Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Genetics and population analysis

xQTLImp: efficient and accurate xQTL summary statistics imputation

Tao Wang [1,†], Quanwei Yin [2,3,†], Yongzhuang Liu [1], Jin Chen [4], Yadong Wang [1,*] and Jiajie Peng [2,3,*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China, [2]School of Computer Science, [3]Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an, 710072, China, [4]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA.

[*]To whom correspondence should be addressed. [†]Equal contributor.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Quantitative trait locus (xQTL) analysis to multi-omic molecular features, such as gene transcription (eQTL), DNA methylation (mQTL) and histone modification (haQTL), is being widely used to decipher the effect of genomic variations on multi-level molecular activities. However, missing genotypes and limited effective sample size largely reduce the power of detecting significant xQTLs in single study or integrating multi-studies for meta-analysis. While existing hidden Markov models (HMMs) based imputation approaches require individual-level genotypes and molecular traits, there is still no available implementation suitable for imputation of xQTL summary statistics, which is becoming widely available and useful.

Results: We developed xQTLImp, a C++ software specially designed to efficiently impute xQTL summary statistics based on multivariate Gaussian approximation. Experiments on a single cell eQTL dataset demonstrated that considerable amount of novel significant eQTL associations can be rediscovered by xQTLImp.

Availability: Software is available at https://github.com/hitbc/xQTLimp.

Contact: ydwang@hit.edu.cn or jiajiepeng@nwpu.edu.cn

Supplementary information: Supplementary data are available online.

1 Introduction

Quantitative trait locus (QTL) analyses to multi-omic phenotypes have been proven useful in understanding the complex functionally regulatory natures of disease associated variants unequivocally identified by genome wide association studies (GWAS) (Ng *et al.*, 2017). Summary statistics of xQTL studies are becoming widely useful in estimating causal effects of GWAS variants such as in mendelian randomization analysis or meta-analysis to boost the statistical power (Consortium *et al.*, 2017). In analyzing summary statistics of multiple xQTL studies, one of the critical problems is missing xQTL statistics in each study, which may bias the findings and cause false negative results because of power lost (Furukawa *et al.*, 2006). The missing statistics can be mainly caused by two reasons. First, missing of genotypes may exist in different studies because of different genotyping platforms, genotype imputation methods, or quality control procedures. Although there are excellent hidden Markov models (HMMs)-based genotype imputation tools, individual-level genotypes are obligatory which in practice are usually hard to access due to ethic limitations (Marchini and Howie, 2010). Second, xQTL analysis require both genotypes and molecular traits for the same individual, which might cause limited effective sample size due to factors like cost or sample availability. Small sample size will limit the MAF lower bound in association tests, which will cause statistics missing even if the genotype is present. For example, the recent single cell eQTL studies can only perform QTL analyses on variants with MAF ≥ 0.1 (van der Wijst *et al.*, 2018) and 0.2 (Kang *et al.*, 2018) with effective sample size to be 45 and 23 respectively.

To the best of our knowledge, there is no existing tool that is designed for xQTL summary statistics imputation. To fill this gap, we implemented xQTLImp, a C++ tool specially designed for efficiently imputing xQTL summary statistics. xQTLImp models statistics of variants in linkage

disequilibrium (LD) and associated with same molecular feature by using multivariate Gaussian model (Supp. materials), which has been successfully applied in GWAS era, such as multiple correlated marker correction (Conneely and Boehnke, 2007; Han *et al.*, 2009), fine mapping (Zaitlen *et al.*, 2010) and GWAS imputation (Wen and Stephens, 2010; Pasaniuc *et al.*, 2014; Kwan *et al.*, 2016; Julienne *et al.*, 2018). Experiments on real eQTL, mQTL and haQTL datasets demonstrated high imputation accuracy of xQTLImp and ability to discover novel signals which will enhance discoveries of xQTL studies.

2 Methods

The detailed methods can be found in supplementary notes. In brief, suppose $\boldsymbol{Z_u}|G$ and $\boldsymbol{Z_k}|G$ are two vectors with length n and m, representing unknown/missing and known xQTL statistics (i.e. Z-scores) of n and m variants associated with molecular trait G respectively. Our purpose is to impute $\boldsymbol{Z_u}|G$ from $\boldsymbol{Z_k}|G$ given known LD matrix $\boldsymbol{\Sigma}$, representing the LD correlations among $n+m$ variants. We estimated $\boldsymbol{Z_u}|G = \boldsymbol{\Sigma_{uk}}\boldsymbol{\Sigma_{kk}^{-1}}\boldsymbol{Z_k}|G$, derived by modeling Z scores of $n+m$ variants associated with G with multivariate normal distribution (MVN). And we adopt an imputation quality score $r2pred \in [0,1)$ to measure the xQTL imputation accuracy.

3 Results

3.1 Imputation accuracy

To test the imputation accuracy of xQTLImp, we first randomly masked various percent of summary statistics of eQTL from GTEx study v7 (Brain Amygdala, $n = 88$) (Consortium *et al.*, 2017) , haQTL and mQTL from brain xQTL study (Brain frontal cortex, $n = 433$ and 468 respectively) (Ng *et al.*, 2017), with variants having MAF ≥ 0.01 on chromosome 1. Then, xQTLImp was applied to impute those concealed statistics based on 1000G phase3 reference panel (EUR population, $n = 503$). Results in Figure 1 showed xQTLImp achieved high imputation accuracy with Pearson's $r > 0.95$ in all masking percentages under the thresholds of $r2pred \geq 0.6$. And more than 90% masked associations can be imputed with $r2pred = 0.6$. Similar results were also observed in other chromosomes (Supp. Figure 1).

