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

Wenhan Chen[1], Yang Wu[1], Zhili Zheng[1], Ting Qi[1,2], Peter M Visscher[1], Zhihong Zhu[1], Jian Yang[1,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia

[2]Present address: School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China

[*]Correspondence: Jian Yang (jian.yang.qt@gmail.com)

Abstract

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 (FPR reduced from >28% to <2% in the presence of ancestral difference between GWAS and LD reference). 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 quantitative trait locus data.

Introduction

Genome-wide association studies (GWASs) have been extraordinarily successful in uncovering genetic variants associated with complex human traits and diseases[1,2]. Summary statistics available from GWASs have facilitated the development of various summary-data-based methods[3] such as those for fine-mapping[4-9], imputing summary statistics at untyped variants[10,11], estimating SNP-based heritability[12-14], assessing causal or genetic relationship between traits[15-17], prioritizing candidate causal genes for a trait[18-21], and polygenetic risk prediction[8,22,23]. Most of the summary-data-based methods require linkage disequilibrium (LD) structure of the variants used, which are not available in the summary data but can be estimated from a reference cohort with individual-level genotypes assuming a homogeneous LD structure between the GWAS and reference cohorts. Hence, summary-data-based analyses can be affected by not only errors in the GWAS and LD reference data sets but also differences between them for the following reasons. First, there are often errors in GWAS summary statistics resulting from the data generation and analysis processes (e.g., genotyping/imputation errors and genetic variants with mis-specified effect alleles)[24,25], some of which are not easy to detect, even if individual-level data are available. Second, there is often heterogeneity between data sets (e.g., between the discovery GWAS and LD reference) because of differences in ancestry, and genotyping platform, analysis pipeline. Although the recommended practice is to use an ancestry-matched reference cohort, samples with similar ancestries, such as populations of European ancestry, can still have discernable differences in LD structure[26], and the effects of such differences on summary-data-based analyses are largely unexplored. To the best of our knowledge, there is no existing method specifically designed to detect data heterogeneity that affect summary data-based analyses.

In this study, we propose a quality control (QC) method to identify errors in GWAS summary data and heterogeneity between summary data and LD reference by testing the difference between the observed z-score of each variant and its predicted value from the surrounding variants. The method has been implemented in a software tool named DENTIST (detecting errors in analyses of summary statistics). We show by simulation that DENTIST can effectively detect simulated errors of several kinds. We then demonstrate the utility of DENTIST as a QC step for multiple, frequently-used, summary data-based methods, including the conditional and joint analysis (COJO)[6] of summary statistics, LD score regression[12], and heterogeneity in dependent instruments (HEIDI) test[21].

Results

Overview of the DENTIST method

Details of the methodology can be found in the Methods section. 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 ($\tilde{z}_i$) based on z-scores of an array of variants $\boldsymbol{t}$ in S2 (**Methods**).

$$T_{d(i)} = \frac{(z_i - \tilde{z}_i)^2}{\mathbf{1} - \mathbf{R}_{i\boldsymbol{t}}\mathbf{R}_{\boldsymbol{tt}}^{-1}\mathbf{R}_{i\boldsymbol{t}}'} \text{ with } \tilde{z}_i = \mathbf{R}_{i\boldsymbol{t}}\mathbf{R}_{\boldsymbol{tt}}^{-1}\boldsymbol{z_t} \quad (1)$$

where $\boldsymbol{z_t}$ is a vector of z-scores of variants $\boldsymbol{t}$ in S2, and $\mathbf{R}$ is the LD correlation matrix calculated from a reference sample with $\mathbf{R}_{\boldsymbol{tt}}$ to denote the LD between variants $\boldsymbol{t}$ and $\mathbf{R}_{i\boldsymbol{t}}$ to denote the LD between variant i and variants $\boldsymbol{t}$. T_d follows approximately a χ^2 distribution with 1 degree of freedom. Note that methods that leverage LD to predict GWAS test-statistic of a variant (i.e., $\tilde{z}_i$) from test-statistics of its adjacent variants (i.e., $\boldsymbol{z_t}$) have been developed in prior work[10,11]. A significant difference between the observed and predicted z-scores indicates errors in the discovery GWAS or LD reference, or heterogeneity between them. If the difference between z_i and $\tilde{z}_i$ is due to error in z_i, the power of T_d depends on how z_i deviates from its true value and how well variant i is tagged by variants $\boldsymbol{t}$. We conduct a truncated singular value decomposition (SVD) on $\mathbf{R}_{\boldsymbol{tt}}$ to mitigate the sampling noise in LD estimated from the reference that is often independent from the discovery GWAS and to perform a pseudo inverse when $\mathbf{R}_{\boldsymbol{tt}}$ is singular[13] (see **Methods**).

One challenge for the DENTIST method is that errors can be present in both S1 and S2, and errors in S2 can inflate T_d statistics of the variants in S1. To mitigate this issue, we propose an iterative partitioning approach. In each iteration, we partition the variants at random into two subsets (S1 and S2) and remove variants with $P_{DENTIST} < 5\times10^{-8}$ (capped at 0.5% variants with the smallest P-values). This step is to create a more reliable set of variants for the next iteration. The problematic variants are prioritized and filtered out in the first few iterations so that the prediction of $\tilde{z}_i$ (Equation 1) becomes more accurate in the following iterations. We set the number of iterations to 10 in practice. All variants with $P_{DENTIST} < 5\times10^{-8}$ are removed in the final step.

Detecting simulated errors in GWAS data

To assess the performance of DENTIST in detecting errors, we simulated GWAS data with genotyping errors and allelic errors (i.e., variants with the effect allele mis-labelled) using whole genome sequence (WGS) data of chromosome 22 on 3,642 unrelated individuals from the UK10K project[27,28] (denoted by UK10K-WGS). A descriptive summary of all the data sets used in

this study can be found in **Supplementary Table 1** and the Methods section. We simulated a trait affected by 50 common, causal variants with effects drawn from $N(0,1)$, which together explained 20% of the phenotypic variation (proportion of variance explained by a causal variant, denoted by q^2, was 0.4%, on average). Prior to the simulations with errors, we showed by a simulation under the null (i.e., simulating a scenario without errors and applying DENTIST using the discovery GWAS as the reference) that the DENTIST test-statistics were well calibrated, meaning that DENTIST will only remove a very small proportion of variants if there are no errors and heterogeneity in the data (**Supplementary Figure 1**). We then simulated genotyping and allelic errors at 0.5% randomly selected variants respectively. Genotyping errors of each of these variants were simulated by altering the genotypes of a certain proportion (f_{error} = 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. The simulation was repeated 200 times with the causal and erroneous variants re-sampled in each simulation. We then ran DENTIST using UK10K-WGS or an independent sample (UKB-8K-1KGP) as the LD reference after standard QCs of the discovery GWAS: removing variants with a Hardy-Weinberg Equilibrium (HWE) P-value < 10^{-6} using the individual-level data or ΔAF > 0.1 with ΔAF being the difference in allele frequency (AF) between the summary data and reference sample. The independent sample UKB-8K-1KGP is referred to as a set of 8000 unrelated individuals from the UK Biobank[29] (UKB) with variants imputed from the 1000 Genomes Project (1KGP). The statistical power (sensitivity) was measured by the proportion of erroneous variants in the data that can be detected from QC. We also computed the fold enrichment in probability of an erroneous variant being detected from QC compared to a random guess (i.e., the ratio of the percentage of true erroneous variants in the variants detected by DENTIST to that in all variants).

