medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Inferring causal cell types of human diseases and risk variants from candidate regulatory elements

Artem Kim [1,2], Zixuan Zhang [1,2], Come Legros [1,2], Zeyun Lu [1,2], Adam de Smith [1,2], Jill E. Moore [4], Nicholas Mancuso [1,2,3], Steven Gazal [1,2,3]

[1] Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[2] Center for Genetic Epidemiology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[3] Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

[4] Department of Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA, USA

Correspondence should be addressed to A. K. (artemkim@usc.edu) or S.G. (gazal@usc.edu).

Abstract

The heritability of human diseases is extremely enriched in candidate regulatory elements (cRE) from disease-relevant cell types. Critical next steps are to infer which and how many cell types are truly causal for a disease (after accounting for co-regulation across cell types), and to understand how individual variants impact disease risk through single or multiple causal cell types. Here, we propose CT-FM and CT-FM-SNP, two methods that leverage cell-type-specific cREs to fine-map causal cell types for a trait and for its candidate causal variants, respectively. We applied CT-FM to 63 GWAS summary statistics (average N = 417K) using nearly one thousand cRE annotations, primarily coming from ENCODE4. CT-FM inferred 81 causal cell types with corresponding SNP-annotations explaining a high fraction of trait SNP-heritability (~2/3 of the SNP-heritability explained by existing cREs), identified 16 traits with multiple causal cell types, highlighted cell-disease relationships consistent with known biology, and uncovered previously unexplored cellular mechanisms in psychiatric and immune-related diseases. Finally, we applied CT-FM-SNP to 39 UK Biobank traits and predicted high confidence causal cell types for 2,798 candidate causal non-coding SNPs. Our results suggest that most SNPs impact a phenotype through a single cell type, and that pleiotropic SNPs target different cell types depending on the phenotype context. Altogether, CT-FM and CT-FM-SNP shed light on how genetic variants act collectively and individually at the cellular level to impact disease risk.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Introduction

Understanding how genetic variants act collectively and individually at the cellular level to impact disease risk is critical to improve our understanding of disease biology [1–3]. Previous studies have integrated genome-wide association studies (GWASs) with cell-type-specific (CTS) functional annotations to identify cell types presenting significant association with the disease or complex trait [4–10]. However, most of these associations are not truly causal (we define here a *causal cell type* as a cell type where altered gene regulation impacts disease risk) but only 'tag' a disease causal cell type as a consequence of shared regulatory patterns (e.g., gene expression and regulatory elements) across cell types and tissues [11–15]. This issue has hindered our ability to pinpoint precise cell types underlying the disease biology, and to determine the number of distinct cell types collectively targeted by disease variants. In addition, it is unclear how individual disease variants act at the cellular level (i.e., does it impact disease risk by acting on an individual causal cell type, or on multiple causal cell types simultaneously?), and if pleiotropic variants act on multiple traits through the same or different cell types [16]. For example, it remains uncertain which brain cell types are targeted by variants impacting body mass index (BMI) and obesity, as well as the additional role of immune and adipose cell types [17,18]; similarly, it is also unclear through which cell types variants in the *FTO* locus confer obesity risk [19]. Altogether, these limitations hinder our understanding of how genetic variants act at the cellular level to confer disease risk, and on how to translate GWAS findings into discovery of actionable drug targets.

Methods accounting for gene co-regulation in expression quantitative trait loci (eQTLs) datasets have already been proposed to fine-map causal tissues of human diseases [14] and their risk variants [20]. However, current eQTLs datasets have been generated on bulk tissues that do not capture cell-type-specific effects and often have limited overlap with GWAS results [21–24], limiting insights from these methods. Unlike eQTLs, candidate regulatory elements (cREs) are already available for thousands of cell types and conditions [25–29], and are extremely enriched in disease SNP-heritability (h^2) when disease-relevant cell types have been assayed [6,28–31]. However, we lack methods to fine-map causal cell types of human diseases and risk variants from CTS cREs, while accounting for shared cREs across cell types.

Here, we propose CT-FM and CT-FM-SNP, methods that fine-map causal cell types at genome-wide and single variant level from CTS SNP-annotations, respectively. Both methods jointly analyze GWAS summary statistics with a set of CTS SNP-annotations and output probabilities for each cell type to be causal for a trait (CT-FM) or a particular GWAS candidate

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

variant (CT-FM-SNP), as well as credible sets (CSs) reflecting the number of independent causal signals while accounting for uncertainty in cell type prioritization. Both methods were applied using 927 curated CTS SNP-annotations coming primarily from ENCODE4 (ref. [32]) and single-cell ATAC-seq data from 30 tissues [29]. We first validated and benchmarked CT-FM and CT-FM-SNP using simulations and GWAS of blood cell traits that biologically correspond to specific immune cell types. We applied CT-FM to a set of 63 independent GWASs (average *N* = 417K), identified expected causal cell types, and highlighted previously unexplored cellular mechanisms involved in psychiatric and immune-related diseases. Finally, we applied CT-FM-SNP to 6,975 candidate causal {non-coding SNP, trait} pairs identified via SNP-fine-mapping within 39 UK Biobank traits [33,34]. Results suggest that most individual SNPs impact a phenotype through a single cell type, and that pleiotropic SNPs might target different cell types depending on the phenotype context. Overall, CT-FM and CT-FM-SNP can be leveraged to infer disease causal cell types and uncover new cellular mechanisms involved in complex traits.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Overview of methods

We developed CT-FM, a method aiming to fine-map the causal cell types of a trait, and CT-FM-SNP, a method aiming to fine-map the causal cell types of a trait-associated SNP. They take as input a GWAS summary statistics with a matching LD reference panel, a set of CTS SNP-annotations, and (for CT-FM-SNP only) a list of trait's candidate causal SNPs (**Fig. 1**).

CT-FM aims to identify CTS SNP-annotations that are the most likely to explain the h^2 observed across a large set of CTS SNP-annotations. Formally, CT-FM models the vector of per-normalized-genotype effect sizes β as a mean 0 vector with variance $Var(\beta_j) = \Sigma_{b\in B} a_b(j)\tau_b + \Sigma_{c\in C} a_c(j)\tau_c$, where *B* represents a set of background (non-CTS) SNP-annotations (including the coding, enhancer, and promoter from the baseline model [4,6]), $a_b(j)$ is the indicator variable of variant *j* for SNP-annotation *b*, τ_b is the contribution of *b* to $Var(\beta_j)$, and *C* represents the set of CTS SNP-annotations. Rather than assuming that all *C* SNP-annotations contribute to $Var(\beta_j)$, CT-FM assumes that at most L SNP-annotations contribute positively to $Var(\beta_j)$ (i.e., $\tau_c > 0$). Specifically, we model the vector τ of $\tau_{c\in C}$ as $\tau = \sum_{l=1}^{L}\tau_l$ as the sum of L single-effects $\tau_l = \gamma_l \cdot b_l$ where γ_l is a $C \times 1$ binary vector indicating which CTS SNP-annotation is causal for the lth effect, and b_l is a scalar quantifying the contribution of the causal l^{th} effect to $Var(\beta_j)$. We caution here that CT-FM specifically assumes that all disease-relevant cell types have corresponding SNP-annotations among *C*, while they might not have been assayed in practice (see **Discussion**). CT-FM applies the following procedure to infer L and the most likely causal cell types. First, for each CTS SNP-annotation c, it applies stratified LD score regression (S-LDSC) [30,35,36] with the *B* background SNP-annotations to estimate their marginal effect $\tilde{\tau}_c$. Second, it fine-maps SNP-annotations from the vector of marginal $\tilde{\tau}$ Z-scores using an extension of the Sum of Single Effects (SuSiE) model [37,38] on a co-regulation matrix defined as the correlation of the LD scores of the CTS SNP-annotations adjusted on the *B* background SNP-annotations. CT-FM outputs posterior inclusion probabilities (PIP) for every CTS SNP-annotation, which correspond to probabilities of each cell type to be causal for the trait. It also regroups putative causal cell types within independent credible sets (CSs) allowing to estimate the number of independent causal signals while accounting for uncertainty in cell type prioritization. Here, CT-FM defined *candidate* causal cell type when a corresponding SNP-annotation was found in a CS, and *highly-confident* causal cell type when a corresponding SNP-annotation had a PIP ≥ 0.5; in instances where no *highly-confident* causal cell type was

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

detected but multiple CTS SNP-annotations from the same cell type were detected in the same CS (for example, several SNP-annotations corresponding to B cells stimulated in different context), we reported a combined PIP (cPIP) for the cell type (i.e., B cells). Finally, CT-FM defined the number of causal cell types L for a trait as the number of CSs detected.

Because disease risk can be influenced by SNPs acting through different causal cell types, it is critical to infer which cell types are targeted by a particular causal SNP to understand how genetic variants act at the cellular level to confer disease risk. To infer the most likely causal cell types of a trait candidate causal SNP, CT-FM-SNP leverages both polygenic enrichment within overlapping SNP-annotations and co-regulation within cREs. Specifically, it applies the same workflow as CT-FM, but differs in that it restricts the fine-mapping step to marginal $\tilde{\tau}$ Z-scores of CTS SNP-annotations that overlap the candidate SNP. Similar to CT-FM, CT-FM-SNP defined *candidate* causal cell types for a trait candidate SNP when their corresponding SNP-annotations were found in a CS, and *highly-confident* causal cell types for a trait candidate SNP when their corresponding SNP-annotations presented a PIP ≥ 0.5. We note that while the inference of multiple CSs by CT-FM suggests that multiple cell types are causal for the disease (as it relies on polygenicity), the inference of multiple CSs by CT-FM-SNP suggests that the candidate SNP overlap cREs from multiple causal cell types but does not infer whether the candidate SNP impacts disease risk by disrupting gene regulation within a single or multiple cell types.

We applied CT-FM and CT-FM-SNP with a total of 927 CTS SNP-annotations (capturing, on average, 1.1% of common SNPs) derived from cREs of human cell-types and human-derived cell lines (see **Fig. 1** and **Methods**), classified into 9 biological groups for visualization (**Supplementary Fig. 1**). CTS SNP-annotations were candidate regulatory elements from ENCODE4 (ref. [32]) (653), CATlas single-cell ATAC-seq data [29] (222) and the ABC method [27] (52) (**Supplementary Table 1**). CT-FM was successfully applied (i.e., detected at least one CS) on a set of 63 independent and well-powered GWASs prioritizing diseases over quantitative traits (average N = 417K; **Supplementary Table 2**), and CT-FM-SNP was applied to 6,975 {non-coding variant, trait} pairs obtained via SNP-fine-mapping on 39 UK Biobank traits [33,34]. We investigated the contribution of each CS to trait h^2 by creating SNP-annotations where we merged its constituent SNP-annotations.

Further details are provided in **Methods**. We have released open-source software implementing our framework (see **Code availability**) and have made all CTS SNP-annotations and GWAS summary statistics analyzed publicly available (see **Data availability**).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Validating and benchmarking CT-FM using simulations

We performed extensive simulations to assess CT-FM power. We simulated GWAS summary statistics using h^2 = 0.5, different sample sizes *N*, realistic LD patterns from the UK Biobank, and effect sizes depending on per-SNP h^2 estimated by S-LDSC on a height GWAS [39]. We considered one scenario with a single causal cell type (osteoblast), and one scenario with two causal cell types (osteoblast and fibroblast). We evaluated and benchmarked CT-FM by estimating the mean PIP of the causal SNP-annotation(s) across simulations, the proportion of causal SNP-annotations identified as a candidate causal cell type (candidate causal cell type sensitivity), and the proportion of SNP-annotations identified as a highly-confident causal cell type that are truly causal (highly-confident causal cell type precision). In preliminary analyses, we determined that these three metrics were optimized when restricting CT-FM fine-mapping step to GWAS summary statistics where at least one out of the CTS SNP-annotations has an S-LDSC marginal $\tilde{\tau}$ Z-score > 4 (**Supplementary Fig. 2**); we thus decided to restrict all further CT-FM and CT-FM-SNP analyses to summary statistics with at least one CTS SNP-annotation found in CS and presenting S-LDSC marginal $\tilde{\tau}$ Z-score > 4 (hereafter referred to as well-powered GWAS). Further details are provided in **Methods**.

CT-FM power and accuracy were fairly high for well-powered GWASs (**Fig. 2** and **Supplementary Table 3**). Specifically, in simulations with *N* = 350K (same order of magnitude than our analyzed GWASs) and with one causal cell type, we estimated that the mean PIP of the true causal cell type was 0.77 ± 0.01, that the sensitivity to detect candidate causal cell type was 93 ± 1%, and that the precision of the highly-confident causal cell type inference was 88 ± 1%. In simulations where the causal SNP-annotation did not have the highest PIP, its correlation with the SNP-annotation with the highest PIP had a median of 0.51 (mean = 0.56), demonstrating that when CT-FM fails to identify the causal cell type it still pinpoints a highly correlated cell type (**Supplementary Fig. 3**). In simulations with two causal cell types, we observed a decrease (but still fairly high values) for the mean PIP of the true causal cell types (0.54 ± 0.01) and the candidate causal cell type sensitivity (71 ± 1%), while the highly-confident causal cell type precision was similar (87 ± 1%). Finally, we observed that CT-FM power and accuracy was reduced for *N* = 100K, and that increasing sample size to *N* = 1M did not improve the results.

Altogether, these results demonstrate that CT-FM is a powerful and precise method to infer the causal CTS SNP-annotations from well-powered GWAS.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Validating and benchmarking CT-FM and CT-FM-SNP on blood traits

As proof of principle, we applied CT-FM and CT-FM-SNP to five blood cell traits that biologically correspond to specific immune cell types [5,40,41] (**Fig. 3**).

We applied CT-FM to the five traits using GWAS summary statistics from the UK Biobank [39] (average N = 444K) and identified expected causal cell types (**Fig. 3a** and **Supplementary Table 4**). Specifically, CT-FM identified one CS per trait (4.4 CTS SNP-annotations per trait on average), all highlighting a high-confidence causal cell type (PIP ≥ 0.5), such as T cells for lymphocyte count (PIP = 0.59 for CD8+ T cells) and monocytes for monocyte count (PIP = 0.71); overall, results were consistent with ref. [5]. CT-FM greatly reduced the number of candidate causal cell types compared to the initial association signal inferred by S-LDSC (**Fig. 3b** and **Supplementary Table 5**), highlighting the need and the benefit of cell type fine-mapping. Additional benchmarking analyses investigating the number of CSs, SuSiE parameters L and sample size n, the use of SuSiE-inf [42], and quality control on CSs are described in **Supplementary Fig. 4**, **Supplementary Tables 4** and **6**, and in **Methods**, respectively.

