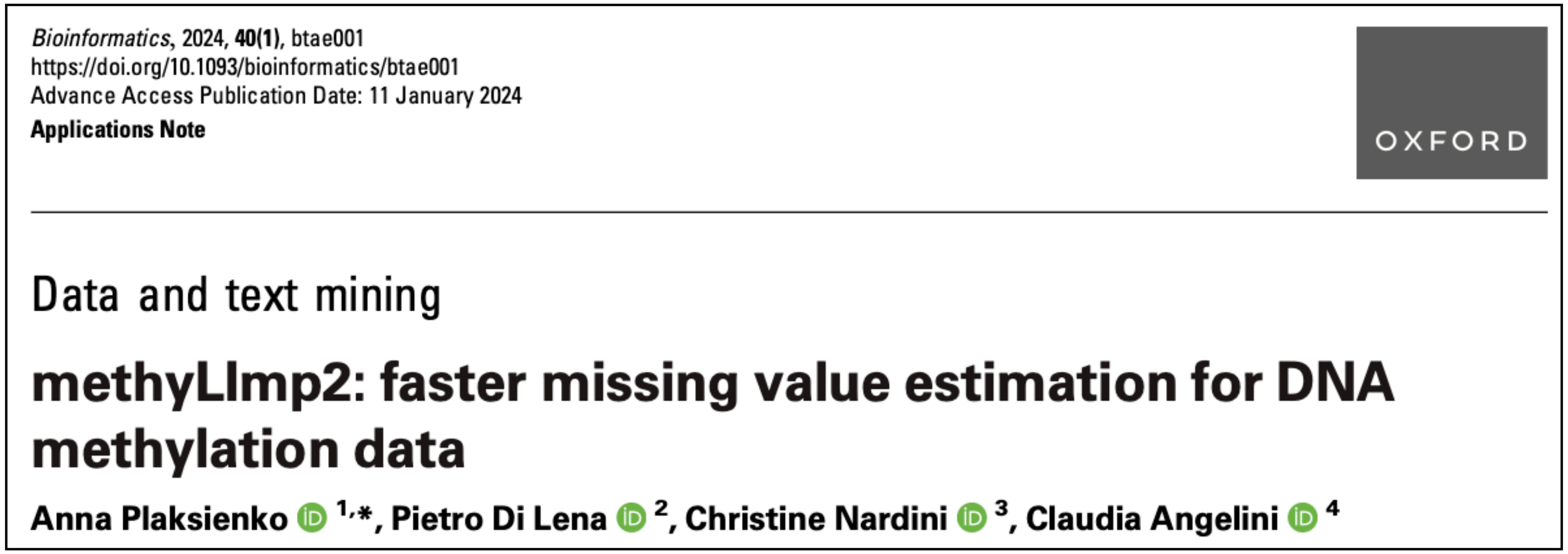
BMC Bioinformatics

RESEARCH ARTICLE **Open Access**

Methylation data imputation performances under different representations and missingness patterns

Check for updates

Pietro Di Lena[1]*, Claudia Sala[2], Andrea Prodi[3] and Christine Nardini[4]*



Bioinformatics, 2024, **40(1)**, btae001
https://doi.org/10.1093/bioinformatics/btae001
Advance Access Publication Date: 11 January 2024
Applications Note

OXFORD

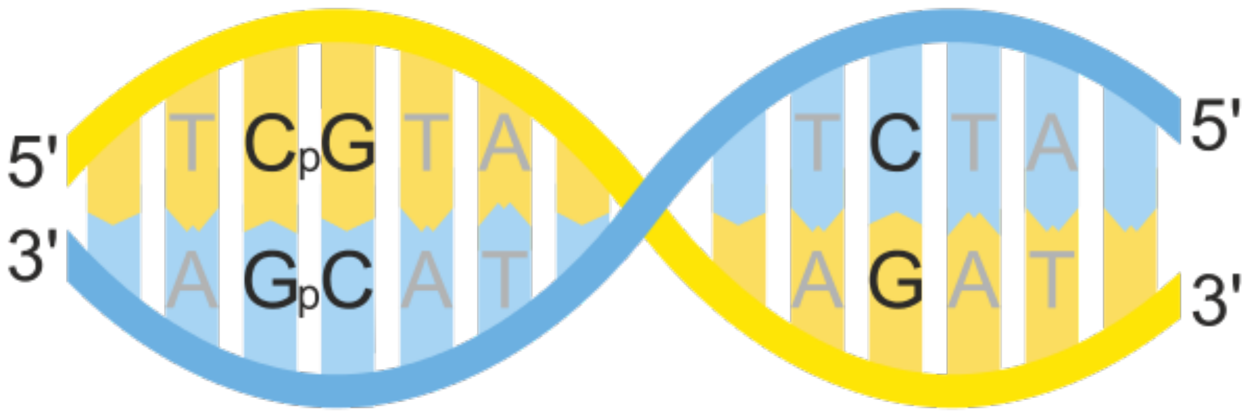
Data and text mining

methyLImp2: faster missing value estimation for DNA methylation data

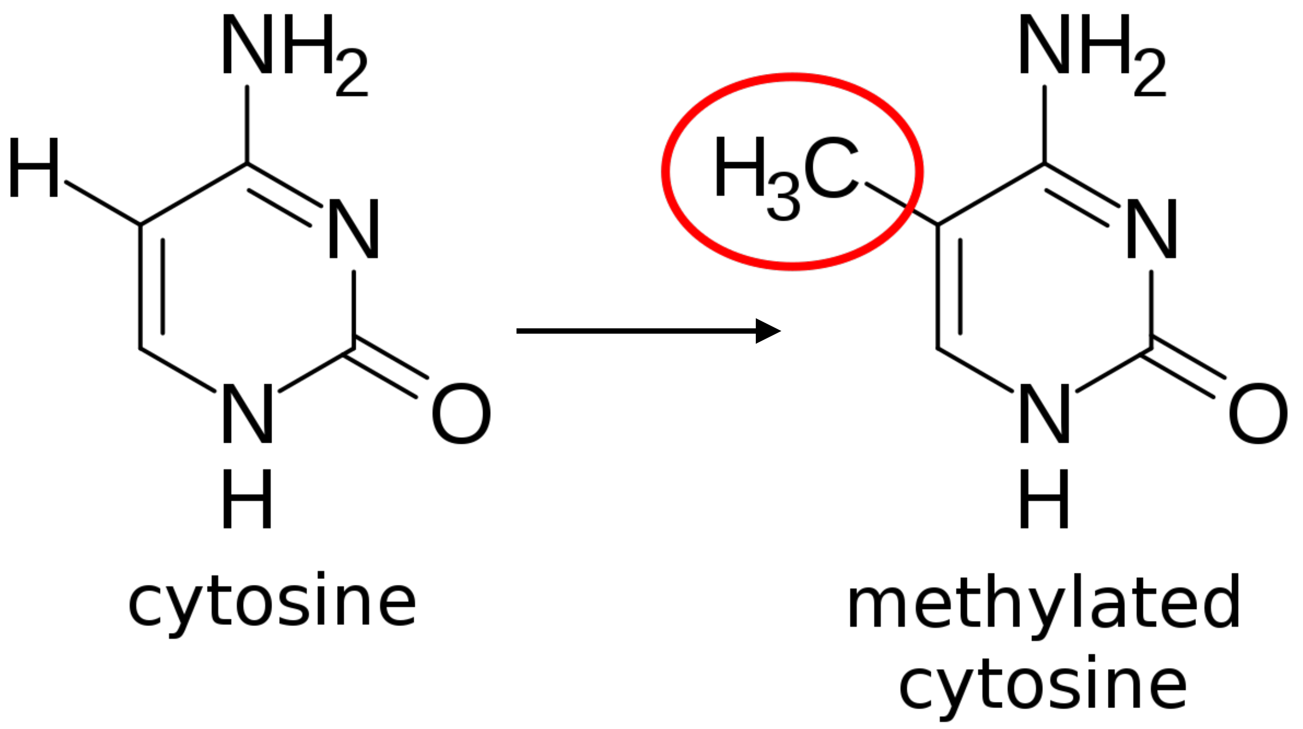
Anna Plaksienko [1,]*, Pietro Di Lena [2], Christine Nardini [3], Claudia Angelini [4]

Methylation data

DNA methylation –
epigenetic modification of DNA,
addition of methyl group (CH_3)
to 5’-carbon of cytosine in
a CpG (cytosine-phosphate-guanine) dinucleotide.



Wiki Helixitta



Wiki Mariuswalter

Methylation data

- Illumina Infinium assay with 27, 450 or 850 K probes – **high dimension**!
- Measures a ***proportion* of intensities** of signal of the methylated and both methylated and unmethylated alleles at each CpG site across all cells in the sample tissue. These are β**-values:**

$$\beta = \frac{\text{intensity of methylated allele}}{\text{intensity of methylated allele} + \text{intensity of unmethylated allele} + \text{const*}}$$

**to not divide by 0*

$\beta = 0$ means that all the copies of the CpG site in the sample are completely unmethylated,
$\beta = 1$ means that all the copies of the CpG site in the sample are methylated.

$$0 \leq \beta \leq 1$$

Methylation data

- β**-values:**

$$\beta = \frac{\text{intensity of methylated allele}}{\text{intensity of methylated allele} + \text{intensity of unmethylated allele} + \text{const}}$$

$$0 \leq \beta \leq 1$$

- Another used measure is M**-values**:

$$M = log_2\left(\frac{\text{intensity of methylated allele + const}}{\text{intensity of unmethylated allele} + \text{const}}\right)$$

$$-\infty \leq M \leq \infty$$

- $\beta = \dfrac{2^M}{2^M + 1}$ and $M = log_2\left(\dfrac{\beta}{1 - \beta}\right)$

Methylation data

$$\beta = \frac{\text{intensity of methylated allele}}{\text{intensity of methylated allele} + \text{intensity of unmethylated allele} + \text{const}}$$

$$M = log_2\left(\frac{\text{intensity of methylated allele + const}}{\text{intensity of unmethylated allele} + \text{const}}\right)$$

$$\beta = \frac{2^M}{2^M + 1} \quad \text{and} \quad M = log_2\left(\frac{\beta}{1-\beta}\right)$$

β-values

+ more intuitive

+ encouraged by array producers

+ produce more accurate results in imputation

- *heteroscedastic*, i.e. the standard deviations are compressed in the low and high ranges and larger in the middle ranges

M-values

+ more popular for differential methylation analysis

Types of missingness

1. **Missing completely at random (MCAR)** – probability of data missing does not depend on either the observed or the missing values. **Ignorable**;
   Example: student sneezed over your samples and some of them randomly got sneeze particles that led to bad measurements. The cause has nothing to do with the samples at all.

