# Improved Use of Small Reference Panels for Conditional and Joint Analysis with GWAS Summary Statistics

Yangqing Deng[1], Wei Pan[1,2]

[1]*Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA*

[2]*Corresponding author:* weip@biostat.umn.edu

February 9, 2018; Revised April 3, 2018

## Abstract

Due to issues of practicality and confidentiality of genomic data-sharing on a large scale, typically only meta- or mega-analyzed GWAS summary data, not individual-level data, are publicly available. Re-analyses of such GWAS summary data for a wide range of applications have become more and more common and useful, which often require the use of an external reference panel with individual-level genotypic data to infer linkage disequilibrium (LD) among genetic variants. However, with a small sample size in only hundreds, as for the most popular 1000 Genomes Project European sample, estimation errors for LD are not negligible, leading to often dramatically increased numbers of false positives in subsequent analyses of GWAS summary data. To alleviate the problem in the context of association testing for a group of SNPs, we propose an alternative estimator of the covariance matrix with an idea similar to multiple imputation. We use numerical examples based on both simulated and real data to demonstrate the severe problem with the use of the 1000 Genomes Project reference panels, and the improved performance of our new approach.

**Key words**: 1000 Genomes Project; COJO analysis; Gene-based testing; Multiple imputation; Multiple SNPs; Type I error; Wald test.

## Introduction

Due to logistic reasons and privacy concerns, often individual-level genotypic and phenotypic data from large genome-wide association studies (GWAS) are not publicly available; in contrast, GWAS summary statistics, in the form of the z-scores and/or p-values of single SNPs based on their univariate/marginal associations with a GWAS trait, are publicly available. Interestingly, as first demonstrated by Yang et al. (2012), one can combine GWAS summary statistics with a reference panel with individual-level genotypic data (that mimic the original population for the GWAS summary data) to conduct conditional and joint (COJO) analyses; that is, one can estimate and test the joint effects of multiple SNPs in a genomic region such as a gene or a pathway, which may be more powerful and/or interpretable than the standard/marginal single SNP-based analysis. In addition, based on a joint regression model for multiple SNPs, one can conduct COJO analysis: conditional on (i.e. after accounting for the effects of) other SNPs, one can estimate and test for possible association of one SNP (or multiple SNPs), which is useful in sorting out multiple causal SNPs as in fine mapping (Hormozdiari et al. 2014; Kichaev et al. 2014; Chen et al. 2015). The usefulness of such analyses and other ones on some publicly available GWAS summary datasets has been nicely reviewed in Pasaniuc and Price (2017).

A critical issue in these approaches with GWAS summary statistics is to estimate linkage disequilibrium (LD) among the SNPs in a genomic region using a reference panel, which is necessary for estimating the correlation or covariance matrices of various parameter estimates and their associated testing statistics in any subsequent conditional or joint analysis. As pointed out by Pasaniuc and Price (2017), “Conditional association and imputation using summary statistics crucially rely on accurate LD information from a population reference panel. Even in the best case, when the reference population closely matches the GWAS population, the relatively small size of reference panels for which LD information is publicly available (typically hundreds or at most thousands of individuals)

makes accurate estimation of a large number of LD parameters a challenge." Although regularization-based methods for estimating LD or covariance matrices have been discussed in other related contexts, as to be shown, it is unclear how to choose tuning parameters associated with any regularization-based method in the current context of statistical inference, not estimation or prediction (Pasaniuc et al 2014; Kichaev and Pasaniuc 2015; Shi et al 2016). In spite of the obvious importance of the issue, however, it has been largely ignored in practice: although a small reference panel with a sample size in hundreds, such as any of the 1000 Genomes Project (1000G) racial/ethnic group-specific panels, is often used, strikingly there has been barely any assessment on its effects on subsequent statistical inference. As to be shown here, both expectedly and surprisingly, it often leads to dramatically inflated type I error rates and thus large numbers of false positives. We also point out that Yang et al. (2012) used a reference sample of size over 6000, more than 10 times larger than that of the popular choice with the 1000G data in practice. Even though the total sample size of the 1000G data (The 1000 Genomes Project Consortium 2015) is more than 3000, due to the presence of multiple populations or ethnic groups, its sample size for a single population is no more than a few hundreds, which is often used in practice.

We emphasize that the underlying issue discussed here is quite general and wide-ranging: although our focus is on conditional and joint analyses of multiple SNPs such as in gene-based testing and fine mapping, any approach using GWAS summary data and a reference panel may suffer from the same problem, no matter it is polygenic risk prediction (Vilhjalmsson et al. 2015), or inferring genetic correlations among complex traits (Bulik-Sullivan et al. 2015), or Mendelian randomization for causal inference (Burgess et al. 2013). Very recently Benner et al (2017) demonstrated the severe problem in the context of fine-mapping, while we consider both conditional and global testing with a group of SNPs. More importantly, we propose a new method to alleviate the problem. We note that, even if a reference panel comes from the same population of the GWAS data, using the reference data with a small sample size may still lead to increased numbers of false positives. Of course, if the reference sample is from a

different population, the situation becomes worse; here we mainly focus on the former case. The main issue is the ignorance of the small sample size of the reference panel, and thus its associated estimation errors or uncertainties. Accordingly, we propose using an idea similar to multiple imputations (Rubin 1996) to alleviate the problem. We provide numerical examples based on both simulated and real data to show the impact of small reference panels, even when they are drawn from the same GWAS population, and the effectiveness of our proposed multiple imputation-type (MI-type) approach.

