nature portfolio

https://doi.org/10.1038/s41588-022-01087-y

Supplementary information

Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity

Supplementary material for

Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity

Steven Gazal[1,2,3,4*], Omer Weissbrod[3,4], Farhad Hormozdiari[3,4], Kushal Dey[3,4], Joseph Nasser[4], Karthik Jagadeesh[3,4], Daniel Weiner[4], Huwenbo Shi [3,4], Charles Fulco[4,5,6], Luke O'Connor[4], Bogdan Pasaniuc[7], Jesse M. Engreitz[4,8,9], Alkes L. Price[3,4,10*]

1. Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
2. Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
3. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
4. Broad Institute of MIT and Harvard, Cambridge, MA, USA
5. Department of Systems Biology, Harvard Medical School, Boston, MA, USA
6. Present address: Bristol Myers Squibb, Cambridge, MA, USA
7. Departments of Computational Medicine, Human Genetics, Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA
8. Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA
9. BASE Initiative, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford University School of Medicine, Stanford, CA, USA
10. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Correspondence should be addressed to S.G. (gazal@usc.edu) or A.L.P. (aprice@hsph.harvard.edu).

Supplementary Note

Description of the 50 SNP-to-gene (S2G) strategies used in this study

We considered 50 S2G strategies; in each case we first considered raw linking values $A_{k,j,g}$, which we next converted into linking scores $\psi_{k,j,g}$:

Exon: for each gene, the list of each exons was extracted from the GENCODE database[1]. We added 20bp flanking windows between the exons to include splice regions.

Promoter: for each gene, we selected the list of TSSs of different transcripts from the Ensembl dataset, added a +/-1kb regions around them, intersected with promoter annotations from the baseline model[2] (including promoters from refs.[3–5]) and from Roadmap[6], and removed regions overlapping exons or splice regions.

Gene body: for each gene, we selected the minimum starting position and maximum ending position across Ensembl, GENCODE, and RefSeq databases.

Gene±100kb: for each gene, we added a +/- 100kb window around its gene body.

Closest TSS and Closest *i^th^* TSS (20 strategies): for each SNP, we defined its closest, second closest, and up to 20th closest TSS based on physical distance with the TSSs of different transcripts from the Ensembl dataset.

Distance constrained closest TSS (7 strategies): we only linked SNPs for which closest TSS was less than 1kb away, between 1kb and 5kb away, between 5kb and 10kb away, between 10kb and 50kb away, between 50kb and 100kb away, between 100kb and 500kb away, and between 500kb and 1,000kb away.

GTEx *cis*-eQTL: we used GTEx v8 significant variant-gene associations for each of the 54 cell-types (17,382 samples in total), kept the minimum variant-gene association *P* value when a variant was link to a gene in multiple cell-types, and used -log10 of this *P* value as raw linking value.

GTEx blood/immune *cis*-eQTL: same as GTEx *cis*-eQTL, but by restricting eQTLs from three blood/immune cell-types (i.e. whole blood, spleen and EBV-transformed lymphocytes).

GTEx fine-mapped *cis*-eQTL: we fine-mapped the GTEx v8 *cis*-eQTLs of each gene in each tissue as in ref.[7], selected the SNPs with a causal posterior probability (CPP) ≥0.05, kept the maximum variant-gene CPP when a variant was fine-mapped to a gene in multiple tissues, and used corresponding CPP as raw linking value.

GTEx blood/immune fine-mapped *cis*-eQTL: same as GTEx fine-mapped *cis*-eQTL, but by restricting fine-mapped cis-eQTL from three blood/immune cell-types.

eQTLGen blood *cis*-eQTLs: we used eQTLGen statistically significant *cis*-eQTLs in blood (31,684 individuals), and used corresponding -log10 *P* value as raw linking value.

eQTLGen fine-mapped blood *cis*-eQTL: we fine-mapped the eQTLGen *cis*-eQTLs of each gene in each tissue as in ref.[7], selected the SNPs with a CPP ≥0.05, and used corresponding CPP as raw linking value.

Roadmap enhancer-gene linking: we used linked Roadmap enhancers[4,6,8] based on expression-enhancer activity correlation across 127 cell-types, kept the maximum correlation when an enhancer was linked to a gene in multiple tissues, and used corresponding correlation as raw linking value.

Roadmap blood/immune enhancer-gene linking: same as Roadmap enhancer-gene linking, but by restricting enhancers from 27 blood/immune cell-types.

EpiMap enhancer-gene linking: we used linked EpiMap enhancers[4,9] based on expression-enhancer activity correlation across 833 cell-types, kept the maximum correlation when an enhancer was linked to a gene in multiple tissues, and used this correlation as raw linking value.

EpiMap blood/immune enhancer-gene linking: same as EpiMap enhancer-gene linking, but by restricting enhancers from 85 blood/immune cell-types.

Activity-by-Contact (ABC): we used ABC links from ref.[10], kept the maximum ABC score across the 167 cell-types when an enhancer was interacting with a gene in multiple cell-types, and used this score as raw linking value.

ABC blood/immune: same as ABC, but by restricting elements from 69 cell-types with blood/immune tissue/cells.

Closest TSS (Hi-C): for each SNP, we defined TSS with the highest Hi-C intensity (averaged across 10 cell-types[11]) with the TSSs of each gene as in ref.[11]. We labeled this strategy closest TSS (Hi-C) as it gave similar results to the closest TSS strategy based on physical distance (Supplementary Table 5).

Hi-C distance: we linked each SNP to its surrounding genes using Hi-C intensity (averaged across 10 cell-types[11]) with the TSS of each gene, as in ref.[11].

Jung PCHi-C: we used PCHi-C links from ref.[12], kept the minimum interaction *P* value across the 27 cell-types when a genomic region was interacting with a gene in multiple cell-types, and used corresponding -log10 *P* value as raw linking value.

Javierre PCHi-C blood: we used PCHi-C links from ref.[13], kept the maximum CHiCAGO score across the 17 cell-types when a genomic region was interacting with a gene in multiple cell-types, and used corresponding CHiCAGO score as raw linking value.

Cicero blood/basal: we used the enhancer-promoter links from ref.[14] (data obtained through personal request to the authors), and used the enhancer-promoter correlation as raw linking value.

GeneHancer: we used the GeneHancer dataset[15] and used their enhancer scores as raw linking values.

Open Targets: we downloaded Open targets annotations, and weight each annotation as suggested in Open Targets Genetic website (i.e. 1.00 for variant effect predictor (VEP), 0.66 for expression and protein QTLs, and 0.33 for distance to TSS, PCHi-C, and DHS-promoter and enhancer-TSS interactions); we used these values as raw linking values.

For Hi-C distance, we created $\psi_{k,j,g}$ linking scores proportional to Hi-C intensities (i.e. $\psi_{k,j,g} = A_{k,j,g} / \sum_g A_{k,j,g}$), unless the sum of the intensities across genes was below 0.2 (in that case $\psi_{k,j,g} = A_{k,j,g}/0.2$), in order to have $\sum_g \psi_{k,j,g} < 1$ for SNPs j in gene deserts.

Lines of evidence to validate our precision metric

We report three lines of evidence to further validate our precision metric. First, we verified that the estimated precision of the Closest TSS strategy (0.34 ± 0.03) is consistent with previous studies (0.34 in ref.[16], ~0.50 in ref.[17], 0.29 in ref.[18], 0.35 in ref.[19], and 0.27 ± 0.06 in ref.[20]), and that the estimated recall (h^2 coverage times precision) of the GTEx fine-mapped *cis*-eQTL strategy (0.13 ± 0.01) is consistent with the proportion of SNP-heritability mediated by gene expression in GTEx tissues estimated using a different approach[21] (0.11 ± 0.02). Second, we verified that estimates of precision (and hence recall) were similar when estimated using the (non-trait-specific) training critical gene set (used to optimize cS2G; see below) instead of the (trait-specific) validation critical gene sets (Supplementary Figure 3). Third, we verified that estimates of precision were similar for most S2G strategies when using an independent definition of precision (not relying on critical gene sets or polygenic analyses) based on two curated disease-associated lists of 577 linked sentinel SNP-gene pairs with the underlying genes validated with high confidence by Open Targets[22] and 1,668 linked fine-mapped SNP-gene pairs validated using nearby fine-mapped protein-coding variants[23] (see subsection below and Supplementary Figure 4). Despite the overall concordance, we observed large differences in precision estimates for some S2G strategies (e.g. Closest TSS), as the curated SNP-gene pairs were preferentially ascertained for disease-associated SNPs in which the target gene was the closest gene: indeed, we observed an unusually high proportion of SNP-gene pairs involving genes with a small distance (< 10kb) to its closest TSS (57% and 67% for the two curated lists, vs. h^2 coverage = 34% for the Closest TSS <10kb S2G strategy). Thus, we caution that curated disease-associated lists of linked SNP-gene pairs may be non-randomly ascertained, highlighting the potential benefits of polygenic analyses for evaluating S2G strategies.

Curated SNP-gene pairs for validation

We verified that our estimates of precision and recall (based on polygenic analyses of disease SNP-heritability) were similar to independent definitions (not relying on critical gene sets or polygenic analyses) based on two curated SNP-gene pairs. First, we considered the list of linked sentinel SNP-gene pairs with the underlying genes validated with high confidence by Open Targets[22] (see Data Availability). We selected SNP-gene-disease triplets defined with "high" confidence, selected sentinel SNPs with a minor allele count ≥5 in a 1000 Genomes Project European reference panel[24], selected genes in our list of 19,995 genes, and kept unique SNP-gene pairs, leading to a total of 577 pairs. We note that our number of pairs differs from the 445 pairs mentioned in ref.[22], despite having used their latest processed dataset (released on Jan 27, 2020). Second, we used linked fine-mapped SNP-gene pairs validated using nearby fine-mapped protein-coding variants[23]. After a similar control quality procedure, we retained 1,668 pairs. We estimated precision and recall by analyzing only the SNPs of the SNP-gene pairs, and by assuming that all SNP-gene pairs that are not on the list are false positives.

We note that the Open Targets dataset reports sentinel SNPs (rather than causal fine-mapped SNPs), which may be linked to different genes than causal SNPs, as we verified in analyses of UK Biobank traits (Supplementary Table 16). We also note that these two curated SNP-gene pairs were preferentially ascertained for disease-associated SNPs in which the target gene was the closest gene: indeed, we observed an unusually high proportion of SNP-gene pairs involving closest TSS genes with a small distance (< 10kb) to the closest TSS (57% and 67% for the two curated lists, vs. h^2 coverage = 34% for the Closest TSS <10kb S2G strategy). Thus, we caution that curated disease-associated lists of linked SNP-gene pairs may be non-randomly ascertained.

