nature genetics

Article

https://doi.org/10.1038/s41588-022-01187-9

Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics

Received: 17 May 2021

Accepted: 18 August 2022

Published online: 29 September 2022

Check for updates

Karthik A. Jagadeesh[1,2,8] ✉, Kushal K. Dey[1,2,8] ✉, Daniel T. Montoro[1], Rahul Mohan[1], Steven Gazal[2], Jesse M. Engreitz[1,3,4], Ramnik J. Xavier[1], Alkes L. Price[1,2,5,9] ✉ and Aviv Regev[1,6,7,9] ✉

Genome-wide association studies provide a powerful means of identifying loci and genes contributing to disease, but in many cases, the related cell types/states through which genes confer disease risk remain unknown. Deciphering such relationships is important for identifying pathogenic processes and developing therapeutics. In the present study, we introduce sc-linker, a framework for integrating single-cell RNA-sequencing, epigenomic SNP-to-gene maps and genome-wide association study summary statistics to infer the underlying cell types and processes by which genetic variants influence disease. The inferred disease enrichments recapitulated known biology and highlighted notable cell–disease relationships, including γ-aminobutyric acid-ergic neurons in major depressive disorder, a disease-dependent M-cell program in ulcerative colitis and a disease-specific complement cascade process in multiple sclerosis. In autoimmune disease, both healthy and disease-dependent immune cell-type programs were associated, whereas only disease-dependent epithelial cell programs were prominent, suggesting a role in disease response rather than initiation. Our framework provides a powerful approach for identifying the cell types and cellular processes by which genetic variants influence disease.

Genome-wide association studies (GWASs) have successfully identified thousands of disease-associated variants[1–3], but the cellular mechanisms through which these variants drive complex diseases and traits remain largely unknown. This is due to several challenges, including the difficulty of relating the approximately 95% of risk variants that reside in noncoding regulatory regions to the genes they regulate[4–7] and our limited knowledge of the specific cells and functional programs in which these genes are active[8]. Previous studies have linked traits to functional elements[9–15] and to cell types using bulk RNA-sequencing (RNA-seq) profiles[16–18]. Considerable work remains to analyze cell types and states at finer resolutions across a breadth of tissues, incorporate disease tissue-specific gene expression patterns, model cellular

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [3]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [4]BASE Initiative, Betty Irene Moore Children's Heart Center, Lucile Packard Children's Hospital, Stanford University School of Medicine, Stanford, CA, USA. [5]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [6]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Present address: Genentech, South San Francisco, CA, USA. [8]These authors contributed equally: Karthik A. Jagadeesh, Kushal K. Dey. [9]These authors jointly supervised this work: Alkes L. Price, Aviv Regev. ✉e-mail: kjag@cs.stanford.edu; kdey@hsph.harvard.edu; aprice@hsph.harvard.edu; aviv.regev.sc@gmail.com

processes within and across cell types and leverage enhancer–gene links[19–23] to improve power.

Single-cell RNA-seq (scRNA-seq) data provide a unique opportunity to tackle these challenges[24]. Single-cell profiles allow the construction of multiple gene programs to more finely relate GWAS variants to function, including programs that reflect cell-type-specific signatures[25–28], disease-dependent signatures within cell types[29,30] and key cellular processes that vary within and/or across cell types[31]. Initial studies have related single-cell profiles with human genetics in post-hoc analyses by mapping candidate genes from disease-associated genomic regions to cell types by their expression relative to other cell types[32–34]. More recent studies have begun to leverage genome-wide polygenic signals to map traits to cell types from single cells within the context of a single tissue[35–37]. However, focusing on a single tissue could, in principle, result in misleading conclusions, because disease mechanisms span tissue types across the human body. For example, in the context of the colon, a neural gene associated with psychiatric disorders would appear highly specific to enteric neurons, but this cell population may no longer be strongly implicated when the analysis also includes cells from the human central nervous system[38]. Thus, there is a need for a principled method that combines human genetics and comprehensive scRNA-seq applied across multiple tissues and organs.

In the present study, we develop and apply sc-linker, an integrated framework to relate human disease and complex traits to cell types and cellular processes by integrating GWAS summary statistics, epigenomics and scRNA-seq data from multiple tissue types, diseases, individuals and cells. Unlike previous studies, we analyze gene programs that represent different functional facets of cells, including discrete cell types, processes activated specifically in a cell type in disease and processes activated across cells irrespective of cell-type definitions (recovered by latent factor models). We transform gene programs to SNP annotations using tissue-specific enhancer–gene links[19–23] in preference to standard gene window-based linking strategies used in existing gene-set enrichment methods such as MAGMA[39], RSS-E[13] and linkage disequilibrium score regression (LDSC)-specifically expressed genes[18]. We then link SNP annotations to diseases by applying stratified LDSC[11] (S-LDSC) using the baseline-LD model[40,41] to the resulting SNP annotations. We further integrate cellular expression and GWAS to prioritize specific genes in the context of disease-critical gene programs, thus providing new insights into underlying disease mechanisms.

Results

Overview of sc-linker

We developed a framework to link gene programs derived from scRNA-seq with diseases and complex traits (Fig. 1a). First, we use scRNA-seq to construct gene programs, defined as continuous-valued gene sets, that characterize (1) individual cell types, (2) disease-dependent (disease versus healthy cells of the same type) or (3) cellular processes (cell cycling, endoplasmic reticulum stress). (The continuous values are on the probabilistic 0–1 scale but do not formally represent probabilities (Methods).) Then, we link the genes underlying these programs to SNPs that regulate them by incorporating two tissue-specific, enhancer–gene-linking strategies: Roadmap Enhancer-Gene Linking[19–21] and the Activity-by-Contact (ABC) model[22,23]. Finally, we evaluate the disease informativeness of the resulting SNP annotations by applying S-LDSC[11] conditional on a broad set of coding, conserved, regulatory and LD-related annotations from the baseline-LD model[40,41]. Altogether, our approach links diseases and traits with gene programs recapitulating cell types and cellular processes. We have released open-source software implementing the approach (sc-linker; see Code availability), a web interface for visualizing the results (Data availability) and postprocessed scRNA-seq data, gene programs, enhancer–gene-linking strategies and SNP annotations analyzed in the present study (Data availability). A more comprehensive overview is provided in Supplementary Note.

We analyzed a broad range of human scRNA-seq data, spanning 17 datasets from 11 tissues and 6 disease conditions. The 11 nondisease tissues included blood/immune (peripheral blood mononuclear cells (PBMCs)[26,42], cord blood[27] and bone marrow[27]), brain[28], kidney[43], liver[44], heart[25], lung[29], colon[34], skin[45] and adipose tissue[44]. The six disease conditions included multiple sclerosis (MS, brain)[46], Alzheimer's disease (AD, brain)[30], ulcerative colitis (UC, colon)[34], asthma, lung[47], idiopathic pulmonary fibrosis (IPF), lung[29] and COVID-19, bronchoalveolar lavage fluid (BAL)[48] (Extended Data Fig. 1). In total, the scRNA-seq data included 209 individuals, 1,602,614 cells and 256 annotated cell subsets (Methods and Supplementary Table 1). We also compiled publicly available GWAS summary statistics for 60 unique diseases and complex traits (genetic correlation <0.9; average $N = 297{,}000$) (Methods and Supplementary Table 2). We analyzed gene programs from each scRNA-seq dataset together with each of 60 diseases and complex traits, but we primarily reported those that are most pertinent for each program.

Benchmarking the sc-linker

As a proof of principle, we benchmarked the sc-linker by analyzing five blood cell traits that biologically correspond to specific immune cell types (Supplementary Table 2) using immune cell-type programs constructed from scRNA-seq data (Fig. 2a,b and Extended Data Fig. 1). We constructed six immune cell-type programs that were identified across four datasets: two from PBMCs ($k = 4{,}640$ cells, $n = 2$ individuals[26]; $k = 68{,}551$, $n = 8$ (ref. [42])) and one each of cord blood[27] ($k = 263{,}828$, $n = 8$) and bone marrow[27] ($k = 283{,}894$, $n = 8$). We identified enrichment of erythroid cells for red blood cell count, megakaryocytes for platelet count, monocytes for monocyte count and B cells and T cells for lymphocyte percentage (Fig. 2d and Extended Data Fig. 2a); these enrichments reflect known biological roles and have been reported in previous studies[49,50], such that we refer to them as expected enrichments.

We defined a sensitivity/specificity index quantifying the presence of expected enrichments and absence of other enrichments (Methods). A limitation of this index is that other enrichments may be biologically real in some cases; thus, we also consider sensitivity to detect expected enrichments (Methods). The sc-linker outperformed the MAGMA[39] gene-set-level association method in terms of the sensitivity/specificity index (Fig. 2c). Benchmarks on the sc-linker method, the choice of enhancer–gene-linking strategies and cell-type programs are included in Supplementary Note.

Distinguishing the cells involved in immune-related diseases

We next analyzed eleven autoimmune diseases (Supplementary Table 2) using the six immune cell-type programs above (Fig. 2a,b and Extended Data Fig. 1) and ten (intracell and intercell types) immune cellular process programs (Fig. 2f). (Enrichment results for the remaining 49 diseases and traits with immune cell-type programs are reported in Extended Data Fig. 3; we did not construct disease-dependent programs, because these datasets included healthy samples only.) We identified cell-type-disease enrichments that conform to known disease biology (Fig. 2e and Extended Data Fig. 2b), including T cells for eczema[51,52], B and T cells for primary biliary cirrhosis (PBC)[18] and dendritic cells (DCs) and monocytes for AD[53]. In addition, the highly statistically significant enrichments for MS across all six immune cell-type programs analyzed are consistent with previous analyses[18,54–56], supporting the validity of our approach.

Several of the significant cell-type-disease enrichments have limited literature support and may implicate previously unexplored biological mechanisms (Fig. 2e, Table 1 and Extended Data Fig. 2b). For example, we detected significant enrichment in B cells for UC; B cells have been detected in basal lymphoid aggregates in the UC in the colon, but their pathogenic significance remains unknown[57]. In addition, T cells were highly enriched for celiac disease, the top driving genes including *ETS1* (ranked 1), associated with T cell development

https://doi.org/10.1038/s41588-022-01187-9

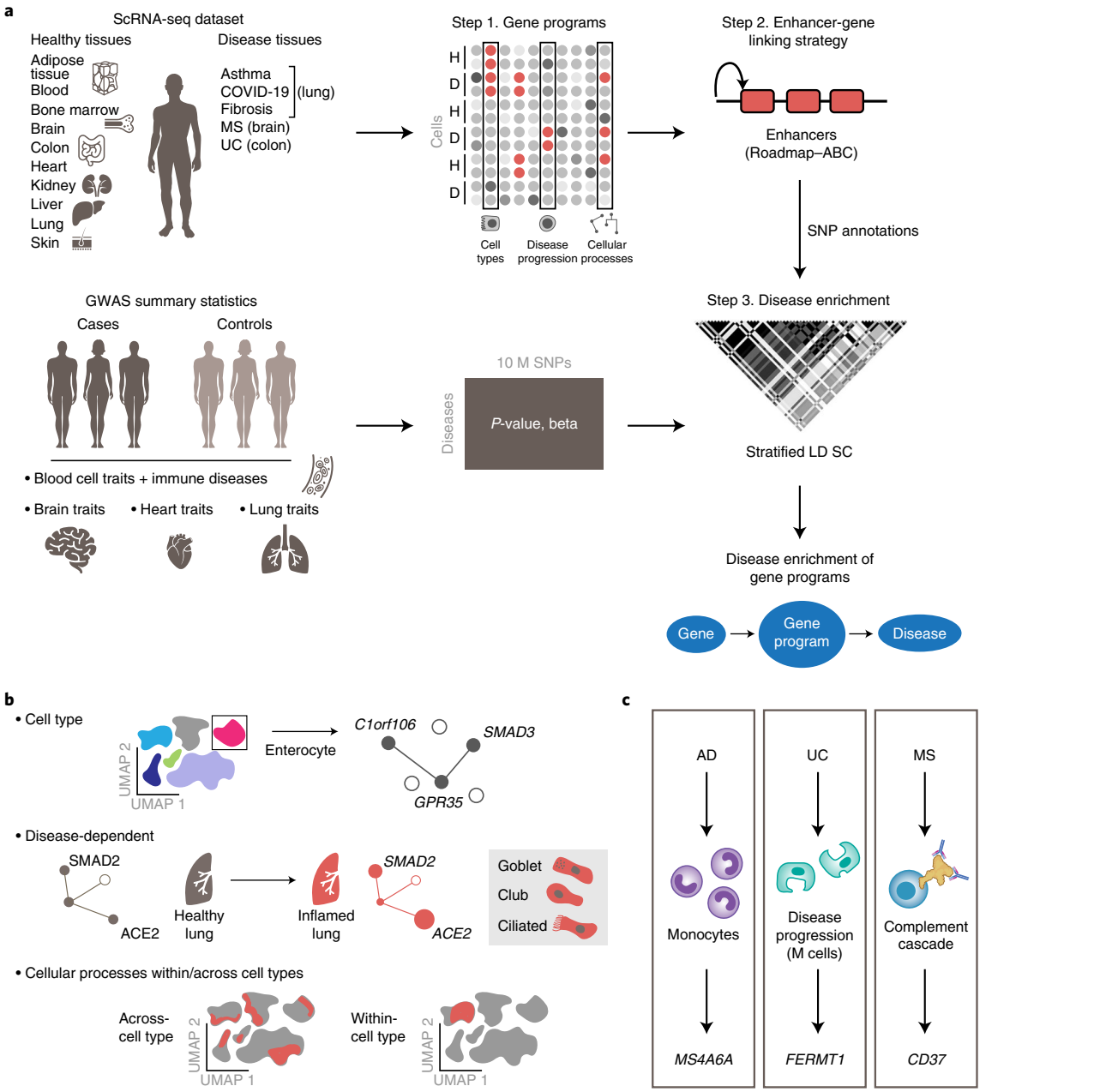


Fig. 1 | Approach for identifying disease-critical cell types and cellular processes by integration of single-cell profiles and human genetics. a, The sc-linker framework. Left: input. scRNA-seq (top) and GWAS (bottom) data. Middle and right: step 1: deriving cell-type, disease-dependent and cellular process gene programs from scRNA-seq (top) and associating SNPs with traits from human GWASs (bottom). Step 2: generation of SNP annotations. Gene programs are linked to SNPs by enhancer–gene-linking strategies to generate SNP annotations. Step 3: S-LDSC is applied to the resulting SNP annotations to evaluate heritability enrichment for a trait. **b**, Constructing gene programs. Top: cell-type programs of genes specifically expressed in one cell type versus others. Middle: disease-dependent programs of genes specifically expressed in cells of the same type in disease versus healthy samples. Bottom: cellular process programs of genes co-varying either within or across cell subsets; these programs may be healthy specific, disease specific or shared. **c**, Examples of disease–gene, program–gene relationships recovered by our framework.

and interleukin (IL)-2 signaling[58], and *CD28* (ranked 3), critical for T cell activation. This suggests that aberrant T cell maintenance and activation may impact inflammation in celiac disease. Recent reports of a permanent loss of resident γδ T cells in the celiac bowel and the subsequent recruitment of inflammatory T cells may further support this hypothesis[59]. These results were recapitulated across an independent immune cell scRNA-seq dataset, in both the gene programs (average correlation: 0.78 for the same cell type) and the disease enrichments (0.86 correlation of the *E*-score over all cell-type and -trait pairs). A cross-trait analysis of the patterns of cell-type enrichments suggests that celiac disease and rheumatoid arthritis involve cell-mediated adaptive immune response, UC and PBC involve antibody-mediated adaptive immune response, AD has a strong signal of innate immune and MS and inflammatory bowel disease (IBD) involve contributions from a wide range of immune cell types (Extended Data Fig. 4).