## Methods

To be concrete and general, we focus on the joint analysis of a group of SNPs in a genomic region, and on the COJO analysis of one of the SNPs (after accounting for the effects of other SNPs) with a single quantitative trait. For a quantitative trait with a normal distribution, although an F-test is exact, due to the large sample size of a typical GWAS, we restrict our attention to the asymptotically equivalent Wald chi-squared test.

Suppose that we are interested in $L$ SNPs and one trait, denoted by $\mathbf{X}_l$ ($l = 1 \ldots L$) and $\mathbf{Y}$, which are $n \times 1$ vectors for $n$ subjects. Given the summary statistics $\widehat{b_l}$ and $\widehat{\text{var}}(\widehat{b_l})$ in marginal analysis

$$\mathbf{Y} = b_l \mathbf{X}_l + \mathbf{e}, \qquad (1)$$

as well as $\mathbf{C}$, an estimate of the correlation matrix of the SNPs from a reference panel, we can obtain the regression coefficient estimates and their covariance matrix, denoted as $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ respectively (Yang et al. 2012), in the joint model

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \ldots + \beta_L \mathbf{X}_L + \boldsymbol{\varepsilon}. \qquad (2)$$

The Wald test statistic is

$$W = \widehat{\boldsymbol{\beta}}' \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\beta}}. \qquad (3)$$

Under the global/overall null hypothesis $H_0$: $\boldsymbol{\beta} = (\beta_1 \quad \ldots \quad \beta_L)' = 0$, $W$ asymptotically (approximately) follows a chi-squared distribution with $L$ degrees of freedom. When the covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is not invertible, we use the Moore-Penrose generalized inverse and modify the degrees of freedom as the rank of the covariance matrix.

From our experience, using the estimated correlations of the SNPs based on a small reference panel may lead to inflated type I errors, since $\widehat{\boldsymbol{\Sigma}}$ may not be accurate enough. From a different angle, we can regard that using the reference sample only once to estimate LD among the SNPs ignores the non-negligible uncertainty in the resulting estimate due to the small sample size. To account for the estimation uncertainty, we borrow the idea of multiple imputations (MI) (Rubin 1996) and propose a MI-type method. Specifically, in addition to using the reference panel to build one model and obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$, we can use the data $(T-1)$ more times to get $(T-1)$ estimates of the coefficients, and then inflate the covariance estimates for a more conservative inference (to battle the issue of inflated type I errors). We denote the parameter estimate and its covariance matrix based on the complete reference panel as $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\Sigma}}_1$. In each imputation $t$ $(t>1)$, sample $n_{\text{ref}}$ subjects from the $n_{\text{ref}}$ subjects in the reference panel with replacement. Take these subjects as the reference sample to build a joint model. Estimate the coefficients as $\widehat{\boldsymbol{\beta}}_t$ and its covariance matrix as $\widehat{\boldsymbol{\Sigma}}_t$. Calculate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ using the following formulas, then carry out the Wald test (3):

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}}_1,\\
\mathbf{V}_w &= \widehat{\boldsymbol{\Sigma}}_1,\\
\mathbf{V}_b &= \tfrac{1}{T-2}\textstyle\sum_{t=2}^{T}(\widehat{\boldsymbol{\beta}}_t-\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_t-\widetilde{\boldsymbol{\beta}})',\\
\widehat{\boldsymbol{\Sigma}} &= \mathbf{V}_w+(1+\tfrac{1}{T})\mathbf{V}_b,\\
\widetilde{\boldsymbol{\beta}} &= \tfrac{1}{t-1}\textstyle\sum_{t=2}^{T}\widehat{\boldsymbol{\beta}}_t.
\end{aligned} \tag{4}$$

From our experience, as in MI, usually setting $T$ no more than 50 should be enough. We can regard the existing approach of using the complete reference panel only once as a special case of our proposed MI-type approach with $T=1$; with a slight abuse of notation, we denote the approach using individual level data as $T=0$.

COJO analysis on an individual SNP can be conducted based on $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$, which may be based on individual-level data, or the complete reference sample, or MI-type estimation. For example, if we'd like to test $H_{0j}$: $\beta_j=0$ against $H_{1j}$: $\beta_j\neq 0$, then the Wald test statistic is

$$W_j=\hat{\beta}_j{}^2/\hat{\Sigma}_{jj}, \tag{5}$$

where $\hat{\Sigma}_{jj}$ is the $j$th diagonal element of $\hat{\boldsymbol{\Sigma}}$. Under the null hypothesis $H_{0j}$, the test statistic follows a chi-squared distribution with 1 degree of freedom.

