Imputing Gene Expression in Uncollected Tissues Within and Beyond GTEx

Jiebiao Wang,[1] Eric R. Gamazon,[2,3] Brandon L. Pierce,[1] Barbara E. Stranger,[4,5] Hae Kyung Im,[4] Robert D. Gibbons,[1] Nancy J. Cox,[2] Dan L. Nicolae,[4,6] and Lin S. Chen[1,*]

Gene expression and its regulation can vary substantially across tissue types. In order to generate knowledge about gene expression in human tissues, the Genotype-Tissue Expression (GTEx) program has collected transcriptome data in a wide variety of tissue types from post-mortem donors. However, many tissue types are difficult to access and are not collected in every GTEx individual. Furthermore, in non-GTEx studies, the accessibility of certain tissue types greatly limits the feasibility and scale of studies of multi-tissue expression. In this work, we developed multi-tissue imputation methods to impute gene expression in uncollected or inaccessible tissues. Via simulation studies, we showed that the proposed methods outperform existing imputation methods in multi-tissue expression imputation and that incorporating imputed expression data can improve power to detect phenotype-expression correlations. By analyzing data from nine selected tissue types in the GTEx pilot project, we demonstrated that harnessing expression quantitative trait loci (eQTLs) and tissue-tissue expression-level correlations can aid imputation of transcriptome data from uncollected GTEx tissues. More importantly, we showed that by using GTEx data as a reference, one can impute expression levels in inaccessible tissues in non-GTEx expression studies.

Introduction

Studies of gene expression in peripheral whole blood, skin, liver, and other tissues have revealed that gene expression and its regulation depend on cell context.[1] The expression of a given gene can vary substantially across tissue types, and the genetic variants that regulate gene expression—expression quantitative trait loci (eQTLs)[2,3]—can have eQTL effects that also vary across tissue types.[4–7] A careful examination of gene expression across human tissues and within target tissues would not only help to answer a wide range of scientific questions related to transcriptional variation but also inform other fundamental aspects of biology and prioritize therapeutic gene targets in the development of precision medicine.[8] The challenge is that many tissues are not regenerative and are difficult to collect (hereinafter referred to as “inaccessible” tissues). To date, most large-scale gene-expression studies have been conducted with RNA extracted from peripheral-blood cells or their derivatives, such as lymphoblastoid cell lines. The blood samples are generally heterogeneous and contain a mixture of different cell types. The expression in the blood cells might not directly inform the expression and its regulatory mechanisms in other target cell types from other tissues.

The NIH Common Fund’s Genotype-Tissue Expression (GTEx) program has generated rich transcriptome data in a wide variety of human tissue types, as well as genome sequencing data from a large number of post-mortem donors, thus allowing researchers to generate knowledge about gene expression across human tissues and also characterize the regulatory role of genetic variation from both cross-tissue and tissue-specific perspectives.[9–11] In May 2015, GTEx released pilot data including transcriptome measurements in 44 reference human tissue types and sequencing data on 175 donors.[11] The GTEx project provides a unique opportunity to systematically evaluate the relationships among transcriptomes of different tissues and inform the design of future studies of multi-tissue gene expression.

One major challenge of conducting similar types of analyses in studies beyond GTEx is tissue accessibility. Despite the importance of obtaining specific target tissues from additional cohorts of interest, it might be difficult to collect multi-tissue expression data in many studies. For example, the collection of inaccessible tissues from living study participants is neither possible nor ethical, certain samples in some existing expression studies might not be available for additional data collection, or certain samples might have only limited tissue biopsies available, etc. In those cases, it would be desirable if available information on the target samples and the rich resources in GTEx could be harnessed for accurate imputation of the expression data in the uncollected or inaccessible tissues. With multi-tissue imputation, we are able to reanalyze and leverage existing single-tissue expression data or design future multi-tissue expression studies with limited resources. Compared with single-tissue expression data, multi-tissue expression data provide a more comprehensive and systematic view of the underlying biological mechanisms. Moreover, the expression levels of a gene in functionally related tissues often show coordinated expression patterns, reflecting shared developmental and genetic factors. By jointly analyzing expression data from multiple tissue

[1]Department of Public Health Sciences, University of Chicago, Chicago, IL 60637, USA; [2]Division of Genetic Medicine, Department of Medicine, Vanderbilt University and Vanderbilt Genetics Institute, Nashville, TN 37232, USA; [3]Academic Medical Center, University of Amsterdam, Amsterdam 1105 AZ, the Netherlands; [4]Section of Genetic Medicine, University of Chicago, Chicago, IL 60637, USA; [5]Institute for Genomics and Systems Biology, University of Chicago, Chicago, IL 60637, USA; [6]Department of Statistics, University of Chicago, Chicago, IL 60637, USA
*Correspondence: lchen@health.bsd.uchicago.edu

http://dx.doi.org/10.1016/j.ajhg.2016.02.020.

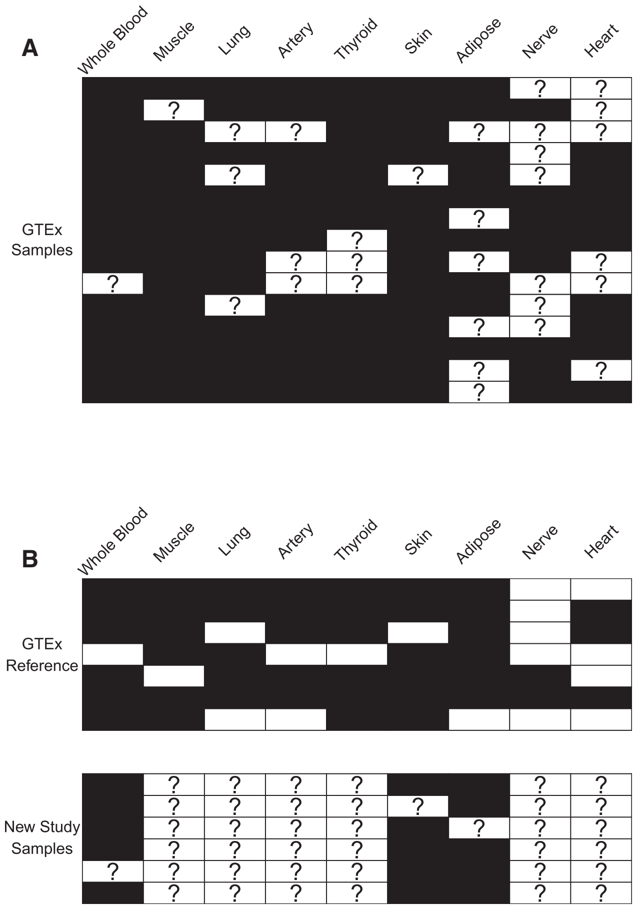


Figure 1. Illustrations of Two Imputation Scenarios
In both scenarios, one can apply the proposed methods to impute the expression in the uncollected or inaccessible tissues of interest. Each row is one individual, and each column is one tissue type. The collected and measured tissues are shown in black, and the uncollected or inaccessible ones are in white. The tissues with question marks are the ones of interest.
(A) Expression in the uncollected GTEx tissues (with question marks) was imputed on the basis of expression in the collected GTEx tissues.
(B) Expression in the uncollected tissues (with question marks), including inaccessible tissue types, was imputed on the basis of collected tissues in a new expression study. GTEx was used as a reference.

types, one can enhance the power to identify biomarkers for complex diseases and traits and facilitate the development of precision medicine.

In this work, we propose harnessing eQTLs and tissue-tissue expression-level correlations for imputing expression data in uncollected or inaccessible tissues. We propose algorithms for multi-tissue imputation based on a mixed-effects model[12] that treats the expression measures from multiple tissues as the outcome and considers as predictors the eQTL genotypes, known covariates, and the estimated tissue-specific top principal components (PCs) of expression data. By borrowing information across genes and across related tissues, the proposed method captures not only the genetic factors influencing gene expression in tissues but also the major developmental and environmental factors. We conducted simulation studies to show the superior imputation performance of the proposed methods over existing imputation approaches[13–19] in multi-tissue expression imputation, as well as the utility of the imputed multi-tissue expression data. Moreover, on the basis of cross-validation (CV) analyses of GTEx pilot data (accession number dbGaP: phs000424.v4.p1), we demonstrated the feasibility of imputing expression in uncollected GTEx tissues (as shown in Figure 1A) and using GTEx data as a reference for imputing expression in inaccessible tissues from samples beyond GTEx (Figure 1B shows an illustration).

Material and Methods

A Mixed-Effects Model for Multi-tissue Imputation

For imputation of gene-expression levels in uncollected or inaccessible tissues, structured information—including the expression levels of the gene of interest in observed tissues, *cis*- (local) and *trans*- (distal) eQTLs, and sample characteristics (gender, age, etc.) shared across genes—is uniquely available in the GTEx data. In GTEx data, we measured the expression levels in multiple tissues from each individual, and the multi-tissue expression measures naturally clustered within individuals.

A natural model to account for these features is a mixed-effects model with expression levels from multiple tissues of a gene as the response, eQTLs and other cross-tissue or tissue-specific covariates as predictors, and random effects (here a random intercept) for each individual:

$$y_{it} = \mu_t + \boldsymbol{\beta}_t^T \mathbf{x}_i + \boldsymbol{\alpha}_t^T \mathbf{c}_i + \gamma_i + \epsilon_{it}. \qquad \text{(Equation 1)}$$

Here, y_{it} is the expression level of a gene in tissue type t $(t = 1, \ldots, T)$ of individual i $(i = 1, \ldots, N)$, μ_t is the tissue-specific mean expression, $\mathbf{x}_i$ is the genotype vector of length K in individual i for K selected eQTLs ($\mathbf{x}_i$ is the same across tissues), $\boldsymbol{\beta}_t$ is a vector of length K and represents the tissue-specific eQTL effects in tissue type t, γ_i is the random intercept for individual i with $\gamma_i \sim N(0, D)$, $\mathbf{c}_i$ is the vector of covariates for individual i with $\boldsymbol{\alpha}_t$ as the corresponding coefficients in tissue type t, and ϵ_{it} is the error term.

