*GENETICS*, 2022, 221(4), iyac088

https://doi.org/10.1093/genetics/iyac088
Advance Access Publication Date: 11 June 2022

# TwinEQTL: ultrafast and powerful association analysis for eQTL and GWAS in twin studies

Kai Xia [1,2,†] Andrey A. Shabalin [3,†] Zhaoyu Yin,[4] Wonil Chung [5] Patrick F. Sullivan [2] Fred A. Wright,[6] Martin Styner [2] John H. Gilmore [2] Rebecca C. Santelli [7] Fei Zou[1,*]

[1]Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA,
[2]Department of Psychiatry, University of North Carolina, Chapel Hill, NC 27599, USA,
[3]Department of Psychiatry, University of Utah, Salt Lake City, UT 84108, USA,
[4]Gilead Sciences, Foster City, CA 94404, USA,
[5]School of Public Health, Harvard, Boston, MA 02115, USA,
[6]Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA,
[7]Department of Pediatrics and Human Development, Michigan State University, East Lansing, MI 48912, USA

*Corresponding author: Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599, USA. E-mail: fzou@bios.unc.edu
[†]These authors contributed equally to this work.

## Abstract

We develop a computationally efficient alternative, TwinEQTL, to a linear mixed-effects model for twin genome-wide association study data. Instead of analyzing all twin samples together with linear mixed-effects model, TwinEQTL first splits twin samples into 2 independent groups on which multiple linear regression analysis can be validly performed separately, followed by an appropriate meta-analysis-like approach to combine the 2 nonindependent test results. Through mathematical derivations, we prove the validity of TwinEQTL algorithm and show that the correlation between 2 dependent test statistics at each single-nucleotide polymorphism is independent of its minor allele frequency. Thus, the correlation is constant across all single-nucleotide polymorphisms. Through simulations, we show empirically that TwinEQTL has well controlled type I error with negligible power loss compared with the gold-standard linear mixed-effects models. To accommodate expression quantitative loci analysis with twin subjects, we further implement TwinEQTL into an R package with much improved computational efficiency. Our approaches provide a significant leap in terms of computing speed for genome-wide association study and expression quantitative loci analysis with twin samples.

**Keywords:** Twin; eQTL; GWAS

## Introduction

For complex psychiatric disorders, such as schizophrenia and major depressive disorder, twin studies have received attention for establishing the general extent to which genes and environment are etiologically important (Neale and Cardon 1992; Boomsma *et al.* 2002; Silventoinen *et al.* 2003; Vaccarino *et al.* 2008; Chou *et al.* 2009; Park *et al.* 2012). Typical twin data include both monozygotic twins (MZ) and dizygotic twins (DZ), plus unpaired individual twins (singletons). Unlike data with independent samples, twin data require more careful statistical modeling since ignoring genetic relatedness and shared environment among twin pairs may lead to high false and/or low true positive findings. Several statistical approaches are available for twin data. One of the most common approaches is the linear mixed-effects model (LMM) where random effects are used to properly account for the correlations among subjects (Carlin *et al.* 2005; Wang *et al.* 2011; Kuna *et al.* 2012). Mixed-effects model have a well-established theory which is familiar to statisticians. Moreover, it is conveniently implemented in most statistical software and can flexibly adjust other nongenetic and genetic covariates (Ghazalpour *et al.* 2008; Rabe-Hesketh *et al.* 2008). Although single GWAS analysis using LMM is feasible by high performance computing (HPC), GWAS analysis of multiple traits such as expression quantitative loci (eQTL) studies where associations between thousands of transcripts and millions of single nucleotide polymorphisms (SNPs) are tested, mixed-effects models are extremely computationally inefficient if not practically impossible. The BOLT-LMM algorithm (Loh *et al.* 2015) rapidly computes statistics for association between phenotype and genotypes using an LMM by assuming a Bayesian mixture-of-normals prior for the random effect attributed to SNPs other than the one being tested, which provides an opportunity for increased power to detect associations while controlling false positives. However, it is only recommended to be used for human genetic datasets containing more than 5,000 samples. MatrixEQTL is an ultrafast R package for genome-wide association study with genetically unrelated

subjects (Shabalin 2012) and it is not readily applicable to twin eQTL data. Previously, we have developed fast eQTL analysis method which use a score statistic that automatically adjusts the (hidden) correlation between the 2 correlated groups (Yin *et al.* 2015). However, this method requires reanalyzing the original data using HPC clusters.

A meta-analysis procedure, which uses only the SNP level GWAS summary statistics, is a popular approach for combining analysis results from multiple studies to increase power for detecting association findings. Meta-analysis has been used to integrate several GWAS studies (Evangelou and Ioannidis 2013) from multiple international institutes, and to GWAS of schizophrenia (Ripke *et al.* 2013, 2014), type 2 diabetes (Zeggini *et al.* 2008), polygenic dyslipidemia (Kathiresan *et al.* 2009), height (Allen *et al.* 2010), etc. It has also been mathematically and numerically proved that there is no efficiency gain in performing mega-analysis where full data with all individual level data are analyzed vs meta-analysis where only summarized test statistics are combined (Lin and Zeng 2010a, 2010b), which further popularizes the application of meta-analysis in GWAS. However, traditional meta-analysis requires independent studies without overlapping subjects or family members. For studies with overlapping subjects, it has been shown that failure to take the overlapping into consideration can lead to inflated type I errors and proper meta-analysis procedures are needed (Lin and Sullivan 2009). More recently, a meta-analysis of correlated traits method was proposed by Zhu *et al.* (2015), whose approach integrates correlated subjects and correlated traits in the same meta-analysis framework using 2 alternative approaches, $S_{Hom}$ for homogeneous traits and $S_{Het}$ for heterogeneous traits. According to the homogeneous assumption, $S_{Hom}$ can be directly applied to studies with correlated subjects such as twins where the correlation of test statistics is estimated empirically by Pearson's correlation coefficient. However, the difference of correlation structures in MZ, DZ, and singleton are ignored and approximated, which could lead to power loss that cannot be afforded for large-scale twin GWAS and eQTL studies. Up to our best knowledge, no meta-analysis method has been specifically designed for elegantly modeling twin structures in eQTL and GWAS studies while reducing computational cost and keeping statistical power as high as using LMM. Cheung (2014, 2018) adapted a structure equation modeling approach to model complicated variance–covariance structure among correlated traits or subjects, which can also be applied to GWAS studies with twins. However, it is always a judgment call to check the assumption of homogeneity or heterogeneity of variance–covariance matrix.

In this study, we propose a computationally efficient and statistical powerful approach in twin GWAS/eQTL analysis, TwinEQTL, that boosts the computing efficiency from thousands of folds (Yin *et al.* 2015) to 10,000 folds compared with LMM. Similar to previous approach (Zhu *et al.* 2015), TwinEQTL algorithm was derived by adjusting variance–covariance structure in the meta-analysis model when combining correlated test statistics. In addition, by modeling the difference of correlation structures in MZ and DZ pairs, the correlation between test statistics can be more accurately estimated using only 1 run of LMM per phenotype, without iterating all the combinations of both phenotype and SNPs. Furthermore, TwinEQTL is implemented in a fashion similar to the currently one of the fastest engines for GWAS, MatrixEQTL, while taking the correlation structure of both MZ and DZ twins into consideration by mimicking meta-analysis procedure in GWAS. In TwinEQTL pipeline, we first split twin pairs and singletons into 2 independent sets, so that within each set the samples are unrelated and a linear regression is performed separately for each set. We then combine the 2 nonindependent test statistics through a meta-analysis where the correlation between the 2 sets of results is adjusted. Our method avoids the iterative use of LMM with substantially reduced computing time. When Pearson's correlation coefficient is used to empirically approximate the correlation between test statistics, TwinEQTL is equivalent to $S_{Hom}$ at the cost of power loss in many scenarios according to simulations. We performed a series of simulations to demonstrate that TwinEQTL is robust to various correlation structures in twin samples while maintaining superior statistical power compared with competing methods.

