Editor summary:

Here the authors identify genetic effectors of the level of inflammation-related plasma proteins and use Mendelian randomization to identify proteins that contribute to immune-mediated disease risk.

Peer Review Information:

Nature Immunology thanks Tom Richardson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Editor recognition statement:

N. Bernard was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

1. Extended Data

Figure or Table # Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	Figure/Table title One sentence only	Filename Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	Figure/Table Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	**Overview of the pQTL analysis**	EDF 1.eps	Schematic of the analysis pipeline.
Extended Data Fig. 2	**Plasma protein abundance and pQTL detection**	EDF 2.eps	**a)** Proteins with low abundance are more likely to have no detectable pQTL. Y-axis: percentage of samples above lower limit of detection for each protein, calculated using the INTERVAL data (n=4,896) for which we had individual-level protein data available. Blue and red points indicate presence or absence of at least 1 significant pQTL in the GWAS meta-analysis, respectively. **b)** Manhattan plot for genetic associations with plasma IL17C, where the red horizontal line indicates the statistical significance threshold ($5x10^{-10}$). P-values from linear regression.
Extended Data Fig. 3	**pQTL replication in the ARISTOTLE cohort**	EDF 3.eps	Comparison of effect sizes between pQTLs from the discovery pQTL meta-analysis (n=14,824) and the ARISTOTLE cohort (n=1,585). Each point represents a genetic variant that was a significant pQTL in the discovery meta-analysis. Effect size = standard deviation (sd) increase in protein per allele.

			174 of 180 genetic variants were available for testing in the ARISTOTLE data. Red= cis, Blue= trans.
Extended Data Fig. 4	**Genetic architecture of circulating inflammation-related proteins**	EDF 4.eps	**Genetic architecture of circulating inflammation-related proteins. a)** Relationship between minor allele frequency (MAF), pQTL effect size and proportion of variance explained (2MAF(1-MAF)Effect^2), for 227 conditionally independent pQTLs (red=cis, blue=trans). **b)** Proportion of variance explained (PVE) by the conditionally independent variants associated with each protein. Proteins are annotated using the gene symbol of their encoding genes. Protein names are coloured in red if over 80% of samples have levels below the lower limit of detection in the INTERVAL dataset.
Extended Data Fig. 5	**Chemokine *trans*-pQTL hotspot**	EDF 5.eps	Forest plot showing the associations for the pleiotropic *trans*-pQTL at rs12075 (GRCh37, 1:158175353-160525679) with plasma levels of chemokines and blood cell counts. Centre of bar = effect size estimate, whiskers = 95% confidence interval (ci). WBC = white blood cell count. P = p-value, b= beta (effect size). SE = standard error. Blood cell association data from Chen *et al*[31]. P-values from linear regression.
Extended Data Fig. 6	**Colocalisation of pleiotropic chemokine trans-pQTL and blood cell count trait signals**	EDF 6.eps	Regional association plots in the region around rs12075 (GRCh37, 1:158175353-160525679). Left side: association with plasma chemokine levels. Right: associations with basophil, monocyte and white blood cell (WBC) counts using data from Chen *et al*[31]. P-values from linear regression.
Extended Data Fig. 7	**Interactions between the candidate mediators for multi-locus-regulated proteins**	EDF 7.jpg	**a)** TNFSF10 (also known as TRAIL), **b)** KITLG (also known as stem cell factor), and **c)** IL12B. The graphs were generated using the STRINGdb (v11.5) webtool. The colouring of the edges indicates the type of evidence supporting an interaction, as shown in the legend above.
Extended Data Fig. 8	**Protein-disease connections from overlap of pQTLs and disease GWAS**	EDF 8.eps	The protein and the corresponding pQTL sentinel variant are indicated in the format of protein-rsid. The nearest gene to the pQTL sentinel variant is shown in brackets. Red lettering= cis-pQTL, blue lettering= trans-pQTL. Asterix indicates the genetic variant lies

			in the *HLA* region. Red squares: genetic susceptibility to increased plasma levels of the protein is associated with increased disease risk. Blue squares: decreased disease risk.
Extended Data Fig. 9	**Protein and immune-mediated disease (IMD) connections from overlap of pQTLs and disease GWAS**	EDF 9.eps	The protein and the corresponding pQTL sentinel variant are indicated in the format of protein-rsid. The nearest gene to the pQTL sentinel variant is shown in brackets. Red lettering= cis-pQTL, blue lettering= trans-pQTL. Asterix indicates the genetic variant lies in the *HLA* region. Red squares: genetic susceptibility to increased plasma levels of the protein is associated with increased disease risk. Blue squares: decreased disease risk.
Extended Data Fig. 10	**Mendelian randomisation analysis for CXCL5 and ulcerative colitis**	EDF 10.eps	**a)** Scatterplot showing the 13 variants used in the GSMR analysis assessing the effect of CXCL5 on ulcerative colitis (UC) risk from the GWAS by de Lange *et al* (ref 49) Each point represents a genetic variant, and indicates the effect size of the variant on CXCL5 levels versus UC risk (log odds ratio). Vertical and horizontal lines represent 95% confidence intervals. **b)** Directional concordance between CXCL5 pQTL and blood and colon tissue eQTLs. Forest plots showing effect size estimates for rs450373 pQTL in plasma (from our discovery meta-analysis) and eQTLs in whole blood and transverse colon tissue (GTEx v8 data). OR= odds ratio, calculated from beta estimate (representing the change in inverse-rank normalised plasma protein level in standard deviations associated with each copy of the effect allele). CI = confidence interval. P = p-value. Centre of bar = OR estimate, whiskers = 95% CI.

2. Supplementary Information:

A. PDF Files

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*

		extension must be .pdf	
Supplementary Information	Yes	Supplementary-merged.pdf	Supplementary Note Supplementary Figures 1-4
Reporting Summary	Yes	NI-A35592B_Peters_RSf.pdf	
Peer Review Information	Yes	NI-A35592B_Peters_TPR.pdf	

B. Additional Supplementary Files

Type	Number Each type of file (Table, Video, etc.) should be numbered from 1 onwards. Multiple files of the same type should be listed in sequence, i.e.: Supplementary Video 1, Supplementary Video 2, etc.	Filename Whole original file name including extension. i.e.: *Smith_Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Video	Supplementary Item 1	SI.html	**3-dimensional interactive genomic map of pQTLs**. The html file (supplied separately) shows pQTL sentinel variant position in relation to the gene encoding the target protein and the strength of the statistical association (two-sided *P*-values are from meta-analysis of linear regression estimates). Hover over a point to see detailed information. The image can be rotated by holding at left clicking the mouse.
Supplementary Table	Supplementary Table 1	Supplementary-Tables.xlsx	Excel file containing multiple tabs (1 summary tab, then 1 tab per Supplementary Table).

Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets

Jing Hua Zhao[1,2,37], David Stacey[1,2,3,4,37], Niclas Eriksson[5], Erin Macdonald-Dunlop[6], Åsa K Hedman[7,8], Anette Kalnapenkis[9,1], Stefan Enroth[10], Domenico Cozzetto[11], Jonathan Digby-Bell[12], Jonathan Marten[1], Lasse Folkersen[13], Christian Herder[14,15,16], Lina Jonsson[17], Sarah E Bergen[7], Christian Geiger[18,19], Elise J Needham[1,2], Praveen Surendran[1,20,21], Estonian Biobank Research Team[9], Dirk S Paul[1,20,2,22], Ozren Polasek[23], Barbara Thorand[16,18], Harald Grallert[16,18,19], Michael Roden[14,24,16], Urmo Võsa[9], Tonu Esko[9], Caroline Hayward[25], Åsa Johansson[10], Ulf Gyllensten[10], Nicholas Powell[11], Oskar Hansson[26,27], Niklas Mattsson-Carlgren[28,29,30], Peter K Joshi[6], John Danesh[1,2,20,21,31,32], Leonid Padyukov[33,34], Lars Klareskog[33,34], Mikael Landén[17,7], James F Wilson[6,25], Agneta Siegbahn[35], Lars Wallentin[35], Anders Mälarstig[7,8], Adam S Butterworth[1,2,20,21,31,38,*], James E Peters[1,21,36,38,*]

[1]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, United Kingdom
[2]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Papworth Road, Cambridge, United Kingdom
[3]Australian Centre for Precision Health, Unit of Clinical and Health Sciences, University of South Australia, Adelaide, South Australia, Australia
[4]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia
[5]Uppsala Clinical Research Center, Uppsala University, Uppsala, Sweden
[6]Centre for Global Health Research, Usher Institute, University of Edinburgh, Teviot Place, Edinburgh, United Kingdom
[7]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[8]Pfizer Worldwide Research, Development and Medical, Stockholm, Sweden
[9]Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
[10]Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, Uppsala, Sweden
[11]Department of Metabolism, Digestion and Reproduction, Faculty of Medicine, Imperial College London, London, United Kingdom
[12]School of Immunology and Microbial Sciences, King's College London, London, United Kingdom
[13]Nucleus Genomics ltd, New York, NY, USA
[14]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[15]Department of Endocrinology and Diabetology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[16]German Center for Diabetes Research (DZD), Munich-Neuherberg, Germany
[17]Institute of Neuroscience and Physiology, University of Gothenburg, Gothenburg, Sweden
[18]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[19]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany
[20]British Heart Foundation Centre of Research Excellence, School of Clinical Medicine, Addenbrooke's Hospital, University of Cambridge, Cambridge, United Kingdom
[21]Health Data Research UK, Wellcome Genome Campus and University of Cambridge, Hinxton, United Kingdom
[22]Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, United Kingdom
[23]University of Split Medical School, University of Split, Split, Croatia
[24]Division of Endocrinology and Diabetology, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany
[25]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Crewe Road, Edinburgh, United Kingdom
[26]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden
[27]Skåne University Hospital, Malmö, Sweden
[28]Wallenberg Centre for Molecular Medicine, Lund University, Lund, Sweden
[29]Clinical Memory Research Unit, Faculty of Medicine, Lund University, Lund, Sweden
[30]Department of Neurology, Skåne University Hospital, Lund University, Lund, Sweden
[31]NIHR Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge, United Kingdom
[32]Department of Human Genetics, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom
[33]Division of Rheumatology, Department of Medicine (Solna), Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden
[34]Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden
[35]Department of Medical Sciences and Uppsala Clinical Research Center, Uppsala University, Uppsala, Sweden
[36]Department of Immunology and Inflammation, Imperial College London, London, United Kingdom
[37]These authors contributed equally
[38]These authors jointly supervised the study

***e-mail:** Adam S Butterworth asb38@medschl.cam.ac.uk or James E Peters j.peters@imperial.ac.uk

ABSTRACT

Circulating proteins have important functions in inflammation and a broad range of diseases. To identify genetic influences on inflammation-related proteins, we conducted a genome-wide protein quantitative trait locus (pQTL) study of 91 plasma proteins measured using the Olink Target platform in 14,824 participants. We identify 180 pQTLs (59 *cis*, 121 *trans*). Integration of pQTL data with eQTL and disease GWAS provided insight into pathogenesis, implicating lymphotoxin-α in multiple sclerosis. Using Mendelian randomisation (MR) to assess causality in aetiology, we identify both shared and distinct effects of specific proteins across immune-mediated diseases, including directionally discordant functions for CD40 in rheumatoid arthritis versus multiple sclerosis and inflammatory bowel disease. MR implicated CXCL5 in the aetiology of ulcerative colitis (UC) and we show elevated gut *CXCL5* transcript expression in patients with UC. These results identify targets for existing drugs and provide a powerful resource to facilitate future drug target prioritisation.

INTRODUCTION

Inflammation is a physiological host response to infection or injury. However, aberrant inflammatory responses result in tissue damage and are central to the pathogenesis of multiple diseases including sepsis, autoimmunity and atherothrombosis. Inflammatory responses are orchestrated by a complex network of cells and mediators, including circulating proteins such as cytokines and soluble receptors. Therefore, discovery of the genetic determinants of abundance of inflammation-related circulating proteins should yield valuable insights into both physiology and the aetiology of a broad range of diseases.

Proteomic studies are informative as proteins are the effector molecules of most biological processes, and from a translational perspective, proteins are the targets of most drugs. The development of high-throughput proteomic technologies now allows for profiling of the plasma proteome at epidemiological scale. Coupling genomic and proteomic data enables identification of genetic variants associated with protein abundance, protein quantitative trait loci (pQTLs). pQTLs provide valuable insights into the molecular basis of complex traits and diseases, by identifying proteins that lie between genotype and phenotype. Recent years have seen a rapid increase in both the number and size of pQTL studies, transforming our understanding of the genetic architecture of the circulating proteome[1-11].

Here we extend previous work by performing pQTL mapping for 91 inflammation-related

proteins in 14,824 participants. We integrate these data with disease genome-wide association studies (GWAS) to characterise the functional effects of disease-associated variants. Using Mendelian randomisation and colocalisation analyses, we identify proteins that play a causal role in immune-mediated disease aetiology. Our results reveal both pathways that are known to be therapeutically important and new putative drug targets, including CD40 in rheumatoid arthritis, lymphotoxin-α (LTA) in multiple sclerosis, and the chemokine CXCL5 in ulcerative colitis.

RESULTS

Genetic architecture of circulating inflammatory proteins. We performed genome-wide pQTL mapping for 91 plasma proteins measured using the Olink Target Inflammation panel in 11 cohorts totalling 14,824 European-ancestry participants (**Supplementary Table 1, Supplementary Note 1**), and meta-analysed the results (**Extended Data Figure 1**). In order to provide a succinct and standardised nomenclature, we report proteins by the non-italicised symbols of the genes encoding them (see **Supplementary Table 2** for a mapping of symbols to full protein names and UniProt identifiers). We identified a total of 180 significant ($P \leq 5 \times 10^{-10}$, fixed-effects meta-analysis) associations between 108 genomic regions (see **Methods** for locus definition) and 70 proteins (**Figure 1, Supplementary Table 3**, **Supplementary Item, Supplementary Figures 1-2**). Of the 180 significant locus-protein associations, 59 (33%) were local-acting ('*cis*' pQTLs; defined here as a genetic variant lying within +/- 1 megabase of the gene encoding the associated protein) and 121 (67%) were distant-acting ('*trans*'). We found evidence of *trans*-pQTL hotspots associated with multiple proteins (e.g. rs3184504 at the *SH2B3* locus was associated with CXCL9, CXCL10, CXCL11, CD5, CD244, and IL12B) (**Figure 2a**).

For 70 (77%) of the 91 proteins studied, we identified at least 1 significant pQTL, including 59 (65%) proteins that had a *cis*-pQTL. Of these 70 proteins, 19 had only *cis*-pQTL(s), 11 had only *trans*-pQTL(s), and 40 had both *cis*- and *trans*-pQTLs. For 18 of the 21 proteins for which no pQTL was detected, >50% of samples had levels below the lower limit of detection (LLOD), suggesting that the lack of genetic signal is due to low protein abundance in plasma (**Extended Data Figure 2a**). The number of genomic loci associated with each protein ranged between 1 and 8 (**Figure 2b**), but was less than 4 for the majority of proteins. Examples of multi-locus-regulated proteins include TNFSF10 and IL12B, both of which had 1 *cis*- and 7 *trans*-pQTLs (**Figure 2c,d**). Conditional analyses revealed the presence of an additional 47 independent signals, raising the total number of pQTL signals from 180 (59 *cis*, 108 *trans*) to 227 (99 *cis*, 128 *trans*) (**Supplementary Table 4**).

To validate our pQTL results, we tested significant associations from our discovery meta-analysis for replication in an independent cohort (ARISTOTLE) comprising 1,585 participants with Olink plasma proteomic data[12]. Of the 180 pQTL signals, we were able to test 174 in the ARISTOTLE data, of which 168 had a directionally consistent effect estimate. There was a strong correlation (Pearson r=0.97) between the pQTL effect estimates in ARISTOTLE and in the discovery meta-analysis; this correlation was consistent for both *cis*- and *trans*-pQTL effect sizes (r=0.99 and r=0.94, respectively) (**Extended Data Figure 3**)**.** Out of the 174 pQTL signals, 32 replicated at $P \leq 5 \times 10^{-10}$ (linear regression) and 72 at $P \leq 2.8 \times 10^{-4}$ (a Bonferroni-corrected threshold), respectively (**Supplementary Table 5**). We also tested our significant pQTLs for replication in 35,556 Icelanders from the deCODE study[9], which assayed plasma proteins using the aptamer-based SomaScan platform (**Supplementary Note 2**). 72 of the 91 proteins in our study were measured in the deCODE study. Of the 158 locus-protein associations that could be tested, 75 were significant at $P \leq 5 \times 10^{-10}$ (linear regression), and 96 were significant at $P \leq 2.8 \times 10^{-4}$. Overall, we replicated 126 (71%) of the 178 testable pQTLs in either ARISTOTLE or deCODE at $P \leq 2.8 \times 10^{-4}$ (linear regression) (**Supplementary Note Table 1**).

In line with other GWAS, we observed an inverse relationship between effect size and minor allele frequency (MAF), with rarer pQTL variants generally showing larger effect sizes (**Extended Data Figure 4a**). The proportion of variance explained by the significant sentinel variants from our discovery meta-analysis varied from 0.003 for NTF3 to 0.285 for CCL8 (**Extended Data Figure 4b**).

Annotation and characterisation of *cis*-pQTLs. Of the 59 *cis*-pQTLs identified, 11 sentinel variants were protein-altering variants (PAVs) (10 missense and 1 splice acceptor). A further 10 sentinel variants were in high linkage disequilibrium (r^2>0.8) with a protein-altering variant (all missense); of these, 7 were variants in the gene encoding the target protein itself and 3 in another nearby gene (**Supplementary Note 3**). PAVs can result in false positive *cis*-pQTL signals by altering protein epitopes recognised by antibodies used in proteomic assays[13]. However, they can also impact the abundance of plasma proteins through several mechanisms, including protein translation, secretion into the circulation, enzymatic cleavage of pre-proteins, and protein clearance and degradation. Alternatively, plasma protein abundance can also be affected by altered transcriptional regulation in blood cells or other tissues.

We next examined the degree to which our 59 *cis*-pQTLs were explained by corresponding

cis-eQTLs, by comparing our findings with publicly available *cis*-eQTL data. In a meta-analysis of whole blood eQTL data from the eQTLGen Consortium[14], we found a genome-wide significant ($P \leq 5 \times 10^{-8}$; meta-analysis) *cis*-eQTL for 40 of the 59 *cis*-pQTLs, where the *cis*-eQTL target gene encodes the *cis*-pQTL target protein. However, statistical colocalisation analyses showed that only six (rs34790908-*TNFSF12*, rs72912115-*TGFA*, rs471994-*MMP1*, rs674379-*CD5*, rs450373-*CXCL5*, rs5744249-*IL18*) of these *cis*-eQTLs colocalised (posterior probability [PP]≥0.8) with their cognate *cis*-pQTLs (**Supplementary Table 6**), indicating that the remaining 34 eQTL-pQTL pairs may not share the same underlying causal genetic variant. Examination of regional association plots confirmed that the majority of blood eQTL and pQTL signals were distinct (**Supplementary Figure 3**)**.** Of the 6 colocalising eQTL-pQTL pairs, 5 were directionally consistent. However, the eQTL and pQTL for IL18 at rs5744249 were oppositely associated with the mRNA and protein levels. rs5744249 resides in intron 2 of *IL18* and is in high LD (r^2>0.8) with a 3' UTR variant (rs5744292, r^2=0.98 | 1000G EUR), but no PAVs. Therefore, the directional discordance is not easily explained either by an artefactual pQTL signal due to altered antibody binding or by a difference in the release of IL18 into the circulation due to differences in protein structure, but may instead relate to differential post-transcriptional regulatory mechanisms or contributions of different cell types to the plasma pQTL versus whole blood eQTL. Indeed, directional uncoupling of eQTL-pQTL pairs has been previously reported[8], and eQTL directional discordance has been observed between different tissues[15] or even within different leucocytes[16].

Since tissues other than blood are the primary source of many plasma proteins, we explored eQTL data across a range of tissues and cell types from the Genotype-Tissue Expression (GTEx) (v8) project[15] and the eQTL Catalogue[17]. Systematic COLOC analyses revealed colocalising (PP≥0.8) *cis*-eQTLs in at least one tissue or cell type for 32 of the 59 *cis*-pQTLs (**Supplementary Tables 7-8**); 16 were highlighted by both eQTL resources, 12 by GTEx only, and the remaining 4 by the eQTL Catalogue only. These included all 6 colocalising *cis*-eQTLs from eQTLGen. These findings suggest that at least 50% of our *cis*-pQTLs may be driven by underlying cognate *cis*-eQTLs. In most cases, colocalisation (PP H4≥0.8) between *cis*-eQTL-pQTL pairs was observed across two or more distinct tissues or cell types, up to a maximum of 41 (for rs1883832-*CD40*). In other cases, colocalisation was observed in just a single tissue or cell type (e.g. the colocalising *cis*-eQTL signal (rs62360376) for *GDNF* was found only in skeletal muscle). Of the 27 *cis*-pQTLs without a corresponding colocalising *cis*-eQTL, for 12 the sentinel variant or a proxy in high LD was a PAV (**Supplementary Note Table 3**).

