



# Single-cell polygenic risk scores dissect cellular and molecular heterogeneity of complex human diseases

Received: 15 May 2024

Accepted: 28 May 2025

Published online: 25 July 2025

Check for updates

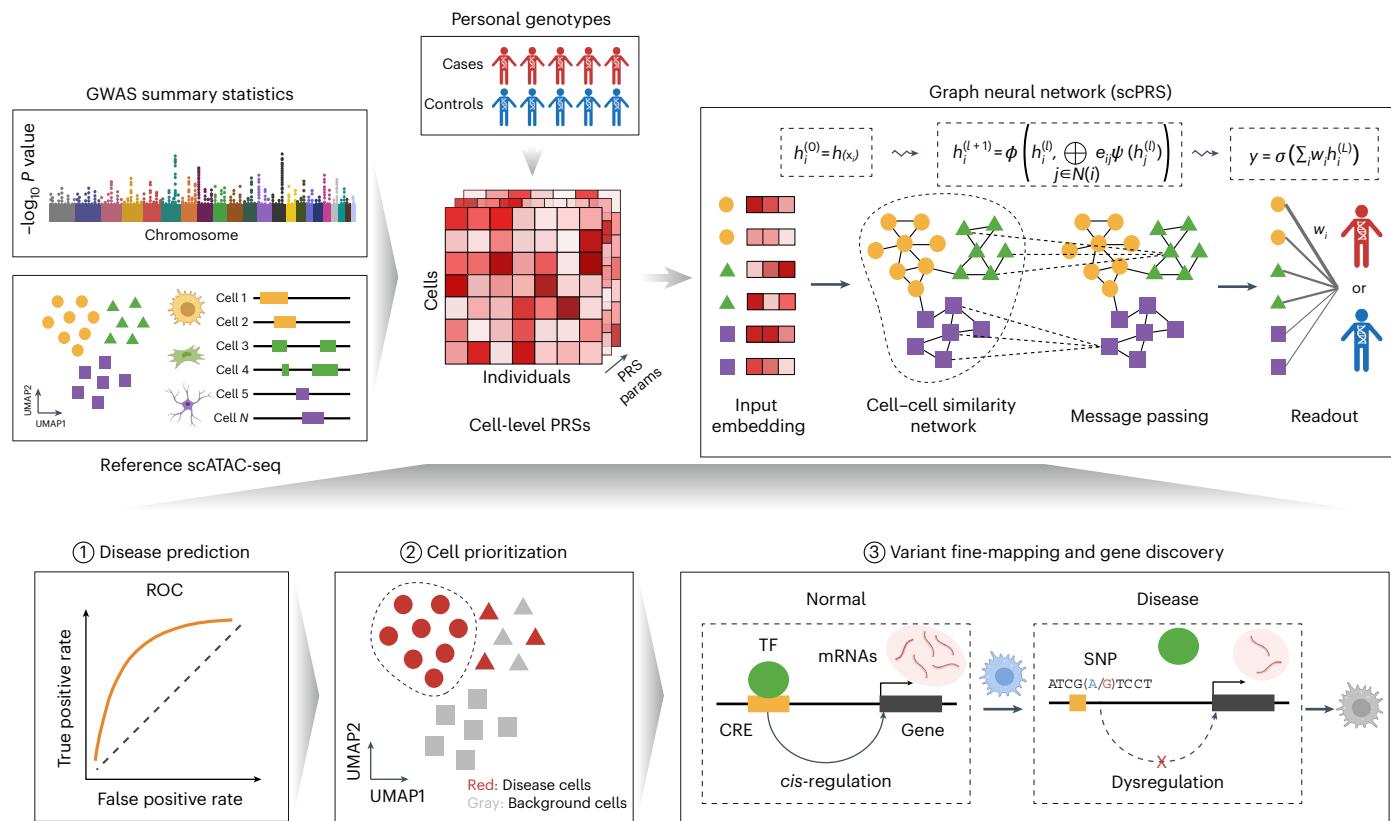
Sai Zhang <sup>1,2,3,4,16</sup>, Hantao Shu <sup>5,16</sup>, Jingtian Zhou <sup>6,7,8,16</sup>, Jasper Rubin-Sigler <sup>9</sup>, Xiaoyu Yang <sup>10</sup>, Yuxi Liu <sup>10</sup>, Johnathan Cooper-Knock <sup>11</sup>, Emma Monte <sup>3</sup>, Chenchen Zhu <sup>3</sup>, Sharon Tu <sup>9</sup>, Han Li <sup>5</sup>, Mingming Tong <sup>3</sup>, Joseph R. Ecker <sup>7,12</sup>, Justin K. Ichida <sup>9</sup>, Yin Shen <sup>10,13</sup>, Jianyang Zeng <sup>14</sup> , Philip S. Tsao <sup>4,15</sup> & Michael P. Snyder <sup>3</sup>

Polygenic risk scores (PRSs) predict an individual's genetic risk for complex diseases, yet their utility in elucidating disease biology remains limited. We introduce scPRS, a graph neural network-based framework that computes single-cell-resolved PRSs by integrating reference single-cell chromatin accessibility profiles. scPRS outperforms traditional PRS approaches in genetic risk prediction, as demonstrated across multiple diseases including type 2 diabetes, hypertrophic cardiomyopathy, Alzheimer disease and severe COVID-19. Beyond risk prediction, scPRS prioritizes disease-critical cells and, when combined with a layered multiomic analysis, links risk variants to gene regulation in a cell-type-specific manner. Applied to these diseases, scPRS fine-maps causal cell types and cell-type-specific variants and genes, demonstrating its ability to bridge genetic risk with cell-specific biology. scPRS provides a unified framework for genetic risk prediction and mechanistic dissection of complex diseases, laying a methodological foundation for single-cell genetics.

Polygenic risk score<sup>1</sup> (PRS), also known as polygenic score<sup>2</sup>, is a widely used approach to predict quantitative traits and disease risk on the basis of an individual's genetic makeup. The method is built upon genetic variants, including single-nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) that are common (minor allele frequency (MAF) > 5%) in the population. PRS is a critical component of precision genomic medicine and has promise in versatile utilities<sup>3</sup>, such as health management, disease screening and therapeutic intervention. Traditionally, PRS computation involves a linear model that sums the genotypes of selected variants, with each variant weighted according to its effect size as estimated by a genome-wide association study<sup>4</sup> (GWAS). The clumping and thresholding (C+T) method serves as the basis of constructing PRSs; however, other advanced approaches<sup>5–8</sup> have also been developed to enhance prediction by considering nuanced genetic architectures. Complex diseases exhibit notable cellular heterogeneity,

involving multiple tissues or cell types in their pathogenesis<sup>9</sup>. Risk variants, particularly noncoding ones, can influence disease susceptibility and phenotypic variability through diverse cellular and molecular processes<sup>10–12</sup>. However, these multiple layers of complexity have been oversimplified in conventional modeling, substantially limiting the predictive power and interpretability of PRS<sup>13</sup>.

In recent years, single-cell sequencing has emerged as a potent tool to dissect cellular and molecular heterogeneity across different tissues and conditions<sup>14</sup>, offering unprecedented opportunities to explore genome function at high resolution. Single-cell profiling data from healthy tissues provide high-resolution annotations of the baseline genome function in which genetic variants are involved. Incorporating functional annotations into PRS calculation will remove confounders such as linked nonfunctional variants, better characterize a disease's genetic architecture and, therefore, improve the predictive accuracy



**Fig. 1 | Overview of scPRS and its applications.** For a given disease, scPRS first leverages GWAS summary statistics obtained from the discovery cohort and the reference scATAC-seq or snATAC-seq dataset to calculate single-cell-level PRSs with different parameters for individuals in the target cohort. Next, scPRS embeds and propagates cell-level PRSs over the cell-cell similarity network using a GNN. The final readout combines smoothed PRSs from all cells to predict

the disease risk. scPRS is trained to minimize the loss between predicted and true disease labels. The trained model can be used to (1) predict disease risk for unseen individuals; (2) prioritize disease-relevant cells and cell types; and (3) fine-map disease risk variants, genes and disrupted genetic regulation in specific cell types. UMAP, uniform manifold approximation and projection. The schematic was created using [BioRender.com](https://biorender.com).

and generalization. This has been demonstrated elsewhere<sup>15</sup>, including our latest study<sup>16</sup>. Moreover, the interpretability of PRS can be considerably enhanced by incorporating functional information, adding biological discovery functionality to predictive methods.

To bridge this gap, we propose a strategy that unifies genetics and single-cell genomics, named single-cell genetics<sup>17</sup>, to study disease genetics at single-cell resolution. In particular, we introduce scPRS, a graph neural network<sup>18</sup> (GNN)-based framework that enables individualized genetic risk prediction at the single-cell level. scPRS leverages the GNN to construct genetic risk score by drawing insights from reference single-cell chromatin accessibility measured by single-cell or single-nucleus sequencing assay for transposase-accessible chromatin<sup>19</sup> (scATAC-seq or snATAC-seq). scATAC-seq or snATAC-seq maps single-cell-resolved candidate *cis*-regulatory elements<sup>20</sup> (cCREs), which are specific DNA regions that potentially regulate the transcription of nearby genes. Beyond enhanced disease prediction, scPRS is empowered with fine-grained model interpretability, which allows for systematic discovery of cell types and cell-type-specific gene-regulatory programs underpinning diseases.

We performed extensive simulation experiments to demonstrate the effectiveness and robustness of scPRS in identifying phenotype-relevant cells. We applied scPRS to four diseases—type 2 diabetes (T2D), hypertrophic cardiomyopathy (HCM), Alzheimer disease (AD) and severe COVID-19—and showcased its superior predictive performance compared to traditional PRS methods. Through model interpretation, scPRS identified known disease-critical cell types as well as previously uncharacterized cell populations. scPRS-powered functional analysis further fine-mapped candidate causal variants, cCREs

and target genes within specific cell types, revealing a cell-type-specific landscape of genetic regulation. Using drug perturbation data, we validated our scPRS-nominated HCM genes, showing that the suppression of these genes in diseased cardiomyocytes (CDMs) was rescued by mavacamten (a US Food and Drug Administration (FDA)-approved HCM drug) treatment. Supported by experiments, we identified a new role of the AD risk variant [rs7922621](https://www.ncbi.nlm.nih.gov/snp/rs7922621) in downregulating *ANXA11* and *TSPAN14*, specifically in microglia. We also demonstrated the negative effect of suppressing these genes on microglial phagocytosis. Taken together, scPRS offers a unified approach that encompasses GNN modeling and GNN-inspired downstream analysis for simultaneous disease prediction and biological discovery, establishing the methodological foundation for single-cell genetics.

## Results

### Overview of scPRS

The design principle of scPRS is to leverage single-cell epigenome profiling to rationalize the calculation of PRS. The approach begins with deconvoluting traditional PRS within individual cells on the basis of their chromatin accessibility profiled by scATAC-seq, followed by the integration of decomposed single-cell-level PRSs into a final score capitalizing on cell-cell similarities (Fig. 1 and Methods). In particular, using GWAS summary statistics derived from a disease cohort (referred to as the discovery cohort) and an scATAC-seq dataset of healthy tissue pertinent to the disease (referred to as the reference scATAC-seq dataset), we compute a conditioned PRS for each individual within our target cohort (independent with the discovery cohort) and for each reference cell, in which we mask out genetic variants located outside

open chromatin regions captured in that specific cell. Recognizing the sparsity of scATAC-seq data, scPRS further refines per-cell PRS features using a GNN<sup>21</sup>. This GNN operation serves the dual purpose of denoising raw PRS features while capturing nonlinear relationships between genetic signals and the cellular epigenome. In the final step, scPRS aggregates smoothed single-cell-level PRSs and yields a final disease risk score. The interpretability of scPRS is achieved by the learned model weights accompanied with single cells that indicate the contribution of different cells to the disease risk.

The functionalities of scPRS are exemplified by three downstream tasks. First, scPRS predicts disease risk for unseen individuals solely on the basis of their genotypes (Fig. 1, step 1). Second, scPRS prioritizes single cells that are relevant to the disease, overcoming the resolution constraint of predefined cell clusters (Fig. 1, step 2). Third, integrated with a multiomic approach, scPRS fine-maps causal variants, genes and genetic regulation within prioritized cell types (Fig. 1, step 3).

### Evaluation of scPRS using simulations

We first performed simulation experiments to evaluate the capacity of scPRS in identifying phenotype-relevant cells. Assuming that the trait 'monocyte count' is fully determined by genetic variants located within monocyte-specific open chromatin regions<sup>22</sup>, we simulated monocyte counts for individuals of a genotyped cohort<sup>23</sup> ( $n = 401$ ). We then asked whether we could use scPRS to recapitulate monocytes as the causal cell type. Specifically, we used a reference scATAC-seq dataset<sup>24</sup> (Extended Data Fig. 1a) of human peripheral blood mononuclear cells (PBMCs) to identify monocyte-specific peaks (Methods). On the basis of a monocyte count GWAS<sup>22</sup> defining variant effect sizes, we simulated the monocyte count for each individual by calculating the C+T PRS using only variants located within monocyte-specific peaks (Methods). Next, we trained an scPRS model to predict simulated monocyte counts from cell-level PRSs computed on all PBMCs. We observed that scPRS predictions were significantly correlated with simulated monocyte counts ( $r = 0.77$ ,  $P < 2.2 \times 10^{-16}$ , Pearson correlation; Extended Data Fig. 1b). The cells prioritized by scPRS (Methods) were significantly enriched within monocytes ( $Z = 39.58$ ,  $P < 1 \times 10^{-50}$ , two-sided Fisher's exact test; Extended Data Fig. 1c), demonstrating that scPRS captured causal cells.

Human phenotypes such as complex diseases can be influenced by various nongenetic factors, including environmental and lifestyle factors<sup>25</sup>. Additionally, the measurement of phenotypes often carries inherent noise. Therefore, it is important to assess the robustness of scPRS by introducing noise and randomness into the simulation (Methods). As expected, we observed a progressive reduction in predictive performance as we introduced larger amounts of noise (Extended Data Fig. 1d). Notably, scPRS sustained its ability in uncovering monocytes even in the presence of considerable noise terms (Extended Data Fig. 1e, f and Supplementary Fig. 1a). For example, when we introduced a noise term with the same amount of variance ( $\sigma = 1$ ) as that of the simulated phenotype, scPRS still accurately identified monocytes (area under the curve (AUC) = 0.812; Extended Data Fig. 1e); the enrichment of monocytes persisted even when three times the amount of variance was added ( $\sigma = 3$ ;  $Z = 2.68$ ,  $P < 1 \times 10^{-50}$ , two-sided Fisher's exact test; Extended Data Fig. 1f).

We further introduced peak noise into simulation by replacing a proportion of randomly selected monocyte-specific peaks with non-monocyte-specific peaks. Using these mixed peaks, we generated noisy monocyte counts for individuals. We then assessed whether scPRS could still identify monocytes from the noisy data. We found that scPRS was able to identify monocytes with peak noise levels up to 90% (Supplementary Fig. 1b, c). We also tested different model hyper-parameter settings and observed no significant variation in predictive performance (Supplementary Fig. 1d). All these results demonstrate the robustness of scPRS against different sources of noise, randomness and model settings.

Lastly, we conducted a negative control experiment by excluding monocyte-related cells, including monocytes and cells containing more than 40 monocyte-specific peaks (-1% of all peaks used for simulating monocyte counts) from the PBMC dataset. Unsurprisingly, the predictive performance of scPRS was significantly reduced ( $r = 0.488$  (mean)  $\pm 0.085$  (s.d.); Supplementary Fig. 1e) compared to scPRS trained on the full dataset. Moreover, scPRS exhibited increased nonspecificity in prioritizing monocyte-count-relevant cells (Supplementary Fig. 1f), showing a similar saturation pattern in large-noise scenarios (Extended Data Fig. 1f).

### scPRS accurately predicts diseases

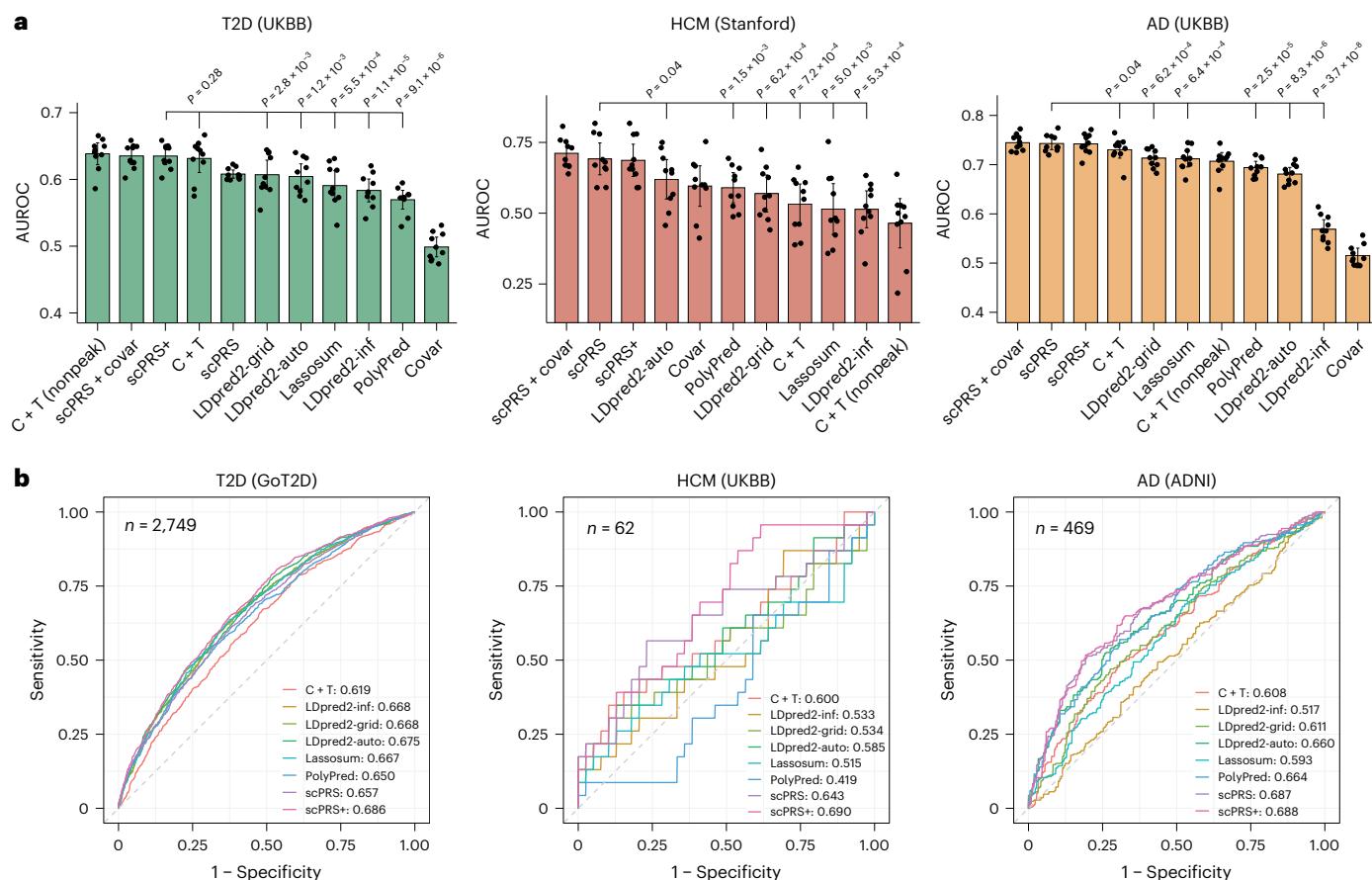
We applied scPRS to multiple diseases, including T2D, HCM, AD and severe COVID-19. We used UK Biobank<sup>26</sup> (UKBB) data to construct target cohorts for T2D and AD and our in-house whole-genome sequencing (WGS) data<sup>27</sup> for HCM (Methods). The severe COVID-19 target cohort was constructed on the basis of the Veterans Affairs (VA) Million Veteran Program<sup>28</sup> (MVP) WGS dataset (Methods). The discovery GWAS dataset<sup>29–32</sup> was carefully chosen to ensure nonoverlap with the target cohort for each disease. Multiple reference scATAC-seq datasets of disease-relevant tissues were used, including the pancreas<sup>33</sup> for T2D, left ventricle<sup>34</sup> for HCM, frontal cortex<sup>35</sup> for AD and lung<sup>34</sup> for severe COVID-19 (Methods).

For benchmarking, we used six well-established PRS methods: C+T (implemented by PLINK<sup>36</sup>), LDpred2 (including LDpred2-inf, LDpred2-grid, and LDpred2-auto)<sup>5</sup>, Lassosum<sup>7</sup> and PolyPred<sup>37</sup> (Methods). Among these baseline methods, PolyPred uses functional annotations to compute prior causal probabilities of variants<sup>38</sup>, for which we used scATAC-seq peaks as the annotation to ensure a fair comparison. To examine the predictability of nonpeak and nongenetic factors, we also built a C+T PRS model on the basis of variants situated beyond open chromatin regions and a logistic regression (LR) model using individual's age, sex and the first ten principal components (PCs) as input features (Methods).

