Improving fine-mapping by modeling infinitesimal effects

Ran Cui[1,2,3], Roy A Elzur[1,2,3], Masahiro Kanai[1,2,3,4,5,6], Jacob C Ulirsch[1,2,3,7], Omer Weissbrod[8], Mark J Daly[1,2,3,5], Benjamin M Neale[1,2,3], Zhou Fan[9,*], Hilary K Finucane[1,2,3,*]

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA, [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [3]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA, [5]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland, [6]Department of Statistical Genetics, Osaka University Graduate School of Medicine, Suita, Japan, [7]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA, [8]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [9]Department of Statistics and Data Science, Yale University, New Haven, CT, USA, [*]These authors jointly supervised this work.

Abstract

Fine-mapping aims to identify genetic variants that causally impact a given phenotype. State-of-the-art Bayesian fine-mapping algorithms (for example: SuSiE[1], FINEMAP[2,3], ABF[4], and COJO[5]-ABF) are widely applied in practice[6–11], but it remains challenging to assess their calibration (i.e., whether or not the posterior probability of causality reflects the true proportion of causal variants) in real data, where model misspecification almost certainly exists and true causal variants are unknown. Here, we present the Replication Failure Rate (RFR), a metric to assess the consistency of fine-mapping results based on downsampling a large cohort. Empirical evaluation of fine-mapping results from SuSiE, FINEMAP and COJO-ABF suggest that these methods may be miscalibrated in the under-conservative direction. Next, we show in simulations that non-sparse genetic architecture can lead to miscalibration, while imputation noise, non-normal effect size distributions, and quality control filters removing potentially causal variants are less likely contributors. Here, we present two new fine-mapping methods, SuSiE-inf and FINEMAP-inf, that extend SuSiE and FINEMAP to incorporate a term for infinitesimal effects in addition to a small number of larger causal effects of interest. Our methods exhibit better calibration in simulations and improved RFR and functional enrichment in real data, with minimal loss of recall and competitive computational cost. Furthermore, using the sparse fine-mapped variants identified by our methods to perform cross-population genetic risk prediction in the UK Biobank, we observed a substantial increase in predictive accuracy over SuSiE and FINEMAP. Our work improves our ability to pinpoint causal variants for complex traits, a fundamental goal of human genetics.

Introduction

Over the past two decades, genome-wide association studies (GWAS) have successfully identified thousands of genetic loci that are associated with disease phenotypes and complex traits[12]. However, refining these associations to determine the specific genetic variants that causally affect traits remains challenging, due to extensive linkage disequilibrium (LD) among associated variants[13]. Many approaches can be taken to help nominate variants that are more likely to be causal, such as overlapping GWAS signals with coding or functional elements of the genome[14], with eQTLs[15], and across populations having different ancestries and patterns of LD[16–18]. Complementary to and in conjunction with these approaches, Bayesian sparse regression and variable selection methods, which aim to identify causal variants and quantify their uncertainty based upon a statistical model, are widely used[19].

The appeal of Bayesian approaches to fine-mapping is two-fold. First, these methods determine a posterior inclusion probability (PIP) for each variant, quantifying the probability that the variant is causal under the model, which can reflect uncertainty due to LD. For example, two variants in perfect LD and harboring a strong association with the phenotype may each have PIP 50%, representing confidence that one but not likely both variants are causal. Second, these methods incorporate assumptions about genetic architecture -- namely, the relative probabilities of different numbers of and configurations of causal SNPs, as reflected by a Bayesian prior -- to improve statistical power for identifying high-confidence variants.

Bayesian fine-mapping methods are correctly calibrated when the PIPs accurately reflect the true proportions of causal variants, e.g. 9 out of 10 variants having PIP 90% are truly causal for the trait. Calibration is ensured when the linear model for genetic effects and Bayesian prior for genetic architecture across loci are both correctly specified, and accurate calibration has also been demonstrated empirically in simulations to be robust under mild model misspecifications[20]. However, the actual calibration and false discovery rates of these methods in real data are not easily determined, as true causal variants and the sources of model misspecification may be unknown.

Here, we propose the Replication Failure Rate (RFR) – a metric that assesses the stability of posterior inclusion probability by evaluating the consistency of PIPs in random subsamples of individuals from a larger well-powered cohort – in this instance for 10 quantitative traits in the UK Biobank,. We found the RFR to be higher than expected across traits for several Bayesian fine-mapping methods. Moreover, variants that failed to replicate at the higher sample size were less likely to be coding. Together these analyses suggest that SuSiE[1], FINEMAP[2], and COJO[5]-ABF[4] may be mis-calibrated on real data. In particular, they may return a disproportionately large number of false discoveries among high-PIP variants.

We performed large-scale simulations to assess the effects of several plausible sources of model misspecification. These simulations — which include, among other factors, varying levels of non-sparsity and stratification — suggest that a denser and more polygenic architecture of genetic effects may be a major contributor to PIP mis-calibration. We thus propose incorporating a model of infinitesimal effects when performing Bayesian sparse fine-mapping, recasting the goal of fine-mapping as the identification of a sparse set of large-effect causal variants among many variants having smaller effects. We therefore develop and implement fine-mapping tools SuSiE-inf and FINEMAP-inf that extend the computational ideas of SuSiE and FINEMAP to model additional infinitesimal genetic effects within each fine-mapped locus.

Applying SuSiE-inf and FINEMAP-inf to the preceding 10 quantitative traits in the UK Biobank, we observe improved RFR. Moreover, high-PIP variants identified by SuSiE-inf but not SuSiE are more enriched for functional annotations than SuSiE-specific high-PIP variants. Using sparse effects estimated by SuSiE-inf/FINEMAP-inf to perform cross-ancestry phenotype prediction, we observe a large improvement over SuSiE/FINEMAP, across 7 traits and 5 held-out cohorts having diverse ancestries in the UK Biobank. These results suggest that explicit modeling of a polygenic genetic architecture, even within individual genome-wide significant loci, may substantially improve fine-mapping accuracy.

Results

Real data benchmarking shows that current fine-mapping methods are likely mis-calibrated

Real-data benchmarking of fine-mapping methods is challenging due to the lack of ground truth. However, downsampling large cohorts offers the opportunity to directly assess the stability of fine-mapping methods. We chose 10 well-powered quantitative phenotypes in the UK Biobank (**Methods**) and computed the Replication Failure Rate (RFR) for SuSiE, FINEMAP as follows (see **Methods** for results related to ABF and COJO-ABF). Our group previously performed fine-mapping (results available in [20]) on a cohort of 366,194 unrelated “white British” individuals defined previously in the Neale Lab UKBB GWAS (https://github.com/Nealelab/UK_Biobank_GWAS). Here we down-sampled this cohort to a random subsample of 100,000 and performed fine-mapping with the same pipeline (**Methods**). We then defined RFR as the proportion of high-confidence (PIP > 0.9) variants fine-mapped in the 100K subsample that failed to replicate (PIP < 0.1) in the full 366K cohort. This RFR is an estimate of the conditional probability $Pr(PIP^{366K} < 0.1 \mid PIP^{100K} > 0.9)$ for a randomly chosen SNP. In a truly sparse causal model, under the assumption that the method is well-powered at sample size 366K to detect true causal variants which are identified with high confidence at 100K, the RFR is an approximate lower bound for the false discovery rate $Pr(not\ causal \mid PIP^{100K} > 0.9)$ (see **Supplementary Note**).

Across all 10 traits, we observed different levels of RFR for different phenotypes, and an aggregated RFR of 15% for SuSiE and 12% for FINEMAP (**Fig. 1a-b**; see **Extended Data Fig. 1** for other PIP thresholds). These values far exceed the false discovery rate expected in a correctly specified sparse Bayesian model (SuSiE 1.8%, FINEMAP 2.0%), which we denote by EPN and estimate from the mean reported PIPs exceeding 0.9. In contrast, ideal simulations under correctly specified models show close agreement between RFR and EPN (**Fig. 1a, Methods,** and **Supplementary Note**).

To gain further insight into whether non-replicating variants (PIP>0.9 at 100K and PIP<0.1 at 366K) are causal, we examined the functional annotations of these variants, focusing mainly on two distinct functionally important categories: coding and putative regulatory (**Methods**). We found a significant depletion of functionally important variants in the non-replicating set compared to the replicated set (P=1.7e-7) (**Fig. 1b**, **Methods**). This further suggests that many non-replicating variants may be non-causal, and that SuSiE and FINEMAP may be miscalibrated when applied in real data (see **Methods** for our investigation into other potential causes for high RFR). We did observe more functional enrichment in the non-replicating set of variants than the background, indicating that PIPs for some variants at N=366K may be too conservative. However, here we focus on investigating the more concerning under-conservative PIPs which can lead to elevated false discovery rate.

Un-modeled non-sparse causal effects can lead to miscalibration

Bayesian sparse variable selection approaches to fine-mapping, including SuSiE and FINEMAP, commonly rely on some of the following assumptions: (1) Within each genome-wide significant locus, one or a small number of variants have a true causal contribution to the phenotype. (2) All true causal variants within the locus are included in, or tagged by a sparse subset of, the given genotypes. (3) The distribution of causal variant effect sizes is well-approximated by a simple, oftentimes Gaussian, prior. (4) There is no uncorrected confounding, and the residual error is uncorrelated with the genotype. (5) There is no imputation noise or error in the genotypes. Violations of any of these assumptions can, in principle, cause the posterior probabilities of Bayesian fine-mapping methods to be miscalibrated, although the severity of such miscalibration under the degrees of violation that are present in real fine-mapping applications is unclear a priori.

We designed large-scale simulations to investigate how SuSiE and FINEMAP may be affected by these five sources of misspecification. Our simulations use UK Biobank genotypes (N=149,630 individuals of white British ancestry) and BOLT-LMM[21] for GWAS, incorporating (1) varying amounts of unmodeled non-sparse causal effects, (2) missing causal variants that are removed by quality-control filtering prior to fine-mapping, (3) effect size distributions for the large and sparse causal variants that reflect estimates from fine-mapping of real traits, (4) varying amounts of uncorrected population stratification, and (5) imputation noise in the input genotypes (see **Methods** for detailed description of our simulations and other misspecifications we considered). In previous work[20], we found that quality control filters and imputation noise did not contribute to miscalibration in simulations similar to the ones we perform here; here we continued to include them while adding non-sparsity, effect size estimates from real data, and uncorrected population stratification as additional sources of miscalibration. We note that the simulations considered here are for fine-mapping a single cohort, without the heterogeneity that often comes with meta-analysis; in meta-analysis fine-mapping, quality control and imputation are important contributors to miscalibration[22]. Moreover, it is possible that factors we do not consider here, such as error in the probabilities outputted by standard imputation software or different types of genotyping error, could contribute to miscalibration even in the absence of heterogeneity.

We found that, within our simulations, missing causal variants due to QC, using a realistic, non-Gaussian effect size distribution estimated from real data, and imputation error did not induce miscalibration, consistent with and extending the results from[20].

However, SuSiE and FINEMAP were both significantly miscalibrated in simulations with non-sparse genetic effects. Specifically, miscalibration increased as we increased the proportion of SNP-heritability (set at 0.5; see **Supplementary Table 1** for common-SNP heritability in real traits) explained by non-sparse effects from 58% to 100% (**Fig 2, Table 1**). For example, when non-sparse causal effects explain 75% of the SNP-heritability, for both SuSiE and FINEMAP only about 80% of variants with PIP ≥ 0.9 are causal, far below the rate of approximately 97% that we would expect given the variants' mean PIP. We emphasize that calibration was measured against the set of all causal variants, including the non-sparse causal effects.

To further confirm that unmodeled non-sparse causal effects, among all the misspecification we incorporated, formed the primary driver of the observed miscalibration, we decomposed the simulated genetic component $X\beta$ of the phenotype into the sum of four sub-components representing sparse causal effects, missing causal variants, uncorrected stratification, and unmodeled non-sparse causal effects. Regressing each of these four sub-components on the true and false positive variants (respectively defined as causal and non-causal variants with PIP ≥ 0.9), false positive variants were significantly more correlated with the non-sparse causal effects than true positive variants (**Fig. 2c, Methods**).

Our simulated population stratification failed to induce miscalibration. However, with our pipeline, which computes association statistics with BOLT-LMM, we were unable to induce uncorrected confounding at high levels within reasonable parameter settings (**Methods**); replacing BOLT-LMM with ordinary least squares for association mapping allowed us to induce higher levels of uncorrected confounding (**Supplementary Table 24**) that did lead to miscalibration (**Fig. 3**, **Methods**), but are less true to the pipeline used in our real data analysis.

In conclusion, non-sparse effects can be a driver of miscalibration for SuSiE and FINEMAP. The stratification we simulated only induced miscalibration when using OLS for association mapping but not when using BOLT-LMM. None of the other sources of misspecification incorporated in our simulations caused miscalibration within our fine-mapping pipeline.

New methods for Bayesian fine-mapping

To address PIP miscalibration that may arise from non-sparse causal effects, we propose to explicitly incorporate a model of broad infinitesimal genetic effects when fine-mapping causal variants. Here, we describe two specific implementations of this idea that extend FINEMAP and SuSiE. We call the resulting methods FINEMAP-inf and SuSiE-inf.

FINEMAP-inf and SuSiE-inf are based on a random-effects linear model y=X(β+α)+ε for observed phenotypes y across n samples, where X is a n by p genotype matrix for p SNPs, β is a vector of sparse genetic effects of interest, α is an additional vector of dense infinitesimal effects, and ε is residual error. In the context of such a model, we define the primary goal of fine-mapping as inferring the non-zero coordinates of the sparse component β. We will refer to these coordinates as the “causal model” and the “causal variants”, although in this model, every variant may have an additional small causal effect on y through the infinitesimal component α.

We model coordinates of α and of the residual error ε as i.i.d. with normal distributions $N(0,\tau^2)$ and $N(0,\sigma^2)$, respectively, where τ^2 is the effect size variance for the infinitesimal effect. For FINEMAP-inf, coordinates of the sparse effects β are also modeled as i.i.d., with point-normal distribution π_0N(0,s^2)+(1-π_0)δ_0. We use a shotgun-stochastic-search (SSS) procedure as in FINEMAP for performing approximate posterior inference of the sparse component β, marginalizing its posterior distribution over both the infinitesimal effects α and the residual errors ε. The SSS is divided into several epochs, and we propose a method-of-moments approach to update estimates of the variance components (σ^2,τ^2) between epochs.

For SuSiE-inf, we follow the approach of SuSiE and instead parametrize the sparse causal effects as a sum of single effects $\beta = \sum_{l=1}^{L} \beta^{(l)}$ for a pre-specified number of causal variants L. As in SuSiE, we perform posterior inference for β using a variational approximation for the joint posterior distribution of $\beta^{(1)}$,..., $\beta^{(L)}$, again marginalizing over both α and ε. The approximation is computed by iterative stepwise optimization of an evidence lower bound (ELBO), where updated estimates of the variance components (σ^2,τ^2) are computed within each iteration using a method-of-moments approach.

The resulting models are similar to linear mixed models commonly used in contexts of association testing and phenotype prediction[21,23–26]. Here, we focus on applications to fine-mapping, which differ from previous uses of these models in that (a) fine-mapping requires inclusion of a dense set of variants in each locus, so that the causal variants are likely to be included; (b) fine-mapping requires accurate inference of posterior inclusion probabilities; (c) fine-mapping is often performed at very large sample sizes; and (d) fine-mapping does not require joint modeling of genome-wide data, which would be computational challenging given the density of variants and typical sample sizes. Because of these factors, we do not apply existing methods for fitting linear mixed models in other contexts; instead, we estimate the infinitesimal variance component separately for each genome-wide significant locus, and extend algorithmic ideas in the fine-mapping literature to estimate the sparse component in the presence of strong LD. We model infinitesimal effects for variants in LD with those of the sparse component, which we believe is important for obtaining improved calibration and fine-mapping accuracy.