Analyzing the ten immune cellular process programs (Fig. 2f) across the eleven immune-related diseases and five blood cell traits, we identified both disease-specific enrichments and others that shared across diseases (Fig. 2g and Table 1). For example, although T cells have been previously linked to eczema, we pinpointed higher enrichment

https://doi.org/10.1038/s41588-022-01187-9

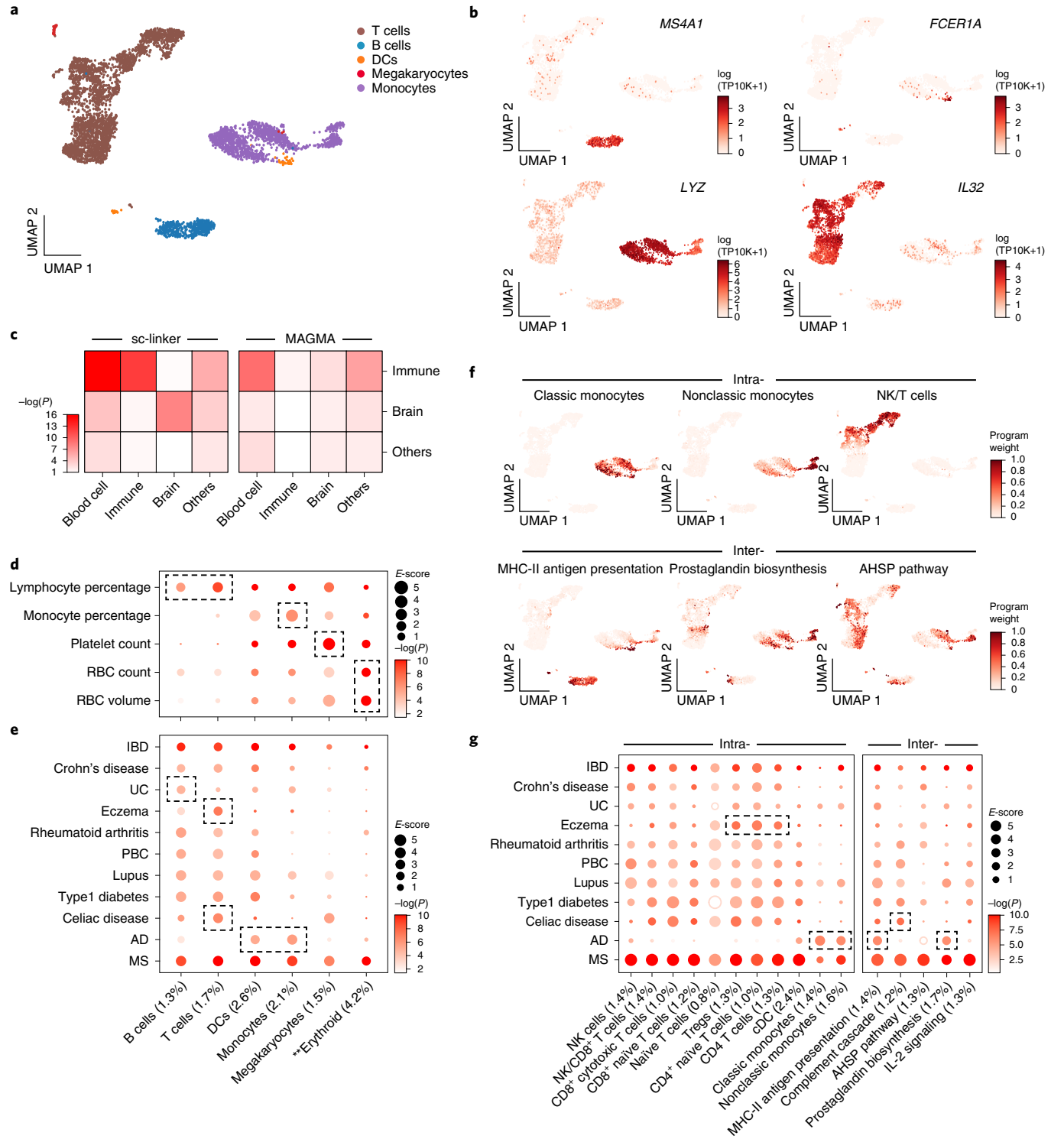


Fig. 2 | Linking immune cell types and cellular processes to immune-related diseases and blood cell traits. a,**b**, Immune cell types. UMAP embedding of PBMC scRNA-seq profiles (dots) colored by cell-type annotations (**a**) or expression of cell-type-specific genes (**b**). **c**, Benchmarking of sc-linker versus MAGMA. Significance (average $-\log_{10}(P)$) of association across immune, brain and other tissue cell-type programs (rows) and blood cell, immune-related, brain-related and other traits (columns) for the sc-linker (left) and MAGMA gene-set analysis (right). Other cell types × other diseases/traits are not included in the specificity calculation, due to the broad set of cell types and diseases/traits in this category. For the MAGMA analysis, the gene program is binarized using a threshold = 0.95; the numerical results for other binarization thresholds and continuous variable-based approaches are reported in Supplementary Data 7. **d**,**e**, Enrichments of immune cell-type programs for blood cell traits and immune-related diseases. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of immune cell-type programs (columns) are shown for blood cell traits (rows, **d**) or immune-related diseases (rows, **e**). **f**, Examples of inter- and intracell-type cellular process programs. UMAP of PBMCs (as in **a**) are colored by each program weight (color bar) from NMF. NK, natural killer. **g**, Enrichments of immune cellular process programs for immune-related diseases. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of cellular process programs (columns) are given for immune-related diseases (rows). In **d**, **e** and **g**, the size of each corresponding SNP annotation (percentage SNPs) is reported in parentheses, and the dashed boxes denote results that are highlighted in the main text. Numerical results are reported in Supplementary Data 1 and 3. Further details of all diseases and traits analyzed are provided in Supplementary Table 2. **Erythroid cells were observed in only bone marrow and cord blood datasets.

https://doi.org/10.1038/s41588-022-01187-9

Table 1 | Notable enrichments from analyses of cell-type, disease-dependent and cellular process gene programs

Cell-type programs						
GWAS disease/trait	**Tissue (scRNA-seq)**	**Cell type**	***E*-score**	***P* (*E*-score)**	***q*-value**	**Top genes**
UC	Blood/Immune	B cells	3.2	1.50×10^{-5}	2.33×10^{-5}	*REL, GPX1, LSP1*
Celiac disease	Blood/Immune	T cells	4.5	2.3×10^{-7}	7.16×10^{-7}	*ETS1, CD247, CD28*
MDD	Brain	GABA-ergic	4	1.00×10^{-4}	3.39×10^{-4}	*TCF4, BEND4, TMX2*
Atrial fibrillation	Heart	Atrial cardiomyocyte	5.6	3.2×10^{-9}	2.2×10^{-8}	*CAV2, PKD2L2, FAM13B*
Blood pressure (dia.)	Heart	Smooth muscle	3.4	2.9×10^{-6}	1.2×10^{-5}	*CACNB2, TMEM165, MRVI1*
Eczema	Skin	Langerhans' cells	3.7	0.004	0.03	*IL1R1, RUNX3, FCER1G*
IBD	Colon	Endothelial	2.8	0.002	0.01	*RHOA, PDLIM4, STARD3*
Disease-dependent programs						
GWAS disease/trait	Tissue (scRNA-seq)	Cell type	*E*-score	*P* (*E*-score)	*q*-value	Top genes
MS	MS, brain	Microglia	11.6	5.70×10^{-6}	3.66×10^{-5}	*PRDX5, RPL5, SKP1,*
AD	AD, brain	Microglia	9.1	7.10×10^{-5}	6.82×10^{-4}	*PICALM, APOE, APOC1*
UC	UC, colon	Enterocytes	2.6	2.70×10^{-7}	1.66×10^{-6}	*RNF186, APEH, DLD*
IBD	UC, colon	M cells	2.2	1.07×10^{-4}	2.2×10^{-4}	*UQCR10, FERMT1, PPP1R1B*
Asthma	Asthma, lung	T cells	12.8	4.82×10^{-5}	3.99×10–4	*FMNL1, RORA, GPR183*
Cellular process programs						
GWAS disease/trait	Tissue (scRNA-seq)	Cellular process	*E*-score	*P* (*E*-score)	*q*-value	Top genes
Eczema	Blood/Immune	$CD4^+$ T cells	3.8	1.32×10^{-7}	4.83×10^{-7}	*IL7R, STMN3, NDFIP1*
Celiac disease	Blood/Immune	Complement cascade	2.8	4.84×10^{-8}	1.92×10^{-7}	*DCC, PDIA5, PPCDC*
AD	Blood/Immune	MHC-II antigen processing	4.9	7.11×10^{-0}	2.08×10^{-6}	*MS4A6A, MS4A4A, CD33*
BMI	Brain	LAMP5	2.7	6.33×10^{-8}	7.01×10^{-7}	*FLRT1, COL4A2, SBF2*
MDD	Brain	SST	3.9	4.37×10^{-5}	1.22×10^{-4}	*TCF4, PCLO, ZNF462*
Years of education	Brain	Electron transport	3.5	4.42×10^{-8}	5.49×10^{-7}	*ATP6V0B, NSF, GPX1*
MS	MS, brain	Complement cascade[a]	4.9	5.49×10^{-11}	9.62×10^{-10}	*CD37, RGS14, NCF4*
AD	AD, brain	Apelin signaling[b]	1.5	9.27×10^{-7}	6.50×10^{-6}	*MS4A6A, SORL1, SYK*
UC	UC, colon	EGFR-1 pathway[b]	3.0	8.81×10^{-4}	2.14×10^{-3}	*C1orf106, SLC26A3, NXPE4*
Asthma	Asthma, lung	Mac-neutrophil trans[b]	6.6	0.002	0.006	*CCL20, IL6, GPR183*

For each notable enrichment, we report the GWAS disease/trait, tissue source for scRNA-seq data, cell type, enrichment score (*E*-score), one-sided S-LDSC *P* value for positive *E*-score and top genes driving the enrichment. Multiple testing correction was performed across cell types and traits at the level of each tissue. Blood pressure (dia.), diastolic blood pressure; mac-neutrophil trans., macrophage–neutrophil transition.[a]Cellular process programs specific to disease states. The full list of genes driving these associations is provided in Supplementary Data 4.[b]Cellular process programs shared across healthy and disease states.

in $CD4^+$ T cells compared with $CD8^+$ T cells. The IL-2 signaling cellular process program in T and B cells was significantly enriched for both eczema and celiac disease, although the genes driving the enrichment were not significantly overlapping ($P = 0.21$). In addition, the complement cascade cellular process program in plasma, B cells and hematopoietic stem cells was most highly enriched among all intercellular programs for celiac disease. For AD, there was a strong enrichment in both classic and nonclassic, monocyte intracell-type cellular programs, and in major histocompatibility complex class II (MHC-II) antigen presentation (intercell type: dendritic cells (DCs) and B cells) and prostaglandin biosynthesis (intercell type: monocytes, DCs, B cells and T cells) programs. Among the notable driver genes were *IL7R* (ranked 1) and *NDFIP1* (ranked 3) for $CD4^+$ T cells in eczema, which respectively play key roles in helper T cell 2 differentiation[60,61] and in mediating peripheral *CD4* T cell tolerance and allergic reactions[62,63], and *CD33* (ranked 1) in MHC-II antigen processing in AD, a microglial receptor strongly associated with increased risk in previous GWASs[64,65].

Linking GABA-ergic and glutamatergic neurons to psychiatric disease

We next focused on brain cells and psychiatric disease, by analyzing 9 cell-type programs (Fig. 3a) and 12 cell process programs (Fig. 3e; 10 intra- and 2 intercell-type programs) from scRNA-seq data of healthy brain prefrontal cortex ($k = 73{,}191$, $n = 10$)[28] (Supplementary Table 1) with 11 psychiatric or neurological diseases and traits (Supplementary Table 2).

Notably, we observed enrichments of major depressive disorder (MDD) and body mass index (BMI) specifically in γ-aminobutyric acid (GABA)-ergic neurons, whereas insomnia, schizophrenia and intelligence were highly enriched, specifically in glutamatergic neurons, and neuroticism was highly enriched in both. GABA-ergic neurons regulate the brain's ability to control stress levels, which is the most prominent vulnerability factor in MDD[66] (Fig. 3b,c, Table 1 and Extended Data Fig. 2c). Among the top genes driving this enrichment were *TCF4* (ranked 1), a critical component for neuronal differentiation that affects neuronal migration patterns[67,68], and *PCLO* (ranked 4), which is important for synaptic vesicle trafficking and neurotransmitter release[69]. Although predominant therapies for MDD target monoamine neurotransmitters, especially serotonin, the enrichment for GABA-ergic neurons is independent of serotonin pathways, suggesting that they might include other therapeutic targets for MDD. These results were robustly detected in an independent brain scRNA-seq dataset, in both the gene programs (average correlation: 0.77 for the same cell type and −0.21 otherwise) and the disease enrichments (0.77 correlation of the

E-score over all cell-type and -trait pairs), including GABA-ergic neurons in MDD and BMI as well as glutamatergic neurons in insomnia and schizophrenia. Enrichment results for the remaining 49 diseases and traits together with brain cell-type programs are reported in Extended Data Fig. 3.

Tissue specificity of both the cell-type program and the enhancer–gene strategy was important for successful linking, which we found by comparing the enrichment of all four possible combinations of immune or brain cell-type programs with immune- or brain-specific enhancer–gene-linking strategies, meta-analyzed across 11 immune-related diseases or 11 psychiatric/neurological diseases and traits (Fig. 3d). This highlights the importance of leveraging the tissue specificity of enhancer–gene strategies.

The 12 brain cellular process programs showed that the significant enrichment of brain-related diseases in the neuronal cell types above is primarily driven by finer programs reflecting neuron subtypes (Fig. 3f, Table 1 and Supplementary Note). For example, the enrichment of GABA-ergic neurons for BMI was driven by programs reflecting *LAMP5*$^+$ and *VIP*$^+$ cell subsets with higher expression of *LAMP5* and *VIP*, respectively. Furthermore, the enrichment of GABA-ergic neurons for MDD reflects *SST*$^+$ and *PVALB*$^+$ cell subsets with higher expression of *SST* and *PVALB*, respectively. We also observed enrichment in more specific cell subsets within glutamatergic neurons (for example, inferior temporal (IT) neurons were enriched for neuroticism).

Linking cell types from diverse human tissues to disease

Analysis of kidney, liver, heart, skin and adipose tissuse cell types (Supplementary Table 1) and corresponding relevant traits (Supplementary Table 2) revealed the role of particular immune, stromal and epithelial cellular compartments across different diseases/traits. For example, kidney and liver cell-type programs (Extended Data Fig. 1) highlighted relations with urine biomarker traits (Fig. 4a and Extended Data Figs. 3 and 5a,b), such as enrichment for creatinine level in kidney proximal and connecting tubule cell types, but not in liver cell types, as expected[70,71], or a significant enrichment for bilirubin level only in liver hepatocytes (driven by *ANGPTL3*; ranked 4)[72,73]. In heart (Fig. 4b, Extended Data Figs. 3 and 5c and Table 1), atrial cardiomyocytes were enriched for atrial fibrillation, and pericytes and smooth muscle cells for blood pressure, consistent with their respective roles in determining heart rhythm through activity[74] of ion channels (top genes included the ion channel genes *PKD2L2* (ranked 2), *CASQ2* (ranked 7) and *KCNN2* (ranked 18)) and blood pressure regulation through vascular tone[75] (top driving genes included adrenergic pathway genes *PLCE1* (ranked 1), *CACNA1C* (ranked 21) and *PDE8A* (ranked 23)). In skin (Fig. 4c, Extended Data Fig. 3 and Table 1), both brain-derived neurotrophic factor signaling and Langerhans' cells were enriched for eczema. Langerhans' cells have been implicated in inflammatory skin processes related to eczema[76] (top driving genes included IL-2-signaling pathway genes (*FCER1G* (ranked 3), *NR4A2* (ranked 26) and *CD52* (ranked 43)), which modulate eczema pathogenesis[77]). In adipose (Fig. 4d and Extended Data Figs. 3 and 5e), adipocytes were enriched for BMI, driven by adipogenesis pathway genes[78] (*STAT5A* (ranked 15), *EBF1* (ranked 29), *LIPE* (ranked 45)) and triglyceride biosynthesis genes[78] (*GPAM* (ranked 14), *LIPE* (ranked 45), both of which contribute to the increase in adipose tissue mass in obesity[79,80]).

We expanded our analysis to evaluate all cell-type programs for all diseases, irrespective of the tissue locus of disease, aiming to identify cell-type enrichments involving 'mismatched' cell-type disease/trait pairs (Supplementary Fig. 5). As expected, in most cases, 'mismatched' cell-type programs and disease/trait pairs do not yield significant association. Notable exceptions included enrichments of skin Langerhans' cells for AD (*E*-score: 15.2, $P = 10^{-4}$), M cells (in colon) for asthma (*E*-score: 2.2, $P = 10^{-4}$) and heart smooth muscle cells for lung capacity (*E*-score: 5.6, $P = 3 \times 10^{-4}$). In some cases, the association may indicate a direct relationship, whereas in others the associated cell type may only 'tag' the causal cell type in the disease tissue, as cell-type programs derived from cells of the same type across tissues were found to be highly correlated (Fig. 4e), with consistent enrichment in these correlated cell-type programs (Extended Data Fig. 3 and Supplementary Note).

Linking neuronal cells to MS and AD progression

We next turned to cases where both healthy and disease tissue have been profiled, allowing us to link disease GWASs to programs associated with disease-specific biology. Such understanding is especially important for identifying therapeutic targets associated with disease development rather than disease-onset mechanisms.

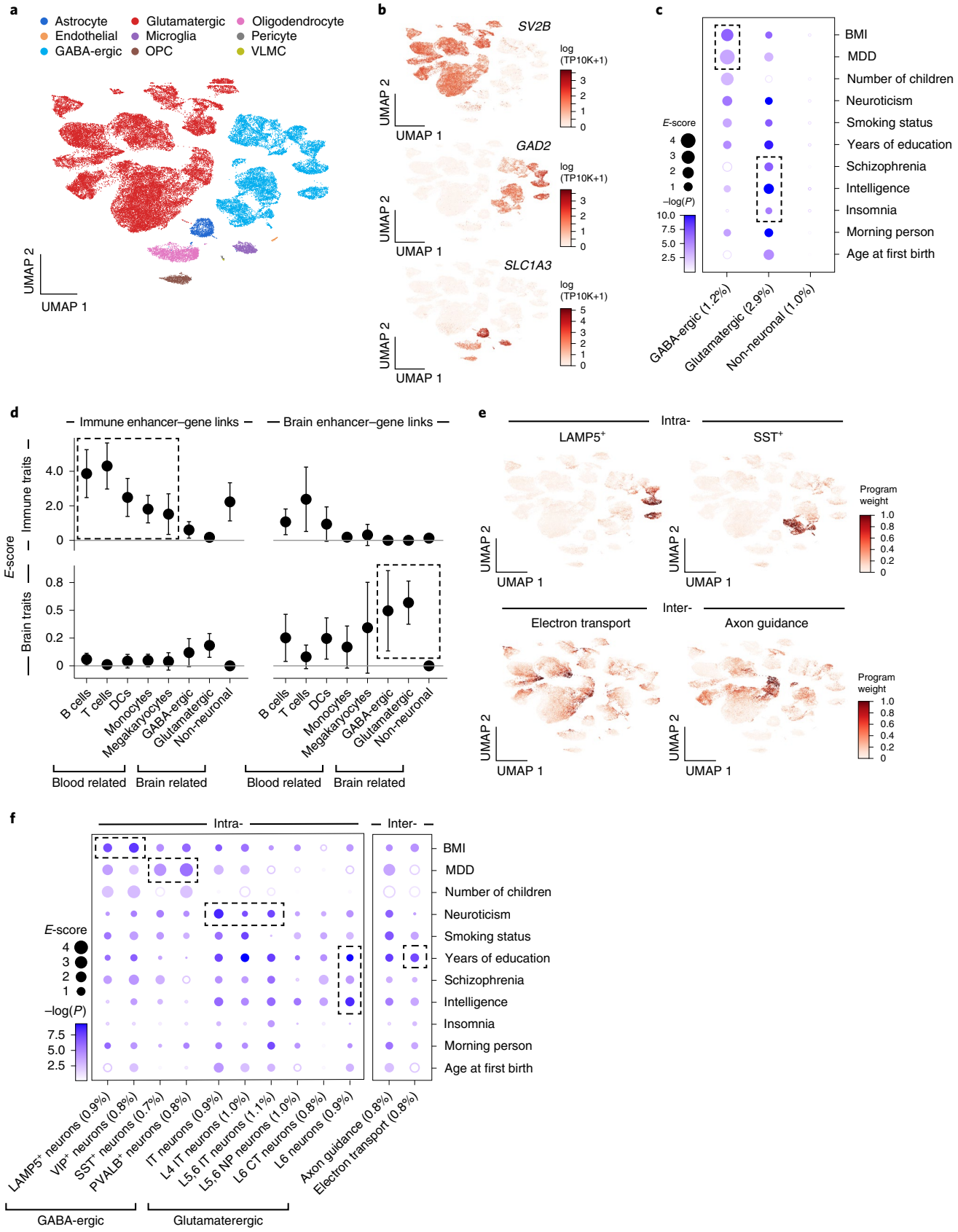
We first examined disease-dependent programs in MS and AD, where aberrant interactions between neurons and immune cells are thought to play an important role. We analyzed MS and AD GWAS data (Supplementary Table 2) along with cell-type, disease-dependent and cellular process programs from scRNA-seq of brains of healthy and MS[46] or AD[30] individuals (Fig. 5a,e and Supplementary Table 1). We considered brain enhancer–gene links, immune enhancer–gene links (because MS and AD are associated with both tissue types) and nontissue-specific enhancer–gene links (Extended Data Fig. 6) and detected the strongest enrichment results for the immune enhancer–gene links. In both MS and AD, disease-dependent programs in each cell type differed substantially from cell-type programs constructed from cells from healthy (average Pearson's $r = 0.16$) or disease (average Pearson's $r = 0.29$) samples alone (Extended Data Fig. 7). Furthermore, we confirmed that disease GWASs matched to the corresponding disease-dependent programs produced the strongest enrichments, although there was substantial cross-disease enrichment (Extended Data Figure 8).

