



Mapping gene regulatory architectures driving human disease using single-cell data

Citation

Dorans, Elizabeth Rose. 2026. Mapping gene regulatory architectures driving human disease using single-cell data. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://dash.harvard.edu/handle/1/42740278>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)

HARVARD
Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Division of Medical Sciences
in the subject of Genetics and Genomics
have examined a dissertation entitled

*Mapping gene regulatory architectures driving human disease using
single-cell data*

presented by Elizabeth Rose Dorans
candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature: Shamil Sunyaev
Shamil Sunyaev (Apr 6, 2026 15:15:31 EDT)

Typed Name: Dr. Shamil Sunyaev

Signature: Alexander Gusev
Alexander Gusev (Apr 6, 2026 17:00:32 EDT)

Typed Name: Dr. Alexander Gusev

Signature: Olivia Corradin
Olivia Corradin (Apr 7, 2026 17:07:14 EDT)

Typed Name: Dr. Olivia Corradin

Signature: Luca Pinello
Luca Pinello (Apr 7, 2026 08:33:30 EDT)

Typed Name: Dr. Luca Pinello

Date: April 06, 2026

Mapping gene regulatory architectures driving human disease using single-cell data

A dissertation presented

by

Elizabeth Rose Dorans

to

The Division of Medical Sciences

in the partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

April 2026

© 2026 Elizabeth Rose Dorans

All rights reserved.

Mapping gene regulatory architectures driving human disease using single-cell data

Abstract

More than 90% of disease-associated genetic variants implicated in genome-wide association studies (GWAS) lie outside of protein-coding regions of the genome and are believed to mediate disease risk by regulating the expression of nearby genes. Because the regulatory regions (“enhancers”) harboring these variants often do not regulate the nearest gene, the biological mechanisms underlying these associations remains poorly understood. Genome-wide “single-cell” technologies assay various cellular modalities at the single-cell level and enable the detection of regulatory relationships at unprecedented resolution. However, existing computational methods for linking enhancers to their target genes using high-dimensional single-cell data fail to identify true regulatory relationships, implicate incorrect regulatory relationships, or do not adequately consider context specificity; indeed, methodological advancements in single-cell data analysis are needed to realize the promise of this powerful new data.

Single-cell “peak-gene linking” methods link enhancers to target genes by measuring associations between enhancer activity (“peaks” of accessible chromatin) and gene expression across single cells, using single-cell “multimodal” data (which simultaneously profile gene expression and chromatin accessibility). However, existing methods exhibit low concordance and power. We developed a statistical framework that integrates existing linking strategies (including genomic distance) with expression quantitative trait loci (eQTL) data to score each candidate link. Our method outperformed existing methods in several orthogonal evaluation tasks, and restricting to a focal cell type improved power to identify cell type-specific links. We constructed enhancer-

gene maps that reveal novel genes and regulatory elements underlying GWAS associations that were not implicated by other methods.

Despite this substantial improvement in power, we hypothesized that correlations among peaks of accessible chromatin may induce non-causal *tagging* enhancer-gene links. Indeed, we demonstrate that tagging effects induced by “co-accessibility” (correlations among peaks) are pervasive, by analyzing tagging correlations in CRISPR-tested enhancer-gene links and co-accessibility stratified by functional peak categories. We show that the binding of transcription factors (TFs) (particularly *pioneer TFs*, which activate repressed chromatin regions) at multiple sites across the genome drives tagging effects. We also determined that statistical “fine-mapping” to distinguish causal from tagging associations improves peak-gene linking. These findings underscore the importance of accounting for tagging effects when linking enhancers to target genes.

We further investigated cell state-dependent genetic effects on gene expression, which may often play an important role in disease etiology. Expression quantitative trait loci (eQTL) studies link regulatory variants to target genes by measuring associations between genotype and gene expression across multiple individuals. Using a single-cell eQTL mapping framework modeling cell state, we demonstrate that gene expression is often regulated by cell state-dependent genetic effects and that these “dynamic” regulatory effects are often independent from “main” effects (detected without accounting for cell state). We provide several examples demonstrating context-dependent regulatory architectures underlying GWAS associations.

Characterizing rich and multi-layered gene regulatory architectures by linking enhancers to target genes is critical to advancing our understanding of disease etiology and elucidating potential drug targets. Here, we leverage the unprecedented power and resolution of genome-wide single-cell data to develop a new method for linking enhancers to their target genes, characterize patterns of enhancer co-accessibility inducing spurious regulatory links, and investigate cell state-dependent genetic effects on gene expression.

Table of Contents

Title page	i
Copyright	ii
Abstract	iii
Table of Contents	v
Acknowledgements	vi
Chapter 1: Introduction	1
Chapter 2: Linking regulatory variants to genes by integrating single-cell multiome methods and genomic distance	11
Figures	26
Tables	32
Chapter 3: Distinguishing causal from tagging enhancers using single-cell multiome data	34
Figures	58
Tables	67
Chapter 4: Robust and fast context-dependent eQTL mapping for single-cell data	68
Figures	78
Chapter 5: Conclusion	82
BACK MATTER	84
Chapter 2 Appendix	84
Methods	85
Supplementary Note	95
Extended Data Figures	113
Supplementary Figures	126
Chapter 3 Appendix	178
Methods	179
Supplementary Figures	192
Chapter 4 Appendix	254
Methods	254
Bibliography	261

Acknowledgements

I would like to deeply thank my advisors, Alkes Price and Soumya Raychaudhuri. Learning from you has been a tremendous privilege. I would like to further thank all members of the Price and Raychaudhuri labs, my dissertation defense committee, my dissertation advisory committee, and all of the other amazing scientists I have worked with throughout my PhD for their guidance and support. I would also like to extend my deepest gratitude to my family, my friends, and my fiancé, Matt. You all mean everything to me, and I feel so lucky to share this chapter of life—and beyond—with you.

Chapter 1: Introduction

Genome-wide association studies (GWAS) identify associations between genotype and phenotype, pinpointing genomic regions underlying particular traits (e.g. height, cardiovascular disease) and nominating disease-relevant genes as potential therapeutic targets^{1,2}. GWAS leverage natural genetic “perturbations” or variations existing in a human population (rather than experimentally perturbing the genome), enabling large-scale investigations into human diseases and traits measured *in vivo*. For example, several coding variants in the *PCSK9* gene were found in GWAS to be associated with both low-density lipoprotein (LDL) levels and coronary artery disease³. *PCSK9*, which encodes an enzyme involved in regulating LDL levels, has been implicated in several cardiovascular diseases⁴ and has been successfully targeted by a class of cholesterol-lowering drugs⁵.

While GWAS have nominated thousands of disease-associated variants, the vast majority of variants implicated in GWAS lie outside protein-coding regions of the genome^{1,2,6}. Most variants significantly associated with disease lie in intergenic or intronic regions, and only a small proportion (~10%) lie in exons⁶. Furthermore, common diseases are often highly polygenic, showing diffuse patterns of variant association rather than strong, concentrated signal around one or a few genes⁶. Because many GWAS variants do not directly implicate a specific gene and numerous genomic regions may be implicated in each GWAS, pinpointing the biological mechanisms underlying GWAS variant-disease associations is often challenging.

These non-coding disease-associated variants often lie within non-coding regulatory elements called enhancers, which are hypothesized to mediate disease risk by modulating the expression of nearby genes^{7–14}. Enhancers, canonically marked by regions of open chromatin and specific histone modifications (e.g. H3K4me1, H3K27ac), have long been experimentally demonstrated to influence the expression of nearby genes^{15,16}. Partitioned heritability analyses suggest that regions of open chromatin account for <20% of common SNPs tested in GWAS but

harbor the vast majority of GWAS signal, reflecting substantial enrichment for causal signal^{17,18}. GWAS variants are particularly enriched in regions with histone modifications associated with enhancer activity (H3K4me1, H3K27ac)¹⁹.

Importantly, enhancer activity is highly cell type-specific^{19–22}. Canonical enhancer marks (open chromatin, histone modifications) often differ greatly across cell types²⁰, and disease association signals are often highly enriched in regions with enhancer-associated marks in disease-relevant cell types¹⁹. Similarly, the genetic architectures driving gene expression shift substantially throughout cellular differentiation and organismal development²². Because enhancer-gene relationships are highly dynamic, they can be difficult to pinpoint (data profiling the relevant context may be unavailable).

Furthermore, non-coding variants often do not regulate the nearest gene^{23–25}. Genomic distance is a powerful predictor of enhancer-gene regulatory interactions and the *cis* (proximal) region surrounding a gene often harbors the most regulatory effects^{25–27}. However, transcriptome-wide association studies (TWAS) demonstrate that the genes whose genetically predicted expression is most strongly associated with disease are often not the closest genes to disease-associated variants²³; the nearest gene may be the target gene for as few as ~50% of disease GWAS variants²⁴. Genetic regulation of gene expression is often orchestrated by distal regulatory elements, precluding the use of a simple nearest-gene method to link GWAS variants to their target genes. Thus, the specific biological mechanisms underlying GWAS associations remain poorly understood^{2,28}.

Outside of the lens of GWAS, mapping enhancer-gene regulatory architectures is critical for understanding the cellular circuitry driving key developmental and disease processes. Indeed, functionally important genomic regions are depleted of variation due to purifying selection; thus, enhancers that do not harbor common genetic variation may be among the most disease-critical regions (despite SNPs within the enhancer being absent or too rare to conduct a well-powered GWAS)^{29–33}. Networks of enhancers, in concert with transcription factors (TFs) and their target

genes, are hypothesized to regulate cellular development, respond to environmental stimuli, and orchestrate a complex array of biological processes that work to maintain homeostasis and fend off internal and external threats^{13,26,34–36}. A nascent category of gene therapy involves targeting enhancers (either by modifying the activity of existing enhancers or by inserting new enhancers in a relevant genomic location), which may provide an avenue for modulating gene expression while avoiding the safety risks of directly manipulating coding sequences^{37,38}. Because enhancers are highly context-specific, enhancer-targeting therapies may be safer and more effective than directly modifying a gene, which can induce off-target effects.

For example, the SNP rs1427407, confidently fine-mapped for red blood cell count, lies in an experimentally validated intronic enhancer of *BCL11A*. *BCL11A* is known to play a critical role in fetal-to-adult hemoglobin switching in erythroid cells^{39,40}, and the disruption of this enhancer has been experimentally shown to result in decreased *BCL11A* expression and elevated fetal hemoglobin⁴¹. Experimental evidence also demonstrates that activity of the focal enhancer is necessary for *BCL11A* expression in erythroid cells but not in other cell types⁴², underscoring the cell type-specificity of trait-relevant enhancer activity.

Previous approaches have used data from DNA sequence and epigenomic assays to approximate the genomic location and activity of enhancers. Because functionally important genomic regions are depleted of genetic variation, scores quantifying DNA sequence conservation such as PhyloP and phastCons can be suggestive of enhancer activity^{29,43}. Several genome-wide biochemical assays can also provide suggestive evidence of enhancer activity. Assay for Transposase-Accessible Chromatin using sequencing⁴⁴ (ATAC-seq) and DNase I hypersensitive site sequencing⁴⁵ (DNase-seq) measure regions of open chromatin, which are biologically available for interactions with transcription factors and other transcription machinery. Histone chromatin immunoprecipitation sequencing (ChIP-seq) maps the locations of DNA histone modifications related to enhancer activity (e.g. H3K27ac, H3K4me1)⁴⁶. Transcription factor ChIP-seq pinpoints the binding sites of transcription factors, which are hypothesized to

affect gene expression in concert with enhancers³⁶. While these methods generate approximate maps of putative enhancers across the genome, they do not intrinsically link enhancers to the genes whose expression they regulate.

Previous methods to link enhancers *to their target genes* include large-scale or targeted experimental validation, genetic association studies, additional genome-wide epigenomic assays, and integrated computational approaches.

CRISPR-based approaches (including CRISPR interference, CRISPR activation, base editing, and Perturb-seq), considered the current “gold standard” for identifying enhancer-gene pairs, perturb a candidate enhancer and detect resulting changes (or lack thereof) in expression of a candidate target gene^{47–59}. However, CRISPR-based approaches often have low efficiency and highly variable effectiveness (efficiency can vary dramatically across guide RNAs targeting the same gene⁶⁰). These experiments are often costly and laborious and may only simultaneously profile a few elements or genes (though recent approaches have increased the number of tested element-gene pairs per experimental input)^{47–59}. Furthermore, these experiments are often restricted to cell lines that are prone to structural abnormalities and that may not reflect disease-relevant *in vivo* contexts⁶¹.

Massively parallel reporter assays (MPRA)⁶² quantify changes in gene expression in response to enhancer activity by cloning the candidate enhancer into a reporter construct and measuring expression of a reporter gene. While MPRA are less expensive and traditionally higher-throughput, they do not reflect gene regulation in an enhancer’s native chromatin context and thus may be prone to false positive or biologically implausible results⁶². Like CRISPR-based approaches, MPRA are also restricted to reporter systems which do not reflect *in vivo* contexts relevant for human disease.

Expression quantitative trait loci (eQTL) measure associations between genotype and gene expression across individuals^{63,64}. eQTL studies are restricted to genomic regions harboring genetic variation and are not well-powered for low-frequency or rare variants^{63,65}; because

functionally important genomic regions are depleted of variation^{29–33}, eQTL studies may not characterize disease-critical regulatory elements. Linkage disequilibrium (LD), correlation between SNPs, also limits specificity of variant-gene associations, with many non-causal SNP-gene associations induced by SNP-SNP associations. Importantly, unlike CRISPR- and MPRA-based studies, eQTL only provide evidence of *association* and no direct evidence that a focal SNP is *causal* in influencing focal gene expression; associations may instead reflect the biological action of a biologically related or co-expressed gene. Recent work has also demonstrated limited overlap between variants implicated by eQTL and GWAS^{66,67}; indeed, only ~10% of disease-causal genes have colocalizing GWAS and eQTL variants²¹. One possible explanation for this discrepancy is context-specific gene regulatory effects.

High-throughput chromatin conformation capture (Hi-C) detects 3D spatial interactions between genomic regions; in particular, promoter capture Hi-C enriches for interactions between non-coding regions and genes, nominating potentially causal enhancer-gene pairs⁶⁸. Further, HiChIP combines Hi-C and ChIP to detect 3D chromatin contacts associated with the binding of a specific protein and can be used to identify putative enhancer-promoter loops in 3D space⁶⁹. However, Hi-C-based methods are difficult to conduct in primary cells, because a large number of cells are required; thus, chromatin contexts detected by Hi-C may not reflect disease-relevant interactions. These also do not imply causal enhancer-gene links and may be confounded by other structural factors influencing 3D genome conformation.

Several computational approaches (e.g. EpiMap²⁰, ABC²⁷, ENCODE-rE2G²⁶) simultaneously leverage multiple data types (e.g. transcriptomic, epigenomic, chromatin accessibility, and 3D contact information) to infer regulatory links between non-coding elements and their target genes. While these computational approaches are powerful in predicting genome-wide enhancer-gene maps, they represent predicted associations and do not themselves provide evidence of causality. They also require inputs from large-scale experiments and/or “gold

standard” validation sets of enhancer-gene links and thus may be difficult to apply to specific or rare disease contexts without significant experimental efforts.

Importantly, most prior methods were developed and applied to *bulk* data (which comprise a single readout per sample representing an aggregate of all cells in the sample). Because enhancer activity is highly cell type-specific^{19–22}, bulk methods may be underpowered to detect many disease-relevant context-specific enhancer-gene regulatory effects³⁴. Bulk data sets are also laborious to assemble, often requiring the recruitment of hundreds or thousands of human participants to achieve meaningful statistical power (e.g. for eQTL studies).

Single-cell data (which provide a single readout for each individual cell within a sample) characterize biological processes at high resolution^{36,70,71}. In their most proximal applications, single-cell measurements enable the precise definition of a homogeneous sample from a biological context of interest (for example, by restricting a single-cell RNA-seq data set to cells that express markers of a particular cell type or disease state) prior to applying traditional bulk methods (“pseudobulk”)^{72,73}. New computational methods have been further extended to not only refine bulk samples but leverage biological variation across individual cells^{74,75}.

In particular, single-cell multiome technologies co-assaying chromatin accessibility and gene expression (scRNA+ATAC-seq) profile the enhancer-gene regulatory landscape at high resolution. By jointly capturing the activity of candidate enhancers (as measured by “peaks” of open chromatin in scATAC-seq) and candidate target genes (as measured by scRNA-seq), these technologies enable the detection of regulatory interactions by associating the accessibility of non-coding regulatory elements with gene expression across individual cells, revealing rich profiles of gene regulatory processes in diverse cell types and states^{76–86}. In addition to providing complementary information on multiple cellular modalities, scRNA+ATAC-seq data capture *in vivo* regulatory contexts and thus may have greater potential to reflect disease-relevant regulatory interactions relative to data collected in cell lines or from experimentally perturbed samples.

‘Peak-gene linking methods,’ computational methods which predict enhancer-gene links from scRNA+ATAC-seq data by correlating the accessibility of ATAC peaks with genes, have shown particular promise in linking regulatory elements to their target genes^{77,80–92}. Other studies have produced methods for mapping peak co-accessibility^{82,83,93} or constructed ‘modules’ of co-accessible peaks^{85,94} from single-cell ATAC-seq data, inferring a larger regulatory network in which enhancer-gene interactions operate and raising the question of coordinated enhancer action^{95–98}.

In addition, single-cell eQTL mapping methods have been developed to capture genetic variants controlling gene expression at a more granular level. These approaches can model gene expression values from individual cells as multiple draws from a distribution, rather than using a single gene expression value aggregated across cells, providing a richer profile of genetic regulatory control^{71,74,99}. Disease-driving genetic variants may only affect gene expression in a specific rare or transitional cell state; thus, their effects may not be detected by eQTL studies using bulk data^{21,22,70,100}. In contrast to bulk and pseudobulk eQTL mapping, single-cell approaches can characterize eQTL in fine-grained cell states (e.g. lymphocyte cytotoxicity, activation, cellular stress, etc.), which are often implicated in disease mechanisms and difficult to precisely define⁹⁹.

While methods characterizing gene regulatory architectures from single-cell data show promise, statistical and biological challenges of analyzing these data persist.

For peak-gene linking, causal inference remains challenging due to a lack of consensus around enhancer-gene modeling approaches as well as biological and technical correlation structure present in single-cell data. Existing methods take varied approaches to selecting candidate peak-gene links, modeling the peak-gene relationship^{77,80–90}. Additionally, previously proposed peak-gene linking methods have attempted to account for false positive links induced by technical and biological confounders^{77,80,81,84,88} and gene co-expression⁸⁷, but have not fully explored how enhancer correlations may confound enhancer-gene linking.

For single-cell eQTL mapping, single-cell data sets used for eQTL mapping are also more complicated in structure than bulk data sets. Each observation in single-cell eQTL mapping is a single cell, and individuals may differ in their overall and cell type-specific cell abundance. Therefore, these approaches often require complex or hierarchical models (e.g. to distinguish cell-level vs. donor-level effects) to avoid false positives (from assuming independence across non-independent observations), which impose a significant computational burden^{71,86,89,101}. Methods that handle the statistical, biological, and computational challenges of genome-wide single-cell data are needed to realize the promise of emerging high-resolution and high-dimensional single-cell eQTL data sets^{36,88,91}.

For both peak-gene linking methods and single-cell eQTL methods, single-cell data is subject to high levels of measurement noise, which often does not follow a normal distribution and is thus not amenable to standard statistical solutions or assumptions^{102,103}. In these conditions, linear models may produce false positive associations. To combat this, single-cell models must implement careful normalization^{81,82,104} or use Poisson or zero-inflated negative binomial models rather than linear models^{80,99}. (Also of note, these methods are purely correlational, measuring associations between ATAC peaks or SNPs and genes; they do not provide causal evidence of a regulatory relationship.)

Characterizing gene regulatory architectures by linking enhancers to their target genes is critical for the functional interpretation of GWAS discoveries and understanding the circuitry driving key developmental and disease processes. In the following chapters, we develop and characterize statistical methods to link enhancers (and the regulatory variants they harbor) to their target genes, carefully considering the challenges and leveraging the advantages of high-dimensional genome-wide single-cell data.

In Chapter 2, we present a new statistical method for linking enhancers to their target genes using single-cell multiome data. While peak-gene linking methods show promise in linking regulatory elements to genes, we show that existing methods exhibit low concordance and do not

capture the effects of genomic distance. We present pgBoost, an integrative modeling framework that trains a non-linear combination of existing linking strategies (including genomic distance) on eQTL data to assign a probabilistic score to each candidate SNP-gene link. We show that pgBoost outperforms existing methods (including genomic distance) on evaluation sets derived from genetic and experimental data. We further demonstrate that restricting pgBoost to features from a focal cell type improved power to identify links relevant to that cell type. We highlight several examples where pgBoost linked fine-mapped GWAS variants to experimentally validated or biologically plausible target genes that were not implicated by other methods.

In Chapter 3, we investigate correlation structure in single-cell multiome data, which may induce false positive enhancer-gene associations, and propose a framework for improving peak-gene linking accuracy in light of this structure. We demonstrate that “tagging” effects (non-causal peak-gene associations induced by peak-peak correlations) are pervasive in peak-gene linking by analyzing peak-peak correlations among CRISPR-tested enhancer-gene pairs. We further show that peak-peak correlations are substantially driven by shared transcription factor (TF) activity within correlated peaks. Consistent with widespread tagging, peak-gene links that we fine-mapped using SuSiE significantly outperformed marginal peak-gene links in evaluation sets derived from genetic and experimental data. We provide examples demonstrating the impact of tagging effects at specific peaks and genes implicated in GWAS of blood cell traits.

In Chapter 4, we describe a new method for mapping cell state-dependent genetic effects on gene expression and characterize the landscape of context-dependent gene regulation. We validate and apply a dynamic single-cell eQTL framework with an interaction term to model continuous cell states across single cells, identifying genetic variants that affect gene expression in a cell state-dependent manner. We discover a substantial number of genes with genetic effects that interact with cell state, and we demonstrate that these effects are sometimes independent from the main genetic effect. We further show that interaction genetic effects colocalize with

GWAS associations and provide examples. Our findings suggest that dynamic eQTL mapping reveals novel regulatory variant-gene relationships.

In conclusion, each of these efforts will advance our understanding of gene regulatory architectures, provide new insights into disease etiology, and elucidate potential therapeutic targets. Training a non-linear combination of linking strategies improves power to identify target genes underlying GWAS associations. Investigating the correlation structure in single-cell multiome data demonstrates the importance of accounting for tagging effects when linking enhancers to target genes, leveraging information from broader enhancer regulatory networks to reduce false positive enhancer-gene associations. Profiling conditional dynamic eQTL reveals previously undiscovered cell state-dependent enhancer-gene relationships. These efforts leverage the unprecedented power and resolution of genome-wide single-cell data to characterize rich and multi-layered gene regulatory architectures and advance our understanding of human disease biology.

Chapter 2: Linking regulatory variants to genes by integrating single-cell multiome methods and genomic distance

Elizabeth Dorans, Karthik Jagadeesh, Kushal Dey*, Alkes L. Price*

*These authors contributed equally

Introduction

More than 90% of disease-associated genetic variants implicated in genome-wide association studies (GWAS) lie outside of protein-coding regions of the genome⁶. However, non-coding variants often do not regulate the nearest gene^{23–25}; thus, the biological mechanisms underlying these associations remains poorly understood^{2,28}. Previous methods have leveraged transcriptomic, epigenomic, chromatin accessibility, and 3D contact information to infer regulatory links between non-coding elements and their target genes^{20,25–27}, but methods analyzing data from bulk tissues or cell lines may be underpowered to detect cell-type-specific patterns³⁴. Nascent single-cell multiome technologies co-assaying chromatin accessibility and gene expression (scRNA/ATAC-seq) enable the detection of context-specific regulatory relationships by correlating the accessibility of candidate regulatory elements (as measured by ATAC-seq peaks) with gene expression (as measured by RNA-seq) across individual cells^{76–86}. While existing peak-gene linking methods have shown promise^{80–86}, we observe low concordance in their linking scores (**Figure 2.1a**), and these methods underperform a simple genomic distance-based linking score in their enrichment for fine-mapped eSNP-eGene pairs (**Figure 2.1b**) and other evaluation sets (see below).

Here, we propose an eQTL-informed gradient boosting¹⁰⁵ approach (pgBoost) that integrates linking scores from existing peak-gene linking methods across cell types and data sets with genomic distance, training on fine-mapped eQTL data to assign a single probabilistic score

to each candidate SNP-gene link. We evaluate the performance of pgBoost and existing methods using sets of SNP-gene links derived from eQTL^{63,64}, Activity-By-Contact (ABC)^{26,27}, CRISPR^{47–55}, and GWAS^{106,107} data. We also investigate whether restricting to single-cell data from a focal cell type can improve power to detect links relevant to that cell type. We further highlight examples demonstrating that our integrative approach can aid in the biological interpretation of GWAS variant-disease associations. Accurately mapping regulatory architectures can improve our understanding of disease mechanisms and highlight regulatory relationships with therapeutic potential.

Results

Overview of Methods

pgBoost integrates information from existing peak-gene linking scores derived from single-cell RNA/ATAC-seq multiome data^{77,80–82} (or scATAC-seq data⁸³), as well as genomic distance, to predict regulatory SNP-gene links (**Figure 2.2**). The existing scores assess the co-activity of ATAC peaks and genes^{77,80–82} (or the co-accessibility of peaks and gene promoters⁸³) across single cells^{77,80,81} (or metacells^{82,83}) and are computed separately for each multiome data set and cell type. pgBoost optimizes a gradient boosting classification task¹⁰⁵, using existing scores from each data set and cell type and genomic distance as features and fine-mapped GTEx eSNP-eGene pairs^{63,64} as training data (**Figure 2.2**). The gradient boosting algorithm, a decision tree-based ensemble approach, learns the non-linear relationships among these features to predict the causal probability of candidate SNP-gene links.

In this study, we computed peak-gene linking scores using the existing single-cell methods SCENT⁸⁰, Signac^{77,81}, ArchR⁸², and Cicero⁸³, restricting to *cis* peak-gene pairs with peak-TSS distance >1kb and <500kb, thus focusing primarily on enhancer-gene links rather than promoter-gene links (**Table 2.1** and **Methods**). To integrate linking scores from existing methods across data sets, we intersected ATAC peaks with 1000 Genomes Project variants with minor allele count ≥ 5 , thus analyzing SNP-gene links instead of peak-gene links. We selected pgBoost features

from the output of SCENT, Signac, ArchR, and Cicero using evaluation sets of SNP-gene links (see below and **Supplementary Note**). After this feature selection step, pgBoost features for each candidate SNP-gene link included 3 features for each data set and cell type—peak-gene effect sizes generated by SCENT, Signac, and Cicero—and 2 distance-based features: SNP-TSS distance and a binary variable indicating whether the candidate SNP links to the gene with closest TSS (**Table 2.2** and **Methods**). As positive and negative training sets, we used fine-mapped eSNP-eGene pairs assigned maximum PIP > 0.2 and maximum PIP < 0.01 across GTEx tissues^{63,64}, respectively, incorporating training data from 49 tissues (**Supplementary Figure 2.1** and **Methods**). To avoid overfitting, we trained pgBoost using a leave-one-chromosome-out (LOCO) approach, generating predictions on held-out chromosomes. We have publicly released open-source software implementing pgBoost (see Code Availability).

We compared pgBoost to each existing peak-gene linking method (specifically, the top score assigned by each method across data sets and cell types analyzed) and to a distance-based method that ranks candidate SNP-gene links based on their genomic distance (**Methods**). For a given evaluation data set and a given method, we computed *average enrichment across recall values*, defined as the enrichment of predicted SNP-gene links in the set of evaluation links, averaged across values of recall and upweighting enrichments at lower values of recall (**Methods**). To evaluate performance at more proximal vs. more distal links, we computed average enrichment when restricting candidate and evaluation SNP-gene links to various distance thresholds ranging from >1kb to >100kb. We conservatively restrict the set of candidate SNP-gene links used in all computations to SNP-gene pairs involving a SNP lying in an ATAC peak (not all *cis* SNP-gene pairs as in ref.⁸⁰), leading to much lower enrichment for all methods.

We analyzed 4 scRNA/ATAC-seq multiome data sets (10x PBMC¹⁰⁸, Luecken BMMC¹⁰⁹, SHARE-seq LCL⁷⁷, Xu K562¹¹⁰) that each spanned 1-5 blood/immune cell types (total of 10 data set-cell type pairs, spanning 7 cell types and 93K cells after QC) (**Table 2.2**, **Supplementary Figure 2.2**, and **Methods**; see **Data Availability**). This resulted in 32 pgBoost features (combined

in a single model), including 2 features based on genomic distance. We computed each of these features for 6,882,594 candidate SNP-gene links. We have publicly released linking scores and percentiles for every method considered in this study (see **Data Availability**).

pgBoost enrichment in eQTL and ABC evaluation sets

We constructed pgBoost predictions and assessed whether SNP-gene links predicted by pgBoost recapitulated SNP-gene links predicted by eQTL (via fine-mapped eSNP-eGene pairs^{63,64}) and Activity-By-Contact (ABC)^{26,27}. We compared pgBoost to existing peak-gene linking methods and to restricted gradient boosting models using linking scores from a single method plus distance as features. We note that although SNP-gene links predicted by eQTL/ABC are only partially accurate and represent statistical predictions rather than gold-standard links, high enrichment with respect to these evaluation sets is still suggestive of high power to detect true links. We also note that, in primary analyses, pgBoost uses the eQTL evaluation set for training. We use a LOCO approach such that overfitting, if present, would reduce (and not improve) model performance on the eQTL evaluation set. However, results on the eQTL evaluation set should still be viewed as a best-case scenario for pgBoost.

Results for 4,434 SNP-gene links implicated by eQTL (fine-mapped eSNP-eGene pairs with maximum PIP > 0.5 across GTEx tissues^{63,64}; see **Data Availability**) are reported in **Figure 2.3a** and **Supplementary Figure 2.3**. pgBoost significantly outperformed the 4 existing scores (SCENT, Signac, ArchR, Cicero) at all distance thresholds, including >10kb (average enrichment 12x vs. 1.9-2.2x, all p-values for difference < 4e-69). pgBoost did not significantly outperform the distance-based method when including all links (distance >1kb; 18.3x vs. 17.7x, p = 0.27), but significantly outperformed it at increasingly distal thresholds, including >10kb (12.0x vs. 10.6x, p = 2e-4) and >100kb (12x vs. 3.9x, p = 2e-5). pgBoost also significantly outperformed a gradient boosting model restricted to SCENT scores plus distance as features, but did not significantly outperform gradient boosting models restricted to Signac, ArchR, or Cicero scores plus distance (**Figure 2.3b**) (but see results for other evaluation sets below). Analyses using secondary

evaluation metrics produced similar findings (**Supplementary Figure 2.5a, Supplementary Figure 2.6a, Supplementary Figure 2.7a**).

Results for 53,701 SNP-gene links implicated by ABC (ABC score > 0.2 in at least one of 344 biosamples^{26,27}) are reported in **Figure 2.3c** and **Supplementary Figure 2.4**. pgBoost significantly outperformed all 4 existing scores (SCENT, Signac, ArchR, Cicero) at >1kb (65x vs. 2.1-4.9x, all p-values for difference < 2e-308) and >10kb (37x vs. 2.9-10x, all p-values for difference < 4e-45), but not at >100kb (12x vs. 3.5-18x, p-values for difference = 0.01-0.72). pgBoost did not significantly outperform the distance-based method when including all links (distance >1kb; 65x vs. 66x, p = 0.25), but significantly outperformed it at increasingly distal thresholds, including >10kb (37x vs. 26x, p = 2e-11) and >100kb (12x vs. 4.6x, p = 0.04); we note that the ABC evaluation set strongly favors links at shorter distances (thus favoring the distance-based method), as it is derived from Hi-C contact frequencies that are largely determined by genomic distance^{52,111} (**Supplementary Figure 2.8**). pgBoost also outperformed gradient boosting models restricted to SCENT, Signac, or ArchR scores plus distance as features, but did not significantly outperform a gradient boosting model restricted to Cicero scores plus distance (**Figure 2.3d**). Analyses using secondary evaluation metrics produced similar findings (**Supplementary Figure 2.5b, Supplementary Figure 2.6b, Supplementary Figure 2.7b**).

We performed 2 secondary analyses. First, we compared pgBoost to a logistic regression model using identical features and determined that gradient boosting significantly outperformed logistic regression on eQTL evaluation data (**Extended Data Figure 2.1a-b**). Second, we compared pgBoost to an eQTL-based linking strategy. We determined that pgBoost significantly outperformed the eQTL-based linking strategy on ABC evaluation data (**Extended Data Figure 2.2a**). Additional secondary analyses are reported in the **Supplementary Note** and **Supplementary Figures 1.9-1.15**.

As more distal links are generally more difficult to detect and are of high biological interest (in particular, causal variants in disease GWAS often do not link to the closest gene^{23–25}), these results indicate substantial potential for pgBoost to identify biologically important SNP-gene links.

pgBoost enrichment in CRISPR and GWAS evaluation sets

We assessed whether SNP-gene links predicted by pgBoost recapitulated relatively smaller sets of gold-standard SNP-gene links that were either experimentally validated using CRISPR^{47–55} or implicated by fine-mapped non-coding variants near genes with fine-mapped coding variants in disease/complex trait GWAS^{106,107}, providing strong evidence of biological causality. We note that the CRISPR SNP-gene links were derived from regulatory element-gene links based on SNPs that reside in a regulatory element.

Results for 892 gold-standard SNP-gene links experimentally validated in 9 CRISPR perturbation studies^{47–55} (see **Data Availability**) are reported in **Figure 2.4a** and **Supplementary Figure 2.16**. pgBoost significantly outperformed the 4 existing scores (SCENT, Signac, ArchR, Cicero) at >1kb (7.6x vs. 1.5-3.3x, all p-values for difference < 2e-12) and >10kb (6.5x vs. 1.5-3.3x, all p-values for difference < 6e-6), but not at >100kb (8.4x vs. 1.5-3.1x, p-values for difference = 0.03-0.09). pgBoost significantly outperformed the distance-based method at all distance thresholds, including >10kb (6.5x vs. 5.0x, p = 4e-3). pgBoost also outperformed gradient boosting models restricted to SCENT, ArchR, or Cicero scores plus distance as features, but did not significantly outperform a gradient boosting model restricted to Signac scores plus distance (**Figure 2.4b**). Analyses using secondary evaluation metrics produced similar findings (**Supplementary Figure 2.18a**, **Supplementary Figure 2.19a**, **Supplementary Figure 2.20a**). We note that all CRISPR evaluation data was derived from the K562 erythroleukemia cell line, a cell line that is widely analyzed in CRISPR experiments⁶¹; while CRISPR experiments are viewed as a general gold-standard^{26,80}, the relevance of these links to individual cell types will be explored below.

Results for 155 gold-standard SNP-gene links implicated by fine-mapped non-coding variants near genes with fine-mapped coding variants¹⁰⁶ in GWAS of 94 diseases/complex traits from the UK Biobank^{107,112} (see **Methods** and **Data Availability**) are reported in **Figure 2.4c** and **Supplementary Figure 2.17**. pgBoost significantly outperformed the 4 existing scores (SCENT, Signac, ArchR, Cicero) at all distance thresholds, including >10kb (6.9x vs. 1.9-2.1x, all p-values for difference < 6e-7). pgBoost and the distance-based method performed similarly when including all links (distance >1kb; 6.8x vs. 6.7x, p = 0.87), but pgBoost performed significantly better at distance >10kb (7.0x vs. 5.2x, p = 3e-3) and >100kb (13x vs. 3.6x, p = 0.04). Differences between pgBoost and gradient boosting models restricted to one existing score plus distance were not statistically significant (**Figure 2.4d**). Analyses using secondary evaluation metrics produced similar findings (**Supplementary Figure 2.18b**, **Supplementary Figure 2.19b**, **Supplementary Figure 2.20b**).

We evaluated the performance of pgBoost trained on other training data sets, using the same set of input features. First, we trained pgBoost using CRISPR links. We found that training on eQTL significantly outperformed training on CRISPR for eQTL and ABC evaluations but yielded similar performance on CRISPR and GWAS evaluations, motivating further research on training using CRISPR data (**Extended Data Figure 2.3**). Second, we trained pgBoost using fine-mapped eSNP-eGene pairs in whole blood (the GTEx tissue closest to the single-cell data) instead of all GTEx tissues. Training on eSNP-eGene pairs from all GTEx tissues significantly outperformed training on blood eSNP-eGene pairs for eQTL, ABC, CRISPR, and GWAS evaluations (**Extended Data Figure 2.4**). Additional training set analyses are reported in the **Supplementary Note** and **Supplementary Figures 1.21-1.25**.

To select features for the main pgBoost model, we performed sequential feature selection, comparing the performance of pgBoost models using different feature sets on eQTL, ABC, CRISPR, and GWAS evaluations (**Supplementary Note**). This approach selected SCENT, Signac, and Cicero features (but not ArchR features), in addition to distance-based features, for

inclusion in the main pgBoost model. pgBoost significantly outperformed restricted gradient boosting models using a subset of existing scores as features and an expanded model adding ArchR features (**Figure 2.3b, Figure 2.3d, Figure 2.4b, Figure 2.4d, Extended Data Figure 2.5, Extended Data Figure 2.6**), for various subsets of evaluation sets. Additional feature set analyses are reported in the **Supplementary Note** and **Supplementary Figures 1.26-1.32**.

We also compared pgBoost to the ENCODE-rE2G method²⁶, a logistic regression model using chromatin activity and 3D interactions as features and using CRISPRi data^{47–55} from K562 cells to guide training (see **Data Availability**). We determined that pgBoost significantly outperformed ENCODE-rE2G on eQTL evaluation data (**Extended Data Figure 2.7a**) and ENCODE-rE2G significantly outperformed pgBoost on CRISPR evaluation data (**Extended Data Figure 2.7b**), consistent with the training data used by each method. In addition, pgBoost significantly outperformed ENCODE-rE2G on longer-range (>50kb and >100kb) GWAS evaluation data (**Extended Data Figure 2.7c**). Further details and additional secondary analyses are reported in the **Supplementary Note** and **Supplementary Figures 1.33-1.34**.

These results indicate substantial power for pgBoost to detect experimentally validated links and links underlying disease/complex trait GWAS associations. In particular, pgBoost strongly outperformed the 4 existing scores and distance-based method in capturing longer-range links.

Identifying SNP-gene links in a focal cell type

pgBoost leverages features from an input set of data set-cell type pairs, which may often span multiple cell types (**Figure 2.2, Table 2.2**). However, in settings where the goal is to predict SNP-gene links in a focal cell type, it may be beneficial to restrict model features to that cell type. We generated independent pgBoost predictions for each of the 4 major cell types of the Luecken BMMC¹⁰⁹ multiome data set (spanning >8,000 cells: T cell, B cell, myeloid, erythroid; **Table 2.2**), K562 cells from the Xu¹¹⁰ data set, and oligodendroglia from an additional brain multiome data set (Meijer¹¹³; see **Data Availability**) by restricting model features to each cell type in turn (as

well as the 2 distance-based features; **Methods**) during model training and prediction. We continued to use fine-mapped eSNP-eGene pairs^{63,64} (maximum PIP > 0.2 vs. maximum PIP < 0.01 across GTEx tissues; **Methods**) as training data. We compared these pgBoost predictions (as well as pgBoost predictions using model features from multiple cell types) using two different types of evaluation involving a focal cell type: *cell-type-specific* evaluation, involving evaluation sets of SNP-gene links that are present in the focal cell type and absent from other cell types, and *cell-type-level* evaluation, involving evaluation sets of SNP-gene links that are present in the focal cell type (irrespective of presence/absence in other cell types).

We first analyzed cell-type-specific ABC evaluation data in 4 blood/immune cell types, defined by SNP-gene links strongly implicated in biosample(s) related to the focal cell type (ABC score > 0.2) and not strongly implicated in other biosamples (ABC score < 0.1; see **Methods**). We compared the performance of pgBoost models restricted to T cells, B cells, myeloid cells, erythroid cells, K562, or the union of these 5 cell types on cell-type-specific ABC evaluation sets from T cells, B cells, myeloid cells, or K562 (excluding erythroid, which was not available). Across the 4 evaluation sets, the pgBoost model restricted to features from the focal cell type significantly outperformed models restricted to other cell types and usually significantly outperformed the model spanning the union of all 5 cell types (**Figure 2.5a**, left panel and **Supplementary Figure 2.35**); differences between models were not statistically significant at other distance thresholds besides >1kb (**Supplementary Figure 2.36**).

We next analyzed tissue-specific ABC evaluation data in blood or brain, similarly defined by SNP-gene links strongly implicated in biosample(s) related to a focal tissue and not strongly implicated in other biosamples. We compared the performance of pgBoost models restricted to K562 or brain oligodendroglia on tissue-specific ABC evaluation sets from blood or brain tissue. Across blood and brain evaluation sets, the pgBoost model restricted to features related to the focal tissue (K562 or brain oligodendroglia, respectively) significantly outperformed the model restricted to features related to the non-focal tissue (**Figure 2.5a**, right panel and

Supplementary Figure 2.37); the model spanning both tissues significantly outperformed both tissue-specific models on blood-specific ABC evaluation data, but significantly underperformed the model restricted to brain oligodendroglia on brain-specific ABC evaluation data. Results were similar at other distance thresholds (**Supplementary Figure 2.36**).

We next analyzed cell-type-level CRISPR evaluation data^{47–55}, defined by SNP-gene links experimentally validated in the K562 cell line. The pgBoost model restricted to features from K562 significantly outperformed models restricted to other cell types (**Figure 2.5b** and **Supplementary Figure 2.38**); the model restricted to K562 did not significantly outperform the model spanning the union of all 5 blood/immune cell types at >1kb, but significantly outperformed it at more distal thresholds (> 5kb to >100kb) (**Supplementary Figure 2.39**). We did not consider cell-type-specific CRISPR evaluation, because all CRISPR evaluation data was derived from K562.

We also analyzed cell-type-level GWAS variant evaluation data, defined by fine-mapped variants for 7 autoimmune diseases and granulocyte-related blood cell traits (for which T cells, B cells and myeloid cells are most relevant) and 7 red blood cell or platelet-related blood cell traits (for which erythroid cells are most relevant) (**Supplementary Note**). For the first (resp. second) set of traits, the pgBoost models restricted to B cells and myeloid cells (resp. erythroid cells) outperformed models restricted to other cell types (**Extended Data Figure 2.8**).

We also analyzed cell-type-level GTEx evaluation data, defined by fine-mapped eSNP-eGene pairs identified in EBV-transformed lymphocytes or LCL. A pgBoost model restricted to features from the SHARE-seq LCL⁷⁷ data set (**Table 2.2**) outperformed models restricted to other cell types (**Extended Data Figure 2.9**). We further analyzed cell-type-level or tissue-level ABC evaluation data, defined by SNP-gene links identified in biosample(s) related to the focal cell type/tissue (ABC score > 0.2), irrespective of other biosamples. Models restricted to the focal cell type/tissue generally significantly outperformed models restricted to other cell types/tissues and the model spanning the union of all cell types (**Extended Data Figure 2.10**). Additional secondary analyses are reported in the **Supplementary Note** and **Supplementary Figures 1.40-1.44**.

We conclude that training pgBoost by restricting to features from a focal cell type generally maximizes power to detect SNP-gene links corresponding to that cell type in cell-type-specific and cell-type-level evaluation data—despite the inclusion of fine-mapped eSNP-eGene pairs from all GTEx tissues in our training data (see **Extended Data Figure 2.4**). However, training pgBoost using all cell types attained similar performance as training pgBoost in the focal cell type in some cell-type-specific and cell-type-level evaluation experiments.

pgBoost links variants to genes mediating GWAS associations

We have shown that SNP-gene links identified by pgBoost may not be identified by other methods (**Figure 2.3** and **Figure 2.4**). Below, we dissect four loci in which pgBoost links fine-mapped GWAS variants to experimentally validated or biologically plausible target genes that are not identified by other single-cell methods and/or the distance-based method. For simplicity, we defined SNP-gene links identified by a given method based on the top 5% of linking scores across all candidate links at distance >1kb. These examples were derived from fine-mapping results for 94 diseases/complex traits (see **Data Availability**).

First, pgBoost linked rs499883 (fine-mapped for apolipoprotein B^{107,112,114}, LDL cholesterol^{107,112,115}, and total cholesterol^{107,112,115}, each with PIP = 1.00) to *PCSK9* (intronic variant 14kb from TSS; closest TSS) (**Figure 2.6a**). This is considered a gold-standard SNP-gene link, based on an independent *PCSK9* GWAS coding variant that was fine-mapped for apolipoprotein B^{106,107}, an essential structural protein of LDL cholesterol. *PCSK9*, which encodes an enzyme involved in regulating LDL levels, has been implicated in several cardiovascular diseases⁴ and has been successfully targeted by a class of cholesterol-lowering drugs⁵. This link also scored in the top 5% of links predicted by the distance-based method, but scored outside the top 5% of links predicted by existing single-cell methods.

Second, pgBoost linked rs12722502 (fine-mapped for deep vein thrombosis with PIP = 0.93 and eosinophil count with PIP = 0.52^{107,112,114}) to *IL2RA* (11kb distance to TSS; closest TSS) (**Figure 2.6b**). rs12722502 lies in an intronic enhancer of *IL2RA* experimentally validated by

CRISPR activation¹¹⁶. *IL2RA* encodes a key component of the interleukin 2 signaling pathway critical for T cell proliferation and homeostasis¹¹⁷ and has been implicated in several autoimmune diseases^{118–120}. This link scored in the top 5% of links predicted by most methods (Signac, Cicero, ArchR, and the distance-based method). However, SCENT, ArchR, and Cicero each assigned highest priority to other genes (*RMB27*, *IL15RA*) and additionally prioritized other genes in the top 5% of links (*GDI2*, *ANKRD16*, *FBXO18*, *PFKFB3*).

Third, pgBoost linked rs12151289 (fine-mapped for monocyte count with PIP = 1.00, basophil count with PIP = 0.98, and mean corpuscular volume with PIP = 0.82^{107,112,114}) to *CEBPA* (42kb distance to TSS, 2nd closest TSS) (**Figure 2.6c**). rs12151289 lies in an enhancer of *CEBPA* experimentally validated by CRISPR¹²¹. *CEBPA* has been widely studied for its role in myeloid cell differentiation and proliferation, and somatic mutations in the *CEBPA* gene are associated with acute myeloid leukemia and other hematological malignancies^{122,123}. This link also scored in the top 5% of links predicted by Cicero, but scored outside of the top 5% of links predicted by the distance-based method and other existing methods.

Fourth, pgBoost linked rs28679793 (fine-mapped for mean corpuscular hemoglobin concentration with PIP = 0.98^{107,112,114}) to *SPTB* (56kb distance to TSS; 3rd closest TSS) (**Figure 2.6d**). *SPTB* encodes a protein critical for red blood cell membrane stability and has been implicated in disorders stemming from erythrocytic abnormalities, including spherocytosis and hemolytic anemia^{124–126}. Thus, while the link between rs28679793 and *SPTB* has not been experimentally tested (to our knowledge), *SPTB* is a biologically plausible target gene. This link also scored in the top 5% of links predicted by SCENT and Signac, but scored outside of the top 5% of links predicted by the distance-based method, ArchR, and Cicero.

Additional examples are discussed in the **Supplementary Note** and **Supplementary Figure 2.45**. These examples demonstrate that pgBoost can combine weak or conflicting evidence from constituent methods to link regulatory variants to experimentally validated or biologically plausible target genes.

Discussion

We developed a gradient boosting¹⁰⁵ framework (pgBoost) integrating single-cell peak-gene linking scores^{80–83} with genomic distance by training on fine-mapped eSNP-eGene pairs^{63,64} to predict the causal probability of candidate SNP-gene links. We demonstrated that pgBoost significantly outperforms existing single-cell methods and genomic distance in enrichment for SNP-gene links derived from eQTL^{63,64}, Activity-By-Contact (ABC)^{26,27}, CRISPRi^{47–55}, and GWAS^{106,107} data. In particular, pgBoost substantially outperformed existing methods in evaluations of longer-range links, which are of high biological importance (for example, non-coding variants in GWAS often do not regulate the closest gene^{23–25}) and are more difficult to capture using distance-based approaches. We further demonstrated that restricting pgBoost to features from a focal cell type improves power to detect regulatory links relevant to that cell type, supporting the future application of this framework to a wider variety of biological contexts. Finally, we applied pgBoost scores to fine-mapping results for 94 diseases/traits to nominate target genes mediating GWAS associations, highlighting several examples where pgBoost links a fine-mapped variant to an experimentally validated or biologically plausible target gene.

Our work has several implications for downstream analyses. First, we have publicly released pgBoost scores as a resource to aid in the functional interpretation of GWAS variant-disease associations or epigenomic assays. While we highlight links scoring in the top 5% by pgBoost score, we note that this is a stringent cutoff at larger distances (for example, only 0.3% of links >100kb are included in the top 5% of all links); thus, this threshold could plausibly be relaxed for longer-range links. Second, pgBoost scores nominate SNP-gene pairs with high causal probability for functional follow-up (e.g. CRISPRi or base editing⁵⁹) to validate predicted links and dissect underlying biological mechanisms. Third, we have publicly released open-source software enabling users to retrain pgBoost while incorporating features from additional scRNA/ATAC-seq multiome data sets (or additional single-cell methods). In settings where the goal is to predict SNP-gene links across a broad set of cell types, or in a cell type (or context/state)

that is not represented in multiome data, we recommend incorporating features from a larger number of multiome data sets (as this approach generally improves pgBoost performance; **Supplementary Figure 2.30** and **Supplementary Figure 2.31**). In settings where the goal is to predict SNP-gene links in a focal cell type, we recommend incorporating features from only the focal cell type (as this approach usually performs best; **Figure 2.5**). Fourth, the eQTL, ABC, CRISPR and GWAS evaluation sets of SNP-gene links that we have used here may prove useful in future studies.

Our work has several limitations and directions for future research. First, we primarily focused our analyses on blood/immune cell types (for which there currently exist multiple high-quality multiome data sets^{77,108–110}). Tissue-specific analyses using a brain data set¹¹³ yielded encouraging results for pgBoost performance in broader biological contexts (**Figure 2.5a**, right panel), but it will be important to apply pgBoost to other tissues/cell types as more multiome data becomes available. Second, a high pgBoost score (for example, in the top 5% of all links) does not imply biological causality, underscoring the importance of experimental validation. Third, pgBoost does not include as features enhancer-associated biochemical and epigenomic annotations; including these features in future extensions of pgBoost may further improve performance. Additional limitations are discussed in the **Supplementary Note**. Despite these limitations, we have shown that pgBoost improves power to link regulatory variants to genes and can aid the functional interpretation of GWAS variant-disease associations.

Acknowledgements

We thank Jesse Engreitz, Robin Andersson, Andreas Gschwind, and Weilin Qiu for valuable feedback on the manuscript and Soumya Raychaudhuri, Shamil Sunyaev, Vijay Sankaran, Saori Sakaue, Carles Boix, Ben Strober, Jordan Rossen, and Martin Zhang for helpful discussions. This research was funded by NIH grants U01 HG012009, R56 HG013083, R01 MH101244, R37 MH107649, R01 HG006399, and R01 MH115676 (A.L. Price). K. Dey is funded

by NIH/NHGRI under grant R00HG012203, NIH/NCI Cancer Center Support Grant P30CA008748, and the Josie Robertson Investigator Program.

Figures

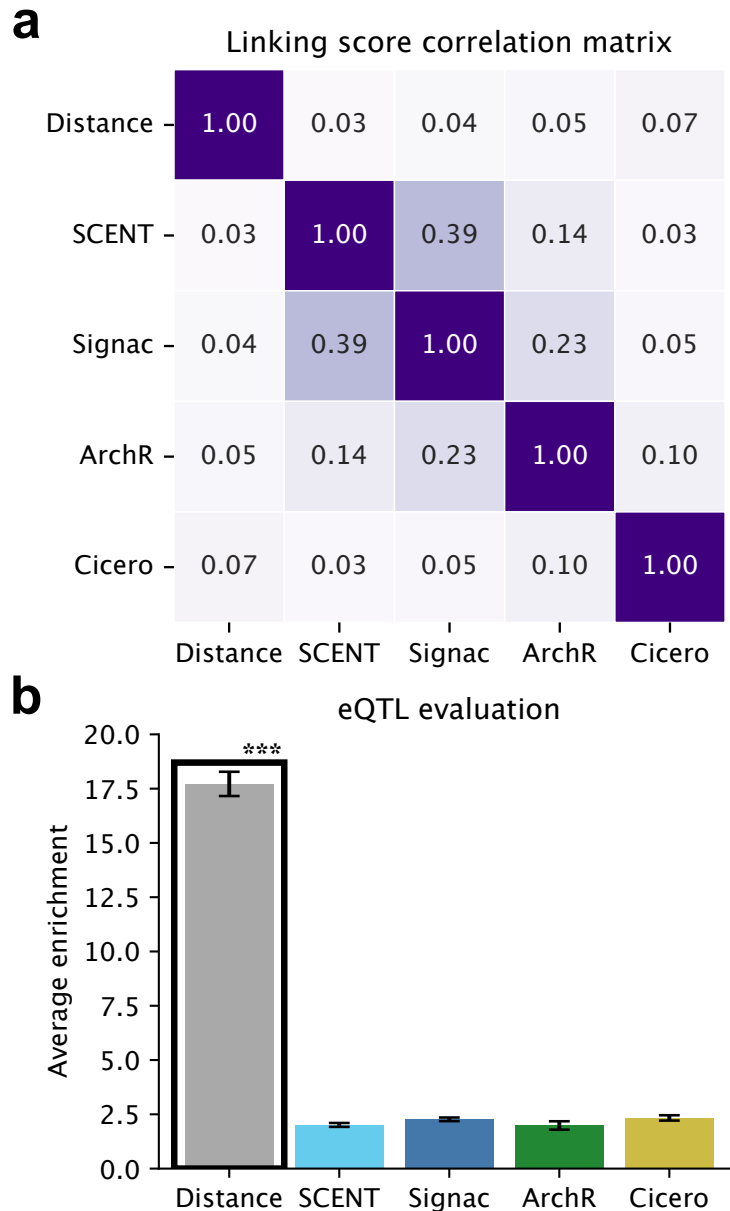


Figure 2.1. Existing single-cell peak-gene linking methods exhibit low concordance and underperform a distance-based linking score. a) Pearson correlation matrix of linking scores predicted by genomic distance from TSS and 4 single-cell peak-gene linking methods (**Table 2.1**) across 2,672,886 candidate SNP-gene links scored by any method in T cells from the 10x PBMC data set (**Table 2**). **b)** Average enrichment across recall values of links predicted by distance and 4 single-cell peak-gene linking methods (across 6,882,594 candidate SNP-gene links scored in 4 single-cell multiome data sets; **Table 2.2**) for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across all GTEx tissues (**Methods**). Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (***) of top method (denoted in black outline) vs. each other method (actual p-values < 2e-152 vs. each method) (**Methods**).

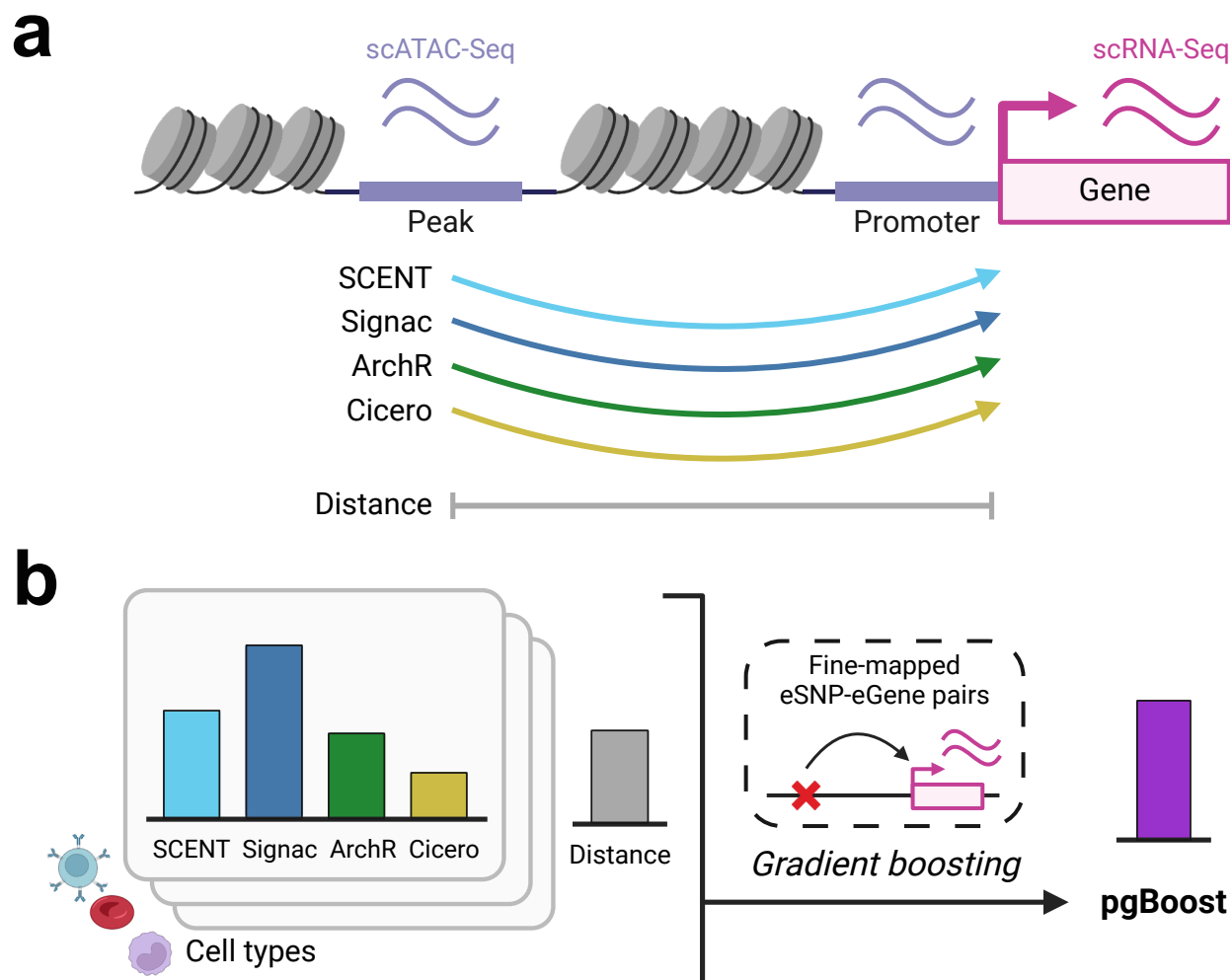


Figure 2.2. Overview of pgBoost method. **a)** Single-cell peak-gene linking methods assign linking scores measuring the co-activity of ATAC peak-gene pairs. **b)** pgBoost learns a non-linear combination of SNP-gene distance and linking scores across methods, cell types, and data sets to assign probabilistic linking scores using a gradient boosting classification framework and a training set of positive and negative links comprised of fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.2 and maximum PIP < 0.01 across all GTEx tissues, respectively. This figure was created with BioRender.com.

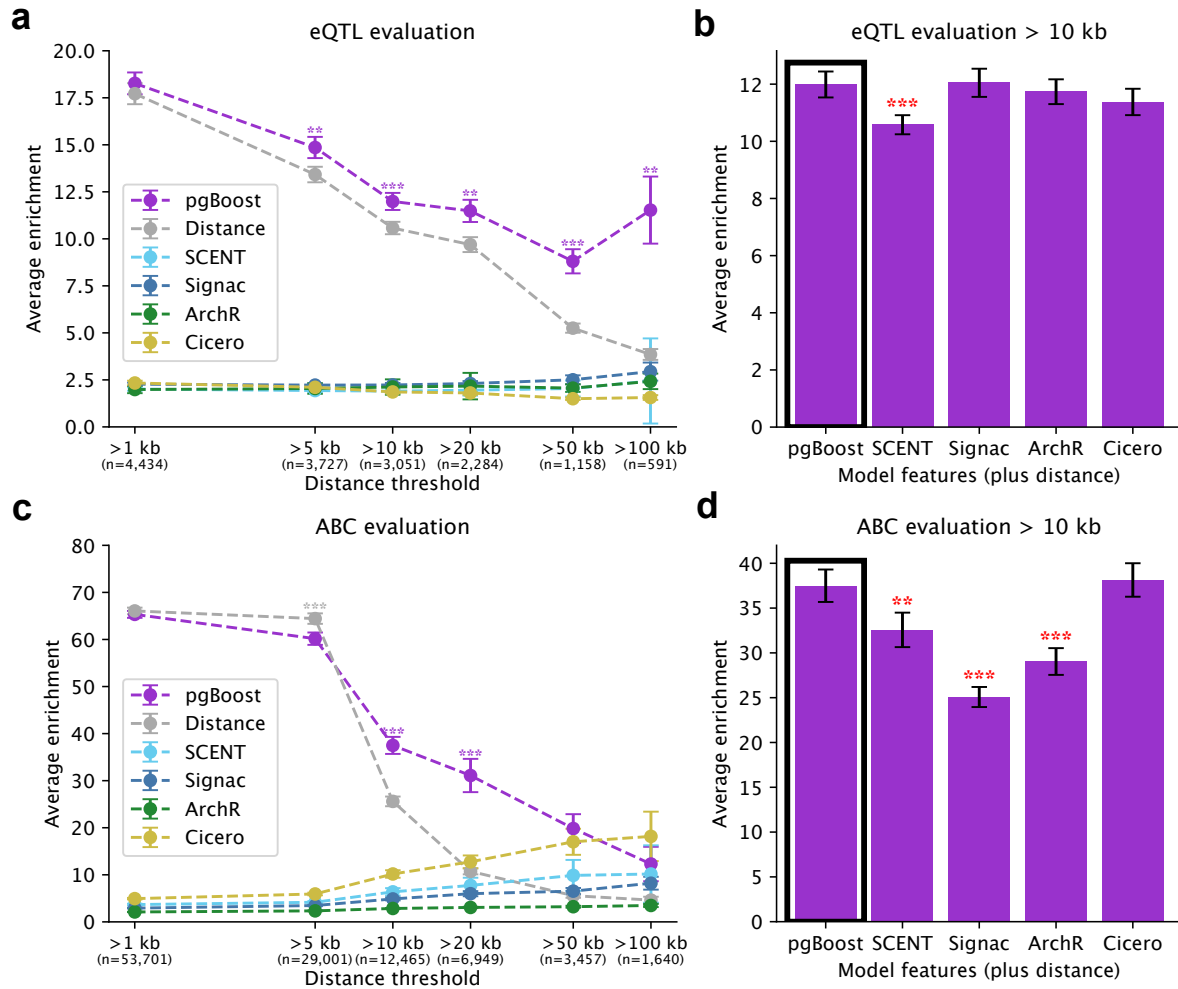


Figure 2.3. Performance of pgBoost and other methods on eQTL and ABC evaluation data sets. **a)** Average enrichment across recall values of links predicted by pgBoost, distance, and 4 existing methods for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, at various distance thresholds. **b)** Average enrichment across recall values of links predicted by pgBoost (denoted in black outline) vs. gradient boosting models restricted to linking scores from a single existing method plus distance for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, at >10kb. **c)** Average enrichment across recall values of links predicted by pgBoost, distance, and 4 existing methods for 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, at various distance thresholds. **d)** Average enrichment across recall values of links predicted by pgBoost (denoted in black outline) vs. gradient boosting models restricted to linking scores from a single existing method plus distance for 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, at >10kb. The number of positive evaluation links at each distance threshold is specified in parentheses in **a** and **c**. All enrichments are measured across 6,882,594 candidate SNP-gene links scored in 4 single-cell multiome data sets (**Table 1.2**). Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the top-performing method (in **a**, **c**) or pgBoost (in **b**, **d**) vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) (**Methods**). Red stars (in **b**, **d**) denote significant underperformance of the focal method vs. pgBoost.

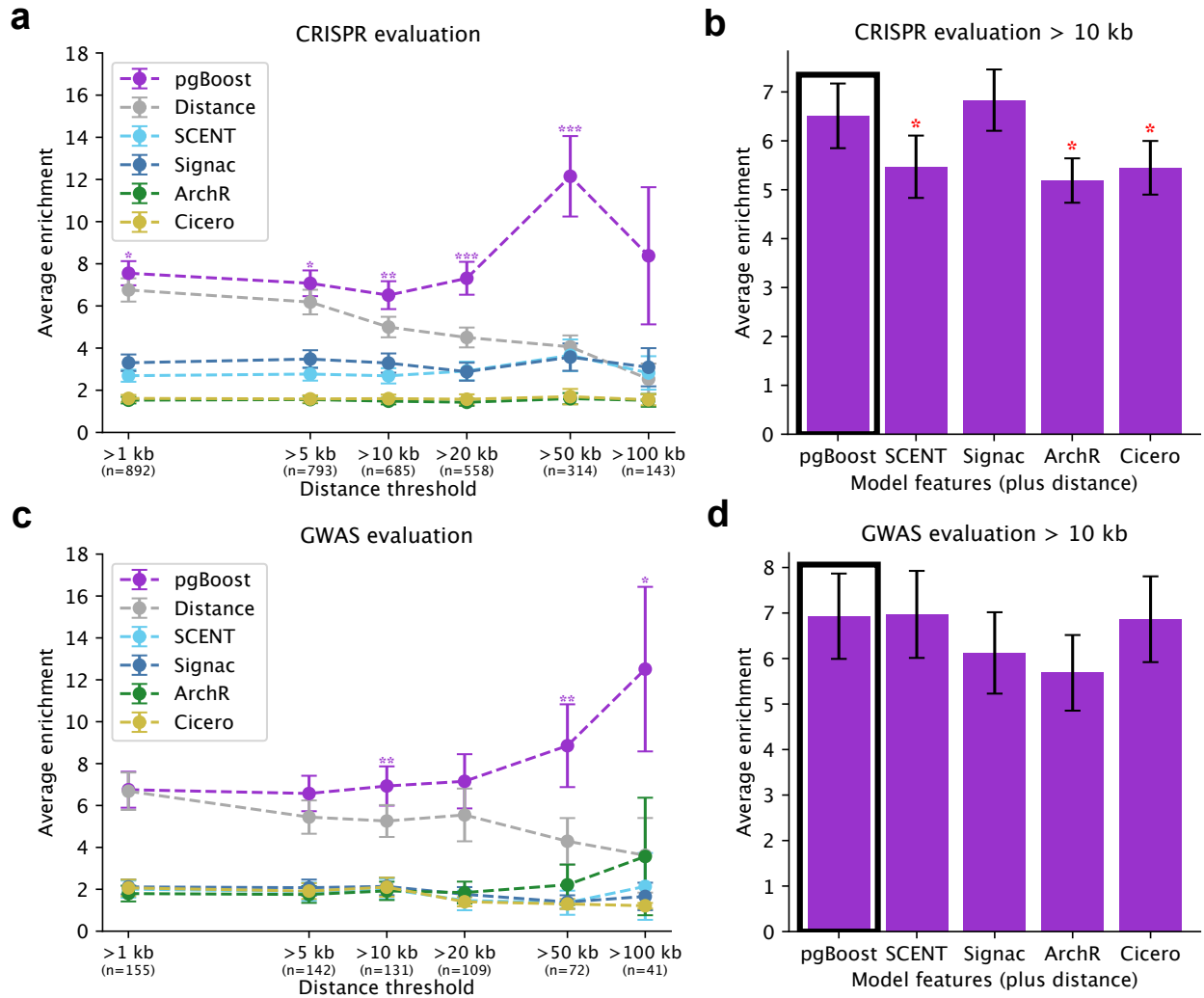


Figure 2.4. Performance of pgBoost and other methods on gold-standard evaluation data sets. **a)** Average enrichment across recall values of links predicted by pgBoost, distance, and 4 existing methods for 892 links validated by CRISPR, at various distance thresholds. **b)** Average enrichment across recall values of links predicted by pgBoost (denoted in black outline) vs. gradient boosting models restricted to linking scores from a single existing method plus distance for 892 links validated by CRISPR, at >10kb. **c)** Average enrichment across recall values of links predicted by pgBoost, distance, and 4 existing methods for 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. **d)** Average enrichment across recall values of links predicted by pgBoost (denoted in black outline) vs. gradient boosting models restricted to linking scores from a single existing method plus distance for 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at >10kb. The number of positive evaluation links at each distance threshold is specified in parentheses in **a** and **c**. All enrichments are measured across 6,882,594 candidate SNP-gene links scored in 4 single-cell multiome data sets (**Table 2.2**). Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the top-performing method (in **a**, **c**) or pgBoost (in **b**, **d**) vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) (**Methods**). Red stars (in **b**, **d**) denote significant underperformance of the focal method vs. pgBoost.

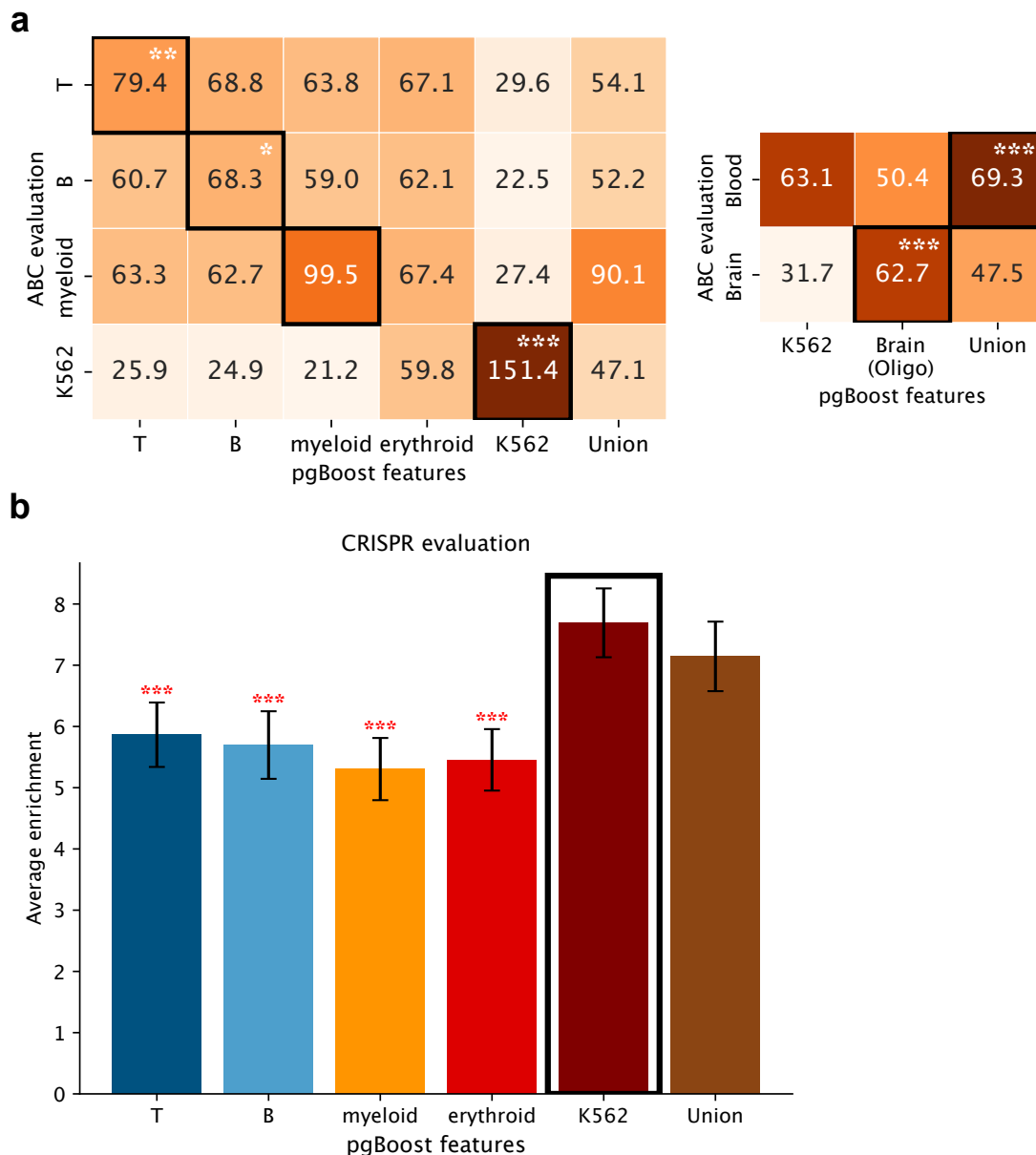


Figure 2.5. Performance of pgBoost restricted to features from a focal or non-focal cell type in evaluation data sets relevant to the focal cell type. a) Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (left) or tissue (right) for training and prediction (columns; Union denotes all features) for cell type/tissue-specific ABC evaluation links (rows). Number of evaluation links in each ABC evaluation set: $n_{\text{blood}} = 10,845$, $n_{\text{brain}} = 6,417$, $n_{\text{T}} = 5,855$, $n_{\text{B}} = 4,404$, $n_{\text{myeloid}} = 2,680$, $n_{\text{K562}} = 1,107$. Number of biosamples in each ABC evaluation set: $N_{\text{blood}} = 74$, $N_{\text{brain}} = 26$, $N_{\text{T}} = 53$, $N_{\text{B}} = 14$, $N_{\text{myeloid}} = 5$, $N_{\text{K562}} = 1$. **b)** Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (columns; Union denotes all features) for cell-type-level CRISPR evaluation links in K562 (number of evaluation links: $n = 892$). Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method (denoted in black outline; defined separately for each row in **a**) vs. each other method (**Methods**). In **b**, red stars denote significant underperformance of the focal method vs. the top method.

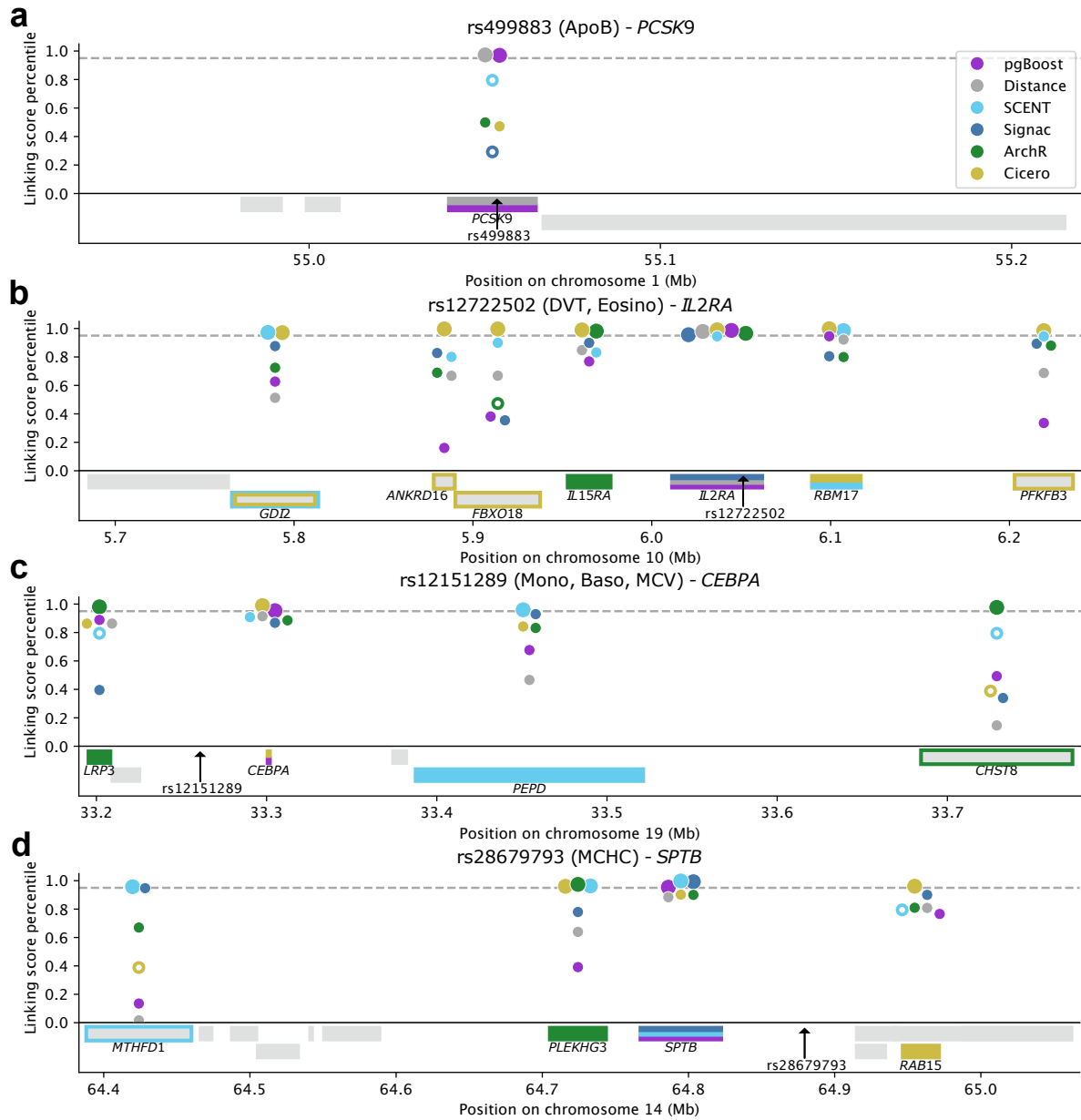


Figure 2.6. Examples of SNP-gene links identified by pgBoost. We report links prioritized by pgBoost, distance, and 4 existing methods for (a) rs499883, (b) rs12722502, (c) rs12151289, and (d) rs28679793. Figure panel headings for each SNP report the corresponding GWAS trait(s) (in parentheses) and the SNP-gene link attaining the highest pgBoost score. Linking scores for each method are reported as dots in the middle of each gene (jittered for easier visualization). Large dots denote links scored in the top 5% by a given method. Hollow dots denote links that were not scored by a given method (with their percentile defined as the percentage of links not scored by the method). Genes scored in the top 5% by any method are reported at the bottom of each panel (with gene name); nearby genes are also displayed (without gene name). Colored boxes denote the gene attaining the highest linking score percentile > 95% by one or more method(s); colored borders denote other genes scored in the top 5% by a given method. ApoB: apolipoprotein B, DVT: deep vein thrombosis, Eosino: eosinophil count, Mono: monocyte count, Baso: basophil count, MCV: mean corpuscular volume, MCHC: mean corpuscular hemoglobin concentration.

Tables

Method	Input data	Scoring procedure	Significance testing	Candidate link criteria
SCENT ⁸⁰	RNA + ATAC	Poisson regression effect size	Bootstrapping across cells	Peak & gene in $\geq 5\%$ of cells
Signac ^{77,81}	RNA + ATAC	Correlation across single cells	Permutation across peaks	Peak & gene in ≥ 10 cells
ArchR ⁸²	RNA + ATAC	Correlation across metacells	Analytical p-value	N/A
Cicero ⁸³	ATAC	Peak-promoter co-accessibility across metacells	N/A	Promoter overlaps a peak

Table 2.1. Overview of 4 single-cell peak-gene linking methods. For each method linking ATAC-seq peaks to genes, we indicate the input data (either scRNA/ATAC-seq multiome data or scATAC-seq data), scoring procedure, significance testing approach (if applicable), and candidate link selection criteria. Further details are provided in **Methods**.

Data set	Cell types	# cells	Experimental protocol	# pgBoost features
10x PBMC ¹⁰⁸	T, B, myeloid	10,561	10x Multiome	9
Luecken BMMC ¹⁰⁹	T, B, myeloid, erythroid, dendritic	57,371	10x Multiome	15
SHARE-seq LCL ⁷⁷	lymphoblastoid cell line (LCL)	17,644	SHARE-Seq	3
Xu ¹¹⁰	K562	7,861	10x Multiome	3
Genomic distance	N/A	N/A	N/A	2
<i>All</i>	<i>7 cell types</i>	<i>93,437 cells</i>		<i>32 features</i>

Table 2.2. Overview of pgBoost features. We indicate the cell types profiled, number of cells passing QC, experimental protocol, and number of pgBoost features derived from each single-cell multiome data set (plus 2 distance-based features). 10x Multiome: 10x Genomics Single Cell Multiome ATAC + Gene Expression.

Chapter 3: Distinguishing causal from tagging enhancers using single-cell multiome data

Elizabeth Dorans, Alkes L. Price

Introduction

Mapping gene regulatory architectures is critical for the functional interpretation of GWAS discoveries and understanding the circuitry driving key developmental and disease processes^{1,2,6,10,13,15,18,19,26,36}. ‘Peak-gene linking methods,’ which predict regulatory relationships from single-cell RNA-seq+ATAC-seq multiome data by correlating the accessibility of peaks of open chromatin in ATAC-seq data with genes in RNA-seq data, have shown particular promise in linking regulatory elements to their target genes^{77,80–92}. Previously proposed peak-gene linking methods have attempted to account for false positive links induced by technical and biological confounders^{77,80,81,84,88} and gene co-expression⁸⁷. Other studies have mapped peak co-accessibility^{82,83,93} or inferred ‘modules’ of co-accessible peaks^{85,94} from single-cell ATAC-seq data, raising the question of how peak co-accessibility might impact peak-gene linking.

We hypothesized that correlations among ATAC-seq peaks may induce non-causal *tagging* peak-gene links, analogous to tagging associations in GWAS (**Figure 3.1**); in particular, we hypothesized that these correlations are induced by the binding of shared transcription factors (TFs) at multiple ATAC peaks, with patterns of co-accessibility reflecting dynamic cell types / states. Here, we show that tagging effects, induced by peak co-accessibility, are pervasive in peak-gene linking. We first define two scores for each ATAC-seq peak: *co-accessibility score*, quantifying peak-peak correlations, and *co-activity score*, quantifying peak-gene correlations. We investigate peak tagging by analyzing CRISPR data and by regressing co-activity scores on stratified co-accessibility scores (S-CASC) to disentangle causal from tagging effects and pinpoint

categories of peaks harboring causal peak-gene links. We then assess TF binding activity as a potential biological mechanism underlying peak co-accessibility. We further investigate tagging peak-gene associations induced by gene-gene co-expression. Finally, we explore statistical fine-mapping¹¹⁴ as a method to distinguish causal from tagging peak-gene links, and we highlight several examples. Understanding the correlation structure among peaks and genes in single-cell multiome data is critical for accurate mapping of regulatory links and functional interpretation of GWAS discoveries.

Results

Overview of methods

To quantify peak-peak and peak-gene correlations for each peak in single-cell multiome data, we define two scores for each peak: *co-accessibility score* (quantifying peak-peak correlations) and *co-activity score* (quantifying peak-gene correlations), each reflecting a sum of squared correlations. We assess the relationship between co-accessibility scores and co-activity scores across peaks. We propose a method, stratified co-accessibility score regression (S-CASC), to identify functional peak categories that are enriched for causal peak-gene associations. We note that co-accessibility scores can be viewed as an analogue of SNP LD scores¹²⁷, co-activity scores can be viewed as an analogue of GWAS associations, and S-CASC can be viewed as an analogue of stratified LD score regression¹⁸.

In detail, we define the *co-accessibility score* of peak j as $x_j = \sum_{k \in K_j} r_{jk}^2$, where K_j is the set of all *cis* peaks located <1Mb from focal peak j (including peak j itself) and r_{jk}^2 are squared peak-peak correlations (from scATAC-seq data) across metacells⁸². We define the *co-activity score* of peak j as $y_j = \sum_{g \in G_j} r_{jg}^2$, where G_j is the set of all *cis* genes with TSS <1Mb from focal peak j and r_{jg}^2 are squared peak-gene correlations (from scATAC-seq and scRNA-seq data) across metacells. In all analyses, we exclude peaks with distance <1kb to any TSS^{90,128} (thus focusing primarily on enhancers rather than promoters), and we correct each r^2 term for upward

bias¹²⁷ (**Methods**). We assess whether peaks with higher co-accessibility scores tend to have higher co-activity scores, as would be expected if peak-gene correlations arise in part due to peak-peak tagging effects.

S-CASC determines that a category of peaks is enriched for causal peak-gene associations if peaks with high co-accessibility to that category have higher co-activity scores than peaks with low co-accessibility to that category. The S-CASC regression relies on the following model:

$$E[y_j] = \tau_0 + \sum_C (\tau_C x_{j,C}) + \gamma |G_j|, \quad [2.1]$$

where y is the co-activity score of peak j , τ_0 is an intercept term, $x_{j,C}$ is the co-accessibility score of peak j with respect to peak category C (defined as $\sum_{k \in (K_j \cap C)} r_{jk}^2$), τ_C is the per-peak contribution to causal peak-gene effect variance for peak category C , and the covariate $|G_j|$ is the number of genes with TSS <1Mb from peak j . We define the enrichment of a peak category as the proportion of causal peak-gene effect size variance from peaks in that category divided by the proportion of peaks in that category. We estimate standard errors by jackknifing across chromosomes; we then meta-analyze enrichments for each peak category across data set-cell type pairs (**Methods**). Due to peak measurement noise, we focus on the enrichment of a peak category, and not on total causal peak-gene effect size variance. We have released open-source software implementing S-CASC.

Although our primary focus is on peak-peak tagging effects, we also explore gene-gene tagging effects by defining the *gene co-expression score* of gene i as $x_i = \sum_{g \in G_i} r_{ig}^2$, where G_i is the set of all genes with TSS <1Mb from the TSS of focal gene g (including gene g itself) and r_{ig}^2 are squared gene-gene correlations across metacells; defining the *gene co-activity score* of gene i as $y_i = \sum_{k \in K_i} r_{ik}^2$, where K_i is the set of all peaks located <1Mb from the TSS of focal gene i

and r_{ik}^2 are squared gene-peak correlations across metacells; and identifying gene categories enriched for causal gene-peak associations using stratified co-expression score regression (**Methods**).

We analyzed 4 scRNA+ATAC-seq multiome data sets (Xu K562¹¹⁰, Satpathy K562¹²⁹, SHARE-seq LCL⁷⁷, and Luecken BMMC¹⁰⁹) that each spanned 1-4 blood or immune cell types (total of 7 data set-cell type pairs, spanning 6 cell types and 86k cells after QC) (**Table 3.1** and **Methods**). We have publicly released peak co-accessibility and co-activity scores and gene co-expression and co-activity scores for each data set-cell type pair analyzed in this study.

We assessed the relationship between co-accessibility scores and co-activity scores across peaks, for each of the 7 data set-cell type pairs (**Table 3.1**). Results for the Xu-K562¹¹⁰ data set are reported in **Figure 3.2a**, and results for all other data set-cell type pairs are reported in **Supplementary Figure 3.1**. Co-accessibility scores and co-activity scores were strongly correlated ($r = 0.57-0.73$ in each data set-cell type pair), and these correlations remained strong after conditioning on the number of genes with TSS <1Mb from the focal peak ($r = 0.48-0.69$ in each data set-cell type pair). This result is potentially consistent with the hypothesis that co-accessibility between peaks induces tagging peak-gene associations (“tagging hypothesis”), but it is important to rule out other explanations (see below). Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.2**).

Co-accessibility between peaks induces tagging peak-gene associations

To evaluate the tagging hypothesis, we analyzed CRISPRi-tested peak-gene pairs^{47–58} in K562 cells, restricting to genes with both an experimentally validated causal (CRISPR-positive) peak that is identified by ArchR⁸² (with peak-gene correlation > 0.45) and at least one experimentally tested non-causal (CRISPR-negative) peak overlapping our single-cell data. We restricted this analysis to the 3 data set-cell type pairs most relevant to K562 cells: Xu-K562¹¹⁰, Satpathy-K562¹²⁹, and Luecken-erythroid¹⁰⁹ (**Table 3.1**). We measured the correlation (across metacells) of each CRISPR-negative peak with both the CRISPR-positive peak and the focal

gene. Correlation with the CRISPR-positive peak was strongly correlated to correlation with the focal gene ($r = 0.92$; **Figure 3.2b**), corroborating the tagging hypothesis. Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.3**).

To further evaluate this hypothesis, we computed the enrichment of causal peak-gene effect size variance for 19 functional categories of peaks using S-CASC; the 19 functional categories of peaks included gene proximity, histone modifications, constraint, and other genic features. We observed significant enrichment of causal effect size variance for peak categories defined by gene proximity (closest to TSS, $TSS \pm 10\text{kb}$, $TSS \pm 100\text{kb}$), activating histone modifications (H3K4me3, H3K9ac, H3K4me2, H3K27ac, H3K4me1, H3K36me3), constraint, and other genic features (3' UTR, 5' UTR, coding) (ranging from 1.09 ± 0.014 to 2.91 ± 0.67), and significant depletion for the repressive histone modification H3K27me3 (0.67 ± 0.11) (**Figure 3.2c**, **Supplementary Figure 3.4**). These findings provide additional support for the tagging hypothesis, as alternative explanations for the correlation between co-accessibility scores and co-activity scores (see below) would not be expected to produce these functional enrichments. Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.5**). Analyses excluding the covariate for the number of genes <1Mb from each peak also produced similar findings. (**Supplementary Figure 3.6**).

We restricted most of our analyses to peaks/genes located <1Mb from the focal peak, but we also assessed the relationship (across focal peaks) between genome-wide co-accessibility scores and genome-wide co-activity scores, defined without this restriction (**Methods**). Genome-wide co-accessibility scores and genome-wide co-activity scores were very strongly correlated (**Figure 3.2d** and **Supplementary Figure 3.7**; $r = 0.94-0.99$ in each data set-cell type pair). Genome-wide co-accessibility scores were strongly correlated with co-accessibility scores restricted to <1Mb from focal peak ($r = 0.73-0.86$), and genome-wide co-activity scores were strongly correlated with co-activity scores restricted to <1Mb from focal peak ($r = 0.58-0.67$).

Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.8**). This is consistent with peak tagging being driven by variation across cells in the abundance of particular transcription factors (TF) that bind at multiple sites across the genome (see below). This could involve the same copy of a TF binding to multiple nearby sites (*cis* TF effect), or different copies of the same TF binding to distal sites (*trans* TF effect), or a combination of *cis* and *trans* effects. The genome-wide correlation between co-accessibility and co-activity scores (**Figure 3.2d**) specifically implicates *trans* TF effects. On the other hand, our observation of higher peak-peak r^2 at shorter genomic distances (**Supplementary Figure 3.9**) specifically implicates *cis* TF effects. We conclude that co-accessibility is driven by a combination of *cis* TF effects and *trans* TF effects.

We performed 5 secondary analyses to rule out alternative explanations for the correlation between co-accessibility scores and co-activity scores. First, to further assess the effects of biological causality, we performed a bivariate regression of co-activity score on co-accessibility score and a binary indicator of CRISPR-positive status (restricting to genes with CRISPR-positive and CRISPR-negative peaks), as well as univariate regressions of co-activity score on each predictor individually. We determined that co-accessibility score alone explained a much higher proportion of the observed variation in co-activity scores, and that the bivariate model did not substantially increase the proportion of variation explained (**Supplementary Figure 3.10**). Second, to assess peak read depth effects, we performed a bivariate regression of co-activity score on co-accessibility score and peak read depth, as well as univariate regressions of co-activity score on each predictor individually. We determined that co-accessibility score alone tended to explain a much higher proportion of the variance in co-activity scores ($r^2=0.34-0.58$) vs. peak read depth alone ($r^2=0.0013-0.15$), and that the bivariate model did not substantially increase the proportion of variation explained ($r^2=0.34-0.61$; **Supplementary Figure 3.11**); we note that peak read depth explained $r^2=0.00046-0.12$ of the variance in co-accessibility scores. Third, to assess cell subtype effects, we computed co-accessibility scores and co-activity scores

using peak-peak and peak-gene correlations conditional on cell subtype proportions within each metacell of the Luecken¹⁰⁹ T cell data set (for which major cell subtype annotations are available^{102,109}). We determined that the correlation between co-accessibility scores and co-activity scores decreased but remained substantial ($r = 0.51$, vs. $r = 0.62$ for unconditioned correlation) (**Supplementary Figure 3.12**). This suggests that, although peak tagging is related to cell subtype / cell state-specific processes (see **Discussion**), major cell subtypes based on available annotations do not explain the bulk of the correlation between co-accessibility and co-activity. Fourth, to assess peak measurement noise effects, we introduced additional artificial noise for a subset of ATAC peaks and recomputed the correlation between co-accessibility scores and co-activity scores. We determined that adding noise did not substantially increase the correlation (**Supplementary Figure 3.13**). Fifth, we assessed the correlation between co-accessibility scores and co-activity scores when conditioning on the number of peaks located <1Mb from the focal peak (in addition to conditioning on the number of nearby genes; see above). We determined that the correlation remained strong (**Supplementary Figure 3.14**). We thus conclude that the correlation between co-accessibility scores and co-activity scores cannot be explained by biological causality (also see **Figure 3.2b**), peak read depth, major cell subtypes, varying peak measurement noise, or the number of nearby peaks.