Both methods take as input either the GWAS data (y, X) or sufficient summary statistics given by the un-standardized per-SNP z-scores z = 1/√n $X^T y$, the in-sample LD correlation matrix LD = 1/n $X^T X$, and the mean-squared phenotype $\langle y^2 \rangle$ = 1/n $y^T y$. Both methods output estimates of (σ^2, τ^2) for each locus fine-mapped, together with a posterior-inclusion-probability (PIP) and posterior-mean effect size estimate for each SNP. Computational cost is reduced by expressing all operations in terms of the eigenvalues and eigenvectors of LD, which may be pre-computed separately for each fine-mapped locus (**Extended Data Fig. 2**). Details of the methods and computations are provided in **Supplementary Note**. We have released open source software implementing these methods (see **Code availability**).

SuSiE-inf and FINEMAP-inf show improved performance

In our simulations, we find that SuSiE-Inf and FINEMAP-Inf have improved calibration over SuSiE and FINEMAP, respectively, except for the simulations using ordinary least squares, which are less relevant to our findings in real data for which we used BOLT-LMM (**Fig. 2a, 3a**). Recall of SuSiE-Inf and FINEMAP-Inf was very similar to, but slightly lower than, that of SuSiE and FINEMAP, respectively (**Fig. 2b, 3b**). Credible sets generated by SuSiE-inf are smaller on average than those generated by SuSiE (**Extended Data Fig. 3**). With improved performance in simulations with non-sparsity, similar performance in simulations with stratification and BOLT-LMM, and worsened performance in simulations with stratification and OLS (**Fig. 3a**), we turned to real data benchmarking to assess whether the new methods improve performance in practice.

Real data benchmarking shows improvements by several metrics. RFR was substantially decreased for SuSiE-inf (**Fig. 4a**). High-PIP variants that are identified by SuSiE-inf but not SuSiE are 58% more enriched in functionally important categories than high-PIP variants identified by SuSiE but not SuSiE-inf (P=6e-4); the analogous difference in functional enrichment for FINEMAP vs. FINEMAP-inf was non-significant (38% more for FINEMAP-inf specific variants, P=0.07, **Fig. 4b**). Compared to SuSiE and FINEMAP, we obtained fewer high-PIP variants (16% reduction aggregated between SuSiE and FINEMAP); however, the reduction is smaller for high-confidence variants, characterized either by replicated variants (11% reduction), or variants achieving PIP>0.9 for both SuSiE-Inf and FINEMAP-Inf/both SuSiE and FINEMAP (11% reduction) (**Extended Data Fig. 4a**). We observed a more substantial reduction of 42% in the number of credible sets when using SuSiE-inf; however, the reduction for smaller credible sets (number of variants < 10) was somewhat smaller (36% reduction). High confidence variants discovered by SuSiE-inf and FINEMAP-inf exhibit higher functional enrichment (**Extended Data Fig. 4b-c, Supplementary Fig. 1**). Together, these results demonstrate both that SuSiE-inf and FINEMAP-inf allow for more confident identification of likely causal variants than the current state of the art, and also that there is room for further methodological improvement.

In simulation, estimates of τ^2 were higher on average for simulation settings with higher true infinitesimal variance (**Extended Data Fig. 5**). In UKBB data, the estimates of the infinitesimal variance τ^2 varied across traits, with height showing the highest estimates and LDL showing the lowest estimates (**Extended Data Fig. 6a-b**). We also found that estimates of τ^2 increased, on average, as the number of credible sets in a locus increased (**Extended Data Fig. 6c**). Estimates of τ^2 varied across loci for a given trait, due either to differences in genetic architecture or to estimation noise.

To further validate our methods in real data, we performed cross-ancestry Polygenic Risk Score (PRS) prediction[27,28], using posterior effect sizes estimated on 366K samples from the "white British" cohort in UK Biobank to predict phenotypes in five held out cohorts of different ancestries[29]: AFR (N=6637), AMR (N=982), CSA (N=8876), EAS (N=2709), and MID (N=1599). Prediction accuracy is measured by "delta R^2" which is the difference in R^2 from a model that includes both the covariates and genotype effects relative to a model that includes the covariates alone. Using posterior mean effect size estimates for the sparse component β in SuSiE-inf/FINEMAP-inf yields, on average, a near 10-fold increase in delta R^2 across these five held out cohorts and across traits compared to using SuSiE/FINEMAP (**Methods, Fig 4c-d**). Here we compute PRS using only the sparse component, to provide a validation metric for the fine-mapped SNPs. We leave an exploration of improved PRS methods that further integrate fine-mapping output with estimates of dense effects in a polygenic architecture to future work.

We have shown previously that combining SuSiE and FINEMAP can yield more reliable PIPs[30]. Here we recommend the general user to take the minimum PIP between SuSiE-inf and FINEMAP-inf for each fine-mapped variant. We henceforth refer to this method as minPIP-inf, and refer to taking the minimum PIP between SuSiE and FINEMAP as minPIP. minPIP-inf has a smaller reduction in the number of high confidence variants than minPIP does, or in other words, SuSiE-inf and FINEMAP-inf agree with each other more than SuSiE and FINEMAP does (**Extended Data Fig. 7**). We observed substantially improved RFR for minPIP-inf over minPIP (**Extended Data Fig. 8a**). Functional enrichment for the top N variants for minPIP-inf is comparable to either SuSiE-inf or FINEMAP-inf individually (**Extended Data Fig. 8b**). Simulation and PRS performance of minPIP-inf are also comparable to those of either method individually (**Fig. 2a-b, Extended Data Fig. 8c-d**). As examples of the utility of the minPIP-inf method, we examined the PCSK9 locus for LDLC, as well as the AK3 locus for Plt, where two non-coding variants have previously been validated as having independent regulatory activity in a reporter assay[9] (**Supplementary Fig. 5-6**).

Discussion

We propose fine-mapping methods that control for infinitesimal causal effects while fine-mapping sparse causal effects. Using our new methods, we observed significant improvements in simulations with non-sparse genetic architecture. but our results when simulating uncorrected stratification were ambiguous: when using BOLT-LMM, stratification did not lead to miscalibration and the new methods performed similarly to the previous methods; however, when using OLS, stratification led to substantial miscalibration that was similar between FINEMAP and FINEMAP-inf and worse for SuSiE-inf than SuSiE. In contrast, our real data benchmarking showed an unambiguous improvement in performance when using SuSiE-inf and FINEMAP-inf over SuSiE and FINEMAP: the new methods led to decreased

RFR, improved functional enrichment of top variants, and large gains in polygenic risk prediction. Put together, our results suggest that the accuracy of identifying sparse causal variants is greatly improved when using the infinitesimal model to finemap, and that there is a need for further methods development to continue to improve fine-mapping accuracy.

The models we propose here are similar to models that have been proposed previously to model genome-wide genetic architecture for risk prediction, heritability estimation, and association mapping[21,23,25,26]. Fine-mapping differs from these other applications in that only one locus is modeled at a time, that a dense set of variants must be jointly modeled, and that extra precision is needed in differentiating the effect sizes of variants in LD and assigning posterior inclusion probabilities to individual variants. Practically, this means that fine-mapping is more sensitive to factors such as meta-analysis heterogeneity, inexact reference panel LD, low variant density, and low sample size, but that genome-wide joint modeling is not necessary. As an example of the differences between fine-mapping and risk prediction, consider the case in which SNP 1 and SNP 2 are in perfect LD with a large marginal effect size and no other variants in LD. In this case, a risk prediction method will perform equally well regardless of whether the effect size is estimated to be large for SNP 1 and zero for SNP 2, large for SNP 2 and zero for SNP 1, or moderate for both SNP 1 and SNP 2. For fine-mapping, though, the desired outcome is a more precise quantification of uncertainty: with high probability, only one of the two SNPs has a non-negligible effect size, equal to the marginal effect size, and it is equally likely that the causal SNP is SNP 1 or SNP 2. Because of these differences, we did not use existing methods for fitting linear mixed models, but instead extended the ideas of previously-validated fine-mapping methods to accommodate a linear mixed model. With careful translation, we anticipate that methodological innovations in risk prediction may continue to lead to advances in fine-mapping and vice versa.

We view our methods as complementary to a body of recent and influential statistical developments that seek to more accurately quantify and control false discoveries under minimal modeling assumptions, using constructions of knock-off variables and related conditional re-randomization ideas.[31–33] Our approach remains largely model-dependent, but we illustrate the potential of improving fine-mapping accuracy and reducing false discovery through incorporating improved models of genetic architecture. We see further exploration of the possible utility of knock-off variables in the context of fine-mapping as a promising direction for future research.

While our work allows for more accurate fine-mapping, further advances are needed in fine-mapping methods development. First, we see further investigation of the effects of stratification and different association mapping methods on fine-mapping as an important direction for future work. More generally, our new methods improve the replication failure rate over the current state-of-the-art, but even the improved RFR is above what is found in simulations with no model misspecification, suggesting that fine-mapping model and methodology can be further improved. In addition to better modeling, independent replication in another biobank is strong evidence for true causality[30]. Functional evidence such as annotations and eQTLs[20] can help boost accuracy of discovery as well. Further methodological advancements, in addition to leading to more accurate identification of likely causal variants, may also contribute to further improvements in cross-population polygenic risk prediction. Such methodological advancements may come from more flexible models of genetic architecture or from further study of the effects of uncorrected confounding on fine-mapping. In the meantime, we recommend SuSiE-Inf and FINEMAP-Inf for Bayesian fine-mapping when genotype data or in-sample LD data are available.

Methods

Selection of UKBB phenotypes and downsampling analysis

The selection of 10 phenotypes for which to perform downsampling analyses was mainly based on the combined number of high-PIP (PIP > 0.9) variants fine-mapped at N = 366K samples using both SuSiE and FINEMAP[30]. From the top 15 phenotypes (out of 94) with the highest number of high-PIP variants (**Supplementary Table 22**) we selected: Height, estimated heel bone mineral density (eBMD), platelet count (Plt), hemoglobin A1c (HbA1c), red blood cell count (RBC), alkaline phosphatase (ALP), insulin-like growth factor 1 (IGF1), low density lipoprotein cholesterol (LDLC), lymphocyte count (Lym), and estimated glomerular filtration rate based on serum creatinine (eGFR) to perform down-sampling analyses. For phenotype definitions and processing see[30].

We downsampled from N=366K to a random subset of N=100K twice (to increase the number of discoveries and therefore statistical power for RFR analyses) and performed GWAS and fine-mapping on both set of the N=100K individuals using the same pipeline used at N=366K (see below for pipeline description).

Comparison with ABF and COJO-ABF

Approximate Bayes' factors (ABF[4]) and conditional and joint analysis followed by ABF (COJO[5]-ABF) are two commonly used Bayesian fine-mapping methods. ABF is a single-causal-variant fine-mapping method where only one causal variant is modeled in a given fine-mapped region. It is most commonly applied when only summary statistics are available since it does not require LD. COJO-ABF first uses conditional analysis to infer independent associations, then performs ABF on each identified association, conditioning on the others. This approach is not model-based, unlike SuSiE and FINEMAP.

COJO-ABF has RFR 11% in real data, 2.6% RFR in ideal simulations, and a similar functional enrichment profile to SuSiE and FINEMAP (**Extended Data Fig. 9**). However we observed severe miscalibration of COJO-ABF in our ideal simulations and lower recall than the other model-based multiple-causal-variant fine-mapping methods we tested (**Extended Data Fig. 10a-b**). This is consistent with existing literature showing that conditional analysis has suboptimal precision and power/sensitivity[20,34]. We observed that the regions with multiple causal variants are more likely to harbor false positive variants (**Extended Data Fig. 10c**) a possible reason that previous simulations with fewer causal variants did not show such severe miscalibration[20].

ABF has a relatively higher RFR of 17%, however functional annotations of the non-replicating variants show comparable enrichment to the replicated variants (**Extended Data Fig. 9b**). ABF also has higher RFR in ideal simulations (12%), where 70% of the non-replicating variants are true causal variants. Based on the evidence from both real data and simulations, we hypothesize that the high RFR for ABF in real data is mostly due to different causal variants being prioritized at different sample sizes and it mostly reflects the non-discovery rate of single-causal-variant fine-mapping in a locus with multiple causal variants, instead of a high false discovery rate (**Supplementary Note**). As expected, ABF exhibits much lower recall than the multiple-causal-variant methods (**Extended Data fig. 10b**).

We recommend using model-based multiple-causal-variant methods for fine-mapping when in-sample LD or genotype/phenotype data are available for better calibration and recall.

Fine-mapping pipeline

GWAS and fine-mapping in this paper were performed following the pipeline described in. Briefly, GWAS summary statistics were computed using BOLT-LMM with covariates including sex, age, age2, age and sex interaction term, age2 and sex interaction term, and top 20 genotype Principal Components (PCs). Fine-mapping regions were defined using a 3Mb window around each lead GWAS variant, with merging of overlapping regions. Fine-mapping was performed with in-sample LD computed using LDstore v2.0[35]. We refer to[30] for additional details.

Excessively large merged regions that could not be fine-mapped due to computational limitations were tiled with overlapping 3Mb loci, with 1Mb spacing between the start points of consecutive loci. For these tiled regions, we computed a PIP for each SNP based on the 3Mb locus whose center was closest to the SNP. This tiling approach was previously applied in[36].

To investigate the effects of uncorrected population stratification, we also performed a few simulations using GWAS summary statistics computed by ordinary least squares regression instead of BOLT-LMM (see **Population stratification** below).

Ideally, variants in low-complexity regions (LCR) would be filtered out before imputation due to higher chance of genotyping error and therefore worse imputation performance. However, we found that these variants are in the imputed genotypes provided by UK Biobank, therefore they are included in our fine-mapping pipeline. Around 5% of total variants included in our GWAS are in LCR, and around 6% of total fine-mapped variants are in LCR. We provide a list of variants that are in LCR and obtained nontrivial PIP (>0.1) from any of the six fine-mapping methods (SuSiE, FINEMAP, COJO-ABF, ABF, SuSiE-inf, and FINEMAP-inf) in **Supplementary Table 23**. Since the function of LCR is mostly unknown and the accuracy of genotyping in these regions are not provided, we recommend caution when interpreting results at or near these variants

(**Supplementary Fig. 7-8** show different fine-mapping results with and without LCR variants at the APOE locus for LDLC). This does not affect the overall message of our manuscript since most featured results are based on comparison between methods where the same sets of variants are included.

Ideal simulations

To establish reference Replication Failure Rate (RFR) and calibration for all tested methods, we performed ideal simulations without model misspecification using UK Biobank genotypes. For simulating RFR, we performed two sets of simulations each at sample size N = 366K and subsample size N = 100K. We used UK Biobank imputed dosages as true genotypes, and only selected “white British” individuals defined previously in the Neale lab GWAS. We drew 1000 causal variants per simulation uniformly randomly from a total of 6.6 million common (MAF ≥ 1%) imputed variants genome-wide. We standardized genotypes to mean 0 and variance 1, and drew per-standardized-genotype causal effect sizes from the same normal distribution N (0, 0.5/1000) for all selected causal variants. We then added errors randomly drawn from a normal distribution N (0, 0.5) to simulate phenotypes. For comparison of calibration with our simulations under model misspecifications, three additional sets of ideal simulations at a matching sample size N = 150K were performed. Phenotypes were generated similarly, with 700 uniformly sampled true causal variants having effect sizes drawn from N (0, 0.5/700).

Functional enrichment

We analyzed functional annotations to gain insights into the potential causal status of non-replicating variants (defined in the main text and in the next paragraph). We define three main disjoint functional categories: coding, putative regulatory, and non-genic. These categories are derived from the seven main functional categories defined in[30]. The “coding” category is the union of pLoF and missense categories; the “putative regulatory” category is the union of synonymous, 5’ UTR, 3’ UTR, promoter and cis-regulatory element (CRE) categories; the “non-genic” category is identical to the “non-genic” category defined in[30]. We compare the proportion of non-genic variants in the following groups of variants:

1. Non-replicating, defined as the variants with PIP ≥ 0.9 at N = 100K and PIP ≤ 0.1 at N = 366K.
2. Replicated, defined as the variants with PIP ≥ 0.9 at N = 100K and PIP ≥ 0.9 at N = 366K.
3. Matched on PIP at 100K, defined as the group of replicated variants closest resembling the non- replicating variants in terms of PIP at N = 100K. For each non-replicating variant with PIP / = 1, we find a replicated variant whose PIP is the closest as its match, and the matched variant is removed for future matches. If the non-replicating variant has PIP = 1, we match a random (if there are multiple) replicated variant with PIP = 1. If there are more non-replicating variants with PIP = 1 than there are replicated variants with PIP = 1, we do not remove the matched replicated variant from future matches, resulting in repeated matches.