In MS, there was enrichment in disease-dependent programs in GABA-ergic neurons and microglia (Fig. 5b and Extended Data Fig. 9), as well as in layer 2 and 3 glutamatergic neurons and the complement cascade (in multiple cell types; Fig. 5d). The specific enrichment of the GABA-ergic neuron, disease-dependent program (but not the healthy cell-type program) for MS is consistent with the observation that inflammation inhibits GABA transmission in MS[81]. The GABA-ergic disease-dependent program was enriched with hydrogen ion transmembrane transporter activity genes, whereas the GABA-ergic cell-type program was enriched in genes with general

Fig. 3 | Linking neuron cell subsets and cellular processes to brain-related diseases and traits. a,**b**, Major brain cell types. UMAP embedding of brain scRNA-seq profiles (dots) colored by cell-type annotations (**a**) or expression of cell-type-specific genes (**b**). **c**, Enrichments of brain cell-type programs for brain-related diseases and traits. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of brain cell-type programs (columns) are shown for brain-related diseases and traits (rows). **d**, Comparison of immune versus brain cell-type programs, enhancer–gene-linking strategies and diseases/traits. Magnitude (*E*-score and s.e.m.) of the heritability enrichment of immune versus brain cell-type programs (columns) is constructed using immune versus brain enhancer–gene-linking strategies (left and right panels) for immune-related ($n = 11$) versus brain-related ($n = 11$) diseases and traits (top and bottom panels). Data are presented as mean values ± s.e.m. **e**, Examples of inter- and intracell-type cellular processes. UMAP (as in **a**) is colored by each program weight (color bar) from NMF. **f**, Enrichments of brain cellular process programs for brain-related diseases and traits. Each of the cellular process programs is constructed using NMF to decompose the cells using a genes matrix into two matrices, cells by programs and programs by genes (NP = neural progenitor, CT = corticothalamic). Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of cellular process programs (columns) are shown for brain-related diseases and traits (rows). In **c** and **f**, the size of each corresponding SNP annotation (percentage SNPs) is reported in parentheses. Numerical results are reported in Supplementary Data 1 and 3. Further details of all diseases and traits analyzed are provided in Supplementary Table 2.

https://doi.org/10.1038/s41588-022-01187-9

neuronal functions (Supplementary Data 10). The enrichment of the microglia disease-dependent program for MS is consistent with the role of microglia in inflammation and demyelination in MS lesions[82,83] and highlights a contribution of microglia to both disease onset and response. The top driving genes for the microglia disease-dependent program enrichment included *MERTK* (ranked 2) and *TREM2* (ranked 4), both having roles in myelin destruction in MS patients[84,85]. Supporting this finding, there is a significant increase in the number of microglia



https://doi.org/10.1038/s41588-022-01187-9

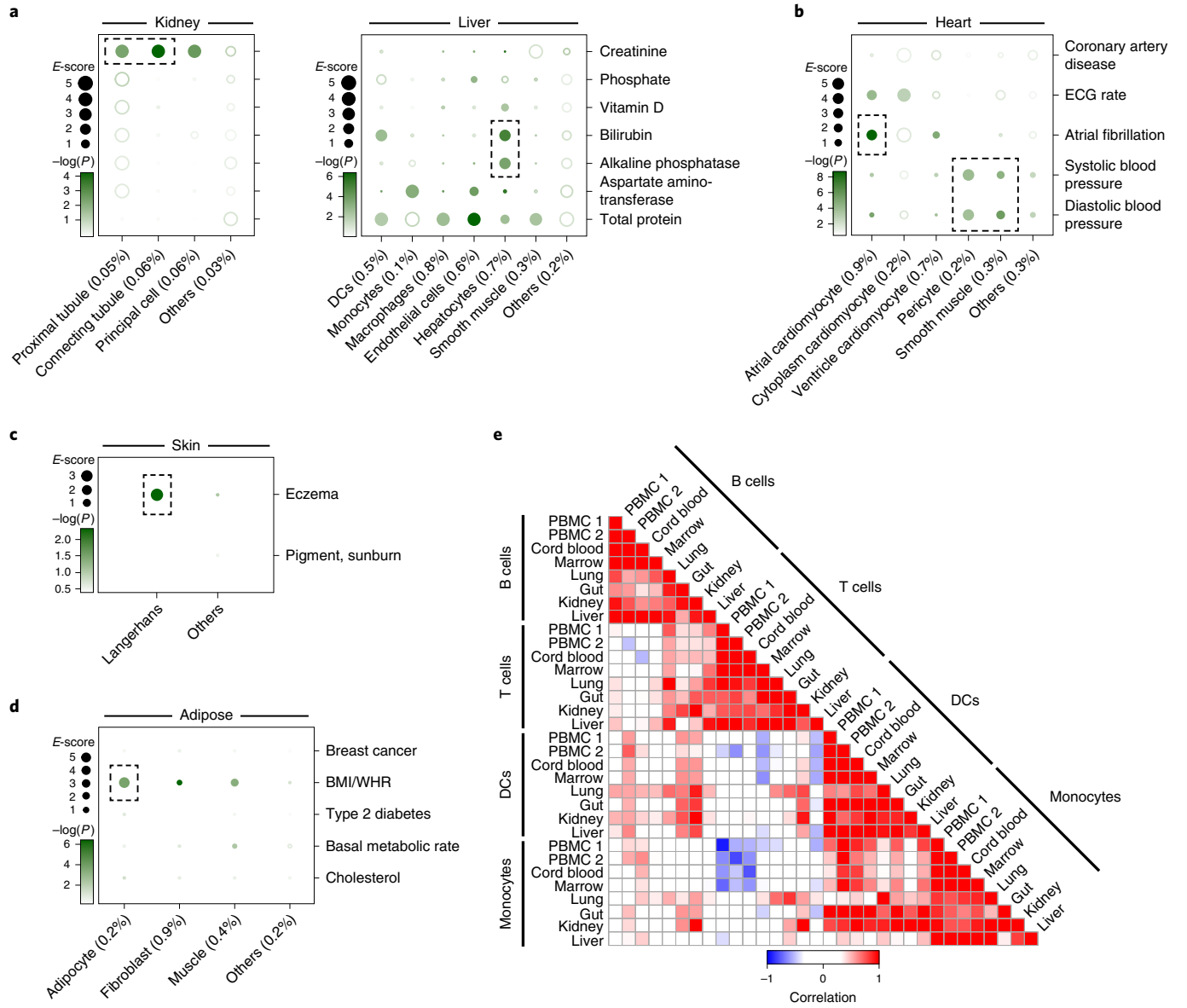


Fig. 4 | Linking cell types from diverse human tissues to disease. a–d, Enrichments of cell-type programs for corresponding diseases and traits. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of cell-type programs (columns) are shown for diseases and traits relevant to the corresponding tissue (rows) for kidney and liver (**a**), heart (**b**), skin (**c**) and adipose tissue (**d**). The size of each corresponding SNP annotation (percentage SNPs) is reported in parentheses. WHR, Waist:hip ratio. The numerical results are reported in Supplementary Data 1. Further details of all traits analyzed are provided in Supplementary Table 2. **e**, Correlation of immune cell-type programs across tissues. Pearson's correlation coefficients (color bar) of gene-level program memberships for immune cell-type programs are shown across different tissues (rows, columns), grouped by cell type (labels).

($P = 2 \times 10^{-4}$, Fisher's exact test) and a significant decrease in number of glutamatergic neurons ($P = 8 \times 10^{-5}$) in MS lesions (Fig. 5c and Supplementary Data 11).

In AD, all associations highlighted the central role of microglia, suggesting that different processes may be at play in microglia or microglia subsets in healthy brains and after disease initiation: only the microglia disease-dependent program was enriched out of eight disease-dependent programs tested (Fig. 5e,f and Extended Data Fig. 10), along with the healthy microglia program and the apelin signaling pathway, disease-specific cellular process program (intercell type: GABA-ergic neurons and microglia). The microglia program enrichments are consistent with the contribution of microglia-mediated inflammation to AD progression[86]. Supporting this finding, there is a significant increase in the number of microglia in AD, brain (Fig. 5g and Supplementary Data 11).

Thus, in both MS and AD, heritability was enriched in distinct ways in microglia cell-type, disease-dependent and cellular process programs, suggesting therapeutic opportunities to combat the role of microglia in varying contexts for disease risk.

Linking enterocytes and M cells to UC

We next examined the role of cell-type, disease-dependent and cellular process programs in UC, where failure to maintain the colon's epithelial barrier results in chronic inflammation. We analyzed UC and IBD GWAS data (Supplementary Table 2) with healthy cell-type, UC disease-dependent and UC cellular process programs constructed from scRNA-seq of healthy colon and from matched uninflamed and inflamed colon of UC patients (Fig. 6a and Supplementary Table 1). We compared colon enhancer–gene links (Fig. 6) and nontissue-specific enhancer–gene links (Extended Data Fig. 6) and detected the strongest

https://doi.org/10.1038/s41588-022-01187-9

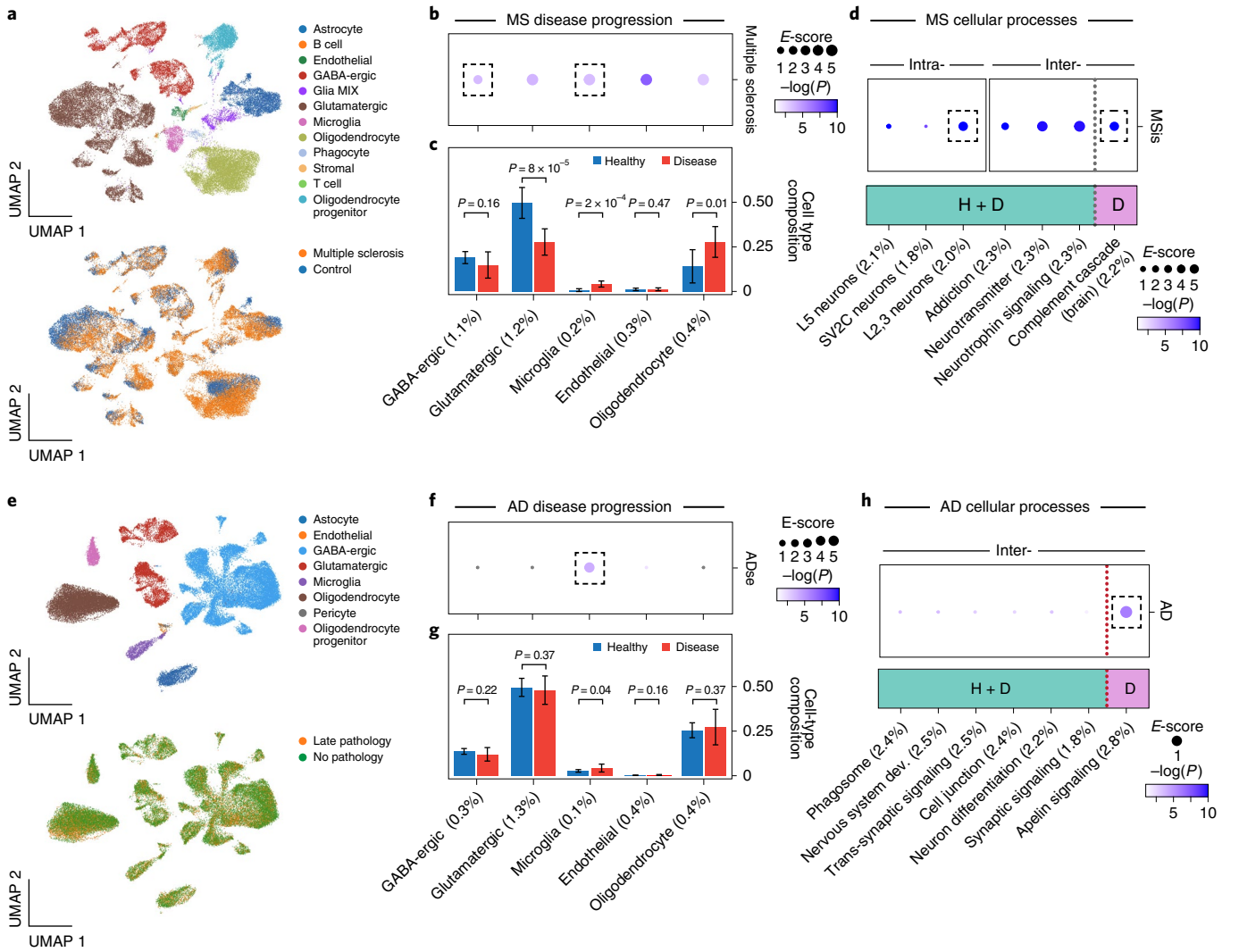


Fig. 5 | Linking MS and AD disease-dependent and cellular process programs to MS and AD. a, UMAP embedding of scRNA-seq profiles (dots) from MS and healthy brain tissue, colored by cell-type annotations (top) or disease status (bottom). **b**, Enrichments of MS disease-dependent programs for MS. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of MS disease-dependent programs (columns) are shown, based on the Roadmap–ABC–immune enhancer–gene-linking strategy. **c**, Proportion (mean and s.e.m.) of the corresponding cell types (columns) in healthy (blue) and MS (red) ($n = 21$ biologically independent brain samples). *P* value is by one-sided Fisher's exact test. **d**, Enrichments of MS cellular process programs for MS. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of intracell (left) or intercell (right) type cellular processes (healthy specific (H), MS specific (D) or shared (H + D)) (columns) are shown, based on the Roadmap–ABC–immune enhancer–gene-linking strategy. **e**, UMAP embedding of scRNA-seq profiles (dots) from AD and healthy brain tissue, colored by cell-type annotations (top) or disease status (bottom). **f**, Enrichments of AD disease-dependent programs for AD. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of AD disease-dependent programs (columns) are shown, based on the Roadmap–ABC–immune enhancer–gene-linking strategy. **g**, Proportion (mean and s.e.m.) of the corresponding cell types (columns) are shown in healthy (blue) and AD (red) samples ($n = 48$ biologically independent brain samples). *P* value is by one-sided Fisher's exact test. **h**, Enrichments of AD cellular process programs for AD. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of intercell-type cellular processes (AD specific (D) or shared (H + D)) (columns) are shown, based on the Roadmap–ABC–immune enhancer–gene-linking strategy. dev., development. In **b**, **c**, **d** and **f**–**h**, the size of each corresponding SNP annotation (percentage SNPs) is reported in parentheses. Numerical results are reported in Supplementary Data 2 and 3. Further details of all traits analyzed are provided in Supplementary Table 2.

enrichment results for the colon enhancer–gene links. As in MS and AD, UC disease-dependent programs in each cell type differed substantially from the corresponding healthy or disease colon cell-type programs (average Pearson's $r = 0.24$; Extended Data Fig. 7 and Supplementary Data 12).

In addition to previously observed enrichments in healthy immune cell-type programs, our analysis highlighted healthy cell-type programs of enteroendocrine and endothelial cells, disease-dependent programs of enterocytes and M cells, as well as the complement cascade (in plasma, B cells, enterocytes and fibroblasts), MHC-II antigen presentation (macrophages, monocytes and DCs) and epidermal growth factor receptor 1 (EGFR-1) signaling (macrophages and enterocytes) in both healthy and disease cells (Fig. 6, Extended Data Fig. 3 and Supplementary Data 1). The strong enrichment in endothelial cells, which comprise the gut vascular barrier, is consistent with their rapid changes in UC[87]; the top driving genes included members of the tumor necrosis factor-α signaling pathway (*EFNA1*, *NFKBIA* and *CD40*, ranked 18, 26 and 29, respectively), a key pathway in UC[88].