Additional experiments assessing the precision and recall of our cS2G strategy

We performed additional experiments assessing the precision and recall of our cS2G strategy over alternative approaches to combine S2G strategies. First, we constructed a new S2G strategy in which we linked all the SNPs linked by cS2G (22% of common SNPs) to the gene with closest TSS. This strategy attained only slightly lower precision and recall than the cS2G strategy (0.70 vs. 0.75 and 0.31 vs. 0.33, respectively; Supplementary Table 8); we note that it is possible for two cS2G strategies to have very different SNP-gene links but attain relatively similar performance, if both strategies have imperfect SNP-gene links. These results indicate that the most important difference between the cS2G and Closest TSS strategies derives from the set of SNPs linked to genes, rather than the linking strategy applied to those SNPs (Supplementary Table 8); evaluation of S2G strategies defined by linking other sets of functional SNPs to the gene with closest TSS is of potential interest. Second, we constructed a combined S2G strategy using only Exon and Promoter, which had the highest weights in our cS2G strategy (Supplementary Table 6). This Exon+Promoter strategy achieved a precision of 1.05 (s.e. 0.07) (statistically indistinguishable from the Exon (1.00, by definition) and Promoter (0.80, s.e. = 0.16) strategies) and recall of 0.16 (s.e. 0.01) (statistically indistinguishable from the sum of recalls attained by Exon (0.10) and Promoter (0.05) strategies), which is much lower than the recall of 0.33 (s.e. 0.03) attained by the cS2G strategy (Supplementary Table 5). Third, we expanded the set of S2G strategies provided as input when constructing the combined S2G strategy by adding the 3 non-functionally informed main S2G strategies from Table 1 (Gene body, Gene±100kb and Closest TSS), thus including all 13 main S2G strategies. The resulting combined strategy was identical to our primary cS2G combined strategy, indicating that Closest TSS provides no additional information. Fourth, we expanded the set of S2G strategies to include all 50 S2G strategies (Supplementary Table 1), not just the 13 main S2G strategies. The resulting combined strategy included 8 S2G strategies (Supplementary Table 9), attained higher recall (0.39, vs. 0.33 for cS2G), but lower precision (0.60 vs. 0.75). (We note that the estimated precision of this combined strategy using the training critical gene set was equal to 0.84 during the optimization process (Supplementary Table 8); the difference between 0.84 and 0.60 can be attributed to higher precision for some constituent S2G strategies in the training critical gene set compared to the validation critical gene sets (Supplementary Figure 3 and Supplementary Table 5). For this reason (and because a fundamental goal of cS2G is to provide functional interpretation of GWAS findings), we recommend the use of the cS2G strategy. Fifth, we evaluated the robustness of our results to overlap between the training gene set and the validation gene sets (Supplementary Figure 1) by removing all 1,760 genes in the training gene set from the set of all genes analyzed in the validation step. We estimated a precision of 0.80 (s.e. = 0.09) for cS2G, which is similar to (and actually slightly higher than) what we reported in our primary analysis (precision of 0.75, s.e. = 0.06), validating the robustness of our results. Sixth, to assess whether training and validating cS2G using different critical gene sets (but the same 63 independent traits) avoids overfitting, we randomly split the set of 63 independent traits in two, and performed training using one set of traits (using the training critical gene set) and validation using the other set of traits (using the validation critical gene sets). This procedure led to combined S2G strategies with high precision (≥0.76, vs. 0.75 for cS2G in the analysis of 63 traits) and recall (≥0.30, vs. 0.33) (Supplementary Table 10), confirming that our primary cS2G combined strategy avoids traits overfitting. Seventh, we constructed a combined S2G strategy by maximizing the recall without constraining precision to be ≥0.75. The resulting combined S2G strategy (including its weights, precision and recall) were unchanged; however, we still recommend constraining precision to be ≥0.75, as high precision is important for maximizing the utility of functional follow-up studies. Eighth, we repeated the experiment of maximizing the recall without constraining precision to be ≥0.75 for the relevant secondary analyses (numbered as “Third”, “Fourth”, and “Sixth” above). Of these, only the analysis numbered as “Third” yielded a different combined S2G strategy: the precision of 0.75 (s.e. 0.06) and recall of 0.33 (s.e. 0.03) changed to a precision of 0.34 (s.e. 0.03) and recall of 0.34 (s.e. 0.03). We consider this to be an unfavorable result, demonstrating the advantages of constraining the precision to be ≥0.75. Finally, we constructed a new combined strategy by optimizing the *F1 score* (the harmonic mean of precision and recall[25]) instead of optimizing the recall while constraining precision to be ≥0.75. Once again, this strategy greatly outperformed the individual strategies (now with respect to the F1 score) in the validation critical gene sets (Supplementary Table 11), confirming that our framework is robust to the choice of optimization metric.

Validating the combined S2G strategy using curated lists of disease-associated SNP-gene pairs

We sought to validate the combined S2G strategy using two curated lists of disease-associated SNP-gene pairs: a small curated list reflecting the very limited set of experimentally validated disease-associated SNP-gene pairs[26], and a larger curated list consisting of disease genes that have been validated with high confidence without strictly requiring experimental validation[22]. We restricted these analyses to SNPs that had a linked gene with cS2G linking score >0.5, consistent with our goal of attaining high precision for each individual SNP analyzed in order to maximize the utility of functional follow-up studies. We note that restricting cS2G to these links led to similar estimates of aggregate precision and recall in polygenic analyses (Supplementary Table 8).

First, we manually curated a list of 17 disease-associated loci (including 12 loci from ref.[26]) containing 25 experimentally validated causal SNP-gene pairs (Table 2). 16 of the 25 pairs had a linked gene with cS2G linking score >0.5. The cS2G prediction of the target gene matched the experimentally validated gene for 11 of these 16 loci, yielding a precision of 11/16 = 0.69 (s.e. = 0.12) and a recall of 11/25 = 0.44 (s.e. = 0.10) (Table 2 and Supplementary Table 12). The precision was lower than our estimate based on validation critical gene sets (0.75) (and lower than the precision of one constituent strategy; 0.78 (s.e. = 0.14) for EpiMap), whereas the recall was higher than our estimate based on validation critical gene sets (0.33) (and higher than the recall of any constituent strategy); however, these differences were not statistically significant due to the small number of experimentally validated SNP-gene pairs (Supplementary Table 13). Interestingly, of the 11 pairs that were correctly linked to the experimentally validated gene, 8 pairs were linked by at least two cS2G constituent strategies, including rs339331-*RFX6* (Prostate cancer[27,28]; EpiMap and ABC), rs356168-*SNCA* (Parkinson disease[29]; EpiMap and ABC), rs11257655-*CAMK1D* (Type 2 diabetes[30]; GTEx fine-mapped *cis*-eQTL and EpiMap), and rs61839660-*IL2RA* (Inflammatory bowel disease[31]; ABC and Cicero). However, we failed to identify the well-studied rs1421085-*IRX5/IRX3* link[32], as none of the constituent S2G strategies linked rs1421085 to either *IRX5* or *IRX3*; we also failed to identify the well-studied rs12740374-*SORT1* link[33], as rs12740374 is an exonic SNP for *CELSR2*, outweighing the link to *SORT1* by the GTEx fine-mapped *cis*-eQTL strategy.

Second, we analyzed a curated disease-associated list of 577 linked sentinel SNP-gene pairs (for ~300 diseases and complex traits, partially distinct from the set of 63 traits used to construct the cS2G strategy) with the underlying genes validated with high confidence by Open Targets[22] (see above). 356 of the 577 SNPs had a linked gene with cS2G linking score >0.5. The cS2G prediction of the target gene matched the Open Targets prediction for 205 of these 356 SNPs (precision = 58%, recall = 36%); these included SNP-gene links between the high-density lipoprotein (HDL) cholesterol sentinel SNP rs4983559 (ref.[34]) with *AKT1* (second closest TSS), and the depression sentinel SNP rs915057 (located in an intron of *SYN2*; ref.[35]) with *ESR2* (second closest TSS) (Supplementary Table 14). We observed similar precision when analyzing a different set of 1,668 linked sentinel SNP-gene pairs from Weeks et al.[23] (Supplementary Table 15). We confirmed the robustness of the precision of cS2G for the Open Targets SNP-gene pairs (and the SNP-gene pairs from Weeks et al.) to overlap of the underlying traits with the 63 traits used to construct the cS2G strategy, as analyses in which SNP-gene pairs for overlapping traits were removed produced similar results (Supplementary Table 15). The discrepancy between 58% and our estimated precision (0.75) can be explained by the fact that Open Targets reports sentinel SNPs (rather than causal fine-mapped SNPs), which may be linked to different genes than causal SNPs, as we verified in analyses of UK Biobank traits (Supplementary Table 16). We thus recommend to apply our cS2G strategy to confidently fine-mapped SNPs in preference to sentinel SNPs.

In summary, these analyses provide a promising validation of the potential of cS2G to pinpoint causal disease genes.

Leveraging the combined S2G strategy to pinpoint disease genes

We predicted target genes of 9,670 predicted causal SNP-disease pairs with PIP >0.5 from PolyFun analyses (7,675 unique SNPs). The SNP-gene-disease triplets predicted by cS2G included 2,163 triplets involving distal regulatory fine-mapped SNPs that were not in the gene body (or promoter) of the target gene, of which 532 were supported by at least 2 of the functionally informed constituent S2G strategies used by cS2G (Supplementary Table 17). We highlight 4 examples (Figure 4), discussed in this Supplementary Note.

First, for type 2 diabetes, two SNPs, rs234866 and rs74046911 ($r^2 = 0.02$ (but $D' = 1$)), both located in an intron of *KCNQ1* (initially reported as a candidate target gene[36,37]), were fine-mapped (PIP = 0.97 and 0.92, respectively) and both linked by cS2G to *CDKN1C* (third closest TSS; cS2G linking scores = 1.00 and 0.96, respectively) (Figure 4a). *CDKN1C* is a gene expressed in pancreas for which a rare coding mutation was previously linked to type 2 diabetes[38], and has been nominated as a candidate target gene at this locus using methylation data[39] and CRISPR-Cas9 genome editing[40]. *CDKN1C* was implicated by 3 of the functionally informed S2G strategies used by cS2G, including EpiMap enhancer-gene linking in endocrine pancreas (for both SNPs), identified by LDSC-SEG[41] as a critical tissue for type 2 diabetes (see Supplementary Table 15).

Second, for asthma, two independent SNPs, rs509399 and rs13099273 ($r^2 < 0.01$), were fine-mapped (PIP = 0.96 and 0.58, resp.) and linked by cS2G to the target gene *BCL6* (third and sixth closest TSS, resp.; cS2G linking scores = 0.96 and 1.00, resp.) (Figure 4b). *BCL6* modulates the response of interleukin 4, known to be involved in asthma[42], and has been linked to asthma in mice[43] and humans[44], but to our knowledge *BCL6* has not previously been implicated as an asthma gene using GWAS data. *BCL6* was implicated by 3 of the functionally informed S2G strategies used by cS2G, including EpiMap enhancer-gene linking in common myeloid progenitor CD34+ cells (for rs509399 only), identified by LDSC-SEG[41] as a critical cell-type for asthma ($P = 1.2 \times 10^{-4}$; see Methods; also see ref.[45]).

Third, for eczema, the SNP rs34290285 was fine-mapped (PIP = 0.99) and linked by cS2G to *PDCD1* (seventh closest TSS; cS2G linking score = 0.73) (Figure 4c). *PDCD1* is an immune-inhibitory receptor expressed in activated T cells, known to be implicated in Eczema. *PDCD1* has previously be linked to skin cancer[46] and autoimmune diseases[47], but to our knowledge *PDCD1* has not previously been implicated as an eczema gene using GWAS data. *PDCD1* was implicated by the 2 blood-informed S2G strategies used by cS2G (eQTLGen fine-mapped blood *cis*-eQTL and Cicero blood/basal). Two others functionally informed S2G strategies (GTEx fine-mapped *cis*-eQTL and EpiMap) linked rs34290285 to different genes (*GAL3ST2* and *D2HGDH*, respectively) with lower cS2G linking scores (0.21 and 0.05, respectively), highlighting the benefits of aggregating evidence from multiple S2G strategies to infer biological mechanisms.

Fourth, for HDL cholesterol, the SNP rs9604045 was fine-mapped (PIP = 0.99) and linked by cS2G to *LAMP1* (closest TSS, although *CUL4A* is the closest gene; cS2G linking score = 0.97) (Figure 4d). While deficiency of lysosome associated membrane proteins (*LAMP1* and *LAMP2*) has been connected to cholesterol accumulation in mice[48,49], to our knowledge *LAMP1* has not previously been implicated as an HDL gene using GWAS data. *LAMP1* was implicated by 3 of the functionally informed S2G strategies used by cS2G (GTEx fine-mapped *cis*-eQTL, eQTLGen fine-mapped blood *cis*-eQTL, and ABC). However, none of the S2G strategies implicating *LAMP1* involved a plausible critical tissue/cell-type for HDL cholesterol (e.g. liver, identified by LDSC-SEG[41]; Supplementary Table 15), despite the availability of S2G links for liver in GTeX and ABC. This result highlights both the benefit of aggregating S2G links across multiple cell-types to infer SNP-gene pairs, and the challenge of identifying the causal cell-type of action.

We extended our analyses to 222,842 potentially causal SNP-disease pairs with PIP>0.05 (instead of PIP>0.50) from functionally informed fine-mapping of 49 UK Biobank diseases/traits[50,51]. Restricting to SNPs that had a linked gene with cS2G linking score >0.5, we predicted 138,716 potentially causal SNP-gene-disease triplets (99,847 unique SNPs, 15,820 unique genes) (see Data Availability). We also analyzed 170,346 SNP-disease pairs from the NHGRI-EBI GWAS catalog[52] (4,688 diseases/traits), with the caveat that these SNPs were not fine-mapped and have only a small probability of being causal. Restricting to SNPs that had a linked gene with cS2G linking score >0.5, we predicted 78,499 potentially causal SNP-gene-disease triplets (49,313 unique SNPs, 13,349 unique genes) (see Data Availability).

Substantial advance of our polygenic framework

Our framework, using polygenic analyses of disease SNP-heritability, is a substantial advance over previous approaches for evaluating S2G strategies using curated lists of disease-associated SNP-gene pairs[22,23]. In particular, curated lists of disease-associated SNP-gene pairs may contain SNPs whose causality has not been quantified (e.g. the Open Targets curated list[22] uses sentinel SNPs) and may contain ascertainment biases. Experimentally validated enhancer-gene pairs[11,53,54] provide an *in vitro* validation in a specific cell type for specific types of S2G strategies (i.e. enhancer-gene links), but this validation may not extend to *in vivo* disease contexts, which may involve cell types and cell states that are different from those assayed in validation experiments[11,21,53–55]; in particular, our genome-wide GWAS-based estimates of precision and recall for linking disease risk variants to disease genes are not comparable to the CRISPR-based estimates of precision and recall in K562 cells in Fig. 3 of ref.[11]. Furthermore, unlike previous approaches, our framework provides a route to optimally combining S2G strategies, greatly improving precision and recall relative to individual strategies (Figure 2) as well as previously proposed combined strategies (Supplementary Table 5).