4. Matched on PIP at 366K, defined as the group of low-PIP variants (PIP ≤ 0.1 at N=366K) closest resembling the non-replicating variants in terms of PIP at N = 366K. Matching is performed the same way as described above, except that there are no repeated matches.
5. Background, defined as the union of all variants included in fine-mapping from all 10 phenotypes.

P-values are reported when assessing the significance of the difference between proportions of non-genic variants in different groups of variants. Fisher's exact test was performed using the R function fisher.test, and one-sided p-values were reported from the output of this function.

Investigating non-replication

BOLT-LMM vs. OLS

We investigated whether the use of BOLT-LMM summary statistics can induce non-replication in downstream fine-mapping, by comparing results with those obtained using summary statistics computed by ordinary least squares (OLS). We performed GWAS using OLS on Height at N = 366K and N = 100K. We obtained near linear (correlation coefficient 0.95) relations between the OLS marginal association chi-squared statistics and the BOLT-LMM chi-squared statistics, with BOLT-LMM chi-squared statistics being larger (**Supplementary Fig. 2a-b**). We observed decreased RFR (SuSiE: 20% → 8% ± 5%, FINEMAP: 28% → 6% ± 5%), but also substantially reduced power, when using OLS summary statistics. To approximately match the OLS analyses on power, we re-performed analyses using BOLT-LMM and SuSiE (omitting other methods for due to computational cost concerns) with reduced sample and subsample sizes of N = 280K and N = 88K (**Supplementary Fig. 2c-d**). In these analyses we observed an RFR reduction (SuSiE: 20% → 9%±5%) similar to the reduction observed using OLS. We conclude that the power difference between BOLT-LMM and OLS, rather than the difference between statistical models, was likely the main contributor to the RFR differences in this investigation.

Region definition differences between sample sizes

In the fine-mapping pipeline of[30], fine-mapped regions are defined by windows around genome-wide-significant variants, and regions defined at N = 366K are often larger (because of increased power and the merging of adjacent regions) than those defined at N = 100K. We investigated whether potentially missing causal variants due to differences in region definition can contribute to non-replication. We re-applied SuSiE for Height at N = 100K using the regions defined at N = 366K, for 35 regions that harbored either non-replicating or replicated variants. We observed only one fewer (12 → 11) non-replicating variant when using the same region definitions at both sample sizes, suggesting that region definition differences are unlikely to be a main contributor to non-replication. We note that using an alternative pipeline described here[36] where the same sets of 3Mb sliding windows are used to perform fine-mapping at different sample sizes, we also observed high levels of RFR (**Supplementary Fig. 3**).

Maximum number of causal variants per region

In our analyses, for both SuSiE and FINEMAP, we set the maximum number of causal variants per region to be 10. We investigated whether increasing this number would reduce RFR that is potentially caused by different prioritizations of true causal variants at different sample sizes. We re-applied SuSiE on the 35 regions for Height defined above, setting the maximum number of causal variants to 20 and 50. Both settings yielded the exact same high-PIP and non-replicating variants, and the RFRs were 23% ± 6.5%.

PC differences between sample sizes

We considered the possibility that PC differences at different sample sizes can potentially introduce uncorrected (or differently corrected) confounding, and therefore lead to inconsistency in fine-mapping results. We re-applied SuSiE on Height using the PCs computed at N = 366K as covariates when performing GWAS at N = 100K. The resulting RFR is 27% ± 6.5%, similar to what we observed when using PCs computed at N = 100K. We therefore rule out this possibility.

SNP properties

To further investigate the non-replicating variants, we attempted to characterize non-replicating and replicated variants using the following properties: (a) Minor Allele Frequency (MAF), (b) imputation INFO score, (c) chi-square statistics, (d) LD score, (e) value of the PIP, (f) SuSiE and FINEMAP PIP difference, (g) posterior expected number of other causal variants within a 100kb window. We measured the ability of each property for distinguishing non-replicating and replicated variants by how well a simple threshold rule using its value can separate these classes, as is commonly done to measure feature importance in binary classification. We found that none of these properties can lead to an effective threshold-based QC process that reduces RFR without significantly compromising power (**Supplementary Fig. 4a**).

Distribution of non-replicating variants under repeated subsampling

We investigated whether most of the non-replication may be attributed to a small number of non-representative regions. To do this, we repeated our analyses for Height using SuSiE and FINEMAP in 10 additional randomly downsampled subsets of N = 100K individuals. We call a region non-replicating if it harbors any non-replicating variant(s). Out of 88 non-replicating regions that harbored a total of 193 non-replicating variants across both methods and all 10 downsampling analyses, only 19 (22%) regions were non-replicating in more than 2 out of 10 downsampling analyses, and only 5 (6%) regions were non-replicating in more than 5 downsampling analyses. In comparison, regions containing replicated variants tended to repeatedly appear in multiple downsampling analyses (**Supplementary Fig. 4b**). We conclude that non-replication is not mainly due to complexities in a few non-representative loci.

Large-scale simulations with misspecification

We selected 149,630 UK Biobank individuals from a set of 366,194 unrelated “white British” individuals defined previously in the Neale lab GWAS for our large-scale simulations. We performed simulations under models that are misspecified in the following ways: (1) genotype imputation noise, (2) non-uniform probabilities for the identities of causal variants, (3) non-sparsity of true causal effects, (4) uncorrected population stratification, and (5) missing causal variants. We performed 9 sets of simulations. All simulations included the same extent of (1) imputation noise, (2) non-uniform prior causal probabilities, and (5) missing causal variants. The first simulation, "baseline misspecification" in **Table 1**, also included a small amount of (4) uncorrected stratification. Another four simulations varied, in addition, (3) the level of non-sparsity of causal effects. Finally, four additional simulations varied (4) the amount of simulated stratification and the methods for correcting this stratification (see **Population stratification** below).

Genotypes

To simulate genotypes for 149,630 individuals, we randomly drew true genotypes for all autosomes based on the genotype probabilities in the imputed bgens provided by UKBB. Briefly, probabilistic true genotypes (pGTs) for a given variant i were computed via

$$pGT_i = \lceil u_i - GP(X_i = 0)\rceil + \lceil u_i - GP(X_i = 0) - GP(X_i = 1)\rceil, \quad (3)$$

where $GP(X_i = k)$, $k \in \{0, 1, 2\}$ represents the genotype probability of having k copies of alternative alleles and $u_i \sim Uniform(0, 1)$ represents a uniform random variable. Phenotypes were generated using the pGTs. In downstream GWAS and fine-mapping, we use imputed genotype dosages provided by UKBB, thus simulating imputation noise. We only included variants with minor allele count > 10, INFO score > 0.2, and Hardy-Weinberg equilibrium p-value > 1e-10 in our simulations.

Causal variants

To incorporate a more realistic non-uniform distribution over causal variants, we simulated sparse causal effects from the SuSiE posterior distribution for UKBB Height, as computed in the larger 366K sample in[30]. Specifically, in each locus, for each credible set CSi outputted by SuSiE, we chose a causal variant according to normalized posterior inclusion probabilities within the corresponding SuSiE single effect (denoted α_{ik} for $k \in CS_i$). We then drew the chosen variant's raw effect size (to be scaled later) from a normal distribution with mean and standard deviation given by the SuSiE posterior mean and standard deviation conditional on inclusion in the model. In total, 1434 sparse causal variants were chosen.

For the 4 sets of simulations that investigated non-sparsity of causal effects, we drew additional causal variants uniformly at random such that approximately 1% of all simulated variants have a non-zero effect. For each selected variant, we sampled its raw effect size (to be scaled later) from $N(0, v)$ where $v = [2p(1 - p)]\alpha$, p represents the MAF, and $\alpha = -0.38$[37]. For all simulation settings, simulated non-sparse effects had an overall effect size standard deviation approximately on the order of 1e-4 units per normalized genotype.

We simulated 4 settings of non-sparsity, where the proportions of total SNP heritability explained by the non-sparse causal variants were 58%, 75%, 83% and 100%, corresponding to heritability ratios between sparse and non-sparse causal effects of 1-to-1.4, 1-to-3, 1-to-5, and 0-to-1. We set the total SNP heritability to be 0.5, which accounts for all simulated causal SNPs and not just the common SNPs. s-LDSC measured common SNP heritability for all the simulations and all 10 UK Biobank phenotypes are available in **Supplementary Table 24,25**. To achieve these heritability proportions, we scaled all of the simulated sparse and non-sparse causal effect sizes by corresponding constants. For all simulation settings, simulated large effects had an overall effect size standard deviation approximately on the order of 1e-2 units per normalized genotype.

Population stratification

To simulate population stratification, we first regressed UKBB Height on the top 20 principal components (PCs) of the genotyped variants for N = 360,415 individuals. We then added the sum of the principal component scores multiplied by their respective regression coefficients to the simulated phenotype, scaling this sum by a factor to vary the amount of simulated stratification. We assessed the amount of stratification by running s-LDSC[38] on the resulting GWAS summary statistics (without using any in-sample PCs as covariates) and examining the fitted intercept.

For the stratification simulations referenced in the main text and **Table 1**, we scaled PC effects by a factor of 5 (resp. 8) for moderate (resp. severe) stratification with BOLT, yielding a phenotype with 16.4% (resp. 42.9%) of its variance explained by stratification. For stratification with OLS, we scaled PC effects by 1 and 2 for moderate and severe stratification, yielding phenotypes with 0.6%, 2.6% of their variance due to stratification, respectively. s-LDSC intercepts of the stratification simulations are available in **Supplementary Table 24**.

Phenotype

Phenotypes were generated as

$$y = X\beta + C\zeta + \varepsilon, \quad (4)$$

where X is the above true genotype (pGT) matrix, β is a vector of the (sparse and non-sparse) causal effects, C is a matrix with top 20 principal components with corresponding effects ζ, and $\varepsilon \sim N(0, \sigma^2 I_n)$ where σ^2 was chosen to yield total phenotypic variance equal to 1.

Missing causal variants

After generating phenotypes and before performing GWAS and fine-mapping, we applied variant-level quality-control criteria as previously defined in the Neale lab GWAS, which retained 13,364,303 variants with INFO > 0.8, MAF > 0.001, and Hardy-Weinberg equilibrium P value > 1e-10, with exception for the VEP-annotated coding variants where we allowed MAF > 1e-6. Notably, this QC step resulted in the exclusion of approximately 71% of the simulated "non-sparse" causal variants.

GWAS and fine-mapping

We performed GWAS on N = 149,630 individuals using BOLT-LMM v2.3.2[21], with corresponding imputed variant dosages from UKBB. We used the top 19 principle components computed in-sample as covariates in the GWAS, except in the population stratification simulations, which included no covariates. For some of the population stratification simulations, we performed GWAS with ordinary least squares regression, rather than BOLT-LMM. We performed OLS using the linear regression rows method in Hail v0.2.93. For fine-mapping we used the pipeline previously described in **Fine-mapping pipeline**.

Interpreting population stratification simulation results

When scaling PC effects by a factor of 5 and computing GWAS summary statistics using BOLT-LMM, we observed an s-LDSC intercept of 1.07, which is comparable to s-LDSC intercepts estimated in real complex traits (**Supplementary Table 24**)**,** and we did not observe significant miscalibration in the downstream fine-mapping results. When we simulated a higher level of uncorrected stratification, scaling PC effects by a factor of 8 (s-LDSC intercept of 1.16, see “Severe stratification with BOLT” in **Table 1**), PIPs obtained in downstream fine-mapping remained well-calibrated (**Fig. 3**).

We hypothesize that the use of BOLT-LMM in our standard fine-mapping pipeline helped to correct for the simulated stratification effects, even though the in-sample PCs were not explicitly provided as covariates. This also likely explains the prima facie surprising recall results in **Fig. 3** where the severe stratification simulations with BOLT have higher recall than the moderate stratification simulations with BOLT. In the severe simulations, stratification accounts for 42.9% of the phenotypic variance, whereas in the moderate simulations, stratification accounts for only 16.4% of phenotypic variance. Because BOLT-LMM likely corrects for much of this simulated stratification, it effectively reduces the residual noise in the associations by much more for the severe simulations than for the moderate ones, allowing fine-mapping to nominate more causal variants.

To investigate stratification effects without using an LMM procedure, we performed 2 additional sets of simulations where GWAS summary statistics were instead computed using ordinary least squares (OLS). In these simulations, scaling PC effects by factors of 1 and 2 yielded average s-LDSC intercepts of 1.055 and 1.295, respectively (**Supplementary Table 24**), and induced significant miscalibration across all methods. This miscalibration was more severe for SuSiE-inf and FINEMAP-inf than for SuSiE and FINEMAP (**Fig. 3**).

It is unclear to us which of these simulation settings may be closer to reflecting the possible effects of uncorrected stratification in real fine-mapping applications, given that common methods of computing GWAS summary statistics do use LMM procedures and, in addition, explicitly control for in-sample PCs as covariates. Our real-data results in UKBB show evidence that SuSiE-inf and FINEMAP-inf are outperforming existing methods in realistic settings. We leave to future work a fuller investigation of the possible effects of uncorrected stratification on downstream fine-mapping, and a potential extension of these methods to address uncorrected stratification.

Regression of phenotype components on high-PIP variants

To identify which of several simulated model misspecifications were responsible for observed miscalibration, we decomposed the simulated genetic component $X\beta$ of the phenotype into the sum of four sub-components representing sparse causal effects, non-sparse causal effects, non-sparse causal effects due to QC, and the effects of stratification. That is,

$$X\beta = X\beta_{sparse} + X\beta_{nonsparse} + X\beta_{missing.nonsparse} + XW\zeta \qquad (1)$$

Where W is an $n \times 20$ matrix of UKBB PC loadings computed at a sample size of 360,415 and ζ is a 20×1 vector of regression coefficients for the top 20 PCs on UKBB Height at 360,415. For each simulation, we regressed each of the four genetic effect sub-components on each of the PIP > 0.9 variants independently, with 19 in-sample (n=149,630) PCs as covariates in the regression (i.e. the same covariates we use in GWAS in our simulations). For example, for the sparse genetic effect component, we compute the regression coefficient b and its associated F-statistic for the following equation:

$$X\beta_{sparse} = X_i b + CA \qquad (2)$$

where variant i is the index of a PIP > 0.9 variant and C is a matrix of 19 in-sample PCs. We then compare the F-statistics of truly causal and non-causal variants.

Polygenic Risk Score (PRS)

Cohort assignment

We used six ancestry groups derived by the Pan UKBB project[29], they are: EUR = European ancestry (N=420531), CSA = Central/South Asian ancestry (N=8876), AFR = African ancestry (N=6636), EAS = East Asian ancestry (N=2709), MID = Middle Eastern ancestry (N=1599), and AMR = Admixed American ancestry (N=980). 1000 Genomes Project and Human Genome Diversity Panel (HGDP) were used as reference panels to assign continental ancestry. For technical details please see the Pan UKBB project[29].

Weights

We chose seven phenotypes: HbA1c, Height, LDLC, Lym, Plt, RBC and eBMD for PRS predictions. We fine-mapped these seven phenotypes on the training cohort: EUR (QC'ed from N=420531 to N=366,194 unrelated “white British” individuals). SuSiE, FINEMAP, SuSiE-inf and FINEMAP-inf posterior effect sizes were obtained and filtered to regions that were successfully fine-mapped by all four methods. PLINK2.0[39] was then used to compute polygenic risk scores for the other five cohorts using these posterior effect sizes. For SuSiE-inf and FINEMAP-inf we assigned weights to variants using the estimated posterior effect sizes from the sparse effects β and did not add the estimated posterior effect sizes of the infinitesimal effects α.

Accuracy metric

We use delta R^2 as our accuracy metric for PRS predictions, as in [40]. To obtain delta R^2 , we fit two models:

- Model 0: a linear model using only covariates as predictor, denoted model0.
- Model 1: a linear model using true phenotype as target and both the PRS generated from multiplying the fine-mapped posterior effect size estimates with the genotypes and the covariates (sex, age, age^2, age and sex interaction term, age^2 and sex interaction term) as predictors

We applied the function *lm* in R and obtained *adj.r.squared* from *summary(model1)* and *summary(model0)*. The difference: *summary(model1)$adj.r.squared* - *summary(model0)$adj.r.squared* is delta R^2.

