Genealogy based trait association with LOCATER boosts power at loci with allelic heterogeneity

Xinxin Wang,[1,2,3] Ryan Christ,[1,2] Erica Young,[4,5,+] Chul Joo Kang,[4] Indraniel Das,[5,++] Edward A. Belter Jr.,[5,+++] Markku Laakso,[6] Louis J.M. Aslett,[7] David Steinsaltz,[8] Nathan O. Stitziel,[4,9] Ira M. Hall[1,2,*]

1. Department of Genetics, Yale University, New Haven, CT, 06520, USA
2. Center for Genomic Health, Yale University, New Haven, CT, 06520, USA
3. Roy and Diana Vagelos Division of Biology and Biomedical Sciences, Washington University School of Medicine, Saint Louis, MO, 63110, USA.
4. Center for Cardiovascular Research, Division of Cardiology, Department of Medicine, Washington University School of Medicine, Saint Louis, MO, 63110, USA.
5. McDonnell Genome Institute, Washington University School of Medicine, Saint Louis, MO, 63108, USA
6. Institute of Clinical Medicine, Internal Medicine, University of Eastern Finland, Kuopio, FI-70211, Finland
7. Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, United Kingdom
8. Department of Statistics, University of Oxford, OX1 4BH, United Kingdom
9. Department of Genetics, Washington University School of Medicine, Saint Louis, MO, 63110, USA.

+ Current affiliation: Food and Drug Administration

++ Current affiliation: Institute for Informatics, Data Science & Biostatistics, Washington University in St. Louis, Saint Louis, MO, 63110, USA

+++ Current affiliation: Collaborative and Integrative Genomics Lab at the McDonnell Genome Institute at Washington University School of Medicine, Saint Louis, MO, 63108, USA

*Correspondence: ira.hall@yale.edu

ABSTRACT

A key methodological challenge for genome wide association studies is how to leverage haplotype diversity and allelic heterogeneity to improve trait association power, especially in noncoding regions where it is difficult to predict variant impacts and define functional units for variant aggregation. Genealogy-based association methods have the potential to bridge this gap by testing combinations of common and rare haplotypes based purely on their ancestral relationships. In parallel work we developed an efficient local ancestry inference engine and a novel statistical method (LOCATER) for combining signals present on different branches of a locus specific haplotype tree. Here, we developed a genome-wide LOCATER analysis pipeline and applied it to a genome sequencing study of 6,795 Finnish individuals with 101 cardiometabolic traits and 18.9 million autosomal variants. We identified 351 significant trait associations at 47 genomic loci and found that LOCATER boosted single marker test (SMT) association power at 5 loci by combining independent signals from distinct alleles. LOCATER successfully recovered known quantitative trait loci not found by SMT, including *LIPG*, recovered known allelic heterogeneity at the *APOE/C1/C4/C2* gene cluster, and suggested one novel association. We find that confounders have a more pronounced effect on genealogy-based methods than SMT; we propose a new randomization approach and a general method for genomic control to eliminate their effects. This study demonstrates that genealogy-based methods such as LOCATER excel when multiple causal variants are present and suggests that their application to larger and more diverse cohorts will be fruitful.

INTRODUCTION

Genome-wide association studies (GWAS) have been extremely successful at identifying variants and genes associated with common human diseases and other complex traits. The vast majority of studies have used one or both of two statistical methods for trait association: common variant association using single marker tests (SMT) and rare variant association using gene-based aggregation tests[1,2]. However, neither method is well suited for identifying rare variant associations in noncoding regions where most causal variants are known to reside[3], or for testing the combined association of multiple independent common and/or rare variant signals (potentially with opposing effects) in cases of allelic heterogeneity. Region-based methods have been adapted to noncoding regions using sliding window approaches, with some success[4], but this approach is limited by two major challenges. First, it is difficult to decide which intervals and sets of variants to test in noncoding regions where knowledge of variant function is limited. Second, including nonfunctional variants in these tests can greatly reduce power.

In theory, genealogy-based methods that seek to associate local ancestral clades with traits have the potential to overcome these limitations through their ability to combine independent and potentially opposing signals present in different regions of the local ancestral tree, without the need to define functional regions or variant sets. Despite notable early progress[5,6], these methods have proven difficult to implement in practice due to the computational challenges of genome-wide population-scale haplotype inference and the statistical challenges of tree-based association testing. Recent advances in haplotype inference have eased the computational burden of building local genealogies[7–11], making genome-wide trait association studies feasible (albeit still computationally expensive). Building on this, three new methods have recently been developed to test local genealogies for trait association[8,12,13].

The first, ARG-Needle[8], builds local genealogies using a scalable ARG-based algorithm, samples a set of inferred clades that may harbor an unobserved variant, and adds the genotypes corresponding to those clades to the list of genotypes tested in genome-wide single marker testing (SMT). Although ARG-Needle is an extremely powerful method for reference free imputation, this "inferred variant" SMT approach is not designed to combine independent genetic signals at loci with allelic heterogeneity.

The second method, developed by Link et al.[12], uses the Li and Stephens (LS) model HMM implemented in Relate[9] to generate local expected genetic relatedness matrices (eGRMs) which are then tested for association with the phenotype. Link et al. showed via simulations that this approach can boost power when multiple independent causal variants are present at locus, and in their analysis of two chromosomes in a Native Hawaiian cohort of 5,384 individuals found evidence for a robust (albeit not genome-wide significant) BMI association signal with allelic heterogeneity. Shortly before the submission of this study, a preprint by Gunnarsson et al.[14] (accompanied by Zhu et al.[15]) proposed a scalable local ancestry inference engine and a statistical testing approach similar to Link et al.[12] However, since their approaches are based purely on a quadratic form class test statistic, they struggle against the enormous multiple-testing burden incurred by attempting to test all of the clades in a local tree.

The third method, LOCATER[13], is a comprehensive genome-wide screening procedure that uses a novel statistical method to test for the association of local genealogical relationships with traits. In addition to employing a quadratic form (QForm) association test similar to the method used in Link et al.[12], LOCATER also applies a novel statistic test that we developed, Stable Distillation (SD)[16], that is much more powerful than QForm at assessing the combined effects of many small clades marked by ultra-rare variants. When used in conjunction with an optimized implementation of the LS model we developed to infer local ancestry, kalis[11], our simulation studies have shown that LOCATER

outperforms SMT in cases of allelic heterogeneity (i.e., when multiple causal variants are present at a locus). Furthermore, that this advantage is more pronounced as the number of causal variants increases and their allele frequency decreases, and that the SD sub-test is the primary driver of power gains[13].

Recent evidence suggests that allelic heterogeneity is widespread in humans. For example, a recent massive parallel reporter assay (MPRA) study estimated that 10%~20% of expression quantitative trait loci (eQTLs) have multiple causal variants in humans[17], and a previous study showed that by inference, the proportion of all loci with allelic heterogeneity is 4-23% in eQTLs, 35% in GWASs of high-density lipoprotein (HDL), and 23% in GWASs of schizophrenia[18]. These and related observations[19] suggest that methods such as LOCATER that leverage allelic heterogeneity to improve power have the potential to discover novel trait associated alleles and genes not found by other methods.

Here, we use LOCATER to screen for trait-associated loci in the METSIM cohort. METSIM is a population sampled cohort of Finnish men with whole-genome sequencing data and a large number of cardiometabolic traits. Prior genome-wide association studies in METSIM have mapped many loci associated with cardiometabolic traits and disease, including studies based on array-based genotype data[20,21], exome sequencing data[22], and whole-genome sequencing data[23,24]. Due to a recent population bottleneck and subsequent expansion, genetic diversity is somewhat reduced in Finland and there is a larger fraction of deleterious variants at intermediate allele frequencies, facilitating trait mapping at relatively modest sample sizes[22,25]. These features, coupled with extensive prior knowledge of Finnish genetics, make this an ideal cohort to test new trait mapping methods such as LOCATER.

RESULTS

Applying LOCATER in this study required overcoming two key methodological challenges that were not addressed in our prior simulation-based work[13]: tuning our ancestry inference engine for use with real-world whole-genome sequencing (WGS) data, specifically from the METSIM cohort in Finland, and fully accounting for the effects of cryptic confounders in haplotype screening. Although the specific details of these analysis steps will need to be worked out for any new association study that uses LOCATER or similar methods on a new dataset, the approaches we describe below may provide a general solution that helps guide future implementation of these methods.

Parameter tuning for local ancestry inference

Although the LOCATER software can in principle use any ancestry inference engine as the substrate for genealogy based trait association, the LOCATER pipeline used for our prior and current work uses the Speidel version[9] of the LS model implemented in kalis[11], which includes recent developments to improve efficiency for genome-wide testing[13]. As described in Aslett & Christ[11], the LS model in kalis has two parameters: the recombination scale parameter (N_e), and the mutation probability (μ). We will refer to $-\log_{10}N_e$ as the recombination penalty parameter and $-\log_{10}\mu$ as the mutation penalty parameter.

Tuning these parameters has been a focus of recent work on the LS model[9,26]. Rather than using the data likelihood and expectation-maximization (EM)[27] to select these two parameters for our METSIM analysis, Relate[9] proposes using a more relevant objective aimed at maximizing the performance of the LS model for their specific purpose: capturing local variants in their inferred ancestral trees. Here, we propose an objective aimed at optimizing the discovery power of LOCATER and other methods aiming to leverage allelic heterogeneity. LOCATER boosts the power of SMT at a given locus by incorporating additional signals from nearby independent causal variants using Stable Distillation (SD)

149 and a quadratic form test (QForm)[13], so we selected HMM parameters that optimize the propagation of
150 nearby signals using the following objective function and sampling scheme.
151
152 We randomly sampled many genomic regions from the dataset and assigned a causal variant and a
153 target variant that are 0.05 cM away (see Methods for details). We then simulated a phenotype vector
154 with a strong effect driven by the causal variant and ran LOCATER at the target variant to measure
155 how well it could capture the signal driven by the nearby causal variant. We calculated relative
156 efficiency as our metric, defined as

157 $$\frac{-log(p_{SD}^{l} p_{Q}^{l})}{-log(p_{SMT}^{m})},$$

158 where l is the index of the target variant and m is the index of the causal variant. p_{SMT} refers to p-
159 values returned by SMT, p_{SD} and p_{Q} refer to p-values returned by the SD and QForm sub-tests in
160 LOCATER respectively. This objective focused on association signal propagation could be used to
161 train parameters for any association method targeting loci with allelic heterogeneity.
162
163 The trimmed mean surface for this relative efficiency showed that multiple parameters have
164 comparable efficiency and formed a plateau (Figure S1) (See Supplementary Methods, see Table S3
165 for HMM parameters evaluated). Our result aligns with a similar finding in Speidel et al.[9] that high
166 mutation penalties combined with low recombination penalties are not well suited for haplotype
167 inference. After ruling out parameters that require a much longer time to run, we randomly picked one
168 parameter (recombination penalty of 6 and a mutation penalty of 8) on the plateau of the surface,
169 averaging across allele frequency bins. We also generated surfaces for different allele frequency bins
170 and confirmed that the shape of the surfaces remains consistent across allele frequency bins for both
171 the QForm and SD association methods.
172
173 **Methodological improvements to the LOCATER pipeline to account for cryptic confounders**