## Materials and methods

### TwinEQTL without covariates

We describe the method first assuming no covariates exist. Extension to situations where one or more covariates exist will be discussed later. Suppose for a given GWAS, there are $n_{MZ}$ pairs of MZ, $n_{DZ}$ pairs of DZ, and $n_{Sg}$ singletons. The total sample size $n = 2 * n_{MZ} + 2 * n_{DZ} + n_{Sg}$. First, we randomly split each twin pair into 2 groups, named group 1 and group 2. We then randomly divide the singletons into half and assign them to group 1 and group 2 separately. Samples in groups 1 and 2 are ordered in such a way that the first $n_{MZ}$ samples in groups 1 and 2 are the paired MZ samples, and the $n_{DZ}$ samples are the paired DZ samples, and the remaining samples are singletons. Now for samples within group $k$ ($k = 1, 2$), they are genetically unrelated, on which the simple linear model below can be performed between a given SNP and the trait:

$$y_{ik} = \mu_k + \beta g_{ik} + \epsilon_{ik}, \tag{1}$$

where $g_{ik}$ and $y_{ik}$ are the corresponding genotype and phenotype of subject $i$ ($i = 1, \ldots, n_k$) in group $k$, and the random error $\epsilon_{ik} \sim N(0, \sigma_k^2)$. For simplification, we assume that both $y$ and $g$ are standardized to have mean 0 and variance 1 within each group. Under null hypothesis ($H_0$), there is no association between a given SNP and responses ($H_0 : \beta = 0$) in each subset.

The maximal likelihood estimate (MLE) of $\widehat{\beta}$ from each dataset $k$ ($k = 1, 2$) therefore equals

$$\widehat{\beta}_k = \frac{\sum_{i=1}^{n_k}(g_{ik} - \overline{g_k})(y_{ik} - \overline{y_k})}{\sum_{i=1}^{n_k}(g_{ik} - \overline{g_k})^2} = \frac{\sum_{i=1}^{n_k} y_{ik} g_{ik}}{n_k - 1}. \tag{2}$$

The correlation between $\epsilon_{i1}$ and $\epsilon_{j2}$ is

$$Corr(\epsilon_{i_1}, \epsilon_{j_2}) = \begin{cases} \rho_{DZ}, & \text{subjects } i \text{ and } j \text{ are a DZ pair} \\ \rho_{MZ}, & \text{subjects } i \text{ and } j \text{ are a MZ pair} \\ 0, & \text{subjects } i \text{ and } j \text{ are unrelated,} \end{cases} \tag{3}$$

according to the common ACE model (Neale and Cardon 1992) for twin data. Here

$$\rho_{DZ} = \left(\frac{1}{2}\sigma_a^2 + \sigma_c^2\right) / \left(\sigma_a^2 + \sigma_c^2 + \sigma_e^2\right) \text{ and}$$

$$\rho_{MZ} = (\sigma_a^2 + \sigma_c^2) / (\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$$

with $\sigma_a^2$ being the additive genetic effect (A), $\sigma_c^2$ being the shared common environment effect (C), and $\sigma_e^2$ being the unique environment effect (E).

The derived metatest statistic (Appendix) can be expressed as

$$Z = \frac{T_1 + T_2}{\sqrt{1 + 1 + 2\mathrm{Corr}(T_1, T_2)}}, \quad (4)$$

where $\mathrm{Corr}(T_1, T_2) = (n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ})/n$, which is the analytic correlation between $T_1$ and $T_2$, where $T_1$ and $T_2$ are the corresponding $t$-tests from the group 1 and group 2 data, respectively. Under $H_0 : Z \sim N(0, 1)$.

In Appendix, we have proved that for a given SNP, the correlation between the 2 $t$ statistics, $T_1$ and $T_2$, only depends on $\rho_{MZ}$, $\rho_{DZ}$ and the numbers of DZ, MZ pairs, and singletons, and is independent of the minor allele frequency (MAF) of the SNP. Thus, the correlation is constant across all SNPs. To estimate the $\rho_{MZ}$ and $\rho_{DZ}$, we fit the following ACE model with all samples

$$\mathrm{var}(y) = \sigma_a^2 + \sigma_c^2 + \sigma_e^2 \quad (5)$$

and estimate $\sigma_a^2$, $\sigma_c^2$, and $\sigma_e^2$, respectively, based on which, we estimate $\mathrm{Corr}(T_1, T_2)$ analytically as $\widehat{\mathrm{Corr}}(T_1, T_2) = (n_{DZ}\widehat{\rho}_{DZ} + 2n_{MZ}\widehat{\rho}_{MZ})/n$ with $\widehat{\rho}_{DZ} = 0.5\widehat{\sigma}_a^2 + \widehat{\sigma}_c^2$ and $\widehat{\rho}_{MZ} = \widehat{\sigma}_a^2 + \widehat{\sigma}_c^2$.

Alternatively, Zhu *et al.* (2015) proposed to empirically estimate $\mathrm{Corr}(T_1, T_2)$ by the sample correlation of $T_1$ and $T_2$ across all tested SNPs. As the number of SNPs can be large for modern GWAS data, we may choose to calculate the sample correlation of $T_1$ and $T_2$ across a few thousands of randomly selected SNPs. The performance of both methods was evaluated through simulations.

**Table 1.** Type I errors of twin data using LM, LMM, and TwinEQTL.

| | | $\alpha = 0.01$ | | | | $\alpha = 0.001$ | | | | $\alpha = 0.0001$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $a^2$ | $c^2$ | LM | LMM | $S_{Hom}$ | TwinEQTL | LM | LMM | $S_{Hom}$ | TwinEQTL | LM | LMM | $S_{Hom}$ | TwinEQTL |
| 0 | 0 | 0.010 | 0.010 | 0.010 | 0.009 | 0.0009 | 0.0010 | 0.0010 | 0.0009 | 0.00008 | 0.00009 | 0.00008 | 0.00007 |
| 0.2 | 0.1 | 0.018 | 0.010 | 0.010 | 0.010 | 0.0024 | 0.0009 | 0.0010 | 0.0010 | 0.00029 | 0.00009 | 0.00009 | 0.00009 |
| 0.5 | 0.1 | 0.027 | 0.010 | 0.010 | 0.010 | 0.0045 | 0.0010 | 0.0010 | 0.0010 | 0.00076 | 0.00009 | 0.00010 | 0.00010 |
| 0.7 | 0.2 | 0.037 | 0.010 | 0.010 | 0.010 | 0.0077 | 0.0010 | 0.0010 | 0.0010 | 0.00161 | 0.00009 | 0.00010 | 0.00010 |
| 0.9 | 0 | 0.036 | 0.010 | 0.010 | 0.010 | 0.0073 | 0.0009 | 0.0010 | 0.0010 | 0.00149 | 0.00009 | 0.00010 | 0.00010 |

