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

Saori Sakaue[1,2,3], Kathryn Weinand[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[1,2,3,4,14,*]

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
14. Centre for Genetics and Genomics Versus Arthritis, University of Manchester, Manchester, UK

*Address correspondence to:
Soumya Raychaudhuri
77 Avenue Louis Pasteur, Harvard New Research Building, Suite 250D
Boston, MA 02446, USA.
soumya@broadinstitute.org
617-525-4484 (tel); 617-525-4488 (fax)

Abstract

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. We developed a new non-parametric statistical method, SCENT (Single-Cell ENhancer Target gene mapping) which models association between enhancer chromatin accessibility and gene expression in single-cell multimodal RNA-seq and ATAC-seq data. We applied SCENT to 9 multimodal datasets including > 120,000 single cells and created 23 cell-type-specific enhancer-gene maps. These maps were highly enriched for causal variants in eQTLs and GWAS for 1,143 diseases and traits. We identified likely causal genes for both common and rare diseases. In addition, we were able to link 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 variant function.

Main

Introduction

Genome-wide association studies (GWAS) have comprehensively mapped loci for human diseases[1–4]. These loci harbor untapped insights about causal mechanisms 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 that may be present in a single locus, only one or a few may be causal; others are associated since they tag causal variants[2,6,7]. Moreover, causal genes may also be challenging to determine, since causal variants lie in non-coding regions in 90% of the time[8–10], may regulate distant genes[11–13], and may employ context-specific regulatory mechanisms[14–17].

To define causal variants and genes, previous studies have used both statistical and experimental approaches. Statistical fine-mapping[18–23] can narrow the set of candidate causal variants, and is more effective when genetic studies include diverse ancestral backgrounds with different allele frequencies and linkage disequilibrium structures (LD)[24–28]. However, these approaches alone are seldom able to identify true causal variants with confidence[7,23,29–32]. To define causal genes, previous studies have built enhancer-gene maps, that can be used to prioritize causal variants in enhancers and link causal variants to genes they regulate. These maps often require large-scale epigenetic and transcriptomic atlases (e.g., Roadmap[33], BLUEPRINT[34], and ENCODE[35]). By using these atlases, enhancer-gene maps have been built by correlating epigenetic activity (i.e., enhancer activity; e.g., histone mark ChIP-seq and bulk ATAC-seq) with gene expression (e.g., RNA-seq)[36,37], by combining epigenetic activity and probability of physical contact with the gene[38,39], or by integrating multiple linking strategies to

create composite scores [40]. However, current methods largely use bulk tissues or cell lines. Bulk data potentially (i) cannot be easily applied to rare cell populations (ii) obscures the cell-type-specific nature of gene regulation and (iii) requires hundreds of experimentally characterized samples, necessitating consortium-level efforts. While perturbation experiments (e.g., CRISPR interference[41] or base editing[42]) can point to causal links between enhancers and genes, they are difficult to scale because they require the development of cell- or tissue-type specific experimental protocols [43].

Advances in single-cell technologies offer new opportunities for building cell-type specific enhancer-gene maps. Multimodal protocols now enable joint capture of epigenetic activity by ATAC-seq alongside early transcriptional activity with nuclear RNA-seq[44–48] . These methods are easily applied at scale to cells in human primary tissues without disaggregation, making it possible to easily query many samples from disease-relevant tissues. Gene-enhancer maps can be built from this data if links between open chromatin and genes can be accurately established. Since each observation is at a cell-level resolution, statistical power should exceed bulk-tissue-based methods. However, the sparse and non-parametric nature of RNA-seq and ATAC-seq in single-cell experiments makes confident identification of these links challenging. To date, most methods use linear regression models to link enhancers and genes (e.g., ArchR[49] and Signac [50]) despite these sparse and non-parametric features or only utilize co-accessibility of regulatory regions from ATAC-seq but not gene expression data from sc-RNA-seq (e.g., Cicero[51]). These previous methods have not widely demonstrated efficacy in practice for fine-mapping of causal variants and genes in complex traits.

In this context, we developed Single-Cell Enhancer Target gene mapping (SCENT), to accurately map enhancer-gene pairs where an enhancer's activity (i.e. peak accessibility) is associated with gene expression across individual single cells. We use Poisson regression and non-parametric bootstrapping[52] to account for the sparsity and non-parametric distributions. We predicted that peaks with gene associations identified by SCENT are more likely to be functionally important. We apply SCENT to 9 multimodal datasets to build 23 cell-type specific enhancer-gene maps. We show that SCENT enhancers are highly enriched in statistically fine-mapped putative causal variants for eQTL and GWAS. We use SCENT enhancer-gene map to define causal variants, genes, and cell types in common and rare disease loci and somatic mutation hotspots, which has not been previously demonstrated by conventional enhancer-gene mapping based on bulk-tissues.

Results

Overview of SCENT

To identify (1) active *cis*-regulatory regions and (2) their target genes (3) in a given cell type, we leveraged single-cell multimodal datasets. SCENT accurately identifies significant association between chromatin accessibility of regulatory regions (i.e., peaks) from ATAC-seq and gene expression from RNA-seq across individual single cells (**Figure 1a**). Those associations can be used for prioritizing (1) putative causal variants if they are in regulatory regions that are associated with expression of a gene, (2) putative causal genes if they are associated with the identified regulatory region and (3) the critical cell types based on which map the relevant

association is identified in. We assessed whether binarized chromatin accessibility in each of the ATAC peaks is associated with expression counts for each gene in *cis* (<500kb from gene body), testing one peak-gene pair at a time in each cell type (see **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.

Since both RNA-seq and ATAC-seq data are generally sparse[50,53–56], we used Poisson regression to model the effect of chromatin accessibility on gene expression accounting for donor effects and cell-level factors capturing data quality, such as pesssrcentage of mitochondrial reads. However, gene expression counts are highly variable across genes (**Figure 1b**; **Supplementary Figure 1a**). Genes that are highly expressed and dispersed might inflate association test statistics. This inflation was even apparent in data where we permuted cell barcodes to disrupt any underlying association between ATAC and RNA profiles (**Supplementary Figure 1b**). Common analytical statistical models (e.g., linear, negative binomial and Poisson regression) all demonstrated inflated statistics (**Supplementary Figure 1c-e**). Therefore, to accurately estimate the error and significance of the effect of each peak-gene pair, we implemented bootstrapping framework (i.e., resampling cells with replacement; see **Methods**). This resulted in well-calibrated statistics with appropriate type I error (**Supplementary Figure 1f**). For each accessible chromatin peak and gene expression in each cell type, SCENT estimates an effect size, reflecting the strength of the regulatory effect, and sign, reflecting enhancing versus silencing effects.

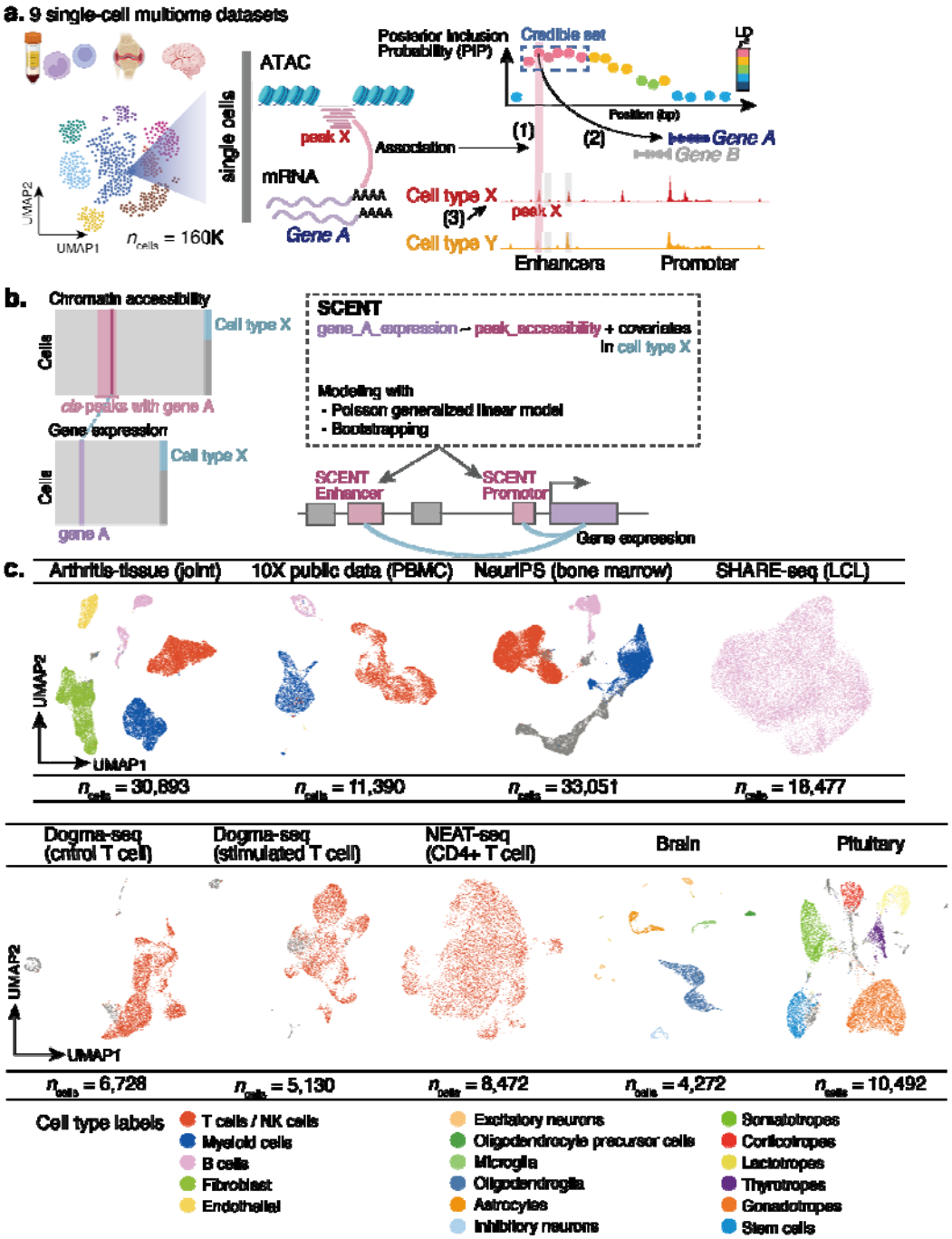

Figure 1. Schematic overview of SCENT and SCENT enhancer-gene pairs across 9 single-cell multimodal datasets. a. SCENT identifies (1) active *cis*-regulatory regions and (2) their target genes in (3) a specific cell type. Those SCENT results can be used to define causal variants, genes, and cell types 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**. 9 single-cell datasets on which we applied SCENT to create 23 cell-type-specific enhancer-gene map. The cells in each dataset are described in UMAP embeddings from RNA-seq and colored by cell types.

