TITLE

Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics

AUTHORS

Karthik A. Jagadeesh[1,*,‡], Kushal K. Dey[2,*,‡], Daniel T. Montoro[1], Steven Gazal[2], Jesse M. Engreitz[1], Ramnik J. Xavier[1], Alkes L. Price[1,2,3,**,‡], Aviv Regev[1,4,5**,‡]

AFFILIATIONS

[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 Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[4]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

[5]Current address: Genentech, 1 DNA Way, South San Francisco, CA, USA

* Equal contribution

** Co-senior authors

‡ To whom correspondence should be addressed: kjag@broadinstitute.org (KAJ), kdey@hsph.harvard.edu (KKD), aprice@hsph.harvard.edu (AP), aregev@broadintitute.org (AR)

ABSTRACT

Cellular dysfunction is a hallmark of disease. Genome-wide association studies (GWAS) have provided a powerful means to identify loci and genes contributing to disease risk, but in many cases the related cell types/states through which genes confer disease risk remain unknown. Deciphering such relationships is important both for our understanding of disease, and for developing therapeutic interventions. Here, we introduce a framework for integrating single-cell RNA-seq (scRNA-seq), epigenomic maps and GWAS summary statistics to infer the underlying cell types and processes by which genetic variants influence disease. We analyzed 1.6 million scRNA-seq profiles from 209 individuals spanning 11 tissue types and 6 disease conditions, and constructed gene programs capturing cell types, disease progression in cell types, and cellular processes both within and across cell types. We evaluated these gene programs for disease enrichment by transforming them to SNP annotations with tissue-specific epigenomic maps and computing enrichment scores across 60 diseases and complex traits (average N=297K). The inferred disease enrichments recapitulated known biology and highlighted novel relationships for different conditions, including GABAergic neurons in major depressive disorder (MDD), disease progression programs in M cells in ulcerative colitis, and a disease-specific complement cascade process in multiple sclerosis. Our framework provides a powerful approach for identifying the cell types and cellular processes by which genetic variants influence disease.

INTRODUCTION

Genome wide association studies (GWAS) 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 non-coding 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 from bulk 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 processes within and across cell types, and leverage enhancer-gene links(*19–23*) to improve power.

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 map GWAS variants to function, including programs that reflect cell-type-specific signatures(*25–28*), disease-specific effects within cell types(*29*, *30*), co-varying gene programs within a cell type, and conditions reflecting key cellular processes(*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 lead to misleading conclusions as 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 (CNS)(*38*). Thus, there is a need for a principled method that combines human genetics and comprehensive scRNA-seq across multiple tissues and organs.

Here, we develop and apply 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 11 tissue types and 6 disease conditions (including COVID-19), spanning 209 individuals and 1.6 million cells. We transform gene programs to SNP annotations using tissue-specific enhancer-gene links(*19–23*) and then link SNP annotations to diseases by applying stratified LD score regression(*11*) (S-LDSC) with the baseline-LD model(*39*, *40*) 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 shedding light on underlying disease mechanisms.

RESULTS

Linking cell types, disease progression and cellular processes to disease

We developed a framework to link gene programs derived from scRNA-seq with diseases and complex traits (**Figure 1A**). Our approach consists of three steps. (**1**) We use scRNA-seq to construct gene programs, defined as probabilistic gene sets that characterize individual cell types (*vs.* other cell types), disease progression in individual cell types (*vs.* healthy cells of the same type), or cellular processes both within and across cell types in both healthy and disease states (**Methods**). (**2**) We linked 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*). (**3**) We evaluated 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(*39*, *40*). 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; **Code Availability**), along with all gene programs, enhancer-gene linking strategies, and SNP annotations analyzed in this study (**Data Availability**).

In step 1, we constructed three types of gene programs from scRNA-seq data (**Figure 1B**): (i) cell type programs representing genes specifically enriched in an individual broad cell type (e.g. T cells) of a tissue compared to other cell types in that tissue; (ii) disease progression cell type programs representing disease-specific differences in gene expression within the same cell type; and (iii) cellular process programs capturing gene co-expression patterns within and across cell types (**Methods**). We constructed cell type programs by assessing the differential expression of each gene for the focal cell type *vs*. all other cell types in the tissue (with cell types defined by clustering(*41*),(*42*) and annotated *post hoc*) and transforming each gene's Z score to a probabilistic score (**Methods**). We constructed disease progression cell type programs by assessing differential expression between cells of the same type in healthy *vs*. disease tissue and transforming each gene's Z score to a probabilistic score (**Methods**), aiming to capture risk genes involved in disease progression and severity of symptoms after onset. Disease progression cell type programs had low correlation on average with healthy cell type programs of the same cell type (Pearson r=0.16; see below). We constructed cellular process programs by non-negative matrix factorization(*43*) (NMF) of normalized gene expression values and a modified NMF (to jointly model both healthy and disease states), with the latent factors (programs) aiming to capture gene expression signatures

within and across cell types. We annotated each cellular process program by the pathway most enriched in the top genes and labeled it as an 'intra-cell type' or 'inter-cell type' cellular process program if highly correlated with only one or multiple cell type programs, respectively (**Methods**). Intra-cell type cellular processes often corresponded to narrower cell types (e.g. CD4 T cells) which were cell subsets of broader cell type gene programs (e.g. T cells).

In step 2, we transformed the genes prioritized by each program into SNP annotations by linking each gene to SNPs that may regulate their activity in *cis* (**Figure 1A**). We generated SNP annotations using an enhancer-gene linking strategy, defined as an assignment of 0, 1 or more linked genes to each SNP, combining Roadmap Enhancer-Gene Linking (Roadmap)(*19*, *21*) and Activity-By-Contact (ABC)(*22*, *23*) strategies (Roadmap∪ABC) in the tissue underlying the program of interest (**Methods**). We used tissue level enhancer-gene links instead of cell type level because in its current form, the cell type enhancer-gene links are more noisy and sparse resulting in SNP annotations that are too sparse for heritability analysis.

In step 3, we evaluated each gene program for disease heritability enrichment by applying S-LDSC(*11*) with the baseline-LD model(*39*, *40*) to the resulting SNP annotations (**Figure 1A**, **Methods**). The S-LDSC analysis was conditioned on 86 coding, conserved, regulatory and LD-related annotations from the baseline-LD model (v2.1)(*39*, *40*) (**Data Availability**), and uses heritability enrichment to evaluate informativeness for disease. Heritability enrichment is defined as the proportion of heritability explained by SNPs in an annotation divided by the proportion of SNPs in the annotation(*11*); this generalizes to annotations with values between 0 and 1(*44*). We further define the enrichment score of a gene program as the difference between the heritability

enrichment of the SNP annotation corresponding to the gene program of interest and the SNP annotation corresponding to a gene program assigning a probabilistic grade of 1 to all protein-coding genes with at least one enhancer-gene link in the relevant tissue (**Methods**). We use the p-value of the enrichment score as our primary metric, assessing statistical significance using a genomic block-jackknife as in our previous work(*11*). We performed this analysis over healthy cell type programs (**data file S1**), disease progression cell type programs (**data file S2**), and cellular process programs (**data file S3**). We identified top genes driving disease enrichments based on proximity based MAGMA (v 1.08) gene-disease association scores(*45*) of genes with high probabilistic grade in each gene program (**Figure 1C, data file S4, Methods**) focusing on genes that are both (i) close to a GWAS signal and (ii) in an enriched gene program.

We analyzed a broad range of human scRNA-seq data, spanning 17 data sets from 11 tissues and 6 disease conditions. The 11 non-disease tissues include immune (peripheral blood mononuclear cells(*26*, *46*), cord blood(*27*), and bone marrow(*27*)), brain(*28*), kidney(*47*), liver(*48*), heart(*25*), lung(*29*), colon(*34*), skin(*49*) and adipose(*48*). The 6 disease conditions include multiple sclerosis (MS) brain(*50*), Alzheimer's disease brain(*30*), ulcerative colitis (UC) colon(*34*), asthma lung(*51*), idiopathic pulmonary fibrosis (IPF) lung(*29*) and COVID-19 bronchoalveolar lavage fluid(*52*) (**Figure S1**). In total, the scRNA-seq data includes 209 individuals, 1,602,614 cells and 256 annotated cell subsets (**Methods**, **table S1**). We also compiled publicly available GWAS summary statistics for 60 unique diseases and complex traits (genetic correlation < 0.9; average N=297K) (**Methods**, **table S2**). We analyzed gene programs from each scRNA-seq dataset in conjunction with each of 60 diseases and complex traits, but we primarily report those that are most pertinent for each program.

Linking B cells, T cells and monocytes to immune-related diseases

We first focused on scRNA-seq data from the immune system (**Figure 2A,B**, **Figure S1**). We constructed 6 immune cell type programs that were identified across 4 data sets – two from PBMC (k=4,640 cells; n=2 individuals(*26*); k=68,551; n=8 individuals(*46*)), and one each of cord blood(*27*) (k=263,828; n=8) and bone marrow(*27*) (k=283,894; n=8) – and 10 (intra-cell type and inter-cell type) immune cellular process programs (**Figure 2E**). (We did not construct disease progression programs, as these datasets included healthy samples only.)

To assess the validity of our approach, we first analyzed 5 blood cell traits that inherently correspond to underlying cell types (**table S2**), using the 6 immune cell type programs. We identified the expected cell type enrichments, including enrichment of erythroid cells for red blood cell count and of B cells and T cells for lymphocyte percentage (**Figure 2C**, **Figure S2A**). The Roadmap∪ABC enhancer-gene linking strategy outperformed other strategies in identifying expected enrichments (**Figure S3**). We observed higher specificity in enrichments of relevant cell type and trait pairs for our polygenic approach compared to functional enrichment of fine-mapped SNPs(*53–55*) (**Figure S4A**).

We next analyzed 11 autoimmune or inflammation-associated diseases (**table S2**) using the 6 immune cell type programs, identifying previously reported cell type-disease enrichments (**Figure 2D**, **Figure S2B**), as well as additional notable findings (see below). Expected enrichments included T cells for eczema(*56*, *57*), B and T cells for primary biliary cirrhosis (PBC)(*18*), and dendritic cells and monocytes for Alzheimer's disease(*58*). The top two genes driving the latter

enrichment were *MS4A6A* and *MS4A4A*; the *MS4A* gene cluster is a modulator of *TREM2*, which plays a critical role in Alzheimer's disease through monocyte-mediated microglial activation and phagocytosis(*59*).

Several of the significant cell type-disease enrichments are not as widely established. These notable findings (**Figure 2D**, **Table 1, Figure S2B**) may implicate previously unexplored biological mechanisms. For example, B cells have been detected in basal lymphoid aggregates in the ulcerative colitis (UC) colon, but their pathogenic significance remains unknown(*60*). We detected significant enrichment in B cells for UC, indicating a potential causal role. In addition, T cells were highly enriched for celiac disease; the top driving genes including *ETS1* (ranked 1), which is associated with T cell development and IL2 signaling(*61*), and *CD28* (ranked 3), which plays a critical role in T cell activation. This suggests that aberrant T cell maintenance and activation may impact the inflammation observed in celiac disease. Recent results reporting a permanent loss of resident gamma delta T cells in the celiac bowel and the subsequent recruitment of inflammatory T cells may further support this hypothesis(*62*). (Enrichment results for the remaining 49 diseases and traits with immune cell type programs are reported in **Figure S5**).

We next analyzed the 10 immune cellular process programs, including 6 intra-cell type and 4 inter-cell type programs (**Figure 2E**), across the 11 immune-related diseases and 5 blood cell traits. We identified several notable findings including both enrichments shared across diseases and disease-specific enrichments (**Figure 2F**, **Table 1**). While T cells have been previously linked to eczema, this analysis provided finer resolution to pinpoint higher enrichment in CD4+ T cells compared to CD8+ T cells. The IL2 signaling cellular process program (inter cell type; T and B cells) was

significantly enriched for both eczema and celiac disease, though the genes driving the enrichment were not significantly overlapping (p-value: 0.21). Additionally, the complement cascade cellular process program (inter cell type; plasma, B, and hematopoietic stem cells (HSCs)) was most highly enriched among all inter cellular programs for celiac disease. For Alzheimer's disease, there was a strong enrichment in both classical and non-classical monocyte intra-cell type cellular programs, and in MHC class II antigen presentation (inter cell type; dendritic cells (DCs) and B cells) and prostaglandin biosynthesis (inter cell 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 Th2 cell differentiation(*63*, *64*) and in mediating peripheral *CD4* T cell tolerance and allergic reactions(*65*, *66*); and *CD33* (ranked 1) in MHC class II antigen processing in Alzheimer's disease, a microglial receptor strongly associated with increased risk in previous GWAS(*67*, *68*).

Linking GABAergic and glutamatergic neurons to psychiatric disease

We analyzed scRNA-seq data from brain prefrontal cortex (k=73,191, n=10)(*28*) tissue (**Figure 3A**). We constructed 9 cell type programs, including GABAergic and glutamatergic neurons (as categorized in the Allen Brain Atlas(*28*)) (**table S1**), as well as 8 brain cellular process programs (**Figure 3E**). (We did not construct disease progression programs, as this dataset included healthy samples only.)

We analyzed 11 psychiatric or neurological diseases and traits (**table S2**) using the 9 brain cell type programs (**Figure 3A,B**), identifying several notable cell type-disease enrichments (**Figure 3C**, **Table 1, Figure S2C)**. In particular, major depressive disorder (MDD) and body mass index

(BMI) were highly enriched specifically in GABAergic neurons, whereas insomnia, schizophrenia (SCZ), and intelligence were highly enriched specifically in glutamatergic neurons, and neuroticism was highly enriched in both. GABAergic neurons regulate the brain's ability to control stress levels, which is the most prominent vulnerability factor in MDD(*69*). Among the top genes driving this enrichment were *TCF4* (ranked 1), a critical component for neuronal differentiation that affects neuronal migration patterns(*70*, *71*), and *PCLO* (ranked 4), which is important for synaptic vesicle trafficking and neurotransmitter release(*72–74*). The enrichment of glutamatergic neurons for SCZ was noted previously(*18*), and their enrichment for insomnia and intelligence is consistent with their role in maintaining electroencephalogram (EEG) synchronization during sleep(*75*), and in learning and memory(*76*, *77*), respectively. (Enrichment results for the remaining 49 diseases and traits in conjunction with brain cell type programs are reported in **Figure S5**).

We assessed the importance of tissue specificity of both the cell type program and the enhancer-gene strategy, 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 (**Figure 3D**). We reached two main conclusions. First, as expected, the immune (resp. brain) cell types produced far stronger signals for the immune (resp. brain) diseases compared to the brain (resp. immune) cell types. Second, linking the immune (resp. brain) cell type programs to SNPs using the immune (resp. brain) enhancer-gene linking strategy generated much stronger enrichment scores (>2x larger on average) across immune (resp. brain) diseases compared to using the brain (resp. immune) enhancer-gene strategy. This highlights the importance of leveraging the tissue specificity of enhancer-gene strategies.

We next analyzed the 12 brain cellular process programs, including 10 intra-cell type and 2 inter-cell type programs (**Figure 3E**), across the 11 psychiatric/neurological diseases and traits. We determined that the significant enrichment of neuronal cell types described previously is primarily driven by finer programs reflecting neuron subtypes (**Figure 3F, Table 1**). For example, the enrichment of GABAergic neurons for BMI was driven by programs reflecting LAMP5$^+$ and VIP$^+$ subsets; the respective top driving genes included *FLRT1* (for LAMP5$^+$ neurons; ranked 1), whose absence reduces intercellular adhesion and promotes premature neuron migration(*78*), and *TIMP2* (for VIP$^+$ neurons; ranked 7), implicated in obesity through hypothalamic control of food intake and energy homeostasis in mice(*79*, *80*). Furthermore, the enrichment of GABAergic neurons for MDD reflects SST$^+$ and PVALB$^+$ subsets; the respective top driving genes included *PCLO* (for SST$^+$ GABAergic neurons; ranked 2), and *ADARB1* (for PVALB$^+$ neurons; ranked 4), encoding an RNA editing enzyme that can edit the transcript for the serotonin receptor 2C with a role in MDD(*81*). We also observed more granular structure in the glutamatergic neurons (IT neurons were enriched for neuroticism, whereas L6 neurons were enriched for years of education and intelligence). Among inter cell type programs (shared across multiple cell types), electron transport cellular process programs (GABAergic and glutamatergic neurons) were enriched for several psychiatric/neurological traits, such as years of education, consistent with previous studies(*82*), with the top driving genes including *ATP6V0B* and *NDUFAF3* (ranked 1, 4).

Linking cell types from diverse human tissues to disease

We analyzed scRNA-seq data from a broad set of 5 additional tissues: kidney, liver, heart, skin and adipose. We constructed cell type programs and cellular process programs for each of these

tissues. (We did not construct disease progression programs, as these datasets included healthy samples only.)

We analyzed 7 urine biomarker traits (**table S2**) using 24 kidney cell type programs and 12 liver cell type programs (**Figure S1**). Results are reported in **Figure 4A, Figure S5**. We detected a significant enrichment for creatinine level in kidney proximal and connecting tubule cell types, but not in any liver cell types, accurately reflecting the physiology of creatinine secretion by kidney tubule cells(*83*, *84*). On the other hand, we detected a significant enrichment for bilirubin level in liver hepatocytes (driven by *ANGPTL3*; ranked 4), but not in any kidney cell types, consistent with bilirubin absorption by liver hepatocytes(*85*, *86*). Results for cellular process programs are reported in **Figure S6A,B**.

We next analyzed 5 heart-related diseases and traits (**table S2**) using 12 heart cell type programs (**Figure S1**). Results are reported in **Figure 4B, Figure S5**, and **Table 1**. We identified significant enrichments of atrial cardiomyocytes for atrial fibrillation, and pericyte, and smooth muscle cells for blood pressure. Atrial cardiomyocytes are believed to play a role in determining heart rhythm through coordinated electrical activity(*87*) of ion channels critical for atrial fibrillation. The top genes driving this enrichment included the ion channel genes *PKD2L2* (ranked 2), *CASQ2* (ranked 7) and *KCNN2* (ranked 18). On the other hand, pericyte and smooth muscle cells are known to regulate blood pressure by modulating vascular tone(*88*). The top genes driving this enrichment included adrenergic pathway genes *PLCE1* (ranked 1), *CACNA1C* (ranked 21), and *PDE8A* (ranked 23), which release adrenaline to increase blood pressure. Results for cellular process

programs are reported in **Figure S6C**; we observed a significant enrichment of dilated cardiomyopathy in heart for atrial fibrillation.

