



Published in final edited form as:

Nat Genet. 2020 December ; 52(12): 1355–1363. doi:10.1038/s41588-020-00735-5.

Functionally-informed fine-mapping and polygenic localization of complex trait heritability

Omer Weissbrod^{1,*}, Farhad Hormozdiari¹, Christian Benner², Ran Cui³, Jacob Ulirsch^{3,4}, Steven Gazal¹, Armin P. Schoech¹, Bryce van de Geijn¹, Yakir Reshef¹, Carla Márquez-Luna⁵, Luke O'Connor³, Matti Pirinen^{2,6,7}, Hilary K. Finucane^{3,8}, Alkes L. Price^{1,3,*}

¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA ²Institute for Molecular Medicine Finland (FIMM), Helsinki Institute of Life Sciences, University of Helsinki, Helsinki, Finland ³Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA ⁴Program in Biological and Biomedical Sciences, Harvard Medical School, Cambridge, MA ⁵The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY ⁶Department of Public Health, University of Helsinki, Helsinki, Finland ⁷Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland ⁸Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA

Abstract

Fine-mapping aims to identify causal variants impacting complex traits. We propose PolyFun, a computationally scalable framework to improve fine-mapping accuracy by leveraging functional annotations across the entire genome—not just genome-wide significant loci—to specify prior probabilities for fine-mapping methods such as SuSiE or FINEMAP. In simulations, PolyFun+SuSiE and PolyFun+FINEMAP were well-calibrated and identified >20% more variants with posterior causal probability >0.95 than their non-functionally informed counterparts. In analyses of 49 UK Biobank traits (average $N=318K$), PolyFun+SuSiE identified 3,025 fine-mapped variant-trait pairs with posterior causal probability >0.95, a >32% improvement vs. SuSiE. We used posterior mean per-SNP heritabilities from PolyFun+SuSiE to perform polygenic localization, constructing minimal sets of common SNPs causally explaining 50% of common SNP heritability; these sets ranged in size from 28 (hair color) to 3,400 (height) to 2 million (number of children). In conclusion, PolyFun prioritizes variants for functional follow-up and provides insights into complex trait architectures.

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

*Corresponding authors: Omer Weissbrod (oweissbrod@hsph.harvard.edu); Alkes Price (aprice@hsph.harvard.edu).

Author contributions

O.W. and A.L.P. designed the study; O.W. and S.G. analyzed data; C.B. extended the FINEMAP software; O.W. and A.L.P. wrote the manuscript with assistance from F.H., C.B., R.C., J.U., S.G., A.P.S., B.v.d.G., Y.R., C.M.L., L.O., M.P., and H.K.F.

Competing interests

The authors declare no competing interests.

Genome-wide association studies of complex traits have been extremely successful in identifying loci harboring causal variants but less successful in identifying the underlying causal variants, making the development of fine-mapping methods a key priority^{1,2}. The power of fine-mapping methods^{3–12} is limited due to strong linkage disequilibrium (LD), but it can be increased by prioritizing variants in functional annotations that are enriched for complex trait heritability^{7,8,10,13–17}. However, previous functionally-informed fine-mapping methods^{18–20} have computational limitations and can only use genome-wide significant loci to estimate functional enrichment (or can only incorporate a small number of functional annotations¹⁰), severely limiting the benefit of functional data.

We propose PolyFun, a computationally scalable framework for functionally-informed fine-mapping that makes full use of genome-wide data by specifying prior causal probabilities for fine-mapping methods such as SuSiE²¹ or FINEMAP^{22,23}. PolyFun estimates functional enrichment using a broad set of coding, conserved, regulatory, MAF and LD-related annotations from the baseline-LF model^{24–26}.

We show in simulations with in-sample LD that PolyFun is well-calibrated and is more powerful than previous fine-mapping methods, with a >20% power increase over non-functionally informed fine-mapping methods. In simulations with mismatched reference LD, PolyFun remains well-calibrated when reducing the maximum number of assumed causal SNPs per locus. We apply PolyFun to 49 complex traits from the UK Biobank²⁷ (average $N=318K$) with in-sample LD and identify 3,025 fine-mapped variant-trait pairs with posterior causal probability >0.95, spanning 2,225 unique variants. 223 of these variants were fine-mapped for multiple genetically uncorrelated traits, indicating pervasive pleiotropy. We further used posterior mean per-SNP heritabilities from PolyFun + SuSiE to perform polygenic localization, finding sets of common SNPs causally explaining 50% of common SNP heritability that range in size across many orders of magnitude, from dozens to millions of SNPs.

Results

Overview of methods

PolyFun prioritizes variants in enriched functional annotations by specifying prior causal probabilities in proportion to predicted per-SNP heritabilities and providing them as input to fine-mapping methods such as SuSiE²¹ or FINEMAP^{22,23}. For each target locus, PolyFun robustly specifies prior causal probabilities for all SNPs on the corresponding odd (resp. even) target chromosome by (1) estimating functional enrichments for a broad set of coding, conserved, regulatory and LD-related annotations from the baseline-LF 2.2.UKB model²⁵ (187 annotations; Methods, Supplementary Table 1) using an L2-regularized extension of S-LDSC¹⁷, restricted to even (resp. odd) chromosomes; (2) estimating per-SNP heritabilities for SNPs on odd (resp. even) chromosomes using the functional enrichment estimates from step 1; (3) partitioning all SNPs into 20 bins of similar estimated per-SNP heritabilities from step 2; (4) re-estimating per-SNP heritabilities for all SNPs on the target chromosome by applying S-LDSC to the 20 bins, restricted to odd (resp. even) chromosomes excluding the target chromosome; and (5) setting prior causal probabilities for SNPs on the target chromosome proportional to per-SNP heritabilities from step 4. The L2 regularization in

step 1 improves the accuracy of per-SNP heritability estimation; the partitioning into odd and even chromosomes in steps 1–2 and the exclusion of the target chromosome in step 4 prevents winner’s curse; and the re-estimation of per-SNP heritabilities in step 4 ensures robustness to model misspecification.

PolyFun specifies prior causal probabilities in proportion to per-SNP heritability estimates:

$$P(\beta_i \neq 0 | \mathbf{a}_i) \propto \text{var}[\beta_i | \mathbf{a}_i], \quad (1)$$

where β_i is the causal effect size of SNP i in standardized units (the number of standard deviations increase in phenotype per 1 standard deviation increase in genotype), \mathbf{a}_i is the vector of functional annotations of SNP i , and $\text{var}[\beta_i | \mathbf{a}_i]$ is the estimated per-SNP heritability of SNP i from step 4 (Methods).

A key distinction between PolyFun and previous functionally-informed fine-mapping methods^{10,18–20} is the use of the entire genome and a large number of functional annotations to estimate prior causal probabilities. We exploited the computational scalability of PolyFun (together with SuSiE²¹) to fine-map up to 2,763 overlapping 3Mb loci spanning the entire genome (Methods). We subsequently used our fine-mapping results to perform polygenic localization, identifying minimal sets of common SNPs causally explaining a given proportion of common SNP heritability. Details of the PolyFun method are provided in the Methods section; we have released open-source software implementing PolyFun in conjunction with SuSiE²¹ and FINEMAP²². In all main simulations and analyses of real traits, we applied PolyFun using summary LD information estimated directly from the target samples (both for running S-LDSC and for running SuSiE or FINEMAP), as previously recommended for fine-mapping methods^{12,28}.

Main simulations

We evaluated PolyFun via simulations using real genotypes from 337,491 unrelated UK Biobank British samples²⁷. We analyzed 10 3Mb loci on chromosome 1, each containing 1,468–27,784 imputed MAF 0.001 SNPs (including short indels; Supplementary Table 2). We estimated prior causal probabilities using 18,212,157 genome-wide imputed MAF 0.001 SNPs with INFO score 0.6. We simulated traits with heritability equal to 25% and genome-wide proportion of causal SNPs equal to 0.5%, with each target locus including 10 causal SNPs jointly explaining heritability of 0.05%. We specified prior causal probabilities using the baseline-LF model²⁵ with meta-analyzed functional enrichments from real data analyses (Supplementary Table 3). We generated summary statistics using $N=320K$ samples. Further details are provided in the Methods section.

We evaluated 10 fine-mapping methods (Methods, Table 1). We assessed calibration via the proportion of false positives among SNPs with posterior causal probability (posterior inclusion probability; PIP) above a given threshold (e.g. PIP>0.95), aggregating the results across all simulations; we refer to this quantity as the false discovery rate (FDR). For each PIP threshold, we estimated the FDR as one minus the PIP threshold, which is more conservative than an exact estimate (Figure 1a–b, Supplementary Note, Supplementary Table 4). No method except CAVIARBF2- and CAVIARBF2 had significantly inflated false

discovery rates, although fastPAINTOR and CAVIARBF1 had suggestive evidence of inflated false discovery rates. We assessed power via the proportion of true causal SNPs with PIP above a given threshold, aggregating the results across all simulations. PolyFun + FINEMAP was the most powerful method, identifying >5% more PIP>0.95 causal SNPs than PolyFun + SuSiE and >20% more PIP>0.95 causal SNPs than FINEMAP; PolyFun + SuSiE was the second most powerful method, identifying >25% more PIP>0.95 causal SNPs than SuSiE (Figure 1c–d, Supplementary Table 4). These results demonstrate the benefits of prioritizing SNPs using functional annotations.

We evaluated the computational cost of each method. SuSiE and PolyFun + SuSiE were much faster than the other methods, fine-mapping a 3Mb locus in 5 minutes on average (excluding fixed preprocessing time; see below) (Figure 1e, Supplementary Table 4). CAVIARBF methods allowing >2 causal SNPs per locus were not evaluated due to prohibitively slow computation time. PolyFun also requires fixed preprocessing time (steps 1–4; see Overview of methods) of 630 minutes on average; when restricting analyses to subsets of loci, PolyFun + SuSiE was still faster than all other functionally-informed methods when analyzing >23 loci (Figure 1f).

We performed additional experiments to assess the robustness of PolyFun to model misspecification of functional architectures, to assess the individual impact of each of steps 1–5 of PolyFun on fine-mapping performance, and to explore additional simulation settings (Supplementary Note, Extended Data Figures 1–5, Supplementary Tables 4–6).

We conclude from these experiments that PolyFun + FINEMAP and PolyFun + SuSiE outperformed all other methods, with a 3.4x faster runtime for PolyFun + SuSiE. Thus, we restricted our analyses in the remainder of this manuscript to SuSiE and PolyFun + SuSiE.

Simulations with mismatched reference LD

Our main simulations used in-sample LD computed directly from the target samples. Although we have publicly released summary LD information for British-ancestry UK Biobank samples as part of this study, there are many settings in which researchers conducting fine-mapping cannot obtain in-sample LD, and instead use LD information from an external LD reference panel²⁹. We performed extensive simulations to assess how fine-mapping performance is impacted by LD mismatch between the target sample and the LD reference panel. We specifically considered (1) non-overlapping target and reference samples; (2) sample sizes of the target sample and reference panel; (3) differences in ancestry; (4) presence of related individuals in the target sample; and (5) SNPs available for analysis in the target sample and reference panel.

We performed 19 experiments, described in detail in Table 3, in the Supplementary Note and in Supplementary Table 7. We quantified how mismatched reference LD impacts fine-mapping performance via the maximum number of assumed causal SNPs per locus (denoted as L) that maintains FDR<0.05 at a PIP=0.95 threshold. Based on these experiments we provide fine-mapping best-practice recommendations: (1) PolyFun + SuSiE should ideally use in-sample LD from the GWAS target sample, with $L=10$; (2) PolyFun + SuSiE can alternatively use a non-overlapping LD reference panel from the target population spanning

10% of the target sample size, with $L=10$; (3), PolyFun + SuSiE can be used without an LD reference panel by specifying $L=1$. We caution that using an LD reference panel with even subtle population differences with $L>1$ may lead to false positive results; (4) PolyFun + SuSiE can be used in the presence of related individuals in the target sample (but these results apply to the typical levels of relatedness observed in UK Biobank); and (5) PolyFun + SuSiE should include as many well-imputed SNPs from the target locus as possible to minimize the risk of omitting causal SNPs. The real-world implications of these best-practice recommendations are discussed in the Discussion.

Functionally informed fine-mapping of 49 complex traits

We applied PolyFun + SuSiE to fine-map 49 traits in the UK Biobank, including 33 traits analyzed in refs. ^{30,31}, 9 blood cell traits analyzed in ref. ¹², and 7 metabolic traits (average $N=318K$; Supplementary Table 8). For each trait we fine-mapped up to 2,763 overlapping 3Mb loci spanning $M=18,212,157$ imputed MAF 0.001 SNPs with INFO score 0.6 (including short indels; excluding three long-range LD regions and loci with close to zero heritability; Methods). We assigned to each SNP its PIP computed using the locus in which it was most central. We have publicly released the PIPs and the prior and posterior means and variances of the causal effect sizes for all SNPs and traits analyzed.

PolyFun + SuSiE identified 3,025 PIP>0.95 fine-mapped SNP-trait pairs, a >32% improvement vs. SuSiE; 9,684 PIP>0.5 SNP-trait pairs, a >59% improvement vs. SuSiE; and 225,153 PIP>0.05 SNP-trait pairs, a >84% improvement vs. SuSiE (Supplementary Table 9). The number of PIP>0.95 SNPs per trait ranged from 0 (number of children) to 407 (height) (Figure 2a, Supplementary Table 9). The 3,025 PIP>0.95 SNP-trait pairs spanned 2,225 unique SNPs, including 532 low-frequency SNPs ($0.005<MAF<0.05$) and 185 rare SNPs ($0.001<MAF<0.005$) (Supplementary Table 10). Only 39% of the 2,225 PIP>0.95 SNPs were also lead GWAS SNPs (defined as MAF>0.001 SNPs with $P<5\times 10^{-8}$ and no MAF>0.001 SNP with a smaller p-value within 1Mb) (Supplementary Table 10), demonstrating the importance of using fine-mapped SNPs rather than lead GWAS SNPs for downstream analysis. 31% of the PIP>0.95 SNPs resided in coding regions and 22% were non-synonymous (broadly consistent with previous fine-mapping studies^{8,12}) (Supplementary Table 10). When restricting the analysis to 16 genetically uncorrelated traits ($|r_g|<0.2$; Methods and Supplementary Tables 11–12) we identified 1,626 PIP>0.95 SNP-trait pairs spanning 1,496 unique SNPs, with a median distance of 9kb between a PIP>0.95 SNP and the nearest lead GWAS SNP for the same trait (Supplementary Table 10). The 16 genetically uncorrelated traits included 5,314 genome-wide significant locus-trait pairs (defined by 1Mb windows around lead GWAS SNPs) harboring 0.28 PIP>0.95 SNPs per locus on average (Supplementary Table 13); 1,080 of the 5,314 locus-trait pairs (20%) harbored 1 PIP>0.95 SNP(s), harboring 1.37 PIP>0.95 SNPs on average (Supplementary Table 13). 150 of the 1,626 SNP-trait pairs identified by PolyFun + SuSiE PIP>0.95 (9.2%) did not lie within genome-wide significant loci, and 161 of the 1,626 SNP-trait pairs (9.9%) had $P>5\times 10^{-8}$ (Supplementary Table 10).

We estimated the SNP-heritability (h_g^2) tagged by PIP>0.95 fine-mapped SNPs (which is likely to be close to the heritability causally explained by these SNPs, if most of the tagged

SNP-heritability originates from PIP>0.95 SNPs). The h_g^2 tagged by PIP>0.95 SNPs captured a large proportion of the h_g^2 tagged by lead GWAS SNPs (median proportion=42%; Figure 2b, Methods, Supplementary Table 14). This proportion was substantially larger than the proportion of GWAS loci harboring PIP>0.95 SNPs (20%; see above), as fine-mapping power is higher at loci with larger causal effects (Supplementary Table 4). However, fine-mapped SNPs tagged a smaller proportion of total MAF>0.001 h_g^2 (median proportion=19%; Figure 2b, Methods, Supplementary Table 14), indicating that substantially larger sample sizes are required to comprehensively fine-map all heritable SNP effects.