Remarkably, scPRS-based methods consistently outperformed all baseline PRS approaches in all diseases (Fig. 2a and Supplementary Fig. 2a, b). In particular, for HCM, AD and severe COVID-19, scPRS achieved superior predictive performance evaluated by both the area under the receiver operating characteristic curve (AUROC; HCM,  $0.692 \pm 0.079$ ; AD,  $0.743 \pm 0.017$ ; severe COVID-19,  $0.591 \pm 0.029$ ) and the area under the precision–recall curve (AUPRC; HCM,  $0.781 \pm 0.062$ ; AD,  $0.751 \pm 0.035$ ; severe COVID-19,  $0.281 \pm 0.034$ ) compared to all baseline PRS methods (adjusted  $P < 0.1$ , Benjamini–Hochberg (BH) correction; Fig. 2a and Supplementary Fig. 2a, b), except for C+T and LDpred2-auto, which yielded comparable AUPRC values in some cases.

For T2D, scPRS presented performance comparable to other methods (AUROC,  $0.608 \pm 0.009$ ; AUPRC,  $0.598 \pm 0.032$ ; Fig. 2a and Supplementary Fig. 2b). Integrating nonpeak C+T PRSs into the scPRS model (referred to as scPRS+; Methods) further boosted its performance (AUROC,  $0.635 \pm 0.018$ ; AUPRC,  $0.633 \pm 0.036$ ), outperforming all baseline methods (adjusted  $P < 0.1$ , BH correction; Fig. 2a and Supplementary Fig. 2b), except for C+T where the AUROC remained comparable. These results suggest that the variants located outside pancreas cCREs, such as protein-coding<sup>39</sup> and splicing<sup>40</sup> variants, or variants within cCREs specific to other tissues<sup>41</sup> may also contribute to T2D susceptibility. This is also supported by the observation that a predictor built solely on nonpeak PRSs (referred to as nonpeak C+T) performed best among all methods (AUROC,  $0.638 \pm 0.023$ ; AUPRC,  $0.633 \pm 0.039$ ; Fig. 2a and Supplementary Fig. 2b).

We also constructed peak PRSs across different cell types annotated in the scATAC-seq datasets (Methods). scPRS outperformed all single-cell-type and multi-cell-type PRSs for all diseases (Supplementary Fig. 3), underscoring the advantage of single-cell-resolved modeling in disease prediction.



**Fig. 2 | Predictive performance comparison between scPRS and baseline methods.** **a**, Bar plots of AUROC values of different models. The training and testing procedure was conducted for ten repeats with different random seeds. Training, validation and test dataset splits were kept identical across different methods to ensure a fair comparison. scPRS+, scPRS model integrating nonpeak PRSs; scPRS+covar, scPRS model integrating nonpeak PRSs and covariates (that

is, age, sex and first ten PCs); C+T (nonpeak), LR model of nonpeak C+T PRSs; Covar, LR model of covariates. Performance comparison was conducted using a one-sided paired *t*-test. The mean and 95% confidence interval (CI) are annotated using the bar plot and error bar, respectively. **b**, ROC curves of different models evaluated on independent target cohorts. The performance of a random predictor is shown by the dashed gray line.

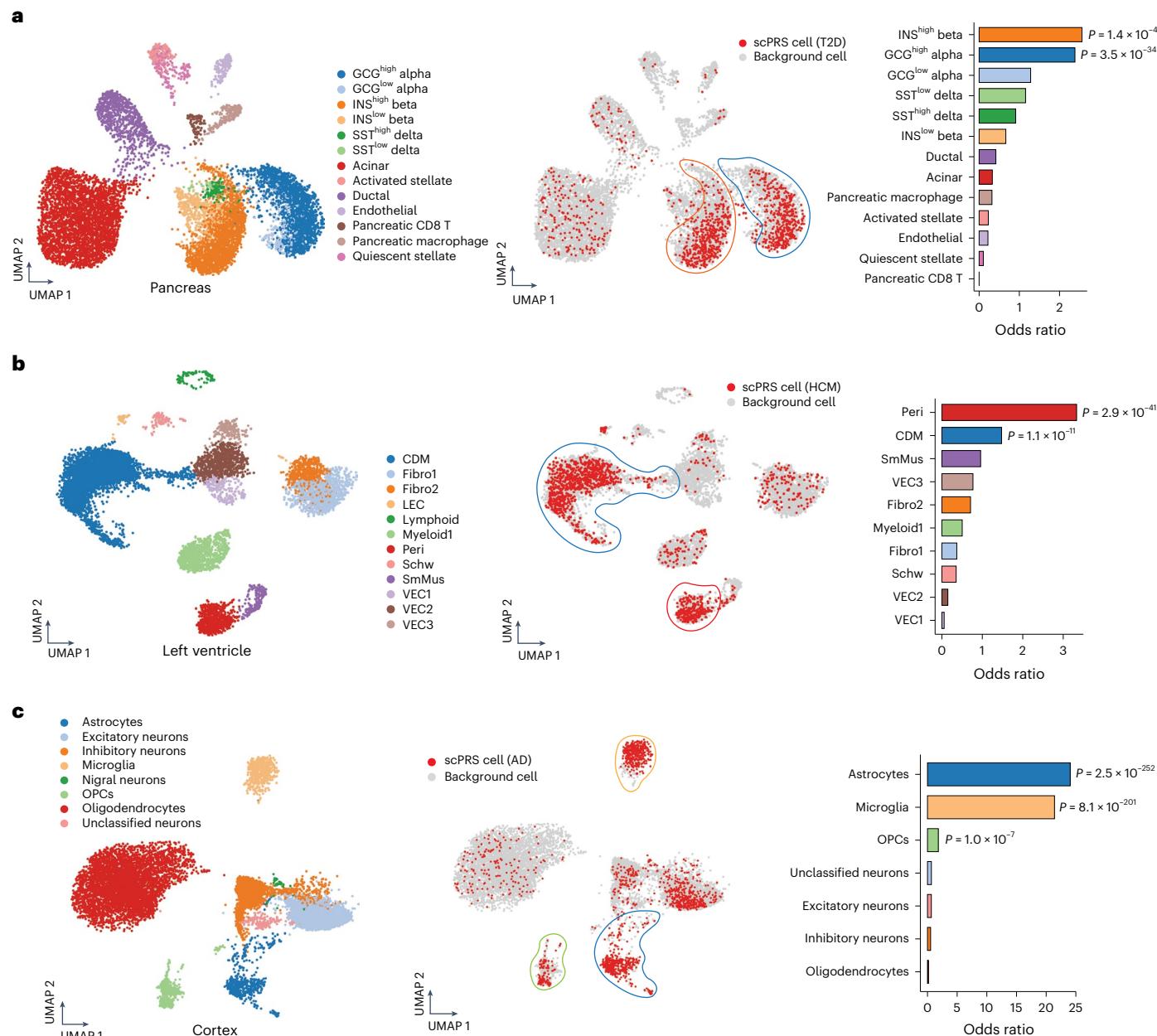
The covariate models exhibited limited predictive power for T2D and AD (Fig. 2a and Supplementary Fig. 2b) because of the fact that we matched age, sex and population between cases and controls in constructing the target cohorts. Not surprisingly, the predictive performance reached a peak for all diseases after integrating all other factors, including nonpeak PRSs and covariates, into the scPRS model (referred to as scPRS+covar; Fig. 2a and Supplementary Fig. 2a,b).

We tested the use of alternative scATAC-seq datasets in scPRS, including those from a different study<sup>42</sup> (for AD), a different donor (for HCM) and a different sampling (for T2D). We found that scPRS yielded comparable predictive performance (Supplementary Fig. 4a,b), demonstrating its robustness against distinct choices of reference single-cell datasets. To examine the impact of cell numbers, we compared the predictive performance of T2D scPRS models using different numbers of cells randomly sampled from the pancreas scATAC-seq dataset. We observed that scPRS exhibited moderately stable predictive performance across a broad range of cell numbers (Supplementary Fig. 4c), with an increase in performance as more cells were sampled. We also assessed the impact of input PRS choices. In particular, we randomly removed input PRSs, in which a certain proportion of randomly selected PRS features were set to zero for all samples in each training–testing procedure. We then evaluated the predictive performance of scPRS across different dropout rates. scPRS yielded stable predictive performance with only a slight decrease as dropout rates increased up to 70% (Supplementary

Fig. 4d), whereas performance was substantially reduced at higher dropout rates.

As a negative control, we chose PBMCs as an unrelated system for T2D. scPRS trained on PBMC scATAC-seq data presented inferior predictive performance compared to the model trained on the pancreas data (Supplementary Fig. 5a), highlighting the importance of choosing reference single-cell data from disease-relevant systems or tissues in scPRS.

Lastly, we sought to evaluate scPRS on independent target cohorts. For T2D, we used the Genetics of T2D Consortium<sup>39</sup> (GoT2D) genotype dataset as the independent cohort; for HCM, because the discovery GWAS was performed on UKBB European (EUR) samples, we constructed an independent cohort comprising non-EUR HCM samples and matched controls from UKBB; for AD, we used the AD Neuroimaging Initiative<sup>43</sup> (ADNI) WGS dataset. We trained scPRS models on the basis of the original target cohorts and all PRS methods were tested on the new independent target cohorts. Notably, scPRS still outperformed all baseline methods for HCM and AD (Fig. 2b and Supplementary Fig. 2c). Similarly, scPRS+ further improved the prediction for T2D, surpassing all other baseline PRS approaches (Fig. 2b and Supplementary Fig. 2c). Interestingly, for HCM, even when scPRS was trained on EUR samples, it performed comparably for non-EUR samples (AUROC, 0.692 (EUR) versus 0.643 (non-EUR); Fig. 2b and Supplementary Fig. 2c), suggesting its portability across different populations, although further validation with additional data is needed.



**Fig. 3 | Disease-critical cells identified by scPRS.** **a–c**, Disease-critical cells identified by scPRS for T2D (a), HCM (b) and AD (c). Left, scATAC-seq or snATAC-seq datasets used by scPRS, along with annotated cell types. Middle, disease-relevant cells prioritized by scPRS (in red). Cell clusters enriched with scPRS-prioritized cells are highlighted in closed curves with corresponding cell type colors. Right, enrichment of scPRS-selected disease cells within each cell type.

ORs and *P* values were determined using a one-sided Fisher's exact test. Cell type abbreviations: Fibro, fibroblast; LEC, lymphatic endothelial cell; Peri, pericyte; Schw, Schwann cell; SmMus, smooth muscle cell; VEC, vascular endothelial cell. For robustness, small cell clusters with fewer than 150 cells were excluded from analysis and visualization for all diseases.

### scPRS prioritizes disease-relevant cells

Next, we sought to examine the disease–cell association using scPRS. For each disease, we first trained 100 scPRS models with different random seeds based on the entire target cohort and then prioritized cells whose model weights consistently exceeded those of background cells, designating them as disease-relevant cells (Methods). We also harnessed the knowledge of annotated cell types to facilitate biological interpretation (Methods).

**T2D.** There were 14 cell types identified in the human pancreas<sup>33</sup> (Fig. 3a, left, and Methods), among which two hormone-high cell types (namely, GCG<sup>high</sup> alpha cells and INS<sup>high</sup> beta cells) were significantly enriched

with scPRS-selected cells (adjusted *P* < 0.1, BH correction; Fig. 3a). The original study<sup>44</sup> that generated the pancreas snATAC-seq dataset had linked INS<sup>high</sup> and INS<sup>low</sup> beta cells to T2D risk using the stratified linkage disequilibrium (LD) score regression<sup>45</sup> (sLDSC). As another benchmark, we applied SCAVENGE<sup>46</sup>, a computational method that also enables single-cell-resolved cell prioritization, to the same data (Methods). In addition to GCG<sup>high</sup> alpha cells and INS<sup>high</sup> beta cells, SCAVENGE prioritized GCG<sup>low</sup> alpha cells (adjusted *P* < 0.1, BH correction; Supplementary Fig. 6a). In comparison, cells selected by PBMC-based T2D scPRS exhibited nonspecificity across cell types (Supplementary Fig. 5b).

While pancreatic beta cell dysfunction and cell death are known as key processes in the development of T2D (ref. 47), it is increasingly

evident that T2D may result from defects in multiple cell types<sup>48</sup>. Notably, the alpha cell, which serves as the counterpart to the beta cell and is responsible for producing glucagon, has been increasingly recognized for its role in T2D pathogenesis<sup>49–51</sup>. Single-cell profiling further revealed the diversity within islet endocrine cells, spanning from fine-grained cell states to a continuous spectrum<sup>44</sup>. Our findings, coupled with prior research<sup>44,52</sup>, underscore the complexity of T2D pathogenesis involving multiple cell types within the pancreatic islets.

**HCM.** In the human left ventricle, a total of 17 cell types were identified (Fig. 3b, left, and Methods). Among these, two cell types, including CDMs and pericytes, presented significant enrichment with scPRS-selected cells (adjusted  $P < 0.1$ , BH correction; Fig. 3b). As comparison, we found no genetic enrichment within snATAC-seq peaks of all left-ventricle cell types using sLDSC (Supplementary Fig. 6b and Methods). SCAVENGE also linked CDMs to HCM (adjusted  $P < 0.1$ , BH correction) but enrichment within pericytes was not observed (Supplementary Fig. 6c). CDMs, the primary cell type involved in the process of hypertrophy and thickening of heart muscle, have a pivotal role in HCM pathogenesis<sup>53</sup>. Pathogenic mutations disrupt the normal function of CDMs, leading to structural and functional abnormalities<sup>53</sup>, such as myocardial hypertrophy and fibrosis, contractile dysfunction and arrhythmias. Our scPRS prediction not only reinforces the association between CDM dysfunction and HCM but also extends this connection from protein function to noncoding gene regulation.

Cardiac pericytes interact with endothelial cells through both physical and paracrine mechanisms and are integral in maintaining cardiac and vascular homeostasis<sup>54</sup>. Despite being relatively understudied, the loss and dysfunction of pericytes have been associated with cardiomyopathy<sup>55–57</sup>. Our results confirm this connection and shed light on the potential causal involvement of pericytes in cardiac hypertrophy. Importantly, this link would not have been identified with either sLDSC or SCAVENGE.

**AD.** Eight major cell types were identified in the human cortex<sup>35</sup> (middle frontal and superior and middle temporal gyri; Fig. 3c, left, and Methods), among which three cell types were significantly enriched with scPRS-prioritized cells (adjusted  $P < 0.1$ , BH correction; Fig. 3c), including microglia, astrocytes and oligodendrocyte progenitor cells (OPCs). It is noteworthy that the original study<sup>35</sup> that generated the brain scATAC-seq dataset linked only microglia to AD using sLDSC. Applying SCAVENGE to the same data revealed the same set of AD-relevant cell types as scPRS (adjusted  $P < 0.1$ , BH correction; Supplementary Fig. 6d).

The relationship between microglia and AD has been well established in the literature<sup>58</sup>. Microglia have diverse roles, including immune response, phagocytosis and synapse modulation, contributing extensively to the development and progression of AD pathology.

Moreover, genetic studies consistently prioritize microglia as the most prominent brain cell type associated with AD<sup>59,60</sup>. In recent years, accumulating evidence has underscored the essential role of astrocytes in AD pathogenesis through their reactivation or dysfunction<sup>61,62</sup>. Additionally, latest research has linked OPCs to AD, likely because of its function in immune modulation and remyelination<sup>63</sup>. Our results reinforce these findings and offer further insights into the cellular heterogeneity of AD pathogenesis.

**Severe COVID-19.** scPRS-prioritized cells were significantly enriched in macrophages, natural killer (NK) cells and monocytes (adjusted  $P < 0.1$ , BH correction; Supplementary Fig. 6e). Dysregulated activation of macrophages contributes to tissue damage and disease progression through excessive cytokine production<sup>64–66</sup>. NK cells, crucial for early defense against viral infections, may exacerbate the cytokine storm when impaired<sup>67–70</sup>. Monocytes, as precursors to macrophages, have also been linked to severe COVID-19 because of their role in inflammation and tissue damage<sup>71–73</sup>. In particular, monocytes were also prioritized by SCAVENGE<sup>46</sup> for severe illness.

Of note, scPRS-prioritized cell types aligned with the top-performing single-cell-type peak PRSSs (Supplementary Fig. 3), providing additional insight into the rationale behind scPRS-based cell prioritization.

### scPRS reveals disease regulatory programs

As per model design, scPRS prioritizes cells that contain disease-associated variants within their differentially accessible chromatin regions. This feature empowers us to delve deeper into the regulatory circuits that contribute upstream of the disease across different cell types. To achieve this, we devised a layered multiomic strategy based on the trained scPRS model to systematically map cell-type-specific gene regulation underlying diseases (Fig. 4a and Methods).

For each disease-relevant cell type nominated by scPRS, we first identified the cCREs that were differentially accessible within scPRS-selected cells. Within these, we further prioritized cCREs (referred to as disease-relevant cCREs) that were significantly enriched with disease-associated variants using MAGMA<sup>74</sup>. To map cCRE–gene interactions, we performed coaccessibility analysis<sup>75</sup> on the basis of the scATAC-seq data, supplemented by the closest-gene strategy given its effectiveness in nominating disease genes<sup>76</sup>. For each cell type, this procedure yielded a set of candidate disease genes associated with the disease-relevant cCREs.

To fine-map causal variants within disease-relevant cCREs, we used a sequence-based deep learning model<sup>77–79</sup> that predicted chromatin accessibility across different cell types from the DNA sequence (Supplementary Fig. 4a and Methods). We trained the model using scATAC-seq data and then used it to predict the functional effects of individual variants on chromatin accessibility across cell types (Supplementary

**Fig. 4 | Cell-type-specific genetic regulation in T2D.** **a**, Schematic of scPRS-based multiomic strategy for uncovering disease-relevant genetic regulation. RNAi, RNA interference. The schematic was created using [BioRender.com](https://biorender.com). **b**, Enrichment of T2D-associated variants within cCREs that were differentially accessible in scPRS-prioritized cells. LD threshold  $r^2 = 0.1$  was used in clumping to retrieve an independent variant set ( $n = 783,082$ ).  $P$  values were determined using a two-sided Fisher's exact test. The  $\log_{10}(\text{OR})$  and 95% CI are annotated by the dots and error bars, respectively. **c**, Candidate T2D genes and GO enrichment analysis results. Significant GO terms (adjusted  $P < 0.1$ , BH correction) with OR  $> 5$  are visualized. **d**, Enrichment of TFBS-disrupting variants within seq-DL-panc-prioritized variants (various thresholds applied). seq-DL-panc, the sequence deep learning model trained on the pancreas snATAC-seq data. Enrichment was estimated by  $t$  statistics, where a total of 6,865,604 variants were tested. The box plot center line, limits and whiskers represent the median, quartiles and 1.5× the interquartile range (IQR), respectively. The dots indicate outliers falling above

or below the end of the whiskers. Crosses indicate adjusted  $P > 0.1$ . **e**, Enrichment of seq-DL-panc-prioritized T2D-associated variants (various thresholds applied) within T2D-cCREs. ORs and CIs were determined using a two-sided Fisher's exact test. The  $\log_{10}(\text{OR})$  is annotated by the solid line and 95% CI is represented by the shaded area. The red dashed line indicates null enrichment. **f**, Illustration of the genetic regulation of rs10811660 in INS<sup>high</sup> beta cells. In the bar plot, the asterisk indicates that the percentage of seq-DL-panc score is greater than 85%. In the gene plot, the mapped target gene is highlighted in red. In the link plot, links with coaccessibility  $> 0.05$  are visualized; Coaccess, coaccessibility. **g**, The UMAP plot of the pancreas snATAC-seq dataset showing the expression of MAFA in individual cells. Gene expression was estimated on the basis of gene activity computed by Signac. INS<sup>high</sup> beta cells are highlighted in the dashed closed curve. **h**, Ratio between observed and expected cell counts in GCG<sup>high</sup> alpha (left) and INS<sup>high</sup> beta (right) cells.  $P$  values were determined using a two-sided chi-square test.

