RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Across-Platform Imputation of DNA Methylation Levels Incorporating Nonlocal Information Using Penalized Functional Regression

Guosheng Zhang,[1,2,3] Kuan-Chieh Huang,[4] Zheng Xu,[1,4,5] Jung-Ying Tzeng,[6] Karen N. Conneely,[7] Weihua Guan,[8] Jian Kang,[9] and Yun Li[1,2,4,5]*

[1]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina, United States of America; [2]Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, North Carolina, United States of America; [3]Department of Statistics, University of North Carolina, Chapel Hill, North Carolina, United States of America; [4]Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, United States of America; [5]Department of Computer Science, University of North Carolina, Chapel Hill, North Carolina, United States of America; [6]Department of Statistics, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America; [7]Department of Human Genetics, School of Medicine, Emory University, Atlanta, Georgia, United States of America; [8]Division of Biostatistics, School of Public Health, University of Minnesota, Minnesota, United States of America; [9]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States of America

Received 17 September 2015; Revised 3 February 2016; accepted revised manuscript 18 February 2016.
Published online 7 April 2016 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21969

ABSTRACT: DNA methylation is a key epigenetic mark involved in both normal development and disease progression. Recent advances in high-throughput technologies have enabled genome-wide profiling of DNA methylation. However, DNA methylation profiling often employs different designs and platforms with varying resolution, which hinders joint analysis of methylation data from multiple platforms. In this study, we propose a penalized functional regression model to impute missing methylation data. By incorporating functional predictors, our model utilizes information from nonlocal probes to improve imputation quality. Here, we compared the performance of our functional model to linear regression and the best single probe surrogate in real data and via simulations. Specifically, we applied different imputation approaches to an acute myeloid leukemia dataset consisting of 194 samples and our method showed higher imputation accuracy, manifested, for example, by a 94% relative increase in information content and up to 86% more CpG sites passing post-imputation filtering. Our simulated association study further demonstrated that our method substantially improves the statistical power to identify trait-associated methylation loci. These findings indicate that the penalized functional regression model is a convenient and valuable imputation tool for methylation data, and it can boost statistical power in downstream epigenome-wide association study (EWAS).

Genet Epidemiol 40:333–340, 2016. © 2016 Wiley Periodicals, Inc.

KEY WORDS: DNA methylation; imputation; penalized functional regression; epigenome-wide association study

Introduction

DNA methylation is an important epigenetic modification involved not only in normal development [Smith and Meissner, 2013], but also in risk and progression to many diseases [Bergman and Cedar, 2013]. It has been shown to play a key role in the regulation of gene transcription, X-inactivation, cellular differentiation, and other critical processes such as aging [Bird, 2002; Gonzalo, 2010]. Recently, the emergence of powerful technologies such as microarray-based DNA methylation studies [Bibikova et al., 2011] and whole-genome bisulfite sequencing [Harris et al., 2010] has enabled the profiling of DNA methylation levels at high resolution. Numerous studies employed these high-throughput approaches to characterize changes in DNA methylation patterns and their corresponding tissue- and disease-specific differentially methylated regions on a genome-wide scale [Berman et al., 2012; Chen, Ning, Hong, & Wang, 2014; Horvath, 2013; Varley et al., 2013].

As new technologies emerge, researchers tend to replace older methylation profiling platforms with new ones. However, different platforms can target CpG sites at different locations and with varying resolutions, which hinders the joint analysis of data from multiple platforms. For instance, the Illumina HumanMethylation27 (HM27) and HumanMethylation450 (HM450) BeadChip [Bibikova et al., 2011] are two common microarrays used by The Cancer Genome Atlas (TCGA) project. HM27 investigates 27,578 CpG sites predominantly located near CpG islands, while HM450 provides broader coverage with 485,577 probes spanning 96% of CpG islands and 92% of CpG shores across a larger number of genes [Bibikova et al., 2011]. Several TCGA studies have

Supporting Information is available in the online issue at wileyonlinelibrary.com.

*Correspondence to: Yun Li, Department of Genetics, Campus Box 7264, University of North Carolina, Chapel Hill, NC 27599. E-mail: yunli@med.unc.edu; Or Jian Kang, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109; jiankang@umich.edu

used HM450 to generate methylation profile data for more recently collected samples while still using HM27 to measure DNA methylation in the older test subjects. These mixed profiles compel researchers to focus on those probes shared between the two platforms when using the data for downstream analysis, as reevaluating all samples using HM450 is not only expensive, but also time-consuming [Getz et al., 2013; Koboldt et al., 2012; The Cancer Genome Atlas Research Network, 2012, 2013].

Imputation has been successfully employed in many genetic, genomic, and epigenomic contexts [Donner et al., 2012; Ernst and Kellis, 2015; Jewett et al., 2012; Li et al., 2009; Zhang et al., 2015]. For methylation profiling, multiple methods have been proposed to impute methylation levels across tissue types [Ma et al., 2014] or employing various genomic and epigenomic features, including DNA sequence context, genomic position, predicted DNA structure, GC content, and DNA regulatory elements [Bock et al., 2006; Das et al., 2006; Zhang et al., 2015]. However, most of these methods dichotomize methylation status. More importantly, no cross-platform imputation methods have been proposed for predicting methylation levels at unassayed CpG sites. On the other hand, for genotypes, imputation of untyped SNPs has become a standard procedure used both to resolve similar inconsistencies between genotyping arrays and to increase the resolution of genotype data collected in genome-wide association studies [Li et al., 2009]. Here, we propose the application of a similar concept to impute data in DNA methylation profiles from a subset of probes. Although DNA methylation does not exhibit as clear or strong a correlation structure as LD blocks among SNPs, we observe local correlation among neighboring probes similar as reported by others [Eckhardt et al., 2006; Zhang et al., 2015]. Importantly, we have found nonlocal correlations among probes falling into the same functional categories that have not been employed in the literature. Therefore, we adopt a penalized functional regression model [Goldsmith et al., 2011], which uses functional predictors to capture these nonlocal correlations. Our study demonstrates that this model can impute an HM27 dataset into an HM450 dataset effectively and accurately, and using these imputed values can improve the statistical power of downstream epigenome-wide association study (EWAS).