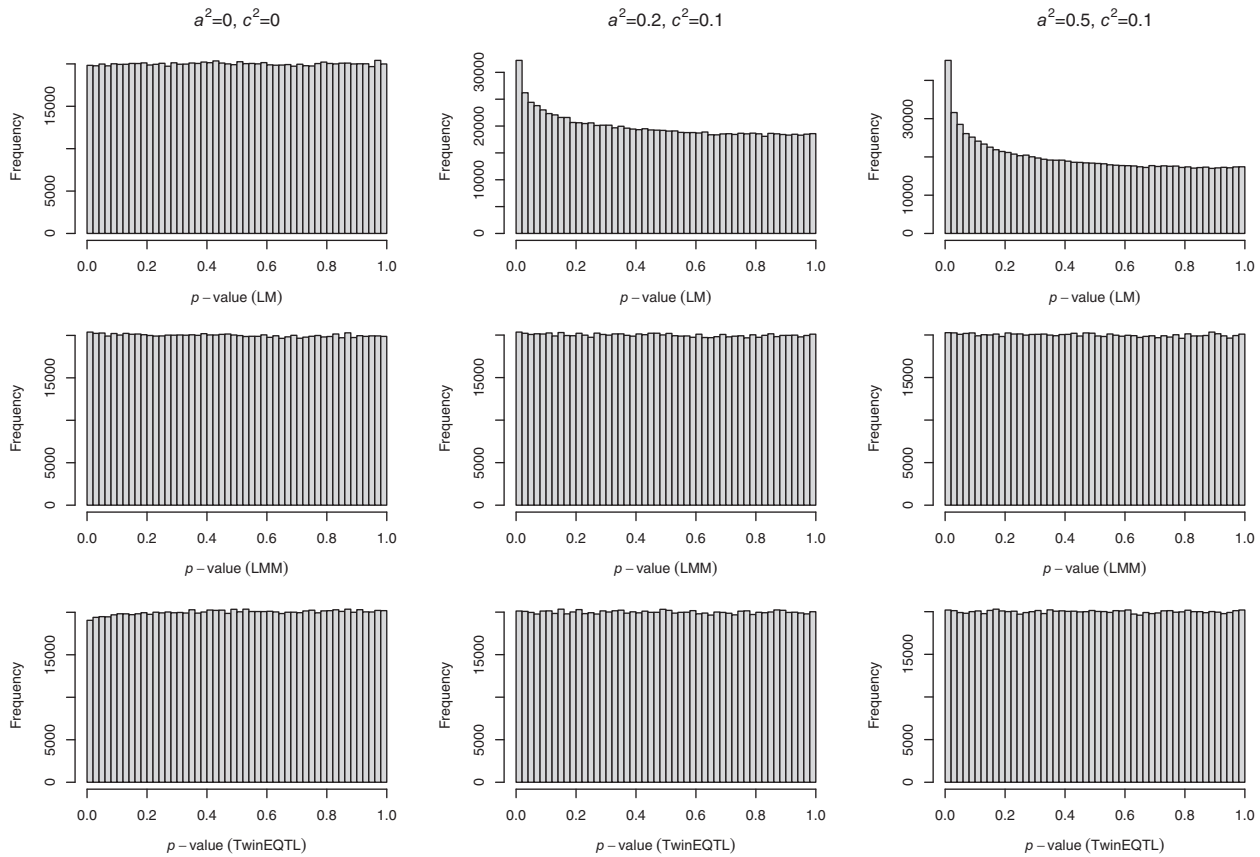


**Fig. 1.** Distributions of $P$-value under $H_0$ for different scenarios of $a^2$ and $c^2$. The comparisons of distributions of $P$-value under different scenarios of additive genetic effect ($a^2$) and shared environment effect ($c^2$) using linear model (top panel), LMM (middle panel) and TwinEQTL (bottom panel). For linear model method, both subjects in each twin pair and all the singletons were used ignoring correlation structure within each twin pair. The number of MZ and DZ are balance ($n_{MZ} = n_{DZ} = 500$) and the number of singletons are relatively small ($n_{Sg} = 100$) in this case.

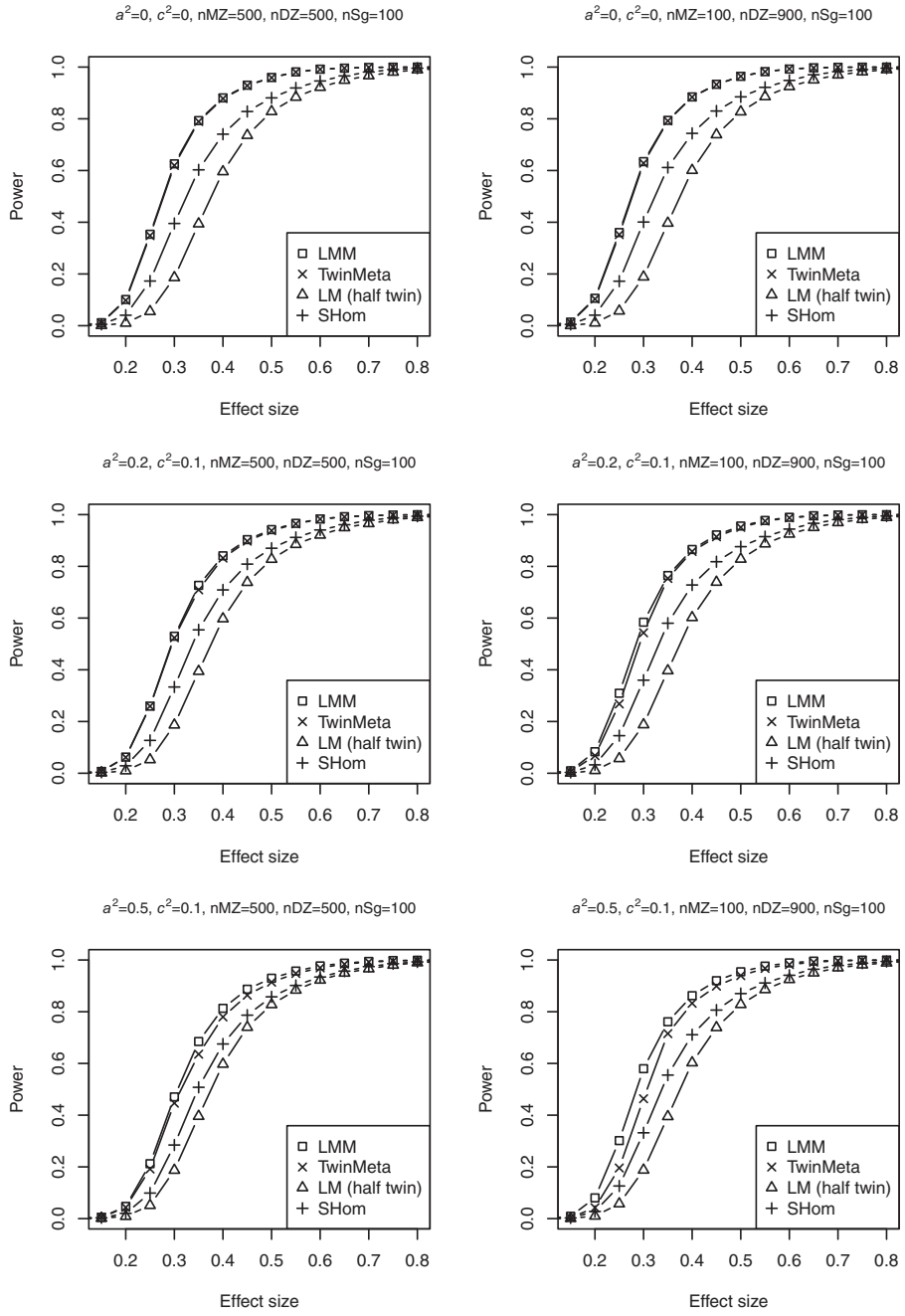


**Fig. 2.** Power estimations for twin data with different scenarios of $a^2$ and $c^2$. The power analysis at significance level of $\alpha = 5 \times 10^{-8}$ under different scenarios of sample size, effect size, additive genetic effect ($a^2$), and shared environmental effect ($c^2$). For linear model method, only one subject per twin pair and all the singletons were used in the regression model. The left panel shows the estimated power under circumstance that the number of MZ and DZ are balance ($n_{MZ} = n_{DZ} = 500$ and $n_{Sg} = 100$). The right panel shows the estimated power under circumstance that the number of MZ and DZ are imbalance ($n_{MZ} = 100$, $n_{DZ} = 900$, and $n_{Sg} = 100$)

## TwinEQTL with covariates

As discussed in Shabalin (2012), the multiple linear regression can be reduced to simple linear regression when testing for the SNP effect by regressing the response variable and the SNP genotypes relative to the other covariates, from which the residuals are obtained and used for the subsequent simple linear regression. Following the similar argument, we can show that the correlation between the 2 sets of $t$ statistics will be constant across all SNPs and is independent of the MAF of each SNP, when the covariates and the SNP genotypes are not correlated, which should be true for most of the SNPs, or approximately true for all SNPs.

## Linear mixed-effect model

LMM is considered as a gold standard for analyzing twin data and multivariate phenotypes. In this paper, we compare the performances of the proposed method with LMM. For twin data, we implemented LMM as described in Wright *et al.* (2014) which

distinguishes MZ and DZ twins by considering additive genetic effect (A), shared environment effect (C), and unique environment effect [aka the ACE model (Neale and Cardon 1992)].

## GWAS of brain volume in neonates

The GWAS study of early brain development study (EBDS) at UNC-Chapel Hill was to investigate how genetic variation impacts prenatal and early postnatal brain development of global brain tissue volumes in a unique cohort of infants who received high-resolution MRI scans of the brain around 5 weeks of age (Xia *et al.* 2017).