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

We obtained a total of nine single-cell multimodal datasets from diverse human tissues representing 13 cell-types (immune-related, hematopoietic, neuronal, and pituitary). Since we are interested in rheumatoid arthritis and other autoimmune diseases, we created a disease-relevant inflammatory tissue dataset by obtaining inflamed synovial tissues from ten rheumatoid arthritis (RA) and two osteoarthritis (OA) patients (arthritis-tissue dataset; n_{donor} = 12). Applying stringent QC to these multimodal data, we obtained information on 30,893 cells (see **Methods**). In addition, we obtained eight public data sets with 129,672 cells. In total we had data from 160,565 cells [46,57–61]. We analyzed 16,621 genes and 1,193,842 open chromatin peaks in *cis* after QC (4,753,521 peak-gene pairs; **Figure 1c**, **Supplementary Table 1**). After clustering cells and annotating them with cell labels, we applied SCENT individually to each of the cell types with n_{cells} > 500 within each dataset to construct 23 enhancer-gene maps. SCENT identified 87,648 cell-type-specific peak-gene links (false discovery rate (FDR) < 10%, **Figure 2a**, **Supplementary Figure 2**). Each gene had variable number of associated peaks in *cis* (from 0 to 97, mean = 4.13, **Supplementary Figure 3a**).

To assess replicability of SCENT peak-gene links across datasets, we used arthritis-tissue dataset, which had the largest number of significant peak-gene pairs as a stringent discovery dataset. We compared the effects from the arthritis-tissue dataset in the same cell-type with those from other datasets (B cell, T/NK cell and myeloid cell; **Supplementary Table 2a;** see **Methods**). Despite different tissue contexts, we confirmed high directional concordance of the effect of chromatin accessibility on gene expression for

peak-gene pairs significant in both datasets (mean Pearson *r* = 0.62 of effect sizes, 99% mean concordance across all the datasets: **Supplementary Figure 3b**). For comparison, we tested two popular linear parametric single-cell multimodal methods, ArchR[56] or Signac[50]. When comparing with the same discovery and replication data (i.e., arthritis-tissue dataset as a discovery and public PBMC as a replication), we noted lower directional concordance and effect correlation in these previous methods than in SCENT (*55*)(mean Pearson's *r* = 0.31, 62% mean directional concordance in ArchR and *r* = 0.24, 98% mean directional concordance in Signac; **Supplementary Table 2b** and **c**). These results argue that SCENT can more reproducibly detect enhancer-gene links compared with previous parametric methods for single-cell multimodal data.

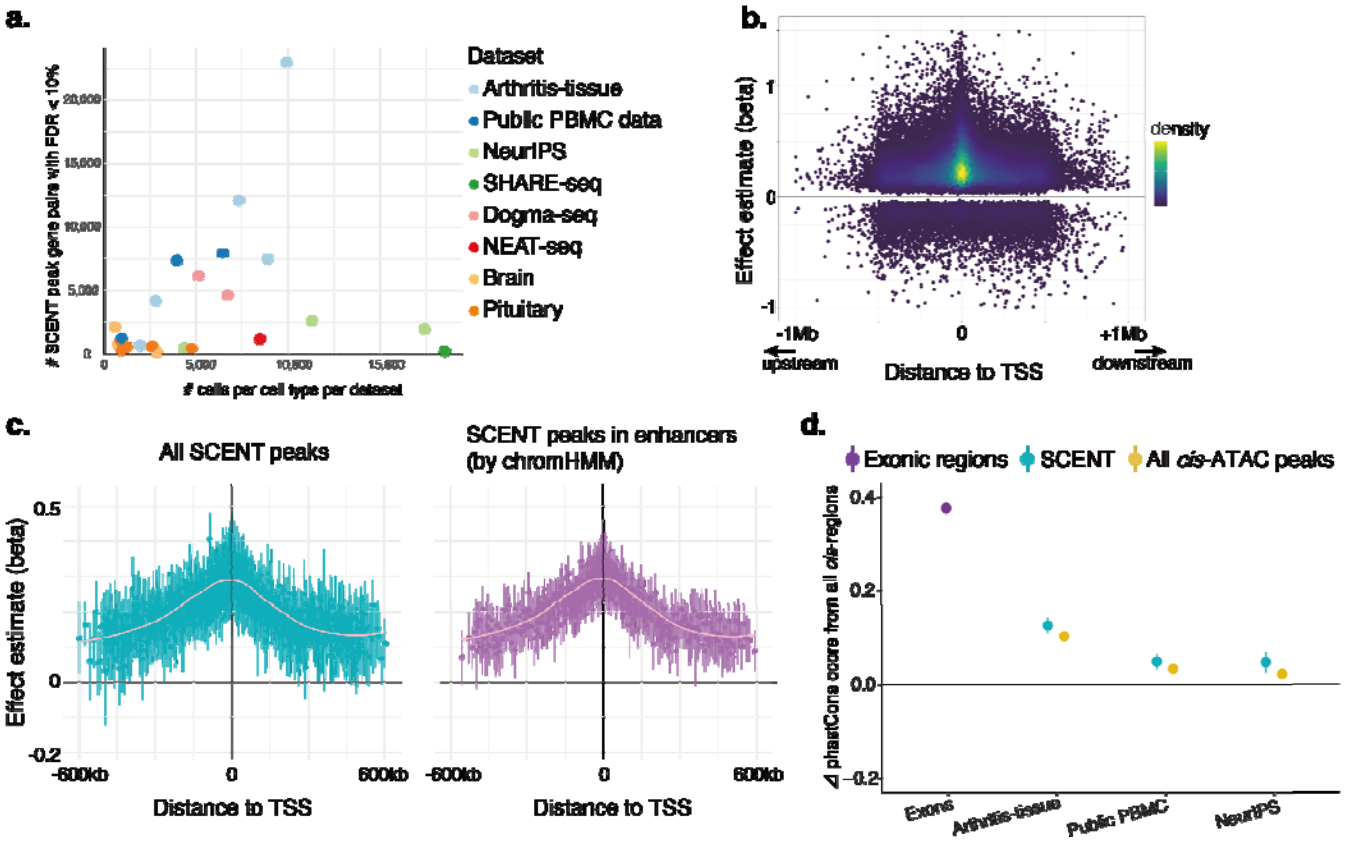

Figure 2. SCENT identified functionally active and evolutionary conserved *cis*-regulatory regions from single-cell multimodal data.

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 (beta) of chromatin accessibility on the gene expression from Poisson regression (y-axis). Each dot is a significant gene-peak pair and plotted against the distance between the peak and the transcription start site (TSS) of the gene, colored as a density plot. **c.** The mean effect size (beta) of chromatin accessibility on the gene expression in arthritis-tissue dataset within each bin of TSS distance. Left; all significant gene-peak links. Right; SCENT peaks within enhancers identified using chromHMM in immune-related tissues. **d**. Mean phastCons score difference (phastCons score) between each annotated region and all *cis*-regulatory non-coding regions. We show the phastCons score for exonic regions (purple) as a reference, and for SCENT (green) and all *cis*-ATAC peaks (yellow) enhancers in each multimodal dataset.

To assess if peaks (i.e., *cis*-regulatory regions) identified through SCENT were functional, we examined if (1) they co-localized with conventional *cis*-regulatory annotation, (2) the effect of peaks on expression was greater for closer peak-gene pairs, (3) SCENT peaks had high sequence conservation, and (4) peak-gene connections were more likely to be validated experimentally.

First, we tested how many of the SCENT peaks overlapped with an ENCODE cCRE[62], a conventional *cis*-regulatory annotation defined by bulk-based epigenomic datasets. We observed that 98.0% of the SCENT peaks overlapped with ENCODE cCRE on average, compared to 23.3% of random *cis*-regions matched for size and 89.0% of non-SCENT peaks (**Supplementary Figure 3c**).

Second, we examined the strength and direction of enhance-gene links, hypothesizing that stronger links would be more proximal to the transcription start site (TSS) of target genes. We observed that the regression coefficient β_{peak} (the effect size of peak accessibility on gene expression) becomes larger and more positive as the SCENT *cis*-regulatory elements (peaks) get closer to the TSS (**Figure 2b** and **Figure 2c,** left panel), consistent with previous observations[56,63]. We annotated *cis*-regulatory regions identified by SCENT with 18-state chromHMM results from 41 immune-related samples in ENCODE consortium[37]. When we subset peaks to those within enhancer annotations, we observed a clearer decay in effect size as a function of TSS distance (**Figure 2c,** right panel).