In Equation 1, the effect of each eQTL can vary across tissues. Some eQTLs consistently regulate the expression of a gene across multiple tissues and are considered cross-tissue eQTLs, whereas others show eQTL effects only in certain tissue types and are considered tissue specific.[4–6,20] Even for cross-tissue eQTLs, the effect sizes $\boldsymbol{\beta}_t$ can vary by tissue type (similar to an interaction effect of eQTL and tissue type).

To estimate the tissue-specific eQTL effects, we need to estimate a total of $T \times K$ parameters in Equation 1. To reduce the number of parameters, we further employ an adaptive weighting scheme:[21,22] we regress the gene expression in tissue type t on the k^{th} eQTL and let the marginal eQTL effect be the adaptive weight, w_{kt}. This strategy implicitly assumes that the tissue-specific eQTL effects in different tissues in Equation 1 are proportional to the marginal tissue-specific eQTL effects. In the GTEx data, we observed empirical evidence supporting the validity of this assumption (see Supplemental Data for details). The pre-specified adaptive weights in the following model allow us to account for tissue-specific eQTL effects with only one parameter θ_k for the k^{th} eQTL, thereby reducing the total number of parameters for eQTL effects from $T \times K$ to K:

$$y_{it} = \mu_t + \sum_k \theta_k \cdot (w_{kt} x_{ki}) + \boldsymbol{\alpha}_t^T \mathbf{c}_i + \gamma_i + \epsilon_{it}. \qquad \text{(Equation 2)}$$

A Mixed-Model-Based Random-Forest Approach

To obtain the predicted values of y_{it} with weighted genotypes and other covariates as predictors, we propose a mixed-model-based random-forest (MixRF) approach. Random forest is an ensemble learning method that operates by constructing a multitude of regression trees,[23] each of which considers a subset of model predictors and a subset of samples. To learn a regression tree for a continuous outcome on the basis of some predictors, one can employ a recursive binary partitioning algorithm.[24] At each partitioning, the algorithm splits the response variable on the basis of a binary (or dichotomized) predictor in the current node such that the reduction in the sum of squares for values in the node is maximized. The split continues until the tree is too complex or the number of observations in the current node is too small. A regression tree is a non-linear model that predicts the value of a target variable. Predictions based on a single regression tree can be unstable. By aggregating over many regression trees, a random-forest approach intrinsically constitutes a multiple-imputation scheme[16] and provides a more robust prediction that minimizes the overall CV prediction (i.e., imputation) errors.[23–25]

Most existing random-forest approaches[26,27] ignore the clustered data structure. With the proposed MixRF algorithm, we obtain the predictive values by using the following steps: for each gene, we obtain the externally defined eQTLs or select the eQTLs on the basis of the current data and assign the adaptive weight to each eQTL genotype in each tissue type. We set the initial values of $\gamma_i^{(0)} = 0$. Given the estimated random effects at the j^{th} iteration, we build a random forest with $u_{it}^{(j)} = y_{it} - \widehat{\gamma}_i^{(j)}$ as the response and with weighted genotypes in each tissue type and other covariates as predictors, $u_{it}^{(j)} = f(w_{1t}x_{1i}, \ldots, w_{Kt}x_{Ki}, \mathbf{c}_i) + \delta_{it}$, where δ_{it} is the error term. We obtain the predicted value $\widehat{u}_{it}^{(j)}$. In re-estimating the random effects, we let $\omega_{it}^{(j)} = y_{it} - \widehat{u}_{it}^{(j)}$ and fit a linear random-effect model with $\omega_{it}^{(j)} = \gamma_i^{(j)} + \epsilon_{it}$ to obtain the estimated random effect $\widehat{\gamma}_i^{(j)}$. The proposed MixRF algorithm iterates through estimating the random effect γ_i in the linear mixed-effects model[12] and constructing a random forest[26] for the new response variable u_{it} until the change in the likelihood at successive iterations is small (< 0.001). The proposed MixRF often converges quickly in a few iterations, and the prediction is not sensitive to the specified initial values. We summarize MixRF in algorithm 1 in Appendix A.

Our random-forest-based prediction model is a non-linear function of the predictors in Equation 2: $\widehat{y}_{it} = \widehat{f}(w_{1t}x_{1i}, \ldots, w_{Kt}x_{Ki}, \mathbf{c}_i) + \widehat{\gamma}_i$. It can automatically capture the potential non-linear effects of the predictors and the interaction effects among the predictors on the outcome. In the multi-tissue expression GTEx data, we observed that the eQTL effects on gene expression levels could be additive, dominant, or recessive (such that 58%, 38%, or 4% of the eQTL expression pairs better fit an additive, dominant, or recessive eQTL model, respectively).

In addition, we also observed eQTL-eQTL interaction effects and gender-specific eQTLs (gender-eQTL interactions)[28] on many genes. The proposed random-forest-based prediction model would be helpful in capturing those effects and would improve the imputation performance. Moreover, because the random-forest-based prediction model allows higher-order interactions among the predictors, it is more flexible than Lasso-type penalized regression-based predictions and would not induce biased prediction.[29]

An Extension to Capture the Effects of Major Developmental and Environmental Factors in the Imputation

We further propose an extension—MixRF + iPC, where iPC stands for PCs constructed from imputed and observed expression data. Specifically, we propose (1) imputing selected gene-expression levels with multiple eQTLs (~1,000 genes with at least three eQTLs) by using MixRF with adaptively weighted genotypes and other known covariates as predictors, (2) constructing tissue-specific PCs by performing singular value decomposition (SVD) on the combined observed and imputed expression data on the selected genes within each tissue type and keeping the top five PCs for each tissue type, and (3) incorporating the tissue-specific PCs with adaptively weighted genotypes and other known covariates as predictors in MixRF + iPC for imputing or re-imputing gene-expression levels in the genome.

Most of the differences in gene expression among tissues and many of the correlations in gene expression across tissues are driven by the sets of genes that are not expressed in many of the same tissues but rather are expressed in other tissues. Their expression levels are so correlated across tissues not because of shared genetic architecture but because they are completely and invariantly not expressed in so many of the same tissues. Human developmental profiles are invariantly shared within our species, and major developmental information is important information that augments the genetic information. By borrowing information across genes, the top PCs within each tissue type partially capture major developmental factors, as well as the tissue-specific effects of major environmental factors. By incorporating the top PCs from each tissue type as predictors, the extension MixRF + iPC improves the multi-tissue imputation for genes with no eQTLs or low heritability. We summarize MixRF + iPC in algorithm 2 in Appendix A.

In addition to predicting values of multi-tissue expression levels, MixRF and MixRF + iPC provide a measure of imputation quality—the estimated imputation correlation ($\widehat{r}_{\text{imp}}$). It is estimated on the basis of a 10-fold CV analysis of the currently observed data. One splits the data into ten subsamples and each time uses nine subsamples as training data and the rest as testing data. One then applies MixRF to the training data to impute the testing data and repeats this until all the data have been imputed once. In the end, one calculates the correlation between the observed expression levels and the imputed expression levels for each gene.

On the basis of simulation studies, we suggest excluding the imputed expression levels for genes with estimated imputation correlations less than 0.3 in the subsequent analyses, although there is no universal cutoff value for post-imputation exclusion or filtering. The appropriate threshold for a specific analysis might differ.

With parallel computing, imputing 10,000 genes in nine tissue types from about 150 individuals and obtaining the 10-fold CV-based measures of imputation quality could be completed within 30 hr with a 40-node cluster (3.0 GHz Intel Xeon E7 processor) and 16.5 GB of memory.

The overall computation time of MixRF and MixRF + iPC increases linearly with the number of genes and the number of eQTLs and other covariates. The computation complexity of random-forest-based approaches is also dependent on the total number of observed tissues, N_T, a summation of the observed tissues for all individuals. The runtime of MixRF and MixRF + iPC scales with a complexity of $\mathcal{O}(N_T \log N_T)$ in the total number of tissues.[30] The computation is highly parallelizable.

Selecting eQTLs

To obtain the eQTLs for each gene, one can use the reported eQTL lists from other independent data. However, most of the published eQTLs are mapped in whole blood or lymphoblastoid cell lines and might not show eQTL effects in other tissues. In our CV analyses, we did not use the eQTLs reported in the GTEx project.[9,11] Those eQTLs were calculated on the basis of all of GTEx tissues, whereas in each round of our CV analyses, we treated a certain proportion of GTEx tissues as "uncollected," imputed the expression in those tissues, and evaluated the imputation performance. Using eQTLs that were calculated on the basis of all tissues to impute the expression in the "uncollected" tissues would have overestimated the imputation performance.

We propose selecting eQTLs for each gene on the basis of the observed data (for example, the training data in the CV analysis). The selection of eQTLs might affect the predictors used in the imputation and therefore the imputation performance. Nevertheless, the selection can be viewed as a pre-screening of predictors before imputation, and this step will not lead to biased imputation assessment yet will greatly reduce the computational burden. When using GTEx data as a reference for imputing expression in the uncollected tissues from other studies, one can combine the GTEx data with data from non-GTEx samples to obtain the eQTLs used in the imputation.

In each round of CV in our data analyses, we calculated and selected eQTLs on the basis of only the "observed" (i.e., training) data. Given the limited sample size in the GTEx pilot project, we selected only the cross-tissue *cis*- and *trans*-eQTLs and ignored the tissue-specific ones because of low power to detect the latter.

