Supplementary Note

Extended Overview of Methods

We constructed three kinds of gene programs from scRNA-seq data (**Figure 1b**): (i) cell type programs that represent genes specifically enriched in an individual broad cell type of a tissue (*e.g.*, colon T cells) compared to other cell types in that tissue; (ii) disease-dependent cell type programs that represent disease specific differences in gene expression within the same cell type (*e.g.*, colon T cells in UC *vs*. healthy colon); and (iii) cellular process programs that capture gene co-variation patterns within and across cell types (*e.g.*, MHC class II antigen presenting process varying across dendritic cells and B cells) (**Methods**). We constructed (healthy) cell type programs by assessing the differential expression of each gene for the focal cell type *vs*. all other cell types in the tissue using healthy individuals (with cell types defined by clustering[1] and annotated *post hoc*) and transforming each gene's Z score to a continuous-valued score on the probabilistic (0-1) scale (**Methods**). (Analogous to healthy cell type programs generated from healthy tissues, we also generated disease cell type programs from cell profiles from disease tissues.) We constructed disease-dependent cell type programs by assessing differential expression between cells of the same type in disease *vs*. healthy tissue and transforming each gene's Z score to a continuous-valued score (**Methods**), aiming to capture genes involved in disease and symptoms after onset. (We caution that disease-dependent programs may also capture genes reflecting genetic susceptibility to disease, rather than progression.) On average, disease-dependent cell type programs had low correlation with healthy cell type programs of the same cell type (Pearson r=0.16 across tissues; see below) compared to the much higher correlation between disease and healthy cell type programs (average r=0.62 across tissues); thus, we did not consider disease cell type programs in any of our primary analyses. Finally, independently of predefined cell type subsets, we constructed cellular process programs using unsupervised learning, via non-negative matrix factorization[2] (NMF) and a modified NMF (to jointly model both healthy and disease states) of normalized gene expression values, with the latent factors (programs) representing variation across continuums of cell types or processes active in multiple cell types. We computed the correlations between weights of each latent factor across cells and each gene's expression across cells and then transformed them to a 0-1 continuous-valued scale to define each cellular process program. We annotated each cellular process program by its most enriched pathways (**Methods**) and labeled it as 'intra-cell type' or 'inter-cell type' if highly correlated with only one or multiple cell type programs, respectively (**Methods**). Intra-cell type cellular processes can correspond to narrower cell types (*e.g.*, CD4 T cells) reflecting cell subsets of broader cell type categories (*e.g.*, T cells) or variation within a cell type continuum, whereas inter-cell type cellular process programs can reflect shared processes or transitions.

Next, 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)[3,4] and Activity-By-Contact (ABC)[5,6] strategies (Roadmap∪ABC) in the tissue underlying the program of interest (**Methods**). We used tissue level enhancer-gene links instead of cell type level enhancer-gene links because they generated more significant associations in benchmarking experiments based on current data (see below). We primarily focused on linking genes to non-coding regulatory variants (which may drive cell-type specific differences in expression), based on the results of our benchmarking experiments (see below).

Finally, we evaluated each gene program for disease heritability enrichment by applying S-LDSC[7] with the baseline-LD model[8,9] 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)[8,9] (**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[7]; this generalizes to annotations with values between 0 and 1[10]. We further define the Enrichment score (E-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 value 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 E-score as our primary metric, assessing statistical significance using a genomic block-jackknife as in our previous work[7], because the p-values can be compared across datasets, whereas the E-score magnitude can vary substantially in gene programs dominated by a smaller (or larger) number of genes. We primarily focus on E-scores greater than 2, because E-scores that are statistically significant but small in magnitude may have more limited biological importance, as the cell types underlying these E-scores may be tagging other causal cell types (**Methods**). We performed this analysis over healthy cell type programs (**Supplementary Data 1**), disease-dependent programs (**Supplementary Data 2**), and cellular process programs (**Supplementary Data 3**). If the effect has the same magnitude for pathways in the cell type *vs*. disease-specific pathways in the cell type, then we might observe a cell type enrichment but not disease dependent enrichment. If the effect has higher magnitude for disease-specific vs. cell type pathways, then we might observe a disease-dependent enrichment but not a cell type enrichment. Biologically, if only the cell type program is enriched, it could be because the native processes in this cell are impacted by genetic variants, and thus, for example, could impact disease initiation or onset. If only disease dependent programs are enriched, it could be because the process affected by genetic variants in that cell type is more active during disease or even only clearly apparent in the cell in the context of other tissue processes following onset of disease. Such genes are still often detectable by GWAS because in many common complex diseases, disease processes are gradual, or are processes that occur in healthy tissue under challenge, and thus the genes/process would still affect overall disease risk. We identified the top 50 genes driving disease enrichments with highest proximity based MAGMA (v 1.08) gene-disease association scores[11] of genes with high grade in each gene program (**Figure 1C, Supplementary Data 4, Methods**) focusing on genes that are both (i) close to a GWAS signal and (ii) in an enriched gene program.