Identifying the mediators of *trans*-pQTLs. We sought to identify the most likely gene mediators of the *trans*-pQTLs using the ProGeM bioinformatics tool[18], which utilises genomic

(e.g., *cis*-eQTL) and biological (e.g., gene ontology and pathways) annotation data from multiple sources. For some *trans*-pQTLs, we identified strong evidence to implicate a gene encoded near the pQTL as mediating the distant association with the target protein. Examples included receptor-ligand pairs such as IL-6-IL6-R, IL-10-IL-10RA, CCL2-CCR2, CCL4-CCR5, and CCL11-CCR3. We also identified genes mediating pQTLs through intracellular signalling pathways rather than direct ligand-receptor interactions. An example is rs385076, an intronic variant in *NLRC4*, which is a *trans*-pQTL for IL-18. IL-18 is synthesised as an inactive precursor (pro-IL-18), which is cleaved by caspase-1 in the NLRC4 inflammasome to produce the active form of IL-18 (**Figure 3a**). Since rs385076 is also a *cis*-eQTL for the inflammasome gene *NLRC4* (**Figure 3b**), together, these QTL data suggest that genetic variation in *NLRC4* alters its expression and thereby inflammasome activity, with consequent effects on circulating IL-18 levels.

Following a manual literature review to refine the ProGeM output, we narrowed down the most likely mediating gene(s) to either one or two candidates for 100 of the 121 *trans*-pQTLs (**Supplementary Table 9**). For 94, one of the three nearest genes to the sentinel variant was the primary candidate. In several instances where either one or two candidate genes were prioritised, ProGeM revealed functional links between both (i) the sentinel variant and the nearby candidate mediating gene (e.g., *cis*-eQTL) and (ii) the same candidate mediating gene and the *trans*-affected protein(s) (e.g., through protein-protein interaction). We have previously shown that such convergence on the same gene is indicative of a strong candidate[18]. An example of this is the *trans*-pQTL at rs12075, which is associated with multiple chemokines (CCL2, CCL7, CCL8, CCL11, CCL13, CXCL6) that attract and activate leucocytes. rs12075 is a missense variant and a *cis*-eQTL in whole-blood for the *DARC* gene, which encodes the atypical chemokine receptor 1 (ACKR1) protein. STRINGdb analysis revealed that ACKR1 is an interacting partner for 3 (CCL2, CCL7, CCL8) of the 6 *trans*-affected chemokines. Previous studies have shown that ACKR1 acts as a negative regulator of inflammation by non-specifically binding both the CCL and CXCL family of chemokines[19], suggesting an explanation for the multiple chemokine associations at this variant. Potentially downstream of its effects on chemokines, rs12075 is also associated with white blood cell count, as well as monocyte and basophil count[20] (**Extended Data Figures 5-6**).

We found that plasma levels of some proteins were associated with numerous genetic loci, with IL12B, KITLG, and TNFSF10 regulated by seven genetic loci each. We hypothesised that the mediating genes at each of the regulatory loci for a given protein might be functionally related, enabling identification of shared pathways and/or the most likely mediating gene(s).

We therefore generated protein-protein interaction networks for each of these multi-locus-regulated proteins and their respective candidate mediating genes (**Extended Data Figure 7**). For TNFSF10, the network analysis linked genetic regulators of TNFSF10 to the plasminogen-activating system (**Extended Data Figure 7a**, **Supplementary Note 4**). For KITLG, a driver of hematopoiesis[21], we found a cluster of interacting proteins, including PON1, ABCA1, PLTP, (**Extended Data Figure 7b**) converging on cholesterol metabolism. Supporting this, we found that 5 of the 7 *trans*-pQTLs for KITLG were significantly ($P \leq 5x10^{-8}$, linear regression) associated with levels of either HDL or LDL cholesterol, and some with other lipids such as triglycerides and blood cell traits (**Supplementary Table 10**). Our findings therefore suggest a link between plasma KITLG levels, cholesterol metabolism and altered hematopoiesis.

Overlap with GWAS of traits and diseases. Genome-wide association studies (GWAS) have identified thousands of genomic regions associated with common diseases[22], including immune-mediated diseases (IMDs). Many of these disease-associated loci lie outside protein-coding regions, leaving the effector molecules and pathways by which these genetic variants confer disease risk unclear. Integration of pQTL and GWAS data can help bridge this knowledge gap by linking disease risk loci to specific proteins. To this end, we looked for overlap between pQTLs, or proxy variants in high LD ($r^2 \geq 0.8$) with our sentinel variants, and disease-associated variants from GWAS. This revealed overlap between our pQTLs and disease-associated variants for 73 diseases (**Extended Data Figure 8, Supplementary Table 11**). Examples of genetically anchored protein-disease connections included: TNFSF11 (RANKL) with osteoporosis and hypothyroidism; NGF (nerve growth factor) with migraine; TNFSF12 (TWEAK) with hypertension; and FGF5 with hypertension and cardiovascular diseases.

We next focussed on IMDs in more detail, intersecting our pQTLs with IMD GWAS data to identify proteins linking genotype and disease phenotypes. We found that 31 of our pQTLs overlap GWAS hits for at least one common IMD, with 76 unique pQTL-protein-disease associations (**Supplementary Table 12, Extended Data Figure 9**). For example, we observed that a *cis*-pQTL for IL-10 was also associated with risk of inflammatory bowel disease (IBD), with the allele associated with higher plasma IL-10 correlating with reduced IBD risk, consistent with the anti-inflammatory effects of IL-10. Some pQTLs showed diverging directions of effect on different diseases (e.g. the *trans*-pQTL at *IL6R* for plasma IL-6 levels described earlier had opposing directions of effect on risk of rheumatoid arthritis and allergic diseases (**Extended Data Figure 9**), as previously described[23,24]).

***Trans*-pQTL implicates the *LTBR-LTA* axis in multiple sclerosis.** We identified a t*rans*-pQTL for lymphotoxin α (LTA, also known as TNF-β) at rs2364485 on chromosome 12 (**Table 1**), an intergenic variant previously found to be associated with multiple sclerosis[25]. We found that the multiple sclerosis risk allele, rs2364485:A, was associated with higher plasma levels of LTA. We next applied the ProGeM algorithm which revealed two candidate genes in the region near the pQTL that might mediate the *trans*-pQTL: *TNFRSF1A* (encoding tumour necrosis factor receptor 1, TNFR1) and *LTBR* (encoding lymphotoxin beta receptor, LTBR). LTA is a ligand for TNFR1, but also can bind the membrane-bound receptor LTBR when in complex with LTB. Functional studies have shown that *TNFRSF1A* is the causal gene underlying a neighbouring independent multiple sclerosis association in the region, about 70kb upstream from rs2364485. The sentinel variant at this neighbouring signal, rs1800693, results in an alternative *TNFRSF1A* isoform due to skipping of exon 6[26]. We therefore sought to determine if *TNFRSF1A* is also the likely mediating gene for the LTA *trans*-pQTL at rs2364485, or whether *LTBR* is the more likely candidate. Through mining of eQTL databases, we found that rs2364485 is a *cis*-eQTL for *LTBR* (but not *TNFRSF1A*) in multiple tissues, including in the eQTLGen consortium meta-analysis of whole-blood[14], with the multiple sclerosis risk allele (rs2364485:A) associated with reduced *LTBR* mRNA. Pairwise statistical colocalisation analyses using conditioned *LTBR* eQTL data (from eQTLGen) and multiple sclerosis GWAS data[25] (**Methods**) showed that the rs2364485 *trans*-pQTL signal for LTA colocalises with both *LTBR* mRNA expression in whole blood (PP=0.79) and multiple sclerosis (PP=0.86) (**Figure 4**). Taken together, these data are consistent with a pathogenic model whereby the multiple sclerosis risk allele results in lower abundance of LTBR (the receptor) and consequently higher circulating levels of the ligand LTA.

Mendelian randomisation to identify protein drivers of IMDs. Observational studies comparing patients with IMDs to healthy controls have identified many proteins that are dysregulated. However, it is often unclear whether such proteins play causal roles in the disease process or are merely downstream markers. Distinguishing these possibilities is important therapeutically, as pharmacological targeting of the latter is unlikely to be beneficial. We therefore applied Mendelian randomisation (MR), an approach that tests the causal role of a risk factor ('exposure') in a disease in observational data using genetic variants as instrumental variables[27]. We used the 58 proteins with *cis*-pQTLs outside the HLA region in our study as exposures and 14 IMDs as outcomes (**Methods)**. By restricting our use of genetic instruments to *cis*-pQTLs, we reduce the likelihood of violating MR assumptions through horizontal pleiotropy. Using Generalised Summary-data-based MR (GSMR)[28], we found 22 significant (FDR<0.01) putative causal associations (**Figure 5**, **Supplementary Table 13**). To evaluate the robustness of these associations, we performed additional checks including

evaluating the strength of the disease association in the GWAS summary statistics and whether there might be confounding due to LD (**Methods, Supplementary Table 14**). After applying these filters, 10 disease-protein pairs with robust evidence remained (**Table 1**). These results highlighted a number of established links between proteins and inflammatory diseases that are supported by other lines of evidence. For example, our finding that genetic predisposition to higher plasma levels of IL-12B (a subunit of IL-12) was associated with increased risk of IBD is consistent with the therapeutic benefit of ustekinumab (a monoclonal antibody targeting the p40 subunit common to IL-12 and IL-23) in IBD (**Supplementary Table 15**).

Our MR analysis implicated CXCL5, a chemokine that acts on neutrophils, in the aetiology of ulcerative colitis (UC). The plasma *cis*-pQTL for CXCL5, colocalised with *cis*-eQTLs for *CXCL5* in both blood and gut tissue, and with the UC GWAS signal (**Figure 6a**). To further explore the role of CXCL5 in UC, we compared expression of *CXCL5* transcripts in gut samples from patients with IBD and healthy controls using IBD TAMMA, an open-access resource for interrogating transcriptomic data across multiple datasets[29]. We observed that *CXCL5* gene expression was significantly increased in mucosal biopsies from patients with UC in comparison with biopsies from healthy control participants (log_2 fold change 7.07, P $1.98x10^{-174}$, Wald test) (**Figure 6b**). Indeed, *CXCL5* was the third most highly upregulated transcript across the transcriptome (**Figure 6c**). We replicated these findings in three independent datasets (**Figure 6d**). Of note, our MR analysis revealed the association of CXCL5 was restricted to UC (unadjusted P=$2.3x10^{-6}$, GSMR), with no significant association in CD (unadjusted P=0.4) (**Figures 6a,e**). Supporting this specific pathogenic effect, *CXCL5* gene expression in gut samples from IBD patients was higher in UC than in CD (**Figure 6b**). Counter-intuitively (given the upregulation of *CXCL5* in UC patient tissue samples), evaluation of the direction of MR association effect revealed that genetic susceptibility to higher plasma CXCL5 reduces risk of UC (**Figure 6e**). This effect was consistent across 12 of the 13 the individual genetic variants used in the MR score (**Extended Data Figure 10a**). We found consistent directions of effect for the CXCL5 plasma pQTL and the blood and gut eQTLs (**Extended Data Figure 10b**), indicating that our results are generalizable at both the mRNA and protein level and across local and systemic sites. Together these data indicate that genetic tendency to lower CXCL5 is a causal risk factor for development of UC, despite the strong upregulation of CXCL5 once disease develops.

We observed that genetic predisposition to higher plasma CD40 levels was associated with increased rheumatoid arthritis risk, consistent with evidence from both animal models and

humans implicating the CD40 pathway in rheumatoid pathogenesis[30]. In addition, our MR analysis identified a potential causal role for the CD40 pathway in IBD (including both Crohn's and UC) and multiple sclerosis. Interestingly, however, the MR associations for these diseases had the opposite direction of effect compared to rheumatoid arthritis i.e. genetic predisposition to *lower* plasma CD40 levels was associated with *higher* risk of IBD and multiple sclerosis. These findings highlight how the same pathway can have pleiotropic effects on disease susceptibility, but also point to the complexity of immune-mediated disease pathogenesis, with opposing effects on different diseases.

DISCUSSION

Here, we performed a large-scale pQTL GWAS of 91 circulating inflammation-related proteins measured using Olink immunoassays, identifying 180 significant primary pQTL signals (59 *cis*, 121 *trans*). Colocalisation analysis suggested that only a small proportion of the plasma *cis*-pQTLs reported here are underpinned by the same causal genetic variant as the whole blood *cis*-eQTL for the corresponding gene. Of note, the plasma proteome is not the direct corollary of the whole blood transcriptome: plasma pQTL studies examine genetic effects on extracellular protein levels whereas blood eQTL studies examine the effects on intracellular RNA levels (predominantly in leucocytes). This has several implications. First, plasma protein levels can be affected by non-transcriptional mechanisms including cleavage, secretion and clearance. Second, a wide range of tissues other than blood cells (e.g. the liver) contribute to the plasma proteome. This is evident when considering circulating proteins that are measured as biomarkers in clinical practice (e.g. albumin produced by the liver, troponin by the heart, PSA by the prostate). Indeed, by extending our comparison across multi-tissue eQTL databases, we showed that at least 50% of the *cis*-pQTLs we observed are likely driven by cognate *cis*-eQTLs in a diverse range of tissues and cell-types. Blood eQTL studies have been carried out using sample sizes similar to the sample size in our pQTL study. eQTL studies in other tissues tend to be smaller, and so it is likely that some of the plasma *cis*-pQTLs observed here are underpinned by tissue-specific eQTLs that have not yet been detected due to lack of statistical power. Finally, other mechanisms such as alternative splicing might account for some *cis*-pQTLs without corresponding eQTLs.

Our pQTL study identified twice as many *trans* associations compared to *cis* (121 versus 59, respectively), in keeping with other well-powered pQTL studies (e.g., [7-9]). The integration of *cis*-pQTLs (and *cis*-eQTLs) with GWAS data provides useful, if sometimes obvious, insights into the upstream mechanisms of disease, since the mediating gene has usually already been

suspected by virtue of the location of the GWAS signal. In contrast, *trans*-pQTLs represent a double-edged sword for interpreting genetic associations with disease. On the one hand, they often represent a less direct link from genotype to disease than *cis*-pQTLs, and from the perspective of causal inference analysis, are more vulnerable to violating the assumptions of MR through horizontal pleiotropy. On the other hand, they can reveal important molecular mediators of disease encoded by genes distant from the disease GWAS signal. For example, we identified a *trans*-pQTL (rs2364485) for LTA at a multiple sclerosis risk locus. This multiple sclerosis risk locus contains two plausible causal genes (*TNFRSF1A*, *LTBR*) and two independent signals for multiple sclerosis risk (rs1800693, rs2364485). By integrating whole blood eQTL and multiple sclerosis GWAS data, we showed that *LTBR* is the most likely gene meditating our LTA *trans*-pQTL at rs2364485, and one of the multiple sclerosis signals at the locus. LTA is a member of the TNF superfamily of proteins and is the only member of this superfamily that is generated as a secreted protein rather than through cleavage of a membrane-bound protein. The multiple sclerosis risk allele is associated with lower expression of *LTBR* and higher circulating protein levels of LTA, a component of its ligand. This raises the question as to whether elevated LTA is secondary to lower LTBR, or vice versa (e.g. through compensatory receptor downregulation). The distinction between *cis*- and *trans*-QTLs enables us to address this. Given that the eQTL for *LTBR* is *cis*, and the pQTL for LTA is *trans*, it is highly likely that former is the upstream effect, with the higher levels of soluble LTA occurring as a result of reduced binding to its receptor. This demonstrates the value of pairing QTLs for ligands and their receptors for deconvoluting the ordering of biological pathways.

Integration of pQTLs with GWAS disease signals revealed disease-protein connections reflecting both established and plausible putative mechanisms of pathophysiology. For example, a *cis*-pQTL for TNFSF11 (RANKL) overlapped with GWAS signals for osteoporosis and hypothyroidism. The former is consistent with RANKL's well-established role in bone biology and RANKL is the target of the anti-osteoporosis drug denosumab[31]. However, RANKL also plays a role in the immune system[32], and these effects may be relevant to risk of autoimmune hypothyroidism. A *cis*-pQTL for TNFSF12 (TWEAK) was associated with risk of hypertension. TWEAK is a cytokine predominantly produced by leucocytes and has pleiotropic actions, including on the endothelium[33,34], potentially explaining the association with blood pressure. A *cis*-pQTL for FGF5 was also associated with susceptibility to hypertension and cardiovascular diseases, with the allele associated with higher plasma FGF5 levels associating with lower risk of cardiovascular diseases. Consistent with this, there are reports that FGF5 has cardioprotective effects in pig models[35].

31 of our pQTLs overlap GWAS hits for at least one common IMD. Disease-protein links

identified from this analysis highlighted commonalities in pathogenesis between specific IMDs, mirroring the overlap in clinical manifestations. However, the contributions of proteins to IMD risk were sometimes complex, with the same protein conferring risk of one IMD but protecting from another. For example, genetic predisposition to higher levels of soluble IL6 had opposing effects on risk of rheumatoid arthritis and allergic disease. We observed a similar phenomenon for CD40, with genetic predisposition to higher CD40 increasing risk of RA but protecting against IBD and multiple sclerosis.

The development of biologic therapies targeting specific inflammatory proteins has transformed the clinical management of immune-mediated diseases[36]. Understanding which proteins are drivers of disease and distinguishing these from proteins that are simply markers of inflammation is therefore important for development of new treatments. To this end, we used Mendelian randomisation to evaluate the causal contributions of proteins to different IMDs. Our results identify pathways that are already the target of existing drugs (e.g. IL-12B in IBD), providing confirmation of the utility of this approach, and also highlight new potential therapeutic targets.

One such example was the CD40 pathway in rheumatoid arthritis. CD40 is a stimulatory receptor constitutively or inducibly expressed on both immune and non-immune cells [37]. Its ligand, CD40L, is expressed primarily on activated T cells but also on a range of other immune and non-immune cells. CD40L-CD40 binding triggers immune cell activation and proliferation and inflammatory cytokine production, and the differentiation of B cells into IgG-secreting plasma cells, making it central to antibody responses. In a murine model of inflammatory arthritis, knock out or inhibition of the CD40 pathway resulted in reduced inflammation[38]. Observational studies have demonstrated upregulation of CD40L in the blood and tissues of patients with RA and other autoimmune rheumatic diseases[30,39]. These findings previously motivated development of drugs targeting the CD40 pathway in RA and other IMDs, but anti-CD40L therapy was complicated by thrombosis due to cross-linking CD40L on platelets. Therapeutic targeting of CD40 rather than CD40L may avoid this. Our MR results suggest rheumatoid arthritis as a candidate for this approach. However, the directionally discordant effects we observed of CD40 on RA versus multiple sclerosis and IBD raises the possibility of triggering other forms of immune-mediated diseases as a side-effect of anti-CD40 therapy. This has some parallels with therapies targeting TNF, which are effective in rheumatoid arthritis but not in multiple sclerosis, and indeed can worsen multiple sclerosis or provoke *de novo* central nervous system demyelination[40,41].

Our MR findings implicate CXCL5 in the aetiology of UC, where genetic susceptibility to higher

levels of plasma CXCL5 was associated with lower UC risk. Examination of eQTL data revealed this observation was consistent at the RNA level in both the blood and gut tissue. By contrast, in our case-control analysis comparing gut tissue from UC patients versus controls, *CXCL5* is one of the most upregulated transcripts. A previous study reported that serum levels of CXCL5 are higher in IBD patients than controls[42]. Recent studies using UC gut tissue have implicated upregulation of genes encoding neutrophil-targeting chemokines, including *CXCL5*, by non-immune cells as correlating with important histopathological features, such as ulceration, and differentiating patient trajectories, including their responsiveness to different treatments[43,44]. Targeting CXCR2, the receptor for CXCL5, significantly attenuates animal models of UC[44]. One possible explanation that may reconcile these apparently contradictory findings is that genetic tendency to lower CXCL5 expression increases UC risk through impaired mucosal immune homeostasis, but that elevated CXCL5 is an important driver of tissue injury once disease is initiated. By analogy, a non-coding genetic variant associated with lower gene and protein expression of *TNFSF15* (encoding the inflammatory cytokine TL1A) in monocytes and macrophages increases IBD susceptibility[45], but TL1A is upregulated both systemically and in the gut in patients with active IBD[46,47,] and anti-TL1A therapies have recently shown efficacy in IBD in phase 2 randomised trials (NCT05013905 and NCT04996797[48]).

