Bioinformatics, 36(12), 2020, 3788–3794
doi: 10.1093/bioinformatics/btaa239
Advance Access Publication Date: 11 April 2020
Original Paper

Gene expression

Blood-based multi-tissue gene expression inference with Bayesian ridge regression

Wenjian Xu, Xuanshi Liu, Fei Leng and Wei Li*

Beijing Key Laboratory for Genetics of Birth Defects, Beijing Pediatric Research Institute, MOE Key Laboratory of Major Diseases in Children, Genetics and Birth Defects Control Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on September 11, 2019; revised on March 15, 2020; editorial decision on April 5, 2020; accepted on April 6, 2020

Abstract

Motivation: Gene expression profiling is widely used in basic and cancer research but still not feasible in many clinical applications because tissues, such as brain samples, are difficult and not ethnical to collect. Gene expression in uncollected tissues can be computationally inferred using genotype and expression quantitative trait loci. No methods can infer unmeasured gene expression of multiple tissues with single tissue gene expression profile as input.

Results: Here, we present a Bayesian ridge regression-based method (B-GEX) to infer gene expression profiles of multiple tissues from blood gene expression profile. For each gene in a tissue, a low-dimensional feature vector was extracted from whole blood gene expression profile by feature selection. We used GTEx RNAseq data of 16 tissues to train inference models to capture the cross-tissue expression correlations between each target gene in a tissue and its preselected feature genes in peripheral blood. We compared B-GEX with least square regression, LASSO regression and ridge regression. B-GEX outperforms the other three models in most tissues in terms of mean absolute error, Pearson correlation coefficient and root-mean-squared error. Moreover, B-GEX infers expression level of tissue-specific genes as well as those of non-tissue-specific genes in all tissues. Unlike previous methods, which require genomic features or gene expression profiles of multiple tissues, our model only requires whole blood expression profile as input. B-GEX helps gain insights into gene expressions of uncollected tissues from more accessible data of blood.

Availability and implementation: B-GEX is available at https://github.com/xuwenjian85/B-GEX.

Contact: liwei@bch.com.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression profiling can explain biological phenotypes by depicting transcriptional changes of tissues and is a powerful and indispensable tool to assist in disease mechanism study, tumor subtyping and pharmacodynamic evaluation (Bullinger *et al.*, 2004; Byron *et al.*, 2016; Costa *et al.*, 2013; Finak *et al.*, 2008; Kwa *et al.*, 2017; van 't Veer *et al.*, 2002; Wang *et al.*, 2019). Gene expression analysis-based knowledge discovery is greatly benefited from several public reference resources, such as Gene Expression Omnibus, Library of Integrated Network-based Cellular Signatures (LINCS, http://www.lincsproject.org) and Genotype-Tissue Expression (GTEx) (Barrett *et al.*, 2013; GTEx Consortium, 2013; Koleti *et al.*, 2018; Wang *et al.*, 2016).

Despite the above advances, gene expression profiling is limited in many clinical fields. Gene expression is tissue-specific; therefore, biopsy samples have to be from the affected tissues of patients. However, some affected tissues like brain in autism patients are difficult and not ethical to collect. The major obstacle of gene expression profiling analysis is the inaccessibility of human tissues. Therefore, an alternative method to circumvent tissue specimen collection will facilitate clinical application of gene expression profiling.

Blood gene expressing profiling analysis have been widely used to identify RNA biomarkers for cancer subtyping and prognosis, chronic disorders, genetic disorders and chronological age (Best *et al.*, 2015; Iqbal *et al.*, 2014; Jansen *et al.*, 2016; Ju *et al.*, 2015; Kusko *et al.*, 2016; Laing *et al.*, 2019; Miller *et al.*, 2016; Peters *et al.*, 2015; Tang *et al.*, 2018; Weinstein *et al.*, 2013). These studies proved that blood is a valuable resource for gene expression analysis.

Gene expression level of an affected tissue may be inferred from blood, a readily accessible specimen material in clinical field. The gene expression in blood is reported to be moderately correlated with that in brain tissues, and weakly correlated with that in immune and muscle tissues (Sullivan *et al.*, 2006). Later, using lung and blood RNA-seq data of 31 donors in GTEx, researchers established a linear regression model for inferring gene expression in lung using age, gender and gene expression in blood (Halloran *et al.*, 2015). They found expression level of 18% genes have significant correlations between lung and blood. Robust statistical methods are also developed for accounting tissue-wise and gene-wise gene expression correlations (Touloumis *et al.*, 2016; Wang *et al.*, 2019). These findings suggest it is possible to infer gene expression of other tissues using the ‘surrogate’ information of blood gene expression profiles.

Gene expression inference in tissues usually requires prior knowledge of genotype-transcriptome associations. For example, by analyzing GTEx dataset, researchers developed multi-tissue gene expression imputation methods PrediXcan and MixRF (Gamazon *et al.*, 2015; Wang *et al.*, 2016). PrediXcan requires a patient’s genotype inputs to infer genetically regulated component of gene expression profile. PrediXcan has been used to identify disease and drug-associated genes (Gottlieb *et al.*, 2017; Huckins *et al.*, 2019). The performance of imputation relies largely on the completeness of the expression quantitative trait locus (eQTL) reference resource. Because these methods explicitly require genomic features (e.g. SNP genotype and eQTLs) as part of the predictors to infer unmeasured gene expression in target tissues, genotyping patients must be done before gene expression inference.