174 In our initial trait association experiments we observed deviations in the empirical p-value distributions
175 from LOCATER, with cases of both mild inflation and mild deflation across the 101 traits analyzed in
176 this study, despite using standard procedures to correct for population structure using PCA. In
177 contrast, we did not observe inflation when running SMT on the exact same data, nor did we observe
178 inflation when running LOCATER on simulated phenotype data. This suggests that genealogy-based
179 trait association methods are especially sensitive to the effects of cryptic confounders. We believe that
180 this is due to the fact that there is a much greater degree of correlation between nearby genetic
181 markers for tree-based tests than for SMT, and that this causes a much larger fraction of markers to
182 be affected by confounders. Notably, this implies that these confounders are also affecting the SMT
183 results, just in a less readily detectable way.
184
185 To more precisely calibrate LOCATER results, we took inspiration from the phenotype rank-matching
186 procedure used in LOCATER. By default, LOCATER normalizes phenotypes by mapping ranks to
187 simulated Gaussian random variables rather than to fixed quantiles[13]. This approach avoids the subtle
188 dependence induced when mapping to fixed quantiles. Under the assumption that the original
189 residuals are exchangeable, matching to simulated Gaussian random variables indeed yields
190 independent Gaussian phenotypes, which is required by the SD procedure underlying LOCATER.
191
192 Building on this procedure, we repeated the rank-matching process for the same phenotype several
193 times, each version of the phenotype vector is a subtly different perturbation of the original phenotype.
194 We found that when we ran LOCATER across these different perturbations, the amount of inflation
195 observed via Q-Q plots differed moderately across perturbations. We believe this variation reflects the
196 fact that, by chance, different perturbations can have stronger or weaker correlations with
197 confounding factors such as population structure. Thus, the rank-matching procedure allows us to
198 modulate the correlation between the phenotype and confounding variables such that we can identify

199 a rank-matched version that is the least correlated with confounders. We therefore rank matched each
200 phenotype 100 times and chose the version that showed the least alignment with confounders for later
201 adjustment (see Methods). For each version of each phenotype, we conducted an association test for
202 ~30,000 evenly spaced variants and selected the version for which the p-value distribution is the
203 closest to a uniform distribution for both SD and QForm (see Methods). We found that SMT p-values
204 of all versions are consistently well calibrated, and SD and QForm p-values from different versions will
205 have substantially different deviations from the theoretical null (Figures S2, S3 and S4, column "Rank
206 matched chosen", "Rank matched 2" and "Rank matched 3"). We also found that the rank-matched
207 phenotype chosen for the association study indeed has the least deviation and eliminated the inflated
208 body of the QForm p-value distribution. Our study is the first to report genome-wide Q-Q plots for a
209 genealogy-based association method and demonstrate calibration of that method after rigorously
210 adjust for unobserved confounding factors.
211
212 We then sought to apply a more general genomic control to the p-values. In short, we fit a line to the
213 Q-Q plot of LOCATER $-\log_{10}p$ of the selected rank-matched phenotype. We adjusted p-values using
214 the slope and intercept of the fitted line (see Methods). This is similar to traditional genomic control[28],
215 which only fits a slope to the Q-Q plot. By incorporating a non-zero intercept, we can achieve a much
216 more accurate fit to the tail of our null distribution of $-\log_{10}p$. In order to gain a better intuition for the
217 role of this intercept, it is helpful to think in terms of a simple application like single-marker testing
218 where each p-value corresponds to a Z-score. In this setting, fitting a slope via classic genomic
219 control to the Q-Q plot is equivalent to fitting a scale parameter to the distribution of null Z-scores
220 while keeping the location of that null distribution fixed at zero. Introducing a non-zero intercept allows
221 one to fit a null distribution to the Z-scores where the location may also be adjusted. A similar
222 procedure is described in Chapter 6 of Efron[29] and is standard in large-scale testing problems where
223 the null assumptions are not satisfied (e.g. when there is confounding). In Figures S2, S3 and S4, we

224 applied this adjustment procedure to p-value distributions from LOCATER-specific tests, and all
225 distributions align with the expected distribution much better (Figures S2, S3 and S4, row "SD
226 adjusted" and "QForm adjusted").
227
228 **GWAS of 101 traits**
229 We then performed an association analysis of 6,795 METSIM individuals with 101 quantitative
230 metabolic phenotypes (see Table 1 for highlighted associations) using LOCATER. We also analyzed
231 these traits with SMT as our benchmark. All traits were residualized based on trait-specific covariates
232 beforehand, exactly as described in our prior exome based study of this same cohort[22]. After filtering
233 out indels and variants lacking ancestral allele annotation (see Methods), we tested 18,949,137
234 autosomal variants. We used the top 10 Finnish-specific principal components as background
235 covariates for the association test. As mentioned above, during the pre-screening stage we selected
236 the phenotype vector perturbation that yielded the most calibrated Q-Q plot (see Methods). Post-
237 screening we adjusted the resulting p-values based on the slope and intercept (as described above,
238 see Methods) in the Q-Q plot to ensure that both SD and QForm sub-tests in LOCATER were well-
239 calibrated.
240
241 Considering that this genome sequencing dataset contains an abundance of rare variants, applying
242 the canonical genome-wide significance threshold (5×10^{-8}), which assumes one million independent
243 tests, is not appropriate. We used permuted phenotypes to estimate the effective number of
244 independent tests[30,31] (see Methods). Based on the distribution of minimum p-values, we estimated
245 that the effective number of independent tests for SMT (T_{SMT}) is 6,977,438, and for LOCATER
246 ($T_{LOCATER}$) is 3,551,616 ($\alpha = 0.05$). Based on T_{SMT}, the genome-wide significance threshold for SMT
247 should be 7.17×10^{-9}. Note that LOCATER and SMT used the exact same set of variants, but that the
248 p-values reported by LOCATER are more dependent across loci due to shared local genealogies. For

249 a clearer comparison with SMT in visualizations (e.g., Fig 2A), we further standardized the LOCATER
250 p-value with

251 $$p_{LOCATER\ standardized} = p_{LOCATER\ adjusted} * T_{LOCATER}/T_{SMT}$$

252 to make the LOCATER genome-wide Bonferroni threshold match that of SMT (7.17×10^{-9}). This does
253 not change the interpretation of LOCATER results since both the results and the significance cutoff
254 were rescaled to the same extent. To enable a direct comparison between SMT and LOCATOR, all
255 LOCATER results reported below are standardized.
256
257 After the screening, we identified loci of interest as genomic regions with one or more variants with
258 significant p-values by either LOCATER or SMT in any of the 101 traits. To compare LOCATER and
259 SMT signals at these loci, we defined a common shared association interval by merging the set of
260 significantly associated variants identified by either method, where the merging includes a 600 kb
261 flanking region for each variant. For convenience, we used a similar process to merge association
262 intervals across traits to obtain a nonredundant set of genomic regions, although in this case, we note
263 that independent signals for different traits may often be lumped together (see Table S2 for the full
264 set). Altogether, we identified 47 genomic regions with 351 associations across all traits.
265
266 When comparing the most significant signal for LOCATER and SMT at every identified association
267 (identified as min($p_{LOCATER\ standardized}$) < threshold or min($p_{SMT}$) < threshold), we found that LOCATER
268 and SMT identified many associations together (327 out of 351, Figure 1D). Many of these are in
269 canonical regions known to be associated with cardiovascular diseases, such as *PCSK9*, *APOB*, *LPL*,
270 *LIPC*, *CETP*, and the *APOE/C1/C4/C2* gene cluster. A small number of associations are significant
271 only in LOCATER (7 out of 351), and SMT found 17 associations that LOCATER did not (Figure 1B).
272

273 Notably, in the cases where SMT is more significant than LOCATER (321 out of 351 associations), it is
274 typically by a very small margin, whereas in the cases where LOCATER is more significant (30 out of
275 351), it is typically by a relatively substantial margin (Figure 1C). Our interpretation of this result is that
276 LOCATER has less power than SMT at trait associations resulting from a single causal variant because
277 of the statistical penalty incurred by attempting to incorporate nearby signals, but that LOCATER
278 greatly outperforms SMT at loci with multiple causal variants. In total, 5 of the 47 significant loci
279 (10.6%) show a signal boost from LOCATER, indicating that allelic heterogeneity is fairly common,
280 even in a relatively small sample of the Finnish population that is known to be depleted of genetic
281 diversity relative to most other human populations due to historical bottlenecks.
282
283 For the 30 associations with a more significant signal from LOCATER, we inferred the number of
284 independent causal variants based on the Stable Distillation (SD) and quadratic form (QForm) sub-test
285 signals from LOCATER (Table S5). These 30 associations reside in 13 distinct genomic regions. After
286 clumping associations that reside in the same genomic region (with 600 kb flanking regions) and
287 involve traits that are highly correlated to each other (r^2>0.8) in our dataset, there were 21
288 nonredundant associations. Of these, the LOCATER signal boost came from SD in 15 cases, and from
289 QForm in 6 cases.
290
291 Of the 15 nonredundant associations that were boosted by SD, 7 of them were boosted by only one
292 sprig (independent predictors, defined as the smallest possible inferred clades), while the other 8 of
293 them had 2-12 significant sprigs contributing to the signal. All significant sprigs represent distinct
294 haplotype groups in the local ancestry trees. As a result, the total number of inferred causal variants
295 for these 15 associations ranges from 2 to 13. We also report the variants that are completely linked
296 with significant sprigs in Table S4. For the six associations boosted by QForm, we used iterative
297 conditional analysis to infer the number of causal variants. One association became insignificant after

298 accounting for the lead marker, and signals in 5 other associations were diminished after multiple (2-5)
299 rounds of conditional analysis, suggesting that the total number of inferred causal variants for these
300 six associations ranges from 1-5.
301
302 Rare variant association methods such as STAAR[32] also seek to assess the combined effects of
303 multiple causal variants at a locus, albeit using a very different approach than LOCATER. To test
304 whether STAAR could potentially detect the same signals as LOCATER, we ran STAAR (without
305 variant annotations) on the 30 associations mentioned above for which LOCATER had a more
306 significant p-value than SMT. Notably, STAAR did not detect any of these 30 associations at genome-
307 wide significance. It is important to note that there are many different ways to run STAAR based on
308 window size, variant inclusion, variant annotation and weighting criteria, and so we cannot discount
309 the possibility that STAAR might be able to detect some of these signals. Nonetheless, these results
310 suggest that the trait association signals detected by LOCATER are not easily captured by current rare
311 variant association methods such as STAAR.
312
313 Below we discuss some of the trait associations detected by LOCATER. We highlight five known
314 association signals that LOCATER detected but SMT did not, three cases where both LOCATER and
315 SMT detected the association but LOCATER provided a substantial boost in signal strength, and two
316 potentially novel association signals detected solely by LOCATER.
317
318 **LOCATER recovers known associations at the *LIPG* locus**
319 LOCATER recovered several known quantitative trait loci that would otherwise have been missed by
320 SMT in our present study (see Figure 1B), an example being the *LIPG* locus. *LIPG* encodes a well-
321 known member of the triglyceride lipase family of proteins and is primarily involved in the metabolism
322 of HDL[33–36]. LOCATER recovered genome-wide significant associations for triglycerides in medium