When using UKB-8K-1KGP as the reference, ~45% of the genotyping and ~95% of the allelic errors could be removed by the standard QCs (**Supplementary Table 2**). However, the ΔAF approach performed poorly for the very common variants, e.g., the power was ~16% for variants with MAF > 0.45. After the standard QCs, DENTIST was able to detect ~42% of the remaining genotyping and ~78% of the remaining allelic errors (**Figure 1** and **Supplementary Table 3**), with only ~0.3% variants being removed in total (**Supplementary Table 4**). The fold enrichment was 212 for allelic errors and of 112 for genotyping errors (**Supplementary Table 5**), showing good specificity of DENTIST in detecting the simulated errors. Notably, the power to detect allelic errors was ~78% for variants with MAF > 0.45, compensating the low power of the ΔAF approach in this MAF range (note: another shortcoming of the ΔAF approach is that the threshold is heavily sample size dependent, and currently there is no consensus guidance on the choice of a ΔAF threshold in practice). When restricted to variants passing the genome-wide

significance level (i.e., $p < 5\times10^{-8}$), the DENTIST detection power increased to ~87% for the genotyping errors and ~84% for the allelic errors (**Supplementary Table 3**). The power also varied with the genotyping error rate (f_{error}), e.g., the power in the f_{error}=0.05 scenario was, on average, lower than that for f_{error}=0.15 (**Figure 1c** and **Supplementary Table 3**). When using UK10K-WGS as the reference (mimicking the application of DENTIST in a scenario where individual-level data of the discovery GWAS are available), the power remained similar, but the fold enrichment was much higher compared to that using UKB-8K-1KGP (**Supplementary Tables 5** and **6**). In addition, using this same simulation setting, we explored the choice of the parameter θ_k (i.e., the proportion of eigenvectors retained in SVD; see Methods for details) and reference sample size (n_{ref}), and the results suggested a choice of θ_k=0.5 and $n_{\text{ref}} \geq 5000$ in practice (**Supplementary Figure 2**). Together, these results demonstrate the power of DENTIST to identify allelic and genotyping errors even after the standard QCs, suggesting that DENTIST can complement existing QC filters for either individual- or summary-level GWAS data. On the other hand, DENTIST was parsimonious in data filtering, with ~0.3% of the variants being removed in total across all the simulation scenarios (**Supplementary Table 4**). Nevertheless, we acknowledge that this simulation did not cover the full complexity of real case scenarios, which may involve multiple independent samples with heterogeneous LD structures caused by several factors, such as imputation errors or ancestry mismatches (**Supplementary Figure 3**). These cases are difficult to mimic in this simulation but will be assessed in the following analyses.

Applying DENTIST to COJO with simulated phenotypes

COJO[6] is a method that uses summary data from a GWAS or meta-analysis and LD data from a reference sample to run a conditional and joint multi-SNP regression analysis. We used simulations to assess the performance of COJO in the presence of heterogeneity between discovery GWAS and LD reference before and after DENTIST filtering. To mimic the reality that causal signals are often not perfectly captured by imputed variants, 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 312,264 variants, in common with those on an SNP array, to the 1KGP[28,30]; 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≥0.01 (denoted by common-causal) and 0.01>MAF≥0.001 (denoted by rare-causal) to generate a phenotype (note: MAF > 0.001 is equivalent to minor allele count > 7 in this sample). The causal variant q^2 was set to 2% to achieve similar power to a scenario with q^2 = 0.03% and n = 250,000 (note: the mean q^2 of 697 height variants discovered in Wood et al.[31] is 0.03%) because the power of GWAS is determined by $nq^2/(1-q^2)$. Then, we

ran a GWAS using UK10K-1KGP and performed COJO analyses using multiple LD references, including the discovery GWAS sample, UKB-8K-1KGP, the Health Retirement Study (HRS)[32], and the Atherosclerosis Risk In Communities (ARIC) study[33], with different degrees of ancestral differences with UK10K-1KGP (**Supplementary Figure 4**). For a fair comparison, only the variants shared between these reference samples were included. We repeated the simulation 100 times for each autosome and computed the false positive rate (FPR, i.e., the frequency of observing two COJO signals in the scenario where there was only one causal variant) and power (the frequency of observing two COJO signals in the scenario where there were two distinct causal variants with LD r^2 <0.1 between them). It should be noted that the false positive COJO signals defined here are not false associations but falsely claimed as jointly associated (also known as quasi-independent) signals.

When using the discovery GWAS sample as the LD reference, the FPR of COJO was 0.1% for common-causal and 0.2% for rare-causal (**Figure 2; Table 1**), which can be regarded as a baseline for comparison as there was no data heterogeneity in this case. The FPRs were higher than the expected values because the causal variants were not perfectly tagged by the imputed variants (**Supplementary Table 7**). When using UKB-8K-1KGP (i.e., 1KGP-imputed data of 8000 UKB participants with similar ancestry to the UK10K participants as shown in **Supplementary Figure 4**) as the LD reference, the FPR was close to the benchmark for common-causal (1%) and slightly inflated for rare-causal (2.7%) (**Figure 2**). After DENTIST filtering (using UKB-8K-1KGP as the LD reference), the FPR for rare-causal decreased to 1.3%. Moreover, when using LD computed from European-American individuals in HRS or ARIC, the FPR of COJO was strongly inflated in the whole MAF range: >7% for common-causal and >28% for rare-causal, likely because of the difference in ancestry between HRS/ARIC and UK10K-1KGP. DENTIST can effectively control the FPR of COJO to <1% for common-causal and <2% for rare-causal (**Figure 2**). Taken together, the FPR of COJO was reasonably well controlled for common variants but substantially inflated for rare variants especially when there was a difference in ancestry between the GWAS and LD reference samples, and most of the false positive COJO signals could be removed by DENTIST.

The power of COJO (without DENTIST) using in-sample LD from UK10K-1KGP or out-of-sample LD from the other references were similar: 77-81% for common-causal and 26-30% for rare-causal (**Table 1**). The low power for rare-causal was because they were poorly captured by imputation (**Supplementary Table 7**). DENTIST filtering caused a <2% loss of power for common-causal, and 5-10% for rare-causal (**Table 1**). Hence, the control of FPR of COJO by

DENTIST was to some extent at the expense of power although the reduction in FPR was larger than that in power.

We also examined the effect of imputation INFO score-based QC on the FPR and power of COJO. Take the analysis with HRS as an example. By removing variants with INFO-scores < 0.9 from the HRS data, the FPR of COJO decreased to 2.3% for common-causal and 6% for rare-causal (**Supplementary Table 8**), both of which were higher than those using DENTIST (FPR = 0.5% for common-causal and 1.7% for rare-causal) (**Table 1**). Meanwhile, the power of COJO after the INFO score-based QC decreased to 70% for common-causal and 12% for rare-causal (**Supplementary Table 8**), both of which were lower than those using DENTIST (power = 81% for common-causal and 27% rare-causal). The other less stringent INFO-score threshold were even less effective, and the results from analyses using the other references were similar (**Supplementary Table 8**). These results suggest that filtering variants by imputation INFO is less effectively than that by DENTIST.

Applying DENTIST to COJO for real phenotypes

Having assessed the performance of DENTIST in COJO analyses by simulation, we then applied it to COJO analyses for height in the UKB. The height GWAS summary statistics (n = 328,577) were generated from a GWAS analysis of all the unrelated individuals of European ancestry in the UKB (denoted by UKBv3-329K) except 20,000 individuals (denoted by UKBv3-20K), which were used as a non-overlapping LD reference. Genotype imputation of the UKB data was performed by the UKB team with most of the variants imputed from the Haplotype Reference Consortium (HRC)[34]. We performed COJO analyses with a host of references: overlapping in-sample references with sample sizes (n_{ref}) varying from 10,000 to 150,000, non-overlapping in-sample references including UKBv3-8K (n = 8,000) and UKBv3-20K (containing UKBv3-8k), and out-of-sample references including ARIC and HRS (**Supplementary Table 1**). We excluded from the analysis variants with MAF < 0.001 to ensure sufficient number of minor alleles for rare variants in reference samples with n_{ref} < 10k. We first performed a COJO analysis using the actual GWAS sample as the reference and identified 1,279 signals from variants with MAFs >0.01, and 1310 signals from variants with MAFs > 0.001 (**Table 2**). These results can be regarded as a benchmark. When using the overlapping in-sample LD references, the number of COJO signals first decreased as n_{ref} increased and then started to stabilize when n_{ref} exceeded 30,000 (**Supplementary Figure 5**). The results from using the two non-overlapping in-sample references (UKBv3-8K and UKBv3-20K) were comparable to those from using the overlapping in-sample references with similar sample sizes (**Table 2** and **Supplementary Table 9**) because the non-overlapping in-sample references, despite being excluded from the GWAS, were

consistent with the GWAS sample with respect to ancestry, data collection, and analysis procedures.

