Novel Missing Data Imputation Approaches Enhance Quantitative Trait Loci Discovery in Multi-Omics Analysis

Zining Qi[1,2], Alexandre Pelletier[3], Jason Willwerscheid[4], Xuewei Cao[1], Xiao Wen[5], Carlos Cruchaga[6,7,8], Philip De Jager[9,10], Julia TCW[3,*], and Gao Wang[1,11,*]

[1]Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Columbia University, New York, NY, USA
[2]Department of Biostatistics, Columbia University, New York, NY, USA
[3]Department of Pharmacology, Physiology & Biophysics, Chobanian & Avedisian School of Medicine, Boston University, MA, USA
[4]Department of Mathematics & Computer Science, Providence College, Providence, Rhode Island, USA
[5]Data Science Institute, Columbia University, New York, NY, USA
[6]Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA
[7]NeuroGenomics and Informatics Center, Washington University School of Medicine, St. Louis, MO, USA
[8]Charles F. and Joanne Knight Alzheimer Disease Research Center, Washington University School of Medicine, St. Louis, MO, USA
[9]Department of Neurology, Center for Translational & Computational Neuroimmunology, Columbia University Medical Center, New York, NY, USA
[10]Cell Circuits Program, Broad Institute, Cambridge, MA, USA
[11]Department of Neurology, Columbia University, New York, NY, USA
[*]Correspondence: Julia TCW, juliatcw@bu.edu; Gao Wang, wang.gao@columbia.edu

Abstract

Handling missing values in multi-omics datasets is essential for a broad range of analyses. While several benchmarks for multi-omics data imputation methods have recommended certain approaches for practical applications, these recommendations are not widely adopted in real-world data analyses. Consequently, the practical reliability of these methods remains unclear. Furthermore, no existing benchmark has assessed the impact of missing data and imputation on molecular quantitative trait loci (xQTL) discoveries. To establish the best practice for xQTL analysis amidst missing values in multi-omics data, we have thoroughly benchmarked 16 imputation methods. This includes methods previously recommended and in use in the field, as well as two new approaches we developed by extending existing methods. Our analysis indicates that no established method consistently excels across all benchmarks; some can even result in significant false positives in xQTL analysis. However, our extension to a recent Bayesian matrix factorization method, *FLASH*, exhibits superior performance in multi-omics data imputation across various scenarios. Notably, it is both powerful and well-calibrated for xQTL discovery compared to all the other methods. To support researchers in practically implementing our approach, we have integrated our extension to *FLASH* into the R package `flashier` , accessible at `https://github.com/willwerscheid/flashier`. Additionally, we provide a bioinformatics pipeline that implements *FLASH* and other methods compatible with xQTL discovery workflows based on `tensorQTL`, available at `https://cumc.github.io/xqtl-pipeline/code/data_preprocessing/phenotype/phenotype_imputation.html`.

1. Introduction

Technological advancements in recent years have greatly enhanced the availability of high-throughput biological tools for researchers at reduced cost. This has led to a dramatic increase in the rate of data generation particularly in the area of multi-omics data science, where multiple types of molecular phenotypes (or traits) are generated from the same set of samples. Each type of multi-omics data, such as methylation, proteomics, and metabolomics, offers unique insights into different layers of biological processes. Methylation studies, for example, sheds lights into the epigenetic aspects affecting gene

activity. Proteomics focuses on the variety and levels of proteins that emerge from the process of gene expression. Furthermore, metabolomics analyses metabolites as products of synergistic interactions among multiple genes, to provide a broader perspective on metabolic pathways relevant to complex traits and disease phenotypes [1].

Multi-omics phenotypes are now routinely measured in diverse biological samples, spanning various cell types and tissues. These measurements, obtained through different platforms, often contain varying levels of missing data. Such gaps in multi-omics data can occur for multiple reasons: poor sample quality, limited sample size to observe low phenotypic values, technical limitations of measurement platforms, and other reasons for sample drop-out in quality control. Depending on data-type, missing values in multi-omics studies can range from 5% to as high as 80%, according to our literature review detailed in Table S1. Consequently, it warrants thorough investigation how missing data is addressed in multi-omics research. This is particularly crucial given unique nature of each different multi-omics studies that leads to varying degrees and patterns of missing data.

Imputation is a widely accepted strategy for tackling missing data issues, facilitating statistical analyses on complete datasets without excluding samples or features that many statistical methods, not tailored for missing data, cannot handle. Several approaches have been documented in the existing multi-omics literature for missing value imputation [2]. Simple methods like mean imputation replace missing values with the average of observed data, whereas the lowest of detection (LOD) method uses the smallest observed value for filling in missing entries. Mean imputation and its more sophisticated derivatives — such as the K Nearest Neighbors (KNN) that leverages closest observed samples — are extensively utilized in various multi-omics studies [3,4]. Advanced methods — including low-rank approximation via matrix factorization [5–7] along with sophisticated regression methods that combine single or multiple imputation (SI/MI) with machine learning strategies [8] — have shown better performance than KNN in simulation benchmark studies [9], yet they are not as commonly adopted as KNN or other simpler methods in practice. For a comprehensive list of multi-omics imputation method benchmarks and the recommended approaches derived from these benchmarks, please refer to Table S2.

We have identified several gaps between methodological work in the literature and their practical applications to multi-omics data. Firstly, as indicated in Tables S1 and S2, it is evident that real-world multi-omics studies (even high-impact ones) frequently opt for simpler imputation techniques over more advanced methods [10,11]. This trend raises questions about the practical reliability of advanced methods in various multi-omics research contexts, such as clustering, differential expression analysis, and molecular quantitative trait loci (xQTL) discoveries. Secondly, most existing benchmarks focus narrowly on specific types of multi-omics data. There are only a few that comprehensively address a broad spectrum of data types with varying sample sizes, extents, and patterns of missing data [12–14]. Furthermore, these benchmarks often rely on synthetic data such as simulations based on real data to assess methods performance, leaving uncertainties about how these findings translate to reliability and robustness in real-world studies. Lastly, to our knowledge there is not yet a benchmark specifically tailored for xQTL analysis. This gap became apparent in our search for methodological guidance to impute multi-omics data for xQTL analysis, posing challenges in making informed decisions for several of our ongoing research projects.

Driven by the practical need to tackle missing data issues in xQTL analysis, we conducted an extensive evaluation of 16 methods as detailed in Table S3 Promising results from some existing methods in our preliminary benchmarks motivated us to develop new enhancements using similar models and algorithms. We utilized data across three cohorts/biobanks, focusing on three types of multi-omics data known for wide-spread missing values (methylation, proteomics, and metabolomics). Our approach involves conducting simulation studies to assess imputation accuracy, xQTL discovery power and calibration, along with real-world xQTL discovery and subsequent replication studies. We show that our extension to a recent empirical Bayes matrix factorization approach, *FLASH*, outperforms other methods in almost every scenario we examined, is reasonably fast for large-scale multi-omics datasets, and producing outputs that are more straightforward to interpret compared to various machine learning approaches.

2. Results

2.1. Overview of multi-omics imputation and xQTL discovery benchmark design

We developed a comprehensive benchmark for simulation and real data analysis to evaluate methods across three multi-omics modalities — proteomics, metabolomics, and DNA methylation — in three multi-omics resources including the Religious Orders Study/Memory and Aging Project (ROSMAP), Knight Alzheimer Disease Research Center (Knight), and Mount Sinai Brain Bank (MSBB). These datasets underwent standard quality control (QC) and normalization, as previously detailed for these specific datasets[15–17]. By offering a diverse range of sample sizes, molecular features, measurement platforms, and missing data patterns, these datasets enables us to create an extensive framework for missing data methods assessment (Details see Section 4).

As illustrated in Figure 1, the multi-omics imputation and QTL analysis benchmark in our study consists of several key steps. We began by simulating missing data patterns completely at random (MCAR) across various rates, from a minimal 5% to an extreme 50%, reflecting both literature reports (Table S1) and the characteristics of our own data-sets (Figure S1). To more accurately approximate real-world multi-omics data with realistic missing patterns, we developed a descriptive model based on data from ROSMAP, Knight, and MSBB to generate missing values not at random (MNAR) for the three modalities assessed (Section 4.2.1). Beyond benchmarking imputation accuracy, we further conducted numerical studies for xQTL discovery. This involved simulating genotypic associations (xQTL) with realistic molecular phenotypes, and applying statistical fine-mapping to determine the impact of imputation methods on identifying true non-zero genotypic effects. We then applied these methods to discover xQTL in our datasets, focusing on pQTL and metaQTL, and compared significant genes identified by each method. To validate the robustness of our findings, we carried out replication studies for pQTL discovery among the three resources.

