Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data

Martin Jinye Zhang1,2,20✉, Kangcheng Hou3,4,5,20✉, Kushal K. Dey1,2, Saori Sakaue2,2,2,7,8,9, Karthik A. Jagadeesh1,2, Kathryn Weinand2,6,7,8,9, Aris Taychameekiatchail10,11, Poovri Rao10, Angela Oliveira Pisco12, James Zou12,13,14, Bruce Wang10, Michael Gandal15,16,17, Soumya Raychaudhuri2,6,7,8,18, Bogdan Pasaniuc3,4,5,21✉ and Alkes L. Price1,2,19,21✉

Single-cell RNA sequencing (scRNA-seq) provides unique insights into the pathology and cellular origin of disease. We introduce single-cell disease relevance score (scDRS), an approach that links scRNA-seq with polygenic disease risk at single-cell resolution, independent of annotated cell types. scDRS identifies cells exhibiting excess expression across disease-associated genes implicated by genome-wide association studies (GWASs). We applied scDRS to 74 diseases/traits and 1.3 million single-cell gene-expression profiles across 31 tissues/organs. Cell-type-level results broadly recapitulated known cell-type disease associations. Individual-cell-level results identified subpopulations of disease-associated cells not captured by existing cell-type labels, including T cell subpopulations associated with inflammatory bowel disease, partially characterized by their effector-like states; neuron subpopulations associated with schizophrenia, partially characterized by their spatial locations; and hepatocyte subpopulations associated with triglyceride levels, partially characterized by their higher ploidy levels. Genes whose expression was correlated with the scDRS score across cells (reflecting coexpression with GWAS disease-associated genes) were strongly enriched for gold-standard drug target and Mendelian disease genes.

The mechanisms through which risk variants identified by GWASs impact critical tissues and cell types are largely unknown; identifying these tissues and cell types is central to our understanding of disease etiologies and will inform efforts to develop therapeutic treatments. scRNA-seq has emerged as the tool of choice for measuring gene expression abundance at single-cell resolution, providing an increasingly clear picture of which genes are active in which cell types and identifying new cell populations within classically defined cell types. Integrating scRNA-seq with GWAS data offers the potential to identify critical tissues, cell types and cell populations through which GWAS risk variants impact disease13, thus providing finer resolution than studies using bulk transcriptomic data13–19.

Previous studies integrating scRNA-seq with GWAS have largely focused on predefined cell-type annotations (for example, classical cell types defined using known marker genes), aggregating cells from the same cell type followed by evaluating overlap of the cell-type-level information with GWAS14–19. However, this approach overlooks the considerable heterogeneity of individual cells within cell types that has been reported in studies of scRNA-seq data alone13–19; the underlying methods (for example, Seurat cell-scoring function14, Vision15 and VAM16) have sought to explain transcriptional heterogeneity in scRNA-seq data by scoring cells based on predefined gene sets such as cellular pathways but do not consider polygenic disease risk from GWAS and generally do not provide individual-cell-level association P values. Integrating information from scRNA-seq data at fine-grained resolution (for example, individual cells both within and across cell types) with polygenic signals from disease GWAS has considerable potential to produce new biological insights.

1Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. 2Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 3Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA. 4Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. 5Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. 6Center for Data Sciences, Brigham and Women’s Hospital, Boston, MA, USA. 7Division of Genetics, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, USA. 8Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA, USA. 9Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. 10Department of Medicine and Liver Center, University of California, San Francisco, San Francisco, CA, USA. 11Department of Developmental and Stem Cell Biology Graduate Program, University of California, San Francisco, San Francisco, CA, USA. 12Chan Zuckerberg Biohub, San Francisco, CA, USA. 13Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA. 14Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA. 15Department of Psychiatry, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. 16Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. 17Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA. 18Centre for Genetics and Genomics Versus Arthritis, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK. 19Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. 20These authors contributed equally: Martin Jinye Zhang, Kangcheng Hou. 21These authors jointly supervised this work: Bogdan Pasaniuc, Alkes L. Price. ✉ e-mail: jinyezhang@hsph.harvard.edu; houkc@ucla.edu; pasaniuc@ucla.edu; aprice@hsph.harvard.edu
Here, we introduce the single-cell disease relevance score (scDRS), a method to evaluate polygenic disease enrichment of individual cells in scRNA-seq data. scDRS assesses whether a given cell has excess expression levels across a set of putative disease genes derived from GWASs using an appropriately matched empirical null distribution to estimate well-calibrated P values. We performed extensive simulations to assess the calibration and power of scDRS. We applied scDRS to 74 diseases/traits (average GWAS N = 346K where “K” means thousands) and 16 scRNA-seq data sets (including the Tabula Muris Senis (TMS) mouse cell atlas\(^1\)), assessing cell-type–disease associations and within-cell-type association heterogeneity, including heterogeneity of T cells in association with autoimmune diseases, neurons in association with brain-related diseases/traits and hepatocytes in association with metabolic traits. We analyzed a broader set of scRNA-seq data sets to provide validation across species (human versus mouse) and across sequencing platforms and include scRNA-seq data sets with experimentally determined cell types and cell states.

**Results**

**Overview of methods.** scDRS integrates gene expression profiles from scRNA-seq with polygenic disease information from GWASs to associate individual cells to disease, by assessing the excess expression of GWAS putative disease genes in a given cell relative to other genes with similar expression across all cells. scDRS consists of three steps (Fig. 1). First, scDRS constructs a set of putative disease genes from GWAS summary statistics using MAGMA\(^{18}\), an existing gene scoring method (top 1,000 MAGMA genes). Second, scDRS quantifies the aggregate expression of the putative disease genes in each cell to generate cell-specific raw disease scores; to maximize power, each putative disease gene is weighted by its GWAS MAGMA z-score and inversely weighted by its gene-specific technical noise level in scRNA-seq. scDRS also computes a set of 1,000 MC raw control scores for each cell, in each case using a random set of control genes matching the gene set size, mean expression and expression variance of the putative disease genes. (3) scDRS normalizes the raw disease score and raw control scores across gene sets and across cells and then computes a P value for each cell based on the empirical distribution of the pooled normalized control scores across all control gene sets and all cells. The choice of 1,000 for the number of putative disease genes and the choice of 1,000 for the number of control scores are independent.

Fig. 1 | Overview of the scDRS method. scDRS takes a disease GWAS and an scRNA-seq data set as input and outputs individual-cell-level P values for association with the disease. (1) scDRS constructs a set of putative disease genes from GWAS summary statistics by selecting the top 1,000 MAGMA genes; these putative disease genes are expected to have higher expression levels in the relevant cell population. (2) scDRS computes a raw disease score for each cell, quantifying the aggregate expression of the putative disease genes in that cell; to maximize power, each putative disease gene is weighted by its GWAS MAGMA z-score and inversely weighted by its gene-specific technical noise level in scRNA-seq. scDRS also computes a set of 1,000 MC raw control scores for each cell, in each case using a random set of control genes matching the gene set size, mean expression and expression variance of the putative disease genes. (3) scDRS normalizes the raw disease score and raw control scores across gene sets and across cells and then computes a P value for each cell based on the empirical distribution of the pooled normalized control scores across all control gene sets and all cells. The choice of 1,000 for the number of putative disease genes and the choice of 1,000 for the number of control scores are independent.
disease score. Third, we perform gene-level analyses to prioritize disease-relevant genes whose expression is correlated with the scDRS disease score, reflecting coexpression with genes implicated by disease GWASs.

We analyzed publicly available GWAS summary statistics of 74 diseases/traits (average $N = 346K$; Supplementary Table 1) in conjunction with 16 scRNA-seq or single-nucleus RNA sequencing (snRNA-seq) data sets spanning 1.3 million cells from 31 tissues/organs from mouse ($Mus$ $musculus$) and human ($Homo$ $sapiens$) (Supplementary Tables 2–7 and Supplementary Fig. 4). The single-cell data sets include two data sets from the TMS mouse cell atlas$^{17}$ collected using different technologies, the Tabula Sapiens (TS) human cell atlas$^{18}$, and other data sets focusing on specific tissues containing well-annotated cell types/states. We focused on the TMS fluorescence-activated cell sorting (FACS) data in our primary analyses due to its comprehensive coverage of 23 tissues and 120 cell types and more accurate quantification of gene expression levels (via SMART-Seq$^{24}$); we used the other 15 data sets to validate our results. We note the extensive use of mouse gene expression data to study human diseases and complex traits$^{19,20}$ (Supplementary Note).

