



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

Steven Gazal<sup>1,2,3,4</sup>✉, Omer Weissbrod<sup>1,3,4</sup>, Farhad Hormozdiari<sup>3,4</sup>, Kushal K. Dey<sup>1,3,4</sup>, Joseph Nasser<sup>4</sup>, Karthik A. Jagadeesh<sup>3,4</sup>, Daniel J. Weiner<sup>4</sup>, Huwenbo Shi<sup>1,3,4</sup>, Charles P. Fulco<sup>4,5,10</sup>, Luke J. O'Connor<sup>1,4</sup>, Bogdan Pasaniuc<sup>6</sup>, Jesse M. Engreitz<sup>4,7,8</sup> and Alkes L. Price<sup>1,3,4,9</sup>✉

Disease-associated single-nucleotide polymorphisms (SNPs) generally do not implicate target genes, as most disease SNPs are regulatory. Many SNP-to-gene (S2G) linking strategies have been developed to link regulatory SNPs to the genes that they regulate in *cis*. Here, we developed a heritability-based framework for evaluating and combining different S2G strategies to optimize their informativeness for common disease risk. Our optimal combined S2G strategy (cS2G) included seven constituent S2G strategies and achieved a precision of 0.75 and a recall of 0.33, more than doubling the recall of any individual strategy. We applied cS2G to fine-mapping results for 49 UK Biobank diseases/traits to predict 5,095 causal SNP-gene-disease triplets (with S2G-derived functional interpretation) with high confidence. We further applied cS2G to provide an empirical assessment of disease omnigenicity; we determined that the top 1% of genes explained roughly half of the SNP heritability linked to all genes and that gene-level architectures vary with variant allele frequency.

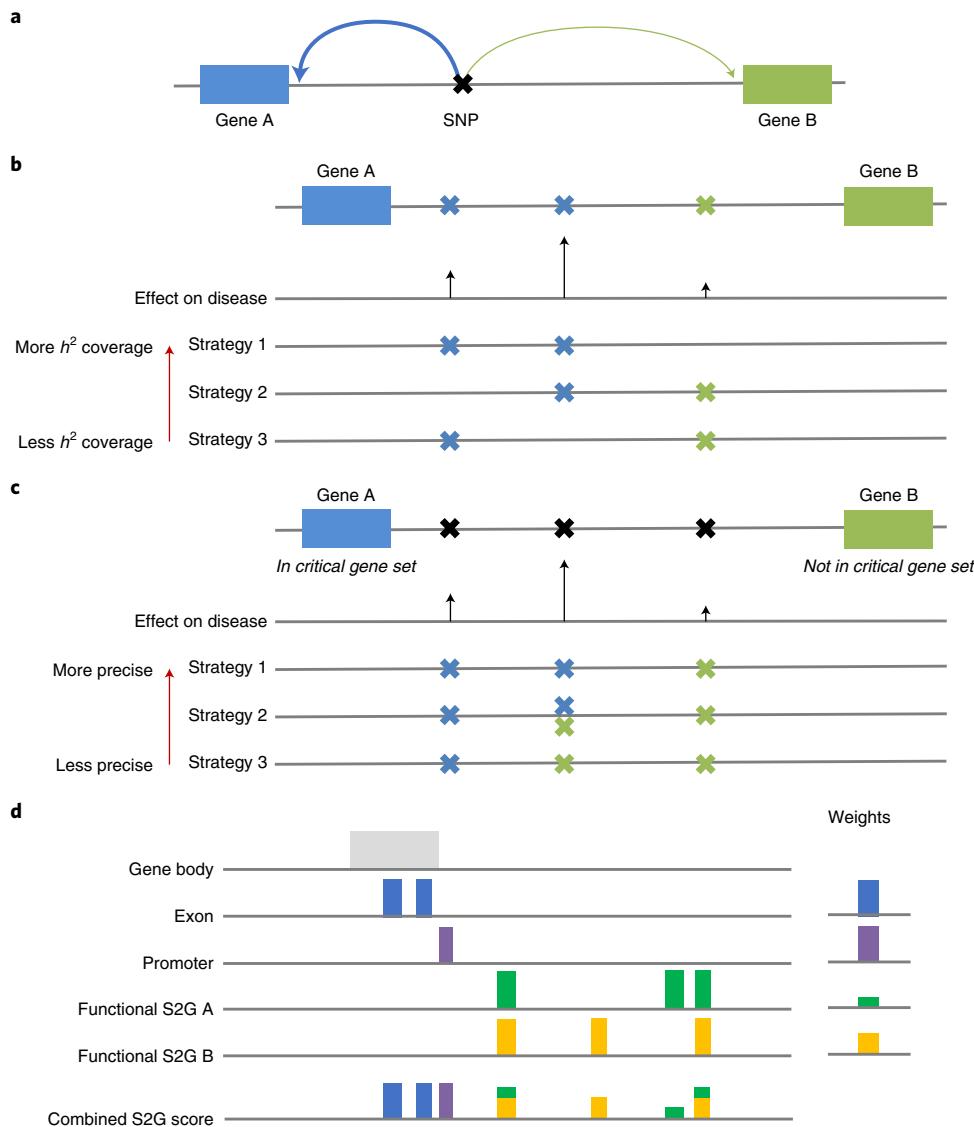
Although genome-wide association studies (GWASs) have successfully identified thousands of loci associated with human diseases and complex traits, they generally do not identify the underlying causal variants, target genes, cell types and biological functions, thus limiting the translation of GWAS findings into discoveries that will enhance disease treatment<sup>1–3</sup>. Although recent advances in fine-mapping techniques have improved our ability to nominate causal variants<sup>4–7</sup>, identifying the underlying target genes remains a critical challenge, as causal variants are predominantly regulatory SNPs<sup>8–11</sup> that do not necessarily regulate the closest genes<sup>12–17</sup>. Large gene expression quantitative trait loci (eQTL) data sets<sup>18,19</sup> have proven useful in linking disease variants to their target genes through colocalization analyses<sup>20–24</sup> or transcriptome-wide association studies<sup>14,17,25</sup> but explain a limited proportion of disease SNP heritability<sup>15,26,27</sup>, likely due to limited representation of disease-relevant cell types or states<sup>28</sup>. Many other functional assays and computational approaches have recently been developed to link regulatory SNPs to their target genes in *cis* in a broad set of cell types<sup>26,29–38</sup>; for example, EpiMap enhancers<sup>37</sup> and ABC enhancers<sup>34,38</sup> are linked to their target genes using correlation of enhancer activity with gene expression across cell types and a quantitative combination of enhancer activity and 3D contact frequencies, respectively. Combining S2G linking strategies has previously been proposed as an appealing approach to improve S2G linking<sup>39–42</sup>. However, it is currently unclear how S2G strategies should be prioritized in the context of GWAS, limiting our ability to pinpoint disease genes.

Here, we developed a framework for evaluating and combining S2G strategies to optimize their informativeness for human disease risk, leveraging polygenic analyses of disease SNP heritability. We applied this framework to GWAS summary statistics for 63 diseases and complex traits, evaluating 50 S2G strategies and constructing an optimal combined S2G strategy (cS2G) informed by GWAS data. We applied cS2G to genome-wide fine-mapping results for 49 diseases and complex traits from the UK Biobank<sup>7,43</sup> to pinpoint disease genes at thousands of GWAS loci and enable an empirical assessment of the ‘omnigenic model’<sup>44–46</sup>.

## Results

**Overview of methods.** An S2G linking strategy is defined as an assignment of raw linking values between each SNP and zero or more candidate target genes, which we converted into linking scores such that each SNP has a sum of linking scores  $\leq 1$  (Fig. 1a and Methods). We analyzed 50 S2G strategies that link SNPs to their target genes in *cis* (Supplementary Table 1 and Supplementary Note), including 13 main strategies (Table 1). Correlations and overlap proportions between S2G strategies (Supplementary Table 2) indicate low concordance between the strategies. In our primary analyses, we restricted each S2G strategy by restricting each SNP to the gene with the highest linking score, as we observed that this method led to slightly higher precision (Extended Data Fig. 1). To evaluate each S2G strategy’s informativeness for pinpointing disease genes, we leveraged polygenic analyses of disease SNP heritability (we denote SNP heritability as  $h^2$  throughout the paper). We aimed

<sup>1</sup>Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. <sup>2</sup>Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. <sup>3</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. <sup>4</sup>Broad Institute of MIT and Harvard, Cambridge, MA, USA. <sup>5</sup>Department of Systems Biology, Harvard Medical School, Boston, MA, USA. <sup>6</sup>Departments of Computational Medicine, Human Genetics, Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. <sup>7</sup>Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. <sup>8</sup>BASE Initiative, Betty Irene Moore Children’s Heart Center, Lucile Packard Children’s Hospital, Stanford University School of Medicine, Stanford, CA, USA. <sup>9</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. <sup>10</sup>Present address: Bristol Myers Squibb, Cambridge, MA, USA. ✉e-mail: [gazal@usc.edu](mailto:gazal@usc.edu); [aprince@hspf.harvard.edu](mailto:aprince@hspf.harvard.edu)



**Fig. 1 | Overview of the S2G framework.** **a**, Toy example of SNP linked to two genes (arrow widths denote linking scores). **b**, Toy example of  $h^2$  coverage. Strategy 1 (which links SNPs with larger effects on disease) has more  $h^2$  coverage than strategy 2, which has more  $h^2$  coverage than strategy 3 (which links SNPs with smaller effects on disease). **c**, Toy example of using critical gene sets to define precision. Strategy 1 (which links the middle SNP with high effect on disease to the gene from the critical gene set) is more precise than strategy 2 (which links the middle SNP to both genes), which is more precise than strategy 3 (which links the middle SNP to the gene that is not from the critical gene set). Recall is defined as the product of the  $h^2$  coverage and precision. **d**, Toy example of a cS2G. The cS2G is a linear combination of constituent S2G strategies.

to define and estimate parameters that correspond to an S2G strategy's  $h^2$  coverage (proportion of total disease SNP heritability that is linked to genes), precision (proportion of disease SNP heritability linked to genes that is linked to the correct target gene) and recall (proportion of total disease SNP heritability that is linked to the correct target gene). First, we defined  $h^2$  coverage as the proportion of SNP heritability explained by all SNPs linked to one or more genes (weighted by their linking scores) (Fig. 1b). Second, we defined precision as the relative excess SNP-heritability enrichment of SNPs linked to a critical gene set (see below) versus SNPs linked to all genes, as compared to the (gold standard) exon S2G strategy; this definition is based on the intuition that a precise S2G strategy is more likely (than an imprecise strategy) to link a disease SNP to a critical gene (Fig. 1c). Third, we defined recall as the product of the  $h^2$  coverage and precision. We estimated these quantities by applying stratified linkage disequilibrium score regression (S-LDSC)

to 63 independent diseases and complex traits (average  $N=314K$  (e.g., 314,000); Supplementary Table 3), meta-analyzing results across traits. We jointly analyzed SNP annotations derived from the 50 S2G strategies for ~10,000,000 SNPs with a minor allele count  $\geq 5$  in a 1000 Genomes Project European reference panel<sup>17</sup>.

Our definitions of precision and recall rely on a critical gene set (see above). We used a non-trait-specific training critical gene set to construct an optimal cS2G (see below) and trait-specific validation critical gene sets to evaluate the optimal cS2G while avoiding overfitting. We defined the training critical gene set as the top 10% of genes with the most highly constrained exons and conserved promoters and the validation critical gene set for a given trait as the top 10% of genes ranked by the PoPS method<sup>48</sup>; the excess overlap between the gene sets was limited (Supplementary Fig. 1 shows the distribution of overlap across diseases/traits) and did not impact our results (Supplementary Note). These gene sets attained high

**Table 1 | Description of the 13 main S2G strategies**

S2G strategy	Description	SNPs linked (%)	$h^2$ coverage (%)
<b>High confidence (2)</b>			
Exon <sup>a</sup>	Exons <sup>60</sup> $\pm$ 20 bp	2.3	9.8
Promoter <sup>a</sup>	TSSs <sup>61</sup> $\pm$ 1 kb $\cap$ promoter annotations <sup>11,30</sup>	2.0	5.8
<b>Non-functionally informed (3)</b>			
Gene body	Gene body <sup>60–62</sup>	43.1	55.4
Gene $\pm$ 100 kb	Gene body <sup>60–62</sup> $\pm$ 100 kb <sup>56</sup>	73.2	85.4
Closest TSS	Gene with closest TSS <sup>61</sup>	100	100
<b>Fine-mapped cis-eQTL (2)</b>			
GTEx fine-mapped cis-eQTL <sup>a</sup>	Fine-mapped GTEx v8 cis-eQTLs <sup>18,26</sup> in 54 cell types	9.5	19.4
eQTLGen fine-mapped blood cis-eQTL <sup>a</sup>	Fine-mapped eQTLGen cis-eQTLs <sup>19,26</sup> in blood	2.2	6.2
<b>Enhancer-gene linking (3)</b>			
Roadmap enhancer-gene linking	Correlation between Roadmap enhancers and gene expression across 127 cell types <sup>29,30,32</sup>	8.	16.3
EpiMap enhancer-gene linking <sup>a</sup>	Correlation between EpiMap enhancers and gene expression across 833 cell types <sup>37</sup>	7.5	13.8
ABC <sup>a</sup>	Hi-C linked enhancers in 167 cell types <sup>34,38</sup>	7.7	17.3
<b>PCHi-C (2)</b>			
Jung PCHi-C	PCHi-C in 27 cell types <sup>35</sup>	39.9	43.9
Javierre PCHi-C blood	PCHi-C in 17 blood cell types <sup>31</sup>	27.2	35.5
<b>scATAC-seq (1)</b>			
Cicero blood/basal <sup>a</sup>	Correlation between scATAC-seq peaks and gene promoter peaks across 61,806 blood/basal cells <sup>33,36</sup>	1.0	2.9

<sup>a</sup>ABC, activity by contact; GTEx, Genotype-Tissue Expression Project; PCHi-C, promoter capture Hi-C; scATAC-seq, single-cell sequencing assay for transposase-accessible chromatin; TSS, transcription start site. <sup>a</sup>Included in our cS2G strategy. For each of 13 main S2G strategies (in six categories), we provide a brief description and report the percentage of SNPs linked (proportion of common SNPs that are linked to at least one gene) and  $h^2$  coverage (meta-analyzed across 63 independent traits). When combining S2G strategies, we did not include the three non-functionally informed strategies because a fundamental goal of cS2G is to provide functional interpretation of GWAS findings. A description of all 50 S2G strategies analyzed is provided in Supplementary Table 1.