323 HDL (p=1.68 x 10^{-9}) and apolipoprotein A-I (p=6.51 x 10^{-9}), the major protein component of HDL
324 particles. These two trait associations are likely to be independent given that the lead markers are in
325 low LD (r^2 of 7.29 x 10^{-4}) and the two traits are not significantly correlated (Pearson correlation of
326 0.132), which is consistent with prior work[21]. Neither of these associations was captured by SMT at
327 genome-wide significance. The smallest SMT p-value for triglycerides in medium HDL within a 1.2 Mb
328 window of *LIPG* was 4.71x10^{-7}, and that for apolipoprotein A-I was 2.15x10^{-8} (Table 1). The lead
329 variant for apolipoprotein A-I was also found in a prior study of Finns (p=2 x 10^{-10}) using many of the
330 same METSIM samples analyzed here[21], and the lead variant for triglycerides in medium HDL was
331 found in a large study of 233 metabolic traits in 33 cohorts[20].
332
333 We first discuss the *LIPG* association with triglycerides in medium HDL, where LOCATER detected a
334 significant signal but SMT did not (Figure 2A). LOCATER's improved power over SMT in this case
335 comes from the Stable Distillation (SD) sub-test (Figure S5C, D), which indicates contributions from
336 multiple ultra-rare causal variants. We confirmed that the p-value distribution after adjustment aligns
337 very well with the expected distribution, and the QQ-plot-based adjustment required to control
338 confounding was minimal (Figure S5A, B).
339
340 The LOCATER SD sub-test has the advantage that p-values from all predictors (sprigs, defined as the
341 smallest possible inferred clades) are independent under the null hypothesis. A Q-Q plot of all -$\log_{10}$
342 sprig p-values calculated at the lead marker shows that the top 12 sprigs significantly deviated from
343 the expected distribution and thus contributed to the SD signal (Figure 2B). To highlight the coalescent
344 path of significant haplotypes, we plotted a dendrogram of haplotypes based on hierarchical
345 clustering of the similarity matrix at the lead marker, which showed that different significant sprigs
346 were present within distinct clades in the local coalescent tree, and that haplotypes in the same sprig
347 were very similar. Haplotypes in the same sprig have a very recent coalescence, and haplotypes from

348 different sprigs have a much more distant coalescence (Figure 2D). This suggests that the SD sub-test
349 captured signals from multiple distinct haplotype groups rather than multiple signals driven by a single
350 variant.
351
352 Notably, SD is able to combine signals from individuals at both extremes of the phenotype
353 distribution, which correspond to alleles with opposing effects. As expected, all samples within
354 significant sprigs have phenotype values that are far from the mean (Figure 2C). Individuals within
355 each sprig resided on the same side of the distribution, but different sprigs could reside on different
356 sides.
357
358 We next sought to visualize variants influencing the SD signal at this locus using residual analysis. The
359 first phase involved conducting SMT based on a phenotype that projected out the lead marker
360 genotype vector, removing its effect on the signal. We defined the p-values resulting from this
361 experiment as p_S. The second phase involved performing SMT based on the residualized phenotype
362 orthogonal against the lead marker SMT and SD signal, yielding a second set of p-values defined as
363 p_D. The difference between p_S and p_D shows the contribution of genomic variants to the SD signal. We
364 plotted $-\log_{10} p_S$ and $-\log_{10} p_D$ on a Manhattan plot (Figure S5E) and a scatter plot (Figure S5F),
365 highlighting variants where $p_S < 1 \times 10^{-3}$ and $p_D > 10 * p_S$. This experiment shows that SD captured
366 signals from many variants scattered in an extensive genomic region (> 10 Mb), supporting the
367 success of our ancestry inference model tuning procedure.
368
369 By plotting principal components (PCs), we confirmed that SD-significant individuals do not form a
370 tight cluster in any specific area of the plot, so this association signal was not obviously confounded
371 by population structure (Figure S5G).
372

373 We now turn to the second *LIPG* association, apolipoprotein A-I. LIPG is known to regulate serum
374 apolipoprotein A-I[33,37]. A previous study in METSIM[21] showed that our lead marker is associated with
375 five HDL subclass traits and apolipoprotein A-I (Figure 3A), and in this study, LOCATER recovered the
376 signal in apolipoprotein A-I, but SMT missed it.
377
378 LOCATER gained its advantage over SMT from SD (Figure S6A). The Q-Q inflation plot of sprig $-\log_{10}$
379 p-values showed that 6 sprigs contributed to the signal (Figure S6B). Similar to the first *LIPG*
380 association result discussed above, haplotypes from the same sprig have a very recent coalescence
381 and those from different sprigs have a more distant coalescence (Figure S6C), and samples within
382 outlying sprigs had phenotypes that were far away from the median and on different sides of the
383 distribution (Figure 3B).
384
385 In summary, LOCATER was able to discover two independent trait associations at *LIPG* based on the
386 presence of allelic heterogeneity, where both trait associations were missed by SMT. In addition to
387 these two examples at *LIPG*, three additional trait associations at other known loci were also detected
388 by LOCATER but not SMT. The association with HDL2 cholesterol on chr11 (Figure S13) and the
389 association of monounsaturated fatty acids (MUFA) on chr7 (Figure S9) were also identified based on
390 SD signals and also showed consistent phenotype values across relevant individuals. Both
391 associations were boosted by only one sprig, and phenotype values of individuals assigned to
392 significant sprigs are all far from the mean of the corresponding phenotype distribution. There was
393 also an association with 'large VLDL particle concentration' on chromosome 7 (Figure S11) that was
394 boosted by QForm; however, this association is somewhat less confident than others given that only
395 one variant crossed the significance threshold and that a substantial post-hoc adjustment for
396 confounding was required to calibrate the corresponding QQ-plot. .
397

398 **A potentially new association on chr11**

399 LOCATER found an association of 'triglycerides in medium VLDL' on chromosome 11, while SMT did
400 not (Figure 4A), and the GWAS catalog[38] did not report any known association with correlated traits.
401 LOCATER was much more significant than SMT due to the SD sub-test, implying a contribution from
402 ultra-rare haplotypes (Figure S7C). The SD signal was distributed evenly across the entire ~1 Mb locus
403 (Figure 4B). Among all sprigs called at the lead marker position, the top five sprigs significantly
404 deviated from the expected distribution (Figure 4C). The phenotype distribution of individuals in
405 significant sprigs showed that they are outliers, and that trait outliers are found at both extremes of the
406 distribution (Figure 4D). The dendrogram from hierarchical clustering of the local similarity matrix, with
407 highlighted coalescent paths for significant haplotypes, showed that the signal was not coming from a
408 larger clade but rather from multiple distinct small groups of haplotypes (Figure 4E). These results
409 suggest that LOCATER is combining association signals from 5 rare haplotype groups with distinct
410 genealogical histories, while these signals were not detectable by standard SMT.

411

412 We performed residual analyses to investigate the contribution of genomic variants to the SD signal
413 and noticed many variants with drastically different p_S and p_D (i.e., substantial contribution to the SD
414 signal). These variants extend across a mega-base long genomic distance, highlighting LOCATER's
415 ability to merge sub-significant signals from a large genomic region.

416

417 In addition, we detected a second potentially novel association with HDL3 cholesterol on chr21 (Figure
418 S14). This association was attributable to SD combining signals from a single sprig. However, in this
419 case, only one variant is significant, and even though the lead marker is genotyped in gnomAD, it
420 resides in a 305bp recent segmental duplication within which variant calling is likely to be prone to
421 artifacts, making this result less confident than others presented here.

422

423 **LOCATER boosted a classic association at the apolipoprotein gene cluster**

424 LOCATER recovered a known association for remnant cholesterol at the well studied *APOE/C1/C4/C2*
425 gene cluster on chromosome 19[20,39]. This same locus is associated with various other traits that are
426 correlated with remnant cholesterol (Table S2). Although SMT also achieved genome-wide
427 significance for remnant cholesterol, LOCATER's advantage over SMT implies that there are additional
428 signals from other haplotypes (Figure 5A).

429

430 In contrast to the other examples outlined above, LOCATER's advantage over SMT in this case came
431 from QForm (Figure 5B, Figure S8C), indicating that the causal haplotypes are likely to be more
432 common (i.e., not ultra-rare). Unlike the SD subtest, QForm does not inherently provide direct insight
433 into the number of distinct haplotypes and how they relate to each other in the genealogy; however,
434 using multiple rounds of conditional analysis with SMT, we confirmed that there are at least four
435 groups of causal variants. We iteratively conditioned on the genotype vector of lead markers, and
436 observed that a significant association signal ($p < 3.52 \times 10^{-5}$) persisted through three rounds of
437 conditional analyses (Figure 5C, D). We found that the variants used as covariates in the conditional
438 analyses are not in LD with each other, and one of them (chr19-44908822-C-T) is the most significant
439 known marker associated with remnant cholesterol (Figure 5D, black arrow). These results show that
440 LOCATER effectively combined signals from four distinct causal haplotypes, resulting in a substantial
441 boost in power.

442

443 There are also two additional examples of known loci found by both LOCATER and SMT, where
444 LOCATER has a more significant p-value (implying the presence of multiple causal variants): the
445 association with 'triglycerides in small HDL' on chromosome 20 (Figure S10) and the association of
446 triglycerides in VLDL on chromosome 7 (Figure S12). In both of these cases the power boost was
447 driven by the SD sub-test.

448

449 **DISCUSSION**

450

451 We have used our new genealogy-based trait association method, LOCATER, to perform a genome-
452 wide screen in a cohort of 6,795 Finnish individuals with deep cardiometabolic trait measurements and
453 whole-genome sequencing data. In total we identified 30 associations at 13 known GWAS loci at
454 which LOCATER was genome-wide significant and provided a clear power boost over SMT, 7
455 associations of which (at 5 loci) were not genome-wide significant by SMT and would have been
456 missed. LOCATER also identified two novel association signals, one of which is fairly compelling
457 based on the underlying haplotype structure. At each locus, dissection of the association signals and
458 underlying haplotype structure revealed evidence for allelic heterogeneity in the form of multiple
459 independent association signals present in distinct portions of the local ancestry tree. Moreover, in the
460 process of optimizing LOCATER's performance on real world genomic data, we made several key
461 methodological improvements, including a novel approach for tuning ancestry inference parameters
462 for trait association and a rigorous approach to account for the effects of cryptic confounders.