Among the 2,225 unique PIP>0.95 SNPs fine-mapped for at least one trait, 223 SNPs were fine-mapped for multiple genetically uncorrelated traits (selecting a different subset of genetically uncorrelated traits for each SNP; Methods), including 55 SNPs fine-mapped for 3 genetically uncorrelated traits, indicating pervasive pleiotropy (Extended Data Figure 6, Supplementary Table 15). 118 pleiotropic SNPs resided in coding regions and 93 were non-synonymous (Supplementary Table 15). The 17 SNPs fine-mapped for at least 4 traits are reported in Table 2. Previous studies have reported that genetically uncorrelated traits often share association signals at the same loci³², but did not fine-map those signals to individual SNPs as performed here.

To better understand the improvement of PolyFun + SuSiE over SuSiE, we examined the 121 loci where PolyFun + SuSiE identified a fine-mapped common SNP (PIP>0.95) but SuSiE did not (PIP<0.5 for all SNPs within 1Mb) (Figure 3 and Supplementary Table 16). In each case, functional annotations prioritized one SNP out of several candidates, greatly improving fine-mapping resolution.

We validated the motivation for performing functionally-informed fine-mapping by verifying that fine-mapped SNPs are enriched for functional annotations, as previously shown for autoimmune diseases^{7,8,10} and blood traits¹² (using non-functionally-informed SuSiE to avoid biasing the results). For each of 50 main binary annotations from the baseline-LF model²⁴, for various PIP ranges, we computed the functional enrichment of fine-mapped common SNPs in the PIP range, defined as the proportion of common SNPs in the PIP range lying in the annotation divided by the proportion of genome-wide common SNPs lying in the annotation, and meta-analyzed the results across genetically uncorrelated traits (Methods, Figure 4, Supplementary Table 17). PIP>0.95 SNPs were strongly and significantly enriched for non-synonymous SNPs (51x enrichment, $P=6.8\times 10^{-185}$) and SNPs in conserved regions (16x enrichment, $P<10^{-300}$), significantly enriched for SNPs in various regulatory annotations (e.g. promoter-ExAC and H3K4me3), and significantly depleted for SNPs in repressed regions, consistent with previous literature on functional enrichment of fine-mapped SNPs^{7,8,10-12} and disease heritability^{17,24,25,33}. We observed qualitatively similar but weaker enrichments at lower PIP ranges (Figure 4, Supplementary Table 17).

We compared our fine-mapping results to two previous studies. First, we compared our results to ref. ¹², which performed non-functionally informed fine-mapping for 9 blood cell traits using approximately 115K of the individuals included in our analyses. PolyFun + SuSiE identified 4.4× more SNPs than ref. ¹², including all four SNPs that were functionally

validated via luciferase reporter assays in ref. ¹² (PIP>0.999 for all four SNPs; Methods, Supplementary Table 18–20). Second, we compared our results to ref. ⁷, which performed non-functionally-informed fine-mapping for 7 of our traits, using a non-functionally informed method (PICS) and independent smaller data sets. PolyFun + SuSiE identified 35x more SNPs than ref. ⁷; Supplementary Tables 21–22). Further details of the comparison are provided in the Supplementary Note.

We further performed 6 secondary analyses, described in the Supplementary Note, in Extended Data Figures 7–9, and in Supplementary Tables 10 and 23–28.

In summary, we leveraged the improved power of PolyFun + SuSiE to robustly identify thousands of fine-mapped SNPs, providing a rich set of potential candidates for functional follow-up. Our results further indicate pervasive pleiotropy, with many SNPs fine-mapped for two or more genetically uncorrelated traits.

Polygenic localization of 49 complex traits

PIP>0.95 SNPs tag a large proportion of the SNP-heritability (h_g^2) tagged by lead GWAS SNPs (median proportion=42%) but a small proportion of total genome-wide h_g^2 (median proportion=19%) (Figure 2b), implying that they causally explain a small proportion of h_g^2 . We thus propose *polygenic localization*, whose aim is to identify a minimal set of common SNPs causally explaining a specified proportion of common SNP heritability. A key difference between polygenic localization and previous studies of polygenicity^{34–38} is that polygenic localization aims to *identify* (not just characterize) such SNPs.

Given a ranking of SNPs by posterior per-SNP heritability (i.e., the posterior mean of their squared effect size; see Methods), we define $M_{50\%}$ as the size of the smallest set of top-ranked common SNPs causally explaining 50% of common SNP heritability (resp. M_p for proportion p of common SNP heritability). We estimate $M_{50\%}$ (resp. M_p) by (1) partitioning SNPs into 50 ranked bins of similar posterior per-SNP heritability estimates from PolyFun + SuSiE and stratifying the lowest-heritability bin into 10 equally-sized MAF bins, yielding 59 bins; (2) running S-LDSC using a different set of samples to re-estimate the average per-SNP heritability in each bin; and (3) computing the number of top-ranked common SNPs (with respect to the original ranking) whose estimated per-SNP heritabilities (from step 2) sum up to 50% (resp. the proportion p) of the total estimated SNP-heritability. We refer to this method as PolyLoc. The analysis of new samples in step 2 of PolyLoc prevents winner's curse; although PolyFun + SuSiE is robust to winner's curse, PolyLoc would be susceptible to winner's curse if it reused the data analyzed by PolyFun + SuSiE. We note that $M_{50\%}$ relies on an empirical ranking and is thus larger than the size of the smallest set of SNPs causally explaining 50% of common SNP heritability, denoted as $M_{50\%}^*$ ($M_{50\%} \geq M_{50\%}^*$). We performed extensive simulations to confirm that PolyLoc produced robust upper bounds of $M_{50\%}^*$ (Supplementary Note, Supplementary Tables 29–30). Further details of PolyLoc are provided in the Methods section; we have released open source software implementing PolyLoc.

We applied PolyLoc to the 49 complex traits from the UK Biobank (Supplementary Table 8). We ranked SNPs using $N=337\text{K}$ unrelated British ancestry samples (steps 1–2) and re-estimated average per-SNP heritabilities in each of 59 SNP bins using S-LDSC applied to $N=122\text{K}$ European-ancestry UK Biobank samples that were not included in the $N=337\text{K}$ set to avoid winner's curse (step 3). Estimates of $M_{50\%}$ ranged widely from 28 (hair color) to 3.4K (height) to 2 million (number of children) (Figure 5, Supplementary Table 31). The median estimate of $M_{50\%}$ across 16 genetically uncorrelated traits was 8.9K; the median estimate of $M_{5\%}$ was 8; and the median estimate of $M_{95\%}$ was 4.4 million (of 7.0 million total common SNPs) (Supplementary Table 31). Pigmentation traits were the least polygenic traits while number of children was the most polygenic trait, having $M_{50\%}$ 3.7x larger than the second most polygenic of the 16 independent traits (chronotype, having $M_{50\%}=553\text{K}$), consistent with ref. ³⁴. We performed 7 secondary analyses, described in the Supplementary Note and in Supplementary Tables 32–33. We note that far fewer than 2 million SNPs may causally explain 50% of the common SNP heritability of number of children, because $M_{50\%}$ is a (possibly loose) upper bound.

Our results demonstrate that half of the common SNP heritability of complex traits is causally explained by typically thousands of SNPs (median $M_{50\%}=8.9\text{K}$), and the remaining heritability is spread across an extremely large number of extremely weak-effect SNPs (median $M_{95\%}=4.4$ million), consistent with extremely polygenic but heavy-tailed trait architectures^{1,34–36,39–43}.

Discussion

We have introduced PolyFun, a framework that improves fine-mapping by prioritizing variants that are a-priori more likely to be causal based on their functional annotations. Across 49 UK Biobank traits, PolyFun + SuSiE confidently fine-mapped 3,025 SNP-trait pairs (PIP > 0.95), a 32% increase over non-functionally informed SuSiE. 223 of the fine-mapped SNPs were fine-mapped for multiple genetically uncorrelated traits, indicating pervasive pleiotropy. We further leveraged the results of PolyFun to perform polygenic localization by constructing minimal SNP sets causally explaining a given proportion of common SNP heritability, demonstrating that 50% of common SNP heritability can be explained by sets ranging in size from 28 (hair color) to 3,400 (height) to 2 million (number of children). We note that these set sizes impose a (possibly loose) upper bound on the size of the smallest sets causally explaining 50% of common SNP heritability. We have publicly released the PIPs and the prior and posterior means and variances of effect sizes for all SNPs and traits analyzed.

We recommend applying PolyFun using in-sample LD from the GWAS target sample (i.e., using exactly the same samples in both the target and reference samples), assuming 10 causal SNPs per locus; we have facilitated this option for UK Biobank researchers by publicly releasing summary LD information for $N=337\text{K}$ British-ancestry UK Biobank samples. As a second-best option we recommend applying PolyFun using LD-reference panel from the target sample population spanning at least 10% of the target sample size, while assuming 10 causal SNPs per locus. However, we caution that even subtle population differences may lead to false positive results. Hence, our published summary LD

information files are unsuitable for analysis of summary statistics involving non-British UK Biobank individuals, or data from other cohorts or consortia^{44–46}. However, researchers may use larger subsets of UK Biobank data to identify genome-wide significant loci, which they can fine-map using summary statistics and LD reference data based on $N=337K$ British-ancestry individuals. In the absence of a reference panel from the target sample population spanning $>10\%$ of the target sample size, we recommend applying PolyFun without using an LD reference panel by restricting it to assume a single causal SNP per locus.

Our fine-mapping analysis differs from several previous fine-mapping studies in two aspects. First, we applied PolyFun genome-wide. However, we envision that the PolyFun software will primarily be used to fine-map genome-wide significant loci, which harbor most $PIP>0.95$ SNPs. We discuss possible reasons for identifying $PIP>0.95$ SNPs with $P>5\times 10^{-8}$ in the Supplementary Note. Second, PolyFun fine-maps all signals in a locus jointly to maximize power^{5,28}. Researchers wishing to use PolyFun for a partitioned analysis⁴⁷ may still do so by first partitioning a locus into multiple signals using a separate tool (e.g. GCTA-COJO⁴⁷) and then applying PolyFun to each signal separately, restricting PolyFun to assume a single causal SNP per signal.

Our results provide several opportunities for future work. First, the fine-mapped SNPs that we have identified can be prioritized for functional follow-up. Second, fine-mapping results (posterior mean effect sizes) can be used to compute trans-ethnic polygenic risk scores⁴⁸ which may be less sensitive to LD differences between populations than existing methods^{49,50}. Third, the proximal pairs of coding and non-coding fine-mapped SNPs that we identified (Supplementary Table 25) may aid efforts to link SNPs to genes^{51–53}. Fourth, SNPs that were fine-mapped for multiple genetically uncorrelated traits may shed light on shared biological pathways⁵⁴. Fifth, sets of SNPs causally explaining 50% of common SNP heritability can potentially be used for gene and pathway enrichment analysis^{55,56}. Finally, PolyFun can incorporate additional functional annotations at negligible additional computational cost, motivating further efforts to identify conditionally informative annotations.

Our work has several limitations. First, our $PIP>0.95$ FDR estimates for PolyFun and for other methods are conservative, demonstrating the challenges of exact calibration in fine-mapping. Second, subtle population stratification may lead to spurious fine-mapping results⁵⁷. However, our fine-mapped SNPs are concentrated in associated loci with larger estimated effects, which are relatively less likely to be spurious. Third, we restricted fine-mapping to $N=337K$ unrelated British-ancestry individuals, consistent with previous studies¹². Hence, our published summary LD information files do not support fine-mapping of UK Biobank data that includes non-British individuals. Fourth, PolyLoc requires analyzing samples distinct from the samples analyzed by PolyFun to avoid winner's curse. Researchers with access to individual-level genetic data can partition the samples as we have done (we recommend using approximately 75% of the data for fine-mapping and 25% for polygenic localization). Fifth, PolyFun does not support X-chromosome analysis. Sixth, PolyLoc only provides an upper bound on the proportion of SNPs causally explaining a given proportion of SNP-heritability. Finally, multi-ethnic fine-mapping⁵⁸ and incorporation

of tissue-specific functional annotations^{9,13,15,17} may further increase fine-mapping power. Incorporating these into our fine-mapping framework is an avenue for future work.

Online Methods

PolyFun fine-mapping method

PolyFun first estimates prior causal probabilities for all SNPs and then applies fine-mapping methods such as SuSiE²¹ or FINEMAP^{22,23} with these prior causal probabilities. Here we describe estimation of the prior causal probabilities.

We model standardized phenotypes y using the linear model $y = \sum_i x_i \beta_i + \epsilon$, where x_i denotes standardized SNP genotypes, β_i denotes effect size, and ϵ is a residual term. We use a point-normal model for β_i :

$$\beta_i | \mathbf{a}_i \sim \begin{cases} \mathcal{N}(0, \text{var}[\beta_i | \beta_i \neq 0]) & \text{with probability } P(\beta_i \neq 0 | \mathbf{a}_i) \\ 0 & \text{otherwise} \end{cases},$$

where \mathbf{a}_i are the functional annotations of SNP i , $P(\beta_i \neq 0 | \mathbf{a}_i)$ is its prior causal probability, and $\text{var}[\beta_i | \beta_i \neq 0]$ is its causal variance, which we assume is independent of \mathbf{a}_i . This assumption is motivated by our recent work showing that functional enrichment is primarily due to differences in polygenicity rather than differences in effect-size magnitude, which is constrained by negative selection³⁴.

The key quantity that PolyFun uses to estimate prior causal probabilities is the per-SNP heritability of SNP i , $\text{var}[\beta_i | \mathbf{a}_i]$ (we refer to this quantity as per-SNP heritability because the total SNP-heritability $\text{var}[\sum_i x_i \beta_i | \mathbf{a}]$ is equal to $\sum_i \text{var}[\beta_i | \mathbf{a}_i]$, assuming that causal SNP effects have zero mean and are uncorrelated with other SNP effects and with other SNPs conditional on \mathbf{a}). PolyFun relates the prior causal probability $P(\beta_i \neq 0 | \mathbf{a}_i)$ to the per-SNP heritability $\text{var}[\beta_i | \mathbf{a}_i]$ via the law of total variance:

$$P(\beta_i \neq 0 | \mathbf{a}_i) = \frac{\text{var}[\beta_i | \mathbf{a}_i]}{\text{var}[\beta_i | \beta_i \neq 0]}. \quad (2)$$

Equation 1 in the main text follows because $P(\beta_i \neq 0 | \mathbf{a}_i)$ is proportional to $\text{var}[\beta_i | \mathbf{a}_i]$ with the proportionality factor $1/\text{var}[\beta_i | \beta_i \neq 0]$.

To derive Equation 2 we define the causality indicator $c_i = \mathbb{I}[\beta_i \neq 0 | \mathbf{a}_i]$ and apply the law of total variance to $\text{var}[\beta_i | \mathbf{a}_i]$:

$$\begin{aligned} \text{var}[\beta_i | \mathbf{a}_i] &= E_{c_i}[\text{var}[\beta_i | \mathbf{a}_i, c_i]] + \text{var}_{c_i}[E[\beta_i | \mathbf{a}_i, c_i]] \\ &= E_{c_i}[\text{var}[\beta_i | \mathbf{a}_i, c_i]] + \text{var}_{c_i}[0] \\ &= P(c_i = 0) \cdot 0 + P(c_i = 1) \text{var}[\beta_i | \mathbf{a}_i, c_i = 1] \\ &= P(\beta_i \neq 0 | \mathbf{a}_i) \text{var}[\beta_i | \mathbf{a}_i, \beta_i \neq 0] \\ &= P(\beta_i \neq 0 | \mathbf{a}_i) \text{var}[\beta_i | \beta_i \neq 0]. \end{aligned}$$

The last equality holds because we assume that the causal effect size variance is independent of functional annotations, as explained above.

