nature genetics

Article

https://doi.org/10.1038/s41588-024-01682-1

Tissue-specific enhancer–gene maps from multimodal single-cell data identify causal disease alleles

Received: 7 March 2023

Accepted: 7 February 2024

Published online: 9 April 2024

Check for updates

Saori Sakaue[1,2,3], Kathryn Weinand[1,2,3,4], Shakson Isaac[1,2,3,4], Kushal K. Dey[3,5], Karthik Jagadeesh[3,5], Masahiro Kanai[3,6,7,8], Gerald F. M. Watts[9], Zhu Zhu[9], Accelerating Medicines Partnership® RA/SLE Program and Network*, Michael B. Brenner[9], Andrew McDavid[10], Laura T. Donlin[11,12], Kevin Wei[9], Alkes L. Price[3,5,13] & Soumya Raychaudhuri[1,2,3,4]

Translating genome-wide association study (GWAS) loci into causal variants and genes requires accurate cell-type-specific enhancer–gene maps from disease-relevant tissues. Building enhancer–gene maps is essential but challenging with current experimental methods in primary human tissues. Here we developed a nonparametric statistical method, SCENT (single-cell enhancer target gene mapping), that models association between enhancer chromatin accessibility and gene expression in single-cell or nucleus multimodal RNA sequencing and ATAC sequencing data. We applied SCENT to 9 multimodal datasets including >120,000 single cells or nuclei and created 23 cell-type-specific enhancer–gene maps. These maps were highly enriched for causal variants in expression quantitative loci and GWAS for 1,143 diseases and traits. We identified likely causal genes for both common and rare diseases and linked somatic mutation hotspots to target genes. We demonstrate that application of SCENT to multimodal data from disease-relevant human tissue enables the scalable construction of accurate cell-type-specific enhancer–gene maps, essential for defining noncoding variant function.

Genome-wide association studies (GWASs) have mapped human diseases loci[1–4] harboring untapped mechanistic insights that can point to novel therapeutics[2,5]. However, only rarely are we able to define causal variants or their target genes. Of the hundreds of associated variants in a single locus, only one or a few may be causal; others simply tag causal variants[2,6,7]. Causal genes are also challenging to determine, since causal variants lie in noncoding regions 90% of the time[8–10], may regulate distant genes[11–13] and employ context-specific regulatory mechanisms[14–17].

To define causal variants and genes, previous studies used statistical and experimental approaches. Statistical fine-mapping[18–23] can narrow the set of candidate causal variants, particularly when GWAS includes diverse ancestral backgrounds[24–28]. However, statistical approaches rarely identify true causal variants with confidence[7,23,29–32]. To define causal genes, previous studies have built enhancer–gene maps that can prioritize causal variants in enhancers and link them to genes they regulate. These maps often require large-scale epigenomic and transcriptomic atlases (for example, Roadmap[33], BLUEPRINT[34] and ENCODE[35]) and are built by correlating enhancer activity with gene expression[36,37], by combining enhancer activity and probability of physical contact with the gene[38,39], or by integrating multiple linking strategies through composite scores[40]. However, current methods largely use bulk tissues or cell lines. Bulk data (1) cannot be easily

A full list of affiliations appears at the end of the paper. e-mail: soumya@broadinstitute.org

applied to rare cell populations, (2) obscure cell-type-specific gene regulation and (3) require hundreds of experimentally characterized samples. While perturbation experiments (for example, CRISPR interference[41] or base editing[42]) can point to links between enhancers and genes, they are difficult to scale because they require cell-type-specific experimental protocols[43].

Advances in single-cell technologies offer new opportunities for building cell-type-specific enhancer–gene maps. Multimodal protocols enable joint capture of epigenomic activity by assay for transposase-accessible chromatin with sequencing (ATAC–seq) alongside transcriptional activity with nuclear RNA sequencing (RNA-seq)[44–48]. These methods are applied at scale to cells in human primary tissues without disaggregation, enabling query of disease-relevant tissues. If we establish accurate links between open chromatin enhancers and genes in single cells or nuclei, statistical power should exceed bulk-tissue-based methods since each observation is at a cell-level resolution. However, the sparse and nonparametric nature of single-cell RNA-seq and ATAC–seq makes confident identification of these links challenging. So far, most methods use parametric linear regression models to link enhancers and genes (for example, ArchR[49] and Signac[46,50]) despite these features or utilize co-accessibility of regulatory regions from ATAC–seq alone (for example, Cicero[51]). These previous methods have not generally demonstrated efficacy in practice for fine-mapping causal variants in complex traits.

In this Article, we developed single-cell enhancer target gene mapping (SCENT) to accurately map enhancer–gene pairs by associating enhancer activity (that is, peak accessibility) with gene expression across multimodal single cells by Poisson regression and nonparametric bootstrapping. We predicted that expression-associated enhancers are more likely to be functionally important. We show that SCENT enhancers are enriched in statistically fine-mapped causal variants. We use SCENT enhancer–gene map to define causal variants, genes and cell types in common and rare disease loci.

Results

Overview of SCENT

SCENT accurately identifies significant association between chromatin accessibility in regulatory regions and expression of individual genes across single cells (Fig. 1a). For each peak–gene pair, we tested association between binarized chromatin accessibility in an ATAC peak with RNA-seq gene counts in *cis* (<500 kb from gene body, Methods). We tested each cell type separately to capture cell-type-specific gene regulation and to avoid spurious peak–gene associations due to gene co-regulation across cell types. Those associations can be used for prioritizing (1) likely causal variants in regulatory regions associated with gene expression, (2) likely causal genes if they are associated with the identified regulatory region and (3) the critical cell types based on cell type the association is identified in.

Since both RNA-seq and ATAC–seq data are generally sparse[50,52–55], we used Poisson regression[53,56]:

$$E_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \beta_0 + \beta_{\text{peak}} X_{\text{peak}} + \beta_{\%\text{mito}} X_{\%\text{mito}} + \beta_{n\text{UMI}} X_{n\text{UMI}} + \beta_{\text{batch}} X_{\text{batch}},$$

where E_i is the observed expression count of ith gene, and λ_i is the expected count under Poisson distribution. β_{peak} is the effect of chromatin accessibility of a peak (X_{peak}) on expression of the ith gene; its magnitude reflects the strength of the regulatory effect and its sign indicates enhancing versus silencing effect. We accounted for donor or batch effects (X_{batch}) and cell-level technical factors such as percentage of mitochondrial reads ($X_{\%\text{mito}}$).

However, Poisson regression might be suboptimal for highly expressed and dispersed genes (Fig. 1b and Extended Data Fig. 1a). Unsurprisingly, we observed uncontrolled type I error with Poisson regression in null dataset where we permuted cell barcodes to disrupt ATAC and RNA associations (Extended Data Fig. 1b,c). Moreover, commonly used single-cell analytical models (for example, negative binomial regression and linear regression) demonstrated inflated statistics (Extended Data Fig. 1d,e and Methods). To accurately estimate the error and significance of β_{peak}, we implemented a two-tailed nonparametric bootstrapping framework[57] (Methods and Extended Data Fig. 2) by resampling cells and deriving the empirical significance of β_{peak}. Bootstrapping resulted in calibrated statistics with appropriate type I error (Extended Data Fig. 1f). Therefore we used this model to define statistically significant peak–gene associations. We considered and implemented two alternative models (Extended Data Fig. 1g,h, Supplementary Fig. 1 and Supplementary Note 1).

Discovery of cell-type-specific SCENT enhancer–gene links

We obtained nine single-cell multimodal datasets from diverse human tissues representing 13 cell types (immune-related, hematopoietic, neuronal and pituitary). Since we are interested in autoimmune diseases, we generated an inflammatory tissue dataset from synovial tissues from 11 patients with rheumatoid arthritis and 1 patient with osteoarthritis (arthritis-tissue dataset; $n_{\text{donor}} = 12$, $n_{\text{cells}} = 30{,}893$)[58]. In addition, we obtained eight public datasets with 129,672 cells[46,59–63] (Fig. 1c). We analyzed 16,621 genes and 1,193,842 open chromatin peaks in *cis* after quality control (QC) (4,753,521 peak–gene pairs, 28 median peaks per gene; Supplementary Fig. 2 and Supplementary Table 1). In each dataset, we clustered and annotated cell types. Applying SCENT to each of the cell types with $n_{\text{cells}} > 500$, we constructed 23 enhancer–gene maps with a total of 87,648 peak–gene links (false discovery rate (FDR) <10%, Fig. 2a and Extended Data Fig. 3). Genes had variable number of associated peaks (from 0 to 97, mean 4.13, Extended Data Fig. 4a). After accounting for the number of cells, power to detect peak–gene links was associated with the number of ATAC–seq fragments ($P = 0.045$) and unique RNA molecules ($P = 0.030$; Extended Data Fig. 4b,c).

To assess replicability of SCENT peak–gene links, we compared the effects from the arthritis-tissue dataset (discovery) with those from other datasets in the same cell type (replication) in B cells, T/natural killer (NK) cells and myeloid cells (Supplementary Table 2a). Despite different tissue contexts, we observed high concordance in estimated effect of chromatin accessibility on gene expression for peak–gene pairs significant in both datasets (FDR <10%; mean Pearson's $r = 0.63$ of effect sizes, 99% mean directional concordance across all the datasets: Extended Data Fig. 4d). For comparison, we tested ArchR[55] or Signac[46,50], two popular linear parametric single-cell multimodal methods. In contrast, we noted lower concordance (mean Pearson's $r = 0.19$, 57% mean directional concordance in ArchR and $r = 0.38$, 99% mean directional concordance in Signac; Supplementary Table 2b,c). SCENT detects enhancer–gene links more reproducibly than previous parametric methods.

To assess if SCENT peaks were functional, we examined if (1) they co-localized with conventional *cis*-regulatory annotation, (2) their effect on expression was greater for closer peak–gene pairs, (3) they had high sequence conservation and (4) peak–gene connections had experimental support.

First, we tested the overlap of SCENT peaks with an ENCODE cCRE[64], a *cis*-regulatory annotation devised from bulk epigenomic datasets. We observed that 98.0% of SCENT peaks overlapped with cCRE on average, compared to 23.3% of random size-matched *cis*-regions and 89.0% of non-SCENT peaks (Extended Data Fig. 4e). We also annotated SCENT peaks in immune cell types with 18-state chromHMM results; 97.4% of the SCENT peaks overlapped with promoter or enhancer annotations in aggregate of 41 immune-related samples[37].

Second, we examined the strength of enhancer-gene links, hypothesizing that stronger links would be more proximal to the transcription start site (TSS) of target genes. The regression coefficient β_{peak} (the

https://doi.org/10.1038/s41588-024-01682-1

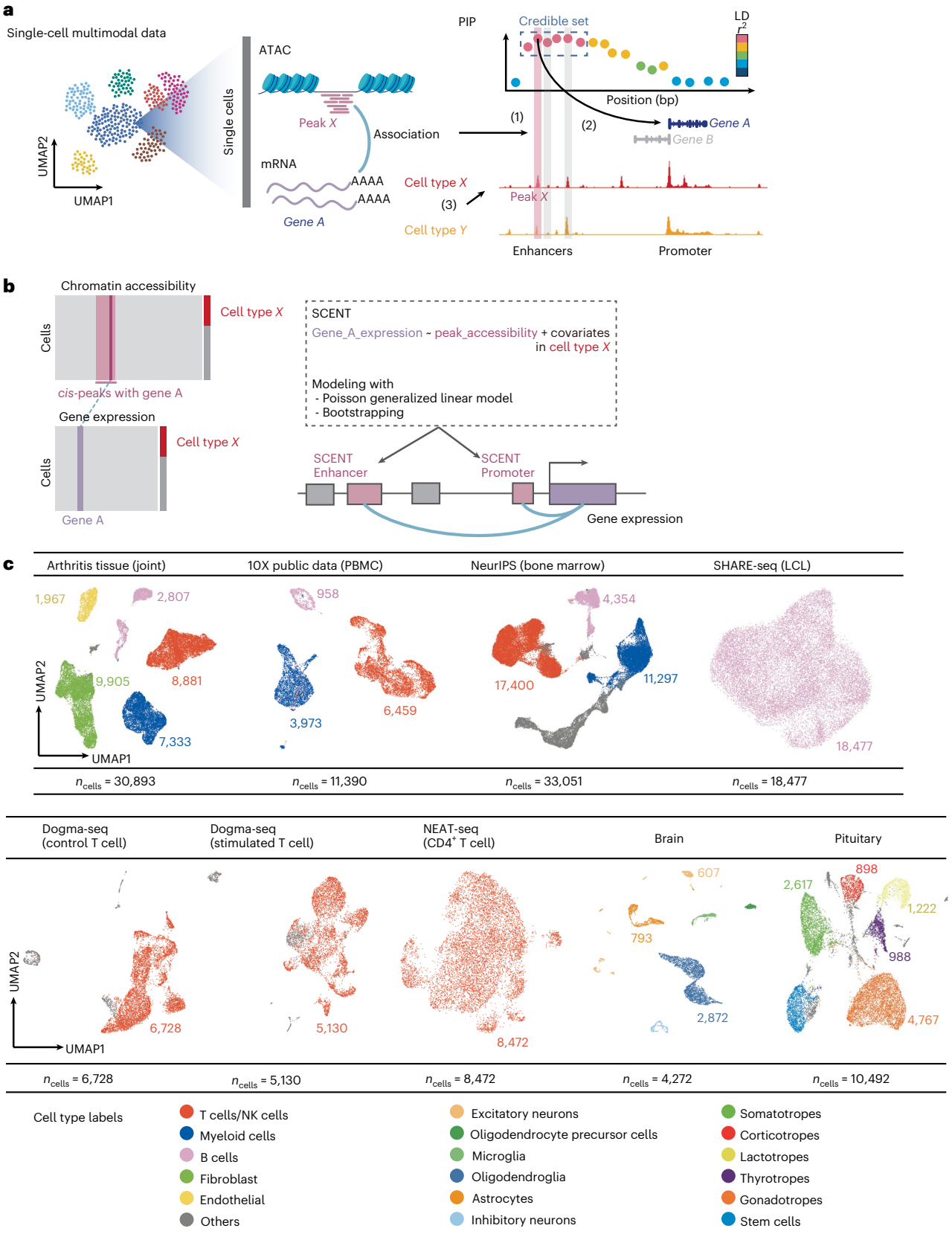


Fig. 1 | Schematic overview of SCENT and SCENT enhancer–gene pairs across nine single-cell multimodal datasets. a, SCENT identifies (1) active *cis*-regulatory regions and (2) their target genes in (3) a specific cell type. Those SCENT results can be used to define likely causal variants, genes and cell types for GWAS loci. **b**, SCENT models association between chromatin accessibility from ATAC–seq and gene expression from RNA-seq across individual cells in a given cell type. **c**, Nine single-cell datasets on which we applied SCENT to create 23 cell-type-specific enhancer–gene maps. The cells in each dataset are described in UMAP embeddings from RNA-seq, colored by cell types ($n_{cells} > 500$) and annotated by cell numbers.

effect size of peak accessibility on gene expression) became larger and more positive as the SCENT peaks got closer to the TSS (Fig. 2b,c), consistent with previous observations[55,65].

Third, we assessed whether SCENT peaks had larger phastCons score[66], reflecting higher sequence conversation across species[67]; evolutionary conserved regulatory regions are functionally active and enriched for complex trait heritability[67]. As expected, exonic regions were much more evolutionary conserved than all noncoding *cis*-region (mean ΔphastCons score 0.38, paired *t*-test $P < 10^{-323}$; Fig. 2d, purple). SCENT regulatory regions were also conserved relative to noncoding *cis*-regions (mean ΔphastCons score 0.13, paired *t*-test $P = 4.2 \times 10^{-42}$ in arthritis-tissue dataset; Fig. 2d, teal). In contrast, the ΔphastCons score between all *cis*-ATAC peaks and all noncoding *cis*-region was more modest (mean ΔphastCons score 0.092, paired *t*-test $P = 8.7 \times 10^{-27}$ in arthritis-tissue dataset; Fig. 2d, yellow). We tested the effect of promoters on the observed higher conservation in SCENT peaks. We confirmed that the ΔphastCons score for SCENT is still larger than for all *cis*-regulatory ATAC–seq peaks, even after excluding promoters. We note that the difference was consistent across multiple datasets with overlapping confidence intervals (CIs) suggesting non-statistically significant differences (Extended Data Fig. 4f). More generally, to test the effect of SCENT peaks' proximity to TSS on the higher conservation (Supplementary Fig. 3a), we matched each of the SCENT peak–gene pairs to one non-SCENT peak–gene pair matching on TSS distance (Supplementary Fig. 3b). SCENT peaks had significantly higher conservation scores than distance-matched non-SCENT peaks (mean ΔphastCons score 0.034, $P = 4.7 \times 10^{-4}$ in arthritis-tissue dataset; Extended Data Fig. 4g and Methods), indicating the functional importance of SCENT regulatory regions not solely driven by TSS proximity.

Finally, we tested whether the target genes from SCENT were enriched for experimentally confirmed enhancer–gene links. First, we used CRISPR-Flow FISH results[39] that included 278 positive and 5,470 negative enhancer–gene connections. The SCENT peaks from tissues relevant to the experiment were significantly enriched for positive connections relative to non-SCENT peaks (for example, Fisher's exact odds ratio 4.5×, $P = 1.8 \times 10^{-9}$ in arthritis-tissue dataset; Methods, Fig. 2e and Supplementary Table 3). Second, we used H3K27ac HiChIP data in naive T cells, Th17 T cells and regulatory T cells up to 1-kb resolution[68]. Across our six T cell datasets, the SCENT peak–gene links were 1.6-fold enriched within H3K27ac HiChIP enhancer–gene loops relative to non-SCENT peaks (Fisher's exact test $P = 2.3 \times 10^{-54}$, Fig. 2f and Methods).

We anticipate that the genes with the largest number of SCENT peaks are likely to be the most constraint and least tolerant to loss of function mutations. These genes included *FOSB* ($n = 97$), *JUNB* ($n = 95$) and *RUNX1* ($n = 77$), highly conserved transcription factors. We assessed mutational constraint based on the absence of deleterious variants within human populations, including the probability of being loss-of-function intolerant (pLI)[69] and the loss-of-function observed/expected upper bound fraction (LOEUF)[70]. The normalized number of SCENT peaks per gene is strongly associated with mean constraint score for the gene ($\beta = 0.37$, $P = 4.9 \times 10^{-90}$ for pLI where higher score indicates more constraint, and $\beta = -0.35$, $P = -0.35 \times 10^{-106}$ for LOEUF where lower score indicates more constraint; Extended Data Fig. 5a,b, respectively). Our results are consistent with prior reports that genes with many regulatory regions from bulk-epigenomic data are enriched for loss-of-function intolerant genes[71].