We next applied CT-FM-SNP on 1,564 candidate causal {non-coding variant, trait} pairs of the 5 blood traits (obtained via SNP-fine-mapping in the UK Biobank [33,34] and by selecting SNPs with a SNP-PIP > 0.5). We identified at least one CS for nearly half of the candidate SNPs (49%, 764 out of 1,564; 835 CSs in total) and detected a high-confidence causal cell type for 90% of these CSs (754 CSs with PIP > 0.5 among the 835 CSs), highlighting a high success rate of CT-FM-SNP (**Fig. 3c** and **Supplementary Table 7**). Among the CT-FM-SNP CSs with a high-confidence causal cell type, 76% (575 out of 754) corresponded to candidate cell types previously identified by CT-FM, demonstrating high consistency of causal cell types inferred by CT-FM and CT-FM-SNP methods (**Fig. 3d**, **Supplementary Table 8**). For the platelet count trait, we identified 67 CSs with a high-confidence causal cell type inferred by CT-FM-SNP but not found in CT-FM CSs (**Fig. 3d**); approximately half of these high confidence causal cell types (52%, 35 out of 67) corresponded to the megakaryocyte cell type (**Supplementary Table 9**, see also **Supplementary Fig. 4**), indicating that in cases where CT-FM fails to detect a cell type (for example, if targeted by a limited number of causal variants) this cell type can be captured by CT-FM-SNP. Similar conclusions for the red blood cell (RBC) volume trait are discussed in **Supplementary Table 9**. Similar patterns were observed when restricting analyses to SNPs with SNP-PIP > 0.95 (282 SNPs in total), demonstrating that our conclusions are robust to the imperfect selection of candidate causal SNPs (**Supplementary Fig. 5**).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Altogether, our results demonstrate that CT-FM and CT-FM-SNP were able to capture known cellular mechanisms of five blood cell traits.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Applying CT-FM to 63 GWASs

We successfully applied CT-FM to a total of 63 independent well-powered GWASs (average *N* = 417K, see **Methods**) and reported our main findings in **Fig. 4**. CT-FM identified 81 CSs (each CS containing an average of 7.07 CTS SNP-annotations) and 59 high-confidence causal cell types (PIP > 0.5) (**Fig. 4a** and **Supplementary Table 10**); for the 22 CSs without a high confidence causal cell type, we were able to identify 8 cell types with a cPIP > 0.5 (**Supplementary Table 11**). CT-FM identified an average of 1.29 CS and 9.09 candidate CTS SNP-annotations per trait, representing a ~20x decrease from candidate SNP-annotations identified by S-LDSC (false discovery rate (FDR) $P < 0.05$) (**Fig. 4b, Supplementary Table 12**). CT-FM identified high confidence causal cell types consistent with known biology, such as glutamatergic neurons for schizophrenia (SCZ) [43,44] (PIP = 0.98), B cells and/or T cells in multiple immune-related diseases, cardiomyocytes in atrial fibrillation [45] (PIP = 0.99), pancreatic cells for type 2 diabetes [46] (PIP = 0.51) and prostate epithelial cells for prostate cancer [47] (PIP = 0.99) (**Fig. 4c**); new biological insights for individual psychiatric and immune-related diseases is provided in the next section.

To evaluate the h^2 captured by CTS SNP-annotations identified by CT-FM, we constructed 81 SNP-annotations representing the union of the CTS SNP-annotations within each CS (see **Methods**). These SNP-annotations captured on average 2.7% of common SNPs, had h^2 enrichment of 14.1 ± 0.6x and explained a third (33.9 ± 1.5%) of trait h^2 (results meta-analyzed across the 81 SNP-annotations; **Fig. 4d** and **Supplementary Table 13**). After combining SNP-annotations for traits where CT-FM identified more than one CS (see below), we observed that 39.4 ± 1.9% of trait h^2 was concentrated in CTS SNP-annotations, which represents around 2/3 of the h^2 explained by SNPs in existing cREs (i.e., cREs from ENCODE4, ABC, CATLas, EpiMap [26] and DHS [28], as well as fine-mapped eQTLs from GTeX [48,49]; 38.9 % of common SNPs explaining 60.6 ± 1.5% of trait h^2) (**Supplementary Table 14**). Highest enrichments were observed for blood and immune SNP-annotations (21.3 ± 2.1x when meta-analyzed across 8 immune-related diseases, and 25.0 ± 3.3x when meta-analyzed across 8 blood-related traits), while low enrichments were observed for brain SNP-annotations tend to have lower enrichments (7.8 ± 0.9x when meta-analyzed across 6 psychiatric diseases); these results either highlight different functional architectures between immune- and brain-related traits, a higher informativeness of functional data in immune cell types that are available for a wide range of conditions (i.e., *in vitro* stimulations; 552 SNP-annotations analyzed in this study),

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

a lower informativeness of functional data in brain cell types (often collected in pre-natal or post-mortem tissues), and/or higher confounding in GWAS results of brain-related traits [50].

CT-FM identified multiple CSs with distinct causal cell types for 25% of the traits (16 out of 63), such as osteoblast (PIP = 0.99) and fibroblast (PIP = 0.89) in height [51], and radial glial cells (PIP = 0.91) and liver hepatocytes (PIP = 0.99) in body mass index (BMI) [52]. It also highlighted multiple CSs for psychiatric diseases (bipolar disorder [53] and SCZ [54]) and immune-related diseases (multiple sclerosis [55], inflammatory bowel disease [56] and rheumatoid arthritis [57]) (see next section). We validated that each CS corresponds to one independent causal signal by performing S-LDSC analyses where we jointly analyze SNP-annotations corresponding to each CS conditionally to each other (**Supplementary Table 15**).

We next investigated whether traits with independent causal cell types tend to have causal variants acting jointly on multiple cell types. To test this hypothesis, we focused on the 16 traits where CT-FM identified multiple CSs and investigated whether the h^2 explained by SNPs within multiple CSs was higher than expected. Specifically, for each trait, we ran S-LDSC with a model containing one SNP-annotation for each of the identified CSs, and one interaction SNP-annotation that exclusively contained SNPs overlapping multiple CSs (see **Methods**). The 16 interaction SNP-annotations captured on average 1.1% of common SNPs, were expected to have a h^2 enrichment of 16.7 ± 1.4x under a model with no interaction, and had an observed h^2 enrichment of 15.3 ± 1.4x with the interaction (P = 0.12 for interaction) (**Supplementary Table 16**). Overall, these results suggest that, even if SNPs falling within cREs of two causal cell types have a higher probability to be causal, it is more likely that these SNPs will impact disease risk through one of these cell types than through both cell types.

Finally, we evaluated the relevance of the three sources of cREs used in this study. Out of the 59 high-confidence causal cell types, 36 were coming from ENCODE4, and 23 from CATlas. Next, we replicated CT-FM analyses by leveraging CTS SNP-annotations from each sources independently. We identified 66, 57 and 19 CSs and 47, 52 and 19 high-confidence causal cell types for ENCODE4, CATlas and ABC sources respectively, which were consistent with main CT-FM results (**Supplementary Fig. 6** and **Supplementary Table 17**). These results demonstrate the benefits of leveraging CTS SNP-annotations from different sources.

Overall, CT-FM allowed us to refine the number of candidate causal cell types per trait (~20x decrease compared to S-LDSC), identified highly-confident causal cell types corresponding to known and novel cellular mechanisms, and confirmed the large fraction of h^2 explained by relevant CTS SNP-annotations (39.4 ± 1.9% of trait h^2). Finally, for 16 traits with

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

multiple independent causal cell types, we did not find evidence that genetic variants impact disease risk or complex trait variability by jointly targeting multiple cell types.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

CT-FM refines cellular mechanisms of psychiatric and immune-related diseases

We focused on CT-FM results on 6 independent psychiatric diseases and 8 independent immune-related diseases (**Fig. 4c**). We note that among the 927 SNP-annotations analyzed by CT-FM, 71 corresponded to brain cell types, and 552 to immune/blood cell types, and that the major histocompatibility complex (MHC) locus was removed from all analyses.

Within psychiatric diseases, CT-FM detected 3 CSs for SCZ [54] (highest number in this study), significantly decreasing the initial association signal of S-LDSC (487 associated SNP-annotations with FDR $P < 0.05$; **Fig. 4f** and **Supplementary Table 18**), and providing new insights into underlying SCZ cellular mechanisms. Two CSs corresponded to fetal and adult excitatory neurons (PIP = 0.99 and 0.98, respectively; both SNP-annotations came from single-cell ATAC-seq data [29]). While the role of excitatory/glutamatergic excitatory neurons in SCZ has been previously highlighted using gene expression datasets [5,6], detecting two independent CSs within fetal and adult samples would suggest either that different gene regulation mechanisms in excitatory neurons impacting SCZ risk at early- and late-life stages (as suggested in ref. [58]), either that both fetal and post-mortem adult brain samples are needed to characterize gene regulation in brain living samples [59]. The two corresponding SNP-annotations (1.05% and 1.26% of common SNPs, respectively) were highly enriched in SCZ h^2 (12.8 ± 1.4x and 12.7 ± 1.4x, respectively), and remained highly enriched when keeping SNPs specific to their CS and not present in other CSs (8.9 ± 2.3x and 10.9 ± 1.6x, respectively), confirming the needs of both fetal and post-mortem adult brain samples to dissect SCZ risk variants (**Supplementary Table 19**). The third CS corresponded to an ENCODE4 annotation of liver hepatocytes derived from H9 cell line (PIP = 0.99). While epidemiological studies have reported a risk of developing liver-mediated diseases among patients with SCZ [60,61], few genetic studies have discussed the impact of risk variants through hepatocytes and liver gene regulation. To further confirm this result, we created a hepatocyte SNP-annotation where we removed SNPs overlapping the SNP-annotations of the two excitatory neuron CSs; this SNP-annotation contained 0.80% of common variants, and had an enrichment of 6.82 ± 1.22x in SCZ h^2, demonstrating the independent action of variants within hepatocyte regulatory elements in SCZ risk. We also replicated a significant effect of the hepatocyte SNP-annotation (conditioned to the two excitatory neuron CSs SNP-annotations) in an independent East-Asian GWAS (one-sided $P = 5.7 \times 10^{-3}$) (**Supplementary Table 20**). These results could summarize the link between SCZ and the level of C-reactive protein (a protein synthesized by the liver in response to inflammation) [62,63]. Indeed, a recent C-reactive protein GWAS study highlighted an enrichment

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

of associated variants in liver SNP-annotations and supported a causal association with SCZ [63]. In addition, we observed that the significant enrichment of SCZ h^2 in immune-related SNP-annotations was not significant after conditioning on the hepatocyte SNP-annotation (**Supplementary Fig. 7**), suggesting that the link between immunity and SCZ risk is mediated by hepatocytes. However, we note that these findings could also reflect a selection bias in patients with SCZ (for example by selecting only patients that do not respond to existing drugs through different genetics of gene regulation in the liver) or . Altogether, our results show that SCZ risk variants are independently enriched within SNP-annotations from fetal and post-mortem excitatory neurons and from liver hepatocytes, suggesting different cellular causal mechanisms.

CT-FM also highlighted adult glutamatergic neurons (PIP = 0.99) and liver hepatocytes (PIP = 0.86) for bipolar disorder [53], which has a high genetic correlation [64] with SCZ (r_g = 0.71 +/- 0.02). Interestingly, the fetal glutamatergic neuron SNP-annotation was selected by CT-FM for SCZ but not for bipolar disorder, consistent with recent reports [65,66]. For other psychiatric diseases, CT-FM identified excitatory neurons (either from fetal and/or adult samples) for insomnia [67] (cPIP = 0.47), depression [68] (PIP = 0.45), and attention-deficit/hyperactivity disorder [69] (PIP = 0.58), consistent with published literature [5,6], and inhibitory neurons for anorexia nervosa [70] (PIP = 0.64), which has been previously discussed but remains unelucidated [71,72].

When applied to 8 immune-related diseases, CT-FM was able to recapitulate known findings, such as the role of CD4+ T cells in asthma [73] (cPIP = 0.61), celiac disease [74] (cPIP = 0.55), and type 1 diabetes [75] (PIP = 0.70), the role of B cells in lupus [76] (cPIP = 0.42) and primary biliary cirrhosis [77] (cPIP = 0.66), the role of both CD4+ T cells and B cells in rheumatoid arthritis [57] (cPIP = 0.50 and PIP = 0.95, respectively), and the role of both CD4+ T cells and macrophages in multiple sclerosis [55] (PIP = 0.90 and 0.99, respectively). Interestingly, CT-FM was able to identify two CSs for multiple sclerosis, while previous approaches highlighted enrichments in a variety of immune cell types but were unable to distinguish between them [5,6]. Similarly, for rheumatoid arthritis, previous works have mainly characterized heritability enrichment within CD4+ T cells [78–80], while CT-FM was able to also detect a role of variants in regions active in B cells. Indeed, B cells play critical roles in rheumatoid arthritis pathogenesis by secreting physiologically important proteins such as rheumatoid factors, anti-citrullinated protein antibodies and pro-inflammatory cytokines [81]. CT-FM results on an alternative rheumatoid arthritis GWAS dataset [79] are discussed in **Supplementary Fig 8**.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Finally, CT-FM identified two CSs for inflammatory bowel disease [56] (IBD): the first CS corresponded to monocyte-derived cells (max PIP = 0.63 for macrophages), the second CS corresponded to various types of T cells (cPIP = 0.29 and 0.44 for CD4+ and CD8+ T cells, respectively). Interestingly, when performing additional analyses on Crohn's disease (CD) and ulcerative colitis (UC) [56] (two main subtypes of IBD), CT-FM identified one of these CS for each subtype: monocyte-derived cell types for CD (max PIP = 0.98 for macrophages), and T cells for UC (PIP = 0.59 and 0.10 for CD4+ and CD8+ T cells, respectively) (**Fig. 4c**). Our findings are consistent with several previous reports implicating macrophages and T cells in both CD and UC [82,83]. Interestingly, drastically elevated levels of monocytes were observed in patients with CD but not UC, suggesting that monocytes may play a more important role in the etiology of CD [84,85]; several studies report elevated/decreased levels of various T cell subsets and their effects in IBD patients compared with healthy controls, with differences more pronounced in UC rather than CD patients [86,87].

Altogether, CT-FM identified multiple causal cell types for psychiatric and immune-related diseases. For SCZ, it identifies distinct roles of fetal and adult excitatory neurons, suggesting different gene regulation mechanisms in excitatory neurons impacting SCZ risk at early- and late-life stages [58], and hepatocytes, consistent with the role of C-reactive protein in SCZ risk [62,63]. For IBD, CT-FM identified two CSs, each corresponding to the CS identified by CT-FM for UC and CD, consistent with a model where diseases are biologically heterogeneous, and where different disease subtypes are impacted by genetic risk variants through different cell types.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Applying CT-FM-SNP to 39 UK Biobank complex traits

We successfully applied CT-FM to 39 UK Biobank traits (53 CSs; **Supplementary Table 21**), and CT-FM-SNP to 6,975 candidate causal {non-coding SNP, trait} pairs (5,819 unique SNPs) of these traits identified via SNP-fine-mapping [33,34]. CT-FM-SNP was able to assign at least one CS to nearly half (3,154/6,975; ~46%) of the pairs (3,840 CSs in total), detecting a high confidence causal cell types for most of them (2,798/3,154; 89%) (**Fig. 5A** and **Supplementary Table 22**). We identified expected high confidence causal cell types for a high number of {non-coding SNP, trait} pairs, such as CD4+ T cells for eczema candidate SNPs (24 SNPs with PIP > 0.5 for CD4+ T cells among 40 analyzed, average PIP = 0.90), erythroblasts for mean corpuscular hemoglobin candidate SNPs [88] (98/198, average PIP = 0.94) and osteoblasts for bone density candidate SNPs [89,90] (76/241, average PIP = 0.94) (**Fig. 5B**). While high confidence causal cell types detected by CT-FM-SNP were consistent with CT-FM results, CT-FM-SNP identified trait-relevant cell types for a smaller fraction of variants, such as vascular smooth muscle cells (VSMCs) for systolic blood pressure (7/70, average PIP = 0.71) and bone marrow stromal cells (BMSCs) for height (25/417, average PIP = 0.77).

To further predict causal {non-coding SNP, cell type, gene, trait} quadruplets, we leveraged predictions from both CT-FM-SNP and SNP-gene links from our cS2G strategy [91] and predicted 2,751 causal quadruplets (**Supplementary Table 23**). To validate these predictions, we performed gene ontology enrichment analyses of genes assigned to the same cell types across the 39 traits [92] (**Fig. 5c** and **Supplementary Table 24**). We identified expected enrichments, such as enrichment of bone- and osteoblast-related processes for genes linked to SNPs targeting the osteoblast cell type (minimum FDR $P = 1.8 \times 10^{-10}$), enrichment of immune- and T-cell related processes for genes linked to SNPs targeting T cells (minimum FDR $P = 2.4 \times 10^{-16}$), and enrichment of epithelial-related processes for genes linked to SNPs targeting epithelial cells (minimum FDR $P = 6.1 \times 10^{-7}$). These results underline the capacity of CT-FM-SNP and cS2G to perform {non-coding SNP, cell type, gene, trait} quadruplet predictions consistent with known biological mechanisms.