We next analyzed 2 skin-related diseases and traits (**table S2**) using 13 skin cell type programs (**Figure S1**). Results are reported in **Figure 4C, Figure S5**, and **Table 1**. We identified a significant enrichment of Langerhans cells for eczema. Langerhans cells were previously implicated in inflammatory skin processes related to eczema(*89*), and the top genes driving this enrichment included IL-2 signaling pathway genes (*FCER1G* (ranked 3)*, NR4A2* (ranked 26)*,* and *CD52* (ranked 43)), which have a well-established role in modulating eczema pathogenesis(*90*). Results for cellular process programs are reported in **Figure S6D**; we observed a significant enrichment of BDNF signaling in skin for eczema.

We also analyzed 5 adipose-related diseases and traits (**table S2**) using 13 adipose cell type programs (**Figure S1**). Results are reported in **Figure 4D** and **Figure S5**. We identified a significant enrichment of adipocytes for BMI; this enrichment was driven by genes in the adipogenesis pathway(*91*) (*STAT5A* (ranked 15)*, EBF1* (ranked 29), *LIPE* (ranked 45) and genes in the triglyceride biosynthesis pathway(*91*) (*GPAM* (ranked 14), *LIPE* (ranked 45), both of which contribute to the increase in adipose tissue mass in obesity(*92*, *93*)). Results for cellular process programs are reported in **Figure S6E**.

Cell type programs derived from cells of the same type across tissues were found to be highly correlated (**Figure 4E**). Consequently, enrichments in these correlated cell type programs were also largely consistent, especially for immune cells (**Figure S5**; e.g. T cells in PBMC and lung).

Additionally, we observed many enrichments which were largely unique to one tissue (**Figure S5;** *e.g.*, cardiomyocyte in heart). We observed more significant enrichments using tissue-specific enhancer-gene linking strategies compared to a cross-tissue aggregate linking strategy (**Figure S7**), again highlighting the importance of leveraging the tissue specificity of enhancer-gene strategies.

Linking neurons, microglia, and the complement and apelin signaling pathways to MS and AD progression

Aberrant interactions between neurons and immune cells are thought to contribute to the disease progression of multiple sclerosis (MS) and Alzheimer's disease (AD)(*94*). To focus on risk genes that may contribute to disease progression, rather than those that may be active prior to disease onset, we constructed disease progression and cellular process programs in MS brain(*50*) and AD brain(*95*), by combining disease and healthy tissue samples in each case (**Figure 5A,E, table S1**). In both MS and AD, disease progression programs in each cell type differed substantially from corresponding cell type programs generated using either disease or healthy brain tissue exclusively (**Figure S8, data file S5**). We analyzed MS GWAS data using MS disease progression and MS cellular process programs, and analyzed AD GWAS data using AD disease progression and AD cellular process programs. We considered both brain enhancer-gene links (since MS and AD are neurological diseases, and we analyzed scRNA-seq data from MS brain and AD brain) and immune enhancer-gene links (since MS and AD are immune-related diseases); we observed stronger enrichment results for the immune enhancer-gene links. Furthermore, we confirmed that disease GWAS matched to the corresponding disease progression programs produced the strongest enrichments, although there was substantial cross-disease enrichment (**Figure S9**).

We first analyzed MS GWAS data (**table S2**) using 12 MS disease progression programs. Results using immune enhancer-gene links are reported in **Figure 5B**, (we also report results using brain enhancer-gene links in **Figure S10**). The GABAergic neuron disease progression program (but not the GABAergic neuron healthy cell type program; **Figure S5**) was enriched for MS consistent with the observation that inflammation inhibits GABA transmission in MS(*96*). Additionally, the microglia disease progression program (and the microglia healthy cell type program; **Figure S5**) was enriched for MS, possibly driven by the role of microglia in inflammation and demyelination in MS lesions(*97*, *98*). The top driving genes for the microglia disease progression enrichment included *MERTK* (ranked 2), a regulator of myelin phagocytosis in myeloid cells that can impact lesion and disease evolution in MS(*99*, *100*), and *TREM2* (ranked 4), which is highly expressed in microglia of active demyelinating lesions in MS patients(*101*). Interestingly, we observed a significant increase in the number of microglia and oligodendrocytes and a significant decrease in number of glutamatergic neurons in MS lesions in the brain (**Figure 5C, data file S6**).

We next analyzed 7 MS cellular process programs: 6 shared (healthy and disease) and 1 disease-specific. Results are reported in **Figure 5D.** We detected a significant enrichment for the Layer 2,3 neurons shared cellular process program (in Glutamatergic neurons); the top driving genes included *GAP43* (ranked 16), a marker of axonal regeneration with mutations inhibiting axon regeneration in MS patients(*102*). We also detected a significant enrichment for the complement cascade disease-specific cellular process program (in B cells and microglia), consistent with studies showing that Complement activity is a marker for MS progression(*103–105*); the top

driving genes included *CD37*, *FCRL2* and *FCRL1* (ranked 1, 10, 14): *CD37* activates complement through FC-mediated clustering, driven by *FC* receptors, including *FCRL1* and *FCRL2*(*106*).

We first analyzed AD GWAS data (**table S2**) using 8 AD disease progression programs (**Figure 5E**). Results using immune enhancer-gene links are reported in **Figure 5F and Figure S11** (we also report results using brain enhancer-gene links). Only the microglia disease progression program (and the microglia healthy cell type program; **Figure S5**) was enriched for AD, consistent with the contribution of microglia-mediated inflammation to AD progression(*107*, *108*); the top genes driving enrichment specifically in the disease progression program (but not the healthy cell type program) included *PICALM1, APOC1*, *APOE* and *TREM2* (ranked 1, 2, 3 and 8). *APOE* regulates microglial responses to Alzheimer's related pathologies(*109–111*), *APOC1* is a an *APOE*-dependent suppressor of glial activation(*112*), and *TREM2* modulates microglial morphology and neuroinflammation in Alzheimer's disease pathogenesis models(*113*, *114*). Interestingly, we observed a significant increase in the number of microglia in AD brain (**Figure 5G, data file S6**).

Finally, we analyzed 7 AD cellular process programs: 6 shared (healthy and disease) and 1 disease-specific. Results are reported in **Figure 5H**. We detected a significant enrichment for the apelin signaling pathway disease-specific cellular process program (inter cell type; GABAergic neurons and microglia). Recent studies have implicated this pathway in reducing neuroinflammation in the context of cognitive deficit in animal models of Alzheimer's disease, and suggested that it may be important for prevention and treatment(*115*, *116*). The top genes driving the enrichment included

SORL1 and *SYK* (ranked 2 and 3). *SORL1* expression levels are significantly reduced in Alzheimer's disease patients, and has also been implicated by rare variant analyses(*117–119*).

Linking enterocytes, M cells, and T cells to ulcerative colitis disease progression

The colon is under constant pathogenic exposure, and failure to maintain its epithelial barrier in inflammatory bowel disease and ulcerative colitis (UC) results in chronic inflammation. To explore the biological basis of UC genetic variation, we constructed disease progression and cellular process programs in UC colon, by combining UC inflamed and uninflamed tissue samples (**Figure 6A, table S1**). UC disease progression programs in each cell type differed substantially from corresponding cell type programs generated using either healthy or disease colon tissue exclusively (average Pearson r=0.24; **Figure S8, data file S5**). We analyzed both UC and IBD GWAS data using healthy cell type, UC disease progression and UC cellular process programs. We primarily used colon enhancer-gene links, but also report results using immune enhancer-gene links. As noted above, we confirmed that disease GWAS matched to the corresponding disease progression programs produced the strongest enrichments, (**Figure S9**).

We first analyzed UC and IBD GWAS data (**table S2**) using 20 healthy cell type programs in colon. Results using colon enhancer-gene links are reported in **Figure 6B** and **Figure S5** (we also report results using immune enhancer-gene links in **data file S1**). We detected significant enrichments both in colon immune cell types (as expected; **Figure S5**), and in less appreciated cell types from the colon, including enteroendocrine and endothelial cells (**Figure 6B**). Endothelial cells comprise the gut vascular barrier that controls the passage of immune cells and pathogens between the intestine and the bloodstream. The strong enrichment in endothelial cells is consistent

with their rapid changes in IBD(*120*); the top driving genes included members of the TNF-α signaling pathway (*EFNA1*, *NFKBIA*, *CD40*, ranked 18, 26, 29), a key pathway in IBD(*121*).

We next analyzed IBD and UC GWAS data using 20 UC disease progression programs. Results using colon enhancer-gene links are reported in **Figure 6C, Table 1, Figure S11,** (we also report results using immune enhancer-gene links in **Figure S10**). We detected significant enrichments in enterocyte, M cell, and T cell disease progression programs, but no enrichment in cell type programs corresponding to enterocyte and M cells (**Figure 6B**), potentially implicating processes activated in enterocytes and M cells only after disease activation. Top driving genes for the enterocyte disease progression enrichment included *C1orf106* and *RNF186* (ranked 1, 11), important for maintenance of epithelial barrier function(*122*, *123*); a compromised epithelial barrier is a potential trigger of IBD and a hallmark of the disease(*124*). M cells surveil the lumen for pathogens and play a key role in immune–microbiome homeostasis(*125*). M cells expand dramatically in UC colon(*34*) (they are typically nearly absent in healthy colon; **Figure 6D, data file S6**), and stop-gain and frameshift mutations in *FERMT1*, a top driving gene in the M cell disease progression program (ranked 3), cause Kindler syndrome, a monogenic form of IBD with UC-like gastrointestinal symptoms(*126–128*). For T cells, both the healthy and disease progression cell type programs were significantly enriched (**Figure 6B,C**), despite low correlations between them (**Figure S8**). Top genes driving the T cell disease progression enrichment included IL-7 signaling genes (e.g. *IL2RA*, ranked 24), highlighting the expansion of T regulatory cells in inflamed colon(*34*).

Finally, we analyzed 7 UC cellular process programs: 4 shared (healthy and disease), 1 healthy-specific and 2 disease-specific (**Figure 6E**). Results using colon enhancer-gene links are reported

in **Figure 6F** and **Table 1** (we also report results using immune enhancer-gene links in **data file S3**). The complement cascade shared cellular process program (inter cell type; plasma, B cells, enterocytes and fibroblasts) was enriched for IBD, the MHC-II antigen presentation shared cellular process program (inter cell type; macrophages, monocytes and dendritic cells) was enriched for both UC and IBD (as in our immune cell analysis; **Figure 2F**), and the EGFR1 shared cellular process program (inter cell type; macrophages and enterocytes) was enriched for UC. EGFR1 has been implicated in UC given its role in promoting epithelial cell proliferation and development(*129*), but the underlying mechanism is largely unknown.

Linking immune and connective tissue cell types to asthma, fibrosis and COVID-19 disease progression

The elaborate cellular organization of the lung supports the purification of inhaled air through its branched airways that lead to elastic alveoli that conduct gas exchange. Narrowing of the airways, thickening of the alveolar walls, and loss of lung elasticity are characteristics of chronic lung diseases like asthma and idiopathic pulmonary fibrosis (IPF). Some risk genes may be active prior to disease onset, while others may be involved in disease progression and severity of symptoms after onset. To better understand the biological mechanisms driving disease in the lung, we constructed disease progression and cellular process programs in disease lungs, by combining asthma, IPF, COVID-19 and healthy (lower lung lobes) tissue samples (**Figure 7A,C,F**). Disease progression programs differed substantially from corresponding cell type programs generated using either healthy or disease tissue exclusively (average r=0.15; **Figure S8D-F, data file S5**). We analyzed GWAS data for asthma (and lung capacity, a related trait), IPF, and COVID-19 (both general COVID-19 and severe COVID-19) using the respective disease progression and cellular

process programs. We primarily used lung enhancer-gene links, but also report results using immune enhancer-gene links. As noted above, we confirmed that disease GWAS matched to the corresponding disease progression programs produced the strongest enrichments, (**Figure S9**).

We first analyzed GWAS data for asthma and lung capacity (height-adjusted FEV1adjFVC, a standard metric of lung expiratory volume capacity) (**table S2**) using 19 healthy cell type programs in lower lung lobes(*29*). Results using lung enhancer-gene links are reported in **Figure 7B** (we also report results using immune enhancer-gene links in **data file S1**). For asthma, we detected significant enrichment in T cells, consistent with the contribution of T cell-driven inflammation to the development of asthma pathologies, including airway hyper-responsiveness and tissue remodeling(*130*). The top driving genes included genes in the IL2 signaling pathway (*CAMK4*, *FMNL1*, *RORA*, ranked 1, 5, 9); IL2 is a T cell growth factor that increases airway response to allergens(*131*) and is essential for differentiation of Th2 cells which are known cellular drivers of asthma(*132*). For lung capacity, we detected significant enrichments in stromal cell programs, especially fibroblasts. Fibroblasts produce extracellular matrix (ECM), whose overproduction and pathological alteration contribute to the reduced lung capacity and elasticity characteristic of fibrosis(*133*). The top driving genes included *LOX* (ranked 1), which alters ECM mechanical properties via collagen cross-linking(*134*), and *TGFBR3* (ranked 37) which regulates the pool of available TGFβ. TGFβ is a master regulator of lung fibrosis, impacting many of the altered pathways, including cell proliferation, differentiation, and inflammation. Thus, mutations that change the expression or function of *LOX* or *TGFBR3* may change lung capacity by altering the mechanical properties of ECM and lung fibrotic pathways(*135*, *136*).

We next analyzed asthma and lung capacity GWAS data using 26 asthma disease progression programs (**Fig, 7C**). Results using lung enhancer-gene links are reported in **Figure 7D** and **Table 1** (we also report results using immune enhancer-gene links in **Figure S10C**). For asthma, we detected a significant enrichment in the T cell disease progression program, consistent with the role of T cells in airway hyper responsiveness and tissue remodeling in asthma(*137*). The top genes driving this enrichment again included genes in the interleukin-2 (IL2) signaling pathway (*FMNL1*, *RORA*, *GPR183* and *CD52*, ranked 1, 2, 3, 5; only partially overlapping (34%) with top driving genes for the enrichment in the healthy T cell program). For lung capacity, we detected significant enrichments in basal cell and fibroblast disease progression programs. Thus, genetic variants impacting genes expressed in basal cells impact a phenotype (lung capacity) that is important in asthma patients, but are not related to asthma risk; interestingly, we observed a significant increase in the number of basal cells in asthma vs. healthy lungs (**Figure 7E**).

We next analyzed asthma and lung capacity GWAS data using 5 asthma cellular process programs: 3 shared (healthy and disease), 1 healthy-specific and 1 disease-specific (**Figure 7F**). Results using lung enhancer-gene links are reported in **Figure 7G** and **Table 1**. For asthma, we detected significant enrichments in the macrophage-neutrophil transition shared cellular process program (inter cell type; macrophages and neutrophils). The interaction between macrophages and neutrophils is known to promote IL1-β maturation, which plays a key role in changing airway smooth muscle responsiveness in asthma(*138*, *139*). The top genes driving this enrichment included *CCL20* and *IL6* (ranked 1 and 2). *CCL20* is a macrophage inflammatory chemokine expressed by neutrophils, and plays a key role in initiation and maintenance of airway immune responses underlying asthma(*140*, *141*). *IL6R* is a known regulator of the transition from neutrophil

to monocyte recruitment during inflammation, and has been implicated as a prominent biomarker for asthma(*142*, *143*). For lung capacity, we detected a significant enrichment in the MAPK signaling pathway shared cellular process program (inter cell type: basal, club, fibroblast and endothelial). The top driving genes included *FOXA3* and *PDE2A* (ranked 1 and 2). Knockdown of *PDE2A* in mouse lung has been associated with alveolar inflammation(*144*), and *FOXA3* plays a key role in allergic airway inflammation(*145*).

Finally, we analyzed IPF and COVID-19 GWAS data (**table S2**) using the respective disease progression and cellular process programs defined using scRNA-seq from bronchoalveolar lavages. Results are reported in **data file S2,3**. For IPF, a disease characterized by mucociliary dysfunction(*146*), the mucous disease progression program was most enriched, and nominally significant (p = 0.04) (but not FDR significant). The top genes driving the enrichment included *DSP* (ranked 1), a cell-cell adhesion molecule critical in stabilizing tissue architecture and overexpressed in IPF lung(*147*), and *MUC5B* (ranked 2), in which mutations may lead to impaired mucous and ciliary function, retention of foreign particles and lung injury(*148*, *149*). For COVID-19(*150*), the macrophage disease progression program was enriched, and nominally significant (p = 0.01 for severe COVID-19) (but not FDR significant). The top driving genes included *OAS3* and *OAS1* (ranked 1, 3), which are key antiviral enzyme activators(*151*, *152*), and *CCR5,* a chemokine receptor in which therapeutic intervention has been associated with improved prognosis in severe COVID-19 patients, including decreased inflammatory cytokines and reduced SARS-COV2 RNA in plasma(*153*). Further analyses of a meta-atlas of COVID-19 scRNA-seq from lung, liver, kidney and heart autopsy tissue in conjunction with COVID-19 GWAS data are described elsewhere(*154*).

Our nominally significant findings should be interpreted cautiously, but the analyses will become more powerful as IPF and COVID-19 GWAS sample sizes grow.

DISCUSSION

Prior 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 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. To this end, we integrated scRNA-seq, epigenomics and GWAS summary statistic data sets to infer the underlying cell types and cellular processes through which genetic variants influence disease. To ensure a broad context, we analyzed 11 tissues and 6 disease conditions, constructing gene programs reflecting cell types, disease progression in individual cell types, and cellular processes within and across cell types. We identified disease-critical cell types and cellular processes across the human body in both healthy and disease contexts, including both well-established cell type-trait pairs and new links, such as those between B cells and UC, microglia disease progression and the apelin cellular process and Alzheimer's disease, and the complement cascade cellular process and MS.

Our work has several limitations. First, the enhancer-gene linking strategies from Roadmap and Activity-By-Contact (ABC) models are limited in the tissues and cell types/states represented. More fine-grained enhancer-gene linking strategies will likely prove beneficial, but the strategies that we used here provide a clear improvement over standard gene window-based approaches. We did not perform a comprehensive evaluation of enhancer-gene linking strategies and methods to

combine them, which will be provided elsewhere(*155*, *156*) (S. Gazal, unpublished data). Second, 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. Third, although all studies considered in this work profiled large numbers of cells (up to 300,000 in some tissues), some rare cell types and processes may not yet be adequately sampled due to the number of cells or their tissue distribution(*157*, *158*), or may only be apparent in a disease context, as we observe for disease progression for rare M cells in UC. Fourth, we have focused on human scRNA-seq data(*33*); however, incorporating data from animal models, as discussed in prior work(*36*), would allow experimental validation of disease mechanisms in model organisms. Fifth, the disease progression cell type programs that we link to disease may not be causal for disease, as they may reflect disease-induced changes(*159*, *160*). However, our findings clearly validate the relevance of these gene programs to disease as observed in disease progression M cells and UC(*34*).