PolyFun avoids directly estimating the proportionality factor $1/\text{var}[\beta_j|\beta_j = 0]$ by constraining the prior causal probabilities $P(\beta_j = 0|a_j)$ in each tested locus to sum to 1.0. This constraint implies that each locus is a-priori expected to harbor one causal SNP, consistent with previous fine-mapping methods^{5,6,22} (this constraint is ignored by PolyFun + SuSiE because it is invariant to scaling of prior causal probabilities). Hence, the main challenge is estimating the per-SNP heritabilities $\text{var}[\beta_j|a_j]$.

To estimate $\text{var}[\beta_j|a_j]$, PolyFun incorporates a regularized extension of S-LDSC with the baseline-LF model^{17,24–26}, which we extend to a new version 2.2.UKB (Supplementary Table 1, see below). S-LDSC uses the linear model $\text{var}[\beta_i|a_i] = \sum_c \tau^c a_i^c$ and jointly estimates all τ^c parameters by minimizing the term $\sum_i (\chi_i^2 - n \sum_c \tau^c l(i, c) - nb - 1)^2$, where c are functional annotations, τ^c is the coefficient of annotation c , χ_i^2 is the χ^2 statistic of SNP i , n is the sample size, b measures the contribution of confounding biases, and $l(i, c) = \sum_j r_{ij}^2 a_j^c$.

While S-LDSC produces robust estimates of functional enrichment, it has two limitations in estimating $\text{var}[\beta_j|a_j]$: (i) these estimates can have large standard errors in the presence of many annotations, and (ii) the model may not be robust to model misspecification. To address the first limitation, PolyFun incorporates an L2-regularized extension of S-LDSC. To address the second limitation, PolyFun employs special procedures to ensure robustness to model misspecification. The key idea is to approximate arbitrary complex functional forms of $\text{var}[\beta_j|a_j]$ via a piecewise-constant function. To do this, PolyFun partitions SNPs with similar estimated values of $\text{var}[\beta_j|a_j]$ (estimated via a possibly misspecified model) into non-overlapping bins; estimates the SNP-heritability causally explained by each bin b ; and specifies $\text{var}[\beta_j|a_j]$ for SNPs in bin b as the SNP-heritability causally explained by bin b divided by the number of SNPs in bin b . PolyFun avoids winner's curse by using different data for partitioning SNPs and for per-bin heritability estimation.

In detail, PolyFun robustly specifies prior causal probabilities for all SNPs on a target locus on a corresponding odd (resp. even) target chromosome via the following procedure:

1. Estimate annotation coefficients $\hat{\tau}_{\text{even}}^c$ and intercepts \hat{b}_{even} using only SNPs in even chromosomes via an L2-regularized extension of S-LDSC that minimizes $\sum_i (\chi_i^2 - n \sum_c \hat{\tau}_{\text{even}}^c l(i, c) - n\hat{b}_{\text{even}} - 1)^2 + \lambda \sum_c (\hat{\tau}_{\text{even}}^c)^2$ (resp. using $\hat{\tau}_{\text{odd}}^c$ and \hat{b}_{odd}). We select the regularization strength λ from a geometrically-spaced grid of 100 values ranging from 10^{-8} to 100, selecting the one that minimizes the average out-of-chromosome error $\sum_r \sum_{i \in r} (\chi_i^2 - n \sum_c \hat{\tau}_{\text{even}\setminus r}^c l(i, c) - n\hat{b}_{\text{even}\setminus r} - 1)^2$, where r iterates over even (resp. odd) chromosomes, and $\hat{\tau}_{\text{even}\setminus r}^c$, $\hat{b}_{\text{even}\setminus r}$ are the S-LDSC τ and b estimates, respectively, when applied to all SNPs on even chromosomes except for chromosome r (resp. for odd chromosomes).

2. Compute per-SNP heritabilities $\widehat{\text{var}}[\beta_i | \mathbf{a}_i] = \sum_c \hat{\tau}_{\text{even}}^c a_i^c$ for each SNP i in an odd chromosome (resp. $\sum_c \hat{\tau}_{\text{odd}}^c a_i^c$).
3. Partition all SNPs into 20 bins with similar values of $\widehat{\text{var}}[\beta_i | \mathbf{a}_i]$ using the Ckmedian.1d.dp method⁵⁹. This method partitions SNPs into 20 maximally homogenous bins such that the average distance of $\widehat{\text{var}}[\beta_i | \mathbf{a}_i]$ to the median $\widehat{\text{var}}[\beta_i | \mathbf{a}_i]$ of the bin of SNP i is minimized. Even though this step uses functional annotations data of the target chromosome it does not use the summary statistics of SNPs in the target chromosome, which ensures robustness to winner's curse.
4. Apply S-LDSC with non-negativity constraints to estimate per-SNP heritabilities in each of the 20 bins of all SNPs in odd (resp. even) chromosomes except for the target chromosome r (to avoid using the same data that will be used in fine-mapping), denoted $\hat{\sigma}_{r,1}^2, \dots, \hat{\sigma}_{r,20}^2$. Afterwards, regularize the estimates by setting all values smaller than $q \cdot \max_b(\hat{\sigma}_{r,b}^2)$ to $q \cdot \max_b(\hat{\sigma}_{r,b}^2)$, using $q = 1/100$ by default, and rescaling the $\hat{\sigma}_{r,b}^2$ estimates to have the same sum (over all genome-wide SNPs) as before. The regularization prevents SNPs from having a zero per-SNP heritability, which would exclude them from fine-mapping. We did not apply L2-regularization in this step because we require approximately unbiased estimates, and because standard errors are relatively small under a small number of non-overlapping annotations.
5. Specify a prior causal probability proportional to $\hat{\sigma}_{r,b}^2$ to each SNP that is in bin b and that resides in a target locus in chromosome r , such that the prior causal probabilities in the target locus sum to one.

PolyFun uses version 2.2.UKB of the baseline-LF model, which differs from the original baseline-LF model²⁵ by including MAF 0.001 SNPs and several new annotations, and omitting annotations that could not be easily extended to account for MAF < 0.005 SNPs (Supplementary Table 1). Briefly, we use 187 overlapping functional annotations, including 10 common MAF bins (MAF 0.05); 10 low-frequency MAF bins (0.05 > MAF 0.001); 6 LD-related annotations for common SNPs (levels of LD, predicted allele age, recombination rate, nucleotide diversity, background selection statistic, CpG content); 5 LD-related annotations for low-frequency SNPs; 40 binary functional annotations for common SNPs; 7 continuous functional annotations for common SNPs; 40 binary functional annotations for low-frequency SNPs; 3 continuous functional annotations for low-frequency SNPs; and 66 annotations constructed via windows around other annotations¹⁷. We did not include a base annotation that includes all SNPs, because such an annotation is linearly dependent on all the MAF bins when S-LDSC uses the same set of SNPs to compute LD-scores and to estimate annotation coefficients.

Main fine-mapping simulations

We simulated summary statistics for 18,212,157 genotyped and imputed MAF 0.001 autosomal SNPs with INFO score 0.6 (including short indels, excluding three long-range LD regions; see below), using $N=337,491$ unrelated British-ancestry individuals from UK

Biobank release 3. In most simulations we computed an effect variance β_i for every SNP i with annotations \mathbf{a}_i using the baseline-LF (version 2.2.UKB) model, $\text{var}[\beta_i|\mathbf{a}_i] = \sum_c \tau^c a_i^c$, where c are annotations and τ^c estimates are taken from a fixed-effects meta-analysis of 16 well-powered genetically uncorrelated ($|r_g| < 0.2$) UK Biobank traits, scaled such that $\sum_i \text{var}[\beta_i|\mathbf{a}_i]$ is the same across all traits (Supplementary Table 3). In some simulations we generated values of $\text{var}[\beta_i|\mathbf{a}_i]$ under alternative functional architectures to evaluate the robustness of PolyFun to modeling misspecification (Supplementary Note). Each SNP was set to be causal with probability proportional to $\text{var}[\beta_i|\mathbf{a}_i]$, such that the average causal probability was equal to the desired proportion of causal SNPs. We provide technical details about the simulations in the Supplementary Note.

We performed fine-mapping in each of the 10 selected 3Mb loci on chromosome 1 using methods based on SuSiE²¹, FINEMAP^{22,23}, CAVIARBF²⁰ and fastPAINTOR¹⁹. Following previous literature^{12,28} all methods used in-sample LD (i.e., summary LD information based on the genotypes of the same 337,491 individuals used to generate summary statistics), computed via LDstore²⁸. For fastPAINTOR-, fastPAINTOR, SuSiE, and PolyFun + SuSiE, we specified a causal effect size variance using an estimator that we developed based on a modified version of HESS⁶⁰ rather than using the estimator implemented in these methods, because it improved false discovery rate and power in most simulation settings (Supplementary Note, Supplementary Table 4).

We ran SuSiE 0.7.1.0487 with default values for all parameters except the following: (1) We used 10 causal SNPs per locus; and (2) we estimated a per-locus causal effect size variance (the scaled_prior_variance parameter) via our modified HESS approach. We specified prior causal probabilities via the prior_weights parameter. We modified the SuSiE source code to avoid performing the LD matrix diagnostics (positive-definiteness and symmetry) because they greatly increased memory consumption.

We ran FINEMAP 1.3.1.b with a maximum of 10 causal SNPs per locus and with default settings for all other parameters. We specified prior causal probabilities via the -prior-snps argument.

We ran CAVIARBF 0.2.1 with an AIC-based parameter selection, using ridge regression with regularization parameter λ selected from $\{2^{-10}, 2^{-5}, 2^{-2.5}, 2^0, 2^{2.5}, 2^5, 100, 1000, 10000, 100000\}$, with a single locus and with up to either 1 or 2 causal SNPs per locus, owing to computational limitations.

We ran fastPAINTOR 3.1 in MCMC mode. We specified a per-locus causal effect size variance (specified via the -variance argument) using our modified HESS approach (as in PolyFun + SuSiE). We avoided truncating the LD matrix (using prop_ld_eigenvalues=1.0) because we used in-sample summary LD information. As fastPAINTOR is generally not designed to work with >10 annotations^{18,19} (and was too slow in our simulations to estimate the significance of each annotation and include only conditionally significant annotations as done in ref. ¹⁸), we selected a subset of 10 highly informative annotations by (1) scoring each annotation based on its average contribution to effect variance $|a_i^c \tau^c|$ across all SNPs, using the true τ^c of the generative model; (2) iteratively selecting top-ranked annotations

such that no annotation has correlation >0.3 (in absolute value) with a previously selected annotation, until selecting 10 annotations. We determined that 10 annotations yielded approximately optimal power while maintaining correct calibration (Supplementary Table 4).

For each PIP threshold, we conservatively estimated false discovery rates by setting all PIPs greater than the threshold to the threshold, yielding a uniform false-discovery threshold (Supplementary Note, Supplementary Table 4).

We computed p-values of FDR differences and of power differences of analyses with perturbed PolyFun steps via a Wald test, using a jackknife over simulated datasets to estimate standard errors (Supplementary Note).

Simulations with mismatched reference LD

Our mismatched reference LD simulations differed from our main simulations in several ways: (i) we generated summary statistics using up to $N=44K$ unrelated (or related) European-ancestry (British or non-British) UK Biobank target samples in most experiments, compared with $N=320K$ in our main simulations, because the UK Biobank includes only 44K unrelated UK Biobank individuals of non-British European ancestry (we used $N=293K$ unrelated British-ancestry UK Biobank target samples in a subset of experiments to more closely match our main simulations); (ii) we computed summary LD information using either $N=400$, $N=4,000$, or $N=44K$ unrelated British-ancestry UK Biobank reference samples (either non-overlapping or overlapping with the target samples), or using $N=3,567$ reference samples from the UK10K cohort⁶¹ (compared with in-sample LD based on the target samples in the main simulations); (iii) we generated summary statistics using individual level genotypes rather than summary LD information (as required when the target sample and the LD reference panel are not the same); (iv) we simulated 3 causal SNPs per locus that jointly explain 0.5% of trait variance, compared with 10 causal SNPs that jointly explain 0.05% of trait variance in our main simulations, to obtain sufficient power despite having a smaller sample size; and (v) in some experiments we used a subset of SNPs for generating causal SNPs or for fine-mapping analysis. We provide technical details of these simulations in the Supplementary Note.

Functionally informed fine-mapping of 49 complex traits in the UK Biobank

We applied SuSiE and PolyFun + SuSiE to fine-map 49 traits in the UK Biobank, using the same data and the same parameter settings described in the Fine-mapping simulations section. We performed basic QC on each trait as described in our previous publications^{30,31}. Specifically, we removed outliers outside the reasonable range for each quantitative trait, and applied quantile normalizing within sex strata after correcting for covariates for non-binary traits with non-normal distributions. We computed summary statistics with BOLT-LMM v2.3.3³¹ adjusting for sex, age and age squared, assessment center, genotyping platform, and the top 20 principal components (computed as described in ref. ³¹), and dilution factor for biochemical traits. As the non-infinitesimal version of BOLT-LMM does not estimate effect sizes, we computed z-scores for fine-mapping by taking the square root of the BOLT-LMM

χ^2 statistics and multiplying them by the sign of the effect estimate from the infinitesimal version of BOLT-LMM.

We partitioned all autosomal chromosomes into 2,763 overlapping 3Mb-long loci with a 1Mb spacing between the start points of consecutive loci. We computed a PIP for each SNP based on the locus whose center was closest to the SNP (excluding SNPs >1Mb away from the closest center and loci wherein all SNPs had squared marginal effect sizes smaller than 0.00005). We excluded the MHC region (chr6 25.5M-33.5M) and two other long-range LD regions (chr8 8M-12M, chr11 46M-57M)⁶² from all analyses, following our observations that both FINEMAP and SuSiE tend to produce spurious results in these regions, finding many PIP=1 SNPs across many traits regardless of their BOLT-LMM p-values. We verified that other previously reported long-range LD regions⁶² do not harbor a disproportionate number of PIP>0.95 SNPs. We specified per-locus causal effect variances for SuSiE and PolyFun + SuSiE via our modified HESS approach. For all S-LDSC and fine-mapping analyses we specified a sample size corresponding to the BOLT-LMM effective sample size³¹ (given by the true sample size multiplied by the median ratio between χ^2 statistics of BOLT-LMM and linear regression across SNPs having BOLT-LMM $\chi^2 > 30$).

All S-LDSC analyses used LD scores computed from in-sample summary LD information (based on imputed SNP dosages rather than sequenced genotypes as in previous publications²⁴⁻²⁶, assigning to each SNP the LD score computed in the locus in which it was most central) because they provide better coverage of low-frequency SNPs and are consistent with the fine-mapping analyses. We computed genetic correlations with LDSC, using the same summary statistics used for fine-mapping and restricting the analysis to common SNPs.

We selected a subset of 16 genetically uncorrelated traits by ranking all traits according to the number of PolyFun + SuSiE PIP>0.95 SNPs and greedily selecting top-ranked traits such that no selected trait has $|r_g| > 0.2$ with a previously selected trait, excluding traits having either (1) h_g^2 estimates <0.05 in either the PolyFun dataset ($N=337K$) or in the PolyLoc dataset ($N=122K$) (see h_g^2 estimation description below); or (2) traits with an effective sample size <100K in the $N=337K$ dataset (using $4/(1/\#cases + 1/\#controls)$ for binary traits).