We report quadruplets consistent with known biology or highlighting new biology in **Table 1**. Specifically, for the type 2 diabetes candidate causal variant rs571342427, we inferred pancreatic islet cells and pancreatic beta cells as two high confidence cell types, and the insulin gene (*INS*) [93,94] as a target gene. For the bone density candidate causal variant rs10130587, we inferred osteoblasts as a high confidence cell type, and the bone morphogenetic protein 4 gene (*BMP4*; a gene previously implicated in body stature [95,96], involved in bone formation and

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

stimulation of osteoblast differentiation [97–99]) as a target gene. Finally, for the neuroticism candidate causal variant rs191480627, we inferred hepatocyte cell type as a high confidence cell type, and *AGPAT3* as a target gene. These results are consistent with our previous findings suggesting the link between hepatocytes and C-reactive protein in other psychiatric disorders (see above), and with the role of the hepatic *AGPAT3* gene in regulating levels of polyunsaturated fatty acids (PUFAs) [100], as low levels of PUFA have been previously linked to the pathogenic mechanisms underlying psychiatric disorders with limited literature support [101,102] and to the regulation of C-reactive protein levels [103–105]. Additional details of quadruplets highlighted in **Table 1** are provided in the **Supplementary Note**.

CT-FM-SNP identified 668/3,154 SNPs with at least two CSs. In practice, it is very challenging to infer if a regulatory SNP overlapping cRE from diverse cell types impacts disease risk by acting on only one of these cell types, or by acting jointly on multiple cell types. Here, we leveraged CT-FM-SNP results to tackle this question. First, we validated that CT-FM-SNP does not overestimate the number of SNPs assigned to at least two CSs by comparing its results with h^2 results estimated by S-LDSC. We focused on 7 UK Biobank traits for which CT-FM identified two CSs and CT-FM-SNP identified at least one CS for at least 50 SNPs; CT-FM-SNP identified 736 SNPs assigned to one or two cell types previously identified by CT-FM (**Supplementary Fig. 9**). On average, 37 ± 5% of these SNPs were assigned by CT-FM-SNP to both CT-FM CSs (285 SNPs in total), which was consistent with the fraction of h^2 explained by SNPs in the intersection of CT-FM CSs by the h^2 explained by SNPs in the union of CT-FM CSs (**Supplementary Fig. 10, Supplementary Table 25 and Supplementary Table 26**). Second, for the same set of traits, we observed that the number of SNPs assigned to both CT-FM CSs (285) was smaller than what is expected under a model where SNPs in both CT-FM CSs impact disease risk through a single cell type (395, more information in **Supplementary Table 27** caption). This result suggests that, among SNPs where CT-FM-SNP identified two CSs, it is extremely likely that most of them impact disease risk through a single cell type.

Next, we investigated whether a SNP impacting multiple traits (i.e., pleiotropic SNP), tends to target the same or different cell types. We analyzed 104 candidate causal SNPs for which CT-FM-SNP identified a candidate causal cell type in at least two genetically uncorrelated UK Biobank traits (**Supplementary Table 28**). We found that approximately half of these pleiotropic SNPs (46/104; 44%) shared at least 1 candidate cell type across different traits (**Fig. 5c**). Shared {non-coding SNP, cell type} pairs across traits included SNPs acting on hematopoietic multipotent progenitors (HMP) in blood/immune traits and SNPs acting on

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

osteoblasts and bone marrow stromal cells in height and bone density traits (**Fig. 5d**). Strikingly, more than half of the pleiotropic SNPs (58/104; 56%) were assigned to different cell types by CT-FM-SNP (**Fig. 5e**). For example, the pleiotropic SNP rs1887428 was assigned to HMP and erythroblast cell types for mean corpuscular hemoglobin, but to megakaryocytes for mean platelet volume; this result is consistent with the known role of its candidate target gene *JAK2* on the regulation of both erythroid progenitors [106] and megakaryocyte morphology [107]. The pleiotropic SNP rs3918226 was assigned to fibroblasts, endothelial and CD4+ T cells for height, diastolic blood pressure and eczema, respectively; this result is consistent with the known role of its candidate target gene *NOS3* (ref. [91]), which is produced in endothelial cells and plays a key role in regulating blood pressure [108,109], but was also shown to regulate CD4+ T cell production [110,111] and collagen synthesis in fibroblasts [112]. Finally, the pleiotropic SNP rs75475627 was assigned to osteoblasts and bone marrow stromal cells for bone density trait, and to macrophages for monocyte count trait; this result is consistent with the known role of its candidate target gene *SPTBN1*, which is involved in osteoblast differentiation and function [113,114], but was also shown to regulate monocyte differentiation into macrophages [115].

Altogether, CT-FM-SNP results shed light on how variants individually impact disease risk. We predicted 2,751 {non-coding SNP, cell type, gene, trait} quadruplets, representing one of the largest catalogs of this type to date. Finally, while we observed that most individual SNPs impact a phenotype through a single cell type, pleiotropic SNPs might target different cell types depending on the phenotype context.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

We developed CT-FM and CT-FM-SNP, two methods that leverage CTS cRE annotations to fine-map causal cell types for a trait and for its candidate causal variants, respectively. These methods were validated using simulations and application to five blood cell traits that biologically correspond to specific immune cell types. By applying CT-FM to 63 GWASs, we identified 81 CSs with corresponding SNP-annotations explaining a high fraction of trait h^2 (39.4 ± 1.9%, which corresponds to ~2/3 of the h^2 explained by existing cREs), identified 16 traits with multiple causal cell types, and identified cell-disease relationships consistent with known biology and highlighted previously unexplored cellular mechanisms involved in psychiatric and immune-related diseases. Notably, for SCZ, we identified three credible sets corresponding to fetal and adult excitatory neurons and hepatocytes. For IBD, we identified two credible sets corresponding to macrophages and T cells; interestingly, each of these cell types were inferred as causal for CD and UC, respectively, suggesting disease heterogeneity for diseases with multiple causal cell types. By applying CT-FM-SNP to 39 UK Biobank traits, we suggest that most individual SNPs impact a phenotype through a single cell type (as also suggested by h^2 analyses), and that pleiotropic SNPs might target different cell types depending on the phenotype context.

Compared to existing methods, CT-FM and CT-FM-SNP propose several conceptual advances. First, it fine-maps cell types of diseases and risk variants from cREs rather than gene expression QTLs datasets [14,20], allowing to capture more disease h^2. Indeed, the h^2 explained by the cREs identified by CT-FM is nearly four times higher than the h^2 explained by fine-mapped eQTLs [22] or mediated by gene expression [23] from all GTeX tissues. Second, employing statistical fine-mapping on S-LDSC results of cRE SNP-annotations allows disentangling shared regulatory patterns across cell types, greatly reducing the number of resulting candidate SNP-annotations (~20x decrease). For example, CT-FM was able to identify two CSs for multiple sclerosis, while previous approaches highlighted enrichments in a variety of immune cell types but were unable to distinguish between them [5,6]. Third, CSs identified by CT-FM allow distinguishing between co-regulated and conditionally independent causal cell types. In SCZ, we were able to identify two independent causal signals in brain and excluded immune cell types which co-regulated with hepatocytes. Finally, CT-FM-SNP is (to our knowledge) the first method prioritizing causal cell type of individual variants by leveraging both polygenic enrichment and co-regulation within cREs.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Our findings have several implications for downstream analyses. First, it highlights the benefits of leveraging cREs to infer causal cell types of diseases and risk variants. Their advantages include a high fraction of h^2 explained, cost effective experiments targeting multiple cell types and conditions (ENCODE4 provides candidate *cis*-regulatory elements for 1,680 cell types and cell lines; single-cell ATAC-seq technologies are now allow to infer cREs specific to a cell state [80]), and high potential for linking rare variants to cell types. Second, our application of CT-FM on 63 GWASs uncovered previously unexplored cellular mechanisms in psychiatric and immune-related diseases. Specifically, identifying hepatocytes as a causal cell type for SCZ was consistent with the link between SCZ risk variants and C-reactive protein reported in ref. [63]; identifying genetic variants impacting SCZ risk through hepatocytes could lead to new candidate disease genes and drug targets. Third, our results are consistent with a model where diseases are biologically heterogeneous, and where different disease subtypes are impacted by genetic risk variants through different cell types. For example, CT-FM identified two CSs for IBD, each corresponding to the CSs identified for CD and UC (macrophages and T cells, respectively). Identifying multiple causal cell types for a trait could thus allow constructing CTS polygenic risk scores to build distinct profiles of patients [116]. However, we underline that these results do not indicate a disease-specific action of different cell types but rather their more pronounced role in different subtypes which involves a complex interplay between different causal cell types. Finally, CT-FM-SNP provides insights into functions of disease candidate variants and may also be helpful for choosing the relevant cell type for *in vitro* experiments.

We note several limitations of our work, related to CT-FM and CT-FM-SNP power to perform causal inferences. First, CT-FM and CT-FM-SNP specifically assume that all disease-relevant cell types have corresponding SNP-annotations, while they might not have been assayed in practice. This is consistent with the observation that cREs are available for a wide range of immune cell types and conditions (i.e., *in vitro* stimulations; 552 SNP-annotations analyzed in this study) and presented the largest h^2 enrichments (21.3 ± 2.1x when meta-analyzed across 8 immune-related diseases vs. 7.8 ± 0.9x when meta-analyzed across 6 psychiatric diseases). Generating cRE catalogs in diverse cell types and conditions should fill the gap between the total trait h^2, and the h^2 explained by SNP-annotations from putative causal cell types. Nevertheless, CT-FM and CT-FM-SNP can still indicate the “best proxy” for the truly causal cell type. Second, as highlighted in simulations, power to detect disease causal cell types depends on GWAS sample size and trait h^2, which product needs to be large. While CT-FM has been successful at detecting at least one CS on 63 GWASs and at least two CSs for 16 traits, it still might have missed some causal cell types due to lack of statistical power. However,

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

we detected similar causal cell types with a height GWAS performed in the UK Biobank (*N* = 460K) than with a height GWAS with a significantly larger sample size (*N* = 1.6M; ref. [51]), suggesting that there should not be many causal cell types for a complex trait. In addition, we highlighted that CT-FM-SNP can detect relevant cell types that have not been highlighted by CT-FM. Third, another limitation related to sample size is that we were unable to analyze GWAS of non-European ancestries, which have significantly lower sample size than GWAS of European ancestry [117]. Specifically, we extended our CT-FM approach to analyze 31 GWASs of East-Asian ancestry [118], and identified only five traits well-powered GWAS (i.e., at least 1 S-LDSC marginal $\tilde{\tau}$ Z-score > 4) (**Supplementary Table 29**). However, despite sample size differences of these traits between East-Asian and European GWASs (average *N* = 111K vs *N* = 795K), CT-FM results were consistent across ancestries (**Supplementary Table 30**), confirming similar CTS genetic architectures across these two ancestries [79]. Fourth, CT-FM power is also reduced when cell types highly co-regulate with each other. Credible sets inferred by SuSiE account for uncertainty in causal cell type selection but are unable to distinguish how many causal signals are contained within a CS. For example, in many immune-related diseases, we identified CS related to T cells and containing both CD4+ and CD8+ T cell SNP-annotations, without being able to dissociate the cases where only one or both of them are causal. Improved cRE definition, such as dynamic regulatory elements defining precise cell types or cell states [80], should improve CT-FM ability to identify precise causal cell types or cell states. Finally, CT-FM-SNP relies on a list of candidate SNPs identified via SNP-fine-mapping, whose power is still limited [33].

Despite these limitations, our results demonstrate the advantage of CT-FM and CT-FM-SNP to infer independent causal cell types from GWASs, providing novel insights into the complex cell biology underlying the genetic architecture of human diseases and complex traits.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

CT-FM method

CT-FM is a method that fine-maps the causal cell types of a trait. It takes as inputs GWAS summary statistics with a matching LD reference panel, and a set of CTS SNP-annotations (**Fig. 1**). Below we describe its model and parameters estimation.

We assume the infinitesimal linear model $Y = X\beta + \varepsilon$, where Y is a vector of quantitative phenotypes, X is a matrix of standardized genotypes, β is the vector of per-normalized-genotype effect size, and ε is a mean 0 vector of residuals. We model the variance of the causal effect size β_j of each variant *j* as an extension of S-LDSC baseline model [4,6]. Specifically, we model

$$Var(\beta_j) = \Sigma_{b\in B} a_b(j)\tau_b + \Sigma_{c\in C} a_c(j)\tau_c,$$

where *B* represents a set of background SNP-annotations from the baseline model (i.e., coding, enhancer, promoter, …), $a_b(j)$ is the indicator variable of variant *j* for SNP-annotation *b*, $\tau_b$is the contribution of *b* to $Var(\beta_j)$, and *C* represents the whole set of CTS SNP-annotations. Rather than assuming that all *C* SNP-annotations contribute to $Var(\beta_j)$, we assume that at most L SNP-annotations contribute positively to $Var(\beta_j)$ (i.e., $\tau_c > 0$). Specifically, we model the vector τ of $\tau_{c\in C}$ as $\tau = \sum_{l=1}^{L} \tau_l$ as the sum of L single-effects $\tau_l = \gamma_l \cdot b_l$ where γ_l is a $C \times 1$ binary vector indicating which CTS SNP-annotation is causal for the lth effect, and b_l is a scalar quantifying the contribution of the causal l^{th} effect to $Var(\beta_j)$. We caution that CT-FM specifically assumes that all disease-relevant cell types have corresponding SNP-annotations among *C*, while they might not have been assayed in practice (see **Discussion**).

In practice, we obtain marginal effects $\tilde{\tau}_c$ by applying S-LDSC on GWAS χ^2 summary statistics from the following model:

$$E[\chi_j^2] = N\Sigma_{b\in B}\tau_b l(j,b) + N\tilde{\tau}_c l(j,c) + Nd + 1$$

where N is the GWAS sample size, d is a term measuring the contribution of confounding biases [119] and $l(j,b) = \Sigma_k a_b(k) r_{jk}^2$ is the LD score of SNP *j* with respect to the value of SNP-annotation a_b and r_{jk}^2 is the correlation between SNPs *j* and *k.* We can then model the vector $\tilde{\tau}$ of $\tilde{\tau}_{c\in C}$ as $\tilde{\tau} \sim N(R\tau, R)$, where τ is the sum of single effects from above, and R is the $C \times C$ co-regulation matrix defined as the residual correlation matrix of LD scores

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

$$R = {A_C}^t A_C - {A_C}^t A_B ({A_B}^t A_B)^{-1} {A_B}^t A_C$$

where A_B and A_C are the SNP x B and SNP x C SNP-annotation matrices of the LD scores of the baseline model SNP-annotations and CTS SNP-annotations, respectively. After estimating the marginal effects $\tilde{\tau}_c$ and their corresponding Z-scores using the S-LDSC framework [30,35,36], and computing the co-regulation matrix, we can estimate the number of single effects $l \in L$ and corresponding γ_l and b_l values from the SuSiE-RSS function from the SuSiE-R package [38]. In practice SuSiE-RSS will estimate the PIP of each SNP-annotation c by modeling γ_l as $Multi(1, \alpha_l)$, where α_l is the probability of each CTS SNP-annotation to explain a single l^{th} effect, and by defining PIP_c as

$$PIP_c = P(\tau_c \neq 0 \,|\tilde{\tau}, A_C) = 1 - \prod_{l=1}^{L}(1 - \alpha_{l,c})$$

We set SuSiE parameters *L* to 10 (default) and *n* to 1,190,321 matching the number of SNPs used by S-LDSC. Of note, varying these parameters did not change the results (identical PIP values and CSs).

We next compute η-CSs for each l, where η represents the desired probability that the set contains causal CTS SNP-annotations (0.95 in this study). To infer a CS for each l, we decreasingly sort α_l and take a greedy approach to include CTS SNP-annotations until their cumulative sum exceeds η.