There are four ways to conduct a COJO analysis on a SNP (conditioning on other SNPs) or a group of SNPs. The first, denoted "Ind", is based on using individual-level data. The second, denoted uses summary statistics but the LD matrix X'X is estimated from the original data (i.e. assuming the availability of LD from the original data); it was confirmed to give the results exactly the same as that of the first method "Ind", and hence is omitted in the sequel. The third one is the naive method of using summary statistics with a reference panel to estimate LD or the matrix X'X only once, i.e. with $T = 1$. The fourth method, denoted "Sum-MI", is our proposed new MI-type method with $T > 1$.

Alternatively, a general class of approaches to better estimating LD or covariance matrices is to apply regularizations: one is to truncate the eigen-values of a matrix based on its singular value decomposition (SVD) (Shi et al 2016), while the other is to impose a penalty like the ridge penalty (Pasaniuc et al 2014; Kichaev and Pasaniuc 2015), which have been studied in other contexts. The general ideas can be applied here. Specifically, we can decompose $\hat{\mathbf{C}}$, estimated from the complete reference panel (with T=1), as

$$\hat{\mathbf{C}} = \sum_{s=1}^{S} \boldsymbol{u}_s d_s \boldsymbol{u}_s', \qquad (6)$$

where $d_l$, $\boldsymbol{u}_l$ are the $l$th largest eigenvalue and the $l$th eigenvector of $\hat{\mathbf{C}}$. Applying the ridge penalty is equivalent to replacing $\hat{\mathbf{C}}$ with $\sum_{s=1}^{S} \boldsymbol{u}_s (d_s + \lambda) \boldsymbol{u}_s'$, while the truncation is to replace $\hat{\mathbf{C}}$ with $\sum_{s=1}^{S} \boldsymbol{u}_l d_l \boldsymbol{u}_l'$; $\lambda \geq 0$ and $0 < E < L$ are the corresponding tuning parameters. Instead of $\hat{\mathbf{C}}$, one can also apply each of the two regularization methods to the estimated X'X, the LD matrix. Then we can carry out the Wald test as usual. As expected and to be shown, the performance of the regularization methods critically depends on the choice of the tuning parameters; however, differing from estimation and prediction, it is quite difficult and largely unknown how to choose tuning parameters for a regularization method in the current context of hypothesis testing.

## Data Availability

The 2013 lipid data (Willer et al. 2013) is publicly available at http://csg.sph.umich.edu/abecasis/public/lipids2013/. The Lung Health Study data can be downloaded from the dbGaP database (accession: phs000335.v3.p2) by request. Information on the WTCCC data and how to apply for access can be found at https://www.wtccc.org.uk/info/access_to_data_samples.html. The method is implemented and freely available in R package *jointsum* at https://github.com/yangq001/conditional. The package will also be available on CRAN soon.

## Results

### Simulations

To investigate the reference panels' impact on the testing performance, we first did some simulation studies. To be as realistic as possible, we used the individual-level genotypic data of 2938 subjects in the control group from the WTCCC data (Burton et al 2007). We randomly chose some SNPs in genomic regions on chromosome 19 so that none of the pair-wise (absolute) correlations was greater than 0.9. For power study, we needed to specify effect sizes. So we used the lipid data (Willer et al. 2013) to build a joint model for triglycerides (TG) versus SNPs. Denote its coefficients by $\boldsymbol{\beta}^*$. Since the lipid data only contains summary statistics, we used the correlations of the SNPs estimated from the WTCCC data. We scaled the significant effects (p-value < 5e-8) while forcing insignificant effects as zero to obtain a true regression model. Then we generated a quantitative trait for the 2938 subjects using the model (2) with no intercept and $\beta_i = k\beta_i^*$ if $\beta_i^*$ (for SNP i) was significant; $\beta_i = 0$ otherwise, and the error term $\boldsymbol{\varepsilon}$ was an independent normal random variable with mean 0 and variance obtained from the joint model. For each replication, we randomly chose $n_{\text{ref}}$ and $n$ subjects from $N = 2938$ subjects as the reference panel and the GWAS

sample respectively.

For the approaches based on summary statistics, in addition to a subsample of the WTCCC data, we also chose the 1000G data as our reference panel. We used the 379 CEU (Utah Residents with Northern and Western Ancestry) samples from the 1000G Phase I version 3 Shapeit2 Reference data from the KGG software website (Li et al. 2012), denoted as 1000G A. We calculated the rejection rates based on 10000 replications for the null case.

