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

Wenhan Chen[1,2], Yang Wu[1], Zhili Zheng[1], Ting Qi[1,3,4], Peter M Visscher[1], Zhihong Zhu[1], Jian Yang[1,3,4*]

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

[2]Epigenetics Research Laboratory, Genomics and Epigenetics Theme, Garvan Institute of Medical Research, Sydney, NSW, 2010, Australia

[3]School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China

[4]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China

[*]Correspondence: Jian Yang (jian.yang@westlake.edu.cn)

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 heterogeneity between GWAS and LD reference). We further show that DENTIST can improve other summary-data-based analyses such as fine-mapping 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[4-11], imputing summary statistics at untyped variants[12,13], estimating SNP-based heritability[14-16], assessing causal or genetic relationship between traits[17-19], prioritizing candidate causal genes for a trait[20-23], and polygenetic risk prediction[8,24,25]. 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 biased 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 data generation and analysis processes (e.g., genotyping/imputation errors and genetic variants with mis-specified effect alleles)[26,27], 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, imputation quality, and 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[28], and the effects of such differences on summary-data-based analyses are largely unexplored.

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 after standard QC. We then demonstrate the utility of DENTIST as an additional QC step for multiple, frequently-used, summary data-based methods, including the conditional and joint analysis (COJO)[6] of summary statistics, FINEMAP[9], LD score regression[14], and heterogeneity in dependent instruments (HEIDI) test[23].

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**).

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

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 $\mathbf{R}_{\boldsymbol{tt}}$ being the LD between variants $\boldsymbol{t}$ and $\mathbf{R}_{\boldsymbol{it}}$ being the LD between variant *i* and variants $\boldsymbol{t}$. 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[12,13]. T_d follows approximately a χ^2 distribution with 1 degree of freedom. We denote *P*-value calculated from T_d by P_{DENTIST}. A significant difference between the observed and predicted z-scores indicates error(s) in the discovery GWAS or LD reference, and/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 $\boldsymbol{t}$. Following prior work[15], we use the truncated singular value decomposition (SVD) method to compute a pseudo inverse of $\mathbf{R}_{\boldsymbol{tt}}$ to improve computational stability (**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 variant partitioning approach (**Methods**). In each iteration, we partition the variants into two subsets (S1 and S2) at random, remove variants with $P_{\text{DENTIST}} < 5\times10^{-8}$ from the variant group with $P_{\text{GWAS}} > 0.01$, and apply the same proportion of variants removed to the variant group with $P_{\text{GWAS}} < 0.01$ (note: the proportion of variants removed is capped at 0.5% in each group). 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 $P_{\text{DENTIST}} < 5\times10^{-8}$ are removed in the final step without the constraints.

Detecting simulated errors in GWAS data after standard QC

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 on chromosome 22 of 3,642 unrelated individuals from the UK10K project after QC[29,30] (denoted by UK10K-WGS). A descriptive summary of the data sets used in this study can be found in **Supplementary Table 1** and the Methods section. We

100 performed standard QC on the imputed data to exclude variants with an imputation INFO score
101 < 0.3 or a Hardy-Weinberg Equilibrium (HWE) P-value <10^{-6}. We simulated a trait affected by 50
102 common, causal variants with effects drawn from $N(0,1)$, which together explained 20% of the
103 phenotypic variation (proportion of variance explained by a causal variant, denoted by q^2, was
104 0.4%, on average). Prior to the simulations with errors, we performed a benchmark simulation
105 (i.e., simulating a scenario without errors and applying DENTIST using the discovery GWAS
106 sample as the LD reference) to show that the DENTIST test statistics were well calibrated, given
107 the small fraction of variants removed by DENTIST in each simulation replicate and the small
108 proportion of causal variants removed across simulation replicates, in the absence of errors and
109 LD heterogeneity (**Supplementary Figure 1**).
110
111 We then simulated genotyping and allelic errors at 0.5% randomly selected variants
112 respectively. Genotyping errors of each of these variants were simulated by altering the
113 genotypes of a certain proportion (f_{error} = 0.05, 0.1 or 0.15) of randomly selected individuals, and
114 allelic error of each of the variants was introduced by swapping the effect allele by the other
115 allele. The simulation was repeated 200 times with the causal and erroneous variants re-
116 sampled in each simulation. We then ran DENTIST using UK10K-WGS or an independent sample
117 (UKB-8K-1KGP) as the LD reference after standard QC of the discovery GWAS: removing
118 variants with an HWE P-value < 10^{-6} using the individual-level data or ΔAF > 0.1 where ΔAF is
119 the difference in allele frequency (AF) between the summary data and reference sample. The
120 independent sample UKB-8K-1KGP refers to a set of 8,000 unrelated individuals from the UK
121 Biobank[31] (UKB) with variants imputed from the 1000 Genomes Project (1KGP). The statistical
122 power (sensitivity) was measured by the proportion of erroneous variants in the data that can
123 be detected by DENTIST. We also computed the fold enrichment in probability of an erroneous
124 variant being detected by DENTIST compared to a random guess (i.e., the ratio of the percentage
125 of true erroneous variants in the variants detected by DENTIST to that in all variants).
126
127 When using UKB-8K-1KGP as the reference, ~45% of the genotyping and ~95% of the allelic
128 errors could be removed by the standard QC (**Supplementary Table 2**). After the standard QC,
129 DENTIST was able to detect ~27% of the remaining genotyping and ~73% of the remaining
130 allelic errors (**Figure 1** and **Supplementary Table 3**), with only ~0.2% variants being removed
131 in total (**Supplementary Table 4**). The fold enrichment of errors in the removed variants was
132 366 for allelic errors and of 137 for genotyping errors (**Supplementary Table 5**). Notably, the
133 power to detect allelic errors was ~73% for variants with MAF > 0.45, compensating the low
134 power of the ΔAF approach in this MAF range (~16%). When restricted to variants passing the
135 genome-wide significance level (i.e., $P_{GWAS} < 5\times10^{-8}$), the power of DENTIST increased to ~93%

