The Impact of Stability Considerations on Genetic Fine-Mapping

Alan Aw[1,2], Lionel Chentian Jin[3], Nilah Ioannidis[2,4,*], Yun S. Song[1,2,4,*]

[1] Department of Statistics, University of California, Berkeley
[2] Center for Computational Biology, University of California, Berkeley
[3] McKinsey & Company, Seattle
[4] Computer Science Division, University of California, Berkeley

March 23, 2023

Abstract

Fine-mapping methods, which aim to identify genetic variants responsible for complex traits following genetic association studies, typically assume that sufficient adjustments for confounding within the association study cohort have been made, e.g., through regressing out the top principal components (i.e., residualization). Despite its widespread use, however, residualization may not completely remove all sources of confounding. Here, we propose a complementary stability-guided approach that does not rely on residualization, which identifies consistently fine-mapped variants across different genetic backgrounds or environments. We demonstrate the utility of this approach by applying it to fine-map eQTLs in the GEUVADIS data. Using 378 different functional annotations of the human genome, including recent deep learning-based annotations (e.g., Enformer), we compare enrichments of these annotations among variants for which the stability and traditional residualization-based fine-mapping approaches agree against those for which they disagree, and find that the stability approach enhances the power of traditional fine-mapping methods in identifying variants with functional impact. Finally, in cases where the two approaches report distinct variants, our approach identifies variants comparably enriched for functional annotations. Our findings suggest that the stability principle, as a conceptually simple device, complements existing approaches to fine-mapping, reinforcing recent advocacy of evaluating cross-population and cross-environment portability of biological findings. To support visualization and interpretation of our results, we provide a Shiny app, available at: https://alan-aw.shinyapps.io/stability_v0/.

*To whom correspondence should be addressed: nilah@berkeley.edu, yss@berkeley.edu

1 Introduction

An important challenge faced by computational precision health research is the lack of generalizability of biological findings, which are often obtained by studying cohorts that do not include particular communities of individuals. Known as the *cross-population generalizability* or *portability problem*, the challenge persists in multiple settings, including gene expression prediction (Keys et al. 2020) and polygenic risk score prediction (Mostafavi et al. 2020). Generalizable biological signals, such as the functional impact of a variant, are important, as they ensure that general conclusions drawn from cohort-specific analyses are not based on spurious discoveries. Portability problems are potentially attributable to cohort-biased discoveries being treated as generalizable true signals. Efforts to address such problems have included the use of diverse cohorts typically representing multiple population ancestries (Márquez-Luna et al. 2017), meta-analyses of earlier studies across diverse cohorts (Turley et al. 2021; Han and Eskin 2012; Morris 2011; Willer et al. 2010), or focusing on biological markers that are more likely *a priori* to play a causal role, e.g., rare variants (Zaidi and Mathieson 2020) or the transcriptome (Liang et al. 2022).

In this paper, we consider an approach based on the notion of *stability* to improve generalizability. Being a pillar of veridical data science (Yu and Kumbier 2020), stability refers to the robustness of statistical conclusions to *perturbations* of the data. Perturbations are not arbitrary, but instead they encode the practitioner's beliefs about the quality of the data and the nature of relationship between variables. Well-chosen perturbation schemes will help the practitioner obtain statistical conclusions that are robust, in the sense that they are generalizable rather than spurious findings, and therefore more likely to capture the true signal. For example, in sparse linear models, prioritizing the stability of effect sizes to cross-validation "perturbations" leads to a much smaller set of selected features without reducing predictive performance (Lim and Yu 2016). In another example involving the application of random forests to detect higher-order interactions between gene regulation features (Basu et al. 2018), interactions stable to bootstrap perturbations are also largely consistent with known physical interactions between the associated transcription factor binding or enhancer sites.

To further investigate the utility of the stability approach, we focus on a procedure known as (genetic) fine-mapping (Schaid et al. 2018). Fine-mapping is the task of identifying genetic variants that causally affect some trait of interest. From the viewpoint of stability, previous works focusing on cross-population stability of fine-mapped variants typically use Bayesian linear models. The linear model has allowed modeling of heterogeneous effect sizes across different user-defined populations (e.g., ethnic groups, ancestrally distinct populations, or study cohorts), and it is common to assume that causal variants share correlated effects across populations (see, e.g., eq. (24) of LaPierre et al. 2021 or eq. (9) of Wen et al. 2015).

Unlike the parametric approaches described above, we here apply a *non-parametric* fine-mapping method to GEUVADIS (Lappalainen et al. 2013), a database of gene expression traits measured across individuals of diverse genetic ancestries and from different geographical environments, whose accompanying genotypes are available through the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015). We apply fine-mapping through two approaches, one that is commonly used in practice and another that is guided by stability. We evaluate the agreement of fine-mapped variants between the two approaches, and then measure the functional significance of variants picked by both approaches using a much wider range of functional annotations than considered in previous studies. By performing various statistical tests on the functional annotations, we evaluate the advantages brought about by the incorporation of stability. Finally, to support visualization of results at both the single gene and genome-wide levels, we provide an interactive Shiny app which is available at: `https://alan-aw.shinyapps.io/stability_v0/`. Our app is open-source

and designed to support geneticists in interpreting our results, thereby also addressing a growing demand for accessible, *integrative* software to interpret genetic findings in the age of big data and variant annotation databases.

2 Results

2.1 Experimental Design

We apply the fine-mapping method PICS (Taylor et al. 2021; Farh et al. 2015; see Algorithm 1 in Subsection 4.1) to the GEUVADIS data (Lappalainen et al. 2013), which consists of $T = 22,720$ gene expression traits measured across $N = 445$ individuals, with their accompanying genotypes obtained from the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015). These individuals are of either European or African ancestry, with about four fifths of the cohort made up of individuals of (self-identified) European descent. In particular, these ancestrally different subpopulations have distinct linkage disequilibrium patterns and environmental exposures, which constitute potential confounders that we wish to stabilize the fine-mapping procedure against.

PICS, like many eQTL analysis methods, requires the lead variant at a locus to compute posterior probabilities, so we perform marginal regressions of each gene expression trait against variants within the fine-mapping locus. Our implementation of PICS generates three sets of variants, which represent putatively causal variants that are (marginally) highly associated, moderately associated and weakly associated with the gene expression phenotype. (See Section 4, Materials and Methods for details.) We compare two versions of the fine-mapping procedure — one that is typically performed in practice, and another that is motivated by the stability approach. See Figure 1A.

Stable variant. Specifically, to incorporate stability into fine-mapping and obtain what we call the *stable variant* or stable SNP, we encode into the algorithm our belief that for a gene expression trait, a causal variant would presumably act through the same mechanism regardless of the population from which they originate. Indeed, if a set of genetic variants were causal, the same algorithm should report it, if run on subsets of the data corresponding to heterogeneous populations. This belief is consistent with recent simulation studies showing that the inclusion of GWAS variants discovered in diverse populations both mitigates false discoveries driven by linkage disequilibrium differences (Li et al. 2022) and improves generalizability of polygenic score construction (Cavazos and Witte 2021). Hence, the stable variant is the variant with the highest probability of being causal, conditioned on being reported by PICS in the most number of subpopulations. Section 4.2 provides a formal definition of the stable variant.

Top variant. The stability consideration we have described is closely related to a popular procedure known as correction for population structure, which residualizes the trait using measured confounders (e.g., top principal components computed from the genotype matrix). We call the variant returned by the residualization approach the *top variant* or top SNP. The residualization approach removes the effects of genetic ancestry and environental exposures on a trait, and is used to avoid the risk of low statistical power borne by stratified analyses. The top variant is formally defined towards the end of Section 4.1. We remark that the top variant is a function of the residualized trait, as opposed to the stable variant, which is a function of the unresidualized trait (see Figure 1A).

Comparison. We compare the results obtained by using each approach, across a range of categories of functional annotations including conservation, pathogenicity, chromatin accessibility, transcription factor binding and histone modification, and across biological assays and computational

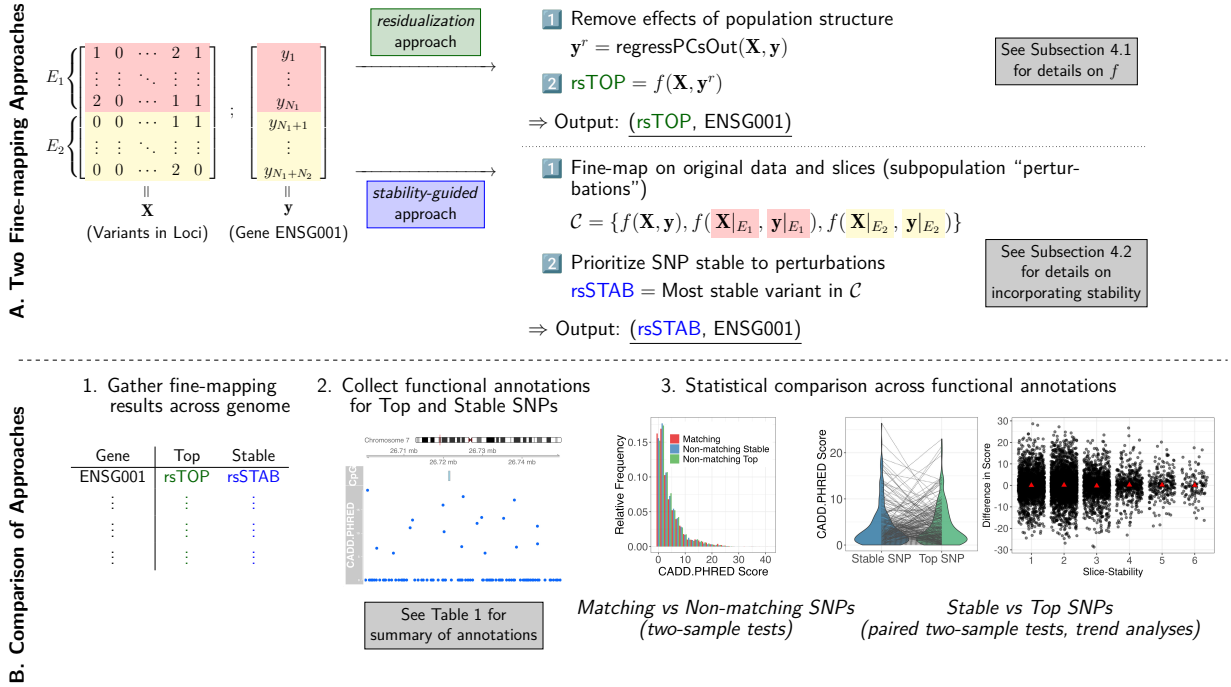

Figure 1: An overview of our study of the impact of stability considerations on genetic fine-mapping. **A.** The two ways in which we perform fine-mapping, the first of which (colored in green) prioritizes the stability of variant discoveries to subpopulation perturbations. The data illustrates the case where there are two distinct environments, or subpopulations (denoted E_1 and E_2), that split the observations. **B.** Key steps in our comparison of the stability-guided approach with the popular residualization approach.

predictions. See Table 1 for full list of annotations covered. (Supplementary Material Section S2 contains a full description of each quantity and its interpretation.)

