Imputation of human methylation array based on KNN algorithm

KATY MARTINSON[1,2], AIDAN ZHANG[3], Emily Maciejewski[4], Jason Ernst[4,5,6]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 Department of Biology, Bucknell University
3 Computational and Systems Biology Interdepartmental Program, UCLA
4 Department of Computer Science, UCLA
5 Department of Biological Chemistry in the David Geffen School of Medicine, UCLA
6 Department of Computational Medicine, UCLA

Abstract

Whole-genome bisulfite sequencing (WGBS) is a powerful and expensive tool that provides genome-wide single-base resolution of methylated cytosines. Methylation arrays are a cheaper alternative and are commonly used for cohort and EWAS studies, but they suffer from low CpG site coverage. Imputation of missing CpGs is necessary to meet the same coverage of WGBS data. Using the k-nearest neighbors algorithm (KNN), we can accurately extend methylation arrays using IHEC WGBS[1] methylation data as a reference. We calculated nearest neighbors and distances from the reference and used these to predict methylation values for CpG sites not located on the array. We also transferred these calculations to a different platform, the EPIC BeadChip array[2] downloaded using the recountmethylation package. Both studies using the KNN algorithm demonstrated higher correlations to the ground-truth than when compared to a baseline. Using algorithms to impute methylation values rather than depending on WGBS data vastly reduces costs and efforts for EWAS studies.

Background

WGBS data, while extremely valuable for methylation studies, is not the most practical method of collecting methylation data. This is simply due to its high costs and difficulty to obtain. This creates a restriction on what kinds of studies can be done with this data and who can do those studies. In order to make this kind of research more feasible, machine-learning methods are used to make predictions of WGBS data from methylation arrays, which are much cheaper and easier to gather data with. The combination of imputation and arrays allows labs to have WGBS-like data of similar quality at a fraction of the cost and efforts.

Methods

The k-nearest neighbors algorithm, or KNN, is a learning algorithm that makes predictions of a single point based on the average of the k-nearest surrounding values of its group. Using WGBS data as a reference, we calculated the genomic distance between all IHEC CpG sites that overlap with the Illumina Epic BeadChip array and those that do not overlap. Based on these distances, all probes were ranked by closest-neighbor order. We selected a k value of 32 based on its performance on a tuning set of data (Figure 1). The prediction for each non-probe site becomes the average of k nearest probes on the array. Site-wise Pearson correlations and RMSE values were calculated to analyze the accuracy of these predictions.

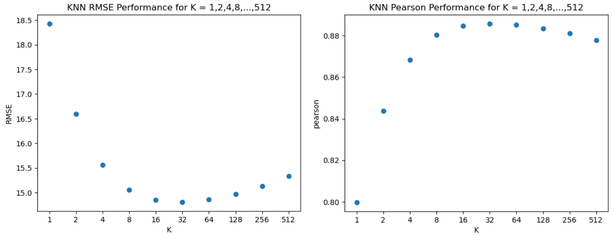

Figure 1. K-selection performance on tuning set of data to ensure no over-fitting.

Results

IHEC WGBS Data

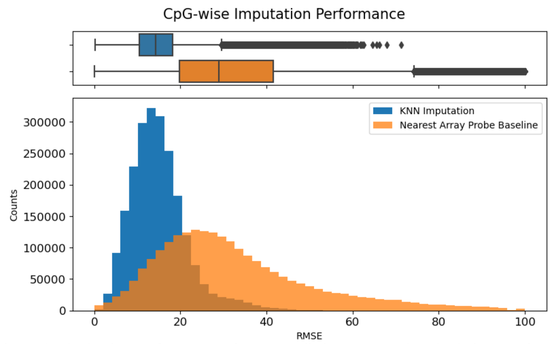

Figure 2. CpG-wise performance of KNN predictions on IHEC WGBS data against a nearest-array probe baseline. Predicted methylation values range between 0 and 100. Predicted values were significantly different compared to the nearest array probe baseline (p-value < 0.01).