As xQTLImp does not require individual-level datasets, it has the potential to impute novel associations with MAF lower than the MAF lower bound in original xQTL studies. To test the imputation accuracy on lower MAF levels, we adopted two WGS-based eQTL datasets from GTEx study v7 (Brain Amygdala and Muscle Skeletal, $n = 88$ and 491 respectively). EQTL associations with MAF > 0.1 and MAF > 0.05 were used to impute associations with $0.05 \leq$ MAF ≤ 0.1 and $0.01 \leq$ MAF ≤ 0.05 in Brain Amygdala and Muscle Skeletal respectively using same reference panel as above. xQTLImp still achieved high accuracy with Pearson's r = 0.95, 0.92 respectively by setting a stringent r2pred cutoff as 0.9 (Supp. Figure 2).

We also compared xQTLImp with HMM-based genotype imputation method, always regarded as the golden standard, by performing eQTL analysis using individual-level genotypes and RNA-seq profiles from ROSMAP study (Mostafavi *et al.*, 2018) after stringent quality controls (Supp. Notes). We first performed cis-eQTL association analysis on genotyped variants, and then applied xQTLImp for summary statistics imputation. For comparison, we also performed HMM-based genotype imputation on Michigan Imputation Server (Das *et al.*, 2016), and then eQTL analysis was performed on variants with imputed genotypes. The result showed high consistency (Pearson'r = 0.996) between statistics from xQTLImp imputation and HMM-based imputation (Supp. Figure 3). It indicates that xQTLImp can accurately obtain xQTL statistics without individual-level genotype data.

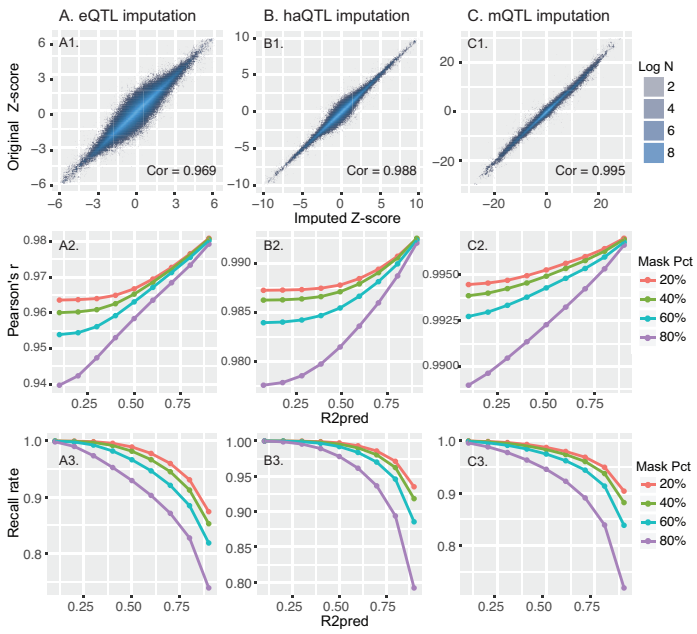

Fig. 1. Accuracy of xQTLImp on eQTL, haQTL and mQTL statistics imputation. – window_size was set to 500kb, 500kb and 100kb for eQTL, haQTL and mQTL respectively. For each type/column of QTL imputation, the first row (A1, B1 and C1) shows correlation between masked Z-statistics and imputed Z-statistics by setting mask percentage as 40% and r2pred cutoff as 0.6. The color legend shows density of associations in a local area. The second row (A2, B2 and C2) shows the Pearson's correlations under various masking percentages and r2pred cutoffs. The third row (A3, B3 and C3) shows the recall rates, defined as percentage of masked associations that can be recovered by imputation (with imputation quality better than certain r2pred thresholds).

3.2 Case study on single cell eQTL datasets

Recently, researchers start to pay attention to cell type specific xQTL based on single cell multiomics sequencing(van der Wijst *et al.*, 2018;Schmiedel *et al.*, 2018;Sarkar *et al.*, 2019). We applied xQTLImp on summary statistics from a single cell RNA sequencing based eQTL study with six major cell types (CD4+ T cells, CD8+ T cells, natural killer (NK) cells, monocytes, dendritic cells (DCs) and B cells) detected from peripheral blood mononuclear cells (PBMCs) from 45 donors (van der Wijst *et al.*, 2018). The MAF lower bound was 0.1 in original study due to limited sample size, and we used xQTLImp to impute novel eQTL associations with MAF ≥ 0.01 using 1000G reference panel. By applying a stringent r2pred cutoff as 0.9, and same permutation-based significance P-value cutoffs (FDR < 0.05) as original study, we detected 0.64∼1.48 times of novel significant eQTL associations compared with those in original results (Supp. Table 1). And considerable amount of novel significant cell-type specific eQTL associations were revealed (Supp. Figure 4). And no statistical difference was found in number of significant associations that can be replicated between original findings and our novel findings, in an independent CD4+ T cell-type specific eQTL dataset (Kasela *et al.*, 2017) (Supp. Figure 5).

Regarding the performance, xQTLImp took $3 \sim 4$ hours and <4GB memory for whole genome wide imputation on each cell-type in discovery study by using 20 threads and other settings in default (Supp. Table 2).