Figure 1B summarizes the key steps of our investigation. To be clear, a first comparison is between the set of genes for which the top and stable variants match and the set of genes for which the top and stable variants disagree; comparisons are made between matching variants and one of the non-matching sets of variants (top or stable). A second comparison is restricted to the set of genes with non-matching top and stable variants, i.e., between *paired* sets of variants fine-mapped to a gene, where one set of a pair corresponds to the output of the residualization approach whereas the other corresponds to the output of the stability-guided approach. We additionally compare across three fine-mapping output sets, corresponding to variants that have a high, moderate or weak marginal association with the expression phenotype. For brevity, we term these three sets Potential Set 1, Potential Set 2 and Potential Set 3, respectively. (See Algorithm 1 for details.) For each Potential Set, we find the top and stable variants as described in Section 4.2.

2.2 Stability-guided variants frequently do not match residualization variants

As a preliminary analysis, when interrogating if the residualization and stability-guided approaches produce the same candidate causal variants, we find that for Potential Set 1, which corresponds to the set typically reported in fine-mapping studies, 56.2% of genes had matching top and stable variants (see Figure 2). Moving down potential sets, we find less agreement (Potential Set 2: 36.2%,

Table 1: A list of 378 functional annotations across which the biological significances of stable and top fine-mapped single nucleotide polymorphisms are compared. Annotations that report multiple scores have the total number of scores reported shown in parentheses. Scores mined from the FAVOR database (Zhou et al. 2022) are indicated by an asterisk. (TSS = Transcription Start Site, bp = base pair)

Functional Annotation Type	**Functional Annotation**
Ensembl	Distance to Canonical TSS (Cunningham et al. 2022) Regulatory Features (6; Cunningham et al. 2022)
Computational Predictions	CADD* (2; Rentzsch et al. 2019) SIFTVal* (Ng and Henikoff 2003) FATHMM-XF* (Rogers et al. 2018) LINSIGHT* (Huang et al. 2017) Polyphen* (Adzhubei et al. 2010) PhyloP* (3; Pollard et al. 2010) Gerp* (2; Davydov et al. 2010) B Statistic* (McVicker et al. 2009) FunSeq2* (Fu et al. 2014) ALoFT* (Balasubramanian et al. 2017) Percent CpG in $75bp$ window* (Rentzsch et al. 2019) Percent GC in $75bp$ window* (Rentzsch et al. 2019) FIRE (Ioannidis et al. 2017) Enformer (177 tracks $\times$ 2 scores per track; Avsec et al. 2021)

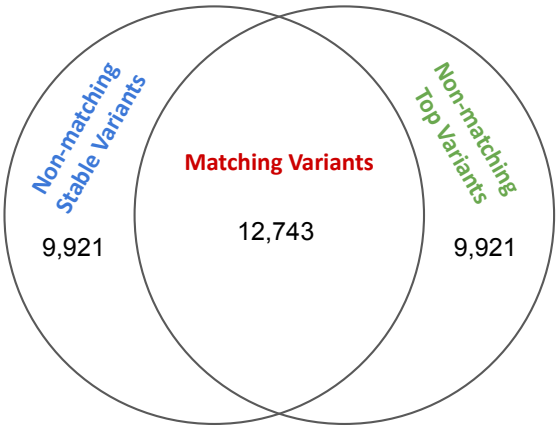

Figure 2: Venn diagram showing the number of matching and non-matching variants for Potential Set 1.

Potential Set 3: 25.6%), providing evidence that as the marginal association of a variant with the expression phenotype decreases, the two approaches prioritize different signals when searching for putatively causal variants.

For the rest of this Section, we report only results for Potential Set 1, given that it corresponds to the set typically reported in fine-mapping. Results for the other potential sets are provided in Supplementary Material Sections S4 and S5.

2.3 Matching versus Non-Matching Variants

For each gene, our algorithm finds the top eQTL variant and the stable eQTL variant, which may not coincide. We thus run (one-sided) unpaired Wilcoxon tests on matching and non-matching sets of variants, to detect significant functional enrichment of one set of variants over the other. We find that the top variants that are also stable for the corresponding gene ($N_{\text{match}} = 12,743$ maximum annotatable) score significantly higher in functional annotations than the top variants that are not stable ($N_{\text{non-match}} = 9,921$ maximum annotatable). Notably, 361 out of 378 functional annotations report one-sided greater p-values < 0.05 for the matching (i.e., both top and stable) variants after correcting for multiple testing using the Benjamini-Hochberg (BH) procedure, with many of these annotations measuring magnitudes of functional impact or functional enrichment (e.g., Enformer perturbation scores, FATHMM.XF score). Amongst the 17 remaining functional annotations, none has significantly lower scores for the matching variants. (Supplementary Material Section S4 lists these functional annotations in detail.) We also find that empirically, the matching variants tend to have greater agreement in posterior probabilities than non-matching variants (Supplementary Figures S1 and S2).

As an example of a significant functional annotation, consider raw CADD scores (Rentzsch et al. 2019) — a higher value of which indicates a greater likelihood of deleterious effects. Out of the $22,559$ genes for which both the top and the stable variant are annotatable, looking at the distribution of top variant scores, the one corresponding to the $12,685$ genes with matching top and stable variants stochastically dominates the one corresponding to the remaining $9,874$ genes (Figure 3A). This relationship is more pronounced when we inspect PHRED-scaled CADD scores, where we apply a sliding cutoff threshold for calling variant deleteriousness (i.e., potential pathogenicity — see Rentzsch et al. 2019). We find that a greater proportion of matching variants than of either non-matching variant is classified as deleterious under the typical range of deleteriousness cutoffs (Figure 3B).

Another example demonstrating significant functional enrichment of matching variants over non-matching variants is the perturbation scores on H3K9me3 ChIP-seq peaks, as predicted by the Enformer (Avsec et al. 2021). Out of the $6,364$ genes for which the distance of both the top and the stable variant to the TSS are within the Enformer input sequence length constraint, looking at the the distribution of top variant scores, the one corresponding to the $4,491$ genes with matching top and stable variants stochastically dominates the one corresponding to the remaining $1,873$ genes (Figure 3C,D). This relationship is true regardless of whether perturbation scores are calculated from an average of input sequences centered on the gene TSS and its two flanking positions (Figure 3C), or from input sequences centered on the gene TSS only (Figure 3D). (See Supplementary Material Section S3 for details on Enformer annotation calculation.)

Similarly, amongst all stable variants, those variants that match the top variant are significantly more enriched in functional annotations: 363 functional annotations report one-sided greater BH-adjusted p-values < 0.05 for the matching variants, and none of the remaining 15 annotations present significant depletion for the matching variants set.

2.4 Top versus Stable variants when they do not match

Focusing on the genes for which the top and stable variants are different, we run (one-sided) paired Wilcoxon tests to detect significant functional enrichment of one set of variants over the other. We find in general that for some genes stable variants can carry more functional impact than top variants, and for other genes top variants carry more functional impact — although neither of these patterns is statistically significant genome-wide after multiple testing correction using the