Types of missingness

1. **Missing completely at random (MCAR)** – probability of data missing does not depend on either the observed or the missing values. **Ignorable**;
   Example: student sneezed over your samples and some of them randomly got sneeze particles that led to bad measurements. The cause has nothing to do with the samples at all.

2. **Missing at random (MAR)** – probability data missing does not depend on the value that is missing but it may depend on the observed values. **Ignorable**;
   Example: men in general are less likely to fill out depression surveys, regardless of their level of depression. So missing data doesn't depend in value that is missing (level of depression) but depends on observed variable (gender).

Types of missingness

1. **Missing completely at random (MCAR)** – probability of data missing does not depend on either the observed or the missing values. **Ignorable**;
   Example: student sneezed over your samples and some of them randomly got sneeze particles that led to bad measurements. The cause has nothing to do with the samples at all.

2. **Missing at random (MAR)** – probability data missing does not depend on the value that is missing but it may depend on the observed values. **Ignorable**;
   Example: men in general are less likely to fill out depression surveys, regardless of their level of depression. So missing data doesn't depend in value that is missing (level of depression) but depends on observed variable (gender).

3. **Missing not at random (MNAR)** – probability data missing depends on the value that is missing. **Not ignorable** since the imputation process needs to model explicitly the missing data mechanism in order to avoid biased estimations.
   Example: Patients with severe form of disease come to visits less regularly. Their results would probably be different because of the severity of the disease and it's the some reason why the data is missing.

Types of missingness

1. **Missing completely at random (MCAR)** – probability of data missing does not depend on either the observed or the missing values. Ignorable;
2. **Missing at random (MAR)** – probability data missing does not depend on the value that is missing but it may depend on the observed values. Ignorable;
3. **Missing not at random (MNAR)** – probability data missing depends on the value that is missing. Not ignorable since the imputation process needs to model explicitly the missing data mechanism in order to avoid biased estimations.

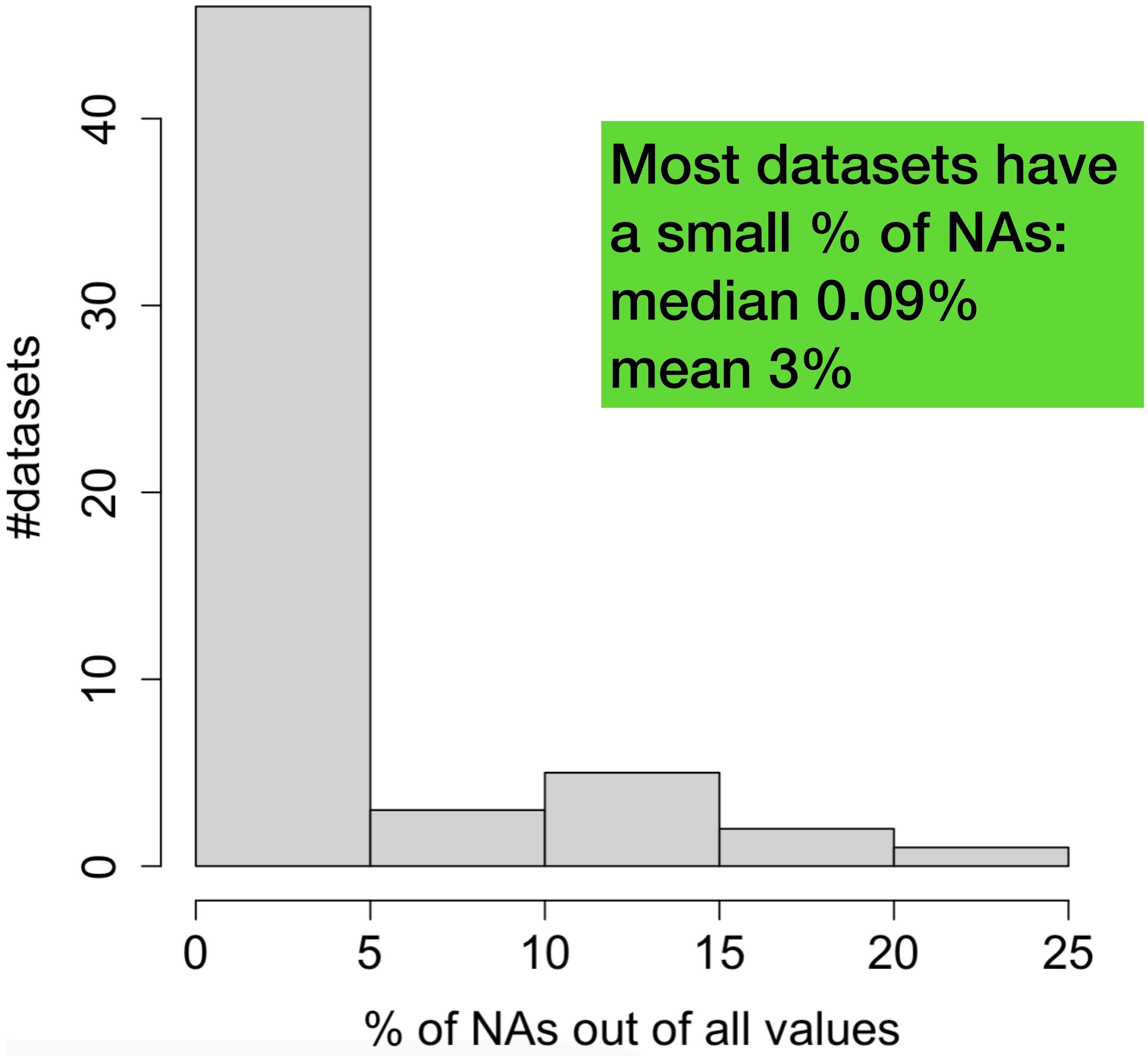
- **No statistical way** to determine type missingness;
- Explore if samples or sites with most NAs have something in common;
- Usually assume MCAR/MAR and model and impute accordingly.

How many missing values are there (usually)?

Di Lena et. al (2019) considered 58 Illumina 450k Human Beadchip platform datasets: 1881 samples total, patients and controls, different tissues

How many missing values are there (usually)?

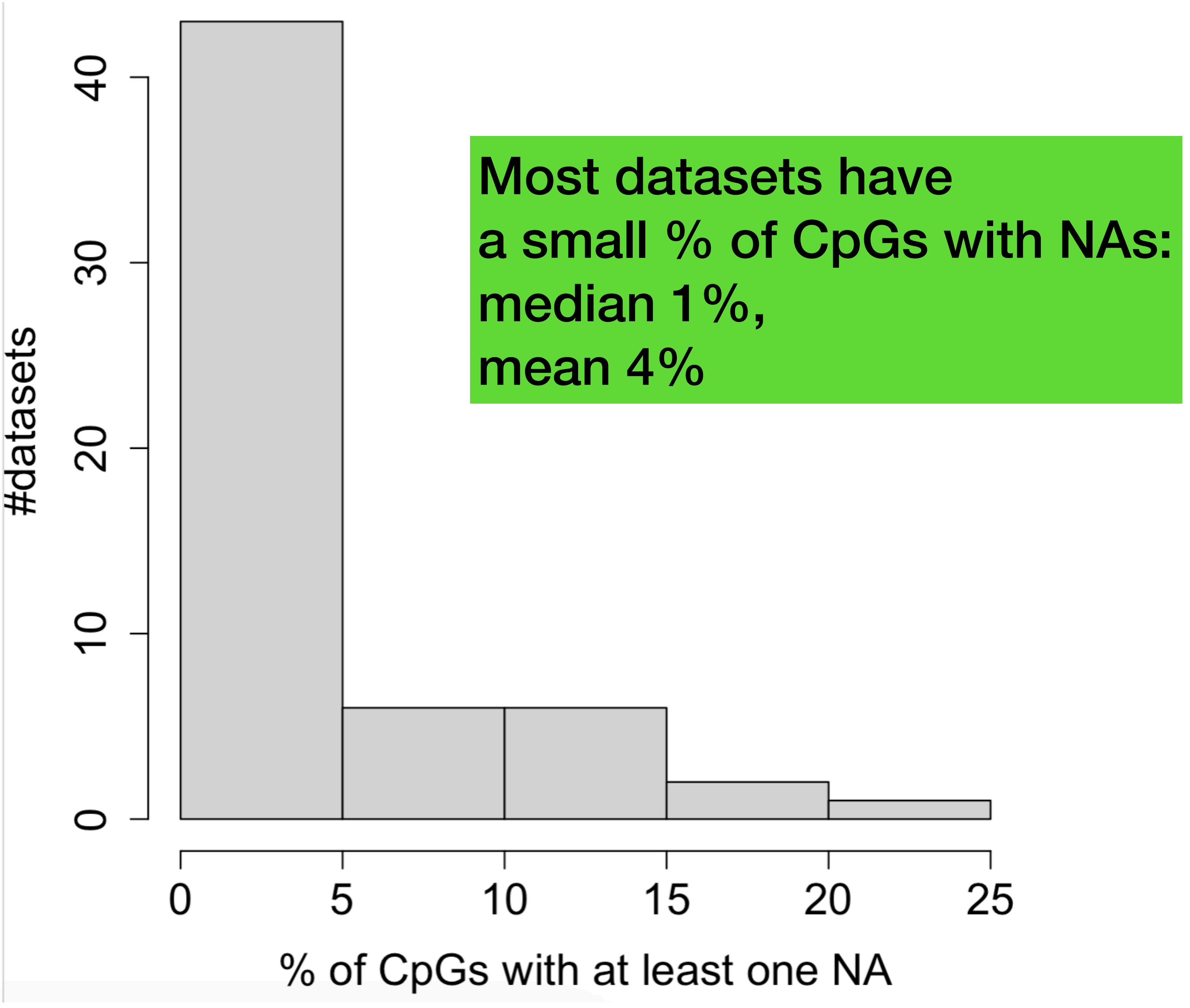
Di Lena et. al (2019) considered 58 Illumina 450k Human Beadchip platform datasets: 1881 samples total, patients and controls, different tissues



Most datasets have a small % of NAs:
median 0.09%
mean 3%

How many missing values are there (usually)?