When using LD from an out-of-sample reference (either HRS or ARIC), there was substantial inflation in the number of COJO signals compared to the benchmark (by 15.5-16.1% for common variants and 18.7-25.6% for all variants), with a few variants in weak LD with those identified from the benchmark analysis (**Supplementary Figure 6**). The results from using the two out-of-sample references became more consistent with the benchmark after DENTIST filtering, with the inflation reduced to 4.6-5.8% for common variants and 2.7-6.7% for all variants, comparable to the results using an in-sample LD reference with a similar sample size (**Table 2**). Polygenic score analysis shows that the reduction in the number of COJO signals owing to DENTIST QC had almost no effect on the accuracy of using the COJO signals to predict height in HRS (**Supplementary Table 10**), suggesting the redundancy of the COJO signals removed by DENTIST. We further found that compared to using the imputed data from ARIC or HRS, using UK10K-WGS (n=3,642) as the reference showed lower inflation (10% for common variants and <12.4% for all variants) before DENTIST QC but larger inflation after DENTIST QC (**Table 2**), suggesting a large reference sample size is essential even for WGS data. In all the DENTIST analyses above, the total number of removed variants varied from 0.05% to 0.98% (**Supplementary Table 11**). All these results are consistent with what we observed from simulations, demonstrating the effectiveness of DENTIST in eliminating heterogeneity between GWAS and LD reference samples.

We further applied DENTIST to Educational Attainment (EA), Coronary Artery Disease (CAD), Type 2 Diabetes(T2D), Crohn's Disease (CD), Major Depressive Disorder (MDD), Schizophrenia (SCZ), Ovarian Cancer (OC), Breast Cancer (BC), Height and Body Mass Index (BMI) using GWAS summary data from the public domain[35-43] (**Supplementary Table 12**) and three LD reference samples (i.e., ARIC, HRS, and UKBv3-8K). Since these published studies focus on common variants (rare variants are not available in most of the data sets), we used a MAF threshold of 0.01 in this analysis. When using ARIC as the LD reference, the proportion of variants removed by DENTIST QC ranged from 0.02% (BMI) to 0.94% (CAD) with a median of 0.28% (**Supplementary Table 13**), and the reduction in the number of COJO signals for common variants ranged from 0% (OC and MDD) to 11.9% (CAD) with a median of 1.5% (**Supplementary Table 14**). The results from using the other two references are similar (**Supplementary Table 13** and **14**).

Improved HEIDI test with DENTIST

The summary data-based Mendelian randomization (SMR) is a method that integrates summary-level data from a GWAS and an expression quantitative trait loci (eQTL) study to test pleiotropic associations between a trait and expression levels of genes[21]. It features the HEIDI test that utilizes multiple cis-eQTL variants at a locus to distinguish pleiotropy (the trait and expression level of a gene are affected by the same causal variants) from linkage (causal variants for the trait are in LD with a distinct set of causal variants affecting gene expression). The HEIDI test uses summary data from two studies and LD from a reference so that any errors in and heterogeneity between the GWAS, eQTL and reference samples can cause inflated HEIDI test statistics, giving rise to more SMR associations being rejected than expected by chance[21,44]. Here, we performed simulations to assess the effect of data heterogeneity on HEIDI and sought to mitigate it using DENTIST. We first generated a trait based on a causal variant (q^2 = 1%) randomly sampled from the variants on chromosome 22 in the ARIC data. To simulate a pleiotropic model, we used the same causal variant to simulate the expression level of a gene in a subset of the HRS data (n = 3,000; denoted by HRS-3K) with q^2 for the gene expression level randomly sampled from the eQTL q^2 distribution reported by the Consortium for the Architecture of Gene Expression (CAGE)[45]. 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 the CAGE. In addition to the two-sample scenario above, we also simulated a one-sample scenario in which both the trait and gene expression level were generated in the HRS-3K sample. The UKB-8K-1KGP sample was used as the LD reference for both the SMR-HEIDI and DENTIST analyses. For each scenario, the simulation was repeated 4000 times with the causal variants re-sampled in each replicate. The FPR of the HEIDI test was calculated as the proportion of pleiotropic models detected with $P_{\text{HEIDI}} < 0.05$, and the power was defined as the proportion of linkage models detected with $P_{\text{HEIDI}} < 0.05$.

We found that the FPR of HEIDI was close to the expected value (5%) in the one-sample scenario (5.8%) but inflated (9.8%) in the two-sample scenario (**Figures 3a** and **3b**, and **Supplementary Table 15**). To mitigate the inflation, we performed DENTIST in both the GWAS and eQTL summary data using UKB-8K-1KGP as the reference. After DENTIST filtering, the FPR of HEIDI in the two-sample scenario decreased to 7.6%; the decrease was small but statistically significant ($P_{difference}$ = 0.002). The results remained similar when the discovery GWAS sample (i.e., HRS-3K) was used as the reference (**Supplementary Table 15)**. These results suggest that the HEIDI test statistic was inflated in the two-sample scenario likely because of LD heterogeneity between the GWAS and eQTL samples. The power of HEIDI to detect the linkage model remained almost the same before and after DENTIST filtering (**Figure 3c** and **Supplementary Table 15**). To further validate if the inflation of HEIDI FPR was due to heterogeneity between the GWAS and eQTL

samples, we increased the difference in ancestry between the two discovery samples by simulating GWAS and eQTL data from UKB-8K-1KGP and HRS-3K, respectively, and performed the HEIDI analysis using ARIC as the reference. In this case, the FPR of HEIDI increased to 12.0% before and to 9.1% after DENTIST filtering (**Supplementary Table 15**). All these results suggest that DENTIST slightly improved the FPR of HEIDI in the presence of data heterogeneity at almost no expense of power and that in the presence of ancestry difference between the GWAS and eQTL samples, HEIDI tends to be conservative (rejecting more SMR associations than expected by chance) even after DENTIST filtering.

Improved LD score regression analysis with DENTIST

LD score regression (LDSC) is an approach which was originally developed to distinguish polygenicity from population stratification in GWAS summary data set by a weighted regression of GWAS χ^2 statistics against LD scores computed from a reference[12] but has often been used to estimate the SNP-based heritability (h^2_{SNP}). We investigated the impact of DENTIST on LDSC using the height GWAS summary data generated using the UKBv3-329K sample along with several reference samples including four imputation-based samples (i.e., the discovery GWAS sample, HRS, ARIC and UKB-8K-1KGP) and two WGS-based samples (i.e., UK10K-WGS and the European individuals from the 1KGP (1KGP-EUR)). We performed the one- and two-step LDSC analyses using LD scores of the variants, in common with those in the HapMap3, computed from each of the references before and after DENTIST-based QC. Note that DENTIST was performed for all common variants but only those overlapped with HapMap3 were included in the LDSC analyses.

Using the discovery GWAS sample as the reference, the estimates of h^2_{SNP} and regression intercept from the one-step LDSC were 46% (SE = 0.02) and 1.13 (SE = 0.04) respectively (**Table 3**). When using the other reference samples, the results were very close to the benchmark except for a noticeably larger estimate of regression intercept using HRS (1.24, SE = 0.04). After DENTIST filtering, the intercept estimate using HRS decreased to 1.15 (SE = 0.04) with little difference in $\hat{h}^2_{SNP}$ (increased from 45% to 46%) (**Table 3**). To better understand the effect of DENTIST QC on LDSC using HRS, we plotted the mean χ^2-statistic against the mean LD score across the LD score bins. We found that the GWAS mean χ^2-statistic in the bin with the smallest mean LD score deviated from the value expected from a linear relationship between the LD score and χ^2-statistic (**Figure 4a**), and the deviation was removed by filtering out a small proportion of variants with very small LD scores but large χ^2-statistic by DENTIST (**Supplementary Figure 7**). These results show how the quality of LD reference can impact the LDSC analysis, but such effect is typically unknown *a priori* and varied across different reference

samples. We also re-ran the analyses using the two-step LDSC, where the intercept was estimated using the variants with χ^2 values < 30 in the first step and constrained in the second step to estimate h^2_{SNP} using all the variants. Compared to the one-step approach, the two-step approach provides larger estimates of the intercepts and smaller estimates of h^2_{SNP} either before or after DENTIST filtering. It is noteworthy that when using 1KGP-EUR as the LD reference, DENTIST suggested many more variants for removal compared to that using the other references, which caused a substantially smaller estimate of h^2_{SNP} (**Table 3**). This is because LD correlations computed from references of small sample size are noisy due to sampling variation, which cause inflated test-statistic and thereby elevated FPR of DENTIST (**Supplementary Figure 2**). This result cautions the use of DENTIST with LD references with small sample sizes (e.g., $n < 5000$). In addition, we applied LDSC to the 10 published GWAS data sets mentioned above (**Supplementary Table 12**) using ARIC and UKBv3-8K as the reference. The improvement of LDSC by DENTIST QC was small particularly for traits whose LDSC intercepts were close to 1 before DENTIST QC (**Supplementary Table 16**), demonstrating the robustness of LDSC to data heterogeneity and errors.

