Supplementary Information for: Functionally-informed fine-mapping and polygenic localization of complex trait heritability

Supplementary Table Legends

Table S1: Description of baseline-LF model annotations. For each annotation we report #SNPs (unless it is a continuous-valued annotation), #common SNPs, whether it is binary or continuous-valued, and a literature reference.

Table S2: Description of 10 3Mb loci used in simulations. We report the start and end position of each locus (using hg19 coordinates), as well as the number of SNPs in that locus.

Table S3: Functional enrichments used to generate simulated data. For each annotation, we report the coefficient τ used in the generative model, as computed via a meta-analysis across 16 real traits.

Table S4: Numerical results of fine-mapping simulations. Each row reports simulation results for a unique combination of a method, method settings, simulation settings and a PIP cutoff. The columns are (**A**) Method: Method name; (**B**) Functional: the type of functional data used (None: non-functionally informed analysis; PolyFun: PolyFun; default: default settings for CAVIARBF and fastPAINTOR; S-LDSC priors: priors from standard S-LDSC; S-LDSC+L2 priors: Priors from L2-regularized S-LDSC; cheat: fine-mapping with true prior causal probabilities; PolyFun (no L2): PolyFun that does not apply L2-regularization); (**C**) PIP cutoff: The PIP cutoff evaluated; (**D**) N: simulated sample size; (**E**) h2: The SNP-heritability causally explained by SNPs at the target locus; (**F**) c: number of simulated causal SNPs at the target locus; (**G**) functional architecture: the function that relates functional annotations to prior causal probabilities (additive, multiplicative, or sub-additive); (**H**) q: the ratio between the highest and lowest prior causal probability; (**I**) Max #causal SNPs: The maximum number of causal SNPs allowed (or the exact number for SuSiE and PolyFun + SuSiE); (**J**) Genome-wide h2: the SNP-heritability causally explained by all genome-wide SNPs; (**K**) Genome-wide sparsity: The proportion of genome-wide non-causal SNPs; (**L**) Max #annotations: The maximum number of annotations allowed (only relevant for fastPAINTOR); (**M**) variance estimator: determines whether the (method-specific) default estimator or the HESS-based estimator of causal effect size variance was used for fine-mapping; (**N**) FDR: false discovery rate under the evaluated PIP cutoff; (**O**) FDR (s.e.): the standard error of the FDR (evaluated via jackknife); (**P**) power: empirical power under the evaluated PIP cutoff; (**Q**) power (s.e.): the standard error of the power (evaluated via jackknife); (**R**) #experiments: The number of experiments performed for this combination of settings; (**S**) Avg. #fine-mapped SNPs: The average number of SNPs with PIP greater than the cutoff; (**T**) FDR threshold: the exact FDR threshold, given by one minus the average PIP among SNPs with PIP greater than the cutoff; (**U**) Time (avg): The average time required to perform fine-mapping (minutes); (**V**) Time (s.d.): The standard deviation of time required to perform fine-mapping (minutes); (**W**) 95% CS size: The size of the union of all 95% credible sets, averaged across simulations (only relevant for FINEMAP and SuSiE-based methods); (**X**) %Causal SNPs in 95% CS: The proportion of causal SNPs found in the union of all 95% credible sets (averaged across simulations).

Table S5: Description of 47 UK Biobank traits analyzed. For each trait we report its group (used in Figure 4); whether it is binary or continuous-valued; sample size analyzed by PolyFun (max N=337K unrelated British-ancestry samples) and PolyLoc (max N=122K non-overlapping samples); the number of cases in the PolyFun and PolyLoc sample sets (for binary traits); and the heritability causally explained by MAF>0.001 SNPs (using observed-scale heritability for binary traits), and its standard error, in the PolyFun dataset (max N=337K) and in the PolyLoc dataset (max N=122K).

