Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Original paper

OXFORD

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,*]

[1]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy.
[2]Department of Physics and Astronomy,University of Bologna, Bologna, Italy.
[3]Smart Cities Living Lab, Institute of Organic Synthesis and Photoreactivity, CNR, Bologna, Italy.
[4]Department of Laboratory Medicine, Karolinska Institutet, Stockholm, Sweden and CNR IAC "Mauro Picone", Roma, Italy, and Personal Genomics S.r.l., Verona, Italy.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: DNA methylation is a stable epigenetic mark with major implications in both physiological (development, aging) and pathological conditions (cancers and numerous diseases). Recent research involving methylation focuses on the development of molecular age estimation methods based on DNA methylation levels (mAge). An increasing number of studies indicate that divergences between mAge and chronological age may be associated to age-related diseases. Current advances in high-throughput technologies have allowed the characterization of DNA methylation levels throughout the human genome. However, experimental methylation profiles often contain multiple missing values, that can affect the analysis of the data and also mAge estimation. Although several imputation methods exist, a major deficiency lies in the inability to cope with large datasets, such as DNA methylation chips. Specific methods for imputing missing methylation data are therefore needed.

Results: We present a simple and computationally efficient imputation method, *metyhLImp*, based on linear regression. The rationale of the approach lies in the observation that methylation levels show a high degree of inter-sample correlation. We performed a comparative study of our approach with other imputation methods on DNA methylation data of healthy and disease samples from different tissues. Performances have been assessed both in terms of imputation accuracy and in terms of the impact imputed values have on mAge estimation. In comparison to existing methods, our linear regression model proves to perform equally or better and with good computational efficiency. The results of our analysis provide recommendations for accurate estimation of missing methylation values.

Availability and implementation: The R package *methyLImp* is freely available at `https://github.com/pdilena/methyLImp`.

Contact: name@bio.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Among the plethora of omics that are now available for the study of biological systems, epigenomics is gaining strength. Epigenomics is concerned with the study of alterations in gene expression that do not imply DNA base sequence changes, but imply instead stable yet reversible modifications that chemically alter such bases or histones, and that are mitotically (and sometimes meiotically) heritable. This heritability makes epigenetics a crucial factor to be considered in the study of causative biological mechanisms, and a desirable mean of investigation, owing to its stability. Epigenetics is more specifically concerned with three types of modifications: DNA methylation, histones' modification and post-transcriptional changes such as the activity of miRNAs. This work focuses on DNA methylation, enabled by cost-effectiveness of highly parallel

assays (methylation arrays) and by robustness of the signals, owing to the reversible, enzyme-dependent addition (DNA methylatranspherase) of a methyl group to the 5'-end of the CpG dinucleotide.

DNA methylation is crucial in embryonic development and in aging (Nardini *et al*., 2018), leading to tissue-specific cells' identity, by selective and stable, although reversible, silencing of portions of the genome. Importantly, methylation is linked to the accessibility to the DNA molecule, normally tightly packed around histones. Therefore, methylation patterns present with regularity across individuals owing to the underlying biophysical regularity of the packaging, observable even in the unfolded methylated DNA. Methylation is also involved in phenomena with higher dynamics, tied to the reversibility of epigenomic changes, such as the modifications of immune cells upon environmental challenges, for periods that can last from weeks to years and up to a lifetime, fundamental in the setup of the so-called cell memory, crucial in vaccination (Ciabattini *et al*., 2018). Recently, research involving methylation has focused on establishing correlations between changes in chronological and methylation age (mAge), assessed via a number of epigenetic clocks (Garagnani *et al*., 2012; Hannum *et al*., 2013; Horvath, 2013; Weidner *et al*., 2014). The rationale for epigenetic clocks descends from the observation that divergence form the expected correlation between chronological and the expected methylation at specific CpGs (base of the computation of mAge) is proven to be: i) indicative and sometimes predictive (Durso *et al*., 2017) of a number of diseases in case of acceleration (i.e. mAge higher than chronological age), such as in Down Syndrome, Parkinson's and Alzheimer's diseases, cancers (Horvath and Raj, 2018); ii) indicative of successful aging in case of deceleration, such as in centenarians and their off-springs (Horvath *et al*., 2015).

Given the relevance of the computation of mAge for application in areas that go from cancer prevention to healthy aging, we here focus on the optimization of such calculation, intimately tied to the presence of valid values at the clocks' specific CpGs. Similarly to other omics, missing data represent an issue researchers have been coping with via a number of imputation approaches (see Section 2.2). With reference to the computation of mAge the issue is more pressing, as relevant information does not have the dimensionality of the whole chip, but is restrained to the CpGs necessary to compute the clock, from a handful to some hundreds. Based on this need and on the peculiar characteristics of methylation (stability and robustness) resulting in inter-sample correlation, we can successfully exploit simple multiple linear regression models for limited-range missing values imputation. Although the more advanced approaches for data imputation are able to deal with heterogeneous (continuous and categorical) variables, this usually comes at the cost of being computationally intensive. Despite linear regression models may not be suitable for imputation in heterogeneous datasets, we show how, in this context, they represent the best compromise between accuracy and computational efficiency for data imputation in general, and for assessment of mAge, in particular.

