



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

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

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 a posterior causal probability >0.95 than identified in their nonfunctionally informed counterparts. In analyses of 49 UK Biobank traits (average $n = 318,000$), PolyFun + SuSiE identified 3,025 fine-mapped variant-trait pairs with posterior causal probability >0.95, a >32% improvement versus 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.

Genome-wide association studies (GWAS) 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 use only genome-wide-significant loci to estimate functional enrichment (or can incorporate only 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 and FINEMAP^{22,23}. PolyFun estimates functional enrichment using a broad set of coding, conserved, regulatory, minor allele frequency (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 more powerful than previous fine-mapping methods, with a >20% power increase over nonfunctionally 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 $n=318,000$) with in-sample LD and identify 3,025 fine-mapped variant-trait pairs with posterior causal probability >0.95, spanning 2,225

unique variants. Of these variants, 223 were fine-mapped for multiple genetically uncorrelated traits, indicating pervasive pleiotropy. We further used the 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 and FINEMAP^{22,23}. For each target locus, PolyFun robustly specifies prior causal probabilities for all SNPs on the corresponding odd (respectively 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 and Supplementary Table 1) using an L2-regularized extension of S-LDSC¹⁷, restricted to even (respectively odd) chromosomes; (2) estimating per-SNP heritabilities for SNPs on odd (respectively 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 (respectively 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.

¹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ²Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. ³Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ⁴Program in Biological and Biomedical Sciences, Harvard Medical School, Cambridge, MA, USA. ⁵The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ⁶Department of Public Health, University of Helsinki, Helsinki, Finland. ⁷Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland. ⁸Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.
 ☎e-mail: oweihsbrod@hsph.harvard.edu; aprince@hsph.harvard.edu

Table 1 | Summary of methods evaluated in main simulations

Method	Functional data	Max. no. of annotations	Max. no. of causal SNPs	Reference
fastPAINTOR–	No	N/A	Unlimited	¹⁹
fastPAINTOR	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 article
SuSiE	No	N/A	10	²¹
PolyFun + SuSiE	Yes	Unlimited	10	This article

For each method we report whether it incorporates functional data, the maximum number of functional annotations that we specified under default simulation settings (for fastPAINTOR 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 fastPAINTOR 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 introduced here that incorporates prior causal probabilities. N/A, not available.

The L2 regularization in step 1 improves the accuracy of per-SNP heritability estimation. The partitioning into odd and even chromosomes in steps 1 and 2, and the exclusion of the target chromosome in step 4 prevent winner's curse. 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 s.d. increases in phenotype per 1 s.d. 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 3-Mb 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 Methods; we have released open-source software implementing PolyFun together 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 (for running both S-LDSC and 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 3-Mb loci on chromosome 1, each containing 1,468–27,784 imputed SNPs with MAF ≥ 0.001 and INFO score ≥ 0.6 (including short indels; Supplementary Table 2). We estimated prior causal probabilities using 18,212,157 genome-wide imputed SNPs with MAF ≥ 0.001 and INFO score ≥ 0.6 . We simulated traits with heritability equal to 25% and a 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 $n=320,000$ samples. Further details are provided in Methods.

We evaluated ten fine-mapping methods (Methods and 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 (for example, 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 the $1 - \text{PIP}$ threshold, which is more conservative than an exact estimate (Fig. 1a,b, Supplementary Note and Supplementary Table 4). No method except CAVIARBF2– and CAVIARBF2 had significantly inflated FDRs, although fastPAINTOR and CAVIARBF1 had suggestive evidence of inflated FDRs. We assessed power via the proportion of true causal SNPs with a PIP above a given threshold, aggregating the results across all simulations. PolyFun + FINEMAP was the most powerful method, identifying $>5\%$ more causal SNPs with a PIP > 0.95 than PolyFun + SuSiE and $>20\%$ more causal SNPs with a PIP > 0.95 than FINEMAP; PolyFun + SuSiE was the second most powerful method, identifying $>25\%$ more causal SNPs with a PIP > 0.95 than SuSiE (Fig. 1c,d and 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 3-Mb locus in 5 min on average (excluding fixed preprocessing time; see below) (Fig. 1e and Supplementary Table 4). CAVIARBF methods allowing more than two causal SNPs per locus were not evaluated, owing to prohibitively slow computation time. PolyFun also requires fixed preprocessing time (steps 1–4; see Overview of methods) of 630 min on average; when restricting analyses to subsets of loci, PolyFun + SuSiE was still faster than all other functionally informed methods when analyzing >23 loci (Fig. 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 Figs. 1–5 and Supplementary Tables 4–6).

