An Effector Index to Predict Causal Genes at GWAS Loci

Vincenzo Forgetta*[1], Lai Jiang*[1,8], Nicholas A. Vulpescu*[2], Megan S. Hogan[2], Siyuan Chen[1,8], John A. Morris[1,3,4,10], Stepan Grinek[2], Christian Benner[5] , Mark I McCarthy[6‡], Eric Fauman[7], Celia MT Greenwood[†,1,8,9,10], Matthew T. Maurano[†,2], J. Brent Richards[†,1,8,10,11,12]

[1]Centre for Clinical Epidemiology, Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Canada
[2]Institute for Systems Genetics and Department of Pathology, NYU School of Medicine, New York, USA
[3]New York Genome Center, New York, USA
[4]Department of Biology, New York University, New York, USA
[5]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00014 Helsinki, Finland
[6]Wellcome Centre for Human Genetics, University of Oxford, UK
[7]Internal Medicine Research Unit, Pfizer Worldwide Research, Development and Medical, USA
[8]Departments of Medicine, Epidemiology and Biostatistics, McGill University, Montreal, Canada
[9]Gerald Bronfman Department of Oncology, McGill University, Montréal, Canada
[10]Department of Human Genetics, McGill University, Montréal, Canada
[11]Department of Medicine, McGill University, Montréal, Canada
[12]Department of Twin Research, King's College London, London. United Kingstom

[‡]Current address: Genentech, 340 Point San Bruno Boulevard, South San Francisco, CA 94080.

*These authors contributed equally to this work.

[†] Correspondence:
Celia MT Greenwood
celia.greenwood@mcgill.ca

Matthew T Maurano
maurano@nyu.edu

J. Brent Richards, MD, MSc
Professor of Medicine, McGill University
Senior Lecturer, King's College London (Honorary)
Pavillon H-413, Jewish General Hospital
3755 Cote Ste Catherine
Montreal, QC, Canada, H3T 1E2
brent.richards@mcgill.ca
www.mcgill.ca/genepi

ABSTRACT

Drug development and biological discovery require effective strategies to map existing genetic associations to causal genes. To approach this problem, we identified a set of positive control genes for 12 common diseases and traits that cause a Mendelian form of the disease or are the target of a medicine used for disease treatment. We then identified a simple set of genomic features enriching GWAS single nucleotide variants for these positive control genes. Using these features, we trained and validated the Effector Index (*Ei*), a causal gene mapping algorithm using the 12 common diseases and traits. The area under *Ei's* receiver operator curve to identify positive control genes was 80% and area under the precision recall curve was 29%. Using an enlarged set of independently curated positive control genes for type 2 diabetes which included genes identified by large-scale exome sequencing, these areas increased to 85% and 61%, respectively. The best predictors were coding or transcript altering variation, distance to gene and open chromatin-based metrics. We provide the *Ei* algorithm for its widespread use. *Ei* provides a simple, understandable tool to prioritize genes at GWAS loci for functional follow-up and drug development.

INTRODUCTION

The majority of late-stage drug development programs fail.[1–3] The most common cause of these failures is a lack of efficacy of the medicine on the disease outcome.[2] Such failures are due, in part, to unreliable drug target identification and validation.[4] Recent evidence has suggested that when drug development programs have support from human genetics the probability of success increases.[5–7] Yet, this evidence relies upon retrospective surveys of drug development programs and assumes that human genetic associations, which are most often non-coding, could be used to identify the specific causal gene(s) at an associated locus—while in reality this is challenging. Moving forward, the use of human genetics for target identification and validation will require that a genetic association be mapped to specific genes. Therefore, strategies to reliably map the thousands of GWAS associations to causal genes are required to realize the potential of human genetics to deliver medicines to the clinic.

An idealized strategy to map genetic associations to causal genes should have several characteristics. First, the predictive features that enable mapping a genetic association to a causal gene would be easily and reliably measured, shared across diseases and available from public datasets. By using a parsimonious set of simply measured predictors, such a mapping tool would be more easily applicable. Second, the strategy would be generalizable across diseases, enabling its use even in diseases where few causal genes are known. Third, the strategy would be evaluated by testing its predictive ability against a set of truly causal positive control genes. Without such a set of causal positive control genes, the value of any strategy cannot be properly examined. Last, the strategy would be evaluated in a dataset distinct from the training dataset. We therefore developed the "Effector Index (*Ei*)", an algorithm which generates the probability of causality for all genes at a GWAS locus. Specifically, the *Ei* aims to answer the question, "What is the probability of causality for each gene at a locus which harbors genome-wide significant SNVs for a disease or trait?".

Here we present this *Ei* and describe its ability to identify positive control causal genes. To generate this index, we first carefully defined a set of positive control causal genes for 12 diseases and traits by relying only upon data from humans, defining positive control genes as genes whose perturbation causes a Mendelian form of the common disease, or whose encoded protein acts as a drug target for the common disease. We next assessed the genomic annotations at GWAS loci that enriched for positive control genes. Then, to predict each gene's probability of causality at a GWAS locus we used both locus-level features, such as the number of genes at a locus, and gene-level features, such as distance of a gene to nearest associated SNV. We trained a gradient boosted trees algorithm, as implemented in XGBoost[8] to generate the probability of causality for each gene at each GWAS locus for 11 diseases and tested the resulting model on the 12th disease, iterating this process across all 12 diseases. As a sensitivity analysis, for type 2 diabetes (T2D) we tested an enlarged a set of positive control genes that, in addition to the positive controls as described above, included independently and manually curated positive controls, incorporating recent evidence from coding variants arising from exome array and exome sequencing studies[9] (http://www.type2diabetesgenetics.org/).

The *Ei* provides a rapid and readily understandable way to assess the probability that a gene at a GWAS locus is causal for a disease or trait. It requires relatively simple annotations from GWAS summary statistics, genomic positions and DNaseI hypersensitive sites (DHS). Not only do we expect these findings to accelerate drug development, by providing guidance on which gene(s) at a GWAS locus are likely to be causal, but they should also assist in prioritizing genes for downstream functional genomic and biological experiments. Below we present the key results of the study.

RESULTS

Diseases and Traits Studied, and Positive Control Genes

We selected a broad panel of diseases and quantitative traits which have been the subject of large-scale GWASs, such as those from UK Biobank or large-scale international GWAS consortia[10] (**Table 1**). The diseases and traits studied included: T2D, low density lipoprotein (LDL) cholesterol level, adult height, calcium level, hypothyroidism, triglyceride (Tg) level, estimated bone mineral density (eBMD), glucose level, red blood cell count (RBC) systolic blood pressure (SBP), diastolic blood pressure (DBP) and direct bilirubin level.

We required that for each disease or trait, a set of "positive control genes" could be identified as the targets of drugs used to treat the disease (or influence the trait) (**Table S1**), and/or genes known to cause Mendelian forms of the disease (or influence the trait) (**Tables S2 & S3**). We stringently defined as "positive control genes" those genes meeting at least one of two criteria: 1) perturbation of the gene is known to cause of a Mendelian form of the disease (or influences the trait); or 2) the gene's protein is the target of a therapy successfully developed to treat the disease, or influence the trait. To identify Mendelian disease genes, we first used the Human Disease Ontology[11] database to identify diseases influencing traits studied. Using a resultant list of curated ontological terms we obtained a list of Mendelian disease genes from the Online Mendelian Inheritance in Man (OMIM; see **Online Methods**).[12] Positive control drug targets were identified by first collecting guideline-suggested medications for each trait or disease, as described in UpToDate®, an online decision tool written and edited by medical experts,[13] then we gathered the known targets of these medicines using DrugBank.[14]

We identified 494 positive control genes across the 12 diseases and traits, 381 known to cause Mendelian forms of the disease (or influence the trait) and 113 drug targets (**Tables S3 & S4**). This ranged from two for direct bilirubin to 66 for RBC, with an average of 32 per disease/trait. Of the 113 positive control genes from drug targets, we found an average of 9 drug classes per disease trait. This represented 55 unique drug targets, since different classes of clinically used medicines may have the same drug target (**Table S4**).

Fine-mapping GWAS loci

We applied a statistical fine mapping pipeline to prioritize SNVs from previously published GWASs for T2D,[10] eBMD[15] and *de-novo* GWASs for the other diseases and traits from the UK Biobank (**Fig. 1a**) (**Table 1**).[16] Fine-mapping is more helpful in this context than conditional analyses since it provides probabilistic measures of causality for SNVs.[17] We selected GWASs

that used only individuals of European descent (see **Online Methods** for further details) to enable fine-mapping. SNVs passing quality control and having a minor allele frequency > 0.005 were retained for LD clumping, followed by merging with adjacent association signals (see **Online Methods** for further details). We applied the program FINEMAP[17] using a matching LD panel comprised of 50,000 individuals of white-British ancestry from UK Biobank. These GWASs yielded between 78 and 1,011 independent loci where the average was 366 (**Table 1**). Therefore, each disease or trait had a large number of loci which could be used to train or test the predictive model.

Fine-mapping dramatically reduced the size of the putatively causal SNV set relative to the initial GWAS (**Supplementary Fig. 1a**). Fully 48% of most strongly associated lead SNVs also had the highest FINEMAP $\log_{10}$(Bayes Factor (BF)) at a locus, and an average of 55% of these lead SNVs were also in the fine-mapped credible set (**Supplementary Fig. 1b-c**). There was an average of 3.7 fine-mapped SNVs per locus (**Supplementary Fig. 2**), suggesting a large degree of allelic heterogeneity. Consequently, this fine-mapping step substantially reduced the number of SNVs to be considered for mapping to causal genes.

Statistical fine mapping strongly enriches for positive control genes

We assessed the utility of fine-mapping for enrichment for positive control genes by comparing enrichment using our approach to previously published studies for eBMD[15] and T2D[10]. SNVs achieving a $\log_{10}$(BF) > 2 upon fine-mapping demonstrated increased enrichment relative to published credible sets for our positive control gene sets (**Supplementary Fig. 3**). We next compared the enrichment of SNVs for proximity to positive control genes (within 25 Kbp) when SNVs were limited to only those which were genome-wide significant ($P < 5 \times 10^{-8}$), compared to fine-mapped SNVs. In every case, the fine-mapped SNVs showed higher enrichment for positive control genes (**Fig. 1b, Supplementary Fig. 4**) compared to genome-wide significant SNVs. Taken together, these findings suggest that our fine-mapping approach decreases the total number of SNVs to be mapped to genes and helps to identify positive control genes.

We next analyzed the genomic relationship between fine-mapped SNVs and positive control genes. An average of 6.6% of GWAS loci had a positive control gene within 1 Mbp of an SNV with a $\log_{10}$(BF) > 2. Assuming that each GWAS locus reflects a causal biological signal, these findings indicate that most of the causal genes at GWAS loci are not currently known to cause Mendelian forms of the disease, or act as drug targets, thereby providing the opportunity to identify novel causal genes.