Third, we assessed whether SCENT peaks had higher sequence conversation across species. We reasoned that SCENT captures functionally important and thus evolutionary

conserved regions. The evolutionary conserved regulatory regions are known to be enriched for complex trait heritability[64]. We used evolutionary conservation metric, phastCons[65] to assess sequence conservation at SCENT peaks. As expected, exonic regions were much more evolutionary conserved than all non-coding *cis*-region (mean Δ phastCons score = 0.38, paired t-test $P < 10^{-323}$; **Figure 2d**, purple). The SCENT regulatory regions were also conserved relative to non-coding *cis*-regions (mean Δ phastCons score = 0.13, paired t-test $P = 4.2 \times 10^{-42}$ in arthritis-tissue dataset; **Figure 2d**, green). In contrast, the Δ phastCons score between all *cis*-ATAC peaks and all non-coding *cis*-region was more modest (mean Δ phastCons score = 0.092, paired t-test $P = 8.7 \times 10^{-27}$ in arthritis-tissue dataset; **Figure 2d**, yellow). To test if the higher conservation in SCENT peaks were driven by their proximity to TSS, we assessed Δ phastCons score between SCENT peaks and non-SCENT peaks with matching peaks on TSS distance. SCENT peaks had significantly higher conservation scores than the non-SCENT peaks when we matched on TSS distance (mean Δ phastCons score = 0.034, $P = 4.7 \times 10^{-4}$ in arthritis-tissue dataset; **Supplementary Figure 3d**; see **Methods**). The higher sequence conservation suggested 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. We used Nasser et al.[39] CRISPR-Flow FISH results which included 278 positive enhancer-gene connections where perturbation decreased connections, and 5,470 negative connections. We observed that the SCENT peaks were > 4-fold enriched relative to non-SCENT peaks for positive connections (4.5X, Fisher's exact

P=1.8x10^{-9}, arthritis-tissue dataset and 4.5X, *P*=1.0x10^{-8} in public PBMC dataset; **Methods, Supplementary Table 3**).

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. The genes with the most SCENT peaks per gene included *FOSB* (*n* = 97), *JUNB* (*n* = 95), and *RUNX1* (*n* = 77), critical and highly conserved transcription factors. We used mutational constraint metrics based on the absence of deleterious variants within human populations (i.e., the probability of being loss-of-function intolerant (pLI) [66] and the loss-of-function observed/expected upper bound fraction (LOEUF)[67]). We observed that the normalized number of SCENT *cis*-regulatory elements per gene is strongly associated with mean constraint score for the gene (beta = 0.37, *P* = 4.9x10^{-90} for pLI where higher score indicates more constraint, and beta = -0.35, *P* = -0.35x10^{-106} for LOEUF where lower score indicates more constraint; **Supplementary Figure 4a** and **4b**, respectively). Previous results have shown that genes with many regulatory regions based on bulk-epigenomic data had been enriched for loss-of-function intolerant genes[68]. We were able to replicate the same trend by using the single-cell multimodal datasets and SCENT.

Enrichment of eQTL putative causal variants in SCENT peaks

We sought to examine whether the SCENT peaks are likely to harbor statistically fine-mapped putative causal variants for expression quantitative loci (eQTL). We analyzed tissue-specific eQTL fine-mapping results from GTEx consortium across 49 tissues[69]. We used statistical fine-mapping results (posterior inclusion probability [PIP] > 0.2) to define putative causal

variants. We then tested enrichment statistics within ATAC peaks or SCENT peaks (see **Methods**). Unsurprisingly, all the accessible regions defined by ATAC-seq in *cis*-regions were modestly enriched in fine-mapped variants by 2.7X (yellow, **Figure 3a**). However, SCENT peaks were more strikingly enriched in fine-mapped variants by 9.6X on average across all datasets (green, **Figure 3a**). Using more stringent PIP threshold cutoffs (0.5 and 0.7) to define putative causal variants resulted in even stronger enrichments (**Supplementary Figure 5**).

Since many SCENT peaks are close to TSS regions, we considered whether this enrichment might be driven by TSS proximity (**Figure 2c**, **Supplementary Figure 6a**). To test this, we matched each of the SCENT peak-gene pairs to one non-SCENT peak-gene pair that had the most similar TSS distance (**Supplementary Figure 6b**). We compared the fine-mapped variant enrichment between those two sets of peak-gene pairs with matched TSS distance. We observed that SCENT peaks consistently had higher enrichment in all analyzed datasets (**Supplementary Figure 6c**) than TSS-distance-matched non-SCENT peaks (e.g., 12.3X in SCENT vs. 9.64X in distance-matched non-SCENT in arthritis-tissue dataset). This suggests that SCENT has additional information in identifying biologically important cis-regulatory regions beyond TSS distance.

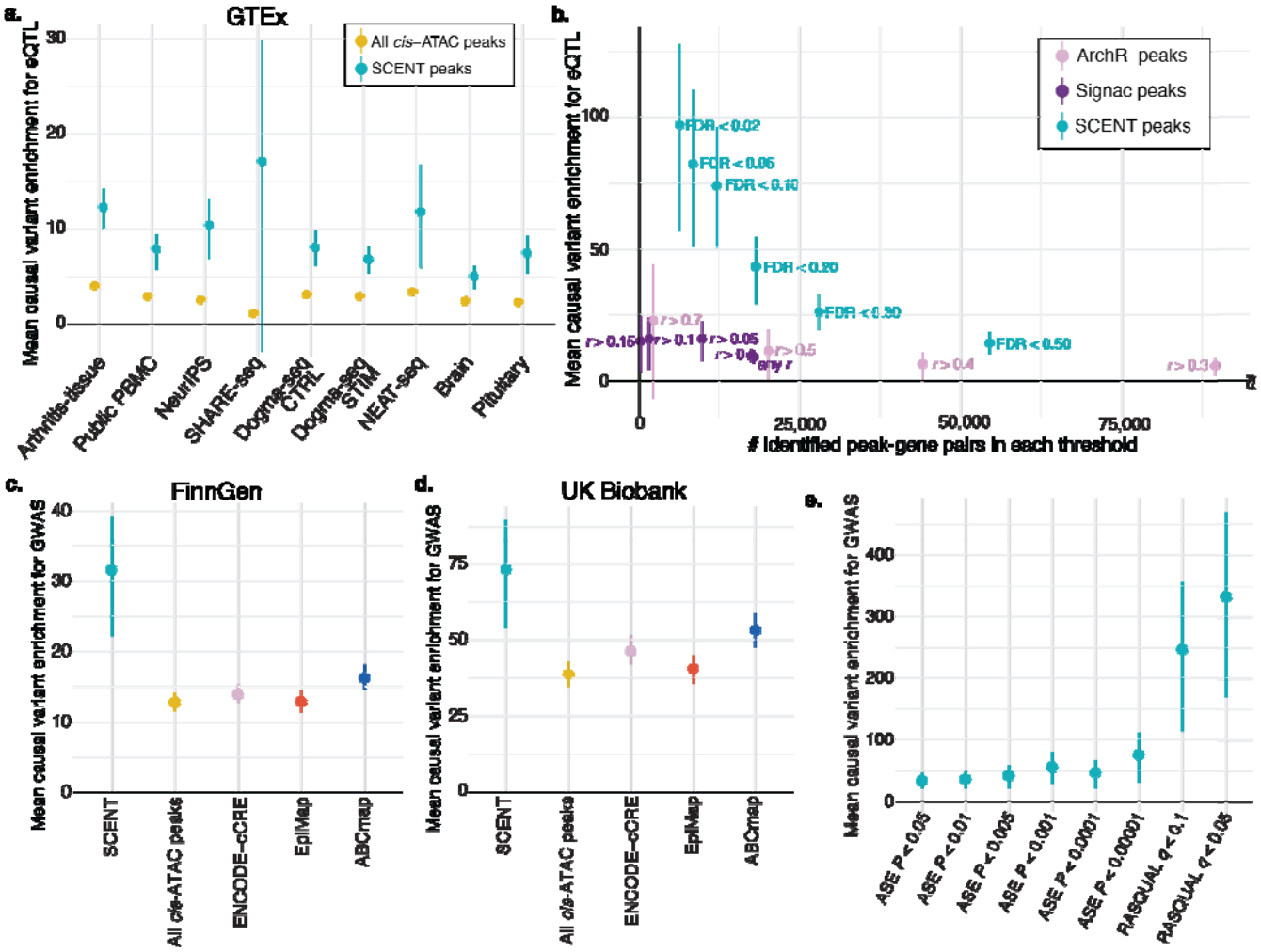

Figure 3. SCENT enhancers are enriched in causal variants of eQTL and GWAS.
a. The mean causal variant enrichment for eQTL within SCENT peaks or all ATAC-seq peaks in each of the 9 single-cell datasets. The bars indicate 95% confidence intervals by bootstrapping genes. **b**. Comparison of the mean causal variant enrichment for eQTL (y-axis) between SCENT (green), ArchR (pink), and Signac (purple) as a function of the number of significant peak-gene pairs at each threshold of significance. The bars indicate 95% confidence intervals by bootstrapping genes. The ArchR results with > 100,000 peak-gene linkages are omitted, and full results are in **Supplementary Figure 6d**. **c** and **d**. The mean causal variant enrichment for GWAS within SCENT enhancers (green), 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% confidence intervals by bootstrapping traits. **e**. The mean causal variant enrichment for FinnGen GWAS within intersection of SCENT enhancers and caQTL enhancers at each threshold of significance. The bars indicate 95% confidence intervals by bootstrapping traits.

We next compared the enrichment for eQTL causal variants in SCENT peaks to peaks identified by linear parametric methods, ArchR[56] and Signac[50] using myeloid cells in the arthritis-tissue dataset. ArchR and Signac peaks had substantially lower causal variant enrichment for eQTL in blood (1.4X and 9.3X, respectively) compared to SCENT peaks (74.1X) in arthritis-tissue dataset. By varying the thresholds to define significant peak-gene associations (correlation *r* in ArchR and FDR in SCENT), we assessed the number of peak-gene pairs and the causal variant enrichment for eQTLs within these peaks (**Figure 3b** and **Supplementary Figure 6d**). Indeed, SCENT peaks consistently demonstrated higher causal variant enrichment than ArchR peaks and Signac peaks.

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 per-cell-type SCENT enhancers across all datasets. We observed that the cell-type-specific SCENT enhancers (e.g., SCENT B cell peaks) were most enriched in putative causal eQTL variants within relevant samples in GTEx (e.g., EBV-transformed lymphocytes; **Supplementary Figure 6e**).

These results suggest that SCENT can prioritize regulatory elements harboring putative causal eQTL variants in a cell-type-specific manner, with higher precision than the previous single-cell based method.