Limitations of this work

We note several limitations of our work. First, our definition of precision assumes that the Exon strategy has a precision of 1, but this is an approximation as the link between exonic SNPs and target genes is likely to be imprecise in some cases (see rs12740374-*SORT1* example[33] in Table 2). Second, our estimates of precision had large standard errors for S2G strategies linking a limited fraction of SNPs to genes (Supplementary Figure 2 and Supplementary Table 5), such that evaluation of these S2G strategies was imprecise; however, for the cS2G strategy, estimates of precision (0.75, s.e. 0.06) and recall (0.33, s.e. 0.03) were reasonably precise. Third, we restricted each S2G strategy to the gene(s) with the highest linking score, as we observed that this led to slightly higher precision (Extended Data Figure 1). This does not reflect biological reality, in which a regulatory element may target more than one gene[9,10,56]; refinements to this choice are a direction for future research. Fourth, we included all available tissues and cell types for constituent S2G strategies of cS2G, as we observed that this led to higher precision (Extended Data Figure 3), perhaps due to limited biosample size. However, S2G links involving disease-critical tissues/cell-types are central to understanding biological mechanisms (Figure 4, Supplementary Table 20, Supplementary Note). As larger data sets become available, it may become practical to define disease-specific combined S2G strategies that restrict to disease-critical tissues and cell types, furthering the goal of pinpointing the causal cell-types of action of SNP-gene-disease triplets; these can be evaluated under our framework by meta-analyzing results of disease-specific combined S2G strategies across diseases/traits to obtain precise estimates of the precision and recall of a specific approach (Extended Data Figure 3). Fifth, our cS2G strategy uses a linear combination of S2G strategies with genome-wide weights, which may not optimize performance at a given locus. Exploring ways to optimally use locus-specific information is a promising direction of future research. Sixth, our cS2G strategy is derived from functionally informed S2G strategies that are primarily based on functional experiments in European-ancestry samples, and our evaluation of cS2G and its constituent strategies focused on European-ancestry GWAS, which are currently available in far greater sample sizes than GWAS from other populations[57,58]. Assessing the transferability of cS2G to non-European populations is a critical future research direction; we note that previous studies have generally reported high transferability of functional enrichments across populations[59–62]. Seventh, our results on disease omnigenicity are difficult to empirically verify; however, the very distinct patterns of disease omnigenicity inferred using the cS2G vs. Closest TSS strategies (Figure 5a) strongly support the use of functionally informed S2G strategies in such analyses. Finally, our analyses using cS2G (and its constituent S2G strategies[1,2,6,7,9–11,14,56,63,64]) pertain exclusively to SNP-gene pairs *in cis*, and do not capture *trans* effects, which may contribute substantially to disease omnigenicity[65]; in particular, this may explain why the disease SNP-heritability linked to genes in *cis* using the cS2G strategy (h^2_{gene}) represents only roughly half of total SNP-heritability (h^2).

Supplementary Table Legends

Supplementary Table 1: Description of the 50 SNP-to-gene (S2G) strategies.

Supplementary Table 2: Overlap proportions and correlations between 34 SNP-to-gene (S2G) strategies. (**a**) For each strategy (row), we report the proportion of SNP-gene links that overlap with another S2G strategy (column) (these values are not symmetric). (**b**) We report the correlation between the linking scores across 34 S2G strategies. We omitted 6th closest TSS to 20th closest strategy and Hi-C due to computational constraints. Correlations were computed on all SNP-gene links observed by at least one of the 34 SG strategies (i.e. most of the observed scores were 0 for the two S2G considered in the correlation computation). We observed low concordance between the S2G strategies: the average overlap fraction is 0.14 across the 34 S2G strategies, 0.19 across the 13 main strategies, and 0.08 across the 10 main functionally informed strategies; similarly, the average correlation is 0.10 across the 34 S2G strategies, 0.09 across the 13 main strategies, and 0.05 across the 10 main functionally informed strategies.

Supplementary Table 3: List of 63 diseases/traits used to estimate h^2 coverage, precision and recall. We defined a list of 63 summary statistics with independent association data (labeled as independent traits) by excluding genetically correlated traits in overlapping samples by measuring the intercept of cross-trait LD score regression[66], as previously described[2]; for traits with summary statistics computed from UK Biobank data, we also excluded traits with a squared genetic correlation[67] greater than 0.1 (similar to the squared phenotypic correlation threshold used in ref.[68]). The 63 datasets included six traits that were duplicated in two different datasets (genetic correlation of at least 0.9). Thus, we analyzed 57 independent diseases and complex traits. Traits were prioritized using the z-score for nonzero SNP-heritability computed using S- LDSC with the baseline-LD model (minimum of 6, as in ref.[69]). We also considered the 11 autoimmune diseases and blood cell traits, as in refs.[7,70].

Supplementary Table 4: Estimates of h^2 enrichment and gene enrichment for the validation and training critical gene sets for 50 S2G strategies. We report the h^2 enrichment, gene enrichment, and corresponding standard errors, meta-analyzed across 63 traits, for 50 S2G strategies, for both the (trait-specific) validation critical gene sets and the training critical gene set. h^2 enrichment is defined as the proportion of common variant heritability linked to the critical gene set, divided by the proportion of common SNPs linked to the critical gene set. Gene enrichment is defined as the fraction of common variant heritability linked to the critical gene set and to all genes, divided by the fraction of common SNPs linked to the critical gene set and to all genes. We also report estimates for validation critical gene sets constructed using default PoPS score (i.e. creating gene-level association statistics using gene body S2G strategy, rather than using Exon and Promoter S2G strategies); results were similar to the default validation critical gene sets.

Supplementary Table 5: Estimates of h^2 coverage, precision and recall for the validation and training critical gene sets for 50 S2G strategies and the cS2G strategy. We report the h^2 coverage, precision and recall, and corresponding standard errors, meta-analyzed across 63 traits, for 50 S2G strategies and the combined cS2G strategy, for both the (trait-specific) validation critical gene sets and the training critical gene set.

Supplementary Table 6: Weights of constituent S2G strategies in the combined cS2G strategy. We report the weights of each constituent S2G strategy in the combined cS2G strategy. We allowed weights to have a maximum value of 100, to prioritize S2G strategies with higher precision in the case where two S2G strategies link the same SNP to different genes. For example, if a SNP is linked to gene A through the Exon S2G strategy (weight = 100), and to gene B through the Cicero S2G strategy (weight = 1), then the cS2G linking score is 100/101 for gene A (stronger evidence from Exon), and 1/101 for gene B. We note that weights of 10 and 0.1 for Exon and Cicero (rather than 100 and 1), would have assigned the same linking scores in the case of the SNP described above, but would have assigned lower linking scores to SNPs that are linked to genes only through Cicero.

Supplementary Table 7: Overlap of SNP-gene links between the constituent strategies of cS2G. For each constituent strategy of cS2G (row), we report the proportion of SNP-gene links that overlap with another S2G strategy (column) (these values are not symmetric). For example, 4.6% of links from the GTEx fine-mapped *cis*-eQTL strategy are also in the Exon strategy, and 17.6% of links from the Exon strategy are also in the GTEx fine-mapped *cis*-eQTL strategy. All the numbers are restricted to links involving common SNPs.

Supplementary Table 8: Estimates of h^2 coverage, precision and recall for 4 different combined S2G strategies. We report the h^2 coverage, precision and recall, and corresponding standard errors, meta-analyzed across 63 traits, for 5 combined cS2G strategies. First, we considered our cS2G strategy. Second, we considered a combined S2G strategy where we linked all the SNPs linked by the cS2G (22% of common SNPs) to the gene with the closest TSS (ClosestTSS-cS2GSNPs). Third, we considered a combined S2G strategy maximizing recall (with precision > 0.75) when including all 50 S2G strategies (see Supplementary Table 9 for the 8 selected S2G strategies) (cS2G – 50 S2G). Fourth, we considered a combined S2G strategy where we restricted the SNPs linked by the cS2G strategy that had a linked gene with cS2G linking score >0.5 (99% of all linked SNPs and 82% of all SNP-gene pairs) (cS2G-score>0.50). Fifth, we considered a combined S2G strategy with the same 7 S2G strategies than cS2G, but give them the exact same weight (i.e. 1) (cS2G – same weights). We report values estimated using both validation and training critical gene sets. For cS2G and cS2G – 50 S2G, we also report values estimated during the optimization algorithm using training critical gene sets. We note that the precision of the combined S2G strategies in the training critical gene set tend to be large (>0.85), which is very likely due to higher precision for some constituent S2G strategies in the training critical gene set compared to the validation critical gene set (such as promoter, Closest TSS (1kb-5kb), or GTEx fine-mapped *cis*-eQTL; see Supplementary Table 5). We sought to further investigate the reduction in precision of ClosestTSS-cS2GSNPs vs. cS2G. We estimated the reduction in precision of ClosestTSS-cS2GSNPs as "proportion of h^2 coverage explained by cS2G SNPs not linked to Closest TSS by cS2G" * ("precision of cS2G for cS2G SNPs not linked to Closest TSS by cS2G" – "precision of Closest TSS for cS2G SNPs not linked to Closest TSS by cS2G") = 0.33 * (0.64 – 0.55) = 0.03, consistent with the observed reduction of 0.05.

Supplementary Table 9: Combined S2G strategy obtained when including all 50 S2G strategies. We report the selected S2G strategies and their corresponding weights that maximize recall (with precision > 0.75) when including all 50 S2G strategies. The resulting combined strategy included 8 S2G strategies: 4 that were included in our primary cS2G strategy (Exon, Promoter, eQTLGen blood fine-mapped *cis*-eQTL, GTEx fine-mapped *cis*-eQTL), as well as EpiMap and ABC restricted to blood and immune cell-types and tissues, Closest TSS (1-5kb), and GTEx all *cis*-eQTL.

Supplementary Table 10: Combined S2G strategies using different diseases/traits for training and validation. We split the set of 63 diseases/traits in 2 (1st half and 2nd half), built a combined S2G strategy using each of those (cS2G - 1st half, and cS2G - 2nd half, respectively), and report their h^2 coverage, precision and recall, and corresponding standard errors, meta-analyzed across each set of diseases/traits. cS2G - 1st half includes all the constituent S2G strategies of cS2G except EpiMap. cS2G - 2nd half includes all the constituent S2G strategies of cS2G except ABC and Cicero blood/basal. In all scenarios we observed that the combined strategies (cS2G - 1st half and cS2G - 2nd half) have a high precision (>0.76) and recall (>0.30). Note that in all scenarios, precision and recall were higher when using the training critical gene set than when using the validation critical gene set.

Supplementary Table 11: F1 scores of 50 S2G strategies and the combined strategy obtained when maximizing the F1 score. We report the *F1 score*, the harmonic mean of precision and recall[25], for the 50 S2G strategies and a combined strategy maximizing the F1 score in the training critical gene set (cS2G-F1). cS2G-F1 contains 5 out of 7 S2G strategies of cS2G (all but EpiMap and ABC). F1 scores were computed based on precision and recall estimated on the validation critical gene sets, and meta-analyzed across 63 traits. Strategies are ranked based on their F1 score. We observed that cS2G-F1 optimizes the F1 score over the 50 SG strategies, showing that our framework is robust to the quantity to maximize. For comparison purposes, we also report the F1 score of the cS2G strategy and observed that cS2G F1 score is slightly higher than cS2G-F1 F1 score. This result may be due to precision and recall heterogeneity across the training and validation critical gene sets, as cS2G-F1 maximizes the F1 score over cS2G during the optimization procedure (0.49 vs. 0.47 for cS2G).