Assessment of Gene Expression

Tissue-specific expression has also been used as a method to identify putatively causal genes at GWAS loci,[18,19] and we tested this in a sensitivity analysis as an alternative source for positive control genes. To do so we developed tissue-selective gene sets based on expression using RNA-seq data for a variety of tissues from the GTEx project and purified hematopoietic cells from the ENCODE project (**Supplementary Fig. 5a, Online Methods**). We similarly observed that fine-mapped SNVs enriched for tissue-specific expression gene sets (**Supplementary Fig. 5b**), when compared to SNVs surpassing a P-value threshold $< 5 \times 10^{-8}$. However,

fold-enrichment was substantially higher for the set of positive control genes derived from Mendelian forms of disease and drug targets, than gene sets identified through tissue-selective gene expression (**Supplementary Fig 4**). Specifically, enrichment of $\log_{10}$(BF) > 2 SNVs for the positive control genes identified using Mendelian disease and drug targets was, on average, 6.8-fold higher than enrichment for the expression-derived genes (**Supplementary Fig. 4**), but the tissue-specific expression sets contained 5.4-142 times more genes. Given the clearly stronger enrichment for positive control genes, we did not further consider genes identified through tissue-selective expression as positive control genes.

Fine-mapping enriches for cell-type selective DNase I hypersensitive sites

Trait-associated variants localize to genomic regulatory regions of relevant cell and tissue types,[20] suggesting that this enrichment can be used as an orthogonal method to validate assumptions inherent to fine-mapping. We analyzed potential local regulatory effects of non-coding disease and trait-associated SNVs by comparison with DNase-seq data from a broad set of cell and tissue types generated the ENCODE and Roadmap Epigenomics projects.[20,21] We also generated DNase-seq data for Saos-2 and U2OS osteosarcoma cell lines and downloaded published accessible sites for pancreatic islets.[22] All data were analyzed using a uniform mapping and peak-calling algorithm (**Online Methods**).

We then calculated the enrichment for disease-and trait-associated SNVs in DHSs for each cell or tissue type (**Supplementary Fig. 6**) for progressively increasing $\log_{10}$(BF) thresholds. We identified strong enrichment at higher $\log_{10}$(BF) cutoffs for DHSs from cell types relevant to the trait. We then selected cell and tissue types which showed strong enrichment for each trait (**Table S5**). By comparing enrichment in DHSs for SNVs with P-value < 1×10^{-8} or $\log_{10}$(BF) > 2, enrichment was an average of 2-fold higher after fine-mapping (**Fig. 1c)**. These results again provide additional orthogonal evidence for the value of statistical fine-mapping in order to map genetic associations to positive control genes.

Genomic landscape annotations that enrich for positive control genes at GWAS loci

Fine-mapped SNVs were mapped to genes based on several different types of genomic annotations: (i) when the SNV alters amino acid sequence or transcript structure, (ii) overlap of the SNV with DHSs in relevant cell or tissue types, (iii) overlap of the SNV with an eQTL in a relevant tissue, and (iv) promoter-capture Hi-C in a relevant tissue. Most positive control genes in the proximity of fine-mapped SNVs were not targeted directly by a protein or transcript-altering SNV (**Fig. 2a**). But while a sizable proportion of positive control genes were only found by looking at longer range or with chromatin interaction data (**Fig. 2a**), enrichment decreased with distance to TSS (**Fig. 2b**). High-resolution examination of enrichment showed a strong inflection point at close range to the TSS (<50 Kbp); however, enrichment remained substantial even at distances >250 Kbp (**Fig. 2c**, **Supplementary Fig. 7**). These characteristics became even clearer upon fine-mapping and subsequent restriction to SNVs in cell-type specific DHSs. Enrichment for non-coding SNVs in DHSs was highest at medium range (<25 Kbp away) (**Fig. 2c**). However, while enrichment for more distant SNVs was lower, 69% of positive control genes at GWAS loci were > 25 Kbp from the nearest $\log_{10}$(BF) > 2 SNV (**Fig. 2a**). eQTL and promoter capture Hi-C data showed enrichment even after accounting for distance to TSS (**Supplemen-**

tary Fig. 8), but the overall magnitude of enrichment was considerably lower than when using simpler distance to gene metrics.

Taken together, these findings suggest that the genomic annotations that strongly enrich when mapping SNVs to positive control genes at GWAS loci include protein or transcript altering fine-mapped SNVs, distance and DHSs. These annotations were therefore used to develop the features in for the predictive model.

Generating the *Ei* predictive model

Given the observed enrichment of certain genomic annotations with positive control genes, we next sought to develop a predictive model using the enriched features (described in the **Online Methods** and summarized in **Fig. 1a**). Briefly, after defining GWAS loci and obtaining a set of fine-mapped SNVs, each SNV was first assessed for a set of annotations, including functional protein-coding or non-coding impact. All annotations are shown in **Table S6**. Then, these annotations were used to map SNVs to genes at each locus (e.g. distance from SNV to each TSS for each gene at the locus; see **Online Methods**). Finally, gene-level summary features were developed for all SNVs paired to a given gene to capture both the overall "intensity" of an annotation (e.g. the minimum, mean or maximum) as well as how these intensities varied with distance to gene, measured directly or inversely. This process resulted in a set of primary features used in the prediction model (**Table S7**).

We next took several precautions to ensure the validity and utility of the predictive model. First, since each locus contains features that are shared across all genes at the locus, such as the number of genes at the locus, we incorporated such locus-level features into the model. For instance, if there is only one gene at a GWAS locus, the probability of that gene being causal is higher than if there are a dozen genes at the same GWAS locus. To control for factors associated with locus-wide probability of causality, we included locus-level metrics as features and tested whether the model outperformed these predictors. Second, some diseases and traits studied share GWAS loci and/or positive control genes, such as glucose level and T2D. To avoid cross-contamination of the training and validation sets,[23] the genes were randomly retained for only one trait. Therefore, the same positive control genes were not used for correlated traits, such as glucose level and T2D, thereby preventing over-fitting of the model. Similarly, positive control genes were randomly selected to be included in analysis of only one trait.

We then trained models using two classifier methods, including logistic regression and XGBoost. The analysis was restricted to loci with at least one positive control gene. The training was conducted using a leave-one-out approach (i.e. leaving out one disease or trait) and the final predictions were then aggregated across all 12 diseases and traits. For comparison with established eQTL-based approaches, we ran eCAVIAR[24] using relevant tissues for each trait from GTEx, and assessed whether the top posterior probability for a gene across tissues predicted positive control gene status (**Supplementary Table 5, Online Methods**). We recognize that eCAVIAR is meant to identify eQTLs, rather than positive control genes as we have defined them. Nevertheless, this comparison allows for a calibration of performance relative to this known metric.

Aggregating predictions across all 12 traits revealed that XGBoost performed substantially better than logistic regression, achieving an area under the curve (AUC) for the receiver operator curve (AUC-ROC) of 0.80 versus 0.61, and an AUC for the precision recall curve (AUC-PRC) of 0.29 versus 0.12 or 0.01 (**Table 2**). Performance was also higher than eCAVIAR, which achieved an AUC-ROC of 0.70 and AUC-PRC of 0.01 (**Table 2**). These results outperformed simpler approaches, such as selecting the gene nearest the most strongly associated SNV (**Fig. 3**). As a result, we refer to the algorithm for generating features and the prediction model from XGBoost as the Effector index (*Ei*) (**Fig. 3**). The PRC of the final model suggests that in order to have an average positive predictive value (PPV) of ~60% across all traits, the *Ei* would generate an average sensitivity to identify positive control genes of ~20% (**Fig. 3**). These findings suggest that the *Ei* has reasonable prediction accuracy to identify causal genes for common diseases and traits. Disease and trait-specific AUC-ROC and AUC-PRC from the leave one out analysis are shown in **Table S8**.

Ei Validation

We next measured the predictive performance of the *Ei* against an augmented set of positive control genes for T2D, which were selected using a complementary approach to select positive control genes. Independently, Mahajan and McCarthy generated a list of 35 positive control genes for T2D[9] (**Table S9**) that included T2D drug targets, genes causing Mendelian forms of T2D, and coding evidence from large-scale exome arrays studies and strong evidence from gene-based burden tests from large-scale whole exome sequencing studies (**Methods Section** and **Table S9**)[25]. The performance of the *Ei* was evaluated in this validation set by first training the *Ei* on all other 11 diseases and traits and testing its performance for T2D using this augmented set of 35 genes as the positive control genes, 24 of which were present within a GWAS locus for T2D as defined in this study.

Using this enlarged and independent list of positive control genes, the AUC-ROC for T2D increased from 0.81 to 0.85 (95% CI: 0.76-0.94) and the AUC-PRC improved substantially from 0.37 to 0.62 (95% CI: 0.55-0.68). These results indicate that in order to have a positive predictive value of ~80%, the *Ei* would provide a sensitivity of ~40% to identify positive control genes for T2D (**Fig. 3c**). The *Ei* probabilities of causality between these two sets of positive control genes were similar (**Table S10**). Specifically, 61% of the original positive control genes had an *Ei* probability of >0.8, whereas 63% of the Mahajan and McCarthy positive control genes had an Ei probability of >0.8. This suggests that the *Ei* is able to assign high probabilities to causal genes that are not necessarily known drug targets or causes of Mendelian forms of disease. These findings also demonstrate that as the number of positive control genes identified through large-scale whole-exome sequencing and exome arrays increases, the utility of the *Ei* for mapping GWAS associations to positive control genes is likely to improve. We conclude that applying the *Ei* to data from currently available large-scale GWAS can help to anticipate the results from exome sequencing studies.

Features influencing the *Ei*'s performance

Given the validation of the *Ei* and its favorable prediction performance, we next asked which features received the largest importance in the model (**Table S7 and Fig. 4)**. Amongst the top 20 features, 4 were influenced by the physical distance of SNVs to genes. The second highest ranked feature was the rank of predicted gene impact from SNPEff (**Table S6** and **Online Methods**). Enrichment analysis of these ranked gene impacts, comparing genes containing SNVs with only lowest rank (MODIFIER) to genes with one more SNVs with higher ranks (LOW, MODERATE and HIGH), revealed that higher ranks are more predictive of positive control genes (OR 9.21, $p < 2 \times 10^{-16}$). As expected, this demonstrates that coding and transcript altering variation is predictive of causal genes, as these properties are within the higher impact ranks from SNPEff. Conversely, low impact rank is less predictive, as it includes SNVs in introns or within close proximity to the gene and also accounts for the majority of SNVs with SNPEff predictions (**Table S11**). Another predictive feature was fine-mapped SNVs within DHSs in disease-relevant tissue types that are nearest to a gene.