Discussion

In this study, we developed DENTIST, an QC tool for summary data-based analyses, which leverages LD from a reference sample to detect and filter out problematic variants by testing the difference between the observed z-score of a variant and a predicted z-score from the neighboring variants. From simulations and real data analyses, we show that some of the commonly-used analyses including the COJO, SMR-HEIDI and LDSC, can be biased to various extents in the presence of data heterogeneity, e.g., inflated number of COJO signals, elevated rate of rejecting pleiotropic models for SMR-HEIDI, or biased estimates of regression intercept and h^2_{SNP} for LDSC. For most of these analyses, DENTIST-based QC can substantially mitigate the biases.

Our results suggest that summary-data-based analyses are generally well calibrated in the absence of data heterogeneity but biased otherwise. For example, we showed that the mismatch in ancestry between the discovery GWAS and LD reference (e.g., European vs. British ancestry) caused inflated FPRs of both COJO and SMR-HEIDI analyses (**Table 2** and **Supplementary Table 15**). Also, we found that the FPR of COJO for rare variants was much higher than that for common variants likely because rare variants are more difficult to impute such that they are more likely to have discrepancy in LD between two imputed data sets. It should be clarified that the false-positive COJO signals as defined here are not false associations but falsely claimed as jointly significant associations. DENTIST substantially reduced false-positive detections from

COJO analyses especially for rare variants even when there was a difference in ancestry between the GWAS and LD reference samples. This extends the utility of COJO, which was originally developed for common variants, to rare variants. This message is important for the field because more and more GWASs and meta-analyses have started included rare variants from imputation. The FPR of HEIDI was only marginally reduced by DENTIST but at almost no cost of power. It should also be clarified again that the inflated HEIDI test-statistics would not lead to false discoveries because higher HEIDI test-statistics correspond to higher probability of rejecting SMR associations rather than tending to claim more significant associations. Among all the methods tested, LDSC was least affected by errors or data heterogeneity, but in one case where HRS was used as the LD reference for the analysis of the UKB height summary data, the estimates were biased but could be corrected by DENTIST (**Figure 4a**). DENTIST has the unique feature to detect heterogeneity between a GWAS summary data set and an LD reference. The benefit of using DENTIST as a QC tool has been demonstrated in the three case studies above, but we believe that it can potentially be applied to all GWAS summary data-based analyses that require a LD reference such as fine mapping methods[7,9,46] and joint modeling of all variants for polygenic risk prediction[22,23,47] . We have also shown by simulation that DENTIST can even add value to the standard GWAS QC process in a single-cohort-based GWAS to detect genotyping/imputation errors.

Given that a QC step can potentially remove true signals, we make sure that DENTIST is conservative in filtering variants. We show by simulation that in the absence of errors and data heterogeneity, the DENTIST test-statistics were not inflated, and on average, only <0.05% variants were filtered out by DENTIST (**Supplementary Figure 1**). In practice, we implemented two strategies to avoid widespread inflation of the DENTIST statistics in the presence of data errors or heterogeneity: 1) we used the SVD truncation method to control for sampling variation in LD estimated from the reference; 2) we applied an iterative approach to prioritize the elimination of larger outliers in earlier iterations (Methods). We optimized the parameters related to these two steps through simulations (**Supplementary Figure 2)**. Throughout all the analyses performed in this study, we found in no cases DENTIST degraded the results, and DENTIST often only needed to remove a very small proportion of the variants to correct or alleviate the biases (**Table 3** and **Supplementary Tables 4, 11, 13**). To the contrary, filtering variants based on imputation INFO score[48] caused significant loss of power in the COJO analysis when a stringent INFO score threshold was used to achieve a similar level of FPR as that using DENTIST.

Our method is an early attempt at QC for summary-data-based analyses. To avoid misuse, we summarize the usages and limitations, in addition to the features mentioned above. Firstly, DENTIST is a QC method for detecting not only errors in summary-data but also heterogeneity between discovery and reference data. As shown from our simulation, DENTIST does not guarantee the filtering of all the errors but most of them with large GWAS z-scores and a large proportion of them with small z-scores. Secondly, DENTIST can identify errors that passed the standard QC approaches (such as HWE test and allelic frequency checking), which makes it a good complementary method to existing QC filters. We suggest that DENTIST-based QC should be applied after the standard QC as DENTIST is more powerful when the proportion of errors is smaller (**Supplementary Table 3**). DENTIST can also be used as a method for checking summary data sanity by running it with a reliable LD reference sample. Thirdly, regarding the choice of a reference sample, DENTIST expects unrelated individuals from a closely matched ancestry, with a large sample size ($n > 5{,}000$). From simulations, we found that small reference sample size biased the DENTIST test statistics leading to significantly elevated FPR (**Supplementary Figure 2**). Lastly, DENTIST assumes the test statistics of different variants have similar sample sizes; violation of this assumption will lead to variants with significantly smaller or larger sample sizes being mistakenly recognized as problematic variants by DENTIST.

In summary, we have proposed a new QC approach to improve summary-data-based analyses that are potentially affected by the errors in summary data or heterogeneity between data sets. This method has been implemented in a user-friendly software tool DENTIST. The software tool is multi-threaded so that it is computationally efficient when enough computing resources are available. For example, when running each chromosome in parallel, it took <1h to run DENTIST for all variants with MAF > 1% and <5h for all variants with MAF > 0.01% on each chromosome (**Supplementary Table 16**).

Methods

The DENTIST test-statistic

Given an ancestrally homogeneous sample and a genotype matrix **X** consisting of *n* unrelated individuals genotyped/imputed at *m* variants, an association study is carried out at each variant by performing a linear regression between the variant and a phenotype of interest. This provides a set of summary data that include the estimate of variant effect, the corresponding standard error, and thereby the z-statistic. Under the null hypothesis of no association, the z-scores of *m* variants follow a multivariate normal distribution, $\boldsymbol{Z} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{Z} = (Z_1, Z_2, \dots, Z_m)$, with $\boldsymbol{\Sigma}$ a LD correlation matrix of the variants.

The aim of this method is to test the difference between the z-statistic of a variant and that predicted from adjacent variants. To do this, we use a sliding window approach to divide genome into 2Mb segments with a 500kb overlap between one another and randomly partition the variants in a segment into two subsets, S1 and S2, with similar numbers of variants. We then use variants in S2 to predict those in S1 and vice versa. According to previous studies[10,11] , the distribution of z-statistic of a variant *i* from S1, conditional on the observed z-scores of a set of variants from S2 is

$$Z_i|\mathbf{Z_t} = \mathbf{z_t} \sim N(\boldsymbol{\Sigma}_{it}\boldsymbol{\Sigma}_{tt}^{-1}\,\mathbf{z}_t,\ \Sigma_{ii} - \boldsymbol{\Sigma}_{it}\boldsymbol{\Sigma}_{tt}^{-1}\boldsymbol{\Sigma}_{it}') \quad (1),$$

where $\boldsymbol{\Sigma}_{it}$ denotes the correlation of z-scores between variant *i* from S1 and variants $\boldsymbol{t}$ from S2, and $\boldsymbol{\Sigma}_{tt}$ is the correlation matrix of variants $\boldsymbol{t}$. We use the correlation matrix calculated from an ancestry-matched reference sample (denoted by **R**) to replace that in the discovery sample if individual-level genotypes of in the discovery GWAS are unavailable. In this case, **Equation 1** can be rewritten as

$$Z_i|\mathbf{z_t} \sim N(\mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\,\boldsymbol{z}_t, \mathbf{1} - \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\mathbf{R}_{it}') \quad (2).$$

We can use $E(Z_i|\mathbf{z_t})$ as a predictor of Z_i, i.e., $\tilde{z}_i = \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\,\boldsymbol{z}_t$, and can therefore use the test-statistic below to test the difference between the observed and predicted z-scores

$$T_{d(i)} = (z_i - \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\,\boldsymbol{z}_t)^2/(\mathbf{1} - \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\mathbf{R}_{it}') \quad (3)$$

which approximately follows a χ^2 distribution with 1 degree of freedom. A deviation of $T_{d(i)}$ from χ_1^2 can be attributed to 1) errors in the summary data; 2) errors in the reference data; or 3) heterogeneity between the two data sets. Using **Equation 3**, the test statistic T_d can be calculated for each variant in S1 given z-scores from S2. As in previous studies[10,11], the method is derived under the null hypothesis of no association, but the test-statistics are well calibrated in the presence of true association signals (**Supplementary Figure 2**).