Di Lena et. al (2019) considered 58 Illumina 450k Human Beadchip platform datasets: 1881 samples total, patients and controls, different tissues



Most datasets have
a small % of CpGs with NAs:
median 1%,
mean 4%

How many missing values are there (usually)?

Di Lena et. al (2019) considered 58 Illumina 450k Human Beadchip platform datasets: 1881 samples total, patients and controls, different tissues

Most datasets have a small % of NAs: median 0.09% mean 3%

Most datasets have a small % of CpGs with NAs: median 1%, mean 4%

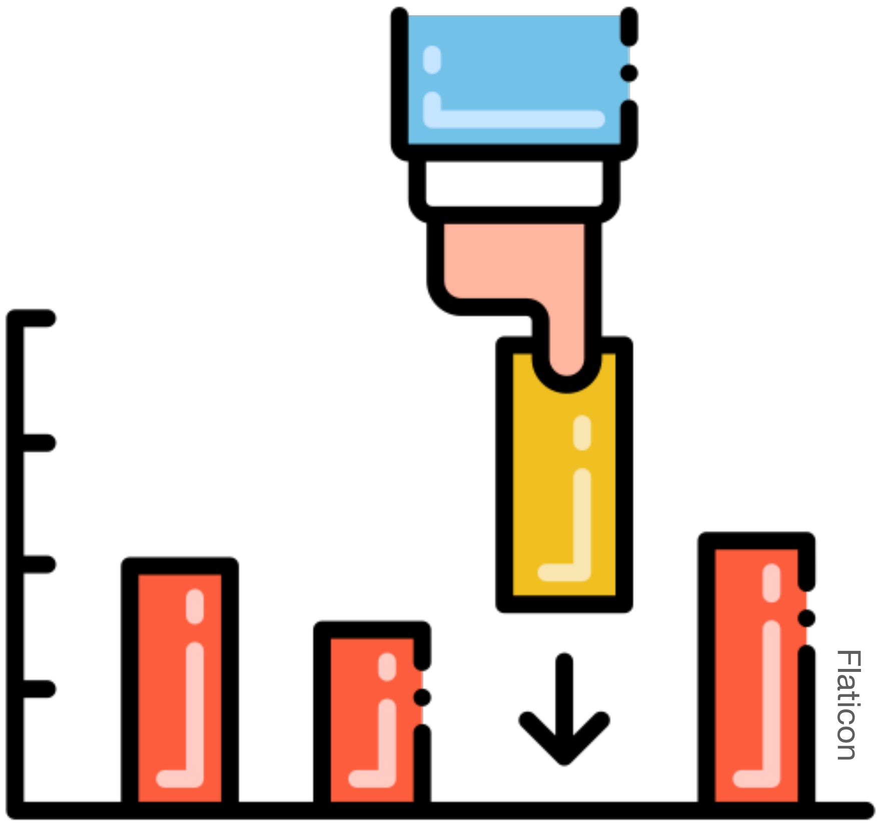
So usually, not that much in %.

There is not theoretical upper boundary on the max number of NAs after which analysis is meaningless.

Why impute and why do it well?

Most methods require full datasets. Dropping samples/probes with missing values = losing your precious data.

Small perturbations in your data may affect the results of your analysis, so it’s better to **impute accurately**.



When to do imputation?

Immediately after reading .idat files into R.

Before any kind of normalization, analysis, etc.

Exception: filter out samples with low quality before imputation.

When to do imputation?

Immediately after reading .idat files into R.

Before any kind of normalization, analysis, etc.

Exception: filter out samples with low quality before imputation.

Be careful!

Some packages have imputation step as default **in** their reading function! It's not necessarily bad but you have to be aware of it.

Be careful: imputation during import

package	import function	Imputation
`ChAMP`	`champ.load`	By default `autoimpute = TRUE`, uses `impute.knn(n = 3)`
`RnBeads`	`rnb.run.import`	By default `rnb.options(imputation.method = “none”)`
`minfi`	`read.metharray.exp` and others	none by default
`missMethyl`	none	

Imputation methods

Points to remember when choosing a method:

– If data is NOT missing at random: must explicitly model the missing mechanism -> generic methods are NOT suitable;

– Method has to be suited for **continuous limited range** variables (since β-values are between 0 and 1);

– Method has to be suited for large datasets (since you'll have at least **450K columns**).

There are 2 imputation methods categories:
single imputation (when you just impute NA once) or
multiple imputation (when you obtain multiple estimates and then combine them) –> probably not suitable for methylation data

Imputation methods

I describe here approaches for which there exist a comparison of performance and runtime specifically for methylation data.

Mean-value imputation approaches

- Mean: replace the missing value by the mean of all the known values for that variable.
- `impute.knn`: replace the missing value by the mean of its nearest neighbours.

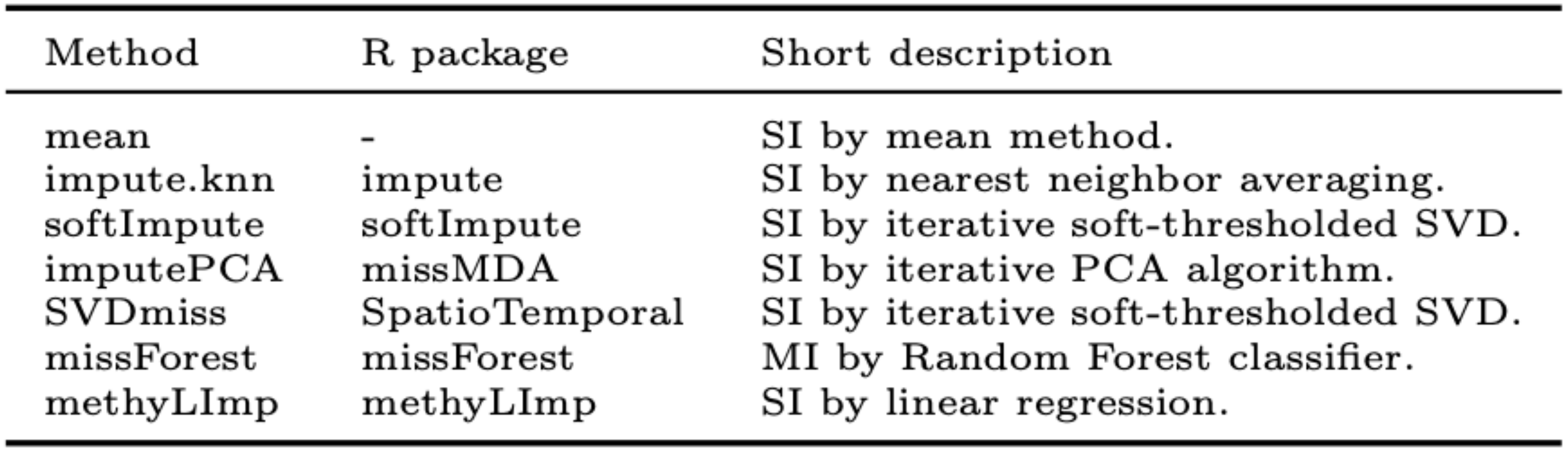
Iterative soft-thresholding approaches: replace missing values with some initial guess and then iteratively update, up to convergence, the missing elements with values generated by low-rank approximation of the input matrix.

- `SVDmiss`: uses soft-thresholding singular value decomposition (SVD) of the input matrix.
- `softImpute`: uses soft-thresholding singular value decomposition (SVD) of the input matrix.
- `imputePCA`: implements a low-rank approximation version of the iterative principal component analysis (PCA) algorithm.