Our study has several limitations. Our pQTL analysis was restricted to 91 proteins, limiting the generalisability of our findings, particularly with regards to genetic architecture. Since this was a pQTL meta-analysis, study-level technical variation resulted in heterogeneity, which necessitated filtering out of potentially spurious associations that were inconsistent across cohorts. There is a risk that some true biological signals were also removed in this process. Very large single cohorts with standardised sample processing such as UK Biobank will avoid this issue. Our meta-analysis consisted predominantly of general population cohorts without inflammatory disease. There may be context-specific pQTLs that are only present during infection or inflammation, which our study may not have detected. By analogy, eQTL studies using human immune cells stimulated *in vitro* (e.g. with lipopolysaccharide or interferon) have demonstrated eQTLs that are not present in resting cells but become apparent in the context of cellular activation[49,50]. Conducting well-powered pQTL studies in patients with inflammation will be an important future research endeavour. Where proteins exist in both membrane-bound and cleaved states, it is not always clear whether plasma proteomic assays are exclusively capturing the soluble form or also protein from cell membranes (e.g. arising from in vivo sources such as exo-/ectosomes, or *ex vivo* processes such as venepuncture or sample processing). This complicates the interpretation of the direction of effect from MR analysis. Future well-powered studies examining genetic determinants of cell surface protein

expression measured through flow cytometry would provide valuable complementary information to aid the interpretation of plasma pQTL studies. Finally, as with all epidemiological-scale pQTL studies, proteins were measured in plasma (i.e. the extracellular component of blood), which may not always be the disease-relevant biological compartment, and where the direction of genotype-expression association may even be opposite to the site of inflammation. Thus, future tissue- and cell-specific pQTL studies will be valuable to understand differences in genetic signals across tissues.

In summary, we have used a large international consortium to identify the genetic determinants of a set of inflammation-related proteins, providing insight into the aetiology of immune-mediated diseases. The pQTL summary statistics generated in this study will be a valuable resource for interrogating future disease GWAS and guiding drug target identification and prioritisation.

Acknowledgements

This work was performed under the auspices of the SCALLOP Consortium.
We thank:
-study participants from the contributing cohorts;
-the International Multiple Sclerosis Genetics Consortium, who provided multiple sclerosis GWAS summary statistics used in our analyses;
-Aikaterina Siopi and Dianna McLeod for support with SCALLOP Consortium administration;
-the authors of the GCTA software for advice;
-Bram Prins for help with INTERVAL study genotype data quality control;
-Arianne Richard for comments on the manuscript.

J.E.P was supported by a grant and a fellowship from the Medical Research Foundation grant (MRF-042-0001-RG-PETE-C0839, MRF-057-0003-RG-PETE-C0799). E.J.N. was supported by the Schmidt Science Fellows, in partnership with the Rhodes Trust. P.S. was supported by a Rutherford Fund Fellowship from the Medical Research Council (grant no. MR/S003746/1). J.D. holds a British Heart Foundation Professorship and a NIHR Senior Investigator Award*. C.H. is supported by an MRC University Unit Programme Grant “QTL in Health and Disease" (U.MC_UU_00007/10).

Funding of the GWAS and proteomics studies of STABILITY and ARISTOTLE were supported by GlaxoSmithKline, BristolMyersSquibb, and the Swedish Foundation for Strategic Research (grant number RB13-0197).

The Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to J.F.W., the MRC Human Genetics Unit quinquennial programme “QTL in Health and Disease”, Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). DNA extractions were performed at the Edinburgh Clinical Research Facility, University of Edinburgh. We would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney. For the purpose of open access, the author has applied a Creative

Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Participants in the INTERVAL trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping were co-funded by the National Institute for Health and Care Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The academic coordinating centre for INTERVAL was supported by core funding from the: NIHR Blood and Transplant Research Unit (BTRU) in Donor Health and Genomics (NIHR BTRU-2014-10024), NIHR BTRU in Donor Health and Behaviour (NIHR203337), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and NIHR Cambridge BRC (BRC-1215-20014; NIHR203312) [*] and has received funding from an EC-Innovative Medicines Initiative (BigData@Heart). The academic coordinating centre thank blood donor centre staff and blood donors for participating in the INTERVAL trial. This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Estonian Biobank work was supported by the European Regional Development Fund and the programme Mobilitas Pluss (MOBTP108), No. 2014-2020.4.01.15-0012 GENTRANSMED and 2014-2020.4.01.16-0125 This study was also funded by the EU H2020 grant 692145, by the Estonian Research Council Grant PUT1660 and by the Estonian Research Council grant PUT (PRG1291). Data analyzes with Estonian datasets were carried out in part in the High-Performance Computing Center of University of Tartu.

The SWEBIC biobank was supported by the Stanley Medical Research Institute. The proteomic analyses in SWEBIC was funded by the Swedish foundation for Strategic Research (KF10-0039). For RECOMBINE and SWEBIC, the data handling and analysis were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

The CROATIA-Vis study was funded by grants from the Medical Research Council (UK), from the Republic of Croatia Ministry of Science, Education and Sports (108-1080315-0302; 216-1080315-0302) and the Croatian Science Foundation (8875). We thank the staff of several institutions in Croatia that supported the field work, including Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb, the recruitment team from the Croatian Centre for Global Health, University of Split and all the study participants.

The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ.

The measurement of inflammatory biomarkers was funded by a grant from the German Center for Diabetes Research (DZD; to C. Herder and B. Thorand). This work was also supported by the Ministry of Culture and Science of the State of North Rhine-Westphalia and the German Federal Ministry of Health. This study was supported in part by a grant from the German Federal Ministry of Education and Research to the German Center for Diabetes Research

(DZD).

NP is supported by a Wellcome Trust Discovery award (225875/Z/22/Z). DC is supported by the NIHR Imperial Biomedical Research Centre (BRC). Infrastructure support for this research was provided by the NIHR Imperial Biomedical Research Centre (BRC).

Support for title page creation and format was provided by AuthorArranger, a tool developed at the National Cancer Institute.

Author Contributions

J.E.P., N.E., E.M-D., A.K.H., A.K., S.E., L.F., C.H., L.J., S.E.B., P.S. conducted study-level analyses.
C.G., D.S.P., O.P., B.T., H.G., M.R., U.V., T.O., C.H., A.J., U.G., N.P., O.H., N.M-C., P.K.J., J.D., L.P., L.K., M.L., J.F.W., A.S., L.W., A.M., A.S.B., and J.E.P. provided data and study supervision.
D.C., J.D-B., N.P. collected IBD samples and generated and analysed the IBD RNA-seq data.
J.H.Z., D.S., A.K., J.M., P.S, conducted the meta-analysis and downstream analyses.
J.H.Z, D.S., E.N., A.S.B, and J.E.P. drafted the manuscript.
J.F.W., A.M., A.S.B., and J.E.P. conceived the project.
All authors critically reviewed the manuscript and gave final approval to publish.

Competing interests

J.D. serves on scientific advisory boards for AstraZeneca, Novartis, and UK Biobank, and has received multiple grants from academic, charitable and industry sources outside of the submitted work. A.S.B. has received grants unrelated to this work from AstraZeneca, Bayer, Biogen, BioMarin, Bioverativ, Novartis and Sanofi. J.E.P. has received hospitality and travel expenses to speak at Olink-sponsored academic meetings (none within the past 5 years). During the drafting of the manuscript, D.S.P. became a full-time employee of AstraZeneca, and P.S. became a full-time employee of GlaxoSmithKline. M.L. has received lecture honoraria from Lundbeck pharmaceutical. The other authors declare no competing interests.

Protein	Disease	OR (95% CI)	P
CD40	Rheumatoid arthritis	1.28 (1.21-1.37)	$1.4x10^{-15}$
CD40	Multiple sclerosis	0.75 (0.70-0.82)	$1.2x10^{-12}$
CD40	Crohn's disease	0.81 (0.75-0.87)	$2.2x10^{-8}$
CD40	IBD	0.87 (0.82-0.92)	$1.9x10^{-6}$
CD5	Primary sclerosing cholangitis	0.50 (0.35-0.70)	$8.1x10^{-5}$
CD6	IBD	1.10 (1.06-1.14)	$2.1x10^{-7}$
CXCL5	Ulcerative colitis	0.79 (0.72-0.87)	$2.3x10^{-6}$
IL-12B	IBD	1.38 (1.31-1.46)	$1.5x10^{-30}$
IL-12B	Ulcerative colitis	1.38 (1.29-1.48)	$4.7x10^{-20}$
IL-18R1	Eczema	1.15 (1.10-1.20)	$2.1x10^{-10}$

Table 1. Putative causal protein-disease associations from MR analysis. IBD = inflammatory bowel disease (i.e. based on GWAS where Crohn's disease and ulcerative colitis cases are grouped together). P = two-sided *P*-value for the causal estimate of protein on disease from the GSMR package. FDR = false discovery rate. OR = odds ratio associated with a 1 standard deviation increase in the protein level. OR>1 indicates genetic propensity to higher levels of the plasma protein is associated with higher disease risk, and OR<1 with reduced risk. CI = confidence interval.

Figure 1. **Genomic map of genetic determinants of inflammation-related proteins.** Circos plot linking the location of pQTLs to the gene encoding their associated proteins. Labels for the *cis*-pQTLs (red) indicate the gene encoding the target protein. For the *trans*-pQTLs (blue), the gene symbols of the target proteins are indicated along with the putative mediating gene(s) at the *trans*-pQTLs in square brackets. $-\log_{10}$(*P*-values) are capped at 150 for readability. Two-sided *P*-values are from meta-analysis of linear regression estimates.

Figure 2. **Genetic architecture of 91 inflammation-related proteins. a)** Distribution of the number of identified pQTLs per protein. The HLA region was treated as a single region. **b)** Circos plot showing the six *trans*-pQTLs for the *SH2B3* 'hotspot' locus on chromosome 12. **c)** Manhattan plots showing genetic associations with plasma abundance of IL-12B and **d)** TNFSF10 (TRAIL). The horizontal red line indicates statistical significance ($P=5x10^{-10}$). Two-sided *P*-values are from meta-analysis of linear regression estimates. Nearest genes in the region of pQTL signals are annotated.

Figure 3. Genetic regulation of the inflammasome affects plasma IL-18 levels. a) Schematic illustrating the cleavage of pro-IL-18 by caspase 1 and subsequent secretion of mature IL18 from the cell into the extracellular space. **b)** Regional association plots around *NLRC4* showing: the *trans*-pQTL signal for plasma IL18 protein (top) from this study (n=14,824), and the *cis*-eQTL signal for *NLRC4* (bottom) in whole blood from the eQTLGen study (n=31,684)[14]. Purple diamond = sentinel pQTL variant. Other variants coloured by LD to

sentinel pQTL. Two-sided *P*-values are from meta-analysis of linear regression estimates.

Figure 4. **The *LTBR-LTA* axis in the aetiology of multiple sclerosis. a)** Unconditioned and **b)** conditioned regional association plots at the *TNFRSF1A*-*LTBR* locus (rs2364485 +/- 100kb) for MS (top), plasma LTA protein levels (middle), and *LTBR* mRNA expression in whole blood from eQTLGen[14] (bottom). MS associations were conditioned on rs1800693 (the strongest disease signal in the region). *LTBR* mRNA expression levels were conditioned on the following independent eQTLs: rs3759322, rs1800692, rs2228576, rs10849448, rs2364480, and rs12319859. Purple diamond = sentinel pQTL variant. Other variants coloured by LD to sentinel pQTL. Two-sided *P*-values are from meta-analysis of linear regression estimates.

Figure 5. MR analysis of circulating proteins in immune-mediated disease aetiology. GSMR analysis[28] using *cis*-pQTLs as genetic instruments to test the causal role of plasma proteins across IMDs. Cells are coloured according to effect size and direction: red = higher genetically-predicted plasma protein levels are associated with higher risk of disease; blue = higher genetically-predicted plasma protein levels are associated with lower risk of disease; grey = fewer than 3 variants available to run the GSMR analysis. Associations with FDR≤0.01 are denoted with dots, with filled circles indicating those that are robust to confounding by LD and open circles indicating those that were not.

Figure 6. **CXCL5 in UC pathogenesis**. **a)** Genetic associations in the *CXCL5* gene region. From top to bottom: plasma CXCL5 pQTL (n=14,824 participants), whole blood eQTL (from eQTLGen data, n=31,684 participants), colon tissue eQTL (GTEx, n=368 individuals), ulcerative colitis (cases=12,366, controls=33,609) and Crohn's disease (cases=12,194, controls=28,072)(from the IBD Genetics Consortium[51]). Purple diamond = sentinel pQTL variant. Other variants coloured by LD to sentinel pQTL. *P*-values from linear regression. **b)** Violin plots showing *CXCL5* expression in gut mucosal samples from patients with UC or CD and healthy controls (HC) in IBD TaMMA. **c)** Volcano plot showing differential expression analysis comparing colonic tissue from UC versus healthy controls (IBD TaMMA). Red and blue points represent significantly (5% FDR) up- and down-regulated transcripts, respectively. Grey = non-significant. P_{BH} = Benjamini-Hochberg adjusted *P*-values. *P*-values in c-d from Wald test (two-sided). **d)** Replication. Left hand panel: CXCL5 differential expression in colon biopsies in UC versus HC from transcriptome-wide analysis across 3 cohorts. GSE numbers are Gene Expression Omnibus accession numbers. Imperial = Imperial UC cohort. Each lollipop represents a separate cohort. GSE16879 n=24 UC patients versus n=6 HC. GSE206285 n=550 UC patients versus n=18 HC. Imperial n=16 UC vs 6 HC. Circle colour indicates the $\log_2$ fold change (FC) in *CXCL5* expression between UC and HC. Right hand panel: CXCL5 expression in colon biopsies sampled at baseline during the UNIFI clinical trial. Each point represents an individual. **e)** Forest plot showing Mendelian randomisation analysis for UC and CD. OR = odds ratio for the risk associated with a 1 standard deviation increase in the level of the protein. Centre of bar = point estimate for OR, whiskers = 95% confidence intervals (CI).

REFERENCES

1. Sun, B.B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018).

2. Enroth, S., Johansson, A., Enroth, S.B. & Gyllensten, U. Strong effects of genetic and lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat Commun* **5**, 4684 (2014).

3. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun* **8**, 14357 (2017).

4. Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769-773 (2018).

5. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* **4**, e1000072 (2008).

6. Lourdusamy, A. *et al.* Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum Mol Genet* **21**, 3719-3726 (2012).

7. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* **2**, 1135-1148 (2020).

8. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).

9. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics* **53**, 1712-1721 (2021).

10. Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nature Genetics* **54**, 593-602 (2022).

11. Gudjonsson, A. *et al.* A genome-wide association study of serum proteins reveals shared loci with common diseases. *Nature Communications* **13**, 480 (2022).

12. Siegbahn, A. *et al.* Multiplex protein screening of biomarkers associated with major bleeding in patients with atrial fibrillation treated with oral anticoagulation. *J Thromb Haemost* **19**, 2726-2737 (2021).

13. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nature Communications* **12**, 6822 (2021).

14. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300-1310 (2021).

15. The GTEx Consortium *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

16. Peters, J.E. *et al.* Insight into Genotype-Phenotype Associations through eQTL

Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS Genet* **12**, e1005908 (2016).

17. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* **53**, 1290-1299 (2021).

18. Stacey, D. *et al.* ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. *Nucleic Acids Res* **47**, e3 (2019).

19. Rappoport, N., Simon, A.J., Amariglio, N. & Rechavi, G. The Duffy antigen receptor for chemokines, ACKR1,- 'Jeanne DARC' of benign neutropenia. *Br J Haematol* **184**, 497-507 (2019).

20. 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.e1114 (2020).

21. Hassan, H.T. & Zander, A. Stem cell factor as a survival and growth factor in human normal and malignant hematopoiesis. *Acta Haematol* **95**, 257-262 (1996).

22. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179-189 (2020).

23. Ferreira, R.C. *et al.* Functional IL6R 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS Genet* **9**, e1003444 (2013).

24. Rosa, M. *et al.* A Mendelian randomization study of IL6 signaling in cardiovascular diseases, immune-related disorders and longevity. *NPJ Genom Med* **4**, 23 (2019).

25. Patsopoulos, A. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).

26. Gregory, A.P. *et al.* TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. *Nature* **488**, 508-511 (2012).

27. Smith, G.D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1-22 (2003).

28. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9**, 224 (2018).

29. Massimino, L. *et al.* The Inflammatory Bowel Disease Transcriptome and Metatranscriptome Meta-Analysis (IBD TaMMA) framework. *Nature Computational Science* **1**, 511-515 (2021).

30. Croft, M. & Siegel, R.M. Beyond TNF: TNF superfamily cytokines as targets for the

treatment of rheumatic diseases. *Nat Rev Rheumatol* **13**, 217-233 (2017).

31. Yasuda, H. Discovery of the RANKL/RANK/OPG system. *J Bone Miner Metab* **39**, 2-11 (2021).

32. Walsh, M.C. & Choi, Y. Biology of the RANKL-RANK-OPG System in Immunity, Bone, and Beyond. *Front Immunol* **5**, 511 (2014).

33. Jakubowski, A. *et al.* Dual role for TWEAK in angiogenic regulation. *J Cell Sci* **115**, 267-274 (2002).

34. Donohue, P.J. *et al.* TWEAK is an endothelial cell growth and chemotactic factor that also potentiates FGF-2 and VEGF-A mitogenic activity. *Arterioscler Thromb Vasc Biol* **23**, 594-600 (2003).

35. Domouzoglou, E.M. *et al.* Fibroblast growth factors in cardiovascular disease: The emerging role of FGF21. *Am J Physiol Heart Circ Physiol* **309**, H1029-1038 (2015).

36. Schett, G., McInnes, I.B. & Neurath, M.F. Reframing Immune-Mediated Inflammatory Diseases through Signature Cytokine Hubs. *N Engl J Med* **385**, 628-639 (2021).

37. Peters, A.L., Stunz, L.L. & Bishop, G.A. CD40 and autoimmunity: the dark side of a great activator. *Semin Immunol* **21**, 293-300 (2009).

38. Durie, F.H. *et al.* Prevention of collagen-induced arthritis with an antibody to gp39, the ligand for CD40. *Science* **261**, 1328-1330 (1993).

39. Guo, Y. *et al.* CD40L-Dependent Pathway Is Active at Various Stages of Rheumatoid Arthritis Disease Progression. *J Immunol* **198**, 4490-4501 (2017).

40. The Lenercept Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group. TNF neutralization in MS: results of a randomized, placebo-controlled multicenter study. The Lenercept Multiple Sclerosis Study Group and The University of British Columbia MS/MRI Analysis Group. *Neurology* **53**, 457-465 (1999).

41. Bosch, X., Saiz, A., Ramos-Casals, M. & Group, B.S. Monoclonal antibody therapy-associated neurological disorders. *Nat Rev Neurol* **7**, 165-172 (2011).

42. Singh, U.P. *et al.* Chemokine and cytokine levels in inflammatory bowel disease patients. *Cytokine* **77**, 44-49 (2016).

43. Friedrich, M. *et al.* IL-1-driven stromal-neutrophil interactions define a subset of patients with inflammatory bowel disease that does not respond to therapies. *Nat Med* **27**, 1970-1981 (2021).

44. Pavlidis, P. *et al.* Interleukin-22 regulates neutrophil recruitment in ulcerative colitis

and is associated with resistance to ustekinumab therapy. *Nat Commun* **13**, 5820 (2022).

45. Richard, A.C. *et al.* Reduced monocyte and macrophage TNFSF15/TL1A expression is associated with susceptibility to inflammatory bowel disease. *PLoS Genet* **14**, e1007458 (2018).

46. Bamias, G. *et al.* Differential expression of the TL1A/DcR3 system of TNF/TNFR-like proteins in large vs. small intestinal Crohn's disease. *Dig Liver Dis* **44**, 30-36 (2012).

47. Bamias, G. *et al.* High intestinal and systemic levels of decoy receptor 3 (DcR3) and its ligand TL1A in active ulcerative colitis. *Clin Immunol* **137**, 242-249 (2010).