2 Background

2.1 DNA methylation

DNA methylation is an epigenetic modification involving the covalent addition of a methyl group to the 5'-carbon of cytosine in a CpG dinucleotide, making DNA CpGs rich areas more prone to methylation. The Illumina Infinium assay (Illumina 27k and 450k and now 850k Human Beadchip) is currently the most cost-effective technology for estimating the methylation level of DNA samples. Infinium assays utilize a pair of probes to measure the intensities of the methylated and unmethylated alleles at each CpG site. The methylation level is then estimated based on the measured intensities of this pair of probes, across all cells in the sample tissue.

Two measures are commonly used to quantify the methylation level. The first one, called β-value (ranging from 0 to 1), is the ratio between the methylated probe intensity and the overall intensity of both methylated and unmethylated alleles. The second method, called M-value (ranging from $-\infty$ to ∞), is the log2 ratio of the β-value, i.e $\log_2(\beta/(1-\beta))$. Despite the desirable statistical properties of M-value metric for differential analysis of methylation levels (Du *et al*., 2010), β-value is the predominantly used metric owing to its intuitive biological interpretation, and it is recommended by array producers (Bibikova *et al*., 2006).

Age-related changes in DNA methylation has recently become an active field of investigation. In particular, several studies describe mAge estimators, based on β-value representation of methylation levels from a limited number of CpG sites across the human genome. Garagnani *et al*. (2012) identified the methylation of the EOVL2 gene as an epigenetic marker of aging; Hannum developed an age predictor based on 71 CpG markers from the whole blood tissue (Hannum *et al*., 2013); Weidner's predictor (Weidner *et al*., 2014) is based on the methylation level of three CpGs in blood samples; Horvath's model is a multi-tissue predictor relying on 353 CpGs markers, trained with elastic-net regression that includes both a ridge and a lasso penalization (Horvath, 2013). Given the limited set of markers considered by such epigenetic clocks, the mAge estimation accuracy is crucially dependent on the availability of the methylation levels for all the selected CpGs. Due to its widespread usage and validation in a variety of tissues, and to the number of CpGs, we limit our analysis to Horvath's clock.

2.2 Missing Data imputation

Missing value imputation involves the estimation of the missing entries in a data matrix of experimental or analytical measurements by exploiting information about available data. Missing data are a common issue in numerous scientific research domains from Biology (Troyanskaya *et al*., 2001; Stekhoven and Bülmann, 2012; Severson *et al*., 2017) to Medicine (Donders *et al*., 2006) and Social Sciences (Durrant, 2009). Many approaches have been proposed and there exists a vast literature on this topic (Enders, 2010).

Missing data are categorized into the following three types (Little and Rubin, 1986): a value is i) *missing completely at random* (MCAR) if the probability of it being missing does not depend on either the observed or the missing values; ii) *missing at random* (MAR) if the the probability of it being missing does not depend on the value that is missing but it may depend on the observed values; iii) *missing not at random* (MNAR) if the probability of it being missing depends on the value that is missing. Imputation methods require assumptions about the underlying missing data mechanism. Unfortunately, there is no statistical way to determine in which category of missingness the data falls. Assumptions are generally made based upon the knowledge of the data and data collection procedure. There is no reference in literature that addresses the missingness patterns in DNA methylation data. Our assumption is that missing values are MCAR/MAR due to random experimental errors and technology-related errors, e.g. some probes fail to capture target sequences, hence for some probes there is a slightly higher chance to have a missing value (independently of the value itself).

In general terms, imputation approaches can be classified into methods that deal with continuous or categorical variables (or both). Some approaches can handle also limited-range variables, avoiding out-of-range imputations. Another substantial distinction can be done between single (SI) and multiple imputation (MI) methods. SI methods replace a missing value with one reasonable value. MI methods perform several single imputations and average the parameters' estimates across the multiple

imputations to produce a single estimate. The most basic imputation method is the *mean*, where the missing value of a variable is replaced with the mean of the non-missing observations for that variable. The *mean* approach is the simplest example of SI method that can handle limited-range continuous variables. Under the MCAR/MAR assumptions, the most common (SI or MI) imputation methods can deal with the missing data (Bennet, 2001). In MNAR data, the missingness needs to be modelled explicitly.

Although all imputation methods that can deal with continuous variables are, in principle, suitable for DNA methylation data imputation, in this domain a comparative study of missing value imputation methods, as well as specific imputation methods, has not been presented yet, to the best of our knowledge. We only mention two recent works related to this topic. In Zhang *et al*. (2016), the authors exploit a penalized functional regression model to estimate the Illumina 450k Human Beadchip β-values from 27k Human BeadChip samples. In Wu *et al*. (2016), the authors propose a site selection method followed by MI to estimate missing covariate values and to perform association tests in epigenome-wide association studies.

3 Materials and Methods

3.1 Linear regression model for β-value imputation

We exploit a simple SI multiple linear regression model for limited-range continuous variables. In detail, the missing values are imputed by iteratively performing linear regression with pseudo-inverse on the available (observed) data. The restriction of the range is specified within the imputation procedure. The rationale of this approach is that methylation levels exhibit both long- and short-range correlations (Lövkvist *et al*., 2016; Zhang *et al*., 2017) that can be captured by a simple linear regression.

The pseudo-code of the approach is described in Algorithm 1 (see Supplementary Information 1). The method takes as input an $n \times m$ data matrix containing missing values. The columns of the matrix correspond to variables (CpGs) and the rows to observations (samples). In the following we assume that the matrix does not contain samples with only missing values and that it does contain only values in the range $[0, 1]$ (e.g β-values). However, the interface of the R-package implementation *methyLImp* works also with unrestricted-range continuous values.