CT-FM then applies several steps to refine the CS inferences. First, to remove any potential false positive CSs, it removes CSs in which no CTS SNP-annotations present a high S-LDSC association signal (S-LDSC Z-score > 4), criteria similar to our definition of a 'well-powered' GWAS. Second, we introduce several modifications to the default SuSiE-RSS approach to detect potentially missed CSs (i.e., false negatives). By default, SuSiE applies the "purity" approach which removes CSs whose minimum pairwise correlation R between its components is less than a given threshold (0.5 by default). However, this pruning method may falsely remove CSs containing two causal signals exhibiting equal effect size but presenting low pairwise correlation R. To remedy this, we heuristically introduce a new criterion "divergence" measuring the Kullback–Leibler (KL)-divergence between the distribution of α_l values and the null distribution for each l. We keep the CSs whose either purity is greater than 0.3 (threshold corresponding to the average absolute pairwise correlation between two CTS SNP-annotations in our dataset) or the KL divergence is greater than 3. The KL divergence of a CS significantly correlated with the maximum S-LDSC Z-score observed in a CS (r^2 = 0.79, **Supplementary Fig. 11**), and the threshold of 3 was empirically defined as it optimizes the selection of well-

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

powered CSs (defined as CSs containing at least one CTS SNP-annotation with a S-LDSC Z-score > 4). Third, in very rare cases, SuSiE-RSS can output two CSs both containing a "duplicate" CTS SNP-annotation (i.e., present in both CS1 and CS2). To address these particular cases, we remove the CSs containing such duplicates if the sum of their PIP values exceeds 0.1. In practice, only 2 of the 135 analyzed CSs contained duplicate CTS SNP-annotations (1/53 for 39 UK Biobank traits and 1/82 for 63 GWASs analyzed in this study), with only 1 case where we removed the CS as the sum of the corresponding PIP values exceeded 0.1 (second GWAS dataset of Rheumatoid arthritis, see **Supplementary Fig. 8**). Finally, for CSs where no highly-confident (PIP > 0.5) causal cell type was detected but multiple CTS SNP-annotations correspond to the same cell type, we calculate a combined PIP (cPIP) defined as

$$cPIP_{ct,l} = 1 - \prod_{cts \in l} (1 - \alpha_{l,cts})$$

where ct, l is the cell type in a credible set l, and $\alpha_{l,cts}$ are the α values of the CTS SNP-annotations *cts* present in the credible set l and corresponding to the cell type ct (8 cases of cPIP > 0.5 calculated across 63 GWASs, **Supplementary Table 10**).

CT-FM-SNP method

While CT-FM was designed to infer causal cell types underlying the overall GWAS signal, CT-FM-SNP is a method that fine-maps the causal cell types targeted by a particular causal SNP. It takes as inputs GWAS summary statistics with a matching LD reference panel, a set of CTS SNP-annotations, and a list of candidate causal SNPs of the trait, and outputs CSs and PIPs for those SNPs (**Fig. 1**). Briefly, for each analyzed SNP, we restrict CT-FM model to the subset of cRE annotations overlapping the SNP and perform the same inference and filtering of CSs as described above. We additionally retain the results of SNPs that present a high confidence causal signal (PIP > 0.5) for a well-powered CTS SNP-annotation (S-LDSC Z-score > 4) even if they were not identified in a CS (67 such SNPs in total among 3,154 analyzed, **Supplementary Table 19**). We note that while the inference of multiple CSs by CT-FM suggests that multiple cell types are causal for the disease (as it relies on polygenicity), the inference of multiple CSs by CT-FM-SNP suggests that the candidate SNP overlap cREs from multiple causal cell types, but does not infer whether the candidate SNP impacts disease risk by disrupting gene regulation within a single or multiple cell types.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Selection of 927 CTS annotations for CT-FM and CT-FM-SNP

Our dataset of 927 CTS SNP-annotations (1.1% of common SNPs on average) was derived from three catalogs of cREs: ENCODE4, CATlas and ABC (average genome coverage = 32.5 Mb). For ENCODE4, we manually curated the dataset of 1,680 cell-type-specific cREs by primarily excluding cREs corresponding to tissues and cancer cell lines. For each resulting cRE file, we excluded marks corresponding to "*low DNase" and* "*Unclassified"* state. ENCODE4 data was converted from hg38 to hg19 using the *liftOver* tool from UCSC [120]. For ABC, we retrieved a dataset of gene-enhancer predictions for 131 tissues/cell types and excluded all tissue-level annotations. CATlas data (epigenomic annotations of 222 cell types) was retrieved and converted to hg19 (liftOver) without any additional filtering. The resulting bed files were homogenized (*sortBed* function of *bedtools* [121]) and checked for errors (*bedops --ec --merge* function of *BEDOPS* [122]). For each dataset, we additionally create one baseline annotation corresponding to the union of all candidate regulatory regions across all cell types (see *CT-FM analyses* in the Methods section).

Simulations

To assess CT-FM power under realistic scenarios, we simulated GWAS summary statistics for the 1,187,349 HapMap 3 SNPs used by S-LDSC. We considered one scenario with a single causal cell type (the osteoblast SNP-annotation identified in CT-FM first CS), and one scenario with two causal cell types (the osteoblast and fibroblast SNP-annotations identified by CT-FM CSs). First, we used S-LDSC regression coefficients estimated on height GWAS using the baseline model SNP-annotations and the osteoblast SNP-annotation for the first scenario, and using the baseline model SNP-annotations, the osteoblast SNP-annotation, and the fibroblast SNP-annotation for the second scenario. Then we computed expected per-SNP h^2 for every SNPs using the corresponding annotations. Second, we randomly selected 10,000 causal SNPs (same order of magnitude as the number of causal SNPs per GWAS in ref. [33]), and simulated effect sizes proportional to the per-SNP h^2 computed above while setting h^2 to 0.5. Because expected per-SNP h^2 can be negative, we initially set these values to 0, and rescaled positive per-SNP h^2 so that expected h^2 enrichment of each annotation in the model was similar to the ones observed on the height GWAS. Third, we simulated a vector of GWAS Z-scores using a multivariate normal distribution (*mvrnorm R* function), an LD matrix estimated on 337,491 unrelated British-ancestry individuals from UK Biobank release 3 (ref. [33]), and a GWAS sample size of 100K, 350K (main simulation), and 1M. 500 simulations were performed for each

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

scenario and sample size (4,500 simulations in total). Fourth, we ran CT-FM using 200 CTS SNP-annotations with varying degrees of residual correlation R with the osteoblast SNP-annotation in the first simulation scenario, and added 200 CTS SNP-annotations with varying degrees of residual correlation R with the fibroblast SNP-annotation in the second simulation scenario. Specifically, for the osteoblast cell type, we selected 40 CTS SNP-annotations presenting a weak correlation with the causal SNP-annotation (R between -0.11 and 0.10), 80 CTS SNP-annotations presenting a low-to-moderate correlation with the causal SNP-annotation (R between 0.10 and 0.32), 40 CTS SNP-annotations presenting a moderate-to-high correlation with the causal SNP-annotation (R between 0.32 and 0.45) and 40 CTS SNP-annotations presenting a high correlation with the causal SNP-annotation (R between 0.45 and 0.81). Similar selection was performed for additional 200 CTS SNP-annotations with varying degrees of correlation with the causal fibroblast CTS SNP-annotation (minimum R = -0.11; maximum R = 0.78; mean R = 0.12) (**Supplementary Table 31**).

CT-FM analyses

We successfully ran CT-FM on GWAS summary statistics of five blood traits from the UK Biobank [39], 63 independent GWAS summary statistics, and CD and UC summary statistics [56] (see **Supplementary Table 9 and 20**). All GWASs were performed on individuals of European ancestry. S-LDSC Z-scores were obtained by running S-LDSC with the with the baseline v1.2 SNP-annotations; to make sure that our analyses capture CTS information we added six SNP-annotations corresponding to the union of all CTS ENCODE4, CATlas, ABC, EpiMap [26], and DHS [28] cREs, as well as GTeX fine-mapped eQTLs (SuSiE PIP > 0.05 in at least one tissue) [48,49]. We used Europeans from the 1000 Genomes Project as the reference panel [123].

To benchmark CT-FM, we first let vary SuSiE parameters *l* and *n* (default values set to *L*=10; and n=1,190,321 matching the number of SNPs used in S-LDSC model) on the 5 blood traits, and observed identical results (i.e., same CSs and PIP values for each trait). Second, we compared CT-FM results on the five blood traits while keeping all CTS SNP-annotations (including the ones with negative □, which were artificially set to 0) and while restricting the CTS SNP-annotations to those presenting a positive S-LDSC □ (default approach of CT-FM). We observed no significant differences on the resulting CT-FM PIP values (correlation = 0.99) and CT-FM credible sets (22/22 CTS SNP-annotations identified in CS by both methods) (**Supplementary Fig. 12**). Third, we ran SuSiE-inf [42] instead of SuSiE-RSS and observed a

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

lower number of CSs (3 with SuSiE-inf vs. 5 with SuSiE-RSS) and a lower number causal cell types with PIP ≥ 0.5 (3 with SuSiE-inf vs. 5 with SuSiE-RSS), arguing in favor of using SuSiE-RSS for cell-type fine-mapping. In particular, SuSiE-inf (and not SuSiE-RSS) failed to detect monocytes as the top causal cell type for the monocyte count trait (**Supplementary Table 5**). Finally, we applied a recommended fine-mapping diagnostic procedure [38,124] and observed a high concordance between expected and observed Z-scores across traits (mean r^2 = 0.95 across the 63 GWASs, **Supplementary Fig. 13**), guaranteeing reliable fine-mapping results by CT-FM.

S-LDSC analyses

To evaluate the h^2 captured by CTS SNP-annotations identified by CT-FM, we constructed 81 SNP-annotations representing the union of the CTS SNP-annotations within each CS, and estimated the h^2 they explained using S-LDSC with baseline-LD model v2.2 and the six additional functional SNP-annotations; we note that we followed the recommendation to use baseline model SNP-annotations for identifying critical cell types, and the baseline-LD model SNP-annotations for estimating h^2 enrichment [125]. To evaluate the h^2 captured by SNPs in existing cREs we merged the six additional functional SNP-annotations to a single annotation. Meta-analyses were performed using random-effects with the R package *rmeta*.

To investigate whether traits with independent causal cell types tend to have causal variants acting jointly on multiple cell types, we ran S-LDSC with the baseline model SNP-annotations, the six additional functional SNP-annotations, one SNP-annotation corresponding to each CS and one interaction SNP-annotation corresponding to the intersection of the multiple CSs. We meta-analyzed standardized effects □* of the interaction SNP-annotation, which is defined as □* = □ x M x sd / h^2, where M is the number of common SNPs in the reference panel, sd is the standard deviation of the SNP-annotation, and h^2 is the estimate of SNP-heritability from S-LDSC). We compared the h^2 enrichment of the interaction SNP-annotation under a model with no interaction (i.e., obtained from the □ estimated by S-LDSC with the baseline-LD model, the six additional functional SNP-annotations, and one SNP-annotation corresponding to each CS) and with interaction (i.e., estimated directly from S-LDSC by adding the interaction SNP-annotation to the previous model).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

CT-FM-SNP analyses

We ran CT-FM on the 49 UK Biobank traits analyzed in ref. [33] and identified at least one CS for 39 traits. We next leveraged functionally-informed SNP-fine-mapping results from PolyFun for these 39 traits [33], and identified 6,975 candidate causal {non-coding SNP, trait} pairs (5,819 unique SNPs); we note that no CTS SNP-annotations were included in PolyFun functional priors. We applied CT-FM-SNP to all of these pairs. To link non-coding SNP to their target genes, we leveraged predictions from our cS2G strategy [91] and retained one gene per SNP-trait pair by taking the gene with the highest cS2G score as defined in ref. [91]. GO enrichment analyses were performed using the *GOseq* R package [92] for cell types presenting at least 50 SNP-Cell-type pairs identified in this study. For the analyses presented in **Fig. 5c,** we regrouped cell types corresponding to T cells (such as CD4+ and CD8+ T cells) and cell types corresponding to epithelial cells (such as epithelial fibroblasts, choroid plexus epithelial cells and mammary epithelial cells). We retained all GO terms corresponding to the *Biological Process* category, containing between 10 and 1,000 genes, and containing at least 5 genes from the tested gene set. All FDR significant enrichments (FDR P < 0.05) are reported in **Supplementary Table 24**. For heritability analyses presented in **Supplementary Fig. 9**, we selected 7 UK Biobank traits with two identified CT-FM CSs and at least 50 SNPs successfully analyzed by CT-FM-SNP. For presentation purposes of CT-FM-SNP results on BMI, we leveraged the 2 CSs inferred by CT-FM on the BMI GWAS summary statistics of ref. [52] (used in *Applying CT-FM to 63 GWASs* section) rather than the 3 CSs inferred by CT-FM on UK Biobank BMI summary statistics (**Supplementary Table 21**). For analyses related to pleiotropic SNPs, genetically uncorrelated UK Biobank traits were defined by prioritizing UK Biobank traits with the highest SNP-heritability Z-score, and by removing traits with a squared genetic correlation [64] > 0.05.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Data availability

S-LDSC reference files and GWAS summary statistics used in this study are available at https://zenodo.org/records/10515792. S-LDSC CTS SNP-annotations used in this study are available at https://zenodo.org/records/11194201.

Code availability

CT-FM/CT-FM-SNP softwares and the code to replicate our analyses are available at https://github.com/ArtemKimUSC/CTFM.

Acknowledgments

We thank G. Lettre, P. F. Sullivan, P. Dieudé, M. Vujkovic, A. Vaissié and members of the Gazal and Mancuso labs for helpful discussions. We thank the ENCODE consortium and the International Multiple Sclerosis Genetics Consortium (IMSGC) for sharing multiple sclerosis GWAS summary statistics. This research has been funded by the National Institutes of Health grant R35 GM147789.

Competing interests

S.G reports consulting fees from Eleven Therapeutics unrelated to the present work. The other authors declare no competing interests.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figures