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

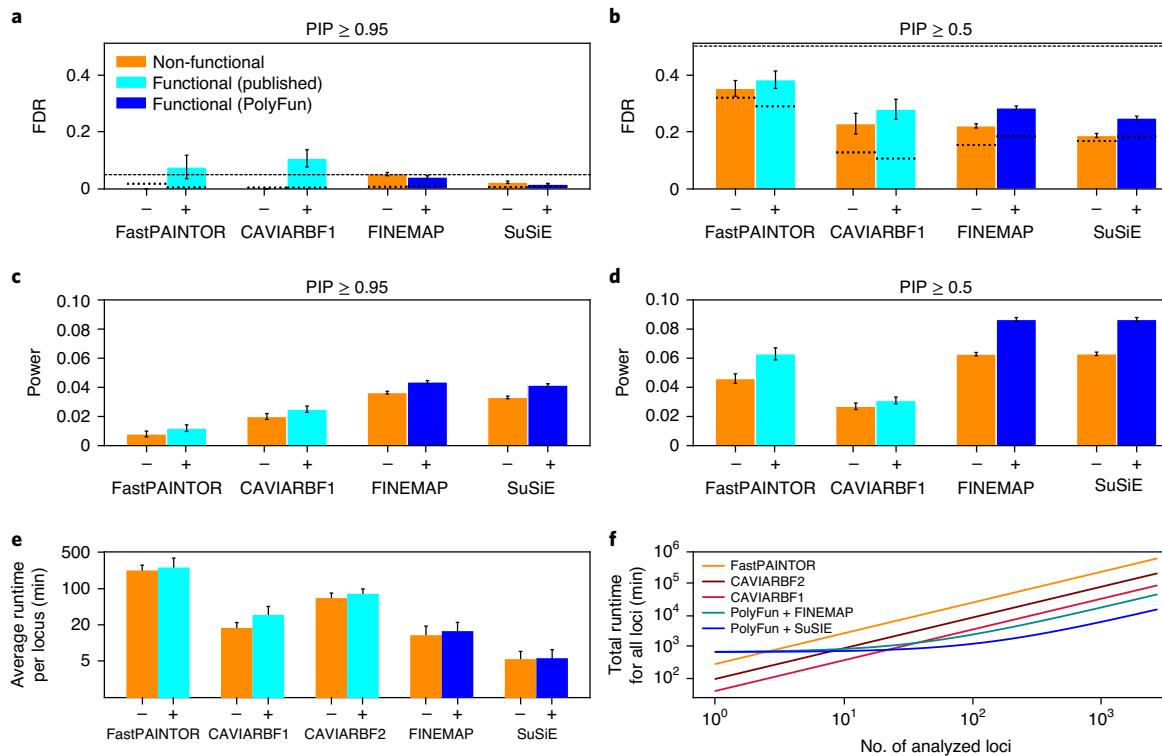


Fig. 1 | Calibration, power and computational cost of fine-mapping methods in main simulations. a,b, FDR at $\text{PIP} = 0.95$ (a) and $\text{PIP} = 0.5$ (b). Upper dashed horizontal lines denote conservative FDR estimates. Lower dotted horizontal lines denote anti-conservative FDR estimates, which are not recommended (Supplementary Note). **c,d**, Power at $\text{PIP} = 0.95$ (c) and $\text{PIP} = 0.5$ (d). The first bar of each method uses nonfunctionally informed fine-mapping (denoted as $-$) and the second uses functionally informed fine-mapping (denoted as $+$). **e**, The average runtime required to fine-map a 3-Mb locus in a genome-wide analysis (log scale). The first bar of each method uses nonfunctionally informed fine-mapping (denoted as $-$) and the second uses functionally informed fine-mapping (denoted as $+$). **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 sublinear because they include the fixed preprocessing cost of computing prior causal probabilities (630 min). Error bars denote the s.e. Numerical results, including results for CAVIARBF2– and CAVIARBF2, and the results in **f** for nonfunctionally informed methods, are reported in Supplementary Table 4.