SNP-heritability enrichment (Supplementary Table 4), validating their informativeness for disease.

We constructed combined S2G strategies as linear combinations of linking scores from constituent S2G strategies (Fig. 1d). We developed an optimization framework maximizing the recall while constraining precision (defined using the training critical gene set) to be  $\geq 0.75$ ; high precision is important for maximizing the utility of functional follow-up studies. The precision and recall of the resulting combined S2G strategies were subsequently evaluated using the validation critical gene sets.

Further details are provided in Methods. We have released open-source software implementing our framework (see Code availability) and have made all S2G strategies, SNP annotations and critical gene sets analyzed publicly available (Data availability).

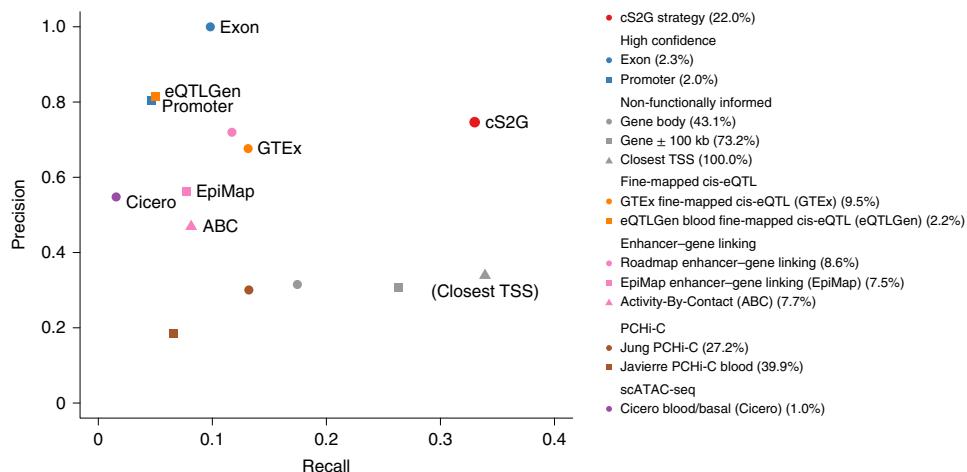
**Evaluation of S2G strategies.** We estimated the  $h^2$  coverage, precision and recall for the 50 S2G strategies (Supplementary Table 5); we used the (trait-specific) validation critical gene sets to perform these evaluations. Results for the 13 main S2G strategies are reported in Fig. 2 and Supplementary Fig. 2. The exon and promoter strategies attained high precision (1.00 for exons (by definition) and 0.80 for promoters) but low  $h^2$  coverage (0.06–0.10) and thus low recall (0.05–0.10). On the other hand, the closest TSS strategy attained low precision (0.34) but the highest  $h^2$  coverage (1.00) and recall (0.34). In addition to exon and promoter strategies, five other main S2G strategies attained high precision ( $>0.5$ ) but low recall (0.02–0.13): the two fine-mapped cis-eQTL strategies, the two enhancer-gene linking strategies informed by gene expression and the scATAC-seq strategy. Interestingly, S2G strategies using fine-mapped cis-eQTLs<sup>26</sup> attained significantly higher precision than S2G strategies using all

cis-eQTL (0.68  $\pm$  0.07 versus 0.40  $\pm$  0.04 in GTEx, 0.81  $\pm$  0.11 versus 0.29  $\pm$  0.03 in eQTLGen), consistent with previous reports of low precision for strategies using all cis-eQTL<sup>17,49</sup> and emphasizing the advantage of fine-mapped cis-eQTL for more precise analyses of GWAS data.

The 50 S2G strategies included 27 S2G strategies based on physical distance to the TSS. As expected, we observed that proximal closest TSSs are likely to implicate target genes (for example, precision of 0.78 for closest TSS range <1 kb), whereas distal closest TSSs are much less likely to implicate target genes (for example, precision of 0.15 for closest TSS range 100–500 kb) (Extended Data Fig. 2 and Supplementary Table 5). We further determined that the closest TSSs are moderately likely to systematically implicate target genes: precision of 0.34, decreasing to 0.17 for 2nd closest TSS and 0.062 for 5th closest TSS. The mean value of 0.043 for 6th–20th closest TSS suggests that genes located in the same regions as causal disease genes have a slightly elevated probability of being causal.

We next investigated whether functionally informed S2G strategies restricted to trait-specific tissues and cell types were more precise for the corresponding traits. We determined that S2G strategies defined using all available tissues and cell types achieved higher precision than S2G strategies restricted to blood and immune cell types (Extended Data Fig. 3) in analyses restricted to 11 autoimmune diseases and blood cell traits (average  $N=257$  K; Supplementary Table 3), perhaps due to limited biosample size; these results support including all available cell types in current efforts to pinpoint disease genes<sup>17,27</sup> (Discussion).

Lines of evidence to further validate our precision metric are provided in the Supplementary Note and Supplementary Figs. 3 and 4.



**Fig. 2 | Accuracy of individual S2G strategies and the 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). Colored font denotes the cS2G strategy and its seven constituent S2G strategies (gray font in parentheses denotes the closest TSS strategy). Numbers in parentheses in legend denotes the proportion of common SNPs that are linked to at least one gene (as in Table 1). We note that our evaluation of these S2G strategies is impacted by their widely varying underlying biosample sizes (see Methods), in addition to differences in functional assays and S2G linking methods. Standard errors are reported in Supplementary Figure 2, and numerical results are reported in Supplementary Table 5; standard errors for all S2G strategies linking >2.5% of common SNPs were  $\leq 0.12$  for precision and  $\leq 0.03$  for recall, with smaller standard errors for S2G strategies linking larger proportions of common SNPs.

In summary, we developed and validated a polygenic framework for evaluating S2G strategies, inferring high precision but low recall for many functionally informed S2G strategies and low precision but relatively high recall for non-functionally informed strategies (such as closest TSS).

**Combining S2G strategies.** We constructed an optimal cS2G by considering linear combinations of linking scores from ten functionally informed main S2G strategies (Table 1), maximizing the recall while constraining precision to be  $\geq 0.75$  using the (non-trait-specific) training critical gene set. The resulting cS2G strategy included seven constituent S2G strategies: exon, promoter, eQTLGen blood fine-mapped *cis*-eQTL, GTEX fine-mapped *cis*-eQTL, EpiMap enhancer–gene linking, ABC and Cicero blood/basal (ordered from highest to lowest weight; Supplementary Table 6). The cS2G strategy linked 22% of common SNPs (minor allele frequency (MAF)  $\geq 5\%$ ) to at least one gene and 99.6% of genes to at least one common SNP (average of 1.20 genes per linked common SNP and 79 common SNPs per linked gene). Despite the high weights for exon and promoter, 43% of linked common SNPs were not linked to the gene with closest TSS; the mean distance to the gene TSS for all cS2G links involving common SNPs was 96 kb (Supplementary Fig. 5). The number of common SNPs linked per gene was correlated to gene-body length ( $r^2 = 0.13$ ) but uncorrelated to gene-body length after correcting for exon and promoter length ( $r^2 = 0.01$ ). Strikingly, only 18% (respectively (resp.) 3.3%) of the cS2G links were supported by at least two (resp. 3) of seven constituent S2G strategies (Supplementary Table 7), consistent with the low correlations between the constituent S2G strategies (Supplementary Table 2). This provides a strong motivation for combining multiple S2G strategies.

We evaluated the cS2G strategy using the (trait-specific) validation critical gene sets and meta-analyzed the results across the 63 independent diseases and complex traits. The cS2G strategy attained  $h^2$  coverage of 0.44 (standard error (s.e.) 0.01), precision of 0.75 (s.e. 0.06) and recall of 0.33 (s.e. 0.03) (Fig. 2 and Supplementary Table 5), more than doubling the precision and/or recall of any individual strategy; this implies that 33% of SNP heritability can be linked to causal genes with 75% confidence using cS2G. Notably,

cS2G attained much higher precision than two previously proposed combined strategies, GeneHancer<sup>39</sup> (0.14) and Open Targets<sup>42</sup> (0.33) (Supplementary Table 5). Additional experiments comparing cS2G strategy to alternative approaches for combining S2G strategies are described in the Supplementary Note and Supplementary Tables 8–11.

In summary, we constructed an optimal cS2G, incorporating seven constituent S2G strategies, which more than doubles the precision and/or recall of any individual strategy; our evaluation of cS2G was based on validation critical gene sets that were distinct from the training critical gene set used to optimize cS2G.

**Validating the cS2G strategy using curated examples.** We sought to validate the cS2G strategy using a manually curated list of 17 disease-associated loci (including 12 loci from ref. <sup>50</sup>) containing 25 experimentally validated causal SNP–gene pairs (Table 2), and reflecting the very limited set of experimentally validated disease-associated SNP–gene pairs<sup>50</sup>. 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 to maximize the utility of functional follow-up studies; 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. However, we failed to identify the well-studied rs1421085–IRX5/IRX3 link<sup>12</sup>, 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<sup>51</sup>, as rs12740374

**Table 2 | Validation of the cS2G strategy using experimentally validated causal SNP–gene pairs associated to disease**

Position (hg19)	SNP	Gene	Disease/trait	cS2G prediction		
				Gene	Score	Annotations
1:109,817,590	rs12740374	SORT1	LDL <sup>50,51</sup>	<b>CELSR2</b> <sup>a</sup>	0.91	Exon
1:162,020,969	rs7539120	NOS1AP	QT interval <sup>63</sup>	-	-	-
2:60,718,043	rs1427407	BCL11A	Fetal hemoglobin level <sup>50,64</sup>	-	-	-
2:60,725,451	rs7606173			<b>BCL11A</b>	1.00	EpiMap, ABC, Cicero
3:38,767,315	rs6801957	SCNSA	QRS prolongation <sup>65</sup>	<b>SCNSA</b>	1.00	EpiMap, ABC
4:90,674,431	rs356168	SNCA	Parkinson's disease <sup>50,66</sup>	<b>SNCA</b>	1.00	EpiMap, ABC
5:56,031,822	rs17432750	MAP3K1	Breast cancer <sup>50,67</sup>	<b>MAP3K1</b>	0.66	EpiMap
5:56,052,695	rs62355900			-	-	-
5:56,053,479	rs74345699			-	-	-
5:56,134,276	rs16886397			<b>MAP3K1</b>	1.00	ABC
6:12,903,957	rs9349379	EDN1	Vascular diseases <sup>68</sup>	<b>PHACTR1</b>	1.00	GTEX
6:105,706,878	rs1743292	BVES	Cardiac QT interval and QRS duration <sup>50,69</sup>	-	-	-
6:105,720,538	rs1772203			<b>POPDC3</b>	1.00	ABC
6:117,210,052	rs339331	RFX6	Prostate cancer <sup>50,70,71</sup>	<b>RFX6</b>	1.00	EpiMap, ABC
6:135,418,635	rs7775698			<b>MYB</b>	1.00	EpiMap
6:135,418,637	rs66650371	MYB	Fetal hemoglobin level <sup>50,72</sup>	-	-	-
6:135,431,640	rs9494142			<b>HBS1L</b>	1.00	ABC
7:12,284,008	rs1990620	TMEM106B	Frontotemporal dementia <sup>50,73</sup>	-	-	-
8:11,351,220	rs1382568	BLK	Systemic lupus <sup>50,74</sup>	-	-	-
8:11,351,912	rs922483			<b>BLK</b>	0.76	Exon, ABC, Cicero
8:81,290,387	rs2370615	PAG1	Allergy <sup>50,75</sup>	<b>ZBTB10</b>	0.51	ABC, Cicero
10:12,307,894	rs11257655	CAMK1D	Type 2 diabetes <sup>50,76</sup>	<b>CAMK1D</b>	0.90	GTEX, EpiMap
10:6,094,697	rs61839660	IL2RA	Inflammatory bowel disease <sup>77</sup>	<b>IL2RA</b>	1.00	ABC, Cicero
16:53,800,954	rs1421085	IRX5/IRX3	Obesity <sup>12,50</sup>	-	-	-
20:55,990,405	rs737092	RBM38	Red blood cell count <sup>78</sup>	<b>RBM38</b>	0.96	eQTLGen, ABC

<sup>a</sup>rs12740374 was linked to SORT1 using GTEX with a cS2G linking score of 0.07 (which is less than 0.5). Predicted genes that match the experimentally validated gene are denoted in bold font. For each of 25 experimentally validated causal SNP–gene pairs at 17 disease-associated loci, we report the cS2G predictions: predicted gene with cS2G linking score >0.5 (if applicable), corresponding cS2G linking score, and constituent S2G annotation(s). Further details are provided in Supplementary Table 12. GTEX, GTEX fine-mapped *cis*-eQTL; eQTLGen, eQTLGen blood fine-mapped *cis*-eQTL; EpiMap, EpiMap enhancer-gene linking; Cicero: Cicero blood/basal; LDL, low-density lipoprotein.

is an exonic SNP for *CELSR2*, outweighing the link to *SORT1* by the GTEX fine-mapped *cis*-eQTL strategy.