136 for the genotyping errors and 100% for the allelic errors (**Supplementary Table 3**). The power
137 also varied with the genotyping error rate (f_{error}), e.g., the power in the f_{error}=0.05 scenario was,
138 on average, lower than that for f_{error}=0.15 (**Figure 1c** and **Supplementary Table 3**). When using
139 UK10K-WGS as the reference (mimicking the application of DENTIST in a scenario where
140 individual-level data of the discovery GWAS are available), the power remained similar, but the
141 fold enrichment was much higher compared to that using UKB-8K-1KGP (**Supplementary**
142 **Tables 5** and **6**). The percentage of causal variants removed from DENTIST was 0.19% using in-
143 sample LD from UK10K-WGS, <1% using out-of-sample LD from UKB-8K-1KGP
144 (**Supplementary Table 4**). Moreover, using the same simulation setting, we explored the choice
145 of the parameter θ_k (i.e., the proportion of eigenvectors retained in the truncated SVD method;
146 see **Methods** for details) and reference sample size (n_{ref}); the results suggested that a choice of
147 θ_k=0.5 and $n_{\mathrm{ref}} \geq 5000$ is appropriate in practice (**Supplementary Figure 2**). In addition, the
148 power of DENTIST to detect genotyping and allelic errors was robust to the randomness of
149 variant splitting (**Supplementary Figure 3**), or whether the GWAS analysis was corrected for
150 the variant-derived principal components in the presence of population stratification effect
151 (**Supplementary Figure 4**). Altogether, these results demonstrate the power of DENTIST to
152 identify allelic and genotyping errors after the standard QC, suggesting that DENTIST can
153 complement existing QC filters for either individual- or summary-level GWAS data. Nevertheless,
154 we acknowledge that this simulation did not cover the full complexity of real case scenarios,
155 which may involve multiple independent samples with heterogeneous LD structures caused by
156 several factors, such as imputation errors or ancestry mismatches (**Supplementary Figure 5**).
157 These cases are difficult to mimic in this simulation but will be assessed in the following
158 analyses.
159
160 **Applying DENTIST to COJO with simulated phenotypes**
161 COJO is a method that uses GWAS summary statistics and reference LD to run a conditional and
162 joint multi-SNP regression analysis[6]. We used simulations to assess the performance of COJO in
163 the presence of heterogeneity between discovery GWAS and LD reference samples before and
164 after DENTIST filtering. To mimic the reality that causal signals are often not perfectly captured
165 by imputed variants, we simulated a phenotype affected by one or two sequence variants using
166 WGS data (i.e., UK10K-WGS) and performed association analyses using imputed data of the same
167 individuals (imputing 312,264 variants, in common with those on an SNP array, to the 1KGP[30,32];
168 denoted by UK10K-1KG). More specifically, we first randomly selected one or two sequence
169 variants with MAF≥0.01 (denoted by common-causal) or 0.01>MAF≥0.001 (denoted by rare-
170 causal) from a locus as causal variant(s) to generate a phenotype (note: MAF > 0.001 is
171 equivalent to minor allele count > 7 in this sample). The proportion of phenotypic variance

172 explained by the causal variant (denoted by q^2) was set to 2% to achieve similar power to a
173 scenario with q^2 = 0.03% and n = 250,000 (note: the mean q^2 of 697 height variants discovered
174 in Wood et al.[33] is 0.03%) because the power of GWAS is determined by $nq^2/(1-q^2)$. Then, we
175 ran a GWAS using UK10K-1KGP and performed COJO analyses using multiple LD references,
176 including the discovery GWAS sample (i.e., UKB-8K-1KGP), the Health Retirement Study (HRS)[34],
177 and the Atherosclerosis Risk In Communities (ARIC) study[35], with different degrees of ancestral
178 differences with UK10K-1KGP (**Supplementary Figure 6**). For a fair comparison, only the
179 variants shared between these reference samples were included. We repeated both the
180 common-causal and rare-causal simulations 100 times for each autosome and computed the
181 false-positive rate (FPR, i.e., the frequency of observing two COJO signals in the scenario where
182 there was only one causal variant) and power (the frequency of observing two COJO signals in
183 the scenario where there were two distinct causal variants with LD r^2 <0.1 between them). It
184 should be noted that the false-positive COJO signals defined here are not false associations but
185 falsely claimed as jointly associated (also known as quasi-independent) signals.
186
187 When using the discovery GWAS sample as the LD reference, the FPR of COJO was 0.1% for
188 common-causal and 0.2% for rare-causal (**Figure 2** and **Table 1**), which can be regarded as a
189 baseline for comparison as there was no data heterogeneity in this case. The FPRs were higher
190 than the expected values because the causal variants were not perfectly tagged by the imputed
191 variants (**Supplementary Table 7**). When using UKB-8K-1KGP (i.e., 1KGP-imputed data of
192 8,000 UKB participants with similar ancestry to the UK10K participants as shown in
193 **Supplementary Figure 6**) as the LD reference, the FPR was close to the benchmark for
194 common-causal (1%) and slightly inflated for rare-causal (2.7%) (**Figure 2**). After DENTIST
195 filtering (using UKB-8K-1KGP as the LD reference), the FPR for rare-causal decreased to 1.6%.
196 When using LD computed from European-American individuals in HRS or ARIC, the FPR of COJO
197 was strongly inflated in the whole MAF range: >7% for common-causal and >28% for rare-
198 causal, likely because of the difference in ancestry between HRS/ARIC and UK10K-1KGP.
199 DENTIST can effectively control the FPR of COJO to <1% for common-causal and <3.5% for rare-
200 causal (**Figure 2** and **Table 1**). We performed additional simulations with three causal variants
201 at a locus and observed similar patterns of changes in FPR before and after DENTIST QC
202 (**Supplementary Figure 7**). Taken together, the FPR of COJO was reasonably well controlled for
203 common variants but substantially inflated for rare variants, especially when there was a
204 difference in ancestry between the GWAS and LD reference samples, and most of the false-
205 positive COJO signals could be removed by DENTIST.
206