We estimated h_g^2 tagged by PIP>0.95 SNPs and by lead GWAS SNPs via a multivariate linear regression. We regressed all the covariates used in BOLT-LMM out of the phenotypes, performed multivariate linear regression on the residuals (using all PIP>0.95 SNPs as explanatory variables) and reported the adjusted R^2 as the h_g^2 tagged by these SNPs. We verified that the results remained nearly identical regardless of whether we excluded related individuals (Supplementary Table 14). We estimated MAF>0.001 SNP-heritability for trait selection and for Figure 2b by running S-LDSC with all the baseline-LF annotations. We overrode the automatic removal of very large effect SNPs employed by S-LDSC for hair color, because this removal led to h_g^2 estimates that were smaller than the linear regression-based estimates, due to the large proportion of SNP-heritability originating from very large-effect SNPs.

We defined top annotations for Table 2, Figure 3, and Supplementary Tables 15–16 by first ranking all annotations according to their functional enrichment among PIP>0.95 SNPs (as in Figure 4; see below), and associating each SNP with its top ranked annotation, using meta-analyzed enrichment.

We selected a subset of genetically uncorrelated traits for each SNP (used in Extended Data Figure 6, Table 2, and Supplementary Table 15), aiming to select traits from a diverse set of groups as possible (anthropometric, lipids/metabolic, blood, cardiovascular/metabolic disease, other; Extended Data Figure 6, Supplementary Table 8). To this aim, we iterated over trait groups cyclically. For each group containing 1 unselected traits with PIP>0.95 for the analyzed SNP, we selected the trait having the smallest average $|r_g|$ with unselected traits from other groups (if there remained any) or from all remaining traits (otherwise), selecting among all traits having $|r_g|<0.2$ with previously selected traits, until no more eligible traits remained. We plotted the ideogram in Extended Data Figure 6 with the PhenoGram⁶³ software.

We computed enrichment of functional annotations among fine-mapped SNPs (Figure 4) as the ratio between the proportion of common SNPs with PIP above a given threshold having a specific annotation and the proportion of common SNPs having the annotation. We excluded continuous annotations and annotations constructed via windows around other annotations, and merged concordant annotations for common and low-frequency variants. We computed P-values using Fisher's exact test (meta-analyzed across traits via Fisher's method). We computed standard errors by (1) computing the standard error s of the log of the enrichment via the standard formula for the standard error of relative risk (exploiting the fact that enrichment and relative risk are both ratios of proportions); and (2) computing the standard error of the enrichment via $\sqrt{r^2(e^{s^2} - 1)}$ (i.e., the standard deviation of the exponent of a normal random variable), where r is the original enrichment estimate (meta-analyzed across traits using a fixed-effects meta-analysis). We excluded traits having <10 PIP>0.95 SNPs from the meta-analysis. The annotations shown in Figure 4 are non-synonymous, Conserved_LindbladToh (denoted Conserved), Human_Promoter_Villar_ExAC (denoted Promoter-ExAC), H3K4me3_Trynka (denoted H3K4me3), and Repressed_Hoffman (denoted Repressed) (see Supplementary Table 1 for details).

To compare our fine-mapping results with those of refs.^{7,12}, we restricted the comparison to SNPs that were not excluded from our fine-mapping procedure (SNPs having MAF 0.001 in the UK Biobank $N=337K$ dataset, INFO score 0.6, distance <1Mb away from the closest locus center, and not residing in one of the excluded long-range LD regions). When the same SNP had multiple reported PIPs in ref.¹², we used the entry with the larger PIP. We caution that the comparison with ref.¹² is not a replication analysis because the datasets of ref.¹² and of PolyFun + SuSiE are correlated.

We selected five traits for down-sampling analysis (analyzing $N=107K$ individuals) as the set of traits having (1) the largest number of 3Mb loci harboring a genome-wide significant SNP; (2) >10 PIP>0.95 SNPs in the SuSiE $N=107K$ analysis; and (3) $|r_g|<0.2$ with another selected trait.

Polygenic localization

Polygenic localization aims to identify a minimal set of SNPs causally explaining a given proportion of common SNP heritability. To define polygenic localization, we first define M_p^* , as the smallest integer k such that $\sum_{i \in \{s_1, \dots, s_k\}} \beta_i^2 / \sum_{i=1}^m \beta_i^2 \geq p$, where β_i are standardized SNP effect sizes, s_j denotes a ranking of β_i^2 such that $\beta_{s_1}^2 \geq \beta_{s_2}^2 \geq \dots \geq \beta_{s_m}^2$, and m is the number of common SNPs. Unfortunately, β_i^2 is unknown in practice. Polygenic localization therefore estimates an upper-bound of M_p^* , denoted as M_p . We define M_p as the smallest integer k' such that $\sum_{i \in \{s'_1, \dots, s'_{k'}\}} \beta_i^2 / \sum_{i=1}^m \beta_i^2 \geq p$, where s' is a possibly non-optimal ranking of SNPs. We note that $M_p \geq M_p^*$ by construction. We provide a full derivation of Polygenic localization in the Supplementary Note.

We now provide a brief conceptual description of PolyLoc (a full description is provided in the Supplementary Note). Briefly, PolyLoc proceeds by (1) partitioning SNPs with similar β_i^2 posterior mean estimates (using PolyFun + SuSiE estimates) into bins; (2) treating β_i as a zero-mean random variable and jointly estimating $\text{var}[\beta_i]$ in every bin using S-LDSC; and (3) finding the smallest integer k such that $\sum_{i \in \{\hat{s}_1, \dots, \hat{s}_k\}} \text{var}[\beta_i] / \sum_{i=1}^m \text{var}[\beta_i] \geq p$, where \hat{s}_j denotes the original ranking of β_i^2 posterior mean estimates from PolyFun + SuSiE. The use of $\text{var}[\beta_i] = E[\beta_i^2] - E[\beta_i]^2$ instead of β_i^2 uses the assumption that β_i has zero mean in each bin. The partitioning into bins in step 1 induces a piecewise-linear approximation of the function $(k) = \sum_{i \in \{\hat{s}_1, \dots, \hat{s}_k\}} \beta_i^2 / \sum_{i=1}^m \beta_i^2$. We use different datasets to estimate β_i^2 posterior means and to estimate $\text{var}[\beta_i]$ to prevent winner's curse. Our approach is conservative by design due to using an imperfect ranking compared to the true ranking s_1, \dots, s_m . The degree of conservativeness is a function of fine-mapping power, and thus depends on factors affecting fine-mapping power such as sample size, levels of LD at causal SNPs, MAFs of causal SNPs, and trait polygenicity.

In secondary analyses, we compared PolyLoc to an alternative method that performs polygenic localization based on prior estimates of per-SNP heritability from functional annotations, rather than posterior estimates. This alternative method uses per-SNP heritability estimates and SNP bins from step 4 of PolyFun, based only on the $N=337K$ dataset (noting that it does not suffer from winner's curse because PolyFun applies a partitioning into odd and even chromosomes).

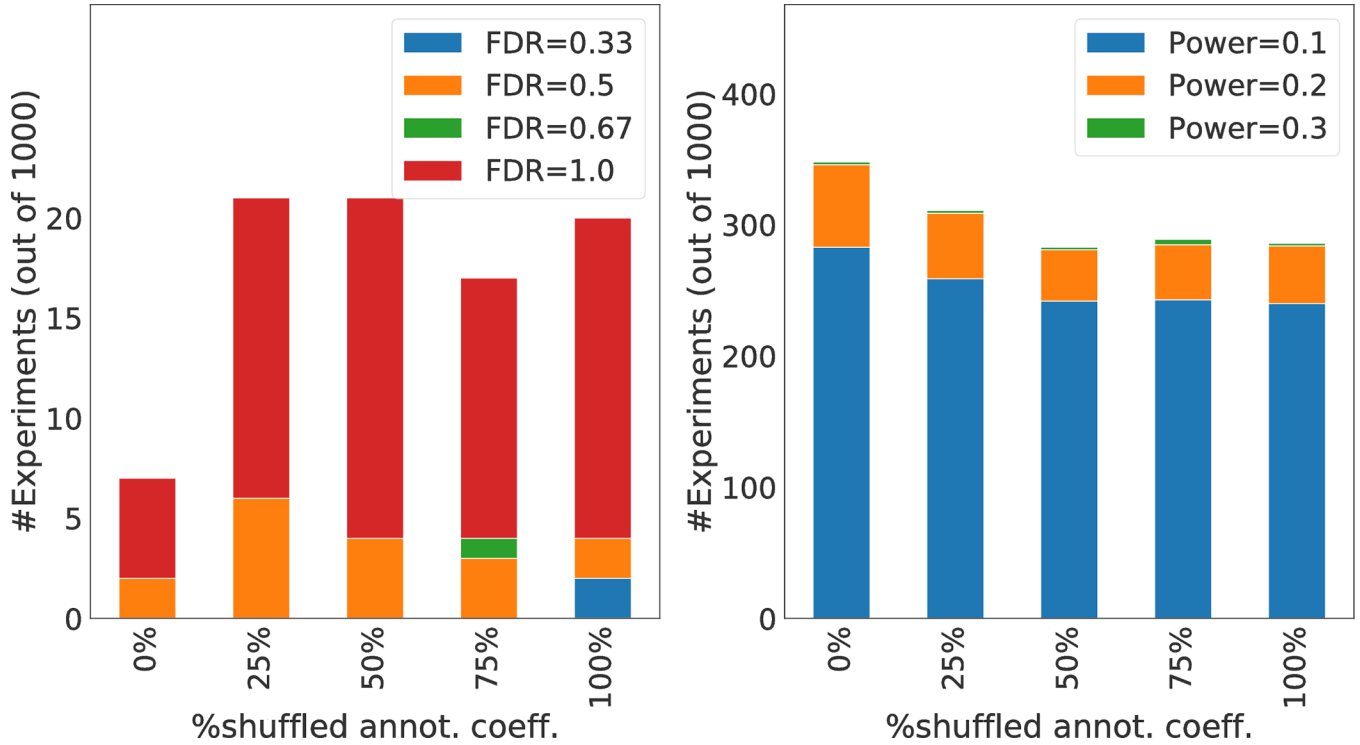
Data availability

PolyFun fine-mapping results generated in this study are available for public download at http://data.broadinstitute.org/alkesgroup/polyfun_results. Summary LD information generated in this study is available for public download at https://data.broadinstitute.org/alkesgroup/UKBB_LD. Baseline-LF v2.2.UKB annotations and LD-scores for UK Biobank SNPs are available at https://data.broadinstitute.org/alkesgroup/LDSCORE/baselineLF_v2.2.UKB.tar.gz. Access to the UK Biobank resource is available via application (<http://www.ukbiobank.ac.uk>).

Code availability

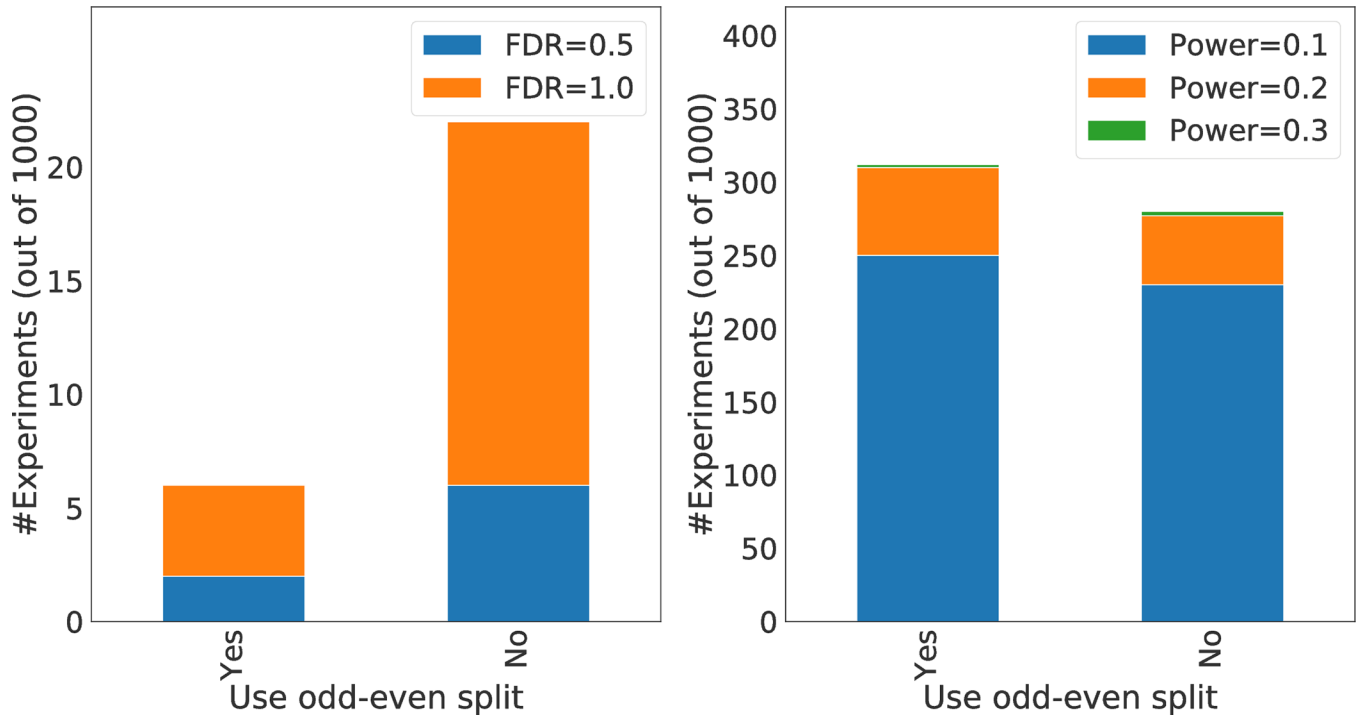
PolyFun and PolyLoc software is available at <https://github.com/omerwe/polyfun>. SuSiE software is available at <https://github.com/stephenslab/susieR>. FINEMAP software is available at <http://www.christianbenner.com/#>.