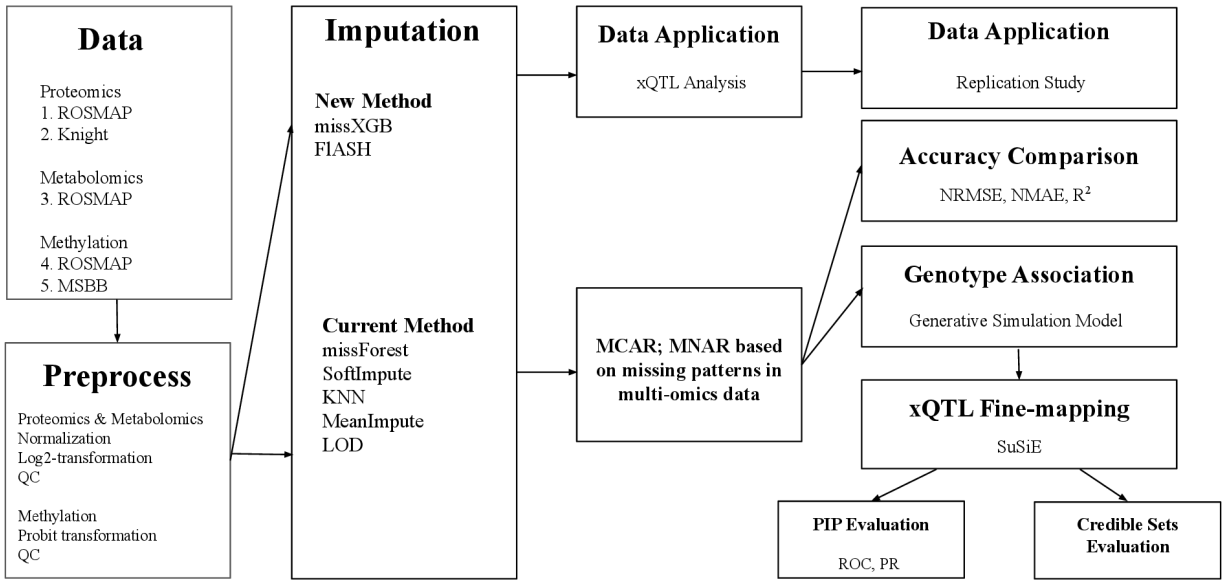

Fig. 1. Overview of multi-omics imputation methods and xQTL discovery benchmark. This workflow summarize the datasets, imputation methods and key steps of evaluating imputation methods in the context of xQTL studies, using both simulated and real-world multi-omics data.

We assessed imputation accuracy, xQTL detection power, false discovery rate control, real-world xQTL discovery, replication, and runtime across 16 methods (Table S3). This selection represents major imputation techniques in the multi-omics field (Table S1 and S2), covering: 1) observed value based methods such as *MeanImpute*, *LOD*, *KNN*; 2) low-rank approximation methods such as *SoftImpute*[5]

and *FLASH*[18]; and 3) machine learning regression approaches incorporated to single or multiple imputation frameworks[19,20] such as *MissForest*[8], and LR[21].

We also introduce two new methods by extending existing ones. The first, *MissXGB*, is an adaptation of *MissForest*, using XGBoost[22] for machine learning regression instead of traditional random forest regression trees. The second, and optimized version of *FLASH*, enhances computational efficiency and imputation accuracy through a genome-wide *FLASH* approach that explicitly models chromosome-specific and global factors. Further details about *MissXGB* and the modified *FLASH* are provided in Section 4.1.

After exploring various parameter settings for the methods under consideration, such as the number of neighbors (K) in KNN, the regularization and threshold parameters in *SoftImpute*, and the choice of priors in *FLASH*, we finalized a selection of seven methods for comprehensive benchmarking as described earlier. These methods are summarized with their respective abbreviations in the box "Imputation" on Figure 1. Key findings from our simulation studies are summarized in Figures 2, 3, 4, with additional details in Table S4. Real-world data analysis results are summarized in Figure 5 and Table 1, with additional details in Figures S1 and S2.

2.2. FLASH outperforms other methods in imputation accuracy for multi-omics data

We gauged the efficacy of various imputation methods using metrics such as NRMSE, NMAE, and R^2, the definitions of which are provided in Section 4.2.2. Our evaluation focused on individual molecular features, including protein levels for specific genes, metabolite measurements, and CpG site intensities. Aggregated performance across multiple molecular features are shown as box plots (Figure 2 and 3). For each dataset we evaluated a range of missing rates of MCAR scenarios, along with dataset-specific MNAR patterns. Interestingly, we observed that the pattern of missing data was less influential than the sample size and the number of features. Therefore, in our analysis, we present MCAR scenarios at 50% as an example of a moderate to high level of missing data, a situation frequently encountered in real-world settings (Table S1, Figures S1 and S2).

We found that widely used observation based methods like KNN (demonstrated here as KNN_{10} with $K = 10$) and *MeanImpute* consistently underperform across various scenarios, especially when compared to low-rank approximation and machine learning regression techniques. The new imputation tool we developed, *MissXGB*, showed comparable or slightly superior performance to *MissForest*. We also noted that *SoftImpute* is effective for larger datasets (in terms of samples and features) compared to *MissForest* and *MissXGB*. However, its performance suffers in datasets with fewer samples and features, such as Knight proteomics and ROSMAP metabolomics, though it still surpasses KNN_{10} and *MeanImpute*. Our enhanced *FLASH* method proved capable of efficiently handling both large and small multi-omics datasets, consistently outperforming other methods in all evaluated scenarios.

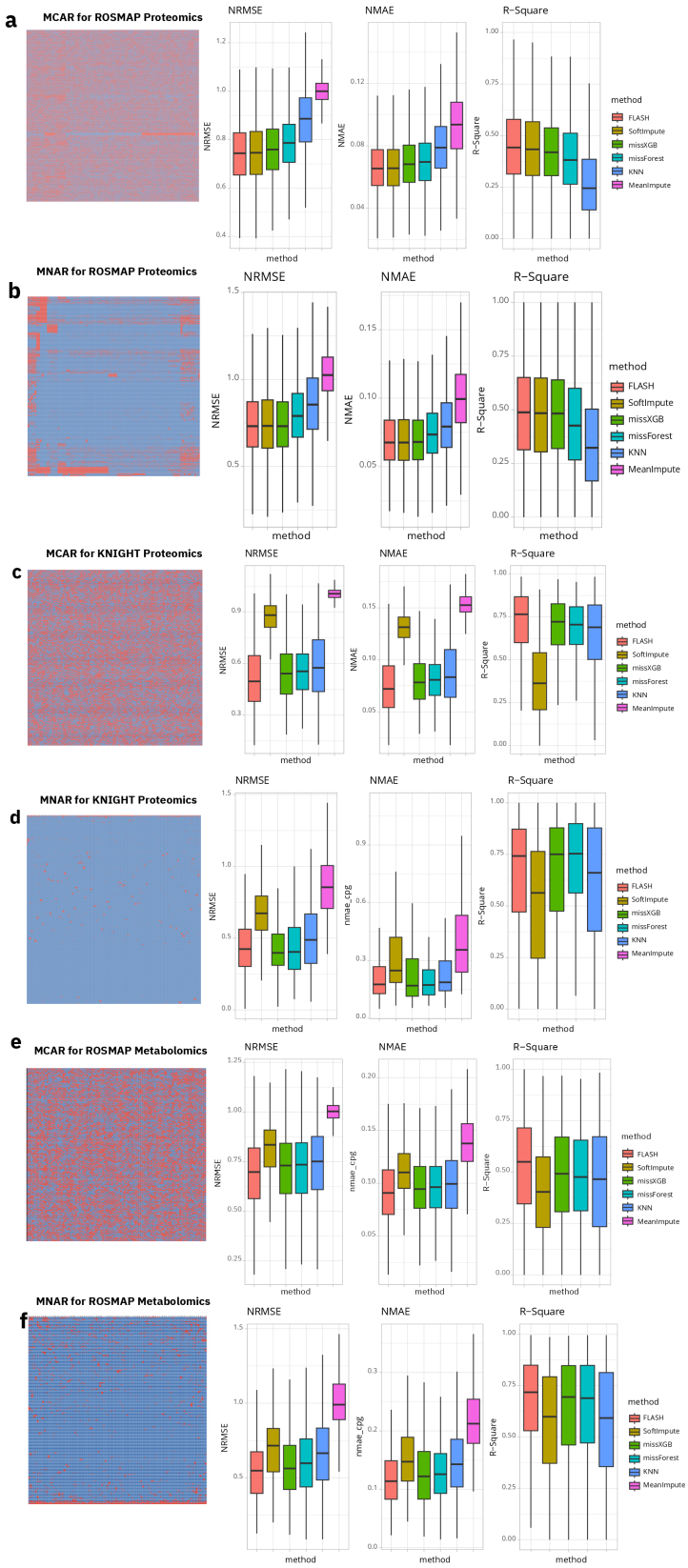

Fig. 2. Imputation accuracy for Proteomics and Metabolomics data. Panel a-f summarize the performance of methods on imputation accuracy for proteomics and metabolomics data. Missing patterns are displayed for different scenario and datasets, followed by boxplots showing feature-wise accuracy using 3 metrics. Panel a, c, e illustrate ROSMAP proteomics, Knight proteomics, and ROSMAP metabolomics data with MCAR. Panel b, d, f illustrate ROSMAP proteomics, Knight proteomics, and ROSMAP metabolomics data with MNAR.