207 The power of COJO (without DENTIST) using in-sample LD from UK10K-1KGP or out-of-sample
208 LD from the other references were similar: 77-80% for common-causal and 26-30% for rare-
209 causal (**Table 1**). The low power for rare causal variants was because they were poorly captured
210 by imputation (**Supplementary Table 7**). DENTIST filtering caused a <1% loss of power for
211 common-causal and 3-4% for rare-causal (**Table 1**), which is acceptable because the reduction
212 in FPR was larger than that in power, especially for rare variants.
213
214 **Applying DENTIST to COJO for real phenotypes**
215 Having assessed the performance of DENTIST in COJO analyses by simulation, we then applied it
216 to COJO analyses for height in the UKB. The height GWAS summary statistics ($n = 328{,}577$) were
217 generated from a GWAS analysis of all the unrelated individuals of European ancestry in the UKB
218 (denoted by UKBv3-329K) except 20,000 individuals (denoted by UKBv3-20K), deliberately left-
219 out to be used as a non-overlapping LD reference. Genotype imputation of the UKB data was
220 performed by the UKB team with most of the variants imputed from the Haplotype Reference
221 Consortium (HRC)[36]. We performed COJO analyses with a host of references: overlapping in-
222 sample references with sample sizes (n_{ref}) varying from 10,000 to 150,000, non-overlapping in-
223 sample references including UKBv3-8K ($n = 8{,}000$) and UKBv3-20K (containing UKBv3-8k), and
224 out-of-sample references including ARIC and HRS (**Supplementary Table 1**). We excluded from
225 the analysis variants with MAF < 0.001 to ensure sufficient minor alleles for rare variants in
226 reference samples with n_{ref} < 10k. We first performed a COJO analysis using the actual GWAS
227 sample as the reference and identified 1,279 signals from variants with MAFs >0.01, and 1,310
228 signals from variants with MAFs > 0.001 (**Table 2**). These results can be regarded as a
229 benchmark. When using the overlapping in-sample LD references, the number of COJO signals
230 first decreased as n_{ref} increased and then started to stabilize when n_{ref} exceeded 30,000
231 (**Supplementary Figure 8**). The results from using the two non-overlapping in-sample
232 references (UKBv3-8K and UKBv3-20K) were comparable to those from using the overlapping
233 in-sample references with similar sample sizes (**Table 2** and **Supplementary Table 8**) because
234 the non-overlapping in-sample references, despite being excluded from the GWAS, were
235 consistent with the GWAS sample with respect to ancestry, data collection, and analysis
236 procedures.
237
238 When using LD from an out-of-sample reference (either HRS or ARIC), there was substantial
239 inflation in the number of COJO signals compared to the benchmark (by 15.5-16.1% for common
240 variants and 18.7-25.6% for all variants), with a few variants in weak LD with those identified
241 from the benchmark analysis (**Supplementary Figure 9**). The results from using the two out-of-
242 sample references became more consistent with the benchmark after DENTIST filtering, with the

243 inflation reduced to 5.0-7.2% for common variants and 3.9-4.4% for all variants, comparable to
244 the results using an in-sample LD reference with a similar sample size (**Table 2**). Polygenic score
245 analysis shows that the reduction in the number of COJO signals owing to DENTIST QC had
246 almost no effect on the accuracy of using the COJO signals to predict height in HRS
247 (**Supplementary Table 9**), suggesting the redundancy of the signals removed by DENTIST. We
248 further found that compared to using the imputed data from ARIC or HRS, using UK10K-WGS
249 ($n = 3{,}642$) as the reference showed lower inflation (10.9% for common variants and <8.5% for
250 all variants) before DENTIST QC but larger inflation after DENTIST QC (**Table 2**), suggesting a
251 large reference sample size is essential even for WGS data. In all the DENTIST analyses above,
252 the total number of removed variants varied from 0.03% to 1.4% (**Supplementary Table 10**).
253 All these results are consistent with what we observed from simulations, demonstrating the
254 effectiveness of DENTIST in eliminating heterogeneity between GWAS and LD reference
255 samples.
256
257 We further applied DENTIST to Educational Attainment (EA), Coronary Artery Disease (CAD),
258 Type 2 Diabetes(T2D), Crohn's Disease (CD), Major Depressive Disorder (MDD), Schizophrenia
259 (SCZ), Ovarian Cancer (OC), Breast Cancer (BC), Height, and Body Mass Index (BMI) using GWAS
260 summary data from the public domain[37-45] (**Supplementary Table 11**) and three LD reference
261 samples (i.e., ARIC, HRS, and UKBv3-8K). Since these published studies focus on common
262 variants (rare variants are not available in most of the data sets), we used an MAF threshold of
263 0.01 in this analysis. When using ARIC as the LD reference, the proportion of variants removed
264 by DENTIST QC ranged from 0.07% (BMI) to 0.86% (CAD) with a median of 0.21%
265 (**Supplementary Table 12**), and the reduction in the number of COJO signals for common
266 variants ranged from 0% (OC and MDD) to 12.5% (CAD) with a median of 1.6%
267 (**Supplementary Table 13**). The results from using the other two references were similar
268 (**Supplementary Table 12** and **13**).
269
270 **Improved FINEMAP analysis by DENTIST**
271 FINEMAP[9] is a method used for prioritizing causal variants from GWAS summary statistics. We
272 assessed the effect of DENTIST QC on an FINEMAP analysis using the data simulated above with
273 one, two or three common causal variants at a 1-Mb locus. The impact of the use of an out-of-
274 sample LD reference on FINEMAP was evaluated by the estimated number of causal variants
275 before or after DENTIST QC, in comparison with the benchmark analysis using in-sample LD. We
276 observed that FINEMAP using out-of-sample LD tended to overestimate the number of causal
277 variants (**Supplementary Figures 10** and **11**). For example, given a posterior inclusion
278 probability threshold of 95%, FINEMAP with out-of-sample LD from the HRS cohort provided

279 the correct number of causal variants in 78.4%, 49.8%, and 44.8% of the simulation replicates in
280 the one-, two-, and three-causal scenarios, respectively (**Figure 3a**). After DENTIST QC, the
281 proportions of times when FINEMAP identified the correct number of causal variants increased
282 substantially to 87.3%, 78.5%, and 63.4% in the one-, two-, and three-causal scenarios,
283 respectively, almost comparable to the corresponding proportions (i.e., 91.4%, 87.2%, and
284 68.2%) from the benchmark analysis (**Figure 3a**). The results remained similar when using
285 ARIC as the LD reference (**Figure 3b**). These results demonstrate a substantial improvement of
286 the performance of FINEMAP after DENTIST QC.
287
288 **HEIDI test with DENTIST**
289 The summary data-based Mendelian randomization (SMR) is a method that integrates summary-
290 level data from a GWAS and an expression quantitative trait loci (eQTL) study to test pleiotropic
291 associations between a trait and expression levels of genes[23]. It features the HEIDI test that
292 utilizes multiple cis-eQTL variants at a locus to distinguish pleiotropy (the trait and expression
293 level of a gene are affected by the same causal variants) from linkage (causal variants for the
294 trait are in LD with a distinct set of causal variants affecting gene expression). The HEIDI test
295 uses summary data from two studies and LD from a reference so that any errors in and
296 heterogeneity between the GWAS, eQTL and reference samples can cause inflated HEIDI test
297 statistics, giving rise to more SMR associations being rejected than expected by chance[23,46]. Here,
298 we performed simulations to assess the effect of data heterogeneity on HEIDI and sought to
299 mitigate it using DENTIST. We first generated a trait based on a causal variant (q^2 = 1%)
300 randomly sampled from the variants on chromosome 22 in the ARIC data. To simulate a
301 pleiotropic model, we used the same causal variant to simulate the expression level of a gene in a
302 subset of the HRS data (n = 3,000; denoted by HRS-3K) with q^2 for the gene expression level
303 randomly sampled from the eQTL q^2 distribution reported by the Consortium for the
304 Architecture of Gene Expression (CAGE)[47]. To simulate a linkage model, a second causal variant
305 in LD with the trait causal variant ($0.25 < r^2 \leq 1$) was selected to generate the gene expression
306 level, again with the eQTL q^2 value sampled from the CAGE. In addition to the two-sample
307 scenario above, we also simulated a one-sample scenario in which both the trait and gene
308 expression level were generated in the HRS-3K sample. The UKB-8K-1KGP sample was used as
309 the LD reference for both the SMR-HEIDI and DENTIST analyses. For each scenario, the
310 simulation was repeated 4,000 times with the causal variants re-sampled in each replicate. The
311 FPR of the HEIDI test was calculated as the proportion of pleiotropic models detected with P_{HEIDI}
312 < 0.05, and the power was defined as the proportion of linkage models detected with P_{HEIDI} <
313 0.05.
314