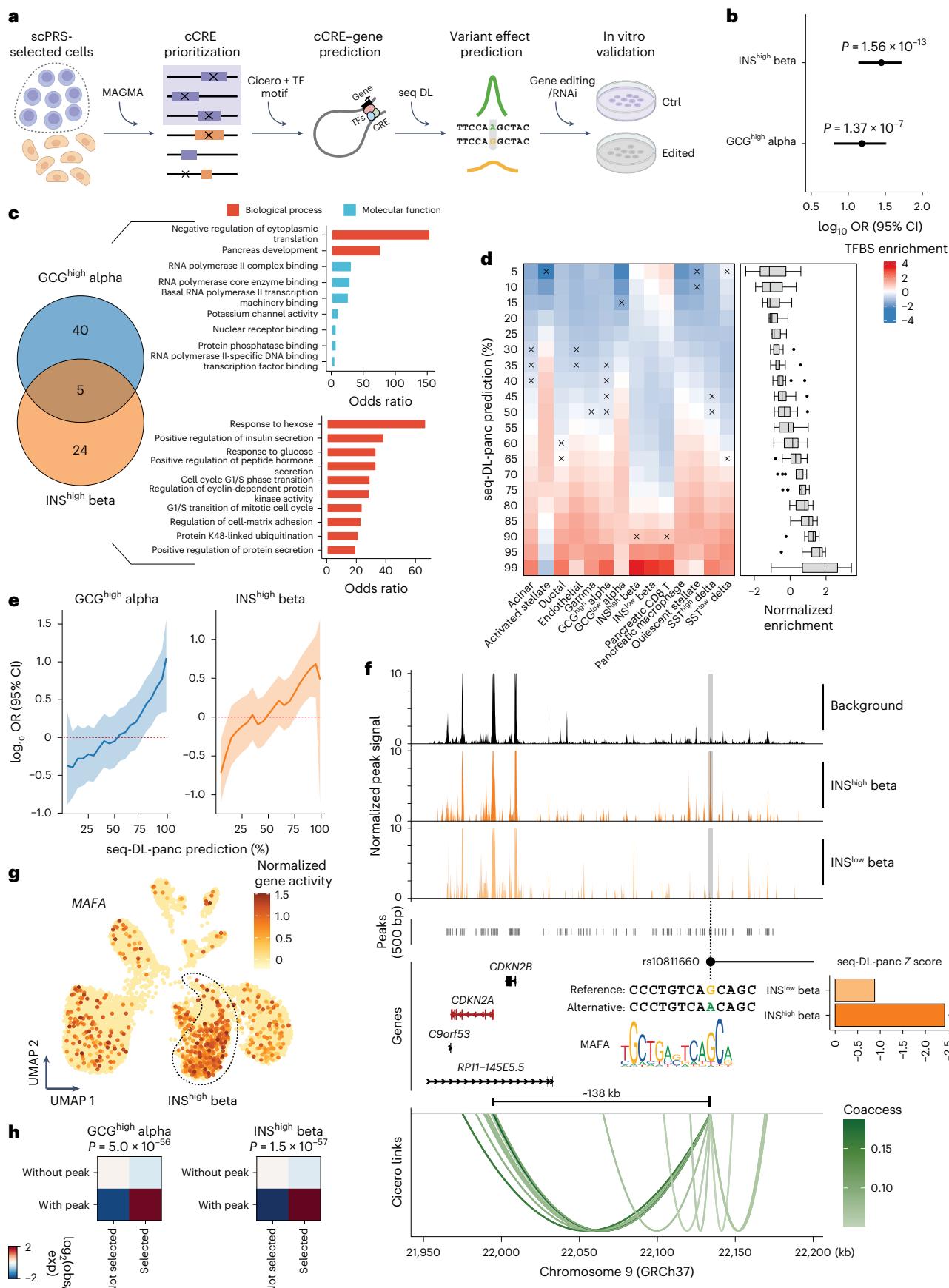


Fig. 7a and Methods). This completed the map of disease-relevant regulatory circuits composed of variant–cCRE–gene trios. Follow-up experiments were carried out in corresponding cell types to validate our predictions.

**T2D.** We first observed a significant enrichment of T2D-associated variants (GWAS  $P < 5 \times 10^{-8}$ ) within differentially accessible cCREs for scPRS-prioritized cells ( $P < 1 \times 10^{-6}$ , two-sided Fisher's exact test; Fig. 4b, Supplementary Fig. 7b and Methods). Using MAGMA, we identified 19 and 22 T2D-relevant cCREs (referred to as T2D-cCREs) in GCG<sup>high</sup> alpha and INS<sup>high</sup> beta cells, respectively (Supplementary Fig. 7c and Supplementary Table 1). Motif enrichment analysis for T2D-cCREs uncovered transcription factors (TFs) of functional importance in corresponding cell types (Supplementary Fig. 7c and Methods). For example, TEAD1 is a critical beta cell TF necessary for coordinating various aspects of adult beta cell function, including proliferative quiescence, mature identity and functional competence to uphold glucose homeostasis<sup>80,81</sup>. MAFB, whose motif is enriched in both cell types, is another pivotal TF in the islet. It is essential for the production and secretion of glucagon in alpha cells<sup>82</sup> and for the maturation of beta cells<sup>83</sup>. A recent study demonstrated that XBP1 has a vital role in maintaining beta cell identity and repressing beta-to-alpha cell transdifferentiation, and is required for beta cell compensation and the prevention of diabetes in insulin resistance states<sup>84</sup>.

By mapping target genes of T2D-cCREs, we identified 45 and 29 candidate risk genes in GCG<sup>high</sup> alpha and INS<sup>high</sup> beta cells, respectively (Fig. 4c and Supplementary Table 1). The function of alpha cell genes was enriched with 'pancreas development' (GO:0031016) and 'RNA polymerase core enzyme binding' (GO:0043175) (adjusted  $P < 0.1$ , BH correction), whereas the function of beta cell genes was enriched with 'response to hexose' (GO:0009746), 'positive regulation of insulin secretion' (GO:0032024) and 'response to glucose' (GO:0009749) (adjusted  $P < 0.1$ , BH correction).

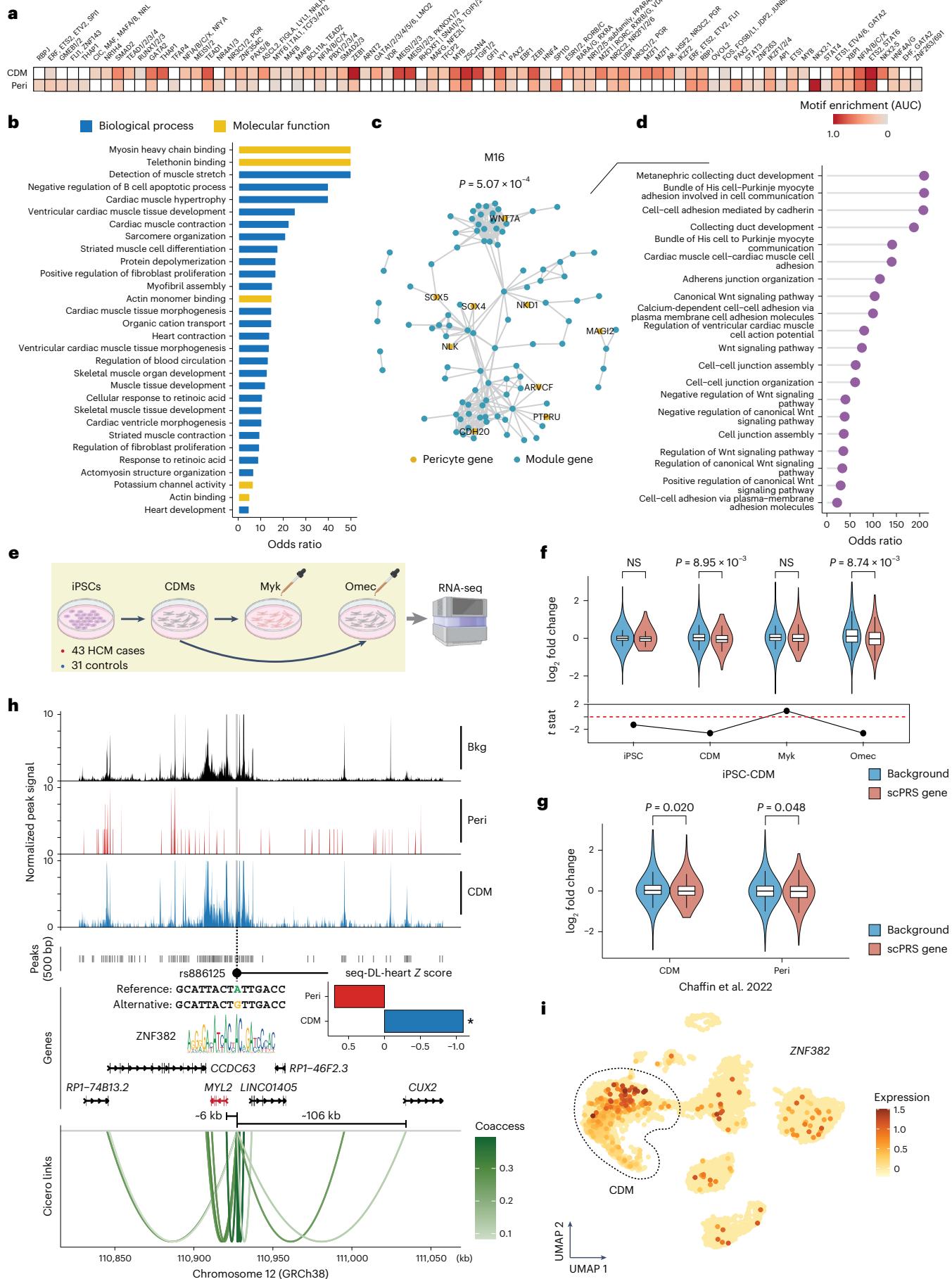
Trained on the pancreas snATAC-seq data, the sequence model exhibited high accuracy in peak prediction (AUROC,  $0.819 \pm 0.011$ ; AUPRC,  $0.639 \pm 0.044$ ; Supplementary Fig. 7d). We validated our variant effect prediction using two different approaches: expression quantitative trait locus (eQTL) analysis and TF-binding site (TFBS) prediction (Methods). Leveraging eQTL datasets generated in relevant tissues<sup>85–88</sup>, we observed that eQTLs tended to display larger effects on the basis of deep learning prediction in related cell types compared to non-eQTLs (Supplementary Fig. 7e). Additionally, variants with larger effects were more likely to alter TF binding<sup>89</sup> (Fig. 4d). These results indicate that the sequence model had captured underlying gene regulation mechanisms. We also examined functional effects of T2D-associated variants (GWAS  $P < 0.05$ ) located within T2D-cCREs in GCG<sup>high</sup> alpha and INS<sup>high</sup> beta cells (Methods). Variants with larger effect sizes showed higher enrichment in T2D-cCREs in corresponding cell types (Fig. 4e), providing additional support for the functional importance of T2D-cCREs we identified.

Combining multiomic evidence from eQTLs, TF binding and sequence model prediction fine-mapped T2D risk variants with functional implications (Supplementary Fig. 7f,g and Supplementary Table 1). One variant of particular interest is **rs10811660**, a T2D GWAS SNP<sup>31</sup> (GWAS  $P = 1.30 \times 10^{-11}$ ,  $\beta = -0.13$ , effect/alternative allele is A) residing within an INS<sup>high</sup> beta cell-specific T2D-cCRE (chr9:22,133,835–22,134,336;  $P = 1.91 \times 10^{-14}$ , log<sub>2</sub> fold change (FC) = 4.99; Fig. 4f). We predicted that the alternative allele specifically reduced the cCRE accessibility in INS<sup>high</sup> beta cells (INS<sup>high</sup> beta cell  $Z = -2.43$ , percentile = 96.84%; Fig. 4f). Furthermore, the affected cCRE was found to be coaccessible with *CDKN2A* (coaccessibility = 0.159; Fig. 4f). Previous studies demonstrated that the p16 inhibitor of cyclin-dependent kinase (p16<sup>INK4A</sup>), encoded by *CDKN2A*, restricts beta cell proliferation during aging, restricts beta cell regeneration, mediates overnutrition-related senescence and reduces insulin secretory function<sup>90</sup>. While **rs10811660** has also been linked to a *CDKN2A* paralog, *CDKN2B*, because of their distance proximity<sup>90</sup>, our coaccessibility analysis suggested that this association might be a false-positive nomination (Fig. 4f). This conclusion was further supported by the islet eQTL data<sup>88</sup>, wherein **rs10811660** was significantly associated with the expression of *CDKN2A* ( $P = 9.94 \times 10^{-4}$ ,  $Z = 3.29$ ) rather than that of *CDKN2B* ( $P > 0.05$ ,  $Z = 0.40$ ; Supplementary Fig. 7h). Additionally, we found that the alternative allele A disrupted the binding motif of MAFA ( $P < 1 \times 10^{-4}$ , motifbreakR<sup>91</sup>; Fig. 4f and Methods), a critical regulator of pancreatic beta cell function<sup>92</sup>, which was more highly expressed in beta cells (Fig. 4g). Collectively, our analysis suggests a genetic regulation influencing T2D risk; the T2D risk allele G (**rs10811660**) increases the abundance of MAFA binding, which further upregulates *CDKN2A* expression in INS<sup>high</sup> beta cells. This aligns with previous evidence implicating that higher expression of *CDKN2A* may increase T2D risk<sup>90</sup>.

Lastly, we sought to characterize scPRS-selected cells beyond the resolution of predefined cell types. In particular, we compared selected cells to unselected ones from the same cell type. Differential accessibility analysis identified two peaks (chr10:94,479,864–94,480,365 and chr10:114,780,533–114,781,034) that were significantly enriched in scPRS-selected GCG<sup>high</sup> alpha cells and three peaks (chr9:22,133,835–22,134,336, chr10:114,758,079–114,758,580 and chr10:114,780,533–114,781,034) enriched within INS<sup>high</sup> beta cells. These marker peaks defined novel cell populations relevant to T2D (Supplementary Fig. 8), as informed by genetic risk. Consistent with this, we further classified each of these two cell types into subtypes on the basis of the accessibility of marker peaks and observed significant enrichment of scPRS-selected cells in the marker-defined subtype ( $P < 1 \times 10^{-50}$ , two-sided chi-square test; Fig. 4h). Of note, the peak chr10:114,780,533–114,781,034 was shared between the selected populations of alpha and beta cells; all marker peaks contained at least one T2D GWAS variant (Supplementary Table 1). The marker peaks of the selected GCG<sup>high</sup> alpha cells were linked to genes such as *TCF7L2* and *CPEB3*, with *TCF7L2* also pinpointed in the selected INS<sup>high</sup> beta cells, suggesting shared T2D biology across these two cell populations. The peak chr9:22,133,835–22,134,336 highlighted

**Fig. 5 | Cell-type-specific genetic regulation in HCM.** **a**, Motif enrichment within HCM-cCREs identified in two HCM-relevant cell types including CDMs and pericyte. Motif enrichment was measured by AUC. Row-wise standardization was performed. Only significant enrichment (adjusted  $P < 0.1$ , Bonferroni correction) is colored. **b**, Bar plot of GO enrichment for CDM HCM risk genes. Significant GO terms (adjusted  $P < 0.1$ , BH correction) with OR  $> 5$  are shown. **c**, The network module M16 enriched with pericyte HCM genes. *P* values were determined using a one-sided hypergeometric test. Edges between module genes are shown. **d**, Lollipop chart of GO enrichment (biological process) for M16 genes. Significant GO terms (adjusted  $P < 0.1$ , BH correction) are shown. **e**, Schematic of iPS cell RNA-seq experiments. Myk, mavacamten; Omec, omecamtiv mecarbil. The schematic was created using [BioRender.com](https://biorender.com). **f**, Expression FC comparison between HCM risk genes and the background transcriptome in CDMs across different conditions. The box plot center line, limits and whiskers represent the

median, quartiles and  $1.5 \times$  the IQR, respectively. *P* values were determined using a two-sided *t*-test ( $n = 16,160$ ). NS, not significant; stat, statistics. **g**, Expression FC comparison between HCM risk genes and the background transcriptome in HCM-relevant cell types based on an HCM snRNA-seq study. The box plot center line, limits and whiskers represent the median, quartiles and  $1.5 \times$  the IQR, respectively. *P* values were determined using a two-sided *t*-test ( $n = 11,683$ ). **h**, Illustration of the genetic regulation of **rs886125** in CDMs. In the bar plot, the asterisk indicates a seq-DL-heart score percentage greater than 85%; seq-DL-heart, the sequence deep model trained on the left-ventricle snATAC-seq data. In the gene plot, differentially expressed target genes are mapped (in red). Bkg, background. **i**, The UMAP plot of the left-ventricle snRNA-seq dataset showing the expression of *ZNF382* in individual cells. Expression was estimated by normalized gene count. CDMs are highlighted in the dashed closed curve.



above (Fig. 4f) was one of the marker peaks of the selected  $\text{INS}^{\text{high}}$  beta cells; this peak contained the T2D risk SNP **rs10811660**, underscoring the cellular specificity of **rs10811660** in impacting T2D risk.

**HCM.** We identified 137 and 358 HCM-relevant cCREs (referred to as HCM-cCREs) that were linked to 199 and 492 target genes in CDMs and pericytes, respectively (Supplementary Table 2). We observed only minimal overlap, with just one cCRE and 24 genes shared between these two cell types, highlighting their cell type specificity.

Our motif enrichment analysis for HCM-cCREs revealed TFs that have critical roles in corresponding cell types (Fig. 5a). For instance, TEAD1 is a pivotal regulator involved in maintaining the proper functioning of adult CDMs, whose loss of function has been associated with dilated cardiomyopathy<sup>93</sup>. GATA4 exerts notable control over cardiac gene expression, impacting embryonic development, CDM differentiation and stress responsiveness of the adult heart<sup>94</sup>. NKX2-5 is a central regulator of heart development and pathogenic mutations within it contribute to progressive cardiomyopathy and conduction defects<sup>95</sup>. Additionally, RBPJ inactivation has been linked to the development of disease-promoting properties in brain pericytes<sup>96</sup>. STAT3 serves as a key regulator of cell–cell communication within the heart, a critical aspect of pericyte functionality<sup>97</sup>.

HCM risk genes identified in CDMs exhibited functional importance in CDMs and cardiomyopathy, such as ‘myosin heavy chain binding’ (GO:0032036), ‘cardiac muscle contraction’ (GO:0060048) and ‘sarcomere organization’ (GO:0045214) (adjusted  $P < 0.1$ , BH correction; Fig. 5b). No Gene Ontology (GO) enrichment was observed for pericyte genes, suggesting a marked functional diversity within this gene set. To better dissect this heterogeneity, we carried out a network analysis on the basis of the protein–protein interactions (PPIs)<sup>98</sup> (Methods), in which one module M16 was significantly enriched with HCM pericyte genes ( $P = 5.07 \times 10^{-4}$ , hypergeometric test; adjusted  $P = 0.034$ , BH correction; Fig. 5c). Genes within this module displayed GO enrichment in various pericyte functions, such as ‘cell–cell adhesion mediated by cadherin’ (GO:0044331), ‘cell–cell junction assembly’ (GO:0007043) and ‘cadherin binding’ (GO:0045296) (adjusted  $P < 0.1$ , BH correction; Fig. 5d and Supplementary Fig. 9a).

To better understand the gene function in the disease context, we analyzed an RNA sequencing (RNA-seq) dataset<sup>27</sup> of induced pluripotent stem cell (iPS cell)-derived CDMs obtained from 43 HCM cases and 31 healthy controls (Fig. 5e). Bulk RNA-seq profiling was conducted under four conditions: iPS cells, differentiated CDMs, mavacamten-treated<sup>99</sup> (an HCM drug recently approved by FDA) CDMs

and omecamtiv mecarbil<sup>100</sup> (a heart failure drug serving as the negative control) treated CDMs. Notably, although the CDM HCM genes exhibited no expression difference in iPS cells between HCM cases and healthy controls, their expression was significantly reduced in differentiated HCM CDMs compared to control cells ( $P = 8.95 \times 10^{-3}$ , two-sided  $t$ -test; Fig. 5f, Supplementary Table 3 and Methods). Intriguingly, the downregulation of HCM genes was rescued by mavacamten treatment ( $P = 0.017$ , two-sided  $t$ -test) but persisted in omecamtiv mecarbil treatment ( $P > 0.05$ , two-sided  $t$ -test; Fig. 5f). The reduced expression of HCM genes identified in CDMs and pericytes was also confirmed in corresponding cell types using an independent HCM single-cell transcriptome dataset<sup>101</sup> (CDM  $P = 0.02$ , pericyte  $P = 0.048$ , two-sided  $t$ -test; Fig. 5g), while showing cell type specificity (Supplementary Fig. 9b,c). These results demonstrate the disease relevance of our HCM genes.

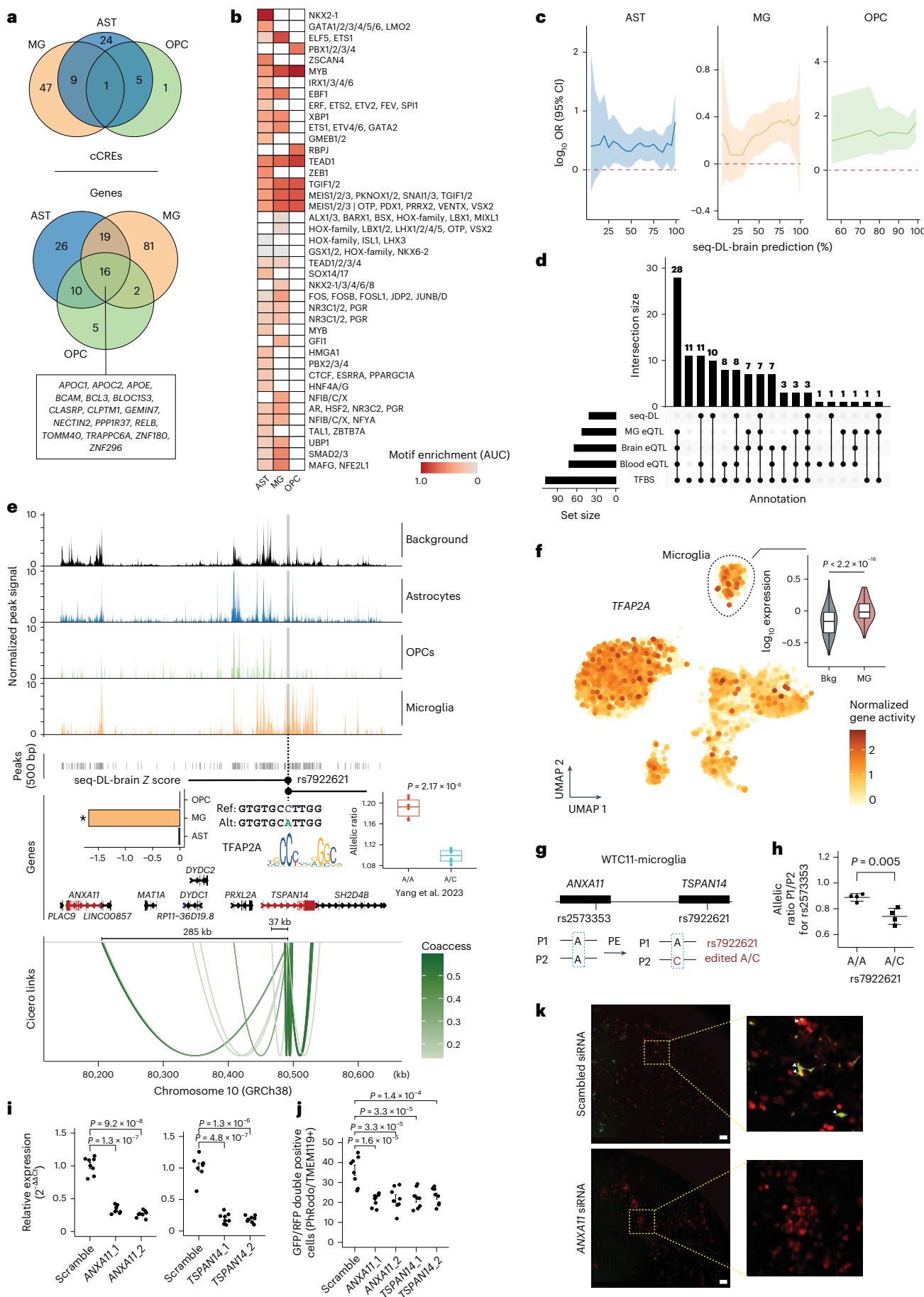
We trained a different sequence deep learning model on the basis of the snATAC-seq dataset of the left ventricle (AUROC,  $0.846 \pm 0.019$ ; AUPRC,  $0.658 \pm 0.032$ ; Supplementary Fig. 9d). Variant effects predicted by the model agreed well with eQTL profiling<sup>85,102</sup> and TFBS prediction (Supplementary Fig. 9e,f). HCM-cCREs presented increased enrichment of HCM-associated variants (GWAS  $P < 0.05$ ) with larger effects (Supplementary Fig. 9g).