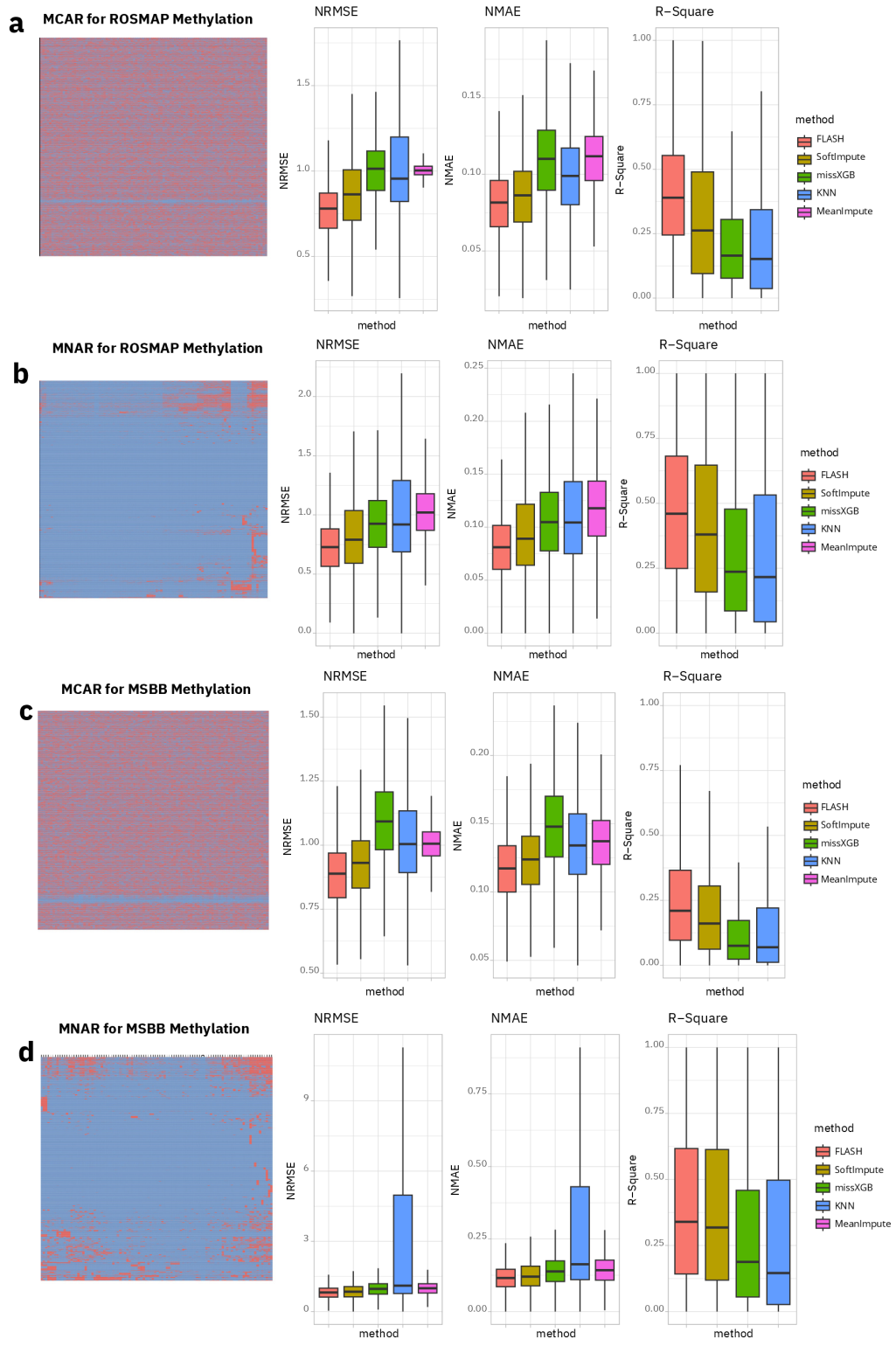

Fig. 3. Imputation accuracy for Methylation data. Similar to Figure 2, panel a and c illustrates ROSMAP methylation and MSBB methylation with MCAR. Panel b and d illustrate ROSMAP methylation, MSBB methylation with MNAR.

2.3. FLASH demonstrates increased power and improved false discovery control in xQTL studies

After conducting fine-mapping on simulated xQTL data, as explained in Section 4.3, we assessed power and false discovery rates using precision-recall and ROC curves at various posterior inclusion probability thresholds. This was done under the challenging conditions of 50% MCAR settings, reflecting moderate to high missing rates commonly seen in practice. Consistent with our predictions, the effectiveness of imputation methods in xQTL discovery power aligns with our earlier evaluation of their imputation performance. For moderately sized multi-omics data, *FLASH* outperforms other methods, closely followed by *SoftImpute*. Both *MissForest* and *MissXGB* show similar results and are notably more effective than KNN_{10}, *MeanImpute*, and *LOD* (Figures 4, panel a, d, e). Performance of *SoftImpute* declines in smaller data-sets (Figures 4, panel b, c). We noticed that in simulations of Knight-based pQTL studies with small sample and feature sizes, KNN_{10} outperforms *MissXGB* in terms of power.

We also evaluated the calibration of fine-mapping credible sets (CS) by their coverage, defined as the proportion of CS capturing the true simulated effects. For well-calibrated 95% CS, the coverage is expected to be at least 95%. Our findings indicate that not all imputation methods lead to well-calibrated CS. As shown in Figure 4, panel a-c, *LOD* and *MeanImpute* have inadequate false discovery rate (FDR) control. *SoftImpute* also fails to control FDR in smaller datasets. Even KNN_{10} and *MissXGB* exhibit slightly inflated 95% CS under certain conditions. Reassuringly, both *FLASH* and *MissForest* demonstrate well-controlled FDR. Moreover, compared to *MissForest*, 95% CS from *FLASH* attain higher observed coverage.

The outcomes of our numerical xQTL studies highlight that *FLASH*-based imputation of molecular phenotype measurements is the most powerful for xQTL discovery, while also being the most conservative in terms of FDR.

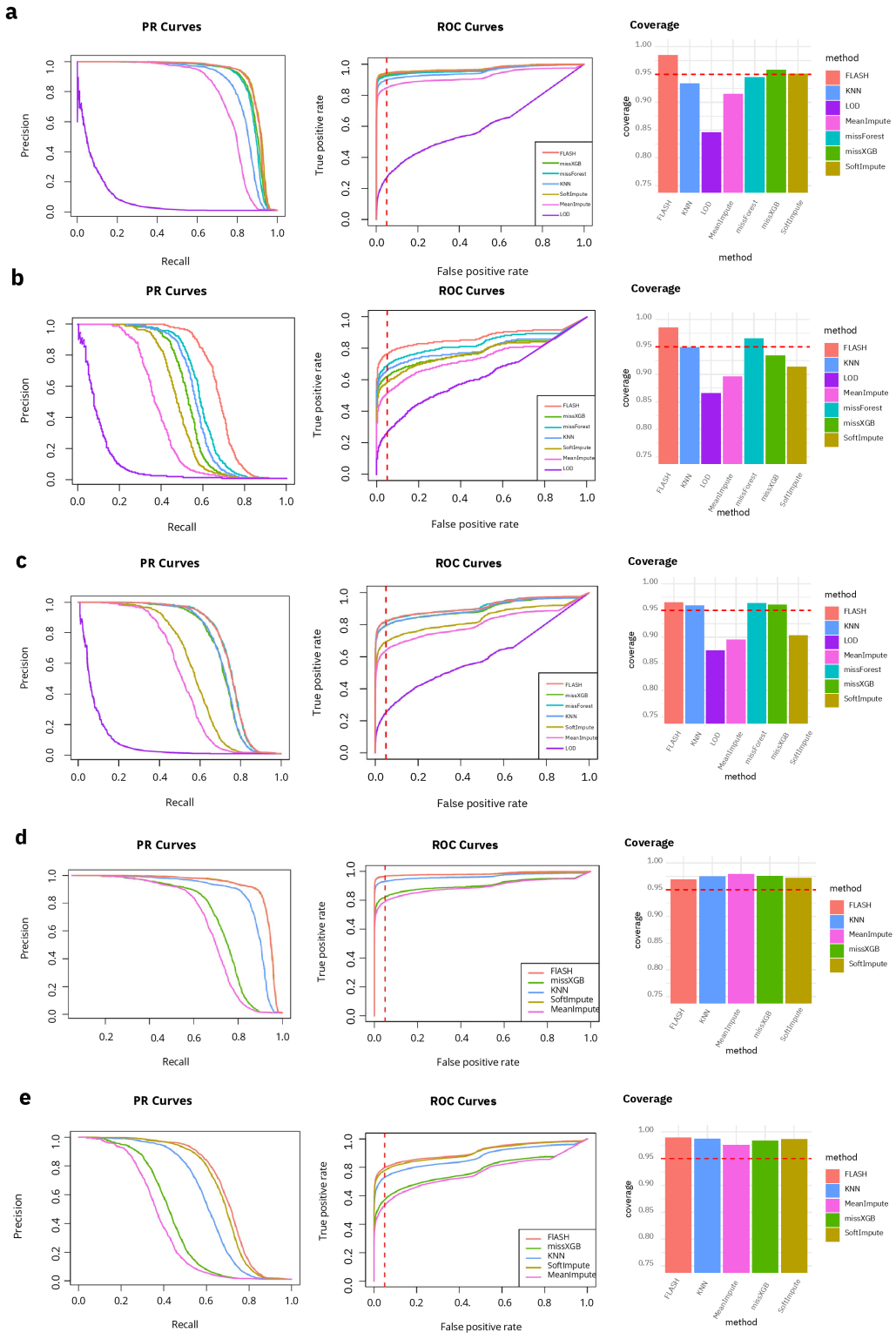

Fig. 4. Impact of imputation methods on xQTL discoveries. Fine-mapping posterior inclusion probability and credible sets (CS) comparison of imputation methods are summarized as PR/ROC curves and bar plots. Panel a is results for ROSMAP Proteomics. The PR and ROC curves are calculated based on PIP result from SuSiE[23]. The red dashed line on ROC curves are false positive rate at 0.05. Coverage can be interpreted as $1 - FDR$; the red dashed line on the coverage plots is 0.95 corresponding to FDR at 0.05 for fine-mapping CS. Panel b, c, d, e summarize the performance for Knight proteomics, ROSMAP metabolomics, ROSMAP methylation, and MSBB methylation respectively.