Data availability

The main fine-mapping results at N=100K sample size produced by this study are publicly available at https://doi.org/10.5281/zenodo.7055906. The fine-mapping results at N=366K previously produced by our group is available at https://www.finucanelab.org/data. The UKBB individual-level data is accessible on request through the UK Biobank Access Management System (https://www.ukbiobank.ac.uk/). The UKBB analysis in this study was conducted via application number 31063.

Code availability

Software implementing SuSiE-inf and FINEMAP-inf are publicly available: https://github.com/FinucaneLab/fine-mapping-inf. The code to generate all figures in this manuscript is available at https://github.com/cuiran/improve-fine-mapping.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 31063. We acknowledge all the participants of UK Biobank. We thank all the members of Finucane lab and Analytic and Translational Genetics Unit (ATGU) at Massachusetts General Hospital for their helpful feedback.

Competing interests

J.C.U. is an employee of Illumina. O.W. is an employee and holds equity in Eleven Therapeutics. B.M.N. is a member of the scientific advisory board at Deep Genomics and

Neumora, consultant of the scientific advisory board for Camp4 Therapeutics and consultant for Merck.B.M.N. is a member of the scientific advisory board at Deep Genomics and Neumora, consultant of the scientific advisory board for Camp4 Therapeutics and consultant for Merck.

Figures and tables

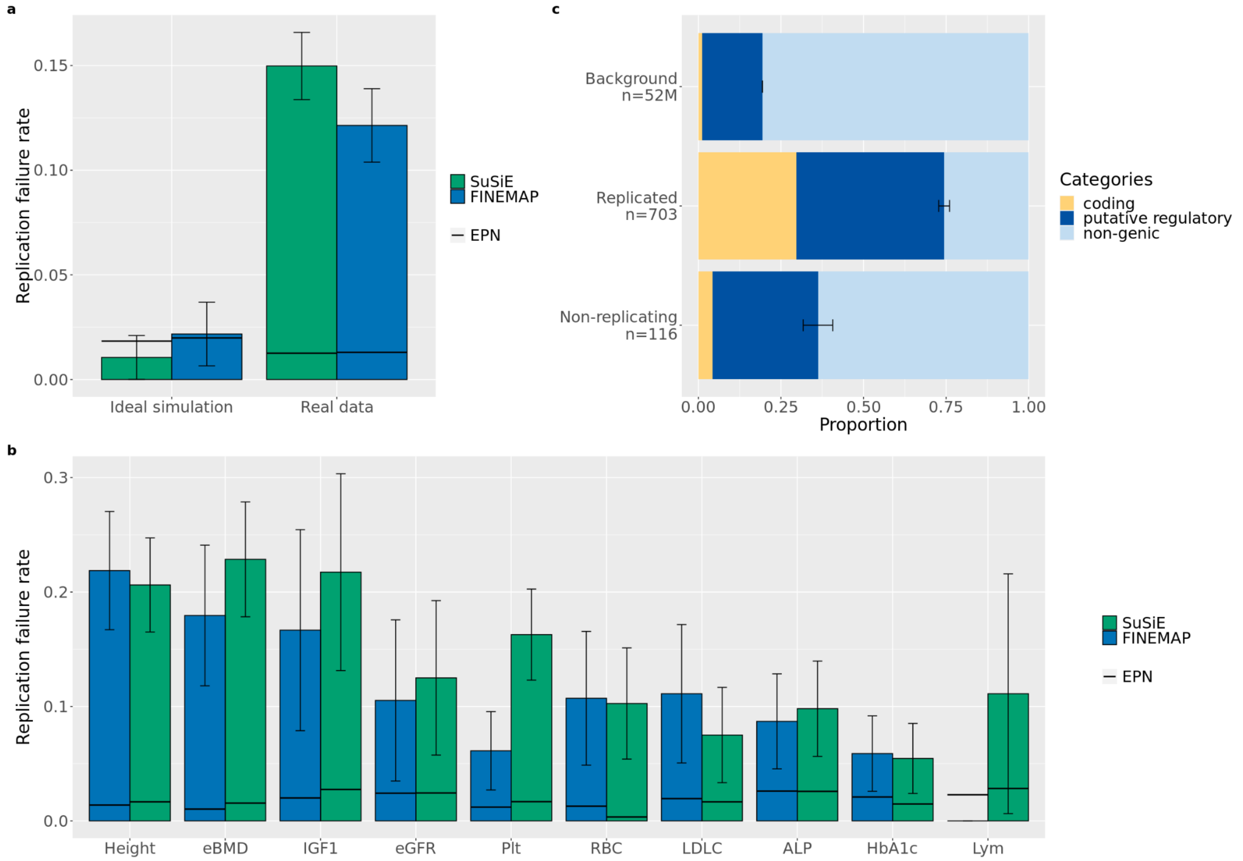


Fig 1 | Replication failure rates and functional enrichments. a. RFRs for SuSiE and FINEMAP aggregated across 10 UKBB quantitative phenotypes and in ideal simulations. **b.** Trait-separated RFRs for SuSiE and FINEMAP. **c.** Functional annotations in 3 disjoint categories: coding, putative regulatory and non-genic (see **Methods** for detailed definitions). Variants are aggregated between SuSiE and FINEMAP. Non-replicating: the set of non-replicating variants (PIP>0.9 at N=100K and PIP<0.1 at N=366K); Replicated: the set of replicated variants (PIP>0.9 at both N=100K and N=366K); Background: the set of all variants included in the fine-mapping analysis, aggregated across 10 traits. n denotes the total number of variants in each set. See **Extended Data Fig. 9 b-e** for method-separated plots and more categories including replicated variants matching PIP with non-replicated variants (**Methods**). Numerical results are available in **Supplementary Table 1,2**.

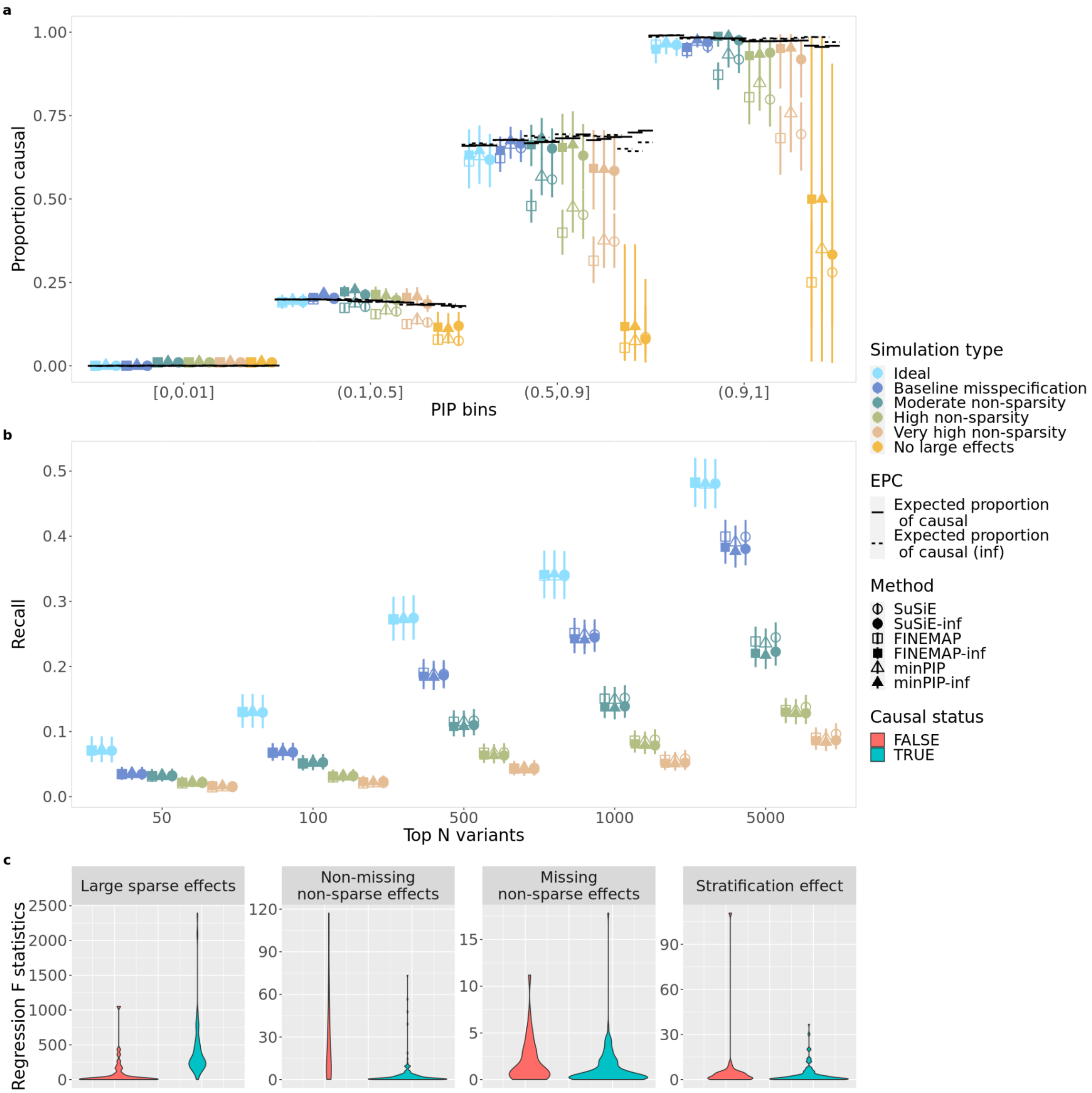


Fig 2 | Non-sparsity simulation. a. Calibration for SuSiE, FINEMAP, minPIP, and corresponding "inf" methods under non-sparsity simulations settings detailed in **Table 1, Methods**. minPIP and minPIP-inf are aggregating methods: minPIP-inf = taking min(PIP) between SuSiE-inf and FINEMAP-inf; minPIP = min(PIP) between SuSiE and FINEMAP. **b.** Recall for these same methods, defined as the percentage of simulated large effects among the top N variants when ranked by PIP. Error bars on calibration and recall plots correspond to 95% Wilson confidence interval. Note that "No large effects" simulations are not shown on the recall plot because there are zero simulated large effects. **c.** Regressing sub-components of "high non-sparsity" phenotype on true vs. false positives (variants with PIP > 0.9 that are either causal or non-causal). Numerical results are available in **Supplementary Table 3-5**.

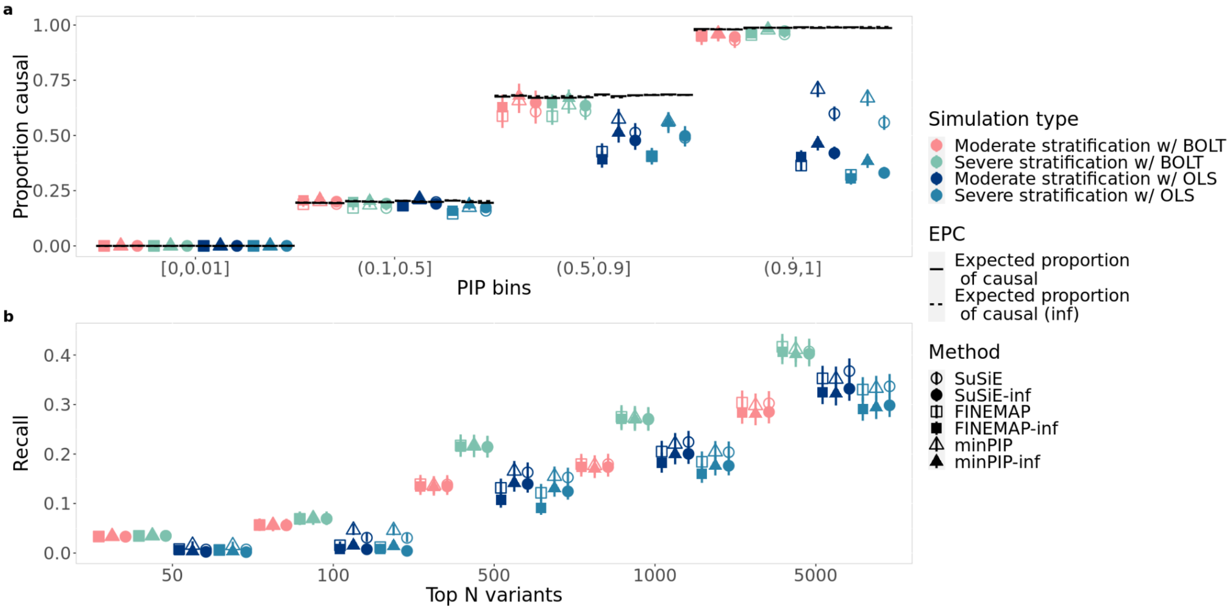


Fig. 3 | Calibration and recall for stratification simulation. a. Calibration plot for six methods in four stratification simulation settings (**Table 1**). **b.** Recall for the same methods and simulations. Numerical results are available in **Supplementary Table 3-4.**

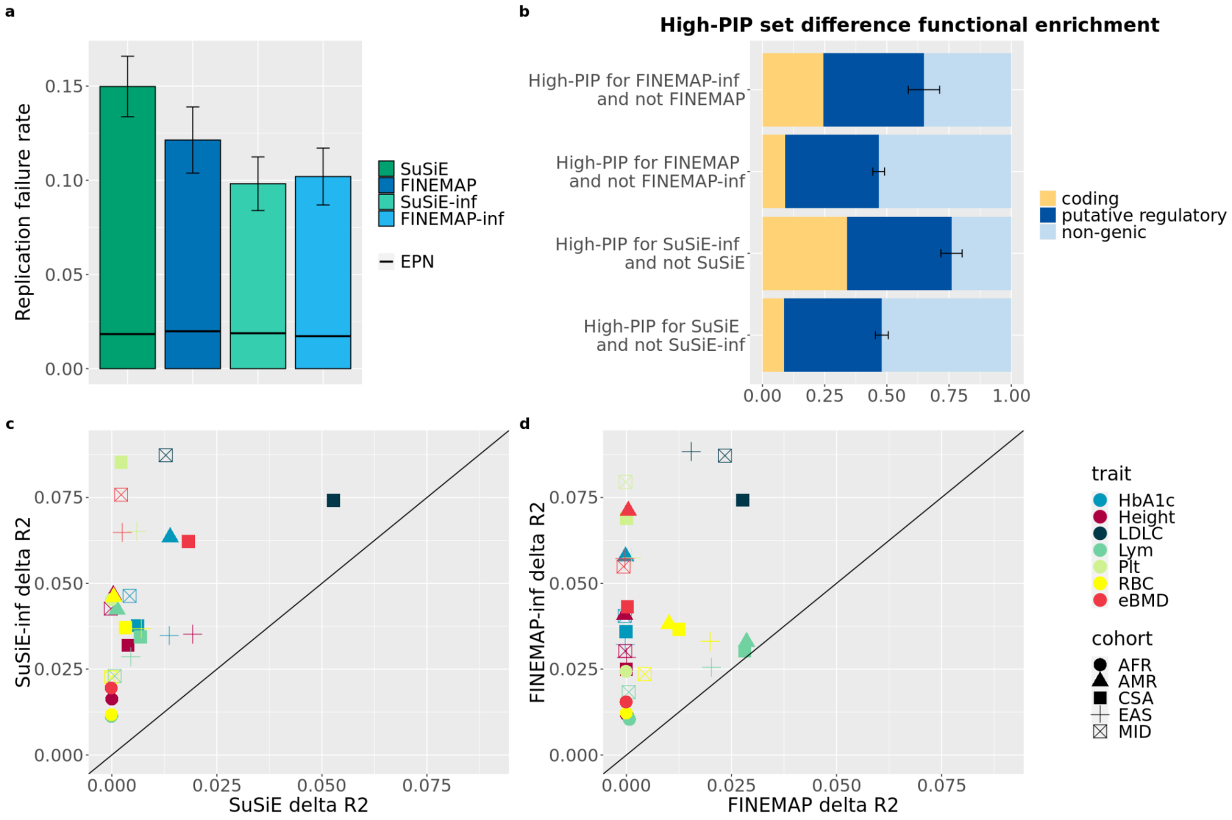
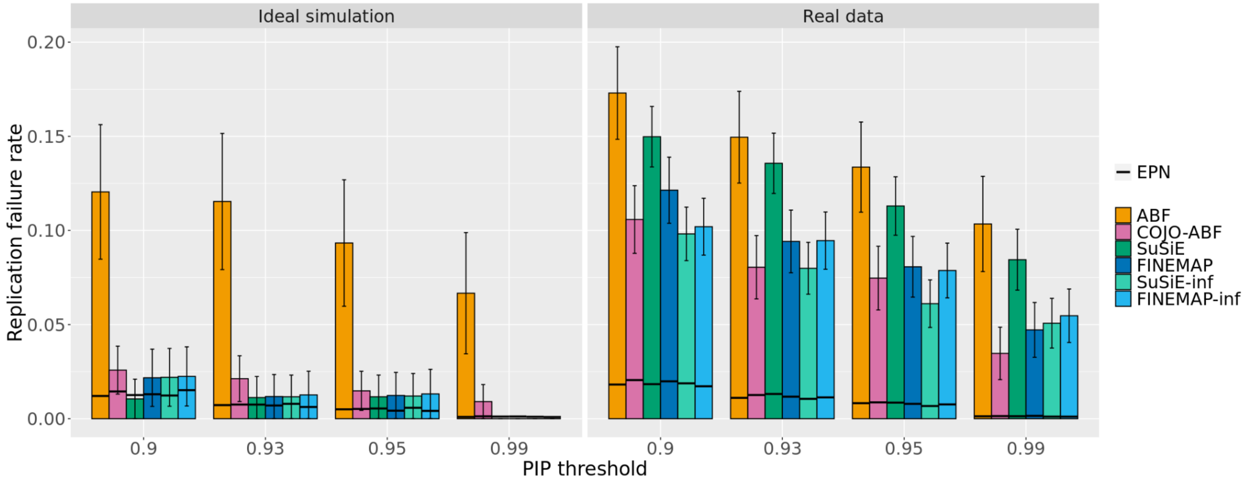