The sequence deep learning prediction, together with eQTL and TFBS analyses, fine-mapped novel cell-type-specific HCM risk variants (Supplementary Fig. 9h,i and Supplementary Table 2). As an example, the CDM-specific HCM-cCRE (chr12:110,927,025–110,927,526;  $P = 2.5 \times 10^{-3}$ ,  $\log_2 \text{FC} = 1.94$ ; Fig. 5h) contained a nominally significant GWAS<sup>30</sup> variant **rs886125** (GWAS  $P = 0.019$ ,  $\beta = -0.149$ , effect/alternative allele = G) and was coaccessible (coaccessibility = 0.367) with *MYL2*, a widely recognized HCM gene<sup>80</sup>. On the basis of our predictions, the alternative allele G specifically decreased the cCRE within CDMs (CDM  $Z = -1.10$ , percentile = 87.62%; Fig. 5h) and it disrupted the TFBS of ZNF382 ( $P < 1 \times 10^{-4}$ , motifbreakR; Fig. 5h), which is known as a transcriptional repressor<sup>103</sup>. These results together suggest that the risk-increasing allele A, bound by ZNF382, would lower the expression of *MYL2* in CDMs. This was supported by the eQTL data<sup>86</sup> in which the risk allele A was associated with decreased expression of *MYL2* ( $P = 0.011$ ,  $\beta = 0.125$ ; GTEx artery aorta). Additionally, using our paired snRNA-seq data, we found that ZNF382 was more highly expressed in CDMs (Fig. 5i), highlighting its cell-type-specific role in gene regulation.

**AD.** We first confirmed a significant enrichment of AD-associated variants (GWAS  $P < 5 \times 10^{-8}$ ) within differentially accessible cCREs in scPRS-prioritized cells ( $P < 5 \times 10^{-3}$ , two-sided Fisher’s exact test;

**Fig. 6 | Cell-type-specific genetic regulation in AD.** **a**, Venn diagram of AD-relevant cCREs (top) and genes (bottom) identified by the scPRS-based multiomic strategy. AST, astrocyte; MG, microglia. **b**, Motif enrichment within AD-cCREs across different cell types. Motif enrichment was measured by AUC. Column-wise standardization was performed. Only significant enrichment (adjusted  $P < 0.1$ , Bonferroni correction) is colored.  $P$  values were determined using a hypergeometric test. **c**, Enrichment of seq-DL-prioritized AD-associated variants (various thresholds applied) within AD-cCREs. ORs and CIs were determined using a two-sided Fisher’s exact test. The  $\log_{10}(\text{OR})$  is annotated by the solid line and the 95% CI is represented by the shaded area. The red dashed line indicates null enrichment. **d**, Summary statistics of fine-mapped AD risk variants in microglia using different annotations. **e**, Illustration of the genetic regulation of **rs7922621** in microglia. Box plot: the box plot center line, limits, and whiskers represent the median, quartiles and  $1.5 \times$  the IQR, respectively.  $P$  values were determined using a two-sided  $t$ -test. ref, reference; alt, alternative. In the bar plot, the asterisk indicates a seq-DL-brain score percentage greater than 85%; seq-DL-brain, the sequence deep model trained on the cortex scATAC-seq data. In the gene plot, differentially expressed target genes are mapped (in red). In the link plot, links with coaccessibility  $> 0.05$  are shown. Coaccess, coaccessibility. **f**, The UMAP plot of the cortex scATAC-seq dataset showing the expression of **TFAP2A** in individual cells. In the violin plot,  $P$  values were determined using a two-sided  $t$ -test. Gene expression was estimated on the basis of the gene

activity computed by Signac. Microglia are highlighted in the dashed closed curve. **g**, Diagram showing the haplotypes of variants in wild-type and **rs7922621** prime-edited WTC11-derived microglia. The P1 allele has the risk allele (A), while the P2 allele has the nonrisk allele (C). PE, prime editing. **h**, Allelic imbalance between P1 and P2 alleles for *ANXA11* quantified by **rs2573353** in **rs7922621** wild-type (A/A) and prime-edited (A/C) WTC-derived microglia ( $n = 4$  replicates). The center line and error bar represent the mean and s.d., respectively.  $P$  values were determined using a two-sided  $t$ -test. **i**, RT-qPCR quantification of relative mRNA levels in iMGs treated with siRNAs targeting AD genes or scrambled siRNA ( $n = 2$  siRNAs for each gene;  $n = 8$  replicates for each condition). mRNA levels were normalized to GAPDH.  $P$  values were determined using a two-sided  $t$ -test. Data are presented as the mean  $\pm$  standard error. **j**, Quantification of the number of *TMEM119*<sup>+</sup> cells colocalized with pHrodo particles indicating phagocytosed beads ( $n = 2$  siRNAs for each gene;  $n = 8$  replicates for each condition). A one-way ANOVA with Tukey’s HSD test was used for comparison between siRNA targeting AD genes and scrambled siRNA. Data are presented as the mean  $\pm$  standard error. **k**, Representative images of *TMEM119*<sup>+</sup> (red) iMGs treated with *ANXA11* siRNA or scrambled siRNA showing colocalization of phagocytosed pHrodo particles (green, highlighted with arrows). Images were captured 2 h after incubation with pHrodo. Parts of the images are zoomed in for better visualization. Scale bar, 100  $\mu\text{m}$ .



Supplementary Fig. 10a). We identified 39, 57 and 6 AD-relevant cCREs (referred to as AD-cCREs) that were linked to 71, 118 and 33 target genes in astrocytes, microglia and OPCs, respectively (Fig. 6a and Supplementary Table 4). Numerous AD-cCREs and genes were shared across different cell types, among which we recognized multiple well-established AD genes, such as the *APOE* region genes (*BCAM*, *NECTIN2*, *TOMM40*, *APOE* and *APOC1*), *BCL3* and *PPP1R37*. This signifies their versatile roles in AD pathogenesis.

Next, we examined the function of AD-cCREs and candidate genes in corresponding cell types. We found that AD-cCREs were enriched with binding motifs of cell-type-critical TFs (Fig. 6b). For example, astrocyte AD-cCREs displayed exclusive motif enrichment for GATA4, a regulator of astrocyte cell proliferation and apoptosis<sup>104</sup>; microglia AD-cCREs exhibited significant motif enrichment for SMAD3, which cooperates with PU.1 to enable transcription of some microglia-specific genes<sup>105</sup>; OPC AD-cCREs were exclusively enriched with the RBPJ motif, which is a repressor of OLIG2, a major determinant of oligodendrocyte differentiation and myelination<sup>106</sup>. Additionally, the AD candidate genes also presented cell-type-specific functions. For instance, astrocyte AD genes were enriched with the function of 'regulation of complement activation, classical pathway' (GO:0030450), microglia AD genes displayed enrichment in 'negative regulation of endocytosis' (GO:0045806) and OPC AD genes exhibited significant enrichment in 'IkB kinase and NF- $\kappa$ B signaling' (GO:0007249) (adjusted  $P < 0.1$ , BH correction; Supplementary Table 5).

To characterize the variant effect within AD-cCREs, we trained a sequence deep learning model on the basis of the cortex scATAC-seq data (AUROC,  $0.916 \pm 0.017$ ; AUPRC,  $0.795 \pm 0.059$ ; Supplementary Fig. 10b). We confirmed the agreement in variant effect prediction between the sequence model and two other approaches, including QTL (expression and chromatin accessibility) analysis and TFBS prediction (Supplementary Fig. 10c,d). We also uncovered an enrichment of large-effect AD-associated variants (GWAS  $P < 0.05$ ) within AD-cCREs across all three relevant cell types, where the enrichment was positively correlated with variant effect (Fig. 6c).

We fine-mapped AD risk variants by combining multiomic evidence (Fig. 6d, Supplementary Fig. 10e,f and Supplementary Table 4). Among the prioritized variants, we recognized numerous cell-type-specific risk loci that were previously reported in the literature. For example, the AD risk variant rs10792832 (GWAS<sup>29</sup>  $P = 7.56 \times 10^{-16}$ ,  $\beta = -0.12$ , effect allele/reference = A) was associated with the deactivation of a microglia-specific cCRE for *PICALM*<sup>60</sup>, aligning with our prediction (microglia  $Z = -1.98$ , *PICALM* coaccessibility = 0.246). Another AD risk variant rs13025717 (GWAS  $P = 2.98 \times 10^{-15}$ ,  $\beta = 0.13$ , effect/alternative allele = T), which represses a microglia cCRE for *BIN1* (ref. 35), was also prioritized by our analysis (microglia  $Z = -2.60$ , *BIN1* coaccessibility = 0.382). A recent study validated the role of rs1532278 (GWAS  $P = 3.27 \times 10^{-16}$ ,  $\beta = -0.13$ , effect/reference allele = T) in modulating *CLU* expression in astrocytes<sup>107</sup>, supporting our findings (astrocyte  $Z = -0.498$ , *CLU* coaccessibility = 0.356;).

In addition to known AD risk loci and genes, our analysis discovered novel genetic factors. One of particular interest is rs7922621, which is nominally significant across the genome<sup>29</sup> (GWAS  $P = 2.78 \times 10^{-5}$ ,  $\beta = 0.08$ , effect/alternative allele = A). This variant resides within a microglia-specific AD-cCRE (chr10:82,251,479–82,251,979;  $P = 1.99 \times 10^{-19}$ ,  $\log_2$  FC = 2.39; Fig. 6e). According to the sequence model prediction, rs7922621 diminished the accessibility of this cCRE exclusively in microglia but not in other cell types (microglia  $Z = -1.68$ , percentile = 96.63%; Fig. 6e). Coaccessibility analysis further predicted that this cCRE regulated the expression of two genes: *ANXA11* and *TSPAN14* (Fig. 6e). Importantly, a recent study reported a reduction in local chromatin accessibility associated with rs7922621 in human PS cell-derived microglia<sup>108</sup>. They further validated the reduced expression of *TSPAN14* caused by rs7922621 using prime editing ( $P = 2.17 \times 10^{-6}$ , two-sided  $t$ -test; Fig. 6e). Of note, another variant, rs7910643, located

within the same cCRE and in strong LD with rs7922621 ( $r^2 = 1.0$ , estimated in the 1,000 Genomes EUR population), was shown to be nonfunctional<sup>108</sup>, consistent with our prediction (microglia  $Z = 0.29$ , percentile < 85%; Supplementary Table 4).

To further elucidate the regulatory program involving rs7922621, we conducted TF motif analysis and identified one TF, TFAP2A, whose binding site was disrupted by rs7922621 ( $P < 1 \times 10^{-4}$ , motifbreakR; Fig. 6e). The TFAP2 family is known for its pivotal role in regulating both embryonic and oncogenic development<sup>109</sup>. Furthermore, *TFAP2A* expression showed a significant elevation in microglia compared to other cell types ( $P < 2.2 \times 10^{-16}$ , two-sided  $t$ -test; Fig. 6f), suggesting its functional importance in microglia, although further evidence is required to validate these conclusions.

### Prime editing of rs7922621 alters expression of both *ANXA11* and *TSPAN14* in microglia

Our scPRS-based analysis pinpointed rs7922621 (chr10:82,251,544:C>A) as a candidate AD risk variant and predicted that it regulates two genes (*ANXA11* and *TSPAN14*) by altering the accessibility of a microglia-specific cCRE (chr10:82,251,479–82,251,979; Fig. 6e). Our prior study<sup>108</sup> validated the association between rs7922621 and this cCRE and further demonstrated that the prime editing of rs7922621, converting the risk allele (A) to the nonrisk allele (C) in WTC11 (A/A to A/C)-derived microglia (a male iPS cell line), led to an increase in *TSPAN14* expression. Leveraging the rs7922621-edited clones<sup>108</sup>, we further examined its regulatory role on *ANXA11* (Fig. 6g and Methods). We observed a similar trend in the allelic expression changes of *ANXA11* associated with rs7922621 in WTC11-derived microglia, with the edited nonrisk allele upregulating *ANXA11* compared to the risk allele ( $P = 0.005$ , two-sided  $t$ -test; Fig. 6h). We note that, in contrast to *TSPAN14*, *ANXA11* exhibits a long-range interaction (-285 kb) with rs7922621 (Fig. 6e). Altogether, these results suggest an upstream role of rs7922621 in modulating the expression of both *ANXA11* and *TSPAN14* in microglia, with the AD risk allele (A) reducing their expression.

### Suppression of *ANXA11* and *TSPAN14* impairs microglial phagocytosis

To elucidate the function of *ANXA11* and *TSPAN14* in microglia, we examined the effect of knockdown of these genes on microglial phagocytic activity. In particular, we individually suppressed *ANXA11* and *TSPAN14* in iPS cell-derived microglia-like cells<sup>110,111</sup> (iMGs) using small interfering RNA (siRNA), in which two different siRNAs were tested for each gene. Phagocytosis activity was measured using a fluorescent readout of pHrodo particles. A reduction in expression following siRNA treatment was confirmed for both genes ( $P < 1 \times 10^{-4}$ , two-sided  $t$ -test; Fig. 6i). Notably, suppression of these two genes resulted in significantly decreased iMG uptake of pHrodo particles compared to scrambled siRNA treatments (adjusted  $P < 1 \times 10^{-3}$ , one-way analysis of variance (ANOVA) with Tukey's honestly significant difference (HSD) test; Fig. 6j). These results were consistent across treatments using different siRNAs. Our experimental results validated the functional importance of *ANXA11* and *TSPAN14*, showing that their suppression impaired microglial phagocytosis, thus supporting the pivotal role of rs7922621 in modulating AD risk through its impact on microglial function.

## Discussion

GWAS has substantially advanced our understanding of the genetic basis of complex human diseases<sup>112</sup>. Traditionally, these studies aim to identify genetic loci that reach genome-wide significance (that is, GWAS  $P < 5 \times 10^{-8}$ ). However, for many diseases, the best predictive performance is only achieved by including nominally significant or even nonsignificant variants in PRS calculation<sup>113</sup>. This suggests that the genetic factors contributing to diseases extend beyond those genome-wide significant loci and cannot be fully uncovered by conventional approaches<sup>114</sup>. While scientists have been calling for larger

GWAS consortia and meta-analyses to identify more disease risk loci<sup>15</sup>, it remains an open question how to increase the discovery power given relatively limited sample size. Incorporating prior knowledge or multi-omic data into genetic association analysis has proven to be an effective solution<sup>6,16</sup>.

PRS has been demonstrated as a powerful tool to predict an individual's disease risk. However, it lacks the ability to provide insights into disease mechanisms. From the perspective of modern machine learning, model interpretation is critical in uncovering latent features that contribute to prediction and understanding how models make decisions<sup>17</sup>. As a score computed by aggregating a wide range of variants, PRS offers limited knowledge on the significance of each variant in prediction. Moreover, distinguishing causal variants from statistically correlated elements poses an even greater challenge. For example, a variant can be associated with the disease through its linkage with the causal variant, yet both are treated equivalently within a PRS model. This lack in biology-informed model interpretability can, in turn, constrain predictive performance such as generalizability<sup>15</sup>.

We designed scPRS, a deep learning-based PRS framework, to address these challenges. scPRS leverages single-cell epigenetic data to dissect the genome-wide PRS and then integrates single-cell-level PRSs using a GNN. By breaking down PRS into higher-resolution components informed by cellular functions, scPRS not only enhances its predictive power but also allows for a systematic exploration of cellular and molecular basis for diseases. Applications to various diseases have shown that scPRS outperformed a variety of existing PRS methods. Importantly, this superior predictive performance of scPRS was achieved using less than 11% of all the variants (that is, variants located within open chromatin regions; Supplementary Table 6), highlighting the importance of incorporating functional data<sup>15</sup> and suggesting a notable contribution of noncoding variants to disease risk<sup>18</sup>.

We showcased the effectiveness of scPRS in identifying disease-critical cells. Our method is not confined to cell clustering and predefined cell types, offering an unbiased, agnostic analysis. Through single-cell-resolved modeling, scPRS can discover disease-relevant cell populations by integrating genetic insights. This was demonstrated in identifying previously uncharacterized T2D-related alpha and beta cell populations defined by scPRS-selected cells. Similar analysis was performed for other three diseases but no significance was observed. Unlike the selected cells enriched within highlighted cell types, the model-prioritized cells in other cell types were sparsely distributed in the epigenome space, suggesting less homogeneous cellular functions. The selection of these cells was likely because of the randomness of model initialization and training, as well as the technical noise inherent in single-cell sequencing. Therefore, we recommend considering these cells as background.

The cell type prioritization results agreed well between scPRS and sLDSC but the most notable difference occurred for HCM, where sLDSC failed to identify any relevant cell types. This lack in power could be explained by the difference between bulk and individual-level or single-cell-level modeling. First, sLDSC estimates heritability enrichment across an ensemble of open chromatin regions within a specific cell type but this bulk-level approach does not account for the variation among individual cells. This limitation can lead to confounding by non-disease-relevant regions, reducing its ability to identify disease-critical cells with a high sensitivity. In contrast, scPRS weighs the importance of each cell relative to others, providing a global model that captures cross-cell variation in heritability enrichment. The subsequent cell enrichment analysis within each cell type enables an effective disease–cell association discovery. Another advantage of this single-cell-resolved approach is its ability to identify novel disease-relevant cell populations that are not annotated in the single-cell dataset, which has been demonstrated for T2D. Second, sLDSC works in the GWAS space and its performance can be influenced by the power of the original GWAS. This may explain why sLDSC

identified zero relevant cell types for HCM, as the HCM GWAS was relatively underpowered, with only two genome-wide significant variants<sup>30</sup> ( $P < 5 \times 10^{-8}$ ). Although GWAS summary statistics are part of the input to scPRS, it selects cells whose cell-level PRSs best differentiate individual patients from controls rather than relying on the overall GWAS performance. This individual-level modeling further increases the power to identify disease-relevant cells. Of note, the cell type enrichment *P* values for sLDSC, SCAVENGE and scPRS are not directly comparable. Significant cell types were identified by comparing *P* values derived from the same method. Hence, the conclusions regarding cell type enrichment are comparable across different approaches.

Several recent studies<sup>46,119,120</sup> have also achieved prioritization of disease-relevant cells at the single-cell level. However, these approaches rely on GWAS summary statistics and, thus, lack predictive power. Moreover, superior to these methods, scPRS enables pinpointing disease risk variants, genes and regulatory programs across different cellular contexts, substantially enhancing the power and resolution of genetic discovery. This advancement is exemplified by rs7922621, which was pinpointed by scPRS-based analysis as a candidate AD risk variant but missed by GWAS because of its nominal significance. rs7922621 was also nominated in two recent studies<sup>108,121</sup>, where it was mapped to *TSPAN14* in microglia as the target gene. Our scPRS-based analysis further linked rs7922621 to another gene *ANXA11*. The understanding of the role of *ANXA11* in neurodegenerative disease is rapidly evolving; it was first implicated as a genetic cause of amyotrophic lateral sclerosis<sup>122</sup> (ALS) and later corticobasal syndrome<sup>123</sup>. Recently, *ANXA11* was also revealed as a central pathology in specific subtypes of frontotemporal dementia<sup>124</sup>. It is interesting that *ANXA11* pathology extends beyond neurons; for example, muscle pathology has been observed as part of a multisystem proteinopathy with prominent myopathy<sup>125</sup>. A gap remains in our understanding of the biology underlying *ANXA11* dysfunction. *ANXA11* protein is involved in the tethering of RNA granules, including lysosomes with a role in RNA transport<sup>126</sup>. However, this mechanism does not easily account for all of the pathological observations made. Our results suggest a role of *ANXA11* in microglia that underpins AD risk. This is reminiscent of observations of *TBK1*, another ALS gene where distinct pathological mechanisms have been observed in neurons and microglia<sup>127</sup>. We experimentally validated the regulatory relationship between rs7922621 and *ANXA11* and the function of these two genes (*ANXA11* and *TSPAN14*) in maintaining microglial phagocytosis. Our data support a model where rs7922621 increases AD risk by reducing a microglia cCRE targeting *ANXA11* and *TSPAN14* and then suppressing their expression, which impairs microglial phagocytosis.