2.4. FLASH yields additional and robust xQTL discoveries in real-world multi-omics data analysis

2.4.1. Statistical analysis for xQTL association

We applied seven imputation methods to proteomics data from ROSMAP (totaling 7,712 genes) and Knight (1,181 genes), along with metabolomics data from ROSMAP (635 metabolites). These datasets were then processed using the FunGen-xQTL analysis protocol described in Section 4.4. Following standard xQTL study practices, we present a gene-level summary of pQTL discovery. This involves permutation testing for each gene within the cis-window to adjust for multiple testing of many genetic variants[24,25], and genome-wide false discovery rate control at gene level using the qvalue method[26]. For the metabolomics data, we conducted a GWAS for each metabolite, counting significant metaQTLs as those with p-values below the standard GWAS threshold of 5×10^{-8}, without additional adjustments for multiple testing at metabolites level. xQTL discoveries are summarized in Table 1. We selected KNN_{10} as a baseline for comparison due to its popularity in multi-omics literature, and relatively calibrated results across evaluated scenarios as shown in Figure 4.

In the ROSMAP proteomics data, *FLASH* led to the identification of the most pGenes (genes with at least one pQTL), closely followed by *SoftImpute*, *MissXGB*, and *MissForest*. While *MeanImpute* and *LOD* detected more pGenes than KNN_{10}, these results warrant caution given their potential for inflated false discoveries as previously demonstrated. In the smaller Knight dataset, *FLASH* proved to be the most conservative, yielding the fewest pGenes. Conversely, *SoftImpute*, which has shown a propensity for false positives, identified nearly double the pGenes compared to *FLASH*, while *MeanImpute* found over three times more. The most striking result was from *LOD*, reporting 1,056 pGenes — 91.7% of all genes in the Knight proteomics dataset — which likely includes many false positives.

Table 1. Analysis of pQTL data: discovery from ROSMAP. The first column summarizes the number of pQTL genes identified by each method and their proportion compared to the total gene counts in ROSMAP. The second column compares each method to KNN_{10}, showing the percentage increase or decrease in the number of significant genes identified. The last column summarizes genes uniquely identified by each method.

Method	ROSMAP		
	#(%) sig. genes	diff. from KNN_{10}	method specific
FLASH	2,811(37.8%)	18.8%	930
MissXGB	2,640(35.5%)	11.58%	759
MissForest	2,625(35.3%)	10.95%	744
KNN_{10}	2,336(31.4%)	0.0%	455
SoftImpute	2,654(35.7%)	12.17%	773
MeanImpute	2,585(34.8%)	9.27%	704
LOD	2,527(34.0%)	6.80%	646

2.4.2. Replication analysis for pQTL results between Knight and ROSMAP

In the context of real-world xQTL discoveries, the absence of known ground truth poses a challenge in validating the robustness different imputation methods for the xQTL signals obtained through them. To address this, we conduct replication analyses between Knight and ROSMAP data-sets. Specifically, we designated the pQTL data from the smaller Knight proteomics sample as the discovery set and the pQTL from the larger ROSMAP sample as the replication set. We incorporated cross-method results when defining these sets, aiming to create both reasonable discovery and robust replication data. The criteria for these sets were established as follows:

- Method-specific discovery set: for genes that exist in both ROSMAP and Knight, we evaluate significant genes (permutation q-value < 0.05) in Knight pQTL reported by each of the 7 methods, which gives us 7 discovery sets.
- Baseline discovery set: for genes that exist in both ROSMAP and Knight, we consider those reported significant (permutation q-value < 0.05) in Knight pQTL by 6 methods excluding *LOD*. This gives us 76 genes.
- Joint discovery set: for genes that exist in both ROSMAP and Knight, we consider those reported significant (permutation q-value < 0.05) in Knight pQTL by any of those methods that shows well

calibrated coverage in simulation studies, including *FLASH*, *MissForest*, *MissXGB* and KNN_{10}. This is a total of 78 genes.

- Replication set: for genes that exist in both ROSMAP and Knight, we consider those reported significant (permutation q-value < 0.05) in ROSMAP pQTL by 6 methods excluding *LOD*. This is a total of 171 genes.
- We compute replicate rate as $\frac{\#\text{genes in both discovery and replication sets}}{\#\text{genes in discovery set}}$. This gives us baseline replication rate of 57.9% and joint replication rate of 56.4%.
- We define relative replication rate as method specific replication rates divided by the joint replication rate.

Method-specific replication rates are shown in Figure 5. In these comparisons, *FLASH* continues to outperform, demonstrating the highest replication rates among the methods evaluated.

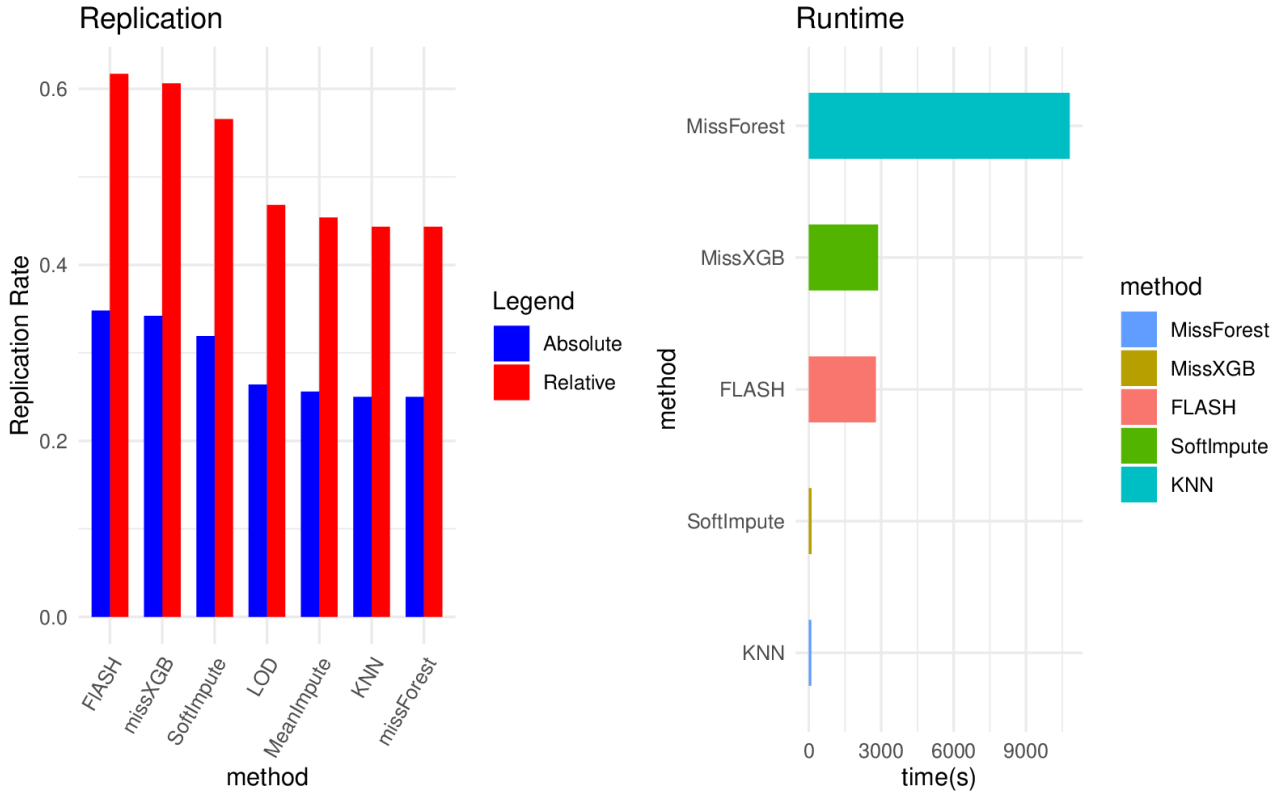

Fig. 5. Analysis of pQTL data: replication and runtime. Panel a summarizes the absolute and relative replication rate for each method in the pQTL replication study, comparing discoveries in Knight to those in ROSMAP. The relative replication rate is calculated by absolute replication over replication of joint discovery set. Panel a shows the runtime of each method for imputation on ROSMAP proteomics data.

Runtime benchmarks for all imputation methods were conducted using the ROSMAP proteomics data. Our findings indicate that while *FLASH* and *MissXGB* take longer to process than simpler methods like *MeanImpute*, *LOD*, and KNN_{10}, they are still significantly faster than *MissForest*, and are practical for analyzing real-world datasets.