We performed 3 additional secondary analyses. First, we estimated the proportion of gene expression variance explained by measured *cis*-peak accessibility using co-accessibility score regression (without stratifying by functional category). Our results suggested that the proportion of gene expression variance explained by measured *cis*-peak accessibility may exceed 0.5, but we caution that our estimates have limited interpretability (**Supplementary Figure 3.15**). Second, we assessed how much peak-peak correlation patterns vary across cell types by computing correlations across peaks between co-accessibility scores from each data set-cell type pair. We determined that co-accessibility scores are partially concordant across cell types ($r = 0.08$ - 0.35) and show stronger concordance in more similar cell types (**Supplementary Figure 3.16**),

providing additional evidence that peak tagging is related to cell subtype / cell state-specific processes (see **Discussion**). Third, we recomputed standard errors and p-values for enrichments of peak categories using 200 jackknife blocks defined via *k*-means clustering on ATAC data, instead of via chromosomes. We determined that standard errors and p-values were similar using these two approaches (**Supplementary Table 3.6**).

These findings demonstrate the pervasive effects of peak-peak tagging on peak-gene associations.

Co-accessibility between peaks is driven by transcription factor binding activity

We hypothesized that TF binding at multiple sites (including sites that do not impact gene expression in the focal cell type) could induce tagging correlations between pairs of ATAC-seq peaks—either via the same copy of a TF binding to multiple nearby sites (*cis* TF effect), or via different copies of the same TF binding to distal sites (*trans* TF effect), or via a combination of *cis* and *trans* effects—providing a mechanistic explanation for the relationship between peak tagging and cell subtypes / cell states (see **Discussion**). Under this hypothesis, *pioneer TFs*^{130,131} (a class of TFs with the ability to bind and activate closed chromatin) would be particularly powerful in driving tagging correlations. To evaluate the TF binding hypothesis, we analyzed transcription factor binding sites (TFBS) detected by ChIP-seq for 303 TFs in K562 cells and 148 TFs in GM12878 cells, including 9 known pioneer TFs in K562 and 8 known pioneer TFs in GM12878^{130,131} (**Supplementary Table 3.7** and **Methods**). For each ATAC-seq peak, we counted the number of TFs that bind within the peak. For each pair of ATAC-seq peaks, we counted the number of TFs that bind within both peaks. For each peak-gene pair, we counted the number of TFs that bind within both the peak and the promoter of the gene (defined as TSS $\pm$ 1kb¹²⁸).

We computed the correlation across peaks between the number of TFs that bind within the peak and the co-accessibility score. Results for the Xu-K562¹¹⁰ data set are reported in **Figure 3.3a**, left panel, and results for the Satpathy-K562¹²⁹ and SHARE-seq-LCL⁷⁷ data sets are reported in **Supplementary Figure 3.17a**, left panel and **Supplementary Figure 3.18a**, left

panel. The number of TFs binding within a peak was moderately correlated to the co-accessibility score ($r = 0.18-0.35$ in each data set-cell type pair). To compare the effects of pioneer and non-pioneer TFs, we performed a bivariate linear regression of the co-accessibility score on the number of pioneer TFs that bind within a peak and the number of non-pioneer TFs that bind within a peak. We determined that the effect size per pioneer TFs was much larger than the effect size per non-pioneer TF (0.068 ± 0.028 vs. 0.012 ± 0.0047 , p-value for difference = 0.028) (**Figure 3.3a**, right panel, **Supplementary Figure 3.17a**, right panel, **Supplementary Figure 3.18a**, right panel, and **Supplementary Figure 3.19a**, right panel). These results are consistent with peak co-accessibility being driven by TF binding activity (particularly of pioneer TFs). Analyses using peak-peak correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.20a**).

We computed the correlation across peak-peak pairs between the number of TFs that bind within both peaks and the squared peak-peak correlation (across cells in scATAC-seq data). Results for the Xu-K562¹¹⁰ data set are reported in **Figure 3.3b**, left panel, and results for the Satpathy-K562¹²⁹ and SHARE-seq-LCL⁷⁷ data sets are reported in **Supplementary Figure 3.17b**, left panel and **Supplementary Figure 3.18b**, left panel. The number of TFs that bind within both peaks was moderately correlated to the squared peak-peak correlation ($r = 0.12-0.21$ in each data set-cell type pair). To compare the effects of pioneer and non-pioneer TFs, we performed a bivariate linear regression of squared peak-peak correlation on the number of pioneer TFs and the number of non-pioneer TFs that bind within both peaks. We determined that the effect size per pioneer TF was much larger than the effect size per non-pioneer TF (0.0025 ± 0.00062 vs. 0.00046 ± 0.00016 , p-value for difference = $2e-5$) (**Figure 3.3b**, right panel, **Supplementary Figure 3.17b**, right panel, **Supplementary Figure 3.18b**, right panel, and **Supplementary Figure 3.19b**, right panel). These results further corroborate the hypothesis that peak co-accessibility is driven by TF binding activity (particularly of pioneer TFs). Analyses using peak-

peak correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.20b**).

We computed the correlation across peaks between the number of TFs that bind within the peak and the co-activity score. Results for the Xu-K562¹¹⁰ data set are reported in **Figure 3.3c**, left panel, and results for the Satpathy-K562¹²⁹ and SHARE-seq LCL⁷⁷ data sets are reported in **Supplementary Figure 3.17c**, left panel and **Supplementary Figure 3.18c**, left panel. The number of TFs binding within a peak was moderately correlated to the co-activity score ($r = 0.14-0.25$ in each data set-cell type pair). To compare the effects of pioneer and non-pioneer TFs, we performed a bivariate linear regression of co-activity score on the number of pioneer TFs that bind within a peak and the number of non-pioneer TFs that bind within a peak. We determined that the point estimate of the effect size per pioneer TF was larger than the point estimate of the effect size per non-pioneer TF, although the difference was not statistically significant due to noisy estimates (p-value for difference = 0.29) (**Figure 3.3c**, right panel, **Supplementary Figure 3.17c**, right panel, **Supplementary Figure 3.18c**, right panel, and **Supplementary Figure 3.19c**, right panel). These results are consistent with peak-gene co-activity being impacted by TF-driven peak co-accessibility. Analyses using peak-gene correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.20c**).

Finally, we computed the correlation across peak-gene pairs between the number of TFs that bind within both the peak and the promoter of the gene and the squared peak-gene correlation (across cells in scRNA-seq and ATAC-seq data). Results for the Xu-K562¹¹⁰ data set are reported in **Figure 3.3d**, left panel, and results for the Satpathy-K562¹²⁹ and SHARE-seq-LCL⁷⁷ data sets are reported in **Supplementary Figure 3.17d**, left panel and **Supplementary Figure 3.17d**, left panel. The number of TFs binding within both the peak and the promoter of the gene was somewhat correlated to the squared peak-gene correlation ($r = 0.03-0.09$ in each data set-cell type pair). To compare the effects of pioneer and non-pioneer TFs, we performed a bivariate linear regression of squared peak-gene correlation on the number of pioneer TFs and the number

of non-pioneer TFs that bind within both the peak and promoter of the gene. We determined that the effect size per pioneer TF was much larger than the effect size per non-pioneer TF (0.0016 ± 0.00069 vs. $8.4e-5 \pm 4.4e-5$, p-value for difference = 0.019) (**Figure 3.3d**, right panel, **Supplementary Figure 3.17d**, right panel, **Supplementary Figure 3.18d**, right panel, and **Supplementary Figure 3.19d**, right panel). These results further corroborate the hypothesis that peak-gene co-activity is impacted by TF-driven peak co-accessibility. Analyses using peak-gene correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.20d**).

We repeated the above analyses (**Figure 3.3a-d**) using genome-wide peak-peak and peak-gene pairs (and genome-wide co-accessibility and co-activity scores), instead of restricting to pairs with distance <1Mb. Correlations across peaks between the number of TFs that bind within a peak and either the co-accessibility score (**Figure 3.3a**) or the co-activity score (**Figure 3.3c**) became stronger when using genome-wide scores, with a larger and more significant difference for pioneer TFs vs. non-pioneer TFs (**Supplementary Figure 3.21a,c**). Correlations across peak-peak pairs (resp. peak-gene pairs) between the number of TFs that bind within both peaks (resp. the peak and the gene's promoter) and the squared peak-peak correlation (**Figure 3.3b**) (resp. squared peak-gene correlation (**Figure 3.3d**)) were similar when using genome-wide pairs, with similar differences between pioneer TFs vs. non-pioneer TFs (**Supplementary Figure 3.21b,d**). These results further corroborate the *trans* TF binding mechanism of peak-peak tagging; as noted above, our observation of higher peak-peak r^2 at shorter genomic distances (**Supplementary Figure 3.9**) also supports the *cis* TF binding mechanism.

We performed 3 secondary analyses. First, we repeated the bivariate regression of the co-accessibility score on the number of pioneer TFs that bind within a peak and the number of non-pioneer TFs that bind within a peak (**Figure 3.3a**, right panel) while adding covariates that bin TFs that bind each peak by their average binding and expression. We determined that the effect size per pioneer TF was still much larger than the effect size per non-pioneer TF

(**Supplementary Figure 3.22**). Second, we repeated each bivariate regression comparing pioneer vs. non-pioneer TF effects (right panels of **Figure 3.3a-d**) using Perturb-multiome data to define pioneer TFs (TFs whose perturbation is associated with a significant change in peak accessibility or gene expression in scRNA+ATAC-seq multiome data¹³²), and drew similar conclusions (**Supplementary Figure 3.23**). Third, we repeated the above analyses (**Figure 3.3a-d**) using sequence motif data¹³³ instead of ChIP-seq data to quantify TF activity, and reached similar conclusions (**Supplementary Figures 2.24-2.28**).

These analyses suggest that TF binding (particularly of pioneer TFs) at multiple sites may drive peak-peak tagging.

Co-expression between genes induces tagging peak-gene associations

We shifted our focus to gene-gene tagging, a potential alternative source of non-causal peak-gene correlations, using gene co-expression scores and gene co-activity scores (defined above). We assessed the relationship between gene co-expression scores and gene co-activity scores across genes, for each of the 7 data set-cell type pairs (**Table 3.1**). Results for the Xu-K562¹¹⁰ data set are reported in **Figure 3.4a**, and results for all other data set-cell type pairs are reported in **Supplementary Figure 3.29**. Gene co-expression scores and gene co-activity scores were strongly correlated ($r = 0.25-0.76$ in each data set-cell type pair), and these correlations remained strong after conditioning on the number of peaks <1Mb from the focal gene ($r = 0.26-0.65$ in each data set-cell type pair). This result is potentially consistent with the hypothesis that co-expression between genes also induces tagging peak-gene associations (“gene tagging hypothesis”), in addition to tagging associations induced by co-accessibility between peaks. Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.30**).

To evaluate the gene tagging hypothesis, we analyzed CRISPRi-tested peak-gene pairs^{47–58} in K562 cells, analyzing multiple genes per peak (instead of multiple peaks per gene; **Figure 3.2b**). We restricted to peaks with both an experimentally validated target gene that is

identified by ArchR⁸² and at least one experimentally tested non-target gene overlapping our single-cell data, restricting to the same 3 data set-cell type pairs most relevant to K562 cells (**Table 3.1**). We measured the correlation (across metacells) of each CRISPR-negative gene with both the CRISPR-positive gene and the focal peak. Correlation with the CRISPR-positive gene was strongly correlated to correlation with the focal peak (**Figure 3.4b**), corroborating the gene tagging hypothesis. Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.31**).

To further evaluate the gene tagging hypothesis, we computed the enrichment of causal peak-gene effect size variance for 18 functional categories of genes by modeling gene co-activity score as a function of co-expression scores stratified by functional gene category, analogous to S-CASC (**Methods**); the 18 functional categories of genes included various metrics for gene essentiality and disease relevance. We observed significant enrichment of causal effect size variance for gene categories defined by gene conservation (high pLI, high pHI) and proximity to peaks (peak $\pm$ 10kb) (ranging from 1.06 $\pm$ 0.021 to 1.20 $\pm$ 0.047), and significant depletion for genes with many nearby SNPs (e.g. most independent SNPs within 10kb, 0.76 $\pm$ 0.090) (**Figure 3.4c** and **Supplementary Figure 3.32**). These findings provide additional support for the gene tagging hypothesis. Analyses using correlations across single cells instead of metacells were generally underpowered (**Supplementary Figure 3.33**). Analyses excluding the covariate for the number of peaks <1Mb from each gene also produced similar findings (**Supplementary Figure 3.34**).

We also assessed the relationship (across focal genes) between genome-wide gene co-expression scores and genome-wide gene co-activity scores, defined without restricting to gene-gene or peak-gene pairs with distance <1Mb. Genome-wide gene co-expression scores and genome-wide gene co-activity scores were very strongly correlated (**Figure 3.4d** and **Supplementary Figure 3.35**; $r = 0.72$ - 0.99 in each data set-cell type pair). Analyses using correlations across single cells instead of metacells produced similar findings (**Supplementary Figure 3.36**).

Although TF binding within a peak explained substantial variance in peak co-accessibility (and peak co-activity) (**Figure 3.3a-b**), TF binding within a gene promoter explained very little variance in gene co-expression (and gene co-activity) (data not shown). Instead, gene co-expression is driven by other biological factors^{36,134,135} (as well as gene read depth; see below). In particular, gene co-expression may be driven by peak co-accessibility, as hypothesized by ref.⁹⁴. Indeed, we observed that the average gene-gene r^2 for pairs of genes linked to highly co-accessible pairs of peaks was substantially higher ($r^2 = 0.28$) than for all gene-gene pairs ($r^2 = 0.047$; p-value for difference = $4e-15$), supporting this hypothesis (**Supplementary Figure 3.37**).

We performed 5 secondary analyses to rule out alternative explanations for the correlation between gene co-expression scores and gene co-activity scores. First, to further assess the effects of biological causality, we performed a bivariate regression of gene co-activity score on gene co-expression score and a binary indicator of CRISPR-positive status (restricting to genes with CRISPR-positive and CRISPR-negative peaks), as well as univariate regressions of gene co-activity score on each predictor individually. We determined that gene co-expression score alone explained a much higher proportion of the observed variation in gene co-activity scores, and that the bivariate model did not substantially increase the proportion of variation explained (**Supplementary Figure 3.38**). Second, to assess gene read depth effects, we performed a bivariate regression of gene co-activity score on gene co-expression score and gene read depth, as well as univariate regressions of gene co-activity score on each predictor individually. We determined that gene co-expression score alone tended to explain a higher proportion of the variance in gene co-activity scores ($R^2=0.062-0.57$) vs. gene read depth alone ($R^2=0.12-0.27$), and that the bivariate model did not substantially increase the proportion of variation explained ($R^2=0.20-0.58$; **Supplementary Figure 3.39**); we note that gene read depth explained $R^2=0.036-0.24$ of the variance in gene co-expression scores. Third, to assess cell subtype effects, we computed gene co-expression scores and gene co-activity scores using gene-gene and peak-gene correlations conditional on cell subtype proportions within each metacell of the Luecken¹⁰⁹

T cell data set. We determined that the correlation between gene co-expression scores and gene co-activity scores remained substantial ($r = 0.66$, vs. $r = 0.57$ for unconditioned correlation), suggesting that major cell subtypes do not explain this correlation (**Supplementary Figure 3.40**). Fourth, to assess gene measurement noise effects, we introduced additional artificial noise for a subset of genes and recomputed the correlation between gene co-expression scores and gene co-activity scores. We determined that adding noise did not substantially increase the correlation (**Supplementary Figure 3.41**). Fifth, we assessed the correlation between gene co-expression scores and gene co-activity scores when conditioning on the number of genes with TSS <1Mb from the focal gene's TSS (in addition to conditioning on the number of nearby peaks; see above). We determined that the correlation remained strong (**Supplementary Figure 3.42**). We thus conclude that the correlation between gene co-expression scores and gene co-activity scores cannot be explained by biological causality (also see **Figure 3.4b**), gene read depth, major cell subtypes, varying gene measurement noise, or the number of nearby genes.

These findings demonstrate the pervasive effects of gene-gene tagging on peak-gene associations.

Accounting for co-accessibility between peaks via fine-mapping improves peak-gene linking

We sought to assess whether peak-level fine-mapping could distinguish causal peaks from tagging peaks, analogous to SNP-level fine-mapping in GWAS^{114,136,137}. For each target gene (in all 7 data set-cell type pairs; **Table 3.1**), we applied SuSiE¹¹⁴ to our multiome data (processed using ArchR⁸² or Signac^{77,81} pipelines) to compute posterior inclusion probabilities (PIPs) for each peak, and multiplied marginal peak-gene linking scores (squared peak-gene correlations from ArchR⁸² or Signac^{77,81}) by PIPs to obtain fine-mapped peak-gene linking scores. We assumed either a single causal peak or multiple causal peaks per gene, and used either flat priors or functionally informed priors obtained from S-CASC (analogous to ref.¹³⁸). We evaluated the fine-mapped peak-gene linking scores using evaluation sets derived from CRISPR^{47–58} and eQTL^{63,64} data, computing average enrichment across recall values⁹⁰. Further details are provided in the

Methods section. We caution that we view these fine-mapping analyses as a proof of concept that accounting for tagging can improve upon existing approaches, and not as a resolution of the state-of-the-art.

We first evaluated fine-mapped peak-gene linking scores using 448 gold-standard links validated by CRISPR^{47–58}. Results for ArchR⁸² are reported **Figure 3.5a** and **Supplementary Figure 3.43**. Fine-mapped peak-gene linking scores substantially outperformed marginal peak-gene linking scores, particularly for single causal peak methods (average enrichment 2.48-2.51x vs. 1.1x for marginal, each p-value for difference < 8e-15); we observed little benefit from including functional priors. Results for Signac^{77,81} are reported in **Figure 3.5b** and **Supplementary Figure 3.44**. Again, fine-mapped peak-gene linking scores substantially outperformed marginal peak-gene linking scores, particularly for single causal peak methods (average enrichment 3.49-3.50x vs. 2.2x for marginal, each p-value for difference < 2e-10), with little benefit from including functional priors.

We next evaluated fine-mapped peak-gene linking scores using 39,194 predicted SNP-gene links implicated by fine-mapped eQTL (eSNP–eGene pairs with maximum PIP > 0.5 across GTEx tissues^{63,64}). Results for ArchR⁸² are reported in **Figure 3.5c** and **Supplementary Figure 3.45**. Fine-mapped peak-gene linking scores substantially outperformed marginal peak-gene linking scores, with functionally informed multiple causal peak fine-mapping performing best (average enrichment 1.43x vs. 1.16x for marginal, p-value for difference = 1e-48; all fine-mapped peak-gene linking scores outperform marginal, p-values for difference < 6e-8). Results for Signac^{77,81} are reported in **Figure 3.5d** and **Supplementary Figure 3.46**. Functionally informed multiple causal peak fine-mapping again performed best, but with a much smaller improvement (average enrichment 1.69x vs. 1.64x for marginal, p-value for difference = 2e-5). Non-functionally informed multiple causal peak fine-mapping also slightly outperformed marginal peak-gene linking scores (p-value for difference = 0.0057), while single causal peak methods underperformed marginal peak-gene linking scores.

We also sought to assess whether gene-level fine-mapping could pinpoint target genes. For each peak, we applied SuSiE¹¹⁴ to our multiome data to compute PIPs for each gene, and multiplied marginal peak-gene linking scores (squared peak-gene correlation from ArchR⁸² or Signac^{77,81}) by PIPs to obtain fine-mapped gene-peak linking scores, as above (**Methods**). We determined that both fine-mapped ArchR⁸² peak-gene linking scores and fine-mapped Signac^{77,81} peak-gene linking scores significantly outperformed marginal peak-gene linking scores, on both CRISPR^{47–58} and eQTL^{63,64} evaluation data (**Supplementary Figure 3.47**).

We performed 3 secondary analyses. First, we performed multiple causal peak fine-mapping of peak-gene correlations across metacells (from ArchR⁸²) using peak-peak correlations computed across single cells (from Signac^{77,81}), and vice versa. We determined that these approaches did not outperform estimating peak-peak correlations across the same data types (metacells or single cells, respectively) as the focal peak-gene linking method (**Supplementary Figure 3.48**). Second, we performed fine-mapping restricting to candidate peak-gene pairs with strong correlations instead of all peak-gene pairs. We determined that this approach did not outperform using all candidate peak-gene pairs in fine-mapping (**Supplementary Figure 3.49**). Third, we compared marginal and fine-mapped scores to a ‘top peak’ approach, retaining only peak-gene links for which the peak attains the highest correlation to the focal gene. We determined that the top-peak approach attained similar enrichment to single causal peak fine-mapping but attained very low recall, and is thus a less desirable approach (**Supplementary Figure 3.50**).

These initial efforts at fine-mapping peak-gene links using standard approaches^{114,138} suggest that accounting for peak-peak co-accessibility has substantial potential to improve peak-gene links.

Peak-peak and gene-gene tagging impact peak-gene links at GWAS loci

We have shown that peak-peak tagging and gene-gene tagging effects are present in single-cell multiome data and induce tagging peak-gene associations (**Figure 3.2** and **Figure 3.4**,

resp.). Below, we dissect two examples illustrating peak-peak and gene-gene tagging effects in the Xu K562¹¹⁰ data set. We selected these examples from 90 peak-gene links (marginal ArchR linking score (r^2) > 0.2025, corresponding to ArchR threshold $r > 0.45^{82,90}$) involving a peak harboring a fine-mapped GWAS variant (PIP > 0.5 across 32 blood cell traits¹³⁹), based on single causal variant fine-mapping (as multiple causal variant fine-mapping is not recommended in the absence of in-sample LD¹³⁸).

First, we dissect an example involving fine-mapping the causal peak for a given gene. ArchR linked the gene *RAP1GAP* to 16 peaks, including the ATAC peak spanning chr1:21687033-21688350 (assigned the highest marginal ArchR linking score among 171 peaks <1Mb from the *RAP1GAP* TSS) (**Figure 3.6a**, upper panel). The linked peak harbors the fine-mapped GWAS variant rs17423390, fine-mapped to high light scatter reticulocyte percentage of red blood cells (HLRP) with PIP = 0.63^{139,140}. *RAP1GAP* has been well studied for its role in the differentiation and survival of erythroid cells (and other hematopoietic cell lineages)^{59,141–144}. HLRP reflects erythroid cell maturity and is an established clinical indicator for anemia^{145–147}. Therefore, *RAP1GAP* is a biologically plausible driver of this GWAS association, supporting the link between *RAP1GAP* and chr1:21687033-21688350; an erythroid/hematopoietic progenitor cell line such as K562^{61,148} is also a biologically plausible context for detecting this link. We measured the correlation of each other peak (<1Mb from the *RAP1GAP* TSS) with both chr1:21687033-21688350 and *RAP1GAP*. Squared correlation with chr1:21687033-21688350 was strongly correlated ($r = 0.90$) with linking score to *RAP1GAP* (**Figure 3.6a**, upper panel and **Supplementary Figure 3.51**), suggesting the presence of peak-peak tagging. (We note that measured peak-peak r^2 are likely underestimates of true biological values, due to peak measurement noise.)

We determined above that pioneer TFs were important for explaining peak co-accessibility and tagging peak-gene links (**Figure 3.3**). For each tagging peak, we thus measured the number of pioneer TFs binding both that tagging peak and the focal peak chr1:21687033-21688350 (of 6

total pioneer TFs^{130,131} that bind chr1:21687033-21688350). The number of shared pioneer TFs was moderately correlated to the tagging peak's squared correlation with chr1:21687033-21688350 ($r = 0.45$) and the tagging peak's linking score to *RAP1GAP* ($r = 0.42$) across peaks (**Supplementary Figure 3.52**). In particular, GATA1 is critical for hematopoiesis^{149–151} and has been shown to affect the accessibility of chr1:21687033-21688350 and the expression of *RAP1GAP* in Perturb-multiome experiments¹³². GATA1 binding within a peak was moderately correlated with squared correlation to chr1:21687033-21688350 ($r = 0.35$) and linking score to *RAP1GAP* ($r = 0.32$) across peaks. Other TFs were also moderately correlated with both chr1:21687033-21688350 and *RAP1GAP*, but were not implicated by Perturb-multiome¹³² (data not shown). These findings provide additional support for this peak-gene link and for the TF-driven origins of peak-peak tagging.

We examined fine-mapped ArchR linking scores for *RAP1GAP*. Fine-mapped ArchR scores (assuming a single causal peak and flat priors) correctly prioritized the link to chr1:21687033-21688350 (the 4th closest peak) and assigned very low linking scores to other peaks (**Figure 3.6a**, lower panel), demonstrating the ability of fine-mapping to pinpoint causal peak-gene links. While the marginal ArchR method correctly prioritized the link to chr1:21687033-21688350, marginal ArchR linked 15 other peaks to *RAP1GAP* (**Figure 3.6a**, upper panel). Fine-mapping approaches assuming multiple causal peaks linked 3 additional peaks to *RAP1GAP*; while none harbor fine-mapped GWAS variants, these may be true causal peaks in this or other cellular contexts (**Supplementary Figure 3.53**). Functionally-informed approaches yielded highly similar results (**Supplementary Figure 3.53**). Overall, all fine-mapping approaches generated a smaller and thus 'de-noised' set of peaks relative to marginal peak-gene linking scores.

Second, we dissect an example involving fine-mapping the target gene for a given peak. ArchR linked the ATAC peak spanning chr19:8417808-8418540 to 7 genes, including the gene *ANGPTL4* (assigned the highest linking score among 35 genes with TSS <1 Mb from the focal peak) (**Figure 3.6b**, upper panel). The peak chr19:8417808-8418540 harbors the fine-mapped

GWAS variant rs12150986, fine-mapped to mean platelet volume with PIP > 0.64^{139,140}. *ANGPTL4* encodes an angiopoietin-like protein, a family of proteins well-studied for their role in hematopoiesis (production of erythroid and other blood cells from hematopoietic stem and progenitor cells)^{152,153}, and has been broadly implicated in disease^{154–156}. *ANGPTL4* has been shown to support the survival and expansion of hematopoietic progenitors^{157,158}, with *ANGPTL4* treatment linked to increased platelet production in mouse experiments¹⁵⁹. Therefore, *ANGPTL4* is a biologically plausible driver of this GWAS association, supporting the link between *ANGPTL4* and chr19:8417808-8418540; a hematopoietic progenitor cell line such as K562^{61,148} is also a biologically plausible context for detecting this link. We measured the correlation (across metacells) of each other gene (with TSS <1Mb from the focal peak) with both *ANGPTL4* and the focal peak. Linking score to *ANGPTL4* was strongly correlated ($r = 0.82$) to squared correlation with the focal gene (**Figure 3.6b**, upper panel and **Supplementary Figure 3.54**), suggesting the presence of gene-gene tagging. (We note that measured gene-gene r^2 are likely underestimates of true biological values, due to peak measurement noise.)

We determined above that pioneer TF binding did not explain gene co-expression and that co-expression may instead be explained by gene read depth (see **Supplementary Figure 3.39**) and other biological factors^{36,134,135} (including peak co-accessibility, see **Supplementary Figure 3.37**). We determined that gene read depth was only moderately correlated to the tagging gene's squared correlation with *ANGPTL4* ($r = 0.13$) and the tagging gene's linking score to chr19:8417808-8418540 ($r = 0.29$) (**Supplementary Figure 3.55**) across genes. We also determined that the average squared correlation with *ANGPTL4* was substantially higher for tagging genes strongly linked to peaks co-accessible with chr19:8417808-8418540 ($r^2 = 0.24$) than for all tagging genes ($r^2 = 0.070$) (**Supplementary Figure 3.56**), supporting the hypothesis that peak co-accessibility underlies gene co-expression⁹⁴.

We examined fine-mapped ArchR linking scores for chr19:8417808-8418540 (from fine-mapping genes to a focal peak). Fine-mapped ArchR scores (assuming a single causal gene and

flat priors) correctly prioritized the link to *ANGPTL4* (the 3rd closest gene) and assigned very low linking scores to other genes (**Figure 3.6b**, lower panel), demonstrating the ability of fine-mapping to pinpoint causal peak-gene links. While the marginal ArchR method correctly prioritized the link to *ANGPTL4*, marginal ArchR linked 6 other genes to chr19:8417808-8418540 (**Figure 3.6b**, upper panel). Fine-mapping approaches assuming multiple causal genes linked 2 additional genes to chr19:8417808-8418540 (*CERS4*, with known roles in ceramide synthesis and cancer progression^{160,161}, and *OR7G1*, an olfactory receptor gene involved in odor perception¹⁶²); while these genes seem less functionally relevant, they may be truly causal target genes in this or other cellular contexts (**Supplementary Figure 3.57**). Functionally-informed approaches yielded highly similar results (**Supplementary Figure 3.57**). Overall, all fine-mapping approaches generated a smaller and thus ‘de-noised’ set of genes relative to marginal peak-gene linking scores.

These examples demonstrate the effects of peak-peak and gene-gene tagging at specific peaks and genes implicated in GWAS.

Discussion

In this study, we hypothesized that co-accessibility among scATAC-seq peaks may induce *tagging* peak-gene associations in single-cell multiome data (analogous to tagging associations in GWAS). We examined patterns of peak co-accessibility to assess this hypothesis. We determined that the extent of a peak’s correlation with nearby peaks is strongly related to the extent of its correlation to nearby genes, consistent with tagging and not with other explanations. CRISPRi data and a regression of co-activity scores on co-accessibility scores stratified by functional peak categories (S-CASC) provided additional support for the tagging hypothesis. We explored a TF-driven mechanism for peak tagging and determined that peak co-accessibility is strongly related to TF binding activity (with a particularly strong effect for pioneer TFs, which induce accessibility of closed chromatin). Indeed, although major cell subtypes based on available annotations do not explain the bulk of the correlation between co-accessibility and co-activity, we showed that peak tagging is related to cell subtype / cell state-specific processes (likely reflecting

TF activity), and it is possible that tagging effects would largely disappear in the extreme limit of a perfectly homogeneous cell state across cells. We demonstrated that accounting for tagging effects using statistical fine-mapping improves peak-gene linking. We observed similar patterns when investigating gene-gene tagging effects, suggesting that gene co-expression can also impact peak-gene linking. Finally, we highlighted examples in which peak tagging or gene tagging obscures a causal peak-gene link, such that fine-mapping yields more precise predictions of peak-gene links. Overall, we find that tagging has a major impact on efforts to identify peak-gene links. We propose that tagging peak-gene links should be viewed as spurious links, because they implicate incorrect gene regulatory mechanisms (in contrast to tagging associations in GWAS, which are generally viewed as true associations because they implicate correct GWAS loci).

Our work has several implications for downstream analyses. First, researchers applying existing peak-gene linking methods should be aware that non-causal peak-gene links may be induced by tagging peaks or genes. Examining patterns of peak co-accessibility and gene co-expression in the focal data set or incorporating other types of data (e.g. transcription factor binding data or experimental follow-up via CRISPRi^{47–57} or base editing⁵⁹) may help distinguish causal from tagging links. Second, researchers developing new single-cell peak-gene linking methods should consider how to explicitly model peak co-accessibility and/or gene co-expression to more effectively identify causal links; extensions of the fine-mapping approaches explored here¹¹⁴, as well as other approaches for causal inference^{163–165}, could plausibly improve performance. In particular, peak-gene linking studies are often motivated by the functional interpretation of GWAS associations. A possible future method could jointly examine patterns of peak co-accessibility, gene co-expression, and GWAS summary statistics at a GWAS locus to nominate the causal peak-gene links driving the underlying genetic association.

Our work has several limitations and directions for future research. First, we primarily focused our analyses on blood/immune cell types (for which large and high-quality multiome data sets are available^{77,109,110,129}); in particular, we focused our analyses of transcription factor binding

patterns on the K562 and GM12878 cell lines^{77,110,129}, for which abundant TF ChIP-seq data are available. While our findings were consistent in all cell types analyzed, exact patterns of peak co-accessibility and gene co-expression vary among cell types (**Supplementary Figure 3.16**); it will therefore be important to extend these analyses to other multiome data sets^{91,92}. Second, in our fine-mapping analysis, we applied fine-mapping software designed for use on genotype data (which is generally highly accurate or well-imputed). While these analyses demonstrated in principle that accounting for tagging effects can improve peak-gene linking, extensions of the fine-mapping algorithm that explicitly model measurement noise in single-cell data may improve performance and should be further explored in future studies^{166–168}; indeed, we present the fine-mapping approaches described here as a proof of concept rather than state-of-the-art methods (e.g. functionally informed priors do not always substantially improve performance). Likewise, due to noisy single-cell measurements, relative causal effect size enrichments from S-CASC are likely to be more reliable than absolute causal effect sizes variances. Third, a high fine-mapped linking score does not guarantee biological causality. While fine-mapping linking scores improve upon marginal linking scores (**Figure 3.5**), experimental validation^{47–57,59} is still needed to conclusively demonstrate a causal link. Fourth, we do not include a large set of existing correlation-based peak-gene linking methods in our analyses^{80,88,89}. While our findings were highly consistent using linking scores computed from two foundational and distinct approaches (ArchR⁸² and Signac^{77,81}), it would be valuable to perform these analyses using additional peak-gene linking methods. Fifth, we did not consider TF footprinting^{169,170}, i.e. inferring TF binding sites from the strength and position of ATAC reads, in our TF analyses. TF footprinting requires high-resolution ATAC data, which can be challenging to obtain in a single-cell multiome data set, and achieves variable precision^{169,171,172}. While our findings were consistent using ChIP-seq data (which provide context-specific, ‘gold standard’ readouts for a smaller set of TFs¹⁶⁹) and DNA sequence motifs (which provide non-context-specific predictions for a larger set of TFs¹³³) to define TF binding sites, it would be valuable to perform these analyses using TF footprinting data, particularly as higher

quality multiome data sets amenable to precise footprinting become available^{91,92}. Despite these limitations, we have shown that peak co-accessibility and gene co-expression, driven by TF activity and other biological factors, can induce non-causal peak-gene links in single-cell data, and that accounting for these ‘tagging’ effects is crucial for identifying causal peak-gene links.

Acknowledgements

We thank Benedikt Geiger, Sasha Gusev, Gaspard Kerner, Soumya Raychaudhuri, Ben Strober, and Shamil Sunyaev for helpful discussions, and Gaspard Kerner for providing GWAS fine-mapping results. This research was funded by NIH grants U01 HG012009, R01 MH101244, R37 MH107649, R01 HG006399, R01 MH115676, and R01 HG013083.

Figures

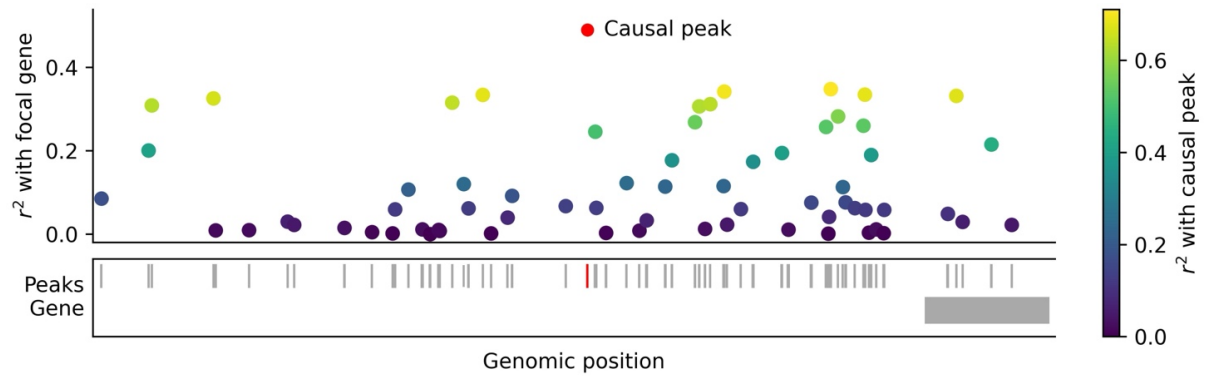


Figure 3.1. Toy example illustrating the effects of peak tagging on peak-gene links. We plot simulated squared peak-gene correlations for each peak <1Mb from the TSS of the focal gene. r^2 values are reported as dots in the middle of each peak, with the causal peak colored in red and other peaks colored based on their squared correlation to the causal peak.

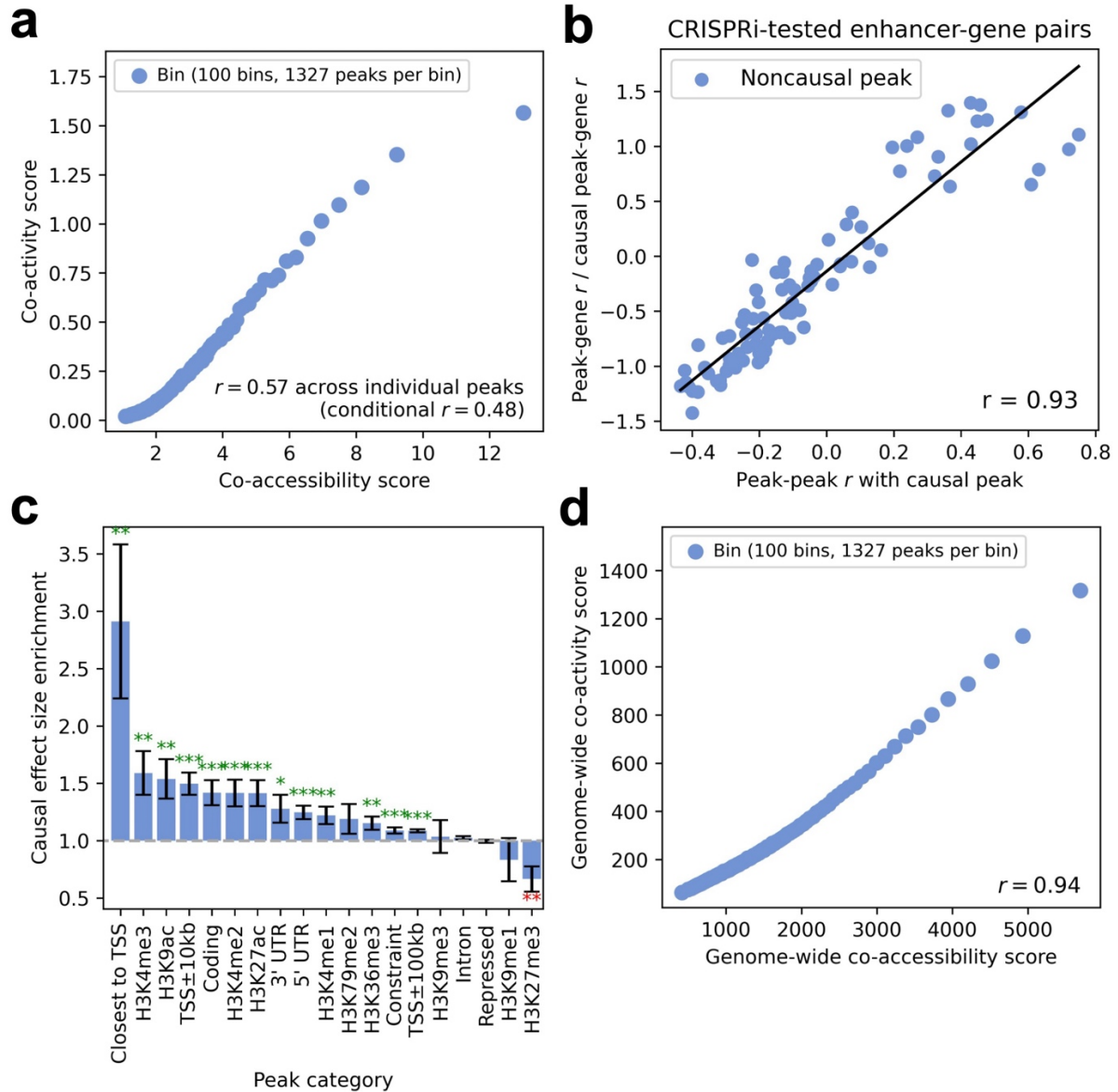


Figure 3.2. Relationship between peak-peak co-accessibility and peak-gene co-activity. a) Relationship between co-accessibility score and co-activity score across 132,736 peaks in the Xu K562¹¹⁰ data set, computed using peaks and genes within the *cis* window (<1Mb) of each focal peak. ‘Conditional r ’ denotes correlation conditioned on number of genes <1Mb from the focal peak. **b)** Relationship between correlation with the CRISPR-validated causal peak and correlation with the target gene (relative to the causal peak’s correlation with the target gene) across non-causal peaks tested by CRISPRi, restricted to CRISPR positive-negative peaks where ArchR correctly identifies the causal peak-gene link. **c)** Causal effect size enrichment of 19 peak categories in stratified co-accessibility score regression. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). **d)** Relationship between genome-wide co-accessibility score and genome-wide co-activity score across 132,736 peaks in the Xu K562¹¹⁰ data set, computed using all peaks and genes for each focal peak. In **a** and **d**, peaks are partitioned equally into 100 bins (each represented by 1 point) by co-accessibility score and genome-wide co-accessibility score, respectively.

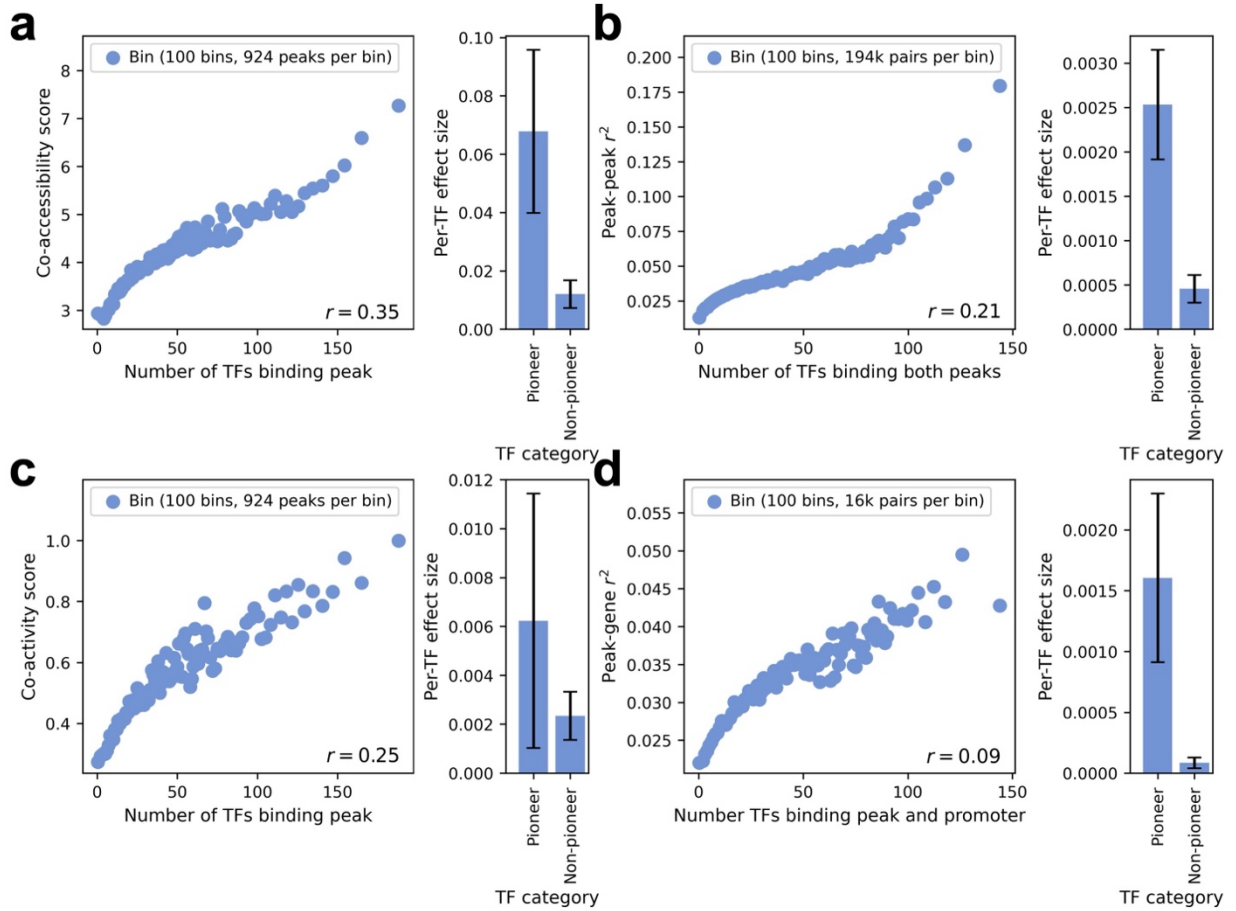


Figure 3.3. Peak co-accessibility and peak-gene co-activity are related to transcription factor binding activity. **a)** Relationship between number of TFs binding an ATAC peak and co-accessibility score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-accessibility score (right). **b)** Relationship between number of TFs binding both peaks in a peak-peak pair and squared peak-peak correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-peak correlation (right). **c)** Relationship between number of TFs binding an ATAC peak and co-activity score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-activity score (right). **d)** Relationship between number of TFs binding both the peak and the gene (promoter) in a peak-gene pair and squared peak-gene correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-gene correlation (right). In left subpanels of **a-d**, correlations are computed across 132,736 peaks in the Xu K562¹¹⁰ data set, and peaks (or peak-peak / peak-gene pairs) are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. In right subpanels of **a-d**, regression effect sizes are meta-analyzed across 3 data set-cell type pairs (Xu K562¹¹⁰, Satpathy-K562¹²⁹, and SHARE-seq-LCL⁷⁷), and confidence intervals denote standard errors.

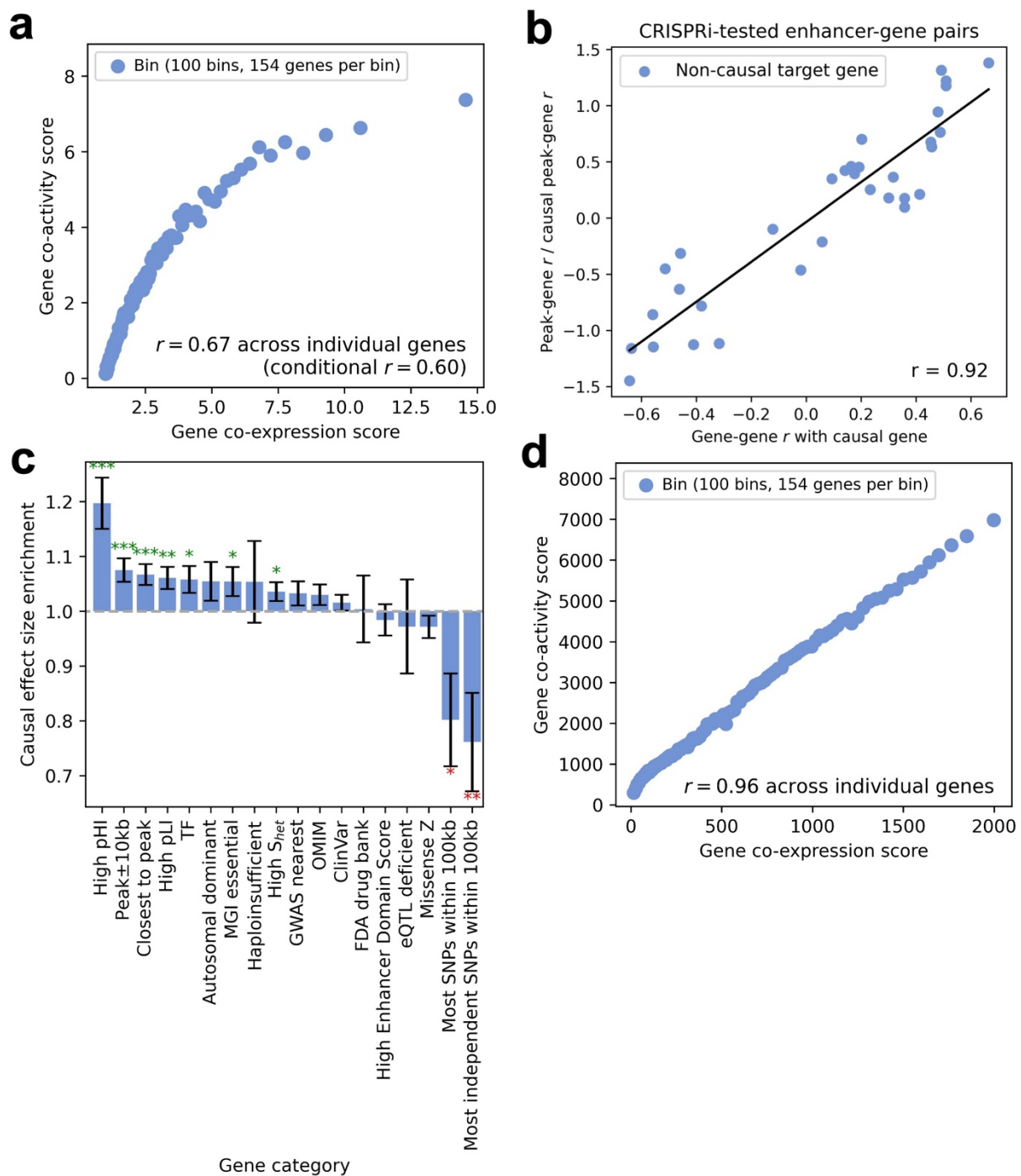


Figure 3.4. Relationship between gene-gene co-expression and peak-gene co-activity. **a)** Relationship between gene co-expression score and gene co-activity score, computed using genes and peaks within the *cis* window (<1Mb) of each focal gene, across 15,460 genes in the Xu K562¹¹⁰ data set. ‘Conditional r ’ denotes correlation conditioned on number of peaks <1Mb from the focal gene. **b)** Relationship between correlation with the CRISPR-validated causal target gene and correlation with the causal peak (relative to the causal target gene’s correlation with the causal peak) across non-causal target genes tested by CRISPR, restricted to CRISPR positive-negative peaks where ArchR correctly identifies the causal peak-gene link. **c)** Causal effect size

enrichment of 18 gene categories in stratified co-expression score regression. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). **d)** Relationship between genome-wide gene co-expression score and genome-wide gene co-activity score, computed using all peaks and genes for each focal gene, across 15,460 genes in the Xu K562¹¹⁰ data set. In **a** and **d**, genes are partitioned equally into 100 bins (each represented by 1 point) by gene co-expression score and genome-wide gene co-expression score, respectively.

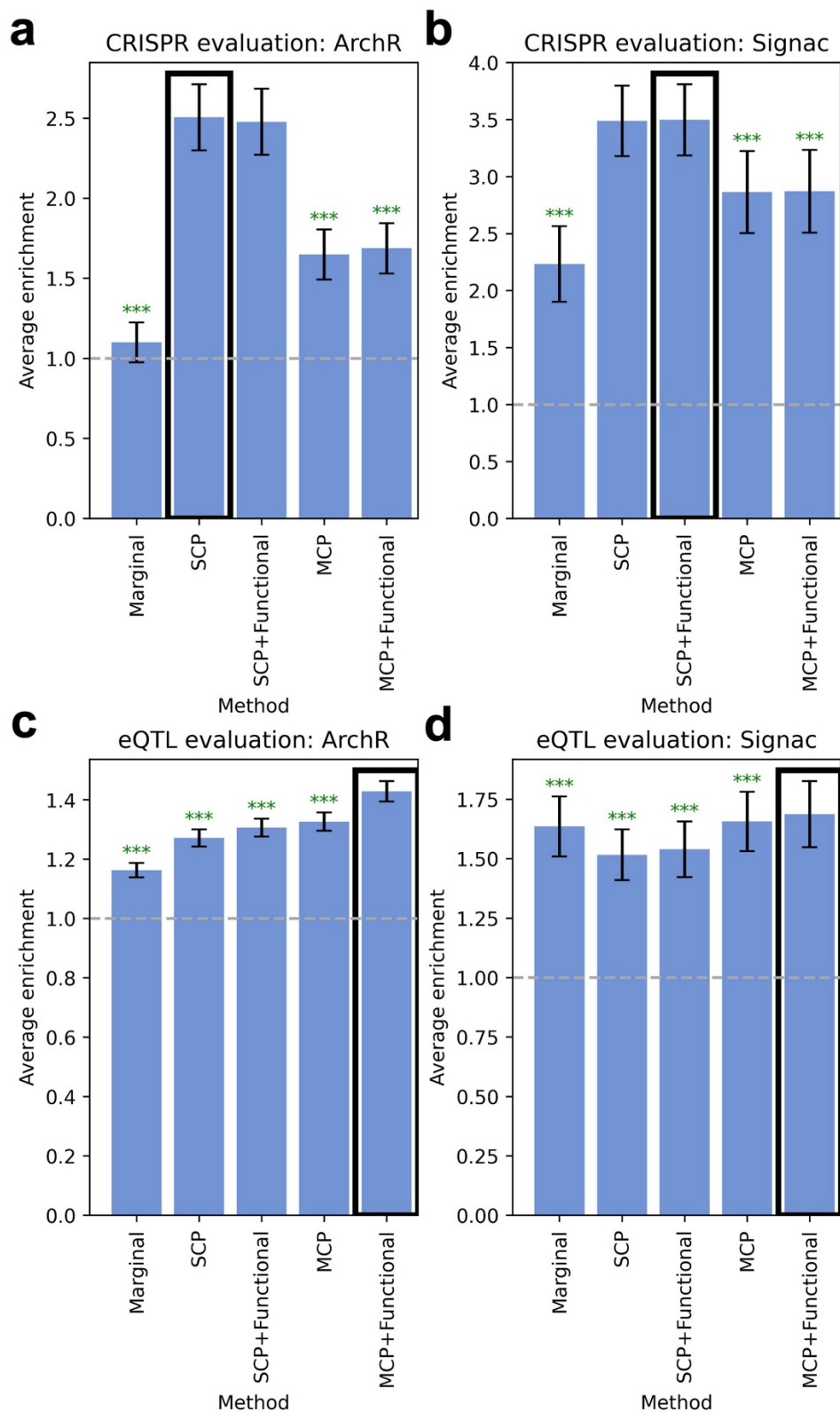


Figure 3.5. Fine-mapping improves peak-gene linking methods. a) Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores for 448 links validated by CRISPR. **b)** Average enrichment across recall values of links predicted by Signac

vs. fine-mapped Signac scores for 448 links validated by CRISPR. **c)** Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. **d)** Average enrichment across recall values of links predicted by Signac vs. fine-mapped Signac scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping; SCP + Functional, single causal peak functionally informed fine-mapping; MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.

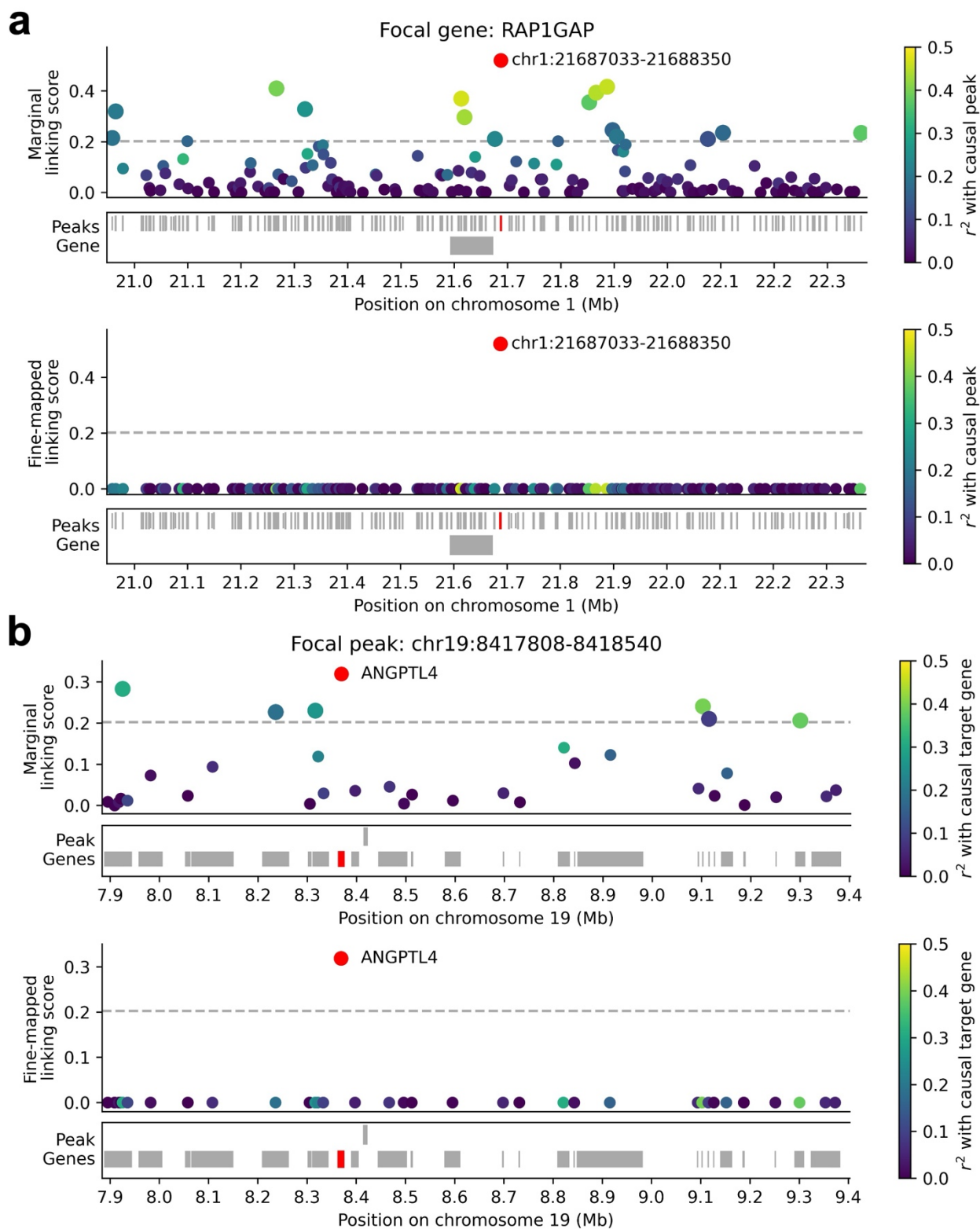


Figure 3.6. Examples of peak-peak and gene-gene tagging. a) Upper panel: Marginal linking score (squared ArchR peak-gene correlation) for each peak <1Mb from the TSS of the gene *RAP1GAP*. r^2 values are reported as dots in the middle of each peak, with the top correlated

peak colored in red and other peaks colored based on their squared correlation to the top correlated peak (non-focal genes not shown). Lower panel: Fine-mapped linking score (squared ArchR peak-gene correlation multiplied by fine-mapped PIP) for each peak <1Mb from the TSS of the gene *RAP1GAP*. **b)** Upper panel: Marginal linking score for each gene with TSS <1Mb from the peak spanning chr19:8417808-8418540. r^2 values are reported as dots in the middle of each gene, with the top correlated gene colored in red and other genes colored based on their squared correlation to the top correlated gene (non-focal peaks not shown). Lower panel: Fine-mapped linking score for each gene with TSS <1Mb from the peak spanning chr19:8417808-8418540. In each panel, dashed line denotes linking score threshold corresponding to the ArchR linking threshold.

Tables

Data set	Cell type(s)	Number of cells
Xu ¹¹⁰	K562*†	7,861
Satpathy ¹²⁹	K562*†	5,359
SHARE-seq LCL ⁷⁷	lymphoblastoid cell line (LCL)*	17,644
Luecken BMMC ¹⁰⁹	T, B, myeloid, erythroid	55,321
<i>All</i>	<i>6 cell types</i>	<i>86,185 cells</i>

Table 3.1. Overview of single-cell multiome data sets. We report the cell types profiled and number of cells passing QC for each single-cell multiome data set. *: TF ChIP-seq data is available for that cell type. †: CRISPR data is available for that cell type.

Chapter 4: Robust and fast context-dependent eQTL mapping for single-cell data

Elizabeth Dorans, Jose Alquicira-Hernandez, Aparna Nathan, Soumya Raychaudhuri

Introduction

Genome-wide association studies (GWAS) identify many non-coding genetic variants associated with disease, but the functional mechanisms by which these variants mediate complex disease risk remain poorly characterized^{15,17,19}. Expression quantitative trait loci (eQTL) studies link regulatory variants to target genes by measuring associations between genotype and gene expression across individuals^{63,64,173}. However, recent work has demonstrated limited overlap between variants implicated by eQTL and GWAS^{66,67}; indeed, only ~10% of disease-causal genes have colocalizing GWAS and eQTL variants²¹. One possible explanation for this discrepancy is context-specific gene regulatory effects. Disease-driving genetic variants may only affect gene expression in a specific rare or transitional cell state; thus, their effects may not be detected by eQTL studies using bulk RNA sequencing data for heterogeneous tissues^{21,22,70,100}. Methods to investigate genetic effects on molecular phenotypes at granular cell state resolution are needed to fully characterize the functional impact of disease variants.

Single-cell eQTL mapping has recently emerged as a powerful strategy for linking disease variants with context-specific effects to their target genes. While previous eQTL studies of context-dependent effects (“dynamic” eQTL) have focused on discrete cell type definitions⁷⁰, single-cell data have the potential to elucidate continuous cell states (e.g. immune activation) at high resolution⁹⁹. In contrast to bulk eQTL mapping, single-cell approaches characterize eQTL in fine-grained cell states, which are often implicated in disease mechanisms (e.g. cytotoxicity, activation, cellular stress, etc.) and difficult to precisely define. Existing single-cell eQTL models^{99,104,174} (e.g. a single-cell Poisson Mixed Effect (scPME) model⁹⁹) have shown promise in identifying disease-

associated variants whose regulatory effects were mediated by continuous cell states. However, the computational complexity of these models is an obstacle to their application; genome-wide eQTL scans in data sets with large sample sizes are often computationally intractable. Existing approaches also often rely on mixed-effect modeling (to account for repeated measurements per donor, cellular heterogeneity, and data sparsity), which can be sensitive to model misspecification and computational instability^{175–177}.

Due to computational constraints, previous work has focused on interaction eQTL mapping at genes with strong main effects, often restricted to the lead main effect variant, instead of locus-wide eQTL scans⁹⁹. There has thus been little exploration of interaction eQTL at genes with no detected main effect. Further, there can be multiple *independent* eQTL effects at the same locus (separate biological signals)^{178,179}. Even genes with strong main effects may harbor interaction effects that are independent from the main effect, which would not be detected by testing only lead main effect variants. Therefore, a systematic investigation into the overlap of genes with main and interaction effects, as well as the independence of main and interaction effects at the same loci, is needed.

Here, we propose a robust and scalable dynamic single-cell eQTL mapping framework modeling cell-state-dependent genetic effects in T cells. We first identify genetic variants that interact with continuously defined cell states (e.g. cytotoxicity and regulatory function). We then assess whether these dynamic eQTL effects tend to occur at previously reported “main effect” genes. We further explore the overlap of interaction eQTL effects with variant-disease associations from GWAS and provide an example of a context-dependent eQTL that may drive autoimmune disease. Profiling cell-state-dependent eQTL at cellular resolution is critical for profiling the full landscape of gene regulatory effects and understanding the biological mechanisms of disease.

Results

Overview of Methods

Dynema performs single-cell dynamic eQTL mapping to identify variant-gene pairs where continuous cell state modulates genetic effect on gene expression. This approach provides robust p-values across genes and multiple cell states, scales to larger datasets, and enables locus-wide eQTL searches.

Dynema uses a Poisson Generalized Linear Model (Poisson GLM) to model single-cell gene expression as a function of SNP genotype dosage, cell state(s), and genotype x cell state interaction(s), while controlling for sample and single-cell covariates (**Methods**). *Dynema* takes as input: (1) a single-cell gene expression matrix, (2) a genotype matrix, (3) a context matrix representing discrete and/or continuous cellular phenotypes (e.g. cell states), and (4) a covariate matrix including cell-level and sample-level covariates. Cell states (pre-defined by the user) might include discrete phenotypes (e.g. cell type, pathway activation, surface marker expression) or continuous phenotypes describing dynamic cell processes (e.g. T cell cytotoxicity, differentiation, activation).

In detail, we modeled single-cell gene expression counts using a Poisson Generalized Linear Model:

$$\log(E) = \beta_0 + \beta_G G + \sum_{i=1}^n \beta_{CV_i} CV_i + \sum_{i=1}^n \beta_{G \times CV_i} (G \times CV_i) + \mu, \quad [4.1]$$

where E is the gene expression count, β_0 is the intercept term, G is the alternate allele dosage of the tested SNP, CV is the cell state variable of interest, $G \times CV$ is the interaction term between genotype and cell state, and μ are the residuals. We also include biological and technical covariates to account for possible confounding (see **Methods**).

After fitting the Poisson GLM, *Dynema* uses cluster-robust variance estimators (CRVE) to empirically measure the residual variance at the donor level and generate calibrated p-values in the presence of correlated cells within donors and moderate model misspecification. *Dynema* also offers a flexible framework to test single or multiple model parameters jointly, including eQTL main (i.e., cell-state-independent) or interaction (i.e., cell-state-dependent) effects.

We analyzed an existing cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) data set of 500,089 memory T cells from 259 healthy Peruvian individuals with a previous tuberculosis infection¹⁸⁰ (**Methods**). These data contain measured gene expression and T cell surface proteins, allowing for single-cell-level eQTL mapping with single-cell-level cell state definitions. We quantified 3 independent continuous cell state representations using the top 3 canonical variates (CVs) from canonical correlation analysis on both modalities, corresponding to cytotoxicity (CV₁), Treg/activation (CV₂), and central memory (CV₃) phenotypes, as previously described^{99,180} (which play key roles in immune function and autoimmune disease risk). We jointly tested for significant interaction effects across the 3 CVs (testing for nonzero variance explained by any of the 3 CVs).

Dynema is well-calibrated to identify genetic effects that interact with cell state

We evaluated the calibration of p-values for genetic interaction effects computed by *Dynema*. To quantify model calibration, we selected 2,202 genes and their lead eQTL SNPs from our previous pseudobulk analysis of the TBRU data set. We then permuted genotypes to generate a “null” effect scenario and tested for an eQTL interaction effect with CV₁ (T cell cytotoxicity). We quantified model inflation by computing a lambda inflation factor λ . *Dynema* produced well-calibrated p-values across all genes ($\lambda = 1.04$).

We further assessed *Dynema*’s calibration in extreme scenarios with smaller sample size and cell count imbalance across donors. We generated downsampled gene expression for each of the 2,202 SNP-gene pairs by downsampling donors (selecting a random subset of 50,

100, or 200 donors), downsampling cells (selecting a random subset of 10%, 25%, or 50% of cells for half of donors, and retaining all cells for half of donors), or both (selecting a random subset of 50 donors and downsampling cells as previously described). We observed moderately robust calibration of Dynema p-values in each scenario ($\lambda = 1.04$ -1.22). We conclude that Dynema is generally robust to low sample size and cell count imbalance.

Dynema identifies genetic effects that interact with cell state

We first defined a set of eGenes with any significant genetic effect for use in all analyses. We ran Dynema on 6,019/6,267 genes from the TBRU data set expressed in at least 50% of donors (due to computational constraints, analyses are ongoing). We first tested “total” genetic effect of all *cis* variants within 1 Mb of each gene’s TSS using a 4-degree of freedom Lagrange multiplier test jointly testing the main effect and interaction effects for CVs 1-3 (from CCA on the TBRU data set, as in ref.⁹⁹) (**Methods**). We defined a conservative Bonferroni-corrected p-value threshold of $0.05 / 11,721,353$ (the total number of variant-gene pairs tested in this study). We restricted subsequent analyses to 776 genes with at least one variant whose “total” genetic effect passed this Bonferroni-corrected p-value threshold.

For each of these 776 genes, we then separately ran Dynema testing the main genetic effect and the interaction genetic effect (jointly testing CVs 1-3). For each locus-wide eQTL scan, we generated FDR-corrected p-values (using Benjamini-Hochberg correction across all *cis* variants at a locus) and defined a “significant” variant-gene effect as $FDR < 0.05$ for either a main effect ($FDR < 0.05$ for main genotype effect) or an interaction ($FDR < 0.05$ for joint CVs 1-3 interaction). We defined a main eGene as any gene with a significant main effect eQTL and an interaction eGene as any gene with a significant interaction eQTL. We discovered 768 main effect eGenes and 285 interaction eGenes (**Figure 4.1a**). While a substantial proportion of main eGenes were not interaction eGenes (490/768), the majority of interaction eGenes were also main eGenes (278/285) (**Figure 4.1b**).

Genetic effects that interact with cell state are often independent from main genetic effects

We next assessed whether cell state interaction genetic effects are independent from main genetic effects for 211/285 interaction eGenes (due to computational constraints; analyses are ongoing). For 11/211 interaction eGenes, the lead interaction SNP was identical to the lead main effect SNP, and for 200/211 interaction eGenes, the lead interaction SNP and lead main effect SNP were different. To determine whether an interaction genetic effect controlling a gene's expression was independent from the main genetic effect, we performed conditional eQTL mapping: for each interaction eGene, we retested all *cis* SNPs for their joint interaction genetic effect, conditioned on the main and interaction effects of the lead main effect SNP, computing conditional p-values for each SNP. We defined an interaction eGene as having an “independent” interaction effect if the FDR-corrected p-value for the original lead interaction SNP remained < 0.05 after conditioning and the correlation between log-transformed interaction p-values across SNPs before and after conditioning was > 0.9 (**Figure 4.2a**).

We determined that 21 genes (out of 211 interaction eGenes) had interaction effects independent from the main effect (**Figure 4.2b**). We further restricted to genes with significant interaction and main effects and determined that 17 genes (out of 205 genes that were both interaction eGenes and main effect eGenes) had non-independent interaction and main effects (**Figure 4.2b**). These results suggest that cell state interaction eQTL may sometimes be independent from main effect eQTL and may often provide mechanistic insights into the specific contexts driving a main effect eQTL, motivating further investigation into the disease relevance of these variants.

We highlight *CAPN2* (calpain-2), an example of an eGene with independent main and interaction effects (**Figure 4.2c**). The lead main effect variant at the *CAPN2* locus, located at position 223756117 on chromosome 1, attains a p-value of $5e-12$ (and FDR-corrected p-value of $5e-9$). The lead interaction effect variant at the *CAPN2* locus, located at position 223712428 on

chromosome 1, attains a p-value of $2e-5$ (and FDR-corrected p-value of 0.0013). After conditioning on the lead main effect variant, this variant attains an interaction p-value of $3e-4$ (FDR-corrected p-value of 0.0064), and the correlation between log-transformed interaction p-values across variants before and after conditioning was 0.95, suggesting independent interaction and main effects.

CAPN2 is a calcium-dependent cysteine protease with established roles in immune synapse formation and NF- κ B signaling^{181,182}. In T cells, calpain activity is linked to T cell receptor (TCR)-induced calcium influx, translating antigen recognition into changes in signaling and cell motility^{183,184}. A cell state interaction eQTL for CAPN2 is biologically plausible because its functional impact would be pronounced during T cell activation, which induces calcium-related pathways; indeed, we observe the largest interaction effect size for the interaction of genotype with CV₂, previously shown to quantify T cell regulatory vs. activated state⁹⁹. Examination of total SNP effect sizes per cell also suggests that eQTL effects vary with the value of CV2 (**Figure 4.2d-e**). This example demonstrates how genetic variation may selectively modulate gene expression dependent on T cell activation dynamics. This example also underscores the importance of conducting a principled test for effect independence rather than relying on heuristics; while locus plots look visually similar and may suggest the main and interaction signal arise from the same causal variant, statistical investigation strongly suggests independence.

Interaction genetic effects drive disease associations in GWAS

We tested for colocalization of interaction and main eQTL effects with autoimmune disease GWAS associations. We used publicly available summary statistics for rheumatoid arthritis (RA)¹⁸⁵, systemic lupus erythematosus (SLE)¹⁸⁶, and Type 1 diabetes (T1D)¹⁸⁷. We ran coloc¹⁸⁸ for all eGenes with each of the 3 diseases, defining colocalization as probability of a shared causal variant (PPH4) > 0.5, as in ref.⁹⁹ Out of all 776 eGenes, 14-29 genes had main effects that colocalized with GWAS associations for each disease and 6-13 genes had interaction

effects that colocalized with GWAS associations for each disease (**Figure 4.3a**). Below, we highlight an example of an interaction eQTL effect that colocalized with disease GWAS signal.

We observed strong colocalization of the interaction effect for the *ORMDL3* gene with a GWAS association for Type 1 diabetes¹⁸⁷ (PPH4 = 0.92) (**Figure 4.3b**). The *ORMDL3* locus harbors a strong interaction eQTL effect (lead variant located at position 39912129 on chromosome 17, p-value = 4e-7; **Figure 4.3c**, upper panel). We also observed a strong main effect eQTL for *ORMDL3* (lead main variant located at position 39888640 on chromosome 17, attains a p-value of 2e-27; **Figure 4.3c**, lower panel). We determined that the interaction effect was not independent from the main effect at this locus (correlation between log-transformed interaction p-values across variants before and after conditioning on lead main variant = -0.023); indeed, we also observed colocalization of the main effect with T1D GWAS (PPH4 = 0.61). Thus, the *ORMDL3* locus may harbor a T1D-driving genetic association with both a cell-state-independent genotype effect and a cell state interaction effect. *ORMDL3* regulates cellular stress responses and sphingolipid metabolism, which are both critical for T cell receptor (TCR) signaling and cytokine production, and plays a known role in other immune-related diseases (e.g. allergic asthma)^{189–191}. Higher *ORMDL3* expression may drive cytotoxicity and inflammatory cytokine release, increasing pancreatic β -cell destruction in T1D¹⁹² and magnifying genetic effects in cytotoxic cells; indeed, we observe the largest interaction effect size for the interaction of genotype with CV₁, previously shown to quantify T cell cytotoxicity⁹⁹. Examination of total SNP effect sizes per cell also suggests that eQTL effects vary with the value of CV₁ (**Figure 4.3d**). This example illustrates how, in addition to revealing novel genetic effects, cell state interaction eQTL can nominate biological mechanisms for known genetic effects.

Discussion

In this study, we developed a statistically robust and computationally efficient method for dynamic eQTL mapping using single-cell gene expression data. We applied this method to a

paired genotype and CITE-seq data set from healthy individuals with a previous tuberculosis infection. We demonstrated that genes may often harbor eQTL effects that interact with cell state. These interaction effects may exist with or without a corresponding main effect, and they may or may not be independent from the main effect. We most frequently observe interaction effects for eGenes harboring main effects, and interaction and main effects are most often not independent. We identify examples of colocalization between interaction eQTL effects and GWAS variant-disease associations, demonstrating how interaction eQTL can reveal novel regulatory effects or provide mechanistic insights into known genetic effects.

Our work has several implications for downstream analyses. First, Dynema software is publicly available for users to apply to additional data sets with paired genotype and CITE-seq or scRNA-seq. Because dynamic eQTL are context-specific and require many donors/cells to detect, there are likely many more dynamic eQTL (in the T cell context and beyond) that explain disease signal than those reported here. Second, Dynema SNP-gene associations nominate SNP-gene pairs for functional follow-up (e.g. via CRISPRi^{47–54} or base editing⁵⁹) to validate predicted links and investigate the cell type-specific mechanisms of the genetic associations identified.

Our work has several limitations and directions for future research. First, our analyses have thus far been restricted to the TBRU data set¹⁸⁰; while this data set is well-powered and characterized by highly interpretable CVs, it will be important to validate our results in additional (particularly, larger) data sets⁹¹. Second, we currently rely on a maximally stringent p-value correction procedure to select genes of interest, which may dampen true biological signal, especially for interaction eQTL which may lack power; we plan to develop a permutation-based approach to construct a per-gene null distribution and more accurately quantify the significance of genetic associations. We will also explore larger data-set and potential meta-analyses to maximize power. Third, we have conducted GWAS colocalization analyses for a limited number of immune-related diseases (RA¹⁸⁵, SLE¹⁸⁶, and T1D¹⁸⁷); it will be important to conduct these

analyses in a larger and broader set of diseases and traits with publicly available GWAS summary statistics. Fourth, it would be of interest to conduct large-scale validation of Dynema on real data (in addition to calibration simulations); while gold-standard sets of dynamic eQTL effects are not widely available, gene-level or variant-level validations (e.g. disease-relevance or context specificity scores^{33,106,193–195}) could be useful in characterizing the regulatory relationships uniquely captured by Dynema. Fifth, our analyses focus on T cells and well-defined (from prior work¹⁸⁰) cell states; it will be of great importance to expand this work beyond the T cell context to new data sets profiling other cell types. This extension will require further consideration of the dimension reduction approaches for each biological context (and careful interpretation of the biological dimensions of variation captured), as CCA may not be suitable for non-immune contexts.

Despite these limitations, we have shown that Dynema robustly and efficiently captures genetic effects on gene expression that interact with cell state, suggests biological mechanisms underlying known regulatory effects, and reveals new disease-relevant gene regulatory effects.

Author Contributions

J. A.-H., A.N., and S.R. conceived this study. J. A.-H. developed the Dynema model and conducted calibration analyses. E.D. conducted conditional interaction eQTL analyses and GWAS colocalization analyses. E.D. and J. A.-H. contributed to the written report. A.N. and S.R. provided critical scientific guidance.

Figures

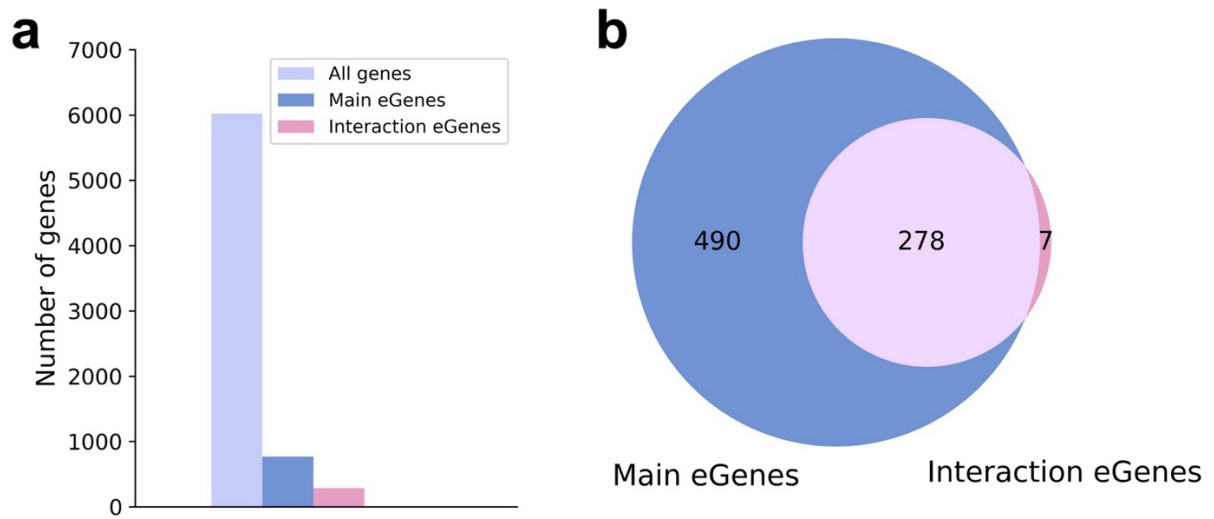


Figure 4.1. Dynama identifies genetic effects on gene expression that interact with cell state. **a)** Number of genes expressed in at least 50% of donors ('All genes'), genes that have at least one variant within 1Mb with a significant main genotype effect ('Main eGenes'), and genes that have at least one variant within 1Mb with a significant interaction between genotype and cell state ('Interaction eGenes') in the TBRU data set. **b)** Set overlap between main eGenes and interaction eGenes.

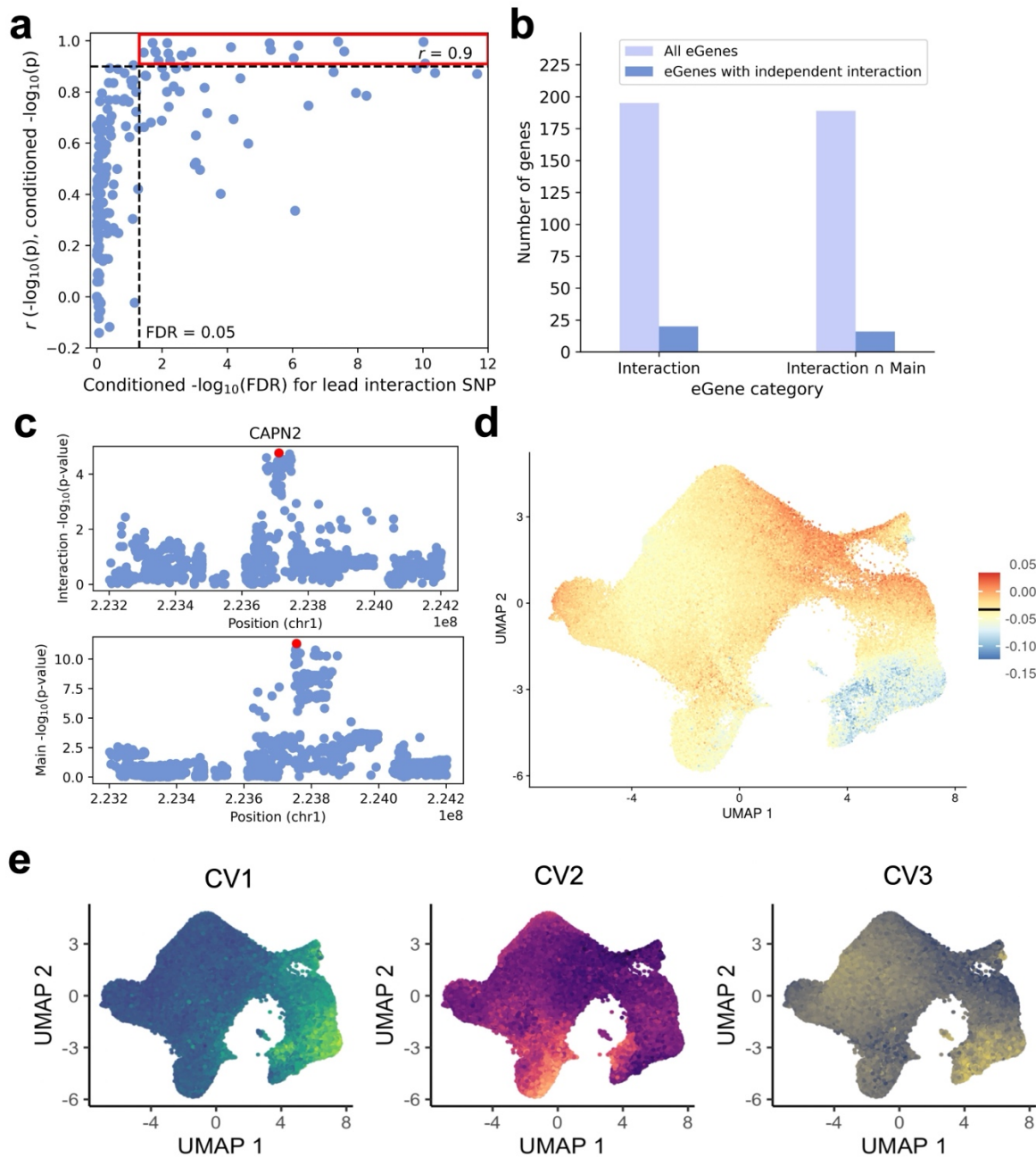


Figure 4.2. Independence of main and cell state interaction genetic effects. **a)** Relationship between conditional lead interaction FDR-corrected p-values (lead interaction effect p-values after conditioning on the lead main effect SNP, corrected via FDR; x-axis) and the correlation across SNPs of interaction effect p-values with conditional interaction p-values (y-axis), across 211 interaction eGenes. Red outline denotes interaction eGenes whose interaction effect is independent from the gene's main effect (defined by conditional lead interaction p-value < 0.001 and $r(\text{unconditional interaction p-values, conditional interaction p-values}) > 0.9$, denoted by black dashed lines). **b)** Number of interaction eGenes (left) and interaction and main eGenes (right) whose interaction effects are tested for independence from the main effect ('All genes') and whose interaction effects are determined to be independent from the main effect ('eGenes with independent interaction'). **c)** Locus plots mapping the joint interaction effects (top) and main

effects (bottom) for all variants <1 Mb from the *CAPN2* TSS. Y-axes display $-\log_{10}$ -transformed p-values for either joint interaction p-value (top) or main effect p-value (bottom). Red points highlight the variant with the most significant p-value. **d)** UMAP embedding of cells in the TBRU data set, colored by total effect on *CAPN2* expression for each cell (main effect size plus interaction effect sizes multiplied by CV values for each cell). **e)** UMAPs of cells in the TBRU data set, colored by values of CV1 (left), CV2 (middle), and CV3 (right).

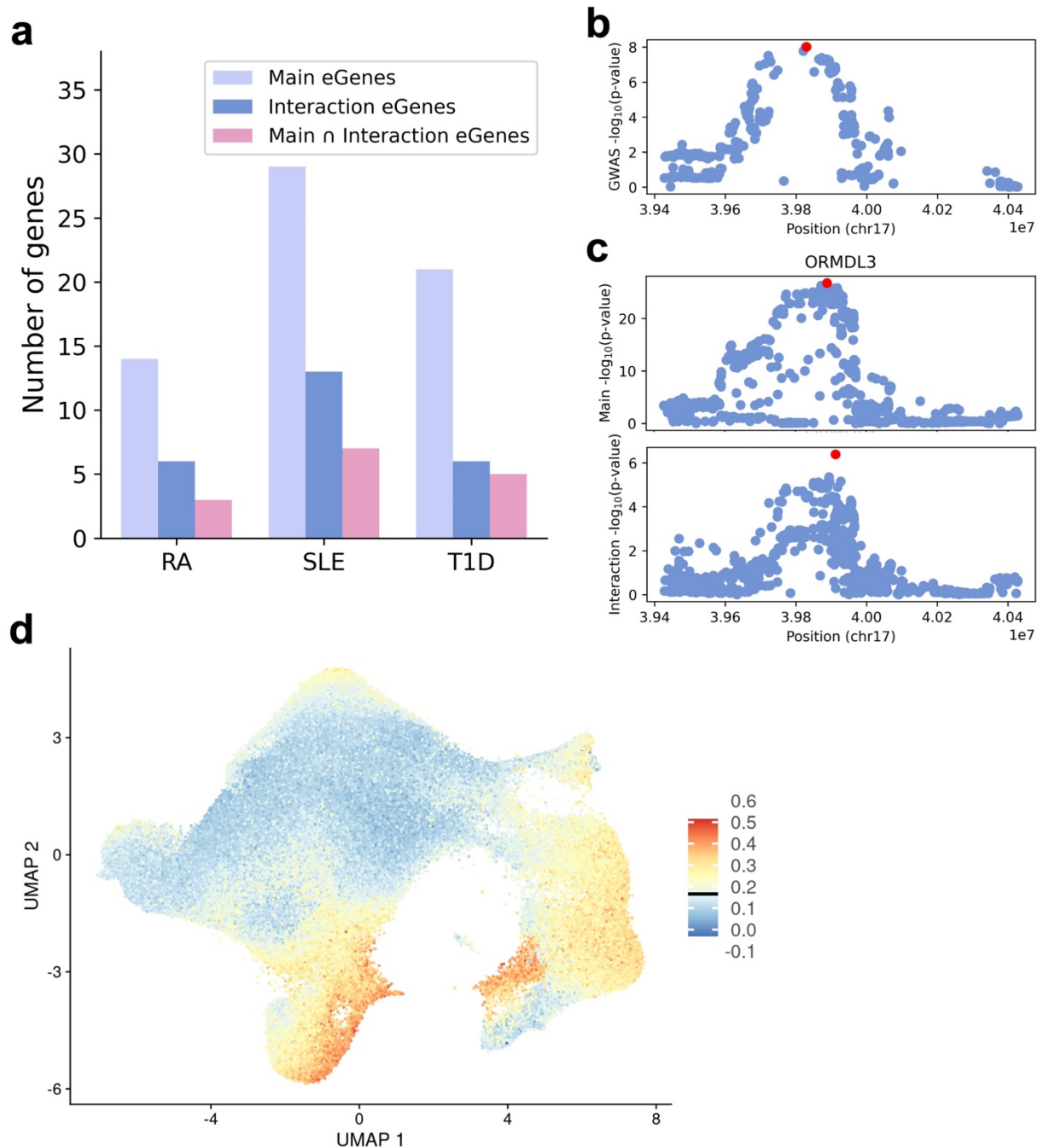


Figure 4.3. Colocalization of cell state interaction eQTL with GWAS associations. **a)** Number of genes with GWAS-colocalizing main effects, interaction effects, or interaction and main effects in each of 3 diseases analyzed (RA, rheumatoid arthritis; SLE, systemic lupus erythematosus; T1D, type 1 diabetes). **b)** Locus plot mapping the $-\log_{10}$ -transformed GWAS p-values for all variants <1 Mb from the *ORMDL3* TSS with summary statistics reported by ref.¹⁸⁷ **c)** Locus plots mapping the $-\log_{10}$ -transformed p-values for the main effect (top) and joint interaction effect (bottom) for all variants <1 Mb from the *ORMDL3* TSS. **d)** UMAP embedding of cells in the TBRU data set, colored by total effect on *ORMDL3* expression for each cell (main effect size plus interaction effect sizes multiplied by CV values for each cell). Red points in **b** and **c** highlight the variant with the most significant p-value.

Chapter 5: Conclusion

Characterizing enhancer-gene regulatory architectures is critical for understanding the biological mechanisms underlying GWAS discoveries and understanding the circuitry driving key biological processes. We have shown that combining evidence from genomic distance and single-cell methods improves power to link enhancers to genes, that enhancer correlations have a major impact on the identification of enhancer-gene links, and that dynamic eQTL are an underexplored and fruitful source of gene regulatory effects.

In Chapter 2, we developed an integrative modeling framework (pgBoost) combining single-cell peak-gene linking scores with genomic distance and genetic evidence to predict causal SNP-gene links. We demonstrated that pgBoost significantly outperforms existing single-cell methods and genomic distance in multiple evaluation analyses. In particular, pgBoost substantially outperformed existing methods in evaluations of longer-range links, which are of high biological importance in GWAS. We further demonstrated that restricting pgBoost to features from a focal cell type improves power to detect regulatory links relevant to that cell type, supporting the future application of this framework to a wider variety of biological contexts.

In Chapter 3, we demonstrated that co-accessibility among scATAC-seq peaks induces false positive *tagging* peak-gene associations in single-cell multiome data (analogous to tagging associations in GWAS). CRISPR analyses and a stratified regression of peak-gene associations on functional peak annotations both support pervasive tagging. We presented a TF-driven mechanism for peak tagging and determined that peak co-accessibility is strongly related to TF binding activity (with a particularly strong effect for pioneer TFs, which induce accessibility of closed chromatin). We showed that accounting for tagging effects through statistical fine-mapping improves peak-gene linking. We observed similar patterns when investigating gene-gene tagging effects.

In Chapter 4, we developed a statistically robust and computationally efficient method for dynamic eQTL mapping using single-cell gene expression data. We demonstrated that genes may often harbor eQTL effects that interact with cell state. These interaction effects may exist with or without a corresponding main effect, and they may or may not be independent from the main effect. There are likely many more dynamic eQTL that help explain variant-disease associations than those reported here, and it will be important to perform dynamic eQTL mapping in broader disease-related contexts.

