

Cell-type specific analyses in blood and gut identify cis-eQTL matching 140 IBD risk loci and entrectinib as repurposing candidate

Received: 4 November 2024

Accepted: 4 August 2026

Published online: 01 September 2026

 Check for updates

Hélène Perée¹, Viacheslav A. Petrov^{1,28,29}, Yumie Tokunaga^{1,29}, Alexander Kvasz¹, Frédéric Farnir², Sophie Vieujean³, Sarah Regimont¹, Myriam Mni¹, Marie Wéry¹, Samira Azarzar², Sophie Jacques¹, Nicolas Fouillien¹, Latifa Karim⁴, Manon Deckers⁴, Emilie Detry⁴, Alice Mayer⁵, Raafat Stephan⁶, Keith Harshman⁷, Yasutaka Mizoro¹, Catherine Reenaers³, Catherine Van Kemseke³, Odile Warling³, Virginie Labille³, Sophie Kropp³, Maxime Poncin³, Anne Catherine Moreau³, Benoit Servais³, Jean-Philippe Loly³, SYSCID Consortium*, Wouter Coppieters⁴, Emmanouil Dermitzakis⁷, Edouard Louis³, Michel Georges^{1,8}✉, Haruko Takeda¹ & Souad Rahmouni¹✉

Genes whose expression is affected, in a consistent manner, by GWAS-identified risk variants and the disease process, constitute preferred drug targets. We herein combine cis-eQTL analysis in 27 sorted blood cell populations and 43 intestinal cell types identified by single cell RNA-Seq in the ileum, colon and rectum, with information on gene expression in patients, to search for putative drug targets for inflammatory bowel disease. We detect >95 K cis-eQTL that affect >13 K e-genes and cluster in >24 K regulatory modules. We uncover matching regulatory modules for 140 risk loci, implicating >300 e-genes not previously connected with inflammatory bowel disease, and find 152 matching e-genes whose expression is perturbed in the blood or gut of patients. We identify entrectinib, a small molecule inhibiting the NLRP3 inflammasome by binding NEK7, as a possible repurposing candidate.

The growing prevalence of common complex diseases (CCD) threatens the sustainability of health care systems. There is a need for more effective prevention and treatment^{1–3}.

Predisposition to most CCD has a considerable inherited component⁴. Accordingly, GWAS have systematically uncovered tens to hundreds of risk loci that explain up to ~50% of the heritability^{5,6}.

Identifying the genes that are perturbed by the risk variants in these loci is a key goal, as these constitute preferred drug targets^{7–10}.

A common feature of CCD is that most associations are driven by regulatory as opposed to coding variants. For example, amongst the > 200 risk loci mapped for inflammatory bowel disease (IBD), only 14 encompass ORF-altering variants in the credible set (i.e., the variants in

¹Unit of Animal Genomics, GIGA Institute & Faculty of Veterinary Medicine, University of Liège, Brussels, Belgium. ²Unit of Biostatistics and Bioinformatics. FARA & Faculty of Veterinary Medicine, University of Liège, Brussels, Belgium. ³Department of Gastroenterology, Faculty of Medicine, CHU & GIGA Institute, University of Liège, Brussels, Belgium. ⁴Genomics Core Facility, GIGA Institute, University of Liège, Brussels, Belgium. ⁵GIGA Bioinformatics Core Facility, GIGA Institute, University of Liège, Brussels, Belgium. ⁶GIGA in Vitro Imaging and Cell Sorting Core Facility, GIGA Institute, University of Liège, Brussels, Belgium. ⁷Genomics Core Facility, University of Geneva, Switzerland, Switzerland. ⁸Welbio Research Institute, Brussels, Belgium. ²⁸Present address: Systems Ecology group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. ²⁹These authors contributed equally: Viacheslav A. Petrov, Yumie Tokunaga.*Lists of authors and their affiliations appear at the end of the paper.

✉ e-mail: michel.georges@uliege.be; srahmouni@uliege.be

high linkage disequilibrium (LD) with high posterior probability to be causal), thereby convincingly pinpointing causative genes and hence putative drug targets¹¹. The prevailing hypothesis is that for the remaining loci, regulatory variants perturb the expression of causative genes in *cis*. The challenges are to identify the causative regulatory variants, and—most importantly—the genes they perturb in *cis*. A common approach to achieve the latter is eQTL analyses in disease relevant cell types collected from healthy individuals. Indeed, it is reasonable to assume that: (i) the effect of most regulatory variants will manifest by genotype-specific differences in messenger levels (i.e., eQTL effects), and (ii) the corresponding eQTL, caused by common variants, are detectable in healthy individuals. If a *cis*-eQTL underpins a GWAS-identified risk locus it will exhibit an association pattern (vector of association $\log(1/p)$ values of all the SNPs in the locus with the gene's expression levels, henceforth referred to as EAP for expression association pattern) that is very similar to the association pattern of the risk locus (henceforth referred to as DAP for disease association pattern), provided that GWAS and eQTL studies were conducted in cohorts of same ethnicity (and hence same LD structure). A number of colocalization methods have been developed to quantify the resemblance between DAP and EAP^{12–15}.

Several datasets have been assembled towards that goal. The most prominent is GTEx that provides eQTL information for 52 tissues in up to 838 subjects¹⁶. A common observation from colocalization experiments is that “matching” eQTL are only found for ~25% of risk loci, raising questions about the molecular underpinnings of the remaining loci¹⁷. For example, we previously identified matching eQTL in 63 of 200 analyzed IBD risk loci using 23,650 eQTL identified in six circulating immune cell types and intestinal biopsies at three locations (CEDAR-1 dataset)¹⁵.

A plausible explanation for this high proportion of “orphan” risk loci, is that the corresponding eQTL remain to be discovered. This could be because the relevant cell populations were underrepresented in the samples studied to date, because the manifestation of the eQTL is context dependent (f.i. after initiation of the disease process), or because the eQTL discovered thus far (with small sample sizes when compared to disease GWAS) are distinct from the ones that drive CCD¹⁸.

In this work, we followed up on the first hypothesis, i.e., the relevant cell populations are underrepresented in the samples studied to date, in the context of IBD. Towards that goal we established the CEDAR-2 eQTL dataset based on (i) bulk transcriptome data from 27 circulating immune cell populations for 200 healthy individuals, and (ii) single-cell RNA (scRNA) transcriptome data for intestinal biopsies collected at three locations (ileum, colon and rectum) from 60 healthy individuals. We report the identification of matching *cis*-eQTL for 140 of 206 examined IBD risk loci, identify more than 300 novel candidate causal genes, and pinpoint opportunities for drug repurposing.

Results

60,113 *cis*-eQTL affecting 11,874 eQTL genes in 27 circulating immune cell populations cluster in 22,067 *cis*-acting regulatory modules

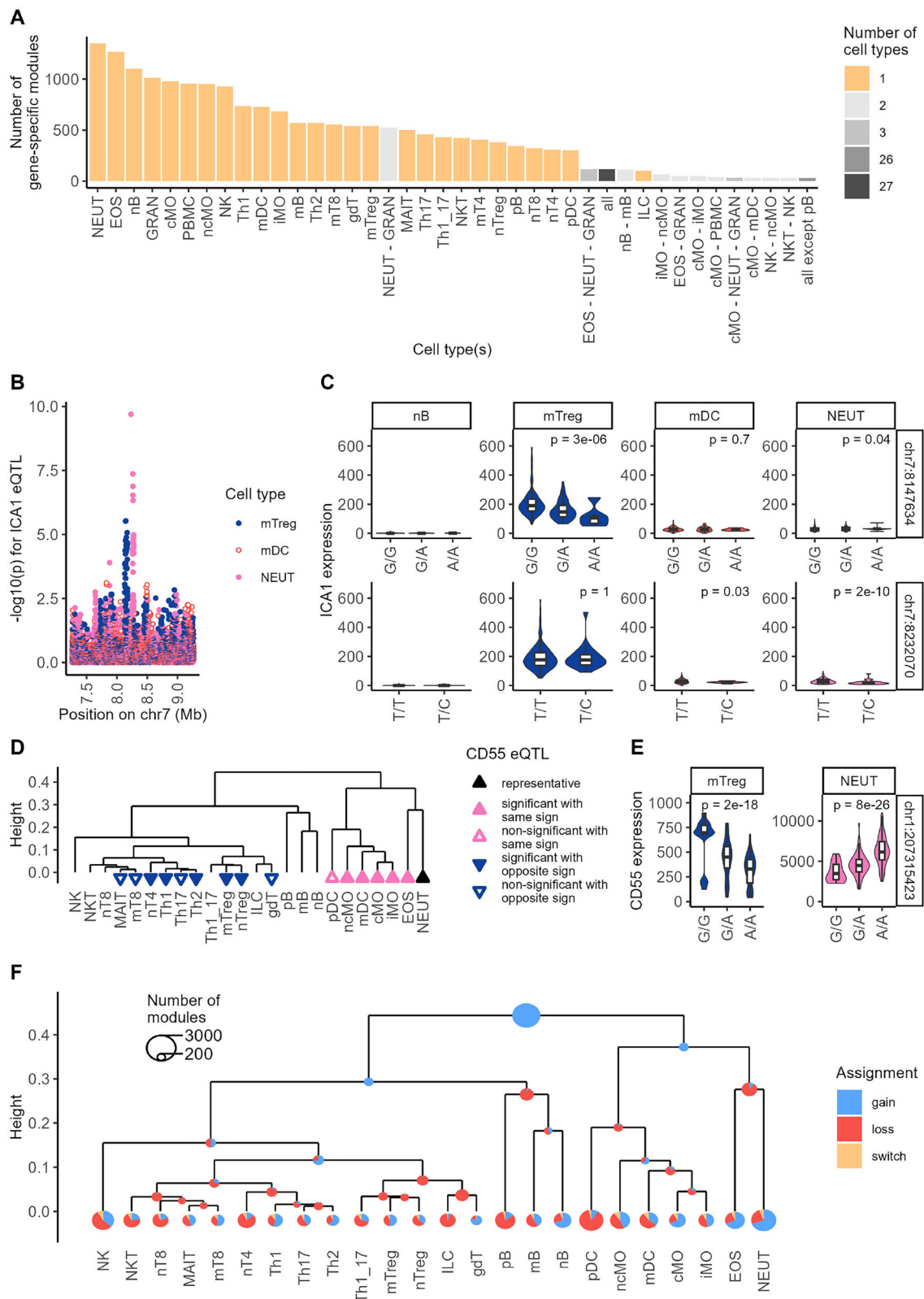
We collected blood from 251 healthy individuals of both sexes (Supplementary Data 1). We genotyped all individuals with Illumina's OmniExpress array interrogating ~700 K SNPs, and augmented genotype data to ~6.3 million (M) variants by imputation. For each individual, in addition to collecting peripheral blood mononuclear cells (PBMC), we sorted 26 distinct immune cell populations by fluorescent activated or magnetic cell sorting (FACS or MACS) (SFig. 1 and Supplementary Data 2). We produced next generation sequencing (NGS) libraries for each of the 27 cellular fractions of each individual (5292 mRNA SMART-Seq HT libraries), and generated an average of ~12.6 M, 2 × 150 bp (3180 libraries) or ~13 M, 2 × 50 bp (2112 libraries) paired-end reads per library on Illumina instruments (Supplementary Data 3).

Transcriptome data were used to check the assignment of libraries to individuals and cell types, and errors corrected (SFig. 2A, B). The numbers of genes detected averaged 16,615 per cell type (range: 12,776–19,042) (Supplementary Data 4). Hierarchical clustering of cell types based on average gene expression levels generated a dendrogram largely consistent with hematopoietic ontogeny (SFig. 2C).

