HGG Advances

ARTICLE

Building an optimal predictive model for imputing tissue-specific gene expression by combining genotype and whole-blood transcriptome data

Sunwoo Jung,[1,5] Cue Hyunkyu Lee,[2,5] Jae Hoon Sul,[3] and Buhm Han[1,4,6,*]

Summary

Accurate imputation of tissue-specific gene expression can be a powerful tool for understanding the biological mechanisms underlying human complex traits. Existing imputation methods can be grouped into two categories according to the types of predictors used. The first category uses genotype data, while the second category uses whole-blood expression data. Both data types can be easily collected from blood, avoiding invasive tissue biopsies. In this study, we attempted to build an optimal predictive model for imputing tissue-specific gene expression by combining the genotype and whole-blood expression data. We first evaluated the imputation performance of each standalone model (using genotype data [GEN model] and using whole-blood expression data [WBE model]) using their respective data types across 47 human tissues. The WBE model outperformed the GEN model in most tissues by a large gain. Then, we developed several combined models that leverage both types of predictors to further improve imputation performance. We tried various strategies, including utilizing a merged dataset of the two data types (MERGED models) and integrating the imputation outcomes of the two standalone models (inverse variance-weighted [IVW] models). We found that one of the MERGED models noticeably outperformed the standalone models. This model involved a fixed ratio between the two regularization penalty factors for the two predictor types so that the contribution of the whole-blood transcriptome is upweighted compared with the genotype. Our study suggests that one can improve the imputation of tissue-specific gene expression by combining the genotype and whole-blood expression, but the improvement can be largely dependent on the combination strategy chosen.

Introduction

Transcriptomics has provided important information for understanding the physiological mechanisms involved in human traits and diseases.[1,2] As gene expression is associated with cellular activity and the environment, individual-level transcriptome profiles can be used for various research purposes.[3,4] Because of the complex genetic regulatory mechanisms that govern gene expression, transcriptome profiles vary greatly among different organs and tissues.[5] For this reason, an accurate assessment of the transcriptome profile of the relevant tissue is necessary when used for clinical purposes. However, obtaining tissue-specific transcriptome data most often involves an invasive biopsy of target tissues, which is not feasible for many inaccessible tissues, such as the brain.

Recently, several methods for imputing the transcriptome profile of a specific tissue have been proposed. These methods can be divided into two main categories according to the types of predictors used. The first category of methods imputes tissue-specific gene expression using genotype data as predictors.[6–8] These studies have shown that genetic variants have tissue-specific effects on gene expression and can therefore be used as predictors for imputing the tissue-specific transcriptome profile. With these methods, the upper bound of imputation accuracy is determined by the heritability of the expression trait. Most of these methods use *cis* variants close to a gene rather than exploiting genome-wide variants to reduce computational costs and risk of overfitting. The second category of methods imputes the gene expression levels of target tissues using the whole-blood transcriptome profile as a predictor.[9–11] The process of collecting the whole-blood transcriptome is much less invasive than biopsy of major tissues. Basu et al.[11] constructed a model that uses the genotype data and the whole-blood transcriptome data, primarily to examine how helpful the genotype data would be when added to the whole-blood transcriptome data in predicting tissue-specific gene expression.

In this study, we aimed to build an optimal predictive model for imputing the transcriptome profile of inaccessible tissues using the genotype data and the whole-blood expression data. We first evaluated the imputation performance of the standalone predictive model using the genotype data (GEN model) as predictors and the standalone predictive model using the whole-blood

[1]Interdisciplinary Program in Bioengineering, Seoul National University, Seoul, Republic of Korea; [2]Department of Biostatistics, Columbia University, New York, NY, USA; [3]Department of Psychiatry, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA; [4]Department of Biomedical Sciences, BK21 Plus Biomedical Science Project, Seoul National University College of Medicine, Seoul, Republic of Korea
[5]These authors contributed equally
[6]Lead contact
*Correspondence: buhm.han@snu.ac.kr
https://doi.org/10.1016/j.xhgg.2023.100223.

expression data (WBE model) as predictors. In our analysis, we used regularized linear regression to impute tissue-specific gene expression. We assessed the imputation performance using the mean R^2, considering all available genes in each of the 47 GTEx tissues. We found that the WBE model outperformed the GEN model by a large gain across the 47 tissues. Then, we attempted to combine the two standalone models (the GEN model and the WBE model) into one through several different strategies. Our strategies can be grouped into two categories. In the first category, we merged the genotype data and the whole-blood transcriptome data into one large dataset and used this merged dataset as a predictor (MERGED models). In the second category, we used the inverse variance-weighted (IVW) polymerization method to combine the imputation outcomes of the two standalone models (IVW models). We investigated whether these combined models would improve the overall performance for imputation of tissue-specific gene expression compared with the standalone models. We found that different combination strategies can give different performance outcomes. Notably, we observed a considerable improvement when we used the MERGED model that incorporates all available predictors and employs a fixed ratio between the two regularization penalty factors for the two types of predictors. This approach upweights the contribution of the whole-blood transcriptome data compared with the genotype data through the fixed ratio parameter. Our results suggest that one can indeed improve imputation performance by utilizing two types of data simultaneously, but the combination strategy can be an important factor that affects the final performance.

Material and methods

Data collection and preprocessing

We accessed all genotype data and transcriptome data used in this study from the GTEx v.7 database (dbGaP Accession phs000424.v7.p2).[12] The genotype data originally consisted of 635 whole-genome sequencing (WGS) samples that passed the quality control procedure according to the standard protocol described in the GTEx portal.[12] The samples were aligned against the human reference genome panel of GRCh37 (hg19). The transcriptome data consisted of the gene expression profiles from 714 donors, which were obtained via bulk RNA sequencing (RNA-seq) on 52 tissues. The number of samples available varied from tissue to tissue. We transformed the gene expression data into transcripts per million (TPM). We removed samples with missing data on either the genotype or the whole-blood transcriptome profile so that we could appropriately compare the imputation accuracy of the models based on the two data types. Of 52 tissues provided by GTEx, we excluded five of them that contained fewer than 40 samples and were left with 47 tissues. For the genes in the 47 tissues, we included only the autosomal protein-coding genes while excluding pseudo-genes, mitochondrial genes, and genes in sex chromosomes. Table S1 illustrates the summary of the data we generated and used for our study.

Regularized linear model for gene expression imputation

The standalone imputation model based on a single data type can be expressed as

$$y = X\beta + \epsilon,$$

where y denotes a $N\times1$ vector of the expression level of a target gene, adjusted for non-genetic covariates (Figure S1); X denotes a $N\times M$ matrix of the predictors, either the genotype or the transcriptome profile of the whole-blood tissue; β denotes the effect size of the predictors; and ϵ denotes residual error. We normalized y and each column of X by subtracting the mean and dividing by the standard deviation.