Notably, four locus-level features, such as number of genes at the locus, which results in the same value for all genes at a locus, were also of high importance to prediction (**Fig 4**). To investigate the ability of the *Ei* to identify positive control genes over and above the performance provided by only locus-level features, we trained models and generated predictions for only the 14 locus-level features (**Table S7**). We found that the *Ei* substantially out-performed locus level features (**Fig 4b & 4c**), in that models built with only locus-level features provided an AUC-ROC of only 0.73 (95% CI: 0.69-0.77) and an AUC-PRC of only 0.14 (95% CI: 0.08 0.23). We conclude that the *Ei* provides enhanced capabilities to discriminate between genes within a locus.

We next sought to test whether the *Ei* could predict a subset of genes at a locus at elevated causal probability, rather than simply assigning the same probability to all genes at a locus. To do so, we first determined the predicted probability cutoff that lead to optimal enrichment of positive control genes and then used this cutoff to select putatively causal genes at each locus. We determined the optimal cutoff to be 0.46, a threshold above which we observe 99 of 159 (62%) positive control genes (**Fig. 4c**).

When applying this cutoff of 0.46 to all loci, we observed that the number of genes considered per locus reduces substantially. Specifically, we found that after applying this cut-off, 78% of loci harbored four or fewer genes. Without such a cutoff, only 28% of loci had four or fewer genes to consider for potential causality. This demonstrates that the *Ei* can enrich for known causal genes and is able to considerably reduce the number of genes to be considered for biological interrogation at a locus.

DISCUSSION

We have developed the *Ei* which provides helpful predictive accuracy for prioritizing causal genes across a range of traits and diseases. The most useful predictive features were simple metrics such as protein-coding or transcript alterations, fine-mapped SNVs in DHSs and distance from fine-mapped SNVs to genes—features which are often already available in public datasets. We further demonstrated that the *Ei*'s performance increased in a larger and independently curated list of causal genes for T2D, which included genes identified through whole

exome sequencing in large cohorts. Importantly, the *Ei* outperforms locus-specific features and simpler algorithms such as nearest gene to the lead GWAS SNV. Last, the *Ei* can substantially reduce the number of genes to be considered for biological validation at GWAS loci. Taken together, the *Ei* provides a publicly accessible algorithm which can be applied to GWAS datasets to enable functional biological dissection and drug development.

A major strength of this study was the use of positive control genes that were derived from Mendelian forms of disease and the targets of clinically useful medicines. While this provides a shorter list of positive control genes than would have resulted from using other methods to establish causality, such as murine or cellular models, it does yield a list of genes whose probability of causality is high. A second major strength is the generalizability of the method across different traits and diseases, influencing different organ systems. Despite the biological differences in the types of conditions studied, the AUC-ROC and AUC-PRC relatively stable.

To improve the generalizability of the *Ei,* we have chosen to focus on a select set of readily available genomic annotations, allowing for its application across many traits and diseases. The annotations we have used can be generated for every GWAS, provided that the imputation reference panel is available to enable accurate fine-mapping and DHS maps from a disease-relevant cell type are available. DHSs have been measured for many tissues and cells via the ENCODE and Roadmap Epigenomics programs, and extension to rarer cell and tissue populations and states is well underway.

We have not used any prior biological information to prioritize genes at loci. While this could be implemented, for example through the use of text-mining tools, we caution that one of the main advantages of the agnostic GWAS approach is the ability to identify genes whose function was not previously known. We note that eCAVIAR, which relies on colocalization of GWAS and eQTL data, did not identify positive control genes as readily as the *Ei*. However, this method is not intended to identify disease causal genes selected through Mendelian inheritance and drug targets, and therefore does not offer a direct comparison. Since our goal is to identify positive control genes, unbiased by prior biological data, we chose to first prioritize genes which can then be complemented by a thorough assessment of the literature.

The relative importance of different predictors in the final *Ei* model is informative. The most important predictor was the simple count of genes at a GWAS locus. This is sensible, since if there is only one gene at a locus, its probability of causality will be higher than if there are 20 genes at the locus. Other informative features were simple metrics such as protein-coding or transcript alterations, distance from fine-mapped SNVs to genes and overlap with DHSs. Previously we have shown that distance to gene is a strong predictor of causal genes in the field of metabolomics,[26] yet there are examples of causal genes at GWAS loci that lie hundreds of kilobases away from the lead SNV.[27,28] Further, we have also demonstrated that the nearest gene approach under-performs when compared to the *Ei*, which instead considers proximity in the context of other relevant factors. Given that the majority of positive control genes were associated through non-coding variation, the *Ei* will benefit from improved genomics approaches to infer

long-range variant-to-gene links, promising improved prioritization of causal genes for functional dissection and drug development.

This work has important limitations. We have only used large GWASs and emphasize that the accuracy of the *Ei* may not perform well in smaller studies. Further, when implementing fine-mapping, we have been careful to use the same, or very similar reference panels. We would not expect this step of the *Ei* algorithm to perform well if inappropriate reference panels were used. Since fine-mapping strongly enriches for positive control genes, care is warranted when using reference panels different than the original GWAS.

In summary, we have developed an implementable algorithm that can identify positive control genes at GWAS loci with reasonable accuracy and can therefore be used to prioritize genes for functional dissection and drug development programs.

Method Overview

Here we describe the overall study plan, which was divided into five steps. Further information is provided in the **Online Methods**. First, a set of positive control genes was identified by including genes that either encode a protein which is the target for a medication used in humans to treat the disease or trait, or the gene causes a Mendelian form of the disease (or influences the trait). Second, for each trait of interest, GWAS results were obtained and significant loci identified (defined as $P \leq 5 \times 10^{-8}$). The studies from which the summary statistics were taken are available in **Table 1**. When appropriate studies were not available for the trait of interest, we undertook a GWAS using data from UK Biobank. Third, at each locus, we undertook statistical fine-mapping[17] and mapped putative causal SNVs to genes, by annotating the SNVs for putative biological function, such altering the transcript or protein, associated with gene expression (eQTL), or intersection with non-coding elements (DHSs). Annotations were assessed for their ability to enrich for positive control genes. Each annotation was then summarized for all SNVs per gene using min, max, sum, and mean, followed by scaling by distance to gene for relevant annotations. These features were then used for model building and testing. Fourth, using only loci known to contain at least one positive control gene, we built predictive models using logistic regression and XGBoost. We tested the performance of the models on traits and genes not included in the model building. Specifically, measurement of model performance was assessed using a leave-one-out approach, wherein the model was built iteratively using eleven traits and tested on the twelfth. To further assess model performance, we then generated a model training on all traits and diseases, except T2D. We then tested this model on an independently curated set of positive control genes for T2D, which included drug targets, OMIM genes and genes identified by large-scale exome sequencing and exon arrays. The *Ei* was further tested against simpler approaches including the gene nearest the lead SNV.

FUNDING AND ACKNOWLEDGEMENTS

The funding agencies had no role in the design, implementation or interpretation of this study. The views expressed in this article are those of the author(s) and not necessarily those of funders. The Richards research group is supported by the Canadian Institutes of Health Research (CIHR), the Lady Davis Institute of the Jewish General Hospital, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK and the Fonds de Recherche Québec Santé (FRQS). JBR is supported by a FRQS Clinical Research Scholarship. JBR has served as an advisor to GlaxoSmithKline. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. MMcC is funded by the NHS, the NIHR, and the Department of Health. MMcC has served on advisory panels for Pfizer, NovoNordisk and Zoe Global, has received honoraria from Merck, Pfizer, Novo Nordisk and Eli Lilly, and research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier, and Takeda. As of June 2019, MMcC is an employee of Genentech, and a holder of Roche stock. MMcC has received funding from the NIH: U01-DK105535 and the Wellcome Trust: Wellcome: 090532, 098381, 106130, 203141, 212259. The Greenwood lab acknowledges support from Compute

Canada (RAPI: nzt-671-aa). EF is an employee of Pfizer. This research has been conducted using the UK Biobank Resource using project number 27449.

Data availability statement

All accession codes and URLs for publicly available data are provided in the Online Methods. The data that support the findings of this study are available from GitHub, https://github.com/richardslab/Ei. This includes raw data underlying the figures.

Code availability statement

The code that support the findings of this study are available from GitHub, https://github.com/richardslab/Ei. This includes code underlying the figures.

REFERENCES

1. Arrowsmith, J. Trial watch: Phase II failures: 2008-2010. *Nat. Rev. Drug Discov.* **10**, 328–329 (2011).
2. Arrowsmith, J. & Miller, P. *Trial Watch: Phase II and Phase III attrition rates 2011-2012*. *Nature Reviews Drug Discovery* **12**, 569 (2013).
3. Arrowsmith, J. Trial watch: Phase III and submission failures: 2007-2010. *Nat. Rev. Drug Discov.* **10**, 87 (2011).
4. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
5. Cook, D. *et al.* Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–31 (2014).
6. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
7. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLOS Genet.* **15**, e1008489 (2019).
8. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-Augu**, 785–794 (ACM Press, 2016).
9. Mahajan, A. & McCarthy, M. I. Predicted Type 2 Diabetes Effector Genes. (2019). Available at: https://s3.amazonaws.com/broad-portal-resources/effector_predictions_documentation.pdf.
10. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0241-6
11. Disease Ontology. Available at: http://disease-ontology.org/about/. (Accessed: 24th January 2020)
12. Online Mendelian Inheritance in Man®. (2020). Available at: omim.org. (Accessed: 3rd February 2020)
13. UpToDate. Available at: https://www.uptodate.com/home/about-us.
14. DrugBank.ca.
15. Morris, J. A. *et al.* An atlas of genetic influences on osteoporosis in humans and mice. *Nat. Genet.* 338863 (2018). doi:10.1038/s41588-018-0302-x
16. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
17. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
18. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
19. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9**, 4361 (2018).
20. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).
21. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
22. Greenwald, W. W. *et al.* Pancreatic islet chromatin accessibility and conformation reveals distal enhancer networks of type 2 diabetes risk. *Nat. Commun.* **10**, 2078 (2019).
23. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer-promoter

interaction-prediction methods. *Nat. Genet.* **51**, 1196–1198 (2019).
24. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
25. O'Seaghdha, C. M. *et al.* Meta-Analysis of Genome-Wide Association Studies Identifies Six New Loci for Serum Calcium Concentrations. *PLoS Genet.* **9**, e1003796 (2013).
26. Stacey, D. *et al.* ProGeM : a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res.* 1–13 (2018). doi:10.1093/nar/gky837
27. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
28. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).