Table S6: Comparison of PolyFun + SuSiE vs. SuSiE fine-mapping results for UK Biobank traits. We report the number of fine-mapped SNPs found by PolyFun + SuSiE and by SuSiE for various trait and PIP cutoffs. For every trait and one of three evaluated PIP cutoffs (0.05 0.5, 0.95) we report the number of SNPs with PIP greater than this cutoff identified by PolyFun + SuSiE, SuSiE, and by both (i.e. SNPs at the intersection of these two SNP sets).

Table S7: List of all fine-mapped SNPs for UK Biobank traits. We report the list of all SNPs with PIP>0.95 in either PolyFun + SuSiE or SuSiE. For every SNP we report its rsid, position (using hg19 coordinates), major (A1) and minor (A2) alleles, PIP, posterior per-SNP h^2 (i.e., the sum of the squared posterior mean and posterior variance of its causal effect size), the posterior mean and standard deviation (sd) of its cal effect size, the index of one of the 95% credible sets to which it belongs under each of the two methods, its MAF in the entire UK Biobank and in the N=337K dataset we analyzed, its distance to the nearest lead GWAS SNP, the center of the locus in which it was analyzed, its BOLT-LMM p-value, its BOLT-LMM marginal effect size estimate (BETA), its prior causal probability, and whether it is coding and non-synonymous.

Table S8: Genetic correlations between UK Biobank traits. We report a matrix of r_g estimates, computed using N=337K UK Biobank individuals (see Methods).

Table S9: List of 15 genetically uncorrelated UK Biobank traits. We report the 15 genetically uncorrelated traits selected for many of our primary analyses. All the traits have pairwise $|r_g|$<0.2, h_g^2>0.05 in the two different subsets of the UK Biobank, and an effective sample size >100K in the N=337K data set (see Methods).

Table S10: Distribution of PIP>0.95 SNPs across GWAS loci for UK Biobank traits. For each lead GWAS SNP, we report the number of PIP>0.95 SNPs within distance of at most 0.5Mb from the lead SNP.

Table S11: Numerical values of heritability tagged by PIP>0.95, heritability tagged by fine-mapped SNPs, and heritability causally explained by MAF>0.001 SNPs for UK Biobank traits. For each of the 15 genetically uncorrelated traits we report the heritability causally explained by MAF>0.001 SNPs (as estimated via S-LDSC), the heritability tagged by PIP>0.95 SNPs (estimated via the adjusted R^2 of a multivariate linear regression) and the heritability tagged by lead GWAS SNPs (estimated via the adjusted R^2 of a multivariate linear regression), using observed-scale heritability for binary traits. All analyses were performed using a set of (max N=122K) individuals not analyzed by PolyFun + SuSiE, to prevent winner's curse. We also report the corresponding estimates obtained via a multivariate linear regression using a subset of N=45K unrelated individuals, which were nearly identical to the N=122K results.

Table S12: List of all pleiotropic fine-mapped SNPs (PIP>0.95) identified by PolyFun + SuSiE for UK Biobank traits. The table is analogous to Table 2 but includes results for all 223 PIP>0.95 pleiotropic SNPs. The nearest gene column was specified via (1) the GWAS catalog "mapped gene" entry if the SNP

appeared in the GWAS catalog; (2) the dbSNP gene entry for the SNP, if such an entry exists; or (3) a manual inspection in the UCSC genome browser. The nearest gene column for intergenic SNPs sometimes includes two genes separated by a dash, representing the two closest flanking genes.

Table S13: Examples of the advantages of functionally-informed fine-mapping for UK Biobank traits. For all 121 loci where PolyFun + SuSiE identified a PIP>0.95 SNP and SuSiE did not identify a single PIP>0.5 SNP within 500kb of that SNP, we report the prior causal probability and annotations of the fine-mapped SNP and of the top SuSiE SNP in this locus.

Table S14: Numerical values of functional enrichment of SuSiE fine-mapped common SNPs for UK Biobank traits. For each main functional annotation (Methods), PIP range, and trait, we report enrichment, its standard error, and its p-value. We report meta-analysis results in separate rows with the trait name "Meta-analysis".