Most of the *cis*-eQTLs are cross-tissue[4–6] and can potentially be replicated in different cell contexts or even across ethnicities.[31,32] To obtain the cross-tissue *cis*-eQTLs, we used MatrixEQTL[33] to calculate the tissue-specific *cis*-eQTL effects, used Stouffer's method[34] to combine the Z statistics from the nine tissue types, and selected the *cis*-eQTLs with Stouffer's p values $< 10^{-6}$. For *trans*-eQTLs, we selected the *trans*-eQTLs with tissue-specific p values ≤ 0.05 in at least eight out of nine tissues. These selected cross-tissue *trans*-eQTLs have Stouffer's p values of less than 10^{-8}. The omission of tissue-specific *trans*-eQTLs in our analysis might have hurt the imputation performance, but this can be improved with the later phase of GTEx data, in which the project will scale up donor collection to 900, and all 44 tissue types will have reasonably large sample sizes.

Ethics Statements

All individuals who donated adipose and muscle biopsies in the IS-MA (insulin-sensitivity muscle-adipose) study[35] provided written informed consent under protocols originally approved by the institutional review board (IRB) at the University of Arkansas for Medical Sciences.

The GTEx project involves recruitment, IRB approval, and consent issues for deceased donors and their families. The collection of tissues from deceased donors is not legally classified as human subjects research under 45 CFR 46 in the Code of Federal Regulations; nevertheless, sites were required to obtain written or recorded verbal authorization from the next-of-kin for deceased donor participation in GTEx.

Processing GTEx Data

Our analyses of GTEx data focused on the expression data from nine tissue types each with $\geq$80 collected samples. We restricted the analyses to the 150 samples with at least four observed tissues, such that in each subsample of the CV data, each individual had at least two observed tissues.

We applied standard data pre-processing and quality-control procedures to both DNA and RNA sequencing data. We considered only the 10,919 genes that were expressed in all nine tissues with a tissue-specific $\log_2$ (mean expression level) significantly greater (according to a one-sided t test) than the $\log_2$ of five read counts. We normalized each gene expression in each tissue and removed the batch effects. For genotype data, we excluded the single-nucleotide variants (SNVs) with minor allele frequencies less than 5% or with p values of Hardy-Weinberg equilibrium test $\leq$ 0.001 and used PLINK[36] to prune the SNVs with a linkage-disequilibrium (LD) threshold of 0.5. After filtering and pruning, we considered 282,295 variants as potential eQTLs in the imputation analyses.

Results

Simulations: Methods Comparison on Imputation Performance

In order to evaluate the imputation performance of our proposed methods and other competing imputation methods, we simulated gene-expression data for 150 individuals and nine tissue types on the basis of Equation 2. We simulated the expression levels of 1,000 genes each with zero, one, two, five, and ten eQTLs. We examined the imputation performance of competing methods when the "heritability" (the percentage of expression variation explained by genetic factors, here the eQTLs) ranged from 0% to 80%, a wide range commonly observed in eQTL studies.[20] We simulated the random intercept $\gamma_i \sim N(0, 1.26^2)$ and the error term $\epsilon_{it} \sim N(0, 3^2)$. Given the SDs of γ_i and ϵ_{it}, the intra-class correlation was 0.15. Additionally, we simulated two cross-tissue covariate $\mathbf{c}_i$ values with various effects on the simulated gene-expression levels. The input parameters for the simulations, including eQTL count, eQTL effect sizes, tissue-tissue expression-level correlations, and covariate effect sizes, reflected what we observed in the real data from the GTEx pilot project.

We randomly treated 30% of all tissues as "uncollected" and set their gene-expression data as "missing." We applied eight imputation methods to the simulated dataset to impute the missing gene-expression data. Those eight competing methods were k-NN,[17] missForest,[16] MICE,[37] linear regression (lm), liner mixed-effects model (lmer),[12] REEMtree,[25] MixRF, and MixRF + iPC. The true eQTLs were used as predictors in the five regression-based methods, lm, lmer,[12] REEMtree,[25] MixRF, and MixRF + iPC. The median of gene-level true imputation correlations of the 1,000 genes was used for evaluating the imputation performance. Note that here, the gene-level true imputation

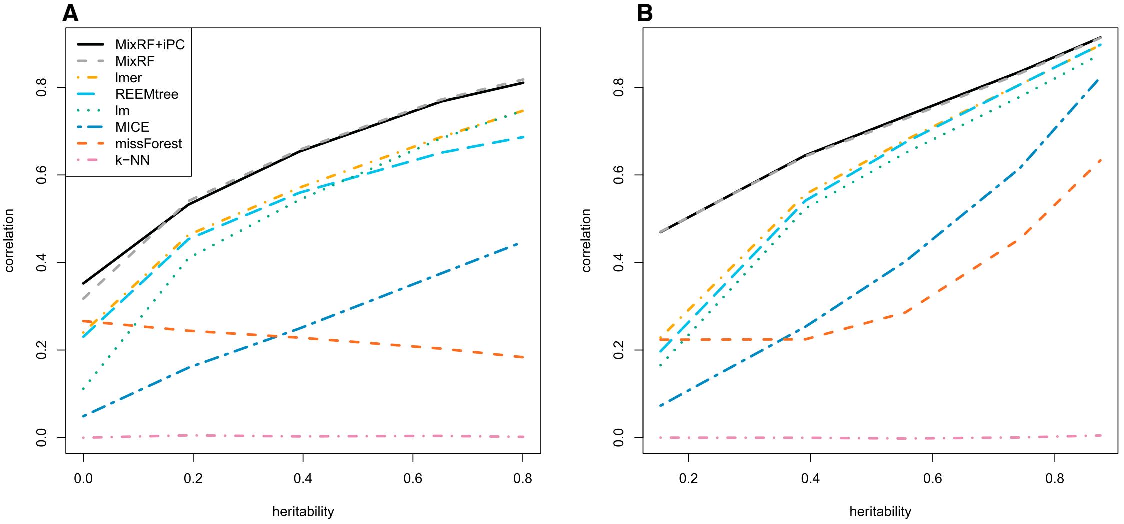


Figure 2. Methods Comparison of Imputation Performance Based on Simulations
Competing methods included k-NN, missForest, MICE, lm, lmer, REEMtree, MixRF, and MixRF + iPC.
(A) We simulated the expression levels of 1,000 genes each with zero, one, two, five, and ten eQTLs.
(B) Each expression level was simulated to be affected by two eQTLs and their interaction.
We simulated 1,000 gene-expression levels each for five varying heritability levels. We used the median of gene-level true imputation correlations of the 1,000 genes to evaluate imputation performance. The difference between the median imputation correlations of MixRF and those of the best alternative approach was highly significant in all scenarios ($p < 2e-16$).

correlation was calculated as the Spearman's correlation between the true and imputed values of a given gene in a specific tissue type. A true correlation is distinct from the estimated imputation correlation based on CV, $\widehat{r}_{\mathrm{imp}}$.

As shown in Figure 2A, our proposed methods MixRF and MixRF + iPC outperformed other imputation methods, and MixRF + iPC showed an advantage over MixRF for imputing gene expression with zero eQTLs. The five regression-based methods incorporated eQTL effects and performed better than other methods. The imputation methods k-NN, missForest, and MICE were designed for single-tissue imputation—whereby selected gene-expression levels are used for imputing the rest of the expression levels from the same tissues—and performed less competitively in the multi-tissue imputation.

In Figure 2B, we simulated another setting, in which each expression level was affected by two eQTLs and an interaction effect between them (a gene-gene interaction effect). In this setting, we simulated 1,000 gene-expression levels each for five varying "heritability" levels from 15% to 87%. Our proposed methods MixRF and MixRF + iPC showed more obvious advantages over other competing methods when the heritability was low. The likely reasons for the observed advantages are that our methods are based on random-forest approaches and are thus capable of capturing the non-linear effects of predictors and their interactions with minor extra computation burdens.

Simulations: Incorporating Imputed Data to Improve the Power to Detect Phenotype-Expression Correlations

When directly collecting certain tissues in a specific cohort is challenging and when resources are available, one can impute expression data on inaccessible tissues by using available information and potentially GTEx as a reference. We argue that the imputed data can be treated as supplemental data or supporting data to enhance the primary analysis on the basis of the observed expression data. To support this claim, we took the expression data on whole blood, adipose tissue, and nerve tissue and the genotype data in the GTEx pilot project and then simulated phenotypes that were correlated (at 0.25 and 0.3) with gene-expression levels in the nerve tissues. We treated 50% of the nerve tissues as "uncollected" and set the expression levels in those tissues as missing.

By applying the proposed MixRF + iPC method to the 10,919 genes in the observed data (with blood, adipose tissue, and 50% nerve tissue) and estimating the imputation correlation for each gene, we obtained 1,537, 762, and 324 genes with estimated imputation correlations ($\widehat{r}_{\mathrm{imp}}$) greater than 0.3, 0.4, and 0.5, respectively. At the significance thresholds of 5% and 10% false-discovery rates (FDRs), we compared the power to detect the phenotypes associated with the nerve expression levels on the basis of (1) only the observed nerve expression data (50% of the complete nerve data), (2) the combined observed and imputed nerve expression data with varying imputation quality ($\widehat{r}_{\mathrm{imp}} \geq$ 0.3, 0.4, and 0.5), and (3) the complete GTEx nerve expression data with 95 samples.

