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

Martin Jinye Zhang[1,2,*], **Kangcheng Hou**[3-5,*], **Kushal K. Dey**[1,2], **Karthik A. Jagadeesh**[1,2], **Kathryn Weinand**[2,6-9], **Saori Sakaue**[2,6-9], **Aris Taychameekiatchai**[10,11], **Poorvi Rao**[10], **Angela Oliveira Pisco**[12], **James Zou**[12-14], **Bruce Wang**[10], **Michael Gandal**[15-17], **Soumya Raychaudhuri**[2,6-9,18], **Bogdan Pasaniuc**[3-5,†], **and Alkes L. Price**[1,2,19,†]

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[3]Bioinformatics Interdepartmental Program, University of California, Los Angeles, Los Angeles, CA, USA
[4]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[5]Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[6]Center for Data Sciences, Brigham and Women's Hospital, Boston, MA, USA
[7]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[8]Division of Rheumatology, Inflammation, and Immunity, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[9]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[10]Department of Medicine and Liver Center, University of California San Francisco, San Francisco, CA, USA
[11]Developmental and Stem Cell Biology Graduate Program, University of California San Francisco, San Francisco, CA, USA
[12]Chan Zuckerberg Biohub, San Francisco, CA, USA
[13]Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA
[14]Department of Biomedical Data Science, Stanford University, Palo Alto, CA, USA
[15]Department of Psychiatry, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[16]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[17]Program in Neurobehavioral Genetics, Semel Institute, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[18]Versus Arthritis Centre for Genetics and Genomics, Centre for Musculoskeletal Research, Manchester Academic Health Science Centre, The University of Manchester, Manchester, UK
[19]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[*]Equal contribution: M. J. Zhang jinyezhang@hsph.harvard.edu, K. Hou houkc@ucla.edu
[†]Co-senior authors: A. L. Price aprice@hsph.harvard.edu, B. Pasaniuc pasaniuc@ucla.edu

ABSTRACT

Gene expression at the individual cell-level resolution, as quantified by single-cell RNA-sequencing (scRNA-seq), can provide unique insights into the pathology and cellular origin of diseases and complex traits. Here, we introduce single-cell Disease Relevance Score (scDRS), an approach that links scRNA-seq with polygenic risk of disease at individual cell resolution; scDRS identifies individual cells that show excess expression levels for genes in a disease-specific gene set constructed from GWAS data. We determined via simulations that scDRS is well-calibrated and powerful in identifying individual cells associated to disease. We applied scDRS to GWAS data from 74 diseases and complex traits (average N =341K) in conjunction with 16 scRNA-seq data sets spanning 1.3 million cells from 31 tissues and organs. At the cell type level, scDRS broadly recapitulated known links between classical cell types and disease, and also produced novel biologically plausible findings. At the individual cell level, scDRS identified subpopulations of disease-associated cells that are not captured by existing cell type labels, including subpopulations of CD4$^+$ T cells associated with inflammatory bowel disease, partially characterized by their effector-like states; subpopulations of hippocampal CA1 pyramidal neurons associated with schizophrenia, partially characterized by their spatial location at the proximal part of the hippocampal CA1 region; and subpopulations of hepatocytes associated with triglyceride levels, partially characterized by their higher ploidy levels. At the gene level, we determined that genes whose expression across individual cells was correlated with the scDRS score (thus reflecting co-expression with GWAS disease genes) were strongly enriched for gold-standard drug target and Mendelian disease genes.

Introduction

The mechanisms through which risk variants identified by genome-wide association studies (GWASs) impact critical tissues and cell types are largely unknown[1,2]; identifying these tissues and cell types is central to our understanding of disease etiologies and will inform efforts to develop therapeutic treatments[3]. Single-cell RNA sequencing (scRNA-seq) has emerged as the tool of choice for measuring gene abundances at single-cell resolution[4,5], providing an increasingly clear picture of which genes are active in which cell types and also being able to identify novel 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 disease[6–8], thus providing finer resolution than studies using bulk transcriptomic data[9–12].

Previous studies integrating scRNA-seq with GWAS have largely focused on predefined cell type annotations (e.g., 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 GWAS[6–8]. However, this approach overlooks the considerable heterogeneity within cell types that has been reported in studies of scRNA-seq data alone[13–18]; the underlying methods (e.g., Seurat cell-scoring function[15], Vision[16], and VAM[18]) have sought to explain transcriptional heterogeneity in scRNA-seq data by scoring cells based on predefined gene sets such as pathway gene sets, 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 (e.g., individual cells) with polygenic signals from disease GWAS has considerable potential to produce new biological insights.

Here, we introduce *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 GWAS, using an appropriately matched empirical null distribution to estimate well-calibrated p-values. To our knowledge, scDRS is the first method to associate individual cells in scRNA-seq data to disease GWAS. We performed extensive simulations to assess the calibration and power of scDRS. We then applied scDRS to 74 diseases and complex traits (average GWAS N =341K) in conjunction with 16 scRNA-seq data sets (including the Tabula Muris Senis (TMS) mouse cell atlas[19]), assessing cell type-disease associations and within-cell type association heterogeneity, including heterogeneity of T cells in their association with inflammatory bowel disease (IBD) and other autoimmune diseases, neurons in their association with schizophrenia (SCZ) and other brain-related diseases/traits, and hepatocytes in their association with triglyceride levels (TG) and other metabolic traits; we analyzed a broader set of scRNA-seq data sets to provide validation across species (human vs. mouse) and across sequencing platforms, and to 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 GWAS to associate individual cells to disease, by assessing the excess expression of putative disease genes from GWAS. scDRS consists of three steps (Fig. 1, Methods, and Supp. Note). First, scDRS constructs a set of putative disease genes from GWAS summary statistics using MAGMA[20], an existing gene scoring method (top 1,000 MAGMA genes; see Methods for other choices evaluated). 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 inversely weighted by its gene-specific technical noise level in the single-cell data, estimated via modeling the mean-variance relationship across genes[18,21] (alternative choices of cell scores are evaluated in Methods). To determine statistical significance, scDRS also generates 1,000 sets of cell-specific *raw control scores* at Monte Carlo (MC) samples of matched control gene sets (matching the gene set size, mean expression, and expression variance of the putative disease genes); cell-specific MC p-values are defined as the proportion of the 1,000 raw control scores for a given cell exceeding the raw disease score for that cell[22]. Third, scDRS approximates the ideal MC p-values (obtained using $\gg$1,000 MC samples) by pooling control scores across cells. Specifically, it normalizes the raw disease score and raw control scores for each cell (producing the *normalized disease score* and *normalized control scores*), and then computes cell-level p-values based on the empirical distribution of the pooled normalized control scores across all control gene sets and all cells; this approximation relies on the assumption that the raw control score distributions (across the 1,000 control gene sets, for each cell) are from the same parametric distribution (e.g., normal distributions with different parameters, such that the normalization procedure can align these distributions across cells), a reasonable assumption when the disease gene set is neither too small nor too large (e.g., $>$50 genes and $<$20% of all genes; Methods). Importantly, scDRS does not use cell type or other cell-level annotations, although these annotations can be of value when interpreting its results.

scDRS outputs individual cell-level p-values, normalized disease scores, and 1,000 sets of normalized control scores (referred to as "disease scores" and "control scores" in the rest of the paper) that can be used for a wide range of downstream applications (Methods). Here, we focus on three downstream analyses. First, we perform *cell type-level* analyses to associate predefined cell types to disease and assess heterogeneity in association to disease across cells within a predefined cell type. Second, we perform *individual cell-level* analyses to associate individual cells to disease and correlate individual cell-level variables to the scDRS disease score. Third, we perform *gene-level* analyses to prioritize disease-relevant genes whose expression is correlated with the scDRS disease score, reflecting co-expression with genes implicated by disease GWAS. In the cell type-level analyses, we compute a single MC p-value across the focal set of cells by comparing the diseases scores to the 1,000 sets of control scores (Methods), avoiding the assumption that the cells are independent—a strong assumption in scRNA-seq analyses, e.g., when analyzing cells in the same cluster that are dependent due to the clustering process.

We analyzed publicly available GWAS summary statistics of 74 diseases and complex traits (average N=341K; Supp. Table 1) in conjunction with 16 scRNA-seq or single-nucleus RNA-seq (snRNA-seq) data sets spanning 1.3 million cells from 31 tissues and organs from mouse (*mus musculus*) and human (*homo sapiens*) (Supp. Table 2; 15 out of 16 data sets publicly available; Data Availability). The single-cell data sets include two mouse cell atlases from the Tabula Muris Senis (TMS)[19] collected using different technologies (fluorescence-activated cell sorting followed by Smart-seq2 amplification[23] for the TMS FACS data and 10x microfluidic droplet capture and amplification[24] for the TMS droplet data), the unpublished Tabula Sapiens (TS) human cell atlas[25], and other data sets focusing on specific tissues containing well-annotated cell types and cell states. We focused on the TMS 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-seq2); 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 (see Bryois et al.[8], other studies[6,7,9,12,26], and Discussion).

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)[15], Vision[16], and VAM[18]. To our knowledge, VAM is the only method for scoring individual cells that provides cell-level association p-values; Seurat and Vision provide quantitative cell-level scores that we transformed to p-values based on the standard normal distribution (Methods).

First, we evaluated each method in null simulations in which 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. scDRS and Seurat produced well-calibrated p-values, Vision suffered slightly inflated type I error, and VAM suffered severely inflated type I error (Fig. 2A and Supp. Table 9). The slight miscalibration of Vision may be due to the mismatch between the normal distribution used for computing p-values and the actual null distribution of the cell-level scores. The poor calibration of VAM may be because it uses a permutation-based test that assumes independence between genes under the null, an assumption that is likely to be violated in scRNA-seq data. We performed 3 secondary analyses pertaining to null simulations. First, we considered other numbers of putative disease genes (100 or 500, instead of 1,000). We determined that scDRS remained well-calibrated, VAM continued to suffer from severely inflated type I error, and Seurat and Vision suffered increased type I error at 500 genes and severely inflated type I error at 100 genes (Supp. Fig. 2A-C). Second, we considered biased sets of putative disease genes (randomly selected from genes with high mean expression, genes with high expression variance, or overdispersed genes (genes with high expression variance but normal levels of technical noise[27])). We determined that scDRS remained reasonably well-calibrated, VAM continued to suffer from severely inflated type I error, and Seurat and

Vision were conservative for high-expression genes and high-variance genes but suffered inflated type I error for overdispersed genes (Supp. Fig. 2D-L). Third, we assessed calibration of our MC test for *cell type*-disease association based on the output of scDRS, using the same subsampled data (and 1,000 putative disease genes). We confirmed that this test was well-calibrated (Supp. Table 10).

Next, we evaluated scDRS, Seurat and Vision in causal simulations in which 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 times for different simulations) across the 1,000 causal disease genes, and randomly selected 1,000 putative disease genes (provided as input to each method) with 25% overlap with the 1,000 causal disease genes. We determined that scDRS attained higher power than Seurat and Vision to detect individual cell-disease associations at FDR$<$0.1 (Fig. 2B and Supp. Table 11); the improved power of scDRS may be due to its incorporation of gene-specific weights that down-weight genes with higher levels of technical noise. We performed 4 secondary analyses pertaining to causal simulations. First, we considered other levels of overlap between the 1,000 causal genes and 1,000 putative disease genes (from 5% to 50%, instead of 25%). We determined that scDRS continued to attain higher power than Seurat and Vision (Supp. Fig. 3B). Second, we considered causal simulations in which we selected all 528 B cells in the subsampled data as causal cells, instead of randomly selecting 500 causal cells. We observed a similar improvement in power of scDRS over Seurat and Vision (Supp. Fig. 3C). Third, we computed the actual FDR for each method in each of the above causal simulations. We determined that scDRS attained well-calibrated FDR across all parameter settings, whereas Seurat and Vision suffered from inflated type I error at smaller effect sizes ($\leq$1.1 times higher expression for causal cells) and lower levels of overlap ($\leq$15%) (Supp. Fig. 3D-F). Fourth, since Seurat and Vision were not initially designed to produce calibrated p-values, we also evaluated each method's area under the receiver operating characteristic curve (AUC) in distinguishing causal from non-causal cells. We determined that scDRS attained more accurate classification than Seurat and Vision under this metric (Supp. Fig. 3G-I).

In summary, scDRS is well-calibrated in null simulations and attains higher power to detect causal cells than Seurat and Vision in causal simulations.

Results across 120 TMS cell types for 74 diseases and complex traits

We analyzed GWAS data from 74 diseases and complex traits (average N=341K; Supp. Table 1) in conjunction with the TMS FACS data with 120 cell types (cells from different tissues were combined for a given cell type; Supp. Table 3). We first report scDRS results for individual cells aggregated at the cell type level; individual cell-level results are discussed in subsequent sections. Results for a subset of 20 representative cell types and 21 representative diseases/traits are reported in Fig. 3 (complete results in Supp. Fig. 4 and Supp. Table 12). Within this subset, scDRS identified 70 associated cell type-disease pairs (FDR$<$0.05; squares in Fig. 3) and detected significant heterogeneity in association with disease across individual cells within cell type for 37 of these 70 associated cell type-disease pairs (FDR$<$0.05; cross symbols in Fig. 3; 247 of 577 associated cell type-disease pairs across all pairs of 120 cell types and 74 diseases/traits). We also report the proportion of significantly associated individual cells for each cell type-disease pair (FDR$<$0.1, a less stringent threshold as false positive associations of individual cells are less problematic and we do not focus on the results for any one specific cell; heatmap colors in Fig. 3). We note these associated cell type-disease pairs (and individual cell-disease associations discussed in subsequent sections) may reflect indirect tagging of causal cell types rather than direct causal associations, analogous to previous work (see Discussion).

For cell type-disease associations, as expected, scDRS broadly associated blood/immune cell types with blood/immune-related diseases/traits, brain cell types with brain-related diseases/traits, and other cell types with other diseases/traits (block-diagonal pattern in Fig. 3). Interestingly, there were 3 exceptions to the block-diagonal pattern, involving 4 diseases/traits (Fig. 3). First, hepatocytes (in addition to proerythroblasts) were associated with red blood cell distribution width (RDW), possibly because liver malfunction affects RDW. This association is consistent with the observation of increased RDW values in patients with liver disease[28], but to our knowledge has not been reported in previous genetic studies. Second, ventricular myocytes (in addition to immune cell types) were associated with lymphocyte count. This association is consistent with the prognostic value of relative lymphocyte concentration in patients with symptomatic heart failure[29], but to our knowledge has not been reported in previous genetic studies. Third, pancreatic beta cells (in addition to brain cell types) were associated with SCZ and body mass index (BMI); risk variants for SCZ and BMI are reported to be enriched in pancreatic islet-specific epigenomic regulatory elements[30,31].

We discuss 3 main findings for the blood/immune-related diseases/traits (upper left block in Fig. 3). First, different blood/immune cell types were associated with the corresponding blood cell traits, including proerythroblasts with RDW, classical monocytes with monocyte count, and adaptive immune cells with lymphocyte count. We detected significant heterogeneity across cells for the proerythroblast-RDW association, which may correspond to erythrocytes at different differentiation stages[32] (see Supp. Fig. 5). Second, immune cell types were associated with immune diseases, including

dendritic cells, CD4$^+$ α/β T cells, CD8$^+$ α/β T cells, and/or regulatory T cells with rheumatoid arthritis (RA), multiple sclerosis (MS), and IBD, consistent with previous findings[12,33]. We detected significant heterogeneity across cells for many of these cell type-disease associations, consistent with the known diversity within the T cell population (see "Heterogeneous subpopulations of T cells associated with autoimmune disease" section). Third, granulocyte monocyte progenitors (GMP) were strongly associated with MS, highlighting the role of myeloid cells in MS[34,35].