Enrichment of eQTL putative causal variants in SCENT peaks

We examined whether the SCENT peaks harbor statistically fine-mapped putative causal variants for expression quantitative loci (eQTL). We used eQTL from GTEx across 49 tissues[72] and defined putative causal variants as those with posterior inclusion probability (PIP) >0.2. Unsurprisingly, all accessible regions defined by ATAC–seq in *cis*-regions were modestly enriched in fine-mapped variants by 2.7× (yellow, Fig. 3a). Strikingly, SCENT peaks were more enriched in fine-mapped variants by 9.6× averaged across all datasets (teal, Fig. 3a). More stringent PIP threshold cutoffs yielded stronger enrichments (Supplementary Fig. 4).

Since many SCENT peaks are close to TSS regions, we considered if the enrichment was driven by TSS proximity (Supplementary Fig. 3a). The TSS distance is one of the most important features for causal eQTL variants[73–78]. We confirmed the relationship between proximity to TSS and causal variant enrichment in GTEx (Supplementary Fig. 5a,b). We therefore tested if the higher eQTL causal variant enrichment remained (1) after excluding promoter peaks from SCENT peak–gene linkages or (2) after matching the TSS distance. Excluding promoters, SCENT still consistently had higher enrichment in all analyzed datasets than all *cis*-regulatory ATAC–seq peaks (Extended Data Fig. 6a). We observed higher enrichment in SCENT peaks compared to TSS-distance-matched non-SCENT peaks while the differences now became insignificant (Extended Data Fig. 6b). This suggests that SCENT has additional information in identifying functional *cis*-regulatory regions beyond TSS distance.

We next compared eQTL variant enrichment in SCENT peaks to peaks identified by two published linear parametric methods using single-cell multimodal data, ArchR[55] and Signac[46,50]. Statistically significant peak–gene links defined by the threshold of FDR <10% in ArchR and Signac without filtering with correlation *r* had lower causal variant enrichment (4.9× and 19.7×, respectively) compared to SCENT peaks (74.1×) with the same FDR threshold (Extended Data Fig. 6c). Given the large differences in the number of peak–gene links with FDR <10% among methods (4,330 in SCENT, 817,000 in ArchR and 9,840 in Signac on average), we were concerned that performance differences may reflect recall differences. By varying the thresholds of correlation *r* in ArchR and Signac (Methods), we called peak–gene pairs with different levels of stringency and tested causal variant enrichment (that is, recall–precision tradeoff; Fig. 3b and Extended Data Fig. 6c). SCENT peaks demonstrated higher causal variant enrichment than ArchR and Signac peaks across different recall values. Additional benchmarking with existing methods including Cicero[51] is described in Supplementary Note 2 (Extended Data Fig. 6d–i).

We assessed whether the Poisson regression or the bootstrapping in SCENT was driving the improved performance over linear parametric methods. We observed lower causal variant enrichment in peaks identified with Poisson regression alone compared to SCENT (14.4× versus 74.1× at FDR <10%, respectively; Extended Data Fig. 6h). This underscored the importance of accounting for variable gene count distribution by nonparametric bootstrapping.

SCENT can detect *cis*-regulatory regions in a cell-type-specific manner. We created cell-type-specific enhancer–gene maps in four major cell types with >5,000 cells across datasets; for each cell type we took the union of SCENT enhancers across datasets. The cell-type-specific SCENT enhancers were most enriched in eQTL variants within relevant samples in GTEx (for example, B cell SCENT peaks and eQTLs in Epstein–Barr-virus-transformed lymphocytes; Extended Data Fig. 6j).

These results showcased SCENT's prioritization of causal eQTL variants in a cell-type-specific manner with higher precision than the previous single-cell methods.

Enrichment of GWAS causal variants in SCENT enhancers

SCENT can be used to build disease-specific enhancer–gene maps by applying it to multimodal data from disease-relevant tissues. We examined whether SCENT peaks can be used to prioritize disease causal variants. We obtained candidate causal with PIP >0.2 (ref. 28) from GWASs in two large-scale biobanks, including 1,046 disease traits from FinnGen[79] and 35 binary and 59 quantitative traits from UK Biobank[80]. The aggregated SCENT enhancers were strikingly enriched in causal GWAS variants in FinnGen (31.6× on average; Fig. 3c) and UK Biobank (73.2× on average; Fig. 3d) while cell-type-specific SCENT tracks had

https://doi.org/10.1038/s41588-024-01682-1

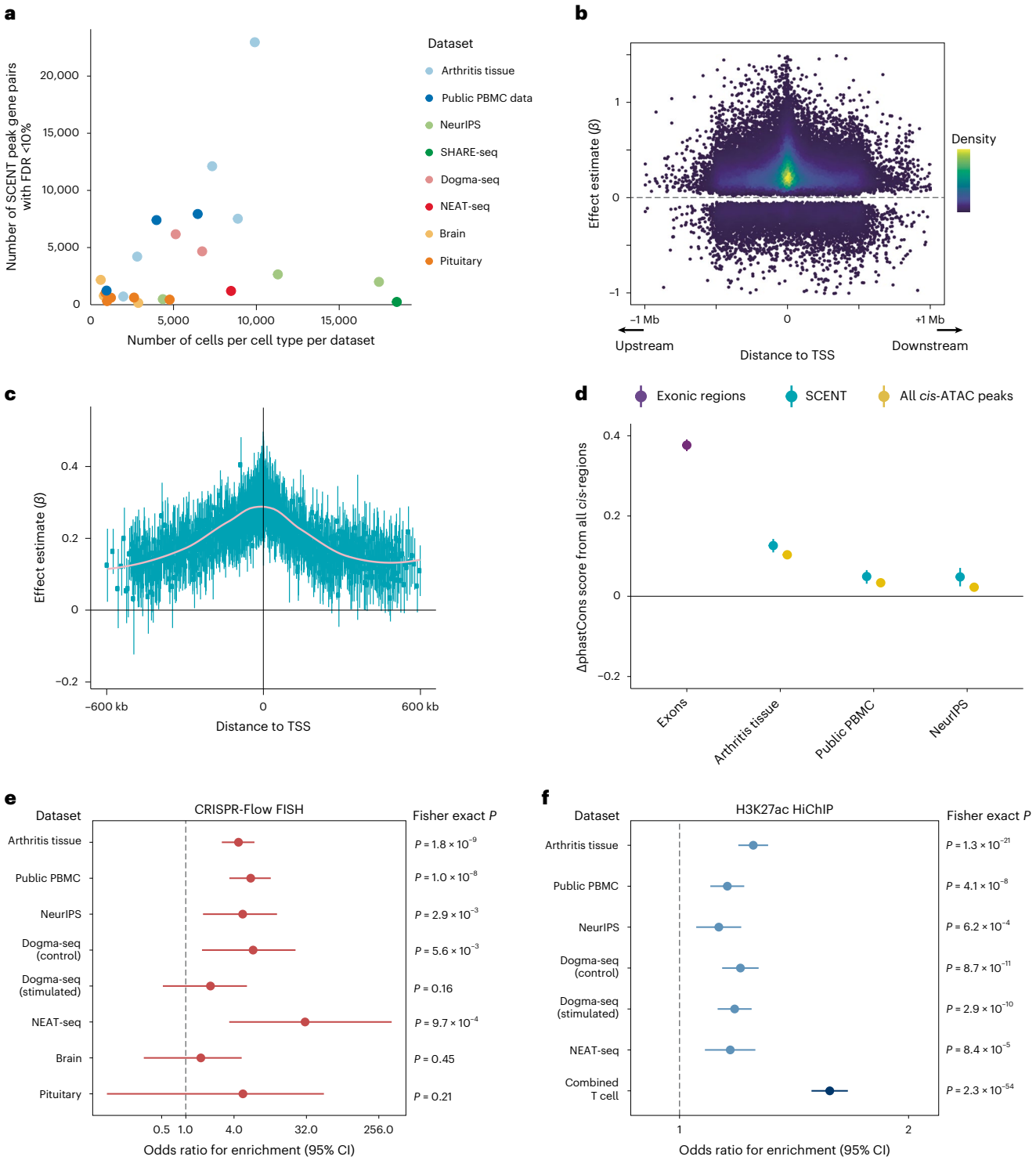


Fig. 2 | SCENT identified functionally active and evolutionarily conserved *cis*-regulatory regions from single-cell multimodal data. a, The number of significant gene–peak pairs discovered by SCENT with FDR <10%. Each dot represents the number of significant gene–peak pairs in a given cell type in a dataset (*y* axis) as a function of the number of cells in each cell type in a dataset (*x* axis), colored by the dataset. **b**, The effect size (β) of chromatin accessibility on the gene expression from Poisson regression (*y* axis). Each dot is a significant gene–peak pair (FDR <10%) and plotted against the distance between the peak and the TSS of the gene, colored as a density plot. **c**, The mean effect size (β) of chromatin accessibility on the gene expression for SCENT enhancer–gene links in arthritis-tissue dataset within each bin of TSS distance. The lines indicate 95% CIs in each bin with the bin size either 47 or 48. **d**, The mean phastCons score difference (ΔphastCons score) between each annotated region and all *cis*-regulatory noncoding regions. We show the ΔphastCons score for exonic regions (purple) as a reference, and for SCENT (teal) and all *cis*-ATAC peaks (yellow) in each multimodal dataset. The bars indicate 95% CIs by bootstrapping ($n_{bootstrap}$ = 1,000). **e**,**f**, Enrichment test for SCENT enhancer–gene links within enhancer–promoter contacts based on CRISPR-Flow FISH and H3K27ac HiChIP. We plot the odds ratio (a point estimate as a dot and 95% CI as a bar) and *P* values from the two-sided Fisher's exact enrichment test for SCENT enhancer–gene links within enhancer–gene connections based on CRISPR-Flow FISH in cell lines (**e**) and enhancer–promoter contact loops based on H3K27ac HiChIP in T cells (**f**). In **f**, light-blue rows show the results from each of the six datasets that include T cells, and the dark-blue row at the bottom show the result from combined SCENT track across the six datasets.

https://doi.org/10.1038/s41588-024-01682-1

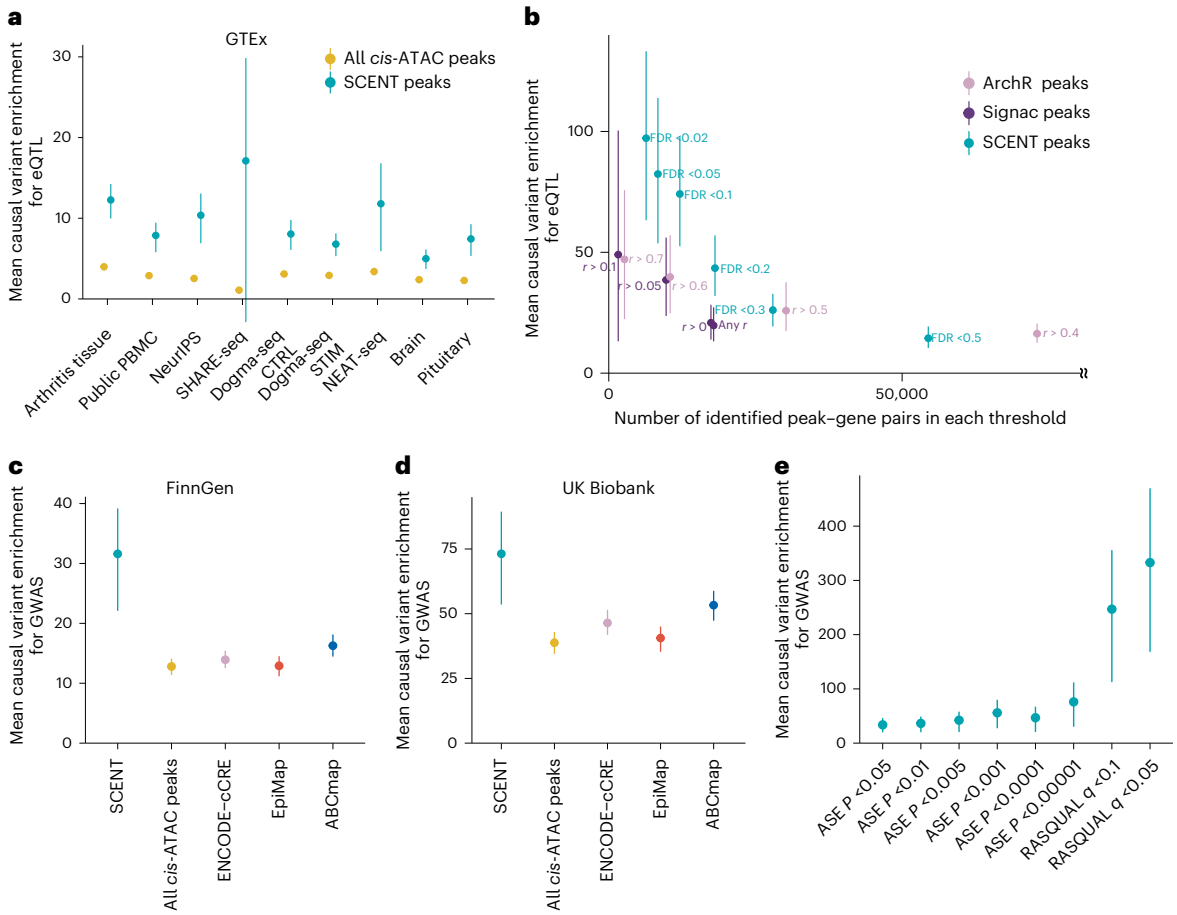


Fig. 3 | SCENT enhancers are enriched in putative causal variants of eQTL and GWAS. a, The mean causal variant enrichment for eQTL within SCENT peaks or all ATAC–seq peaks in each of the nine single-cell datasets. The bars indicate 95% CIs by bootstrapping genes ($n_{bootstrap}$ = 1,000). CTRL, control; STIM, stimulated. **b**, Comparison of the mean causal variant enrichment for eQTL (*y* axis) between SCENT (teal), ArchR (pink) and Signac (purple) as a function of the number of significant peak–gene pairs at each threshold of significance (FDR and Pearson's correlation *r*). The bars indicate 95% CIs by bootstrapping genes ($n_{bootstrap}$ = 1,000). The ArchR results with >100,000 peak–gene linkages are omitted, and full results are shown in Extended Data Fig. 6c. **c**,**d**, The mean causal variant enrichment for GWAS within SCENT enhancers (teal), all *cis*-ATAC peaks (yellow), ENCODE cCREs (pink), EpiMap enhancers across all groups (red) and ABC enhancers across all samples (blue). GWAS results were based on FinnGen (**c**) and UK Biobank (**d**). The bars indicate 95% CIs by bootstrapping traits ($n_{bootstrap}$ = 1,000). **e**, The mean causal variant enrichment for FinnGen GWAS within intersection of SCENT enhancers and caQTL enhancers at each threshold of significance (either by binomial test for allele-specific effect (ASE) followed by combining *P* values by Fisher's method or by RASQUAL). The bars indicate 95% CIs by bootstrapping traits ($n_{bootstrap}$ = 1,000).

variable enrichment (Extended Data Fig. 7a,b). This enrichment was again much larger than all *cis*-ATAC peaks (12.8× in FinnGen and 38.8× in UK Biobank). The enrichment in SCENT peaks remained higher than all peaks even after removing promoter regions or conditioning on TSS distance; in some datasets, the difference was not significant with overlapping CIs (Extended Data Fig. 7c,d). The target genes of the likely causal variants for autoimmune diseases identified by SCENT peaks in immune cell types had higher fraction (10.8%) of know genes implicated in Mendelian disorders of immune dysregulation (n_{gene} = 550)[81,82] than SCENT peaks in fibroblasts (3.8%; Extended Data Fig. 7e).

We compared SCENT to alternative genome annotations and enhancer–gene maps from bulk tissues. Causal variant enrichment in FinnGen and UK Biobank was higher in SCENT (31.6× and 72.3×, respectively) than the conventional bulk-based annotations such as ENCODE cCREs (13.9× and 46.5×), ABC (16.3× and 53.3×) and EpiMap (12.9× and 40.6×) (Fig. 3c,d and Extended Data Fig. 7a,b). We again assessed the tradeoff between recall and enrichment (precision). We constructed SCENT from 9 datasets and 23 cell types with only 28 samples, substantially less than the 833 samples and tissues used to construct EpiMap and 131 samples and cell lines for the ABC model. Despite the smaller dataset, SCENT peaks demonstrated higher enrichment of GWAS variants at a given number of identified peak–gene linkages than ABC model and EpiMap (Extended Data Fig. 8a). More stringent PIP threshold increased the enrichment (Extended Data Fig. 8b). The target genes for autoimmune disease by SCENT in immune-related cell types had higher fraction (10.8%) of known Mendelian genes of immune dysregulation[81,82] than EpiMap (8.6%) and ABC model (4.4%) (Extended Data Fig. 7e). GWAS variants were also more enriched in SCENT enhancers than ArchR and Signac (Extended Data Fig. 8c,d). These results demonstrate the benefit of accurately modeling association between chromatin accessibility and gene expression at the single-cell resolution.

We hypothesized that putative causal variants identified by SCENT modulate chromatin accessibility (for example, transcription factor binding affinity). If so, the intersection of the SCENT enhancers and chromatin accessibility quantitative trait loci (caQTL) may be further enriched for GWAS causal variants[83–86]. To test this, we used single-cell ATAC–seq samples with genotype data[58] (n_{donor} = 17; Methods) and performed caQTL mapping by leveraging allele-specific chromatin accessibility (binomial test followed by meta-analysis across donors) or by combining allele-specific with inter-individual differences (RASQUAL[87]). We observed higher enrichment within intersected regions

with SCENT and caQTL than those with SCENT alone. The enrichment increased as we used more stringent threshold for caQTL peaks, reaching as high as 333-fold (Fig. 3e). As an illustrative example, an asthma GWAS locus at 15q22.33 included a SCENT enhancer within an intron of *SMAD3* gene that harbored a putative causal variant rs17293632 (PIP 0.34; Extended Data Fig. 9a). This SCENT enhancer (Extended Data Fig. 9b) had a significant caQTL effect, from both (1) allele-specific effect (meta-analyzed binomial-test $P = 2.7 \times 10^{-4}$; Extended Data Fig. 9c) and (2) inter-individual differences in chromatin accessibility ($P = 6.0 \times 10^{-5}$; Extended Data Fig. 9d). The alternative allele T reduced the chromatin accessibility and was reported to disrupt a conserved AP-1 consensus site[30]. The allele T also decreased *SMAD3* expression ($\beta = -0.0687$, $P = 3.3 \times 10^{-13}$ from eQTL catalog[88]). *SMAD3*, the target gene identified by SCENT, is involved in TGF-β signaling, which remodels airways in asthma[89].

Together, SCENT demonstrated the potential to further enrich causal variants by integrating caQTLs.

Defining mechanisms of GWAS loci by SCENT

Finally, we sought to use SCENT to define disease causal mechanisms. We analyzed the fine-mapped variants from GWAS (FinnGen, UK Biobank and GWAS cohorts of rheumatoid arthritis[26], inflammatory bowel disease[29] and type 1 diabetes[90]). SCENT linked 4,124 putative causal variants (PIP >0.1) to their potential target genes across 1,143 traits (Supplementary Table 4). These target genes were mostly close to the causal variant, with 20% of them being the closest gene to the causal variant (Supplementary Fig. 6a,b). However, 30.6% of SCENT-linked genes were more than 300 kb away from the causal variants.