We performed *cis*-eQTL analyses using a custom-made pipeline including QTLtools¹⁹, regressing expression level on alternate allele dosage and including an average of 24.4 expression principal components (PC) in the model (Supplementary Data 4; M&M). We identified 60,113 eQTL (within cell type FDR ≤ 0.05), influencing 11,874 eQTL genes (e-genes) (Supplementary Data 5). The number of eQTL detected per cell type (average: 2226; range: 608–3448) was largely determined by the number of samples available for analysis ($r^2 = 0.67$) (SFig. 3A). The number of eQTL detected in PBMC was larger than expected given sample numbers, consistent with PBMC encompassing multiple cell types. The number of eQTL detected in plasmacytes was lower than expected. Controlling for the variable abundance of immunoglobulin transcripts did not increase the number of detected plasmacyte eQTL (SFig. 3B). On average, *cis*-eQTL explained 19.7% (range: 3.2–83.6%) of variance in gene expression (SFig. 3C). Secondary *cis* signals (range: 2–5) were detected for 6% of *cis*-eQTL, when repeating eQTL analysis conditional on the previous signals (SFig. 3D).

eQTL affecting the same gene in multiple cell types were merged if sharing a similar association pattern (EAP) as evaluated using theta ($|\theta| \geq 0.6$)¹⁵, yielding 23,831 gene-specific modules (Supplementary Data 6). Modules were augmented with 5840 tier-2 eQTL that would match ($|\theta| \geq 0.6$) at least one significant eQTL in the module (see M&M). Modules can be characterized by a vector of 0 and 1 that indicates in which cell type(s) the module is active. The 31 most common vectors were dominated by 17,439 modules (i.e., 73.2%) that are active in only one of each of the 27 cell types, and included 117 modules that are—a *contrario*—active in all 27 cell types (Fig. 1A). The overdispersion of the number of active cell types (i.e., excess of modules active either in very few or many cell types) was highly significant ($p < 0.001$) (see M&M).

A module can be active in cell type A but not in cell type B because: (i) the gene is expressed in cell type A but not in cell type B, (ii) the gene is expressed in both cell types but the eQTL is only active in cell type A, or (iii) the gene is in distinct modules in cell type A and B (i.e., different EAP in cell types A and B). Expression of the gene was below detection levels in cell type B (i.e., first scenario) in 17.7% of cases. Analysis of the modules indicated that the third scenario (module switch) accounts for 8.5% of cases. We performed an interaction test to evaluate the importance of scenario 2 (see M&M) in the remaining 73.8% of the cases. We estimated the proportion of alternative hypothesis (π_1) following Storey et al.²⁰ at 94%, hence indicating that the eQTL was indeed not active in cell type B in 69.6% of cases despite the fact that the gene was expressed in cell type B (Fig. 1B–C). Estimates of π_1 were nearly identical when considering genes that were more strongly expressed in cell type A (95%), or genes that were more strongly expressed in cell type B (93%) (SFig. 3E), hence arguing against reduced gene expression generating false positive scenarios 2 by affecting statistical power. For modules active in more than one cell type, the direction of the effects was the same across cell types in 96.1% of the cases, yet there were 250 cases for which the sign of the effect switched between cell types (Fig. 1D–E; Supplementary Data 6). The proportion of shared eQTL with opposite sign increased with ontogenic distance between cell types ($p_{\text{logit}} = 4 \times 10^{-33}$; Supplementary Data 7). Modules were assigned to nodes and leaves in the ontogenic dendrogram using the patterns of module sharing across cell types (vectors of 0's and 1's) (see M&M), suggesting extensive remodeling of eQTL activity during hematopoiesis, including the common loss of progenitor cell eQTL and gain of new eQTL by differentiated cells (Fig. 1F). Genes with *cis*-eQTL assigned towards the root (presumed to be active in progenitor



cells, $n = 2776$) were enriched in GO terms: antigen processing and presentation, amino-acid metabolism, organic acid and small molecule metabolic processes, and cell motility, while genes with *cis*-eQTL assigned towards the leaves ($n = 9201$) did not show enrichment in any GO terms (Supplementary Data 8). Sharing of modules by cell types was largely determined by their ontogenic proximity ($p = 9 \times 10^{-68}$; SFig. 3F). However, this may reflect excess sharing of gene expression

as much as excess sharing of polymorphic gene regulation between closely related cell types. To distinguish between these two factors, we restricted the analyses to genes subject to multiple regulatory modules. For those we measured (i) the excess sharing of the same module between cell types, and (ii) the excess switching of modules between cell types. We showed that B cells, T cells (including CD4 and CD8) monocytes/dendritic cells, and granulocytes were forming four

Fig. 1 | Blood regulatory eQTL modules. **A** Gene-specific modules grouped by cell type in which they are active, ordered according to the frequency of occurrence (40 most frequent combinations out of 3,572 observed). The most common modules are dominated by cell type-specific ones, i.e., only active in one cell type (27 yellow bars). Also, amongst the top 40 combinations, are modules that are active in all (i.e., ubiquitous eQTL; black bar). The remaining correspond to modules that are active in 2 or 3 closely related cell types. Cell type abbreviations are as in Supplementary Data 2. **B, C** Example of a gene (*IC41*) controlled by two modules operating respectively in memory regulatory T cells (mTreg, module #12,606, blue in EAP and violin plots) and neutrophils (NEUT, module #16,344, pink in EAP and violin plots). Neither is active in myeloid dendritic cells (mDC, orange EAP and violin plot) despite the fact that the gene is expressed at high levels in this cell type (scenario 2). The gene is expressed at low levels in naïve B cells (nB) in which neither module is active either (scenario 1). Individuals are sorted by genotype at the lead SNPs for modules #12,605 (upper row; $n_{GG} \geq 128$, $n_{GA} \geq 44$, $n_{AA} \geq 5$) and #16,344 (lower row; $n_{TT} \geq 164$, $n_{TC} \geq 20$). Boxplots in the violins mark median, interquartile range (IQR), and 1.5xIQR distance from IQR. **D, E** Example of a gene-specific module (*CD55*, module #678) that is active in 16 cell types, including all

myeloid cell types, and nine of 12 types of T-cells (filled triangles: significant eQTL; empty triangles (tier-2): non-significant eQTL but matching ($|\theta| \geq 0.6$) at least one of the significant eQTL in the module). The sign of the eQTL is opposite in myeloid cells (pink, upward pointing triangles) and T lymphocytes (blue, downward pointing triangles). The reference sign is determined by the most significant, “representative” eQTL in the module (black triangle). Violin plots show the distribution of *CD55* expression (DESeq2 normalized) in memory regulatory T cells (left, blue) and neutrophils (right, pink) for individuals sorted by genotype at the lead variant for the consensus EAP ($n_{GG} \geq 13$, $n_{GA} \geq 69$, $n_{AA} \geq 103$). Boxplots as in (C). **F** Gains of gene-specific modules were assigned to the “most recent common ancestor” (MRCA) of all nodes/leaves in which the module is active (blue segments). Losses of gene-specific modules were assigned to the MRCA of all descendant nodes/leaves in which the module was not active (red segments). If the loss of a module coincided with the gain of a module for the same gene, we assumed that a module switch occurred (yellow segments). The diameter of the circles is indicative of the corresponding numbers, *per* legend. All reported *p*-values are from eQTL analyses computed as described in Methods; they are nominal and two-sided. Source data are provided as a Source data file.

clusters, with excess module sharing within cluster, and excess module switching between clusters (SFig. 3G).

We merged gene-specific modules with matching EAP “across genes”, yielding a total of 22,067 regulatory modules. Modules were augmented with 7402 tier-2 eQTL matching at least one significant *cis*-eQTL in the module, recruiting 759 extra genes (M&M; Supplementary Data 9). We observed 2470 modules (11.2%) comprising more than one gene. On average such multigenic modules encompassed 2.9 genes (range: 2–32) and were active in 9.3 cell types (SFig. 3H). Gene-specific modules active in multiple (but not all) cell types had more chance to join a multigenic module than gene-specific modules active in only one cell type (SFig. 3I). eQTL effects with opposite sign (negative θ) were observed for 48.8% of multigenic modules. For the 5,856 modules encompassing more than one EAP (whether from the same or different genes), we combined constituent association patterns in a consensus EAP representing the module (see M&M). We developed a web browser to visualize the activity of the regulatory modules across the genome and cell types (SFig. 4; <https://tools.giga.uliege.be/cedar/publihpq>).

Single-cell RNA Seq analysis of intestinal biopsies reveals 35,010 eQTL affecting 3007 genes and clustering in 3337 modules

We collected biopsies in the terminal ileum (IL), transverse colon (TC), and rectum (RE) (locations) of 60 healthy individuals. Epithelium and lamina propria (fractions) were separated, location- and fraction-specific cell suspensions hash-tagged, and subjected to scRNA-Seq using a 10X Chromium platform and Illumina sequencers. We obtained quality-filtered sequence data for a total of 293,801 cells from 57 individuals (5154 cells per individual on average). The number of reads per cell averaged 48,628, the number of unique molecular identifiers (UMI) per cell 7422, and the number of genes detected per cell 2,035 (Supplementary Data 10). Cells were assigned to one of nine sets corresponding to anatomical location (IL, TC, RE) and cell category (epithelial, immune, stromal) (see M&M). K-means clustering of the cells, by set, yielded a total of 276 clusters. Samples were merged using Harmony²¹, and a hierarchical tree (of clusters) constructed using the Euclidean distances between the clusters’ centroids in Harmony space. Leaves (corresponding to clusters in the original nine cell sets) and nodes in the tree were assigned to 13 epithelial, 14 lymphoid, 6 myeloid, and 10 stromal cell types (43 cell types in total) using cell-type specific gene signatures from the literature^{22–29} (Fig. 2; SFig. 5 and 6; and Supplementary Data 11). Numbers of cells were relatively evenly distributed across locations (Fig. 2B). Epithelial cells were more abundant than lymphoid, myeloid and stromal fractions combined (Fig. 2C). Myeloid, lymphoid and endothelial cells from the three locations overlapped well in UMAP space, while absorptive and secretory epithelial cells as well as fibroblast from distinct locations did

not, supporting larger effects of location on the transcriptome for the latter (Fig. 2E). Paneth cells were exclusively observed in ileal samples as expected, while most other cell types were present in the three locations.

Epithelial, stromal, myeloid, and lymphoid cell types clustered in the tree as expected (except for glia and mast cells) (Fig. 2F). The segregation of ileal, colonic and rectal clusters occurred mostly at terminal branches of the tree, with the exception of absorptive epithelium, in agreement with the overlap in UMAP space (SFig. 6J).

We conducted eQTL analyses by leaf and node across the tree (551 analyses). Analyses were performed using the same custom-made pipeline including QTLtools, a pseudo-bulk approach, and leaf/node-specific PEER factors³⁰ to correct for hidden confounders (including variable cell-type proportions) (SFig. 5). We detected 35,010 *cis*-eQTL (within leaf/node FDR ≤ 0.05) affecting 3007 e-genes (Supplementary Data 12). The number of eQTL detected per leaf/node was largely determined by the number of cells in the leaf/node with, however, a higher yield per cell for epithelial than for other categories. It plateaued at ~1100 presumably limited by sample size (57 individuals) (Fig. 3A). A second independent effect was detected for 259 *cis*-eQTL, and a third for two (Supplementary Data 12). As for circulating immune populations, we merged *cis*-eQTL in 3345 gene-specific modules when sharing similar EAP ($|\theta| \geq 0.6$), and augmented modules with 22,904 tier-2 eQTL that matched at least one significant eQTL in the module (see M&M) (Supplementary Data 13). We assigned modules to the most recent common ancestor (MRCA) of all active nodes/leaves. Most modules mapped towards the root of the tree, as expected as this is where the number of cells per node is largest and hence detection power highest (Fig. 3B). However, the numbers of leaves/nodes in which a module was active was over-dispersed: modules tended to be either active in fewer leaves/nodes or in more leaves/nodes than expected assuming random assortment ($p < 0.001$; see M&M) (Fig. 3C), reminiscent of circulating immune cells (Fig. 1A). This suggests that numerous cell-type specific eQTL also exist in the gut. Accordingly, we observed 524 modules that were only active in one of the 43 cell-types, of which 429 were also location-specific. The latter were mainly eQTL that were active either in enterocytes or their precursors from the small intestine (SI = IL) or in colonocytes or their precursors from the large intestine (LI = TC + RE), respectively (Fig. 3D). We developed a 3D application to visualize the activity of eQTL on a UMAP, vividly illustrating the location- and cell-type-specific activity of some modules (Fig. 3E). For gene-specific modules active in more than one leaf/node, the sign of the eQTL effect was the same in all leaves/nodes for 99.2% of the modules. For 27 genes, however, the effect differed depending on the leaf/node. In a number of instances, this clearly corresponded to a