To the best of our knowledge, no available methods can infer gene expression of multiple tissues without genotype. Here, we present a machine-learning-based method for gene expression inference of multiple uncollected tissues using blood gene expression profile (B-GEX). B-GEX is a set of tissue-specific multi-task linear regression model. We define multiple genes in blood as feature variables and each gene in another tissue as one target. To achieve this goal, we built an independent model for each target. We used ℓ_1-norm-based feature selection method to select the most significant correlated genes expressed in blood specific to the target gene. We adopted Bayesian ridge regression (BayR) to model the cross-tissue gene expression correlations between the features and the target. We evaluated the performances of B-GEX and least square regression (LSR), LASSO regression and ridge regression on GTEx RNAseq dataset of 16 tissues. Our results show that B-GEX outperforms the other three linear regression models in the majority of tissues and target genes. Finally, we explored the most frequently selected blood feature genes to interpret the clustering patterns revealed in the tissue-level comparisons.

2 Materials and methods

2.1 Datasets

GTEx RNAseq expression dataset (2016-01-15 v7) was downloaded from GTEx Portal (https://gtexportal.org). Original data consist of 11 688 gene expression profiles across 53 tissues from 714 individuals. Besides whole blood (B), there are 52 tissues in GTEx data. Each profile contains expression level of 56 202 transcripts in a tissue. We consider as one individual as one sample. Gene expression profiles of multiple tissues are associated with each sample. The usable samples for tissue (T_k) gene expression inference are defined as the samples in which both B and T_k gene expression data are available. A total of 26 tissues have more than 100 usable samples and were chosen for further analysis (Table 1). Our goal is to infer gene expression in T_k ($k = 1,\ 2,\ \cdots,\ 26$) with gene expression in B. The inference models are tissue-specific, so we built an independent dataset for each tissue as follows.

For each tissue T_k, we made a dataset for training and test purposes. First, get the sample IDs of its usable samples. Second, extract gene expression data of B associated with these sample IDs as *raw feature* data, and extract gene expression data of T_k associated with these sample IDs as *raw target* data. Third, remove the *raw feature* genes not expressed in all usable samples to make *clean feature* data, meanwhile remove the *raw target* genes not expressed in all usable samples to make *clean target* data. The dimension of clean feature data X and clean target data Y is in the range of 12 000–13 000 and 14 000–24 000, respectively. Next, randomly separate samples into training set (90%) and test set (10%). In training set preprocessing, logarithm of transcripts per million is transformed to standardized values by subtracting the mean μ and dividing the SD δ of each gene. These values of μ and δ are stored and used later because test data (and other future data) should be preprocessed in the same way as the training set.

Table 1. Summary of usable GTEx tissues and samples

Tissue	#Sample	#Target gene	#Feature gene
Adipose—subcutaneous	278	16 688	12 453
Adipose—visceral (omentum)	197	16 953	13 001
Adrenal gland	121	17 845	13 187
Artery—tibial	283	15 712	12 684
Artery—aorta	202	16 980	12 793
Artery—coronary	116	17 422	13 514
Brain—cerebellum	101	19 531	13 555
Breast—mammary tissue	173	17 261	12 983
Colon—transverse	173	17 388	13 542
Colon—sigmoid	135	17 722	13 604
Esophagus—mucosa	247	17 075	12 615
Esophagus—muscularis	224	17 151	12 799
Esophagus–gastroesophageal junction	139	17 537	13 668
Heart—left ventricle	196	15 208	12 855
Heart—atrial appendage	171	16 277	13 456
Liver	105	16 444	13 610
Lung	271	17 363	12 407
Muscle—skeletal	341	14 505	12 322
Nerve—tibial	257	17 678	12 597
Pancreas	156	17 077	13 110
Pituitary	104	19 662	13 632
Skin—sun exposed (lower leg)	289	16 896	12 387
Skin—not sun exposed (suprapubic)	203	17 487	12 919
Stomach	165	16 800	13 527
Testis	163	24 093	12 892
Thyroid	256	18 077	12 415

To show the model can learn meaningful structure rather than spurious structure of dataset, the candidate model is also evaluated on a random-generated data of similar numeric structure with GTEx data. For example, X is a gene x individual table. The random data X_{rand} is constructed by randomly permutated each row of X using a different seed.

2.2 Framework of multi-tissue gene expression inference

In last section, we have described the preparation of feature data X and target data Y of tissue T_k. Each tissue corresponds to a pair of target data and feature data. We formulate an inference framework consisting of multiple tissue-specific models (Fig. 1). For tissue T_k, assume there are N training samples (paired feature data X and target data Y), P blood expressing feature genes in data X, Q tissue expressing target genes in data Y. The training dataset (X, Y) is expressed as $D = \{x_i,\ y_i\}_{i=1}^{N}$, where $x_i \in \mathcal{R}^P$ denotes expression profile of feature genes in the i-th sample and $y_i \in \mathcal{R}^Q$ denotes the expression profile of target genes in the i-th sample. In this framework, model$_k$ is a multi-task regression model for inferring gene expressions of tissue T_k using gene expressions of tissue B as input. The whole inference model$_k$ is a functional mapping $\mathcal{F}_k : \mathcal{R}^P \to \mathcal{R}^Q$ that fits the dataset $D = \{x_i,\ y_i\}_{i=1}^{N}$.

model$_k$ consists of many basic tasks, which corresponds to inference of a single gene’s expression $y_{i(q)} (q = 1, 2, \ldots, Q)$ based on the input x_i. Each basic task $y_{i(q)}$ can be solved independently. Now, a basic regression task $\mathcal{F}_{(q)}$ for each target gene q is independently formulated as: $\mathcal{F}_{k(q)} : \mathcal{R}^P \to \mathcal{R}^1$, and to be trained with dataset