Materials and Methods

Data

We evaluated our imputation model using DNA methylation data from TCGA acute myeloid leukemia (AML) samples [Ley et al., 2013]. The dataset contains DNA methylation data of tumor tissues from 194 patients with AML and is one of the largest methylation datasets from the TCGA project. All samples were evaluated using both HM27 and HM450. We transformed the raw β values into M values, defined as $M = log_2[\beta/(1-\beta)]$, as the M values better follow a Gaussian distribution [The Cancer Genome Atlas Research Network, 2013]. Our goal is to impute the HM27 dataset into an HM450 dataset to get an expanded view of the epigenomic landscape. The dataset is publicly available at the TCGA data portal (https://tcga-data.nci.nih.gov/tcga/).

Because imputation of sporadic missing data is not the focus of this work, we removed all probes with at least one missing values for the sake of convenience. However, these missing values can be imputed by applying similar methods developed for gene expression profiles [Bo, Dysvik, & Jonassen, 2004; Kim et al., 2005; Liew et al., 2011; Troyanskaya et al., 2001] to generate data without missing values. Additionally, we removed 743 probes designed in HM27 but not in HM450. In total, the HM27 dataset consisted of 20,794 probes passing TCGA quality control (QC) criteria [Ley et al., 2013] and the HM450 dataset consisted of 393,152 QC+ probes. The latter set contained all 20,794 probes in HM27, leaving the remaining 373,358 as our potential imputation targets.

When training and using our model, we required data from HM450 and HM27, respectively. However, we noted that as HM27 and HM450 employ different biochemical methods to measure methylation levels, platform-specific effects might negatively impact imputation performance. To alleviate this systematic effect, we fitted a LOESS (locally weighted scatterplot smoothing) regression model [Cleveland, 1979] between two platforms, stratified by the number of CpGs in the probe (#CpG = 0, 1, 2, 3, 4, 5, 6, 7+), using 14 randomly chosen samples and normalized the HM27 data against the HM450 data [The Cancer Genome Atlas Research Network, 2013].

Penalized Functional Regression Model

We employed the penalized functional regression model [Goldsmith et al., 2011] with minor modifications detailed below to quantify the relationship between DNA methylation from HM450 probes and the DNA methylation density function estimated from HM27 probes together with other covariates. Specifically, assume for each target HM450 probe, we have n observations and for each sample $i = 1, 2, \ldots, n$, we have data $[Y_i, X_i(t), Z_i]$, where Y_i is the transformed DNA methylation level at the target HM450 probe, $X_i(t)$ is the sample-specific density function of the DNA methylation level measured by HM27 probes, denoted as T_i, and Z_i is a p-dimensional vector of covariates. We consider a functional linear regression model:

$$Y_i = \alpha + \int_0^1 X_i(t)\beta(t)dt + Z_i\gamma + \varepsilon_i.$$

Here, α is the overall mean, $\beta(t)$ is the functional coefficient that characterizes the effect of density function $X_i(t)$ when $T_i = t$, γ is the regression coefficient vector for covariates, and $\varepsilon_i \sim N(0, \sigma^2)$.

To improve imputation accuracy, we incorporated functional predictors $X_i(t)$ into our model to capture information such as nonlinear relationships from nonlocal probes. Based on the assumption that probes with similar properties tend to show similar methylation profiles, we divided the probes into several property groups. Here, we divided the probes among five groups according to their relative location to a CpG island. The five groups are "CpG Island," "North Shore," "South Shore," "North Shelf," and "South Shelf" [Bibikova et al., 2011]. Then, we estimated the DNA methylation function $X_i(t)$ for a particular target probe with the DNA methylation data from HM27 probes in the same group as the target probe. Assume the target probe is in group g and there are q HM27 probes in the same group. The observed DNA methylation data are denoted as $\tau_i^g = (t_1^g, \ldots, t_q^g)$, where t_j^g is the DNA methylation value at jth HM27 probe in group g and $j = 1, \ldots, q$. Instead of estimating $X_i(t)$ by expanding into the principal component basis obtained from its covariance matrix [Goldsmith et al., 2011], we used the kernel density estimation to obtain $X_i(t)$ with τ_i^g so that it is specific to group g.

To perform the model fitting, the functional coefficient $\beta(t)$ was expanded by a linear spline basis $\beta(t) = b_1 + b_2 t + \sum_{k=3}^{K_b} b_k (t - \delta_k)_+$, where δ_k is the knot along the interval [0,1] and $(t - \delta_k)_+$ is an indicator function, taking a value of 1 if $t > \delta_k$ and 0 if $t \leq \delta_k$. We further defined a spline basis vector $\varphi(t) = \{\varphi_1(t), \varphi_2(t), \ldots, \varphi_{K_b}(t)\} = \{1, t, (t - \delta_3)_+, \ldots, (t - \delta_{K_b})_+\}$ and a coefficient vector $b = (b_1, \ldots, b_{K_b})'$ so that we may induce smoothing by assuming $b \sim N(0,D)$, where D is a penalty matrix corresponding to the particular spline basis $\phi(t)$.