It is worth noting that we identified rs7922621 in microglia by starting with a comparison analysis of cells prioritized by scPRS, without which the subsequent analysis would not have been possible (Fig. 4a). While microglia are a well-recognized cell type in AD, we pinpointed this AD-microglia linkage without incorporating any prior knowledge, using an agnostic, unbiased approach. This serves as a positive control to demonstrate the effectiveness of scPRS in identifying disease-critical cells. The nomination of new disease-relevant cell populations that are not annotated in the original single-cell dataset (Fig. 4h) underscores the discovery power of scPRS beyond simply reaffirming known disease–cell associations.

HCM is a genetic condition with a heritability of up to 50% in its familial form<sup>53</sup> and an estimated SNP-based heritability ranging from 0.17 to 0.29 (ref. 128). The genetic study of HCM has been traditionally focused on rare pathogenic coding variants<sup>53</sup>. However, approximately 40% of persons with HCM remained unexplained by known pathogenic variants. Previous HCM GWASs for common variants have been underpowered, likely because of the limited number of participants recruited, resulting in an incomplete knowledge of the genetic architecture<sup>129</sup>. Our scPRS-based analysis greatly expands our understanding of HCM genetics, highlighting the critical role of common noncoding

variants in influencing disease risk. Our findings underscore the importance of regulatory variants that have been largely overlooked in the HCM field. These variants may act as risk modifiers through modulating the expression of their target genes, including known HCM risk genes such as *MYL2*. Although further validations are necessary, our results shed light on the complexity of HCM genetics and biology.

We also constructed C+T PRSs using disease-associated variants (GWAS  $P < 0.05$ ) located within disease-relevant cCREs. Disease cCREs from different prioritized cell types were aggregated. We observed that these selected variants dominated scPRS prediction, showing comparable performance for HCM and T2D (Supplementary Fig. 11). This result supports the effectiveness of our scPRS-based framework in identifying cell-type-specific disease-related variants. However, we caution that it cannot be concluded that this PRS, built on selected variants, is comparable to scPRS in terms of prediction, as it was derived from scPRS after explicitly seeing all samples, which may have caused the overfitting issue.

Single-cell genetics is an emerging field that is reshaping our understanding of genotype–phenotype relationships<sup>17</sup>. By integrating single-cell genomic data into genetic analysis, single-cell genetics provides a novel instrument to link genetic variants to diverse cellular processes. This is well exemplified by single-cell eQTL studies<sup>130–132</sup>, which enable the identification of context-dependent eQTLs that vary across cell states or cell types. scPRS lays the methodological foundation of single-cell genetics, marking a step toward mapping the genetic basis of complex diseases in a single-cell-resolved context.

We note that scATAC-seq only annotates genomic regions that are potentially involved in transcriptional regulation (for example, promoters, enhancers and silencers), whereas other layers of functions, such as proteins, translation and post-transcription, are not considered in our current modeling. Considering the heterogeneity and complexity of a disease's genetic architecture, the prediction of scPRS could be suboptimal for certain diseases wherein coding, splicing or other variants have an important role. Indeed, we observed better predictive performance of C+T over scPRS for T2D. As a compensation, we further incorporated nonpeak PRSs constructed using variants located outside scATAC-seq peaks into scPRS, resulting in scPRS+, which outperformed all baseline PRS methods across the board. Following the same design principle, scPRS can be extended to include a wider range of variants by integrating additional modalities, such as scRNA-seq<sup>133,134</sup> and single-cell DNA methylation<sup>135,136</sup>. This will be explored in our future work.

Considering both time and space complexities, we recommend starting with a moderately large number of cells, such as the 10,000 used in this manuscript, when applying scPRS in practice. This approach ensures coverage of cases where the disease or phenotype is driven by rare cell types or populations. Moreover, this strategy aligns with the exploratory nature of scPRS, as it is a discovery process in which the disease-relevant cells are largely unknown until analysis, typically requiring multiple iterations of testing.

In summary, scPRS stands as a versatile framework for simultaneous disease prediction and biological discovery, enabling the dissection of the genetic, cellular and molecular heterogeneity underlying complex 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/s41587-025-02725-6>.

## References

- Ding, Y. et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618**, 774–781 (2023).
- Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
- Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
- Uffelmann, E. et al. Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 59 (2021).
- Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).
- Choi, S. W. & O'Reilly, P. F. PRSice-2: polygenic risk score software for biobank-scale data. *Gigascience* **8**, giz082 (2019).
- Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
- Thompson, D. J. et al. A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLoS ONE* **19**, e0307270 (2024).
- Choi, Y. H. & Kim, J. K. Dissecting cellular heterogeneity using single-cell RNA sequencing. *Mol. Cells* **42**, 189–199 (2019).
- Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* **1842**, 1896–1902 (2014).
- Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
- Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
- Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annu. Rev. Biomed. Data Sci.* **5**, 293–320 (2022).
- Piwecka, M., Rajewsky, N. & Rybak-Wolf, A. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nat. Rev. Neurol.* **19**, 346–362 (2023).
- Amariuta, T. et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat. Genet.* **52**, 1346–1354 (2020).
- Li, H., Zeng, J., Snyder, M. P. & Zhang, S. Modeling gene interactions in polygenic prediction via geometric deep learning. *Genome Res.* **35**, 178–187 (2025).
- Cuomo, A. S. E., Nathan, A., Raychaudhuri, S., MacArthur, D. G. & Powell, J. E. Single-cell genomics meets human genetics. *Nat. Rev. Genet.* **24**, 535–549 (2023).
- Hamilton, W. L., Ying, R. & Leskovec, J. Representation learning on graphs: methods and applications. *IEEE Data Eng. Bull.* **40**, 52–74 (2017).
- Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
- The ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
- Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. Preprint at <https://doi.org/10.48550/arxiv.1609.02907> (2016).
- Vuckovic, D. et al. The polygenic and monogenic basis of blood traits and diseases. *Cell* **182**, 1214–1231 (2020).
- Li, J. et al. Decoding the genomics of abdominal aortic aneurysm. *Cell* **174**, 1361–1372 (2018).
- Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
- Li, J., Li, X., Zhang, S. & Snyder, M. Gene–environment interaction in the era of precision medicine. *Cell* **177**, 38–44 (2019).

26. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
27. Monte, E. et al. Personalized transcriptome signatures in a cardiomyopathy stem cell biobank. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.05.10.593618> (2024).
28. Gaziano, J. M. et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
29. Kunkle, B. W. et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A $\beta$ , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
30. Sakae, S. et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
31. Scott, R. A. et al. An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).
32. COVID-19 Host Genetics Initiative Mapping the human genetic architecture of COVID-19. *Nature* **600**, 472–477 (2021).
33. Chiou, J. et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).
34. ENCODE Project Consortium et al. Perspectives on ENCODE. *Nature* **583**, 693–698 (2020).
35. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
36. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
37. Weissbrod, O. et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat. Genet.* **54**, 450–458 (2022).
38. Weissbrod, O. et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
39. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
40. Zhou, B. et al. Identification of a splicing variant that regulates type 2 diabetes risk factor CDKAL1 level by a coding-independent mechanism in human. *Hum. Mol. Genet.* **23**, 4639–4650 (2014).
41. Diamanti, K. et al. Organ-specific metabolic pathways distinguish prediabetes, type 2 diabetes, and normal tissues. *Cell Rep. Med.* **3**, 100763 (2022).
42. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001 (2021).
43. Petersen, R. C. et al. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. *Neurology* **74**, 201–209 (2010).
44. Chiou, J. et al. Single-cell chromatin accessibility identifies pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. *Nat. Genet.* **53**, 455–466 (2021).
45. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
46. Yu, F. et al. Variant to function mapping at single-cell resolution through network propagation. *Nat. Biotechnol.* **40**, 1644–1653 (2022).
47. Eizirik, D. L., Pasquali, L. & Cnop, M. Pancreatic  $\beta$ -cells in type 1 and type 2 diabetes mellitus: different pathways to failure. *Nat. Rev. Endocrinol.* **16**, 349–362 (2020).
48. Suzuki, K. et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, 347–357 (2024).
49. Tellez, K. et al. In vivo studies of glucagon secretion by human islets transplanted in mice. *Nat. Metab.* **2**, 547–557 (2020).
50. Dai, X.-Q. et al. Heterogenous impairment of  $\alpha$  cell function in type 2 diabetes is linked to cell maturation state. *Cell Metab.* **34**, 256–268 (2022).
51. Gilon, P. The role of  $\alpha$ -cells in islet function and glucose homeostasis in health and type 2 diabetes. *J. Mol. Biol.* **432**, 1367–1394 (2020).
52. Wang, G. et al. Integrating genetics with single-cell multiomic measurements across disease states identifies mechanisms of beta cell dysfunction in type 2 diabetes. *Nat. Genet.* **55**, 984–994 (2023).
53. Marian, A. J. & Braunwald, E. Hypertrophic cardiomyopathy: genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circ. Res.* **121**, 749–770 (2017).
54. Avolio, E., Campagnolo, P., Katare, R. & Madeddu, P. The role of cardiac pericytes in health and disease: therapeutic targets for myocardial infarction. *Nat. Rev. Cardiol.* **21**, 106–118 (2024).
55. Cattaneo, M. et al. The longevity-associated *BPIFB4* gene supports cardiac function and vascularization in ageing cardiomyopathy. *Cardiovasc. Res.* **119**, 1583–1595 (2023).
56. Rolle, I. G. et al. Heart failure impairs the mechanotransduction properties of human cardiac pericytes. *J. Mol. Cell. Cardiol.* **151**, 15–30 (2021).
57. Chintalgiattu, V. et al. Coronary microvascular pericytes are the cellular target of sunitinib malate-induced cardiotoxicity. *Sci. Transl. Med.* **5**, 187ra69 (2013).
58. Hansen, D. V., Hanson, J. E. & Sheng, M. Microglia in Alzheimer's disease. *J. Cell Biol.* **217**, 459–472 (2018).
59. Lopes, K. et al. Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nat. Genet.* **54**, 4–17 (2022).
60. Kosoy, R. et al. Genetics of the human microglia regulome refines Alzheimer's disease risk loci. *Nat. Genet.* **54**, 1145–1154 (2022).
61. Bellaver, B. et al. Astrocyte reactivity influences amyloid- $\beta$  effects on tau pathology in preclinical Alzheimer's disease. *Nat. Med.* **29**, 1775–1781 (2023).
62. Price, B. R., Johnson, L. A. & Norris, C. M. Reactive astrocytes: the nexus of pathological and clinical hallmarks of Alzheimer's disease. *Ageing Res. Rev.* **68**, 101335 (2021).
63. Buchanan, J., da Costa, N. M. & Cheadle, L. Emerging roles of oligodendrocyte precursor cells in neural circuit development and remodeling. *Trends Neurosci.* **46**, 628–639 (2023).
64. Sefik, E. et al. Inflammasome activation in infected macrophages drives COVID-19 pathology. *Nature* **606**, 585–593 (2022).
65. Chen, S. T. et al. A shift in lung macrophage composition is associated with COVID-19 severity and recovery. *Sci. Transl. Med.* **14**, eabn5168 (2022).
66. Lian, Q. et al. Differential effects of macrophage subtypes on SARS-CoV-2 infection in a human pluripotent stem cell-derived model. *Nat. Commun.* **13**, 2028 (2022).
67. Zhang, S. et al. Multiomic analysis reveals cell-type-specific molecular determinants of COVID-19 severity. *Cell Syst.* **13**, 598–614 (2022).
68. Miorin, L. et al. SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling. *Proc. Natl. Acad. Sci. USA* **117**, 28344–28354 (2020).
69. Vietzen, H. et al. Deletion of the NKG2C receptor encoding *KLRC2* gene and HLA-E variants are risk factors for severe COVID-19. *Genet. Med.* **23**, 963–967 (2021).
70. Maucourant, C. et al. Natural killer cell immunotypes related to COVID-19 disease severity. *Sci. Immunol.* **5**, eabd6832 (2020).
71. Wack, A. Monocyte and dendritic cell defects in COVID-19. *Nat. Cell Biol.* **23**, 445–447 (2021).
72. Saichi, M. et al. Single-cell RNA sequencing of blood antigen-presenting cells in severe COVID-19 reveals multi-process defects in antiviral immunity. *Nat. Cell Biol.* **23**, 538–551 (2021).
73. Arunachalam, P. S. et al. Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* **369**, 1210–1220 (2020).

74. de Leeuw, C. A., Mooij, J. M. & Heskes, T. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
75. Pliner, H. A. et al. Cicero predicts *cis*-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
76. Gazal, S. et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
77. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
78. Zhou, J. et al. Deep learning sequence-based *ab initio* prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
79. Chen, K. M., Cofer, E. M., Zhou, J. & Troyanskaya, O. G. Selene: a PyTorch-based deep learning library for sequence data. *Nat. Methods* **16**, 315–318 (2019).
80. De Bortoli, M. et al. Novel missense variant in *MYL2* gene associated with hypertrophic cardiomyopathy showing high incidence of restrictive physiology. *Circ. Genom. Precis. Med.* **13**, e002824 (2020).
81. Lee, J. et al. Tead1 reciprocally regulates adult  $\beta$ -cell proliferation and function. *eLife* **13**, RP95603 (2024).
82. Katoh, M. C. et al. MafB is critical for glucagon production and secretion in mouse pancreatic  $\alpha$  cells *in vivo*. *Mol. Cell Biol.* **38**, e00504-17 (2018).
83. Artner, I. et al. MafB is required for islet beta cell maturation. *Proc. Natl. Acad. Sci. USA* **104**, 3853–3858 (2007).
84. Lee, K. et al. XBP1 maintains beta cell identity, represses beta-to-alpha cell transdifferentiation and protects against diabetic beta cell failure during metabolic stress in mice. *Diabetologia* **65**, 984–996 (2022).
85. Võsa, U. et al. Large-scale *cis*- and *trans*-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
86. GTEx Consortium The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
87. Kerimov, N. et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
88. Alonso, L. et al. TIGER: the gene expression regulatory variation landscape of human pancreatic islets. *Cell Rep.* **37**, 109807 (2021).
89. Kumar, S., Ambrosini, G. & Bucher, P. SNP2TFBS—a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* **45**, D139–D144 (2017).
90. Kong, Y., Sharma, R. B., Nwosu, B. U. & Alonso, L. C. Islet biology, the *CDKN2A/B* locus and type 2 diabetes risk. *Diabetologia* **59**, 1579–1593 (2016).
91. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–3849 (2015).
92. Liang, J., Chirikjian, M., Pajvani, U. B. & Bartolomé, A. MafA regulation in  $\beta$ -cells: from transcriptional to post-translational mechanisms. *Biomolecules* **12**, 535 (2022).
93. Liu, R. et al. Tead1 is required for maintaining adult cardiomyocyte function, and its loss results in lethal dilated cardiomyopathy. *JCI Insight* **2**, e93343 (2017).
94. Pikkarainen, S., Tokola, H., Kerkelä, R. & Ruskoaho, H. GATA transcription factors in the developing and adult heart. *Cardiovasc. Res.* **63**, 196–207 (2004).
95. Pashmforoush, M. et al. Nkx2-5 pathways and congenital heart disease; loss of ventricular myocyte lineage specification leads to progressive cardiomyopathy and complete heart block. *Cell* **117**, 373–386 (2004).
96. Diéguez-Hurtado, R. et al. Loss of the transcription factor RBPJ induces disease-promoting properties in brain pericytes. *Nat. Commun.* **10**, 1–19 (2019).
97. Haghikia, A., Ricke-Hoch, M., Stapel, B., Gorst, I. & Hilfiker-Kleiner, D. STAT3, a key regulator of cell-to-cell communication in the heart. *Cardiovasc. Res.* **102**, 281–289 (2014).
98. Szklarczyk, D. et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
99. Nag, S., Gollapudi, S. K., Del Rio, C. L., Spudich, J. A. & McDowell, R. Mavacamten, a precision medicine for hypertrophic cardiomyopathy: from a motor protein to patients. *Sci Adv* **9**, eab07622 (2023).
100. Teerlink, J. R. et al. Cardiac myosin activation with omecamtiv mecarbil in systolic heart failure. *N. Engl. J. Med.* **384**, 105–116 (2021).
101. Chaffin, M. et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180 (2022).
102. GTEx Consortium Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
103. Ma, L. et al. ZNF382 controls mouse neuropathic pain via silencer-based epigenetic inhibition of *Cxcl13* in DRG neurons. *J. Exp. Med.* **218**, e20210920 (2021).
104. Agnihotri, S., Wolf, A., Picard, D., Hawkins, C. & Guha, A. GATA4 is a regulator of astrocyte cell proliferation and apoptosis in the human and murine central nervous system. *Oncogene* **28**, 3033–3046 (2009).
105. Crotti, A. & Ransohoff, R. M. Microglial Physiology and pathophysiology: insights from genome-wide transcriptional profiling. *Immunity* **44**, 505–515 (2016).
106. Fujimoto, M. et al. RBP-J promotes neuronal differentiation and inhibits oligodendroglial development in adult neurogenesis. *Dev. Biol.* **332**, 339–350 (2009).
107. Liu, Z. et al. Astrocytic response mediated by the CLU risk allele inhibits OPC proliferation and myelination in a human iPSC model. *Cell Rep.* **42**, 112841 (2023).
108. Yang, X. et al. Functional characterization of Alzheimer’s disease genetic variants in microglia. *Nat. Genet.* **55**, 1735–1744 (2023).
109. Jin, C. et al. Crucial role of the transcription factors family activator protein 2 in cancer: current clue and views. *J. Transl. Med.* **21**, 371 (2023).
110. Shi, Y. et al. Identification and therapeutic rescue of autophagosome and glutamate receptor defects in C9ORF72 and sporadic ALS neurons. *JCI Insight* **5**, e127736 (2019).
111. Shi, Y. et al. Haploinsufficiency leads to neurodegeneration in C9ORF72 ALS/FTD human induced motor neurons. *Nat. Med.* **24**, 313–325 (2018).
112. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
113. Wang, Y. et al. Global Biobank analyses provide lessons for developing polygenic risk scores across diverse cohorts. *Cell Genom.* **3**, 100241 (2023).
114. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
115. Zhou, W. et al. Global Biobank Meta-analysis Initiative: powering genetic discovery across human disease. *Cell Genom.* **2**, 100192 (2022).
116. Zhang, S. et al. Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis. *Neuron* **110**, 992–1008 (2022).

117. Li, X. et al. Interpretable deep learning: interpretation, interpretability, trustworthiness and beyond. *Knowl. Inf. Syst.* **64**, 3197–3234 (2022).
118. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015).
119. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
120. Ma, Y. et al. Polygenic regression uncovers trait-relevant cellular contexts through pathway activation transformation of single-cell RNA sequencing data. *Cell Genom.* **3**, 100383 (2023).
121. Li, Y. E. et al. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
122. Smith, B. N. et al. Mutations in the vesicular trafficking protein annexin A11 are associated with amyotrophic lateral sclerosis. *Sci. Transl. Med.* **9**, eaad9157 (2017).
123. Snyder, A. et al. An ANXA11 P93S variant dysregulates TDP-43 and causes corticobasal syndrome. *Alzheimers. Dement.* **20**, 5220–5235 (2024).
124. Arseni, D. et al. Heteromeric amyloid filaments of ANXA11 and TDP-43 in FTLD-TDP type C. *Nature* **634**, 662–668 (2024).
125. Leoni, T. B. et al. A novel multisystem proteinopathy caused by a missense ANXA11 variant. *Ann. Neurol.* **90**, 239–252 (2021).
126. Liao, Y.-C. et al. RNA granules hitchhike on lysosomes for long-distance transport, using annexin A11 as a molecular tether. *Cell* **179**, 147–164 (2019).
127. Gerbino, V. et al. The loss of TBK1 kinase activity in motor neurons or in all cell types differentially impacts ALS disease progression in SOD1 mice. *Neuron* **106**, 789–805 (2020).
128. Tadros, R. et al. Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. *Nat. Genet.* **53**, 128–134 (2021).
129. Harper, A. R. et al. Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. *Nat. Genet.* **53**, 135–142 (2021).
130. Yazar, S. et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
131. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
132. Kang, J. B., Raveane, A., Nathan, A., Soranzo, N. & Raychaudhuri, S. Methods and insights from single-cell expression quantitative trait loci. *Annu. Rev. Genomics Hum. Genet.* **24**, 277–303 (2023).
133. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eab4290 (2022).
134. Siletti, K. et al. Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
135. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
136. Tian, W. et al. Single-cell DNA methylation and 3D genome architecture in the human brain. *Science* **382**, eadf5357 (2023).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

<sup>1</sup>Department of Epidemiology, University of Florida, Gainesville, FL, USA. <sup>2</sup>Departments of Biostatistics & Biomedical Engineering, UF Genetics Institute, University of Florida, Gainesville, FL, USA. <sup>3</sup>Department of Genetics, Center for Genomics and Personalized Medicine, Stanford University School of Medicine, Stanford, CA, USA. <sup>4</sup>VA Palo Alto Healthcare System, Palo Alto, CA, USA. <sup>5</sup>Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China. <sup>6</sup>Arc Institute, Palo Alto, CA, USA. <sup>7</sup>Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA.