4 Conclusion

In conclusion, we present xQTLImp, an efficient and accurate tool specially designed for xQTL summary statistics imputation. We demonstrate in real dataset analyses that xQTLImp can not only impute missing xQTL statistics with high accuracy but also has the ability to help break through lower bound of MAF in original xQTL study, and discover novel significant xQTL signals, which will enhance xQTL discoveries.

Funding

This work has been supported by National Key Research and Development Program of China (Nos: 2017YFC0907503 and 2017YFC1201201).

References

Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, **81**(6), 1158–1168.

Consortium, G. *et al.* (2017). Genetic effects on gene expression across human tissues. *Nature*, **550**(7675), 204.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., *et al.* (2016). Next-generation genotype imputation service and methods. *Nature genetics*, **48**(10), 1284.

Furukawa, T. A., Barbui, C., Cipriani, A., Brambilla, P., and Watanabe, N. (2006). Imputing missing standard deviations in meta-analyses can provide accurate results. *Journal of clinical epidemiology*, **59**(1), 7–10.

Han, B., Kang, H. M., and Eskin, E. (2009). Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS genetics*, **5**(4), e1000456.

Julienne, H., Shi, H., Pasaniuc, B., and Aschard, H. (2018). Raiss: Robust and accurate imputation from summary statistics. *bioRxiv*, page 502880.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., *et al.* (2018). Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, **36**(1), 89.

Kasela, S., Kisand, K., Tserel, L., Kaleviste, E., Remm, A., Fischer, K., Esko, T., Westra, H.-J., Fairfax, B. P., Makino, S., *et al.* (2017). Pathogenic implications for autoimmune mechanisms derived by comparative eqtl analysis of cd4+ versus cd8+ t cells. *PLoS genetics*, **13**(3), e1006643.

Kwan, J. S., Li, M.-X., Deng, J.-E., and Sham, P. C. (2016). Fapi: Fast and accurate p-value imputation for genome-wide association study. *European Journal of Human Genetics*, **24**(5), 761.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, **11**(7), 499.

Mostafavi, S., Gaiteri, C., Sullivan, S. E., White, C. C., Tasaki, S., Xu, J., Taga, M., Klein, H.-U., Patrick, E., Komashko, V., *et al.* (2018). A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimerâŁ™s disease. *Nature neuroscience*, **21**(6), 811.

Ng, B., White, C. C., Klein, H.-U., Sieberts, S. K., McCabe, C., Patrick, E., Xu, J., Yu, L., Gaiteri, C., Bennett, D. A., *et al.* (2017). An xqtl map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature neuroscience*, **20**(10), 1418.

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L. (2014). Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, **30**(20), 2906–2914.

Sarkar, A. K., Tung, P.-Y., Blischak, J. D., Burnett, J. E., Li, Y. I., Stephens, M., and Gilad, Y. (2019). Discovery and characterization of variance qtls in human induced pluripotent stem cells. *PLoS genetics*, **15**(4), e1008045.

Schmiedel, B. J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A. G., White, B. M., Zapardiel-Gonzalo, J., Ha, B., Altay, G., Greenbaum, J. A., McVicker, G., *et al.* (2018). Impact of genetic polymorphisms on human immune cell gene expression. *Cell*, **175**(6), 1701–1715.

van der Wijst, M. G., Brugge, H., de Vries, D. H., Deelen, P., Swertz, M. A., and Franke, L. (2018). Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls. *Nature genetics*, **50**(4), 493.

Wen, X. and Stephens, M. (2010). Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, **4**(3), 1158.

Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E., and Halperin, E. (2010). Leveraging genetic variability across populations for the identification of causal variants. *The American Journal of Human Genetics*, **86**(1), 23–33.

xQTLImp: efficient and accurate xQTL summary statistics imputation

Supplementary Materials

Tao Wang[1,†], Quanwei Yin[2,3,†], Yongzhuang Liu[1], Jin Chen[4], Yadong Wang[1,*], and Jiajie Peng[2,3,*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China
[3]Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an, 710072, China
[4]Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40536, USA

CONTENTS

1 Statistical model 3
2 Implementation features 3
3 eQTL analysis of ROSMAP datasets 4
3.1 Genotype data process 4
3.2 Gene expression data process 4
3.3 Cis-eQTL association analysis 5
4 Datasets availability 5
5 Acknowledgement 5

LIST OF FIGURES

Figure S1 Imputation accuracy on different chromosomes. 6
Figure S2 Imputation accuracy on lower MAF levels. 7
Figure S3 Imputation accuracy compared with HMM-based genotype imputation method. 8
Figure S4 Upset diagram for significant eQTL associations in six cell types. 9
Figure S5 Diagram of two-way contingency table for Fisher's exact test. 10

LIST OF TABLES

Table S1 Summary of novel significant eQTL associations found by xQTLImp compared with that in original single cell eQTL datasets of six cell-types. 11

* *Corresponding authors: ydwang@hit.edu.cn or jiajiepeng@nwpu.edu.cn.*
† *Equal contributor.*

Table S2 Running time and memory cost on single cell eQTL datasets of six cell-types. 12