We discuss 2 main findings for brain-related diseases/traits (middle block in Fig. 3). First, neuronal cell types, including medium spiny neurons (MSNs), interneurons, and neurons (neuronal cells with undetermined subtypes), were associated with schizophrenia (SCZ), major depressive disorder (MDD), college education (ECOL), and several other brain-related traits; the role of MSN in SCZ, MDD and ECOL is supported by previous genetic studies[8,26,36]. We detected significant heterogeneity across neurons in their association with most brain-related diseases/traits (see "Heterogeneous subpopulations of neurons associated with brain-related diseases and traits" section). Second, oligodendrocytes, oligodendrocyte precursor cells (OPCs) were also associated with multiple brain-related diseases/traits. These associations are less clear in existing genetic studies[6,8,26,37], but are biologically plausible, consistent with the increasingly discussed role of oligodendrocyte lineage cells in brain diseases/traits: the differentiation and myelination of oligodendrocyte lineage cells are important to maintain the functionality of neuronal cells[38,39]. We detected significant heterogeneity across OPCs in their association with many brain-related diseases/traits, consistent with recent evidence of functionally diverse states of OPCs[40], traditionally considered to be a homogeneous population (see Supp. Fig. 6).

We discuss 3 main findings for other diseases/traits (lower right block in Fig. 3). First, hepatocytes were associated with several metabolic traits including TG and testosterone (TST) (and other lipid traits; Supp. Fig. 4); hepatocytes are known to play an important role in metabolism[41]. We detected significant heterogeneity across hepatocytes in their association with TG and TST (see "Heterogeneous subpopulations of hepatocytes associated with metabolic traits" section). Second, pancreatic beta cells were associated with glucose and type 2 diabetes (T2D). We detected significant heterogeneity across pancreatic beta cells in their association with glucose, which could be due to different insulin-producing beta cell states[42,43] (see Supp. Fig. 7). Third, pancreatic PP cells (in addition to chondrocytes and bladder cells) were associated with bone mineral density heel T-score (BMD-HT), consistent with the fact that osteoblast and osteoclast cells (which form and reabsorb bones, respectively) are regulated by pancreatic polypeptide[44], which are produced by pancreatic PP cells. To our knowledge, this finding has not been reported in previous genetic studies.

We performed 2 secondary analyses to assess robustness of these results. First, we performed the same analyses on a mouse cell atlas assayed with a different technology (TMS droplet) and a human cell atlas assayed using the same technology (TS FACS) to provide comparisons of the results across technologies and across species. Results are reported in Supp. Fig. 8. We determined that the associations are highly consistent across technologies (r =0.90 for association $-\log_{10}$ p-value across cell type-disease pairs; $P = 1.7 \times 10^{-26}$, Fisher's exact test) and reasonably consistent across species (r =0.65 for association $-\log_{10}$ p-value; $P = 7.5 \times 10^{-7}$, Fisher's exact test). Second, we analyzed the same 120 TMS FACS cell types and 74 diseases using LDSC-SEG[12] for comparison purposes (Methods). Results are reported in Supp. Fig. 9. We determined that the cell type-disease associations identified by the two methods are highly consistent (r =0.65 for association $-\log_{10}$ p-value; $P = 8.7 \times 10^{-295}$, Fisher's exact test). Interestingly, `scDRS` identified some biologically plausible associations that were missed by LDSC-SEG, including pancreatic PP cells and BMD-HT (`scDRS` FDR=0.046 vs. LDSC-SEG FDR=1.000; see ref.[44]), GMPs and MS (FDR 0.020 vs. 0.270; see ref.[34,35]), and OPCs and MDD (FDR 0.020 vs. 0.100; see ref.[45]).

We performed 3 secondary analyses to assess alternative versions of `scDRS`. First, we considered using an unweighted average for the cell-level score (without weighting genes by gene-specific technical noise). We determined that our default weighted score achieved moderately higher power than the unweighted score in detecting disease-associated cells (Supp. Fig. 10A). Second, we considered an overdispersion score capturing both overexpression and underexpression of putative disease genes in the relevant cell population (whereas the default weighted score only captures overexpression; Methods). We determined our default weighted score achieved substantially higher power than the overdispersion score, suggesting that most putative disease genes are overexpressed in the relevant cell population (Supp. Fig. 10B). Third, we investigated other choices of MAGMA gene window size for mapping SNPs to genes (0 kb or 50 kb, instead of 10 kb) and other numbers of putative disease genes (100, 500, or 2,000, instead of 1,000) and determined that results were not sensitive to the choices of these parameters (Supp. Fig. 11 and Supp. Table 16), although the optimal number of putative disease genes was significantly correlated with trait polygenicity[46] (r =0.54, P =0.011, Supp. Fig. 12).

In summary, cell type-disease associations identified by `scDRS` recapitulate known biology but also produced novel, biologically plausible findings. Many cell type-disease associations were heterogeneous across individual cells, strongly motivating analysis at the level of individual cells instead of cell types; further investigation of three examples of heterogeneity is provided in the remaining sections.

Heterogeneous subpopulations of T cells associated with autoimmune disease

We sought to further understand the heterogeneity across T cells in the TMS FACS data in their association with autoimmune diseases (Fig. 3). We jointly analyzed all T cells in the TMS FACS data (3,769 cells, spanning 15 tissues). Since the original study clustered cells from different tissues separately[19], we reclustered these T cells, resulting in 11 clusters (Fig. 4A; Methods); we verified that cells from different tissues, age, or sex clustered together (Supp. Fig. 13). We considered 10 autoimmune diseases: IBD, Crohn's disease (CD), ulcerative colitis (UC), RA, MS, AIT, hypothyroidism (HT), eczema, asthma (ASM), and respiratory and ear-nose-throat diseases (RR-ENT) (Supp. Table 1); we also considered height as a negative control trait.

We focused on individual cells associated with IBD, a representative autoimmune disease (Fig. 4B; results for the other 9 autoimmune diseases and height are reported in Supp. Fig. 14). The 357 IBD-associated cells (FDR<0.1) formed subpopulations of 4 of the 11 T cell clusters; we characterized these subpopulations based on marker gene expression and overlap of specifically expressed genes in each subpopulation with T cell signature gene sets (Methods). First, the subpopulation of 120 IBD-associated cells in cluster 3 (labeled as "Treg") had high expression of regulatory T cell (Treg) marker genes (e.g., *FOXP3*$^+$, *CTLA4*$^+$, *LAG3*$^+$; Supp. Fig. 15A), and their specifically expressed genes significantly overlapped with Treg signatures ($P = 3.9 \times 10^{-7}$, Fisher's exact test; Supp. Fig. 15B), suggesting these cells had Treg immunosuppressive functions. Interestingly, these 120 IBD-associated cells were non-randomly distributed in cluster 3 on the UMAP plot ($P < 0.001$, MC test; Methods). Genes specifically expressed in these IBD-associated cells are preferentially enriched (compared to the 509 non-IBD-associated cells in the same cluster) in pathways defined by response to lipopolysaccharide, T helper cell differentiation, and tumor necrosis factor-mediated signaling (Supp. Fig. 15D); these pathways are closely related to inflammation, a distinguishing feature of IBD[47]. Second, the 75 IBD-associated cells in cluster 4 (*IL1RL1*$^+$ *KLRG1*$^+$ *AREG*$^+$; labeled as "Effector-like Treg") were characterized as effector-like Tregs, which have active functions in Treg differentiation, immunosuppression, and tissue repair[48]; to our knowledge, this subpopulation of effector-like Tregs has only been studied in the context of lung cancer[48], but their role in IBD is not surprising given the strong connection between Treg functions and IBD[49,50]. Third, the 61 IBD-associated cells in cluster 5 (*IL23R*$^+$ *RORC*$^+$ *IL17A*$^+$; labeled as "Th17-like") were characterized as having T helper 17 (Th17) proinflammatory functions. Interestingly, drugs targeting *IL17A* (secukinumab and ixekizumab) have been considered for treatment of IBD but their use was associated with the onset of paradoxical effects (disease exacerbation after treatment with a putatively curative drug); the mechanisms underlying these events are not well understood[51]. Fourth, the 38 IBD-associated cells in cluster 9 (*IFNG*$^+$ *GZMB*$^+$ *FASL*$^+$; labeled as "Effector-like CD8$^+$") were characterized as having effector CD8$^+$ (cytotoxic) T cell functions. Overall, these findings are consistent with previous studies associating subpopulations of effector T cells to IBD, particularly Tregs and Th17 cells[47,49,50,52].

We investigated whether the heterogeneity of T cells in association with autoimmune diseases was correlated with T cell effectorness gradient, a continuous classification of T cells defined by naive T cells on one side (immunologically naive T cells matured from the thymus) and effector T cells on the other (differentiated from naive T cells upon activation and capable of mediating effector immune responses); we hypothesized that such a correlation might exist given the effector-like T cell subpopulations associated to IBD above. Following a recent study[53], we separately computed the effectorness gradients for CD4$^+$ T cells (1,686 cells) and CD8$^+$ T cells (2,197 cells) using pseudotime analysis[54] (Supp. Fig. 16A,B; Methods), and confirmed that the inferred effectorness gradients were significantly negatively correlated with naive T cell signatures and positively correlated with memory and effector T cell signatures (Supp. Fig. 16C,D; Methods). We assessed whether the CD4 (resp., CD8) effectorness gradient was correlated with `scDRS` disease scores for IBD or other autoimmune diseases, across CD4$^+$ T cells (resp., CD8$^+$ T cells). Results are reported in Fig. 4C and Supp. Table 17. We determined that the CD4 effectorness gradient was strongly associated with IBD, CD, UC, and AIT ($P < 0.005$, MC test; 18%-31% of variance in `scDRS` disease score explained by CD4 effectorness gradient), weakly associated with HT, Eczema, ASM, and RR-ENT ($P < 0.05$, MC test; 6%-10% variance explained), but not significantly associated with RA or MS. This implies that these autoimmune diseases are associated with more effector-like CD4$^+$ T cells. We also determined that the CD8 effectorness gradient was weakly associated with IBD and CD ($P < 0.05$, MC test; 10%-11% variance explained), but not significantly associated with the other autoimmune diseases, suggesting that CD4$^+$ effector T cells may be more important than CD8$^+$ effector T cells for these diseases. Notably, after conditioning on the 11 cluster labels, the associations with CD4 effectorness gradient remained significant for IBD, CD, AIT ($P < 0.005$, MC test), and UC ($P < 0.05$, MC test), and the associations with CD8 effectorness gradient remained significant for IBD and CD ($P < 0.05$, MC test), indicating that `scDRS` distinguishes effectorness gradients within clusters. In addition, as a negative control, height was not significantly associated in any of these analyses. The association of T cell effectorness gradients with autoimmune diseases has not previously been formally evaluated, but is consistent with previous studies linking T cell effector functions to autoimmune disease[55,56]; the results also suggest that different subpopulations of effector T cells share certain similarities in their association with autoimmune diseases, consistent with previous studies characterizing the similarities among different subtypes of effector T cells, such as an increase in the expression of cytokines and chemokines[53,57,58].

Finally, we prioritized disease-relevant genes by computing the correlation (across all 110,096 TMS FACS cells) between

the expression of a given gene and the scDRS score for a given disease; this approach identifies genes that are co-expressed with genes implicated by disease GWAS. We compared the top 1,000 genes prioritized using this approach with gold-standard disease-relevant genes based on putative drug targets from Open Targets[59] (phase 1 or above; 8 gene sets with 27-250 genes; used for 8 autoimmune diseases except RR-ENT and HT; Supp. Table 18) or genes known to cause a Mendelian form of the disease[60] (550 genes corresponding to "immune dysregulation", used for RR-ENT and HT; Supp. Table 18). Results are reported in Fig. 4D and Supp. Table 19. We determined that scDRS attained a more accurate prioritization of disease-relevant genes compared to the top 1,000 MAGMA genes (median ratio of (excess overlap − 1) was 2.02, median ratio of $-\log_{10}$ p-value was 2.76; see Methods), likely by capturing disease-relevant genes with weak GWAS signal[46]. For example, *ITGB7* was prioritized by scDRS for association with IBD (rank 9) but was missed by MAGMA (rank 10565, MAGMA P =0.54); *ITGB7* impacts IBD via controlling lymphocyte homing to the gut and is a drug target for IBD (using vedolizumab)[61,62]. In addition, *JAK1* was prioritized by scDRS for association with RA (rank 386) but was missed by MAGMA (rank 5228, MAGMA P =0.26); *JAK1* plays a role in regulating immune cell activation and is a drug target for RA (using tofacitinib, baricitinib, or upadacitinib)[63,64].

We performed 4 secondary analyses. First, we assessed cell type-disease associations using scDRS for two human scRNA-seq data sets (Cano-Gamez & Soskic et al.[53] and Nathan et al.[65]; Supp. Table 2) and each of the 10 autoimmune diseases (and height, a negative control trait); we focused on cell type-disease associations because these data sets contain well-annotated T cell subtypes and states. Results are reported in Supp. Table 20. In the Cano-Gamez & Soskic et al. data, cytokine-induced Tregs, cytokine-induced Th17 cells, and activated natural Tregs were significantly associated with IBD (FDR<0.05, MC test). In the Nathan et al. data, RORC$^+$ Tregs, Th17 cells, CD161$^+$ Th2 cells, Th2 cells, Th1 cells, and activated CD4$^+$ T cells were significantly associated with IBD (FDR<0.05, MC test). These findings are consistent with our discoveries in TMS FACS linking activated T cells, particularly Tregs and Th17 cells, to IBD. In addition, as a negative control, no cell type was significantly associated with height in these two data sets. Second, we compared scDRS to cluster-level analyses using LDSC-SEG at various clustering resolutions. Results are reported in Supp. Fig. 17. We determined that both methods produced similar results at the cluster level, but the cluster-level analyses failed to recapitulate the individual cell-disease associations detected in the scDRS individual cell-level analysis (even when clustering at a very high resolution). Third, we investigated alternative disease gene prioritization methods, including prioritizing genes based on specific expression in the disease-critical T cell population (differentially expressed genes for comparing T cells vs. other cells in the TMS FACS data) and based on correlating the expression level of a given gene with scDRS disease scores across T cells, CD4$^+$ T cells, or CD8$^+$ T cells (instead of all TMS FACS cells). We determined that our primary approach provided a more accurate prioritization of gold-standard disease-relevant genes (Supp. Fig. 18A-J). Fourth, we extended our prioritization of disease-relevant genes to all 74 diseases/traits. We compared the prioritized genes with drug target genes for 27 diseases and Mendelian disease genes for 45 diseases (Supp. Table 18). We determined that our approach attained a similar improvement over MAGMA across this broader set of diseases/traits (Supp. Fig. 18K-N).

We conclude that T cells exhibit strong heterogeneity in association with autoimmune diseases. This heterogeneity can be partially explained by T cell effectorness gradients (e.g., 31% variance explained by the CD4 effecterness gradient for IBD), with stronger associations for effector-like T cells. In addition, genes whose expression across individual cells is correlated with the scDRS disease score are strongly enriched for gold-standard drug target and Mendelian disease genes.