Enrichment of GWAS causal variants in SCENT enhancers

Given that the combination of SCENT and multimodal data from disease-relevant tissues can be used to quickly build disease-specific enhancer-gene maps, we sought to examine whether SCENT peaks can be used for the more difficult task of prioritizing disease causal variants. To identify candidate causal variants (PIP>0.2), we used fine-mapping results from complex trait loci from GWAS in two large-scale biobanks (FinnGen[70] [1,046 disease traits] and UK Biobank[71] [35 binary traits and 59 quantitative traits])[28]. We defined enrichment statistics for GWAS causal variants within SCENT enhancers (see **Methods**). The SCENT enhancers were strikingly enriched in causal GWAS variants in FinnGen (31.6X on average; 1046 traits; **Figure 3c** and **Supplementary Figure 7a**) and UK Biobank (73.2X on average; 94 traits; **Figure 3d** and **Supplementary Figure 7b**). This enrichment was again much larger than all *cis*-ATAC peaks (12.8X in FinnGen and 38.8X in UK Biobank). Moreover, the target genes of causal variants for autoimmune diseases (AID) identified by SCENT enhancer-gene map in immune-related cell types had higher fraction (10.8%) of know genes implicated in Mendelian disorders of immune dysregulation (n_{gene} = 550)[72,73] than SCENT enhancer-gene map in fibroblast (3.8%; **Supplementary Figure 7c**).

We compared SCENT to other alternative genome annotations and enhancer-gene maps that have recently emerged. Causal variant enrichment in SCENT was much higher than the conventional bulk-based annotations such as ENCODE cCREs (13.9X in FinnGen and 46.5X in UK Biobank), ABC (16.3X in FinnGen and 53.3X in UK Biobank) and EpiMap, (12.9X in FinnGen and 40.6X in UK Biobank; **Figure 3c** and **3d, Supplementary Figure 7a** and **b)**. We varied thresholds to assess recall and precision tradeoffs (FDR in SCENT, ABC score in

ABC model and EpiMap correlation score in EpiMap) for identifying causal GWAS variants (**Supplementary Figure 8a**). We constructed SCENT from 9 datasets with only 28 samples, substantially less than the 833 samples used to construct EpiMap and 131 samples for the ABC model. Despite this, SCENT peaks consistently demonstrated higher enrichment of causal GWAS variants at a similar number of identified peak-gene linkages than ABC model and EpiMap. A more stringent PIP threshold (0.5 and 0.7) to define putative causal variants further increased the enrichment statistics while maintaining the higher enrichment in SCENT than bulk methods (**Supplementary Figure 8b**). We tested whether causal genes are being identified by examining a set of known genes implicated in Mendelian disorders of immune dysregulation. We observed that the target genes for AID identified by SCENT enhancer-gene map in immune-related cell types had higher fraction (10.8%) of implicated known Mendelian genes [72,73] than EpiMap (8.6%) and ABC model (4.4%) enhancer-gene map (**Supplementary Figure 7c**). These results demonstrate the power SCENT achieved by accurately modeling association between chromatin accessibility and gene expression at the single-cell resolution.

We hypothesized that disease-causal variants prioritized by SCENT would likely modulate chromatin accessibility (e.g., transcription factor binding affinity). If so, the intersection of the SCENT enhancers and chromatin accessibility quantitative trait loci (caQTL) could further enrich the causal GWAS variants[74–77]. To test this hypothesis, we used single-cell ATAC-seq samples with genotype (n_{donor} = 17; arthritis-tissue dataset) from the same tissue and performed caQTL mapping by leveraging allele-specific (AS) chromatin accessibility (binomial test followed by meta-analysis across donors) or by combining AS with inter-individual differences

(RASQUAL[78]; see **Methods**). We then defined caQTL ATAC peaks with variable thresholds and intersected them with the SCENT enhancers. We calculated the causal GWAS variant enrichment within these intersected regions. We observed drastically higher enrichment of causal variants as we used more stringent threshold in defining caQTL peaks, reaching as high as 333-fold enrichment (**Figure 3e**). This suggested that SCENT efficiently prioritized causal GWAS variants in part by capturing regulatory regions of which chromatin accessibility is perturbed by genetic variants and modulates gene expression. SCENT demonstrated a potential to further enrich causal variants by using caQTLs if the multimodal data has matched genotype data.

Defining mechanisms of GWAS loci by SCENT

We finally sought to use SCENT enhancer-gene links to define causal mechanisms of complex trait GWAS. We used the fine-mapped variants from GWASs (FinnGen, UK Biobank traits and GWAS fine-mapping results of rheumatoid arthritis (RA)[26], inflammatory bowel disease[29] and type 1 diabetes (T1D)[79]). SCENT linked 4,124 putative causal variants (PIP > 0.1) to their potential target genes across 1,143 traits in total (**Supplementary Table 4**).

We first focus on autoimmune loci, given that our current SCENT tracks are largely derived from immune cell types and inflammatory tissues. We prioritized a single well fine-mapped variant rs72928038 (PIP > 0.3) at 6q15 locus in multiple autoimmune diseases (RA, T1D, atopic dermatitis and hypothyroidism), within the T-cell-specific SCENT enhancer (T cells in Public PBMC datasets and Dogma-seq control and stimulated T cells; **Figure 4a**). This

enhancer was linked to *BACH2*, which was also the closest gene to this fine-mapped variant. Notably, base-editing in T cells has confirmed that this variant affects *BACH2* expression[80]. Moreover, editing of this variant into CD8 T cells skewed naive T cells toward effector T cell fates[80].

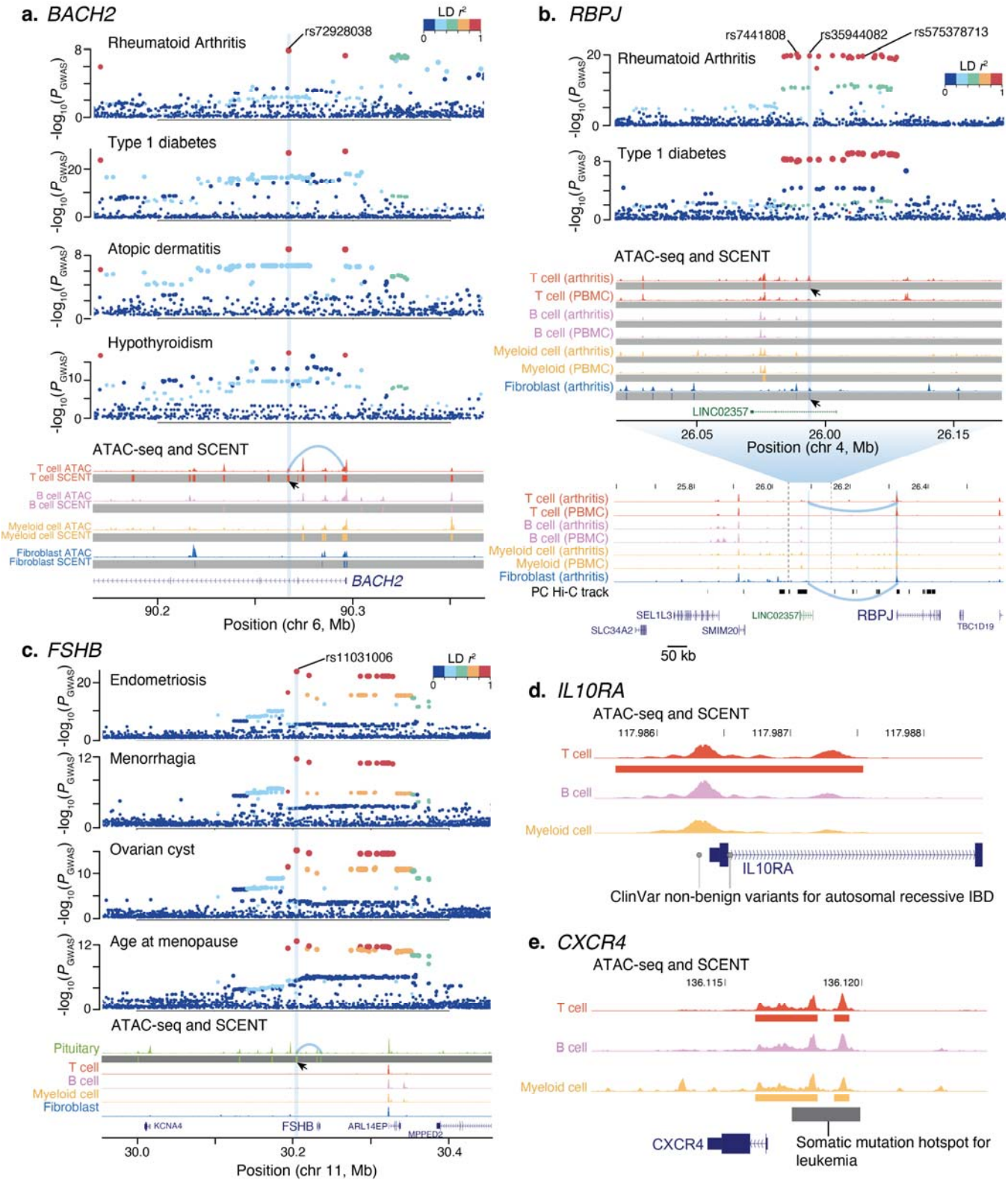

Figure 4. SCENT defined causal variants and genes in complex trait GWAS.
a. Rs72928038 at *BACH2* locus was prioritized by T-cell-specific SCENT enhancer-gene map, being for RA, T1D, Atopic dermatitis and hypothyroidism. The top four panels are GWAS

regional plots, with x-axis representing the position of each genetic variant. The color of the dots represent 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 grey stripes) in each cell type (public PBMC dataset for immune cell types and arthritis-tissue dataset for fibroblast). An arrow head indicates the SCENT peak overlapping with fine-mapped variant. **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 arthritis-tissue dataset. The top two panels are GWAS regional plots similarly to panel **a**. ATAC-seq and SCENT tracks are shown similarly to panel **a**, but 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 panel **a**. ATAC-seq and SCENT tracks are shown similarly to panel **a**, and include tracks from pituitary dataset. There were no SCENT peaks in cell types except for pituitary. **d**. ATAC-seq and SCENT tracks for *IL10RA* locus, where non-coding ClinVar variants (grey dots) colocalized with T-cell SCENT track. **e**. ATAC-seq and SCENT tracks for *CXCR4* locus, where somatic mutation hotspot for leukemia colocalized with T-cell and myeloid-cell SCENT tracks.