Table S15: Summary of comparison between our results and those of Ulirsch *et al.* 2019 Nat Genet. The table reports the number of SNPs fine-mapped (PIP>0.95) by Ulirsch *et al.* for 9 blood cell traits, the number of SNPs fine-mapped (PIP>0.95) by one of our evaluated methods/datasets (a different method/dataset in every row) and the size of the intersection of the two sets.

Table S16: Detailed comparison between our results and those of Ulirsch *et al.* 2019 Nat Genet. The table includes the identities and PIPs of each SNP having PIP>0.95 in at least one of the evaluated methods.

Table S17: Detailed comparison between our results and those of Ulirsch *et al.* 2019 Nat Genet for four SNPs that were functionally validated via luciferase reporter assays. The table includes a subset of Table S16 for only the four SNPs of interest.

Table S18: Summary of comparison between our results and those of Farh et al. 2015 Nature. The table is analogous to Table S15 but compares our results to Farh *et al.* rather than Ulirsch *et al.*

Table S19: Detailed comparison between our results and those of Farh et al. 2015 Nature. The table is analogous to Table S16 but compares our results to Farh *et al.* rather than Ulirsch *et al.*

Table S20: Comparison of PolyFun + SuSiE vs. SuSiE results for UK Biobank traits when down-sampling the data to N=107K individuals from the UK Biobank interim release. For each of five selected traits (Methods) and all combinations of pairs of analyses (SuSiE or PolyFun + SuSiE, using either N=337K or N=107K subsets of the UK Biobank), we report the number PIP>0.95 SNPs identified by each method, and the number of PIP>0.95 SNPs identified by both methods.

Table S21: Comparison of PolyFun + SuSiE vs. SuSiE 95% credible set sizes across all locus-trait pairs for UK Biobank traits. We report the sizes of 95% credible sets for SuSiE and PolyFun + SuSiE for each locus-trait pair analyzed.

Table S22: List of pairs of coding and non-coding fine-mapped SNPs within 1Mb of each other. We report all pairs of fine-mapped coding and non-coding SNPs within 1Mb of each other, which could aid in linking regulatory variants to genes.

Table S23: Numerical values of functional enrichment of PolyFun + SuSiE fine-mapped common SNPs for UK Biobank traits. The table is analogous to Table S14 but uses PolyFun + SuSiE instead of SuSiE to compute PIPs.

Table S24: Numerical values of functional enrichment of SuSiE fine-mapped MAF>0.001 SNPs for UK Biobank traits. The table is analogous to Table S14 but uses MAF>0.001 SNPs instead of only common (MAF>0.05) SNPs.

Table S25: Numerical values of functional enrichment of SuSiE fine-mapped low-frequency and rare SNPs for UK Biobank traits. The table is analogous to Table S14 but uses 0.05>MAF>0.001 SNPs instead of common (MAF>0.05) SNPs.

Table S26: Numerical polygenic localization results for 47 UK Biobank traits. For each trait we report M_p estimates (and their standard errors) at multiple values of p.

Table S27: Comparison between polygenic localization and M_e estimates for 13 UK Biobank traits. We report M_e and $M_{50\%}$ estimates for 13 genetically uncorrelated traits for which there exist published M_e estimates (from O'Connor et al. 2019).

Table S28: Polygenic localization results for secondary analyses of UK Biobank traits. The table reports the results of all the secondary analyses described in the polygenic localization section of the paper. The analysis type is specified in the "Analysis" column.

Table S29: Results of polygenic localization simulations. Each row contains results for a specific combination of settings: u – simulated power (u=1 means perfect power and u=0 means no power; see Methods); h2 – simulated SNP-heritability; p – simulated proportion of causal SNPs; n – simulated sample size for the PolyFun dataset; num_bins – the number of bins used in the simulation; logM*p_mean - the number of SNPs causally explaining proportion *p* of SNP-heritability under the optimal ranking, averaged across 10 simulations; logM*p_sd – the standard deviation of logM*p across 10 simulations; logMp_mean – the average estimate of logMp across 10 simulations; logMp_sd – the standard deviation of estimates of logMp under 10 simulations; logMp_diff – the average difference between the estimates and true values of logMp under 10 simulations (positive values indicate conservative estimates); logMp_sd – the standard deviation of the difference between the estimates and true values of logMp across 10 simulations. All the logarithm in this table are log10 based. We note that logMp_diff is evaluated with respect to the true value of logMp rather than logM*p (see Methods).