Supplementary Table 12: Validation of combined S2G (cS2G) strategy using experimentally validated SNP-gene pairs. We manually curated a list of 17 disease-associated loci (including 12 loci from ref.[26]) containing 25 experimentally validated causal SNP-gene pairs. For each experimentally validated causal SNP-gene pairs, we report the cS2G predictions: predicted gene with cS2G linking score > 0.5 (if applicable), corresponding cS2G linking score, and constituent S2G annotation(s), as well as the gene with the closest TSS. Predicted genes that match the experimentally validated gene are denoted in bold font. Experimental validation is detailed in column "Functional experiment". Of the 25 pairs, 16 are annotated with cS2G linking score >0.5, of which 11 are accurately linked to the validated causal gene. We thus estimated precision as 11/16 = 0.69 (s.e. = 0.12), and recall as 11/25 = 0.44 (s.e. = 0.10). This is a lower precision and higher recall than our estimates based on validation critical gene sets (0.75 for precision, 0.33 for recall), but the differences were not statistically significant due to the small number of experimentally validated SNP-gene pairs. We disclose that the causal SNP rs1421085 was not annotated with cS2G linking score >0.5, as it was linked to 2 genes (*FTO* and *RPGRIP1L*) each with a score of 0.5, and we decided to prioritize SNPs linked to a gene with cS2G linking score >0.5 (if we had instead used cS2G linking score ≥0.5 as our threshold, the impact on results would be minimal, with precision = 11/17 = 0.65 instead of 0.69). We note that it is also possible to compute precision and recall across the 17 loci (instead of across the 25 SNPs). Of the 17 loci, 14 had at least one causal SNP annotated with cS2G linking score >0.5, of which 10 had a least one causal SNP accurately linked to the validated causal gene. For the *MYB* locus, one SNP was accurately linked to *MYB*, and one was inaccurately linked to another gene. By assigning a precision of 0.5 at this locus, we estimated am locus-based precision of 9.5/14 = 0.68 (s.e. = 0.12) and recall of 9.5/17 = 0.56 (s.e. = 0.12). GTEx: GTEx fine-mapped *cis*-eQTL; eQTLGen: eQTLGen blood fine-mapped *cis*-eQTL; EpiMap: EpiMap enhancer-gene linking; ABC: Activity-By-Contact; Cicero: Cicero blood/basal.

Supplementary Table 13: Precision and recall of cS2G and its constituent strategies for 25 experimentally validated SNP-gene pairs. We report the precision and recall (and corresponding binomial standard errors) of cS2G, its constituent strategies, and the Closest TSS strategy for 25 experimentally validated causal SNP-gene pairs. We observed that cS2G obtained the 2nd best precision and the best recall when compared to its constituent strategies, but differences were not statistically significant due to the small number of experimentally validated SNP-gene pairs. In addition, none of the precision and recall values were significantly different from the genome-wide estimates from polygenic analyses.

Supplementary Table 14: cS2G predictions for a curated list of 577 sentinel SNP-gene pairs with genes curated with high confidence by Open Targets. For each SNP in this list, we report the genes targeted by cS2G, their score, their corresponding annotations, and if the SNP-gene pair has been validated by Open Targets (column Validated). We also report the validated target gene if the cS2G lining score was 0. Note that the validated column has values different to 1 if multiple genes were assigned for one causal SNP. We warn here that while the causal gene has been validated with high confidence, the causal SNP(s) might be less confident as none of the 577 examples nominated causal SNPs using rigorous fine-mapping and/or functional follow-up, thus impacting the precision of the linking strategies. 356 of the 577 causal SNPs had a linked gene with a cS2G linking score >0.5, enabling us to predict the target gene. 205 of 356 predicted genes matched the Open Targets gene (precision = 205/356 = 0.58; we note that this precision definition, restricted to links with a cS2G linking score >0.5, is different from the one used in Extended Data Figure 3, which weights links by their cS2G linking score), and 205 of 577 causal SNP-gene-disease triplets were correctly identified (recall = 205/577 = 0.36) (see also Extended

Data Figure 3). Of these 205 triplets, 178 involved the gene with closest TSS, and 88 (resp. 20) involved variants in exons (resp. promoters), illustrating that curated SNP-gene-disease triplets that can be validated using functionally informed S2G strategies are preferentially ascertained for triplets in which the target gene is the gene with the closest TSS and/or implicated by a high-confidence S2G strategy. Of the 97 SNP-gene-disease triplets involving distal regulatory variants (defined here as not lying in exons or promoters), only 31 were functionally supported by at least 2 of the 5 remaining functionally informed S2G strategies.

Supplementary Table 15: Precision and recall of cS2G in the Open Targets and Weeks et al. curated datasets. We report precision and recall of cS2G in the Open Targets and Weeks et al. curated datasets when considering all SNP-gene pairs and when manually removing all pairs for obtained from diseases/traits with a name identical or similar to the names of our 63 diseases/traits. We determined that results were little changed in both datasets. We note that although our manual curation step to remove SNP-gene pairs that overlap our 63 diseases/traits is likely to be incomplete, and some overlapping or genetically correlated traits may remain, the fact that results were not sensitive to a substantial removal of overlapping diseases/traits strongly supports the overall robustness of our results.

Supplementary Table 16: Sentinel SNPs are linked to different genes than causal SNPs in analyses of fine-mapped UK Biobank traits. We found 114 SNP-gene-trait triplets with the underlying genes validated with high confidence by Open Targets[22] and with available fine-mapping results in our analyses of 49 UK Biobank traits. For each triplet, we report the sentinel SNP, gene and trait defined by the Open Target list (4 first columns), the trait ID in our UK Biobank analyses (5th column), if the SNP-gene pair was validated by cS2G (6th column), and the posterior inclusion probability (PIP) in our fine-mapping analyses (7th column). 58/114 pairs were validated using cS2G (precision = 51%). The mean and median PIP of the 58 validated pairs were 0.260 and 0.045, respectively, against 0.055 and 0.004 for 56 unvalidated pairs (*P* Wilcoxon test = 1.6 x 10^{-3}). The precision is 93% (resp. 71%) when restricted to the 14 (resp. 41) triplets with PIP >0.50 (resp. >0.05). These results highlight that sentinel SNPs that are fine-mapped are more likely to be linked to the accurate target gene using cS2G, and potentially explain the discrepancy between the precision estimated using 577 Open Targets pairs (58%) and our estimated precision (75%).

Supplementary Table 17: Causal SNP-gene-disease triplets predicted by application of cS2G strategy to 9,670 fine-mapped SNP-disease pairs. We report cS2G predictions for 9,670 predicted causal SNP-trait pairs with a posterior inclusion probability (PIP) > 0.50 from functionally informed fine-mapping of 49 UK Biobank diseases/traits[50,51]. Using cS2G linking scores >0.5, we predicted 7,111 causal SNP-gene-disease triplets (see column "In 7111") and report their corresponding confidence score. We also predicted 2,163 triplets involving distal regulatory fine-mapped SNPs that were not in the gene body (or promoter) of the target gene (see column "In 2163"), of which 532 were supported by at least 2 of the functionally informed constituent S2G strategies used by cS2G (see column "In 532"). We report the cS2G linking score as well as the score before normalization (column cS2G score*). We also report for each SNP-gene pair the annotations of the 7 constituent S2G strategies. Fine-mapped SNPs that were not linked to genes with cS2G have a value NA for all the columns "Gene", "cS2G score*", "cS2G score", and "cS2G triplet confidence score". SNP-gene pairs with a cS2G score <0.5 have value NA for the column "cS2G triplet confidence score".

Supplementary Table 18: cS2G strategy predicts more correct SNP-gene-disease triplets than other S2G strategies. We report for the 13 main S2G strategies, the Exon + Promoter strategy, and the cS2G strategy the number of inferred SNP-gene-disease triplets from 9,670 predicted causal SNP-trait pairs with a posterior inclusion probability (PIP) > 0.50 from functionally informed fine-mapping of 49 UK Biobank diseases/traits[50,51]. For each strategy, we report the number of unique SNPs and genes in all the triplets, the mean PIP across all the triplets, the mean confidence score across all the triplets, and the number of correct SNP-gene-disease triplets (obtained by multiplying the number of inferred triplets by the mean confidence score). We observed that the cS2G links at least 1.6 times more unique genes and predict at least 2.0 times more correct SNP-gene-disease triplets than any of the other 10 functionally informed S2G strategies.

Supplementary Table 19: Number of unique fine-mapped SNPs linked to each of the 3,401 unique genes in the 7,111 predicted SNP-gene-disease triplets. In many instances, multiple causal SNPs were linked to the same gene. For example, 119 genes were each linked to at least 5 fine-mapped SNPs, illustrating that a single gene can be causal for different diseases/traits using different causal SNP-gene links.

Supplementary Table 20: Cell-types identified by the functionally informed S2G strategies used by cS2G in the 4 examples of high-confidence SNP-gene-disease triplets identified by cS2G. For the 6 SNPs involved in the 4 examples of Figure 5, we report all the cell-types identified by the functionally informed S2G strategies used by cS2G, their linking score, and their *P* value in LDSC-SEG analyses. Results are ordered by LDSC-SEG *P* value significance.

Supplementary Table 21: List of 49 UK Biobank diseases/traits used to empirically assess the omnigenic model. We used the set of 49 traits and 16 independent traits as in ref. [51] (also same as in the fine-mapping analyses). We report the sample size used to estimate posterior mean squared causal effect sizes of genome-wide SNPs (*N*=337K British UK Biobank samples), and the sample size used to compute the summary statistics on European-ancestry UK Biobank samples that were distinct from the *N*=337K British UK Biobank samples (*N*=122K) for S-LDSC analyses.

Supplementary Table 22: Numerical results of assessment of disease omnigenicity using cS2G. We report the proportion of SNP-heritability (h^2) and the proportion of SNP-heritability linked to genes (h^2_{gene}) explained by the top 100, 200, 500, 1,000, 2,000, 5,000, 10,000 and all (19,995) genes using different linking strategies (cS2G and Closest TSS) (see Figure 5a). The standard error of the proportion of h^2_{gene} explained by the X top genes was computed as the standard error of the proportion of h^2 explained by the X top genes divided by h^2_{gene} (viewing the denominator h^2_{gene} as a constant); we believe this to be a reasonable approximation, as the numerator has greater uncertainty than the denominator (except when including all genes), and the errors are correlated such that this approximation is conservative. We note that a ratio of meta-analyzed values (meta-analyzed proportion of SNP-heritability explained by X top genes divided by meta-analyzed proportion of SNP-heritability explained by all genes) is more robust than a meta-analyzed value of ratios (meta-analyzing the proportion of SNP-heritability explained by X top genes divided by the proportion of SNP-heritability explained by all genes). For X<19,995, we note that values greater than 1 are outside the biologically plausible 0-1 range, but allowing point estimates outside the biologically plausible 0-1 range is necessary to ensure unbiasedness. The top genes were defined using posterior mean squared causal effect sizes estimated on *N*=337K British UK Biobank samples, and reported proportions were estimated using S-LDSC on the N=122K samples and meta-analyzed across 16 independent traits. For the cS2G, we also report the proportion of SNP-heritability linked to genes (h^2_{gene}) estimated using posterior mean squared causal effect sizes estimated on the *N*=337K samples; we observed that from 200 genes, h^2_{gene} S-LDSC estimates (based on *N*=122K samples) were not significantly different from the estimates directly based on the *N*=337K samples, implying minimal effects of winner's curse for a small number of genes.

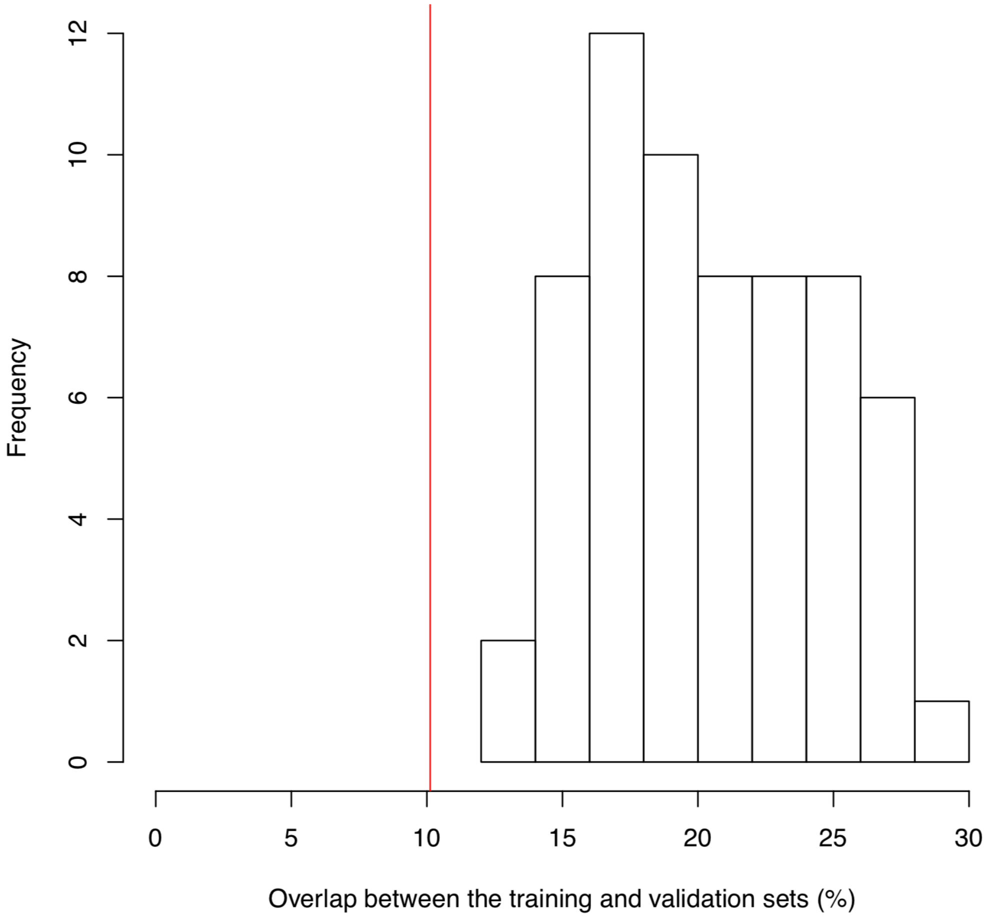
Supplementary Table 23: SNP-heritability explained by top genes with the highest per-gene heritabilities for each disease/trait. For each of the 49 traits, we report the proportion of SNP-heritability (h^2) and the proportion of SNP-heritability linked to genes (h^2_{gene}) explained by the top 100, 200, 500, 1,000, 2,000, 5,000, 10,000 and all (19,995) genes using the cS2G strategy. The standard error of the proportion of h^2_{gene} explained by the X top genes was computed as the standard error of the proportion of h^2 explained by the X top genes divided by h^2_{gene} (viewing the denominator h^2_{gene} as a constant); we believe this to be a reasonable approximation, as the numerator has greater uncertainty than the denominator (except when including all genes), and the errors are correlated such that this approximation is conservative. We note that a ratio of meta-analyzed values (meta-analyzed proportion of SNP-heritability explained by X top genes divided by meta-analyzed proportion of SNP-heritability explained by all genes) is more robust than a meta-analyzed value of ratios (meta-analyzing the proportion of SNP-heritability explained by X top genes divided by the proportion of SNP-heritability explained by all genes). For X<19,995, we note that values greater than 1 are outside the biologically plausible 0-1 range, but allowing point estimates outside the biologically plausible 0-1 range is necessary to ensure unbiasedness.