Another locus for RA and T1D at chr 4p15.2 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 (**Figure 4b**). SCENT linked this variant to *RBPJ*, which was the 3rd closest gene to this variant located 235kb away. This variant-gene link was supported by a physical contact from promotor-capture Hi-C contact data measured in hematopoietic cells[81]. *RBPJ* (recombination signal binding protein for immunoglobulin kappa J region) is a transcription factor critical for NOTCH signaling, which has been implicated in RA tissue inflammation through functional studies[82,83]. *Rbpj* knockdown in mice resulted in abnormal T cell differentiation and disrupted regulatory T cell phenotype[84,85], consistent with a plausible role gene in autoimmune diseases. Intriguingly, we observed no SCENT peaks in T cells from PBMC or blood to prioritize causal variants at this locus. This linkage was not present in EpiMap. ABC map prioritized another variant, rs7441808 at this locus and linked it non-specifically to 16 genes including *RBPJ*, making it difficult to define the true causal gene.

In a final example, we highlight the value of using SCENT to build enhancer-gene maps from disease-critical tissues. We examined the enhancer-gene map produced from single-cell multimodal pituitary data[61] to assess the 11p14.1 locus for multiple gynecological traits (endometriosis, menorrhagia, ovarian cyst and age at menopause). Our map connected rs11031006 to *FSHB* (follicle stimulating hormone subunit beta) (**Figure 4c**), which is specifically expressed in the pituitary[69,86] and enables ovarian folliculogenesis to the antral follicle stage[87]. Rare genetic variants within *FSHB* are known to cause autosomal recessive

hypogonadotropic hypogonadism[88]. However, the other multimodal datasets and bulk-based methods (ABC model and EpiMap) were unable to prioritize and connect this variant to *FSHB*. *FSHB* example showed the potential of SCENT for defining causal variants and genes by being applied to disease-relevant tissues.

Mendelian-disease variants and somatic mutations in cancer within SCENT enhancers

Having established the SCENT's utility in defining causal variants and genes in complex diseases, we examined rare non-coding genetic variants causing Mendelian diseases. Currently, causal mutations and genes can only be identified in ~30–40% of patients with Mendelian diseases [89–91]. Consequently, many variants identified in case individuals are annotated as variants of uncertain significance (VUS). The VUS annotation is especially challenging for non-coding variants. We asked whether our SCENT enhancers overlapped with clinically reported non-benign non-coding variants by ClinVar[92] (400,300 variants in total). The SCENT enhancers harbored 2.0 times ClinVar variants on average than all the ATAC regions with the same genomic length across all the datasets we investigated (**Supplementary Figure 9**). This density of ClinVar variants was 3.2 times and 12 times on average larger than that in ENCODE cCREs and of all non-coding regions, respectively. We thus defined 3,724 target genes for 33,618 non-coding ClinVar variants by SCENT in total (**Supplementary Table 5**). As illustrative examples, we found 40 non-coding variants linked to *LDLR* gene causing familial hypercholesterolemia 1[92], 3 non-coding variants linked to *IL10RA* causing autosomal recessive

early-onset inflammatory bowel disease 28 (**Figure 4d**)[93], and an intronic variant rs1591491477 linked to *ATM* gene causing hereditary cancer-predisposing syndrome[92].

Finally, we used SCENT to connect non-coding somatic mutation hotspots to target genes. Recently, somatic mutation analyses across the entire cancer genome revealed possible driver non-coding events[94]. Among 372 non-coding somatic hotspots in 19 cancer types, SCENT enhancers included 193 cancer-mutation hotspot pairs (**Supplementary Table 6**). SCENT enhancer-gene linkage successfully linked those hotspots to known driver genes (e.g., *BACH2*, *BCL6*, *BCR*, *CXCR4* (**Figure 4e**), and *IRF8* in leukemia). In some instances, SCENT nominated different target genes for those 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 *IGHA1* (Immunoglobulin Heavy Constant Alpha 1) which might be more biologically relevant than *ADAM6* nominated by ABC model. These results implicate broad applicability of SCENT for annotating all types of human variations in non-coding regions.

Augmenting SCENT enhancer-gene maps with more samples

While the recall for enhancer-gene maps defined by SCENT was lower than that by bulk-tissue-based methods, we felt that this might be a function of current limited sample sizes. We wanted to assess if the addition of more cells into SCENT might lead to the higher recall for enhancer-gene maps while retaining the precision. To assess this, we downsampled 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 sc-multimodal datasets increases (**Supplementary Figure 10**). We considered the possibility that enhancer-gene maps with greater numbers of cells might capture spurious associations; if this was the case, we would expect more long-range associations, which are more likely to be false positives with greater cell numbers. In contrast, we observed that shorter-range and longer-range associations were both equivalently represented as we added additional cells, suggesting the robustness of our discovery.

Discussion

In this study, we presented a novel statistical method, SCENT, to create a cell-type-specific enhancer-gene map by using single-cell multimodal data. Single-cell RNA-seq and ATAC-seq are both sparse and have variable count distributions, which requires non-parametric bootstrapping to connect chromatin accessibility with gene expression. The SCENT model demonstrated well-controlled type I error, outperforming commonly used statistical models which showed inflated statistics. SCENT mapped enhancers that showed strikingly high enrichment for putative causal variants in eQTLs and GWASs and outperformed previous methods analyzing single-cell multimodal data (e.g., ArchR[49] and Signac[50]). Despite using substantially lower number of samples (28 from 9 datasets in total), enhancers defined by SCENT had equivalent or even higher enrichment for causal variants than bulk-tissue-based

methods with more than 100 samples (e.g., EpiMap and ABC model), by modeling single-cell level observations instead of obscuring them into sample-level association.

As potential limitations, first, our enhancer-gene maps had relatively fewer enhancers identified compared to other resources (**Figure 2a**). However, downsampling experiments showed a clear linear relationship between the number of cells per cell type per dataset and the number of significant SCENT peak-gene links. It follows that SCENT applied to larger datasets from a diverse set of tissues will further expand the current enhancer-gene map. In contrast, bulk-tissue-based enhancer-gene map might have an upper limit of discovery by the number of samples generated by each consortium (e.g., ENCODE). Second, SCENT focuses on gene cis-regulatory mechanisms to fine-map disease causal alleles, while there could be other causal mechanisms that explain disease heritability, such as alleles that act through trans-regulatory effects, splicing effects, or post-transcriptional effects[95].

We argue that the real utility of SCENT is that it enables the rapid construction of disease-tissue-relevant enhancer-gene maps. Multimodal single cell data can be easily obtained from a wide range of primary human tissues. Since these assays query nuclear material, data can be obtained without disaggregating tissues. Hence, the data is likely robust to the effects of disaggregation which is employed for assays that need intact cells from tissue. Therefore, it is possible to build relevant tissue-specific enhancer-gene maps that are necessary to understand the causal mechanisms of common diseases from GWAS, rare diseases from mapping Mendelian diseases, and somatic non-coding mutations in cancers. For

example, understanding the *FSHB* locus in gynecological traits specifically required a pituitary map, and *RBPJ* locus in RA specifically required a synovial tissue map.

In summary, our method SCENT is a robust, versatile method to efficiently define causal variants and genes in human diseases and will fill the gap in the current enhancer-gene map built from genomic data in bulk tissues.

Data Availability

The publicly available datasets were downloaded via Gene Expression Ombibus (accession codes: GSE140203, GSE156478, GSE178707, 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, https://openproblems.bio/neurips_docs/data/dataset/). The raw data for arthritis-tissue dataset (single-cell multimodal RNA/ATAC-seq and single-cell ATAC-seq) will be publicly available before the acceptance of this manuscript.

Code Availability

The computational scripts related to this manuscript are available at https://github.com/immunogenomics/SCENT.

519 **Methods**

520 *Data and sample in arthritis-tissue dataset*

521 This study was performed in accordance with protocols approved by the Brigham and Women's
522 Hospital and the Hospital for Special Surgery institutional review boards. Synovial tissue from
523 patients with RA and OA were collected from synovectomy or arthroplasty procedures followed
524 by cryopreservation as previously described[96]. RA samples with high levels of lymphocyte
525 infiltration (as scored by a pathologist on histologic sections) were identified as "inflamed" and
526 used for downstream analysis. Next, cryopreserved synovial tissue fragments were dissociated
527 by a mechanical and enzymatic digestion[96], followed by flow sorting to enrich for live synovial
528 cells. For each tissue sample, 15,000 viable cells were isolated and lysed to extract nuclei
529 according to manufacturer protocol (10X Genomics). Joint sc-RNA- and sc-ATAC-seq libraries
530 were prepared using the 10x Genomics Single Cell Multiome ATAC + Gene Expression kit
531 according to manufacturer's instructions. Libraries were sequenced with paired-end 150-bp
532 reads on an Illumina Novaseq to a target depth of 20,000 read pairs per nuclei for mRNA
533 libraries and 25,000 read pairs per nucleus for the ATAC libraries. Demultiplexed scRNA-seq
534 fastq files were inputted into the Cell Ranger ARC pipeline (version 2.0.0) from 10x Genomics
535 to generate barcoded count matrix of gene expression. For ATAC-seq, we trimmed adaptor and
536 primer sequences and mapped the trimmed reads to the hg38 genome by BWA-MEM with
537 default parameters. To deduplicate reads from PCR amplification bias within a cell while
538 keeping reads originating from the same positions but from different cells, we used in-house
539 scripts (manuscript in preparation).