**Simulations assessing calibration and power.** We performed null simulations and causal simulations to assess the calibration and power of scDRS, comparing scDRS to three state-of-art methods for scoring individual cells with respect to a specific gene set: Seurat (cell-scoring function)$^{19}$, Vision$^{11}$ and VAM$^{10}$ (Methods).

First, we evaluated each method in null simulations where no cells have systematically higher expression across the putative disease genes analyzed. We subsampled 10,000 cells from the TMS FACS data and randomly selected 1,000 putative disease genes. We simulated GWAS gene weights for scDRS matching the MAGMA $z$-score distributions in real traits and used binary disease gene sets for the other three methods. scDRS and Seurat produced well-calibrated $P$ values, Vision showed slightly inflated type I errors and VAM produced severely inflated type I errors (Fig. 2a and Supplementary Table 8).

Next, we evaluated scDRS, Seurat and Vision in causal simulations where a subset of causal cells has systematically higher expression across putative disease genes (we did not include VAM, which was not well calibrated in null simulations). We used the same 10,000 cells subsampled from the TMS FACS data, randomly selected 1,000 causal disease genes, randomly selected 500 of the 10,000 cells as causal cells and artificially perturbed their expression levels to be higher (1.05–1.50× for different simulations) across the 1,000 causal disease genes, and we randomly selected 1,000 putative disease genes (provided as input to each method) with 25% overlap with the 1,000 causal disease genes. We used the binary gene set for all three methods because there were no GWAS weights involved in data generation. We determined that scDRS attained higher power than Seurat and Vision to detect individual-cell–disease associations at false discovery rate (FDR) < 0.1 (Fig. 2b and Supplementary Table 9); the improved power of scDRS may be due to its incorporation of gene-specific weights that downweight genes with higher levels of technical noise.

Results of additional null and causal simulations are reported in the Supplementary Note, Extended Data Figs. 1, 4 and 5; Supplementary Tables 13–16.

**Results across 120 TMS cell types for 74 diseases and traits.** We analyzed GWAS data from 74 diseases/traits (average $N = 346K$; Supplementary Tables 1 and 11) in conjunction with the TMS FACS data with 120 cell types (cells from different tissues were combined for a given cell type; Supplementary Table 5). We first report scDRS cell-type-level results, aggregated for each cell type from the scDRS individual-cell-level results; the individual-cell-level results are discussed in subsequent sections. Results for a representative subset of 19 cell types and 22 diseases/trait are reported in Fig. 3. Within this subset, scDRS identified 80 associated cell-type–disease pairs (FDR < 0.05; squares in Fig. 3) and detected significant within-cell-type disease-association heterogeneity for 43 of these 80 associated cell-type–disease pairs (FDR < 0.05; cross symbols in Fig. 3) (273 of 597 across all pairs of the 120 cell types and 74 diseases/trait; Extended Data Figure 3 and Supplementary Table 12).

For cell-type–disease associations, as expected, scDRS broadly associated blood/immune cell types with blood/immune-related diseases/trait, brain cell types with brain-related diseases/trait and other cell types with other diseases/trait (block-diagonal pattern in Fig. 3). Most scDRS discoveries recapitulated well-established cell-type–disease associations, including blood/immune cell types with blood cell traits, immune cell types with immune diseases, neuronal cell types with brain-related traits/diseases$^{22,26}$ and hepatocytes with metabolic traits$^{26}$. In addition, chondrocytes, bladder cells, ventricular myocytes and pancreatic beta cells were associated with their corresponding expected diseases/trait$^{27,28}$, Exceptions to the block-diagonal pattern and further details are discussed in the Supplementary Note.

We also discuss several less well-established results. First, granulocyte monocyte progenitors were strongly associated with multiple sclerosis (MS), highlighting the role of myeloid cells in MS$^{27}$. Second, oligodendrocytes and oligodendrocyte precursor cells (OPCs) were associated with multiple brain-related diseases/trait. These associations are less clear in existing genetic studies$^{3,4,20,26}$, but are biologically plausible, consistent with the increasingly discussed role of oligodendrocyte lineage cells in brain diseases/trait; the differentiation and myelination activity of oligodendrocyte lineage cells are important to maintain the functionality of neuronal cells$^{11}$. We detected significant heterogeneity across OPCs in their association with many brain-related diseases/trait, consistent with recent evidence of functionally diverse states of OPCs$^{26}$, traditionally considered to be a homogeneous population (Supplementary Fig. 6). Third, we detected significant heterogeneity across cells for the association between proerythroblasts and red cell distribution width, which may correspond to erythroblasts at different differentiation stages$^{41}$ (Supplementary Fig. 7). We also detected other instances of significant within-cell-type association heterogeneity, including T cells with immune diseases, neurons with brain-related diseases/trait and hepatocytes with metabolic traits (discussed in subsequent sections).

Additional secondary analyses assessing robustness of the results and alternative versions of scDRS are reported in Methods; Supplementary Note; Extended Data Figs. 1, 4 and 5; Supplementary Figs. 8–11; and Supplementary Tables 13–16.

**Heterogeneous T cells subpopulations in autoimmune disease.** We investigated the heterogeneity across TMS FACS T cells in association with autoimmune diseases (Fig. 3). We jointly analyzed all TMS FACS T cells (3,769 cells spanning 15 tissues). Because the original study clustered cells from different tissues separately$^{17}$, we reclustered these T cells, resulting in 11 clusters (Fig. 4a and Methods); we verified that batch effects were not observed for tissue, age or sex (Supplementary Fig. 12). We considered ten autoimmune diseases: inflammatory bowel disease (IBD), Crohn's disease (CD), ulcerative colitis (UC), rheumatoid arthritis (RA), MS, autoimmune traits (AITs; a general term for autoimmune diseases), hypothyroidism (HT), eczema, asthma (ASM) and respiratory and ear-nose-throat diseases (RR-ENT) (Supplementary Table 1); we considered height as a negative control trait.

We focused on individual cells associated with IBD, a representative autoimmune disease (Fig. 4b). The 387 IBD-associated cells (FDR < 0.1) formed subpopulations of 4 of the 11 T cell clusters. The subpopulation of 123 IBD-associated cells in cluster 3 (labeled...
T_{reg}-like) had high expression of regulatory T (T_{reg}) cell marker genes (FOXP3, CTLA4 and LAG3; Supplementary Fig. 17a), and their specifically expressed genes significantly overlapped with T_{reg} signatures ($P=6.0\times10^{-4}$ for MSigDB signatures and $P=4.0\times10^{-38}$ for an effector-like T_{reg} program\(^4\), two-sided Fisher’s exact test; Supplementary Fig. 17c,d), suggesting these cells had T_{reg} immune-suppressive functions. Interestingly, these 123 IBD-associated cells were nonrandomly distributed in cluster 3 on the uniform manifold approximation and projection (UMAP) plot ($P<0.001$, MC test; Methods). Genes specifically expressed in these IBD-associated cells were preferentially enriched (compared to the 506 non-IBD-associated cells in the same cluster) in pathways defined by nuclear factor kB signaling, T helper cell differentiation and tumor necrosis factor-mediated signaling (Supplementary Fig. 17c); these pathways are closely related to inflammation, a distinguishing feature of IBD\(^3\). In addition, the subpopulations of IBD-associated cells in clusters 4, 5 and 9 were labeled as “Th2/T_{reg}-like”, “Th17-like” and “CD8\(^+\) effector-like”, respectively, consistent with previous studies associating subpopulations of effector T cells to IBD, particularly T_{reg} and Th17 cells\(^1\). Results for other autoimmune diseases, details for annotating disease-associated T cell subpopulations and replication analyses on two human scRNA-seq data sets\(^{36,37}\) are reported in Extended Data Fig. 2.

We further compared the individual T cell associations of IBD with HT, another representative autoimmune disease (Fig. 4c,d). The top four HT-associated subpopulations included not only three IBD-associated subpopulations (cells in clusters 3, 4 and 9; Fig. 4c) but also a unique subpopulation of cells in cluster 4 (labeled as “Th2/T_{reg}-like”; Fig. 4d). Additional results are reported in the Supplementary Note, Extended Data Figure 6, Supplementary Fig. 18 and Supplementary Table 19.