48. Sands, B. *et al.* OP40 PRA023 Demonstrated Efficacy and Favorable Safety as Induction Therapy for Moderately to Severely Active UC: Phase 2 ARTEMIS-UC Study Results 2023 [cited 2023 2/7]Available from: https://www.ecco-ibd.eu/publications/congress-abstracts/item/op40-pra023-demonstrated-efficacy-and-favorable-safety-as-induction-therapy-for-moderately-to-severely-active-uc-phase-2-artemis-uc-study-results.html

49. Fairfax, B.P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).

50. Lee, M.N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343**, 1246980 (2014).

51. 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).

METHODS

Cohorts. We recruited 11 cohorts, totalling 14,824 participants, with genome-wide genetic data and plasma proteomic data measured using the Olink Target Inflammation panel. All participants provided written, informed consent. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to or larger than those reported in previous publications [1-4,7-9]. Cohort details are provided in the **Supplementary Note 1**.

Protein assays. Plasma proteins were measured using the Olink Target-96 Inflammation multiplexed immunoassay panel, which measures 92 inflammation-related proteins. Proteomic data for each cohort were generated at Olink laboratories in Uppsala. During the course of the project, BDNF was removed from the inflammation panel by Olink due to assay problems therefore 91 proteins were included in our study (**Supplementary Table 2**). Normalised Protein eXpression (NPX), is Olink's normalised relative units in $\log_2$ scale. Olink

defines the lower limit of detection (LLOD) for quantification of each protein as 3 standard deviations above background (determined using blank control samples), but provides NPX as continuous data which can include values below the calculated LLOD. We had access to individual-level data for INTERVAL, the largest contributing cohort (n= 4,896), and used this to calculate the proportion of samples <LLOD for each protein (**Extended Data Figure 2a**).

Genotyping. Each cohort was genotyped on a SNP (single nucleotide polymorphism) array and imputed using either a 1000Genomes or Haplotype Reference Consortium (HRC) panel (**Supplementary Table 1**).

Cohort-level pQTL mapping. In each cohort, a GWAS analysis was run for each protein using linear regression (additive genetic association model) with protein level as the dependent variable. Proteins were inverse-rank normalised prior to linear regression, and thus met the assumptions of the statistical test. Population substructure was adjusted for by including genetic principal components as covariates. We also included age, sex and other study-specific covariates in the model (see **Supplementary Table 1**). To avoid proteins with truncated distributions due to LLOD with multiple tied values that would violate linear regression assumptions, pQTL analysis was performed using continuous protein values (including those below the LLOD where relevant). We illustrate the value of this approach in recovering biological signals in **Extended Data Figure 2b**.

pQTL meta-analysis. We meta-analysed pQTL summary statistics from each cohort (**Supplementary Table 1**), representing a total of 14,824 participants. A schematic of our analysis pipeline is shown in **Extended Data Figure 1**. Prior to the meta-analysis, we applied cohort-level filters to pQTL GWAS summary statistics with respect to MAF (≥0.001), HWE ($P>10^{-6}$), and imputation score (r^2≥0.3 or SNPTEST proper_info≥0.4). For each cohort, we generated QQ plots and Manhattan plots for visual examination using the R packages qqman v0.1.4 and QCGWAS v1.0-8. We performed the fixed-effects meta-analysis with the METAL software (version 28.8.2018), using inverse-variance weighted analysis of regression betas and standard errors from the cohort-level summary statistics. From the meta-analysis summary statistics, we calculated the genomic inflation factor for each protein GWAS, and generated QQ and Manhattan plots (**Supplementary Figure 1**). We generated Forest plots to examine inter-cohort heterogeneity using the gap package v1.2.3-6. Regional association plots were generated using LocusZoom 1.4 (**Supplementary Figure 2**). We defined statistical significance as $P\leq5\times10^{-10}$ (based on Bonferroni correction of the conventional 'genome-wide' significance threshold $P\leq5\text{x}10^{-8}$ for approximately 100 proteins).

To remove potentially erroneous meta-analysis signals arising due to a strong association in

a single cohort, we examined the meta-analysis results at each sentinel variant by visual inspection of the forest plot and imposed the following criteria: 1) to be included in the meta-analysis, a variant was required to be available in at least three studies and in at least 3,500 participants; 2) in order to be declared significant, we required a meta-analysis $P \leq 5 \times 10^{-10}$, and, if there was evidence of heterogeneity with $I^2 > 30\%$, then we required the *P*-value in at least three studies to be <0.05 and the direction of effect in those studies to be consistent with the overall meta-analysis results. These were implemented through modification of METAL source code.

Replication cohort. We compared the results from our primary meta-analysis to pQTL results generated in an independent set of 1,585 participants from the ARISTOTLE study[12,52].

Definition of pQTL sentinel variants and regions. We defined a pQTL as a genetic locus significantly ($P \leq 5 \times 10^{-10}$) associated with protein abundance. We defined the sentinel variant at a locus as the variant with the lowest *P*-value in the region for a given protein. We used the following approach for each protein to define genomic regions and the sentinel variant in each: 1) we first obtained a list of significant ($P \leq 5 \times 10^{-10}$) variants and the flanking region (+/-1Mb) for each variant; 2) overlapping regions were then iteratively merged until no overlapping regions remained; 3) the most significant variant in each resulting region was then defined as the sentinel variant. This approach has the flexibility to cope with long stretches of LD whilst avoiding the drawback of setting a longer than necessary region for all variants. The algorithm was implemented using bedtools v2.27.0. Signals within 1Mb of the transcription start site (TSS) of the gene encoding the target protein were defined as *cis* and those beyond 1 Mb as *trans*.

Protein variance explained by pQTLs. We used the following equation to estimate the proportion of variance explained (PVE) by (*T*) pQTLs from the meta-analysis summary statistics for each protein:

$$\text{PVE} = \sum_{i=1}^{T} \frac{{\chi_i}^2}{{\chi_i}^2 + N_i - 2} \quad (1)$$

where ${\chi_i}^2$ is the chi-squared statistic based on each associated variant's effect size and its standard error and N_i is the associated sample size.

Conditional analysis. To identify conditionally independent signals within a genomic region, we performed approximate stepwise conditional analyses using GCTA v1.93.0beta with the '--cojo-slct' option, using effect sizes and standard errors from the meta-analysis. We estimated the correlation between variants using individual-level data from the INTERVAL study. As GCTA imputes LD from mean genotypes when they are missing, to avoid bias we excluded

variants with MAF<1% (unless they were sentinel variants). For stepwise selection, we considered only those variants passing the genome-wide threshold ($P \leq 5 \times 10^{-10}$), rather than all variants in the region. As in certain cases GCTA conditional analysis yielded results involving pairs of variants in relatively high LD ($r^2 \geq 0.7$), we restricted to independent genetic variants ($r^2 \leq 0.1$)[53] in the INTERVAL imputed genotype data whilst forcing the inclusion of the sentinel variants in the pruned set[54] (**Supplementary Table 4**).

Identification of known pQTLs. To identify previously reported pQTLs, we manually curated published results from literature obtained from the NCBI web interface (https://pubmed.ncbi.nlm.nih.gov/) through its Entrez programming utility R/rentrez[55], PhenoScanner v2[56], and the NHGRI-EBI GWAS Catalog with phenotypes mapped to the Experimental Factor Ontology (EFO) EFO_0004747 (protein measurement), restricting to associations reported in European-ancestry populations. We selected variants that reached the conventional genome-wide significance threshold $P \leq 5 \times 10^{-8}$ and that were in high LD ($r^2 \geq 0.8$) with the meta-analysis pQTL sentinel variant.

Variant annotation. We obtained the absolute distance of sentinel variants to the TSS of the gene encoding the target protein using the rGREAT (Genomic Regions Enrichment of Annotations Tool)[57] R package. We annotated sentinel variants and LD proxies (defined as $r^2 \geq 0.8$, using the INTERVAL dataset as the LD reference panel) using Ensembl's Variant Effect Predictor (VEP, v98.3) including the LOFTEE plugin.

eQTL-pQTL colocalisation analysis. We performed pairwise statistical colocalisation analyses of *cis*-pQTLs identified in the meta-analysis with cognate *cis*-eQTL data from eQTLGen[14,58], the eQTL Catalogue[17], and GTEx v8[15]. We extracted the meta-analysis summary statistics for each *cis*-pQTL sentinel and their +/-1Mb flanking regions, then extracted the same genomic windows from their cognate *cis*-eQTL data. eQTLGen comprises eQTL data from 31,684 participants on 19,250 genes that are robustly expressed in blood (https://www.eqtlgen.org/cis-eqtls.html). Of our 59 *cis*-pQTLs, there was genome-wide significant ($P \leq 5 \times 10^{-8}$) *cis*-eQTL for 40 genes in the eQTLGen data. 1 gene (*TGFB1*) had a *cis*-eQTL at FDR<0.05 but that was not genome-wide significant (p 1.8×10^{-7}), and 2 had no eQTL association (*IL17C*, *TNFSF11*). 16 genes had no eQTL data in the eQTLGen summary statistics, presumably due to lack of robust expression in blood. These were: *CCL11*, *CCL13*, *CCL19*, *CCL20*, *CCL7*, *CST5*, *CX3CL1*, *CXCL11*, *DNER*, *FGF21*, *FGF5*, *GDNF*, *IL12B*, *MMP10*, *NGF*, *TNFRSF11B*.

For GTEX v8 and eQTL Catalogue, all 59 *cis*-pQTLs had corresponding eQTL summary statistics available for colocalisation testing across one or more tissues. We performed colocalisation analyses using the coloc R package as implemented in v.5.2.2 of the eQTL

Catalogue/colocalisation workflow[17] (https://github.com/kauralasoo/eQTL-Catalogue-resources). Coloc returns posterior probabilities indicating the likelihood that the following scenarios are true: (i) there is no association at the locus with either protein or mRNA (PP0), (ii) there is an association with either the protein (PP1) or the mRNA (PP2), but not both, (iii) there is an association with both the protein and the mRNA but with distinct causal variants (PP3) or with a shared causal variant (PP4). We considered a PP4≥0.8 to be robust evidence of colocalisation between a *cis*-pQTL and its cognate *cis*-eQTL. Since eQTLGen data only provides allele frequency (*f*) and z-score statistic (z) for a particular variant, we obtained the effect size (*b*) and its standard error (*se*) as follows[59]:

$$b = z/d \quad (2)$$

$$se = 1/d \quad (3)$$

where

$$d = \sqrt{2f(1-f)(z^2+N)} \quad (4)$$

and N is the sample size.

Prioritizing likely mediating genes at *trans*-pQTLs. To prioritize likely mediating genes at *trans*-pQTLs, we used the ProGeM tool[18]. To identify *cis*-eQTLs that could mediate our *trans*-pQTLs, we queried the *trans*-pQTL sentinel variants in eQTLGen[14], the eQTL Catalogue[17], and the Genotype-Tissue Expression project (GTEx, v8) data. To determine whether the *trans*-pQTL sentinels are likely to be causal *cis*-eQTL variants in the eQTL Catalogue and GTEx data, we used the fine-mapped eQTL credible sets available at the eQTL Catalogue (https://www.ebi.ac.uk/eqtl/Data_access/). For the eQTLGen data, where credible sets were not available, we used a manual approach whereby we: (i) first defined a region around each *trans*-pQTL sentinel variant of +/-500kb; (ii) identified the variant with the lowest *cis*-eQTL *P*-value in this region for the *cis*-affected gene(s); and (iii) checked to see whether this sentinel *cis*-eQTL variant is the same sentinel variant for the *trans*-pQTL, or if the two are in high LD (r^2≥0.8).

For the "top-down" component of ProGeM, we first identified locally-encoded genes using a window around each *trans*-pQTL sentinel variant of +/-500kb. We then probed the proteins encoded by these local genes using: (1) protein:protein interaction (PPI) data; and (2) data from functional annotation databases. With the PPI data we aimed to determine whether there is evidence to indicate that genes residing close to each sentinel variant might physically interact with the corresponding *trans*-affected protein. We used the Bioconductor package STRINGdb (v2.8.4) to identify any pairwise interactions. We used data from functional annotation databases to determine whether any local genes encode proteins that might be functionally related to the corresponding *trans*-affected protein(s). For both the *trans*-affected proteins and locally encoded proteins, all assigned Gene Ontology terms, Reactome pathways, and KEGG pathways were extracted using the Bioconductor biomaRt (v2.52) and

KEGGREST (v1.36) packages. To assess whether there is significant overlap between the functional annotation terms/pathways assigned to locally encoded proteins and the corresponding *trans*-affected proteins, we determined the number of shared and non-shared terms for each local gene and the corresponding *trans*-affected protein. Fisher's exact test was then applied for each local gene/*trans*-protein combination, and *P*-values were Bonferroni-corrected for the number of local genes at each given *trans*-pQTL. The background set of terms for each *trans*-pQTL was composed of all terms assigned to all local genes at the locus (i.e., all protein-coding genes within 500kb from the sentinel variant).

To determine the most likely mediating genes for the multi-locus regulated proteins IL12B, KITLG, and TNFSF10 (TRAIL), we used the STRINGdb webtool to identify interactions or functional relationships between genes residing at distinct loci. This is based on the assumption that if the mediating genes at distinct loci are all associated with plasma levels of the same protein, then they may share some other functional relationship. As input to STRINGdb, we used all proteins encoded by candidate mediating genes identified by ProGeM (**Supplementary Table 9**) at each of the loci for a given protein, as well as the relevant *trans*-affected protein. We deemed clusters of proteins residing at distinct loci with multiple functional interactions to be the most likely mediating genes at their respective loci. We performed a phenome-scan of the *trans*-pQTLs for *KITLG* using the Open Targets Genetics webtool[60].

Overlap of pQTL and disease traits. We used a PhenoScanner v2-based R code to look up associations of our pQTL sentinels and their LD proxies (r^2≥0.8) in disease GWAS summary statistics.

To investigate potential colocalisation between a *trans*-pQTL (rs2364485) for LTA identified in our meta-analysis, a multiple sclerosis GWAS signal[25], and a *cis*-eQTL for *LTBR* from eQTLGen[14], we used HyPrColoc for multi-trait colocalisation[61]. We obtained multiple sclerosis summary statistics (MSchip, "discovery_metav3.0.meta.gz") from Patsopoulos *et al* by request to the International Multiple Sclerosis Genetics Consortium (IMSGC). Due to a lack of genotype coverage at the *LTBR*/*TNFRSF1A* locus in the extended and replication samples from Patsopoulos *et al*, we selected the summary statistics from the "discovery" sample (*n*=41,505) for colocalisation analyses, not the full meta-analysis. As a result, the *P*-value for association between the variant of interest (rs2364458) and multiple sclerosis in the discovery subset (P=5.78x10^{-6}, logistic regression) was higher than reported in Patsopoulos *et al*[25] (P=2.0x10^{-20}, fixed-effects meta-analysis). We then extracted summary statistics for rs2364458 (+/-1Mb) (chr12:5514963-7514963) from each of the 3 datasets, and performed conditional analyses to adjust for any independent signals at the locus using GCTA-COJO. We ran this using a two-step approach: we first used the COJO-slct function to identify

independent signals at the locus, and then for datasets with independent signals (i.e., in addition to rs2364485), we used COJO-cond to generate conditioned summary statistics for use in HyPrColoc. HyPrColoc returns the posterior probability that 2 or more traits colocalise, akin to PP4 from coloc. We considered a PP≥0.8 as robust evidence of colocalisation between traits.

Mendelian randomisation analyses. We performed MR analyses using the proteins with *cis*-pQTLs identified in this meta-analysis as exposures, and immune-mediated diseases (IMDs) as outcomes. All MR analyses were run using the Generalized Summary-data-based Mendelian Randomisation (GSMR) method[28], which implements two-sample MR accounting for correlation between variants. For each protein analysed, we defined a +/- 1 Mb window around the gene encoding it and extracted pQTL summary statistics for this region. For outcome data, we downloaded GWAS summary statistics for IMDs from OpenGWAS (https://gwas.mrcieu.ac.uk/datasets/) or from the GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads) where studies with larger sample sizes or more variants were available. For IMDs with several alternative datasets available, we selected the dataset with the largest number of cases, provided it (i) had genotype data with sufficient coverage at the loci of interest, (ii) was generated in European-ancestry samples so that it matched the ancestry of the participants in our pQTL meta-analysis, and (ii) had effect estimates and standard errors either available or calculable. Proteins encoded by genes in the *HLA* region were excluded because MR analysis would be confounded by complex LD. The analysis involved 57 proteins and 14 diseases. We used the GSMR implementation in GCTA with the following parameters: 1) at least 3 (--gsmr-snp-min 3) genome-wide significant (--gwas-thresh 5e-8), quasi-independent variants (--clump-r2 0.1); 2) difference in the allele frequency of each effect allele between the GWAS summary datasets and the LD reference sample is at most 0.4 (--diff-freq 0.4); 3) a *P*-value threshold of 0.05 for the HEIDI-outlier filtering analysis (--heidi-thresh 0.05), which is used to identify potential confounding by LD (https://yanglab.westlake.edu.cn/software/gcta/#Mendelianrandomisation). *P*-values were corrected for the number of models tested using Benjamini-Hochberg correction, with FDR<0.01 used to define statistical significance.

To evaluate the robustness of significant associations, we performed additional checks. First, we checked the strength of the disease association in the GWAS summary statistics. Of the 22 significant protein-disease MR associations, we eliminated 5 due to the lack of convincing disease association (smallest *P*-value at the locus $P>1x10^{-4}$). For the remaining 17 MR associations, we then evaluated whether there might be confounding due to LD. We first evaluated the r^2 between the sentinel pQTL and the disease-associated variant. For 12 of 17 disease-protein pairs, r^2 was >0.8 (**Supplementary Table 14**). We next performed visual

inspection of regional association plots of these 12 pQTL–disease pairs (**Supplementary Figure 4**) and colocalisation testing using PWCoCo[62,63] which accounts for the presence of multiple independent signals within a locus (see below).

Pairwise conditional colocalisation analyses. PairWise COnditional and COlocalisation analysis (PWCoCo)[62,63] integrates conditional analyses (GCTA-COJO) to identify independent signals for each of two tested traits associated with a genomic region, followed by pairwise colocalisation analyses (COLOC) to test all possible pairs of conditionally independent signals across the traits. We ran PWCoCo for the 12 significant protein-disease pairs that resulted from our MR filtering steps using the default parameters, detailed as follows: 1) *P*-value cutoff for variants to be selected by the stepwise selection process, --p_cutoff 5e-8 for disease and protein summary statistics; 2) a large number of variants subject to the stepwise selection process, --top_snp 1e10; 3) distance beyond which variants are treated as in linkage equilibrium, --ld_window 1e7 (kb); 4) collinearity threshold for variants, --collinear 0.9; 5) variant frequency filter for the reference dataset according to this threshold, --maf 0.1; 6) exclusion threshold for variants with allele frequency difference between the phenotype and the reference datasets, --freq_threshold 0.2; 7) stop criteria, --init_h4 80 (i.e. 80%); 8) the three prior probabilities, --coloc_pp 1e-4, 1e-4 and 1e-5.

***CXCL5* differential expression analysis in UC cohorts.** Changes in *CXCL5* gene expression levels were evaluated in four independent cohorts, including the IBD Transcriptome and Metatranscriptome Meta-Analysis (TaMMA) platform[29], the Gene Expression Omnibus (GEO) series GSE16879, GSE206285, as well as the Imperial UC cohort. IBD TaMMA (https://ibd-meta-analysis.herokuapp.com/) gives access to 3,853 transcriptomic profiles from 26 independent studies including IBD and control samples across different tissues, all processed with the same pipeline and batch-corrected[29]. Pre-computed differential expression results between colon biopsies from UC patients versus healthy donors were downloaded and plotted.

Data from GEO Series GSE16879 used in this study consist of colonic mucosa microarray expression profiles from healthy donors (n=6) and UC patients (n=24) sampled before first infliximab treatment[64]. CEL file import into R, and background correction, normalization and RMA calibration of the raw intensity data were carried out using the oligo package. Only probe sets with median expression greater than 4 and associated to only one ENTREZ gene identifier were kept for analysis. Intensity data for different probe sets mapped to the same ENTREZ gene identifier were combined by taking the geometric mean sample wise. Tests of differential gene expression of ulcerative colitis samples compared to healthy control samples

were performed with the limma package. *P*-values were adjusted for multiple testing with the Benjamini and Hochberg procedure.