**Table 1** Type I error rates for simulations in a region with 8 SNPs, none of which was associated with the trait. The nominal significance level was at 0.05; in each set-up there were 10000 replications; the GWAS sample size was $n = 1000$.

| $n_{\text{ref}}$ | $H_0: \beta_1 = 0$ | | | $H_0: \boldsymbol{\beta} = \mathbf{0}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $T = 1$ | $T = 5$ | $T = 10$ | $T = 1$ | $T = 5$ | $T = 10$ | $T = 20$ | $T = 30$ | $T = 50$ |
| 900 | 0.0531 | 0.0504 | 0.0502 | 0.0621 | 0.0518 | 0.0507 | 0.0512 | 0.0509 | 0.0510 |
| 500 | 0.0559 | 0.0504 | 0.0506 | 0.0694 | 0.0520 | 0.0500 | 0.0507 | 0.0499 | 0.0507 |
| 379 (1000G A) | 0.0548 | 0.0516 | 0.0520 | 0.0092 | 0.0062 | 0.0055 | 0.0057 | 0.0057 | 0.0057 |
| Ind (individual) | 0.0515 | | | 0.0515 | | | | | |

**Table 2** Type I error rates of the regularization methods for simulations in a region with 8 SNPs, none of which was associated with the trait. The null hypothesis tested was $H_0: \boldsymbol{\beta} = \mathbf{0}$ with the nominal significance level at 0.05; in each set-up there were 10000 replications; the GWAS sample size was $n = 1000$.

| $n_{\text{ref}}$ | ridge penalty | | | | SVD-truncation (# eigenvalues) | | |
|---|---|---|---|---|---|---|---|
| | $\lambda = 0$ | $\lambda = 0.05$ | $\lambda = 0.1$ | $\lambda = 0.2$ | $S = 5$ | $S = 6$ | $S = 7$ |
| 900 | 0.0621 | 0.0248 | 0.0164 | 0.0108 | 0.920 | 0.0140 | 0.0362 |
| 500 | 0.0694 | 0.0270 | 0.0175 | 0.0115 | 0.833 | 0.0162 | 0.0400 |
| 379 (1000G A) | 0.0092 | 0.0048 | 0.0048 | 0.0042 | 0.127 | 0.0022 | 0.0046 |
| Ind | 0.0515 | | | | | | |

In a representative region with 8 SNPs, as shown in Table 1, in both conditional and joint analyses, using a reference sample of size 500 or 900 drawn from the same population led to inflated type I error rates, while our proposed approach largely corrected the problem with a small T=10. It is noteworthy that using the 1000G reference panel also gave an inflated type I error rate for the COJO analysis in the naive approach, but yielded very conservative global testing. One possible explanation for the latter is the possible difference inherent between the 1000G data and the WTCCC data: the correlation structure of the 8 SNPs in the reference data was different from that of the WTCCC data, leading to a huge difference in the test statistics. Since the Wald test statistic involves the inverse of the correlation matrix, we examined the eigen-values of the inverse correlation matrices estimated from the individual-level WTCCC data and that from the 1000G reference data: their largest eigen-values were 17.1 and 14.6 respectively, explaining why using the 1000G reference data led to a lower rejection rate than that of the nominal level. For this situation, it is unknown how to avoid conservative inference; our method cannot avoid it either.

We also considered the two regularization methods. Table 2 shows the results for regularizing X'X. As expected, the performance critically depends on the choice of the tuning parameter, which however is unknown. The same conclusion can be drawn on regularizing the covariance matrix $\hat{\mathbf{C}}$. For example, for the SVD-truncation method keeping E= 3, 5 and 7 top eigen-values: the empirical type I error rates were 1) 0.002, 0.012 and 0.039, respectively, with the reference panel of 900 subjects; 2) 2e-4, 0.001 and 0.005, respectively, with the 1000G A reference panel.

**Table 3** Empirical power for simulations in a region with 8 SNPs. The nominal significance level was at 0.05; in each set-up there were 3000 replications; the GWAS sample size was n=1000.

| $\beta_1$ | $n_{\text{ref}}$ | $H_0: \beta_1 = 0$ | | | $H_0: \boldsymbol{\beta} = \mathbf{0}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $T = 1$ | $T = 5$ | $T = 10$ | $T = 1$ | $T = 5$ | $T = 10$ | $T = 30$ | $T = 50$ |
| 0.01 | 900 | 0.738 | 0.739 | 0.739 | 0.408 | 0.395 | 0.393 | 0.398 | 0.399 |
| | 500 | 0.738 | 0.737 | 0.737 | 0.415 | 0.388 | 0.388 | 0.388 | 0.389 |
| | 379 (1000G A) | 0.736 | 0.734 | 0.738 | 0.279 | 0.259 | 0.257 | 0.249 | 0.250 |
| | Ind | 0.731 | | | 0.405 | | | | |
| 0.015 | 900 | 0.970 | 0.970 | 0.970 | 0.808 | 0.798 | 0.800 | 0.798 | 0.798 |
| | 500 | 0.971 | 0.970 | 0.971 | 0.810 | 0.793 | 0.794 | 0.793 | 0.792 |
| | 379 (1000G A) | 0.970 | 0.970 | 0.970 | 0.722 | 0.711 | 0.710 | 0.702 | 0.702 |
| | Ind | 0.970 | | | 0.806 | | | | |

For empirical power, as shown in Table 3, in all situations corresponding to the anti-conservative inference of the naive approach, our proposed method barely lost power as compared to the individual-level data-based method (or the naive method "Sum"). On the other hand, for the global testing, due to the its conservativeness with the use of the 1000G A reference panel (for its possible difference from the WTCCC data), there was some power loss from the naive and our new methods based on summary statistics as compared to the individual-level data-based method; nevertheless, at least compared to the naive method, our method lost only minimal power.