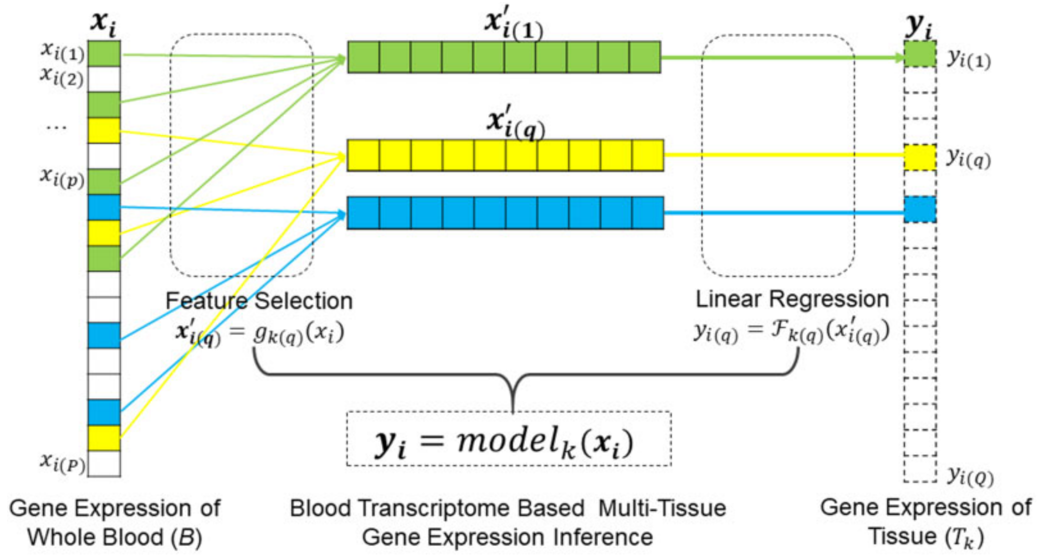
x_i
$x_{i(1)}$
$x_{i(2)}$
…
$x_{i(p)}$
$x_{i(P)}$
$x'_{i(1)}$
$x'_{i(q)}$
Feature Selection
$x'_{i(q)} = g_{k(q)}(x_i)$
Linear Regression
$y_{i(q)} = \mathcal{F}_{k(q)}(x'_{i(q)})$
$y_i = model_k(x_i)$
y_i
$y_{i(1)}$
$y_{i(q)}$
$y_{i(Q)}$
Gene Expression of Whole Blood (B)
Blood Transcriptome Based Multi-Tissue Gene Expression Inference
Gene Expression of Tissue (T_k)

Fig. 1. Illustration of B-GEX model architecture. model$_k$ is a multi-task regression model for inferring gene expressions of tissue T_k. The model has two parts: feature selection function $g_{k(q)}$ and linear regression function $\mathcal{F}_{k(q)}$. $x_i \in \mathcal{R}^P$ denotes expression profile of feature genes in the i-th sample and $y_i \in \mathcal{R}^Q$ denotes the expression profile of target genes in the i-th sample. For each gene q, $x'_{i(q)}$ is a subset of $P'_{(q)}$ most relevant predictive features, which are generated from x_i by feature selection function $g_{k(q)}$. The $\mathcal{F}_{k(q)}$function takes $x'_{i(q)}$ vector as input and output inferred expression scalar value $y_{i(q)}$

$D_q = \{x_i,\ y_{i(q)}\}_{i=1}^N$, where $x_i \in \mathcal{R}^P$and $y_{i(q)} \in \mathcal{R}^1$. Since $y_{i(q)}$ is a subset of y_i, $\mathcal{F}_{k(q)}$ is simpler to solve than $\mathcal{F}_k$. We can train $\mathcal{F}_{k(q)}$ one by one and assemble them as $\mathcal{F}_k$.

Next, we consider how to solve the basic inference task $\mathcal{F}_{k(q)}$. Its feature length P is two orders of magnitudes larger than sample size N. So $\mathcal{F}_{k(q)}$ is a typical high-dimensional low-sample-size problem. We have to reduce the dimension of the feature space before we actually train regression predictive model. A subet consists of $P'_{(q)}$ most relevant predictive feature set $x'_{i(q)}$ from x_i is independently selected for target gene q, where $x'_{i(q)} \in \mathcal{R}^{P'_{(q)}}$ and $P'_{(q)} < N \ll P$. Finally, the basic regression task for each target gene q is reformulated as: $\mathcal{F}_{k(q)} : \mathcal{R}^{P'_{(q)}} \rightarrow \mathcal{R}^1$, and trained with $D_q = \left\{x'_{i(q)},\ y_{i(q)}\right\}_{i=1}^N$, where $x'_{i(q)} \in \mathcal{R}^{P'_{(q)}}$ and $y_{i(q)} \in \mathcal{R}^1$.

2.3 Evaluation metrics

We use mean absolute error (MAE) to evaluate the predictive performance of a target gene in each tissue under 5-fold cross-validation scheme.

$$\text{MAE}_{(q)} = \frac{1}{N'} \sum\nolimits_{i=1}^{N'} \left(y_{i(q)} - \hat{y}_{i(q)} \right) \quad (1)$$

where N' is the number of test sample size and $\hat{y}_{i(q)}$ is the model predicted expression values of target gene q in the i-th sample.