1 STATISTICAL MODEL

Similar as the statistical model described in GWAS imputation works [1, 2, 3, 4], we model xQTL summary statistics, given by Z scores, of variants in LD associated with a molecular trait G, by using multivariate normal distribution (MVN). Suppose $n+m$ variants are in association with molecular trait G, and Z scores are represented as $\mathbf{Z}|G = (x_1, ..., x_n, y_1, ..., y_m)^T$, where $\mathbf{Z_u}|G = (x_1, ..., x_n)^T$ represents unknown xQTL statistics of n variants associated with G, which needs to be imputed, and $\mathbf{Z_k}|G = (y_1, ..., y_m)^T$ represents known xQTL statistics of m variants associated with G. We model $\mathbf{Z}|G$ by MVN, i.e. $\mathbf{Z}|G \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ represents the vector of means with length $n+m$, and $\boldsymbol{\Sigma}$ represents the covariance matrix of $n+m$ variants with $\boldsymbol{\Sigma}_{i,j}$ represents LD correlation r between SNP_i and SNP_j, which can be calculated from reference panel such as 1000G and HapMap. By denoting expections $E(\mathbf{Z_u}|G) = \boldsymbol{\mu_u}$, and $E(\mathbf{Z_k}|G) = \boldsymbol{\mu_k}$, we have $\boldsymbol{\mu} = (\boldsymbol{\mu_u}, \boldsymbol{\mu_k})^T$, and similarly, $\boldsymbol{\Sigma}$ can be partitioned into four sub-matrices based on two groups of variants, i.e. $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_{uu}} & \boldsymbol{\Sigma_{uk}} \\ \boldsymbol{\Sigma_{ku}} & \boldsymbol{\Sigma_{kk}} \end{bmatrix}$, where $\boldsymbol{\Sigma_{ku}} = \boldsymbol{\Sigma_{uk}^T}$.

Our purpose is to impute $\mathbf{Z_u}|G$ from $\mathbf{Z_k}|G$ given known LD matrix $\boldsymbol{\Sigma}$. This problem can be solved by the conditional distribution of $\mathbf{Z_u}|G$ given $\mathbf{Z_k}|G$, represented as $(\mathbf{Z_u}|\mathbf{Z_k})|G$. It has been proven that the conditional distribution for $\mathbf{Z_u}|G$ given $\mathbf{Z_k}|G$ is n dimensional normal with expectation and covariance matrix as follows:

$$\begin{aligned} E[(\mathbf{Z_u}|\mathbf{Z_k})|G] &= \boldsymbol{\mu}_{\mathbf{Z_u}|\mathbf{Z_k}} = \boldsymbol{\mu_u} + \boldsymbol{\Sigma_{uk}}\boldsymbol{\Sigma_{kk}^{-1}}(\mathbf{Z_k}|G - \boldsymbol{\mu_k}) \\ &= \boldsymbol{\Sigma_{uk}}\boldsymbol{\Sigma_{kk}^{-1}}\mathbf{Z_k}|G \end{aligned} \tag{1}$$

$$\begin{aligned} Cov[(\mathbf{Z_u}|\mathbf{Z_k})|G] &= \boldsymbol{\Sigma_{u|k}} = \boldsymbol{\Sigma_{uu}} - \boldsymbol{\Sigma_{uk}}\boldsymbol{\Sigma_{kk}^{-1}}\boldsymbol{\Sigma_{ku}} \\ &= \boldsymbol{\Sigma_{uu}} - \boldsymbol{\Sigma_{uk}}\boldsymbol{\Sigma_{kk}^{-1}}\boldsymbol{\Sigma_{uk}^T} \end{aligned} \tag{2}$$

We then estimate unknown statistics $\mathbf{Z_u}|G = \boldsymbol{\Sigma_{uk}}\boldsymbol{\Sigma_{kk}^{-1}}\mathbf{Z_k}|G$, which can be interpreted that the unknown statistics associated with G is a weighted linear combination of known statistics of variants associated with G, with the weights reflecting LD relationships among variants. We adopted the metric $r2pred = 1 - \boldsymbol{\Sigma_{u|k}}$ to measure the xQTL imputation accuracy, which represents variances of unknown xQTL statistics $\mathbf{Z_u}|G$ explained by observed statistics $\mathbf{Z_k}|G$.

2 IMPLEMENTATION FEATURES

XQTLImp is specially designed to impute QTL statistics in association with thousands or even millions of molecular features of different kinds of molecular types such as gene expression (eQTL), DNA methylation (mQTL) and histone acetylation (haQTL), which differs from GWAS imputation where only one phenotype is considered. The main challenge of taking into account of tremendous molecular features is the computational cost in time and memory due to huge amount of associations that need to be considered during imputation process. For example, the summary statistics of GTEx cis-eQTL study [5] of Brain Amygdala contain 170 million eQTL associations in genome-wide, with 23.6 thousand unique genes and 10 million unique variants. Thus, during implementation, we adopt several strategies to accelerate the imputation process. First, based on the observation that multiple features, such as genes, might locate in same LD, xQTLImp preserves a dynamic local LD matrix in memory to avoid redundant LD calculations. Second, xQTLImp allows user to execute imputation process in parallel with *–num_threads N* parameter, during which associations on each chromosome will be broken into N chunks. And user can also specify chromosome number for chromosome-wide imputation

using *–chr* parameter, which is practically useful in taking advantage of high performance cluster. Third, xQTLImp also allow user to specify genome regions to exclude using *–exclude or –exclude_file*, such as complex HLA region, and to control MAF lower boundary of variants in reference panel to achieve better efficiency. Besides, xQTLImp is also memory efficient by adaptively reading xQTL records and genotypes in reference panel into memory and only saving local LD matrix. Performance of xQTLImp on real datasets is shown in Table S2.