**Table 4** Empirical type I error rate (with $\beta_1$=0) and power (with $\beta_1 = 0.001$) for simulations in 100 regions with 5~37 SNPs (17.8 on average), none or only the first one of which was associated with the trait. The nominal significance level was at 0.05; in each set-up there were 3000 replications (30 per region); the GWAS sample size was $n = 1000$.

| $\beta_1$ | $n_{\text{ref}}$ | $H_0: \beta_1 = 0$ | | | $H_0: \boldsymbol{\beta} = \mathbf{0}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $T = 1$ | $T = 5$ | $T = 10$ | $T = 1$ | $T = 5$ | $T = 10$ | $T = 30$ | $T = 50$ |
| 0 | 900 | 0.089 | 0.024 | 0.022 | 0.162 | 0.020 | 0.016 | 0.014 | 0.013 |
| | 500 | 0.107 | 0.022 | 0.020 | 0.197 | 0.022 | 0.014 | 0.012 | 0.011 |
| | 379 (1000G A) | 0.171 | 0.048 | 0.045 | 0.352 | 0.085 | 0.072 | 0.064 | 0.063 |
| | Ind | 0.053 | | | 0.055 | | | | |
| 0.001 | 900 | 0.580 | 0.579 | 0.577 | 0.827 | 0.742 | 0.736 | 0.732 | 0.733 |
| | 500 | 0.605 | 0.577 | 0.576 | 0.847 | 0.740 | 0.732 | 0.727 | 0.724 |
| | 379 (1000G A) | 0.596 | 0.597 | 0.589 | 0.822 | 0.759 | 0.745 | 0.738 | 0.735 |
| | Ind | 0.495 | | | 0.794 | | | | |

We did another simulation with 100 randomly selected regions, each including 5 to 37 SNPs. Most of the regions were larger than the region in Table 1. As shown in Table 4, again the naïve method could not control the type I error rate while the new method performed much better, though the new method became conservative as $T$ went up. A possible explanation is that the sample size needed to estimate the LD accurately for a larger number of SNPs should be larger. With relatively small reference samples, the estimation of the regression coefficients is unstable, leading to large $\mathbf{V}_b$ and thus less significant test statistics. Nevertheless, the performance improved as the reference sample size increased from 500 to 900 with little loss of power.

## LHS data

Next we applied the methods to the Lung Health Study (LHS) data with 4387 subjects and 5112 SNPs on chromosome 19, downloaded from the dbGaP database (accession: phs000335.v3.p2). Our trait of interest was forced expiratory volume (FEV) at the baseline, FEVAC112. First, to adjust for non-genetic covariates, we built a linear model: FEVAS112 ~ AGE + SEX + PACKYEAR. Then we treated the residuals as the quantitative trait Y for the SNPs. We obtained the summary statistics of the marginal effects for each individual SNP on Y after centering the data at 0.

After choosing 4132 subjects with complete outcomes and 5111 SNPs that were present in both the LHS and 1000G data, we tested each single SNP and found none of them marginally significant. Then we used a sliding window approach to test the association between the trait and the SNPs inside each sliding window in a joint linear models (with the trait versus multiple SNPs). In each window, we selected SNPs so that none of their pair-wise correlation absolute values was greater than 0.95. We used two window sizes of 20 and 50 with two moving step-sizes/gaps of 1 and 20 respectively.

For the global/overall testing, as shown in Table 5, the Wald test based on the individual-level data detected no significant association regardless of the window size and moving step size; in contrast, the naive method based on the summary statistics (T=1) reported many significant associations, which (or at least most of which) are most likely to be false positives. Our new method with T=30 or larger eliminated all the false positives. The QQ plot in Figure 1 also demonstrates the problem of the naive method with an inflation factor lambda=1.49, much larger than 1, while the new method might be a bit conservative with an inflation factor less than 1 (Devlin and Roeder 1999).

Similarly, Table 6 shows in the COJO analysis on the first SNP inside each window, the individual-level data-based method identified no significant association.

Again the naive method with summary statistics detected three significant ones, most likely false positives; two or all three could be eliminated by the new method.

**Table 5** Numbers of the significant sliding windows for global testing with the LHS data. The nominal significance was at 0.05 with the Bonferroni adjustment with a cutoff 0.05/#windows. The reference sample size was $n_{\text{ref}} = 379$ based on the 1000G A reference panel. $U$, $\xi$, $f$ and Gap were the total number of SNPs, window size, the number of windows and the moving-step/gap size, respectively.

| $U$ | $\xi$ | Gap | $f$ | Ind | $T = 1$ | $T = 5$ | $T = 10$ | $T = 20$ | $T = 30$ | $T = 50$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 5111 | 20 | 1 | 5092 | 0 | 20 | 4 | 0 | 0 | 0 | 0 |
| | | 20 | 255 | 0 | 2 | 1 | 0 | 0 | 0 | 0 |
| | 50 | 1 | 5062 | 0 | 103 | 51 | 17 | 4 | 0 | 0 |
| | | 50 | 102 | 0 | 7 | 3 | 2 | 1 | 0 | 0 |