Our work suggests several directions for future research. In the area of peak-gene linking, it will be important to apply pgBoost and other methods to other non-blood-related cell types as higher quality multiome data from such contexts become available. In addition, statistical fine-mapping (and other algorithms adapted from the GWAS toolkit in Chapter 3) can be extended to explicitly model measurement noise in single-cell data, which may further improve performance and biological insight. Importantly, strong peak-gene linking scores represent predictions and do not imply biological causality, and downstream experimental validation will be crucial. In the area of single-cell eQTL mapping, it will be important to explore larger immune cell data sets and potential meta-analyses to maximize power for detecting interaction eQTL related to immune function. It will also be of great interest to expand these analyses to diverse cell types and states, given the broad and largely unknown set of contexts relevant for human disease.

To conclude, we have developed and characterized statistical methods to link enhancers (and the regulatory variants they harbor) to their target genes, carefully navigating the challenges and leveraging the advantages of high-dimensional single-cell genomic data. Each of these efforts improves enhancer-gene maps, aids in the functional interpretation of GWAS variant-disease associations, and facilitates downstream drug target identification.

Back Matter

Chapter 2 Appendix

Methods

Existing single-cell peak-gene linking methods

We applied 4 existing single-cell peak-gene linking methods (SCENT⁸⁰, Signac^{77,81}, ArchR⁸², Cicero⁸³; **Table 2.1**) separately to each cell type in each data set (**Table 2.2**), generating 10 total sets of predictions from each method (1 for each data set-cell type pair). We selected candidate links as peak-gene pairs with an ATAC-seq peak centered $\pm 500\text{kb}$ from the gene's TSS (as in refs.^{77,80,81,83}) and not overlapping any promoter region (defined as $\pm 1\text{kb}$ from the gene's TSS as in ref.¹²⁸). Some methods implement additional requirements for candidate peak-gene pairs to test for association (for example, SCENT only tests peak-gene pairs from which the peak and gene are both active in $\geq 5\%$ of cells). To integrate linking scores across data sets and facilitate downstream variant-level analyses (e.g. interpreting GWAS variant disease associations), we intersected peaks with 1000 Genomes Project variants with minor allele count ≥ 5 (following refs.^{25,26,86,196,197}), to define SNP-gene linking scores from peak-gene linking scores; we consider SNP-gene pairs for which the focal SNP lies in any peak centered $\pm 500\text{kb}$ from the gene's TSS (we do not restrict to peaks tested by single-cell peak-gene linking methods). We also restricted all analyses to the set of autosomal genes analyzed in ref.²⁶ (see Data Availability).

Brief descriptions of each linking method are provided below (additional implementation details are provided in the **Supplementary Note**):

SCENT⁸⁰ (v0.1.0) uses Poisson regression to test the association of each candidate peak-gene pair (while controlling for batch and cell-level covariates) and quantifies the significance of associations by bootstrapping across cells.

Signac^{77,81} (v1.10.0) computes the correlation between binarized peak accessibility and normalized gene expression across single cells for each candidate peak-gene pair and quantifies significance by permuting across off-chromosome peaks matched to the focal peak for peak accessibility, peak length, and GC content.

ArchR⁸² (v1.0.2) defines metacells (low-level cell aggregates) using k -nearest neighbors, aggregates RNA and ATAC peak counts across single cells within each metacell, and computes the correlation between normalized aggregated RNA and ATAC peak counts across metacells. ArchR outputs an FDR computed from the analytical correlation p-value.

Cicero⁸³ (v1.3.9), which uses as input scATAC-seq data only, computes partial correlations across metacells between pairs of ATAC peaks, quantifying the association between two peaks after accounting for accessibility of all other peaks ± 500 kb. Cicero uses the Graphical LASSO¹⁶³ with a distance-dependent penalty term to prioritize local *cis*-regulatory interactions. Our results show that this penalty term does not effectively capture the impact of genomic distance (**Figure 2.3** and **Figure 2.4**). These peak-peak associations can be interpreted as peak-gene associations when one peak overlaps a gene promoter (we define the promoter region as ± 1 kb from the gene's TSS as in ref.¹²⁸).

pgBoost framework

pgBoost is a gradient boosting classification model (implemented using XGBoost software¹⁰⁵) which takes as input (features) linking scores from existing methods and genomic distance and outputs (predictions) probabilistic scores for each pair.

The gradient boosting algorithm iteratively builds regression trees to capture relationships between features x_i and observed “truth” values y_i , minimizing the following objective function at each iteration t :

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \quad [2.1]$$

where $l(y_i, \hat{y}_i)$ is a loss function measuring the difference between “truth” (y_i) and prediction ($\hat{y}_i$) for instance i , $f_t(x_i)$ is a regression tree fit to the errors from the previous iteration, and $\Omega(f_t)$ is a

regularization term which prevents the model from overfitting to the training data. The model predicts the causal probability of candidate links by summing over tree outputs: $\hat{y}_i = \sum_{t=1}^T f_t(x_i)$. The regularization term $\Omega(f_t)$ can be expressed as a sum of regularization costs across trees at each iteration t :

$$\Omega(f_t) = \sum_{i=1}^t \omega(f_i) \quad [2.2]$$

where ω denotes the “complexity” of tree f , defined as:

$$\omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad [2.3]$$

where γ is the “gamma” parameter, T denotes the number of leaves, λ is an additional L2 regularization parameter, and $\sum_{j=1}^T w_j^2$ denotes the sum of squared prediction scores within each leaf. The “gamma” parameter prevents overfitting by specifying a minimum reduction in loss required to make an additional split, thus penalizing more complex trees. We report gradient boosting hyperparameters used in the **Supplementary Note**.

Across candidate SNP-gene links, pgBoost considers 3 separate features for each cell subset (one cell type in one data set)—SCENT regression effect size, Signac correlation, Cicero peak-promoter co-accessibility—as well as 2 distance-based features: SNP-TSS distance, and a binary indicator for “closest TSS”. A feature matrix of 6,882,594 candidate SNP-gene links x 32 features was thus defined for candidate SNP-gene links profiled in any of the 4 multiome data sets (**Table 2.2**). If a candidate link was not scored by a given method in a given cell subset, an NA was entered as the corresponding feature. (We note that the main pgBoost model excludes features from ArchR, which requires separate processing of raw fragment/BAM files; see *pgBoost*

feature selection and evaluation of feature importance). Further details on pgBoost feature selection and evaluation of feature importance are provided in the **Supplementary Note**.

We defined a training set comprised of fine-mapped eSNP-eGene pairs attaining a posterior inclusion probability (PIP) > 0.2 in any GTEx tissue as positives and maximum PIP < 0.01 across all GTEx tissues as negatives^{63,64} (**Supplementary Figure 2.1**). In detail, for each SNP-gene pair, we computed the fine-mapped PIP in each GTEx tissue and determined the maximum PIP attained across all GTEx tissues. We then defined positive SNP-gene pairs as those attaining maximum PIP > 0.2 across tissues and negative SNP-gene pairs as those attaining maximum PIP < 0.01 across tissues. We selected negative SNP-gene pairs by restricting to genes that included a positive SNP-gene pair in order to avoid uninformative (and potentially false negative) SNP-gene pairs that might arise due to gene-specific factors not captured by pgBoost model features (e.g. read mappability). We intersected this set of SNP-gene pairs with candidate SNP-gene links profiled in any of the 4 multiome data sets (**Table 2.2**) and removed any “duplicate” SNP-gene pairs receiving the same scores across data set-cell type pairs (intuitively, limiting to 1 entry per uniquely-defined regulatory element). In total, pgBoost was trained on 11,926 positive SNP-gene pairs and 99,772 negative SNP-gene pairs.

Single-cell data sets

We downloaded 4 publicly available scRNA/ATAC-seq multiome data sets (10x PBMC¹⁰⁸, Luecken BMMC¹⁰⁹, SHARE-seq⁷⁷, and Xu K562¹¹⁰; **Table 2.2**) (see Data Availability) that span a large number of cells and reflect immune/blood cell biology. For each data set, we analyzed major cell types spanning >500 cells (in accordance with ref.⁸⁰), yielding 10 data set-cell type pairs. (No statistical methods were used to pre-determine the number of data sets or number of data set-cell type pairs analyzed). We also downloaded 1 publicly available scRNA/ATAC-seq multiome data set from brain tissue (Meijer brain¹¹³). We obtained publicly available QCed gene expression and ATAC peak matrices for the Luecken BMMC¹⁰⁹ and Meijer brain¹¹³ data sets, and processed the 10x PBMC¹⁰⁸, SHARE-seq⁷⁷, and Xu K562¹¹⁰ data sets using both Seurat 4.3.0.1¹⁹⁸ and ArchR

1.0.2⁸² (ArchR requires raw fragment/BAM files as input and thus separate processing). We retained all cells from each data set passing QC filters on total ATAC and RNA reads, nucleosome signal, and TSS enrichment in accordance with ref.⁸⁰ We next identified “peaks” of open chromatin in each data set using macs2¹⁹⁹ and excluded non-standard chromosomes and blacklist regions. To ensure uniform cell and peak sets were used by all methods, we imported the set of ATAC peaks identified using macs2 into ArchR using the addPeakSet function.

These steps yielded matrices of gene expression and ATAC peak counts in individual cells for each data set. We also generated binarized ATAC peak matrices, in which an entry of 1 for a given peak-cell pair indicates that ≥ 1 ATAC fragments overlap the peak and an entry of 0 indicates that no ATAC fragments overlap the peak. Finally, we added cell type annotations. For the Luecken BMMC¹⁰⁹, SHARE-seq⁷⁷, Xu K562¹¹⁰, and Meijer brain¹¹³ data sets, we used cell type labels provided by the authors. For the 10x PBMC¹⁰⁸ data set, we performed Seurat reference-query mapping with a PBMC reference object to define cell type labels, as described in the “Multimodal reference mapping” vignette at https://satijalab.org/seurat/articles/multimodal_reference_mapping.html. In detail, we normalized and scaled 10x PBMC gene expression data using the Seurat SCTransform function¹⁶⁶, downloaded a PBMC reference data set (previously annotated with cell type labels and normalized using SCTransform) from ref.¹⁹⁸, then assigned cell type labels to the 10x PBMC data set by projecting the PCA structure of the reference data set onto the 10x PBMC data set.

Defining single-cell and distance-based SNP-gene linking scores

Running existing peak-gene linking methods separately in each of the 10 data set-cell type pairs (**Table 2.2**) typically yielded multiple scores per method for a single candidate SNP-gene link. To facilitate downstream analyses, we assigned a single score per method to each candidate link in accordance with ref.⁸⁰ For SCENT, we assigned each candidate SNP-gene link the minimum FDR observed across cell subsets. For Signac, we assigned each candidate SNP-gene link the maximum |score| across cell subsets. For ArchR, we assigned each candidate SNP-gene

link the maximum |correlation| across cell subsets. For Cicero, we assigned each candidate SNP-gene link the maximum |co-accessibility| across cell subsets.

To compare the performance of pgBoost and existing methods to a distance-based strategy, we defined a distance-based linking score by ranking all candidate SNP-gene links by SNP-TSS distance (such that more proximal SNP-gene links are assigned higher linking scores).

Evaluation data sets

We defined 4 main evaluation sets derived from eQTL^{63,64}, Activity-By-Contact (ABC)^{26,27}, CRISPR^{47–55}, and GWAS^{106,107} (see Data Availability). We intersected all evaluation sets with candidate SNP-gene links (see *Existing single-cell peak-gene linking methods*).

Brief descriptions of each evaluation set are provided below:

The eQTL evaluation set is comprised of 4,434 fine-mapped eSNP-eGene pairs attaining a posterior inclusion probability (PIP) > 0.5 in any GTEx tissue^{63,64}. Fine-mapping was conducted by ref.⁶⁴ (by applying the SuSiE¹¹⁴ fine-mapping method to 49 GTEx tissues⁶³).

The Activity-By-Contact evaluation set is comprised of 53,701 SNP-gene links attaining ABC score > 0.2 in any of 344 biosamples passing QC^{26,27}. Activity-By-Contact scores²⁷ were computed and made publicly available by ref.²⁶ (344 biosamples were obtained by removing duplicate assays corresponding to same cell type from the 352 biosamples released by ref.²⁶).

The CRISPR evaluation set is comprised of 892 SNP-gene links experimentally validated by CRISPR interference experiments^{47–55}. Validated links are defined as those marked “TRUE” in the “Regulated” column of the combined CRISPR evaluation data set (see Data Availability). We derived 892 CRISPR-validated SNP-gene links from these 472 element-gene links based on SNPs that reside in a regulatory element. CRISPR results were assembled and made publicly available by ref.²⁶

The GWAS-derived evaluation set is comprised of 155 SNP-gene pairs defined by non-coding variants fine-mapped to a focal trait (PIP > 0.1) with a coding variant for exactly one candidate gene $\pm$ 1 Mb from the non-coding variant attaining PIP > 0.5 for the same trait. This

evaluation set was defined by ref.¹⁰⁶ (based on fine-mapping results from ref.¹⁰⁷ on 94 diseases and complex traits from the UK Biobank¹¹²).

Average enrichment metric

We used an enrichment-based metric, *average enrichment across recall values*, to assess the ability of each linking method to prioritize (assign higher linking scores to) “true” links over “false” links (as defined by an evaluation set of SNP-gene links).

For a given evaluation set and given method, we ranked candidate SNP-gene links by linking score. At each unique linking score c assigned to any candidate SNP-gene link, we computed enrichment and recall as follows:

$$\text{Enrichment} = \frac{|\text{predicted SNP-gene links} \cap \text{true SNP-gene links}| / |\text{true SNP-gene links}|}{|\text{predicted SNP-gene links}| / |\text{candidate SNP-gene links}|} \quad [2.4]$$

$$\text{Recall} = \frac{|\text{predicted SNP-gene links} \cap \text{true SNP-gene links}|}{|\text{true SNP-gene links}|} \quad [2.5]$$

where “predicted SNP-gene links” denotes the binarized set of SNP-gene links attaining linking score $\geq c$, “true SNP-gene links” denotes the binarized set of “true” SNP-gene links defined by the evaluation set (intersected with the set of “candidate SNP-gene links”), and “candidate SNP-gene links” denotes SNP-gene pairs profiled in at least one single-cell multiome data set (involving a SNP lying in an ATAC-seq peak and a gene expressed in the focal data set; **Table 2.2**). When computing enrichment for evaluation sets derived from CRISPR^{47–55} and GWAS^{106,107} we further restricted the set of “candidate SNP-gene links” to SNP-gene pairs tested by a CRISPR experiment^{47–55} (for CRISPR evaluation) or SNP-gene pairs involving a SNP with a “true” GWAS-derived link (for GWAS evaluation). We did not further restrict the set of “candidate SNP-gene links” for evaluation sets derived from eQTL^{63,64} and ABC^{26,27}, as these methods can detect SNP-gene associations genome-wide. Intuitively, equation 4.4 quantifies the extent of overlap between

predicted SNP-gene links and evaluation SNP-gene links relative to the overlap expected by chance.

For a given evaluation set and set of methods, we used equation 4.4 and equation 4.5 to construct enrichment-recall curves. For each method, we considered different values of the number of evaluation links prioritized (k), corresponding to a recall and enrichment of $(\text{recall}_k, \text{enrichment}_k)$, and performed linear extrapolation to complete the enrichment-recall curve between the points $(\text{recall}_k, \text{enrichment}_k)$ and $(1, 1)$, such that the enrichment-recall curve spanned values of recall in the range $[0, 1]$. We then defined *average enrichment across recall values* (or *average enrichment*, shortened) as follows:

$$\text{Average enrichment} = (\sum_{i=1}^n \text{enrichment}_i * (1 - \text{recall}_i)) / n \quad [2.6]$$

where n denotes the total number of evaluation links and $(\text{recall}_i, \text{enrichment}_i)$ for $i \in \{1 \dots n\}$ denote the values of each point on the enrichment-recall curve. Multiplying enrichment_i by $(1 - \text{recall}_i)$ upweights enrichments at lower recall values, prioritizing method performance at high score thresholds. Intuitively, this metric can be interpreted as a weighted area under the enrichment-recall curve.

We obtained standard errors on average enrichment for each method, as well as standard errors and p-values on the difference in average enrichment between two methods, by bootstrapping genes using 1,000 bootstrap iterations. (The distribution of enrichment differences was assumed to be normal, but this was not formally tested.)

Training pgBoost models in a focal cell type

To investigate whether restricting model features to a focal cell type improves power to detect SNP-gene links relevant to that cell type, we trained individual pgBoost models and generated predictions for T cells, B cells, myeloid cells, and erythroid cells in the Luecken

BMMC¹⁰⁹ multiome data set (each spanning > 8,000 cells), K562 cells from the Xu¹¹⁰ data set, and oligodendroglia from the Meijer brain¹¹³ data set (**Table 2.2**). We trained pgBoost models in a focal cell type by restricting model features to the 3 linking scores derived from the focal cell type plus 2 distance-based features, totaling 5 features. We used fine-mapped eSNP-eGene pairs (maximum PIP > 0.2 vs. maximum PIP < 0.01 across GTEx tissues, see *pgBoost framework* above) as training data.

ABC evaluation data in a focal cell type or tissue

We defined evaluation SNP-gene links relevant to each of the 4 focal cell types (T cells, B cells, myeloid cells, and erythroid cells; see *Training pgBoost models in a focal cell type*) from ABC data^{26,27}. We first assigned 344 biosamples profiled by ABC data^{26,27} to either 1 focal cell type (if a biosample is closely biologically related to that cell type) or 0 focal cell types (if a biosample is not related to any of the 4 focal cell types), yielding 53 biosamples related to T cells, 14 biosamples related to B cells, 6 biosamples related to myeloid cells, and 1 biosample related to K562 cells. We then defined 4 *cell-type-level* sets of SNP-gene links strongly implicated in any biosample related to each focal cell type ($\max(\text{ABC}) > 0.2$ across biosamples). We finally defined *cell-type-specific* sets of SNP-gene links as the set of *cell-type-level* SNP-gene links in each focal cell type that are not as strongly implicated in the 3 non-focal cell types (defining “not as strongly implicated” via $\max(\text{ABC}) < 0.1$ across biosamples). We compared the performance of pgBoost models restricted to T cell features, B cell features, myeloid features, erythroid features, or K562 features (see *Training pgBoost models in a focal cell type*) on cell-type-specific and cell-type-level ABC evaluation data sets.

We also defined evaluation SNP-gene links relevant to each of 2 focal tissues (blood, brain; see *Training pgBoost models in a focal cell type*) from ABC data^{26,27}. We first assigned 344 biosamples profiled by ABC data^{26,27} to either 1 focal tissue (if a biosample is closely biologically related to that tissue) or 0 focal tissues (if a biosample is not related to either focal tissue), yielding 74 biosamples related to blood and 26 biosamples related to brain. We then defined 2 *tissue-level*

sets of SNP-gene links strongly implicated in any biosample related to each focal tissue ($\max(\text{ABC}) > 0.2$ across biosamples). We finally defined *tissue-specific* sets of SNP-gene links as the set of *tissue-level* SNP-gene links in each focal tissue that are not as strongly implicated in the non-focal tissue (defining “not as strongly implicated” via $\max(\text{ABC}) < 0.1$ across biosamples). We compared the performance of pgBoost models restricted to K562 features or brain oligodendroglia features (see *Training pgBoost models in a focal cell type*) on tissue-specific and tissue-level ABC evaluation data sets. We selected K562 as the blood-related cell type in these analyses (instead of T/B/myeloid/erythroid cells) because the K562 data set¹¹⁰ is most similar to the oligodendroglia data set¹¹³ in number of cells (8k cells for K562, vs. 3k cells for oligodendroglia and 9-23k cells for T/B/myeloid/erythroid cells) and number of ABC biosamples derived from the focal cell type (1 biosample for K562, vs. 0 biosamples for oligodendroglia and 6-53 biosamples for T/B/myeloid/erythroid cells).

Because cell-type-specific evaluation data is often limited, we used all evaluation SNP-gene links (distance $> 1\text{kb}$) in primary analyses to maximize power; we consider other distance thresholds in secondary analyses.

Supplementary Note

Additional secondary analyses using eQTL and ABC evaluation data

We performed 9 additional secondary analyses involving eQTL and ABC evaluation data. First, we compared the performance of pgBoost and existing methods on eQTL and ABC evaluation data using secondary metrics area under the precision-recall curve (AUPRC), area under the receiver operating characteristic curve (AUROC), and average enrichment across recall values without linear extrapolation or weighting (see below), each stratified by distance threshold. These analyses produced similar findings to our main enrichment metric (**Supplementary Figure 2.5a-b**, **Supplementary Figure 2.6a-b**, **Supplementary Figure 2.7a-b**). Second, we assessed the enrichment of the sets of SNP-gene links scored by each method. We determined that, for eQTL and ABC evaluations, differences in performance of existing methods were partially driven by the subset of candidate links scored by each method, particularly for SCENT (**Supplementary Figure 2.9a-b**). SCENT employs stringent criteria for selecting candidate peak-gene links to test (**Table 2.1**), thereby pre-selecting a smaller set of scored links with higher enrichments for eQTL and ABC evaluations. Third, we assessed how pgBoost scores vary with SNP-gene distance and determined that pgBoost scores generally decrease with distance, even at larger distances (**Supplementary Figure 2.10**). Fourth, we compared the performance of pgBoost and existing single-cell methods on an evaluation set of 41,029 SNP-gene links implicated by promoter capture Hi-C²⁰⁰. We determined that pgBoost significantly outperformed existing single-cell methods but significantly underperformed the distance-based method across distance thresholds (**Supplementary Figure 2.11**); we note that Hi-C contact frequencies are largely determined by genomic distance^{52,111,201} (even more so than ABC; **Supplementary Figure 2.8**) and are thus expected to favor the distance-based method. Fifth, we compared pgBoost to an ABC-based linking strategy (ranking candidate SNP-gene links by maximum ABC score across 344

biosamples^{26,27}) and determined that pgBoost significantly outperformed ABC on eQTL evaluation data (**Supplementary Figure 2.12a**). Sixth, we assessed the sensitivity of pgBoost and the 4 existing scores to variation in cell count and sequencing depth by downsampling to 50% of the input data (see *Methods: Downsampling analyses varying cell count and sequencing depth* below). We observed that the performance of pgBoost and the 4 existing scores in eQTL evaluation data generally does not significantly differ between 2-fold downsampled and non-downsampled conditions (**Supplementary Figure 2.13**). Seventh, we assessed the calibration of the pgBoost probabilistic score. The relationship between the proportion of predicted links that were concordantly predicted by eQTL or ABC vs. the pgBoost score was monotonic but not perfectly linear, supporting the prioritization of top-scoring links rather than probabilistic scores in enrichment computations and downstream analyses (**Supplementary Figure 2.14**). Eighth, we assessed the performance of pgBoost trained using alternative gradient boosting hyperparameter values. We determined that pgBoost performance was generally robust to alternative values of fixed hyperparameters, and our approach for selecting values of the remaining hyperparameters performed at least as well as alternative approaches (**Supplementary Figure 2.15**; see *Methods: Gradient boosting parameters* below). Ninth, we assessed the computational cost of pgBoost and determined that the runtime and memory requirements of training and inference are generally modest (e.g. <2 hours and <15G of memory for the data sets analyzed).

Evaluations of alternative training data sets

We performed 5 additional secondary analyses evaluating the performance of pgBoost trained on alternative training data sets, using the same set of input features; as noted above, each link set that we analyze can be considered as a prediction, used for model training, or used for evaluation. First, we trained pgBoost using ABC links (ABC score > 0.2 in at least one of 344 biosamples^{26,27}). Training using eQTL and ABC links attained the highest enrichments for eQTL and ABC evaluations, respectively, but training on eQTL links significantly outperformed training on ABC links for CRISPR and GWAS evaluations (**Supplementary Figure 2.21**). Second, we

trained pgBoost using fine-mapped eSNP-eGene pairs from alternative eQTL catalogs instead of GTEx, including eQTLGen¹⁷³ or the union of GTEx and eQTLGen fine-mapped eSNP-eGene pairs (see Data Availability). We determined that the GTEx training set outperformed the eQTLGen training set and performed similarly to the union of training sets, supporting our primary use of the GTEx training set (**Supplementary Figure 2.22**). Because restricting training data to eQTLs in whole blood (from either eQTLGen or GTEx) significantly underperformed training using eQTL from all GTEx tissues, we did not consider more granular cell-type-specific eQTL catalogs (e.g. OneK1K²⁰²). Third, we trained pgBoost using random subsamples of fine-mapped eSNP-eGene pairs from our primary GTEx training set. We determined that mild downsampling did not substantially impact performance, suggesting that there may be diminishing returns from incorporating additional training data (**Supplementary Figure 2.23**). Fourth, we trained pgBoost using fine-mapped eSNP-eGene pairs at other PIP thresholds, instead of $PIP > 0.2$. We determined that more stringent PIP thresholds yielded smaller training data sets and lower pgBoost enrichments for eQTL, ABC, CRISPR, and GWAS evaluations (**Supplementary Figure 2.24**). We did not train pgBoost using the set of gold-standard SNP-gene links identified by GWAS, as this set contained only 155 links. Fifth, we trained pgBoost using fine-mapped eQTL spanning all SNPs in GTEx, instead of restricting to SNPs with minor allele count ≥ 5 in 1000 Genomes. We determined that including all GTEx SNPs yielded a slightly larger training set but did not impact pgBoost performance on eQTL, ABC, CRISPR, and GWAS evaluations (**Supplementary Figure 2.25**).

Evaluations of alternative feature sets

To investigate which features contributed most to the classification of SNP-gene links, we performed 5 secondary analyses comparing the performance of pgBoost using different feature sets. First, we compared pgBoost to restricted gradient boosting models using a subset of linking scores as features. We determined that pgBoost significantly outperformed models restricted to a single existing score or distance (**Extended Data Figure 2.5**) or a single existing score plus

distance (**Figure 3b, Figure 3d, Figure 4b, Figure 4d**), for various subsets of evaluation sets. We further determined that models ablating a single constituent score or distance generally performed significantly worse than pgBoost across the 4 evaluation sets, particularly when ablating distance (**Extended Data Figure 2.6**). Second, we compared pgBoost to expanded models including features from additional constituent methods. We determined that the model adding features from ArchR did not significantly outperform pgBoost on eQTL, ABC, CRISPR, and GWAS evaluations (**Extended Data Figure 2.6**). We also considered including features from SCARlink⁸⁶, a recently published method for linking regulatory elements to genes using scRNA/ATAC-seq multiome data. We analyzed SCARlink predictions on the 10x PBMC¹⁰⁸ data set (**Table 2.2**), restricted to ATAC peaks (**Supplementary Figure 2.26**). We first determined that SCARlink generally performed similarly to the 4 other existing methods (**Supplementary Figure 2.27**). We then compared pgBoost to an expanded model including features from SCARlink and determined that the model including SCARlink features did not significantly outperform pgBoost (**Supplementary Figure 2.28**). Third, we compared pgBoost to expanded models including p-values from SCENT and Signac as features (we note that Cicero does not output p-values). We determined that the model adding significance features did not significantly outperform pgBoost (**Supplementary Figure 2.29**). Fourth, we compared pgBoost to restricted gradient boosting models using linking scores from a subset of the 4 multiome data sets. We determined that pgBoost significantly outperformed models restricted to a single multiome data set (**Supplementary Figure 2.30**) and models ablating a single multiome data set (**Supplementary Figure 2.31**). Fifth, we conducted a SHAP²⁰³ analysis of feature importance. As expected, we determined that genomic distance attained the highest feature importance, followed by the “closest TSS” binary indicator, but features based on constituent scores were also broadly important (**Supplementary Figure 2.32**); we caution that SHAP analysis of feature importance can be misleading due to correlation among pgBoost features (**Figure 1a**).

Additional secondary analyses using CRISPR and GWAS evaluation data

We performed 7 additional secondary analyses involving CRISPR and GWAS evaluation data. First, we compared the performance of pgBoost and existing methods on CRISPR and GWAS evaluation data using secondary metrics area under the precision-recall curve (AUPRC), area under the receiver operating characteristic curve (AUROC), and average enrichment across recall values without linear extrapolation or weighting (see below), each stratified by distance threshold. These analyses produced similar findings to our main enrichment metric (**Supplementary Figure 2.5c-d, Supplementary Figure 2.6c-d, Supplementary Figure 2.7c-d**). Second, we compared pgBoost to the ENCODE-rE2G method²⁶, a logistic regression model using chromatin activity and 3D interactions as features and using CRISPRi data^{47–55} from K562 cells to guide training (see Data Availability). Specifically, we ranked candidate SNP-gene links by maximum ENCODE-rE2G score across 344 biosamples and performed evaluations on eQTL, CRISPR, and GWAS data (we exclude ABC evaluation data from these comparisons, because the ABC evaluation data were used as features to generate ENCODE-rE2G predictions). We determined that pgBoost significantly outperformed ENCODE-rE2G on eQTL evaluation data (**Extended Data Figure 2.7a**) and ENCODE-rE2G significantly outperformed pgBoost on CRISPR evaluation data (**Extended Data Figure 2.7b**), consistent with the training data used by each method. In addition, pgBoost significantly outperformed ENCODE-rE2G on longer-range (>50kb and >100kb) GWAS evaluation data (**Extended Data Figure 2.7c**). We also compared pgBoost restricted to features from the Xu K562²⁰⁴ data set to ENCODE-rE2G scores in K562 and reached similar conclusions, when well-powered to detect a difference (**Supplementary Figure 2.33**). Third, we compared pgBoost to a logistic regression model using identical features (see above) and determined that gradient boosting significantly outperformed logistic regression on longer-range (>50kb) CRISPR evaluation data but performed similarly on GWAS evaluation data (**Extended Data Figure 2.1c-d**). Fourth, we compared pgBoost to an eQTL-based linking strategy (see above) and determined that pgBoost significantly outperformed eQTL on CRISPR and GWAS evaluation data (**Extended Data Figure 2.2b-c**). Fifth, we compared pgBoost to an ABC-

based linking strategy (see above) and determined that pgBoost significantly outperformed ABC on CRISPR and GWAS evaluation data (**Supplementary Figure 2.12b-c**). Sixth, we compared gradient boosting models restricted to features from one existing method (across data sets and cell types) to the maximum score of the same existing method (across data sets and cell types). We determined that these reduced gradient boosting models sometimes significantly outperformed the maximum score of the existing method across data sets (**Supplementary Figure 2.34**), but the main pgBoost model still significantly outperformed these reduced models (**Extended Data Figure 2.5**). Seventh, we assessed the enrichment of the sets of SNP-gene links scored by each method. We determined that, for CRISPR and GWAS evaluations, differences in performance of existing methods were partially driven by the subset of candidate links scored by each method (but less so than for eQTL and ABC evaluations, because CRISPR and GWAS enrichments are computed against restricted backgrounds of candidate links tested by CRISPR and candidate links including a variant with a gold-standard GWAS link, respectively; see **Methods**) (**Supplementary Figure 2.9c-d**).

Additional secondary analyses of pgBoost performance in a focal cell type

We performed 5 additional secondary analyses involving pgBoost restricted to a focal cell type. First, we analyzed tissue-specific GTEx evaluation data, defined by fine-mapped eSNP-eGene pairs identified in brain or blood tissue ($PIP > 0.5$) and not strongly implicated in the other tissue ($PIP < 0.1$). On both evaluation sets, the pgBoost model spanning both K562 and brain oligodendroglia features and the model restricted to brain oligodendroglia features performed similarly, and both models significantly outperformed the model restricted to K562 features (**Supplementary Figure 2.40**). The underperformance of the model restricted to K562 features on eQTL evaluation data sets from whole blood may reflect more pronounced differences between K562 (a cultured cell line) and *in vivo* human blood regulatory circuitry at the level of downstream regulatory effects (e.g. eQTL) vs. upstream molecular interactions (e.g. ABC). Second, instead of restricting features to cell type X for training and prediction and evaluating in

cell type/tissue Y in cell-type-specific ABC evaluation data (**Figure 5a**, left panel) or tissue-specific ABC evaluation data (**Figure 5a**, right panel), we restricted features to cell type X for training, used features in cell type/tissue Y for prediction, and evaluated in cell type/tissue Y in cell-type-specific or tissue-specific ABC evaluation data. We determined that the cell type used for training no longer significantly impacts performance (**Supplementary Figure 2.41**), suggesting that the cell-type-specific patterns of enrichment in **Figure 5a** were driven by the use of the focal cell type for prediction rather than the use of the focal cell type for training. Third, we trained pgBoost using fine-mapped eSNP-eGene pairs from blood or brain tissue instead of all GTEx tissues and evaluated performance on tissue-specific ABC evaluation data. Across blood and brain evaluation sets, the pgBoost model restricted to features related to the focal tissue (K562 or brain oligodendroglia, respectively) again significantly outperformed the model restricted to features related to the non-focal tissue (**Supplementary Figure 2.42**). Fourth, we generated independent pgBoost predictions for each of the 3 major cell types shared by the Luecken BMMC¹⁰⁹ and 10x PBMC¹⁰⁸ multiome data sets (T cell, B cell, myeloid; **Table 2.2**) by restricting model features to linking scores from both data sets for each cell type (as well as the 2 distance-based features). The pgBoost model restricted to myeloid features significantly outperformed the other two models and the model spanning the union of all 3 cell types on myeloid-specific ABC evaluation data; models restricted to T cells or B cells attained similar enrichments and significantly outperformed the myeloid model and the model spanning the union of all 3 cell types in T cell-specific and B cell-specific ABC evaluation data (**Supplementary Figure 2.43**). Fifth, we analyzed cell-type-specific or tissue-specific ABC evaluation data defined by SNP-gene links identified in biosample(s) related to the focal cell type or tissue (ABC score > 0.2) but below a more stringent threshold in non-focal cell types or tissues (ABC score < 0.05) and observed qualitatively similar patterns to **Figure 5a** (**Supplementary Figure 2.44**). We did not consider evaluation sets of SNP-gene links implicated by fine-mapped non-coding and coding variants in GWAS^{106,107} (**Figure 4c**)

restricted to sets of traits most relevant to a focal cell type(s), as this would result in prohibitively small evaluation sets.

Additional examples of pgBoost-predicted SNP-gene links

pgBoost linked rs78507984 (fine-mapped for uric acid with PIP = 1.00^{107,112,114}) to *SLC2A9* (12kb distance to TSS; closest TSS) (**Supplementary Figure 2.45a**). This is considered a gold-standard SNP-gene link, based on an independent *SLC2A9* GWAS coding variant that was fine-mapped for uric acid^{106,107}. *SLC2A9* encodes an enzyme critical for the reabsorption of uric acid and glucose from the kidneys into the bloodstream^{205,206}, and is thus a biologically plausible target gene. Dysregulation of *SLC2A9* has been implicated in hyperuricemia²⁰⁷ and gout²⁰⁵. This link also scored in the top 5% of links predicted by the distance-based method, but scored outside the top 5% of links predicted by other single-cell methods.

pgBoost linked rs6116189 to *RNF24* (67kb distance to TSS, 2nd closest TSS) (**Supplementary Figure 2.45b**). rs6116189 lies in an enhancer of *RNF24* experimentally validated by CRISPR interference⁴¹. Knockout of *RNF24* resulted in decreased hemoglobin content and erythrocyte cell number in mice^{208,209}, suggesting that this gene may play an important role in red blood cell function; however, rs6116189 was not fine-mapped for any of the 94 diseases/traits analyzed. This link also scored in the top 5% of links predicted by Signac, SCENT, and ArchR, but scored outside of the top 5% of links predicted by the distance-based method and Cicero. Existing single-cell methods assigned highest priority to other genes (*CDC25B*, *PANK2*) and additionally prioritized other genes in the top 5% of links (*SIGLEC1*, *C20orf7*, *SPEF1*, *AP5S1*, *MAVS*, *SMOX*).

pgBoost linked rs7665090 to *NFKB1* (130kb distance to TSS; closest TSS) (**Supplementary Figure 2.45c**). rs7665090 is a known risk variant for multiple sclerosis (MS)²¹⁰; however, rs7665090 was not fine-mapped for any of the 94 diseases/traits analyzed. *NFKB1* encodes a subunit of the NF- κ B transcription factor, which has been implicated in MS^{210,211} and other autoimmune diseases²¹². While this link has not been tested via CRISPR (to our

knowledge), previous studies have reported an association between rs7665090 and *NFKB1* activation^{213,214}; thus, *NFKB1* is a biologically plausible target gene. Notably, rs7665090 is approximately equidistant to *MANBA* (131kb distance to TSS; 1.0kb distance to TES vs. 13kb for *NFKB1*). A recent study also reported an association of rs7665090 with *MANBA*, a lysosomal enzyme primarily studied for its role in kidney disease^{215–217} and rare lysosomal disorders²¹⁸, in specific cell types and stimulatory conditions²¹⁹. While ref.²¹⁴ did not observe this association, it is possible that pgBoost failed to detect a true link between rs7665090 and *MANBA* if the relevant biological context was not represented in our multiome data (see below). This link also scored in the top 5% of links predicted by SCENT and Signac, but scored outside of the top 5% of links predicted by the distance-based method and other existing methods.

Additional limitations and directions for future research

Our work has several additional limitations and directions for future research. First, because pgBoost uses as features linking scores computed from ATAC peaks, predictions are largely limited to SNPs lying in regions of accessible chromatin. However, chromatin accessibility is a known predictor of regulatory activity, making ATAC peaks a well-supported prior for regulatory regions^{20,27,220}. In addition, it would be straightforward to apply pgBoost using a genomic tiling approach to define candidate regulatory regions instead of ATAC peaks (although we anticipate limited benefit from this approach; **Supplementary Figure 2.26**). We note that we have performed our analyses at the level of SNP-gene links, but users can also generate pgBoost predictions for peak-gene links by computing constituent peak-gene linking scores using peaks that are merged across data sets and defining positive and negative training sets of peak-gene links based on the maximum PIP (of fine-mapped eSNP-eGene pairs) across SNPs within a focal peak. Second, pgBoost may fail to identify links specific to rare cell types or disease contexts. Indeed, pgBoost results are a function of the multiome data sets included as features and cannot be perfectly reproduced when different multiome data sets are included as features (thus, different pgBoost models are recommended for predicting SNP-gene links in different cell types).

Encouragingly, pgBoost identifies links specific to blood/immune or brain cell types when including fine-mapped eSNP-eGene pairs from all GTEx tissues in model training (**Figure 5**). However, it will be important to investigate pgBoost performance in rare contexts as more multiome data from such contexts becomes available. Third, pgBoost does not exhaustively include all single-cell peak-gene linking methods^{77,80–86}. However, our analyses suggest that there will be diminishing returns from including additional single-cell peak-gene linking methods; indeed, we determined that including ArchR⁸² or SCARlink⁸⁶ scores as features did not significantly improve pgBoost performance relative to the main pgBoost model incorporating 3 peak-gene linking methods^{77,80,81,83} (**Extended Data Figure 2.6, Supplementary Figure 2.28**). Fourth, our default set of positive SNP-gene links predicted by pgBoost (top 5% of linking scores across all candidate links at distance >1kb) prioritizes only 50% and 37% of SNP-gene links from the gold-standard CRISPR and GWAS evaluation sets, respectively; our primary enrichment metric is a relative measure and does not reflect the total rate of discovery, but our secondary metrics (including AUPRC and AUROC) provide complementary information. Fifth, the eQTL and ABC evaluation sets that we analyzed are only partially accurate and represent statistical predictions rather than gold-standard links; however, high enrichment with respect to these evaluation sets is still suggestive of high power to detect true links. In addition, the eQTL evaluation set should be viewed as a best-case scenario for pgBoost because the pgBoost model uses (off-chromosome) eQTL data for training. Sixth, we have not conclusively established whether it is better to use eQTL data or CRISPR data for training (considering comparable performance on GWAS-derived evaluation data); further, we caution that pgBoost causal probabilities are estimated with respect to the training data and may not reflect biological causality. It will be of interest to further explore training on CRISPR-validated links as more CRISPR data sets become available. Seventh, we have not conclusively established whether longer-range links reflect *cis*-regulatory or *trans*-regulatory biology, but we hypothesize that they are predominantly *cis*-regulatory, as pgBoost scores generally decrease with distance, even at larger distances (**Supplementary Figure 2.10**).

Eighth, pgBoost does not include as features linking scores derived from bulk data sets (e.g. ABC⁵², evaluated in **Supplementary Figure 2.12**); including these features in future extensions of pgBoost may further improve performance.

Methods: Implementation details for existing single-cell peak-gene linking methods

To generate peak-gene links from SCENT (version 0.1.0), we first ran the CreateSCENTObj function on gene expression count matrices and binarized ATAC peak matrices (specifying number of unique molecular identifiers / UMIs and percentage of mitochondrial reads as cell-level covariates). We then generated peak-gene links by running the SCENT_algorithm function on the resulting SCENT object (with parameters: reg = “poisson”, bin = TRUE). We then computed FDR (Benjamini Hochberg correction) from bootstrapped p-values as in ref.⁸⁰

To generate peak-gene links from Signac, we first normalized and scaled gene expression data using the Seurat SCTransform function¹⁶⁶ and binarized the ATAC peak matrix. We then ran the Signac (version 1.10.0) LinkPeaks function on the normalized gene expression and binarized ATAC peak matrices (with parameters: peak.assay = “peaks”, expression.assay = “SCT”, peak.slot = “counts”, expression.slot = “data”, method = “pearson”, distance = 5e+05, min.cells = 10, n_sample = 200, pvalue_cutoff = 1, score_cutoff = 0). We specified these non-default values of pvalue_cutoff and score_cutoff to obtain scores for all candidate links (instead of limiting output to links passing specific score thresholds).

To generate peak-gene links from ArchR (version 1.0.2), we first performed dimension reduction on RNA and ATAC data separately using the addIterativeLSI function and then generated a joint dimension reduction from RNA and ATAC using the addCombinedDims function. We then ran the addPeak2GeneLinks function (with parameters: reducedDims = “LSI_Combined”, useMatrix = “GeneExpressionMatrix”, dimsToUse = 1:30, scaleDims = NULL, corCutOff = 0.75, k = 100, knnIteration = 500, overlapCutoff = 0.8, maxDist = 5e+05, log2Norm = TRUE, predictionCutoff = 0.4). Finally, we output peak-gene links by running the getPeak2GeneLinks function (with parameters corCutOff = -1.5, FDRCutOff = 1, varCutOffATAC

= 0.25, varCutOffRNA = 0.25, resolution = 1). We specified these non-default values of corCutOff and FDRCutOff to obtain scores for all candidate links (instead of limiting output to links passing specific score thresholds).

To generate peak-gene links from Cicero (version 1.3.9), we ran the run_cicero function on binarized ATAC peak matrices (with parameters: window = 5e+05, sample_num = 100).

Methods: Gradient boosting hyperparameters

We specified the following model parameters (passed to the xgboost¹⁰⁵ function): learning rate (0.05), gamma (10), maximum tree depth (10, 15, 25), minimum child weight (6, 8, 10), and subsampling ratio (0.6, 0.8, 1). We selected these values of gamma and learning rate in accordance with previous studies^{221–225} to minimize overfitting. During each round of inference, pgBoost selects values of maximum tree depth, minimum child weight, and subsampling ratio via a randomized search with five-fold cross validation (over the sets of potential values given above, selected in accordance with refs^{224,225}). The parameter set attaining the highest AUC is selected and used to generate predictions. Descriptions of each parameter are given by ref.¹⁰⁵ Briefly: learning rate defines a shrinkage factor for feature weights after each iteration (a lower value makes the boosting process more conservative), the gamma parameter is described in **Methods** (equation 4.3), maximum tree depth defines the maximum number of edges from the “root” to any “node” of the tree constructed during each iteration (a lower value makes the boosting process more conservative), minimum child weight defines the minimum instance weight (intuitively, sample size) in a node required to conduct further partitioning (a larger value makes the boosting process more conservative), and subsampling ratio defines the proportion to which pgBoost will randomly subsample training data to prevent overfitting (a lower values makes the boosting process more conservative).

Methods: pgBoost feature selection and evaluation of feature importance

To select existing scores to include as pgBoost features, we conducted analyses adding sets of features to the pgBoost framework and evaluating model performance on eQTL, ABC,

CRISPR, and GWAS evaluation data sets. We trained pgBoost using features from a single existing score (SCENT⁸⁰, Signac^{77,81}, ArchR⁸², or Cicero⁸³) plus distance, resulting in 4 models each spanning 12 features (1 constituent score for each of 10 data set-cell type pairs plus 2 distance-based features). We selected the top-performing model among these 4 models on eQTL, ABC, CRISPR, and GWAS evaluations. For a given evaluation set, we identified the top-performing model(s): the “winning” model attaining the highest average enrichment plus any other models that did not significantly underperform the “winning” model. We intersected the top-performing model(s) from the 4 evaluations to determine the overall top-performing model. We then conducted sequential feature selection, iteratively adding existing methods as pgBoost features only if doing so significantly improved model performance across the 4 evaluation sets. We ended the feature selection procedure once the top-performing model could not be significantly improved by adding more features. Whenever the 4 evaluation sets yielded discordant results, we prioritized results on the eQTL, CRISPR, and GWAS evaluations over the ABC evaluation. The eQTL evaluation set is large (and therefore well-powered to detect differences in method performance) and provides orthogonal genetic evidence of regulatory links. The CRISPR and GWAS evaluations represent “gold-standards” with stronger evidence of biological causality, although they are smaller and therefore less well-powered to detect differences. The ABC evaluation is large and well-powered but may be susceptible to biases related to genomic distance. If $n > 1$ top-performing feature sets (with equal numbers of features) were identified, we split the feature selection procedure into n separate paths, separately testing further additions to each top-performing feature set. (In our analyses, each path of feature selection yielded the same final feature set).

To determine which features contribute most to pgBoost performance, we conducted analyses ablating sets of features from the pgBoost framework and evaluating performance on eQTL, ABC, CRISPR, and GWAS evaluation data sets. We trained pgBoost using features from genomic distance, a single existing peak-gene linking score, or a single existing score plus

distance (resulting in models spanning either 2 distance-based features; 10 features derived from SCENT⁸⁰, Signac^{77,81}, ArchR⁸², or Cicero⁸³; or 12 features using one existing score plus distance-based features). We similarly trained gradient boosting models with features from a single existing score ablated, resulting in models spanning 30 features (with 2 distance-based features ablated) or 22 features (with SCENT⁸⁰, Signac^{77,81}, ArchR⁸², or Cicero⁸³ features ablated). We then trained pgBoost using linking scores computed in one data set (**Table 2.2**), resulting in models spanning 11, 17, or 5 features (linking scores computed in the 10x PBMC¹⁰⁸, Luecken BMMC¹⁰⁹, or SHARE-seq LCL⁷⁷/Xu K562²⁰⁴ data sets, respectively, plus distance-based features). We similarly trained gradient boosting models with features from one data set ablated, resulting in models spanning 23, 17, or 29 features (ablating linking scores computed in the 10x PBMC¹⁰⁸, Luecken BMMC¹⁰⁹, or SHARE-seq LCL⁷⁷/Xu K562²⁰⁴ data sets, respectively).

We also applied Shapley Additive Explanation (SHAP), a tool for interpreting complex nonlinear models^{203,224}. SHAP uses positive/negative training labels and the trained model to quantify the importance of individual features driving pgBoost predictions while accounting for nonlinear interactions. We computed SHAP scores using SHAPforxgboost software²²⁶.

Methods: GWAS variant enrichment analyses

We applied the *average enrichment across recall values* metric (**Methods**) to assess the ability of pgBoost models trained in focal cell types to prioritize (assign higher linking scores to) causal variants fine-mapped to cell type-relevant traits in GWAS.

We analyzed GWAS fine-mapping results for 7 red blood cell or platelet-related blood cell traits (platelet count, hemoglobin A1c, hemoglobin, red blood cell count, mean corpuscular volume, mean corpuscular hemoglobin, mean corpuscular hemoglobin concentration) and 7 autoimmune diseases and granulocyte-related blood cell traits (white blood cell count, monocyte count, eosinophil count, autoimmune disease, lymphocyte count, neutrophil count, basophil count) from the UK Biobank^{107,112}. 455 variants (resp. 488 variants) scored by pgBoost were fine-mapped to any red blood cell or platelet-related blood cell traits (resp. autoimmune disease or

granulocyte-related blood cell trait) with PIP > 0.2. (We selected a PIP threshold of > 0.2 instead of > 0.5 to specify a larger variant evaluation set, as approximately 200 variants were fine-mapped to each category of GWAS traits with PIP > 0.5.)

For a given evaluation set of fine-mapped GWAS variants and a given pgBoost model corresponding to a focal cell type, we ranked candidate SNPs by the maximum linking score across all candidate SNP-gene links involving the focal SNP. At each unique score c assigned to any SNP, we computed variant-level enrichment and recall as follows:

$$\text{Enrichment} = \frac{|\text{linked SNPs} \cap \text{causal SNPs}| / |\text{causal SNPs}|}{|\text{linked SNPs}| / |\text{candidate SNPs}|} \quad [2.7]$$

$$\text{Recall} = \frac{|\text{linked SNPs} \cap \text{causal SNPs}|}{|\text{causal SNPs}|} \quad [2.8]$$

where “linked SNPs” denotes the binarized set of SNPs attaining maximum linking score $\geq c$, “causal SNPs” denotes the binarized set of SNPs attaining PIP > 0.2 for any trait in the focal category, and “candidate SNPs” denotes SNPs lying in ATAC-seq peaks in the Luecken BMMC¹⁰⁹ or Xu K562¹¹⁰ data sets (**Table 2.2**).

For a given evaluation set and set of methods, we used equation 2.7 and equation 2.8 to construct variant-level enrichment-recall curves. For each method, we considered different values of the number of evaluation SNPs prioritized (k), corresponding to a recall and enrichment of $(\text{recall}_k, \text{enrichment}_k)$, and performed linear extrapolation to complete the enrichment-recall curve between the points $(\text{recall}_k, \text{enrichment}_k)$ and $(1, 1)$, such that the enrichment-recall curve spanned values of recall in the range $[0, 1]$. We then computed *average enrichment across recall values* using equation 2.6 (**Methods**).

We obtained standard errors on average enrichment for each method, as well as standard errors and p-values on the difference in average enrichment between two methods, by bootstrapping genes using 1,000 bootstrap iterations.

Methods: AUPRC metric

As a secondary evaluation metric, we computed *area under the precision-recall curve* (AUPRC) or, equivalently, *average precision across recall values*. At each unique linking score c assigned to any candidate SNP-gene link, we computed precision as follows:

$$\text{Precision} = \frac{|\text{predicted SNP-gene links} \cap \text{true SNP-gene links}|}{|\text{predicted SNP-gene links}|} \quad [2.9]$$

where “predicted SNP-gene links” denotes the binarized set of SNP-gene links attaining linking score $\geq c$, “true SNP-gene links” denotes the binarized set of “true” SNP-gene links as defined by the evaluation set, and the set of “candidate SNP-gene links” is defined depending on the evaluation set (see *Average enrichment metric* in **Methods**).

For a given evaluation data set and set of methods, we used equation 2.9 and equation 2.5 to construct precision-recall curves. For each method, we considered different values of the number of evaluation links prioritized (k), corresponding to a recall and precision of $(\text{recall}_k, \text{precision}_k)$, and performed linear extrapolation to complete the precision-recall curve between the points $(\text{recall}_k, \text{precision}_k)$ and $(1, 0)$, such that the precision-recall curve spanned values of recall in the range $[0, 1]$. We then computed average precision across each unique level of recall in $[0, 1]$, thus defining *AUPRC*.

We obtained standard errors on AUPRC for each method, as well as standard errors and p-values on the difference in AUPRC between two methods, by bootstrapping genes using 1,000 bootstrap iterations.

Methods: AUROC metric

As a secondary evaluation metric, we computed *area under the receiver operating characteristic curve* (AUROC) or, equivalently, *average sensitivity across specificity values*. At each unique linking score c assigned to any candidate SNP-gene link, we computed sensitivity and specificity as follows (noting that sensitivity is equivalent to recall in equation 2.5):

$$\text{Sensitivity} = \frac{|\text{predicted positive SNP-gene links} \cap \text{true positive SNP-gene links}|}{|\text{true positive SNP-gene links}|} \quad [2.10]$$

$$\text{Specificity} = \frac{|\text{predicted negative SNP-gene links} \cap \text{true negative SNP-gene links}|}{|\text{true negative SNP-gene links}|} \quad [2.11]$$

where “predicted positive SNP-gene links” and “predicted negative SNP-gene links” denote the binarized set of SNP-gene links attaining linking score $\geq c$ or $< c$, respectively, and “true positive SNP-gene links” and “true negative SNP-gene links” denote the binarized set of “true” positive and negative SNP-gene links defined by the evaluation set (with the background set of candidate SNP-gene links, defined depending on the evaluation set; see *Average enrichment metric* in **Methods**).

For a given evaluation data set and set of methods, we used equation 2.10 and equation 2.11 to construct sensitivity-specificity curves. For each method, we considered different values of the number of evaluation links prioritized (k), corresponding to a specificity and sensitivity of $(\text{specificity}_k, \text{sensitivity}_k)$, and performed linear extrapolation to complete the sensitivity-specificity curve between the points $(\text{specificity}_k, \text{sensitivity}_k)$ and $(1, 0)$, such that the sensitivity-specificity curve spanned values of specificity in the range $[0, 1]$. We then computed average sensitivity across each unique level of specificity in $[0, 1]$, thus defining *AUROC*.

We obtained standard errors on AUROC for each method, as well as standard errors and p-values on the difference in AUPRC between two methods, by bootstrapping genes using 1,000 bootstrap iterations.

Methods: Unweighted average enrichment metric

As a secondary average enrichment metric, we computed *unweighted average enrichment across shared recall values*, with two modifications to the main enrichment metric. First, for a given set of methods, we computed $k_{\min}$, the minimum number of evaluation links scored by any method, and averaged enrichment across shared recall values $[0, \text{recall}_{k_{\min}}]$,

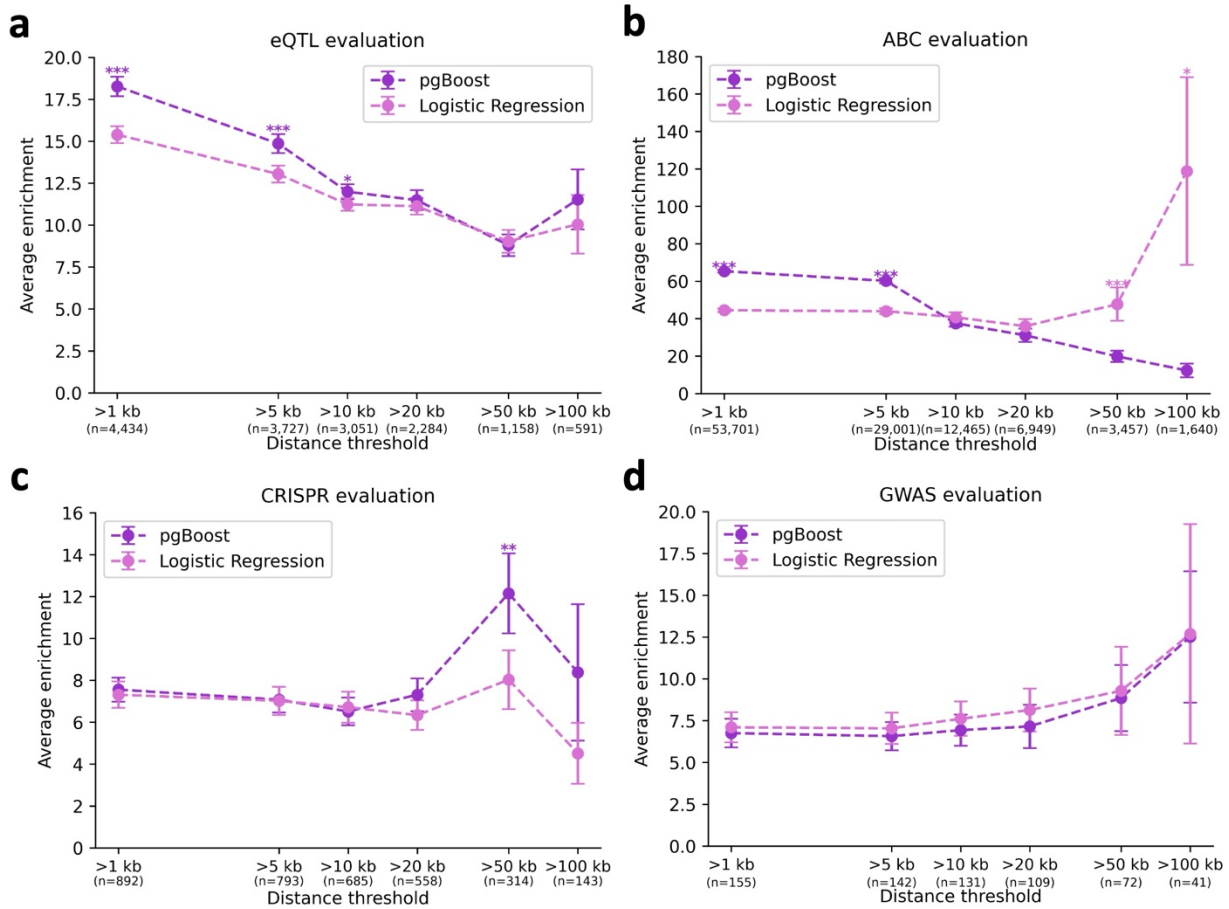
instead of $[0, 1]$. Second, we computed a flat average of enrichment values, instead of weighting by $(1 - \text{recall})$ as in equation 2.6 (**Methods**).

Methods: Downsampling analyses varying cell count and sequencing depth

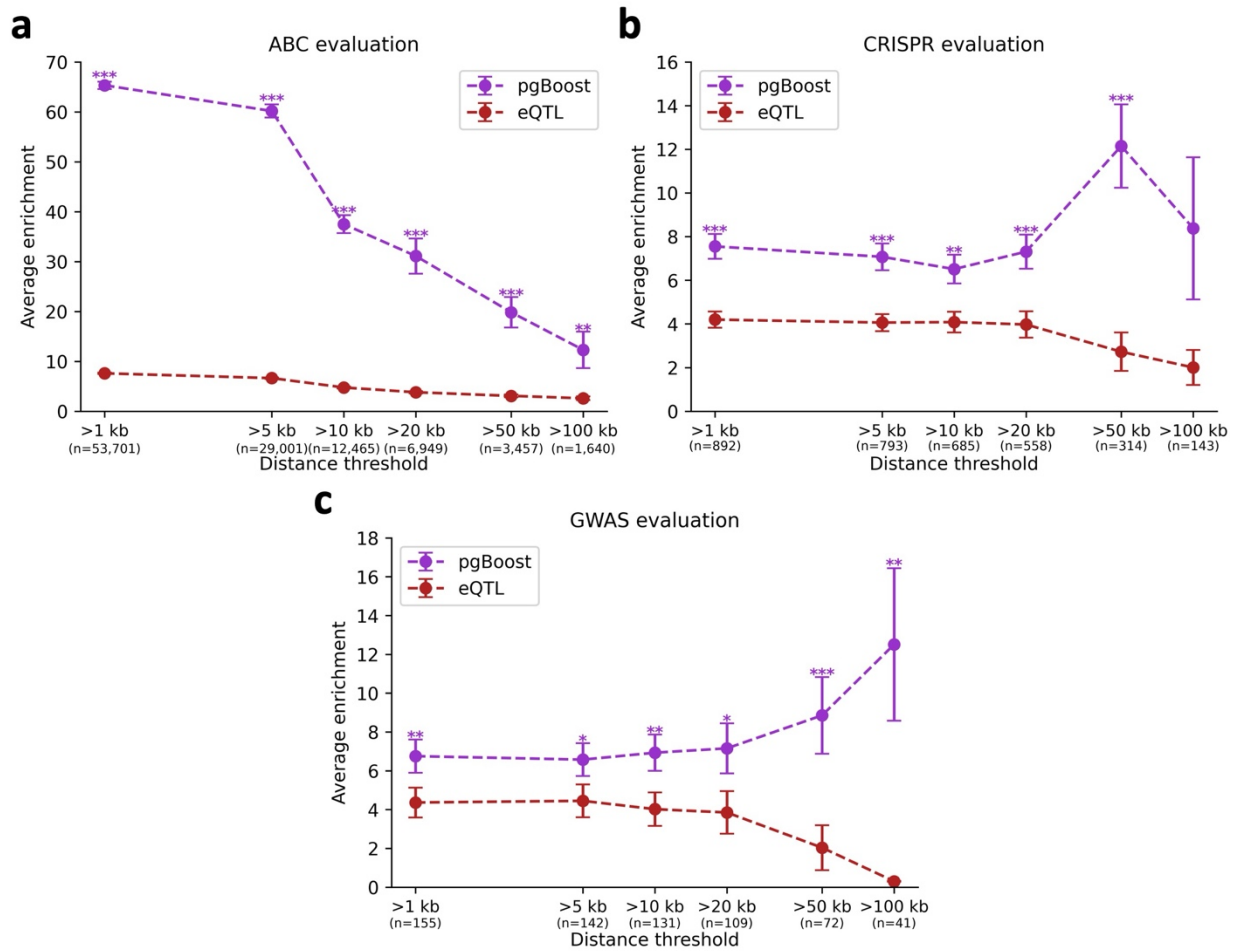
We performed downsampling analyses assessing the sensitivity of pgBoost and existing methods to variation in cell count, RNA count, ATAC count, and RNA + ATAC count simultaneously. We conducted all downsampling analyses in T cells from the NeurlPS data set (**Table 2.2**). We generated matrices reflecting 4 downsampling conditions (50% of cells; 50% of RNA counts and 100% of ATAC peak counts; 100% of RNA counts and 50% of ATAC peak counts; 50% of RNA counts and 50 % of ATAC peak counts), ran the 4 existing methods and pgBoost in each of these downsampled conditions, and computed average enrichment across recall values (**Methods**) of each method in downsampled and non-downsampled conditions. (We note that Cicero only uses ATAC-seq data as input and thus is unaffected by downsampling of ATAC peak counts.)

In each downsampled condition, pgBoost is trained on the 2 distance-based features and constituent linking scores (SCENT, Signac, Cicero) computed on downsampled data. Because ArchR software⁸² does not support modification of the ATAC peak count matrix, we did not consider ArchR in experiments downsampling ATAC or RNA counts (**Supplementary Figure 2.13b-d**).

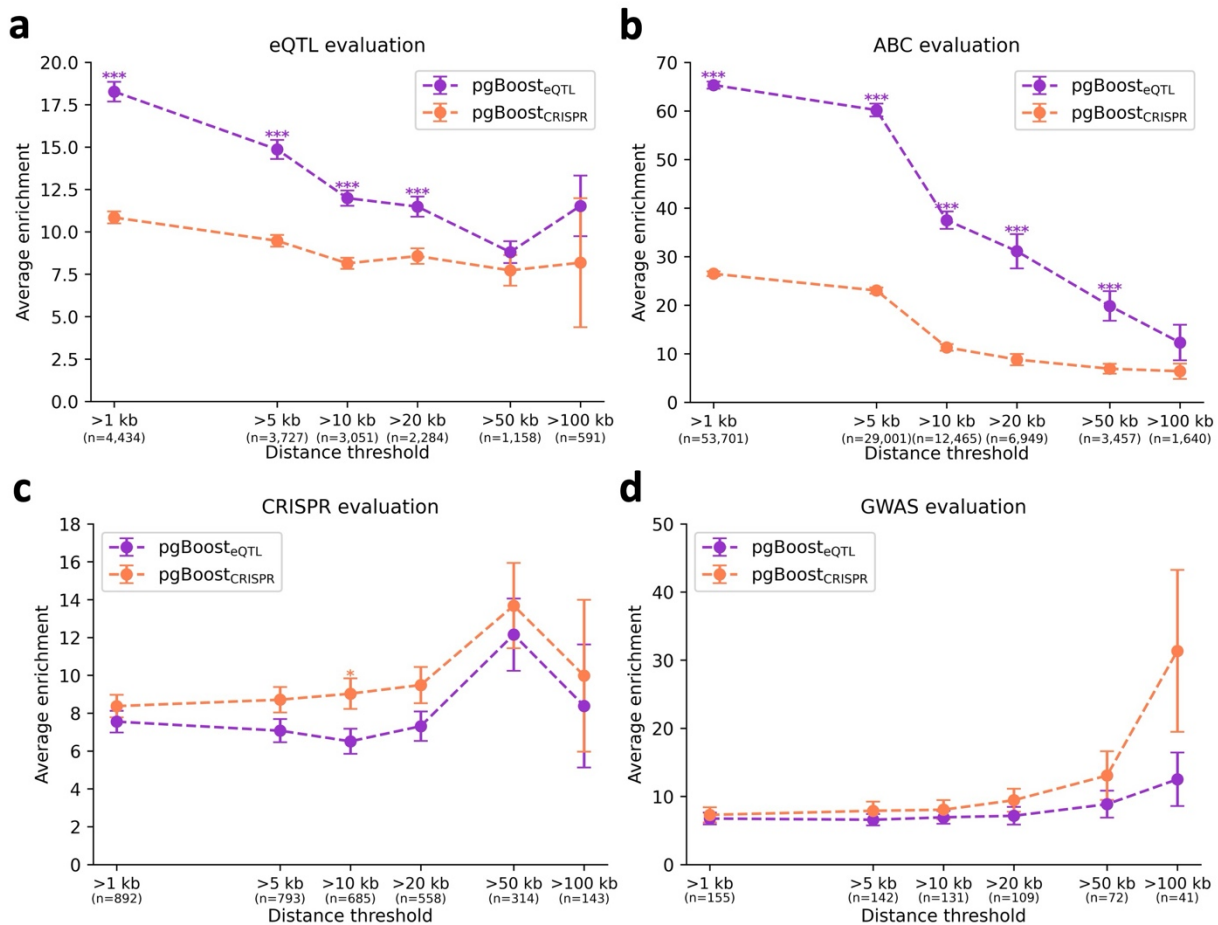
Extended Data Figures



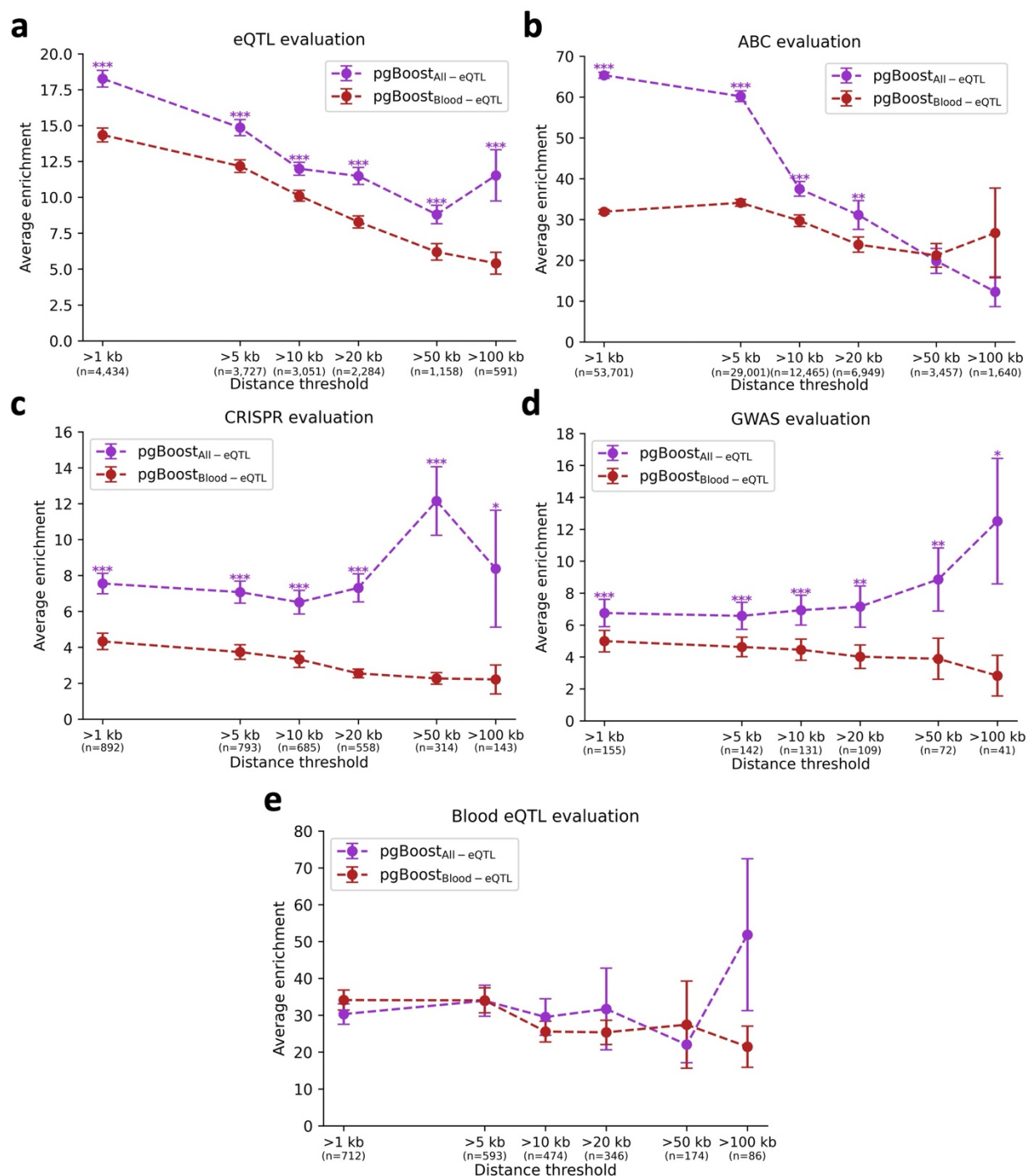
Extended Data Figure 2.1. Average enrichment across recall values of links predicted by pgBoost vs. a logistic regression model with identical features for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method (**Methods**). Gradient boosting significantly outperformed logistic regression on eQTL evaluation data and CRISPR evaluation data >50kb. Gradient boosting did not significantly outperform logistic regression on GWAS evaluation data. While gradient boosting significantly outperformed logistic regression on all ABC evaluation data (>1kb and >5kb), gradient boosting significantly underperformed logistic regression on longer-range ABC evaluation data (>50kb and >100kb); this could be due to the gradient boosting model assigning higher weight to distance-based features than the logistic regression model (however, we do not observe this behavior in other evaluation sets).



Extended Data Figure 2.2. Average enrichment across recall values of links predicted by pgBoost vs. fine-mapped eQTL (ranking candidate SNP-gene links by maximum PIP across GTEx tissues) for **(a)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(b)** 892 links validated by CRISPR, and **(c)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method (**Methods**).

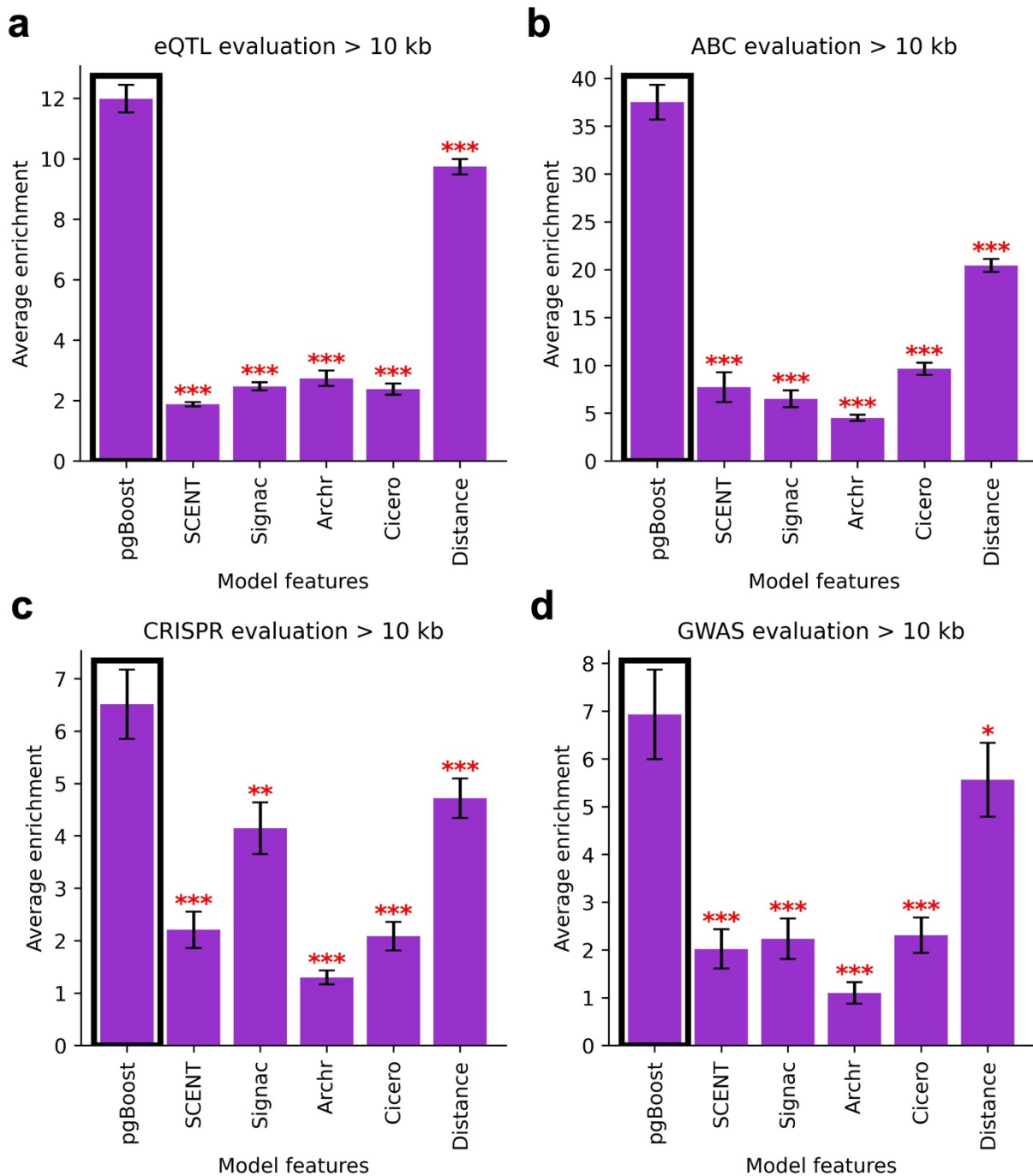


Extended Data Figure 2.3. Average enrichment across recall values of links predicted by pgBoost trained on fine-mapped eSNP-eGene pairs vs. CRISPR-validated links for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. While other regulatory linking methods have used CRISPR links for model training, pgBoost may perform better when trained on eQTL due to the limited amount of CRISPR data, pervasive aneuploidy and structural variants in the K562 cell line, and/or better representation of multiple cell types in GTEx eQTL data. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method (**Methods**). For pgBoost_{CRISPR}, we defined a training set comprised of 892 SNP-gene links validated by CRISPR as positives and 10,345 links tested but not validated by CRISPR as negatives.



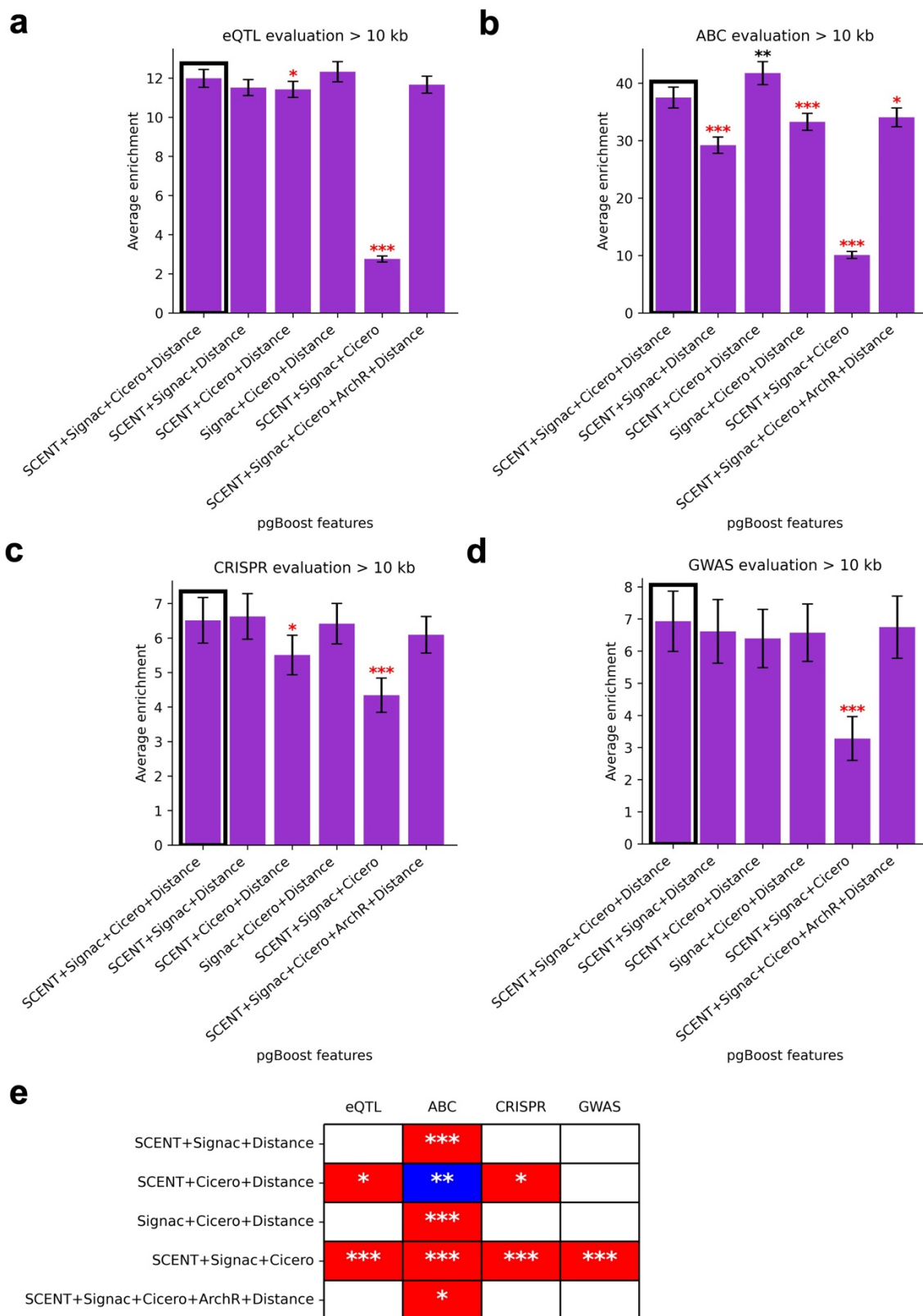
Extended Data Figure 2.4. Average enrichment across recall values of links predicted by pgBoost trained on fine-mapped eSNP-eGene pairs from all GTEx tissues vs. fine-mapped eSNP-eGene pairs from whole blood for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, and **(e)** 712 fine-

mapped eSNP-eGene pairs attaining $PIP > 0.5$ in GTEx whole blood, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method (**Methods**). For $pgBoost_{\text{Blood-eQTL}}$, we defined a training set of 1,518 eSNP-eGene pairs attaining $PIP > 0.2$ in whole blood as positives and 6,997 eSNP-eGene pairs including a gene in the positive set and attaining $PIP < 0.01$ in whole blood as negatives.



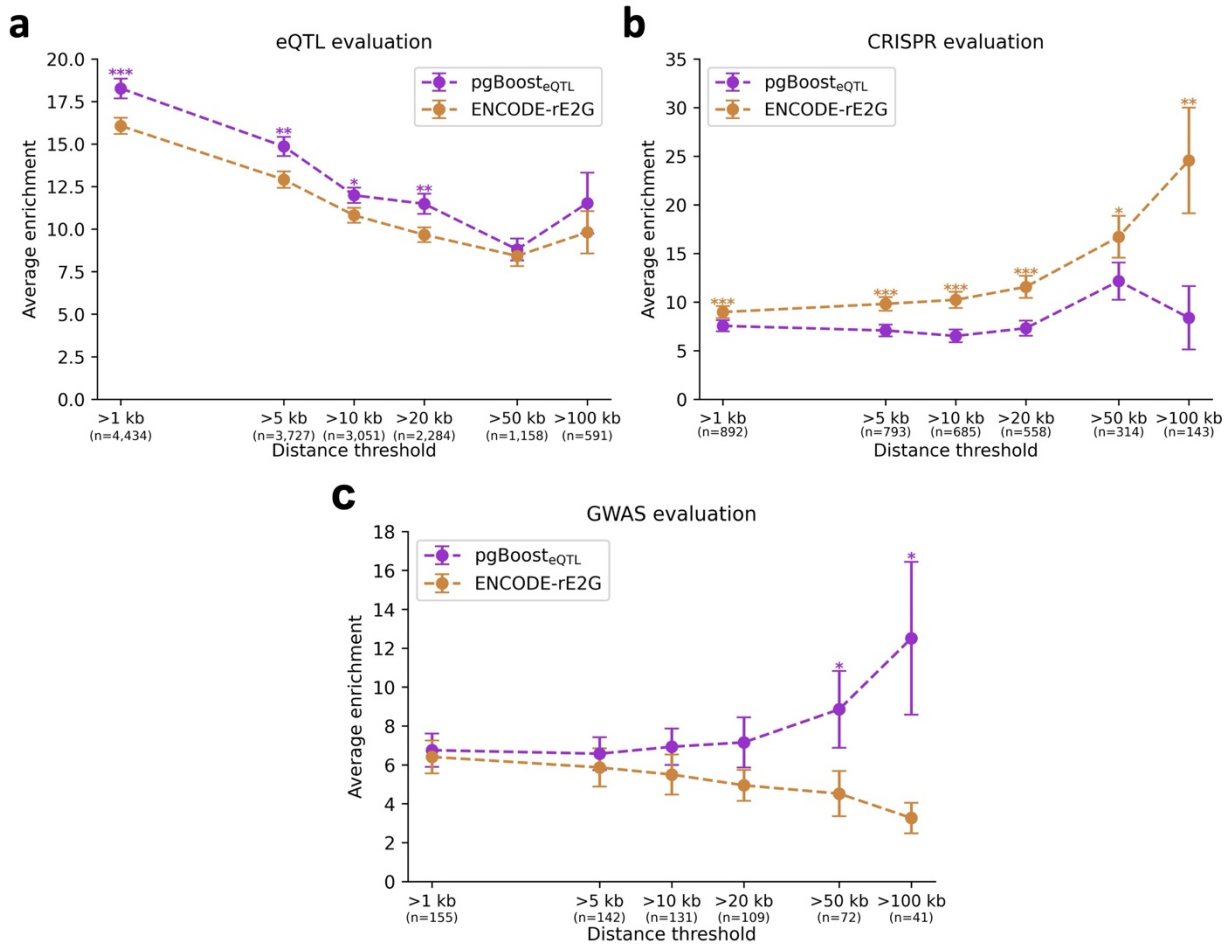
Extended Data Figure 2.5. Average enrichment across recall values of links predicted by pgBoost restricted to linking scores from a single existing method or genomic distance for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar

across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the main pgBoost model (denoted in black outline) vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) (**Methods**). Red stars indicate that the focal method significantly underperforms pgBoost.

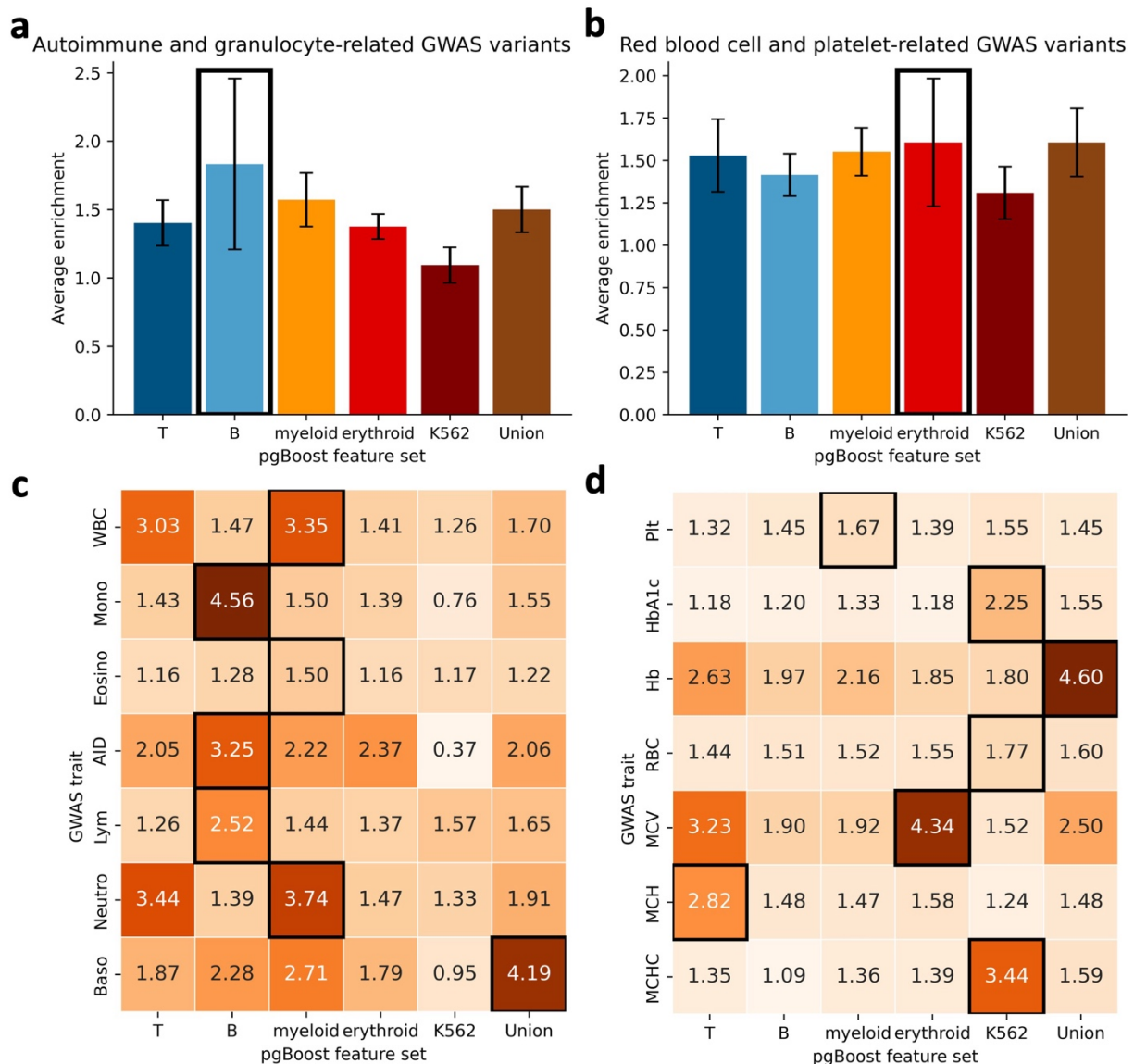


Extended Data Figure 2.6. Average enrichment across recall values of links predicted by pgBoost (denoted in black outline) vs. pgBoost models ablating features from one

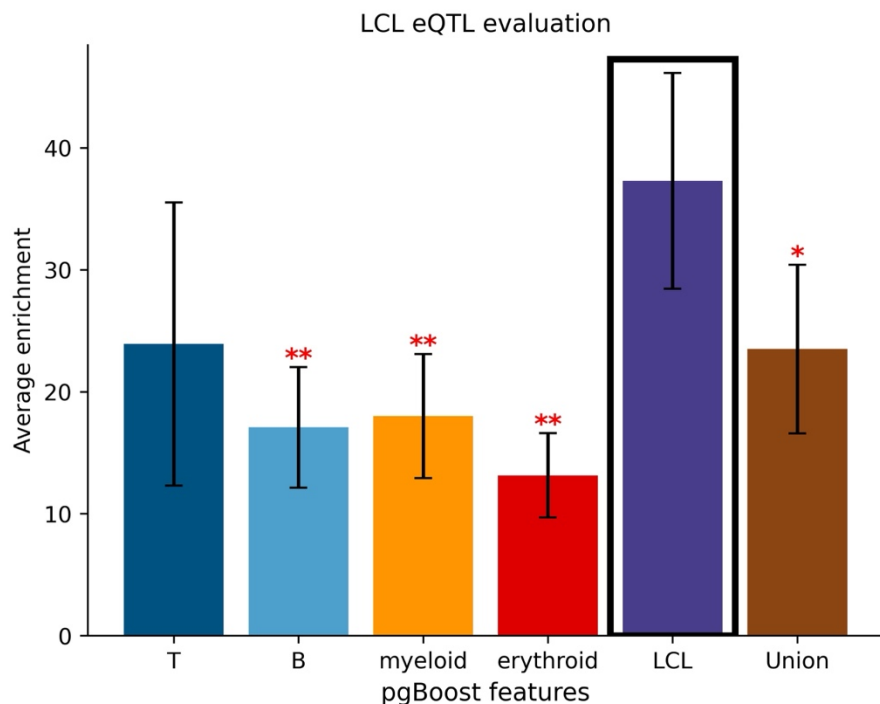
existing peak-gene linking method (SCENT, Signac, Cicero), ablating distance-based features, or adding ArchR features for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. **(e)** Table summarizing the significance of differences in average enrichments between pgBoost and each other model on the 4 evaluation sets. Because enrichment differences are highly similar across distance thresholds in **a-d**, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of focal method vs. main pgBoost model (**Methods**). Red stars (in **a-d**) or red shading (in **e**) indicates that the focal method significantly underperformed pgBoost. Black stars (in **a-d**) or blue shading (in **e**) indicates that the focal method significantly outperformed pgBoost. While the main pgBoost model significantly underperformed the model ablating Signac on ABC evaluation data, we prioritized results on eQTL and CRISPR evaluations (see **Supplementary Note**). The main pgBoost model significantly outperformed each ablated model on at least 1 evaluation data set and never significantly underperformed an expanded model, supporting our choice to include SCENT, Signac, Cicero, and distance as features (and exclude ArchR).



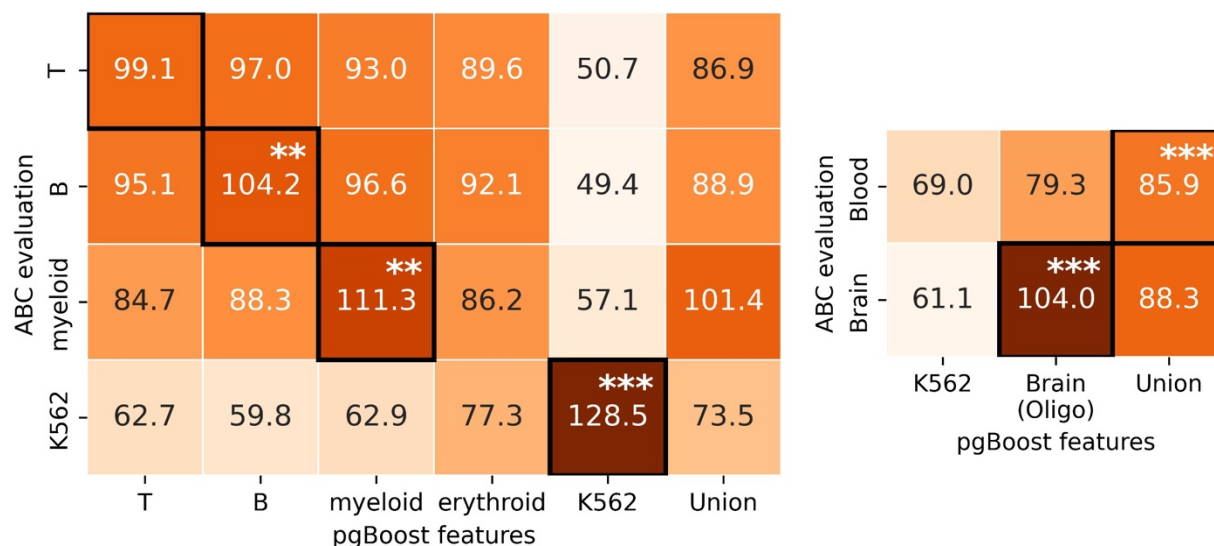
Extended Data Figure 2.7. Average enrichment across recall values of links predicted by pgBoost vs. ENCODE-rE2G (maximum score across biosamples) for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds (we exclude ABC evaluation data from these comparisons, because the ABC evaluation data were used as features to generate ENCODE-rE2G predictions). The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method (**Methods**).