463

464 Genealogy based trait association has been a topic of interest for nearly two decades. Seminal early
465 work established its potential value using theory, simulations, and single-locus analysis[5,6], yet the
466 practical advantages of these methods are only now becoming accessible due to recent advances in
467 scalable tree inference, "clade association" methods capable of testing unobserved variants inferred
468 by reference-free imputation, and "global tree association" methods capable of combining association
469 signals across multiple causal variants[12,13,16]. Our work makes several key contributions that promise to
470 push the field forward. In particular, our genome-wide analysis of 101 traits across 6,795 individuals
471 greatly exceeds the scale and statistical power of the two prior efforts[8,12] that applied global tree
472 association to human data, both of which focused on smaller cohorts, a single trait, and a subset of

473 the genome – a single locus in one case, and two chromosomes in the other – and failed to detect any
474 trait associations at genome-wide significance. Two recent preprints[14,15] from the same group
475 proposed a genome-wide ancestral recombination graph inference engine and performed a gene-
476 based association study with multiple traits in the UK Biobank. This method is promising in terms of
477 improving power for gene-based testing; however, it remains unclear whether the approach is well
478 calibrated on real-world data with population structure, and whether it can be corrected for cryptic
479 confounders. To our knowledge, ours is the first genome-wide study to use global tree association,
480 the first to use Stable Distillation, and the first to identify genome-wide significant trait associations
481 with rigorous empirical correction for cryptic confounders. In the process, we encountered and
482 overcame several key methodological obstacles related to parameter tuning and statistical calibration
483 that have not previously been examined at this level of detail. These lessons are applicable to any
484 haplotype association method, not just LOCATER, and thus this work may serve as a roadmap for
485 future efforts. More specifically, this work provides a detailed framework for how to perform genome-
486 wide screens using LOCATER, the only software of which we are aware that is capable of genome-
487 wide global tree association in large cohorts.
488
489 Our work also demonstrates that genealogy based methods such as LOCATER show considerable
490 promise for increasing the power of genetic association studies. LOCATER's ability to combine
491 multiple signals improved power over SMT at 8.5% of genome-wide significant associations and for
492 10.6% of loci, and the difference in p-value exceeded an order of magnitude difference in 7 of these
493 associations at 5 loci. The fact that such notable power gains are observed in the Finnish population –
494 one of the least diverse human populations studied to date – suggests that our results are likely to be
495 an underestimate of the performance improvements possible in more diverse populations that harbor
496 more causal alleles per locus. LOCATER may be especially valuable for multi-ethnic studies, where

497 allelic heterogeneity is greatest, and where standard association methods have shown poor
498 performance.
499
500 Notably, few of the 30 association signals for which LOCATER provided a power boost could have
501 been captured by other existing methods. None of these signals were captured by STAAR using the
502 window-based screening approach that has been employed previously for association studies,
503 indicating that genealogy-based methods provide unique value. Moreover, although two recently
504 developed methods[12,14,15] have some similarities with LOCATER, they almost certainly would not have
505 been able to detect most of the 30 associations reported here. LOCATER employs two distinct
506 statistical tests, QForm and Stable Distillation (SD)[16], of which only the former corresponds to the
507 testing employed by the other two methods[12,14,15]. That test (QForm) contributed minimal benefit above
508 SMT in the analyses presented here, and in simulated data[13]; most of the power gains come from the
509 unique ability of SD to efficiently combine association signals across many clades. This suggests that
510 LOCATER's use of the recently developed SD test makes it uniquely powerful for haplotype
511 association. Despite these promising results, global tree association remains a difficult statistical
512 problem, and we expect that future work in this area will continue to yield significant power
513 improvements.
514
515 Tree-based trait association also presents unique challenges relative to SMT and gene-based testing.
516 Perhaps the most difficult aspect of this study was the detailed work required to optimize haplotype
517 inference and to control for cryptic confounders, both of which are important practical considerations
518 for future studies. First, it is well known that the performance of the LS model HMMs is sensitive to the
519 mutation and recombination penalty parameters, and prior studies have used different approaches for
520 selecting them. We developed a novel tuning approach that optimizes regional trait association power
521 at a distance of 0.05 cM rather than local variant imputation as in Relate[9], and we explored a broad

522 range of potential parameters to optimize haplotype inference for this specific population and dataset.
523 Optimizing for trait association power rather than imputation ensures that the resulting haplotype
524 representations used carry long-range information required to combine independent signals at loci
525 with allelic heterogeneity.
526
527 Second, we found that genealogy-based methods such as LOCATER are extremely sensitive to
528 cryptic confounders, much more so than standard SMT, and that special measures are required to
529 control for these effects. We studied this issue in detail and devised two new approaches to control
530 for type I error: permuted rank matching of phenotype data to simulated Gaussian random variables to
531 produce independent Gaussian phenotypes with minimal confounding, and a more general type of
532 genomic control that fits both a slope and intercept. In combination, these measures led to a well
533 calibrated analysis in our study and are likely to be applicable to future genealogy based screens as
534 well. Interestingly, the confounding effect of cryptic confounders was only apparent in our LOCATER
535 analysis, whereas SMT appeared to be well calibrated after standard PCA-based measures. We
536 hypothesize that LOCATER is more sensitive to confounders than SMT because of the high correlation
537 between proximal inferred genealogies along the genome. For intuition, consider the simple case
538 where there is a small sub-population within a dataset that has some environmental exposure that
539 affects the phenotype of interest. There will be variants that tag that sub-population throughout the
540 genome. Every local ancestral tree inferred near those confounded variants will have a clade marking
541 that sub-population and thus have inflated test statistics. Thus, any potential confounders will affect
542 LOCATER at many more markers than SMT, making the resulting inflation obvious in Q-Q plots.
543 Importantly, these effects are expected to be equally problematic for SMT in terms of false positives, if
544 more difficult to discern; thus, a SMT study may appear to be well calibrated even while it has false
545 positives, especially in rare variants, due to latent sub-populations or other unobserved
546 confounders[40,41].

547

548 Although there is more work to be done before these methods are mature, the work presented here –
549 in combination with our prior simulation based results[13] and recent work from others[12,14,15] – suggests
550 that genealogy-based trait association methods such as LOCATER are finally ready to fulfill their long-
551 promised potential as practical tools for genome-wide association studies.

552

553 **METHODS**

554

555 **The METSIM study**

556 METSIM is a single-site study investigating cardiometabolic disorders and related traits in 10,197 men
557 randomly selected from the population register of Kuopio, Eastern Finland, aged 45 to 73 years at
558 initial examination from 2005 to 2010[42]. All participants provided informed consent. The phenotype
559 data used for the work described here was adapted from Locke et al.[22], in which the authors
560 accounted for different factors (including trait specific background covariates) during linear regressing
561 raw phenotypes, and used rank-based inverse normal transformation.

562

563 **Whole-genome sequencing, data processing, and variant calling**

564 DNA samples were extracted from blood. We constructed DNA libraries with automated Kapa Hyper
565 PCR free, automated TruSeq PCR free, Kapa Hyper PCR free, or TruSeq PCR free kit, with target
566 insert size varying from 260 to 475. The libraries were sequenced with Illumina HiSeq 2000, 2500, X10,
567 or NovaSeq 6000, generating 2 x 151-bp paired-end sequencing data.

568

569 We performed alignment and data processing using the "functional equivalence" pipeline[43]. Briefly, we
570 aligned reads to the GRCh38 human reference genome using bwa-mem (v.0.7.15) and used Picard
571 MarkDuplicates (v.2.4.1; http://broadinstitute.github.io/picard) to remove duplicate reads. We excluded

572 samples that had estimated contamination >5% or that were likely to represent sample swaps

573 (verifyBamID v.1.1.3). We also required a discordant rate of <5%, haploid coverage ≥19.5X, inter-

574 chromosomal rate of <5%, and first-of-pair mismatch rate of <5%.

575

576 We performed variant calling with HaplotypeCaller in GATK (v.3.5) and concatenated the output into
577 full single sample GVCFs (Picard MergeVcfs, v2.4.1). Since standard joint genotyping with GATK
578 GenotypeGVCFs function could not scale to our CCDG callset, we used ReblockGVCFs (GATK
579 v.4.2.2.0) to decrease GVCF file sizes for future joint analysis in Hail. We used the ValidateVariants
580 function to ensure the quality of the reblocking process. We then used the VariantDatasetCombiner
581 function in Hail (v.0.2.78) to combine GVCFs from each sample into multi-sample VariantDataset (VDS)
582 files before running GnarlyGenotyper (unpublished version from Docker image gcr.io/broad-dsde-
583 methods/gnarly_genotyper:hail_ukbb_300K, image hash ID: 7cc8cfa6e9af; created April 2020;
584 received August 2021) for joint genotyping and VQSR to annotate variant quality. Finally, we converted
585 VDS files into MatrixTable (MT) files (Hail v.0.2.97) and decomposed multi-allelic variants into bi-allelic
586 variants.

587

588 To QC the variant callset, we excluded samples that had a low het/hom ratio (<5 MADs less than the
589 median), low sequencing depth (number of bases with depth >10 is <20 MADs less than the median),
590 or excessive number of singleton variants (>20 MADs more than the median), where each of these
591 criteria was applied separately to each self-reported ancestry group. We also removed samples with
592 fewer than 580,000 insertions or deletions in joint variant calling, genetic-phenotypic sex mismatches,
593 withdrawal of consent, sample swaps, or inheritance inconsistencies and other sample identity issues.
594 The maximal independent set of these samples was calculated in Hail using
595 `hl.maximal_independent_set` and individuals up to second-degree related were removed. For variant

level QC, we flagged genotype calls with genotype quality (GQ) < 20, depth (DP) < 10, and heterozygous calls with allele balance (AB) ≤ 0.2 or ≥ 0.8 as low quality, and filtered out variants that had AS_VQSLOD < 0 or that had a high proportion (>95%) of missing or low quality genotypes.

Phasing and variant annotation

We performed a more stringent variant quality control for phasing. After exporting Hail MT files to VCFs (Hail 0.2.95), we selected PASS and non-singleton variants and filtered out sites with high quality genotype call rate < 90% or Hardy-Weinberg $p < 10^{-7}$ (one-sided p-value for excess heterozygotes). METSIM samples were phased with other WashU CCDG samples without a reference panel with Eagle2 (v.2.4.1). Due to memory restraints, we divided chromosomes into 20 Mb chunks with 2 Mb overlaps on both ends. Phasing was done with default options except for a bigger `Kpbwt` value (200k) for better phasing accuracy. Missing genotypes were imputed during phasing.