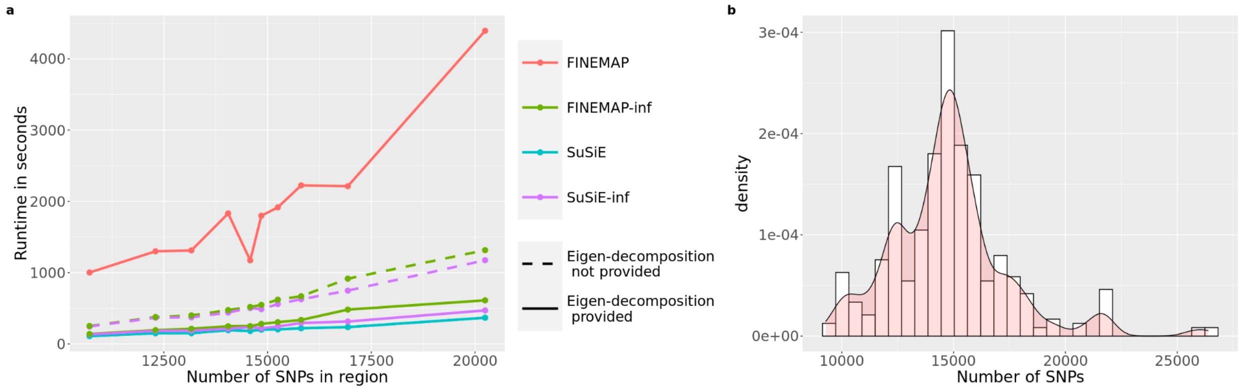
Fig 4 | Real data performance improvements. a. Replication Failure Rates for SuSiE, FINEMAP, SuSiE-inf and FINEMAP-inf aggregated across 10 UKBB traits (**Supplementary Table 1**). **b.** Functional enrichment of the set difference between SuSiE and SuSiE-inf, FINEMAP and FINEMAP-inf (numerical results are available in **Supplementary Table 6**). **c-d.** Delta R^2 comparison of PRS predictions using SuSiE v.s. SuSiE-inf and FINEMAP v.s. FINEMAP-inf (numerical results are available in **Supplementary Table 7**).

	Imputation noise	Sparse causal prior	20 PC effects multiplier	PCs corrected in GWAS	Non-sparse causal effects	Missing causal effects
Ideal	No	Uniform	0	0	None	None
Baseline misspecification	Yes	SuSiE Height posterior	1	19 out of 20	None	None
Moderate stratification w/ BOLT	Yes	SuSiE Height posterior	5	0 out of 20	None	None
Severe stratification w/ BOLT	Yes	SuSiE Height posterior	8	0 out of 20	None	None
Moderate stratification w/ OLS	Yes	SuSiE Height posterior	1	0 out of 20	None	None
Severe stratification w/ OLS	Yes	SuSiE Height posterior	2	0 out of 20	None	None
Moderate non-sparsity	Yes	SuSiE Height posterior	1	19 out of 20	58% of h2	Yes
High non-sparsity	Yes	SuSiE Height posterior	1	19 out of 20	75% of h2	Yes
Very high non-sparsity	Yes	SuSiE Height posterior	1	19 out of 20	83% of h2	Yes
No large effects	Yes	SuSiE Height posterior	1	19 out of 20	100% of h2	Yes

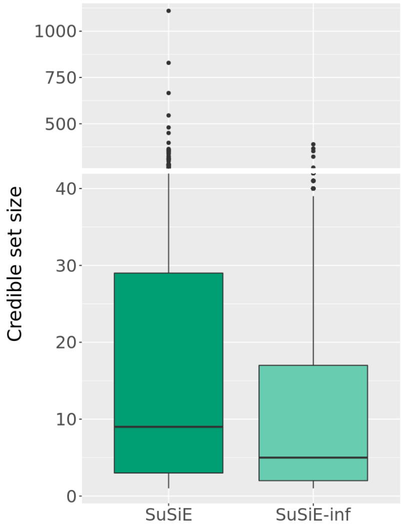
Table 1 | Parameters for large scale simulations. Different parameter settings for ten sets of simulations mentioned in the main text. Note that PCs corrected in GWAS used in-sample (N=150K) PCs as covariates for phenotypes generated with full sample (N=366K) PCs. See **Methods** for details on how each misspecification is incorporated.



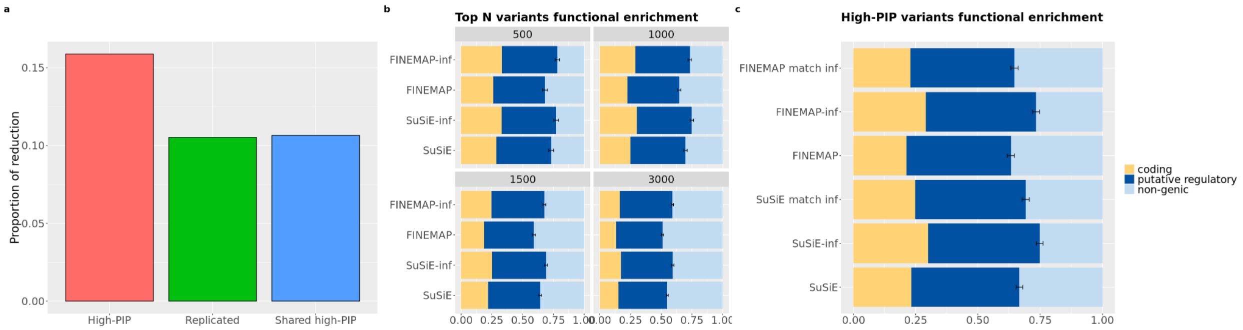
Extended Data Fig. 1 | Replication failure rates at different PIP thresholds. RFR and EPN for six methods in ideal simulations and in real data aggregated across 10 UKBB phenotypes. High-PIP variants defined at four different PIP thresholds: 0.9, 0.93, 0.95, and 0.99. Numerical results available in **Supplementary Table 8**.



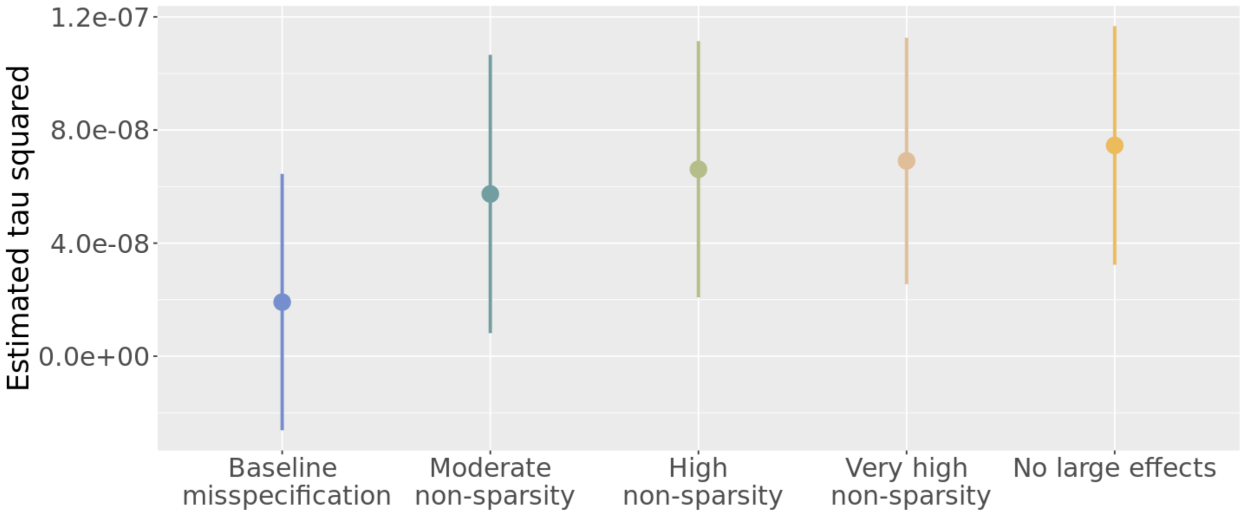
Extended Data Fig. 2 | Runtime comparison. a. Average runtime in ten quantiles based on number of SNPs in fine-mapped region for SuSiE, SuSiE-inf, FINEMAP and FINEMAP-inf, as well as SuSiE-inf and FINEMAP-inf without provided eigen-decomposition of the LD matrix. **b.** Distribution of locus sizes in terms of number of SNPs, aggregated across 10 UKBB phenotypes and across two sample sizes: N=100K and N=366K. Numerical results available in **Supplementary Table 9**.



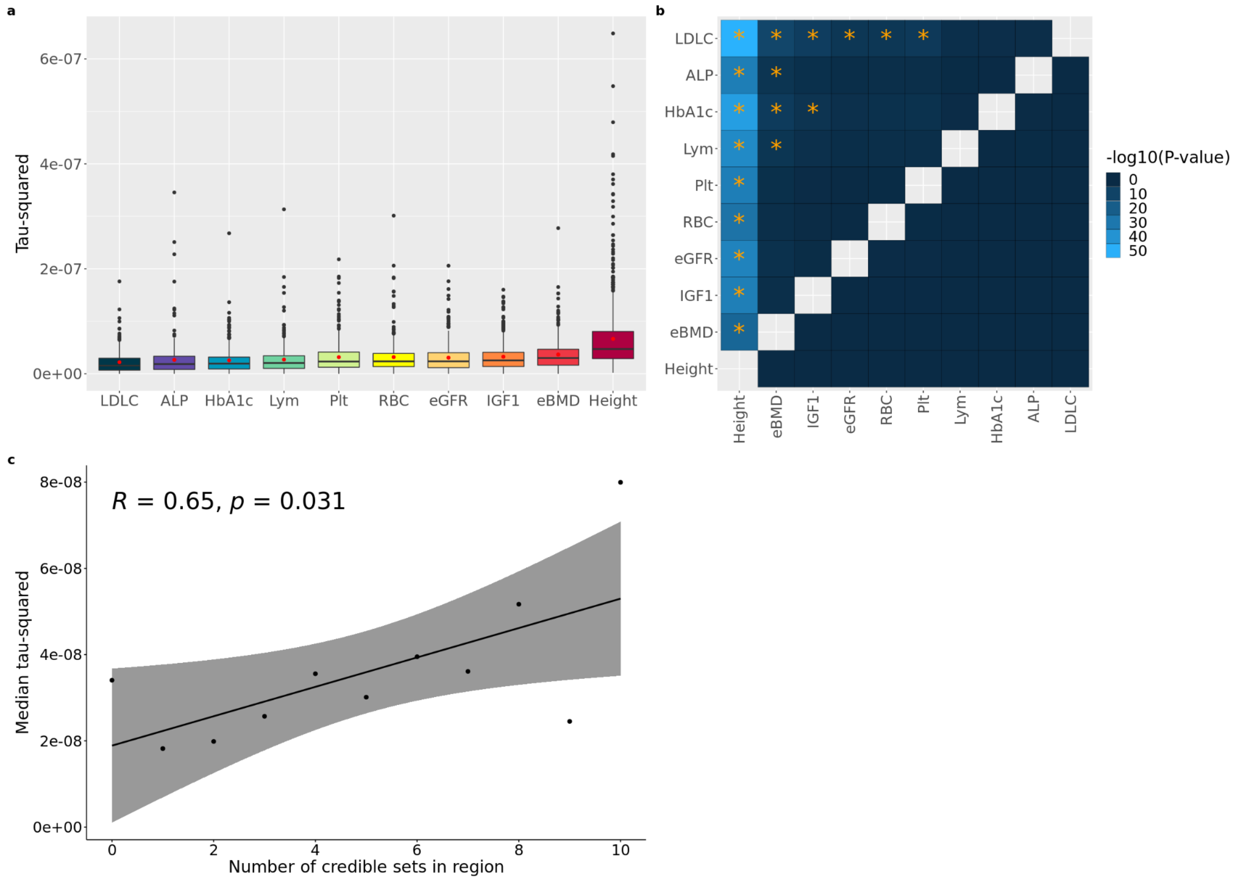
Extended Data Fig. 3 | Credible set sizes. Credible set sizes for SuSiE and SuSiE-inf, aggregated across all fine-mapped regions of 10 UKBB phenotypes. Numerical results available in **Supplementary Table 10**.