Motivated by the effector-like T cell subpopulations associated with IBD, we investigated whether the heterogeneity of T cells in association with autoimmune diseases was correlated with T cell effectorness gradient, a continuous classification from naive to effector T cells. We separately computed the effectorness gradients for CD4\(^+\) T cells (1,686 cells) and CD8\(^+\) T cells (2,197 cells) using pseudotime analysis\(^{36,38}\) and assessed whether the CD4 (respectively, CD8) effectorness gradient was correlated with scDRS disease scores for the ten autoimmune diseases across CD4\(^+\) T cells (respectively, CD8\(^+\) T cells). Results are reported in Fig. 4e and Supplementary Table 20. The CD4 effectorness gradient had strong associations with IBD, CD, UC,AIT and HT ($P<0.005$, MC test), weak associations with eczema and ASM ($P<0.05$ and nonsignificant associations with RA, MS and RR-ENT, implying these autoimmune diseases are associated with more effector-like CD4\(^+\) T cells. The CD8 effectorness gradient had weaker associations ($P<0.05$ for IBD and CD and AIT but nonsignificant for others), suggesting that CD4\(^+\) effector T cells may be more important than CD8\(^+\) effector T cells for these diseases. The association of T cell effectorness gradients with autoimmune diseases has not been formally evaluated previously but is consistent with previous studies linking T cell effector functions to autoimmune disease\(^39\) or characterizing similarities among effector T cell subtypes\(^36,40\). Additional results on T cell effectorness gradient and comparison to cluster-level LDSC-SEG\(^10\) are reported in the Supplementary Note, Supplementary Figs. 19 and 20 and Supplementary Tables 17 and 20.

Finally, we prioritized disease-relevant genes by computing the correlation (across all TMS FACS cells) between the expression of a given gene and the scDRS score for a given disease; this approach identifies genes coexpressed with genes implicated by disease GWAS. We compared the top 1,000 genes prioritized using this approach with putative drug targets\(^41\) (for eight autoimmune diseases and excluding RR-ENT and HT) or genes known to cause a Mendelian form of the disease\(^1\) (for RR-ENT and HT, for which drug targets are not available; Supplementary Table 21). Results are

---

**Fig. 2 | Results for null and causal simulations.** a. Quantile-quantile plot for null simulations using 1,000 randomly selected genes as the putative disease genes. Random GWAS gene weights were used for scDRS matching the MAGMA z-score distributions in real traits while binary gene sets were used for the other three methods. $x$ axis denotes theoretical $-\log_{10} P$ value quantiles, and $y$ axis denotes actual $-\log_{10} P$ value quantiles for different methods. Error bars denote 95% confidence intervals around the mean of 100 simulation replicates (with 10,000 cells per simulation replicate); all error bars are <0.05 from the point estimates. Numerical results are reported in Supplementary Table 8, and additional results are reported in Extended Data Fig. 1 and Supplementary Fig. 5. b, Power for casual simulations with perturbed expression of causal genes in causal cells. We report the power at FDR = 0.01 for different methods and different effect sizes. Error bars denote 95% confidence intervals around the mean of 100 simulation replicates (with 10,000 cells per simulation replicate); all error bars are <0.02 from the point estimates. Numerical results are reported in Supplementary Table 9, and additional results are reported in Extended Data Fig. 2.
We determined that scDRS attained a more accurate prioritization of disease-relevant genes compared to the top 1,000 MAGMA genes (2.07 for median ratio of excess overlap − 1); Methods), likely by capturing disease-relevant genes with weak GWAS signal43. For example, scDRS prioritized ITGB7 for IBD (rank 11; drug target for IBD using vedolizumab44) and JAK1 for RA (rank 358; drug target for RA using baricitinib45), both missed by MAGMA (ranks 10,565 and 5,228, P = 0.54, 0.26, respectively). Additional results are reported in the Supplementary Note, Extended Data Figure 7 and Supplementary Table 21.

**Heterogeneous neuron subpopulations in brain traits.** We investigated the heterogeneity across neurons in association with brain-related diseases/trait (Fig. 3). We considered seven brain-related diseases/trait: schizophrenia (SCZ), major depressive disorder (MDD), bipolar disorder (BP), neuroticism (NRT), college education (ECOL), body mass index (BMI) and smoking status (smoking). Complete results for 120 cell types and 74 diseases/trait are reported in Extended Data Figure 3 and Supplementary Table 12.

**Fig. 3 | Disease associations at the cell-type level.** We report scDRS results for individual cells aggregated at the cell-type level for a subset of 19 cell types and 22 diseases/trait in the TMS FACS data. Each row represents a disease/trait and each column represents a cell type (with number of cells indicated in parentheses). Heatmap colors for each cell-type–disease pair denote the proportion of significantly associated cells (“Prop. of sig. cells” in the figure legend; FDR < 0.1 across all cells for a given disease). Squares denote significant cell-type–disease associations (FDR < 0.05 across all pairs of the 120 cell types and 74 diseases/trait; P values via MC test; Methods). Cross symbols denote significant heterogeneity in association with disease across individual cells within a given cell type (FDR < 0.05 across all pairs; P values via MC test; Methods). Heatmap colors (>10% of cells associated) and cross symbols are omitted for cell-type–disease pairs with nonsignificant cell-type–disease associations via MC test (heatmap colors omitted for one pair (dendritic–ASM) and cross symbols omitted for six pairs (CD4+α–β T-MONO, CD8+α–β T-MONO, bladder cell–RA, bladder cell–ASM, oligodendrocyte–bipolar disorder and dendritic–heel T score)). Autoimmune traits (AITs) represent a collection of diseases in the UK Biobank that characterize autoimmune physiopathogenic etiology62 (Supplementary Table 1). Abbreviated cell-type names include red blood cells, granulocyte monocyte progenitors, medium spiny neuron (MSN) and OPCs. Neuron refers to neuronal cells with undetermined subtypes (whereas MSN and interneuron (nonoverlapping with neuron) refer to neuronal cells with those inferred subtypes). Complete results for 120 cell types and 74 diseases/trait are reported in Extended Data Figure 3 and Supplementary Table 12.
**Fig. 4 | Associations of T cells with autoimmune diseases.** a, UMAP visualization of T cells in TMS FACS. Cluster labels are based on annotated TMS cell types in the cluster. b, c, Subpopulations of T cells associated with IBD and HT, respectively. Significantly associated cells (FDR < 0.05) are denoted in red, with shades of red denoting scDRS disease scores; other cells are denoted in gray. Cluster boundaries indicate the corresponding T cell clusters from panel a. Clusters are annotated based on the putative identities of associated cells in the cluster, for the top 4 clusters (out of 11) with the strongest level of association (highest average disease score for associated cells in the cluster). The number of disease-associated cells and the number of all cells in the cluster are provided in parentheses. d, Differences in individual-cell-level associations between IBD and HT. Differentially associated cells (absolute scDRS disease score difference >2) are denoted in red and blue, with shades of colors denoting scDRS disease score differences; other cells are denoted in gray. Cluster boundaries indicate the corresponding T cell clusters from panel a. Clusters are annotated the same as in panels b and c. The numbers of IBD-enriched cells, HT-enriched cells and all cells in the cluster are provided in parentheses. For panels b-d, results for the other eight autoimmune diseases are reported in Supplementary Figs. 13 and 18. e, Association between scDRS disease score and CD4 effectorness gradient across CD4+ T cells for five representative autoimmune diseases and height, a negative control trait. x axis denotes CD4 effectorness gradient quintile bins, and y axis denotes the average scDRS disease score in each bin for each disease. *P < 0.05 and, **P < 0.005 (one-sided MC test). Numerical results for all ten autoimmune diseases are reported in Supplementary Table 20. f, Excess overlap with gold-standard gene sets. x axis denotes excess overlap of genes prioritized by MAGMA, and y axis denotes excess overlap of genes prioritized by scDRS, for each of the ten autoimmune diseases. Median ratio of (excess overlap − 1) for scDRS versus MAGMA is 2.07. Numerical results are reported in Supplementary Table 22.

level in previous genetic studies. Results for TMS FACS neurons are reported in the Supplementary Note, Extended Data Figure 6, Supplementary Figs. 21 and 22, and Supplementary Table 19.
We used the results of scDRS for individual cells to assess whether the inferred spatial scores for each of the six spatial regions (dorsal–ventral/proximal–distal/superficial–deep) were correlated to the inferred spatial coordinates for each cell in terms of six continuous individual-cell-level scores for these spatial regions (Supplementary Figs. 23c and 24, Supplementary Table 24 and Methods). The inferred spatial scores for the dorsal–ventral and proximal–distal axes varied along the top two UMAP axes, providing visual evidence of stronger neuron–SCZ associations in dorsal and proximal regions (Fig. 5a and Supplementary Fig. 23).