The iterative partitioning approach

One challenge of using **Equation 3** is that errors in $\mathbf{z}_t$ or discrepancy between $\mathbf{R}_{it}$ and $\mathbf{\Sigma}_{it}$ can affect the accuracy of predicting $\tilde{z}_i$. To mitigate this, we use an iterative partitioning approach. That is, in each iteration, we randomly partition the variants into two sets, S1 and S2, predict the z-statistic of each variant in S1 using its adjacent variants in S2 and vice versa, and run the T_d test to remove a small fraction of variants with $P_{\text{DENTIST}} < 5\times10^{-8}$ (capped at 0.5% variants with the smallest P_{DENTIST} if more than 0.5% of the variants exceeding this threshold). The default number of iterations is set to 10. In this iterative process, variants with very large errors or LD heterogeneity between the discovery and LD reference samples are prioritized for removal in the first few iterations so that the prediction accuracy increases in the following iterations. After the iterations are completed, any SNPs with $P_{\text{DENTIST}} < 5\times10^{-8}$ are removed in the final step.

Accounting for sampling noise in LD

A simple replacement of the LD correlation matrix $\mathbf{\Sigma}$ by $\mathbf{R}$ introduces additional noises, which can inflate T_d, because the sampling variations in $\mathbf{R}$ differ from those $\mathbf{\Sigma}$. Therefore, we adopt a truncated singular value decomposition (SVD) approach used in a previous study[13] to suppress the sampling noises. The essential idea was to remove variance components of $\mathbf{R}_{tt}$ that corresponded to the smallest singular values in SVD, as these variance components were likely to be induced by sampling noises. Given the equivalence between SVD and eigen decomposition of $\mathbf{R}_{tt}$, we perform pseudoinverse of $\mathbf{R}_{tt}$ using eigen decomposition, set small eigen values to 0, and retain only k components with large eigen values.

$$\mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\,\mathbf{z}_t = \mathbf{R}_{it}\mathbf{R}_{tt}^{+}\,\mathbf{z}_t = \sum_{\mathbf{1}..k} 1/w_k(\mathbf{R}_{it}\boldsymbol{v}_k)(\boldsymbol{v}_k'\mathbf{z}_t) \quad (4)$$

$$\mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\mathbf{R}_{it}' = \mathbf{R}_{it}\mathbf{R}_{tt}^{+}\mathbf{R}_{it}' = \sum_{\mathbf{1}..k} 1/w_k(\mathbf{R}_{it}\boldsymbol{v}_k)^2 \quad (5)$$

$\mathbf{R}_{tt}^{+}$ denotes the pseudo inversion of $\mathbf{R}_{tt}$. The scalars $w_1,...w_k$ correspond to the largest k eigenvalues, and vectors $\boldsymbol{v}_1...\boldsymbol{v}_k$ are the corresponding k eigenvectors. Given q = rank ($\mathbf{R}_{tt}$), the suggested value of k is $k \ll q$. Let $\theta_k = k/q$. We show by simulation that $\theta_k = 0.5$ appears to be a good choice meanwhile a large reference sample size (e.g., $n_{\text{ref}} \geq 5000$) is need (**Supplementary Figure 2**). This pseudoinverse also prevents the problem of rank deficiency due to strongly correlated variants when computing $\mathbf{R}_{tt}^{-1}$.

According to the term $\mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\,\mathbf{z}_t$ in **Equation 3**, which is a weighted sum of multiple z-scores, a variant displaying a strong correlation with i can overrule the information from the rest of the variants in S2. This would affect the robustness of our method. Therefore, we prune the variants for LD with an r^2 threshold of 0.95 (note: we do not actually remove variants in this case). For a set of variants in high LD ($r^2 > 0.95$), variants pruned out by this pruning process are assigned with the same T_d value as that of the variant retained.

Genotype data sets

This study is approved by the University of Queensland Human Research Ethics Committee (approval number: 2011001173). A summary of the genotype data sets used in this study as well as their relevant information can be found in **Supplementary Table 1**. These data are from four GWAS cohorts of European descendants, including the Health Retirement Study (HRS)[32], Atherosclerosis Risk in Communities (ARIC) study[33], UK10K[27], and UK Biobank (UKB)[29]. The samples were genotyped using either WGS or SNP array technology (**Supplementary Table 1**). Imputation of the UKB data had been performed in a previous study[49] using the Haplotype Reference Consortium (HRC)[34] and UK10K reference panels[29,50]. We used different subsets of the imputed UKB data as the LD reference in this study, denoted with the prefix "UKBv3", such as UKBv3-unrel (all the unrelated individuals of European ancestry, n = 348,577), UKBv3-329K (a subset of 328,577 individuals of UKBv3-unrel), UKBv3-20K (another subset of 20,000 individuals of UKBv3-unrel, independent of UKBv3-392K) and UKBv3-8K (a subset of 8,000 individuals of UKBv3-20K). HRS, ARIC and UK10K cohorts were imputed to the 1KGP reference panel in prior studies[28,30], and a subset of 8000 unrelated individuals from UKB were imputed to the 1KGP reference panel in this study (referred to as UKB-8K-1KGP). The UK10K variants in common with those on an Illumina CoreExome array were used for 1KGP imputation[30]. The imputation dosage values were converted to best-guess genotypes in all the data sets except for UKBv3-all, in which the hard-called genotypes were converted from the imputation dosage values using PLINK2 --hard-call-threshold 0.1 (Ref[51]). For all the data set, standard QCs were performed to remove variants with HWE test *P*-value < 10^{-6}, imputation INFO score <0.3, or MAF < 0.001. Since the hard-called genotypes had missing values, in UKBv3 and its subsets, we further removed variants with missingness rate > 0.05.

Genome-wide association analysis for height using the UKB data

We performed a genome-wide association analysis for height using the genotype data of UKBv3-329K, i.e., all the unrelated individuals of European ancestry in the UKB (n=328,577) except for 20000 individuals randomly selected to create a non-overlapping reference sample (i.e., UKBv3-20K). The height phenotype was pre-adjusted for sex and age. We conducted the association analysis using the simple linear regression model in fastGWA[52] with the first 10 principle components (PCs) fitted as covariates.

Data availability

All the data sets used in this study are available in the public domain (**Supplementary Table 12**).

Code availability

The software tool DENTIST was written in C++ as a command-line tool. The source code and pre-compiled executable for 64-bit Linux distributions are available at https://github.com/Yves-CHEN/DENTIST/.

Acknowledgements

We are very grateful for constructive comments from Naomi Wray, Loic Yengo, Ying Wang and Jian Zeng and technical supports from Allan McRae, Julia Sidorenko, and Futao Zhang. This research was supported by the Australian Research Council (FT180100186, FL180100072), the Australian National Health and Medical Research Council (1113400 1107258), and the Sylvia & Charles Viertel Charitable Foundation.

Author Contributions

JY conceived and supervised the study. WC, ZZh and JY developed the method. WC, YW, ZZh, and JY designed the experiment. WC performed the simulations and data analyses under the assistance and guidance from YW, ZZl, TQ, PMV, ZZh and JY. WC developed the software tool. PMV and JY contributed funding and resources. WC and JY wrote the manuscript with the participation of all authors. All authors reviewed and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

References

1. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
2. Visscher, P.M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
3. Pasaniuc, B. & Price, A.L. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet* **18**, 117-127 (2017).
4. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491-504 (2018).
5. Dadaev, T. *et al.* Fine-mapping of prostate cancer susceptibility loci in a large meta-analysis identifies candidate causal variants. *Nat Commun* **9**, 2256 (2018).
6. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
7. Chen, W. *et al.* Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719-736 (2015).
8. Zhu, X. & Stephens, M. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann Appl Stat* **11**, 1561 (2017).