The *methyLImp* algorithm contains a (parallelizable) loop (lines 7-18) that evaluates a non-empty list L of variables containing missing values. The body of the loop selects the first variable in the L list and extracts the list of all observations R_{NA} (i.e. rows of the matrix) whose value is missing (line 8). In order to lower the number of linear regressions to be performed and, thus, speed up the computation, the complete list of variables C_{NA} (i.e. columns of the matrix) for which only the selected observations R_{NA} have missing values is also computed (line 9). In a single step (line 15) the algorithm imputes all the values in the submatrix $M[R_{NA}, C_{NA}]$ and then removes the C_{NA} variables from the evaluation list L (line 17). The approach does not work for those variables whose value is missing in all samples. Such variables are simply ignored (line 11). A linear regression with pseudo-inverse is used to compute the coefficients, x, that solve the linear system:

$$A \cdot x = \texttt{logit}(B) \Longrightarrow x = A^{-1} \cdot \texttt{logit}(B)$$

and the missing values are then imputed by $\texttt{logit}^{-1}(X \cdot x)$, where:

- A is the submatrix of all non-missing variables not in C_{NA} for all the observations not in R_{NA} (line 12);
- B is the submatrix of all non-missing variables in C_{NA} for all the observations not in R_{NA} (line 13);
- X is the submatrix of all non-missing variables not in C_{NA} for all the observations in R_{NA} (line 14).

The $\texttt{logit(p)} = \log(p/(1-p)), p \in [0,1]$ function and its inverse logistic function $\texttt{logit}^{-1}(q) = 1/(1-\exp(-q))), q \in (-\infty, \infty)$ are used to ensure that the imputed values are in the range $[0, 1]$. The range of the actual implementation of the logit function does not span from $-\infty$ to ∞ but is fixed to a limited range. The pseudo-inverse A^{-1} is computed using the singular value decomposition of A. Although the implementation allows for dimensionality reduction of the pseudo-inverse, in practice we experimented that low rank representations lower the imputation accuracy (data not shown).

We remark that our regression model uses only completely observed submatrices. Thus, the approach cannot be applied if each variable in the matrix contains at least one missing value. However, in practice this does not happen, since due to the pattern distribution of missing values in real DNA methylation data, the regression submatrices (A and X) are typically very large, involving thousands of variables (see Supplementary Information 1).

Algorithm 1 Impute missing values in a matrix of β-values

1: **function** methyLImp($M \in [0,1]^{n \times m}$)
2: $M' \leftarrow M$
3: $R \leftarrow \{1, ..., n\}$ ▷ All row indexes
4: $C \leftarrow \{1, ..., m\}$ ▷ All column indexes
5: $NA \leftarrow \{c \mid \exists r, M[r,c] \text{ is missing }\}$ ▷ Cols with missing values
6: $L \leftarrow NA$
7: **while** $L \neq \emptyset$ **do**
8: Select some $col \in L$
9: $R_{NA} \leftarrow \{r \mid M[r, col] \text{ is missing }\}$
10: $C_{NA} \leftarrow \{c \mid M[r,c] \text{ is missing } \iff r \in R_{NA}\}$
11: **if** $R \setminus R_{NA} \neq \emptyset$ **and** $C \setminus NA \neq \emptyset$ **then**
12: $A \leftarrow M[R \setminus R_{NA}, C \setminus NA]$
13: $B \leftarrow M[R \setminus R_{NA}, C_{NA}]$
14: $X \leftarrow M[R_{NA}, C \setminus NA]$
15: $M'[R_{NA}, C_{NA}] \leftarrow \texttt{logit}^{-1}(X \cdot [A^{-1} \cdot \texttt{logit}(B)])$
16: **end if**
17: $L \leftarrow L \setminus C_{NA}$
18: **end while**
19: **return** M'
20: **end function**

3.2 Missing data imputation software

In order to cover representative techniques described in literature, our linear regression model's performances were benchmarked against six R-implemented imputation methods. We focused only on methods that can handle continuous variables and large amount of data without requiring massive computational power. The complete list is available in the Supplementary Information 1.

Briefly, two of the six methods, *mean* and *impute.knn* (Troyanskaya *et al*., 2001), are based on the classical and computationally efficient mean-value imputation technique. The *mean* approach consists of simply replacing the missing value of a given variable by averaging all the known values for that variable. The *impute.knn* approach, used in Horvath (2013), replaces a missing element for a variable by averaging the non-missing values of its nearest neighbours. Three methods, *SVDmiss* (Fuentes *et al*., 2006), *softImpute* (Mazumder *et al*., 2010) and *imputePCA* (Josse and Husson, 2013), are roughly based on iterative soft-thresholding of the input matrix. The techniques in this class first replace the missing values with some initial guess and then iteratively update, up to convergence, the missing elements with values generated by a low-rank approximation

of the input matrix. *SVDmiss* and *softImpute* iteratively perform a soft-thresholding SVD (Singular Value Decomposition) of the input matrix. *imputePCA* implements a low-rank approximation version of the iterative PCA (Principal Component Analysis) algorithm, also known as EM-PCA (Dempster *et al*., 1977). Finally, *missForest* (Stekhoven and Bülmann, 2012) is an iterative method based on random forest regression trees. All such methods have been run with the default parameters provided in their R implementations.