The SNP genotypes were generated from Affymetrix Axiome World Array 4.0 on 852 infants with their buccal samples. A total of 756 infants and 854,979 SNPs were obtained after the following quality control steps [see more details in Xia *et al.* (2017)]: we excluded samples with low DishQC (<0.82), low call rates (<95%), outliers for homozygosity, sex, or zygosity from genotypes inconsistent with reported phenotypes, ancestry outliers, excessive relatedness, and unexpected relatedness. We removed individual SNPs that deviated from Hardy–Weinberg equilibrium (HWE) ($P_{HWE} < 1 \times 10^{-8}$), had low call rate (<95%), high Mendelian error rate (>0.1, based on 5 parent–child trios), high deviation of allele frequency compared with European American and African American subsets from the 1000 Genomes Project. Imputation was performed with MACH-Admix using the 1000 Genomes Project (1000G) reference panel (phase1 release v3.20101123). To evaluate the quality of imputed SNPs, we computed mean $R^2$ for varying MAF categories and $R^2$ cutoffs. We retained SNPs with mean $R^2 > 0.8$ and excluded SNPs with MAF $< 0.01$. A total of 561 infants (300 male, 261 female) between 0 and 24 weeks of age, encompassing 295 singletons or unpaired twins, 17 sibling pairs and 232 twins [61 same-sex dizygotic (DZ) pairs, 37 same-sex monozygotic (MZ) pairs and 18 opposite-sex DZ pairs]. Overall, 63% of subjects are Europeans with remaining subjects being primarily Africans. The MRI measurements of brain volume include white matter (WM), intracranial volume (ICV), gray matter (GM), and total cerebrospinal fluid (CSF).

The genome-wide association analysis using linear regression is not valid due to the correlation among twins. Instead, the LMM was used in the original study by utilizing HPC cluster. The first 3 genotypic principal components (PCs) were included as covariates to control for population stratification, so was scanner type to control for potential scanner bias. ICV was included as a covariate for GM, WM, and CSF. Additional covariates were selected via adaptive LASSO from a comprehensive set of demographic and medical history variables (Xia *et al.* 2017) including birth weight, gestational age at birth, sex, and age at MRI.

## Netherlands Twin Registry eQTL Study

In Netherlands Twin Registry (NTR) twin study (dbGaP study accession number: phs000486.v1.p1), 2,752 individuals had their SNP genotypes and gene expression data measured (Wright *et al.* 2014) on Affymetrix Genome-Wide Human SNP Array 6.0 and Affymetrix u219 array, respectively. One of the goals of this project is to identify a comprehensive list of eQTL in peripheral blood and their biological significance. After a series of quality control steps described previously in Wright *et al.* (2014), a total of 642,489 autosomal SNPs and 47,495 transcripts on 2,561 individuals were kept for the eQTL analysis, which include 641 MZ pairs, 564 DZ pairs, and 151 singletons. A total of 16 covariates including the age at blood sampling, sex, smoking status, body mass index, hematocrit count, hemoglobin count and total white and red cell counts, 5 PCs from the gene expression data, and 3 PCs derived from the pruned genotype data are also included as covariates.

# Results

## Simulation studies

For simulation analysis, genotypes of MZ and DZ twins were simulated with the following manners: For a given SNP, its MAF was first sampled from a uniform distribution of $U(0.05, 0.5)$. Then for each twin pair, 2 maternal alleles and 2 paternal alleles were independently and randomly generated from the Bernoulli distribution with the sampled MAF. If the twin pair is an MZ pair, their identical genotype was generated by combining one randomly

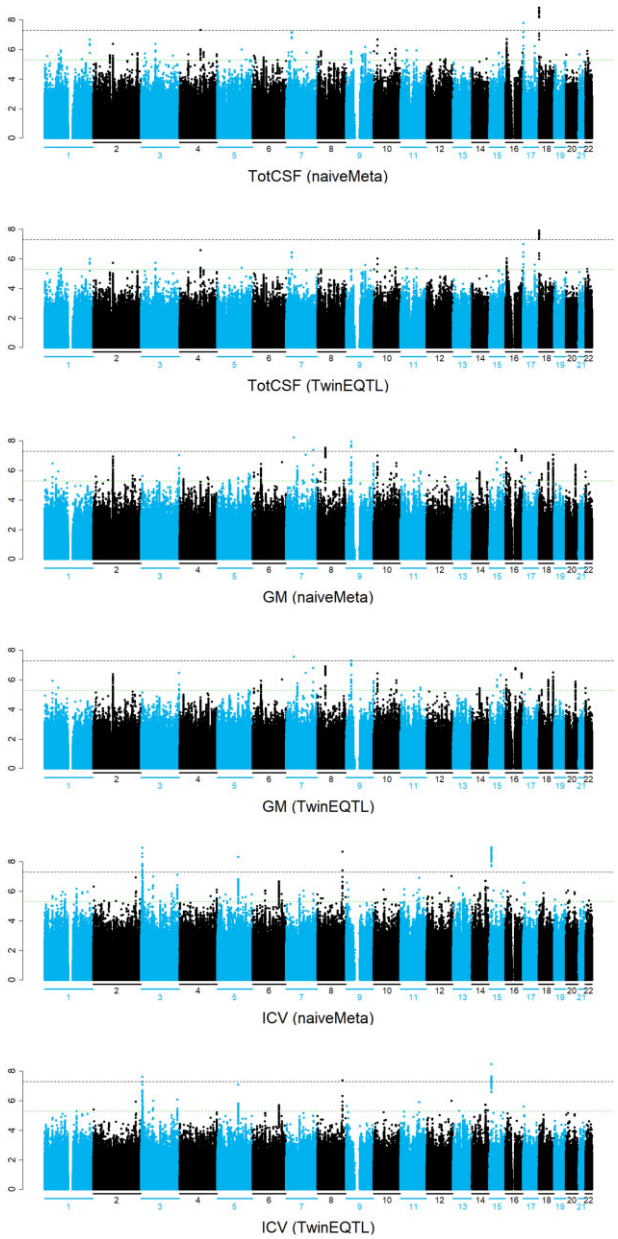


**Fig. 3.** Manhattan plot of GWAS study of neonatal CSF (top panels), GM (middle panels), and ICV (bottom panels) using either naive meta-analysis (not controlling for correlation structure between twins) or TwinEQTL (controlling for correlation structure between twins). The results using naive meta-analysis show overestimated significant findings due to inflated type I error.