The results are presented in Table 1. Incorporating reasonably imputed data helped to improve the power to detect phenotype-expression correlations even when the phenotype-expression correlations were not strong and/or when the quality of the imputed data was not superb. As the imputation quality improved, the power improvement became more substantial. Analyses based on poorly

Table 1. Power Comparison for Detecting Phenotype-Expression Correlations on the Basis of the Observed, Observed and Imputed, and Complete Expression Data

No. of Genes Passing Estimated Imputation-Correlation ($\widehat{r}_{\text{imp}}$) Thresholds	Phenotype-Expression Correlations	FDR	Observed Data Only	Observed and Imputed Data	Complete Data
1,537 genes ($\widehat{r}_{\text{imp}} \geq 0.3$)	0.25	0.05	0.269	0.377	0.953
		0.1	0.455	0.586	1
	0.3	0.05	0.548	0.738	1
		0.1	0.729	0.898	1
762 genes ($\widehat{r}_{\text{imp}} \geq 0.4$)	0.25	0.05	0.230	0.652	0.933
		0.1	0.391	0.839	1
	0.3	0.05	0.613	0.734	1
		0.1	0.778	0.887	1
324 genes ($\widehat{r}_{\text{imp}} \geq 0.5$)	0.25	0.05	0.454	0.534	1
		0.1	0.688	0.744	1
	0.3	0.05	0.676	0.784	1
		0.1	0.883	0.926	1

More specifically, the three sources of expression data were (1) only the observed nerve expression data (from which 50% of GTEx nerve tissue was missing), (2) the observed and imputed data with varying imputation quality, and (3) the complete GTEx nerve expression data. The significance thresholds were 5% and 10% FDRs. We assessed the power comparison when the phenotype-expression correlations were 0.25 and 0.3 for three groups of genes with estimated imputation correlations of at least 0.3, 0.4, and 0.5 (representing fair, moderate, and good imputation quality, respectively).

imputed genes might not help or might even hurt the power of the analyses. Although only a small proportion of imputed gene-expression levels might be retained in the subsequent analyses after exclusion of the poorly imputed expression levels, those genes are often affected by multiple eQTLs and/or related to other factors in functional pathways and, as such, are often of biological interest.

Analyses of GTEx Data: Imputing Uncollected GTEx Tissues

The GTEx project is collecting 44 human tissue types, and most of them are difficult to access. In the pilot data, only nine tissue types were collected in more than 80 out of 175 donors, and the remainder yielded tissue-specific sample sizes of less than 40. We sought to impute the uncollected GTEx tissues by using MixRF + iPC (Figure 1A). We conducted a 10-fold CV analysis focusing on the nine tissue types to evaluate the imputation performance within GTEx. Specifically, we randomly split the GTEx transcriptome data on the nine tissue types into ten subsamples each containing data on one-tenth of the collected tissues from each tissue type. In each round of CV analysis, we treated one subsample of the transcriptome data as unobserved and uncollected and the other nine subsamples as observed or collected. For each gene, we imputed the unobserved expression levels in uncollected GTEx tissues by using the expression levels in the collected tissues (the imputation scheme is illustrated in Figure 1A). We repeated the exercise for each subsample of data, combined the imputed data, and evaluated the true tissue-specific imputation correlations.

The imputation performance of MixRF was generally comparable with that of its extension, MixRF + iPC, although the latter performed better in imputing gene expression with no eQTLs or in the blood tissue (Table S1). We also compared the true imputation correlations from MixRF + iPC with the standard practice of using blood expression as a surrogate for target-tissue expression (hereafter referred to as "blood surrogate") (Figure 3). The imputation performance of MixRF + iPC largely relied on the heritability and tissue-tissue expression-level correlations for each gene, both of which were directly related to the number of cross-tissue eQTLs. For genes with five or more combined *cis-* and *trans-*eQTLs, the median true imputation correlation was 0.48. For genes with 10+ and 30+ eQTLs, the median true imputation correlation increased to 0.55 and 0.63, respectively. Note that although we used PLINK[36] to perform LD pruning (with a LD threshold of 0.5 and a window size of 50 bp) on the SNVs, moderate LD could still remain among the eQTLs. Among the 10,919 expressed genes that we considered, the genes with 5+, 10+, and 30+ eQTLs in at least one subsample of the CV data numbered 1,065 (9.8%), 465 (4.3%), and 170 (1.6%), respectively.

Generally, expression in whole blood is weakly correlated (with a median correlation of 0.1) with expression in other tissues and is a poor surrogate for the latter. We compared the imputation performance of the proposed methods with that of additional competing imputation methods for genes with different numbers of eQTLs (Table S1). Additionally, we evaluated the sample-level true imputation correlations (Figure S1), which were calculated as the correlations between the observed and imputed gene ranks for each sample in each tissue. The imputed gene ranks could be useful in analyses of gene-set enrichment.[38]

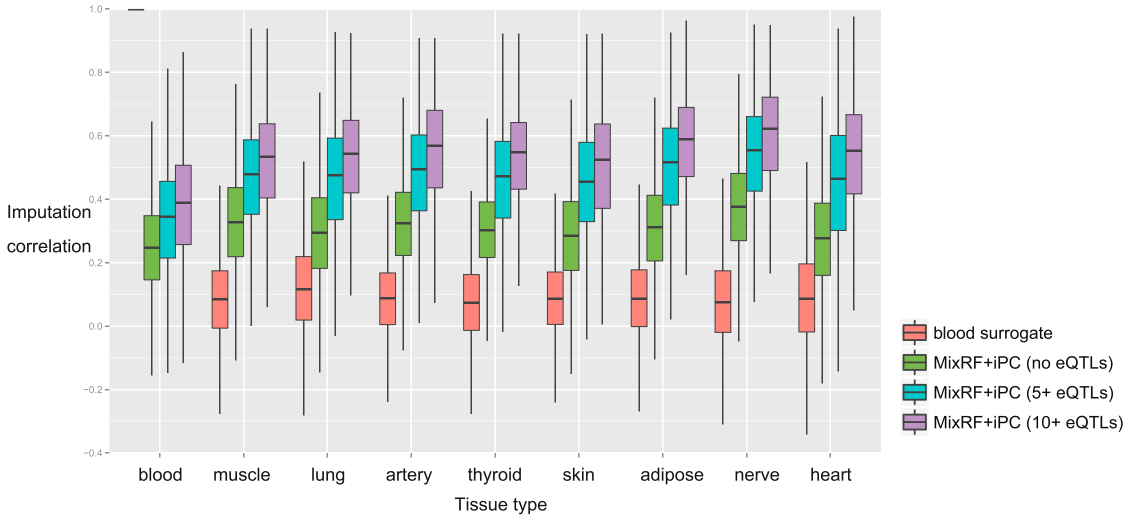


Figure 3. Boxplots of Gene-Level True Imputation Correlation by Tissue Type
The results are based on a 10-fold CV analysis within GTEx tissues. Specifically, we randomly split the GTEx transcriptome data into ten subsamples, each of which contained one-tenth of the collected tissues from each tissue type. In each round of CV analysis, we treated one subsample of the transcriptome data as unobserved and the other nine subsamples as observed. We then imputed the unobserved data. Figure 1A illustrates the imputation scheme in one round of the CV analysis. We repeated the analysis for each subsample of data. Each correlation was calculated as the Spearman's correlation between the observed and combined imputed values of a given gene in the current tissue. We compared the true imputation correlations based on MixRF + iPC for genes with zero, at least five, and at least ten eQTLs with the correlations based on blood surrogate. Note that the eQTLs for each gene can be in moderate LD.

To evaluate the impact of sample size on imputation, we performed a 3-fold CV analysis within GTEx tissues and compared the results with those obtained from the 10-fold CV analysis (Table 2). In each round of the 3-fold and 10-fold CV analyses, two out of three and nine out of ten data subsamples, respectively, were treated as "observed," yielding average tissue-specific sample sizes of 73 and 99, respectively. We found that sample size substantially affected imputation performance largely because sample size substantially affected the power to detect cross-tissue eQTLs. With a 36% sample-size increase in the 10-fold CV analysis and the same significance criteria, we detected 65% more cross-tissue *cis*-eQTLs (8,792 versus 5,332) and 225% more cross-tissue *trans*-eQTLs (12,884 versus 3,976). As a result, the median true imputation correlation across the genome improved from 0.305 to 0.349. Additional simulations are presented in the Supplemental Data to further demonstrate the impact of sample size on imputation. When more GTEx samples become available, we expect further improvement in imputation performance.

Overall, both eQTL and tissue-tissue expression-level correlations play a major role in multi-tissue imputation. The average estimated heritability for expressed genes was reported to be 0.14 ~ 0.26 for different tissue types in other studies,[39,40] which roughly corresponds to an imputation correlation of 0.37 ~ 0.51 if the appropriate SNVs were selected in the imputation. According to our results, the median true imputation correlation based on linear regressions that use only eQTLs as predictors (see "lm" in Table S1) was much lower (~0.2), indicating that the current imputation results could be improved if sample size were to increase and more eQTLs were detected and used in the imputation. Additional comparison of linear-regression and mixed-effects models (Table S1) showed that information on tissue-tissue expression-level correlation helped improve the absolute median imputation correlation by 0.1 ~ 0.3 for genes with at least five or at least ten eQTLs. For genes with no eQTLs, the median imputation correlation with mixed-effects models was nearly 0.3 and was higher than that from using blood expression as a surrogate.

Analyses of GTEx Data: Using GTEx as a Reference to Impute Other Studies

Tissue accessibility often limits the feasibility and scale of multi-tissue expression studies in specific cohorts. Multi-tissue expression imputation would be helpful when direct measurements in specific tissues are limited or not available and when expression data on related tissues are existing or accessible. Incorporating expression in secondary and related tissue types into the primary data might enhance the power to detect differentially expressed genes under different phenotypic conditions and provide insights into disease etiology from a multi-tissue perspective. Multi-tissue imputation could impute expression in the uncollected tissues, which could be used as supplemental data to be combined with the primary observed data in the secondary data analysis. In those imputation scenarios, one can use GTEx as a reference and impute gene expression in the uncollected tissues or tissue types in non-GTEx samples. Figure 1B shows an example of such imputation scenarios.