We obtained similar results when analyzing a larger curated list consisting of disease genes validated with high confidence without strictly requiring experimental validation<sup>42</sup> (Supplementary Note and Supplementary Tables 14–16).

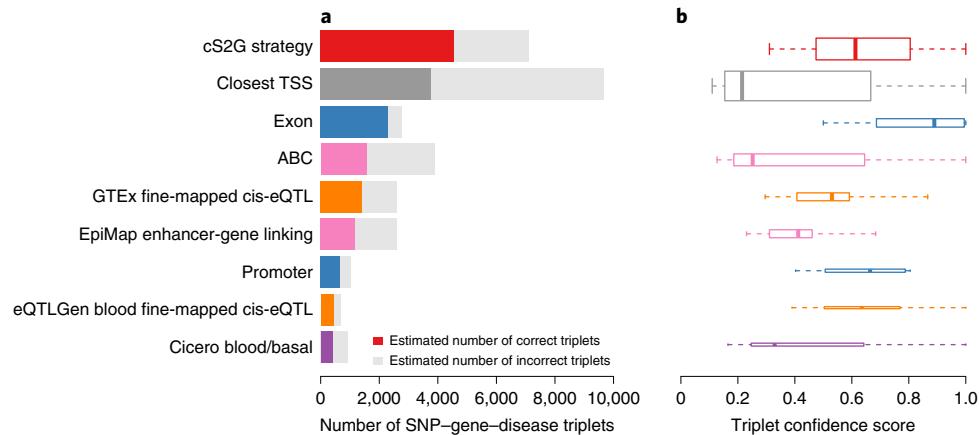
In summary, these analyses provide a promising validation of the potential of cS2G to pinpoint causal disease genes.

**Leveraging the cS2G strategy to pinpoint disease genes.** In previous work, we showed that functionally informed fine-mapping under the PolyFun framework robustly identifies a large number of fine-mapped SNPs for 49 UK Biobank diseases/traits (using  $N=337\text{K}$  unrelated British samples<sup>7,43</sup>) with well-calibrated posterior inclusion probabilities (PIPs)<sup>7</sup>. Here, we leveraged the high precision and recall of our cS2G strategy to predict target genes of 9,670 predicted causal SNP–disease pairs with PIP > 0.5 from PolyFun analyses (7,675 unique SNPs) (Supplementary Table 17).

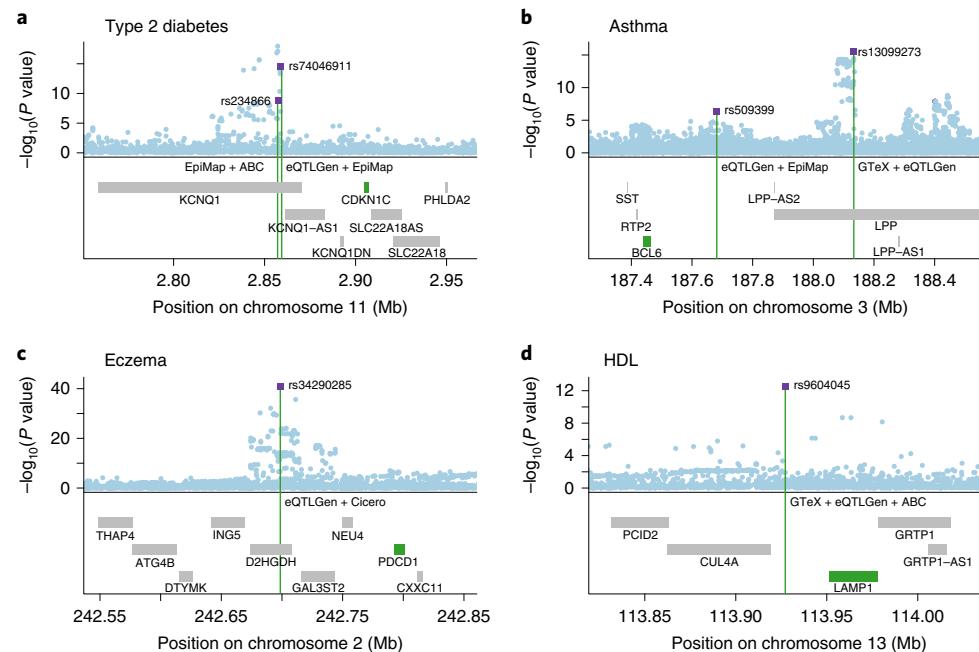
Restricting to SNPs that had a linked gene with cS2G linking score >0.5, we predicted 7,111 causal SNP–gene–disease triplets (5,384 unique SNPs; 3,401 unique genes) (Fig. 3a and Supplementary Table 17). The mean PIP of the 7,111 SNP–gene–disease triplets was equal to 0.80. We further assigned a confidence

score to each of the triplets by multiplying the PIP of the constituent SNP and its precision (Methods and Supplementary Tables 17 and 18). The average confidence score was 0.64, implying that 64% of the 7,111 SNP–gene–disease triplets (4,554 triplets) predict the correct causal SNP and target gene (Fig. 3a and Supplementary Table 18). Distributions of confidence scores are reported in Fig. 3b. Using cS2G, we detected 5,095 triplets with confidence score >0.5 (72% of its detected triplets), which is 1.84 times larger than what could be attained using any individual functionally informed constituent S2G strategy (only 2,763 for exon, even though the exon strategy had the highest confidence scores, consistent with its high precision) and 1.58 times higher than what could be attained using the closest TSS strategy (Fig. 3b and Supplementary Table 18). In many instances, multiple causal SNPs were linked to the same gene (for example, 119 genes were each linked to at least five different fine-mapped SNPs for five different diseases/traits; Supplementary Table 19), implying that a single gene can be causal for different diseases/traits using different causal SNP–gene links; we note that SNPs linked to the same gene are generally not in LD with each other (mean  $r^2=0.09$  across 3,900 pairs).

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



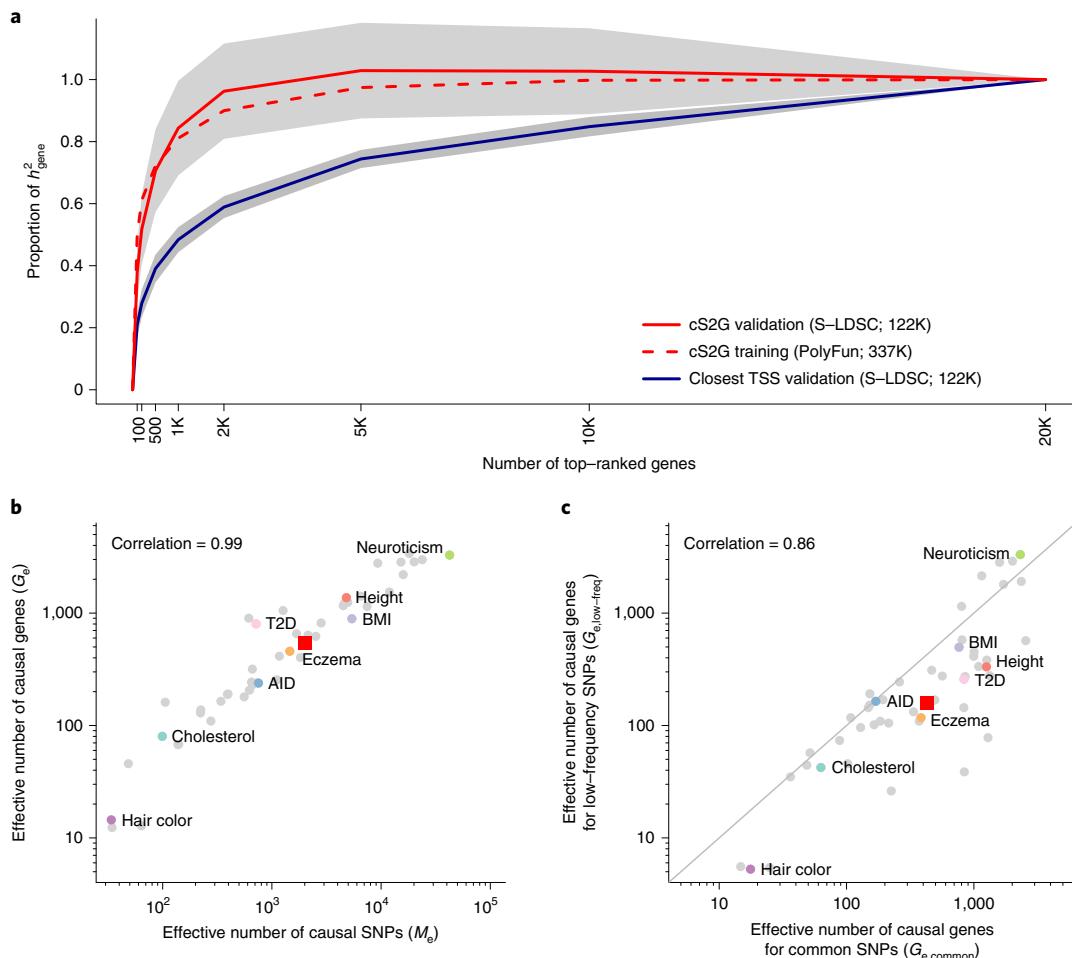
**Fig. 3 | SNP-gene-disease triplets identified by cS2G and other S2G strategies. a,** We report the number of SNP–gene–disease triplets identified by cS2G, its seven constituent strategies and the closest TSS S2G strategy. For each strategy, we estimated the number of correct triplets based on the mean confidence score across triplets; the estimated number of correct triplets is denoted as a colored bar, and the estimated number of incorrect triplets is denoted as a gray bar. **b,** We report the distribution of confidence scores of SNP–gene–disease triplets for each S2G strategy. The median value of confident scores is displayed as a band inside each box; boxes denote values in the second and third quartiles; the length of each whisker is 1.5 times the interquartile range, defined as the width of each box; the height of each box is proportional to the total number of triplets linked by each strategy (7,111, 9,664, 2,763, 3,889, 2,589, 2,604, 1,029, 674 and 943 for the nine plotted S2G strategies). The list of SNP–gene–disease triplets predicted by cS2G is reported in Supplementary Table 17. Numerical results are reported in Supplementary Table 18.



supported by at least two of the functionally informed constituent S2G strategies used by cS2G (Supplementary Table 17). We highlight four examples (Fig. 4). Briefly, these examples included instances where two independent fine-mapped SNPs were linked to the same gene (Fig. 4a and Fig. 4b) and instances where the functionally informed constituent S2G strategies either implicated (Fig. 4c) or failed (Fig. 4d) to implicate a plausible critical tissue/cell type

(Supplementary Table 20), highlighting 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. Further details of each example are provided in the Supplementary Note.

In summary, our cS2G strategy predicted 5,095 causal SNP–gene–disease triplets with high confidence, far more than any other S2G strategy; these triplets represent one of the largest high-confidence



**Fig. 5 | Empirical assessment of disease omnigenicity using cS2G.** **a**, We report the proportion of SNP heritability linked to genes ( $h_{\text{gene}}^2$ ) explained by genes ranked by top per-gene  $h^2$ , as inferred using three approaches (see text). Gray shading denotes 95% confidence intervals for cS2G validation and closest TSS validation around meta-analyzed values. We forced the standard error of the proportion of  $h_{\text{gene}}^2$  explained by all genes to be 0 (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. Results were meta-analyzed across 16 independent UK Biobank traits. **b**, We report the effective number of causal SNPs<sup>54</sup> ( $M_e$ ) and the effective number of causal genes ( $G_e$ ) for 49 UK Biobank diseases/traits, with representative traits in colored font. **c**, We report the effective number of causal genes for per-gene  $h^2$  linked to common SNPs ( $G_{e,\text{common}}$ ) and the effective number of causal genes for per-gene  $h^2$  linked to low-frequency SNPs ( $G_{e,\text{low-freq}}$ ) for 49 UK Biobank diseases/traits, with representative traits in colored font. In **b** and **c**, red squares denote median values across 16 independent traits and correlations are computed on log-scale values. Numerical results are reported in Supplementary Tables 22 and 24. AID, autoimmune disease; BMI, body mass index; cholesterol, total cholesterol; T2D, type 2 diabetes.

SNP–gene–disease resources with S2G-derived functional interpretation to date.