Heterogeneous subpopulations of neurons associated with brain-related diseases and traits

We sought to further understand the heterogeneity across neurons (in the non-myeloid brain tissue) in the TMS FACS data (484 cells labeled as "neuron") in association with brain-related diseases and traits (Fig. 3). We considered 6 brain-related diseases and traits: SCZ, MDD, neuroticism (NRT), ECOL, BMI, Smoking (Supp. Table 1); we also considered height as a negative control trait. The TMS FACS data includes a partition of neurons into four brain subtissues (cerebellum, cortex, hippocampus, and striatum), but significant heterogeneity remained when we stratified our heterogeneity analyses by subtissue (Supp. Fig. 19). Since the TMS FACS data has limited coverage of neuronal subtypes, we focused our subsequent analyses on a separate mouse brain scRNA-seq data set[66] (Zeisel & Muñoz-Manchado et al.[66]; 3,005 cells), which has better coverage of neuronal subtypes and has been analyzed at cell type level in several previous genetic studies[8,26,67]. We first investigated cell type-trait associations using scDRS, which associated several neuronal subtypes (CA1 pyramidal neurons, SS pyramidal neurons, and interneurons) with the 6 brain-related traits (Supp. Fig. 20A, Supp. Table 21), consistent with previous genetic studies[8,26,67]. We focused on the CA1 pyramidal neurons from the hippocampus (827 cells), which exhibited the strongest within-cell type heterogeneity (FDR< 0.005 for all 6 brain traits, MC test; Supp. Table 21). Individual cell-trait associations for SCZ are reported in Fig. 5A, and individual cell-trait associations for all 6 brain-related traits are reported in Supp. Fig. 20B. We observed a continuous gradient of CA1 pyramidal neuron-SCZ associations, with similar results for other traits.

We investigated whether the heterogeneity observed in Fig. 5A was correlated with spatial location; we hypothesized that such a correlation might exist because of the known location-specific functions of hippocampal neurons[17,68]. We inferred

spatial coordinates of the CA1 pyramidal neurons along the natural CA1 spatial axes[69] (dorsal-ventral long axis, proximal-distal transverse axis, and superficial-deep radial axis) for each cell in terms of continuous individual cell-level scores for these 6 spatial regions by applying scDRS to published spatial signature gene sets (instead of MAGMA putative disease gene sets; Supp. Fig. 20C and Supp. Table 8; Methods). We verified that the inferred dorsal and ventral scores obtained by applying this procedure to independent mouse[70] and human[71] data sets with annotated spatial coordinates on the long axis were significantly correlated with the annotated spatial coordinates (r =0.54, 0.65 for the mouse data, r =0.18, 0.20 for the human data, P <0.01 each, MC test; Supp. Fig. 21); annotated spatial coordinates on the transverse and radial axes were not available in these data sets. The inferred spatial scores for the long (dorsal, ventral) and transverse (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, Supp. Fig. 20).

We used the results of scDRS for individual cells to assess whether the inferred spatial scores for each of the 6 spatial regions (dorsal/ventral/proximal/distal/superficial/deep) were correlated to the scDRS disease scores for each of the 6 brain-related traits (and height, a negative control trait) across CA1 pyramidal neurons (Methods). Results are reported in Fig. 5B (for the proximal region, which had the strongest associations), Supp. Fig. 22 and Supp. Table 22. We determined that proximal score was strongly associated with all 6 brain-related traits (all P <0.001 except P =0.002 for MDD; P =0.008 for height, non-significant after Bonferroni correction; MC test; 15%-29% of variance in scDRS disease score explained by proximal scores across 6 brain-related traits), 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[72,73].

We reapplied scDRS to three additional mouse single-cell data sets[70,74,75] and three human single-cell data sets[71,76,77] (Supp. Table 2), computing both spatial scores and disease scores for each cell as above. Results are reported in Supp. Fig. 22. We determined that the proximal score was consistently associated with the 6 brain-related traits across these 7 data sets (while the distal score was consistently non-associated). For the long (dorsal-ventral) and radial (superficial-deep) axes, while the dorsal and deep scores were consistently associated with the 6 brain-related traits across the 7 data sets, the corresponding ventral and superficial scores were consistently associated across the 3 human data sets but consistently non-associated across the 4 mouse data sets, possibly due to differences in brain biology between human and mouse[68,78].

We conclude that CA1 pyramidal neurons exhibit strong heterogeneity in association with brain-related diseases and traits. This heterogeneity can be partially explained by inferred spatial coordinates and may reflect the underlying functional organization of CA1 pyramidal neurons.

Heterogeneous subpopulations of hepatocytes associated with metabolic traits

Finally, we sought to further understand the heterogeneity across hepatocytes (in the liver) in the TMS FACS data in their association with metabolic traits (Fig. 3). Since the original study clustered all cells from the liver together[19] (limiting the resolution for distinguishing cell states within hepatocytes), we reclustered the hepatocytes alone, resulting in 6 clusters (1,102 cells, Fig. 5C; Methods). We considered 9 metabolic traits: TG, high-density lipoprotein (HDL), low-density lipoprotein (LDL), total cholesterol (TC), TST, alanine aminotransferase (ALT), alkaline phosphatase (ALP), sex hormone-binding globulin (SHBG), and total bilirubin (TBIL) (Supp. Table 1); we also considered height as a negative control trait.

We focused on individual cells associated with TG, a representative metabolic trait (Fig. 5C; results for the other 8 metabolic traits and height are reported in Supp. Fig. 23). The 530 TG-associated cells (FDR<0.1) formed subpopulations of 5 of the 6 hepatocyte clusters; we characterized these subpopulations based on ploidy level (number of sets of chromosomes in a cell) and zonation (pericentral/mid-lobule/periportal spatial location in the liver lobule), which have been extensively investigated in previous studies of hepatocyte heterogeneity[79–81]. We inferred the ploidy level and zonation for each individual cell in terms of a polyploidy score, a pericentral score, and a periportal score by applying scDRS to published polyploidy/zonation signature gene sets[82–84] (instead of MAGMA putative disease gene sets; Supp. Fig. 24; Methods); we verified that the inferred high-ploidy hepatocytes had higher expression levels of the *Xist* (X-inactive specific transcript) non-coding RNA gene (for hepatocytes in female mice) and higher numbers of expressed genes (Supp. Fig. 24H,I), two distinguishing features of high-ploidy hepatocytes[83,85]. We further verified that the inferred polyploidy score obtained by applying this procedure to independent data[83] with experimentally determined ploidy level annotation were significantly correlated with the experimentally determined annotation (r =0.28, P <0.001, MC test), and that the inferred zonation scores obtained by applying this procedure to independent data[84] with experimentally determined zonation annotations were significantly correlated with the experimentally determined annotations (r =0.42, P <0.001 for pericentral score, r =0.45, P <0.001 for periportal score, MC test). The inferred ploidy level and zonation varied across clusters, providing visual evidence of stronger cell-TG associations in high-ploidy clusters (cluster 1,2), particularly the periportal high-ploidy cluster (cluster 2; Fig. 5C).

We used the results of scDRS for individual cells to assess whether the inferred polyploidy, pericenteral and periportal

scores were correlated to the scDRS disease score for each of the 9 metabolic traits (and height, a negative control trait) across hepatocytes; we jointly regressed the scDRS disease score for each trait on the polyploidy score, pericentral score, and periportal score (because the polyploidy score was positively correlated with the other 2 scores; Methods). Results are reported in Fig. 5D (for the polyploidy score which had the strongest associations), Supp. Fig. 25 and Supp. Table 23. The polyploidy, pericentral, and periportal scores jointly explained 42%-63% of variance of the scDRS disease scores across the 9 metabolic traits. We determined that the polyploidy score was strongly associated with all 9 metabolic traits (all $P < 0.005$ except $P = 0.01$ for LDL; $P = 0.62$ for height; 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[81,82], and supports the hypothesis that liver functions are enhanced in polyploid hepatocytes[81]. In addition, the periportal score was associated with the 9 metabolic traits ($P < 0.005$ for ALT and ALP, all $P < 0.05$ except $P = 0.19$ for TBIL; $P = 0.24$ for height; MC test). While the pericentral score was not significantly associated with these traits in the TMS FACS data, we detected significant associations across multiple other data sets (see below). These results suggest that these metabolic traits are impacted by complex processes involving both pericentral and periportal hepatocytes.

We performed 3 secondary analyses. First, we reapplied scDRS to 4 additional mouse single-cell data sets[83,84,86] and 1 human single-cell data set[87] (Supp. Table 2). Results are reported in Supp. Fig. 25A,B. The results suggest consistent association of the polyploidy score and both the pericentral and periportal scores with the 9 metabolic traits. Second, given that scDRS associated both pericentral and periportal hepatocytes to metabolic traits, we assessed whether scDRS is able to detect pericentral-specific and periportal-specific effects. We analyzed all 6 hepatocyte scRNA-seq data sets using 8 metabolic pathway gene sets[88,89] (instead of MAGMA genes from GWAS; Methods) whose zonation patterns are well-understood (4 pericentral-specific pathways and 4 periportal-specific pathways[80]). Results are reported in Supp. Fig. 25C,D. We determined that pericentral-specific pathways generally exhibited pericentral-specific effects, and periportal-specific pathways generally exhibited periportal-specific effects. Third, we assessed the robustness of our polyploidy score by inferring the ploidy level of hepatocytes using 3 additional sets of polyploidy signatures and 3 additional sets of diploidy signatures[82] (expected to be negatively correlated with the polyploidy score; Methods) for each of the 6 data sets. Results are reported in Supp. Table 24. We determined that the polyploidy score is strongly positively correlated with scores obtained using the additional polyploidy signatures ($P < 0.005$ for 17/18 correlations, MC test) and strongly negatively correlated with scores obtained using the additional diploidy signatures ($P < 0.005$ for 10/18 correlations, MC test).

We conclude that hepatocytes exhibit strong heterogeneity in association with metabolic traits. This heterogeneity can be partially explained by inferred ploidy levels and zonation patterns, with stronger associations for hepatocytes with higher ploidy level and hepatocytes located both in pericentral or periportal regions (instead of the mid-lobule region).

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. We applied scDRS to 74 diseases and complex traits in conjunction with 16 scRNA-seq data sets and detected extensive heterogeneity in disease associations of individual cells within classical cell types, including subpopulations of T cells associated with IBD partially characterized by their effector-like states, subpopulations of neurons associated with SCZ partially characterized by their spatial location, and subpopulations of hepatocytes associated with TG partially characterized by their higher ploidy levels. These findings have improved our understanding of these diseases/traits, and 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.

scDRS does not rely on annotations of classical cell types based on known marker genes, a standard approach for integrating GWAS with scRNA-seq data[6–8] (and bulk gene expression data[9–12]; see Supp. Note). Thus, scDRS is particularly well-suited for analyzing data sets that are less well-annotated (e.g., large-scale cell atlases[19,25]) or contain less well-studied cell populations. In addition, scDRS characterizes heterogeneity across individual cells in their associations to common diseases and complex traits, providing a unique perspective relative to studies of single-cell transcriptional heterogeneity focusing on scRNA-seq data alone[13–16,18,90,91]; it also improves upon recent methods for scoring individual cells with respect to a given gene set (e.g., Seurat[15], Vision[16], and VAM[18]) by providing robust individual cell-level association p-values and higher detection power (see Supp. Note).

We have demonstrated the value of scDRS 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. We anticipate that application of scDRS to future scRNA-seq/snRNA-seq and GWAS data sets 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 work[6,7,12,20]. 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 relevant cell-level variables that we identified (e.g., T cell effectorness gradients for autoimmune diseases) only partially explain the heterogeneity across individual cells in their association to disease; there are likely more cell-level variables driving this heterogeneity that remain to be identified. Third, we primarily used mouse RNA-seq data (TMS FACS) to study human diseases and complex traits, but there are biological differences between human and mouse. Arguments in favor of using mouse RNA-seq data to study human diseases and complex traits include (1) it is easier to obtain high-quality atlas-level scRNA-seq data from mice, (2) our key findings were replicated in human data, (3) we evaluated only protein-coding genes with 1:1 orthologs between mice and humans, which are highly conserved, (4) we used a large number of genes to associate cells to diseases (1,000 MAGMA putative disease genes), minimizing potential bias due to individual genes differentially expressed across species (see Bryois et al.[8] and other studies[6,7,12,26] for additional discussion). However, it is possible that some cell types are less conserved across species[8,92] (e.g., our results for CA1 pyramidal neurons along the long and radial axes (Supp. Fig. 22) seem to indicate different disease association patterns between human and mouse), motivating follow-up analyses involving human scRNA-seq data (including those that we have performed here). Fourth, we identified putative disease genes using MAGMA, a widely used method[20]. However, it may be possible to construct more accurate sets of disease genes by incorporating other types of data, such as protein-protein interaction data[93] or functionally informed SNP-to-gene linking strategies[94]; we caution that such efforts must strive to avoid biases towards well-studied tissues. Fifth, `scDRS` detects overexpression of putative disease genes (analogous to previous works[7,8,12]), but is not designed to detect underexpression. Our initial implementation of an overdispersion score was less well-powered than `scDRS` in analyses of real disease/traits (Supp. Fig. 10), but further efforts to combine directional and overdispersion scores may be warranted[95]. Sixth, `scDRS` results for a given cell depends on the other cells in the data set through both the estimation of technical noise levels and the selection of matched control genes; however, both steps depend only on gene-specific quantities averaged across all cells (gene-specific expression mean and expression variance) and are thus robust to inclusion or exclusion of a small set of cells (or a large random subset of cells). Seventh, 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 Tregs (one of the causal cell types for IBD) will likely yield largely non-significant results. This limitation did not impact our main analyses, because the TMS data includes 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 (e.g., the TMS or TS data). Eighth, we have only analyzed scRNA-seq data from control samples. Extending `scDRS` to analyze scRNA-seq data from case-control samples or experimentally perturbed samples[96], perhaps by applying `scDRS` and comparing disease scores of cells from different conditions, may provide further insights about disease. Despite all these limitations, `scDRS` is a powerful method for distinguishing disease associations of individual cells in single-cell RNA-seq data.

Methods

scDRS method

We consider a scRNA-seq data set with n_{cell} cells and n_{gene} genes. We denote the cell-gene matrix as $\mathbf{X} \in \mathbb{R}^{n_{\text{cell}} \times n_{\text{gene}}}$, where X_{cg} represents the expression level of cell c and gene g. We assume that $\mathbf{X}$ is size-factor-normalized (e.g., 10,000 counts per cell) and log-transformed ($\log(x+1)$) from the original raw count matrix[21]. We regress the covariates out from the normalized data[21] (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, which is needed for estimating the gene-specific technical noise levels (see Supp. Note).

The `scDRS` algorithm is described in Box 1. Given a disease GWAS and a 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 B =1,000) that can be used for data visualization and Monte Carlo-based statistical inference (see Downstream applications and MC test). `scDRS` consists of three steps. First, `scDRS` constructs a set of putative disease genes from the GWAS summary statistics. Second, `scDRS` computes a 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. These steps are detailed below.

Step 1: Constructing disease gene set. We use MAGMA[20] to compute gene-level association p-values from disease GWAS summary statistics (Box 1, step 1). We use a reference panel based on individuals of European ancestry in the 1000 Genomes Project[97]. We use a 10-kb window around the gene body to map SNPs to genes. We select the top 1,000 genes based on MAGMA p-values as putative disease genes. We denote the disease gene set as $G \subset \{1, 2, \cdots, n_{\text{gene}}\}$. Alternative parameter

Box 1 Single-cell disease relevance score (`scDRS`)

Input: Disease GWAS summary statistics (or putative disease gene set G), scRNA-seq data $\mathbf{X} \in \mathbb{R}^{n_{\text{cell}} \times n_{\text{gene}}}$.
Parameters: Number of MC samples of control gene sets B (default 1,000).