Uniform processing 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 ($n_{dataset}$ = 9, as of April 2022). We processed these downloaded count matrices of gene expression and ATAC data. Briefly, we applied QC to both the nuclear RNA data and the ATAC data based on RNA counts, ATAC fragments, nucleosome signal, and TSS enrichment (**Supplementary Table 7**). 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 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. Gene expression counts were normalized using the NormalizeData function (Seurat[97]), scaled using the ScaleData function (Seurat), and batch corrected using Harmony[98]. We visualized the cells in two low-dimensional embeddings with UMAP by using 20 batch-corrected principal components from these normalized gene expression matrices (**Figure 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 bp from the target gene's TSS. We modeled the association between peak's binarized accessibility and the target gene's expression with Poisson distribution:

$$E_i \sim Poisson(\lambda_i)$$

$$log(\lambda_i) = \beta_0 + \beta_{peak} X_{peak} + \beta_{\%mito} X_{\%mito} + \beta_{nUMI} X_{nUMI} + \beta_{batch} X_{batch}$$

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 i th gene expression. $\beta_{\%mito}$, β_{nUMI}, and β_{batch} each represents the effect of covariates, percentage of mitochondrial reads per cell as a mesure of cell quality, the number of UMIs in the cell, and the batch, respectively. To assess significance of β_{peak}, we used bootstrapping procedures, where we resampled cells with replacement in each procedure and estimate β'_{peak} within those cells. To reduce computational burden, we adaptively increased the number of from at least 100 and up to 10,000, depending on the significance of β_{peak} in each chunk of bootstrapping trials. To avoid spurious associations from rare ATAC peak and rare gene expression, we QCed 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 FDR (Benjamini & Hochberg correction).

When we tested the calibration of statistics from SCENT or other regression strategies (**Supplementary Figure 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 analyzed arthritis-tissue dataset with ArchR[49] and Signac[50] for single-cell multimodal data, 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 log2-normalized peak count and gene count. Signac computes the Pearson correlation coefficient *r* (corSparse function in R) for each gene and for each peak within 500kb of the gene TSS. Signac then compares the observed correlation coefficient with an expected correlation coefficient for each peak given the GC 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 cut-off for correlation *r* or *P* values. We used FDR in the output from ArchR software, or computed FDR using *P* values in the output from Signac software by Benjamini & Hochberg correction. We defined significant peak-gene linkages as those with FDR < 0.10, and used varying correlation *r* to assess the precision and recall in the causal variant enrichment analysis (see later sections in **Method**).

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 most stringent FDR threshold for defining an enhancer-gene map in arthritis-tissue dataset (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 (i.e., 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. For the largest replication dataset of Public PBMC, we 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[65] score. We downloaded the phastCons score for multiple alignments of 99 vertebrate genomes from https://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/. We lifted them over to GRCh38 by LiftOver software. 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*-non-coding regions (< 500kb 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 ($\Delta\,\mathrm{phastCons\ score}$), and computed the mean differences across all the genes ($\mathrm{mean}\ \Delta\ \mathrm{phastCons\ score}$) as follows.

$$\mathrm{mean}\ \Delta\ \mathrm{phastCons\ score} = \frac{1}{n_{gene}} \sum_{gene} \overline{(phastCons}_{g,in\_annot} - \overline{phastCons}_{g,non-coding})$$

By bootstrapping the genes, we calculated the 95% CI of the mean Δ phastCons score. If this metric is positive, that indicates that the annotated regions are more conserved than non-coding regions.

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

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

By bootstrapping the genes, we again calculated the 95% CI of the mean Δ phastCons 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 Figure 5b**).

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 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 (https://gnomad.broadinstitute.org/downloads). 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 rightly skewed (**Supplementary Figure 3a**). 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[66,67].

Validation with CRISPR-Flow FISH results

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]. We used "Perturbation Target" as candidate enhancer elements. We defined 283 positive enhancer element-gene links when they are

"TRUE" for "Regulated" column (i.e., 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 two most powered datasets, arthritis-tissue and Public PBMC datasets. For each dataset, we used "bedtools 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.

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 (**Supplementary Figure 6e**), we used fine-mapped variants from each tissue separately. We intersected these putatively causal variants with our annotation (SCENT peaks, ArchR peaks or Signac peaks). We then retained any variants which the linking method (SCENT, ArchR and Signac) connected to the same gene as GTEx phenotype gene.

$$Enrichment_{gene\_i} = \frac{\#\ causal\_var\_in\_annot_{gene\_i} \Big/ \sum common\_var\_in\_annot_{gene\_i}}{\#\ causal\_var_{gene\_i} \Big/ \sum common\_var\_in\_cis_{gene\_i}}$$

$$Overall\_Enrichment = \frac{1}{n}\sum_{i=1}^{n} Enrichment_{gene\_i}$$

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 1000 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 (i.e., 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 socre {any, 0, 0.05, 0.1, 0.15} in Signac). We then used each set of peak-gene linkages to re-calculate causal variant enrichment $Overall\_Enrichment$ score (**Figure 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 re-defined 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.

GWAS fine-mapping results

We used GWAS fine-mapping results in FinnGen release 6[70] upon registration and publicly available GWAS fine-mapping results in UK Biobank[71] (https://www.finucanelab.org/data). 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 non-coding fine-mapped loci which did not include any non-synonymous 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 non-coding fine-mapped loci which did not include any non-synonymous 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 RA[26], T1D[79], and IBD[29], given our special interest in immune-mediated traits. We similarly selected non-coding fine-mapped loci which did not include any non-synonymous or splicing variants with PIP > 0.5, and lifted the results over to GRCh38 by using LiftOVer software. We defined 117 loci for RA, 77 loci for T1D and 86 loci for IBD.

Causal variant enrichment analysis using GWASs

We defined a causal enrichment for GWAS within SCENT enhancers and other annotations by using statistically fine-mapped variants from FinnGen[70] and UK Biobank[71] which we described in the previous section. We selected variants with PIP > 0.2 as putatively causal variants for primary analyses.

$$Enrichment_{trait\_i} = \frac{\#\ causal\_var\_in\_annot_{trait\_i} \Big/ \sum common\_var\_in\_annot_{trait\_i}}{\#\ causal\_var_{trait\_i} \Big/ \sum common\_var\_across\_loci_{trait\_i}}$$

$$Overall\_Enrichment = \frac{1}{n}\sum_{i=1}^{n} Enrichment_{trait\_i}$$

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.

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

We downloaded per-group EpiMap enhancer-gene links from https://personal.broadinstitute.org/cboix/epimap/links/pergroup/. We lifted the genomic coordinates to GRCh38 by using LiftOver software. When we assessed aggregated EpiMap enhancer-gene links across all the groups, we used “bedtools merge” function for each gene to create a union of all enhancer-gene links. 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 ftp://ftp.broadinstitute.org/outgoing/lincRNA/ABC/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz. We lifted the genomic coordinates to GRCh38 by using LiftOver software. When we assessed cell-type-specific ABC model with SCENT enhancers, we aggregated cell lines or cell types to be corresponding with our cell types (B cell, T cell, Myeloid cells, and

fibroblasts). 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. 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 (i.e., EpiMap and ABC model), we used sets of significant peak-gene linkages in each method with varying thresholds (i.e., FDR in SCENT {0.5, 0.3, 0.2, 0.1, 0.05, 0.02}, EpiMap 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 re-calculate causal variant enrichment for GWAS (**Figure 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 re-defined 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.

caQTL analysis using scATAC-seq samples with genotype

We used independent arthritis-tissue dataset with single-cell unimodal ATAC-seq data with genotype (*n* = 17, *manuscript in preparation*) to define chromatin accessibility QTLs (caQTLs). We used two methods, binomial test and RASQUAL. Briefly, we genotyped donors by using Illumina Multi-Ethnic Genotyping Array. We performed quality control of genotype by sample call rate > 0.99, variant call rate > 0.99, minor allele frequency > 0.01, and $P_{HWE} > 1.0\times10^{-6}$. We performed haplotype phasing with SHAPEIT2 software[99] and performed whole-genome

776 imputation by using minimac3 software[100] with a reference panel of 1000 Genomes Project
777 phase 3[101]. After imputation, we selected variants with imputation *Rsq* > 0.7 as post-imputation
778 QC. We next created a merged bam file of ATAC-seq for each donor and each cell type by
779 aggregating all the reads. Using the imputed genotype for each donor and aggregated bam files
780 for each donor and cell type, we applied WASP[102] to correct any bias in read mapping toward
781 reference alleles to accurately quantify allelic imbalance. We thus created a bias-corrected bam
782 files for each donor and cell type.

783 For binomial tests, we ran ASEReadCounter module in GATK software[103] using the
784 bias-corrected bam files as input to quantify allelic imbalance in heterozygous sites with read
785 count > 4 within ATAC peak counts. We first performed one-sided binomial tests in each donor,
786 and meta-analyzed the statistics across donors by Fisher's method if multiple donors shared
787 the same heterozygous site. For RASQUAL, we created a VCF file containing both genotype
788 dosage and allelic imbalance from ASEReadCounter. We quantified the read coverage for each
789 peak and for each donor by "bedtools coverage" function. We created a peak by donor matrix
790 with read coverage. We QCed samples with log(total mapped fragments) fewer than mean –
791 2SD across samples in each cell type. We QCed peaks so that at least two individuals have any
792 fragments for the peak. We then ran RASQUAL software with the inter-individual differences in
793 ATAC peak counts (in a peak by donor matrix) and intra-individual allelic imbalance (in VCF),
794 with accounting for chromatin accessibility PCs (the first *N* components whose explained
795 variances are greater than those from permutation result), 3 genotype PCs, sample site and sex
796 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 (i.e., the number of *cis*-SNPs in the region). To correct for genome-wide multiple testing, we ran the RASQUAL with random permutation, where 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.

ClinVar analysis

We downloaded the latest clinically reported variant list registered at ClinVar from https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/clinvar.vcf.gz. 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 non-coding and non-benign variants within each annotation x 1,000 divided by the total length (bp) of each annotation.