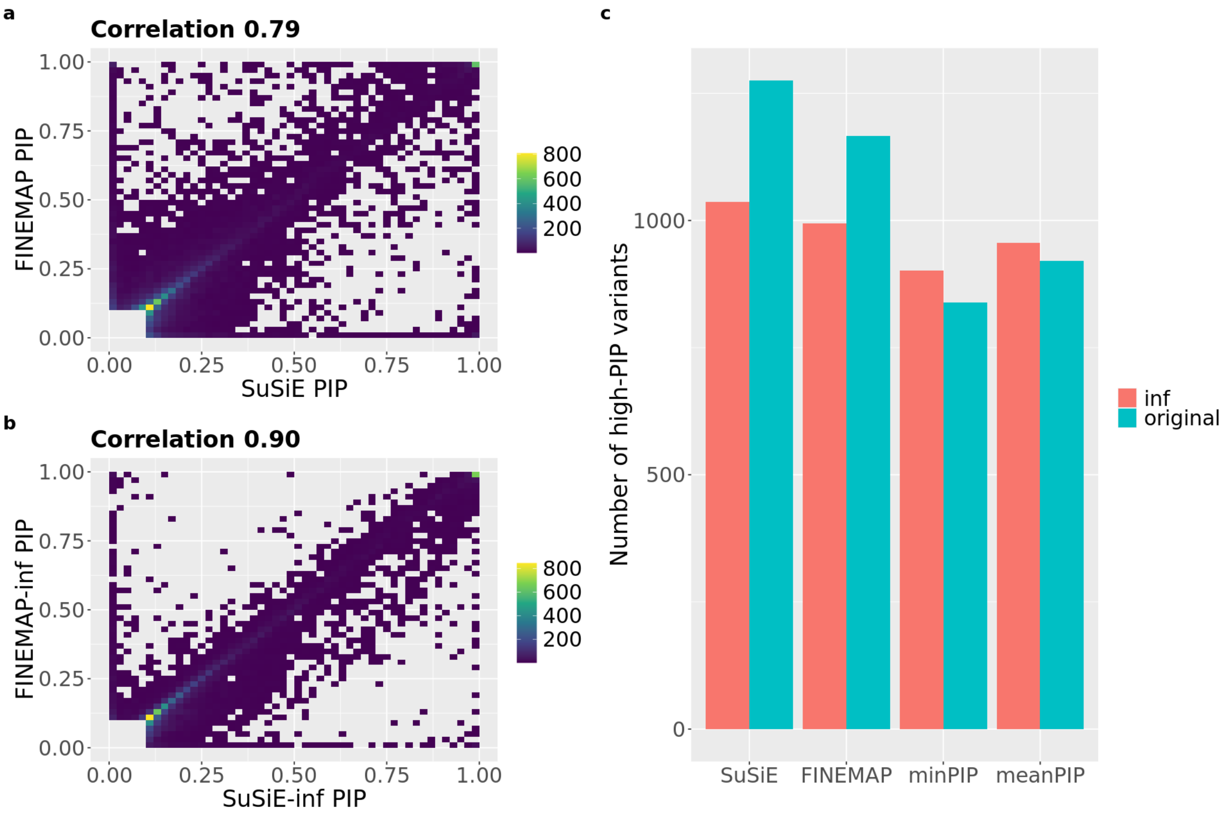
Extended Data Fig. 4 | Performance when taking min(PIP) between methods. a. The proportion of reduction for the number of high-PIP variants aggregated across SuSiE and FINEMAP at N=100K when using either SuSiE-inf or FINEMAP-inf than using either SuSiE or FINEMAP. **b.** Functional enrichment of top 500, 1000, 1500, and 3000 highest PIP variants from different methods. **c.** Functional enrichment of high-PIP variants (PIP>0.9) for SuSiE, FINEMAP, SuSiE-inf and FINEMAP-inf. Additionally, functional enrichment of top N variants of SuSiE (resp. FINEMAP) were plotted, where N matches the number of high-PIP variants of SuSiE-inf (resp. FINEMAP-inf). Numerical results available in **Supplementary Table 11-13**.



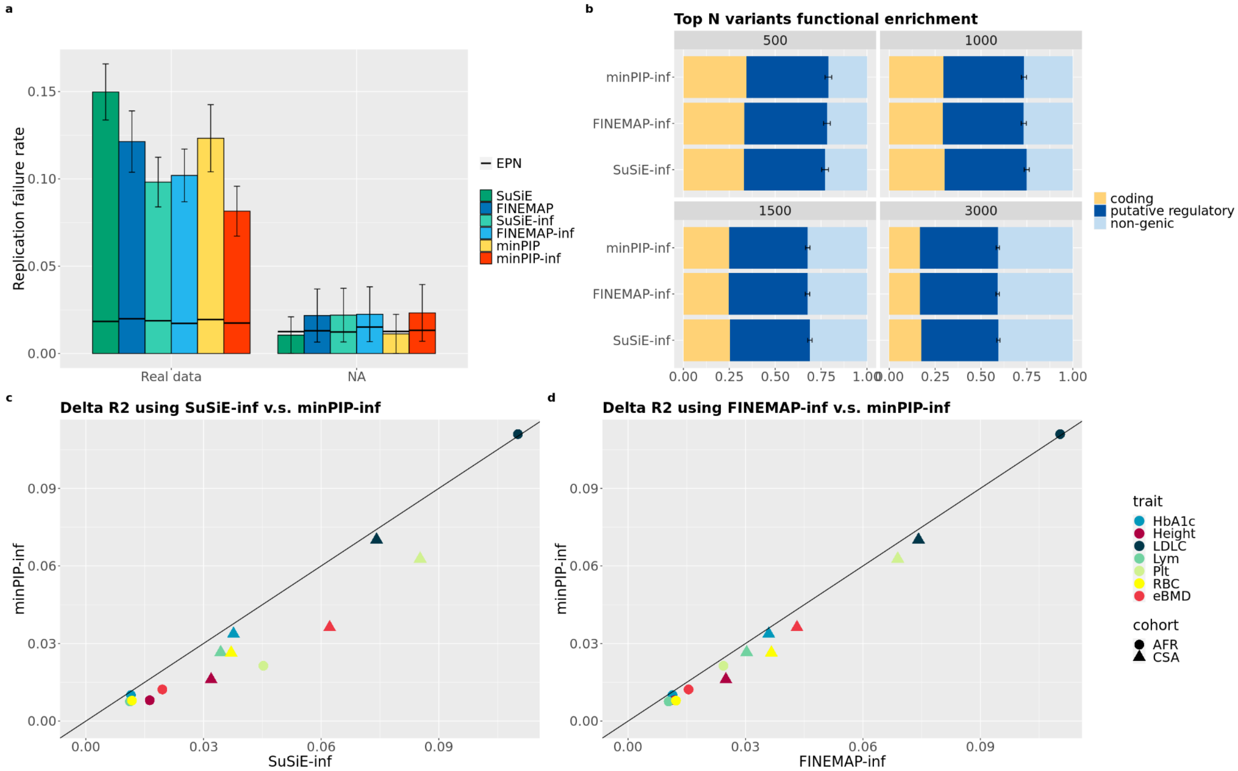
Extended Data Fig. 5 | Estimated tau-squared in non-sparsity simulations. The estimated tau-squared in all regions are plotted for each non-sparse simulation setting. (**Table 1, Methods**). Numerical results available in **Supplementary Table 14**.



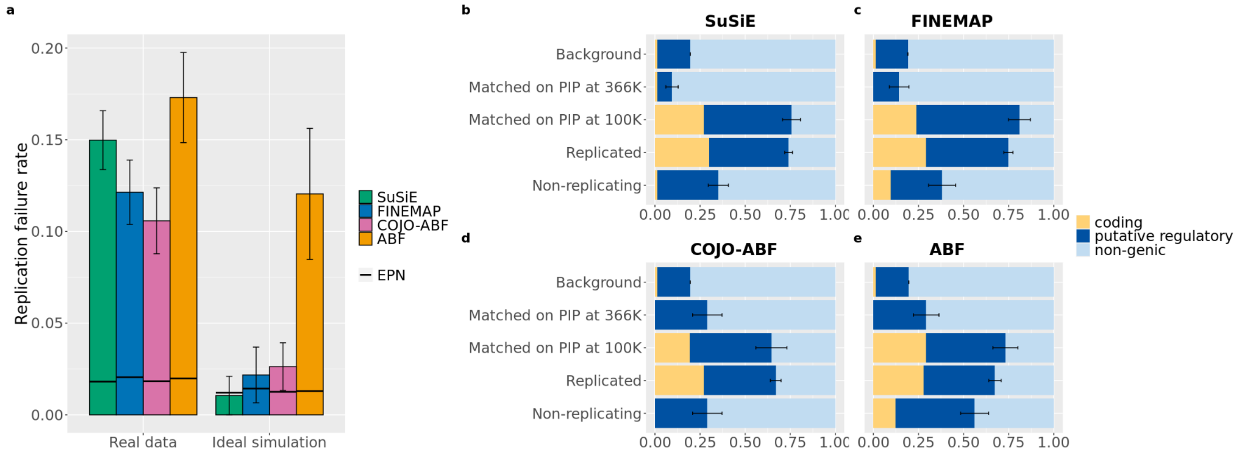
Extended Data Fig. 6 | Estimated tau-squared (infinitesimal variance) in UKBB. a. Boxplot for estimated tau-squared in all fine-mapped regions for 10 tested phenotypes. The line across the boxplots denotes the median, the red dot denotes the mean. **b.** Two-sample T-test with alternative hypothesis: mean of estimated tau-squared for all fine-mapped regions for trait1 (x-axis) is greater than that of trait2 (y-axis). P-value cutoff is set to be 0.05/90 = 5.5e-4, correcting for the total number of trait pairs tested. Stars indicate P-values pass the significant threshold. **c.** Correlation between number of credible sets and the infinitesimal variance. Medians of tau-squared are computed for regions with the same number of credible sets. R is the Pearson correlation, p is the correlation p-value. The 95% confidence interval is shown on the plot as the gray shaded area. Numerical results available in **Supplementary Table 15-17**.



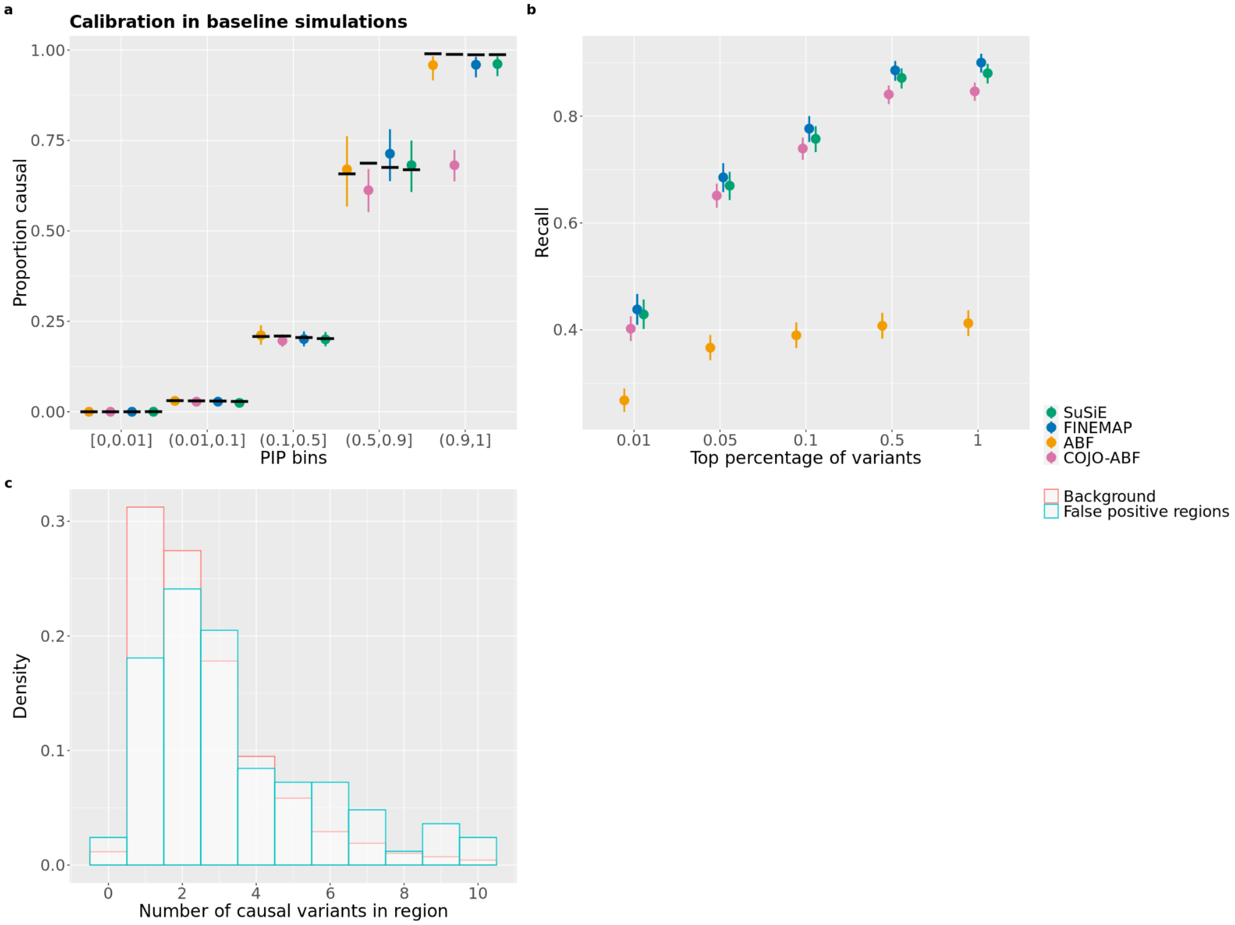
Extended Data Fig. 7 | Agreement between SuSiE and FINEMAP v.s. SuSiE-inf and FINEMAP-inf. a-b. Density plot of PIPs from 10 UKBB traits fine-mapped at N=366K. All variants with PIP>=0.1 for either method are shown on the density plots. **c.** Number of high-PIP (PIP>0.9) variants identified by SuSiE, SuSiE-inf, FINEMAP, FINEMAP-inf, minPIP, minPIP-inf, meanPIP and meanPIP-inf, where meanPIP(-inf) is defined as taking the average PIP between SuSiE and FINEMAP (resp. SuSiE-inf and FINEMAP-inf). Data aggregated across 10 UKBB traits fine-mapped at N=366K. Numerical data available in **Supplementary Table 18**.



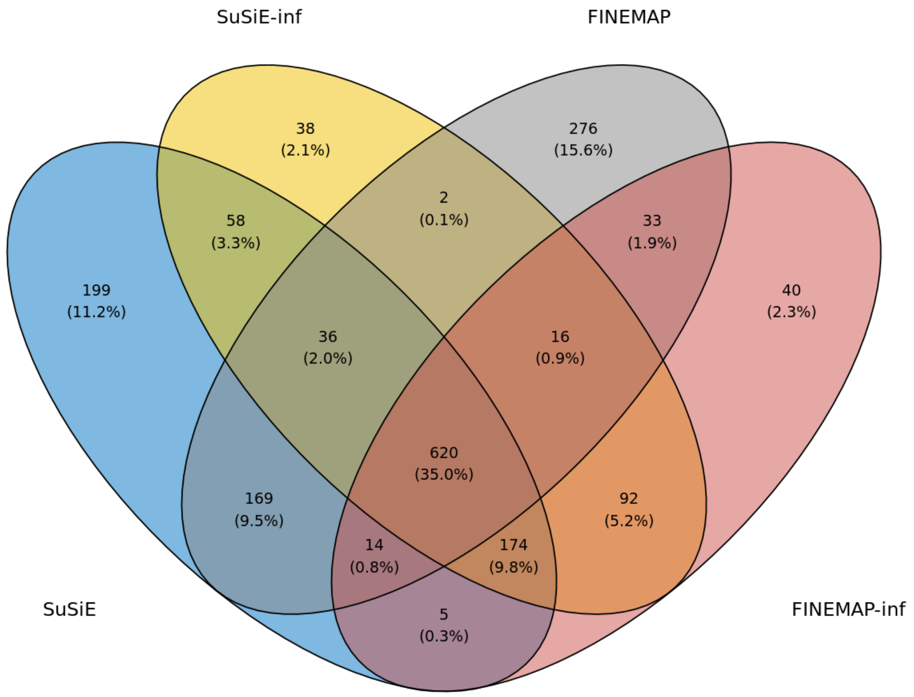
Extended Data Fig. 8 | minPIP-inf performance a. Replication failure rate of minPIP and minPIP-inf compared to other methods (**Supplementary Table 1**). minPIP = taking min(PIP) between SuSiE & FINEMAP; minPIP-inf = taking min(PIP) between SuSiE-inf and FINEMAP-inf. **b.** Functional enrichment of top 500, 1000, 1500, and 3000 highest PIP variants from different methods (**Supplementary Table 12**). **c-d.** Comparison of PRS accuracy (measured by Delta R^2) of selected cohorts between minPIP-inf and SuSiE-inf (resp. FINEMAP-inf) (**Supplementary Table 7**).