Finally, we had $\int_0^1 X_i(t)\beta(t)dt = \int_0^1 f_{T_i}(t)\phi(t)b\,dt = \int_0^1 f_{T_i}(t)\phi(t)dt \cdot b$. For ease of notation, we denoted $J_{X\phi}$ as the $n \times K_b$ matrix with the (i,k)th entry equal to $\int_0^1 f_{Ti}(t)\phi_k(t)dt$ and Z as the $n \times p$ matrix with the ith row equal to Z_i, where p is the number of covariates. The model can be written in matrix format as:

$$Y|X(t) = \left[1_n, J_{X\varphi}, Z\right]\left[\alpha', b', \gamma'\right]' + \varepsilon,$$

$$b \sim N(0,D).$$

This is a mixed effect model with K_b random effects b and penalty matrix:

$$D = \begin{bmatrix} 0_{2\times 2} & 0_{2\times(K_b-2)} \\ 0_{(K_b-2)\times 2} & I_{(K_b-2)\times(K_b-2)} \end{bmatrix}.$$

Typically, $K_b = 30$ is sufficient to avoid under-smoothing in most applications [Goldsmith et al., 2011]. Consistent with previous work [Fan et al., 2015a,2015bb], choice of K_b has little impact on performance (Supplementary Fig. S2).

Selection of Local Covariates

We exploited linear correlation with neighboring probes by including methylation values of HM27 probes near the target HM450 probe as local covariates Z in our imputation model. For simplicity, we selected the five nearest upstream probes and the five nearest downstream probes to each target probe as these local covariates.

Quality Filter

Because most probes showed nearly constant methylation levels across samples, we found for many probes, the imputation model is formed without sufficient information. Thus, it tends to be underfitted and yields inaccurate imputation results. It is therefore desirable to have quality metrics for gauging the imputation quality. As such a quality metric, we proposed an under-dispersion measure defined as the ratio of the variance of fitted methylation values to its expected value (the variance of the true methylation values in the training set). If this ratio is below a certain threshold for a probe, it indicates an underfitted model for that probe, and we discard imputed values for the probe before subsequent analysis. A more stringent threshold can provide more accurate results, although at the cost of more probes discarded after imputation.

Imputation Quality Assessment

We assessed imputation quality using fivefold cross-validation. Within each split, the full dataset was randomly divided into a training set consisting of 80% of the samples and a testing set comprised the remaining 20%. For each testing set, we only retained HM27 data that contain a subset of HM450 probes, and masked methylation values of other HM450-specific probes. For the training set, we used methylation measurements on probes shared between the two arrays as predictors to impute methylation values at HM450-specific probes. Because most HM27 probes were measured by both HM27 and HM450, the predictors used in our model can be methylation levels for these shared probes measured from either array. Note that our prediction model was built under the realistic (more challenging) scenario where we used as predictors the measurements from HM450 array instead of those from HM27 array, which would require the training dataset had measurements from both arrays. Specifically, we fitted the functional regression model based on the training set, learned the relationship between methylation values of the shared and HM450-specific probes, and used the fitted model to impute the masked values of HM450 probes from the HM27 data in the testing set. Finally, we evaluated the imputation performance by averaging quality measures across splits.

As quality measures, we selected the mean squared error (MSE) and the squared Pearson correlation (R^2) between the imputed and the true methylation values in the testing sets. Although R^2 is a more intuitive measure of quality directly related to power and sample size in downstream analysis, we would like to note that this metric could easily be affected by a few outliers. Additionally, if the variance of methylation values for a specific probe is small, R^2 can be dramatically affected even by small imputation errors.

Simulation of Association Study

To assess the potential improvement of statistical power when using well-imputed methylation values for epigenetic association studies, we performed several simulated association studies for continuous and binary traits. Specifically, we randomly selected 100 HM450 probes with imputation R^2 between 0.1 and 0.3 based on our functional