Extended Data Figure 2.8. Average enrichment across recall values of variants linked to genes by pgBoost restricted to features from a focal cell type (Union denotes all features) for **(a)** $n = 488$ fine-mapped variants (PIP > 0.2) for 7 autoimmune diseases or granulocyte-related blood cell traits, **(b)** $n = 455$ fine-mapped variants (PIP > 0.2) for 7 red blood cell or platelet-related blood cell traits, **(c)** fine-mapped variants for individual traits from **(a)**, **(d)** fine-mapped variants for individual traits from **(b)**. Confidence intervals denote standard errors, and stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of focal method vs. top method (denoted in black outline) (**Methods**). Plt: platelet count, HbA1c: hemoglobin A1c, Hb: hemoglobin, RBC: red blood cell count, MCV: mean corpuscular volume, MCH: mean corpuscular hemoglobin, MCHC: mean corpuscular hemoglobin concentration, WBC: white blood cell count, Mono: monocyte count, Eosino: eosinophil count, AID: autoimmune disease, Lym: lymphocyte count, Neutro: neutrophil count, Baso: basophil count.

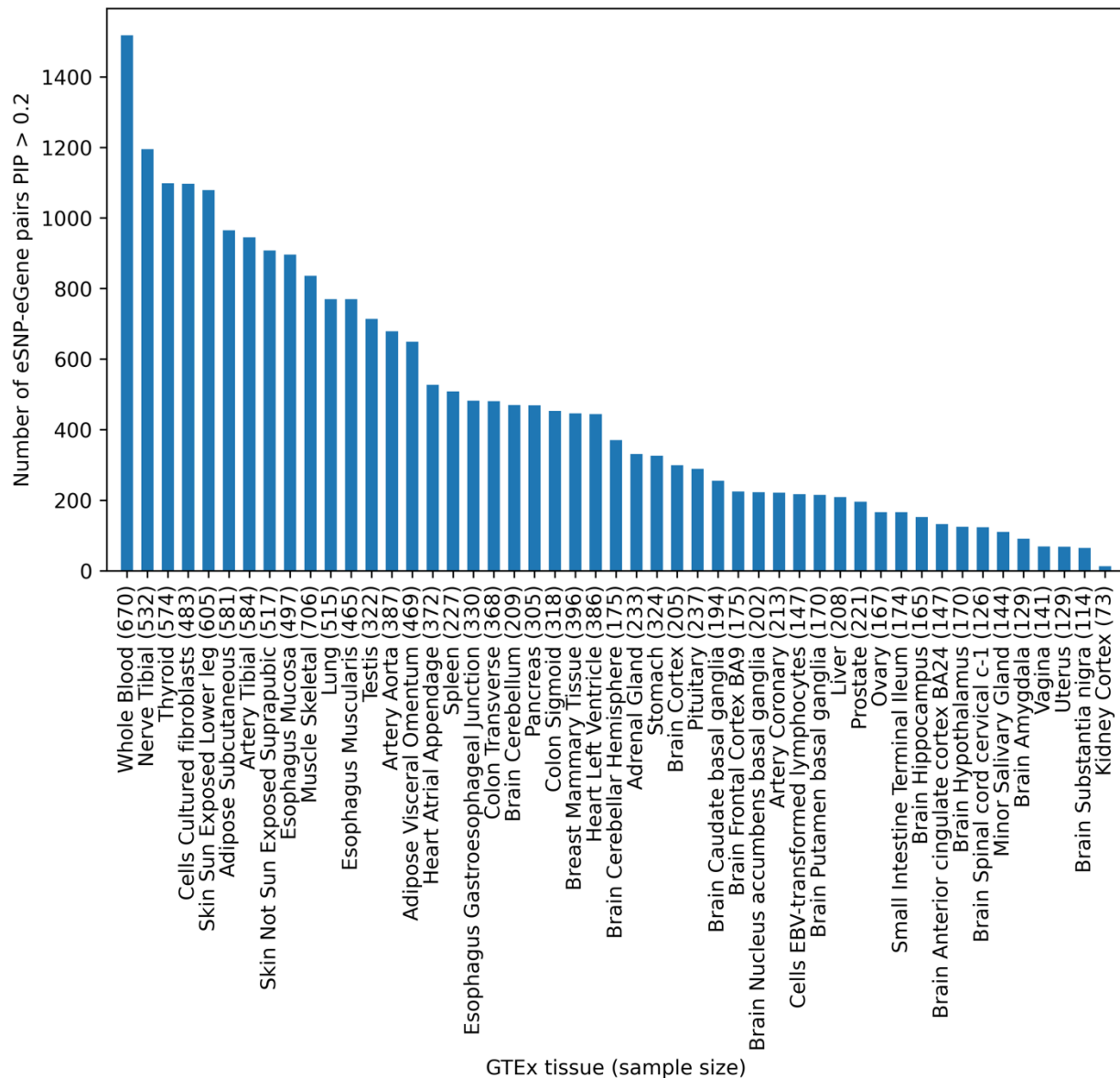


Extended Data Figure 2.9. Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (columns; Union denotes all features) for fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 in GTEx EBV-transformed lymphocytes or LCL (number of evaluation links: $n = 79$). We expanded the set of pgBoost models to include a model restricted to LCL features from the SHARE-seq LCL data set (**Table 2**). The pgBoost model restricted to LCL significantly outperformed models restricted to B cell, myeloid, or erythroid features and the model spanning the union of all 5 cell types, and performed similarly to the model restricted to T cell features. (We did not analyze cell-type-specific GTEx evaluation data for T cells, B cells, myeloid cells, or erythroid cells, because these cell types are not represented in GTEx). Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method (denoted in black outline) vs. each other method (**Methods**). Red stars denote significant underperformance of the focal method vs. the top method.

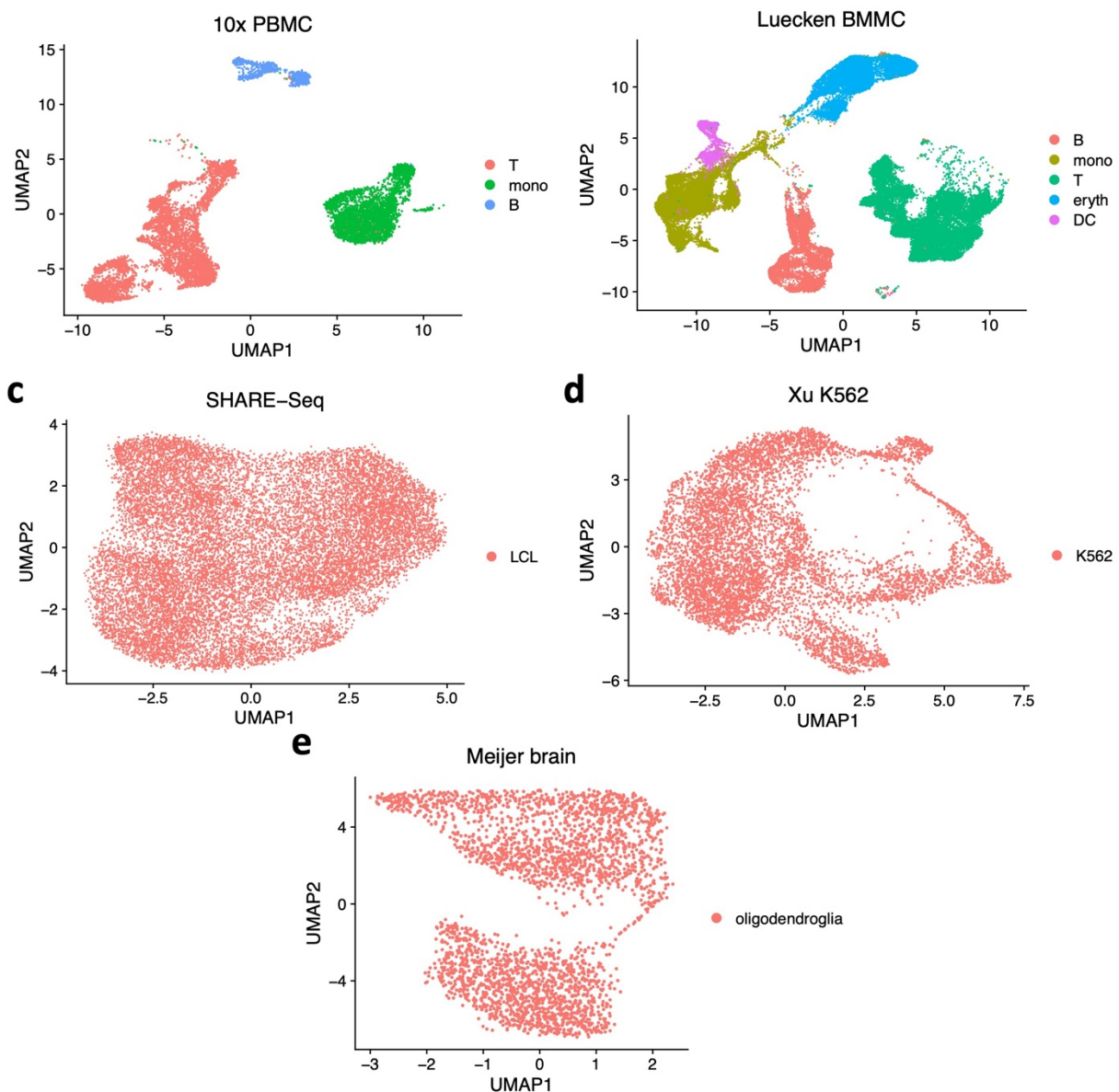


Extended Data Figure 2.10. Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (columns; Union denotes all features) for cell-type-level (left) or tissue-level (right) ABC evaluation links (rows) defined by SNP-gene links attaining ABC score > 0.2 in biosample(s) related to the focal cell type or tissue (irrespective of ABC scores attained in other biosamples). Number of evaluation links in each ABC evaluation set: $n_T = 19,301$, $n_B = 17,638$, $n_{\text{myeloid}} = 10,632$, $n_{K562} = 3,793$, $n_{\text{blood}} = 33,867$, $n_{\text{brain}} = 19,623$. Number of biosamples in each ABC evaluation set: $N_{\text{blood}} = 74$, $N_{\text{brain}} = 26$, $N_T = 53$, $N_B = 14$, $N_{\text{myeloid}} = 5$, $N_{K562} = 1$. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method (denoted in black outline; defined separately for each row) vs. each other method (**Methods**). For cell-type-level evaluation data sets (left), pgBoost models restricted to the focal cell type significantly outperformed each other model, in general (except for the T cell-level ABC evaluation data set, for which the model restricted to T cell features did not significantly outperform the model restricted to B cell features, but significantly outperformed each other model). On brain-level evaluation data (right), the pgBoost model restricted to brain oligodendroglia features significantly outperformed the model restricted to K562 features and the model spanning both cell types. On blood-level evaluation data (right), the pgBoost model spanning the union of both cell types significantly outperformed both other models.

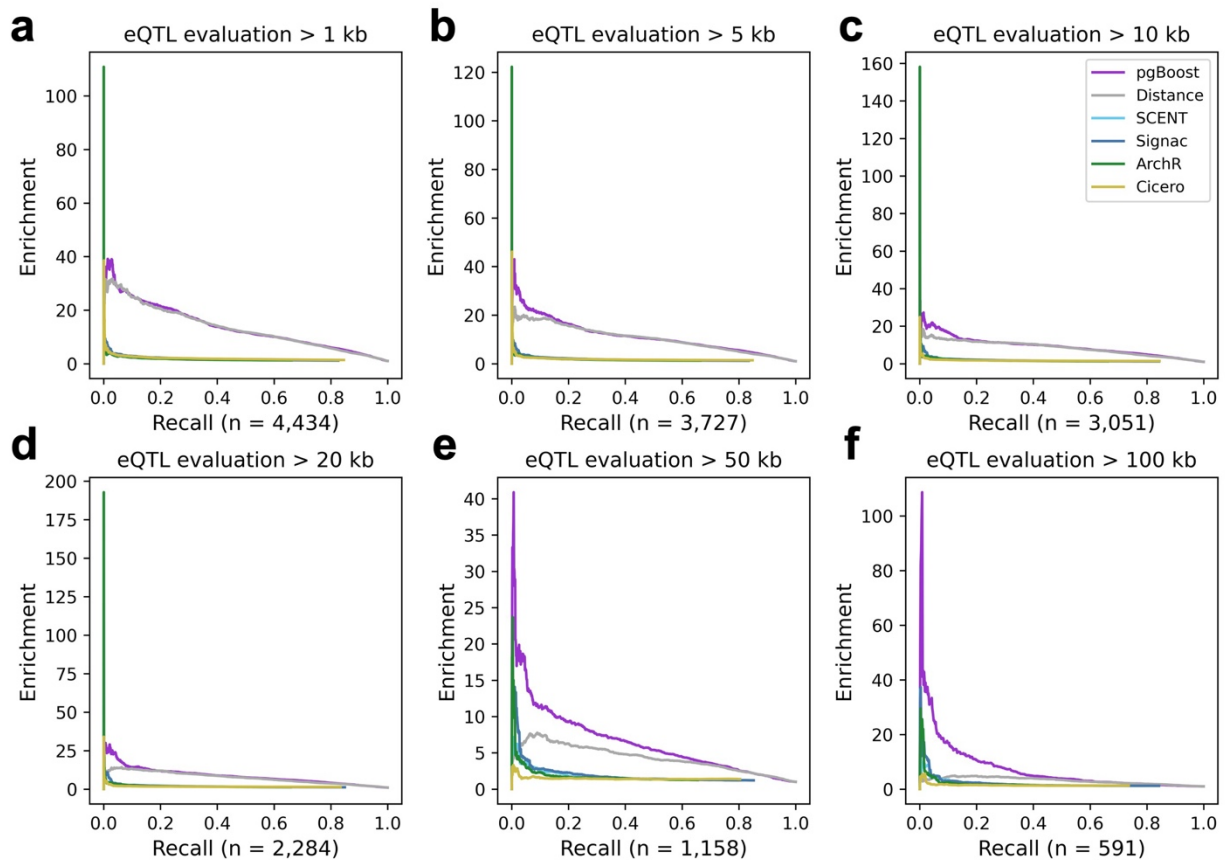
Supplementary Figures



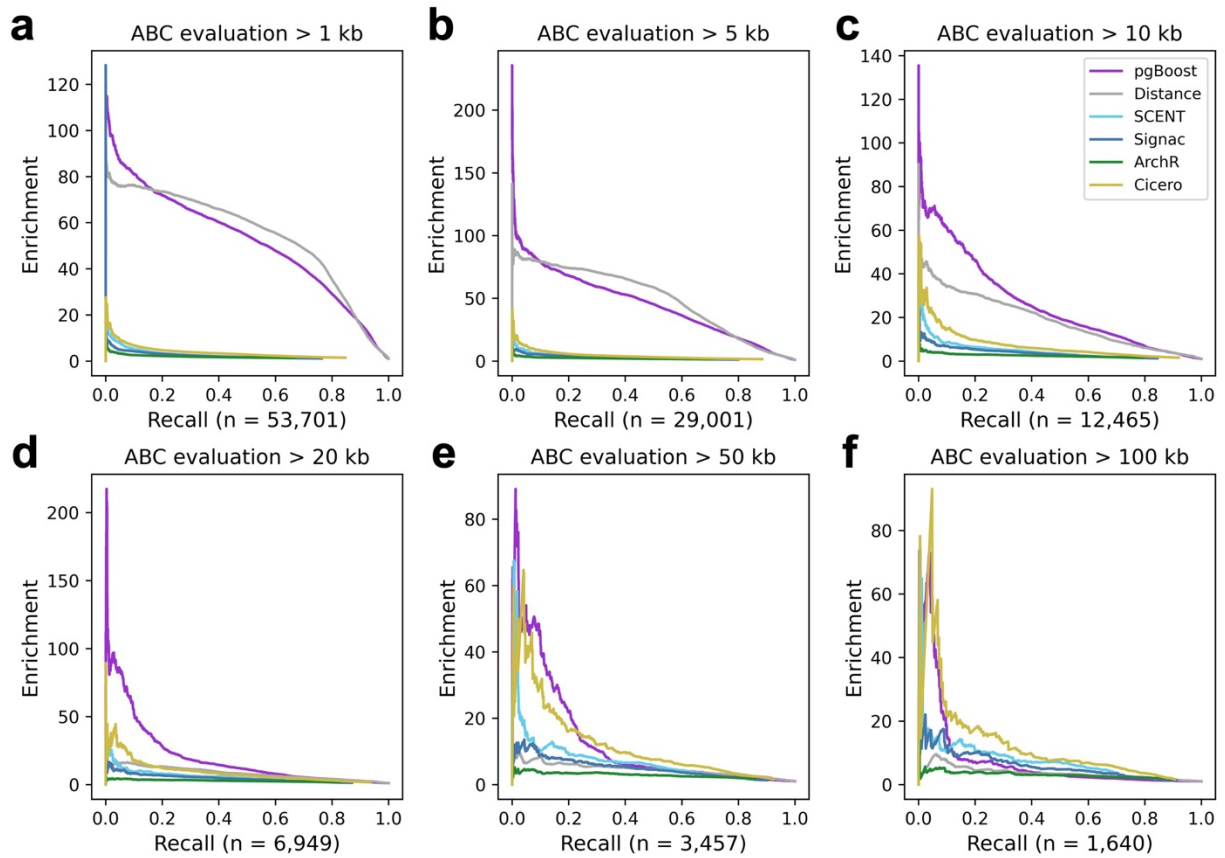
Supplementary Figure 2.1: Tissue representation in eQTL training set. Number of positive training eSNP-eGene pairs (fine-mapped PIP > 0.2) detected in each GTEx tissue. The sample size of each GTEx tissue is specified in parentheses. We note that each positive eSNP-eGene pair can be discovered in multiple tissues. In general, more positive eSNP-eGene pairs were detected within tissues with higher sample size. While whole blood harbors the largest number of positive eSNP-eGene pairs (1,518, vs. 370 for the median tissue), 10,866 / 11,926 positive eSNP-eGene pairs were discovered in at least one tissue other than whole blood.



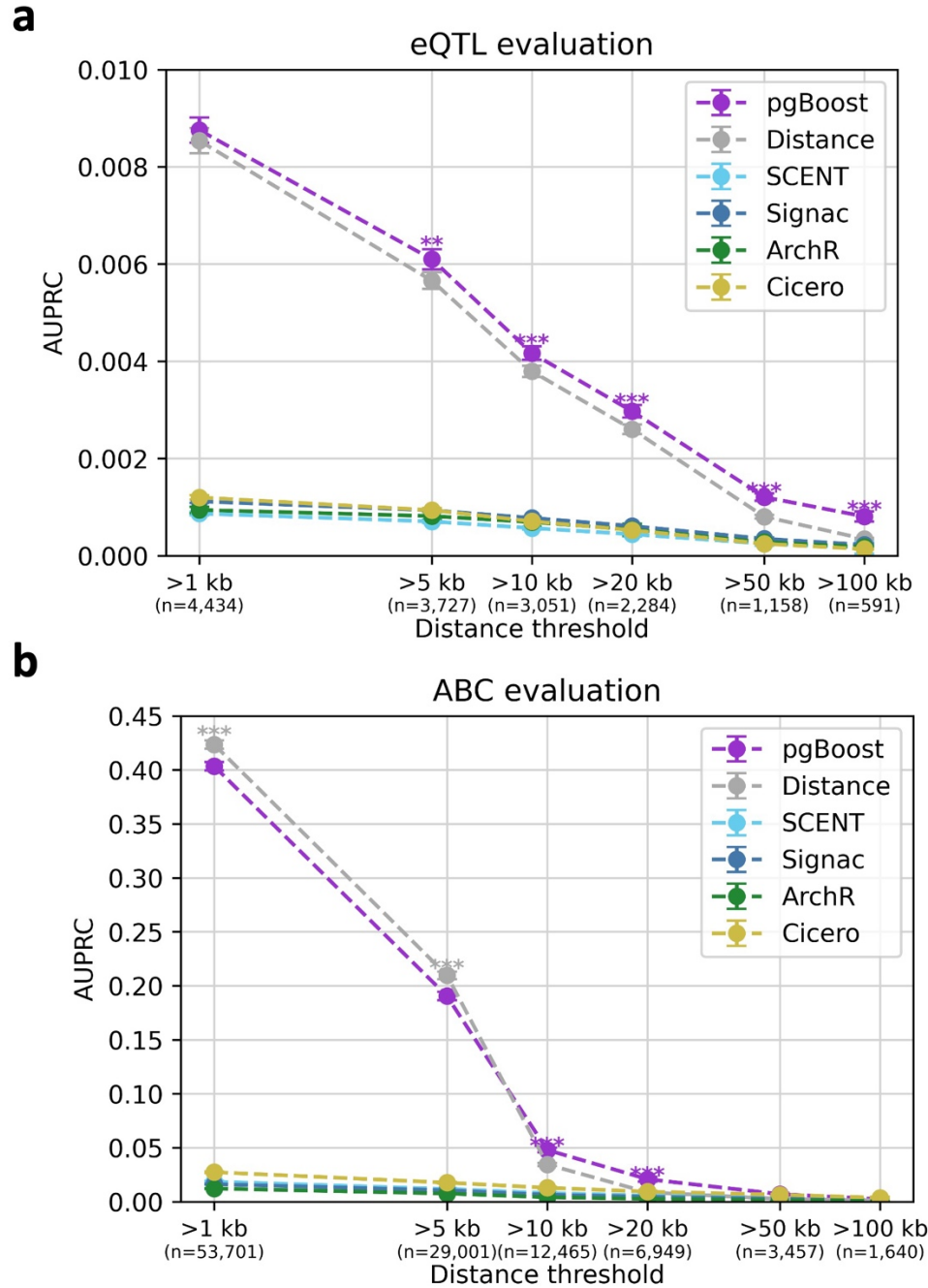
Supplementary Figure 2.2: UMAP plots of single-cell data sets. Cells in 5 single-cell RNA/ATAC-seq multiome data sets summarized in a joint UMAP visualization representing both gene expression and chromatin accessibility (built on a joint weighted nearest neighbor graph²²⁷) and colored by cell type.



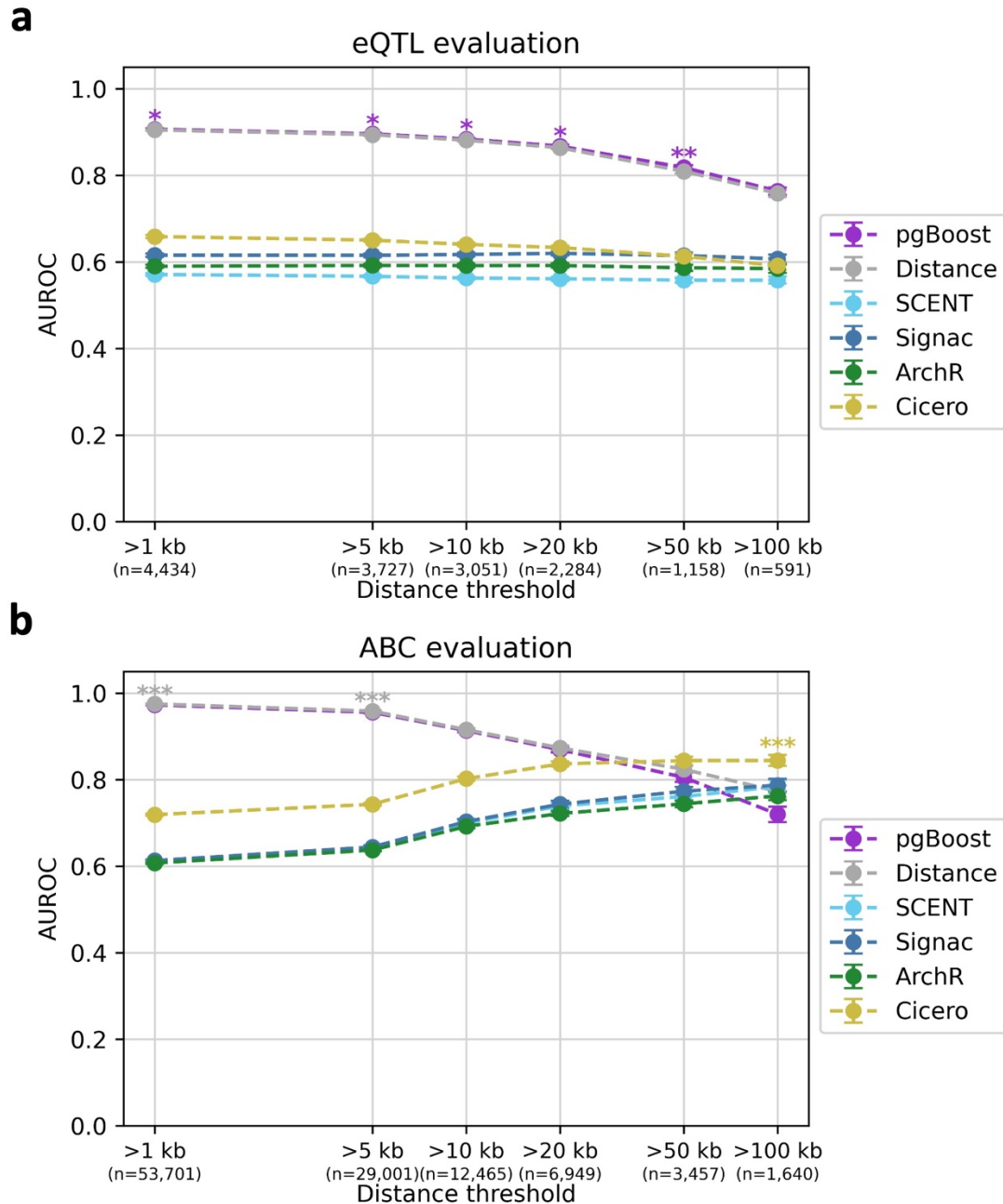
Supplementary Figure 2.3: Enrichment-recall curves of pgBoost and other methods on eQTL evaluation data set. Enrichment-recall curves of links predicted by pgBoost, distance, and 4 existing methods for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses.



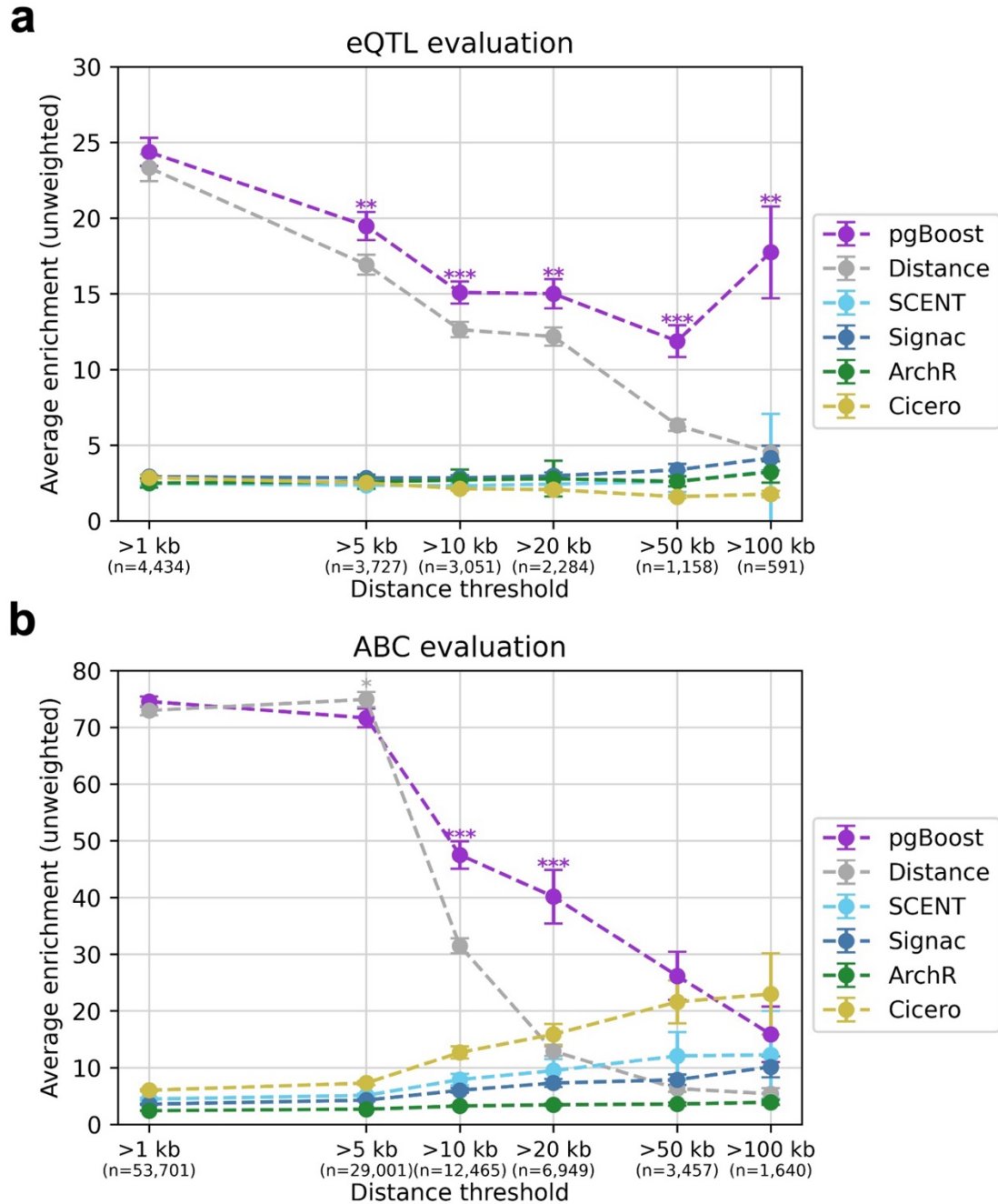
Supplementary Figure 2.4: Enrichment-recall curves of pgBoost and other methods on ABC evaluation data set. Enrichment-recall curves of links predicted by pgBoost, distance, and 4 existing methods for 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses.



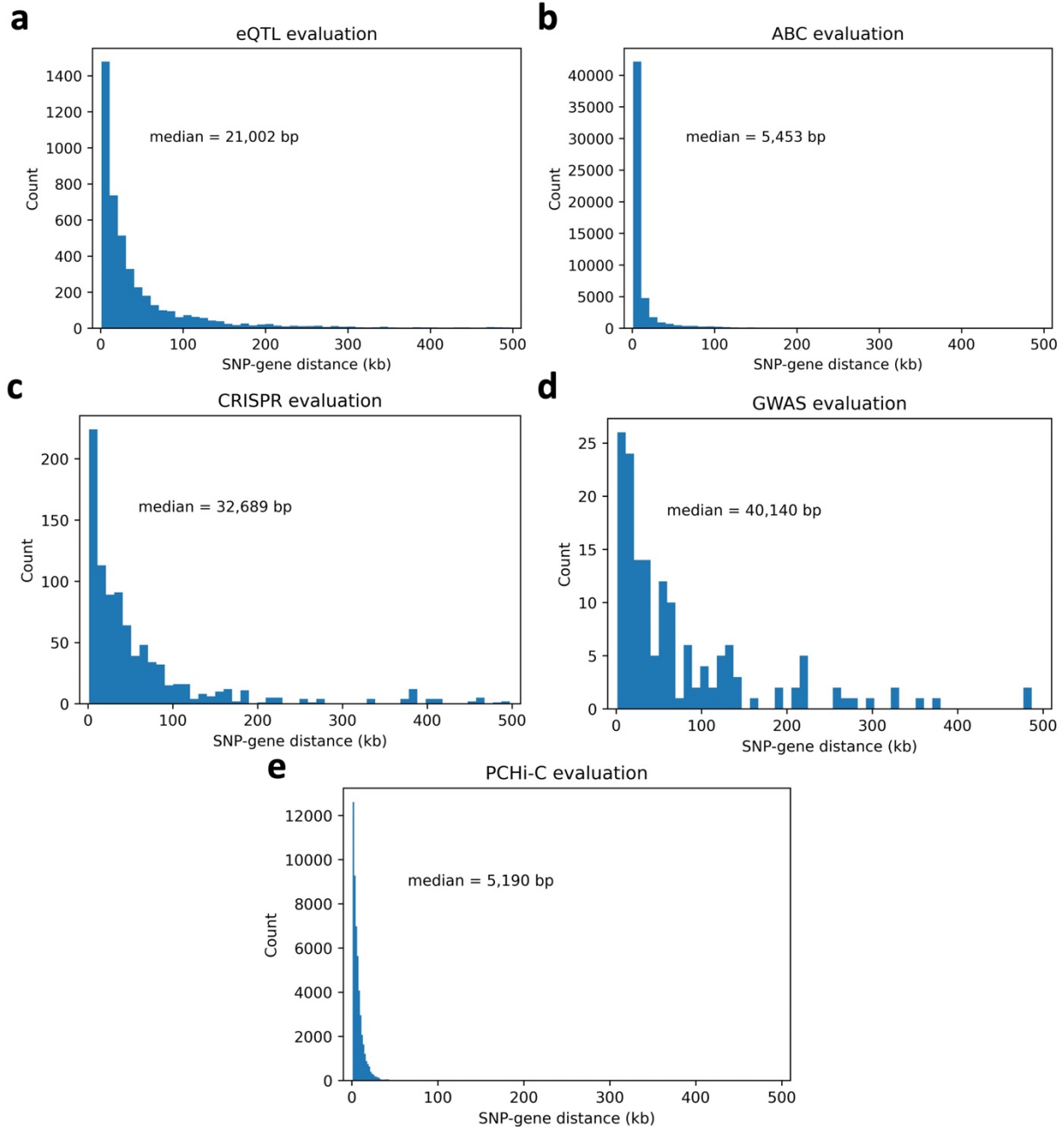
Supplementary Figure 2.5: AUPRC of pgBoost and other methods on eQTL and ABC evaluation data sets. a) Area under the precision-recall curve (AUPRC) of links predicted by pgBoost, distance, and 4 existing methods for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, at various distance thresholds. **b)** Area under the precision-recall curve (AUPRC) of links predicted by pgBoost, distance, and 4 existing methods for 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method.



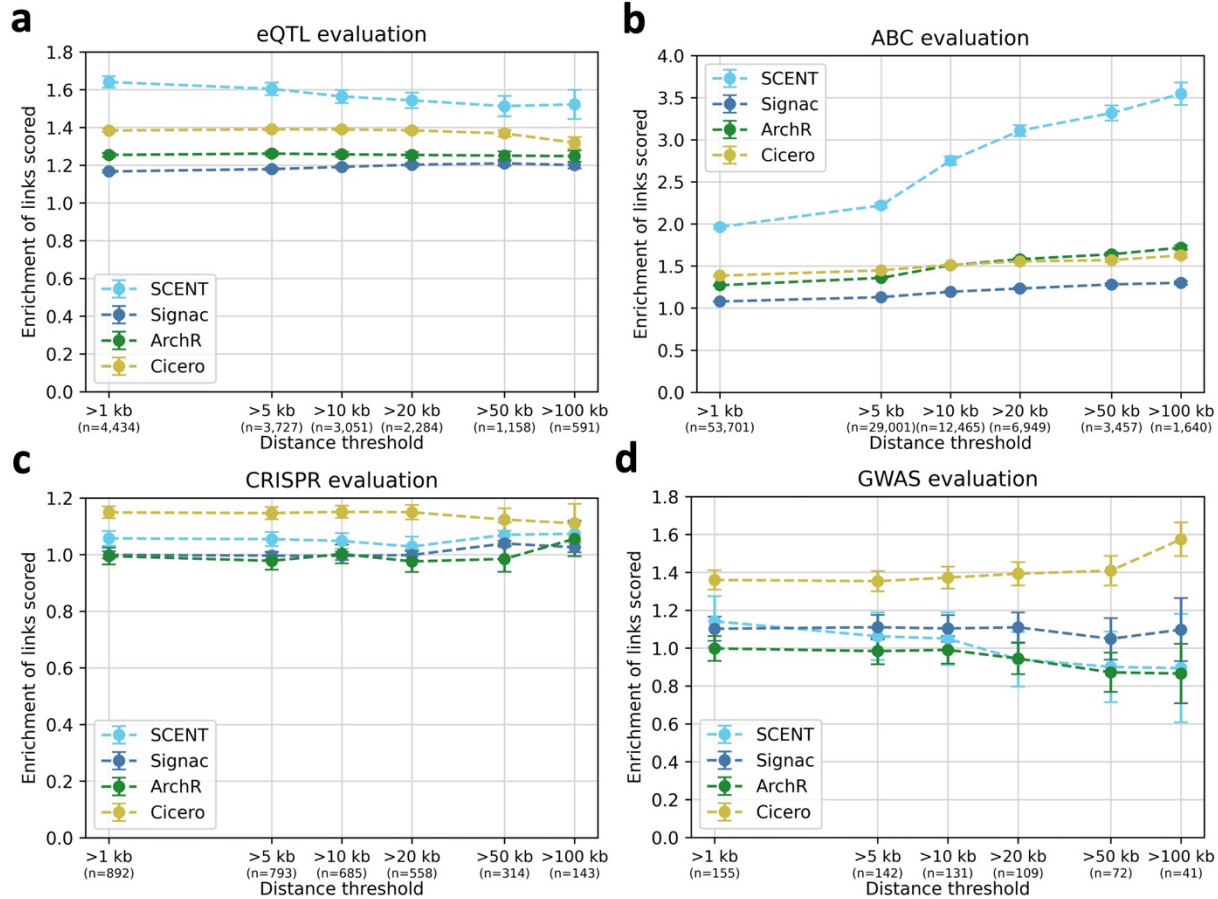
Supplementary Figure 2.6: AUROC of pgBoost and other methods on eQTL and ABC evaluation data sets. a) Area under the receiver operating characteristic curve (AUROC) of links predicted by pgBoost, distance, and 4 existing methods for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, at various distance thresholds. **b)** Area under the receiver operating characteristic curve (AUROC) of links predicted by pgBoost, distance, and 4 existing methods for 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. We note that AUPRC is often preferred over AUROC in imbalanced data with a small proportion of true positives.



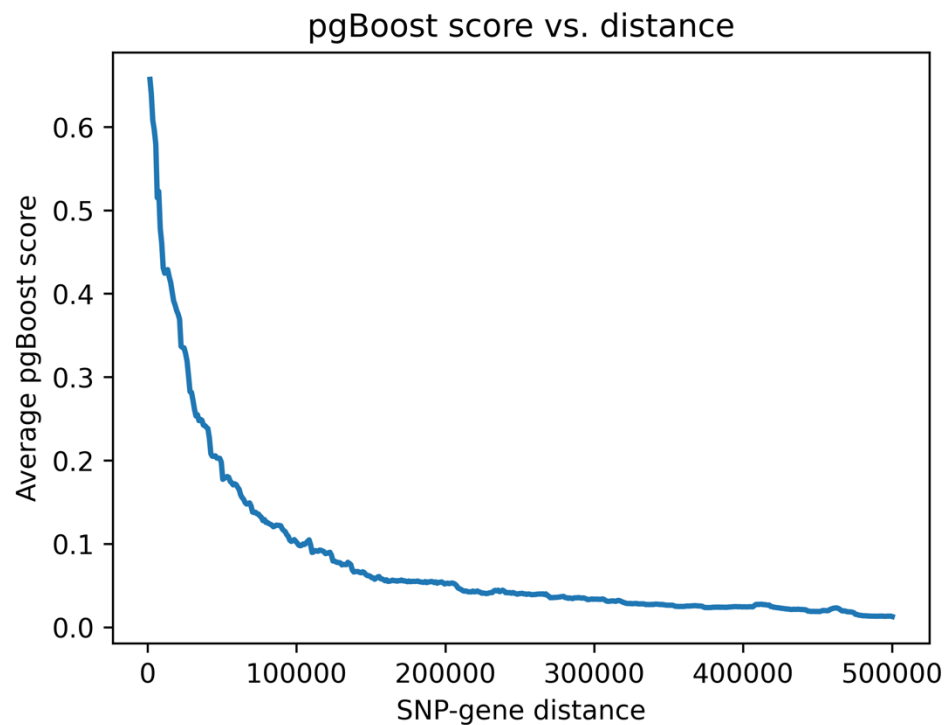
Supplementary Figure 2.7: Unweighted average enrichment of pgBoost and other methods on eQTL and ABC evaluation data sets. a) Unweighted average across recall values in $[0, 0.3]$ of links predicted by pgBoost, distance, and 4 existing methods for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, at various distance thresholds. **b)** Unweighted average across recall values in $[0, 0.4]$ of links predicted by pgBoost, distance, and 4 existing methods for 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. Average enrichments were computed over a range of recall attained by all methods (no linear extrapolation) and without weighting (**Methods**).



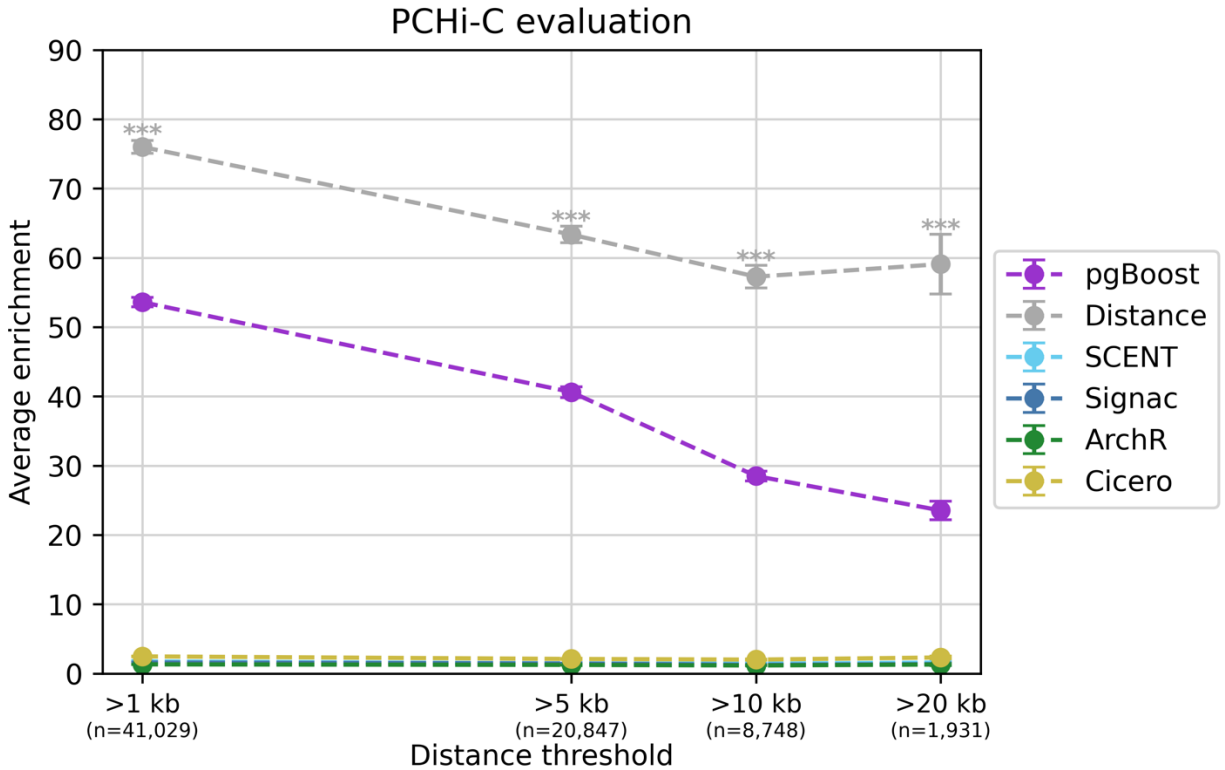
Supplementary Figure 2.8: Distance distributions of eQTL, ABC, CRISPR, GWAS, and PCHI-C evaluation data sets. SNP-gene distance distributions of evaluation sets comprised of **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, and **(e)** 41,029 SNP-gene pairs implicated by PCHI-C. The median SNP-gene distance for positive SNP-gene pairs in each evaluation data set is reported within each plot.



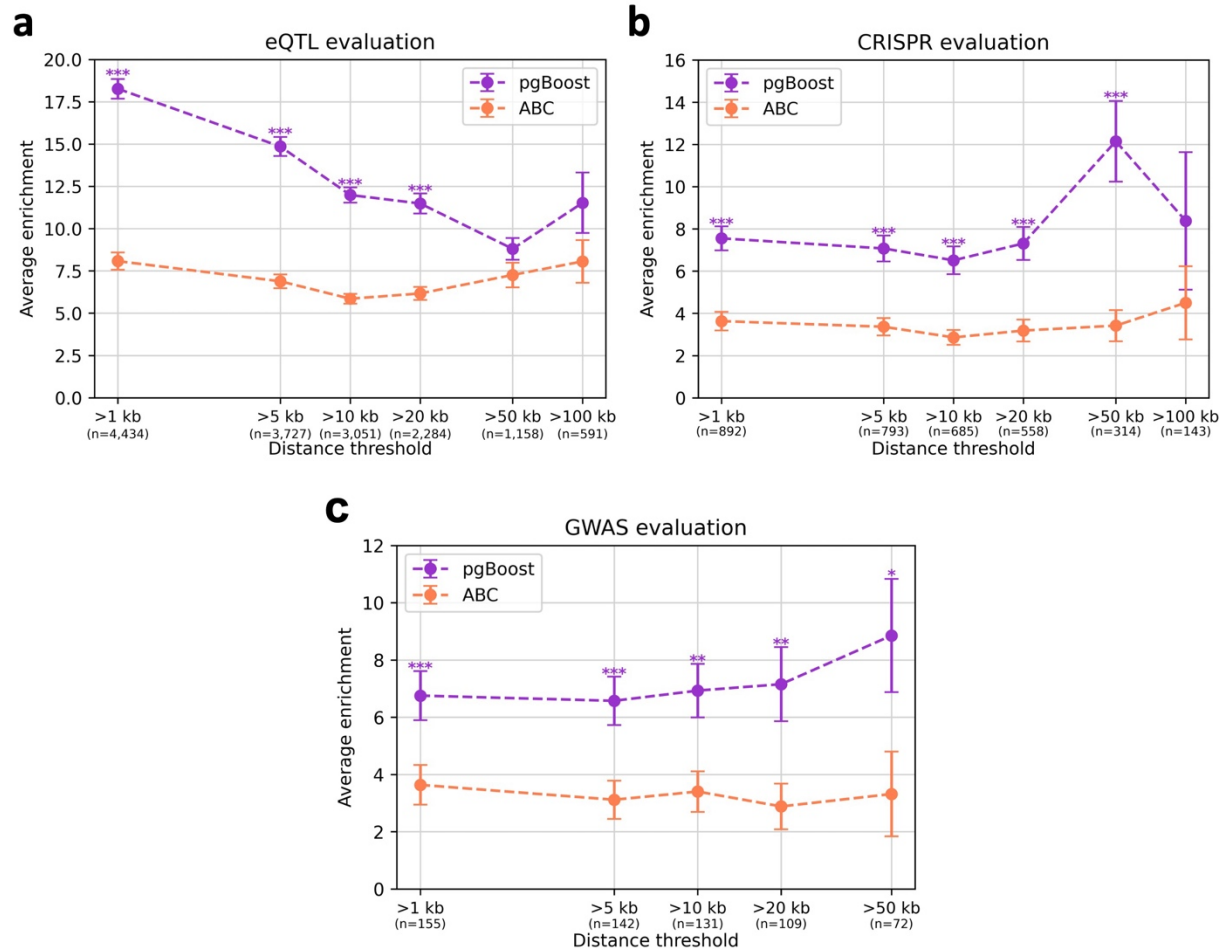
Supplementary Figure 2.9: Enrichment of candidate links scored for eQTL, ABC, CRISPR, and GWAS evaluation data sets. Enrichment of the subset of candidate SNP-gene links scored by 4 existing methods for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors.



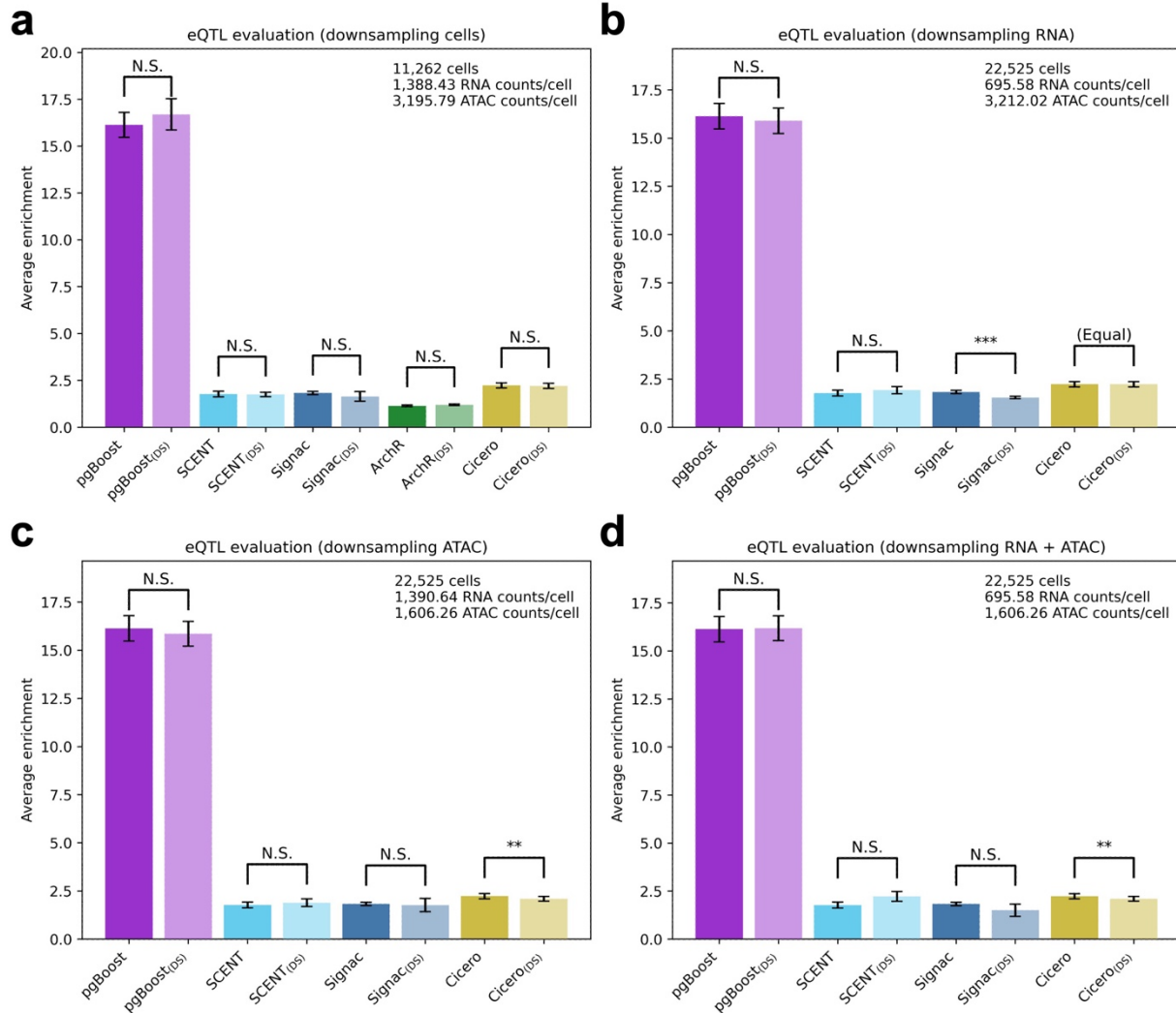
Supplementary Figure 2.10: Average pgBoost scores at different linking distances. Relationship between average SNP-gene distance and average pgBoost score across 6,882,594 candidate SNP-gene links. Candidate links are partitioned equally into 500 bins by SNP-gene distance (each represented by 1 point). In general, more proximal SNP-gene links are assigned higher pgBoost scores.



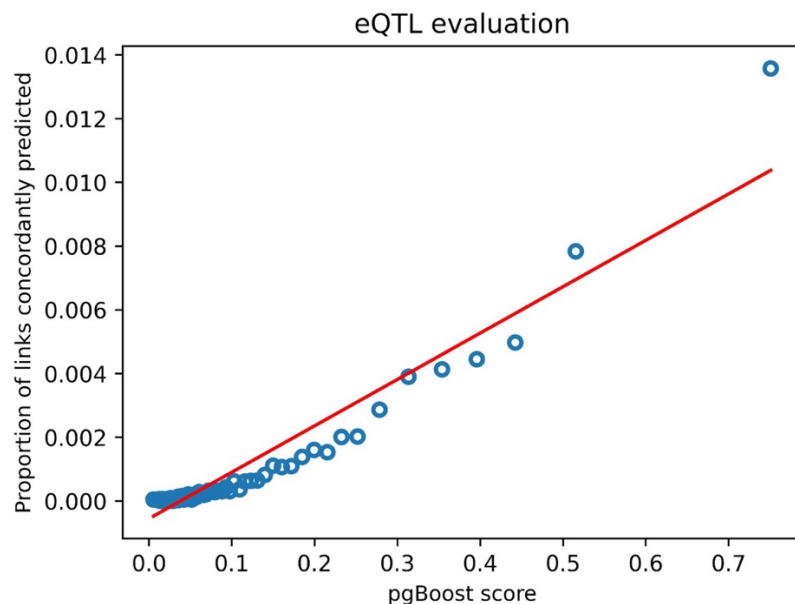
Supplementary Figure 2.11: Performance of pgBoost and other methods on promoter capture Hi-C evaluation data sets. Average enrichment across recall values of links predicted by pgBoost, distance, and 4 existing methods for 41,029 SNP-gene pairs implicated by PCHi-C (attaining maximum CHiCAGO score ≥ 5 across 17 hematopoietic cell types²⁰⁰), at various distance thresholds. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the top-performing method vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Results are reported for distance thresholds ranging from >1kb to >20kb. (We did not measure enrichments at >50kb or >100kb, because there were too few evaluation SNP-gene links at these distance thresholds (48 links >50kb and 0 links >100kb) to perform bootstrapping across genes.)



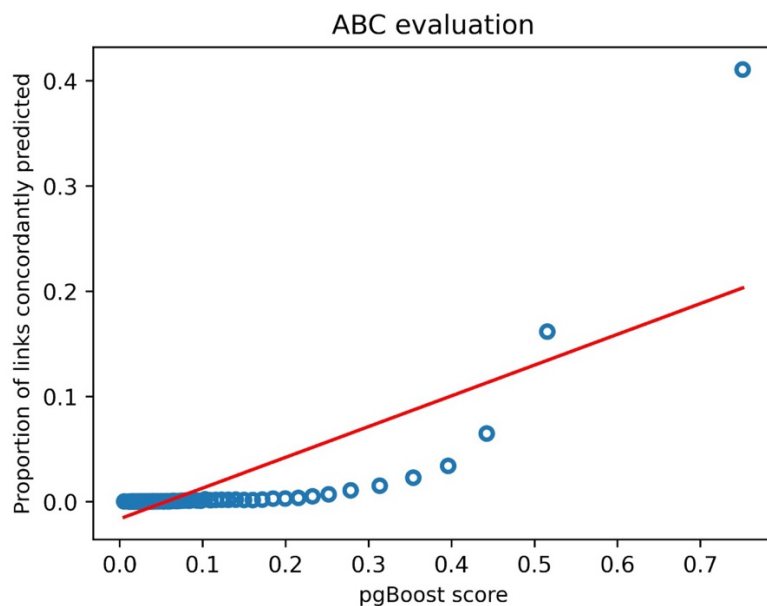
Supplementary Figure 2.12: Performance of pgBoost vs. ABC. Average enrichment across recall values of links predicted by pgBoost vs. ABC (maximum score across 344 biosamples) for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 892 links validated by CRISPR, and **(c)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method.



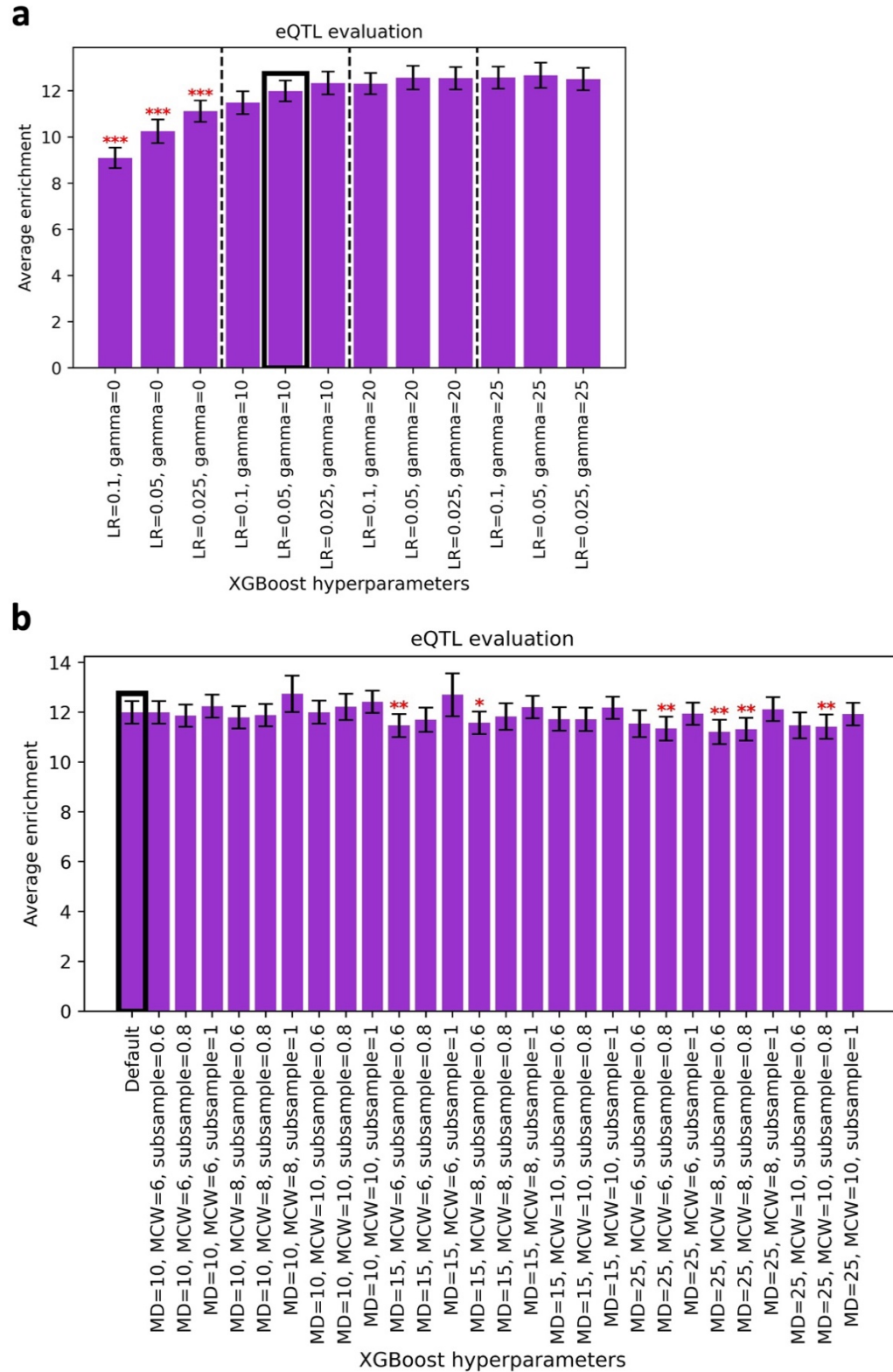
Supplementary Figure 2.13: Performance of pgBoost and existing methods in downsampled conditions. Average enrichment across recall values of links predicted by pgBoost and 4 existing methods using all data from NeurIPS T cells (**Table 2**) vs. downsampling to **(a)** 50% of cells, **(b)** 50% of RNA counts, **(c)** 50% of ATAC counts, **(d)** 50% of RNA counts and 50% of ATAC counts, for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Dark and light bars report average enrichments for each method in non-downsampled and downsampled (“DS”) conditions, respectively. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, N.S.: non-significant) of focal method in downsampled vs. non-downsampled conditions. Downsampling conditions (number of cells, average per-cell RNA and ATAC counts) are reported within each plot. Because ArchR software does not support modification of the ATAC peak count matrix, ArchR is excluded from analyses **b-d**. We note that Cicero performance is unaffected by RNA count downsampling **(b)** because Cicero only uses ATAC-seq data as input.



b

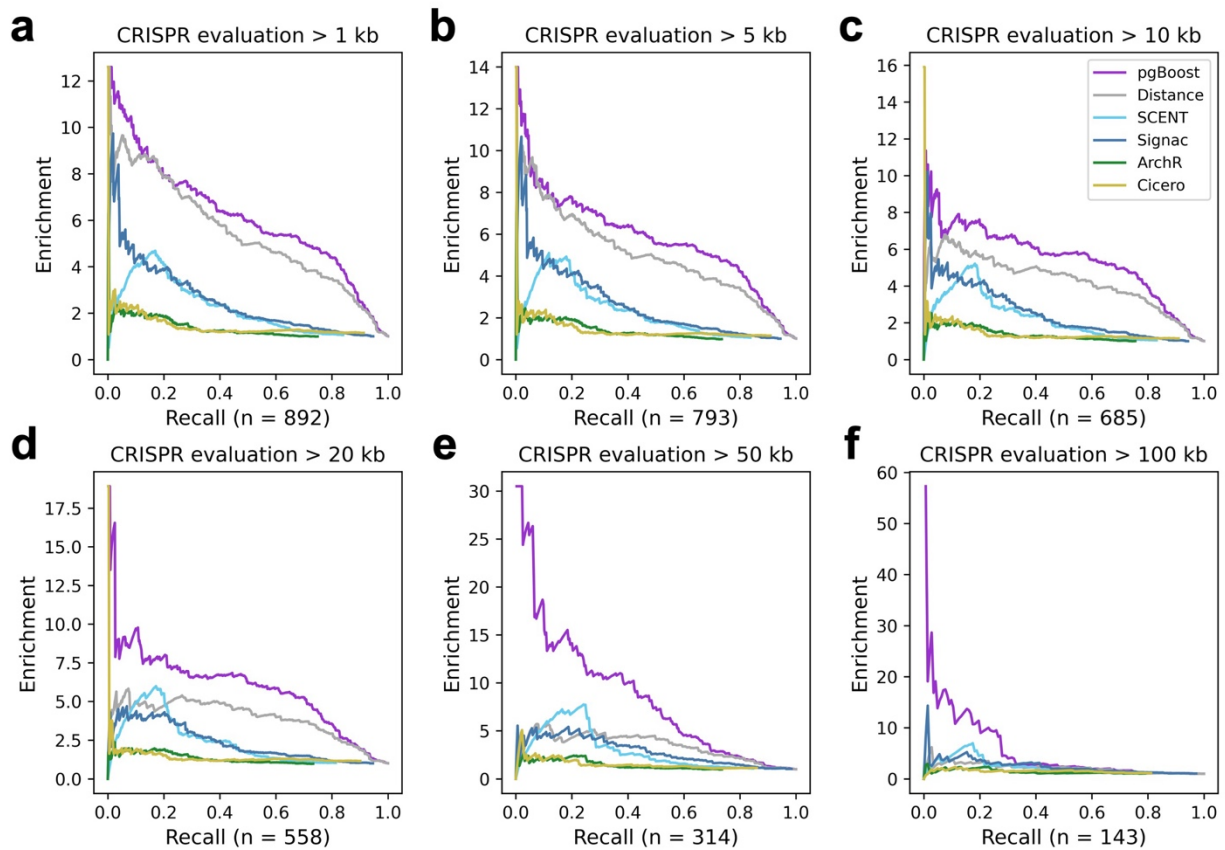


Supplementary Figure 2.14: Calibration of the pgBoost probabilistic score. Relationship between average pgBoost score and the proportion of links concordantly predicted by **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues and **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples. All 6,882,594 candidate links are partitioned equally into 100 bins by pgBoost score (each represented by 1 point). The discrepancy between predicted probability and actual proportion of concordantly predicted links in **a** is expected given the different PIP threshold used to define model training (PIP > 0.2) vs. model evaluation (PIP > 0.5) sets as well as the different definitions of “negative” SNP-gene links in model training (link includes a gene in the set of positive links) vs. enrichment computations (any candidate SNP-gene link that is not a positive link) (**Methods**). The relationship between proportion of links concordantly predicted and pgBoost score is not perfectly linear, supporting the use of binarized sets of top-scoring links predicted by pgBoost rather than the probabilistic score in downstream analyses.

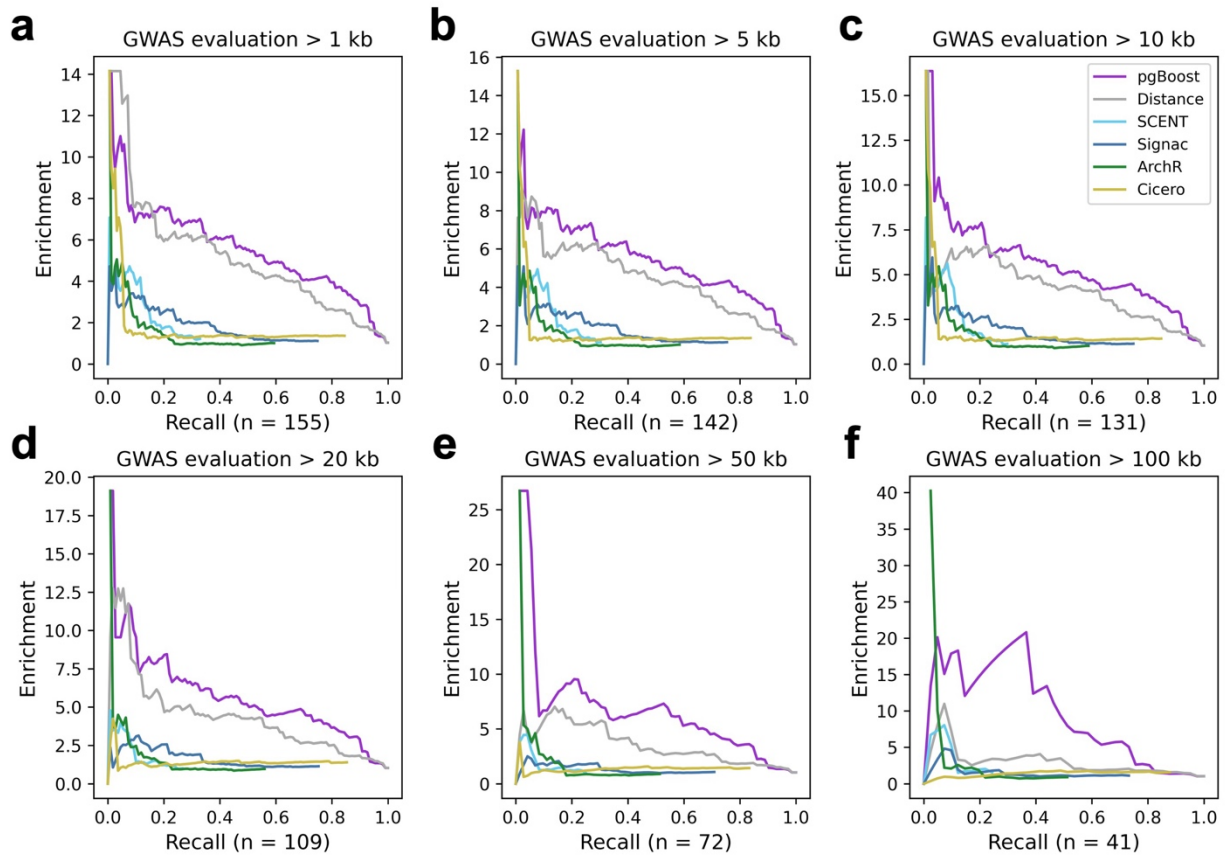


Supplementary Figure 2.15: Performance of pgBoost trained using alternative gradient boosting hyperparameters. a) Average enrichment across recall values of links predicted by pgBoost trained using alternative values for the hyperparameters learning rate and gamma for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. b)

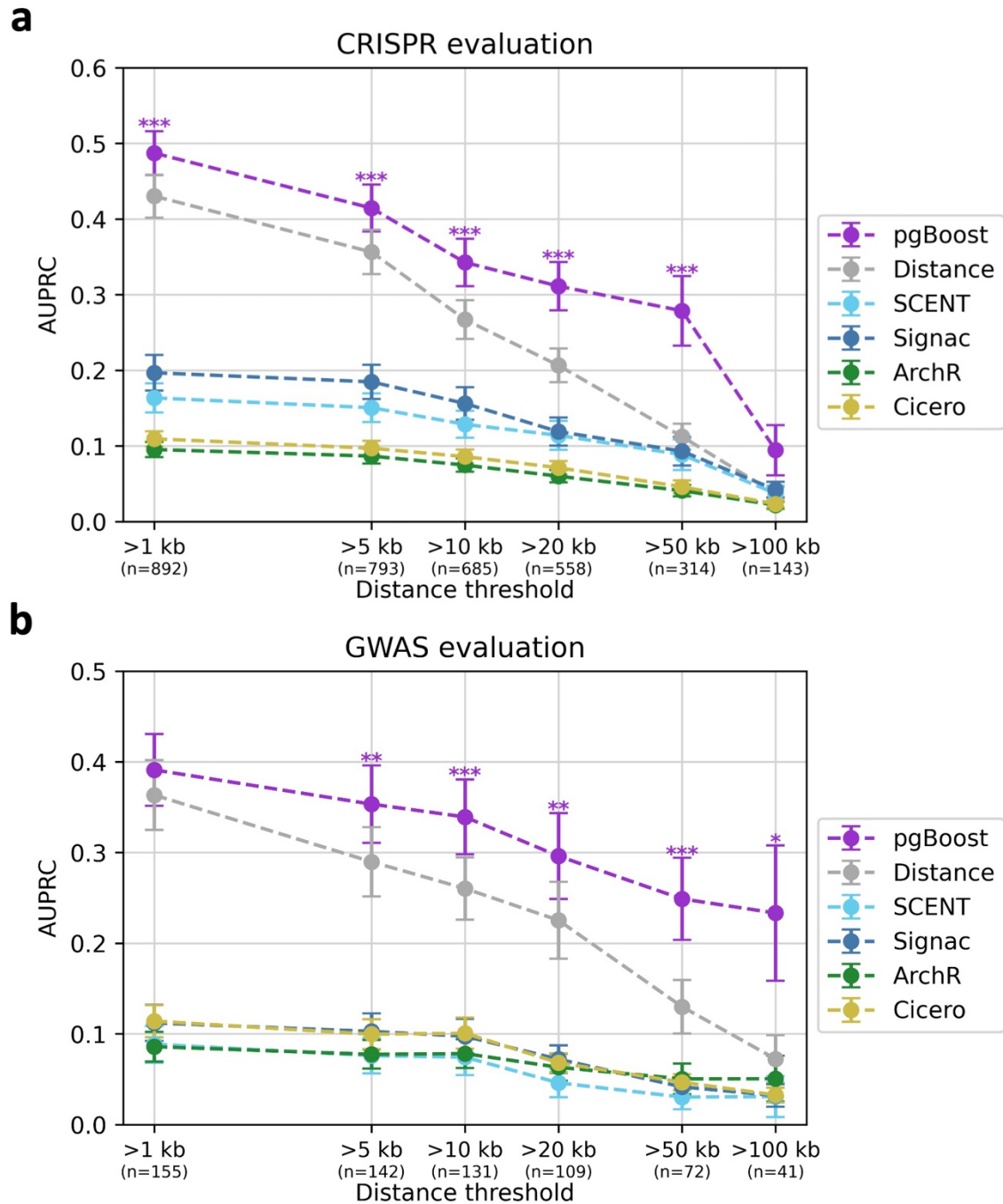
Average enrichment across recall values of links predicted by pgBoost trained using alternative values for the hyperparameters maximum tree depth, minimum child weight, and subsampling ratio for 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the default pgBoost hyperparameters (denoted in black outline) vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Red stars indicate that the focal method significantly underperforms pgBoost. LR = learning rate, MD = maximum tree depth, MCW = minimum child weight. In **a**, the default values of learning rate (0.05) and gamma (10) significantly outperformed models trained using a gamma value of 0 and generally performed similarly to other values of gamma and learning rate; a higher value of gamma implements stronger regularization and prevents overfitting. In **b**, the default approach (denoted in black outline) for selecting values of maximum tree depth, minimum child weight, and subsampling ratio (via randomized search and cross validation (**Methods**)) did not significantly underperform any fixed combination of these parameter values.



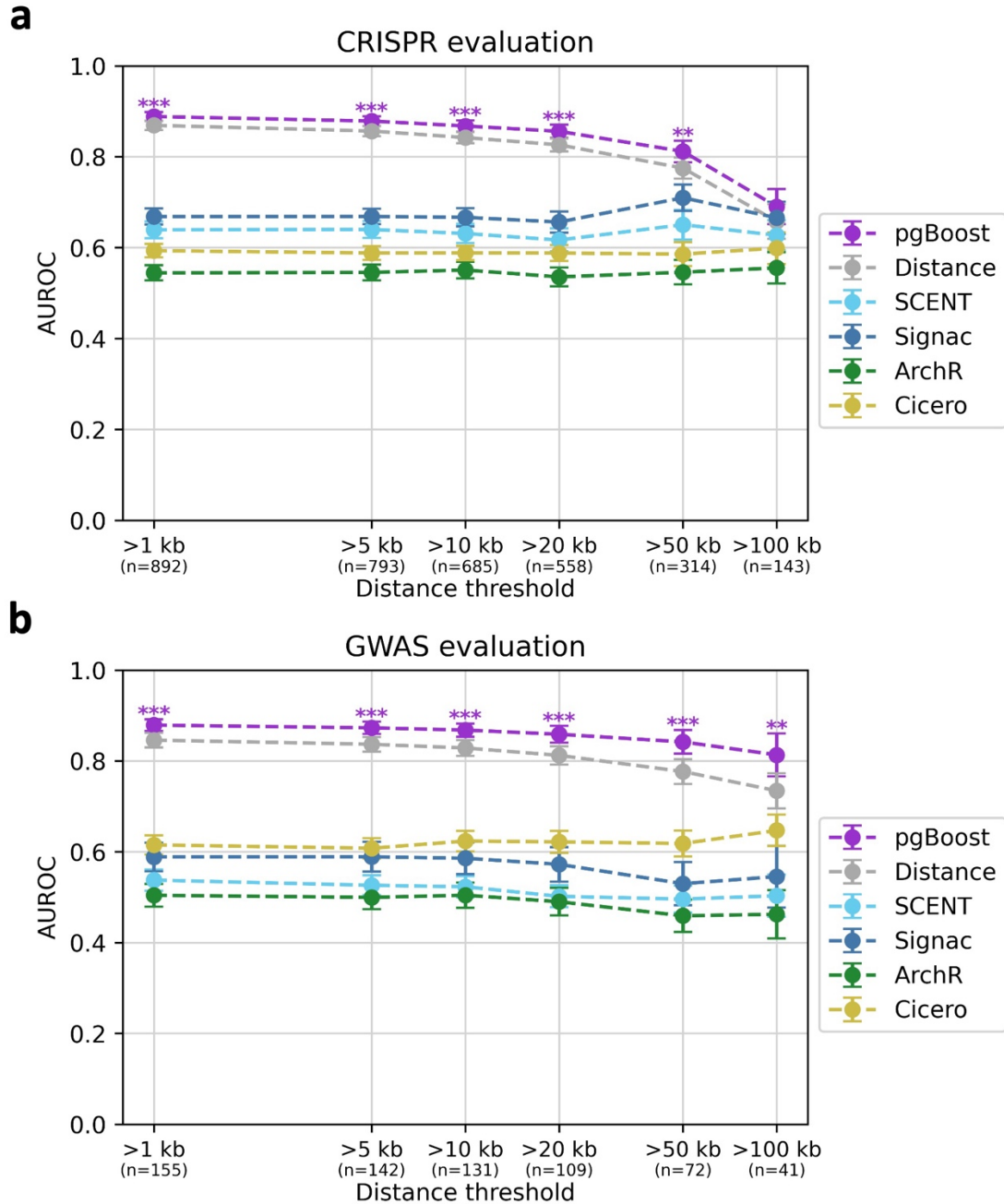
Supplementary Figure 2.16: Enrichment-recall curves of pgBoost and other methods on CRISPR evaluation data set. Enrichment-recall curves of links predicted by pgBoost, distance, and 4 existing methods for 892 links validated by CRISPR, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses.



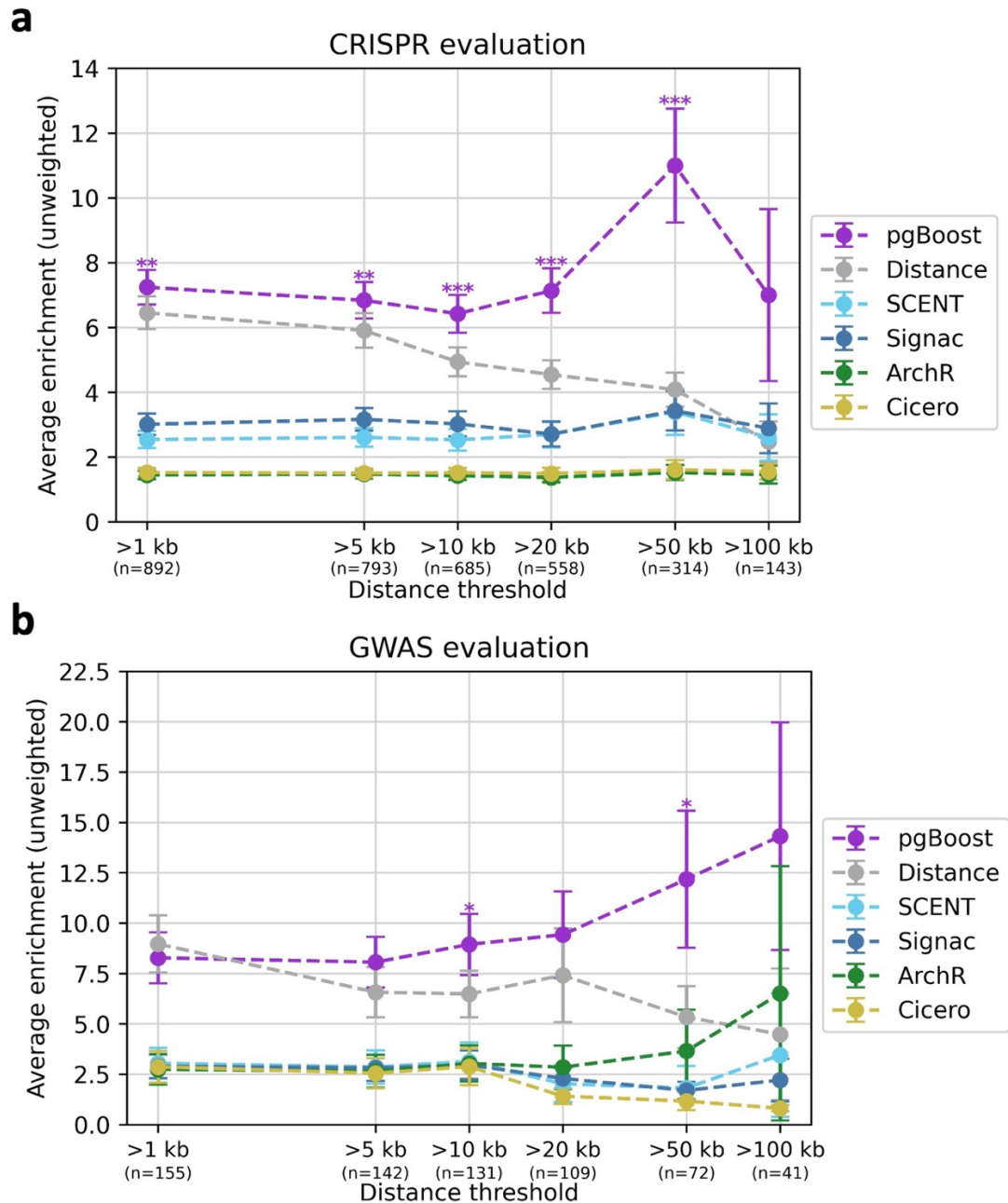
Supplementary Figure 2.17: Enrichment-recall curves of pgBoost and other methods on GWAS-derived evaluation data set. Enrichment-recall curves of links predicted by pgBoost, distance, and 4 existing methods for 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses.



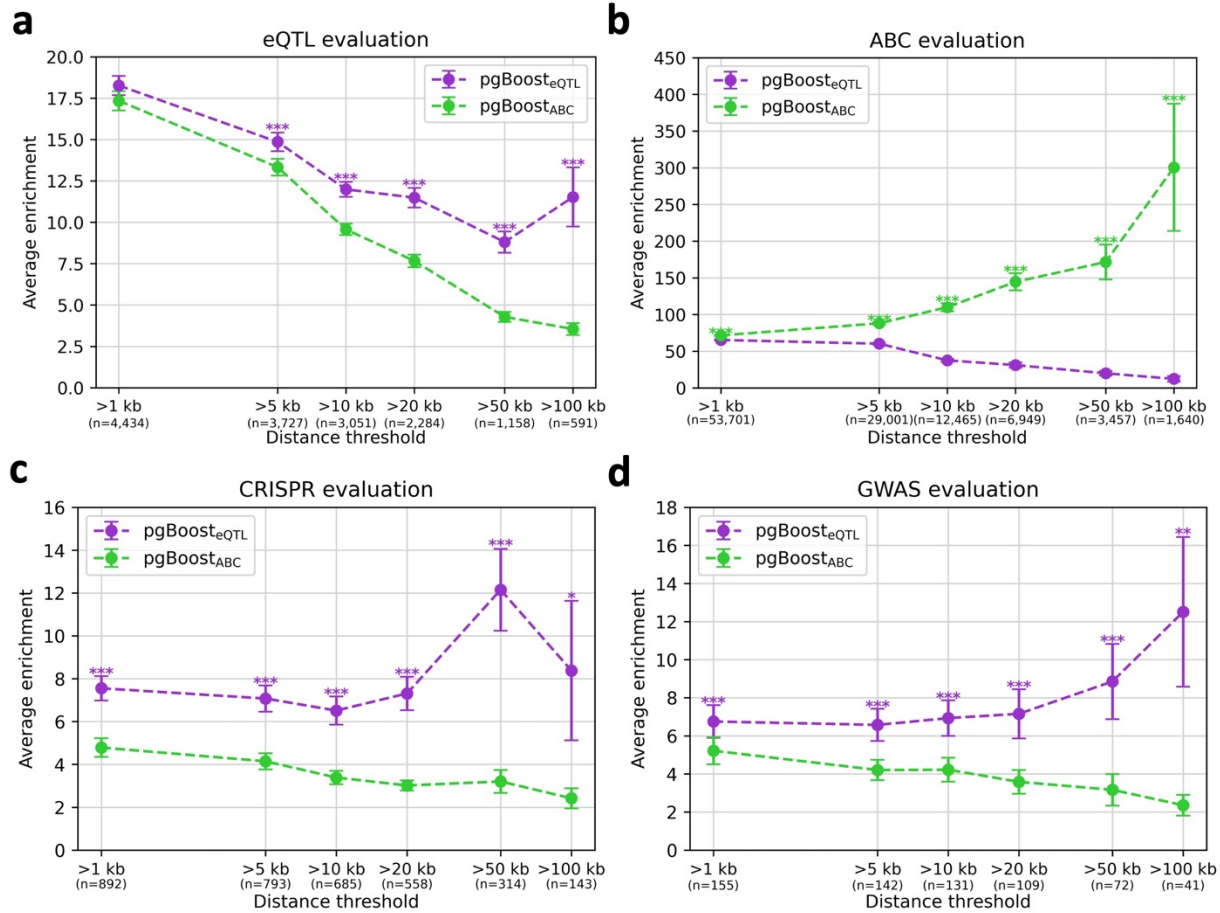
Supplementary Figure 2.18: AUPRC of pgBoost and other methods on gold-standard evaluation data sets. a) Area under the precision-recall curve (AUPRC) of links predicted by pgBoost, distance, and 4 existing methods for 892 links validated by CRISPR, at various distance thresholds. **b)** Area under the precision-recall curve (AUPRC) of links predicted by pgBoost, distance, and 4 existing methods for 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method.



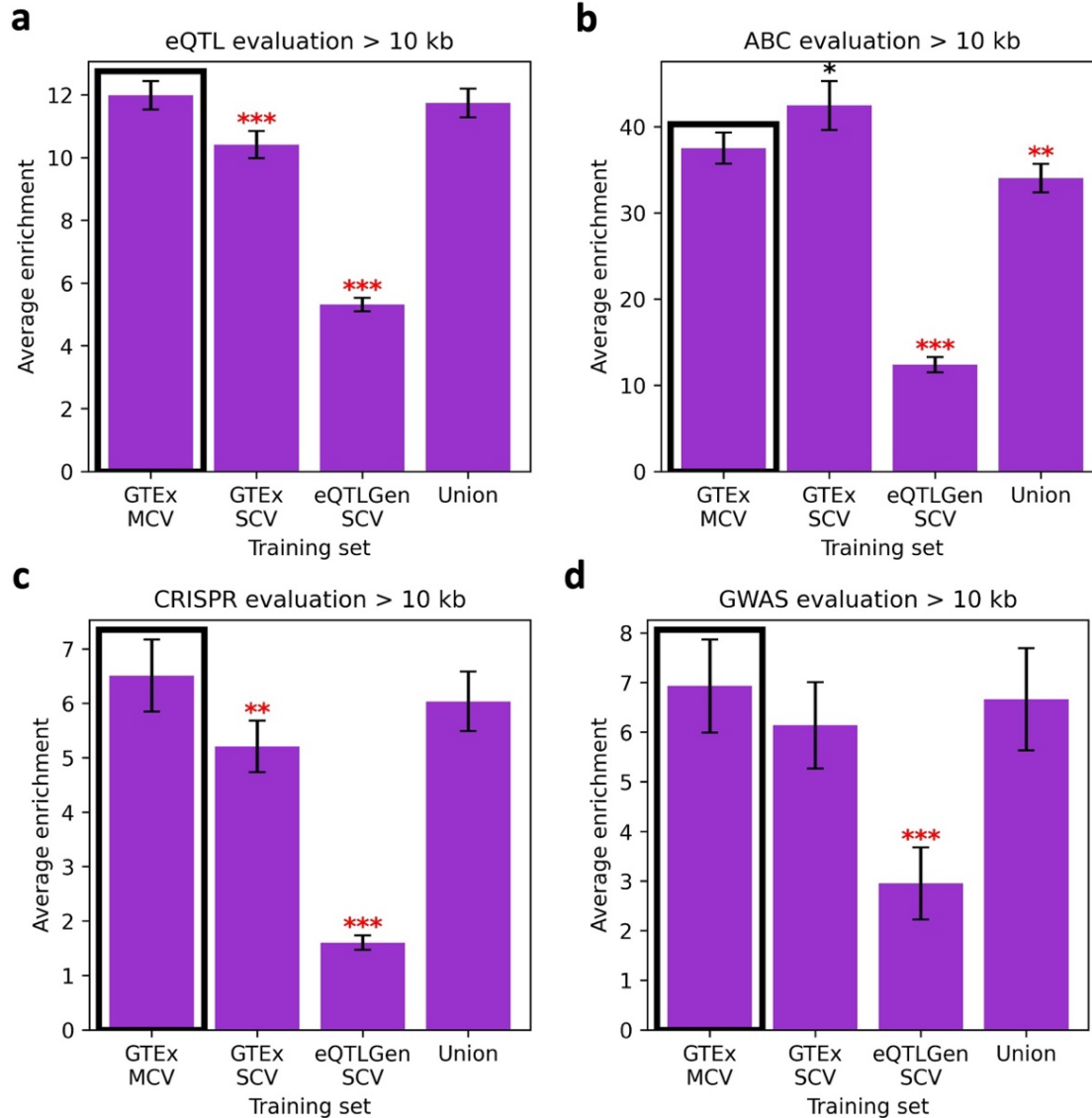
Supplementary Figure 2.19: AUROC of pgBoost and other methods on gold-standard evaluation data sets. a) Area under the receiver operating characteristic curve (AUROC) of links predicted by pgBoost, distance, and 4 existing methods for 892 links validated by CRISPR, at various distance thresholds. **b)** Area under the receiver operating characteristic curve (AUROC) of links predicted by pgBoost, distance, and 4 existing methods for 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. We note that AUPRC is often preferred over AUROC in imbalanced data with a small proportion of true positives.



Supplementary Figure 2.20: Unweighted average enrichment of pgBoost and other methods on gold-standard evaluation data sets. a) Unweighted average across recall values in $[0, 0.73]$ of links predicted by pgBoost, distance, and 4 existing methods for 892 links validated by CRISPR, at various distance thresholds. **b)** Unweighted average across recall values in $[0, 0.25]$ of links predicted by pgBoost, distance, and 4 existing methods for 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. Average enrichments were computed over a range of recall attained by all methods (no linear extrapolation) and without weighting (**Methods**).



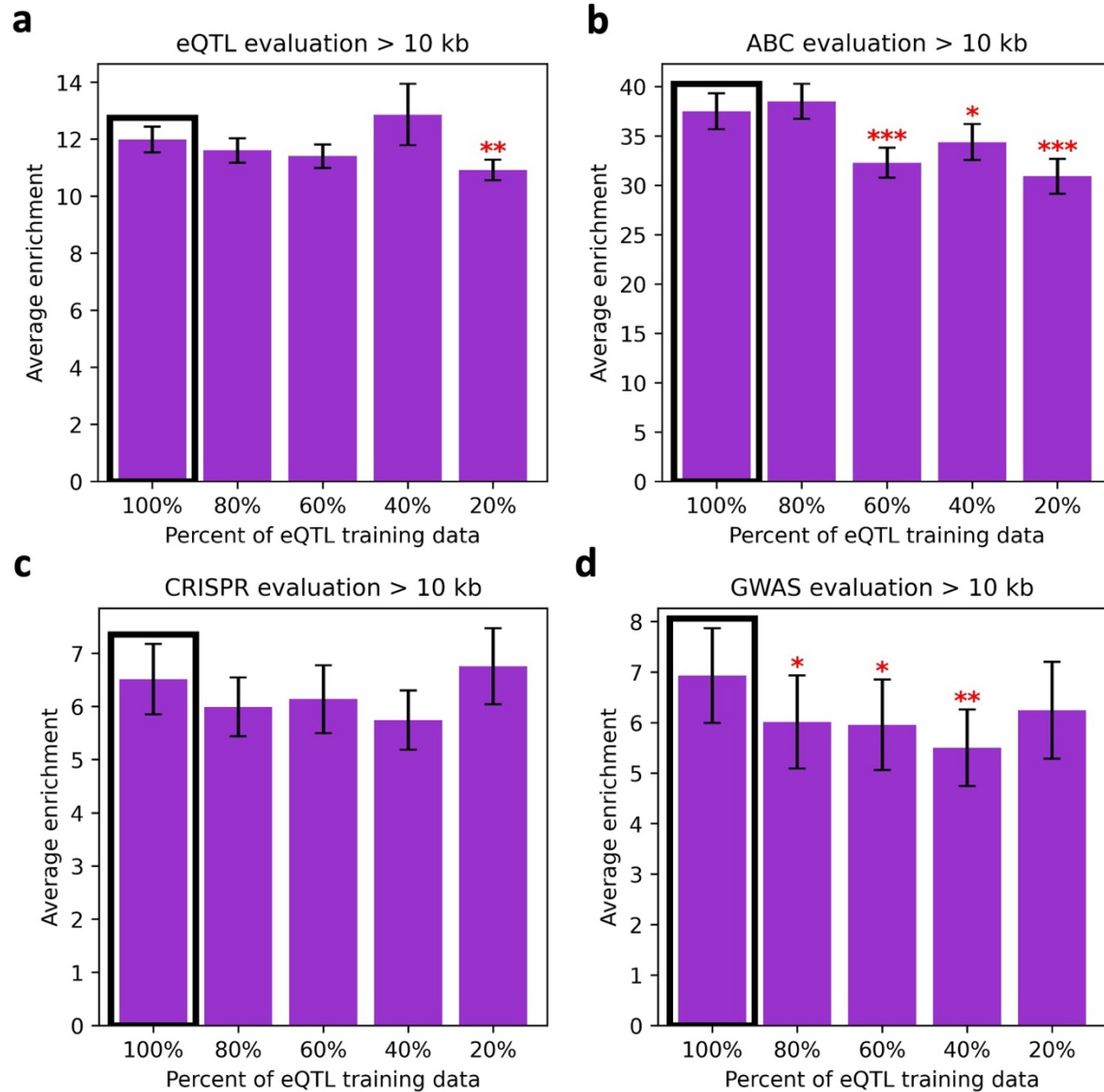
Supplementary Figure 2.21: Performance of pgBoost trained on eQTL vs. ABC data. Average enrichment across recall values of links predicted by pgBoost trained on fine-mapped eSNP-eGene pairs vs. ABC links for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. For pgBoost_{ABC}, we defined a training set comprised of 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples as positives and 1,207,294 SNP-gene pairs including a gene in the positive set and attaining maximum ABC score < 0.015 across 344 biosamples as negatives.