Extended benchmarking of sc-linker

The Roadmap∪ABC enhancer-gene linking strategy outperformed every other enhancer-gene linking strategy we tested in identifying these expected enrichments, including its constituent Roadmap and ABC strategies, the standard 100kb window-based approach used in LDSC-SEG[12] (**Supplementary Fig. 1a, 2a-c**), and other SNP-gene linking strategies (**Supplementary Fig. 1b and Supplementary Data 5**). Additionally, the tissue-specific Roadmap∪ABC-immune enhancer-gene linking strategy outperformed cell-type-specific enhancer-gene linking strategies, supporting the use of tissue-specific enhancer-gene linking (**Supplementary Fig. 2l**). This trend may stem from existing cell-type-specific enhancer-gene links being noisier, due to the limited amount of underlying cell-type-specific data, or because tissue-specific enhancer-gene links may tag enhancer-gene links in causal cell types that were not assayed (distinct from tagging captured by cell type programs).

The cell type programs were robust to the number of cells and individuals. Specifically, cell type programs and their corresponding enrichment results were robust (correlation of *r*=0.91) to changes in the number of profiled cells for scRNA-seq datasets with greater than 500 cells (**Supplementary Fig. 2f-h**); larger scRNA-seq datasets can uncover cell populations and states that may be missed in smaller datasets, due to sampling power. The cell type programs were also highly similar across different sets of individuals (*r*=0.96 on average between programs of the same cell type generated from different samples, with consistent specificity in expected enrichments; **Supplementary Fig. 2i-k**).

We observed higher values of sensitivity/specificity index for enrichments of expected cell type-trait pairs for our polygenic approach based on specifically expressed genes *vs*. other cell types compared to several other approaches including (i) functional enrichment of fine-mapped SNPs[13] (**Supplementary Fig. 3a**); (ii) all expressed genes in a cell type, defined across several thresholds (**Supplementary Fig. 4**); (iii) specifically expressed genes *vs*. other genes in the same cell type; or (iv) specifically expressed genes *vs*. other genes in the same cell type, after normalizing each gene across cell types (**Supplementary Fig. 5 and Supplementary Data 6**). We hypothesize that the “*all expressed genes*” approach greatly underperforms sc-linker because, for a given expressed gene, centrality of function in a cell is often reflected in its level of expression compared to that in other cells[14,15].

Sc-linker also outperformed two methods that use the MAGMA software[11]. First, we compared sc-linker to a baseline method for scoring cell types by scoring each cell using MAGMA *gene-level* associations to a trait and averaging across all cells of a cell type. We scored each cell for a trait using the top 200 MAGMA genes with highest score for the trait, computing the average expression over all the genes and subtracting an expression-matched control gene set. Sc-linker attained a higher sensitivity/specificity index compared to this baseline (**Supplementary Fig. 2n**). Second, we compared sc-linker to MAGMA *gene set-level* association, either applied to binarized gene programs at different gene value thresholds ranging from 0.20 to 0.95 or applied to gene programs treated as continuous variables on the probability scale or negative log odds of the probability scale (**Methods**, **Supplementary Data 7**). Sc-linker slightly outperformed MAGMA gene set-level association, with a sensitivity/specificity index of 6.29 for sc-linker versus 4.76-5.83 for MAGMA (across different binarization thresholds and continuous variable based approaches; the binary threshold of 0.95 performed best) (**Supplementary Fig. 6**, **Supplementary Data 7**; sensitivity of 8.78 for sc-linker versus 7.55-8.68 for MAGMA). This was further underscored by a comparison across a broader set of cell types and diseases/traits. Specifically, we analyzed 3 major cell type categories (immune, brain, other) and 4 major categories of diseases/traits (blood cell traits, immune-related diseases, brain-related diseases, other diseases/traits), and used the most plausible pairings (immune cell types x blood cell traits, immune cell types x immune-related diseases, and brain cell types x brain-related diseases, which have previously been reported to include many true enrichments[12]), to define a sensitivity/specificity index (**Methods**); as in the analysis of blood cell traits, we caution that a limitation of this index is that other enrichments may be biologically real in some cases; thus, we also consider sensitivity to detect expected enrichments. Sc-linker attained a higher sensitivity/specificity index (9.47) compared to MAGMA gene set-level association (1.78-3.68 at different binarization thresholds and continuous variable based approaches; the negative log odds of the probability scale performed best) (**Methods, Figure 2c, Supplementary Data 7**). The difference in performance was primarily due to the higher sensitivity of sc-linker (sensitivity of 12.2 for sc-linker versus 4.50-7.70 for MAGMA). (We also compared sc-linker to FUMA[16], a web interface that applies (gene set-level) MAGMA using precompiled scRNA-seq data (distinct from the data in our study). FUMA underperformed both sc-linker and gene set-level MAGMA (**Supplementary Data 8 and Supplementary Data 9**), but we caution that this is not a fair comparison due to the different underlying scRNA-seq data used by FUMA.)