Supplementary Table 24: Estimates of the effective number of causal genes. For each of the 49 traits, we report its effective number of causal SNPs (M_e) and causal genes (G_e) (see Figure 5b), its effective number of causal genes explained by common variants ($G_{e,common}$) and low-frequency variants ($G_{e,low\text{-}frequency}$) (see Figure 5c), its correlation between per-gene heritability explained by common and low-frequency variants across all the genes (r_{20K}), its correlation between per-gene heritability explained by common and low-frequency variants restricted to genes in the top 200 (i.e. 1%) of per-gene heritability explained by common and low-frequency variants (r_{top200}), and the shared number of genes in the top 200 (i.e. 1%) of per-gene heritability explained by common and low-frequency variants ($shared_{top200}$). We also report the median values across 16 independent traits.

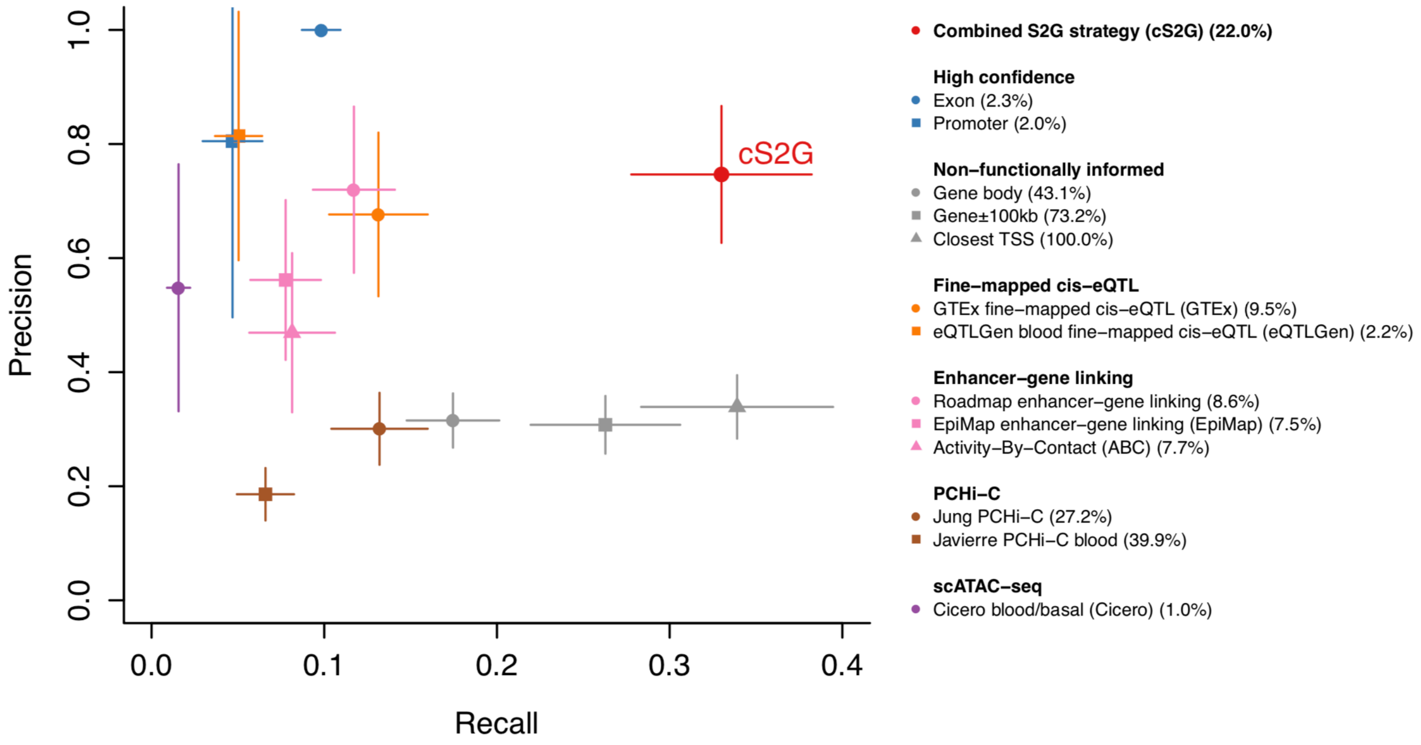
Supplementary Table 25: Top genes contributing to both common and low-frequency variant heritability linked to genes. Across all 49 traits, we report the 19 triplets (13 unique genes) where the gene is in the top 3 genes contributing to the common and low-frequency variant heritability linked to genes ($h^2_{gene,common}$ and $h^2_{gene,low\text{-}freq}$, respectively). We note that our results include *CDKN1C* for type 2 diabetes, further validating *CDKN1C* as the causal gene at this locus.

Supplementary Table 26: SNP-gene links for all datasets used to create S2G strategies. We report the SNP-gene links for all datasets used to create S2G strategies.

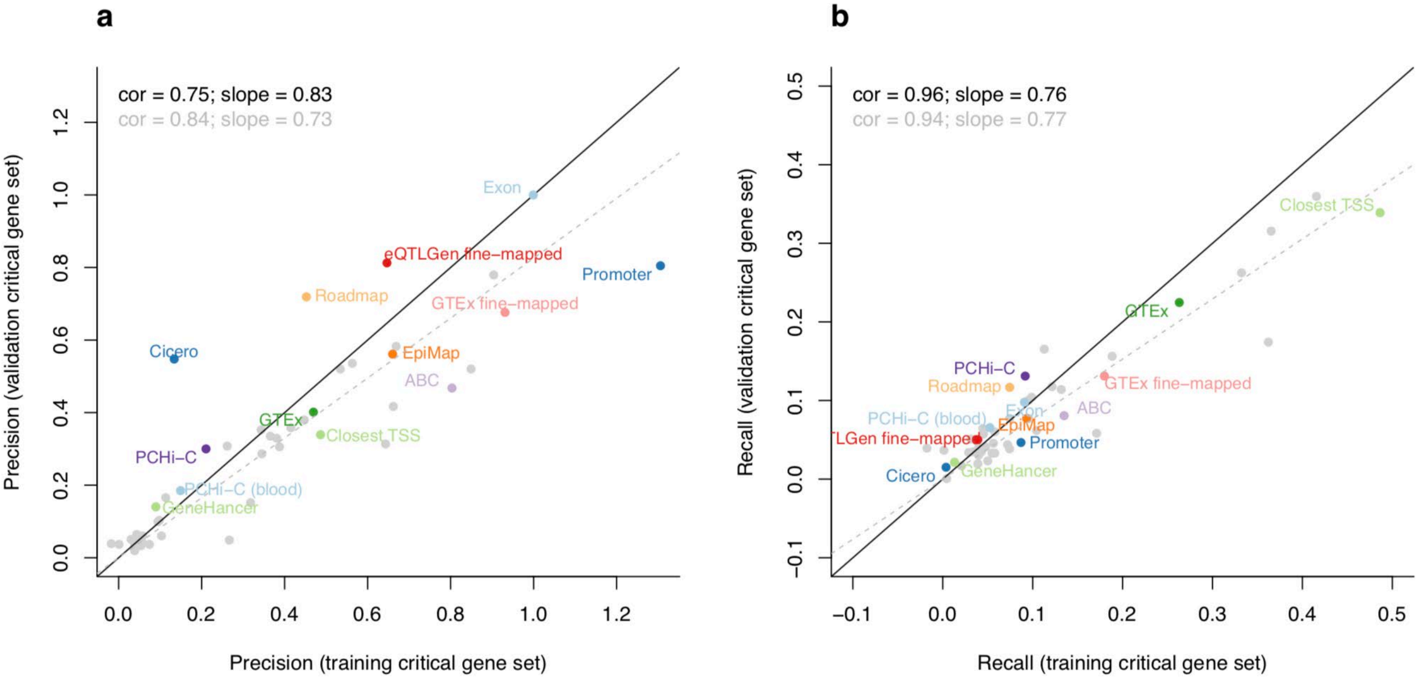
Supplementary Figures



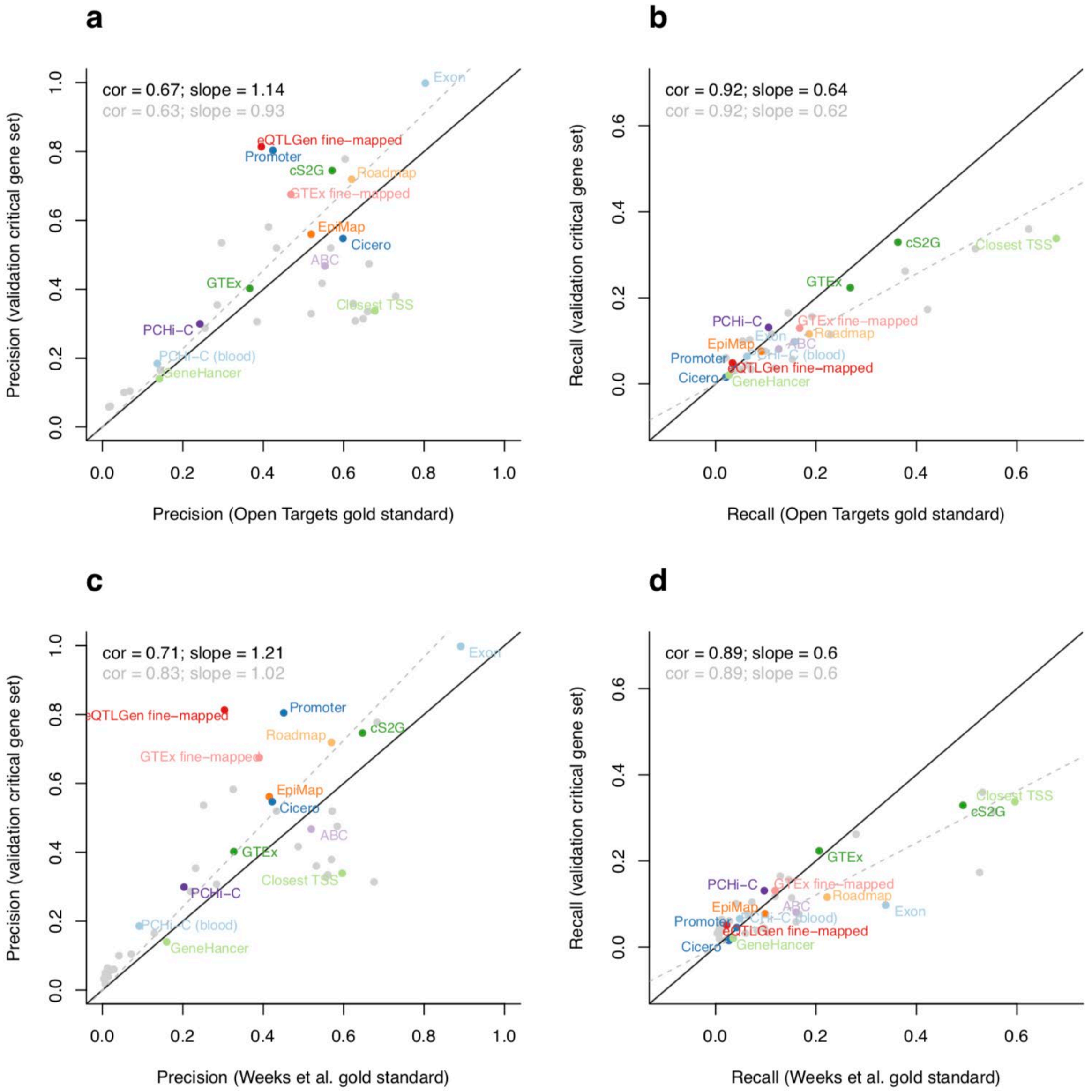
Supplementary Figure 1: Overlap between the training gene set and the validation gene sets. We report the distribution of the overlap between the training gene set (which does not vary across disease/traits) and the validation gene sets (which does vary across diseases/traits) across the 63 diseases/traits analyzed. It has a mean of 20% (vs. 10% expected by chance), mean of 20%, standard deviation of 4.1%, and range from 13%-28%.