Ancestral allele encoding

The Speidel version of the LS model[9] assumes the ancestral allele is known for all variants and uses that information in its ancestry inference. For this study, we used ancestral allele calls obtained via a 10-way EPO alignment of primates from Ensembl v106 [44]. With `bcftools`, we updated the REF and ALT allele and the genotype fields in VCF files to make the REF allele the ancestral allele. Encoding VCFs with the ancestral allele in large-scale data requires ancestral allele calls in FASTA format. Explicitly, we downloaded the human ancestral genome FASTA from 10-way EPO alignment of primates from Ensembl v106 available at https://ftp.ensembl.org/pub/release-106/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh38.tar.gz created on March 19, 2022. In the FASTA file, lowercase indicated lower quality. For simplicity, all lowercase were converted to

620 uppercase. Since the conversion works best when multiallelic variants are merged, we merged them
621 with `bcftools norm -m +any` (bcftools v1.9). We then used `bcftools norm --check-ref s --fasta-ref
622 {fasta_file} ` (bcftools v1.9) to edit the REF allele in VCF files to be ancestral allele, which automatically
623 updated the genotype field. The files also went through `bcftools +fill-tags` (bcftools v1.16) to make
624 sure the AC tag of the INFO column in VCF files is correct. Finally, we split multiallelic sites into
625 biallelic variants with `bcftools norm -m -any` (bcftools v1.16).
626

627 **The LOCATER pipeline**

628 We performed ancestry-based association testing using the LOCATER pipeline, which is described in
629 detail elsewhere, including methodological details and an in-depth evaluation of simulated data[13]. The
630 detailed description of the real data pipeline of this study is described here: https://github.com/Xinxin-
631 Wang-0128/LOCATER_real_data_vignette. Briefly, the first step is to run local ancestry inference at
632 each genetic marker included in the study. For this we used the newest version (v2) of our local
633 ancestry inference engine, kalis[11], which is an optimized implementation of the Speidel version of the
634 LS model[9]. Notably, unlike other LS model implementations, kalis v2 uses an optimal checkpointing
635 algorithm which allows it to be run on arbitrarily large sequences (e.g., whole chromosomes). Here, in
636 the interests of computational efficiency, we only performed local ancestry inference on genomic
637 segments that contained one or more single markers with a promising p-value ($P < 10^{-3}$ in this study).
638 In total, this included 5.7% of the genome. This step produces an N x N matrix (where N is the number
639 of haplotypes; 13,590 in this case) of genomic distances at each genetic marker.
640

641 The second step is to identify small clades, which we refer to as “sprigs”. LOCATER includes a sprig
642 calling algorithm that uses a multithreaded partial sorting algorithm to cluster the genomic distances
643 into level sets separately for each haplotype, followed by a greedy clique finding procedure to rapidly
644 call sprigs.

645

646 Then, at each genetic marker, LOCATER performs three types of association tests: (1) a standard SMT
647 to measure the contribution of that specific marker; (2) Stable Distillation (SD) to measure the
648 contribution of the inferred small clades; and (3) a quadratic-form based test to measure the
649 contribution of remaining ancestral relationships present at deeper portions of the tree[13]. Steps 2 and
650 3 use a residualized phenotype vector from the prior step. This, combined with the independence
651 guarantees of SD, ensures that the resulting three p-values at a given site are mutually independent
652 under the null hypothesis. We then combine the three p-values using an adapted version of Fisher's
653 method that we call Maximizing over Subsets of Summed Exponentials (MSSE)[13].

654

655 **Rank matching and selection**

656 The SD procedure used in LOCATER requires that all phenotypes have a Gaussian distribution. As we
657 show in Figures S2, S3 and S4, since the quadratic form testing procedure is downstream of SD,
658 violating this assumption can affect the null distribution of both SD and QForm. We ensured the
659 Gaussianity of our phenotype vector by rank-matching our phenotypes to a vector of simulated
660 independent Gaussian random variables.

661

662 During preliminary analyses, we repeated this rank-matching process for the same phenotype several
663 times, yielding a set of vectors, each a subtly different perturbation of the original phenotype. When
664 we ran LOCATER across these different perturbations, the amount of inflation observed via the Q-Q
665 plot of genome-wide LOCATER p-values differed moderately across perturbations (see examples in
666 Figures S2, S3, and S4). We believe this variation reflects the fact that, by chance, different
667 perturbations can have stronger or weaker correlations with confounding processes such as
668 population structure. Note that genome-wide SMT p-values are consistently well-calibrated for all
669 perturbations; thus, no adjustments were needed for SMT. Based on this observation, we simulated

100 perturbations of each phenotype and selected the perturbation that minimized the deviation from the expected tail distribution under the null hypothesis. Explicitly, for all perturbations of each phenotype, we ran LOCATER across evenly spaced variants (~30,000 variants) along the genome, and after plotting the Q-Q inflation plot for -$\log_{10}p$ of SD and QForm in LOCATER, we fitted a least-squares line over the $x \in$ [2,2.5] domain of the resulting Q-Q inflation plot. This corresponds to fitting the tail of the null distribution based on all p-values in [0.01,0.0032]. We then used the parameters of each least square line (the slope and intercept) to select the perturbation that was closest to the expected null distribution. More explicitly, let the slopes of SD and QForm Q-Q plots be m_{SD} and m_Q, while intercepts are b_{SD} and b_Q, respectively. First, we selected all perturbations satisfying the following boolean expression $m_{SD} \in$ [0.8, 1.1] AND $b_{SD} \in$ [-0.1, 0.1] AND (($m_Q \in$ [0.7, 1.2] AND $b_Q \in$ [-0.1, 0.1]) OR ($m_Q \in$ [0.6, 0.8] AND $b_Q \in$ [0, 0.4])). If multiple perturbations met this requirement, the perturbation with the largest $min(m_{SD}, m_Q)$ was chosen. If no perturbation met this requirement, which happened for 15 of our 101 phenotypes, the standard rank-based inverse normalized phenotype was used.

After the screening, we adjusted the p-value for each sub-test based on the estimated tail parameters, m and b, via

$$-log\left(P_{SD\ adjusted}\right) = \frac{-log(P_{SD}) - b_{SD}}{m_{SD}} \quad \text{and} \quad -log\left(P_{Q\ adjusted}\right) = \frac{-log(P_Q) - b_Q}{m_Q}.$$

Since SMT is consistently well-calibrated, no adjustments were applied to SMT p-values. To combine the resulting SMT, SD, and QForm p-values returned by LOCATER, we used the MSSE method in Christ et al.[13].

Tuning ancestry inference for trait association

691 The LS model at the heart of the ancestry inference we used in this study, kalis, is a hidden Markov
692 model (HMM) with two parameters that can be interpreted as tolerance for recombination and
693 mutation, respectively. As we will delineate below, rather than using expectation-maximization or other
694 more standard tuning objectives to select our recombination and mutation parameters, we chose
695 parameters to optimize the propagation of proximal association signals along the genome in order to
696 maximize LOCATER's power.
697
698 In our tuning procedure, we randomly sampled core genomic regions with at least 15,000 variants,
699 each with flanking regions of 5,000 variants on both sides. These flanking regions served as "burn-in"
700 regions to ensure accurate ancestry inference along the full length of the core region. We then
701 selected a variant in the middle of the core region as our target variant and used kalis to perform
702 ancestry inference at that site.
703
704 We then selected a causal variant 0.05 (± 0.005) cM away. In LOCATER, as in all GWAS studies, any
705 association signals driven by variants that are colinear with the background covariates are assumed to
706 be attributable to confounding processes. Thus, we only chose causal variants that were not colinear
707 with the background covariates (multiple $r^2 \leq 0.02$). We then simulated a pseudo-phenotype vector
708 with a strong effect driven by the causal variant. In order to ensure that the strength of this causal
709 variant's signal (in terms of the observed $-\log_{10}$ p-value) would be roughly consistent across
710 simulations, we increased the strength of the causal variant's effect as a function of the number of
711 clades inferred at the target locus: more null clades yield a higher multiple testing burden for the
712 causal variant to overcome.
713 Given our pseudo-phenotype vector, we ran LOCATER at our target variant *l*, yielding two p-values
714 ${p_{SD}}^l$ and ${p_Q}^l$. We assessed how these two p-values captured the signal at the nearby causal variant
715 using

$$\frac{-log(p_{SD}^{l} p_{Q}^{l})}{-log(p_{SMT}^{m})}$$

as our relative efficiency metric. Here p_{SMT}^{m} is the SMT p-value at the causal variant.

We calculated this relative efficiency metric for 14 parameter settings (see Table S3) across 144 distinct core regions sampled from our METSIM dataset. Altogether, this procedure allowed us to select LS HMM parameters to maximize LOCATER's power in this METSIM study.

Estimating the effective number of independent tests

Our dataset has a huge number of rare variants, and the canonical 5 x 10^{-8} threshold for p-value was based on 1 million independent tests. This effective number of tests was identified in the array genotyping era, and a couple of studies[45,46] argued that this was an overly generous threshold. There are papers describing methods to calculate the effective number of independent tests[30], and they agreed that the gold standard is the permutation-based method[31]. Permutation-based methods permute the phenotype N times and do the association analysis for those permuted phenotypes. Then the minimum p-value (p_{min}) was recorded for all these permuted phenotypes. If we want to control the type I error rate to be 0.05, the new threshold should be the 95th percentile of the p_{min} distribution. Here we used N = 1000 permutations and targeted a type I error rate of 0.05[30].

We simulated 1000 normally distributed quantitative traits, ran SMT first, and used the subset of variants with $p_{SMT} < 10^{-4.5}$ for LOCATER association. Based on the p_{min} distribution, the experiment-wise significance threshold should be 7.17 x 10^{-9} for p_{SMT} and 1.41 x 10^{-8} for $p_{LOCATER}$ if the type I error rate is 0.05. To directly compare LOCATER and SMT, the final p-values of LOCATER were standardized so that its effective number of tests matches with SMT and both tests use the same threshold 7.17 x 10^{-9}.

741 **Association Screening**

742 Since the Speidel version of the LS model requires ancestral allele calls[9], we only included SNPs with
743 ancestral allele calls in the 10-way EPO alignment of primates from Ensembl v106. Indels were ignored
744 due to their lower quality of phasing and variant calling. After removing monomorphic sites and
745 singletons, we obtained a final dataset of 18.9 million variants.

746

747 Kinship among individuals was calculated and only unrelated individuals were included. Ethnicity-
748 specific principal component (PC) outliers were also removed. This yielded a final dataset of 101 traits
749 and 6,795 individuals. Background covariates in this study were the top 10 principal components.

750

751 For computational efficiency, we adopted the following three strategies. First, we divided the whole
752 genome into 4,587 segments, with ± ~6,000 variants of overlap between segments. Second, while all
753 of our variants were used for ancestry inference by kalis, we only ran LOCATER on a relatively small
754 subset of target variants. We skipped variants where the SMT p-values from all phenotypes were
755 greater than 10^{-3} while ensuring that the recombination distance between consecutive target variants
756 was at most 0.1 cM. Third, we avoided any expensive eigendecomposition steps in calculating p_Q at
757 our target loci by using the Satterthwaite approximation to p_Q[13] in our first round of screening.

758

759 We identified any variant with a LOCATER combined p-value smaller than 7.17×10^{-8} in our first round
760 of screening as putative variants. We then merged putative variants in all the phenotypes to generate
761 putative loci (see main text) for follow-up. During association follow-up, we only focused on 10 cM
762 regions centered around putative loci, which doubly ensured reliable ancestry inference. During
763 LOCATER testing, we used partial eigendecomposition and Chi-square-based approximation on the
764 local relatedness matrix to obtain precise p_Q.