9. Wang, G., Sarkar, A.K., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine-mapping. *bioRxiv*, 501114 (2019).
10. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906-14 (2014).
11. Lee, D., Bigdeli, T.B., Riley, B.P., Fanous, A.H. & Bacanu, S.A. DIST: direct imputation of summary statistics for unmeasured SNPs. *Bioinformatics* **29**, 2925-7 (2013).
12. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
13. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet* **99**, 139-153 (2016).
14. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228 (2015).
15. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun* **9**, 224 (2018).
16. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
17. Hartwig, F.P., Davies, N.M., Hemani, G. & Davey Smith, G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol* **45**, 1717-1726 (2016).
18. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
19. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).
20. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-52 (2016).
21. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
22. Vilhjalmsson, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-92 (2015).
23. Lloyd-Jones, L.R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* **10**, 5086 (2019).
24. Johnson, E.O. *et al.* Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. *Hum Genet* **132**, 509-22 (2013).
25. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
26. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98 (2008).
27. UK10K consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
28. Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* **47**, 1114-20 (2015).
29. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* **12**, e1001779 (2015).
30. Wu, Y., Zheng, Z., Visscher, P.M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol* **18**, 86 (2017).
31. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
32. Sonnega, A. *et al.* Cohort profile: the health and retirement study (HRS). *Int J Epidemiol* **43**, 576-585 (2014).
33. ARIC INVESTIGATORS. The atherosclerosis risk in communit (aric) stui) y: Design and objectwes. *American Journal of Epidemiology* **129**, 687-702 (1989).
34. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279 (2016).

35. Lee, J.J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**, 1112-1121 (2018).
36. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ Res* **122**, 433-443 (2018).
37. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet* **50**, 1505-1513 (2018).
38. Liu, J.Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet* **47**, 979-986 (2015).
39. Wray, N.R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* **50**, 668-681 (2018).
40. Pardinas, A.F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat Genet* **50**, 381-389 (2018).
41. Phelan, C.M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **49**, 680-691 (2017).
42. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94 (2017).
43. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641-3649 (2018).
44. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**, 918 (2018).
45. Lloyd-Jones, L.R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* **100**, 228-237 (2017).
46. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).
47. Robinson, M.R. *et al.* Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1**(2017).
48. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
49. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
50. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature communications* **6**, 1-9 (2015).
51. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**, 559-575 (2007).
52. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749-1755 (2019).

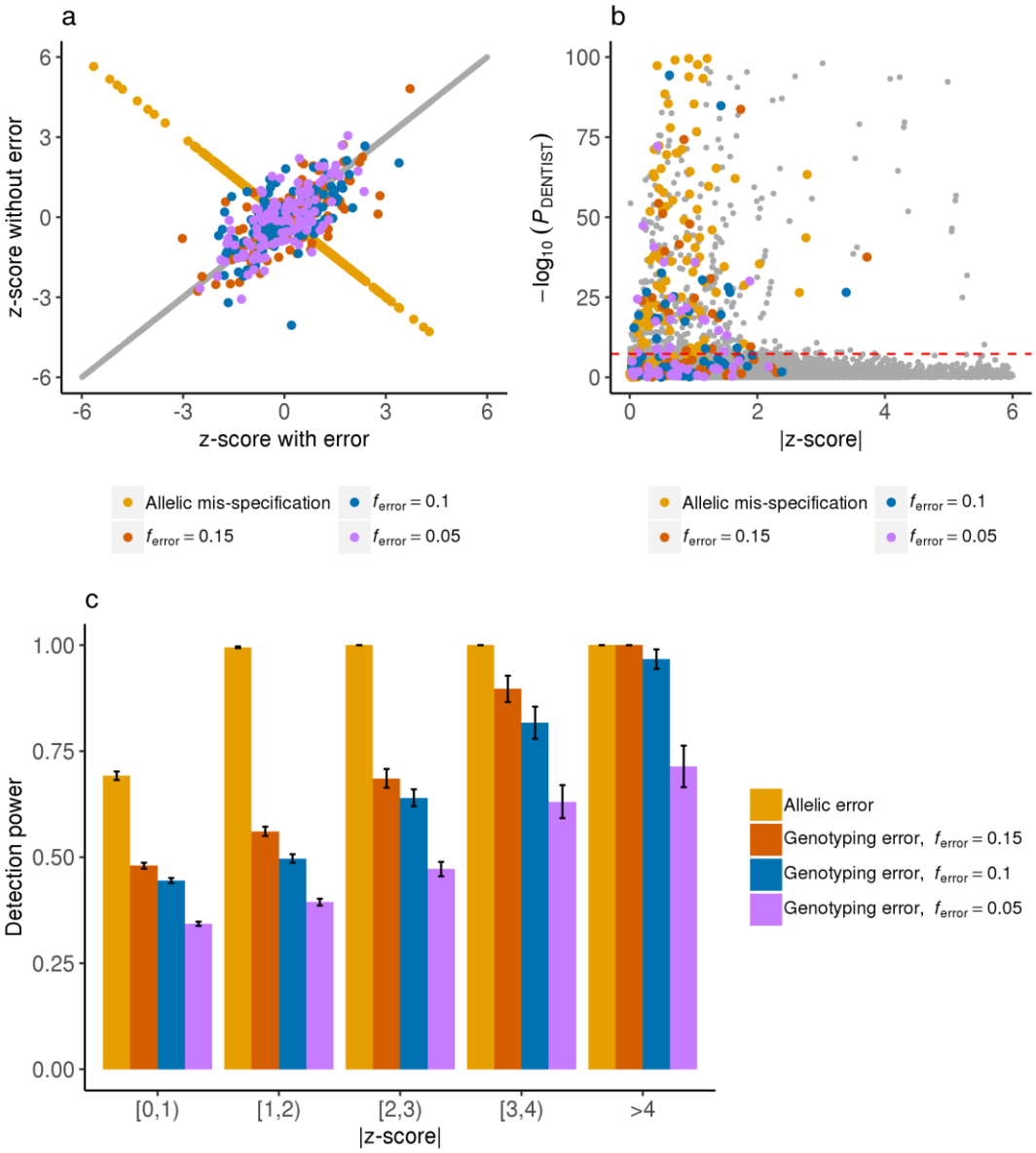


Figure 1. Detecting simulated allelic and genotyping errors using DENTIST. We assessed the power of DENTIST in detecting allelic and genotyping errors. There are three levels of genotyping error rate (f_{error}=0.15, 0.1 or 0.05), defined as the proportion of individuals with erroneous genotypes for a variant. Panel a) is a plot of the GWAS z-scores from data with simulated errors against those from data without such errors. The gray dots in the diagonal represents z-scores of variants without errors. In panel b), the DENTIST *P*-values are plotted against the absolute values of the GWAS z-scores for all the variants. The horizontal dashed line corresponds to $P = 5 \times 10^{-8}$. In panel c), to demonstrate the power in each |z-score| bin, we pooled the results from 200 simulations. Each bar plot (+/- s.e.) represents power computed from the pooled results.