3. Discussion

Missing data in multi-omics studies is a common issue that poses practical challenges. Various algorithms, statistical models, and machine learning methods are theoretically available to impute missing multi-omics data, but beyond simulation studies for imputation accuracy, the impact of missing

data and the choice of methods in specific contexts, such as molecular xQTL analysis, is not well-understood. This work demonstrates, through both simulation and data applications, the significant impact of missing data handling on xQTL discoveries. We recommend a Bayesian matrix factorization approach, *FLASH*, as a default choice for imputing missing data in certain types of multi-omics studies including proteomics, metabolomics and methylation.

In our in-depth comparison of the selected seven imputation methods, we evaluated performance across three multi-omics data types with varying sample sizes, number of features, and missing patterns (MAR and MNAR). Our findings align with previously published benchmark on methylation data, which advised against using KNN for imputation and suggested regression-based imputation as a powerful strategy [27]. The pattern of missingness appears less critical than the nature of the data (types and size of multi-omics data). For xQTL studies, we found that the performance in power is consistent with accuracy. However, methods like *MeanImpute*, *LOD*, and sometimes KNN_{10}, showed inflated type I errors, potentially leading to spurious xQTL signals in practice. In real data analysis, we examined xQTL in multi-omics data from three sources, finding that the choice of imputation method can lead to different genes being identified. In all scenarios and datasets, *FLASH* consistently outperformed other methods in imputation accuracy, power for xQTL discovery, calibration of FDR control, and attained the highest xQTL discovery replication rate. Notably, inferences drawn from *FLASH* were more interpretable than those from other methods, shedding lights into the inherent hidden structures in high-dimensional multi-omics datasets.

The development of new approaches — *MissXGB* and an extension to *FLASH*— was motivated by the need to efficiently analyze large-scale datasets in light of initial promising results from the benchmark. Both *MissForest* and *FLASH*, while more accurate compared to other methods, are slow for large datasets. For example, for the ROSMAP methylation dataset with 721 individuals and 450K CpGs, it takes 7 days for *MissForest* and 2 days for *FLASH* implemented in R package `flashier`) to complete the analysis. Our new *MissXGB* approach and extension to *FLASH* are much faster while maintaining comparable accuracy. In particular, *FLASH*, as a novel method in the context of multi-omics data imputation, is highly recommended for practical applications.

Our work, while comprehensive, has several limitations that may require further investigation. Firstly, none of the methods or benchmarks we used take advantage of information across different multi-omics modalities, especially when data for the same set of individuals is available, as in the case of the ROSMAP and MSBB datasets in our study with RNA-seq, methylation and proteomics features measured. More sophisticated methods might offer improved results in such scenarios [13], though we did not test their practical performance. Secondly, our analysis was confined to proteomics, metabolomics, and methylation data. We did not assess the impact of missing data and imputation methods on other xQTL data types, particularly single cell or single nucleotide RNA-seq, though there are existing benchmarks and recommendations for these data types [28]. Our framework for evaluating xQTL power and FDR could be adapted to these benchmarks to determine whether certain methods can improve xQTL discoveries. Thirdly, the imputation methods we employed are generic. There is potential for future development of methods tailored to specific data types, for example matrix factorization methods — possibly by extending *FLASH*— for functional data such as methylation. We did not explore domain-specific approaches, such as penalized functional regression, which might offer better results for methylation imputation [29]. Finally, while our benchmark design is sophisticated and includes both simulation and real-data applications, the methods we considered are limited to those that are popular and computationally feasible. In particular, we did not systematically evaluate deep learning approaches, which might be effective when executed properly [30,31]. In fact, we did implement and benchmark a variational auto-encoder approach [32], but found its performance to be highly sensitive to parameter tuning and generally less satisfactory than simpler methods, leading us to exclude it from our in-depth comparison.

Implementation of our benchmark outlined in Figure 1 is available at `https://github.com/zq2209/omics-imputation-paper`. The R package *MissXGB* is available at `https://github.com/zq2209/missXGB`. Our extension to *FLASH* is available at `https://github.com/willwerscheid/flashier`. Imputation pipeline for real-world data analysis implementing seven methods can be found at `https://cumc.github.io/xqtl-pipeline/code/data_preprocessing/phenotype/phenotype_imputation.html`.

4. Material and Methods

4.1. Multi-omics missing data imputation methods

In this manuscript, we focus on several key methods, including *MeanImpute*, *LOD*, KNN_{10}, *SoftImpute*, and *MissForest*, each with different parameterizations as detailed in Table S3. Additionally, we introduce two new approaches, *MissXGB* and *FLASH*, detailed in the subsequent sections.

4.1.1. The MissXGB algorithm

We developed the *MissXGB* algorithm and software package in `R`, which follows from the *MissForest* [8] framework but using an XGBoost model trained on the observed portions of a dataset to predict missing values. Specifically, a data matrix $D_{n\times p}$ can be divided into four parts based on any given j-th variable D_j (a length-n vector): Y_{obs}, Y_{miss}, X_{obs}, and X_{miss}. Here, Y_{obs} represents the observed values of D_j, while Y_{miss} denotes its unobserved values. Variables other than j in the dataset are partitioned into X_{obs} amd X_{miss} for observed and missing values respectively. Overall, for each column X_{miss}, a XGBoost model is trained with response Y_{obs} and predictors X_{obs}. The trained model is then applied to predict Y_{miss} with X_{miss}. The iteration will stop as soon as the difference between the newly imputed data matrix and the previous one increases for the first time. Algorithm 1 outlines the *MissXGB* implementation.

Algorithm 1 *MissXGB* Algorithm

Require: The $n \times p$ matrix D; stopping criteria δ_s
Ensure: Y_{obs}, Y_{miss}, X_{obs}, and X_{miss}
  Sort columns of D w.r.t increasing amount of missing values
  Initialized missing values by *MeanImpute*
  while not δ_s **do**
    X^{old}: the previous imputed dataset
    for s in p **do**
      Train an XGBoost model: $Y_{obs} \sim X_{obs}$
      Predict Y_{miss} using X_{miss}
      Update X^{new} using Y_{miss}
    end for
    update $\delta = \frac{\sum_n (X^{new} - X^{old})^2}{\sum_n (X^{new})^2}$
  end while

4.1.2. Extension to FLASH

FLASH— the Factors and Loadings by Adaptive SHrinkage — is an Empirical Bayes Matrix Factorization (EBMF) model that can be applied to perform low-rank approximation for high-dimensional data. We apply *FLASH* to identify hidden structures in datasets and use them to impute missing values. Unlike standard matrix factorization methods like PCA or SVD, *FLASH* offers a more versatile Empirical Bayes approach. This flexibility allows the model to handle various levels of sparsity in the loadings (L) and factors (F). One advantages is its adaptive shrinkage feature, which automatically selects the number of factors for L and F, saving users from having to manually choose and adjust this parameter.

4.2. Numerical Study on imputation accuracy

4.2.1. Simulating realistic missing patterns in multi-omics data

In our study, we assessed the performance of different methods using simulated datasets with missing values categorized as either completely at random (MCAR) or missing not at random (MNAR) [33–35]. The MNAR simulations were crafted to mirror missing patterns found in actual multi-omics data. We implemented a Bayesian approach to create these missing values, taking into account both the observed data and the identified missing patterns. By clustering samples and features based on their

missing profiles, we identified major patterns of missingness in data. The probability of occurrence of missing patterns was used as a prior. This, in conjunction with the observed probability of data being missing, enabled us calculate the probability of missingness taking into consideration of information both from other variables (common in missing at random, MAR, scenarios) and from the variable itself (a characteristic of MNAR situations).

To implement this, consider data matrix D with dimensions $n \times p$. This matrix includes n observations and p features, and contains the missing data patterns we aim to capture using simulated data. To create generative models from these patterns, we characterize them using two steps of clustering. First, we group the p features into g clusters using a K-means algorithm. For instance, in methylation data, we use $g = 100$. This gives us a new matrix of size $n \times g$, showing the average missingness for each cluster of features. Next, we cluster this matrix by samples using hierarchical clustering with a cutree algorithm [36], forming k sample clusters. In the case of methylation data, we set $k = 10$. This creates a total of $g \times k$ unique clusters, each representing a different missing data pattern.

To apply the derived missing pattern cluster to a new dataset D', we first assign each individual in D' to one of the k sample clusters based on the relative frequency of occurrence of the sample cluster in dataset D. Then, for each of the g feature clusters, we compute the probability of a data point from cluster (k, g) being missing as

$$P(m_{kg} = 1|x_{kg}) = \frac{P(x_{kg}|m_{kg} = 1) \times P(m_{kg} = 1)}{P(x_{kg})}$$

where $P(m_{kg} = 1)$ is the probability of missing data in cluster (k, g), estimated by the proportion of missingness in this cluster. $P(x_{kg})$ is the probability density of values in cluster (k, g), which can be approximated from the observed data. $P(x_{kg}|m_{kg} = 1)$ is the probability density of missing values, approximated using features with higher missing rates within the cluster, defined as those having missing rates above the median for features with at least one missing value. To compute $P(x_{kg})$ and $P(x_{kg}|m_{kg} = 1)$, we consider the area under the curve (AUC) of the empirical density function, within one standard deviation around the mean values of these variables, i.e., AUC within the range of $E(x_{kg}) \pm \sqrt{Var(x_{kg})}$ and $E(x_{kg}|m_{kg} = 1) \pm \sqrt{Var(x_{kg}|m_{kg} = 1)}$, respectively. This process is iteratively applied to each feature within every cluster, until a sample is simulated to accurately reflects the intended missing pattern for the features. Other samples in D' can be simulated similarly.