The disease-dependent programs (Fig. 6c, Table 1 and Extended Data Figs. 9 and 10) highlighted M cells, a rare cell type in healthy colon that increases in UC[34] (Fig. 6d and Supplementary Data 11). M cells surveil the lumen for pathogens and play a key role in immune–microbiome

https://doi.org/10.1038/s41588-022-01187-9

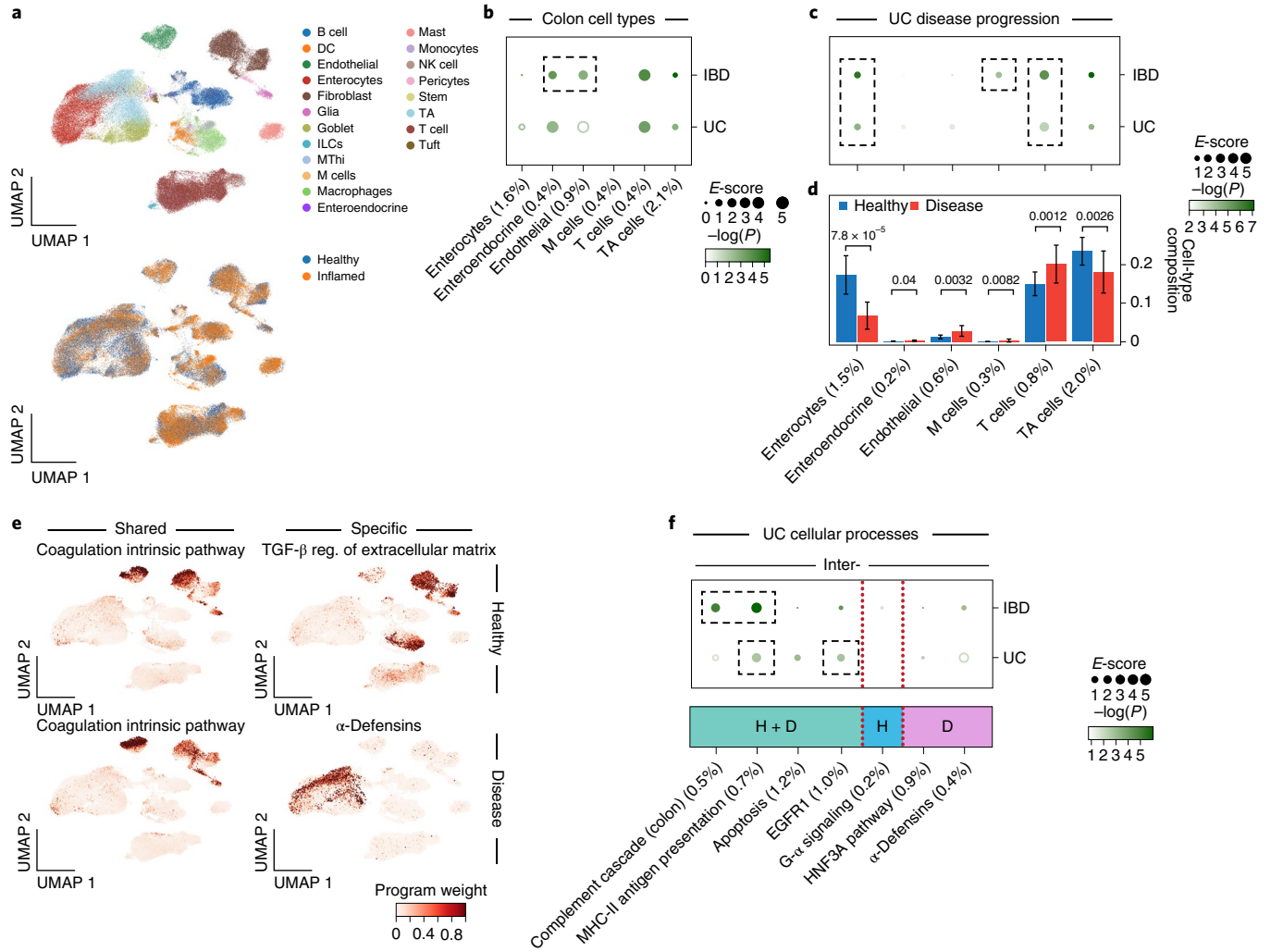


Fig. 6 | Linking UC disease-dependent and cellular process programs to UC and IBD. a, UMAP embedding of scRNA-seq profiles (dots) from UC and healthy colon tissue, colored by cell-type annotations (top) or disease status (bottom) (TA = Transit Amplifying, MThi = mitochondrial high, ILCs = immune-like cells). **b**, Enrichments of healthy colon cell types for disease. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of colon cell-type programs (columns) are shown for IBD or UC (rows). Results for additional cell types, including immune cell types in the colon, are reported in Extended Data Fig. 3 and Supplementary Data 1. **c**, Enrichments of UC disease-dependent programs for disease. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of UC disease-dependent programs (columns) are shown for IBD or UC (rows). **d**, Proportion (mean and s.e.m.) of the corresponding cell types (columns) in healthy (blue) and UC (red) samples is shown ($n = 36$ biologically independent colon samples). *P* value is by one-sided Fisher's exact test. **e**, Examples of shared (healthy and disease), healthy-specific and disease-specific cellular process programs. UMAP (as in **a**) is colored by each program weight (color bar) from NMF. TGF-β, transforming growth factor-β. **f**, Enrichments of UC cellular process programs for disease. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of intercell-type cellular processes (shared (H + D), healthy specific (H) or disease specific (D)) (columns) are shown for IBD or UC (rows). In **b**–**d** and **f**, the size of each corresponding SNP annotation (percentage SNPs) is reported in parentheses. Numerical results are reported in Supplementary Data 1–3. Further details of all traits analyzed are provided in Supplementary Table 2.

homeostasis[89]. Supporting this finding, mutations in *FERMT1*, a top driving gene in the M-cell disease-dependent program (ranked 3), cause Kindler's syndrome, a monogenic form of IBD with UC-like symptoms[90]. Notably, there was no enrichment in M-cell healthy cell-type programs (Fig. 6b), emphasizing that M cells are activated specifically in UC disease, as their proportions increase ($P = 0.008$; Fig. 6d).

Immune and connective tissue cell types linked to asthma

We analyzed GWAS data for asthma, idiopathic pulmonary fibrosis (IPF), COVID-19 (both general COVID-19 and severe COVID-19) and lung capacity (Supplementary Table 2) with healthy cell-type, disease-dependent and cellular process programs from asthma, IPF, COVID-19 and healthy[29] (lower lung lobes) tissue scRNA-seq (Fig. 7a,c,f, Supplementary Figs. 13d–f and 15 and Supplementary Data 12), using either lung enhancer or immune enhancer–gene links. For asthma, there was significant enrichment for healthy cell-type and disease-dependent programs in T cells (Supplementary Note). For lung capacity (height-adjusted forced expiratory volume in 1 s (FEV_{1adj}), relaxed vital capacity (RVC)), there was significant enrichment for healthy cell-type and disease-dependent programs in fibroblasts (Fig. 7b and Supplementary Data 1) and the MAPK cellular process program (in basal, club, fibroblast and endothelial cells) (Fig. 7f,g and Table 1). Genes driving these enrichments and enrichment results for IPF and COVID-19 are detailed in Supplementary Note.

Discussion

Previous work on identifying disease-critical tissues and cell types by combining expression profiles and human genetics signals has largely focused on the direct mapping of the expression of individual genes[34] and genome-wide polygenic signals[18,36] to discrete cell categories. Our

https://doi.org/10.1038/s41588-022-01187-9

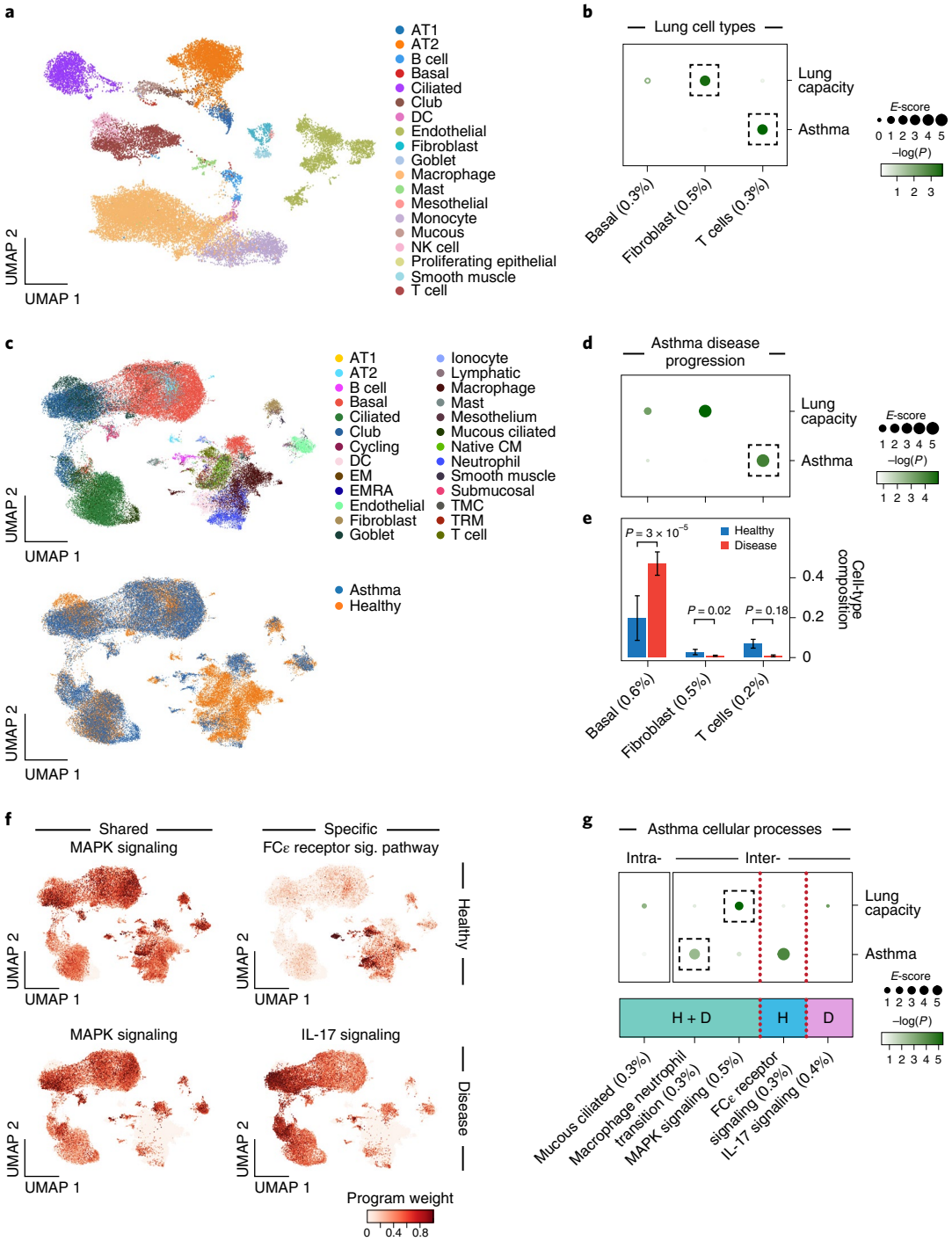


Fig. 7 | Linking asthma disease-dependent and cellular process programs to asthma and lung capacity. a, UMAP embedding of healthy lung scRNA-seq profiles (dots) colored by cell-type annotations. **b**, Enrichments of healthy lung cell types for disease. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of healthy lung cell-type programs (columns) are shown for lung capacity or asthma (rows). **c**, UMAP embedding of scRNA-seq profiles (dots) from asthma and healthy lung tissue, colored by cell-type annotations (top) or disease status (bottom) (AT1 = Alveolar Type 1, AT2 = Alveolar Type 2, EM = effector memory T cell, EMRA = effector memory re-expressing CD45RA T cell, TMC = tissue migratory $CD4^+$ T cells, CM = central memory T cells, TRM = tissue resident memory T cell). **d**, Enrichments of asthma disease-dependent programs for disease. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of asthma disease-dependent programs (columns) are shown for lung capacity or asthma (rows). **e**, Proportion (mean and s.e.m.) of the corresponding cell types (columns), in healthy (blue) and asthma (red) samples ($n = 54$ biologically independent lung samples). *P* value is by one-sided Fisher's exact test. **f**, Examples of shared (healthy and disease), healthy-specific and disease-specific cellular process programs. sig., signaling. UMAP (as in **c**) is colored by each program weight (color bar) from NMF. **g**, Enrichments of asthma cellular process programs for disease. Magnitude (*E*-score, dot size) and significance ($-\log_{10}(P)$, dot color) of the heritability enrichment of intracell (left) and intercell (right)-type cellular processes (shared (H + D), healthy specific (H) or disease specific (D)) (columns) are shown for lung capacity and asthma GWAS summary statistics (rows). In **b**, **d**, **e** and **g**, the size of each corresponding SNP annotation (percentage SNPs) is reported in parentheses. Numerical results are reported in Supplementary Data 1–3. Further details of all traits analyzed are provided in Supplementary Table 2.

study demonstrates that there is much to be gained by linking inferred representations of the underlying biological processes beyond cell types in different cell and tissue contexts with genome-wide polygenic disease signals, by integrating scRNA-seq, epigenomic and GWAS datasets.

Our work introduces three main conceptual advances: first, by integrating scRNA-seq data and GWAS summary statistics using tissue-specific enhancer–gene-linking strategies, we detect subtle differences in SNP-to-gene mapping between tissues which, on aggregation over the full GWAS signal, produce strong differences in disease heritability across cell types. Second, by constructing disease-dependent programs comparing cells of the same type in disease versus healthy tissue, we project GWAS signals across disease-specific cell states. Third, by using non-negative matrix factorization (NMF) to construct cellular process programs that do not rely on known cell-type categories, we identify cellular mechanisms that vary across a continuum of cells of one type or are shared between cells of different types, such as the mitogen-activated protein kinase (MAPK) signaling pathway identified in the lung.

Leveraging these advances, we identified notable enrichments (Table 1) that have not previously been identified using GWAS data and are biologically plausible but not clearly expected, thus supporting the potential of the sc-linker to identify new knowledge. We also observed patterns across datasets that offer additional insights. For example, we observed that disease-dependent programs, but not healthy cell-type programs, of epithelial cells (M cells and basal cells) tend to be enriched in autoimmune diseases (UC and asthma). In contrast, for immune cells, healthy and disease-dependent programs tended to be similarly enriched. We posit that this suggests a role for epithelial cells in development, rather than initiation, of disease. Future studies are required to experimentally validate these hypotheses.

Our work has several limitations that highlight directions for future research. First, the cell types and states covered in this work are not exhaustive, and there will continue to be other cell types and more granular cell states uncovered as the scale of sequencing continues to grow. Second, the enhancer–gene-linking strategies can continue to be improved beyond the Roadmap and ABC models incorporated here. Finally, we focus on genome-wide disease heritability (rather than a particular locus); however, our approach can be used to implicate specific genes and gene programs. Additional limitations are discussed in Supplementary Note.

Looking forward, the gene program–disease links identified by our analyses can be used to guide downstream studies, including designing systematic perturbation experiments[91,92] in cell and animal models for functional follow-up. In the long term, with the increasing success of phenome-wide association studies and the integration of multimodal single-cell resolution epigenomics, this framework will continue to be useful in identifying biological mechanisms driving a broad range of diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01187-9.

References

1. Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
2. Visscher, P. M. et al. 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
3. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
4. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
5. Price, A. L., Spencer, C. C. A. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B Biol. Sci.* **282**, 20151684 (2015).
6. Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic medicine—progress, pitfalls, and promise. *Cell* **177**, 45–57 (2019).
7. Zeggini, E., Gloyn, A. L., Barton, A. C. & Wain, L. V. Translational genomics and precision medicine: moving from the lab to the clinic. *Science* **365**, 1409–1413 (2019).
8. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* **21**, 137–150 (2020).
9. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
10. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
11. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
12. Zhou, J. et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).
13. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9**, 4361 (2018).
14. Wang, Q. et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat. Neurosci.* **22**, 691–699 (2019).
15. Fang, H. et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082–1091 (2019).
16. Calderon, D. et al. Inferring relevant cell types for complex traits by using single-cell gene expression. *Am. J. Hum. Genet.* **101**, 686–691 (2017).
17. Ongen, H. et al. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676–1683 (2017).
18. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
19. Ernst, J. et al. Systematic analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
20. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
21. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol.* **18**, 193 (2017).
22. Fulco, C. P. et al. Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669 (2019).
23. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
24. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
25. Tucker, N. et al. Transcriptional and cellular diversity of the human heart. *Circulation* **142**, 466–482 (2020).
26. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
27. Kowalczyk, M. S. Census of immune cells. *Human Cell Atlas Data Portal* https://data.humancellatlas.org/explore/projects/cc95ff89-2e68-4a08-a234-480eca21ce79 (2018).