Extended analysis of disease critical brain cellular processes

The 12 brain cellular process programs showed that the significant enrichment of neuronal cell types above 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[17], and *TIMP2* (for VIP$^+$ neurons; ranked 7), implicated in obesity through hypothalamic control of food intake and energy homeostasis in mice[18,19]. 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[20]. We also observed enrichment in more specific cell subsets within 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, 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[21], with the top driving genes including *ATP6V0B* and *NDUFAF3* (ranked 1, 4).

“Tagging” cell types from one tissue to disease in another

The enrichment of Langerhans cells for AD is plausible given that Langerhans cells respond differently to Aβ peptides, which has implications in AD immunotherapy[22]. On the other hand, the enrichment of colon M cells for asthma may suggest a role for lung-resident M cells, which have not been identified to date but are expected to be in the lung, as M cells stimulate IgA antibody production as an immune response[23], while selective IgA immunodeficiency increases risk for asthma[24]. Similarly, the heart smooth muscle cell program may merely mirror that of airway smooth muscle cells, whose function is a pivotal determinant of lung capacity[25].

Gene driving enrichment in Alzhemiers and microglia disease-dependent programs

The top genes driving enrichment specifically in the disease-dependent microglia 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[26], *APOC1* is a an *APOE*-dependent suppressor of glial activation[27], and *TREM2* modulates microglial morphology and neuroinflammation in Alzheimer’s disease pathogenesis models[28].

Cellular processes involved in MS and Alzheimer's Disease

For MS, there was enrichment for the complement cascade disease-specific cellular process program (in B cells and microglia; the top driving genes included FC-complement genes *CD37*, *FCRL2* and *FCRL1* (ranked 1, 10, 14) consistent with studies showing that Complement activity is a marker for MS progression[29,30]. For Alzheimer's disease, the apelin signaling pathway disease-specific cellular process program is consistent with recent studies implicating this pathway in reducing neuroinflammation in animal models of Alzheimer's disease[31]. 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[32].

Role of healthy and disease-dependent T cells in Asthma

For example, healthy cell type and disease-dependent T cell programs were enriched in asthma, consistent with the contribution of T cell-driven inflammation to airway hyper-responsiveness and tissue remodeling[33]. From a pathway enrichment analysis, we identified that healthy T cell program overlapped with T cell receptor signaling, while the T cell disease-dependent program overlapped with RNA binding (**see Supplementary Data 10**). These partially overlapping programs both included IL2 signaling pathway genes; IL2 is a T cell growth factor that increases airway response to allergens[34] and drives differentiation of Th2 cells linked to asthma[35].

Disease critical cell types in IPF and COVID-19

For asthma, we looked into the gene driving the enrichments that were observed. For example, both healthy and disease-dependent fibroblast/stromal programs were enriched for lung capacity (but not asthma), consistent with the adverse impact of overproduction of extracellular matrix (ECM) on the reduced lung capacity and elasticity characteristic of fibrosis[36]. In the cell type program, top driving genes included *LOX* (ranked 1), which alters ECM mechanical properties via collagen cross-linking[37], and *TGFBR3* (ranked 37) which regulates the pool of available TGFβ, a master regulator of lung fibrosis. Notably, the enrichment of basal cell disease-dependent programs in lung capacity are supported by the significant increase (p-value: $3x10^{-5}$) in basal cells in asthma *vs*. healthy lungs (**Figure 7e**). Expanding the analysis to cellular process programs, the top driving genes of the enrichment of a MAPK signaling pathway program for lung capacity, include *FOXA3* (ranked 1), which plays a key role in allergic airway inflammation[38], and *PDE2A* (ranked 2), which has been associated with alveolar inflammation[39].

For IPF, a disease characterized by mucociliary dysfunction[40], the mucous disease-dependent program was most enriched, and nominally significant ($p = 0.04$, not FDR significant), with top driving genes including *DSP* (ranked 1), a cell-cell adhesion molecule linked to tissue architecture in IPF lung[41], and *MUC5B* (ranked 2), the well characterized genetic risk factor for IPF that likely increases mucinous expression in terminal airways of the lung[40].