1: **Construct putative disease gene set**
a: Construct putative disease gene set $G \subset \{1, 2, \cdots, n_{\text{gene}}\}$ from GWAS summary statistics using MAGMA.

2: **Compute disease scores and control scores**
a: Sample B sets of control genes $G_1^{\text{ctrl}}, \ldots, G_B^{\text{ctrl}}$ matching mean expression and expression variance of disease genes.
b: Estimate gene-specific technical noise level $\sigma^2_{\text{tech},g}$, $\forall g \in \{1, \cdots, n_{\text{gene}}\}$.
c: Compute raw disease score and B raw control scores for each cell $c = 1, \cdots, n_{\text{cell}}$,

$$\text{raw disease score: } s_c = \frac{\sum_{g \in G} \sigma^{-1}_{\text{tech},g} X_{cg}}{\sum_{g \in G} \sigma^{-1}_{\text{tech},g}}, \qquad B \text{ raw control scores: } s^{\text{ctrl}}_{cb} = \frac{\sum_{g \in G^{\text{ctrl}}_b} \sigma^{-1}_{\text{tech},g} X_{cg}}{\sum_{g \in G^{\text{ctrl}}_b} \sigma^{-1}_{\text{tech},g}}, \ \forall\, b \in \{1, \cdots, B\} \quad (1)$$

3: **Compute disease association p-values**
a: First gene set alignment by mean and variance. Let σ^2_g be the expression variance of gene g. For each cell c,

$$s_c \leftarrow s_c - \frac{1}{n_{\text{cell}}} \sum_{c'=1}^{n_{\text{cell}}} s_{c'}, \qquad s^{\text{ctrl}}_{cb} \leftarrow \left(s^{\text{ctrl}}_{cb} - \frac{1}{n_{\text{cell}}} \sum_{c'=1}^{n_{\text{cell}}} s^{\text{ctrl}}_{c'b} \right) \frac{\sum_{g \in G^{\text{ctrl}}_b} \sigma^{-1}_{\text{tech},g}}{\sum_{g \in G} \sigma^{-1}_{\text{tech},g}} \sqrt{\frac{\sum_{g \in G} \sigma^{-2}_{\text{tech},g} \sigma^2_g}{\sum_{g \in G^{\text{ctrl}}_b} \sigma^{-2}_{\text{tech},g} \sigma^2_g}}, \ \forall\, b \in \{1, \cdots, B\} \quad (2)$$

b: Cell-wise standardization for each cell c by the mean $\hat{\mu}^{\text{ctrl}}_c$ and variance $\hat{\sigma}^{\text{ctrl}}_c$ of control scores $s^{\text{ctrl}}_{c1}, \cdots, s^{\text{ctrl}}_{cB}$ of that cell,

$$s_c \leftarrow (s_c - \hat{\mu}^{\text{ctrl}}_c)/\hat{\sigma}^{\text{ctrl}}_c, \qquad s^{\text{ctrl}}_{cb} \leftarrow (s^{\text{ctrl}}_{cb} - \hat{\mu}^{\text{ctrl}}_c)/\hat{\sigma}^{\text{ctrl}}_c, \ \forall\, b \in \{1, \cdots, B\} \quad (3)$$

c: Second gene set alignment by mean. For each cell c,

$$s_c \leftarrow s_c - \frac{1}{n_{\text{cell}}} \sum_{c'=1}^{n_{\text{cell}}} s_{c'}, \qquad s^{\text{ctrl}}_{cb} \leftarrow s^{\text{ctrl}}_{cb} - \frac{1}{n_{\text{cell}}} \sum_{c'=1}^{n_{\text{cell}}} s^{\text{ctrl}}_{c'b}, \ \forall\, b \in \{1, \cdots, B\} \quad (4)$$

d: Compute cell-level p-values based on the empirical distribution of the pooled normalized control scores for each cell c,

$$p_c = \left[1 + \sum_{c'=1}^{n_{\text{cell}}} \sum_{b=1}^{B} \mathbb{I}(s_c \le s^{\text{ctrl}}_{c'b}) \right] \Big/ (1 + n_{\text{cell}} B) \quad (5)$$

Output: cell-level p-values p_c, normalized disease scores s_c, and normalized control scores $s^{\text{ctrl}}_{c1}, \cdots, s^{\text{ctrl}}_{cB}$.

choices and methods for constructing putative disease gene sets are considered below (see Alternative versions of `scDRS` method).

Step 2: Computing disease scores and control scores. We construct B sets of control genes $G_1^{\text{ctrl}}, \ldots, G_B^{\text{ctrl}}$ by randomly selecting genes matching the mean expression and expression variance of the disease genes (Box 1, step 2a). Specifically, each control gene set G_b^{ctrl} has the same size as the disease gene set G and is constructed by first dividing all genes into 20×20 equal-sized mean-variance bins and then for each gene in the disease gene set, randomly sampling a control gene from the same bin (containing the disease genes) without replacement. Next, we estimate the technical noise level for each gene $\sigma^2_{\text{tech},g}$ in the scRNA-seq data using a procedure similar to previous works[18,21], and compute the raw disease score and raw control scores for each cell as weighted average expression of genes in the corresponding gene set (Box 1, steps 2b-2c, Supp. Note). The weight for gene g is proportional to $\sigma^{-1}_{\text{tech},g}$, which down-weights genes with higher levels of technical noise to increase detection power. The weighting strategy was adapted from VAM[18], where the cell-specific score is proportional to $\sum_{g \in G} \sigma^{-2}_{\text{tech},g} X^2_{cg}$ and was shown to have a superior classification accuracy. Alternative cell scores (instead of the weighted average score) are evaluated below (see Alternative versions of `scDRS` method).

Step 3: Computing disease-association p-values. We first describe the alternative distribution that `scDRS` aims to detect. Since the control genes match the mean expression and expression variance of the disease genes, it can be shown that the raw disease score has the same mean but a higher variance compared to each set of raw control scores; the higher variance is

because the disease genes are more positively correlated with each other due to co-expression in the associated cell population (Supp. Fig. 1A-C). As a result, the disease-relevant cells, with high expression of the disease genes, are expected to have larger raw disease scores than raw control scores. Please see Supp. Note for more details.

The first gene set alignment (Box 1, step 3a) corrects for the potential mismatch of control gene sets by first centering the scores and then aligning the variance level for each gene set. The variance of the raw disease score is estimated as $\sum_{g\in G} w_g^2 \sigma_g^2$ and similarly for the raw control scores, with σ_g^2 being the expression variance of gene g and $w_g = \sigma_{\text{tech},g}^{-1} / \sum_{g\in G} \sigma_{\text{tech},g}^{-1}$ the corresponding weight; this heuristic assumes independence of the genes (or different gene sets have similar levels of gene-gene correlation), and consequently avoids down-weighting the raw disease score due to the higher correlation between disease genes (Supp. Fig. 1D, Supp. Note). After adjusting the control gene sets, the gold standard MC p-values, based on comparison to B MC samples of raw control scores of the same cell, can be written as[22]

$$p_c^{\text{MC}} = \frac{1+\sum_{b=1}^{B} \mathbb{I}(s_c \leq s_{cb}^{\text{ctrl}})}{1+B}, \ \forall\, c \in \{1, \cdots, n_{\text{cell}}\}. \tag{6}$$

This finite-sample MC p-value is a conservative estimate of the ideal MC p-value obtained via an infinite number of MC samples[22]. However, as Eq. (6) suggests, an MC test with B MC samples can only produce an MC p-value no smaller than $1/(1+B)$. Instead of using a large number of MC samples which is computationally intensive, we approximate the ideal MC p-value by pooling the control scores across cells. Specifically, we first align the control score distributions (across the B control gene sets, for each cell) by matching their means and variances, followed by re-centering the mean scores of different gene sets (Box 1, steps 3b-3c, Supp. Fig. 1E,F, Supp. Note). This procedure produces a normalized disease score and B normalized control scores for each cell. Finally, we compute the `scDRS` p-values based on the empirical distribution of the pooled normalized control scores across all control gene sets and all cells (Box 1, step 3d). The pooling procedure assumes that the raw control score distributions (across the B control gene sets, for each cell) are from the same location-scale family (e.g., the family of all normal distributions or that of all student's t-distributions) such that they can be aligned by matching the first two moments; it is a reasonable assumption when the number of disease genes is neither too small nor too large (e.g., $50 < |G| < 20\% n_{\text{gene}}$), where the control score distributions are close to normal distributions by the central limit theorem (Supp. Note). As shown in Supp. Fig. 1G-I, the `scDRS` p-values with B =1,000 is indeed able to well approximate the MC p-values obtained using a much larger number of MC samples (B =20,000).

Downstream applications and MC test

`scDRS` outputs individual cell-level p-values, (normalized) disease scores, and (normalized) control scores that can be used for a wide range of downstream applications: assessing association between a given cell type and a given disease; assessing heterogeneity in association with a given disease across a given set of cells; and assessing association between a cell-level variable and a given disease across a given set of cells. We use a unified MC test for these 3 analyses based on the disease score and control scores. Specifically, let t be the test statistic computed from the disease score of the given set of cells (the 3 analyses differ by the test statistics they use) and let $t_1^{\text{ctrl}}, \cdots, t_B^{\text{ctrl}}$ 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

$$p^{\text{MC}} = \frac{1+\sum_{b=1}^{B} \mathbb{I}(t \leq t_b^{\text{ctrl}})}{1+B}. \tag{7}$$

The MC test avoids the assumption that the cells are independent—a strong assumption in scRNA-seq analyses, e.g., when analyzing cells in the same cluster that are dependent due to the clustering process. We can also compute an MC z-score as $z^{\text{MC}} = \left[t - \text{Mean}\left(\{t_b^{\text{ctrl}}\}_{b=1}^{B}\right)\right] / \text{SD}\left(\{t_b^{\text{ctrl}}\}_{b=1}^{B}\right)$; this MC z-score is not restricted by the MC limit of $1/(1+B)$ but relies the assumption that the control test statistics $\{t_b^{\text{ctrl}}\}_{b=1}^{B}$ approximately follow a normal distribution. Below, we describe the test statistics used by the 3 analyses listed above. We note that the MC test can in principle be extended to any analysis that computes a test statistic from the disease scores of a set of cells.

Assessing association between a given cell type and a given disease. We use the top 5% quantile of the disease scores of cells from the given cell type as the test statistic. This test statistic is robust to annotation outliers, e.g., a few misannotated but highly significant cells. One can also use other test statistics such as the top 1% quantile or the maximum.

Assessing heterogeneity in association with a given disease across a given set of cells. We use Geary's C[16,98] as the test statistic. Geary's C measures the spatial autocorrelation of the disease score across a set of cells (e.g., cells from the same cell type or cell cluster) with respect to a cell-cell similarity matrix. Given a set of n cells, the corresponding disease scores $s_1, \cdots, s_n$, and the cell-cell similarity matrix $W \in \mathbb{R}^{n\times n}$, Geary's C is calculated as

$$C = \frac{(n-1)\sum_{i,j} W_{ij}(s_i - s_j)^2}{2(\sum_{i,j} W_{ij})\sum_i (s_i - \bar{s})^2}, \tag{8}$$

where $\bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$. We use the cell-cell connectivity matrix for the similarity matrix like previous works[16], which corresponds to the "connectivities" output from the scanpy function "scanpy.pp.neighbors"[99]. A value significantly lower than 1 indicates positive spatial autocorrelation, suggesting cells close to each other on the similarity matrix have similar disease scores, forming subclusters of cells with similar levels of disease association. This indicates a high level of disease association heterogeneity across the given set of cells. We use this test to assess within-cell type disease association heterogeneity and within-cluster association disease heterogeneity.

Assessing association between a cell-level variable and a given 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.

Alternative versions of `scDRS` method

We consider 3 alternative versions of `scDRS`, involving (1) other choices of MAGMA gene window size and number of putative disease genes, (2) other choices of cell-level raw scores, (3) other strategies of constructing putative disease genes.

Other choices of MAGMA gene window size and number of putative disease genes. We evaluated other choices of the MAGMA gene window size for mapping SNPs to genes (0 kb or 50 kb, instead of 10 kb) and other numbers of the putative disease genes (100, 500, or 2,000, instead of 1,000). We considered 6 configurations (varying gene window size while fixing the putative disease gene set size as 1,000 or varying disease gene set size while fixing the gene window to be 10 kb) and evaluated the performance of each configuration using 5 traits each with a positive control and a negative control cell type (Supp. Table 15). Results are reported in Supp. Table 16. We determined that the default setting (10 kb window size and 1,000 disease genes) attained a reasonable overall performance relative to other parameter choices. As a secondary analysis, we investigated if the optimal number of putative disease genes is trait-specific and specifically if it depends on the trait heritability (Supp. Table 1) or polygenicity[46] (M_e for common SNPs; Supp. Table 1). We obtained the optimal number of disease genes (from 500, 1,000, or 2,000 while using 10-kb MAGMA gene window) for each trait that yields most significantly associated cells (FDR$<$0.1). We then correlated the optimal gene set size with heritability and polygenicity across traits. Results are reported in Supp. Fig. 12. We determined that the optimal disease gene set size is significantly correlated with trait polygenicity but non-significantly correlated with trait heritability.

Other choices of cell-level raw scores. We evaluated two alternative cell-level raw scores, namely

$$\text{unweighted average: } s_c = \frac{1}{|G|}\sum_{g\in G} X_{cg}, \qquad \text{overdispersion score: } s_c = \frac{\sum_{g\in G}\left[(X_{cg}-\mu_g)^2 - \sigma^2_{\text{tech},g}\right]/\sigma^2_{\text{tech},g}}{\sum_{g\in G} 1/\sigma^2_{\text{tech},g}}, \tag{9}$$

where μ_g and $\sigma^2_{\text{tech},g}$ are the average expression and technical noise level of gene g respectively. The overdispersion score tests for both overexpression and underexpression of the putative disease genes in the relevant cell population (unlike the weighted average score which only tests for overexpression of the disease genes). We applied the two alternative scores to the TMS FACS data and the 74 diseases and complex traits. Results are reported in Supp. Fig. 10. We determined that the weighted average score (used by `scDRS`) attained high power than the two alternative scores. It further suggests that the GWAS putative disease genes are mostly overexpressed in the relevant cell population.

Other strategies of constructing putative disease genes. While we constructed putative disease genes using GWAS data and mapped SNPs to genes based on genomic locations, it is potentially possible to obtain a more accurate disease gene set by either incorporating data from other sources like protein-protein interaction data[93] or using a more sophisticated SNP-to-gene linking strategy[94]; we did not use these approaches which may be biased towards well-studied tissues.