4.2.2. Imputation efficacy metrics

In our analysis, we evaluated the effectiveness of various imputation methods using three key metrics: normalized root mean square error (NRMSE) [37], normalized mean absolute error (NMAE) [38], and the squared Pearson correlation coefficient (R^2) that we computed at feature level. NRMSE, defined as the root mean square error normalized by the standard deviation of the observed value of the feature, is calculated as $NRMSE = \frac{\sqrt{\sum_{i=1}^{N}(\hat{x}_i - x_i)^2}}{\sqrt{Var(x)}}$ where N is the number of observations with missing value, x is the observed values, and $\hat{x}$ is the predicted values. NMAE, the mean absolute error normalized by the range of the observed value, is given by $NMAE = \frac{\sum_{i=1}^{N}|\hat{x}_i - x_i|}{\max(x) - \min(x)}$. The R^2 metric measures the linear correlation between observed and predicted values, represented by the squared Pearson correlation, defined as $R^2 = \left(\frac{\sum_{i=1}^{N}(\hat{x}_i - \bar{x})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{N}(\hat{x}_i - \bar{x})^2 \sum_{i=1}^{N}(x_i - \bar{x})^2}} \right)^2$.

4.3. Numerical study on xQTL discovery

4.3.1. Simulate genotype and molecular phenotype association

To accurately capture the molecular phenotype structure found in real multi-omics data, we devised a reverse simulation approach generating genotype of variants in cis-regulatory regions in linkage disequilibrium (LD) with each other, and their association with molecular features observed in multi-omics resources previously described. We considered five phenotype datasets, listed in "Data" of Figure 1, to conduct the simulation study. Following basic processing procedures (elaborated in Figure 1), we select subsets of molecular features for proteomics and metabolomics data requiring that 1) features

have no missing entry for MCAR based simulation scenarios; 2) features with less than 5% missing rate in the MNAR scenario. For methylation data containing many CpG sites, we compute PCA of CpGs in each topologically associating domain (TAD)[39] and select up to 10 CpG most correlated to the first PC with $R^2 > 0.75$.

For each molecular feature (a gene, a metabolite or a CpG site), we considered the generative model $y = Xb + Zc + \epsilon$, where X is an $N \times P$ genotype matrix for P variants, b is a $P \times 1$ vector of the effect size on molecular genotype on y, and $\epsilon \sim N(0, \sigma^2)$ is other random effects. After regressing out known and inferred covariates Z from y[40] we generate genotype by using this residual y_{res}[41]. Specifically,

$$X_{sim} = y_{res}\beta + T,$$

where $T \sim MVN_p(0, \Sigma)$ is a multivariate normal distribution with mean 0 and covariance matrix Σ being the LD matrix of genotype. For each molecular phenotype, we randomly drew $M_{block} \in [5, 15]$ LD blocks and $P_m \in [10, 50]$ variants in each LD block for $m = 1, \cdots, M_{block}$. Note that $\sum_{m=1}^{M_{block}} P_m = P$. For m-th LD block, we used the following procedure to generate LD structure:

1. Generate a Barabasi-Albert network (BAN) to mimic LD among variants[42]. If two variants are connected in the BAN, these two variants are considered to be in LD; otherwise, these two variants are considered independent. The variants may be connected only if they are in the same LD block.

2. Let $\Theta = (\theta_{pk})$ be an initial concentration matrix for $p, k = 1, \ldots, P_m$. Here, θ_{pk} is set as 0 if two variants (p, k) are not connected in BAN; θ_{pk} from a uniform distribution on the domain of $[-0.9, -0.1] \cup [0.1, 0.9]$ if two variants (p, k) are connected in BAN.

3. Rescale the non-zero elements in Θ to assure its positive definiteness, that is, we divide each off-diagonal element by λ-fold of the sum of the corresponding row, where $\lambda > 1.5$ is the rescale rate. Then, the rescaled matrix is averaged by its transpose to ensure symmetry.

4. Denote $W = (\omega_{pk})$ as the inverse of Θ after rescaling and averaging. Elements Σ_{pk} in the covariance matrix Σ is determined by $\Sigma_{pk} = \omega_{pk}\sqrt{\omega_{pp}\omega_{kk}}$.

We randomly selected one to five linkage disequilibrium (LD) blocks from M_{block} to establish the cis-windows for xQTL analysis. Within each block, one variant is randomly chosen as the true causal variant, with a fixed effect size $\beta = 1$. The residual variance σ^2 is used to adjust the proportion of variance explained (PVE) by genetic variants for a molecular phenotype. By definition, for a generative model $y_{sim} = y_{res} + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$, PVE $= \frac{Var(y_{res})}{Var(y_{res})+\sigma^2}$. In our reverse simulation model, $y_{sim} = \frac{X_{sim}-T}{\beta} + \epsilon$, leading to PVE $= \frac{Var(y_{res})}{Var(y_{res})+\sigma^2} = \frac{(Var(X_{sim})+1)/\beta^2}{(Var(X_{sim})+1)/\beta^2+\sigma^2}$. For proteomics and metabolomics in our simulation model, we assume a total PVE of 25% (per variant PVE is 5%), while for the methylation study, a 50% PVE[43,44] is assumed (per variant PVE is 10%). This approach allows us to simulate a realistic scenario reflecting the genetic contributions to molecular phenotypes in these studies.

We conducted multiple independent simulations to evaluate power and false discovery rate (FDR). Each molecular feature in the simulation is treated as an independent analysis unit. For the ROSMAP proteomics data, we considered 3,851 features across 3 replicates, which amounts to 11,553 association analysis units for assessing power and FDR. The Knight proteomics dataset, having a much fewer number of features 49, required 10 replicates to ensure a robust evaluation of power and FDR for this particular dataset with a total of 490 analysis units. The ROSMAP metabolomics dataset has 369 features across 5 replicates, amounting to 1845 associations analysis units. We analyzed 1,381 TADs for methylation data, with each TAD containing 10 CpG sites. This results in a total of 13,810 association analysis units to robustly evaluate methylation QTL. We combine these molecular features into matrices $Y = [y_1, y_2, \ldots, y_R]$ and assign missing data to them based on MCAR model with 50% missing rate, to assess a moderate to high missing data scenario.

386 *4.3.2. Statistical analysis for xQTL association*

387 Statistical fine-mapping of the simulated genotype across the three molecular phenotype types was
388 conducted on each replicate using SuSiE [23]. Fine-mapping serves to address multiple testing issues in
389 cis-xQTL associations, and to differentiate causal variants from other variants in LD with them. The
390 outputs of fine-mapping are posterior inclusion probabilities (PIPs) and credible sets each capturing a
391 single causal variants. To evaluate these results, we generated precision-recall (PR) and ROC curves
392 using the R package `ROCR` [45] at various PIP thresholds. Additionally, we assessed the coverage of
393 credible sets to evaluate the false discovery rate for single effects fine-mapped. Further details on these
394 evaluation methods are available in the SuSiE manuscript [23].

References

[1] Gary J. Patti, Oscar Yanes, and Gary Siuzdak. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4):263–269, March 2012. ISSN 1471-0080. doi: 10.1038/nrm3314. URL `http://dx.doi.org/10.1038/nrm3314`.

[2] Meng Song, Jonathan Greenbaum, Joseph Luttrell, Weihua Zhou, Chong Wu, Hui Shen, Ping Gong, Chaoyang Zhang, and Hong-Wen Deng. A review of integrative imputation for multi-omics datasets. *Frontiers in Genetics*, 11, October 2020. ISSN 1664-8021. doi: 10.3389/fgene.2020.570255. URL `http://dx.doi.org/10.3389/fgene.2020.570255`.

[3] Trinh Nguyen, Xiaopeng Bian, David Roberson, Rakesh Khanna, Qingrong Chen, Chunhua Yan, Rowan Beck, Zelia Worman, and Daoud Meerzaman. Multi-omics pathways workflow (mopaw): An automated multi-omics workflow on the cancer genomics cloud. *Cancer Informatics*, 22, January 2023. ISSN 1176-9351. doi: 10.1177/11769351231180992. URL `http://dx.doi.org/10.1177/11769351231180992`.

[4] Zhiguang Huo, Yun Zhu, Lei Yu, Jingyun Yang, Philip De Jager, David A. Bennett, and Jinying Zhao. Dna methylation variability in alzheimer's disease. *Neurobiology of Aging*, 76:35–44, April 2019. ISSN 0197-4580. doi: 10.1016/j.neurobiolaging.2018.12.003. URL `http://dx.doi.org/10.1016/j.neurobiolaging.2018.12.003`.

[5] Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *J. Mach. Learn. Res.*, 16(1):3367–3402, jan 2015. ISSN 1532-4435.

[6] Julie Josse and François Husson. missmda: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1), 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i01. URL `http://dx.doi.org/10.18637/jss.v070.i01`.