Supplementary Figure 2.22: Performance of pgBoost trained on different eQTL data sets.

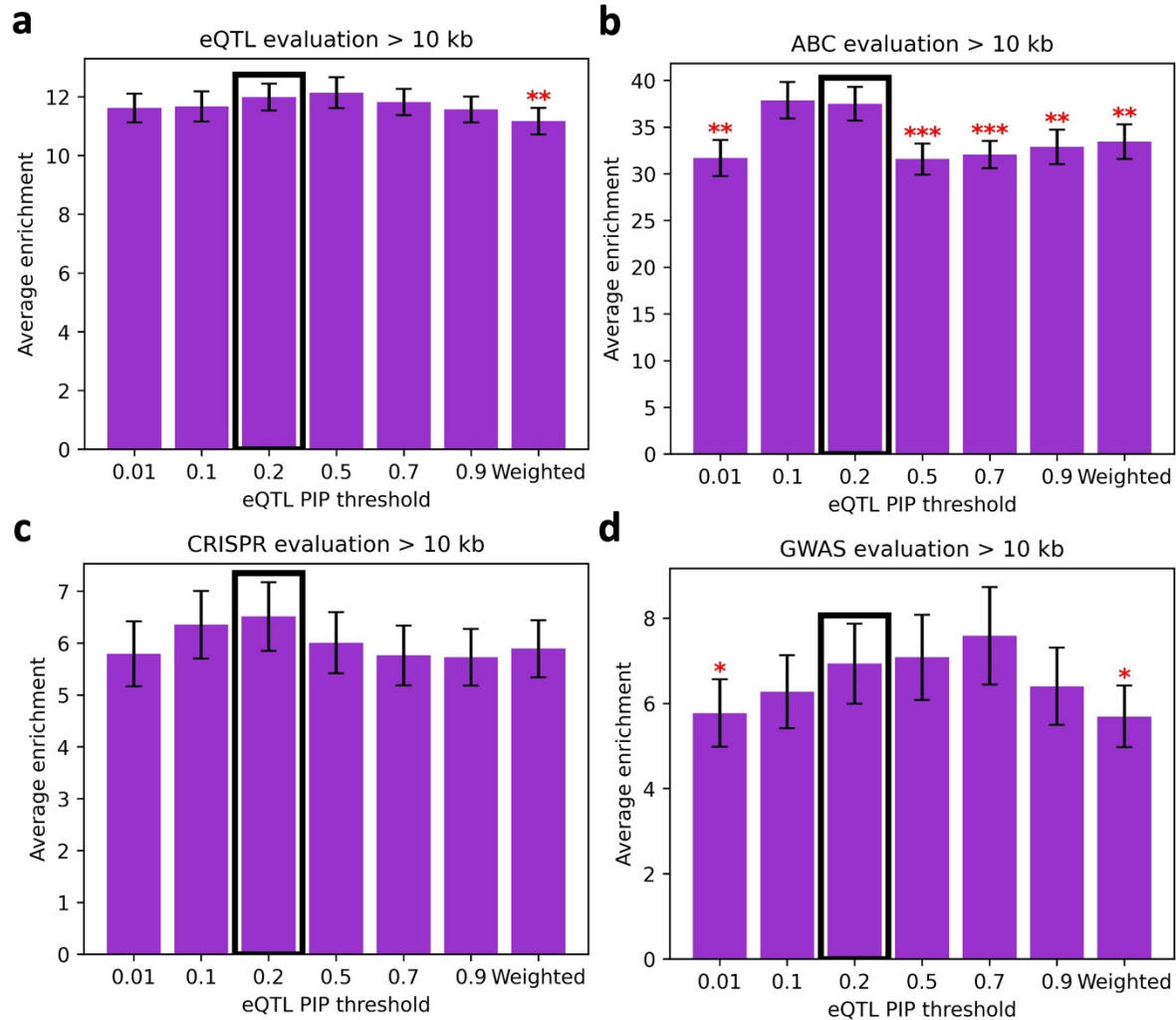
Average enrichment across recall values of links predicted by pgBoost trained on fine-mapped eSNP-eGene pairs from different eQTL catalogs for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the main pgBoost model (denoted in black outline) vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Red (resp. black) stars indicate that the focal method significantly underperforms (resp. outperforms) pgBoost. In addition to our primary training set from GTEx multiple causal variant fine-mapping (GTEx-MCV), we considered a GTEx single causal variant fine-mapping training set (GTEx-SCV), an eQTLGen¹⁷³ single causal variant fine-mapping training set (eQTLGen-SCV), and the union of all three training sets (Union). We performed fine-mapping using SuSIE¹¹⁴ to compute PIPs for the GTEx-SCV and eQTLGen-SCV training sets (see Data Availability); we used approximate Bayes factors²²⁸ to construct the eQTLGen-SCV training set.

For each training set, we defined eSNP-eGene pairs attaining maximum PIP > 0.2 as positives and eSNP-eGene pairs including a gene in the positive set and attaining maximum PIP < 0.01 as negatives (yielding 11,926 positives for GTEx-MCV; 8,279 positives for GTEx-SCV; 2,296 positives for eQTLGen-SCV; and 13,457 positives for the Union training set). GTEx-MCV significantly outperformed GTEx-SCV on eQTL and CRISPR evaluations; while GTEx-MCV significantly underperformed GTEx-SCV on ABC evaluation data, we prioritized results on eQTL and CRISPR evaluations (see **Methods**). GTEx-MCV significantly outperformed eQTLGen-SCV on all 4 evaluations. GTEx-MCV outperformed the Union training set on ABC evaluation data but performed similarly on eQTL, CRISPR, and ABC evaluations, supporting our choice to train using fine-mapped eSNP-eGene pairs from GTEx. (We note that eQTLGen-MCV is not feasible because multiple causal variant fine-mapping requires in-sample LD from individual-level data¹³⁸, which is not available for eQTLGen.) (We also note that GTEx-SCV is an approximate subset of GTEx-MCV.)



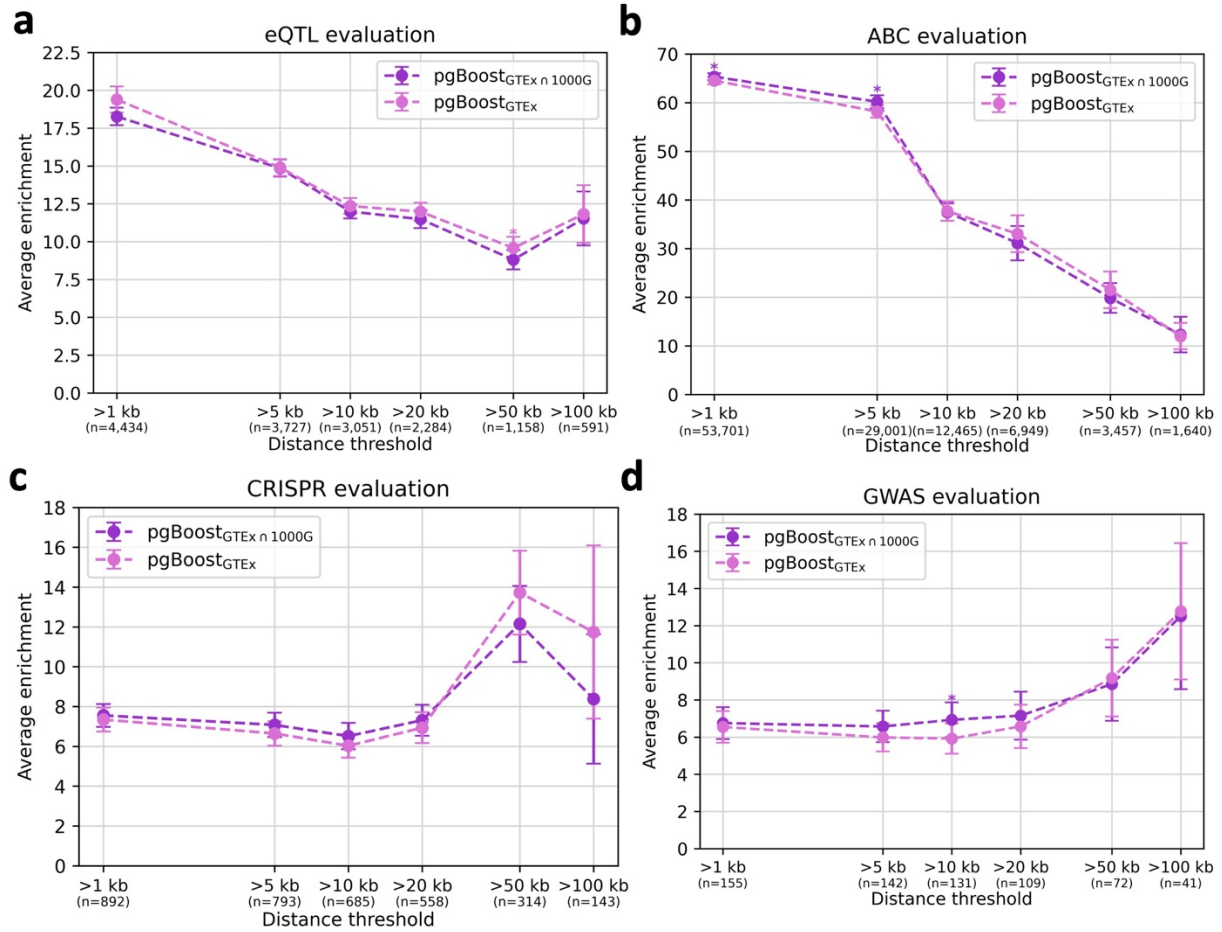
Supplementary Figure 2.23: Performance of pgBoost trained on downsampled eQTL data.

Average enrichment across recall values of links predicted by pgBoost trained on random subsamples (20-80%) of GTEx fine-mapped eSNP-eGene pairs for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference of the main pgBoost model (denoted in black outline) vs. each other method (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Red stars indicate that the focal method significantly underperforms pgBoost. While more severe downsampling of the training data yielded significant reductions in pgBoost performance, mild downsampling did not substantially impact performance. This suggests that there may be diminishing returns from incorporating additional training data (consistent with our results for the Union training set in **Supplementary Figure 2.22**).

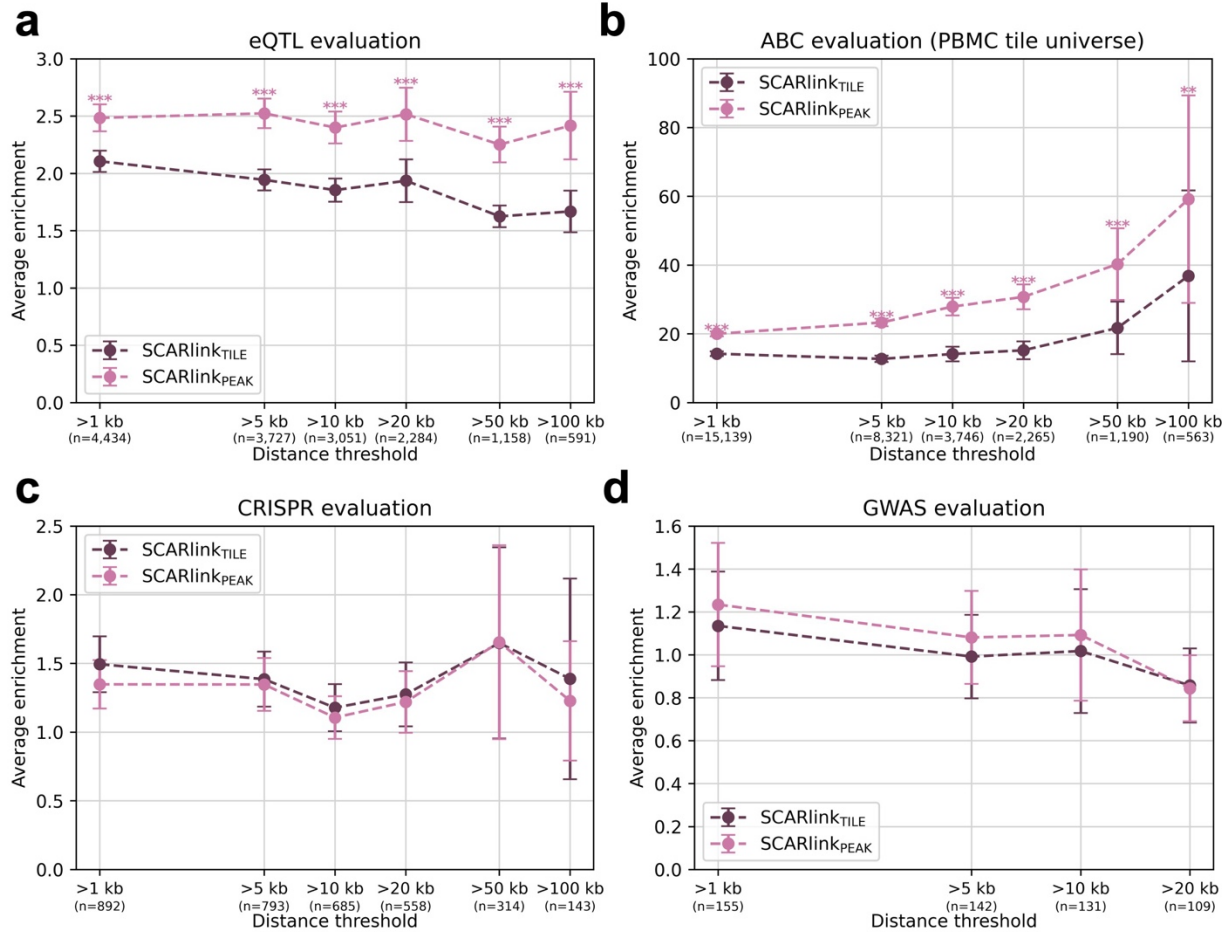


Supplementary Figure 2.24: Performance of pgBoost trained using various eQTL PIP thresholds. Average enrichment across recall values of links predicted by pgBoost trained on eSNP-eGene pairs passing varying PIP thresholds for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of focal method vs. top method (denoted in black outline). Red stars indicate that the focal method significantly underperforms pgBoost. We defined training sets comprised of SNP-gene links attaining maximum PIP across GTEx tissues above each threshold as positives and SNP-gene links including a gene in the positive set and attaining maximum PIP < 0.1 across GTEx tissues as negatives. Using this strategy, we defined training sets of fine-mapped eSNP-eGene pairs passing PIP thresholds of 0.01 (124,472 positives), 0.1 (22,668 positives), 0.2 (11,926 positives), 0.5 (4,435 positives), 0.7 (3,033 positives), 0.9 (2,008 positives). We also assessed a “weighted” training set where SNP-gene pairs attaining PIP > 0.8 were assigned a weight of 4, SNP-gene pairs attaining PIP > 0.4 and < 0.8 were assigned a weight of 2, and SNP-gene pairs attaining PIP > 0.1 and < 0.4 were assigned a weight of 1 (and SNP-gene links including a gene in the positive set and attaining maximum PIP

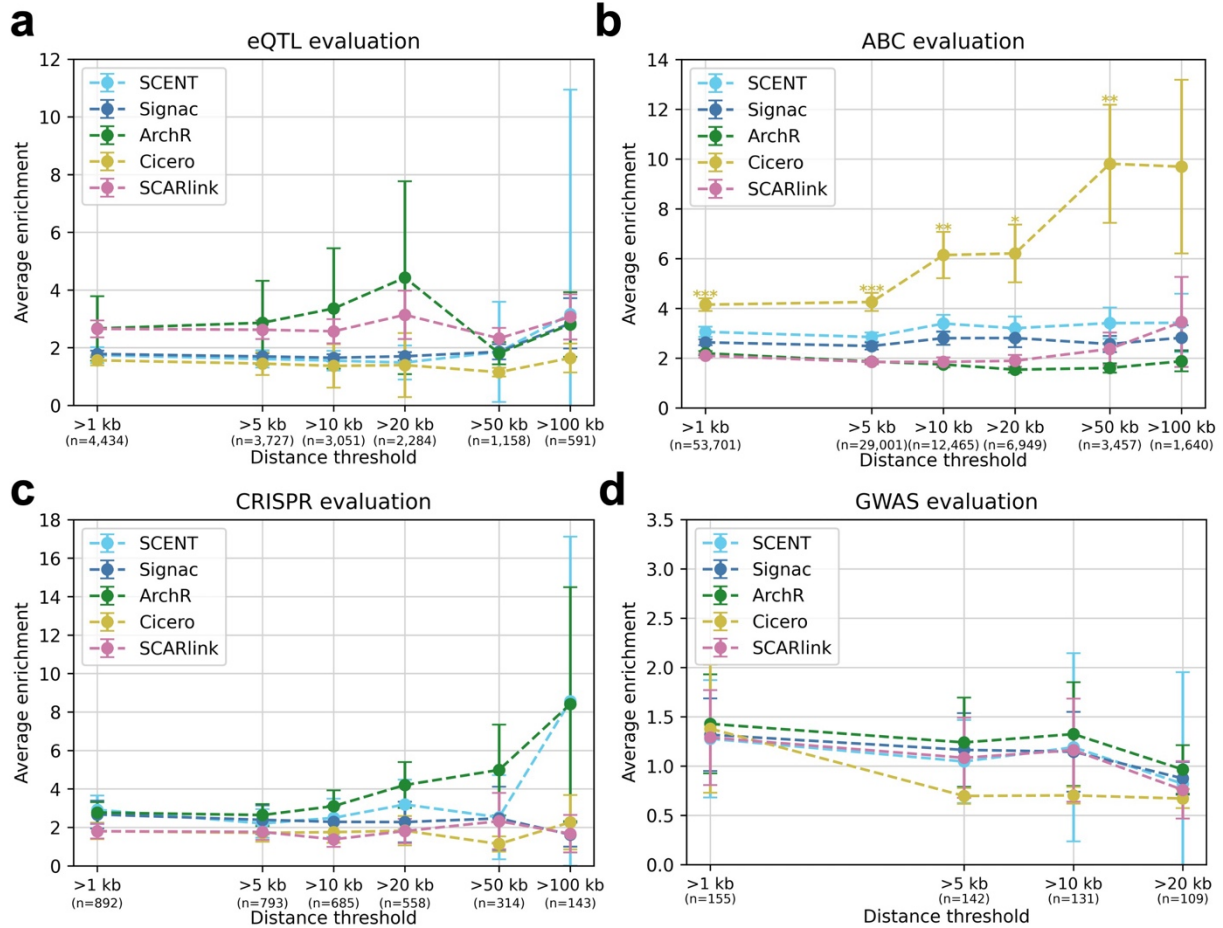
< 0.1 were defined as negatives). We determined that: the threshold of $PIP > 0.2$ significantly outperformed the less stringent PIP threshold of $PIP > 0.01$ on ABC and GWAS evaluations; $PIP > 0.2$ significantly outperformed the more stringent PIP thresholds of $PIP > 0.5$, $PIP > 0.7$, and $PIP > 0.9$ on ABC evaluation data; and $PIP > 0.2$ significantly outperformed the weighted training strategy on eQTL, ABC, and GWAS evaluations. Average enrichment differences between training sets were not statistically significant on CRISPR evaluation data. While thresholds of $PIP > 0.2$ and $PIP > 0.1$ yielded statistically similar performance on all evaluation sets, we elected to train using $PIP > 0.2$ to limit computational cost.



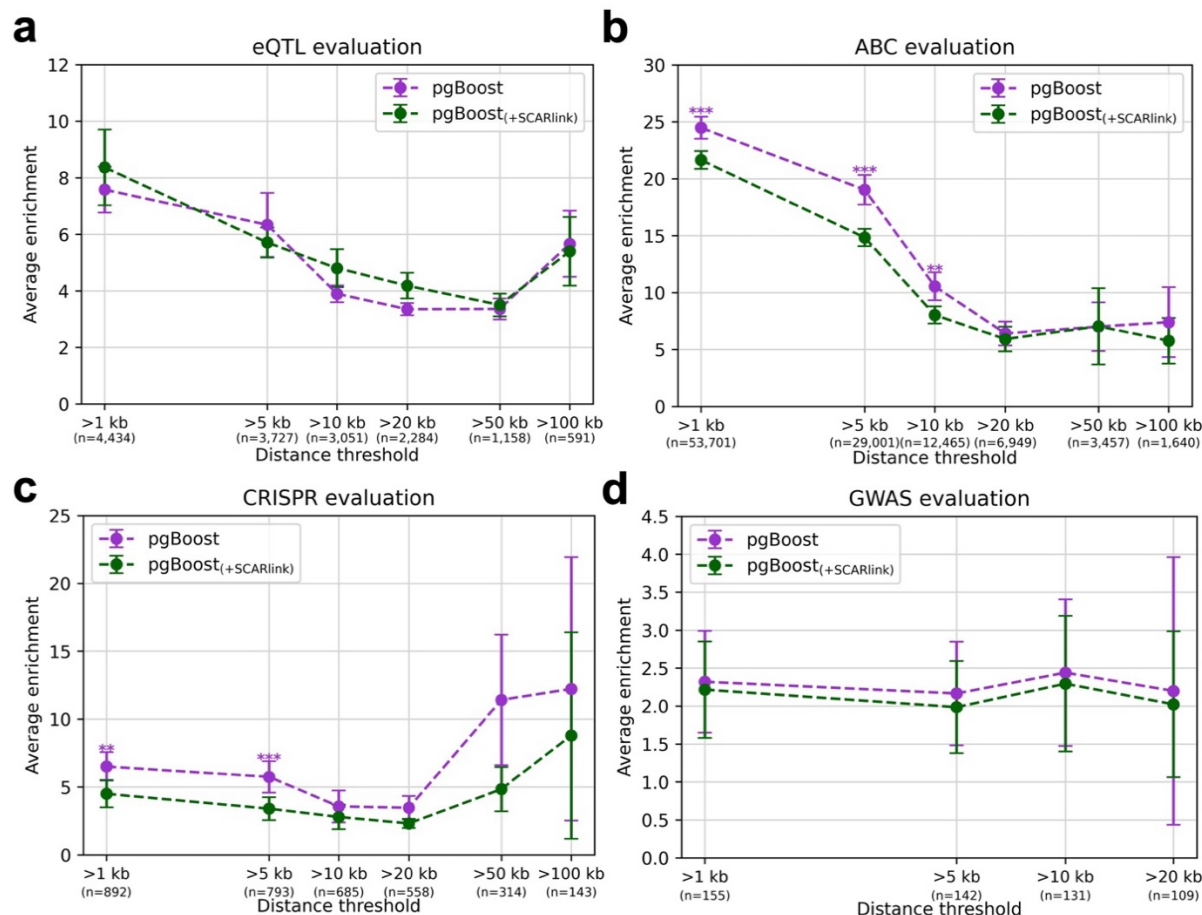
Supplementary Figure 2.25: Performance of pgBoost trained using common SNPs from 1000 Genomes vs. all SNPs from GTEx. Average enrichment across recall values of links predicted by the main pgBoost model (trained using GTEx SNPs with minor allele count ≥ 5 in 1000G, labeled pgBoost_{GTEX ∩ 1000G}) vs. pgBoost trained using all SNPs present in GTEx (labeled pgBoost_{GTEX}) for (a) 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, (b) 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, (c) 892 links validated by CRISPR, and (d) 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. For both models, we used a training set comprised of SNP-gene links attaining maximum PIP > 0.2 across GTEx tissues (as positives) or maximum PIP < 0.01 across GTEx tissues (as negatives) (Methods). We trained pgBoost_{GTEX} without restricting to SNPs with minor allele count ≥ 5 in 1000G (resulting in a training set of 12,698 positive SNP-gene pairs and 112,927 negative SNP-gene pairs).



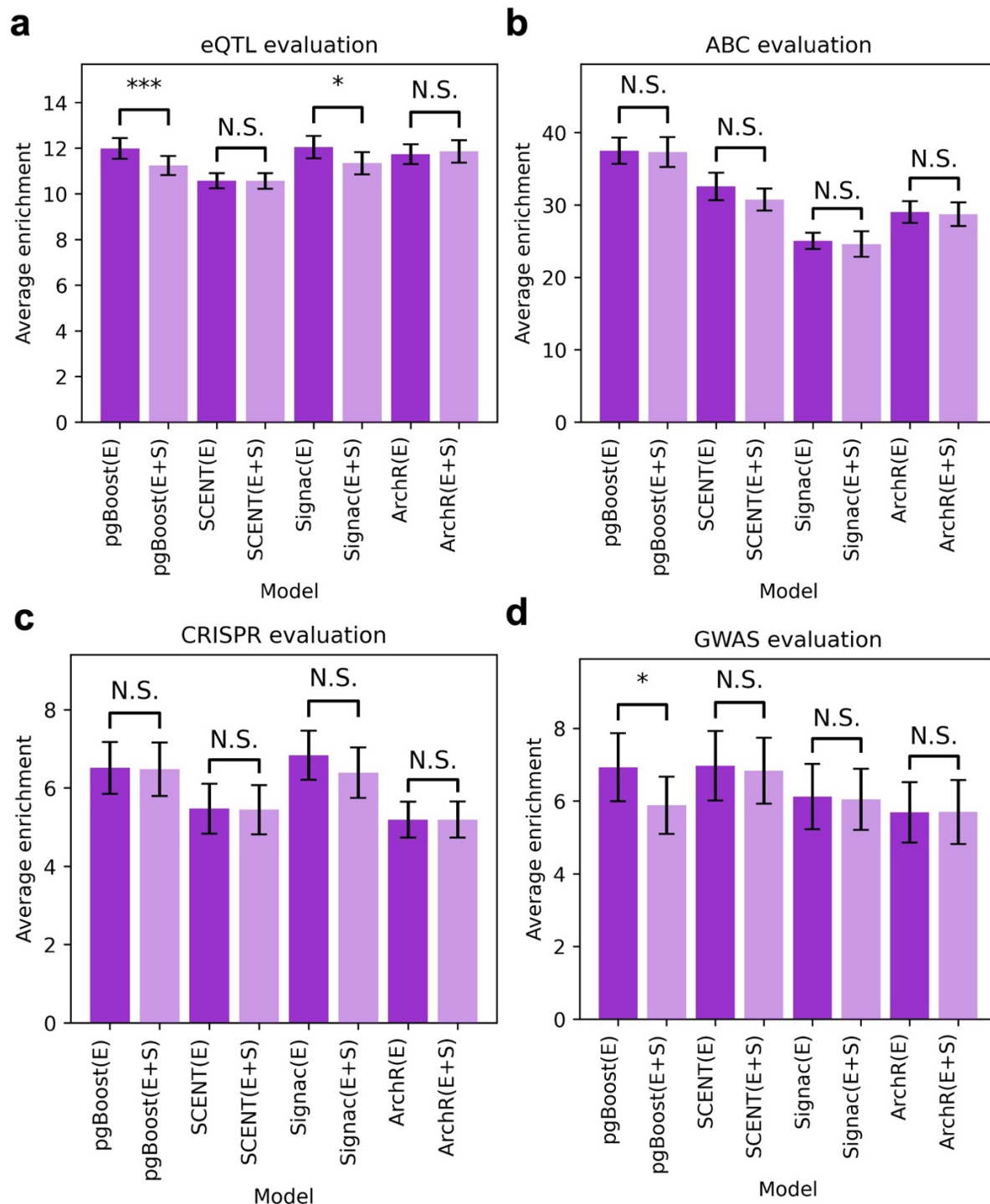
Supplementary Figure 2.26: Performance of SCARlink restricted to ATAC peaks. Average enrichment across recall values of links predicted by SCARlink including all scored regions vs. restricted to ATAC peaks for (a) 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, (b) 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, (c) 892 links validated by CRISPR, and (d) 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. By default, SCARlink⁸⁶ uses a “genomic tiling” approach, scoring 500bp tiles spanning the 250kb *cis* window of each gene. To compare with peak-gene linking methods and incorporate SCARlink into pgBoost, we first defined SCARlink_{PEAK} by restricting SCARlink predictions across all tiles (SCARlink_{TILE}) to tiles overlapping ATAC peaks. Here, we show that SCARlink_{PEAK} performed at least as well as SCARlink_{TILE} across eQTL, ABC, CRISPR, and GWAS evaluations, justifying our choice to restrict SCARlink predictions to ATAC peaks in downstream analyses. We restricted analyses to the 10x PBMC data set and to candidate links with SNP-TSS distance ≤ 250kb, because SCARlink predictions for this data set spanning 250kb *cis* windows around genes are publicly available from ref.⁸⁶ The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: p < 0.05, **: p < 0.01, ***: p < 0.001) of top method vs. each other method. (We did not measure GWAS enrichments at >50kb or >100kb, because there were too few evaluation SNP-gene links profiled in the 10x PBMC data set at these distance thresholds to perform bootstrapping across genes.)



Supplementary Figure 2.27: Performance of SCARlink vs. other existing methods. Average enrichment across recall values of links predicted by SCARlink vs. other existing peak-gene linking methods for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. We restricted analyses to the 10x PBMC data set and to candidate links with SNP-TSS distance ≤ 250 kb, because SCARlink predictions for this data set spanning 250kb *cis* windows around genes are publicly available from ref.⁸⁶ We further restricted analyses to ATAC peaks (see **Supplementary Figure 2.26**). For each method, we analyzed the top score (for SCARlink, highest Z-score) across cell types in the 10x PBMC data set. In general, SCARlink performed similarly to other existing methods across the 4 evaluation sets (and significantly underperformed Cicero and ArchR on CRISPR and GWAS evaluations, respectively). (We did not measure GWAS enrichments at >50kb or >100kb, because there were too few evaluation SNP-gene links profiled in the 10x PBMC data set at these distance thresholds to perform bootstrapping across genes.)

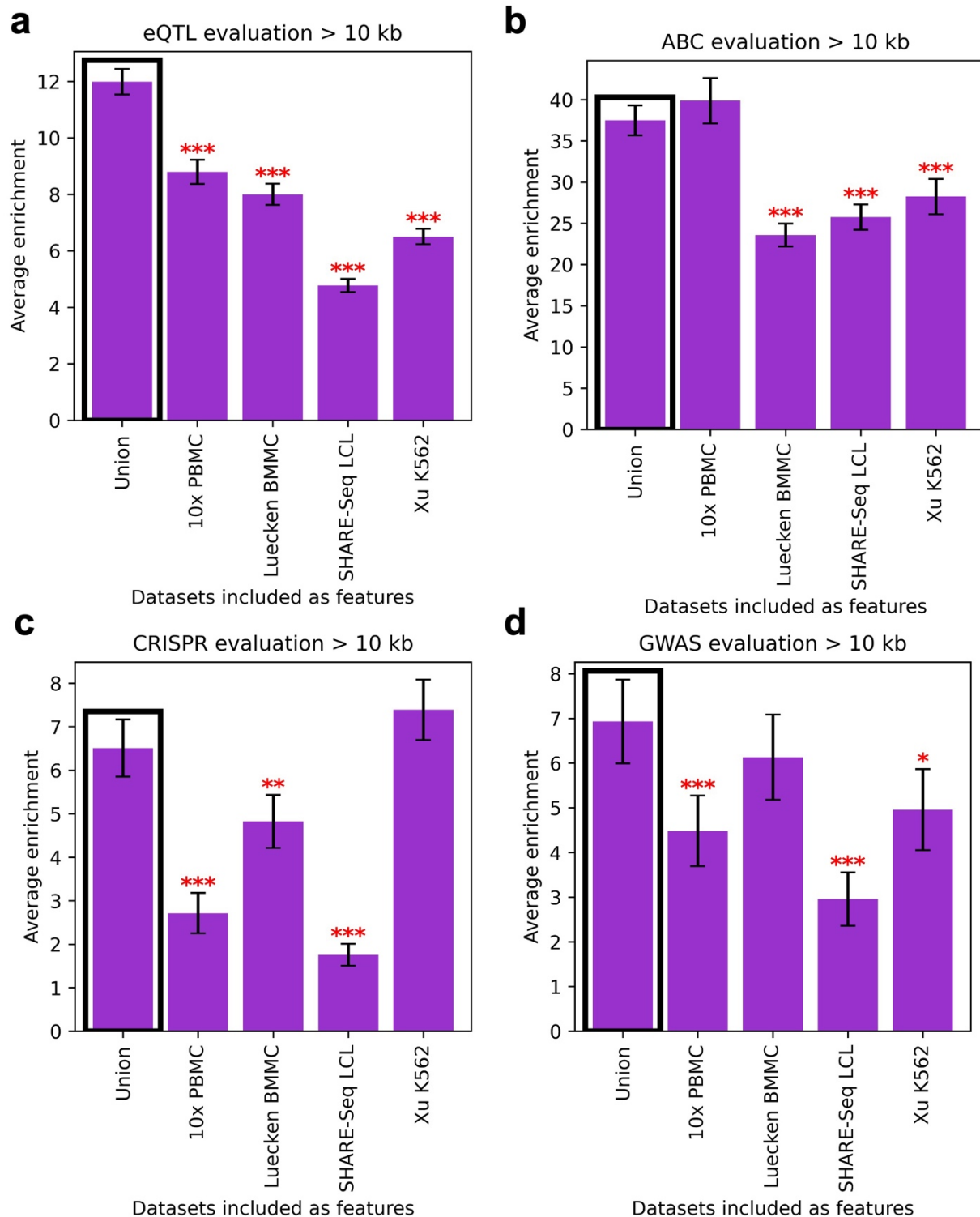


Supplementary Figure 2.28: Performance of pgBoost adding SCARlink as a constituent method. Average enrichment across recall values of links predicted by pgBoost vs. a model incorporating additional features from SCARlink for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. We restricted analyses to the 10x PBMC data set and to candidate links with SNP-TSS distance ≤ 250 kb, because SCARlink predictions for this data set spanning 250kb *cis* windows around genes are publicly available from ref.⁸⁶ We further restricted analyses to ATAC peaks (see **Supplementary Figure 2.26**). The main pgBoost model restricted to 10x PBMC data includes 11 total features: SCENT, Signac, and Cicero scores for each of the 3 cell types in the 10x PBMC data set, as well as 2 distance-based features (**Table 2**). The pgBoost model adding SCARlink as a constituent method includes 14 total features: SCENT, Signac, Cicero, and SCARlink scores (Z-score) for each of the 3 cell types in the 10x PBMC data set, as well as 2 distance-based features. The main pgBoost model significantly outperformed the model adding SCARlink as a constituent method on ABC and CRISPR evaluation data sets, and differences between these models on eQTL and GWAS evaluations were not significant, supporting our choice to exclude SCARlink from the set of pgBoost features. (We did not measure GWAS enrichments at >50kb or >100kb, because there were too few evaluation SNP-gene links profiled in the 10x PBMC data set at these distance thresholds to perform bootstrapping across genes.)



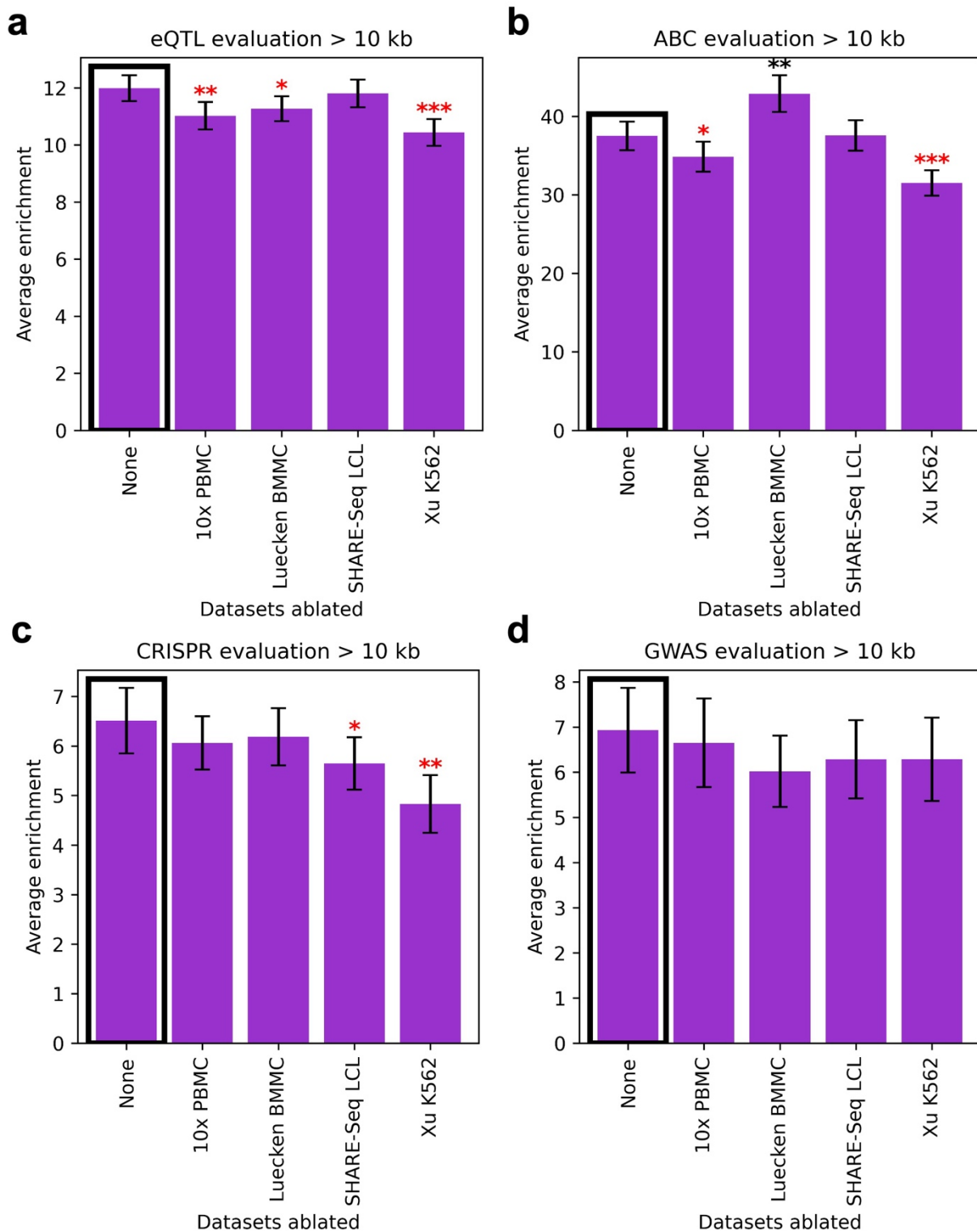
Supplementary Figure 2.29: Performance of pgBoost models incorporating significance features. Average enrichment across recall values of links predicted by pgBoost (or models restricted to linking scores from a single existing method plus genomic distance) using effect size features (E) or effect size and significance features (E+S) for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-

mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for the difference between each pair of models (denoted with black brackets) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Models incorporating only effect size metrics (SCENT regression coefficient, Signac correlation, ArchR correlation, and/or Cicero co-accessibility) plus distance as features are denoted by dark purple bars. Models incorporating effect size metrics, significance metrics (SCENT p-value, Signac p-value, and/or ArchR FDR), and distance features are denoted by light purple bars and labeled (E+S). The main pgBoost model significantly outperformed a model adding significance features on eQTL and GWAS evaluations, and differences between (E) and (E+S) models were otherwise generally not significant across the 4 evaluation sets, suggesting that significance metrics do not provide additional useful information for predicting SNP-gene links when effect size metrics are already included as features. These results are consistent with ref.²²⁹, which suggested that Pearson correlation between peak accessibility and gene expression may be a better predictor of regulatory links than a permutation-based p-value.



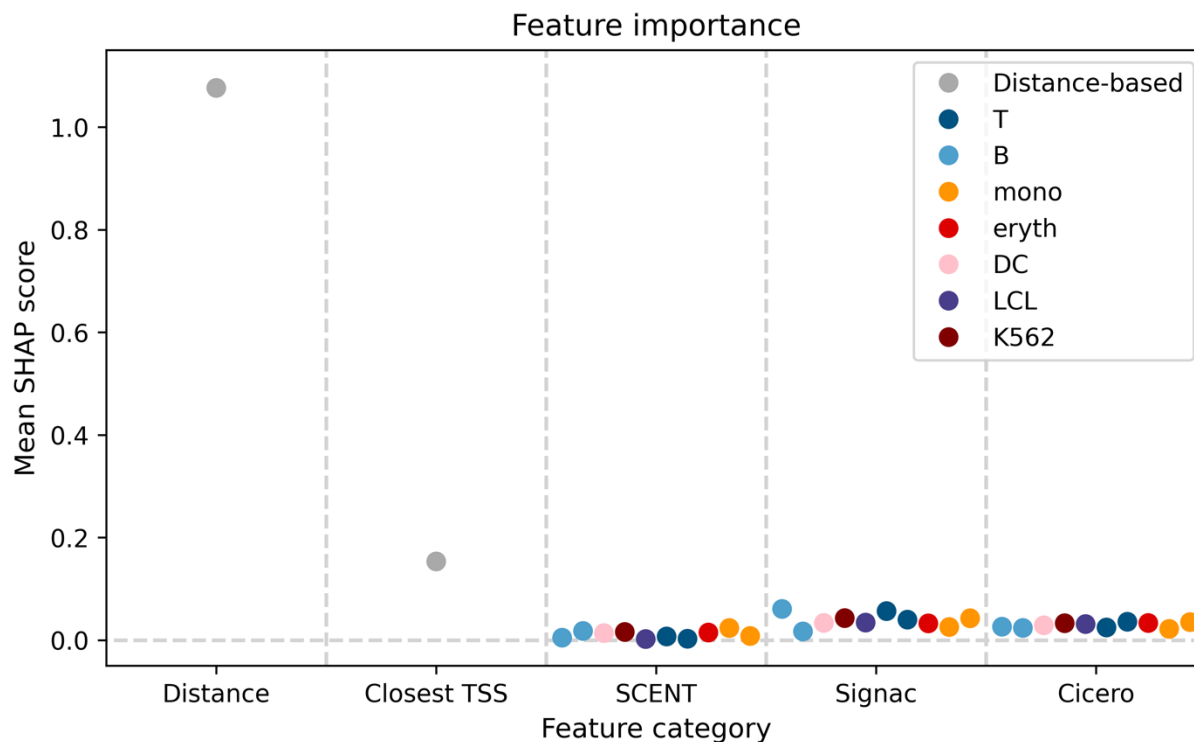
Supplementary Figure 2.30: Performance of pgBoost models restricted to a single data set. Average enrichment across recall values of links predicted by pgBoost vs. pgBoost models restricted to features from a single multiome data set for (a) 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, (b) 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, (c) 892 links validated by CRISPR, and (d)

155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of focal method vs. top method (denoted in black outline). Red stars indicate that the focal method significantly underperforms pgBoost.

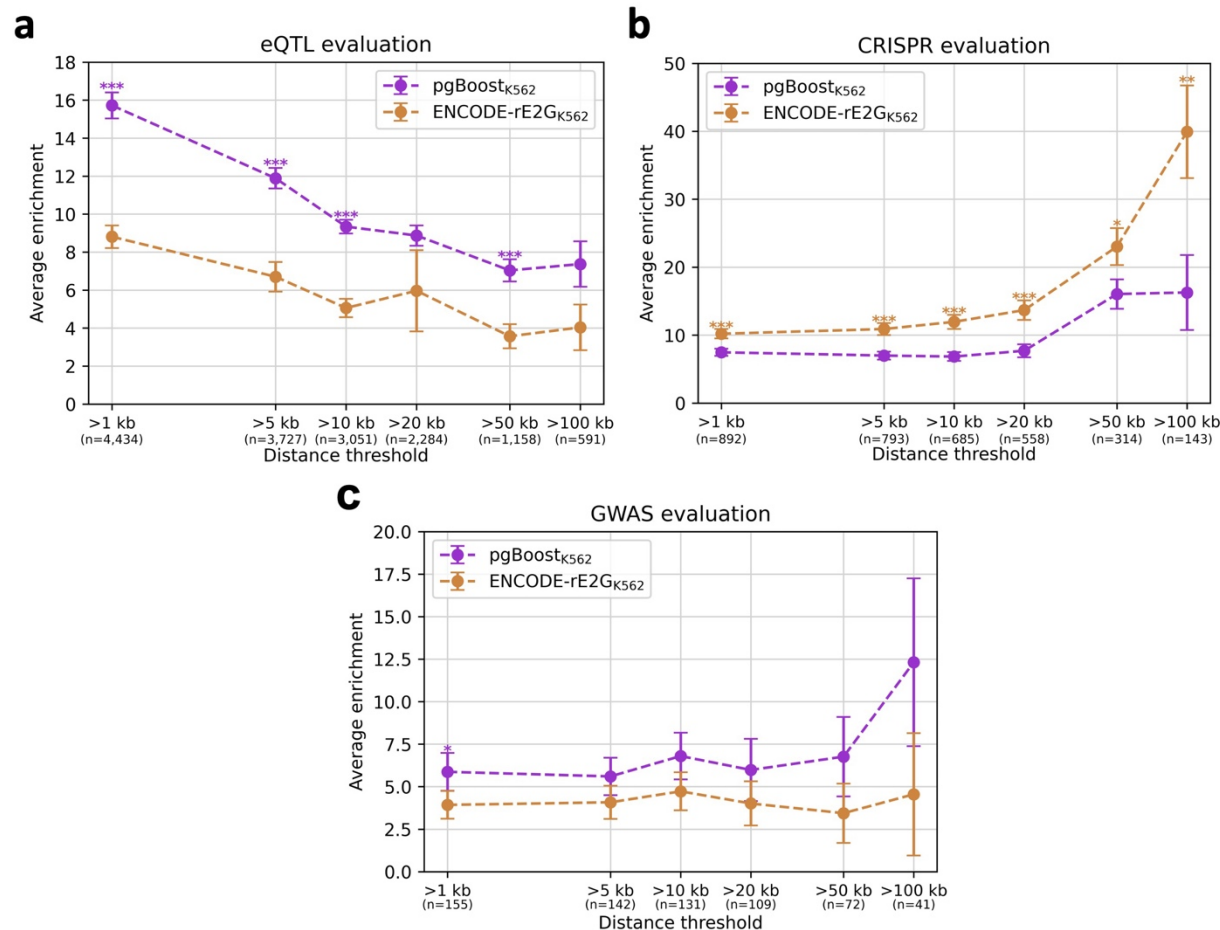


Supplementary Figure 2.31: Performance of pgBoost models ablating a single data set. Average enrichment across recall values of links predicted by pgBoost vs. pgBoost models ablating features from a single multiome data set for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum

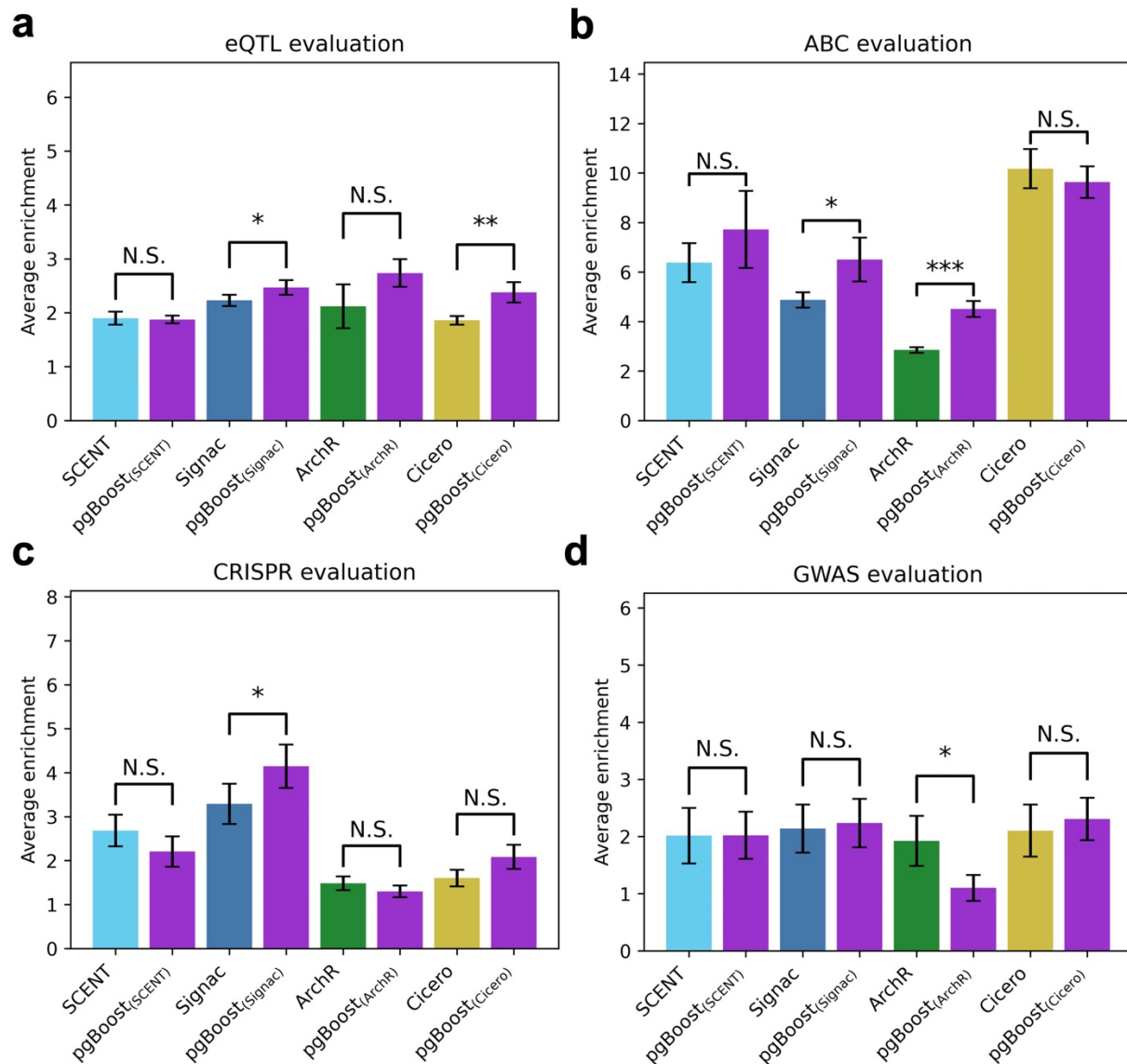
ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at $>10\text{kb}$ are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of focal method vs. top method (denoted in black outline). Red stars indicate that the focal method significantly underperforms pgBoost. While the main pgBoost model significantly underperformed the model ablating the Luecken BMMC data set on ABC evaluation data, we prioritized results on eQTL and CRISPR evaluations (see **Methods**).



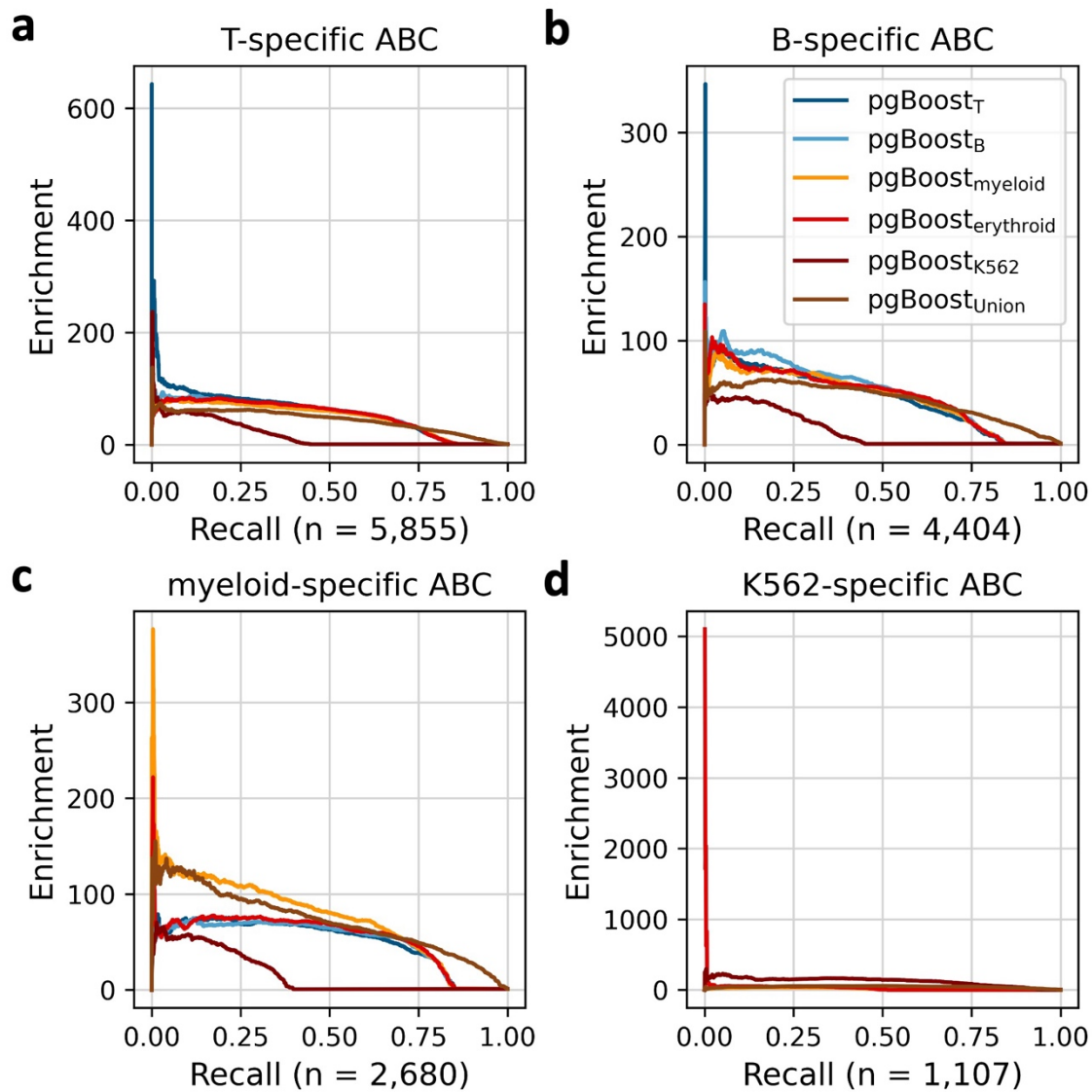
Supplementary Figure 2.32: SHAP analysis of pgBoost feature importance. Mean SHAP scores of pgBoost features, categorized by feature category (denoted by x-axis position) and color-coded by cell type (if applicable). Vertical grey dashed lines separate feature categories. Distance attains the highest mean SHAP score followed by the binary indicator for “closest TSS,” indicating that these distance-based features contribute most to linking predictions. Signac-derived features attain slightly higher mean SHAP scores than SCENT and Cicero (possibly because Signac scores more candidate links than SCENT and Cicero). We note that SHAP score assessment of the relative importance of pgBoost features may be misleading due to correlation among features (**Figure 1a**).



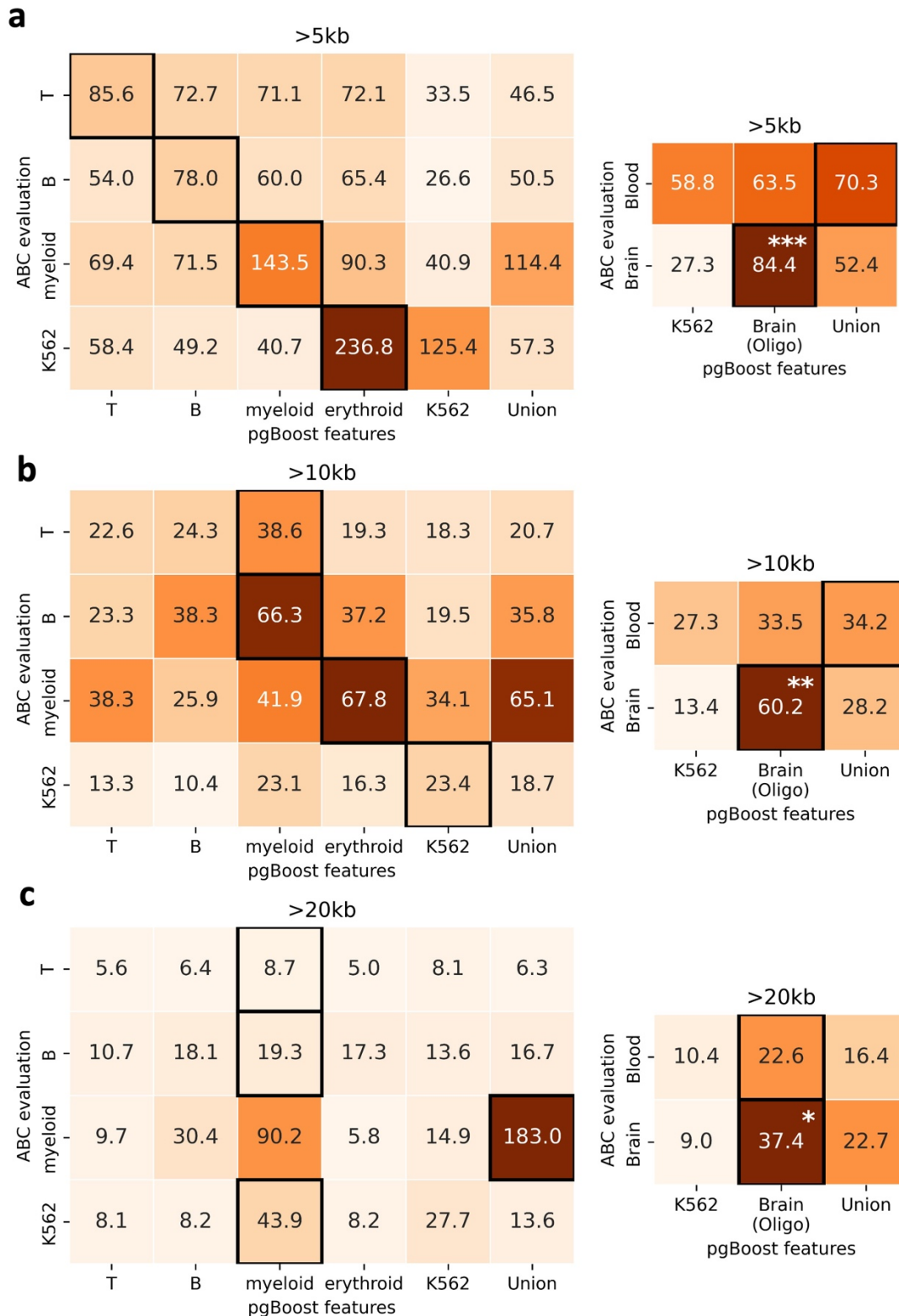
Supplementary Figure 2.33: Performance of pgBoost vs. ENCODE-rE2G in K562. Average enrichment across recall values of links predicted by pgBoost (restricted to linking scores from the Xu K562 data set plus distance features) vs. ENCODE-rE2G (score assigned to K562 cells) for (a) 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, (b) 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, (c) 892 links validated by CRISPR, and (d) 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window, at various distance thresholds (we exclude ABC evaluation data from these comparisons, because the ABC evaluation data were used as features to generate ENCODE-rE2G predictions). In particular, we compared pgBoost to predictions made by the ENCODE-rE2G^{Extended} model, which incorporates features derived from DNase-seq, histone ChIP-seq, and HiC data²⁶ for K562. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method.



Supplementary Figure 2.34: Performance of each existing method vs. pgBoost models restricted to features from each existing method. Average enrichment across recall values of links predicted by a single existing method (taking the best score across data sets and cell types) vs. pgBoost restricted to linking scores from the focal method as features for **(a)** 4,434 fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across GTEx tissues, **(b)** 53,701 SNP-gene pairs attaining maximum ABC score > 0.2 across 344 biosamples, **(c)** 892 links validated by CRISPR, and **(d)** 155 non-coding SNP-gene pairs derived from fine-mapped GWAS variants with a unique fine-mapped coding variant within a 2 Mb window. Because enrichment differences are highly similar across distance thresholds, results at >10kb are shown as barplots to aid in visual comparison. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, N.S.: non-significant) for each comparison.

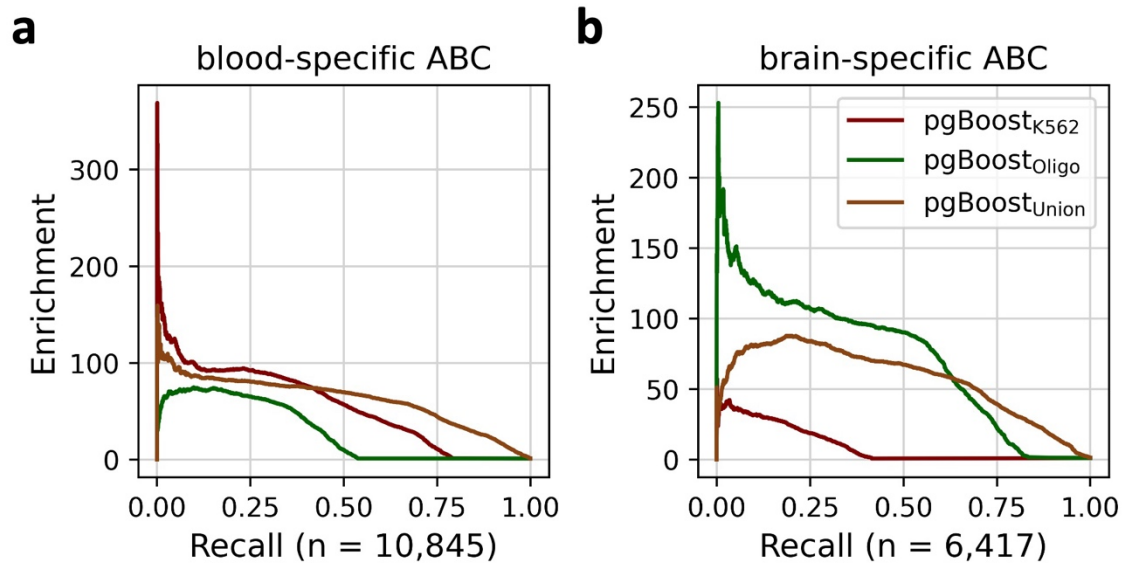


Supplementary Figure 2.35: Enrichment-recall curves of pgBoost restricted to features from a focal cell type in cell-type-specific ABC evaluation data sets. Enrichment-recall curves of links predicted by pgBoost restricted to features from a focal cell type (Union denotes all features) for cell-type-specific ABC evaluation links detected in biosamples related to **(a)** T cells, **(b)** B cells, **(c)** myeloid cells, and **(d)** K562 cells. The number of positive links in each evaluation set is specified in parentheses.

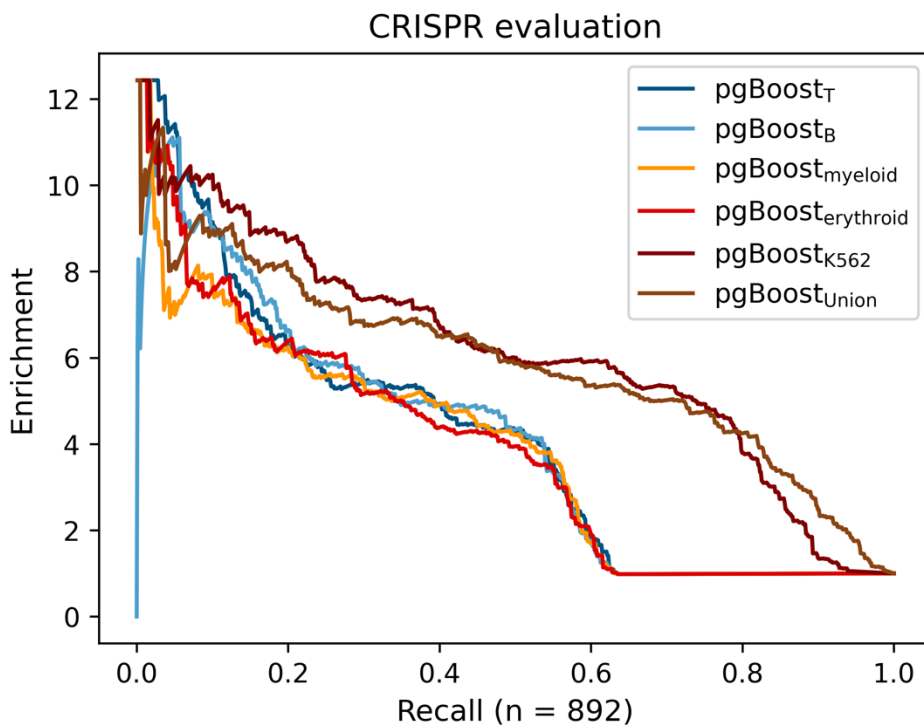


Supplementary Figure 2.36: Performance of pgBoost restricted to features from a focal cell type in cell-type-specific ABC evaluation data sets. Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (left) or tissue (right) for training and prediction (columns; Union denotes all features) for cell type/tissue-specific ABC

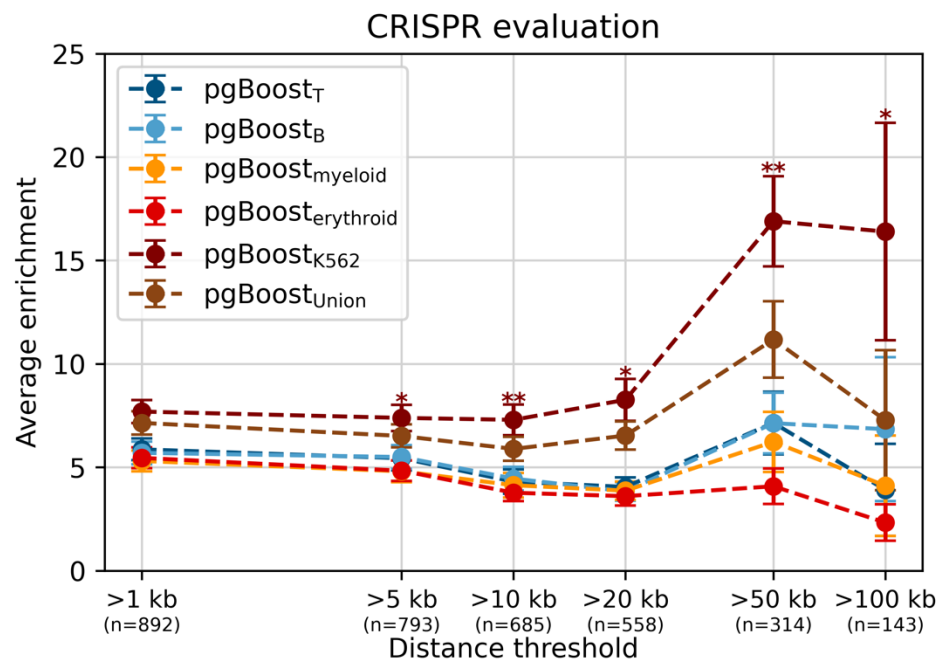
evaluation links (rows) at **(a)** >5kb, **(b)** >10kb, and **(c)** >20kb. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method (defined separately for each row and denoted in black outline) vs. each other method. We were not well-powered to detect differences in the performance of pgBoost models at higher distance thresholds because evaluation sets were small (e.g. number of evaluation links at >20kb: $n_T = 662$, $n_B = 542$, $n_{\text{myeloid}} = 134$, $n_{K562} = 59$). (We did not measure enrichments at >50kb or >100kb, because there were too few evaluation SNP-gene links at these distance thresholds to perform bootstrapping across genes.)



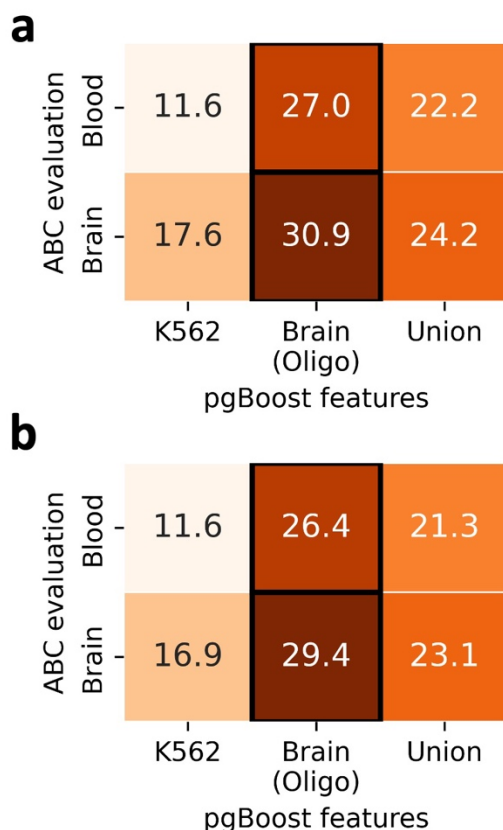
Supplementary Figure 2.37: Enrichment-recall curves of pgBoost restricted to K562 and/or brain oligodendroglia features in tissue-specific ABC evaluation data sets. Enrichment-recall curves of links predicted by pgBoost restricted to features from a focal cell type (Union denotes all features) for tissue-specific ABC evaluation links detected in biosamples related to **(a)** blood and **(b)** brain. The number of positive links in each evaluation set is specified in parentheses.



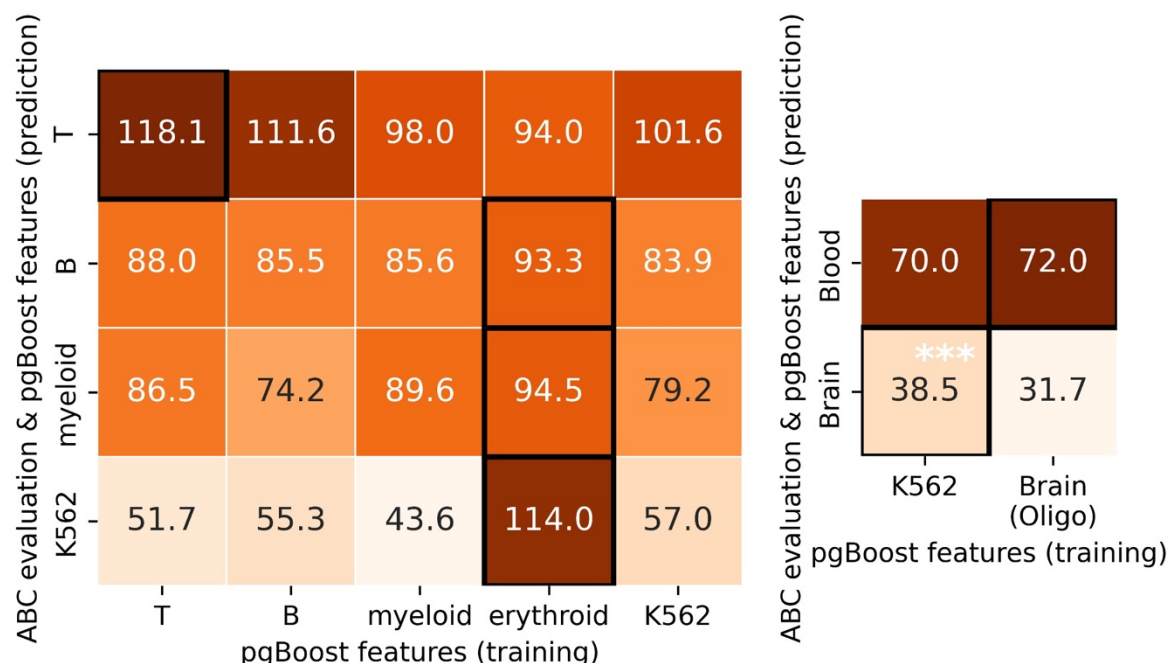
Supplementary Figure 2.38: Enrichment-recall curves of pgBoost restricted to features from a focal cell type on CRISPR evaluation data set. Enrichment-recall curves of links predicted by pgBoost restricted to features from a focal cell type (Union denotes all features) for links validated by CRISPR. The number of positive evaluation links is specified in parentheses.



Supplementary Figure 2.39: Performance of pgBoost restricted to features from a focal cell type on CRISPR evaluation data set. Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (Union denotes all features) for 892 links validated by CRISPR, at various distance thresholds. The number of positive evaluation links at each distance threshold is specified in parentheses. Confidence intervals denote standard errors. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method.

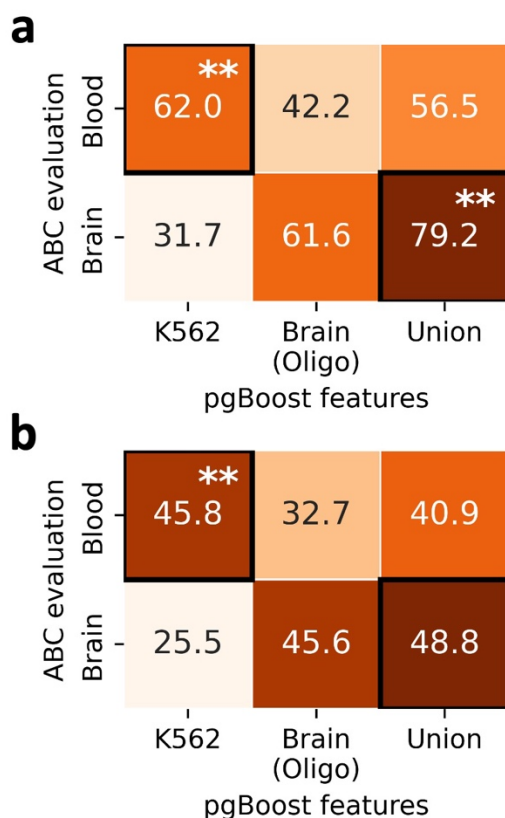


Supplementary Figure 2.40: Performance of pgBoost restricted to K562 and/or brain oligodendroglia features on tissue-specific and tissue-level eQTL evaluations. Average enrichment across recall values of links predicted by pgBoost restricted to features from K562 or brain oligodendroglia (columns; Union denotes all features) for **(a)** fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across brain-related tissues (resp. whole blood) and maximum PIP < 0.1 in whole blood (resp. brain-related tissues) and **(b)** fine-mapped eSNP-eGene pairs attaining maximum PIP > 0.5 across brain-related tissues or whole blood (irrespective of PIPs attained in the other tissue). Number of evaluation links in tissue-specific evaluation sets in **(a)**: $n_{\text{blood}} = 344$, $n_{\text{brain}} = 427$. Number of evaluation links in tissue-level evaluation sets in **(b)**: $n_{\text{blood}} = 392$, $n_{\text{brain}} = 460$. The GTEx brain evaluation set spans the following tissues: amygdala, anterior cingulate cortex, caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens basal ganglia, putamen basal ganglia, and substantia nigra. For all 4 eQTL evaluation data sets, the pgBoost model spanning brain oligodendroglia features did not significantly outperform the model spanning both K562 and brain oligodendroglia features, and both of these models significantly outperformed the model restricted to K562 features; results of tissue-specific and tissue-level analyses were similar because these evaluation sets were similar. The underperformance of the model restricted to K562 features on eQTL evaluation data sets from whole blood may reflect more pronounced differences between K562 (a cultured cell line) and *in vivo* human blood regulatory circuitry at the level of downstream regulatory effects (e.g. eQTL) vs. upstream molecular interactions (e.g. ABC).

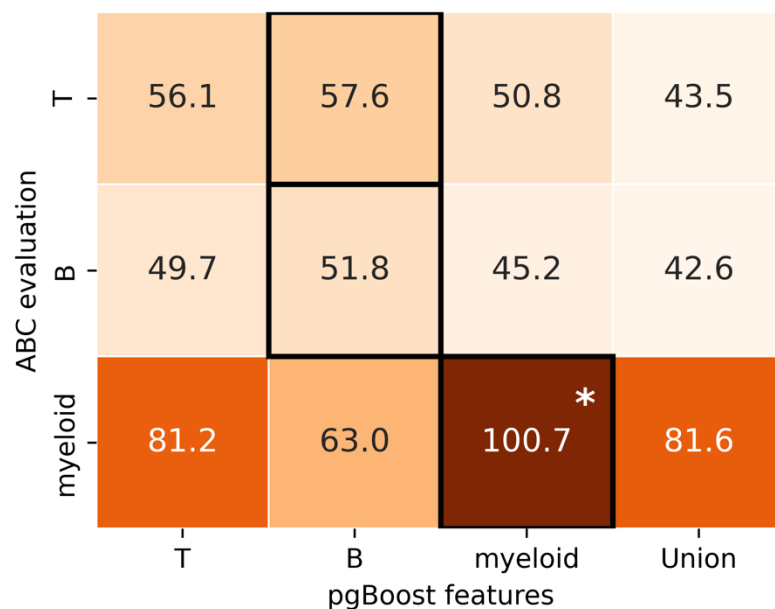


Supplementary Figure 2.41: Performance of pgBoost restricted to features from a focal or non-focal cell type for training only in evaluation data sets relevant to the focal cell type.

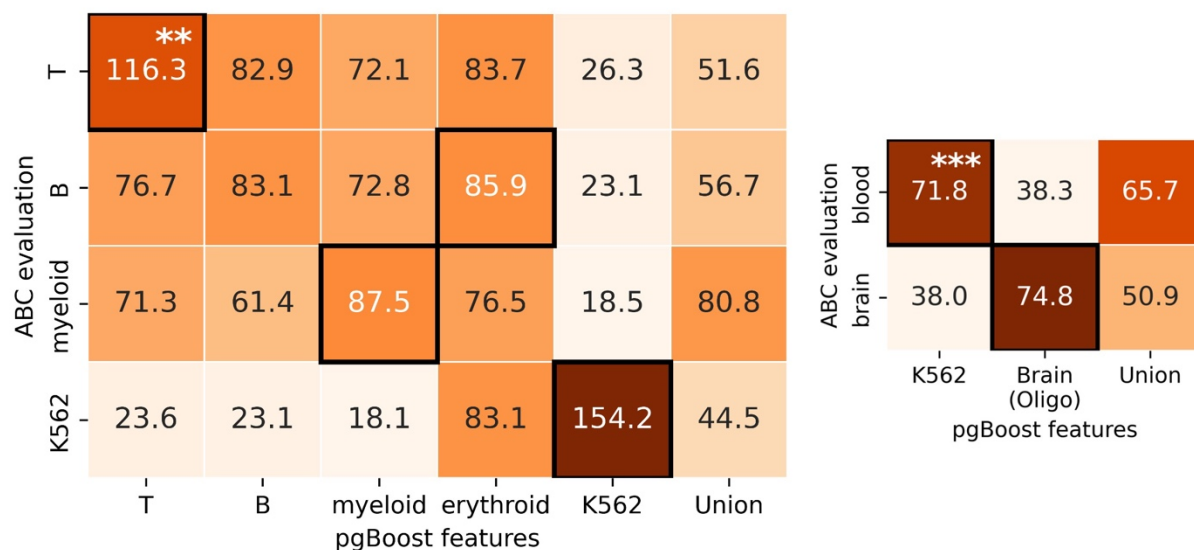
Average enrichment across recall values of links predicted by a pgBoost model restricted to features from an individual cell type (left) or tissue (right) during training (columns) for cell-type/tissue-specific ABC evaluation links (rows), where predictions are always generated by applying models to linking scores from the evaluation cell type. Number of evaluation links in each ABC evaluation set: $n_{\text{blood}} = 10,845$, $n_{\text{brain}} = 6,417$, $n_{\text{T}} = 5,855$, $n_{\text{B}} = 4,404$, $n_{\text{myeloid}} = 2,680$, $n_{\text{K562}} = 1,107$. Enrichment differences between models in each row of **(a)** were generally not statistically significant, with a few exceptions. First, the pgBoost model restricted to K562 features during training significantly underperformed each other model on T-specific, B-specific, and myeloid-specific ABC evaluations. Second, the model restricted to erythroid features did not significantly underperform the model restricted to K562 features during training on K562-specific evaluation data, and both models significantly outperformed each other model. In **(b)**, the difference between pgBoost models restricted to K562 and brain oligodendroglia features during training on blood-specific ABC evaluation data was not statistically significant, while the model restricted to K562 features during training significantly outperformed the model restricted to brain oligodendroglia features on brain-specific ABC evaluation data. These results suggest that, while there is sometimes a significant improvement in performance from training in the focal cell type (e.g. on K562 evaluation data), pgBoost models trained using linking scores from individual cell types are generally portable across cell types.



Supplementary Figure 2.42: Performance of pgBoost restricted to K562 and/or brain oligodendroglia features and trained on eQTL from blood and/or brain tissues on tissue-specific ABC evaluation data sets. Average enrichment across recall values of links predicted by pgBoost trained on GTEx eSNP-eGene pairs from **(a)** whole blood or **(b)** brain tissues, restricted to features from a focal cell type (columns; Union denotes all features), for tissue-specific ABC evaluation links (rows). Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method (defined separately for each row and denoted in black outline) vs. each other method. In both tissue-specific evaluations, the pgBoost models restricted to the related cell type (K562 for blood, oligodendroglia for brain) significantly outperformed a model restricted to the unrelated cell type. For the blood eQTL training set in **(a)**, we defined 1,518 eSNP-eGene pairs attaining $PIP > 0.2$ in whole blood as positives and 6,997 eSNP-eGene pairs including a gene in the positive set and attaining $PIP < 0.01$ in whole blood as negatives. For the brain eQTL training set in **(b)**, we defined 1,334 eSNP-eGene pairs attaining $PIP > 0.2$ across brain tissues as positives and 12,399 eSNP-eGene pairs including a gene in the positive set and attaining maximum $PIP < 0.01$ across brain tissues as negatives. The pgBoost model restricted to K562 (resp. brain oligodendroglia) features significantly outperformed the model restricted to brain oligodendroglia (resp. K562) on blood-specific (resp. brain-specific) evaluation data in both **(a)** and **(b)**. The pgBoost model spanning both cell types significantly underperformed the model restricted to K562 on blood-specific ABC evaluation data in both **(a)** and **(b)**, but outperformed (in **a**) or performed similarly to (in **b**) the model restricted to brain oligodendroglia features in brain-specific ABC evaluation data.

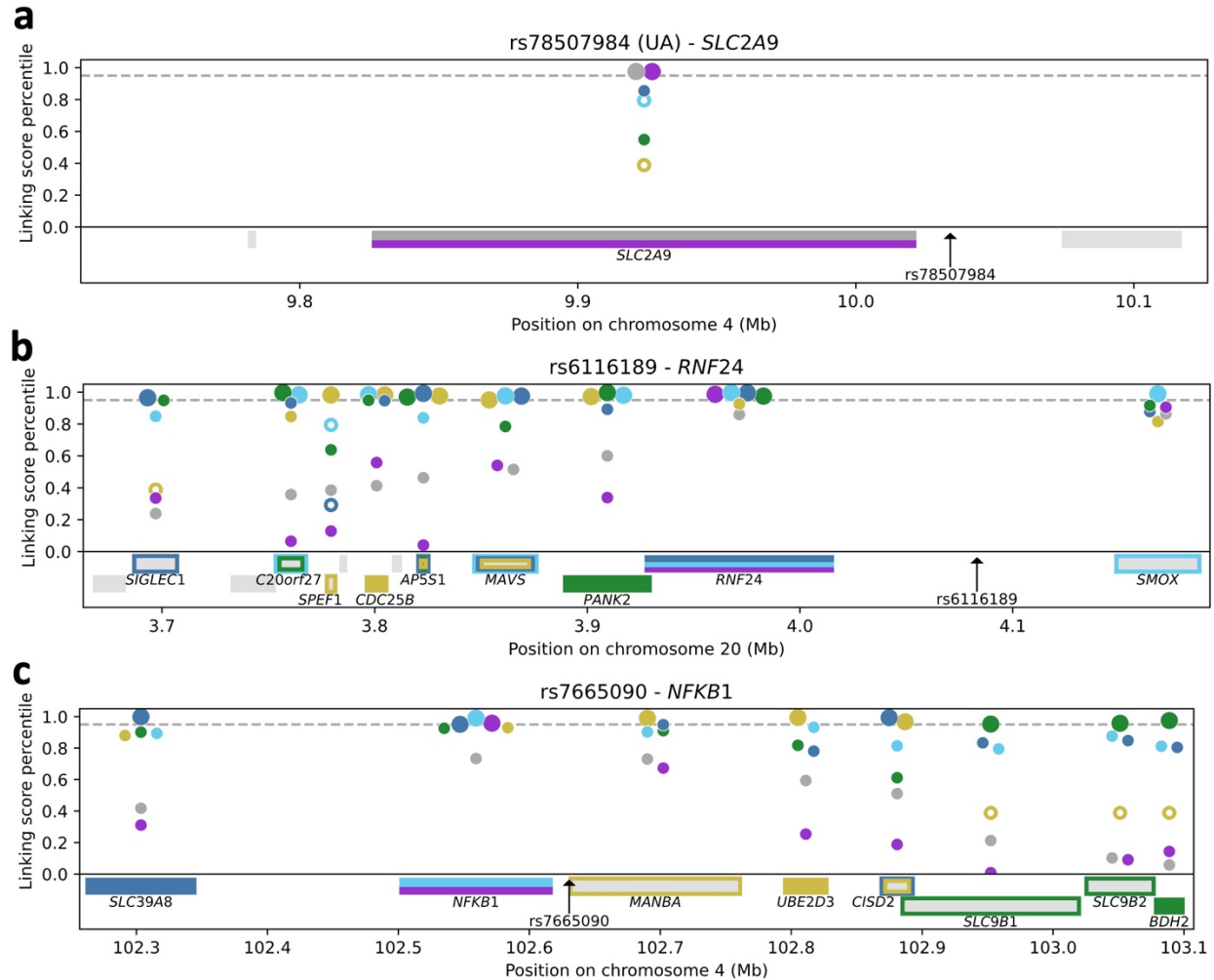


Supplementary Figure 2.43: Performance of pgBoost restricted to features from a focal cell type across two single-cell data sets in cell-type-specific ABC evaluation data sets. Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (columns; Union denotes all features) included in the Luecken BMMC and 10X PBMC data sets (**Table 2**) for cell-type-specific ABC evaluation links (rows). Top method is defined separately for each row and denoted in black outline. Stars denote 2-sided bootstrap p-values for difference (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$) of top method vs. each other method. The pgBoost model restricted to myeloid features significantly outperformed all other models on myeloid-specific ABC evaluation data. Models restricted to T cells or B cells attained statistically similar enrichments and both significantly outperformed the model restricted to myeloid features and the model spanning all 3 cell types in T cell-specific and B cell-specific ABC evaluation data.



Supplementary Figure 2.44: Performance of pgBoost restricted to features from a focal cell type in ABC evaluation links passing a stringent threshold for cell-type/tissue-specificity.

Average enrichment across recall values of links predicted by pgBoost restricted to features from a focal cell type (left) or tissue (right) (columns; Union denotes all features) for cell-type/tissue-specific ABC evaluation links (rows) defined by SNP-gene links attaining ABC score > 0.2 in biosample(s) related to the focal cell type/tissue and below a more stringent threshold in non-focal cell types/tissues (ABC score < 0.05). Number of evaluation links in each ABC evaluation set: $n_T = 2,223$, $n_B = 1,747$, $n_{\text{mono}} = 1,204$, $n_{K562} = 708$. Top method is defined separately for each row and denoted in black outline. For T cell-specific evaluation data, the pgBoost model restricted to T cell features significantly outperformed each other model. For B cell-specific evaluation data, the pgBoost model restricted to erythroid features did not significantly outperform the models restricted to B cell features or myeloid features, but significantly outperformed the model restricted to T cell features, the model restricted to K562 features, and the model spanning the union of all 5 cell types. For myeloid-specific evaluation data, the pgBoost model restricted to myeloid features significantly outperformed models restricted to T cell, B cell, or K562 features, but did not significantly outperform the model restricted to erythroid features or the model spanning the union of all 5 cell types. For K562-specific evaluation data, the pgBoost model restricted to K562 features did not significantly outperform the model restricted to erythroid features, but significantly outperformed each other model. For brain-specific evaluation data, the pgBoost model restricted to brain oligodendroglia features significantly outperformed the model restricted to K562 features, but did not significantly outperform the model spanning both cell types. For blood-specific evaluation data, the pgBoost model restricted to K562 features significantly outperformed both other models.



Supplementary Figure 2.45: Additional examples of variant-gene links predicted by pgBoost. We report links prioritized by pgBoost, distance, and 4 existing methods for (a) rs78507984, (b) rs6116189, and (c) rs7665090. Panel headings for each SNP report the corresponding GWAS trait(s) (in parentheses) and the SNP-gene link attaining the highest pgBoost score. Linking scores for each method are reported as dots in the middle of each gene (jittered for easier visualization). Large dots denote links scored in the top 5% by a given method. Hollow dots denote links that were not scored by a given method (with their percentile defined as the percentage of links not scored by the method). Genes scored in the top 5% by any method are reported at the bottom of each panel (with gene name); nearby genes are also displayed (without gene name). Colored boxes denote the gene attaining the highest linking score percentile > 95% by one or more method(s); colored borders denote other genes scored in the top 5% by a given method. UA: uric acid.

Chapter 3 Appendix

Methods

Defining peak co-accessibility and peak co-activity

Computing co-accessibility scores. We defined the *co-accessibility score* of peak j as $x_j = \sum_{k \in K_j} (r_{jk}^2 - b_x)$, where K_j is the set of all *cis* peaks centered <1Mb from the center of peak j (including peak j itself) (average number of *cis* peaks per focal peak = 124-147 across data set-cell type pairs), r_{jk}^2 are squared peak-peak correlations (from scATAC-seq data) across metacells, and b_x is a term that captures expected bias in measured peak-peak r^2 due to finite sample size (see below). We defined *genome-wide co-accessibility score* analogously, letting K_j be the set of all peaks genome-wide. Co-accessibility scores can be viewed as an analogue of SNP LD scores¹²⁷. We computed peak-peak correlations across metacells using the ArchR⁸² function `addCoAccessibility` (with parameter `maxDist` = 1e+06) followed by `getCoAccessibility` (with parameter `corCutOff` = -1.5, ensuring that all peak-peak correlations $\in [-1,1]$ are returned). We excluded peaks overlapping any promoter region (defined as <1kb to any gene's TSS^{90,128}) from all analyses.

Bias correction. We computed the expected bias b_x (see above) for a given data set-cell type pair as follows: we first randomly selected 100,000 peak-peak pairs and computed a_x , the average peak-peak r^2 across the selected peak-peak pairs. We then generated a randomly downsampled ATAC matrix (containing 50% of metacells) and computed $a_{x, \text{downsampled}}$, the average peak-peak r^2 (measured in the 50% downsampled metacells) across the selected peak-peak pairs. Finally, we computed $b_x = a_{x, \text{downsampled}} - a_x$. (Because squared correlations measured across 50% of metacells should contain 2x the amount of upward bias due to noise, this value of b_x provides an estimate of the amount of upward bias due to noise in all metacells.)