Looking forward, the gene program-disease links identified by our analyses can be used to guide downstream studies, including designing systematic perturbation experiments(*161*, *162*) in cell and animal models(*163*) for functional follow up. Additionally, the continued growth of the Human Cell Atlas(*164*) as a reference as well as cell atlases across many diseases, will help to more specifically map the cellular mechanisms underlying a disease across varying contexts. Integrating single cell profiles from different modalities, including scATAC-seq(*165*), SHARE-Seq(*166*), CITE-seq(*167*), and Perturb-seq(*161*) will greatly aide our ability to construct modular gene programs to capture interactions between cell types and map genetic interactions. In the long term, with the increasing success of PheWAS and the integration of multi modal single cell

resolution epigenomics, this framework will continue to be useful in identifying the biological mechanisms driving a broad range of diseases.

Online methods

scRNA-seq data pre-processing

All scRNA-seq datasets in this study(*25–30*, *34*, *46–52*) are publicly available cell by gene expression matrices that are aligned to the hg38 human transcriptome (**table S1**). 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, and standardized the normalization approach across all datasets to account for differences in sequencing depth across cells by normalizing by the total number of UMIs per cell, converting to transcripts-per-10,000 (TP10K) and taking the log of the result to obtain log(10,000*UMIs/total UMIs + 1) "$\log_2$(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(*41*) *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 principle 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(*168*) 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(*169*) with resolution 1. For each dataset, individual single-cell profiles were visualized using the Uniform Manifold Approximation and Projection (UMAP)(*170*). If prior annotations were available they are used as a reference to annotate each cell in each dataset. If prior 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 non-parametric Wilcoxon rank sum test for differential expression (DE) between each cell type *vs*. 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 follow a χ_2^2 distribution, and these transformed values to a grade between 0 and 1 using the min max normalization g = (X – min (X))/(max(X) – min(X)) resulting in a relative weighting of genes in each program.

Disease progression cell type gene programs

We constructed disease progression cell type programs for each cell type observed in both healthy and matching disease tissue. For each cell type, we computed a gene-level non-parametric Wilcoxon rank sum DE test between cells from healthy and 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 differential expression between viral infected and uninfected cells of the same cell type in COVID-19 disease

individuals. We observed low correlation between healthy cell type gene programs and disease progression cell type programs (see **Figure S8** and **data file S5**).

Cellular process gene programs

Using latent factors derived from non-negative matrix factorization (NMF)(*43*) (see below), we define 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). We then annotated each factor (program) by the pathway most enriched in the top driving genes for the factor, and labeled each as an 'intra-cell type' or 'inter-cell type' latent factor if the factor was highly correlated with only one or multiple cell type programs, respectively.

We constructed cellular process programs using an unsupervised approach, by applying non-negative matrix factorization (NMF)(*43*) to the scRNA-seq cells-by-genes matrix. The solution to this formulation can be identified by solving the following minimization problem:

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

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 to be the number of annotated cell types in the dataset plus 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 above 0.8 are collapsed to only consider a single latent factor.

Cellular process gene programs constructed from healthy and disease tissues

For scRNA-seq from healthy and disease tissue contexts, we propose a modified NMF approach to construct gene programs that are either shared across both tissues, specific to healthy tissue or specific to disease tissue. Let H_{P*N_1} be the observed gene expression data for a tissue T from a healthy individual and D_{P*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 a non-negative matrix factorization for H and D as follows

$$H_{\mathrm{PX}N_1} \approx \left[L^{CH}_{\mathrm{PX}K_C} L^{UH}_{\mathrm{PX}K_H}\right] F^{H}_{(K_C+K_H)\mathrm{X}N_1} L^{CH}, L^{UH}, F^{H} \quad (2)$$

$$D_{\mathrm{PX}N_2} \approx \left[L^{CD}_{\mathrm{PX}K_C} L^{UD}_{\mathrm{PX}K_D}\right] F^{D}_{(K_C+K_D)\mathrm{X}N_2} L^{CD}, L^{UD}, F^{D} \quad (3)$$

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 like 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. We frame this in the form of the following optimization problem

$$\underset{L^H, L^D, F^H, F^D}{\operatorname{argmin}} \frac{1}{2}\|H - L^H F^H\|_F{}^2 + \frac{1}{2}\|D - L^D F^D\|_F{}^2 + \frac{v}{2}\left(\|L^H\|_F{}^2 + \|L^D\|_F{}^2\right) + \frac{\gamma}{2}\left(\|L^{CH} - L^{CD}\|_F{}^2\right) \quad (4)$$

where γ is a tuning parameter that controls how close L^{CH} is to L^{CD}. 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 as Q:

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

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

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

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

Following the multiplicative update rules of NMF as per Lee and Seung (NIPS 2001), we get the following iterative updates

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

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

$$L_{ij}^H \leftarrow L_{ij}^H \frac{\left(HF^{H^T} + \gamma[L^{CD}0]\right)_{ij}}{\left(L^H F^H F^{H^T} + \mu L^H\right)_{ij}} \tag{11}$$

$$F_{ij}^H \leftarrow F_{ij}^H \frac{\left(L^{H^T} H\right)_{ij}}{\left(L^{H^T} L^H F^H\right)_{ij}} \tag{12}$$

$$F_{ij}^D \leftarrow F_{ij}^D \frac{\left(L^{D^T} D\right)_{ij}}{\left(L^{D^T} L^D F^D\right)_{ij}} \tag{13}$$

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(*171*). Here, we primarily considered an enhancer-gene linking strategy defined by the union of the Roadmap(*21*, *172*) and Activity-By-Contact (ABC)(*22*, *23*) 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(*155*). 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 non-tissue specific Roadmap and ABC strategies (**Figure S7**). Besides this enhancer-gene linking strategy, we also considered a standard 100kb 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 (*e.g.*, 1000 Genomes Project(*171*); see Data Availability). Binary annotations can have value 0 or 1 only; continuous-valued annotations can have any real value; 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(*39*, *40*) (v.2.1) contains 86 functional annotations (see Data Availability), including binary coding, conserved, and regulatory annotations (*e.g.*, promoter, enhancer, histone marks, TFBS) and continuous-valued linkage disequilibrium (LD)-related annotations.

Stratified LD score regression

Stratified LD score regression (S-LDSC) assesses the contribution of a genomic annotation to disease and complex trait heritability(*11*). S-LDSC 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.

$$var(ß_j) = \sum_{c}^{C} a_{jc}\, t_c \tag{14}$$

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:

$$E(X_j^2) = N \sum_c l\,(j,c)\, t_c + 1 \tag{15}$$

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 Enrichment score (Escore), which relies on the enrichment of annotation c (E_c), defined for binary annotations as follows (for binary and probabilistic 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 Enrichment score (E_{score}) is defined as the difference between the enrichment for annotation *c* corresponding to a particular program against a SNP annotation for all protein coding genes with a predicted enhancer-gene link in the relevant tissue. The Escore metric generalizes to probabilistic annotations with values between 0 and 1(*44*). We primarily focus on the p-value for nonzero enrichment score (see below).

The second metric is standardized effect size (τ^*), the proportionate change in per-SNP heritability associated with a one standard deviation increase in the value of the annotation, conditional on other annotations included in the model(*39*).

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

where sd_c is the standard error 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 one standard deviation in the value of an 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)(*173*). We used the p-value for nonzero enrichment score as our primary

metric, because τ^* is often non-significant for small cell-type-specific annotations when conditioned on the baseline-LD model(*174*).

GWAS summary statistics

We analyzed publicly available GWAS summary statistics for 60 unique diseases and traits with genetic correlation less than 0.9. Each trait passed the filter of being well powered enough for heritability studies (z score for observed heritability > 5). We used the summary statistics for SNPs with minor allele count >5 in a 1000 Genomes Project European reference panel(*171*). The lung FEV1FVC trait was corrected for height data. For COVID-19, we analyzed two phenotypes – general COVID-19 (covid *vs*. population, liability scale heritability h^2 = 0.05, se. = 0.01), and severe COVID-19 (hospitalized covid vs population, liability scale heritability h^2 = 0.03, se. = 0.01)(*175*) (meta-analysis round 4, October 20, 2020, https://www.covid19hg.org/).

Identifying genes driving heritability enrichment

For each gene program, we first subset the full gene list to only consider genes with greater than 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.

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, as an increase in the number of cells of one cell types implies 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 R package.

ACKNOWLEDGMENTS

We thank Leslie Gaffney for assistance with preparing figures as well as Chris Smillie, Basak Eraslan, Alok Jaiswal, and the entire Price and Regev groups for helpful scientific discussions. This work was funded through (K.A.J) NIH F32 Fellowship, (A.L.P) NIH grants U01 HG009379, R01 MH101244, R37 MH107649, R01 MH115676 and R01 MH109978, and (A.R.) Klarman Cell Observatory, HHMI, the Manton Foundation and NIH grant 5U24AI118672. K.A.J., K.K.D, A.L.P and A.R designed the study. K.A.J., 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 highlighted biological insights. J.M.E. provided Activity-by-Contact 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. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov. From August 1, 2020, A.R. is an employee of Genentech.

Data and code availability

All gene programs, enhancer-gene linking annotations, supplementary data files and high-resolution figures are available online at https://data.broadinstitute.org/alkesgroup/LDSCORE/Jagadeesh_Dey_sclinker. 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/. This work uses the S-LDSC software (https://github.com/bulik/ldsc) as well as MAGMA v1.08 for *post-hoc* analysis (https://ctg.cncr.nl/software/magma). Code for constructing cell type, disease progression 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. Code for processing gene programs and combining with enhancer-gene links can be found at https://github.com/kkdey/GSSG.

REFERENCES

1. Consortium, S. W. G. of the P. G. *et al.* Biological Insights From 108 Schizophrenia-Associated Genetic Loci. *Nature* **511**, 421 (2014).
2. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5 (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 (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**, (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**, (2013).
10. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *Am. J. Hum. Genet.* **95**, 126 (2014).
11. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (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**, (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 (2019).

15. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082 (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 (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 (2018).

19. Ernst, J. *et al.* Systematic analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43 (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 (2019).

23. Nasser, J. *et al.* Genome-wide maps of enhancer regulation connect risk variants to disease genes. *bioRxiv* 2020.09.01.278093 (2020) doi:10.1101/2020.09.01.278093.

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* (2020) doi:10.1161/CIRCULATIONAHA.119.045401.

26. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single cell RNA sequencing. *bioRxiv* 742320 (2020) doi:10.1101/742320.

27. Kowalczyk, M. S. Census of Immune Cells (Human Cell Atlas). 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**, (2020).

30. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332 (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. Gazal, S. *et al.* Linkage disequilibrium dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421 (2017).

40. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202 (2019).

41. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

42. Shekhar, K. *et al.* Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell* **166**, 1308-1323.e30 (2016).

43. Lee, D. D. & Seung, H. S. Algorithms for non-negative matrix factorization. in *Proceedings of the 13th International Conference on Neural Information Processing Systems* 535–541 (MIT Press, 2000).

44. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041 (2018).

45. 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).

46. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, (2017).

47. Stewart, B. J. *et al.* Spatio-temporal immune zonation of the human kidney. *Science* **365**, 1461 (2019).

48. Muus, C. *et al.* Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells. *bioRxiv* 2020.04.19.049254 (2020) doi:10.1101/2020.04.19.049254.

49. Cheng, J. B. *et al.* Transcriptional Programming of Normal and Inflamed Human Epidermis at Single-Cell Resolution. *Cell Rep.* **25**, 871 (2018).

50. Schirmer, L. *et al.* Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75 (2019).

51. Braga, F. *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, (2019).

52. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).

53. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

54. Huang, H. *et al.* Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).

55. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).

56. 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**, (2015).

57. Hennino, A. *et al.* Skin-Infiltrating CD8+ T Cells Initiate Atopic Dermatitis Lesions. *J. Immunol.* **178**, 5571–5577 (2007).

58. Thériault, P., ElAli, A. & Rivest, S. The dynamics of monocytes and microglia in Alzheimer's disease. *Alzheimers Res. Ther.* **7**, (2015).

59. Deming, Y. *et al.* The MS4A gene cluster is a key modulator of soluble TREM2 and Alzheimer's disease risk. *Sci. Transl. Med.* **11**, (2019).

60. Yeung, M. *et al.* Characterisation of mucosal lymphoid aggregates in ulcerative colitis: immune cell phenotype and TcR-γδ expression. *Gut* **47**, 215–227 (2000).

61. Mouly, E. *et al.* The Ets-1 transcription factor controls the development and function of natural regulatory T cells. *J. Exp. Med.* **207**, 2113 (2010).

62. Mayassi, T. *et al.* Chronic Inflammation Permanently Reshapes Tissue-Resident Immunity in Celiac Disease. *Cell* **176**, 967-981.e19 (2019).

63. Pandey, A. *et al.* Cloning of a receptor subunit required for signaling by thymic stromal lymphopoietin. *Nat. Immunol.* **1**, 59–64 (2000).

64. 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).

65. 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.* **111**, 2067–2074 (2014).

66. Yip, K. H. *et al.* The Nedd4-2/Ndfip1 axis is a negative regulator of IgE-mediated mast cell activation. *Nat. Commun.* **7**, (2016).

67. 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).

68. Efthymiou, A. G. & Goate, A. M. Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. *Mol. Neurodegener.* **12**, (2017).

69. Luscher, B., Shen, Q. & Sahir, N. The GABAergic Deficit Hypothesis of Major Depressive Disorder. *Mol. Psychiatry* **16**, 383–406 (2011).

70. 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**, (2018).

71. Li, L. *et al.* Disruption of TCF4 regulatory networks leads to abnormal cortical development and mental disabilities. *Mol. Psychiatry* **24**, (2019).

72. Minelli, A. & Scassellati, C. PCLO gene: its role in vulnerability to major depressive disorder - PubMed. *J. Affect. Disord.*

73. Hek, K. *et al.* The PCLO gene and depressive disorders: replication in a population-based study. *Hum. Mol. Genet.* **19**, 731–734 (2010).

74. Mbarek, H. *et al.* Genome-Wide Significance for PCLO as a Gene for Major Depressive Disorder. *Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud.* **20**, (2017).

75. Shi, Y.-F. & Yu, Y.-Q. [The roles of glutamate in sleep and wakefulness]. *Zhejiang Xue Xue Bao Yi Xue Ban J. Zhejiang Univ. Med. Sci.* **42**, 583–590 (2013).

76. McEntee, W. J. & Crook, T. H. Glutamate: its role in learning, memory, and the aging brain. *Psychopharmacology (Berl.)* **111**, 391–401 (1993).

77. Audet, J.-N. *et al.* Divergence in problem-solving skills is associated with differential expression of glutamate receptors in wild finches. *Sci. Adv.* **4**, eaao6369 (2018).

78. del Toro, D. *et al.* Regulation of Cerebral Cortex Folding by Controlling Neuronal Migration via FLRT Adhesion Molecules. *Cell* **169**, 621-635.e16 (2017).

79. Jaworski, D. M. *et al.* Sexually dimorphic diet-induced insulin resistance in obese tissue inhibitor of metalloproteinase-2 (TIMP-2)-deficient mice. *Endocrinology* **152**, 1300–1313 (2011).

80. Stradecki, H. M. & Jaworski, D. M. Hyperphagia and leptin resistance in Tissue Inhibitor of Metalloproteinase-2 (TIMP-2) deficient mice. *J. Neuroendocrinol.* **23**, 269–281 (2011).

81. Barbon, A. & Magri, C. RNA Editing and Modifications in Mood Disorders. *Genes* **11**, (2020).

82. Rezin, G. T., Amboni, G., Zugno, A. I., Quevedo, J. & Streck, E. L. Mitochondrial dysfunction and psychiatric disorders. *Neurochem. Res.* **34**, 1021–1029 (2009).

83. Ciarimboli, G. *et al.* Proximal Tubular Secretion of Creatinine by Organic Cation Transporter OCT2 in Cancer Patients. *Clin. Cancer Res.* **18**, 1101 (2012).

84. Zhang, X. *et al.* Tubular secretion of creatinine and kidney function: an observational study. *BMC Nephrol.* **21**, (2020).

85. 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**, (2001).

86. Wang, X., Chowdhury, J. R. & Chowdhury, N. R. Bilirubin metabolism: Applied physiology. *Curr. Paediatr.* **16**, 70–74 (2006).

87. Barth, A. S. & Tomaselli, G. F. Cardiac metabolism and arrhythmias. *Circ. Arrhythm. Electrophysiol.* **2**, 327–335 (2009).

88. Yamazaki, T. & Mukouyama, Y. Tissue Specific Origin, Development, and Pathological Perspectives of Pericytes. *Front. Cardiovasc. Med.* **5**, (2018).

89. Deckers, J., Hammad, H. & Hoste, E. Langerhans Cells: Sensing the Environment in Health and Disease. *Front. Immunol.* **9**, (2018).

90. Hsieh, K. H., Chou, C. C. & Huang, S. F. Interleukin 2 therapy in severe atopic dermatitis. *J. Clin. Immunol.* **11**, 22–28 (1991).

91. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-97 (2016).

92. Attie, A. D. & Scherer, P. E. Adipocyte metabolism and obesity. *J. Lipid Res.* **50**, S395–S399 (2009).

93. Chen, H. C. & Farese, R. V. Inhibition of triglyceride synthesis as a treatment strategy for obesity: lessons from DGAT1-deficient mice. *Arterioscler. Thromb. Vasc. Biol.* **25**, 482–486 (2005).

94. Heneka, M. T. An immune-cell signature marks the brain in Alzheimer's disease. *Nature* **577**, 322–323 (2020).

95. Mathys, H. *et al.* Single-cell transcriptomic analysis of Alzheimer's disease. *Nature* **570**, 332–337 (2019).

96. Rossi, S. *et al.* Inflammation inhibits GABA transmission in multiple sclerosis. *Mult. Scler. Houndmills Basingstoke Engl.* **18**, 1633–1635 (2012).

97. 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. U. S. A.* **95**, 10100–10105 (1998).

98. 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).

99. Healy, L. M. *et al.* MerTK Is a Functional Regulator of Myelin Phagocytosis by Human Myeloid Cells. *J. Immunol.* **196**, 3375–3384 (2016).

100. Healy, L. M. *et al.* MerTK-mediated regulation of myelin phagocytosis by macrophages generated from patients with MS. *Neurol. Neuroimmunol. Neuroinflammation* **4**, (2017).

101. Cignarella, F. *et al.* TREM2 activation on microglia promotes myelin debris clearance and remyelination in a model of multiple sclerosis. *Acta Neuropathol. (Berl.)* **140**, 513–534 (2020).

102. Rot, U., Sandelius, Å., Emeršič, A., Zetterberg, H. & Blennow, K. Cerebrospinal fluid GAP-43 in early multiple sclerosis. *Mult. Scler. J. - Exp. Transl. Clin.* **4**, (2018).

103. Watkins, L. M. *et al.* Complement is activated in progressive multiple sclerosis cortical grey matter lesions. *J. Neuroinflammation* **13**, 161 (2016).