Simulations

We performed simulations on a data set with 10,000 cells subsampled from the TMS FACS data. In null simulations, we randomly selected putative disease genes from a set of non-informative genes. We considered three numbers of putative disease genes (100, 500, or 1,000) and four types of genes to sample from: (1) the set of all genes, (2) the set of top 25% genes with high mean expression, (3) the set of top 25% genes with high expression variance, (4) the set of top 25% overdispersed genes, where the level of overdispersion is calculated as the difference between the actual variance and the estimated technical variance in the log scale data. For the MC test for cell type-disease association, we used the top 5% quantile as the test statistic and computed the MC p-values for each cell type and each set of random putative disease genes by comparing the test statistic from the disease scores to those computed from the 1,000 sets of control scores (see Monte-Carlo-based downstream analyses above). In causal simulations, we 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 (at various effect sizes) across the 1,000 causal disease genes, and randomly selected 1,000 putative disease genes (provided as input to `scDRS` and other methods) with various levels of overlap with the 1,000 causal disease genes. Here, the effect size corresponds to the fold change of expression of the causal genes in the causal cells (multiplicative in the original count space and additive in the log space). We performed three sets of causal simulations: (1) varying effect size from 5% to 50% while fixing 25% overlap, (2) varying level of overlap from 5% to 50% while fixing 25% effect size, (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). The FDR and power reported in Fig. 2B and Supp. Fig. 3 are based on applying the B-H procedure[100] to all cells at nominal FDR=0.1. All experiments were repeated 100 times and confidence intervals were computed based on the normal distribution. We considered three methods for comparison, namely Seurat[15] ("score_genes" as implemented in scanpy[99]), Vision[16], and VAM[18]. 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 and 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 and can be easily converted to p-values; we again added the standardization step because otherwise the results of Vision were highly unstable. We did not include other methods like PAGODA[14] or AUCell[14] because it is not straightforward to convert their outputs to cell-level association p-values and also because the z-scoring methods (e.g., Vision) outperformed other methods in a comprehensive evaluation in Frost et al.[18]

GWAS summary statistic data sets

We analyzed GWAS summary statistics of 74 diseases and complex traits from the UK Biobank[101] (47 of the 74 diseases/traits with average N=415K) and other publicly available sources[102–124] (27 of the 74 diseases/traits with average N=212K); average N=341K for all 74 diseases/traits; Supp. Table 1). All diseases and traits were well-powered (heritability z-score>6), except celiac disease (Celiac), systemic lupus erythematosus (SLE), multiple sclerosis (MS), primary biliary cirrhosis (PBC), subject well being (SWB), fasting glucose (FG), and type 1 diabetes (T1D), which were included due to their clinical importance. The major histocompatibility complex (MHC) region was removed from all analyses because of its unusual LD and genetic architecture[125].

scRNA-seq data sets

We analyzed 16 scRNA-seq or snRNA-seq data sets (Supp. Table 2). We included 3 atlas-level data sets (TMS FACS, TMS droplet, and TS FACS) to broadly associate diverse cell types and cell populations to disease; these 3 data sets cover different species (mouse and human) and different technologies (FACS and droplet), which allows us to assess the robustness of our results across different species and technologies. We included another 13 data sets that focus on a single tissue and contain finer-grained annotations of cell types and cell states. Notably, several of these data sets contain experimentally determined annotations which allow us to better validate our results, including Cano-Gamez & Soskic et al. data[53] containing experimentally perturbed $CD4^+$ T cell states, Nathan et al. data[65] containing T cells states determined by profiling surface markers using CITE-seq, Habib & Li et al. data[70] containing experimentally determined spatial locations for CA1 pyramidal neurons based on ISH of spatial landmark genes, Ayhan et al. data[71] containing experimentally determined spatial locations for CA1 pyramidal neurons (dorsal and ventral) based on surgical resection, and Richter & Deligiannis et al. data[83] containing experimentally determined hepatocyte ploidy levels based on Hoechst staining.

Analysis of T cells and autoimmune diseases

We collectively analyzed all T cells from the TMS FACS data (4,125 cells labeled as $CD4^+$ α-β T cell, $CD8^+$ α-β T cell, regulatory T cell, mature NK T cell, mature α-β T cell, or T cell in the TMS data; Supp. Table 3); the more general terms like "T cell" and "mature α-β T cell" were used for cells whose more specific identities were not clear. We processed the T cells following the standard procedure using scanpy[99]. First, we performed size factor normalization (10,000 counts per cell) and log transformation. Second, we selected highly variable genes and computed the batch-corrected PCA embedding using Harmony[126], treating each mouse as a batch. Finally, we constructed KNN graphs and clustered the cells using the Leiden algorithm[127] (resolution=0.7), followed by computing the UMAP embedding. We removed 376 cells either from small clusters (less than 100 cells) or whose identities are ambiguous, resulting in 3,769 cells. We annotated the clusters based on the major TMS cell types in the cluster; the label "mature α-β T cell" was omitted because a more specific TMS cell type label (e.g., "$CD8^+$ α-β T") was available in the corresponding cluster. We considered cells from clusters 1-4 as clear $CD4^+$ T cells (1,686 cells) and cells from clusters 1, 2, 7-9 as clear $CD8^+$ T cells (2,197 cells; the shared clusters 1 and 2 contain a mix of naive $CD4^+$ and $CD8^+$ T cells). We used diffusion pseudotime (DPT)[54] to assign effectorness gradient for $CD4^+$ and $CD8^+$ T cells separately, where we used the leftmost cell in cluster 2 on the UMAP as the root cell (clearly naive T cell).

We used MSigDB[88,89] (v7.1) to curate T cell signature gene sets, including naive CD4, memory CD4, effector CD4, naive CD8, memory CD8, effector CD8, Treg, Th1 (T helper 1), Th2 (T helper 2), and Th17 (T helper 17) signatures. For each T cell

signature gene set, we identified a set of relevant MSigDB gene sets (22-34, Supp. Table 7), followed by selecting the top 100 most frequent genes in these MSigDB gene sets as the T 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 (Supp. Table 8). For gold-standard gene sets used in the analysis of disease gene prioritization, we curated 27 putative drug target gene sets from Open Targets[59] (mapped to 27 of the 74 diseases/traits considered in the paper; Supp. Table 18); 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 diseases gene sets from Freund et al.[60] (mapped to 45 of the 74 diseases/traits considered in the paper; Supp. Table 18). For comparison of two gene sets, the p-value is based on 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 $-\log_{10}$ p-value increases with the size of the reference gene set; we report both excess overlap and $-\log_{10}$ p-value while using the former as our primary metric, which is more interpretable.

For the analysis of individual cells associated with IBD, we considered 4 major clusters of T cells with >25 IBD-associated cells (FDR<0.1) and inferred the identities of the subpopulations of IBD-associated cells in these 4 clusters based on the expression of marker genes and overlap of the specifically expressed genes in each of these subpopulations with T cell signatures. First, the subpopulation of 120 IBD-associated cells in cluster 3 (which consisted of 629 cells with TMS cell type labels "CD4$^+$ α-β T" or "regulatory T") were labeled as "Treg" as described in the main paper. Second, the 75 IBD-associated cells in cluster 4 (which consisted of 165 cells with TMS cell type label "CD4$^+$ α-β T") had specifically expressed genes overlapping with a *KLRG1*$^+$ *AREG*$^+$ effector-like Treg program[48] characterized by high expression levels of *IL1RL1* (*ST2*), *KLRG1*, and *AREG* ($P = 1.3 \times 10^{-50}$, Fisher's exact test; Supp. Fig. 15A,C), suggesting these cells had active functions for Treg differentiation, immunosuppression, and tissue repair[48] (labeled as "Effector-like Treg" in Fig. 4B). Third, the 61 IBD-associated cells in cluster 5 (which consisted of 370 cells with TMS cell type label "T cell") had high expression of Th17 marker genes (e.g., *IL23R*, *RORC*, *IL17A*; Supp. Fig. 15A) and their specifically expressed genes significantly overlapped with Th17 signatures ($P = 2.0 \times 10^{-6}$, Fisher's exact test; Supp. Fig. 15B) and a Th17-like Treg program[48] ($P = 1.9 \times 10^{-24}$, Fisher's exact test; Supp. Fig. 15C), suggesting Th17 proinflammatory functions (labeled as "Th17-like" in Fig. 4B). Finally, the 38 IBD-associated cells in cluster 9 (consisting of 499 cells with TMS cell type label "CD8$^+$ α-β T") had high expression of genes related to cytotoxicity (e.g., *IFNG*, *GZMB*, *FASL*; Supp. Fig. 15A,B), and their specifically expressed genes significantly overlapped with effector CD8$^+$ T cell signatures ($P = 1.4 \times 10^{-7}$, Fisher's exact test; Supp. Fig. 15B), suggesting cytotoxic T cell functions (labeled as "Effector-like CD8$^+$" in Fig. 4B).

Analysis of neurons and brain-related diseases/traits

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 827 CA1 pyramidal neurons ("level1class" label "pyramidal CA1"). For inferring spatial coordinates, we curated differentially expressed genes for each of the 6 spatial regions (dorsal vs. ventral, ventral vs. dorsal, proximal vs. distal, distal vs. proximal, deep vs. superficial, and superficial vs. deep) using the gene expression data from Cembrowski et al.[69] (GEO GSE67403; gene sets in Supp. Table 8). For each differential gene expression analysis, we selected genes based on FPKM>10 for the average expression in the enriched region (e.g., dorsal for the dorsal vs. ventral comparison), q-value<0.05, and $\log_2$(fold change) >2. We used `scDRS` and these signature gene sets to assign 6 spatial scores for each cell. For the regression analysis, we separately regressed the `scDRS` disease scores for each of the 6 brain-related diseases/traits (and height, a negative control trait) on each of the 6 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 (e.g., dorsal vs. ventral) were highly collinear (strongly negatively correlated), and the inferred spatial scores for dorsal, proximal, and deep regions (which had strong marginal associations to diseases) had very low pairwise correlations (average $|r| = 0.10$; Supp. Fig. 20D), suggesting these associations were independent. We reported correlation p-values (MC test) and variance explained for each of the 6 spatial scores.

Analysis of hepatocytes and metabolic traits

We considered all hepatocytes in the TMS FACS data (1,162 cells) and reprocessed them following the same procedure as we did for the T cells. We further filtered out low-quality cells (mitochondrial proportion≥0.3; likely to be apoptotic or lysing cells), resulting in 1,102 hepatocytes (Fig. 5C). We curated signature gene sets for ploidy level, zonation, and putative zonated pathways. We curated 4 sets of polyploidy signatures, including differentially expressed genes (DEGs) for partial hepatectomy (PH) vs. pre-PH[82] (used for the polyploidy score), Cdk1 knockout (case) vs. control[82], 4n vs. 2n hepatocytes[83], large vs. small hepatocytes[82]. We curated 3 sets of diploidy signatures, including DEGs for pre-PH vs. PH[82], control vs. Cdk1 knockout[82], and 2n vs. 4n hepatocytes[83]. We curated signature gene sets for pericentral (CV) and periportal (PN) hepatocytes from Halpern et al.[84]. We curated gene sets for putative zonated pathways from MSigDB[88,89] (v7.1), including glycolysis (pericentral), bile acid

production (pericentral), lipogenesis (pericentral), xenobiotic metabolism (pericentral), beta-oxidation (periportal), cholesterol biosynthesis (periportal), protein secretion (periportal), and gluconeogenesis (periportal). All signature gene sets are reported in Supp. Table 8. For the joint regression analysis of `scDRS` disease score on ploidy and zonation scores, we regressed the polyploidy 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 polyploidy score was positively correlated with the pericentral and periportal scores (unlike the analysis in the neuron section where the 3 sets of scores had low correlations).

Data availability

We release our data at https://figshare.com/projects/Single-cell_Disease_Relevance_Score_scDRS_/118902 (instructions at https://github.com/martinjzhang/scDRS), including GWAS summary statistics of the 74 diseases/traits, 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/traits). The 16 scRNA-seq data sets were obtained as follows. The TMS FACS data and TMS droplet data[19] was downloaded from the official release https://figshare.com/articles/dataset/Processed_files_to_use_with_scanpy_/8273102. The TS FACS data[25] was downloaded from the official release https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219. The Cano-Gamez & Soskic et al. data[53] was downloaded from https://www.opentargets.org/projects/effectorness. The Nathan et al. data[65] was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158769. The Zeisel & Muñoz-Manchado et al. data[66] was downloaded from http://linnarssonlab.org/cortex/. The Zeisel et al. data[74] was downloaded from http://mousebrain.org/downloads.html. The Habib & Li et al. data[70] and Habib, Avraham-Davidi, & Basu et al. data[76] were downloaded from https://singlecell.broadinstitute.org/single_cell. The Ayhan et al. data[71] was downloaded from https://cells.ucsc.edu/human-hippo-axis/. The Yao et al. data[75] was downloaded from https://assets.nemoarchive.org/dat-jb2f34y. The Zhong et al. data[77] was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119212. The Aizarani et al. data[87] was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124395. Halpern & Shenhav et al. data[84] was downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84498. The Richter & Deligiannis et al. data[83] (annotated count matrix) was obtained via communication with the authors (raw data publicly available via links in the paper). The Taychameekiatchai et al. data[86] is not publicly available, but was obtained via communication with the authors.

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.

Acknowledgements

We thank Huwenbo Shi, Katherine Siewert-Rocks, Tiffany Amariuta, and Xiaoyu Xu for helpful suggestions. This research was funded by NIH grants U01 HG009379, R01 MH101244, R37 MH107649, and R01 MH115676.

Competing interests

The authors declare no competing interests.

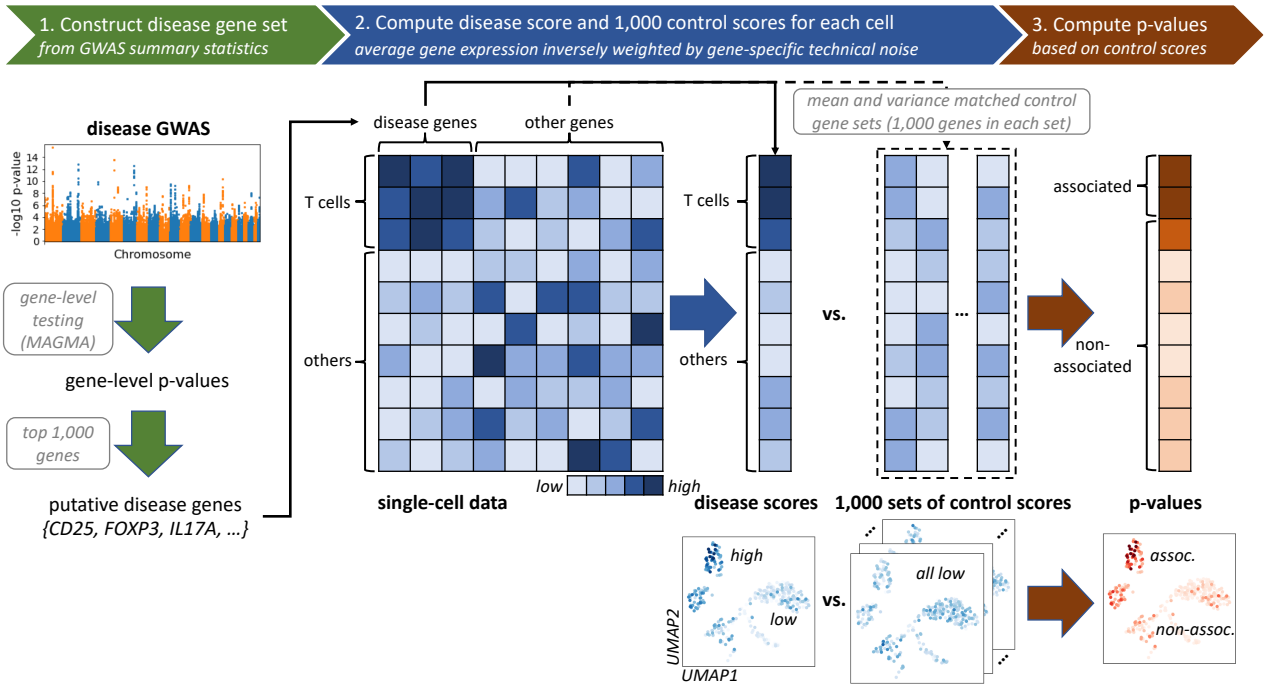

Figure 1. Overview of `scDRS` method. `scDRS` takes disease GWAS and scRNA-seq data sets as input and outputs individual cell-level p-values for association with the disease. **(1)** `scDRS` constructs a set of putative disease genes from GWAS summary statistics by selecting the top 1,000 MAGMA genes; 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 inversely weighted by its gene-specific technical noise level in the scRNA-seq data. `scDRS` also computes a set of 1,000 Monte Carlo 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.