Extended Data



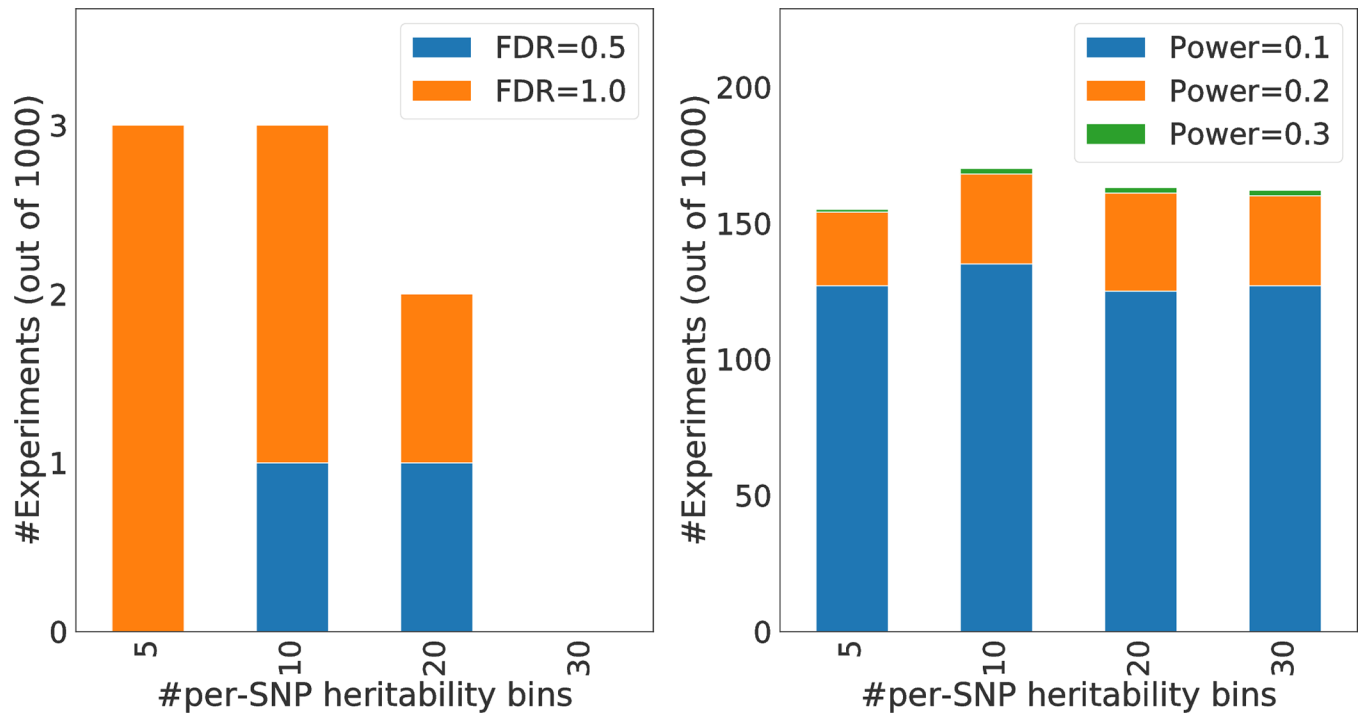
Extended Data Fig. 1: Assessing the individual impact of step 1 of PolyFun (estimating functional enrichment) via perturbation analysis, by randomly shuffling different proportions of annotation coefficient estimates.

For each evaluated value of the proportion of shuffled annotation coefficient estimates, we report the number of experiments having each obtained FDR level >0 (left panel) and the number of experiments having each obtained power level >0 (right panel), out of 1000 experiments. FDR and power are reported with respect to identifying PIP 0.95 SNPs. Experiments with FDR=0 (resp. power=0) are not reported in the left panel (resp. right panel) to improve clarity. Numerical reports are provided in Supplementary Table 6.



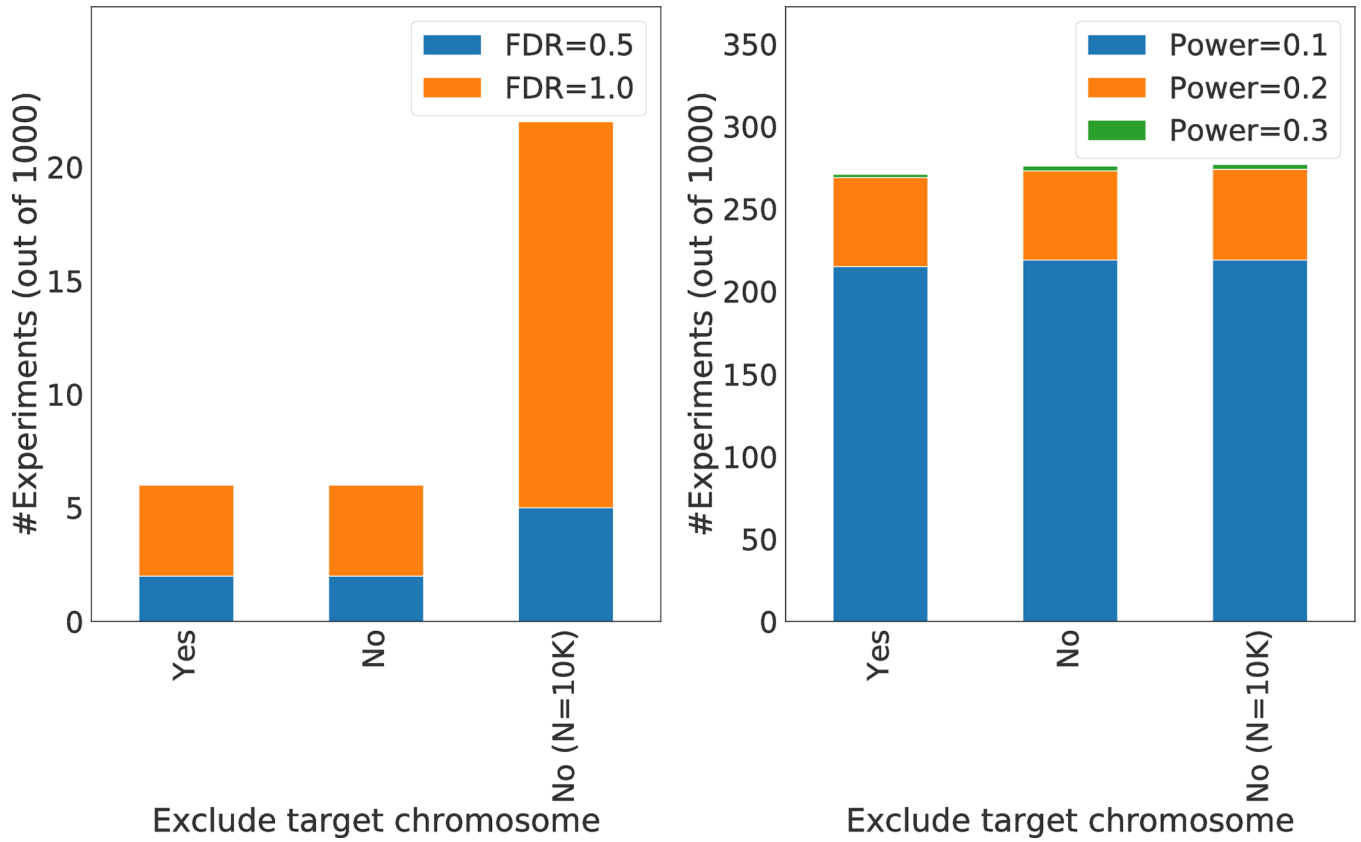
Extended Data Fig. 2: Assessing the individual impact of step 2 of PolyFun (estimating per-SNP heritabilities on odd/even chromosomes) via perturbation analysis, by using both odd and even chromosomes to estimate functional enrichment.

The figure is similar to Extended Data Figure 1 but applies a different perturbation (using both odd and even chromosomes to estimate functional enrichment). Numerical reports are provided in Supplementary Table 6.



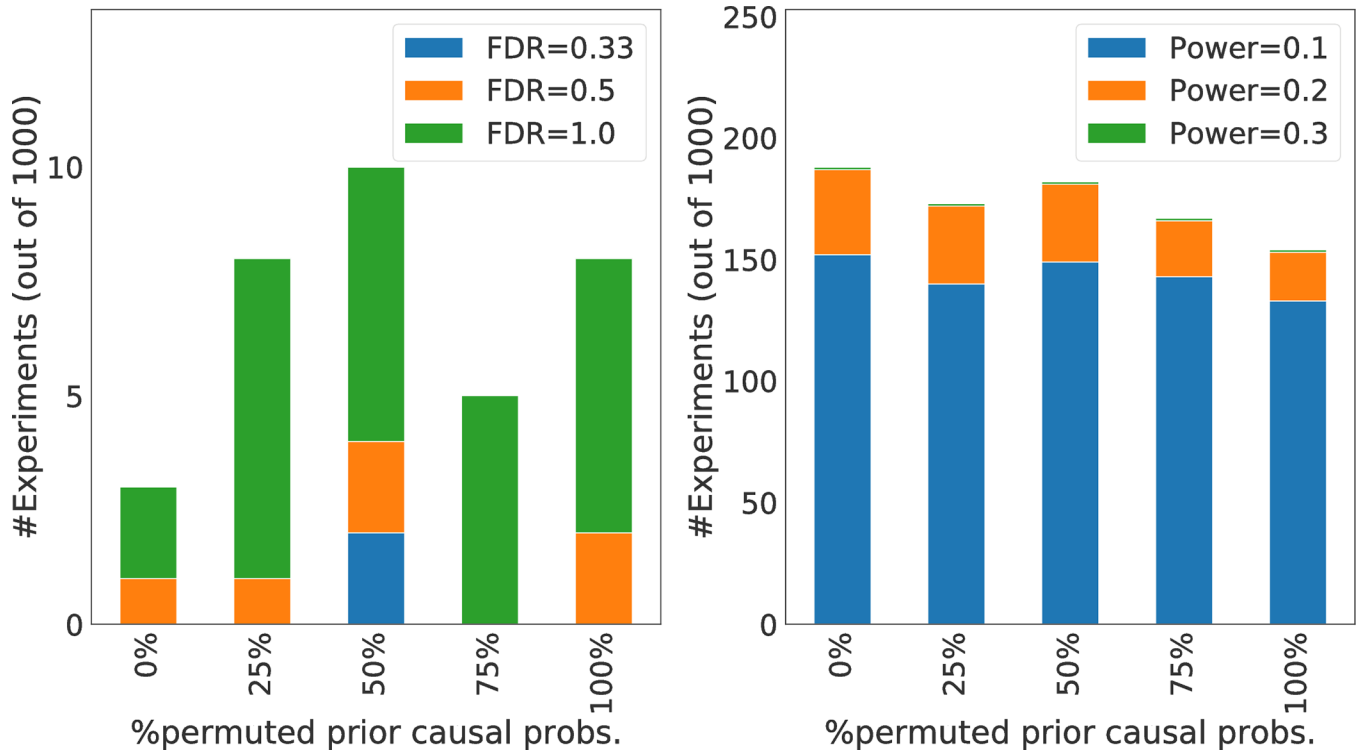
Extended Data Fig. 3: Assessing the individual impact of step 3 of PolyFun (partitioning all SNPs into 20 bins of similar per-SNP heritability) via perturbation analysis, by varying the number of per-SNP heritability bins.

The figure is similar to Extended Data Figure 1 but applies a different perturbation (changing the number of per-SNP heritability bins). Numerical reports are provided in Supplementary Table 6.



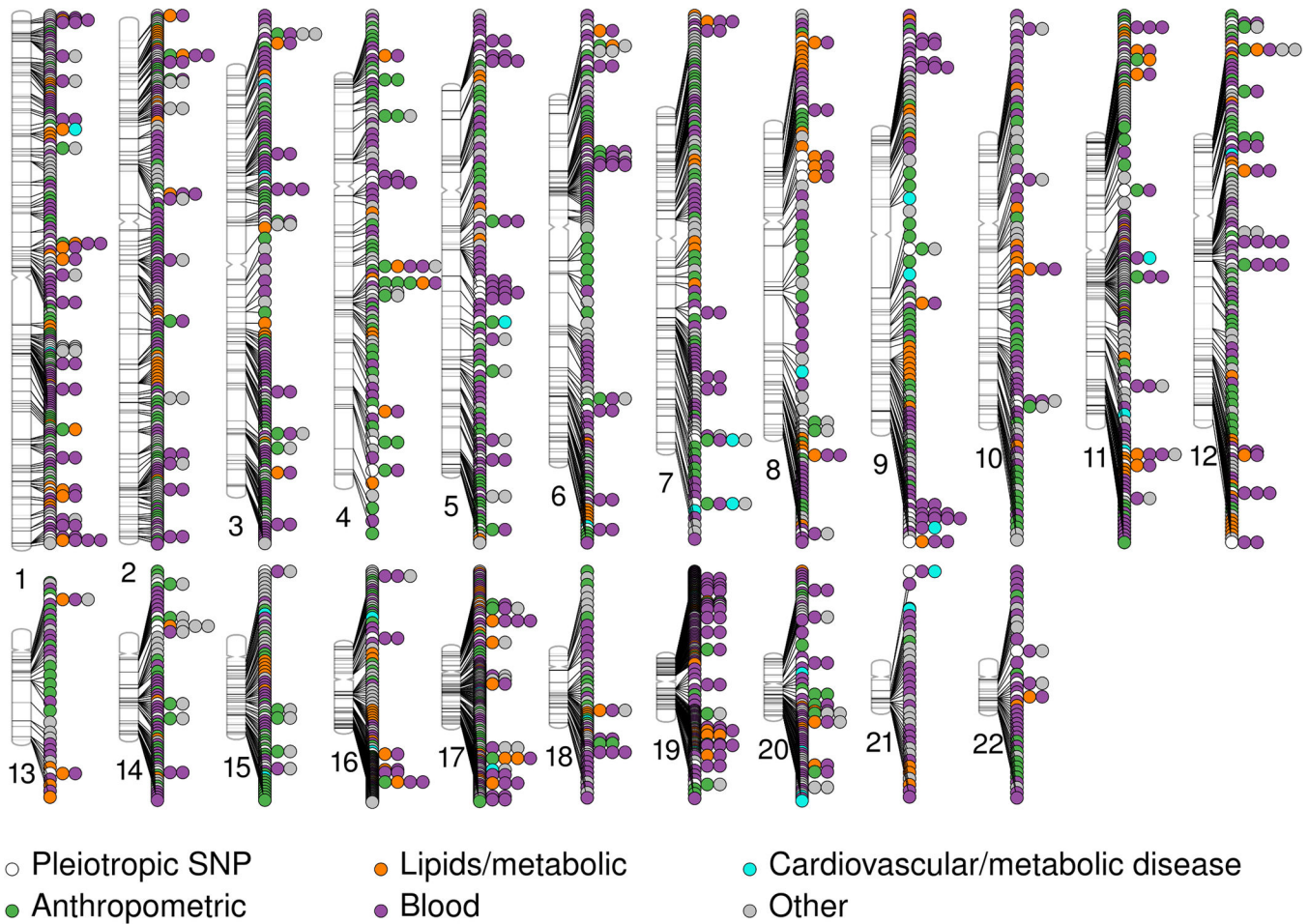
Extended Data Fig. 4: Assessing the individual impact of step 4 of PolyFun (re-estimating per-SNP heritabilities within each bin excluding the target chromosome) via perturbation analysis, by not excluding the target chromosome from the re-estimation procedure.

The figure is similar to Extended Data Figure 1 but applies a different perturbation (disables the exclusion of the target chromosome, either when using the default sample size $N=320K$ or when using a smaller sample size of $N=10K$). Numerical reports are provided in Supplementary Table 6.