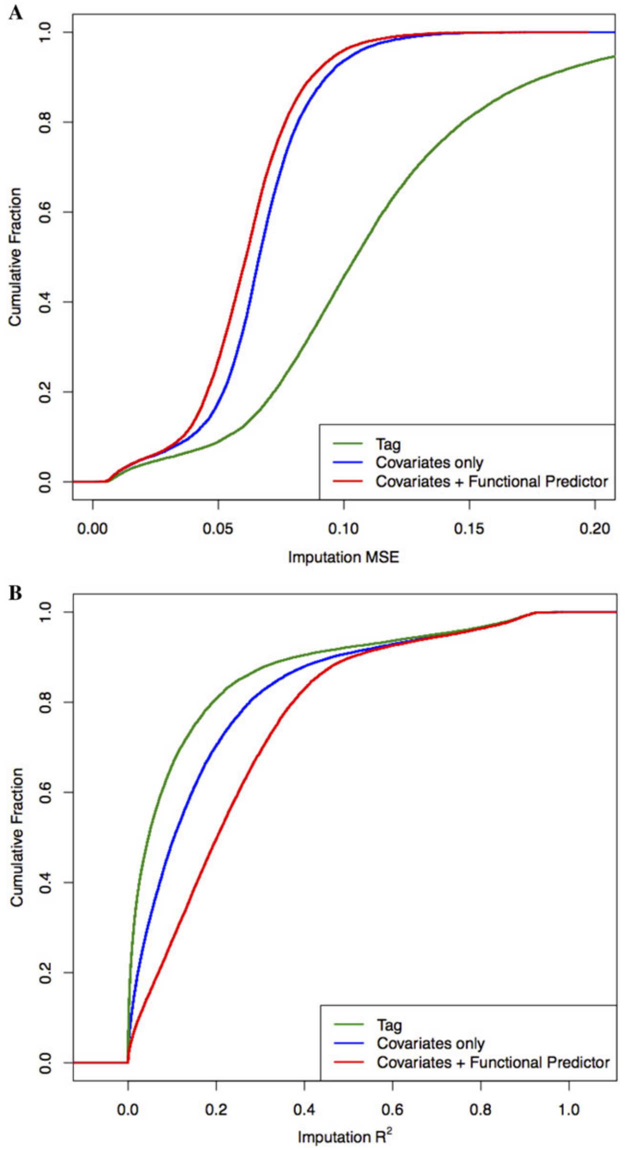

Figure 1. Empirical cumulative density function of (A) imputation MSE and (B) imputation R^2 for probes showing large variations in the AML dataset.

model, and simulated a dataset with 180 samples for each probe. In the continuous trait setting, for each probe, a trait value Y_i^* was simulated from the methylation level of this probe according to the linear model $Y_i^* = c\beta_i^* + \varepsilon_i$ for sample i, where β_i^* is true methylation β value, the effect size $c \in \{0, 0.1, 0.2, \ldots, 0.9, 1.0\}$, and $\varepsilon_i \sim N(0, 2s_{\beta_i^*})$, where $s_{\beta_i^*}$ is the sample standard deviation of β_i^*. In the binary trait setting, we first calculated $\eta_i^* = c(\beta_i^* - \bar{\beta}^*)$, $p_i^* = \frac{e^{\eta_i^*}}{e^{\eta_i^*}+1}$, and simulated Y_i^* from Bernoulli(p_i^*), where $\bar{\beta}^*$ is the mean value of β_i^*, and the effect size $c \in \{0, 0.5, 1.0, \ldots, 4.5, 5.0\}$.

We repeated the simulation 2000 times. For each simulated dataset, we performed association tests (linear regression for the continuous trait, and logistic regression for the binary trait) based on the true methylation values, as well as imputed values from the simple linear model and our proposed penalized functional model. The empirical power of each method was calculated as the proportion of observed P values that fall below the significance threshold, $\alpha = 0.05$. Finally, we evaluated the empirical power for each effect size c by averaging results across 100 probes.

Table 1. Quantiles of imputation MSE and R^2

	Imputation MSE			Imputation R^2		
	Q1	Median	Q3	Q1	Median	Q3
Covariates only	0.0553	0.0662	0.0781	0.0326	0.1040	0.2321
Covariates + functional predictor	0.0489	0.0610	0.0731	0.0907	0.2015	0.3375
Improvement	12%	8%	6%	178%	94%	45%

Results

Evaluation of Imputation Quality

Most probes showed nearly constant methylation levels in populations, making imputation trivial for them. We therefore focused on probes showing large variations and chose the top 20,000 such probes to evaluate the imputation quality. The time complexity of our method increases linearly with the number of target probes. However, since the imputation for each target probe is independent, we can accelerate it by running imputation in parallel. In the fivefold cross-validation experiment, 14 samples used for normalization were removed at first. Among the remaining 180, 144 individuals were chosen at random as the training set and 36 as the testing set within each split. The empirical cumulative distribution of imputation MSE and R^2 are shown in Figure 1. The baseline method we used is the "tag" approach, where for each target probe, we calculated the Euclidean distance between the target probe and local probes, chose the local probe with the smallest distance as the tag probe, and directly copied its methylation values as imputed values for the target probe. We also compared the two models with and without functional predictors and found that incorporating functional predictors lead to significantly improved imputation MSE and R^2 ($P < 2.2 \times 10^{-16}$ for both metrics, paired Wilcoxon test). Table 1 summarizes some basic statistics. As expected, the "tag" method performs worst and we have therefore focused in subsequent text only the two models with and without functional predictors.

We used the target probe cg00288598 as an example to illustrate how the functional predictors improve the imputation quality. As shown in Figure 2A, the selected local probes showed much smaller variation than the target probe, leading to an underfitted linear regression model and thus low imputation quality. In contrast, the methylation profile of the target probe is strongly associated with the distribution of methylation levels from all HM27 probes in its assigned North Shelf group, as indicated in Figure 2B. Therefore, after the functional predictors are added, the model can utilize the information from these nonlocal probes, including probes