Root-mean-squared error (RMSE) and the Pearson correlation coefficient (r) is used as the secondary evaluation metrics:

$$\text{RMSE}_{(q)} = \sqrt{\frac{1}{N'} \sum\nolimits_{i=1}^{N'} \left(y_{i(q)} - \hat{y}_{i(q)} \right)^2} \quad (2)$$

$$r_{(q)} = \frac{\sum_{i=1}^{N'} \left(y_{i(q)} - \bar{y}_{i(q)} \right) \left(\hat{y}_{i(q)} - \bar{\hat{y}}_{i(q)} \right)}{\sqrt{\sum_{i=1}^{N'} \left(y_{i(q)} - \bar{y}_{i(q)} \right)^2} \sqrt{\sum_{i=1}^{N'} \left(\hat{y}_{i(q)} - \bar{\hat{y}}_{i(q)} \right)^2}} \quad (3)$$

We define overall error of model$_k$ as the average error of all target genes of tissue T_k.

2.4 Feature selection

To reduce the dimension of the feature space, we quantify the linear dependencies of target genes in tissue T_k (response variables) on feature genes in B (predictor variables) and select a small set of the most relevant B features for each target gene q expressed in T_k. This part only uses training set.

Step 1: for each target gene q, we compute coefficient of variation (cv $= \delta/\mu$) of each features and rank them in descending order. We select those top 10% features to construct a low-dimensional new subsets $D_q^0 = \left\{x_{i(q)}^0,\ y_{i(q)}\right\}_{i=1}^N$, where $x_{i(q)}^0 \in \mathcal{R}^{P^0}$, $P^0 < N$.

Step 2: next, we further reduce the dimension of feature space. $y_{i(q)}$ of a group of samples forms an expression value vector $y_{(q)}$ and expression value vectors of feature genes form the columns of matrix $\boldsymbol{x}_{(q)}^0$. We assume that one feature gene is important when the absolute of cosine similarity between vector $y_{(q)}$ and the corresponding column of $\boldsymbol{x}_{(q)}^0$ is close to 1. We can reduce feature number to any arbitrary number with this cosine similarity approach. By adjusting this hyperparameter appropriately and assessing the predictive performance of baseline model (see Section 2.5), we can determine the optimal number of feature genes associated with T_k and construct a new subset $D_q^1 = \left\{x_{i(q)}^1,\ y_{i(q)}\right\}_{i=1}^N$, where $x_{i(q)}^1 \in \mathcal{R}^{P^1}$, $P^1 < P^0 < N$.

2.5 Baseline and candidate regression models

We use LSR as the baseline method.

$$w_{(q)}, b_{(q)} = \arg\min_{w,\ b} \frac{1}{N} \sum_{i=1}^N \left(y_{i(q)} - w_{(q)}^T x_{i(q)}^1 - b_{(q)} \right)^2 \quad (4)$$

where $w_{(q)} \in \mathcal{R}^{P^1}, b_{(q)} \in \mathcal{R}^1$ are parameters associated with each target gene q.

We also include three candidate models in the model comparison, which are LSR-L1:

$$w_{(q)}, b_{(q)} = \arg\min_{w,\ b} \frac{1}{N} \sum_{i=1}^N \left(y_{i(q)} - w_{(q)}^T x_{i(q)}^1 - b_{(q)} \right)^2 + \alpha ||w_{(q)}||_1 \quad (5)$$

LSR with ℓ_2-norm regularization (LSR-L2, ridge regression):

$$w_{(q)}, b_{(q)} = \arg\min_{w,\ b} \frac{1}{N}\sum_{i=1}^{N}\left(y_{i(q)} - w_{(q)}^{T}x_{i(q)}^{1} - b_{(q)}\right)^{2} + \alpha||w_{(q)}||_{2} \quad (6)$$

and BayR, which is similar to LSR-L2 except the regularization parameter α and noise precision parameter λ are chosen from gamma distribution and is estimated jointly during model fitting. The probabilistic model of y is

$$p\left(y_{i(q)}|x_{i(q)}^{1}, w_{(q)}, b_{(q)},\ \beta\right) = N\left(y_{i(q)}|w_{(q)}^{T}x_{i(q)}^{1} + b_{(q)},\ \alpha^{-1}\boldsymbol{I}\right) \quad (7)$$

Gaussian prior of coefficients $w_{(q)}$ and bias $b_{(q)}$ is

$$p\left(w_{(q)}, b_{(q)}|\alpha\ \right) = N\left(w_{(q)}, b_{(q)}|0,\ \lambda^{-1}\boldsymbol{I}\right) \quad (8)$$

The hyperparameters of BayR model are gamma priors over α and λ, which are usually non-informative. LSR-L2 models use α=0.04, max_iter=1e5. Elsewhere default values of hyperparameters provided by scikit-learn linear models (*LinearRegression*, *Lasso*, *Ridge* and *BayesianRidge*) are used.

We implement feature selection and predictive models using scikit-learn v0.21 (Pedregosa *et al.*, 2011). The total computation time of cosine similarity-based feature selection (10 features for each target gene) and BayR model training for 26 tissues is 16 min on linux x86_64 server [2 x Intel(R) Xeon(R) Gold 6144 CPU 3.50 GHz, 128 Gb RAM].

3 Results

3.1 Construction of optimal feature sets

We use the baseline prediction model LSR to help selecting most relevant features from all features and construct an optimal feature set. The way to find the optimal feature set is extracting different number of features in cosine similarity feature selection and compares their performances.

Feature space is reduced to 10% of original one after feature selection step 1 with the whole-training set. Next, we randomly split the training set into 90% ‘inner’ training set and 10% validation set. We use MAE and *r* metrics of the baseline model to evaluate gene expression prediction errors (on validation set) with a series of selected feature sets generated in feature selection step 2 (on ‘inner’ training set). Averaged metric values across five trials with different randomization seeds are recorded to stabilize measurement of MAE and *r*.