We considered the natural CA1 spatial axes (dorsal–ventral long axis, proximal–distal transverse axis and superficial–deep radial axis) and inferred spatial coordinates for each cell in terms of six continuous individual-cell-level scores for these spatial regions (Supplementary Figs. 23c and 24, Supplementary Table 24 and Methods). The inferred spatial scores for the dorsal–ventral and proximal–distal axes varied along the top two UMAP axes, providing visual evidence of stronger neuron–SCZ associations in dorsal and proximal regions (Fig. 5a and Supplementary Fig. 23).

We observed a continuous gradient of CA1 pyramidal neuron associations with the seven brain-related traits.

Motivated by known location-specific functions of hippocampal neurons, we investigated whether the heterogeneity observed in Fig. 5a was correlated with spatial location. We considered the natural CA1 spatial axes (dorsal–ventral long axis, proximal–distal transverse axis and superficial–deep radial axis) and inferred spatial coordinates for each cell in terms of six continuous individual-cell-level scores for these spatial regions (Supplementary Figs. 23c and 24, Supplementary Table 24 and Methods). The inferred spatial scores for the dorsal–ventral and proximal–distal axes varied along the top two UMAP axes, providing visual evidence of stronger neuron–SCZ associations in dorsal and proximal regions (Fig. 5a and Supplementary Fig. 23).
Extended Data Figure 8 and Supplementary Table 25. The proximal score was strongly associated with all seven brain-related traits (all $P < 0.002$, MC test), suggesting proximal CA1 pyramidal neurons may be more relevant to these brain-related traits (instead of distal CA1 pyramidal neurons). The association between the proximal region and brain-related traits is consistent with the fact that the proximal region of the hippocampus receives synaptic inputs in the perforant pathway, which is the main input source of the hippocampus\(^6\). Validations of the spatial scores using independent data and results on other spatial scores and six additional mouse and human data sets\(^{10,12,13}\) are reported in the Supplementary Note and Extended Data Figure 8.

### Heterogeneous hepatocyte subpopulations in metabolic traits.

We investigated the heterogeneity across TMS FACS hepatocytes (in the liver) in their association with metabolic traits (Fig. 3). We considered nine metabolic traits: total cholesterol (TG), high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), testosterone (TST), alanine aminotransferase (ALT), alkaline phosphatase (ALP), sex hormone-binding globulin (SHBG) and total bilirubin (TBIL) (Supplementary Table 1); we considered height as a negative control trait.

We focused on individual cells associated with TG, a representative metabolic trait (Fig. 5c; results for other traits in Supplementary Fig. 25). The 530 TG-associated cells ($FDR < 0.1$) formed subpopulations of five of the six hepatocyte clusters; we characterized these subpopulations based on ploidy level (number of sets of chromosomes in a cell) and zonation (pericentral/periportal location in the liver lobule), which have been extensively investigated in previous studies of hepatocyte heterogeneity\(^{4,5}\). We inferred the ploidy level and zonation for each individual cell in terms of continuous individual-cell-level polyploidy, pericentral and peripheral scores (Supplementary Fig. 26 and Methods). The inferred ploidy level and zonation varied across clusters, providing visual evidence of stronger cell–TG associations in high-ploidy clusters (clusters 1 and 2), particularly the periporal high-ploidy cluster (cluster 2; Fig. 5c).

We used the results of scDRS for individual cells to assess whether the inferred polyploidy, pericentral and peripheral scores were correlated to the scDRS disease score for each of the nine metabolic traits across hepatocytes; we jointly regressed the scDRS disease score for each trait on the polyploidy score, pericentral score and peripheral score (because the polyploidy score was positively correlated with the other two scores; Methods). Results are reported in Fig. 5d (for the polyploidy score which had the strongest associations), Extended Data Figure 9 and Supplementary Table 26. The polyploidy score was strongly associated with all nine metabolic traits (all $P < 0.007$, MC test), suggesting that high-ploidy hepatocytes may be more relevant to these metabolic traits. The association between ploidy level and metabolic traits is consistent with previous findings that ploidy levels are associated with changes in the expression level of genes for metabolic processes such as de novo lipid biosynthesis and glycolysis\(^{47,48}\) and supports the hypothesis that liver functions are enhanced in polyploid hepatocytes\(^5\). In addition, the peripheral score was associated with the nine metabolic traits (all $P < 0.005$ except $P = 0.04,0.02,0.24$ for HDL, SHBG and TBIL, MC test). Although the pericentral score was not significantly associated with these traits in TMS FACS, we detected significant associations across multiple other data sets (Supplementary Note). These results suggest that these metabolic traits are impacted by complex processes involving both pericentral and peripheral hepatocytes.

Validation of the polyploidy and zonation scores using independent data and results on five additional mouse and human data sets\(^{15,19,41}\) (the unpublished Taychameekitchai et al. data were provided by A.T., P.R. and B.W.) are reported in the Supplementary Note, Extended Data Figure 9, Supplementary Fig. 27 and Supplementary Tables 24 and 27.

### Discussion

We have introduced scDRS, a method that leverages polygenic GWAS signals to associate individual cells in scRNA-seq data with diseases and complex traits. We showed via extensive simulations that scDRS is well calibrated and powerful in realistic scenarios. We applied scDRS to 74 diseases/trait in conjunction with 16 scRNA-seq data sets and detected extensive heterogeneity in disease associations of individual cells within classical cell types. These findings may prove useful for targeting the relevant cell populations for in vitro experiments to elucidate the molecular mechanisms through which GWAS risk variants impact disease.

We have demonstrated the value in associating individual cells to disease, assessing the heterogeneity across individual cells within predefined cell types in their association to disease, identifying cell-level variables partially characterizing the individual cells that are associated to disease, and broadly associating predefined cell types to disease. Analyses of larger scRNA-seq/sRNA-seq and GWAS data sets using approaches such as scDRS will continue to further these goals.

We note several limitations and future directions of our work. First, identifying a statistical correlation between individual cells (or cell types) and disease does not imply causality but may instead reflect indirect tagging of causal cells/cell types, analogous to previous works\(^{4,5,10,18}\). However, even in such cases, the implicated cells/cell types are likely to be closely biologically related to the causal cells/cell types, based on their similar expression patterns. Second, the fact that scDRS assesses the statistical significance of an individual cell’s association to disease by implicitly comparing it to other cells via matched control genes may reduce power if most cells in the data are truly causal. For example, association with IBD in a data set containing only Treg cells (one of the causal cell types for IBD) will likely yield largely nonsignificant results. This limitation did not impact our main analyses, because the TMS data include a broad set of cell types; in more specialized data sets (which may be preferred in some settings due to the more comprehensive profiling of the focal cell population), this limitation can potentially be addressed by selecting matched control genes based on a broad cell atlas (for example, the TMS or TS data). Third, although we have primarily focused on the associations involving a single disease/trait, further investigation of differences between diseases/trait within the same category is an important future direction. Additional limitations are discussed in the Supplementary Note. Despite all these limitations, scDRS is a powerful method for distinguishing disease associations of individual cells in scRNA-seq data.

### Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at [https://doi.org/10.1038/s41588-022-01167-z](https://doi.org/10.1038/s41588-022-01167-z).

Received: 23 September 2021; Accepted: 19 July 2022; Published online: 1 September 2022.

### References

1. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
2. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* **21**, 137–150 (2020).
3. Repov, A. et al. The Human Cell Atlas. *Cell* **6**, e227841 (2017).
4. Calderon, D. et al. Inferring relevant cell types for complex traits by using single-cell gene expression. *Am. J. Hum. Genet.* **101**, 686–699 (2017).
Methods

Ethics statement. This study analyzed publicly available data sets and hence did not require ethical approval.

scDRS method. We consider an scRNA-seq data set with \( n_{cell} \) cells (not cell types) and \( n_{gene} \) genes. We denote the cell–gene matrix as \( X \in \mathbb{R}^{n_{cell} \times n_{gene}} \), where \( X_{ig} \) represents the expression level of cell \( i \) and gene \( g \). We assume that \( X \) is size-factor-normalized (for example, 10,000 counts per cell) and log-transformed (\( \log(x + 1) \)) from the original raw count matrix \( x \). We regress the covariates out from the normalized data \( x \) (with a constant term in the regressors to center the data), before adding the original log mean expression of each gene back to the residual data. Such a procedure preserves the mean-variance relationship in the covariate-corrected data, necessary for estimating the gene-specific technical noise levels (Supplementary Note). Gene-level statistics for the scRNA-seq data are reported in Supplementary Data, necessary for estimating the gene-specific technical noise levels (Supplementary Data).