Computing co-activity scores. We defined the *co-activity score* of peak j as $y_j = \sum_{g \in G_j} (r_{jg}^2 - b_y)$, where G_j is the set of all *cis* genes with TSS <1Mb from the center of the focal

peak j (average number of *cis* genes per focal peak = 8-12 across data set-cell type pairs), r_{jg}^2 are squared peak-gene correlations (from scATAC-seq and scRNA-seq data) across metacells, and b_y is a term that captures expected bias in measured peak-gene r^2 due to finite sample size (analogous to bias correction procedure described above). We also defined *genome-wide co-activity score* as above, where G_j is the set of all genes genome-wide. We computed peak-gene correlations across metacells using the ArchR⁸² function `addPeak2GeneLinks` (with parameter `maxDist = 1e+06`) followed by `getPeak2GeneLinks` (with parameters `corCutOff = -1.5` and `FDRCutOff = 1`, ensuring that all peak-gene correlations $\in [-1,1]$ are returned). We excluded peaks overlapping any promoter region (defined as <1kb to any gene's TSS^{90,128}) from all analyses.

Stratified co-accessibility score regression

S-CASC model. We propose stratified co-accessibility score regression (S-CASC) as a method to identify functional peak categories that are enriched for causal peak-gene associations and assess the tagging hypothesis. The S-CASC regression relies on the following model:

$$E[y_j] = \tau_0 + \sum_C (\tau_C x_{j,C}) + \gamma |G_j|, \quad [3.2]$$

where y is the co-activity score of peak j , τ_0 is an intercept term, $x_{j,C}$ is the co-accessibility score of peak j with respect to peak category C , τ_C is the per-peak contribution to causal peak-gene effect variance for peak category C , and the covariate $|G_j|$ is the number of genes with TSS <1Mb from peak j . $x_{j,C}$ is defined as $\sum_{k \in K_j} (I_{Ck} r_{jk}^2 - b_x)$, where K_j is the set of all *cis* peaks centered <1Mb from the center of peak j (including peak j itself) and not overlapping any promoter region, I_{Ck} is the value of peak category C for peak k (for binary peak categories, I_{Ck} equals 1 if peak k belongs to category C and 0 otherwise; for probabilistic peak categories, $I_{Ck} \in [0,1]$; see below),

and other terms are defined as above. S-CASC can be viewed as an analogue of stratified LD score regression¹⁸.

Estimating per-peak causal effect size variance. We defined per-peak causal effect size variance for peak j as $\text{Var}(\beta_j) = \sum_C I_{Cj} \tau_C$. After computing co-activity scores (y_j) and stratified co-accessibility scores ($x_{j,C}$), we performed the multiple regression in equation 3.2 (regressing y_j on $x_{j,C}$) to obtain estimates $\hat{\tau}_C$ of τ_C for each peak category C . We then estimated per-peak causal effect size variances $\widehat{\text{Var}}(\beta_j)$ using $\hat{\tau}_C$.

Estimating causal effect size enrichment of a peak category. We defined the total causal effect size variance in peak category C as $v_C = \sum_k I_{Ck} \text{Var}(\beta_k)$. We then defined the causal effect size enrichment of peak category C as follows:

$$E_C = \frac{v_C}{v_{\text{all peaks}}} / \frac{M_C}{M_{\text{all peaks}}}, \quad [3.3]$$

where $v_{\text{all peaks}}$ is the total causal effect size variance of all peaks, M_C is defined as $\sum_j I_{Cj}$ (for a binary peak category, M_C is the number of peaks in category C), and $M_{\text{all peaks}}$ is the total number of peaks. We use our estimates of $\widehat{\text{Var}}(\beta_j)$ to estimate $\hat{v}_C$ and $\hat{v}_{\text{all peaks}}$. We then use these estimates to estimate $\hat{E}_C$ for each peak category C . (Due to peak measurement noise, we focus on $\hat{E}_C$ rather than $\hat{v}_C$ to assess the relative importance of each peak category.)

Computing jackknife standard errors. For a given data set-cell type pair, we estimated standard errors for each $\hat{E}_C$ using a jackknife with blocks defined by chromosomes, analogous to previous work^{18,127}. (We first obtained jackknifed estimates of $\hat{\tau}_C$, then used these estimates to obtain jackknifed estimates of $\widehat{\text{Var}}(\beta_j)$, $\hat{v}_C$, and $\hat{v}_{\text{all peaks}}$, and $\hat{E}_C$.)

Computing meta-analyzed estimates and standard errors. We computed meta-analyzed $\hat{E}_C$ estimates and standard errors via random effect meta-analysis across data set-cell type pairs.

Defining peak categories. We implemented S-CASC using 19 functional peak categories (including 18 binary categories and 1 probabilistic peak annotation), in addition to a category defined by all peaks, as described in ref.¹⁸ For each binary category, we defined an ATAC peak belonging to that category as a peak overlapping any genomic region in that category.

The functional peak categories were defined as follows:

- 5 binary genomic categories (“Coding”, “3’ UTR”, “5’ UTR”, “Intron”, “Repressed”) from the S-LDSC¹⁸ baseline model.
- 10 binary categories defined by activating or repressive histone modifications (7 activating: H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K36me3, H3K79me2; and 3 repressive: H3K27me3, H3K9me1, H3K9me3), as measured by histone ChIP-seq experiments in K562 cells publicly on the ENCODE portal (for each experiment, we used the ‘pseudoreplicated peaks’ file in ‘bed narrowPeak’ format).
- 2 binary gene proximity-based categories (“TSS±10kb”, “TSS±100kb”) defined by any overlap between a peak and the union of genomic windows TSS±10kb (or TSS±100kb, respectively) across all genes.
- 1 binary category defined as “Closest to TSS”. For each gene, we identified the single ATAC peak closest to the focal gene’s TSS. We then defined the set of these peaks as the “Closest to TSS” category.
- 1 probabilistic constraint-based annotation (“Constraint”) defined by the proportion of bases in an ATAC peak overlapping any constrained region defined by ref.¹⁹⁵

Genome-wide S-CASC. As stated above, to compute $x_{j,C}$, we restricted to the set of all *cis* peaks centered <1Mb from peak j . Computing $x_{j,C}$ without this restriction (and instead using all peaks genome-wide) yielded $x_{j,C}$ that were highly correlated across peak categories, leading to substantial multicollinearity in the S-CASC regression and inflated standard errors for $\hat{E}_C$, such that all $\hat{E}_C$ were statistically indistinguishable from 1.

Gene co-expression and gene co-activity scores

Computing gene co-expression scores. We defined the *gene co-expression score* of gene i as $x_i = \sum_{g \in G_i} (r_{ig}^2 - b_e)$, where G_i is the set of all genes with TSS <1Mb from the TSS of focal gene i (including gene i itself) (average number of *cis* genes per focal gene = 31-44 across data set-cell type pairs), r_{ig}^2 are squared gene-gene correlations (from scRNA-seq data) across metacells, and b_e is a term that captures expected bias in measured gene-gene r^2 due to finite sample size (analogous to the bias correction procedure described above). We also defined *genome-wide gene co-expression score* as above, where G_i is the set of all genes genome-wide. We computed gene-gene correlations using the `cor()` function on ArchR⁸² gene expression metacell matrices.

Computing gene co-activity scores. We defined the *gene co-activity score* of gene i as $y_i = \sum_{k \in K_i} (r_{ik}^2 - b_y)$, where K_i is the set of all peaks centered <1Mb from the TSS of gene i (average number of *cis* peaks = 64-102 across data set-cell type pairs), r_{ik}^2 are squared gene-peak correlations (from scATAC-seq and scRNA-seq data) across metacells, and b_y is a term that captures expected bias in measured peak-gene r^2 due to finite sample size (analogous to the bias correction procedure described above; we note that the same b_y value is used to compute peak co-activity scores and gene co-activity scores). We also defined *genome-wide gene co-activity score* analogously, letting K_i be the set of all peaks genome-wide. We excluded peaks overlapping any promoter region (defined as <1kb to any gene's TSS^{90,128}) from all analyses. We computed peak-gene correlations across metacells as described above.

Stratified co-expression score regression model. We implemented stratified co-expression score regression as a method to identify functional gene categories that are enriched for causal peak-gene associations and assess the gene tagging hypothesis. This regression relies on a model analogous to the S-CASC model:

$$E[y_i] = \tau_0 + \sum_C (\tau_C x_{i,C}) + \gamma |K_i|, \quad [3.4]$$

where y is the gene co-activity score of gene i , τ_0 is an intercept term, $x_{i,C}$ is the co-expression score of gene i with respect to gene category C , τ_C is the per-peak contribution to causal peak-gene effect variance for gene category C , and the covariate $|K_i|$ is the number of peaks centered <1Mb from the TSS of gene i . $x_{i,C}$ is defined as $\sum_{g \in G_i} (I_{Cg} r_{ig}^2 - b_e)$, where G_i is the set of all *cis* genes with TSS <1Mb from the TSS of gene i (including gene i itself), I_{Cg} is the value of gene category C for gene i (for binary gene categories, 1 if peak k belongs to category C and 0 otherwise), and other terms are defined as above.

Estimating per-gene causal effect sizes and causal effect size enrichment. We defined and estimated per-gene causal effect size variances, total causal effect size variance in each gene category, and causal effect size enrichment for each gene category, as described above. We also computed jackknife standard errors and meta-analyzed enrichments across data set-cell type pairs, as described above.

Defining gene categories. We implemented S-CESC using 18 functional gene categories, in addition to a category defined by all genes. These categories spanned 16 categories described in ref.²³⁰ and 2 peak proximity-based categories; all were binary categories.

The 16 functional gene categories from ref.²³⁰ are briefly described below; we excluded the gene category “Olfactory receptors”²³⁰ from our analyses because the Luecken BMMC¹⁰⁹ data set did not include any genes in this category:

- 6 categories reflecting gene essentiality or selection against protein-altering variants (“High pLI”, “High pHI”, “High S_{het} ”, “Haploinsufficient”, “Missense Z”, “MGI essential”).
- 3 categories reflecting clinical significance or pathogenicity (“OMIM”, “ClinVar”, “Autosomal dominant”).

- 2 categories defined by proximity to a large number of SNPs (“Most SNPs within 100kb”) or independent ($r < 0.1$) SNPs (“Most independent SNPs within 100kb”).
- 2 categories defined by the presence or absence of genetic associations (“GWAS nearest”, “eQTL-deficient”).
- 1 category defined by genes encoding transcription factors (“TFs”).
- 1 category defined by high predicted connectedness to enhancers (“High Enhancer Domain Score”).
- 1 category defined by established drug targets (“FDA drug bank”).

The 2 peak proximity-based categories are briefly described below:

- 1 category (“Peak $\pm$ 10kb”) defined by the overlap of a gene’s TSS with the 10kb genomic window surrounding any peak. (We note that the “Peak $\pm$ 100kb” category included most genes and was thus uninformative.)
- 1 category defined as “Closest to peak”. For each peak, we identified the single gene with TSS closest to the focal peak. We then defined the set of these genes as the “Closest to peak” category.

Single-cell multiome data sets

We downloaded 4 publicly available scRNA+ATAC-seq multiome data sets (Xu K562¹¹⁰, Satpathy K562¹²⁹, SHARE-seq LCL⁷⁷, and Luecken BMMC¹⁰⁹; **Table 2.1**) that spanned immune and blood cell types. For the Luecken BMMC¹⁰⁹ data set, we analyzed major cell types that spanned >5,000 cells, yielding 7 data set-cell type pairs.

We obtained publicly available quality controlled gene expression and ATAC peak matrices for the Luecken BMMC data sets, and processed the Xu K562¹¹⁰, Satpathy K562¹²⁹, SHARE-seq LCL⁷⁷ data sets using both Seurat (v5.3.0)¹⁹⁸ and ArchR (v1.0.3)⁸². We retained all cells from each data set passing quality control filters on total ATAC and RNA reads, nucleosome signal, and TSS enrichment, following ref.⁹⁰ We next identified ‘peaks’ of open chromatin in each

data set using macs2¹⁹⁹ and excluded non-standard chromosomes and blacklist regions. For the Luecken BMMC data set, which spans multiple cell types, we defined cell types using labels provided by the authors¹⁰⁹. These steps yielded matrices of gene expression and ATAC peak counts in individual cells for each data set-cell type pair.

We pre-processed each data set-cell type pair using the ArchR⁸² pipeline. We first performed iterative LSI-based dimensionality reduction separately on RNA and ATAC peak matrices using `addIterativeLSI()` with default parameters. We then combined the RNA and ATAC LSI-based reductions into a joint dimension reduction using `addCombinedDims()` with default parameters. When computing peak-peak and peak-gene correlations (see above), ArchR⁸² generates metacell-level RNA and ATAC matrices from the joint dimension reduction, by defining metacells (low-level cell aggregates) using *k*-nearest neighbors and aggregating RNA and ATAC peak counts across single cells within each metacell. After computing peak-gene correlations, we extracted RNA and ATAC metacell matrices from the metadata of the output from peak-gene linking.

Transcription factor binding site data

ChIP-seq data. We obtained publicly available ChIP-seq data for 303 transcription factors (TFs) in K562 cells and 148 TFs in GM12878 cells from the ENCODE portal, using the following search terms: “TF ChIP-seq” for assay title, “*Homo sapiens*” for organism, “K562” or “GM12878” for biosample, “Not perturbed” for perturbation, and “GRCh38” for genome assembly. We further restricted to assays with biosample summary either “*Homo sapiens* K562” or “*Homo sapiens* GM12878” (removing assays with “genetically modified using CRISPR” in the biosample summary). For each experiment, we downloaded the “IDR thresholded peaks” file in “bed narrowPeak” format, which contained quality-controlled ChIP-seq peaks.

Defining TFBS for ATAC peaks. We analyzed transcription factor binding sites (TFBS) in 3 data set-cell type pairs (Xu K562^{110,129} and Satpathy K562¹²⁹ for K562 TFs, SHARE-seq LCL⁷⁷ for GM12878 TFs; **Table 2.1**). For a given TF and data set-cell type pair, we defined a TF binding

in an ATAC peak as any overlap of the ATAC peak with a ChIP-seq peak for that TF in the corresponding cell type; we thus generated a binarized TFBS matrix for ATAC peaks in each data set-cell type pair (in which an entry of 1 for a given ATAC peak-TF pair indicates that ≥ 1 ChIP-seq peaks for the focal TF overlapped the focal ATAC peak, and an entry of 0 indicates that no ChIP-seq peaks for the focal TF overlapped the ATAC peak). We further computed the number of TFs that bind in each ATAC peak by summing binarized matrix entries across TFs. For a given peak-peak pair, we also computed the number of TFs that bind both peaks as the dot product between the binarized TFBS vectors for both peaks.

Defining TFBS for genes. For a given TF and data set-cell type pair, we also defined a TF binding a gene as any overlap of the gene's promoter ($\text{TSS} \pm 1\text{kb}$) with a ChIP-seq peak for that TF in the corresponding cell type; we thus generated a binarized TFBS matrix for genes in each data set-cell type pair. We further computed the number of TFs that bind in each gene's promoter by summing binarized matrix entries across TFs. For a given peak-gene pair, we also computed the number of TFs that bind both the peak and the gene's promoter as the dot product between the binarized TFBS vectors for the focal peak and gene.

Defining pioneer TFs. We obtained a list of 43 established pioneer TFs from refs.^{130,131} We intersected this set with the set of 303 TFs in our K562 ChIP-seq data and 148 TFs in our GM12878 ChIP-seq data to define 9 known pioneer TFs in K562 and 8 known pioneer TFs in GM12878. We define 'non-pioneer' TFs as TFs included in our ChIP-seq data but not included in the set of established pioneer TFs^{130,131}; we note that the label 'non-pioneer' does not exclude the possibility that these TFs do exhibit pioneer activity in these or other cell types. We further note that the distinct sets of pioneer TFs in each cell type are a function of ChIP-seq data availability and do not necessarily reflect cell type-specific TF abundance or activity.

Fine-mapping peak-gene links

Computing posterior inclusion probabilities. We performed peak-level fine-mapping using SuSiE¹¹⁴ (v0.14.2) to distinguish causal from tagging peaks. For a focal gene, we ran the function

susie(), using as input standardized expression of the focal gene (“y” parameter) and standardized accessibility of candidate peaks (“X” parameter), to generate posterior inclusion probabilities (PIPs) for each candidate peak.

Fine-mapping approaches. We performed fine-mapping assuming either a single causal peak (with susie() parameter $L = 1$) or multiple causal peaks (with susie() parameter $L = 10$ or the number of candidate peaks, if there were <10 candidate peaks). We also performed fine-mapping assuming either flat priors or functionally informed priors, defined by per-peak causal effect size variances from S-CASC and supplied to the susie() parameter prior_weights (see above; analogous to ref.¹³⁸). We thus defined 4 peak-gene fine-mapping approaches (single causal peak, single causal peak informed, multiple causal peak, multiple causal peak informed).

Defining marginal and fine-mapped linking scores. We define “marginal peak-gene linking score” as squared peak-gene correlation (from ArchR⁸² or Signac^{77,81}; see below). We define “fine-mapped peak-gene linking score” as marginal peak-gene linking score multiplied by the fine-mapped PIP for the focal peak-gene pair. We computed marginal and fine-mapped linking scores for each candidate peak-gene link in each data set-cell type pair. We define “candidate peak-gene links” as the set of peak-gene pairs with peak-TSS distance $<1\text{Mb}$ (for analyses involving the CRISPR^{47–58} evaluation set, we further restricted the set of “candidate peak-gene links” to peak-gene pairs tested by a CRISPR experiment^{47–58}).

Generating marginal and fine-mapped linking scores from Signac. For each data set-cell type pair (**Table 2.1**), we generated separate fine-mapped scores for the ArchR⁸² peak-gene linking method (using metacell-level data pre-processed by ArchR⁸² and standardized across metacells) and the Signac^{77,81} peak-gene linking method (using single-cell-level data pre-processed by the Signac^{77,81} pipeline and standardized across single cells). To compute Signac linking scores, we first generated binarized ATAC matrices (in which an entry of 1 for a given peak-cell pair indicates that ≥ 1 ATAC fragments overlap the peak and an entry of 0 indicates that no ATAC fragments overlap the peak) and normalized gene expression matrices (by running

SCTransform¹⁶⁶ on the RNA counts matrix). We computed peak-gene correlations by running the Signac^{77,81} LinkPeaks() function (with parameters: 1e+06, pvalue_cutoff = 1, score_cutoff = 0; these values of “pvalue_cutoff” and “score_cutoff” ensured that all peak-peak correlations $\in [-1,1]$ were returned).

Gene-level fine-mapping. We also performed gene-level fine-mapping using SuSiE¹¹⁴ to distinguish causal from tagging target genes. For a focal peak, we ran the function susie(), using as input standardized accessibility of the focal peak (“y” parameter) and standardized expression of candidate genes (“X” parameter), to generate posterior inclusion probabilities (PIPs) for each candidate gene. As above, we performed fine-mapping assuming either a single causal target gene or multiple causal target genes (with susie() parameter L = 1 or 10, respectively). We also performed fine-mapping assuming either flat priors or functionally informed priors, defined by per-peak causal effect size variances from stratified co-expression score regression. We defined “marginal peak-gene linking score” and “fine-mapped peak-gene linking score” as above.

Evaluating fine-mapped peak-gene links

To assess the performance of peak-gene linking methods, we used an enrichment-based evaluation framework and peak-gene link evaluation sets described in ref.⁹⁰ We defined 2 evaluation sets of peak-gene links derived from eQTL^{63,64} and CRISPR^{47–58} data (also described in ref.⁹⁰). Brief descriptions of the evaluation sets are provided below:

The eQTL evaluation set is comprised of 39,194 fine-mapped eSNP-eGene pairs attaining a posterior inclusion probability (PIP) > 0.5 in any GTEx tissue^{63,64}. Fine-mapping was conducted by ref.⁶⁴ (by applying the SuSiE¹¹⁴ fine-mapping method to 49 GTEx tissues⁶³). For each data set-cell type pair, we defined a peak-gene pair as a “true” link if the focal peak harbored any SNP with PIP > 0.5 to the focal gene; we classified all other peak-gene pairs as “false” links.

The CRISPR evaluation set (defined by ref.⁵⁸) is comprised of 448 enhancer-gene links experimentally validated by CRISPR interference experiments^{47–57}. Validated links are defined as those marked “TRUE” in the “Regulated” column of the combined CRISPR evaluation data set.

For each data set-cell type pair, we defined a peak-gene pair as a “true” CRISPR link if the focal peak overlapped any enhancer with a validated causal effect on the focal gene; we defined a peak-gene pair as a “false” CRISPR link if the focal peak overlapped any genomic region experimentally tested for an effect on the focal gene and found to be non-causal.

We used an enrichment-based metric (also described in ref.⁹⁰), *average enrichment across recall values*, to assess the ability of each linking method to prioritize (assign higher linking scores to) “true” links over “false” links (as defined by an evaluation set of peak-gene links).

For a given evaluation set and given method, we ranked candidate peak-gene links by linking score. At each unique linking score c assigned to any candidate peak-gene link, we computed enrichment and recall as follows:

$$\text{Enrichment} = \frac{|\text{predicted peak-gene links} \cap \text{true peak-gene links}| / |\text{true peak-gene links}|}{|\text{predicted peak-gene links}| / |\text{candidate peak-gene links}|} \quad [3.5]$$

$$\text{Recall} = \frac{|\text{predicted peak-gene links} \cap \text{true peak-gene links}|}{|\text{true peak-gene links}|} \quad [3.6]$$

where “predicted peak-gene links” denotes the binarized set of peak-gene links attaining linking score $\geq c$, “true peak-gene links” denotes the binarized set of “true” peak-gene links defined by the evaluation set (intersected with the set of “candidate peak-gene links”), and “candidate peak-gene links” denotes the set of peak-gene pairs with peak-TSS distance $< 1\text{Mb}$ (as noted above, for analyses involving the CRISPR^{47–58} evaluation set, we further restricted the set of “candidate peak-gene links” to peak-gene pairs tested by a CRISPR experiment^{47–58}). We restricted the set of “candidate peak-gene links” to links involving genes with ≥ 1 “true” link. Intuitively, equation 3.4 quantifies the extent of overlap between predicted peak-gene links and evaluation SNP-gene links relative to the overlap expected by chance.

For a given evaluation set and set of methods, we used equation 3.5 and equation 3.6 to construct enrichment-recall curves and observed $R \leq 1$, the maximum recall achieved by all

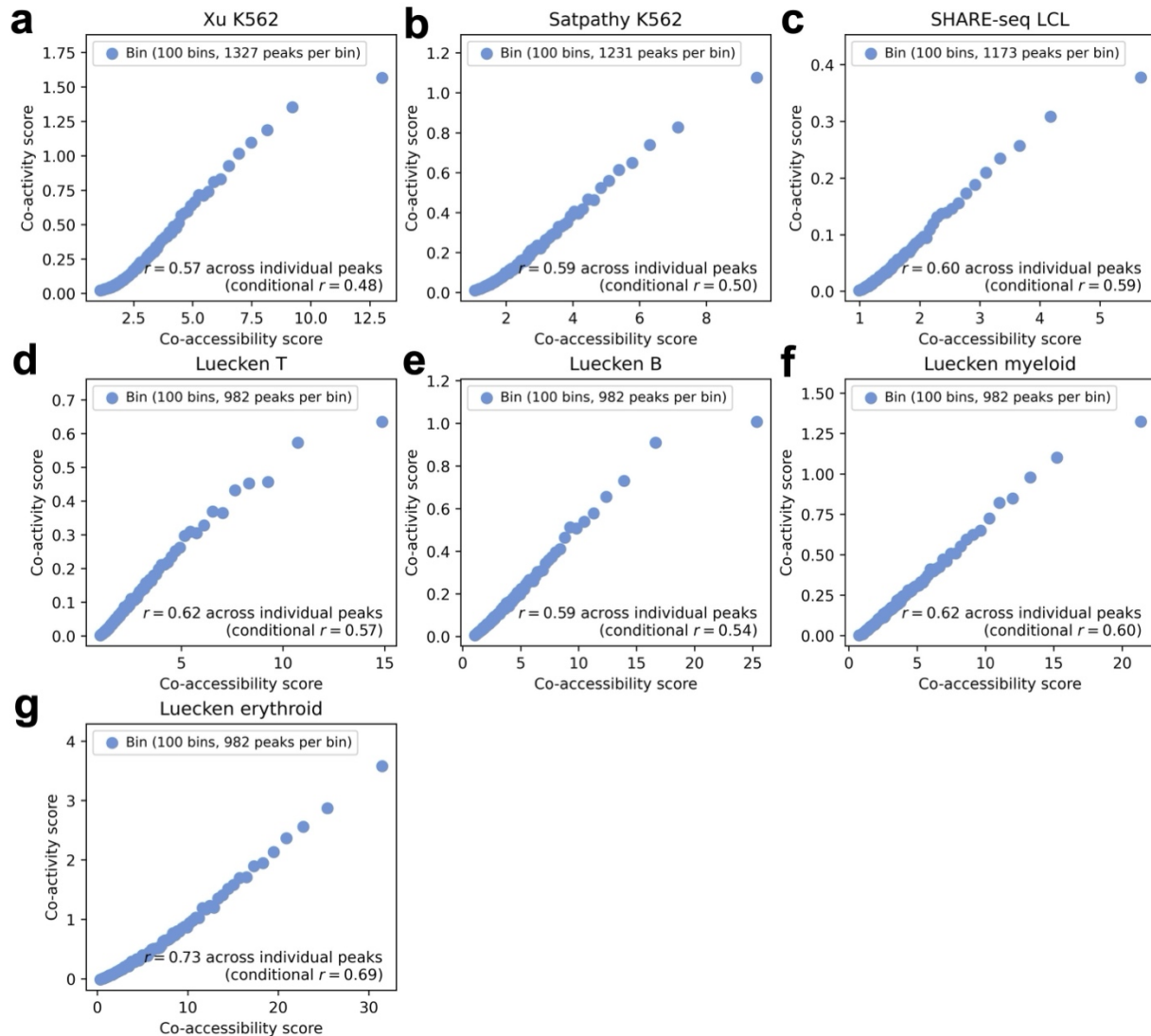
methods (e.g. if ArchR and fine-mapped ArchR attained maximum recall values of R_1 and R_2 , respectively, we defined R as the minimum of R_1 and R_2 . For each method, we then defined *average enrichment across recall values* (or *average enrichment*, shortened) as follows:

$$\text{Average enrichment} = (\sum_{i=1}^{n_R} \text{enrichment}_i * (1 - \text{recall}_i)) / n_R \quad [3.7]$$

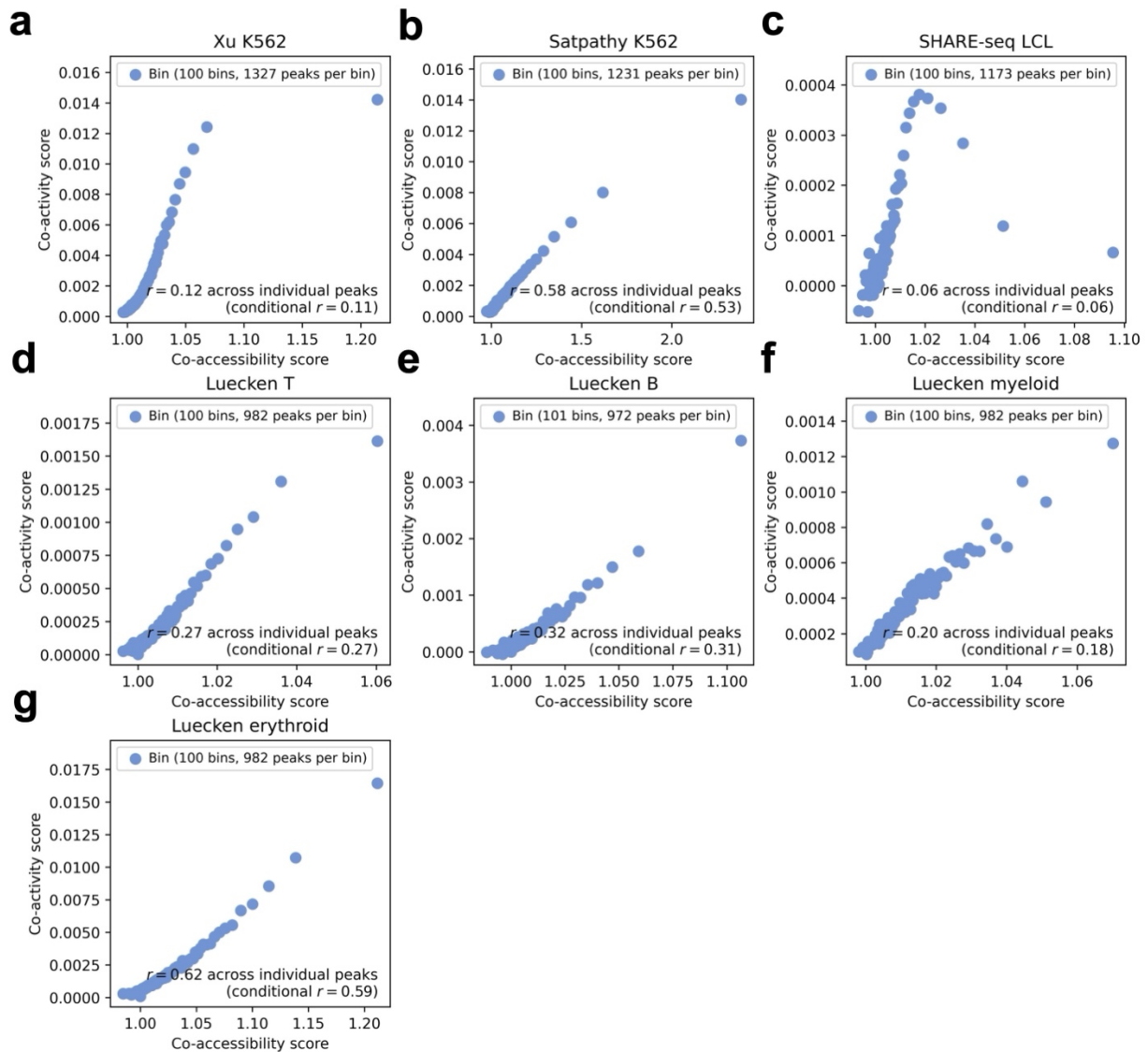
where n_R denotes the total number of “true” evaluation links corresponding to recall R and $(\text{recall}_i, \text{enrichment}_i)$ for $i \in \{1 \dots n\}$ denote the values of each point on the enrichment-recall curve. Multiplying enrichment_i by $(1 - \text{recall}_i)$ upweights enrichments at lower recall values, prioritizing method performance at high score thresholds. Intuitively, this metric can be interpreted as a weighted area under the enrichment-recall curve.

We obtained standard errors on average enrichment for each method, as well as standard errors and p-values on the difference in average enrichment between two methods, by bootstrapping genes using 1,000 bootstrap iterations (as in ref.⁹⁰).

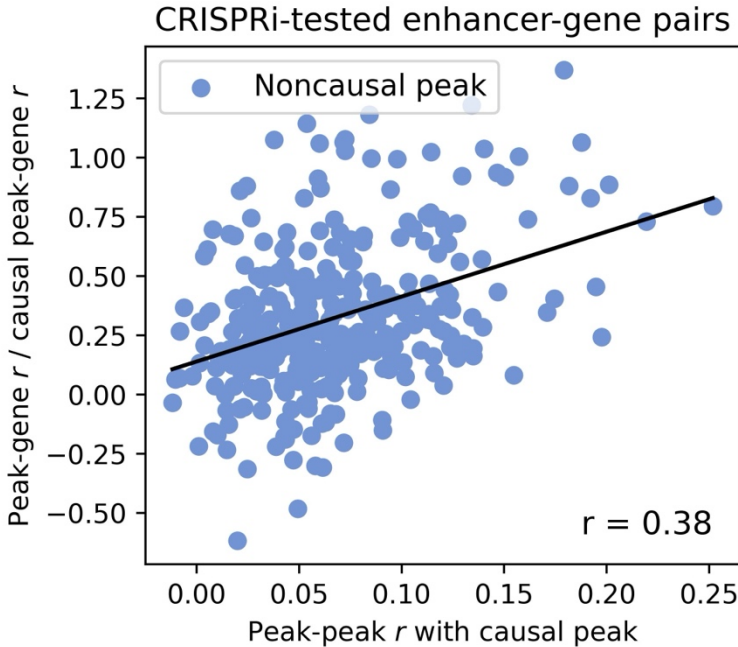
Supplementary Figures



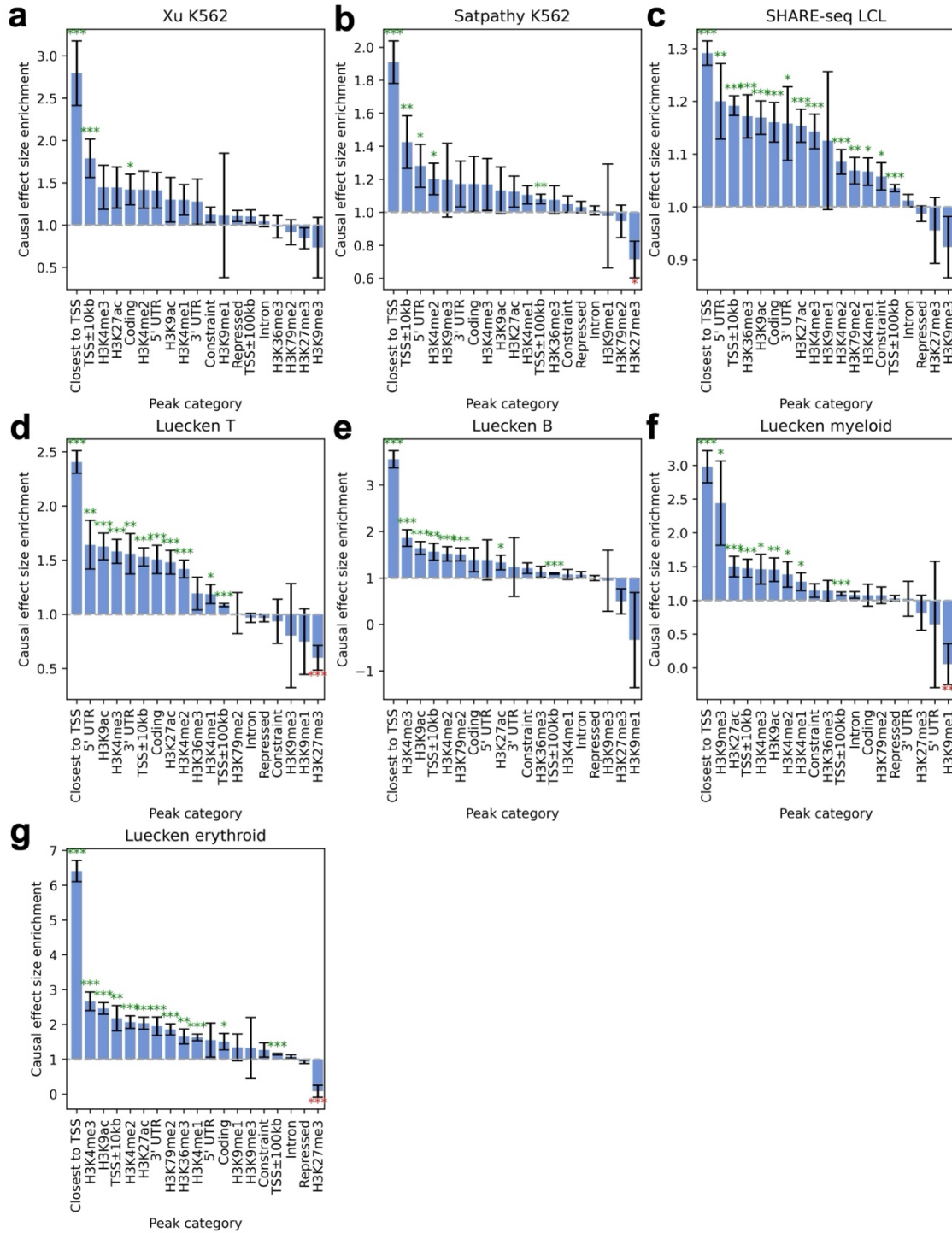
Supplementary Figure 3.1. Relationship between co-accessibility score and co-activity score. Relationship between co-accessibility score and co-activity score across peaks in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal peak. ‘Conditional r ’ denotes correlation conditioned on number of genes <1Mb from the focal peak. Peaks are partitioned equally into 100 bins (each represented by 1 point) by co-accessibility score.



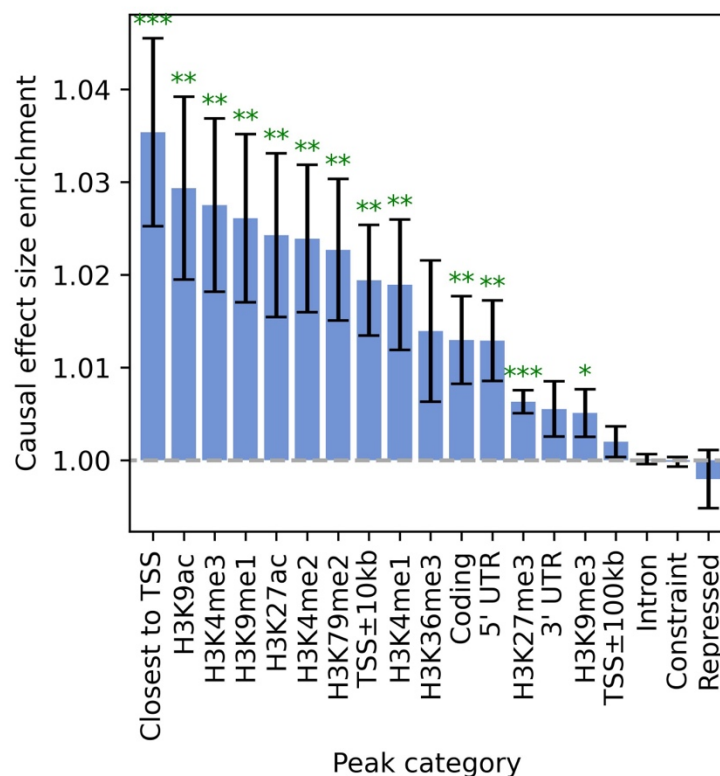
Supplementary Figure 3.2. Relationship between co-accessibility score and co-activity score computed using correlations across single cells (instead of metacells). Relationship between co-accessibility score and co-activity score across peaks in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal peak, with peak-peak and peak-gene correlations computed across single cells (instead of metacells). ‘Conditional r ’ denotes correlation conditioned on number of genes <1Mb from the focal peak. Peaks are partitioned equally into 100 bins (each represented by 1 point) by co-accessibility score.



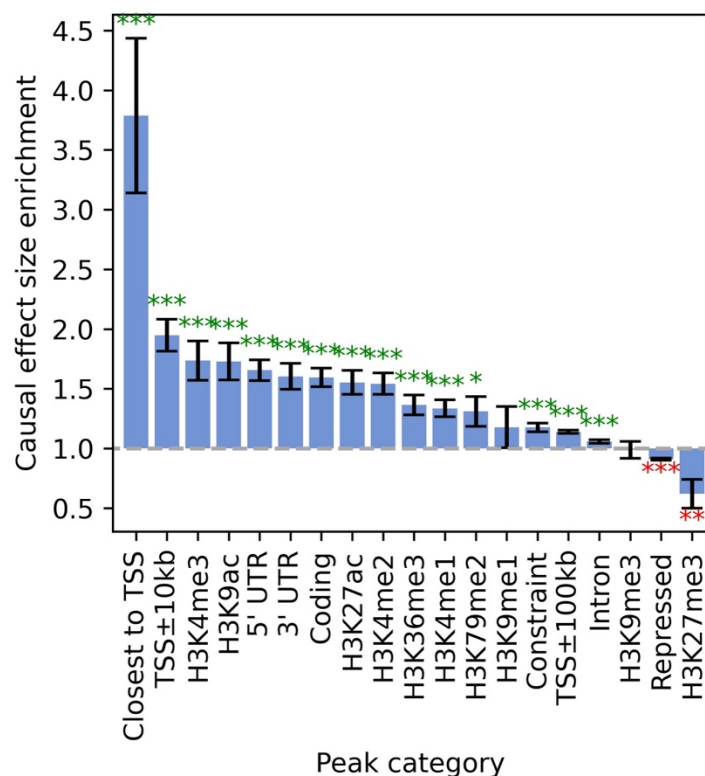
Supplementary Figure 3.3. Relationship between non-causal peak-gene correlation and tagging peak correlation using correlations computed across single cells (instead of metacells). Relationship between correlation with the CRISPR-validated causal peak and correlation with the target gene (relative to the causal peak's correlation with the target gene) across non-causal peaks tested by CRISPRi, with correlations computed across single cells (instead of metacells). We analyzed CRISPRi-tested peak-gene pairs, restricting to 28 genes with both an experimentally validated causal (CRISPR-positive) peak that is identified by Signac (with peak-gene correlation > 0.05 and $p\text{-value} < 0.05$) and at least one experimentally tested non-causal (CRISPR-negative) peak overlapping our single-cell data. We restricted this analysis to the 3 data set-cell type pairs most relevant to K562 cells: Xu K562, Satpathy K562, and Luecken erythroid. We measured the correlation of each CRISPR-negative peak with both the CRISPR-positive peak and the focal gene. Correlation with the CRISPR-positive peak was strongly correlated to correlation with the focal gene.



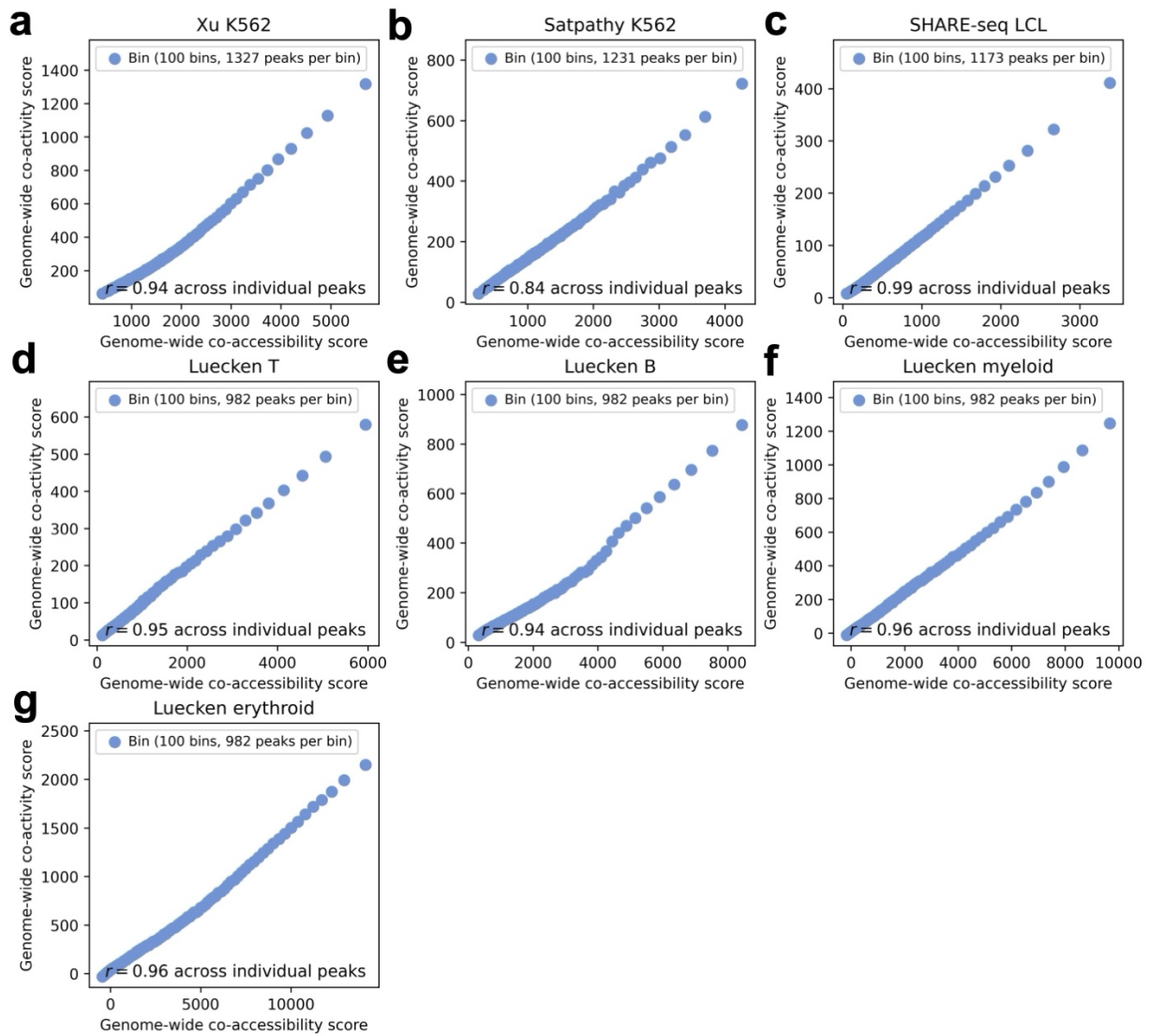
Supplementary Figure 3.4. S-CASC causal effect size enrichments. Causal effect size enrichment of 19 peak categories in stratified co-accessibility score regression in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



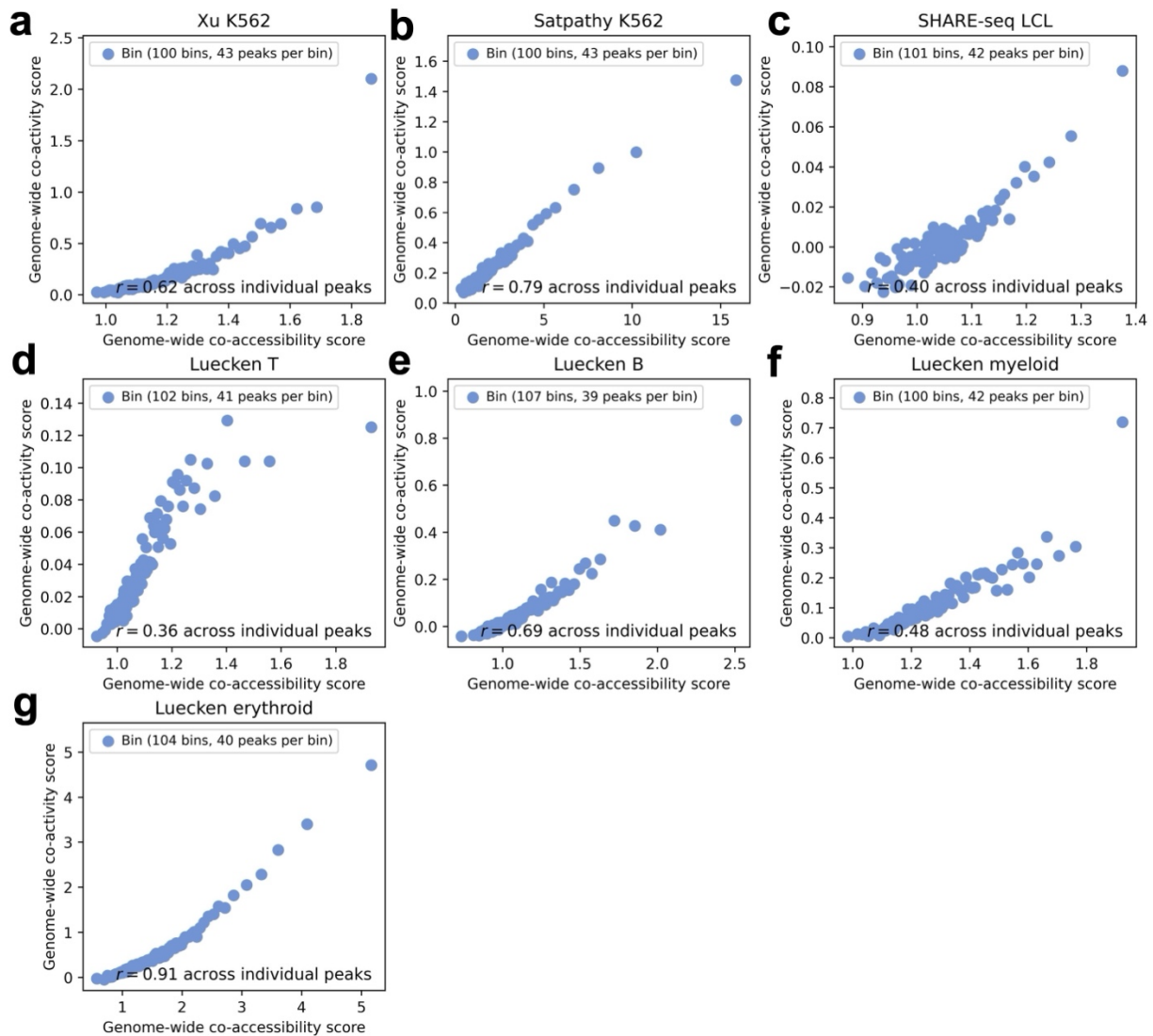
Supplementary Figure 3.5. S-CASC causal effect size enrichments computed using correlations across single cells (instead of metacells). Causal effect size enrichment of 19 peak categories in stratified co-accessibility score regression, implemented using peak-peak and peak-gene correlations computed across single cells (instead of metacells) and meta-analyzed across all 7 data set-cell type pairs. Results are qualitatively similar to **Figure 3.2c**, except for significant enrichments for the H3K27me3 and H3K9me1 categories (repressive histone marks) that were not observed at the metacell level. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



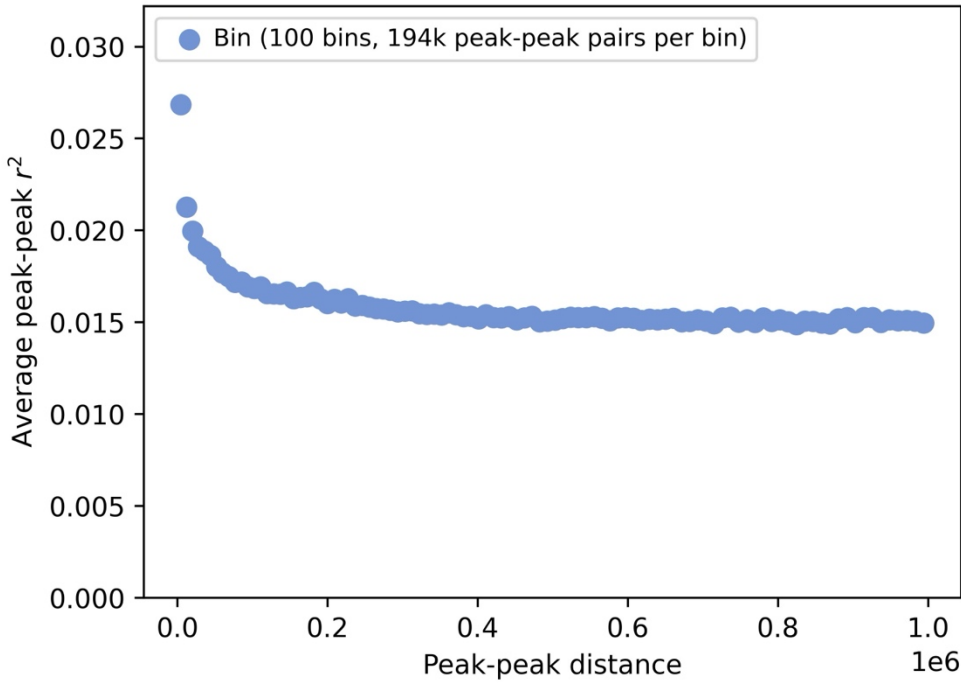
Supplementary Figure 3.6. S-CASC causal effect size enrichments without conditioning on number of nearby genes. Causal effect size enrichment of 19 peak categories in stratified co-accessibility score regression across peaks, implemented without the covariate for the number of genes <1Mb from each peak and meta-analyzed across all 7 data set-cell type pairs. Results are very similar to **Figure 3.2c**; differences include a small but significant enrichment for the “Intron” category and a small but significant depletion for the “Repressed” category (both nonsignificant in **Figure 3.2c**). Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



Supplementary Figure 3.7. Relationship between genome-wide co-accessibility score and genome-wide co-activity score. Relationship between genome-wide co-accessibility score and genome-wide co-activity score across peaks in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs, computed using all peaks and genes genome-wide. Peaks are partitioned equally into 100 bins (each represented by 1 point) by co-accessibility score.

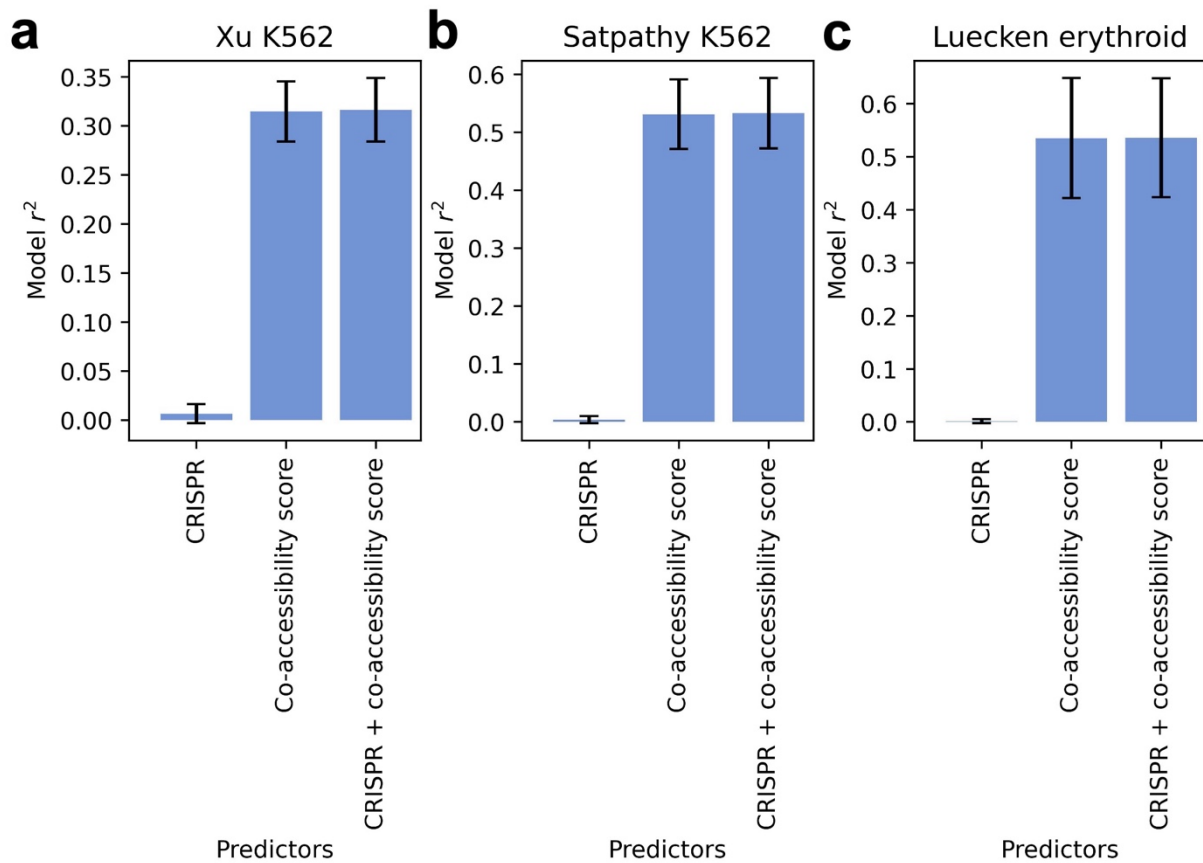


Supplementary Figure 3.8. Relationship between genome-wide co-accessibility score and genome-wide co-activity score computed using correlations across single cells (instead of metacells). Relationship between genome-wide co-accessibility score and genome-wide co-activity score (using a subset of 5000 peaks and 5000 genes) across peaks in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed using all peaks and genes genome-wide, with peak-peak and peak-gene correlations computed across single cells (instead of metacells). We downsampled to 5000 randomly selected peaks and 5000 randomly selected genes when computing scores (due to the high computational cost of computing correlations for every peak-peak, gene-gene, and peak-gene pair genome-wide across single cells). Peaks are partitioned equally into 100 bins (each represented by 1 point) by genome-wide co-accessibility score.

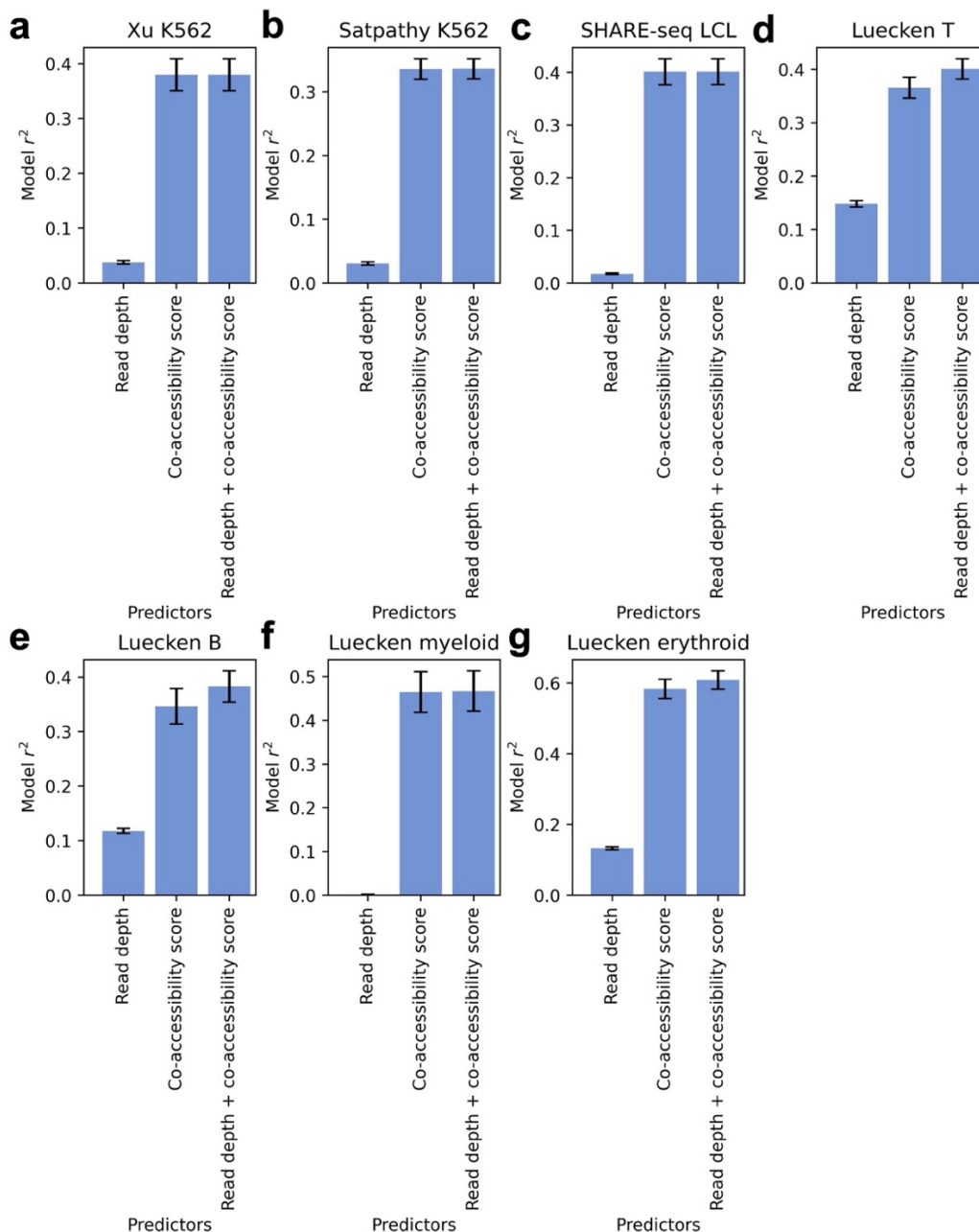


Supplementary Figure 3.9. Average peak-peak r^2 at different peak-peak distances.

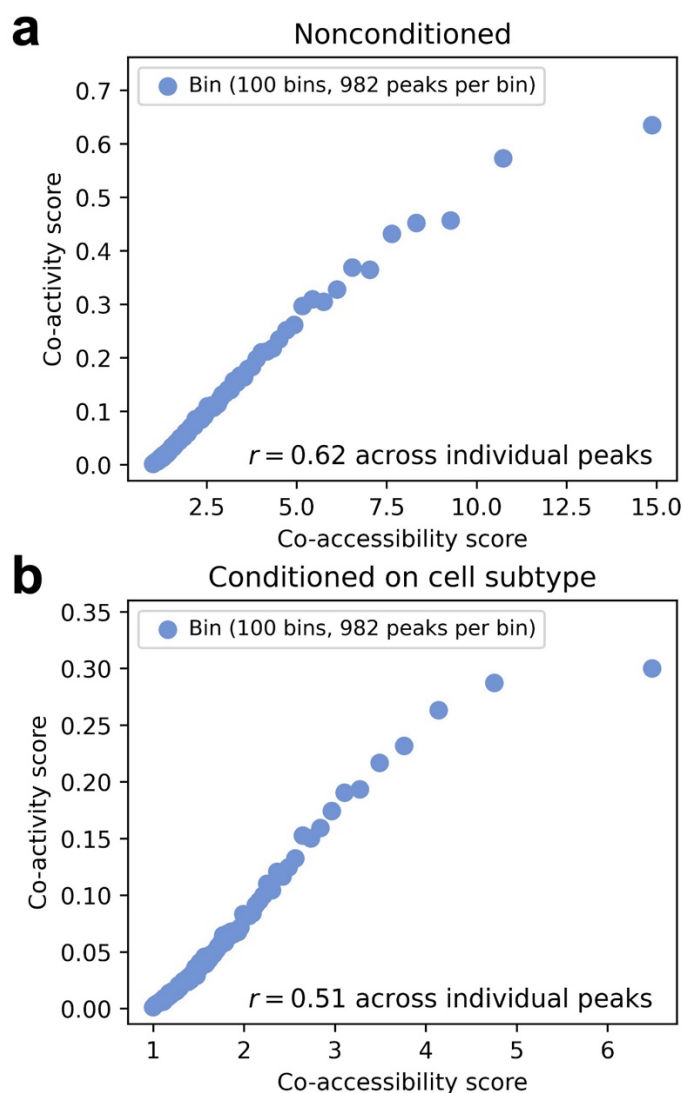
Relationship between average peak-peak squared correlation and peak-peak distance across 132,736 peaks in the Xu K562 data set. Peaks are partitioned equally into 100 bins (each represented by 1 point) by peak-peak distance. Peak-peak r^2 is moderately higher at shorter peak-peak distances, possibly implicating *cis* TF effects (the same copy of a TF binding to multiple nearby sites).



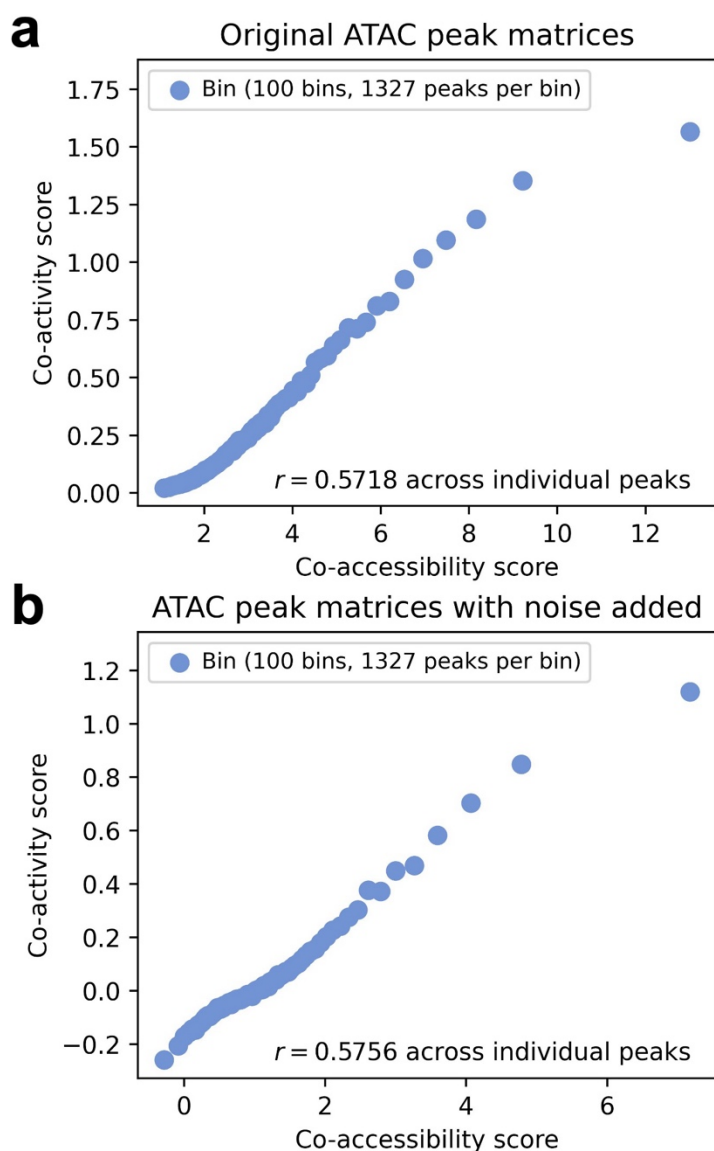
Supplementary Figure 3.10. Regression of co-activity score on CRISPR status and co-accessibility score. Model r^2 for linear regressions of co-activity score on CRISPR-positive status ("CRISPR"), co-accessibility score, or both, for the **a)** Xu K562, **b)** Satpathy K562, and **c)** Luecken erythroid data set-cell type pairs. Peaks in more 'active' regulatory regions (higher co-accessibility scores) may be more likely to truly affect gene expression. To assess the role of biological causality in driving the correlation between co-activity score and co-accessibility score, we performed a bivariate linear regression of co-activity score on co-accessibility score and a binary indicator of CRISPR-positive status (1 if a peak has at least one causal target gene validated by CRISPR, 0 otherwise) across peaks. We also performed univariate regressions of co-activity score on co-accessibility score and CRISPR-positive status individually. We restricted this analysis to the 3 data set-cell type pairs most relevant to K562 cells: Xu-K562, Satpathy-K562, and Luecken-erythroid. Within each data set-cell type pair, we further restricted to genes with ≥ 1 peak validated by CRISPR (CRISPR-positive) and ≥ 1 CRISPR-tested non-causal (CRISPR-negative) peak. Bars and confidence intervals denote estimates and standard errors, respectively.



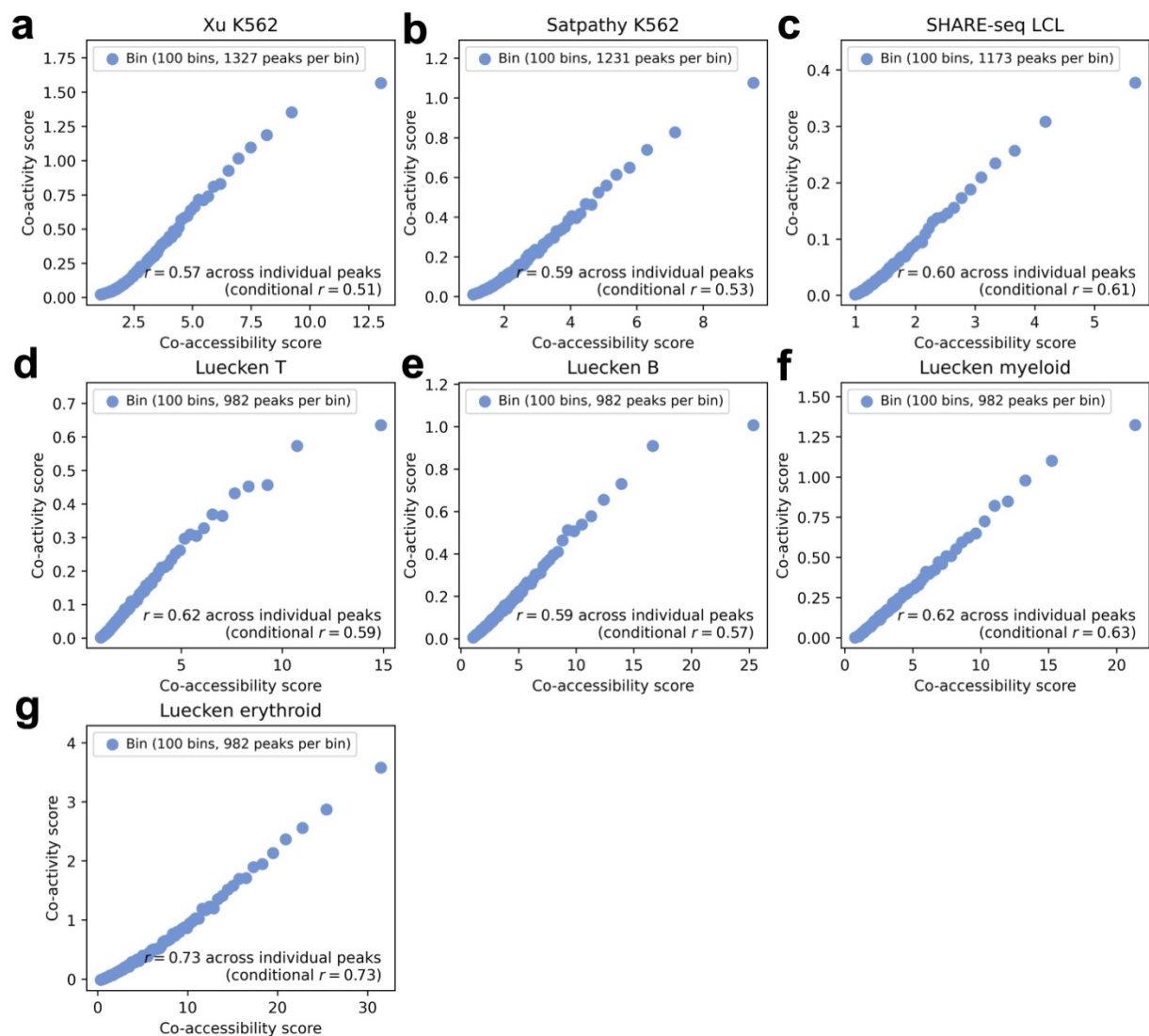
Supplementary Figure 3.11. Regression of co-activity score on peak read depth and co-accessibility score. Model r^2 for linear regressions of co-activity score on peak read depth, co-accessibility score, or both in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs. Peaks with higher read depth may have higher measured correlations with other peaks and genes, generating higher co-accessibility scores and higher co-activity scores. To assess the role of read depth in driving the correlation between co-activity score and co-accessibility scores, we performed a bivariate linear regression of co-activity score on co-accessibility score and $\log(\text{read depth})$ across peaks. We also performed univariate regressions of co-activity score on co-accessibility score and read depth individually. Bars and confidence intervals denote estimates and standard errors, respectively. (We note that log-transformed read depth attained higher r^2 than raw read depth.)



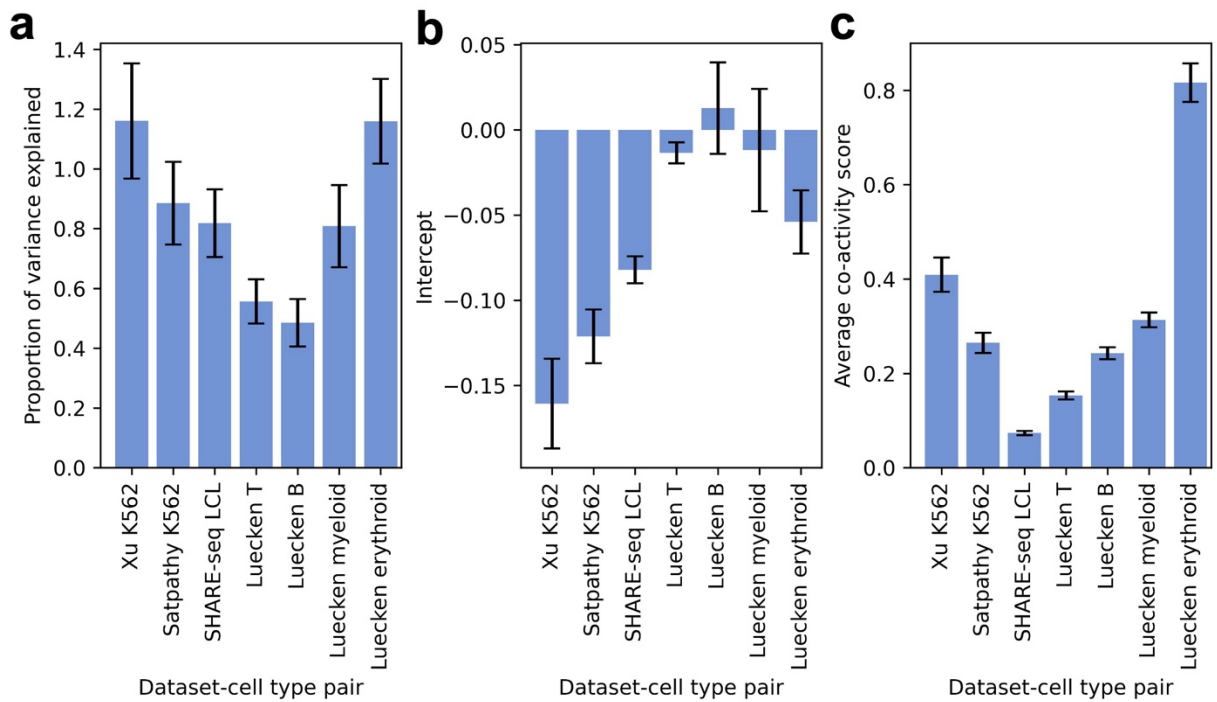
Supplementary Figure 3.12. Relationship between co-accessibility score and co-activity score conditioned on cell subtype. Relationship between co-accessibility score and co-activity score across peaks in the Luecken T cell data set computed using **a)** unconditioned peak-peak and peak-gene correlations and **b)** peak-peak and peak-gene correlations conditioned on cell subtype proportions. Cell subtypes may be associated with varying levels of average peak accessibility and gene expression, driving a correlation between co-accessibility score and co-activity score. To assess the role of cell subtypes, we analyzed T cells from the Luecken data set, which span 4 labeled T cell subtypes (activated CD4+, naive CD4+, CD8+, and naive CD8+; each represented by 1-12k cells). We separately residualized gene expression and ATAC peak accessibility by cell type proportions in each metacell (via linear regression across metacells), recomputed peak-peak correlations and peak-gene correlations using residualized data, and recomputed co-accessibility scores and co-activity scores by summing these conditional correlations. The correlation between co-activity score and co-accessibility score was lower after accounting for cell subtypes, suggesting that, while major cell subtypes do not fully explain the correlation, peak tagging may be related to cell subtype/state. Peaks are partitioned equally into 100 bins (each represented by 1 point) by co-accessibility score.



Supplementary Figure 3.13. Relationship between co-accessibility score and co-activity score with ATAC measurement noise added. Relationship between co-accessibility score and co-activity score across peaks in the Xu K562 data set computed using **a)** the original ATAC peak matrix and **b)** an ATAC peak matrix with added measurement noise. ATAC peaks with higher amounts of measurement noise may have lower measured correlations with other peaks and genes, possibly driving the correlation between co-activity score and co-accessibility score. Because it is difficult to directly assess the amount of measurement noise in ATAC data, we introduced additional artificial noise to 50% of ATAC peaks (by randomly setting 20% of entries equal to 0, retaining the original data for the remaining 50% of peaks), and recomputed co-accessibility and co-activity scores.

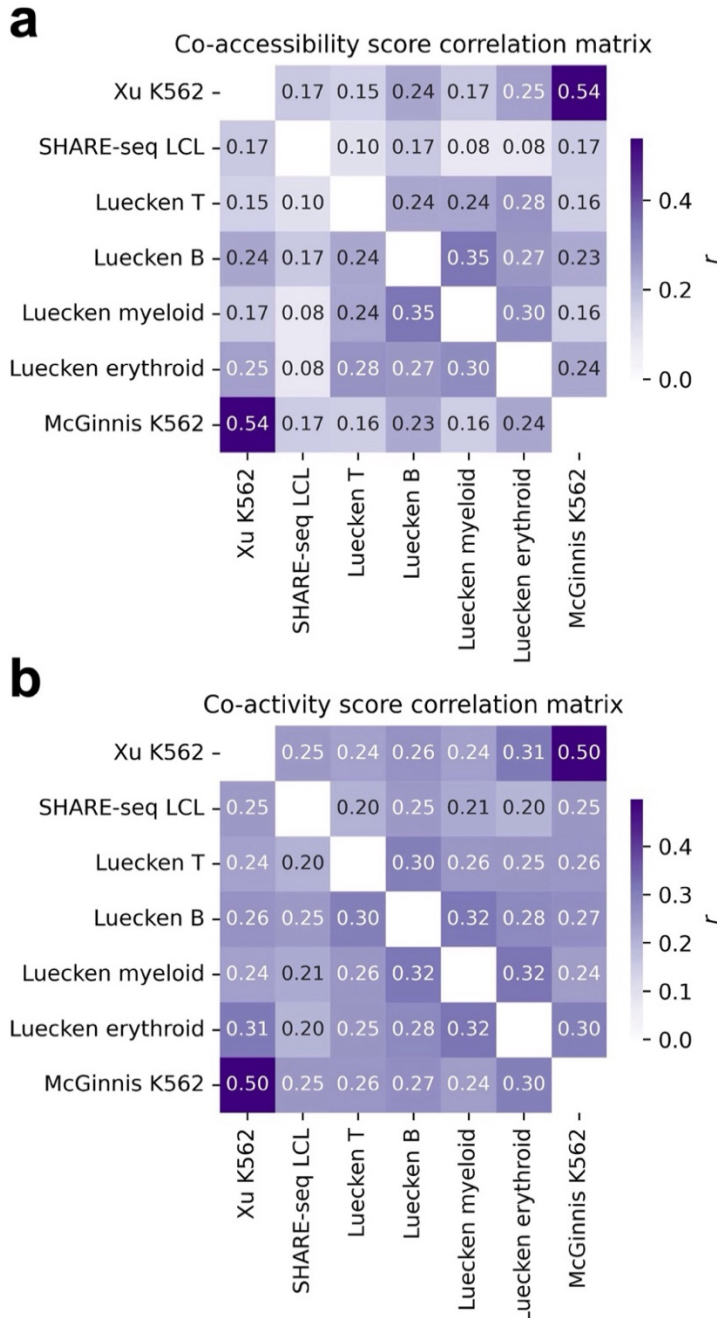


Supplementary Figure 3.14. Relationship between co-accessibility score and co-activity score conditioned on number of nearby genes and peaks. Relationship between co-accessibility score and co-activity score across peaks in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal peak. ‘Conditional r ’ denotes correlation conditioned on both the number of genes <1Mb from the focal peak and the number of peaks <1Mb from the focal peak (instead of only the number of genes <1Mb from the focal peak). Peaks are partitioned equally into 100 bins (each represented by 1 point) by co-accessibility score.

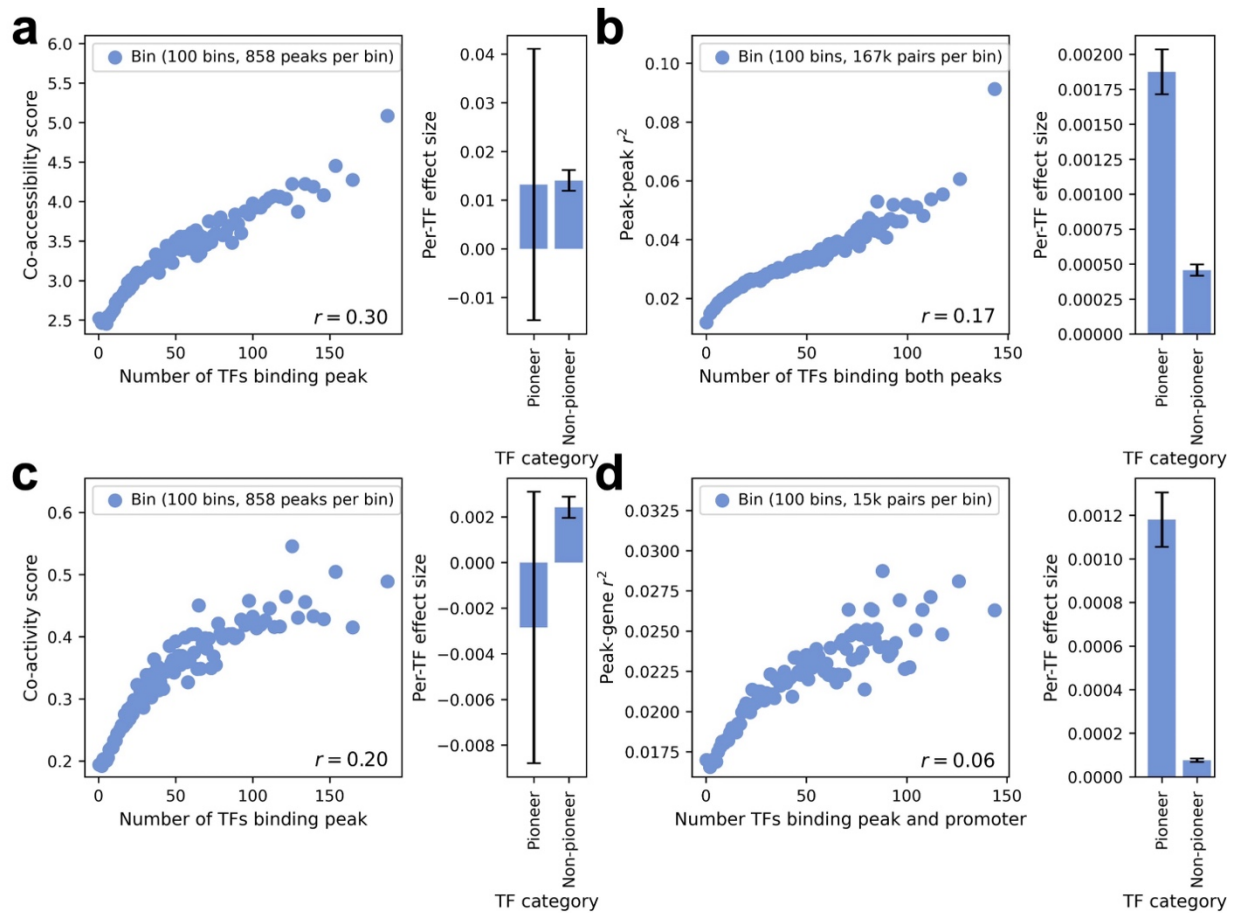


Supplementary Figure 3.15. Estimates of proportion of gene expression variance explained by measured *cis*-peak accessibility from co-accessibility score regression. a)

Estimates of proportion of variance in gene expression explained by measured *cis*-peak accessibility, **b)** regression intercepts, and **c)** average co-activity score from co-accessibility score regression (performed without stratifying by functional category) in each data set-cell type pair. We regressed co-activity score on co-accessibility score across peaks (equivalent to performing stratified co-accessibility score regression with a single peak category defined by all peaks). We then estimated the proportion of variance explained by measured *cis*-peak accessibility as $\tau/k * M$, where τ is the slope of the regression, k is the average number of genes <1Mb from the focal peak (across peaks), and M is the average number of peaks <1Mb from the focal peak (across peaks). Estimating proportion of gene expression variance explained by peak accessibility is analogous to estimating heritability using LD score regression (LDSC)¹²⁷, with 2 changes: we divide by k because co-activity score is the sum of r^2 across multiple genes instead of r^2 with a single trait as in ref.¹²⁷, and we do not divide by N (number of metacells) because co-activity scores are sums of squared correlations and not χ^2 statistics (which are equal to squared correlations multiplied by N) as in ref.¹²⁷. Importantly, we note that regression slopes may be overestimated when regression intercept is negative relative to the average value of co-activity score (leading to inflated estimates of proportion of gene expression variance explained); indeed, estimates in **a)** are largest for data set-cell type pairs with the most negative intercepts. The estimates in **a)** could be impacted by downward bias due to measurement noise or bias in either direction due to various forms of model misspecification, analogous to LDSC/S-LDSC^{18,32,127}. Bars and confidence intervals denote estimates and standard errors, respectively.

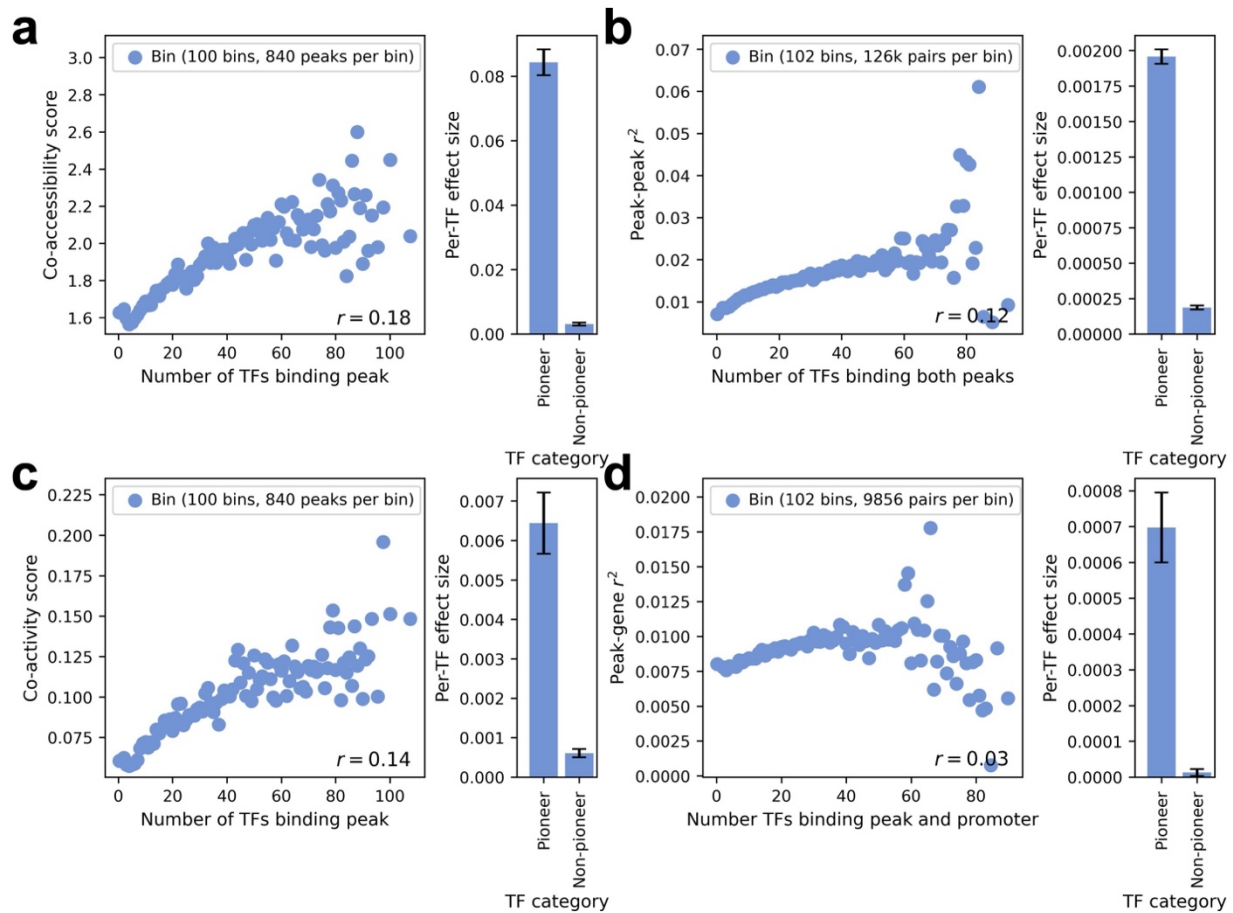


Supplementary Figure 3.16. Correlations between co-accessibility scores and co-activity scores across data set-cell type pairs. a) Pearson correlation matrix of co-accessibility scores computed in 7 data set-cell type pairs across ATAC peaks. **b)** Pearson correlation matrix of co-activity scores computed in 7 data set-cell type pairs across ATAC peaks. For each entry in **a)** and **b)** representing two data set-cell type pairs, correlation is computed across peaks called in both data set-cell type pairs (defined by any overlap between peaks in each data set-cell type pair). Scores tend to show stronger concordance in more similar cell types. For example: scores computed in K562 show the strongest concordance with each other, followed by scores computed in Luecken erythroid (the primary cell type most related to K562), and scores computed in SHARE-Seq LCL are more concordant with scores computed in Luecken B cells (the primary cell type from which LCL are derived) vs. other cell types.



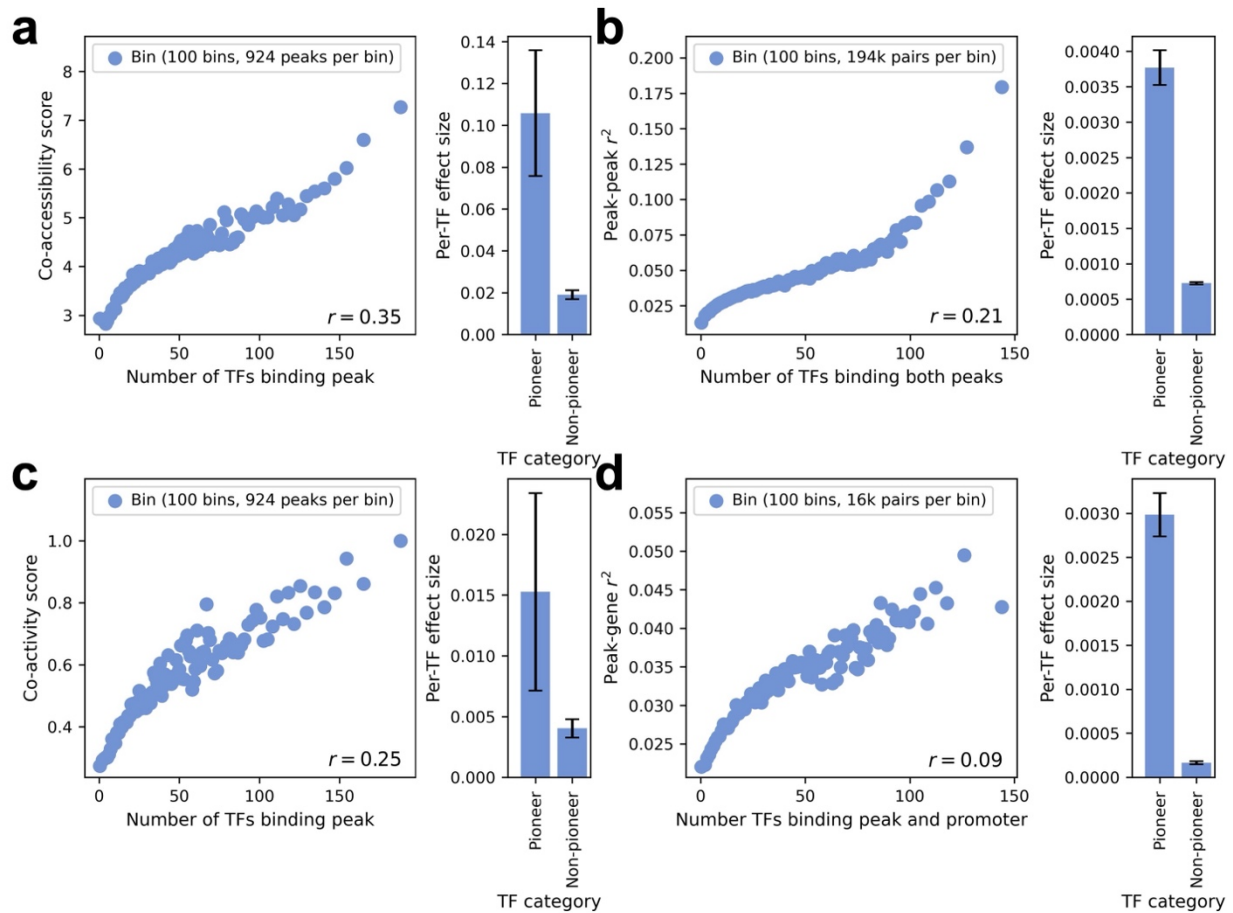
Supplementary Figure 3.17. Relationship between peak-peak co-accessibility, peak-gene co-activity, and transcription factor binding activity in Satpathy K562 data set. a)

Relationship between number of TFs binding an ATAC peak and co-accessibility score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-accessibility score (right). **b)** Relationship between number of TFs binding both peaks in a peak-peak pair and squared peak-peak correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-peak correlation (right). **c)** Relationship between number of TFs binding an ATAC peak and co-activity score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-activity score (right). **d)** Relationship between number of TFs binding both the peak and the gene (promoter) in a peak-gene pair and squared peak-gene correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-gene correlation (right). Correlations and regression effect sizes are computed using 123,113 peaks in the Satpathy K562 data set. In left subpanels of **a-d**, peaks (or peak-peak / peak-gene pairs) are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. In right subpanels of **a-d**, confidence intervals denote standard errors.