3.3 Benchmark data

Benchmark datasets, were chosen from the NCBI database Gene Expression Omnibus (GEO) (Edgar *et al*., 2002), filtering by platform GPL13534 (450k Human Beadchip) available up to July 1st 2017. The datasets were split into smaller sets, separating healthy controls from disease case samples. Healthy and disease datasets were further split in order to separate samples from different tissues. Only healthy controls with age information and datasets containing at least four samples were retained. The healthy control set consists of 37 benchmark datasets with overall 1,495 samples. The disease case set consists of 21 benchmark datasets with overall 386 samples. Due to the high cpu time required by some imputation methods to process the largest datasets, we chose to pre-filter all datasets in order to decrease their size. We reduced the dataset information by keeping only the methylation sites in the intersection between the Illumina 27k and 450k Human Beadchips, approximatively 21k sites, already selected by Horvath for training his epigenetic clock.

Each dataset originally contains a variable number of missing values. On average, the number of variables (CpGs) that contain missing informations is typically quite small. The full list with detailed informations is available in the Supplementary Information 1.

3.4 Experimental setup

The general workflow of our analysis is as follows. Given a DNA methylation dataset, we introduced in the data varying percentages of missing values, under the MCAR assumption of random experimental error. The selected methods were used to impute the missing values in the dataset and Horvath's clock was used to estimate the mAge after imputation.Calculation of mAge was done using the wateRmelon (Pidsley *et al*., 2013) implementation of the Horvath's epigenetic clock (Horvath, 2013). Performances were evaluated in terms of difference between imputed and original β-values, as well as difference in mAge estimation between original and imputed data. For performance evaluation we adopted the most popular evaluation metrics for missing data imputation (see Supplementary Information 1): RMSE (Root Mean Square Error), MAE (Mean Absolute Error), PCC (Person Correlation Coefficient) and MAPE (Mean Average Percentage Error).

Since one of our primary interest is investigating to which extent data imputation affects the mAge estimation, we introduced missing values uniquely for the 353 CpG markers used in Horvath's clock: inaccurate imputations for different markers do not affect Horvath's mAge estimation. In more detail, we randomly introduced 10% (35), 30% (106), 50% (176) and 70% (247) missing values among the Horvat's 353 markers. In order to make the analysis more robust, we repeated such random introduction of missing values 10 times for each percentage. These 40 random sets of CpG markers are stable throughout the analysis. In each test, missing values were introduced only for a single sample at a time, leaving the remaining samples in the dataset unchanged. Thus, for each sample we performed 40 different tests with varying percentages and types of missing values. Each imputation method was then run a total of $40 \cdot (1,495 + 386) = 75,240$ times.

All the experiments were performed on a computer with a double Intel Xeon® E5-2683 v4 CPU @ 2.10GHz with a total of 64 computing cores and 256 GB of RAM at the Smart Cities Living Lab, CNR.

4 Results

4.1 Imputation accuracy assessment

We compare here the performances of the imputation methods in terms of imputation accuracy per CpG site.

Table 1 and Table 2 show the overall performances with respect to the percentage of missing values among the 353 CpGs in Horvath's clock on healthy and disease samples, respectively. The evaluation metric scores in Tables 1 and 2 have been averaged over the ten random replicas. The error measures and the detailed results per dataset are available in Supplementary Information 2.

Since, in some cases *softImpute*, *imputePCA* and *SVDmiss* impute values outside the $[0,1]$ range of β-values, we consider also the performances of the three methods after a post-imputation truncation of the overflowed values. As shown in Supplementary Information 2, the accuracy of the three methods (*SVDmiss*^{T}, *softImpute*^{T} and *imputePCA*^{T}) is not significantly affected by out-of-range imputations.

From Table 1 and Table 2, we can notice that the performances of most methods do not decrease with increasing percentages of missing markers, likely owing to the high intercorrelation among methylated regions of the genome. On the other end, independently from the specific approach, imputations are less accurate on disease-related samples. In particular, the mean absolute error per site is more than doubled in datasets related to cancer (see Supplementary Information 2), the most common disease type in our benchmark set. This is somewhat expected, due to the known heterogeneity of tumors and the associated higher inter-sample methylation variability (Lomberk *et al*., 2018; Klughammer *et al*., 2018).

In Fig 1, we can notice that absolute errors are not equally distributed over the range of β-values, being smaller at the extremes. Although Fig. 1 shows performances on healthy samples only, this general trend is clearly visible also for the disease-related samples and in each single tissue-specific dataset (see Supplementary Information 2), as well as for all metrics but MAPE (which is, not surprisingly, higher for smaller β-values). This behaviour can be explained by the already discussed heteroscedasticity of β-values (Du *et al*., 2010), i.e. the standard deviations of β-values are compressed in the low and high ranges. The error distribution of *mean* imputations in Fig. 1 are a clear evidence of such variability. Furthermore, this behaviour corresponds to the strength of the biological signal, as 1 and 0 correspond to situations where all or none of the cells in the tissue under study are methylated, indicating a very robust biological condition, easier to predict than a situation where only a proportion of the cells presents a given methylation pattern (intermediate values). Unexpectedly, *impute.knn* is the only method for which the errors increase with increasing range of the β-values. This seems to be particularly evident in the smaller datasets, independently from the specific tissue (see Supplementary Information 2). This behaviour affects the overall performances of the *impute.knn* method, which, especially on control samples, behaves slightly worse than the baseline *mean* method.