We first focused on autoimmune loci, since our SCENT tracks were largely derived from immune cell types. We prioritized a single fine-mapped variant rs72928038 (PIP >0.3) within the T-cell-specific SCENT enhancer at a locus in 6q15 for multiple autoimmune diseases (rheumatoid arthritis, type 1 diabetes, atopic dermatitis and hypothyroidism; Fig. 4a). This enhancer was linked to *BACH2*, the closest gene to this fine-mapped variant. Base-editing the protective allele to the risk allele in T cells has confirmed the effect of this variant on *BACH2* expression[91]. Moreover, rs72928038-deleted naive CD8 T cells were more prone to differentiate into effector T cells in mice[91].

A 4p15.2 locus for rheumatoid arthritis and type 1 diabetes harbored 21 candidate variants, each with low PIPs (<0.14). SCENT prioritized a single variant rs35944082 in T cells and fibroblasts only within the arthritis-tissue dataset from inflamed synovial tissue (Fig. 4b). SCENT linked this variant to *RBPJ*, which was the third closest gene located 235 kb away. This variant-gene link was supported by promoter-capture Hi-C data in hematopoietic cells[92] and by H3K27ac HiChIP data in T cells[68]. The *RBPJ* transcription factor is critical for NOTCH signaling, which has been implicated in rheumatoid arthritis tissue inflammation through functional studies[93,94]. *Rbpj* knockdown in mice resulted in abnormal T cell differentiation and disrupted regulatory T cell phenotype[95,96], consistent with a plausible role in autoimmune diseases. Intriguingly, we did not observe this enhancer–gene link in T cells from peripheral blood mononuclear cells (PBMC), blood nor in EpiMap. ABC map prioritized another variant, rs7441808, at this locus and linked it nonspecifically to 16 genes including *RBPJ*, making it difficult to define the causal gene. The Hi-C and H3K27ac HiChIP data nominated the gene *RBPJ*, but due to the limited resolution of the contact maps, they could not prioritize a single causal variant.

As an example of the power of SCENT to build enhancer–gene maps from disease-critical tissues, we examined single-cell data from pituitary[63]. We assessed a 11p14.1 locus for multiple gynecological traits (endometriosis, menorrhagia, ovarian cyst and age at menopause). Our map connected rs11031006 to *FSHB* (Fig. 4c), which is specifically expressed in the pituitary[72,97] and enables ovarian folliculogenesis[98]. Rare *FSHB* variants cause autosomal recessive hypogonadotropic hypogonadism[99]. Multimodal data from other tissues and bulk-based methods were unable to prioritize this variant since they missed pituitary, the most disease-relevant tissue. These results underscored the importance of creating enhancer–gene links using causal cell types, in the instances where links exist only in disease-relevant tissues.

Rare disease variants and somatic mutations within SCENT enhancers

Having established SCENT's utility in defining causal variants and genes in complex diseases, we examined rare noncoding variants causing Mendelian diseases. Currently, causal mutations can be identified in only ~30–40% of patients with Mendelian diseases[100–102]. Consequently, many variants in cases are annotated as variants of uncertain significance. The variants of uncertain significance annotation is especially challenging for noncoding variants. We examined the overlap of clinically reported nonbenign noncoding variants by ClinVar[103] (400,300 variants in total) within SCENT enhancers. The SCENT enhancers harbored 2.0× ClinVar variants on average than the ATAC regions with the same genomic length (Supplementary Fig. 7). This density of ClinVar variants was 3.2× and 12× larger than that in ENCODE cCREs and all noncoding regions, respectively. We defined 3,724 target genes for 33,618 noncoding ClinVar variants (Supplementary Table 5). As illustrative examples, we found 40 noncoding variants linked to *LDLR* gene causing familial hypercholesterolemia 1 (ref. 103), 3 noncoding variants linked to *IL10RA* causing autosomal recessive early-onset inflammatory bowel disease 28 (Fig. 4d)[104], and an intronic variant rs1591491477 linked to *ATM* gene causing hereditary cancer-predisposing syndrome[103].

We also used SCENT to connect noncoding somatic mutation hotspots to target genes. Recently, somatic mutation analyses across the entire cancer genome revealed possible driver noncoding events[105]. Among 17 noncoding mutation hotspots in leukemia, SCENT enhancers from blood-related cell types included 12 hotspots (Supplementary Table 6). SCENT enhancer–gene linkage linked those hotspots to known driver genes (for example, *BACH2*, *BCL6*, *BCR*, *CXCR4* (Fig. 4e) and *IRF8* in leukemia). In some instances, SCENT nominated different target genes for these mutation hotspots from those based on ABC model used in the original study. For example, SCENT connected a somatic mutation hotspot in leukemia at chr14:105568663-106851785 to immunoglobulin heavy chain related genes such as *IGHA1*, which might be more biologically relevant than *ADAM6* nominated by ABC model. These results implicate broad applicability of SCENT for annotating human variations in noncoding regions.

Augmenting SCENT enhancer–gene maps with more samples and cells

While the number for enhancer–gene links by SCENT was smaller than that by bulk-tissue-based methods, this might be a function of current limited sample sizes. By downsampling of our multimodal single cell dataset, we observed that the number of significant gene–peak pairs increased linearly to the number of cells per cell type in a given dataset, suggesting that SCENT will be even better powered as the size of multimodal datasets increases (Supplementary Fig. 8).

In SCENT association, we used cells from a specific cell type to identify cell-type-specific gene regulation. While association across cells from different cell types might increase the number of significant peak–gene linkages due to greater variance in chromatin accessibility and gene expression, this strategy could yield false-positive enhancer–gene associations by increasing the chances of connecting enhancers that are merely 'correlated' with gene expression (Extended Data Fig. 10a). By simulation and real data analyses, we confirmed that the cell-type-specific analysis was better calibrated to reject false enhancer–gene links by correlation and powered to detect experimentally validated enhancer–gene links, while it was less powered to detect promoter–gene links when compared with multiple-cell-type analysis (Supplementary Note 3 and Extended Data Fig. 10b–e). We anticipate that we might be able to obtain even higher signal-to-noise ratio with

https://doi.org/10.1038/s41588-024-01682-1

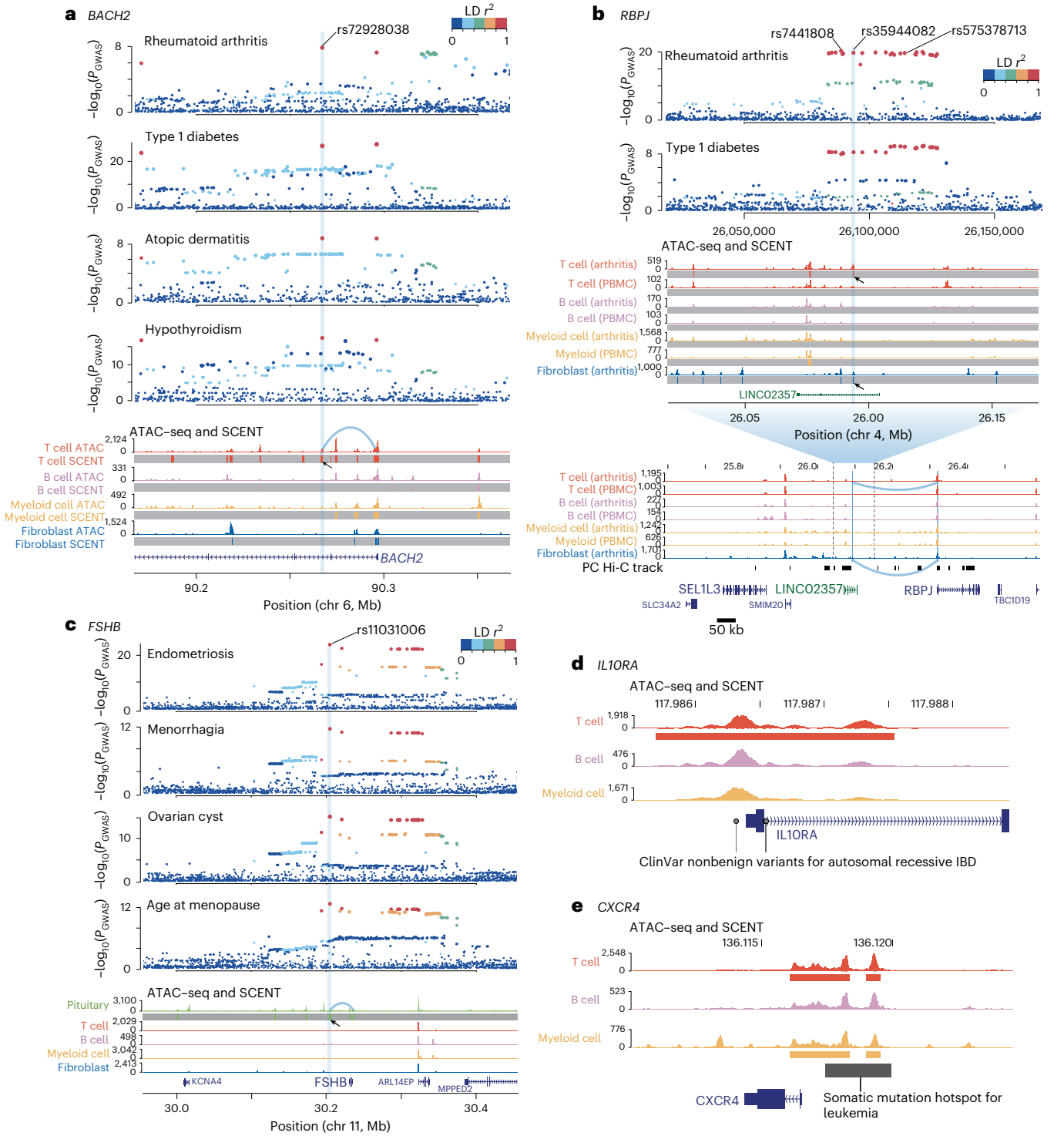


Fig. 4 | SCENT defined causal variants and genes in complex trait GWAS. **a**, rs72928038 at *BACH2* locus was prioritized by T-cell-specific SCENT enhancer–gene map for rheumatoid arthritis (RA), type 1 diabetes (T1D), atopic dermatitis and hypothyroidism. The top four panels are GWAS regional plots, with *x* axis representing the position of each genetic variant and *y* axis representing $-\log_{10}(P)$ from GWAS (two-sided *P* values). The color of the dots represents linkage disequilibrium (LD) r^2 from the prioritized variant (highlighted by light-blue stripe). ATAC–seq and SCENT tracks represent aggregated ATAC–seq tracks (top) and SCENT peaks (bottom with gray stripes) in each cell type (public PBMC dataset for immune cell types and arthritis-tissue dataset for fibroblast). An arrow indicates the SCENT peak overlapping with fine-mapped variant. LD, linkage disequilibrium. **b**, rs35944082 for RA and T1D was prioritized and connected to *RBPJ* by long-range interaction from T cell and fibroblast SCENT enhancer–gene map using inflamed synovium in the arthritis-tissue dataset. The top two panels are GWAS regional plots similarly to **a**. ATAC–seq and SCENT tracks represent aggregated ATAC–seq tracks (top) and SCENT peaks (bottom with gray stripes) using both public PBMC and arthritis-tissue datasets. **c**, rs11031006 was prioritized and connected to *FSHB* for multiple gynecological traits by using pituitary-derived single-cell multimodal dataset. The top four panels are GWAS regional plots similarly to **a**. ATAC–seq and SCENT tracks represent aggregated ATAC–seq tracks (top) and SCENT peaks (bottom with gray stripes). There were no SCENT peaks in cell types except for pituitary. **d**, ATAC–seq (top) and SCENT tracks (bottom) for *IL10RA* locus, where noncoding ClinVar variants (gray dots) colocalized with T cell SCENT track. **e**, ATAC–seq (top) and SCENT tracks (bottom) for *CXCR4* locus, where somatic mutation hotspot for leukemia colocalized with T cell and myeloid cell SCENT tracks.

more fine-grained cell-state-specific analyses in the future as better powered datasets with more cells become available; we conjecture that in these datasets there will be fewer false negatives while maintaining the high precision of enhancer–gene calls.

Discussion

We presented a statistical method, SCENT, to create a cell-type-specific enhancer–gene map from single-cell multimodal data. Single-cell RNA-seq and ATAC–seq are both sparse and have variable count distributions, which requires nonparametric bootstrapping to connect chromatin accessibility with gene expression. The SCENT model demonstrated well-controlled type I error, outperforming commonly used statistical models. SCENT mapped enhancers that showed high enrichment for putative causal variants in eQTLs and GWAS and outperformed previous methods. Despite using substantially fewer samples, enhancers defined by SCENT had equivalent or higher enrichment for causal variants than bulk-tissue-based methods with many samples. SCENT benefits from modeling at the single-cell level instead of obscuring associations by aggregating cells into individual samples.

As limitations, first, SCENT enhancer–gene maps had relatively fewer enhancers compared to other resources (Fig. 2a). However, a linear relationship between the number of cells and the number of significant SCENT peak–gene links (Supplementary Fig. 8) indicates that application of SCENT to larger datasets will expand the current enhancer–gene map. In contrast, bulk-tissue-based enhancer–gene map might be more challenging to expand given the number of samples required. Second, SCENT focuses on gene *cis*-regulatory mechanisms to fine-map disease causal alleles. However, there could be other causal mechanisms, such as alleles that act through *trans*-regulatory effects, splicing effects or post-transcriptional effects[106]. Third, to prove the causality of the alleles prioritized by SCENT, experimental validation using gene editing technology[107–109] is necessary. Fourth, due to Poisson regression and bootstrapping, SCENT is more computationally intensive than the previous methods (for example, 1.5×10^7 CPU seconds in SCENT, 2.5×10^5 in Signac, and 2.2×10^2 in ArchR; Supplementary Table 7). We implemented multi-threading and parallelization options in SCENT, which lead to linearly faster computation but at the cost of additional computational resources. Algorithmic improvements, such as downsampling or aggregating cells, may be useful for extremely large datasets.

We argue that the real utility of SCENT is that it enables the construction of disease-tissue-relevant enhancer–gene maps. Multimodal single-cell data can be obtained from a wide range of primary human tissues. It can be applied to those that are difficult to disaggregate since these multimodal data can be obtained from nuclear material without tissue disaggregation. Therefore, it is possible to build relevant tissue-specific enhancer–gene maps that are necessary to understand disease causal mechanisms. For example, understanding the *FSHB* locus in gynecological traits specifically required a pituitary map, and *RBPJ* locus in rheumatoid arthritis specifically required a synovial tissue map.

In summary, SCENT is a robust, versatile method to define causal variants and genes in human diseases.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01682-1.

References

1. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
2. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet* **101**, 5–22 (2017).
3. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
4. Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
5. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov.* **12**, 581–594 (2013).
6. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic medicine—progress, pitfalls, and promise. *Cell* **177**, 45–57 (2019).
7. 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).
8. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
9. Edwards, S. L., Beesley, J., French, J. D. & Dunning, M. Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet* **93**, 779–797 (2013).
10. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
11. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
12. Smemo, S. et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371–375 (2014).
13. Won, H. et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).
14. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
15. Cuomo, A. S. E. et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).
16. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
17. Nathan, A. et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature* **606**, 120–128 (2022).
18. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
19. Maller, J. B. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
20. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
21. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
22. 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. Ser. B* **82**, 1273–1300 (2020).
23. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
24. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
25. Chen, M. H. et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213.e14 (2020).

26. Ishigaki, K. et al. Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).
27. Kichaev, G. & Pasaniuc, B. Leveraging functional-annotation data in trans-ethnic fine-mapping studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).
28. Kanai, M. et al. Insights from complex trait fine-mapping across diverse populations. Preprint at *medRxiv* https://doi.org/10.1101/2021.09.03.21262975 (2021).
29. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
30. Farh, K. K. H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2014).
31. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
32. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
33. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
34. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414.e24 (2016).
35. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
36. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
37. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
38. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664 (2019).
39. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
40. Gazal, S. et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
41. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
42. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
43. Baglaenko, Y., Macfarlane, D., Marson, A., Nigrovic, P. A. & Raychaudhuri, S. Genome editing to define the function of risk loci and variants in rheumatic disease. *Nat. Rev. Rheumatol.* **17**, 462–474 (2021).
44. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
45. 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).
46. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
47. Allaway, K. C. et al. Genetic and epigenetic coordination of cortical interneuron development. *Nature* **597**, 693–697 (2021).
48. Trevino, A. E. et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069.e23 (2021).
49. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
50. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
51. Pliner, H. A. et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871.e8 (2018).
52. Lähnemann, D. et al. Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
53. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat. Genet.* **53**, 770–777 (2021).
54. Chen, H. et al. Assessment of computational methods for the analysis of single-cell ATAC–seq data. *Genome Biol.* **20**, 1–25 (2019).
55. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
56. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. *Genome Biol.* **20**, 1–16 (2019).
57. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, 1994).
58. Weinand, K. et al. The chromatin landscape of pathogenic transcriptional cell states in rheumatoid arthritis. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.07.536026 (2023).
59. Luecken, M. D. et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (NeurIPS, 2021).
60. Mimitou, E. P. et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
61. Chen, A. F. et al. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nat. Methods* **19**, 547–553 (2022).
62. Meijer, M. et al. Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. *Neuron* **110**, 1193–12 (2022).
63. Zhang, Z. et al. Single nucleus transcriptome and chromatin accessibility of postmortem human pituitaries reveal diverse stem cell regulatory mechanisms. *Cell Rep.* https://doi.org/10.1016/J.CELREP.2022.110467 (2022).
64. Abascal, F. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
65. Westra, H. J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* **1842**, 1896–1902 (2014).
66. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
67. Hujoel, M. L. A., Gazal, S., Hormozdiari, F., van de Geijn, B. & Price, A. L. Disease heritability enrichment of regulatory elements is concentrated in elements with ancient sequence age and conserved function across species. *Am. J. Hum. Genet* **104**, 611–624 (2019).
68. Mumbach, M. R. et al. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).
69. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
70. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
71. Wang, X. & Goldstein, D. B. Enhancer domains predict gene pathogenicity and inform gene discovery in complex disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).