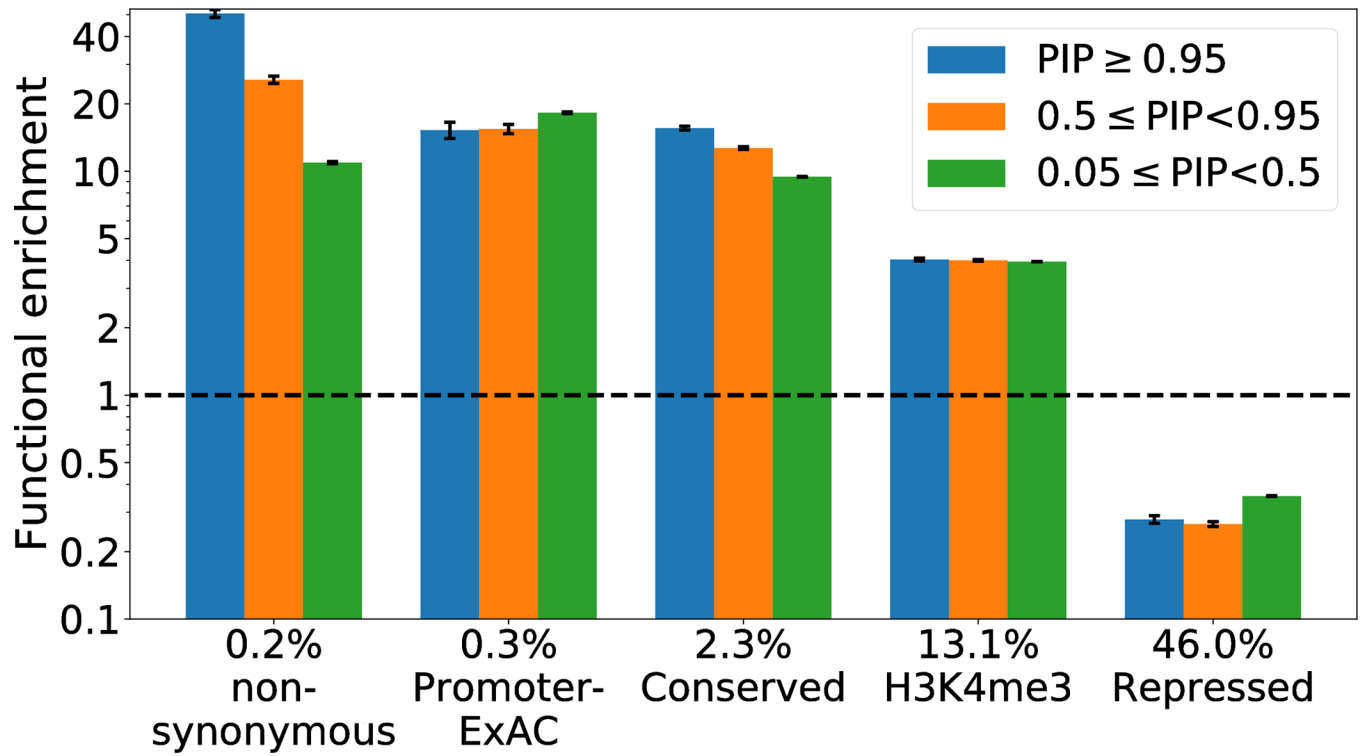
Extended Data Fig. 5: Assessing the individual impact of step 5 of PolyFun (specifying prior causal probabilities in proportion of the re-estimated per-SNP heritabilities) via perturbation analysis, by randomly permuting estimated prior causal probabilities.

The figure is similar to Extended Data Figure 1 but applies a different perturbation (randomly permuting estimated prior causal probabilities). Numerical reports are provided in Supplementary Table 6.



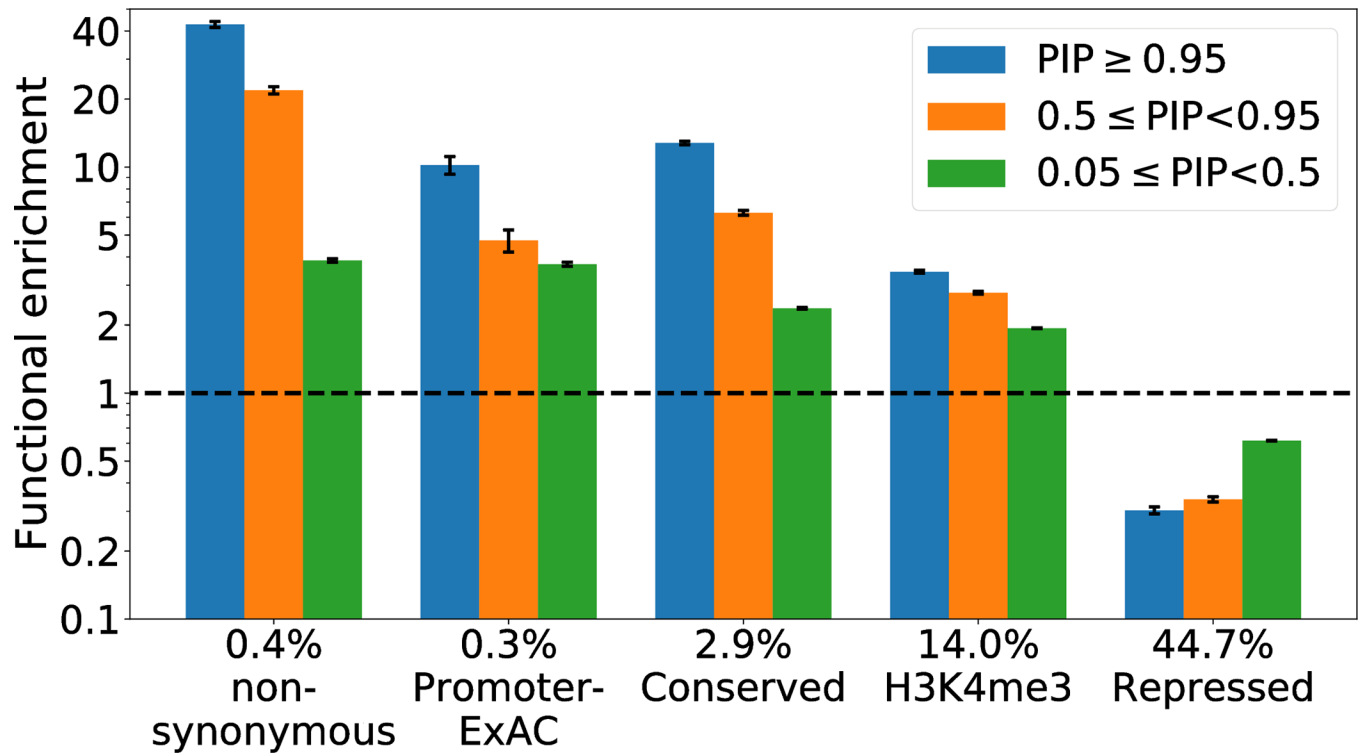
Extended Data Fig. 6: Visualization of fine-mapping results for UK Biobank traits.

We display an ideogram of all 2,225 $PIP > 0.95$ fine-mapped SNPs identified by PolyFun + SuSiE across 49 UK Biobank traits. Traits are color-coded into groups (see legend and Supplementary Table 8). White circles indicate SNPs that are pleiotropic for 2 genetically uncorrelated traits, with circles to the right of a white circle denoting the genetically uncorrelated traits (max of 5 colored circles due to space limitations). Numerical results are reported in Supplementary Table 10.



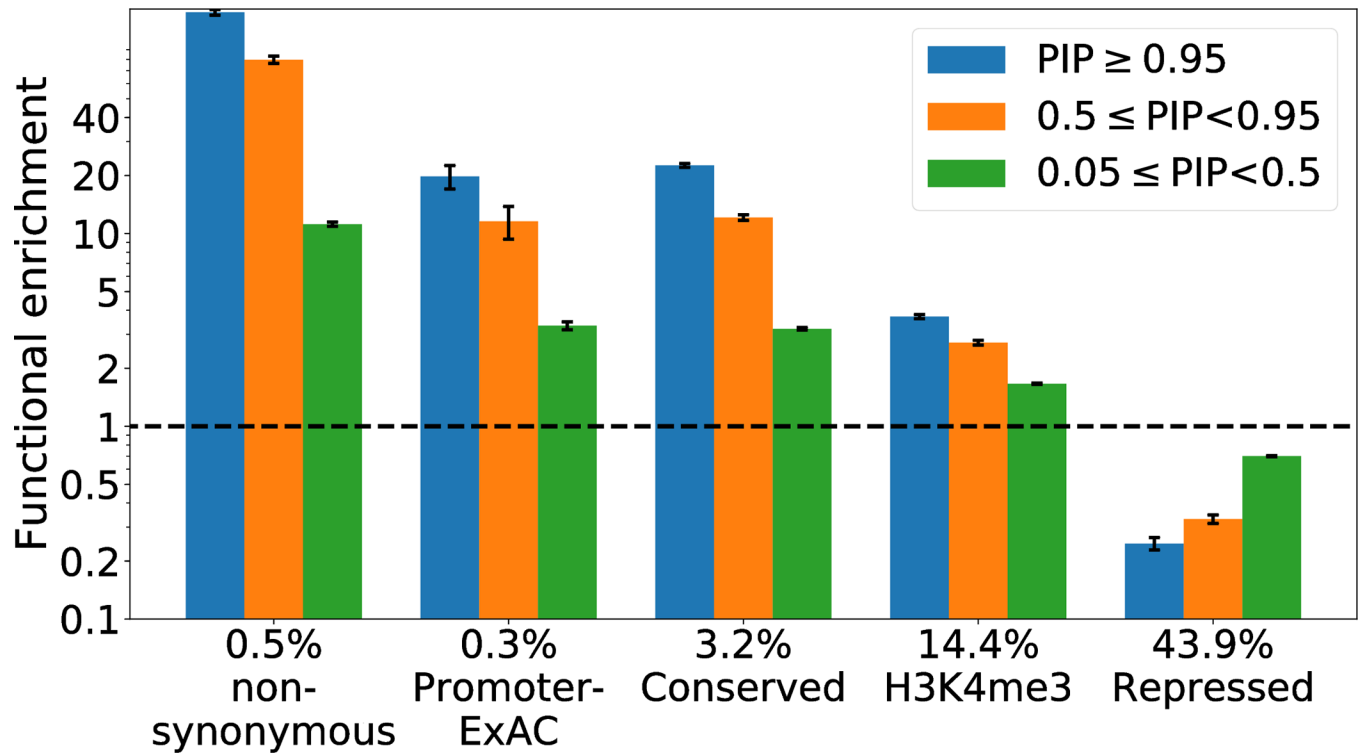
Extended Data Fig. 7: Functional enrichment of PolyFun + SuSiE fine-mapped common SNPs for UK Biobank traits.

The figure is analogous to Figure 4 but uses PIPs computed by PolyFun + SuSiE instead of SuSiE. Numerical results are reported in Supplementary Table 26.



Extended Data Fig. 8: Functional enrichment of SuSiE fine-mapped MAF>0.001 SNPs for UK Biobank traits.

The figure is analogous to Figure 4 but uses MAF>0.001 SNPs instead of common (MAF>0.05) SNPs. Numerical results are reported in Supplementary Table 27.



Extended Data Fig. 9: Functional enrichment of SuSiE fine-mapped low-frequency and rare SNPs for UK Biobank traits.

The figure is analogous to Figure 4 but uses only low-frequency and rare SNPs ($0.05 > \text{MAF} > 0.001$) instead of common ($\text{MAF} > 0.05$) SNPs. Numerical results are reported in Supplementary Table 28.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank B. Pasaniuc, G. Kichaev, M. Stephens, G. Wang, M. Kanai, B.M. Schilder and T. Raj for helpful discussions. This research was conducted using the UK Biobank Resource under Application #16549 and was funded by NIH grants U01 HG009379, R37 MH107649, R01 MH101244 and R01 HG006399, and by the Academy of Finland grants 288509 and 312076. HKF is supported by Eric and Wendy Schmidt. Computational analyses were performed on the O2 High-Performance Compute Cluster at Harvard Medical School.

References

1. Visscher PM et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22 (2017). [PubMed: 28686856]
2. Shendure J, Findlay GM & Snyder MW Genomic medicine—progress, pitfalls, and promise. *Cell* 177, 45–57 (2019). [PubMed: 30901547]
3. Schaid DJ, Chen W & Larson NB From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19, 491–504 (2018). [PubMed: 29844615]
4. The Wellcome Trust Case Control Consortium et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* 44, 1294–1301 (2012). [PubMed: 23104008]

5. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B & Eskin E Identifying causal variants at loci with multiple signals of association. *Genetics* 198, 497–508 (2014). [PubMed: 25104515]
6. Chen W et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* 200, 719–736 (2015). [PubMed: 25948564]
7. Farh KK-H et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337 (2015). [PubMed: 25363779]
8. Huang H et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* 547, 173–178 (2017). [PubMed: 28658209]
9. Mahajan A et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* 50, 559 (2018). [PubMed: 29632382]
10. 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 (2018). [PubMed: 30297969]
11. 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). [PubMed: 30224649]
12. Ulirsch JC et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0362-6.
13. Maurano MT et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* 337, 1190–1195 (2012). [PubMed: 22955828]
14. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74 (2012). [PubMed: 22955616]
15. Trynka G et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124 (2013). [PubMed: 23263488]
16. The Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330 (2015). [PubMed: 25693563]
17. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235 (2015). [PubMed: 26414678]
18. Kichaev G et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* 10, e1004722 (2014). [PubMed: 25357204]
19. Kichaev G et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* 33, 248–255 (2017). [PubMed: 27663501]
20. Chen W, McDonnell SK, Thibodeau SN, Tillmans LS & Schaid DJ Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics* 204, 933–958 (2016). [PubMed: 27655946]
21. 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. Ser. B Stat. Methodol.* n/a, (2020).
22. Benner C et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* 32, 1493–1501 (2016). [PubMed: 26773131]
23. Benner C, Havulinna A, Salomaa V, Ripatti S & Pirinen M Refining fine-mapping: effect sizes and regional heritability. *bioRxiv* 318618 (2018).
24. Gazal S et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* 49, 1421–1427 (2017). [PubMed: 28892061]
25. Gazal S et al. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* 50, 1600–1607 (2018). [PubMed: 30297966]
26. Gazal S, Marquez-Luna C, Finucane HK & Price AL Reconciling S-LDSC and LDK functional enrichment estimates. *Nat. Genet.* 51, 1202–1204 (2019). [PubMed: 31285579]
27. Bycroft C et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203 (2018). [PubMed: 30305743]
28. 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). [PubMed: 28942963]

29. Pasaniuc B & Price AL Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* 18, 117–127 (2017). [PubMed: 27840428]
30. Marquez-Luna C et al. Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* 375337 (2018).
31. Loh P-R, Kichaev G, Gazal S, Schoech AP & Price AL Mixed-model association for biobank-scale datasets. *Nat. Genet.* 50, 906–908 (2018). [PubMed: 29892013]
32. Pickrell JK et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* 48, 709 (2016). [PubMed: 27182965]
33. Hujoel ML, Gazal S, Hormozdiari F, van de Geijn B & Price AL Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *Am. J. Hum. Genet.* 104, 611–624 (2019). [PubMed: 30905396]
34. O'Connor LJ et al. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am. J. Hum. Genet.* 105, 456–476 (2019). [PubMed: 31402091]
35. Zeng J et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet.* 50, 746 (2018). [PubMed: 29662166]
36. Zhang Y, Qi G, Park J-H & Chatterjee N Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* 50, 1318 (2018). [PubMed: 30104760]
37. Zhu X & Stephens M Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* 9, 4361 (2018). [PubMed: 30341297]
38. Moser G et al. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS Genet* 11, e1004969 (2015). [PubMed: 25849665]
39. Yang J et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* 47, 1114–20 (2015). [PubMed: 26323059]
40. Purcell SM et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752 (2009). [PubMed: 19571811]
41. Yang J et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569 (2010). [PubMed: 20562875]
42. Loh P-R et al. Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* 47, 1385–1392 (2015). [PubMed: 26523775]
43. Boyle EA, Li YI & Pritchard JK An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169, 1177–1186 (2017). [PubMed: 28622505]
44. Wuttke M et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* 51, 957 (2019). [PubMed: 31152163]
45. Landi MT et al. Genome-wide association meta-analyses combining multiple risk phenotypes provide insights into the genetic architecture of cutaneous melanoma susceptibility. *Nat. Genet.* 52, 494–504 (2020). [PubMed: 32341527]
46. Vujkovic M et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* 1–12 (2020). [PubMed: 31911675]
47. Yang J et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* 44, 369 (2012). [PubMed: 22426310]
48. 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). [PubMed: 27140283]
49. Márquez-Luna C, Loh P-R, Consortium, S. A. T. 2 D. (SAT2D), Consortium, S. T. 2 D. & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* 41, 811–823 (2017). [PubMed: 29110330]
50. Martin AR et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584 (2019). [PubMed: 30926966]
51. Claussnitzer M et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* 373, 895–907 (2015). [PubMed: 26287746]