3 EQTL ANALYSIS OF ROSMAP DATASETS

Religious Orders Study(ROS) and Memory and Aging Project(MAP) are two longitudinal cohort studies of aging that include de-identified clinical, neuropathological, brain RNA-seq and SNP array data for AD cases and controls [6]. In ROSMAP study, mRNA was extracted from the homogenate of the dorsolateral prefrontal cortex, and sequenced on the Illumina HiSeq platform. We downloaded read count summary table, array based genotyping data and clinical tables from Synapse platform (`https://www.synapse.org/#!Synapse:syn3219045`) with approval. The summary of gene expression levels covered 640 individuals and nearly 5,6000 genes. We used genotype data measured by Affymetrix GeneChip 6.0 at Broad Institute which covered 1,708 individuals and 636,236 SNPs.

3.1 Genotype data process

We applied PLINK2 (v1.9beta) [7] and in house scripts to perform rigorous subject and SNP quality control (QC) for genotype dataset that included: (1) remove subjects with call rate < 95%, (2) remove subjects with gender misidentification, (3) remove SNPs with genotype call rate < 95% (4), remove SNPs with Hardy-Weinberg Equilibrium testing P value < 1E-6, (5) remove SNPs with Test-mishap P value < 1E-9, (6) remove SNPs with minor allele frequency (MAF) < 0.05, (7) remove inbreeding subjects based on heterozygosity F score, and (8) IBS/IBD filtering (9) remove population outliers using smartPCA [8]. After QC, 572,307 variants and 1,664 individuals were left for downstream analysis.
Genotype imputation process was performed on Michigan Imputation Server (`https://imputationserver.sph.umich.edu/index.html#!`) , where Minimac3 was used for phasing and imputation [9], and 1000G phase3 was selected as reference panel. Prior imputation, we used imputation preparation checking tool (`http://www.well.ox.ac.uk/~wrayner/tools/`) to perform external quality controls to fit the requirements of imputation server. Variants with MAF $\geqslant 0.05$ and imputation quality score $R^2 \geqslant 0.3$ were used for downstram eQTL analysis.

3.2 Gene expression data process

We applied stringent quality control (QC) procedures for ROSMAP RNA-seq dataset. Read count table was first transformed into TPM (Transcripts Per Million) table. Then, we used Relative Log Expression (RLE) analysis, spearman correlation based hierarchical clustering, D-statistics analysis to remove sample outliers with problematic gene expression profile [10]. Sample mix-ups were excluded by comparing the reported sex with the sex determined by the expression of female-specific XIST gene and male-specific Y-chromosome gene. TPM values were log transformed after adding a pseudocount and quantile normalization was applied after excluding genes with low expression level. SVA [11] was applied to adjust covariates of batch, age, gender, RIN, PMI and recognition levels (cogdx score).

3.3 Cis-eQTL association analysis

After QC, 26,662 genes and 565,765 variants prior genotype imputation and 6,330,993 post imputation of 292 individuals with both genotypes and RNA-seq data, were used for cis-eQTL association analyses. MatrixEQTL [12] with the additive linear model were used for eQTL association analyses. And SNPs within 1Mb with the TSS of a gene were included.

4 DATASETS AVAILABILITY

To reproduce experiments in this study, we listed datasets and links used for analyses as follows:

- eQTL summary statistics of GTEx study [5]: GTEx protal (`https://gtexportal.org/home/index.html`).
- eQTL, mQTL, haQTL summary statistics of brain xQTL study [13]: Brain xQTL server (`http://mostafavilab.stat.ubc.ca/xqtl/`).
- Genotype and RNA-seq data of ROSMAP study [6] (in control use): Synapse platform (`https://www.synapse.org/#!Synapse:syn3219045`).
- Single cell eQTL summary statistics: (`https://genenetwork.nl/scrna-seq/`)
- CD4+ T cell eQTL summary statistics for replication study [14]: (`https://genenetwork.nl/cd4cd8eqtlbrowser/`)

5 ACKNOWLEDGEMENT

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The summary statistics used for the analyses described in this manuscript were obtained from the GTEx Portal V7.

The results published here arein part based on data obtained from the AMP-AD Knowledge Portal (doi:10.7303/syn2580853). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161, R01AG15819, R01AG17917, R01AG30146, R01AG36836, U01AG32984, U01AG46152, the Illinois Department of Public Health, and the Translational Genomics Research Institute.

SUPPLEMENTARY FIGURES

Figure S1

Multiple percent (10% ~ 60%) of eQTL associations from GTEx Brain Amygdala chromosome 1 ~ 22 were randomly masked pior imputation and Pearson's correlation coefficients between concealed Z-statistics and imputed Z-statistics on those hided variant-gene pairs were calculated for measurement under different r2pred thresholds. The recall rate is defined as percentage of masked associations that with imputation quality score (r2pred) larger than certain thresholds. Each boxplot represents the range of Pearson's r or recall rate among 22 chromosomes under center r2pred threshold. 1000G EUR population was selected as reference panel with MAF $\geqslant$ 0.01 during imputation, and HLA region (chr6:25M-35M) on chromosome 6 was excluded.

To be noted, to measure the effect of the random masking process, we repeated the random masking process on chromosome 1 for five times, and the results of Pearson's r and recall rate show very small deviations (SDs range from 10^{-5} to 10^{-4}).