The optimal feature sets of gene expression inference are tissue-specific. We consider a given feature set as the best for a specific tissue when the baseline model trained with this reduced feature sets have both the smallest average MAE and the largest average *r* in the serial experiments of this tissue’s target gene expression inference tasks.

The relative trend of MAE and *r* with respect to the number of final select features is shown in Figure 2. For each tissue, we do experiments with feature number range of [5, 10, 15, 20, 25, 30], rank the experiments by average MAE in ascending order (or by average *r* in descending order), then color-code the experiments by rankings #1 ~ #7 as blue ~ white (Fig. 2b and c). The ranking #1 experiment represents the optimal feature set in each tissue with respect to MAE or *r*. We observe that the optimal feature number is constantly 5 according to MAE metric and 5 ~ 15 according to *r* metric. The optimal feature number is not apparently correlated with sample size (Fig. 2a). A total of 10 feature genes per target gene are acceptable for the most tissues and chosen as the optimal feature set for all tissues.

Now, we have constructed a low-dimensional optimal feature set for gene expression inference models to use.

3.2 Usage pattern of selected blood features

We next ask whether any common feature genes exist in 26 tissue-specific feature sets. To do this, we define the feature gene usage as how many times a specific feature gene has been selected across all target genes of a tissue. We calculate the usage of blood genes in each target tissue then categorize 2942 blood genes as 5 groups by usage counts: ‘0’, ‘11~10’, ‘11~100’, ‘101~400’ and ‘401~’. The usage pattern of feature genes with respect to 26 tissues suggests

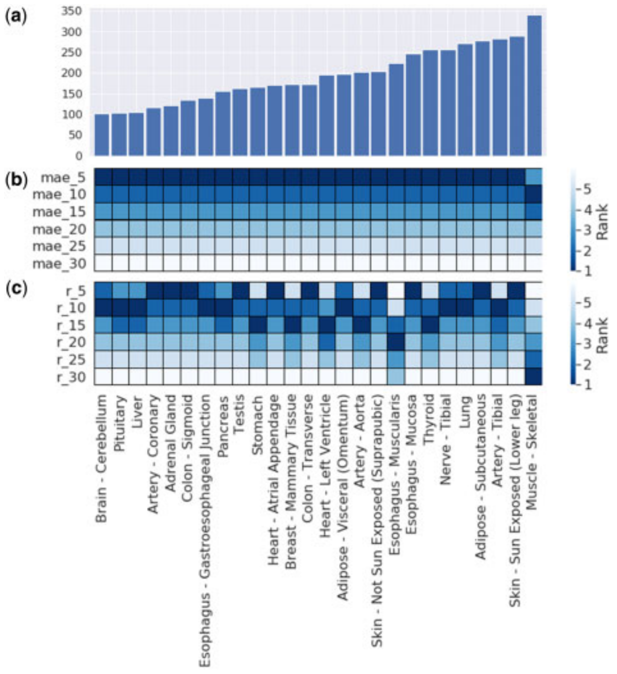

Fig. 2. Determination of optimal feature length of tissues. (**a**) The sample size of tissues. For each tissue, selected feature number is in range of [5, 10, 15, 20, 25, 30]. (**b**) The heatmap of MAE rankings with respect to feature length. Rank the experiments by average MAE in ascending order, then color-code the experiments by rankings #1 ~ #7 as blue ~ white. (**c**) The heatmap of *r* rankings with respect to feature length. Rank the experiments by average *r* in descending order and color-code by rankings #1 ~ #7 as blue ~ white. (Color version of this figure is available at *Bioinformatics* online.)

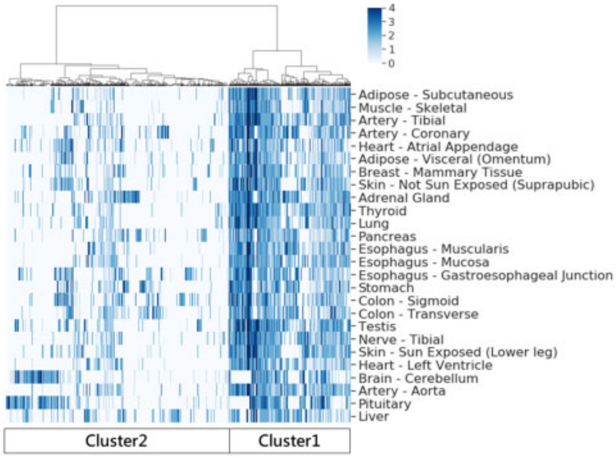

Fig. 3. Hierarchical clustering of the gene expression relevance between blood genes and 26 tissues. We calculate the usage of 2942 blood genes in each target tissue then categorized the blood genes by usage counts: ‘0’, ‘1~10’, ‘11~100’, ‘101~400’ and ‘401~’. The usage metric category of a blood gene represents its expression relevance to the tissues. Usage of blood feature genes with respect to different tissues is plotted. The usage metric vectors of blood genes are column vectors; tissues are row vectors. Calculate the pairwise Euclidean distances between row vectors and visualize hierarchical clustering pattern. Cluster linkage method for rows vectors is Ward variance minimization

Table 2. Overall errors of LSR, LSR-L1, LSR-L2 and BayR in tissue gene expression inference