In general terms, the imputation accuracy of our *methyLImp* linear model is higher in comparison to the other methods and with respect to all selected metrics. In order to test whether the performances of *methyLImp* are statistically significantly higher than those of the other methods, we use the one-tailed Wilcoxon paired rank-sum test with Benjamini-Hochberg multiple test correction. In Tables 1 and 2 we indicate with the † symbol only those cases for which the null hypothesis of having the same accuracy could not be rejected (p-value$\geq$0.05). In summary, all metrics show that *methyLImp* behaves statistically significantly better than the other methods (p-value<0.05) for all the percentages of missing value, with the

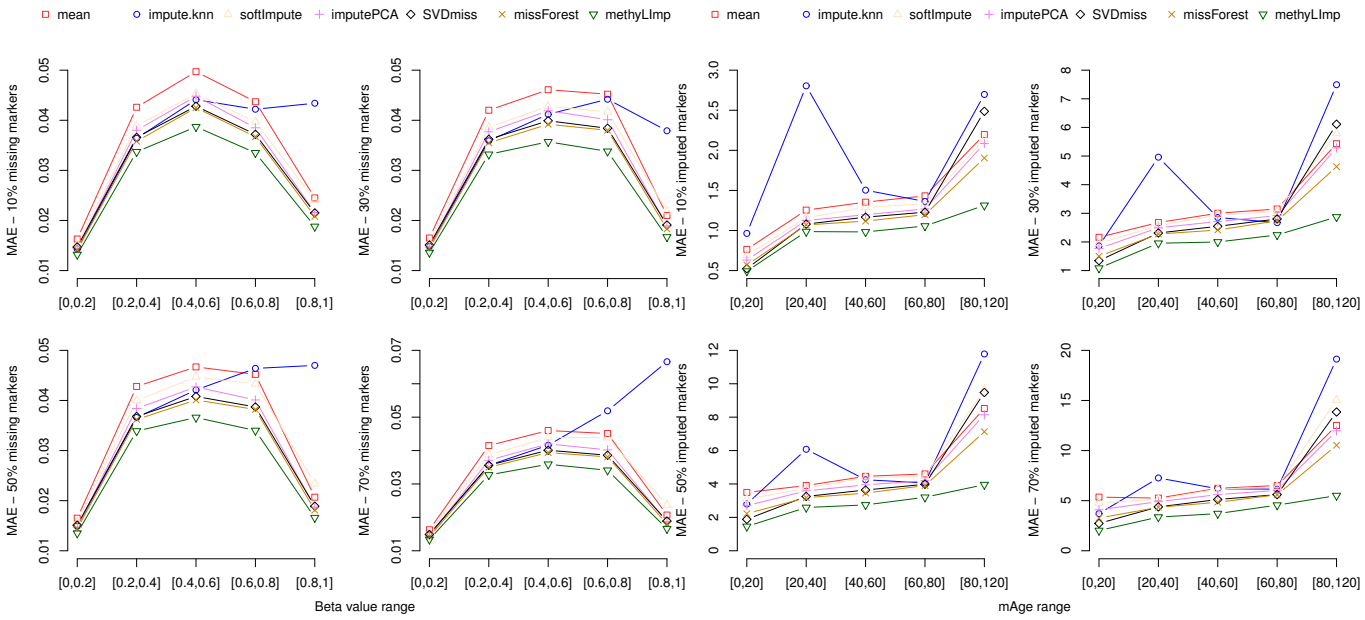

Fig. 1. Global CpG imputation performances in terms of MAE and with respect to β-value range. Healthy control samples.

Fig. 2. Global mAge accuracy from imputed values in terms of MAE and with respect to mAge range. Healthy control samples.

exception of the MAPE in the case of 10% missing values, for which the null hypothesis of *methyLImp* having the same MAPE than *impute.knn*, *imputePCA* and *missForest* could not be rejected.

Finally, recall that, in each test we artificially introduce a maximum of 247 missing values on a single sample. In order to stress the imputation capabilities of the benchmarked methods, we replicated the same experiments for much larger sets of missing values (see Supplementary Informations 3). On such larger sets the imputation accuracies remain unaffected for all methods but *softImpute* and *impute.knn*, which show considerably lower performances.

4.2 DNA methylation age estimation from imputed data

We compare here the performances of the imputation methods in terms of the impact they have on mAge estimation.

In first place, we analyze how much the mAge estimation diverges from the expected measure at increasing percentage of imputed values. This represents the final aim of our work, i.e. identification of the appropriateness of imputing methylation values in order to gain insight into the mAge value. The overall performances on healthy and disease samples are shown in Tables 3 and 4, respectively. As expected, the performances decrease uniformly for all methods at increasing percentages of imputed values. Furthermore, the deviations are higher for disease samples than for control samples, mirroring the known clinical and biological difficulties in capturing regularities among the heterogeneity of tumor samples. This is consistent with the results in Tables 1 and 2, where imputation accuracy is shown to be considerably lower in disease-related datasets. Overall, *methyLImp* imputed values induce smaller mAge approximation errors in comparison to all other methods. In particular, *methyLImp* performances are almost everywhere significantly better than those of the other methods (Wilcoxon paired rank test at significance level < 0.05 with Benjamini-Hochberg multiple test correction), except for very few cases for which there is no significant difference (indicated with † in the tables).