**Leveraging the cS2G strategy to assess disease omnigenicity.** Previous work has proposed an ‘omnigenic model’ in which all genes expressed in disease-critical cell types impact the function of core disease genes and thus impact disease SNP heritability<sup>44–46</sup>. This work raises intense interest in estimating components of disease SNP heritability contributed by each gene, but this has yet to be empirically assessed, due to the challenges of linking SNPs to genes. We leveraged our cS2G strategy to estimate the SNP heritability causally explained by SNPs linked to each gene for 49 UK Biobank diseases/traits (Supplementary Table 21) with functionally informed genome-wide fine-mapping results (not restricted to GWAS loci) available for all  $\text{MAF} \geq 0.1\%$  SNPs<sup>7</sup>. We re-estimated the SNP heritability linked to ranked gene sets by running S-LDSC<sup>11,52,53</sup> on summary statistics computed from  $N = 122$  K European-ancestry UK Biobank validation samples that were distinct from the training

samples used for fine-mapping (to avoid winner’s curse; Methods and ref. <sup>7</sup>). We also estimated the effective number of causal genes ( $G_e$ ) for each trait (using fourth moments of per-gene effects, analogous to  $M_e$  for SNPs<sup>54</sup>), and estimated this quantity separately for per-gene  $h^2$  linked to common ( $\text{MAF} \geq 5\%$ ) and low-frequency ( $0.1\% \leq \text{MAF} < 5\%$ ) SNPs.

The top 200 (resp. top 2,000) genes explained  $52\% \pm 6\%$  (resp.  $96\% \pm 8\%$ ) of the disease SNP heritability linked to genes in *cis* using the cS2G strategy ( $h_{\text{gene}}^2$ , which captures  $53\% \pm 3\%$  of  $h^2$ ), meta-analyzed across a subset of 16 independent diseases/traits (Fig. 5a and Supplementary Tables 22 and 23). Estimates directly based on the  $N = 337$  K training samples used for fine-mapping were very similar, implying minimal effects of winner’s curse (Fig. 5a). Results were similar when restricting to genes expressed in disease-critical cell types<sup>55</sup> (as proposed in ref. <sup>44</sup>) (Supplementary Fig. 6) and when using a restricted set of  $N = 49$  K validation samples that were unrelated to the  $N = 337$  K training samples (Supplementary Fig. 7). Interestingly, repeating the analysis using the closest TSS S2G

strategy implicated a far more polygenic gene-level architecture that required the top 1,000 (resp. top 10,000) genes to explain  $48\% \pm 2\%$  (resp.  $85\% \pm 2\%$ ) of  $h_{gene}^2$  (Fig. 5a and Supplementary Table 22); these results demonstrate the benefits of using more precise S2G strategies to infer more accurately gene-level architectures. We caution that the primary analysis using cS2G may still slightly overestimate gene-level polygenicity, because even the cS2G strategy is not perfectly precise. We further caution that all our findings pertain to effects of SNPs in *cis* (Discussion), potentially leading to underestimation of gene-level polygenicity.

We estimated the effective number of causal genes ( $G_e$ ) for each trait. Estimates of  $G_e$  varied widely, from 3,289 (neuroticism) to 1,375 (height) to 80 (total cholesterol) with a median of 540 (across 16 independent traits) and were strongly correlated (log-scale  $r=0.99$ ) to estimates of the effective number of independently associated SNPs ( $M_e$ ; median = 1,991), a SNP-based measure of disease/trait polygenicity<sup>54</sup> (Fig. 5b and Supplementary Table 24; the strong correlation provides a validation of both  $G_e$  and  $M_e$ ).

We further estimated  $G_{e,common}$  (resp.  $G_{e,low-freq}$ ) by restricting per-gene  $h^2$  explained by causal SNPs to the SNP heritability causally explained by common (resp. low-frequency) SNPs linked to each gene. Gene-level architectures were more polygenic for common versus low-frequency SNPs, with median  $G_{e,common}$  of 427 vs. median  $G_{e,low-freq}$  of 157 (median ratio of 2.8) across the 16 independent traits (Fig. 5c and Supplementary Table 21), consistent with more polygenic SNP architectures for common versus low-frequency SNPs due to the action of negative selection ( $M_{e,common}/M_{e,low-freq}=3.9$  in ref. <sup>54</sup>). Surprisingly, there was low concordance between genes underlying gene-level architectures for common versus low-frequency SNPs (Extended Data Fig. 4 and Supplementary Tables 24 and 25). However, we observed consistent excess overlap between the top 200 genes contributing to  $h_{gene,common}^2$  (resp.  $h_{gene,low-freq}^2$ ) and two disease-specific gene sets<sup>56,57</sup> (Extended Data Figs. 5 and 6), suggesting that common and low-frequency variant gene-level architectures are driven by different genes pertaining to similar biological processes.

In summary, our cS2G strategy provided a quantitative assessment of the ‘omnigenic model’<sup>44–46</sup>, implicating a far less polygenic gene-level architecture than the closest TSS S2G strategy. We inferred a more polygenic gene-level architecture for common variants as compared to low-frequency variants, with little overlap between genes underlying these gene-level architectures despite shared biological processes.

## Discussion

We developed a polygenic framework to evaluate and combine S2G strategies; in particular, our framework is a substantial advance over previous approaches for evaluating S2G strategies using curated lists of disease-associated SNP–gene pairs<sup>42,48</sup> (Supplementary Note). We applied our framework to construct a cS2G that achieved a precision of 0.75 and a recall of 0.33, more than doubling the precision and/or recall of any individual strategy. We applied cS2G to fine-mapping results for 49 UK Biobank diseases/traits to predict 5,095 causal SNP–gene–disease triplets (with S2G-derived functional interpretation) with high confidence, including 2,163 triplets involving distal regulatory fine-mapped SNPs that were not in the gene body (or promoter) of the target gene; notable examples included *CDKN1C* in type 2 diabetes, *BCL6* in asthma, *PDCD1* in eczema and *LAMP1* in high-density lipoprotein, all of which were supported by multiple S2G strategies. We further applied cS2G to provide a quantitative assessment of the ‘omnigenic hypothesis’<sup>44–46</sup>, concluding that the top 200 (1%) of ranked genes explained roughly half of the SNP heritability linked to all genes; this implies that gene-level architectures in *cis* are largely driven by a relatively modest number of top genes.

Our findings have several implications for downstream analyses. First, we recommend that GWAS fine-mapping studies use cS2G to powerfully link fine-mapped SNPs to their target genes;

we note that, as with previous S2G approaches, cS2G can be combined with similarity-based approaches leveraging genome-wide patterns of associated genes (as proposed in ref. <sup>48</sup>). Second, our framework can be used to optimize (and combine) S2G strategies that may be developed in the future; the development of new S2G strategies remains a key priority, as our cS2G strategy—despite its large improvement over other S2G strategies—attained a modest recall of 33%, implying that only 1/3 of disease SNP heritability can be explained by causal disease SNPs linked to their correct target genes. Third, our results highlight the advantages of enhancer-gene linking strategies such as EpiMap and ABC in future efforts to improve S2G strategies; their advantages include cost effective experiments targeting multiple cell types (EpiMap and ABC provide links for 833 and 167 cell types, respectively) and high potential for linking rare variants to genes. Fourth, our findings support the hypothesis that rare variant association studies<sup>58,59</sup> will provide biological insights complementary to those of GWAS, both because we observed little overlap between genes underlying common variant and low-frequency variant gene-level architectures and because we determined that low-frequency variant gene-level architectures were less polygenic (median ratio of 2.8); we expect these differences to be even more pronounced for rare variant architectures. Finally, cS2G can improve identification of gene sets that are enriched for disease SNP heritability (although it has not been optimized for this specific purpose; Supplementary Fig. 8, also see ref. <sup>55</sup>); we further note the importance of including appropriate SNP annotations in the model used by S-LDSC in analyses of enriched gene sets to avoid biased enrichment estimates (Methods and Supplementary Fig. 9). Investigating the relative performance of different combined S2G strategies in analyses of gene sets that are enriched for disease SNP heritability is a direction for future research.

We note several limitations of our work. We included all available tissues and cell types in the constituent S2G strategies of cS2G, as we observed that this led to higher precision (Extended Data Fig. 3), perhaps due to limited biosample size. However, S2G links involving disease-critical tissues/cell types are central to understanding biological mechanisms (Fig. 4, Supplementary Table 20 and 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. Additional limitations are discussed in the Supplementary Note. Despite these limitations, our results convincingly demonstrate both the advantages of using our polygenic framework to evaluate and combine S2G strategies and the effectiveness of using our cS2G strategy to pinpoint disease genes.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-022-01087-y>.

Received: 29 July 2021; Accepted: 27 April 2022;  
Published online: 6 June 2022

## References

- Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- 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).
- Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).

4. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
5. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
6. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
7. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
8. Hindorff, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* **106**, 9362–9367 (2009).
9. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
10. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
11. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
12. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
13. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
14. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
15. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
16. Porcu, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 3300 (2019).
17. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
18. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
19. Väistö, 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).
20. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
21. Lee, D. et al. JEPAGE: a summary statistics based tool for gene-level joint testing of functional variants. *Bioinformatics* **31**, 1176–1182 (2015).
22. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
23. Chun, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. *Nat. Genet.* **49**, 600–605 (2017).
24. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768–769 (2019).
25. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 10 (2015).
26. 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).
27. 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).
28. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
29. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
30. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
31. Javierre, B. M. et al. Lineage-Specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
32. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol.* **18**, 193 (2017).
33. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
34. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
35. Jung, I. et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442–1449 (2019).
36. 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).
37. 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).
38. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
39. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **2017**, bax028 (2017).
40. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
41. GEMO Study Collaborators et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat. Genet.* **52**, 56–73 (2020).
42. 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).
43. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
44. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
45. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common disease is more complex than implied by the core gene omnigenic model. *Cell* **173**, 1573–1580 (2018).
46. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034.e6 (2019).
47. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
48. Weeks, E. M. et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. Preprint at medRxiv <https://doi.org/10.1101/2020.09.08.20190561> (2020).
49. Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).
50. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
51. Musunuru, K. et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
52. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
53. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).
54. O'Connor, L. J. et al. Extreme polygenicity of complex traits is explained by negative selection. *Am. J. Hum. Genet.* **105**, 456–476 (2019).
55. 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. Preprint at bioRxiv <https://doi.org/10.1101/2021.03.19.436212> (2021).
56. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
57. Freund, M. K. et al. Phenotype-Specific enrichment of mendelian disorder genes near GWAS Regions across 62 Complex Traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).
58. Povysil, G. et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat. Rev. Genet.* **20**, 747–759 (2019).
59. Van Hout, C. V. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
60. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
61. Yates, A. D. et al. Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).
62. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
63. Kapoor, A. et al. An enhancer polymorphism at the cardiomyocyte intercalated disc protein NOS1AP locus is a major regulator of the QT interval. *Am. J. Hum. Genet.* **94**, 854–869 (2014).
64. Bauer, D. E. et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* **342**, 253–257 (2013).
65. van den Boogaard, M. et al. A common genetic variant within SCN10A modulates cardiac SCN5A expression. *J. Clin. Invest.* **124**, 1844–1852 (2014).
66. Soldner, F. et al. Parkinson-associated risk variant in distal enhancer of α-synuclein modulates target gene expression. *Nature* **533**, 95–99 (2016).
67. Glubb, D. M. et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am. J. Hum. Genet.* **96**, 5–20 (2015).
68. Gupta, R. M. et al. A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell* **170**, 522–533 (2017).
69. Wang, X. et al. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *eLife* **5**, e10557 (2016).

70. 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).
71. 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).
72. Stadhouders, R. et al. HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. *J. Clin. Invest.* **124**, 1699–1710 (2014).
73. Gallagher, M. D. et al. A dementia-associated risk variant near TMEM106B alters chromatin architecture and gene expression. *Am. J. Hum. Genet.* **101**, 643–663 (2017).
74. Guthridge, J. M. et al. Two functional lupus-associated BLK Promoter variants control cell-type- and developmental-stage-specific transcription. *Am. J. Hum. Genet.* **94**, 586–598 (2014).
75. Vicente, C. T. et al. Long-range modulation of PAG1 expression by 8q21 allergy risk variants. *Am. J. Hum. Genet.* **97**, 329–336 (2015).
76. 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 Genet.* **10**, e1004633 (2014).
77. Simeonov, D. R. et al. Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).
78. Ulirsch, J. C. et al. Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

## Methods

**Ethics statement.** This study relied on analyses of publicly available genetic and genomic data sets and so did not require ethical approval.

**S2G strategies.** An S2G linking strategy  $k$  is defined as an assignment of a raw linking value  $A_{k,j,g}$  between each SNP  $j$  and zero or more candidate target genes  $g$ , which we converted into a linking score  $\psi_{k,j,g}$  such that each SNP has a sum of linking scores  $\leq 1$  (we allowed  $\sum_g \psi_{k,j,g} < 1$  to allow for incomplete S2G linking; see below). We considered only links related to a list of 19,995 genes, including 17,871 protein-coding genes, that pass our quality control procedure (Data availability). Specifically, we selected genes in Ensembl<sup>61</sup> (grch37, accessed on 2019-04-24), Gencode<sup>60</sup> (release 19), and RefSeq<sup>62</sup> (refGene, version 2017-03-08) databases that have a unique identifier in all the data sets, overlapping starting and ending gene positions, and similar strand information. We verified that restricting our analysis to 17,871 protein-coding genes (instead of 19,995 protein-coding and non-protein-coding genes) had little impact on our results (Supplementary Fig. 1).