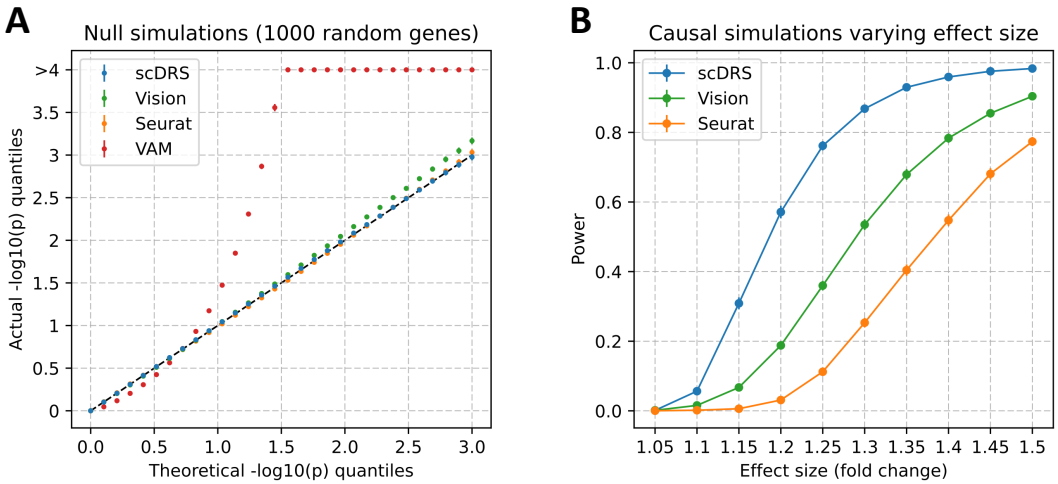

Figure 2. Results for null and causal simulations. (A) Q-Q plot for null simulations using 1,000 randomly selected genes as the putative disease genes. The x-axis denotes theoretical $-\log_{10}$ p-value quantiles and the y-axis denotes actual $-\log_{10}$ p-value quantiles for different methods. Each point is based on 100 simulation replicates (with 10,000 cells per simulation replicate); error bars denote 95% confidence intervals (all error bars are <0.05 from the point estimate). Numerical results are reported in Supp. Table 9 and additional results are reported in Supp. Fig. 2. **(B)** Power for casual simulations with perturbed expression of causal genes in causal cells. We report the power at FDR=0.1 for different methods and different effect sizes. Each point is based on 100 simulation replicates (with 10,000 cells per simulation replicate); error bars denote 95% confidence intervals (all error bars are <0.02 from the point estimate). Numerical results are reported in Supp. Table 11 and additional results are reported in Supp. Fig. 3.

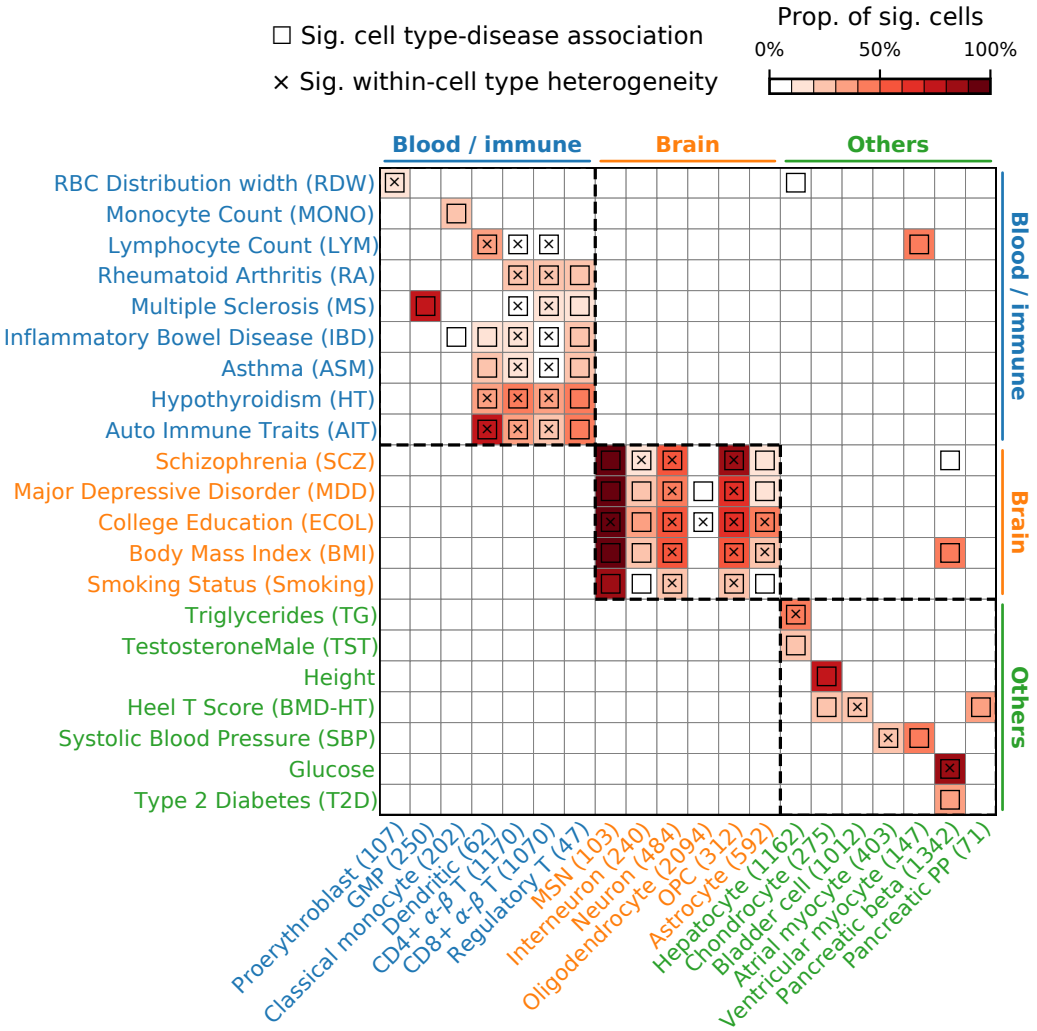

Figure 3. Disease associations at the cell type-level. We report scDRS results for individual cells aggregated at the cell type-level for a subset of 20 cell types and 21 diseases/traits in the TMS FACS data. Each row represents a disease/trait and each column represents a cell type (with number of cells indicated in parentheses). Heatmap colors for each cell type-disease pair denote the proportion of significantly associated cells (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; 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 non-significant cell type-disease associations via MC test (heatmap colors omitted for 2 pairs (GMP-MONO and Dendritic-MS) and cross symbols omitted for 6 pairs ($CD4^+$ α-β T-MONO, $CD8^+$ α-β T-MONO, oligodendrocyte-MONO, bladder cell-ASM, hepatocyte-ECOL, and dendritic-BMD-HT)). Auto Immune Traits (AIT) represents a collection of diseases in the UK Biobank that characterize autoimmune physiopathogenic etiology[128,129]. Abbreviated cell type names include red blood cell (RBC), granulocyte monocyte progenitor (GMP), medium spiny neuron (MSN), and oligodendrocyte precursor cell (OPC). Neuron refers to neuronal cells with undetermined subtypes (whereas MSN and interneuron (non-overlapping with neuron) refer to neuronal cells with those inferred subtypes). Complete results for 120 cell types and 74 diseases/traits are reported in Supp. Fig. 4 and Supp. Table 12.

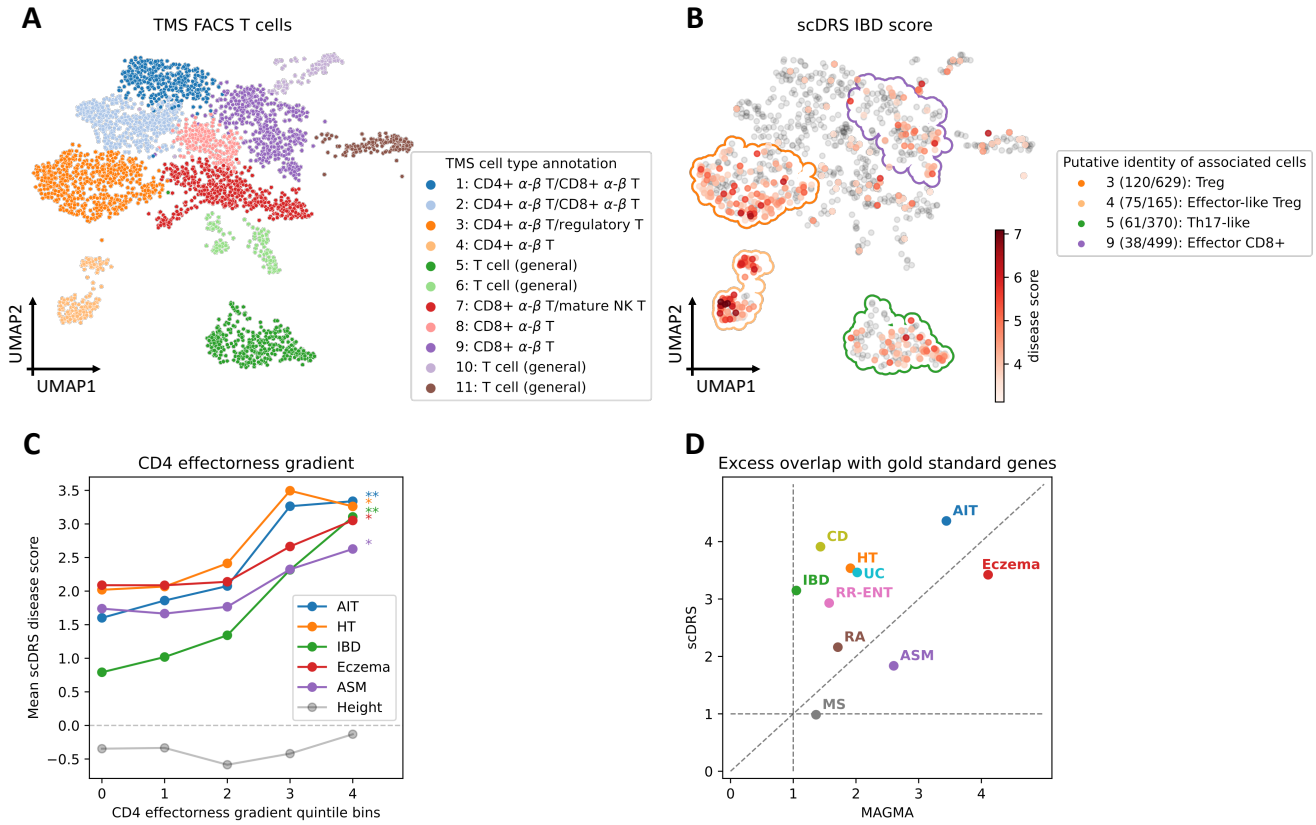

Figure 4. Associations of T cells with autoimmune diseases. (A) UMAP visualization of T cells in the TMS FACS data. In the legend, cluster labels are based on annotated TMS cell types in the cluster. Composition of tissue, sex, and age of cells in each cluster are reported in Supp. Fig. 13. **(B)** Subpopulations of T cells associated with IBD. Significantly associated cells (FDR<0.1) are denoted in red, with shades of red denoting `scDRS` disease scores; non-significant cells are denoted in grey. Cluster boundaries indicate the corresponding T cell clusters from panel A. In the legend, numbers in parentheses denote the number of IBD-associated cells vs. the total number of cells and cluster labels are based on the putative identity of the IBD-associated subpopulations, for the 4 of 11 clusters with IBD-associated cells. Results for the other 9 autoimmune diseases and height are reported in Supp. Fig. 14. **(C)** Association between `scDRS` disease score and CD4 effectorness gradient across $CD4^+$ T cells for 5 representative autoimmune diseases and height, a negative control trait. The x-axis denotes CD4 effectorness gradient quintile bins and the y-axis denotes average `scDRS` disease score in each bin for each disease. * denotes P <0.05 and ** denotes P <0.005 (MC test). Numerical results for all 10 autoimmune diseases are reported in Supp. Table 17. **(D)** Excess overlap of genes prioritized by `scDRS` with gold standard gene sets. The x-axis denotes the excess overlap of genes prioritized by MAGMA and the y-axis denotes the excess overlap of genes prioritize by `scDRS`, for each of 10 autoimmune diseases. The median ratio of (excess overlap − 1) for `scDRS` vs. MAGMA was 2.02. Numerical results are reported in Supp. Table 19.