104. Tatomir, A. *et al.* The complement system as a biomarker of disease activity and response to treatment in multiple sclerosis. *Immunol. Res.* **65**, 1103–1109 (2017).

105. Ingram, G., Hakobyan, S., Robertson, N. P. & Morgan, B. P. Complement in multiple sclerosis: its role in disease and potential as a biomarker. *Clin. Exp. Immunol.* **155**, 128–139 (2009).

106. Oostindie, S. C. *et al.* CD20 and CD37 antibodies synergize to activate complement by Fc-mediated clustering. *Haematologica* **104**, 1841–1852 (2019).

107. Mandrekar, S. & Landreth, G. E. Microglia and Inflammation in Alzheimer's Disease. *CNS Neurol. Disord. Drug Targets* **9**, 156–167 (2010).

108. Hemonnot, A.-L., Hua, J., Ulmann, L. & Hirbec, H. Microglia in Alzheimer Disease: Well-Known Targets and New Opportunities. *Front. Aging Neurosci.* **11**, (2019).

109. Yamazaki, Y., Zhao, N., Caulfield, T. R., Liu, C.-C. & Bu, G. Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies. *Nat. Rev. Neurol.* **15**, 501–518 (2019).

110. Safieh, M., Korczyn, A. D. & Michaelson, D. M. ApoE4: an emerging therapeutic target for Alzheimer's disease. *BMC Med.* **17**, 64 (2019).

111. Ulland, T. K. & Colonna, M. TREM2 — a key player in microglial biology and Alzheimer disease. *Nat. Rev. Neurol.* **14**, 667–675 (2018).

112. Cudaback, E. *et al.* Apolipoprotein C-I is an APOE genotype-dependent suppressor of glial activation. *J. Neuroinflammation* **9**, 192 (2012).

113. Zheng, H. *et al.* TREM2 in Alzheimer's Disease: Microglial Survival and Energy Metabolism. *Front. Aging Neurosci.* **10**, (2018).

114. Karanfilian, L., Tosto, M. G. & Malki, K. The role of TREM2 in Alzheimer's disease; evidence from transgenic mouse models. *Neurobiol. Aging* **86**, 39–53 (2020).

115. Masoumi, J. *et al.* Apelin, a promising target for Alzheimer disease prevention and treatment. *Neuropeptides* **70**, 76–86 (2018).

116. Luo, H. *et al.* Apelin-13 Suppresses Neuroinflammation Against Cognitive Deficit in a Streptozotocin-Induced Rat Model of Alzheimer's Disease Through Activation of BDNF-TrkB Signaling Pathway. *Front. Pharmacol.* **10**, (2019).

117. Scherzer, C. R. *et al.* Loss of apolipoprotein E receptor LR11 in Alzheimer disease. *Arch. Neurol.* **61**, 1200–1205 (2004).

118. Sager, K. L. *et al.* Neuronal LR11/sorLA expression is reduced in mild cognitive impairment. *Ann. Neurol.* **62**, 640–647 (2007).

119. Verheijen, J. *et al.* A comprehensive study of the genetic impact of rare variants in SORL1 in European early-onset Alzheimer's disease. *Acta Neuropathol. (Berl.)* **132**, 213–224 (2016).

120. 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. WJG* **17**, 578–593 (2011).

121. Ruder, B., Atreya, R. & Becker, C. Tumour Necrosis Factor Alpha in Intestinal Homeostasis and Gut Related Diseases. *Int. J. Mol. Sci.* **20**, (2019).

122. Fujimoto, K. *et al.* Regulation of intestinal homeostasis by the ulcerative colitis-associated gene RNF186. *Mucosal Immunol.* **10**, 446–459 (2017).

123. Mohanan, V. *et al.* C1orf106 is a colitis risk gene that regulates stability of epithelial adherens junctions. *Science* **359**, 1161 (2018).

124. Martini, E., Krug, S. M., Siegmund, B., Neurath, M. F. & Becker, C. Mend Your Fences: The Epithelial Barrier and its Relationship With Mucosal Immunity in Inflammatory Bowel Disease. *Cell. Mol. Gastroenterol. Hepatol.* **4**, 33–46 (2017).

125. Graham, D. B. & Xavier, R. J. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* **578**, 527–539 (2020).

126. Bianco, A. M., Girardelli, M. & Tommasini, A. Genetics of inflammatory bowel disease from multifactorial to monogenic forms. *World J. Gastroenterol.* **21**, 12296–12310 (2015).

127. Jobard, F. *et al.* Identification of mutations in a new gene encoding a FERM family protein with a pleckstrin homology domain in Kindler syndrome. *Hum. Mol. Genet.* **12**, 925–935 (2003).

128. Siegel, D. H. *et al.* Loss of Kindlin-1, a Human Homolog of the Caenorhabditis elegans Actin–Extracellular-Matrix Linker Protein UNC-112, Causes Kindler Syndrome. *Am. J. Hum. Genet.* **73**, 174–187 (2003).

129. Dubé, P. E. *et al.* Epidermal growth factor receptor inhibits colitis-associated cancer in mice. *J. Clin. Invest.* **122**, 2780–2792 (2012).

130. Ishmael, F. T. The inflammatory response in the pathogenesis of asthma. *J. Am. Osteopath. Assoc.* **111**, S11-17 (2011).

131. Nag, S., Lamkhioued, B. & Renzi, P. M. Interleukin-2-induced increased airway responsiveness and lung Th2 cytokine expression occur after antigen challenge through the leukotriene pathway. *Am. J. Respir. Crit. Care Med.* **165**, 1540–1545 (2002).

132. Hondowicz, B. D. *et al.* Interleukin-2-Dependent Allergen-Specific Tissue-Resident Memory Cells Drive Asthma. *Immunity* **44**, 155–166 (2016).

133. Herrera, J., Henke, C. A. & Bitterman, P. B. Extracellular matrix as a driver of progressive fibrosis. *J. Clin. Invest.* **128**, 45–53 (2018).

134. Cox, T. R. *et al.* LOX-mediated collagen crosslinking is responsible for fibrosis-enhanced metastasis. *Cancer Res.* **73**, 1721–1732 (2013).

135. Aschner, Y. & Downey, G. P. Transforming Growth Factor-β: Master Regulator of the Respiratory System in Health and Disease. *Am. J. Respir. Cell Mol. Biol.* **54**, 647–655 (2016).

136. Meng, X., Nikolic-Paterson, D. J. & Lan, H. Y. TGF-β: the master regulator of fibrosis. *Nat. Rev. Nephrol.* **12**, 325–338 (2016).

137. Lloyd, C. M. & Hawrylowicz, C. M. Regulatory T Cells in Asthma. *Immunity* **31**, 438 (2009).

138. Sadatomo, A. *et al.* Interaction of Neutrophils with Macrophages Promotes IL-1β Maturation and Contributes to Hepatic Ischemia–Reperfusion Injury. 17.

139. Whelan, R. *et al.* Role and regulation of interleukin-1 molecules in pro-asthmatic sensitised airway smooth muscle. *Eur. Respir. J.* **24**, 559–567 (2004).

140. Scapini, P. *et al.* Neutrophils produce biologically active macrophage inflammatory protein-3alpha (MIP-3alpha)/CCL20 and MIP-3beta/CCL19. *Eur. J. Immunol.* **31**, 1981–1988 (2001).

141. Reibman, J., Hsu, Y., Chen, L. C., Bleck, B. & Gordon, T. Airway epithelial cells release MIP-3alpha/CCL20 in response to cytokines and ambient particulate matter. *Am. J. Respir. Cell Mol. Biol.* **28**, 648–654 (2003).

142. Kaplanski, G., Marin, V., Montero-Julian, F., Mantovani, A. & Farnarier, C. IL-6: a regulator of the transition from neutrophil to monocyte recruitment during inflammation. *Trends Immunol.* **24**, 25–29 (2003).

143. Rincon, M. & Irvin, C. G. Role of IL-6 in Asthma and Other Inflammatory Pulmonary Diseases. *Int. J. Biol. Sci.* **8**, 1281–1290 (2012).

144. Rentsendorj, O. *et al.* Knockdown of lung phosphodiesterase 2A attenuates alveolar inflammation and protein leak in a two-hit mouse model of acute lung injury. *Am. J. Physiol. - Lung Cell. Mol. Physiol.* **301**, L161–L170 (2011).

145. Park, S.-W. *et al.* Distinct Roles of FOXA2 and FOXA3 in Allergic Airway Disease and Asthma. *Am. J. Respir. Crit. Care Med.* **180**, 603–610 (2009).

146. Martinez, F. J. *et al.* Idiopathic pulmonary fibrosis. *Nat. Rev. Dis. Primer* **3**, 17074 (2017).

147. Mathai, S. K. *et al.* Desmoplakin Variants Are Associated with Idiopathic Pulmonary Fibrosis. *Am. J. Respir. Crit. Care Med.* **193**, 1151–1160 (2016).

148. Hancock, L. A. *et al.* Muc5b overexpression causes mucociliary dysfunction and enhances lung fibrosis in mice. *Nat. Commun.* **9**, 5363 (2018).

149. Ridley, C. & Thornton, D. J. Mucins: the frontline defence of the lung. *Biochem. Soc. Trans.* **46**, 1099–1106 (2018).

150. 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.* 1–4 (2020) doi:10.1038/s41431-020-0636-6.

151. Sadler, A. J. & Williams, B. R. G. Interferon-inducible antiviral effectors. *Nat. Rev. Immunol.* **8**, 559–568 (2008).

152. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in Covid-19. *medRxiv* 2020.09.24.20200048 (2020) doi:10.1101/2020.09.24.20200048.

153. Patterson, B. K. *et al.* CCR5 Inhibition in Critical COVID-19 Patients Decreases Inflammatory Cytokines, Increases CD8 T-Cells, and Decreases SARS-CoV2 RNA in Plasma by Day 14. *Int. J. Infect. Dis.* (2020) doi:10.1016/j.ijid.2020.10.101.

154. Delorey, T. M. *et al.* A single-cell and spatial atlas of autopsy tissues reveals pathology and cellular targets of SARS-CoV-2. *bioRxiv* 2021.02.25.430130 (2021) doi:10.1101/2021.02.25.430130.

155. Dey, K. K. *et al.* Unique contribution of enhancer-driven and master-regulator genes to autoimmune disease revealed using functionally informed SNP-to-gene linking strategies. *bioRxiv* 2020.09.02.279059 (2020) doi:10.1101/2020.09.02.279059.

156. Dey, K. K. *et al.* Integrative approaches to improve the informativeness of deep learning models for human complex diseases. *bioRxiv* 2020.09.08.288563 (2020) doi:10.1101/2020.09.08.288563.

157. Strober, B. J. *et al.* Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).

158. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

159. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, (2017).

160. Cho, Y. *et al.* Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework. *Nat. Commun.* **11**, (2020).

161. 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).

162. Ursu, O. *et al.* Massively parallel phenotyping of variant impact in cancer with Perturb-seq reveals a shift in the spectrum of cell states induced by somatic mutations. *bioRxiv* 2020.11.16.383307 (2020) doi:10.1101/2020.11.16.383307.

163. Jin, X. *et al.* In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* **370**, (2020).

164. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).

165. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

166. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).

167. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).

168. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

169. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

170. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv180203426 Cs Stat* (2020).

171. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

172. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

173. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).

174. 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).

175. 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.* 1–4 (2020) doi:10.1038/s41431-020-0636-6.

176. Weissbrod, O. *et al.* Functionally-informed fine-mapping and polygenic localization of complex trait heritability. *bioRxiv* 807792 (2020) doi:10.1101/807792.

TABLES

Table 1

Cell type programs

GWAS disease/trait	Tissue (scRNA-seq)	Cell Type	Escore	p(Escore)	q-value	Top genes
Ulcerative colitis	Blood	B cells	3.2	1.50E-05	2.33E-05	REL,GPX1,LSP1
Celiac disease	Blood	T cells	4.5	2.30E-07	7.16E-07	ETS1,CD247,CD28
MDD	Brain	GABAergic	4	1.00E-04	3.39E-04	TCF4,BEND4,TMX2
Insomnia	Brain	Glutamatergic	0.50	2.4E-06	5.2E-06	LIN28B,NMT1,DNPH1
Atrial fibrillation	Heart	Atrial cardiomyocyte	5.6	3.2E-09	2.2E-08	CAV2,PKD2L2,FAM13B
Blood pressure(dia)	Heart	Smooth muscle	3.4	2.9E-06	1.2E-05	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 progression programs

GWAS disease/trait	Tissue (scRNA-seq)	Cell Type	Escore	p (Escore)	q-value	Top genes
Multiple sclerosis	MS Brain	Microglia	11.6	5.70E-06	3.66E-05	PRDX5,RPL5,SKP1,
Alzheimer's disease	AD Brain	Microglia	9.1	7.10E-05	6.82E-04	PICALM, APOE, APOC1
Ulcerative colitis	UC Colon	Enterocytes	2.6	2.70E-07	1.66E-06	RNF186,APEH,DLD
IBD	UC Colon	M cells	2.2	1.07E-04	2.2E-04	UQCR10,FERMT1,PPP1R1B
Asthma	Asthma Lung	T cells	12.8	4.82E-05	3.99E-04	FMNL1,RORA,GPR183

Cellular process programs

GWAS disease/trait	Tissue (scRNA-seq)	Cellular process	Escore	p (Escore)	q-value	Top genes
Eczema	Blood	CD4+ T cells	3.8	1.32E-07	4.83E-07	IL7R,STMN3,NDFIP1
Celiac disease	Blood	Complement cascade	2.8	4.84E-08	1.92E-07	DCC,PDIA5,PPCDC
Alzheimer's disease	Blood	MHC-II antigen processing	4.9	7.11E-0	2.08E-06	MS4A6A,MS4A4A,CD33
BMI	Brain	LAMP5	2.7	6.33E-08	7.01E-07	FLRT1,COL4A2,SBF2
MDD	Brain	SST	3.9	4.37E-05	1,22E-04	TCF4,PCLO,ZNF462
Years of education	Brain	Electron Transport	3.5	4.42E-08	5.49E-07	ATP6V0B,NSF,GPX1
Multiple sclerosis	MS Brain	Complement cascade**	4.9	5.49E-11	9.62E-10	CD37,RGS14,NCF4
Alzheimer's disease	AD Brain	Apelin signaling*	1.5	9.27E-07	6.50E-06	MS4A6A,SORL1,SYK
Ulcerative colitis	UC Colon	EGFR1 pathway*	3.0	8.81E-04	2.14E-03	C1orf106,SLC26A3,NXPE4
Asthma	Asthma Lung	Mac-neutrophil trans.*	6.6	0.002	0.006	CCL20,IL6,GPR183

Table 1. Notable enrichments from analyses of cell type, disease progression and cellular process gene programs. For each notable enrichment, we report the GWAS disease/trait, tissue source for scRNA-seq data, cell type, enrichment score (Escore), 1-sided p-value for positive Escore, and top genes driving the enrichment. Nominally significant enrichments for diseases with limited GWAS sample size are colored in grey. MDD is an abbreviation for major depressive disorder, blood pressure (dia.) is an abbreviation for diastolic

blood pressure, mac-neutrophil trans. is an abbreviation for macrophage-neutrophil transition. * denotes cellular process programs shared across healthy and disease states. ** denotes cellular process programs specific to disease states.

FIGURES

Figure 1.

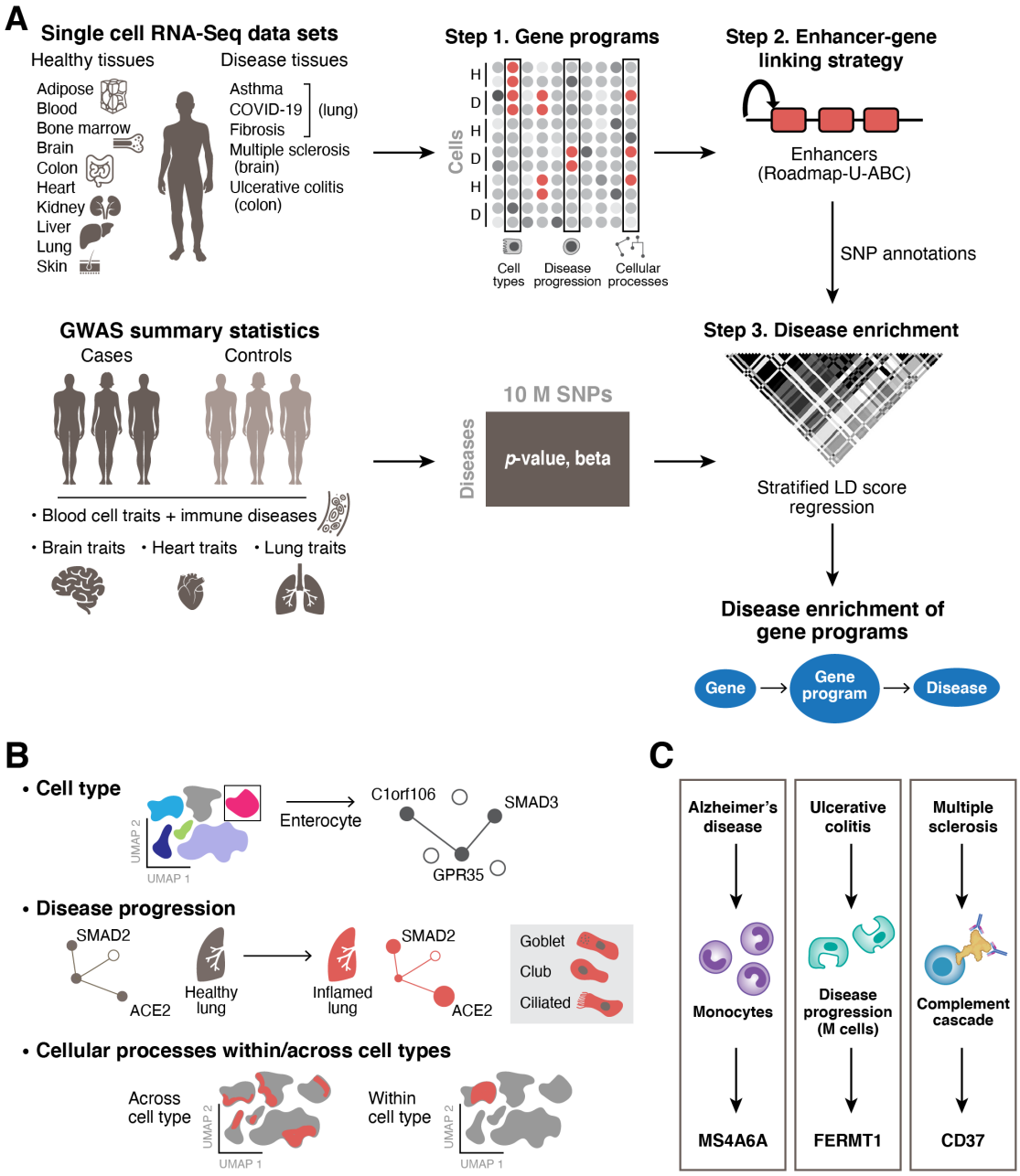

Figure 1. Approach for identifying disease-critical cell types and cellular processes by integration of single-cell profiles and human genetics. a. sc-linker framework. Left: Input. scRNA-seq (top) and GWAS (bottom) data. Middle and right: Step 1: Deriving cell type, disease

progression, and cellular process gene programs from scRNA-seq (top) and associating SNPs with traits from human GWAS (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 *vs*. others. Middle: disease progression programs of genes specifically expressed in cells of the same type in disease *vs*. 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.