The scDRS algorithm is described below. Given a disease GWAS and an scRNA-seq data set, scDRS computes a \( P \) value for each individual cell for association with the disease. scDRS also outputs cell-level normalized disease scores and \( B \) sets of normalized control scores (default \( P \) value for each individual cell for \( B \)). scDRS computes an association \( P \) value for each cell by comparing its normalized disease score to the empirical distribution of the pooled normalized control scores across all control gene sets and all cells.

We discuss guidelines for using scDRS. First, scDRS relies on assumptions that require the disease gene set to have a moderate size (for example, \( > 50 \) genes and \( < 200 \) of all genes). Second, to ensure a reasonable number of scDRS discoveries, we recommend using GWAS data with a heritability \( z \)-score greater than 5 or a sample size greater than 100K. We also recommend using scRNA-seq data with a diverse set of cells potentially relevant to disease, although a smaller number of cells should not affect the scDRS power. However, scDRS will not produce false positives for less ideal GWAS or single-cell data sets. Third, scDRS is computationally efficient and scales linearly with the number of cells and number of control gene sets for both computation and memory use; it takes around \( 3 \) h and 60 GB for a single-cell data set with a million cells. Fourth, scDRS can be used in conjunction with any gene sets (instead of the MAGMA GWAS gene sets) to evaluate the enrichment of aggregate expression for cells in a single-cell data set. Fifth, we provide an option to adjust for cell–type proportions (or any cell group annotations), recommended only for extremely unbalanced data sets. Further details and evaluations of alternative versions of scDRS are provided in the Supplementary Note.

Algorithm description. We describe the scDRS algorithm.

**Input:** Disease GWAS summary statistics (for putative disease gene set \( G \) with GWAS gene weights \( \{w_{g}\}_{g \in G} \), scRNA-seq data \( X \in \mathbb{R}^{n_{cell} \times n_{gene}} \).

**Parameter:** Number of MC samples of control gene sets \( B \) (default 1,000).

1. **Construct putative disease gene set**

   a. Construct putative disease gene set \( G \subset \{1, 2, ..., n_{gene}\} \) from GWAS summary statistics using MAGMA.

2. **Compute disease scores and control scores**

   a. Sample \( B \) sets of control genes \( G_{ \hat{b}} \), \( b \in \{1, 2, ..., B\} \), matching mean expression and expression variance of disease genes.

   b. Sample gene-specific technical noise level \( \sigma_{tech}^{2} \) \( \forall g \in \{1, ..., n_{gene}\} \).

   c. Compute raw disease score and \( B \) MC samples of raw control scores for each cell. Third, after gene set-wise and cell-wise normalization, scDRS computes an association \( P \) value for each cell by comparing its normalized disease score to the empirical distribution of the pooled normalized control scores across all control gene sets and all cells.

3. **Compute disease–association \( P \) values**

   a. First gene set alignment by mean and variance. Let \( \sigma_{g}^{2} \) be the expression variance of gene \( g \). For each cell \( c \),

   \[
   s_{c} \leftarrow \bar{s}_{cell} - \frac{1}{n_{cell}} \sum_{i=1}^{n_{cell}} s_{i,c} - \frac{1}{n_{gene}} \sum_{g=1}^{n_{gene}} \sigma_{g}^{2} - \frac{1}{n_{gene}} \sum_{g=1}^{n_{gene}} \left( \bar{x}_{g,c} - \frac{1}{n_{cell}} \sum_{i=1}^{n_{cell}} x_{i,g} \right), \forall b \in \{1, ..., B\}
   \]

   b. Cell-wise standardization for each cell \( c \) by the mean \( \mu_{c}^{cell} \) and SD \( \sigma_{tech}^{cell} \) of control scores \( s_{c,1}^{cell}, ..., s_{c,B}^{cell} \) of that cell,

   \[
   s_{c,b} \leftarrow (s_{c} - \mu_{c}^{cell})/\sigma_{tech}^{cell}, \forall b \in \{1, ..., B\}
   \]

   c. Second gene set alignment by mean. For each cell \( c \),

   \[
   s_{c,b} \leftarrow -\frac{1}{n_{cell}} \sum_{i=1}^{n_{cell}} s_{i,c} + \frac{1}{n_{gene}} \sum_{g=1}^{n_{gene}} \sigma_{g}^{2} - \frac{1}{n_{gene}} \sum_{g=1}^{n_{gene}} \left( \bar{x}_{g,c} - \frac{1}{n_{cell}} \sum_{i=1}^{n_{cell}} x_{i,g} \right), \forall b \in \{1, ..., B\}
   \]

   d. Compute cell-level \( P \) values based on the empirical distribution of the pooled normalized control scores for each cell \( c \),

   \[
   P_{c} = \frac{1 + \sum_{b=1}^{B} \frac{1}{B} \mathbb{1}_{\left\{s_{c,b} \leq \bar{s}_{cell}\right\}}}{1 + n_{cell}B}
   \]

Output: cell-level \( P \) values \( P_{c} \), normalized disease scores \( s_{c,b} \), and normalized control scores \( s_{c,1}^{cell}, ..., s_{c,B}^{cell} \).

Downstream applications and MC test. We use a unified MC test for the scDRS downstream analyses based on the (normalized) disease and control scores. Specifically, let \( b \) be a test statistic computed from the disease score of the given set of cells (different downstream analyses differ by the test statistics), and let \( \bar{s}_{cell} \) be the same test statistics computed from the \( B \) sets of control scores of the same set of cells. The MC \( P \) value can be written as:

\[
P_{MC} = \frac{1 + \sum_{b=1}^{B} \frac{1}{B} \mathbb{1}_{\left\{b_{b} \leq \bar{s}_{cell}\right\}}}{1 + B}
\]

The MC test avoids the assumption that the cells are independent—a strong assumption in scRNA-seq analyses, for example, when analyzing cells in the same cluster that are dependent due to the clustering process. However, the MC \( P \) value \( P_{MC} \) cannot be smaller than \( 1/(1+B) \) by equation (6), limiting its ability in distinguishing highly significant associations. We can also compute an MC \( z \)-score as \( z_{MC} = \left[ t - \left( \left\{ \left\{ \bar{s}_{b}^{cell} \right\}_{b=1}^{B} \right\} / \left( \left\{ \sigma_{b}^{cell} \right\}_{b=1}^{B} \right) \right) \right] / \left( \left\{ \left\{ \sigma_{b}^{cell} \right\}_{b=1}^{B} \right\} / \left\{ \left\{ \bar{s}_{b}^{cell} \right\}_{b=1}^{B} \right\} \right) \); this MC \( z \)-score is not restricted by the MC limit of \( 1/(1+B) \) but relies on the assumption that the control test statistics \( \bar{s}_{b}^{cell} \) approximate a normal distribution. We use斯坦 MC \( P \) values to determine statistical significance and use MC \( z \)-scores to further prioritize associations whose MC \( P \) values have reached the MC limit. Below, we describe the test statistics used by the three analyses listed above. Besides the three analyses below, the MC test can in principle be extended to any analysis that computes a test statistic from the disease score of a set of cells.

**Associating cell type with disease.** We use the top 5% quantile of the disease score of cells from the given cell type as the test statistic. This test statistic is robust to annotation outliers, for example, a few misannotated but highly significant cells. One can also use other test statistics such as the top 1% quantile or the maximum.

**Assessing within-cell-type heterogeneity in association with disease.** We use Geary’s \( C^{(i)} \) as the test statistic. Geary’s \( C^{(i)} \) measures the spatial autocorrelation of the disease score across a set of cells from a single cell type or cell cluster) with respect to a cell–cell similarity matrix. Given a set of \( n \) cells, the corresponding disease scores \( s_{1}, ..., s_{n} \), and the cell–cell similarity matrix \( W \in \mathbb{R}^{n \times n} \), Geary’s \( C^{(i)} \) is calculated as

\[
C = \frac{(n - 1) - \sum_{i=1}^{n} W_{i} (s_{i} - s_{\bar{i}})^{2}}{2 \left( \sum_{i=1}^{n} W_{i} (s_{i} - s_{\bar{i}})^{2} \right)}
\]

where \( s_{\bar{i}} = \frac{1}{n} \sum_{i=1}^{n} s_{i} \). A value significantly lower than 1 suggests a high level of disease–association heterogeneity across the given set of cells. Details are provided in the Supplementary Note.