In Fig. 2, we can observe how the absolute errors are distributed with respect to the mAge ranges in control samples. In short, imputation errors seem to affect more the mAge evaluation in elderly than in young individuals. Interestingly, as previously reported (Horvath, 2013; Horvath *et al.*, 2015), the mAge estimation for such individuals is not very accurate also with complete data. Approximatively the same trend can be observed in disease case samples as well (see Supplementary Information 2). In Fig. 2 we can also notice that *impute.knn* imputations tend to heavily affect the mAge estimation in the range 20-40. This may be the consequence of the inaccurate imputations of high range β-values, as shown in Fig. 1. However, this behaviour is not observed in disease-related samples (see Supplementary Information 2).

In order to assess how reliable the mAge evaluated from imputed values is, we measured the concordance between mAge and chronological age in healthy samples. We do not consider here the predictive accuracy in disease-related samples since, in this setting, mAge is expected to indicate a deviation from chronological age. The most commonly adopted accuracy measure between mAge and chronological age is the Pearson correlation coefficient, usually denoted as *age correlation*. The age correlation of Horvath's clock for the original 1495 healthy control samples is 0.914 (p-value $< 10^{-15}$). Table 5 shows the age correlation of the Horvath's clock with respect to the percentage of values imputed by the benchmarked methods. The age correlations in Table 5 have been averaged over the ten random tests. Although all correlations in Table 5 are statistically significant (pvalue $< 10^{-15}$), a statistical test of the difference between paired correlations with Benjamini-Hochberg multiple test correction (see Supplementary Information 2) shows that with 70% imputed values the age correlations deviates significantly from the original age correlation. On the other end, up to 30% imputed values do not statistically significantly affect Horvath's clock mAge estimation, almost independently of the specific imputation method. Among all methods, the imputations of *methyLImp* and *missForest* induce smaller deviations in the age correlation, consistently with their performances shown in Tables 1 and 3.

4.3 Computational efficiency

The average memory requirements and running times over the 75k runs performed with each imputation method are shown in Table 6. The fastest methods are *mean*, *impute.knn*, *softImpute* and *imputePCA*, while

	10% missing markers				30% missing markers				50% missing markers				70% missing markers			
Method	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE
mean	0.043	0.026	**0.99**	20.8	0.043	0.026	**0.99**	19.4	0.043	0.026	**0.99**	20.4	0.042	0.026	**0.99**	19.9
impute.knn	0.057	0.027	0.98	16.8	0.055	0.026	0.98	17.0	0.063	0.028	0.98	18.2	0.074	0.030	0.97	18.3
softImpute	0.042	0.024	**0.99**	19.5	0.043	0.025	**0.99**	18.3	0.049	0.025	**0.99**	19.5	0.048	0.025	**0.99**	19.0
imputePCA	0.039	0.023	**0.99**	18.3	0.039	0.024	**0.99**	17.3	0.039	0.023	**0.99**	18.3	0.039	0.023	**0.99**	17.9
SVDmiss	0.039	0.023	**0.99**	18.7	0.040	0.023	**0.99**	17.6	0.040	0.023	**0.99**	18.6	0.039	0.023	**0.99**	18.1
missForest	0.037	0.022	**0.99**	19.2	0.037	0.022	**0.99**	17.4	0.038	0.022	**0.99**	18.7	0.037	0.022	**0.99**	18.1
methyLImp	**0.035**	**0.021**	0.99	**15.1**	**0.035**	**0.021**	0.99	**14.7**	**0.035**	**0.021**	0.99	**15.0**	**0.035**	**0.021**	0.99	**14.9**

Table 1. Global CpG imputation performances on 1495 **healthy control samples** with respect to percentage of missing markers. The results have been averaged over the 10 random replicas per percentage of missing markers. Standard errors (see Supplementary Information 2) are in the order of magnitude 10^{-4} for RMSE, MAE, PCC and on the order of 10^{-1} for MAPE. Best results per metric are highlighted in bold. Results for which the Wilcoxon paired rank-sum test p-value is ≥ 0.05 are indicated with the † symbol.

	10% missing markers				30% missing markers				50% missing markers				70% missing markers			
Method	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE
mean	0.090	0.053	0.95	33.3	0.088	0.052	0.95	32.3	0.089	0.052	0.95	32.2	0.089	0.052	0.95	32.7
impute.knn	0.084	0.047	0.96	**25.3**†	0.084	0.047	0.96	25.2	0.088	0.048	0.96	26.0	0.092	0.049	0.95	26.5
softImpute	0.087	0.049	0.95	29.1	0.085	0.048	0.96	28.8	0.087	0.049	0.96	29.1	0.089	0.049	0.95	29.5
imputePCA	0.080	0.046	0.96	27.9†	0.079	0.045	0.96	27.4	0.079	0.046	0.96	27.5	0.079	0.045	0.96	27.7
SVDmiss	0.084	0.046	0.96	29.0	0.081	0.045	0.96	28.6	0.082	0.046	0.96	28.6	0.081	0.045	0.96	28.8
missForest	0.075	0.044	0.96	27.3†	0.074	0.043	**0.97**	26.9	0.075	0.043	**0.97**	27.2	0.075	0.043	**0.97**	27.3
methyLImp	**0.071**	**0.039**	**0.97**	25.4	**0.070**	**0.038**	**0.97**	**23.4**	**0.071**	**0.039**	**0.97**	**24.4**	**0.070**	**0.038**	**0.97**	**24.5**

Table 2. Global CpG imputation performances on 386 **disease case samples** with respect to percentage of missing markers. The results have been averaged over the 10 random replicas per percentage of missing markers. Standard errors (see Supplementary Information 2) are in the order of magnitude 10^{-4} for RMSE, MAE, PCC and on the order of 10^{-1} for MAPE. Best results per metric are highlighted in bold. Results for which the Wilcoxon paired rank-sum test p-value is ≥ 0.05 are indicated with the † symbol.