For both predictor types, we found that the number of predictors was much greater than the number of samples. In such a case, the general linear regression model can become vulnerable to collinearity, which can make the effect size estimates by the ordinary least squares (OLS) unreliable. In addition, interdependence between multiple predictors is highly likely to prevail within the genotype data because of linkage disequilibrium (LD). To ameliorate these problems, we adopted regularized regression as our predictive method. We tried three well-known regularization methods: least absolute shrinkage and selection operator (LASSO), ridge regression, and elastic net. The analysis was done using the R-glmnet package v.4.1.2.[13]

As shown under results, we found that the three regularization methods showed a similar imputation accuracy (mean R^2). We decided to use ridge regression as our regularization method because ridge yielded the minimal number of uninformative models (the models whose coefficients of predictors are all zero). The objective function of ridge regression can be expressed as follows:

$$\widehat{\beta} = \min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta^T X_i\right)^2 + \lambda\sum_{j=1}^{M}\beta_j^2\right\},$$

where λ denotes a penalty term for the number of predictors. The optimal λ is manually selected after trying the values in a range $(10^{-3}, 10^3)$ in a way that minimizes the mean squared error via the 10-fold cross-validation (CV).

With the estimated effect sizes $\widehat{\beta}$, we could build the standalone imputation model to impute the individual-level transcriptome profile of genes belonging to a tissue of interest. We define the standalone model based on the genotype data as the "GEN model" and the standalone model based on the whole-blood expression data as the "WBE model." Let X^{GEN} be the matrix of the genotype data, and let X^{WBE} be the matrix of the whole-blood transcriptome data. The imputed gene expression from each type of predictors can be expressed as follows:

$$\widehat{y}^{GEN} = X^{GEN}\widehat{\beta}^{GEN},$$

$$\widehat{y}^{WBE} = X^{WBE}\widehat{\beta}^{WBE}$$

In our imputation, we split the entire data into the training set and test set with an 8:2 ratio. We fitted regularized regression using the training set and imputed the transcriptome profile of target tissues using the test set.

Using the merged dataset for gene expression imputation

We developed combined models that leverage the genotype data and the whole-blood transcriptome data. Our first attempt was

to merge the datasets of the two predictor types and use this merged dataset for gene expression imputation. To account for the difference in the contribution of the two predictor types, we implemented separate regularization penalty factors. Specifically, we assumed a ratio φ between the two regularization penalty factors. We define this strategy as the MERGED model. To find the optimal form of the MERGED model, we generated and tested four different approaches. These four approaches differed by (1) whether we performed feature selection on the genotype predictors to reduce overfitting and simplify the model and (2) whether we fixed φ to a predefined value or flexibly searched for the best φ for each gene.

The first approach, referred to as MERGED_fixed, uses the merged dataset including all available predictors of both predictor types. It employs a fixed ratio φ between the two regularization penalty factors for the two types of predictors. The second approach, referred to as MERGED_fixed_filtered, is similar to MERGED_fixed but uses the merged dataset that has undergone feature selection on the genotype predictors. The third approach, referred to as MERGED_flexible, flexibly searches for the best ratio φ for each gene while using all available predictors of both predictor types. The fourth approach, referred to as MERGED_flexible_filtered, is similar to MERGED_flexible but uses the merged dataset that has undergone feature selection on the genotype predictors. A summary table of model abbreviations used in our study is provided in Figure S9.

Now we describe the type of feature selection we performed for pre-filtering the genotype predictors (in models with postfix "_filtered"). We fitted LASSO using the training set of the genotype data. Then we only kept the SNPs that were not discarded by LASSO. Because LASSO applies stronger regularization than ridge, this feature selection allowed us to focus on a smaller set of SNPs. We applied this feature selection only to the genotype data because the effect of the feature selection was only marginal in the WBE model compared with the GEN model.

Below, we describe how we applied the separate regularization penalty factors to each data type. The objective function for the approaches that apply separate penalty factors to two data types can be expressed as follows:

$$\widehat{\beta} = \min_{\beta}\left\{\sum_{i=1}^{N}\left(y_i - \beta^T X_i\right)^2 + \lambda\sum_{p=1}^{M_{WBE}}\beta_p^2 + \lambda\varphi\sum_{q=1}^{M_{GEN}}\beta_q^2\right\},$$

where φ denotes the regularization ratio parameter supplied only to the genotype predictors to differentiate the regularization penalty applied to each predictor type. We considered a value of φ from the set $\{10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}\}$. For the models that implemented a fixed value for φ (models with postfix "_fixed"), we denoted the fixed value of φ in the model's name by appending an additional postfix. For example, MERGED_fixed_100 indicates that φ was set to 100. Because φ can be thought of as the ratio of the penalty (λ) for the two data types, a large value of φ means that we penalize the genotype data more so that the contribution of the whole-blood transcriptome data is upweighted. We also tried a flexible model that determines the best φ for each gene separately using the validation data (models with postfix "_flexible").

Using the weighted sum of the imputation outcomes of the two standalone models

As another attempt to develop a combined model, we integrated the imputation outcomes of the GEN model and WBE model using the IVW polymerization method. We define this strategy as the IVW model. IVW is a widely used method for integrating the estimates from different sources because weighting the estimates with the inverse of their variances can minimize the variance of the final integrated estimate.[14] To find the optimal form of the IVW model, we generated and tested four different approaches. These four approaches differed by (1) whether we tried to conditionally integrate the outcomes of the two standalone models for the genes for which the GEN model outperformed the WBE model according to the validation set and (2) whether we calculated the inverse variance weights using the 10-fold CV method or bootstrap method.

The first approach, referred to as IVW_CV, integrates the imputation outcomes of the two standalone models for all available genes and calculates the inverse variance weights using the 10-fold CV. The second approach, referred to as IVW_CV_cond, integrates the imputation outcomes of the two standalone models only for genes for which the GEN model outperformed the WBE model according to the validation set and calculates the inverse variance weights using the 10-fold CV. The third approach, referred to as IVW_Bootstrap, integrates the imputation outcomes of the two standalone models for all available genes and calculates the inverse variance weights using the bootstrap method. The fourth approach, referred to as IVW_Bootstrap_cond, integrates the imputation outcomes of the two standalone models only for genes for which the GEN model outperformed the WBE model according to the validation set and calculates the inverse variance weights using the bootstrap method.

We employed the empirical approaches (10-fold CV or bootstrap) to obtain the variance of the imputation outcomes needed for IVW. This was because analytically obtaining the actual variance of the estimates was challenging, particularly for regularized regression.[13]

When using the 10-fold CV, we calculated the difference between the true gene expression and the imputed expression across samples and obtained the variance of these differences. Then, we averaged the results over 10 folds to get the final variance estimate. The estimated variance for gene i can be expressed as follows:

$$\widehat{\sigma}_i^2 = \frac{1}{10}\sum_{K=1}^{10} var\left(y_{K,i} - \widehat{y}_{K,i}\right),$$

where $y_{K,i}$ is the actual gene expression level of gene i for the K th set of the 10-fold CV, and $\widehat{y}_{K,i}$ is the imputed value from the corresponding predictor set X_K.

When using the bootstrap, we calculated the difference between the true gene expression and the imputed expression across samples and obtained the variance of these differences. Then, we averaged the results over 40 samplings to get the final variance estimate. The estimated variance for gene i can be expressed as follows:

$$\widehat{\sigma}_i^2 = \frac{1}{40}\sum_{B=1}^{40} var\left(y_{B,i} - \widehat{y}_{B,i}\right),$$

where $y_{B,i}$ is the actual gene expression level of gene i for the Bth sampled set, and $\widehat{y}_{B,i}$ is the imputed value from the corresponding predictor set X_B.