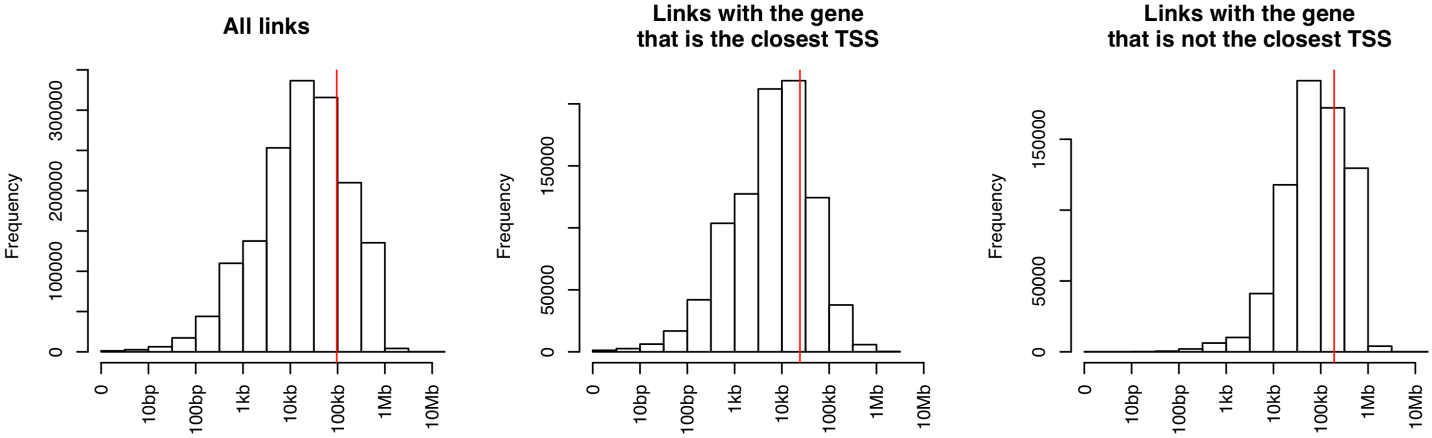
Supplementary Figure 2: Accuracy (with standard error) of individual S2G strategies and combined S2G (cS2G) strategy. We report the precision and recall of the 13 main S2G strategies from **Table 1** and the cS2G strategy (estimated using trait-specific validation critical gene sets and meta-analyzed across 63 independent traits). Reported results are identical to Figure 2, except that they include error bars representing 95% confidence intervals around meta-analyzed values. Our estimates of precision have large standard errors for S2G strategies linking a limited fraction of SNPs to genes; however, for the cS2G strategy, estimates of precision (0.75, s.e. 0.06) and recall (0.33, s.e. 0.03) were reasonably precise.



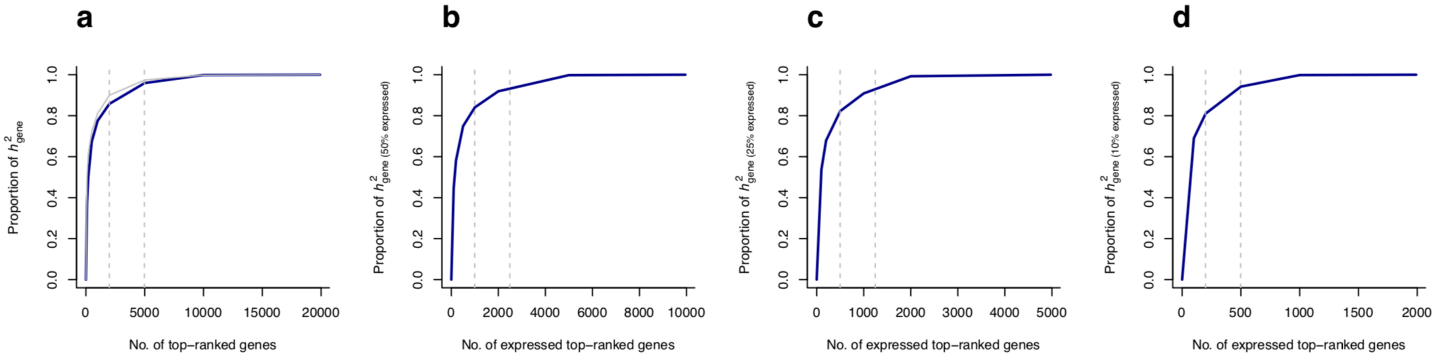
Supplementary Figure 3: Precision and recall in training and validation critical gene sets. We compared precision (**a**) and recall (**b**) estimated in our (non-trait-specifics) training critical gene set and (trait-specific) validation critical gene sets. Correlation (cor) and regression coefficient (slope) were computed either using the 13 highlighted independent S2G strategies (see Methods) (results in black), or using all 50 S2G strategies (results in grey). We observed high correlations and slopes for both precision and recall. GTEx: GTEx *cis*-eQTL; GTEx fine-mapped: GTEx fine-mapped *cis*-eQTL; eQTLGen fine-mapped: eQTLGen fine-mapped blood *cis*-eQTL; Roadmap: Roadmap enhancer-gene linking; EpiMap: EpiMap enhancer-gene linking; ABC: Activity-By-Contact



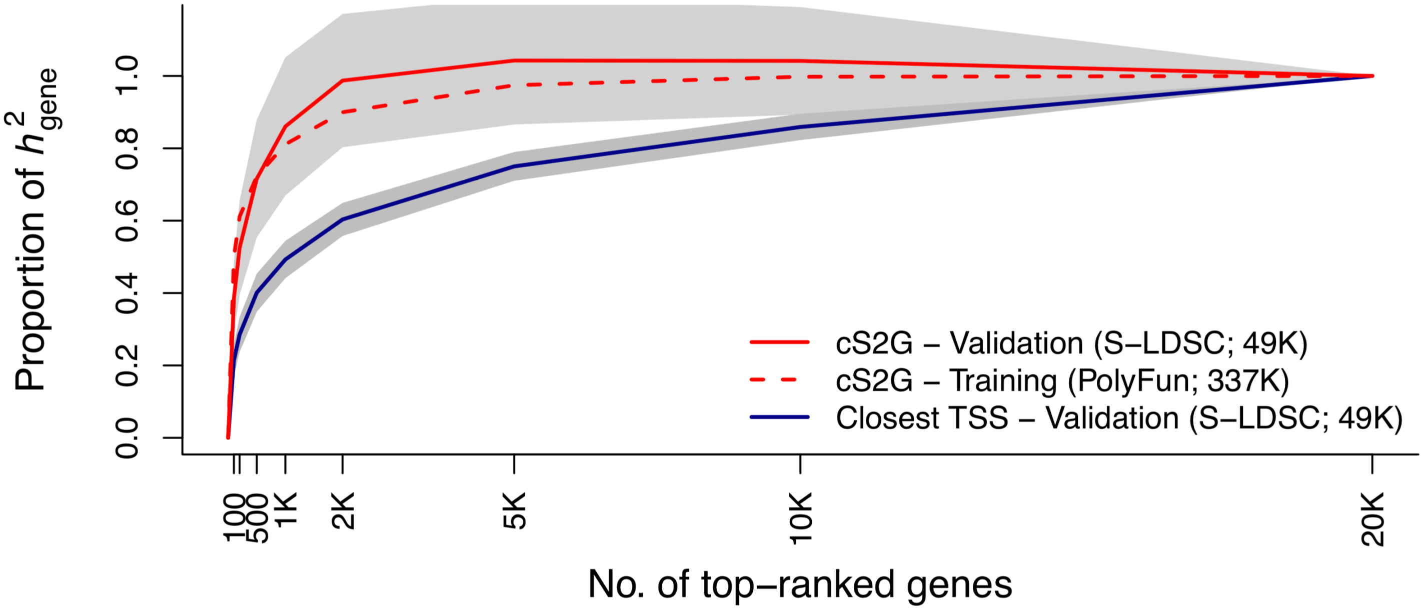
Supplementary Figure 4: Comparison of precision and recall estimates to independent definitions based on two curated disease-associated lists of SNP-gene pairs. We compared our estimates of precision and recall to independent definition of precision (i.e. not relying on critical gene sets or polygenic analyses) based on 577 linked sentinel SNP-gene pairs validated with high confidence by Open Targets[22] (**a, b**) and 1,668 linked fine-mapped SNP-gene pairs validated using nearby fine-mapped protein-coding variants[23] (**c, d**). Correlation (cor) and regression coefficient (slope) were computed either using the 13 highlighted independent S2G strategies (see Methods) and the cS2G strategy (results in black), or using all 50 S2G and cS2G strategies (results in grey). We observed high correlations and slopes for both precision and recall. (Note that since recall is the product of h^2 coverage and precision, differences in recall basically inherit the differences in precision) Despite the overall concordance, we observed large differences in precision and recall estimates for some S2G strategies (e.g. Exon, Closest TSS), as the curated causal SNP-gene pairs were preferentially ascertained for causal SNPs in which the target gene were the closest one: indeed, we observed an unusually high proportion of pairs involving genes with a short distance (< 10kb) to its closest TSS (57%/67% using both curated lists, vs h^2 coverage = 34% for the Closest TSS <10kb S2G strategy). Thus, we caution that curated disease-associated lists of linked SNP-gene pairs may be non-randomly ascertained, highlighting the potential benefits of polygenic analyses for evaluating S2G strategies. GTEx: GTEx *cis*-eQTL; GTEx fine-mapped: GTEx fine-mapped *cis*-eQTL; eQTLGen fine-mapped: eQTLGen fine-mapped blood *cis*-eQTL; Roadmap: Roadmap enhancer-gene linking; EpiMap: EpiMap enhancer-gene linking; ABC: Activity-By-Contact.