Figures

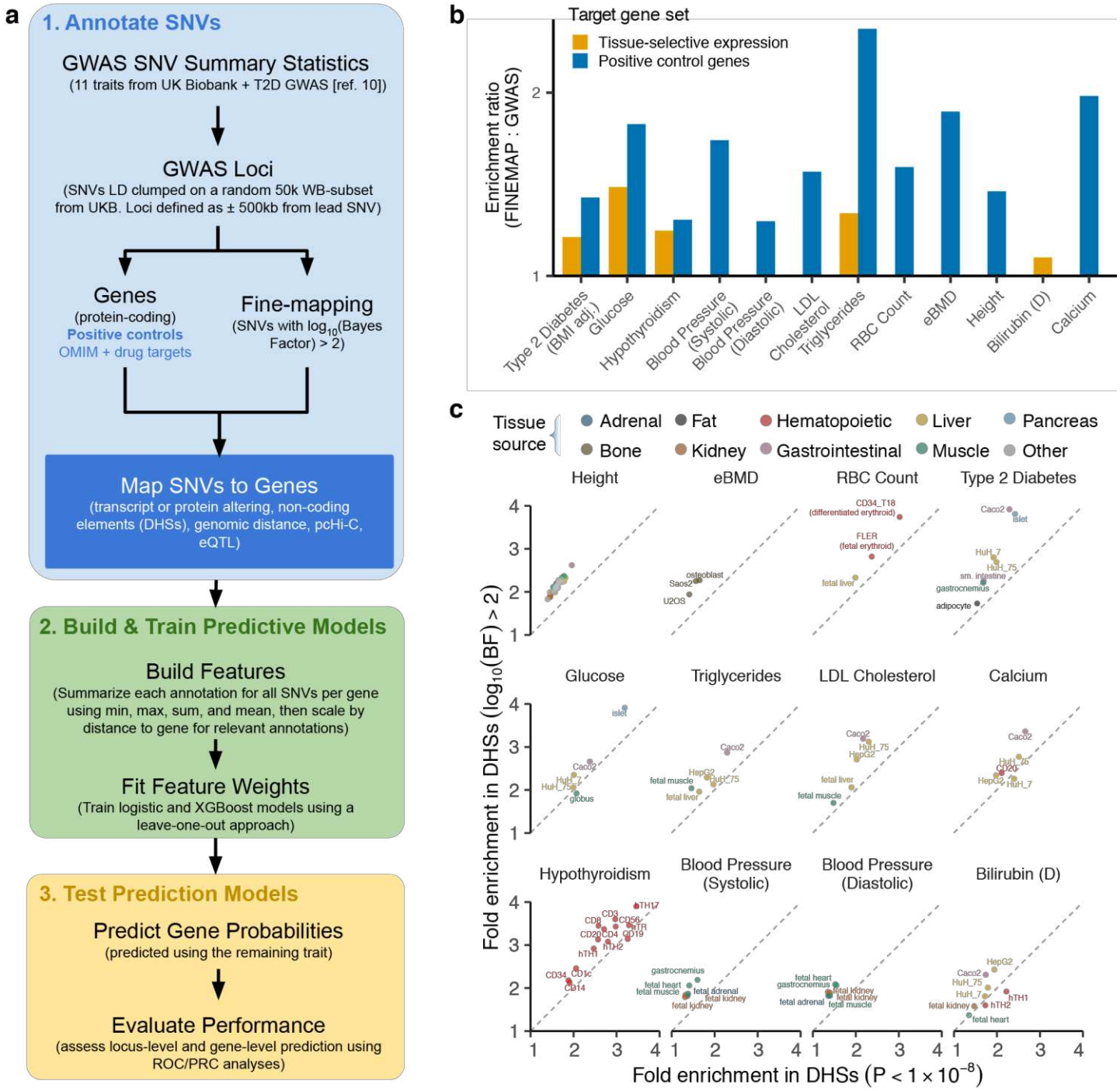


Fig. 1. Building the Effector Index and enrichment for likely target genes by statistical fine-mapping. (**a**) Flow diagram depicting: (1) how data was generated using fine-mapping of GWAS summary statistics, followed by SNV annotation and pairing to genes at each GWAS locus (2) how this data is used to generate gene- and locus-level features followed by fitting their feature weights within the models using a leave-one-out analysis, and (3) assessing the performance of the models to predict target genes for loci containing positive control genes. (**b**) Ratio of enrichment for positive control genes within ±25 kbp of genome-wide significant SNVs ($P < 5 \times 10^{-8}$) compared to SNVs having $\log_{10}(BF) > 2$ after fine-mapping. Fold enrichment was calculated as the proportion of positive control genes targeted to the proportion of all genes targeted. (**c**) Comparison of enrichments for genome-wide significant SNVs (x-axis) vs. SNVs with $\log_{10}(BF) > 2$ SNVs (y-axis) within trait-specific DHS sites (see **Online Methods**).

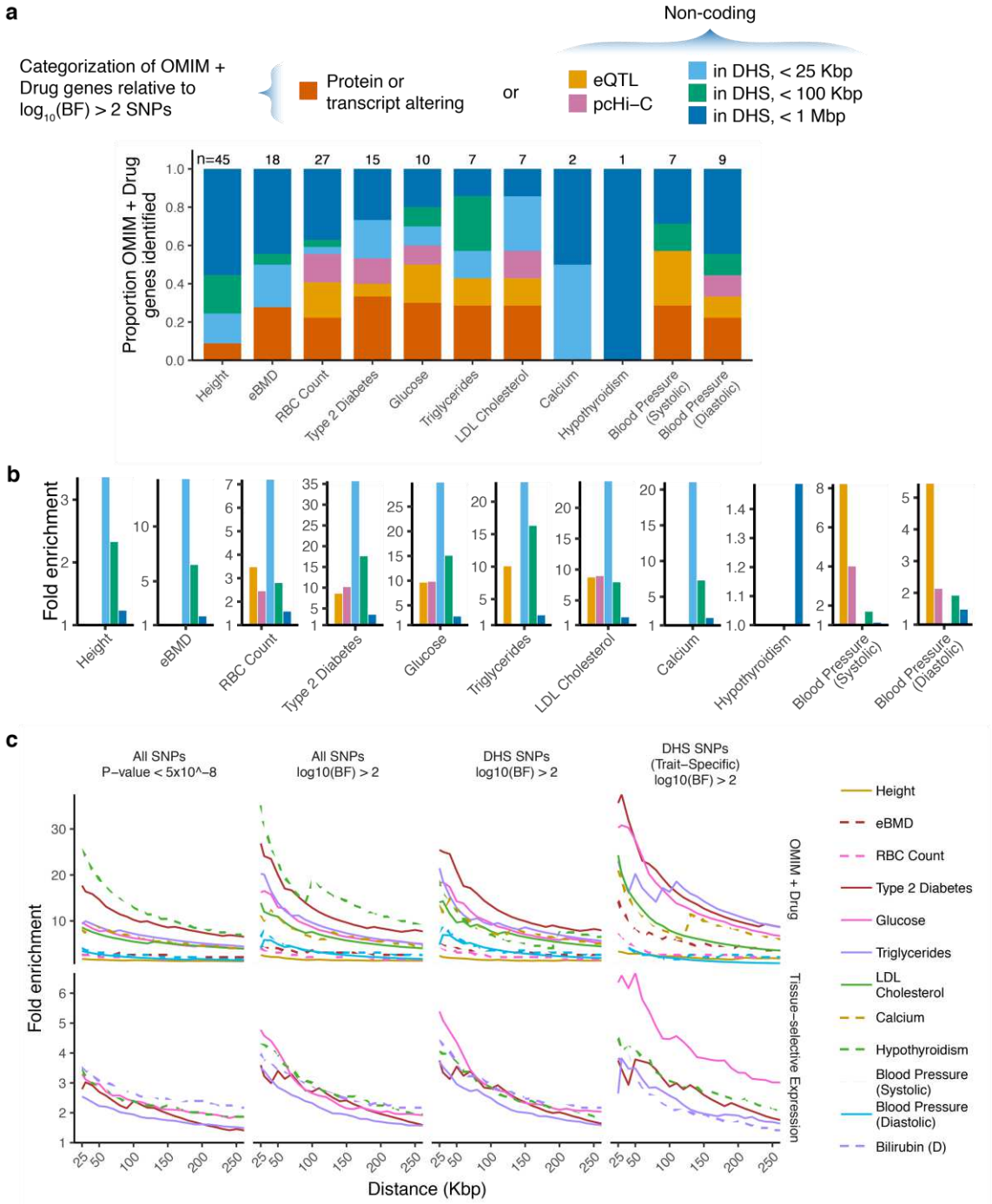


Fig. 2. Enrichment of genomic landscape features with positive control genes

Genes with protein or transcript altering SNVs were assessed separately. Non-coding SNVs were classified by overlap with trait-specific DHSs, distance to the TSS, and eQTL or pcHi-C evidence. (**a**) Summary of positive control genes at GWAS loci by relation to $\log_{10}$(BF) > 2 SNVs. Bar charts demonstrate the proportion of positive control genes identified by intersection of fine-mapped SNVs with different genomic landscape features. Results are separated by trait/disease. Genes were attributed to a single genomic landscape category in the order listed in figure legend above the plot. (**b**) Enrichments for each category of non-coding SNVs for positive control genes segregated by trait/disease. Enrichment for protein or transcript altering variants was excluded for legibility. (**c**) Enrichment of positive control genes by distance to non-coding SNVs (x-axis) for all traits. Fold enrichment was calculated as the ratio between the proportion of positive control genes targeted to the proportion of all genes targeted. The SNVs with $\log_{10}$(BF) > 2 were further overlapped with a master list of DHSs in any cell or tissue type (**Table S5**).