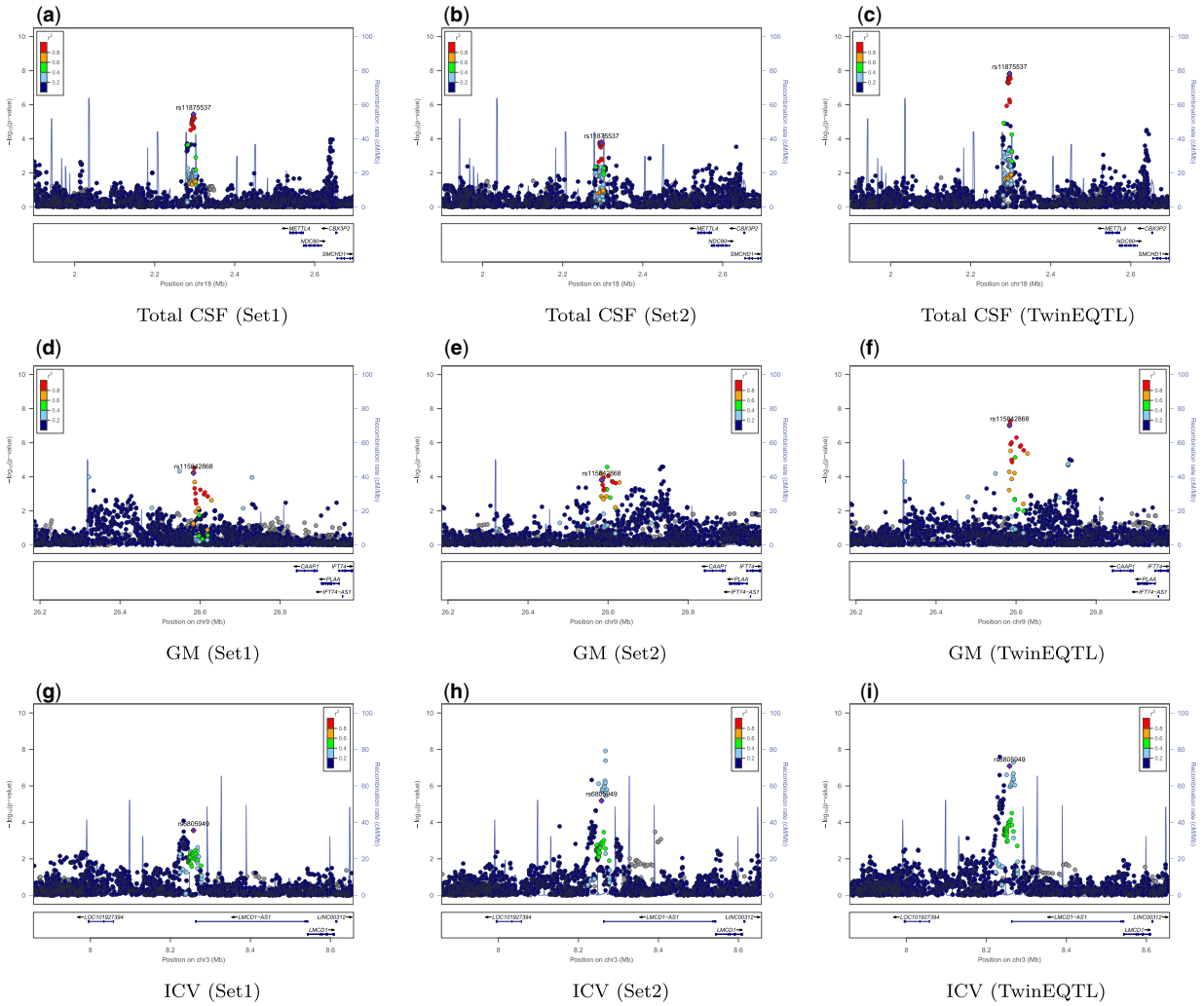


**Fig. 4.** LocusZoom plot of genome-wide association analysis of neonatal GM, CSF, and ICV. GWAS *P*-values of neonatal CSF ( top panels: a-c), GM (middle panels: d-f), and ICV (bottom panels: g-i) in EBDS study. The GWAS analysis performed separately using independent set 1 (left panels), independent set 2 (middle panels) and then combined set 1 and set 2 using TwinEQTL (right panels). The significance of *P*-values in set1 and set2 is augmented by TwinEQTL even though they were not significant in results using only set1 or set2.

sampled maternal allele and one randomly sampled paternal allele. If it is a DZ pair, each sample's genotype was separately created by merging one randomly selected maternal allele and one randomly selected paternal allele.

## Type I error simulations.

We simulated 500 MZ twins, 500 DZ twins and 100 or 1,000 singletons under different scenario of genetic ($a^2$) and shared environmental effect ($c^2$). Table 1 shows the empirical type I errors of $S_{Hom}$, TwinEQTL, LMM, and naive linear method (LM) without correcting for correlation among twin subjects at different significance levels $\alpha$ and 6 scenarios of different genetic and environmental effects. One million simulations were done in each scenario. The results suggest that when naive LM is directly applied to simulated twin samples, type I errors can be highly inflated if additive genetic effects and/or shared environmental effects are present. The empirical type I error also increases as $a^2$ and $c^2$ increase. In contrast, $S_{Hom}$, TwinEQTL, and LMM control the type I errors well under different scenarios. The histograms of the *P*-values from the 3 approaches are shown in Fig. 1. Clearly, that *P*-values from TwinEQTL and LMM are uniformly distributed while the *P*-values from naive LM are enriched closed to 0.

## Power simulations

Similar setups as above in type I error investigation are also used to estimate statistical power. For twin data, since naive LM has inflated power, for each simulated data, we randomly select a subset of samples with a largest number of unrelated samples (i.e. 600 subjects including 100 singletons and one half of total 1,000 twins) on which the linear regression model is valid. Power estimations of LM from the independent subset, $S_{Hom}$ TwinEQTL and LMM are shown in Fig. 2. A total of 10,000 simulations were done in each scenario. Significant power loss is observed in LM where only a subset of subjects is analyzed. We choose significance level of $\alpha = 5 \times 10^{-8}$ in the power analysis in order to recreate the scenarios close to real data analysis in GWAs/eQTL. The power estimates of TwinEQTL and LMM are similar to each other in most cases while $S_{Hom}$ has clear power disadvantage, especially when additive genetic effect is relatively low and when the

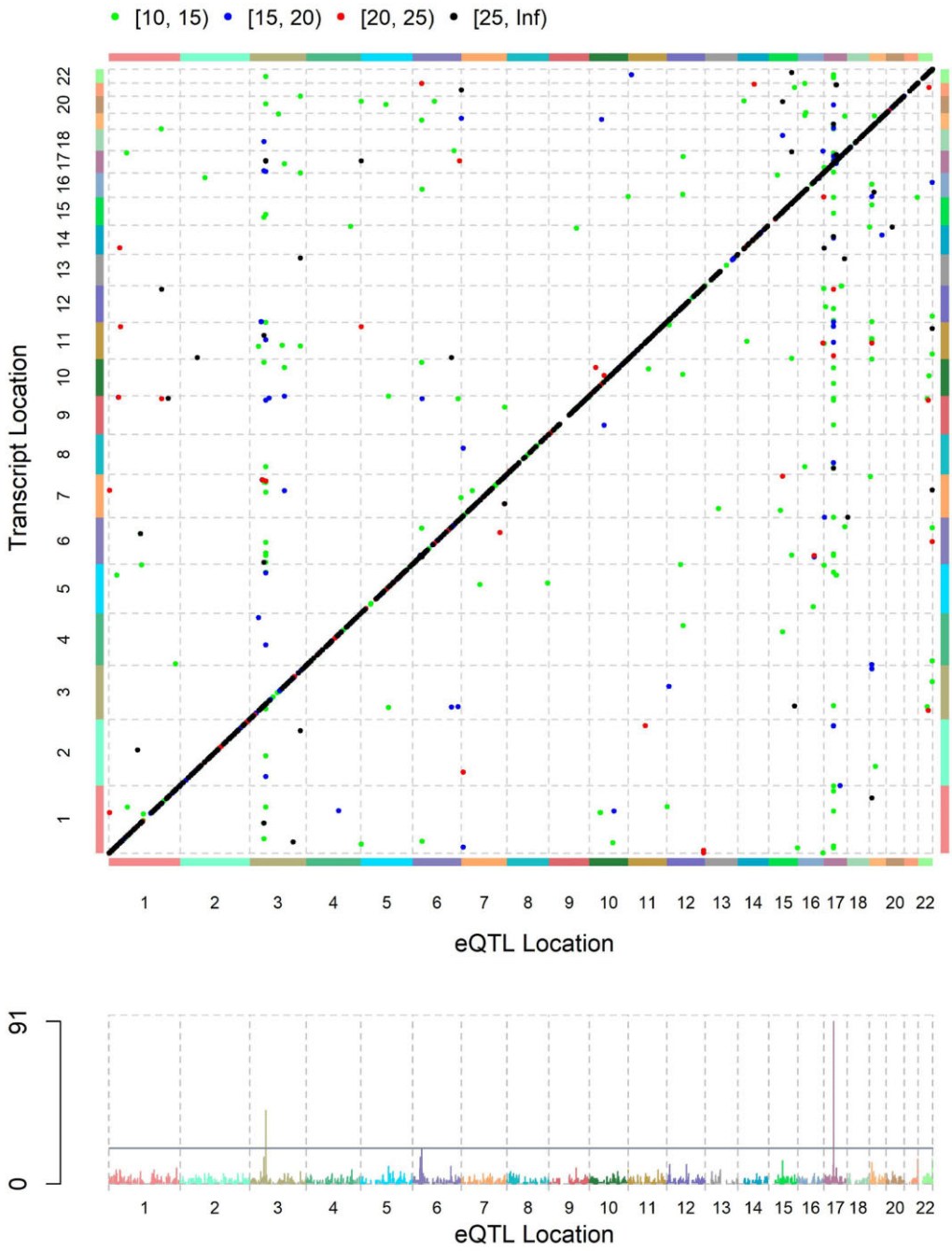


**Fig. 5.** eQTL plot for the SNP–transcript association using TwinEQTL method. The diagonal line shows genome-wide strong local eQTLs and other scattered dots are intra- or interchromosomal distal eQTLs.

DZ are dominate in the samples. Furthermore, we performed additional simulations under various scenarios of $a^2$ and $c^2$, DZ/MZ ratios, significance levels. The results indicate TwinEQTL is a good alternative to LMM at much lower computational cost (Supplementary Fig. 1).