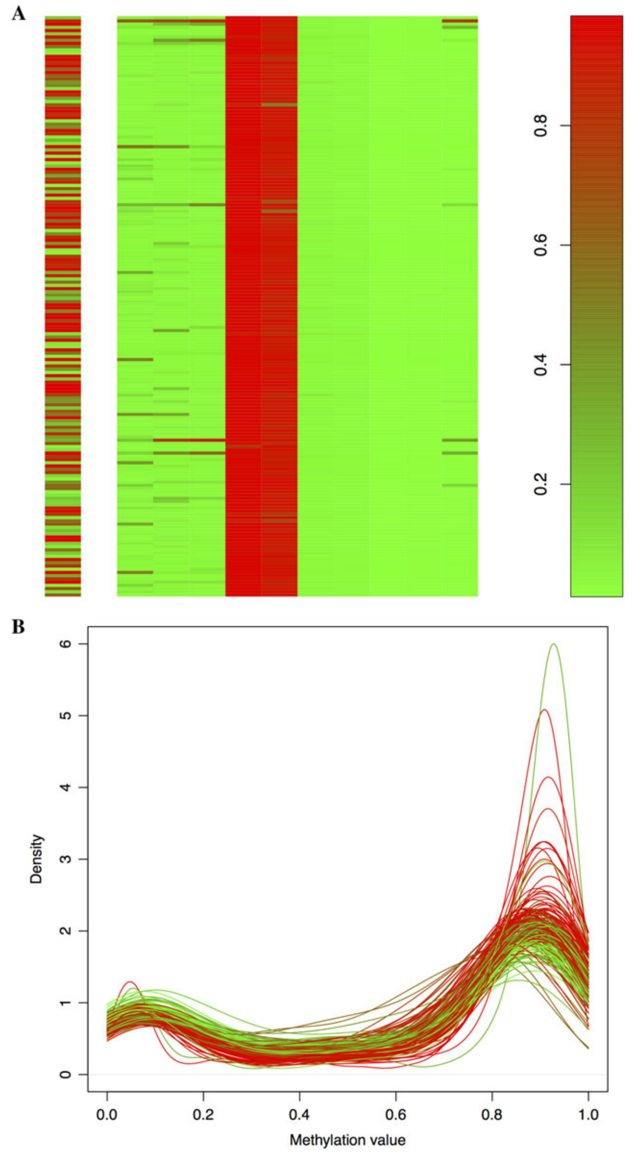

Figure 2. (A) Methylation profiles of a North Shelf probe cg00288598 (left) and 10 selected local probes (middle). (B) The individual-specific density plot of methylation values from all HM27 probes in North Shelf regions. Each line represents one individual and is colored based on the methylation level of the cg00288598 probe.

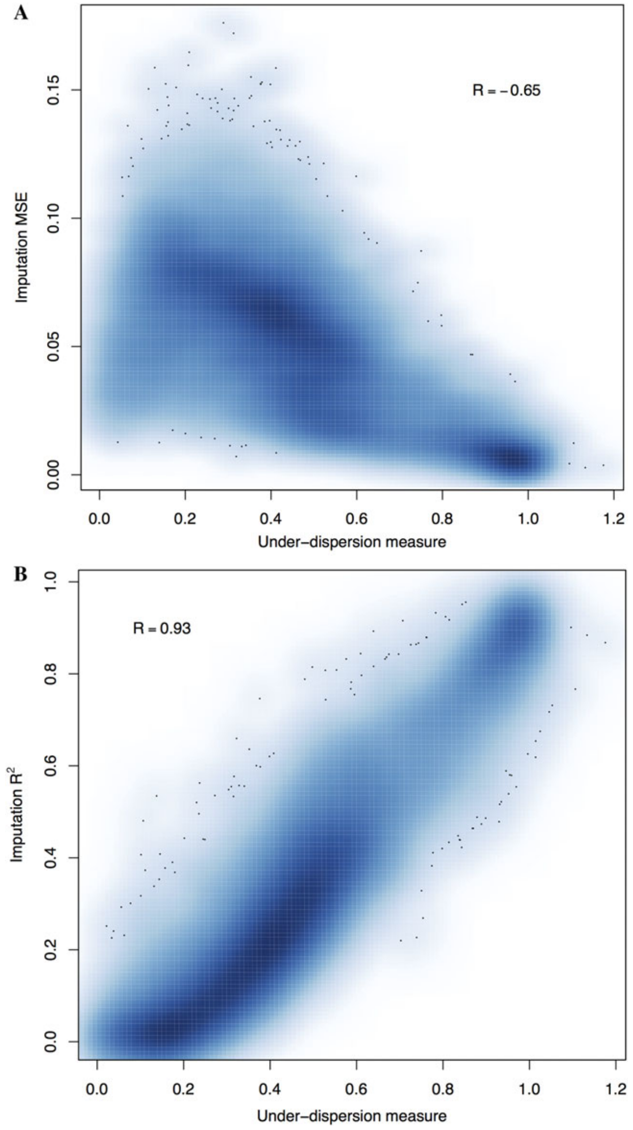

Figure 3. Scatter plot of under-dispersion measure and (A) imputation MSE and (B) imputation R^2.

on different chromosomes, to alleviate the underfitting problem.

Performance of Quality Metrics

Because not all target probes can be imputed with the same level of accuracy, we tried to use the under-dispersion measure described in the Methods section to filter out inaccurate imputation results. We examined the relationship between imputation MSE/R^2 and the under-dispersion measure. We observed a negative correlation between the imputation MSE and this quality measure (Fig. 3A, Pearson correlation coefficient, $R = -0.65$), and a positive correlation between imputation R^2 and the measure (Fig. 3B, Pearson correlation coefficient, $R = 0.93$). Therefore, when performing imputation, we can calculate the under-dispersion measure and use it to filter out low-quality imputation results. Figure 3 indicates that by choosing an appropriate threshold, we can remove most probes imputed with low-quality while simultaneously retaining nearly all probes imputed with high-quality. Based on our results, we suggest a threshold of 0.8 for the under-dispersion measure, which removes all badly imputed probes (defined as true $R^2 < 0.2$) at the cost of 1.24% well-imputed probes (true $R^2 > 0.8$). Table 2 shows the number of probes passing post-imputation quality filter at varying thresholds of the under-dispersion measure and we see that our penalized functional model results in up to 86.0% more probes that can be used for further analysis.