315 We found that the FPR of HEIDI was close to the expected value (5%) in the one-sample scenario
316 (5.8%) but inflated (9.8%-12.0%) in the two-sample scenario (**Figures 4a**). We then performed
317 DENTIST in both the GWAS and eQTL summary data using UKB-8K-1KGP as the reference. After
318 DENTIST filtering, the FPRs of HEIDI in the two-sample scenario decreased to 8.4%-9.2%; the
319 changes were small but consistent across scenarios (**Figure 4a**). On the other hand, the power of
320 HEIDI to detect the linkage model remained almost the same before and after DENTIST filtering
321 (**Figure 4b**). All these results suggest that DENTIST slightly improved the FPR of HEIDI in the
322 presence of data heterogeneity at almost no expense of power and that in the presence of
323 heterogeneity between the GWAS and eQTL samples, HEIDI tends to be conservative (rejecting
324 more SMR associations than expected by chance) even after DENTIST filtering.
325
326 **LD score regression analysis with DENTIST**
327 LD score regression (LDSC) is an approach which was originally developed to distinguish
328 polygenicity from population stratification in GWAS summary data set by a weighted regression
329 of GWAS χ^2 statistics against LD scores computed from a reference[14] but has often been used to
330 estimate the SNP-based heritability (h^2_{SNP}). We investigated the impact of DENTIST on LDSC
331 using the height GWAS summary data generated using the UKBv3-329K sample along with
332 several reference samples including four imputation-based samples (i.e., the discovery GWAS
333 sample, HRS, ARIC and UKB-8K-1KGP) and two WGS-based samples (i.e., UK10K-WGS and the
334 European individuals from the 1KGP (1KGP-EUR)). We performed the one- and two-step LDSC
335 analyses using LD scores of the variants, in common with those in the HapMap3, computed from
336 each of the references before and after DENTIST-based QC. Note that DENTIST was performed
337 for all common variants but only those overlapped with HapMap3 were included in the LDSC
338 analyses.
339
340 Using the discovery GWAS sample as the reference, the estimates of h^2_{SNP} and regression
341 intercept from the one-step LDSC were 46% (SE = 0.02) and 1.13 (SE = 0.04) respectively
342 (**Supplementary Table 14**). When using the other reference samples, the results were very
343 close to the benchmark except for a noticeably larger estimate of regression intercept using HRS
344 (1.24, SE = 0.04). After DENTIST filtering, the intercept estimate using HRS decreased to 1.18 (SE
345 = 0.04) with little difference in $\hat{h}^2_{SNP}$ (increased from 45% to 47%) (**Supplementary Table 14**).
346 To better understand the effect of DENTIST QC on LDSC using HRS, we plotted the mean χ^2-
347 statistic against the mean LD score across the LD score bins. We found that the GWAS mean χ^2-
348 statistic in the bin with the smallest mean LD score deviated from the value expected from a
349 linear relationship between the LD score and χ^2-statistic (**Supplementary Figure 12**), and the
350 deviation was removed by filtering out a small proportion of variants with very small LD scores

351 but large χ^2-statistic by DENTIST (**Supplementary Figure 13**). These results show how the
352 quality of LD reference can impact the LDSC analysis, which is typically unknown *a priori* and
353 varied across different reference samples. We also re-ran the analyses using the two-step LDSC,
354 where the intercept was estimated using the variants with χ^2 values < 30 in the first step and
355 constrained in the second step to estimate h^2_{SNP} using all the variants. Compared to the one-step
356 approach, the two-step approach provides larger estimates of the intercepts and smaller
357 estimates of h^2_{SNP} either before or after DENTIST filtering. In addition, we applied LDSC to the 10
358 published GWAS data sets mentioned above (**Supplementary Table 11**) using ARIC, UKB-8K-
359 1KG and UK10K-WGS as the reference. The improvement of LDSC by DENTIST QC was small
360 particularly for traits whose LDSC intercepts were close to 1 before DENTIST QC
361 (**Supplementary Table 15**), demonstrating the robustness of LDSC to data heterogeneity and
362 errors.
363
364 **Discussion**
365 In this study, we developed DENTIST, a QC tool for summary data-based analyses, which
366 leverages LD from a reference sample to detect and filter out problematic variants by testing the
367 difference between the observed z-score of a variant and a predicted z-score from the
368 neighboring variants. From simulations and real data analyses, we show that some of the
369 commonly used analyses can be biased to various extents in the presence of data heterogeneity,
370 e.g., inflated number of COJO signals, overestimated number of causal variants from FINEMAP, or
371 elevated rate of rejecting pleiotropic models for SMR-HEIDI. For most of these analyses,
372 DENTIST-based QC can mitigate the biases to some extent.
373
374 Our results suggest that summary-data-based analyses are generally well calibrated in the
375 absence of data heterogeneity but biased otherwise despite the use of multiple standard QCs.
376 For example, heterogeneity between the discovery GWAS and LD reference caused inflated FPR
377 of both COJO and SMR-HEIDI analyses (**Table 2** and **Figure 4**). Also, we found that the FPR of
378 COJO for rare variants was much higher than that for common variants likely because rare
379 variants are more difficult to impute such that they are more likely to have discrepancy in LD
380 between two imputed data sets. It should be clarified that the false-positive COJO signals as
381 defined here are not false associations but falsely claimed as jointly significant associations.
382 DENTIST substantially reduced false-positive detections from COJO analyses, especially for rare
383 variants, even when there was a difference in ancestry between the GWAS and LD reference
384 samples. This extends the utility of COJO, which was originally developed for common variants,
385 to rare variants. This message is important for the field because more and more GWASs and
386 meta-analyses have started to include imputed/sequenced rare variants. It is also encouraging