For severe COVID-19[42], the macrophage disease-dependent program was enriched, and nominally significant ($p = 0.01$, not FDR significant), with top driving genes including key antiviral enzyme activators[43,44] *OAS3* and *OAS1* (ranked 1, 3), and *CCR5,* a chemokine receptor in which therapeutic intervention has been associated with improved prognosis in severe COVID-19 patients[45]. Further analyses of a meta-atlas of COVID-19 scRNA-seq in conjunction with COVID-19 GWAS data are described elsewhere[46]. Our nominally significant findings should be interpreted cautiously, but should become more powered as IPF and COVID-19 GWAS sample sizes grow.

Extended discussion of limitations

First, the enhancer-gene linking strategies from Roadmap and Activity-By-Contact (ABC) models are limited in the tissues and cell 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 a standard gene window-based approach. 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, sc-linker does not distinguish whether two cell types (or more generally, gene programs) implicated in disease exhibit conditionally independent signals. Assessing this via a conditional S-LDSC analysis of the corresponding SNP annotations is likely to be underpowered, as the gene programs (and SNP annotations) may be highly correlated. A more powerful approach may be to define cell type programs based on specific expression relative to a narrower set of cells[12]. This approach should be particularly useful in analyses of fine-grained cell types, in which overlapping signals between related cell type programs is a particular concern. Fourth, the continuous-valued scores of the sc-linker gene programs were transformed to the 0-1 scale using min/max normalization[47], but further investigation of the choice of scale remains as a future direction. Fifth, 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[48], or may only be apparent in a disease context, as we observe for rare M cells in UC. Sixth, we have focused on human scRNA-seq data[49]; however, incorporating data from animal models, as discussed in prior work[50], would allow experimental validation of disease mechanisms in model organisms. Seventh, the disease-dependent programs that we link to disease may not be causal for disease, but rather reflect disease-induced changes or genetic susceptibility to disease[51,52]. However, our findings clearly validate the relevance of these gene programs to disease as observed in M cells and UC[53]. Eighth, the LD score regression framework[7] is primarily applicable to common and low-frequency variants, and less applicable to rare variant enrichments. Ninth, we capture programs by cell category or gene co-variation, whereas future work could extend beyond these to capture dynamic cellular transitions[54].

References

1. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
2. 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).
3. Liu, Y., Sarkar, A., Kheradpour, P., Ernst, J. & Kellis, M. Evidence of reduced recombination rate in human regulatory domains. *Genome Biol.* **18**, 193 (2017).
4. Ernst, J. *et al.* Systematic analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43 (2011).
5. Fulco, C. P. *et al.* Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664 (2019).
6. 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.
7. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228 (2015).
8. Gazal, S. *et al.* Linkage disequilibrium dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421 (2017).
9. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202 (2019).
10. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041 (2018).

11. 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).

12. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621 (2018).

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

14. Guo, Y. E. *et al.* Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **572**, 543–548 (2019).

15. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).

16. 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).

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

18. 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).

19. 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).

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

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

22. Cheng, J. *et al.* Dendritic and Langerhans cells respond to Aβ peptides differently: implication for AD immunotherapy. *Oncotarget* **6**, 35443–35457 (2015).

23. Rios, D. *et al.* Antigen sampling by intestinal M cells is the principal pathway initiating mucosal IgA production to commensal enteric bacteria. *Mucosal Immunol.* **9**, 907–916 (2016).

24. Celani, C. *et al.* Selective IgA deficiency and the risk of asthma. *Eur. Respir. J.* **42**, (2013).

25. Irvin, C. G. Lung volume: a principle determinant of airway smooth muscle function. *Eur. Respir. J.* **22**, 3–5 (2003).

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

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

28. 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).

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

30. 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).

31. 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).

32. 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).

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

34. 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).

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

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

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

38. 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).

39. 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).

40. Yang, I. V., Fingerlin, T. E., Evans, C. M., Schwarz, M. I. & Schwartz, D. A. MUC5B and Idiopathic Pulmonary Fibrosis. *Ann. Am. Thorac. Soc.* **12**, S193–S199 (2015).

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

42. 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.

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

44. 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.

45. 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.

46. 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.

47. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082 (2019).

48. 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).

49. Peng, Y.-R. *et al.* Molecular Classification and Comparative Taxonomics of Foveal and Peripheral Cells in Primate Retina. *Cell* **176**, 1222-1237.e22 (2019).

50. 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).

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

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

53. Smillie, C. S. *et al.* Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714-730.e22 (2019).

54. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

55. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

56. Jung, I. *et al.* A Compendium of Promoter-Centered Long-Range Chromatin Interactions in the Human Genome. *Nat. Genet.* **51**, 1442–1449 (2019).

57. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

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

SUPPLEMENTARY MATERIALS

Supplementary Table 1-2
Supplementary Figure 1-7
Supplementary Data File Legends 1-12

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

Supplementary Table 1. 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

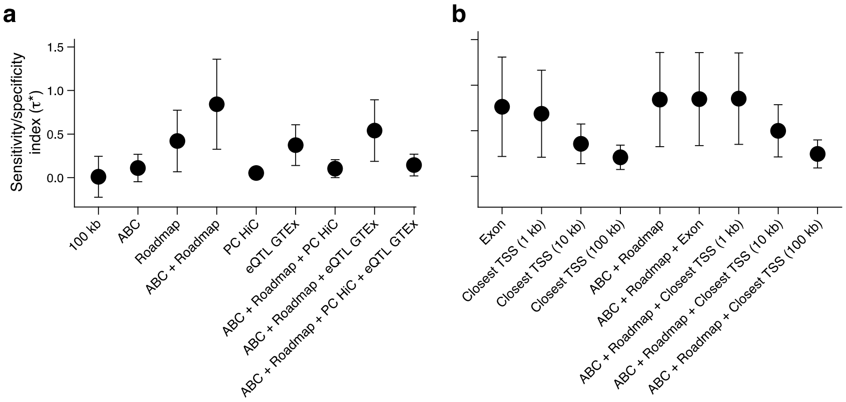
Trait category	**Trait**	**Source**	**Sample size (N)**
Blood cell traits	Lymphocyte percentage	UK Biobank	444502
	Monocyte percentage	UK Biobank	439938
	Platelet count	UK Biobank	444382
	Red blood cell count	UK Biobank	445174
	Red blood cell volume	UK Biobank	442700
	Eosinophil count	UK Biobank	439938
	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

Supplementary Table 2. 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.