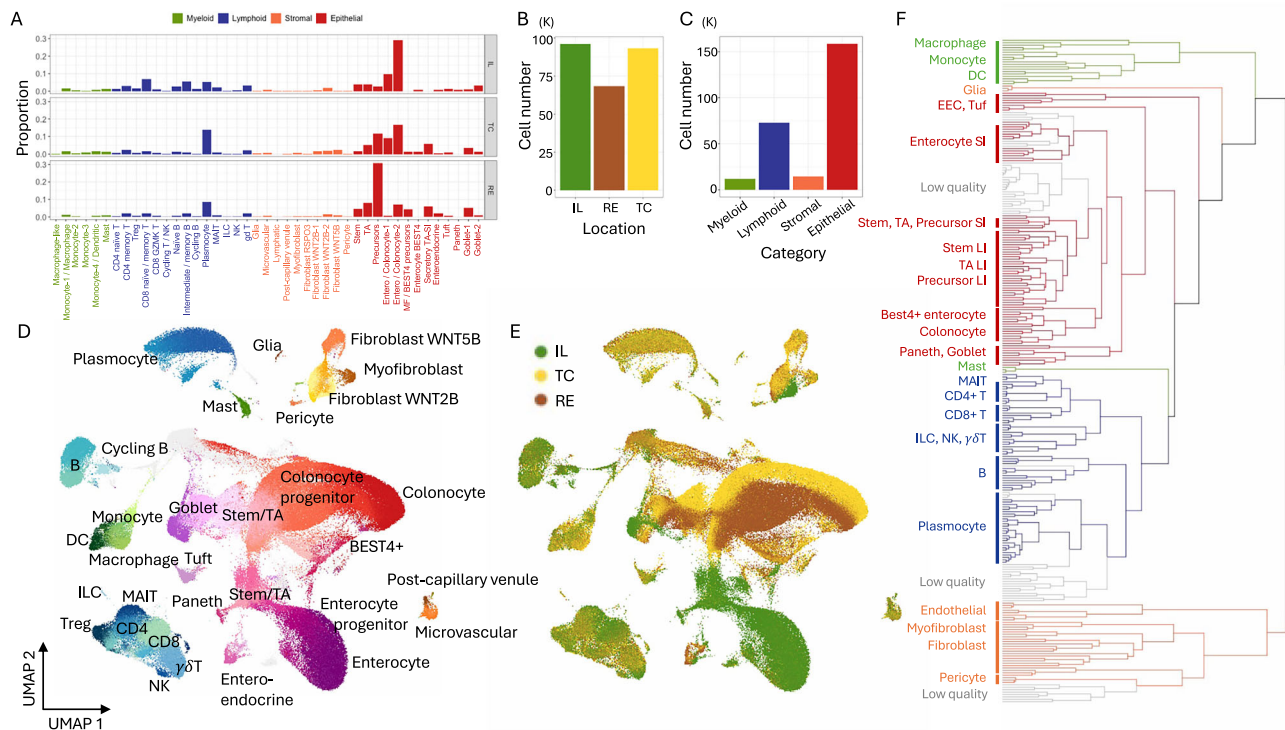


Fig. 2 | Single-cell RNA-seq analysis of the gut. A Proportion of cells from ileal (IL), colonic (TC), and rectal (RE) biopsies assigned to 6 myeloid cell types (green), 14 lymphoid cell types (blue), 10 stromal cell types (orange) and 13 epithelial cell types (red). **B** Number of recovered quality-controlled (QC-ed) cells by anatomical location. **C** Number of recovered QC-ed cells by cell category. Colors are as in (A). **D** UMAP of 293,801 QC-ed cells labeled by cell category (myeloid, lymphoid,

stromal, epithelial) and cell-type within category. **E** Same UMAP as in (D), labeled by anatomical location (colors as in B). **F** Hierarchical tree of 276 cell clusters (with 275 nodes). Four cell categories (myeloid, lymphoid, stromal, epithelial) are color-coded as in (A, C). The main cell types within categories are labeled. Source data are provided as a Source data file.

distinct effect depending on cell type (Fig. 3F and Supplementary Data 13).

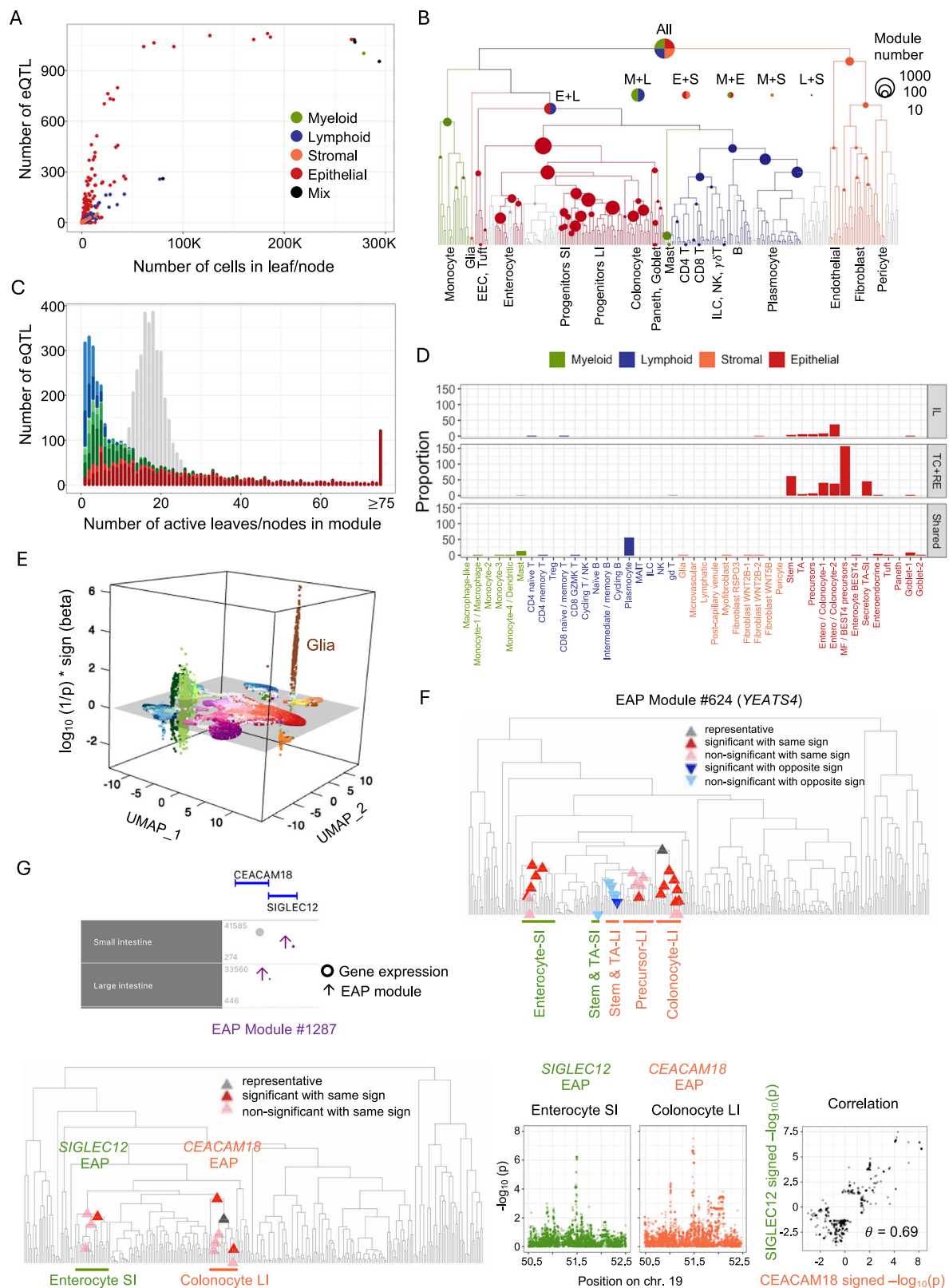
We then merged intestinal gene-specific modules characterized by similar EAP as measured by θ . Across-gene modules were augmented with 24,333 tier-2 eQTLs (see M&M) recruiting 715 extra genes. This yielded 666 modules encompassing EAP from more than one gene (21.6%), and 2415 modules that remained monogenic. The number of genes in multigenic modules averaged 2.8, ranging from 2 to 16 (Supplementary Data 14). The proportion of multigenic modules encompassing eQTL effects with opposite sign was 53%. We observed a number of instances where distinct genes were controlled by the same variants (i.e., were assigned to the same regulatory module) but in distinct cell types (Fig. 3G). All intestinal eQTL/module information in their genomic context is browsable using the Cedar2 website (SFig. 7; <https://tools.giga.uliege.be/cedar/publihpq>).

Merging blood and intestinal cis-eQTL modules reveals eQTLs that are specific for gut-resident immune cells

We then merged EAP in regulatory modules for blood and biopsies combined (Supplementary Data 15 and 16). This yielded 24,745 across-gene modules, of which 8472 were active in more than one cell type. Of the latter, 2172 were found to be active in both blood and biopsies (Fig. 4A). These included 310,783 links ($|\theta| \geq 0.6$) between a significant blood eQTL and a gut eQTL (of which 68% significant and 32% not), and 224,758 links between a significant gut eQTL and a blood eQTL (of which 95% significant and 5% not). Permuting the gut data reduces the number of links with a significant blood eQTL to 7744 (corresponding to 2.5% of 310,783). Permuting the blood data reduces the number of links with a significant gut eQTL to 3342 (corresponding to 1.5% of 224,758). Amongst the 2172 modules active in both blood and biopsies, there was a significant excess of modules that were active in multiple

cell types in both blood and biopsies ($p = 1.2 \times 10^{-5}$), as well as modules that were lymphoid-specific in both blood and biopsies ($p = 5.7 \times 10^{-4}$). Modules that were active in multiple cell types in the blood had more chance to be detected in biopsies than cell-type-specific blood modules ($p = 3.2 \times 10^{-54}$), while modules that were enterocyte-specific in biopsies had less chance to be detected in blood ($p < 1.6 \times 10^{-11}$), all as expected (Fig. 4B and Supplementary Data 17).

We mined the gene-specific catalogue for modules that were labeled lymphoid- or myeloid-specific in biopsies (two cell type categories that are also present in blood), but were “Not detected” in blood (despite the higher statistical power in blood versus gut samples). We reasoned that such a list might be enriched in *cis*-eQTL that would not be active in circulating immune cell population(s), but would reveal themselves in the same immune cell population(s) once becoming gut-resident. Hundred-thirty-two modules matched this pattern, of which 97 assigned to the lymphoid lineage and 35 to the myeloid lineage in the gut (Supplementary Data 18). Fourteen of the latter were mastocyte-specific eQTL, a cell type which was not represented in our blood collection, hence explaining why these were only found in the gut. For the remaining 118, we compared the intestinal EAP with that of the cognate circulating immune cells. The number and type of circulating immune cells considered for comparison depended on the cell type to which the module was assigned in the gut (Supplementary Data 16). For example, if a module was assigned to B cells in the gut, its EAP would be compared to naïve B cells, memory B cells and plasmacytes in the blood, while if a module was assigned to plasmacytes in the gut, it would only be compared to plasmacytes in the blood. The number of comparisons per gut EAP averaged 4.7 (range: 1–18). Scenarios 1, 2 and 3 accounted for 35.1, 56.7 and 8.2% of comparisons in the myeloid lineage, and 20.2, 56.4 and 23.4% in the lymphoid lineage, and these proportions did differ significantly ($p = 0.0002$). Moreover, both



differed very significantly from the equivalent proportions in blood ($p_{\text{myel}} = 3.8 \times 10^{-5}$, $p_{\text{lymph}} < 10^{-16}$). Differences were primarily driven by a higher proportion of scenario 1 (induction of gene expression), for gut-specific eQTL in the myeloid lineage, and a higher proportion of scenario 3 (module switch) for gut-specific eQTL in the lymphoid lineage (Fig. 4C).

Allele-specific mapping bias could have generated false positive compartment-specific eQTL. We excluded this for all but two genes (*LGALS9* and *COL16A1*), by applying WASP³¹ to the intestinal data (Sfig. 8).