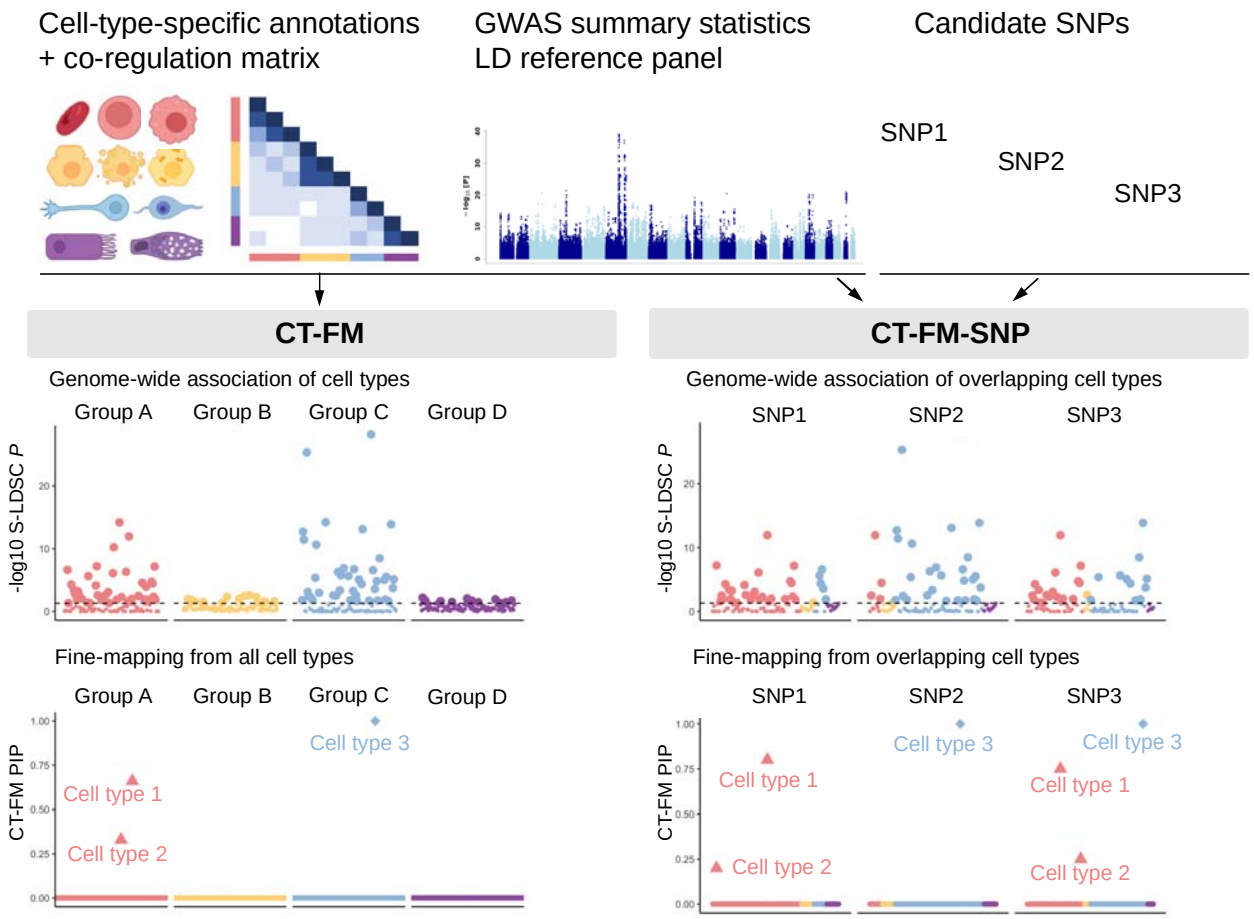


Figure 1. Overview of the CT-FM framework. CT-FM and CT-FM-SNP are two methods that fine-map the causal cell types of a trait and its candidate causal SNPs, respectively. They take as inputs a set of cell-type-specific (CTS) SNP-annotations, GWAS summary statistics with a matching LD reference panel, and (for CT-FM-SNP only) a list of trait's candidate causal SNPs. First, CT-FM estimates the significance of the marginal effect on SNP-heritability of each CTS SNP-annotation by applying stratified LD score regression (S-LDSC) [30,35,36]. Then, CT-FM fine-maps the causal cell types and outputs posterior inclusion probability (PIP) and credible sets (CSs) by leveraging S-LDSC Z-scores and a co-regulation matrix of CTS SNP-annotations. In our toy example, CT-FM reduces the number of SNP-annotations significantly associated with the trait to two CSs (triangle and diamond), each corresponding to a distinct biological group (A and C). CT-FM-SNP leverages the same workflow as CT-FM, in the difference that it restricts the fine-mapping step to CTS SNP-annotations that overlap the candidate SNP. In our toy example, CT-FM-SNP fine-maps the causal cell types of 3 GWAS candidate SNPs, and assigns the first candidate SNP to cell types from the biological group A, the second candidate SNP to a cell type of the biological group C, and the third candidate SNP to cell types of the biological groups A and C. The dashed horizontal line represents the S-LDSC significance threshold. LD: linkage disequilibrium; PIP: posterior inclusion probabilities.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

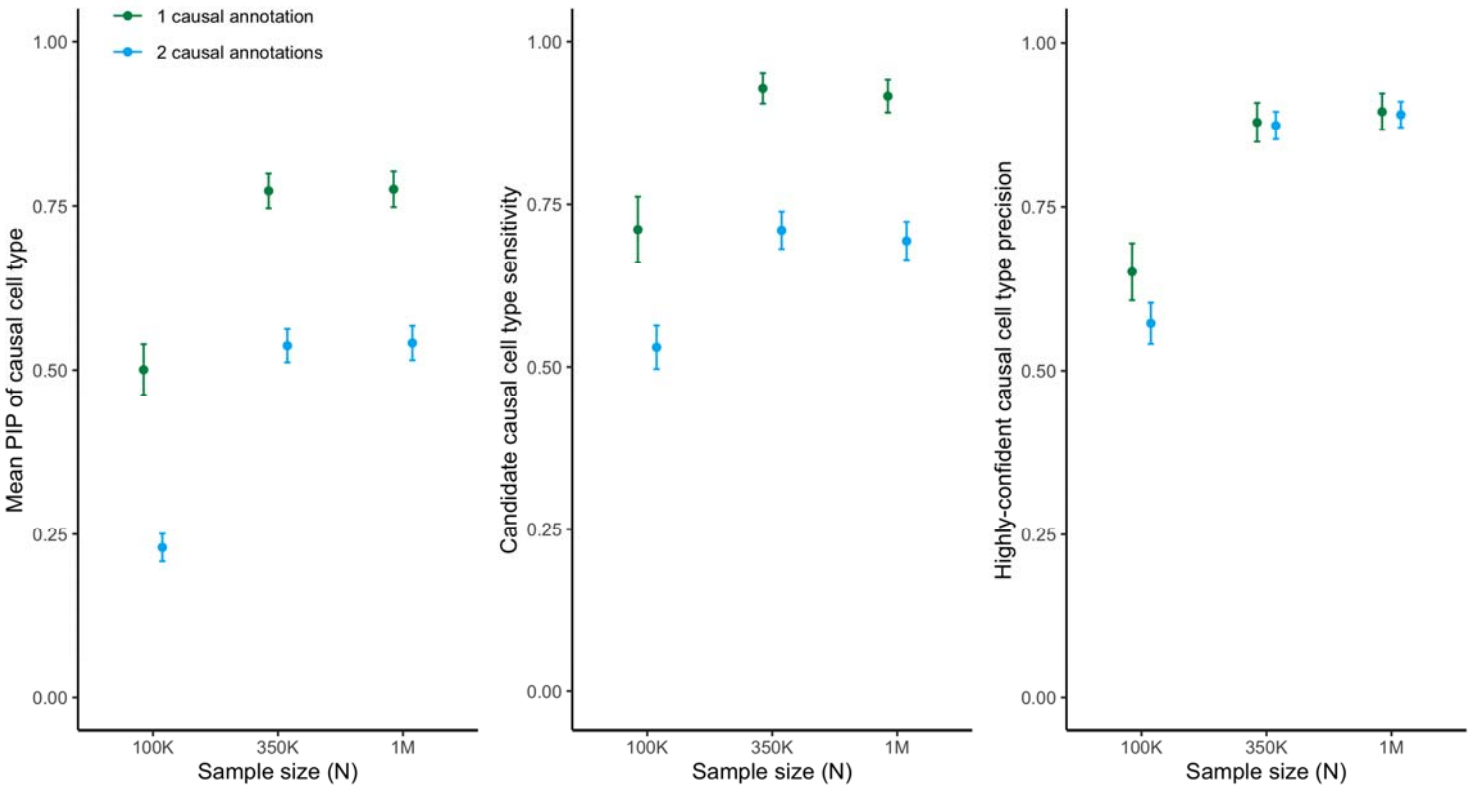


Figure 2. Simulations to assess CT-FM power and accuracy. We report the mean PIP of the causal SNP-annotation(s), the proportion of causal SNP-annotations identified as a candidate causal cell type (candidate causal cell type sensitivity), and the proportion of SNP-annotations identified as a highly-confident causal cell type that are truly causal (highly-confident causal cell type precision). Simulations with one causal cell type (green) and two causal cell types (blue), and with different GWAS sample size *N*. For each scenario, we performed 500 simulations, and applied CT-FM in simulations where S-LDSC maximum Z-score was > 4. Error bars represent 95% confidence intervals. Numerical results are reported in **Supplementary Table 3**.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

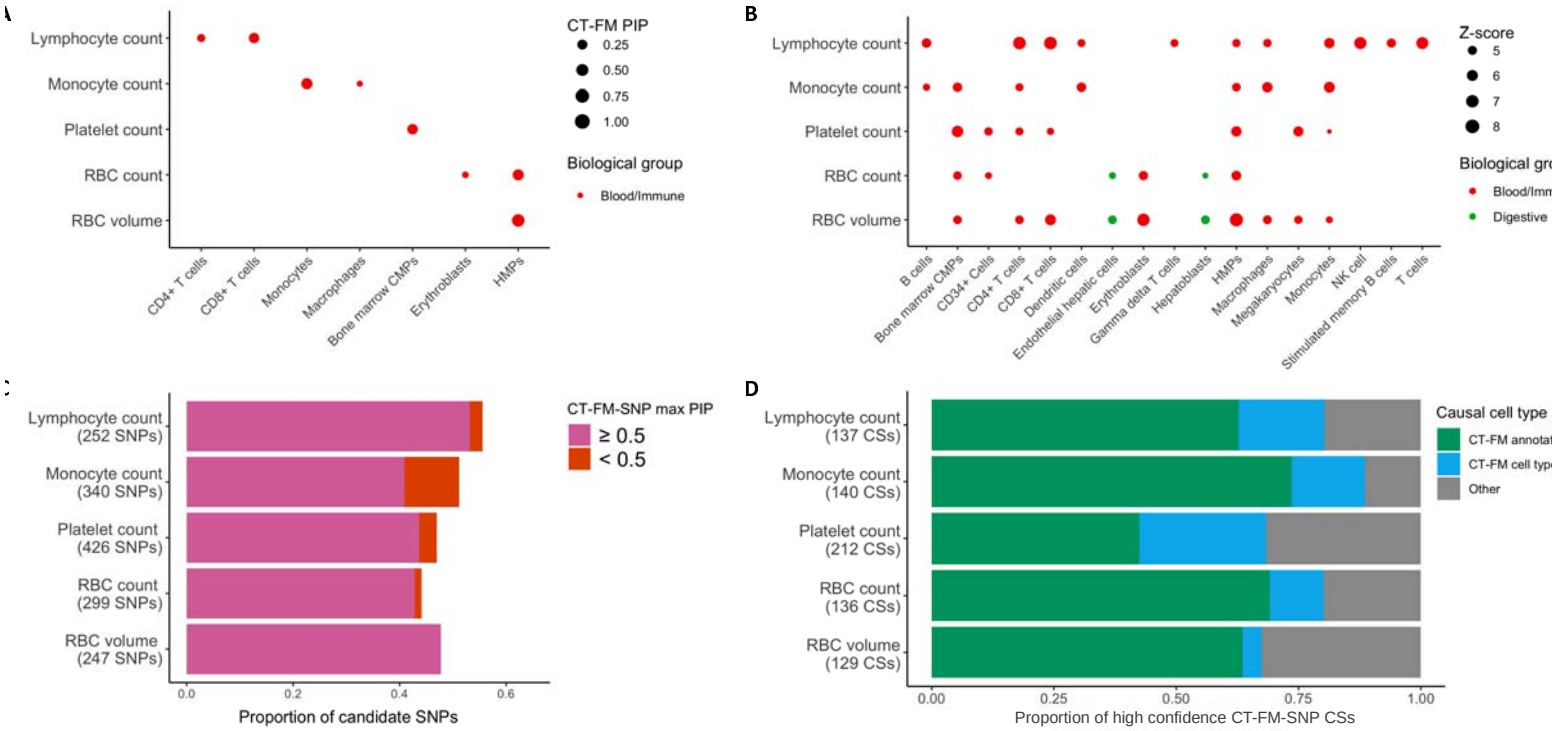


Figure 3. Benchmarking CT-FM and CT-FM-SNP on five blood cell traits. (a) We report cell types identified by CT-FM as causal for each trait. CT-FM PIP value for each cell type (dot size), the corresponding biological group (dot color) and CT-FM credible set (shape) are shown for each trait (rows) and cell type (columns). Numerical results are reported in **Supplementary Table 4**. **(b)** We report cell types identified by CT-FM as significantly associated with each trait. S-LDSC Z-scores (dot size) and the corresponding biological group (dot color) of CTS SNP-annotations are shown for each trait (rows) and cell type (columns). Only CTS SNP-annotations presenting S-LDSC Z-score > 4 are shown. Numerical results are reported in **Supplementary Table 5**. **(c)** We report the proportion of candidate causal SNPs that were linked to at least one causal cell type by CT-FM-SNP. CT-FM-SNP results for all candidate variants are reported in **Supplementary Table 7**. **(d)** We report the proportion of high confidence {non-coding SNP, cell type, trait} triplets inferred by CT-FM-SNP where the cell type is consistent with CT-FM results. We highlight triplets where the causal CTS SNP-annotation was also found in CT-FM CSs (green), triplets where the causal CTS SNP-annotation was not found in CT-FM CSs, but corresponds to the same cell type (blue), and triplets where the causal CTS SNP-annotation was not found in CT-FM CSs (grey). Numerical results are reported in **Supplementary Table 8**. RBC: red blood cell; CMPs: common myeloid progenitor cells; HMPs: hematopoietic multipotent progenitors; NK: natural killer.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

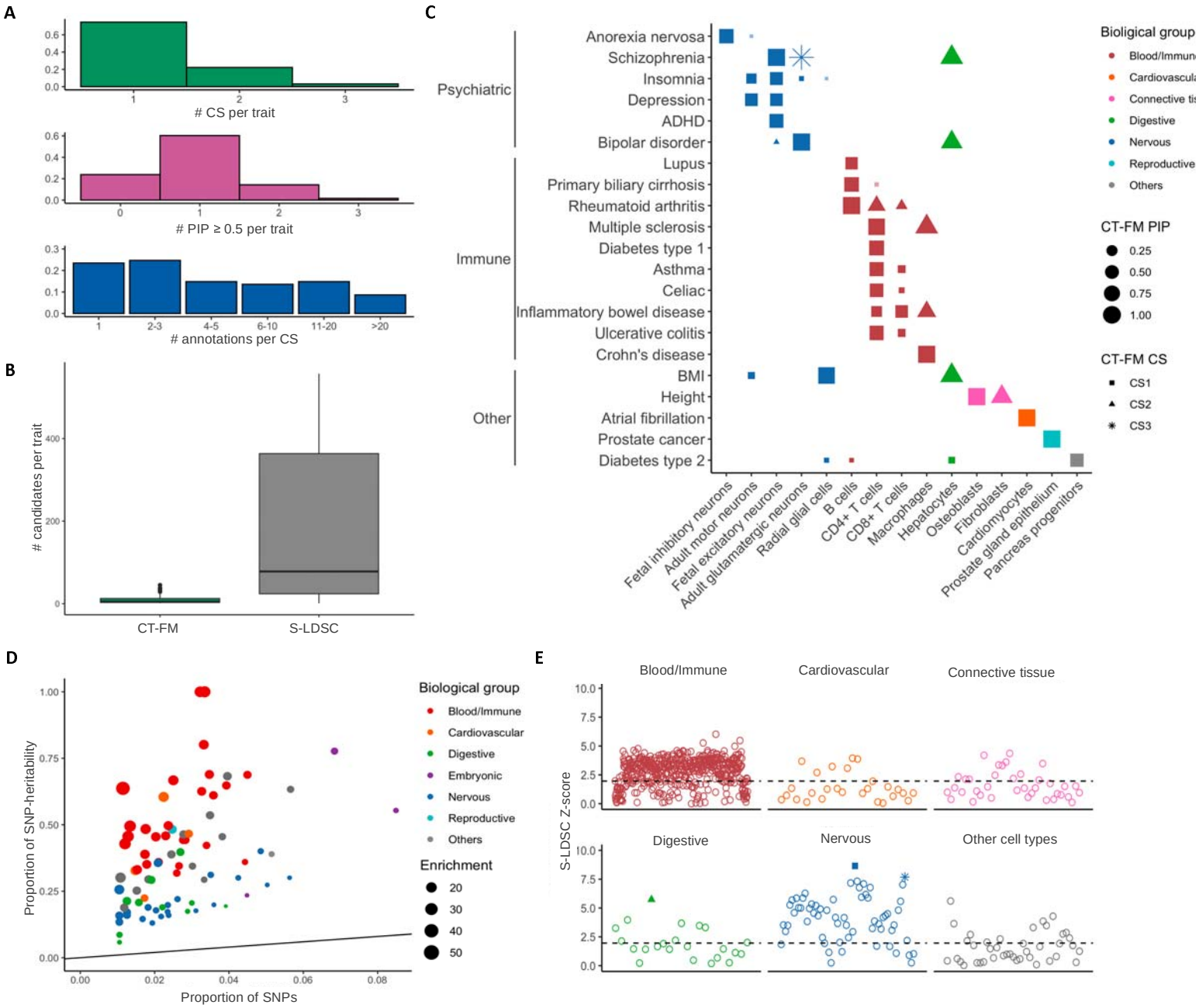


Figure 4. Application of CT-FM to 63 GWAS summary statistics. (a) We report the number of credible sets (CSs) per trait, the number of highly-confident causal cell types (PIP > 0.5) per trait, and the number of CTS SNP-annotations per CS. Causal cell types with PIP < 0.5 but cPIP > 0.5 were not reported. Numerical results are reported in **Supplementary Table 10**. **(b)** We report the distribution of candidate causal CTS SNP-annotations inferred by CT-FM and S-LDSC. The median value of confident scores is displayed as a band inside each box; boxes denote values in the second and third quartiles; the length of each whisker is 1.5 times the interquartile range, defined as the width of each box. Numerical results are reported in **Supplementary Table 12**. **(c)** We report notable candidate causal cell types (columns) inferred by CT-FM for different complex traits (rows). For each {cell type, trait} pair, CT-FM PIP values