Table 2. Comparing 10-fold and 3-fold CV Analyses within GTEx Tissues Shows the Impact of Sample Size

	10-fold CV Analysis			3-fold CV Analysis		
Average sample size across tissues	98.9			73.2		
Average no. of *cis*-eQTLs	8,792			5,332		
Average no. of *trans*-eQTLs	12,884			3,976		
Average median true imputation correlation	0.349			0.305		
	No. of Genes	**Median True Imputation Correlation**	**No. of Genes with True Imputation Correlation ≥ 0.5 (%)**	**No. of Genes**	**Median True Imputation Correlation**	**No. of Genes with True Imputation Correlation ≥ 0.5 (%)**
Genes with no eQTLs	9,240	0.307	207 (2.2)	10,063	0.271	88 (0.9)
Genes with one eQTL	5,062	0.338	250 (4.9)	2,521	0.317	62 (2.5)
Genes with two eQTLs	2,486	0.368	228 (9.2)	767	0.355	63 (8.2)
Genes with three eQTLs	1,394	0.386	191 (13.7)	371	0.390	44 (11.9)
Genes with four eQTLs	883	0.412	169 (19.1)	225	0.401	48 (21.3)
Genes with five to nine eQTLs	839	0.430	191 (22.8)	275	0.454	90 (32.7)
Genes with at least ten eQTLs	465	0.521	270 (58.1)	185	0.578	128 (69.2)

Increasing the sample size affected the power to detect cross-tissue eQTLs and thus imputation results. The number of genes with x eQTLs is counted as the number of genes with x eQTLs in at least one subsample of the CV data. For example, genes might have no eQTLs in one or several subsamples of the CV data and have one or two eQTLs in other subsamples of the CV data. We calculated the true imputation correlation for genes with no eQTLs by only considering the subsamples of the CV data in which the gene had no eQTLs. As such, there was overlap among genes with zero, one, or two eQTLs, etc.

We conducted another 10-fold CV analysis to evaluate the feasibility of such imputation. Unlike in the 10-fold CV analysis conducted in the previous section, here we split the GTEx individuals into ten subsamples. In each round of the current CV analysis, we treated nine subsamples of the GTEx individuals as the "GTEx reference" and the other subsample as testing samples from a new study. In the new samples, we only observed the transcriptome data in the three accessible tissues and used the data on the three tissues with GTEx as a reference to impute the expression in the uncollected tissues in the new samples (Figure 1B).

We used MixRF + iPC to evaluate the tissue-specific gene-level true imputation correlations in the six inaccessible tissue types (Figure 4). Blood surrogate achieved a median correlation of only ~0.1. In contrast, even for genes with no eQTLs, MixRF + iPC achieved a median true imputation correlation of 0.17–0.27 in different tissue types. For genes with at least ten eQTLs, the median true imputation correlation increased to ~0.4 across tissue types. The imputation performance was better in nerve tissue than in other tissues in that it achieved a median correlation of 0.37, 0.42, and 0.54 for genes with 5+, 10+, and 30+ eQTLs, respectively. This might be attributable to the relatedness between nerve tissue and adipose tissue and skin or to its reaction to stimuli. We also assessed the sample-level true imputation correlations (Figure S2), and the conclusions were similar.

Additionally, one can also use multi-tissue imputation to build on existing single-tissue expression and eQTL data. One can collect the tissues of interest in a small set of new samples in the specific cohorts as the learning tissues and then use those tissues together with the GTEx reference samples to impute the samples with expression data only on a single tissue and not available for additional data collection.

The multi-tissue imputation strategy can also be used in designing future multi-tissue expression studies in certain populations or ethnicities or with specific phenotypes. One can utilize the GTEx resource and conduct CV analyses on the GTEx tissues. By leveraging tissue availability and predictability, one can select the tissue types that are most relevant and predictive for the target tissue types.

Using GTEx as a Reference in the Presence of Potential Study Heterogeneity and a Validation Analysis

The performance of the proposed multi-tissue imputation methods primarily depends on the predictive ability of eQTLs and the tissue-tissue expression-level correlations. We suggest including a reference-sample indicator variable in the MixRF as a covariate when using GTEx as a reference for imputing other non-GTEx samples with potential study heterogeneity. When the eQTL effects or effects of other covariates are sufficiently different among the GTEx reference and the non-GTEx samples, the interaction terms of the reference indicator and the eQTLs or other covariates will be selected in building the random forest. As such, in the presence of study heterogeneity, the estimation of eQTL effects in the non-GTEx samples will be based primarily on the non-GTEx samples only.

Recent studies have shown that the predictive ability of eQTLs can be replicated across GTEx and other studies,[20] and the expression patterns of many pharmacogenes

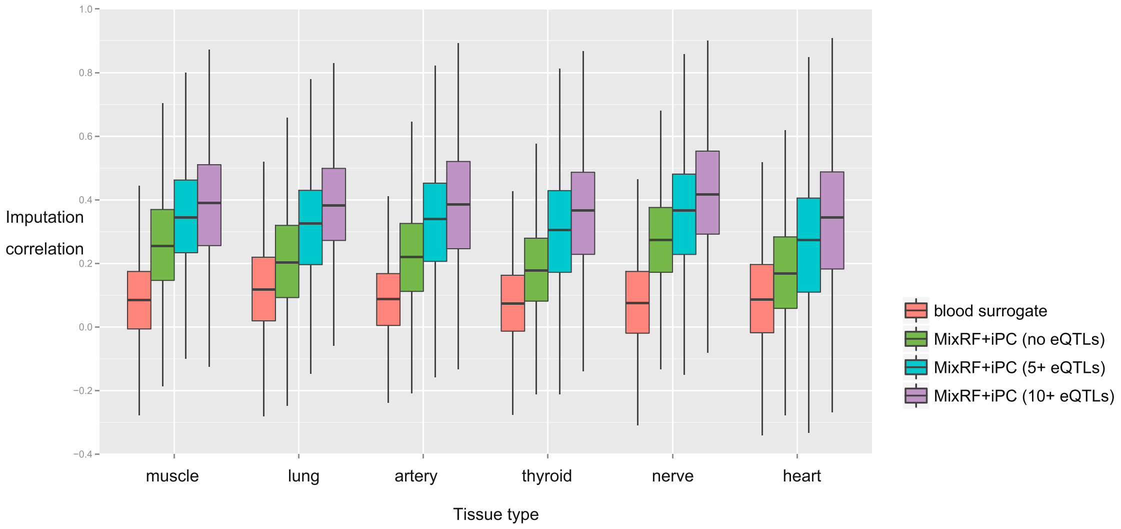


Figure 4. Boxplots of Gene-Level True Imputation Correlation in Inaccessible Tissues in a New Study
The results are based on a 10-fold CV analysis of imputing uncollected tissues in the new samples while using GTEx as a reference. Specifically, we split the GTEx individuals into ten subsamples. In each round of CV analysis, we used nine subsamples as a reference and treated the other subsample as new. With GTEx data as a reference, we imputed the transcriptome data in the inaccessible tissues on the basis of the accessible ones (blood, skin, and adipose) in the new samples. Figure 1B illustrates the imputation scheme in one round of CV analysis. We repeated the analysis for each subsample of data. We compared the true imputation correlations based on MixRF + iPC for genes with zero, at least five, and at least ten eQTLs with the correlations based on blood surrogate.

investigated by the Pharmacogenomics Research Network project can also be validated in the GTEx samples.[41]

To further validate the utility of the proposed methods and of GTEx data as a reference in multi-tissue imputation for non-GTEx samples, we applied MixRF to the IS-MA study on insulin (*INS* [MIM: 176730]) sensitivity (accession number GEO: GSE40234).[35] Fifty-nine samples at the tails of the distribution of insulin sensitivity were selected in the study. The expression levels on adipose and muscle tissues and genotype data are available on those 59 samples. We considered 229 genes with preserved Ensembl IDs in both GTEx and the IS-MA study. Such genes are likely to have completely preserved gene structure across the two datasets. We normalized the expression levels of each gene within each study. We focused on imputing the expression levels of those 229 genes in the muscle tissues from the IS-MA samples.

We compared the performance of the following analyses to impute the muscle-tissue expression levels in the IS-MA samples: (1) imputing with GTEx reference and adipose expression levels and eQTLs from the IS-MA study; (2) imputing with GTEx reference and adipose expression levels from the IS-MA study, but not eQTLs; and (3) imputing on the basis of the eQTLs, but not GTEx as a reference, from the IS-MA study. We calculated the imputation correlations of measured muscle-tissue expression levels and the imputed values on the basis of the three sets of analyses. Figure S3 shows the quantile-quantile plot of the three sets of imputation correlations against the null correlations. Including GTEx as a reference greatly improved the imputation performance, and the mean imputation correlation of those 229 genes according to analysis 1 was 0.313. When imputation was based only on tissue-tissue expression-level correlations (analysis 2) or eQTL genotypes (analysis 3), the imputation correlations substantially deviated from the null correlations. This implies that both tissue-tissue expression-level correlations and eQTL genotypes help in multi-tissue imputation. MixRF with GTEx as a reference combines the two sources of information and improves the overall imputation.

Discussion

The joint analysis of transcriptome data from multiple tissues would enhance the power of analyzing expression data and ultimately improve our understanding of biological mechanisms from a systems perspective. The bottleneck that limits the feasibility and scale of studies of multi-tissue expression is tissue accessibility. When a tissue is not accessible in an individual, the gene-expression levels in that tissue are not available and are considered "missing." We propose algorithms for imputing multi-tissue expression data. The proposed approaches can be used for imputing expression on uncollected tissues in the GTEx project to facilitate downstream analyses and, moreover, for imputing inaccessible tissues in other expression studies while using GTEx as a reference. Different from methods that predict expression levels on the basis of eQTL information,[20] our proposed methods impute multi-tissue expression levels on the basis of eQTLs, tissue-tissue expression-level correlations, and tissue-specific PCs of expression data and harness genetic factors, major developmental biological factors, and environmental factors. Additionally, our MixRF approach captures the dominant and recessive eQTL effects, as well as the interactions among eQTLs, tissue types, and other factors. Most existing single-tissue imputation methods rely on gene-gene correlations, which can be unstable. Our methods outperform existing imputation methods in multi-tissue imputation.