765

766 Because of the stochastic nature of the SD procedure, the p-values p_{SD} and p_Q returned by LOCATER
767 are a function of the seed of the R environment. Setting the same seed for all segments along the
768 genome is not recommended, since it will cause additional correlation between variants, which will
769 cause the p-value distribution to deviate from the expected uniform distribution. Instead, we strongly
770 suggest setting different seeds for different segments along the genome. However, for reproducibility
771 or for follow up experiments of the same region or segment, we suggest using the same seed across
772 experiments for consistent results.
773
774 To assess the novelty of the association results we used the GWAS catalog (April 22nd, 2024 version;
775 https://www.ebi.ac.uk/gwas/home)
776
777 **Rare variant association experiment with STAAR**
778 We applied STAAR with the sliding window method using default parameters. The window size is 2 kb
779 and the sliding step length is 1 kb. For each window, we included all rare variants with AF < 0.01 and
780 removed any window with less than two rare variants. We did not apply functional annotation data and
781 we reported the STAAR-O p-value.
782
783 **Hierarchical clustering and visualization of local distance matrices**
784 At loci with significant LOCATER associations driven by SD, we constructed a local tree based on the
785 local relatedness matrix obtained by kalis in order to understand the relative placement of the sprigs
786 driving that SD signal. Following the method used in Speidel et al[9], we did this using mean-based
787 hierarchical clustering (UPGMA) implemented in the `fastcluster` R package[47]. For plotting clarity, 95%
788 of haplotypes under insignificant sprigs were pruned from the displayed dendrograms.
789
790 **Residual analysis to visualize SD signals**

We employed a residual analysis procedure to visualize the association signals that contributed to the SD signal at a locus. Recall that LOCATER is a three-stage procedure where a phenotype vector is passed between each step. In order to isolate the signals extracted by SD, we made a local Manhattan plot regressing the phenotype vector passed to SD (above we refer to the resulting p-value at a given variant as p_S) and overlaid it with a Manhattan plot regressing the phenotype vector returned by SD (above we refer to the resulting p-value at a given variant as p_D). We highlighted variants where p_S <1 x 10^{-3} and p_D > 10 * p_S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ACKNOWLEDGEMENTS

RC and XW were supported by NIH grants R01HG013371 and UM1HG008853 to IH. LJMA was partially supported by the EPSRC research grant "PINCODE", reference EP/X028100/1, and UKRI grant, "OCEAN", reference EP/Y014650/1. DS was supported by BBSRC research grant BB/S001824/1. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

XW, RC and IMH conceived and designed the analysis. IMH and NOS designed the sequencing study. ML created the METSIM cohort. EY, CJK, ID, EABJ generated the variant callset, EY curated the phenotype data, XW, RC, LJMA and DS contributed analysis tools, and XW performed the analysis. XW, RC and IMH wrote the paper with input from NOS, DS and LJMA.

DATA AND CODE AVAILABILITY

Data are available from dbGaP (accession number for genotype data: phs001579; phenotype data: phs000752). The code used to perform the analyses in this paper is available on a GitHub page: https://github.com/Xinxin-Wang-0128/LOCATER_real_data_vignette. The link to the LOCATER software will be public once the corresponding study is published.

REFERENCES

1. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
2. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. Li, Z. *et al.* Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *Am. J. Hum. Genet.* **104**, 802–814 (2019).
5. Zöllner, S. & Pritchard, J. K. Coalescent-Based Association Mapping and Fine Mapping of Complex Trait Loci. *Genetics* **169**, 1071–1092 (2005).
6. Minichiello, M. J. & Durbin, R. Mapping Trait Loci by Use of Inferred Ancestral Recombination Graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
7. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
8. Zhang, B. C., Biddanda, A., Gunnarsson, Á. F., Cooper, F. & Palamara, P. F. Biobank-scale inference of ancestral recombination graphs enables genealogical analysis of complex traits. *Nat. Genet.* **55**, 768–776 (2023).
9. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).

10. Kelleher, J. *et al.* Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
11. Aslett, L. J. M. & Christ, R. R. kalis: a modern implementation of the Li & Stephens model for local ancestry inference in R. *BMC Bioinformatics* **25**, 86 (2024).
12. Link, V. *et al.* Tree-based QTL mapping with expected local genetic relatedness matrices. *Am. J. Hum. Genet.* **110**, 2077–2091 (2023).
13. Christ, R., Wang, X., Aslett, L. J. M., Steinsaltz, D. & Hall, I. Clade Distillation for Genome-wide Association Studies. *bioRxiv* 2024.09.30.615852 (2024) doi:10.1101/2024.09.30.615852.
14. Gunnarsson, Á. F. *et al.* A scalable approach for genome-wide inference of ancestral recombination graphs. *bioRxiv* 2024.08.31.610248 (2024) doi:10.1101/2024.08.31.610248.
15. Zhu, J. *et al.* Fast variance component analysis using large-scale ancestral recombination graphs. *bioRxiv* 2024.08.31.610262 (2024) doi:10.1101/2024.08.31.610262.
16. Christ, R., Hall, I. & Steinsaltz, D. Stable Distillation and High-Dimensional Hypothesis Testing. *arXiv [stat.ME]* (2022).
17. Abell, N. S. *et al.* Multiple causal variants underlie genetic associations in humans. *Science* **375**, 1247–1254 (2022).
18. Hormozdiari, F. *et al.* Widespread Allelic Heterogeneity in Complex Traits. *Am. J. Hum. Genet.* **100**, 789–802 (2017).
19. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
20. Karjalainen, M. K. *et al.* Genome-wide characterization of circulating metabolic biomarkers. *Nature* **628**, 130–138 (2024).
21. Davis, J. P. *et al.* Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in Finnish men from the METSIM study. *PLoS Genet.* **13**, e1007079 (2017).

866 22. Locke, A. E. *et al.* Exome sequencing of Finnish isolates enhances rare-variant association power.
867 *Nature* **572**, 323–328 (2019).
868 23. Chen, L. *et al.* Association of structural variation with cardiometabolic traits in Finns. *Am. J. Hum.*
869 *Genet.* **108**, 583–596 (2021).
870 24. Ganel, L. *et al.* Mitochondrial genome copy number measured by DNA sequencing in human
871 blood is strongly associated with metabolic traits via cell-type composition differences. *Hum.*
872 *Genomics* **15**, 34 (2021).
873 25. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753
874 (2009).
875 26. Jin, Y. & Terhorst, J. The solution surface of the Li-Stephens haplotype copying model.
876 *Algorithms Mol. Biol.* **18**, 12 (2023).
877 27. Baum, L. E., Petrie, T., Soules, G. & Weiss, N. A Maximization Technique Occurring in the
878 Statistical Analysis of Probabilistic Functions of Markov Chains. *aoms* **41**, 164–171 (1970).
879 28. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
880 29. Efron, B. *Large-Scale Inference*. (Cambridge University Press, 2010).
881 30. Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association
882 studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**, 361–369 (2008).
883 31. Hoh, J., Wille, A. & Ott, J. Trimming, Weighting, and Grouping SNPs in Human Case-Control
884 Association Studies. *Genome Res.* **11**, 2115–2119 (2001).
885 32. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare
886 variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**,
887 969–983 (2020).
888 33. Jaye, M. *et al.* A novel endothelial-derived lipase that modulates HDL metabolism. *Nat. Genet.* **21**,
889 424–428 (1999).
890 34. Jin, W., Millar, J. S., Broedl, U., Glick, J. M. & Rader, D. J. Inhibition of endothelial lipase causes

891 increased HDL cholesterol levels in vivo. *J. Clin. Invest.* **111**, 357–362 (2003).
892 35. Ma, K. *et al.* Endothelial lipase is a major genetic determinant for high-density lipoprotein
893 concentration, structure, and metabolism. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 2748–2753 (2003).
894 36. Strauss, J. G. *et al.* Endothelial cell-derived lipase mediates uptake and binding of high-density
895 lipoprotein (HDL) particles and the selective uptake of HDL-associated cholesterol esters
896 independent of its enzymic activity. *Biochem. J* **368**, 69–79 (2002).
897 37. Ishida, T. *et al.* Endothelial lipase is a major determinant of HDL level. *J. Clin. Invest.* **111**, 347–355
898 (2003).
899 38. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic*
900 *Acids Res.* **51**, D977–D985 (2023).
901 39. Richardson, T. G. *et al.* Characterising metabolomic signatures of lipid-modifying therapies
902 through drug target mendelian randomisation. *PLoS Biol.* **20**, e3001547 (2022).
903 40. Persyn, E., Redon, R., Bellanger, L. & Dina, C. The impact of a fine-scale population stratification
904 on rare variant association test results. *PLoS One* **13**, e0207677 (2018).
905 41. Bouaziz, M. *et al.* Controlling for human population stratification in rare variant association
906 studies. *Sci. Rep.* **11**, 19015 (2021).
907 42. Laakso, M. *et al.* The Metabolic Syndrome in Men study: a resource for studies of metabolic and
908 cardiovascular diseases. *J. Lipid Res.* **58**, 481–493 (2017).
909 43. Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables
910 harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
911 44. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
912 *Nature* **526**, 68–74 (2015).
913 45. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of
914 thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom.* **2**, 100168 (2022).
915 46. Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies

novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).

47. Müllner, D. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines forRandPython. *J. Stat. Softw.* **53**, (2013).