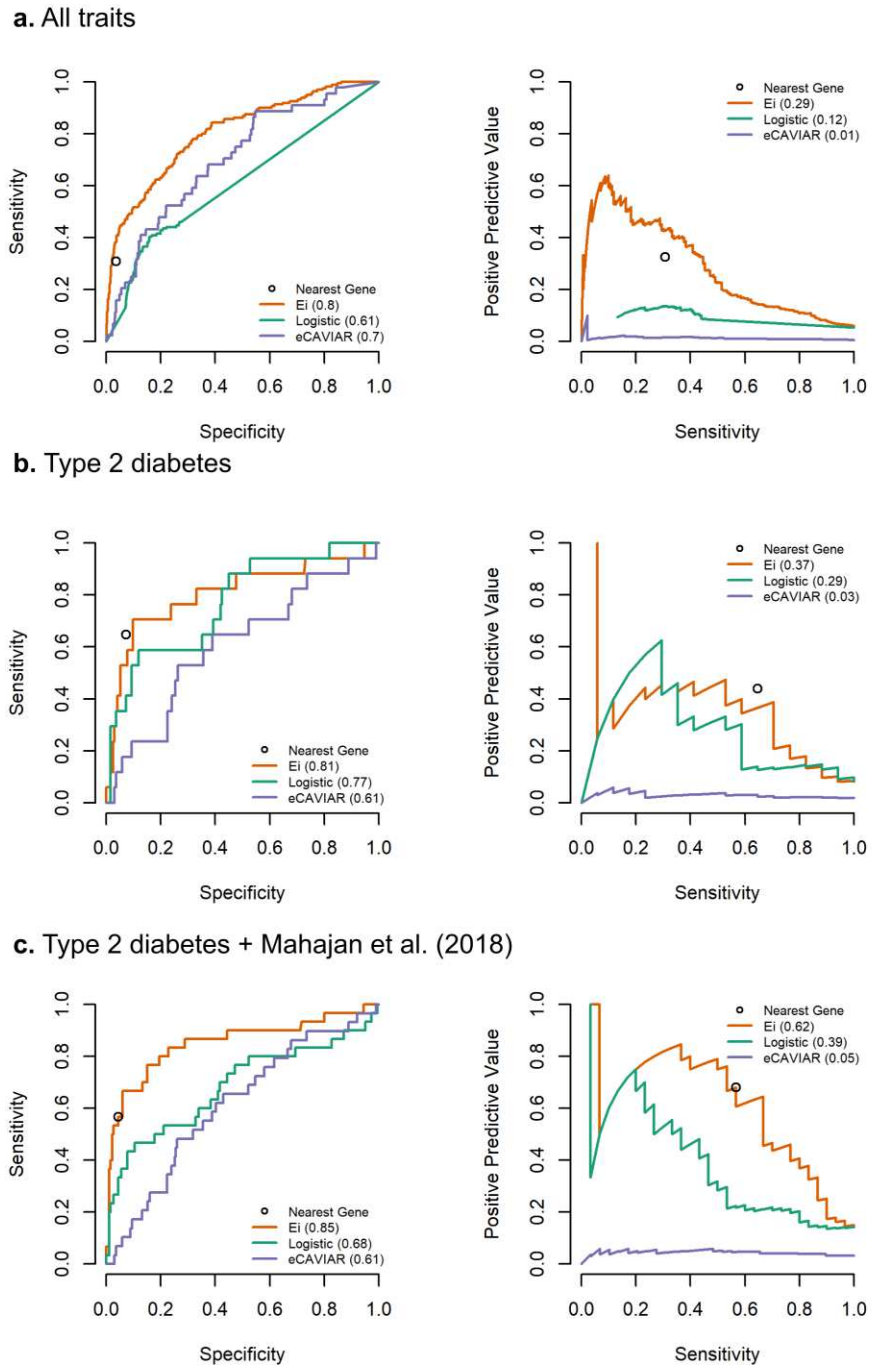


Fig. 3. Performance of the Effector index at loci containing positive control genes.
The performance of the Effector index compared to logistic regression and eCAVIAR for predicting positive control genes for: (**a**) all 12 traits, (**b**)**.** type 2 diabetes only, and (**c**) type 2 diabetes with the addition of manually curated causal genes from large-scale exon array and exome sequencing studies.[9] Area under the curves are provided in parentheses and are segregated by trait/disease **Table 2**. Performance using the nearest gene to the lead SNV by P-value is also shown by open circles.

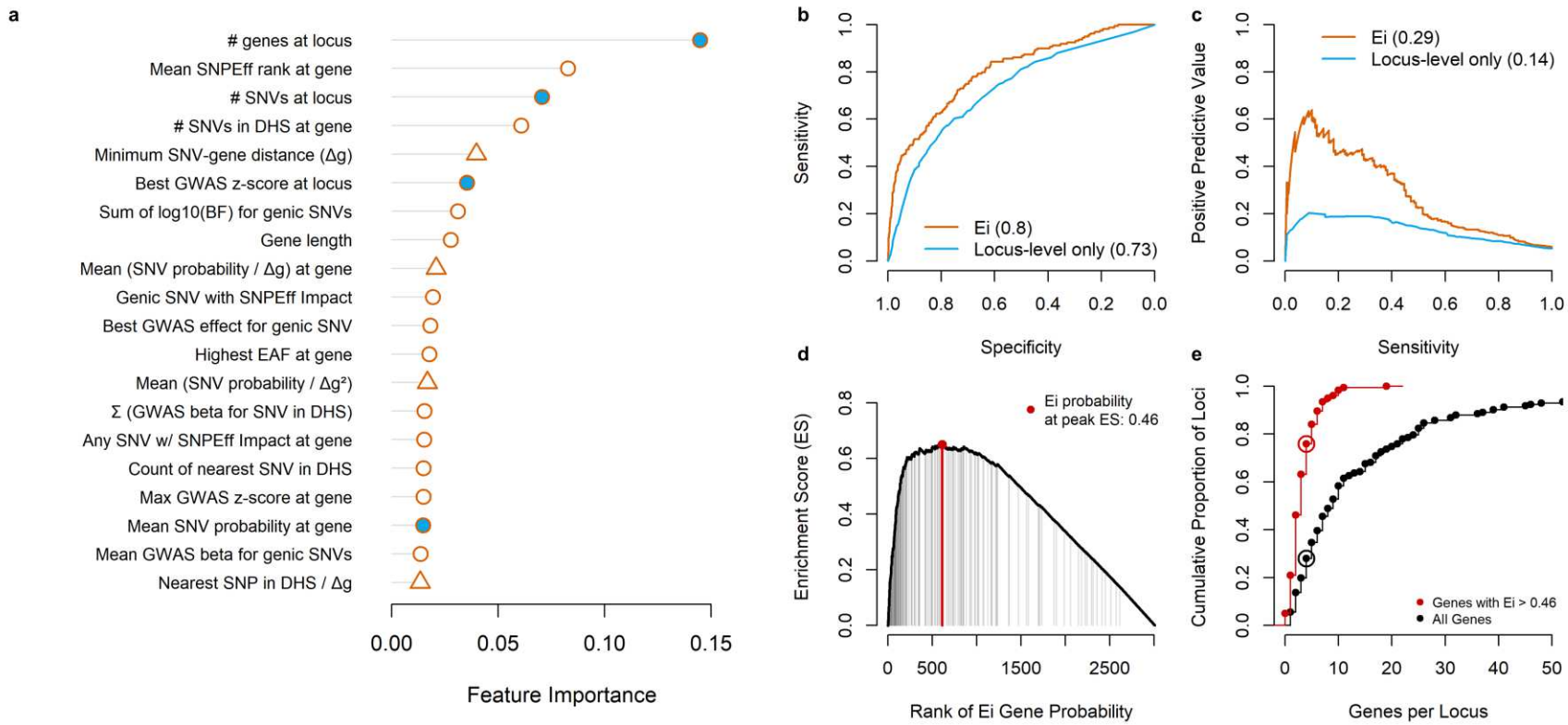
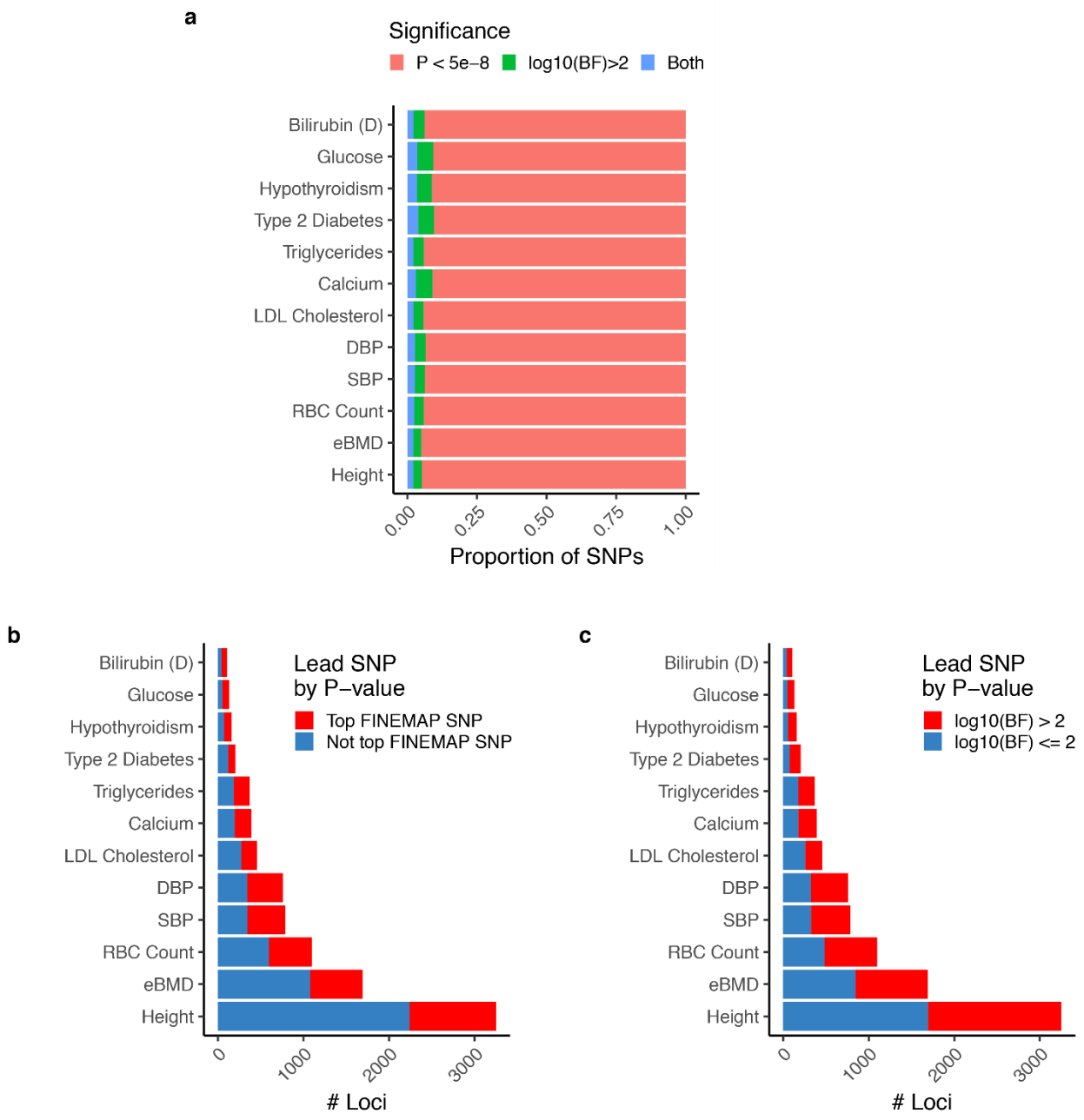
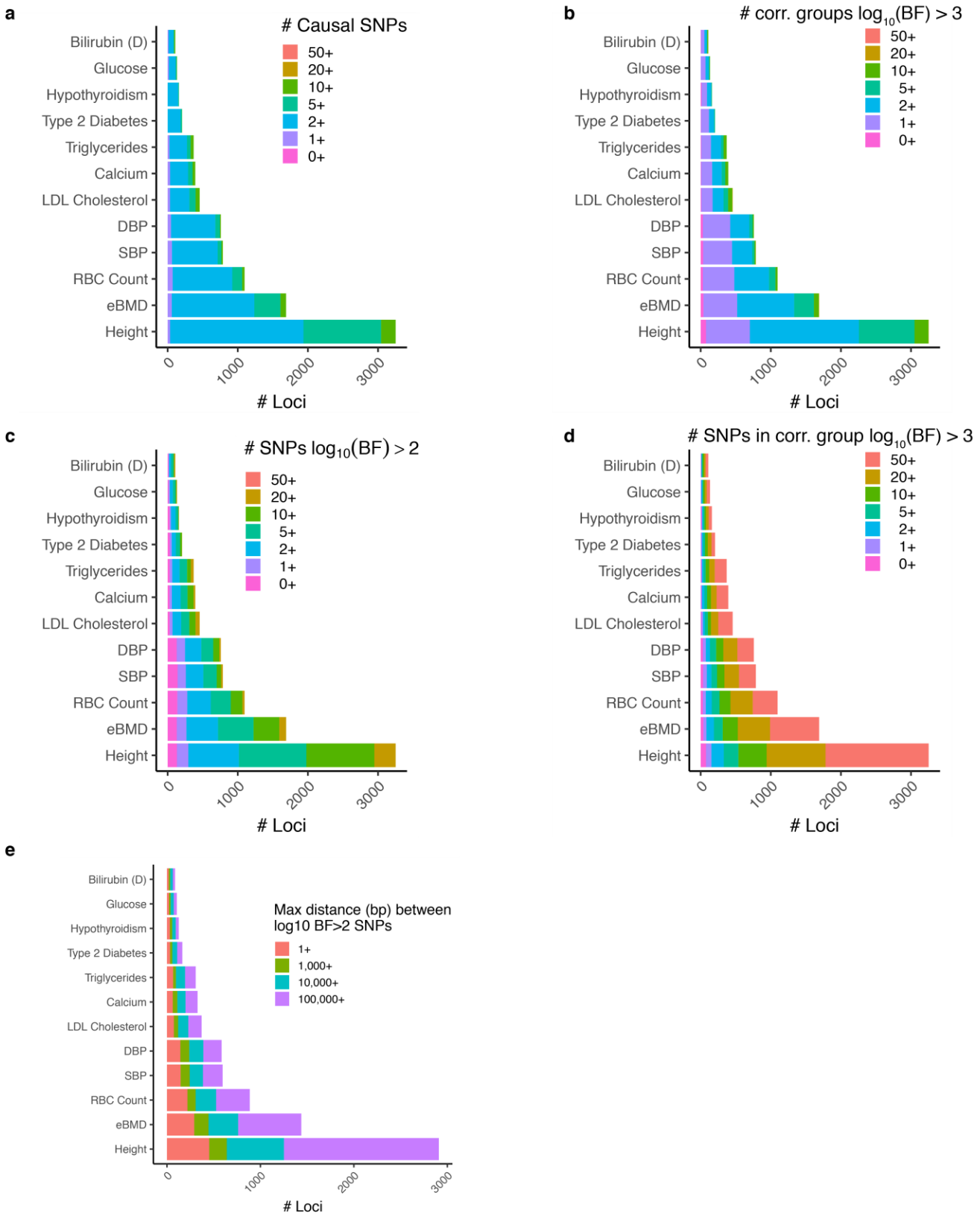