for British-ancestry UK Biobank samples as part of the present 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) nonoverlapping 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 2, the Supplementary Note and 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) which maintains $\text{FDR} < 0.05$ at a $\text{PIP} = 0.95$ 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 nonoverlapping LD reference panel from the target population spanning $\geq 10\%$ 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 use of 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

the 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 dealt with 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 $n = 318,000$; Supplementary Table 8). For each trait we fine-mapped up to 2,763 overlapping 3-Mb loci spanning 18,212,157 imputed SNPs with $\text{MAF} \geq 0.001$ and INFO score ≥ 0.6 (including short indels; excluding three long-range LD regions and loci with close to zero heritability; Methods). To each SNP we assigned 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 fine-mapped SNP-trait pairs with $\text{PIP} > 0.95$, a $>32\%$ improvement versus SuSiE; 9,684 SNP-trait pairs with $\text{PIP} > 0.5$, an improvement of $>59\%$ versus SuSiE; and 225,153 SNP-trait pairs with $\text{PIP} > 0.05$, an improvement of $>84\%$ versus SuSiE (Supplementary Table 9). The number of SNPs with $\text{PIP} > 0.95$ per trait ranged from 0 (number of children) to 407 (height) (Fig. 2a and Supplementary Table 9). The 3,025 SNP-trait pairs with $\text{PIP} > 0.95$ spanned 2,225 unique SNPs, including 532 low-frequency SNPs ($0.005 < \text{MAF} < 0.05$) and 185 rare SNPs ($0.001 < \text{MAF} < 0.005$) (Supplementary Table 10). Only 39% of the

Table 2 | Summary of mismatched reference LD simulations

Expt	GWAS	LD	Generative SNPs	SNPs analyzed	Max. L
a	44K UK	44K UK (44,000 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 (4,000 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 (22,000 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	UK10KnUKB	-
p	44K UK	3.6K UK10K	UK10KnUKB	UK10KnUKB	2
q	44K UK	3.6K UK10K	UK10KnUKBnINF	UK10KnUKBnINF	10
r	44K UK	3.6K UK10K	UK10KnUKBnCOM	UK10KnUKBnCOM	1
s	44K UK	4K UK	UK10KnUKB	UK10KnUKB	10

For each experiment (Expt) we report: GWAS: the sample size and population of the target sample (UK denotes British-ancestry individuals from the 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. Max. L: the maximum number of causal SNPs per locus assumed by PolyFun + SuSiE that maintain FDR <0.05 at a threshold of PIP=0.95 (selected from the options 1, 2, 3 and 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.

2,225 SNPs with PIP>0.95 were also lead GWAS SNPs (defined as SNPs with MAF>0.001 and $P<5\times 10^{-8}$ and no SNP with MAF>0.001 and a smaller P value within 1 Mb) (Supplementary Table 10), demonstrating the importance of using fine-mapped SNPs rather than lead GWAS SNPs for downstream analysis. Of the SNPs with PIP>0.95, 31% resided in coding regions and 22% were nonsynonymous (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 and 12) we identified 1,626 SNP-trait pairs with PIP>0.95 spanning 1,496 unique SNPs, with a median distance of 9 kb between an SNP with PIP>0.95 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 1-Mb windows around lead GWAS SNPs) harboring 0.28 SNPs with PIP>0.95 per locus on average (Supplementary Table 13); 1,080 of the 5,314 locus-trait pairs (20%) harbored ≥ 1 SNP(s) with PIP>0.95, harboring 1.37 SNPs with PIP>0.95 on average (Supplementary Table 13). Of the 1,626 SNP-trait pairs identified by PolyFun + SuSiE, 150 (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 fine-mapped SNPs with PIP>0.95 (which is probably close to the heritability causally explained by these SNPs, if most of the tagged SNP heritability originates from SNPs with PIP>0.95). The h_g^2 tagged by SNPs with PIP>0.95 captured a large proportion of the h_g^2 tagged by lead GWAS SNPs (median proportion = 42%; Fig. 2b; Methods and Supplementary Table 14). This proportion was substantially larger than the proportion of GWAS loci harboring SNPs with