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

(dot size), different CSs (dot shape) and the corresponding biological group (dot color) are indicated. CT-FM results for all GWASs are reported in **Supplementary Tables 10-11**. **(d)** We report the proportion of SNP-heritability explained by CT-FM CSs. Proportion for lupus (1.03 ± 0.23) and celiac (1.18 ± 0.31) diseases were rounded to 1 for visualization purposes; we note that values greater than 1 are outside the biologically plausible 0–1 range, but allowing point estimates outside the biologically plausible 0–1 range is necessary to ensure unbiasedness. Numerical results are reported in **Supplementary Table 13**. **(e)** We report S-LDSC and CT-FM results for Schizophrenia [54]. For each CTS SNP-annotation, we report S-LDSC Z-score on the *y* axis (only SNP-annotations with Z-score > 0 were presented), and CT-FM CS through different shapes (square for fetal excitatory neurons, asterisk for adult glutamatergic neurons, and triangle for hepatocytes; SNP-annotations not in CS are represented with open circle). The dashed horizontal line represents the S-LDSC FDR significance threshold. S-LDSC results for the 927 CTS SNP-annotations are reported in **Supplementary Table 18**. ADHD: attention-deficit / hyperactivity disorder; BMI: body mass index.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

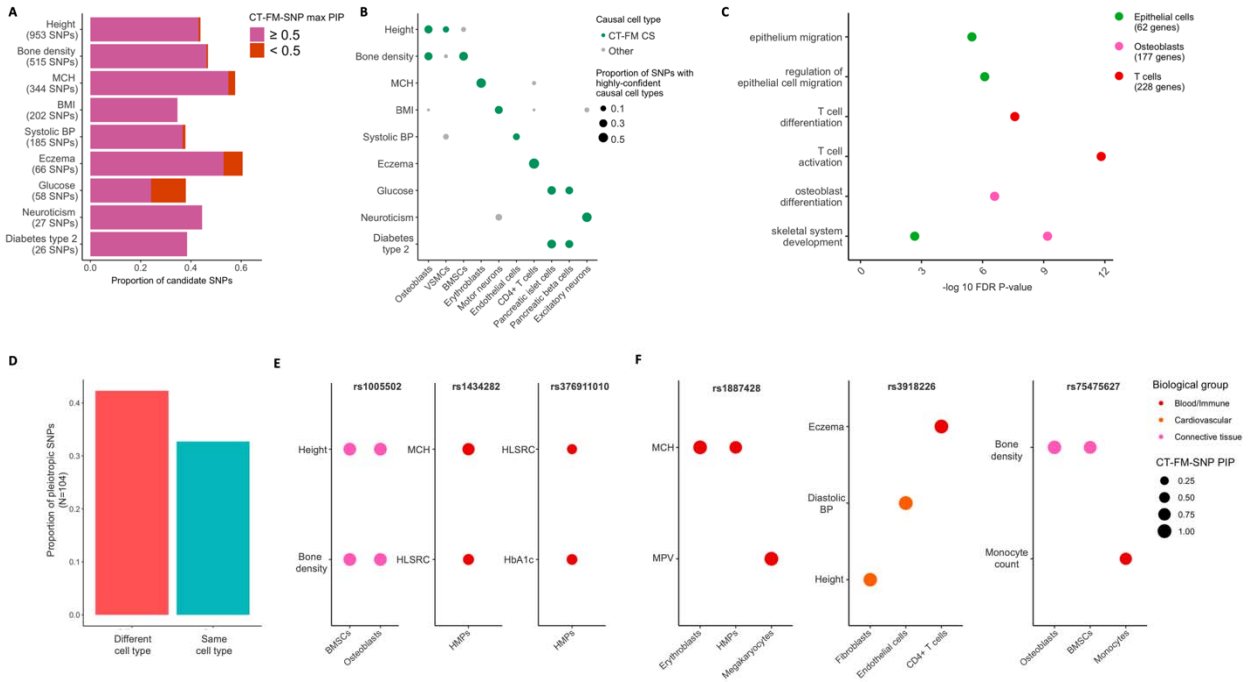


Figure 5. Application of CT-FM-SNP to fine-mapped SNPs of 39 UK Biobank traits. (a) We report the proportion of candidate causal SNPs that were linked to at least one causal cell type by CT-FM-SNP for 9 representative traits. CT-FM-SNP results for all candidate variants of the 39 traits are reported in **Supplementary Table 22**. **(b)** We report the proportion of SNPs with CT-FM-SNP high-confidence causal cell type in different cell types for 9 representative traits. Cell types identified within CT-FM CSs were represented by green dots, cell types not identified by CT-FM were represented by grey dots. **(c)** We report the enrichment of biologically relevant processes (x axis) for genes linked to SNPs assigned to osteoblasts, epithelial and CD4+ T cells by CT-FM-SNP. For each process and cell type, we report the FDR corrected enrichment p-value (y axis). Only biologically relevant gene ontology processes with FDR P < 0.01 are shown. Full gene ontology enrichment results are available in **Supplementary Table 24**. **(d)** We report CT-FM-SNP results for 107 pleiotropic SNPs identified across 18 genetically uncorrelated UK Biobank traits. The proportion of pleiotropic SNPs assigned to different candidate cell types is represented with a red bar, while the proportion of SNPs sharing at least 1 candidate causal cell type across traits is represented with a blue bar. CT-FM-SNP results for the 107 pleiotropic SNPs are reported in **Supplementary Table 27**. **(e,f)** We report three examples where pleiotropic SNPs were assigned to the same cell type **(e)**, and three examples where pleiotropic SNPs were assigned to different cell types **(f)**. MCH: mean corpuscular hemoglobin; BMI: body mass index; BP: blood pressure; VSMCs: vascular smooth muscle cells; BMSC: bone marrow stromal cells; HLSRC: high light scatter reticulocyte count; HbA1c: hemoglobin A1c; MPV: mean platelet volume.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

Trait	SNP	CT-FM-SNP PIP	Cell type	Gene (cS2G)
Asthma	rs479844	0.99	CD4+ T cells	*OVOL1*
Bone density	rs10130587	0.88	Osteoblasts	*BMP4*
Chronotype	rs13081924	0.92	Fetal Excitatory Neurons	*WNT7A*
Lymphocyte count	rs35592432	0.57	CD8+ T cells	*FOXP1*
Neuroticism	rs191480627	0.99	Hepatocytes	*AGPAT3*
	rs34272688	0.61	Fetal Excitatory Neurons	*ATAD2B*
Platelet count	rs117672662	0.98	Megakaryocytes	*ACTN1*
Platelet volume		1	Megakaryocytes	
Platelet volume	rs998908	1	Megakaryocytes	*CD9*
Type 2 diabetes	rs571342427	0.68	Pancreatic beta cells	*INS*
		0.94	Fetal pancreatic islet cells	

Table 1. Notable {non-coding SNP, cell type, gene, trait} quadruplets. Results for rs571342427, rs10130587 and rs191480627 are discussed in the main text. Results for other candidate SNPs are discussed in the **Supplementary Note**.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
2. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
3. Hekselman, I. & Yeger-Lotem, E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat. Rev. Genet.* **21**, 137–150 (2020).
4. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
5. Jagadeesh, K. A. *et al.* Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nat. Genet.* **54**, 1479–1492 (2022).
6. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
7. Aygün, N. *et al.* Inferring cell-type-specific causal gene regulatory networks during human neurogenesis. *Genome Biol.* **24**, 130 (2023).
8. Ma, Y. *et al.* Polygenic regression uncovers trait-relevant cellular contexts through pathway activation transformation of single-cell RNA sequencing data. *Cell Genomics* **0**, (2023).
9. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676–1683 (2017).
10. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, (2013).
11. Varshney, A. *et al.* Cell Specificity of Human Regulatory Annotations and Their Genetic Effects on Gene Expression. *Genetics* **211**, 549–562 (2019).
12. Trynka, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).

13. Hao, X., Zeng, P., Zhang, S. & Zhou, X. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet.* **14**, e1007186 (2018).
14. Amariuta, T., Siewert-Rocks, K. & Price, A. L. Modeling tissue co-regulation estimates tissue-specific contributions to disease. *Nat. Genet.* **55**, 1503–1511 (2023).
15. Zhang, Z. *et al.* A scalable approach to characterize pleiotropy across thousands of human diseases and complex traits using GWAS summary statistics. *Am. J. Hum. Genet.* **110**, 1863–1874 (2023).
16. Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, (2019).
17. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. *Nat. Genet.* **49**, (2017).
18. Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. *Nat. Rev. Genet.* **23**, 120–133 (2021).
19. Sobreira, D. R. *et al.* Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5. *Science* **372**, (2021).
20. Strober, B. J., Zhang, M. J., Amariuta, T., Rossen, J. & Price, A. L. Fine-mapping causal tissues and genes at disease-associated loci. *medRxiv* (2023) doi:10.1101/2023.11.01.23297909.
21. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).
22. Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
23. Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

mediated by assayed gene expression levels. *Nat. Genet.* **52**, 626–633 (2020).

24. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
25. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
26. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300–307 (2021).
27. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).
28. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
29. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001.e19 (2021).
30. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
31. Gazal, S. *et al.* Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat. Genet.* **50**, 1600–1607 (2018).
32. SCREEN: Search Candidate cis-Regulatory Elements by ENCODE. http://screen-beta.wenglab.org/.
33. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
34. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
35. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

36. Gazal, S., Marquez-Luna, C., Finucane, H. K. & Price, A. L. Reconciling S-LDSC and LDAK functional enrichment estimates. *Nat. Genet.* **51**, 1202–1204 (2019).

37. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

38. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the 'Sum of Single Effects' model. *PLoS Genet.* **18**, e1010299 (2022).

39. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

40. Chen, M.-H. *et al.* Trans-ethnic and Ancestry-Specific Blood-Cell Genetics in 746,667 Individuals from 5 Global Populations. *Cell* **182**, 1198–1213.e14 (2020).

41. Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat. Genet.* **51**, 683–693 (2019).

42. Cui, R. *et al.* Improving fine-mapping by modeling infinitesimal effects. *Nat. Genet.* (2023) doi:10.1038/s41588-023-01597-3.

43. Kruse, A. O. & Bustillo, J. R. Glutamatergic dysfunction in Schizophrenia. *Transl. Psychiatry* **12**, 500 (2022).

44. Marsman, A. *et al.* Glutamate in schizophrenia: a focused review and meta-analysis of ^{1}H-MRS studies. *Schizophr. Bull.* **39**, 120–129 (2013).

45. Brandenburg, S., Arakel, E. C., Schwappach, B. & Lehnart, S. E. The molecular and functional identities of atrial cardiomyocytes in health and disease. *Biochim. Biophys. Acta* **1863**, 1882–1893 (2016).

46. Marchetti, P., Bugliani, M., Boggi, U., Masini, M. & Marselli, L. The pancreatic beta cells in human type 2 diabetes. *Adv. Exp. Med. Biol.* **771**, (2012).

47. Park, J. W. *et al.* Prostate epithelial cell of origin determines cancer differentiation state in an organoid transformation assay. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 4482–4487 (2016).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

48. Kanai, M. *et al.* Insights from complex trait fine-mapping across diverse populations. *medRxiv* 2021.09.03.21262975 (2021) doi:10.1101/2021.09.03.21262975.

49. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* (2020) doi:10.1126/science.aaz1776.

50. Young, A. I. *et al.* Mendelian imputation of parental genotypes improves estimates of direct genetic effects. *Nat. Genet.* **54**, 897–905 (2022).

51. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).

52. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. *Hum. Mol. Genet.* **27**, 3641–3649 (2018).

53. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat. Genet.* **53**, 817–829 (2021).

54. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).

55. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, (2019).

56. de Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).

57. Saevarsdottir, S. *et al.* Multiomics analysis of rheumatoid arthritis yields sequence variants that have large effects on risk of the seropositive subset. *Ann. Rheum. Dis.* **81**, 1085–1095 (2022).

58. Schipper, M. *et al.* Gene prioritization in GWAS loci using multimodal evidence. *medRxiv* 2023.12.23.23300360 (2024) doi:10.1101/2023.12.23.23300360.

59. Liharska, L. E. *et al.* A study of gene expression in the living human brain. *medRxiv* 2023.04.21.23288916 (2023) doi:10.1101/2023.04.21.23288916.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

60. Fuller, B. E. *et al.* Prevalence of liver disease in veterans with bipolar disorder or schizophrenia. *Gen. Hosp. Psychiatry* **33**, (2011).

61. Hsu, J. H., Chien, I. C., Lin, C. H., Chou, Y. J. & Chou, P. Increased risk of chronic liver disease in patients with schizophrenia: a population-based cohort study. *Psychosomatics* **55**, (2014).

62. Miller, B. J., Culpepper, N. & Rapaport, M. H. C-reactive protein levels in schizophrenia: a review and meta-analysis. *Clin. Schizophr. Relat. Psychoses* **7**, (2014).

63. Said, S. *et al.* Genetic analysis of over half a million people characterises C-reactive protein loci. *Nat. Commun.* **13**, 2198 (2022).

64. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).

65. Li, M. *et al.* Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science* **362**, (2018).

66. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162–183 (2019).

67. Watanabe, K. *et al.* Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. *Nat. Genet.* **54**, 1125–1132 (2022).

68. Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat. Neurosci.* **22**, 343–352 (2019).

69. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat. Genet.* **55**, 198–208 (2023).

70. Watson, H. J. *et al.* Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nat. Genet.* **51**, 1207–1214 (2019).

71. Surgenor, L. J. & Maguire, S. Assessment of anorexia nervosa: an overview of universal issues and contextual challenges. *Journal of Eating Disorders* **1**, 29 (2013).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

72. Pappaianni, E. *et al.* Initial evidence of abnormal brain plasticity in anorexia nervosa: an ultra-high field study. *Sci. Rep.* **12**, (2022).

73. Tsuo, K. *et al.* Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. *Cell Genom* **2**, 100212 (2022).

74. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).

75. Chiou, J. *et al.* Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. *Nature* **594**, 398–402 (2021).

76. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47**, 1457–1464 (2015).

77. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6**, 8019 (2015).

78. Amariuta, T. *et al.* IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *Am. J. Hum. Genet.* **104**, (2019).

79. Ishigaki, K. *et al.* Multi-ancestry genome-wide association analyses identify novel genetic mechanisms in rheumatoid arthritis. *Nat. Genet.* **54**, 1640–1651 (2022).

80. Gupta, A. *et al.* Dynamic regulatory elements in single-cell multimodal data implicate key immune cell states enriched for autoimmune disease heritability. *Nat. Genet.* **55**, 2200–2210 (2023).

81. Yap, H.-Y. *et al.* Pathogenic Role of Immune Cells in Rheumatoid Arthritis: Implications in Clinical Treatment and Biomarker Development. *Cells* **7**, (2018).

82. Dharmasiri, S. *et al.* Human Intestinal Macrophages Are Involved in the Pathology of Both Ulcerative Colitis and Crohn Disease. *Inflamm. Bowel Dis.* **27**, 1641–1652 (2021).

83. Maul, J. *et al.* Peripheral and intestinal regulatory CD4+ CD25(high) T cells in inflammatory

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

bowel disease. *Gastroenterology* **128**, 1868–1878 (2005).