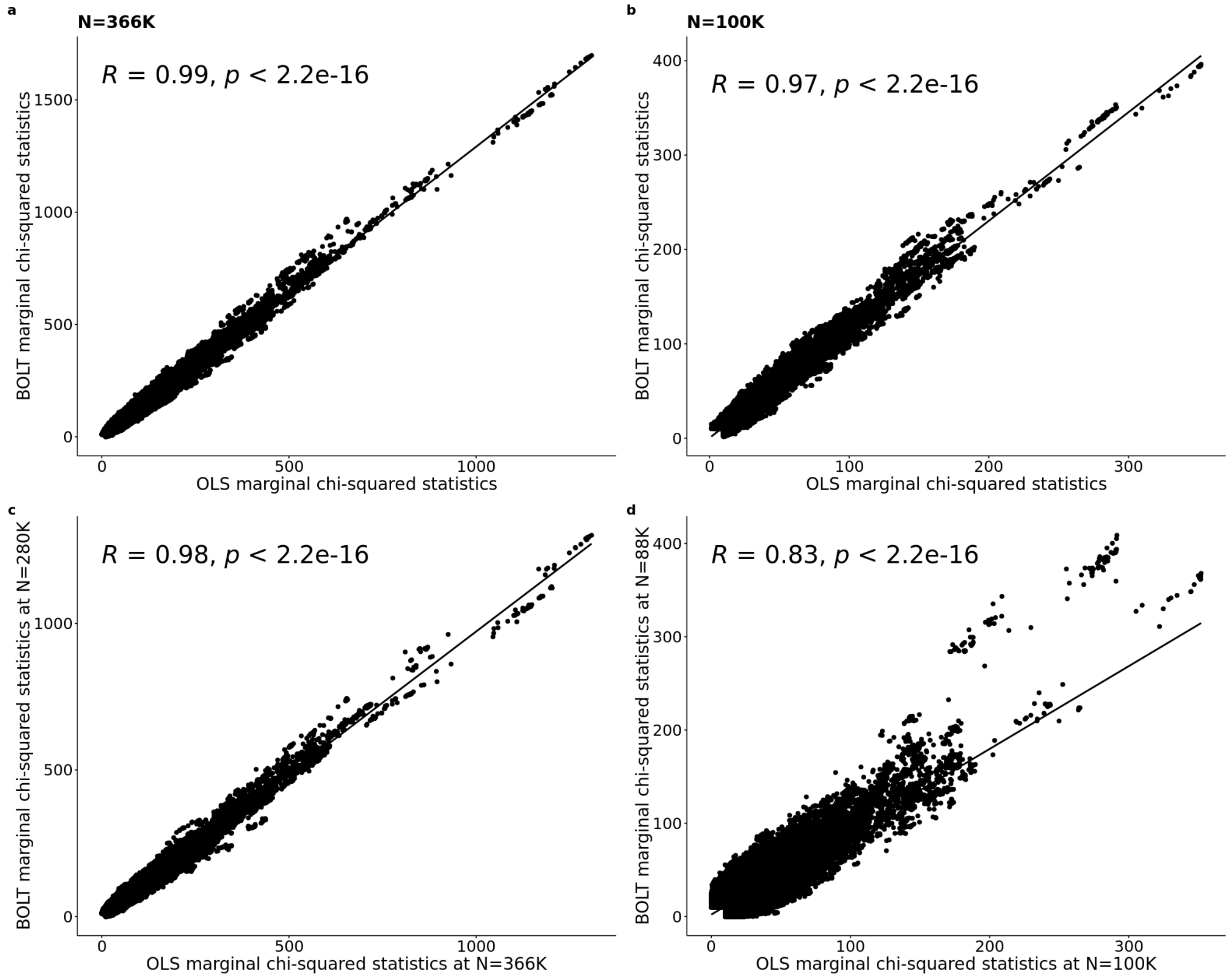
Extended Data Fig. 9 | RFR and functional enrichment of SuSiE, FINEMAP, ABF and COJO-ABF. a. RFR in real data (aggregated across 10 UKBB phenotypes) and in ideal simulations. **b.** Functional enrichment for 5 groups of variants. See **Supplementary Methods** for the definitions of these groups. (**Supplementary Table 1-2**).



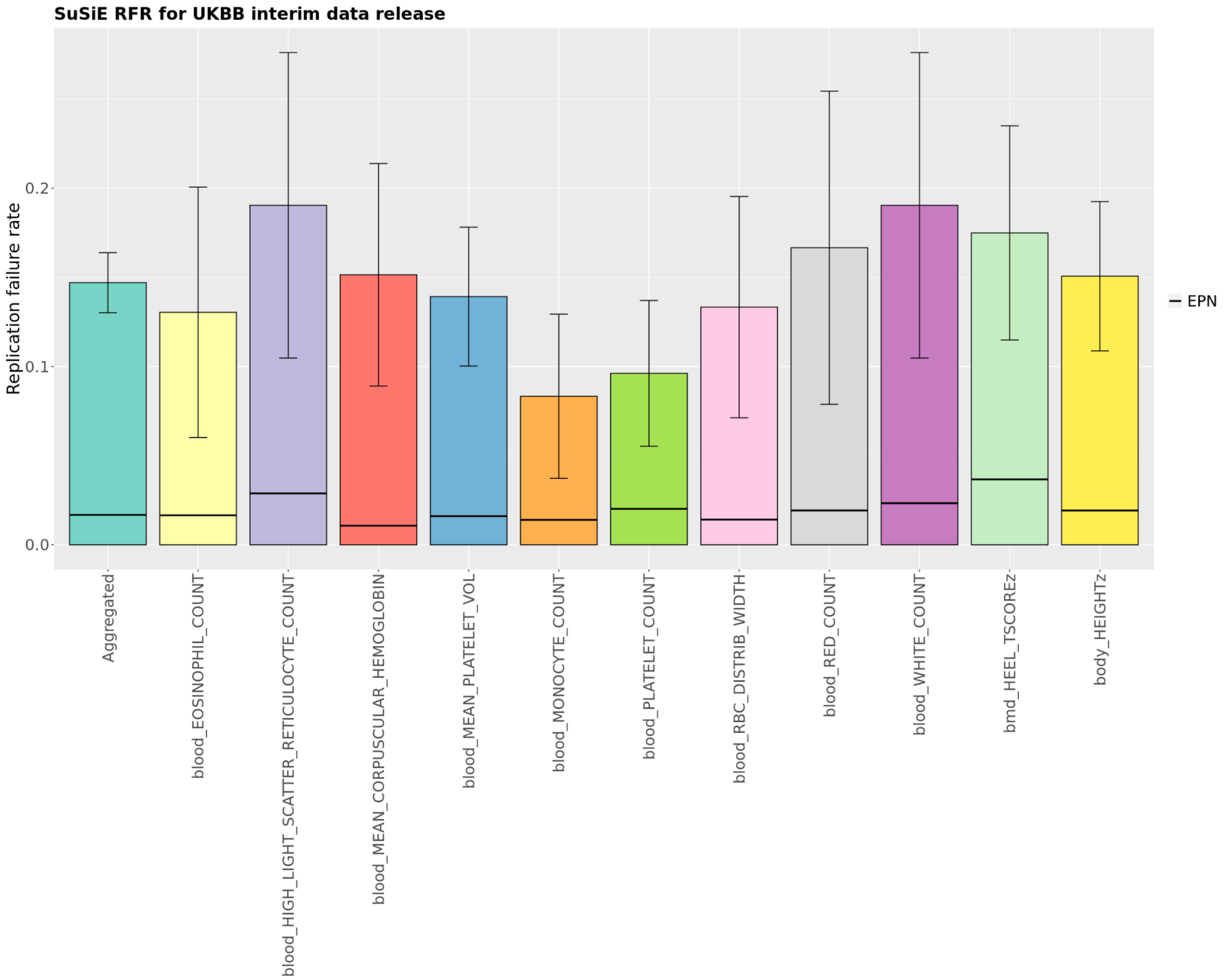
Extended Data Fig. 10 | ABF and COJO-ABF in ideal simulations. a. Calibration of ABF and COJO-ABF in ideal simulations. **b.** Recall of ABF and COJO-ABF for the top 0.01%, 0.05%, 0.1%, 0.5% and 1% SNPs, ordered by PIP. **c.** Distribution of the number of causal variants per region in regions containing COJO-ABF false positive SNPs compared to all regions. Numerical results available in **Supplementary Table 19-21**.



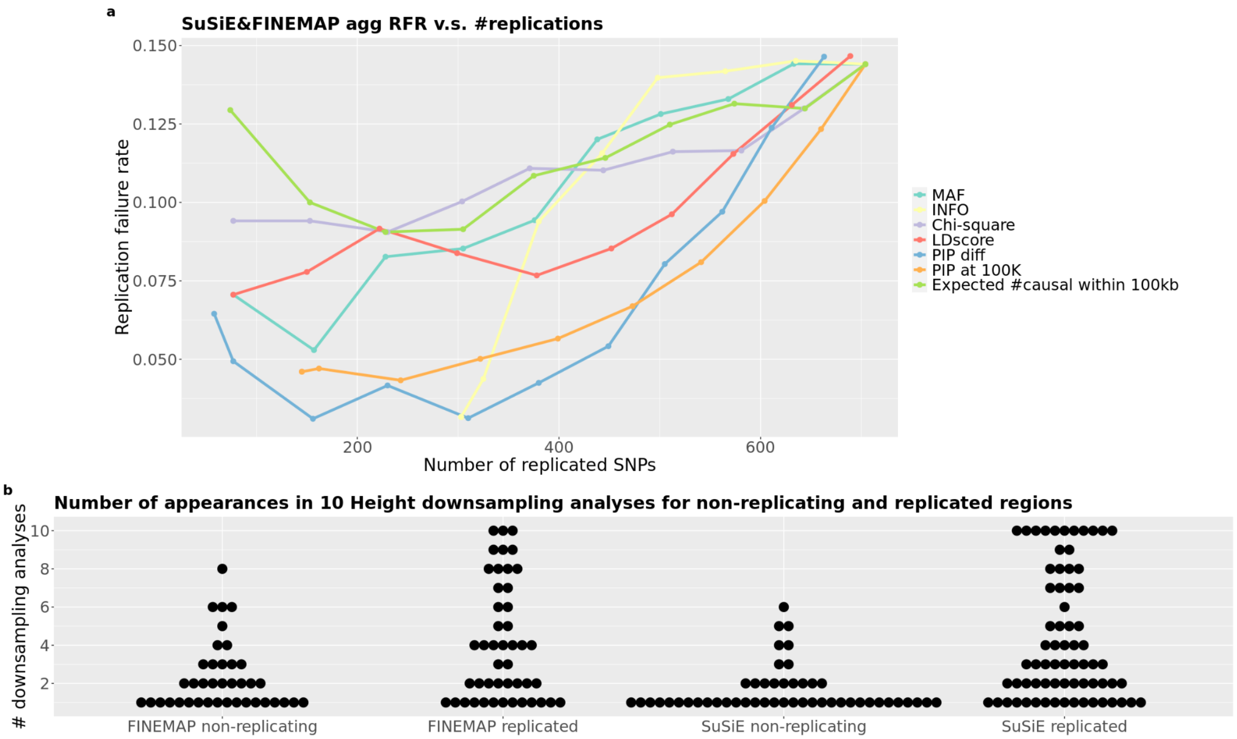
Supplementary Fig. 1 | Venn diagram of high-PIP variants for four methods. High-PIP is defined as PIP>0.9, aggregated across 10 UKBB phenotypes at N=366K. Data available in **Supplementary Table 25**.



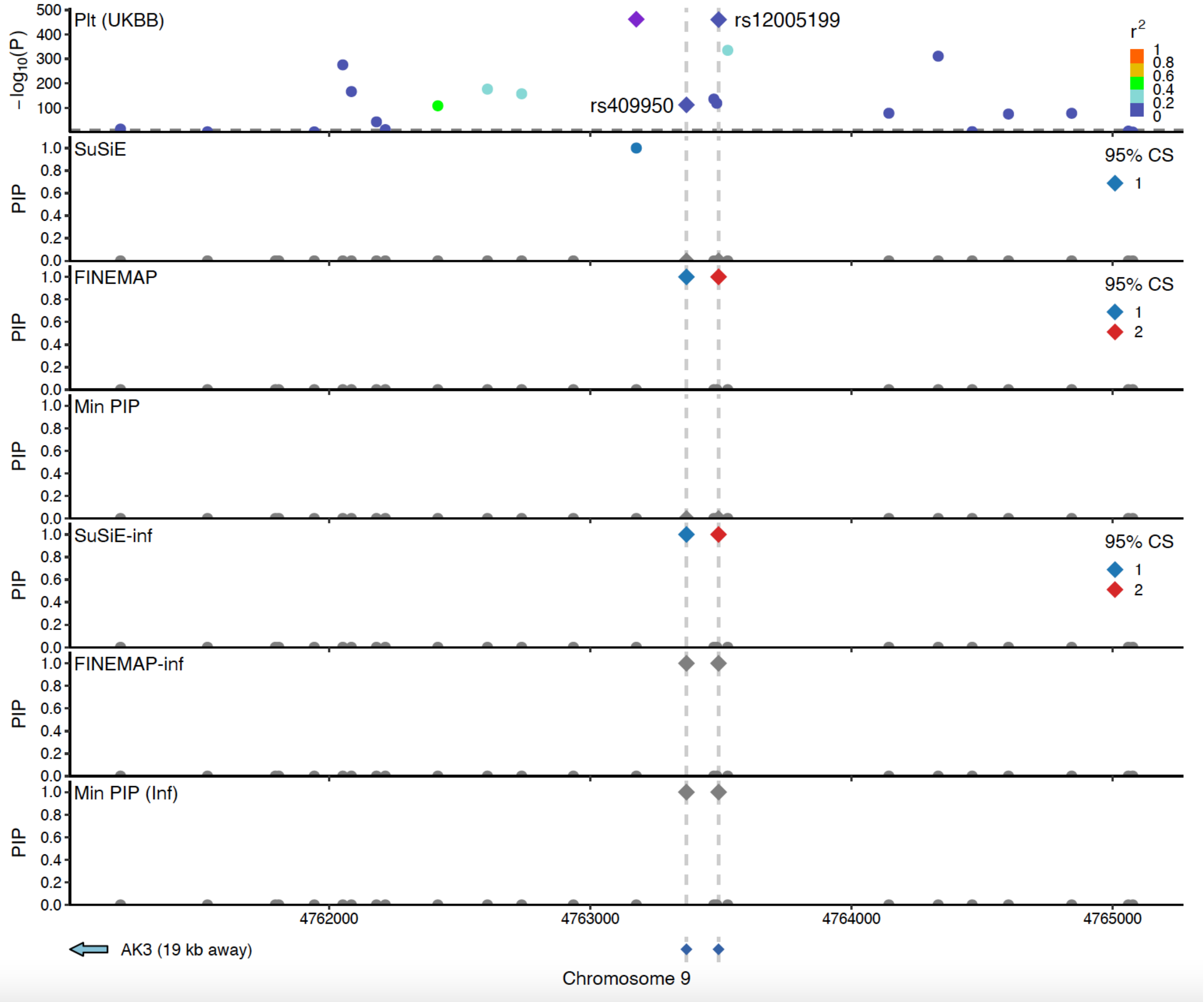
Supplementary Fig. 2 | Marginal chi-squared statistics comparison at different sample sizes. a-b. Marginal chi-squared statistics at N=366K and N=100K using OLS and BOLT-LMM. **c.** Marginal chi-squared statistics using OLS at N=366K and using BOLT at N=280K. **d.** Marginal chi-squared statistics using OLS at N=100K and using BOLT at N=88K. Filtered to Chi-squared statistics greater than 10 for either method.