387 to see the substantial improvement of FINEMAP after DENTIST QC, implying the utility of
388 DENTIST for fine-mapping methods. As for the HEIDI method, the FPR was only marginally
389 reduced by DENTIST but at almost no cost of power. It should also be clarified again that the
390 inflated HEIDI test-statistics would not lead to false discoveries because higher HEIDI test-
391 statistics correspond to higher probability of rejecting SMR associations rather than tending to
392 claim more significant associations. Among all the methods tested, LDSC was the least affected
393 by errors or data heterogeneity (**Figure 4a**). In general, improvement of a summary data-based
394 analysis before and after DENTIST QC depends on how susceptible the methods are to errors
395 and LD heterogeneity. Methods based on signed LD correlations are more prone to errors and
396 LD heterogeneity than those based on unsigned LD correlations or LD scores. For instance, COJO
397 is relatively susceptible to errors and LD heterogeneity because it relies on signed LD
398 correlations in a multiple regression framework. Suppose the effect allele of an SNP is
399 mislabeled (equivalent to the effect size changed to the opposite direction) in the summary data.
400 In that case, the effect sizes of the SNP and another SNP in LD with it will be heavily
401 overestimated in a joint or conditional analysis because the input effect sizes and LD pattern
402 appear to suggest that the effects of the two SNPs mask each other in the marginal analysis. Even
403 in the absence of errors in the summary data, the joint or conditional effects of two SNPs in LD
404 can be estimated with substantial biases in the presence of large LD heterogeneity between the
405 discovery and the reference. Compared to common variants, rare variants are less well imputed
406 and thus are more likely to have LD heterogeneity between data sets, which may explain the
407 remarkable improvement of COJO for rare variants before and after DENTIST QC. DENTIST has
408 the unique feature to detect heterogeneity between a GWAS summary data set and an LD
409 reference. The benefit of using DENTIST as a QC tool has been demonstrated in the four case
410 studies above, but we believe that it can potentially be applied to all GWAS summary data-based
411 analyses that require a LD reference such as other fine-mapping methods[7,48], and joint modeling
412 of all variants for polygenic risk prediction[24,25,49]. In addition to dealing with data heterogeneity,
413 the power of DENTIST also comes from detecting genotyping or imputation errors that passed
414 standard QCs, making it a complementary method to existing QC methods.
415
416 Given that a QC step can potentially remove true signals, we make sure that DENTIST is
417 conservative in filtering variants. We show by simulation that in the absence of errors and data
418 heterogeneity, the DENTIST test statistics were not inflated, and on average, only <0.05%
419 variants were filtered out by DENTIST (**Supplementary Figure 1**). In practice, we implemented
420 an iterative variant partitioning approach to avoid widespread inflation of the DENTIST
421 statistics in the presence of data errors or heterogeneity. Throughout all the analyses performed
422 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 (**Supplementary Tables 4, 10, 12, 14**).

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. First, DENTIST is a QC method for detecting not only errors in summary data but also heterogeneity between discovery and reference data. As shown in our simulation, DENTIST does not guarantee to filter all the errors but most of them with large GWAS z-scores and a large proportion of them with small z-scores. Second, DENTIST can identify errors that passed the standard QC approaches (such as HWE test and allelic frequency checking), which makes it a 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**). Third, regarding the choice of a reference sample, DENTIST expects unrelated individuals from a closely matched ancestry, with a large sample size (n > 5,000). Third, prior work suggests that n_{ref} needs to scale with GWAS sample size[50]. We have quantified the performance of DENTIST with respect to reference sample size (n_{ref}) and shown that a reference sample size of >5000 is recommended and the relationship between the performance of DENTIST and n_{ref} does not seem to depend on GWAS sample size (**Supplementary Figure 2**). Fourth, we may use DENTIST to choose from multiple LD reference samples the best matching reference for an analysis that requires reference LD. It can also be used as a diagnostic tool for sanity check of the GWAS summary data from a meta-analysis to identify potential issues in the contributing data sets or the analysis pipeline. 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 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 errors in GWAS summary or LD reference 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 with 4 CPUs for all variants with MAF > 1% and <5h for all variants with MAF > 0.01% (**Supplementary Table 16**).

Methods

The DENTIST test-statistic

Given an ancestrally homogeneous sample and a genotype matrix $\mathbf{X}$ consisting of n unrelated individuals genotyped or imputed at m variants, an association analysis 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, $\mathbf{Z} \sim MVN(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{Z} = (Z_1, Z_2, \dots, Z_m)$, with $\mathbf{\Sigma}$ 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[12,13] , under the null hypothesis of no association, 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

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

where $\mathbf{\Sigma}_{it}$ denotes the correlation of z-scores between variant i from S1 and variants $\boldsymbol{t}$ from S2, and $\mathbf{\Sigma}_{tt}$ is the correlation matrix of variants $\boldsymbol{t}$. We use the correlation matrix calculated from an ancestry-matched reference sample (denoted by $\mathbf{R}$) 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

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

As shown in previous studies[12,13], $E(Z_i|\mathbf{z_t})$ can be used to predict Z_i, i.e., $\tilde{z}_i = \mathbf{R}_{it}\mathbf{R}_{tt}^{-1}\boldsymbol{z_t}$. In a special case when there is only one variant in vector $\boldsymbol{t}$, this model is the same as the conditional model in CAVIAR[11]. Hence, we can use the 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. These are likely to be the major but not the only factors that cause a discrepancy between the observed and imputed z-statistics, and other factors such as covariate-adjustment may also lead to a difference.

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[12,13], the method is derived under the null hypothesis without imposing any specific assumption about the genetic architecture of the trait, although our simulations show that the method performs well under the alternative hypothesis (**Supplementary Figure 1**). The exact model (not practically usable) for predicting the z-statistic under the alternative and the difference between the exact and approximate models can be found in the **Supplementary Note**. We show by numerical calculation that under the alternative, the test-statistics computed based on the approximate model were highly consistent with those based on the exact model (**Supplementary Figure 14**). Furthermore, we assessed DENTIST under a host of MAF- and LD-dependent architectures, ranging from oligogenic (m_{causal} = 50) to highly polygenic models (m_{causal} = 15,000, 13.1% of all the variants) and found that the DENTIST test was still well calibrated (**Supplementary Figure 15**).