84. Koch, S., Kucharzik, T., Heidemann, J., Nusrat, A. & Luegering, A. Investigating the role of proinflammatory CD16+ monocytes in the pathogenesis of inflammatory bowel disease. *Clin. Exp. Immunol.* **161**, 332–341 (2010).
85. Hanai, H. *et al.* Adsorptive depletion of elevated proinflammatory CD14+CD16+DR++ monocytes in patients with inflammatory bowel disease. *Am. J. Gastroenterol.* **103**, 1210–1216 (2008).
86. Smids, C. *et al.* Intestinal T Cell Profiling in Inflammatory Bowel Disease: Linking T Cell Subsets to Disease Activity and Disease Course. *J. Crohns. Colitis* **12**, 465–475 (2018).
87. Heller, F. *et al.* Interleukin-13 is the key effector Th2 cytokine in ulcerative colitis that affects epithelial tight junctions, apoptosis, and cell restitution. *Gastroenterology* **129**, 550–564 (2005).
88. Zivot, A., Lipton, J. M., Narla, A. & Blanc, L. Erythropoiesis: insights into pathophysiology and treatments in 2017. *Mol. Med.* **24**, 11 (2018).
89. Lee, W. Role of Osteoblasts in the Formation of New Bone Tissue and Maintenance of Healthy Bones. *Journal of Osteoporosis and Physical Activity* **11**, 1–1 (2023).
90. Ottewell, P. D. The role of osteoblasts in bone metastasis. *J Bone Oncol* **5**, 124–127 (2016).
91. Gazal, S. *et al.* Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat. Genet.* **54**, 827–836 (2022).
92. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).
93. Campbell, J. E. & Newgard, C. B. Mechanisms controlling pancreatic islet cell function in insulin secretion. *Nat. Rev. Mol. Cell Biol.* **22**, 142–158 (2021).
94. Whelan, J., Cordle, S. R., Henderson, E., Weil, P. A. & Stein, R. Identification of a pancreatic beta-cell insulin gene transcription factor that binds to and appears to activate

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

cell-type-specific expression: its possible relationship to other cellular factors that bind to a common insulin gene sequence. *Mol. Cell. Biol.* **10**, 1564–1572 (1990).

95. Reis, L. M. *et al.* BMP4 loss-of-function mutations in developmental eye disorders including SHORT syndrome. *Hum. Genet.* **130**, 495–504 (2011).
96. Kim, H.-N. *et al.* Genetic variants that affect length/height in infancy/early childhood in Vietnamese-Korean families. *J. Hum. Genet.* **55**, 681–690 (2010).
97. Yamaguchi, A. *et al.* Effects of BMP-2, BMP-4, and BMP-6 on osteoblastic differentiation of bone marrow-derived stromal cell lines, ST2 and MC3T3-G2/PA6. *Biochem. Biophys. Res. Commun.* **220**, 366–371 (1996).
98. Chang, S.-F. *et al.* BMP-4 induction of arrest and differentiation of osteoblast-like cells via p21 CIP1 and p27 KIP1 regulation. *Mol. Endocrinol.* **23**, 1827–1838 (2009).
99. Martinovic, S. *et al.* Requirement of a bone morphogenetic protein for the maintenance and stimulation of osteoblast differentiation. *Arch. Histol. Cytol.* **69**, 23–36 (2006).
100. Hishikawa, D. *et al.* Hepatic Levels of DHA-Containing Phospholipids Instruct SREBP1-Mediated Synthesis and Systemic Delivery of Polyunsaturated Fatty Acids. *iScience* **23**, 101495 (2020).
101. Thesing, C. S. *et al.* Fatty acids and recurrence of major depressive disorder: combined analysis of two Dutch clinical cohorts. *Acta Psychiatr. Scand.* **141**, 362–373 (2020).
102. Thesing, C. S., Bot, M., Milaneschi, Y., Giltay, E. J. & Penninx, B. W. J. H. The association of omega-3 fatty acid levels with personality and cognitive reactivity. *J. Psychosom. Res.* **108**, 93–101 (2018).
103. Reinders, I., Virtanen, J. K., Brouwer, I. A. & Tuomainen, T.-P. Association of serum n-3 polyunsaturated fatty acids with C-reactive protein in men. *Eur. J. Clin. Nutr.* **66**, 736–741 (2012).
104. Muka, T. *et al.* Polyunsaturated fatty acids and serum C-reactive protein: the Rotterdam study. *Am. J. Epidemiol.* **181**, 846–856 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

105. Mazidi, M., Gao, H.-K., Vatanparast, H. & Kengne, A. P. Impact of the dietary fatty acid intake on C-reactive protein levels in US adults. *Medicine* **96**, e5736 (2017).

106. Shi, J., Yuan, B., Hu, W. & Lodish, H. JAK2 V617F stimulates proliferation of erythropoietin-dependent erythroid progenitors and delays their differentiation by activating Stat1 and other nonerythroid signaling pathways. *Exp. Hematol.* **44**, 1044–1058.e5 (2016).

107. Ghai, S. & Rai, S. Megakaryocytic morphology in Janus kinase 2 V617F positive myeloproliferative neoplasm. *South Asian J Cancer* **6**, 75–78 (2017).

108. Oliveira-Paula, G. H., Lacchini, R. & Tanus-Santos, J. E. Endothelial nitric oxide synthase: From biochemistry and gene structure to clinical implications of NOS3 polymorphisms. *Gene* **575**, 584–599 (2016).

109. Cortese-Krott, M. M. & Kelm, M. Endothelial nitric oxide synthase in red blood cells: key to a new erythrocrine function? *Redox Biol* **2**, 251–258 (2014).

110. Gilkeson, G. S. *et al.* Endothelial nitric oxide synthase reduces crescentic and necrotic glomerular lesions, reactive oxygen production, and MCP1 production in murine lupus nephritis. *PLoS One* **8**, e64650 (2013).

111. Nagy, G. & Perl, A. The role of nitric oxide in abnormal T cell signal transduction in systemic lupus erythematosus. *Clin. Immunol.* **118**, 145–151 (2006).

112. Witte, M. B., Thornton, F. J., Efron, D. T. & Barbul, A. Enhancement of fibroblast collagen synthesis by nitric oxide. *Nitric Oxide* **4**, 572–582 (2000).

113. Xu, X. *et al.* SPTBN1 Prevents Primary Osteoporosis by Modulating Osteoblasts Proliferation and Differentiation and Blood Vessels Formation in Bone. *Front Cell Dev Biol* **9**, 653724 (2021).

114. Calabrese, G. M. *et al.* Integrating GWAS and Co-expression Network Data Identifies Bone Mineral Density Genes SPTBN1 and MARK3 and an Osteoblast Functional Module. *Cell Syst* **4**, 46–59.e4 (2017).

115. Yang, P. *et al.* βII spectrin (SPTBN1): biological function and clinical potential in cancer and

medRxiv preprint doi: https://doi.org/10.1101/2024.05.17.24307556; this version posted May 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

other diseases. *Int. J. Biol. Sci.* **17**, 32–49 (2021).

116. Suzuki, K. *et al.* Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature* **627**, 347–357 (2024).

117. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, (2019).

118. Shi, H. *et al.* Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* **12**, 1098 (2021).

119. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

120. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–8 (2006).

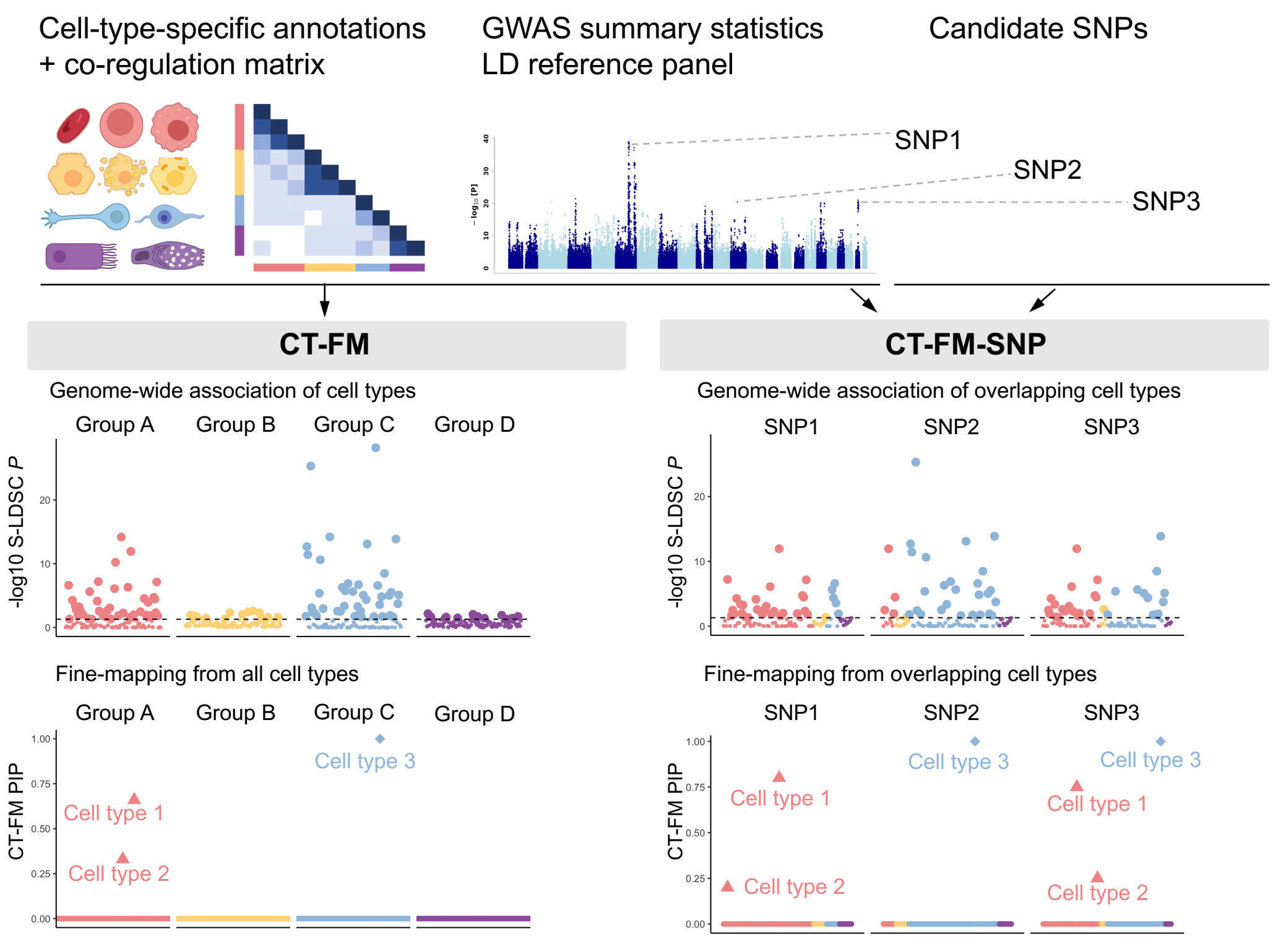
121. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

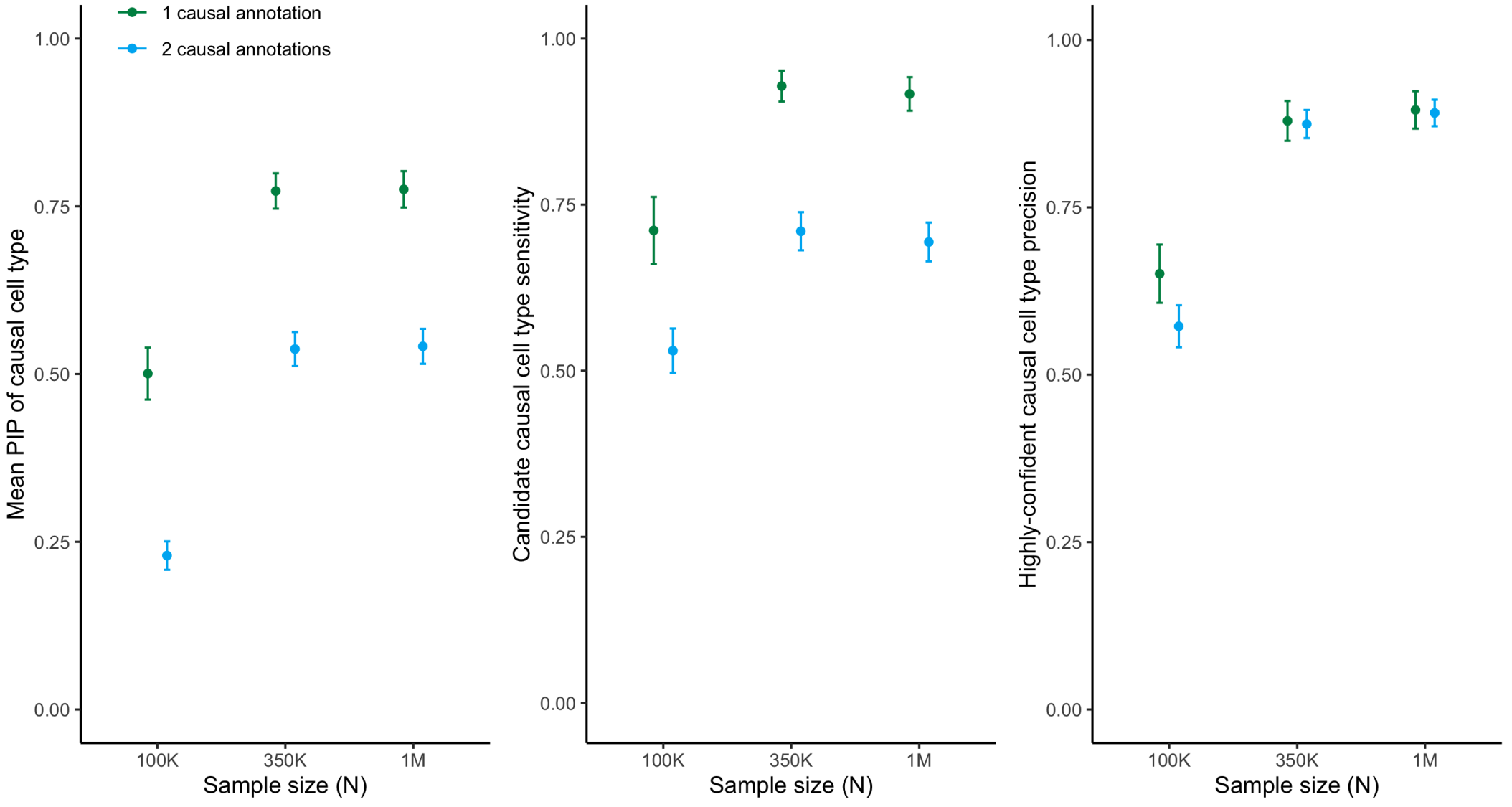
122. Neph, S. *et al.* BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).