	MAE			
Tissue	LSR	LSR-L1	LSR-L2	BayR
Adipose—visceral (omentum)	0.759±0.145	0.747±0.143	0.758±0.145	0.747±0.144
Adrenal gland	0.816±0.189	0.795±0.186	0.813±0.189	0.790±0.184
Artery—aorta	0.779±0.146	0.767±0.143	0.778±0.146	0.766±0.143
Artery—coronary	0.811±0.197	0.785±0.187	0.807±0.196	0.778±0.184
Breast—mammary tissue	0.776±0.157	0.758±0.154	0.775±0.157	0.758±0.153
Colon—sigmoid	0.704±0.164	0.674±0.158	0.700±0.164	0.668±0.156
Esophagus—mucosa	0.676±0.129	0.666±0.125	0.676±0.129	0.665±0.125
Esophagus—muscularis	0.676±0.124	0.667±0.122	0.675±0.124	0.667±0.122
Heart—atrial appendage	0.672±0.128	0.658±0.124	0.670±0.128	0.655±0.124
Heart—left ventricle	0.608±0.121	0.598±0.120	0.607±0.121	0.599±0.120
Liver	0.788±0.237	0.766±0.232	0.785±0.236	0.764±0.229
Lung	0.736±0.131	0.727±0.128	0.735±0.131	0.726±0.128
Muscle—skeletal	0.676±0.104	0.672±0.103	0.676±0.104	0.671±0.103
Nerve—tibial	0.673±0.116	0.666±0.114	0.673±0.116	0.664±0.114
Pancreas	0.729±0.141	0.717±0.138	0.728±0.141	0.713±0.137
Thyroid	0.670±0.123	0.657±0.120	0.669±0.123	0.656±0.120

Note: Numbers before and after '±' sign are average MAEs and SDs of MAEs of all target genes in one specific tissue.

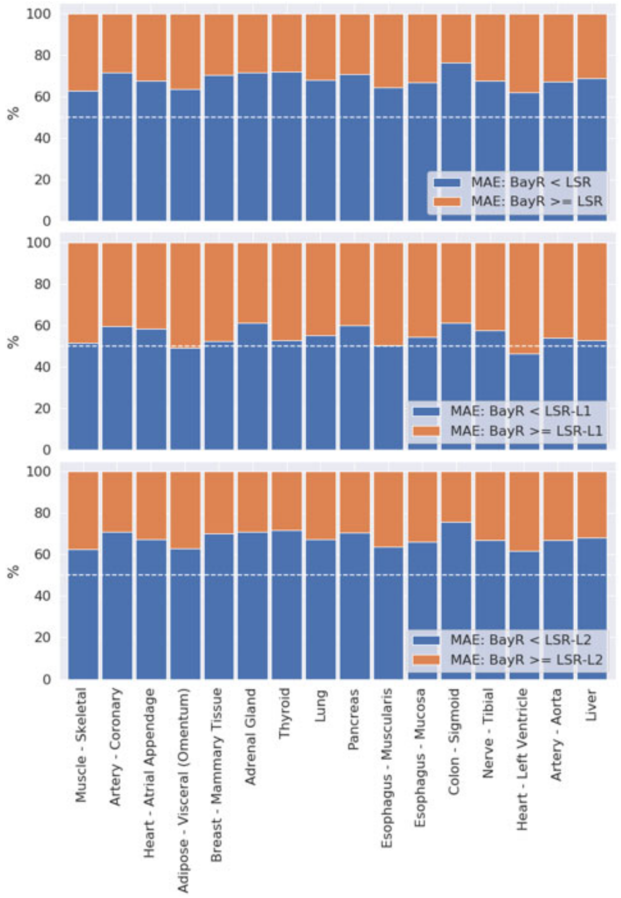

Fig. 4. Gene-wise comparision of the model performances with MAEs of all target genes. Top, middle and bottom plots illustrate the gene-wise MAE comparison of BayR versus LSR, BayR versus LSR-L1 and BayR versus LSR-L2, respectively. In these three stacked barplots, the proportion of target genes in which BayR outperforms the other model is blue; the rest of target genes are orange. (Color version of this figure is available at *Bioinformatics* online.)

that there are two clusters (Fig. 3). Cluster #1 contains blood feature genes have both strong and broad gene expressing correlations in almost all tissues. Cluster #2 contains blood feature genes specific to a couple of tissues. The results agree with the fact that blood expressing genes have intricate correlations with tissue expressing genes.

We then focus on the most frequently selected feature genes by each tissue. Taking a union of top 10 feature genes from each tissue, a representative set of 112 most frequently selected feature genes was made. Here, 97.1% of most frequently selected blood genes are shared by two tissues; 90.2% of most frequently selected blood genes are shared by three and more tissues. In contrast, the ratio for all blood features is 86.1% and 75.5%, respectively. The results suggest that a small proportion of blood genes of highest usage indeed has both strong and broad gene expressing correlation with multiple tissues.

3.3 Performance of inference models

We train four inference models as described in Section 2.5 and evaluate model performances on the test set. Gene expression inference using blood gene expression profile may not be successful for all genes and all tissues. As quality check point, we designed two assessment rules to help decide whether blood-based inference models are satisfactory for each gene and each tissue. Rule 1, if target gene's MAE <0.7 and $r > 0.3$, the gene is predictable. Rule 2, if the ratio of predictable genes to total genes >0.2, the tissue is predictable. A total of 16 out of 26 tissues is considered predictable (Supplementary Table S1) and used in following analysis.

Table 2 and Supplementary Table S2 show the overall performances of LSR, LSR-L1, LSR-L2 and BayR inference models in terms of MAE and r metric. Relative improvement of overall inference error MAE of BayR is 0.7%~75.1% over LSR, 0.77%~4.66% over LSR-L2; the performance of BayR is better than LSR-L1 in 14 tissues and worse than LSR-L1 in the other 2 tissues where BayR is $-0.007\% \sim -0.1380070138\%$ over LSR-L1. The same trend is observed when we use RMSE metric (Supplementary Table S2). We also evaluate model performances with r metric: BayR is the best in 15 out of 16 predictable tissues (Supplementary Table S2). In all, BayR inference model outperforms the other candidate models in most of our experiments. A BayR inference model trained with random generated data of similar numeric structure with GTEx data performs much worse in all metrics (Supplementary Fig. S1). This evidence suggests that our model learns real not spurious structure of input data.