**Correlating a cell-level variable to disease across a given set of cells.** For associating a single-cell-level variable with disease, we use the Pearson’s correlation between the cell-level variable and the disease score across the given set of cells as the test statistic. For jointly associating multiple cell-level variables with disease, we use the regression \( t \)-statistic as the test statistic, obtained from jointly regressing the disease score against the cell-level variables.

**Simulations.** We considered three comparison methods: Seurat13 ("score_genomes" as implemented in Scanpy\(^{v1.6.0}\), Vision\(^{v}\) (implemented in scDRS v0.1.0) and VAM\(^{\ast}\) (0.5.1). To our knowledge, VAM is the only published cell-scoring method that provides cell-level association \( P \) values. We chose to include Seurat due to its wide use. We standardized its output cell-level scores (mean 0 and SD 1) before computing the cell-level \( P \) values based on the standard normal distribution.
We chose to include Vision because its outputs are nominal cell-level z-scores that can be easily converted to P values; we also added the standardization step, because otherwise, the Vision results were highly unstable. We did not include other methods like AGODA or AUCCell to estimate the distance of T cells from the cell signature genes; a gene was required to appear at least twice and genes appearing the same number of times were all included, resulting in 62 to 513 genes for the 10 T cell signature gene sets (Supplementary Table 24). For gold-standard gene sets used in the analysis of disease gene prioritization, we curated 27 putative drug target gene sets from Open Targets1 (mapped to 27 of the 74 diseases/trait); for a given disease, we selected all genes with drug score >0 (clinical trial phase 1 and above) and only considered diseases with at least 10 putative drug target genes. We curated 16 Mendelian disease genes from Freund et al. (mapped to 45 of the 74 diseases/trait) (Supplementary Table 21). For comparison of two gene sets, the P value is based on two-sided Fisher's exact test and excess overlap is defined as the ratio between the observed overlap of the two gene sets and the expected overlap (by chance). Of note, for a given query gene set with a fixed size and a fixed level of excess overlap with the reference gene set, the −log10 P value increases with the size of the reference gene set; we report both excess overlap and −log10 P value using the former as our primary metric, which is more interpretable.

Analysis of neurons and brain-related diseases/trait. For the TMS FACs data, we focused on the 484 neurons (TMS label "neuron", excluding cells with TMS label "medium spiny neuron" or "interneuron"). For the Zeisel & Muñoz-Manchado et al. data, we applied scDRS to all 3,005 cells and then focused on the C1a1/class I label. We included genes that were inferred from spatial coordinates, we curated differentially expressed genes for each of the six spatial regions (dorsal versus ventral, ventral versus dorsal, proximal versus distal, distal versus proximal, deep versus superficial and superficial versus deep) using the gene expression data from Cembrowski et al. (GEO GSE67403; gene sets in Supplementary Table 24). For each differential gene expression analysis, we selected genes based on false discovery rate (FDR) correction of the transformed data (log10 fold change > 1.0). We used scDRS and these signature gene sets to assign six spatial scores for each cell. For the regression analysis, we separately regressed the scDRS disease scores for each of the seven brain-related traits (and height, a negative control trait) on each of the six spatial scores. We performed marginal regression instead of joint regression for these spatial scores because the inferred spatial scores for opposite regions on the same axis (for example, dorsal versus ventral) were highly collinear (strongly negatively correlated), and the inferred spatial scores for dorsal, proximal and deep regions (which had strong marginal associations to disease) had very low pairwise correlations (average r = 0.10; Supplementary Fig. 23d), suggesting these associations were independent.

Analysis of hepatocytes and metabolic traits. We considered all hepatocytes in the TMS FACs data (1,162 cells). Because the original study clustered all cells from the liver together (limiting the resolution for distinguishing cell states within hepatocytes), we reprocessed and reclustering these cells following the same procedure as we did for the T cells. We further filtered out low-quality cells (proportion of mitochondrial gene counts ≥ 20.5; likely apoptotic or lysing), resulting in 1,102 hepatocytes. We computed the polyplody, pericentral and periportal zonation scores by applying scDRS to published polyplody/zonation signature gene sets (instead of MAGMA putative disease gene sets). We curated signature gene sets for ploidy level, zonation (pericentral/periportal) and putative zonated pathways. We curated four sets of polyplody signatures, including differentially expressed genes for partial hepatocyte (PH) versus pre-PH1 (used for the polyplody score), Cdk1 knockout (case) versus control4, 4n versus 2n hepatocytes3, large versus small hepatocytes10. We curated three sets of diplody signatures, including differentially expressed genes for pre-PH versus PH4, control versus Cdk1 knockout10 and 2n versus 4n hepatocytes10. We curated signature gene sets for pericentral and periportal hepatocytes from Halpern et al.1. We curated gene sets for putative zonated pathways from MSigDB17 (v7.1), including gluconeogenesis (pericentral), xenobiotic metabolism (pericentral), beta oxidation (periportal), cholesterol biosynthesis (periportal), protein secretion (periportal) and glucoseogenesis (periportal) (Supplementary Table 24). For the joint regression analysis of scDRS disease score on ploidy and zonation scores, we regressed the polyplody score out of both the pericentral and periportal score before the joint regression, because the ploidy level confounded both zonation scores. We performed joint regression instead of marginal regression here (unlike the regression analysis in the neuron section), because the polyplody score was positively correlated with the pericentral and periportal scores (unlike the analysis in the neuron section where the three sets of scores had low correlations).

Statistics and reproducibility. We analyzed only existing data sets. No statistical method was used to predetermine sample size. No data were excluded from the analyses. We did not use any study design that required randomization or blinding. We replicated our results by performing the same analyses on additional independent data sets; all attempts at replication were successful.
**Data availability**

We release our data at [https://figshare.com/projects/Single-cell_Disease_Relevance_Score_scDRS](https://figshare.com/projects/Single-cell_Disease_Relevance_Score_scDRS) (118992) (ref. 7) (instructions at [https://github.com/martinjzhang/scDRS](https://github.com/martinjzhang/scDRS)), including GWAS summary statistics of the 74 diseases/trait, TMS FACS scRNA-seq data, reprocessed TMS FACS data (for T cells and hepatocytes), MAGMA and gold-standard gene sets and scDRS results for TMS FACS (disease scores and control scores for the 74 diseases/trait). The 16 scRNA-seq data sets were obtained as follows (15 out of 16 publicly available). The TMS FACS data and TMS droplet assay data were downloaded from the official release [https://figshare.com/articles/dataset/Processed_files_to_use_with_scarpy/#8273102](https://figshare.com/articles/dataset/Processed_files_to_use_with_scarpy/#8273102). The TS FACS data were downloaded from the official release [https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/#14267219](https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/#14267219). The Cane-Gamer & Sosic et al. data were downloaded from [https://www.opentargets.org/projects/effectors](https://www.opentargets.org/projects/effectors). The Nathan et al. data were downloaded from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158769](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158769). The Zeisel & Muñoz-Manchado et al. data were downloaded from [http://linnarssonlab.org/cortex/](http://linnarssonlab.org/cortex/). The Zeisel et al. data were downloaded from [http://mousebrain.org/adolescent/downloads.html](http://mousebrain.org/adolescent/downloads.html). The Habib et al. data and Habib, Avraham-David, & Basu et al. data were downloaded from [https://singlecell.broadinstitute.org/single_cell](https://singlecell.broadinstitute.org/single_cell). The Abyhan et al. data were downloaded from [https://cells.ucsc.edu/human-hippo-axis/](https://cells.ucsc.edu/human-hippo-axis/). The Yao et al. data were downloaded from [https://assets.nemochrome.org/dat-jb2f3y](https://assets.nemochrome.org/dat-jb2f3y). The Zhong et al. data were downloaded from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119212](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119212). The Aizumai et al. data were downloaded from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124395](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124395). The Halpern & Shenhab et al. data were downloaded from [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84498](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84498). The Richter & Delgiannis et al. data (annotated count matrix) were obtained via communication with the authors (raw data publicly available in links in the paper). The Taychameekitchai et al. data are not publicly available and were provided by A.T., P.R. and B.W. MSigDB (v7.1) (ref. 74). The Open Targets data were downloaded from the official website [https://www.opentargets.org/](https://www.opentargets.org/).