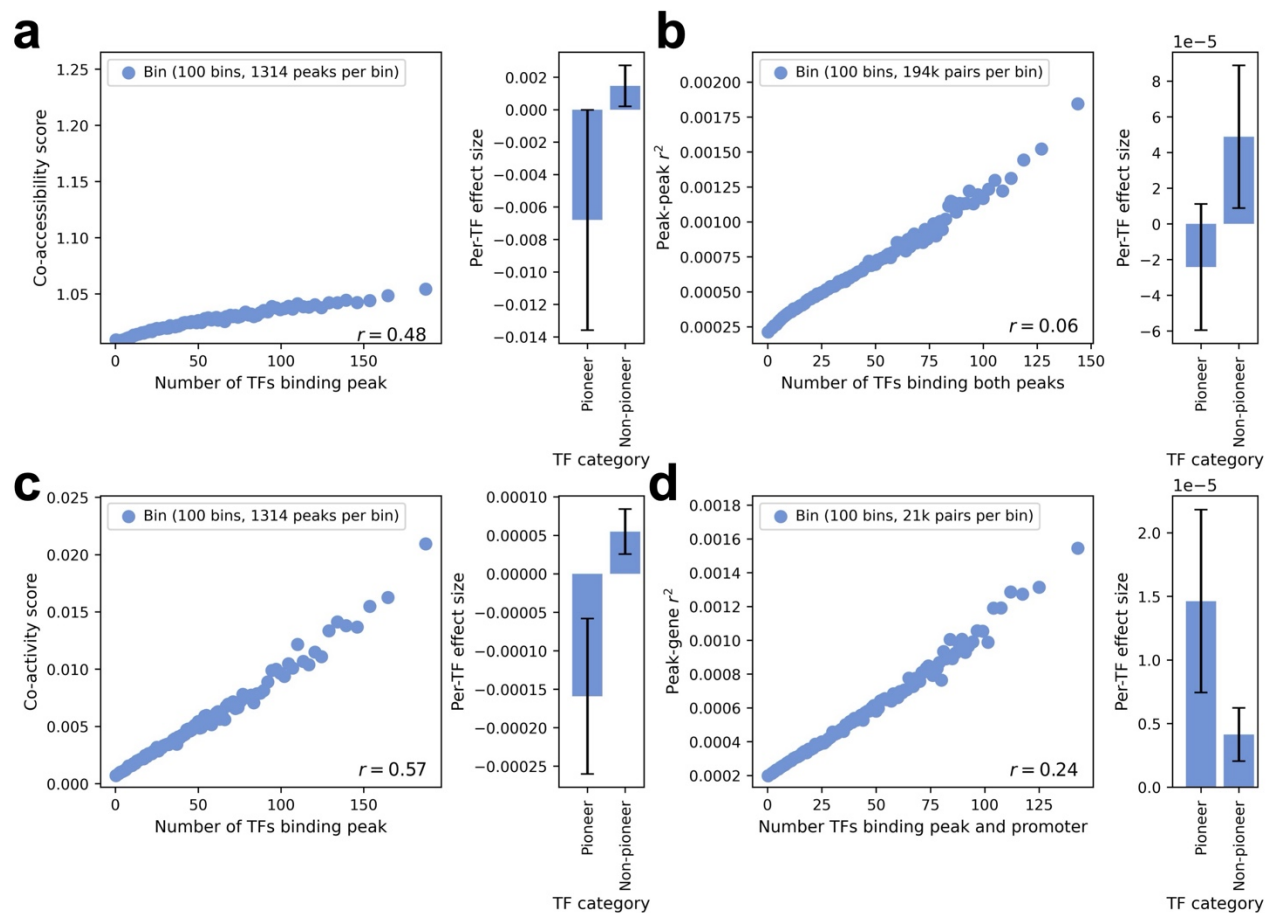


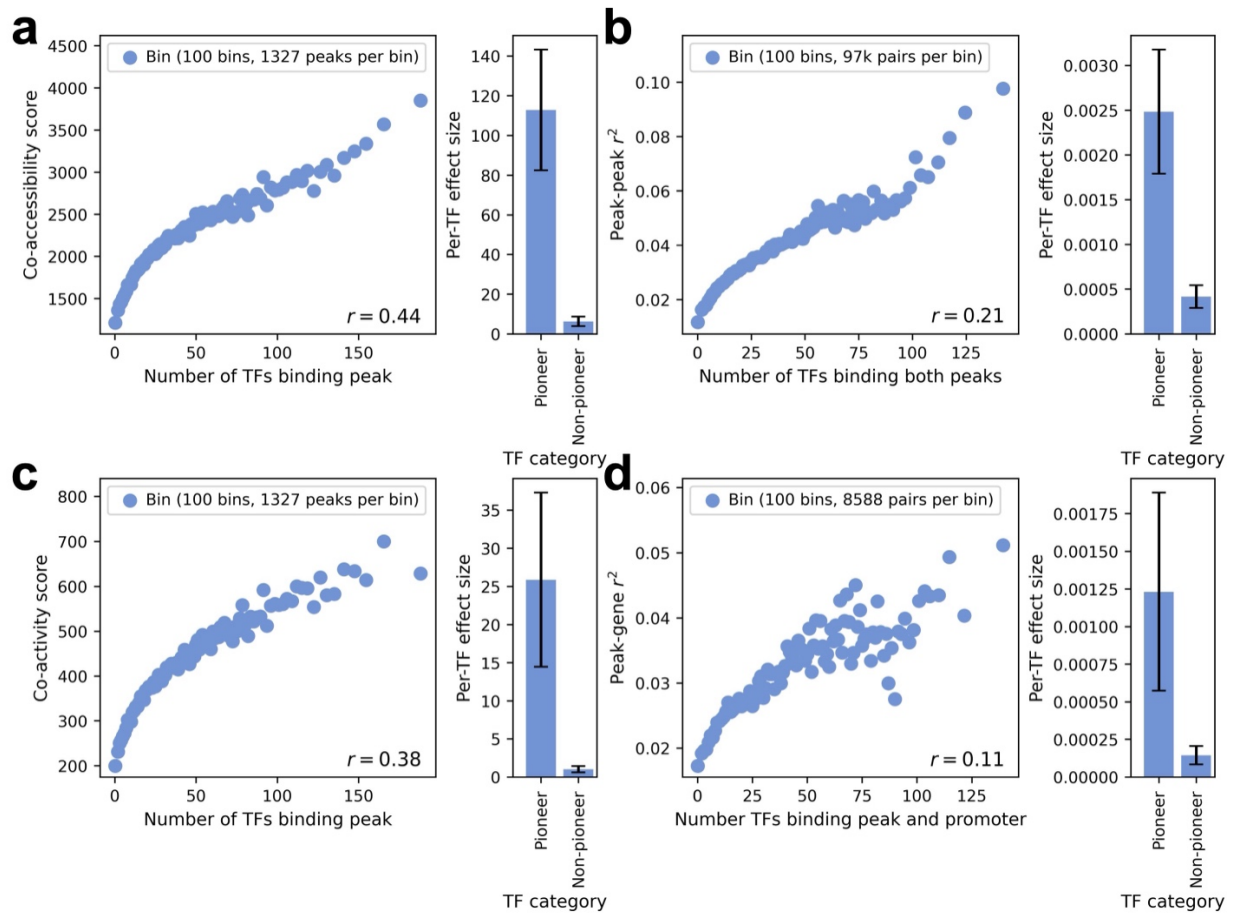
Supplementary Figure 3.18. Relationship between peak-peak co-accessibility, peak-gene co-activity, and transcription factor binding activity in SHARE-seq LCL data set. a)

Relationship between number of TFs binding an ATAC peak and co-accessibility score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-accessibility score (right). **b)** Relationship between number of TFs binding both peaks in a peak-peak pair and squared peak-peak correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-peak correlation (right). **c)** Relationship between number of TFs binding an ATAC peak and co-activity score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-activity score (right). **d)** Relationship between number of TFs binding both the peak and the gene (promoter) in a peak-gene pair and squared peak-gene correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-gene correlation (right). Correlations and regression effect sizes are computed using 117,332 peaks in the SHARE-seq LCL data set. In left subpanels of **a-d**, peaks (or peak-peak / peak-gene pairs) are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. In right subpanels of **a-d**, confidence intervals denote standard errors.

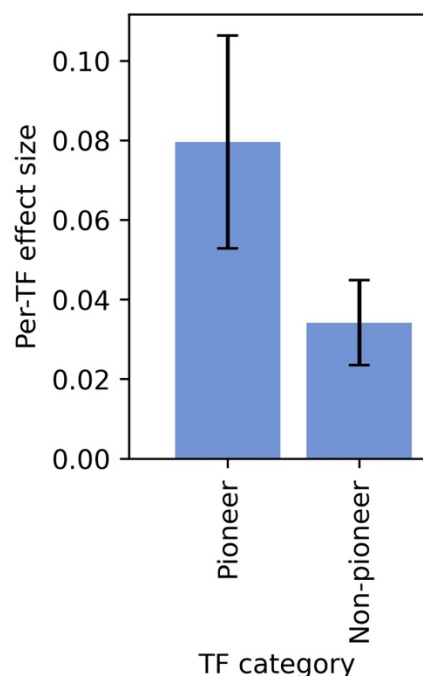


Supplementary Figure 3.19. Relationship between peak-peak co-accessibility, peak-gene co-activity, and transcription factor binding activity in Xu K562 data set. **a)** Relationship between number of TFs binding an ATAC peak and co-accessibility score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-accessibility score (right). **b)** Relationship between number of TFs binding both peaks in a peak-peak pair and squared peak-peak correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-peak correlation (right). **c)** Relationship between number of TFs binding an ATAC peak and co-activity score (left). Per-TF effect for pioneer vs. non-pioneer TFs on co-activity score (right). **d)** Relationship between number of TFs binding both the peak and the gene (promoter) in a peak-gene pair and squared peak-gene correlation (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-gene correlation (right). Correlations and regression effect sizes are computed using 132,736 peaks in the Xu K562 data set. In left subpanels of **a-d**, peaks (or peak-peak / peak-gene pairs) are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. In right subpanels of **a-d**, confidence intervals denote standard errors.

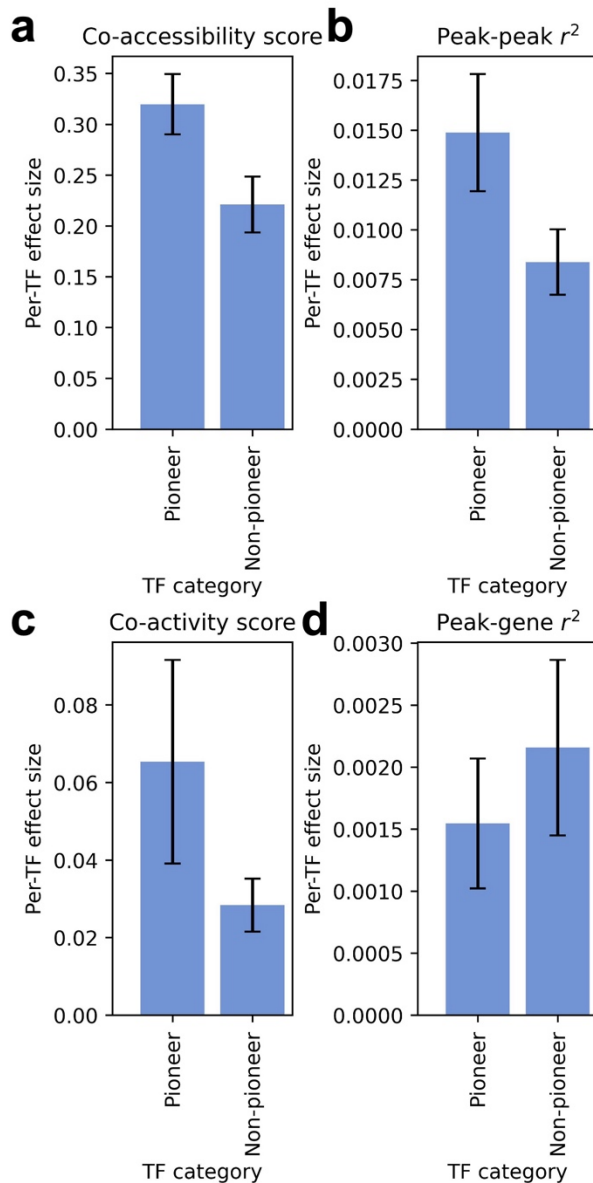




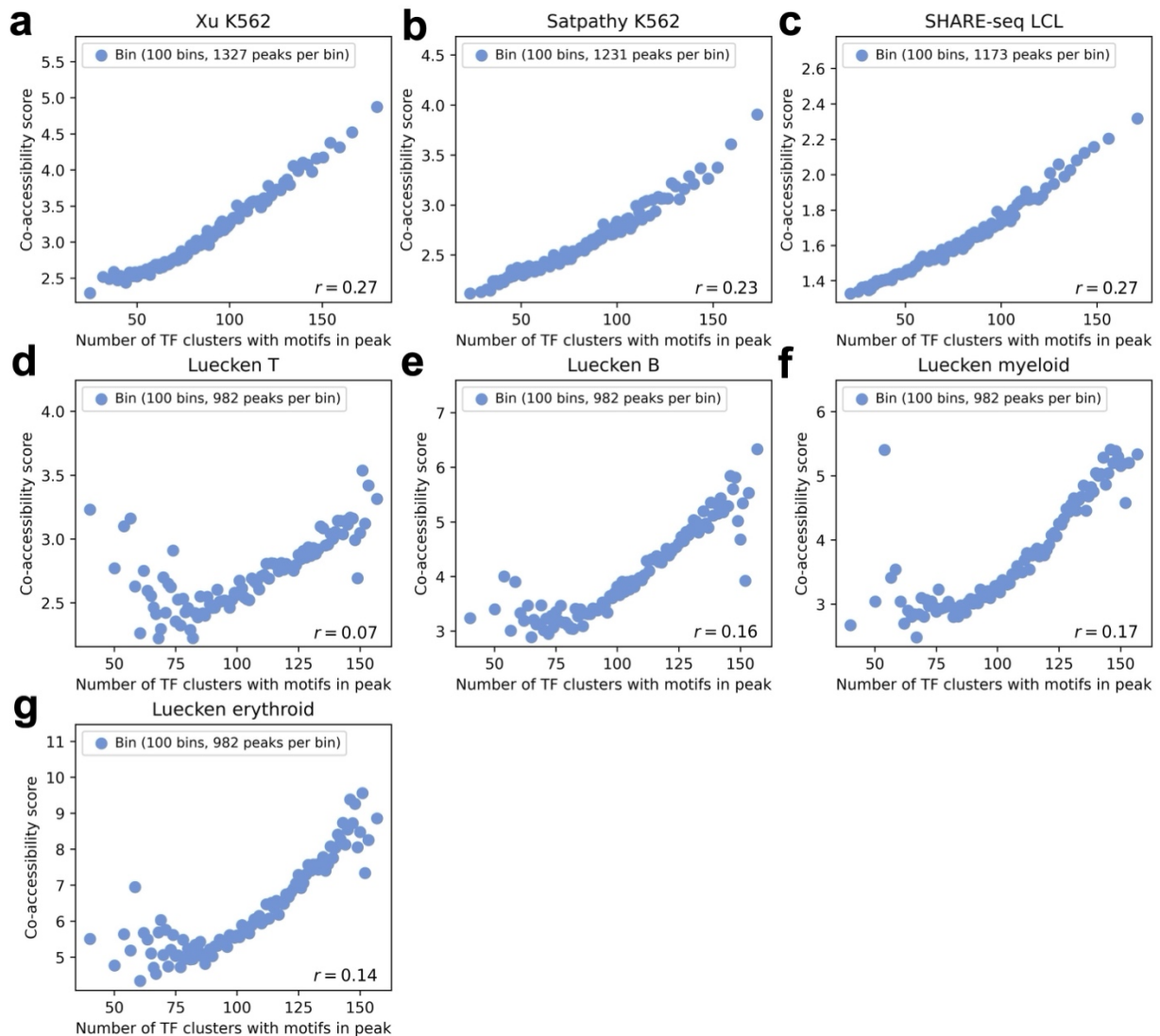
Supplementary Figure 3.21. Relationship between genome-wide peak-peak co-accessibility, genome-wide peak-gene co-activity, and transcription factor binding activity. **a)** Relationship between number of TFs binding an ATAC peak and genome-wide co-accessibility score (left). Per-TF effect for pioneer vs. non-pioneer TFs on genome-wide co-accessibility score (p-value for difference = 0.0002) (right). **b)** Relationship between number of TFs binding both peaks in a peak-peak pair and squared peak-peak correlation, across all peak-peak pairs (not restricted by peak-peak distance) from a random subset of 5,000 ATAC peaks (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-peak correlation, across all peak-peak pairs (not restricted by peak-peak distance) from a random subset of 5,000 ATAC peaks (p-value for difference = 0.0003) (right). **c)** Relationship between number of TFs binding an ATAC peak and genome-wide co-activity score (left). Per-TF effect for pioneer vs. non-pioneer TFs on genome-wide co-activity score (p-value for difference = 0.024) (right). **d)** Relationship between number of TFs binding both the peak and the gene (promoter) in a peak-gene pair and squared peak-gene correlation, across all peak-gene pairs (not restricted by peak-gene distance) from a random subset of 5,000 ATAC peaks and 5,000 genes (left). Per-TF effect for pioneer vs. non-pioneer TFs on squared peak-gene correlation, across all peak-gene pairs (not restricted by peak-gene distance) from a random subset of 5,000 ATAC peaks and 5,000 genes (p-value for difference = 0.070) (right). In left subpanels of **a-d**, correlations are computed across 132,736 peaks in the Xu K562 data set, and peaks (or peak-peak / peak-gene pairs) are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. In right subpanels of **a-d**, regression effect sizes are meta-analyzed across 3 data set-cell type pairs (Xu K562, Satpathy-K562, and SHARE-seq-LCL), and confidence intervals denote standard errors.



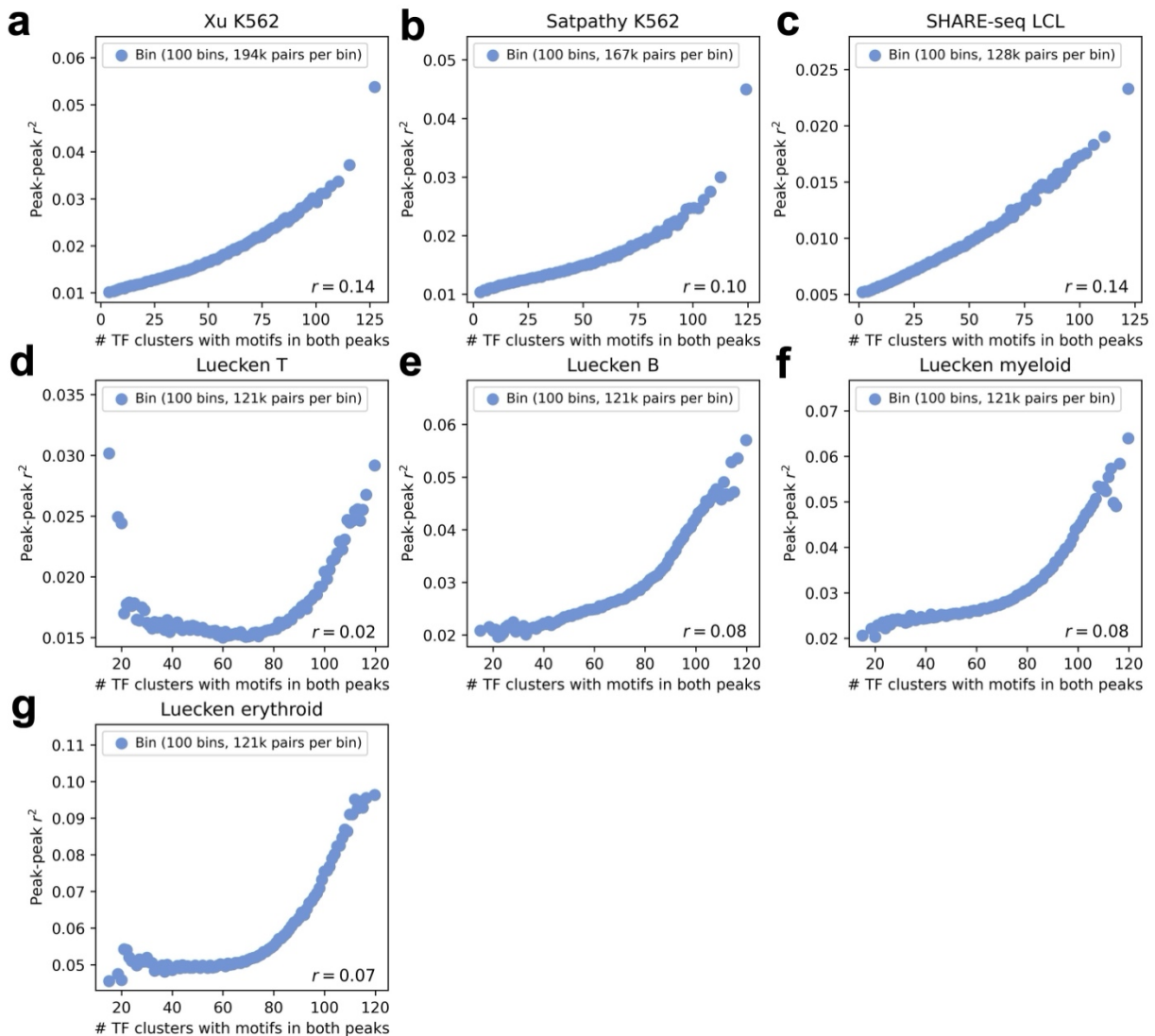
Supplementary Figure 3.22. Pioneer vs. non-pioneer TF effects after conditioning on TF expression and binding. Per-TF effect for pioneer vs. non-pioneer TFs on co-accessibility score in the Xu K562 data set after conditioning on overall TF expression and binding (p-value for difference = 0.034). We partitioned TFs equally into 10 bins by total expression (defined as expression of the focal TF in Xu K562 RNA-seq data) and 10 bins by total binding (defined as number of peaks in the Xu K562 data bound by the focal TF). We performed a linear regression of co-accessibility score on: 1 predictor for the number of pioneer TFs binding in a peak, 1 predictor for the number of non-pioneer TFs binding in a peak, 10 predictors for the number of TFs in each expression bin binding in a peak, and 10 predictors for the number of TFs in each binding bin binding in a peak. Plotted are the effect sizes for pioneer and non-pioneer TFs (conditioned on TF expression and binding information). Confidence intervals denote standard errors.



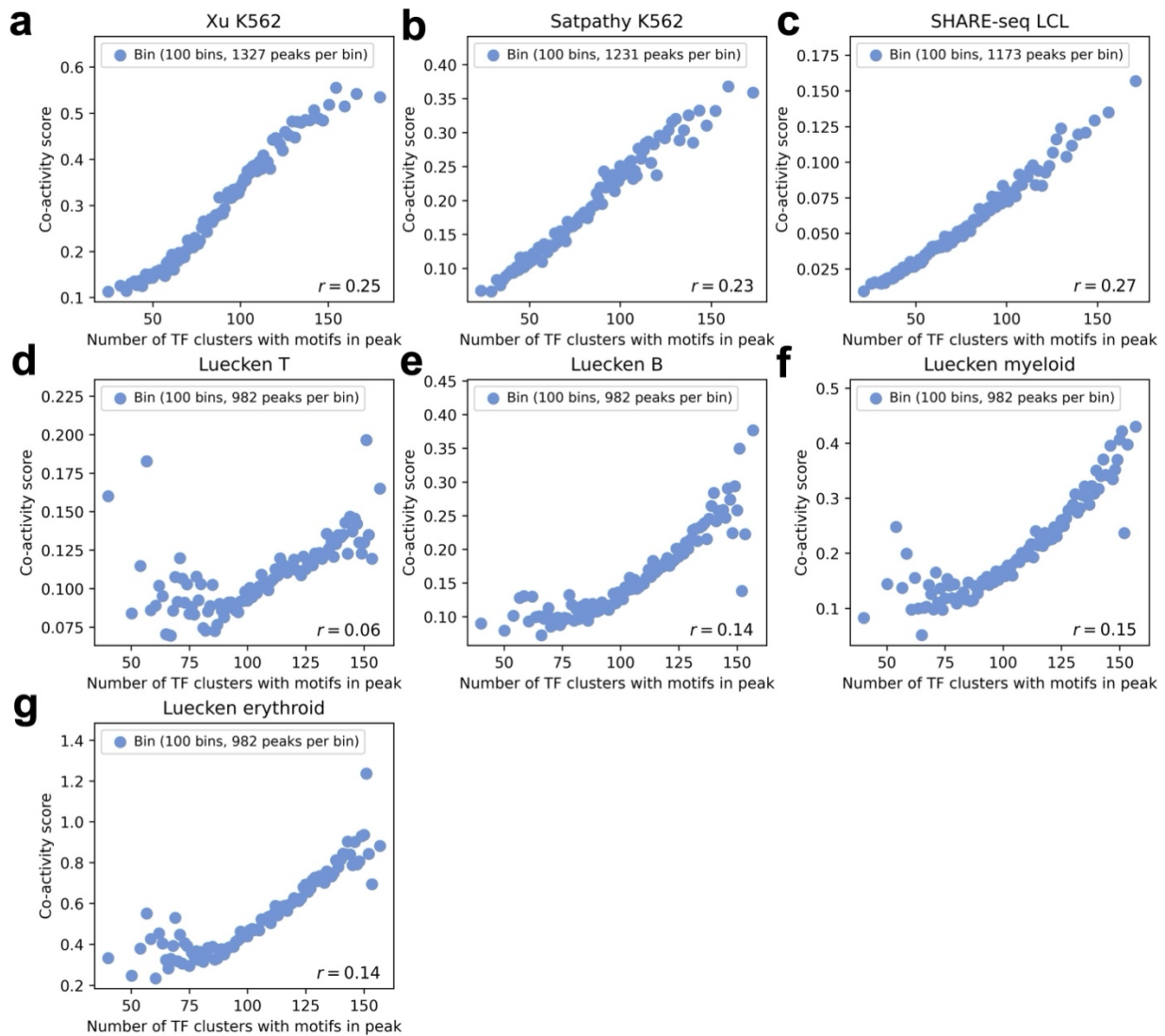
Supplementary Figure 3.23. Pioneer vs. non-pioneer TF effects using Perturb-multiome data to define pioneer TFs. Per-TF effect for pioneer vs. non-pioneer TFs on **a)** co-accessibility score (p-value for difference = $4e-12$), **b)** squared peak-peak correlation (p-value for difference = $8e-7$), **c)** co-activity score (p-value for difference = 0.067), **d)** squared peak-gene correlation (p-value for difference = 0.61). We used Perturb-multiome data from ref.¹³² to define pioneer TFs as TFs whose perturbation is associated with a significant change in peak accessibility or gene expression in scRNA+ATAC-seq multiome data and repeated each bivariate regression comparing pioneer vs. non-pioneer TF effects (right panels of **Figure 3.3a-d**). Confidence intervals denote standard errors. Regression effect sizes are meta-analyzed across 2 data set-cell type pairs most similar to the HSPCs assayed in ref.¹³² (Xu K562 and Satpathy K562). Bars and confidence intervals denote estimates and standard errors, respectively.



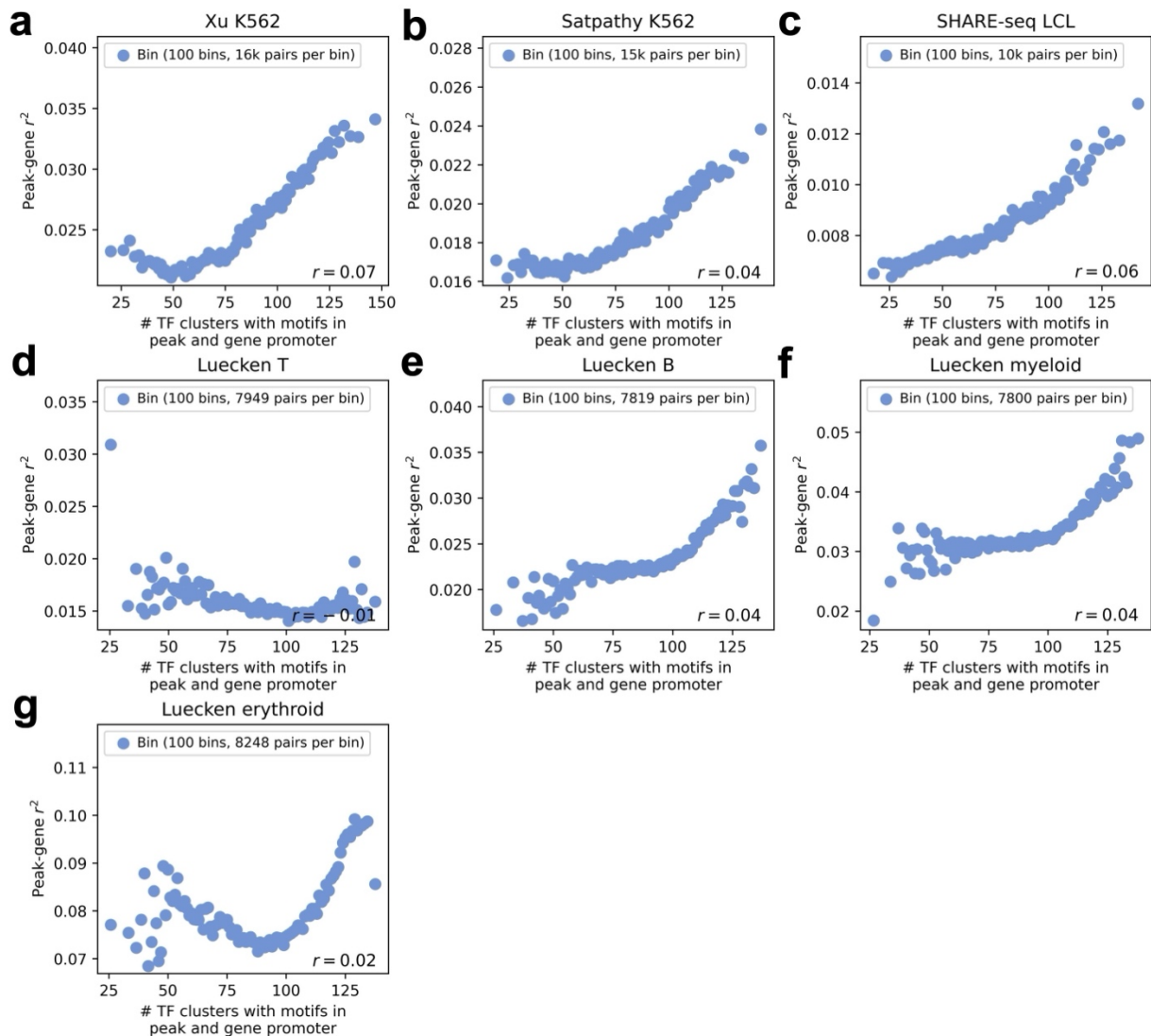
Supplementary Figure 3.24. Relationship between co-accessibility score and transcription factor binding activity as assessed by transcription factor motifs. a) Relationship between number of TF clusters with DNA sequence motifs in an ATAC peak and co-accessibility score across peaks in the a) Xu K562, b) Satpathy K562, c) SHARE-seq LCL, d) Luecken T, e) Luecken B, f) Luecken myeloid, and g) Luecken erythroid data set-cell type pairs. Peaks are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. We obtained genomic coordinates for TF binding motifs (sites with predicted TF binding activity based on DNA sequence) from ref.¹³³, which grouped 2,179 TFs into 286 TF clusters with highly similar binding motifs. For a given TF cluster and data set-cell type pair, we computed the number of TF clusters with motifs in each ATAC peak (analogous to ChIP-seq analyses, using sequence motifs instead of ChIP-seq peaks to define genomic regions of TF activity).



Supplementary Figure 3.25. Relationship between squared peak-peak correlation and transcription factor binding activity as assessed by transcription factor motifs. a) Relationship between number of TF clusters with DNA sequence motifs in both ATAC peaks in a peak-peak pair and peak-peak r^2 , across peak-peak pairs in the a) Xu K562, b) Satpathy K562, c) SHARE-seq LCL, d) Luecken T, e) Luecken B, f) Luecken myeloid, and g) Luecken erythroid data set-cell type pairs. Peak-peak pairs are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. We obtained genomic coordinates for TF binding motifs (sites with predicted TF binding activity based on DNA sequence) from ref.¹³³, which grouped 2,179 TFs into 286 TF clusters with highly similar binding motifs. For a given TF cluster and data set-cell type pair, we computed the number of TF clusters with motifs in each ATAC peak. For each peak-peak pair with peak-peak distance <1Mb, we then computed the number of TF clusters with motifs in both peaks (analogous to ChIP-seq analyses, using sequence motifs instead of ChIP-seq peaks to define genomic regions of TF activity).

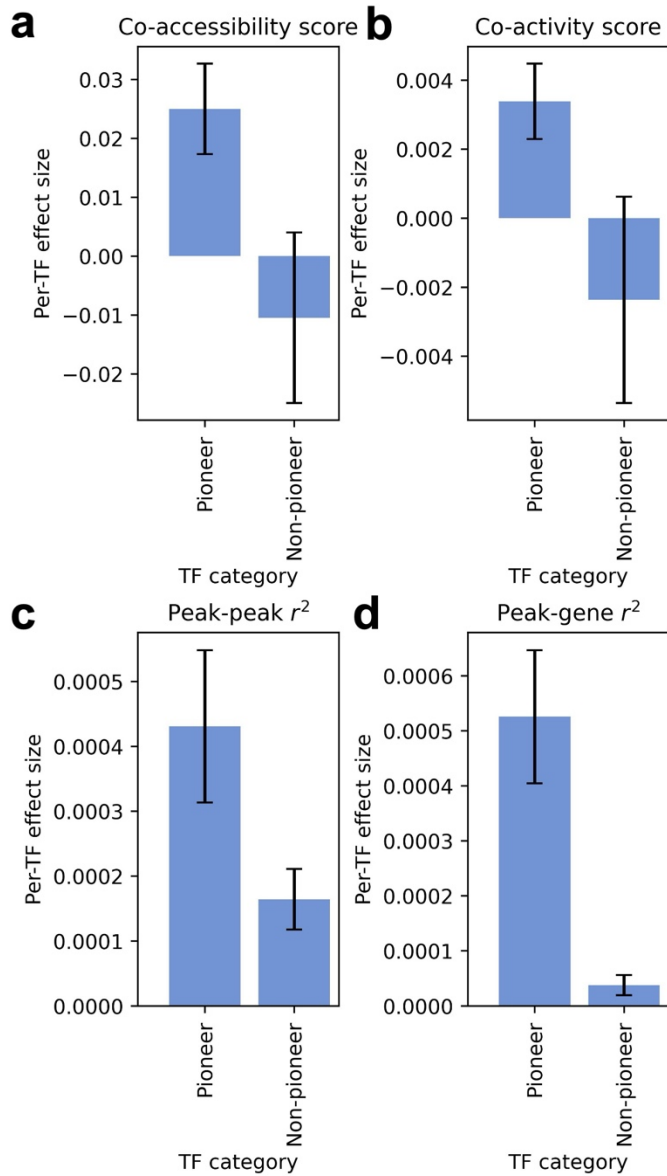


Supplementary Figure 3.26. Relationship between co-activity score and transcription factor binding activity as assessed by transcription factor motifs. a) Relationship between number of TF clusters with DNA sequence motifs in an ATAC peak and co-activity score across peaks in the a) Xu K562, b) Satpathy K562, c) SHARE-seq LCL, d) Luecken T, e) Luecken B, f) Luecken myeloid, and g) Luecken erythroid data set-cell type pairs. Peaks are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. We obtained genomic coordinates for TF binding motifs (sites with predicted TF binding activity based on DNA sequence) from ref.¹³³, which grouped 2,179 TFs into 286 TF clusters with highly similar binding motifs. For a given TF cluster and data set-cell type pair, we computed the number of TF clusters with motifs in each ATAC peak and each gene's promoter (analogous to ChIP-seq analyses, using sequence motifs instead of ChIP-seq peaks to define genomic regions of TF activity).

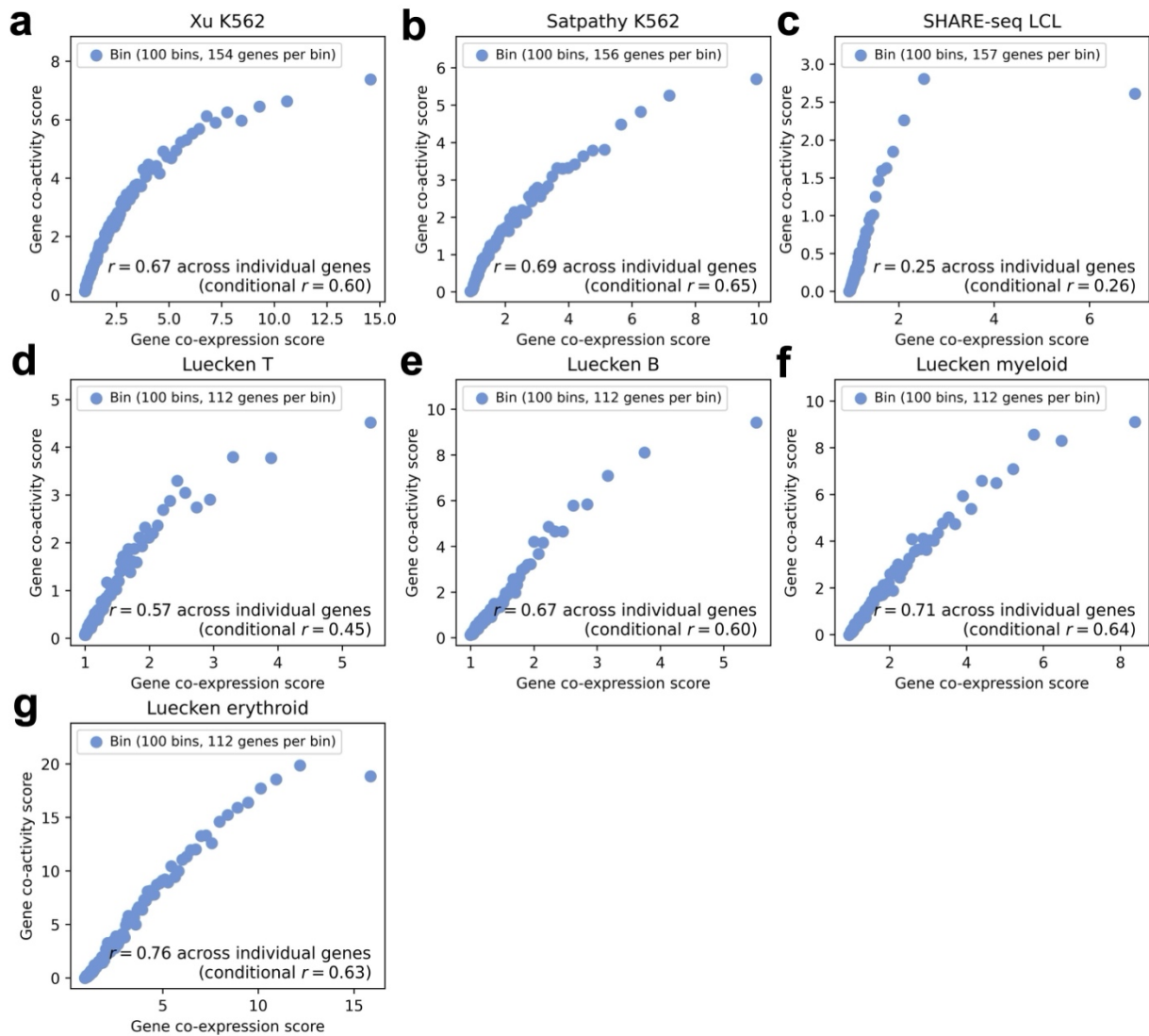


Supplementary Figure 3.27. Relationship between squared peak-gene correlation and transcription factor binding activity as assessed by transcription factor motifs. a)

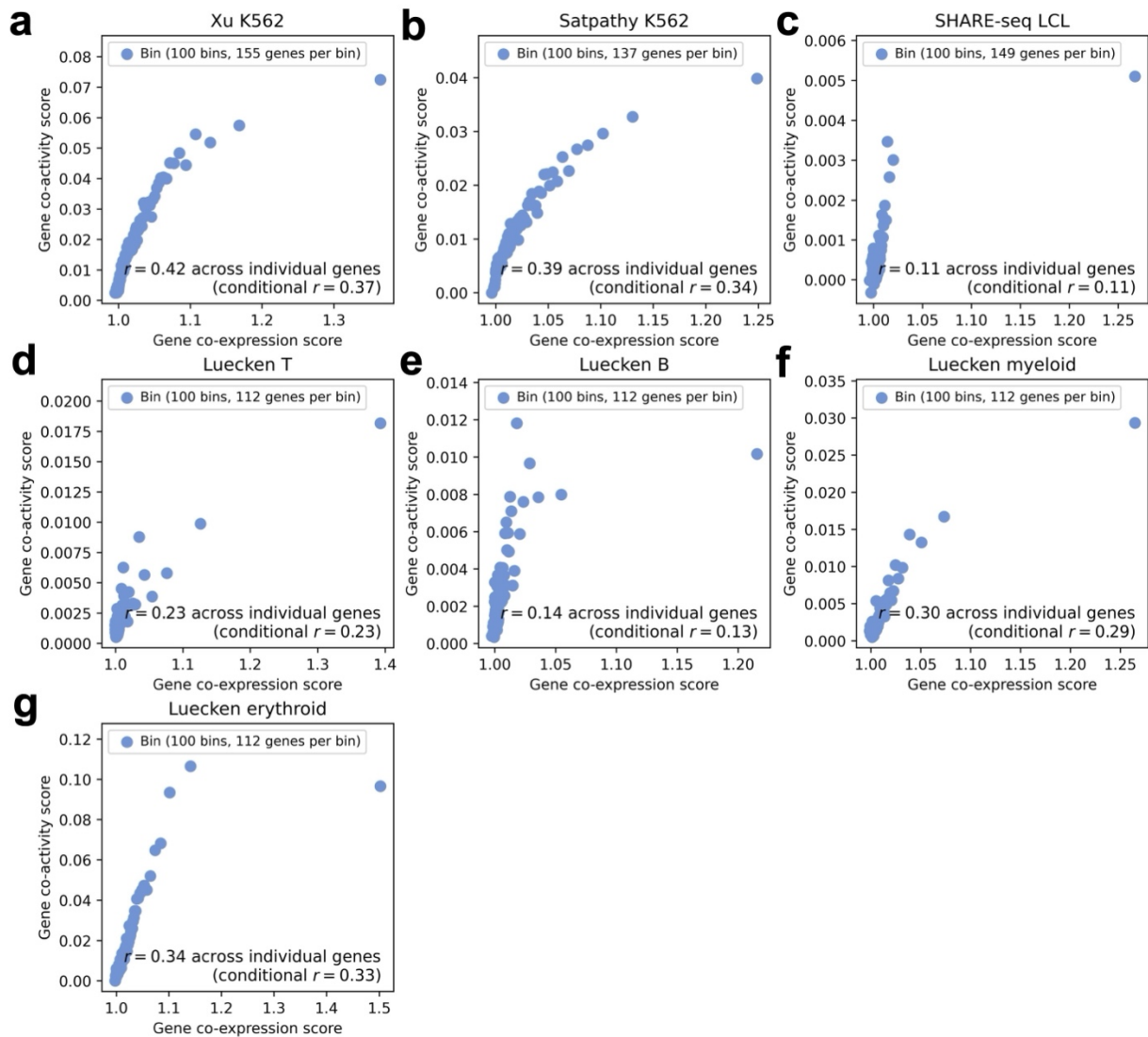
Relationship between number of TF clusters with DNA sequence motifs in both the ATAC peak and the gene's promoter for a peak-gene pair and peak-gene r^2 , across peak-gene pairs in the **a) Xu K562, b) Satpathy K562, c) SHARE-seq LCL, d) Luecken T, e) Luecken B, f) Luecken myeloid, and g) Luecken erythroid** data set-cell type pairs. Peak-gene pairs are partitioned equally into 100 bins (each represented by 1 point) by x-axis value. We obtained genomic coordinates for TF binding motifs (sites with predicted TF binding activity based on DNA sequence) from ref.¹³³, which grouped 2,179 TFs into 286 TF clusters with highly similar binding motifs. For a given TF cluster and data set-cell type pair, we computed the number of TF clusters with motifs in each ATAC peak and each gene's promoter. For each peak-gene pair with peak-gene distance <1Mb, we then computed the number of TF clusters with motifs in both the peak and gene's promoter (analogous to ChIP-seq analyses, using sequence motifs instead of ChIP-seq peaks to define genomic regions of TF activity).



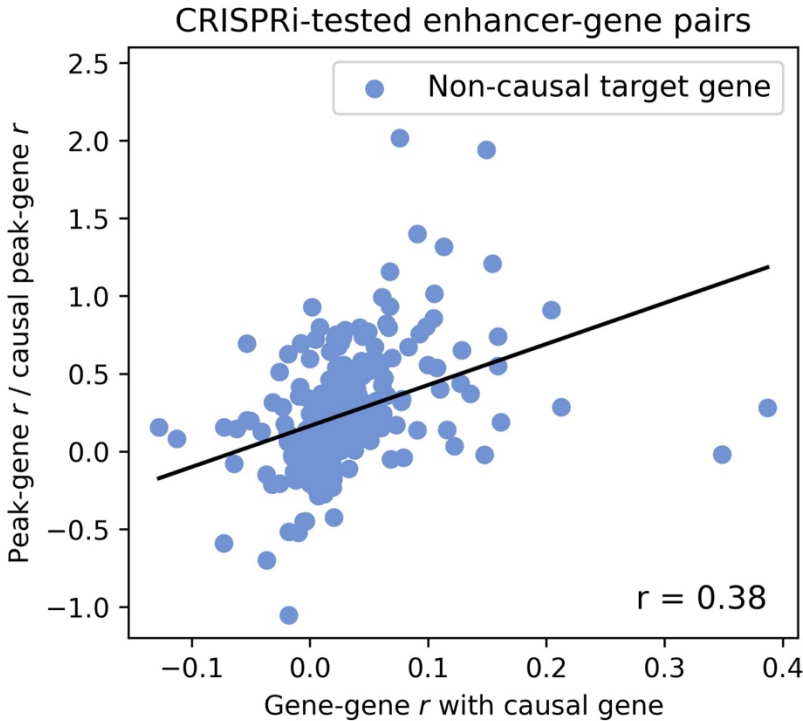
Supplementary Figure 3.28. Pioneer vs. non-pioneer TF effects using transcription factor binding activity as assessed by transcription factor motifs. Per-TF effect for pioneer vs. non-pioneer TF motifs on **a)** co-accessibility score (p-value for difference = $4e-12$), **b)** squared peak-peak correlation (p-value for difference = $8e-7$), **c)** co-activity score (p-value for difference = 0.067), **d)** squared peak-gene correlation (p-value for difference = 0.61). We obtained genomic coordinates for TF binding motifs (sites with predicted TF binding activity based on DNA sequence) from ref.¹³³, which grouped 2,179 TFs into 286 TF clusters with highly similar binding motifs. For a given TF cluster and data set-cell type pair, we computed the number of TF clusters with motifs in each ATAC peak and each gene's promoter. For each peak-gene pair with peak-gene distance <1Mb, we then computed the number of TF clusters with motifs in both the peak and gene's promoter (analogous to ChIP-seq analyses, using sequence motifs instead of ChIP-seq peaks to define genomic regions of TF activity). We defined pioneer TF motif clusters as clusters containing a TF from our main set of pioneer TFs (obtained from refs.^{130,131}). Confidence intervals denote standard errors.



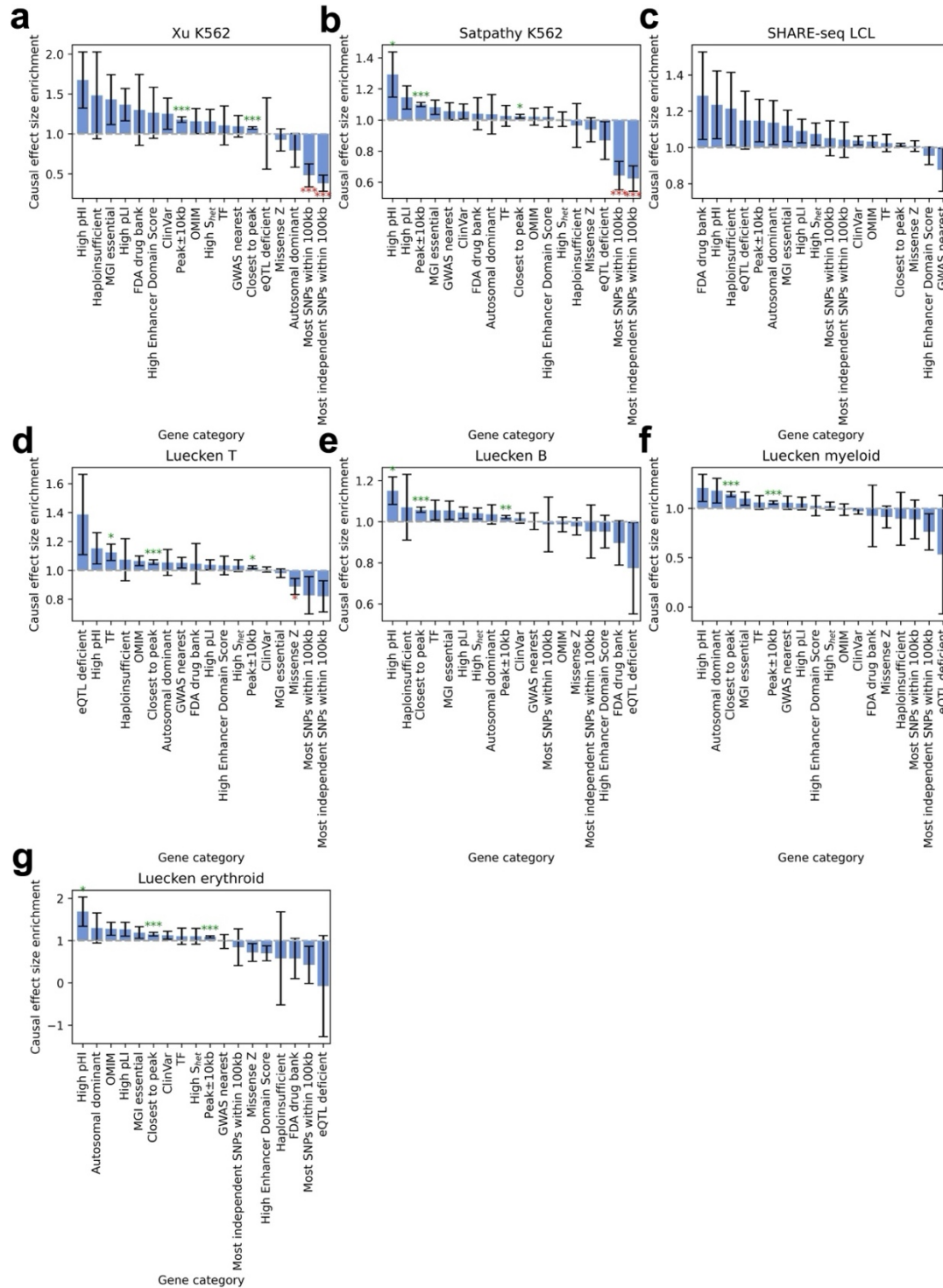
Supplementary Figure 3.29. Relationship between gene co-expression score and gene co-activity score. Relationship between gene co-expression score and gene co-activity score across genes in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal gene. 'Conditional r ' denotes correlation conditioned on number of peaks <1Mb from the focal gene. Genes are partitioned equally into 100 bins (each represented by 1 point) by co-expression score.



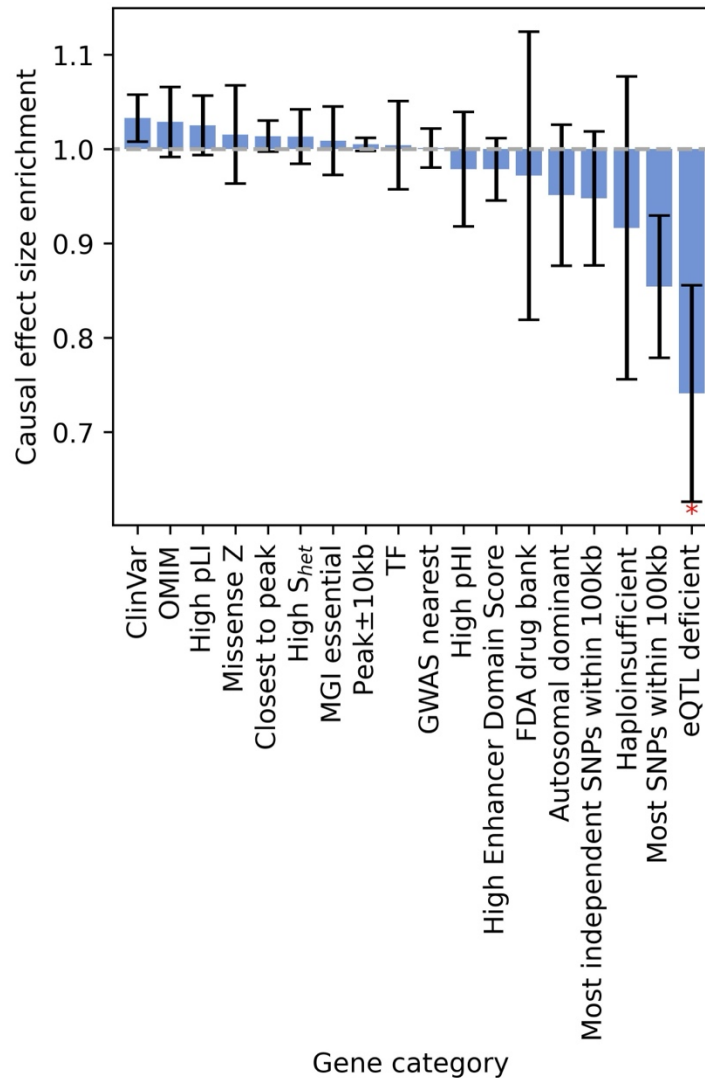
Supplementary Figure 3.30. Relationship between gene co-expression score and gene co-activity score using correlations computed across single cells (instead of metacells). Relationship between gene co-expression score and gene co-activity score across genes in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal gene, with peak-peak and peak-gene correlations computed across single cells (instead of metacells). ‘Conditional r ’ denotes correlation conditioned on number of peaks <1Mb from the focal gene. Genes are partitioned equally into 100 bins (each represented by 1 point) by co-expression score.



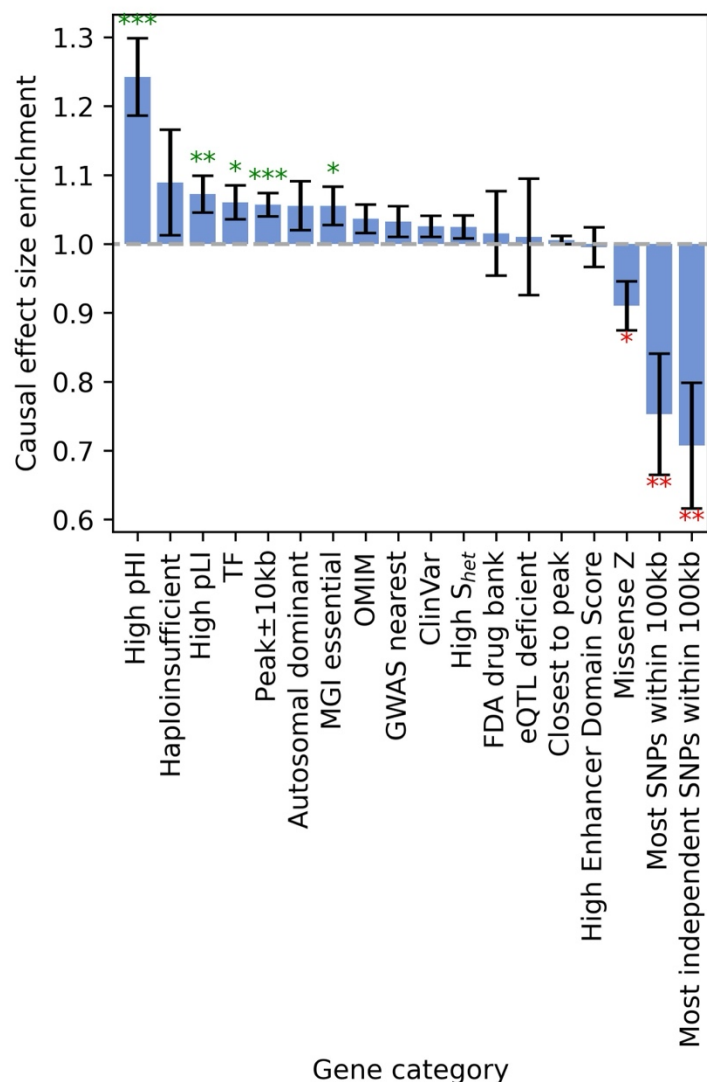
Supplementary Figure 3.31. Relationship between non-causal peak-gene correlation and tagging gene correlation using correlations computed across single cells (instead of metacells). Relationship between correlation with the CRISPR-validated causal target gene (x-axis) and correlation with the causal peak (relative to the causal target gene's correlation with the causal peak) (y-axis) across non-causal target genes tested by CRISPRi, with correlations computed across single cells (instead of metacells). We analyzed CRISPRi-tested peak-gene pairs, restricting to 86 peaks with both an experimentally validated causal (CRISPR-positive) target gene that is identified by Signac (with peak-gene correlation > 0.05 and p-value < 0.05) and at least one experimentally tested non-causal (CRISPR-negative) gene overlapping our single-cell data. We restricted this analysis to the 3 data set-cell type pairs most relevant to K562 cells: Xu K562, Satpathy K562, and Luecken erythroid. We measured the correlation of each CRISPR-negative gene with both the CRISPR-positive gene and the focal peak. Correlation with the CRISPR-positive gene was strongly correlated to correlation with the focal peak.



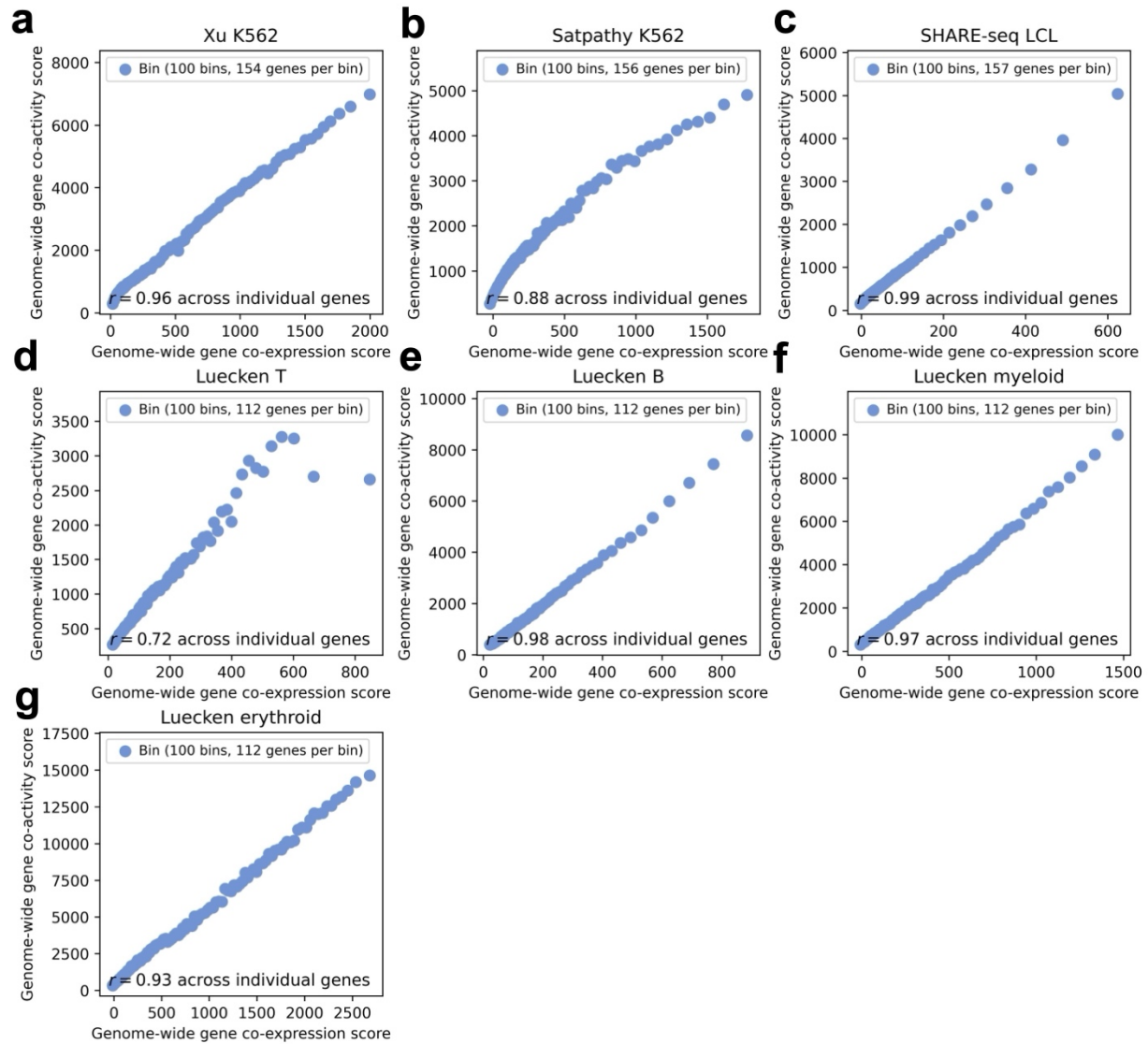
Supplementary Figure 3.32. S-CESC causal effect size enrichments. Causal effect size enrichment of 18 gene categories in stratified co-expression score regression in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



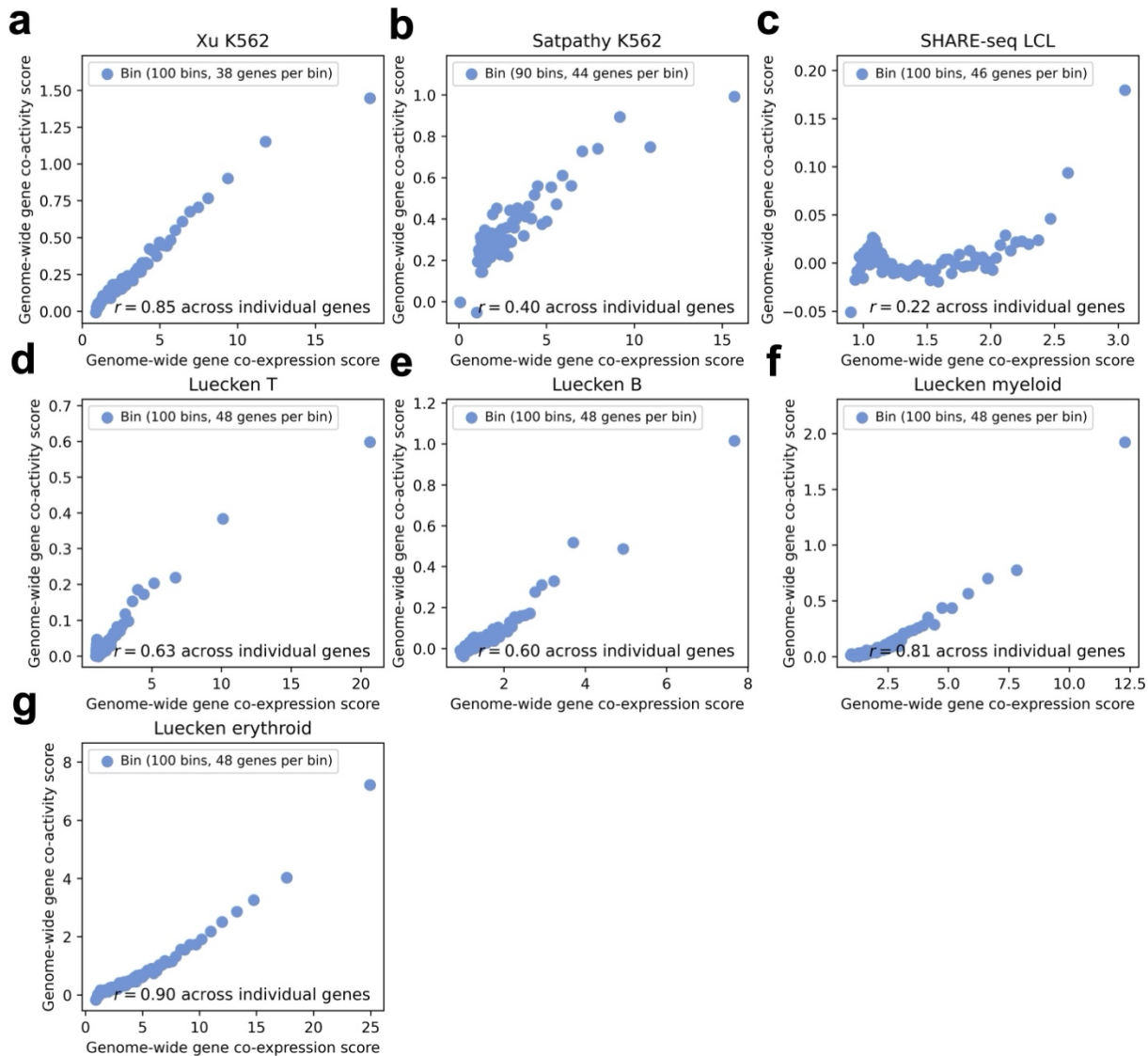
Supplementary Figure 3.33. S-CESC causal effect size enrichments computed using correlations across single cells (instead of metacells). Causal effect size enrichment of 18 gene categories in stratified co-expression score regression, implemented using gene-gene and peak-gene correlations computed across single cells (instead of metacells) and meta-analyzed across all 7 data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



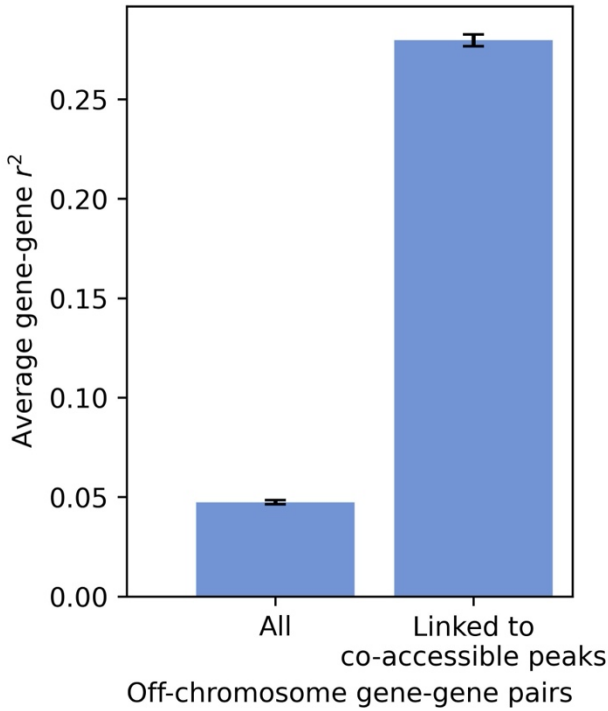
Supplementary Figure 3.34. S-CESC causal effect size enrichments without conditioning on number of nearby peaks. Causal effect size enrichment of 18 peak categories in stratified co-expression score regression across genes, implemented without the covariate for the number of peaks <1Mb from each gene and meta-analyzed across all 7 data set-cell type pairs. Results are generally similar to **Figure 3.4c**. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Stars denote meta-analyzed p-values for significant enrichment (green) or depletion (red) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).



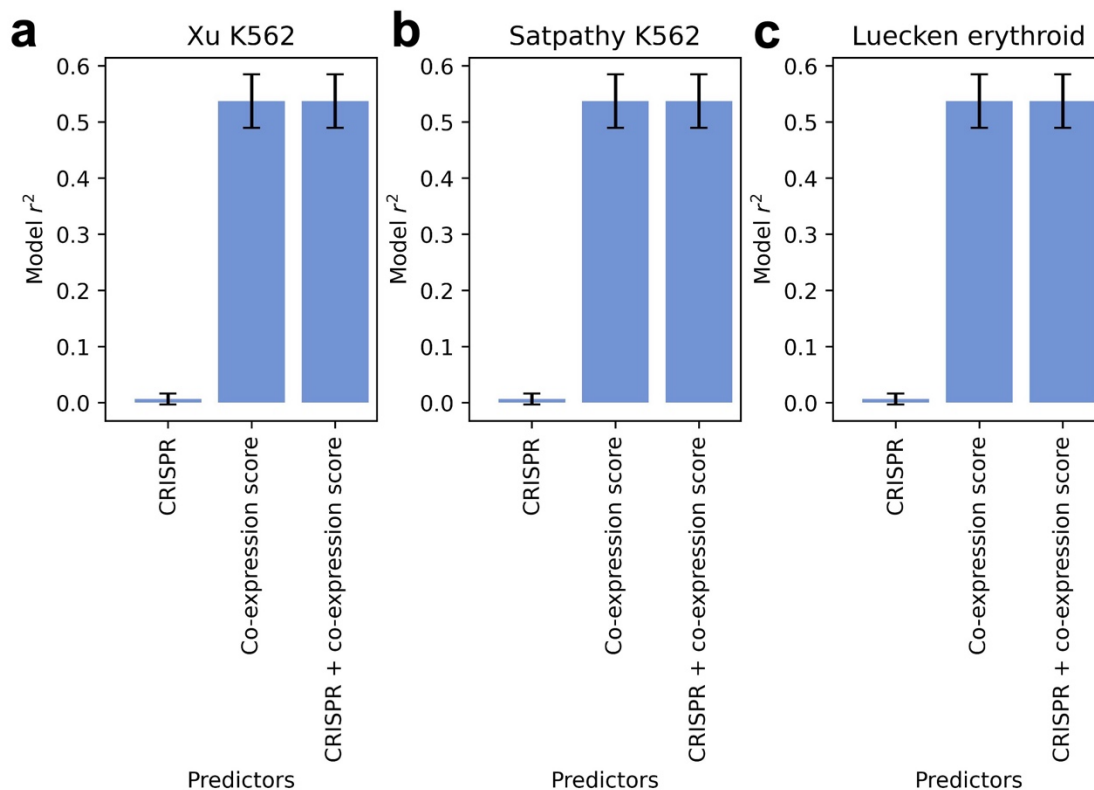
Supplementary Figure 3.35. Relationship between genome-wide gene co-expression score and genome-wide gene co-activity score. Relationship between genome-wide gene co-expression score and genome-wide gene co-activity score across peaks in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal gene. 'Conditional r ' denotes correlation conditioned on number of peaks <1Mb from the focal gene. Genes are partitioned equally into 100 bins (each represented by 1 point) by co-expression score.



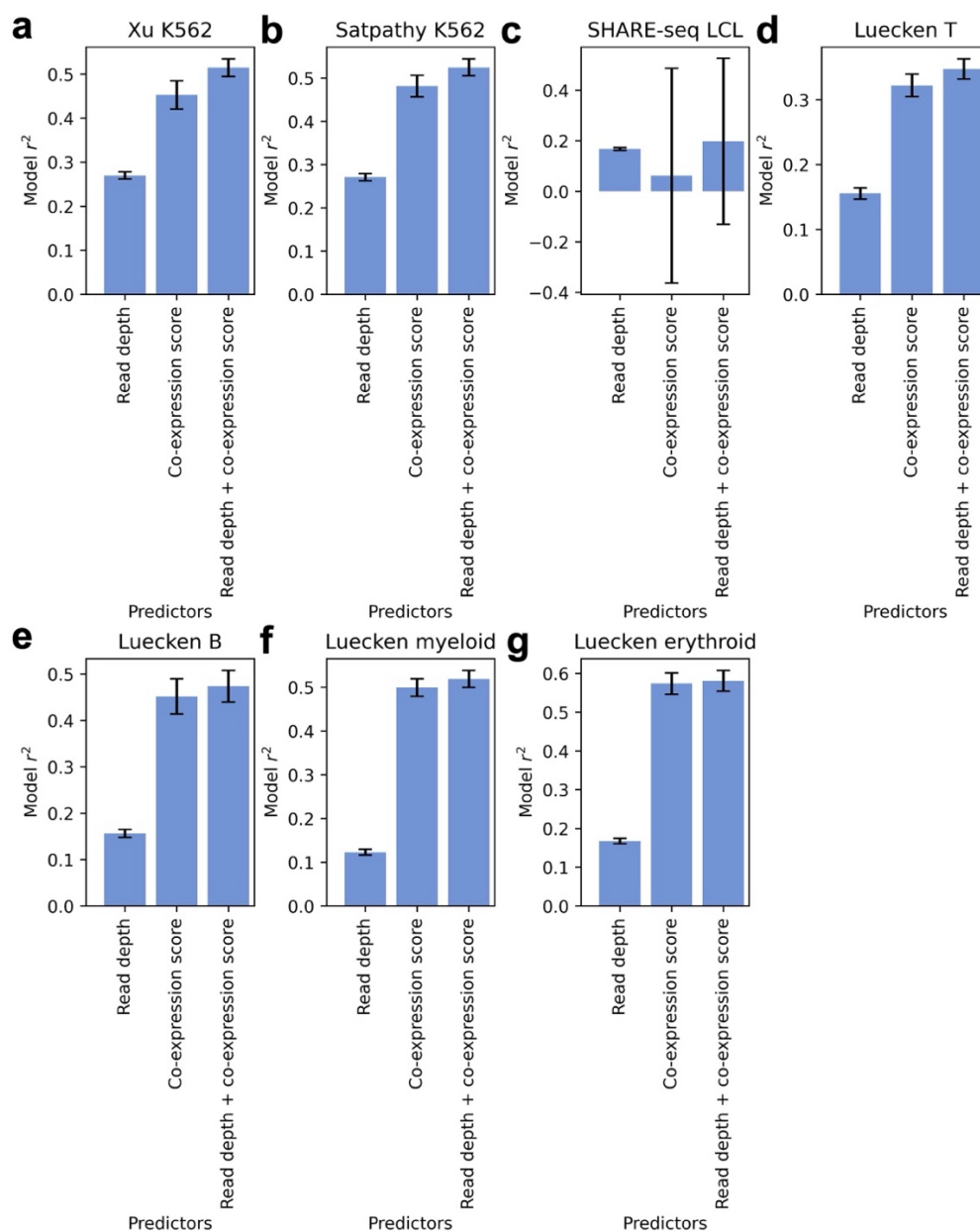
Supplementary Figure 3.36. Relationship between genome-wide gene co-expression score and genome-wide gene co-activity score computed using correlations across single cells (instead of metacells). Relationship between genome-wide gene co-expression score and genome-wide gene co-activity score (using a subset of 5000 peaks and 5000 genes) across peaks in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed across single cells (instead of metacells) using all peaks and genes genome-wide. We downsampled to 5000 randomly selected peaks and 5000 randomly selected genes when computing scores (due to the high computational cost of computing correlations for every peak-peak, gene-gene, and peak-gene pair genome-wide across single cells). Genes are partitioned equally into 100 bins (each represented by 1 point) by co-expression score.



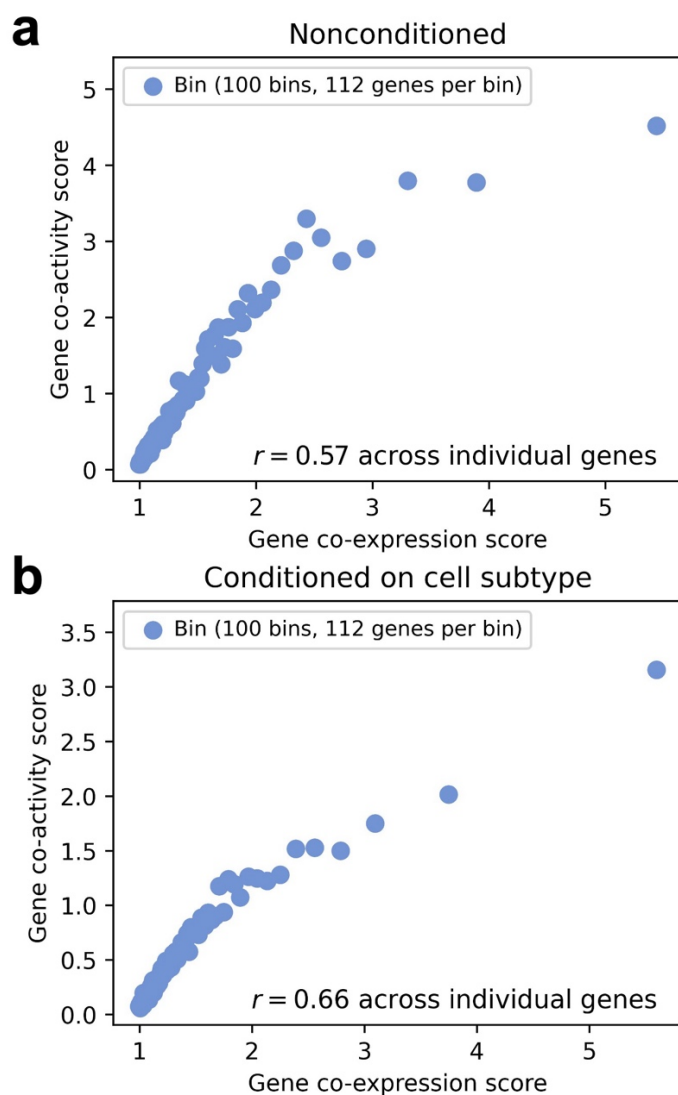
Supplementary Figure 3.37. Average squared gene-gene correlation for off-chromosome pairs of genes linked to co-accessible peaks vs. all genes. We identified 2,108 proximal peak-gene links in the Xu K562 data set (defining “proximal” as peak-gene distance <100kb and “links” as ArchR linking score (r^2) > 0.2025, corresponding to ArchR threshold r > 0.45). From these peak-gene links, we identified 644,981 highly co-accessible pairs of off-chromosome linked peaks (involving a peak from chromosome 1 and a peak from a different chromosome, defining “highly co-accessible” as peak-peak r^2 > 0.2025). We then identified 124,714 pairs of genes linked to peaks forming an off-chromosome co-accessible peak-peak pair (such that each gene-gene pair also has a gene on chromosome 1 and a gene on a different chromosome). We then computed the average gene-gene r^2 across these gene-gene pairs (right) vs. all gene-gene pairs formed from the set of genes involved in a proximal peak-gene link (left). Average gene-gene r^2 for pairs of genes linked to highly co-accessible pairs of peaks was substantially higher (r^2 = 0.28) than for all gene-gene pairs (r^2 = 0.047; p-value for difference = 4e-15). Confidence intervals denote standard errors.



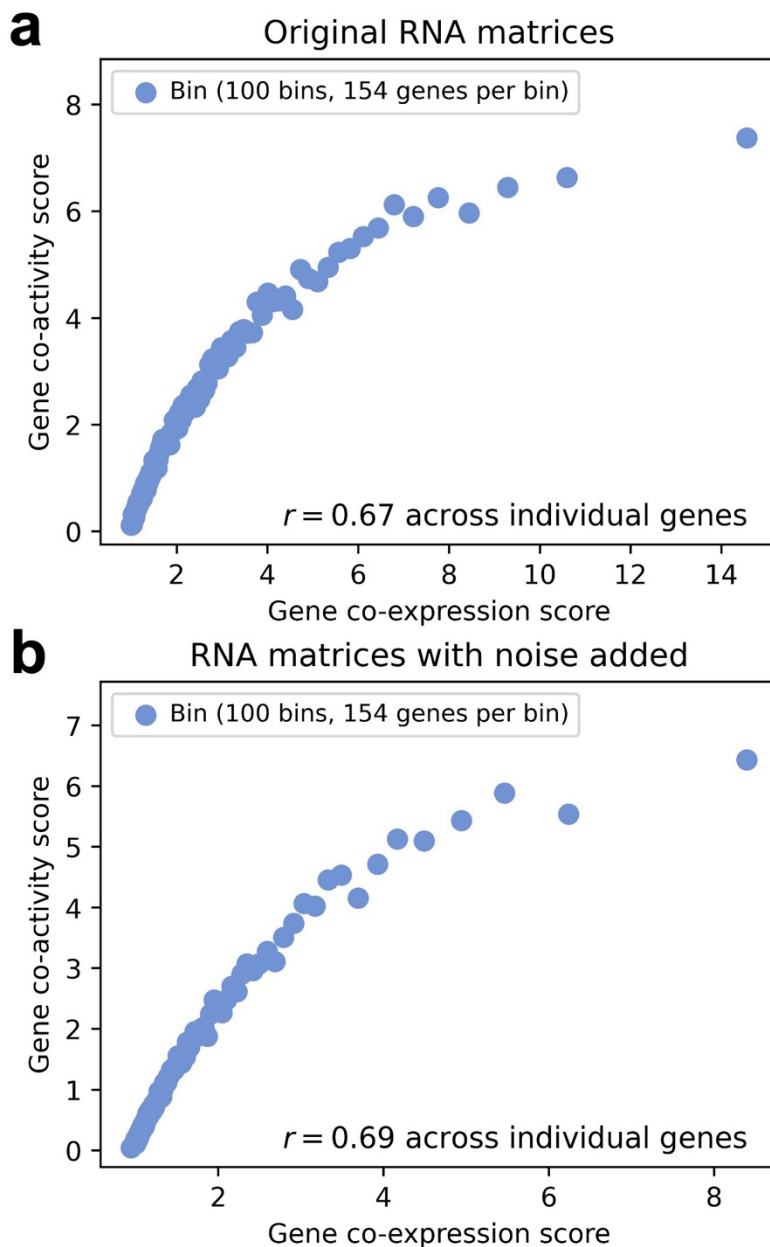
Supplementary Figure 3.38. Regression of gene co-activity score on CRISPR status and gene co-expression score. Model r^2 for linear regressions of gene co-activity score on CRISPR-positive status (“CRISPR”), gene co-expression score, or both, for the **a)** Xu K562, **b)** Satpathy K562, and **c)** Luecken erythroid data set-cell type pairs. Genes in more “active” regulatory regions (higher gene co-expression scores) may be more likely to be true causal target genes. To assess the role of biological causality in driving the correlation between gene co-activity score and gene co-expression score, we performed a bivariate linear regression of gene co-activity score on gene co-expression score and a binary indicator of CRISPR-positive status (1 if a gene has at least one causal peak validated by CRISPR, 0 otherwise) across genes. We also performed univariate regressions of gene co-activity score on gene co-expression score and CRISPR-positive status individually. We restricted this analysis to the 3 data set-cell type pairs most relevant to K562 cells: Xu-K562, Satpathy-K562, and Luecken-erythroid. Within each data set-cell type pair, we further restricted to peaks with ≥ 1 causal target gene validated by CRISPR (CRISPR-positive) and ≥ 1 CRISPR-tested non-target (CRISPR-negative) gene. Confidence intervals denote standard errors.



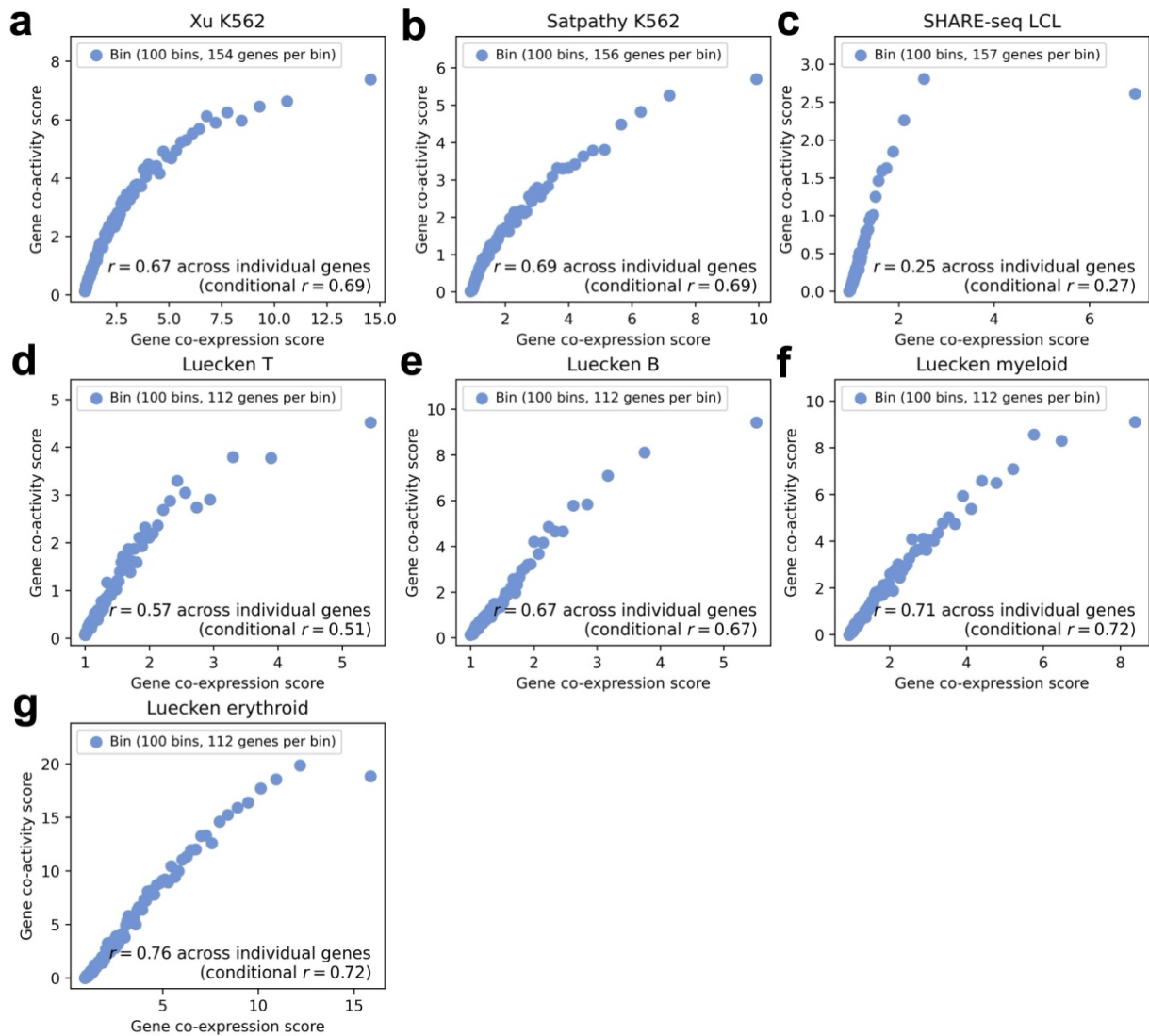
Supplementary Figure 3.39. Regression of gene co-activity score on gene read depth and gene co-expression score. Model r^2 for linear regressions of gene co-activity score on gene read depth, gene co-expression score, or both in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs. Genes with higher read depth may have higher measured correlations with other peaks and genes, generating higher gene co-expression scores and gene co-activity scores. To assess the role of gene read depth in driving the correlation between gene co-activity score and gene co-expression scores, we performed a bivariate linear regression of gene co-activity score on gene co-expression score and log(read depth) across genes. We also performed univariate regressions of gene co-activity score on gene co-expression scores and read depth individually. Confidence intervals denote standard errors. (We note that log-transformed read depth attained higher r^2 than raw read depth.)



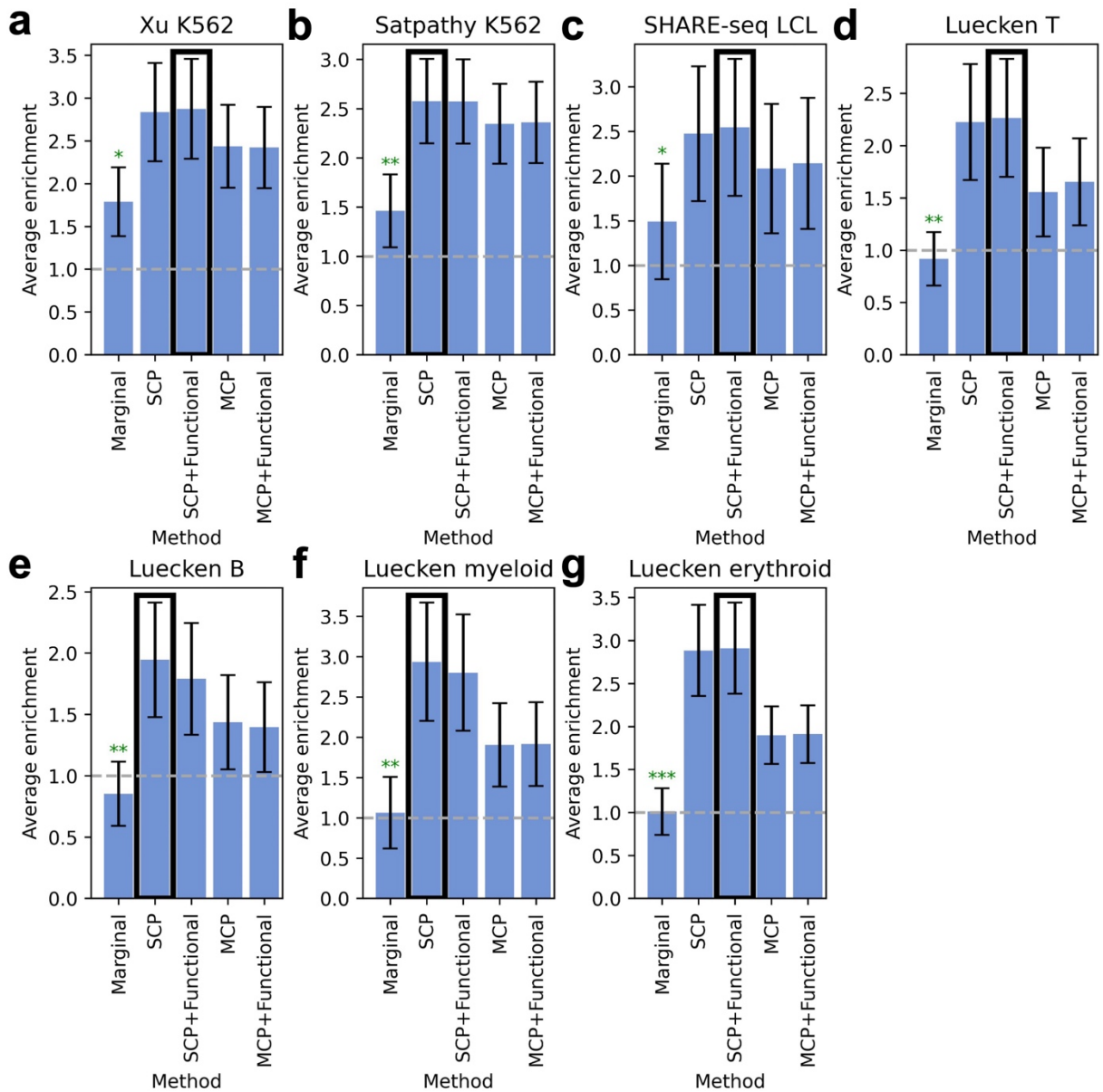
Supplementary Figure 3.40. Relationship between gene co-expression score and gene co-activity score conditioned on cell subtype. Relationship between gene co-expression score and gene co-activity score across peaks in the Luecken T cell data set computed using **a)** unconditioned peak-peak and peak-gene correlations and **b)** gene-gene and peak-gene correlations conditioned on cell subtype proportions. Cell subtypes may be associated with varying levels of average peak accessibility and gene expression, driving a correlation between gene co-expression score and gene co-activity score. To assess the role of cell subtypes, we analyzed T cells from the Luecken BMMC data set, which span 4 labeled T cell subtypes (activated CD4+, naive CD4+, CD8+, and naive CD8+; each represented by 1-12k cells). We separately residualized gene expression and ATAC peak accessibility by cell type proportions in each metacell (via linear regression across metacells), recomputed gene-gene correlations and peak-gene correlations using residualized data, and recomputed gene co-expression scores and gene co-activity scores by summing these conditional correlations. The correlation between gene co-activity score and gene co-expression score did not decrease after accounting for cell subtypes, suggesting that major cell subtypes do not fully explain the correlation. Genes are partitioned equally into 100 bins (each represented by 1 point) by gene co-expression score.