52. Jung I et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* 1–8 (2019). [PubMed: 30578416]
53. Zeggini E, Gloyn AL, Barton AC & Wain LV Translational genomics and precision medicine: Moving from the lab to the clinic. *Science* 365, 1409–1413 (2019). [PubMed: 31604268]
54. Solovieff N, Cotsapas C, Lee PH, Purcell SM & Smoller JW Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* 14, 483 (2013). [PubMed: 23752797]
55. Wang K, Li M & Hakonarson H Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* 11, 843 (2010). [PubMed: 21085203]
56. De Leeuw CA, Neale BM, Heskes T & Posthuma D The statistical properties of gene-set analysis. *Nat. Rev. Genet.* 17, 353 (2016). [PubMed: 27070863]
57. Haworth S et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nat. Commun.* 10, 333 (2019). [PubMed: 30659178]
58. Kichaev G & Pasaniuc B Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* 97, 260–271 (2015). [PubMed: 26189819]

Methods-only References

59. Wang H & Song M Ckmeans.1d.dp: Optimal k-means Clustering in One Dimension by Dynamic Programming. *R J* 3, 29–33 (2011). [PubMed: 27942416]
60. Shi H, Kichaev G & Pasaniuc B Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* 99, 139–53 (2016). [PubMed: 27346688]
61. The UK10K Consortium et al. The UK10K project identifies rare variants in health and disease. *Nature* 526, 82 (2015). [PubMed: 26367797]
62. Price AL et al. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* 83, 132–135 (2008). [PubMed: 18606306]
63. Wolfe D, Dudek S, Ritchie MD & Pendergrass SA Visualizing genomic information across chromosomes with PhenoGram. *BioData Min.* 6, 18 (2013). [PubMed: 24131735]
64. Welter D et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42, D1001–D1006 (2014). [PubMed: 24316577]

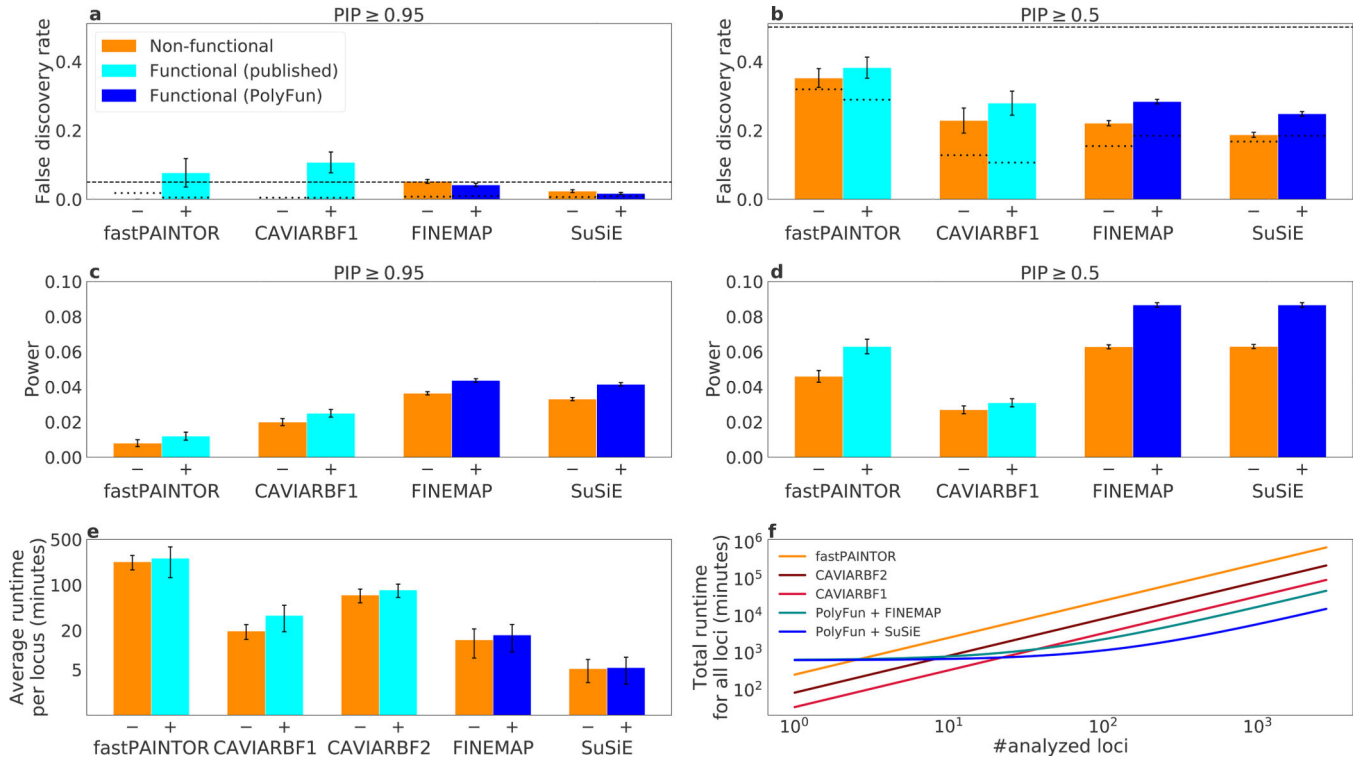


Figure 1: Calibration, power and computational cost of fine-mapping methods in main simulations.

(a-b) FDR at PIP=0.95 (a) and PIP=0.5 (b). Upper dashed horizontal lines denote conservative FDR estimates. Lower dotted horizontal lines denote anti-conservative FDR estimates, which are not recommended (Supplementary Note). (c-d) Power at PIP=0.95 (c) and PIP=0.5 (d). The first bar of each method uses non-functionally informed fine-mapping (denoted -), and the second uses functionally informed fine-mapping (denoted +). (e) The average runtime required to fine-map a 3Mb locus in a genome-wide analysis (log scale). The first bar of each method uses non-functionally informed fine-mapping (denoted -), and the second uses functionally informed fine-mapping (denoted +). (f) The total runtime required to fine-map different numbers of loci, for functionally informed fine-mapping methods only (log scale). The runtimes of PolyFun + SuSiE and PolyFun + FINEMAP are sub-linear because they include the fixed preprocessing cost of computing prior causal probabilities (630 minutes). Error bars denote standard errors. Numerical results, including results for CAVIARBF2- and CAVIARBF2, and including panel (f) results for non-functionally informed methods, are reported in Supplementary Table 4.

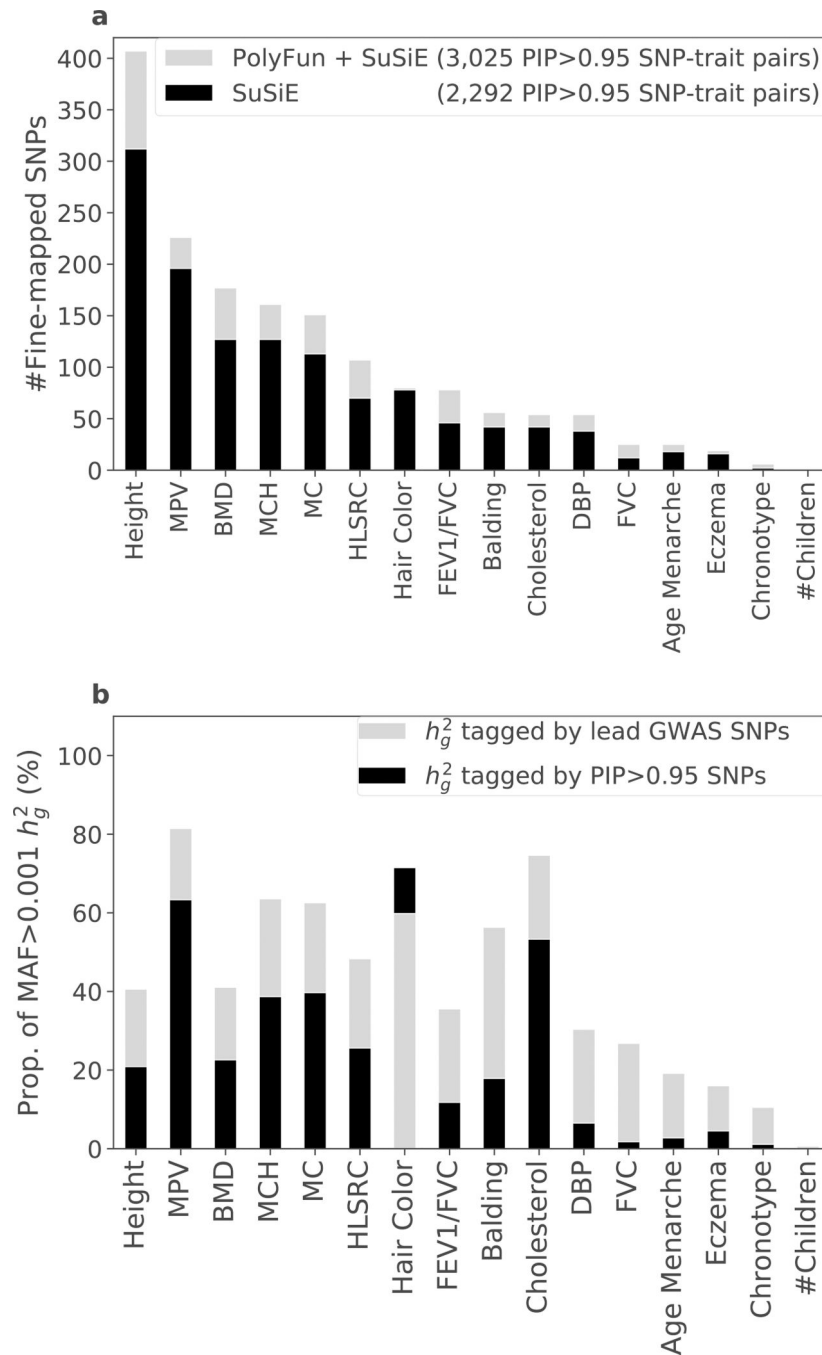


Figure 2: Summary of fine-mapping results for UK Biobank traits.

(a) the number of SNPs with PIP > 0.95 identified by SuSiE (black bars) and PolyFun + SuSiE (gray bars) across 16 genetically uncorrelated traits in the UK Biobank. Traits are ordered by PolyFun + SuSiE results. The numbers in the legend refer to the sum of all 49 traits analyzed. (b) The proportion of MAF > 0.001 SNP-heritability (h_g^2) tagged by lead GWAS SNPs (gray bars) and by PolyFun + SuSiE PIP > 0.95 SNPs (black bars). Traits are ordered as in panel (a). For hair color, the h_g^2 tagged by PIP > 0.95 SNPs is greater than h_g^2 tagged by lead GWAS SNPs. MPV: Mean platelet volume; BMD: bone mineral density;

MCH: mean corpuscular hemoglobin; MC: monocyte count; HLSRC: high light scatter reticulocyte count; FEV1/FVC: ratio of forced expiratory volume to forced vital capacity; DBP: diastolic blood pressure; FVC: forced vital capacity. Numerical results are reported in Supplementary Tables 9,14.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

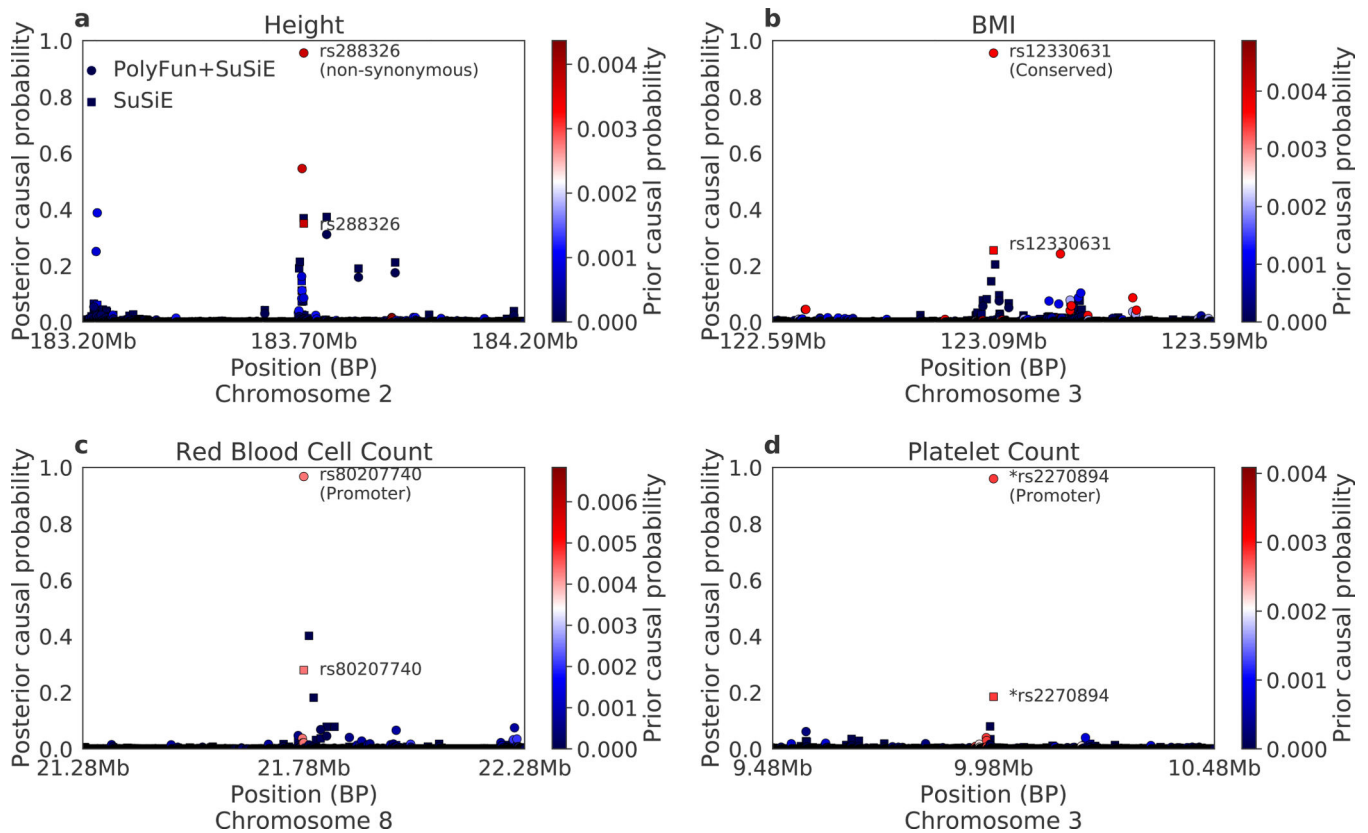


Figure 3: Examples of the advantages of functionally-informed fine-mapping for UK Biobank traits.

We report four examples where PolyFun + SuSiE identified a fine-mapped common SNP (PIP>0.95) but SuSiE did not (PIP<0.5 for all SNPs within 1Mb). Circles denote PolyFun + SuSiE PIPs and squares denote SuSiE PIPs. SNPs are shaded according to their prior causal probabilities as estimated by PolyFun. The top PolyFun + SuSiE SNP is labeled (next to its PolyFun + SuSiE PIP and its SuSiE PIP). The annotation of each top PolyFun + SuSiE SNP that is most enriched among SuSiE PIP>0.95 SNPs (Methods) is reported in parentheses below its label. Asterisks denote lead GWAS SNPs. Numerical results are reported in Supplementary Table 16.