Thus, it appears that there not only exist many cell-type-specific eQTL, but that—for a given cell type—eQTL can be context-specific, f.i. manifest in some anatomical compartments (resident) but not in

Fig. 3 | Gut regulatory eQTL modules. **A** Number of eQTL detected as a function of the number of cells in the corresponding leaf/node. Leaves/nodes are colored by category (myeloid: green, lymphoid: blue, stromal: orange, epithelial: red, mix (= multiple categories): black). **B** Assignment of 3345 gene-specific regulatory modules to the MRCA of all active leaves/nodes. Leaves/nodes are colored by cell-type category as in A. The surface of the circles is proportionate to the log of the number of modules assigned to the leaf/node. Modules assigned near the tree's root were assigned to pairs of categories when possible, corresponding to the bisected circles. **C** X-axis: number of active leaves/nodes in modules. Y-axis: number of observations. The modules are color-coded according to the leaf/node to which they were assigned. Modules assigned to the root are in dark red, modules assigned to a pair of cell-type categories are in red, modules assigned to one cell-type category are in green (shades of green (from dark to light) correspond to increasing levels of cell-type specificity in the category), modules assigned to a specific anatomical location are shown in blue. The grey distribution was obtained by randomly permuting activity status across modules, yet keeping the number of significant eQTL per leaf/node (see M&M). **D** Number of modules that are specific

for one of the 43 most granular cell-types, whether location specific (IL versus TC + RE) or shared across locations. Cell categories are labeled as before.

E Example of a highly cell type-specific eQTL (*GFR2* in glia (brown)). The x- and y-axes correspond to UMAP 1 & 2 axes, while the z-axis measures the strength of the association ($\log(1/p)$ multiplied by the sign of β). **F** Example of a gene-specific regulatory module (gene: *YEATS4*) for which the sign of the eQTL effects differs between cell types. The module is primarily active in absorptive intestinal epithelium in small (SI = IL, green) and large (LI = TC + RE, orange) intestine. The sign of the eQTL effect switches upon transition from stem/TA cells to precursor enterocytes in both SI and LI. **G** Example of two adjacent genes that are controlled by the same *cis*-acting regulatory module yet in different cell types: *SIGLEC12* in enterocytes of the small intestine (SI: green) and *CEACAM18* in enterocytes of the large intestine (LI: orange). The positions on the tree where the RM regulates the corresponding genes are shown as triangles (left panel). The corresponding EAPs and theta-plot are shown (right panels). P-values are from eQTL analyses computed as described in Methods; they are nominal and two-sided. Source data are provided as a Source data file.

others (circulating). *CCL20* and *CCL24* constitute two interesting such examples (Fig. 4D–G).

To gain additional insights in what may underpin compartment-specific eQTL, we generated scRNA-Seq data using the same chemistry (GEM-X Flex gene expression, 10X Genomics) for whole blood and intestinal biopsies (IL, TC and RE) from a single healthy individual. GEM-X Flex interrogates the transcriptome of formalin-fixed samples, and is assumed to better reflect native cell type proportions by freezing biological state and preserving fragile cells. We obtained QC-ed data for a total of 23,261 cells (blood: 6883, IL: 6909, TC: 4374, RE: 5095). The four datasets were integrated into one object with Harmony²¹, clustered and annotated using Seurat²⁴ (SFig. 9A). Comparing the relative abundance of the different immune cell populations between compartments revealed a number of noteworthy features (SFig. 9B): (i) granulocytes amount to a much higher proportion of immune cells in the blood (~57%) than in the gut where they exhibit a rostro-caudal gradient (IL~2%, TR~6%, RE~11%); they are primarily composed of neutrophils in the blood (~98%) and of mastocytes in the gut (~93%); (ii) monocytes account for a higher proportion of immune cells in the blood (~16%), where they are mainly composed of CD14⁺ monocytes (~88%), than in the gut (~6%), where they are mainly composed of macrophages and dendritic cells (~83%), (iii) B lymphocytes account for a higher proportion of immune cells in the gut (~44%) than in the blood (~2.4%); plasmocytes account for ~86% of B lymphocytes in LI (= TC + RE), for ~14% in SI (= IL), and ~9% in blood; (iv) T lymphocytes are more abundant in the SI (~53%) than in the LI (~37%) and in the blood (~25%). We then examined the reasons underlying the compartment-specificity of the eQTL referred to in Supplementary Data 18. For the eQTL that were active in myeloid cells in the gut but not in the blood, the main reasons were (i) mastocyte-specific eQTL, a cell type that was not isolated and analyzed per se in the blood, and (ii) macrophage- or dendritic cell-specific eQTL, cell types that derive from monocytes upon migration in the gut, yet are absent in the blood. The latter differentiation process appears to be accompanied by the induction of new eQTL whether due to induction of gene expression (scenario 1), of novel gene regulation (scenario 2), or of module switch (scenario 3) (see above). For the eQTL that were active in lymphoid cells in the gut but not in blood, we did not observe similar shifts in cell identity. This suggests that upon migration into the gut, the transcriptome of T and B lymphocytes responds to a new microenvironment, which reveals new eQTL, yet without causing marked changes in cell identity (SFig. 9C). Of note, a large proportion of the lymphoid gut-specific eQTL were detected in plasmocytes.

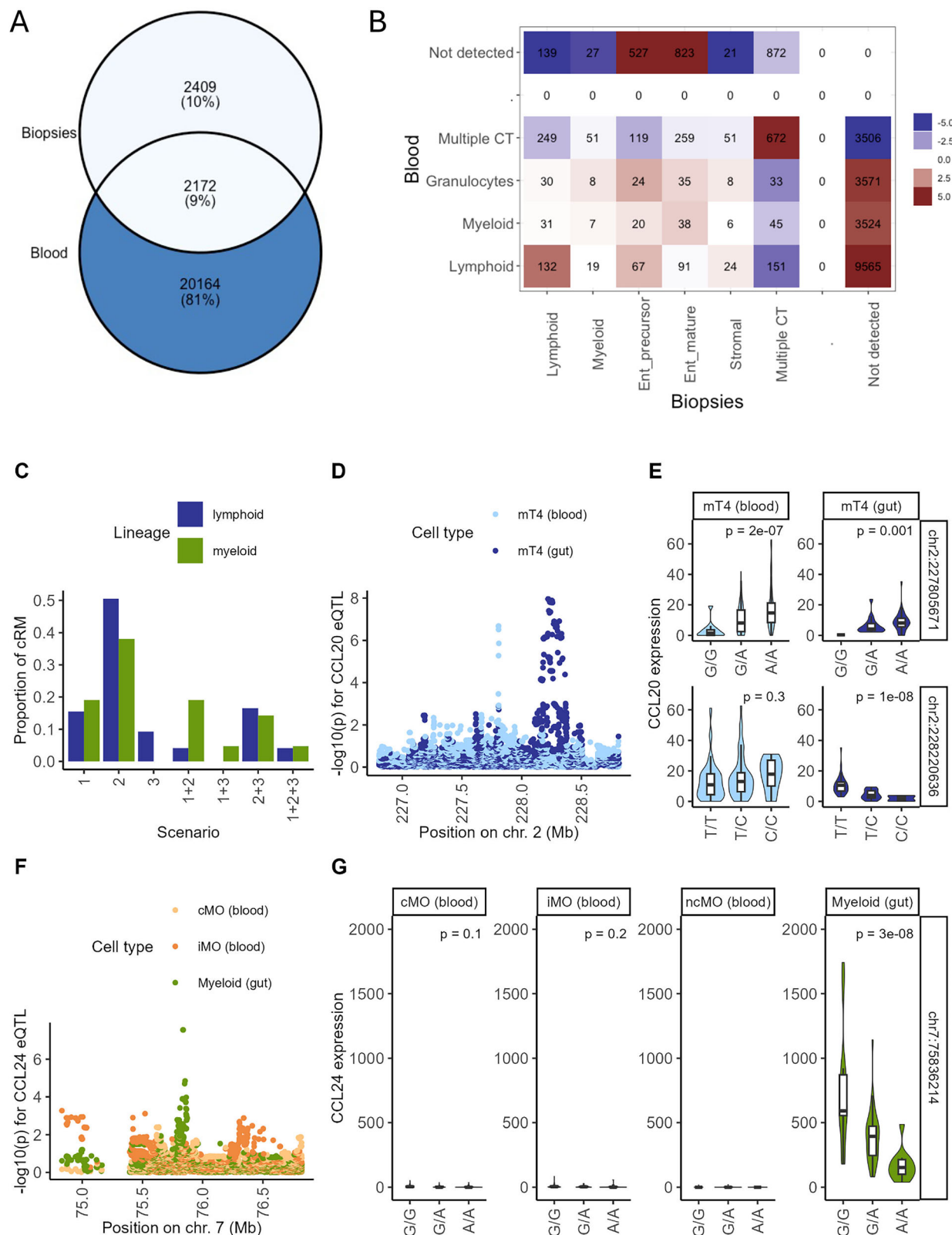
Identifying *cis*-eQTL driving inherited predisposition to IBD

We then mined our database of blood and intestinal *cis*-eQTL for EAP matching DAP for IBD. We defined the boundaries of 206 risk loci

reported by de Lange et al.³², including 173 IBD loci (considering CD and UC patients jointly in GWAS), 157 CD loci and 125 UC loci, by visual examination of the local association patterns obtained with (Europeans-only) ~25 K cases and ~35 K controls (Supplementary Data 19). Thirty-two risk loci that encompass composite peaks were further subdivided in sub-risk loci, to be confronted separately to *cis*-eQTL in addition to the complete risk locus. It is indeed conceivable that larger, multi-peak DAP reflect the compound effects of multiple *cis*-eQTL either of the same or different genes in the same or different cell types. Splitting the corresponding risk loci may reveal distinct matching eQTL. Colocalization analyses were conducted using the θ metric, as described⁴⁵. To define suitable thresholds for significance, we performed parallel analyses using permuted genotype data, separately for blood and biopsies (genome-wide *cis*-eQTL analysis in all cell types, leaves and nodes). Confronting results obtained with the real versus permuted data allowed us to compute an FDR for each DAP-EAP pair as a function of the value of $|\theta|$ (≥ 0.6), its *p*-value, as well as the *p*-value for the eQTL (see M&M). This yielded matching DAP-EAP with FDR ≤ 0.05 for 379 genes in 119 risk loci (tier-1), or 556 genes in 140 risk loci when considering matching DAP-EAP with FDR ≤ 0.10 (tier-1 + 2) (Fig. 5A and Supplementary Data 20–22). No credible extra DAP-EAP matches were detected when using the modules' consensus EAP.

The numbers above correspond to ~51% of the 173 IBD risk loci, ~51% of the 157 CD risk loci and ~49% of the 125 UC risk loci with at least one DAP-EAP match with FDR ≤ 0.10 , and to ~1/375 (IBD), ~1/341 (CD) and ~1/621 (UC) of realized DAP-EAP confrontations resulting in a match (FDR ≤ 0.10). We repeated the same analysis for 58 risk loci for dilated cardiomyopathy (DCM) uncovered in large European case-control cohorts³³. Risk for DCM is a priori less likely (than risk for IBD) to be influenced by perturbation of gene expression in blood and gut. Thus, the number of matches between DCM DAP and blood/gut EAP should be lower than for IBD. Indeed, applying the exact same procedure to DCM, detected matching DAP-EAP for ~17% of DCM risk loci, corresponding to one match every ~1923 realized DAP-EAP confrontations, i.e., approximately five times less compared to IBD, CD and UC ($p \leq 1.4 \times 10^{-14}$). Moreover, the number of realized DAP-EAP confrontations per Kb of risk loci was 3.1 for DCM versus ~4.6 for IBD, indicating that fewer genes in DCM risk loci than in IBD risk loci are expressed in blood and/or gut. Taken together, this strongly supports the biological relevance of the DAP-EAP matches for IBD, CD and UC (Supplementary Data 23).