Table S30: Results of polygenic localization simulations using ranking based on magnitude of summary statistics. The table is analogous to Table S29 but uses a ranking of SNPs based solely on the magnitude of their summary statistics.

Supplementary Figures

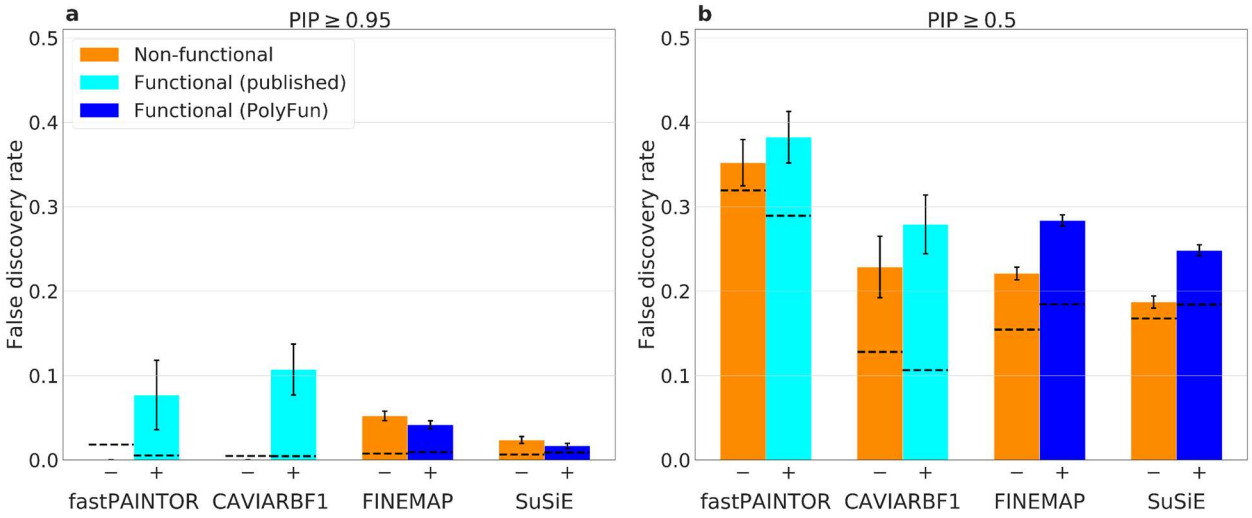

Figure S1: **False discovery rates in simulations using exact false-discovery thresholds.** The figure is the same as Figure 1a-b, except that the false-discovery thresholds (horizontal dashed lines) are lower than before.