Now, let $\widehat{y}_i^{GEN}$ and $\widehat{y}_i^{WBE}$ be the gene expression estimates of the GEN model and WBE model, respectively, for gene i. The combined gene expression estimate, $\widehat{y}_i^{IVW}$, is then calculated as follows:

$$\widehat{y}_i^{IVW} = \frac{\widehat{y}_i^{GEN}\frac{1}{\widehat{\sigma}_i^{2\,\mathrm{GEN}}} + \widehat{y}_i^{WBE}\frac{1}{\widehat{\sigma}_i^{2\,\mathrm{WBE}}}}{\frac{1}{\widehat{\sigma}_i^{2\,\mathrm{GEN}}} + \frac{1}{\widehat{\sigma}_i^{2\,\mathrm{WBE}}}},$$

where $\hat{\sigma}^2_i{}^{\mathrm{GEN}}$ and $\hat{\sigma}^2_i{}^{\mathrm{WBE}}$ are the empirically estimated variances for weighting $\hat{\gamma}_i^{GEN}$ and $\hat{\gamma}_i^{WBE}$, respectively.

Preparing two sets of genes based on tissue specificity of expression

Highly tissue-specific genes

We used the same procedure as Basu et al.[11] for obtaining highly tissue-specific genes for each of the 47 tissues. For each gene in a target tissue, we calculated its tissue specificity score as the log_2 of ratio of its mean gene expression in the target tissue to its mean gene expression in the rest of the tissues. Then, for each tissue, we obtained genes with tissue specificity scores in the top 20th percentile.

Highly conserved genes

For each gene, we calculated the variance of its mean gene expression across the 47 tissues. Then we obtained genes with variances in the bottom 20th percentile.

Results

Optimal regularization method for each type of predictor

We first attempted to figure out an optimal predictive method for each type of predictor (genotype and the whole-blood transcriptome) for imputing tissue-specific gene expression. The number of samples in the genotype dataset and the whole-blood transcriptome dataset was the same because we collected samples for which the genotype data and the transcriptome data were available (Table S1). Figure S1 illustrates the summary of the data used in our study. The number of genes with available transcriptome profiles varied across tissues in a range from 13,533 (skeletal muscle, the minimum) to 21,343 (testis, the maximum) (Figure S1). Each gene differed in the number of *cis* variants located around it, the minor allele frequencies of its *cis* variants, and the gene length (Figure S1). Because the GTEx project collected multiple tissues from the same individual, tissue datasets had sample overlap (Figure S1). For the GEN model, we included SNPs within the 1-Mb window of a target in the predictor set. For the WBE model, we included all available genes from the whole-blood tissue in the predictor set. The simplest form of the WBE model could be constructed by using only the matched gene from the whole-blood tissue as a predictor for imputing the expression of the corresponding gene in a target tissue. Yet, the imputation performance of this simplest form was significantly lower compared with using all available genes from the whole-blood tissue as predictors (Figure S2; Table S2).

Because our datasets had far more predictors than samples, we decided to use regularized regression to reduce overfitting and make the regression fit stable. LASSO, ridge regression, and elastic net are three widely used regularized regression methods, and we tested these three for imputing tissue-specific gene expression using either the genotype data or the whole-blood transcriptome data. For this analysis, we split the data into a training set and a test set with an 8:2 ratio so that we fit regression using the training set and imputed the gene expression levels of the 47 tissues in the test set. The regularization parameter (λ) was optimized within the training set via CV.

Figure 1 illustrates the comparison of the three regularization methods on the imputation of gene expression across the 47 tissues using either the genotype data or the whole-blood transcriptome data as predictors. For the metric of imputation accuracy, we calculated the mean of R^2 between the true expression level and imputed expression level across samples over all available genes belonging to each tissue (mean R^2; Figure 1A). We observed that the three regularization methods showed a similar imputation accuracy when the genotype data were used as predictors (Figure 1A; Table S3). The tissue-wise average of the mean R^2 over the 47 tissues was 0.064 for all three regularization methods. In contrast, we observed that ridge performed slightly better than LASSO and elastic net when the whole-blood transcriptome data were used as predictors (Figure 1A; Table S3). The tissue-wise average of the mean R^2 over the 47 tissues was 0.118 for LASSO, 0.120 for elastic net, and 0.129 for ridge. Ridge outperformed LASSO and elastic net in 39 of 47 tissues and 38 of 47 tissues, respectively, with the whole-blood transcriptome data. Ridge is known to penalize the parameters less strictly than LASSO. Therefore, the superior performance of ridge regression in the WBE model may suggest that the predictive information of the whole-blood transcriptome on tissue-specific gene expression might be dispersed over many different genes.

Aside from the mean R^2, we also assessed for how many genes each regularization method resulted in an informative or uninformative model. Regularization methods can result in a model where the regression coefficients of all predictors are shrunk to zero. Then, the imputed gene expression levels of such a model are no longer influenced by the predictors and thus can be considered uninformative. LASSO and elastic net tended to produce uninformative models for a large proportion of genes across the 47 tissues in the GEN model (44.1% and 42.3% on average, respectively), possibly because of their stronger penalization than ridge (Figure 1B; Table S3). This phenomenon was more severe in the GEN model but still observed in the WBE model when LASSO or elastic net was used (16.1% and 15.5% on average, respectively). On the contrary, ridge did not produce any uninformative model for either predictor type (Figure 1B; Table S3).

We wanted to consider the mean R^2 and the proportion of the uninformative model when selecting the best regularization method for the GEN model and WBE model. For the GEN model, although the three regularization methods showed the same imputation accuracy across all 47 tissues, LASSO and elastic net produced a large number of uninformative models. For the WBE model, ridge was