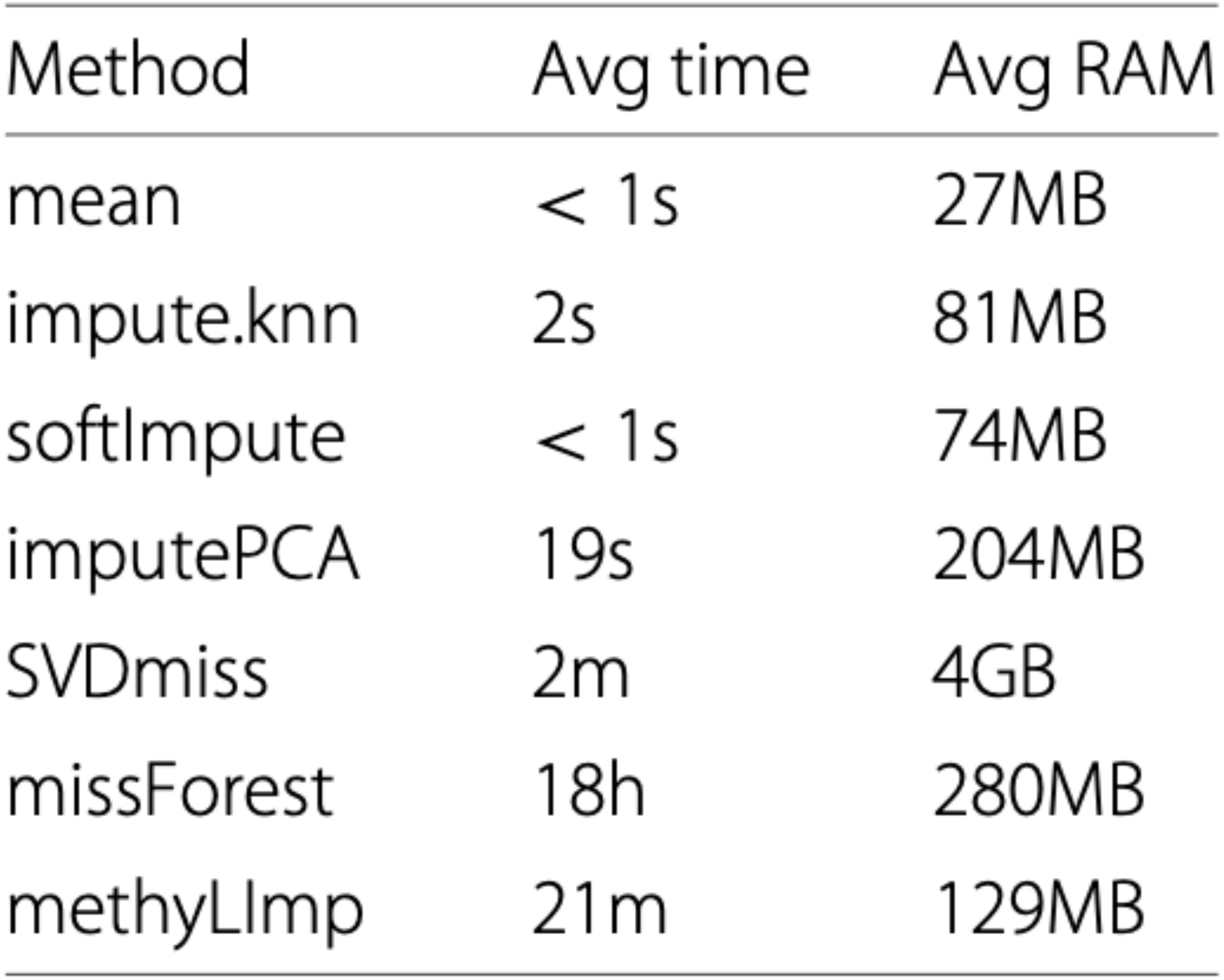
Regression-based imputation approaches: build a regression model from observed data

- `missForest`: builds a random forests regression trees
- `methyLImp`: builds a linear model with observed data.

Imputation methods: time and memory



Method	R package	Short description
mean	-	SI by mean method.
impute.knn	impute	SI by nearest neighbor averaging.
softImpute	softImpute	SI by iterative soft-thresholded SVD.
imputePCA	missMDA	SI by iterative PCA algorithm.
SVDmiss	SpatioTemporal	SI by iterative soft-thresholded SVD.
missForest	missForest	MI by Random Forest classifier.
methyLImp	methyLImp	SI by linear regression.

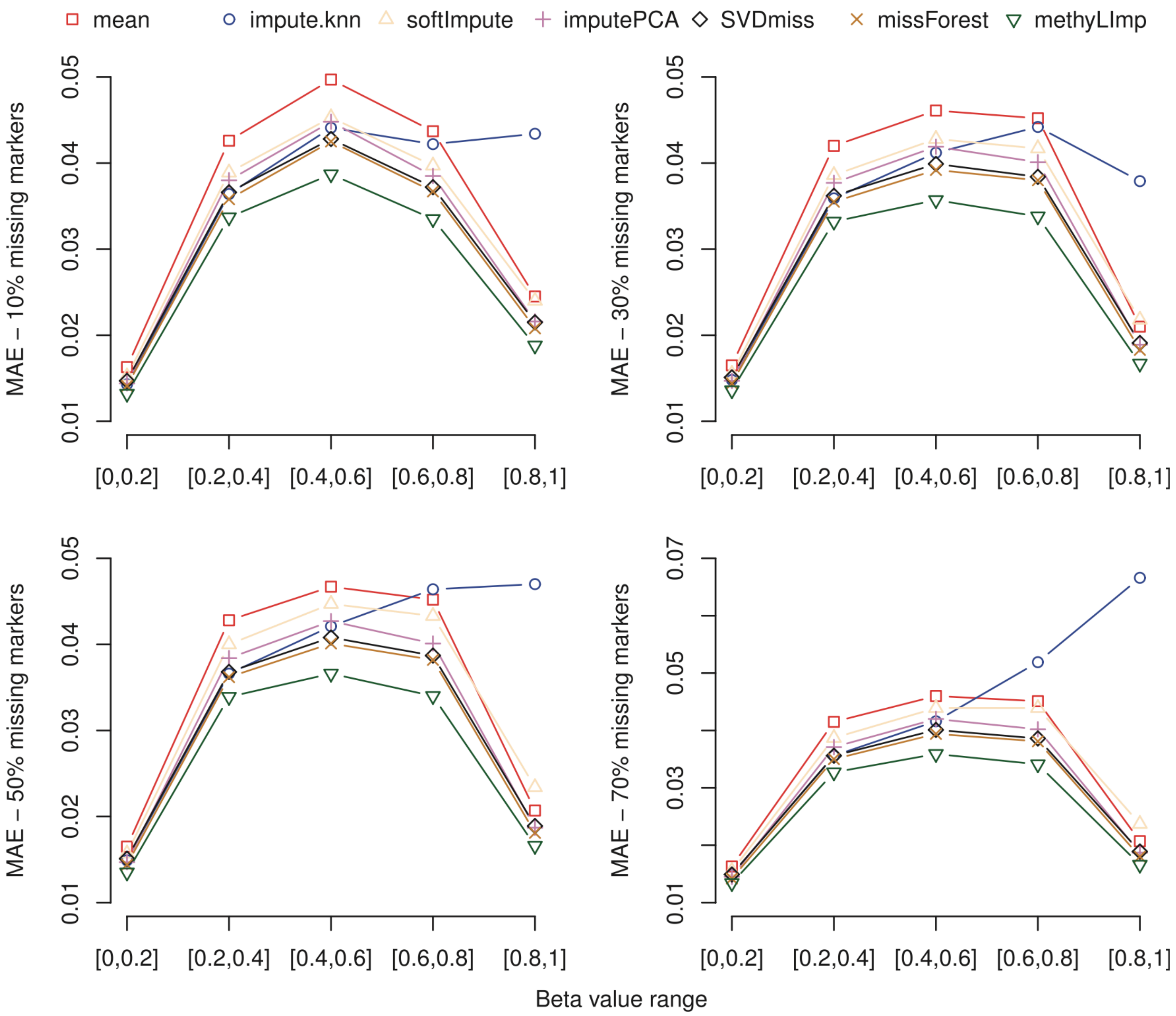


Method	Avg time	Avg RAM
mean	< 1s	27MB
impute.knn	2s	81MB
softImpute	< 1s	74MB
imputePCA	19s	204MB
SVDmiss	2m	4GB
missForest	18h	280MB
methyLImp	21m	129MB

Di Lena et. al (2019)

Imputation methods: performance

Di Lena et. al (2019)



Mean absolute error with respect to β-value range (i.e, you want the line to be as close to zero as possible).

Different panels = different % of missing values
Different colours = different methods

Regression-based methods (i.e. `methyLImp` and `missForest`) are the best performing ones

methyLImp

Our sub-plan here

1. General algorithm
2. Improvement 1: chromosome parallelization
3. Improvement 2: mini-batch for large number of samples
4. Hand-on try of the package

R package `methyLImp2` is available on Bioconductor
https://www.bioconductor.org/packages/release/bioc/html/methyLImp2.html

methyLImp

Methylation data shows high degree of inter-sample correlation -> we can use simple linear regression to impute missing variables. Impute groups of probes with the **same missingness pattern**.