GEO Series GSE206285 contains array-based transcriptomic data collected at baseline as part of UNIFI, a randomised placebo-controlled phase 3 clinical trial evaluating the efficacy and safety of ustekinumab[65]. RMA signal intensity profiles and associated donor information were downloaded from NCBI GEO. Only probe sets associated to only one ENTREZ gene identifier were kept for analysis. Intensity data for different probe sets mapped to the same ENTREZ gene identifier were combined by taking the geometric mean sample wise. Genes with median expression greater than 3 across all samples were tested for differential expression between ulcerative colitis samples (n=550) versus healthy control samples (n=18) using the R limma package. *P*-values were adjusted for multiple testing with the Benjamini and Hochberg procedure.

The Imperial UC cohort includes whole tissue biopsies from ulcerative colitis patients (n=16) and healthy volunteers (n=6). RNA was extracted (Qiagen RNeasy mini kit) and sequencing libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) following manufacturer's recommendations. Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer（5X). First strand cDNA was synthesised using random hexamer primer and M-MuLV Reverse Transcriptase（RNase H-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. Library fragments were purified with AMPure XP system (Beckman Coulter, Beverly, USA) and treated with 3 μl USER Enzyme (NEB, USA) at 37°C for 15 min, followed by 5 min at 95 °C. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. Library quality was assessed on Agilent Bioanalyzer 2100 and on Nanodrop ND-1000 Spectrophotometer. The library preparations were sequenced on an Illumina HiSeq platform, generating 150 bp paired end reads. The resulting fastq files were processed with trimmomatic[66] (v. 0.39) to remove adaptor contamination and poor-quality bases. The output read files were mapped to the GRCh38 assembly of the human genome using Hisat2[67] (v. 2.2.1) with default parameters. The number of reads mapping to the genomic features annotated in Ensembl with a MAPQ score higher than or equal to 10 was calculated for all samples using htseq-count[68](v. 0.11.3) with default parameters. Data for Ensembl genes

with no associated ENTREZ gene identifier were discarded; the read counts for Ensembl genes mapped to the same ENTREZ gene identifier were summed up sample wise. Differential expression analysis between ulcerative colitis versus healthy biopsies was performed in R (v. 3.6.1) using the Wald test as implemented in DESeq2. Only genes with an average read count across samples greater than or equal to 10 were tested for differential expression. *P*-values were adjusted for multiple testing with the Benjamini-Hochberg procedure.

Data availability. Full per-protein GWAS summary statistics are available for download at https://www.phpc.cam.ac.uk/ceu/proteins/ and the EBI GWAS Catalog https://www.ebi.ac.uk/gwas/studies/GCST90270765 (accession numbers GCST90270765-GCST90270855). Individual-level genetic and proteomic data available for the INTERVAL cohort are deposited in the European-Genome Phenome Archive under the accession code https://ega-archive.org/studies/EGAS00001002555. Gene expression data are in Gene Expression Omnibus (GEO) under the accession code GSE16879 for mucosal expression in IBD patients (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16879), GSE206285 for the UNIFI trial (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE206285), and the IBD Transcriptome and Metatransciptome Meta-Analysis (TaMMA - https://ibd-meta-analysis.herokuapp.com/). Whole blood *cis*-eQTL summary statistics from the eQTLGen Consortium were downloaded from https://www.eqtlgen.org/cis-eqtls.html. Fine-mapped eQTL credible sets were downloaded from the eQTL Catalogue (https://www.ebi.ac.uk/eqtl/Data_access/). MR GWAS summary statistics for IMDs were downloaded from OpenGWAS (https://gwas.mrcieu.ac.uk/datasets/) or from the GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads).

Code availability. GitHub: https://jinghuazhao.github.io/INF/**,** cambridge-ceu: https://cambridge-ceu.github.io/public (modified METAL, pQTLtools)

Methods only REFERENCES

52. Hijazi, Z. *et al.* Screening of Multiple Biomarkers Associated With Ischemic Stroke in Atrial Fibrillation. *J Am Heart Assoc* **9**, e018984 (2020).

53. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature Genetics* **51**, 600-605 (2019).

54. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics* **44**, 369-375 (2012).

55. Winter, D.J. rentrez: an R package for the NCBI eUtils API. *The R Journal* **9**, 520-526 (2017).

56. Kamat, M.A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851-4853 (2019).

57. McLean, C.Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495-501 (2010).

58. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*, 447367 (2018).

59. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-487 (2016).

60. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug-target identification and prioritisation. *Nucleic Acids Res* **49**, D1302-D1310 (2021).

61. Foley, C.N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications* **12**, 764 (2021).

62. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics* **52**, 1122-1131 (2020).

63. Robinson, J., W. *et al.* An efficient and robust tool for colocalisation: Pair-wise Conditional and Colocalisation (PWCoCo). *bioRxiv*, 2022.2008.2008.503158 (2022).

64. Arijs, I. *et al.* Mucosal gene expression of antimicrobial peptides in inflammatory bowel disease before and after first infliximab treatment. *PLoS One* **4**, e7984 (2009).

65. Sands, B.E. *et al.* Ustekinumab as Induction and Maintenance Therapy for Ulcerative Colitis. *N Engl J Med* **381**, 1201-1214 (2019).

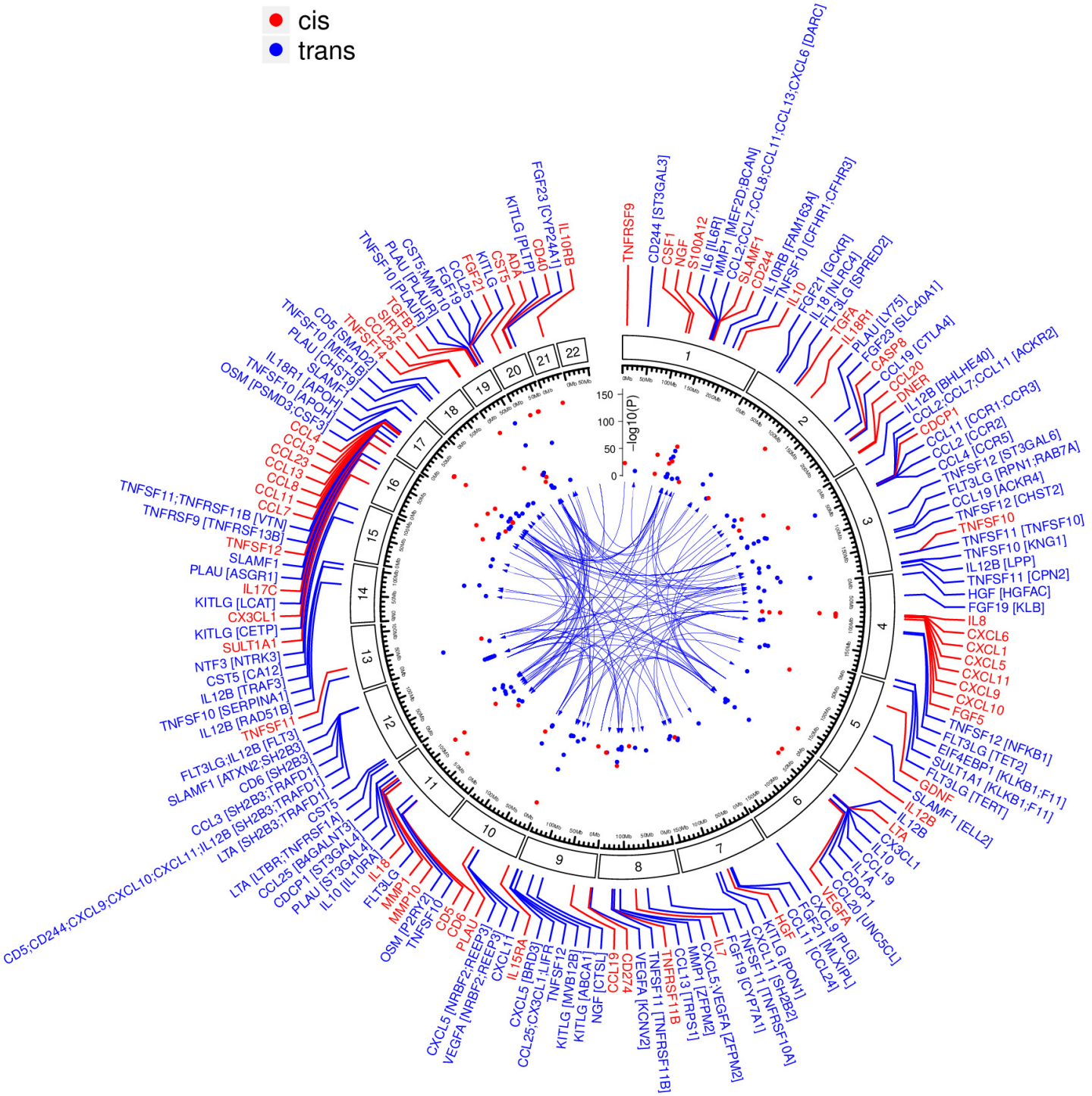
66. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

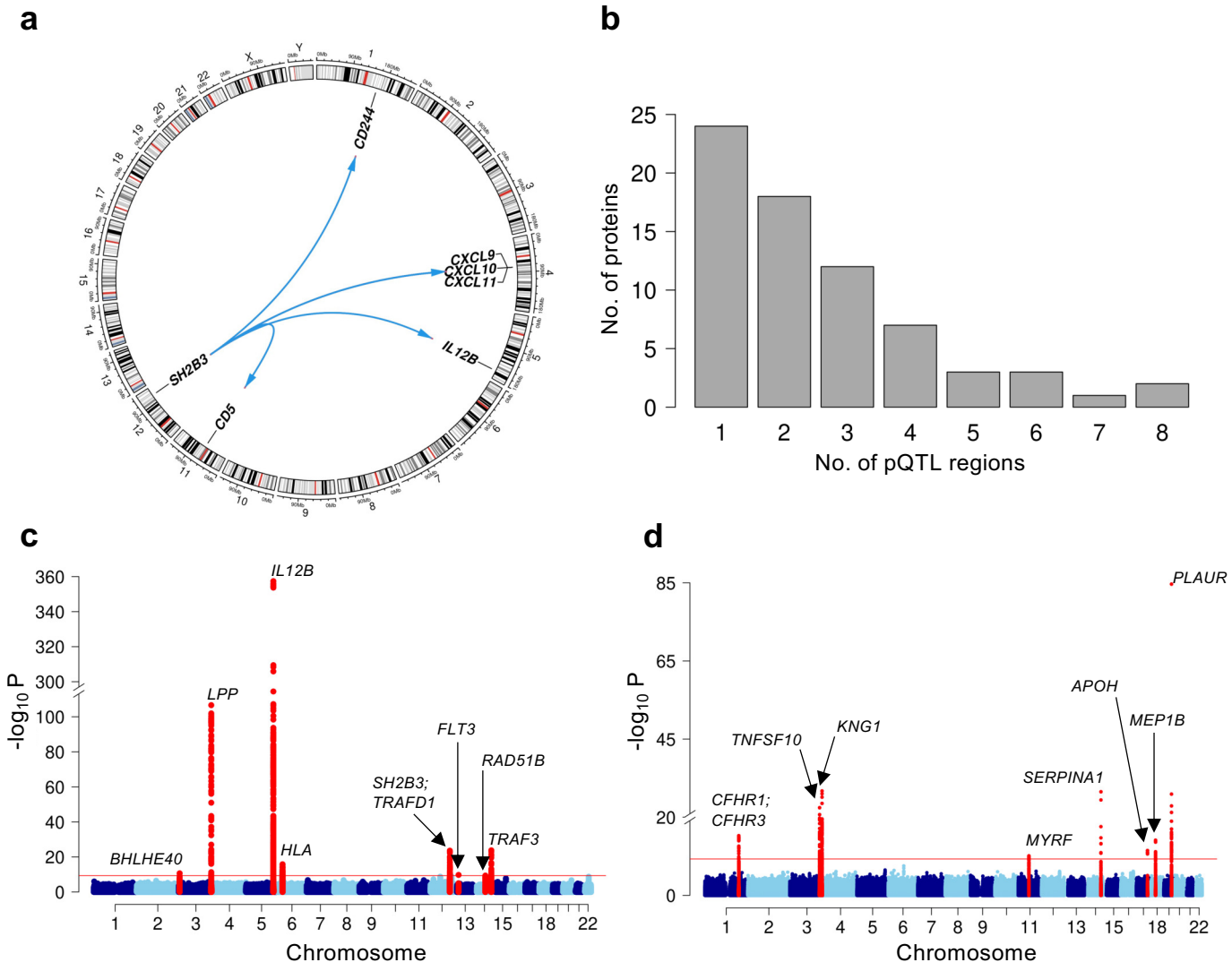
67. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).

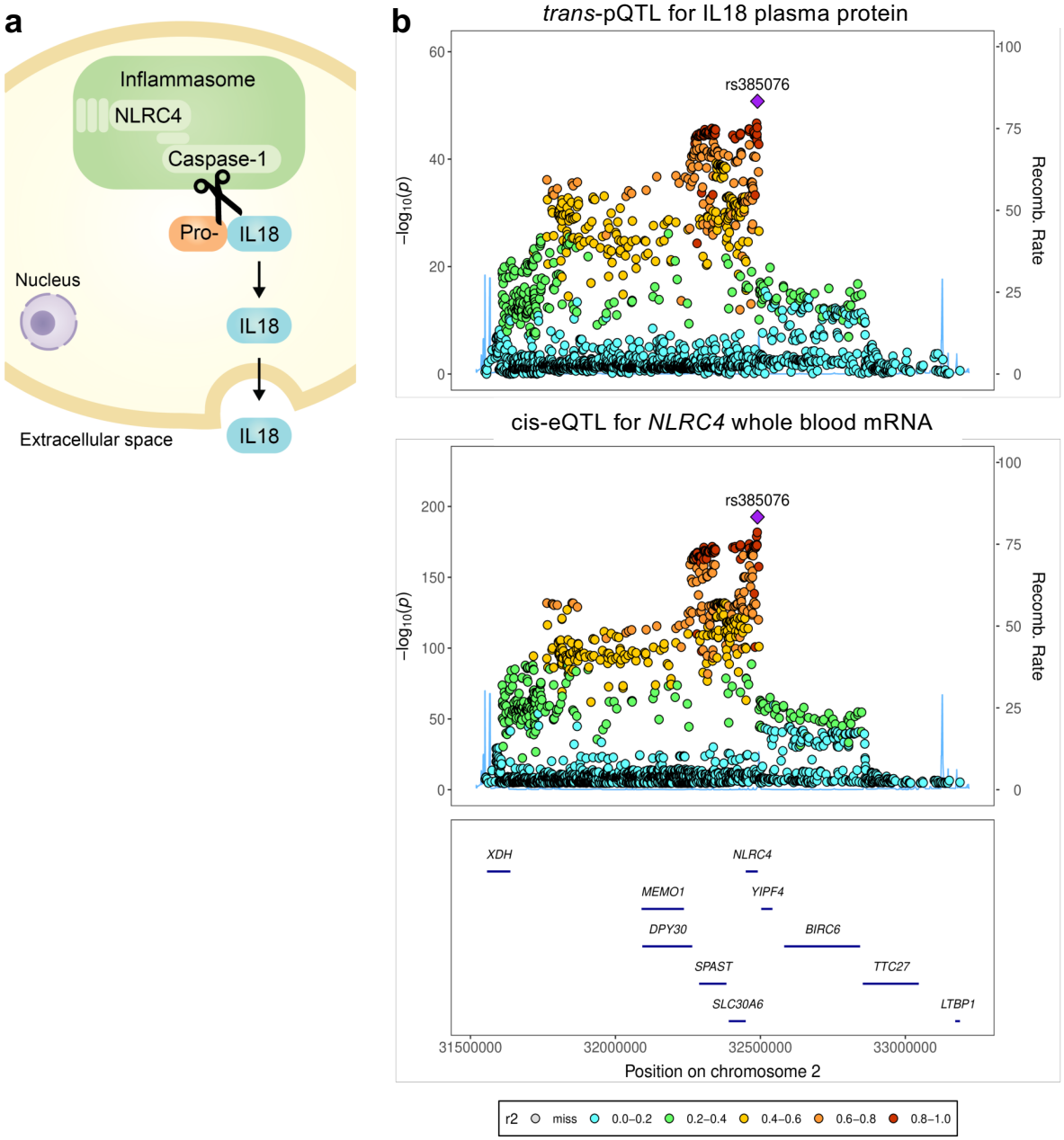
68. Anders, S., Pyl, P.T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

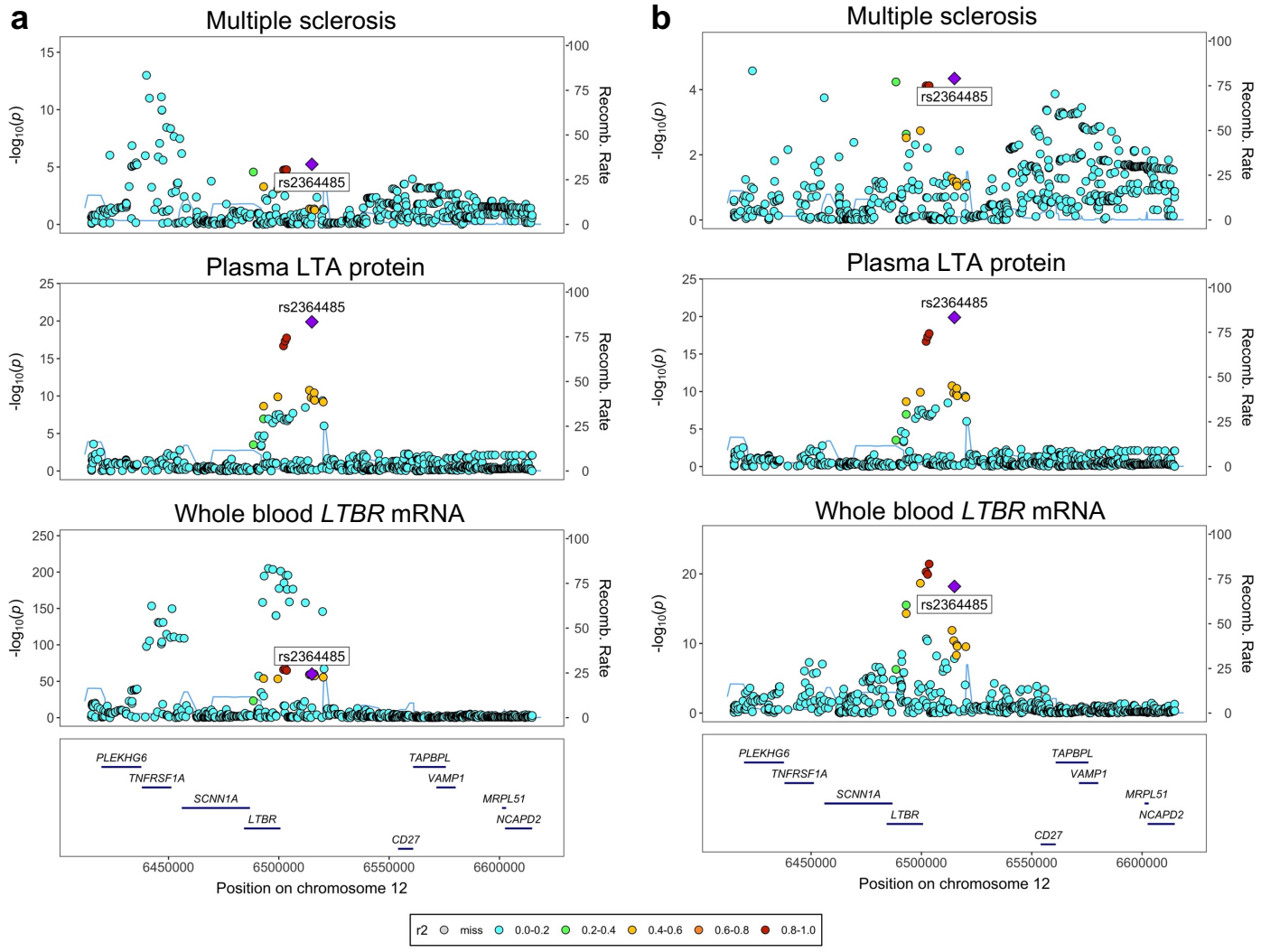
Estonian Biobank Research Team:

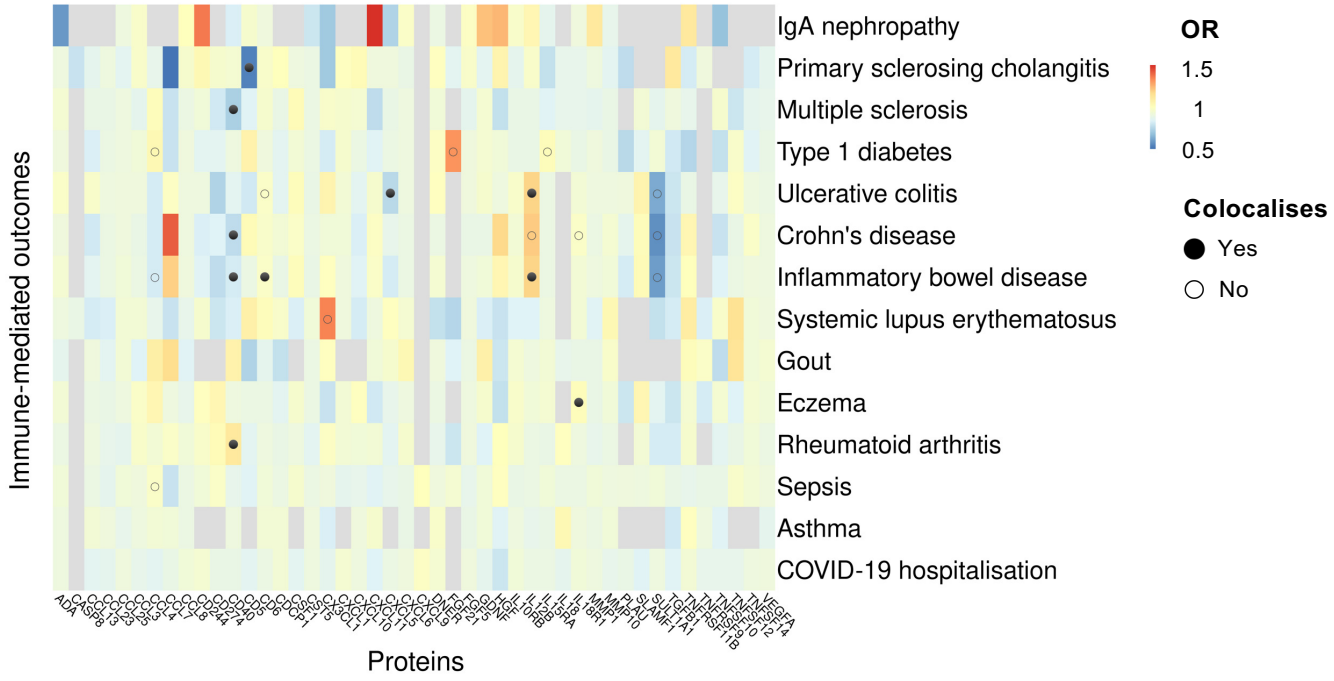
Andres Metspalu[9]
Lili Milani[9]
Reedik Mägi[9]
Mari Nelis[9]
Georgi Hudjašov[9]

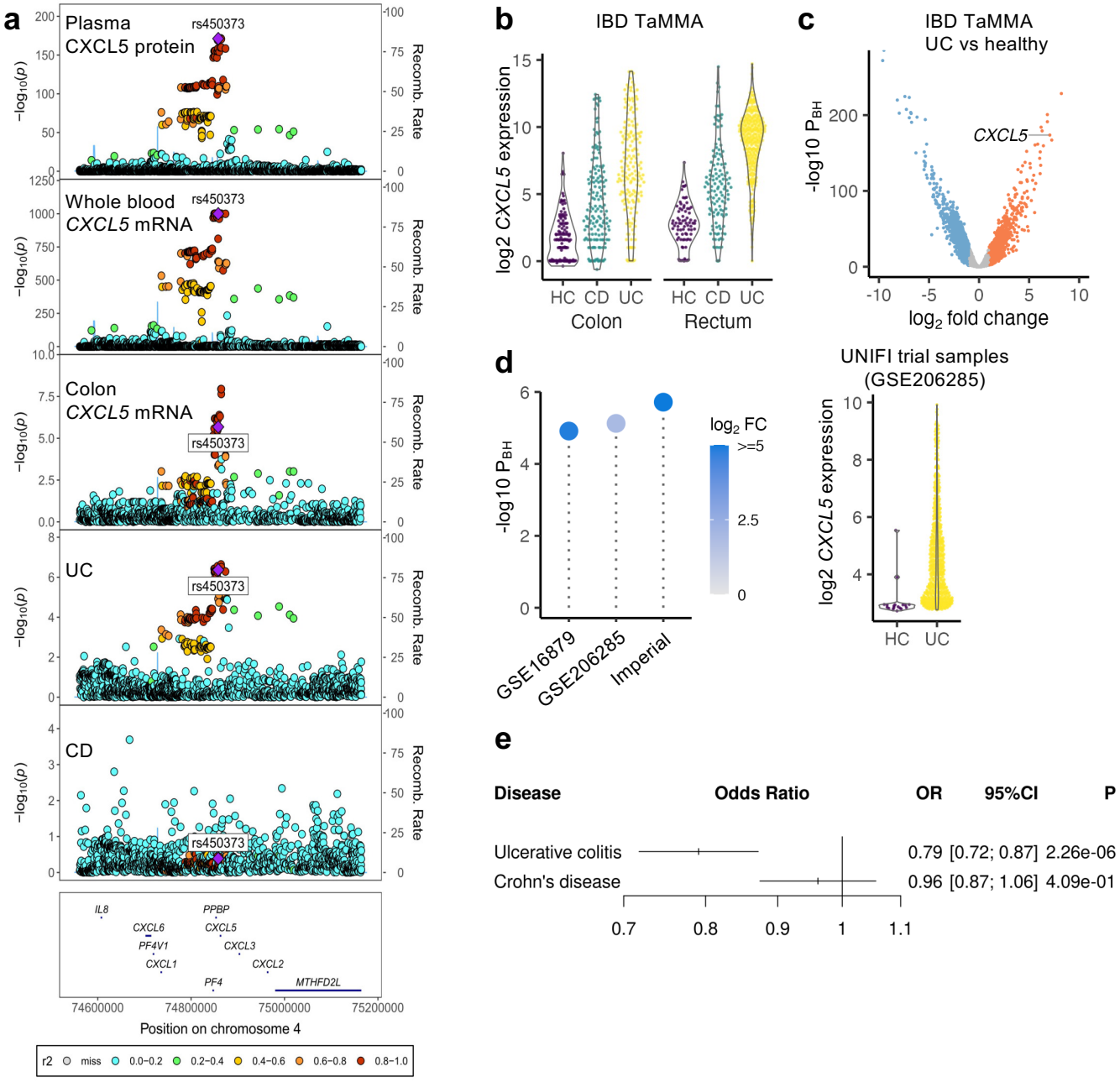

cis
trans
−log10(p)
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
TNFRSF9
CD244 [ST3GAL3]
CSF1
NGF
S100A12
IL6 [IL6R]
MMP1 [MEF2D;BCAN]
CCL2;CCL7;CCL8;CCL11;CCL13;CXCL6 [DARC]
SLAMF1
CD244
IL10RB [FAM163A]
TNFSF10 [CFHR1;CFHR3]
IL10
FGF21 [GCKR]
IL18 [NLRC4]
FLT3LG [SPRED2]
TGFA
IL18R1
PLAU [LY75]
FGF23 [SLC40A1]
CASP8
CCL19 [CTLA4]
CCL20
DNER
IL12B [BHLHE40]
CCL2;CCL7;CCL11 [ACKR2]
CDCP1
CCL11 [CCR1;CCR3]
CCL2 [CCR2]
CCL4 [CCR5]
TNFSF12 [ST3GAL6]
FLT3LG [RPN1;RAB7A]
CCL19 [ACKR4]
TNFSF12 [CHST2]
TNFSF10
TNFSF11 [TNFSF10]
TNFSF10 [KNG1]
IL12B [LPP]
TNFSF11 [CPN2]
HGF [HGFAC]
FGF19 [KLB]
IL8
CXCL6
CXCL1
CXCL5
CXCL11
CXCL9
CXCL10
FGF5
TNFSF12 [NFKB1]
FLT3LG [TET2]
EIF4EBP1 [KLKB1;F11]
SULT1A1 [KLKB1;F11]
FLT3LG [TERT]
SLAMF1 [ELL2]
GDNF
IL12B
LTA
CX3CL1
IL10
CCL19
IL1A
CDCP1
CCL20 [UNC5CL]
VEGFA
CXCL9 [RLG]
FGF21 [MLXIPL]
CCL11 [CCL24]
HGF
KITLG [PON1]
CXCL11 [SH2B2]
TNFSF11 [TNFRSF10A]
FGF19 [CYP7A1]
IL7
CXCL5;VEGFA [ZFPM2]
MMP1 [ZFPM2]
CCL13 [TRPS1]
TNFRSF11B
TNFSF11 [TNFRSF11B]
VEGFA [KCNV2]
CD274
CCL19
NGF [CTSL]
KITLG [ABCA1]
KITLG [MVB12B]
TNFSF12
CXCL5;CX3CL1;LIFR
CCL25;CX3CL1
IL15RA
CXCL5 [BRD3]
VEGFA [NRBF2;REEP3]
CXCL5 [NRBF2;REEP3]
TNFSF10
OSM [P2RY2]
MMP1
IL18
FLT3LG
PLAU [IL10RA]
IL10 [IL10RA]
CDCP1 [ST3GAL4]
CCL25 [B4GALNT3]
LTA [LTBR;TNFRSF1A]
CST5
LTA [SH2B3;TRAFD1]
CCL3 [SH2B3;TRAFD1]
CD6 [SH2B3]
SLAMF1 [ATXN2;SH2B3]
FLT3LG;IL12B [FLT3]
TNFSF11
IL12B [RAD51B]
TNFSF10 [SERPINA1]
IL12B [TRAF3]
CST5 [CA12]
NTF3 [NTRK3]
SULT1A1
KITLG [CETP]
CX3CL1
KITLG [LCAT]
IL17C
SLAMF1
PLAU [ASGR1]
TNFSF12
TNFRSF9 [TNFRSF13B]
TNFSF11;TNFRSF11B [VTN]
CCL7
CCL11
CCL8
CCL13
CCL23
CCL3
CCL4
OSM [PSMD3;CSF3]
TNFSF10 [APOH]
IL18R1 [APOH]
SLAMF1
PLAU [CHST9]
TNFSF10 [MEP1B]
CD5 [SMAD2]
TNFSF14
CCL23
TGFB1
SIRT2
TNFSF10 [PLAUR]
PLAU [PLAUR]
CST5;MMP10
FGF19
CCL25
FGF21
KITLG
CST5
ADA
CD40
KITLG [PLTP]
FGF23 [CYP24A1]
IL10RB
CD5;CD244;CXCL9;CXCL10;CXCL11;IL12B [SH2B3;TRAFD1]


a
CD244
CXCL9
CXCL10
CXCL11
IL12B
SH2B3
CD5
b
No. of proteins
No. of pQTL regions
c
IL12B
LPP
FLT3
RAD51B
SH2B3;
TRAFD1
TRAF3
HLA
BHLHE40
-log10 P
Chromosome
d
PLAUR
APOH
MEP1B
TNFSF10
KNG1
SERPINA1
CFHR1;
CFHR3
MYRF
-log10 P
Chromosome

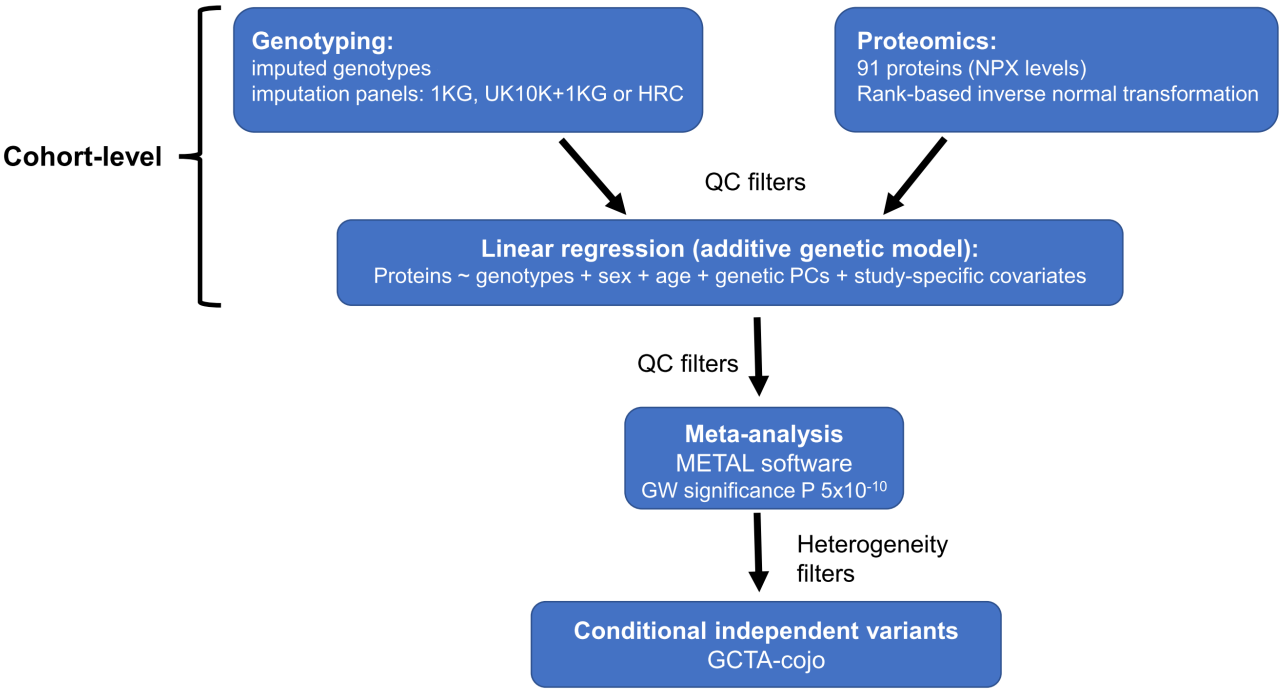

a
Inflammasome
NLRC4
Caspase-1
Pro-
IL18
Nucleus
IL18
IL18
Extracellular space
b
trans-pQTL for IL18 plasma protein
rs385076
−log10(p)
Recomb. Rate
cis-eQTL for NLRC4 whole blood mRNA
rs385076
XDH
NLRC4
MEMO1
YIPF4
DPY30
BIRC6
SPAST
TTC27
SLC30A6
LTBP1
31500000
32000000
32500000
33000000
Position on chromosome 2
r2
miss
0.0–0.2
0.2–0.4
0.4–0.6
0.6–0.8
0.8–1.0

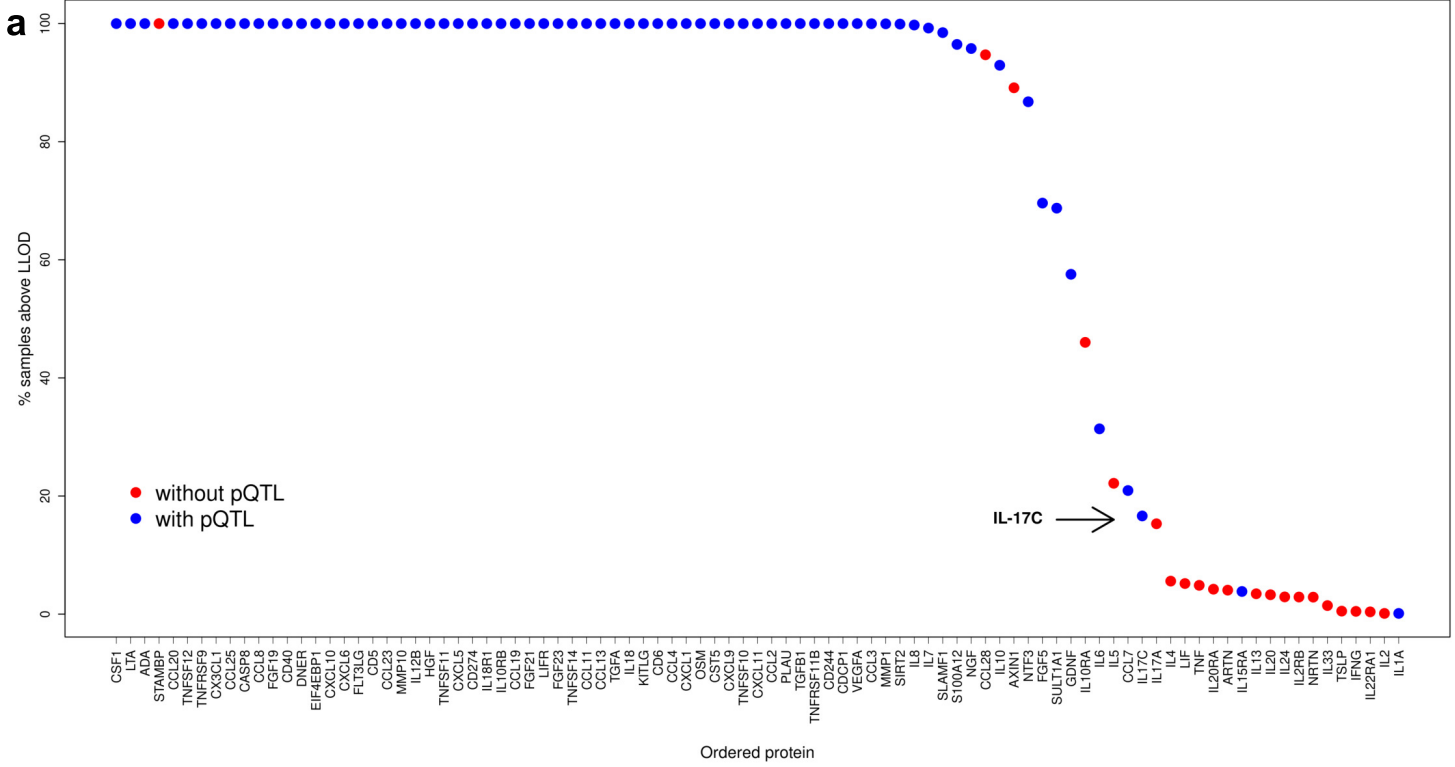

a
Multiple sclerosis
Plasma LTA protein
Whole blood LTBR mRNA
b
Multiple sclerosis
Plasma LTA protein
Whole blood LTBR mRNA
rs2364485
-log10(p)
Recomb. Rate
PLEKHG6
TNFRSF1A
SCNN1A
LTBR
TAPBPL
VAMP1
MRPL51
NCAPD2
CD27
6450000
6500000
6550000
6600000
Position on chromosome 12
r2
miss
0.0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0

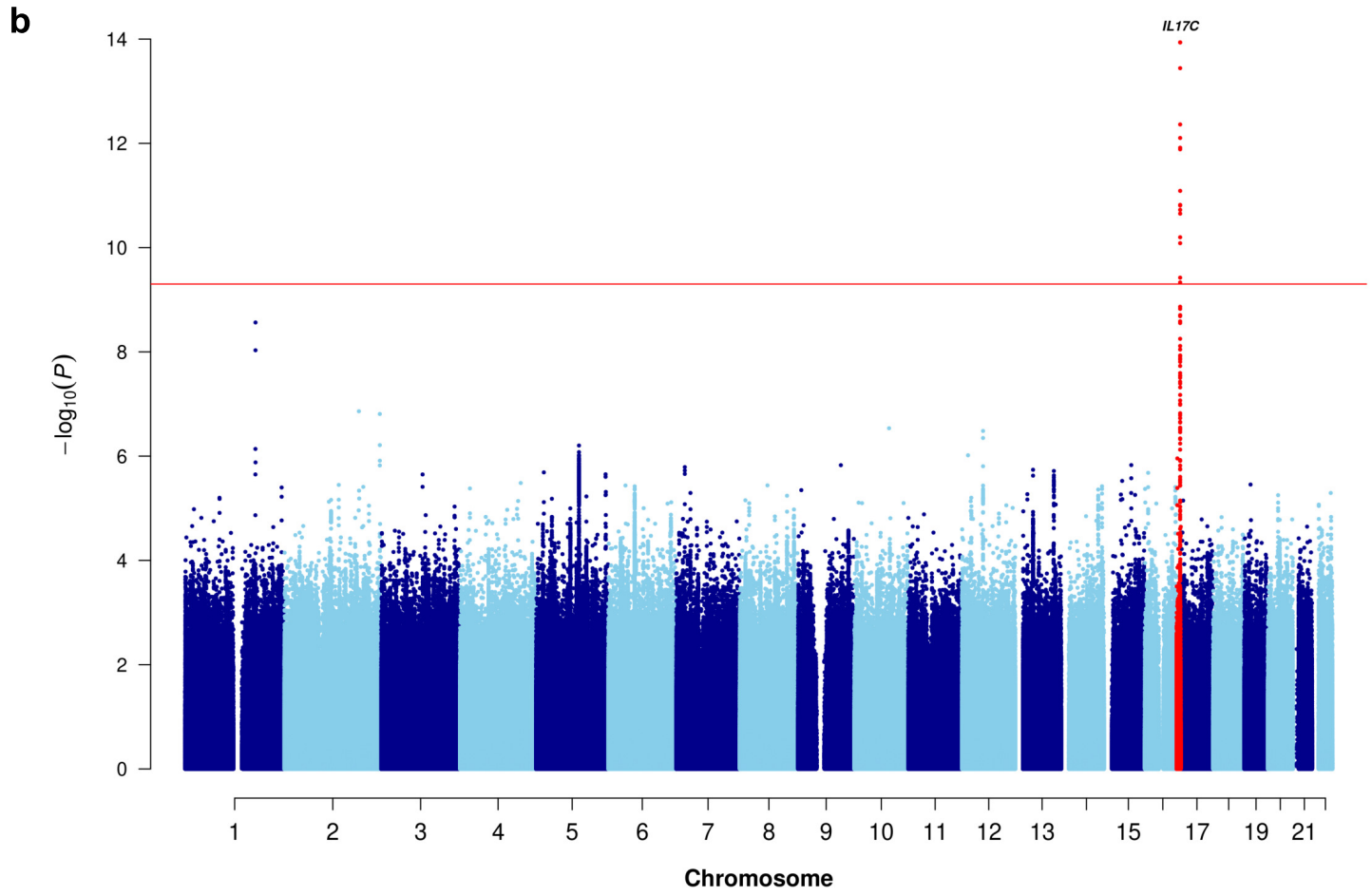

Immune-mediated outcomes
IgA nephropathy
Primary sclerosing cholangitis
Multiple sclerosis
Type 1 diabetes
Ulcerative colitis
Crohn's disease
Inflammatory bowel disease
Systemic lupus erythematosus
Gout
Eczema
Rheumatoid arthritis
Sepsis
Asthma
COVID-19 hospitalisation
Proteins
OR
1.5
1
0.5
Colocalises
Yes
No