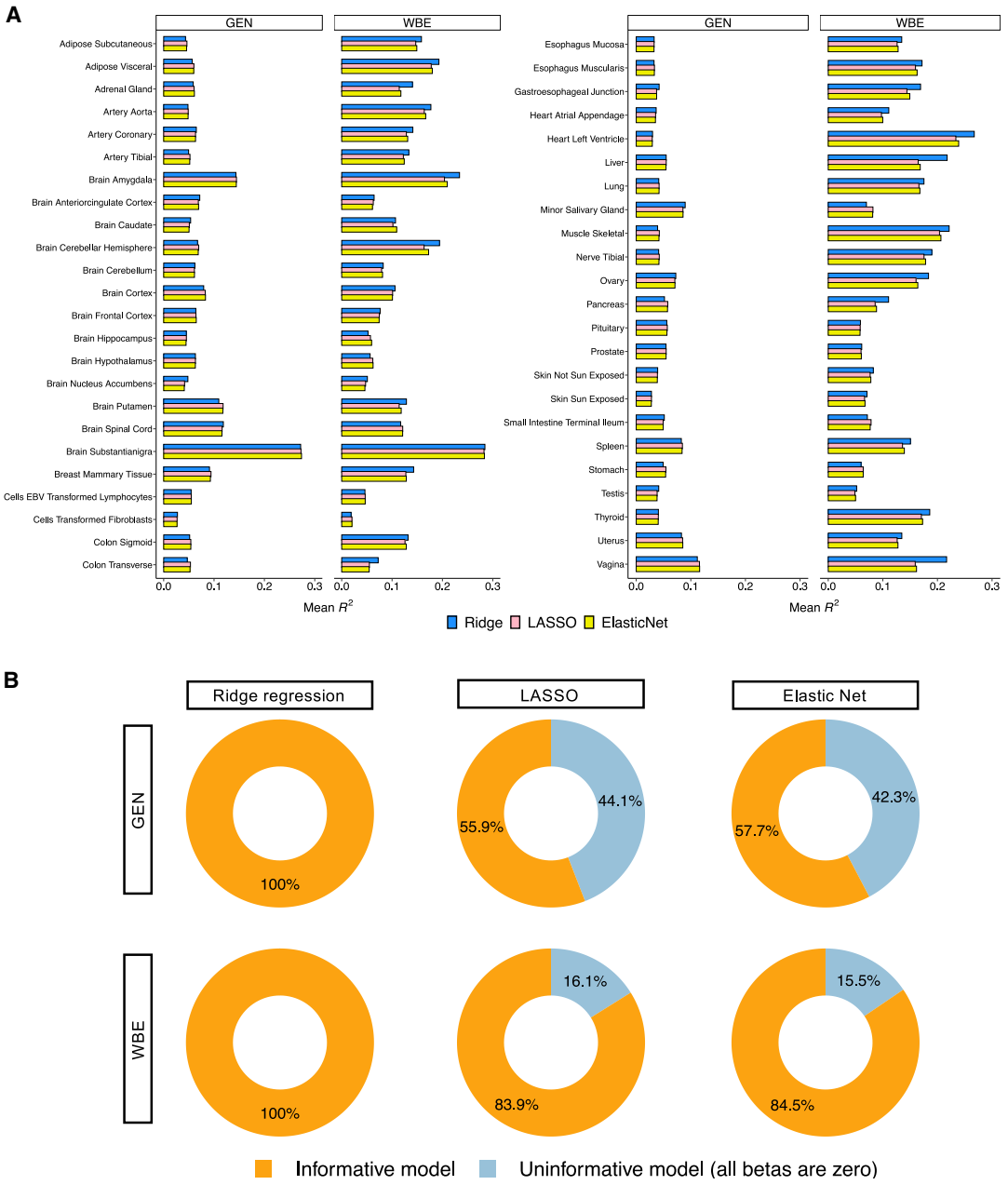

Figure 1. Comparison of three regularized regression methods for the imputation of tissue-specific gene expression using each type of predictor

Shown are imputation performance reports for the three regularized regression methods (ridge regression, LASSO, and elastic net) using either the genotype or whole-blood expression.

(A) The bar plots show the imputation accuracy of ridge, LASSO, and elastic net for gene expression across the 47 tissues, using each type of predictor. The metric for imputation accuracy is mean R^2. We used different color schemes for ridge (blue), LASSO (pink), and elastic net (yellow).

(B) The pie charts show the proportion of the informative models and the uninformative models that resulted from each regularized regression method using each type of predictor.

superior to LASSO and elastic net in terms of the mean R^2 and the proportion of the uninformative model. Summing these up, we decided to use ridge regression for the genotype data and the whole-blood expression data in our subsequent analyses.

Here the question arose whether the genes for which LASSO or elastic net produced the uninformative models are truly uninformative. To this end, we tested how ridge performed for genes that resulted in uninformative models by LASSO. The imputation accuracy of ridge on those genes was relatively low compared with the rest of the genes, but the mean R^2 was significantly greater than zero (Figure S3). This result suggested that, even for those genes, ridge had predictive power.

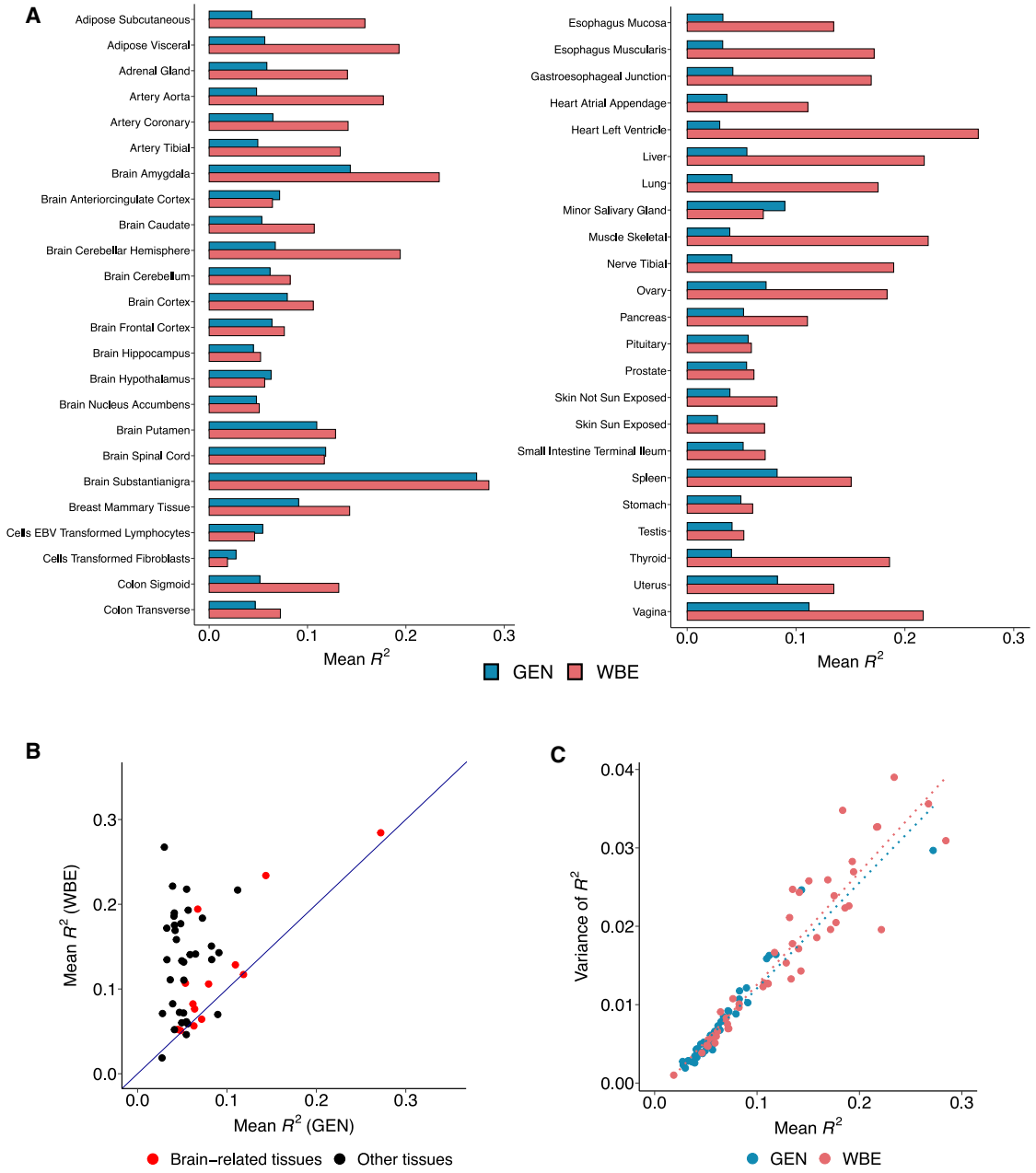

Figure 2. Comparison of the GEN model and WBE model

Shown are imputation performance reports for the GEN model and WBE model.