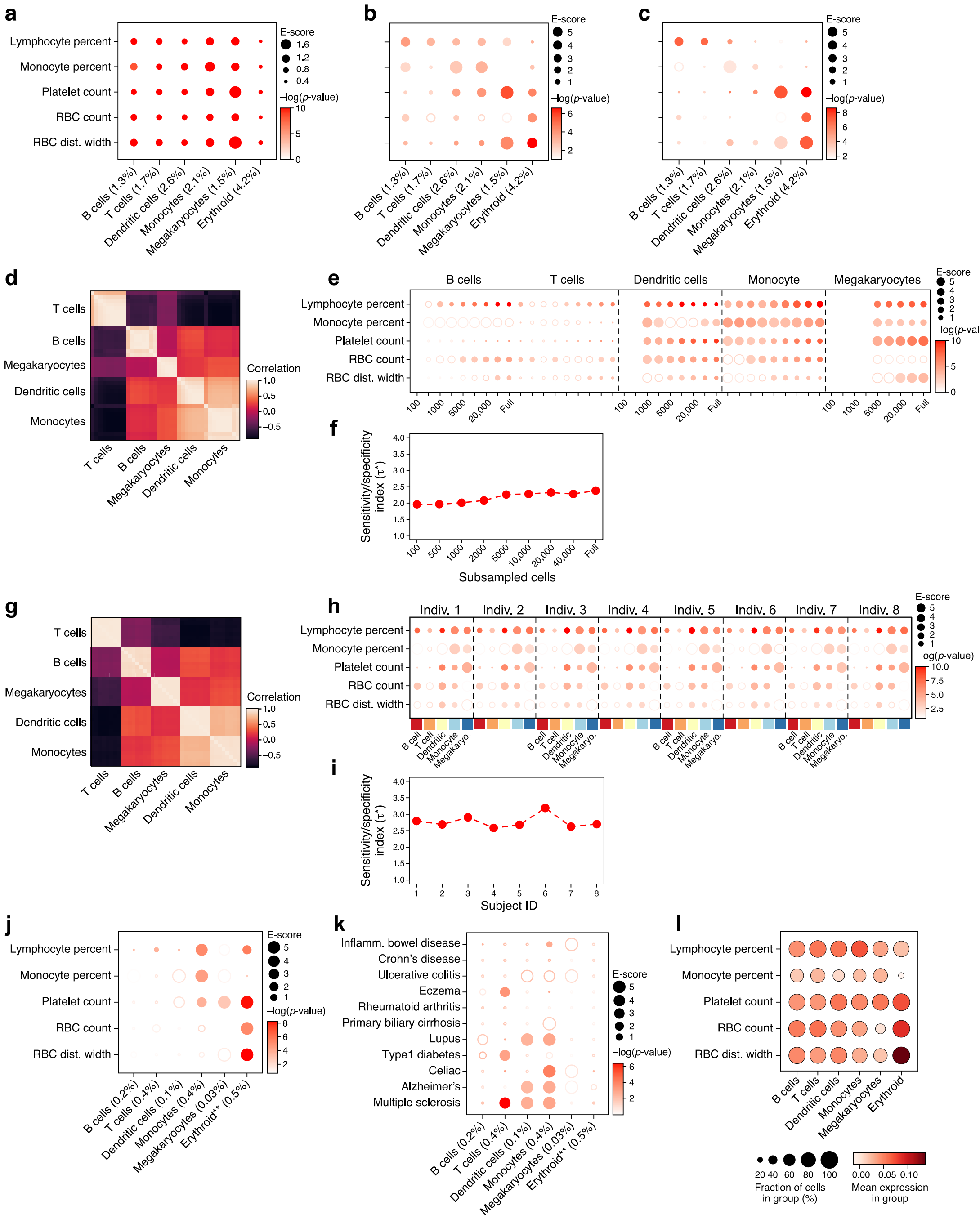
SUPPLEMENTARY FIGURES

Supp. Fig. 1



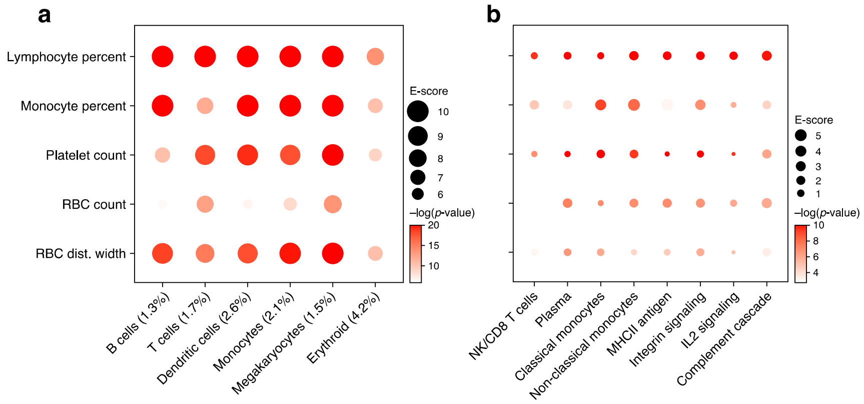
Supplementary Fig. 1. Roadmap ∪ ABC outperforms other strategies. The metric for this comparison is the sensitivity/specificity index. The sensitivity/specificity index (y axis, mean and s.e.) of immune programs and blood cell traits for different choices of regulatory regions linked to genes (*x* axis), including Roadmap ∪ ABC enhancer-gene strategy (ABC+Roadmap) and its constituent ABC and Roadmap strategies, promoter capture Hi-C (PC-HiC)[55,56] and eQTLs from the GTEx data[47], and combination of Roadmap ∪ ABC with PCHiC (Roadmap+ABC+PCHiC), Roadmap ∪ ABC with eQTL (Roadmap+ABC+eQTLGTEx) and both PCHiC and eQTL (Roadmap+ABC+PCHiC+eQTLGTEx) (x axis, **a**), or closest TSS linking strategy between SNPs and genes at different distances (1kb, 10kb and 100kb), and their combinations with Roadmap ∪ ABC. Data are presented as mean values +/- SEM. Numerical results are reported in **Supplementary Data 5**.

Supp. Fig. 2



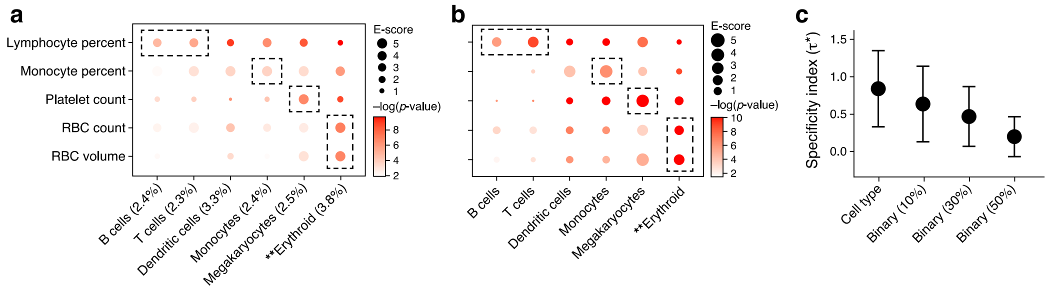
Supplementary Fig. 2. Benchmarking sc-linker across immune cell type programs and blood cell traits. a-c. Magnitude (E-score, dot size) and significance (-log$_{10}$(P-value), dot color) of the heritability enrichment of immune cell type programs (columns) aggregated over 4 scRNA-seq datasets (PBMC (2), 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$\cup$ABC-immune strategy in **Figure 2b**). **d.** Pairwise correlation heat map between all cell type programs computed for each sample separately. **e.** Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune cell type programs constructed for each sample. **f.** Sensitivity/specificity index (y axis; see Methods) for immune cell type programs generated from each individual. **g.** Pairwise correlation heat map between all cell type programs computed for each dataset size separately. **h.** Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune cell type programs constructed for each dataset size. **i.** Sensitivity/specificity index (y axis; see Methods) for immune cell type programs generated from subsampled PBMC scRNA-seq data at varying numbers of cells. **j,k.** Magnitude (E-score, dot size) and significance (-$\log_{10}$(P-value), dot color) of the heritability enrichment of immune cell type programs (columns) for 5 blood cell traits (j) and 11 autoimmune traits (k). **l.** Mean gene set expression score (dot color) from the baseline cell scoring approach. Comparison of panels l,m and n remains subjective, as the two metrics plotted (E-score/p.E-score in **j,k**; cell scores in **l**) are in different types of scoring schemes. Details for all traits analyzed are in **Supplementary Table 2.**