Multi-tissue imputation can be helpful when direct measurements in the desired tissues are uncollected or difficult to collect, and one can use the imputed data as supplement data to support scientific findings from observed data. Within the GTEx project, we can impute the expression in the uncollected tissues and use imputed expression data to enhance the detection of protein QTLs or facilitate the construction of integrative genomics networks. More importantly, by using GTEx as a reference, we can potentially impute inaccessible tissues in other expression studies, impute and recapitalize on existing data, design effective multi-tissue expression studies in other populations or ethnicities, and further inform disease-related tissues. We anticipate that our multi-tissue imputation method will initiate research on methods development and enable the discovery of scientific findings with the use of multi-tissue expression data within and beyond the GTEx project.

One caveat of the current analyses is that we used only cross-tissue eQTLs in the imputation. The sample size in the GTEx pilot data limits the power to detect tissue-specific eQTLs. We believe that a larger sample size in the later phase of GTEx data will bring increased power to detect both cross-tissue and tissue-specific eQTLs and thereby substantially improve imputation performance. An alternative strategy for selecting eQTLs is to combine the eQTLs reported in other studies, which ideally involve multiple tissue types.

We anticipate that the later phase of GTEx data will bring additional challenges to methods development, e.g., the scalability of the approaches and the selection of the accessible tissues for maximizing imputation accuracy. In addition to enabling multi-tissue imputation, it is desirable to develop methods that account for observed and imputed expression values in the subsequent disease- or trait-related analyses and to enable multi-tissue network and integrative analyses.

Appendix A

Algorithm 1: MixRF, a Mixed-Model-Based Random-Forest Approach for Imputing Multi-tissue Expression

```
1. For each gene, use externally defined eQTLs or select eQTLs on the basis of the currently observed data. Obtain the adaptive weights ($w_{kt}$) for each eQTL in each tissue type.
2. Initialize the random-effects estimate in Equation 2, $\widehat{\gamma}_i^{(0)} = 0$.
3. At the $j^{\text{th}}$ iteration, let $u_{it}^{(j)} = y_{it} - \widehat{\gamma}_i^{(j-1)}$. Build a random forest with $u_{it}^{(j)}$ as the response and weighted genotypes in each tissue and other covariates ($\mathbf{c}_i$) as predictors, $u_{it}^{(j)} = f(w_{1t}x_{1i}, \ldots, w_{Kt}x_{Ki}, \mathbf{c}_i) + \delta_{it}$. Obtain the predicted value $\widehat{u}_{it}^{(j)}$.
4. Let $\omega_{it}^{(j)} = y_{it} - \widehat{u}_{it}^{(j)}$. Fit a linear random-effects-only model with $\omega_{it}^{(j)}$ as the response, $\omega_{it}^{(j)} = \gamma_i^{(j)} + \epsilon_{it}$. Obtain the estimated random effect $\widehat{\gamma}_i^{(j)}$.
5. Iterate through steps 3 and 4 until the change in the likelihood is small.
```

Algorithm 2: MixRF + iPC, a MixRF Extension Incorporating PCs of Expression Data

```
0. Select eQTLs.
1. For each tissue type, construct the top PCs of combined observed and imputed expression data on selected genes to capture unknown sample characteristics and tissue-specific major developmental patterns and environmental effects.
    i. Impute the selected gene-expression levels (here, we imputed the expression levels of ~1,000 genes with at least three eQTLs) by using MixRF with adaptively weighted eQTL genotypes and other known covariates as predictors.
    ii. For each tissue type, perform SVD on the combined observed and imputed data on the selected genes, and keep the top five PCs. (Note that the results based on the top ten PCs are similar.)
2. Apply MixRF to each gene by using gene expression as the response and using adaptively weighted eQTL genotypes, other known covariates, and the constructed tissue-specific PCs from step 1 as predictors.
```

Acknowledgments

We thank the GTEx Consortium. J.W. is supported by R01GM108711. L.S.C. is supported by R01GM108711, R03CA174984, R01MH101820, and U01HG007601.

Received: September 3, 2015
Accepted: February 22, 2016
Published: March 31, 2016

Web Resources

The URLs for data presented herein are as follows:

dbGaP, http://www.ncbi.nlm.nih.gov/gap
Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/
OMIM, http://www.omim.org
R software package for MixRF and MixRF + iPC, https://github.com/randel/MixRF

References

1. Schadt, E.E., Woo, S., and Hao, K. (2012). Bayesian method to predict individual SNP genotypes from gene expression data. Nat. Genet. *44*, 603–608.
2. Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. Science *296*, 752–755.

3. Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nat. Rev. Genet. *7*, 862–872.
4. Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. *9*, e1003486.
5. Torres, J.M., Gamazon, E.R., Parra, E.J., Below, J.E., Valladares-Salgado, A., Wacher, N., Cruz, M., Hanis, C.L., and Cox, N.J. (2014). Cross-tissue and tissue-specific eQTLs: partitioning the heritability of a complex trait. Am. J. Hum. Genet. *95*, 521–534.
6. Li, G., Shabalin, A.A., Rusyn, I., Wright, F.A., and Nobel, A.B. (2016). An empirical Bayes approach for multiple tissue eQTL analysis. arXiv, arXiv:1311.2948, http://arxiv.org/abs/1311.2948.
7. Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science *344*, 519–523.
8. Collins, F.S., and Varmus, H. (2015). A new initiative on precision medicine. N. Engl. J. Med. *372*, 793–795.
9. GTEx Consortium (2013). (The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
10. Keen, J.C., and Moore, H.M. (2015). The Genotype-Tissue Expression (GTEx) Project: Linking Clinical Data with Molecular Analysis to Advance Personalized Medicine. J. Pers. Med. *5*, 22–29.
11. GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
12. Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. Biometrics *38*, 963–974.
13. Celton, M., Malpertuy, A., Lelandais, G., and de Brevern, A.G. (2010). Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics *11*, 15.
14. Liew, A.W.-C., Law, N.-F., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief. Bioinform. *12*, 498–513.
15. Donner, Y., Feng, T., Benoist, C., and Koller, D. (2012). Imputing gene expression from selectively reduced probe sets. Nat. Methods *9*, 1120–1125.
16. Stekhoven, D.J., and Bühlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics *28*, 112–118.
17. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics *17*, 520–525.
18. Brock, G.N., Shaffer, J.R., Blakesley, R.E., Lotz, M.J., and Tseng, G.C. (2008). Which missing value imputation method to use in expression profiles: a comparative study and two selection schemes. BMC Bioinformatics *9*, 12.
19. Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., Sciurba, F.C., and Tseng, G.C. (2014). Missing value imputation in high-dimensional phenomic data: imputable or not, and how? BMC Bioinformatics *15*, 346.
20. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., Nicolae, D.L., Cox, N.J., and Im, H.K.; GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
21. Tukey, J.W. (1949). One degree of freedom for non-additivity. Biometrics *5*, 232–242.
22. Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U., and Wacholder, S. (2006). Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. Am. J. Hum. Genet. *79*, 1002–1016.
23. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). Classification and Regression Trees (Chapman and Hall/CRC).
24. Friedman, J. (1977). A recursive partitioning decision rule for nonparametric classification. IEEE Trans. Comput. *4*, 404–408.
25. Sela, R.J., and Simonoff, J.S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. Mach. Learn. *86*, 169–207.
26. Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.
27. Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest (R news), pp. 18–22. https://www.r-project.org/doc/Rnews/Rnews_2002-3.pdf.
28. Dimas, A.S., Nica, A.C., Montgomery, S.B., Stranger, B.E., Raj, T., Buil, A., Giger, T., Lappalainen, T., Gutierrez-Arcelus, M., McCarthy, M.I., and Dermitzakis, E.T.; MuTHER Consortium (2012). Sex-biased genetic effects on gene regulation in humans. Genome Res. *22*, 2368–2375.
29. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. R. Statist. Soc. B *58*, 267–288.
30. Stephan, J., Stegle, O., and Beyer, A. (2015). A random forest approach to capture genetic effects in the presence of population structure. Nat. Commun. *6*, 7432.
31. Pierce, B.L., Tong, L., Chen, L.S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H.-J., Franke, L., et al. (2014). Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. PLoS Genet. *10*, e1004818.
32. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. PLoS Genet. *8*, e1002639.
33. Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353–1358.
34. Stouffer, S.A., Suchman, E.A., Devinney, L.C., Star, S.A., and Williams, R.M., Jr. (1949). The American Soldier: Adjustment during Army Life, Studies in Social Psychology in World War II, *Vol. 1* (Princeton University Press).
35. Elbein, S.C., Gamazon, E.R., Das, S.K., Rasouli, N., Kern, P.A., and Cox, N.J. (2012). Genetic risk factors for type 2 diabetes: a trans-regulatory genetic architecture? Am. J. Hum. Genet. *91*, 466–477.
36. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
37. Van Buuren, S., and Groothuis-Oudshoorn, K. (2011). Multivariate Imputation by Chained Equations. J. Stat. Softw. *45*, 1–67.
38. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

39. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., et al.; Multiple Tissue Human Expression Resource (MuTHER) Consortium (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. *44*, 1084–1089.

40. Wright, F.A., Sullivan, P.F., Brooks, A.I., Zou, F., Sun, W., Xia, K., Madar, V., Jansen, R., Chung, W., Zhou, Y.-H., et al. (2014). Heritability and genomics of gene expression in peripheral blood. Nat. Genet. *46*, 430–437.