Fig 4. Features selected by the Effector index and comparison of the Effector index to use of only locus-level features. (**a**) Top 20 features selected by the Effector index where each importance value provided is the absolute mean importance of that feature across the 12 traits. Locus-level features shown (in blue) are those do not vary across genes at a locus. Features that incorporate distance to gene are displayed using triangles (Δg denotes SNV-gene distance; 'genic' denotes that SNV overlaps gene body). (**b-c**) The ROC (**b**) and PRC (**c**) curves for only locus-level features versus the use of all features in the *Ei* model. Areas under the curve are provided in parentheses. (**d**) Leading edge analysis shows the peak enrichment score for positive control genes occurs an *Ei* probability of 0.46 (red point); a threshold above which we observe 99 of 159 positive control genes (vertical grey lines). (**e**) Using the peak *Ei* threshold of 0.46 considerably reduces the number of genes per locus. For instance, 78% of loci contain 4 or fewer genes with the *Ei* > 0.46 (red open circle), whereas only 28% of loci contain 4 or fewer genes when no threshold is applied (black line open circle).

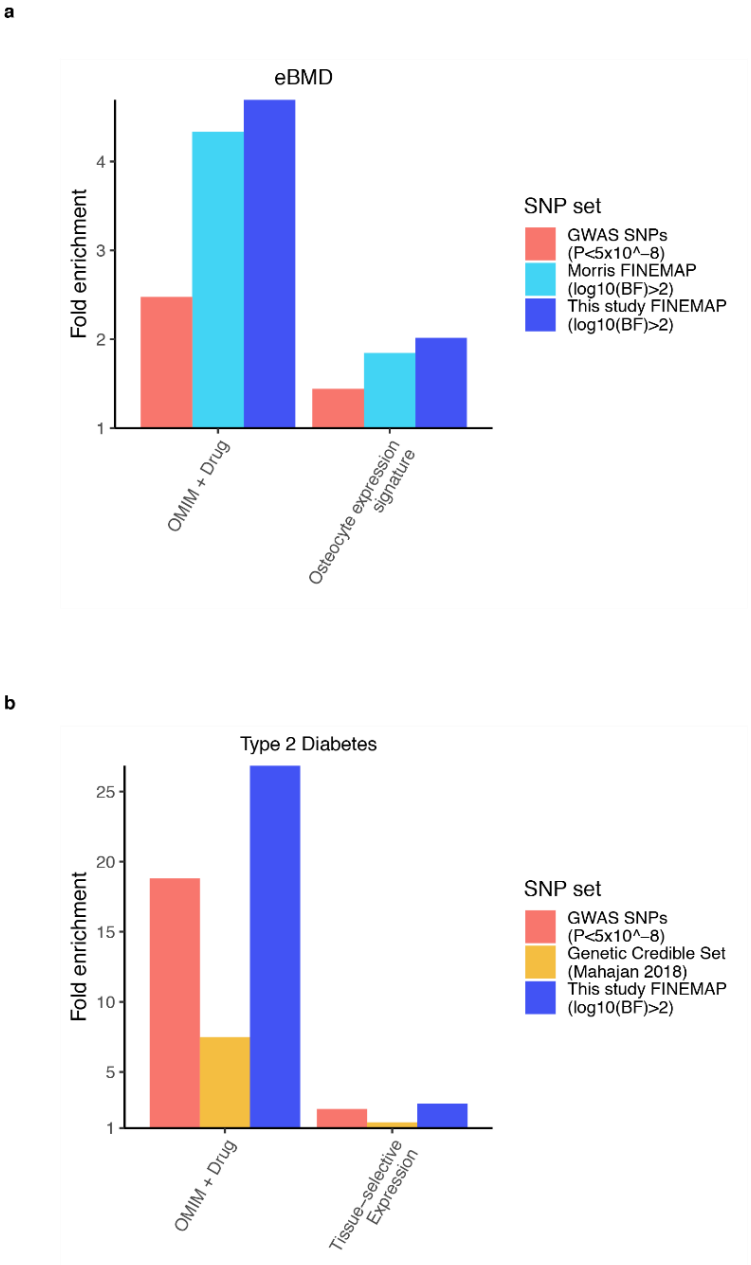
Supplemental Figures



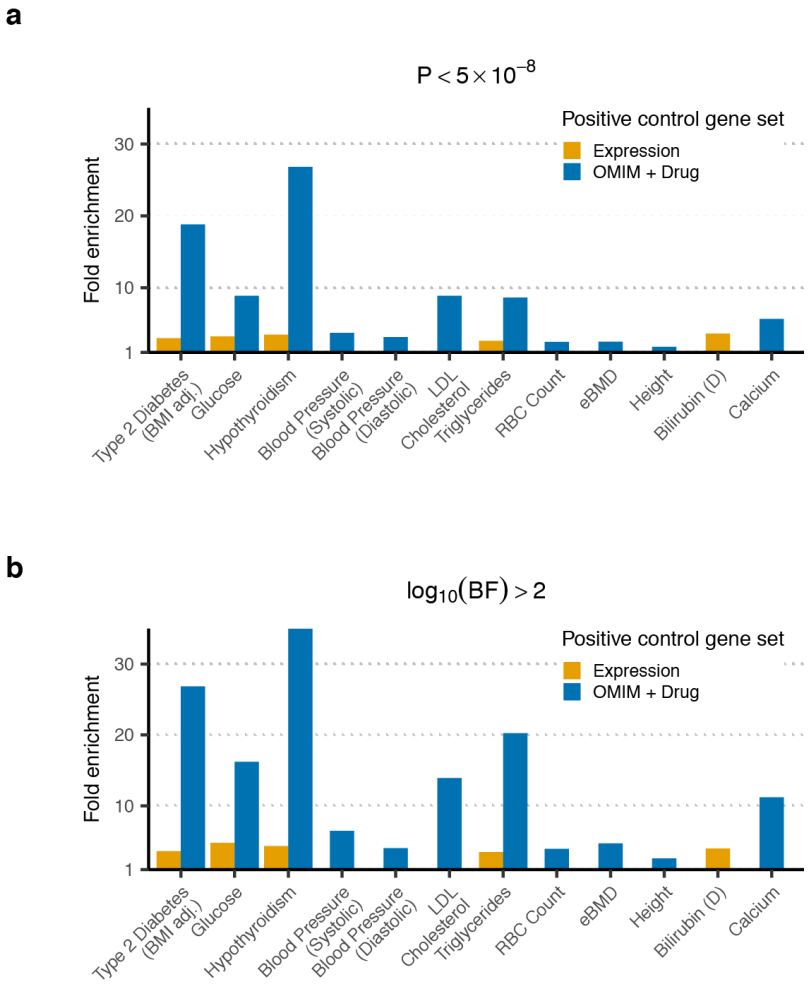
Supplementary Fig. 1. Summary of statistical fine mapping results.
(**a**) Composition and overlap of genome-wide significantly associated sets and FINEMAP credible set. (**b-c**) Frequency of loci where lead GWAS SNV by P-value is also (**b**) lead SNV by $\log_{10}$(BF) or (**c**) has $\log_{10}$(BF) > 2.