To improve computational stability, we do not include variants in near-perfect LD with variant *i* (e.g., $r^2 > 0.95$) in the variant set for predicting $\tilde{z}_i$. However, even by doing so, the LD correlation matrix $\mathbf{R}_{\boldsymbol{tt}}$ can be rank-deficient. Hence, we use a truncated singular value decomposition (SVD) approach[15] to compute a pseudoinverse of $\mathbf{R}_{\boldsymbol{tt}}$, i.e., performing an eigen decomposition of $\mathbf{R}_{\boldsymbol{tt}}$ and retaining only the first *k* eigenvalues ranked by eigenvalues.

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

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

$\mathbf{R}_{tt}^{+}$ denotes the pseudo inversion of $\mathbf{R}_{\boldsymbol{tt}}$. The scalars $w_1, \ldots, w_k$ correspond to the largest *k* eigenvalues, and vectors $\boldsymbol{v_1}, \ldots, \boldsymbol{v_k}$ are the corresponding *k* eigenvectors. Given q = rank ($\mathbf{R}_{\boldsymbol{tt}}$), 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**).

The iterative variant partitioning approach

A challenge in applying **Equation 3** to real data is that errors in $\mathbf{z}_{\boldsymbol{t}}$ or discrepancy between $\mathbf{R}_{\boldsymbol{it}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{it}}$ can affect the accuracy of predicting $\tilde{z}_i$. We propose an iterative variant partitioning approach to mitigate this issue. 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 compute a T_d statistic for each of the variants. To mitigate the problem that variants with higher GWAS test statistics tend to have higher DENTIST test statistics, we stratify all the variants into two groups by a GWAS cutoff p-value of 0.01 and rank the variants by P_{GWAS} from the smallest to the largest in either group. We first apply a P_{DENTIST} threshold of 5×10^{-8} to variants with $P_{\text{GWAS}} > 0.01$ and compute the proportion of variants removed (denoted by x%).

528 We apply the same P_{DENTIST} threshold to variants with $P_{\text{GWAS}} < 0.01$ but capped the proportion of
529 variants removed at the top x%. We impose an additional constraint that the proportion of
530 variants removed is capped at the top 0.5% in either group. The default number of iterations is
531 set to 10. In this iterative process, variants with large errors or LD heterogeneity between the
532 discovery and LD reference samples are prioritized for removal in the first few iterations so that
533 the prediction accuracy increases in the following iterations. After the iterations are completed,
534 we remove all variants with $P_{\text{DENTIST}} < 5\times10^{-8}$ in the final step without the constraints.
535

536 **Genotype data sets**

537 This study is approved by the University of Queensland Human Research Ethics Committee
538 (approval number: 2011001173) and the Westlake University Ethics Committee (approval
539 number: 20200722YJ001). A summary of the genotype data sets used in this study as well as their
540 relevant information can be found in **Supplementary Table 1**. These data are from four GWAS
541 cohorts of European descendants, including the Health Retirement Study (HRS)[34], Atherosclerosis
542 Risk in Communities (ARIC) study[35], UK10K[29], and UK Biobank (UKB)[31]. The samples were
543 genotyped using either WGS or SNP array technology (**Supplementary Table 1**). Imputation of
544 the UKB data had been performed in a previous study[51] using the Haplotype Reference
545 Consortium (HRC)[36] and UK10K reference panels[31,52]. We used different subsets of the imputed
546 UKB data as the LD reference in this study, denoted with the prefix "UKBv3", such as UKBv3-unrel
547 (all the unrelated individuals of European ancestry, n = 348,577), UKBv3-329K (a subset of
548 328,577 individuals of UKBv3-unrel), UKBv3-20K (another subset of 20,000 individuals of
549 UKBv3-unrel, independent of UKBv3-392K) and UKBv3-8K (a subset of 8,000 individuals of
550 UKBv3-20K). HRS, ARIC and UK10K cohorts were imputed to the 1KGP reference panel in prior
551 studies[30,32], and a subset of 8,000 unrelated individuals from UKB were imputed to the 1KGP
552 reference panel in this study (referred to as UKB-8K-1KGP). The UK10K variants in common with
553 those on an Illumina CoreExome array were used for 1KGP imputation[32]. The imputation dosage
554 values were converted to best-guess genotypes in all the data sets except for UKBv3-all, in which
555 the hard-called genotypes were converted from the imputation dosage values using PLINK2 --
556 hard-call-threshold 0.1 (Ref[53]). For all the data sets, standard QC was performed to remove
557 variants with HWE test P-value < 10^{-6}, imputation INFO score <0.3, or MAF < 0.001. Since the hard-
558 called genotypes had missing values, in UKBv3 and its subsets, we further removed variants with
559 missingness rate > 0.05.
560

561 **Genome-wide association analysis for height using the UKB data**

562 We performed a genome-wide association analysis for height using the genotype data of UKBv3-
563 329K, i.e., all the unrelated individuals of European ancestry in the UKB (n = 328,577) except for

20,000 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[54] 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 11**).

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), the Sylvia & Charles Viertel Charitable Foundation, and the Westlake Education Foundation. This study makes use of data from the UK Biobank (project ID: 12514), HRS (dbGaP accession: phs000428), ARIC (dbGaP accession: phs000280) and UK10K (EGA accessions: EGAS00001000108 and EGAS00001000090). A full list of acknowledgements to these data sets can be found in the **Supplementary Note**.

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. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).
10. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273-1300 (2020).
11. Hormozdiari, F., Kostem, E., Kang, E.Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497-508 (2014).
12. Pasaniuc, B. *et al.* Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics* **30**, 2906-14 (2014).
13. 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).
14. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
15. 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).
16. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228 (2015).
17. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun* **9**, 224 (2018).
18. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
19. 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).
20. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
21. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).
22. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-52 (2016).
23. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
24. Vilhjalmsson, B.J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am J Hum Genet* **97**, 576-92 (2015).
25. Lloyd-Jones, L.R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat Commun* **10**, 5086 (2019).

26. 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).
27. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
28. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98 (2008).
29. UK10K consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
30. 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).
31. 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).
32. 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).
33. 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).
34. Sonnega, A. *et al.* Cohort profile: the health and retirement study (HRS). *Int J Epidemiol* **43**, 576-585 (2014).
35. ARIC INVESTIGATORS. The atherosclerosis risk in communit (aric) stui) y: Design and objectwes. *American Journal of Epidemiology* **129**, 687-702 (1989).
36. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279 (2016).
37. 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).
38. 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).
39. 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).
40. 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).
41. 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).
42. 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).
43. Phelan, C.M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* **49**, 680-691 (2017).
44. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92-94 (2017).
45. 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).
46. Wu, Y. *et al.* Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat Commun* **9**, 918 (2018).
47. Lloyd-Jones, L.R. *et al.* The Genetic Architecture of Gene Expression in Peripheral Blood. *Am J Hum Genet* **100**, 228-237 (2017).
48. 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).
49. Robinson, M.R. *et al.* Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1**(2017).