Table 2. Number of probes passing post-imputation quality filter

Under-dispersion measure threshold	0.6	0.7	0.8	0.9
Among top 20,000 probes				
Covariates only	2,113	1,592	1,174	681
Covariates + functional predictor	2,677	1,691	1,226	719
Improvement	26.7%	6.2%	4.4%	5.6%
Among all probes				
Covariates only	14,479	8,796	5,123	2,417
Covariates + functional predictor	26,924	13,117	6,526	2,684
Improvement	86.0%	49.1%	27.4%	11.1%

Power Gain in Association Study

It is not surprising to find relatively little difference in the performance of the two models at the two ends of the distribution (Fig. 1A and B) because of probes that are either trivial or impossible to impute. Therefore in our work, we focus on the ~34% probes with imputation R^2 between 0.1 and 0.3, where our model demonstrates advantages over simpler models. As shown in Figure 4, using imputed values from the penalized functional model for association tests is consistently more powerful than using values from the simple linear model, while the type I error rate (when $c = 0$) was still under proper control. These results suggest that even using probes with moderate imputation quality can substantially improve the statistical power of association test while maintaining the desired type I error rate.

Discussion

In summary, we propose a penalized functional regression framework for across-platform imputation of methylation probes. Although a number of methods exist for predicting methylation levels at single CpG resolution, none of these directly apply to the across-platform imputation that we consider in this work. Moreover, we model information from nonlocal probes and have found such information considerably increase imputation performance. Our real data analysis demonstrates that by incorporating functional predictors from these nonlocal probes, our model can produce accurate imputation results when the reference panel (training set) and target panel (testing set) characterize the same tissue under similar conditions.

Because DNA methylation profiles are highly tissue and condition specific [Laurent et al., 2010; Lister et al., 2009; Varley et al., 2013], our method will not work well if the two datasets are from different tissues or very different conditions. Recent studies suggest some statistical models to predict methylation profile in target tissue from a surrogate tissue [Ma et al., 2014], which might be helpful in this case. Moreover, other systematic errors such as batch effect may also harm imputation quality. Therefore, we suggest using techniques such as principal component analysis to check for obvious discrepancies between reference and target panels before applying our method.

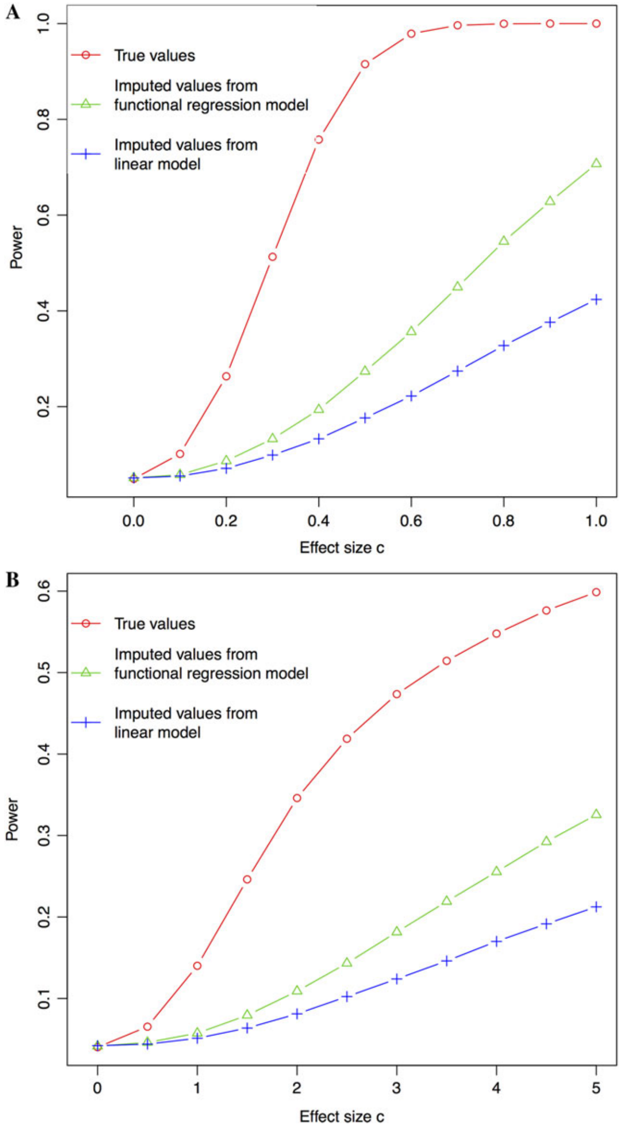

Figure 4. Empirical power of simulated association tests for (A) continuous trait and (B) binary trait across a spectrum of effect size *c*.

In various settings, a different way to construct predictors may further improve the performance of our model. For example, nonlocal probes can be categorized based on other properties, such as their relative location to a gene [Bibikova et al., 2011]. Another possible approach to select nonlocal probes is to choose HM27 probes highly correlated with the target probe (see Supplementary Methods). Supplementary Figure S1 shows that this approach can lead to better imputation performance, but the computational cost will be much higher. We can also explore other approaches to select local covariates, such as using a different number of probes, or choosing the local covariates as the 10 local probes that have the highest correlation with the target probe.