	10% imputed markers				30% imputed markers				50% imputed markers				70% imputed markers			
Method	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE
mean	1.8	1.3	**0.99**	4.3	3.8	2.9	0.98	10.8	5.6	4.3	0.96	17.4	7.9	6.0	0.92	26.6
impute.knn	2.7	1.8	**0.99**	5.8	4.7	3.4	0.98	11.1	6.4	4.7	0.95	15.4	8.8	6.4	0.91	20.5
softImpute	1.7	1.2	**0.99**	3.7	3.7	2.8	0.98	9.3	5.6	4.1	0.96	14.6	8.0	5.8	0.92	22.0
imputePCA	1.6	1.1	**0.99**	3.6	3.5	2.6	**0.99**	8.8	5.0	3.8	0.97	13.4	7.2	5.4	0.94	19.9
SVDmiss	1.6	1.1	**0.99**†	3.2	3.5	2.4	0.98	7.1	5.0	3.4	0.97	10.0	7.0	4.8	0.94	14.0
missForest	1.5†	1.1	**0.99**†	3.3	3.2	2.3	**0.99**	7.9	4.5	3.4	**0.98**	11.8	6.3	4.7	0.95	17.4
methyLImp	**1.4**	**0.9**	**0.99**	**2.8**	**2.8**	**1.9**	**0.99**	**5.7**	**3.8**	**2.6**	**0.98**	**7.8**	**5.1**	**3.6**	**0.97**	**10.5**

Table 3. Global mAge accuracy from imputed values on 1495 **healthy control samples** with respect to percentage of imputed markers. The results have been averaged over the 10 random replicas per percentage of imputed markers. Standard errors (see Supplementary Information 2) are in the order of magnitude 10^{-1} for RMSE, MAE, MAPE and on the order of 10^{-4} for PCC. Best results per metric are highlighted in bold. Results for which the Wilcoxon paired rank-sum test p-value is ≥ 0.05 are indicated with the † symbol

	10% imputed markers				30% imputed markers				50% imputed markers				70% imputed markers			
Method	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE	RMSE	MAE	PCC	MAPE
mean	3.8	2.8	**0.99**	6.4	7.1	5.2	0.96	19.0	9.3	7.0	0.93	42.2	12.5	9.3	0.86	79.1
impute.knn	3.5	2.4	**0.99**	5.5†	6.3	4.5	**0.97**	13.4†	8.8	6.3	0.93	23.3†	12.3	8.8	0.87	**36.2**†
softImpute	3.9	2.6	**0.99**	6.1	7.1	4.9	0.96	17.0	9.4	6.7	0.92	35.1	12.8	9.1	0.85	66.8
imputePCA	3.5	2.5	**0.99**	5.8	6.4	4.6	**0.97**	14.2	8.7	6.4	0.94	26.5	11.8	8.8	0.88	44.7
SVDmiss	3.8	2.5	**0.99**	6.1	6.5	4.6	0.96	17.7	9.0	6.5	0.93	39.1	12.0	8.9	0.87	74.1
missForest	3.2	2.3	**0.99**	5.7	5.9	4.3	**0.97**	16.7	8.0	5.9	**0.95**	35.8	11.0	8.1	0.89	69.2
methyLImp	**3.1**	**2.1**	0.99	**5.1**	**5.5**	**3.9**	**0.97**	**12.1**	**7.4**	**5.3**	**0.95**	**22.1**	**10.1**	**7.3**	**0.91**	39.3

Table 4. Global mAge accuracy from imputed values on 386 **disease case samples** with respect to percentage of imputed markers. The results have been averaged over the 10 random replicas per percentage of imputed markers. Standard errors (see Supplementary Information 2) are in the order of magnitude 10^{-1} for RMSE, MAE, MAPE and on the order of 10^{-4} for PCC. Best results per metric are highlighted in bold. Results for which the Wilcoxon paired rank-sum test p-value is ≥ 0.05 are indicated with the † symbol

missForest is the slowest, requiring the equivalent of 2 years of total calculation on a single processor machine. In terms of extra memory usage (other than the dataset memory requirements), *SVDmiss* is the most demanding method. *SVDmiss* is virtually unusable on large datasets with full 450k CpG sites. In comparison to the other methods *methyLImp* is reasonably efficient in terms of memory requirements and acceptable in terms of running time. A significant speed-up in computation time can be gained by using the parallelized version of the algorithm, at the cost of higher memory requirements.

Method	Avg time (sec)	Avg RAM (Mb)
mean	< 1	63
impute.knn	3	259
softImpute	< 1	238
imputePCA	20	450
SVDmiss	160	4072
missForest	879	426
methyLImp	185	281

Table 6. Average time and memory usage over 75,240 runs.