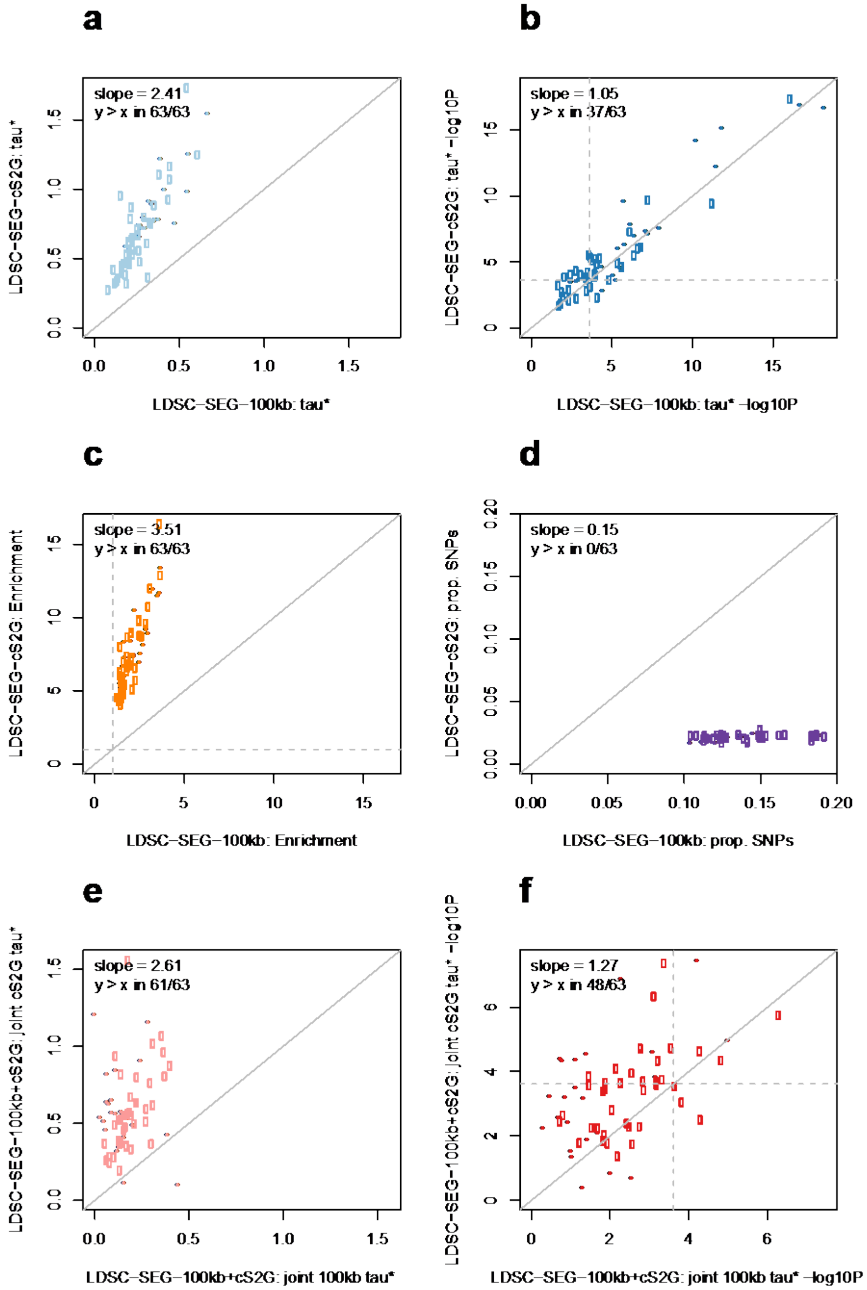
Supplementary Figure 5: Distribution of the distance to the gene TSS for all cS2G links involving common SNPs. We report the distribution of the distance to the gene TSS for all cS2G links (left), for links with the gene that is the closest TSS (middle), and for links with the gene that is not the closest TSS (right). The mean distance to the gene TSS for all cS2G links was 96kb (mean of 24kb for the 57% of linked SNPs linked to the gene with closest TSS, mean of 192kb for the 43% of linked SNPs not linked to the gene with closest TSS).



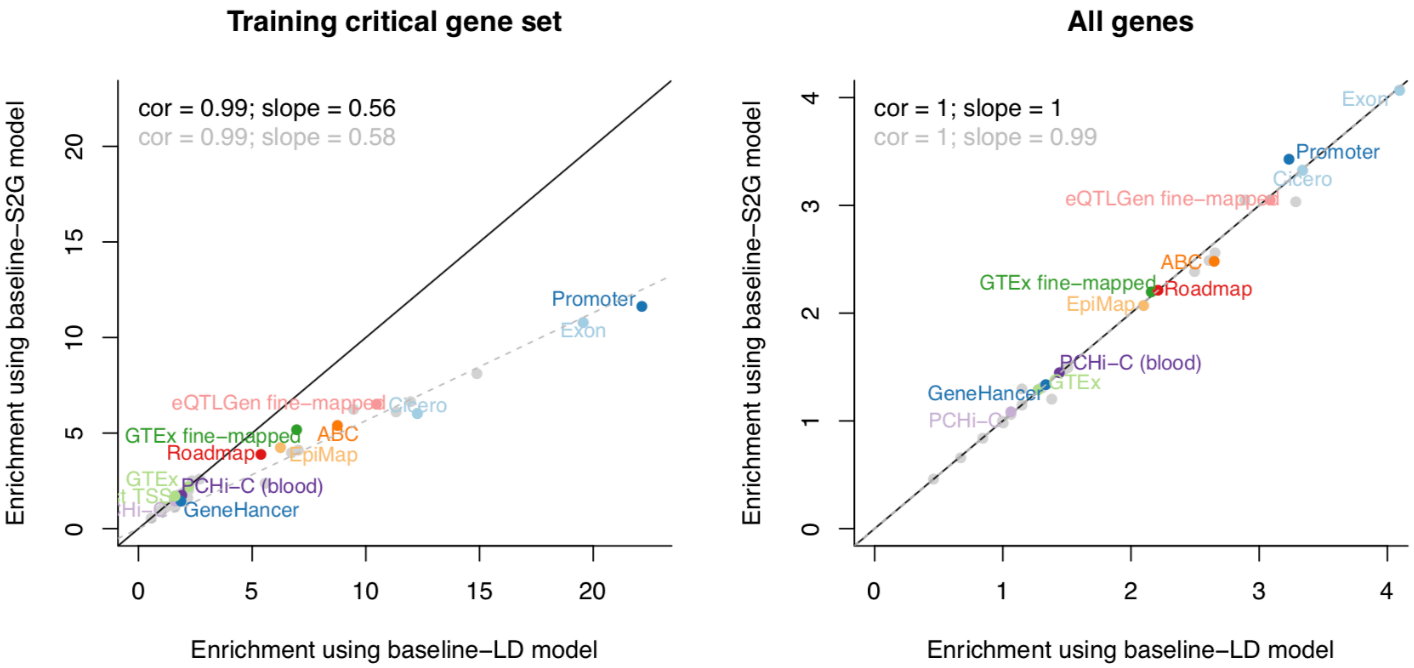
Supplementary Figure 6: SNP-heritability linked to genes expressed in disease-critical cell-types. We report the proportion of SNP-heritability linked to all genes (**a**; similar than Figure 5a) and linked to genes expressed in disease-critical cell-types (**b,c,d**) explained by genes with the top per-gene h^2. We restricted these analyses to 7 of the 16 independent traits that were analyzed in ref.[71] (Chronotype, Diastolic blood pressure, Eczema, Forced Vital Capacity, Mean Platelet Volume, Monocyte Count, and #Children) and plotted the median values. Disease-critical cell-types were selected as in ref.[71] and were glutamatergic for Chronotype, pericyte for Diastolic blood pressure, T-cells for Eczema, smooth muscle for Forced Vital Capacity, megakaryocytes for Mean Platelet Volume, monocytes for Monocyte Count, and GABAergic for #Children. We selected genes expressed in disease-critical cell-types based on the proportion of cells expressed in the cell-types. We selected 50% of the genes with the highest fraction of cells expressed (**b**), 25% of the genes (**c**), and 10% of the genes (**d**). Vertical grey lines indicate 10% and 25% of the genes selected in the analyses and were plotted for comparison purposes. Grey curve in (**a**) indicates results computed on the 16 independent traits (as in Figure 5a) and are similar to the ones computed on the restricted 7 traits. Overall, restricting analyses from Figure 5a to genes expressed in disease-critical cell-types had little impact on the proportion of retained SNP-heritability linked to genes explained by the top 10% of retained genes



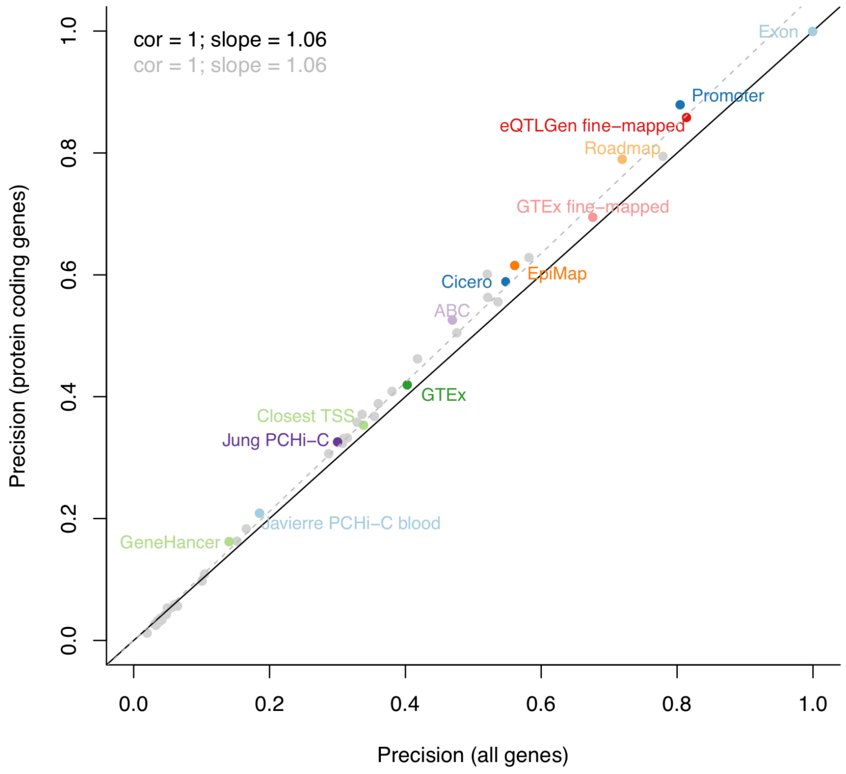
Supplementary Figure 7: Proportion of heritability linked to genes when using a restricted set of *N*=49K validation samples that were unrelated to the *N*=337K training samples. The figure is analogous to Figure 5a, except that a restricted set of validation samples is used. We report the proportion of SNP-heritability linked to genes (h^2_{gene}) explained by genes ranked by top per-gene h^2, as inferred using three approaches. Grey shading denotes 95% confidence intervals for cS2G-validation and Closest TSS-validation around meta-analyzed values. We forced the s.e. of the proportion of h^2_{gene} explained by all genes to be 0 (see Methods). We note that values greater than 1 are outside the biologically plausible 0-1 range, but allowing point estimates outside the biologically plausible 0-1 range is necessary to ensure unbiasedness. These proportions were inferred using three approaches, including two approaches using S-LDSC from an *N*=49K non-British European-ancestry UK Biobank validation sample not related to the *N*=337K training sample. Grey shading denotes 95% confidence intervals for cS2G-validation and Closest TSS-validation. Results were meta-analyzed across 16 independent UK Biobank traits. We observed very similar results for cS2G when using the *N*=49K validation samples and the *N*=122K validation samples used in main analyses: the top 200 (resp. top 2,000) genes explained 52±7% (resp. 99±9%) of the disease heritability linked to genes in cis using the cS2G strategy (h^2_{gene}, which captures 55±4% of h^2) when using the *N*=49K validation samples, versus 52±6% (resp. 96±8%) of the disease heritability linked to genes in *cis* using the cS2G strategy (h^2_{gene}, which captures 53±3% of h^2) when using the *N*=122K validation samples. For the Closest TSS strategy, the top 1,000 (resp. top 10,000) genes to explain 49±3% (resp. 86±2%) of h^2_{gene} when using the *N*=49K validation samples, versus 48±2% (resp. 85±2%) of h^2_{gene} when using the *N*=122K validation samples.