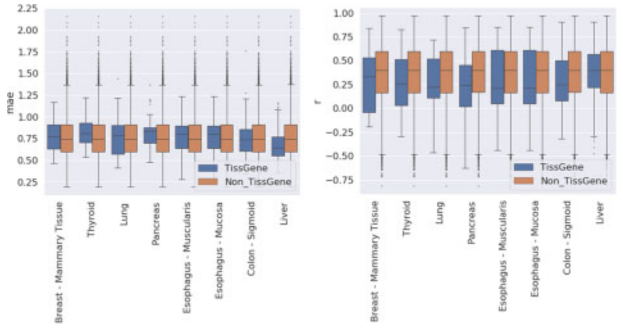

Fig. 5. Gene expression inference performance of B-GEX model on tissue-specific genes of eight tissues. Group 'TissGene' represents a set of tissue-specific genes supported by GTEx data and either HPA data or TiGER data. Group 'Non_TissGene' represents the rest of genes in tissues. Metric value distributions of two groups are shown side by side in eight tissues

3.4 Gene-wise comparison of inference models

The models' performance is further analyzed at gene level. Gene-wise comparison results show that BayR is better than LSR and LSR-L2 in more than 61% target genes in all tissues, better than LSR-L1 in above 50% of target genes in 14 tissues in terms of both MAE and RMSE (Fig. 4 and Supplementary Fig. S2). Similarly, BayR also outperforms the other models in the 14 tissues in *r* metric.

3.5 B-GEX inference of tissue-specific genes

The BayR-based multi-tissue Gene EXpression inference model is abbreviated as B-GEX. To better understand the tissue-specific performance of B-GEX, we look closer at B-GEX performance on a published set of more than 2000 tissue-specific genes named TissGene (http://zhaobioinfo.org/TissGDB)(Kim *et al.*, 2018). The researchers collected the TissGene list from three representative tissue-specific gene expression resources: the Human Protein Atlas (HPA) (Uhlen *et al.*, 2015), Tissue-specific Gene Expression and Regulation (TiGER) (Liu *et al.*, 2008) and GTEx. We found that the gene-level MAE and *r* distribution of one tissue's TissGenes under B-GEX inference model are comparable to that of the tissue's non-TissGenes (Fig. 5). This observation shows B-GEX can infer the expression level of tissue-specific genes as well as other target genes and further validated the B-GEX model has no bias toward genes.

3.6 Usage guide of B-GEX

Our results suggest that the method can partially infer gene expression of 16 predictable tissues out of 26 available tissues and robustly infer gene expression of 20%∼60% genes in a given tissue. For convenience, we separate genes into predictable and non-predictable genes according to empirical quality check results. B-GEX outputs inferred gene expression values from blood gene expression profiles together with empirical quality check marks (MAE, RMSE, *r* and predictability). Users can filter the output by these marks and should be cautious when the genes of interest fall into the non-predictable category.

4 Discussion

In this article, we present a machine-learning method for gene expression inference in multiple tissues using gene expression level in whole blood. Compared to previous eQTLs-based methods for cross-tissue gene expression inference, B-GEX achieves one major improvement: B-GEX depends on blood-tissue transcriptome correlation and only requires user to input gene expression profile of blood, a clinical sample mostly convenient to collect. Moreover, with the advance of next-generation sequencing technology, gene expression profiling may soon become a routine laboratory test to support blood specimen-based clinical test. Our model accepts a patient's blood expression profile, decomposes the gene expression information and outputs inferred gene expression profiles of multiple tissues. Therefore, B-GEX can produce multiple transcriptomes at no cost once blood-based RNAseq is done.

B-GEX model used multi-task linear regression to address 'cross-tissue' inference. Similar design has recently been used to do 'within-tissue' gene expression inference. 'within-tissue' inference focus on how to complete an expression profile based on partially measured expression profiles. Researchers from the LINCS program recently developed low-cost transcriptomics L1000, which experimentally measure expression values of 978 preselected landmark genes and computationally infer the expression profiles of remaining target genes using multi-task linear regression (Subramanian *et al.*, 2017). Later, two deep-learning-based methods D-GEX and GGAN make use of the hierarchical non-linear relation among genes and achieved more than ∼15% improvement on the L1000 inference task (Chen *et al.*, 2016; Wang *et al.*, 2018). B-GEX suggests that multi-task linear regression is also a good starting point for 'cross-tissue' inference.

By comparing four inference models, we have shown that B-GEX outperforms the baseline LSR and its two derivative models on GTEx RNAseq data. LSR-L2 inference model is a competitive model in target genes of several tissues. We believe more accurate inferences of gene expression can be achieved if complicated models or latent variable based models are developed in future.

In future, we plan to improve gene expression inference models using other feature selection and feature extraction techniques. The over-fitting problem is probably not totally resolved at this moment due to the low sample size of usable dataset. In addition, it should be cautioned that the performance evaluation of gene expression inference model solely relies on GTEx RNAseq dataset. Independent validation test on other datasets is necessary in follow-up studies, especially on those in-house datasets.