**Figure 1** QQ plots for the LHS data. $\xi$ and the gap size are both 20.

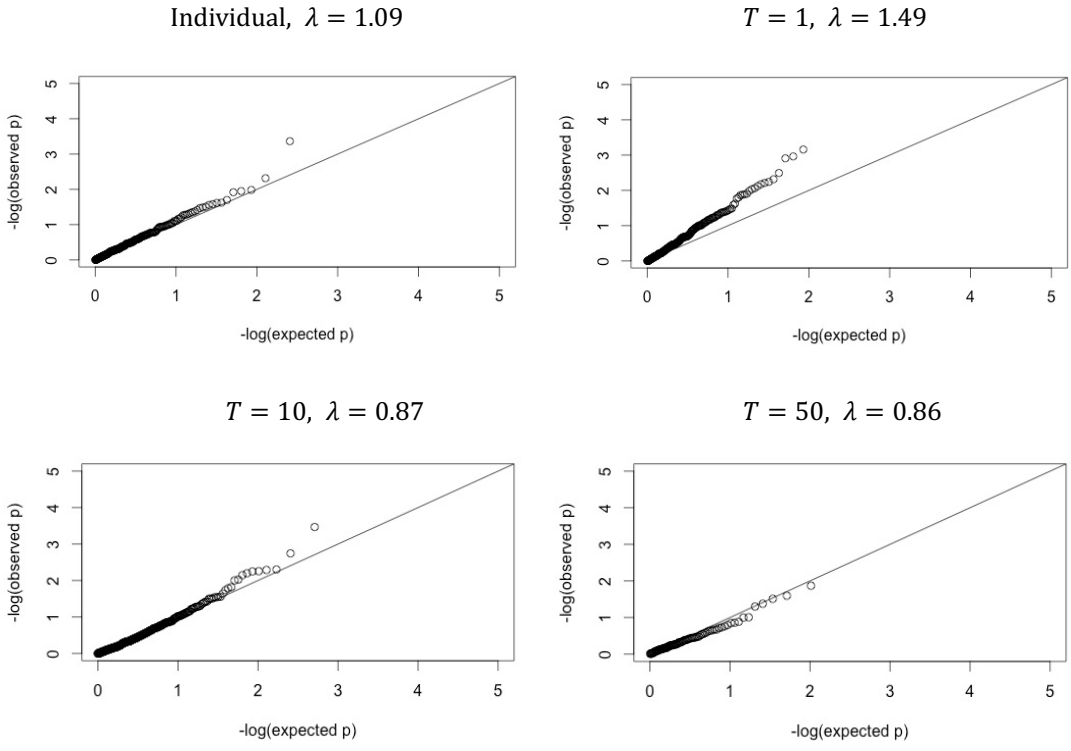

**Table 6** Numbers of the windows with the first SNP being significant in COJO analysis with the LHS data. The statistical significance cutoff was 0.05/#windows. $U$, $\xi$, $f$ and Gap were the total number of SNPs, window size, the number of windows and the moving-step/gap size, respectively.

| $U$ | $\xi$ | Gap | $f$ | Ind | T=1 | $T=5$ | $T=10$ | $T=30$ | $T=50$ |
|---|---|---|---|---|---|---|---|---|---|
| 5111 | 20 | 1 | 5092 | 0 | 1 | 0 | 0 | 0 | 0 |
| | | 20 | 255 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 50 | 1 | 5062 | 0 | 2 | 0 | 1 | 1 | 1 |
| | | 50 | 102 | 0 | 0 | 0 | 0 | 0 | 0 |

## Lipid data

We applied the methods to the 2010 and 2013 lipid data (Teslovich et al. 2010; Willer et al. 2013), testing the association between TG and SNPs on chromosome 19 that are present in both datasets. To save space, we only present the results for the 2013 lipid data in the following. We chose 7366 SNPs that were present in both the lipid data and the 1000 Genomes Phase 3 data with 503 subjects from the European population as the reference panel (denoted as 1000G B reference panel in the following), with minor allele frequencies larger than 0.01. First, we looked at the marginal p-values of each SNP, and found 86 of the 7366 SNPs with p-values less than 0.05/7366, and 911 with p-values less than 0.05. The estimated inflation factor was 1.0.

In addition to the 1000G B reference panel, we also used various subsets of the LHS data as a reference panel. We randomly sampled $n_{\text{ref}} = 1000\ to\ 4000$

subjects from the 4136 subjects in the LHS data as the reference data before applying the sliding window approach to the 4364 overlapping SNPs in the 2013 lipid data. As shown in Table 7 for global testing, as expected, the naive method gave much larger numbers of significant associations than that of the proposed new method, in which T=30 or larger seemed to give stable results. The same conclusion can be drawn for the COJO analysis as shown in Table 8. In summary, we expect that the naive method gave too many false positives.