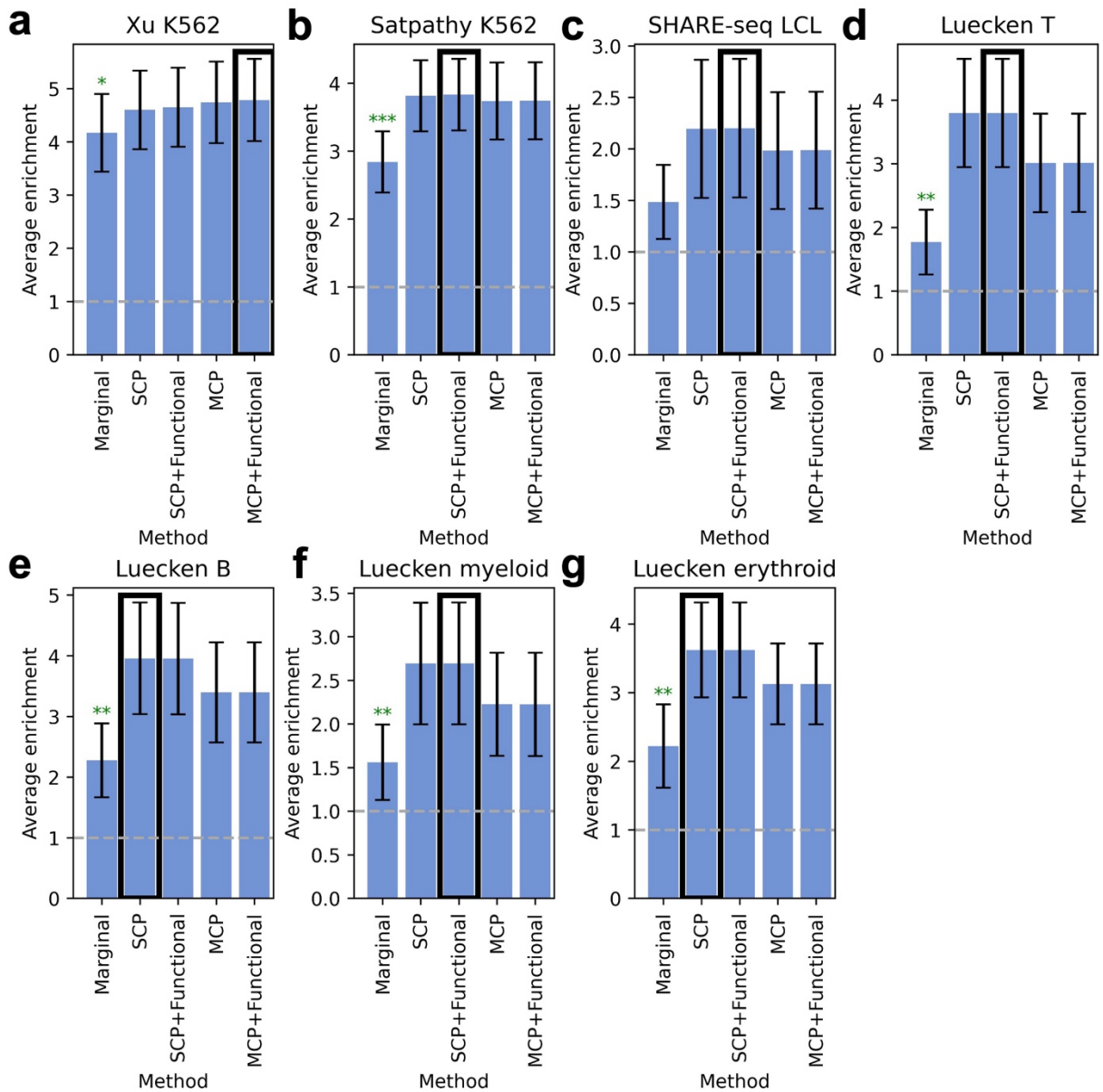
Supplementary Figure 3.41. Relationship between gene co-expression score and gene co-activity score with RNA measurement noise added. Relationship between gene co-expression score and gene co-activity score across peaks in the Xu K562 data set computed using **a)** the original RNA matrix and **b)** an RNA matrix with added measurement noise. Genes with higher amounts of measurement noise may have lower measured correlations with other peaks and genes, possibly driving the correlation between gene co-activity score and gene co-expression score. Because it is difficult to directly assess the amount of measurement noise in gene expression data, we introduced additional artificial noise to 50% of genes (by randomly setting 20% of entries equal to 0, retaining the original data for the remaining 50% of genes), and recomputed gene co-expression and gene co-activity scores.



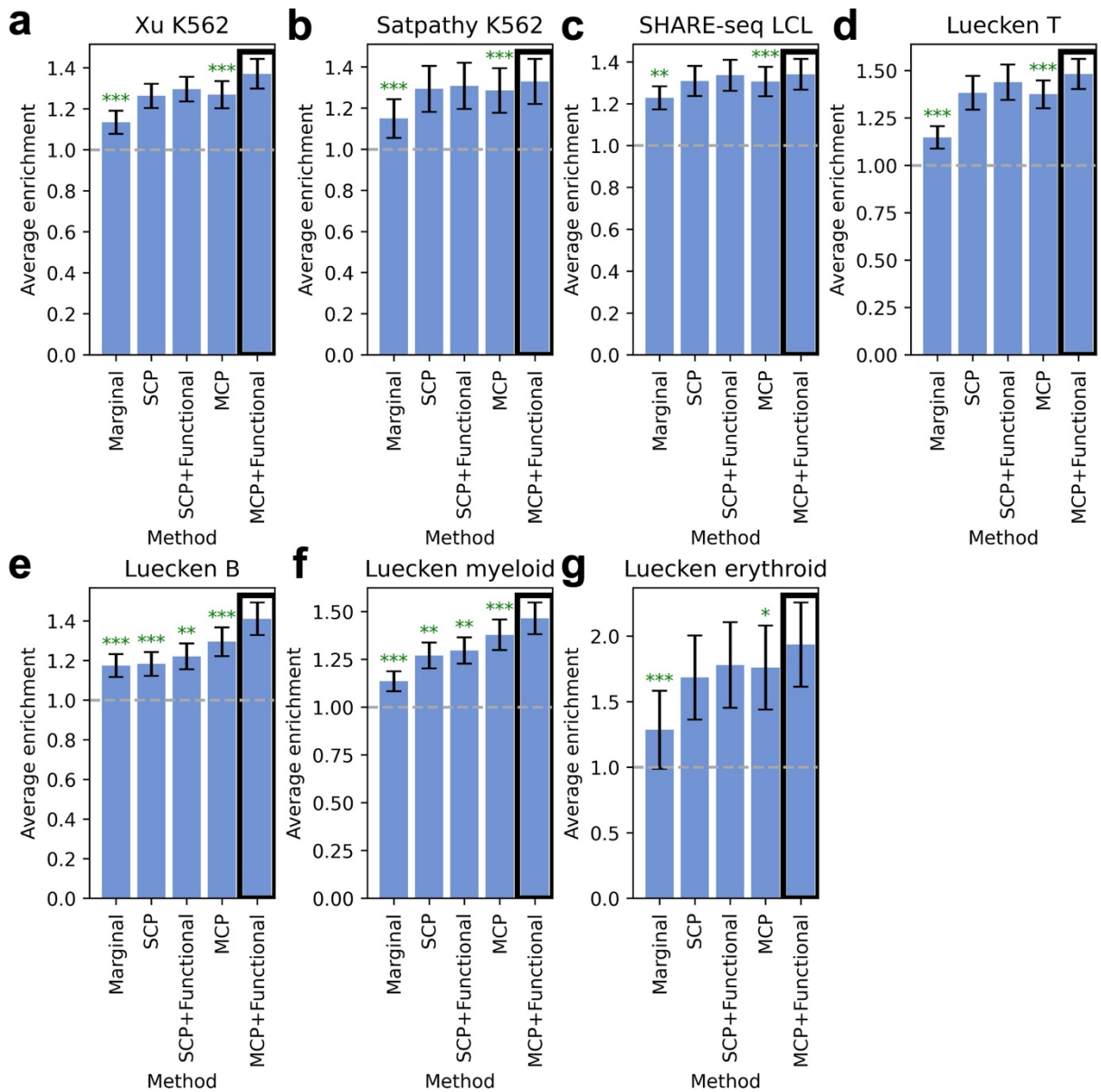
Supplementary Figure 3.42. Relationship between gene co-expression score and gene co-activity score conditioned on number of nearby peaks and genes. Relationship between gene co-expression score and gene co-activity score across genes in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs, computed using peaks and genes within the *cis* window (<1Mb) of each focal gene. 'Conditional r ' denotes correlation conditioned on both the number of peaks <1Mb from the focal gene and the number of genes <1Mb from the focal gene (instead of only the number of peaks <1Mb from the focal gene). Genes are partitioned equally into 100 bins (each represented by 1 point) by gene co-expression score.



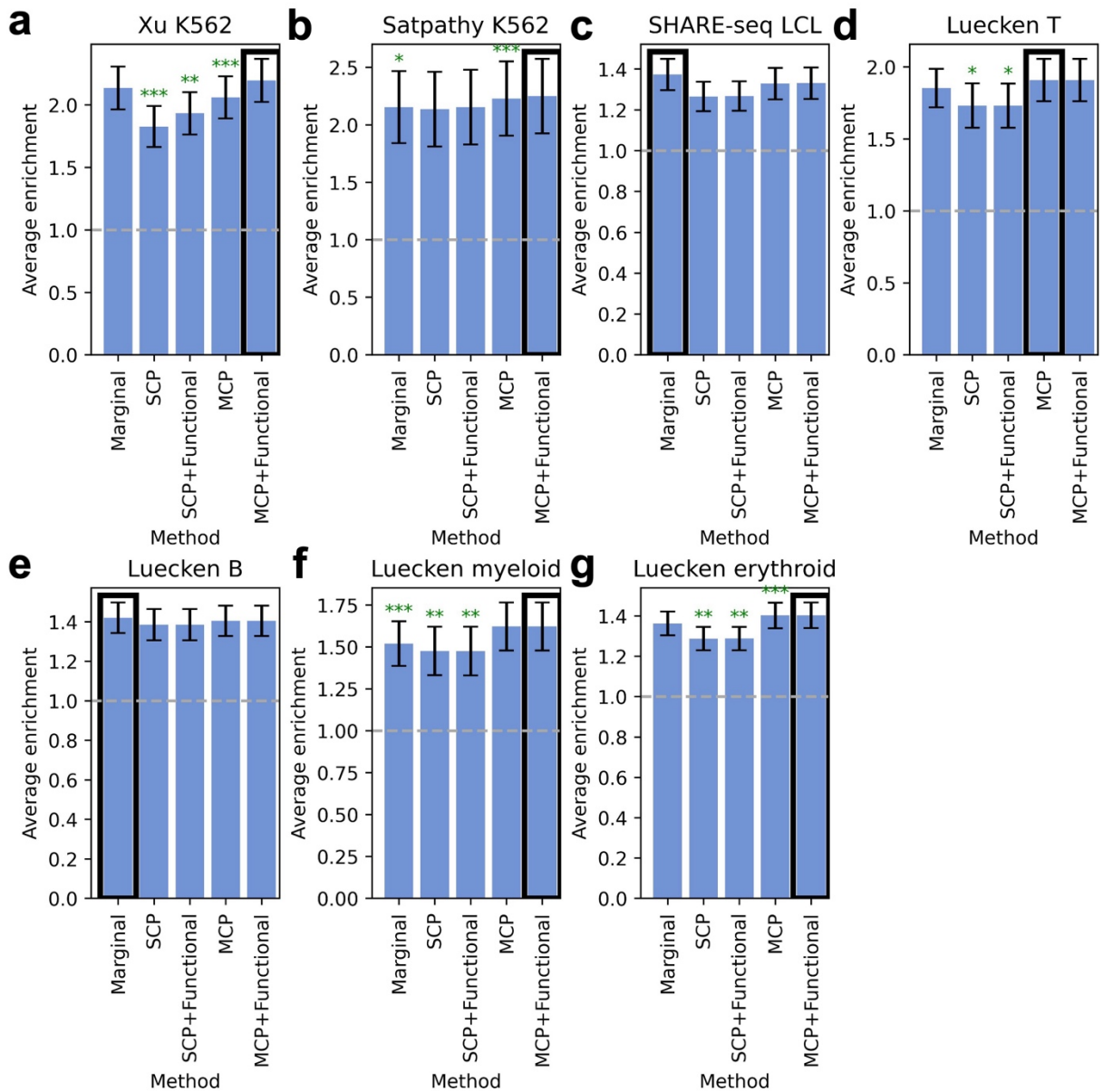
Supplementary Figure 3.43. Evaluation of ArchR vs. fine-mapped ArchR scores on CRISPR data. Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores for 448 links validated by CRISPR in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping; SCP + Functional, single causal peak functionally informed fine-mapping; MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.



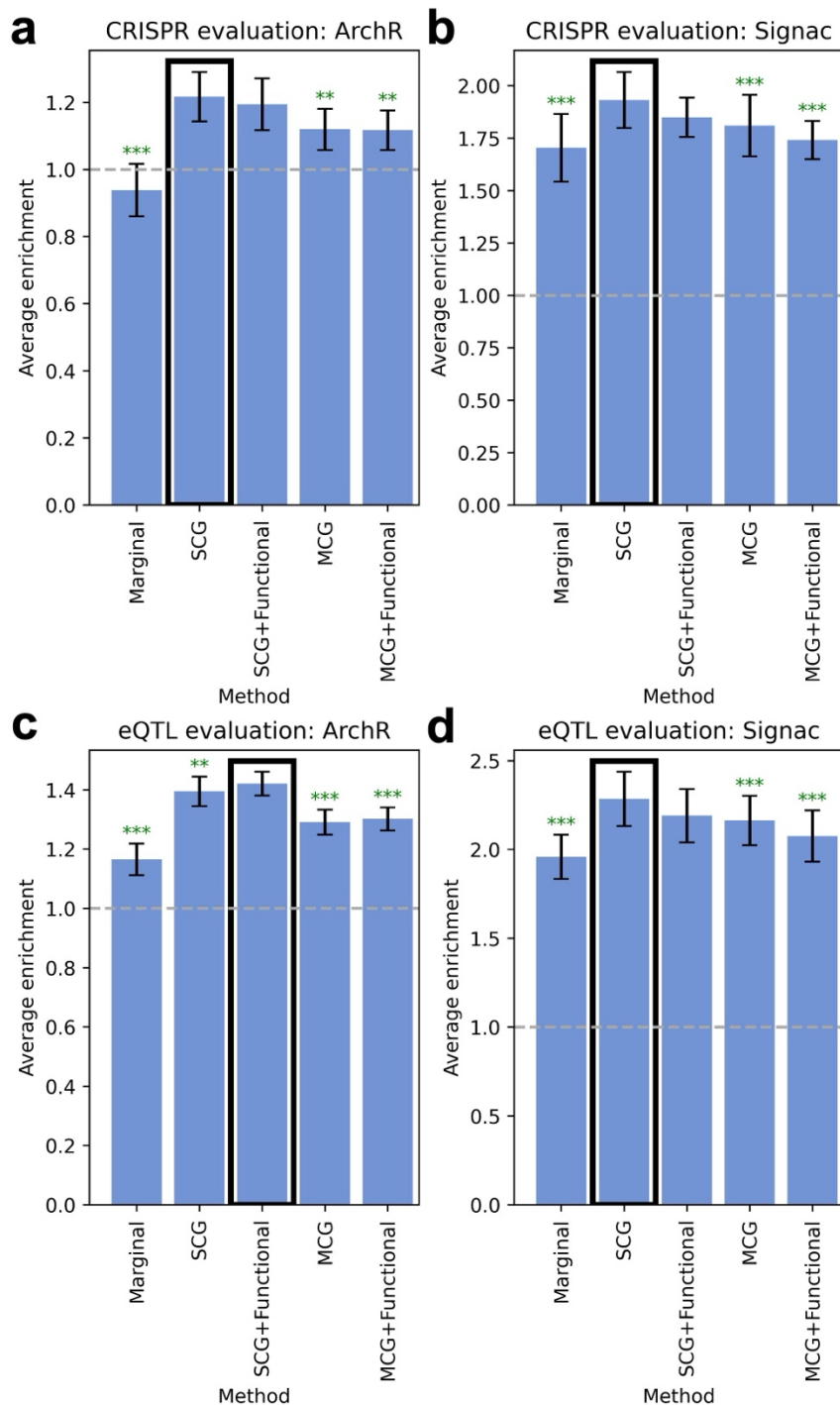
Supplementary Figure 3.44. Evaluation of Signac vs. fine-mapped Signac scores on CRISPR data. Average enrichment across recall values of links predicted by Signac vs. fine-mapped Signac scores for 448 links validated by CRISPR in the **a)** Xu K562, **b)** Satpathy K562, **c)** SHARE-seq LCL, **d)** Luecken T, **e)** Luecken B, **f)** Luecken myeloid, and **g)** Luecken erythroid data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping; SCP + Functional, single causal peak functionally informed fine-mapping; MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.



Supplementary Figure 3.45. Evaluation of ArchR vs. fine-mapped ArchR scores on eQTL data. Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping; SCP + Functional, single causal peak functionally informed fine-mapping; MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.

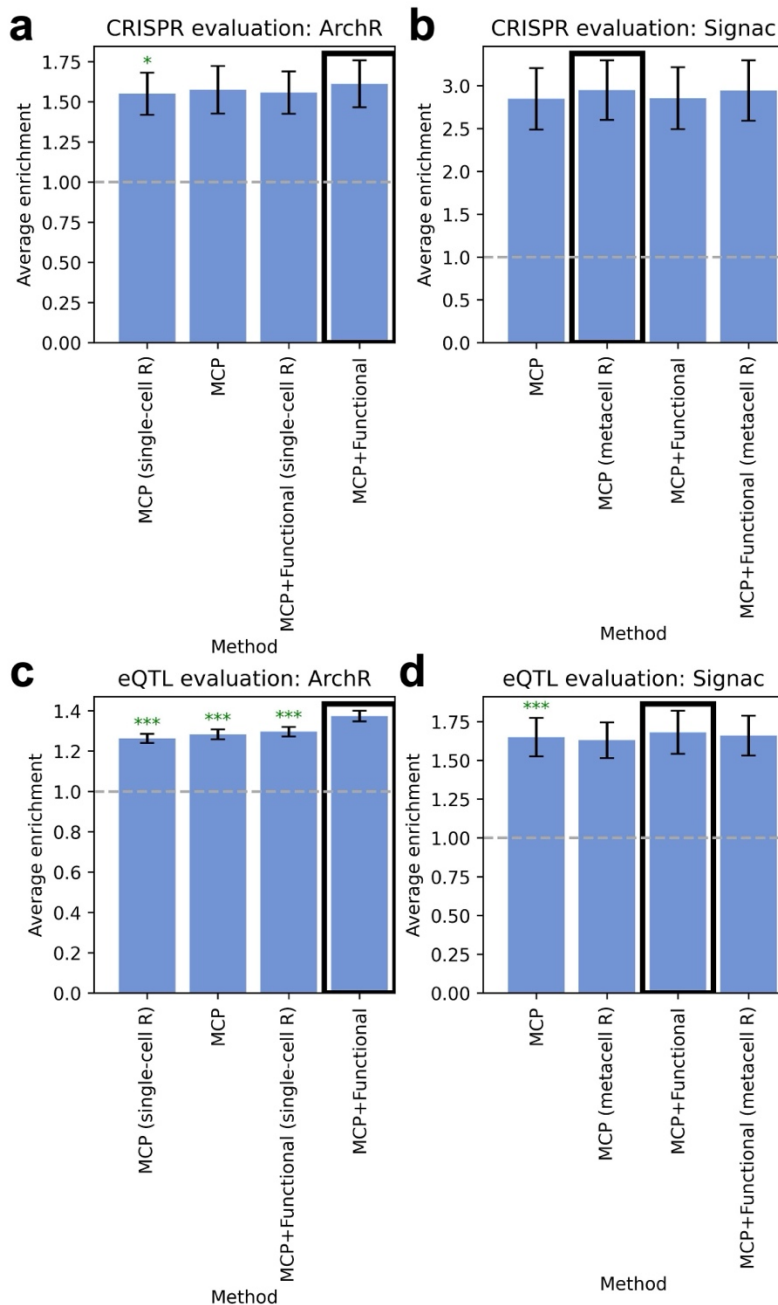


Supplementary Figure 3.46. Evaluation of Signac vs. fine-mapped Signac scores on eQTL data. Average enrichment across recall values of links predicted by Signac vs. fine-mapped Signac scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues in the **a**) Xu K562, **b**) Satpathy K562, **c**) SHARE-seq LCL, **d**) Luecken T, **e**) Luecken B, **f**) Luecken myeloid, and **g**) Luecken erythroid data set-cell type pairs. Bars and confidence intervals denote estimates and standard errors, respectively. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping; SCP + Functional, single causal peak functionally informed fine-mapping; MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.



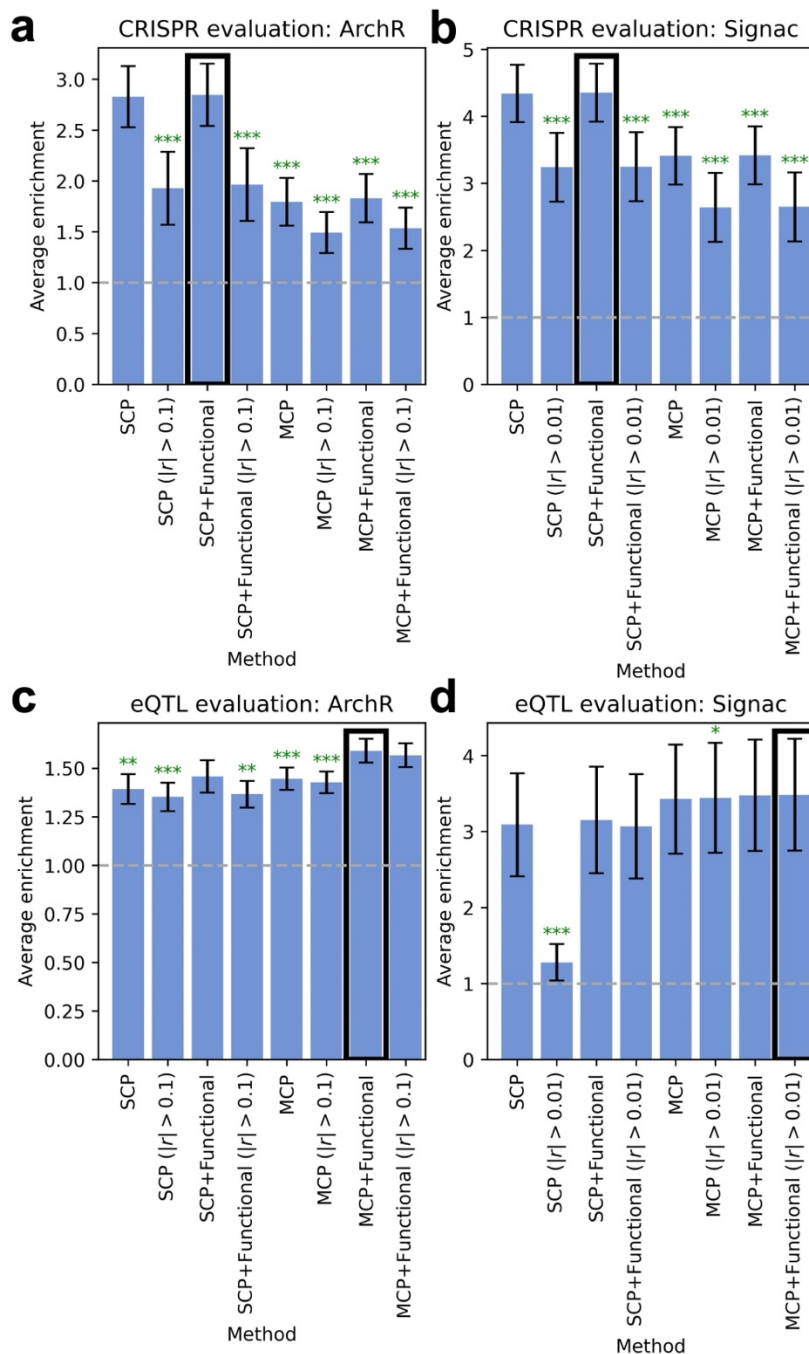
Supplementary Figure 3.47. Evaluations of gene-level fine-mapping on CRISPR and eQTL data. **a)** Average enrichment across recall values of links predicted by ArchR vs. gene-level fine-mapped ArchR scores for 448 links validated by CRISPR. **b)** Average enrichment across recall values of links predicted by Signac vs. gene-level fine-mapped Signac scores for 448 links validated by CRISPR. **c)** Average enrichment across recall values of links predicted by ArchR vs. gene-level fine-mapped ArchR scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. **d)** Average enrichment across recall values of links predicted by Signac vs. gene-level fine-mapped Signac scores for 39,194 fine-mapped eSNP–

eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. Details about gene-level fine-mapping are provided in the **Methods** section. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCG, single causal gene fine-mapping; SCG + Functional, single causal gene functionally informed fine-mapping; MCG, multiple causal gene fine-mapping; MCG + Functional, multiple causal gene functionally informed fine-mapping.



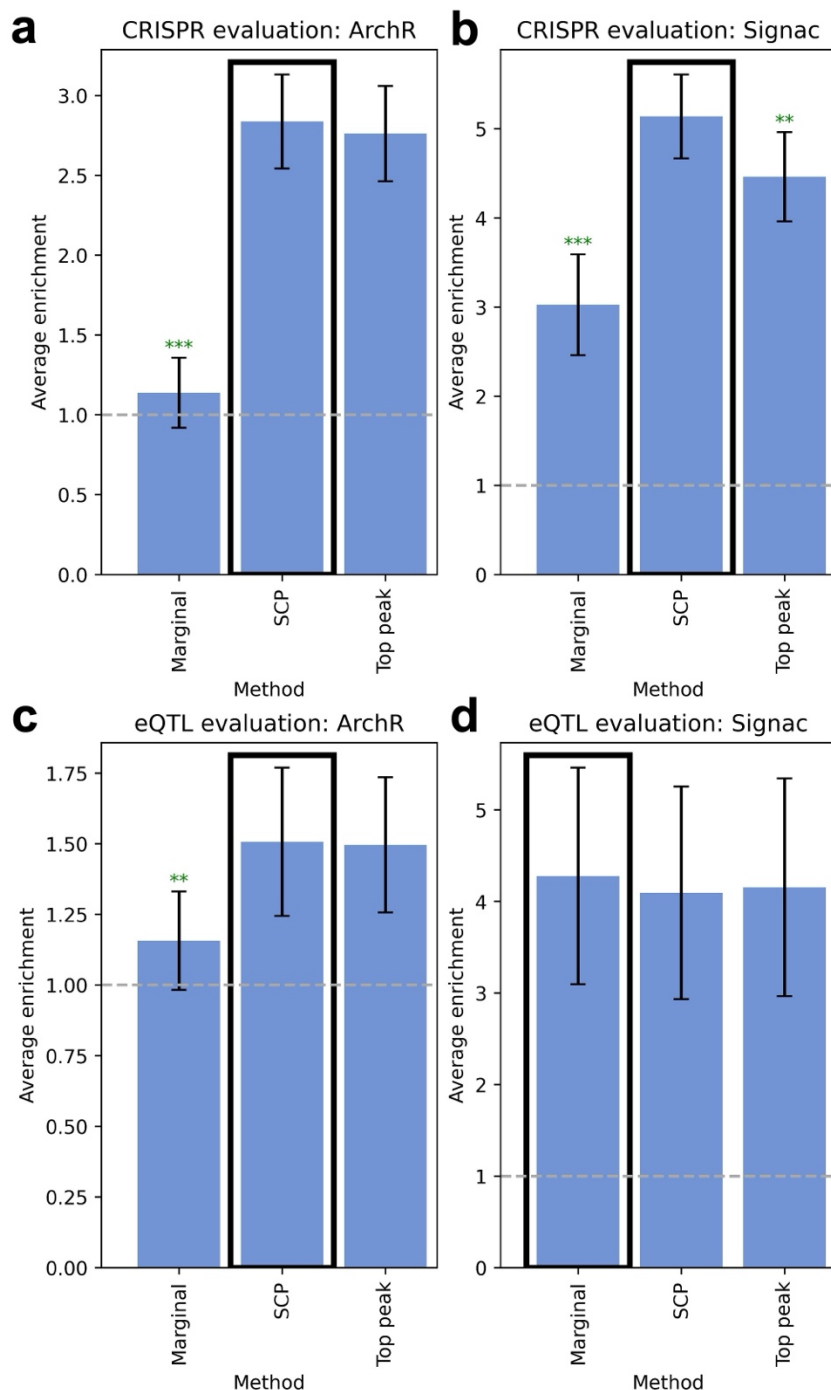
Supplementary Figure 3.48. Evaluations of multiple causal peak fine-mapping using peak-peak correlations computed across single cells vs. metacells. **a)** Average enrichment across recall values of links predicted by multiple causal peak fine-mapped ArchR scores using for 448 links validated by CRISPR. **b)** Average enrichment across recall values of links predicted by multiple causal peak fine-mapped Signac scores for 448 links validated by CRISPR. **c)** Average enrichment across recall values of links predicted by multiple causal peak fine-mapped ArchR scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. **d)** Average enrichment across recall values of links predicted by multiple causal peak fine-mapped Signac scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. In primary analyses, SuSiE used peak-peak correlations computed within the susie() function across metacells (for ArchR) or single cells (for

Signac) to perform multiple causal peak fine-mapping. We also assessed fine-mapping using alternatively computed peak-peak correlations (across single cells for ArchR and across metacells for Signac), which we implemented by supplying a peak correlation matrix (across single cells or metacells) to the “R” argument of the `susie_rss()` function. Fine-mapped scores using alternative peak-peak correlations are noted in parentheses on the x-axis (e.g. “metacell R” or “single-cell R”). Fine-mapping using alternative peak-peak correlations did not outperform fine-mapping using peak-peak correlations from the focal peak-gene linking method. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.



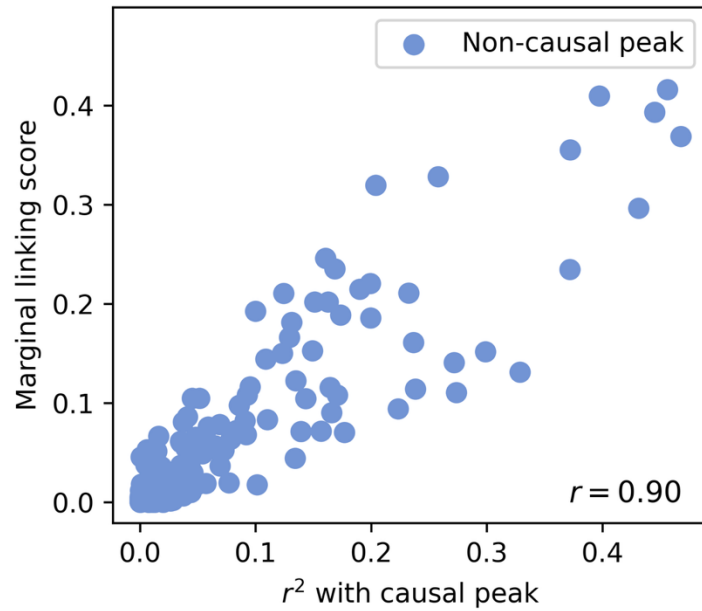
Supplementary Figure 3.49. Evaluations of fine-mapping restricting to peak-gene pairs with strong marginal correlations. **a)** Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores for 448 links validated by CRISPR. **b)** Average enrichment across recall values of links predicted by Signac vs. fine-mapped Signac scores for 448 links validated by CRISPR. **c)** Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. **d)** Average enrichment across recall values of links predicted by Signac vs. fine-mapped Signac scores for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. For all fine-mapping methods labeled “($|r| > X$)”, we restricted the fine-mapping algorithm to candidate peak-gene pairs with

marginal peak-gene $|r| > X$ (with $X = 0.1$ (for ArchR) and $X = 0.01$ (for Signac)). Restricting to peak-gene pairs with strong marginal correlations yielded similar or lower average enrichments on both CRISPR and eQTL evaluation sets vs. fine-mapping without this restriction. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping; SCP + Functional, single causal peak functionally informed fine-mapping; MCP, multiple causal peak fine-mapping; MCP + Functional, multiple causal peak functionally informed fine-mapping.

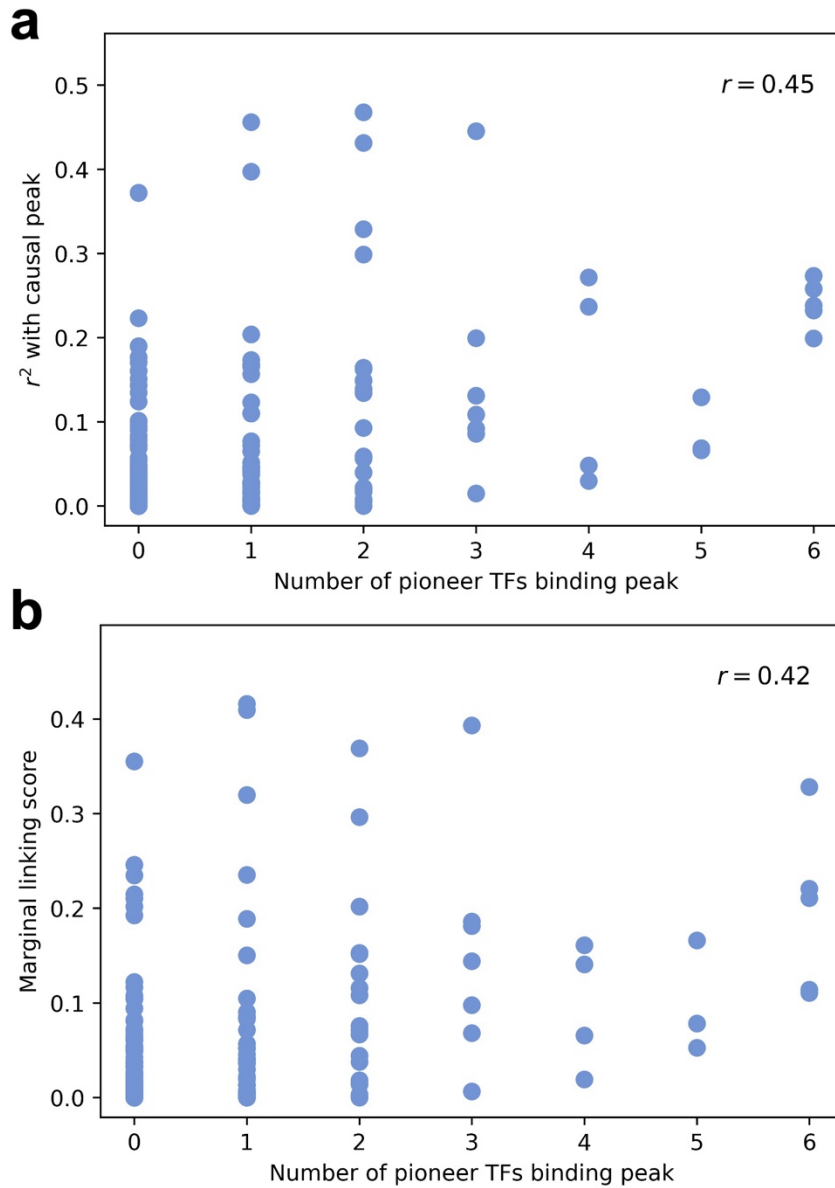


Supplementary Figure 3.50. Evaluations of fine-mapping vs. a ‘top peak’ fine-mapping approach. **a)** Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores vs. a ‘top peak’ fine-mapping approach for 448 links validated by CRISPR. **b)** Average enrichment across recall values of links predicted by Signac vs. fine-mapped Signac scores vs. a ‘top peak’ fine-mapping approach for 448 links validated by CRISPR. **c)** Average enrichment across recall values of links predicted by ArchR vs. fine-mapped ArchR scores vs. a ‘top peak’ fine-mapping approach for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. **d)** Average enrichment across

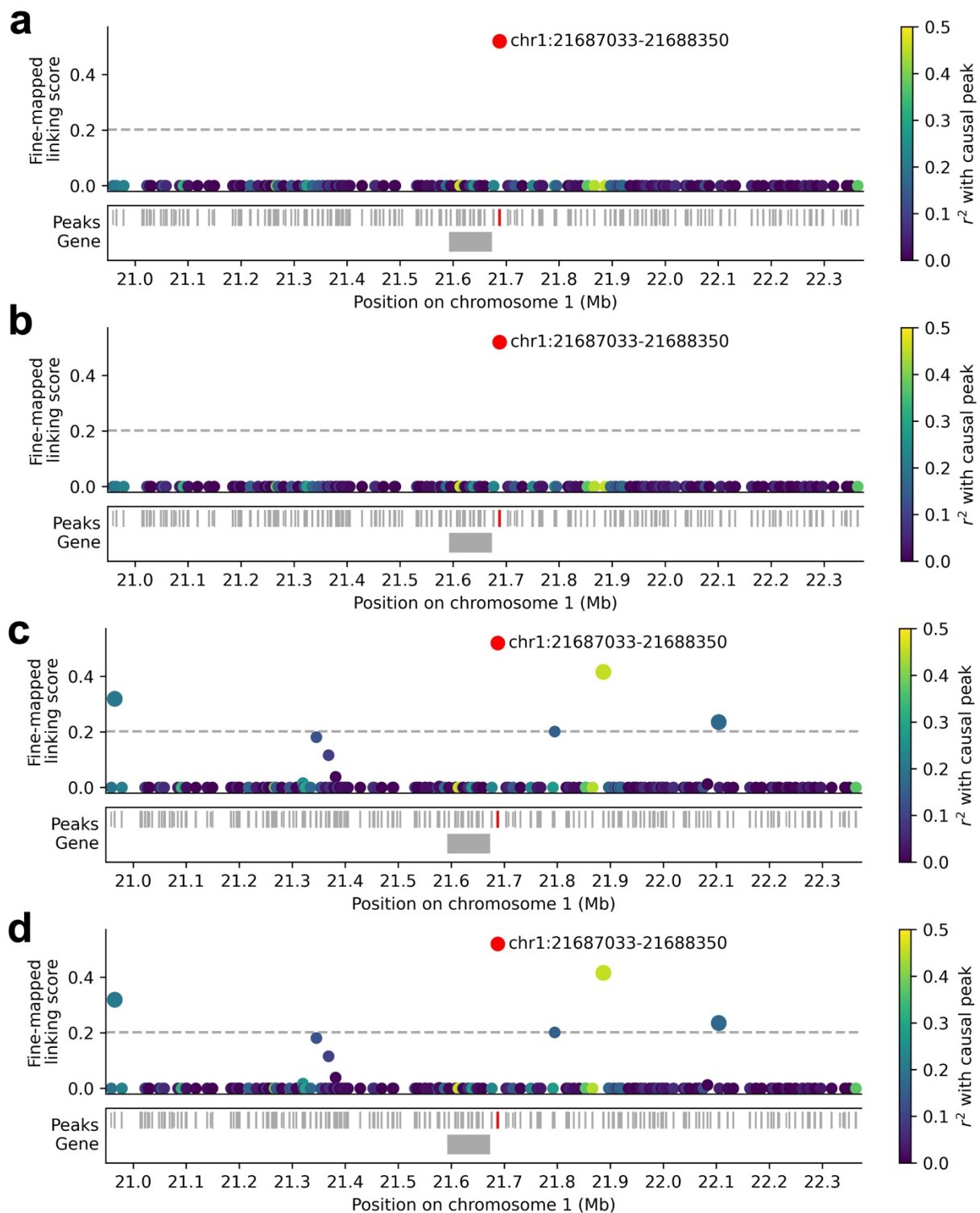
recall values of links predicted by Signac vs. fine-mapped Signac scores vs. a ‘top peak’ fine-mapping approach for 39,194 fine-mapped eSNP–eGene pairs attaining maximum PIP > 0.5 across GTEx tissues. We defined the ‘top peak’ fine-mapping approach as follows: for each gene, we assigned the peak with the highest r^2 to the gene a posterior inclusion probability (PIP) of 1, and assigned PIP of 0 to all other candidate peaks. We then defined fine-mapped linking scores as PIP x marginal peak-gene linking score (as for all other fine-mapping approaches). The ‘top peak’ fine-mapping approach significantly underperformed SCP fine-mapped Signac scores on CRISPR evaluation data, but otherwise attained similar average enrichments across shared values of recall. However, ‘top peak’ fine-mapping yielded very low recall, for both the eQTL evaluation (recall 0.0053-0.025 for ‘top-peak’ fine-mapped methods vs. 0.33-0.84 for marginal and SCP fine-mapped methods) and the CRISPR evaluation (recall 0.07-0.2 for ‘top-peak’ fine-mapped methods vs. 0.35-0.98 for marginal and SCP fine-mapped methods) (because the number of peak-gene pairs with PIP > 0 is strictly equal to the number of genes); thus, ‘top-peak’ fine-mapping is a less desirable approach. Bars and confidence intervals denote estimates and standard errors, respectively, meta-analyzed across data set-cell type pairs. Green asterisks denote significant underperformance of the focal method vs. the top-performing method (denoted with a black outline) (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). Grey horizontal dashed lines mark 1.0x enrichment. SCP, single causal peak fine-mapping.



Supplementary Figure 3.51. Relationship between marginal linking score and r^2 with causal peak across non-causal peaks near *RAP1GAP*. Relationship between marginal linking score (squared correlation) with *RAP1GAP* and squared correlation with the “causal peak” (peak attaining highest linking score with *RAP1GAP*). Each point represents 1 “non-causal” peak (peak <1Mb from the TSS of *RAP1GAP* that does not attain the highest *RAP1GAP* linking score).

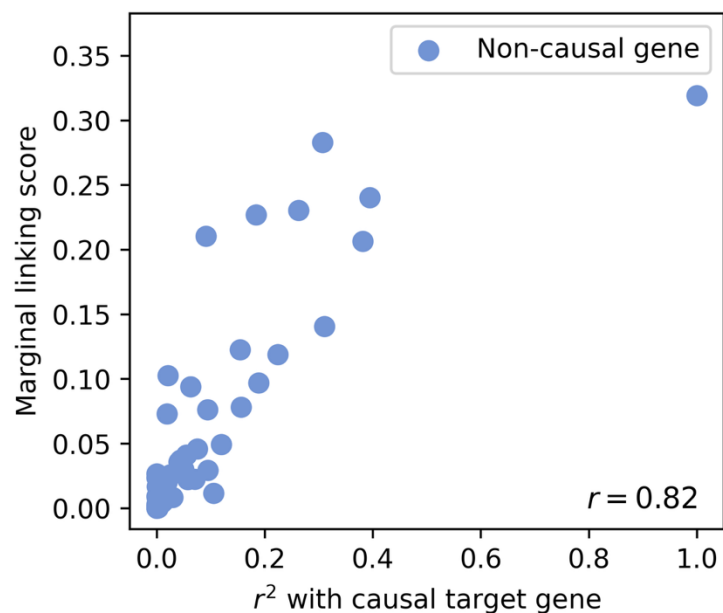


Supplementary Figure 3.52. Relationship between pioneer TF activity, peak-peak co-accessibility, and peak-gene co-activity in the *cis* window of *RAP1GAP*. **a)** Relationship between squared correlation with the “causal peak” and the number of pioneer TFs binding both the causal peak and the focal peak, across “non-causal” peaks. **b)** Relationship between marginal linking score (squared correlation) with *RAP1GAP* and the number of pioneer TFs binding both the focal peak and the causal peak, across “non-causal” peaks. The “causal peak” (chr1:21687033-21688350) is defined as the peak attaining the highest linking score with *RAP1GAP*, and “non-causal peaks” are defined as peaks <1Mb from the TSS of *RAP1GAP* that do not attain the highest *RAP1GAP* linking score. We measured the number of pioneer TFs that bind each non-causal peak and the causal peak (of 6 total pioneer TFs that bind the causal peak).



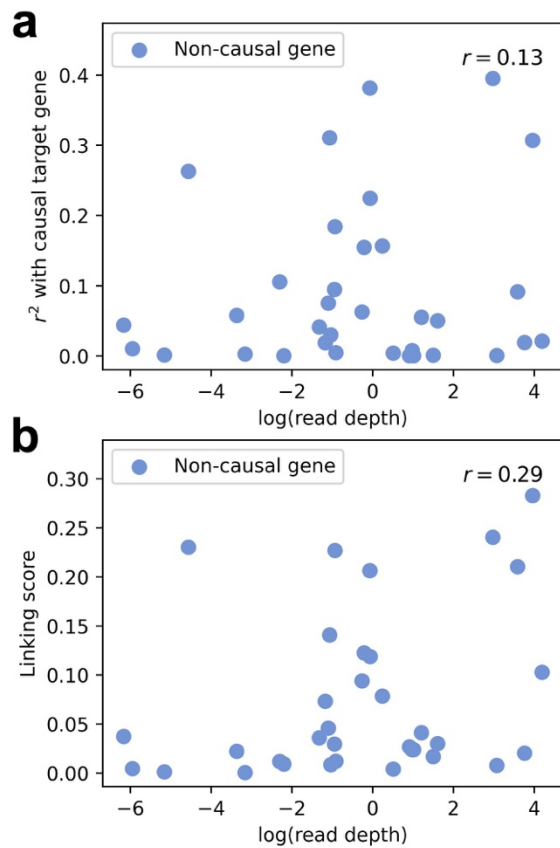
Supplementary Figure 3.53. Fine-mapped linking scores for *RAP1GAP*. Fine-mapped linking scores for each peak <1Mb from the TSS of the gene *RAP1GAP* from **a**) single causal peak fine-mapping, **b**) functionally informed single causal peak fine-mapping, **c**) multiple causal peak fine-mapping, and **d**) functionally informed multiple causal peak fine-mapping. r^2 values are reported as dots in the middle of each peak, with the top correlated peak colored in red and

other peaks colored based on their squared correlation to the top correlated peak. In each panel, dashed line denotes linking score threshold corresponding to the ArchR linking threshold. Fine-mapping approaches assuming multiple causal peaks linked 3 additional peaks to *RAP1GAP*. None of these peaks harbored fine-mapped GWAS variants; nevertheless, these may be true causal peaks in this or other contexts. Functionally-informed approaches yielded highly similar results to non-functionally-informed approaches.

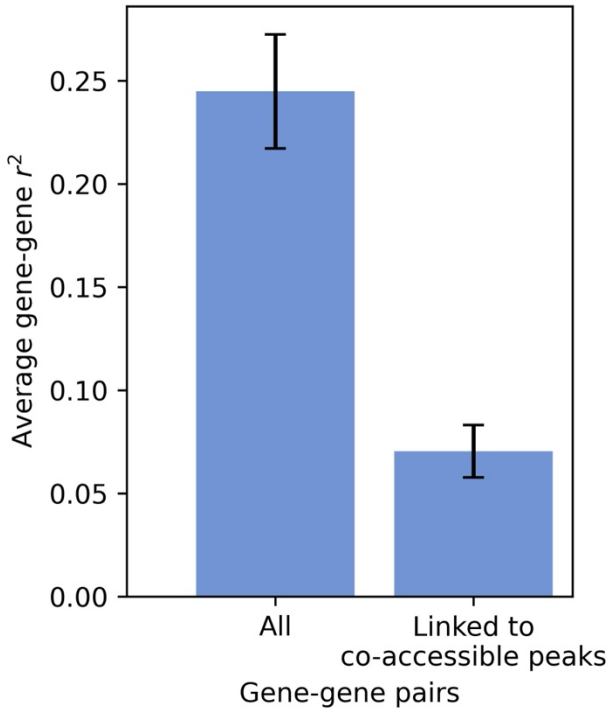


Supplementary Figure 1.54. Relationship between marginal linking score and r^2 with causal target gene across non-causal genes near the peak chr19:8417808-8418540.

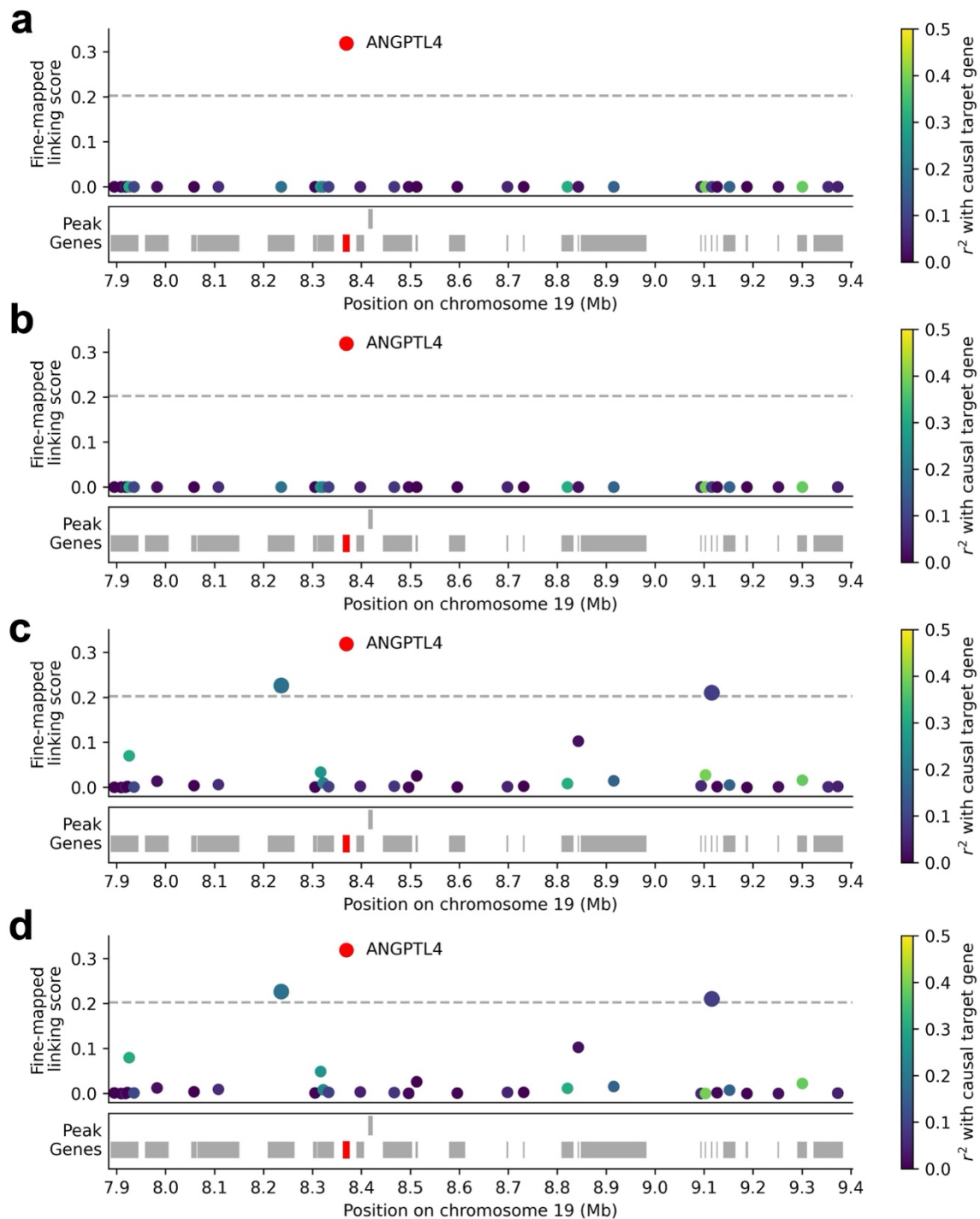
Relationship between marginal linking score (squared correlation) with the focal peak (chr19:8417808-8418540) and squared correlation with the causal target gene (gene attaining highest linking score with the focal peak), across non-causal target genes. Each point represents 1 “non-causal” gene (gene <1Mb from the focal peak that does not attain the highest linking score).



Supplementary Figure 1.55. Relationship between gene read depth and gene-gene tagging for the peak chr19:8417808-8418540. a) Relationship between squared correlation with the causal target gene (*ANGPTL4*) and log(read depth), across non-causal genes. **b)** Relationship between marginal linking score (squared correlation) with the focal peak chr19:8417808-8418540 and log(read depth) across non-causal genes. The causal gene (*ANGPTL4*) is defined as the gene attaining the highest linking score with the focal peak, and non-causal genes are defined as genes <1Mb from the focal peak that do not attain the highest linking score. We also note that *ANGPTL4* had only moderate read depth (with percentile = 0.66 among all genes <1Mb from chr19:8417808-8418540).



Supplementary Figure 1.56. Average r^2 with *ANGPTL4* for genes linked to peaks co-accessible with peak chr19:8417808-8418540 vs. all *cis* genes. We identified 51 peaks linked to any gene <1Mb from the peak chr19:8417808-8418540 (besides *ANGPTL4*) (defining “linked” as ArchR linking score (r^2) > 0.2025). From these peak-gene links, we identified 7 genes linked to peaks highly co-accessible with chr19:8417808-8418540 (defining “highly co-accessible” as peak-peak r^2 > 0.2025). We then computed the average r^2 between *ANGPTL4* and each of these genes (right) vs. the average r^2 between *ANGPTL4* and each gene <1Mb from *ANGPTL4*. Average r^2 with *ANGPTL4* for genes linked to peaks highly co-accessible with chr19:8417808-8418540 was substantially higher ($r^2 = 0.24$) than for all genes ($r^2 = 0.070$). Confidence intervals denote standard errors (from jackknifing across individual genes).



Supplementary Figure 1.57. Gene-level fine-mapped linking scores for the peak chr19:8417808-8418540. Gene-level fine-mapped linking scores for each gene <1Mb from the peak spanning chr19:8417808-8418540 from **a)** single causal gene fine-mapping, **b)** functionally informed single causal gene fine-mapping, **c)** multiple causal gene fine-mapping, and **d)** functionally informed multiple causal gene fine-mapping. r^2 values are reported as dots in the middle of each gene, with the top correlated gene colored in red and other peaks colored based on their squared correlation to the top correlated gene. In each panel, dashed line denotes linking score threshold corresponding to the ArchR linking threshold.

Chapter 4 Appendix

Methods

Genetic data generation, processing, and quality control

The data used in this study were originally generated from a Peruvian tuberculosis progression cohort. Details of the genotyping process can be found in ref.²³¹, but in brief, 4,002 admixed Peruvian individuals were genotyped on a custom Affymetrix array (LIMAArray) designed based on whole exome sequencing from 116 individuals from the same Peruvian cohort.

We used the processed genotyping data from ref.⁹⁹, which included 675,358 genotyped variants after removing variants that are significantly associated with batch ($p < 1 \times 10^{-5}$), duplicate variants, variants with a low call rate ($< 95\%$), variants with significant differences in the missingness rate between cases and controls ($> 10^{-5}$), or variants with Hardy-Weinberg p value $< 10^{-5}$ in controls. These 675,358 variants were mapped to GRCh37/b37, pre-phased with SHAPEIT2, and imputed with IMPUTE2 using the 1000 Genomes Project Phase 3 as the reference panel. After removing SNPs with an INFO score $< .9$, minor allele frequency < 0.05 , or deletions, the remaining variants were converted to GRCh38 using liftOver, resulting in 5,486,956 variants for eQTL analysis.

Genotyping principal components were previously calculated as described in ref.⁹⁹, by combining autosomal genotyping data from the full cohort ($n = 4,002$) with whole-genome sequencing data from the 1000 Genomes Project Phase 3 ($n = 2,045$ individuals from 26 populations); matching variants with MAF $> 1\%$ based on position, reference, and alternate allele; removing highly correlated variants ($r^2 > 0.2$, window size = 50, and step size = 5); and conducting PCA with PLINK.

Single-cell CITE-seq data generation, processing, and quality control

The memory T cell CITE-seq dataset used in this study was previously generated as part of a study of a subset of the genotyped Peruvian tuberculosis progression cohort¹⁸⁰. All samples were collected from individuals after treatment and resolution of *M. tuberculosis* infection. Detailed experimental methods are available in ref.¹⁸⁰. In summary, we collected PBMCs from 259 genotyped Peruvian individuals, isolated memory T cells through negative selection on CD45RA, and carried out an optimized version of the CITE-seq protocol with a panel of 31 surface proteins with TotalSeq-A (BioLegend) oligonucleotide-labeled antibodies. We conducted 10x Genomics library preparation and sequencing for pools of six donors and subsequently demultiplexed donors using Demuxlet.

In the present study, we used single-cell sequencing data from ref.¹⁸⁰ that was aligned to GRCh38 with Cell Ranger (3.1.0). Antibody tags were aligned to a dictionary of oligonucleotides. We followed the same quality control steps as in the original study: we removed doublets labeled by Demuxlet and other genotype mismatches, cells with <500 genes expressed or >20% of mitochondrial unique molecular identifiers, and cells lacking memory T cell surface markers (CD3 and CD45RO). After QC, 500,089 cells remained. We log-normalized gene expression data and applied centered log-ratio transformation to surface marker UMI counts for each of these cells.

We used canonical variates (CVs) reflecting an RNA- and protein-informed low-dimensional embedding of cells that were published in ref.¹⁸⁰ to represent cell states. These CVs were defined using canonical correlation analysis (CCA) on single-cell expression of genes (union of the top 1,000 most variable genes per sequencing batch, excluding TCR genes) and non-control surface markers.

Robust single-cell-resolution eQTL model

To accurately test associations between SNPs and single-cell gene expression, we developed a strategy based on the cluster-robust variance estimator (CRVE) and the score/Lagrange multiplier test to robustly estimate calibrated p-values in the presence of

correlated cells within donors and moderate model misspecification (e.g. over- or under-dispersion). The method requires donor-level genotyping data, single-cell RNA expression data, cell state(s) (e.g. low-dimensional embeddings or cell weights), and donor- and cell-level covariates.

First, we modeled single-cell gene expression counts using a Poisson Generalized Linear Model:

$$\log(E) = \beta_0 + \beta_G G + \sum_{i=1}^n \beta_{CV_i} CV_i + \sum_{i=1}^n \beta_{G \times CV_i} (G \times CV_i) + \mu, [4.2]$$

where E is the gene expression count, β_0 is the intercept term, G is the alternate allele dosage of the tested SNP, CV is the cell state variable of interest, $G \times CV$ is the interaction term between genotype and cell state, and μ are the residuals.

We added biological and technical covariates to equation 4.2 to account for possible confounding:

$$\begin{aligned} \log(E) = & \beta_0 + \beta_G G + \sum_{i=1}^n \beta_{CV_i} CV_i + \sum_{i=1}^n \beta_{G \times CV_i} (G \times CV_i) + \\ & \beta_{age} age + \beta_{sex} sex + \sum_{i=1}^5 \beta_{gPC_i} gPC_i + \\ & \beta_{nUMI} nUMI + \beta_{mito} mito + \sum_{i=1}^5 \beta_{ePC_i} ePC_i + \mu \end{aligned} [4.3]$$

At the donor-level, age is the scaled age (mean = 0 and s.d. = 1), sex is a binary dummy variable for sex, and gPC are genotype PCs to account for genetic ancestry and population stratification. At the single-cell level, $nUMI$ is the scaled and natural log-transformed number of

UMIs, *mito* is the proportion of mitochondrial expression, and *ePC* are the expression PCs. To simplify the notation in the following steps, we can express an unrestricted (full) model (equation 4.3) in its matrix form as:

$$\log(\mathbf{E}) = \mathbf{X}\beta + \mu \quad [4.4]$$

Assuming n cells and p predictor variables, $\mathbf{E}$ is an $n \times 1$ column vector containing the gene expression counts, $\mathbf{X}$ is an $n \times p$ design matrix where columns each correspond to a predictor included in equation 4.3, β is a $p \times 1$ vector of coefficients, and μ is a column vector with the residuals.

For inferential purposes, we fitted a restricted (null) model by setting q terms of interest to be equal to zero:

$$\log(\mathbf{E}) = \mathbf{X}\tilde{\beta} + \tilde{\mu}, \quad [4.5]$$

where $\tilde{\beta}$ is a vector $(p - q) \times 1$ containing the parameter coefficients and $\tilde{\mu}$ is a column vector containing the residuals, under the null model.

Computation of single-cell-level contribution to score function

To perform a Lagrange multiplier test, we first calculated the contribution of each single cell to the score function (i.e., gradient of the Poisson GLM objective log-likelihood), in accordance with refs.^{232,233} The Poisson probability mass function is defined as:

$$f(Y_i = y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad [4.6]$$

where λ is the mean (and variance) parameter of the Poisson distribution and y is a discrete random variable.

The Poisson GLM log-likelihood function for each cell indexed by i is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \ln \left(\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \right) = \sum_{i=1}^n [y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)] \quad [4.7]$$

In a Poisson GLM with a log-link function, we can relate λ to the predictors as the expected value $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$; therefore, we can rewrite the log-likelihood as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i(\mathbf{x}_i^T \boldsymbol{\beta}) - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \ln(y_i!)] \quad [4.8]$$

The scores $U_i(\boldsymbol{\beta})$ can be calculated by differentiating the log-likelihood function with respect to the vector $\boldsymbol{\beta}$:

$$\begin{aligned} \mathbf{U}_i(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} [y_i(\mathbf{x}_i^T \boldsymbol{\beta}) - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \ln(y_i!)] \\ &= y_i \mathbf{x}_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = (y_i - \lambda_i) \mathbf{x}_i, \end{aligned} \quad [4.9]$$

where $y_i - \lambda_i$ are the residuals μ . For parameter inference via the score test/Lagrange multiplier (LM), we built observation-level scores under the null (equation 4.5) using the restricted model residuals as follows:

$$\tilde{\mathbf{S}} = \text{diag}(\tilde{\boldsymbol{\mu}}) \mathbf{X}, \quad [4.10]$$

where $\tilde{\mathbf{S}}$ is an observation-level scores matrix of dimensions $n \times p$ and $\tilde{\mu}$ are the residuals from the restricted model. $\tilde{\mathbf{S}}$ contains the contribution of all predictors from the full model to the gradient of the log-likelihood function evaluated at the restricted model solution.

Cluster-robust variance estimation

The naive variance-covariance matrix of the parameters from a Poisson GLM assumes all observations are independent from each other and the model is accurately specified. In single-cell data, cells from the same donor are not independent observations. Furthermore, gene expression might be difficult to model under stringent assumptions (e.g. the mean and variance of gene expression may not be the same as assumed in a Poisson GLM). We relaxed the assumptions of the Poisson GLM to handle the correlation of cells within each donor and allow for potential model misspecification by applying the cluster-robust variance estimator (CRVE) as described in ref.²³³:

$$\tilde{\mathbf{V}}_{CRVE} = m\tilde{\mathbf{H}}^{-1} \left\{ \sum_{d=1}^D \tilde{\mathbf{s}}_d(\tilde{\mathbf{s}}_d)' \right\} \tilde{\mathbf{H}}^{-1}, \quad [4.11]$$

where $\tilde{\mathbf{s}}_d$ (a vector of length p) is the sum of all the centered single-cell scores from donor d for each predictor, m is a finite-sample correction factor $D/(D-1)$ (where D is defined as the number of donors), and $-\tilde{H}^{-1}$ is the classical Poisson GLM variance-covariance matrix of the full model evaluated at the restricted model solution:

$$\tilde{\mathbf{V}} = (-\tilde{\mathbf{H}})^{-1} = (\mathbf{X}'\tilde{\mathbf{W}}\mathbf{X})^{-1}, \quad [4.12]$$

where $\mathbf{X}$ is the design matrix with all predictors from the unrestricted model and $\tilde{\mathbf{W}} = \text{diag}(\tilde{\mu}_1, \dots, \tilde{\mu}_n)$ contains the residuals from the restricted model.

Lagrange multiplier (score) test

We tested for single or multiple betas jointly by applying general and multiple exclusion linear restrictions. For each model, we tested the null hypothesis:

$$H_0 : \mathbf{R}\beta = \mathbf{0}_q, \quad [4.13]$$

where $\mathbf{R}$ is a matrix with q linear restrictions as rows and k parameters as columns. $\mathbf{R}$ contains 1s for the parameter(s) being tested. For example, for the model:

$$\log(E) = \beta_0 + \beta_G G + \beta_{CV_1} CV_1 + \beta_{G \times CV_1} (G \times CV_1) + \mu \quad [4.14],$$

testing a single restriction $H_0 : \beta_{G \times CV_1} = 0$ requires a row matrix $\mathbf{R} = [0, 0, 0, 1]$. Likewise, testing the joint restriction $H_0 : \beta_G = 0, \beta_{G \times CV_1} = 0$ requires $\mathbf{R} = [0, 1, 0, 0; 0, 0, 0, 1]$

We defined the generalized Lagrange multiplier test as proposed in ref.²³⁴:

$$LM = \left(\sum_{i=1}^N \tilde{s}_i \right)' \tilde{\mathbf{H}}^{-1} \mathbf{R}' \left[\mathbf{R} \tilde{\mathbf{V}}_{CRVE} \mathbf{R}' \right]^{-1} \mathbf{R} \tilde{\mathbf{H}}^{-1} \left(\sum_{i=1}^N \tilde{s}_i \right), \quad [4.15]$$

where the first and last components on the right-hand side of the formula are the sum of uncentered restricted scores for each predictor across all cells. We derived a p-value from the LM statistic based on the area under the upper tail of the χ^2 distribution curve with q degrees of freedom. When $q = 1$, the LM can be linearized by taking its square root to derive a Z-score.

Bibliography

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *The American Journal of Human Genetics* **110**, 1–16 (2023).
3. Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New England Journal of Medicine* **254**, 1264 (2006).
4. Dutka, M., Zimmer, K., Ćwiertnia, M., Ilczak, T. & Bobiński, R. The role of PCSK9 in heart failure and other cardiovascular diseases—mechanisms of action beyond its effect on LDL cholesterol. *Heart Fail. Rev.* **29**, 917–937 (2024).
5. Bao, X. *et al.* Targeting proprotein convertase subtilisin/kexin type 9 (PCSK9): from bench to bedside. *Signal Transduct. Target. Ther.* **9**, (2024).
6. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *PNAS* **106**, 9362–9367 (2009).
7. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
8. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21**, 3255–3263 (2012).
9. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
10. Farh, K. K. H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
11. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).
12. Reyes-Palomares, A. *et al.* Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension. *Nat. Commun.* **11**, (2020).
13. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).
14. Yang, S. Y. C. *et al.* Pan-cancer analysis of longitudinal metastatic tumors reveals genomic alterations and immune landscape dynamics associated with pembrolizumab sensitivity. *Nat. Commun.* **12**, (2021).
15. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (1979)*. **337**, 1190–1195 (2012).
16. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
17. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
18. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
19. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).

20. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
21. Kim-Hellmuth, S. *et al.* Cell type-specific genetic regulation of gene expression across human tissues. *Science* (1979). **369**, (2020).
22. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* (1979). <https://www.science.org> (2019).
23. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
24. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
25. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
26. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. *bioRxiv* <https://doi.org/10.1101/2023.11.09.563812>.
27. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
28. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
29. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
30. Pennacchio, L. A. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
31. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
32. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
33. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
34. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).
35. Huang, J. *et al.* Dynamic Control of Enhancer Repertoires Drives Lineage and Stage-Specific Transcription during Hematopoiesis. *Dev. Cell* **36**, 9–23 (2016).
36. Badia-i-Mompel, P. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
37. Gamage, U., Warnakulasuriya, K., Hansika, S. & Silva, G. N. CRISPR Gene Therapy: A Promising One-Time Therapeutic Approach for Transfusion-Dependent β -Thalassemia—CRISPR-Cas9 Gene Editing for β -Thalassemia. *Thalassemia Reports* **13**, 51–69 (2023).
38. Yang, Z., Su, B., Cao, C. & Wen, J.-R. Regulatory DNA sequence Design with Reinforcement Learning. <http://arxiv.org/abs/2503.07981> (2025).
39. Blein, T. *et al.* A conserved molecular framework for compound leaf development. *Science* (1979). **322**, 1835–1839 (2008).
40. Sankaran, V. G. *et al.* Developmental and species-divergent globin switching are driven by BCL11A. *Nature* **460**, 1093–1097 (2009).
41. Bauer, D. E. *et al.* An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science* (1979). **342**, 253–257 (2013).

42. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).
43. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
44. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
45. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
46. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
47. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nat. Methods* **12**, 1143–1149 (2015).
48. Fulco, C. P. *et al.* Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science (1979)*. **354**, 769–773 (2016).
49. Ulirsch, J. C. *et al.* Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* **165**, 1530–1545 (2016).
50. Wakabayashi, A. *et al.* Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *PNAS* **113**, 4434–4439 (2016).
51. Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, 943 (2018).
52. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
53. Gasperini, M. *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).
54. Schraivogel, D. *et al.* Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat. Methods* **17**, 629–635 (2020).
55. Klann, T. S. *et al.* CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat. Biotechnol.* **35**, 561–568 (2017).
56. Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).
57. Xie, Z. X. *et al.* Rapid and efficient CRISPR/Cas9-based mating-type switching of *Saccharomyces cerevisiae*. *G3: Genes, Genomes, Genetics* **8**, 173–183 (2018).
58. Sheth, M. U. *et al.* Mapping enhancer-gene regulatory interactions from single-cell data. *bioRxiv* <https://doi.org/10.1101/2024.11.23.624931> (2024)
doi:10.1101/2024.11.23.624931.
59. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).
60. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
61. Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019).

62. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
63. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* (1979). **369**, 1318–1330 (2020).
64. Wang, Q. S. *et al.* Leveraging supervised learning for functionally informed fine-mapping of cis-eQTLs identifies an additional 20,913 putative causal eQTLs. *Nat. Commun.* **12**, 3394 (2021).
65. Huang, Q. Q., Ritchie, S. C., Brozynska, M. & Inouye, M. Power, false discovery rate and Winner’s Curse in eQTL studies. *Nucleic Acids Res.* **46**, (2018).
66. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
67. Connally, N. *et al.* The missing link between genetic association and regulatory function. *Elife* **11**, (2022).
68. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
69. Mumbach, M. R. *et al.* HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
70. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-023-00599-5> (2023) doi:10.1038/s41576-023-00599-5.
71. Kang, J. B., Raveane, A., Nathan, A., Soranzo, N. & Raychaudhuri, S. Annual Review of Genomics and Human Genetics Methods and Insights from Single-Cell Expression Quantitative Trait Loci. *Annu. Rev. Genomics Hum. Genet.* **43**, 19 (2023).
72. Jang, B. *et al.* A meta-analysis of single-nucleus expression quantitative trait loci linking genetic risk to brain disorders. *Nat. Genet.* <https://doi.org/10.1038/s41588-026-02541-x> (2026) doi:10.1038/s41588-026-02541-x.
73. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
74. Valencia, C. *et al.* Modeling heterogeneity in single-cell perturbation states enhances detection of response eQTLs. *Nat. Genet.* **57**, 2882–2890 (2025).
75. Zheng, Y., Shen, S. & Keleş, S. Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D. *Genome Biol.* **23**, (2022).
76. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
77. Ma, S. *et al.* Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
78. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
79. Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat. Rev. Genet.* **24**, 535–549 (2023).
80. Sakaue, S. *et al.* Tissue-specific enhancer–gene maps from multimodal single-cell data identify causal disease alleles. *Nat. Genet.* **56**, 615–626 (2024).
81. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).

82. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
83. Pliner, H. A. *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
84. Kartha, V. K. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genomics* **2**, 100166 (2022).
85. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).
86. Mitra, S. *et al.* Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nat. Genet.* **56**, 627–636 (2024).
87. Xie, F. *et al.* Robust enhancer-gene regulation identified by single-cell transcriptomes and epigenomes. *Cell Genomics* 100342 (2023) doi:10.1016/j.xgen.2023.100342.
88. Liu, H. *et al.* Kidney multiome-based genetic scorecard reveals convergent coding and regulatory variants. *Science* (1979). **387**, (2025).
89. Su, C., Lee, D., Jin, P. & Zhang, J. scMultiMap: Cell-type-specific mapping of enhancers and target genes from single-cell multimodal data. *Nature Communications* **16**, (2025).
90. Dorans, E., Jagadeesh, K., Dey, K. & Price, A. L. Linking regulatory variants to target genes by integrating single-cell multiome methods and genomic distance. *Nat. Genet.* **57**, 1649–1658 (2025).
91. Xue, A. *et al.* Genetic regulation of cell type-specific chromatin accessibility shapes immune function and disease risk. *medRxiv* <http://medrxiv.org/lookup/doi/10.1101/2025.08.27.25334533>.
92. Kanai, M. *et al.* Population-scale multiome immune cell atlas reveals complex disease drivers. *medRxiv* <http://medrxiv.org/lookup/doi/10.1101/2025.11.25.25340489>.
93. Ziyani, C., Delaneau, O. & Ribeiro, D. M. Multimodal single cell analysis infers widespread enhancer co-activity in a lymphoblastoid cell line. *Commun. Biol.* **6**, (2023).
94. Arthur, T. D. *et al.* Complex regulatory networks influence pluripotent cell state transitions in human iPSCs. *Nat. Commun.* **15**, (2024).
95. Chatterjee, S. *et al.* Enhancer Variants Synergistically Drive Dysfunction of a Gene Regulatory Network In Hirschsprung Disease. *Cell* **167**, 355-368.e10 (2016).
96. Factor, D. C. *et al.* Cell Type-Specific Intralocus Interactions Reveal Oligodendrocyte Mechanisms in MS. *Cell* **181**, 382-395.e21 (2020).
97. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48**, 758–767 (2016).
98. Bahr, C. *et al.* A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies. *Nature* **553**, 515–520 (2018).
99. Nathan, A. *et al.* Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120–128 (2022).
100. Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
101. Cheung Kwok, A. W., Shim, H. & McCarthy, D. J. A hierarchical, count-based model highlights challenges in scATAC-seq data analysis and points to opportunities to extract finer-resolution information. *Genome Biol.* **26**, (2025).
102. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, (2019).

103. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
104. Cuomo, A. S. E. *et al.* CellRegMap : a statistical framework for mapping context-specific regulatory variants using scRNA -seq . *Mol. Syst. Biol.* **18**, (2022).
105. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
106. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet.* **55**, 1267–1276 (2023).
107. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv* <https://doi.org/10.1101/2021.09.03.21262975> doi:10.1101/2021.09.03.21262975.
108. 10x Genomics. PBMC from a healthy donor (10k), Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger 2.0.0. (2021).
109. Luecken, M. D. *et al.* A sandbox for prediction and integration of DNA, RNA, and protein data in single cells. in *35th Conference on Neural Information Processing Systems Track on Datasets and Benchmarks* (2021).
110. Xu, W. *et al.* ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 1243–1249 (2022).
111. Yardımcı, G. G. *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* **20**, 1–12 (2019).
112. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
113. Meijer, M. *et al.* Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. *Neuron* **110**, 1193-1210.e13 (2022).
114. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).
115. Benner, C. *et al.* FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
116. Simeonov, D. R. *et al.* Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature* **549**, 111–115 (2017).
117. Burchill, M. A., Yang, J., Vang, K. B. & Farrar, M. A. Interleukin-2 receptor signaling in regulatory T cell development and homeostasis. *Immunol. Lett.* **114**, 1–8 (2007).
118. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
119. Martin, J. E. *et al.* The autoimmune disease-associated IL2RA locus is involved in the clinical manifestations of systemic sclerosis. *Genes Immun.* **13**, 191–196 (2012).
120. Lowe, C. E. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.* **39**, 1074–1082 (2007).
121. Avellino, R. *et al.* An autonomous CEBPA enhancer specific for myeloid-lineage priming and neutrophilic differentiation. *Blood* **127**, 2991–3003 (2016).
122. Leroy, H. *et al.* CEBPA point mutations in hematological malignancies. *Leukemia* **19**, 329–334 (2005).

123. Pabst, T. & Mueller, B. U. Complexity of CEBPA dysregulation in human acute myeloid leukemia. *Clinical Cancer Research* **15**, 5303–5307 (2009).
124. Du, Z., Luo, G., Wang, K., Bing, Z. & Pan, S. Identification of a novel heterozygous SPTB mutation by whole genome sequencing in a Chinese patient with hereditary spherocytosis and atrial septal defect: a case report. *BMC Pediatr.* **21**, (2021).
125. Li, S. *et al.* A novel SPTB mutation causes hereditary spherocytosis via loss-of-function of β -spectrin. *Ann. Hematol.* **101**, 731–738 (2022).
126. Wang, Y. *et al.* A novel variant in the SPTB gene underlying hereditary spherocytosis and a literature review of previous variants. *BMC Med. Genomics* **17**, (2024).
127. Bulik-Sullivan, B. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
128. Vernimmen, D. & Bickmore, W. A. The Hierarchy of Transcriptional Activation: From Enhancer to Promoter. *Trends in Genetics* **31**, 696–708 (2015).
129. Samer, E. K. *et al.* Deciphering the impact of genomic variation on function. *Nature* **633**, 47–57 (2024).
130. Mayran, A. & Drouin, J. Pioneer transcription factors shape the epigenetic landscape. *Journal of Biological Chemistry* **293**, 13795–13804 (2018).
131. Lemma, R. B. *et al.* Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers. *Epigenetics Chromatin* **15**, (2022).
132. Martin-Rufino, J. D. *et al.* Transcription factor networks disproportionately enrich for heritability of blood cell phenotypes. *Science* (1979). **388**, 52–59 (2025).
133. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
134. Ribeiro, D. M. *et al.* The molecular basis, genetic control and pleiotropic effects of local gene co-expression. *Nat. Commun.* **12**, (2021).
135. Zinani, O. Q. H., Keseroğlu, K. & Özbudak, E. M. Regulatory mechanisms ensuring coordinated expression of functionally related genes. *Trends in Genetics* **38**, 73–81 (2022).
136. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
137. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
138. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
139. Kerner, G., Kamitaki, N., Strober, B. & Price, A. L. Mapping disease loci to biological processes via joint pleiotropic and epigenomic partitioning. *medRxiv* <http://medrxiv.org/lookup/doi/10.1101/2025.05.05.25327017>.
140. Vuckovic, D. *et al.* The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* **182**, 1214–1231.e11 (2020).
141. Murphy, M. M. *et al.* Role of Rap1 in promoting sickle red blood cell adhesion to laminin via BCAM/LU. *Blood* **105**, 3322–3329 (2005).
142. Stork, P. J. S. & Dillon, T. J. Multiple roles of Rap1 in hematopoietic cells: Complementary versus antagonistic functions. *Blood* **106**, 2952–2961 (2005).
143. Qiu, T., Qi, X., Cen, J. & Chen, Z. Rap1GAP alters leukemia cell differentiation, apoptosis and invasion in vitro. *Oncol. Rep.* **28**, 622–628 (2012).

144. Khattar, E. *et al.* Rap1 regulates hematopoietic stem cell survival and affects oncogenesis and response to chemotherapy. *Nat. Commun.* **10**, (2019).
145. Piva, E., Brugnara, C., Chiandetti, L. & Plebani, M. Automated reticulocyte counting: State of the art and clinical applications in the evaluation of erythropoiesis. *Clin. Chem. Lab. Med.* **48**, 1369–1380 (2010).
146. Kumar Sah, A. & Srinivasa Rao, D. Clinical Significance of Reticulocytes. in *Red Blood Cells - Properties and Functions* (IntechOpen, 2024). doi:10.5772/intechopen.1006826.
147. Zhang, L., Han, X., Zhu, Q., Qin, Y. & Jia, Y. Clinical utility of hematological parameters in aplastic anemia. *Sci. Rep.* **15**, (2025).
148. Rutherford, T. *et al.* Embryonic erythroid differentiation in the human leukemic cell line K562. *PNAS* **78**, 348–352 (1981).
149. Crispino, J. D. GATA1 in normal and malignant hematopoiesis. *Semin. Cell Dev. Biol.* **16**, 137–147 (2005).
150. Hwang, D. *et al.* GATA1 controls numbers of hematopoietic progenitors and their response to autoimmune neuroinflammation. *Blood Adv.* **6**, 5980–5994 (2022).
151. Abunimye, D. A., Okafor, I. M., Okorowo, H. & Obeagu, E. I. The role of GATA family transcriptional factors in haematological malignancies: A review. *Medicine (United States)* **103**, E37487 (2024).
152. Zhang, C. C. *et al.* Angiopoietin-like proteins stimulate ex vivo expansion of hematopoietic stem cells. *Nat. Med.* **12**, 240–245 (2006).
153. Zhang, C. C., Kaba, M., Iizuka, S., Huynh, H. & Lodish, H. F. Angiopoietin-like 5 and IGFBP2 stimulate ex vivo expansion of human cord blood hematopoietic stem cells as assayed by NOD/SCID transplantation. *Hematopoiesis and Stem Cells* **111**, 3415–3423 (2008).
154. Georgiadi, A. *et al.* Overexpression of angiopoietin-like protein 4 protects against atherosclerosis development. *Arterioscler. Thromb. Vasc. Biol.* **33**, 1529–1537 (2013).
155. Romeo, S. *et al.* Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *Journal of Clinical Investigation* **119**, 70–79 (2009).
156. Ramos, P. *et al.* ANGPTL4: A Comprehensive Review of 25 Years of Research. *Cancers (Basel)* **17**, (2025).
157. Akhter, S., Rahman, M. M., Lee, H. S., Kim, H. J. & Hong, S. T. Dynamic roles of angiopoietin-like proteins 1, 2, 3, 4, 6 and 7 in the survival and enhancement of ex vivo expansion of bone-marrow hematopoietic stem cells. *Protein Cell* **4**, 220–230 (2013).
158. Aryal, B. *et al.* ANGPTL4 deficiency in haematopoietic cells promotes monocyte expansion and atherosclerosis progression. *Nat. Commun.* **7**, (2016).
159. Schumacher, A. *et al.* Angptl4 is upregulated under inflammatory conditions in the bone marrow of mice, expands myeloid progenitors, and accelerates reconstitution of platelets after myelosuppressive therapy. *J. Hematol. Oncol.* **8**, (2015).
160. Wegner, M. S. *et al.* Ceramide synthases CerS4 and CerS5 are upregulated by 17 β -estradiol and GPER1 via AP-1 in human breast cancer cells. *Biochem. Pharmacol.* **92**, 577–589 (2014).
161. Zhang, M. *et al.* The ceramide synthase (CERS/LASS) family: Functions involved in cancer progression. *Cellular Oncology* **46**, 825–845 (2023).
162. Ren, H., Zhang, R., Zhang, H. & Bian, C. Ecnomotopic olfactory receptors in metabolic regulation. *Biomedicine and Pharmacotherapy* **179**, (2024).

163. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
164. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301–320 (2005).
165. Li, S., Hua, H. & Chen, S. Graph neural networks for single-cell omics data: a review of approaches and applications. *Brief. Bioinform.* **26**, (2025).
166. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
167. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).
168. Strober, B. J., Zhang, M. J., Amariuta, T., Rossen, J. & Price, A. L. Fine-mapping causal tissues and genes at disease-associated loci. *Nat. Genet.* **57**, 42–52 (2025).
169. Everitt, A., Whalen, S. & Pollard, K. S. Comparative evaluation of genomic footprinting algorithms for predicting transcription factor binding sites in single-cell data. *bioRxiv* <https://doi.org/10.1101/2025.08.07.669008>.
170. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nat. Commun.* **11**, (2020).
171. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: A hitchhiker’s guide to ATAC-seq data analysis. *Genome Biol.* **21**, (2020).
172. Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. & Ohler, U. Reproducible inference of transcription factor footprints in ATAC-seq and DNase-seq datasets using protocol-specific bias modeling. *Genome Biol.* **20**, (2019).
173. Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
174. Kumasaka, N. Genetic association mapping leveraging Gaussian processes. *J. Hum. Genet.* **69**, 505–510 (2024).
175. Litière, S., Alonso, A. & Molenberghs, G. The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Stat. Med.* **27**, 3125–3144 (2008).
176. Lee, Y. R. & Pustejovsky, J. E. Comparing Random Effects Models, Ordinary Least Squares, or Fixed Effects With Cluster Robust Standard Errors for Cross-Classified Data. *Psychol. Methods* **29**, 1084–1099 (2024).
177. Peng, H. & Lu, Y. Model selection in linear mixed effect models. *J. Multivar. Anal.* **109**, 109–129 (2012).
178. Zeng, B. *et al.* Comprehensive multiple eQTL detection and its application to GWAS interpretation. *Genetics* **212**, 905–918 (2019).
179. Cantor, R. M., Pan, C. & Siegmund, K. Genetic complexity at expression quantitative trait loci. *BMC Proc.* **10**, 85–89 (2016).
180. Nathan, A. *et al.* Multimodally profiling memory T cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nat. Immunol.* **22**, 781–793 (2021).
181. Goll, D. E., Thompson, V. F., Li, H., Wei, W. & Cong, J. The Calpain System. *Physiol. Rev.* <https://doi.org/10.1152/physrev.00029.2002>.-The (2003) doi:10.1152/physrev.00029.2002.-The.

182. Ono, Y., Saido, T. C. & Sorimachi, H. Calpain research for drug discovery: Challenges and potential. *Nat. Rev. Drug Discov.* **15**, 854–876 (2016).
183. Stewart, M. P., Mcdowall, A. & Hogg, N. LFA-1-mediated Adhesion Is Regulated by Cytoskeletal Restraint and by a Ca²⁺-dependent Protease, Calpain. *J. Cell Biol.* **140**, 699–707 (1998).
184. Rock, M. T., Dix, A. R., Brooks, W. H. & Roszman, T. L. β 1 Integrin-mediated T cell adhesion and cell spreading are regulated by calpain. *Exp. Cell Res.* **261**, 260–270 (2000).
185. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
186. Khunsriraksakul, C. *et al.* Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus. *Nat. Commun.* **14**, (2023).
187. Crouch, D. J. M. *et al.* Bayesian Effect Size Ranking to Prioritise Genetic Risk Variants in Common Diseases for Follow-Up Studies. *Genet. Epidemiol.* **49**, (2025).
188. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, (2014).
189. Miller, M., Rosenthal, P., Beppu, A., Gordillo, R. & Broide, D. H. Oroscomucoid like protein 3 (ORMDL3) transgenic mice have reduced levels of sphingolipids including sphingosine-1-phosphate and ceramide. *Journal of Allergy and Clinical Immunology* **139**, 1373-1376.e4 (2017).
190. Cantero-Recasens, G., Fandos, C., Rubio-Moscardo, F., Valverde, M. A. & Vicente, R. The asthma-associated ORMDL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress. *Hum. Mol. Genet.* **19**, 111–121 (2009).
191. Moffatt, M. F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
192. Todd, J. A. Etiology of Type 1 Diabetes. *Immunity* vol. 32 457–467 Preprint at <https://doi.org/10.1016/j.immuni.2010.04.001> (2010).
193. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
194. Siewert-Rocks, K. M., Kim, S. S., Yao, D. W., Shi, H. & Price, A. L. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. *Am. J. Hum. Genet.* **109**, 393–404 (2022).
195. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
196. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
197. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
198. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
199. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
200. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).
201. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *PNAS* **112**, E6456 (2015).

202. Yazar, S. *et al.* Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* (1979). **376**, (2022).
203. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A unified approach to interpreting model predictions. in *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017).
204. Xu, W. *et al.* ISSAAC-seq enables sensitive and flexible multimodal profiling of chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 1243–1249 (2022).
205. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* **40**, 437–442 (2008).
206. Auberson, M. *et al.* SLC2A9 (GLUT9) mediates urate reabsorption in the mouse kidney. *Pflugers Arch.* **470**, 1739–1751 (2018).
207. Nie, Q. *et al.* The effects of hyperuricemia on endothelial cells are mediated via GLUT9 and the JAK2/STAT3 pathway. *Mol. Biol. Rep.* **48**, 8023–8032 (2021).
208. Dickinson, M. E. *et al.* High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).
209. Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res.* **51**, D1038–D1045 (2023).
210. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1362 (2013).
211. Yue, Y., Stone, S. & Lin, W. Role of nuclear factor κ b in multiple sclerosis and experimental autoimmune encephalomyelitis. *Neural Regen. Res.* **13**, 1507–1515 (2018).
212. Ilchovska, D. (Daisy) & Barrow, D. M. An Overview of the NF- κ B mechanism of pathophysiology in rheumatoid arthritis, investigation of the NF- κ B ligand RANKL and related nutritional interventions. *Autoimmun. Rev.* **20**, (2021).
213. Housley, W. J. *et al.* Genetic variants associated with autoimmunity drive NF κ B signaling and responses to inflammatory stimuli. *Sci. Transl. Med.* **7**, (2015).
214. Ponath, G. *et al.* Enhanced astrocyte responses are driven by a genetic risk allele associated with multiple sclerosis. *Nat. Commun.* **9**, (2018).
215. Ko, Y. A. *et al.* Genetic-Variation-Driven Gene-Expression Changes Highlight Genes with Important Functions for Kidney Disease. *Am. J. Hum. Genet.* **100**, 940–953 (2017).
216. Gu, X. *et al.* Kidney disease genetic risk variants alter lysosomal beta-mannosidase (MANBA) expression and disease severity. *Sci. Transl. Med.* **13**, 1458 (2021).
217. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* **53**, 185–194 (2021).
218. Michalski, J.-C. & Klein, A. Glycoprotein lysosomal storage disorders: a- and B-mannosidosis, fucosidosis and a-N-acetylgalactosaminidase deficiency. *Biochim. Biophys. Acta* 69–84 (1999).
219. González-Jiménez, A. *et al.* Impact of Multiple Sclerosis Risk Polymorphism rs7665090 on MANBA Activity, Lysosomal Endocytosis, and Lymphocyte Activation. *Int. J. Mol. Sci.* **23**, (2022).
220. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
221. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).

222. Wells, A. *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* **10**, 5241 (2019).
223. Caron, B., Luo, Y. & Rausell, A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. *Genome Biol.* **20**, (2019).
224. Kim, S. S. *et al.* Improving the informativeness of Mendelian disease-derived pathogenicity scores for common disease. *Nat. Commun.* **11**, (2020).
225. Dey, K. K. *et al.* Evaluating the informativeness of deep learning annotations for human complex diseases. *Nat. Commun.* **11**, (2020).
226. Liu, Y. & Just, A. SHAPforxgboost: SHAP Plots for ‘XGBoost’. R package version 0.0.3. <https://github.com/liuyanguu/SHAPforxgboost>. (2019).
227. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
228. Chen, W. *et al.* Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics* **200**, 719–736 (2015).
229. Leblanc, F. J. A. & Lettre, G. Major cell-types in multiomic single-nucleus datasets impact statistical modeling of links between regulatory sequences and target genes. *Sci. Rep.* **13**, (2023).
230. Kim, S. S. *et al.* Genes with high network connectivity are enriched for disease heritability. *Am. J. Hum. Genet.* **104**, 896–913 (2019).
231. Luo, Y. *et al.* Early progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. *Nat. Commun.* **10**, (2019).
232. Kline, P. & Santos, A. A Score Based Approach to Wild Bootstrap Inference. *J. Econom. Method.* (2011).
233. Roodman, D., MacKinnon, J. G., Nielsen, M. Ø. & Webb, M. D. Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal* **19**, 4–60 (2019).
234. Wooldridge, J. M. *Econometric Analysis of Cross Section and Panel Data*. (MIT Press, 2010).