72. Aguet, F. et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
73. 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).
74. Zou, J. et al. Leveraging allelic imbalance to refine fine-mapping for eQTL studies. *PLoS Genet.* **15**, e1008481 (2019).
75. Chen, W., McDonnell, S. K., Thibodeau, S. N., Tillmans, L. S. & Schaid, D. J. Incorporating functional annotations for fine-mapping causal variants in a Bayesian framework using summary statistics. *Genetics* **204**, 933–958 (2016).
76. Gaffney, D. J. et al. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol.* **13**, R7 (2012).
77. Göring, H. H. H. et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat. Genet.* **39**, 1208–1216 (2007).
78. Wen, X., Luca, F. & Pique-Regi, R. Cross-population joint analysis of eQTLs: fine mapping and functional annotation. *PLoS Genet.* **11**, e1005176 (2015).
79. Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
80. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
81. Dey, K. K. et al. SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell Genomics* **2**, 100145 (2022).
82. Freund, M. K. et al. Phenotype-specific enrichment of mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).
83. Gate, R. E. et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* **50**, 1140–1150 (2018).
84. Khetan, S. et al. Type 2 diabetes-associated genetic variants regulate chromatin accessibility in Human Islets. *Diabetes* **67**, 2466–2477 (2018).
85. Alasoo, K. et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.* **50**, 424–431 (2018).
86. Currin, K. W. et al. Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am. J. Hum. Genet.* **108**, 1169–1189 (2021).
87. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC–seq. *Nat. Genet.* **48**, 206–213 (2015).
88. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
89. Sagara, H. et al. Activation of TGF-β/Smad2 signaling is associated with airway remodeling in asthma. *J. Allergy Clin. Immunol.* **110**, 249–254 (2002).
90. Chiou, J. et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).
91. Mouri, K. et al. Prioritization of autoimmune disease-associated genetic variants that perturb regulatory element activity in T cells. *Nat. Genet.* **54**, 603–612 (2022).
92. 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).
93. Radtke, F., Fasnacht, N. & MacDonald, H. R. Notch signaling in the immune system. *Immunity* **32**, 14–27 (2010).
94. Wei, K. et al. Notch signalling drives synovial fibroblast identity and arthritis pathology. *Nature* **582**, 259–264 (2020).
95. Delacher, M. et al. Rbpj expression in regulatory T cells is critical for restraining TH2 responses. *Nat. Commun.* **10**, 1621 (2019).
96. Blake, J. A. et al. Mouse Genome Database (MGD): knowledgebase for mouse–human comparative biology. *Nucleic Acids Res.* **49**, D981–D987 (2021).
97. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
98. Hillier, S. G. Gonadotropic control of ovarian follicular growth and development. *Mol. Cell. Endocrinol.* **179**, 39–46 (2001).
99. Rubinstein, W. S. et al. The NIH genetic testing registry: a new, centralized database of genetic tests to enable access to comprehensive information and improve transparency. *Nucleic Acids Res.* **41**, D925–D935 (2013).
100. Retterer, K. et al. Clinical application of whole-exome sequencing across clinical indications. *Genet. Med.* **18**, 696–704 (2016).
101. Adams, D. R. & Eng, C. M. Next-generation sequencing to diagnose suspected genetic disorders. *N. Engl. J. Med.* **379**, 1353–1362 (2018).
102. Srivastava, S. et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet. Med.* **21**, 2413–2421 (2019).
103. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
104. Glocker, E.-O. et al. Inflammatory bowel disease and mutations affecting the interleukin-10 receptor. *N. Engl. J. Med.* **361**, 2033–2045 (2009).
105. Dietlein, F. et al. Genome-wide analysis of somatic noncoding mutation patterns in cancer. *Science* **376**, eabg5601 (2022).
106. Connally, N. et al. The missing link between genetic association and regulatory function. *eLife* **11**, e74970 (2022).
107. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
108. 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).
109. Morris, J. A. et al. Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science* **380**, eadh7699 (2023).

[1]Center for Data Sciences, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [2]Divisions of Genetics and Rheumatology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [5]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [6]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [7]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Center for Computational and Integrative Biology, Massachusetts General Hospital, Boston, MA, USA. [9]Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [10]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA. [11]Hospital for Special Surgery, New York, NY, USA. [12]Weill Cornell Medicine, New York, NY, USA. [13]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: soumya@broadinstitute.org

Accelerating Medicines Partnership® RA/SLE Program and Network

Jennifer Albrecht[14], Jennifer H. Anolik[14], William Apruzzese[15], Nirmal Banda[16], Jennifer L. Barnas[14], Joan M. Bathon[17], Ami Ben-Artzi[18], Brendan F. Boyce[19], David L. Boyle[20], S. Louis Bridges Jr[11,12], Vivian P. Bykerk[11,12], Debbie Campbell[14], Hayley L. Carr[21], Arnold Ceponis[20], Adam Chicoine[9], Andrew Cordle[22], Michelle Curtis[1,2,3], Kevin D. Deane[23], Edward DiCarlo[24], Patrick Dunn[25,26], Andrew Filer[21], Gary S. Firestein[20], Lindsy Forbess[21], Laura Geraldino-Pardilla[17], Susan M. Goodman[11,12], Ellen M. Gravallese[9], Peter K. Gregersen[27], Joel M. Guthridge[28], Maria Gutierrez-Arcelus[1,2,3,29], Siddarth Gurajala[1,2,3], V. Michael Holers[23], Diane Horowitz[27], Laura B. Hughes[30], Kazuyoshi Ishigaki[1,2,3,31], Lionel B. Ivashkiv[11,12], Judith A. James[28], Anna Helena Jonsson[9], Joyce B. Kang[1,2,3,4], Gregory Keras[9], Ilya Korsunsky[1,2,3], Amit Lakhanpal[11,12], James A. Lederer[32], Zhihan J. Li[9], Yuhong Li[9], Katherine P. Liao[4,9], Arthur M. Mandelin II[33], Ian Mantel[11,12], Mark Maybury[21], Joseph Mears[1,2,3], Nida Meednu[14], Nghia Millard[1,2,3,4], Larry W. Moreland[23,34], Aparna Nathan[1,2,3,4], Alessandra Nerviani[35], Dana E. Orange[11,36], Harris Perlman[33], Costantino Pitzalis[35], Javier Rangel-Moreno[14], Deepak A. Rao[9], Karim Raza[21], Yakir Reshef[1,2,3], Christopher Ritchlin[14], Felice Rivellese[35], William H. Robinson[37], Laurie Rumker[1,2,3,4], Ilfita Sahbudin[21], Jennifer A. Seifert[23], Kamil Slowikowski[4,38,39], Melanie H. Smith[11], Darren Tabechian[14], Dagmar Scheel-Toellner[21], Paul J. Utz[37], Dana Weisenfeld[9], Michael H. Weisman[18,37], Qian Xiao[1,2,3] & Fan Zhang[1,2,3,40]

[14]Division of Allergy, Immunology and Rheumatology, Department of Medicine, University of Rochester Medical Center, Rochester, NY, USA. [15]Accelerating Medicines Partnership® Program: Rheumatoid Arthritis and Systemic Lupus Erythematosus (AMP® RA/SLE) Network, Boston, MA, USA. [16]Division of Rheumatology., University of Colorado School of Medicine, Aurora, CO, USA. [17]Division of Rheumatology, Columbia University College of Physicians and Surgeons, New York, NY, USA. [18]Division of Rheumatology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [19]Department of Pathology and Laboratory Medicine, University of Rochester Medical Center, Rochester, NY, USA. [20]Division of Rheumatology, Allergy and Immunology, University of California, San Diego, La Jolla, CA, USA. [21]Rheumatology Research Group, Institute for Inflammation and Ageing, University of Birmingham, NIHR Birmingham Biomedical Research Center and Clinical Research Facility, University of Birmingham, Queen Elizabeth Hospital, Birmingham, UK. [22]Department of Radiology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA. [23]Division of Rheumatology, University of Colorado School of Medicine, Aurora, CO, USA. [24]Department of Pathology and Laboratory Medicine, Hospital for Special Surgery, New York, NY, USA. [25]Division of Allergy, Immunology, and Transplantation, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD, USA. [26]Northrop Grumman Health Solutions, Rockville, MD, USA. [27]Feinstein Institute for Medical Research, Northwell Health, Manhasset, New York, NY, USA. [28]Department of Arthritis & Clinical Immunology, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [29]Division of Immunology, Department of Pediatrics, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. [30]Division of Clinical Immunology and Rheumatology, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA. [31]Laboratory for Human Immunogenetics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [32]Department of Surgery, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [33]Division of Rheumatology, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [34]Division of Rheumatology and Clinical Immunology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [35]Centre for Experimental Medicine and Rheumatology, William Harvey Research Institute, Queen Mary University of London, London, UK. [36]Laboratory of Molecular Neuro-Oncology, The Rockefeller University, New York, NY, USA. [37]Division of Immunology and Rheumatology, Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, CA, USA. [38]Center for Immunology and Inflammatory Diseases, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [39]MGH Cancer Center, Boston, MA, USA. [40]Division of Rheumatology and the Center for Health Artificial Intelligence, University of Colorado School of Medicine, Aurora, CO, USA.

Methods

Ethics statement

This study complies with all relevant ethical regulations as outlined and approved by the institutional review board of Mass General Brigham (protocol approval number 2021P001846) and the institutional review board of the Hospital for Special Surgery (#2014-233). Written informed consent was obtained from all study participants.

Data and sample in arthritis-tissue dataset

Synovial tissues from patients with rheumatoid arthritis and osteoarthritis were collected from synovectomy or arthroplasty procedures followed by cryopreservation[110]. For rheumatoid arthritis samples, we examined histologic sections of synovial tissue and selected samples with inflammatory features. Sex and age of participants can be found in Supplementary Table 8. Next, cryopreserved synovial tissue fragments were dissociated by a mechanical and enzymatic digestion[110], followed by flow sorting to enrich for live synovial cells. For each tissue sample, the viable cells were isolated and lysed to extract and load approximately 10,000 nuclei according to manufacturer protocol (10x Genomics). Joint single-cell (sc)RNA-seq and scATAC–seq libraries were prepared using the 10x Genomics Single Cell Multiome ATAC + Gene Expression kit according to manufacturer's instructions. Libraries were sequenced with paired-end reads on an Illumina Novaseq to a target depth of 30,000 read pairs per nucleus both for messenger RNA and for ATAC libraries. scRNA-seq fastq files were inputted into the Cell Ranger ARC pipeline (version 2.0.0) from 10x Genomics to generate barcoded count matrix of gene expression. For the scATAC–seq fastq files, we used Cell Ranger ARC to process barcodes and to map the reads to the hg38 genome by BWA-MEM with default parameters. To deduplicate reads from polymerase chain reaction amplification bias within a cell while keeping reads originating from the same positions but from different cells, we used in-house scripts. More detailed information on data and QC steps is described in ref. 58.

Data acquisition and QC of single-cell multimodal datasets

In addition to our arthritis-tissue multimodal dataset, we downloaded all publicly available multimodal RNA-seq/ATAC–seq datasets from adult human tissues ($n_{\text{dataset}} = 9$, as of April 2022). We processed the downloaded matrices of RNA-seq and ATAC–seq data and fragment files if available by using Signac (version 1.9.0), without re-aligning the original reads to reference genome due to lack of availability of raw sequence data. We applied QC to both the nuclear RNA data and the ATAC data based on RNA counts, ATAC fragments, nucleosome signal and TSS enrichment. We defined the QC threshold based on the distribution of these metrics in each of the datasets as described in Supplementary Table 9. We only kept cells that had passed QC in both RNA-seq and ATAC–seq. Then to identify open chromatin regions (peaks), we used macs2 (version 2.2.7.1) to call open chromatin peaks using post-QC ATAC–seq data. We thus obtained count matrices of gene expression and ATAC peaks with corresponding cell barcodes. When cell barcodes after the QC from original publications were available or when fragment files are unavailable, we used the downloaded post-QC matrices for downstream analyses. Gene expression counts were normalized using the NormalizeData function (Seurat[111] version 4.3.0), scaled using the ScaleData function (Seurat), and batch corrected using Harmony[112] (version 0.1.1). We visualized the cells in two low-dimensional embeddings with uniform manifold approximation and projection (UMAP) by using 20 batch-corrected principal components from these normalized gene expression matrices (Fig. 1c). When original cell labels are provided by the authors, we used those labels to obtain broad cell type categories. When they are not available, we performed reference-query mapping by Seurat and PBMC reference object to define broad cell type labels. ATAC peak matrix was binarized to have 1 if a count is >0 and 0 otherwise.

SCENT method

We defined *cis*-peaks as any peaks whose center is within the window ±500 kb from a given gene body. We modeled the association between peak's binarized accessibility and the target gene's expression with Poisson distribution:

$$E_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \beta_0 + \beta_{\text{peak}} X_{\text{peak}} + \beta_{\%\text{mito}} X_{\%\text{mito}} + \beta_{n\text{UMI}} X_{n\text{UMI}} + \beta_{\text{batch}} X_{\text{batch}} \quad (1)$$

where E_i is the observed expression count of ith gene, and λ_i is the expected count under Poisson distribution. β_{peak} indicates the effect of chromatin accessibility of a peak on ith gene expression. $\beta_{\%\text{mito}}$, $\beta_{n\text{UMI}}$ and β_{batch} each represents the effect of covariates, percentage of mitochondrial reads per cell as a measure of cell quality, the number of unique molecular identifiers (UMIs) in the cell, and the batch, respectively. To empirically assess error and significance of β_{peak} for each peak–gene combination, we used bootstrapping procedures. We resampled cells with replacement in each bootstrapping procedure and re-estimated β'_{peak} within those resampled cells. We repeated this procedure N times, where we adaptively increased N (that is, the total number of bootstrapping) from at least 100 and up to 50,000, depending on the significance of β_{peak} in each chunk of bootstrapping trials to reduce the computational burden. After N times of bootstrapping, we assessed the distribution of N β'_{peak}s against null hypothesis ($\beta'_{\text{peak}} = 0$) to derive the significance of β_{peak} (that is, two-sided bootstrapping-based P value for this peak–gene combination by counting the instances where the statistics are equal or more extreme than the null hypothesis of $\beta'_{\text{peak}} = 0$; Extended Data Fig. 2).

To avoid spurious associations from rare ATAC peak and rare gene expression, we quality controlled the *cis*-peak–gene pairs we test so that both peak and gene should have been expressed in at least 5% of the cells we analyze. We finally defined a set of significant peak–gene pairs for each cell type based on bootstrapping-based P values and FDR correction for multiple testing (Benjamini–Hochberg correction).

When we tested the calibration of statistics from SCENT or other regression strategies (Extended Data Fig. 1), we used null dataset where we randomly permuted cell labels in the ATAC–seq and ran the regression model we tested.

ArchR peak2gene and Signac LinkPeaks method

We also analyzed single-cell multimodal datasets with ArchR[49] (version 1.0.2) and Signac[46,50] (version 1.9.0), which both have a function to define peak–gene linkages. In brief, ArchR takes multimodal data and creates low-overlapping aggregates of single cells based on k-nearest neighbor graph. Then it correlates peak accessibility with gene expression by Pearson correlation of aggregated and $\log_2$-normalized peak count and gene count. Signac computes the Pearson or Spearman correlation coefficient r (corSparse function in R) for each gene and for each peak within 500 kb of the gene TSS. Signac then compares the observed correlation coefficient with an expected correlation coefficient for each peak given the guanine–cytosine content, accessibility and length of the peak. Signac defines P value for each gene–peak links from the z score based on this comparison. We ran both methods on arthritis-tissue dataset with default parameters. We output statistics for all peak–gene pairs we tested without any cutoff for correlation r or P values. To obtain significant peak–gene linkages in ArchR and Signac that are comparable with those defined by SCENT with FDR <10%, we used the FDR value in the output from the peak2gene function of ArchR software and selected any linkages with FDR <10% as a significant set of linkages. Because Signac does not directly output FDR values, we computed FDR using P values in the output from Signac LinkPeaks function (either method = 'pearson' or method = 'spearman') by Benjamini–Hochberg correction and used this FDR to define a significant set of linkages with FDR <10%. We thus defined significant peak–gene linkages as those with FDR <10% and used varying correlation r to assess the precision and recall in the causal variant enrichment analysis (see later sections in Methods).

Because ArchR requires fragment files from ATAC–seq to run the basic functions, we ran ArchR and Signac for eight multimodal datasets in which fragment files are available.

Replication across datasets

Since we have the same immune-related cell types across different multimodal datasets, we evaluated the concordance of enhancer–gene map in a discovery dataset (arthritis-tissue dataset) when compared with other replication datasets including immune-related cell types (Public PBMC, NeurIPS, SHARE-seq and NEAT-seq datasets). To this end, we used the most stringent FDR threshold for defining an enhancer–gene map in arthritis-tissue dataset that had the largest number of significant peak–gene linkages (FDR <1%). We then used more lenient threshold for defining an enhancer–gene map in replication datasets (FDR <10%), which is a similar strategy used in assessing replication in GWAS. For each cell type and for each replication dataset, we took the intersection of enhancer-gene links defined as significant in both datasets. We assessed the directional concordance (that is, concordance of the sign of β_{peak}) and the Pearson's correlation r of β_{peak} between the discovery and the replication for these peak–gene pairs. We then performed the same analysis for enhancer-gene map from ArchR and Signac software.

Conservation score analysis

To compare the evolutional conservation across species between our annotated peaks and the other peaks, we used phastCons[66] score. We downloaded the phastCons score for multiple alignments of 99 vertebrate genomes from ref. 113. We lifted them over to GRCh38 by LiftOver software (version 2016). We used SCENT results for arthritis-tissue, Public PBMC and NeurIPS for conservation score analysis as representative datasets with the largest numbers of cells. Because each gene should have variable functional importance and conservation, we assessed each gene separately. For each gene, we took (1) an annotation of interest for the gene and (2) all *cis*-noncoding regions (<500 kb from a gene), and computed the mean phastCons score of each of two sets of the peaks. As annotations to be tested, we used (a) exonic regions of the gene, (b) SCENT peaks for the gene and (c) all ATAC peaks in *cis*-regions from the gene (<500 kb). Then, we took the difference between two mean differences (ΔphastCons score), and computed the mean differences across all the genes (mean ΔphastCons score) as follows:

$$\text{mean ΔphastCons score} = \frac{1}{n_{\text{gene}}} \sum_{\text{gene}} \left(\overline{\text{phastCons}}_{g,\text{in\_annot}} - \overline{\text{phastCons}}_{g,\text{noncoding}} \right).$$

By bootstrapping the genes, we calculated the 95% CI of the mean ΔphastCons score. If this metric is positive, that indicates that the annotated regions are more conserved than noncoding regions.

We also calculated similar ΔphastCons score by comparing the SCENT peaks with TSS-distance-matched non-SCENT peaks in each dataset.

$$\text{mean ΔphastCons score} = \frac{1}{n_{\text{gene}}} \sum_{\text{gene}} \left(\overline{\text{phastCons}}_{g,\text{peak\_in\_SCENT}} - \overline{\text{phastCons}}_{g,\text{peak\_non\_SCENT\_matched}} \right)$$

By bootstrapping the genes, we again calculated the 95% CI of the mean ΔphastCons score. If this metric is positive, that indicates that SCENT peaks are more conserved than TSS-distance-matched non-SCENT peaks.

Construction of a set of TSS-matched non-SCENT peaks

To assess the effect of TSS distance when comparing SCENT peaks with non-SCENT peaks, we matched each one of the SCENT peak–gene pairs to one non-SCENT peak–gene pair, where the peak had the most similar TSS distance to the same gene among all the ATAC peaks in *cis* in each of the dataset. We confirmed that the resulting TSS-distance-matched non-SCENT peak–gene pairs demonstrated the similar distributions of TSS distance when compared with the SCENT peak–gene pairs (Supplementary Fig. 3b).