Because most CpG sites display stable DNA methylation levels, imputation error is low on average (the median imputation root MSE for beta values of all probes is ~0.05). Dichotomizing at beta value of 0.5 following Zhang et al. [2015], our prediction accuracy is 94.9%, largely consistent with their reported 92% prediction accuracy. However, researchers may consider dynamic CpG sites to be of more interest, as these sites often colocalize with key regulators, such as enhancers and transcription factor binding sites [Ziller et al., 2013]. Therefore, we calculated quality metrics for individual probes, facilitating the evaluation of imputation quality for each probe and removing probes with low imputation quality for downstream analysis.

For probes showing a large variation of methylation levels, we notice that even after incorporating functional predictors, the imputation quality is still low for a significant portion of these probes. Possible reasons are the following: first, the DNA methylation profile alone does not provide sufficient information for accurate imputation. We may need to incorporate other information to improve imputation quality, such as local DNA context and the binding profile of regulatory proteins [Bhasin et al., 2005; Bock et al., 2006; Zheng et al., 2013], although this requires additional data sources in the same or similar tissue type that are rarely available. Second, HM27 has a much lower resolution than HM450. In addition, a large proportion of HM27 probes showed nearly constant methylation levels across samples. As such, an extreme case is that if the target HM450 probe is not correlated with any HM27 probes, the model will be underfitted with the predicted methylation levels for all samples close to the average, thus leading to smaller variance than expected, similar to under-dispersion observed with imputed SNP data [Li et al., 2009]. We expect to observe better performance if we impute from a denser microarray. For example, researchers are now replacing the HM450 array with the Illumina EPIC 850K array. We anticipate that imputation from 450K probes to 850K probes will exhibit a much better quality. Third, our normalization procedure does not fully eliminate the inconsistency of measurements between HM27 and HM450, which also affects the performance of our model. Here, we assumed only HM450 data are available for the training dataset, which is a more realistic setting. However, if the training set contains both HM27 and HM450 data in a real case, we can treat HM450 data as response and use HM27 data to construct predictors. Thus, predictors from both training and testing set are constructed from HM27 data and the inconsistency between HM27 and HM450 is automatically learned by the model. In this case, our model will show higher imputation accuracy.

Because a considerable proportion of CpG probes on HM450 overlap with SNPs (hereafter referred to as SNP-probes), we also examined whether imputation quality for these SNP-probes differs from that for non-SNP probes. Our annotation [Barfield et al., 2014] includes 98,741 CpGs that have an SNP somewhere underneath the 50 bp probe, among which 62,777 are QC+ HM450-specific sites. We found that the SNP-probes are slightly less varying than the non-SNP probes (e.g., median variance of β values is 0.00310 and 0.00356, respectively; Table 3). Analogous to rarer variants in SNP imputation [Duan et al., 2013; Li et al., 2009; Liu et al., 2012; Pistis et al., 2015], it is not surprising to find that these SNP-probes appear slightly easier to impute when measured using MSE (e.g., median MSE is 0.00236 and 0.00263, respectively), but actually slightly more challenging to impute when measured using the more honest information content R^2 metric (median R^2 is 0.162 and 0.182, respectively).

Table 3. Imputation quality of SNP probes versus non-SNP probes

	Variance in β measurement		Imputation MSE		Imputation R^2	
	Mean	Median	Mean	Median	Mean	Median
SNP probe	0.0131	0.00310	0.0110	0.00236	0.206	0.162
Non-SNP probe	0.0140	0.00356	0.0115	0.00263	0.223	0.182

The focus of the present work is on imputation per se rather than association analysis. After accurate imputation, we can combine data from multiple platforms to obtain methylation levels of more CpG sites for downstream analysis such as detecting methylation quantitative trait loci or EWAS [Heyn and Esteller, 2012; Rakyan et al., 2011]. Such analysis can take imputation uncertainty into account similarly as for imputed SNPs [Huang et al., 2014]. In this work, we evaluated the statistical power under the mostly commonly observed change in mean values, however, other forms of changes have been observed. For example, several studies [Gervin et al., 2011; Hansen et al., 2011] reported differences in the variation (in addition to the mean) of methylation values between cancer and healthy groups. Our simulation studies (Supplementary Methods S2) show a power improvement even using the standard logistic regression to test the mean difference under such variation differences. Regardless of the epigenetic architecture of the phenotype, we expect our imputation method, by allowing in higher resolution and more powerful exploration of the epigenome, will lead to rapid advances in understanding the functional role of normal DNA methylation and the impact of its aberration. Our method is implemented in R and freely available at https://github.com/Leonardo0628/pfr.

Acknowledgments

We thank the Cancer Genome Atlas (TCGA) project for generating the data. The research is supported by R01HG006292, R01HG006703 (awarded to Y.L.) R01MH105561 (awarded to J.K.) and P01 CA142538 (awarded to J.Y.T.).

The authors declare no conflict of interest.

References

Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP and others. 2014. Accounting for population stratification in DNA methylation studies. *Genet Epidemiol* 38(3):231–241.

Bergman Y, Cedar H. 2013. DNA methylation dynamics in health and disease. *Nat Struct Mol Biol* 20(3):274–281.

Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu YP, Noushmehr H, Lange CPE, van Dijk CM, Tollenaar RAEM and others. 2012. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. *Nat Genet* 44(1):40–46.

Bhasin M, Zhang H, Reinherz EL, Reche PA. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett* 579(20):4302–4308.

Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL and others. 2011. High density DNA methylation array with single CpG site resolution. *Genomics* 98(4):288–295.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev* 16(1):6–21.