We considered 50 S2G strategies (Table 1, Supplementary Table 1 and Supplementary Note); in each case, we first considered raw linking values  $A_{k,j,g}$  which we next converted into linking scores  $\psi_{k,j,g}$ . For all but one strategy (Hi-C distance; Supplementary Note), we converted raw linking values  $A_{k,j,g}$  (as defined above) into linking scores  $\psi_{k,j,g}$  such that each SNP has a sum of linking scores over genes being 0 or 1 (linking score  $\psi_{k,j,g}$  should not be interpreted as probabilities). We note that some S2G strategies include instances of SNP–gene links with low linking scores, based on the definitions of raw linking values for each strategy. When an S2G strategy linked a SNP to multiple genes, we restricted each S2G strategy such that each SNP was restricted to the gene with the highest linking score (regardless of whether this linking score was high or low in absolute terms; no specific threshold), as we observed that this led to slightly higher precision (Extended Data Fig. 1).

Correlations and overlap proportions between the S2G strategies were computed on all SNP–gene links observed by at least one of 34 S2G strategies (we omitted the sixth closest TSSs to 20th closest strategy and Hi-C due to computational constraints) (Supplementary Table 2). Overlap proportion between a strategy  $k$  and a strategy  $k'$  is defined as the proportion of SNP–gene links reported by strategy  $k$  that are also reported by strategy  $k'$  (these values are not symmetric). We defined a subset of 13 independent S2G strategies (different from the 13 main strategies) with  $r^2 < 0.1$  when comparing  $h^2$  coverage, precision or recall estimates.

We note that the functionally informed S2G strategies are derived from functional assays with widely varying biosample sizes. For example, PCHi-C data sets used 17/27 cell types, enhancer maps such as EpiMap or ABC used multiple functional assays for 127 and 833 cell types, and *cis*-eQTLs such as GTEx and eQTLGen used 17,382 and 31,684 RNA-sequencing samples. Thus, our evaluation of S2G strategies should not be viewed as an evaluation of the underlying functional assays.

**Evaluation of S2G strategies.** To evaluate each S2G strategy's informativeness for pinpointing disease genes, we aimed to define and estimate parameters that correspond to an S2G strategy's  $h^2$  coverage (proportion of total disease SNP heritability that is linked to genes), precision (proportion of linked disease SNP heritability that is linked to the correct target gene) and recall (proportion of total disease SNP heritability that is linked to the correct target gene).

First, we defined  $h^2$  coverage as the proportion of SNP heritability explained by all SNPs linked to one or more genes (weighted by their linking scores):

$$h^2 \text{ coverage}(k) = h^2_{(\text{genes:all,S2G:k})}/h^2 \quad (1)$$

where  $h^2_{(\text{genes:all,S2G:k})}$  is the SNP heritability explained by common SNPs linked all genes using  $k$  and  $h^2$  is the SNP heritability explained by common SNPs.

Second, we defined precision as the relative excess SNP heritability enrichment of SNPs linked to a critical gene set (see below) versus SNPs linked to all genes as compared to the (gold standard) exon S2G strategy. More precisely, the precision of an S2G strategy  $k$  was defined as

$$\text{precision}(k) = \frac{\text{geneenrichment}(k) - 1}{\text{geneenrichment}(\text{exon}) - 1} \quad (2)$$

with

$$\text{geneenrichment}(k) = \frac{h^2_{(\text{genes:critical,S2G:k})}/M_{(\text{genes:critical,S2G:k})}}{h^2_{(\text{genes:all,S2G:k})}/M_{(\text{genes:all,S2G:k})}} \quad (3)$$

where  $h^2_{(\text{genes:critical,S2G:k})}$  is the SNP heritability explained by common SNPs linked to the critical gene set using  $k$ ,  $M_{(\text{genes:critical,S2G:k})}$  is the number of common SNPs linked to the critical gene set using  $k$  and  $M_{(\text{genes:all,S2G:k})}$  is the number of common SNPs linked to all genes using  $k$ . We note that this definition relies on the hypothesis that genes in the critical gene set are enriched for causal disease genes (as observed empirically, see below) and the hypothesis that the exon S2G strategy

is a perfectly precise strategy (even though it suffers from low  $h^2$  coverage). Third, we defined recall as the product of the  $h^2$  coverage and precision.

We estimated these quantities using polygenic analyses of disease SNP heritability by applying S-LDSC (v1.0.1) with the baseline-LD model (v2.2)<sup>11,52,53</sup> to 63 independent diseases and complex traits (average  $N = 314$  K; Supplementary Table 3), meta-analyzing results across traits. All traits had  $z\text{-score} > 6$  for non-zero SNP heritability, following previous recommendations<sup>52</sup>. We removed the major histocompatibility complex (MHC) region during the regression step because of its unusual LD patterns and genetic architecture<sup>11</sup>. We analyzed SNP annotations for ~10,000,000 SNPs with a minor allele count  $\geq 5$  in a 1000 Genomes Project European reference panel<sup>47</sup>. We jointly considered the 97 SNP annotations of the baseline-LD model v2.2 (refs. <sup>52,53</sup>), 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 19,995 genes (we did not include SNP annotations constructed using all genes for the 20 closest TSS S2G strategies, as these would include all SNPs), for a total of 177 SNP annotations. Jointly considering all these S2G-derived SNP annotations was crucial to maximize the accuracy of  $h^2_{\text{critical}\cap k}$  (Supplementary Fig. 9). This strongly demonstrates the importance of including appropriate SNP annotations in the model used by S-LDSC in analyses estimating the proportion of SNP heritability explained by enriched gene sets<sup>55</sup> to avoid biased enrichment estimates. We note that precisions from preliminary analyses of Extended Data Fig. 1 were estimated using the baseline-LD model and S2G-derived SNP annotations constructed from the exon, promoter, gene body, gene body  $\pm 100$  kb and closest TSS S2G strategies using this restricted set of S2G strategies attenuated the bias of  $h^2_{\text{critical}\cap k}$ .

We estimated values of  $h^2$  coverage and gene enrichment for each disease/trait, estimated their standard errors using a genomic block-jackknife with 200 blocks, meta-analyzed results across the 63 independent traits using a fixed-effect meta-analysis and used these values to estimate precision and recall. Precision and recall were estimated from meta-analyzed  $h^2$  coverage and gene-enrichment values (instead of meta-analyzing precision and recall across traits) to guarantee robust estimates. Precision and recall standard errors were estimated by using the 200  $h^2$  coverage and gene-enrichment estimates from the block-jackknife procedure but meta-analyzed using the same  $h^2$  coverage and gene-enrichment standard errors. We note that performing fixed-effect meta-analyses when computing overall estimates of precision/recall assigns low weights to traits with large standard errors. For example, including five traits with low SNP heritability ( $< 0.02$ , despite  $z\text{-score} > 6$  for non-zero SNP heritability) had little impact on our results; when removing these traits, estimates of precision changed from 0.747 (s.e. 0.061) to 0.753 (s.e. 0.062), and estimates of recall changed from 0.330 (s.e. 0.027) to 0.332 (s.e. 0.027).

**Training and validation critical gene sets.** Our definitions of precision and recall rely on a critical gene set. We used a non-trait-specific training critical gene set to construct an optimal cS2G and trait-specific validation critical gene sets to evaluate the optimal cS2G while avoiding overfitting (for comparison purposes, we also used the validation critical gene sets to evaluate individual S2G strategies). Training and validation critical gene sets rely on information from exons and promoters to guarantee high-confidence SNP–gene links.

We defined a non-trait-specific training critical gene set as the top 10% of genes with the most highly constrained exons and conserved promoters. Specifically, for each gene, we multiply its probability of being loss of function intolerant (or pLI score)<sup>79</sup> (estimating gene probability to be intolerant to loss-of-function mutations) by the fraction of bases of its promoter (defined using the promoter S2G strategy) that is conserved (defined using four baseline-LD conserved SNP annotations<sup>80,81</sup>) (note that 17,554 out of our 19,995 genes had a pLI score). As we observed a correlation between this score and gene-body length ( $r = 0.18$ ), we created ten bins of genes based on their gene-body length and selected the genes with the top 10% of this score (1,760 genes in total).

We defined the validation critical gene set for a given trait as the top 10% of genes ranked by the PoPS method<sup>48</sup> (we note that PoPS gene scores are based on a leave-one-chromosome-out approach, implying that gene scores should be independent of their surrounding SNPs). By default, the initial step of PoPS is to apply MAGMA<sup>82</sup> to compute gene-level association statistics, which relies on linking SNPs to genes using a gene-body S2G strategy. To limit the impact of the gene-body S2G strategy in our analyses, we modified PoPS to only link SNPs that are in exons or promoters (note that this led to nearly similar SNP heritability enrichment and gene-enrichment values; Supplementary Table 4); 16,728 out of our 19,995 genes had a PoPS score, leading to validation critical gene sets with 1,673 genes.

Across the 63 diseases/traits analyzed, 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) had a mean of 20% (versus 10% expected by chance), mean of 20%, standard deviation of 4.1% and range from 13%–28% (Supplementary Fig. 1). We also observed substantial excess overlap of housekeeping genes (1,997 genes) in the training critical gene set (357 genes; excess overlap = 2.0) and validation critical gene sets (median of 281 genes across 63 traits; median excess overlap = 1.7).

**Combining S2G strategies.** We constructed combined S2G strategies as linear combinations of linking scores from  $K$  S2G strategies. Specifically, for each SNP  $j$  and gene  $g$ , we computed a combined S2G linking score

$$\Psi_{j,g} \propto \sum_k w_k \times \psi_{k,j,g} \quad (4)$$

where  $\psi_{k,j,g}$  is the linking score between SNP  $j$  and gene  $g$  for S2G strategy  $k$  and  $w_k$  is the weight associated to strategy  $k$ . We used  $\Psi_{j,g} = \sum_k w_k \times \psi_{k,j,g}$  when  $\sum_g \sum_k w_k \times \psi_{k,j,g} < 1$  and  $\Psi_{j,g} = \sum_k w_k \times \psi_{k,j,g} / \sum_g \sum_k w_k \times \psi_{k,j,g}$  otherwise; allowing  $\sum_g \Psi_{j,g}$  to be  $< 1$  allows to give small combined S2G linking scores to SNPs with linking score available only for imprecise S2G strategies. Here, we allowed weights  $w_k$  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 (see below).

To estimate the optimal weights  $w_k$ , we developed an optimization framework to identify the weights maximizing the recall while constraining precision (defined using the training critical gene set) to be  $\geq 0.75$ . Indeed, providing high precision maximizes the utility of functional follow-up studies. First, we computed for each SNP its expected per-SNP heritability by meta-analyzing across the 63 independent traits the S-LDSC regression coefficients estimated with the baseline-LD model and the 80 S2G-derived SNP annotations of the training gene set and all genes. Second, we defined a function taking as input a vector of weights  $w$ , computed for each SNP the expected per-SNP heritability linked to the critical gene set and to all genes using the 80 S2G-derived SNP annotations and outputting precision and recall. Finally, we found the vector  $w$  maximizing recall while constraining precision to be  $\geq 0.75$ . (We note that the precision of 0.75 for the cS2G strategy computed using the validation critical gene sets is independent of the threshold of precision  $\geq 0.75$  in the training critical gene set used to optimize the cS2G strategy (estimated precision using the training critical gene set was equal to 0.81 during the optimization process; Supplementary Table 8).) Specifically, we considered a grid of values for  $w$ , going from 0 to 2.5 with a 0.1 step, and the values 5, 7.5, 10, 25, 50, 75 and 100 (33 total values). We created a custom optimization framework that (a) starts by giving a weight of 1 to the exon strategy and 0 to all other S2G strategies investigated, (b) computes precision and recall by increasing a single value of  $w$  at a time (the higher weight on the grid), and keeps for next step the weight vector maximizing recall (constraining precision to be  $\geq 0.75$ ), (c) computes precision and recall by decreasing a single value of  $w$  at a time (the lower weight on the grid) and keeps for next step the weight vector maximizing recall (constraining precision to be  $\geq 0.75$ ), (d) computes precision and recall by randomly modifying a single value of  $w$  at a time and keeps for next step the weight vector maximizing recall (constraining precision to be  $\geq 0.75$ ) and (e) restarts from (b) till the recall does not improve. We repeated this algorithm five times and kept the weight vector providing the maximum recall. We note that we investigated five different optimization algorithms from the R software (methods Nelder-Mead, BFGS, CG, L-BFGS-B and SANN from the optim function; R version 3.6.1 was used in our analyses), but none of them reached higher recall than our algorithm. When giving as a starting point the outputs of our custom algorithm, these five algorithms converged to nearly identical weight vectors and recalls.

We note that 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, in the case of our cS2G strategy (Supplementary Table 6), 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. However, we note that fixing the weights of the seven constituent S2G strategies of the cS2G strategy to the same value only slightly underperformed cS2G in both precision and recall (0.71 versus 0.75 and 0.31 versus 0.33, respectively; Supplementary Table 8), as expected given the low overlap of SNPs annotated by these seven strategies (Supplementary Table 7).