PIP>0.95 (20%; see above), because fine-mapping power is higher at loci with larger causal effects (Supplementary Table 4). However, fine-mapped SNPs tagged a smaller proportion of the total h_g^2 causally explained by all genome-wide SNPs with MAF>0.001 (median proportion = 19%; Fig. 2b; Methods and Supplementary Table 14), indicating that substantially larger sample sizes are required to comprehensively fine-map all heritable SNP effects.

Among the 2,225 unique SNPs with PIP>0.95 fine-mapped for at least one trait, 223 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 Fig. 6 and Supplementary Table 15); 118 pleiotropic SNPs resided in coding regions and 93 were nonsynonymous (Supplementary Table 15). The 17 SNPs fine-mapped for at least 4 traits are reported in Table 3. 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 1 Mb) (Fig. 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 nonfunctionally informed SuSiE to avoid biasing the results). For each of 50 main

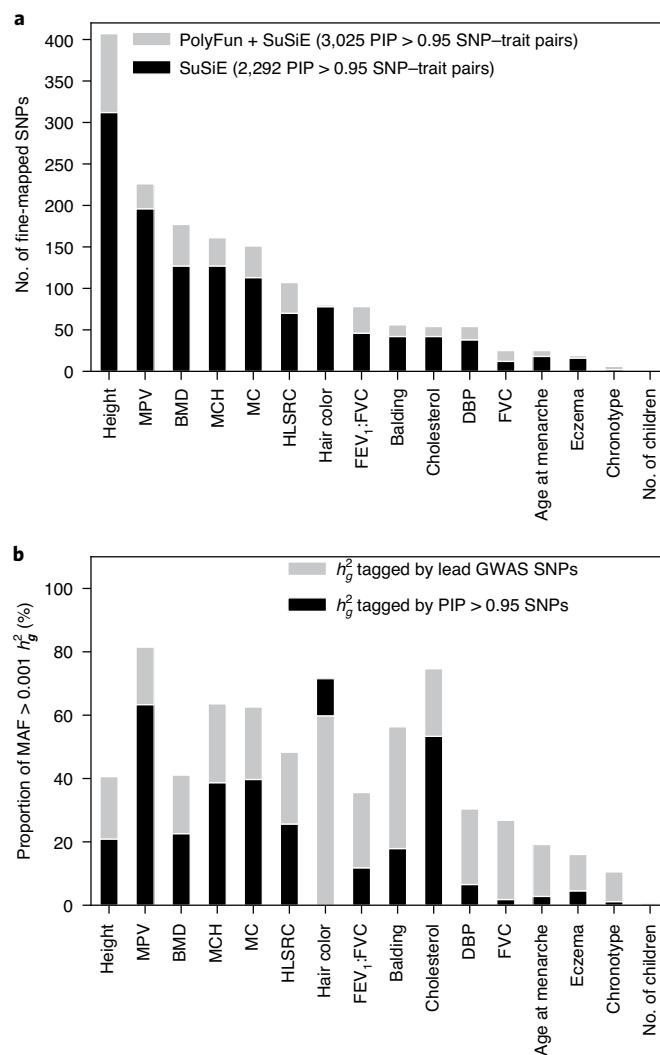


Fig. 2 | Summary of fine-mapping results for UK Biobank traits. **a**, The number of SNPs with PIP > 0.95 identified by SuSiE (black bars) and PolyFun + SuSiE (gray bars) across 16 genetically uncorrelated traits in the UK Biobank. Traits are ordered by PolyFun + SuSiE results. The numbers in the legend refer to the sum of all 49 traits analyzed. **b**, The proportion of SNP heritability causally explained by all genome-wide SNPs with MAF > 0.001 (h_g^2) that is tagged by lead GWAS SNPs (gray bars) and by PolyFun + SuSiE SNPs with PIP > 0.95 (black bars). Traits are ordered as in **a**. For hair color, the h_g^2 tagged by SNPs with PIP > 0.95 is $> h_g^2$ tagged by lead GWAS SNPs. BMD, bone mineral density; DBP, diastolic blood pressure; FEV₁:FVC, ratio of forced expiratory volume in 1 s to forced vital capacity; HLSRC, high light scatter reticulocyte count; MC, monocyte count; MCH: mean corpuscular hemoglobin; MPV, mean platelet volume. Numerical results are reported in Supplementary Tables 9 and 14.

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; Fig. 4 and Supplementary Table 17). SNPs with PIP > 0.95 were strongly and significantly enriched for nonsynonymous SNPs (51× enrichment, $P=6.8 \times 10^{-185}$) and SNPs in conserved regions (16× enrichment, $P<10^{-300}$), significantly enriched for SNPs in various regulatory annotations (for example, 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 (Fig. 4 and Supplementary Table 17).