(A) The bar plots show the imputation accuracy of the GEN model and WBE model for gene expression across the 47 tissues. The metric for imputation accuracy is mean R^2. We used a different color scheme for the GEN model (blue) and WBE model (red).

(B) The dot plot shows the relationship between the imputation results of the GEN model and WBE model across the 47 tissues. The imputation results for 13 brain-related tissues are indicated in red. The dark blue line represents the y = x line. Many of the brain-related tissues (red dots) are near the y = x line, indicating that the performance difference between the GEN model and WBE model was relatively small for these tissues.

(C) The dot plot shows the relationship between the variance of R^2 and the mean R^2 across the 47 tissues for the GEN model (blue) and WBE model (red). For the variance of R^2, we calculated the variance of R^2 values over all genes in each tissue.

Comparing two types of predictors: Genotype and whole-blood expression

Using ridge regression as the common regularization method, we wanted to see which type of predictors would result in better imputation performance and how large the performance difference would be, considering all available genes in the 47 human tissues. Figure 2 illustrates the comparison between the imputation performance of the GEN model and WBE model across the 47 tissues. We observed that the WBE model considerably outperformed the GEN model in 41 of 47 tissues, and the difference in performance of the two models was large

in many tissues (>2-fold difference in mean R^2 in 22 of 47 tissues; Table S4). The tissue-wise average of the mean R^2 over the 47 tissues was 0.129 for the WBE model. In contrast, the tissue-wise average of the mean R^2 over the 47 tissues was only 0.064 for the GEN model, which was lower than half that of the WBE model. The tissue-wise average of the differences of the mean R^2 between the GEN model and WBE model over the 47 tissues was 0.068 (Table S4). The difference was the largest in "Heart Left Ventricle" (GEN, 0.030; WBE, 0.267) and the smallest in "Brain Spinal Cord" (GEN, 0.118; WBE, 0.117). There were 6 cases where the GEN model outperformed the WBE model, but the difference was small in these cases, with the tissue-wise average of the differences being only 0.009. 3 of 6 cases where the GEN model outperformed the WBE model were found in brain-related tissues ("Brain Anteriorcingulate Cortex," "Brain Hypothalamus," and "Brain Spinal Cord"; Table S4). When we examined 13 brain-related tissues, the performance difference between the GEN model and the WBE model was relatively small compared with other tissues, with the tissue-wise average of the differences being 0.030 (Figure 2B; Table S4). The variance of R^2 over all genes in each tissue tended to increase according to their mean for the GEN model and WBE model (Figure 2C).

These results were based on a single split of the training (80%) and the test (20%) set. Because how the data were split could affect the imputation results, we further randomly split the entire data into 5 folds and measured the performance of the GEN model and WBE model over the 5 trials, using each fold as the test set. We found the consistent result that the WBE model outperformed the GEN model by a large gain (Figure S4) across the 5 trials, suggesting that how the data were split did not much affect our results. We found that the tissue-wise averages of the standard error of the mean R^2 over the 5-fold CV were small enough (GEN, 0.001; WBE, 0.001; Figure S4; Table S5) to exclude the possibility that the observed difference between the GEN model and WBE model was due to the sampling error.

Gene-specific penalization for the GEN model

Before proceeding to build a combined model, we wanted to try gene-specific penalization in the GEN model. In contrast to the WBE model whose predictors are the same for all target genes, in the GEN model, the predictors (*cis* SNPs) are different across genes. Thus, the optimal regularization method may vary from gene to gene. We attempted to build a gene-specific GEN model by allowing the model to select the optimal regularization method (LASSO, ridge, or elastic net) for each target gene according to the validation set (1/4 of the training set). In such a way, the model may well reflect the unique genetic architecture of the *cis* SNPs of a given gene. We found that this gene-specific approach ("gene_specific_penalization") did not improve the imputation performance of the original GEN model (Figure S5). The tissue-wise average of the mean R^2 over the 47 tissues was 0.064 for this approach, consistent with using ridge alone (Table S6).

Combined models using both types of predictors

Because the genotype data and the whole-blood transcriptome data could present independent information, combining these two data sources might provide an opportunity to further improve imputation performance. To leverage the genotype data and the whole-blood transcriptome data, we considered several approaches that can be grouped into two categories.

The first category used the merged dataset of the genotype data and the whole-blood transcriptome data for gene expression imputation (MERGED model). We generated and evaluated four different approaches of the MERGED model (MERGED_fixed, MERGED_fixed_filtered, MERGED_flexible, and MERGED_flexible_filtered; material and methods). The four approaches either included all available predictors of the genotype data or only some selected genotype predictors ("_filtered" postfix). Also, they either used a fixed value for the regularization ratio parameter φ ("_fixed") or flexibly selected a value of φ for each gene based on the validation data ("_flexible"). A detailed description of the four approaches is provided in material and methods.

The second category integrated the imputation outcomes of the GEN model and WBE model using the IVW polymerization method and used this weighted sum as the final estimate for gene expression imputation (IVW model). Here again, we generated and evaluated four different approaches of the IVW model (IVW_CV, IVW_CV_cond, IVW_Bootstrap, and IVW_Bootstrap_cond; material and methods). The four approaches integrated the imputation outcomes of the two standalone models either for all available genes or only for the genes for which the GEN model outperformed the WBE model according to the validation set ("_cond" postfix). Also, they calculated the inverse variance weights using either the 10-fold CV ("_CV") or the bootstrap ("_Bootstrap"). A detailed description of the four approaches is provided in material and methods.

Figure 3 illustrates the imputation performance of all combined models evaluated in our study across the 47 tissues. We observed that imputation performance greatly varied depending on the combination strategy chosen. Notably, the MERGED model that used all available predictors of both predictor types and employed a fixed value of 100 for the regularization ratio parameter φ (MERGED_fixed_100) showed a noticeable improvement in imputation performance. The MERGED_fixed_100 model outperformed the WBE model in 43 of 47 tissues. The tissue-wise average of the mean R^2 over the 47 tissues was 0.136 for the MERGED_fixed_100 model, which was greater than 0.129 of the standalone WBE model (Figure 4A; Table S7). The top tissues for which the MERGED_fixed_100 model yielded the largest performance gain over the WBE model were "Cells Transformed