We looked into the usage pattern of blood feature genes using hierarchical clustering, and revealed the presence of a set of blood genes having wide gene expression correlation with multiple tissues. For example, the representative set of most frequently selected 112 (union of top 10) blood genes have clear pattern of tissue specificity (Supplementary Fig. S3). Many tissues, including muscle—skeletal, breast—mammary tissue, heart—left ventricle and brain—cerebellum, form tissue-specific tight cluster of blood gene markers. We observe discernible pattern of gene expression level association with blood marker genes in the above tissue groups. In contrast, two skin tissues (skin—sun exposed and skin—not sun exposed) and two adipose tissues (adipose—visceral and adipose—subcutaneous) do not show a clean specific cluster but a wide spectrum of relevant blood gene markers. The weaker non-specific cluster of tissues may be explained by the tissue correlation matrix as reported in Touloumis *et al.* (2016). However, B-GEX can only infer expression value of predictable genes rather than all genes, the real source of the cluster pattern is difficult to reveal in this moment. A systematic analysis of blood feature genes will provide insights into gene expression regulatory paradigm between blood and other tissues. It is also interesting to investigate those blood feature genes, which are unique to a tissue and figure out why they have correlation with so many target genes in a specific tissue.

Acknowledgements

The authors are thankful to Dr Zhenhua Hao, Dr Yefeng Yuan, Hao Hong and Zilin Ren for helpful discussions.

Funding

This work was supported by grants from the Ministry of Science and Technology of China [2016YFC1000306] and the Beijing Municipal Science and Technology Commission Foundation [Z181100001918003].

Conflict of Interest: none declared.

References

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.

Best,M.G. *et al.* (2015) RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell*, **28**, 666–676.

Bullinger,L. *et al.* (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N. Engl. J. Med.*, **350**, 1605–1616.

Byron,S.A. *et al.* (2016) Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.*, **17**, 257–271.

Chen,Y. *et al.* (2016) Gene expression inference with deep learning. *Bioinformatics*, **32**, 1832–1839.

Costa,V. *et al.* (2013) RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.*, **21**, 134–142.

Finak,G. *et al.* (2008) Stromal gene expression predicts clinical outcome in breast cancer. *Nat. Med.*, **14**, 518–527.

Gamazon,E.R. *et al.*; GTEx Consortium. (2015) A gene-based association method for mapping traits using reference transcriptome data. **47**, 1091–1098.

Gottlieb,A. *et al.* (2017) Cohort-specific imputation of gene expression improves prediction of warfarin dose for African Americans. *Genome Med.*, **9**, 98.

GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Halloran,J.W. *et al.* (2015) Prediction of the gene expression in normal lung tissue by the gene expression in blood. *BMC Med. Genomics*, **8**, 77.

Huckins,L.M. *et al.*; CommonMind Consortium. (2019) Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.*, **51**, 659–674.

Iqbal,J. *et al.* (2014) Gene expression signatures delineate biological and prognostic subgroups in peripheral T-cell lymphoma. *Blood*, **123**, 2915–2923.

Jansen,R. *et al.* (2016) Gene expression in major depressive disorder. *Mol. Psychiatry*, **21**, 339–347.

Ju,W. *et al.* (2015) Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Am. J. Respir. Crit. Care Med.*, **7**, 316ra193.

Kim,P. *et al.* (2018) TissGDB: tissue-specific gene database in cancer. *Nucleic Acids Res.*, **46**, D1031–D1038.

Koleti,A. *et al.* (2018) Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res.*, **46**, D558–D566.

Kusko,R.L. *et al.* (2016) Integrated genomics reveals convergent transcriptomic networks underlying chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.*, **194**, 948–960.

Kwa,M. *et al.* (2017) Clinical utility of gene-expression signatures in early stage breast cancer. *Nat. Rev. Clin. Oncol.*, **14**, 595–610.

Laing,E.E. *et al.* (2019) Identifying and validating blood mRNA biomarkers for acute and chronic insufficient sleep in humans: a machine learning approach. *Sleep*, **42**, zsy186.

Liu,X. *et al.* (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.

Miller,J.R. *et al.* (2016) RNA-Seq of Huntington's disease patient myeloid cells reveals innate transcriptional dysregulation associated with proinflammatory pathway activation. *Hum. Mol. Genet.*, **25**, 2893–2904.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Peters,M.J. *et al.*; NABEC/UKBEC Consortium. (2015) The transcriptional landscape of age in human peripheral blood. *Nat. Commun.*, **6**, 8570.

Subramanian,A. *et al.* (2017) A next generation connectivity map: l 1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.

Sullivan,P.F. *et al.* (2006) Evaluating the comparability of gene expression in blood and brain. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, **141B**, 261–268.

Tang,X.R. *et al.* (2018) Development and validation of a gene expression-based signature to predict distant metastasis in locoregionally advanced nasopharyngeal carcinoma: a retrospective, multicentre, cohort study. *Lancet Oncol.*, **19**, 382–393.

Touloumis,A. *et al.* (2016) HDTD: analyzing multi-tissue gene expression data. *Bioinformatics*, **32**, 2193–2195.

Uhlen,M. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Wang,J. *et al.* (2016) Imputing gene expression in uncollected tissues within and beyond GTEx. *Am. J. Hum. Genet.*, **98**, 697–708.

Wang,J. *et al.* (2019) RNA sequencing (RNA-Seq) and its application in ovarian cancer. *Gynecol. Oncol.*, **152**, 194–201.

Wang,M.Y. *et al.* (2019) Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. *Ann. Appl. Stat.*, **13**, 1103–1127.

Wang,X. *et al.* (2018) Conditional generative adversarial network for gene expression inference. *Bioinformatics*, **34**, i603–i611.

Weinstein,J.N. *et al.*; The Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.