Gene's constraint and the number of significant SCENT peaks for a gene

We sought to investigate the relationship between the number of significant SCENT peaks for each gene and the gene's evolutionary constraint. We used pLI (the probability of being loss-of-function intolerant) and LOEUF as metrics for the gene's loss-of-function intolerance within human population. We downloaded both pLI and LOEUF scores from gnomAD browser[114]. We inverse-normal transformed the raw number of significant SCENT peaks for each gene, since the raw number of significant SCENT peaks for each gene is skewed toward zero (Extended Data Fig. 4a). We performed linear regression between the normalized number of significant SCENT peaks and pLI or LOEUF score with accounting for gene length, which could be potential confounding factor for pLI and LOEUF[69,70].

Validation with CRISPR-Flow FISH data

To validate our SCENT enhancer–gene links, we used published CRISPR-Flow FISH experiments as potential ground-truth positive enhancer element-gene links and negative enhancer element–gene links. We downloaded the experimental results from the Supplementary Table 5 of original publication[39] conducted in multiple cell lines and cell types (K562, Jurkat, THP1, GM12878, BJAB, NCCIT, hepatocytes and LNCAP). We used 'Perturbation Target' as candidate enhancer elements. We defined 283 positive enhancer element–gene links when they are 'TRUE' for 'Regulated' column (that is, the element–gene pair is significant and the effect size is negative) and 5,472 negative enhancer element–gene links when they are 'FALSE' for 'Regulated' column. We lifted them over to GRCh38 and obtained final sets of 278 positive links and 5,470 negative links.

We used the SCENT enhancer–gene maps from eight multimodal datasets, excluding SHARE-seq dataset in which we were unable to perform statistical enrichment test due to low overlap with the designed target element–gene pairs in the CRISPR-Flow FISH experiments. For each dataset, we used 'bedtools (version 2.26.0) intersect' to categorize SCENT peak–gene links and non-SCENT ATAC peak–gene pairs into either CRISPR-positive or CRISPR-negative groups, based on whether these peaks overlapped with positive or negative CRISPR-Flow FISH links for the same gene (Supplementary Table 3). We finally performed two-sided Fisher's exact test to assess the enrichment of CRISPR-positive links within SCENT peak–gene links in each dataset.

Validation with H3K27ac HiChIP data

To assess if the SCENT enhancer–gene linkage is more likely within the contact map of active enhancers and target genes constructed from H3K27ac HiChIP experiment, we downloaded high-confidence contact loops by Hi-C combined with enhancer activity marked by H3K27ac level in naive T cells, Th17 T cells and regulatory T cells from Supplementary Table 2 of ref. 68. We lifted the genomic coordinates to GRCh38 in each cell type, and annotated the promoter regions with gene names to be used as putative target genes if they fall within 1 kb from gene's TSS. We took the union of the enhancer–promoter contacts across these three cell subtypes as experimentally validated T cell linkage. We analyzed six datasets that included T cells (that is, arthritis-tissue, public PBMC, NeurIPS, Dogma-seq (stimulated and control) and NEAT-seq datasets) to test whether the T-cell-specific enhancer–gene linkage from SCENT was enriched for the H3K27ac HiChIP enhancer–gene links. We additionally combined the significant linkages across all six datasets to create union SCENT peak–gene linkages in T cells. For each dataset and the combined linkage, we used 'bedtools intersect' to categorize SCENT peak–gene links and non-SCENT ATAC peak–gene pairs based on whether these peaks overlapped with H3K27ac HiChIP contact loops for the same gene. We performed two-sided Fisher's exact test to assess the enrichment of SCENT peak–gene links within H3K27ac HiChIP contact loops in each dataset and the combined dataset.

Cell-type-specific SCENT tracks and aggregated SCENT tracks

For cell types with more than 5,000 cells across datasets, we concatenated SCENT peak–gene linkages across all the datasets to create cell-type-specific SCENT tracks. We collected a set of SCENT peak–gene linkages for the same cell type and used 'bedtools merge' function (for each gene) to obtain a union of SCENT peaks for each gene. Similarly, we created aggregated SCENT tracks across all the cell types and all datasets. We collected all sets of SCENT peak–gene linkages and used 'bedtools merge' function (for each gene) to obtain a union of SCENT peaks for each gene across all the cell types and all datasets.

Causal variant enrichment analysis using eQTLs

We defined a causal enrichment for eQTL within SCENT enhancers and other annotations by using statistically fine-mapped variant–gene combinations from GTEx. We used publicly available statistics analyzed by CAVIAR software[20], and selected variants with PIP >0.2 as putatively causal (fine-mapped) variants for primary analyses. For the primary enrichment analysis, we aggregated fine-mapped variants from all the 49 tissues. For cell-type-specific SCENT enrichment analysis (Extended Data Fig. 6j), we used fine-mapped variants from each tissue separately. We intersected these putatively causal variants with our annotations. We then retained any variants which the linking method (SCENT, ArchR, Signac and Cicero) connected to the same gene as GTEx phenotype gene.

$$\text{Enrichment}_{\text{gene\_}i} = \frac{\#\text{causal\_var\_in\_annot}_{\text{gene\_}i} \Big/ \sum \text{common\_var\_in\_annot}_{\text{gene\_}i}}{\#\text{causal\_var}_{\text{gene\_}i} \Big/ \sum \text{common\_var\_in\_cis}_{\text{gene\_}i}}$$

$$\text{Overall Enrichment} = \frac{1}{n}\sum_{i=1}^{n} \text{Enrichment}_{\text{gene\_}i}$$

For each gene *i* (expression phenotype), we divided the number of putatively causal variants within an annotation normalized by the number of common variants within an annotation by the number of all causal variants for gene *i* normalized by the number of all common variants within *cis*-region from for gene *i*. To calculate common variants within annotation or within locus, we used 1,000 Genomes Project genotype. We selected any variants with minor allele frequency >1% in European population as a set of common variants to be intersected with each annotation. To derive Overall Enrichment score, we took the mean across all the genes.

To have further insights into precision and recall and compare against ArchR peak2gene and Signac LinkPeaks functions, we varied the threshold for defining a set of significant peak–gene linkages in each software (that is, FDR in SCENT {0.50, 0.30, 0.20, 0.10, 0.05, 0.02}, Peason's correlation *r* {any, 0, 0.1, 0.3, 0.5, 0.7} in ArchR, and correlation score {any, 0, 0.05, 0.1, 0.15} in Signac). We used the same myeloid cells in the arthritis-tissue dataset and a set of eQTL fine-mapped variants in GTEx blood tissue for this benchmark across all three methods. We then used each set of peak–gene linkages to recalculate causal variant enrichment Overall Enrichment score (Fig. 3b).

We also assessed the impact of PIP threshold in defining a set of statistically fine-mapped variants on the causal variant enrichment analysis. To do so, we redefined the set of putative causal variants with more stringent PIP thresholds (PIP >0.5 and PIP >0.7), and re-computed the calculate causal variant enrichment Overall Enrichment score.

Distance from cis region to gene's TSS and causal variant enrichment for eQTL

We sought to evaluate the effect of regulatory elements' proximity to TSS on enrichment for eQTL causal variants. We created annotated genome regions based on TSS distance to a given gene (for example, within 1 kb, from 1 kb to 10 kb, and so on). We concatenated these elements across all 23,715 genes to create genome-wide annotations reflecting TSS proximity (Supplementary Fig. 5a). We used the eQTL fine-mapped variants with PIP >0.2 from GTEx for all tissues as putative causal variants. We computed Overall Enrichment described above as the degree of causal variant enrichment within these regulatory elements in each TSS distance bin.

Peak–gene linkage using Poisson regression alone

As other benchmarking for assessing the effect of the components of SCENT on the causal variant enrichment, we also created peak–gene linkage using the Poisson regression but without nonparametric bootstrapping for the same dataset of myeloid cells in the arthritis-tissue dataset. We used the nominal *P* values for the term X_{peak} from the Poisson regression (equation (1)) to perform FDR correction to obtain significant peak–gene pairs using the Poisson regression alone. We then used the FDR thresholds {0.30, 0.20, 0.10, 0.05, 0.02, 0.01} for assessing the recall–precision tradeoffs as described in the previous section.

GWAS fine-mapping results

We used GWAS fine-mapping results in FinnGen release 6 (ref. 79) upon registration and publicly available GWAS fine-mapping results in UK Biobank[80,115]. For FinnGen traits, we downloaded all the fine-mapping results by SuSIE software[22] and systematically selected any traits with case count >1,000. We then selected noncoding fine-mapped loci that did not include any nonsynonymous or splicing variants with PIP >0.5. We thus analyzed 1,046 traits and 5,753 loci in total after QC. For UK Biobank, we analyzed the fine-mapping results by SuSIE software for all 94 traits including binary and quantitative traits. Since the genomic coordinates for the UK Biobank fine-mapping results were hg19, we lifted them over to GRCh38 by using LiftOver software. We again selected noncoding fine-mapped loci that did not include any nonsynonymous or splicing variants with PIP >0.5. We thus analyzed 7,274 loci in total after QC.

We analyzed three additional autoimmune GWAS fine-mapping results for rheumatoid arthritis[26], type 1 diabetes[90] and inflammatory bowel disease[29], given our special interest in immune-mediated traits. We similarly selected noncoding fine-mapped loci that did not include any nonsynonymous or splicing variants with PIP >0.5, and lifted the results over to GRCh38 by using LiftOver software. We defined 117 loci for rheumatoid arthritis, 77 loci for type 1 diabetes and 86 loci for inflammatory bowel disease.

Causal variant enrichment analysis using GWAS

We defined causal variant enrichment statistics for GWAS within SCENT enhancers and other annotations by using statistically fine-mapped variants from FinnGen[79] and UK Biobank[80] that we described in the previous section. We selected variants with PIP >0.2 as putatively causal variants for primary analyses.

$$\text{Enrichment}_{\text{trait\_}i} = \frac{\#\text{causal\_var\_in\_annot}_{\text{trait\_}i} \Big/ \sum \text{common\_var\_in\_annot}_{\text{trait\_}i}}{\#\text{causal\_var}_{\text{trait\_}i} \Big/ \sum \text{common\_var\_across\_loci}_{\text{trait\_}i}}$$

$$\text{Overall Enrichment} = \frac{1}{n}\sum_{i=1}^{n} \text{Enrichment}_{\text{trait\_}i}$$

For each trait *i*, we divided the number of putatively causal variants within an annotation (across all loci for trait *i*) normalized by the number of common variants within an annotation by the number of all causal variants for trait *i* normalized by the number of all common variants within all significant loci analyzed for the trait *i*. To calculate common variants within annotation or within locus, we again used 1000 Genomes Project

variants with minor allele frequency >1% in European population. To derive Overall Enrichment score, we took the mean across all the traits.

For each trait *i* and putative causal gene pair, we calculated the distance between the TSS of the gene and the most likely causal variant which had the largest PIP when multiple variants were nominated for a single gene by SCENT (Supplementary Fig. 6a). For each putative causal gene for the trait *i*, we also sorted all the genes on the basis of the distance between the gene's TSS and the most likely causal variant (from the smallest to the largest). We then obtained the rank of the putative causal gene from SCENT among the sorted gene list to see how often the SCENT gene is the closest gene from the most likely causal variant.

Comparison with bulk-tissue-based regulatory annotation and enhancer–gene maps

We downloaded per-group EpiMap enhancer–gene links from ref. 116. We lifted the genomic coordinates to GRCh38 by using LiftOver software. When we assessed aggregated EpiMap enhancer–gene links across all the 31 tissue groups, we used 'bedtools merge' function for each gene to create a union of all enhancer–gene links (Fig. 3c,d). For tissue-specific enrichment analyses, we analyzed the 31 group-specific tracks separately (Extended Data Fig. 7a,b). To benchmark the precision and recall, we used EpiMap correlation scores to define variable sets of enhancer–gene links from EpiMap based on the threshold of EpiMap correlation score.

We downloaded ABC predictions in 131 cell types and tissues from ref. 117. We lifted the genomic coordinates to GRCh38 by using LiftOver software. When we assessed aggregated ABC enhancer–gene links across all the groups, we used 'bedtools merge' function for each gene to create a union of all enhancer-gene links across 131 cell types (Fig. 3c,d). For cell-type-specific analyses, we aggregated cell lines or cell types to be corresponding with our cell types and analyzed each of these tracks separately (B cell, T cell, myeloid cells and fibroblasts; Extended Data Fig. 7a,b). To benchmark the precision and recall, we used ABC scores to define variable sets of enhancer–gene links from ABC model based on the threshold of ABC score.

To assess precision and recall and compare against bulk-tissue based methods (that is, EpiMap and ABC model), we used sets of significant peak–gene linkages in each method with varying thresholds (that is, FDR in SCENT {0.5, 0.3, 0.2, 0.1, 0.05, 0.02}, EpiMap correlation score {0, 0.4, 0.8, 0.9} in EpiMap, and ABC score {0, 0.05, 0.1, 0.2} for ABC model). We then used each set of peak–gene linkages to recalculate causal variant enrichment for GWAS (Fig. 3d).

We also assessed the impact of PIP threshold in defining a set of statistically fine-mapped variants on the causal variant enrichment analysis. To do so, we redefined the set of putative causal variants with more stringent PIP thresholds (PIP >0.5 and PIP >0.7), and recomputed the calculate causal variant enrichment Overall Enrichment score.

Comparison with other single-cell-based enhancer–gene maps

To compare the capability of identifying putative causal variants for GWAS between SCENT and previous single-cell-based enhancer–gene maps by ArchR or Signac (Extended Data Fig. 8c,d), we created an integrated enhancer–gene maps across available datasets for these methods. To this end, we used eight multimodal datasets (excluding NeurIPS dataset) in which fragment files for ATAC–seq in addition to count matrices of RNA-seq and ATAC–seq are available to run ArchR. For each of these methods (SCENT, ArchR and Signac), we took the union of significant enhancer–gene linkages with FDR <10% from eight datasets by using 'bedtools merge' function. We then calculated causal variant enrichment statistics for GWAS within SCENT, ArchR and Signac enhancers by using statistically fine-mapped variants from FinnGen and UK Biobank (PIP >0.2) as we described above.

caQTL analysis using scATAC–seq samples with genotype

As part of the AMP consortium, we generated an independent arthritis-tissue dataset with single-cell unimodal ATAC–seq data with genotype ($n = 17$; one sample without genotype data was excluded)[58] to define caQTLs. We used two methods, binomial test for allele-specific chromatin accessibility and RASQUAL. Briefly, we genotyped donors in the AMP Network for rheumatoid arthritis and systemic lupus erythematosus including 17 donors in this study by using Illumina Multi-Ethnic Genotyping Array across three batches. We performed quality control of genotype by sample call rate >0.99, variant call rate >0.99, minor allele frequency >0.01, and $P_{HWE} > 1.0 \times 10^{-6}$. We performed haplotype phasing with SHAPEIT2 software (2.727)[118] and performed whole-genome imputation by using minimac3 software (version 2.0.1)[119] with a reference panel of 1000 Genomes Project phase 3 (ref. 120). After imputation, we selected variants with imputation $R^2 > 0.7$ as post-imputation QC. We next created a merged bam file of ATAC–seq for each donor and each cell type by aggregating all the reads. Using the imputed genotype for each donor and aggregated bam files for each donor and cell type, we applied WASP[121] to correct any bias in read mapping toward reference alleles to accurately quantify allelic imbalance. We thus created a bias-corrected bam files for each donor and cell type.

For binomial tests for allele-specific chromatin accessibility, we ran ASEReadCounter module in GATK software (version 4.1.9.0)[122] using the bias-corrected bam files as input to quantify allelic imbalance in heterozygous sites with read count >4 within ATAC peak counts. We first performed one-sided binomial tests in each donor, and meta-analyzed the statistics across donors by Fisher's method if multiple donors shared the same heterozygous site. For RASQUAL, we created a VCF file containing both genotype dosage and allelic imbalance from ASEReadCounter. We quantified the read coverage for each peak and for each donor by 'bedtools coverage' function. We created a peak by donor matrix with read coverage. We QCed samples with log(total mapped fragments) fewer than mean − 2 × standard deviation across samples in each cell type. We QCed peaks so that at least two individuals have any fragments for the peak. We then ran RASQUAL software with the inter-individual differences in ATAC peak counts (in the peak by donor matrix) and intra-individual allelic imbalance (in the VCF), with accounting for chromatin accessibility PCs (the first *N* components whose explained variances are greater than those from permutation result), 3 genotype PCs, sample site and sex as covariates. RASQUAL output chi-squared statistics and *P* values. We computed FDR from these raw *P* values by Benjamini & Hochberg correction on local multiple test burden (that is, the number of *cis*-SNPs in the region). To correct for genome-wide multiple testing, we ran the RASQUAL with random permutation, in which the relationship between sample labels and the count matrix was broken. Thus, we derived *q* values for each candidate caQTL.

We finally intersected these peaks with significant caQTL effect in each significance threshold with SCENT peaks and assessed causal variants enrichment within these peaks for GWAS as explained in the previous sections.

In the example *SMAD3* locus, we visualized the allele-specific effect by creating an aggregated bam files for each donor and using Integrative Genomics Viewer (Extended Data Fig. 9c). We also visualized the inter-individual effect by taking the residuals after regressing out the covariates from the logged count per million of the ATAC read coverage for the peak in each donor (Extended Data Fig. 9d).

ClinVar analysis

We downloaded the latest clinically reported variant list registered at ClinVar from ref. 123. We then screened the variants to exclude (1) exonic variants and (2) variants categorized as 'benign'. We defined the ClinVar variant density as the number of the noncoding and nonbenign variants within each annotation × 1,000 divided by the total length (bp) of each annotation.

Somatic mutation analysis

We used a list of somatic noncoding mutation hotspots for leukemia in Supplementary Table 13 of the original publication[105]. We lifted

the genomic coordinates to GRCh38 by using LiftOver software. We then intersected the noncoding somatic mutation hotspots with our cell-type-specific SCENT peaks in immune cells (that is, T cells, B cells and myeloid cells). We compared the intersected elements' target genes by SCENT with the 'Annotate_Gene' column from the original publication.

Downsampling experiments

To evaluate the effect of cell numbers on the statistical power in detecting significant SCENT enhancer–gene linkages, we performed downsampling experiments in fibroblast (the most abundant cell type in arthritis-tissue dataset, n_{cell} = 9,905). We randomly samples cells (n_{cell} = 500, 1000, 2,500, 5,000 and 7,500). We then applied SCENT to each of the subset groups of cells and defined significant peak–gene links with FDR <10%. We counted the number of significant peak–gene links in each of the subset groups of cells, and annotated peaks based on the distance to the TSS to the target gene.

Statistics and reproducibility

No sample data were excluded from the single-cell multimodal analyses. Neither randomization, sample size predetermination nor blinding of investigators was applicable to this study.