50. Benner, C. *et al.* Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *The American Journal of Human Genetics* **101**, 539-551 (2017).
51. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
52. Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature communications* **6**, 1-9 (2015).
53. 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).
54. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics* **51**, 1749-1755 (2019).

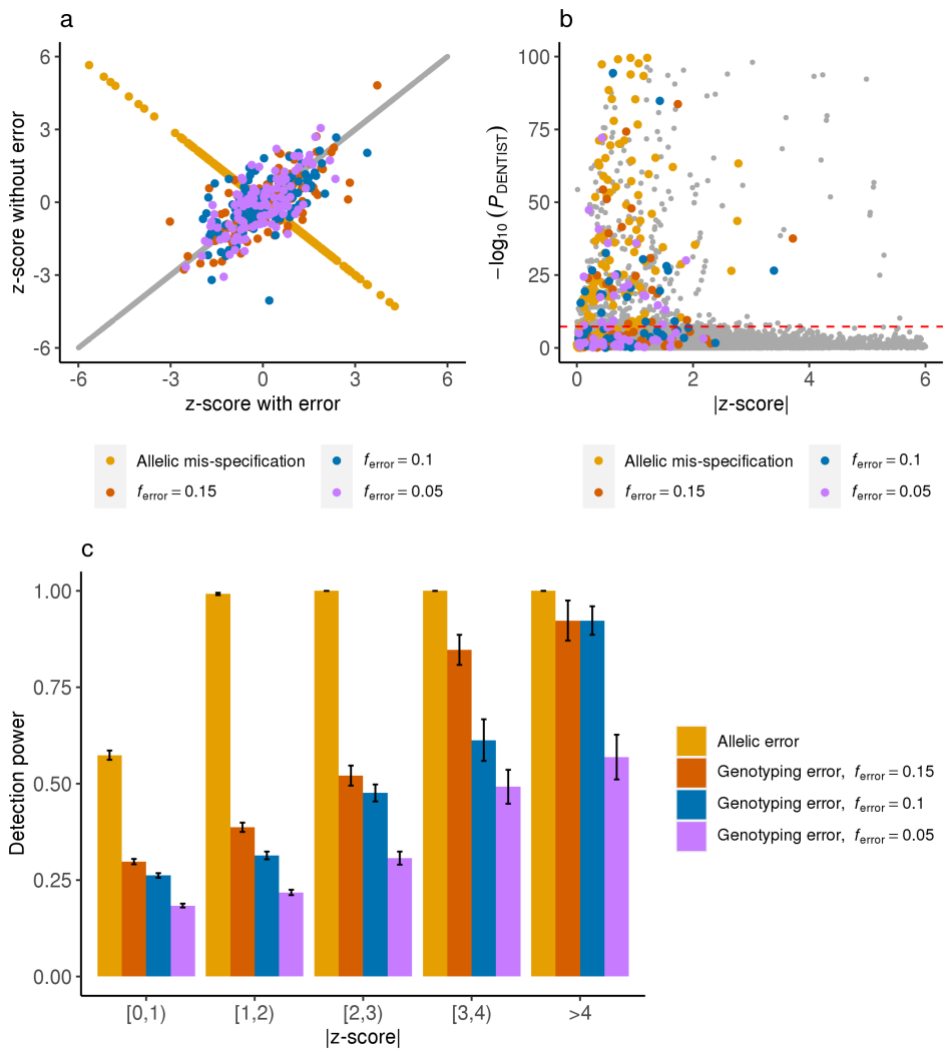

Figure 1. Detecting simulated allelic and genotyping errors using DENTIST. We assessed the power of DENTIST in detecting allelic and genotyping errors by simulation. There are three levels of genotyping error rate (f_{error}=0.15, 0.1 or 0.05), defined as the proportion of individuals with erroneous 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.

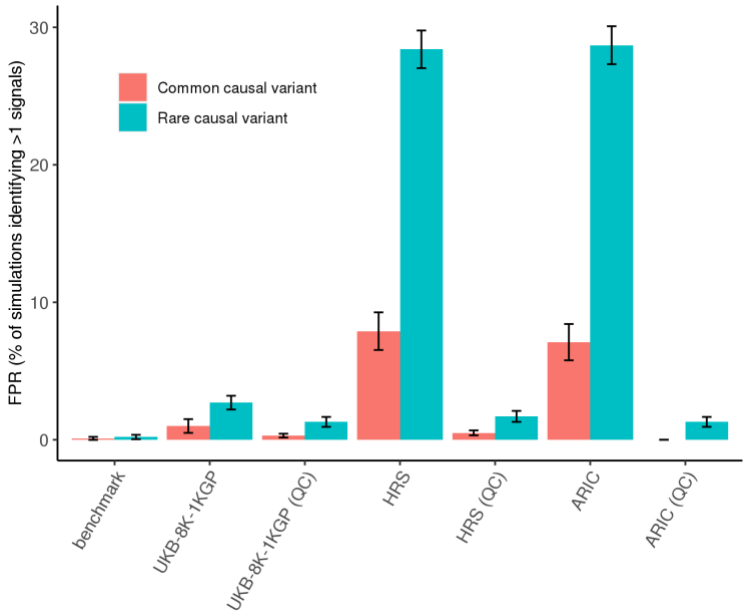

Figure 2. FPR of COJO with and without DENTIST. Based on simulations with one causal variant, we assessed the FPR of COJO when performed with and without DENTIST-based QC (FPR is defined as the frequency of observing more than one COJO signal in the scenario where only one causal variant was simulated). The x-axis labels indicate the LD reference samples, and those performed after DENTIST QC are labeled with "QC" in the parentheses. Each error bar represents the standard error of the FPR calculated from 2,200 simulation replicates with the causal variant resampled in each replicate.