Matching EAP were detected in both blood and gut for 77 (55%) risk loci, only in blood for 33 (24%), and only in the gut for 30 (21%) (SFig. 10A). The number of DAP-matching e-genes averaged 4 per risk locus, ranging from 1 to 34 (i.e., for the 140 risk loci with at least one match). Local gene density explained only ~25% of the differences in number of matching e-genes per risk locus. For risk loci with matching



DAP-EAP in both blood and gut, the e-genes involved were generally different. As a consequence, the number of e-genes matching DAP in both blood and gut was 57 (11%), only in blood 212 (41%), and only in gut 246 (48%) (SFig. 10B). Modules that were active only in the gut made the strongest contribution to matching DAP-EAP (Fig. 5B, C), despite the fact that they only account for ~10% of the modules (Fig. 4A). There was a strong correlation between the number of

matching e-genes in blood and gut for a given risk locus, despite the fact that the genes involved mostly differed (SFig. 10C, D). Within the gut, all cell types contributed approximately evenly to DAP-EAP matches, except enterocyte precursors for CD and UC, and myeloid cells for UC, which contributed less (Fig. 5B, C). In circulating immune cells, regulatory modules active in all cell types contributed most to DAP-EAP matches, reminiscent of a previous report¹⁵ (SFig. 10H, I). We note

Fig. 4 | Merging regulatory modules across blood and gut samples. **A** When merging all blood ($n = 60,113$) and gut ($n = 35,010$) eQTL jointly, we obtained 24,745 across-gene modules, including 8472 active in several cell types of which 2172 (9%) with eQTL from blood and gut. **B** Modules were assigned to cell types, separately for blood and gut, as described before. Blood cell types were grouped in lymphoid, monocytes/dendritic cells (myeloid), granulocytes, or multiple of these cell types. Intestinal cell types were grouped in lymphoid, myeloid, enterocyte precursors, mature enterocytes, stromal, or multiple of these cell types. “Not detected” indicates that the module is not active in the corresponding sample type (blood or gut). The numbers in the tiles correspond to the number of observations for the corresponding combinations. The colors correspond to $-\log(p)$ of a test of independence (i.e., to what extent do the observed numbers deviate from expectation assuming that the proportions of the different categories in blood and gut are independent; see Methods). Red: excess. Blue: depletion. **C** Molecular mechanisms underlying the compartment-specificity: scenario 1: no detectable expression in blood, scenario 2: detectable expression but no detectable eQTL effect in blood, scenario 3: a different regulatory module affects the gene in gut and blood (module switch). If the intestinal EAP was compared to EAPs in multiple cell types, multiple mechanism may co-exist, explaining the 1 + 2, 1 + 3, 2 + 3, and 1 + 2 + 3 categories. **D** Example of a gene (*CCL20*, C-C Motif Chemokine Ligand 20) that is expressed in circulating as well as gut-resident memory CD4 T lymphocytes (mT4). The gene is subject to two distinct eQTL in these compartments (light blue:

blood mT4 EAP, dark blue: gut mT4 EAP). The EAP observed in circulating cells matches a DAP for UC (Table 21). The corresponding EAP is also detectable in one intestinal mT4 leaf, which could correspond to blood present in the biopsy. **E** Violin plot showing the expression levels of *CCL20* in blood mT4 (left panels in light blue) and gut-resident mT4 (right panels in dark blue) for individuals sorted by genotype for the top SNP of the light blue blood EAP (upper panels; $n_{GG}^{blood} = 12$, $n_{GA}^{blood} = 65$, $n_{AA}^{blood} = 113$, $n_{GG}^{gut} = 1$, $n_{GA}^{gut} = 17$, $n_{AA}^{gut} = 38$) and the top SNP of the dark blue gut EAP (lower panels; $n_{GG}^{blood} = 117$, $n_{TC}^{blood} = 61$, $n_{CC}^{blood} = 12$, $n_{TT}^{gut} = 33$, $n_{TC}^{gut} = 19$, $n_{CC}^{gut} = 4$). Boxplots as in Fig. 1. **F** Example of a gene (*CCL24*) that is expressed in gut-resident myeloid cells (including monocytes and dendritic cells) but undetectable in the equivalent circulating cells (shown for the three types of monocytes: conventional (cMO), intermediate (iMO), and non-conventional (ncMO). *CCL24* is subject to an eQTL that is detectable in gut-resident myeloid cells (green EAP) but not detectable in circulating monocytes (yellowish EAP). **G** Violin plots showing the expression of the *CCL24* gene in circulating monocytes (panels on the left), and in gut-resident myeloid cells (panel on the right) for individuals sorted by genotype for the top SNP of the gut EAP in (**E**) ($n_{GG}^{blood} \geq 60$, $n_{GA}^{blood} \geq 86$, $n_{AA}^{blood} \geq 36$, $n_{GG}^{gut} = 13$, $n_{GA}^{gut} = 32$, $n_{AA}^{gut} = 12$). Boxplots as in Fig. 1. All reported p -values are from eQTL analyses computed as described in Methods; they are nominal and two-sided. Source data are provided as a Source data file.

the modest enrichment (1.3-fold) of modules active only in circulating natural killer (NK) cells. Only 12.5% of DAP matching e-genes were shared between the two diseases. This proportion only increased to 31.8% when restricting the analysis to the 25 risk loci associated with both CD and UC. This possibly underscores the distinct molecular determinism of the two pathologies (Sfig. 10E–G).

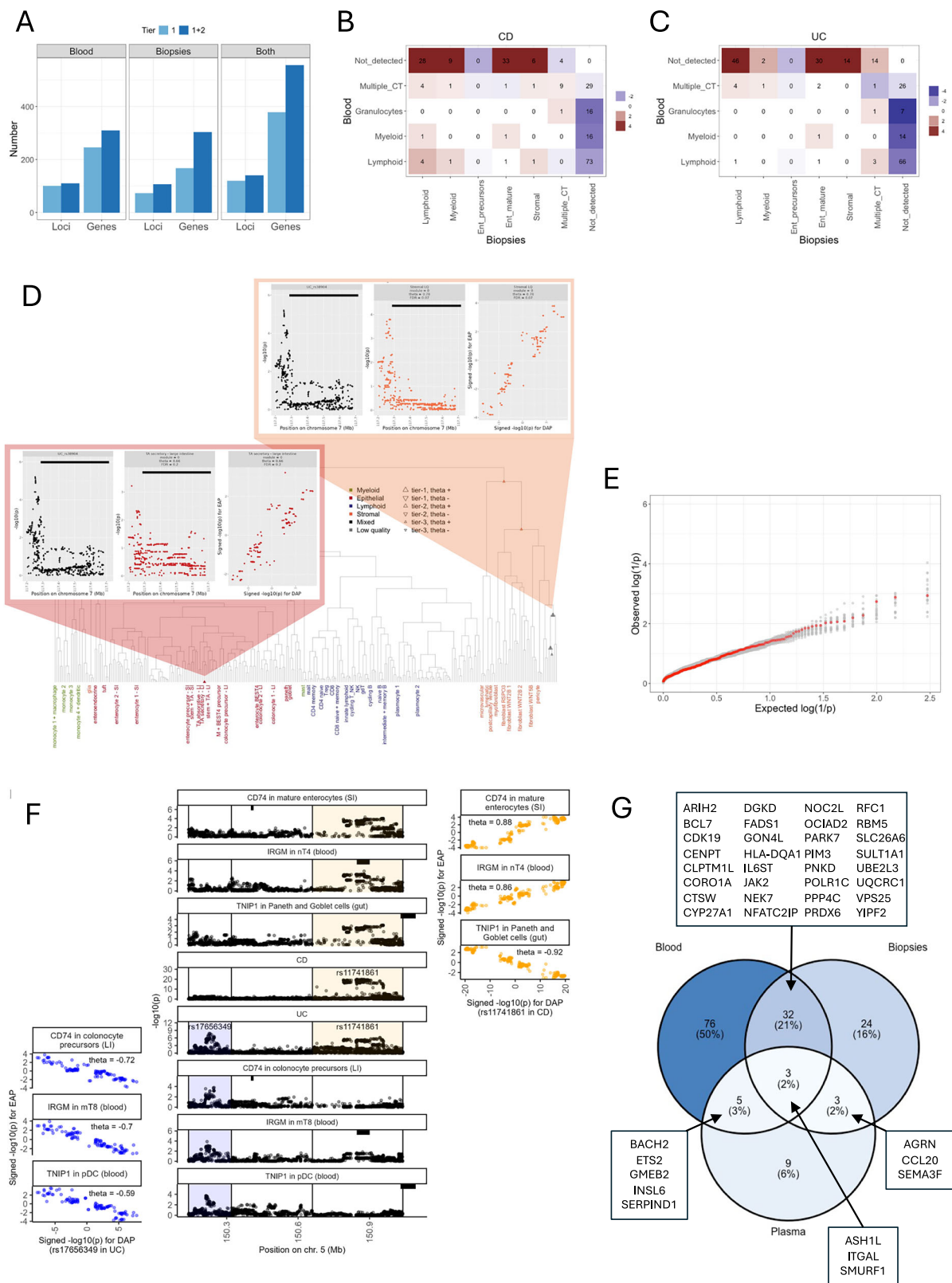
To the best of our knowledge, no eQTL-based connection with IBD was previously reported for 366 of the 556 genes (Supplementary Data 22). Reactome analysis conducted with the complete gene list highlighted four pathways: interferon gamma signaling (found entities: *CIITA*, *IRF5*, *IRF6*, *PTPN2* and *MHC class I* and *II* genes), interleukin-6 signaling (*IL6ST*, *IL6R* and *JAK2*), chemokines and their receptors (*CXCR1*, *CXCR2*, *CCR2*, *CCR6*, *CXCL2*, *CXCL5*, *CCL20*), and RUNX regulated immune response and cell migration (*ITGA4*, *ITGAL*) (Supplementary Data 24). We scanned the literature, for the 481 protein coding genes out of the list, for functional evidence (other than association or differential expression-based) regarding epithelial barrier function, innate or adaptive immunity that would be considered as support for causality if generated as follow-up of the GWAS and eQTL colocalization. We found such support for 221 of the 556 genes (Table 1 and Supplementary Data 25 and 26). This included six genes associated with monogenic forms of human inflammatory bowel disease (*BACH2*, *CARMIL2*, *PMVK*, *SLC9A3*, *TMEM50B* and *TNIP1*). We submitted the 481 DAP-matching protein-coding e-genes mixed with 481 non-DAP-matching control protein-coding e-genes blindly to two trained scientists, and asked them to scan the literature for evidence that genetic perturbation of the corresponding genes affects susceptibility to colitis in rodents. The first person detected 84 such cases in the DAP-matching e-genes versus 36 in the control e-genes ($p = 1.4 \times 10^{-6}$), while the second detected 97 cases in DAP-matching versus 42 in control e-genes ($p = 2.3 \times 10^{-7}$). This strongly suggests that our list of DAP-matching e-genes is indeed enriched in causative genes for IBD. After a final review of the literature, we kept 88 DAP-matching e-genes that upon genetic perturbation affect susceptibility to colitis in a rodent model (Supplementary Data 26). Of note, we observed a significant correlation ($p = 0.012$) between the best θ -value for a gene and the reporting of an effect of genetic perturbation of that gene on susceptibility to colitis in the mouse. Thus, the genes with the strongest colocalization signals for IBD had a higher probability to impact susceptibility to induced colitis upon perturbation in the mouse.

The number of genes with incriminating functional evidence averaged 1.46 per risk locus (across the 140 risk loci), with a maximum

of 15 for the rs3197999 locus on chromosome 3 (48.48–50.23 Mb). There were multiple risk loci with more than one strongly supported candidate gene. For example, the chr1:rs3180018 locus harbors the DAP-matching *IL6R* e-gene, member of the inflammatory cytokine pathway highlighted by the Reactome analysis, but also *PMVK* causing IBD-like manifestations in a compound heterozygote³⁴. Along similar lines, the chr5:rs17656349 locus harbors the DAP-matching *IRGM* gene, regulating autophagy and response to various pathogen-associated molecular patterns (PAMPs), and also *TNIP1* coding for the “TNFAIP3 interacting protein 1”, knowing that de novo mutations in *TNFAIP3* are associated with juvenile IBD^{35,36}. Similarly, the chr16:rs28449958 locus encompasses *CARMIL2* causing very early onset IBD^{37–39}, and also *SMPD3*, known to regulate TNF- α response in macrophages and B cells, and to influence the severity of DSS-induced colitis when modulated in mice^{40–42}. There was no correlation between the number of genes with supporting functional evidence in a risk locus and its odds ratio on disease.

Particularly noteworthy is the observation that variants increasing the expression of the cystic fibrosis-causing *CFTR* gene in stromal and/or epithelial precursor cells increase the risk for UC while possibly decreasing the risk for CD (Fig. 5B). This may be related to recent reports that loss-of-function coding variants in *CFTR* protect against CD⁴³. In addition, we found that variants affecting the expression of *PRKAA1* (shown to phosphorylate and modulate *CFTR* activity⁴⁴), *CDK19* (shown to control the *CFTR* pathway in the intestinal epithelium of mice⁴⁵), *ADCY3* and *PRKAR2A* (both linked to β_2 adrenergic-dependent *CFTR* expression⁴⁶) also affect IBD susceptibility, although these effects were often detected in cells other than the intestinal epithelium. We further observed that variants that decrease *SLC9A3* expression in enterocytes increase UC risk. Loss-of-function mutations in the *SLC9A3* sodium-proton antiporter cause congenital diarrhea 3 and 8⁴⁷.