Somatic mutation analysis

We used a list of somatic mutation hotspot in Supplementary Table 2-20 of the original publication[94]. We lifted the genomic coordinates to GRCh38 by using LifOver software. We then intersected the non-coding somatic mutation hotspots with our cell-type-specific SCENT peaks. We compared the intersected elements’ target genes by SCENT with the “Annotate_Gene” column from the original publication.

818

819 *Downsampling experiments*

820 To evaluate the effect of cell numbers on the statistical power in detecting significant SCENT
821 enhancer-gene linkages, we performed downsampling experiments in fibroblast (the most
822 abundant cell type in arthritis-tissue dataset, n_{cell} = 9,905). We randomly samples cells (n_{cell} =
823 500, 1000, 2500, 5000, and 7500). We then applied SCENT to each of the subset groups of
824 cells and defined significant peak-gene links with FDR < 10%. We counted the number of
825 significant peak-gene links in each of the subset groups of cells, and annotated peaks based on
826 the distance to the TSS to the target gene.

827

828 **References**

829 1. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of
830 SNP-trait associations. *Nucleic Acids Res*. 2014;42(Database issue):D1001-6.
831 doi:10.1093/nar/gkt1229
832 2. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function,
833 and Translation. *Am J Hum Genet*. 2017;101(1):5-22. doi:10.1016/J.AJHG.2017.06.005
834 3. Buniello A, Macarthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS Catalog of published
835 genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic*
836 *Acids Res*. 2019;47(D1):D1005-D1012. doi:10.1093/NAR/GKY1120
837 4. Claussnitzer M, Cho JH, Collins R, et al. A brief history of human disease genetics. *Nature*
838 *2020 577:7789*. 2020;577(7789):179-189. doi:10.1038/s41586-019-1879-7
839 5. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human
840 genetics. *Nature Reviews Drug Discovery 2013 12:8*. 2013;12(8):581-594.
841 doi:10.1038/nrd4051
842 6. Shendure J, Findlay GM, Snyder MW. Genomic Medicine–Progress, Pitfalls, and Promise.
843 *Cell*. 2019;177(1):45-57. doi:10.1016/J.CELL.2019.02.003

7. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics 2018 19:8*. 2018;19(8):491-504. doi:10.1038/s41576-018-0016-z
8. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science (1979)*. 2012;337(6099):1190-1195. doi:10.1126/SCIENCE.1222794/SUPPL_FILE/MAURANO.SM.PDF
9. Edwards SL, Beesley J, French JD, Dunning M. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet*. 2013;93(5):779-797. doi:10.1016/J.AJHG.2013.10.012
10. Trynka G, Sandor C, Han B, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*. 2013;45(2):124-130. doi:10.1038/ng.2504
11. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature 2012 489:7414*. 2012;489(7414):109-113. doi:10.1038/nature11279
12. Smemo S, Tena JJ, Kim KH, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature 2014 507:7492*. 2014;507(7492):371-375. doi:10.1038/nature13138
13. Won H, de La Torre-Ubieta L, Stein JL, et al. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature 2016 538:7626*. 2016;538(7626):523-527. doi:10.1038/nature19847
14. Strober BJ, Elorbany R, Rhodes K, et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*. 2019;364(6447):1287-1290. doi:10.1126/SCIENCE.AAW0040
15. Cuomo ASE, Seaton DD, McCarthy DJ, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications 2020 11:1*. 2020;11(1):1-14. doi:10.1038/s41467-020-14457-z
16. Zhernakova D v., Deelen P, Vermaat M, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*. 2017;49(1):139-145. doi:10.1038/NG.3737
17. Nathan A, Asgari S, Ishigaki K, et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature 2022 606:7912*. 2022;606(7912):120-128. doi:10.1038/s41586-022-04713-1

18. Wakefield J. A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies. *Am J Hum Genet*. 2007;81(2):208. doi:10.1086/519024
19. Maller JB, McVean G, Byrnes J, et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet*. 2012;44(12):1294-1301. doi:10.1038/NG.2435
20. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. *Genetics*. 2014;198(2):497-508. doi:10.1534/GENETICS.114.167908
21. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016;32(10):1493-1501. doi:10.1093/BIOINFORMATICS/BTW018
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 Series B Stat Methodol*. 2020;82(5):1273-1300. doi:10.1111/RSSB.12388
23. Weissbrod O, Hormozdiari F, Benner C, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature Genetics 2020 52:12*. 2020;52(12):1355-1363. doi:10.1038/s41588-020-00735-5
24. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature 2019 570:7762*. 2019;570(7762):514-518. doi:10.1038/s41586-019-1310-4
25. Chen MH, Raffield LM, Mousas A, et al. Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell*. 2020;182(5):1198-1213.e14. doi:10.1016/J.CELL.2020.06.045
26. Ishigaki K, Sakaue S, Terao C, et al. Trans-ancestry genome-wide association study identifies novel genetic mechanisms in rheumatoid arthritis. *medRxiv*. 2021;12:2021.12.01.21267132. doi:10.1101/2021.12.01.21267132
27. Kichaev G, Pasaniuc B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am J Hum Genet*. 2015;97(2):260-271. doi:10.1016/J.AJHG.2015.06.007
28. Kanai M, Ulirsch JC, Karjalainen J, et al. Insights from complex trait fine-mapping across diverse populations. *medRxiv*. Published online September 5, 2021:2021.09.03.21262975. doi:10.1101/2021.09.03.21262975

29. Huang H, Fang M, Jostins L, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*. 2017;547(7662):173-178. doi:10.1038/NATURE22969
30. Farh KKH, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature 2014 518:7539*. 2014;518(7539):337-343. doi:10.1038/nature13835
31. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics 2018 50:11*. 2018;50(11):1505-1513. doi:10.1038/s41588-018-0241-6
32. Kichaev G, Yang WY, Lindstrom S, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*. 2014;10(10). doi:10.1371/JOURNAL.PGEN.1004722
33. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature 2015 518:7539*. 2015;518(7539):317-330. doi:10.1038/nature14248
34. Chen L, Ge B, Casale FP, et al. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*. 2016;167(5):1398-1414.e24. doi:10.1016/j.cell.2016.10.026
35. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/NATURE11247
36. Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature 2011 473:7345*. 2011;473(7345):43-49. doi:10.1038/nature09906
37. Boix CA, James BT, Park YP, Meuleman W, Kellis M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature 2021 590:7845*. 2021;590(7845):300-307. doi:10.1038/s41586-020-03145-z
38. Fulco CP, Nasser J, Jones TR, et al. Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat Genet*. 2019;51(12):1664. doi:10.1038/S41588-019-0538-0
39. Nasser J, Bergman DT, Fulco CP, et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature 2021 593:7858*. 2021;593(7858):238-243. doi:10.1038/s41586-021-03446-x

938 40. Gazal S, Weissbrod O, Hormozdiari F, et al. Combining SNP-to-gene linking strategies to
939 identify disease genes and assess disease omnigenicity. *Nat Genet*. 2022;54(6):827-836.
940 doi:10.1038/S41588-022-01087-Y
941 41. Pickar-Oliver A, Gersbach CA. The next generation of CRISPR–Cas technologies and
942 applications. *Nature Reviews Molecular Cell Biology 2019 20:8*. 2019;20(8):490-507.
943 doi:10.1038/s41580-019-0131-5
944 42. Anzalone A v., Koblan LW, Liu DR. Genome editing with CRISPR–Cas nucleases, base
945 editors, transposases and prime editors. *Nature Biotechnology 2020 38:7*.
946 2020;38(7):824-844. doi:10.1038/s41587-020-0561-9
947 43. Baglaenko Y, Macfarlane D, Marson A, Nigrovic PA, Raychaudhuri S. Genome editing to
948 define the function of risk loci and variants in rheumatic disease. *Nature Reviews*
949 *Rheumatology 2021 17:8*. 2021;17(8):462-474. doi:10.1038/s41584-021-00637-8
950 44. Cao J, Cusanovich DA, Ramani V, et al. Joint profiling of chromatin accessibility and gene
951 expression in thousands of single cells. *Science (1979)*. 2018;361(6409):1380-1385.
952 doi:10.1126/SCIENCE.AAU0730/SUPPL_FILE/AAU0730_TABLESS1_S13.XLSX
953 45. Chen S, Lake BB, Zhang K. High-throughput sequencing of the transcriptome and
954 chromatin accessibility in the same cell. *Nature Biotechnology 2019 37:12*.
955 2019;37(12):1452-1457. doi:10.1038/s41587-019-0290-0
956 46. Ma S, Zhang B, LaFave LM, et al. Chromatin Potential Identified by Shared Single-Cell
957 Profiling of RNA and Chromatin. *Cell*. 2020;183(4):1103-1116.e20.
958 doi:10.1016/J.CELL.2020.09.056
959 47. Allaway KC, Gabitto MI, Wapinski O, et al. Genetic and epigenetic coordination of cortical
960 interneuron development. *Nature 2021 597:7878*. 2021;597(7878):693-697.
961 doi:10.1038/s41586-021-03933-1
962 48. Trevino AE, Müller F, Andersen J, et al. Chromatin and gene-regulatory dynamics of the
963 developing human cerebral cortex at single-cell resolution. *Cell*.
964 2021;184(19):5053-5069.e23. doi:10.1016/J.CELL.2021.07.039
965 49. Granja JM, Corces MR, Pierce SE, et al. ArchR is a scalable software package for
966 integrative single-cell chromatin accessibility analysis. *Nature Genetics 2021 53:3*.
967 2021;53(3):403-411. doi:10.1038/s41588-021-00790-6