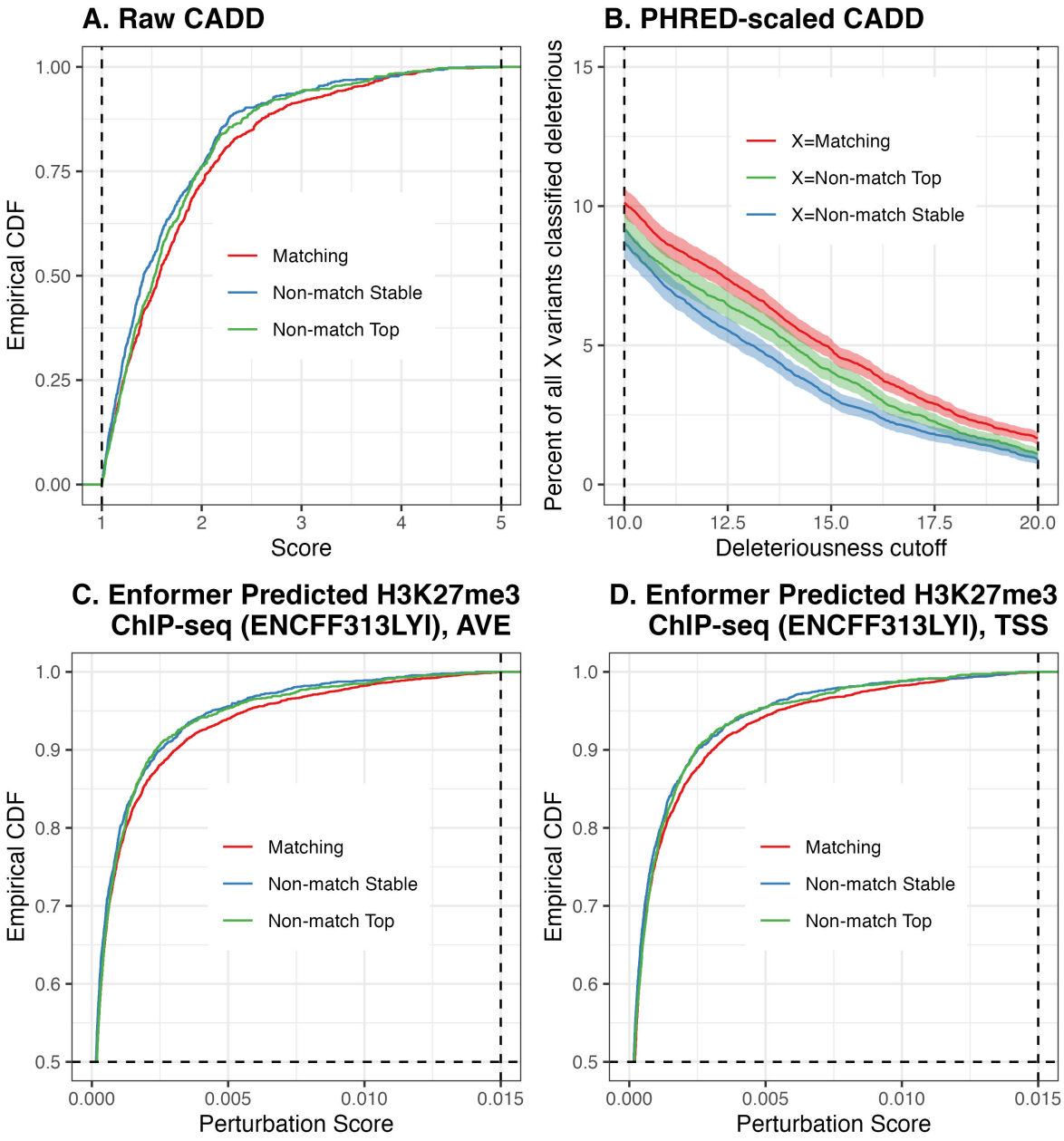

Figure 3: *Top Row.* CADD scores. **A.** Empirical cumulative distribution functions of raw CADD scores of matching and non-matching variants across all genes, for Potential Set 1. Non-matching variants are further divided into stable and top variants, with a score lower threshold of 1.0 and upper threshold of 5.0 used to improve visualization. **B.** For a deleteriousness cutoff, the percent of (i) all matching variants, (ii) all non-matching top variants, and (iii) all non-matching stable variants, which are classified as deleterious. We use a sliding cutoff threshold ranging from 10 to 20 as recommended by CADD authors. *Bottom Row.* Empirical cumulative distribution functions of perturbation scores of Enformer-predicted H3K27me3 ChIP-seq track. Score upper threshold of 0.015 and empirical CDF lower threshold of 0.5 used to improve visualization. **C.** Perturbation scores computed from predictions based on centering input sequences on the gene TSS as well as its two flanking positions. **D.** Perturbation scores computed from predictions based on centering input sequences on the gene TSS only.

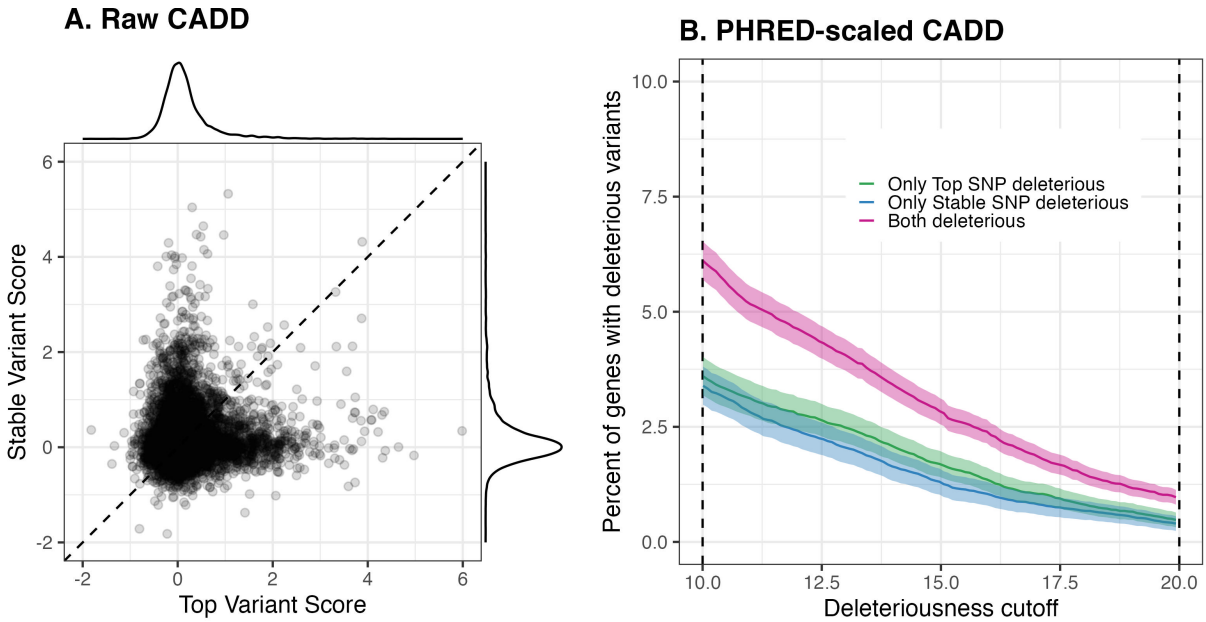

Figure 4: **A.** Paired scatterplot of raw CADD scores of both top and stable variant for each gene, for Potential Set 1. **B.** Percent of genes that are classified as (i) having deleterious top variant only, (ii) having deleterious stable variant only, and (iii) having both top and stable variant deleterious, using a sliding cutoff threshold ranging from 10 to 20 as recommended by CADD authors.

BH procedure. For example, for raw CADD scores, out of the $9,874$ genes for which the top and the stable variants do not match, $4,906$ genes have higher scoring stable variants than top variants, whereas $4,968$ genes have higher scoring top variants than stable variants (Figure 4A). Looking at the accompanying PHRED-scaled CADD scores (CADD PHRED) for Potential Set 1, when applying a sliding cutoff threshold for calling variant deleteriousness, we find that even though a higher fraction of genes have top but not stable variant classified as deleterious, the difference is usually not significant (Figure 4B). Figure 4B also demonstrates that no matter how the cutoff is chosen, there are genes for which the top variant is not classified deleterious while the stable variant is. Taken together, these observations suggest that the stability-guided approach can sometimes be more useful at identifying variants of functional significance, and in a broad sense both fine-mapped variants should be equally prioritized for potential of carrying functional impact.

Because our comparison between the top and stable variants yielded no significant functional enrichment of one over the other, we investigate whether external factors — for example, the posterior probability of the variants — might moderate the relative enrichment of one variant over the other (i.e., we perform *trend analysis* — see Section 4.3 for a list of all moderators). Here, we find that all but one of the moderators considered — namely, the Posterior Probability of Top Variant (see Table 2) — did not produce any significant trends for Potential Set 1. There is a small but significant positive correlation (Pearson's $r = 0.05$) between the posterior probability of the top variant and the difference between the stable variant and top variant FIRE scores (Ioannidis et al. 2017), and a small but significant negative correlation (Pearson's $r = -0.07$) between the posterior probability of the top variant and the difference between the stable variant and top variant Absolute Distance to Canonical TSS. (Details are reported in Supplementary Material Section S5.)

Additional comparisons. We perform various *conditional analyses* to evaluate whether additional restrictions to characteristics of fine-mapping outputs may boost the power of either approach

over the other. Such characteristics include (i) the positive posterior probability support (i.e., how many variants reported a positive posterior probability from fine-mapping), and (ii) the posterior probability of the top or stable variant. Results from our conditional analysis of (i) are reported in Supplementary Material Section S6. As an example, for (ii) we further perform a comparison between top and stable variants, by restricting to genes where the posterior probability of the top variant or the stable variant exceeds 0.9. Such restriction of valid fine-mapped variant-gene pairs is useful in training variant effect prediction models requiring reliable positive and negative examples, as seen in Wang et al. (2021). We find that for genes where the top variant posterior probability exceeds 0.9, there is no significant enrichment of the top variant over the stable variant across the 378 annotations considered (all BH-adjusted p-values exceed 0.05). Interestingly, when focusing on genes where the stable variant posterior probability exceeds 0.9, FIRE scores of the top variant are significantly larger than the stable variant (BH-adjusted p-value $= 1 \times 10^{-6}$). Detailed results are reported in Supplementary Material Section S7.

3 Discussion

Our present work has shown that a stability-guided approach complements existing approaches to detect biologically meaningful variants in genetic fine-mapping. Through various statistical comparisons, we have found that prioritizing the agreement between existing approaches and a stability-guided approach enhances the functional impact of the fine-mapped variant. Incorporating stability into fine-mapping also provides an adjuvant approach that helps discover variants of potential functional impact in case standard approaches fail to pick up variants of functional significance. Our findings are consistent with earlier reports of stable discoveries having the tendency to capture actual physical or mechanistic relationships, potentially making such discoveries generalizable or portable (Basu et al. 2018).

The link between stability and generalizability is not new. In the machine learning and statistics literature, it has been shown that stable algorithms provably lead to generalization errors that are well-controlled (Bousquet and Elisseeff 2002, Theorem 17, p. 510). Furthermore, in certain classes of algorithms (e.g., sparse regression), it has been demonstrated empirically that this mathematical relationship is explained precisely by the stable algorithm removing spurious discoveries (Lim and Yu 2016).