Recently, the whole exome of ~30,000 IBD patients was sequenced, and rare-variant burden tests (MAF < 0.001) conducted for 11,978 genes¹¹. The distribution of burden test p -values for 298 of our DAP-matching e-genes with sequence information did not depart from expectations under the null (Fig. 5C). Nevertheless, our list of 556 includes *TAGAP*, one of nine genes with one associated rare coding variant identified in this study: a rare missense variant (E147K) protecting against CD (OR: 0.786). *TAGAP* has an EAP in enterocyte progenitors that matches the DAP of a CD risk locus on chromosome 6 (rs212388) with positive θ (0.79, FDR = 0.04), hence compatible with



the protective effect of the missense variant. Of note, borderline DAP-EAP hits were observed for two other genes in Sazanovs' list of nine: *CCR7* (memory CD4 in gut) and *RELA* (plasmacytes in gut). However, for both the positive sign of θ did, a priori, not match the increased risk associated with the reported missense variants. One gene, *ATG4C*, was further incriminated in this study based on a mutational burden attributed to three missense variants (suggestive signal). *ATG4C* was

not part of our list, but *ATG16L2*, a paralogue of the *ATG16L1* autophagy gene not previously incriminated in IBD, yielded a borderline signal with a convincing θ of 0.77 despite a modest eQTL signal ($p_{window\ adj} = 0.27$).

We noticed 10 instances where distinct, cell type-specific EAP from the same gene match DAP from IBD risk loci that were considered different albeit adjacent (*QPRT*, *EIF2B4*, *TNIP1*, *PRXL2B*, *SLC35E2B*, *IRGM*,

Fig. 5 | cis-QTL of blood and gut matching IBD risk loci. **A** Numbers of IBD risk loci and genes with matching DAP-EAP in the 27 circulating cell populations (blood), in the 43 intestinal populations (biopsies), and when combining both (Both). Tier 1 corresponds to matching DAP-EAP with $|\theta| \geq 0.6$ and $\text{FDR} \leq 0.05$. Tier 2 corresponds to matching DAP-EAP with $|\theta| \geq 0.6$ and $0.05 < \text{FDR} \leq 0.10$. **B, C** Activity profile, in intestinal biopsies and circulating immune cells (Blood), of gene-specific DAP-matching modules. Modules were assigned to one of five blood categories (Y axis) and one of seven gut categories (X axis). Numbers are shown for CD in the left **B** and UC in the right (**C**). We computed the $-\log_{10}(p)$ -value of the departure from expectation assuming that the distribution would be the same as for all gene-specific modules combined (DAP-matching and not-DAP-matching) using a goodness-of-fit test. The $-\log_{10}(p)$ -value is indicated using a red (excess) – blue (depletion) color gradient. **D** EAP of the cystic fibrosis *CFTR* gene matching the DAP of a UC risk locus (rs38904) in secretory TA precursor cells (red) and intestinal stromal cells (orange). The nodes/leaves with DAP-matching EAP are marked by triangles (large: $\text{FDR} < 0.10$, small: $\text{FDR} > 0.10$). Insets: (left) DAP for UC, (middle) EAP for *CFTR*, (right) θ plot, (red) secretory TA precursor cells, (orange) stromal cells. Boundaries of the *CFTR* gene are marked by the horizontal black line. The positive sign of θ indicates that downregulation of the *CFTR* gene in these cell types may be protective, which is corroborating the recent finding that *CFTR* loss-of-function mutations protect against CD⁴³. All reported p -values are from eQTL

analyses computed as described in Methods; they are nominal and two-sided. **E** QQ plot generated with the burden test p -values obtained for 298 of our 556 DAP matching e-genes by Sazonovs et al.¹¹ (red dots). Grey dots correspond to QQ plots obtained with random sets of 298 genes from the list of genes with data in [11]. **F** Central panel: DAP for CD and UC in two adjacent risk loci on chromosome 5 (rs17656349 shaded in blue, and rs11741861 shaded in yellow), as well as (below) EAP matching the rs17656349 UC DAP for *CD74* (colonocyte precursors in large intestine), *IRGM* (circulating memory CD8), and *TNIP1* (circulating plasmacytoid DC), and (above) EAP matching the rs11741861 CD/UC DAP for *CD74* (mature enterocytes in small intestine), *IRGM* (circulating naïve CD4), and *TNIP1* (paneth and goblet cells). The genomic position of the corresponding genes are marked by black horizontal bars. The corresponding theta plots are shown on the left (rs17656349 in blue) and right (rs11741861 in yellow), respectively. All reported p -values are from eQTL analyses computed as described in Methods; they are nominal and two-sided. **G** DAP matching e-genes whose expression levels are affected by IBD in a direction that is consistent with the effect of risk variants (i.e., both risk variant and disease increase expression, or both risk variant and disease decrease expression) in one or more of 27 circulating immune populations (Blood), in intestinal biopsies²⁸, or in plasma⁴⁸. Source data are provided as a Source data file.

CDK11B, *CD74*, *SLC25A15*, *ZNF589*). For example, an EAP for *IRGM* in memory CD8 cells matches the UC DAP in risk locus chr5:rs17656349, while a distinct *IRGM* EAP in naïve CD4 cells matches the CD DAP in the adjacent chr5:rs11741861 locus. In the same risk locus, a *CD74* EAP in colonocyte precursors of the large intestine matches the UC DAP in the chr5:rs17656349 risk locus, while a distinct *CD74* EAP in mature enterocytes of the small intestine matches the CD DAP in the adjacent chr5:rs11741861 locus (Fig. 5D). Similar observations were made for multiple sub-risk loci as defined above (Supplementary Data 22). Thus, complex “composite” disease association patterns may reflect the effect of distinct risk variants that perturb the expression of the same gene in different cell types, with possibly distinct effects on disease. Also, the DAP for CD and UC, even when overlapping and considered as the same risk locus, may differ and match distinct EAP. For example, the CD DAP in risk locus chr6:rs1819333 matches an EAP for *RNASET2* that is detected in nearly all circulating immune populations, while the distinct UC DAP in the same risk locus matches a *CCR6* EAP active only in NKT cells.

We also observed at least 10 e-genes with DAP-matching EAPs in multiple cell types, for which the sign of θ differed between cell types either for the same (CD: *FADS1*, *IL18R1*; UC: *TOM1*; IBD: *PRXL2B*, *UBE2L3*) or for different forms of IBD (*CD74*, *IL1R2*, *IRGM*, *PRXL2B*, *CD244* (= *SLAMF4*), *SLC25A15*) (f.i. Fig. 5F and Supplementary Data 26). This calls for caution when defining the desired effect of a drug (activator or inhibitor) based on the sign of θ .

The CEDAR2 website [<https://tools.giga.uliege.be/cedar/publihpq>] allows for convenient visual inspection of the matching DAP-EAP patterns underpinning our analyses.

Entrectinib, a small molecule NEK7 inhibitor, is a repurposing candidate for IBD

It is not obvious that pharmacological targeting of causative genes (that upon perturbation by common regulatory variants affect the chance to develop the disease), will succeed in reversing the disease process once initiated, i.e., be curative. Additional prioritization of candidate genes whose expression is also perturbed by the disease process itself may be useful. To that end, we collected blood from 55 active CD patients, performed RNA-Seq on the same 27 fractionated circulating immune cell populations, and performed differential expression analysis between cases and controls (Supplementary Data 27). We additionally consulted lists of genes that were shown, from scRNA-Seq data, to be differentially expressed between intestinal biopsies of IBD cases and controls²⁸, as well as lists of proteins that

were differentially abundant in plasma of IBD cases and controls⁴⁸. The expression of 109 of our 556 e-genes differed in cases in a manner that was consistent with the sign of θ (i.e., both risk variant and disease increase expression, or both risk variant and disease decrease expression) in one of the three data sets (circulating immune populations, biopsies or plasma), of 40 e-genes in two of the three data sets, and of three e-genes in the three data sets (Fig. 5G).

One hundred and eighty-one drugs targeting 179 e-genes are or have been used/tested to treat IBD^{49,50} (Supplementary Data 28). Eleven of these overlapped with our list of 556 e-genes: *CXCR2* (elubrixin: abandoned for lack of efficacy), *IL6R* (TJ301: phase 2 ongoing), *IL6ST* (olamkicept: phase 2 completed with positive results), *IMPDH2* (mycophenolate: abandoned for undocumented reasons), *ITGA4* (multiple in phase 2 and two approved including vedolizumab), *ITGAL* (efalizumab: abandoned for safety issues including multifocal leukoencephalopathy), *JAK2* (multiple with positive results after phase 2 for peficitinib), *NDUFAF1* (metformin: phase 2 ongoing), *PDCD1* (rosnilimab: phase 2 ongoing), *TEC* (ritlectinib: phase 2 completed with positive results for CD and UC) and *TNFSF15* (several: phase 2 ongoing) (Supplementary Data 22). For seven of these (*IL6R*, *IL6ST*, *ITGA4*, *ITGAL*, *JAK2*, *IMPDH2*, *PDCD1*), the effect (presumed activator vs inhibitor) of at least some of the drugs was in agreement with the sign of θ . For six (*CXCR2*, *IL6ST*, *IL6R*, *ITGAL*, *JAK2*, *NDUFAF1*), the effect of disease on expression/abundance level was compatible with the effect of the drug (Supplementary Data 22).

We further identified 45 genes in our list of candidates, targeted by known drugs that—to the best of our knowledge—have not been tested in the context of IBD (Supplementary Data 22). For 22 of these, the activity of at least one drug (likely inhibitor or activator) was consistent with the sign of θ . The corresponding drugs were in phase 1 for three (*PIM3*, *PRKDC*, *RPS6KBI*), phase 2 for ten (*ATP2A1*, *CDK11B*, *CDK19*, *ERAP2*, *HLA-DRB1*, *IL18R1*, *IL18RAP*, *KIR2DL1*, *OPRL1*, *PPP5C*), phase 3 for two (*INPP5D*, *PRKAA1*), and approved for at least one disease other than IBD for seven (*CD19*, *CFTR*, *CYP3A5*, *LAMA2*, *NEK7*, *PTGIR*, *TUBB4B*). For nine of the 22 genes (underlined), expression was affected by the disease process in a manner consistent with θ , at least in one of the three datasets. A mouse model of colitis exists for four of the 22 (bold) (Supplementary Data 22). Detailed examination of the drugs that were in phase 3 or higher (Supplemental Material 1), identifies entrectinib (ENB), in particular, as an intriguing repurposing candidate for IBD. Entrectinib (ENB) is a potent tyrosine multikinase small-molecule inhibitor that targets the NTRK, ROS1 and ALK oncogenes, approved by the FDA for the treatment of various tumors⁵¹. It