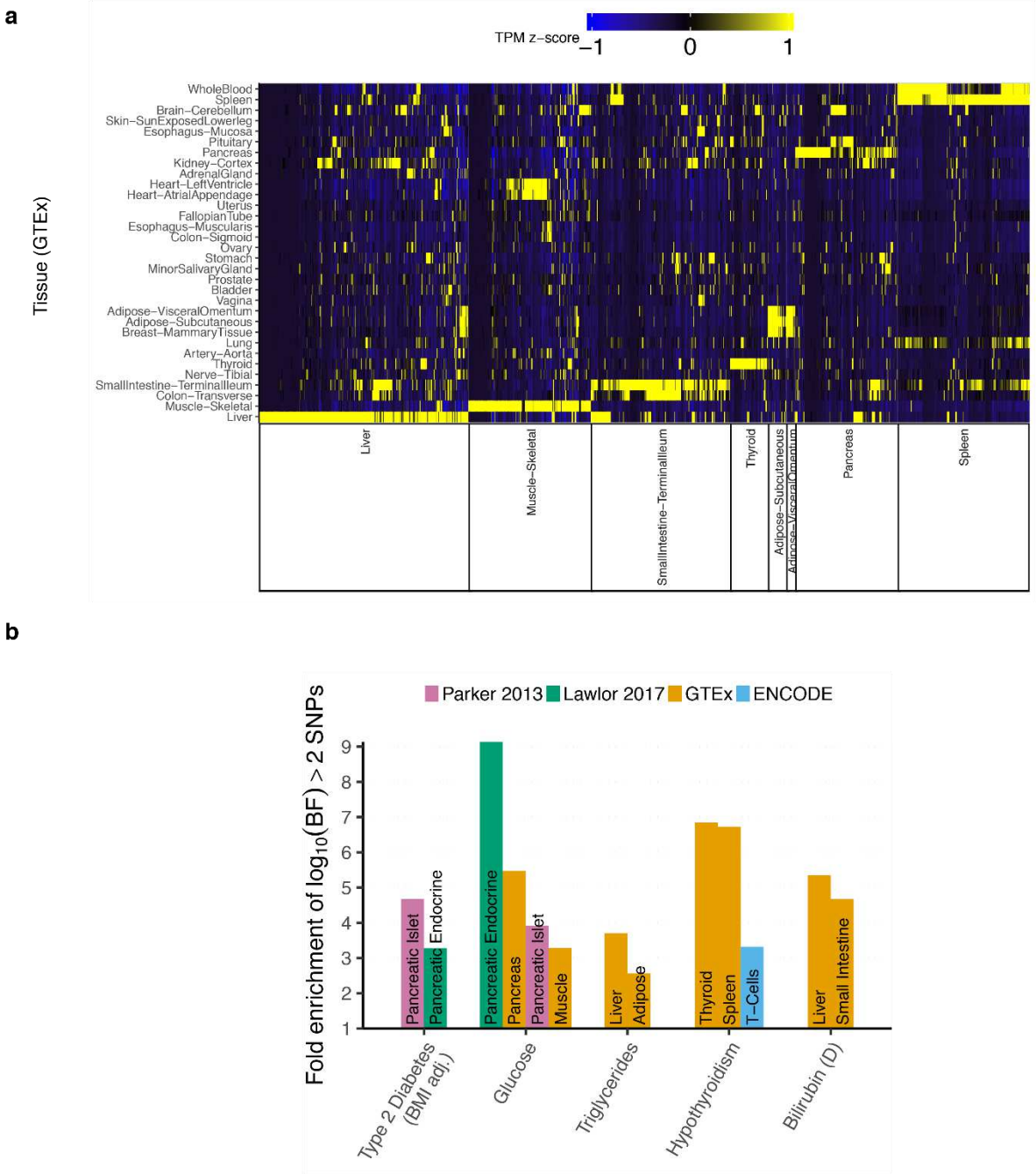
Supplementary Fig. 2. Assessment of fine-mapping GWAS loci.
Overview of allelic heterogeneity at fine-mapped loci (**a-d**). Accounting for the number of causal signals per trait, as measured by (**a**) FINEMAP (k_max); (**b**) number of statistically significant groups of independent SNVs $\log_{10}(BF) > 3$; (**c**) the number of individual SNVs $\log_{10}(BF) > 2$; and, (**d**) the number of SNVs in statistically significant groups of independent SNVs $\log_{10}(BF) > 3$. (**e**) extent of genomic region covered by $\log_{10}(BF) > 3$ at each locus.



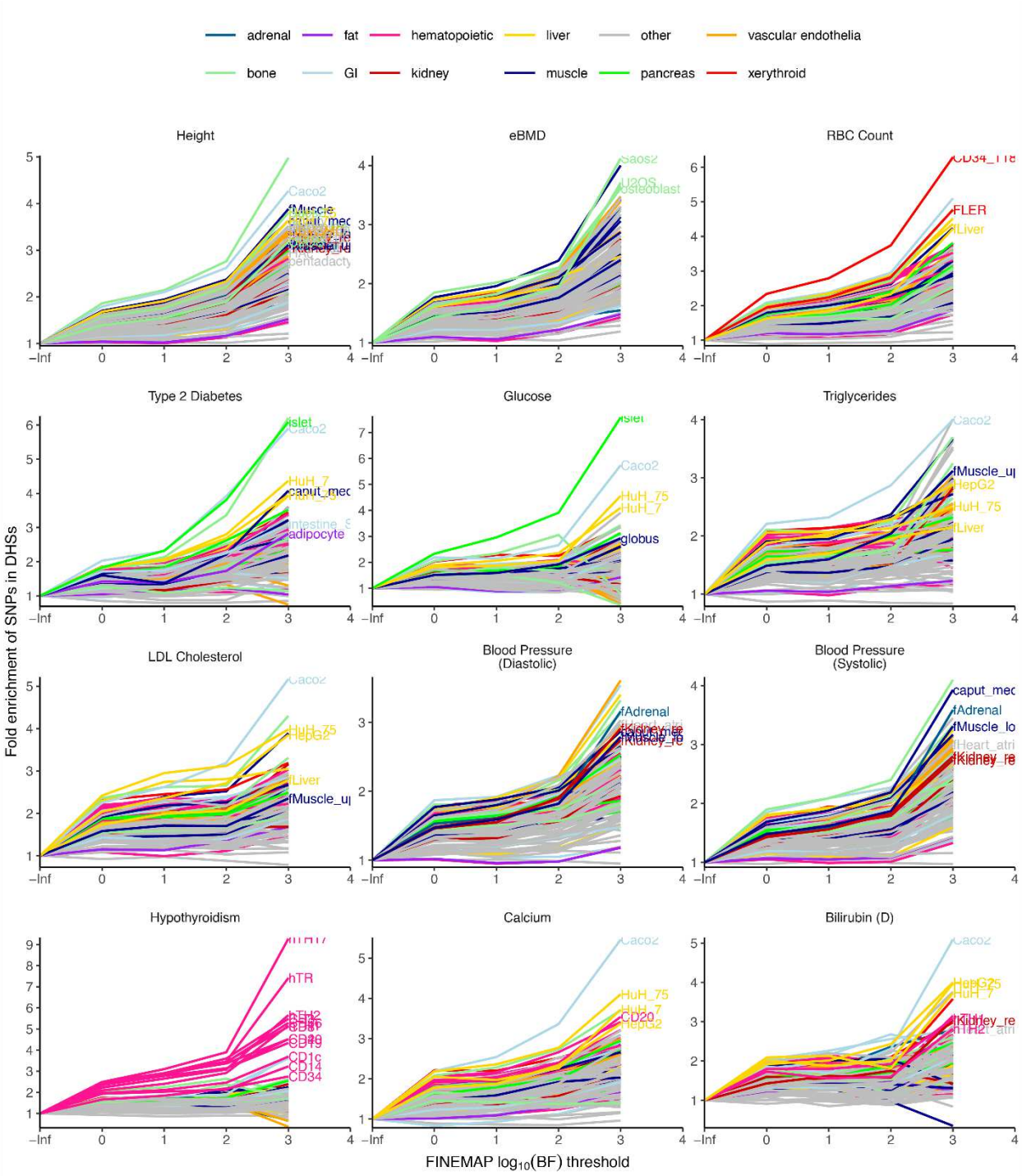
Supplementary Fig. 3. Comparison of fine-mapping credible sets to past studies.
(**a-b**) Enrichment of positive control gene sets. Shown is enrichment of SNV sets for positive control genes within 25 Kbp. (**a**) Comparison to eBMD analysis from Morris et al. 2019[15] in terms of enrichment for OMIM + Drug genes and osteocyte expression genes. The Morris et al. GWAS was used as input to the FINEMAP pipeline in the present study. (**b**) Comparison to published Type 2 Diabetes fine-mapping study[10] in terms of enrichment for OMIM + Drug genes and Expression positive control sets.



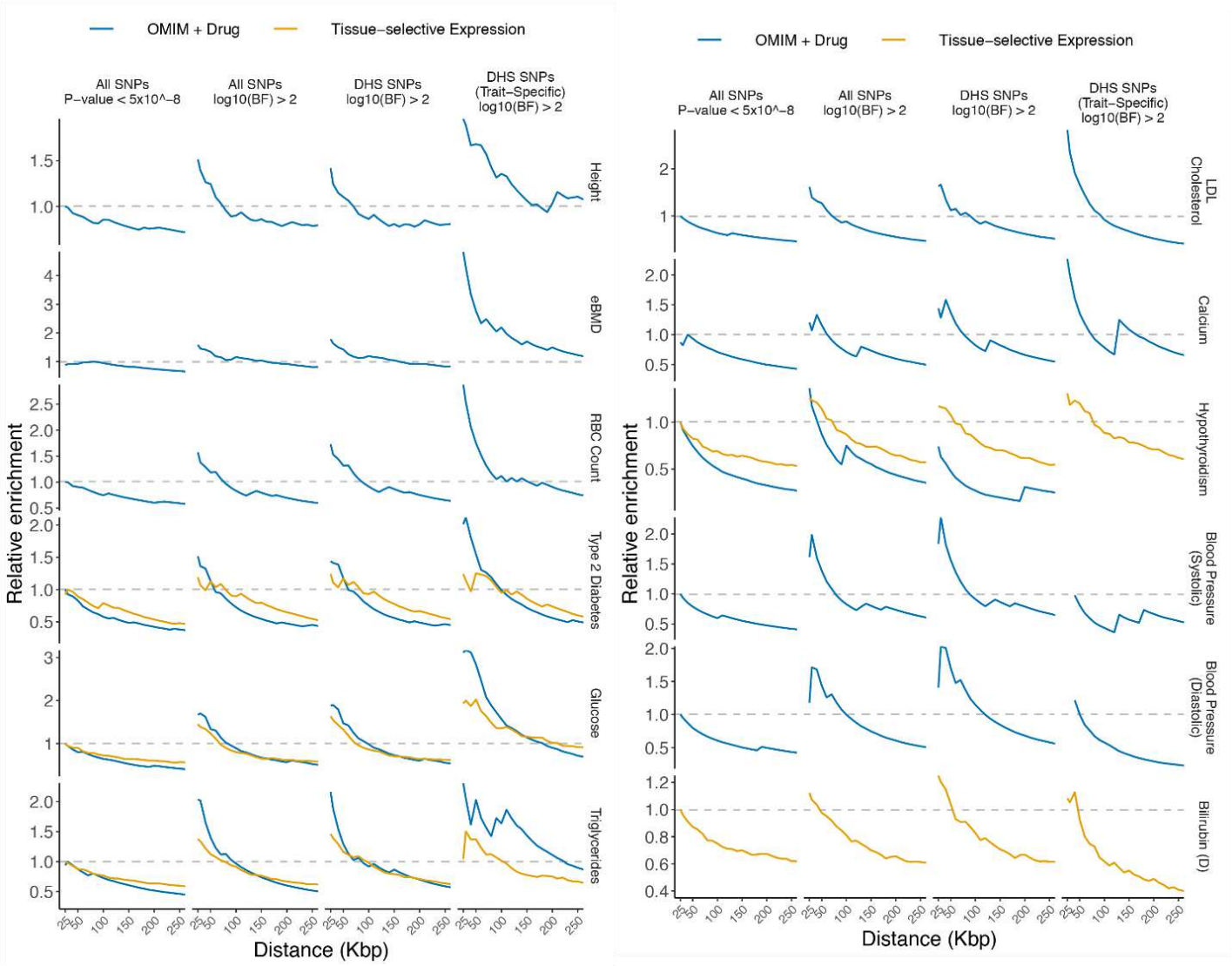
Supplementary Fig. 4 Fine mapping yields increased enrichment for positive control genes. Comparison of fold enrichment for positive control genes within ±25 kbp of (**a**), significant ($P < 5 \times 10^{-8}$) or (**b**), $\log_{10}(BF) > 2$ SNVs. Fold enrichment was calculated as the proportion of positive control genes targeted to the proportion of all genes targeted.