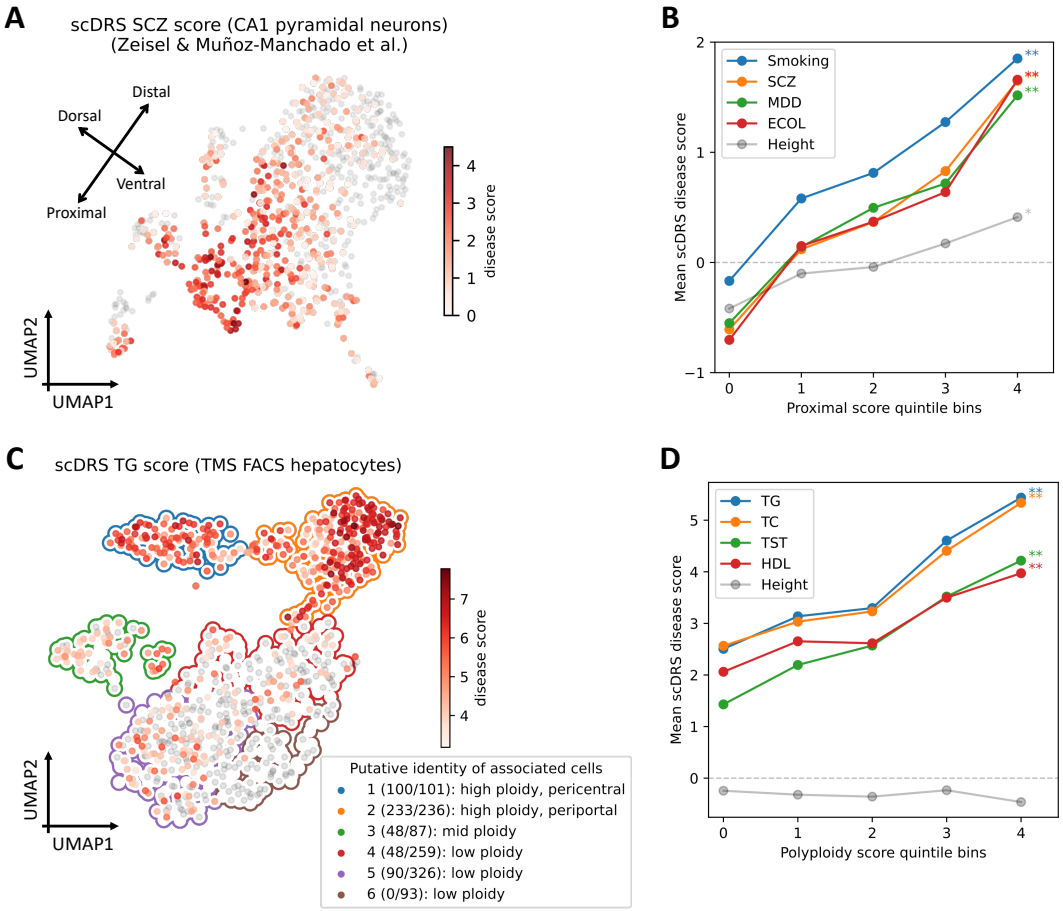

Figure 5. Associations of neurons with brain-related disease/traits and hepatocytes with metabolic traits. (A) Subpopulations of CA1 pyramidal neurons associated with SCZ in the Zeisel & Muñoz-Manchado et al. data. Colors of cells denote scDRS disease scores (negative disease scores are denoted in grey). We include a visualization of putative dorsal-ventral and proximal-distal axes (see text). Results for all 6 brain-related diseases/traits and height are reported in Supp. Fig. 20B. **(B)** Association between scDRS disease score and proximal score across CA1 pyramidal neurons for 4 representative brain-related disease/traits and height, a negative control trait. The x-axis denotes proximal score quintile bins and the y-axis denotes average scDRS disease score in each bin for each disease. * denotes P <0.05 and ** denotes P <0.005 (MC test). Results for all 6 spatial scores and all 6 brain-related diseases/traits are reported in Supp. Fig. 22 and Supp. Table 22. **(C)** Subpopulations of hepatocytes associated with TG in the TMS FACS data. Significantly associated cells (FDR<0.1) are denoted in red, with shades of red denoting scDRS disease scores; non-significant cells are denoted in grey. Cluster boundaries indicate the corresponding hepatocyte clusters. In the legend, numbers in parentheses denote the number of TG-associated cells vs. the total number of cells and cluster labels are based on the putative identity of cells in the cluster. Results for the other 8 metabolic traits and height are reported in Supp. Fig. 23. **(D)** Association between scDRS disease score and polyploidy score for 4 representative metabolic traits and height, a negative control trait. The x-axis denotes polyploidy score quintile bins and the y-axis denotes average scDRS disease score in each bin for each disease. * denotes P <0.05 and ** denotes P <0.005 (MC test). Results for all 3 scores (polyploidy score, pericentral score, periportal score) and all 9 metabolic traits (and height) are reported in Supp. Fig. 25 and Supp. Table 23.

References

1. Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.
2. Melina Claussnitzer, Judy H Cho, Rory Collins, Nancy J Cox, Emmanouil T Dermitzakis, Matthew E Hurles, Sekar Kathiresan, Eimear E Kenny, Cecilia M Lindgren, Daniel G MacArthur, et al. A brief history of human disease genetics. *Nature*, 577(7789):179–189, 2020.
3. Idan Hekselman and Esti Yeger-Lotem. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nature Reviews Genetics*, 21(3):137–150, 2020.
4. Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.
5. Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *elife*, 6:e27041, 2017.
6. Diego Calderon, Anand Bhaskar, David A Knowles, David Golan, Towfique Raj, Audrey Q Fu, and Jonathan K Pritchard. Inferring relevant cell types for complex traits by using single-cell gene expression. *The American Journal of Human Genetics*, 101(5):686–699, 2017.
7. Kyoko Watanabe, Maša Umićević Mirkov, Christiaan A de Leeuw, Martijn P van den Heuvel, and Danielle Posthuma. Genetic mapping of cell type specificity for complex traits. *Nature communications*, 10(1):1–13, 2019.
8. Julien Bryois, Nathan G Skene, Thomas Folkmann Hansen, Lisette JA Kogelman, Hunna J Watson, Zijing Liu, Leo Brueggeman, Gerome Breen, Cynthia M Bulik, Ernest Arenas, et al. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of parkinson's disease. *Nature genetics*, 52(5):482–493, 2020.
9. Xinli Hu, Hyun Kim, Eli Stahl, Robert Plenge, Mark Daly, and Soumya Raychaudhuri. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *The American Journal of Human Genetics*, 89(4):496–506, 2011.
10. Padhraig Gormley, Verneri Anttila, Bendik S Winsvold, Priit Palta, Tonu Esko, Tune H Pers, Kai-How Farh, Ester Cuenca-Leon, Mikko Muona, Nicholas A Furlotte, et al. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature genetics*, 48(8):856–866, 2016.
11. Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature genetics*, 49(12):1676–1683, 2017.
12. Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50(4):621–629, 2018.
13. Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, 13(3):241–244, 2016.
14. Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.
15. Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.
16. David DeTomaso, Matthew G Jones, Meena Subramaniam, Tal Ashuach, J Ye Chun, and Nir Yosef. Functional interpretation of single cell similarity maps. *Nature communications*, 10(1):1–11, 2019.
17. Mark S Cembrowski and Nelson Spruston. Heterogeneity within classical cell types is the rule: lessons from hippocampal pyramidal neurons. *Nature Reviews Neuroscience*, 20(4):193–204, 2019.
18. Hildreth Robert Frost. Variance-adjusted mahalanobis (vam): a fast and accurate method for cell-specific gene set scoring. *Nucleic acids research*, 48(16):e94–e94, 2020.
19. The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583(7817):590–595, 2020.

20. Christiaan A de Leeuw, Joris M Mooij, Tom Heskes, and Danielle Posthuma. Magma: generalized gene-set analysis of gwas data. *PLoS Comput Biol*, 11(4):e1004219, 2015.

21. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

22. Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

23. Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, 9(1):171, 2014.

24. Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

25. Stephen R Quake, Tabula Sapiens Consortium, et al. The tabula sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors. *bioRxiv*, 2021.

26. Nathan G Skene, Julien Bryois, Trygve E Bakken, Gerome Breen, James J Crowley, Héléna A Gaspar, Paola Giusti-Rodriguez, Rebecca D Hodge, Jeremy A Miller, Ana B Muñoz-Manchado, et al. Genetic identification of brain cell types underlying schizophrenia. *Nature genetics*, 50(6):825–833, 2018.

27. Martin Jinye Zhang, Vasilis Ntranos, and David Tse. Determining sequencing depth in a single-cell rna-seq experiment. *Nature communications*, 11(1):1–11, 2020.

28. Zhide Hu, Yi Sun, Qianqian Wang, Zhijun Han, Yuanlan Huang, Xiaofei Liu, Chunmei Ding, Chengjin Hu, Qin Qin, and Anmei Deng. Red blood cell distribution width is a potential prognostic index for liver disease. *Clinical Chemistry and Laboratory Medicine*, 51(7):1403–1408, 2013.

29. Steve R Ommen, David O Hodge, Richard J Rodeheffer, Christopher GA McGregor, Stephen P Thomson, and Raymond J Gibbons. Predictive power of the relative lymphocyte concentration in patients with advanced heart failure. *Circulation*, 97(1):19–22, 1998.

30. Stephan Ripke, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421, 2014.

31. Adam E Locke, Bratati Kahali, Sonja I Berndt, Anne E Justice, Tune H Pers, Felix R Day, Corey Powell, Sailaja Vedantam, Martin L Buchkovich, Jian Yang, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015.

32. Peng Huang, Yongzhong Zhao, Jianmei Zhong, Xinhua Zhang, Qifa Liu, Xiaoxia Qiu, Shaoke Chen, Hongxia Yan, Christopher Hillyer, Narla Mohandas, et al. Putative regulators for the continuum of erythroid differentiation revealed by single-cell transcriptome of human bm and ucb cells. *Proceedings of the National Academy of Sciences*, 117(23):12868–12876, 2020.

33. Karthik A Jagadeesh, Kushal K Dey, Daniel T Montoro, Steven Gazal, Jesse M Engreitz, Ramnik J Xavier, Alkes L Price, and Aviv Regev. Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics. *bioRxiv*, 2021.

34. Noushin Lotfi, Rodolfo Thome, Nahid Rezaei, Guang-Xian Zhang, Abbas Rezaei, Abdolmohamad Rostami, and Nafiseh Esmaeil. Roles of gm-csf in the pathogenesis of autoimmune diseases: an update. *Frontiers in immunology*, 10:1265, 2019.

35. Mirre De Bondt, Niels Hellings, Ghislain Opdenakker, and Sofie Struyf. Neutrophils: Underestimated players in the pathogenesis of multiple sclerosis (ms). *International Journal of Molecular Sciences*, 21(12):4558, 2020.

36. Jonathan RI Coleman, Héléna A Gaspar, Julien Bryois, Enda M Byrne, Andreas J Forstner, Peter A Holmans, Christiaan A de Leeuw, Manuel Mattheisen, Andrew McQuillin, Jennifer M Whitehead Pavlides, et al. The genetics of the mood disorder spectrum: genome-wide association analyses of more than 185,000 cases and 439,000 controls. *Biological psychiatry*, 88(2):169–184, 2020.

37. Devika Agarwal, Cynthia Sandor, Viola Volpato, Tara M Caffrey, Jimena Monzón-Sandoval, Rory Bowden, Javier Alegre-Abarrategui, Richard Wade-Martins, and Caleb Webber. A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nature communications*, 11(1):1–11, 2020.

38. Benjamin Ettle, Johannes CM Schlachetzki, and Jürgen Winkler. Oligodendroglia and myelin in neurodegenerative diseases: more than just bystanders? *Molecular neurobiology*, 53(5):3046–3062, 2016.

39. Andrea G Dietz, Steven A Goldman, and Maiken Nedergaard. Glial cells in schizophrenia: a unified hypothesis. *The Lancet Psychiatry*, 7(3):272–281, 2020.

40. Sonia Olivia Spitzer, Sergey Sitnikov, Yasmine Kamen, Kimberley Anne Evans, Deborah Kronenberg-Versteeg, Sabine Dietmann, Omar de Faria Jr, Sylvia Agathou, and Ragnhildur Thóra Káradóttir. Oligodendrocyte progenitor cells become regionally diverse and heterogeneous with age. *Neuron*, 101(3):459–471, 2019.

41. Michele Alves-Bezerra and David E Cohen. Triglyceride metabolism in the liver. *Comprehensive Physiology*, 8(1):1, 2017.

42. Richard KP Benninger and David J Hodson. New understanding of β-cell heterogeneity and in situ islet function. *Diabetes*, 67(4):537–547, 2018.

43. Joshua Chiou, Chun Zeng, Zhang Cheng, Jee Yun Han, Michael Schlichting, Michael Miller, Robert Mendez, Serina Huang, Jinzhao Wang, Yinghui Sui, et al. Single-cell chromatin accessibility identifies pancreatic islet cell type–and state-specific regulatory programs of diabetes risk. *Nature Genetics*, 53(4):455–466, 2021.

44. Sine Paasch Schiellerup, Kirsa Skov-Jeppesen, Johanne Agerlin Windeløv, Maria Saur Svane, Jens Juul Holst, Bolette Hartmann, and Mette Marie Rosenkilde. Gut hormones and their effect on bone metabolism. potential drug therapies in future osteoporosis treatment. *Frontiers in endocrinology*, 10:75, 2019.

45. Butian Zhou, Zhongqun Zhu, Bruce R Ransom, and Xiaoping Tong. Oligodendrocyte lineage cells and depression. *Molecular psychiatry*, 26(1):103–117, 2021.

46. Luke J O'Connor, Armin P Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, 2019.

47. Clara Abraham and Judy H. Cho. Inflammatory bowel disease. *New England Journal of Medicine*, 361(21):2066–2078, 2009. PMID: 19923578.

48. Amy Li, Rebecca H Herbst, David Canner, Jason M Schenkel, Olivia C Smith, Jonathan Y Kim, Michelle Hillman, Arjun Bhutkar, Michael S Cuoco, C Garrett Rappazzo, et al. Il-33 signaling alters regulatory t cell diversity in support of tumor development. *Cell reports*, 29(10):2998–3008, 2019.

49. Sara Omenetti and Theresa T Pizarro. The treg/th17 axis: a dynamic balance regulated by the gut microbiome. *Frontiers in immunology*, 6:639, 2015.

50. Mei Lan Chen and Mark S Sundrud. Cytokine networks and t-cell subsets in inflammatory bowel diseases. *Inflammatory bowel diseases*, 22(5):1157–1167, 2016.

51. Marine Fauny, David Moulin, Ferdinando D'amico, Patrick Netter, Nadine Petitpain, Djesia Arnone, Jean-Yves Jouzeau, Damien Loeuille, and Laurent Peyrin-Biroulet. Paradoxical gastrointestinal effects of interleukin-17 blockers. *Annals of the rheumatic diseases*, 79(9):1132–1138, 2020.

52. Tanbeena Imam, Sungtae Park, Mark H Kaplan, and Matthew R Olson. Effector t helper cell subsets in inflammatory bowel diseases. *Frontiers in immunology*, 9:1212, 2018.

53. Eddie Cano-Gamez, Blagoje Soskic, Theodoros I Roumeliotis, Ernest So, Deborah J Smyth, Marta Baldrighi, David Willé, Nikolina Nakic, Jorge Esparza-Gordillo, Christopher GC Larminie, et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of cd4+ t cells to cytokines. *Nature communications*, 11(1):1–15, 2020.

54. Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845–848, 2016.

55. David M Gravano and Katrina K Hoyer. Promotion and prevention of autoimmune disease by cd8+ t cells. *Journal of autoimmunity*, 45:68–79, 2013.

56. Stewart Leung, Xuebin Liu, Lei Fang, Xi Chen, Taylor Guo, and Jingwu Zhang. The cytokine milieu in the interplay of pathogenic th1/th17 cells and regulatory t cells in autoimmune disease. *Cellular & molecular immunology*, 7(3):182–189, 2010.

57. Maria Gutierrez-Arcelus, Nikola Teslovich, Alex R Mola, Rafael B Polidoro, Aparna Nathan, Hyun Kim, Susan Hannes, Kamil Slowikowski, Gerald FM Watts, Ilya Korsunsky, et al. Lymphocyte innateness defined by transcriptional states reflects a balance between proliferation and effector functions. *Nature communications*, 10(1):1–15, 2019.

58. Peter A Szabo, Hanna Mendes Levitin, Michelle Miron, Mark E Snyder, Takashi Senda, Jinzhou Yuan, Yim Ling Cheng, Erin C Bush, Pranay Dogra, Puspa Thapa, et al. Single-cell transcriptomics of human t cells reveals tissue and activation signatures in health and disease. *Nature communications*, 10(1):1–16, 2019.

59. Gautier Koscielny, Peter An, Denise Carvalho-Silva, Jennifer A Cham, Luca Fumis, Rippa Gasparyan, Samiul Hasan, Nikiforos Karamanis, Michael Maguire, Eliseo Papa, et al. Open targets: a platform for therapeutic target identification and validation. *Nucleic acids research*, 45(D1):D985–D994, 2017.

60. Malika Kumar Freund, Kathryn S Burch, Huwenbo Shi, Nicholas Mancuso, Gleb Kichaev, Kristina M Garske, David Z Pan, Zong Miao, Karen L Mohlke, Markku Laakso, et al. Phenotype-specific enrichment of mendelian disorder genes near gwas regions across 62 complex traits. *The American Journal of Human Genetics*, 103(4):535–552, 2018.

61. Hailong Zhang, Yajuan Zheng, Youdong Pan, Changdong Lin, Shihui Wang, Zhanjun Yan, Ling Lu, Gaoxiang Ge, Jinsong Li, Yi Arial Zeng, et al. A mutation that blocks integrin α 4 β 7 activation prevents adaptive immune-mediated colitis without increasing susceptibility to innate colitis. *BMC biology*, 18(1):1–15, 2020.