Table 1 | Biological functions of IBD-matching e-genes

Function		Genes
Epithelial homeostasis and barrier function	Epithelial transporters	ABCC2, ADCY3, CDK19, CFTR, MFSD4B, <i>PRKAA1</i> , <i>PRKAR2A</i> , SATB2, SLC26A6, SLC5A6, SLC9A3
	Epithelial integrity and repair	CAMK2g, CD74, CDH3, CERS2, CTSK, <i>ERBB2</i> , HSPD1, INAVA, IRF6, LASP1, MASTL, PARD6A, PEX6, PLAUI, <i>PRKAA1</i> , RAB13, TKT, ZEB2-AS1
	Epithelial stem cells and development	CDX1, HSPG2, LAMA2, NRBP1, SETDB1, SRCAP
Signaling in innate immunity	Pathogen-recognition receptors	ANKRD17-DT, ARIH2, <i>ATG16L1</i> , CARD9, CYLD, <i>FUT2</i> , IL1R1, IL1R2, IL18R1, IL18RAP, INAVA, IPMK, IRF5, LACC1, LAMA2, NEK7, ORMDL3, PRKDC, <i>RFTN2</i> , RNASET2, <i>RNF123</i> , <i>SEC24C</i> , SSC5D, SENP7, <i>TRAF3IP3</i> , <i>TRAIP</i> , TRIM37, IRGM, UBR2, <i>USP19</i> , ZC3H15, ZDHHC11B, ZFP91
	TNF signaling	ADCY7, BABAM2, LTBR, MST1, MST1R, RASSF1, TNFRSF14, TNFSF4, TNFRSF9, TNFSF15, <i>TRAIP</i> , ZFP36L2
	JAK-STAT signaling	FAM220A, JAK2, OCIAD2, <i>PRKAR2A</i> , PTPN2, TYK2
	IL6, SLAMF, Others	ASH1L, BRD7, DGKD, EDC4, GRK6, IL26, IL6R, IL6ST, Ly9, OSM, <i>RASGRP1</i> , <i>SLAMF1</i> , <i>SLAMF4</i>
Cells in innate immunity	Leucocyte extravasation and trafficking	CDH26, <i>CORO1A</i> , CXCR1, CXCR2, CXCL2, CXCL5, GCA, ITGA4, ITGAL, LSP1, NEU4, SEMA3F, TLN1, TMEM87A, TSPAN17
	Macrophages	AUH, CCR2, ETS2, LACC1, PPM1F, <i>PRKAA1</i> , PROK2, QPCTL, SLC11A1, SLC26A6, <i>SP140</i>
	Other or multiple myeloid cells	ARNT, FOXP1-AS1, HINT1, <i>INPP5D</i> , KIR3DL1, KIR3DL3, PLPP6, <i>PPP3CB</i> , <i>PRKAA1</i>
Autophagy and ubiquitin-dependent processes	Autophagy	<i>ATG16L1</i> , CDK11B, CLTC, <i>CORO1A</i> , CUL1, DAP, IPMK, IRGM, NR1D1, PCNP, PIM3, <i>PPP3CB</i> , <i>PRKAA1</i> , PTPN23, SHISA5, TFAM, TUFM, <i>USP19</i> , VPS37C
	Ubiquitin-dependent processes	BBC3, CSID1, CUL2, FBXL20, KBTBD6, LTBR, MIB2, OTUD7B, <i>PARK7</i> , PLA2G15, QRICH1, RFT1, <i>RNF123</i> , <i>SEC24C</i> , SMURF1, TOM1, UBE2L3
Other inflammation mediators	Prostaglandins and glucocorticoids	FADS1, FADS2, HSD11B1, NR5A2, <i>PRKAA1</i> , PTGIR, PRXL2B, <i>PRKAR2A</i>
	Complement system	CD46, CR2, GNA12
Antigen presentation and MHC		CD74, CIITA, CTSS, ERAP2, LNPEP, RFX5, (+MHC)
Signalling in adaptive immunity	BCR &/or TCR	CD19, LAT, LIME1, PTPRC, PTPRH, <i>RFTN2</i> , SLAMF8, THEMIS, <i>TRAF3IP3</i>
	CCR6, Others	ADCY7, CCL20, CCR6, DGKD, DUSP8, ELF3, Ly9, <i>RASGRP1</i> , SH2B3, <i>SLAMF4</i>
Cells in adaptive immunity	T cells	ATXN2L, BCL2L12, CCR2, <i>CORO1A</i> , CREM, CTSK, CTSW, FOSL2, GMEB2, GSDMB, KSR1, MED1, NFATC2IP, PDCD1, PLXNB1, <i>PPP3CB</i> , PPP4C, <i>PRKAA1</i> , RASGRP2, SEMA3G, TAGAP, TEC, TKT, USP4, ZFP90, ZFP91
	B cells and multiple	BACH2, BRWD1, HHEX, IKZF3, <i>INPP5D</i>
Other	Other mechanisms	COQ10B, CYP3A5, <i>FUT2</i> , GPX1, HYAL1, HYAL2, LIG1, NADK, OAT, <i>PARK7</i> , PLCL1, PRDX6, QPRT, RXFP4, SLC22A5, SLC25A15, <i>SP140</i> , SPHK2, TAS1R3, THRA
	Monogenic forms of IBD	BACH2, CARMIL2, PMVK, TMEM50B, TNIP1 (<i>per</i> TNFAIP3)

List of 221 e-genes with DAP-matching EAP and published evidence (documented in Supplementary Data 26 and Supplementary Material) for an effect on one or more functions in one or more cell types participating in epithelial barrier, innate and adaptive immunity, or other IBD-relevant mechanisms. Genes affecting multiple categories of functions are italicized.

was shown to bind to arginine 121 of NEK7, thereby inhibiting its interaction with NLRP3⁵². Downregulation of *NEK7* by intraperitoneal injection of lentiviruses expressing anti-*NEK7* shRNAs was shown to attenuate DSS-induced colitis⁵³. In mouse models, ENB effectively reduced symptoms of NLRP3 inflammasome-related diseases (other than colitis)⁵². ENB has a high safety profile and is well tolerated by almost all patients without cumulative toxicity⁵⁴. Genetic variants that increase *NEK7* expression in naïve B cells increase the risk for IBD ($\theta = 0.94$; FDR = 0.02), while *NEK7* expression is increased in a majority of circulating immune cells of active CD patients, and in non-inflamed colonic epithelium of CD patients when compared to controls²⁸. On the downside, there is some evidence, albeit non-significant, that the same genetic variants decrease *NEK7* expression in some gut-resident immune cells (<https://tools.giga.uliege.be/cedar/publihpq>), while *NEK7* expression is down-regulated in some circulating immune cells of active CD patients (f.i., eosinophils, Supplementary Data 22), and these findings require further scrutiny.

Discussion

Expanding the catalogue of cis-eQTL

We herein increase the number of regulatory modules (RM) in IBD relevant cell types (circulating immune cell populations and intestinal mucosa) from ~9720 in a previous study¹⁵, to ~24,745, i.e., a 2.5-fold

increase. In addition, our data allows us to assign previously recognized RM to specific sub-cell types, which may provide novel insights into gene regulation and pathogenesis. This marked increase in the number of RM, despite analysis of the same tissue samples (blood and biopsies), reflects the high proportion of eQTL that are cell type-specific (reminiscent of those reported in circulating immune cells^{55,56}, and across multiple tissues by GTEx¹⁶), anatomical location-specific (f.i. hundreds of enterocyte eQTL appear to be specific for either the small or large intestine), or even context-specific (f.i. manifest in specific gut-resident cell types but not in their circulating counterparts), and which remained buried in the previous bulk-type analyses.

The degree of cell type-specificity of RM obviously depends on the chosen level of cell type granularity. While for blood, we followed standard FACS/MACS-based immunophenotyping procedures⁵⁷, yielding 26 distinct immune cell types in addition to PBMC, we developed a more dynamic approach for the intestinal scRNA-Seq data. We constructed a hierarchical tree of cell clusters (using transcriptome information), assumed to recapitulate the clusters' ontogenetic relationships. We then performed eQTL analysis separately for each one of 276 leaves and 275 nodes in the tree, and subsequently merged similar EAP into RM across leaves/nodes. We assumed that RM must be turned on at specific developmental time points, i.e., at specific nodes or leaves in the tree, and remain “on” for variable branch

lengths towards the leaves. Accordingly, the RM signal is expected to be strongest in the node in which it was turned “on”. Going up towards the leaves, signal strength should decrease as the number of cells per node diminishes. Going down towards the root, signal strength should decrease as cells in which the RM is active are being diluted by cells in which it is not. In practice, however, we didn’t see the predicted $\log(1/p)$ clines up and downwards from a node of maximum significance (presumably because power is affected by multiple factors). We therefore assigned RM to the MRCA of all active nodes/leaves. This ontogenic analysis framework, that is independent of cell type annotation, considerably increased eQTL yield. For comparison, using this approach we identified 35,010 eQTL affecting 3007 e-genes and clustering in 3081 RM, while a more standard analysis by 43 cell type yielded 3743 eQTL affecting 1309 e-genes and clustering in 1368 RM. It could be applied to several other analyses of single cell data.

We provide strong evidence for the existence of compartment-specific eQTL, and more specifically for eQTL that manifest themselves in gut-resident immune cell populations, while not being detectable in the cognate circulating cells. These included $\sim 1/4$ myeloid-specific eQTL and $\sim 3/4$ lymphoid-specific eQTL. Most gut-specific eQTL of the myeloid lineage were either detected in mastocytes or macrophages, i.e., cell types that acquire this terminal differentiation stage upon migration in tissues, accompanied by profound changes in their transcriptome. Accordingly, the gut-specificity of a larger proportion of these myeloid eQTLs resulted from scenario 1, i.e., induction of gene expression in the gut (35% vs 20%). *A contrario*, gut-specific eQTLs of the lymphoid lineage were detected in cell types that appeared indistinguishable from their circulating counterparts on the basis of their transcriptome. Accordingly, the gut-specificity of a larger proportion of these lymphoid eQTLs resulted from scenario 3, i.e., a switch in regulatory module (23% vs 8%). The procedures used for eQTL detection in blood and gut were not identical, and, hence, one should be attentive to possible technical rather than biological causes of the inferred gut-specificity. The eQTL detection pipeline used for the blood samples included a preprocessing step to avoid variant-dependent mapping bias, which was not applied to the gut samples. We excluded this as a major source of false-positive gut-specific eQTLs, as the support for only one out of 132 candidates was affected by preprocessing of the data with WASP¹³¹ (SFig. 8). We also note that the RNA extraction and cDNA synthesis procedure differed between the blood and gut samples. The blood RNAs were extracted using the Maxwell 16 LEV simplyRNA Tissue Kit (Promega #AS1280) on a Maxwell 16 instrument (Promega) or manually using the AllPrep DNA/RNA Micro Kit (Qiagen #80284) with the QIAshredder (Qiagen #79656) and converted to full-length cDNA using SMART-Seq HT Kit (Takara #634436), which poly-A selects mRNAs, while the gut RNAs were released and cDNA synthesized using the Chromium “Chromium Next GEM Single Cell 3’ Reagent Kits with Feature Barcoding technology for Cell Surface Protein, v3.0 or v3.1” technology. While we used both intronic and exonic reads for eQTL analyses in blood and gut, the gene coverage was more uniform in blood than in gut (typically, most exons equally covered in blood, while 3’ exons more prominently covered in gut). Also, the proportion of intronic versus exonic reads (and hence the proportion of captured pre-mRNAs versus captured mature mRNA) is expected to differ between blood and gut samples. One can imagine scenarios in which these idiosyncrasies generate eQTL that would only be detectable in gut or blood. For instance, imagine that alternative splicing generates transcripts that contain either exon I or exon II. Imagine also, that a “splicing variant” (say allele B) favors the exon II-containing transcripts. A cDNA synthesis method that covers both exons I and II would not detect differences in transcript levels between AA, AB, and BB individuals. However, a method that only interrogates one of the exons (say II) would conclude that the gene’s expression level is higher in BB than in AB than in AA individuals. With regard to the pre-mRNA versus mature mRNA ratio, variants that

would affect the splicing rate of pre-mRNA or the degradation rate of mature mRNA may generate eQTL that are only detectable when interrogating either pre-mRNA or mature mRNA population (see SFig. 11). And, hence, the proportion of pre- versus mature mRNA that is interrogated by the utilized methods may affect the ability to detect the corresponding “eQTL”. However, it is well established that the genome space devoted to the control of transcription rate is substantially larger than the genome space devoted to the control of alternative splicing, splicing, or degradation rate. It is therefore reasonable to speculate, that such mechanisms could only account for a minority of observed gut-specific eQTL. For statistical reasons, we restricted this preliminary search of compartment-specific eQTL to those that are active in the gut but not in the blood. However, there is no reason to believe that the reverse situation, i.e., compartment-specific eQTL that are active in blood but not gut, does not occur as well. Obtaining an accurate estimate of how common compartment-specific eQTL truly are, however, will require larger sample numbers (especially for the biopsies).

Expanding the catalogue of IBD risk loci with matching cis-eQTL

By confronting DAP to this expanded eQTL catalogue, we reveal matching EAP for 140 of 206 studied IBD risk loci, i.e., a ~ 2.3 -fold increase when compared to our previous study¹⁵. eQTL detected in the gut made a large contribution to DAP matches despite the fact that we only analyzed 57 individuals: while gut-specific RM only account for 10% of all RM (Fig. 4), they contribute 21% of the 140 EAP-matching IBD risk loci, and 48% of DAP-matching e-genes (SFig. 10A, B). eQTL that are specific for gut-resident lymphocytes and for mature enterocytes contributed most to DAP-EAP matches (Fig. 5B, C). In addition to uncovering RM that are specific for previously under-represented cell types, splitting risk loci in sub-DAP also proved effective: it revealed DAP-EAP matches for nine of the 32 risk loci on which it was applied.