28. Sunkin, S. M. et al. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, D996 (2013).
29. Habermann, A. C. et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* **6**, eaba1972 (2020).
30. Mathys, H. et al. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).
31. Jerby-Arnon, L. et al. A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell* **175**, 984–997.e24 (2018).
32. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
33. Peng, Y.-R. et al. Molecular classification and comparative taxonomics of foveal and peripheral cells in primate retina. *Cell* **176**, 1222–1237.e22 (2019).
34. Smillie, C. S. et al. Intra- and Inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730.e22 (2019).
35. Watanabe, K., Umićević Mirkov, M., de Leeuw, C. A., van den Heuvel, M. P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat. Commun.* **10**, 3222 (2019).
36. Bryois, J. et al. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nat. Genet.* **52**, 482–493 (2020).
37. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
38. Drokhlyansky, E. et al. The human and mouse enteric nervous system at single-cell resolution. *Cell* **182**, 1606–1622.e23 (2020).
39. Leeuw, C. A., de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
40. Gazal, S. et al. Linkage disequilibrium dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
41. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).
42. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
43. Stewart, B. J. et al. Spatio-temporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).
44. Muus, C. et al. Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nat. Med.* **27**, 546–559 (2021).
45. Cheng, J. B. et al. Transcriptional programming of normal and inflamed human epidermis at single-cell resolution. *Cell Rep.* **25**, 871–883 (2018).
46. Schirmer, L. et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).
47. Braga, F. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
48. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
49. Ulirsch, J. C. et al. Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).
50. Chen, M.-H. et al. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. *Cell* **182**, 1198–1213.e14 (2020).
51. Biedermann, T., Skabytska, Y., Kaesler, S. & Volz, T. Regulation of T cell immunity in atopic dermatitis by microbes: the yin and yang of cutaneous inflammation. *Front. Immunol.* **6**, 353 (2015).
52. Hennino, A. et al. Skin-infiltrating CD8+ T cells initiate atopic dermatitis lesions. *J. Immunol.* **178**, 5571–5577 (2007).
53. Thériault, P., ElAli, A. & Rivest, S. The dynamics of monocytes and microglia in Alzheimer's disease. *Alzheimers Res. Ther.* **7**, 41 (2015).
54. Nuyts, A. H., Lee, W. P., Bashir-Dar, R., Berneman, Z. N. & Cools, N. Dendritic cells in multiple sclerosis: key players in the immunopathogenesis, key players for new cellular immunotherapies? *Mult. Scler.* **19**, 995–1002 (2013).
55. Haschka, D. et al. Expansion of neutrophils and classical and nonclassical monocytes as a hallmark in relapsing–remitting multiple sclerosis. *Front. Immunol.* **11**, 594 (2020).
56. Momeni, A. et al. Fingolimod and changes in hematocrit, hemoglobin and red blood cells of patients with multiple sclerosis. *Am. J. Clin. Exp. Immunol.* **8**, 27–31 (2019).
57. Yeung, M. et al. Characterisation of mucosal lymphoid aggregates in ulcerative colitis: immune cell phenotype and TcR-γδ expression. *Gut* **47**, 215–227 (2000).
58. Mouly, E. et al. The Ets-1 transcription factor controls the development and function of natural regulatory T cells. *J. Exp. Med.* **207**, 2113–2125 (2010).
59. Mayassi, T. et al. Chronic inflammation permanently reshapes tissue-resident immunity in celiac disease. *Cell* **176**, 967–981.e19 (2019).
60. Pandey, A. et al. Cloning of a receptor subunit required for signaling by thymic stromal lymphopoietin. *Nat. Immunol.* **1**, 59–64 (2000).
61. Gao, P.-S. et al. Genetic variants in TSLP are associated with atopic dermatitis and eczema herpeticum. *J. Allergy Clin. Immunol.* **125**, 1403–1407.e4 (2010).
62. Altin, J. A. et al. Ndfip1 mediates peripheral tolerance to self and exogenous antigen by inducing cell cycle exit in responding CD4+ T cells. *Proc. Natl Acad. Sci. USA* **111**, 2067–2074 (2014).
63. Yip, K. H. et al. The Nedd4-2/Ndfip1 axis is a negative regulator of IgE-mediated mast cell activation. *Nat. Commun.* **7**, 13198 (2016).
64. Villegas-Llerena, C., Phillips, A., Garcia-Reitboeck, P., Hardy, J. & Pocock, J. M. Microglial genes regulating neuroinflammation in the progression of Alzheimer's disease. *Curr. Opin. Neurobiol.* **36**, 74–81 (2016).
65. Efthymiou, A. G. & Goate, A. M. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol. Neurodegener.* **12**, 43 (2017).
66. Luscher, B., Shen, Q. & Sahir, N. The GABAergic deficit hypothesis of major depressive disorder. *Mol. Psychiatry* **16**, 383–406 (2011).
67. Mossakowska-Wójcik, J., A, O., M, T., J, S. & P, G. The importance of *TCF4* gene in the etiology of recurrent depressive disorders. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **80**, 304–308 (2018).
68. Li, L. et al. Disruption of TCF4 regulatory networks leads to abnormal cortical development and mental disabilities. *Mol. Psychiatry* **24**, 1235–1246 (2019).
69. Mbarek, H. et al. Genome-wide significance for *PCLO* as a gene for major depressive disorder. *Twin Res. Hum. Genet.* **20**, 267–270 (2017).
70. Ciarimboli, G. et al. Proximal tubular secretion of creatinine by organic cation transporter OCT2 in cancer patients. *Clin. Cancer Res.* **18**, 1101–1108 (2012).
71. Zhang, X. et al. Tubular secretion of creatinine and kidney function: an observational study. *BMC Nephrol.* **21**, 108 (2020).
72. Cui, C., J, K., I, L., U, B. & D, K. Hepatic uptake of bilirubin and its conjugates by the human organic anion transporter SLC21A6. *J. Biol. Chem.* **276**, 9626–9630 (2001).
73. Wang, X., Chowdhury, J. R. & Chowdhury, N. R. Bilirubin metabolism: applied physiology. *Curr. Paediatr.* **16**, 70–74 (2006).
74. Barth, A. S. & Tomaselli, G. F. Cardiac metabolism and arrhythmias. *Circ. Arrhythm. Electrophysiol.* **2**, 327–335 (2009).

75. Yamazaki, T. & Mukouyama, Y. Tissue specific origin, development, and pathological perspectives of pericytes. *Front. Cardiovasc. Med.* **5**, 78 (2018).
76. Deckers, J., Hammad, H. & Hoste, E. Langerhans cells: sensing the environment in health and disease. *Front. Immunol.* **9**, 93 (2018).
77. Hsieh, K. H., Chou, C. C. & Huang, S. F. Interleukin 2 therapy in severe atopic dermatitis. *J. Clin. Immunol.* **11**, 22–28 (1991).
78. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).
79. Attie, A. D. & Scherer, P. E. Adipocyte metabolism and obesity. *J. Lipid Res.* **50**, S395–S399 (2009).
80. Xia, B. Adipose tissue deficiency of hormone-sensitive lipase causes fatty liver in mice. *PLoS Genet.* **13**, e1007110 (2017).
81. Rossi, S. et al. Inflammation inhibits GABA transmission in multiple sclerosis. *Mult. Scler.* **18**, 1633–1635 (2012).
82. Cannella, B. et al. The neuregulin, glial growth factor 2, diminishes autoimmune demyelination and enhances remyelination in a chronic relapsing model for multiple sclerosis. *Proc. Natl Acad. Sci. USA* **95**, 10100–10105 (1998).
83. Horstmann, L. et al. Inflammatory demyelination induces glia alterations and ganglion cell loss in the retina of an experimental autoimmune encephalomyelitis model. *J. Neuroinflammation* **10**, 120 (2013).
84. Healy, L. M. et al. MerTK-mediated regulation of myelin phagocytosis by macrophages generated from patients with MS. *Neurol. Neuroimmunol. Neuroinflamm.* **4**, e402 (2017).
85. Cignarella, F. et al. TREM2 activation on microglia promotes myelin debris clearance and remyelination in a model of multiple sclerosis. *Acta Neuropathol.* **140**, 513–534 (2020).
86. Hemonnot, A.-L., Hua, J., Ulmann, L. & Hirbec, H. Microglia in Alzheimer disease: well-known targets and new opportunities. *Front. Aging Neurosci.* **11**, 233 (2019).
87. Cromer, W. E., Mathis, J. M., Granger, D. N., Chaitanya, G. V. & Alexander, J. S. Role of the endothelium in inflammatory bowel diseases. *World J. Gastroenterol.* **17**, 578–593 (2011).
88. Ruder, B., Atreya, R. & Becker, C. Tumour necrosis factor alpha in intestinal homeostasis and gut related diseases. *Int. J. Mol. Sci.* **20**, 1887 (2019).
89. Graham, D. B. & Xavier, R. J. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* **578**, 527–539 (2020).
90. Bianco, A. M., Girardelli, M. & Tommasini, A. Genetics of inflammatory bowel disease from multifactorial to monogenic forms. *World J. Gastroenterol.* **21**, 12296–12310 (2015).
91. Dixit, A. et al. Perturb-seq: dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e17 (2016).
92. Jin, X. et al. In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, eaaz6063 (2020).

Methods

Ethical approval

This research complies with all relevant ethical regulations, and the research protocols are approved by the Harvard School of Public Health.

ScRNA-seq data pre-processing

All scRNA-seq datasets in the present study[25–30,34,42–48] are publicly available cell-by-gene expression matrices that are aligned to the hg38 human transcriptome (Supplementary Table 1). Each dataset included metadata information for each cell, describing the total number of reads in the cell and which sample the cell corresponds to and, if applicable, its disease status. We transformed each expression matrix to a count matrix by reversing any log(normalization) processing (because each downloaded dataset contained (1) raw counts, (2) normalized $\log_2$(TP10K) or (3) normalized $\log_{10}$(TP10K), where TP10K is transcripts per 10,000 transcripts) and standardized the normalization approach across all datasets to account for differences in sequencing depth across cells by normalizing the total number of unique molecular identifiers (UMIs) per cell, converting to TP10K and taking the log of the result to obtain log(10,000 × UMIs/total UMIs + 1), with '$\log_2$(TP10K + 1)' as the final expression unit.

Dimensionality reduction, batch correction, clustering and annotation of scRNA-seq

The $\log_2$(TP10K + 1) expression matrix for each dataset was used for the following downstream analyses. For each dataset, we identified the top 2,000 highly variable genes across the entire dataset using Scanpy's[93] v.1.7.1 highly_variable_genes function with the sample ID as input for the batch. We then performed a principal component analysis (PCA) with the top 2,000 highly variable genes and identified the top 40 principal components (PCs), beyond which negligible additional variance was explained in the data (the analysis was performed with 30, 40 and 50 PCs and was robust to this choice). We used Harmony[94] v.0.1.1 for batch correction, where each sample was considered its own batch. Subsequently, we built a *k*-nearest neighbors graph of cell profiles ($k = 10$) based on the top 40 batch-corrected components computed by Harmony and performed community detection on this neighborhood graph using the Leiden graph clustering method[95] with resolution 1. For each dataset, individual single-cell profiles were visualized using the Uniform Manifold Approximation and Projection (UMAP)[96]. If previous annotations were available, they are used as a reference to annotate each cell in each dataset. If previous annotations were not available, we used established cell-type-specific expression signatures and gene markers described in the data source to annotate cells at the resolution of Leiden clusters.

Cell-type gene programs

We constructed cell-type programs for every cell type in a given tissue by applying a nonparametric Wilcoxon's rank-sum test for differential expression (DE) between each cell type versus other cell types and computed a *P* value for each gene. Using a previously published strategy[15], we transform these *P* values to $X = -2\log(P)$, which follows a χ^2_2 distribution; these transformed values are converted to a grade between 0 and 1 using the minimum/maximum (min/max) normalization $g = (X - \min(X))/(\max(X) - \min(X))$, resulting in a relative weighting of genes in each program. We note that these scores do not formally represent probabilities. In brief, cell-type programs constructed from healthy cells were termed healthy cell-type programs, and similarly cell-type programs constructed from disease cells were termed disease cell-type programs.

Disease-dependent gene programs

We constructed disease-dependent programs for each cell type observed in both healthy and matching disease tissue. For each cell type, we computed a gene-level, nonparametric, Wilcoxon's rank-sum DE test between cells from healthy and those from disease tissues of the same cell type. The *P* values for each gene were transformed to a grade between 0 and 1 using the same strategy as in the cell-type program to form a relative weighting of genes in each program. In the COVID-19 BAL scRNA-seq, we also constructed viral progression programs based on DE between virally infected and uninfected cells of the same cell type in individuals with COVID-19. We observed low correlation between healthy cell-type gene programs and disease-dependent gene programs (Supplementary Fig. 13 and Supplementary Data 12).

Cellular process gene programs

Using latent factors derived from NMF[97] (see below), we defined a cellular process program based on genes with high correlation (across cells) between their expression in each cell and the contribution of the factor to each cell (collapsing latent factors with high correlation). The correlations were transformed to a continuous-valued scale (between 0 and 1) by scaling their values (negative correlations are assigned to 0). We then annotated each factor (program) by the pathway most enriched in the top driving genes for the factor and labeled each as an 'intracell type' or 'intercell type' latent factor if the pathway was highly correlated with only one or multiple cell-type programs, respectively.

We constructed cellular process programs using an unsupervised approach, by applying NMF[97] to the scRNA-seq cells-by-genes matrix. The solution to this formulation can be identified by solving the following minimization problem:

$$\operatorname{argmin}\left\{\frac{1}{2}\left\|X_{n,m} - \sum_p W_{\{n,p\}} \times H_{p,m}\right\|_F^2 + (1-\alpha)\frac{1}{2}\left\|W_{n,p}\right\| + \frac{1}{2}(1-\alpha)\left\|H_{p,m}\right\| + \alpha\left\|\operatorname{vec}(W_{n,p})\right\|_1 + \alpha\left\|\operatorname{vec}(H_{p,m})\right\|_1\right\} \quad (1)$$

where $X_{n,m}$ represents the log(normalized) expression of gene *m* in sample *n*, $W_{n,p}$ denotes the grade of membership of latent factor *p* in sample *n* and $H_{p,m}$ represents the factor weight of factor *p* in gene *m*. NMF identifies cellular processes as latent factors with a grade of contribution to each cell. For each dataset, we specified the number of latent factors *p* to be the number of annotated cell types in the dataset + 10. For each latent factor, we define a cellular process gene program by identifying genes with high correlation (across cells) between expression in a cell and the contribution of each factor to each cell. Latent factors with correlation >0.8 are collapsed to only consider a single latent factor. We annotated each cellular process program by the pathway most enriched (calculated with the Enrichr database and Fisher's exact test *P* value) in the genes with highest correlation (across cells) between expression levels and factor weights (*H*) underlying the cellular process program (not necessarily the most highly expressed genes; Supplementary Fig. 17) and labeled it as an 'intracell-type' or 'intercell-type' cellular process program if highly correlated with only one or multiple cell-type programs, respectively.

Cellular process gene programs constructed from healthy and disease tissues

For scRNA-seq from healthy and disease tissue contexts, we proposed a modified NMF approach to construct gene programs that are shared across both tissues, specific to either healthy tissue or disease tissue. Let $H_{P\times N_1}$ be the observed gene expression data for a tissue T from a healthy individual and $D_{P\times N_2}$ be the observed gene expression data for the corresponding tissue from a disease individual. *P* is the number of features (genes), and N_1 and N_2 denote the number of samples from the healthy and disease tissues, respectively.

We assume an NMF for H and D as follows:

$$H_{P\times N_1} \approx \left[L^{\mathrm{CH}}_{P\times K_{\mathrm{C}}} L^{\mathrm{UH}}_{P\times K_{\mathrm{H}}}\right] F^{\mathrm{H}}_{(K_{\mathrm{C}}+K_{\mathrm{H}})\times N_1} \text{ where } L^{\mathrm{CH}}, L^{\mathrm{UH}}, F^{\mathrm{H}} > 0 \quad (2)$$

$$D_{P\times N_2} \approx \left[L^{\mathrm{CD}}_{P\times K_{\mathrm{C}}} L^{\mathrm{UD}}_{P\times K_{\mathrm{D}}}\right] F^{\mathrm{D}}_{(K_{\mathrm{C}}+K_{\mathrm{D}})\times N_2} \text{ where } L^{\mathrm{CD}}, L^{\mathrm{UD}}, F^{\mathrm{D}} > 0 \quad (3)$$

where K_{C} is the number of shared programs between the healthy and the disease samples, K_{H} is the number of healthy specific programs and K_{D} is the number of disease-specific programs. L^{CH} and L^{CD} are used to denote the shared programs between healthy and disease states. Therefore, we assume that L^{CH} is very close to L^{CD} but not exact to account for other factors such as experimental conditions perturbing the estimates slightly. On the other hand, L^{UH} and L^{UD} are used to denote the healthy specific and disease-specific programs, respectively. F^{H} and F^{D} denote the program weights in the healthy and disease samples, respectively. This framed in the form of the following optimization problem:

$$\begin{aligned}\underset{L^{\mathrm{H}},L^{\mathrm{D}},F^{\mathrm{H}},F^{\mathrm{D}}}{\operatorname{argmin}} \; & \tfrac{1}{2}\left\|H - L^{\mathrm{H}}F^{\mathrm{H}}\right\|_F^2 + \tfrac{1}{2}\left\|D - L^{\mathrm{D}}F^{\mathrm{D}}\right\|_F^2 \\ & + \tfrac{\mu}{2}\left(\left\|L^{\mathrm{H}}\right\|_F^2 + \left\|L^{\mathrm{D}}\right\|_F^2\right) + \tfrac{\gamma}{2}\left(\left\|L^{\mathrm{CH}}\right\|_F^2 - \left\|L^{\mathrm{CD}}\right\|_F^2\right)\end{aligned} \quad (4)$$

where $L^{\mathrm{H}} = \left[L^{\mathrm{CH}}_{PXK_{\mathrm{C}}} L^{\mathrm{UH}}_{PXK_{\mathrm{H}}}\right]$ and $L^{\mathrm{D}} = \left[L^{\mathrm{CD}}_{PXK_{\mathrm{C}}} L^{\mathrm{UD}}_{PXK_{\mathrm{D}}}\right]$ and γ is a tuning parameter that controls how close L^{CH} is to L^{CD} and μ represents a tuning parameter that controls for the size of the loadings and the factors.