62. Cambrian Y Liu. β7 gives tregs a gut area code. *Cellular and molecular gastroenterology and hepatology*, 9(3):543–544, 2020.

63. Ernest HS Choy, Corinne Miceli-Richard, Miguel A González-Gay, Luigi Sinigaglia, Douglas E Schlichting, Gabriella Meszaros, Inmaculada de la Torre, and Hendrik Schulze-Koops. The effect of jak1/jak2 inhibition in rheumatoid arthritis: efficacy and safety of baricitinib. *Clin Exp Rheumatol*, 37(4):694–704, 2019.

64. Robert Harrington, Shamma Ahmad Al Nokhatha, and Richard Conway. Jak inhibitors in rheumatoid arthritis: an evidence-based review on the emerging clinical data. *Journal of Inflammation Research*, 13:519, 2020.

65. Aparna Nathan, Jessica I Beynor, Yuriy Baglaenko, Sara Suliman, Kazuyoshi Ishigaki, Samira Asgari, Chuan-Chin Huang, Yang Luo, Zibiao Zhang, Kattya Lopez, et al. Multimodally profiling memory t cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nature Immunology*, 22(6):781–793, 2021.

66. Amit Zeisel, Ana B Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

67. Nathan G Skene and Seth GN Grant. Identification of vulnerable cell types in major brain disorders using single cell transcriptomes and expression weighted cell type enrichment. *Frontiers in neuroscience*, 10:16, 2016.

68. Bryan A Strange, Menno P Witter, Ed S Lein, and Edvard I Moser. Functional organization of the hippocampal longitudinal axis. *Nature Reviews Neuroscience*, 15(10):655–669, 2014.

69. Mark S Cembrowski, Julia L Bachman, Lihua Wang, Ken Sugino, Brenda C Shields, and Nelson Spruston. Spatial gene-expression gradients underlie prominent heterogeneity of ca1 pyramidal neurons. *Neuron*, 89(2):351–368, 2016.

70. Naomi Habib, Yinqing Li, Matthias Heidenreich, Lukasz Swiech, Inbal Avraham-Davidi, John J Trombetta, Cynthia Hession, Feng Zhang, and Aviv Regev. Div-seq: Single-nucleus rna-seq reveals dynamics of rare adult newborn neurons. *Science*, 353(6302):925–928, 2016.

71. Fatma Ayhan, Ashwinikumar Kulkarni, Stefano Berto, Karthigayini Sivaprakasam, Connor Douglas, Bradley C Lega, and Genevieve Konopka. Resolving cellular and molecular diversity along the hippocampal anterior-to-posterior axis in humans. *Neuron*, 2021.

72. Menno P Witter, Thanh P Doan, Bente Jacobsen, Eirik S Nilssen, and Shinya Ohara. Architecture of the entorhinal cortex a review of entorhinal anatomy in rodents with some comparative notes. *Frontiers in Systems Neuroscience*, 11:46, 2017.

73. Espen J Henriksen, Laura L Colgin, Carol A Barnes, Menno P Witter, May-Britt Moser, and Edvard I Moser. Spatial representation along the proximodistal axis of ca1. *Neuron*, 68(1):127–137, 2010.

74. Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job Van Der Zwan, Martin Häring, Emelie Braun, Lars E Borm, Gioele La Manno, et al. Molecular architecture of the mouse nervous system. *Cell*, 174(4):999–1014, 2018.

75. Zizhen Yao, Cindy TJ van Velthoven, Thuc Nghi Nguyen, Jeff Goldy, Adriana E Sedeno-Cortes, Fahimeh Baftizadeh, Darren Bertagnolli, Tamara Casper, Megan Chiang, Kirsten Crichton, et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell*, 184(12):3222–3241, 2021.

76. Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, et al. Massively parallel single-nucleus rna-seq with dronc-seq. *Nature methods*, 14(10):955–958, 2017.

77. Suijuan Zhong, Wenyu Ding, Le Sun, Yufeng Lu, Hao Dong, Xiaoying Fan, Zeyuan Liu, Ruiguo Chen, Shu Zhang, Qiang Ma, et al. Decoding the development of the human hippocampus. *Nature*, 577(7791):531–536, 2020.

78. Ruth Benavides-Piccione, Mamen Regalado-Reyes, Isabel Fernaud-Espinosa, Asta Kastanauskaite, Silvia Tapia-González, Gonzalo León-Espinosa, Concepcion Rojo, Ricardo Insausti, Idan Segev, and Javier DeFelipe. Differential structure of hippocampal ca1 pyramidal neurons in the human and mouse. *Cerebral Cortex*, 30(2):730–752, 2020.

79. Miri Adler, Yael Korem Kohanim, Avichai Tendler, Avi Mayo, and Uri Alon. Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. *Cell systems*, 8(1):43–52, 2019.

80. Shani Ben-Moshe and Shalev Itzkovitz. Spatial heterogeneity in the mammalian liver. *Nature Reviews Gastroenterology & Hepatology*, 16(7):395–410, 2019.

81. Romain Donne, Maëva Saroul-Aïnama, Pierre Cordier, Séverine Celton-Morizur, and Chantal Desdouets. Polyploidy in liver development, homeostasis and disease. *Nature Reviews Gastroenterology & Hepatology*, 17(7):391–405, 2020.

82. Teemu P Miettinen, Heli KJ Pessa, Matias J Caldez, Tobias Fuhrer, M Kasim Diril, Uwe Sauer, Philipp Kaldis, and Mikael Björklund. Identification of transcriptional and metabolic programs related to mammalian cell size. *Current Biology*, 24(6):598–608, 2014.

83. ML Richter, IK Deligiannis, K Yin, A Danese, E Lleshi, P Coupland, Catalina A Vallejos, KP Matchett, NC Henderson, M Colome-Tatche, et al. Single-nucleus rna-seq2 reveals functional crosstalk between liver zonation and ploidy. *Nature communications*, 12(1):1–16, 2021.

84. Keren Bahar Halpern, Rom Shenhav, Orit Matcovitch-Natan, Beáta Tóth, Doron Lemze, Matan Golan, Efi E Massasa, Shaked Baydatch, Shanie Landen, Andreas E Moor, et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*, 542(7641):352–356, 2017.

85. Keren Bahar Halpern, Sivan Tanami, Shanie Landen, Michal Chapal, Liran Szlak, Anat Hutzler, Anna Nizhberg, and Shalev Itzkovitz. Bursty gene expression in the intact mammalian liver. *Molecular cell*, 58(1):147–156, 2015.

86. Aris Taychameekiatchai and Bruce Wang. Tentative title. *Manuscript in preparation*, 2021.

87. Nadim Aizarani, Antonio Saviano, Laurent Mailly, Sarah Durand, Josip S Herman, Patrick Pessaux, Thomas F Baumert, Dominic Grün, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature*, 572(7768):199–204, 2019.

88. Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

89. Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

90. Meromit Singer, Chao Wang, Le Cong, Nemanja D Marjanovic, Monika S Kowalczyk, Huiyuan Zhang, Jackson Nyman, Kaori Sakuishi, Sema Kurtulus, David Gennert, et al. A distinct gene module for dysfunction uncoupled from activation in tumor-infiltrating t cells. *Cell*, 166(6):1500–1511, 2016.

91. Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.

92. Jeremy A Miller, Steve Horvath, and Daniel H Geschwind. Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proceedings of the National Academy of Sciences*, 107(28):12698–12703, 2010.

93. Elle M Weeks, Jacob C Ulirsch, Nathan Y Cheng, Brian L Trippe, Rebecca S Fine, Jenkai Miao, Tejal A Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P Fulco, et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *medRxiv*, 2020.

94. Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal Dey, Joseph Nasser, Karthik Jagadeesh, Daniel Weiner, Huwenbo Shi, Charles Fulco, Luke O'Connor, et al. Combining snp-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity. *medRxiv*, 2021.

95. Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

96. Daniel B Burkhardt, Jay S Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio J Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, pages 1–11, 2021.

97. Genomes Project Consortium, A Auton, LD Brooks, RM Durbin, EP Garrison, and HM Kang. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

98. Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

99. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

100. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

101. Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

102. Katrina M De Lange, Loukas Moutsianas, James C Lee, Christopher A Lamb, Yang Luo, Nicholas A Kennedy, Luke Jostins, Daniel L Rice, Javier Gutierrez-Achury, Sun-Gou Ji, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, 49(2):256–261, 2017.

103. Patrick CA Dubois, Gosia Trynka, Lude Franke, Karen A Hunt, Jihane Romanos, Alessandra Curtotti, Alexandra Zhernakova, Graham AR Heap, Róza Ádány, Arpo Aromaa, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nature genetics*, 42(4):295–302, 2010.

104. James Bentham, David L Morris, Deborah S Cunninghame Graham, Christopher L Pinder, Philip Tombleson, Timothy W Behrens, Javier Martín, Benjamin P Fairfax, Julian C Knight, Lingyan Chen, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature genetics*, 47(12):1457–1464, 2015.

105. Farren Briggs, Xiaorong Shao, Benjamin A Goldstein, Jorge R Oksenberg, Lisa F Barcellos, and Philip L De Jager. Genome-wide association study of severity in multiple sclerosis. *Genes & Immunity*, 12(8), 2011.

106. Heather J Cordell, Younghun Han, George F Mells, Yafang Li, Gideon M Hirschfield, Casey S Greene, Gang Xie, Brian D Juran, Dakai Zhu, David C Qian, et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature communications*, 6(1):1–11, 2015.

107. Yukinori Okada, Di Wu, Gosia Trynka, Towfique Raj, Chikashi Terao, Katsunori Ikari, Yuta Kochi, Koichiro Ohmura, Akari Suzuki, Shinji Yoshida, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506(7488):376–381, 2014.

108. Ditte Demontis, Raymond K Walters, Joanna Martin, Manuel Mattheisen, Thomas D Als, Esben Agerbo, Gísli Baldursson, Rich Belliveau, Jonas Bybjerg-Grauholm, Marie Bækvad-Hansen, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics*, 51(1):63–75, 2019.

109. Iris E Jansen, Jeanne E Savage, Kyoko Watanabe, Julien Bryois, Dylan M Williams, Stacy Steinberg, Julia Sealock, Ida K Karlsson, Sara Hägg, Lavinia Athanasiu, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing alzheimer's disease risk. *Nature genetics*, 51(3):404–413, 2019.

110. Eli A Stahl, Gerome Breen, Andreas J Forstner, Andrew McQuillin, Stephan Ripke, Vassily Trubetskoy, Manuel Mattheisen, Yunpeng Wang, Jonathan RI Coleman, Héléna A Gaspar, et al. Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics*, 51(5):793–803, 2019.

111. Mengzhen Liu, Yu Jiang, Robbee Wedow, Yue Li, David M Brazel, Fang Chen, Gargi Datta, Jose Davila-Velderrain, Daniel McGuire, Chao Tian, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics*, 51(2):237–244, 2019.

112. Richard Karlsson Linnér, Pietro Biroli, Edward Kong, S Fleur W Meddens, Robbee Wedow, Mark Alan Fontana, Maël Lebreton, Stephen P Tino, Abdel Abdellaoui, Anke R Hammerschlag, et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature genetics*, 51(2):245–257, 2019.

113. Jeanne E Savage, Philip R Jansen, Sven Stringer, Kyoko Watanabe, Julien Bryois, Christiaan A De Leeuw, Mats Nagel, Swapnil Awasthi, Peter B Barr, Jonathan RI Coleman, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics*, 50(7):912–919, 2018.

114. David M Howard, Mark J Adams, Toni-Kim Clarke, Jonathan D Hafferty, Jude Gibson, Masoud Shirali, Jonathan RI Coleman, Saskia P Hagenaars, Joey Ward, Eleanor M Wigmore, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience*, 22(3):343–352, 2019.

115. Gail Davies, Max Lam, Sarah E Harris, Joey W Trampush, Michelle Luciano, W David Hill, Saskia P Hagenaars, Stuart J Ritchie, Riccardo E Marioni, Chloe Fawns-Ritchie, et al. Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nature communications*, 9(1):1–16, 2018.

116. Aysu Okbay, Bart ML Baselmans, Jan-Emmanuel De Neve, Patrick Turley, Michel G Nivard, Mark Alan Fontana, S Fleur W Meddens, Richard Karlsson Linnér, Cornelius A Rietveld, Jaime Derringer, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics*, 48(6):624–633, 2016.

117. Douglas M Ruderfer, Stephan Ripke, Andrew McQuillin, James Boocock, Eli A Stahl, Jennifer M Whitehead Pavlides, Niamh Mullins, Alexander W Charney, Anil PS Ori, Loes M Olde Loohuis, et al. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, 173(7):1705–1715, 2018.

118. Hassan S Dashti, Samuel E Jones, Andrew R Wood, Jacqueline M Lane, Vincent T Van Hees, Heming Wang, Jessica A Rhodes, Yanwei Song, Krunal Patel, Simon G Anderson, et al. Genome-wide association study identifies genetic loci for self-reported habitual sleep duration supported by accelerometer-derived estimates. *Nature communications*, 10(1):1–12, 2019.

119. Mats Nagel, Philip R Jansen, Sven Stringer, Kyoko Watanabe, Christiaan A De Leeuw, Julien Bryois, Jeanne E Savage, Anke R Hammerschlag, Nathan G Skene, Ana B Muñoz-Manchado, et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature genetics*, 50(7):920–927, 2018.

120. Jonas B Nielsen, Lars G Fritsche, Wei Zhou, Tanya M Teslovich, Oddgeir L Holmen, Stefan Gustafsson, Maiken E Gabrielsen, Ellen M Schmidt, Robin Beaumont, Brooke N Wolford, et al. Genome-wide study of atrial fibrillation identifies seven risk loci and highlights biological pathways and regulatory elements involved in cardiac development. *The American Journal of Human Genetics*, 102(1):103–115, 2018.

121. Heribert Schunkert, Inke R König, Sekar Kathiresan, Muredach P Reilly, Themistocles L Assimes, Hilma Holm, Michael Preuss, Alexandre FR Stewart, Maja Barbalic, Christian Gieger, et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nature genetics*, 43(4):333–338, 2011.

122. Alisa K Manning, Marie-France Hivert, Robert A Scott, Jonna L Grimsby, Nabila Bouatia-Naji, Han Chen, Denis Rybin, Ching-Ti Liu, Lawrence F Bielak, Inga Prokopenko, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics*, 44(6):659–669, 2012.

123. Jonathan P Bradfield, Hui-Qi Qu, Kai Wang, Haitao Zhang, Patrick M Sleiman, Cecilia E Kim, Frank D Mentch, Haijun Qiu, Joseph T Glessner, Kelly A Thomas, et al. A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS genetics*, 7(9):e1002293, 2011.

124. Andrew P Morris, Benjamin F Voight, Tanya M Teslovich, Teresa Ferreira, Ayellet V Segre, Valgerdur Steinthorsdottir, Rona J Strawbridge, Hassan Khan, Harald Grallert, Anubha Mahajan, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nature genetics*, 44(9):981, 2012.

125. Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015.

126. Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, 16(12):1289–1296, 2019.

127. Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

128. Farhad Hormozdiari, Steven Gazal, Bryce Van De Geijn, Hilary K Finucane, Chelsea J-T Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke O'connor, et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature genetics*, 50(7):1041–1047, 2018.

129. Steven Gazal, Po-Ru Loh, Hilary K Finucane, Andrea Ganna, Armin Schoech, Shamil Sunyaev, and Alkes L Price. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nature genetics*, 50(11):1600–1607, 2018.