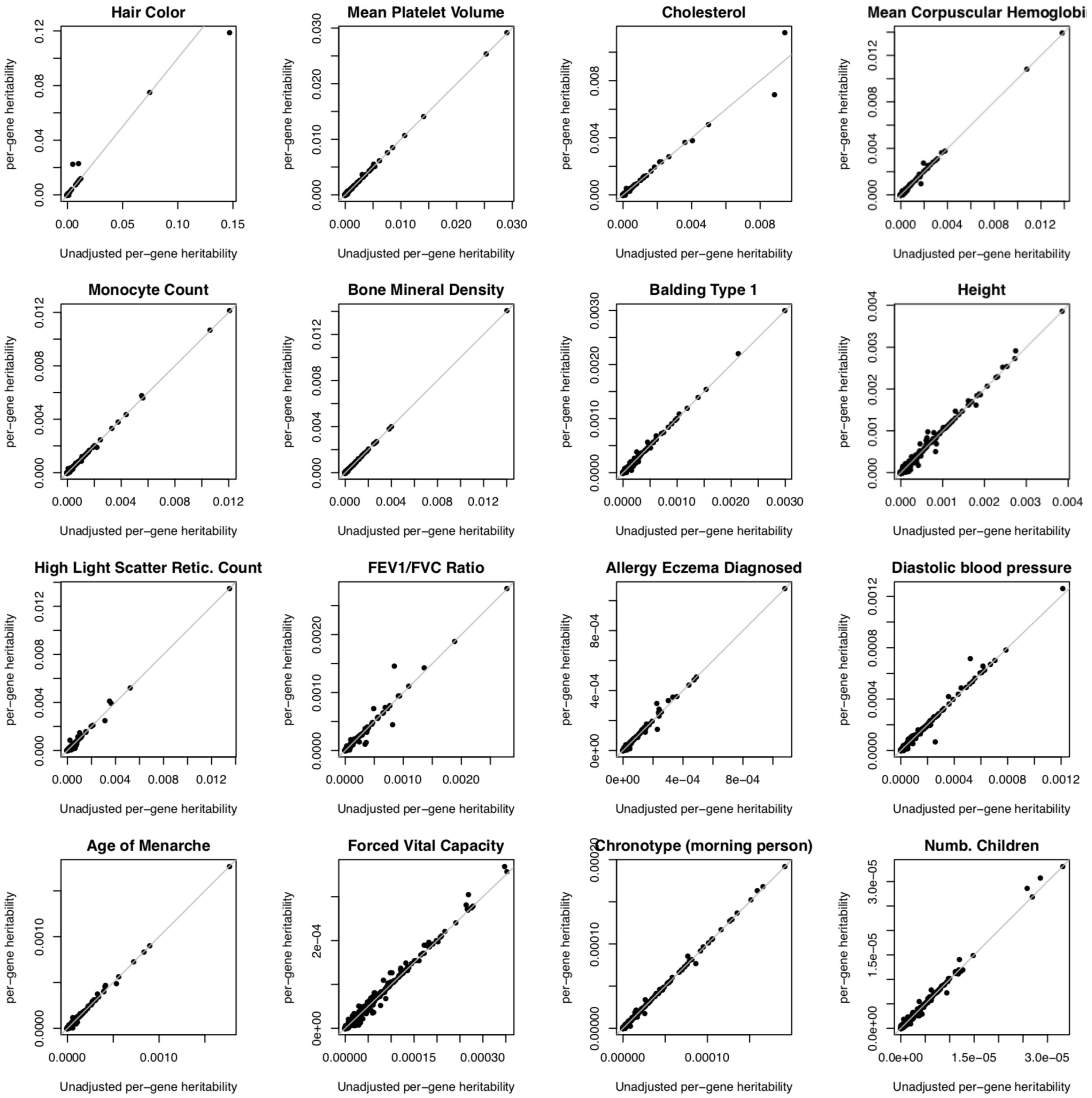
Supplementary Figure 8: Results of using cS2G vs. ±100kb windows to define SNP annotations in analyses of differentially expressed genes that are enriched for disease SNP-heritability. We compared LDSC-SEG[41] results with the default strategy of linking SNPs to a gene set using a Gene body ± 100kb approach (LDSC-SEG-100kb; *x* axis) to an alternative approach leveraging our cS2G strategy (LDSC-SEG-cS2G; *y* axis). For each of the 63 independent traits, we selected the differentially expressed gene set (out of 205), with the smallest LDSC regression coefficient *P* value (traits with the same gene set selected by LDSC-SEG-100kb and LDSC-SEG-cS2G were represented by a filled dot). We reported the per-standardized-annotation effect sizes τ^* (**a**), the per-standardized-annotation effect sizes τ^* two-sided *P* value (**b**), the SNP-heritability enrichment (**c**), and the proportion of SNPs (**d**) for each 63 trait-gene set pairs. We consistently observed higher τ^* and SNP-heritability enrichment for LDSC-SEG-cS2G (**a,c**). LDSC-SEG-cS2G also produces slightly more significant *P* values than LDSC-SEG-100kb (**b**), despite the fact of losing statistical power by creating SNP annotations with an average of 7 times less SNPs than the default approach (**d**). For 37 out of 63 traits, we obtained more significant *P* values with LDSC-SEG-cS2G; 10 traits have a significant *P* value (<0.05/205) with LDSC-SEG-cS2G but not with LDSC-SEG-100kb, and 5 traits have a significant *P* value with LDSC-SEG-100kb but not with LDSC-SEG-cS2G. In further analyses, we considered a joint model with best gene set SNP annotations from LDSC-SEG-100kb and LDSC-SEG-cS2G (LDSC-SEG-100kb+cS2G). We reported the joint per-standardized-annotation effect sizes τ^* of the two gene set SNP annotations (**e**), and corresponding two-sided *P* value (**f**). We observed that once conditioned to the cS2G gene set SNP annotation, the 100kb gene set SNP annotation was rarely significant (P <0.05/205 in 8/63 traits, against 27/63 with cS2G), suggesting that cS2G captures most of the information in a gene body ± 100kb.



Supplementary Figure 9: Comparison of SNP-heritability enrichment estimates of SNP annotations derived from enriched gene sets using different SNP-heritability models. We report the SNP-heritability enrichment of SNP annotations intersecting S2G strategies with constrained genes (left) or all genes (right), estimated by S-LDSC either using the baseline-LD model (and the focal S2G-derived SNP annotation, plus the corresponding S2G-derived SNP annotation for all genes if different) (*x* axis), or using a model with all baseline-LD SNP annotations and 80 S2G-derived SNP annotations (50 S2G-derived SNP annotations constructed by restricting SNPs linked to genes of the critical gene set, and 30 S2G-derived SNP annotations constructed by restricting SNPs linked to all genes (see Methods); baseline-S2G model) (*y* axis). Correlation (cor) and regression coefficient (slope) were computed using either the 13 highlighted independent S2G strategies (see Methods) (results in black), or using all 50 S2G strategies (results in grey). We observed that SNP-heritability enrichment estimates of SNP annotations intersecting S2G strategies with constrained genes were nearly two times higher when using the baseline-LD model than when using the baseline-S2G model. We thus recommend that future S-LDSC SNP-heritability enrichment analyses of gene sets should carefully consider the set of SNP annotations included in the model. We hypothesize that the biases observed under the baseline-LD model are due to tagging effects of unmodeled S2G links; in this case, these biases would not lead to false-positive enriched gene sets. GTEx: GTEx *cis*-eQTL; GTEx fine-mapped: GTEx fine-mapped *cis*-eQTL; eQTLGen fine-mapped: eQTLGen fine-mapped blood *cis*-eQTL; Roadmap: Roadmap enhancer-gene linking; EpiMap: EpiMap enhancer-gene linking; ABC: Activity-By-Contact.



Supplementary Figure 10: Estimates of precision using 17,871 protein-coding genes instead of 19,995 protein-coding and non-protein-coding genes. We compared our default estimates of precision using 19,995 protein-coding and non-protein-coding genes to an estimate restricting analyses to 17,871 protein-coding genes. Correlation (cor) and regression coefficient (slope) were computed either using the 13 highlighted independent S2G strategies (see Methods) and the cS2G strategy (results in black), or using all 50 S2G strategies (results in grey). We observed nearly identical estimates. We note that we also used cS2G to estimate that 1.30% of the SNPs linked to non-protein-coding genes (2,124 out of 19,995 genes) explain 1.55 ± 0.18% of SNP-heritability, justifying their inclusion even if they are not enriched in SNP-heritability (two-sided $P = 0.71$). GTEx: GTEx *cis*-eQTL; GTEx fine-mapped: GTEx fine-mapped *cis*-eQTL; eQTLGen fine-mapped: eQTLGen fine-mapped blood *cis*-eQTL; Roadmap: Roadmap enhancer-gene linking; EpiMap: EpiMap enhancer-gene linking; ABC: Activity-By-Contact.



Supplementary Figure 11: Unadjusted vs. adjusted per-gene heritability estimates. We report unadjusted per-gene heritability and adjusted per-gene heritability estimates across 16 independent UK Biobank traits. Adjusting per-gene heritability impacted estimates of only a small number of genes.

References

1. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766–D773 (2019).
2. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
3. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
4. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
5. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**, 827–841 (2013).
6. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
7. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
8. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biology* **18**, 193 (2017).
9. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
10. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
11. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat Genet* **51**, 1664–1669 (2019).
12. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet* **51**, 1442–1449 (2019).
13. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

14. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936 (2019).

15. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, (2017).

16. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48**, 245–252 (2016).

17. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature Genetics* **50**, 956–967 (2018).

18. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications* **10**, 3300 (2019).

19. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592–599 (2019).

20. Weiner, D. J., Gazal, S., Robinson, E. B. & O'Connor, L. J. Partitioning gene-mediated disease heritability without eQTLs. *bioRxiv* 2021.07.14.452393 (2021) doi:10.1101/2021.07.14.452393.

21. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

22. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021).

23. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv* 2020.09.08.20190561 (2020) doi:10.1101/2020.09.08.20190561.

24. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

25. Van Rijsbergen, C. J. *Information Retrieval. 2nd. Newton, MA.* (1979).

26. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am J Hum Genet* **102**, 717–730 (2018).

27. Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. *Nat Genet* **46**, 126–135 (2014).

28. The GAME-ON/ELLIPSE Consortium *et al.* CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nat Med* **21**, 1357–1363 (2015).

29. Soldner, F. *et al.* Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature* **533**, 95–99 (2016).

30. Fogarty, M. P., Cannon, M. E., Vadlamudi, S., Gaulton, K. J. & Mohlke, K. L. Identification of a Regulatory Variant That Binds FOXA1 and FOXA2 at the CDC123/CAMK1D Type 2 Diabetes GWAS Locus. *PLOS Genetics* **10**, e1004633 (2014).

31. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).

32. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine* **373**, 895–907 (2015).

33. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).

34. Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45**, 1274–1283 (2013).

35. Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics* **50**, 668–681 (2018).

36. Yasuda, K. *et al.* Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nature Genetics* **40**, 1092–1097 (2008).

37. Unoki, H. *et al.* SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nature Genetics* **40**, 1098–1102 (2008).

38. Kerns, S. L. *et al.* A novel variant in CDKN1C is associated with intrauterine growth restriction, short stature, and early-adulthood-onset diabetes. *J Clin Endocrinol Metab* **99**, E2117-2122 (2014).

39. Travers, M. E. *et al.* Insights Into the Molecular Mechanism for Type 2 Diabetes Susceptibility at the KCNQ1 Locus From Temporal Changes in Imprinting Status in Human Islets. *Diabetes* **62**, 987–992 (2013).

40. Chiou, J. *et al.* Single-cell chromatin accessibility identifies pancreatic islet cell type– and state-specific regulatory programs of diabetes risk. *Nature Genetics* **53**, 455–466 (2021).

41. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621–629 (2018).

42. Gour, N. & Wills-Karp, M. IL-4 and IL-13 Signaling in Allergic Airway Disease. *Cytokine* **75**, 68–78 (2015).

43. Ogasawara, T. *et al.* Development of chronic allergic responses by dampening Bcl6-mediated suppressor activity in memory T helper 2 cells. *Proc Natl Acad Sci U S A* **114**, E741–E750 (2017).

44. Revez, J. A. *et al.* Identification of STOML2 as a putative novel asthma risk gene associated with IL6R. *Allergy* **71**, 1020–1030 (2016).

45. Schmidt, M., Sun, G., Stacey, M. A., Mori, L. & Mattoli, S. Identification of Circulating Fibrocytes as Precursors of Bronchial Myofibroblasts in Asthma. *The Journal of Immunology* **171**, 380–389 (2003).

46. Kleffel, S. *et al.* Melanoma Cell-Intrinsic PD-1 Receptor Functions Promote Tumor Growth. *Cell* **162**, 1242–1256 (2015).

47. Prokunina, L. *et al.* A regulatory polymorphism in PDCD1 is associated with susceptibility to systemic lupus erythematosus in humans. *Nature Genetics* **32**, 666–669 (2002).

48. Eskelinen, E.-L. *et al.* Disturbed Cholesterol Traffic but Normal Proteolytic Function in LAMP-1/LAMP-2 Double-deficient Fibroblasts. *MBoC* **15**, 3132–3145 (2004).

49. Eskelinen, E.-L. Roles of LAMP-1 and LAMP-2 in lysosome biogenesis and autophagy. *Molecular Aspects of Medicine* **27**, 495–502 (2006).

50. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

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

52. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).

53. Gasperini, M. *et al.* A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell* **176**, 377-390.e19 (2019).

54. Moore, J. E., Pratt, H. E., Purcaro, M. J. & Weng, Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol* **21**, 17 (2020).

55. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet* **37**, 109–124 (2021).

56. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

57. Hindorff, L. A. *et al.* Prioritizing diversity in human genomics research. *Nat Rev Genet* **19**, 175–185 (2018).

58. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).

59. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am J Hum Genet* **97**, 260–271 (2015).

60. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390–400 (2018).

61. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet* **51**, 1670–1678 (2019).

62. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat Commun* **12**, 1098 (2021).

63. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).

64. Pliner, H. A. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* **71**, 858-871.e8 (2018).

65. Liu, X., Li, Y. I. & Pritchard, J. K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034.e6 (2019).

66. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

67. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

68. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature Genetics* **50**, 1600–1607 (2018).

69. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

70. Dey, K. K. *et al.* Contribution of enhancer-driven and master-regulator genes to autoimmune disease revealed using functionally informed SNP-to-gene linking strategies. *bioRxiv* 2020.09.02.279059 (2021) doi:10.1101/2020.09.02.279059.

71. Jagadeesh, K. A. *et al.* Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics. *bioRxiv* 2021.03.19.436212 (2021) doi:10.1101/2021.03.19.436212.