<sup>8</sup>Bioinformatics and Systems Biology Program, University of California, San Diego, La Jolla, CA, USA. <sup>9</sup>Department of Stem Cell Biology and Regenerative Medicine, Eli and Edythe Broad Center for Regenerative Medicine and Stem Cell Research, University of Southern California, Los Angeles, CA, USA.

<sup>10</sup>Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, USA. <sup>11</sup>Sheffield Institute for Translational Neuroscience, University of Sheffield, Sheffield, UK. <sup>12</sup>Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA, USA. <sup>13</sup>Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. <sup>14</sup>School of Engineering, Research Center for Industries of the Future, Westlake University, Hangzhou, China. <sup>15</sup>Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA.

<sup>16</sup>These authors contributed equally: Sai Zhang, Hantao Shu, Jingtian Zhou.  e-mail: [sai.zhang@ufl.edu](mailto:sai.zhang@ufl.edu); [zengjy@westlake.edu.cn](mailto:zengjy@westlake.edu.cn); [ptsao@stanford.edu](mailto:ptsao@stanford.edu); [mpsnyder@stanford.edu](mailto:mpsnyder@stanford.edu)

## Methods

### Single-cell multiome dataset

Single-cell multiome (snRNA-seq + snATAC-seq) data of the human left ventricle and lung were processed and clustered on the basis of RNA modality using Scanpy<sup>137</sup>. The cells with high-quality RNA information (total detected gene > 500, total unique molecular identifiers < 20,000 and mitochondrial read percentage < 10%) were selected for further analysis. Doublets were filtered using scrublet<sup>138</sup> with parameters min\_counts = 1, min\_cells = 10, min\_gene\_variability\_pctl = 90 and n\_prin\_comps = 30. The thresholds for doublet removal were decided per sample on the basis of the distribution of doublet scores in real versus simulated cells. The top 3,000 highly variable genes were selected by combining the results from each sample separately with seurat\_v3 mode. The cell-by-gene count matrices were normalized and scaled. ALLCools with a Python implementation of Seurat integration was used for correction of batch effect between samples with 50 PCs and 30 canonical correlation dimensions<sup>136,139</sup>. Leiden clustering was performed on a  $k$ -nearest neighbor (kNN;  $k = 25$ ) graph. The cell clusters were annotated and merged to cell types by comparing the expression level of predefined marker genes across clusters. The marker genes in Litviňková et al. (2020) and Tucker et al. (2020)<sup>140,141</sup> were used to annotate the heart cell types.

We also examined the ATAC modality of these cells following the methods described below to ensure that these cells also have high-quality open chromatin information. The cells that did not pass ATAC quality controls (QCs) or constituted an ambiguous cluster in ATAC cell embedding were removed, resulting in 10,233 and 10,330 cells retained for downstream analysis for HCM and severe COVID-19, respectively.

### scATAC-seq datasets

The cell type labels for the human pancreas and cortex in the original datasets<sup>33,35</sup> were used. To generate cell embeddings, scATAC-seq data were processed and clustered using snapATAC2 (ref. 142) and ALLCools<sup>136,139</sup>. The fragment files were processed to generate cell-by-bin matrices at 5-kb resolution using snapATAC2 (ref. 142). The cells with 2,000–50,000 total reads and transcription start site (TSS) enrichment > 5 or 7 according to the distribution in specific samples were retained. The cell embeddings were computed with latent semantic indexing (LSI) and batch effects were corrected using the canonical correlation analysis (CCA) LSI mode in ALLCools. Cell-by-peak matrices at 500-bp resolution were generated by calling peaks per cell cluster using snapATAC2. For cortex data, superior and middle temporal gyri and middle frontal gyrus samples were used for AD analysis, resulting in 11,738 cells. For pancreas data, we randomly sampled 10,000 of 64,948 cells covering all annotated cell types for computational acceleration. The single-cell data<sup>121</sup> we used in the replication experiments were processed and QCed similarly.

### Cell–cell similarity network

Following a previous study<sup>46</sup>, we used the mutual kNN (M-kNN) to measure the similarity between two different cells. We first used LSI to extract low-dimensional embeddings for individual cells. For cortex and left-ventricle datasets encompassing multiple samples, batch effects were corrected using both CCA and Harmony<sup>143</sup> and integrated latent embeddings were adopted. Next, we computed the Euclidean distance for pairs of cells using their embeddings and then constructed the kNN graph  $\hat{G} \in \mathbb{R}^{M \times M}$  on the basis of this distance matrix, in which we defined  $\hat{G}_{i,j} = 1(i, j = 1, \dots, M)$  if cell  $j$  is within the top  $k$  closest cells of cell  $i$  and  $\hat{G}_{i,j} = 0$  otherwise. The M-kNN graph  $G$  was then defined as the graph whose edges connect nodes (that is, cells) that are mutually kNNs of each other, which was calculated by  $G = \hat{G} \circ \hat{G}^T$ , where  $\circ$  denotes the element-wise multiplication.

### Target cohorts for T2D and AD

T2D and AD target cohorts were constructed on the basis of the UKBB. All the disease cases were defined according to the ICD-10 (tenth

revision of the International Statistical Classification of Diseases and Related Health Problems) code. In particular, all Caucasian individuals with a disease ICD-10 code in the inpatient record, death record or diagnosis summary record were defined as the disease participants. We used E11.9 and G30.9 for AD and T2D, respectively. This resulted in 1,096 T2D and 932 AD cases. We randomly sampled an equal number of healthy controls by matching sex, age and ancestry information for each case group. In addition, individuals with a similar or related phenotype with the disease (T2D: E10, E11, E12, E13, E14, E23.2, N08.3, N25.1, O24, P70.2, Z13.1, Z83.3 and R73.9; AD: F00, G30, F01, F02, F03 and F05) were excluded from constructing the control group. In this study, overweight individuals (body mass index (BMI)  $\geq 25$ ) were excluded from constructing the T2D cohort. BMI for each individual was defined as the mean of four BMI measurements in the UKBB Data Field 21001.

### Target cohort for HCM

The recruitment of the HCM cohort was part of our California Institute for Regenerative Medicine (CIRM) cardiomyopathy project<sup>27</sup>. The targeted population constituted persons with various cardiac procedures and noncardiac participants with genetic conditions in clinic who were identified to us by their clinical providers. Noncardiac participants were recruited in person during onsite clinic days or over the phone with permission by the providers. Healthy volunteers were recruited from our cardiovascular prevention clinic (that is, persons with no diagnosis of heart disease).

Library preparation and sequencing was performed by Macrogen (first ten samples) and Novogene on genomic DNA we extracted from iPS cells (Qiagen DNeasy kit). Paired-end 150-bp reads were acquired on the Illumina HiSeq X Ten for a minimum of 90 Gb of data. Reads were processed using Sentieon's FASTQ-to-VCF pipeline (Sentieon version 201808.07)<sup>144</sup>. This pipeline is a drop-in replacement for a Burrows–Wheeler aligner (BWA)<sup>145</sup> plus GATK best-practices<sup>146</sup> pipeline for germline single-nucleotide variations (SNVs) and indels but has been highly tuned for optimal computational efficiency. BWA alignment to hg38 was followed by deduplication, realignment, base quality score recalibration and variant calling to generate g.vcf files for each sample. Coverage was assessed (GATK version 3.7)<sup>27</sup>. Individual sample g.vcf files were joined and variant quality score recalibration was performed.

### Target cohort for severe COVID-19

The VA COVID-19 cohort was derived from the VA MVP. The VA MVP is an ongoing national voluntary research program that aims to better understand how genetic, lifestyle and environmental factors influence veteran health<sup>28</sup>. Briefly, individuals aged 18 to over 100 years old have been recruited from over 60 VA medical centers nationwide since 2011 with current enrollment at >800,000. Informed consent is obtained from all participants to provide blood for genomic analysis and access to their full electronic health record data within the VA before and after enrollment. The study received ethical and study protocol approval from the VA central institutional review board (IRB) in accordance with the principles outlined in the Declaration of Helsinki. COVID-19 cases were identified using an algorithm developed by the VA COVID national surveillance tool based on reverse transcription (RT)–qPCR laboratory test results conducted at VA clinics, supplemented with natural language processing on clinical documents for SARS-CoV-2 tests conducted outside of the VA<sup>147</sup>. This resulted in the VA COVID-19 WGS cohort of 2,716 persons with COVID-19 spanning a wide range of ages and ancestries. We defined severe COVID-19 cases as persons who were hospitalized, received acute care, stayed in the intensive care unit or were deceased and controls as those who did not meet these criteria. To minimize potential confounders, we restricted our analysis to nonelderly individuals (age < 65).

DNA isolated from peripheral blood samples was used for WGS. Libraries were prepared using KAPA hyper prep kits, PCR-free according to manufacturers' recommendations. Sequencing was performed using

an Illumina NovaSeq 6000 System (Illumina) with paired-end  $2 \times 150$ -bp read lengths and Illumina's proprietary reversible terminator-based method. The specimens were sequenced to a minimum depth of  $25 \times$  per specimen and an average coverage of  $30 \times$  per plate.

### Independent target cohorts

The GoT2D cohort including 2,874 individuals was used as the independent target cohort for T2D. Samples were sequenced using three technologies: deep whole-exome sequencing, low-pass ( $4 \times$ ) WGS and OMNI 2.5M genotyping. Genotypes (SNVs, indels and structural variants) were called separately for each technology and then integrated by genotype refinement into a single phased reference panel. More details can be found in a previous study<sup>39</sup>.

The HCM independent target cohort was constructed by extracting non-EUR HCM samples (ICD-10: I42.1/I42.2) and a same number of randomly selected non-EUR controls matching age and sex from the UKBB genotype dataset. This resulted in a total of 152 samples.

The WGS data of the independent target cohort for AD were obtained from the ADNI database. A total of 808 whole genomes were downloaded from ADNI, for which we defined individuals with a diagnosis of 'dementia' as cases and 'cognitively normal' as controls.

### WGS data processing

The WGS data for HCM and COVID-19 were processed using the functional equivalence GATK variant-calling pipeline<sup>48</sup>, which was developed by the Broad Institute and plugged into our data and task management system Trellis. The human reference genome build was GRCh38. We used BWA-MEM (version 0.7.15) to align reads, Picard 2.15.0 to mark PCR duplicates and GATK 4.1.0.0 for base quality score recalibration and variant calling using the 'haplotypeCaller' function. We also used FASTQC (version 0.11.4), SAMtools 'flagstat' (version 0.1.19) and RTG Tools 'vcfstats' (version 3.7.1) to assess the qualities of the FASTQ, BAM and gVCF files, respectively. In addition, we used 'verifybamID' in GATK 4.1.0.0 to estimate DNA contamination rates for individual genomes and removed samples with 5% or more contaminated reads.

### QCs of genotype data

We performed stringent QCs for the genotype data following the PRS tutorial ([https://choisingwan.github.io/PRS-Tutorial/](https://choishingwan.github.io/PRS-Tutorial/)). For the GWAS summary statistics data (also referred to as the discovery or base data), genetic variants with low MAF and imputation information score (INFO) were removed. We used thresholds suggested in corresponding original papers: MAF < 0.0001, 0.001 and 0.0001 and INFO < 0.4, 0.6 and 0.6 for T2D, HCM and AD, respectively. We also excluded duplicated and ambiguous variants to guarantee the accuracy of PRS calculation.

For the individual-level genotype data (also referred to as the target data), we carried out both variant-level and individual-level QCs. For WGS data, we performed pre-QCs: we removed samples with kinship > 0.03, sample call rate < 0.97 or mean sample coverage  $\leq 18 \times$ ; genomic positions resided in low-complexity regions or ENCODE-blacklisted regions were removed; we filtered out genotypes in individual samples that were detected with too low or too high read coverages (read depth < 5 or read depth > 1,500); we required all calls to have genotype quality  $\geq 20$  and, for nonreference calls, a sufficient portion ( $>0.9$ ) of reads was required to cover the alternate alleles.

For all target cohorts, we removed variants with INFO < 0.8 (for UKBB-based cohort), missing call rate > 0.01, MAF < 0.01 or Hardy-Weinberg equilibrium  $< 1 \times 10^{-6}$ . For variants with mismatching alleles between discovery and target data, we strand-flipped these alleles to their complementary ones. We further excluded individuals with genotyping rate < 0.01 or with extreme heterozygosity rate (that is, beyond 3 s.d. from the mean). Individuals with an up-to-second-degree relative ( $\pi > 0.125$ ) within the cohort were also removed to prevent bias in prediction evaluation. Lastly, there were 2,176 ( $n = 1,088$

cases,  $n = 1,088$  controls), 134 ( $n = 81$  cases,  $n = 53$  controls), 1,839 ( $n = 919$  cases,  $n = 920$  controls) and 581 ( $n = 120$  cases,  $n = 461$ ) individuals passing the above QCs for T2D, HCM, AD and severe COVID-19 cohorts, respectively.

All independent target cohorts were processed and QCed using the same pipeline. After sample-level QCs, the final cohorts consisted of 2,749 samples (1,398 cases and 1,351 controls) for GoT2D, 62 samples (23 cases and 39 controls) for non-EUR UKBB and 469 samples (251 cases and 218 controls) for ADNI.

### PC analysis for genotype data

To characterize the population structure of target cohorts, PC analysis was performed after pruning (window size = 200 variants, sliding step size = 50 variants, LD  $r^2$  threshold = 0.25). The first ten PCs were retained as covariates in the downstream analysis.

### PLINK C+T PRS calculation

The cell-level C+T PRS was computed using PLINK, which is given by

$$\text{PRS}_j = \frac{\sum_{i \in \text{cCRE}_j} \beta_i \times G_i}{P \times M},$$

where  $\text{cCRE}_j$  denotes cCREs within cell  $j$ ,  $\beta_i$  is the effect size of variant  $i$ ,  $G_i$  represents the number of effect alleles,  $P$  is the ploidy of the sample (2 for human) and  $M$  is the number of nonmissing variants. In the clumping phase, all index variants were forced to be drawn from the variants located within scATAC-seq peaks of individual cells using the '--clump-index-first' option. Variants within 250 kb of the index variant and three LD thresholds ( $r^2 = 0.1, 0.3$  and  $0.5$ ) were considered for clumping. After constructing the index variant set, we applied multiple  $P$ -value thresholds ( $P = 1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 0.01, 0.05, 0.1$  and  $0.5$ ) to compute PRSs, resulting in 21 PRSs calculated for each cell and each individual. We used the 1,000 Genomes Project samples to estimate the LD (out-sample estimation) for the simulation, HCM and severe COVID-19 cohorts because of their limited sample sizes, while using the target data (in-sample estimation) for other cohorts.

The standard C+T PRS was calculated using the same set of parameters as that used in computing cell-level PRS, except that all variants were considered without conditioning. The  $P$ -value and LD  $r^2$  thresholds were regarded as hyperparameters to be optimized in model selection.

### Model details of scPRS

The cell-level PRS matrix  $X_n \in \mathbb{R}^{M \times 21}$  ( $n = 1, \dots, N$ ) presents single-cell-resolved genetic risk features for each individual and it is input into the scPRS model to predict the disease risk. Here,  $N$  and  $M$  denote the numbers of individuals and cells, respectively.

scPRS consists of three modules (Fig. 1): the feature-embedding module, the graph convolutional network module and the readout module. The feature-embedding module takes normalized cell-level PRS  $X_n$  as the input and uses a one-layer perceptron to reweight and integrate 21 PRS features per cell:

$$h_n^{(0)} = X_n \bullet \text{abs}(W_0),$$

where  $W_0$  denotes learnable model parameters,  $\text{abs}$  represents the absolute function and  $h_n^{(0)} \in \mathbb{R}^M$  represents the integrated features of  $M$  cells for individual  $n$ . According to the definition of PRS, larger values in  $X_n$  indicate higher disease risk. To maintain this interpretability throughout the modeling, we adopt the absolute function  $\text{abs}$  to enforce nonnegativity for  $W_0$ .

We next seek to integrate PRS features across different cells to generate a final risk score. With the consideration of the dropout event and sparsity of scATAC-seq data and assuming that cells with similar low-dimensional embeddings should have comparable epigenomes

and then similar genetic signals, we use a GNN<sup>149</sup> to smooth and denoise single-cell-level PRS features. More specifically, on the basis of the pre-computed M-kNN graph  $G$ , the GNN module is defined as

$$g_v^{(t+1)} = \frac{1}{\deg(v)} \sum_{u \in \mathcal{N}(v)} (\text{abs}(w_1^{(t)}) h_u^{(t)} + \text{abs}(w_2^{(t)}) h_v^{(t)}),$$

$$h_v^{(t+1)} = \text{leaky ReLU}(g_v^{(t+1)}),$$

where  $h_v^{(t)}$  denotes the hidden feature of cell  $v$  at layer  $t$ ,  $w_1^{(t)}$  and  $w_2^{(t)}$  are learnable parameters of layer  $t$ ,  $\deg$  denotes the degree of each node or cell and  $\mathcal{N}(v)$  represents the neighbors of cell  $v$  in the M-kNN graph  $G$ . The leaky ReLU activation function is defined as

$$\text{Leaky ReLU}(x) = \max(\alpha \times x, x),$$

where  $\alpha = 0.1$  is used in this study. Note that the absolute function is also adopted to induce nonnegativity to model weights.

Lastly, we design a readout module to map GNN-smoothed hidden features to the phenotype leveraging a one-layer perceptron:

$$y = \sigma(\beta \cdot h^{(T)} + b),$$

where  $\beta \in \mathbb{R}^M$  represents the learnable regression coefficients indicating cell importance to prediction,  $T$  is the number of total layers in GNN,  $b$  is the bias term and  $\sigma$  is the sigmoid function for binary classification and the identity function for regression.

### Optimization of scPRS

To train scPRS for disease prediction, we adopt the binary cross-entropy (BCE) loss and additional regularization functions for enhancing predictive power and model interpretability. The loss function  $\mathcal{L}$  of scPRS is defined as

$$\mathcal{L} = \frac{1}{N} \sum_n (y_n \log(p_n) + (1 - y_n) \log(1 - p_n)) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 + \lambda_3 \beta^T G_L \beta,$$

where  $y_n \in \{0, 1\}$  is the true disease label for individual  $n$ ,  $p_n \in [0, 1]$  is the scPRS-predicted disease probability and  $\|\cdot\|_1$  and  $\|\cdot\|_2$  represent  $L_1$  and  $L_2$  norms, respectively. We also add a Laplacian regularization term based on the symmetric normalized Laplacian matrix  $G_L$ , which is defined as

$$G_L = D^{\frac{1}{2}} (D - A) D^{-\frac{1}{2}},$$

where  $D$  and  $A$  denote the degree and adjacency matrices of the cell-cell similarity graph  $G$ , respectively. We use hyperparameters  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  to balance across different regularization terms.

scPRS was trained by minimizing the loss  $\mathcal{L}$  using the Adam algorithm<sup>150</sup> with a learning rate of  $1 \times 10^{-3}$  and batch size of 32. We trained scPRS for 200 epochs. Multiple sets of hyperparameters were considered in model selection, including  $T \in \{0, 1, 2\}$ ,  $\lambda_1 \in \{0, 1, 10\}$ ,  $\lambda_2 \in \{1, 10, 50, 100, 250, 500, 750\}$ ,  $\lambda_3 \in \{0.01, 0.1, 0.5, 1, 2.5, 5, 10, 50, 100\}$  and M-kNN neighbor number  $k \in \{25, 50\}$ . We also selected between CCA-based and Harmony-based cell-cell similarity networks for T2D and AD.

In prediction evaluation, we randomly partitioned the dataset into training, validation and testing sets comprising 60%, 20% and 20% of samples, respectively. We trained different scPRS models with all possible combinations of hyperparameters and assessed their performance (measured by AUROC) on the validation dataset. We selected the model yielding the best performance on the validation set and reported its performance on the held-out test set. This process was repeated ten times with different random seeds to assess the robustness of the

model. Predictive performance was evaluated using both the AUROC and the AUPRC.

In cell prioritization, we conducted fivefold cross-validation, which was repeated five times. The best hyperparameter set was then selected on the basis of the average AUROC score. The final model was trained with this optimal hyperparameter set on the entire dataset. To examine the variability of cell weights learned from model training, we trained 100 models using different random seeds.

For the regression task, the mean squared error was used as the loss function instead of BCE. The model performance was evaluated based on the Pearson correlation between true and predicted values.

### Calculation of nonpeak and peak PRS

Similar to the cell-level PRS, the calculation of nonpeak PRS was based on PLINK C+T, using only variants outside of scATAC-seq peaks as the index variants. A total of 21 nonpeak PRSs were computed and integrated in scPRS+, corresponding to different combinations of C+T parameters:  $P \in \{1 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3}, 0.01, 0.05, 0.1, 0.5\}$  and  $r^2 \in \{0.1, 0.3, 0.5\}$ . For scPRS+ (integrating cell-level PRSs and nonpeak PRSs) and scPRS+covar (integrating cell-level PRSs, nonpeak PRSs, age, sex and ten PCs), we concatenated additional features to latent cell features  $h^{(T)}$  at the final GNN layer.

In calculating the single-cell-type peak PRS, only variants located within cell-type peaks were used to select the index variants, where the same 21 combinations of C+T parameters were adopted. A multi-cell-type PRS was further built by combining all single-cell-type PRSs ( $n = 21 \times n_{\text{celltype}}$ ) using LR. LR was trained on the training data and the performance was reported on the testing data.