To determine the multiplicative updates of the NMF optimization problem in equation (4), we compute the derivatives of the optimization criterion with respect to each parameter of interest. We call the optimization criterion Q:

$$\nabla Q(L^{\mathrm{H}}) = -HF^{\mathrm{H}^T} + L^{\mathrm{H}}F^{\mathrm{H}}F^{\mathrm{H}^T} + \mu L^{\mathrm{H}} - \gamma[L^{\mathrm{CD}}0] \quad (5)$$

$$\nabla Q(L^{\mathrm{D}}) = -DF^{\mathrm{D}^T} + L^{\mathrm{D}}F^{\mathrm{D}}F^{\mathrm{D}^T} + \mu L^{\mathrm{D}} - \gamma[L^{\mathrm{CH}}0] \quad (6)$$

$$\nabla Q(F^{\mathrm{H}}) = -L^{\mathrm{H}^T}H + L^{\mathrm{H}^T}L^{\mathrm{H}}F^{\mathrm{H}} \quad (7)$$

$$\nabla Q(F^{\mathrm{D}}) = -L^{\mathrm{D}^T}D + L^{\mathrm{D}^T}L^{\mathrm{D}}F^{\mathrm{D}} \quad (8)$$

Following the multiplicative update rules of NMF as per Lee and Seung[97], we get the following iterative updates and assume convergence has been achieved after 100 iterations or when the reconstruction error is below a user-specified error threshold (here the threshold is taken to be 1×10^{-4}):

$$L^{\mathrm{H}}_{ij} \leftarrow L^{\mathrm{H}}_{ij}\frac{\left(HF^{\mathrm{H}^T} + \gamma\left[L^{\mathrm{CD}}0\right]\right)_{ij}}{\left(L^{\mathrm{H}}F^{\mathrm{H}}F^{\mathrm{H}^T} + \mu L^{\mathrm{H}}\right)_{ij}} \quad (9)$$

$$L^{\mathrm{D}}_{ij} \leftarrow L^{\mathrm{D}}_{ij}\frac{\left(DF^{\mathrm{D}^T} + \gamma[L^{\mathrm{CH}}0]\right)_{ij}}{\left(L^{\mathrm{D}}F^{\mathrm{D}}F^{\mathrm{D}^T} + \mu L^{\mathrm{D}}\right)_{ij}} \quad (10)$$

$$F^{\mathrm{H}}_{ij} \leftarrow F^{\mathrm{H}}_{ij}\frac{\left(L^{\mathrm{H}^T}H\right)_{ij}}{\left(L^{\mathrm{H}^T}L^{\mathrm{H}}F^{\mathrm{H}}\right)_{ij}} \quad (11)$$

$$F^{\mathrm{D}}_{ij} \leftarrow F^{\mathrm{D}}_{ij}\frac{\left(L^{\mathrm{D}^T}D\right)_{ij}}{\left(L^{\mathrm{D}^T}L^{\mathrm{D}}F^{\mathrm{D}}\right)_{ij}} \quad (12)$$

Enhancer–gene-linking strategies

We define an enhancer–gene-linking strategy as an assignment of 0, 1 or more genes to each SNP with a minor allele count >5 in the 1000 Genomes Project European reference panel[98]. In the present study, we primarily considered an enhancer–gene-linking strategy defined by the union of the Roadmap[21,99] and ABC[22,100] strategies. Roadmap and ABC enhancer–gene links are publicly available for a broad set of tissues and have been shown to outperform other enhancer–gene-linking strategies in previous work[101]. We consider tissue-specific Roadmap and ABC enhancer–gene-linking strategies for gene programs corresponding to any of the biosamples (cell types or tissues) associated with the relevant tissue. Based on analysis in immune cell types, 87% of genes expressed in the scRNA-seq were observed to have enhancer–gene links. We also consider nontissue-specific Roadmap and ABC strategies (Supplementary Fig. 12). Besides this enhancer–gene-linking strategy, we also considered a standard 100-kb window-based strategy[13,18].

Genomic annotations and the baseline-LD models

We define an annotation as an assignment of a numeric value to each SNP in a predefined reference panel (for example, 1000 Genomes Project[98]; see Data availability). Binary annotations can have a value of 0 or 1 only, continuous-valued annotations can have any real value and our focus is on continuous-valued annotations with values between 0 and 1. Annotations that correspond to known or predicted functions are referred to as functional annotations. The baseline-LD model[40,41] (v.2.1) contains 86 functional annotations (Data availability), including binary coding, conserved and regulatory annotations (for example, promoter, enhancer, histone marks, transcription factor-binding site) and continuous-valued LD-related annotations.

S-LDSC

S-LDSC assesses the contribution of a genomic annotation to disease and complex trait heritability[11]. It assumes that the per-SNP heritability or variance of effect size (of standardized genotype on trait) of each SNP is equal to the linear contribution of each annotation.

$$\mathrm{var}\left(ß_j\right) = \sum_c^C a_{jc}t_c \quad (14)$$

where a_{jc} is the value of annotation c at SNP j, with the annotation either continuous or binary (0/1), and t_c is the contribution of annotation c to per-SNP heritability conditional on the other annotations. S-LDSC estimates t_c for each annotation using the following equation:

$$E\left(X_j^2\right) = N\sum_c l(j,c)\,t_c + 1 \quad (15)$$

where $l(j,c) = \sum_k a_{ck}r_{jk}^2$ is the stratified LD score of SNP j with respect to annotation c, r_{jk} is the genotypic correlation between SNPs j and k computed using 1000 Genomes Project, and N is the GWAS sample size.

We assess the informativeness of an annotation c using two metrics. The first metric is the enrichment score (E-score), which relies on the enrichment of annotation c (E_c), defined for binary annotations as follows (for binary and continuous-valued annotations only):

$$E_c = \frac{\frac{h_g^2(c)}{h_g^2}}{\frac{\sum_j a_{jc}}{M}} \quad (16)$$

where $h_g^2(c)$ is the heritability explained by the SNPs in annotation c, weighted by the annotation values where M is the total number of SNPs on which this heritability is computed (5,961,159 in our analyses). The E-score is defined as the difference between the enrichment for annotation c corresponding to a particular program against an SNP annotation for all protein-coding genes with a predicted enhancer–gene link in the relevant tissue. The E-score metric generalizes to continuous-valued annotations with values between 0 and 1 (ref. [102]). We primarily focus

on the P value for nonzero E-score >2. We chose the threshold of 2 because it is a round number that is roughly the geometric mean of the value of 1 (no enrichment) and the median value of 3.7 among the notable enrichments highlighted in Table 1.

The second metric is standardized effect size (τ^*), the proportionate change in per-SNP heritability associated with a 1 s.d. in the value of the annotation, conditional on other annotations included in the model[40]:

$$\tau_c^* = \frac{\tau_c \mathrm{sd}_c}{h_g^2/M} \quad (17)$$

where sd_c is the s.e.m. of annotation c, h_g^2 is the total SNP heritability and M is as defined previously. τ_c^* is the proportionate change in per-SNP heritability associated with an increase of 1 s.d. in the value of a annotation.

We assessed the statistical significance of the enrichment score and τ^* via block-jackknife, as in previous work[11], with significance thresholds determined via false discovery rate (FDR) correction (q-value < 0.05)[103]. The FDR was calculated over all relevant relatively independent traits for a tissue and all programs of a particular type (cell-type programs, disease-dependent programs, cellular process programs) derived from that tissue. We used the P value for nonzero enrichment score as our primary metric, because τ^* is often nonsignificant for small cell-type-specific annotations when conditioned on the baseline-LD model[104].

MAGMA gene-level and GSEAs

MAGMA assesses the enrichment of genes and gene sets with disease. MAGMA v.1.08 was run using a 0-kb window around each gene to link SNPs to genes, using all default MAGMA parameters for running the gene-level analysis, and using the 1000 Genomes reference panel for the genotype LD reference. For the gene-set-level analysis, two types of analysis were performed: (1) a binary gene-set analysis by thresholding the gene programs at different thresholds of program score (ranging from 0.2 to 0.95) (using the --set-annot flag in MAGMA) and (2) a continuous variable-based analysis by treating the gene program probabilistic grade or −log(odds) of the probabilistic grade as continuous gene-level variables (using the --gene-covar flag in MAGMA).

GWAS summary statistics

We analyzed publicly available GWAS summary statistics for 60 unique diseases and traits with genetic correlation <0.9. Each trait passed the filter of being well powered enough for heritability studies (z-score for observed heritability >5 as in previous work including Finucane et al.[18]). We used the summary statistics for SNPs with minor allele count >5 in a 1000 Genomes Project European reference panel[98]. The lung FEV_1:forced vital capacity (FVC) trait was corrected for height data. For COVID-19, we analyzed two phenotypes: general COVID-19 (Covid versus population, liability scale heritability, $h^2 = 0.05$, s.e.m. = 0.01) and severe COVID-19 (hospitalized Covid versus population, liability scale heritability, $h^2 = 0.03$, s.e.m. = 0.01)[105] (meta-analysis round 4, 20 October 2020: https://www.covid19hg.org/).

Computing a sensitivity/specificity index

We define a sensitivity/specificity index to benchmark (1) sc-linker versus MAGMA gene-set enrichment analysis (GSEA) and (2) different versions of the sc-linker corresponding to varying ways to define cell-type programs and SNP-to-gene linking strategies.

For the comparison of the sc-linker with MAGMA, we define the sensitivity/specificity index as the difference of (1) the average of $-\log_{10}(P)$ of enrichment score (association) using the sc-linker (MAGMA) for 'expected enrichments' (gene program, trait) combinations (sensitivity) and (2) the average of $-\log_{10}(P)$ of GSEA (association) using the sc-linker (MAGMA) for 'other enrichments' (gene program, trait) combinations (specificity). In Fig. 4e, the expected enrichment combinations include immune programs for blood cell traits and immune diseases, and brain programs for brain-related traits[49,50]; all other combinations are considered to be other enrichments. In Supplementary Fig. 8, the expected enrichment combinations include B and T cells for lymphocyte percentage, monocytes for monocyte percentage, megakaryocytes for platelet count, erythroid for red blood cell (RBC) counts and RBC distribution width; all other combinations of cell types and traits are considered to be other enrichments[49,50]. A limitation of the sensitivity/specificity index is that other enrichments may be biologically real in some cases; thus, we also consider sensitivity to detect expected enrichments.

For the comparison of the different versions of the sc-linker approach using either varying definitions of cell-type programs (Supplementary Figs. 6 and 7) or different ways to link SNPs to genes beyond Roadmap–ABC enhancer–gene-linking strategy (Fig. 3d,e and Supplementary Fig. 3), we use a slightly different definition of sensitivity/specificity index. Instead of the $-\log_{10}(P)$ value, we use the τ^* metric from the S-LDSC method, which evaluates conditional information in the SNP annotation corresponding to a gene program, corrected for the annotation size. This metric is preferred when comparing across cell-type programs or enhancer–gene-linking strategies that are widely different in their corresponding SNP annotation sizes, as is the case in these comparisons (we note that use of this metric is not possible in comparisons involving MAGMA, which does not estimate τ^*).

Identifying genes driving heritability enrichment

For each gene program, we first subset the full gene list to only consider genes with >80% probability grade of membership in the gene program. Subsequently, we ranked all remaining genes using MAGMA (v.1.08) gene-level significance score and considered the top 50 ranked genes for further downstream analysis, which is different from the top 200 genes used for a 'baseline' method for scoring cell-type enrichments for disease that we used as a benchmark for sc-linker.

Identifying statistically significant differences in cell-type proportions

To identify changes in cell-type proportions between healthy and disease tissue, we used a multinomial regression test to jointly test changes across all cell types simultaneously. This helps account for all cell-type changes simultaneously, because an increase in the number of cells of one cell type implies that fewer cells of the other cell type will be captured. This regression model and the associated P values were calculated using the multinom function in the nnet v.7.3–17R package.

Statistics and reproducibility

All data used in the present study were generated and designed by the original studies in which they appear. No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment. All sc-linker heritability enrichment and significance P values are computed using a one-sided S-LDSC test. Multiple hypothesis correction was performed at the level of each scRNA-seq dataset across all cell-type and disease pairs.

Reporting summary

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

Data availability

All postprocessed scRNA-seq data (except for AD; see below) are available through the original publications with PMIDs: 28091601, 33208946, 31316211, 31097668, 31042697, 31348891, 32832598, 31209336, 31604275, 33654293, 32403949 and 30355494. In addition,

gene programs, enhancer–gene-linking annotations, supplementary data files and high-resolution figures are publicly available online at https://data.broadinstitute.org/alkesgroup/LDSCORE/Jagadeesh_Dey_sclinker. The AD scRNA-seq data[30] are available exclusively at https://www.radc.rush.edu/docs/omics.htm per its data usage terms. This work used summary statistics from the UK Biobank study (http://www.ukbiobank.ac.uk). The summary statistics for UK Biobank used in this paper are available at https://data.broadinstitute.org/alkesgroup/UKBB. The 1000 Genomes Project Phase 3 data are available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/2013050. The baseline-LD annotations are available at https://data.broadinstitute.org/alkesgroup/LDSCORE. We provide a web interface to visualize the enrichment results for different programs used in our analysis at https://share.streamlit.io/karthikj89/scgenetics/www/scgwas.py.

Code availability

This work uses the S-LDSC software (https://github.com/bulik/ldsc) to process GWAS summary statistics as well as S-LDSC software and MAGMA v.1.08 (https://ctg.cncr.nl/software/magma) for post-hoc analysis. Code for constructing cell-type, disease-dependent and cellular process gene programs from scRNA-seq data and performing the healthy and disease-shared NMF can be found at https://github.com/karthikj89/scgenetics (https://doi.org/10.5281/zenodo.6516048)[106]. Code for processing gene programs and combining with enhancer–gene links can be found at https://github.com/kkdey/GSSG (https://doi.org/10.5281/zenodo.6513166)[107].

References

93. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
94. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
95. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
96. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for dimension reduction. Preprint at https://doi.org/10.48550/arXiv.1802.03426 (2020).
97. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
98. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
99. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
100. Nasser, J. et al. Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
101. Dey, K. K. et al. SNP-to-gene linking strategies reveal contributions of enhancer-related and candidate master-regulator genes to autoimmune disease. *Cell Genomics* **2**, 100145 (2022).
102. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
103. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the *q*-value. *Ann. Stat.* **31**, 2013–2035 (2003).
104. van de Geijn, B. et al. Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. *Hum. Mol. Genet.* **29**, 1057–1067 (2020).
105. The COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur. J. Hum. Genet.* **28**, 715–718 (2020).https://doi.org/10.1038/s41431-020-0636-6
106. Jagadeesh, K., Dey, K. K. & Mohan, R. karthikj89/scgenetics: v1.0.0. *Zenodo* https://doi.org/10.5281/zenodo.6516048 (2022).
107. Dey, K. K. & Jagadeesh, K. A. kkdey/GSSG: sclinker_NatGenet. *Zenodo* https://doi.org/10.5281/zenodo.6513166 (2022).

Acknowledgements

We thank L. Gaffney for assistance with preparing figures as well as S. Chen, C. Smillie, B. Eraslan, A. Jaiswal and the entire groups of A.L.P and A.R. for helpful scientific discussions. This work was funded through the National Institutes of Health (NIH) F32 Fellowship (to K.A.J.), NIH Pathway to Independence K99/R00 award K99HG012203 (to K.K.D), NHGRI Genomic Innovator award (R35HG011324), by Gordon and Betty Moore, the BASE Research Initiative at the Lucile Packard Children's Hospital at Stanford University, NIH Pathway to Independence award (R00HG009917) (to J.M.E), NIH grants (nos. U01 HG009379, R01 MH101244, R37 MH107649, R01 HG006399, R01 MH115676 and R01 MH109978) to A.L.P. and Klarman Cell Observatory, HHMI, the Manton Foundation and NIH grant (no. 5U24AI118672) to A.R. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

K.A.J., K.K.D., A.L.P. and A.R. designed the study. K.A.J. and K.K.D. developed statistical methodologies and performed all computational analyses. A.L.P. and A.R. provided expert guidance and feedback on analysis and results. D.T.M. interpreted biological signals and guided K.A.J. and K.K.D. on highlighting biological insights. K.A.J. and R.M. designed and developed the web interface to visualize the results. J.M.E. provided ABC mappings. S.G. provided guidance on enhancer–gene-linking strategies. R.J.X. provided guidance on biological interpretations. K.A.J., K.K.D, A.L.P. and A.R. wrote the manuscript with detailed input from D.T.M. and feedback from all authors.

Competing interests

A.R. is a co-founder and equity holder of Celsius Therapeutics and an equity holder in Immunitas and was an SAB member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov. From 1 August 2020, A.R. has been an employee of Genentech. The remaining authors declare no competing interests.

Additional information

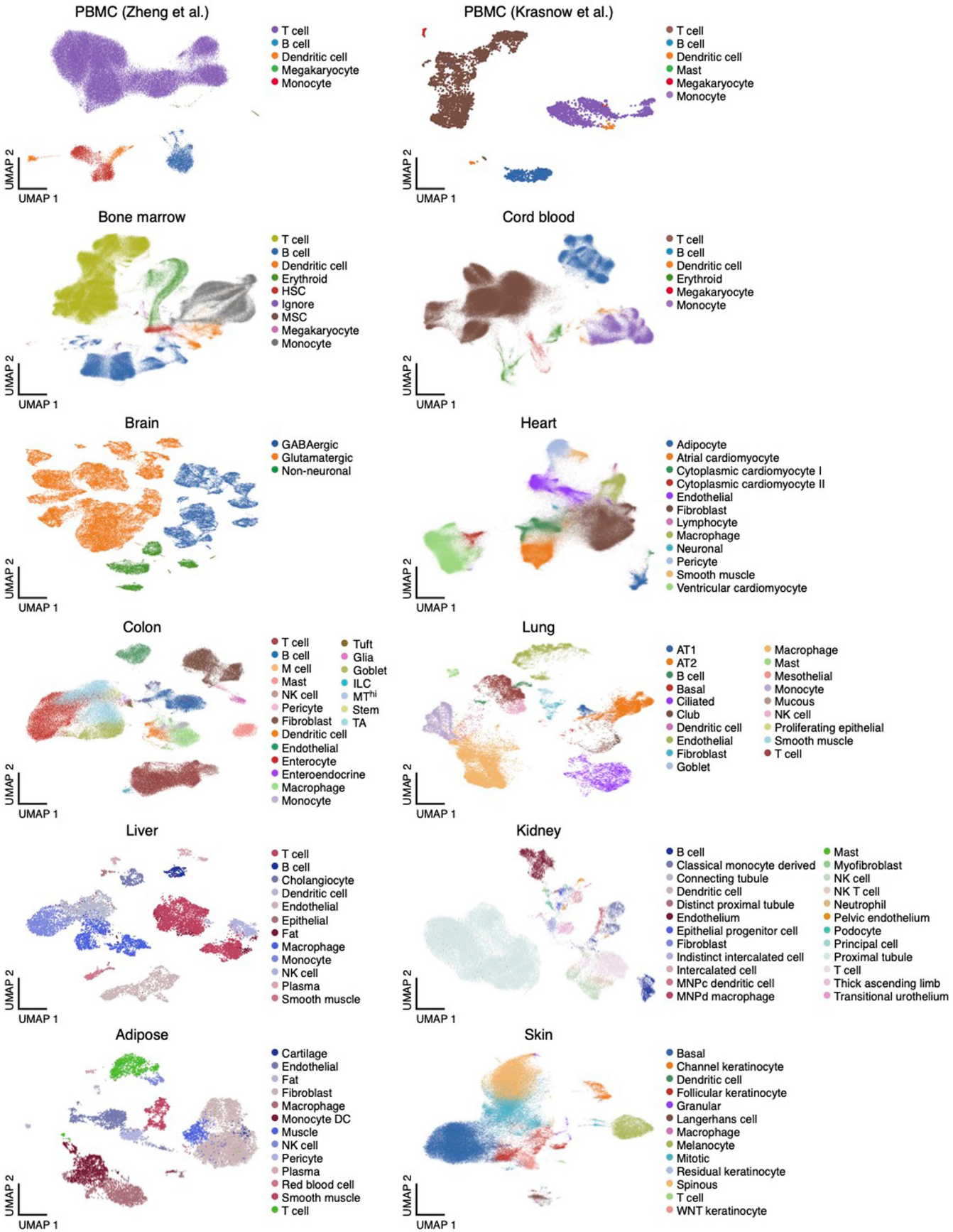
Extended data are available for this paper at https://doi.org/10.1038/s41588-022-01187-9.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01187-9.