**Table 7** Numbers of the significant sliding windows for global testing with the 2013 lipid data, using subsamples of the LHS data as reference with $n_{\text{ref}} \geq 1000$, or using the 1000G B data with $n_{\text{ref}} = 503$. The statistical significance cut-off was 0.05/#windows; $U$, $\xi$ and $f$ were the total number of SNPs, window size, and the number of windows respectively; the moving-step size or gap size was equal to the window size. The numbers of overlapping SNPs between T=1 and others are shown in the parentheses.

| Chr | $\xi$ | $n_{\text{ref}}$ | $U$ $(f)$ | $T = 1$ | $T = 5$ | $T = 10$ | $T = 30$ | $T = 50$ |
|---|---|---|---|---|---|---|---|---|
| 19 | 10 | 503 (1000G B) | 7366 (735) | 57 | 22 (22) | 20 (20) | 22 (22) | 22 (22) |
| | | 1000 | 4364 (435) | 38 | 19 (19) | 15 (15) | 14 (14) | 15 (15) |
| | | 2000 | | 37 | 19 (19) | 17 (17) | 15 (15) | 15 (15) |
| | | 4000 | | 40 | 17 (17) | 16 (16) | 15 (15) | 15 (15) |
| | 20 | 503 (1000G B) | 7366 (367) | 59 | 27 (27) | 20 (20) | 17 (17) | 17 (17) |
| | | 1000 | 4364 (217) | 35 | 23 (23) | 14 (14) | 10 (10) | 9 (9) |
| | | 2000 | | 35 | 17 (17) | 12 (12) | 10 (10) | 10 (10) |
| | | 4000 | | 33 | 18 (18) | 10 (10) | 11 (11) | 11 (11) |

**Table 8** Numbers of the windows with the first SNP being significant in COJO analysis with the 2013 lipid data, using subsamples of the LHS data as reference with $n_{\text{ref}} \geq 1000$, or using the 1000G B data with $n_{\text{ref}} = 503$. The statistical significance cut-off was 0.05/#windows; $U$, $\xi$ and $f$ were the total number of SNPs, window size, and the number of windows respectively; the moving-step size or gap size was equal to the window size. The numbers of overlapping SNPs between T=1 and others are shown in the parentheses.

| Chr | $U$ | $\xi$ $(f)$ | $n_{\text{ref}}$ | $T = 1$ | $T = 5$ | $T = 10$ | $T = 30$ | $T = 50$ | $T = 100$ |
|---|---|---|---|---|---|---|---|---|---|
| 19 | 4364 | 10 (435) | 1000 | 6 | 2 (2) | 1(1) | 1 (1) | 2 (2) | 2 (2) |
| | | | 2000 | 8 | 4 (4) | 2 (2) | 2 (2) | 3 (3) | 3 (3) |
| | | | 4000 | 8 | 2 (2) | 2 (2) | 2 (2) | 2 (2) | 2 (2) |
| | | 20 (217) | 1000 | 6 | 3 (3) | 3 (3) | 1 (1) | 2 (2) | 2 (2) |
| | | | 2000 | 6 | 4 (4) | 3 (3) | 3 (3) | 3 (3) | 3 (3) |
| | | | 4000 | 6 | 2 (2) | 2 (2) | 3 (3) | 2 (2) | 2 (2) |

# Discussion

Using simulated and real data, we have convincingly shown the severe problem of inflated type I error rates in integrating GWAS summary data with small reference panels for joint and conditional analyses, which have been widely

applied in the last few years, ranging from gene-based testing with one or more traits (Kwak and Pan 2016, 2017; Deng and Pan 2017) to fine mapping. In particular, as a gene-based testing approach to integrating eQTL data with GWAS summary data, the recently proposed transcriptome-wide association studies (TWAS) are expected to share the same problem with small reference panels (Gamazon et al 2015; Gusev et al 2016; Xu et al 2017). We emphasize that, although we have focused on conditional and global testing on a group of SNPs, the same issue of using small reference panels persists in many new and existing applications: to name a few, fine mapping (Benner et al 2017), polygenic risk prediction (Vilhjalmsson et al. 2015), inferring genetic correlations among complex traits (Bulik-Sullivan et al. 2015), and Mendelian randomization for causal inference (Burgess et al. 2013). Although standard reference panel samples, as for the 1000G data, are continuing growing with increasing sample sizes, the current and almost exclusive use of the popular 1000G reference panels is expected to suffer from the small sample issue as demonstrated here. Furthermore, even with a larger reference panel, if a GWAS sample size is larger (Benner et al 2017) or if we expand the SNPs to be tested to cover less frequent or rare ones and/or with those in high LD, as in fine mapping with sequencing data, the problem may still arise. Our proposed method, or its idea, could be applied (possibly after suitable modifications) to at least check whether the problem is severe in a given situation. Finally, we note that it is unclear how to deal with the problem if there are genotypic discrepancies between the reference panel and the GWAS data, which may happen in practice, especially with meta-analyzed GWAS summary statistics with multiple racial/ethnic subpopulations, for which any reference sample from a single population may not suffice (for the mixed GWAS population). In this case, perhaps the most straightforward solution is to conserve and share the LD structure from the original GWAS data. This problem is similar to meta-analysis of rare variants with sequencing data (Lee et al 2013). We hope that this study, along with Benner et al (2017), will raise the awareness of and attention to this important and urgent problem in light of the increasing use of GWAS summary data and (small) reference panels.