### Implementation details on LDpred2, Lassosum and PolyPred

We implemented LDpred2 and Lassosum following the bigsnpr tutorial (<https://privetl.github.io/bigsnpr/articles/LDpred2.html>). Three LDpred2 models were implemented—the infinitesimal model (LDpred2-inf), grid model (LDpred-inf) and auto model (LDpred2-auto). All model hyperparameters were selected on the basis of recommendations provided in the tutorial. To ensure a fair comparison, we maintained the same dataset splits (that is, training, validation and test sets) as those used in scPRS. For PLINK C+T, LDpred2-grid and Lassosum, the best model hyperparameters were determined on the basis of predictive performance on the validation dataset.

For a fair comparison, we used scATAC-seq peaks as the functional annotation for variants in PolyPred and adopted the same GWASs as those used in scPRS to compute prior causal probabilities<sup>38</sup>. We implemented PolyPhred following the manual provided by the authors (<https://github.com/omerwe/polyfun/wiki>).

Unlike C+T, more advanced PRS methods, including LDpred2, Lassosum and PolyPred, inherently optimize  $r^2$  and  $P$ -value cutoffs to select an optimal set of variants for PRS computation. This flexibility in optimization is a key innovation of these approaches.

### Benchmark on independent target cohorts

Because the original GWAS discovery cohorts for T2D and AD overlapped with GoT2D and ADNI, respectively, to prevent information leakage, we adopted the UKBB GWAS<sup>151</sup> as new summary statistics for T2D and AD, which were independent from the new target cohorts. We then trained new scPRS models on the basis of original target cohorts. For C+T, LDpred2-grid and Lassosum, model hyperparameters were optimized on the basis of original target cohorts. For scPRS, hyperparameters were selected using fivefold cross-validation of the original target cohorts. All PRS approaches were tested on the basis of new independent target cohorts.

### Prioritization of disease-relevant cells and cell types using scPRS

The mapping from input PRS features  $X$  to latent cell features  $h^{(T)}$  monotonically increases as a result of the design principle of scPRS,

where weights in the embedding and GNN modules are constrained to be nonnegative. This facilitates model interpretation: a larger value of  $\beta_m$  denotes a higher enrichment of genetic risk within that cell, thereby informing disease–cell relevance. To account for the variability of learned cell weights, we trained 100 scPRS models and compared the distribution of  $\beta_m$  for individual cells with that of top-ranking weights (that is, the top 15% of all cell weights per repeat) using a one-sided *t*-test. This comparison was conducted for each cell in the dataset. We defined disease-relevant cells as those cells whose adjusted *P* values (using the Benjamini–Yekutieli procedure) were less than 0.1. Roughly speaking, scPRS prioritizes cells whose weights are consistently larger than those of the majority of cells.

To get more biological insights, we examined the enrichment of scPRS-prioritized cells within each cell type using a Fisher's exact test. The disease-relevant cell types were defined as those cell types whose adjusted *P* values (using the BH procedure) were less than 0.1.

### Simulation details

Using the PBMC multiome data downloaded from 10x Genomics, we first conducted the differential accessibility analysis to identify monocyte-specific scATAC-seq peaks. In this study, we defined monocytes as the total set of CD14/CD16 monocytes and dendritic cells considering their shared heritability<sup>152</sup>. We identified differentially accessible regions (DARs) within monocytes using the top 1,500 marker peaks per cell subtype. Next, leveraging a monocyte count GWAS<sup>22</sup>, we computed PLINK C+T PRS conditioned on the variants located within monocyte DARs for a WGS cohort<sup>23</sup> ( $n = 401$ ). Raw C+T PRS outputs were further standardized to mean = 0 and variance = 1, yielding the 'ground truth' of monocyte count for this cohort.

To introduce randomness, we added a noise term to the simulated monocyte count:

$$\tilde{y} = y + \varepsilon,$$

where  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ . In this study, we used  $\sigma \in \{0, 0.25, 0.5, 1, 3, 5, 7\}$ . We trained scPRS on the basis of these simulation datasets with and without noises to evaluate its capacity in identifying phenotype-associated cells.

### SCAVENGE

We used SCAVENGE<sup>46</sup> as a benchmark for prioritizing disease-relevant cells. Following the SCAVENGE tutorial (<https://sankaranlab.github.io/SCAVENGE/articles/SCAVENGE>), we calculated trait relevance scores (TRSs) for individual cells, indicative of their association with the disease. Cells were prioritized by SCAVENGE if their TRSs were above 95% of all TRSs. As in the scPRS analysis, we evaluated the enrichment of selected cells within each cell type using the Fisher's exact test.

### Stratified LDSC

Partitioned heritability analysis was carried out using sLDSC as previously described<sup>45</sup>. Heritability was quantified within the total set of snATAC-seq peaks identified for each of the left-ventricle cell types. Genetic enrichment for a particular cell type was defined by calculating the captured heritability per unit of sequence within the total set of identified snATAC-seq peaks for that cell type, compared to the genome overall. *P* values were calculated as previously described<sup>45</sup>; nominal significance ( $P < 0.05$ ) was taken to be indicative of true enrichment.

We conducted sLDSC using the same GWAS and scATAC-seq datasets as those used in scPRS for HCM and severe COVID-19, for which no existing sLDSC results were available. For AD, the original sLDSC<sup>35</sup> was performed on the same GWAS and scATAC-seq dataset. For T2D, the original sLDSC<sup>44</sup> was carried out on the same scATAC-seq dataset but used a larger GWAS<sup>153</sup>. We chose to report the results of sLDSC applied to discovery GWAS to optimize its power, given the larger sample size of discovery GWAS compared to target cohort.

### Identification of disease-relevant cCREs

As the first step of the layered multiomic analysis (Fig. 5a), we identified differentially accessible cCREs within each scPRS-prioritized cell type using Signac<sup>154</sup>. Specifically, we used the FindMarker function to compare peaks within scPRS-prioritized cells (per cell type) against all unselected cells in the dataset as background, with parameters test, use = 'LR', latent.vars = 'peak\_region\_fragments', min.pct = 0.02, and logfc.threshold = 0.1. Significant peaks (adjusted *P* < 0.1 based on BH correction) with a positive log<sub>2</sub>FC were defined as differentially accessible cCREs. Next, leveraging the discovery GWAS summary statistics, we conducted MAGMA<sup>74</sup> analysis for these differentially accessible cCREs per cell type, with gene-model = 'multi'. MAGMA is a widely used tool for gene-level and region-level genetic association analysis based on GWAS summary data. It is designed to test genetic associations of predefined genes or regions with diseases or traits by aggregating variant-level GWAS statistics while accounting for LD. We defined disease-relevant cCREs (T2D-cCREs and AD-cCREs) as those cCREs with adjusted MAGMA *P* < 0.1 based on BH correction. We expanded our analysis to involve all nominally significant cCREs (MAGMA *P* < 0.05) for HCM, as no cCRE passed the multiple-testing correction.

### Mapping cCRE–gene links

We mapped cCREs to their target genes on the basis of two complementary strategies. First, we adopted the closest-gene strategy<sup>155</sup> and assigned each cCRE to its closest gene. In addition, we added more distant genes on the basis of a coaccessibility analysis using Cicero<sup>75</sup> and linked each cCRE to those genes whose TSS peak displayed coaccessibility with the cCRE above 80% of all interactions. For each scPRS-prioritized cell type, the expressed genes mapped to disease-relevant cCREs within that cell type defined the repertoire of disease candidate genes.

### Enrichment of disease-associated variants within scPRS-cell-specific peaks

Per disease-relevant cell type, we performed clumping within differentially accessible peaks in scPRS-prioritized cells to remove redundant variants. Multiple LD *r*<sup>2</sup> thresholds ( $r^2 = 0.1, 0.3$  and  $0.5$ ) were tested. Leveraging the clumped variant set, we examined the enrichment of disease-associated variants (GWAS *P* <  $5 \times 10^{-8}$ ) within scPRS-cell-specific peaks by comparing it to the genome-wide distribution.

### TF-binding motif analysis

The TF-binding motif analysis was performed using GimmeMotifs<sup>156</sup>. The differential motifs between disease-relevant cCREs and all peaks within the corresponding cell type were identified using the 'gimme motif' command with options f = 0.5 and s = 0. AUROC was adopted to quantify the motif enrichment.

### Network analysis

We downloaded the human PPIs from STRING (version 12.0)<sup>98</sup>, comprising 19,622 proteins and 6,857,702 interactions. High-confidence PPIs (combined score > 700) were extracted for downstream analysis, including 16,185 proteins and 236,000 interactions. To mitigate bias from hub proteins<sup>157</sup>, we applied the random walk with restart algorithm with a restart probability of 0.5. This produced a smoothed network after retaining the top 5% predicted edges ( $n = 6,243,766$ ). Next, we used the Louvain method<sup>158</sup> to decompose the network into different modules. Following algorithm convergence, we obtained 1,261 modules with an average size of 13 nodes.

The enrichment of genes of interest within each module was tested using the hypergeometric test. Modules with adjusted *P* < 0.1 based on BH correction were considered significant.

### Sequence deep learning model design and training

The sequence-based deep learning model was trained to predict ATAC-seq peaks across various cell types on the basis of the DNA

sequence. Specifically, the sequence model takes a 2,000-bp DNA sequence as the input and outputs the peak status of the centered 200 bp for different cell types. The peak label for a specific cell type is 1 if over 50% of the centered 200 bp is overlapped by an ATAC-seq peak within that cell type and 0 otherwise. Model structure follows the Beluga architecture<sup>78</sup>, except its outputs correspond to different cell types within the tissue of interest.

ATAC-seq peaks within chromosomes 6 and 7 and chromosomes 8 and 9 were held out as validation and test data, respectively. Peaks in other chromosomes were used as training data. Genomic regions annotated by the ENCODE blacklist<sup>159</sup> were excluded from analysis. We adopted the BCE loss as the objective function. The sequence model was trained using the stochastic gradient descent algorithm with a weight decay coefficient of  $1 \times 10^{-6}$ , momentum of 0.9, learning rate of 0.08 and batch size of 64. The model was implemented using Selene<sup>79</sup>, a PyTorch-based library for sequence deep learning modeling. In this study, we trained separate sequence models using different scATAC-seq datasets.

### Prediction of variant effects using sequence deep learning model

We used the sequence model to predict the impact of genetic variants on cCREs across diverse cell types. For a given cell type  $c$  and variant  $v$  (from reference allele to alternative allele), the model predicts the status of cCRE  $y_{ref,c}$  and  $y_{alt,c}$  for sequences centered on the reference and alternative alleles, respectively. We define the functional effect of variant  $v$  in cell type  $c$  as  $y_{v,c} = y_{alt,c} - y_{ref,c}$ , representing how the variant alters cCRE in this cell type. To achieve a global evaluation of functional scores, we introduce the  $Z_{v,c}$  score, which normalizes  $y_{v,c}$  as  $Z_{v,c} = (y_{v,c} - \mu)/\sigma$ , where  $\mu$  and  $\sigma$  denote the mean and s.d. of all variant scores, respectively. The  $Q_{v,c}$  score is further defined as the quantile of  $|Z_{v,c}|$  among all variants. A higher  $Q$  score indicates a larger functional effect within a specific cell type.

### Benchmarking sequence model prediction

To benchmark the sequence model prediction on variant effects against QTL analysis (eQTL or caQTL), we compared the absolute  $Z$  scores computed by the sequence model between QTLs and non-QTLs using a two-sided  $t$ -test. The  $t$  statistics was used to measure the enrichment of functional variants defined by the sequence model within QTLs.

As the second benchmarking, we used SNP2TFBS<sup>89</sup> to predict the effects of variants on altering TFBS affinity. The binding affinities for different TFs were averaged for each studied variant to estimate its overall effect. Given a particular quantile cutoff, variants were split into two groups according to their  $Q$  scores. We then compared the averaged SNP2TFBS scores between these two groups of variants using a two-sided  $t$ -test. We report the  $t$  statistic, which indicates the enrichment of TFBS-disrupting variants within sequence-model-defined functional variants.

### Variant effect within disease-relevant cCREs

We compared the abundance of functional disease-associated variants (GWAS  $P < 0.05$ ) within disease-relevant cCREs against the background using a Fisher's exact test. Similarly, the functional variants were defined as those with  $Q$  scores above a given cutoff (multiple cutoffs applied). The odds ratio (OR) was adopted to measure the enrichment of functional variants within disease cCREs.

### Fine-mapping disease risk variants

We used three approaches to fine-map disease risk variants: the sequence deep learning model, QTL and TFBS. A 0.8 quantile cutoff was adopted to define functional variants on the basis of the sequence model in fine-mapping. In addition to SNP2TFBS, motifbreakR<sup>91</sup> was used to predict variant disruption on TF binding. A positive averaged SNP2TFBS score or a strong-effect motifbreakR score was used to define a disrupting variant. We excluded missense and loss-of-function variants and variants with GWAS  $P \geq 0.05$  from fine-mapping.

### iPS cell reprogramming

iPS cells were reprogrammed from PBMCs using Sendai virus (CytoTune iPS 2.0 Sendai Reprogramming Kit) as previously described<sup>160</sup>. Three clones were generated per subject, karyotyped (KaryoStat, Thermo Fisher Scientific), determined to be free of *Mycoplasma* and evaluated by immunohistochemistry for expression of pluripotency markers TRA-1-60 (LifeTech, MA1023) and SSEA4 (LifeTech, MA1021). Cells were maintained under feed-free conditions in mTeSR (StemCell Technologies, 5850) or Essential 8 medium (Fisher, A1517001) and stored in liquid nitrogen.

### CDM differentiation and drug treatment

As previously described<sup>161</sup>, iPS cells were plated on Matrigel and cultured in StemMACS iPS-Brew XF (MACS Miltenyi Biotec, 130-104-368) until the final passage in Essential 8 medium (Fisher, A1517001). CDM differentiation was induced at 60–80% confluence, with culture in RPMI medium (Gibco/LifeTech, 11875-119) plus B27 supplement lacking insulin (Gibco/LifeTech, A1895601). Then, 6  $\mu$ M CHIR-99021 (Fisher, NC0976209) was added on day 0 and 6  $\mu$ M IWR1 (Fisher, NC1319406) was added on day 3. Beginning on day 7, the medium was changed every other day using RPMI medium supplemented with B27 containing insulin (Gibco/LifeTech 17504-044). Upon commencement of beating (around day 15), cells underwent purification by a 3-day glucose starvation (RPMI medium without glucose (Gibco/LifeTech, 11879-020) supplemented with insulin-containing B27), a 1-day recovery in glucose-containing medium and subsequent replating (dissociated in TrypLE, Fisher, 50-591-353). Cells were then maintained in RPMI medium supplemented with insulin-containing B27 until approximately day 30. After differentiation, drug treatment occurred at 0 and 24 h and samples were assayed at 48 h. Cells were treated with 250 nM MYK-461 (Cayman Chemical, 19216-5mg), 400 nM or 1  $\mu$ M omecamtiv mecarbil (Selleckchem, Fisher, NC1069600) or DMSO.

### RNA-seq library preparation, sequencing, QC and expression matrix generation

RNA was extracted from iPS cells or CDMs (RNeasy, Qiagen). Illumina RNA-seq libraries (TruSeq Stranded Total RNA LP Gold) were prepared on the Bravo (Agilent), pooled and sequenced (NovaSeq 6000, paired-end, 100 bp)<sup>27</sup>. Where possible, drug treatment conditions for the same differentiation were kept together in batches, while replicate differentiations for the same iPS cell lines were split apart and CDM and control samples were distributed across batches. Reads were aligned to hg38 (STAR). PC analysis on CDM and iPS samples separately returned no outlier samples (defined as  $Z$  score of PC1  $> 3$ ). Library QC was assessed using fastp, fastQC, STAR and Picard metrics. Samples were flagged for poor QC by the following metrics: G+C content after filtering outside of 20–80% (fastp), duplication rate greater than 40% (fastp), uniquely mapped read pairs (fragments)  $< 20$  million (STAR), mean reads (average of forward and reverse)  $< 20$  million (fastQC), ribosomal RNA bases  $> 20\%$  (Picard), coding plus UTR (untranslated region)  $< 50\%$  (Picard) and uniquely mapping fragments  $< 60\%$  (STAR). Samples with more than one flag were removed. CDM and iPS cell samples were subsequently processed separately. Reads were computed as counts per million (edgeR), corrected for library preparation batch (combat-seq) and normalized by the trimmed mean of  $M$  values (TMM; edgeR) to generate the final expression matrix. For samples with biological replicates, TMM counts were averaged. PC analysis was performed and PC1 was assessed for Spearman correlation with the following metadata: percent G+C content (fastp), mean reads (average of forward and reverse) in millions (fastQC), percent ribosomal RNA bases (Picard), uniquely mapped fragments in millions (STAR), duplication rate (fastp), percent coding or UTR (Picard), library preparation batch and sequencing pool. The maximum absolute value for spearman correlation between PC1 and the library metadata was 0.08 for CDM

samples, indicating good QC with technical artifacts having minimal influence on the dataset. iPS cell samples had higher correlation for three metrics (0.26 with G+C content, 0.22 with duplication rate and 0.11 with percent coding or UTR), with the remaining having less than an absolute value of 0.04.

### Differential expression analysis

Raw data were input into DESeq2 (ref. 162) as required to compare gene expression between HCM cases and controls across different conditions. Gene counts were averaged across replicates. Sample sex and ancestry were included as covariates in the analysis.

### Allelic imbalance analysis in rs7922621 prime-edited microglia

The **rs7922621** prime-edited WTC11 clones were obtained from our previous study<sup>108</sup> and microglia were differentiated accordingly. Total RNA was isolated from wild-type and prime-edited microglia using the RNeasy plus mini kit (Qiagen, 74034). Briefly, 400 ng of total RNA was reverse-transcribed using the iScript complementary DNA (cDNA) synthesis kit (Bio-Rad, 1708891). The cDNA region containing phased heterozygous SNP of *ANXA11* (**rs2573353** in WTC11)<sup>163</sup> was amplified using the following primers: WTC-ANX-F, AGGTCCAATAATCCCTGCTGA; WTC-ANX-R, CCATGGTGCTGGCTAATT. The PCR products were purified by agarose gel extraction, followed by the addition of Illumina adaptors and deep sequencing. Reads were aligned to the sequence of either allele and counted if the 100-bp regions surrounding **rs2573353** were exactly matched.

### Differentiation of TMEM119-Tdtomato reporter cell line iMGs

iPS cells stably expressing a TMEM119-tdTomato reporter transgene were first differentiated into fibroblast-like cells using a previously established method<sup>110,111</sup>. TMEM119-tdTomato fibroblasts were seeded onto 96-well plates (Corning) coated with 0.1% gelatin and Matrigel in fibroblast medium (DMEM with 10% FBS and 1% penicillin-streptomycin). After 48 h, the cells were transduced with 200  $\mu$ l of two different concentrated retroviruses to overexpress the human *PU.1* and *CEBPA* per 96-well well with 5  $\mu$ g ml<sup>-1</sup> polybrene in fibroblast medium. Then, 24 h after transduction, the medium was switched to DMEM with 5% FBS, 10 ng ml<sup>-1</sup> human macrophage colony-stimulating factor (M-CSF) and 10 ng ml<sup>-1</sup> interleukin 34 (IL-34) and refreshed every 3 days thereafter. iMGs expressing the TMEM119-tdTomato reporter were used for experiments 14 days after viral transduction.

### siRNA transfection

siRNAs (Thermo Fisher) at a concentration of 30 nM were transfected into iMGs on day 14 using Lipofectamine RNAiMAX transfection reagent (Thermo Fisher Scientific, 13778075) in complete iMG medium (DMEM + 5% FBS, 10 ng ml<sup>-1</sup> M-CSF and 10 ng ml<sup>-1</sup> IL-34). After 24 h, the medium was refreshed with complete iMG medium; after an additional 24 h (48 h after transfection), cell cultures were collected for RT-qPCR or pHrodo analysis.

### pHrodo phagocytosis assay

iMGs cultured in 96-well plates (Corning) coated with Matrigel and gelatin were incubated with 10  $\mu$ g of pHrodo green *Escherichia coli* bioparticles (Incyte) for 15 min at 37 °C. Wells were then washed with PBS and were longitudinally imaged with Molecular Devices ImageExpress at 30-min intervals for the initial 2 h and 1-h intervals thereafter up to 24 h after the start. The 2-h time point was selected for downstream analysis. ImageJ software was used for quantification of individual replicates across conditions on the basis of the colocalization of TMEM119-Tdtomato and pHrodo green.