We compared our fine-mapping results with those of two previous studies. First, we compared our results with ref. ¹², which performed nonfunctionally informed fine-mapping for 9 blood cell traits using approximately 115,000 of the individuals included in our analyses. PolyFun + SuSiE identified 4.4× more SNPs than ref. ¹², including all 4 SNPs that were functionally validated via luciferase reporter assays in ref. ¹² (PIP > 0.999 for all 4 SNPs; Methods and Supplementary Tables 18–20). Second, we compared our results with those of ref. ⁷, which performed nonfunctionally informed fine-mapping for seven of our traits, using a nonfunctionally informed method (PICS) and independent smaller datasets. PolyFun + SuSiE identified 35× more SNPs than ref. ⁷ (Supplementary Tables 21 and 22). Further details of the comparison are provided in the Supplementary Note.

We performed a further six secondary analyses, described in the Supplementary Note, Extended Data Figs. 7–9, and 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. SNPs with PIP > 0.95 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%) (Fig. 2b), implying that they causally explain a small proportion of h_g^2 . We thus propose polygenic localization, which has the aim of identifying 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 (that is, the posterior mean of their squared effect size; Methods), we define $M_{50\%}$ as the size of the smallest set of top-ranked common SNPs causally explaining 50% of common SNP heritability (respectively M_p for proportion p of common SNP heritability). We estimate $M_{50\%}$ (respectively 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% (respectively 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 it 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 for $M_{50\%}^*$ (Supplementary Note and Supplementary Tables 29 and 30). Further details of PolyLoc are provided in Methods; 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 $n=337,000$ unrelated British-ancestry samples (steps 1 and 2)

Table 3 | Pleiotropic fine-mapped SNPs for UK Biobank traits

SNP	Position	MAF	Closest gene(s)	Annotation	Traits
rs13107325	chr4:103188709	0.08	SLC39A8	Nonsynonymous	BMI, balding, cholesterol, DBP, FVC, height, RBC, WHR (8)
rs1229984	chr4:100239319	0.02	ADH1B	Nonsynonymous	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	Nonsynonymous	HDL, height, MC, MPV (4)
rs3811444	chr1:248039451	0.33	TRIM58	Nonsynonymous	HLSRC, HbA1c, platelet count, RBC (4)
rs1260326	chr2:27730940	0.39	GCKR	Nonsynonymous	Cholesterol, height, platelet count, RBCDW (4)
rs2270894	chr3:9975386	0.2	CRELD1, IL17RC	Conserved	BMD, FEV ₁ /FVC, height, platelet count (4)
rs11556924	chr7:129663496	0.39	ZC3HC1	Nonsynonymous	Age at 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	Nonsynonymous	Height, LC, RBC, RBCDW (4)
rs72681869	chr14:50655357	0.01	SOS2	Nonsynonymous	FVC, hair color, HbA1c, SBP (4)
rs61745086	chr16:88782050	0.01	PIEZ01, CTU2	Nonsynonymous	HLSRC, HbA1c, height, RBC (4)
rs34557412	chr17:16852187	0.01	TNFRSF13B	Nonsynonymous	HbA1c, MC, MPV, RBC (4)
rs77542162	chr17:67081278	0.02	ABCA6	Nonsynonymous	HbA1c, height, LDL, platelet count (4)