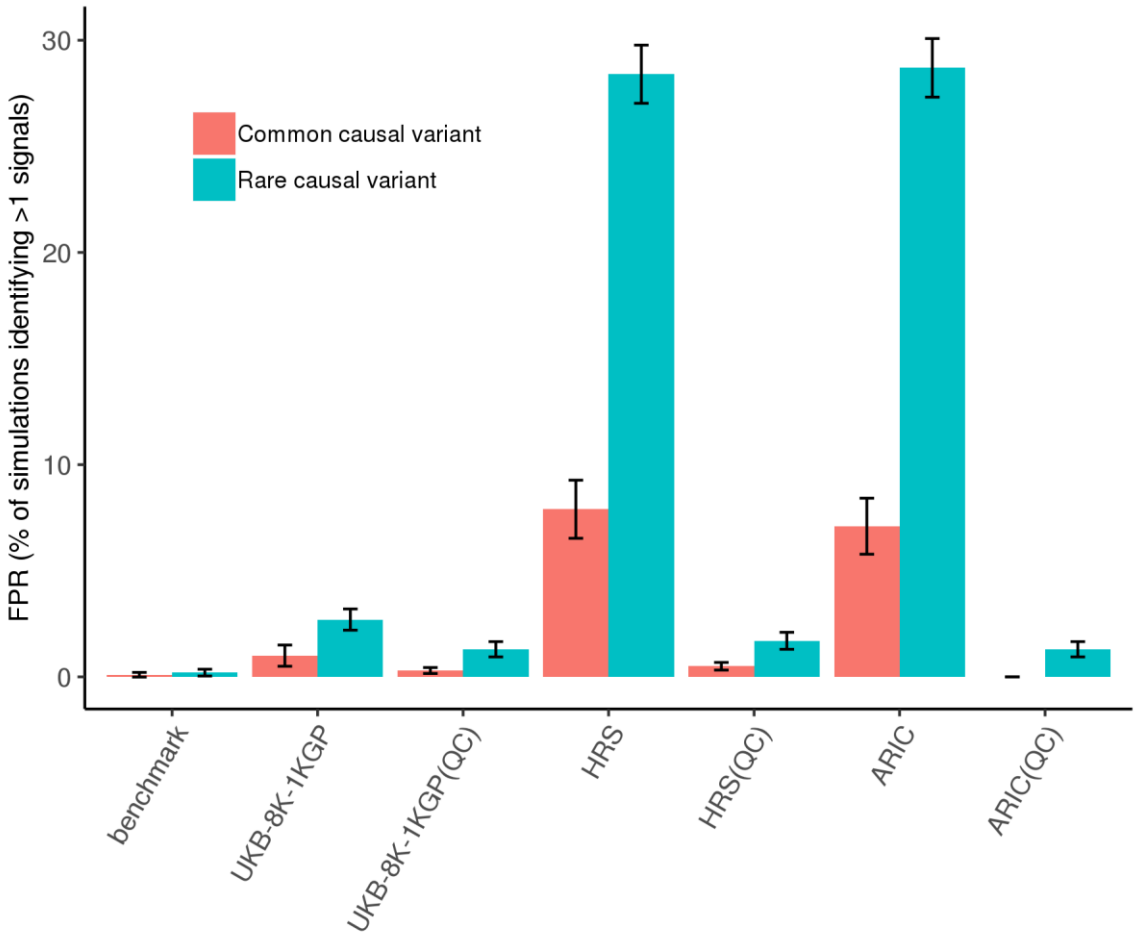


Figure 2. FPRs of COJO with and without DENTIST. Based on simulations with one causal signal, we assessed the FPRs of COJO analyses when performed with and without DENTIST-based QC (FPR is defined as the frequency of observing more than one COJO signals in the scenario in which only one causal variant 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.

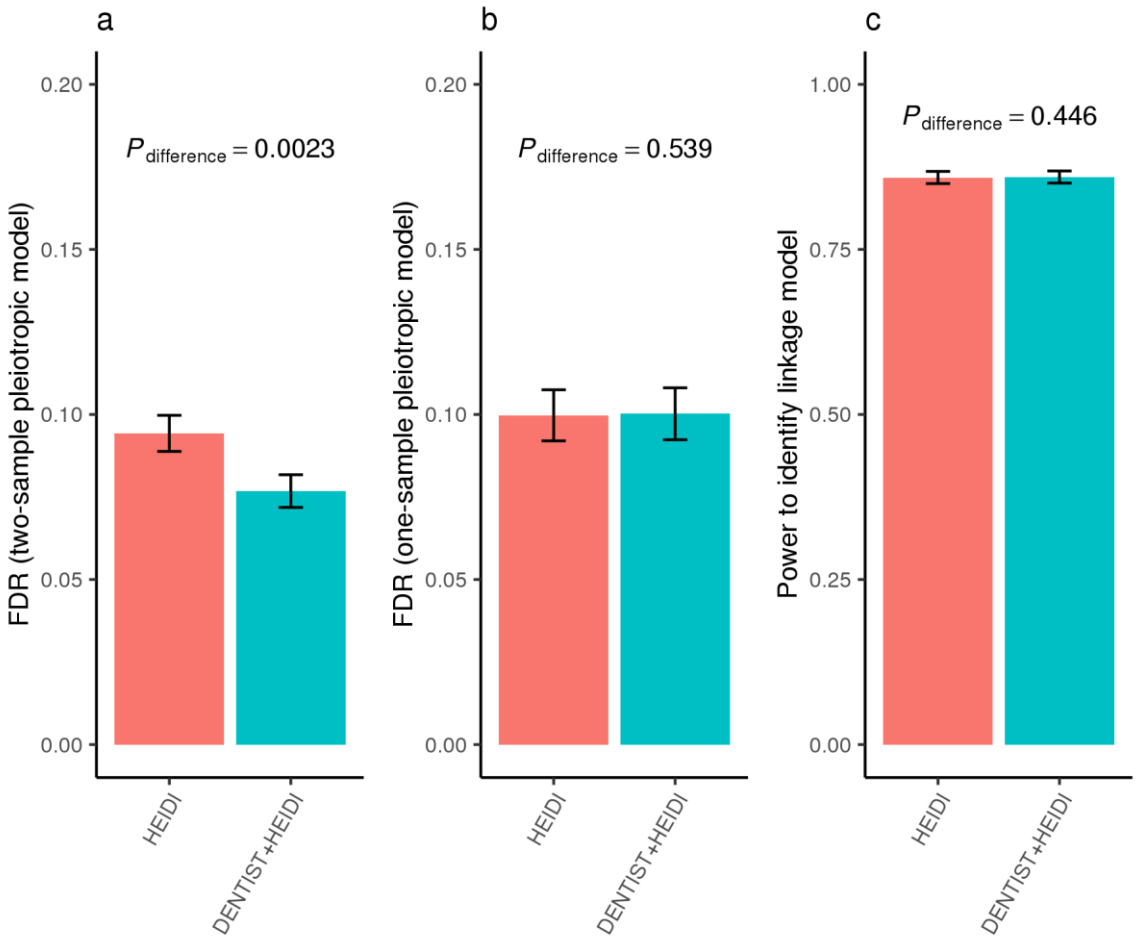
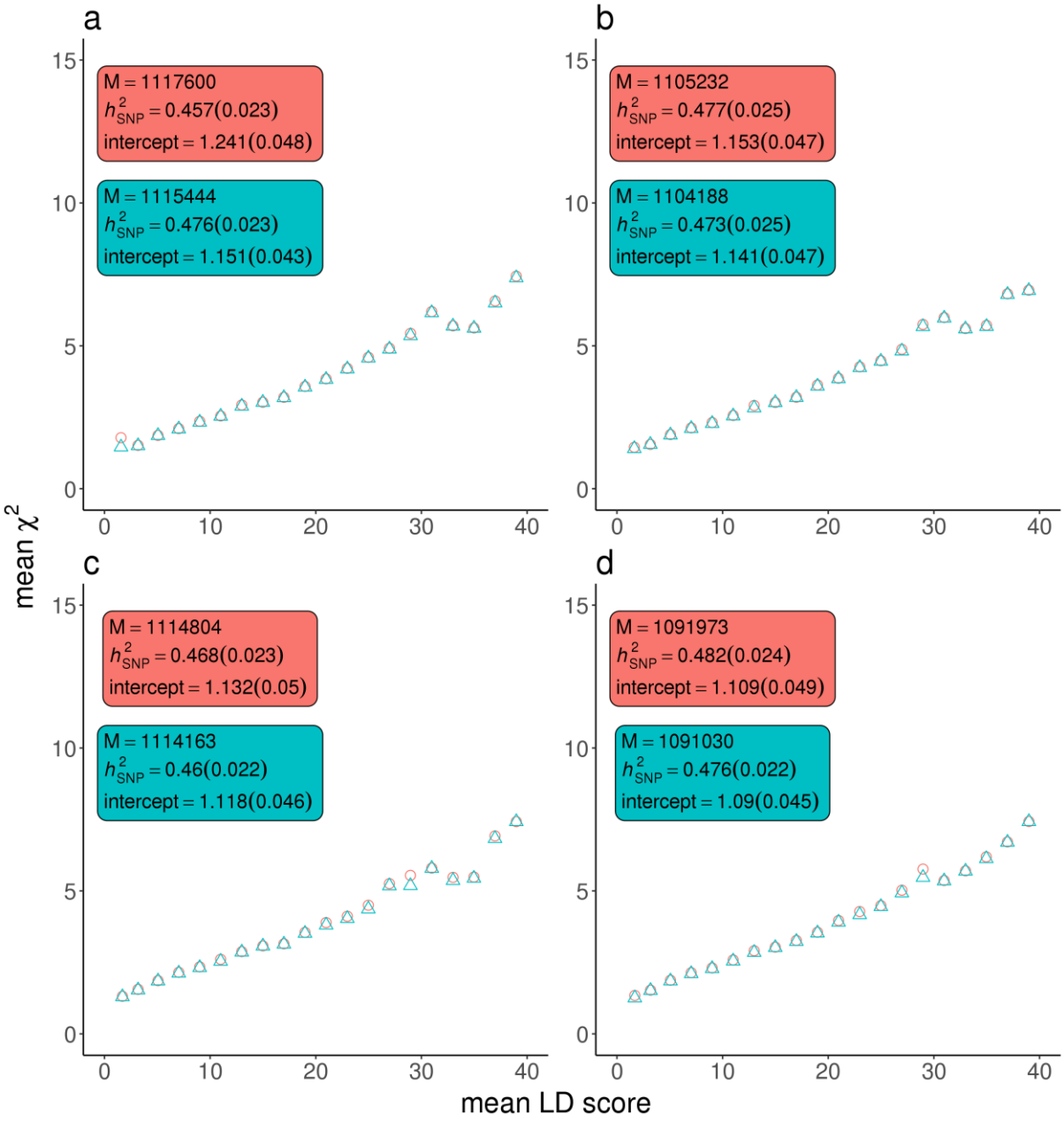