Correspondence and requests for materials should be addressed to Karthik A. Jagadeesh, Kushal K. Dey, Alkes L. Price or Aviv Regev.

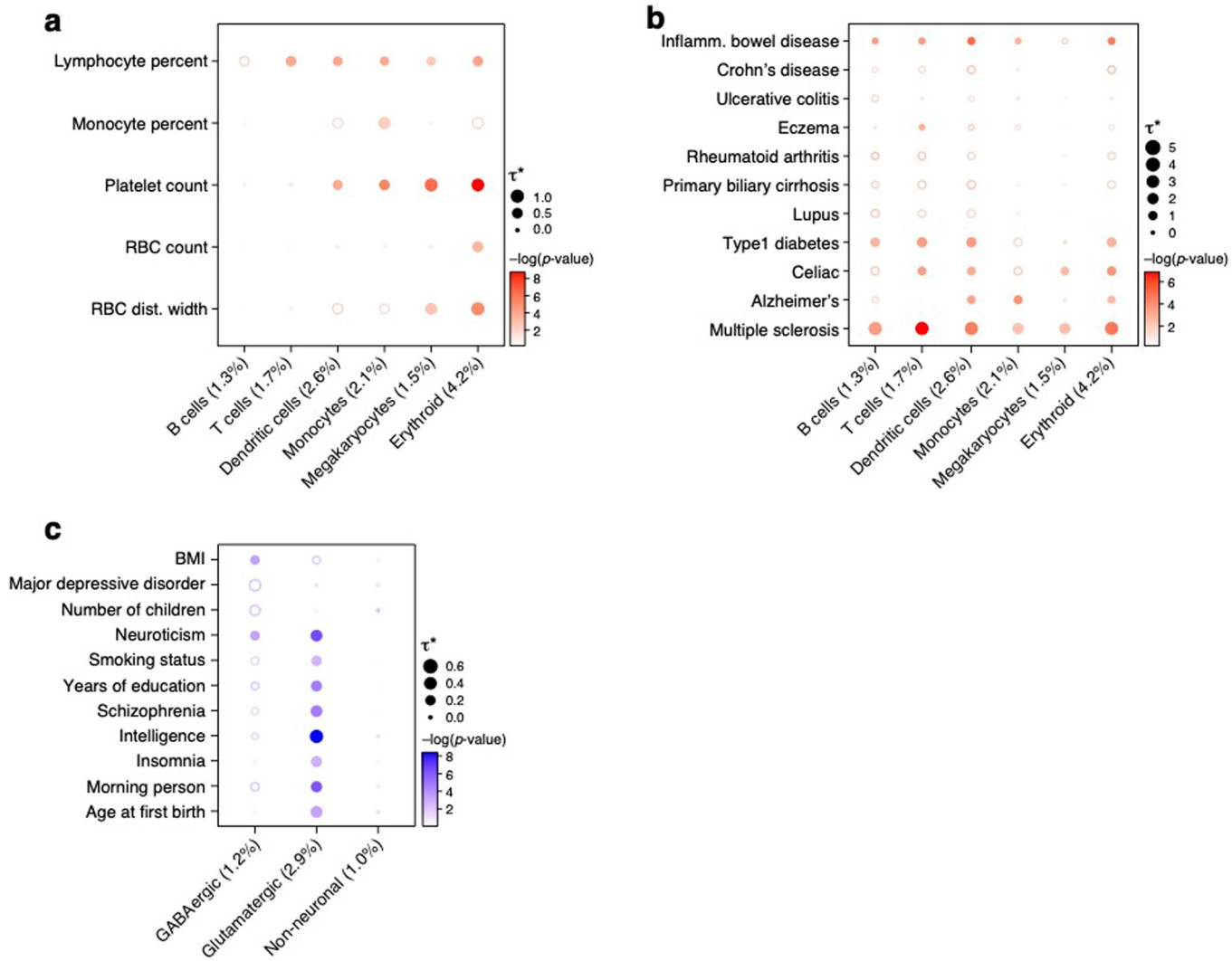
Peer review information *Nature Genetics* thanks Danielle Posthuma, Yukinori Okada, Rachel Brouwer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41588-022-01187-9



Extended Data Fig. 1 | Single-cell RNA-seq datasets. UMAP embedding of scRNA-seq profiles (dots) colored by cell type annotations from 12 datasets (labels on top).

https://doi.org/10.1038/s41588-022-01187-9

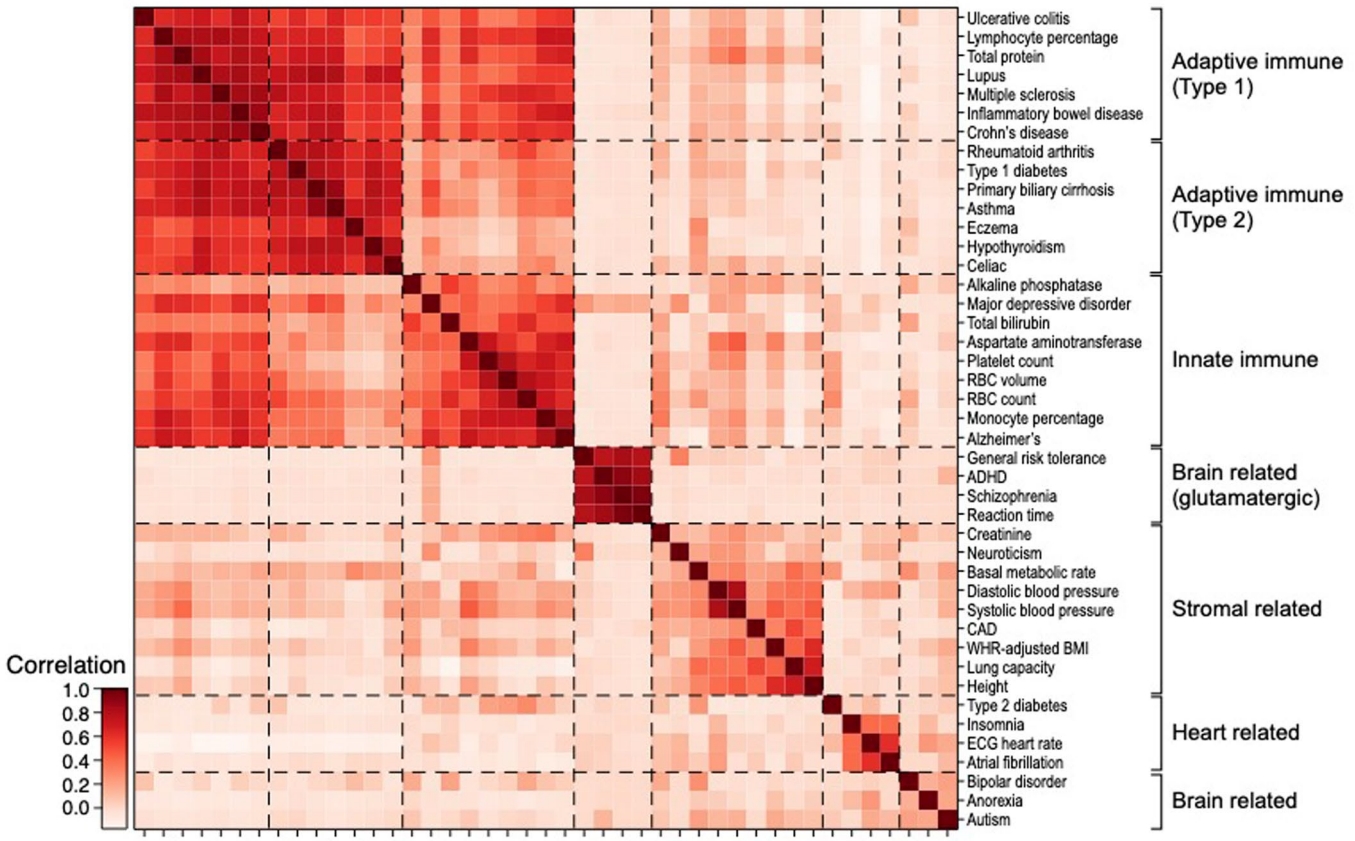


Extended Data Fig. 2 | Standardized effect sizes of immune and brain cell type programs. Standardized effect size (τ^*) (dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of immune (**a**,**b**) or brain (**c**) cell type programs (columns) for blood cell traits (**a**), immune disease traits (**b**), or neurological/psychological related traits (**c**), based on SNP annotations generated with the Roadmap∪ABC-immune (**a**,**b**) or Roadmap∪ABC-brain (**c**) enhancer-gene linking strategy. Numerical results are reported in Supplementary Data 1. Details for all traits analyzed are in Supplementary Table 2.

https://doi.org/10.1038/s41588-022-01187-9

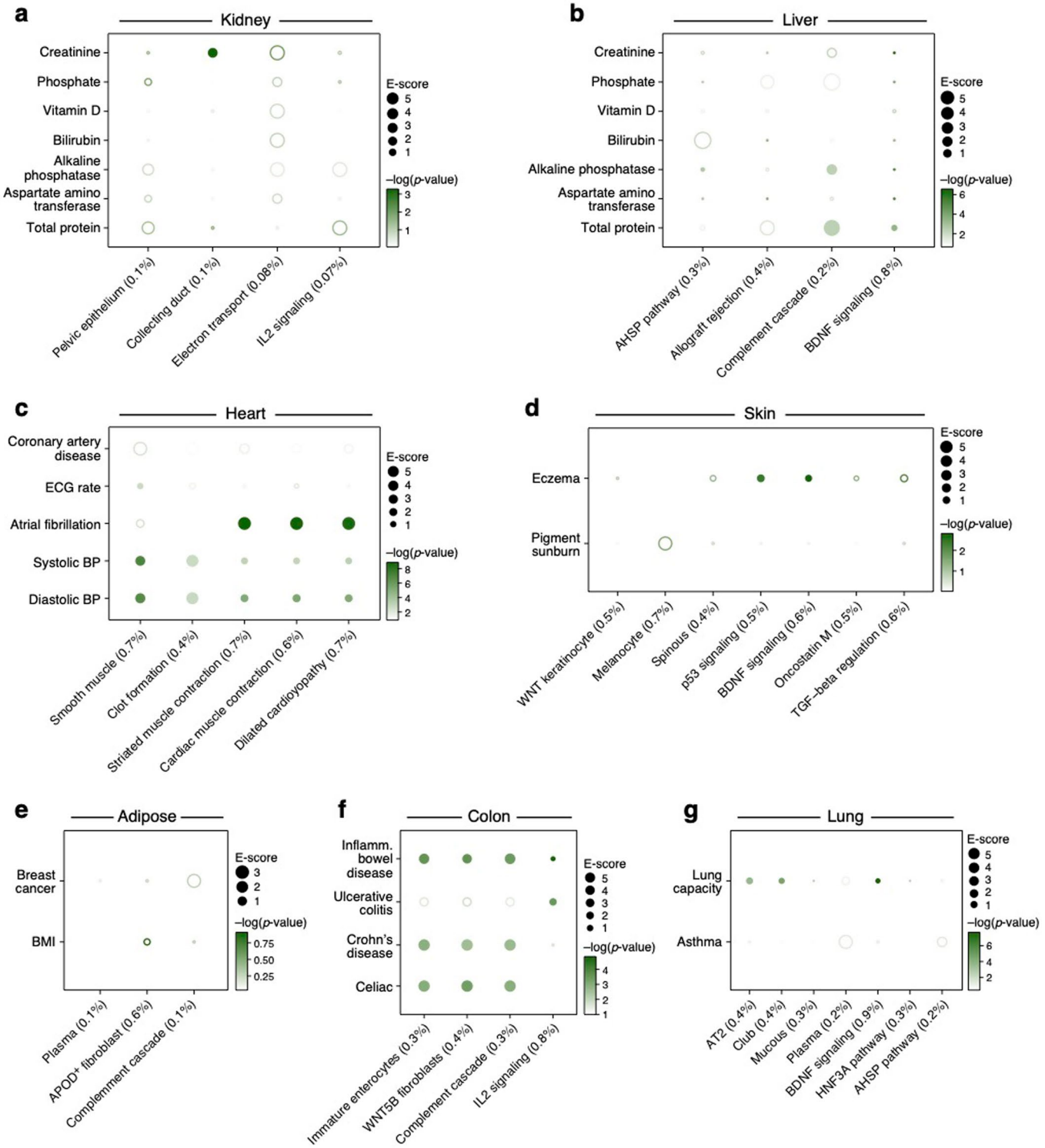
Extended Data Fig. 3 | Linking cell type programs to diseases and traits across all analyzed tissues. Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of cell type programs (columns) from each of nine tissues (color code, legend) for GWAS summary statistics of diverse traits and diseases (rows), based on the Roadmap∪ABC enhancer-gene linking strategy for the corresponding tissue. Details for all traits analyzed are in Supplementary Table 2. See **Data Availability** for higher resolution version of this figure.

https://doi.org/10.1038/s41588-022-01187-9



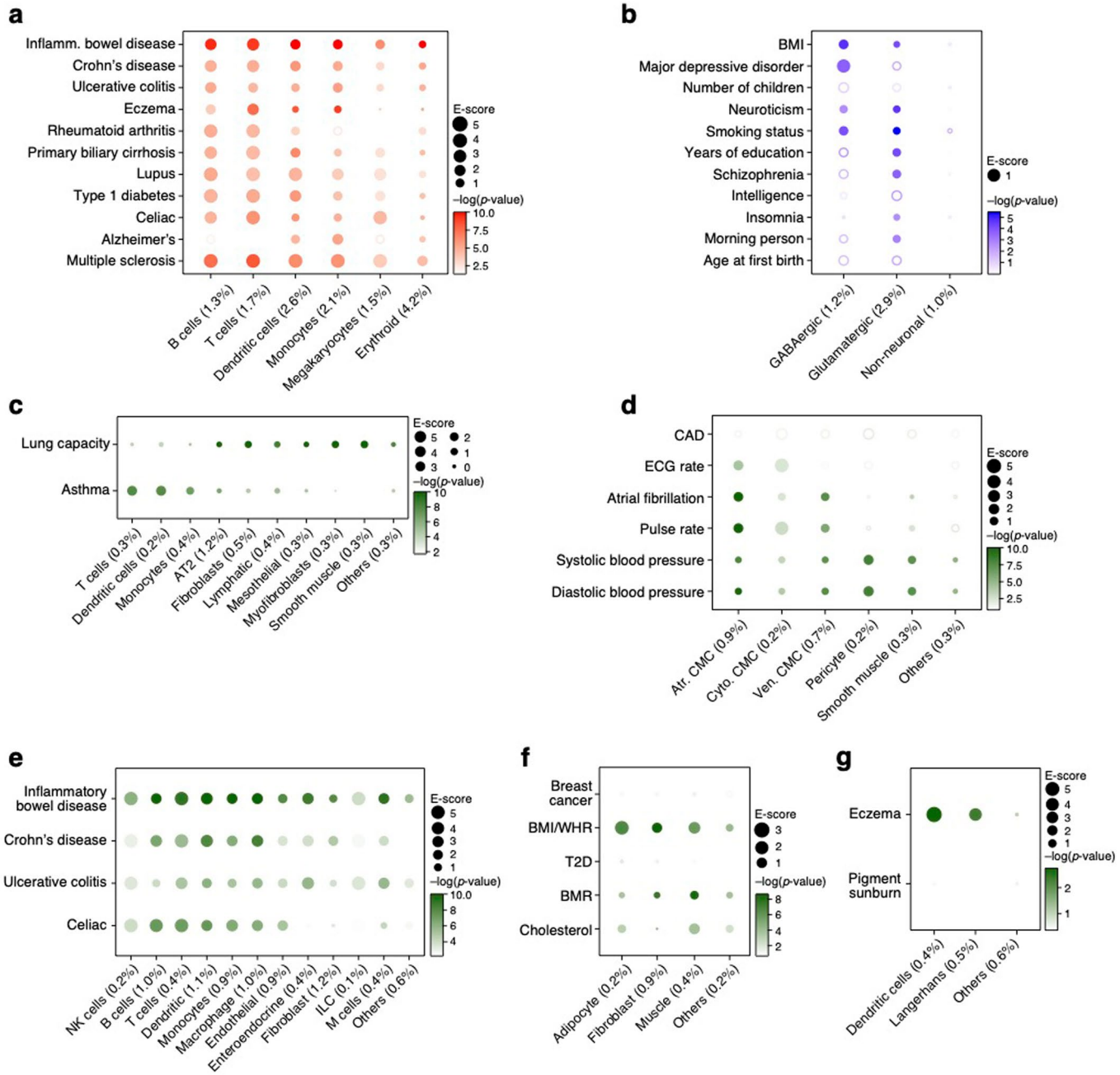
Extended Data Fig. 4 | Cross trait analysis of cell type enrichments. Pearson correlation coefficient (colorbar) between the cell type enrichment profiles of each pair of traits (rows, columns), clustered (dashed lines) hierarchically. Trait clusters labeled by their overall cell type enrichments.