Bo TH, Dysvik J, Jonassen I. 2004. LS impute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res* 32(3):e34.

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* 2(3):e26.

Chen Y, Ning Y, Hong C, Wang S. 2014. Semiparametric tests for identifying differentially methylated loci with case-control designs using Illumina arrays. *Genet Epidemiol* 38(1):42–50.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc* 74(368):829–836.

Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ. 2006. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci USA* 103(28):10713–10716.

Donner Y, Feng T, Benoist C, Koller D. 2012. Imputing gene expression from selectively reduced probe sets. *Nat Methods* 9(11):1120–1125.

Duan Q, Liu EY, Auer PL, Zhang G, Lange EM, Jun G, Bizon C, Jiao S, Buyske S, Franceschini N and others. 2013. Imputation of coding variants in African Americans: better performance using data from the exome sequencing project. *Bioinformatics* 29(21):2744–2749.

Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA and others. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet* 38(12):1378–1385.

Ernst J, Kellis M. 2015. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* 33(4):364–376.

Fan R, Wang Y, Boehnke M, Chen W, Li Y, Ren H, Lobach I, Xiong M. 2015a. Gene level meta-analysis of quantitative traits by functional linear models. *Genetics* 200(4):1089–1104.

Fan R, Wang Y, Chiu CY, Chen W, Ren H, Li Y, Boehnke M, Amos CI, Moore JH, Xiong M. 2015b. Meta-analysis of complex diseases at gene level by generalized functional linear models. *Genetics* 202(2):457–470.

Gervin K, Hammero M, Akselsen HE, Moe R, Nygard H, Brandt I, Gjessing HK, Harris JR, Undlien DE, Lyle R. 2011. Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res* 21(11):1813–1821.

Getz G, Gabriel SB, Cibulskis K, Lander E, Sivachenko A, Sougnez C, Lawrence M, Kandoth C, Dooling D, Fulton R and others. 2013. Integrated genomic characterization of endometrial carcinoma. *Nature* 497(7447):67–73.

Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. 2011. Penalized functional regression. *J Comp Graph Stat* 20(4):830–851.

Gonzalo S. 2010. Epigenetic alterations in aging. *J Appl Physiol* 109(2):586–597.

Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D and others. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* 43(8):768–775.

Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong CB, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao YJ and others. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol* 28(10):1097–1194.

Heyn H, Esteller M. 2012. DNA methylation profiling in the clinic: applications and challenges. *Nat Rev Genet* 13(10):679–692.

Horvath S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol* 14(10):R115.

Huang KC, Sun W, Wu Y, Chen M, Mohlke KL, Lange LA, Li Y. 2014. Association studies with imputed variants using expectation-maximization likelihood-ratio tests. *PLoS One* 9(11):e110679.

Jewett EM, Zawistowski M, Rosenberg NA, Zollner S. 2012. A coalescent model for genotype imputation. *Genetics* 191(4):1239–1255.

Kim H, Golub GH, Park H. 2005. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 21(2):187–198.

Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, Fulton LL, Dooling DJ, Ding L, Mardis ER and others. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61–70.

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Sung KWK, Rigoutsos I, Loring J and others. 2010. Dynamic changes in the human methylome during differentiation. *Genome Res* 20(3):320–331.

Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, Robertson AG, Hoadley K, Triche TJ, Laird PW, Baty JD and others. 2013. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New Eng J Med* 368(22):2059–2074.

Li Y, Willer C, Sanna S, Abecasis G. 2009. Genotype imputation. *Ann Rev Genomics Human Genet* 10:387–406.

Liew AWC, Law NF, Yan H. 2011. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Brief Bioinform* 12(5):498–513.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM and others. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.

Liu EY, Buyske S, Aragaki AK, Peters U, Boerwinkle E, Carlson C, Carty C, Crawford DC, Haessler J, Hindorff LA and others. 2012. Genotype imputation of metabochip SNPs using a study-specific reference panel of ∼4,000 haplotypes in African Americans from the Women's Health initiative. *Genet Epidemiol* 36(2):107–117.

Ma B, Wilker EH, Willis-Owen SA, Byun HM, Wong KC, Motta V, Baccarelli AA, Schwartz J, Cookson WO, Khabbaz K and others. 2014. Predicting DNA methylation level across human tissues. *Nucleic Acids Res* 42(6):3515–3528.

Pistis G, Porcu E, Vrieze SI, Sidore C, Steri M, Danjou F, Busonero F, Mulas A, Zoledziewska M, Maschio A and others. 2015. Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* 23(7):975-83.

Rakyan VK, Down TA, Balding DJ, Beck S. 2011. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 12(8):529-541.

Smith ZD, Meissner A. 2013. DNA methylation: roles in mammalian development. *Nat Rev Genet* 14(3):204-220.

The Cancer Genome Atlas Research Network. 2012. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417):519–525.

The Cancer Genome Atlas Research Network. 2013. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499(7456):43–49.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17(6):520–525.

Varley KE, Gertz J, Bowling KM, Parker SL, Reddy TE, Pauli-Behn F, Cross MK, Williams BA, Stamatoyannopoulos JA, Crawford GE and others. 2013. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res* 23(3):555–567.

Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. 2015. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol* 16:14.

Zheng H, Wu H, Li J, Jiang SW. 2013. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC Med Genomics* 6(Suppl 1):S13.

Ziller MJ, Gu HC, Muller F, Donaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE and others. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* 500(7463):477–481.