Improved analyses of GWAS summary statistics by reducing data heterogeneity and errors

Wenhan Chen[1], Jian Yang[1,2]

[1] Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD, Australia
[2] School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China

Introduction

Summary statistics from genome-wide association studies (GWAS) have facilitated the development of various summary data-based methods, which typically require a reference sample for linkage disequilibrium (LD) estimation. Analyses using these methods may be biased by errors in GWAS summary data and heterogeneity between GWAS and LD reference. Here we propose a quality control method, DENTIST, that leverages LD among genetic variants to detect and eliminate errors in GWAS or LD reference and heterogeneity between the two. Through simulations, we demonstrate that DENTIST substantially reduces false-positive rate (FPR) in detecting secondary signals in the summary-data-based conditional and joint (COJO) association analysis, especially for imputed rare variants. We further show that DENTIST can improve other summary-data-based analyses such as LD score regression analysis, and integrative analysis of GWAS and expression

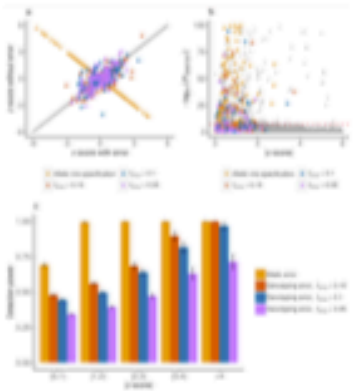


Figure 1: Detecting simulated errors using DENTIST. We simulated genotyping and allelic errors at 0.5% randomly selected variants respectively. Genotyping errors were simulated by altering the genotypes of a certain proportion (*F* = 0.05, 0.1 or 0.15) of randomly selected individuals, and allelic error of each of the variants was introduced by swapping the effect allele by the other allele.

Method

In brief, we first use a sliding window approach to divide the variants into 2Mb segments with a 500kb overlap between two adjacent segments. Within each segment, we randomly partition variants into two subsets, S1 and S2, with an equal number of variants, and apply the statistic below to test the difference between the observed z-score of a variant *i* (z_i) in S1 and its predicted value based on z-scores of an array of variants *t* in S2

$$T_{d(i)} = \frac{(z_i - \hat{z}_i)^2}{1 - \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\mathbf{R}_{it}^{'}} \text{ with } \hat{z}_i = \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\ z_t$$

R is the LD correlation matrix calculated from a reference sample with $\mathbf{R}_{tt}$ to denote the LD between variants *t* and $\mathbf{R}_{it}$ to denote the LD between variant i and variants *t*. Td follows approximately a χ^2 distribution with 1 degree of freedom. The performance of this test was evaluated based on detection of simulated errors as shown in Fig1 above.

Results 1

Applying DENTIST to COJO

1. **Simulation setting**: We simulated a phenotype affected by one or two sequenced variants using WGS data (i.e., UK10K-WGS) and performed association analyses using imputed data of the same individuals (imputing variants, in common with those on an SNP array, to the 1KGP; denoted by UK10K-1KG). More specifically, we first randomly selected one or two variants from two MAF bins as causal variants, i.e., variants with MAF≥0.01and 0.01>MAF≥0.001) to generate a phenotype with q^2 = 2%.We we ran a GWAS using UK10K-1KGP and performed COJO analyses using multiple LD references, including the discovery GWAS sample, UKB-8K-1KGP, HRS, and ARIC cohorts..
2. **Summary**: The between-sample data heterogeneity leads to over-estimation of the number of COJO signals.. This can be effectively QCed by DENTIST