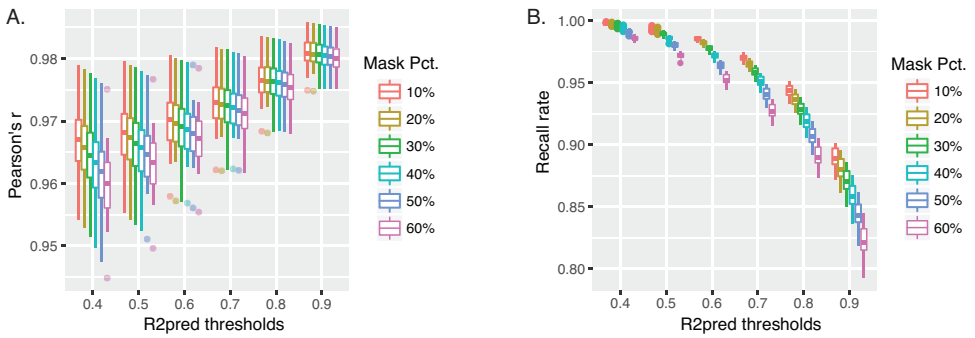

Figure S1: Imputation accuracy on different chromosomes.

Figure S2

To test the imputation accuracy on lower MAF levels, two eQTL datasets, whose genotypes were called from WGS, Brain Amygdala (N = 88) and Muscle Skeletal (N = 491), from GTEx study v7 were used for assessment. In Brain Amygdala dataset, eQTL associations with variants having MAF > 0.1 were used to impute associations with 0.05 ⩽ MAF ⩽ 0.1. Pearson's correlation and recall rate were calculated in same way as above, and the results were shown in Figure S2 panel A. The color in A2/B2 represented density of associations in a tiny local area. And to test the performance on MAF > 0.01, we used the Muscle Skeletal dataset with larger sample size. EQTL associations with variants having MAF > 0.05 were used to impute associations with 0.01 ⩽ MAF ⩽ 0.05. Results were shown in Figure S2 panel B. We can see that xQTLImp still achieved high accuracy with Pearson's r = 0.95, 0.92 respectively by setting a stringent r2pred cutoff as 0.9.

To be noted, eQTL associations on chromosome 1 were used in this experiment, and 1000G EUR population was selected as reference panel with MAF ⩾ 0.01 during imputation. The lower boundaries of MAF in original GTEx study for Brain Amygdala and Muscle Skeletal were set to 0.056 and 0.01 respectively.

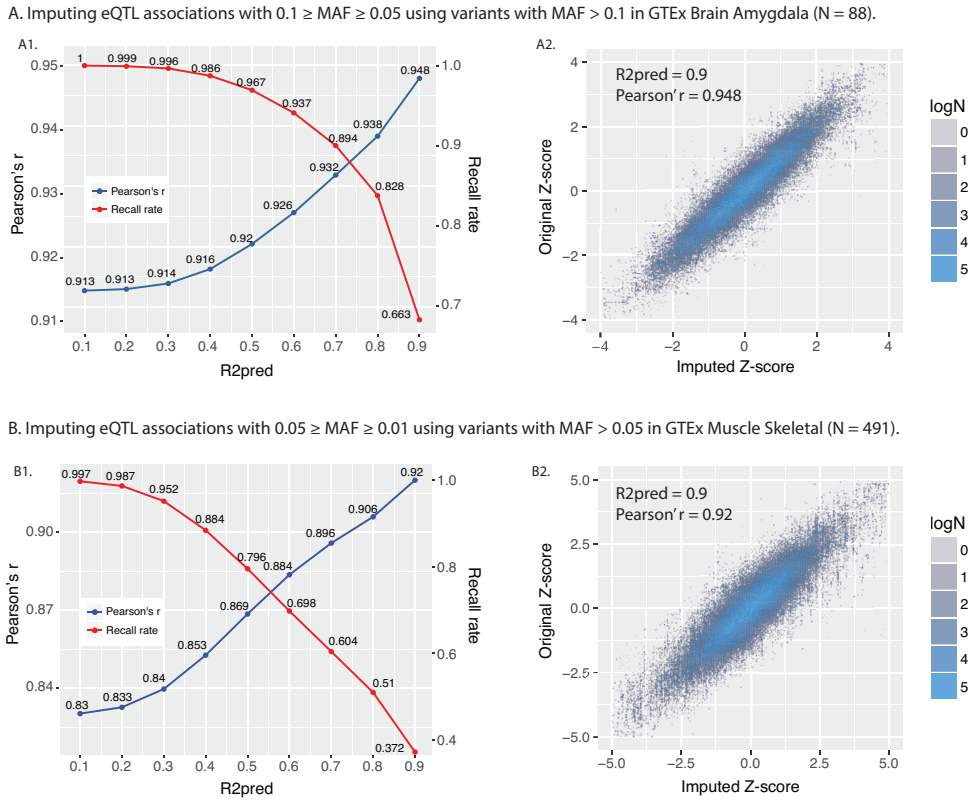

Figure S2: Imputation accuracy on lower MAF levels.