**Leveraging the cS2G strategy to pinpoint disease genes.** We analyzed 9,670 predicted causal SNP-disease pairs with PIP  $> 0.50$  from functionally informed fine-mapping of 49 UK Biobank diseases/traits using PolyFun + SuSiE using  $N=337$  K unrelated British UK Biobank samples<sup>7,43</sup>. SNPs in the MHC region were removed from the fine-mapping analyses<sup>7</sup>. We selected a PIP threshold of 0.5, as this threshold was carefully validated using simulations in ref.<sup>7</sup>, and a threshold of 0.5 was also used for cS2G linking score and triplet confidence score. For this purpose, we created S2G and cS2G strategies for  $\sim 19,000,000$  imputed UK Biobank SNPs with MAF  $\geq 0.1\%$  (we note that these analyses include both SNPs and indels, but we use the term SNP throughout the paper for simplicity). We predicted causal SNP-disease triplets by restricting to SNPs that had a linked gene with cS2G linking score  $> 0.5$  (98% of the cS2G linked SNPs). We note that the proportion of fine-mapped SNPs linked by the cS2G strategy ( $5,384/7,675 = 0.70$ ) was higher than its  $h^2$  coverage (0.44) due to the excess of exon and promoter SNPs

( $2,629/5,384$  SNPs are in exons or promoters, consistent with the use of functional priors in fine-mapping analyses).

We further assigned a confidence score to each SNP-gene-disease triplet with a cS2G linking score  $> 0.5$  by multiplying their corresponding PIP and precision. To account for the excess of exon and promoter SNPs, we assigned the precision of exon (that is, 1.00) if the link was validated through the exon S2G strategy, the precision of promoter (that is, 0.80) if the link was validated through the promoter S2G strategy and an estimated precision for SNPs that are not in exons or promoters (precision<sub>other</sub>) otherwise. We estimated precision<sub>other</sub> for an S2G strategy  $k$  using the following formula:

$$\begin{aligned} \text{precision}(k) = & \text{coverage}_{\text{exon}}(k) \times \text{precision}(\text{exon}) \\ & + \text{coverage}_{\text{promoter}}(k) \times \text{precision}(\text{promoter}) \\ & + \text{coverage}_{\text{other}}(k) \times \text{precision}_{\text{other}}(k) \end{aligned} \quad (5)$$

where  $\text{coverage}_{\text{exon}}$ ,  $\text{coverage}_{\text{promoter}}$  and  $\text{coverage}_{\text{other}}$  are the proportion of  $h^2$  coverage explained by SNPs in exons, promoters and SNPs that are not in exons or promoters, respectively.

To predict the candidate cell type of action for these triplets, we (1) focused on SNP-gene pairs provided by GTEx fine-mapped *cis*-eQTL, EpiMap enhancer-gene linking and/or ABC; (2) restricted to cell types where the SNP-gene pair has been observed (out of 54, 833 and 167, respectively); and (3) reported the cell type with the most significant regression coefficient in an S-LDSC analysis conditioned to the baseline model (as performed in refs. <sup>11,56</sup>).

**Leveraging the cS2G strategy to empirically assess disease omnigenicity.** To estimate the SNP heritability explained by SNPs linked to each gene  $g$  for 49 UK Biobank diseases/traits (per-gene SNP heritability  $h_{\text{gene},g}^2$ ), we used PolyFun + SuSiE estimates of posterior mean squared causal effect sizes for  $\sim 19,000,000$  imputed SNPs with MAF  $\geq 0.1\%$  ( $\beta_j^2$ ) estimated on  $N=337$  K unrelated British UK Biobank samples<sup>43</sup> (SNPs in the MHC region were removed from these analyses<sup>7</sup>). First, we computed unadjusted per-gene SNP heritability  $h_{\text{gene},g}^2 = \sum_j \Psi_{j,g} \times \hat{\beta}_j^2$ , where  $\Psi_{j,g}$  is the cS2G linking score between SNP  $j$  and gene  $g$ . Then, we computed (adjusted) per-gene SNP heritability  $h_{\text{gene},g}^2 = \sum_j \Psi_{j,g}^+ \times \hat{\beta}_j^2$  with

$$\Psi_{j,g}^+ = (\Psi_{j,g} \times h_{\text{gene},g}^2) / \sum_{g'} (\Psi_{j,g'} \times h_{\text{gene},g'}^2)$$

being a trait-specific cS2G linking score. The motivation of this additional step is to improve the cS2G linking scores of SNPs linked to multiple genes by integrating evidence from  $\hat{\beta}_j^2$  linked to a single gene; we note that this step changed per-gene SNP heritability for only a small number of genes, as most SNPs linked using cS2G have large cS2G linking score (87% of the linked SNPs have a maximum cS2G linking score  $> 0.95$ ) (Supplementary Fig. 11). We also estimated per-gene SNP heritabilities linked to common SNPs (MAF  $\geq 5\%$ ) as  $h_{\text{gene,common},g}^2 = \sum_{j \in \text{common SNPs}} \Psi_{j,g} \times \hat{\beta}_j^2$  and per-gene SNP heritabilities linked to low-frequency SNPs ( $0.1\% \leq \text{MAF} < 5\%$ ) as  $h_{\text{gene,low-freq},g}^2 = \sum_{j \in \text{low-freq SNPs}} \Psi_{j,g} \times \hat{\beta}_j^2$ . We note that these estimates of per-gene SNP heritabilities do not account for LD between probabilistically fine-mapped SNPs, leading to genome-wide underestimation of variance explained by SNPs; this underestimation could in principle vary across genes. However, this limitation does not impact our main conclusions, because (1) it impacts only estimates of per-gene SNP heritabilities derived from PolyFun + SuSiE and not S-LDSC estimates of proportions of  $h_{\text{gene}}^2$  explained by the resulting eight gene sets (see next paragraph) as reported in Fig. 5a, and (2) estimates of proportions of  $h_{\text{gene}}^2$  derived from PolyFun + SuSiE (based on  $N = 337$  K training samples) and S-LDSC estimates of proportions of  $h_{\text{gene}}^2$  (based on  $N = 122$  K new validation samples) were strongly concordant (Fig. 5a).

To estimate the proportion of SNP heritability linked to genes ( $h_{\text{gene}}^2$ ) explained by genes with the top per-gene SNP heritability, we partitioned genes into eight gene sets ranked by per-gene SNP heritability explained (top 100, 200, 500, 1,000, 2,000, 5,000, 10,000 and 19,995) and re-estimated the SNP heritability linked to each ranked gene set by running S-LDSC with the baseline-LD model on summary statistics computed from  $N = 122$  K European-ancestry UK Biobank samples that were distinct from the  $N = 337$  K British UK Biobank samples used to estimate  $\hat{\beta}_j^2$  and  $h_{\text{geneg}}^2$  (to avoid winner's curse, analogous to what was performed in ref. <sup>7</sup>). We defined  $h_{\text{gene}}^2$  as the SNP heritability linked to all 19,995 genes. The proportion of  $h_{\text{gene}}^2$  explained by the X top genes was computed as the proportion of SNP heritability explained by the X top genes divided by  $h_{\text{gene}}^2$ . The standard error of the proportion of  $h_{\text{gene}}^2$  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  $h_{\text{gene}}^2$  (viewing the denominator  $h_{\text{gene}}^2$  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). In Fig. 5a, we forced the s.e. of the proportion of  $h_{\text{gene}}^2$  explained by all genes (a quantity that must equal 1) to be 0.

We estimated the effective number of causal genes ( $G_e$ ) for each trait using per-gene SNP heritability and the formula of ref.<sup>34</sup>. Specifically, we defined  $G_e = 3G/\kappa$ , with G the total number of genes and

$$\kappa = E \left[ h_{\text{gene},g}^2 \right] / E \left[ h_{\text{gene},g}^2 \right]^2.$$

These estimates relied on per-gene SNP heritability explained by causal SNPs and were thus directly based on the  $N = 337$  K samples used for fine-mapping; as noted above, the impact of winner's curse on our analyses was minimal. We extended this formula to per-gene  $h^2$  linked to common and low-frequency SNPs to estimate  $G_{e,\text{common}}$  and  $G_{e,\text{low-freq}}$ , respectively. Per-gene  $h^2$  values were directly estimated on the  $N = 337$  K samples used for fine-mapping as we observed that the impact of winner's curse on our analyses was minimal.

**Statistics and reproducibility.** No statistical method was used to predetermine sample size. We restricted our analyses to well-powered ( $z\text{-score} > 6$  for non-zero SNP heritability<sup>53</sup>) GWAS data sets of European ancestry. We removed the MHC region during the S-LDSC regression step<sup>11</sup>. We did not use any study design that required randomization or blinding.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The list of 19,995 genes, summary statistics of the 63 independent traits, training and validation critical gene sets, S2G and cS2G strategies, SNP annotations, predicted causal SNP-disease pairs from UK Biobank fine-mapping analyses and from the NHGRI-EBI GWAS Catalog and SNP heritability causally explained by SNPs linked to each gene have been made publicly available at <https://alkesgroup.broadinstitute.org/cS2G> and <https://doi.org/10.5281/zenodo.6354007>. Links for all data sets used to create S2G strategies are provided in Supplementary Table 26. Access to the UK Biobank resource is available via application at <http://www.ukbiobank.ac.uk/>.

The GWAS Catalog is available at <https://www.ebi.ac.uk/gwas/api/search/downloads/full>.

Open Targets SNP-gene pairs are available at [https://raw.githubusercontent.com/opentargets/genetics-gold-standards/master/gold\\_standards/processed/gwas\\_gold\\_standards.191108.tsv](https://raw.githubusercontent.com/opentargets/genetics-gold-standards/master/gold_standards/processed/gwas_gold_standards.191108.tsv).

SNP-gene pairs from ref.<sup>48</sup> are available at [https://www.dropbox.com/s/kz2c49rpm2yanf5/all\\_byCS\\_rev1.txt?dl=0](https://www.dropbox.com/s/kz2c49rpm2yanf5/all_byCS_rev1.txt?dl=0).

## Code availability

The code to estimate precision and recall of S2G strategies and the code to create combined S2G strategies have been made publicly available at <https://alkesgroup.broadinstitute.org/cS2G/code> and <https://doi.org/10.5281/zenodo.6415925>.

## References

79. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
80. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
81. Lindblad-Toh, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
82. Leeuw, C. A., de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

## Acknowledgements

We thank X. Jiang, C. Boix and M. Kellis for helpful discussion. S.G. is funded by National Institutes of Health grant R00 HG010160. A.L.P. is funded by National Institutes of Health grants U01 HG009379, R01 MH101244, R37 MH107649, R01 MH115676, R01 MH109978, U01 HG012009 and R01 HG006399. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. This research was conducted using the UK Biobank Resource under application 16549.

## Author contributions

S.G. and A.L.P. designed experiments. S.G. performed experiments. S.G., O.W., F.H., K.D., J.N., and K.J. analyzed data. D.W., H.S., C.P.F., L.O.C., B.P. and J.M.E. provided suggestions on the analyses. S.G. and A.L.P., with assistance from all authors, wrote the manuscript.

## Competing interests

C.P.F. is now an employee of Bristol Myers Squibb. The remaining authors declare no competing interests.

## Additional information

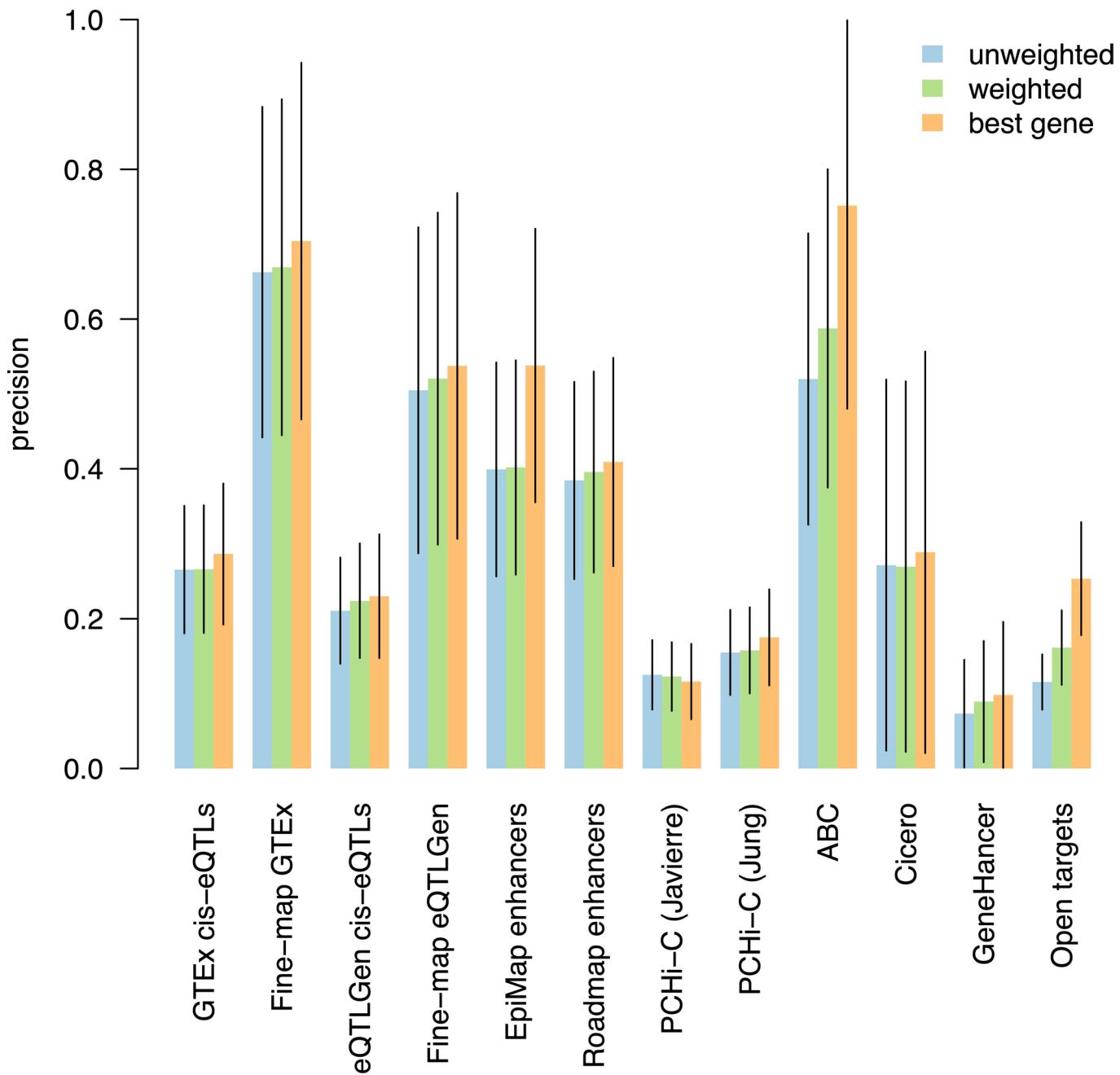
**Extended data** is available for this paper at <https://doi.org/10.1038/s41588-022-01087-y>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41588-022-01087-y>.