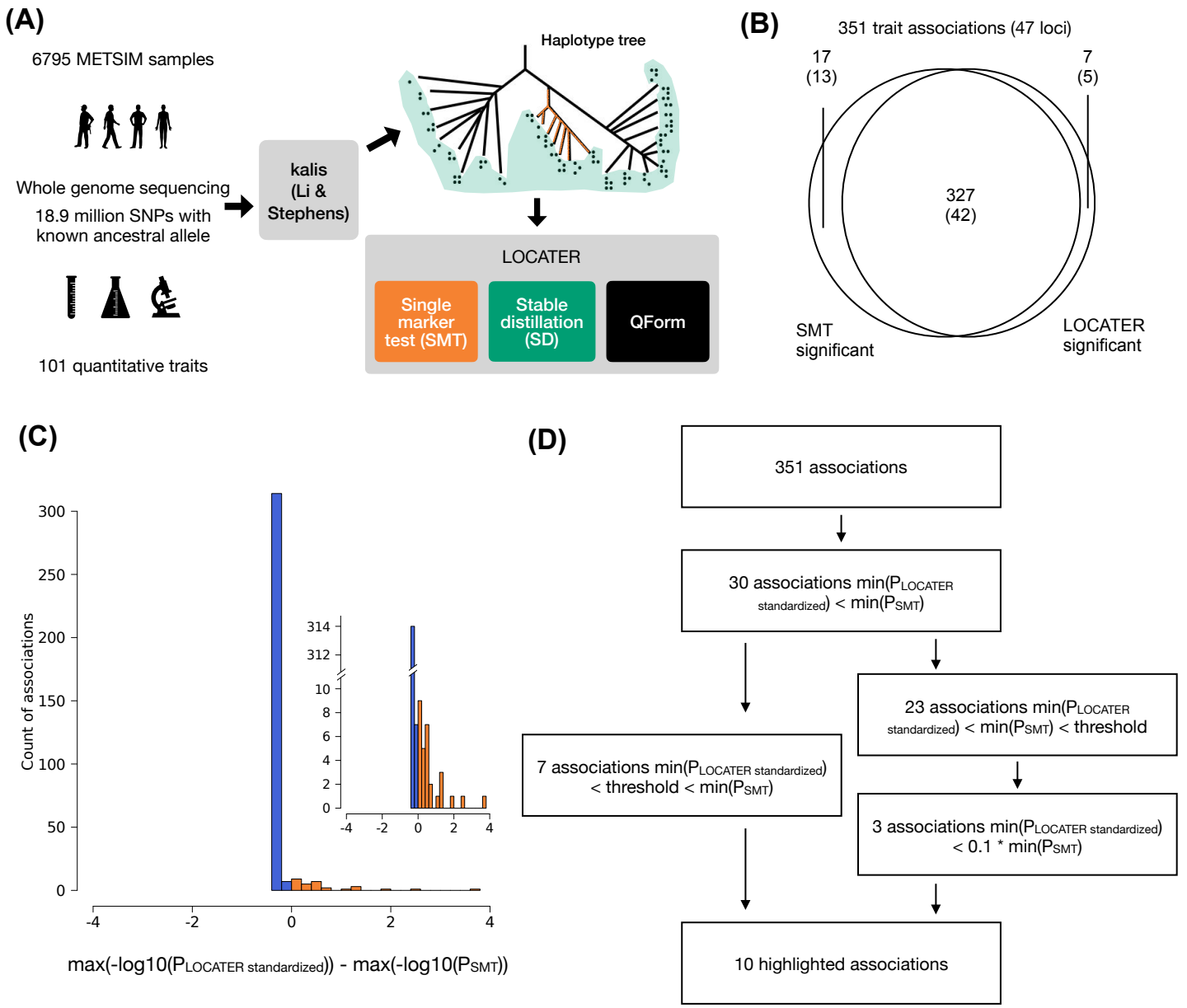


Figure 1. Schematic of our genealogy-based screening procedure, LOCATER, and summary of screening results. (A) Diagram of the experimental design and association testing method. Kalis is an implementation of the Li & Stephens model for local ancestry inference. QForm: quadratic form testing. (B) Venn diagram showing the number of associations and number of loci (in parentheses) with significant SMT and/or significant LOCATER results. Note that a given locus may have distinct associations represented in different parts of the Venn diagram. (C) Distribution of max(-log10($P_{LOCATER\ standardized}$)) - max(-log10($P_{SMT}$)) for all 351 associations. All associations that are genome-wide significant by either SMT or LOCATER are included. (D) Overview of significant associations and highlighted associations. The genome-wide significance threshold is 7.17 x 10^{-9} for SMT and standardized LOCATER.

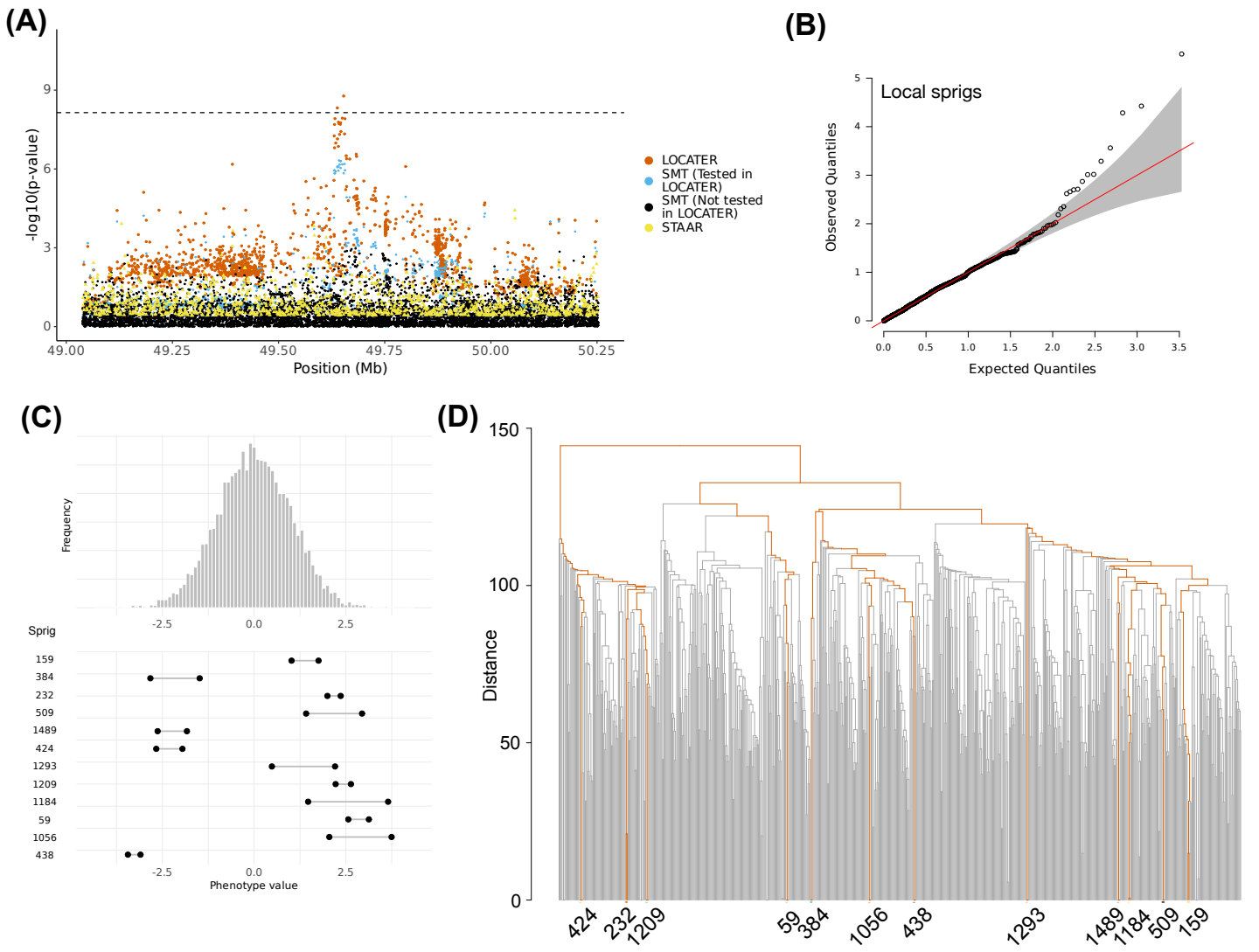
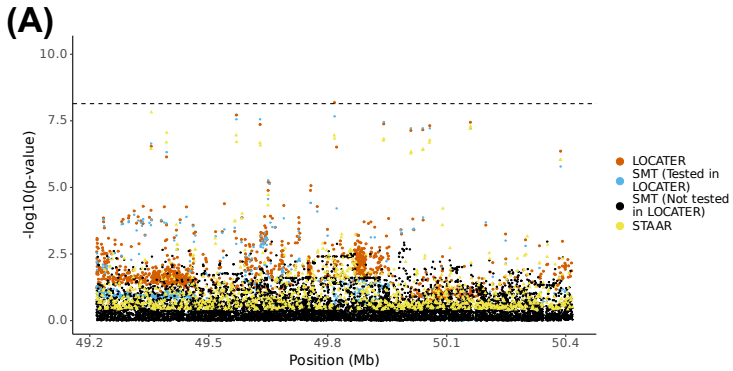


Figure 2. Association of triglycerides in medium HDL at LIPG locus. **(A)** Local Manhattan plot of the association signal for "triglycerides in medium HDL" on chr18:49038347-50253146, including results for single marker test (SMT; blue and black), LOCATER (orange) and STAAR (yellow). Note that LOCATER results are only shown for variants with an SMT p-value less than $1x10^{-3}$, since for computational efficiency only these variants were tested by LOCATER (see Methods). SMT results from variants tested by LOCATER are shown in blue, and those from variants not tested by LOCATER are shown in black. The black dashed line corresponds to the genome-wide significance threshold for SMT, standardized LOCATER, and standardized STAAR. **(B)** Q-Q inflation plot of -log10(p-values) from all "sprigs" at the lead marker chr18:49653146, where "sprigs" are defined as the smallest possible inferred clades. The gray area corresponds to the 95% confidence interval, and the red line denotes x=y. **(C)** Histogram of phenotype values after projecting out the genotype vector of the LOCATER lead marker (chr18:49653146), thus removing signal that can be accounted for by the SMT sub-test. Connected dots show the phenotype value of individuals assigned to significant sprigs. **(D)** Dendrogram generated from the haplotype-level local distance matrix at the lead marker chr18:49653146. The UPGMA method was used for hierarchical clustering. Orange branches highlight the path of all haplotypes in significant sprigs shown previously in **(C)**. Labels at the bottom show the sprig assignment. For plotting clarity, 95% of haplotypes under insignificant sprigs were pruned.



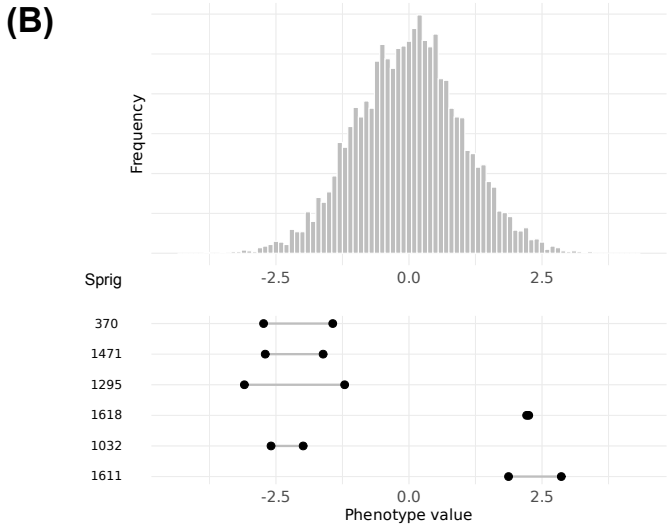


Figure 3. Association of apolipoprotein A1 at LIPG locus. **(A)** Local Manhattan plot of the association signal for apolipoprotein A1 on chr18:49217040-50417040, shown using the exact same data types and color scheme as Figure 2A. **(B)** Histogram of phenotype values after projecting out the genotype vector of the LOCATER lead marker (chr18:49817040), thus removing signal that can be accounted for by the SMT sub-test. Connected dots show the phenotype value of individuals assigned to significant sprigs.