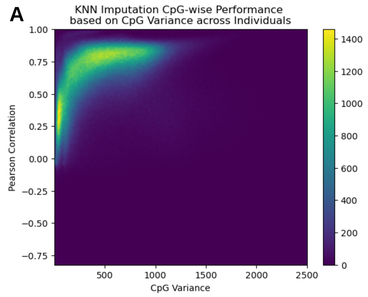

Figure 4. CpG-wise performance of KNN predictions on IHEC WGBS data against the variance of each CpG site. **A.** Pearson correlation values are consistent across medium to high levels of variance. CpG-wise performance was consistent across different variance levels, demonstrating the reliability of the algorithm even when a CpG site is variable across individuals. **B.** Compared to the nearest-array probe baseline, the Pearson correlation trend shows a much higher level of correlation between the predictions and the ground-truth. The clustering of low Pearson correlations in sites with very low variances can be explained by the nature of using Pearson correlations. In cases where there is extremely low variance, RMSE provides a better metric and showed similar quality results in those analyses as well.

Recount Array Data

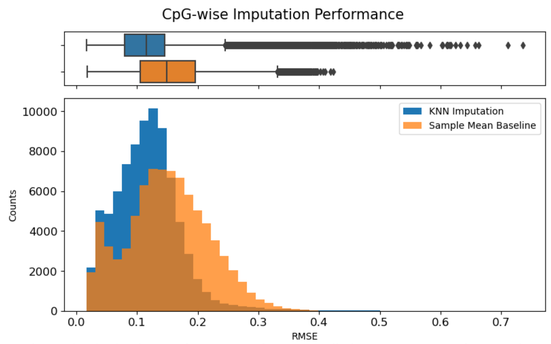

Figure 3. CpG-wise performance of KNN predictions on recount data against a nearest-neighbor baseline. Predicted methylation values range between 0 and 1. Predicted values were significantly different compared to the sample-mean baseline (p-value < 0.01).

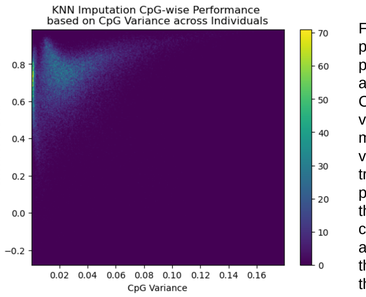

Figure 5. CpG-wise performance of KNN predictions on recount data against the variance of each CpG site. Pearson correlation values are consistent across medium to high levels of variance and match WGBS trend. The trend appears more prominently in the WGBS, but this is expected. A baseline could not be made for this analysis due to the nature of the Pearson correlation and the selection of our baseline.

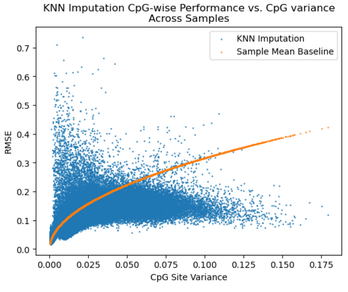

Figure 6. CpG-wise performance of KNN predictions on recount data against the variance of each CpG site. RMSE is a better metric for lower variance sites. 83% of predicted sites performed better than the baseline.

Tissue Analysis

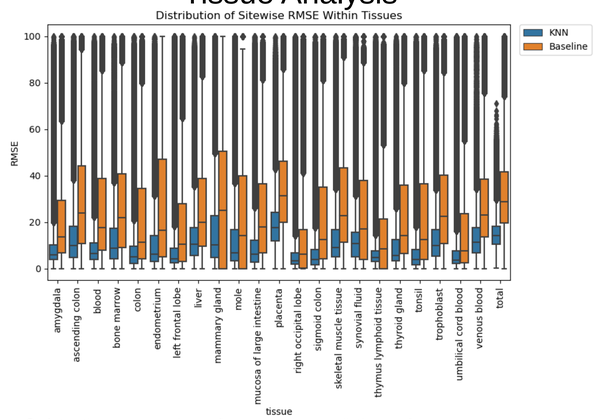

Figure 8. Intra-tissue comparison of sample-wise prediction performance compared to a nearest-neighbor baseline for each tissue. For each tissue, the median site-wise RMSE is lower in the KNN method. For the KNN method, the median site-wise RMSE values in nearly all tissues are lower than the median site-wise RMSE taken over all samples, indicating that the KNN captures individual variation within tissue types.