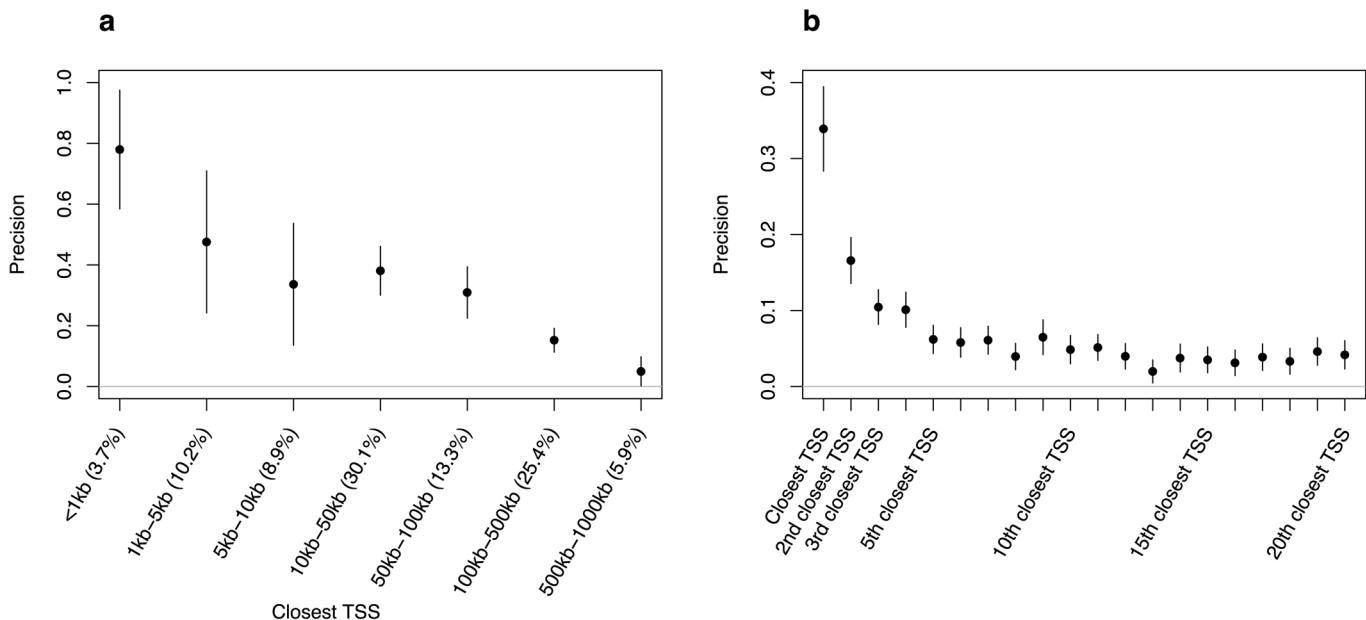
**Correspondence and requests for materials** should be addressed to Steven Gazal or Alkes L. Price.

**Peer review information** *Nature Genetics* thanks Guillaume Lettre and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.

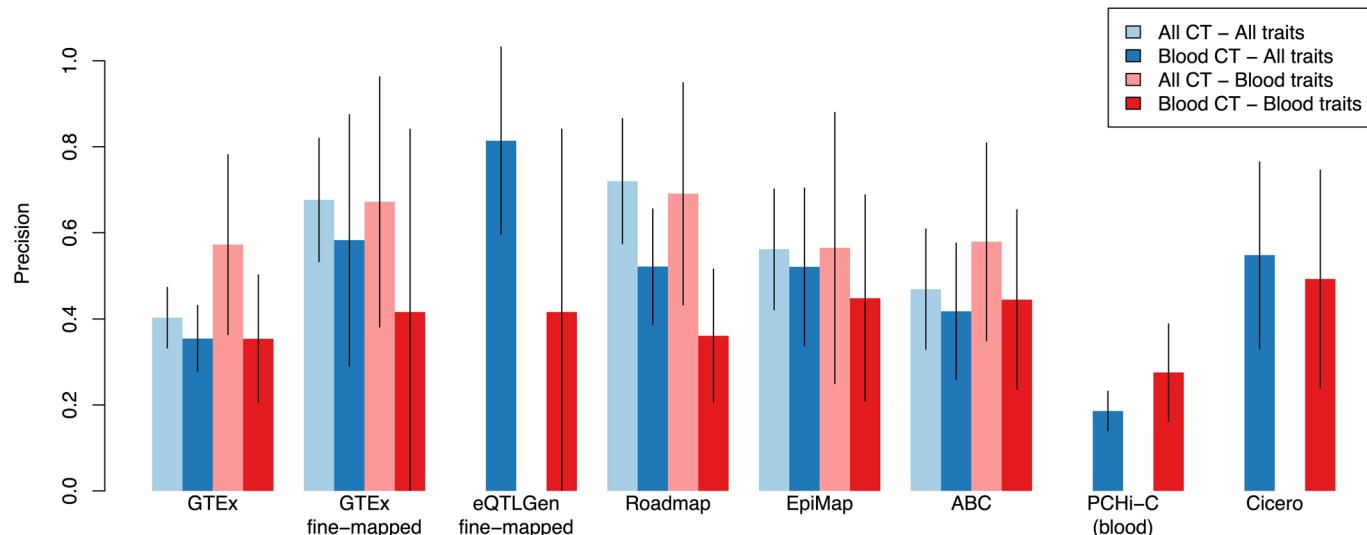
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



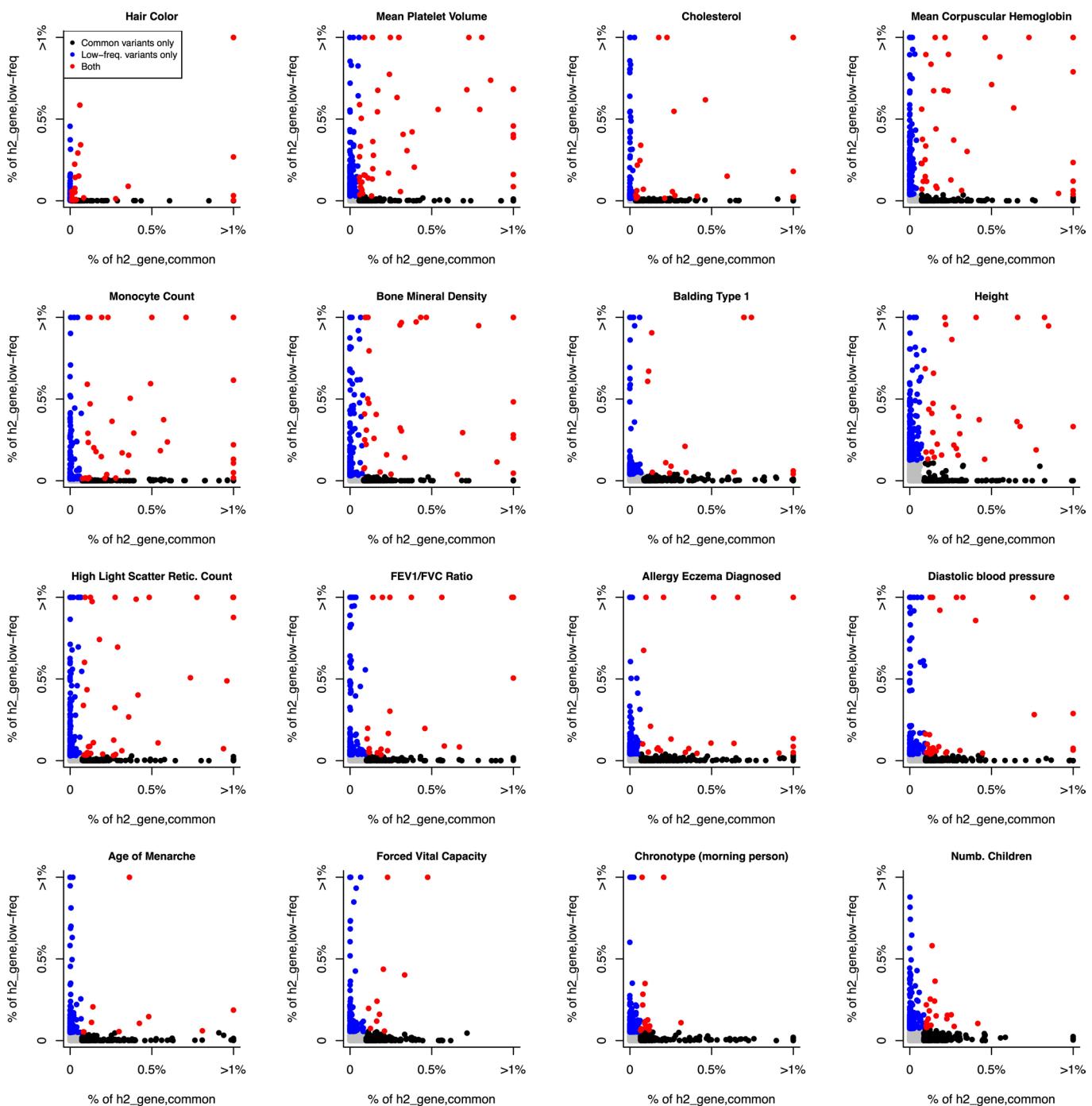
**Extended Data Fig. 1 | S2G strategy linking each SNP to best gene leads to higher precision than linking SNPs to multiple target genes.** We report the precision of S2G strategies linking SNPs to target genes using three difference approaches for converting raw linking values into linking scores: by assigning to each gene with non-zero raw linking value the same linking score (unweighted), by assigning to each gene a linking score proportional to its raw linking value (weighted), and by retaining only the gene(s) with the highest linking score (best gene). Values were estimated using non-trait-specific training critical gene set and meta-analyzed across 63 independent traits. Error bars represent 95% confidence intervals around meta-analyzed values. For most of the S2G strategies the precision was very similar (except for EpiMap, ABC and Open Targets), but the precision was generally highest for the ‘best gene’ strategy. However, we note that this choice does not reflect biological reality, in which a regulatory element may target more than one gene, and that refinements to this choice are a direction for future research.



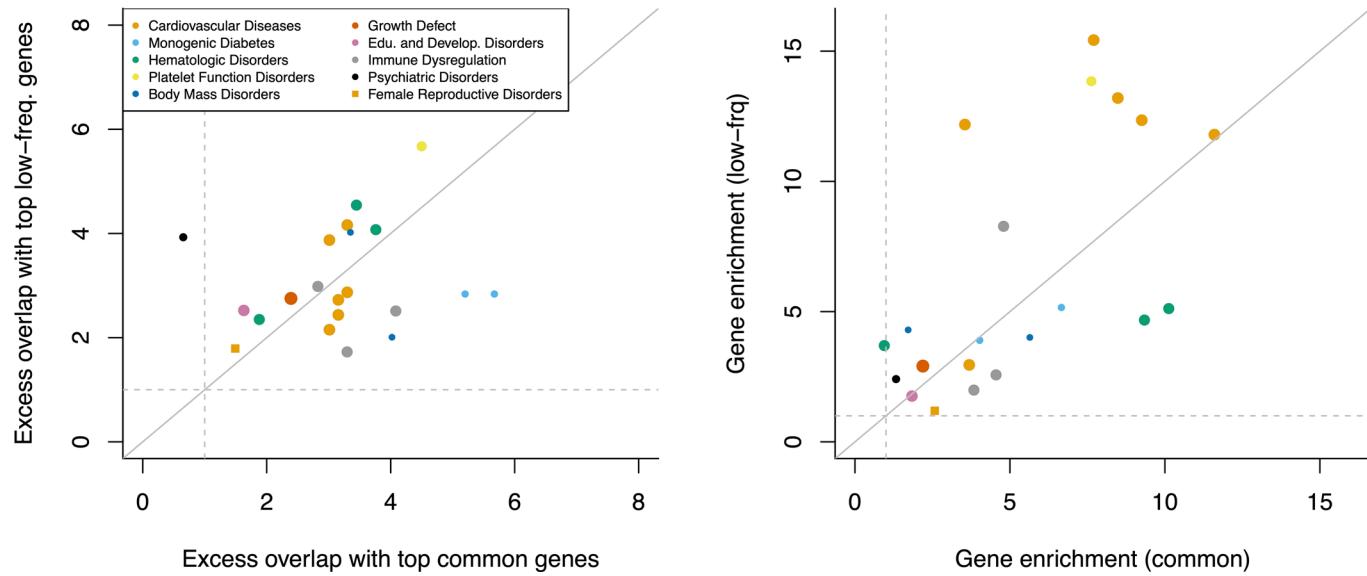
**Extended Data Fig. 2 | Precision of 27 S2G strategies based on physical distance to TSS.** We report precision of the closest TSS strategy as a function of the distance between a SNP and its closest TSS (a) (numbers between parentheses represent the fraction of common SNPs linked by the strategy), and the precision of the  $i^{\text{th}}$  closest TSS (each strategy links 100% of the SNPs) (b). Values were estimated using trait-specific validation critical gene sets and meta-analyzed across 63 independent traits. Error bars represent 95% confidence intervals around meta-analyzed values. The mean value of 0.043 for 6th-20th closest TSS suggests that genes located relatively close to causal disease genes have a slightly elevated probability of being causal. Numerical results including values of recall and corresponding standard errors are reported in Supplementary Table 5.