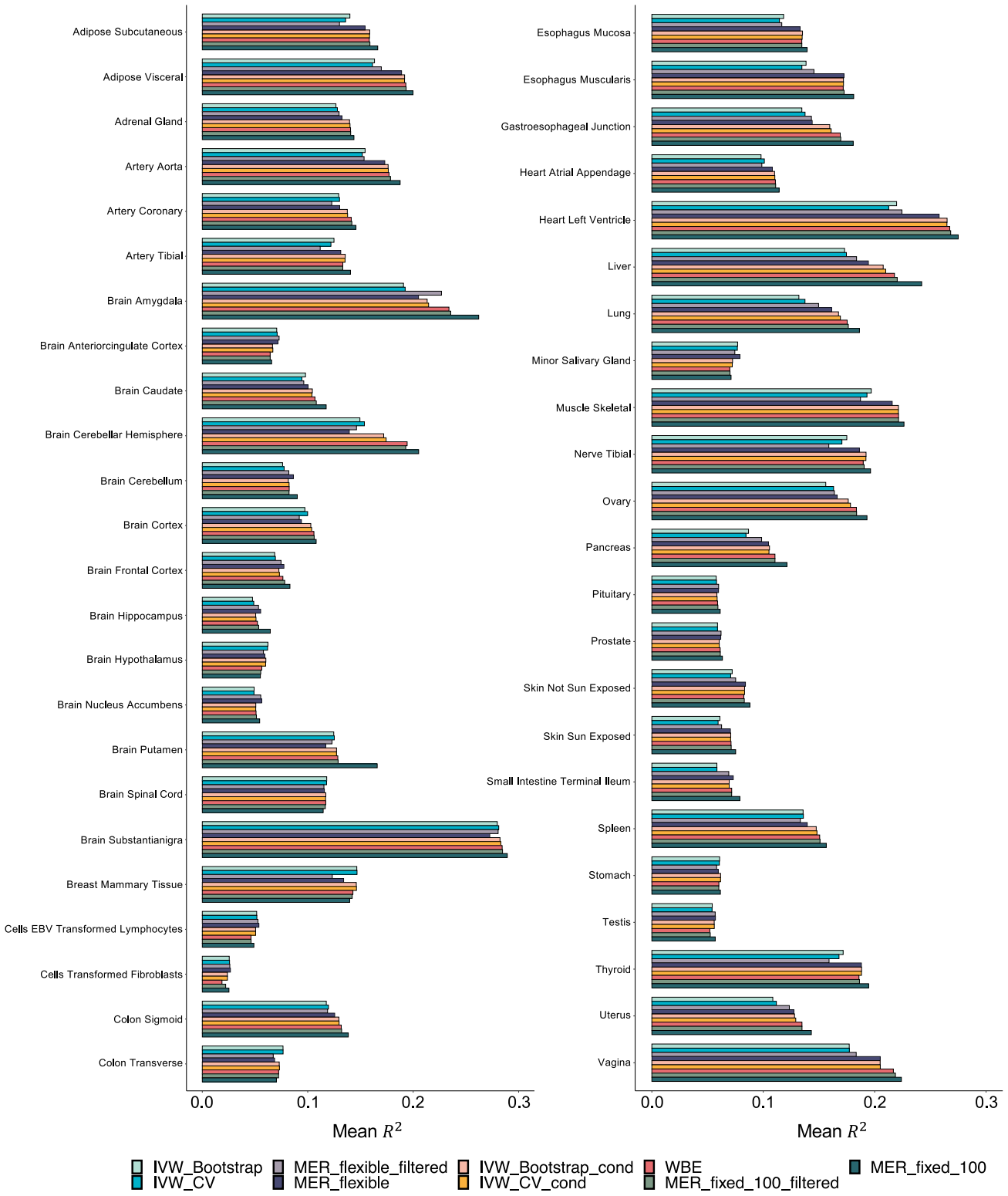

Figure 3. Comparison of the WBE model and various combined models

Shown are imputation performance reports for the standalone WBE model and all combined models evaluated in this study. The bar plots show the imputation accuracy of the WBE model, four different approaches of IVW models, and four different approaches of MERGED models for gene expression across the 47 tissues. The metric for imputation accuracy is mean R^2. We used a different color scheme for each different model, as indicated in the legend.

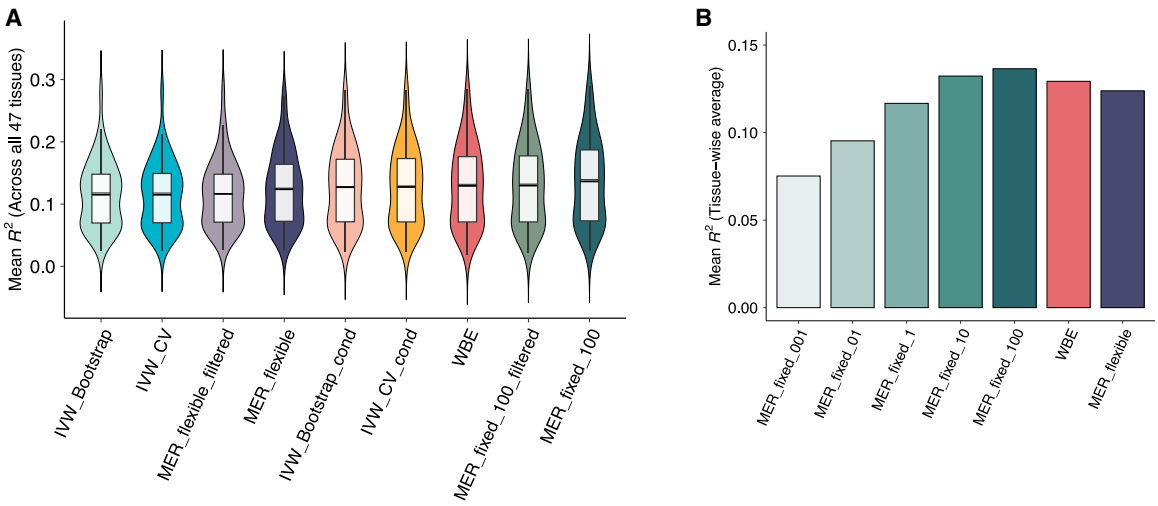

Figure 4. Overall imputation performance of the WBE model and various combined models and comparison of MERGED_fixed models with varying fixed values of the regularization ratio parameter φ

(A) The violin plot shows the distribution of mean R^2 over the 47 tissues for the WBE model and all combined models evaluated in this study. The crossbar in the boxplot indicates the average of the mean R^2 values over the 47 tissues (tissue-wise average).

(B) The bar plot shows the tissue-wise average of the imputation accuracy over the 47 tissues for five variations of MERGED_fixed models with varying fixed values of the regularization ratio parameter φ. The plot also includes the imputation performance of the WBE model and MERGED_flexible model.

Fibroblasts" (approximately 32% increase from 0.019 to 0.025 in mean R^2), "Brain Putamen" (approximately 29% increase from 0.129 to 0.166 in mean R^2), and "Brain Hippocampus" (approximately 23% increase from 0.052 to 0.064 in mean R^2). The MERGED models that flexibly selected the best regularization ratio parameter φ (MERGED_flexible and MERGED_flexible_filtered) did not outperform the WBE model. All IVW models failed to outperform the standalone WBE model (Figures 3 and 4A; Table S7). The two strategies for determining inverse weights (10-fold CV and the bootstrap) did not notably change the results of IVW models.

In the MERGED_fixed model, we considered a fixed value of φ from the set $\{10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}\}$. A larger ratio value of φ indicates a larger regularization penalty applied to the genotype predictors compared with the whole-blood expression predictors. After evaluating each value in the set, we observed that a larger value of φ was associated with higher imputation accuracy (Figure 4B; Table S8). This suggested that upweighting the contribution of the whole-blood expression data was effective in enhancing imputation accuracy because whole-blood expression may have more predictive information compared with the genotype.