[7] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-omics factor analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol*, 14(6):e8124, 2018.

[8] Daniel J. Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL `https://doi.org/10.1093/bioinformatics/btr597`.

[9] Akbar K Waljee, Ashin Mukherjee, Amit G Singal, Yiwei Zhang, Jeffrey Warren, Ulysses Balis, Jorge Marrero, Ji Zhu, and Peter DR Higgins. Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, 3(8):e002847, August 2013. ISSN 2044-6055. doi: 10.1136/bmjopen-2013-002847. URL `http://dx.doi.org/10.1136/bmjopen-2013-002847`.

[10] Simon Davis, Connor Scott, Janina Oetjen, Philip D. Charles, Benedikt M. Kessler, Olaf Ansorge, and Roman Fischer. Deep topographic proteomics of a human brain tumour. *Nature Communications*, 14(1), November 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-43520-8. URL `http://dx.doi.org/10.1038/s41467-023-43520-8`.

[11] Philip L. De Jager, Yiyi Ma, Cristin McCabe, Jishu Xu, Badri N. Vardarajan, Daniel Felsky, Hans-Ulrich Klein, Charles C. White, Mette A. Peters, Ben Lodgson, Parham Nejad, Anna Tang, Lara M. Mangravite, Lei Yu, Chris Gaiteri, Sara Mostafavi, Julie A. Schneider, and David A. Bennett. A multi-omic atlas of the human frontal cortex for aging and alzheimer's disease research. *Scientific Data*, 5(1), August 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.142. URL http://dx.doi.org/10.1038/sdata.2018.142.

[12] Ben Omega Petrazzini, Hugo Naya, Fernando Lopez-Bello, Gustavo Vazquez, and Lucía Spangenberg. Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData Mining*, 14(1), September 2021. ISSN 1756-0381. doi: 10.1186/s13040-021-00274-7. URL http://dx.doi.org/10.1186/s13040-021-00274-7.

[13] Dongdong Lin, Jigang Zhang, Jingyao Li, Chao Xu, Hong-Wen Deng, and Yu-Ping Wang. An integrative imputation method based on multi-omics datasets. *BMC Bioinformatics*, 17(1), June 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1122-6. URL http://dx.doi.org/10.1186/s12859-016-1122-6.

[14] Tommi Välikangas, Tomi Suomi, and Laura L. Elo. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. *Briefings in Bioinformatics*, May 2017. ISSN 1477-4054. doi: 10.1093/bib/bbx054. URL http://dx.doi.org/10.1093/bib/bbx054.

[15] Bernard Ng, Charles C White, Hans-Ulrich Klein, Solveig K Sieberts, Cristin McCabe, Ellis Patrick, Jishu Xu, Lei Yu, Chris Gaiteri, David A Bennett, Sara Mostafavi, and Philip L De Jager. An xqtl map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature Neuroscience*, 20(10):1418–1426, September 2017. ISSN 1546-1726. doi: 10.1038/nn.4632. URL http://dx.doi.org/10.1038/nn.4632.

[16] Chengran Yang, Fabiana H. G. Farias, Laura Ibanez, Adam Suhy, Brooke Sadler, Maria Victoria Fernandez, Fengxian Wang, Joseph L. Bradley, Brett Eiffert, Jorge A. Bahena, John P. Budde, Zeran Li, Umber Dube, Yun Ju Sung, Kathie A. Mihindukulasuriya, John C. Morris, Anne M. Fagan, Richard J. Perrin, Bruno A. Benitez, Herve Rhinn, Oscar Harari, and Carlos Cruchaga. Genomic atlas of the proteome from brain, csf and plasma prioritizes proteins implicated in neurological disorders. *Nature Neuroscience*, 24(9):1302–1312, July 2021. ISSN 1546-1726. doi: 10.1038/s41593-021-00886-6. URL http://dx.doi.org/10.1038/s41593-021-00886-6.

[17] Yun Ju Sung, Chengran Yang, Joanne Norton, Matt Johnson, Anne Fagan, Randall J. Bateman, Richard J. Perrin, John C. Morris, Martin R. Farlow, Jasmeer P. Chhatwal, Peter R. Schofield, Helena Chui, Fengxian Wang, Brenna Novotny, Abdallah Eteleeb, Celeste Karch, Suzanne E. Schindler, Herve Rhinn, Erik C. B. Johnson, Hamilton Se-Hwee Oh, Jarod Evert Rutledge, Eric B. Dammer, Nicholas T. Seyfried, Tony Wyss-Coray, Oscar Harari, and Carlos Cruchaga. Proteomics of brain, csf, and plasma identifies molecular signatures for distinguishing sporadic and genetic alzheimer's disease. *Science Translational Medicine*, 15(703), July 2023. ISSN 1946-6242. doi: 10.1126/scitranslmed.abq5923. URL http://dx.doi.org/10.1126/scitranslmed.abq5923.

[18] Wei Wang and Matthew Stephens. Empirical bayes matrix factorization, 2018. URL https://arxiv.org/abs/1802.06931.

[19] Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, June 1987. ISBN 9780470316696. doi: 10.1002/9780470316696. URL http://dx.doi.org/10.1002/9780470316696.

[20] Sik-Yum Lee. Handbook of latent variable and related models. 2007. URL https://api.semanticscholar.org/CorpusID:119011975.

[21] Huimin Wang, Jianxiang Tang, Mengyao Wu, Xiaoyu Wang, and Tao Zhang. Application of machine learning missing data imputation techniques in clinical decision making: taking the discharge assessment of patients with spontaneous supratentorial intracerebral hemorrhage as an

example. *BMC Medical Informatics and Decision Making*, 22(1), January 2022. ISSN 1472-6947. doi: 10.1186/s12911-022-01752-6. URL http://dx.doi.org/10.1186/s12911-022-01752-6.

[22] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785.

[23] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5):1273–1300, July 2020. ISSN 1467-9868. doi: 10.1111/rssb.12388. URL http://dx.doi.org/10.1111/rssb.12388.

[24] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T. Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, December 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv722. URL http://dx.doi.org/10.1093/bioinformatics/btv722.

[25] Amaro Taylor-Weiner, François Aguet, Nicholas J. Haradhvala, Sager Gosai, Shankara Anand, Jaegil Kim, Kristin Ardlie, Eliezer M. Van Allen, and Gad Getz. Scaling computational genomics to millions of individuals with gpus. *Genome Biology*, 20(1), November 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1836-7. URL http://dx.doi.org/10.1186/s13059-019-1836-7.

[26] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, July 2003. ISSN 1091-6490. doi: 10.1073/pnas.1530509100. URL http://dx.doi.org/10.1073/pnas.1530509100.

[27] Pietro Di Lena, Claudia Sala, Andrea Prodi, and Christine Nardini. Methylation data imputation performances under different representations and missingness patterns. *BMC Bioinformatics*, 21 (1), June 2020. ISSN 1471-2105. doi: 10.1186/s12859-020-03592-5. URL http://dx.doi.org/10.1186/s12859-020-03592-5.

[28] Wenpin Hou, Zhicheng Ji, Hongkai Ji, and Stephanie C. Hicks. A systematic evaluation of single-cell rna-sequencing imputation methods. *Genome Biology*, 21(1), August 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02132-x. URL http://dx.doi.org/10.1186/s13059-020-02132-x.

[29] *Epigenome-Wide Association Studies: Methods and Protocols*. Springer US, 2022. ISBN 9781071619940. doi: 10.1007/978-1-0716-1994-0. URL http://dx.doi.org/10.1007/978-1-0716-1994-0.

[30] Lei Huang, Meng Song, Hui Shen, Huixiao Hong, Ping Gong, Hong-Wen Deng, and Chaoyang Zhang. Deep learning methods for omics data imputation. *Biology*, 12(10):1313, October 2023. ISSN 2079-7737. doi: 10.3390/biology12101313. URL http://dx.doi.org/10.3390/biology12101313.

[31] Mingon Kang, Euiseong Ko, and Tesfaye B Mersha. A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1), November 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab454. URL http://dx.doi.org/10.1093/bib/bbab454.

[32] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

[33] Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. 4. ISSN 2624-909X. URL https://www.frontiersin.org/articles/10.3389/fdata.2021.693674.

[34] Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. 88(15):2909–2930. ISSN 0094-9655. doi: 10.1080/00949655.2018.1491577. URL https://doi.org/10.1080/00949655.2018.1491577. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00949655.2018.1491577.

[35] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: an optimization approach. 18(1):7133–7171. ISSN 1532-4435.

[36] R. Becker. *The New S Language*. CRC Press. ISBN 978-1-351-09188-6. Google-Books-ID: 30paDwAAQBAJ.

[37] S.A. Otto. How to normalize the RMSE. URL `https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/`.

[38] Thi-Thu-Hong Phan, Émilie Poisson Caillault, Alain Lefebvre, and André Bigand. Dynamic time warping-based imputation for univariate time series data. 139:139–147. ISSN 0167-8655. doi: 10.1016/j.patrec.2017.08.019. URL `https://www.sciencedirect.com/science/article/pii/S0167865517302751`.