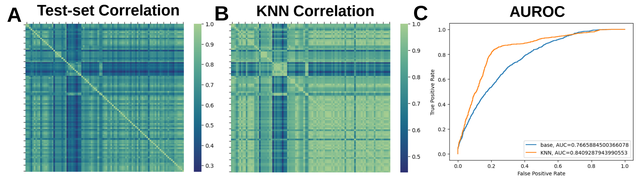

Figure 9. Inter-tissue correlations of **A.** test-set WGBS data and **B.** KNN predictions. Visually, the correlation matrices generated from the predictions on the test set and the test set itself are similar. **C.** Together with the AUROC, this indicates that the KNN may be capturing variation between tissue types and may furthermore be denoising the data (Figure 6). It is also possible that many CpG sites have their nearest neighbors as array probes with methylation values that differ between tissue types.

Cross-Platform Comparison

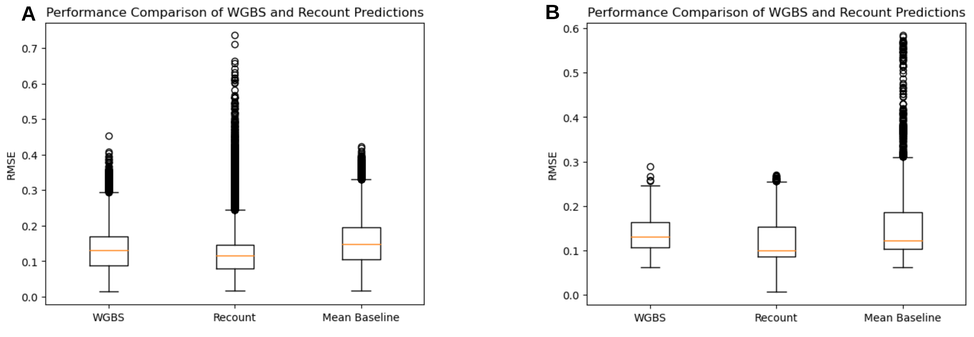

Figure 7. RMSE comparison between **A.** site-wise WGBS performance, recount performance, and a baseline of means for each CpG-site. **B.** Is the same comparison of performances, but sample-wise. In both cases, the WGBS and Recount data are significantly different than the mean baseline (p-value < 0.01).

Discussion

The KNN algorithm, using WGBS data as a reference, is a viable method of imputation for DNAm data. Performance analyses showed high quality levels of predictions with the IHEC data (Figure 2). Distances and nearest-neighbor calculations are transferable between platforms and give similar performance results (Figure 3). The use of Pearson correlations and RMSEs accounted for different performances with more variable CpG-sites across samples (Figure 4, 5, 6). Tissue analyses showed high performance of imputation within tissues as well as similar correlation of inter-tissue methylation values between predictions and ground-truths (Figure 8, 9). Future directions involve EWAS studies with the recount methylation data as well as more tissue analysis in the WGBS data.

Acknowledgements

The Ernst Lab, BIG Summer Program, International Human Epigenome Consortium, National Institutes of Health Gene Expression Omnibus, NIH DP1DA044371, UCLA Jonsson Comprehensive Cancer Center, Eli and Edythe Broad Center of Regenerative Medicine, Stem Cell Research Ablon Scholars Program (J.E.), NIH Training Grant in Genomic Analysis and Interpretation T32HG002536, NSF and REU Award Number 1758002

References

Hendrik G. Stunnenberg et. al, The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery, Cell, Volume 167, Issue 5, 2016, Pages 1145-1149, ISSN 0092-8674, https://doi.org/10.1016/j.cell.2016.11.007.

Maden, S. K., Walsh, B., Ellrott, K., Hansen, K. D., Thompson, R. F., & Nellore, A. (2023). recountmethylation enables flexible analysis of public blood DNA methylation array data. Bioinformatics advances, 3(1), vbad020. https://doi.org/10.1093/bioadv/vbad020