Reporting summary

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

Data availability

The publicly available datasets were downloaded via Gene Expression Omnibus (accession codes GSE140203, GSE156478, GSE178707, GSE194122, GSE193240 and GSE178453) or web repository (https://www.10xgenomics.com/resources/datasets?query=&page=1&configure%5Bfacets%5D%5B0%5D=chemistryVersionAndThroughput&configure%5Bfacets%5D%5B1%5D=pipeline.version&configure%5BhitsPerPage%5D=500&menu%5Bproducts.name%5D=Single%20Cell%20Multiome%20ATAC%20%2B%20Gene%20Expression). The raw data for arthritis-tissue dataset (single-cell multimodal RNA/ATAC–seq and single-cell ATAC–seq) are deposited at the NIH Database of Genotypes and Phenotypes (dbGaP accession number phs003417.v1.p1) and the Gene Expression Omnibus (GEO accession number GSE243917).

Code availability

The computational scripts related to this manuscript are available at https://github.com/immunogenomics/SCENT (https://doi.org/10.5281/zenodo.10452116)[124].

References

110. Donlin, L. T. et al. Methods for high-dimensional analysis of cells dissociated from cyropreserved synovial tissue. *Arthritis Res Ther.* **20**, 139 (2018).
111. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
112. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
113. PhastCons scores for multiple alignments of 99 vertebrate genomes to the human genome. *UCSC Genome Browser* https://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/ (2014).
114. gnomAD database. *Broad Institute* https://gnomad.broadinstitute.org/downloads (2023).
115. GWAS fine-mapping results. *Finucane Lab* https://www.finucanelab.org/data (2019).
116. EpiMap Gene-Enhancer links. *Broad Institute* https://personal.broadinstitute.org/cboix/epimap/links/pergroup/ (2021).
117. ABC predictions across 131 biosamples. *Broad Institute* ftp://ftp.broadinstitute.org/outgoing/lincRNA/ABC/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz (2021).
118. Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
119. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
120. Gibbs, R. A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
121. van de Geijn, B., Mcvicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2015).
122. van der Auwera G. & O'Connor, B. *Genomics in the Cloud* (O'Reilly Media, Inc., 2020).
123. ClinVar variants. *ClinVar* https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar.vcf.gz (2023).
124. Sakaue, S. immunogenomics/SCENT: v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.10452116 (2024).

Acknowledgements

We sincerely thank participants of this study who provided tissue samples. We thank A. Gupta, J. Kang and K. Lagattuta for their comments and helpful discussion on the manuscript. This work is supported in part by funding from the National Institutes of Health (R01AR063759, U01HG012009 and UC2AR081023 to S.R.). S.S. was in part supported by the Uehara Memorial Foundation and The Osamu Hayaishi Memorial Scholarship. K. Weinand was supported by NIH NIAMS T32AR007530. K. Wei was supported by a Burroughs Wellcome Fund Career Awards for Medical Scientists, a Doris Duke Charitable Foundation Clinical Scientist Development Award, a Rheumatology Research Foundation Innovative Research Award, and NIH NIAMS K08AR077037. We thank the Brigham and Women's Hospital Center for Cellular Profiling Single Cell Multiomics Core for experimental design and protocol optimization. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.S. and S.R. conceived the work and wrote the manuscript with critical input from co-authors. S.S. and K. Weinand analyzed the arthritis-tissue dataset and S.S. analyzed publicly available datasets with help and guidance from K.K.D., K.J., M.K., A.M., A.L.P. and S.R. G.F.M.W., Z.Z., M.B.B., L.T.D. and K. Wei provided samples and generated the arthritis-tissue dataset. S.I. refactored the SCENT software implementation as an R package.

Competing interests

S.R. is a founder for Mestag, Inc., a scientific advisor for Rheos, Jannsen and Pfizer, and serves as a consultant for Sanofi and Abbvie. The other authors declare no competing interests.

Additional information

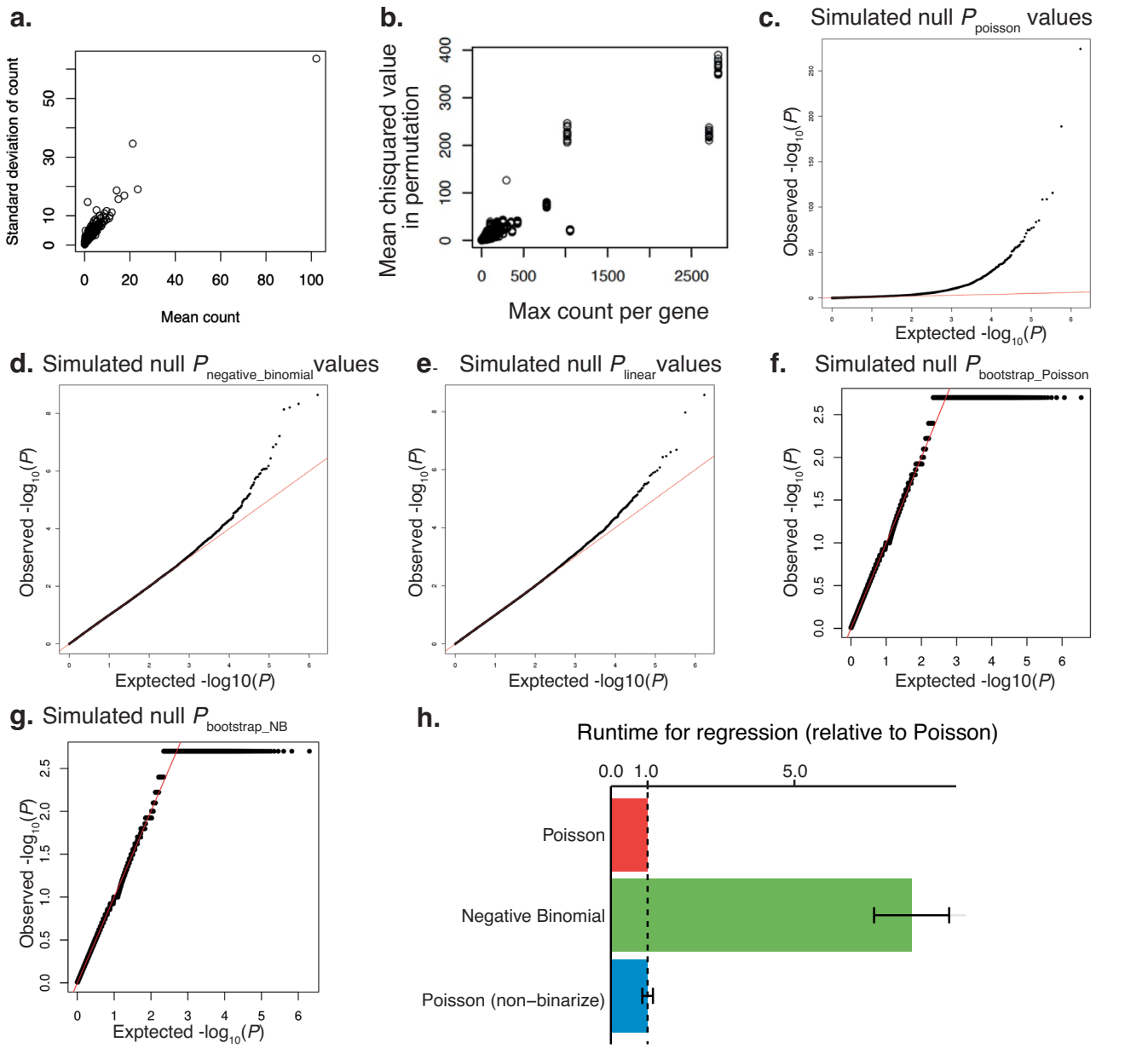
Extended data is available for this paper at https://doi.org/10.1038/s41588-024-01682-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01682-1.

Correspondence and requests for materials should be addressed to Soumya Raychaudhuri.

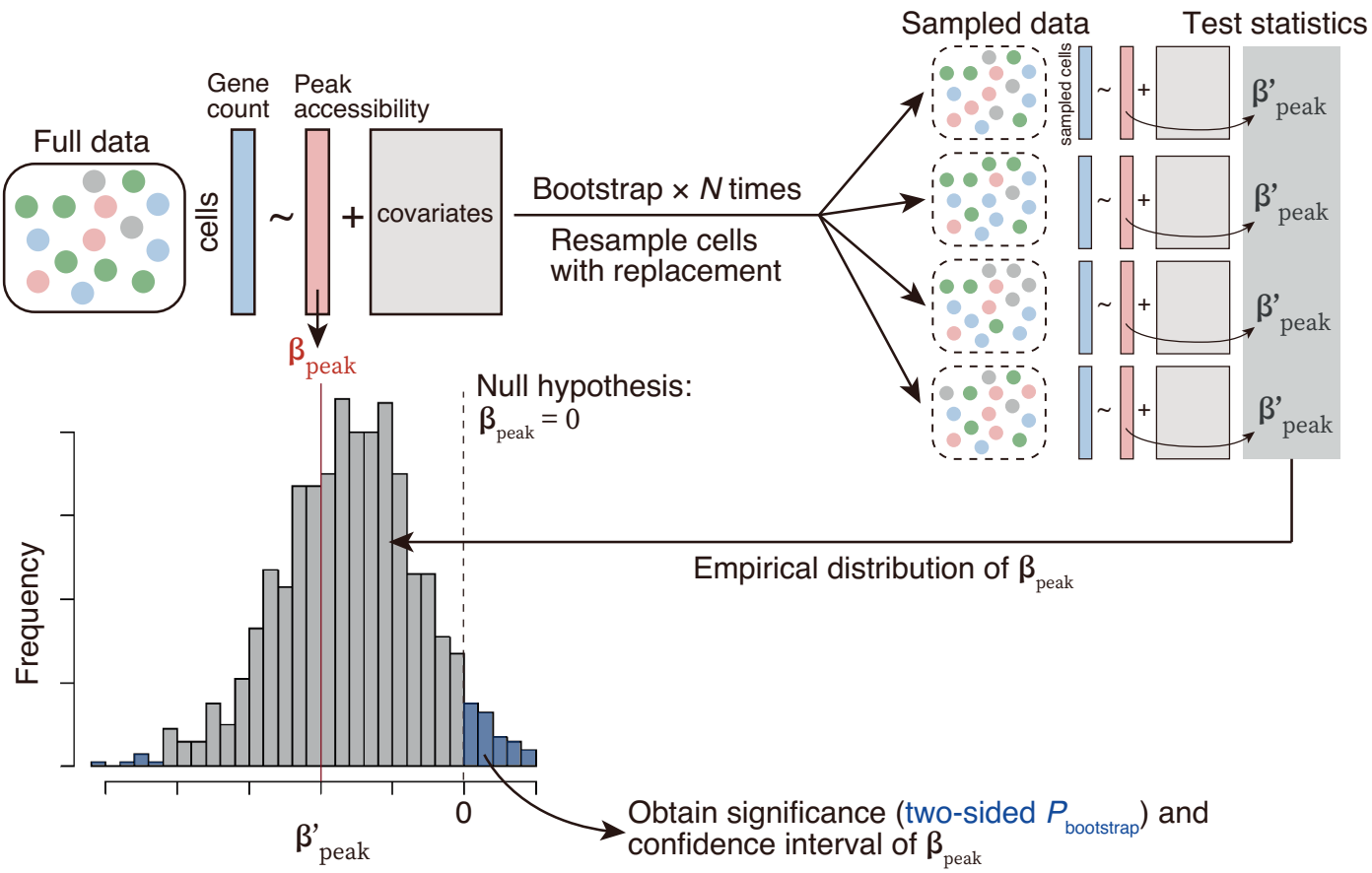
Peer review information *Nature Genetics* thanks Tim Stuart and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-024-01682-1



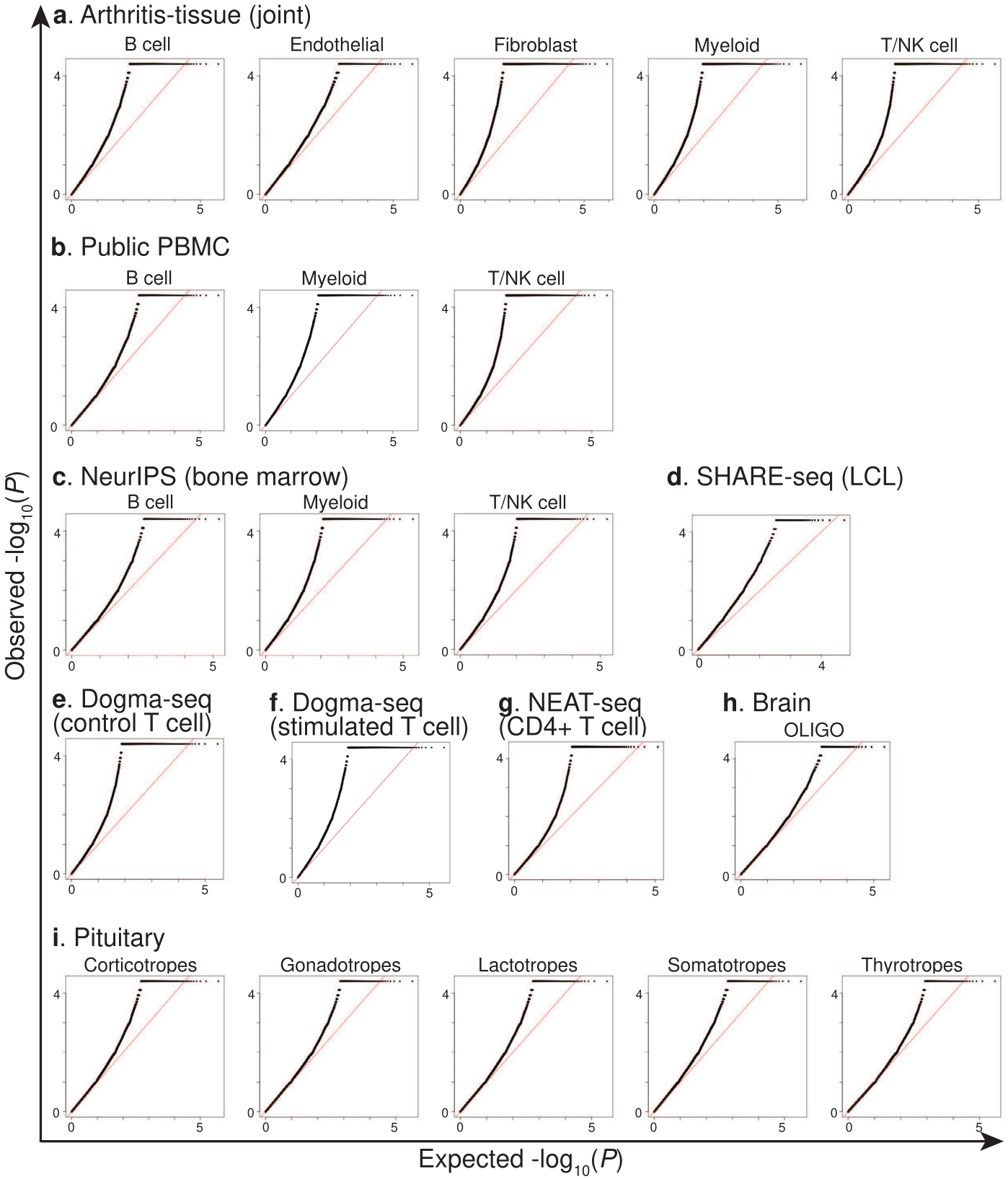
Extended Data Fig. 1 | Distribution of gene expression counts in single-cell RNA-seq and statistics from association between gene expression and chromatin accessibility under null simulation. a. In an example dataset of arthritis-dataset, mean gene count was strongly correlated with standard deviation of the gene count. **b**. The correlation between max expression count per gene (x-axis) and the mean naïve association chi-square values (χ^2) from Poisson regression between gene expression and chromatin accessibility under null simulation (y-axis). **c**. The quantile-quantile (QQ) plot of two-sided *P* values from the Poisson regression between gene expression count and chromatin accessibility under null simulation. **d**. The QQ plot of two-sided *P* values from the negative binomial regression between gene expression count and chromatin accessibility under null simulation. **e**. The QQ plot of two-sided *P* values from the linear regression between log-normalized and inverse-normal-transformed gene expression and chromatin accessibility under null simulation. **f**. The QQ plot of two-sided *P* values estimated from bootstrapping based on the statistics distributions from the Poisson regression between gene expression count and chromatin accessibility under null simulation. **g**. The QQ plot of two-sided *P* values estimated from bootstrapping based on the statistics distributions from the negative binomial regression between gene expression count and chromatin accessibility under null simulation. **h**. Computational runtime benchmarking for Poisson regression with binarized ATAC-seq peak (red), negative binomial regression with binarized ATAC-seq peak (teal), and Poisson regression with non-binarized ATAC-seq peak (blue). The values are relative to the computational time for Poisson regression, and bars are the mean across *n*=100 randomly selected peak-gene pairs. Horizontal lines (error bars) indicate one standard deviation from the mean.