[39] Sudha Rajderkar, Iros Barozzi, Yiwen Zhu, Rong Hu, Yanxiao Zhang, Bin Li, Ana Alcaina Caro, Yoko Fukuda-Yuzawa, Guy Kelman, Adyam Akeza, Matthew J. Blow, Quan Pham, Anne N. Harrington, Janeth Godoy, Eman M. Meky, Kianna von Maydell, Riana D. Hunter, Jennifer A. Akiyama, Catherine S. Novak, Ingrid Plajzer-Frick, Veena Afzal, Stella Tran, Javier Lopez-Rios, Michael E. Talkowski, K. C. Kent Lloyd, Bing Ren, Diane E. Dickel, Axel Visel, and Len A. Pennacchio. Topologically associating domain boundaries are required for normal genome function. *Communications Biology*, 6(1), April 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04819-w. URL `http://dx.doi.org/10.1038/s42003-023-04819-w`.

[40] Hidden factor analysis. `https://cumc.github.io/xqtl-pipeline/code/data_preprocessing/covariate/covariate_hidden_factor.html`.

[41] Xuewei Cao, Ling Zhang, Md Khairul Islam, Mingxia Zhao, Cheng He, Kui Zhang, Sanzhen Liu, Qiuying Sha, and Hairong Wei. Tgpred: efficient methods for predicting target genes of a transcription factor by integrating statistics, machine learning and optimization. *NAR genomics and bioinformatics*, 5(3):lqad083, 2023.

[42] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[43] Margaret Gatz, Chandra A. Reynolds, Laura Fratiglioni, Boo Johansson, James A. Mortimer, Stig Berg, Amy Fiske, and Nancy L. Pedersen. Role of genes and environments for explaining alzheimer disease. *Archives of General Psychiatry*, 63(2):168, February 2006. ISSN 0003-990X. doi: 10.1001/archpsyc.63.2.168. URL `http://dx.doi.org/10.1001/archpsyc.63.2.168`.

[44] Rebecca Sims, Matthew Hill, and Julie Williams. The multiplex model of the genetics of alzheimer's disease. *Nature Neuroscience*, 23(3):311–322, February 2020. ISSN 1546-1726. doi: 10.1038/s41593-020-0599-5. URL `http://dx.doi.org/10.1038/s41593-020-0599-5`.

[45] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, August 2005. ISSN 1460-2059. doi: 10.1093/bioinformatics/bti623. URL `http://dx.doi.org/10.1093/bioinformatics/bti623`.

Supplementary Figures

We validate our realistic missing generation approach on real data. The following is the heatmap of the observed and generated missing patterns on the data.

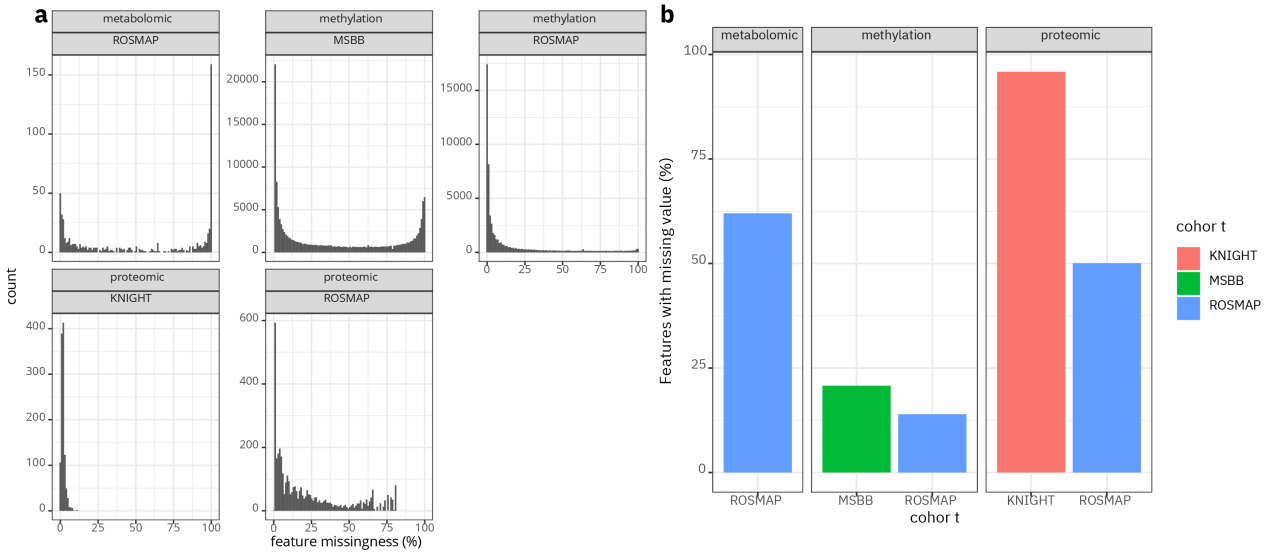

Fig. S1. missing rate distribution. Panel a-b summarize the missing rate distribution for datasets. Pabel a is the summary of distribution of missing rate for features across datasets. Panel b summarize the proportion of features that have at least one missing entry across datasets.

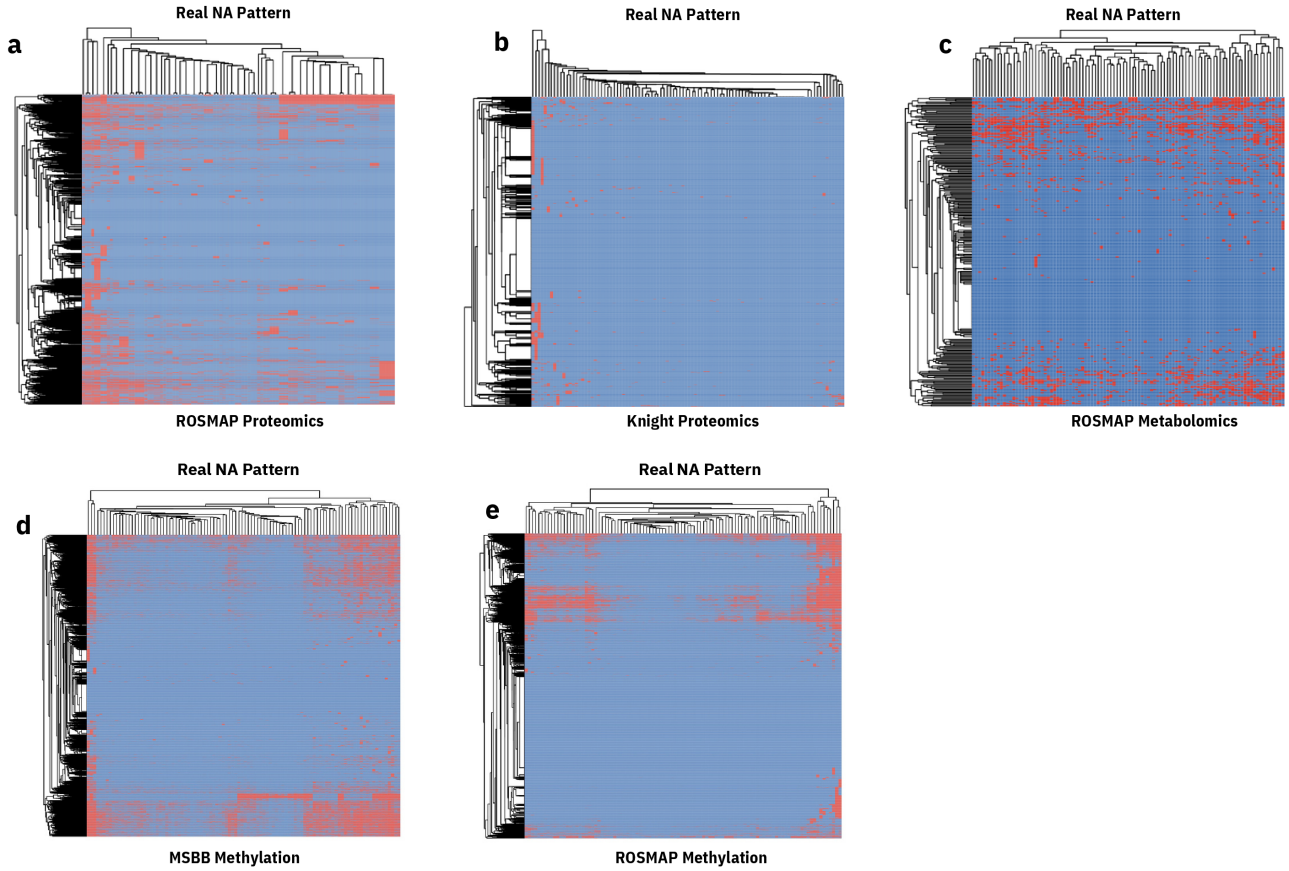

Fig. S2. Real Missing Pattern. The heatmap summarize the onserved missing for each datasets. Panel a is the observed missing pattern for ROSMAP proteomics, panal b for Knight proteomics. Panel c summarize the missing pattern for ROSMAP metabolomics. And panel d-e are missing pattern for MSBB and ROSMAP methylation.

Supplementary Tables

Supplementary Table S1: Multi-omics Application Literature Involving Imputation Methods
Supplementary Table S2: Methodology and Benchmarking Literature for Imputation Methods
Supplementary Table S3: List of Imputation Methods Evaluated
Supplementary Table S4: Sensitivity of Tuning Parameter Settings in Imputation Methods