We report SNPs fine-mapped ($\text{PIP} > 0.95$) 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 number of fine-mapped traits). SNPs are ordered first by the number of fine-mapped traits and then by genomic position. BMI, body mass index; cardiovascular: cardiovascular-related disease; cholesterol: total cholesterol; DBP, diastolic blood pressure; HbA1c, glycated hemoglobin; HDL, HDL-cholesterol; HLSRC, high light scatter reticulocyte count; LC, lymphocyte count; MC, monocyte count; MCH, mean corpuscular hemoglobin; MPV, mean platelet volume; RBC, red blood cell count; RBCDW, red blood cell distribution width; SBP, systolic blood pressure; WHR, waist:hip ratio (adjusted for BMI). Results for all 223 pleiotropic fine-mapped SNPs are reported in Supplementary Table 15.

and re-estimated average per-SNP heritabilities in each of 59 SNP bins using S-LDSC applied to $n = 122,000$ European-ancestry UK Biobank samples that were not included in the $n = 337,000$ set to avoid winner's curse (step 3). Estimates of $M_{50\%}$ ranged widely from 28 (hair color) to 3,400 (height) to 2 million (number of children) (Fig. 5 and Supplementary Table 31). The median estimate of $M_{50\%}$ across 16 genetically uncorrelated traits was 8,900; 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 whereas number of children was the most polygenic trait, having an $M_{50\%}$ 3.7× larger than the second most polygenic of the 16 independent traits (chromotype, having an $M_{50\%} = 553,000$), consistent with ref. ³⁴. We performed seven secondary analyses, described in the Supplementary Note and Supplementary Tables 32 and 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 typically by thousands of SNPs (median $M_{50\%} = 8,900$), 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 initially 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 ($\text{PIP} > 0.95$), a 32% increase over nonfunctionally informed SuSiE. Of the fine-mapped SNPs, 223 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 (that is, using exactly the same samples in both the target and reference samples), assuming ten causal SNPs per locus; we have facilitated this option for UK Biobank researchers by publicly releasing summary LD information for $n = 337,000$ British-ancestry UK Biobank samples. As a second-best option we recommend applying PolyFun using an LD reference panel from the target sample population, spanning at least 10% of the target sample size, while assuming ten 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 $n = 337,000$ 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 SNPs with $\text{PIP} > 0.95$. We discuss possible reasons for identifying SNPs with $\text{PIP} > 0.95$ and $P > 5 \times 10^{-8}$ in