Figure 2.

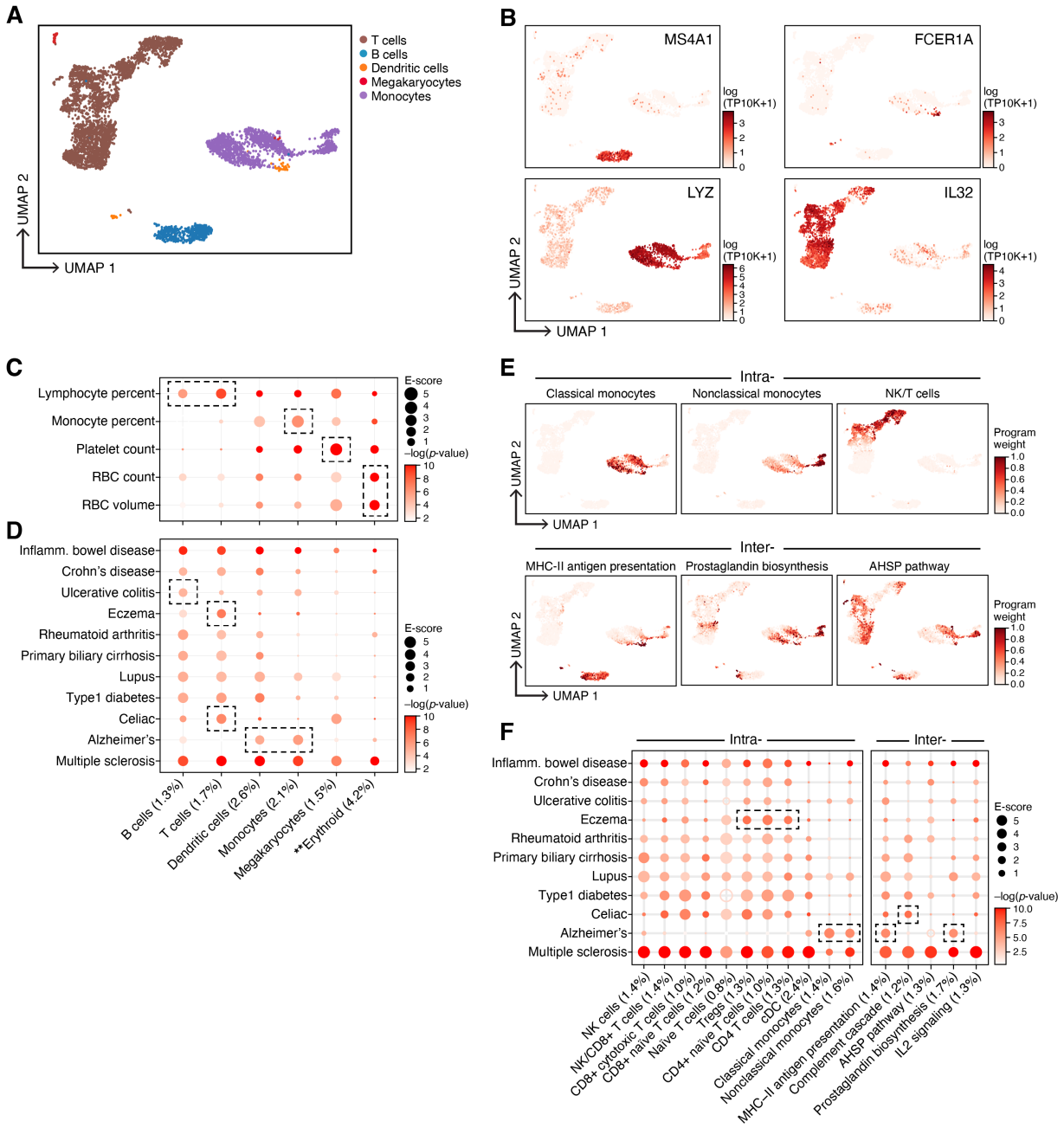

Figure 2. Linking immune cell types and cellular processes to immune-related diseases and blood cell traits. a,b. Immune cell types. Uniform Manifold Approximation and Projection (UMAP) embedding of peripheral blood mononuclear cell (PBMC) scRNA-seq profiles (dots) colored by cell type annotations (a) or expression of cell-type-specific genes (b). **c,d.** Enrichments

of immune cell type programs for blood cell traits and immune-related diseases. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune cell type programs (columns) for blood cell traits (rows, c) or immune-related diseases (rows, d). **e.** Examples of inter- and intra-cell type cellular process programs. UMAP of PBMC (as in a), colored by each program weight (color bar) from non-negative matrix factorization (NMF). **f.** Enrichments of immune cellular process programs for immune-related diseases. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(p-value), dot color) of the heritability enrichment of cellular process programs (columns) for immune-related diseases (rows). In panels c,d,f, the size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. Numerical results are reported in **data file S1,3**. Further details of all diseases and traits analyzed are provided in **table S2**. **Erythroid cells were observed in only bone marrow and cord blood datasets.

Figure 3.

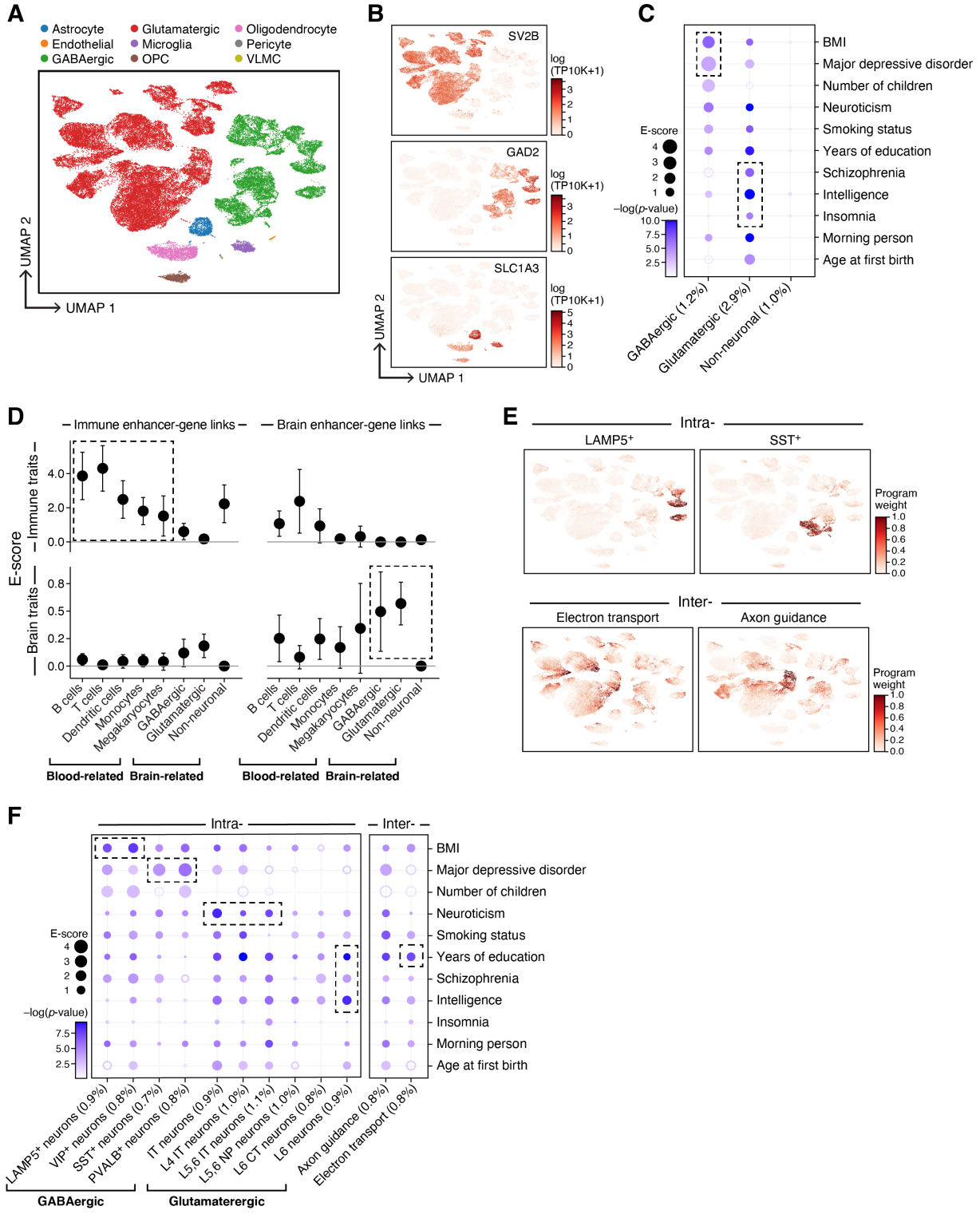
A
Astrocyte
Endothelial
GABAergic
Glutamatergic
Microglia
OPC
Oligodendrocyte
Pericyte
VLMC
UMAP 2
UMAP 1
B
SV2B
GAD2
SLC1A3
log (TP10K+1)
UMAP 2
UMAP 1
C
BMI
Major depressive disorder
Number of children
Neuroticism
Smoking status
Years of education
Schizophrenia
Intelligence
Insomnia
Morning person
Age at first birth
E-score
–log(p-value)
GABAergic (1.2%)
Glutamatergic (2.9%)
Non-neuronal (1.0%)
D
Immune enhancer-gene links
Brain enhancer-gene links
Immune traits
Brain traits
E-score
B cells
T cells
Dendritic cells
Monocytes
Megakaryocytes
GABAergic
Glutamatergic
Non-neuronal
Blood-related
Brain-related
E
Intra-
LAMP5+
SST+
Program weight
Inter-
Electron transport
Axon guidance
F
Intra-
Inter-
BMI
Major depressive disorder
Number of children
Neuroticism
Smoking status
Years of education
Schizophrenia
Intelligence
Insomnia
Morning person
Age at first birth
E-score
–log(p-value)
LAMP5+ neurons (0.9%)
VIP+ neurons (0.8%)
SST+ neurons (0.7%)
PVALB+ neurons (0.8%)
IT neurons (0.9%)
L4 IT neurons (1.0%)
L5,6 IT neurons (1.1%)
L5,6 NP neurons (1.0%)
L6 CT neurons (0.8%)
L6 neurons (0.9%)
Axon guidance (0.8%)
Electron transport (0.8%)
GABAergic
Glutamaterergic

Figure 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-value), dot color) of the heritability enrichment of brain cell type programs (columns) for brain-related diseases and traits (rows). **d.** Comparison of immune vs. brain cell type programs, enhancer-gene linking strategies, and diseases/traits. Magnitude (E-score and SE) of the heritability enrichment of immune vs. brain cell type programs (columns) constructed using immune vs. brain enhancer-gene linking strategies (left and right panels) for immune-related vs. brain-related diseases and traits (top and bottom panels). **e.** Examples of inter- and intra-cell type cellular processes. UMAP (as in a), colored by each program weight (color bar) from non-negative matrix factorization (NMF). **f.** Enrichments of brain cellular process programs for brain-related diseases and traits. Magnitude (E-Score, dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of cellular process programs (columns) for brain-related diseases and traits (rows). In panels c and f, the size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. Numerical results are reported in **data file S1,3**. Further details of all diseases and traits analyzed are provided in **table S2**.

Figure 4.

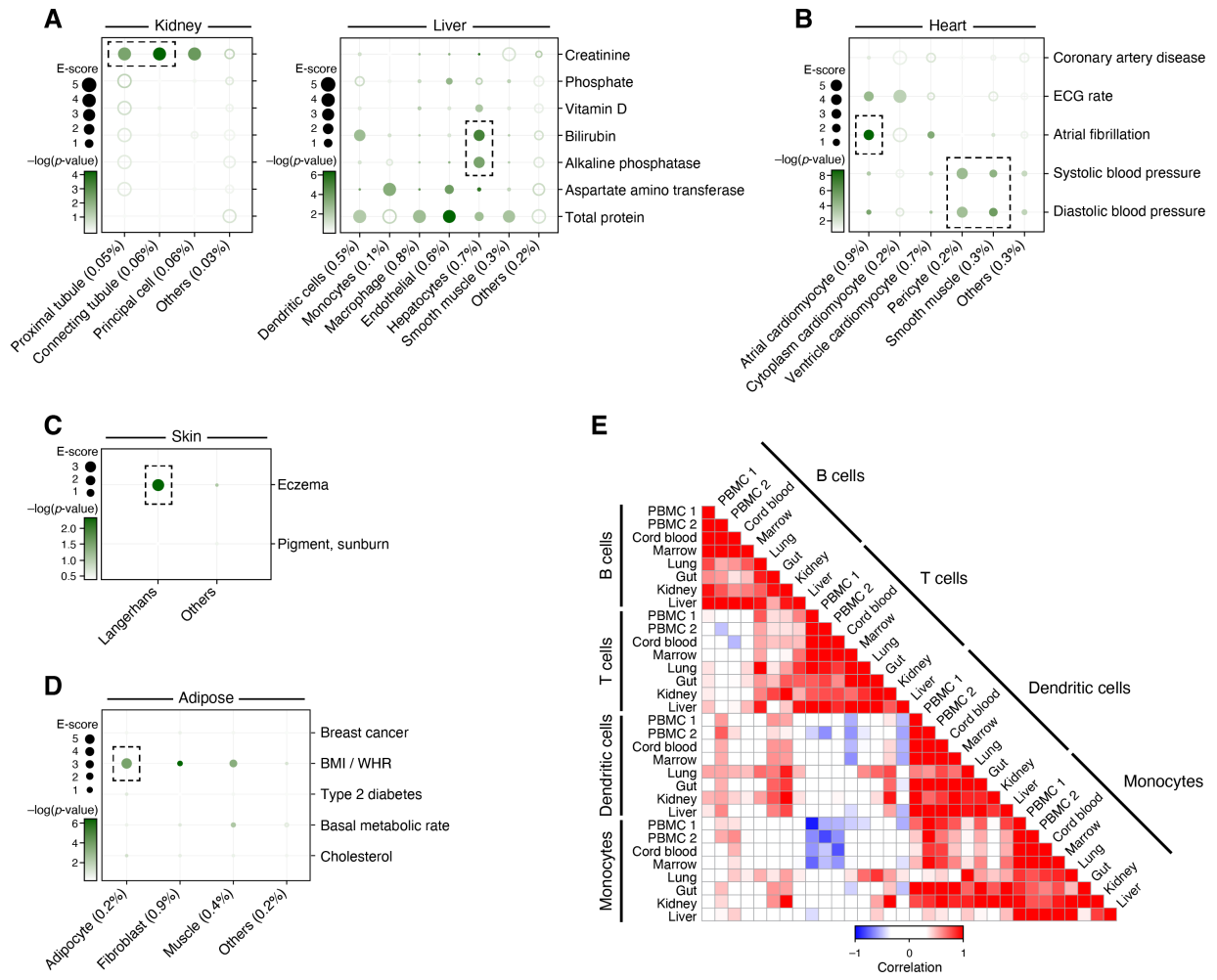

Figure 4. Linking cell types from diverse human tissues to disease

a-d. Enrichments of cell type programs for corresponding diseases and traits. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of cell type programs (columns) for diseases and traits relevant to the corresponding tissue (rows) for kidney and liver (a), heart (b), skin (c) and adipose (d). The size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. **e.** Correlation of immune cell type programs across tissues. We report Pearson correlation coefficients (color bar) of gene-level program memberships for immune cell type programs across different tissues (rows, columns), grouped by cell type (labels).

Numerical results are reported in **data file S1**. Further details of all traits analyzed are provided in **table S2**.

Figure 5.