https://doi.org/10.1038/s41588-022-01187-9



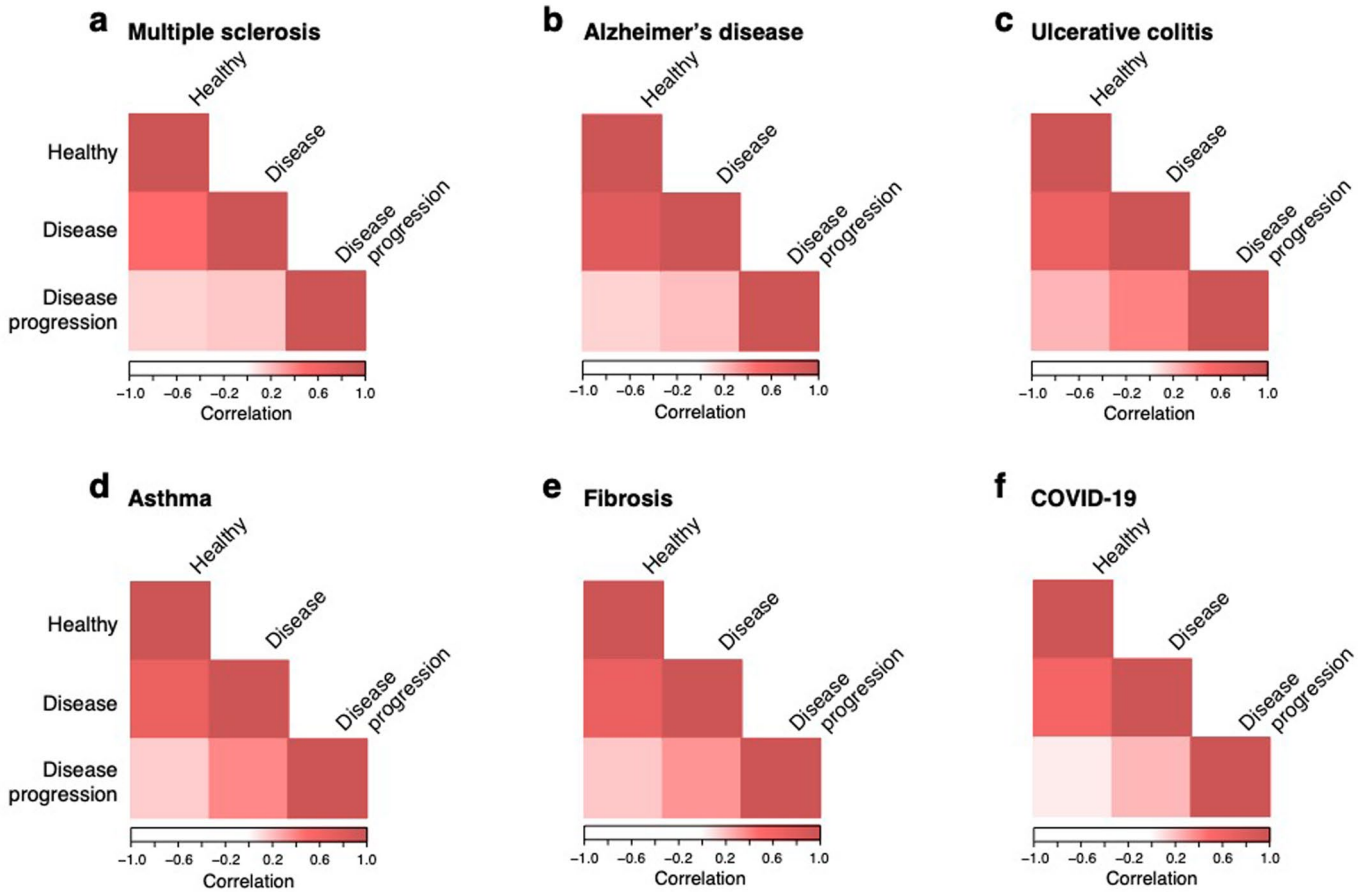
Extended Data Fig. 5 | Linking cellular process programs to relevant diseases and traits in each of six tissues. Magnitude (E-score, dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of cellular process programs (columns; obtained by NMF) in each of seven tissues (label on top) for traits relevant in that tissue (rows) using the Roadmap∪ABC strategy for the corresponding tissue. Details for all traits analyzed are in Supplementary Table 2.

https://doi.org/10.1038/s41588-022-01187-9



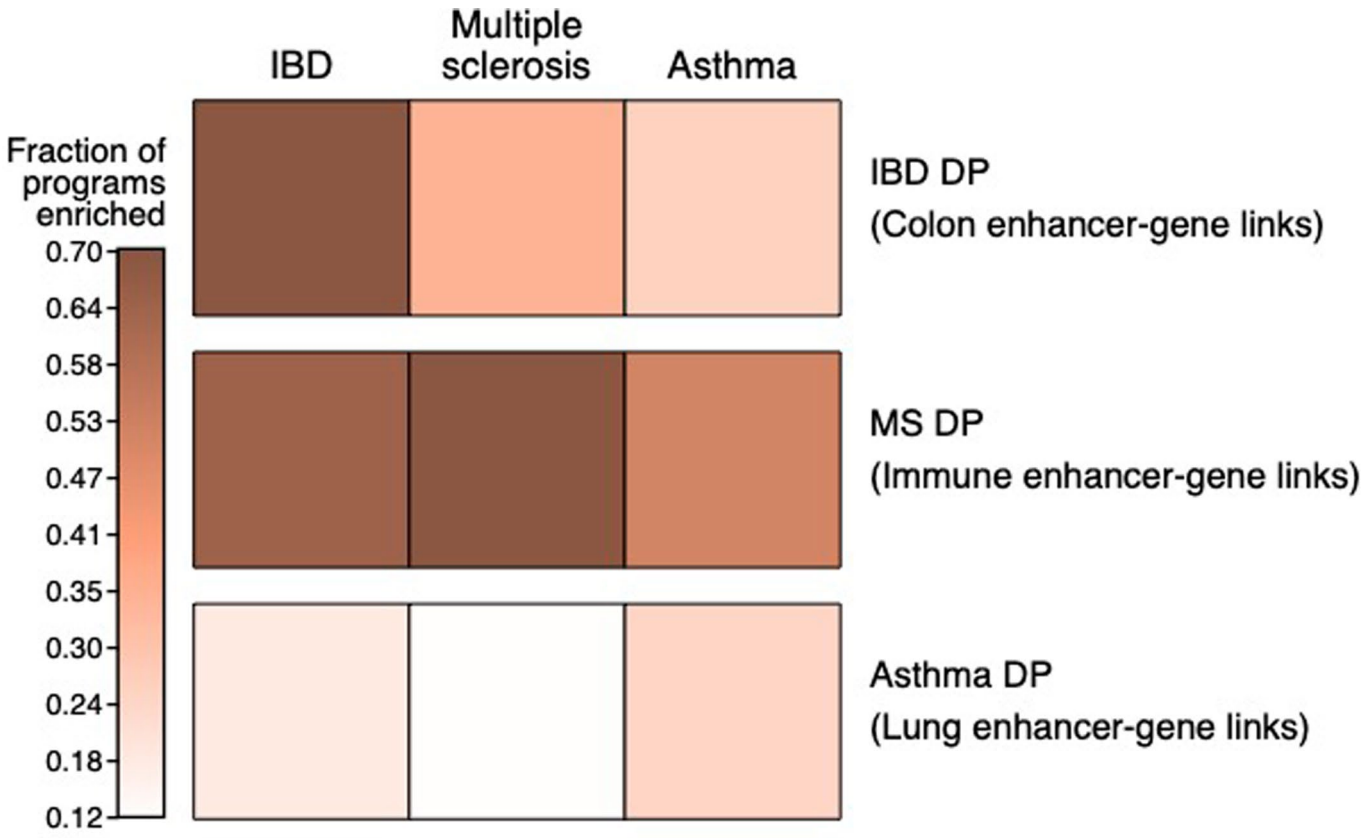
Extended Data Fig. 6 | Analysis of cell type programs using a non-tissue-specific enhancer-gene linking strategy. Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune (**a**), brain (**b**), lung (**c**), heart (**d**), colon (**e**), adipose (**f**) and skin (**g**) cell type programs (columns) for traits relevant in that tissue (rows) using a non-tissue-specific Roadmap∪ABC strategy. Details for all traits analyzed are in Supplementary Table 2.

https://doi.org/10.1038/s41588-022-01187-9



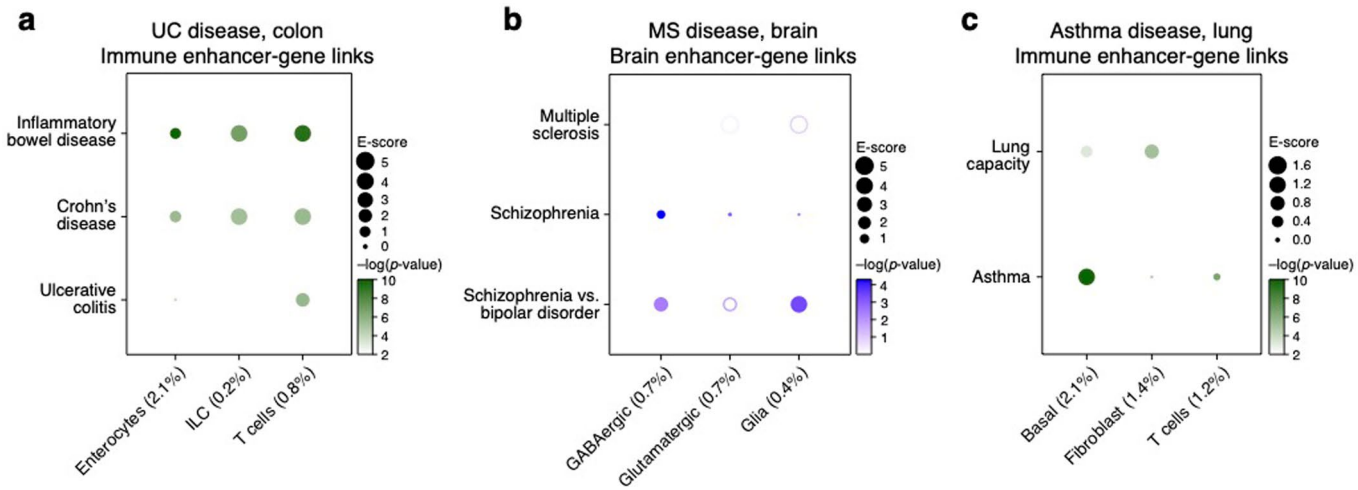
Extended Data Fig. 7 | Disease-dependent programs have low correlations with healthy and disease cell type programs. Pearson correlation coefficient (color bar) of gene program membership vectors between healthy cell type, disease cell type and disease-dependent programs in scRNA-seq studies from a disease tissue (label on top) and the corresponding healthy tissue.

https://doi.org/10.1038/s41588-022-01187-9



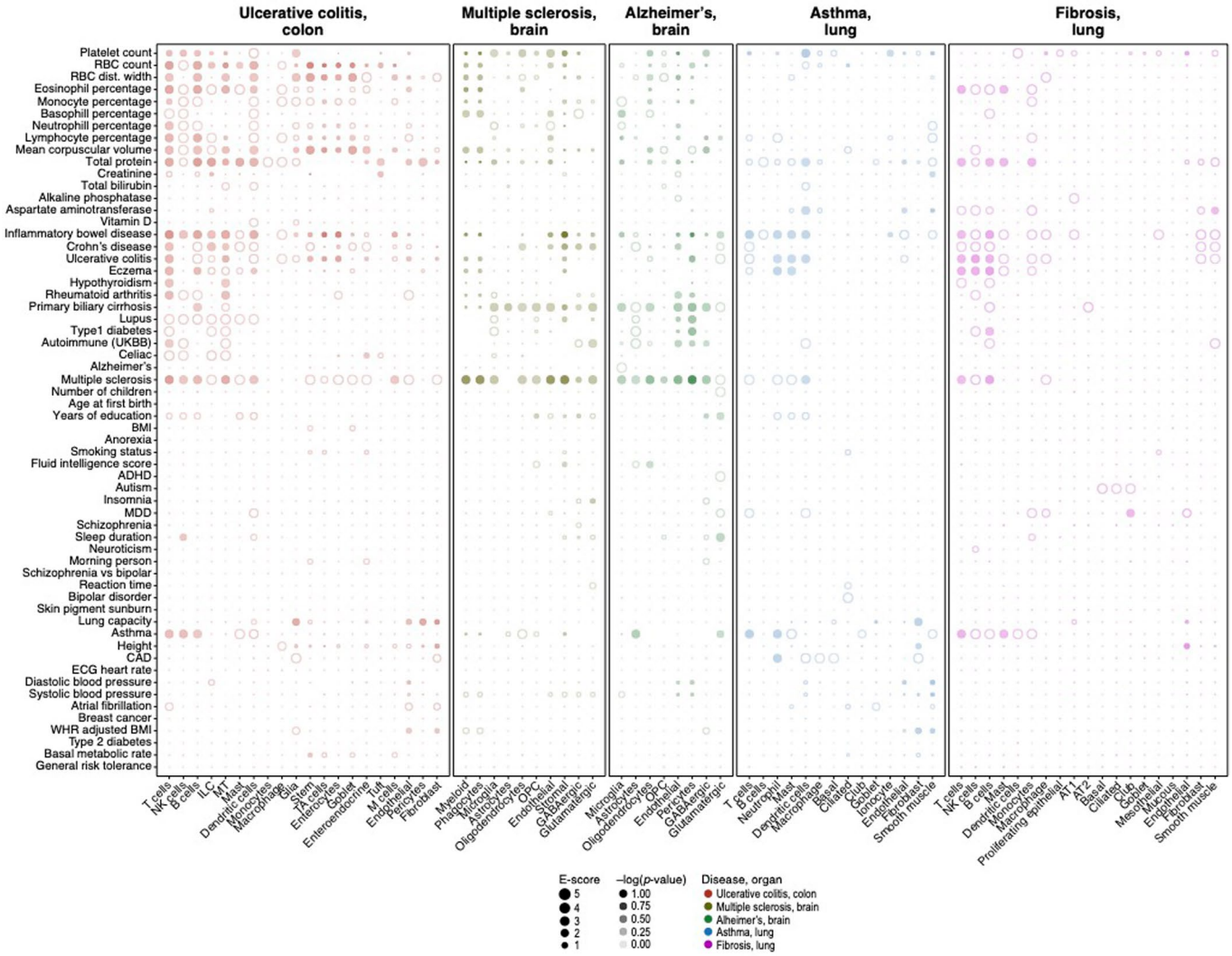
Extended Data Fig. 8 | Disease specificity of disease-dependent programs. Proportion of disease-dependent programs with a -$\log_{10}$(P-value) of enrichment score (p.E-score) > 3 in IBD, MS and asthma GWAS summary statistics (column) for disease-dependent programs from IBD, MS and asthma (columns), when combined with tissue-specific Roadmap∪ABC (row).

https://doi.org/10.1038/s41588-022-01187-9



Extended Data Fig. 9 | Analysis of disease-dependent programs using alternative Roadmap∪ABC enhancer-gene linking strategies. Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of disease-dependent programs (columns) in UC (colon cells) using Roadmap∪ABC-immune (**a**), asthma (lung cells) using Roadmap∪ABC-immune (**b**), and MS (brain cells) using Roadmap∪ABC-brain (**c**). Details for all traits analyzed are in Supplementary Table 2.

https://doi.org/10.1038/s41588-022-01187-9



Extended Data Fig. 10 | Analysis of disease-dependent programs across all tissues and traits. Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of disease-dependent programs (columns) from UC, MS, Alzheimer's, asthma and pulmonary fibrosis (labels on top, color code, legend), for GWAS summary statistics of diverse traits and diseases (rows), based on the Roadmap∪ABC enhancer-gene linking strategy for the corresponding tissue. Details for all traits analyzed are in Supplementary Table 2. See **Data Availability** for higher resolution version of this figure.

nature research

Corresponding author(s): Karthik A. Jagadeesh, Kushal K. Dey, Alkes L. Price, Aviv Regev

Last updated by author(s): May 19, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We used the Harmony python package for single cell batch correction, the scanpy python package for single cell data analysis, the nnet R package, the LDSC package available on github at https://github.com/bulik/ldsc for computing genetic heritability, and the MAGMA gene/gene set prioritization software. Additionally, all custom code developed in this study for analysis of single cell data is available at github at: https://github.com/kkdey/GSSG and https://github.com/karthikj89/scgenetics. This work uses the S-LDSC software (https://github.com/bulik/ldsc) to process GWAS summary statistics as well as S-LDSC software and MAGMA v1.08 (https://ctg.cncr.nl/software/magma) for post-hoc analysis. Code for constructing cell type, disease-dependent and cellular process gene programs from scRNA-seq data and performing the healthy and disease shared NMF can be found at https://github.com/karthikj89/scgenetics (DOI 10.5281/zenodo.6516048). Code for processing gene programs and combining with enhancer-gene links can be found at https://github.com/kkdey/GSSG (DOI 10.5281/zenodo.6513166).

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

Data

Policy information about availability of data

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

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

All postprocessed scRNA-seq data (except for Alzheimer's disease; see below) are available through the original publications with PMIDs: 28091601, 33208946, 31316211, 31097668, 31042697, 31348891, 32832598, 31209336, 31604275, 33654293, 32403949, 30355494. Additionally, gene programs, enhancer-gene linking annotations, supplementary data files and high-resolution figures are publicly available online at https://data.broadinstitute.org/alkesgroup/LDSCORE/Jagadeesh_Dey_sclinker. The Alzheimer's disease scRNA-seq data8 is available exclusively at https://www.radc.rush.edu/docs/omics.htm per its data usage terms. This work used summary statistics from the UK Biobank study (http://www.ukbiobank.ac.uk/). The summary statistics for UK Biobank used in this paper are available at https://data.broadinstitute.org/alkesgroup/UKBB/. The 1000 Genomes Project Phase 3 data are available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/2013050. The baseline-LD annotations are available at https://data.broadinstitute.org/alkesgroup/LDSCORE/. We provide a web interface to visualize the enrichment results for different programs used in our analysis at: https://share.streamlit.io/karthikj89/scgenetics/www/scgwas.py.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	We broadly analyzed across 60 available GWAS data that were picked based on relevance to the scRNA-seq data analyzed.
Data exclusions	We analyzed only autosomes based on pre-established exclusion criteria as seen in Finucane et al 2018.
Replication	Where possible, we replicated our computational results using independent scRNA-seq data sets from the same tissue.
Randomization	We did not allocate samples in experimental groups.
Blinding	There was no group allocation.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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