5 Conclusion

We have designed a novel imputation method, *methyLImp*, for DNA methylation data. Our model implements a simple linear regression approach that exploits the inter-samples correlation descending from the biological nature of the methylated DNA. Imputation performances have been assessed on a variety of samples (healthy and disease) under increasing stress (missing values) and by a variety of accuracy metrics. In comparison to existing methods, our linear model provides a reasonably fast and accurate way to impute missing values in DNA methylation data. Given the relevance of methylation data in the study of the mechanism of action of a variety of biological functions, our approach significantly contributes to this area, allowing the research to be taken forward even in case of imperfect or clearly impaired datasets. However, the results of our tests indicate that it is important to take cautions when drawing conclusions from biological analyses including partially imputed methylation values. In particular, independently from the specific method, imputation accuracy is significantly lower in disease than in normal samples, mirroring the well-known biological and clinical issue associated to the heterogeneity of tumor samples. Similarly, data prediction is more reliable for low and high range methylation levels (i.e., β-values close to 0 and 1), again mirroring the biological event where the signal is coherent for all cells in the tissue (all methylated or non-methylated) versus intermediate situations. Finally, special care needs to be taken when using DNA methylation-based clocks with imputed values, since even few poorly imputed values may severely affect the mAge estimation.

Funding

This work has been partially supported by MIUR's FFABR 2017 (Fondo per il Finanziamento delle Attività di Base di Ricerca)

References

Bennett, D.A. (2001) How can I deal with missing data in my study? Aust N Z J Public Health. 25, 464-469.

Bibikova, M. *et al*. (2006) High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.*, **16**, 383-393.

Ciabattini, A. *et al*. (2018) Vaccination in the Elderly: The Challenge of Immune Changes with Aging. *Seminars in Immunology*, **40**, 83-94

Dempster, A.P., Laird N.M., Rubin D.B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-38.

Donders, A.R. *et al*. (2006) Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, **59**, 1087-1091.

Du, P. *et al*. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, **11**:587.

Durrant, G.B. (2009) Imputation methods for handling item-nonresponse in practice: methodological issues and recent debates. *International Journal of Social Research Methodology* , **12**, 293-304.

Durso, D.F. *et al*. (2017) Acceleration of leukocytes' epigenetic age as an early tumor and sex-specific marker of breast and colorectal cancer. *Oncotarget*, **8**, 23237-23245.

Edgar, R., Domrachev, M, Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207-210.

Enders, C.K. (2010) *Applied Missing Data Analysis.* Guilford Press, New York.

Fuentes, M. , Guttorp, P. , Sampson P.D. (2006) Using Transforms to Analyze Space-Time Processes. In Statistical methods for spatio-temporal systems. In *Statistical Methods for Spatio-Temporal Systems*, pp. 77-150.

Garagnani, P. *et al*., (2012) Methylation of ELOVL2 gene as a new epigenetic marker of age. *Aging Cell.*, **11**, 1132-1134.

Hannum, G *et al*. (2013) Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.*, **49**, 359-367.

Horvath, S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol.*, **14**:R115.

Horvath, S. *et al*. (2015) Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. *Aging*, **7**, 1159-1170.

Horvath, S., Raj, K. (2018) DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet.*, **19**, 371-384.

Little, R.J.A., Rubin D.B (1986) Statistical Analysis with Missing Data. John Wiley & Sons, Inc., New York.

Lomberk, G. *et al*. (2018) Distinct epigenetic landscapes underlie the pathobiology of pancreatic cancer subtypes. *Nat Commun.*, **9**: 1978.

Lövkvist, C. *et al*., (2016) DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Res.*, **44**, 5123-5132.

Josse, J., Husson, F. (2013). Handling missing values in exploratory multivariate data analysis methods. *Journal de la SFdS*, **153**, 79-99.

Klughammer, J. *et al*. (2018) The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat Med.*, **24**, 1611-1624.

Mazumder, R., Hastie, T., Tibshirani, R. (2010) Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.*, **11**, 2287-2322.

Nardini, C. *et al*. (2018) The Epigenetics of Inflammaging – Heterochromatin Loss, Gene-Specific Remodelling, Environmental Stimuli. *Seminars in Immunology* , **40**, 49-60.

Pidsley R. *et al*. (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, **14**:293.

Severson, K.A. *et al*. (2017) A method for learning a sparse classifier in the presence of missing data for high-dimensional biological datasets. *Bioinformatics*, **33**, 2897-2905.

Stekhoven, D.J., Bülmann, P. (2012) MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112-118.

Troyanskaya, *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520-525.

Weidner, C.I. *et al*., (2014) Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology*, **15**:R24.

Wu, C. *et al*. (2016) Imputation of missing covariate values in epigenome-wide analysis of DNA methylation data. *Epigenetics*, **11**, 132-139.

Zhang, G. *et al*. (2016) Across-Platform Imputation of DNA Methylation Levels Incorporating Nonlocal Information Using Penalized Functional Regression. *Genet Epidemiol.*, **40**. 333-340.

Zhang, L. *et al*. (2017) DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells. *Biophys J.*, **113**, 1395-1404.

	Percentage of imputed markers			
Method	10%	30%	50%	70%
mean	0.915	0.913	0.902**	0.877****
impute.knn	0.912	0.911	0.896***	0.893***
softImpute	0.915	0.912	0.900**	0.875****
imputePCA	0.915	0.911	0.902**	0.880****
SVDmiss	0.914	0.909*	0.900***	0.879****
missForest	0.915	0.914	0.909	0.895***
methyLImp	0.914	0.913	0.909	0.900***

Table 5. Pearson correlation coefficient (age correlation) between mAge (predicted age) and chronological age on 1495 healthy control samples, with respect to the percentage of imputed values. Significance levels for the difference between paired correlations, original correlation (0.914) and post-imputation-correlation, are encoded as: * $< 10^{-2}$, ** $< 10^{-4}$, *** $< 10^{-6}$, **** $< 10^{-8}$.