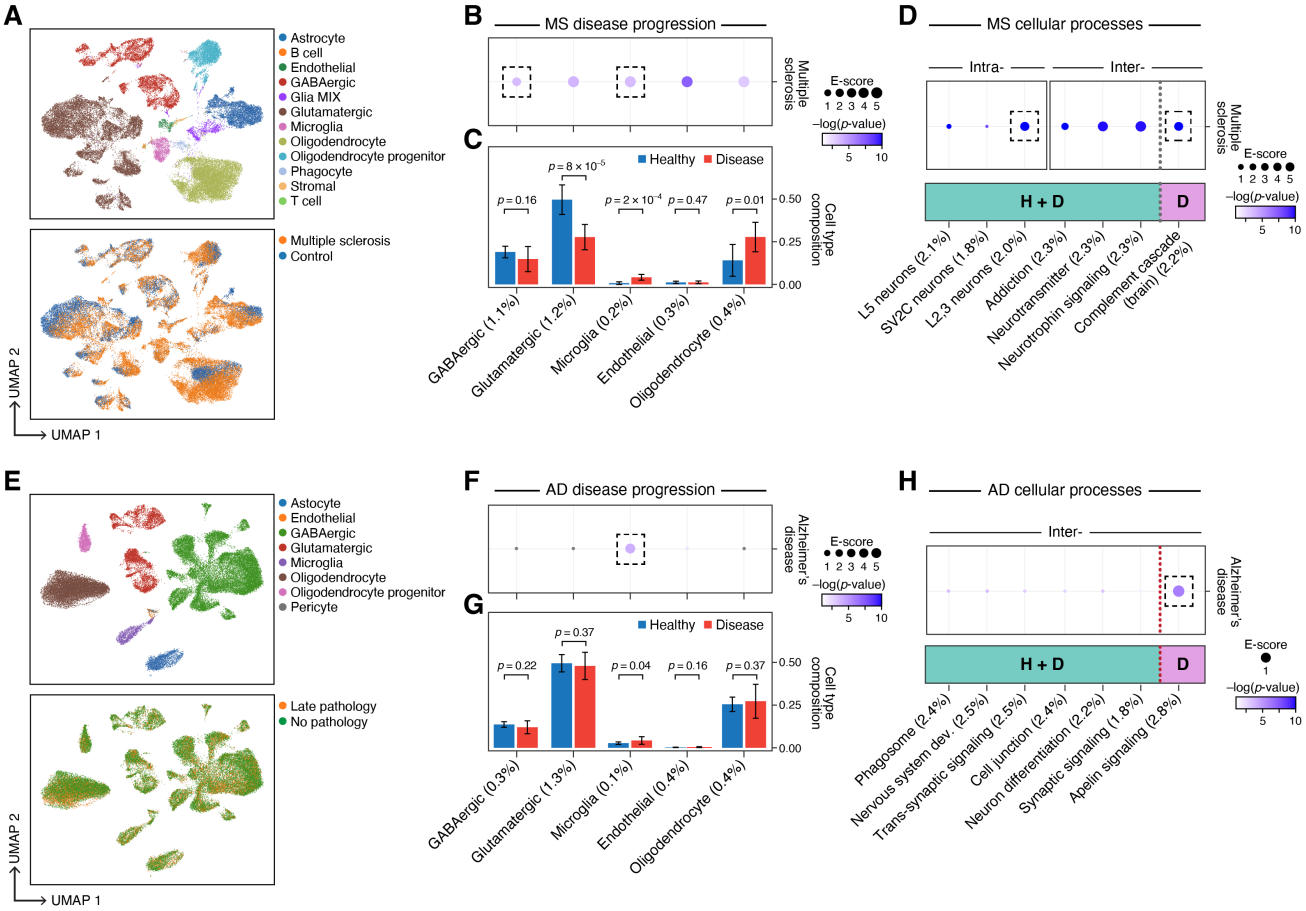

Figure 5. Linking MS and AD disease progression and cellular process programs to MS and AD. a. UMAP embedding of scRNA-seq profiles (dots) from MS and healthy brain tissue, colored by cell type annotations (top) or disease status (bottom). **b.** Enrichments of MS disease progression programs for MS. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of MS disease progression programs (columns), based on the Roadmap∪ABC-immune enhancer-gene linking strategy. **c.** Proportion (mean and SE) of the corresponding cell types (columns) in healthy (blue) and MS (red) brain samples. P-value: Fisher's exact test. **d.** Enrichments of MS cellular process programs for MS. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of intra-cell type (left)

or inter-cell type (right) cellular processes (healthy-specific (H), MS-specific (D) or shared (H+D)) (columns), 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 progression programs for AD. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of AD disease progression programs (columns), based on the Roadmap∪ABC-immune enhancer-gene linking strategy. **g.** Proportion (mean and SE) of the corresponding cell types (columns) in healthy (blue) and AD (red) brain samples. P-value: Fisher's exact test. **h.** Enrichments of AD cellular process programs for AD. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of inter-cell type cellular processes (AD-specific (D) or shared (H+D)) (columns), based on the Roadmap∪ABC-immune enhancer-gene linking strategy. In panels b,c,d,f,g,h, the size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. Numerical results are reported in **data file S2,3**. Further details of all traits analyzed are provided in **table S2**.

Figure 6.

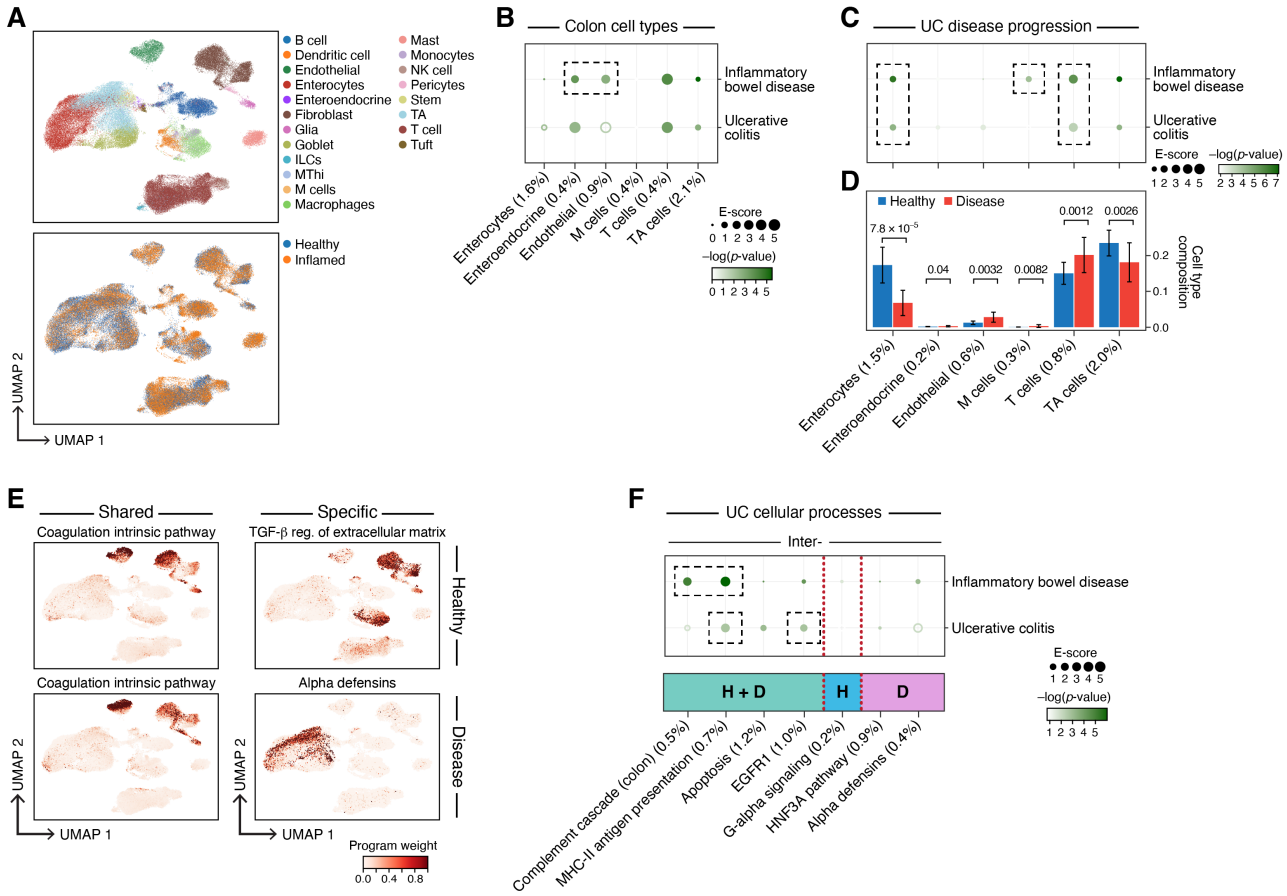

Figure 6. Linking UC disease progression and cellular process programs to UC and IBD. **a.** UMAP embedding of scRNA-seq profiles (dots) from UC and healthy colon tissue, colored by cell type annotations (top) or disease status (bottom). **b.** Enrichments of healthy colon cell types for disease. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of colon cell type programs (columns) for IBD or UC (rows). Results for additional cell types, including immune cell types in colon, are reported in **Figure S5** and **data file S1**. **c.** Enrichments of UC disease progression programs for disease. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of UC disease progression programs (columns) for IBD or UC (rows). **d.** Proportion (mean and SE) of the corresponding cell types (columns) in healthy (blue) and UC (red) colon samples. P-value: Fisher's

exact test. **e.** Examples of shared (healthy and disease), healthy-specific, and disease-specific cellular process programs. UMAP (as in a), colored by each program weight (color bar) from NMF. **f.** Enrichments of UC cellular process programs for disease. Magnitude (E-Score, dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of inter-cell type cellular processes (shared (H+D), healthy-specific (H), or disease-specific (D)) (columns) for IBD or UC (rows). In panels b,c,d,f, the size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. Numerical results are reported in **data file S1,2,3**. Further details of all traits analyzed are provided in **table S2**.

Figure 7.

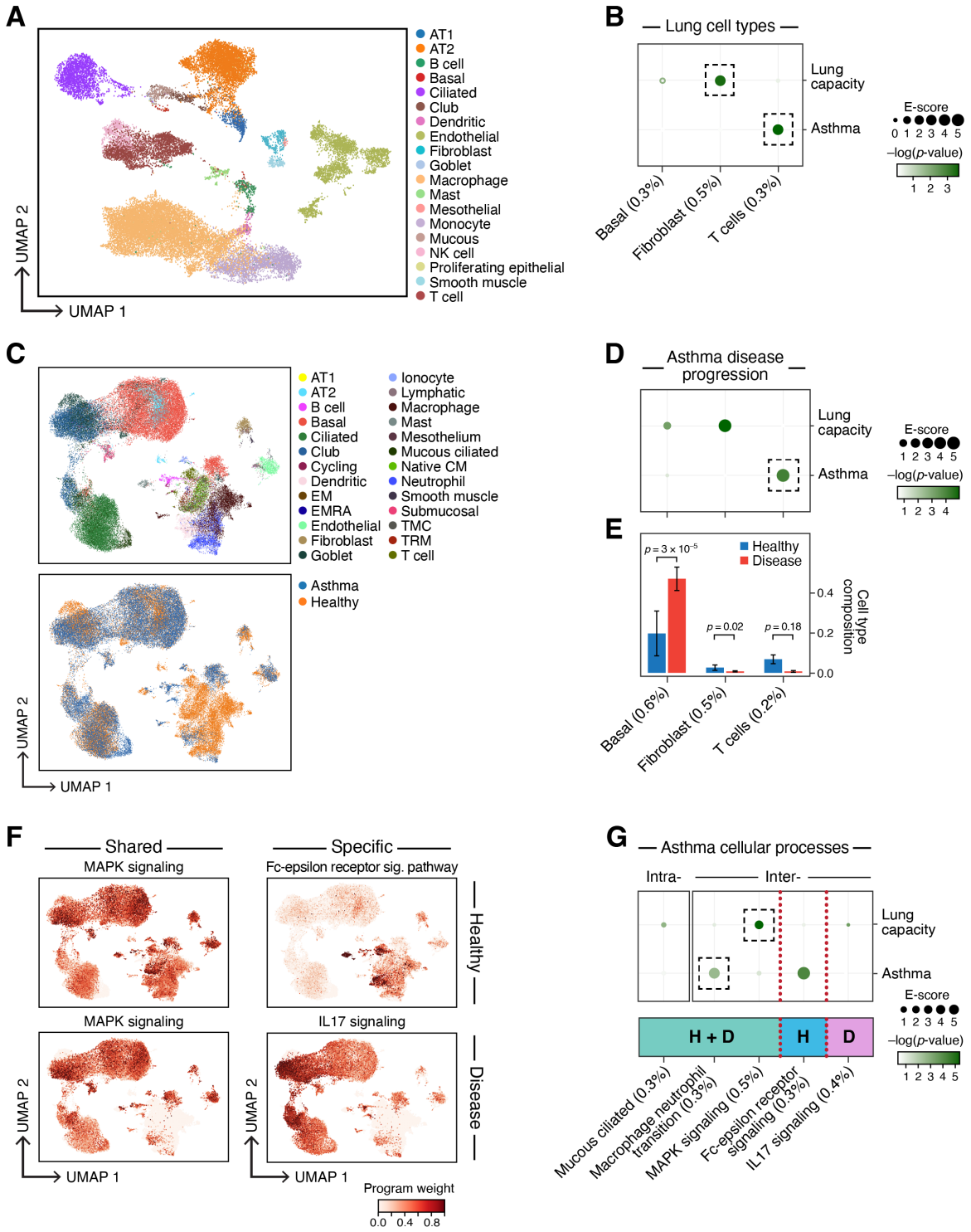

Figure 7. Linking asthma disease progression and cellular process programs to asthma and lung capacity. a. UMAP embedding of healthy lung scRNA-seq profiles (dots) colored by cell

type annotations. **b.** Enrichments of healthy lung cell types for disease. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of healthy lung cell type programs (columns) 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). **d.** Enrichments of asthma disease progression programs for disease. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of asthma disease progression programs (columns) for lung capacity or asthma (rows). **e.**Proportion (mean and SE) of the corresponding cell types (columns), in healthy (blue) and asthma (red) lung samples. P-value: Fisher's exact test. **f.** Examples of shared (healthy and disease), healthy-specific, and disease-specific cellular process programs. UMAP (as in c), 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-value), dot color) of the heritability enrichment of intra-cell type (left) and inter-cell type (right) cellular processes (shared (H+D), healthy-specific (H), or disease-specific (D)) (columns) for lung capacity and asthma GWAS summary statistics (rows). In panels b,d,e,g, the size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. Numerical results are reported in **data file S1,2,3**. Further details of all traits analyzed are provided in **table S2.**

SUPPLEMENTARY TABLES

Supplementary Table 1

Tissue	# of cells	# of individuals	# of cell types
PBMC (Travaglini et al)	4,640	2	6
PBMC (Zheng et al)	68,551	8	6
Cord Blood	263,828	8	6
Bone Marrow	283,894	8	6
Brain	47,509	3	9
Kidney	40,268	13	24
Liver	13,340	4	12
Lung	31,644	10	19
Heart	287,269	7	12
Colon	110,373	12	20
Adipose	11,184	3	13
Skin	71,864	9	13
Colon (healthy + disease)	287,269	20 (healthy), 16 (disease)	20
MS brain (healthy + disease)	48,918	9 (healthy), 12 (disease)	12
Alzheimer's brain (healthy + disease)	70,634	24 (healthy), 24 (disease)	8
Asthma lung (healthy + disease)	67,078	42 (healthy), 12 (disease)	26
Idiopathic pulmonary fibrosis lung (healthy + disease)	114,396	10 (healthy), 20 (disease)	19
COVID-19 BAL (healthy + disease)	43,930	3 (healthy), 6 (disease)	10

Table S1. Description of scRNA-seq datasets analyzed. We report the tissue of origin, number of cells, number of individuals and number of cell type programs analyzed for each single-cell dataset analyzed.

Supplementary Table 2

Trait category	Trait	Source	Sample size (N)
Blood cell traits	Lymphocyte percentage	UK Biobank	444502
	Monocyte percentage	UK Biobank	439938
	Platelet count	UK Biobank	444382
	Red blood cell count	UK Biobank	445174
	Red blood cell volume	UK Biobank	442700
	Eosinophil count	UK Biobank	439938
	Basophil count	UK Biobank	439938
	Neutrophil count	UK Biobank	439938
	Mean corpuscular volume	UK Biobank	442122
Urine biomarkers	Creatinine	UK Biobank	434158
	Vitamin D	UK Biobank	415700
	Bilirubin	UK Biobank	429423
	Alkaline phosphatase	UK Biobank	433862
	Aspartate amino transferase	UK Biobank	430982
	Total protein	UK Biobank	397652
Autoimmune diseases	Inflammatory bowel disease	de Lange et al 2017	59957
	Crohn's disease	de Lange et al 2017	40266
	Ulcerative colitis	de Lange et al 2017	45975
	Eczema	UK Biobank	458699
	Hypothyroidism	UK Biobank	459324
	Rheumatoid Arthritis	Okada et al 2014	37681
	Primary biliary cirrhosis	Cordell et al. 2015	13239
	Lupus	Bentham et al. 2015	14267
	Type 1 diabetes	Bradfield et al. 2011	26890
	All autoimmune traits	UK Biobank	459234
	Celiac disease	Dubois et al. 2010	15283
	Alzheimer's disease	Jansen et al. 2019	450988
	Multiple Sclerosis	Sawcer et al. 2011	27148
Neurological/ Psychiatric	Number of children	UK Biobank	456500
	Anorexia	Boraska et al 2014	32143
	ADHD	Demontis et al 2019	55374
	Autism	PGC cross disorder group	10263
	Sleep duration	Dashti et al 2019	446118
	BMI	UK Biobank	458417
	Major depressive disorder	Wray et al. 2018	173005
	Neuroticism	Nagel et al. 2018	449484
	Smoking status	UK Biobank	457683
	Years of education	UK Biobank	454813
	Intelligence	UK Biobank	117131
	Morning person	UK Biobank	410520
	Insomnia	Jansen et al. 2019	385506
	Schizophrenia	SCZ Working Group 2014	70100
	SCZ v. BD	Ruderfer et al 2018	38855
	Bipolar disorder	PGC bipolar group 2011	16731
	Reaction time	Davies et al 2018	300486
	Age of first birth	Barban et al. 2016	222037

Cardiac related traits	Coronary artery disease	Schunkert et al 2011	77210
	ECG rate	UK Biobank	53777
	Atrial Fibrillation	Nielsen et al. 2018	1030836
	Systolic blood pressure	UK Biobank	422771
	Diastolic blood pressure	UK Biobank	422771
Lung traits	Childhood-Onset-Asthma	Ferreira et al. 2019	314633
	FEV1adjFEVC (lung capacity)	UK Biobank	371949
	Idiopathic Pulmonary Fibrosis	Allen et al. 2020	11259
Other traits	Height	Lango, Allen et al 2010	131547
	Breast Cancer	UK Biobank	459324
	BMI-WHR	UK Biobank	458417
	Type 2 Diabetes	Morris et al 2012	6078
	Basal metabolic rate	UK Biobank	354825
	General risk tolerance	Karlsson Linner et al 2019	466571

Table S2. Diseases and complex traits analyzed. We analyzed 60 diseases and complex traits with genetic correlation <= 0.9 and report the publication and sample size of each study.

SUPPLEMENTARY DATA FILE LEGENDS

Data File S1: Healthy cell type program heritability enrichment results. Numerical values for E-score and significance are reported for all cell type programs and traits analyzed.

Data File S2: Disease progression cell type program heritability enrichment results. Numerical values for E-score and significance are reported for all disease progression cell type programs and traits analyzed.

Data File S3: Cellular process program heritability enrichment results. Numerical values for E-score and significance are reported for all healthy, disease, and shared cellular processes and traits analyzed.

Data File S4: List of genes driving each enrichment. Up to 50 genes with the strongest MAGMA gene score and membership in the gene program.

Data File S5: Correlation between disease progression and healthy cell type program.

Data File S6: Composition of cell types in each tissue. Number of cells of each cell type and condition observed in each single cell dataset.