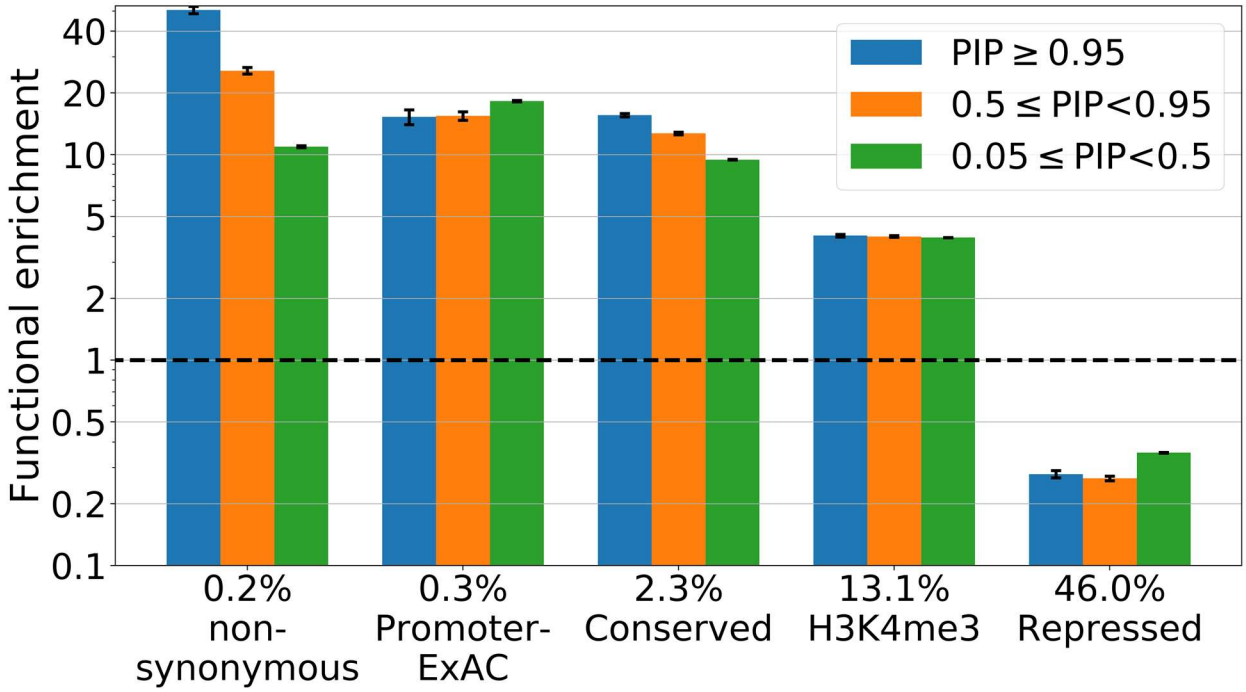

Figure S2: Functional enrichment of PolyFun + SuSiE fine-mapped common SNPs for UK Biobank traits. The figure is analogous to Figure 6 but uses PIPs computed by PolyFun + SuSiE instead of SuSiE.