The stability-guided approach is also distinct from other subsampling approaches, such as the bootstrap (Efron and Tibshirani 1994). First, the bootstrap is more often deployed as a method for calibrating uncertainty surrounding a prediction, which is not the objective of our method. Next, in settings where the bootstrap is deployed as a type of perturbation against which a prediction or estimand is expected to be stable (Basu et al. 2018), a stability threshold is implicitly needed and would require tuning to be chosen (e.g., in Basu et al. 2018 this threshold was chosen to be 50%). Here, we leverage interpretable existing external annotations to define the perturbation against which we expect the fine-mapped variant to be stable — in other words, the user relates what is meant by the fine-mapped variant being stable to a biologically meaningful concept like "portable across environmentally and LD pattern-wise heterogeneous populations."

The last sentence in the preceding paragraph suggests there are teleological similarities between the stability-guided approach and meta-analysis approaches (Turley et al. 2021). We emphasize that meta-analyses rely on already analyzed cohorts, thereby implicitly assuming that *within-cohort* heterogeneities have been sufficiently accounted for prior to the reporting of findings for that cohort. The stability-guided approach, however, is relevant to the *cohort-specific analysis itself*, where existing approaches may present methodological insufficiencies resulting in inflated false discoveries.

In other words, whereas the goal of meta-analysis may be stated as identifying consistent hits across cohorts while also assuming that findings specific to each cohort are reliable, the goal of a stability-guided approach is to search for consistent signals despite the presence of potential confounders within a single cohort. Our analysis has focused on comparing the stability-guided approach against residualization, a convenient and popular approach to account for population structure, which reflects this difference in purpose.

In investigating the impact of stability considerations on the fine-mapping algorithm, our analyses integrated multiple families of functional annotations. These include biological assays and tracks (Ensembl gene regulatory regions, percent GC, etc.), classical variant interpretation scores and statistics (CADD, FIRE, etc.), and modern deep learning-based annotations (Enformer). We further summarize our findings using open-source interactive software, enabling the broader biological community to interpret our results across 378 annotations and all genes, or at the single gene level. Altogether, our work not only provides pristine ways to synthesize new findings from the expanding suite of in-silico mutagenesis models with biological quantities reported from mathematical models in biology and classical statistical principles, but also presents the results in a format accessible to biologists seeking to benefit from computational insights.

Our present work is not without limitations. First, we have chosen to focus on a non-parametric approach to fine-mapping, but multiple parametric fine-mapping approaches have been extended to incorporate cross-population heterogeneity (e.g., Wen et al. 2015; LaPierre et al. 2021; Lu et al. 2022). While our present work is a proof-of-concept of the applicability of the stability to genetic fine-mapping, we believe that future work focusing on comparing functional impact of variants prioritized by our stability-driven approach or these parametric methods will shed more light on the efficacy of the stability principle at detecting generalizable biological signals. Second, our analysis implicitly assumes that there is exactly one cis-eQTL that is prioritized by PICS. It is possible to modify the algorithm to return multiple variants, but this would require a more theoretical investigation, so we defer such technical explorations to future work. Third, while we have relied on numerous computational and experimental annotations to evaluate the functional impact of our fine-mapped variants, there is a possibility that some of the computationally predicted variant annotations (e.g., CADD, Enformer, FIRE) are themselves biased owing to the lack of diversity of training data. Although recent work has evaluated the performance of these computational annotation tools on different tissues and cell types, to our knowledge, there are currently no benchmarks for evaluating the performance of computational annotations on diverse human cohorts. We believe such studies will be important for ensuring that our findings support the theory that stable variants carry generalizable biological signal.

In closing, while our work explores stability to subpopulation perturbations where subpopulations are defined by ancestry, we emphasize that our stability-guided slicing methodology is applicable to all settings where meaningful external labels are available to the data analyst. For instance, environmental or geographical variables, which are well-recognized determinants of some health outcomes (Favé et al. 2018; Abdellaoui et al. 2022) and arguably better measure potential confounders than ancestral population labels, can be the basis on which slices are defined in the stability-guided approach. As the barriers to access larger biobank-scale datasets, which contain such aforementioned variables, continue to be lowered, we expect stability-driven analyses conducted on such data and relying on carefully defined slices will help users better understand genetic drivers of complex traits. Given its utility in our present work, we believe that the stability approach in precision medicine may find uses beyond genetic fine-mapping and few other biological tasks previously studied, ultimately empowering the discovery of veridical effects not previously known.

4 Materials and Methods

We use publicly available GEUVADIS B-lymphocyte RNA-seq measurements from 445 individuals (Lappalainen et al. 2013), whose genotypes are also available from the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015). The 445 individuals come from five populations: Tuscany Italian (TSI, $N = 91$), Great British (GBR, $N = 86$), Finnish (FIN, $N = 92$), Utah White American (CEU, $N = 89$) and Yoruban (YRI, $N = 87$). The mRNA measurements are normalized for library depth, expression frequency across individuals as well as PEER factors, as reported in the GEUVADIS databank[1]. Of the available normalized gene expression phenotypes, only those with non-empty sets of variants lying within 1Mb upstream or downstream of the canonical transcription start site were kept for (cis) fine-mapping. This process yielded $22,664$ genes used in all subsequent analyses involving the PICS fine-mapping algorithm.

4.1 Probabilistic Identification of Causal SNPs

We implement Probabilistic Identification of Causal SNPs (PICS) (Farh et al. 2015), which is based on an earlier genome-wide analysis identifying genetic and epigenetic maps of causal autoimmune disease variants (Farh et al. 2015).

We assume that $\mathbf{X}$ and $\mathbf{y}$ are the $N \times P$ locus haplotype (or genotype) matrix and the $N \times 1$ vector of phenotype values respectively, with $\mathbf{y}$ possibly already adjusted by relevant covariates. For consistency of exposition, let the SNPs be named $A_1, \ldots, A_P$. Recall the goal of fine-mapping is to return information about which SNP(s) in the set $\{A_1, \ldots, A_P\}$ is (are) causal, given the input pair $\{\mathbf{X}, \mathbf{y}\}$ and possibly other external information such as functional annotations or, more generally, prior biological knowledge.

Overview. Probabilistic Identification of Causal SNPs (PICS) is a Bayesian, non-parametric approach to fine-mapping. Given the observed patterns of association at a locus, and furthermore not assuming a parametric model relating the causal variants to the trait itself, one can estimate the probability that any SNP is causal by performing permutations that preserve its marginal association with the trait as well as the LD patterns at the locus.

To see how this is accomplished, suppose that A_1 is the lead SNP. We are interested in $\mathbb{P}(A_i^{\text{causal}} | A_1 = A^{\text{lead}})$, the probability that A_i is causal, for each $i \in [P]$. By Bayes' theorem,

$$\mathbb{P}(A_i^{\text{causal}} | A_1 = A^{\text{lead}}) \propto \mathbb{P}(A_1 = A^{\text{lead}} | A_i^{\text{causal}}) \times \mathbb{P}(A_i^{\text{causal}}). \tag{1}$$

Focusing on the focal SNP i, permute the rows of $\mathbf{X}$ such that the association between the focal SNP and the trait $\mathbf{y}$ is invariant; see Supplementary Material Section S1 for a concrete mathematical description of this set of constrained permutations. Then the first term on RHS of eq. (1), $\mathbb{P}(A_1 = A^{\text{lead}} | A_i^{\text{causal}})$, is estimated by the proportion of all permutations where A_1 emerges as the lead SNP. The second term, $\mathbb{P}(A_i^{\text{causal}})$, is the prior probability of the focal SNP being causal, which the user can choose based on prior knowledge. The default setting in PICS is $\mathbb{P}(A_1^{\text{causal}}) = \ldots = \mathbb{P}(A_P^{\text{causal}})$. Figure 5 provides a visual summary of the method.

By running the permutation procedure across all SNPs in the locus, one obtains $\mathbb{P}(A_i^{\text{causal}} | A_1 = A^{\text{lead}})$ for each i. These posterior probabilities are then normalized so that $\sum_{i=1}^{N} \mathbb{P}(A_i^{\text{causal}} | A_1 = A^{\text{lead}}) = 1$.

Finally, the above procedure is performed after restricting the set of all SNPs to only those with correlation magnitude $|r| > 0.5$ to the lead SNP.