An approach that selects the optimal model for each gene

Previously, when imputing gene expression levels, a single model was consistently used for all genes in a target tissue. Here, we explored an additional approach that selects the most suitable model for each gene based on validation accuracy. For each gene, we evaluated four different models (the GEN model, WBE model, IVW_CV model, and MERGED_fixed_100 model) using the validation set, which was a subset of the original training set, and selected the model with the best validation performance for imputing the expression level. We found that this approach of choosing the optimal model for each gene ("Best_on_validation") did not show better imputation performance compared with the vanilla MERGED_fixed_100 model (Figure 5A). The tissue-wise average of the mean R^2 over the 47 tissues was 0.124 for this additional approach (Table S9), which was lower than 0.136 of the MERGED_fixed_100 model. We found that the Best_on_validation model selected the MERGED_fixed_100 model for approximately 35% of all available genes in the 47 tissues (Figure 5B; Table S9).

Model evaluation considering tissue specificity of gene expression

So far, we have assessed the imputation performance of our models across all available genes in the 47 tissues. Here, we wanted to evaluate the imputation performance of the models by considering the tissue specificity of gene expression. Some genes are very tissue specific in that they are expressed only in specific tissues, carrying out tissue-specific functions. In contrast, some genes are consistently expressed across all tissues, carrying out basal cellular functions required for the survival of cells. Therefore, it is worth evaluating the imputation performance of the models as a function of tissue specificity of gene expression. To this end, we prepared two sets of genes based on their tissue specificity. The first set consisted of genes whose expression was highly tissue specific, while the second set consisted of genes whose expression was highly conserved across the 47 tissues. The procedure for obtaining these two sets of genes is described in material and methods. According to our

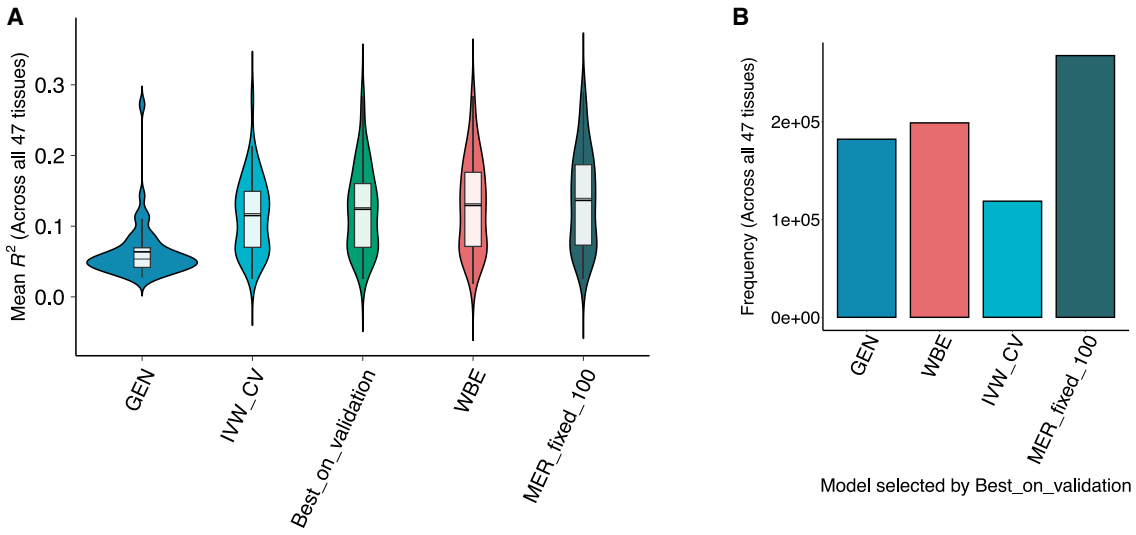

Figure 5. Imputation performance of the approach that selects the optimal model for each gene
(A) The violin plot compares the approach that selects the optimal model for each gene ("Best_on_validation") with some other models. It shows the distribution of mean R^2 over the 47 tissues for each model. The crossbar in the boxplot indicates the average of the mean R^2 values over the 47 tissues (tissue-wise average). The Best_on_validation model is indicated in green.
(B) The bar plot shows the overall distribution of the candidate models selected by the Best_on_validation model based on validation accuracy. For each candidate model, we calculated the sum of all frequencies over all genes in the 47 tissues.

results, the WBE model and MERGED_fixed_100 model performed better for the tissue-specific genes than for the conserved genes (Figure S6). For the WBE model, the tissue-wise average of the mean R^2 over the 47 tissues was 0.141 for the tissue-specific genes and 0.130 for the conserved genes. For the MERGED_fixed_100 model, the tissue-wise average of the mean R^2 over the 47 tissues was 0.149 for the tissue-specific genes and 0.140 for the conserved genes. In contrast, the GEN model performed slightly better for the conserved genes than for the tissue-specific genes (Figure S6). For the GEN model, the tissue-wise average of the mean R^2 over the 47 tissues was 0.063 for the tissue-specific genes and 0.066 for the conserved genes. These results may suggest that the genotype and whole-blood transcriptome profile can have slightly different contributions to the imputation of genes that are conserved across various tissues and genes that are specific to a particular tissue.

Discussion

In this study, we wanted to build the most accurate model possible for imputing the transcriptome profile of inaccessible tissues, leveraging the genotype and whole-blood expression. Tissue-specific transcriptome profiling often requires a biopsy of a target tissue, which is invasive and costly because of the limited accessibility. With an increasing amount of genotype data and RNA-seq profiles available, a suitable strategy using these sources would allow an accurate imputation of the tissue-specific transcriptome profile and facilitate transcriptome-wide association studies (TWASs). With our investigation, we suggest that one can improve imputation performance by utilizing the genotype and whole-blood transcriptome, but the choice of strategy for combining them matters.

One concern of our study is that the imputation outcomes of the GEN model and WBE model cannot be interpreted from the same perspective. The expression estimate obtained using the GEN model represents genetically regulated expression. Also, the R^2 between the expression estimate from the GEN model and the true expression can be interpreted as an estimate of the heritability of the gene expression trait. However, the expression estimate obtained using the WBE model cannot provide the same interpretation. In our study, because the main purpose was to maximize imputation performance, we treated the two types of predictors the same regardless of the difference in their interpretation.

We evaluated many different strategies for combining the two data sources in our study. One method we tried was IVW, which combined the imputation outcomes of the two standalone models using the variance estimates as the weights. This approach assumed that the variance estimates of the two models are compatible. However, the variance estimates of the two models may not be interpreted similarly because of the model differences. If that is the case, then ignoring the variances and simply averaging the point estimates might work better. We tried this strategy as well, but we observed that its imputation accuracy was lower than the imputation accuracy of the IVW strategy (Figure S7).

One limitation of our study is that we only considered one tissue (whole-blood expression) as a predictor along with the genotype. If another tissue is also easily accessible, then we may consider it as our predictor. We additionally

leveraged the expression data of skin because skin is easily accessible using a relatively noninvasive process, like the whole-blood tissue. We obtained the skin expression datasets from the GTEx v.7 database. The imputation models using skin expression as predictors showed performance comparable with the WBE model across the 47 tissues (Figure S8; Table S10), suggesting that use of skin expression can also be considered when imputing the transcriptome profile of inaccessible tissues. Moreover, we used a single data source (GTEx) for the whole-blood transcriptome data, and these data are postmortem. Therefore, our investigation has a limitation in that it may lack robustness across multiple data sources and may not be generalized well to living-donor samples.