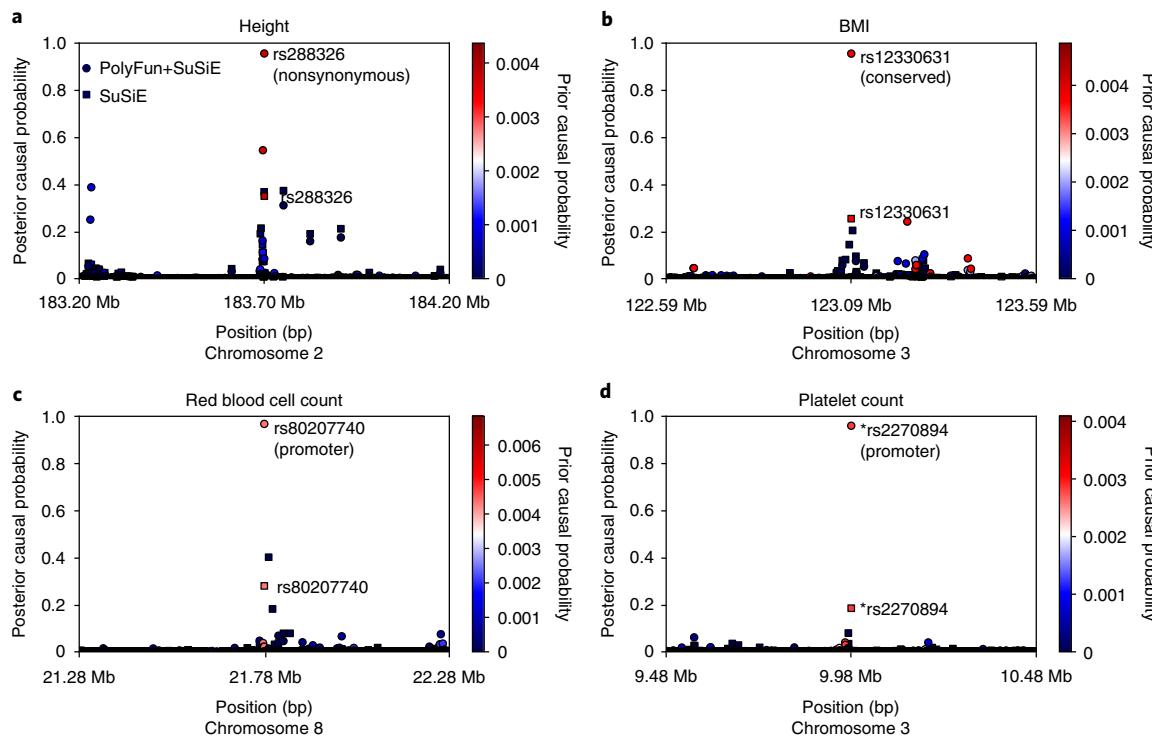


Fig. 3 | Examples of the advantages of functionally informed fine-mapping for UK Biobank traits. **a-d**, We report four examples where PolyFun + SuSiE identified a fine-mapped common SNP ($\text{PIP} > 0.95$) but SuSiE did not ($\text{PIP} < 0.5$ for all SNPs within 1 Mb) in four traits: height (**a**), BMI (**b**), red blood cell count (**c**) and platelet count (**d**). 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 SNPs with $\text{PIP} > 0.95$ (Methods) is reported in parentheses below its label. Asterisks denote lead GWAS SNPs. Numerical results are reported in Supplementary Table 16.

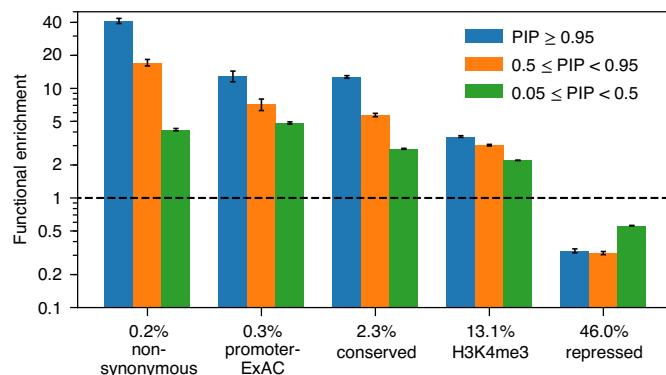


Fig. 4 | Functional enrichment of SuSiE fine-mapped common SNPs for UK Biobank traits. We report the functional enrichment of fine-mapped common SNPs (defined as the proportion of common SNPs with a PIP range lying in an annotation divided by the proportion of genome-wide common SNPs lying in the annotation) for 5 selected binary annotations, meta-analyzed across 14 genetically uncorrelated UK Biobank traits with ≥ 10 SNPs with $\text{PIP} > 0.95$ (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 the s.e. Numerical results for all 50 main binary annotations and all traits are reported in Supplementary Table 17.