Figure 3. FPRs and power of the HEIDI test with and without DENTIST. Shown are the results from simulations to quantify the FPR of HEIDI under a pleiotropic model (panels **a** and **b**) and the power of HEIDI under a linkage model (panel **c**). The two-sample pleiotropic model in panel **a** refers to the scenario where the eQTL and GWAS summary data were simulated based on two different samples (HRS-3K and ARIC). The one-sample scenario in panel **b** refers to the scenario where the GWAS and eQTL data were simulated using the same sample (HRS-3K). In both scenarios, an independent sample (UK10K-1KGP) was used as the LD reference. $P_{\text{difference}}$ is to test if the FPR after QC is significantly different that without QC. $P_{\text{difference}}$ is calculated from a posterior distribution of $k_1 \sim$ Binomial (n, p) with p from a prior distribution of $p \sim Beta(k_2, n - k_2)$, where n is the number of simulation replicates, and k_1 and k_2 are the numbers of simulation replicates in which the HEIDI test correctly identified the right model with and without the DENTIST-based QC, respectively. The error bars correspond to the standard errors of the correspond metrics calculated from 4000 replications with re-sampled causal variants.

721



722

723 **Figure 4. The effect of DENTIST-based QC on LDSC analysis of the UKB height summary**
724 **data.** We assessed the effect of DENTIST on LDSC when different LD references were used,
725 including **a**) HRS, **b**) ARIC, **c**) UKB-8K-1KGP, and **d**) UK10K-WGS. For each reference sample,
726 LDSC was performed before and after DENTIST-based QC, and the corresponding results are
727 shown in the red and cyan text boxes, respectively, on each plot. The variants are binned by their
728 LD scores. Each dot on the plots represents the mean LD score value of each bin on the x-axis
729 and the mean χ^2 value on the y-axis, with those before and after DENTIST-based QC in red and
730 cyan colors respectively. In the textbox, "M" represents the number of variants, "h^2_{SNP}"
731 represents the estimate of SNP-based heritability, and "intercept" represents the LDSC intercept,
732 with the corresponding standard errors given the parentheses.

733 **Table 1.** FPRs and power of COJO before and after DENTIST-based QC in simulations.

Analysis method (LD reference)	FPR (%)		Power (%)	
	Common-causal	Rare-causal	Common-causal	Rare-causal
Benchmark	0.1 ± 0.11	0.2 ± 0.16	78.8 ± 0.9	30.6±1.4
COJO without DENTIST (UKB-8K-1KGP)	1.0 ± 0.50	2.7 ± 0.50	79.0 ± 0.9	28.6±1.9
COJO with DENTIST (UKB-8K-1KGP)	0.3 ± 0.14	1.3 ± 0.36	77.3 ± 0.9	22.2±1.6
COJO without DENTIST (HRS)	7.9 ± 1.37	28.4 ± 1.37	81.8±3.9	27.7±1.4
COJO with DENTIST (HRS)	0.5 ± 0.18	1.7 ± 0.40	81.2±1.3	16.8±1.1
COJO without DENTIST (ARIC)	7.1 ± 1.32	28.7 ± 1.38	77.6±0.9	26.7±1.7
COJO with DENTIST (ARIC)	0 ± 0.00	1.3 ± 0.36	75.8 ± 0.9	17.1±1.4

734 Benchmark: COJO analysis using the discovery GWAS as the reference without DENTIST. Shown
735 are mean ± standard error.

736

737

738 **Table 2.** Numbers of COJO signals from analyses of the UKB height summary data using different
739 LD reference samples with and without DENTIST-based QC.

	Benchmark	**UKBv3-20K (n = 20,000)**	**UKBv3-8K (n = 8,000)**	**HRS (n = 8,557)**	**ARIC (n = 7,703)**	**UK10K-WGS (n = 3,642)**
MAF > 0.01 Without DENTIST	1279	1296 (1.3%)	1337 (4.5%)	1477 (15.5%)	1485 (16.1%)	1417 (10.8%)
MAF > 0.01 With DENTIST	/	1300 (1.6%)	1319 (3.1%)	1338 (4.6%)	1353 (5.8%)	1413 (10.5%)
MAF > 0.001 Without DENTIST	1310	1313 (0.2%)	1337 (2.0%)	1555 (18.7%)	1645 (25.6%)	1473 (12.4%)
MAF > 0.001 With DENTIST	/	1326 (1.2%)	1326 (1.2%)	1346 (2.7%)	1398 (6.7%)	1421 (8.5%)

740 Benchmark: COJO analysis using the discovery GWAS (UKBv3-329K) as the reference without
741 DENTIST. The inflation rate as compared to the benchmark is shown in the parentheses.
742

743 **Table 3.** Estimates from LDSC analyses of the UKB height GWAS summary data with and without
744 DENTIST-based QC.

	Number of variants	One-step approach		Two-step approach	
		h^2_{SNP}	Intercept	h^2_{SNP}	Intercept
Reference = the discovery GWAS sample					
Benchmark	1114780	0.46 (0.023)	1.13 (0.049)	0.41(0.017)	1.34 (0.030)
Reference = HRS					
Without DENTIST	1117600	0.45 (0.022)	1.24 (0.047)	0.42 (0.018)	1.36 (0.028)
With DENTIST	1116249	0.46 (0.022)	1.15 (0.040)	0.41 (0.017)	1.35 (0.027)
Reference = ARIC					
Without DENTIST	1105232	0.47 (0.025)	1.15 (0.047)	0.42 (0.018)	1.34 (0.030)
With DENTIST	1102229	0.47 (0.024)	1.14 (0.047)	0.41 (0.018)	1.34 (0.030)
Reference = UKB-8K-1KGP					
Without DENTIST	1114804	0.46(0.023)	1.13(0.050)	0.41 (0.017)	1.33 (0.030)
With DENTIST	1113260	0.46(0.022)	1.12(0.048)	0.40 (0.016)	1.33 (0.030)
Reference = UK10K-WGS					
Without DENTIST	1091973	0.48 (0.023)	1.10(0.048)	0.42 (0.018)	1.31 (0.029)
With DENTIST	1074399	0.47 (0.022)	1.07 (0.042)	0.41 (0.016)	1.30(0.028)
Reference = 1KGP-EUR					
Without DENTIST	1133151	0.48 (0.024)	1.15 (0.047)	0.43 (0.018)	1.33 (0.028)
With DENTIST	1071447	0.40 (0.018)	1.03 (0.030)	0.35 (0.014)	1.22 (0.024)

745 Standard errors are given in the parentheses.