Another limitation of our study is that we assumed a specific set of values for the regularization ratio parameter φ; namely, $\{10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}\}$. It is possible that there could be better options for the value of φ other than those numbers. Finding the optimal value of φ from the entire range of real numbers would not be computationally feasible, though.

In the MERGED_flexible model, we chose the optimal regularization ratio parameter φ for each gene based on the validation data. Contrary to our expectations, the MERGED_flexible model did not outperform the standalone WBE model. The value of φ selected by the MERGED_flexible model based on the validation data was not consistent with the optimal value of φ for the test data in approximately 73.7% of all genes across the 47 tissues (Table S11). Because the validation data and test data were chosen randomly, the disparities between the two datasets seem to stem from the small sample size. If the sample size to fit model increases in the future, then we expect that the performance of the MERGED_flexible method will increase.

Although we tried many different approaches in our study, we clearly could not try all possible methods, and thus there can be other possible strategies that can outperform our methods. We used regularized linear regression for the imputation of tissue-specific gene expression in our study. Some existing methods used approaches other than regularized linear regression. Bayesian functional genome-wide association study (bfGWAS) used Bayesian variable selection regression (BVSR) to construct a model for gene expression imputation,[15] while transcriptome-integrated genetic association resource (TIGAR) used Bayesian Dirichlet process regression to construct a model for gene expression imputation.[16] BVSR is known to perform well when true causal expression quantitative trait loci (eQTLs) are sparse and have relatively large effect sizes, while Bayesian Dirichlet process regression is preferred when true causal eQTLs exist in a large number and manifest small effect sizes.[17] Using a different approach for the genotype data depending on the scenario of genotypic effect may improve the imputation performance of the GEN model as well as the combined models. Existing gene expression imputation methods that use the genotype data rely on the *cis*-eQTLs within the small window around the transcription start site because of computational burden. If a suitable method that can exploit *trans*-eQTL information becomes readily available, then it would provide an opportunity for improving the current level of gene expression imputation methods.

Data and code availability

We released the imputation models (in Rdata files) of MERGED_fixed_100 for 45 out of 47 tissues in our ZENODO repository (https://doi.org/10.5281/zenodo.8097305). Because of the file size limit (50 GB) of the ZENODO repository, we were unable to release our imputation models for 2 tissues ("Skin Not Sun Exposed" and "Skin Sun Exposed"). We will be glad to share these omitted files with interested readers upon request.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2023.100223.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF; 2022R1A2B5B02001897) funded by the Korean government, Ministry of Science, and ICT. This work was also supported by the Creative-Pioneering Researchers Program funded by Seoul National University (SNU).

Declaration of interests

B.H. is the CTO of Genealogy Inc.

Received: August 2, 2022
Accepted: May 4, 2023

References

1. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63. https://pubmed.ncbi.nlm.nih.gov/19015660/.
2. (2022). Transcriptome: Connecting the Genome to Gene Function. Learn Science at Scitable. https://www.nature.com/scitable/topicpage/transcriptome-connecting-the-genome-to-gene-function-605/.
3. Byron, S.A., van Keuren-Jensen, K.R., Engelthaler, D.M., Carpten, J.D., and Craig, D.W. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat. Rev. Genet. *17*, 257–271. https://pubmed.ncbi.nlm.nih.gov/26996076/.
4. Supplitt, S., Karpinski, P., Sasiadek, M., and Laczmanska, I. (2021). Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. Int. J. Mol. Sci. *22*, 1–22. https://pubmed.ncbi.nlm.nih.gov/33572595/.
5. Breschi, A., Djebali, S., Gillis, J., Pervouchine, D.D., Dobin, A., Davis, C.A., Gingeras, T.R., and Guigó, R. (2016). Gene-specific patterns of expression variation across organs and species. Genome Biol. *17*, 151. https://pubmed.ncbi.nlm.nih.gov/27391956/.

6. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.v., Aquino-Michaels, K., Carroll, R.J., Eyler, A.E., Denny, J.C., GTEx Consortium, Nicolae, D.L., Cox, N.J., and Im, H.K. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098. https://pubmed.ncbi.nlm.nih.gov/26258848/.
7. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., Stahl, E.A., Huckins, L.M., GTEx Consortium, Nicolae, D.L., Cox, N.J., and Im, H.K. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1825. https://pubmed.ncbi.nlm.nih.gov/29739930/.
8. Barbeira, A.N., Pividori, M., Zheng, J., Wheeler, H.E., Nicolae, D.L., and Im, H.K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet. *15*, e1007889. https://pubmed.ncbi.nlm.nih.gov/30668570/.
9. Halloran, J.W., Zhu, D., Qian, D.C., Byun, J., Gorlova, O.Y., Amos, C.I., and Gorlov, I.P. (2015). Prediction of the gene expression in normal lung tissue by the gene expression in blood. BMC Med. Genom. *8*, 77. https://pubmed.ncbi.nlm.nih.gov/26576671/.
10. Xu, W., Liu, X., Leng, F., and Li, W. (2020). Blood-based multi-tissue gene expression inference with Bayesian ridge regression. Bioinformatics *36*, 3788–3794. https://pubmed.ncbi.nlm.nih.gov/32277818/.
11. Basu, M., Wang, K., Ruppin, E., and Hannenhalli, S. (2021). Predicting tissue-specific gene expression from whole blood transcriptome. Sci. Adv. *7*, eabd6991. https://pubmed.ncbi.nlm.nih.gov/33811070/.
12. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585. https://pubmed.ncbi.nlm.nih.gov/23715323/.
13. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Software *33*, 1–22. https://pubmed.ncbi.nlm.nih.gov/20808728/.
14. Lee, C.H., Cook, S., Lee, J.S., and Han, B. (2016). Comparison of Two Meta-Analysis Methods: Inverse-Variance-Weighted Average and Weighted Sum of Z-Scores. Genomics Inform. *14*, 173–180. https://pubmed.ncbi.nlm.nih.gov/28154508/.
15. Yang, J., Fritsche, L.G., Zhou, X., Abecasis, G.; and International Age-Related Macular Degeneration Genomics Consortium (2017). A Scalable Bayesian Method for Integrating Functional Information in Genome-wide Association Studies. Am. J. Hum. Genet. *101*, 404–416. https://pubmed.ncbi.nlm.nih.gov/28844487/.
16. Parrish, R.L., Gibson, G.C., Epstein, M.P., and Yang, J. (2022). TIGAR-V2: Efficient TWAS tool with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8. HGG Adv. *3*, 100068. https://pubmed.ncbi.nlm.nih.gov/35047855/.
17. Luningham, J.M., Chen, J., Tang, S., de Jager, P.L., Bennett, D.A., Buchman, A.S., and Yang, J. (2020 Oct 1). Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. Am. J. Hum. Genet. *107*, 714–726. https://pubmed.ncbi.nlm.nih.gov/32961112/.