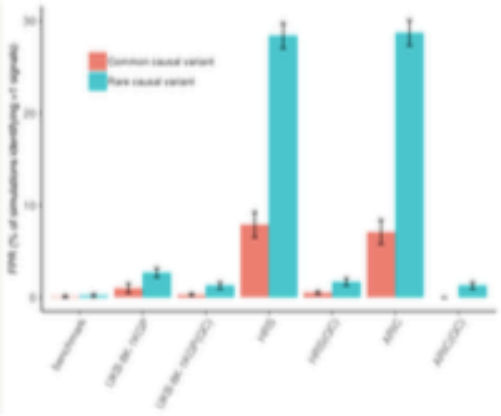


Figure 2: FPRs of COJO with and without DENTIST. Based on simulations with one causal signal, we assessed the FPRs of COJO analyses when performed with and without DENTIST-based QC (FPR is defined as the frequency of observing more than one COJO signals in the scenario in which only one causal variant was simulated). The x-axis labels indicate the LD reference samples used in the COJO analyses, and those performed after DENTIST QC are labeled with "QC" in the parentheses. The error bars correspond to the standard error of FPRs calculated from 2200 replications, each with a re-sampled causal variant.

Results 2

Applying DENTIST to SMR

1. **Simulation setting**: We first generated a trait based on a causal variant (q^2=1%) randomly sampled from variants in the ARIC data. To simulate a pleiotropic model, we used the same causal variant to simulate the a gene expression level using HRS data with q^2 for the expression level randomly sampled from the eQTL q^2 distribution reported by CAGE. To simulate a linkage model, a second causal variant in LD ($r^2 > 0.25$) with the trait causal variant was selected to generate the gene expression level, again with the eQTL q^2 value sampled from CAGE. In addition to the two-sample scenario above, we simulated a one-sample scenario in which both the trait and gene expression level were generated using HRS. The UKB sample was used as the LD reference for both SMR HEIDI and DENTIST analyses.

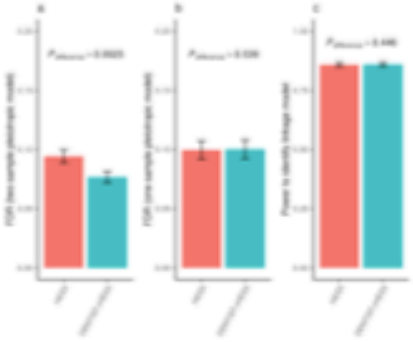
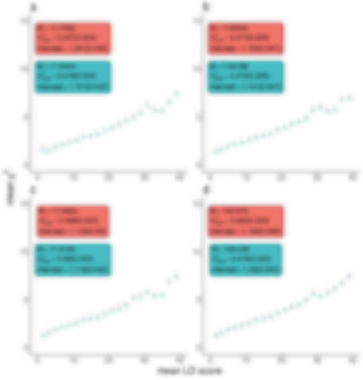


Figure 3: Shown are the results from simulations to quantify the FPR of HEIDI under a pleiotropic model (panels a and b) and the power of HEIDI under a linkage model (panel c). In both scenarios, an independent sample (UK10K-1KGP) was used as the LD reference..

Results 3

Applying DENTIST to LDSC

We assessed the effect of DENTIST on LDSC when different LD references were used, including a) HRS, b) ARIC, c) UKB-8K-1KGP, and d) UK10K-WGS. For demonstration, the following analyses were based GWAS of UKB height. The variants are binned by their LD scores. Each dot on the plots represents the mean LD score value of each bin on the x-axis and the mean χ^2 value on the y-axis, with those before and after DENTIST-based QC in red and cyan colors respectively. In the textbox, "M" represents the number of variants, "h^2_{SNP}" represents the estimate of SNP-based heritability, and "intercept" represents the LDSC intercept.



Conclusions

1. Our results suggest that summary-data-based analyses are generally well calibrated in the absence of data heterogeneity but biased otherwise.
2. DENTIST-based QC can substantially mitigate the biases for different tools tested and in no cases DENTIST degraded the results.

uqwche11@uq.edu.au
Wenhan Chen

DENTIST tool to QC for GWAS

Acknowledgements

Many and many thanks for my brilliant supervisor Jian Yang, co-authors of this paper and wonderful colleagues from the Group PCTG.

The University Of Queensland
AUSTRALIA
CREATE CHANGE