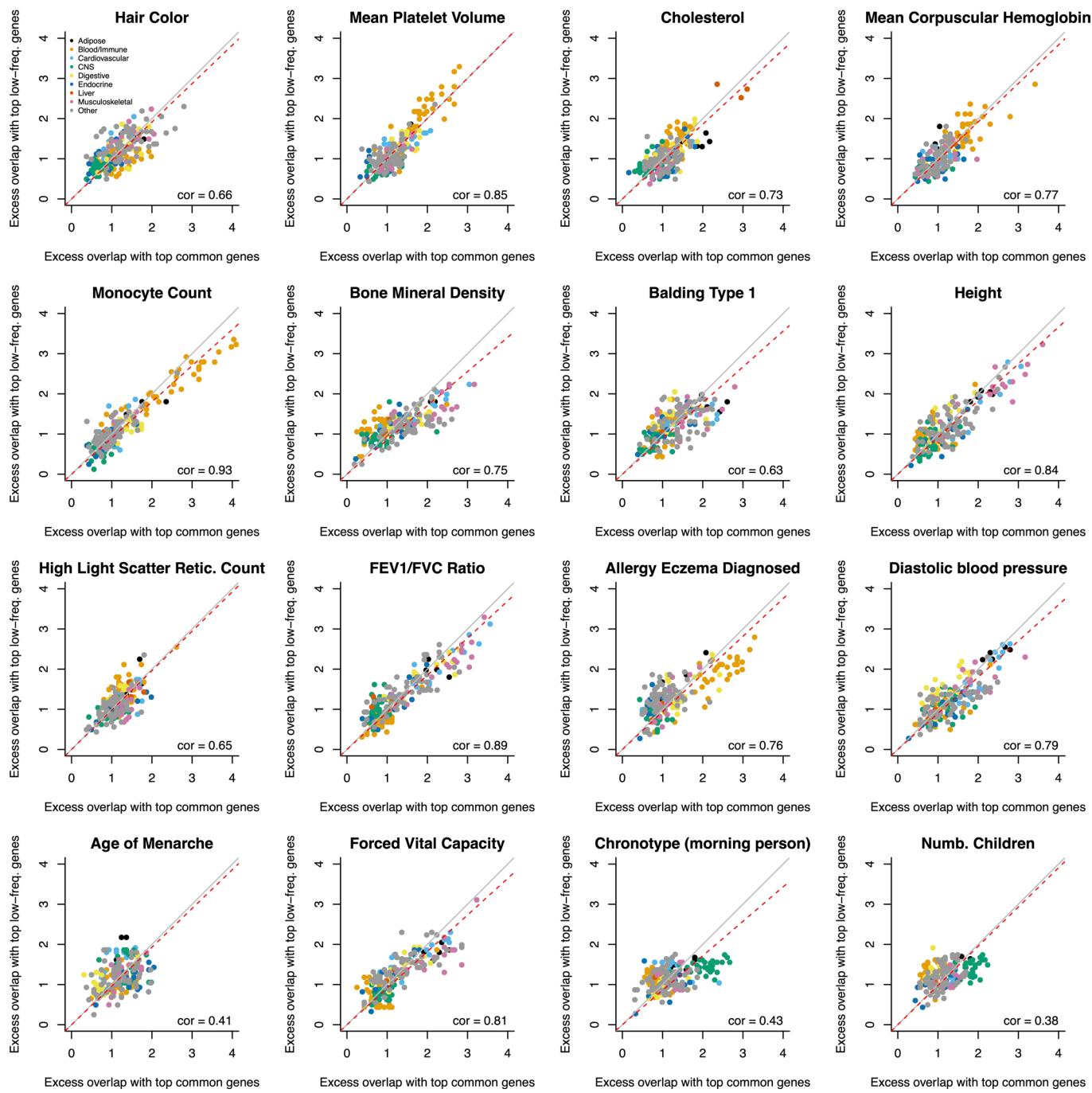
**Extended Data Fig. 3 | Precision of functional S2G strategies using all available cell types and tissues or restricted to blood and immune cell types and tissues.** We report the precisions of functional S2G strategies built using either all available cell types and tissues (All CT; in light color) and/or blood and immune cell types and tissues (Blood CT; in dark color) meta-analyzed across 63 independent traits (All traits; in blue) and 11 blood cell traits and autoimmune diseases (Blood traits; in red) (UK Biobank all autoimmune diseases, Crohn's Disease, Rheumatoid Arthritis, Ulcerative Colitis, Lupus, Celiac, Platelet Count, Red Blood Cell Count, Red Blood Cell Distribution Width, Eosinophil Count, White Blood Cell Count; see Supplementary Table 3). Error bars represent 95% confidence intervals around meta-analyzed values. We considered 5 S2G strategies with data available for cell types and tissues: GTEx *cis*-eQTLs (GTEx), GTEx fine-mapped *cis*-eQTL (GTEx fine-mapped), Roadmap enhancer-gene linking (Roadmap), EpiMap enhancer-gene linking (EpiMap), and Activity-By-Contact (ABC). We considered 3 S2G strategies with data available only for blood and immune cell types and tissues: eQTLGen fine-mapped blood *cis*-eQTL (eQTLGen fine-mapped), PCHi-C (blood), and Cicero blood/basal (Cicero). We observed 1) that S2G strategies using data from all cell types and tissues were more precise than S2G strategies restricted to blood and immune cell types and tissues in both analyses of all traits (light blue vs. dark blue) and blood cell traits and autoimmune diseases (light red vs. dark red), and 2) that S2G strategies using data from blood and immune cell types and tissues are more precise in all traits than in blood cell traits and autoimmune diseases (dark blue vs. dark red).



**Extended Data Fig. 4 | Proportion of common and low-frequency variant heritability linked to genes explained by each individual gene.** We report the proportion of common and low-frequency variant heritability linked to genes ( $h^2_{\text{gene}, \text{common}}$  and  $h^2_{\text{gene}, \text{low-freq}}$ , respectively) explained by each individual gene in 16 independent UK Biobank traits. Genes in the top 200 genes (top 1% of all genes) contributing to both  $h^2_{\text{gene}, \text{common}}$  and  $h^2_{\text{gene}, \text{low-freq}}$  are denoted in red (median of 26 genes across the 16 traits), genes in the top 200 genes contributing to only  $h^2_{\text{gene}, \text{common}}$  (resp.  $h^2_{\text{gene}, \text{low-freq}}$ ) are colored in black (resp. blue) (median of 174 genes each), and remaining genes are colored in gray (median of 19,621 genes, with values close to 0 on both axes). We observe low concordance between per-gene contributions to gene architectures for common vs. low-frequency SNPs.



**Extended Data Fig. 5 | Excess overlap between top genes contributing to common and low-frequency variant heritability linked to genes and disease-specific Mendelian disorder genes.** We report the excess overlap between phenotype-specific Mendelian disorder genes<sup>72</sup> and the top 200 genes contributing to common and low-frequency variant heritability linked to genes (left), and the gene enrichment of disease-specific Mendelian disorder genes (that is  $[\text{SNP heritability linked to Mendelian disorder genes} / \text{SNP heritability linked to all genes}] / [\text{number of Mendelian disorder genes} / \text{total number of genes}]$ ) across common and low-frequency variants (right). Each dot represents a disease/trait - Mendelian disorder gene set pair. These two results suggest that both the set of top 200 genes and the per-gene heritability estimates are unlikely to be driven by noisy estimates arising from finite sample size. We restricted analyses to 21 traits analyzed in ref. <sup>72</sup>.



**Extended Data Fig. 6 | Excess overlap between top genes contributing to common and low-frequency variant heritability linked to genes and differentially expressed gene sets.** We report the excess overlap between 205 differentially expressed gene sets<sup>41</sup> and the top 200 genes contributing to common and low-frequency variants heritability linked to genes across 16 independent UK Biobank traits. Each dot represents a differentially expressed gene set, and is colored by the tissue category. We generally observed excess overlap for disease-critical tissues/cell types. We observed high correlations between excess overlaps for common vs. low-frequency variant architectures, suggesting that common and low-frequency variants architectures are driven by different genes pertaining to similar biological processes.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used for data collection.
Data analysis	Our code is available at <a href="https://alkesgroup.broadinstitute.org/cS2G/code">https://alkesgroup.broadinstitute.org/cS2G/code</a> and <a href="https://doi.org/10.5281/zenodo.6353668">https://doi.org/10.5281/zenodo.6353668</a> . LD score regression software (ldsc) version 1.0.1 is available at <a href="https://github.com/bulik/ldsc">https://github.com/bulik/ldsc</a> PolyFun software version 1.0.0 is available at <a href="https://github.com/omerwe/polyfun">https://github.com/omerwe/polyfun</a> Data analyses were performed using R version 3.6.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The List of 19,995 genes, summary statistics of the 63 independent traits, training and validation critical gene sets, S2G and cS2G strategies, SNP annotations, predicted causal SNP-disease pairs from UK Biobank fine-mapping analyses and from the NHGRI-EBI GWAS catalog, and SNP-heritability causally explained by SNPs linked to each gene have been made publicly available at <https://alkesgroup.broadinstitute.org/cS2G> and <https://doi.org/10.5281/zenodo.6354007>. Links for all datasets used to create S2G strategies are provided in Supplementary Table 26.

Access to the UK Biobank resource is available via application at <http://www.ukbiobank.ac.uk/>.

GWAS catalog <https://www.ebi.ac.uk/gwas/api/search/downloads/full>

Open Targets SNP-gene pairs [https://raw.githubusercontent.com/opentargets/genetics-gold-standards/master/gold\\_standards/processed/gwas\\_gold\\_standards.191108.tsv](https://raw.githubusercontent.com/opentargets/genetics-gold-standards/master/gold_standards/processed/gwas_gold_standards.191108.tsv)

SNP-gene pairs from ref.48 [https://urldefense.proofpoint.com/v2/url?u=https-3A-\\_www.dropbox.com\\_s\\_kz2c49rpm2yanf5\\_all-5FbyCS-5Frev1.txt-3FdI-3D0&d=DwMFaQ&c=WO-RGvefibhHBZq3fL85hQ&r=pj2hZETq-6Xv2-wuSquXm871XqnKfXGPV5duZ9gf88w&m=lWrDyJkE3HPLhLS3pLXW8e80amAyxaNTtTw2ULHrbLA&s=A8IWwTxhGFCV0avT0-G3s0X7Cs1TXIFqvPx8woQ6iiU&e=](https://urldefense.proofpoint.com/v2/url?u=https-3A-_www.dropbox.com_s_kz2c49rpm2yanf5_all-5FbyCS-5Frev1.txt-3FdI-3D0&d=DwMFaQ&c=WO-RGvefibhHBZq3fL85hQ&r=pj2hZETq-6Xv2-wuSquXm871XqnKfXGPV5duZ9gf88w&m=lWrDyJkE3HPLhLS3pLXW8e80amAyxaNTtTw2ULHrbLA&s=A8IWwTxhGFCV0avT0-G3s0X7Cs1TXIFqvPx8woQ6iiU&e=)

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used the larger sample size that was available.
Data exclusions	We restrict our analyses on individuals from European ancestry. We excluded the MHC region from S-LDSC analyses and analyzed only autosomes, as recommended and developed by S-LDSC authors.
Replication	S2G evaluations were replicated on three curated disease-associated lists of SNP-gene pairs. We used training and validation critical gene sets to train and validate the combined S2G strategy. We compared the proportion of heritability linked to genes explained by genes ranked by top per-gene heritability by using per gene heritability that has been used for the ranking (using N=337K British UK Biobank training samples ) and by running S-LDSC on summary statistics computed from N=122K European-ancestry UK Biobank validation samples that were distinct from the N=337K British UK Biobank training samples (to avoid winner's curse).
Randomization	Analyses were performed by maximizing the datasets and samples investigated; randomization is inapplicable here.
Blinding	Blinding is not possible in this study: we did not collect data for this study, but analyzed publicly available summary statistics and functional datasets.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		