[1] https://www.ebi.ac.uk/biostudies/files/E-GEUV-1/E-GEUV-1/analysis_results/GeuvadisRNASeqAnalysisFiles_README.txt

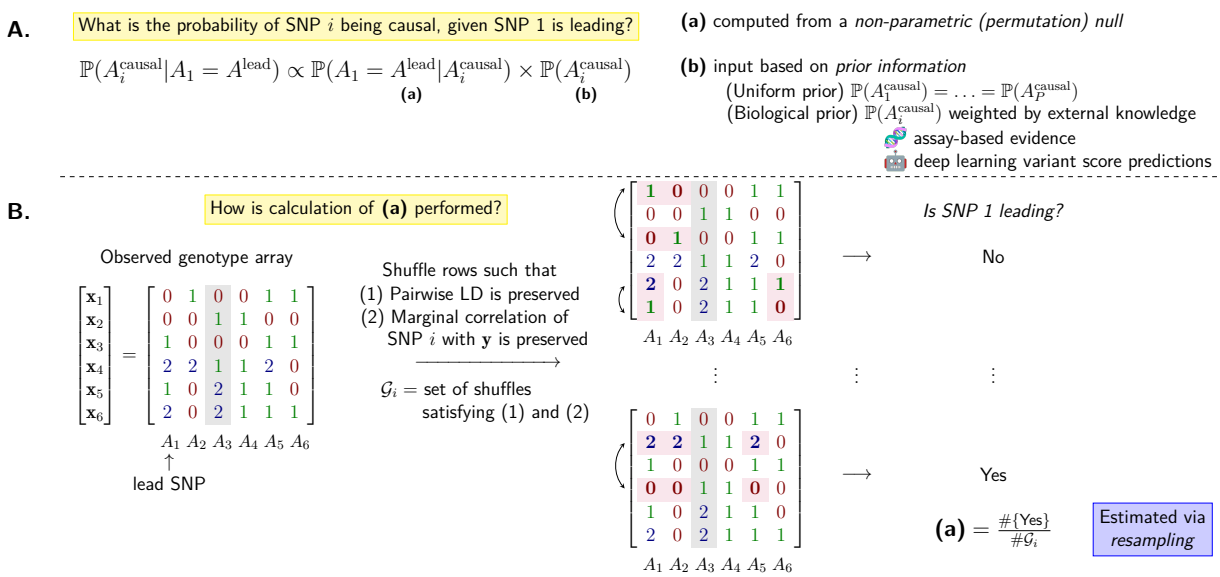

Figure 5: Visual summary of the PICS algorithm described in Subsection 4.1. **A.** Breakdown of the calculation of the probability of a focal SNP A_i being causal. **B.** Illustration of the permutation procedure used to generate the null distribution. An example $N \times P$ genotype array with $N = P = 6$ is used, with two valid row shuffles, or permutations, of the original array shown. Entries affected by the shuffle are highlighted, as is the focal SNP (A_3).

Algorithm. Based on the computation of posterior probabilities described above (eq. (1)), the full PICS algorithm returns putatively causal variants as follows.

Algorithm 1 PICS

Input: Individual-by-genotype array $\underset{N\times P}{\mathbf{X}}$, phenotype array $\underset{N\times 1}{\mathbf{y}}$, LD threshold $|r|$, resampling number R, number of potential sets C, prior causal probabilities $\underset{P\times 1}{\mathbf{p}_0}$.
$c \leftarrow 1$
while $c \leqslant C$ **do**
  $M \leftarrow$ No. columns of $\mathbf{X}$
  Identify lead SNP, $\ell \in \{1, \ldots, M\}$
  Identify neighbouring SNPs, $\mathcal{N}_\ell^c = \{k \in [M] : \text{cor}(A_k, A_\ell)^2 > r^2\}$.
  for $j \in \mathcal{N}_\ell^c$ **do**
    Compute $\widehat{p_j} = \mathbb{P}(A_j^{\text{causal}}|A_\ell = A^{\text{lead}}))$ as described below eq. (1). Use R permutations.
  end for
  Record vector of posterior probabilities, $\overrightarrow{p^c} = [\widehat{p_j} : j \in \mathcal{N}_\ell^c]$
  $\mathbf{X} \leftarrow \mathbf{X}[:, \{1, \ldots, M\} \setminus \mathcal{N}_\ell^c]$
  $c \leftarrow c + 1$
end while
Output: $[\overrightarrow{p^c} : c \leqslant C]$ (a list, where component c corresponds to Potential Set c)

The last line of Algorithm 1 returns a list of posterior probability vectors corresponding to each potential set. For our work, we set $|r| = 0.5$, $C = 3$, $R = 500$ and $\mathbf{p}_0 = (1/P, \ldots, 1/P)$ throughout implementations of Algorithm 1.

As mentioned in the Introduction (Section 1), we apply two different approaches to implementing Algorithm 1. One is guided by stability and will be introduced shortly in Subsection 4.2. The other, which we now describe, is based on regressing out confounders, i.e., residualization. In implementing the residualization approach, we regress the top five principal components, obtained from the genotype matrix $\mathbf{X}$, from the gene expression phenotype, $\mathbf{y}$. We choose five principal components based on the elbow method, an approach described in Brown et al. (2018). This yields residuals $\mathbf{y}^r$ (Figure 1A), which we use as the input phenotype array in Algorithm 1. We subsequently report the variant with highest posterior probability in each potential set as the putatively causal variant for that potential set. This variant is referred to as the top variant.

We remark that there are other ways of using potential confounders in variable selection, such as inclusion as covariates in a linear model. Because our algorithm explicitly avoids assuming a linear model, we have chosen the residualization approach just described.

4.2 Incorporating Stability

Our stability-guided approach to implementing PICS follows the steps outlined below. Let $(\mathbf{X}, \mathbf{y})$ be the genotype array and gene expression array. Assume that there are K subpopulations making up the dataset, and let E_k denote the set of row indices of $\mathbf{X}$ corresponding to individuals from subpopulation k. (In our present work, $K = 5$. The five subpopulations are Utahns ($N_{\text{CEU}} = 89$), Finns ($N_{\text{FIN}} = 92$), British ($N_{\text{GBR}} = 86$), Toscani ($N_{\text{TSI}} = 91$) and Yoruban ($N_{\text{YRI}} = 87$).)

1. Run Algorithm 1 on the pair $(\mathbf{X}, \mathbf{y})$. Obtain list of posterior probability vectors, $\mathcal{L} = [\overrightarrow{p^c} : c \leqslant C]$.

2. For each subpopulation $k = 1, \ldots, K$, run Algorithm 1 on the pair $(\mathbf{X}|_{E_k}, \mathbf{y}|_{E_k})$. Obtain $\mathcal{L}^{(k)} = [\overrightarrow{p_k^c} : c \leqslant C]$ for each k.

3. (Stability-guided choice of putatively causal variants) Collect the probability vectors in lists $\mathcal{L}$ and $\mathcal{L}^{(k)}$ ($k \in [K]$). Operationalizing the principle that a stable variant has positive probability across multiple slices, we pick causal variants as follows. For potential set c, pick the variants that have (i) positive probability in $\overrightarrow{p^c}$ in Step 1, and moreover (ii) have positive probability in the most number of probability vectors; call this set of variants $\mathcal{S}_c$ (note $\mathcal{S}_c$ is a subset of $\overrightarrow{p^c}$). Among members of $\mathcal{S}_c$, select the variant that had the highest posterior probability in $\overrightarrow{p^c}$.

In other words, the stability-guided approach reports the variant that not only appears with positive probability in the most number of subsets including the pooled sample, but also has the largest probability in the pooled sample. This variant is referred to as the stable variant.

To illustrate the last step, suppose there are $K = 2$ subpopulations, $C = 1$ potential set, and $P = 5$ SNPs. Let the outputs from Steps 1 and 2 be

$$\begin{aligned} \mathcal{L}: &\quad \overrightarrow{p^1} = (0.45, 0.43, 0.02, 0.10, 0.00) \\ \mathcal{L}^{(1)}: &\quad \overrightarrow{p_1^1} = (0.00, 0.40, 0.10, 0.25, 0.25) \\ \mathcal{L}^{(2)}: &\quad \overrightarrow{p_2^1} = (0.20, 0.60, 0.00, 0.10, 0.10) \end{aligned}$$

In this example, the second and fourth variants have positive probability in the most number of probability vectors ($\mathcal{S}_1 = \{A_2, A_4\}$). Among A_2 and A_4, A_2 has a higher posterior probability (i.e., 0.43) in $\overrightarrow{p^1}$, so the stable variant reported is A_2. As a side remark, the posterior probability vector

Table 2: List of 6 moderating factors considered.

Moderator	**Quantity/Statistic Computed**
(1) Degree of Stability	No. subpopulations for which stable variant has positive probability
(2) Population Diversity	Maximum of pairwise allele frequency difference between subpopulations for which stable variant has positive posterior probability
(3) Population Differentiation	Maximum F_{ST} between subpopulations for which stable variant has positive posterior probability
(4) Inclusion of Distal Subpopulations (Top)	Whether or not the top variant also had positive probability in Yoruban subpopulation when the stability-guided approach was used
(5) Inclusion of Distal Subpopulations (Stable)	Whether or not the stable variant had positive probability in Yoruban subpopulation when the stability-guided approach was used
(6) Degree of Certainty of Causality Using Residualization Approach	Posterior probability of top variant

$\mathcal{L}$ does not generally agree with the posterior probability vector computed using the residualization approach, because the latter is computed by running Algorithm 1 on the pair $(\mathbf{X}, \mathbf{y}^r)$ rather than $(\mathbf{X}, \mathbf{y})$, as described earlier.

4.3 Statistical Comparison Methodology

We rely on 378 external functional annotations to interpret biological significance of our variants, summarized in Table 1. Supplementary Material Section S2 provides interpretation for the functional annotations, while Supplementary Material Section S3 describes in greater detail how we generate annotations from Enformer predictions.

Our comparison of the top and stable variants is two-fold. First, we evaluate the relative significance of the stable variant against the top variant, by running paired Wilcoxon two-sample tests across all pairs of top and stable variants across all GEUVADIS gene expression phenotypes. We compute one-directional p-values in either direction to check for significant depletion or enrichment of the stable variant with respect to a particular annotation. Raw p-values are adjusted for false discovery rate control by applying the standard Benjamini-Hochberg (BH) procedure (R command `p.adjust(...,method='BH')`) to p-values across all 378 annotations and all potential sets.

To investigate whether various external factors might moderate the differences in functional annotations, we next perform trend tests. Basically, for some external factor F, we run a trend test to see if values of F are associated with attenuation or augmentation of differences in functional annotation of the top and stable variants. The list of all external factors F is provided in Table 2.

For (1) Degree of Stability, we perform a Jonckheere-Terpstra test (R command `clinfun::jonckheere.test(...,nperm=5000)`). For (2) Population Diversity, (3) Population Differentiation and (6) Degree of Certainty of Causality Using Residualization Approach, we perform a correlation test (R command `cor.test(...)`). For (4) Inclusion of Distal Subpopulations (Top) and (5) Inclusion of Distal Subpopulations (Stable), we perform an unpaired Wilcoxon test (R command `wilcox.test(...)`). Finally, for each moderator, we again compute one-directional p-values in either direction, before applying the BH procedure to all p-values across annotations and potential

sets for that moderator only.