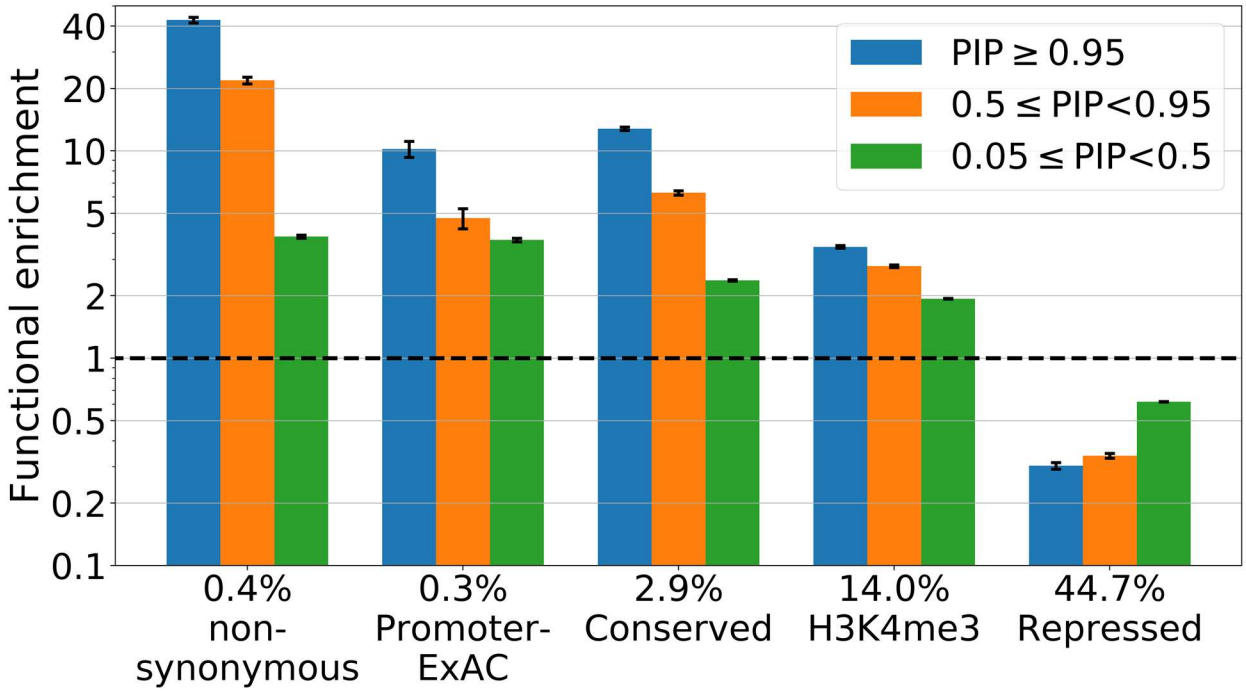

Figure S3: Functional enrichment of SuSiE fine-mapped MAF>0.001 SNPs for UK Biobank traits. The figure is analogous to Figure 6 but uses MAF>0.001 SNPs instead of common (MAF>0.05) SNPs.

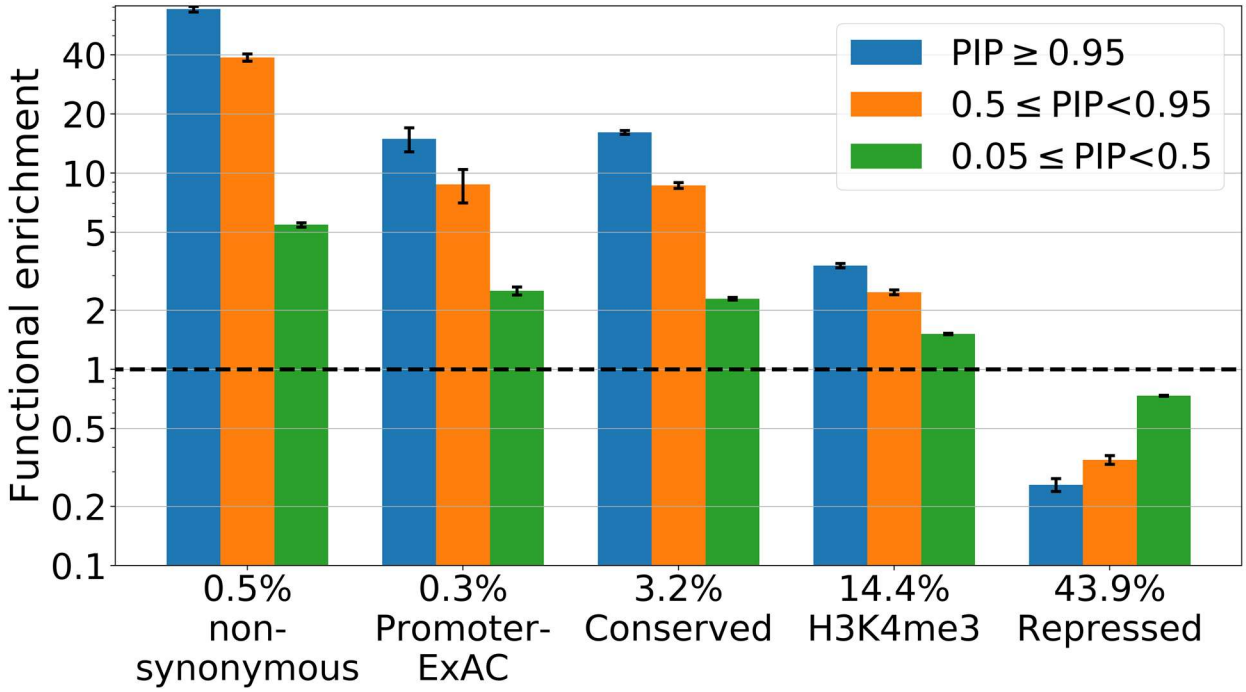

Figure S4: Functional enrichment of SuSiE fine-mapped low-frequency and rare SNPs for UK Biobank traits. The figure is analogous to Figure 6 but uses only low-frequency and rare SNPs (0.05>MAF>0.001) instead of common (MAF>0.05) SNPs.