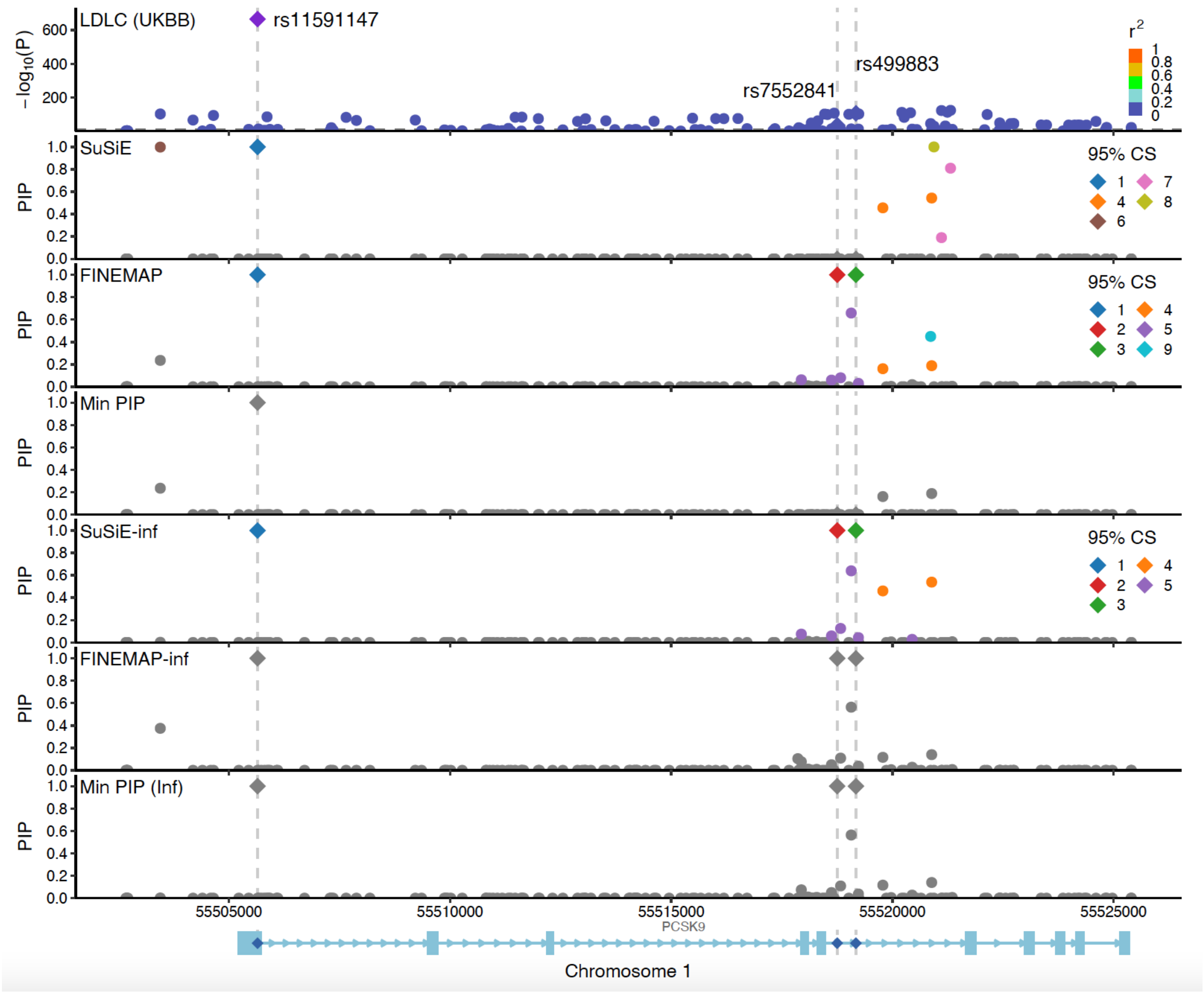
Supplementary Fig. 3 | SuSiE RFR for UK Biobank interim data release. Replication failure rates for 11 traits were computed using data from previously published work[36]. Fine-mapping was performed using SuSiE on the UK Biobank interim release (N=107K) and on the full (defined in [36]) N=337K UK Biobank data. See[36] for phenotype definitions. Numerical results available in **Supplementary Table 26**.



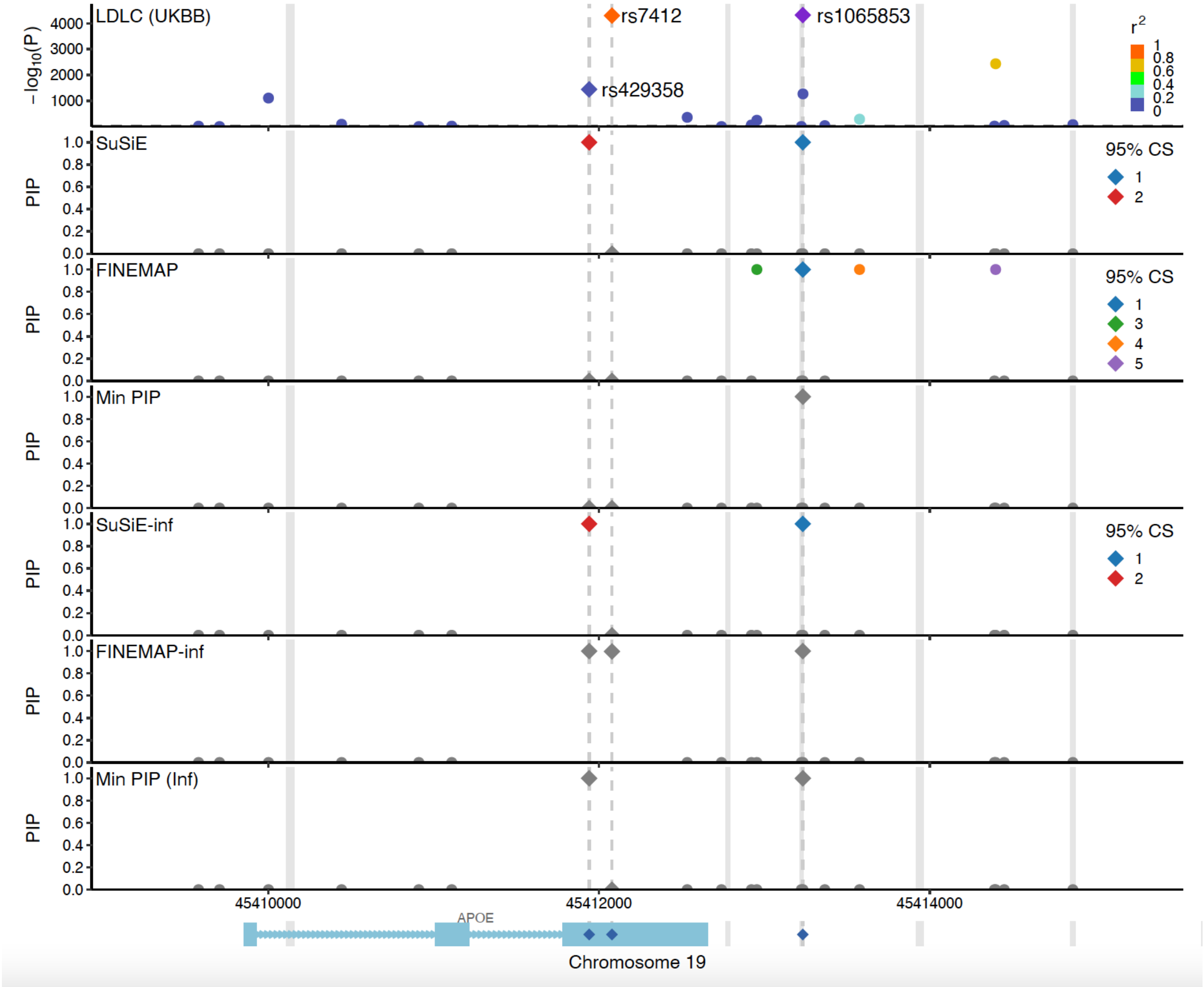
Supplementary Fig 4 | Further investigations into non-replication. a. Number of replicated SNPs and RFR are computed for 10 thresholding values of 7 SNP properties (value as the lower bound: INFO score, Marginal association chi-squared statistic, PIP at 100K, and expected number of causal variants within 100Kb; value as upper bound: MAF, LD score, SuSiE FINEMAP PIP difference. Whether to use value as lower or upper bound was determined by which setting gives better performance). **b.** Dot plot for the number of occurrences in 10 Height downsampling analyses for replicated/non-replicating regions (region definitions taken at N=366K), each dot represents one region. Numerical results available in **Supplementary Table 27-28**.



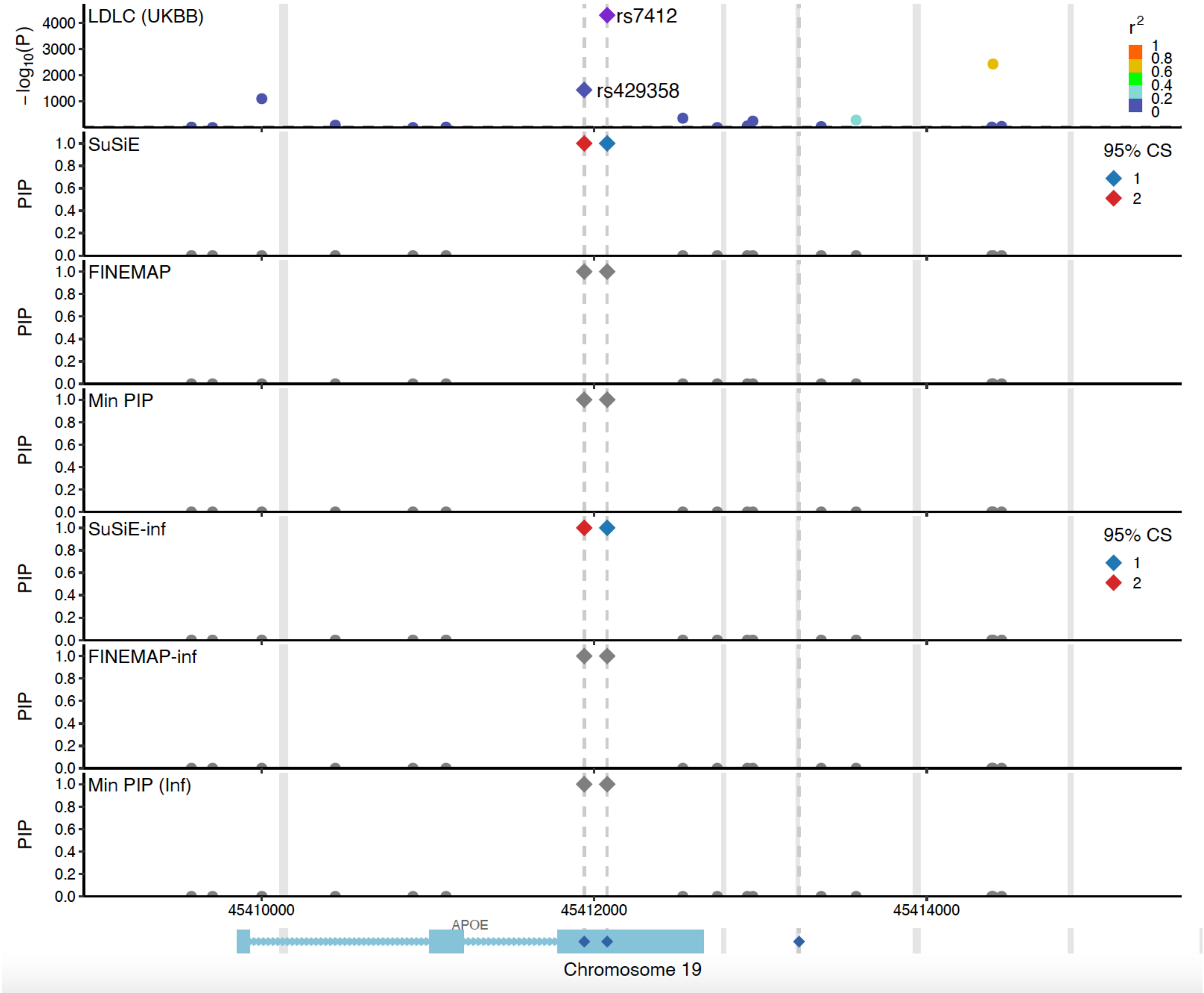
Supplementary Fig. 5 | AK3 locus. 4kbp window near the AK3 gene is shown on the plot, GWAS -log10 P-values for trait Plt are plotted on the top panel, PIPs from 4 fine-mapping methods and 2 aggregating methods are plotted on the subsequent panels. Variants rs12005199 and rs409950 are consistently detected with high confidence by FINEMAP, SuSiE-inf and FINEMAP-inf but not SuSiE, resulting in high confidence when taking min PIP between SuSiE-inf and FINEMAP-inf but low confidence when taking min PIP between SuSiE and FINEMAP. These two variants replicate previous findings in [9] where a luciferase assay was used as orthogonal evidence for variant causality.



Supplementary Fig. 6 | PCSK9 locus. 23kbp window at the PCSK9 gene location is shown on the plot. GWAS -log10 P-values for trait LDLC are plotted on the top panel, PIPs from 4 fine-mapping methods and 2 aggregating methods are plotted on the subsequent panels. In addition to the well-known causal variant rs11591147, SuSiE-inf and FINEMAP-inf consistently identified two intronic variants: rs499883 and rs7552841 with high confidence. SuSiE did not identify variant rs499883, however, previous fine-mapping results in [36] showed high confidence (PIP=1.0) after applying SuSiE with functional priors. Our results replicate this finding without using functional priors.



Supplementary Fig. 7 | APOE locus including variants in LCR. 6kbp window at the APOE gene location is shown on the plot. GWAS -log10 P-values for trait LDLC are plotted on the top panel, PIPs from 4 fine-mapping methods and 2 aggregating methods are plotted on the subsequent panels. Gray areas denote low-complexity regions (LCR). Variant rs1065853 is in LCR and with high LD to the known causal missense variant rs7412. With LCR included, only FINEMAP-inf was able to identify rs7412 as a high confidence variant. The second known causal variant rs429358 is identified by SuSiE, SuSiE-inf and FINEMAP-inf but not FINEMAP. Taking min PIP between SuSiE and FINEMAP did not capture any of the two known causal variants, whereas minPIP-inf captured one of the two.



Supplementary Fig. 8 | APOE locus excluding variants in LCR. 6kbp window at the APOE gene location is shown on the plot. GWAS -log10 P-values for trait LDLC are plotted on the top panel, PIPs from 4 fine-mapping methods and 2 aggregating methods are plotted on the subsequent panels. Gray areas denote low-complexity regions (LCR). When fine-mapping without variants in LCR, all four methods correctly identified rs7412 and rs429358 as causal variants.

References

1. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
2. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
3. Benner, C., Havulinna, A. S., Salomaa, V., Ripatti, S. & Pirinen, M. Refining fine-mapping: effect sizes and regional heritability. *bioRxiv* 318618 (2018) doi:10.1101/318618.
4. Wakefield, J. Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.* **33**, 79–86 (2009).
5. 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).
6. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
7. 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).
8. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
9. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
10. Westra, H.-J. *et al.* Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat. Genet.* **50**, 1366–1374 (2018).

11. Zhang, Z. *et al.* Genetic analyses support the contribution of mRNA N6-methyladenosine (m6A) modification to human disease heritability. *Nat. Genet.* **52**, 939–949 (2020).
12. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
13. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine-Progress, Pitfalls, and Promise. *Cell* **177**, 45–57 (2019).
14. Hukku, A. *et al.* Probabilistic colocalization of genetic variants from complex and molecular traits: promise and limitations. *Am. J. Hum. Genet.* **108**, 25–35 (2021).
15. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
16. LaPierre, N. *et al.* Identifying causal variants by fine mapping across multiple studies. *PLoS Genet.* **17**, e1009733 (2021).
17. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
18. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
19. 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).
20. Ulirsch, J. C. & Kanai, M. An annotated atlas of causal variants for complex human traits. *In prep*.
21. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
22. Kanai, M. *et al.* Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *bioRxiv* (2022) doi:10.1101/2022.03.16.22272457.
23. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).

24. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

25. Erbe, M. *et al.* Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **95**, 4114–4129 (2012).

26. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).

27. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

28. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).

29. Pan-UKB team. https://pan.ukbb.broadinstitute.org (2020).

30. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv* (2021) doi:10.1101/2021.09.03.21262975.

31. Bates, S., Candès, E., Janson, L. & Wang, W. Metropolized Knockoff Sampling. *J. Am. Stat. Assoc.* **116**, 1413–1427 (2021).

32. Sesia, M., Katsevich, E., Bates, S., Candès, E. & Sabatti, C. Multi-resolution localization of causal variants across the genome. *Nat. Commun.* **11**, 1093 (2020).

33. Candès, E., Fan, Y., Janson, L. & Lv, J. Panning for gold: 'model‐X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Series B Stat. Methodol.* **80**, 551–577 (2018).

34. Newcombe, P. J., Conti, D. V. & Richardson, S. JAM: A Scalable Bayesian Framework for Joint Analysis of Marginal SNP Effects. *Genet. Epidemiol.* **40**, 188–201 (2016).

35. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. *Am. J. Hum. Genet.* **101**,

539–551 (2017).

36. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

37. Schoech, A. P. *et al.* Quantification of frequency-dependent genetic architectures in 25 UK Biobank traits reveals action of negative selection. *Nat. Commun.* **10**, 790 (2019).

38. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

39. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* vol. 4 Preprint at https://doi.org/10.1186/s13742-015-0047-8 (2015).

40. Weissbrod, Kanai, Shi, Gazal & Peyrot. Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores. *MedRxiv*.