https://doi.org/10.1038/s41588-024-01682-1



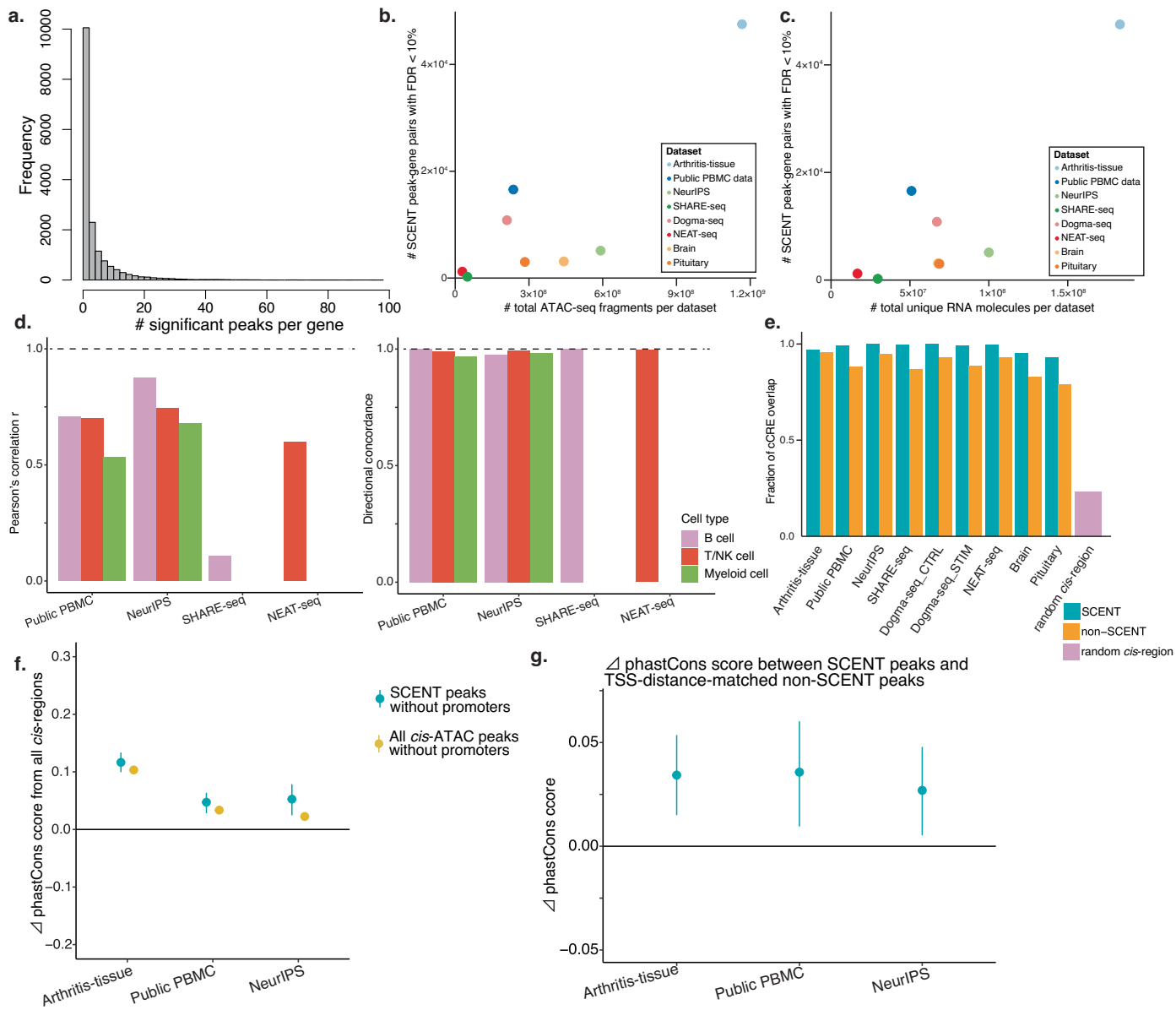
Extended Data Fig. 2 | Schematic overview of SCENT model using Poisson regression and non-parametric bootstrapping. We first run Poisson regression associating the raw gene expression count (RNA-seq) with the peak accessibility (ATAC-seq) accounting for technical covariates across the entire cells in the multimodal data to estimate β_{peak}. Then, we resampled cells with replacement from the full data in each of the bootstrapping round and re-estimated β'_{peak} for N times. We compared this empirical distribution of β'_{peak} against the null hypothesis ($\beta'_{peak} = 0$) to derive the significance of β_{peak} (that is, two-sided bootstrapping-based P value = $P_{bootstrap}$).

https://doi.org/10.1038/s41588-024-01682-1



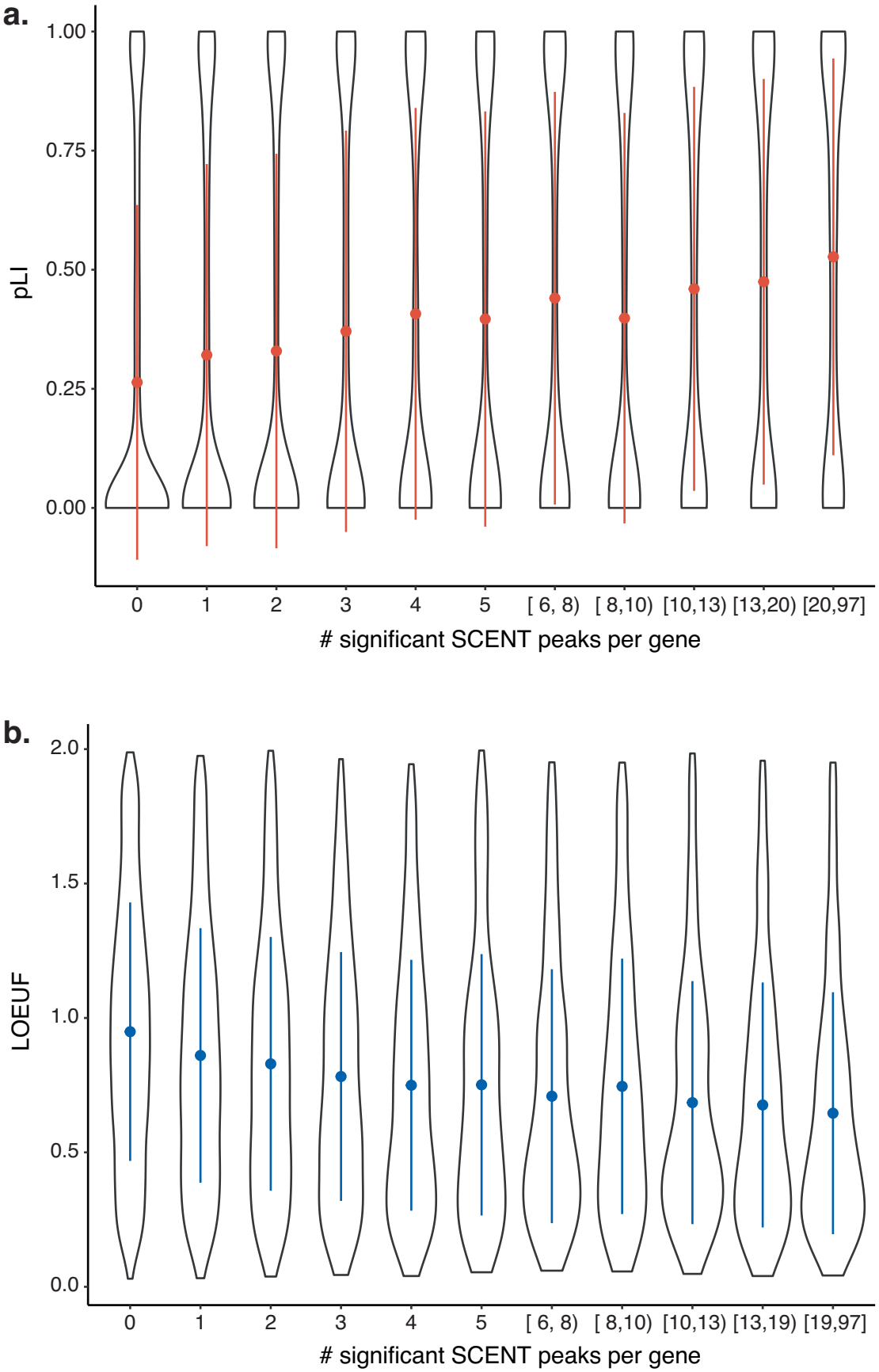
Extended Data Fig. 3 | The QQ plot of SCENT P values by bootstrapping. We applied SCENT to each of 23 broad cell types from 9 single-cell multimodal datasets. Each QQ plot represents two-sided $P_{bootstrap}$ values in each cell type in each dataset (**a**. arthritis-tissue, **b**. public PBMC, **c**. NeurIPS, **d**. SHARE-seq, **e**. Dogma-seq (control), **f**. Dogma-seq (stimulated) **g**. NEAT-seq, **h**. Brain, **i**. Pituitary.

https://doi.org/10.1038/s41588-024-01682-1



Extended Data Fig. 4 | Properties of SCENT peaks. a. The number of significant SCENT peaks per gene across genes we investigated in at least one dataset-cell type pair. **b**. The number of significant gene-peak pairs discovered by SCENT with FDR < 10% in each dataset (y-axis) as a function of the total number of ATAC-seq fragments in each dataset (x-axis), colored by the dataset. **c**. The number of significant gene-peak pairs discovered by SCENT with FDR < 10% in each dataset (y-axis) as a function of the total number of unique RNA molecules in each dataset (x-axis), colored by the dataset. **d**. The effect size correlation *r* by Pearson's correlation between arthritis-tissue dataset and the other dataset for the same cell type (left) and the directional (sign) concordance between arthritis-tissue dataset and the other dataset for the same cell type (right). **e**. Fraction of overlap with ENCODE cCREs in SCENT (teal) or non-SCENT peaks (orange) in each dataset and random set of *cis*-non-coding regions (pink). **f**. The mean Δ phastCons score for SCENT with excluding promoter peaks (teal) and all *cis*-ATAC peaks with excluding promoter peaks (yellow) in each of the three example multimodal datasets. The bars indicate the 95% CI by bootstrapping genes ($n_{\text{bootstrap}}$=1000). **g**. The mean Δ phastCons score between SCENT peaks and TSS-distance-matched non-SCENT peaks across all the genes. The bars indicate the 95% CI by bootstrapping genes ($n_{\text{bootstrap}}$=1000).

https://doi.org/10.1038/s41588-024-01682-1

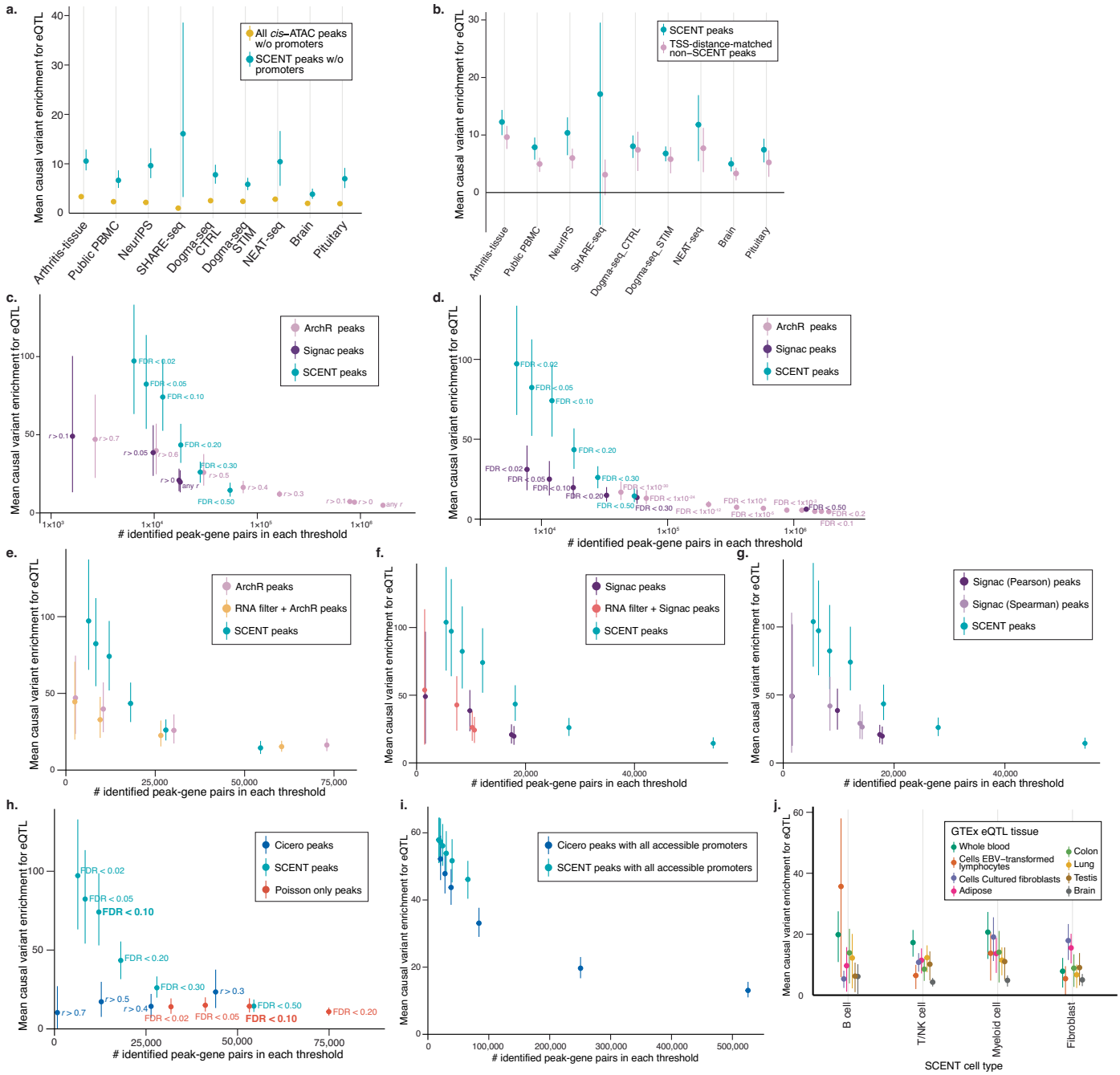


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41588-024-01682-1

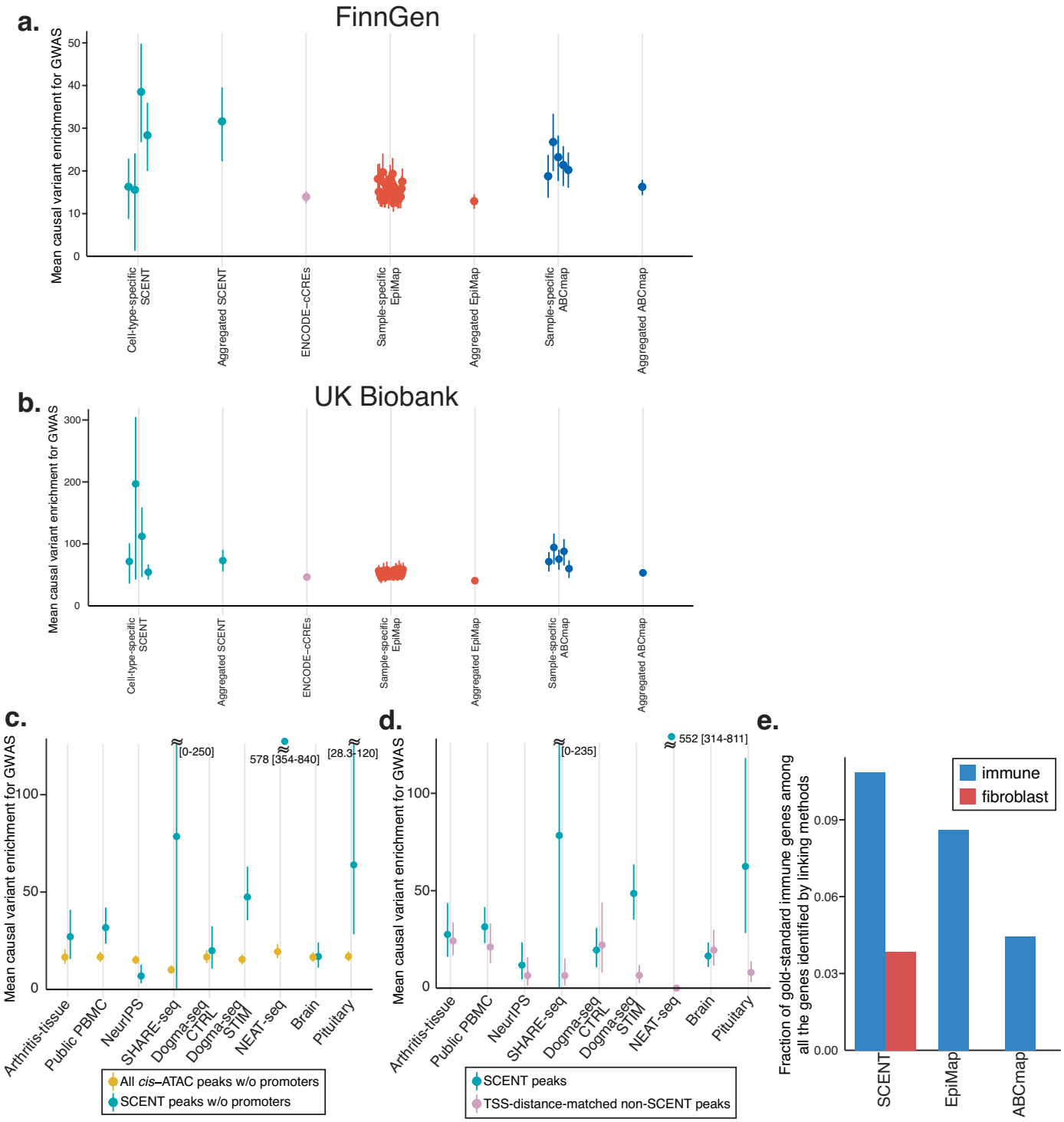
Extended Data Fig. 5 | Mutational constraint on genes with a high number of SCENT peaks. For each gene, the number of SCENT peaks were counted and binned as shown in the x-asis, and mutational constraint metric (pLI (the probability of being loss of function intolerant): **a**, LOEUF (the loss-of-function observed/expected upper bound fraction): **b**) for genes within each bin are shown as a violin plot on the y-axis. The dots indicate the mean score in each bin, and the error bars indicate one standard deviation from the mean. Each bin consists of 555-4071 genes in **a** and 568-4265 genes in **b**.