50. Stuart T, Srivastava A, Madad S, Lareau CA, Satija R. Single-cell chromatin state analysis with Signac. *Nature Methods 2021 18:11*. 2021;18(11):1333-1341. doi:10.1038/s41592-021-01282-5
51. Pliner HA, Packer JS, McFaline-Figueroa JL, et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell*. 2018;71(5):858-871.e8. doi:10.1016/J.MOLCEL.2018.06.044
52. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. *An Introduction to the Bootstrap*. Published online May 14, 1994. doi:10.1201/9780429246593
53. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biology 2020 21:1*. 2020;21(1):1-35. doi:10.1186/S13059-020-1926-6
54. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics 2021 53:6*. 2021;53(6):770-777. doi:10.1038/s41588-021-00873-4
55. Chen H, Lareau C, Andreani T, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol*. 2019;20(1):1-25. doi:10.1186/S13059-019-1854-5/FIGURES/7
56. Granja JM, Klemm S, McGinnis LM, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology 2019 37:12*. 2019;37(12):1458-1465. doi:10.1038/s41587-019-0332-7
57. Luecken MD, Burkhardt DB, Cannoodt R, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. 2021;1. Accessed September 11, 2022. https://openproblems.bio/neurips.
58. Mimitou EP, Lareau CA, Chen KY, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol*. 2021;39(10):1246-1258. doi:10.1038/S41587-021-00927-2
59. Chen AF, Parks B, Kathiria AS, Ober-Reynolds B, Goronzy JJ, Greenleaf WJ. NEAT-seq: simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single cells. *Nature Methods 2022 19:5*. 2022;19(5):547-553. doi:10.1038/s41592-022-01461-y

60. Meijer M, Agirre E, Kabbe M, et al. Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. *Neuron*. 2022;110(7):1193-1210.e13. doi:10.1016/J.NEURON.2021.12.034
61. Zhang Z, Zamojski M, Smith GR, et al. Single nucleus transcriptome and chromatin accessibility of postmortem human pituitaries reveal diverse stem cell regulatory mechanisms. *Cell Rep*. 2022;38(10). doi:10.1016/J.CELREP.2022.110467
62. Abascal F, Acosta R, Addleman NJ, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature 2020 583:7818*. 2020;583(7818):699-710. doi:10.1038/s41586-020-2493-4
63. Westra HJ, Franke L. From genome to function by studying eQTLs. *Biochim Biophys Acta*. 2014;1842(10):1896-1902. doi:10.1016/J.BBADIS.2014.04.024
64. Hujoel MLA, Gazal S, Hormozdiari F, van de Geijn B, Price AL. Disease Heritability Enrichment of Regulatory Elements Is Concentrated in Elements with Ancient Sequence Age and Conserved Function across Species. *Am J Hum Genet*. 2019;104(4):611-624. doi:10.1016/j.ajhg.2019.02.008
65. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15(8):1034-1050. doi:10.1101/GR.3715005
66. Lek M, Karczewski KJ, Minikel E v., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057
67. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7
68. Wang X, Goldstein DB. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. *The American Journal of Human Genetics*. 2020;106(2):215-233. doi:10.1016/J.AJHG.2020.01.012
69. Aguet F, Barbeira AN, Bonazzola R, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318. doi:10.1126/SCIENCE.AAZ1776
70. Kurki MI, Karjalainen J, Palta P, et al. FinnGen: Unique genetic insights from combining isolated population and national health register data. *medRxiv*. Published online March 6, 2022:2022.03.03.22271360. doi:10.1101/2022.03.03.22271360

71. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z
72. Dey KK, Gazal S, van de Geijn B, et al. SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell Genomics*. 2022;2:100145. doi:10.1016/j.xgen.2022.100145
73. Freund MK, Burch KS, Shi H, et al. Phenotype-Specific Enrichment of Mendelian Disorder Genes near GWAS Regions across 62 Complex Traits. *The American Journal of Human Genetics*. 2018;103(4):535-552. doi:10.1016/J.AJHG.2018.08.017
74. Gate RE, Cheng CS, Aiden AP, et al. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nature Genetics 2018 50:8*. 2018;50(8):1140-1150. doi:10.1038/s41588-018-0156-2
75. Khetan S, Kursawe R, Youn A, et al. Type 2 Diabetes-Associated Genetic Variants Regulate Chromatin Accessibility in Human Islets. *Diabetes*. 2018;67(11):2466-2477. doi:10.2337/DB18-0393
76. Alasoo K, Rodrigues J, Mukhopadhyay S, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat Genet*. 2018;50(3):424. doi:10.1038/S41588-018-0046-7
77. Currin KW, Erdos MR, Narisu N, et al. Genetic effects on liver chromatin accessibility identify disease regulatory variants. *Am J Hum Genet*. 2021;108(7):1169-1189. doi:10.1016/J.AJHG.2021.05.001
78. Kumasaka N, Knights AJ, Gaffney DJ. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics 2015 48:2*. 2015;48(2):206-213. doi:10.1038/ng.3467
79. Chiou J, Geusz RJ, Okino ML, et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature 2021 594:7863*. 2021;594(7863):398-402. doi:10.1038/s41586-021-03552-w
80. Mouri K, Guo MH, de Boer CG, et al. Prioritization of autoimmune disease-associated genetic variants that perturb regulatory element activity in T cells. *Nature Genetics 2022 54:5*. 2022;54(5):603-612. doi:10.1038/s41588-022-01056-5
81. Javierre BM, Sewitz S, Cairns J, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*. 2016;167(5):1369-1384.e19. doi:10.1016/J.CELL.2016.09.037

82. Radtke F, Fasnacht N, MacDonald HR. Notch signaling in the immune system. *Immunity*. 2010;32(1):14-27. doi:10.1016/J.IMMUNI.2010.01.004
83. Wei K, Korsunsky I, Marshall JL, et al. Notch signalling drives synovial fibroblast identity and arthritis pathology. *Nature*. 2020;582(7811):259-264. doi:10.1038/S41586-020-2222-Z
84. Delacher M, Schmidl C, Herzig Y, et al. Rbpj expression in regulatory T cells is critical for restraining TH2 responses. *Nature Communications 2019 10:1*. 2019;10(1):1-20. doi:10.1038/s41467-019-09276-w
85. Blake JA, Baldarelli R, Kadin JA, Richardson JE, Smith CL, Bult CJ. Mouse Genome Database (MGD): Knowledgebase for mouse–human comparative biology. *Nucleic Acids Res*. 2021;49(D1):D981. doi:10.1093/NAR/GKAA1083
86. Uhlén M, Fagerberg L, Hallström BM, et al. Tissue-based map of the human proteome. *Science (1979)*. 2015;347(6220). doi:10.1126/SCIENCE.1260419/SUPPL_FILE/1260419_UHLEN.SM.PDF
87. Hillier SG. Gonadotropic control of ovarian follicular growth and development. *Mol Cell Endocrinol*. 2001;179(1-2):39-46. doi:10.1016/S0303-7207(01)00469-5
88. Rubinstein WS, Maglott DR, Lee JM, 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*. 2013;41(D1):D925-D935. doi:10.1093/NAR/GKS1173
89. Retterer K, Juusola J, Cho MT, et al. Clinical application of whole-exome sequencing across clinical indications. *Genet Med*. 2016;18(7):696-704. doi:10.1038/GIM.2015.148
90. Adams DR, Eng CM. Next-Generation Sequencing to Diagnose Suspected Genetic Disorders. *N Engl J Med*. 2018;379(14):1353-1362. doi:10.1056/NEJMRA1711801
91. Srivastava S, Love-Nichols JA, Dies KA, et al. Meta-analysis and multidisciplinary consensus statement: exome sequencing is a first-tier clinical diagnostic test for individuals with neurodevelopmental disorders. *Genet Med*. 2019;21(11):2413-2421. doi:10.1038/S41436-019-0554-6
92. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. doi:10.1093/NAR/GKX1153

93. Glocker EO, Kotlarz D, Boztug K, et al. Inflammatory Bowel Disease and Mutations Affecting the Interleukin-10 Receptor. *New England Journal of Medicine*. 2009;361(21):2033-2045. doi:10.1056/NEJMOA0907206/SUPPL_FILE/NEJM_GLOCKER_2033SA1.PDF
94. Dietlein F, Wang AB, Fagre C, et al. Genome-wide analysis of somatic noncoding mutation patterns in cancer. *Science (1979)*. 2022;376(6589). doi:10.1126/SCIENCE.ABG5601/SUPPL_FILE/SCIENCE.ABG5601_MDAR_REPRODUCIBILITY_CHECKLIST.PDF
95. Connally N, Nazeen S, Lee D, et al. The missing link between genetic association and regulatory function. *medRxiv*. Published online October 13, 2022:2021.06.08.21258515. doi:10.1101/2021.06.08.21258515
96. Donlin LT, Rao DA, Wei K, et al. Methods for high-dimensonal analysis of cells dissociated from cyropreserved synovial tissue. *Arthritis Res Ther*. 2018;20(1):1-15. doi:10.1186/S13075-018-1631-Y/FIGURES/6
97. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/J.CELL.2019.05.031
98. Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods 2019 16:12*. 2019;16(12):1289-1296. doi:10.1038/s41592-019-0619-0
99. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2012;9(2):179-181. doi:10.1038/nmeth.1785
100. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nature Genetics 2016 48:10*. 2016;48(10):1284-1287. doi:10.1038/ng.3656
101. Gibbs RA, Boerwinkle E, Doddapaneni H, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
102. van der Auwera G, O'Connor B, Safari an OMCompany. *Genomics in the Cloud*.; 2020. Accessed December 25, 2021. https://www.oreilly.com/library/view/genomics-in-the/9781491975183/

Acknowledgments

We would like to sincerely thank participants of this study who provided tissue samples. We thank Anika Gupta, Joyce Kang and Kaitlyn 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, UC2AR081023). S.S. was in part supported by the Uehara Memorial Foundation and The Osamu Hayaishi Memorial Scholarship. K.W. is supported by a Burroughs Wellcome Fund Career Awards for Medical Scientists, a Doris Duke Charitable Foundation Clinical Scientist Development Award, and a Rheumatology Research Foundation Innovative Research Award. We would like to thank the Brigham and Women's Hospital Center for Cellular Profiling Single Cell Mulitomics Core for experimental design and protocol optimization.

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.

Competing Financial Interests

We declare no conflict of interest for this study. 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.