Figure S3

To compare with HMM-based genotype imputation method, which is the golden standard when individual-level genotype data is available, we performed eQTL analyses in two ways. First, eQTL analysis was performed on genotyped SNPs only, and xQTLImp was adopted to impute the rest of associations against the 1000G phase3 referece panel (EUR population). Second, genotyped SNPs were used for imputing the genotypes of rest variants also on 1000G phase3 reference panel by Minimac3 on Michigan Imputation Server. Then, eQTL analysis were performed on those imputed variants. The Z scores of those imputed associations were shown in Figure S3, where x-axis represented Z scores obtained from the first way, and y-axis represented Z scores obtained from the second way. Color represented the density of associations in a tiny local area. The result showed high consistency (Pearson'r = 0.996) between statistics from xQTLImp imputation and HMM-based imputation.

To be noted, individual-level genotype data and RNA-seq data were from ROSMAP study downloaded from Synapse platform. Analysing details can be found in section **"EQTL analysis of ROSMAP dataset"**. Chromosome 1 ~ 22 were included for analyses excluding the HLA region (chr6:25M-35M) on chromosome 6 in this experiment. MAF was set to 0.05, and other settings were in default.

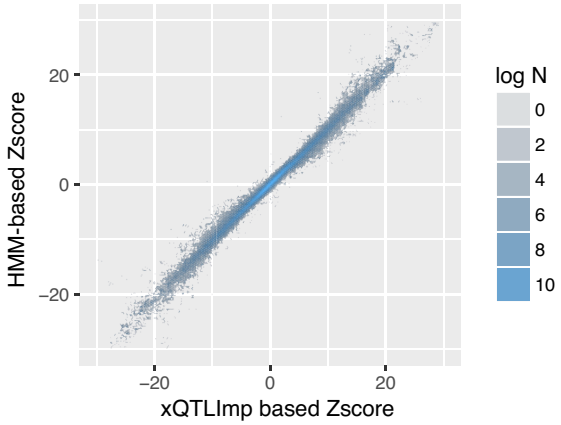

Figure S3: Imputation accuracy compared with HMM-based genotype imputation method.

Figure S4

Upset plot was used for comparing significant eQTL associations among six major cell types in PBMC, which was equivalent to Venn diagram. Upset diagram in panel A showed original significant eQTL associations without eQTL imputation. Upset diagram in panel B showed all significant eQTL associations after eQTL imputation process. Upset plots were made by UpsetR [15].

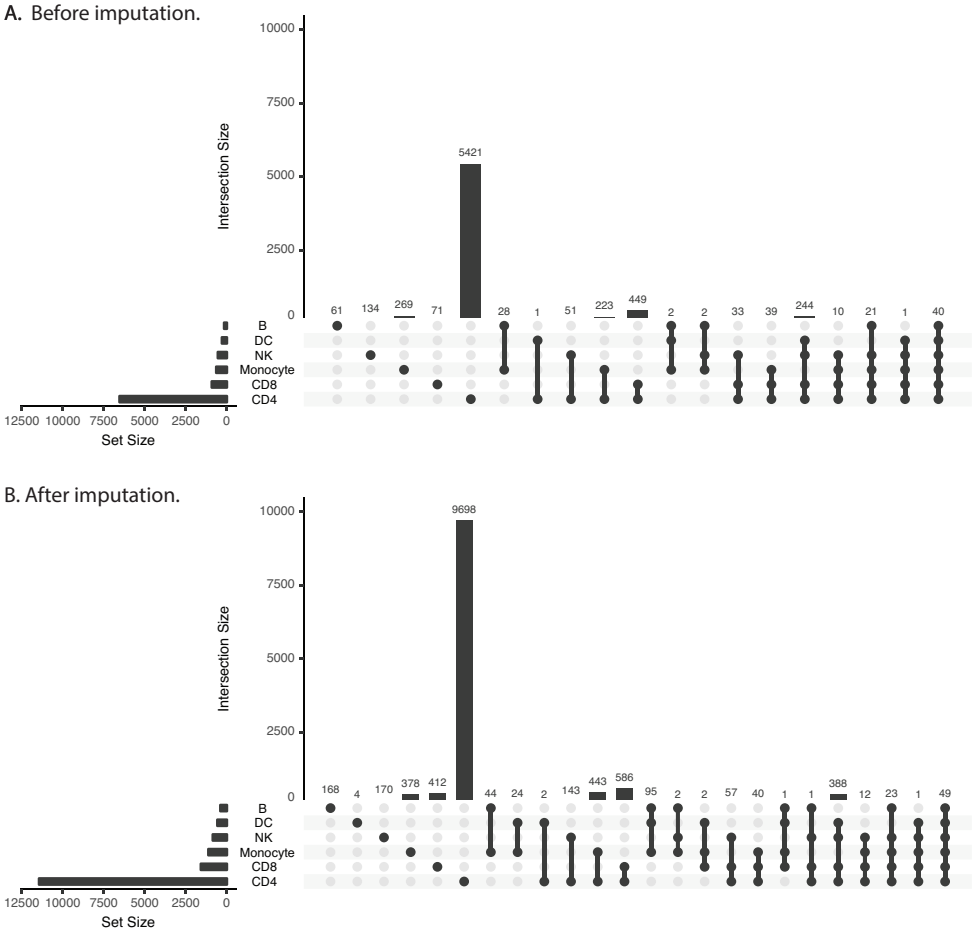

Figure S4: Upset diagram for significant eQTL associations in six cell types.