41. Chhibber, A., French, C.E., Yee, S.W., Gamazon, E.R., Theusch, E., Qin, X., Webb, A., Papp, A.C., Wang, A., Simmons, C.Q., et al. (2016). Transcriptomic variation of pharmacogenes in multiple human tissues and lymphoblastoid cell lines. Pharmacogenomics J. Published online February 9, 2016. http://dx.doi.org/10.1038/tpj.2015.93.

The American Journal of Human Genetics, Volume *98*

Supplemental Data

Imputing Gene Expression in Uncollected Tissues

Within and Beyond GTEx

Jiebiao Wang, Eric R. Gamazon, Brandon L. Pierce, Barbara E. Stranger, Hae Kyung Im, Robert D. Gibbons, Nancy J. Cox, Dan L. Nicolae, and Lin S. Chen

Supplemental Notes

Tissue-specific EQTL Effects Are Proportional to the Marginal Tissue-specific EQTL Effects in the GTEx Data

To assess the assumption of tissue-specific eQTL effects being proportional to the marginal tissue-specific eQTL effects, we conducted the following analysis based on the GTEx pilot data: For a given eQTL and a gene, we calculated the marginal tissue-specific eQTL effects, $(\hat{\alpha}_1, \ldots, \hat{\alpha}_9)^T$, $t = 1, \ldots, 9$, using a linear regression separately on the data for each tissue type. We also calculated the tissue-specific eQTL effects, $(\hat{\beta}_1, \ldots, \hat{\beta}_9)^T$, using a linear mixed effects model (R package `lme4`) with a random intercept for each individual and an interaction effect of the eQTL and each tissue type. In other words, $\hat{\alpha}_t$s are the marginal eQTL effects obtained by analyzing each tissue separately, whereas $\hat{\beta}_t$s are the eQTL effects obtained by jointly considering multiple tissues but allowing each tissue type to have a different eQTL effect. We calculated the correlations of the two vectors of eQTL effects, $\rho = \mathrm{cor}(\hat{\alpha}_t, \hat{\beta}_t)$. We obtained the correlation for each pair of eQTL and gene expression level in the GTEx data. Figure S4a shows the histogram of correlation coefficients and Figure S4b shows the histogram of the p-values of the correlation test. Each test is based on 9 observations (pairs of eQTL effects in 9 tissue types). We observed that for the majority of pairs of eQTLs and genes, the marginal tissue-specific eQTL effects were highly correlated with (i.e. proportional to) the eQTL effects in the mixed effects model that jointly considering individual random intercept and tissue-specific eQTL effects (i.e., interaction effects with tissue type). The mixed effects model requires nine parameters for each eQTL, and our assumption greatly simplifies the model in the GTEx data.

Simulations – the Impact of Sample Size on Imputation Performance

To further assess the impact of sample size (# individual) on the multitissue imputation, we simulated gene expression data for 150, 300, 500, and 1000 individuals in 9 tissue types. We simulated 1000 genes, and the expression levels for each are affected by two cross-tissue eQTLs and their interactions, with average heritability of 0.2, 0.3, and 0.5. Note that the average estimated heritability for expressed genes was reported as 0.14 to 0.26 for different tissue types in other studies.[1,2] In this simulation, we detected the significant cross-tissue eQTLs at the Stouffer's p-value threshold of 10^{-5}.

As shown in Table S2, as sample size increases, we observed improved power to detect true eQTLs. When sample size reaches 500 to 1000, the median estimated heritability is approaching the true heritability. Furthermore, the median imputation correlation for multitissue expression levels improves substantially as sample size increases and more eQTLs are used in the imputation. With a sample size of 1000, 94.6, 67.4, and 24.5% of the genes with high heritability (≥ 0.5) can be imputed with moderate, good, and excellent quality (i.e., imputation correlations of at least 0.3, 0.5 and 0.7), respectively; and majority of the expressed genes with heritability of $0.2 \sim 0.3$ can be imputed with moderate to good quality. The GTEx project will scale up donor collection to 900 individuals in the coming years, and will have decent power to detect eQTLs and will have improved performance in multittisue imputation.

Supplemental Figures and Legends

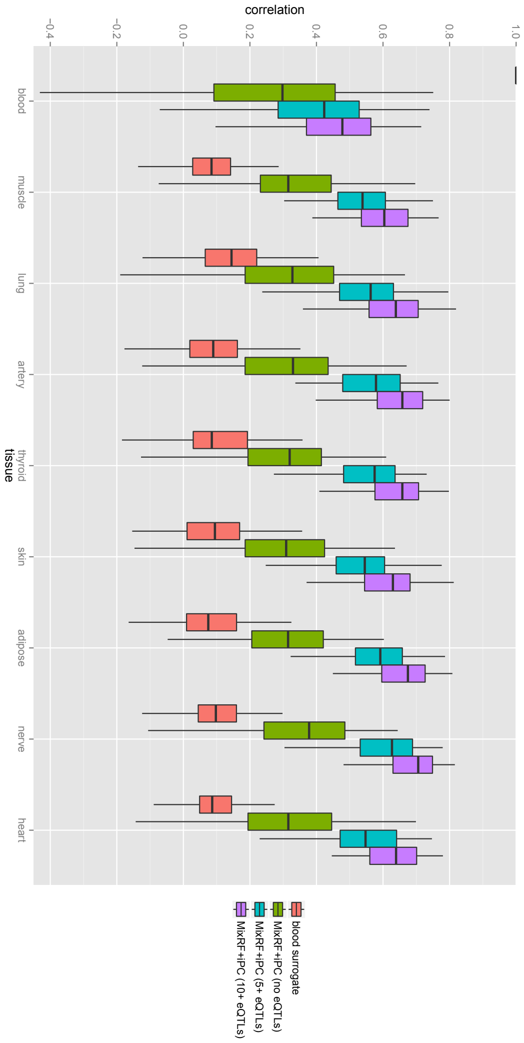


Figure S1: **Boxplots of sample-level true imputation correlation by tissue type.** The results are based on a 10-fold CV analysis within GTEx tissues. Each correlation is calculated as the correlation of the observed versus imputed values of a given gene in the current tissue. We compared the true imputation correlations of gene ranks based on MixRF+iPC among genes with 0, 5+, and 10+ eQTLs and with correlations based on blood surrogate.

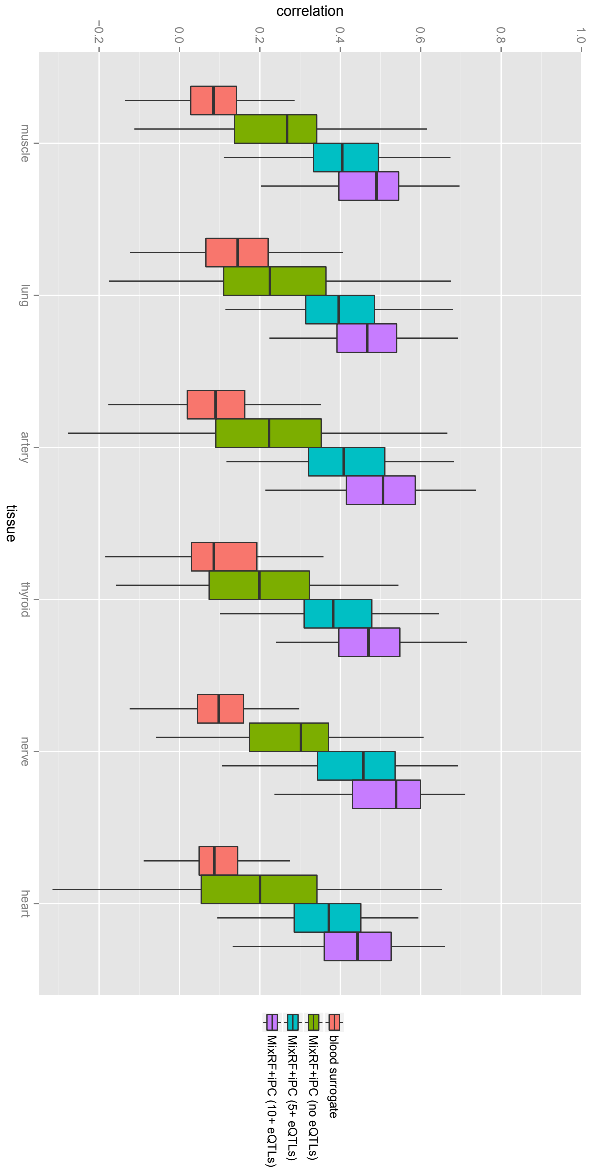


Figure S2: **Boxplots of sample-level true imputation correlation in inaccessible tissues in a new study.** The results are based on a 10-fold CV analysis of imputing uncollected tissues in the new samples, using GTEx as a reference. We compared the true imputation correlations based on MixRF+iPC among genes with 0, 5+, and 10+ eQTLs and with correlations based on blood surrogate.