Supp. Fig. 3



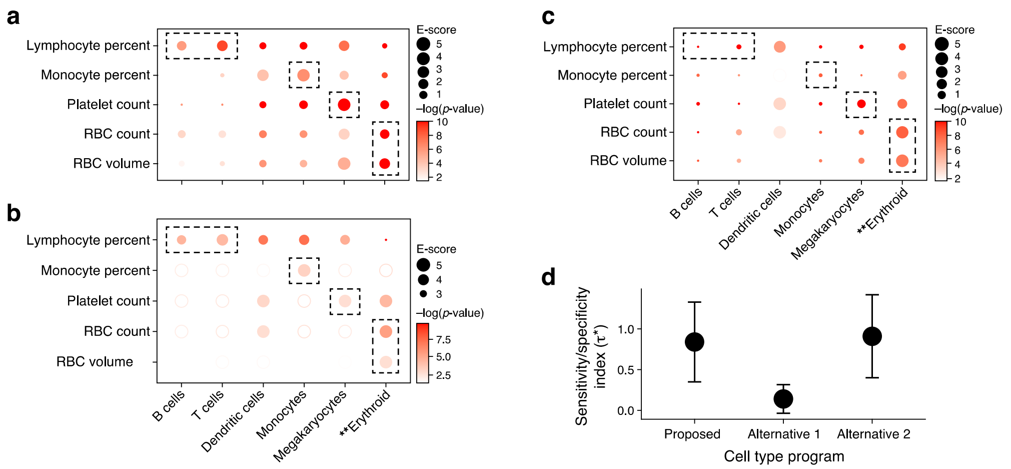
Supplementary Fig. 3. Analysis of functional enrichment of fine-mapped SNPs of immune cell type programs and heritability enrichment of immune cellular process programs. a. Functional enrichment of fine-mapped SNPs of immune cell type programs. Magnitude (Enrichment, dot size) and significance (-$\log_{10}$(P-value), 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. [58]). **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 **Supplementary Table 2.** The metric for comparing overall results in expected enrichments is the sensitivity/specificity index.

Supp. Fig. 4



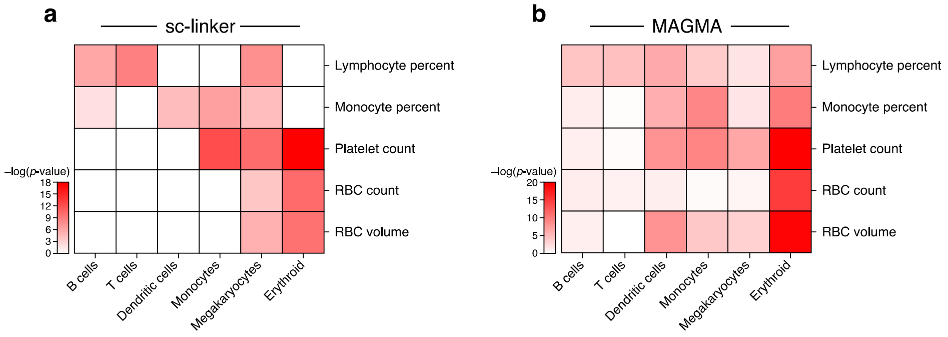
Supplementary Fig. 4: Evaluation of dichotomized gene programs. a,b. Enrichment in blood cell traits for binary and regular cell type programs. Enrichment (E score, dot size; and significance (-$\log_{10}$(P-value), dot color) for blood cell traits (rows) with cell type program defined by genes expressed in more than 10% of cells (**a**) or by our regular approach (**b**, as in **Figure 2d**). The size of each corresponding SNP annotation (% of SNPs) is reported in parentheses. **c.** The metric for this comparison is the sensitivity/specificity index. Regular cell type programs have a higher sensitivity/specificity index than dichotomous ones. The sensitivity/specificity index metric (y axis, mean and s.e.) for blood cell traits and immune cell type programs defined by our regular approach ("cell type") or by genes expressed in more than 10, 30 or 50% of cells of a given type (x axis). Data are presented as mean values +/- SEM. Numerical results are reported in **Supplementary Data 6**.