Data Availability

We provide the following data and scripts on the Github repository https://github.com/songlab-cal/StableFM: fine-mapped variants with their functional annotations and moderator variable quantities, code for reproducing figures in this manuscript and building our visualization app.

Author Contributions

AA, NI, and YSS designed the study. AA performed data analyses and developed the interactive visualization app. LJ developed the interactive visualization app. NI and YSS supervised the research.

Acknowledgements

This research is supported in part by an NIH grant R35-GM134922 and grant number CZF2019-002449 from the Chan Zuckerberg Initiative Foundation. We thank Carlos Albors, Gonzalo Benegas, Ryan Chung, Ziyue Gao, Iain Mathieson and Yutong Wang for helpful discussions, members of the Yu Group at Berkeley for feedback on the work, and Aniketh Reddy for help with data processing.

References

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Abdel Abdellaoui, Conor V Dolan, Karin JH Verweij, and Michel G Nivard. Gene–environment correlations across geographic regions affect genome-wide association studies. *Nature Genetics*, 54(9):1345–1354, 2022.

Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4):248–249, 2010.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021.

Suganthi Balasubramanian, Yao Fu, Mayur Pawashe, Patrick McGillivray, Mike Jin, Jeremy Liu, Konrad J Karczewski, Daniel G MacArthur, and Mark Gerstein. Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes. *Nature Communications*, 8 (1):1–11, 2017.

Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Brielin C Brown, Nicolas L Bray, and Lior Pachter. Expression reflects population structure. *PLoS Genetics*, 14(12):e1007841, 2018.

Taylor B Cavazos and John S Witte. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Human Genetics and Genomics Advances*, 2(1): 100017, 2021.

Fiona Cunningham, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Ruth Bennett, et al. Ensembl 2022. *Nucleic Acids Research*, 50(D1):D988–D995, 2022.

Eugene V Davydov, David L Goode, Marina Sirota, Gregory M Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Computational Biology*, 6(12):e1001025, 2010.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

Kyle Kai-How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J Housley, Samantha Beik, Noam Shoresh, Holly Whitton, Russell JH Ryan, Alexander A Shishkin, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539): 337–343, 2015.

Marie-Julie Favé, Fabien C Lamaze, David Soave, Alan Hodgkinson, Héloïse Gauvin, Vanessa Bruat, Jean-Christophe Grenier, Elias Gbeha, Kimberly Skead, Audrey Smargiassi, et al. Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nature Communications*, 9(1):1–12, 2018.

Yao Fu, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Jasmine Mu, Kevin Y Yip, Ekta Khurana, and Mark Gerstein. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology*, 15(10):1–15, 2014.

Buhm Han and Eleazar Eskin. Interpreting meta-analyses of genome-wide association studies. *PLoS Genetics*, 8(3):e1002555, 2012.

Yi-Fei Huang, Brad Gulko, and Adam Siepel. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics*, 49(4):618–624, 2017.

Nilah M Ioannidis, Joe R Davis, Marianne K DeGorter, Nicholas B Larson, Shannon K McDonnell, Amy J French, Alexis J Battle, Trevor J Hastie, Stephen N Thibodeau, Stephen B Montgomery, et al. FIRE: Functional inference of genetic variants that regulate gene expression. *Bioinformatics*, 33(24):3895–3901, 2017.

Kevin L Keys, Angel CY Mak, Marquitta J White, Walter L Eckalbar, Andrew W Dahl, Joel Mefford, Anna V Mikhaylova, María G Contreras, Jennifer R Elhawary, Celeste Eng, et al. On the cross-population generalizability of gene expression prediction models. *PLoS Genetics*, 16 (8):e1008927, 2020.

Nathan LaPierre, Kodi Taraszka, Helen Huang, Rosemary He, Farhad Hormozdiari, and Eleazar Eskin. Identifying causal variants by fine mapping across multiple studies. *PLoS Genetics*, 17 (9):e1009733, 2021.

Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC t Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

S Li, M Sesia, Y Romano, E Candès, and C Sabatti. Searching for robust associations with a multi-environment knockoff filter. *Biometrika*, 109(3):611–629, 2022.

Yanyu Liang, Milton Pividori, Ani Manichaikul, Abraham A Palmer, Nancy J Cox, Heather E Wheeler, and Hae Kyung Im. Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. *Genome Biology*, 23(1):1–18, 2022.

Chinghway Lim and Bin Yu. Estimation stability with cross-validation (ESCV). *Journal of Computational and Graphical Statistics*, 25(2):464–492, 2016.

Zeyun Lu, Shyamalika Gopalan, Dong Yuan, David V Conti, Bogdan Pasaniuc, Alexander Gusev, and Nicholas Mancuso. Multi-ancestry fine-mapping improves precision to identify causal genes in transcriptome-wide association studies. *The American Journal of Human Genetics*, 109(8): 1388–1404, 2022.

Carla Márquez-Luna, Po-Ru Loh, South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium, and Alkes L Price. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genetic Epidemiology*, 41(8):811–823, 2017.

Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, 5(5):e1000471, 2009.

Andrew P Morris. Transethnic meta-analysis of genomewide association studies. *Genetic Epidemiology*, 35(8):809–822, 2011.

Hakhamanesh Mostafavi, Arbel Harpak, Ipsita Agarwal, Dalton Conley, Jonathan K Pritchard, and Molly Przeworski. Variable prediction accuracy of polygenic scores within an ancestry group. *Elife*, 9:e48376, 2020.

Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814, 2003.

Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

Philipp Rentzsch, Daniela Witten, Gregory M Cooper, Jay Shendure, and Martin Kircher. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, 47(D1):D886–D894, 2019.

Mark F Rogers, Hashem A Shihab, Matthew Mort, David N Cooper, Tom R Gaunt, and Colin Campbell. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics*, 34(3):511–513, 2018.

Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

Kimberly E Taylor, K Mark Ansel, Alexander Marson, Lindsey A Criswell, and Kyle Kai-How Farh. PICS2: next-generation fine mapping via probabilistic identification of causal SNPs. *Bioinformatics*, 37(18):3004–3007, 2021.

Patrick Turley, Alicia R Martin, Grant Goldman, Hui Li, Masahiro Kanai, Raymond K Walters, Jonathan B Jala, Kuang Lin, Iona Y Millwood, Caitlin E Carey, et al. Multi-ancestry meta-analysis yields novel genetic discoveries and ancestry-specific associations. *BioRxiv*, pages 2021–04, 2021.

Qingbo S Wang, David R Kelley, Jacob Ulirsch, Masahiro Kanai, Shuvom Sadhuka, Ran Cui, Carlos Albors, Nathan Cheng, Yukinori Okada, Francois Aguet, et al. Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nature Communications*, 12(1):1–11, 2021.

Xiaoquan Wen, Francesca Luca, and Roger Pique-Regi. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genetics*, 11(4):e1005176, 2015.

Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

Bin Yu and Karl Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 117(8):3920–3929, 2020.

Arslan A Zaidi and Iain Mathieson. Demographic history mediates the effect of stratification on polygenic scores. *Elife*, 9:e61548, 2020.

Hufeng Zhou, Theodore Arapoglou, Xihao Li, Zilin Li, Xiuwen Zheng, Jill E Moore, Abhijith Asok, Sushant Kumar, Elizabeth E Blue, Steve Buyske, et al. FAVOR: Functional annotation of variants online resource and annotator for variation across the human genome. *bioRxiv*, 2022.

Supplementary Material

The Impact of Stability Considerations on Genetic Fine-Mapping

Alan Aw[1,2], Lionel Chentian Jin[3], Nilah Ioannidis[2,4,*], Yun S. Song[1,2,4,*]

[1] Department of Statistics, University of California, Berkeley
[2] Center for Computational Biology, University of California, Berkeley
[3] McKinsey & Company, Seattle
[4] Computer Science Division, University of California, Berkeley

March 23, 2023

S1 Permuting while preserving marginal correlation and LD: a Formal Description

For a focal SNP i, its N entries $(x_{1i}, \ldots, x_{Ni})$ take on only D values, where D is the ploidy of the data (for human genotypes $D = 2$ and for human haplotypes $D = 1$). Suppose $D = 2$ for exposition. For $d = 0, 1, 2$, let $I_d \subset [N]$ denote the indices of those elements of the allelic dosage vector $(x_{1i}, \ldots, x_{Ni})$ taking on value d. Then, any permutation $\sigma \in \mathfrak{S}_N$ of the rows of $\mathbf{X}$ that preserves marginal correlation with the phenotype and LD maps indices from I_d to I_d, because

- Each distinct entry in the allelic dosage vector of the focal SNP is still assigned the the same phenotype entry;
- Row shuffles do not change column-column covariances, the latter of which determines LD.

To be precise, $\sigma \in \mathfrak{S}_{|I_0|} \times \mathfrak{S}_{|I_1|} \times \mathfrak{S}_{|I_2|}$, a direct product of the three permutation groups corresponding to the different allelic dosages.

In practice, we approximate the permutation distribution by sampling permutations uniformly at random from the group defined above. We set the sampling number to 500 for all our fine-mapping experiments.

S2 List of Annotations

We measure the biological significance of a variant using a wide range of available functional annotations (Table S1). The annotations cover potential biological activity in the local vicinity of the variant position along the genome, estimated model-based biological quantities (e.g., selection and conservation scores), and predicted effects of mutagenesis from the reference to the alternate allele at the variant.

*To whom correspondence should be addressed: nilah@berkeley.edu, yss@berkeley.edu

Table S1: List of annotations used in our study, alongside their interpretations.