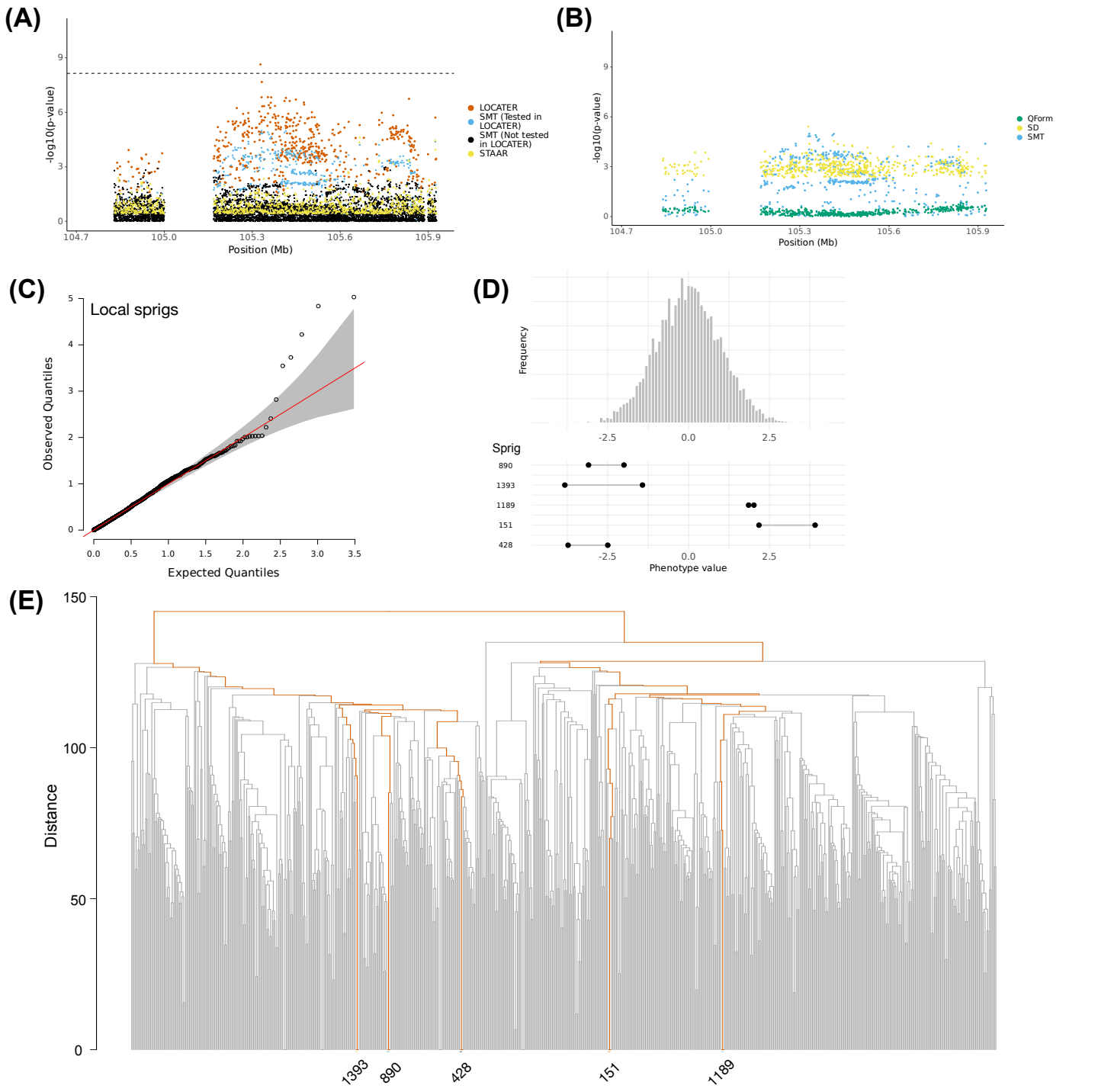


Figure 4. Association of triglycerides in medium VLDL on chr11. **(A)** Local Manhattan plot of the association signal for "triglycerides in medium VLDL" on chr11:104727888-105927888, shown using the exact same data types and color scheme as Figure 2A.**(B)** Local Manhattan plot of "triglycerides in medium VLDL" at chr11:104727888-105927888, showing adjusted -$\log_{10}$(P) for the 3 LOCATER sub-tests. **(C)** Q-Q inflation plot of -log10(p-values) from all "sprigs" at the lead marker chr11:105327888, where "sprigs" are defined as the smallest possible inferred clades. The gray area corresponds to the 95% confidence interval, and the red line denotes x=y. **(D)** Histogram of phenotype values after projecting out the genotype vector of the LOCATER lead marker (chr11:105327888), thus removing signal that can be accounted for by the SMT sub-test. Connected dots show the phenotype value of individuals assigned to significant sprigs. **(E)** Dendrogram generated from the haplotype-level local distance matrix at the lead marker chr11:105327888. The UPGMA method was used for hierarchical clustering. Orange branches highlight the path of all haplotypes in significant sprigs shown previously in part **(D)**. Labels at the right show the sprig assignment. For plotting clarity, 95% of haplotypes under insignificant sprigs were pruned.

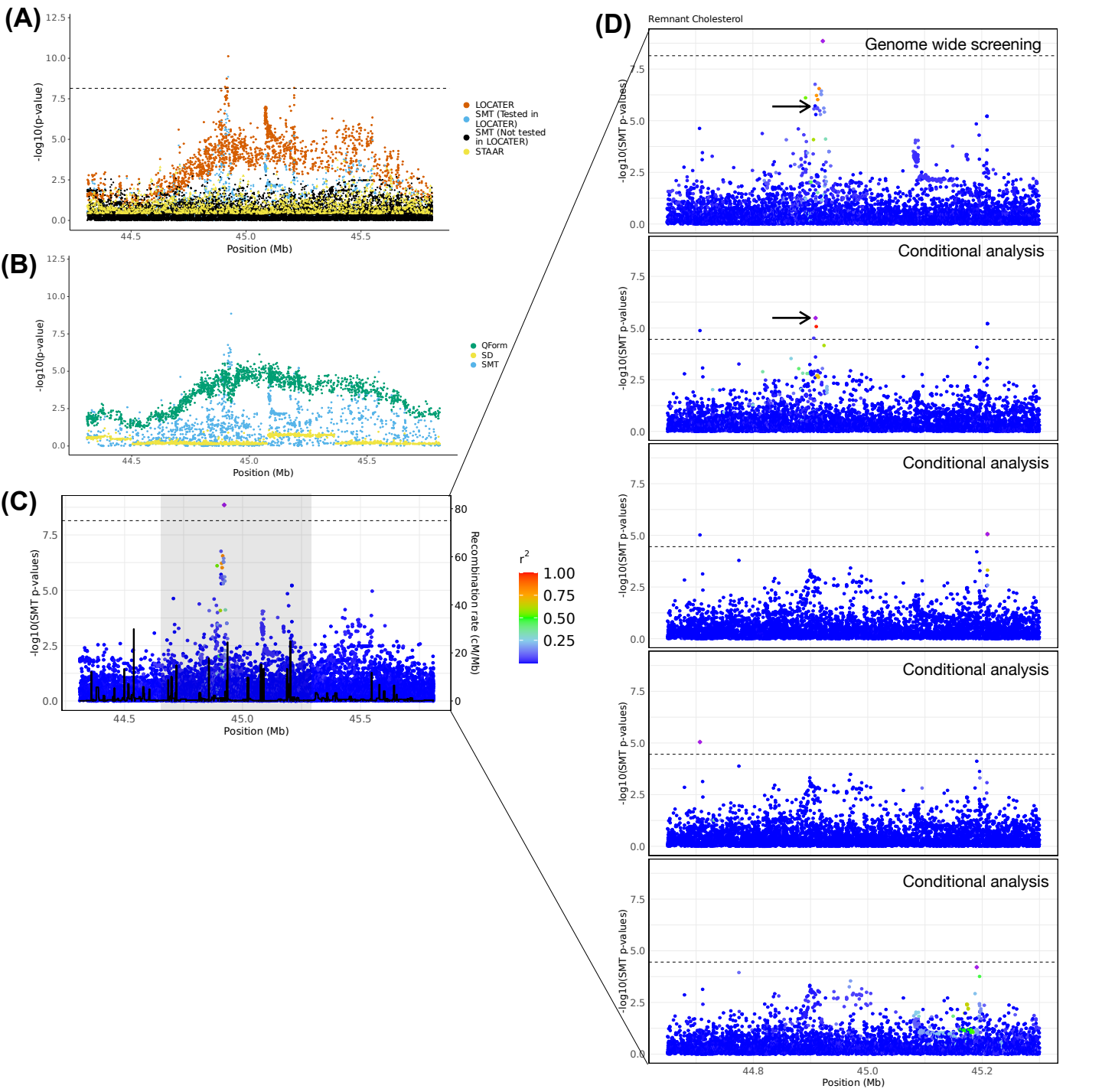


Figure 5. Association of remnant cholesterol at APOE cluster. **(A)** Local Manhattan plot of the association signal for remnant cholesterol on chr19:44308684-45809149, shown using the exact same data types and color scheme as Figure 2A. **(B)** Local Manhattan plot showing adjusted -$\log_{10}$(P) for the 3 LOCATER sub-tests. **(C)** LocusZoom plot of SMT results. Variants are colored based on their r^2 with the SMT lead marker chr19:44922203 (purple diamond), where LD is calculated in the studied samples. The black line shows the recombination rate in Finns (See Methods). Gene annotations are from GENCODE v45. **(D)** Zoomed in LocusZoom plots showing the original association at top, followed by the results from stepwise conditional analysis. Results were zoomed in based on the shaded region in **(C)**. Variants are colored based on their r^2 with the SMT lead marker of each experiment (purple diamond). Black arrows point to the most significant variant from the GWAS catalog. The black dashed line corresponds to the genome-wide significance threshold (7.17×10^{-9}) in the top panel, and the conditional analysis threshold (3.52×10^{-5}) for the rest.

Table 1: Summary of highlighted associations

Variant ID	Mapped Gene	Trait	LOCATER P (adjusted)	LOCATER lead marker MAF	Variant ID of SMT lead marker	SMT P	SMT lead marker MAF	Known hit in GWAS catalog	SMT Lead marker LD w/ GWAS catalog marker	LOCATER Lead marker LD w/ GWAS catalog marker
chr7-73440219-C-T	FZD9, BAZ1B	MUFA	1.57E-09	0.0956	chr7-73440219-C-T*	2.52E-08	0.0956	Yes	0.701	0.701
chr7-73482065-A-C	BAZ1B	Triglycerides in VLDL	1.64E-11	0.117	chr7-73467477-C-T	1.11E-09	0.117	Yes	0.852	0.854
chr7-73643687-A-G	VPS37D, MLXIPL	Concentration of large VLDL particles	2.82E-09	0.118	chr7-73641131-A-C	3.40E-08	0.118	Yes - related trait	0.933	0.932
chr11-61843278-G-A	FADS2	HDL2-C	6.58E-09	0.427	chr11-61798436-T-C	1.65E-08	0.454	Yes - related trait	0.871	0.994
chr11-105327888-G-C	CARD18	Triglycerides in medium VLDL	2.31E-09	0.428	chr11-105331384-A-T	1.05E-05	0.428	No	NA	NA
chr18-49653146-G-A	ACAA2, SMUG1P1	Triglycerides in medium HDL	1.68E-09	0.150	chr18-49642278-G-A	4.71E-07	0.159	Yes	0.864	1
chr18-49817040-T-A	SNHG22	ApoA1	6.51E-09	0.00411	chr18-49817040-T-A*	2.15E-08	0.00411	Yes	1	1
chr19-44922203-A-G	APOC1, APOC1P1	Remnant-C	7.53E-11	0.290	chr19-44922203-A-G*	1.40E-09	0.290	Yes	0.0783	0.0783
chr20-45906012-G-A	PLTP	Triglycerides in small HDL	1.80E-11	0.228	chr20-45923216-T-C	3.10E-10	0.180	Yes	1	0.659
chr21-16318536-C-T	MIR99AHG	HDL3-C	3.87E-09	0.0904	chr21-16318536-C-T*	1.77E-08	0.0904	No	NA	NA

Chromosome positions are based on GRCh38. See Table S1 for trait descriptors, see Table S2 for full results. The genome-wide significance threshold is 7.17×10^{-9}. To allow for a straightforward comparison, the LOCATER p-value was standardized to match the effective number of independent tests for single marker association.
* denotes the associations that SMT lead marker is the same as LOCATER lead marker.