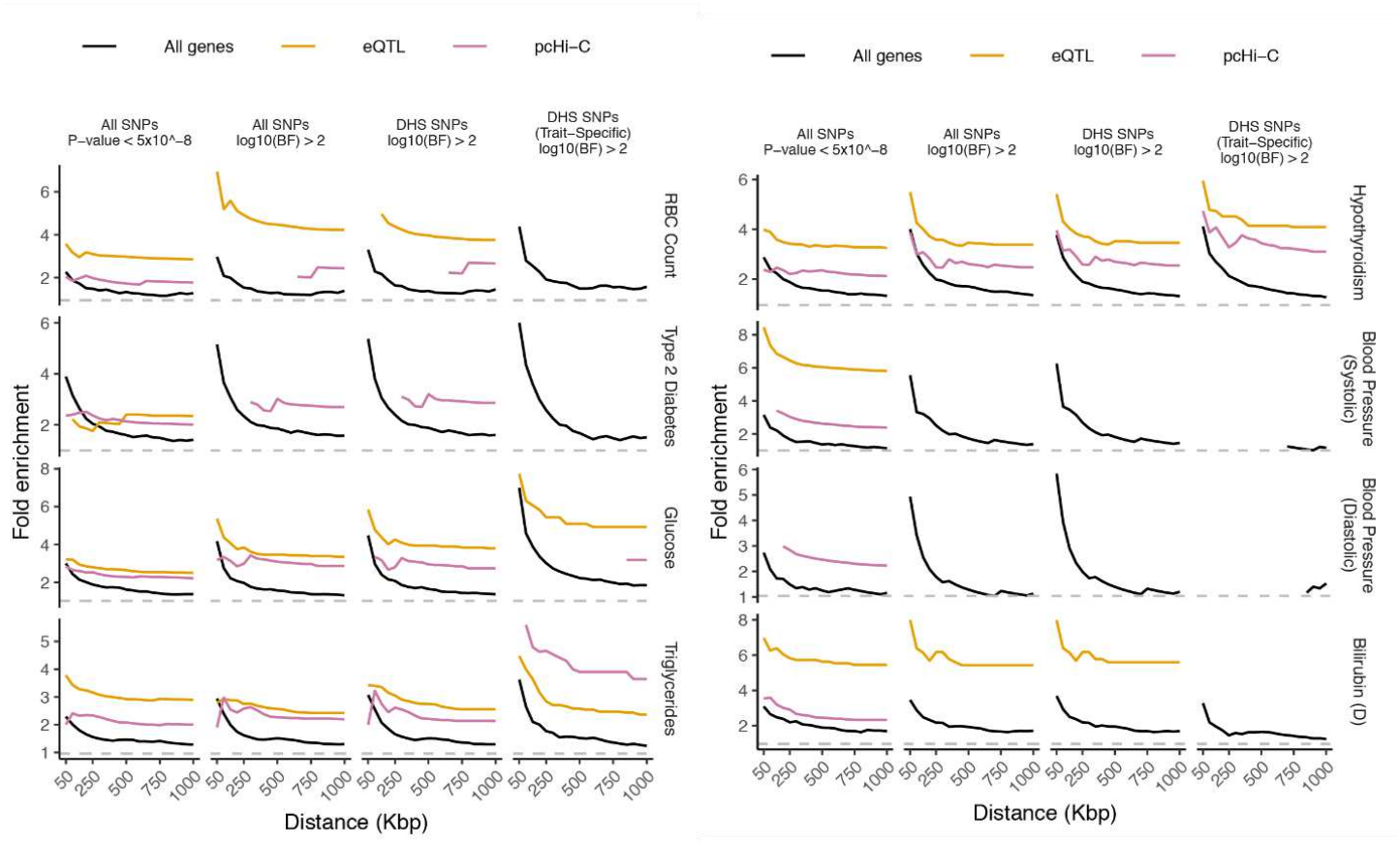
Supplementary Fig. 5. Tissue-selective expression positive control gene sets.
(**a**) Expression of genes by RNA-seq in relevant tissue-selective sets (x-axis) for 32 tissues from GTEx (y-axis). Color indicates TPM level normalized by z-score across tissues.
(**b**) Enrichment for tissue-selective gene sets within 25 Kbp of $\log_{10}$(BF) > 2 SNVs for 5 traits.



Supplementary Fig. 6 Enrichment of SNV sets meeting increasing $\log_{10}$(BF) cutoffs for DHSs. Labeled enrichments correspond to cell and tissue types selected for that trait. Enrichments were calculated as the ratio between the proportion of SNVs above threshold within DHSs and the proportion of all SNVs within DHSs.



Supplementary Fig. 7. Gene targeting enrichment (normalized).
Gene targeting enrichment over distance (x-axis) for all traits. SNV sets, DHS sets, and enrichments are as in **Fig. 2c**, but normalized to the peak enrichment in "All SNVs, P-value < 5×10^{-8}." The OMIM + Drug set for Bilirubin (D) has only 2 genes and is not shown.



Supplementary Fig. 8. Long-range gene targeting using pcHi-C and eQTL data.
Enrichments by distance shown using GTEx eQTLs and pcHi-C (**Table S5**). SNV sets and enrichments are as in **Fig. 2c**. OMIM + Drug and Expression gene sets were combined to serve as positive controls. Enrichments are shown for points where >3 positive control genes are represented.

Table 1. Summary of GWAS studies

Phenotype	Data Source	N individuals	Harmonized SNPs*	Lead SNPs‡	Loci§	Finemapped SNVs¶	Positive Control Genes	
							OMIM	Drug
Type 2 diabetes	Mahajan et al. (2018)	898,130	9,564,286	226	147	679	48	7
Estimated bone mineral density	Morris et al. (2018)	426,824	8,540,200	2,170	700	5,514	54	14
Diastolic blood pressure	Price lab, UK Biobank†	376,437	8,812,132	790	460	2,127	15	18
Height	Price lab, UK Biobank†	408,092	8,811,966	4,155	1,011	11,698	65	2
Hypothyroidism	Price lab, UK Biobank†	459,324	8,811,971	166	125	461	7	6
Red blood cell count	Price lab, UK Biobank†	445,174	8,811,899	1,310	571	3,021	66	6
Systolic blood pressure	Price lab, UK Biobank†	376,437	8,812,132	812	481	2,269	15	18
Calcium	UK Biobank, this study	380,228	10,545,875	438	251	1,976	4	11
Direct bilirubin	UK Biobank, this study	349,743	10,545,843	222	78	561	2	0
Glucose	UK Biobank, this study	375,396	10,545,731	147	106	568	71	20
Low density lipoprotein	UK Biobank, this study	333,541	10,223,520	586	243	2,201	17	5
Triglycerides	UK Biobank, this study	333,891	10,223,638	454	224	1,587	17	6

*Count for SNPs present in UK Biobank and with MAF > 0.005.

†https://data.broadinstitute.org/alkesgroup/UKBB/UKBB_409K/

‡LD independent SNPs at p-value < 5×10^{-8} and r2 < 0.01, excluding SNPs within the MHC locus and those from non-converged FINEMAP loci

§Merged loci containing one or more lead SNPs within 50 kb of each other. Excludes loci that overlap the MHC and those that failed convergence during FINEMAP.

¶FINEMAP SNVs with $\log_{10}(BF) > 2$

OMIM: Count of genes identified through the Online Mendelian Inheritance in Man to cause a Mendelian form of the disease, or influence the trait

Drug: Count of known drug targets used to treat each disease, or influence the trait

Table 2. Performance of the Effector index

	Method	Trait*	auc.u95	AUC 95% CI
PRC	Effector Index	**All Traits**	**0.30**	**(0.27-0.3)**
		Calcium	0.82	(0.54-0.82)
		Bilirubin	0.87	(0.58-0.87)
		Diastolic BP	0.59	(0.31-0.59)
		eBMD	0.24	(0.18-0.24)
		Glucose	0.43	(0.15-0.43)
		Height	0.28	(0.23-0.28)
		LDL	0.59	(0.37-0.59)
		Hypothyroidism	0.57	(0.2-0.57)
		RBC Count	0.40	(0.31-0.4)
		Systolic BP	0.90	(0.64-0.9)
		Type 2 Diabetes	0.44	(0.3-0.44)
		Type 2 Diabetes (AMP)	0.68	(0.55-0.68)
		Triglycerides	0.52	(0.32-0.52)
	Logistic	**All Traits**	**0.15**	**(0.1-0.15)**
	eCAVIAR	**All Traits**	**0.02**	**(0.01-0.02)**
ROC	Effector Index	**All Traits***	**0.84**	**(0.77-0.84)**
		Calcium	1.00	(0.88-1)
		Bilirubin	1.00	(0.89-1)
		Diastolic BP	1.00	(0.75-1)
		eBMD	0.89	(0.77-0.89)
		Glucose	1.00	(0.04-1)
		Height	0.83	(0.66-0.83)
		LDL	0.95	(0.65-0.95)
		Hypothyroidism	1.00	(0.49-1)
		RBC Count	0.90	(0.74-0.9)
		Systolic BP	1.00	(0.71-1)
		Type 2 Diabetes	0.94	(0.67-0.94)
		Type 2 Diabetes (AMP)	0.94	(0.76-0.94)
		Triglycerides	1.00	(0.7-1)
	Logistic	**All Traits**	**0.65**	**(0.57-0.65)**
	eCAVIAR	**All Traits**	**0.78**	**(0.63-0.78)**

*All traits excludes "Type 2 Diabetes (AMP)". Type 2 Diabetes (AMP) refers to the set of positive controls independently identified by Mahajan and McCarthy. The performance metric for each trait was obtained using a leave-one-out analysis, where that trait was left out of model fitting and used only for prediction