## Application in GWAS of brain MRI imaging

We also applied TwinEQTL to the GWAS study of brain volume in neonates, which investigate the association between neonatal total GM volume, CSF volume, and ICV with genome-wide genetic markers in a twin study. A total of 561 infants (300 male, 261 female) between 0 and 24 weeks of age, encompassing 295 singletons or unpaired twins, 17 sibling pairs, and 232 twins. To perform valid genome-wide association analysis without inflated type I error, we first randomly split the subjects into set1 and set2 while restricting twin/sibling pairs to be separated into set1 and set2. So, within each subset, subjects are independent with each other and subjects between subsets are correlated because of shared families. We then performed TwinEQTL by 2 steps: (1) perform linear regression for both set1 and set2 separately; (2) combining the summary statistics from 2 subsets using either TwinEQTL algorithm or naive meta-analysis without controlling for correlation. Figure 3 shows the genome-wide results of naive meta-analysis and TwinEQTL. Overall *P*-values from naive meta-

analysis are much lower compared with those from TwinEQTL due to poor type I error control which we shown in simulations. Figure 4 shows the advantage of TwinEQTL during this procedure. *P*-values from set1 and set2 are suggestive but nowhere near to genome-wide significant when analyzed separately. After meta analyzed using TwinEQTL, association turns out to be genome-wide significant. More specifically, variant rs11875537, which resides downstream of *METTL4*, was found significantly associated with total CSF volume using TwinEQTL. Interestingly, *METTL4* show elevated prenatal expression in a small subset of brain regions. By examining exon level expression data from Brainspan, we found *METTL4* has specific isoforms with differentially regulated expression across the lifespan (Kang *et al.* 2011). Similarly, another variant rs115842868, which resides downstream of *CAAP1*, was found significantly associated with GM volume. By checking with human brain atlas, we also found that *CAAP1* is also preferentially expressed in several brain tissues (Shen *et al.* 2012). We also found that the protein expression level of *LMCD1* is highest in cerebellum among all the tissues (Shen *et al.* 2012). In addition, variant rs6805949, which resides downstream of *LMCD1*, was found significantly associated with ICV. All of these evidences suggest the potential power of TwinEQTL in identifying biological factors related to brain development.

## Application in NTR twin eQTL study

We applied TwinEQTL to our previously published eQTL analysis of NTR twin data (Wright *et al.* 2014). Due to computational limitation of standard LMM in our previous study, we could only afford to apply LMM–ACE model to a subset of SNP–gene pairs. As a compromise, we used 2-step approach to perform eQTL analysis with twin subjects: (1) perform fast eQTL analysis with MatrixEQTL using linear model while ignoring twin status; (2) perform eQTL analysis only on significant eQTLs from first step with valid but slow LMM–ACE model. With TwinEQTL method, we are able to compute both local and distal eQTLs in 1 step and all pairs of SNP–transcripts associations were computed in <2 h. The SNP–transcripts associations with *P*-values passing certain thresholds are shown in the eQTL plot (Fig. 5). In summary, we were able to identify consistent set of eQTL genes with much less computing time using TwinEQTL method.

## Algorithm performance

We performed speed analysis in PC with 64-bit Windows 10, Intel Xeon E3-1225 CPU, 3.30 GHz, 32 GB RAM and MS R Open version 4.0.2 (64 bit). We compared our method with LMM–ACE model using GWAS dataset of UNC-Chapel Hill EBDS with only genotyped SNPs ($m = 854{,}979$). Our estimate shows it took more than 73 h for LMM–ACE using R, and it took about 1 h for BOLT-LMM to complete same task under High Performance Cluster. By implementing TwinEQTL using both accelerated estimation for ACE model (Chen *et al.* 2019) and efficient matrix multiplication method behind MatrixEQTL (Shabalin 2012), the computing time is reduced to only 22 s once the data has been loaded in R, which is more than 10,000 times increase in computational efficiency compared with LMM–ACE model. To assess the computational performance of TwinEQTL with realistic settings, we simulated eQTL data under different sample sizes ($N = 500, 1K, 5K, 10K$), number of SNPs ($P = 500K, 1M, 2M$), and various variance components within a normal range. The average computational time with 20K transcripts is shown in Table 2, suggesting that the computational complexity of TwinEQTL is linear in *N* and *P*, which is practical for real world GWAS and expression data. The results also suggest that it is acceptable to run eQTL analysis when sample size is <2K. However, when sample size is as large as 5K, a few HPC nodes is recommended for such analysis.

**Table 2.** Computational performance of TwinEQTL for eQTL analysis.

| | Estimated time (hours)[a] | | |
|---|---|---|---|
| **Sample size ($N^b$)** | **500K SNPs** | **1M SNPs** | **2M SNPs** |
| 500 | 2.3 | 4.5 | 18.3 |
| 1,000 | 3.8 | 9.9 | 32.1 |
| 2,000 | 9.0 | 15.4 | 36.3 |
| 5,000 | 22.9 | 42.6 | 91.5 |
| 10,000 | 52.3 | 90.0 | 180.5 |

[a] eQTL analysis using 20,000 simulated gene expression values.
[b] The ratio of MZ vs DZ is 1:1 with 80% of twins in total.

# Discussion

We provided a novel and simple approach, TwinEQTL for both GWAS and eQTL studies with correlated twin subjects. Our approach is significantly faster than traditional method such as LMM and our previous fast alternative (Yin *et al.* 2015). TwinEQTL only needs to estimate the correlation once per trait with LMM–ACE, instead of performing LMM each per SNP–trait pair for millions of times in a typical GWAS. Similar to $S_{Hom}$ (Zhu *et al.* 2015), TwinEQTL uses meta-analysis-like approach to combine test statistics from related subjects, while adjusting for correlation between them, but offers a more accurate correlation estimate and a much faster implementation for twin GWAS. The method is especially useful for twin eQTL analysis, where LMM is prohibited.

There are several limitations that must be considered when using TwinEQTL. First, the current implementation of TwinEQTL only deals with continuous traits or gene expressions. Other types of responses could be implemented in the future if there is any demand. Secondly, TwinEQTL does not account for responses with complicated structure such as longitudinal observations or studies with multilevel family structures, which could be estimated through general linear model with correctly specified variance–covariance component such as compound symmetry, autoregressive model, unstructure, etc., or alternatively using more robust estimation method such as generalized estimating equations. Compared with LMM, statistically, TwinEQTL is expected to experience some efficiency loss unless $\rho_{MZ} = \rho_{DZ} = 0$ or $n_{DZ} = 0$, which agrees with the findings on mega-analysis vs meta-analysis in (Lin and Sullivan 2009; Lin and Zeng 2010b). We have conducted extensive simulations for a wide range of values of $\sigma_a^2$ and $\sigma_c^2$ to empirically demonstrate that the power loss of TwinEQTL is acceptable for practical use given its tremendous computational gain (Supplementary Fig. 1). Furthermore, TwinEQTL requires splitting samples into 2 independent subsets, which could potentially create imbalance of MAF between 2 subsets. This can be problematic for variants with low allele frequency and small sample size because rare variants can become rarer in each subset and will have more leverages to be more influential in the association analysis and might potentially create artifacts by having single extreme outlier (Chatterjee and Hadi 1986). Similar issues have been observed in our own analysis when comparing the GWAS results of TwinEQTL with those from LMM (Xia *et al.* 2017), where top hits in CSF GWAS on chromosome 18 from TwinEQTL were not found by LMM and top hit in GM GWAS on chromosome 4 from LMM was missed by TwinEQTL. Lastly, according to our simulations, TwinEQTL

performs best when most of the subjects are twins in the study. However, when more than half of the subjects are independent, neither LMM nor TwinEQTL have overwhelming power advantage over linear regression with only independent subset. As a result, for study that was not designed specifically for twin, the easiest approach to deal with sporadic twin subjects is just simply select one subject from each twin pair followed by linear regression. However, for twin pairs with substantial portion, TwinEQTL should be the most suitable method.

## Data availability

The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables.

TwinEQTL is implemented in R and is available and maintained under GitHub website https://github.com/andreyshabalin/TwinEQTL. The required input data by TwinEQTL includes genotype data, gene expression data, covariates data and twin status data.

Supplemental material is available at *GENETICS* online.

## Conflicts of interest

None declared.

## Literature cited


Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ, Jackson AU, Vedantam S, Raychaudhuri S, *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010;467(7317):832–838.

Boomsma D, Busjahn A, Peltonen L. Classical twin studies and beyond. Nat Rev Genet. 2002;3(11):872–882.

Carlin JB, Gurrin LC, Sterne JA, Morley R, Dwyer T. Regression models for twin studies: a critical review. Int J Epidemiol. 2005;34(5):1089–1099.

Chatterjee S, Hadi AS. Influential observations, high leverage points, and outliers in linear regression. Stat Sci. 1986;1(3):379–393.

Chen X, Formisano E, Blokland GA, Strike LT, McMahon KL, de Zubicaray GI, Thompson PM, Wright MJ, Winkler AM, Ge T, *et al.* Accelerated estimation and permutation inference for ace modeling. Hum Brain Mapp. 2019;40(12):3488–3507.