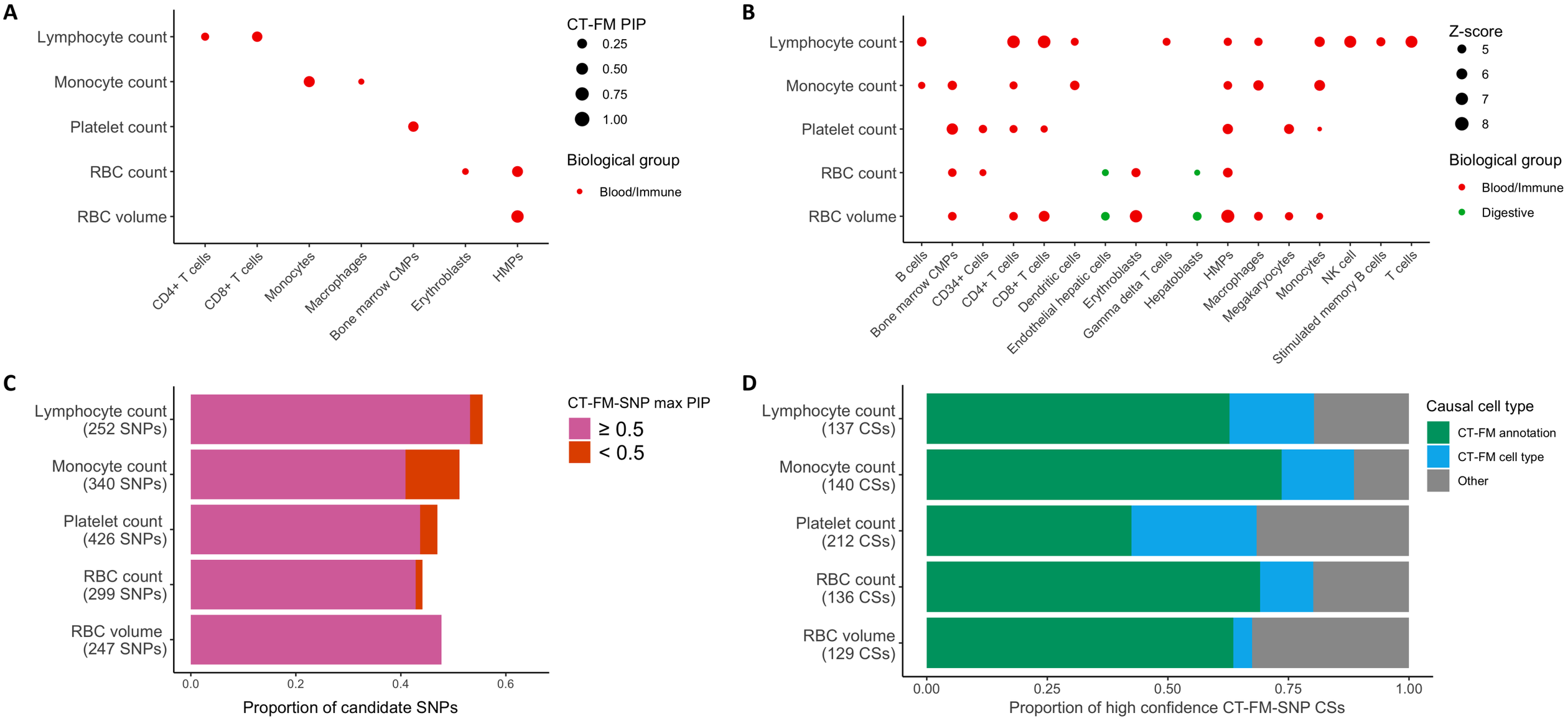
123. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, (2015).

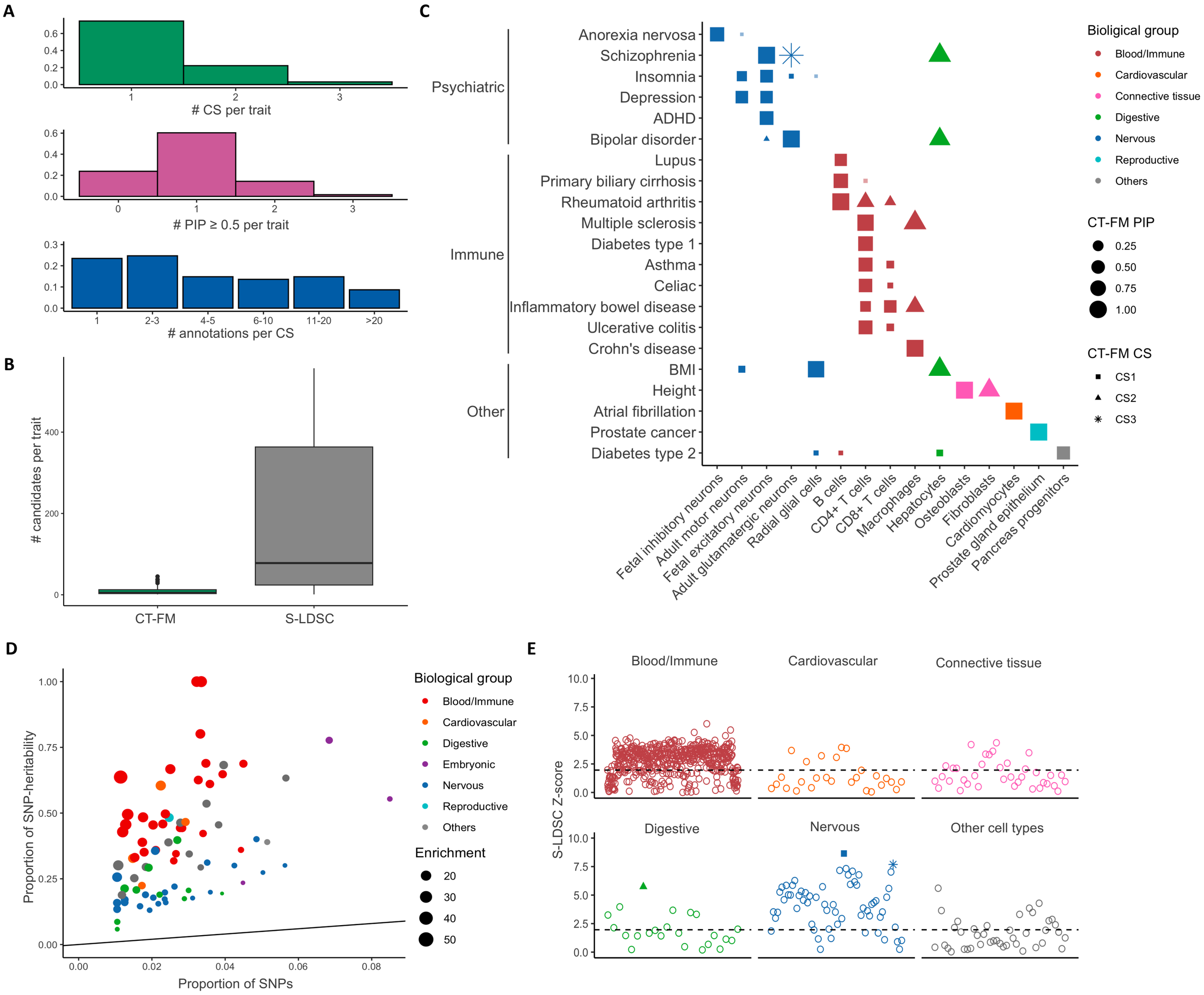
124. Zou, Y. Diagnostic for fine-mapping with summary statistics. https://stephenslab.github.io/susieR/articles/susierss_diagnostic.html.

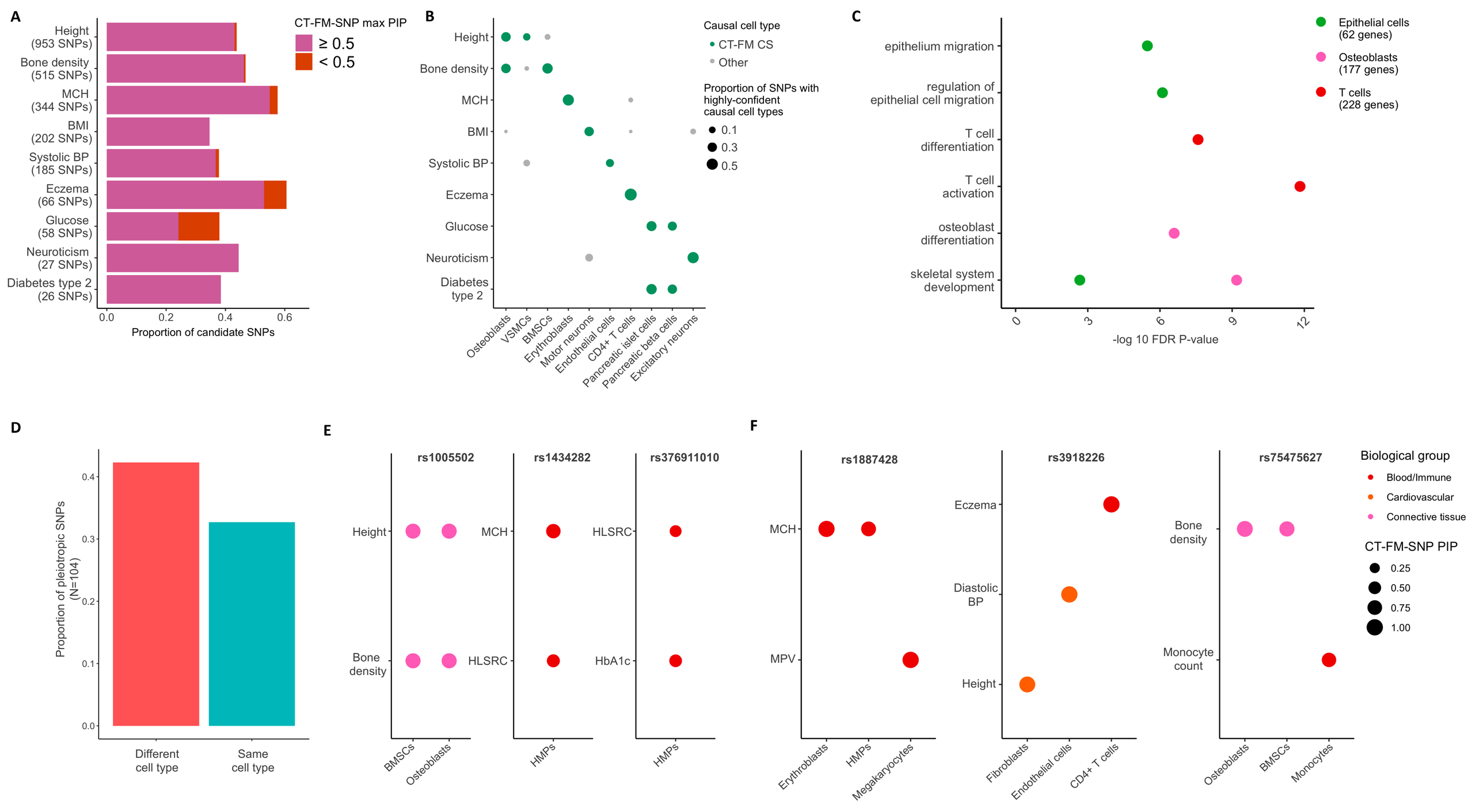
125. van de Geijn, B. *et al.* Annotations capturing cell type-specific TF binding explain a large fraction of disease heritability. *Hum. Mol. Genet.* **29**, 1057–1067 (2019).


Cell-type-specific annotations
+ co-regulation matrix
GWAS summary statistics
LD reference panel
Candidate SNPs
SNP1
SNP2
SNP3
CT-FM
CT-FM-SNP
Genome-wide association of cell types
Group A
Group B
Group C
Group D
-log10 S-LDSC P
Genome-wide association of overlapping cell types
SNP1
SNP2
SNP3
Fine-mapping from all cell types
CT-FM PIP
Cell type 1
Cell type 2
Cell type 3
Fine-mapping from overlapping cell types


1 causal annotation
2 causal annotations
Mean PIP of causal cell type
Candidate causal cell type sensitivity
Highly-confident causal cell type precision
1.00
0.75
0.50
0.25
0.00
100K
350K
1M
Sample size (N)


A
Lymphocyte count
Monocyte count
Platelet count
RBC count
RBC volume
CD4+ T cells
CD8+ T cells
Monocytes
Macrophages
Bone marrow CMPs
Erythroblasts
HMPs
CT-FM PIP
0.25
0.50
0.75
1.00
Biological group
Blood/Immune
B
Lymphocyte count
Monocyte count
Platelet count
RBC count
RBC volume
B cells
Bone marrow CMPs
CD34+ Cells
CD4+ T cells
CD8+ T cells
Dendritic cells
Endothelial hepatic cells
Erythroblasts
Gamma delta T cells
Hepatoblasts
HMPs
Macrophages
Megakaryocytes
Monocytes
NK cell
Stimulated memory B cells
T cells
Z-score
5
6
7
8
Biological group
Blood/Immune
Digestive
C
Lymphocyte count (252 SNPs)
Monocyte count (340 SNPs)
Platelet count (426 SNPs)
RBC count (299 SNPs)
RBC volume (247 SNPs)
0.0
0.2
0.4
0.6
Proportion of candidate SNPs
CT-FM-SNP max PIP
≥ 0.5
< 0.5
D
Lymphocyte count (137 CSs)
Monocyte count (140 CSs)
Platelet count (212 CSs)
RBC count (136 CSs)
RBC volume (129 CSs)
0.00
0.25
0.50
0.75
1.00
Proportion of high confidence CT-FM-SNP CSs
Causal cell type
CT-FM annotation
CT-FM cell type
Other


A
CS per trait
PIP ≥ 0.5 per trait
annotations per CS
1
2-3
4-5
6-10
11-20
>20
B
candidates per trait
CT-FM
S-LDSC
C
Psychiatric
Immune
Other
Anorexia nervosa
Schizophrenia
Insomnia
Depression
ADHD
Bipolar disorder
Lupus
Primary biliary cirrhosis
Rheumatoid arthritis
Multiple sclerosis
Diabetes type 1
Asthma
Celiac
Inflammatory bowel disease
Ulcerative colitis
Crohn's disease
BMI
Height
Atrial fibrillation
Prostate cancer
Diabetes type 2
Fetal inhibitory neurons
Adult motor neurons
Fetal excitatory neurons
Adult glutamatergic neurons
Radial glial cells
B cells
CD4+ T cells
CD8+ T cells
Macrophages
Hepatocytes
Osteoblasts
Fibroblasts
Cardiomyocytes
Prostate gland epithelium
Pancreas progenitors
Bioligical group
Blood/Immune
Cardiovascular
Connective tissue
Digestive
Nervous
Reproductive
Others
CT-FM PIP
0.25
0.50
0.75
1.00
CT-FM CS
CS1
CS2
CS3
D
Proportion of SNP-heritability
Proportion of SNPs
Biological group
Blood/Immune
Cardiovascular
Digestive
Embryonic
Nervous
Reproductive
Others
Enrichment
20
30
40
50
E
Blood/Immune
Cardiovascular
Connective tissue
Digestive
Nervous
Other cell types
S-LDSC Z-score


A
CT-FM-SNP max PIP
≥ 0.5
< 0.5
Height (953 SNPs)
Bone density (515 SNPs)
MCH (344 SNPs)
BMI (202 SNPs)
Systolic BP (185 SNPs)
Eczema (66 SNPs)
Glucose (58 SNPs)
Neuroticism (27 SNPs)
Diabetes type 2 (26 SNPs)
0.0
0.2
0.4
0.6
Proportion of candidate SNPs
B
Causal cell type
CT-FM CS
Other
Proportion of SNPs with highly-confident causal cell types
0.1
0.3
0.5
Height
Bone density
MCH
BMI
Systolic BP
Eczema
Glucose
Neuroticism
Diabetes type 2
Osteoblasts
VSMCs
BMSCs
Erythroblasts
Motor neurons
Endothelial cells
CD4+ T cells
Pancreatic islet cells
Pancreatic beta cells
Excitatory neurons
C
Epithelial cells (62 genes)
Osteoblasts (177 genes)
T cells (228 genes)
epithelium migration
regulation of epithelial cell migration
T cell differentiation
T cell activation
osteoblast differentiation
skeletal system development
0
3
6
9
12
-log 10 FDR P-value
D
Proportion of pleiotropic SNPs (N=104)
0.0
0.1
0.2
0.3
0.4
Different cell type
Same cell type
E
rs1005502
rs1434282
rs376911010
Height
Bone density
MCH
HLSRC
HLSRC
HbA1c
BMSCs
Osteoblasts
HMPs
HMPs
F
rs1887428
rs3918226
rs75475627
MCH
MPV
Eczema
Diastolic BP
Height
Bone density
Monocyte count
Erythroblasts
HMPs
Megakaryocytes
Fibroblasts
Endothelial cells
CD4+ T cells
Osteoblasts
BMSCs
Monocytes
Biological group
Blood/Immune
Cardiovascular
Connective tissue
CT-FM-SNP PIP
0.25
0.50
0.75
1.00