$$\begin{array}{cc} & \text{probes ➡️} \\ \text{samples ⬇️} & \begin{pmatrix} x_{11} & x_{12} & NA & NA & x_{15} & x_{16} & NA & x_{18} \\ NA & x_{22} & x_{23} & x_{24} & x_{25} & NA & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{34} & x_{35} & x_{36} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} & x_{47} & x_{48} \end{pmatrix} \end{array}$$

methyLImp

1. Identify groups of probes with the same pattern of missingness

$$\begin{array}{cc} & \text{probes ➡️} \\ \text{samples ⬇️} & \begin{pmatrix} x_{11} & x_{12} & NA & NA & x_{15} & x_{16} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{24} & x_{25} & NA & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{34} & x_{35} & x_{36} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} & x_{47} & x_{48} \end{pmatrix} \end{array}$$

methyLImp

2. Choose one group and perform imputation on all its elements simultaneously

$$\begin{array}{cc}
 & \text{probes} \rightarrow \\
\text{samples} \downarrow & \begin{pmatrix}
x_{11} & x_{12} & NA & NA & x_{15} & x_{16} & NA & x_{18} \\
x_{12} & x_{22} & x_{23} & x_{24} & x_{25} & NA & x_{27} & x_{28} \\
x_{31} & x_{32} & NA & x_{34} & x_{35} & x_{36} & NA & x_{38} \\
x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} & x_{47} & x_{48}
\end{pmatrix} \\
 & \begin{array}{cccccccc}
 & & \Uparrow & & & & \Uparrow &
\end{array}
\end{array}$$

$$\begin{pmatrix}
NA & NA \\
NA & NA
\end{pmatrix}$$

Y

methyLImp

3. Drop all columns that contain at least one NA
(in 450K or 850K you'll still have plenty left)

$$\begin{array}{cc}
 & \text{probes ➡} \\
\text{samples ⬇} & \begin{pmatrix}
x_{11} & x_{12} & \boxed{NA} & \cancel{NA} & x_{15} & \cancel{x_{16}} & \boxed{NA} & x_{18} \\
x_{12} & x_{22} & x_{23} & \cancel{x_{24}} & x_{25} & \cancel{NA} & x_{27} & x_{28} \\
x_{31} & x_{32} & \boxed{NA} & \cancel{x_{34}} & x_{35} & \cancel{x_{36}} & \boxed{NA} & x_{38} \\
x_{41} & x_{42} & x_{43} & \cancel{x_{44}} & x_{45} & \cancel{x_{46}} & x_{47} & x_{48}
\end{pmatrix} \\
 & \begin{array}{cccccccc} & & \Uparrow & & & & \Uparrow & \end{array}
\end{array}$$

$$\begin{pmatrix} \boxed{NA} & \boxed{NA} \\ \boxed{NA} & \boxed{NA} \end{pmatrix}$$

Y

methyLImp

3. Drop all columns that contain at least one NA
(in 450K or 850K you’ll still have plenty left)

$$\begin{array}{cc} & \text{probes} \rightarrow \\ \text{samples} \downarrow & \begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix} \\ & \begin{matrix} & & \uparrow & & \uparrow & \end{matrix} \end{array}$$

$$\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix}$$

Y

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

$$logit^{-1}(p) = \frac{e^{p}}{e^{p}+1}$$

methyLImp

4. Construct your model

probes ➡️

samples ⬇️

$$\begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix}$$

⬆️ ⬆️

$$Y = logit^{-1}\left(X \cdot (A^{-1} logitB)\right)$$

$$\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix}$$

Y

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

$$logit^{-1}(p) = \frac{e^{p}}{e^{p}+1}$$

methyLImp

4. Construct your model

X: all non-missing values not in columns of interest for samples of interest

probes ➡️

samples ⬇️

$$\begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix}$$

$$\begin{array}{cccccc} \phantom{x_{11}} & \phantom{x_{12}} & \Uparrow & \phantom{x_{15}} & \Uparrow & \phantom{x_{18}} \end{array}$$

$$Y = logit^{-1}\left(X \cdot (A^{-1} logitB)\right)$$

$$\begin{pmatrix} x_{11} & x_{12} & x_{15} & x_{18} \\ x_{31} & x_{32} & x_{35} & x_{38} \end{pmatrix}$$

methyLImp

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

$$logit^{-1}(p) = \frac{e^{p}}{e^{p}+1}$$

4. Construct your model

probes ➡️

samples ⬇️

$$\begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix}$$

$$\begin{array}{cccccc} & & \Uparrow & & \Uparrow & \end{array}$$

A: all non-missing values not in columns of interest for not samples of interest

$$Y = logit^{-1}\left(X \cdot (A^{-1} logitB)\right)$$

$$\begin{pmatrix} x_{12} & x_{22} & x_{25} & x_{28} \\ x_{41} & x_{42} & x_{45} & x_{48} \end{pmatrix}$$

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

$$logit^{-1}(p) = \frac{e^p}{e^p + 1}$$

methyLImp

4. Construct your model

$$\begin{array}{cc} & \text{probes ➡️} \\ \text{samples ⬇️} & \begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix} \\ & \begin{matrix} \phantom{x_{11}} & \phantom{x_{12}} & \Uparrow & \phantom{x_{15}} & \Uparrow & \phantom{x_{18}} \end{matrix} \end{array}$$

$$Y = logit^{-1}\left(X \cdot (A^{-1} logit B)\right)$$

$$\begin{pmatrix} x_{23} & x_{27} \\ x_{43} & x_{47} \end{pmatrix}$$

B: all non-missing values
in columns of interest
for not samples of interest

$$logit(p) = ln\left(\frac{p}{1-p}\right)$$

$$logit^{-1}(p) = \frac{e^p}{e^p + 1}$$

methyLImp

4. Construct your model

probes ➡

samples ⬇

$$\begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix}$$

⬆ ⬆

$$Y = logit^{-1}\left(X \cdot (A^{-1} logit B)\right)$$

X: all non-missing values not in columns of interest for samples of interest

A: all non-missing values not in columns of interest for not samples of interest

B: all non-missing values in columns of interest for not samples of interest

$$\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix} = logit^{-1}(\begin{pmatrix} x_{11} & x_{12} & x_{15} & x_{18} \\ x_{31} & x_{32} & x_{35} & x_{38} \end{pmatrix}(\begin{pmatrix} x_{12} & x_{22} & x_{25} & x_{28} \\ x_{41} & x_{42} & x_{45} & x_{48} \end{pmatrix}^{-1} logit\begin{pmatrix} x_{23} & x_{27} \\ x_{43} & x_{47} \end{pmatrix}))$$

methyLImp2

5. Reduce dimensions

probes ➡️

$$\begin{array}{cc} \text{samples ⬇️} & \begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix} \end{array}$$

⬆️ ⬆️

$$Y = logit^{-1}\left(X \cdot (A^{-1} logit B)\right)$$

$$\underset{\text{2 x 2}}{\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix}} = logit^{-1}(\underset{\text{2 x 4}}{\begin{pmatrix} x_{11} & x_{12} & x_{15} & x_{18} \\ x_{31} & x_{32} & x_{35} & x_{38} \end{pmatrix}}(\underset{\text{4 x 2}}{\begin{pmatrix} x_{12} & x_{22} & x_{25} & x_{28} \\ x_{41} & x_{42} & x_{45} & x_{48} \end{pmatrix}}^{-1} logit \underset{\text{2 x 2}}{\begin{pmatrix} x_{23} & x_{27} \\ x_{43} & x_{47} \end{pmatrix}}))$$

methyLImp2

5. Reduce dimensions

$$\begin{array}{cc} & \text{probes} \rightarrow \\ \text{samples} \downarrow & \begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix} \\ & \begin{array}{cccccc} \phantom{x_{11}} & \phantom{x_{12}} & \uparrow & \phantom{x_{15}} & \uparrow & \phantom{x_{18}} \end{array} \end{array}$$

$$Y = logit^{-1}\left(X \cdot (A^{-1} logit B)\right)$$

$$\underset{2 \times 2}{\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix}} = logit^{-1}(\underset{2 \times 4}{\begin{pmatrix} x_{11} & x_{12} & x_{15} & x_{18} \\ x_{31} & x_{32} & x_{35} & x_{38} \end{pmatrix}}(\underset{4 \times 2}{\begin{pmatrix} x_{12} & x_{22} & x_{25} & x_{28} \\ x_{41} & x_{42} & x_{45} & x_{48} \end{pmatrix}}^{-1} logit \underset{2 \times 2}{\begin{pmatrix} x_{23} & x_{27} \\ x_{43} & x_{47} \end{pmatrix}}))$$

100 samples, 850K probes

methyLImp2

5. Reduce dimensions

$$\begin{array}{cc} & \text{probes} \rightarrow \\ \text{samples} \downarrow & \begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix} \\ & \begin{matrix} \phantom{x_{11}} & \phantom{x_{12}} & \uparrow & \phantom{x_{15}} & \uparrow & \phantom{x_{18}} \end{matrix} \end{array}$$

$$Y = logit^{-1}\left(X \cdot (A^{-1} logit B)\right)$$

$$\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix} = logit^{-1}(\begin{pmatrix} x_{11} & x_{12} & x_{15} & x_{18} \\ x_{31} & x_{32} & x_{35} & x_{38} \end{pmatrix}(\begin{pmatrix} x_{12} & x_{22} & x_{25} & x_{28} \\ x_{41} & x_{42} & x_{45} & x_{48} \end{pmatrix}^{-1} logit \begin{pmatrix} x_{23} & x_{27} \\ x_{43} & x_{47} \end{pmatrix}))$$

2 x 2		2 x 4	4 x 2	2 x 2
	|| over chromosomes	100 samples, 850K probes		
2 x 2		2 x 850K	850K x 98	98 x 2

methyLImp2

5. Reduce dimensions

probes ➡️

$$\begin{array}{cc} \text{samples} \Downarrow & \begin{pmatrix} x_{11} & x_{12} & NA & x_{15} & NA & x_{18} \\ x_{12} & x_{22} & x_{23} & x_{25} & x_{27} & x_{28} \\ x_{31} & x_{32} & NA & x_{35} & NA & x_{38} \\ x_{41} & x_{42} & x_{43} & x_{45} & x_{47} & x_{48} \end{pmatrix} \\ & \begin{array}{cccccc} \phantom{x_{11}} & \phantom{x_{12}} & \Uparrow & \phantom{x_{15}} & \Uparrow & \phantom{x_{18}} \end{array} \end{array}$$

$$Y = logit^{-1}\left(X \cdot (A^{-1} logit B)\right)$$

$$\begin{pmatrix} NA & NA \\ NA & NA \end{pmatrix} = logit^{-1}(\begin{pmatrix} x_{11} & x_{12} & x_{15} & x_{18} \\ x_{31} & x_{32} & x_{35} & x_{38} \end{pmatrix}(\begin{pmatrix} x_{12} & x_{22} & x_{25} & x_{28} \\ x_{41} & x_{42} & x_{45} & x_{48} \end{pmatrix}^{-1} logit\begin{pmatrix} x_{23} & x_{27} \\ x_{43} & x_{47} \end{pmatrix}))$$

2 x 2		2 x 4	4 x 2	2 x 2
	|| over chromosomes	100 samples, 850K probes		
2 x 2	mini-batch	2 x 850K	850K x 98	98 x 2

methyLImp2: || over chromosomes

1. Split the probes across chromosomes: new dimensions are max 80K, min 10K instead of 850K
2. Apply *methyLImp2* to each dataset independently on different cores.