Cheung MWL. metasem: an R package for meta-analysis using structural equation modeling. Front Psychol. 2014;5:1521.

Cheung MWL. Computing multivariate effect sizes and their sampling covariance matrices with structural equation modeling: theory, examples, and computer simulations. Front Psychol. 2018;9:1387.

Chou YY, Leporé N, Chiang MC, Avedissian C, Barysheva M, McMahon KL, de Zubicaray GI, Meredith M, Wright MJ, Toga AW, *et al.* Mapping genetic influences on ventricular structure in twins. Neuroimage. 2009;44(4):1312–1323.

Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379–389.

Ghazalpour A, Doss S, Kang H, Farber C, Wen PZ, Brozell A, Castellanos R, Eskin E, Smith DJ, Drake TA, *et al.* High-resolution mapping of gene expression using association in an outbred mouse stock. PLoS Genet. 2008;4(8):e1000149.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, *et al.* Spatio-temporal transcriptome of the human brain. Nature. 2011;478(7370):483–489.

Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet. 2009;41(1):56–65.

Kuna ST, Maislin G, Pack FM, Staley B, Hachadoorian R, Coccaro EF, Pack AI. Heritability of performance deficit accumulation during acute sleep deprivation in twins. Sleep. 2012;35:1223–1233.

Lin D, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genet Epidemiol. 2010a;34(1):60–66.

Lin D, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika. 2010b;97(2):321–332.

Lin DY, Sullivan PF. Meta-analysis of genome-wide association studies with overlapping subjects. Am J Hum Genet. 2009;85(6):862–872.

Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284–290.

Neale M, Cardon L. Methodology for Genetic Studies of Twins and Families. Number 67. Berlin: Springer Science & Business Media; 1992.

Park JH, Song YM, Sung J, Lee K, Kim YS, Kim T, Cho SI. The association between fat and lean mass and bone mineral density: the healthy twin study. Bone. 2012;50(4):1006–1011.

Rabe-Hesketh S, Skrondal A, Gjessing H. Biometrical modeling of twin and family data using standard mixed model software. Biometrics. 2008;64(1):280–288.

Ripke S, Neale BM, Corvin A, Walters JT, Farh KH, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H, *et al.* Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421.

Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kähler AK, Akterin S, Bergen SE, Collins AL, Crowley JJ, Fromer M, *et al.*; Wellcome Trust Case Control Consortium 2. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet. 2013;45(10):1150–1159.

Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28(10):1353–1358.

Shen EH, Overly CC, Jones AR. The Allen Human Brain Atlas: comprehensive gene expression mapping of the human brain. Trends Neurosci. 2012;35(12):711–714.

Silventoinen K, Sammalisto S, Perola M, Boomsma DI, Cornes BK, Davis C, Dunkel L, De Lange M, Harris JR, Hjelmborg JV, *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res. 2003;6(5):399–408.

Vaccarino V, Brennan ML, Miller AH, Bremner JD, Ritchie JC, Lindau F, Veledar E, Su S, Murrah NV, Jones L, *et al.* Association of major depressive disorder with serum myeloperoxidase and other markers of inflammation: a twin study. Biol Psychiatry. 2008;64(6):476–483.

Wang X, Guo X, He M, Zhang H. Statistical inference in mixed models and analysis of twin and family data. Biometrics. 2011;67(3):987–995.

Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, Madar V, Jansen R, Chung W, Zhou YH, *et al.* Heritability and genomics of gene expression in peripheral blood. Nat Genet. 2014;46(5):430–437.

Xia K, Zhang J, Ahn M, Jha S, Crowley J, Szatkiewicz J, Li T, Zou F, Zhu H, Hibar D, *et al.*; ENIGMA Consortium Genome-wide association analysis identifies common variants influencing infant brain volumes. Transl Psychiatry. 2017;7(8):e1188.

Yin Z, Xia K, Chung W, Sullivan PF, Zou F. Fast eQTL analysis for twin studies. Genet Epidemiol. 2015;39(5):357–365.

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, *et al.*; Wellcome Trust Case Control Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008;40(5):638–645.

Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, Smith JA, Yanek LR, Sun YV, Edwards TL, *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. Am J Hum Genet. 2015;96(1):21–36.

*Communicating editor: H. Zhao*

## Appendix

We describe the method first assuming no covariates exist. Extension to situations where one or more covariates exist will be discussed later. Suppose for a given GWAS, there are $n_{MZ}$ pairs of MZ, $n_{DZ}$ pairs of DZ, and $n_{Sg}$ singletons. The total sample size $n = 2n_{MZ} + 2n_{DZ} + n_{Sg}$. First, we randomly split each twin pair into 2 groups, named group 1 and group 2. We then randomly divide the singletons into half and assign them to group 1 and group 2 separately. Samples in groups 1 and 2 are ordered in such a way that the first $n_{MZ}$ samples in groups 1 and 2 are the paired MZ samples, and the $n_{DZ}$ samples are the paired DZ samples, and the remaining samples are singletons. Now for samples within group $k$ ($k = 1, 2$), they are genetically unrelated, on which the simple linear model below can be performed between a given SNP and the trait:

$$y_{ik} = \mu_k + \beta g_{ik} + \epsilon_{ik} \tag{6}$$

where $g_{ik}$ and $y_{ik}$ are the corresponding genotype and phenotype of subject $i (i = 1, \ldots, n_k)$ in group $k$, and the random error $\epsilon_{ik} \sim N(0, \sigma_k^2)$. For simplification, we assume that both $y$ and $g$ are standardized to have mean 0 and variance 1 within each group.

The MLE of $\widehat{\beta}$ from the 2 subsets ($k = 1, 2$) therefore equals

$$\widehat{\beta}_k = \frac{\sum_{i=1}^{n_k}(g_{ik} - \overline{g_k})(y_{ik} - \overline{y_k})}{\sum_{i=1}^{n_k}(g_{ik} - \overline{g_k})^2} = \frac{\sum_{i=1}^{n_k} y_{ik} g_{ik}}{n_k - 1}. \tag{7}$$

The correlation between $\epsilon_{i1}$ and $\epsilon_{j2}$ is

$$\mathrm{Corr}(\epsilon_{i_1}, \epsilon_{j_2}) = \begin{cases} \rho_{DZ}, & \text{subjects } i \text{ and } j \text{ are a DZ pair} \\ \rho_{MZ}, & \text{subjects } i \text{ and } j \text{ are a MZ pair} \\ 0, & \text{subjects } i \text{ and } j \text{ are unrelated,} \end{cases} \tag{8}$$

according to the common ACE model (Neale and Cardon 1992) for twin data. Here

$$\rho_{DZ} = \left(\frac{1}{2}\sigma_a^2 + \sigma_c^2\right) / \left(\sigma_a^2 + \sigma_c^2 + \sigma_e^2\right) \text{and}$$

$$\rho_{MZ} = (\sigma_a^2 + \sigma_c^2)/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)$$

with $\sigma_a^2$ being the additive genetic effect and $\sigma_c^2$ being the shared common environment effect.

To combine the test results from the 2 subsets, we use the well-known inverse-variance estimator of $\beta$ in traditional meta-analysis (Lin and Zeng 2010b) to estimate the SNP effect $\beta$ as

$$\widehat{\beta} = \frac{\sum_{k=1}^{K} \widehat{\beta}_k / \widehat{\mathrm{var}}(\widehat{\beta}_k)}{\sum_{k=1}^{K} 1/\widehat{\mathrm{var}}(\widehat{\beta}_k)}, \tag{9}$$

which can be further simplified to $\widehat{\beta} = (\widehat{\beta_1} + \widehat{\beta_2})/2$ for our situation since we randomly and evenly split the data, and thus expect the genetic effect estimates from the 2 subsets are equally efficient. Such meta-analysis idea is not new and has been used for integrating GWAS data with shared control samples (Lin and Sullivan 2009).