SUPPLEMENTARY FIGURES

Supplementary Figure 1

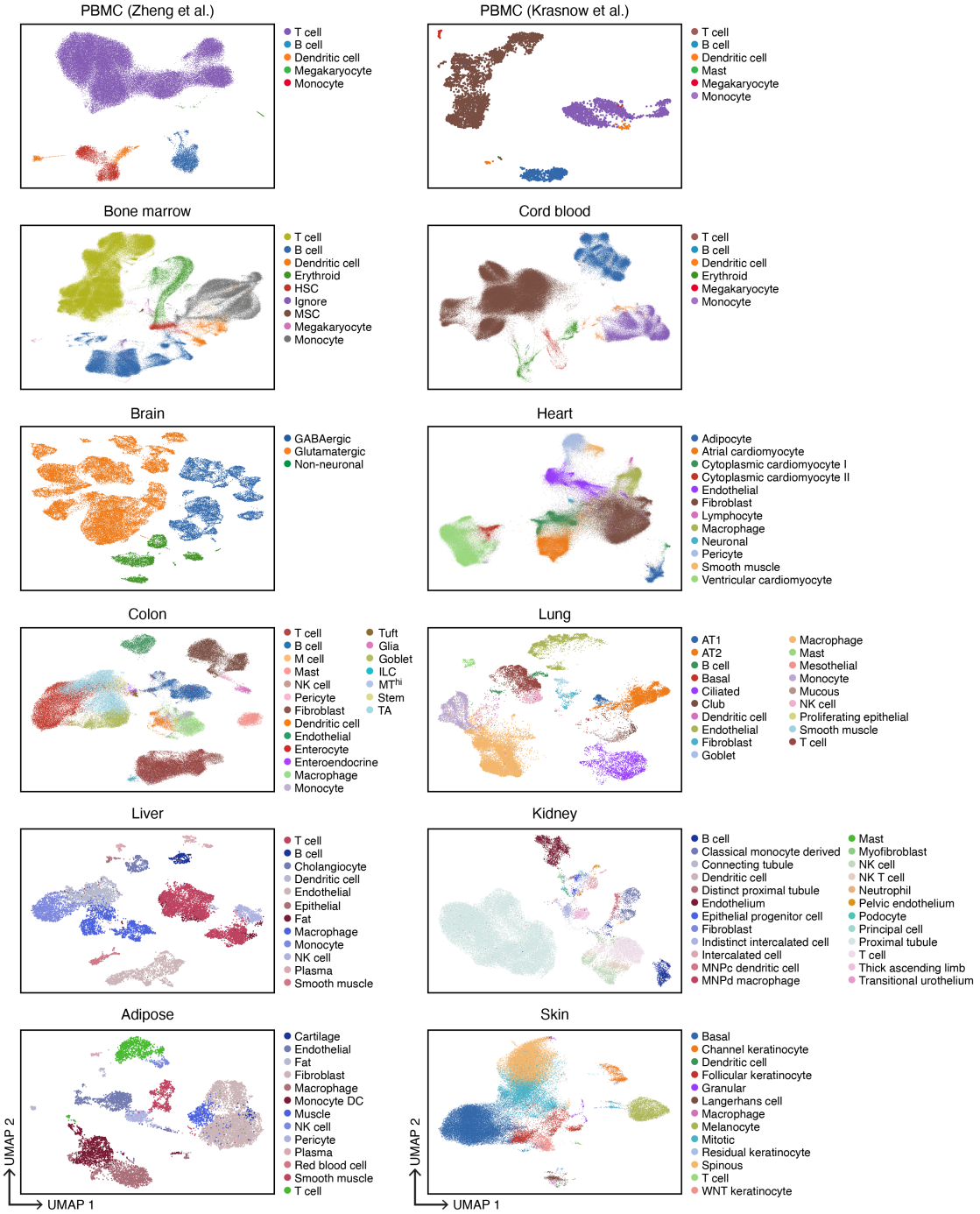

Figure S1. Single-cell RNA-seq datasets. UMAP embedding of scRNA-seq profiles (dots) colored by cell type annotations from 12 datasets (labels on top).

Supplementary Figure 2

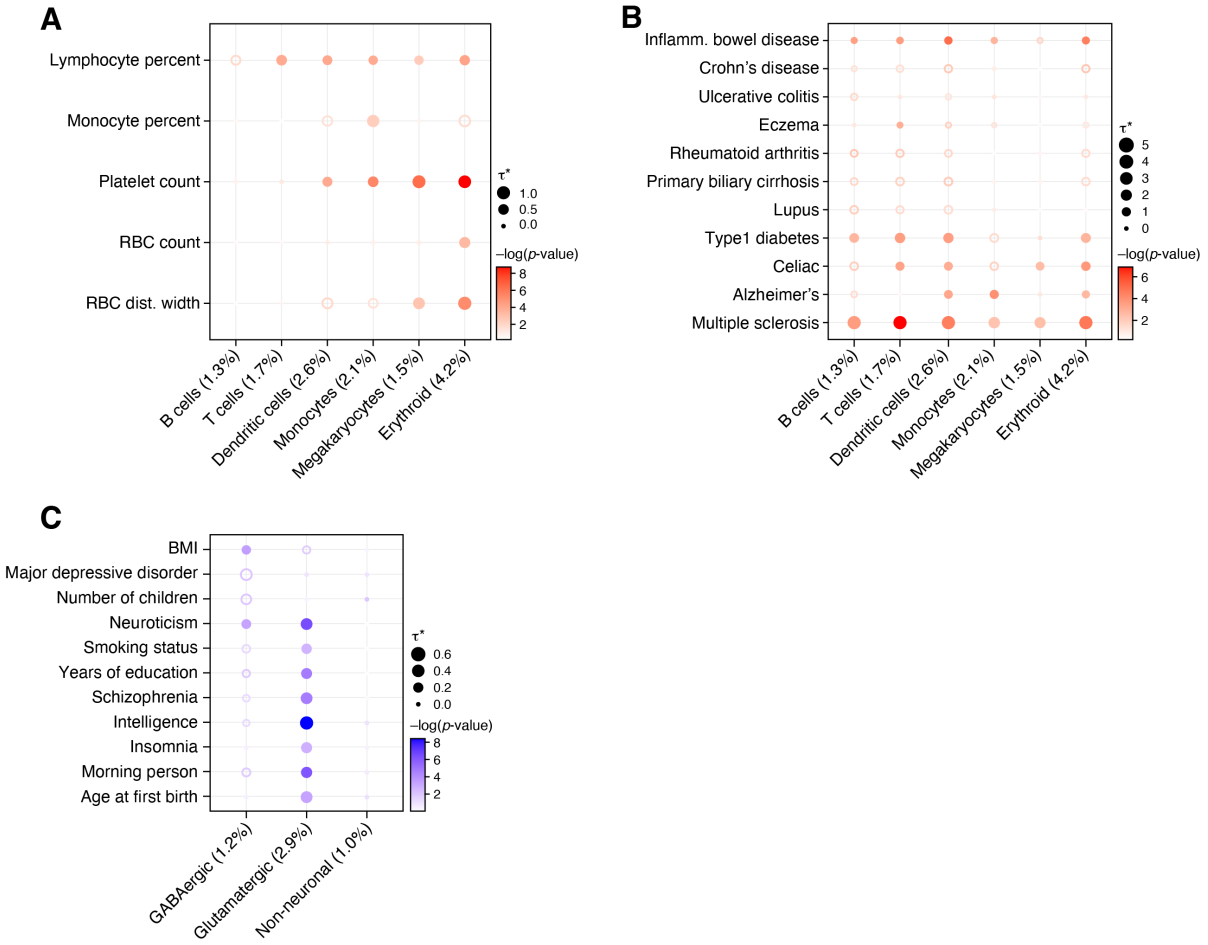

Figure S2. Standardized effect sizes of immune and brain cell type programs. Standardized effect size (τ^*) (dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune (**a,b**) or brain (**c**) cell type programs (columns) for blood cell traits (**a**), immune 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 **data file S1**. Details for all traits analyzed are in **table S2.**

Supplementary Figure 3

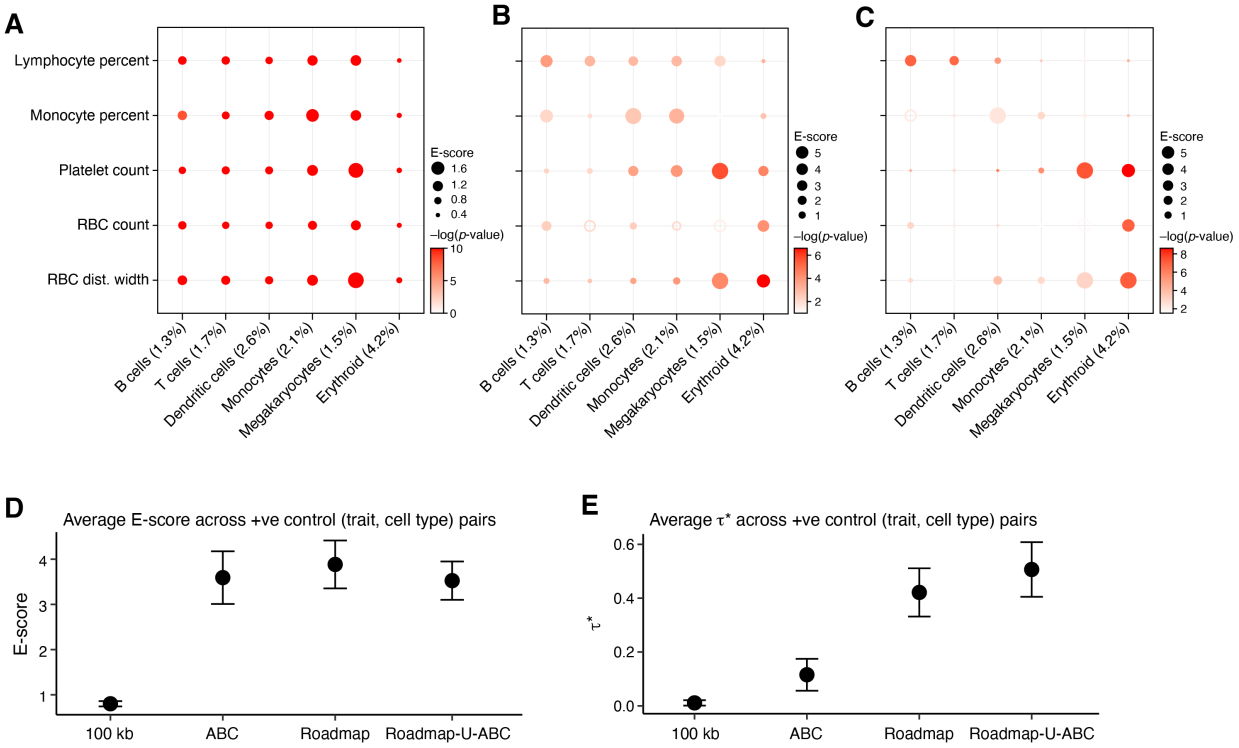

Figure S3. Analysis of immune cell type programs and blood cell traits using different enhancer-gene linking strategies. a-c. Magnitude (E-Score, dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of immune cell type programs (columns) aggregated over 4 scRNA-seq datasets (PBMC (2), cord blood, and bone marrow) for 5 blood cell traits with SNP annotations combined with 100Kb (a), ABC-immune (b) or Roadmap-immune (c) strategies (compare to Roadmap ∪ ABC-immune strategy in **Figure 2b**). **d,e.** Mean E-score (d) or average standardized effect size (τ^*) (e) (y axis) for blood cell traits and immune cell type programs as in **Figure 2b**, with SNP annotations combined with 100Kb, ABC-immune, Roadmap-immune or Roadmap ∪ ABC-immune strategy (x axis). Errors bars: 95% confidence intervals. Details for all traits analyzed are in **table S2.**

Supplementary Figure 4

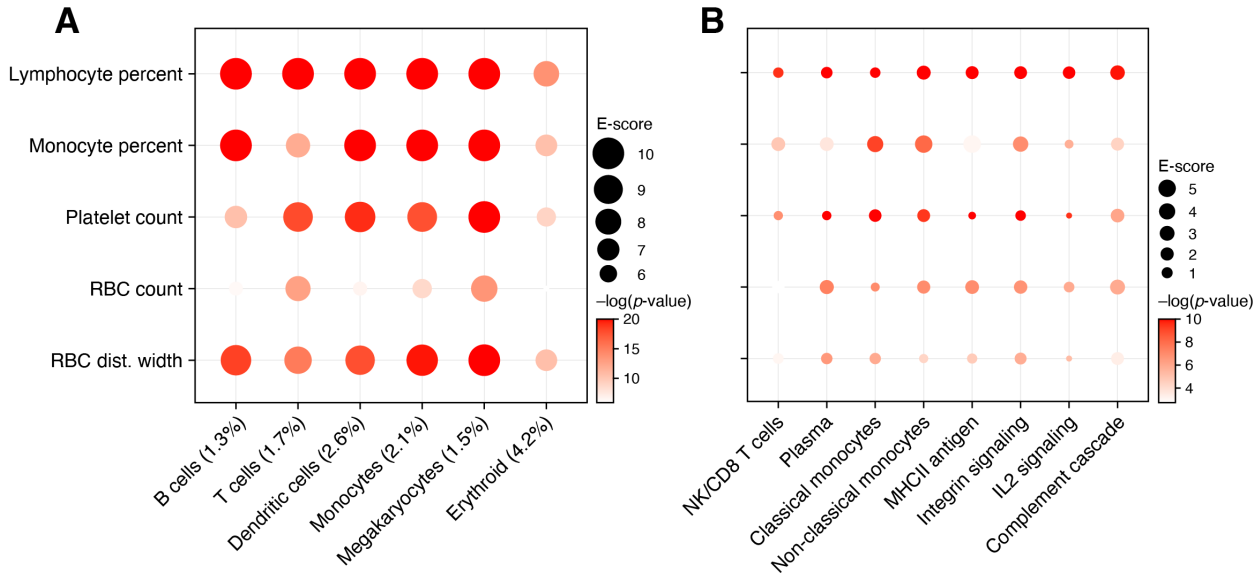

Figure S4. Analysis of functional enrichment of fine-mapped SNPs of immune cell type programs and heritability enrichment of immune cellular process programs. a. Functional enrichment of fine-mapped SNPs of immune cell type programs. Magnitude (Enrichment, dot size) and significance (-$\log_{10}$(P-value), dot color) of SNP annotations corresponding to immune cell type programs (using the Roadmap ∪ ABC-immune enhancer-gene linking strategy) with respect to functionally fine-mapped SNPs (from ref. (*176*)). **b.** Heritability enrichment of cellular process programs for blood cell traits. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune cellular process programs (columns) and blood cell traits (rows). Details for all traits analyzed are in **table S2.**

Supplementary Figure 5

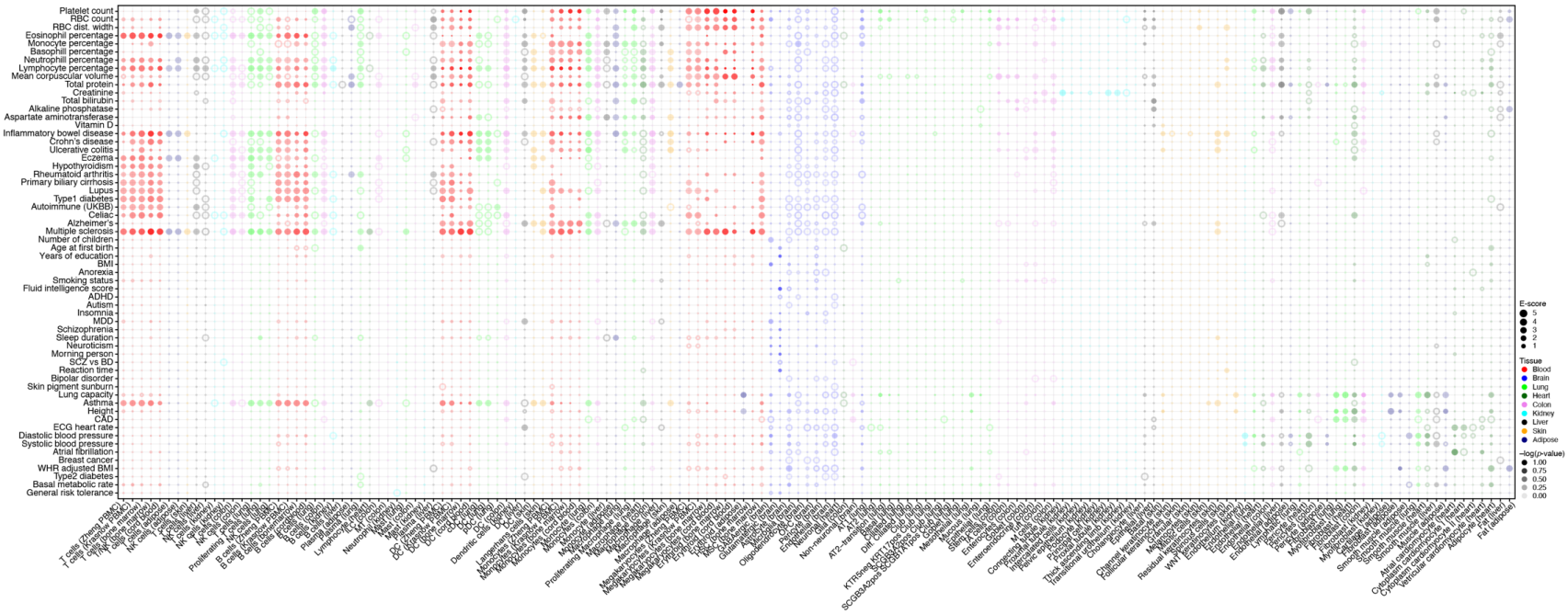

Figure S5. Linking cell type programs to diseases and traits across all analyzed tissues. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of cell type programs (columns) from each of nine tissues (color code, legend) for GWAS summary statistics 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 **table S2.** See **Data Availability** for higher resolution version of this figure.