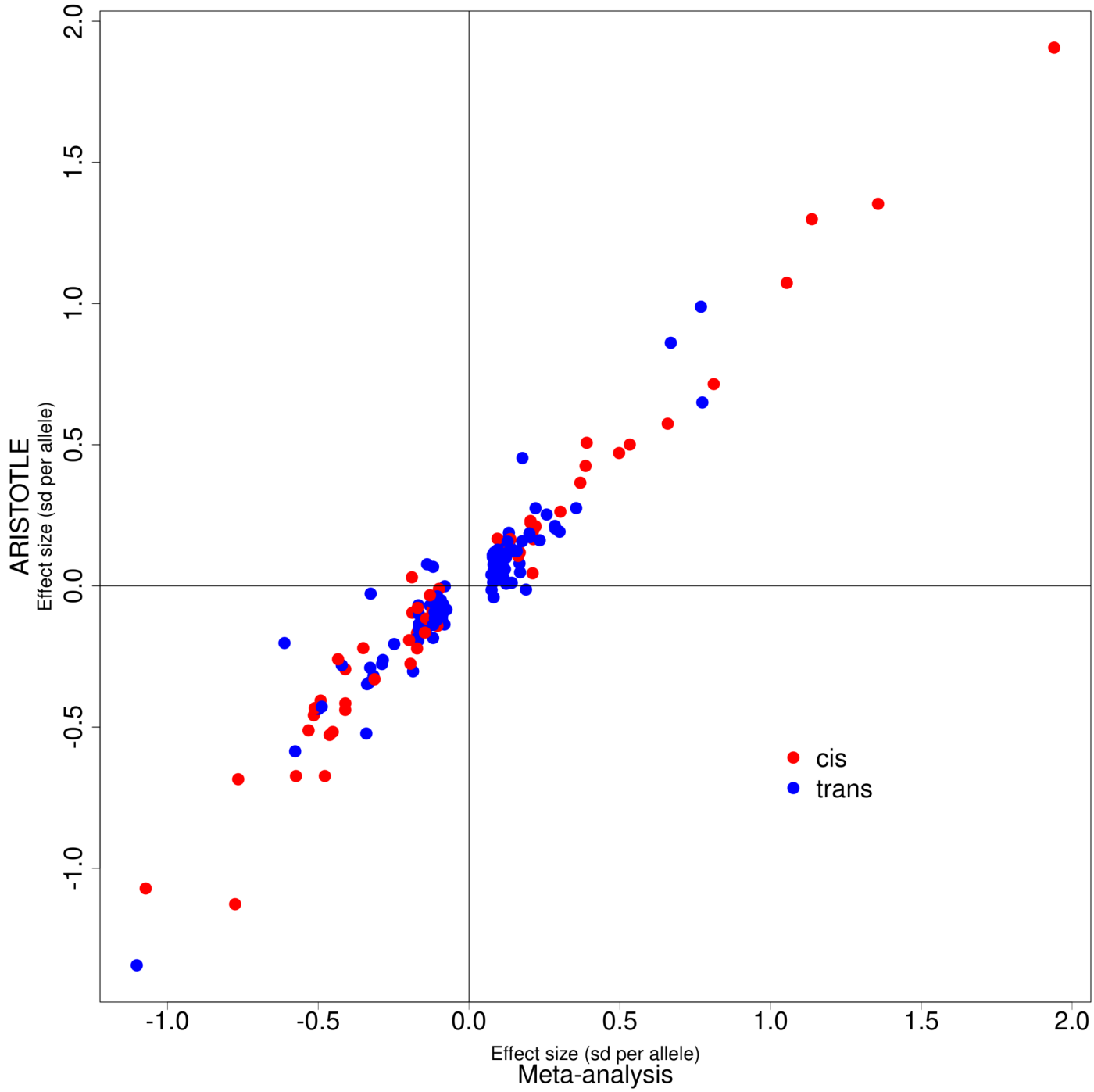

a
Plasma
CXCL5 protein
rs450373
Whole blood
CXCL5 mRNA
Colon
CXCL5 mRNA
UC
CD
−log10(p)
Recomb. Rate
IL8
CXCL6
PF4V1
CXCL1
PPBP
CXCL5
CXCL3
CXCL2
PF4
MTHFD2L
74600000
74800000
75000000
75200000
Position on chromosome 4
r2
miss
0.0–0.2
0.2–0.4
0.4–0.6
0.6–0.8
0.8–1.0
b
IBD TaMMA
log2 CXCL5 expression
HC CD UC
Colon
Rectum
c
IBD TaMMA
UC vs healthy
-log10 PBH
CXCL5
log2 fold change
d
-log10 PBH
log2 FC
>=5
2.5
0
GSE16879
GSE206285
Imperial
UNIFI trial samples
(GSE206285)
log2 CXCL5 expression
HC UC
e
Disease
Odds Ratio
OR
95%CI
P
Ulcerative colitis
0.79 [0.72; 0.87] 2.26e-06
Crohn's disease
0.96 [0.87; 1.06] 4.09e-01
0.7
0.8
0.9
1
1.1