**Code availability**

Software implementing scDRS and its downstream applications and a web interface for interactively exploring results of scDRS are available at [https://github.com/martinjzhang/scDRS](https://github.com/martinjzhang/scDRS). Code for generating all results of the paper is available at [https://github.com/martinjzhang/scDRS_paper](https://github.com/martinjzhang/scDRS_paper) (ref. 7).

**References**

63. Geary, R. C. The contiguity ratio and statistical mapping. *incorporated statistician* 5, 115–146 (1954).
64. Wolf, F. A., Angerer, P. & Theis, F. J. Scany: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 1–5 (2018).
65. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.: Ser. B (Methodol.)* 57, 289–300 (1995).
66. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209 (2018).
67. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* 47, 1228–1235 (2015).
68. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* 16, 1289–1296 (2019).
69. Traag, V. A., Wallman, L. & Eck, N. J. V. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 1–12 (2019).
70. Andreatta, M. et al. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat. Commun.* 12, 1–19 (2021).
71. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740 (2011).
72. Zhang, M. J. & Hou, K. scdrs data release 030122. Figshare, [https://doi.org/10.6084/m9.figshare.19312583.v1](https://doi.org/10.6084/m9.figshare.19312583.v1) (2022).
73. Zhang, M. J. & Hou, K. scdrs software v1.0.1. Zenodo, [https://doi.org/10.5281/zenodo.6615722](https://doi.org/10.5281/zenodo.6615722) (2022).
74. Zhang, M. J. & Hou, K. scdrs data analysis code v1.0.1. Zenodo, [https://doi.org/10.5281/zenodo.6615791](https://doi.org/10.5281/zenodo.6615791) (2022).
75. Sakornsakolpat, P. et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* 51, 494–505 (2019).
76. Kato, K. et al. Pulmonary pericytes regulate lung morphogenesis. *Nat. Commun.* 9, 1–14 (2018).

**Acknowledgements**

We thank H. Shi, K. Siewert-Rocks, T. Amariuta, X. Xu, B. J. Streber and A. Gusev for helpful suggestions. This research was conducted using the UK Biobank Resource (application number 16549). This research was funded by National Institutes of Health (NIH) grants U01 HG009379 (A.L.P. and S.R.), R01 MH101244 (A.L.P.), R37 MH107649 (A.L.P.), R01 MH115676 (A.L.P. and B.P.), U01 HG012009 (A.L.P. and S.R.) and R01 HG006399 (M.G.). The authors were also supported by NIH grants P01-AI148102 (S.R.), UC2-AR081023 (S.R.), R01-AR067677 (S.R.), R01MH112152 (M.G.) and R01MH123922 (M.G.), and NIH National Human Genome Research Institute grants K99HG012203 (K.K.D.) and F32 HG011434-03 (K.A.J.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

**Author contributions**

M.J.Z., K.H., B.P. and A.L.P. designed the study and developed statistical methodologies. M.J.Z. and K.H. analyzed the data with assistance from K.K.D., A.O.P. and K.A.J. S.S., M.G. and B.P. were partially supported by NIH grants K99HG012203 (K.K.D.) and F32 HG011434-03 (K.A.J.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

**Competing interests**

The authors declare no competing interests.

**Additional information**

*Extended data* are available for this paper at [https://doi.org/10.1038/s41588-022-01167-z](https://doi.org/10.1038/s41588-022-01167-z).

*Supplementary information* The online version contains supplementary material available at [https://doi.org/10.1038/s41588-022-01167-z](https://doi.org/10.1038/s41588-022-01167-z).

*Correspondence and requests for materials* should be addressed to Martin Jinye Zhang, Kangcheng Hou, Bogdan Pasaniuc or Alkes L. Price.

*Peer review information* Nature Genetics thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