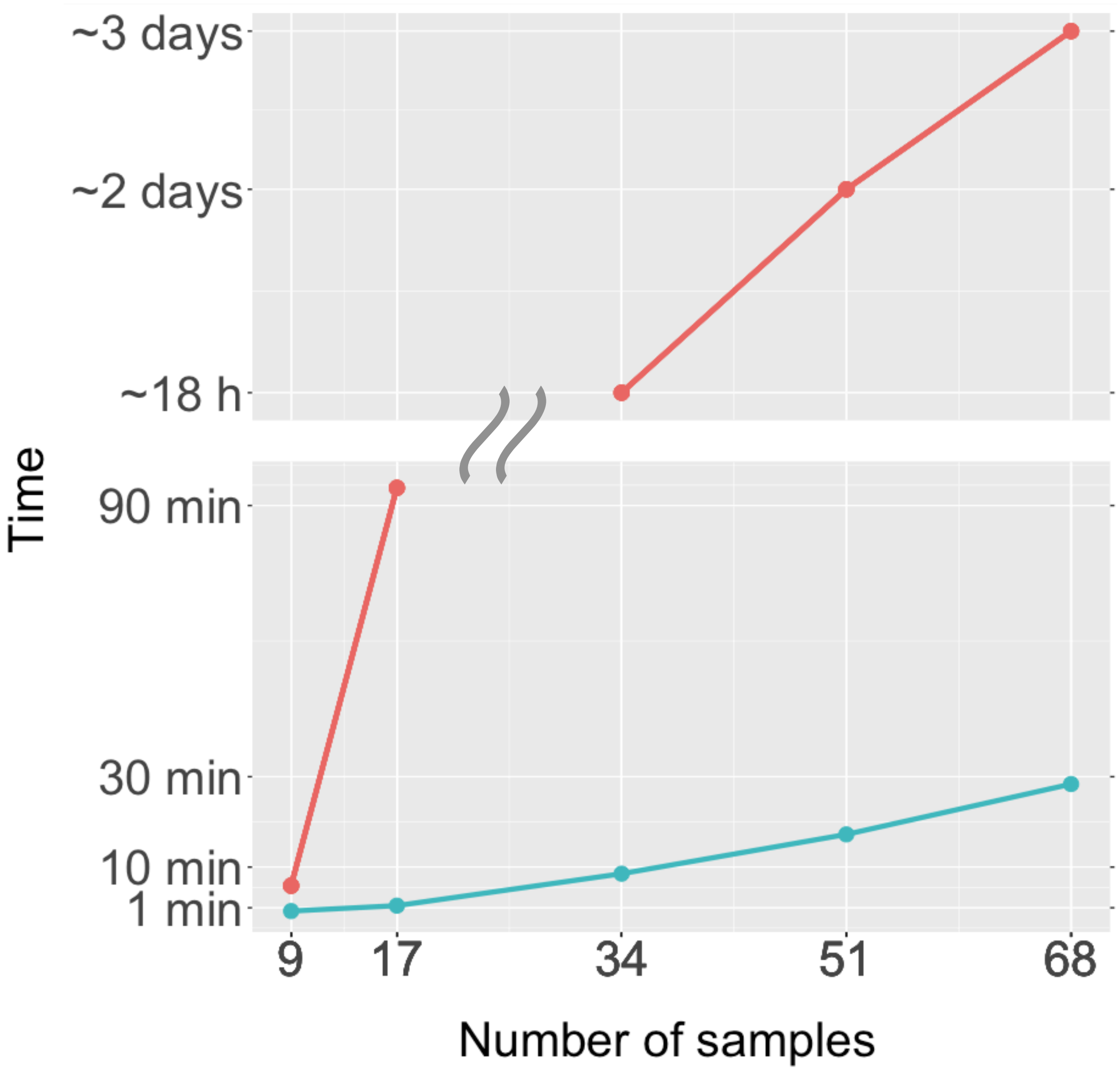
Performance is the same, running time is drastically lower!

methyLImp2: || over chromosomes

1. Split the probes across chromosomes: new dimensions are max 80K, min 10K instead of 850K
2. Apply *methyLImp2* to each dataset independently on different cores.

Performance is the same, running time is drastically lower!