### Reporting summary

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

### Data availability

The PBMC multiome dataset is available from 10x Genomics ([https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc\\_granulocyte\\_sorted\\_10k](https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k)). The single-cell multiome data (snRNA-seq and snATAC-seq coassay) of the human left ventricle and lung are publicly accessible through ENCODE 4 ([https://www.encodeproject.org/single-cell/?type=Experiment&assay\\_slims=SingleCell&status=released](https://www.encodeproject.org/single-cell/?type=Experiment&assay_slims=SingleCell&status=released)). All other scATAC-seq datasets were obtained from their original publications<sup>33,35</sup>. The WGS data used in simulation are available from a previous study<sup>23</sup>. Individual-level genotype–phenotype data for T2D and AD were sourced from the UKBB. The WGS and iPS cell RNA-seq data for HCM are available from a previous study<sup>27</sup>. The COVID-19 WGS and clinical data are available upon request from the corresponding authors (P.S.T. and M.P.S.); these data are not publicly available because of US Government and Department of VA restrictions relating to participant privacy and consent. The independent target cohorts for T2D, HCM and AD are accessible through the European Genome-Phenome Archive ([EGAD00001002247](https://egad00001002247)), UKBB and ADNI (<https://adni.loni.usc.edu/data-samples/adni-data/>), respectively. The HCM snRNA-seq dataset was obtained from a previous study<sup>101</sup>. All GWAS summary statistics data were acquired from their original publications<sup>22,29–31</sup>. The GTEx and islet eQTL datasets were downloaded from the eQTL catalog (<https://www.ebi.ac.uk/eqtl/>). Other eQTL and caQTL datasets were obtained from their original publications<sup>60,85,88,164</sup>. The reference human genomes (hg19 and hg38) are available online (<https://hgdownload.soe.ucsc.edu/downloads.html#human>).

### Code availability

The source code and tutorial for scPRS are available from GitHub (<https://github.com/szhang112/scPRS>).

### References

137. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
138. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell Syst* **8**, 281–291 (2019).
139. Liu, H. et al. Single-cell DNA methylome and 3D multi-omic atlas of the adult mouse brain. *Nature* **624**, 366–377 (2023).
140. Litvánková, M. et al. Cells of the adult human heart. *Nature* **588**, 466–472 (2020).
141. Tucker, N. R. et al. Transcriptional and cellular diversity of the human heart. *Circulation* **142**, 466–482 (2020).
142. Zhang, K., Zemke, N. R., Armand, E. J. & Ren, B. A fast, scalable and versatile tool for analysis of single-cell omics data. *Nat. Methods* **21**, 217–227 (2024).
143. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
144. Kendig, K. I. et al. Sentieon DNaseq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.* **10**, 736 (2019).
145. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
146. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
147. Song, R. J. et al. Phenome-wide association of 1809 phenotypes and COVID-19 disease progression in the Veterans Health Administration Million Veteran Program. *PLoS ONE* **16**, e0251651 (2021).
148. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).

149. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. Preprint at <https://doi.org/10.48550/arXiv.1704.01212> (2017).
150. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at <https://doi.org/10.48550/arXiv.1412.6980> (2014).
151. Karczewski, K. J. et al. Pan-UK Biobank GWAS improves discovery, analysis of genetic architecture, and resolution into ancestry-enriched effects. Preprint at *medRxiv* <https://doi.org/10.1101/2024.03.13.24303864> (2024).
152. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
153. 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).
154. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
155. Fulco, C. P. et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
156. van Heeringen, S. J. & Veenstra, G. J. C. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
157. Krishnan, A. et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci.* **19**, 1454–1462 (2016).
158. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
159. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
160. Gruber, J. J. et al. Chromatin remodeling in response to BRCA2-Crisis. *Cell Rep.* **28**, 2182–2193 (2019).
161. Burridge, P. W. et al. Chemically defined generation of human cardiomyocytes. *Nat. Methods* **11**, 855–860 (2014).
162. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
163. Song, M. et al. Mapping *cis*-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes. *Nat. Genet.* **51**, 1252–1262 (2019).
164. Bryois, J. et al. Cell-type-specific *cis*-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).

## Acknowledgements

We thank R. Corces for providing fragment files for the human brain scATAC-seq data and for his detailed responses to our questions regarding this dataset. This work was supported by National Institutes of Health (NIH; CEGS 5P50HG007735 to M.P.S.; R01AG079291 and RF1AG079557 to Y.S.; 2R01NS097850 and 1R01NS131409 to J.K.I.) and MVP, Office of Research and Development, Veterans Health Administration (MVP001 to P.S.T.). This publication does not represent the views of the Department of Veteran Affairs or the United States Government. This work was also supported by the Tau Consortium and John Douglas French Alzheimer's Foundation (to J.K.I.). J.K.I. is the John Douglas French Alzheimer's Foundation endowed associate professor of stem cell biology and regenerative medicine. We thank Y.-H. Huang for making iMGs in the siRNA experiments. We thank the

UKBB participants for sharing their data, which were accessed under application no. 41751. This research complies with all appropriate ethical regulations. Ethics approval for the UKBB study was obtained from the North West Center for Research Ethics Committee (protocol no. 11/NW/0382). The Stanford CIRM cardiomyopathy project is in compliance with the Stanford Human Research Protection Program guidelines and approved by the Stanford IRB (IRB 30064). In addition, the procedures are in compliance with the International Society of Stem Cell Research guidelines and approved by the Stanford IRB/Stem Cell Research Oversight panel (SCRO 656). This study makes use of data generated by the GoT2D and T2D-GENES consortia. A full list of the investigators who contributed to the generation of data is available online (<https://kp4cd.org/got2d> and <https://kp4cd.org/t2dgenes>). Funding for data generation was provided by National Human Genome Research Institute, National Institute of Diabetes and Digestive and Kidney Diseases, the Wellcome Trust and Helmholtz Zentrum München. A full list of award numbers is available in the primary publication<sup>39</sup>. Data collection and sharing for the ADNI is funded by the National Institute on Aging (NIH U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research and private sector contributions through the Foundation for the NIH including generous contributions from the following: AbbVie, Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Araclon Biotech, BioClinica, Biogen, Bristol-Myers Squibb, CereSpir, Cogstate, Eisai, Elan Pharmaceuticals, Eli Lilly and Company, Euroimmun, F. Hoffmann-La Roche and its affiliated company Genentech, Fujirebio, GE Healthcare, IXICO, Janssen Alzheimer Immunotherapy Research and Development, Johnson & Johnson Pharmaceutical Research and Development, Lumosity, Lundbeck, Merck & Co., Meso Scale Diagnostics, NeuroRx Research, Neurotrack Technologies, Novartis Pharmaceuticals, Pfizer, Piramal Imaging, Servier, Takeda Pharmaceutical and Transition Therapeutics.

## Author contributions

S.Z. conceptualized and designed the study. S.Z. and H.S. developed and implemented scPRS. X.Y. and Y.L. (supervised by Y.S.) conducted prime-editing experiments. J.R.-S. and S.T. (supervised by J.K.I.) performed siRNA and pHrodo phagocytosis experiments. S.Z., H.S., J. Zhou analyzed the data with assistance from J.R.-S., X.Y., Y.L., J.C.-K., E.M. and C.Z. All authors were responsible for data interpretation. S.Z., J. Zeng, P.S.T. and M.P.S. supervised the project. S.Z., H.S., J. Zhou and M.P.S. drafted the manuscript with assistance from other authors. All authors meet the four International Committee of Medical Journal Editors authorship criteria and were responsible for approving the final version for publication and for accuracy and integrity of the work.

## Competing interests

M.P.S. is a cofounder and the scientific advisory board member of Personalis, SensOmics, Qbio, January AI, Fodsel, Filtricine, Protos, RTHM, Iollo, Marble Therapeutics, Crosshair Therapeutics, NextThought and Mirvie. He is a scientific advisor of Jupiter, Neu vivo, Swaza, Mitrix, Yuvan, TranscribeGlass and Applied Cognition. J.K.I. is a cofounder and a scientific advisory board member of AcuraStem and Modulo Bio and a scientific advisory board member of Synapticure and Vessalias Therapeutics. J.K.I. is also an employee of BioMarin Pharmaceutical. The remaining authors declare no competing interests.

## Additional information

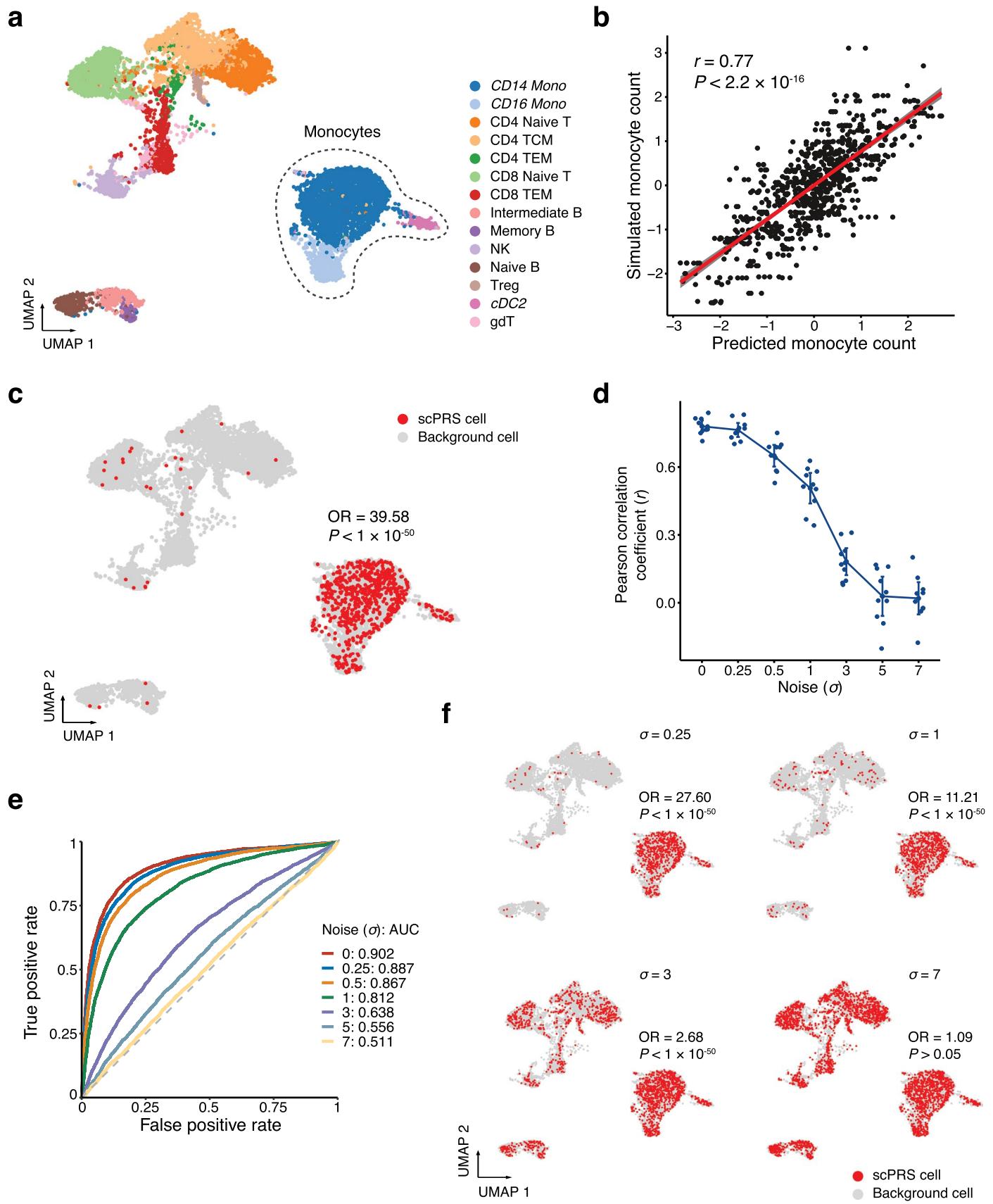
**Extended data** is available for this paper at <https://doi.org/10.1038/s41587-025-02725-6>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41587-025-02725-6>.

**Correspondence and requests for materials** should be addressed to Sai Zhang, Jianyang Zeng, Philip S. Tsao or Michael P. Snyder.

**Peer review information** *Nature Biotechnology* thanks Manu Setty and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



Extended Data Fig. 1 | See next page for caption.

**Extended Data Fig. 1 | Assessing performance of scPRS using simulations.**

**a**, The uniform manifold approximation and projection (UMAP) plot of the human peripheral blood mononuclear cell (PBMC) scATAC-seq dataset. Cell clusters with less than 150 cells are not shown. Monocyte subtypes are highlighted in italic. Mono, monocyte; TCM, memory T cell; TEM, effector memory T cell; NK, natural killer cell; Treg, regulatory T cell; cDC2, conventional type 2 dendritic cell; gdT, gamma-delta T cell. **b**, Pearson correlation between simulated and predicted monocyte counts ( $n=10$  repeats).  $P$ -value by two-sided Pearson correlation. The linear regression line and 95% confidence interval (CI) are annotated in the red line and gray shaded area, respectively.

**c**, Monocyte-count-relevant cells prioritized by scPRS (in red). Odds ratio and  $P$ -value by two-sided Fisher's exact test. OR, odds ratio. **d**, Pearson correlation between simulated and predicted monocyte counts ( $n=10$  repeats) in different noise settings. The mean and 95% CI are annotated in the dot and error bar, respectively.  $\sigma$ , standard deviation. **e**, The receiver operating characteristic (ROC) curves for cell prioritization in different noise settings, wherein monocytes were labeled as "1" and other cells were labeled as "0". AUC, the area under the curve. **f**, Monocyte-count-relevant cells prioritized by scPRS (in red) in different noise settings. Odds ratio and  $P$ -value by two-sided Fisher's exact test.

## Reporting Summary

Nature Portfolio 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 Portfolio 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 dx-toolkit v0.347.0 was used to download data from UKBB RAP.

Data analysis The sequence deep learning model was implemented using Selene v0.2.0 (<https://github.com/FunctionLab/selene>). Baseline PRS methods included C+T (PLINK v1.9), LDpred2 (bigsnp v1.12.2), and Lassosum v0.4.5. We also implemented SCAVENGE v1.0.2 (<https://github.com/sankaranlab/SCAVENGE>) and stratified LDSC v1.0.1 (<https://github.com/bulik/ldsc>). Single-cell data analysis was performed using Scrublet v0.2.2, Seurat v4.3.0, Scanpy v1, Signac v1.11.0, and ALLCools v1.1.0 (<https://github.com/lhqing/ALLCools>). GO analysis was conducted using Enrichr (<https://maayanlab.cloud/Enrichr/>) if not specified. TF motif analysis was carried out using GimmeMotifs v0.18.0(<https://gimmemotifs.readthedocs.io/en/master/>). We performed TF binding site prediction using SNP2TFBS (<https://epd.expasy.org/snp2tfbs/>) and motifbreakR v2.15.5 (<https://github.com/Simon-Coetze/motifBreakR>). All statistical analyses were performed using Python v3 and R v4. scPRS source code can be found at <https://github.com/szhang1112/scPRS>.

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 Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The PBMC multiome dataset is available from 10x Genomics ([https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc\\_granulocyte\\_sorted\\_10k](https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k)). The single-cell multiome data (snRNA-seq and snATAC-seq coassay) of the human left ventricle and lung are publicly accessible through ENCODE 4 ([https://www.encodeproject.org/single-cell/?type=Experiment&assay\\_slims=Single+cell&status=released](https://www.encodeproject.org/single-cell/?type=Experiment&assay_slims=Single+cell&status=released)). All other single-cell ATAC-seq datasets were obtained from their original publications (refs. 7,8). The WGS data used in simulation are available from ref. 28. Individual-level genotype-phenotype data for T2D and AD were sourced from the UK Biobank. The WGS and iPSC RNA-seq data for HCM are available from ref. 14. The COVID-19 WGS and clinical data are available upon request from the corresponding authors (P.S.T. and M.P.S.); these data are not publicly available due to US Government and Department of Veteran's Affairs restrictions relating to participant privacy and consent. The independent target cohorts for T2D, HCM, and AD are accessible through EGA (no. EGAD00001002247), UKBB, and ADNI (<https://adni.loni.usc.edu/data-samples/adni-data/>), respectively. The HCM snRNA-seq dataset was obtained from ref. 52. All GWAS summary statistics data were acquired from their original publications (refs. 27,53–55). The GTEx and islet eQTL datasets were downloaded from the eQTL Catalogue (<https://www.ebi.ac.uk/eqtl/>). Other eQTL and caQTL datasets were obtained from their original publications (refs. 56–59). The reference human genomes (hg19 and hg38) are available from <https://hgdownload.soe.ucsc.edu/downloads.html#human>.

## Research involving human participants, their data, or biological material

### Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

#### Reporting on sex and gender

Sex was included as a covariate in our analysis when possible.

#### Reporting on race, ethnicity, or other socially relevant groupings

We focused on European samples when possible. Mixed populations were analyzed for small cohorts such as HCM and COVID-19. The first 10 genotype principal components (PCs) were included as covariates in the analysis.

#### Population characteristics

The UK Biobank is a large-scale prospective cohort study including 500,000 participants recruited between 2006 and 2010. These participants, aged between 40 and 69, provided extensive biological and medical data, including physical measures, lifestyle information, and health records. We constructed T2D and AD target cohorts based on the UKBB with matched age and sex between cases and controls. Cases were defined by the ICD-10 code in inpatient record, death record, or diagnosis summary record.

HCM target cohort includes 97 healthy controls (age: 52.4 [mean]  $\pm$  18.2 [SD]; male: 54.6%) and HCM cases (age: 54.4  $\pm$  16.3; male: 61.4%). The targeted patient population were patients with various cardiac procedures and non-cardiac patients with genetic conditions in clinic who were identified to us by their clinical providers. Non-cardiac patients were recruited in person during onsite clinic days or over the phone with permission by the providers. Healthy volunteers were recruited from our cardiovascular prevention clinic (i.e., patients with no diagnosis of heart disease).

The VA COVID-19 cohort was derived from the VA Million Veteran Program (MVP). COVID-19 cases were identified using an algorithm developed by the VA COVID National Surveillance Tool based on reverse transcription polymerase chain reaction laboratory test results conducted at VA clinics, supplemented with natural language processing on clinical documents for SARS-CoV-2 tests conducted outside of the VA19. We defined severe COVID-19 cases as patients who were hospitalized, received acute care, stayed in ICU, or deceased, and controls as those who did not meet these criteria. To minimize potential confounders, we restricted our analysis to non-elderly individuals (age < 65).

#### Recruitment

The recruitment of the HCM cohort is part of our CIRM cardiomyopathy project (ref. 14). The targeted patient population were patients with various cardiac procedures and non-cardiac patients with genetic conditions in clinic who were identified to us by their clinical providers. Non-cardiac patients were recruited in person during onsite clinic days or over the phone with permission by the providers. Healthy volunteers were recruited from our cardiovascular prevention clinic (i.e., patients with no diagnosis of heart disease).

The VA COVID-19 cohort was derived from the VA Million Veteran Program (MVP). The VA MVP is an ongoing national voluntary research program that aims to better understand how genetic, lifestyle, and environmental factors influence veteran health. Briefly, individuals aged 18 to over 100 years old have been recruited from over 60 VA Medical Centers nationwide since 2011 with current enrollment at >800,000. Informed consent is obtained from all participants to provide blood for genomic analysis and access to their full electronic health record (EHR) data within the VA prior to and after enrollment. The study received ethical and study protocol approval from the VA Central Institutional Review Board in accordance with the principles outlined in the Declaration of Helsinki. COVID-19 cases were identified using an algorithm developed by the VA COVID National Surveillance Tool based on reverse transcription polymerase chain reaction laboratory test results conducted at VA clinics, supplemented with natural language processing on clinical documents for SARS-CoV-2 tests conducted outside of the VA.

#### Ethics oversight

Ethics approval for the UKB study was obtained from the North West Centre for Research Ethics Committee (protocol no. 11/NW/0382). The Stanford CIRM cardiomyopathy project is in compliance with the Stanford Human Research Protection Program guidelines and approved by the Stanford Institutional Review Board (IRB #30064). In addition, the procedures are in

compliance with the International Society of Stem Cell Research guidelines and approved by the Stanford IRB/Stem Cell Research Oversight panel (SCRO #656). Ethics approval for the MVP COVID-19 study was obtained from the Office of Research and Development, Veterans Health Administration (MVP001)

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

## 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](https://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	All disease cases were utilized in our analysis for each cohort. The same number of control samples were randomly selected from UKBB for T2D and AD.
Data exclusions	Samples that failed quality controls were excluded from the analysis. For severe COVID-19, samples with age $\geq 65$ were excluded to remove confounding factors. Overweight individuals (body mass index (BMI) $\geq 25$ ) were excluded for T2D.
Replication	The siRNA experiment was replicated for 8 times for each siRNA. The prime-editing experiment was replicated for 4 times. All replicates were successful.
Randomization	Random training and testing dataset splits were employed in evaluating model performance.
Blinding	In PRS model training, investigators were not blinded to group allocation because the sample labels were needed for training an ML model. In other scenarios, such as model evaluation and siRNA/prime-editing experiments, the investigators were blinded to group allocation.

## 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
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

### Cell line source(s)

For prime editing experiment, microglia were derived from the WTC11 iPSC line, which was originally established in Dr. Bruce Conklin's lab at UCSF and later shared as a gift with Dr. Yin Shen's lab. Detailed information about this iPSC line can be found in Miyaoka et al., *Nature Methods* (PMID: 24509632) and at <http://hpscreg.eu/cell-line/UCSF1001-A>.

For TMEM119-tdTomato reporter iPSCs, lymphoblastoid cells obtained from a healthy, 56-year-old male, were provided by the NINDS biorepository and reprogrammed into iPSCs in house.

### Authentication

hPSCs are routinely tested for expressing stem cell markers and differentiated microglia are tested by known microglia markers. Confirmation of the knock-in TMEM119-tdTomato construct was obtained by genotyping.

### Mycoplasma contamination

All cells used in the present study were verified as mycoplasma contamination free.

### Commonly misidentified lines (See [ICLAC](#) register)

None of the cell lines used are commonly misidentified lines.

## Plants

Seed stocks

NA

Novel plant genotypes

NA

Authentication

NA