*Reprints and permissions information* is available at [www.nature.com/reprints](http://www.nature.com/reprints).
Extended Data Fig. 1 | Additional null simulations. We performed null simulations for various numbers of putative disease genes (100, 500, 1,000, and 2,000 for the four columns respectively) and various types of genes to randomly sample from: all genes (first row), and top 25% genes with high expression (second row), top 25% genes with high expression variance (third row), top 25% overdispersed genes (fourth row). We considered two additional versions of scDRS: scDRS-bin-gs (binary gene sets instead of MAGMA z-score gene weights) and scDRS-adj-ctp (adjusting for cell-type proportion). For scDRS-adj-ctp, we simulated random biased gene sets (high-mean/high-variance/overdispersed) based on the balanced data (inversely weighting cells by cell-type proportion) to better match the model assumption, namely testing for excess expression relative to cells in the balanced data. In each panel, the x-axis denotes theoretical $-\log_{10} p$-value quantiles and the y-axis denotes actual $-\log_{10} p$-value quantiles for different methods. The 3 versions of scDRS produced well-calibrated p-values in most settings and suffered slightly inflated type I error in panels o,p, possibly because it is hard to match a large number of overdispersed putative disease genes using the remaining set of genes. In comparison, all other methods are less well-calibrated and are particularly problematic when the numbers of putative disease genes are small. Error bars denote 95% confidence intervals around the mean of 100 simulation replicates.
Extended Data Fig. 2 | Additional causal simulations. We performed three sets of causal simulations: (1) varying effect size from 5% to 50% while fixing 25% overlap (first column), (2) varying level of overlap from 5% to 50% while fixing 25% effect size (second column), (3) assigning the 528 B cells in the subsampled data to be causal (instead of the 500 randomly selected cells; varying effect size while fixing 25% overlap; third column). We report the power (first row), FDR (second row), and AUC for classifying causal from non-causal cells based on the p-values (third row). scDRS outperformed other methods under all metrics. Error bars denote 95% confidence intervals around the mean of 100 simulation replicates.
Extended Data Fig. 3 | Complete results for cell type-level disease associations for 74 diseases/traits and TMS FACS 120 cell types. Each row represents a disease/trait and each column represents a cell type (number of cells in parentheses). Heatmap colors denote the proportion of significantly associated cells (FDR < 0.1 across all cells for a given disease). Squares denote significant cell-type–disease associations (FDR < 0.05 across all pairs of the 120 cell types and 74 diseases/traits; 597 significant pairs; MC test; Methods). Cross symbols denote significant heterogeneity in association with disease across individual cells within a given cell type (FDR < 0.05 across all pairs; 273 significant pairs; MC test; Methods). Heatmap colors and cross symbols are omitted for cell-type–disease pairs with nonsignificant cell-type–disease associations. Within the blood/immune block (40 cell types and 21 diseases/traits), 136 of 264 cell-type–disease pairs with significant association also had significant heterogeneity. Within the brain block (11 cell types and 21 diseases/traits), 64 of 133 cell-type–disease pairs with significant association also had significant heterogeneity. Within the other block (69 cell types and 32 diseases/traits), 54 of 146 cell-type–disease pairs with significant association also had significant heterogeneity. We discuss the results for FEV1/FVC. We identified 20 cell types associated with FEV1/FVC (FDR < 0.05), including 5 lung cell types and 15 cell types from other tissues. They can be categorized into 5 sets of associations: (1) type II pneumocyte (2) skin-related cells (3) smooth muscle cells (4) fibroblast-and-MSC-like cells (5) pericycle-like cells. The first 4 sets of associations are consistent with a previous work75. The 5th set of pericycle associations is also plausible because pericytes are known to regulate lung morphogenesis76. We note that the cell type associations from the lung are more likely to be causal and those from the other tissues are more likely tagging the causal cell types due to shared expression. Numerical results are reported in Supplementary Table 12.
Extended Data Fig. 4 | Comparison of cell type-level disease association results between TMS FACS and TMS droplet (different technologies), TS FACS (different species). (a–c) Results for disease association at the cell type-level for TMS FACS, TMS droplet, and TS FACS for diseases and cell types in the blood/immune block (upper left) and the other cell types/diseases block (lower right) in Fig. 3 (TMS droplet and TS FACS do not contain brain data; Supplementary Tables 6, 7). The plotting style is the same as Fig. 3. Heatmap colors for each cell-type–disease pair denote the proportion of significantly associated cells (FDR < 0.1); squares denote significant cell-type–disease associations (FDR < 0.05); and cross symbols denote significant heterogeneity in association with disease across individual cells within a given cell type (FDR < 0.05). Heatmap colors (>10% of cells associated) and cross symbols are omitted for cell-type–disease pairs with nonsignificant cell-type–disease associations via MC test. We matched each TMS FACS cell type using the closest cell type in the TMS droplet and TS FACS data; unmatched cell types were colored in grey. (d) Overlap of significant cell-type–disease associations between TMS FACS and TMS droplet (P = 2.8 × 10^{-24}, two-sided Fisher’s exact test). (e) Pearson’s correlation of −log_{10} p-values for cell-type–disease associations between TMS FACS and TMS droplet. (f) Overlap of significant cell-type–disease associations between TMS FACS and TS FACS (P = 1.3 × 10^{-7}, two-sided Fisher’s exact test). (g) Pearson’s correlation of −log_{10} p-values for cell-type–disease associations between TMS FACS and TS FACS. We determined that the results are highly consistent between TMS FACS and TMS droplet, and are reasonably consistent between TMS FACS and TS FACS. Our method is underpowered in the TS FACS data, possibly due to the smaller sample size (27K cells in TS FACS versus 110K cells in TMS FACS). The current TS FACS data corresponds to the initial data release and there will likely be more cells in future releases.20
Extended Data Fig. 5 | Optimizing parameters of scDRS based on expected and unexpected control cell types across 20 traits. We considered different versions of scDRS by varying methods for selecting (1) putative disease genes (2) weights for the disease genes (3) MAGMA window size. We considered 6 methods for selecting putative disease genes, 4 methods for selecting gene weights, and 3 MAGMA gene window sizes (Supplementary Note). We applied each version of scDRS to the subsampled TMS FACS data (20 repetitions with 10K cells each) and a curated set of 20 traits with expected and unexpected disease-critical cell types (Supplementary Table 15). For a given scDRS version and a given trait, we computed the t-statistic between cells from the expected and unexpected cell types, and divided it by the average t-statistics of results of the given trait from all data sets and all scDRS versions to correct for trait-specific baseline. We evaluated each version by first computing the mean and SE of the normalized t-statistics for a given trait across the 20 repetitions and then combining the estimates across the 20 traits via random-effect meta-analysis. We compared the performance of a pair of scDRS versions by applying the same procedure to the difference of the normalized t-statistics between the two versions. (a) Varying gene selection methods while fixing other parameters as the default. (b) Varying gene weighting methods while fixing other parameters as the default. (c) Varying MAGMA gene window size while fixing other parameters as the default. The default version was denoted in dark blue. Error bars denote 95% confidence intervals around the mean based on meta-analysis across 20 subsampled data sets and 20 traits, using procedures as described above. * denotes \( P < 0.05 \) and ** denotes \( P < 0.001 \) for significant differences relative to the default configuration; one-sided tests based on the estimated mean and CIs. Numerical results are reported in Supplementary Table 16.
Extended Data Fig. 6 | Numbers of overlapping genes (upper triangle) and correlations of the scDRS disease scores across all TMS FACS cells (lower triangle) between the 26 autoimmune, brain, and metabolic traits analyzed in the main paper. Traits are ordered via hierarchical clustering of the scDRS score correlation and the clustering dendrogram was provided. The level of gene set overlap is moderate. scDRS disease score correlations distinguish diseases/trait from the 3 categories as well as subgroups of diseases/trait in the same category.
Extended Data Fig. 7 | Additional results on disease gene prioritization. (a–j) Comparison to alternative disease gene prioritization methods for the 10 autoimmune diseases. The first row shows levels of excess overlap between the prioritized disease genes and the gold standard gene sets while the second row shows the corresponding $-\log_{10}$ p-values for excess overlap. Each dot corresponds to a disease, the y-axis shows results for the proposed prioritization method (correlating gene expression levels with the scDRS disease score across all TMS FACS cells), and the x-axis shows results from comparison methods, including (from left to right) top 1,000 MAGMA genes, top 1,000 genes specifically expressed in T cells (vs. the rest of cells in TMS FACS), prioritization based on correlation across T cells (instead of all TMS FACS cells), prioritization based on correlation across CD4$^+$ T cells (instead of all TMS FACS cells), and prioritization based on correlation across CD8$^+$ T cells (instead of all TMS FACS cells). (k–l) Overlap with drug target genes for 27 diseases. (m–n) Overlap with Mendelian disease genes for 45 diseases. The median ratio of $-\log_{10}$ p-values and (excess overlap − 1) between the y- and x-values (median of ratios) was provided in the figure title. P-values are based on two-sided Fisher’s exact tests.
Extended Data Fig. 8 | Complete results of correlations between scDRS disease scores and inferred spatial coordinates across CA1 pyramidal neurons in 7 single-cell data sets (extending results in Fig. 5b). (a) Results for regressing the scDRS disease scores against the inferred spatial coordinates for each disease/trait and each inferred spatial coordinate. Color represents the t-statistics and stars represent significant associations (* denotes $P < 0.05$ and ** denotes $P < 0.005$, one-sided MC test; Methods). For clarification, Zeisel & Muñoz-Manchado et al. refers to the data from Zeisel & Muñoz-Manchado et al. 2015 Science and Zeisel et al. refers to the data from Zeisel et al. 2018 Cell. (b) Summary of results in panel a. Heatmap color represent the average t-statistics across the 7 brain-related diseases/traits (excluding height) for each data set and stars represent significant associations by combining p-values across data sets using Fisher’s combined probability test. (c) Summary of the association between brain-related diseases and the inferred spatial coordinates for the mouse and human data sets in panel b.
Extended Data Fig. 9 | Complete results of joint regression analysis for GWAS metabolic traits and putative zonated metabolic processes across the 6 data sets (extending results in Fig. 5d). (a, b) Results for the 9 metabolic traits and height, a negative control trait. The polyploidy score (panel a) and both the pericentral and periportal score (panel b) were consistently associated with the 9 metabolic traits across the data sets. The strong association ($P < 0.005$) between the pericentral score and height in the Aizarani et al. data may be because we inferred the pericentral score using mouse gene signatures, which are less conserved in human (as also mentioned in the original paper61). (c, d) Results for the 8 metabolic pathways. Overall, as shown in panel d, the pericentral score was associated with pericentral-specific pathways (first 4 rows) while the periportal score was associated with periportal-specific pathways (last 4 rows). * denotes $P < 0.05$ and ** denotes $P < 0.005$ based on one-sided MC tests.
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.

- [ ] 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
We did not collect data for this study.

Data analysis

Code is available at https://github.com/martinjzheng/scDRS and https://github.com/martinjzheng/scDRS_paper; customized software scdtrs (v1.0.1) was used. We used public softwares MAGMA (v1.08; https://cg.cnic.nl/software/magma) and LDSC (v1.0.1; https://github.com/bulik/lrsc). We used python and public Python packages numpy (v1.21.5), scipy (v1.5.1), pandas (v1.3.5), scpanly (v1.6.0), anddata (v0.7.8), scikit-misc (v0.1.3), statsmodels (v0.11.1), matplotlib (v3.3.0), seaborn (v0.11.0), gseapy (v0.10.2). We used R package VAM (v0.5.1).

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

We release our data at https://figshare.com/projects/Single-cell_Disease_Relevance_Score_scDRS_/118902 (instructions at https://github.com/martinjzheng/scDRS, including GWAS summary statistics of the 74 diseases/trait, TMS FACs scRNA-seq data, reprocessed TMS FACs data for T cells and hepatocytes, MAGMA and gold standard gene sets, and scDRS results for TMS FACs (disease scores and control scores for the 74 diseases/trait). The 16
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.

- [X] 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-list.pdf

Life sciences study design

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

Sample size: We did not collect new data and the sample size corresponds to the actual sample size of the existing data sets.

Data exclusions: No data was excluded from the analyses.

Replication: We replicated our results by applying the same analysis procedures on additional publicly available data sets; all attempts at replication were successful.

Randomization: We did not perform randomization. Covariates such as age, sex, genotyping array, and ancestry PCs were controlled for the GWAS summary statistics in the original studies. We controlled for data-set-specific covariates, including age, sex, number of expressed genes, and mouse ID, for applying scDRS to the 16 single-cell data sets.

Blinding: We did not apply blinding because we only analyzed existing data sets.

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

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

Methods

- [ ] Involved in the study
  - [X] ChiP-seq
  - [ ] Flow cytometry
  - [ ] MRI-based neuroimaging