Functional Annotation	Interpretation
Distance to Canonical Transcription Start Site (TSS)	-
Percent CpG in 75 bp window centered on variant position	-
Percent GC in 75 bp window centered on variant position	-
CTCF Binding Enrichment	whether the variant lies within a CTCF binding site region as predicted by Ensembl
Enhancer Enrichment	whether the variant lies within an enhancer region as predicted by Ensembl
Open Chromatin Enrichment	whether the variant lies within an open chromatin region as predicted by Ensembl
Promoter Enrichment	whether the variant lies within a promoter region as predicted by Ensembl
TF Binding Enrichment	whether the variant lies within a TF binding site region as predicted by Ensembl
Promoter Flanking Enrichment	whether the variant lies within a promoter flanking region as predicted by Ensembl
CADD (2 scores)	whether the variant is likely to be simulated or not, and hence likely deleterious or not. One score is raw while the other is rank-normalized
SIFTVal	whether the variant affects protein function, and hence deleterious
Polyphen2	posterior probability that the variant is damaging
LINSIGHT	probability that the variant site is under selection, thus having functional consequence
PhyloP (3 scores)	substitution rates measuring cross-species evolutionary conservation at the site of the variant. Each score is computed with respect to a clade (vertebrate, mammal, primate)
GerpN	estimated neutral substitution rate at variant position, with higher value implying greater conservation
GerpS	estimated rejected substitution rate at variant position, with positive value implying a deficit in substitutions
B Statistic	background selection at variant position, with smaller value indicating larger impact of selection
FATHMM-XF	integrative score measuring deleteriousness of the variant
Funseq2	integrative score measuring deleteriousness of the variant
ALoft	integrative score measuring loss of function associated with the variant
FIRE	integrative score measuring deleteriousness of the variant
Magnitude of Effect on Enformer Track Prediction (177 tracks)	change in prediction of a gene regulatory track when performing in-silico mutagenesis on the variant in a 196,608 bp sequence

S3 Generating Annotations from Enformer Predictions

The Enformer (Avsec et al., 2021) is a sequence-based prediction model, which leverages the attention mechanism in a transformer neural network to capture long-range effects on gene regulation.

For Enformer predictions, we subset the $5,313$ ChIP-seq, DNase-seq, ATAC-seq and CAGE tracks to only those relevant to the lymphoblastoid cell line (GM12878), the cell line with respect to which GEUVADIS gene expression phenotypes are measured. We also restrict to genes whose top and stable variants are within the $196,608$ bp input length limit of the Enformer from the corresponding GEUVADIS gene's TSS. This restriction allows us to obtain Enformer predictions on three sequences: a *null sequence*, a sequence where the REF allele is replaced with the ALT allele at the top variant (*top sequence*), and a sequence where the REF allele is replaced with the ALT allele at the stable variant (*stable sequence*).

To compare the top and the stable variants with respect to a particular track, we take the predictions obtained from the three input sequences (null prediction, top prediction, stable prediction — a *triplet*), and compute the magnitude of change in predictions between both the top and null and the stable and null. We use the magnitude of change here, rather than the change itself, to capture the impact the mutation associated with the variant has on the track. Our Enformer functional annotations can thus be interpreted as measures of mutagenic impact on a track's prediction, without considering the directionality of the impact. We refer to these annotations as *perturbation scores*.

We compute perturbation scores in two ways. The first way is to center all input sequences on the transcription start site of the gene (occasionally referred to as 'TSS' in our work), thus allowing the model to measure the impact of mutagenesis at the (top or stable) variant on predicted gene expression profile for the sequence. The second way is to average the perturbation scores over three triplets, one centered on the gene TSS, and two centered on the flanking positions of the gene TSS (i.e., the two neighbouring bins). The second way (occasionally referred to as 'AVE' in our work) accounts for errors arising from imprecise TSS positioning and the instability of the model to small changes in input — a technique used in Karollus et al. (2022) and by the Enformer team (Avsec et al., 2021).

S4 Matching versus Non-matching Variant Results

As described in the main text, we compare between one set of genes where the top and stable variants match and another set of genes where the top and stable variants disagree; comparisons are made between matching variants and one of the non-matching sets of variants (top or stable).

Concretely, for Potential Set 1 we compare between $N_{\text{match}} = 12,743$ genes with matching variants and $N_{\text{non-match}} = 9,921$ genes with non-matching variants; for Potential Set 2 we compare between $N_{\text{match}} = 8,197$ genes with matching variants and $N_{\text{non-match}} = 14,452$ genes with non-matching variants; and for Potential Set 3 we compare between $N_{\text{match}} = 5,807$ genes with matching variants and $N_{\text{non-match}} = 16,812$ genes with non-matching variants. Below, we summarize our findings.

Matching Variants vs Non-matching Top Variants. We observe the following significant trends: as mentioned in the main text, for Potential Set 1 361 functional annotations reported significantly higher scores for the matching variants, with all BH-adjusted p-values less than 0.05. The remaining $17 = 378 - 361$ annotations *not* exhibiting significantly higher scores are: Distance to Canonical TSS, CTCF Binding Enrichment, Enhancer Enrichment, Open Chromatin Enrichment, TF Binding Enrichment, Promoter Flanking Enrichment, CADD raw score, PHRED-normmalized

CADD score, SIFTVal, priPhyloP, mamPhyloP, verPhyloP, GerpS, LINSIGHT, Funseq2, ALoft, and FIRE.

For Potential Set 2, 132 functional annotations reported significantly higher scores for the matching variants. For Potential Set 3, 9 functional annotations reported significantly higher scores for the matching variants: B Statistic, percent GC content, and Enformer tracks `ENCFF776DPQ`, `ENCFF831ZHL`, `ENCFF601YET`, `ENCFF945XXY`, `ENCFF676UXN`, `ENCFF151LGF`, `ENCFF876DXW`. For all three potential sets, there were no functional annotations reporting significantly lower scores for the matching variants, even for annotations where a low score implies greater biological functionality (e.g., B Statistic).

Matching Variants vs Non-matching Stable Variants. We observe the following significant trends: as mentioned in the main text, for Potential Set 1 363 functional annotations reported significantly higher scores for the matching variants, with all BH-adjusted p-values less than 0.05. The remaining 15 annotations *not* exhibiting significantly higher scores are: Distance to Canonical TSS, CTCF Binding Enrichment, Enhancer Enrichment, Open Chromatin Enrichment, TF Binding Enrichment, Promoter Flanking Enrichment, CADD raw score, PHRED-normmalized CADD score, SIFTVal, mamPhyloP, verPhyloP, GerpS, LINSIGHT, ALoft, and FIRE.

For Potential Set 2, 359 functional annotations reported significantly higher scores for the matching variants (all BH-adjusted p-values < 0.05), while only one functional annotation — Distance to Canonical TSS — reported significantly lower scores for the matching variants (BH-adjusted p-value $= 8 \times 10^{-4}$). For Potential Set 3, 259 functional annotations reported significantly higher scores for the matching variants (all BH-adjusted p-values < 0.05), while only one functional annotation — Distance to Canonical TSS — reported significantly lower scores for the matching variants (BH-adjusted p-value $= 8 \times 10^{-4}$).

S5 Top Variant versus Stable Variant Results

As described in the main text, we compare, across all genes, the annotations of the top and the stable variant. We reported no significant differences in enrichment or trends in enrichment differences driven by moderators for Potential Set 1. For completeness, we report here significant results for all three potential sets.

Paired analysis. After BH-adjustment, there are no significant differences across all three potential sets using a threshold of $\alpha = 0.05$.

Trend analysis. For the following moderators, we observe some significant trends.

- Inclusion of Distal Subpopulations (Top)

  For Potential Set 3, FIRE scores exhibited a greater difference between stable and top variant scores (i.e., stable score − top score) for genes with top variants not discovered in YRI than those with top variants discovered in YRI (one-sided BH-adjusted $p = 0.03$). So did TSS-centered perturbation scores for 121 Enformer tracks and AVE perturbation scores for 121 Enformer tracks. (These tracks and their corresponding BH-adjusted p-values are available at: https://github.com/songlab-cal/StableFM/blob/master/data/results_with_moderators/p_values/mod_Top%20SNV%20Discovered%20in%20YRI.csv.)

- Posterior Probability of Top Variant

For all potential sets, there are small but significant positive correlations between the posterior probability of the top variant and the difference between the stable variant FIRE score and the top variant FIRE score (Potential Set 1: Pearson's $r = 0.05$, BH-adjusted $p = 3 \times 10^{-5}$; Potential Set 2: Pearson's $r = 0.07$, BH-adjusted $p = 3 \times 10^{-14}$; Potential Set 3: Pearson's $r = 0.05$, BH-adjusted $p = 1 \times 10^{-7}$). For all potential sets, there are small but significant negative correlations between the posterior probability of the top variant and the difference between the stable variant Absolute Distance to TSS and the top variant Absolute Distance to TSS (Potential Set 1: Pearson's $r = -0.07$, BH-adjusted $p = 1 \times 10^{-10}$; Potential Set 2: Pearson's $r = -0.05$, BH-adjusted $p = 9 \times 10^{-7}$; Potential Set 3: Pearson's $r = -0.06$, BH-adjusted $p = 2 \times 10^{-12}$). For Potential Set 2, there are significant negative correlations between the posterior probability of the top variant and the difference between the stable variant and top variant Percent GC (Pearson's $r = -0.04$, BH-adjusted $p = 5 \times 10^{-5}$) and FATHMM-XF scores (Pearson's $r = -0.04$, BH-adjusted $p = 7 \times 10^{-4}$). For Potential Set 3, there is a significant negative correlation between the posterior probability of the top variant and the difference between the stable variant and top variant FATHMM-XF scores (Pearson's $r = -0.04$, BH-adjusted $p = 2 \times 10^{-5}$).