## Acknowledgment

The authors thank the reviewers for many helpful comments, and thank Chong Wu for help with the data. This research was supported by NIH grants R21AG057038, R01HL116720, R01GM113250 and R01HL105397, by NSF grant DMS1711226, and by the Minnesota Supercomputing Institute.

## Literature Cited

Benner, C., A. S., Havulinna, M. R., Järvelin, V., Salomaa, S., Ripatt, *et al.*, 2017 Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am J Hum Genet*. 2017 Oct 5;101(4):539-551.

Bulik-Sullivan, B., H. K., Finucane, V., Anttila, A., Gusev, F. R., Day, *et al.*, 2015 An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47:** 1236–1241.

Burgess, S., A., Butterworth, and S. G., Thompson, 2013 Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37:** 658–665.

Burton, P. R., D. G., Clayton, L. R., Cardon, N., Craddock, P., Deloukas, *et al.,* 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661-678.

Chen, W., B. R., Larrabee, I. G., Ovsyannikova, R. B., Kennedy, I. H., Haralambieva, *et al.,* 2015 Fine mapping causal variants with an approximate bayesian

method using marginal test statistics. *Genetics* **200:** 719–736.

Deng, Y., and W., Pan, 2017 Conditional analysis of multiple quantitative traits based on marginal GWAS summary statistics. *Genet Epidemiol*. **41**: 427-436.

Devlin, B., and K., Roeder, 1999 Genomic control for association studies. *Biometrics* **55:** 997-1004.

Gamazon, E.R., H. E., Wheeler, K. P., Shah, S. V., Mozaffari, K., Aquino-Michaels, *et al,.* 2015 A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47:** 1091-1098.

Gusev, A., A., Ko, H., Shi, G., Bhatia, W., Chung, *et al.,* 2016 Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet.* **48:** 245-252.

Hormozdiari, F., E., Kostem, E. Y., Kang, B., Pasaniuc, and E., Eskin, 2014 Identifying causal variants at loci with multiple signals of association. *Genetics* **198:** 497–508.

Kichaev, G., W. Y., Yang, S., Lindstrom, F., Hormozdiari, E., Eskin, *et al.,* 2014 Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10:** e1004722.

Kichaev, G., and B., Pasaniuc, 2015 Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97:** 260-271.

Kim, J., Y., Bai, and W., Pan, 2015 An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genet Epidemiol* **39**: 651-663.

Kim, J., Y., Zhang, W., Pan, and Alzheimer's Disease Neuroimaging Initiative, 2016 Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. *Genetics* **203**:715-731.

Kwak, I., and W., Pan, 2016 Adaptive gene- and pathway-trait association testing with gwas summary statistics. *Bioinformatics* **32**:1178-1184.

Kwak, I., and W., Pan, 2017 Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics* **33**: 64-71.

Lee, S., T. M., Teslovich, M., Boehnke, and X., Lin, 2013 General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. *American Journal of Human Genetics* **93:** 42-53.

Li, M.-X., J. S. H, Kwan, and P. C., Sham, 2012 HYST: A Hybrid Set-Based Test for Genome-wide Association Studies, with Application to Protein-Protein Interaction-Based Association Analysis. *American Journal of Human Genetics* **91:** 478–488.

Pasaniuc, B., N., Zaitlen, H., Shi, G., Bhatia, A., Gusev, *et al.,* 2014 Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30:** 2906-2914.

Pasaniuc, B., and A. L., Price, 2017 Dissecting the genetics of complex traits using summary association statistics. *Nature Reviews. Genetics **18*:** 117–127.

Rubin, D., 1996 Multiple Imputation After 18 Years. *Journal of the American Statistical Association **91*:** 473-489.

Shi, H., G., Kichaev, and B., Pasaniuc, 2016 Contrasting the genetic architecture of 30 complex traits from summary association data. *Am. J. Hum. Genet.* **99:** 139–153.

Teslovich, T. M., K., Musunuru, A. V., Smith, A. C., Edmondson, I. M., Stylianou, *et al.*, 2010 Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**: 707-13.

The 1000 Genomes Project Consortium, 2015 A global reference for human genetic variation, *Nature* **526**: 68-74.

Vilhjalmsson, B. J., J., Yang, H. K., Finucane, A., Gusev, S., Lindstrom, *et al.*, 2015 Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97:** 576–592.

Willer, C. J., E. M., Schmidt, S., Sengupta, G. M., Peloso, S., Gustafsson, *et al.*, 2013 Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45:** 1274-1283.

Xu, Z., C., Wu, P., Wei, W., Pan, 2017 A Powerful Framework for Integrating eQTL and GWAS Summary Data. *Genetics* **207:** 893-902.

Zhu, X., T., Feng, B. O., Tayo, J., Liang, J. H., Young, *et al.*, 2015 Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet.* **96:**21-36.