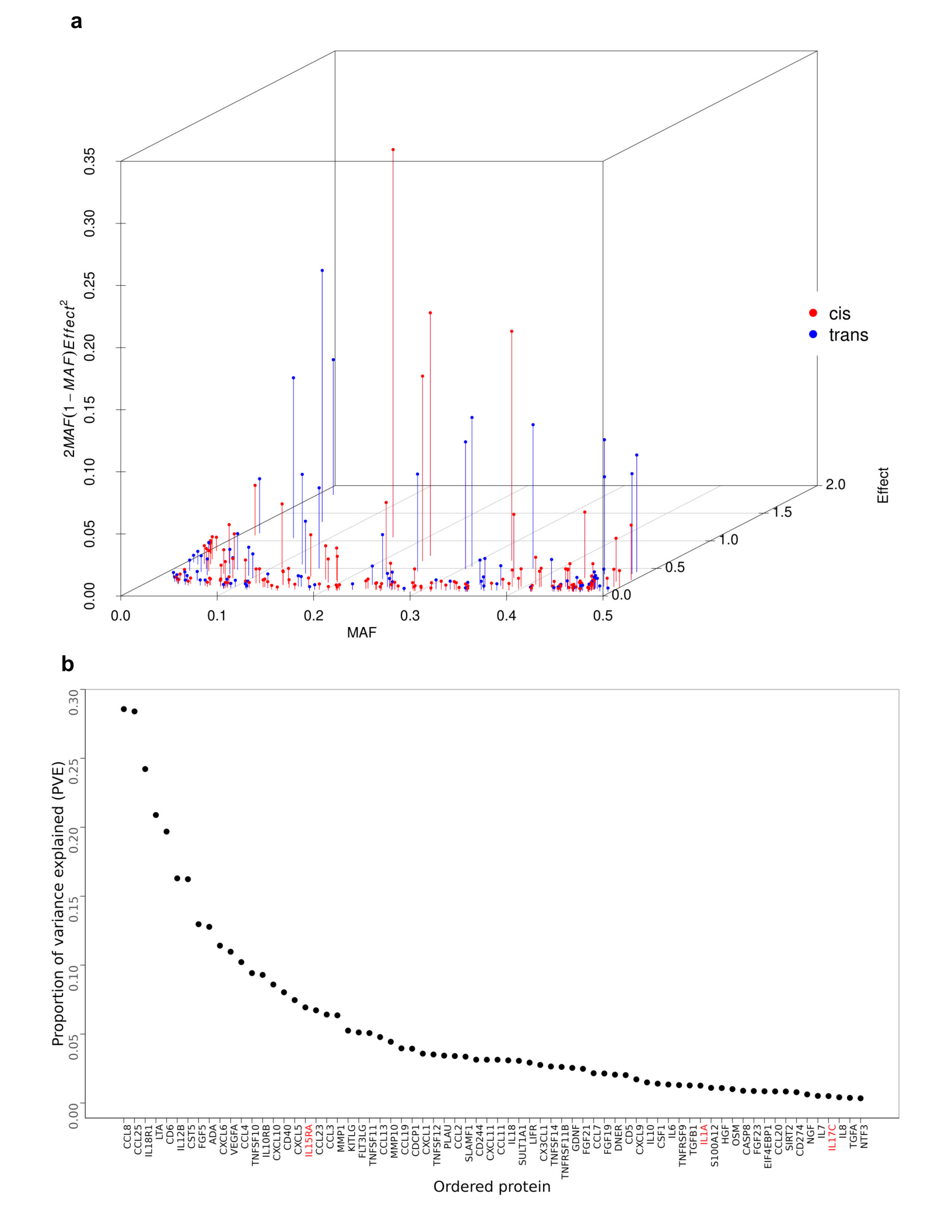
Analysis outline

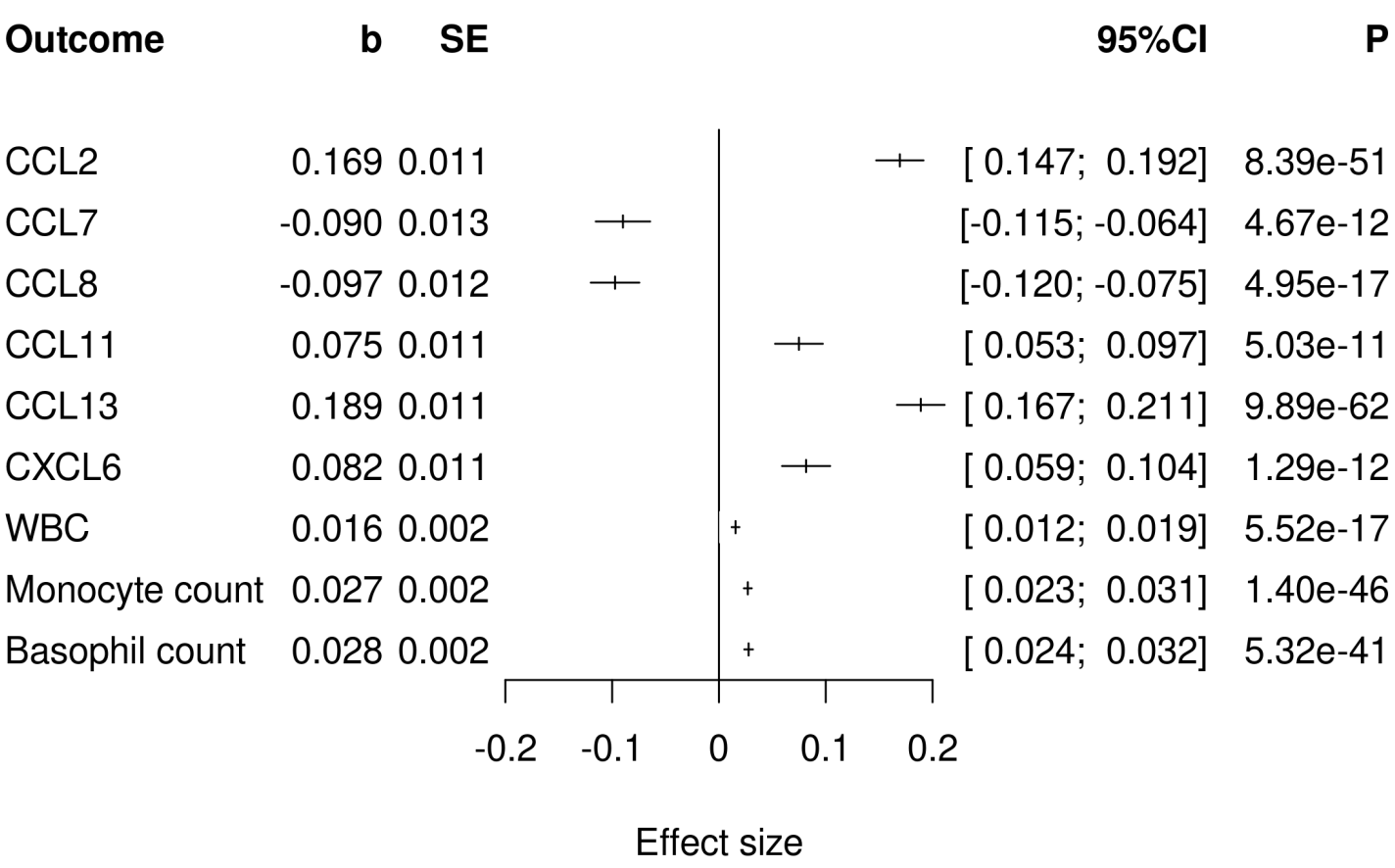


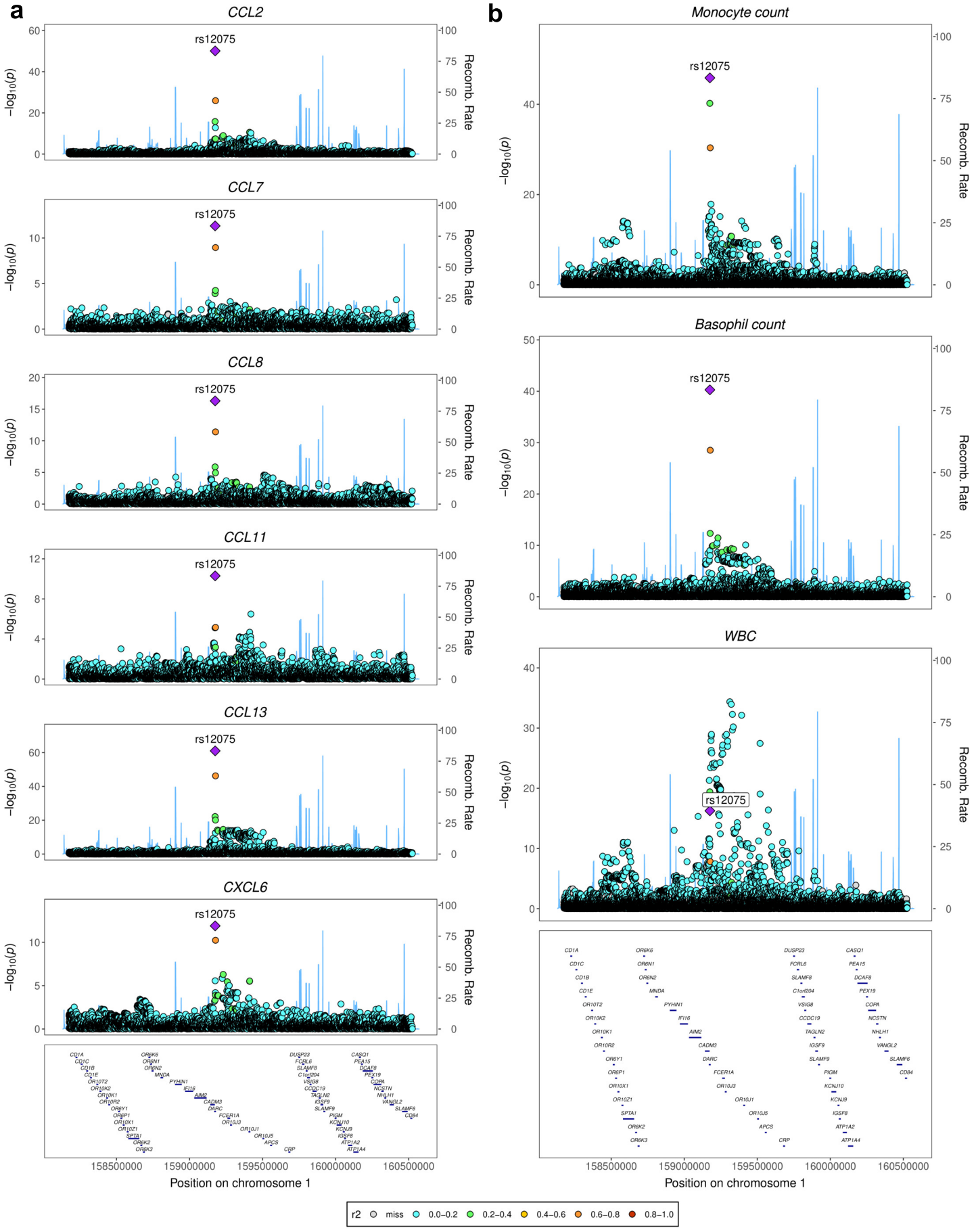

a
% samples above LLOD
without pQTL
with pQTL
IL-17C
Ordered protein

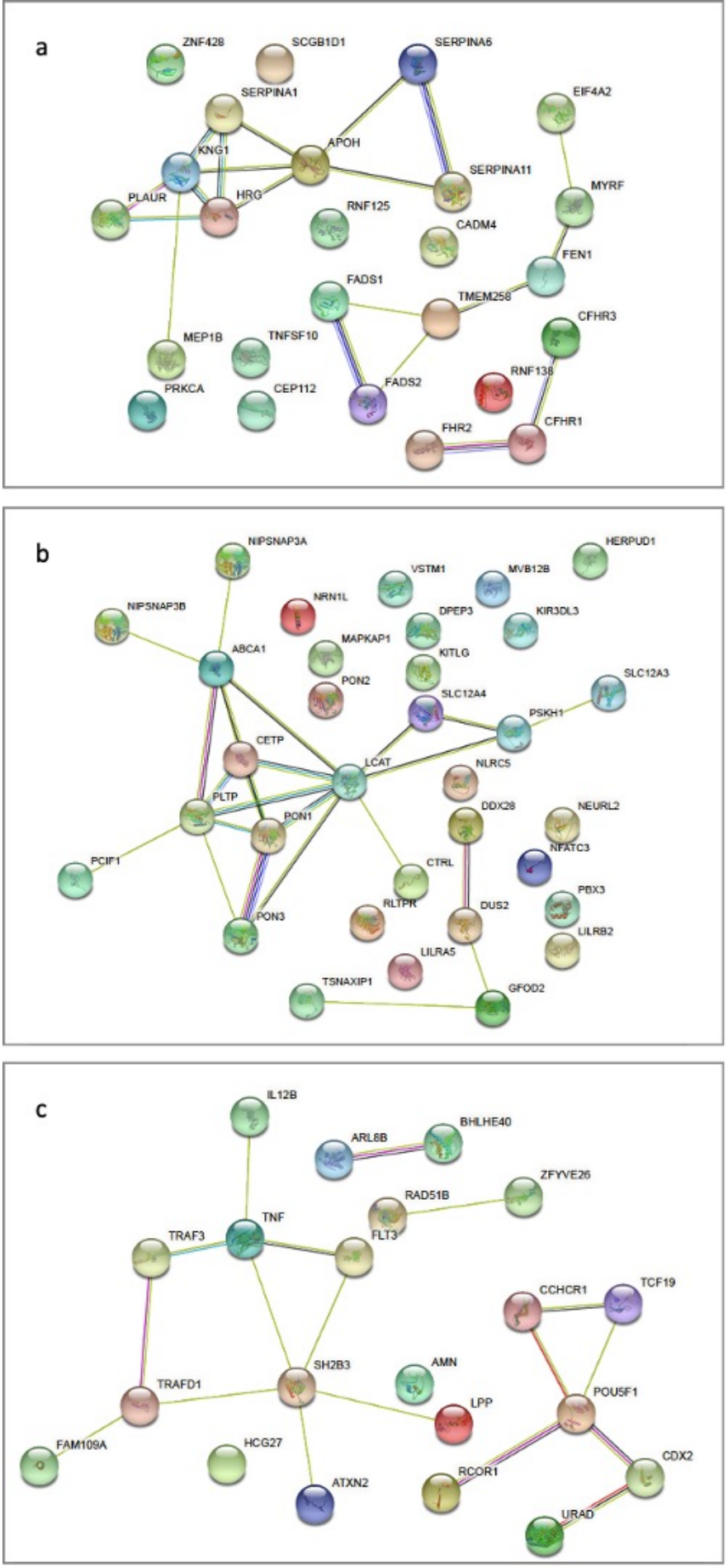

b
IL17C
−log10(P)
Chromosome

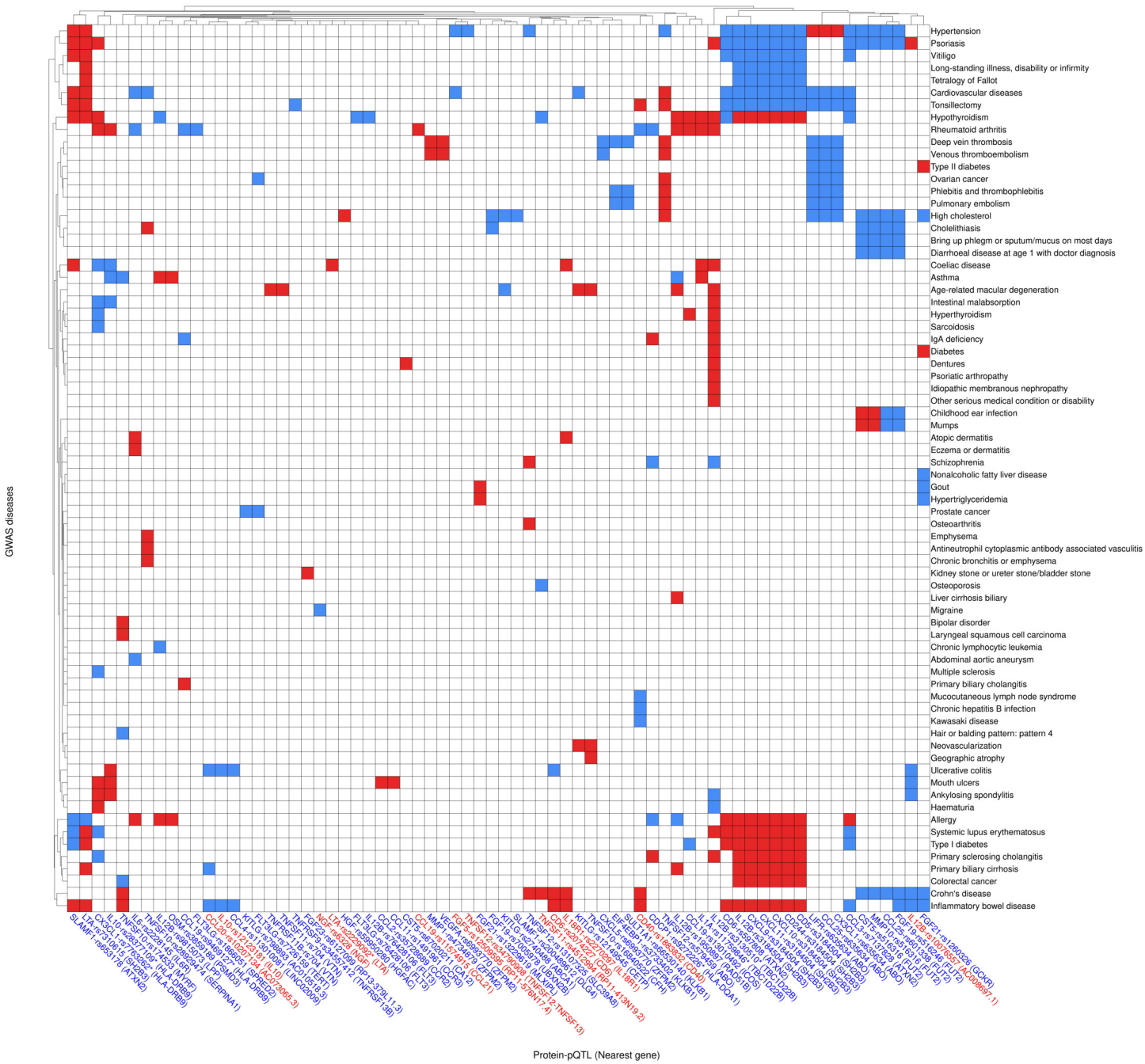

ARISTOTLE
Effect size (sd per allele)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Effect size (sd per allele)
Meta-analysis
cis
trans

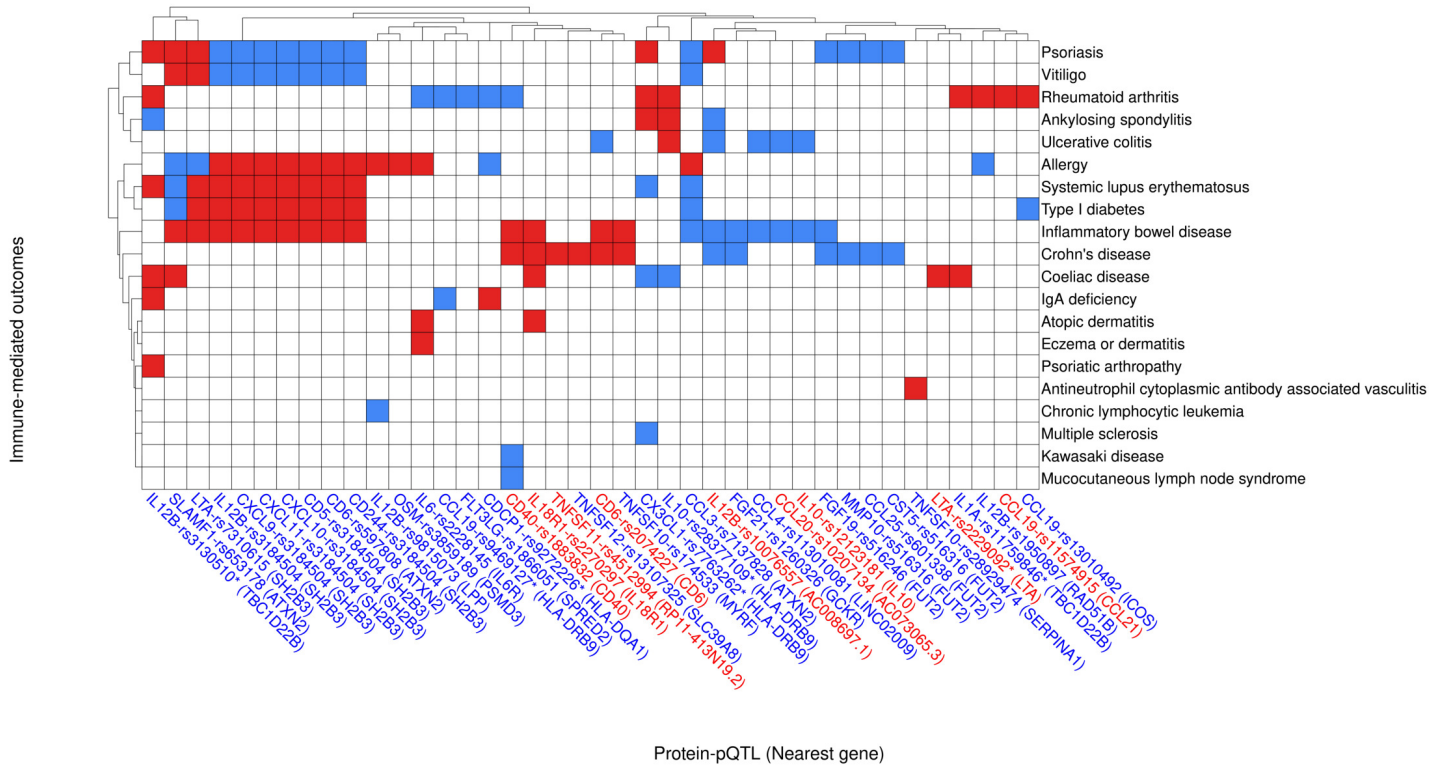

a
$2MAF(1-MAF)Effect^2$
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.0
0.1
0.2
0.3
0.4
0.5
MAF
0.0
0.5
1.0
1.5
2.0
Effect
cis
trans
b
Proportion of variance explained (PVE)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
CCL8
CCL25
IL18R1
LTA
CD6
IL12B
CST5
FGF5
ADA
CXCL6
VEGFA
CCL4
TNFSF10
IL10RB
CXCL10
CD40
CXCL5
IL15RA
CCL23
CCL3
MMP1
KITLG
FLT3LG
TNFSF11
CCL13
MMP10
CCL19
CDCP1
CXCL1
TNFSF12
PLAU
CCL2
SLAMF1
CD244
CXCL11
CCL11
IL18
SULT1A1
LIFR
CX3CL1
TNFSF14
TNFRSF11B
GDNF
FGF21
CCL7
FGF19
DNER
CD5
CXCL9
IL10
CSF1
IL6
TNFRSF9
TGFB1
IL1A
S100A12
HGF
OSM
CASP8
FGF23
EIF4EBP1
CCL20
SIRT2
CD274
NGF
IL7
IL17C
IL8
TGFA
NTF3
Ordered protein

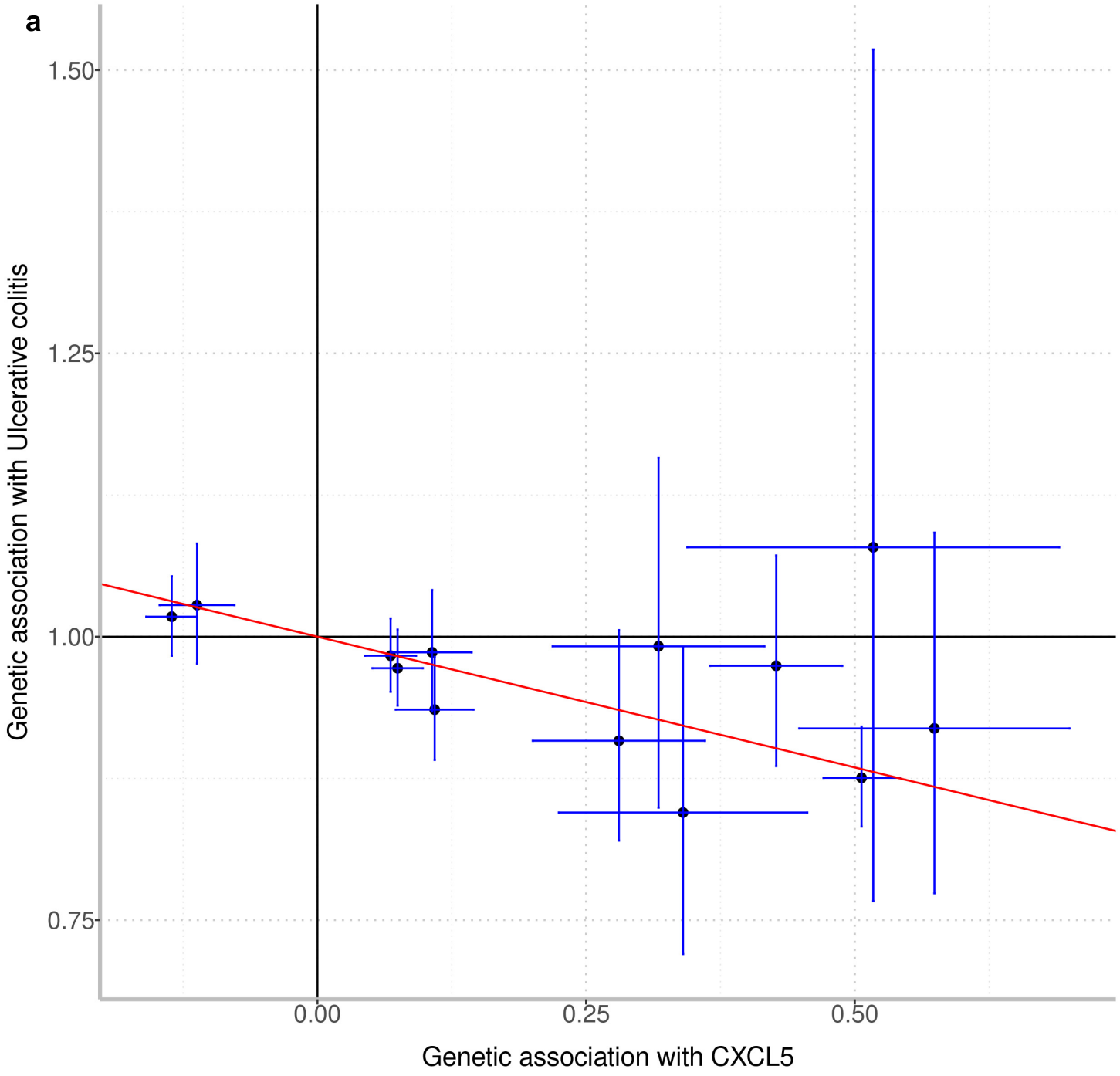

Outcome
b
SE
95%CI
P
CCL2 0.169 0.011 [0.147; 0.192] 8.39e-51
CCL7 -0.090 0.013 [-0.115; -0.064] 4.67e-12
CCL8 -0.097 0.012 [-0.120; -0.075] 4.95e-17
CCL11 0.075 0.011 [0.053; 0.097] 5.03e-11
CCL13 0.189 0.011 [0.167; 0.211] 9.89e-62
CXCL6 0.082 0.011 [0.059; 0.104] 1.29e-12
WBC 0.016 0.002 [0.012; 0.019] 5.52e-17
Monocyte count 0.027 0.002 [0.023; 0.031] 1.40e-46
Basophil count 0.028 0.002 [0.024; 0.032] 5.32e-41
-0.2
-0.1
0
0.1
0.2
Effect size


a
b
CCL2
CCL7
CCL8
CCL11
CCL13
CXCL6
Monocyte count
Basophil count
WBC
rs12075
−log10(p)
Recomb. Rate
Position on chromosome 1
158500000
159000000
159500000
160000000
160500000
r2
miss
0.0–0.2
0.2–0.4
0.4–0.6
0.6–0.8
0.8–1.0


a
ZNF428
SCGB1D1
SERPINA6
SERPINA1
EIF4A2
KNG1
APOH
SERPINA11
MYRF
PLAUR
HRG
RNF125
CADM4
FEN1
FADS1
TMEM258
CFHR3
MEP1B
TNFSF10
RNF138
PRKCA
CEP112
FADS2
FHR2
CFHR1
b
NIPSNAP3A
HERPUD1
VSTM1
MVB12B
NRN1L
NIPSNAP3B
DPEP3
KIR3DL3
MAPKAP1
ABCA1
KITLG
PON2
SLC12A3
SLC12A4
PSKH1
CETP
LCAT
NLRC5
PLTP
PON1
DDX28
NEURL2
NFATC3
PCIF1
CTRL
PBX3
RLTPR
DUS2
PON3
LILRB2
LILRA5
TSNAXIP1
GFOD2
c
IL12B
BHLHE40
ARL6B
ZFYVE26
RAD51B
TNF
TRAF3
FLT3
CCHCR1
TCF19
SH2B3
AMN
TRAFD1
LPP
POU5F1
FAM109A
HCG27
CDX2
RCOR1
ATXN2
URAD
Known Interactions
from curated databases
experimentally determined
Predicted Interactions
gene neighborhood
gene fusions
gene co-occurrence
Others
textmining
co-expression
protein homology


GWAS diseases
Hypertension
Psoriasis
Vitiligo
Long-standing illness, disability or infirmity
Tetralogy of Fallot
Cardiovascular diseases
Tonsillectomy
Hypothyroidism
Rheumatoid arthritis
Deep vein thrombosis
Venous thromboembolism
Type II diabetes
Ovarian cancer
Phlebitis and thrombophlebitis
Pulmonary embolism
High cholesterol
Cholelithiasis
Bring up phlegm or sputum/mucus on most days
Diarrhoeal disease at age 1 with doctor diagnosis
Coeliac disease
Asthma
Age-related macular degeneration
Intestinal malabsorption
Hyperthyroidism
Sarcoidosis
IgA deficiency
Diabetes
Dentures
Psoriatic arthropathy
Idiopathic membranous nephropathy
Other serious medical condition or disability
Childhood ear infection
Mumps
Atopic dermatitis
Eczema or dermatitis
Schizophrenia
Nonalcoholic fatty liver disease
Gout
Hypertriglyceridemia
Prostate cancer
Osteoarthritis
Emphysema
Antineutrophil cytoplasmic antibody associated vasculitis
Chronic bronchitis or emphysema
Kidney stone or ureter stone/bladder stone
Osteoporosis
Liver cirrhosis biliary
Migraine
Bipolar disorder
Laryngeal squamous cell carcinoma
Chronic lymphocytic leukemia
Abdominal aortic aneurysm
Multiple sclerosis
Primary biliary cholangitis
Mucocutaneous lymph node syndrome
Chronic hepatitis B infection
Kawasaki disease
Hair or balding pattern: pattern 4
Neovascularization
Geographic atrophy
Ulcerative colitis
Mouth ulcers
Ankylosing spondylitis
Haematuria
Allergy
Systemic lupus erythematosus
Type I diabetes
Primary sclerosing cholangitis
Primary biliary cirrhosis
Colorectal cancer
Crohn's disease
Inflammatory bowel disease
Protein-pQTL (Nearest gene)


Immune-mediated outcomes
Psoriasis
Vitiligo
Rheumatoid arthritis
Ankylosing spondylitis
Ulcerative colitis
Allergy
Systemic lupus erythematosus
Type I diabetes
Inflammatory bowel disease
Crohn's disease
Coeliac disease
IgA deficiency
Atopic dermatitis
Eczema or dermatitis
Psoriatic arthropathy
Antineutrophil cytoplasmic antibody associated vasculitis
Chronic lymphocytic leukemia
Multiple sclerosis
Kawasaki disease
Mucocutaneous lymph node syndrome
IL12B-rs3130510* (TBC1D22B)
SLAMF1-rs6553178 (ATXN2)
LTA-rs3106615 (SH2B3)
IL12B-rs3184504 (SH2B3)
CXCL9-rs3184504 (SH2B3)
CXCL11-rs3184504 (SH2B3)
CXCL10-rs3184504 (SH2B3)
CD5-rs3184504 (SH2B3)
CD6-rs3184504 (SH2B3)
CD244-rs597808 (ATXN2)
IL12B-rs3184504 (SH2B3)
OSM-rs3859189 (LPP)
IL6-rs2228145 (IL6R)
CCL19-rs946911 (RSMD3)
FLT3LG-rs1886051 (SPRED2)
CDCP1-rs9272226* (HLA-DRB9)
CD40-rs1883832 (CD40)
IL18R1-rs2270297 (IL18R1)
TNFSF11-rs4512994 (RP11-413N19.2)
CD6-rs2014227 (CD6)
TNFSF12-rs1310725 (SLC39A8)
TNFSF10-rs1774533 (MYRF)
CX3CL1-rs7763262* (HLA-DRB9)
IL10-rs2837109* (ATXN2)
CCL3-rs1371828 (ATXN2)
IL12B-rs10076557 (AC008697.1)
FGF21-rs1260326 (GCKR)
CCL20-rs10271081 (LINC02009)
IL10-rs12123161 (IL10)
CCL4-rs11301134 (AC073065.3)
MMP10-rs516246 (FUT2)
FGF19-rs516316 (FUT2)
CST5-rs601338 (FUT2)
TNFSF10-rs8929414 (FUT2)
LTA-rs2229092* (LTA)
IL1A-rs1800587 (IL1A)
IL12B-rs1950897 (RAD51B)
CCL19-rs1574915 (CCL21)
CCL19-rs13010492 (ICOS)
Protein-pQTL (Nearest gene)



b

CXCL5 measurement	Odds Ratio	OR	95%CI	P
Plasma protein		0.60	[0.58; 0.62]	6.44e−172
Whole blood mRNA		0.58	[0.52; 0.64]	8.19e−24
Colon mRNA		0.67	[0.56; 0.79]	2.10e−06