https://doi.org/10.1038/s41588-024-01682-1



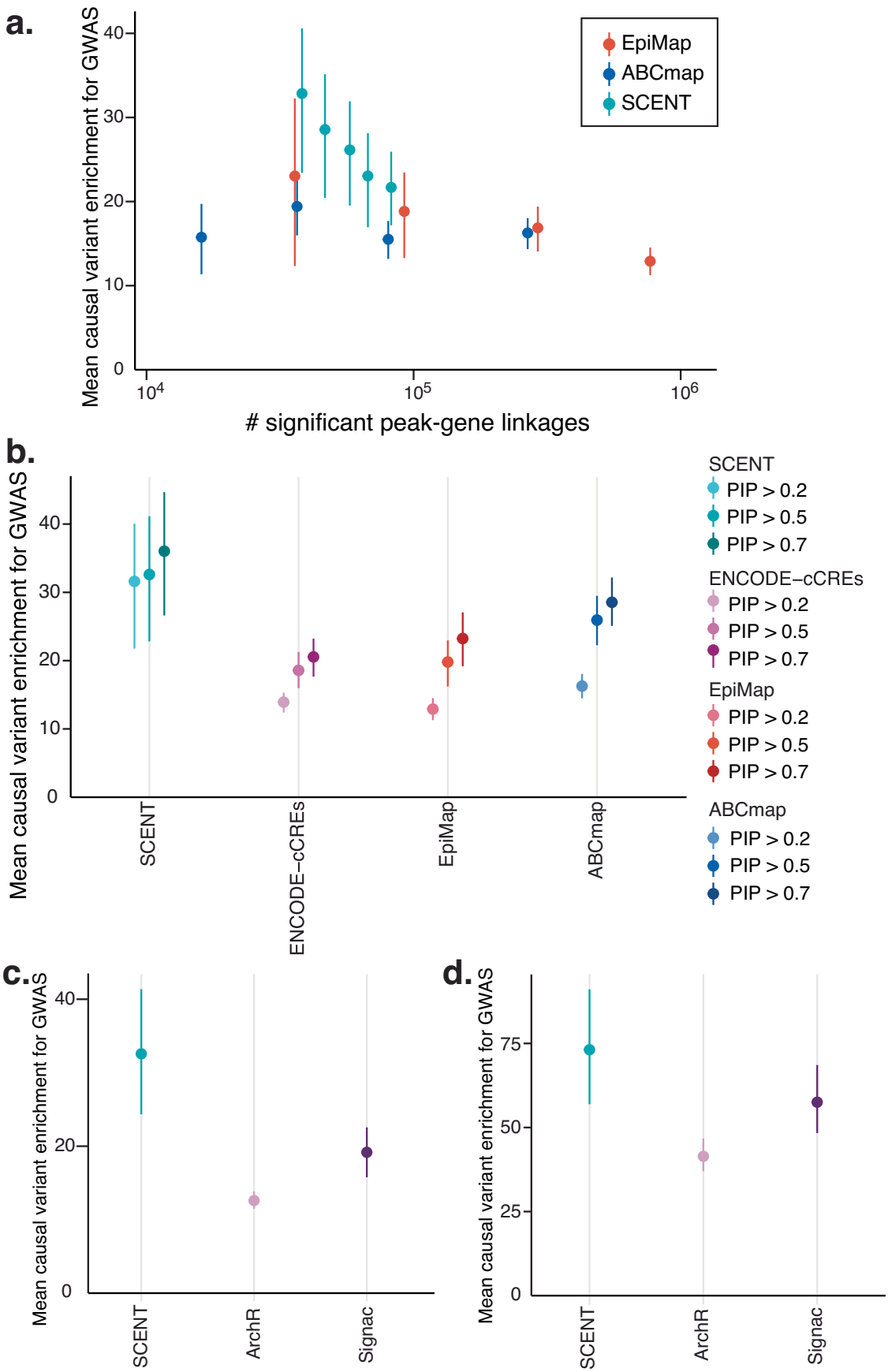
Extended Data Fig. 6 | Causal variant enrichment for eQTLs. a. The mean causal variant enrichment for eQTL within SCENT peaks with excluding all promoters (teal) or *cis*-regulatory ATAC-seq peaks with excluding all promoters (yellow) in each dataset. **b**. The mean causal variant enrichment for eQTL within SCENT peaks (teal) or non-SCENT peaks with matching distance to TSS (pink). **c**. Comparison of the mean causal variant enrichment for eQTL (y-axis) among SCENT (teal), ArchR (pink), and Signac (purple) as a function of the number of significant peak-gene pairs at each threshold of significance by FDR in SCENT and correlation *r* in ArchR and Signac. **d**. Comparison of the mean causal variant enrichment for eQTL among SCENT, ArchR, and Signac as a function of the number of significant peak-gene pairs at each threshold of FDR in SCENT, ArchR and Signac. The ArchR results with > 180,000 peak-gene linkages are omitted. **e**. Comparison of the mean causal variant enrichment for eQTL among SCENT, ArchR, and ArchR filtered on RNA expression as a function of the number of significant peak-gene pairs. **f**. Comparison of the mean causal variant enrichment for eQTL among SCENT, Signac, and Signac filtered on RNA expression as a function of the number of significant peak-gene pairs. **g**. Comparison of the mean causal variant enrichment for eQTL among SCENT, the default Pearson's correlation version of Signac, and the optional Spearman's correlation version of Signac as a function of the number of significant peak-gene pairs. **h**. Comparison of the mean causal variant enrichment for eQTL among original SCENT (Poisson regression + non-parametric bootstrapping), Poisson-only strategy without bootstrapping, and Cicero (correlation method using sc-ATAC-seq alone) as a function of the number of significant peak-gene pairs up to 100,000 peak-gene linkages. **i**. Comparison of the mean causal variant enrichment for eQTL between SCENT and Cicero peaks with adding all accessible promoter regions (1 kb regions from TSS) to account for potential promoter bias. **j**. Tissue-specific causal variant enrichment within SCENT peaks. The dots and lines are colored by the eQTL source tissue in GTEx that we assessed. In all panels, the bars indicate 95% confidence intervals by bootstrapping genes ($n_{bootstrap}$=1000).

https://doi.org/10.1038/s41588-024-01682-1



Extended Data Fig. 7 | Causal variant enrichment for GWAS. a and **b**. The mean causal variant enrichment for GWAS within cell-type-specific and aggregated SCENT enhancers (teal), ENCODE cCREs (pink), group-specific and aggregated EpiMap enhancers (red) and sample-specific and aggregated ABC enhancers (blue). GWAS results were based on FinnGen (**a**) and UK Biobank (**b**). The bars indicate 95% confidence intervals by bootstrapping traits ($n_{bootstrap}$=1000). **c**. The mean causal variant enrichment for FinnGen GWAS (see Methods) within SCENT peaks with excluding all promoters (teal) or *cis*-regulatory ATAC-seq peaks with excluding all promoters (yellow) in each of the 9 single-cell datasets. The bars indicate 95% confidence intervals by bootstrapping traits ($n_{bootstrap}$=1000). **d**. The mean causal variant enrichment for FinnGen GWAS (see Methods) within SCENT peaks (teal) or non-SCENT peaks with matching distance to TSS (pink) in each of the 9 single-cell datasets. The bars indicate 95% confidence intervals by bootstrapping traits ($n_{bootstrap}$=1000). **e**. The fraction of known genes from Mendelian autoimmune diseases among all the genes identified by SCENT, EpiMap, and ABC model. The color of the bars indicates the cell types in each linking method.

https://doi.org/10.1038/s41588-024-01682-1

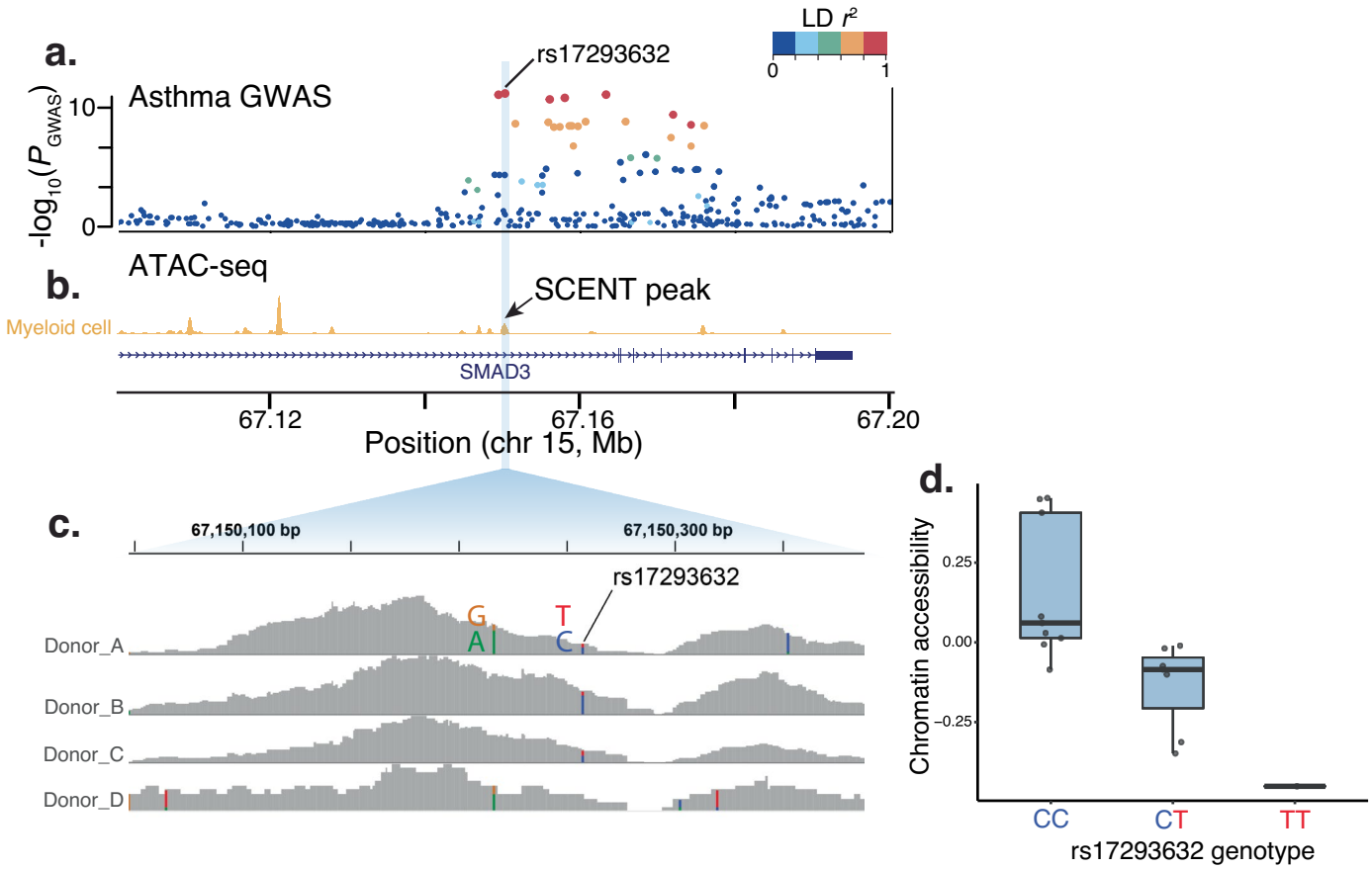


Extended Data Fig. 8 | See next page for caption.

https://doi.org/10.1038/s41588-024-01682-1

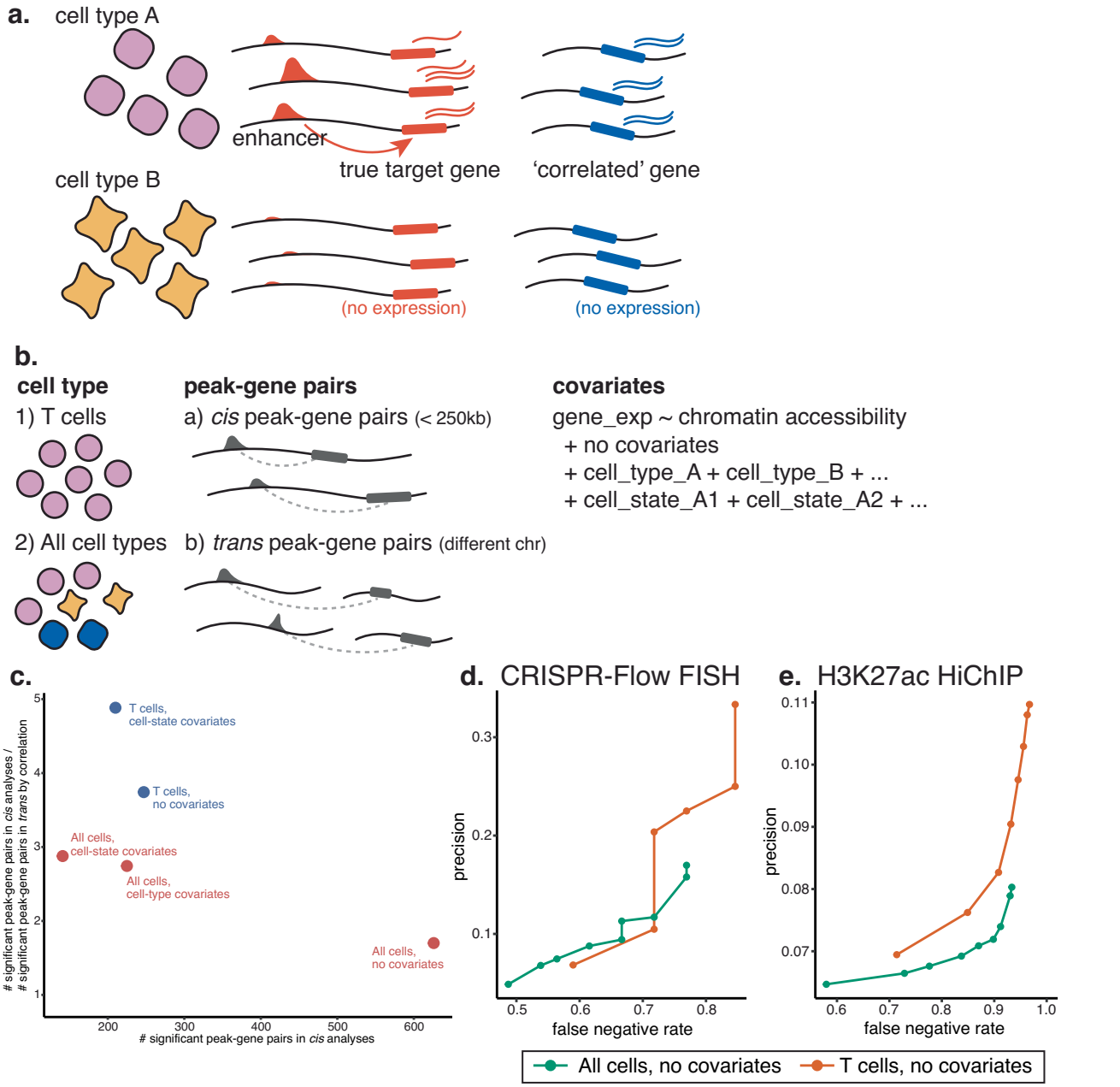
Extended Data Fig. 8 | Causal variant enrichment for GWAS and comparison with published bulk methods and single-cell methods. a. Comparison of the mean causal variant enrichment for FinnGen GWAS (y-axis) among SCENT (teal), EpiMap (red), and ABC model (blue) as a function of the number of significant peak-gene pairs (x-axis) at each threshold of significance. The bars indicate 95% confidence intervals by bootstrapping traits ($n_{bootstrap}$=1000). **b**. We calculated the causal variant enrichment for FinnGen GWAS among SCENT (teal), EpiMap (reds), and ABC model (blues) by changing the PIP thresholds in defining putative causal variants from fine-mapping. The bars indicate 95% confidence intervals by bootstrapping traits ($n_{bootstrap}$=1000). **c** and **d**. The mean causal variant enrichment for GWAS within SCENT enhancers (teal), ArchR (pink) and Signac enhancers (purple). GWAS results were based on FinnGen (**c**) and UK Biobank (**d**) using the FDR < 10% threshold in each software and eight benchmarking datasets (see Methods). The bars indicate 95% confidence intervals by bootstrapping traits ($n_{bootstrap}$=1000).

https://doi.org/10.1038/s41588-024-01682-1



Extended Data Fig. 9 | SMAD3 locus in asthma GWAS. Rs17293632 in asthma GWAS (**a**) was prioritized and connected to *SMAD3* gene by SCENT in myeloid cells (**b**). The panel **a** is a GWAS regional plot, with x-axis representing the position of each genetic variant and y-axis representing -$\log_{10}(P)$ from GWAS (a two-sided *P* value). The rs17293632 has a significant caQTL effect, as shown in **c** and **d**. In panel **c**, the read coverage from single-cell ATAC-seq in each of donors with heterozygous genotype at this accessible region is presented, and at rs17293632, we observed allele-specific increased accessibility with C allele when compared T allele across donors. In panel **d**, normalized chromatin accessibility based on the read coverage for an individual after regressing out covariates is presented by the genotype of rs17293632 (CC, CT and TT). The horizontal bars within boxes indicate the median, and the lower and upper hinges represent 25% and 75% quantile. The upper whisker extends from the hinge to the largest value no further than 1.5 * inter-quartile range (IQR) from the hinge. The lower whisker extends from the hinge to the smallest value at most 1.5 * IQR of the hinge. All individual points are plotted as dots.

https://doi.org/10.1038/s41588-024-01682-1



Extended Data Fig. 10 | Cells to be included in the regression framework. a. An example situation of correlated gene expression without biological regulatory function. **b**. Benchmarking models for statistical power to define biologically plausible peak-gene linkage over false-associations due to correlated genes. **c**. Benchmarking results regarding cells and covariates included in the SCENT regression model. The x-axis represents the number of statistically significant peak-gene linkages among 5,000 randomly selected peak-gene linkages in *cis*, and the y-axis represents the number of statistically significant peak-gene linkages in *cis* divided by the number of statistically significant peak-gene linkages in *trans* among 5,000 randomly selected peak-gene linkages on different chromosomes, as a proxy metric for capability of identifying regulatory elements over 'correlated' elements. Red dots indicate the analyses conducted in all cells including different cell types ($n = 8{,}881$), whereas blue dots indicate the analyses conducted in only T cells ($n = 8{,}881$). **d** and **e**. False positive rate and precision for peak-gene linkages from analyses conducted in all cells (teal) or in only T cells (orange) by using experimentally validated enhancer-gene linkages (that is, CRISPR-Flow FISH data in **d** and H3K27ac data in **e**). False negative rate and precision were defined as follows:

$$false\ negative\ rate = \#\,false\ negative/(\#\,true\ positive + \#\,false\ negative) = 1 - recall$$

nature research

Corresponding author(s): Soumya Raychaudhuri

Last updated by author(s): Jan 4, 2024

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used in data collection.
Data analysis	The computational scripts related to this manuscript are fully available at https://github.com/immunogenomics/SCENT. We also used publicly available software for the data analysis (R v3.6/4.1, bedtools (version 2.26.0), Cell Ranger ARC (version 2.0.0), Seurat (version 4.3.0), Harmony (version 0.1.1), LiftOver (version 2016), ArchR (version 1.0.2), Signac (version 1.9.0), and Cicero (version 1.12.0), SHAPEIT2 (2.727), minimac3 (version 2.0.1), GATK (version 4.1.9.0)).

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The publicly available datasets were downloaded via Gene Expression Ombibus (accession codes: GSE140203, GSE156478, GSE178707, GSE194122, GSE193240, GSE178453) or web repository (https://www.10xgenomics.com/resources/datasets?query=&page=1&configure%5Bfacets%5D%5B0%5D=chemistryVersionAndThroughput&configure%5Bfacets%5D%5B1%5D=pipeline.version&configure%5BhitsPerPage%5D=500&menu%5Bproducts.name%5D=Single%20Cell%20Multiome%20ATAC%20%2B%20Gene%20Expression/). The raw data for arthritis-tissue dataset (single-cell multimodal RNA/ATAC-seq and

single-cell ATAC-seq) is deposited at the NIH Database of Genotypes and Phenotypes (dbGaP accession number: phs003417.v1). The reference human genome GRCh38 genome was used as a reference for read alignment (Ensemble re lease 92).

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	We use all samples post QC for multiome data, both in arthritis-tissue dataset and also in publicly available datasets. We performed downsampling experiments to confirm the statistical power of SCENT we expect at a given number of cells in the multiome data.
Data exclusions	All samples were selected based on quality-control criteria in each cohort, which is summarized in the Method section.
Replication	We compared high replication rate of SCENT across independent different datasets, which is summarized in Supplementary Table 2.
Randomization	We did not need to use randomization in this study as we do not allocate samples into experimental groups. All samples after QC were included in the analysis.
Blinding	We did not apply blinding of the samples because no intervention was conducted in our study.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Synovial tissue samples from 11 RA patients and 1 OA patient were collected from Brigham and Women's Hospital (BWH) and the Hospital for Special Surgery (HSS) for use in the multimodal ATAC + Gene Expression experiments. RA and OA samples with high levels of lymphocyte infiltration (as scored by a pathologist on histologic sections) were identified as "inflamed" and used for downstream analysis. The population characteristics are summarized in Supplementary Table 8.
Recruitment	Synovial tissue samples from 11 RA patients and 1 OA patient were collected and cryopreserved as part of a larger study cohort by the AMP Network for RA and SLE.
Ethics oversight	This study was performed in accordance with protocols approved by the Brigham and Women's Hospital and the Hospital for Special Surgery institutional review boards.

Note that full information on the approval of the study protocol must also be provided in the manuscript.