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

(for example, 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 identified could be prioritized for functional follow-up. Second, fine-mapping results (posterior mean effect sizes) can be used to compute transethnic 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 noncoding 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 estimates of the FDR of $\text{PIP} > 0.95$ SNPs in PolyFun and 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 $n=337,000$ 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 include 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

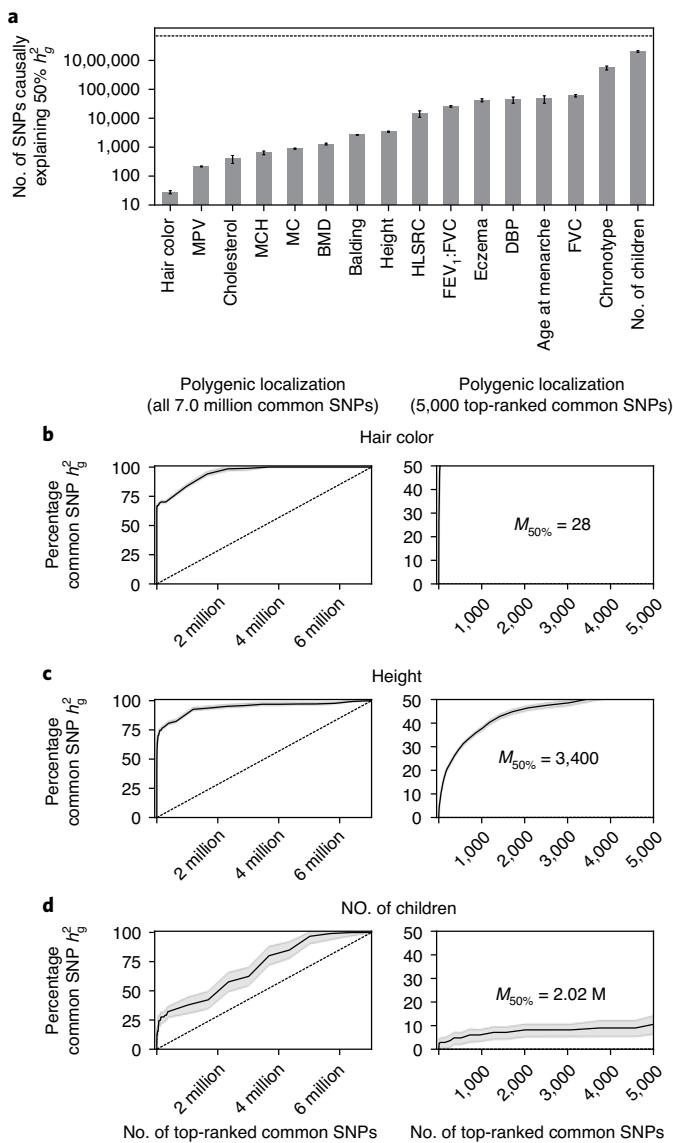


Fig. 5 | Polygenic localization results for UK Biobank traits. **a**, $M_{50\%}$ estimates across 16 genetically uncorrelated traits. For each trait, we report the number of top-ranked common SNPs (using PolyFun+SuSiE posterior per-SNP heritability estimates for ranking) causally explaining 50% of common SNP heritability, and its s.e. (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 hair color (**b**), height (**c**) and number of children (**d**) explained by different numbers of top-ranked SNPs, for all 7.0 million common SNPs (left) and the 5,000 top-ranked common SNPs (right). Gray shading denotes the s.e. 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 as $M_{50\%}$. Discontinuities in the slope indicate transitions between SNP bins. Numerical results for all 49 UK Biobank traits are reported in Supplementary Table 31.

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 provides only an upper bound on the proportion of SNPs causally explaining a given proportion of SNP heritability. Finally, multiethnic 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 content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-020-00735-5>.

Received: 28 October 2019; Accepted: 2 October 2020;
Published online: 16 November 2020

References

- Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic medicine—progress, pitfalls, and promise. *Cell* **177**, 45–57 (2019).
- 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).
- 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).
- Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
- Chen, W. et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).
- Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
- Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
- Mahajan, A. et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat. Genet.* **50**, 559–571 (2018).
- 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).
- 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).
- Ulrirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
- The Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
- Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* **10**, e1004722 (2014).
- Kichaev, G. et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).
- Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J. Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics* **204**, 933–958 (2016).
- 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* <https://doi.org/10.1111/rssb.12388> (2020).
- Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
- Benner, C., Havulinna, A., Salomaa, V., Ripatti, S. & Pirinen, M. Refining fine-mapping: effect sizes and regional heritability. Preprint at *bioRxiv* <https://doi.org/10.1101/318618> (2018).
- Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
- 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).