Supplementary Figure 6

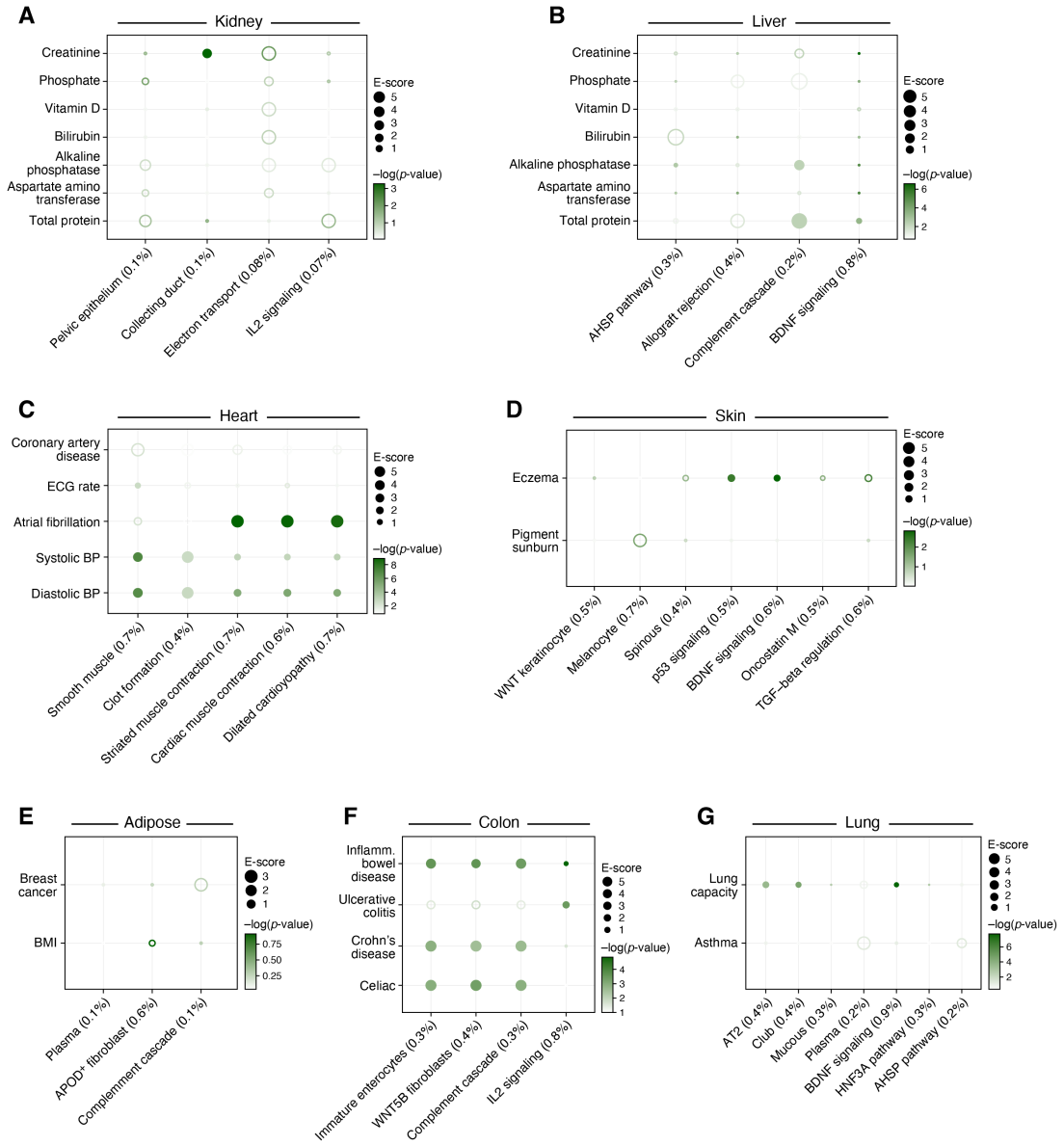

Figure S6. Linking cellular process programs to relevant diseases and traits in each of six tissues. Magnitude (E-Score, dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of cellular process programs (columns; obtained by NMF) in each of seven tissues (label 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 **table S2.**

Supplementary Figure 7

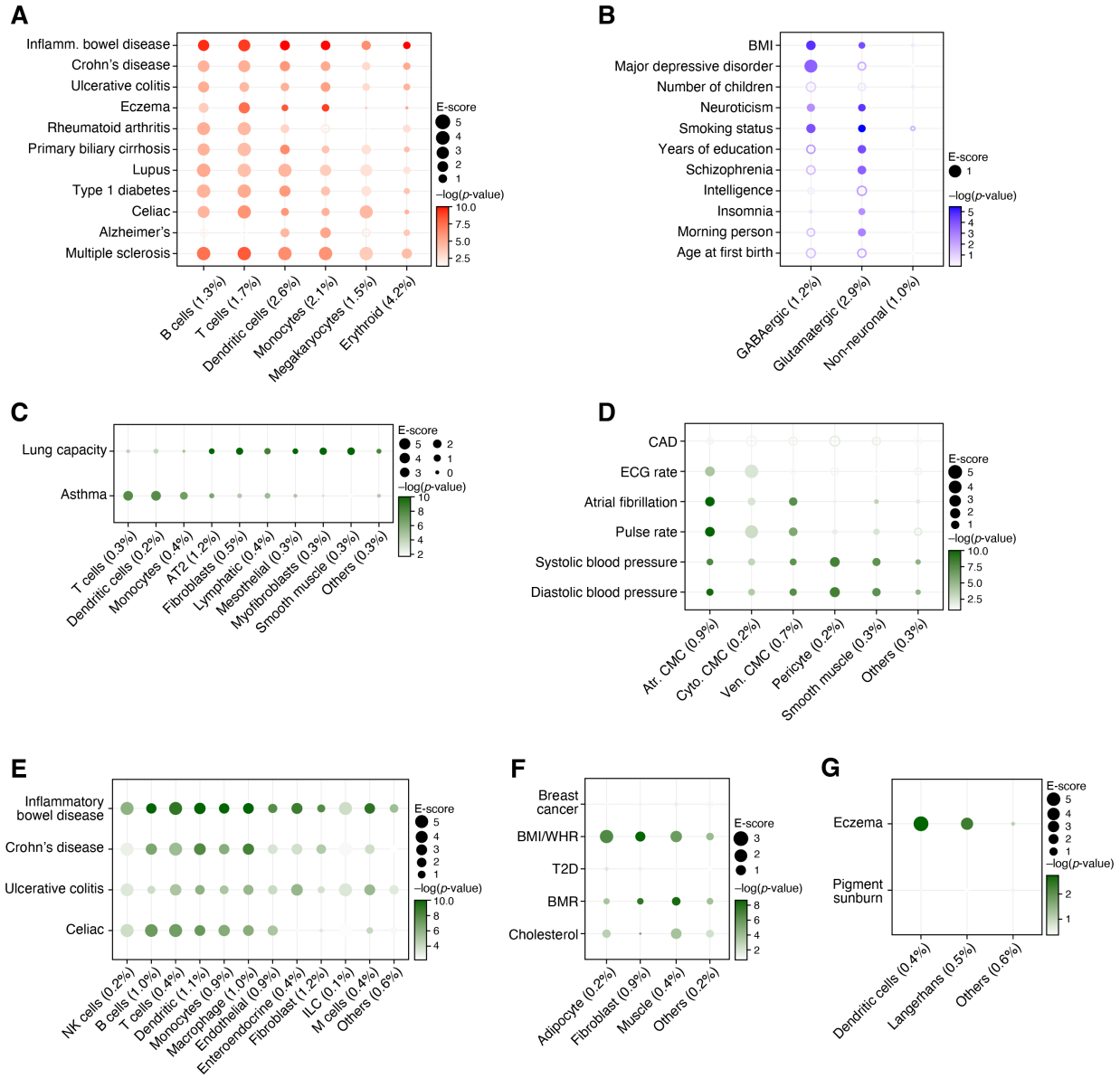

Figure S7. Analysis of cell type programs using a non-tissue-specific enhancer-gene linking strategy. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune (**a**), brain (**b**), lung (**c**), heart (**d**), colon (**e**), adipose (**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 **table S2.**

Supplementary Figure 8

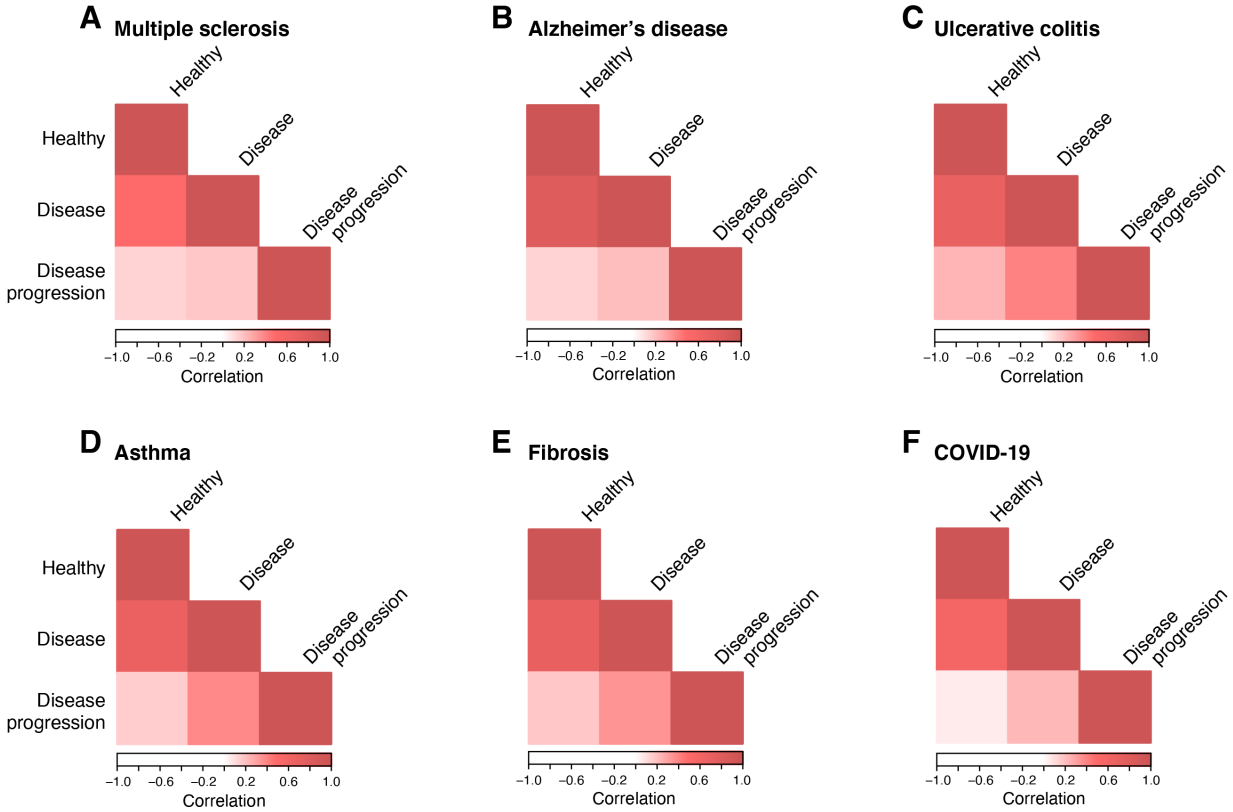

Figure S8. Disease progression programs have low correlations with healthy and disease cell type programs. Pearson correlation coefficient (color bar) of gene program membership vectors between healthy cell type, disease cell type and disease progression programs in scRNA-seq studies from a disease tissue (label on top) and the corresponding healthy tissue.

Supplementary Figure 9

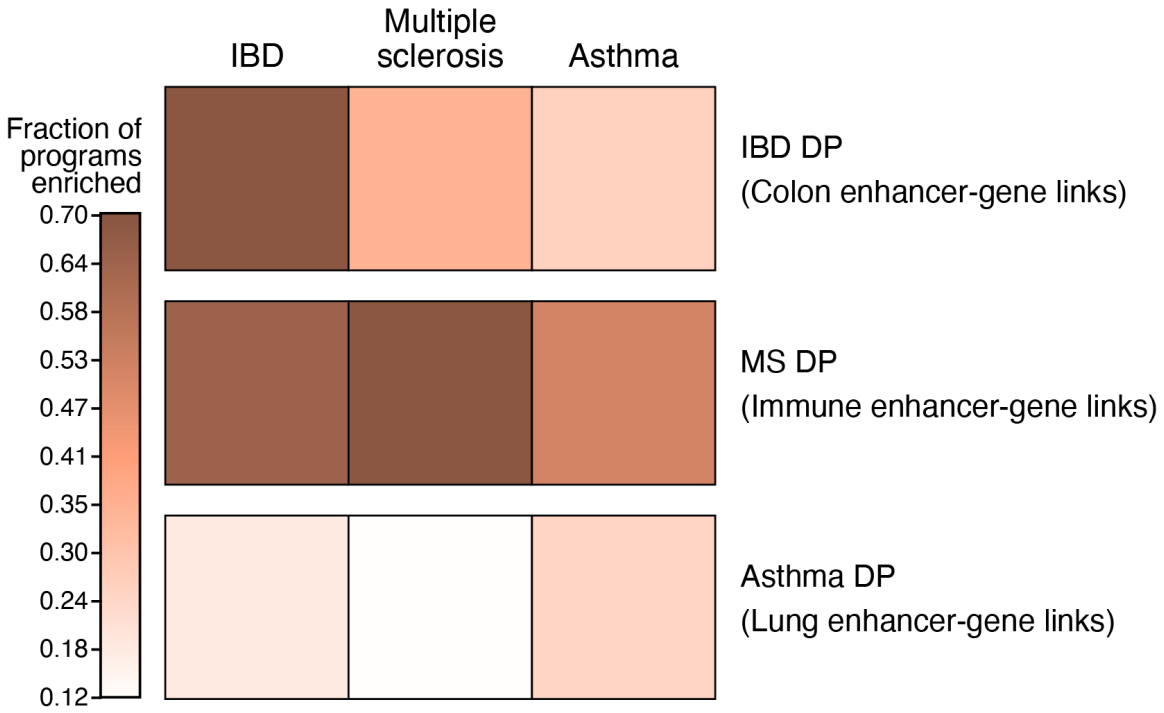

Figure S9. Disease specificity of disease progression programs. Proportion of disease progression programs with a -$\log_{10}$(P-value) of enrichment score (p.Escore) > 3 in IBD, MS and asthma GWAS summary statistics (column) for disease progression cell type programs from IBD, MS and asthma (columns), when combined with Roadmap ∪ ABC-immune strategy (row).

Supplementary Figure 10

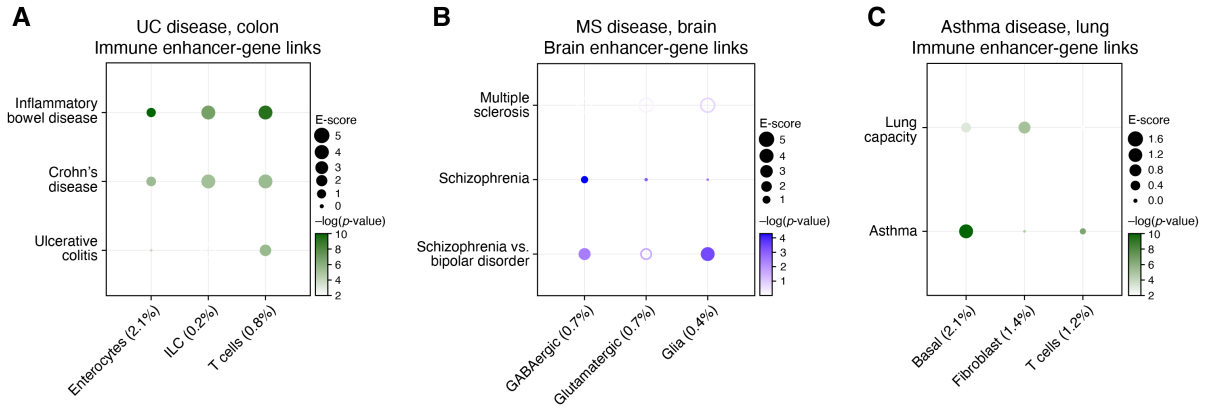

Figure S10. Analysis of disease progression programs using alternative Roadmap∪ABC enhancer-gene linking strategies. Magnitude (E-Score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of disease progression programs (columns) in UC (colon cells) using Roadmap∪ABC-immune (**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 **table S2.**

Supplementary Figure 11

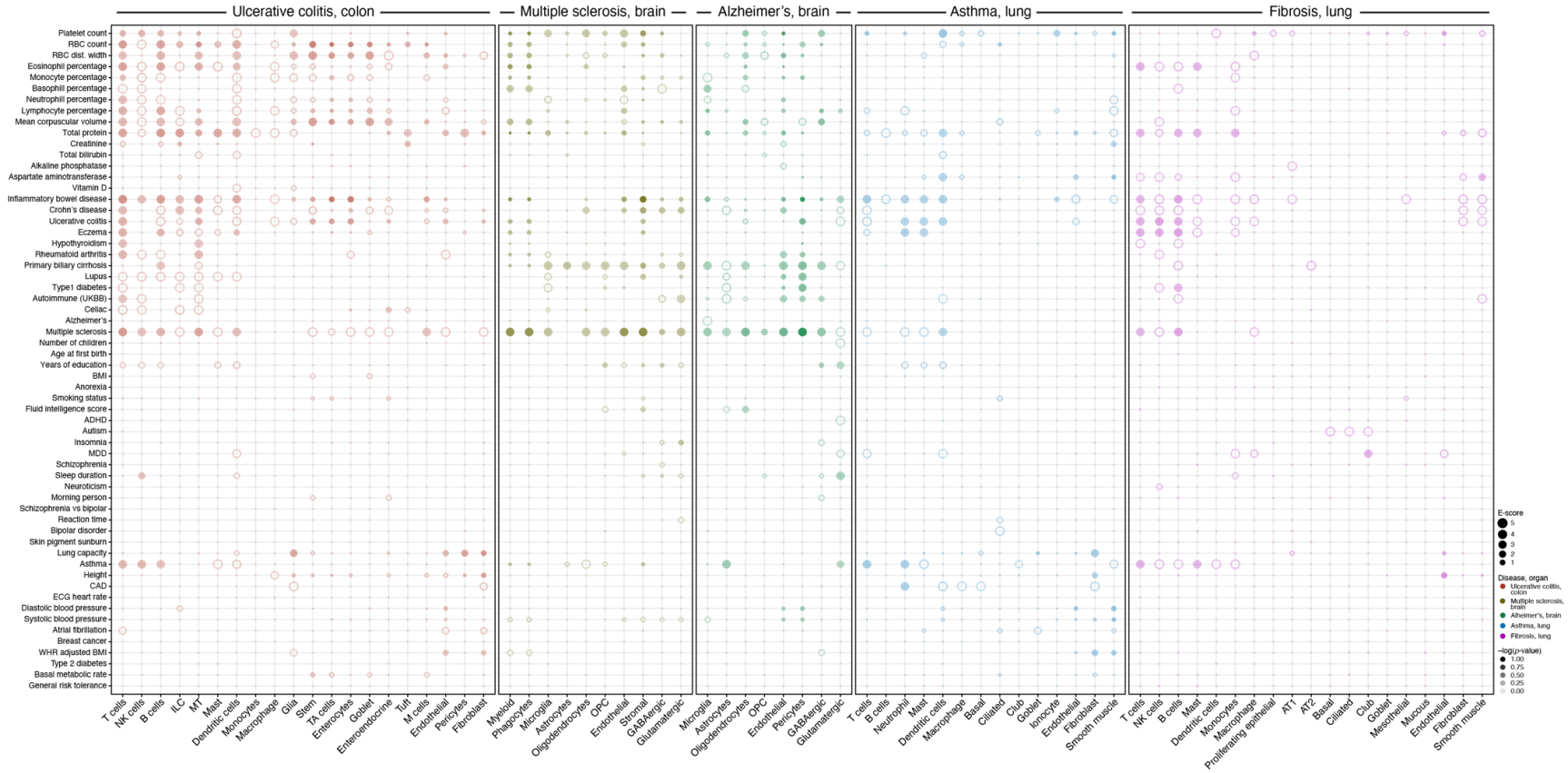

Figure S11. Analysis of disease progression programs across all tissues and traits. Magnitude (E-Score, dot size) and significance ($-\log_{10}$(P-value), dot color) of the heritability enrichment of disease progression programs (columns) from UC, MS, Alzheimer's, asthma and pulmonary fibrosis (labels 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 **table S2.** See **Data Availability** for higher resolution version of this figure.