Running time for original *methyLImp* and for *methyLImp2*

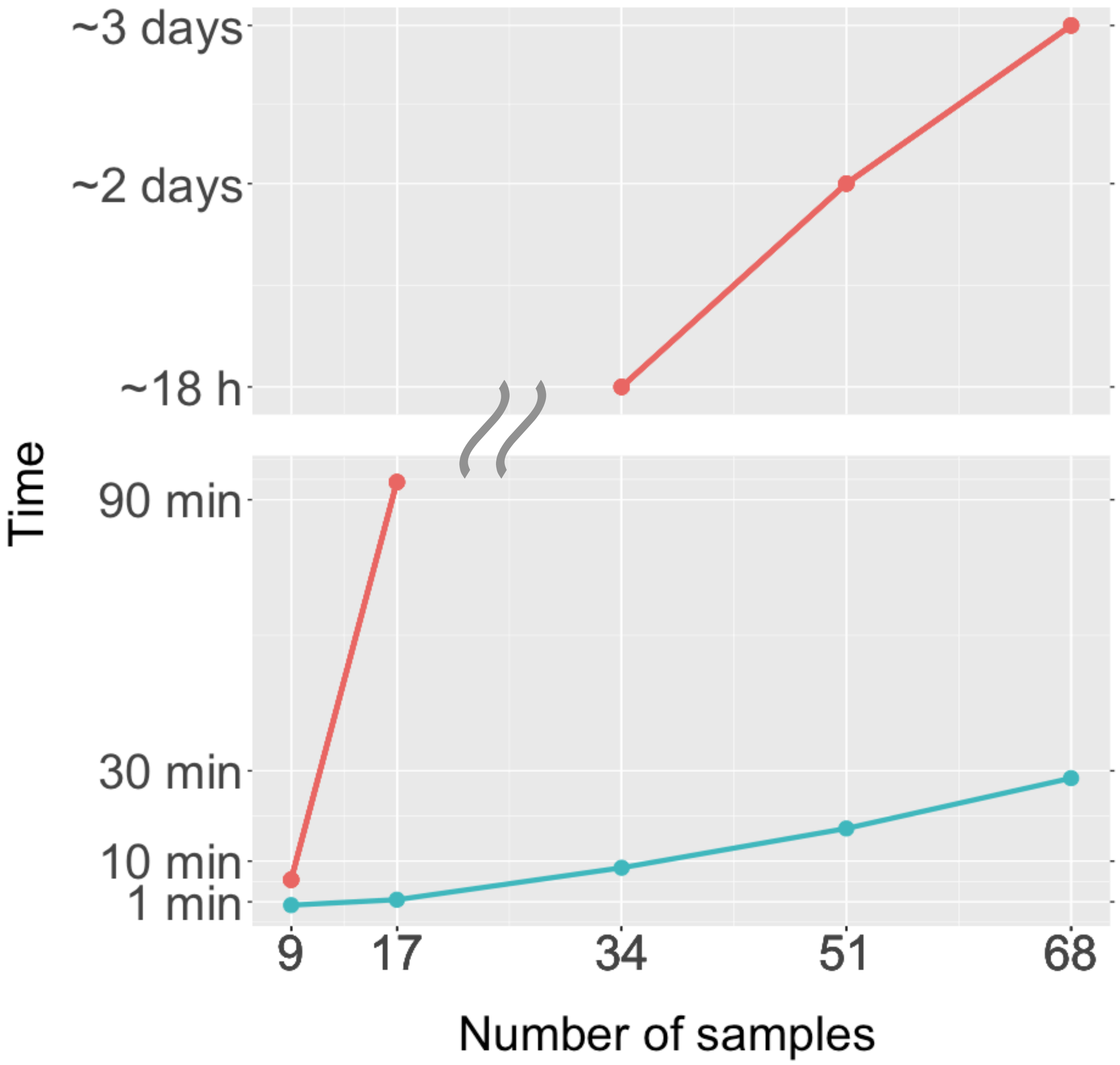


methyLImp2: || over chromosomes

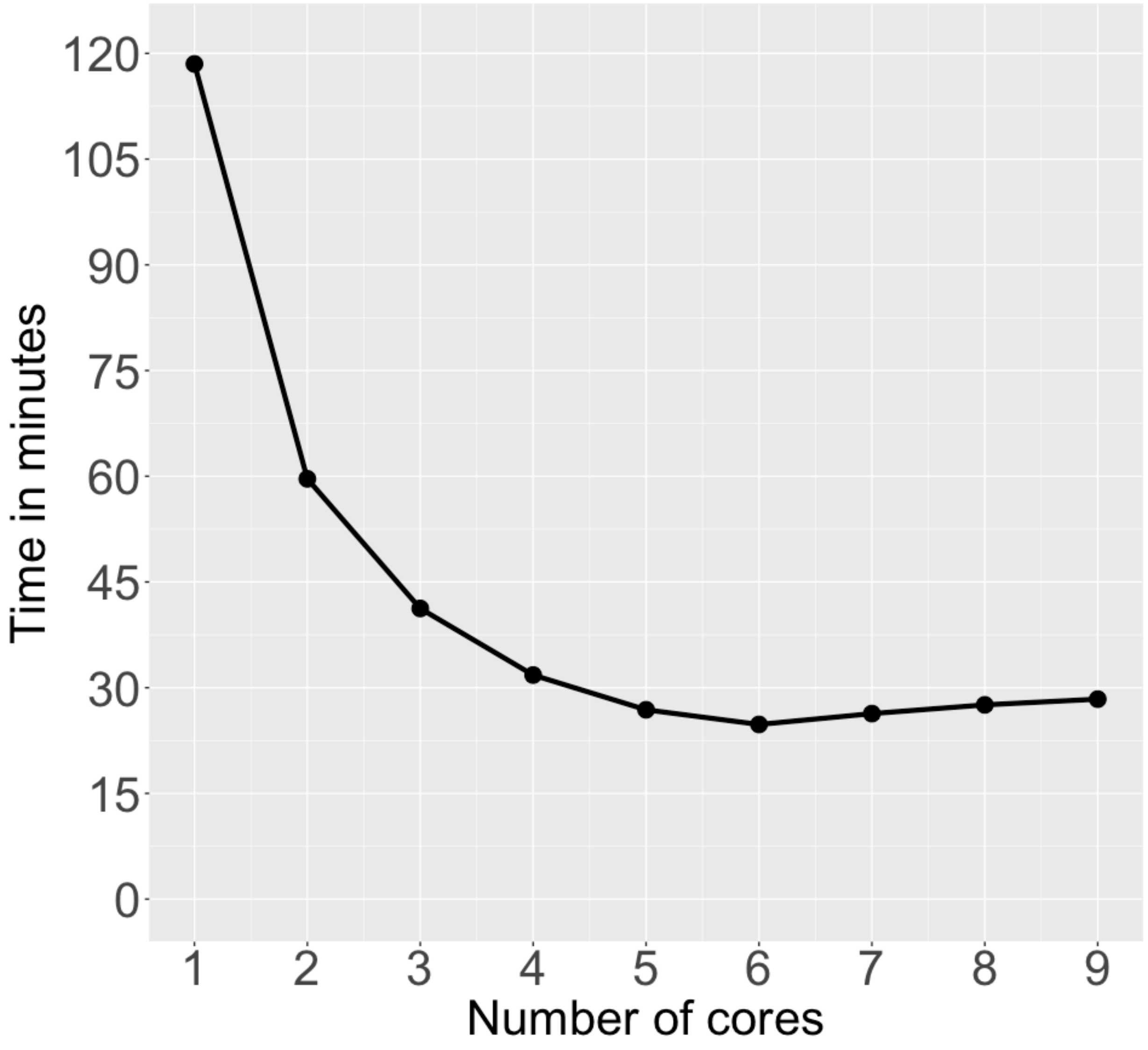
1. Split the probes across chromosomes: new dimensions are max 80K, min 10K instead of 850K
2. Apply *methyLImp2* to each dataset independently on different cores.

Performance is the same, running time is drastically lower!

Running time for original *methyLImp* and for *methyLImp2*



Running time for *methyLImp2* for the varying number of cores



methyLImp2: mini-batch

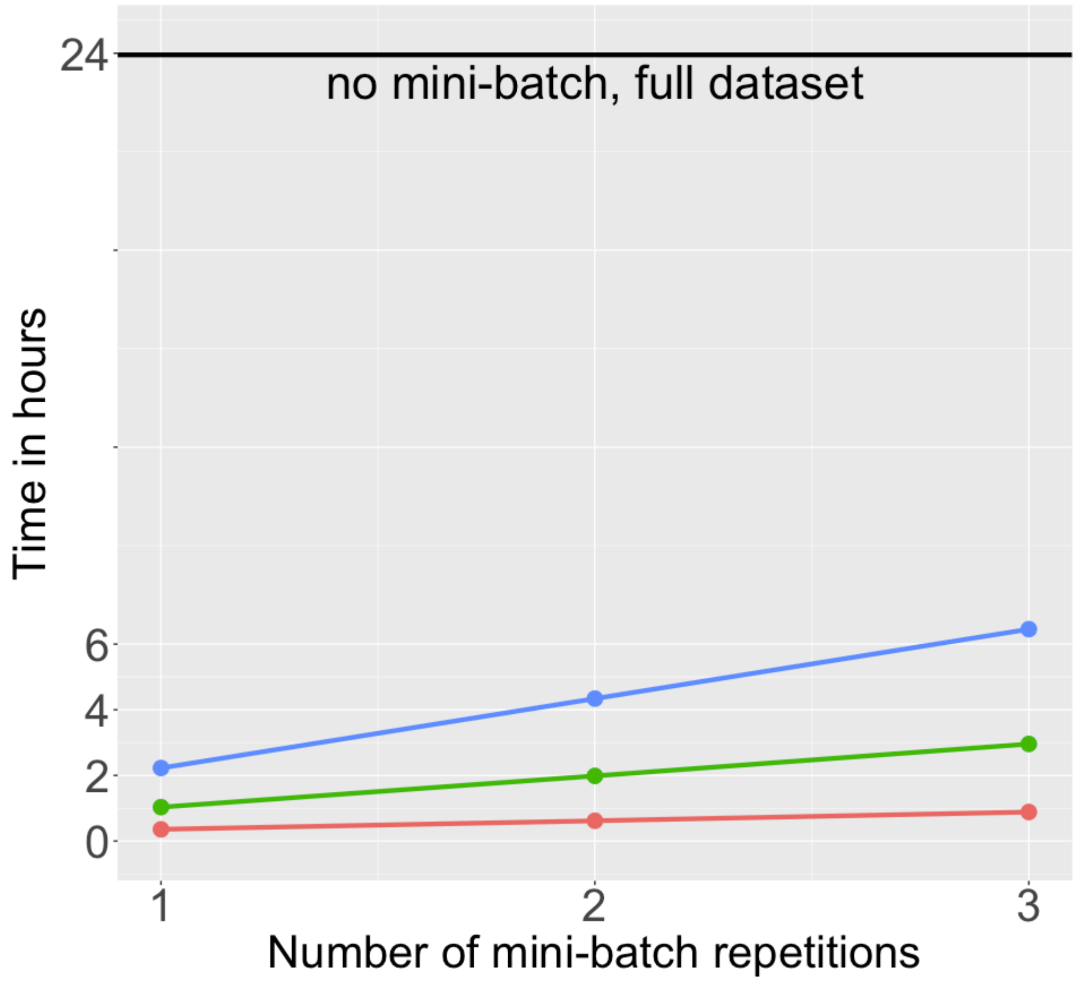
When you have lots of samples, maybe not all of them are needed for the imputation… Try **mini-batch** (for each group of columns with the same pattern that you impute):
– randomly choose some fraction of samples (10/20/30%);
– apply *methyLImp*;
– repeat several times (1/2/3);
– average the results.

methyLImp2: mini-batch

When you have lots of samples, maybe not all of them are needed for the imputation… Try **mini-batch** (for each group of columns with the same pattern that you impute):

– randomly choose some fraction of samples (10/20/30%);
– apply *methyLImp*;
– repeat several times (1/2/3);
– average the results.