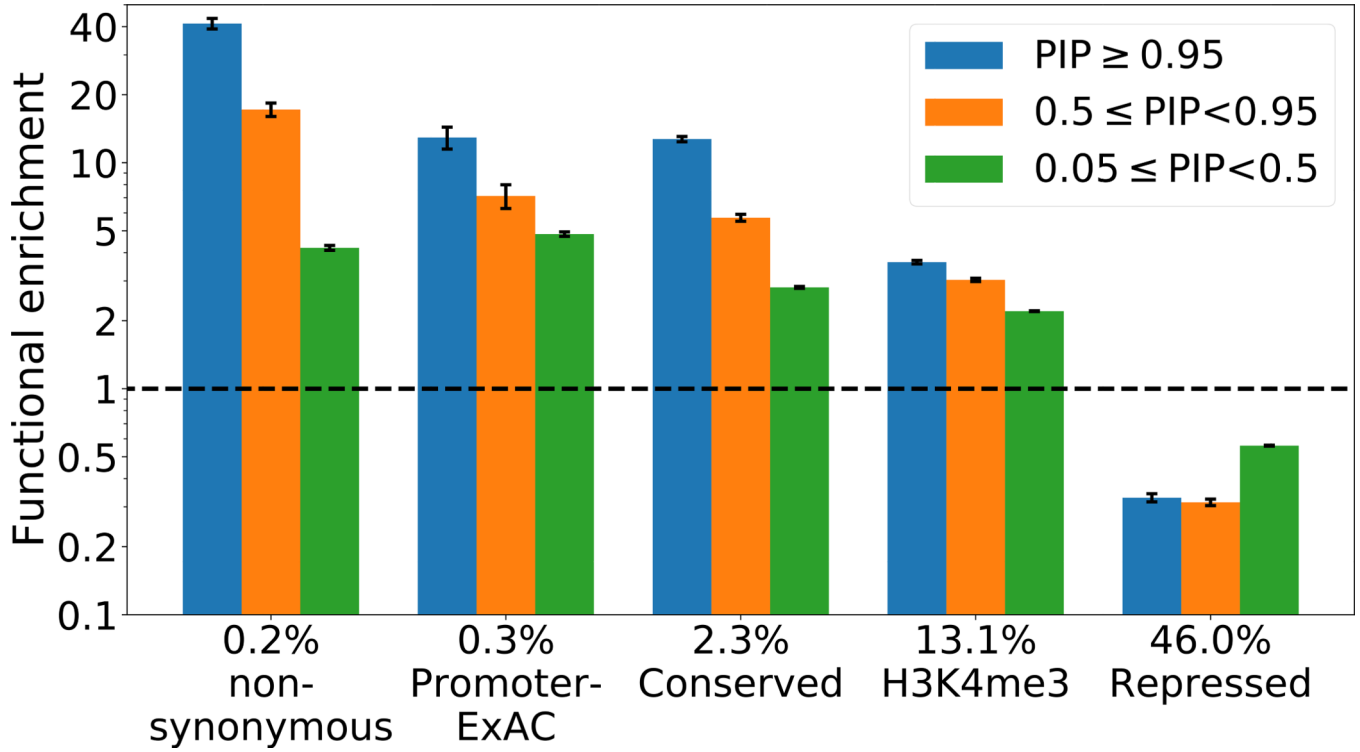


Figure 4: Functional enrichment of SuSiE fine-mapped common SNPs for UK Biobank traits.

We report the functional enrichment of fine-mapped common SNPs (defined as the proportion of common SNPs in a PIP range lying in an annotation divided by the proportion of genome-wide common SNPs lying in the annotation) for 5 selected binary annotations, meta-analyzed across 14 genetically uncorrelated UK Biobank traits with 10 PIP > 0.95 SNPs (log scale). The proportion of common SNPs lying in each binary annotation is reported above its name. The horizontal dashed line denotes no enrichment. Error bars denote standard errors. Numerical results for all 50 main binary annotations and all traits are reported in Supplementary Table 17.

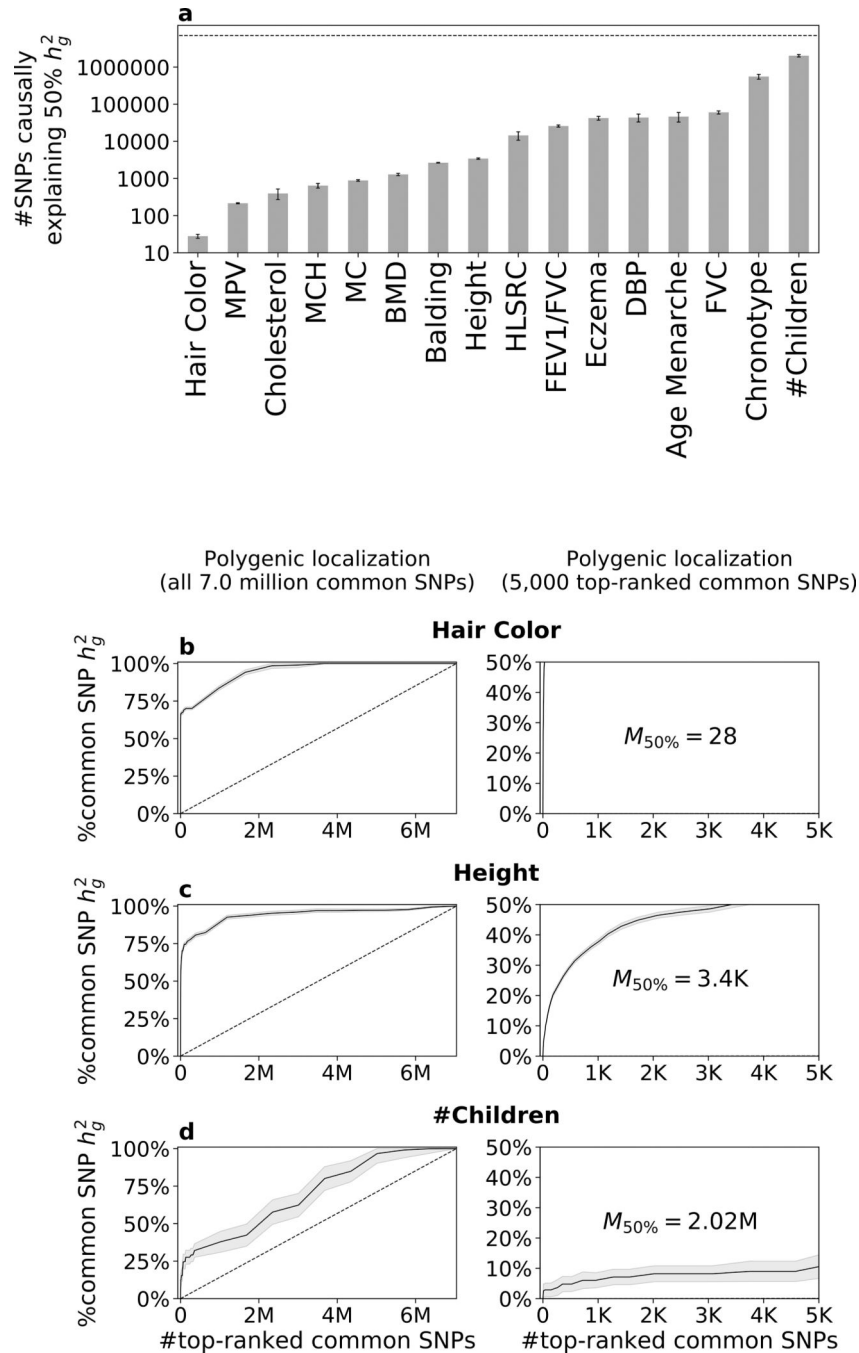


Figure 5: Polygenic localization results for UK Biobank traits.

(a) $M_{50\%}$ estimates across 16 genetically uncorrelated traits. For each trait, we report the number of top-ranked common SNPs (using PolyFun + SuSiE posterior per-SNP heritability estimates for ranking) causally explaining 50% of common SNP heritability, and its standard error (log scale). The horizontal dashed line denotes the total number of common SNPs in the analysis (7.0 million). (b-d) The proportion of common SNP heritability of (b) hair color, (c) height, and (d) number of children explained by different numbers of top-ranked SNPs, for all 7.0 million common SNPs (left) and the 5,000 top-ranked common (right).

Gray shading denotes standard errors. Dashed black lines denote a null model with a constant per-SNP heritability. We also report the number of top-ranked SNPs causally explaining 50% of common SNP heritability, denoted $M_{50\%}$. Discontinuities in the slope indicate transitions between SNP bins. Numerical results for all 49 UK Biobank traits are reported in Supplementary Table 31.

Table 1:
Summary of methods evaluated in main simulations.

For each method we report whether it incorporates functional data, the maximum number of functional annotations that we specified under default simulation settings (for fastPAINOTOR we selected the number of annotations that maximized power while maintaining correct calibration; Methods), the maximum number of causal SNPs modeled per locus (or the exact number for SuSiE and PolyFun + SuSiE), and the corresponding reference. For fastPAINOTOR and CAVIARBF, – denotes the exclusion of functional data. For CAVIARBF, 1 or 2 denotes the maximum number of causal variants. PolyFun + FINEMAP uses a new version of FINEMAP that we introduce here that incorporates prior causal probabilities.

Method	Functional data	Max #annotations	Max #causal SNPs	Ref.
fastPAINOTOR–	No	N/A	Unlimited	19
fastPAINOTOR	Yes	10	Unlimited	19
CAVIARBF1–	No	N/A	1	6
CAVIARBF1	Yes	Unlimited	1	20
CAVIARBF2–	No	N/A	2	6
CAVIARBF2	Yes	Unlimited	2	20
FINEMAP	No	N/A	10	22,23
PolyFun + FINEMAP	Yes	Unlimited	10	This paper
SuSiE	No	N/A	10	21
PolyFun + SuSiE	Yes	Unlimited	10	This paper

Table 2:
Pleiotropic fine-mapped SNPs for UK Biobank traits.

We report SNPs fine-mapped (PIP>0.95) for 4 genetically uncorrelated traits ($|r_g|<0.2$). For each SNP we report its name (SNP), position (hg19), MAF in the UK Biobank, closest gene(s) (using data from the GWAS catalog⁶⁴), top annotation (Methods) and fine-mapped traits (and the number of fine-mapped traits). SNPs are ordered first by the number of fine-mapped traits and then by genomic position. HDL: HDL cholesterol; MC: monocyte count; MPV: mean platelet volume; HLSRC: high light scatter reticulocyte count; Cholesterol: total cholesterol; RBCDW: red blood cell distribution width; FEV1/FVC: ratio of forced expiratory volume to forced vital capacity; MCH: mean corpuscular hemoglobin; SBP: systolic blood pressure; DBP: diastolic blood pressure; FVC: forced vital capacity; Cardiovascular: cardiovascular-related disease; RBC: red blood cell count; LC: lymphocyte count; HbA1c: Hemoglobin A1c; WHR: waist-hip ratio (adjusted for BMI). Results for all 223 pleiotropic fine-mapped SNPs are reported in Supplementary Table 15.

SNP	Position	MAF	Closest gene(s)	Annotation	Traits
rs13107325	chr4:103188709	0.08	SLC39A8	non-synonymous	BMI, Balding, Cholesterol, DBP, FVC, Height, RBC, WHR (8)
rs1229984	chr4:100239319	0.02	ADH1B	non-synonymous	BMI, LDL, MCH, MPV, SBP, Vitamin D (6)
rs76895963	chr12:4384844	0.02	CCND2,CCND2-AS1	Conserved	BMD, Height, RBC, SBP, Triglycerides (5)
rs140584594	chr1:110232983	0.27	GSTM1	non-synonymous	HDL, Height, MC, MPV (4)
rs3811444	chr1:248039451	0.33	TRIM58	non-synonymous	HLSRC, HbA1c, Platelet Count, RBC (4)
rs1260326	chr2:27730940	0.39	GCKR	non-synonymous	Cholesterol, Height, Platelet Count, RBCDW (4)
rs2270894	chr3:9975386	0.2	CRELD1,IL17RC	Conserved	BMD, FEV1/FVC, Height, Platelet Count (4)
rs11556924	chr7:129663496	0.39	ZC3HC1	non-synonymous	Age Menarche, Cardiovascular, Height, Platelet Count (4)
rs3918226	chr7:150690176	0.08	NOS3	Conserved	Eczema, Height, High Cholesterol, MPV (4)
rs150813342	chr9:135864513	0.01	GFI1B	Conserved	Eosinophil Count, HLSRC, MCH, Platelet Count (4)
rs964184	chr11:116648917	0.13	ZPR1	DHS	Cholesterol, MPV, RBCDW, Vitamin D (4)
rs35979828	chr12:54685880	0.07	NFE2	Conserved	Eosinophil Count, Platelet Count, RBC, RBCDW (4)
rs2277339	chr12:57146069	0.1	PRIM1	non-synonymous	Height, LC, RBC, RBCDW (4)
rs72681869	chr14:50655357	0.01	SOS2	non-synonymous	FVC, Hair Color, HbA1c, SBP (4)
rs61745086	chr16:88782050	0.01	PIEZO1,CTU2	non-synonymous	HLSRC, HbA1c, Height, RBC (4)
rs34557412	chr17:16852187	0.01	TNFRSF13B	non-synonymous	HbA1c, MC, MPV, RBC (4)
rs77542162	chr17:67081278	0.02	ABCA6	non-synonymous	HbA1c, Height, LDL, Platelet Count (4)

Table 3:
Summary of mismatched reference LD simulations.

For each experiment (exp) we report: (GWAS) The sample size and population of the target sample (UK denotes British-ancestry individuals from UK Biobank; EUR denotes non-British European-ancestry individuals from UK Biobank, REL indicates that pairs of related individuals are included in the sample); (LD) the sample size and population of the LD reference panel (UK denotes British-ancestry individuals from UK Biobank; UK10K denotes individuals from the UK10K cohort; numbers in parentheses indicate how many individuals overlap the target sample, if any; “none” indicates that there is no LD reference panel); (Generative SNPs) The set of SNPs from which we sampled causal SNPs (UKB: the set of UK Biobank imputed SNPs with INFO score >0.6 and UKB MAF>0.1%; UK10K: the set of UK10K SNPs; INF: the set of UKB imputed SNPs with INFO score >0.9; COM: the set of UKB imputed SNPs with MAF >1% in British-ancestry individuals); (SNPs analyzed) the set of SNPs that was used for fine-mapping; and (max. L) The maximum number of causal SNPs per locus assumed by PolyFun + SuSiE that maintains FDR<0.05 at a PIP=0.95 threshold (selected from the options 1,2,3,10; - indicates that none of these options maintains FDR<0.05). Horizontal lines indicate the partitioning into types of experiments described in the Supplementary Note. Numerical results are reported in Supplementary Table 7.

exp	GWAS	LD	Generative SNPs	SNPs analyzed	max. L
a	44K UK	44K UK (44K overlap)	UKB	UKB	10
b	44K UK	44K UK	UKB	UKB	10
c	44K UK	4K UK	UKB	UKB	10
d	44K UK	400 UK	UKB	UKB	1
e	44K UK	none	UKB	UKB	1
f	293K UK	44K UK	UKB	UKB	10
g	293K UK	4K UK	UKB	UKB	2
h	293K UK	4K UK (4K overlap)	UKB	UKB	2
i	44K EUR	44K UK	UKB	UKB	3
j	44K EUR	4K UK	UKB	UKB	2
k	44K EUR	400 UK	UKB	UKB	1
l	22K EUR+22K UK	44K UK (22K overlap)	UKB	UKB	3
m	44K UK-REL	44K UK	UKB	UKB	10
n	44K EUR-REL	44K UK	UKB	UKB	3
o	44K UK	3.6K UK10K	UKB	UK10K∩UKB	-
p	44K UK	3.6K UK10K	UK10K∩UKB	UK10K∩UKB	2
q	44K UK	3.6K UK10K	UK10K∩UKB∩INF	UK10K∩UKB∩INF	10
r	44K UK	3.6K UK10K	UK10K∩UKB∩COM	UK10K∩UKB∩COM	1
s	44K UK	4K UK	UK10K∩UKB	UK10K∩UKB	10