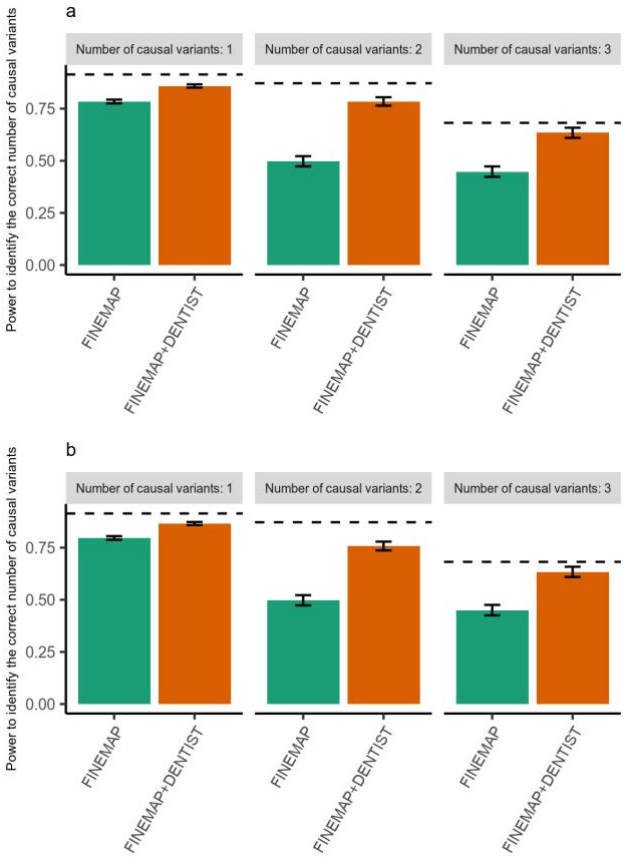

Figure 3. Power of FINEMAP with and without DENTIST QC. The power of FINEMAP is measured by the percentage of simulation replicates in which the number of causal variants is correctly identified under the simulation scenario with 1, 2 or 3 causal variants. Shown in panels **a** and **b** are results from using out-of-sample LD from HRS and ARIC, respectively. The horizontal dashed line in each bar plot shows the benchmark, i.e., the power of FINEMAP when using in-sample LD.

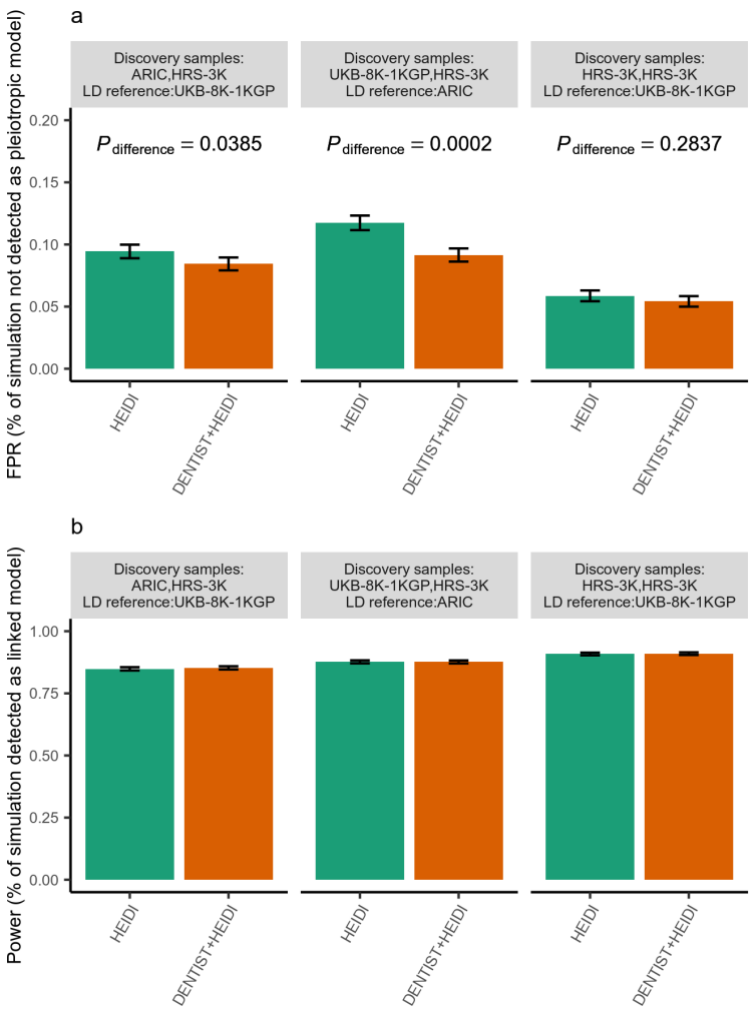

Figure 4. FPR and power of the HEIDI test with and without DENTIST QC. Shown are the results from simulations to quantify the FPR of HEIDI under a pleiotropic model (panel **a**) and the power of HEIDI under a linkage model (panel **b**). The text highlighted in gray on top of each bar plot illustrates the cohorts used in the simulation study, i.e., the GWAS discovery sample, the eQTL discovery sample, and the LD reference sample. $P_{\text{difference}}$ 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 (k_T, p) with p from a prior distribution of $p \sim Beta(k_2, k_T - k_2)$, where k_T is the overall 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 represent the standard errors of the corresponding metrics calculated from 4,000 simulation replicates with the causal variants re-sampled in each replicate.

761 **Table 1.** FPR 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	30.6±3.0
COJO with DENTIST (UKB-8K-1KGP)	0.2 ± 0.13	1.63 ± 0.43	80.0 ± 1.2	29.4±1.6
COJO without DENTIST (HRS)	7.9 ± 1.37	28.4 ± 1.37	78.5 ± 1.2	27.6±2.9
COJO with DENTIST (HRS)	0.5 ± 0.18	2.68 ± 0.55	78.8 ± 1.2	23.0±1.6
COJO without DENTIST (ARIC)	7.1 ± 1.32	28.7 ± 1.38	77.6±0.9	26.3±2.9
COJO with DENTIST (ARIC)	0.4 ± 0.17	3.50 ± 0.63	76.8 ± 1.0	23.2±1.6

762 Benchmark: COJO analysis using the discovery GWAS as the reference without DENTIST. Shown
763 are mean ± standard error.
764

765 **Table 2.** Numbers of COJO signals from analyses of the UKB height summary data using different
766 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	1,279	1,296 (1.3%)	1,337 (4.5%)	1,477 (15.5%)	1,485 (16.1%)	1,417 (10.8%)
MAF > 0.01 With DENTIST	/	1,288 (0.7%)	1,326 (3.6%)	1,344 (5.0%)	1,372 (7.2%)	1,419 (10.9%)
MAF > 0.001 Without DENTIST	1,310	1,313 (0.2%)	1,337 (2.0%)	1,555 (18.7%)	1,645 (25.6%)	1,473 (12.4%)
MAF > 0.001 With DENTIST	/	1,324 (1.0%)	1,325 (1.1%)	1,362 (3.9%)	1,368 (4.4%)	1,422 (8.5%)

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

769

770