Figure S5

Novel significant eQTL associations in CD4+ cell-type (discovery study) were replicated using an independent CD4+ T cell-type specific eQTL dataset (N = 313, replication study) (kasela2017pathogenic). The lower boundary of MAF of variants in replication study was 0.05, and associations with FDR < 0.5 were available to download. By limiting comparisons on common genes and variants between discovery and replication studies, and setting FDR < 0.1 as significance level in replication study, we found 85% of novel significant associations in discovery study can be replicated in the replication study, compared with 85.6% of original significant associations that can be replicated. No statistical difference (fisher's exact test P value = 0.82) was found in number of significant associations that can be replicated between original findings and our novel findings.

	Replicated	Not Replicated	Replication percentage
Original significant eQTL-gene pairs	2,483	419	85.6%
Novel significant eQTL-gene pairs	369	65	85.0%

Fisher's exact test
P value = 0.82

Figure S5: Diagram of two-way contingency table for Fisher's exact test.

SUPPLEMENTARY TABLES

Table S1

Meanings of columns:

- Original sig. pairs: Number of significant eQTL-gene pairs in original eQTL summary dataset.
- Novel sig. pairs: Number of significant eQTL-gene pairs found by xQTLImp, and not shown in original summary dataset.
- Original sig. eQTLs: Number of unique significant eQTLs (SNPs) in original eQTL summary dataset.
- Noval sig. eQTLs: Number of unique significant eQTLS (SNPS) found by xQTLImp, and not shown in original summary dataset.
- P cutoff (FDR $\leqslant$ 0.05): Significance P value cutoff corresponding with permutation based FDR $\leqslant$ 0.05.

Table S1: Summary of novel significant eQTL associations found by xQTLImp compared with that in original single cell eQTL datasets of six cell-types.

Cell type	Original sig. pairs	Novel sig. pairs	Original sig. eQTLs	Novel sig. eQTLs	P cutoff (FDR$\leqslant$ 0.05)
CD4	6489	4910	6320	4870	7.93E-06
CD8	914	664	914	664	1.22E-06
NK	484	310	484	310	4.32E-07
Monocyte	383	479	372	473	6.97E-07
DC	155	229	155	229	5.08E-07
B	290	278	290	278	1.87E-07

Table S2

In this experiment, around 5 million associations across chromosme 1 ~ 22 were used as input for each cell type, and 31.4 million associations on average were outputed for each cell type. 1000G phase3 reference panel of EUR population was used for LD calcualtion, with following additional settings: MAF was set to 0.01, window size was set to 250Kb upstream and downstream from center of a gene, HLA region (chr6:25M-35M) was excluded.

Table S2: Running time and memory cost on single cell eQTL datasets of six cell-types.

Cell type	Threads	Time(H:M:S)	Memory
CD4	20	3:55:29	3.58G
CD8	20	3:15:52	3.54G
NK	20	3:57:20	3.56G
Monocyte	20	3:46:25	3.59G
DC	20	3:51:21	3.66G
B	20	3:45:54	3.65G

REFERENCES

[1] Donghyung Lee, T Bernard Bigdeli, Brien P Riley, Ayman H Fanous, and Silviu-Alin Bacanu. Dist: direct imputation of summary statistics for unmeasured snps. *Bioinformatics*, 29(22):2925–2927, 2013.

[2] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P Strachan, Nick Patterson, and Alkes L Price. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 2014.

[3] Johnny SH Kwan, Miao-Xin Li, Jia-En Deng, and Pak C Sham. Fapi: Fast and accurate p-value imputation for genome-wide association study. *European Journal of Human Genetics*, 24(5):761, 2016.

[4] Hanna Julienne, Huwenbo Shi, Bogdan Pasaniuc, and Hugues Aschard. Raiss: Robust and accurate imputation from summary statistics. *bioRxiv*, page 502880, 2018.

[5] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[6] Sara Mostafavi, Chris Gaiteri, Sarah E Sullivan, Charles C White, Shinya Tasaki, Jishu Xu, Mariko Taga, Hans-Ulrich Klein, Ellis Patrick, Vitalina Komashko, et al. A molecular network of the aging human brain provides insights into the pathology and cognitive decline of alzheimer's disease. *Nature neuroscience*, 21(6):811, 2018.

[7] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):7, 2015.

[8] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.

[9] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, et al. Next-generation genotype imputation service and methods. *Nature genetics*, 48(10):1284, 2016.

[10] Peter AC't Hoen, Marc R Friedländer, Jonas Almlöf, Michael Sammeth, Irina Pulyakhina, Seyed Yahya Anvar, Jeroen FJ Laros, Henk PJ Buermans, Olof Karlberg, Mathias Brännvall, et al. Reproducibility of high-throughput mrna and small rna sequencing across laboratories. *Nature biotechnology*, 31(11):1015, 2013.

[11] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[12] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[13] Bernard Ng, Charles C White, Hans-Ulrich Klein, Solveig K Sieberts, Cristin McCabe, Ellis Patrick, Jishu Xu, Lei Yu, Chris Gaiteri, David A Bennett, et al. An xqtl map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nature neuroscience*, 20(10):1418, 2017.

[14] Silva Kasela, Kai Kisand, Liina Tserel, Epp Kaleviste, Anu Remm, Krista Fischer, Tõnu Esko, Harm-Jan Westra, Benjamin P Fairfax, Seiko Makino, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eqtl analysis of cd4+ versus cd8+ t cells. *PLoS genetics*, 13(3):e1006643, 2017.

[15] Jake R Conway, Alexander Lex, and Nils Gehlenborg. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 2017.