S6 Impact of Positive Posterior Probability Support on Results

One key consideration in statistical fine-mapping is the number of variants possessing positive posterior probability, which we refer to as the *positive posterior probability support* (or *support*, for short). Indeed, a larger support may indicate greater uncertainty in the assignment of causal variant to the allele with largest posterior probability. To evaluate the potential utility of the stability-guided approach in settings where the residualization approach leads to large support, we repeat our analysis on two restricted sets of genes, namely (1) those genes where the top variant has support greater than 10; or (2) those genes where the top variant has support greater than 50.

Paired analysis. *Gene Set (1).* For all Potential Sets, Distance to TSS is significantly smaller for top variant (all three BH-adjusted p-values $< 10^{-6}$). For Potential Set 3, FATHMM-XF score is significantly larger for stable variant (BH-adjusted $p < 10^{-6}$). For all Potential Sets, FIRE score is significantly larger for top variant (BH-adjusted p-values: Potential Set 1 = 0.02, Potential Set 2 = 2×10^{-10}, Potential Set 3 = 1.7×10^{-4}). *Gene Set (2).* For all Potential Sets 2 and 3, FIRE score is significantly larger for top variant (both BH-adjusted p-values are 0.04).

Trend analysis.
Gene Set (1). For the following moderators, we observe some significant trends.

- Posterior Probability of Top Variant

  For Potential Sets 2 and 3, there is a significant negative correlation between the posterior probability of the top variant and the difference between the stable variant FIRE score and the top variant FIRE score (Potential Set 2: Pearson's $r = 0.08$, BH-adjusted $p = 1.9 \times 10^{-5}$; Potential Set 3: Pearson's $r = 0.06$, BH-adjusted $p = 0.003$). For Potential Set 2, there is a significant positive correlation between the posterior probability of the top variant and the difference between the stable variant's FATHMM-XF score and the top variant's FATHMM-XF score (Pearson's $r = -0.06$, BH-adjusted $p = 0.01$).

Gene Set (2). For the following moderators, we observe some significant trends.

- Posterior Probability of Top Variant

For Potential Set 2, TSS-centered perturbation scores for 8 Enformer tracks exhibited a positive correlation between the posterior probability of the top variant and the difference between stable and top variant perturbation scores. Track names (with empirical Pearson's r) are `ENCFF915DFR` (-0.40), `ENCFF107LDM` (-0.41), `ENCFF821PRO` (-0.42), `ENCFF171MDW` (-0.43), `ENCFF170NTY` (-0.43), `ENCFF935KTD` (-0.41), `ENCFF676GTP` (-0.41), `ENCFF013ZOI` (-0.42). Averaged perturbation scores for two Enformer tracks also exhibited negative correlation; these two tracks are `ENCFF107LDM` (-0.40) and `ENCFF170NTY` (-0.40). All BH-adjusted p-values are 0.0495.

S7 Impact of Posterior Probability on Results

We perform a comparison between top and stable variants, by restricting to genes where the posterior probability of the top variant or the stable variant exceeds 0.9. Specifically, we repeat our analysis on two restricted sets of genes, namely (1) those genes where the top variant reported a posterior probability exceeding 0.9; or (2) those genes where the stable variant reported a posterior probability exceeding 0.9. For reference, we plot in Figure S2 the joint distribution of posterior probabilities of top and stable variants, across all genes for which the two fine-mapping approaches returned distinct variants.

Paired analysis. *Gene Set (1).* For Potential Set 3, Distance to TSS is significantly smaller for stable variant (BH-adjusted $p = 0.01$). *Gene Set (2).* For Potential Sets 1 and 2, FIRE scores of the top variant are significantly larger than the stable variant (BH-adjusted p-values: Potential Set $1 = 1 \times 10^{-6}$, Potential Set $2 = 0.02$). For Potential Set 2, Distance to TSS of the top variant is significantly smaller (BH-adjusted $p = 3 \times 10^{-4}$). For Potential Set 3, FATHMM-XF scores of the stable variant are significantly larger (BH-adjusted $p = 0.006$).

Trend analysis.
Gene Set (1). For the following moderators, we observe some significant trends.

- Inclusion of Distal Subpopulations (Top)

  For Potential Set 3, TSS-centered perturbation scores for 9 Enformer tracks exhibited a higher difference between stable and top variant perturbation score for genes with top variants not discovered in YRI versus those discovered in YRI. Track names (with BH-adjusted unpaired Wilcoxon test p-values) are `ENCFF279CYY` (0.03), `ENCFF417WYL` (0.04), `ENCFF782WWH` (0.03), `ENCFF629RRF` (0.03), `ENCFF676GTP` (0.03), `ENCFF984HLU` (0.03), `ENCFF700YOH` (0.03), `ENCFF848LJL` (0.0477), `ENCFF038IYA` (0.03). Averaged perturbation scores for 10 Enformer tracks also a higher difference; these are `ENCFF776DPQ` (0.02), `ENCFF279CYY` (0.02), `ENCFF629RRF` (0.03), `ENCFF319YAI` (0.04), `ENCFF676GTP` (0.046), `ENCFF367WTF` (0.04), `ENCFF984HLU` (0.04), `ENCFF917YSR` (0.03), `ENCFF613CYH` (0.04) and `ENCFF700YOH` (0.04).

Gene Set (2). For the following moderators, we observe some significant trends.

- Posterior Probability of Top Variant

  For Potential Set 3, there is a significant negative correlation between the posterior probability of the top variant and the difference between the stable variant Distance to TSS and the top variant Distance to TSS (Pearson's $r = -0.09$, BH-adjusted $p = 0.004$).

References

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18 (10):1196–1203, 2021.

Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *bioRxiv*, 2022.

Supplementary Figures

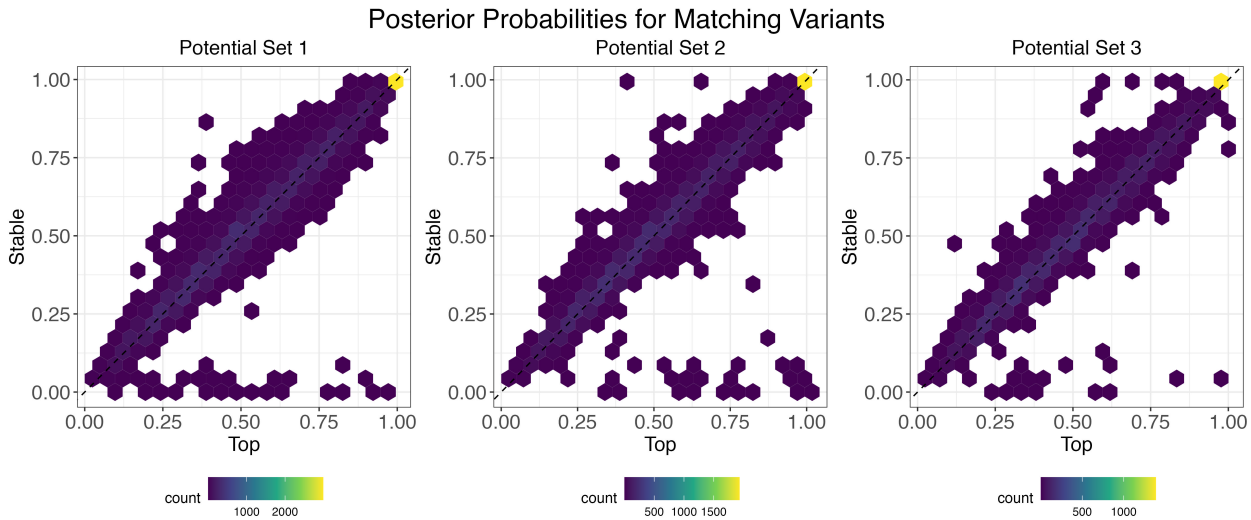

Figure S1: Pair density plot of posterior probabilities of the top variant and the stable variant, in case they match.

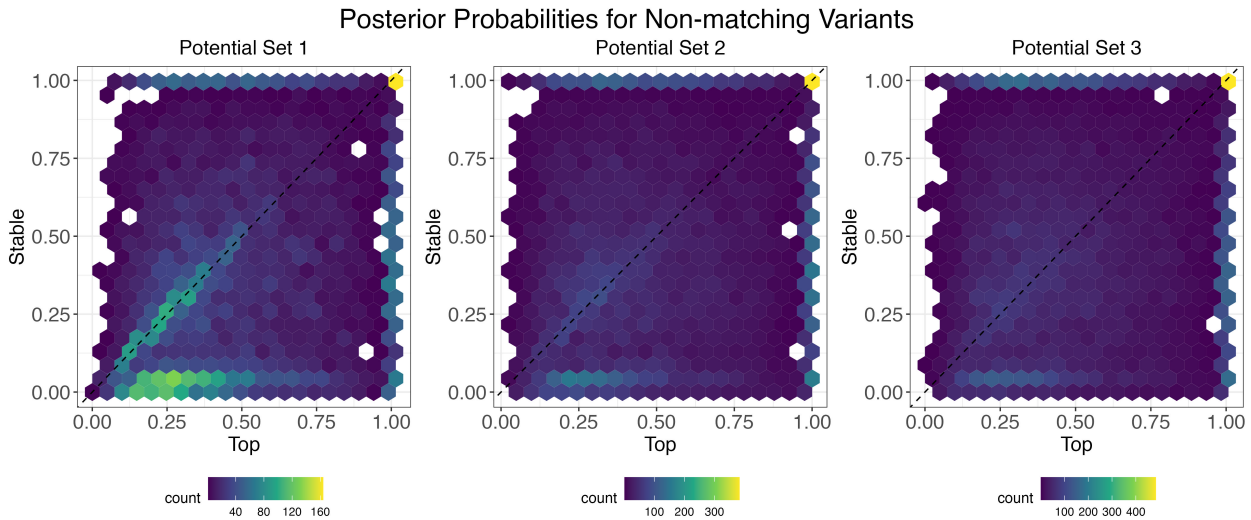

Figure S2: Pair density plot of posterior probabilities of the top variant and the stable variant, in case they do not match.