Running time for 10, 20 and 30% of samples used in the mini-batch approach

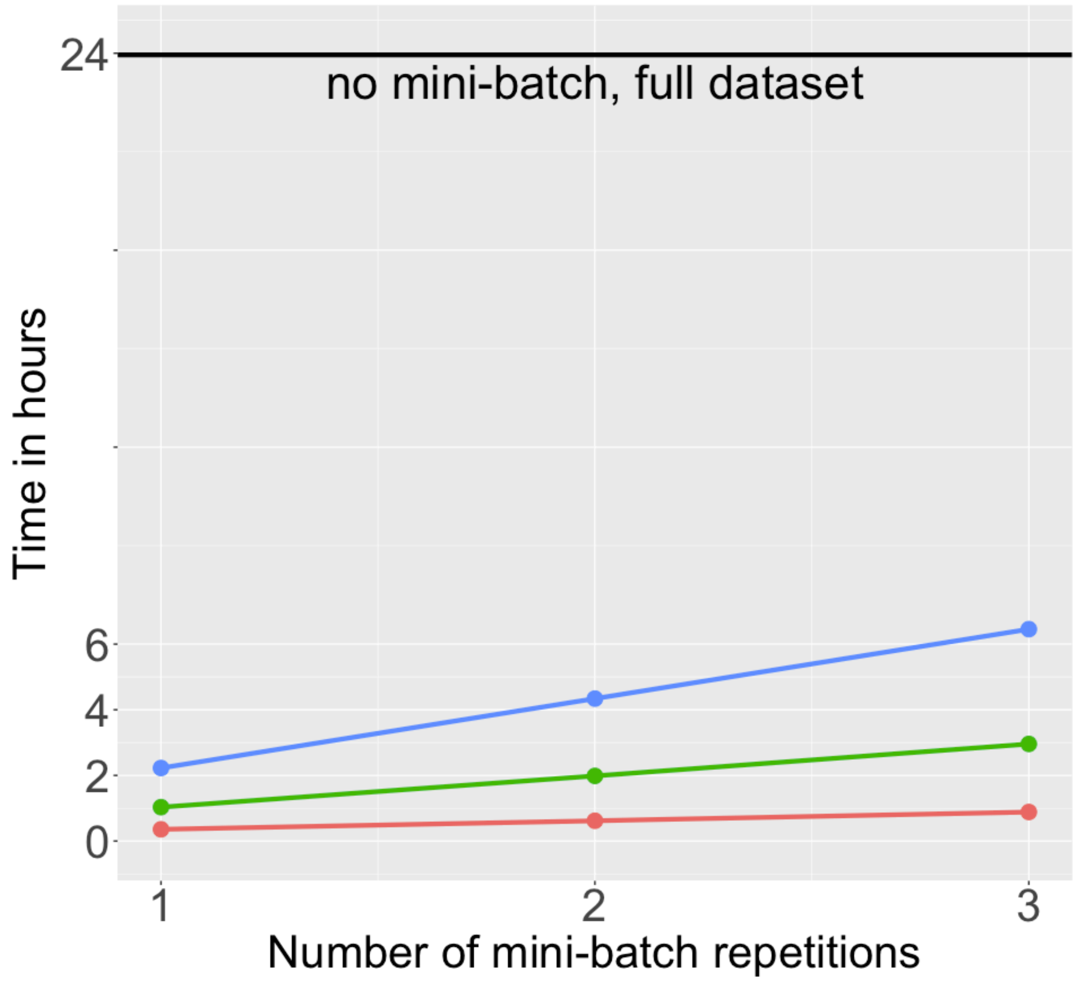


methyLImp2: mini-batch

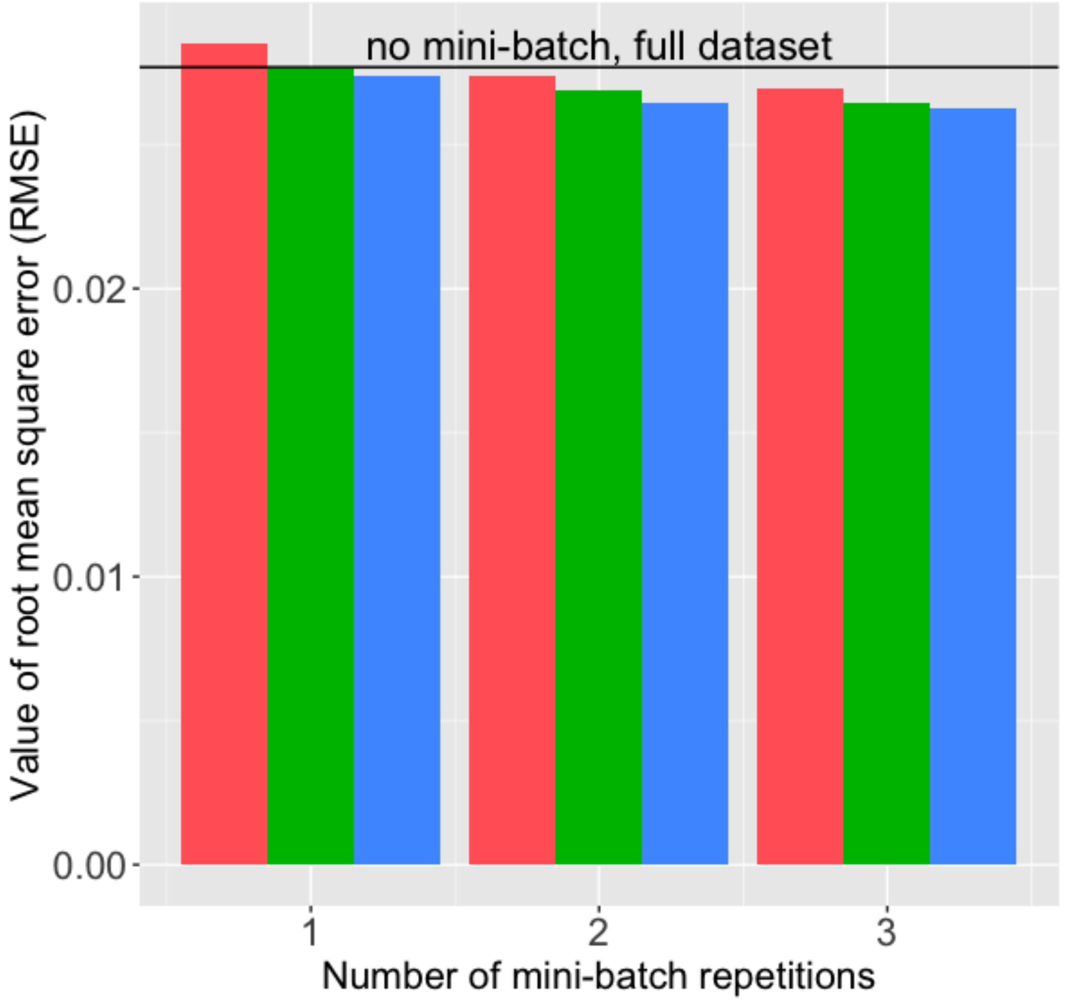
When you have lots of samples, maybe not all of them are needed for the imputation… Try **mini-batch** (for each group of columns with the same pattern that you impute):

– randomly choose some fraction of samples (10/20/30%);
– apply *methyLImp*;
– repeat several times (1/2/3);
– average the results.

Running time for 10, 20 and 30% of samples used in the mini-batch approach



RMSE for 10, 20 and 30% of samples used in the mini-batch approach

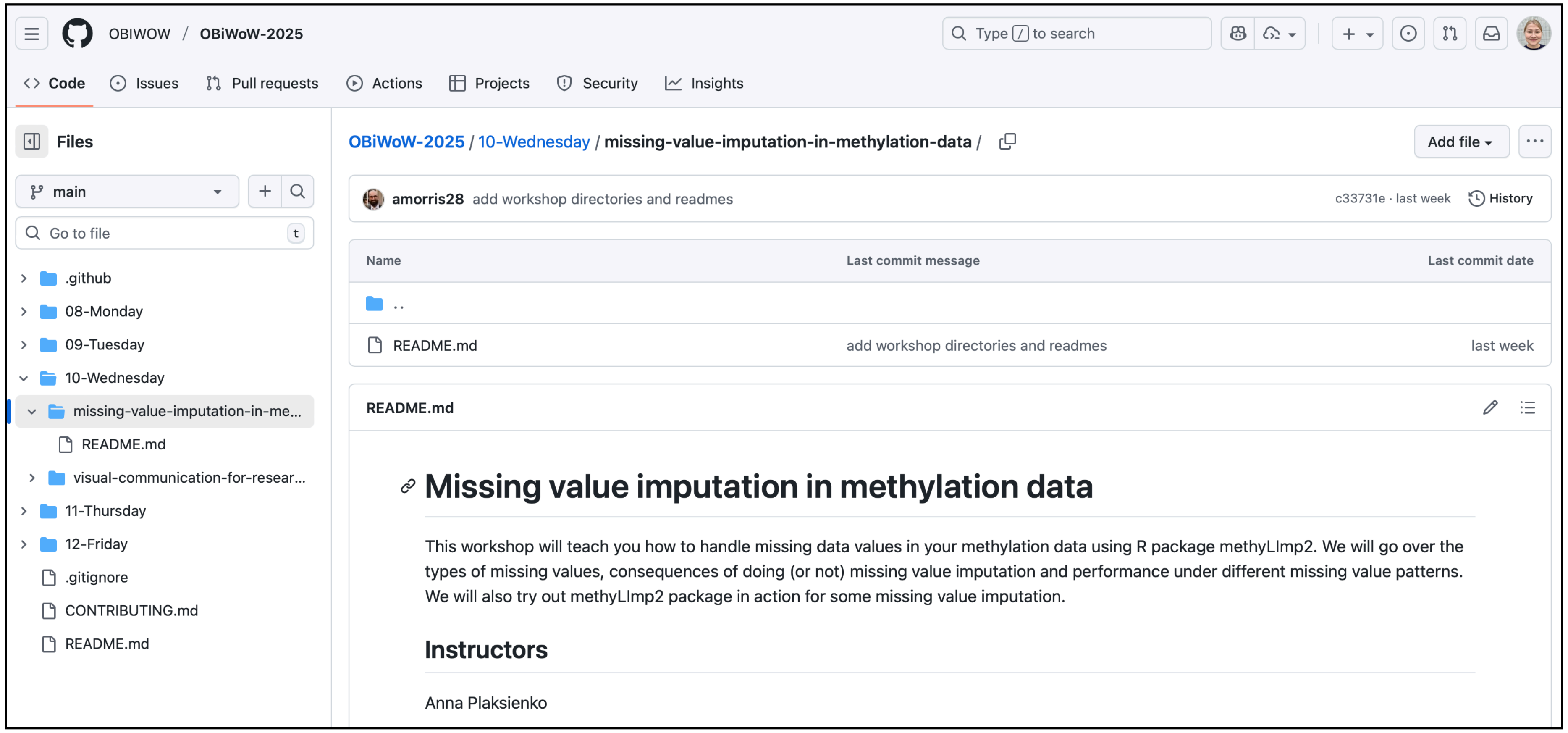


Time to try!

Please access our Rmarkdown for the try-out part of the lecture

https://github.com/OBIWOW/OBiWoW-2025/tree/main/10-Wednesday/missing-value-imputation-in-methylation-data

Download beta matrix is here https://uio-my.sharepoint.com/:u:/g/personal/annapla_uio_no/IQAbtwHa63lVTpsBWvMhf5BxAWsbdbzVeuZ1596Pl2-vQ2c?e=BaclzF



OBiWoW-2025 / 10-Wednesday / missing-value-imputation-in-methylation-data /

amorris28 add workshop directories and readmes c33731e · last week

Name	Last commit message	Last commit date
..		
README.md	add workshop directories and readmes	last week

README.md

Missing value imputation in methylation data

This workshop will teach you how to handle missing data values in your methylation data using R package methyLImp2. We will go over the types of missing values, consequences of doing (or not) missing value imputation and performance under different missing value patterns. We will also try out methyLImp2 package in action for some missing value imputation.

Instructors

Anna Plaksienko

Feedback

Please take 5 minutes to fill out
a **feedback survey**.
It is useful for improvement
of the workshop!