Supp. Fig. 5

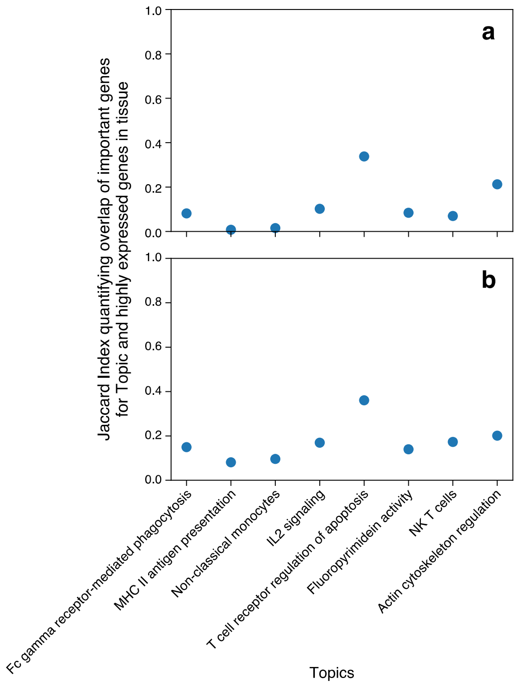


Supplementary Fig. 5. Evaluation of alternative approaches of gene program construction. a-c. Enrichment in blood cell traits for immune cell type programs defined in two different approaches. (**a**) Enrichment (E score, dot size; and significance (-$\log_{10}$(P-value), dot color) for blood cell traits (rows) with cell type programs (columns) defined either by genes differentially enriched in expression in a cell type compared to other genes in the same cell type (**a**), by genes differentially enriched in a cell type compared to their expression in other cell types (**b**, the primary analysis in this study), or by a combination of the previous two strategies (**c**). **d**. The metric for this comparison is the sensitivity/specificity index. Sensitivity/specificity index (y axis, mean and s.e.) for blood cell traits and immune cell type programs for the approaches in a-c. Data are presented as mean values +/- SEM. Numerical results are reported in **Supplementary Data 6**.

Supp. Fig. 6



Supplementary Fig. 6. Comparison of sc-linker and MAGMA. Negative log p-value of immune cell type programs and blood cell traits for (**a**) E-score in sc-linker analysis, and (**b**) MAGMA gene-set level association analysis. For the MAGMA analysis, the gene program is binarized using a threshold=0.95 and numerical results for other binarization thresholds and continuous variable based approaches are reported in **Supplementary Data 7**. Numerical results are reported in **Supplementary Data 9**.



Supplementary Fig. 7: Top genes in blood cellular processes are neither highest expressed in cells nor in the tissue overall. Overlap (Jaccard index, y axis) between the top 200 genes in each blood cellular processes (x axis) and the highest expressed genes in the top 50 cells (based on the weight from the NMF decomposition) associated with the cellular process (**a**) or overall across the tissue (**b**).

EXTENDED DATA FILE LEGENDS

Supplementary Data 1: Healthy cell type program heritability enrichment results. Numerical values for E-score and significance are reported for all cell type programs and traits analyzed.

Supplementary Data 2: Disease-dependent program heritability enrichment results. Numerical values for E-score and significance are reported for all disease-dependent programs and traits analyzed.

Supplementary Data 3: 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.

Supplementary Data 4: List of genes driving each enrichment. Up to 50 genes with the strongest MAGMA gene score and membership in the gene program.

Supplementary Data 5: Heritability enrichment results from eQTL, PCHi-C and other alternative enhancer-gene linking strategies. Numerical values for E-score and significance are reported for all traits analyzed with alternative enhancer-gene linking strategies.

Supplementary Data 6: Heritability enrichment results from alternative approaches for constructing cell type gene programs. Numerical values for E-score and significance are reported for all traits analyzed with the alternative cell type programs.

Supplementary Data 7: MAGMA analysis with alternative input representations. Sensitivity/specificity index, standard error, average sensitivity and average specificity for various binarization thresholds (0.20 to 0.95) and continuous variable approaches (probability scale or negative log odds of the probability scale), for both the analysis of 5 blood cell traits and the analysis of 4 major categories of diseases/traits.

Supplementary Data 8: FUMA enrichments for blood cell traits and immune cell type programs. Numerical values for beta, standard error and p-value for all cell types and traits analyzed.

Supplementary Data 9: MAGMA gene set enrichment results for all cell type programs. MAGMA scores across all traits analyzed.

Supplementary Data 10: Pathway enrichment analysis for each disease-dependent program. Gene overlap, p-value and gene list for each of the enriched pathway ontology terms across KEGG, Wikipathways and Reactome.

Supplementary Data 11: Composition of cell types in each tissue. Proportion of cells observed for each cell type and condition in each of the single cell datasets.

Supplementary Data 12: Correlation between disease-dependent and healthy cell type program.