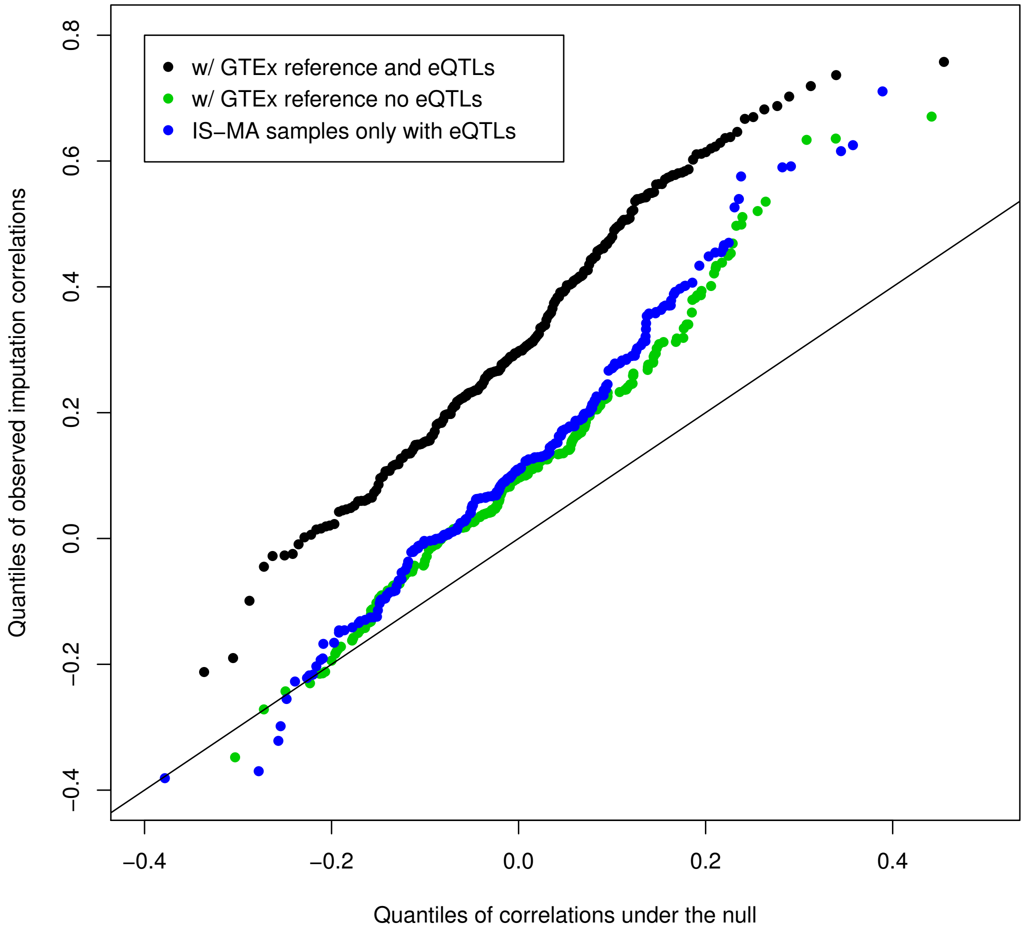
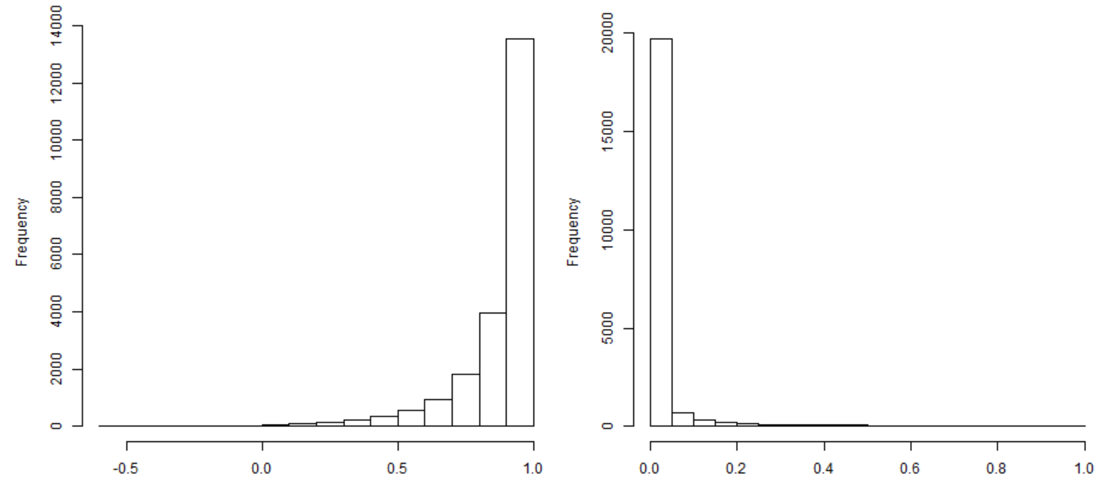


Figure S3: **Quantile-quantile plots of the observed imputation correlations versus the null correlations.** The three sets of observed imputation correlations are based on the analysis with GTEx reference and eQTLs (black), the analysis with GTEx reference and no eQTLs (green), and the analysis only with eQTLs without GTEx reference (blue).



(a) (b)

Figure S4: **Empirical evidence in GTEx data supports the assumption that tissue-specific eQTL effects are proportional to the marginal tissue-specific eQTL effects.** (a) Correlations of marginal versus tissue-specific eQTL effects in the mixed model. (b) P-values of correlations.

Supplemental Tables

	blood			muscle			lung		
# eQTLs	0	5+	10+	0	5+	10+	0	5+	10+
blood surrogate	NA	NA	NA	0.08	0.21	0.27	0.10	0.25	0.32
k-NN	0.02	0.05	0.06	0.14	0.19	0.18	0.10	0.14	0.16
missForest	0.25	0.27	0.26	0.33	0.36	0.33	0.27	0.32	0.31
MICE	0.02	0.09	0.12	0.08	0.22	0.28	0.07	0.22	0.29
lm	NA	0.18	0.27	NA	0.30	0.42	NA	0.32	0.43
lmer	0.09	0.28	0.33	0.26	0.47	0.53	0.27	0.49	0.56
REEMtree	0.10	0.28	0.33	0.26	0.46	0.53	0.27	0.48	0.55
MixRF	0.09	0.27	0.33	0.26	0.45	0.50	0.27	0.47	0.54
MixRF+iPC	0.25	0.35	0.39	0.33	0.48	0.53	0.29	0.48	0.54
	artery			thyroid			skin		
# eQTLs	0	5+	10+	0	5+	10+	0	5+	10+
blood surrogate	0.07	0.21	0.25	0.06	0.22	0.29	0.07	0.22	0.28
k-NN	0.14	0.16	0.15	0.24	0.21	0.18	0.11	0.14	0.14
missForest	0.27	0.32	0.30	0.30	0.35	0.32	0.25	0.27	0.24
MICE	0.10	0.26	0.33	0.07	0.22	0.30	0.06	0.22	0.28
lm	NA	0.34	0.47	NA	0.30	0.42	NA	0.32	0.42
lmer	0.34	0.52	0.59	0.27	0.49	0.55	0.25	0.47	0.54
REEMtree	0.35	0.51	0.58	0.27	0.48	0.55	0.26	0.47	0.53
MixRF	0.34	0.51	0.57	0.27	0.48	0.54	0.25	0.45	0.51
MixRF+iPC	0.32	0.49	0.57	0.30	0.47	0.55	0.28	0.45	0.52
	adipose			nerve			heart		
# eQTLs	0	5+	10+	0	5+	10+	0	5+	10+
blood surrogate	0.08	0.24	0.30	0.05	0.22	0.30	0.07	0.23	0.30
k-NN	0.13	0.15	0.16	0.20	0.20	0.19	0.15	0.15	0.16
missForest	0.28	0.33	0.30	0.34	0.38	0.34	0.30	0.33	0.32
MICE	0.10	0.27	0.35	0.12	0.29	0.37	0.07	0.23	0.30
lm	NA	0.36	0.48	NA	0.35	0.49	NA	0.31	0.43
lmer	0.33	0.55	0.60	0.39	0.58	0.63	0.28	0.49	0.56
REEMtree	0.33	0.54	0.59	0.40	0.58	0.63	0.28	0.49	0.56
MixRF	0.33	0.54	0.58	0.40	0.57	0.62	0.28	0.49	0.55
MixRF+iPC	0.31	0.52	0.59	0.38	0.55	0.62	0.28	0.46	0.55

Table S1: **The comparison of the median gene-level true imputation correlations by different methods.** We compared the median gene-level true imputation correlations for 8 competing methods including the proposed MixRF and MixRF+iPC for three groups of genes – those with 0, 5+, and 10+ eQTLs in any fold of data. The numbers of genes for the three groups are 9,240, 1,065, and 465 respectively. We did not show the imputation results by linear regression (lm) for genes with no eQTLs due to a lack of predictor.

True heritability	Sample size (# individual)	Median estimated heritability	Median imputation correlation (r_{imp})	% genes with $r_{imp} \geq 0.3$	% genes with $r_{imp} \geq 0.5$	% genes with $r_{imp} \geq 0.7$
0.20	150	0.001	0.294	48.6	10.7	0.5
	300	0.042	0.335	59.0	13.6	0.6
	500	0.077	0.359	67.8	15.9	0.8
	1000	0.129	0.383	75.8	18.6	0.9
0.30	150	0.014	0.347	59.4	22.0	2.9
	300	0.148	0.403	72.0	29.0	4.0
	500	0.204	0.434	79.6	34.5	5.1
	1000	0.271	0.451	85.7	37.8	6.0
0.50	150	0.216	0.463	73.4	44.3	14.7
	300	0.413	0.529	86.4	55.4	18.9
	500	0.478	0.566	91.9	63.4	22.6
	1000	0.562	0.581	94.6	67.4	24.5

Table S2: **The impact of sample size on imputation performance.** We simulated the gene expression levels for 1000 genes in 9 tissue types. Each gene expression level is affected by two cross-tissue eQTLs and their interactions, with average heritability of 0.2, 0.3, and 0.5. As sample size (# individual) increases from 150, 300, 500, to 1000, we observed improved power to detect true eQTLs and improved imputation performance.

References

1. E. Grundberg, K. S. Small, A. K. Hedman, A. C. Nica, A. Buil, S. Keildson, J. T. Bell, T.-P. Yang, E. Meduri, A. Barrett, et al., Mapping cis- and trans-regulatory effects across multiple tissues in twins, Nat. Genet. 44 (10) (2012) 1084–1089.

2. F. A. Wright, P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, K. Xia, V. Madar, R. Jansen, W. Chung, Y.-H. Zhou, et al., Heritability and genomics of gene expression in peripheral blood, Nat. Genet. 46 (5) (2014) 430–437.