The variance of $\widehat{\beta}$ therefore equals

$$\begin{aligned} \mathrm{var}\left(\widehat{\beta}\right) &= \mathrm{var}\left(\frac{\widehat{\beta_1} + \widehat{\beta_2}}{2}\right) = \frac{1}{4}\mathrm{var}\left(\widehat{\beta_1}\right) + \frac{1}{4}\mathrm{var}\left(\widehat{\beta_2}\right) + \frac{1}{2}\mathrm{Cov}\left(\widehat{\beta_1}, \widehat{\beta_2}\right) \\ &= \frac{1}{4}\mathrm{var}\left(\widehat{\beta_1}\right) + \frac{1}{4}\mathrm{var}\left(\widehat{\beta_2}\right) + \frac{1}{2}\left(\mathrm{E}\left(\widehat{\beta_1}\widehat{\beta_2}\right) - \mathrm{E}\left(\widehat{\beta_1}\right)\mathrm{E}\left(\widehat{\beta_2}\right)\right) \\ &= \frac{1}{4}\mathrm{var}\left(\widehat{\beta_1}\right) + \frac{1}{4}\mathrm{var}\left(\widehat{\beta_2}\right) + \frac{1}{2}\mathrm{E}\left(\widehat{\beta_1}\widehat{\beta_2}\right) \end{aligned} \tag{10}$$

Under $H_0$, $E(\widehat{\beta_1}) = E(\widehat{\beta_2}) = 0$. Further, due to the special ordering of the samples in groups 1 and 2, we have

$$\begin{aligned} \mathrm{E}\left(\widehat{\beta_1}\widehat{\beta_2}\right) &= \mathrm{E}\left(\left(\frac{1}{n_1}\sum_{i}^{n_1} y_{i1} g_{i1}\right)\left(\frac{1}{n_2}\sum_{j}^{n_2} y_{j2} g_{j2}\right)\right) \\ &= \frac{1}{n_1 n_2}\left(\sum_{i,j} \mathrm{E}(y_{i1} g_{i1} y_{j2} g_{j2})\right) \\ &= \frac{1}{n_1 n_2}\left(\sum_{i,j} \mathrm{E}((\beta_1 g_{i1} + \epsilon_{i1}) g_{i1} (\beta_2 g_{j2} + \epsilon_{j2}) g_{j2})\right) \\ &= \frac{1}{n_1 n_2}\left(\sum_{i,j} \mathrm{E}(\epsilon_{i1} g_{i1} \epsilon_{j2} g_{j2})\right) = \frac{1}{n_1 n_2}\left(\sum_{i,j} \mathrm{E}(\epsilon_{i1}\epsilon_{j2})\mathrm{E}(g_{i1} g_{j2})\right) \\ &= \frac{1}{n_1 n_2}\left(\sum_{i=1}^{n_{MZ}} \mathrm{E}(\epsilon_{i1}\epsilon_{i2})\mathrm{E}(g_{i1} g_{i2}) + \sum_{i=n_{MZ}+1}^{n_{DZ}+n_{MZ}} \mathrm{E}(\epsilon_{i1}\epsilon_{i2})\mathrm{E}(g_{i1} g_{i2})\right) \\ &= \frac{1}{n_1 n_2}\left(\rho_{DZ}\sum_{i=1}^{n_{MZ}} \mathrm{E}(g_{i1} g_{i2}) + \rho_{DZ}\sum_{i=n_{MZ}+1}^{n_{MZ}+n_{DZ}} \mathrm{E}(g_{i1} g_{i2})\right) \\ &= \frac{1}{n_1 n_2}\left(\rho_{MZ} n_{MZ} E_{MZ}(g_1, g_2) + \rho_{DZ} n_{DZ} E_{DZ}(g_1, g_2)\right), \end{aligned} \tag{11}$$

where $g_1$ and $g_2$ are defined as genotypes of a given twin pair.

Instead of calculating $E(g_1, g_2)$ for DZ and MZ pairs, it is easier to estimate $E(g'_{i1}, g'_{i2})$, where $g'_{i1}$ and $g'_{i2}$ are original genotypes coded as 0, 1, and 2 as the number of minor alleles of each genotype. Suppose the MAF of $g'_i$ is $f$, then under HWE, we have

$$E(g'_1 g'_2 | \mathrm{IBD}) = \begin{cases} (2f)^2, & \text{if IBD} = 0 \\ f + 3f^2 & \text{if IBD} = 1 \\ 2f(1-f) + (2f)^2, & \text{if IBD} = 2, \end{cases} \tag{12}$$

since

1) if IBD = 0, $g'_1$ and $g'_2$ are essentially independent of each other so $E(g'_1 g'_2 | \mathrm{IBD} = 0) = E(g'_1)E(g'_2) = (2f)^2$;
2) if IBD = 2, $g'_1$ and $g'_2 Z g'_2$ are identical, so $E(g'_1 g'_2 | \mathrm{IBD} = 2) = E(g_1^2) = \mathrm{var}(g'_1) + E^2(g'_1) = 2f(1-f) + (2f)^2$;
3) if IBD = 1, $g'_1$ and $g'_2$ share 1 allele IBD, so we have

$$\begin{aligned} &E(g'_1 g'_2 | IBD = 1) = E((x + y)(x + z))) = E(x^2 + yx + zx + yz) \\ &= E(x^2) + E(xy) + E(xz) + E(yz) = f + f^2 + f^2 + f^2 = f + 3f^2, \end{aligned} \tag{13}$$

where x, y, and z are the constituting alleles of the 2 twin samples and each follows *Bernoulli*($f$).

Based on the above calculations, we can get

$$\begin{aligned} E_{MZ}(g_1 g_2) &= \frac{E(g'_1 g'_2 | \mathrm{IBD} = 2) - (2f)^2}{2f(1-f)} \\ &= \frac{2f(1-f) + (2f)^2 - (2f)^2}{2f(1-f)} = 1 \end{aligned} \tag{14}$$

For DZ pairs, the IBD can take values 0, 1, or 2 with probabilities 1/4, 1/2, and 1/4, respectively. Thus

$$\begin{aligned} E(g_1 g_2) &= \frac{E(g'_1 g'_2) - (2f)^2}{2f(1-f)} \\ &= \frac{\frac{1}{4}E(g'_1 g'_2|\text{IBD}=0) + \frac{1}{2}E(g'_1 g'_2|\text{IBD}=1)}{2f(1-f)} \\ &\quad + \frac{\frac{1}{4}E(g'_1 g'_2|\text{IBD}=2) - (2f)^2}{2f(1-f)} \\ &= \frac{\frac{1}{4}(4f^2) + \frac{1}{2}(f+3f^2) + \frac{1}{4}\left(2f(1-f)+4f^2\right) - (2f)^2}{2f(1-f)} = \frac{1}{2} \end{aligned} \tag{15}$$

After plugging Equations (14) and (15) into Equation (11), we obtain

$$E\left(\widehat{\beta_1}\widehat{\beta_2}\right) = \frac{1}{n_1 n_2}\left(\frac{1}{2}n_{DZ}\rho_{DZ} + n_{MZ}\rho_{MZ}\right) = \frac{n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ}}{2n_1 n_2}, \tag{16}$$

which can be further plugged into equation 10 to prove that

$$\begin{aligned} \text{var}\left(\widehat{\beta}\right) &= \frac{1}{4}\text{var}\left(\widehat{\beta_1}\right) + \frac{1}{4}\text{var}\left(\widehat{\beta_2}\right) + \frac{1}{2}E\left(\widehat{\beta_1}\widehat{\beta_2}\right) \\ &\approx \frac{1}{4(n_1-1)} + \frac{1}{4(n_2-1)} + \frac{n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ}}{4n_1 n_2} \\ &\approx \frac{n_1 + n_2 + n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ}}{4n_1 n_2}. \end{aligned} \tag{17}$$

Finally, the metatest statistic based on the estimates of $\beta$ from the 2 groups can be expressed as

$$\begin{aligned} Z &= \frac{\widehat{\beta}}{\sqrt{\text{var}(\widehat{\beta})}} = \frac{(\widehat{\beta_1} + \widehat{\beta_2})/2}{\sqrt{(n_1 + n_2 + n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ})/4n_1 n_2}} \\ &= \frac{T_1/\sqrt{n_1} + T_2/\sqrt{n_2}}{\sqrt{(n_1 + n_2 + n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ})/n_1 n_2}} \\ &= \frac{T_1 + T_2}{\sqrt{1 + 1 + 2(n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ})/n}} \\ &= \frac{T_1 + T_2}{\sqrt{1 + 1 + 2\text{Corr}(T_1, T_2)}} \end{aligned} \tag{18}$$

where $\text{Corr}(T_1, T_2) = (n_{DZ}\rho_{DZ} + 2n_{MZ}\rho_{MZ})/n$, which is the correlation between $T_1$ and $T_2$, with $T_1$ and $T_2$ corresponding $t$-statistics from the set1 and set2 data, respectively. Under $H_0: Z \sim N(0,1)$.