It has been argued that it may be more effective—in order to find more DAP-matching eQTL—to perform eQTL analysis in one tissue type (f.i., blood) collected for (tens of) thousands of individuals, rather than in many cell types collected for hundreds of individuals. To test this, we repeated the DAP-EAP matching exercise using PBMC eQTL information from ~ 30 K European individuals⁵⁸. Using this data (Vösa), we identified DAP-matching EAP for 55 loci involving 97 genes. Using our own PBMC information (i.e., 187 individuals), we identified DAP-matching EAP for 32 IBD risk loci involving 42 genes, with 20 overlapping loci and 17 overlapping genes. Using our full blood data set (27 immune cell populations), we identified DAP-matching EAP for 110 loci involving 310 genes, of which 52 loci and 61 genes overlapped with Vösa. Using our complete data set (blood and gut), we detected DAP-matching EAP for 140 loci involving 556 genes, of which 54 loci and 66 genes overlapped with Vösa. Thus, the Vösa data enabled the identification of matching EAP for one IBD risk locus (rs1479918), and 31 e-genes that were missed with our data. However, we identified matching EAP for 86 IBD risk loci, and 490 e-genes that were missed with Vösa’s data. These findings clearly indicate that a large number of additional matches are uncovered when performing eQTL analyses in multiple, disease-relevant cell types.

While our study quite dramatically increases the number of IBD risk loci with matching EAP, at least one third still remain orphan. Common risk variants for CCD, including regulatory ones, are generally assumed to operate in healthy individuals, thereby increasing their probability to develop the disease. This justifies the use of cohorts of healthy individuals to discover eQTL that may underpin the risk to develop CCD, including the GTEx cohort¹⁶. It is becoming increasingly apparent, however, that for at least part of risk loci, the effect of the causative regulatory variants only manifest once the disease process has been initiated. Indeed, performing eQTL analyses in intestinal biopsies from patients has uncovered 190 cis-eQTL (out of 8881) that were inflammation-dependent⁵⁹. Also, performing eQTL

analyses in blood stimulated with TCR and TLR agonists doubled the number of detected cis-eQTL when compared to unstimulated blood, uncovering new DAP-matching EAP, albeit not in extra risk loci⁶⁰. More recently, an eQTL study conducted using colon samples from IBD patients indicated that ~10% of detected eQTL might be disease-specific. This included several instances, where the disease-process activates a distinct regulatory module (i.e., other regulatory variants drive gene expression in healthy versus diseased condition). Performing colocalization with these disease-specific eQTL resulted in the nomination of 16 (out of 194) additional target coding genes for IBD (*ABO*, *APOBR*, *CTSS*, *EFCAB13*, *EMILIN3*, *FERMT1*, *GATD3A*, *HLA-DQA1*, *HLA-DRB1*, *HLA-DRB6*, *FNGR2*, *LRRC25*, *P4HA2*, *PARK7*, *PWP2*, *TNFRSF14*)⁶¹. However, it is noteworthy that four of these (*APOBR*, *CTSS*, *ARK7*, *TNFRSF14*) are also nominated in this and/or a previous study conducted with samples from healthy individuals¹⁵. Also, by using single cell RNA-sequencing of intestinal biopsies, Alegbe et al.⁶² identified 764 “interaction eQTL” modulated by gut inflammation status (primarily active in enterocytes), leading to the nomination of new target genes in two additional IBD risk loci (out of 180). Thus, performing colocalization with eQTL detected in samples from IBD patients indeed uncovers additional target genes, albeit a small proportion. The blood samples of affected individuals used in this work to examine the effect of disease on gene expression will be used, in the future, to uncover affected-specific eQTL.

Expanding the catalogue of DAP-matching e-genes

It was initially assumed that finding a DAP-matching eQTL would be relatively uncommon, yet that it would provide strong evidence for causal involvement of the corresponding e-genes. Contrary to expectations, we herein report as many as 556 DAP-matching e-genes in 140 IBD risk loci, ranging from 1 to 34 e-genes per locus. Are all these genes causally involved in IBD, some or none? In other words, what is the proportion of “red herring” eQTL amongst our findings⁶³?

There are two formal tests of gene causality: the reciprocal hemizygosity test, which is impractical in humans^{64,65}, and the burden test, which has been used to demonstrate the causality of, for instance, *NOD2*^{66,67}. We took advantage of available whole exome sequence data for ~30 K IBD patients to test for an excess burden of ORF-disrupting variants in 298 of our DAP-matching e-genes¹¹. One gene only (*TAGAP*) harbored a single, low-frequency variant significantly associated with CD. There was no other signal, whether considering the e-genes individually or as a group (Fig. 5C). While the burden test is in principle appealing, it has so far yielded few causative genes despite the considerable investments made, whether for IBD¹¹ or other polygenic diseases⁶⁸. This is possibly due to the fact that the corresponding CCD result from regulatory variants that perturb the expression of the causal genes in specific, disease-relevant cell types, while coding variants in these genes, which affect their function in all tissues, cause distinct pathologies. Thus, we don't believe that the overall negative burden test conclusively excludes the genes' causality, but rather that it leaves the field with no effective test of gene causality.

Amongst the strongest arguments in favor of the fact that several DAP-matching e-genes in our list are causally involved in IBD are (i) the observation that the DAP-EAP “hit rate” with the CEDAR-2 data set (focusing on immune and intestinal cell types) is ~5-fold lower for dilated cardiomyopathy (CDM) when compared to inflammatory bowel disease (IBD) ($p \leq 1.4 \times 10^{-14}$), (ii) the observation that the scientific literature reports an effect on colitis-susceptibility following genetic perturbation in mice ~2.3 times more often for DAP-matching than for non-DAP-matching e-genes ($p \leq 1.4 \times 10^{-6}$), and (iii) the observation of a correlation between the strength of the colocalization (θ -value) and the reporting of a mouse colitis model ($p = 0.012$).

It is often assumed that there is one causative gene to be found for each risk locus. Yet, we herein present quite compelling evidence that there may very well be risk loci with multiple causative genes

controlled by the same or distinct sets of regulatory variants (see examples, in results, of loci with strong functional candidates that additionally harbor less well-known genes underpinning early onset, Mendelian forms of IBD). This is somewhat reminiscent, yet at a different scale, of the former demonstration that QTL effects may arise under an infinitesimal architecture⁶⁹. An additional level of complexity may result from the fact that a single gene may affect disease outcome through multiple pathways, operating for instance in different cell types, triggered by the same or by different regulatory modules. For example, *INAVA* was shown to play a pivotal role in PRR-induced signaling, cytokine secretion and bacterial clearance in peripheral macrophages and intestinal myeloid cells⁷⁰, yet at the same time to regulate the stability of adherens junctions of intestinal epithelial cells (where we see the best DAP-match)⁷¹. Also, *ATG16L1* is increasingly understood to affect disease outcome through autophagy-dependent but also -independent mechanisms⁷², in agreement with our observation of a DAP-match in eosinophils but also colonocyte precursors.

A contrario, arguments suggesting that not all DAP matching e-genes are causally involved in IBD pathogenesis, are (i) the relatively modest enrichment of Reactome pathways (both in terms of number of highlighted pathways and significance of the enrichments; Supplementary Data 24), and (ii) the troubling observation that when the DAP of a risk locus matches EAP in both blood and gut, the e-genes involved are generally different. It is worth noting that the omnigenic model^{73,74} asserts that—at least for some risk loci—the effects of regulatory variants on disease are not mediated by the perturbation of (a) causative gene(s) *in cis*, a sobering thought for positional cloners. This does not mean that one will not observe DAP-matching e-genes, but that these e-genes do not directly participate in the disease process.

Insights into the pathogenesis of IBD

An important observation of this study is that eQTL that are only active in intestinal tissue make a disproportionate contribution to DAP-matching e-genes as compared to eQTL that are only active in circulating immune cells, in agreement with the predominant intestinal disease manifestations of IBD. Within the intestinal compartment, however, no specific cell type stood out, other than (i) the absence of any DAP-matching eQTL in enterocyte precursors, and (ii) the reduced participation of eQTL active in gut-resident myeloid cell types in UC risk loci when compared to CD risk loci (Fig. 5B, C). Unsupervised Reactome analysis of the 556 DAP-matching e-genes reveals four enriched pathways, highlighting the importance of inflammation, innate immunity and the cross-talk between innate and adaptive cellular immunity, as expected⁷⁵ (Supplementary Data 26). Further “manual” consultation of the literature uncovered published, experimentally-supported links for nearly half of DAP-matching e-genes covering nearly all facets of mucosal immunity including epithelial homeostasis and barrier function, multiple types of pattern-recognition receptors, signaling in innate immunity including IL6- and TNF α -signaling, biology of macrophages and dendritic cells, leucocyte extravasation and trafficking, autophagy and ubiquitin-dependent processes, inflammatory mediators including prostaglandins, glucocorticoids, and the complement system, antigen presentation, signaling in adaptive immunity including BCR- and TCR-dependent, multiple classes of T and B lymphocytes, as well as other mechanisms including oxidative stress, metabolism of sphingosine and non-proteinogenic amino acids, gut motility, and host-microbiome interactions (Table 1 and Fig. 6). Jointly, perturbation of components of adaptive immunity seem to predominate. It is important to recognize, however, that the interrogation of biological functions is inherently biased by assay availability and “en vogue” paradigms, providing what is certainly a partial if not biased view. What is certain, however, is that IBD results from the perturbation of multiple distinct immunological pathways as abundantly recognized⁷⁵.

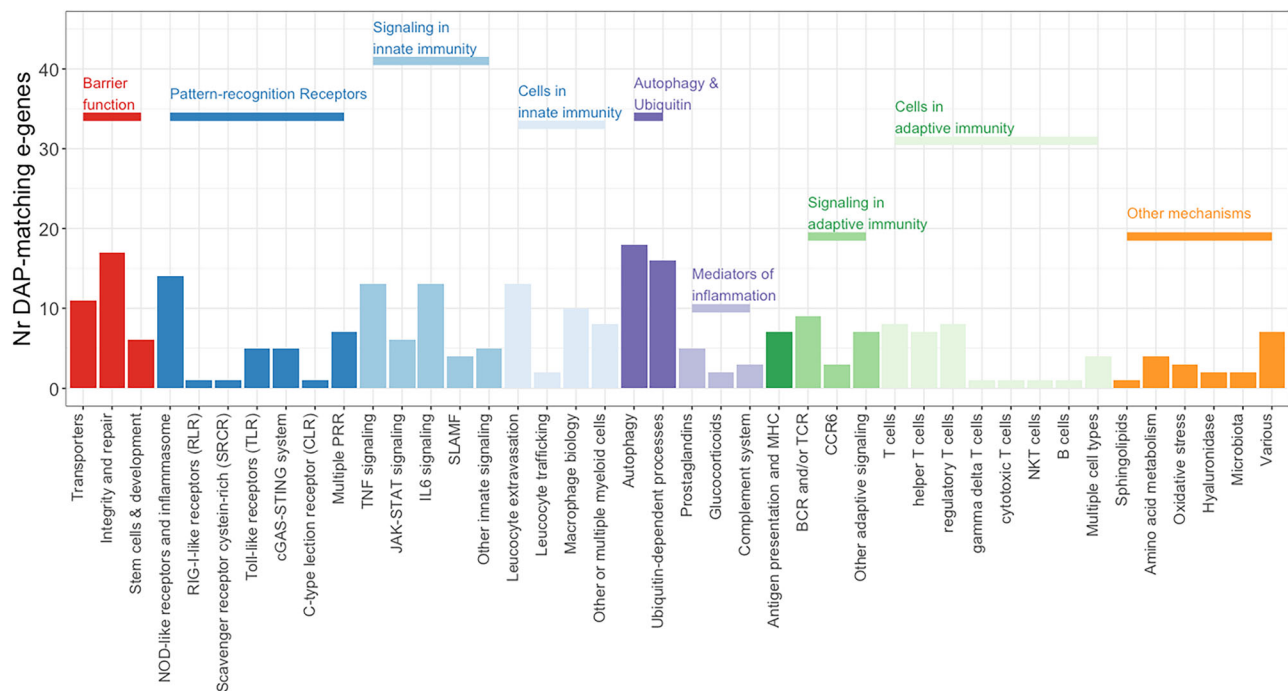


Fig. 6 | Biological functions of IBD-matching e-genes. Assignment of 220 of the 481 protein-coding DAP-matching e-genes to pathways related to epithelial homeostasis and barrier function, pattern-recognition receptors, signaling in innate immunity, cell biology in innate immunity, autophagy and ubiquitin-

dependent processes, inflammatory, antigen presentation, signaling in adaptive immunity, biology of T and B lymphocytes, as well as other mechanisms, providing hints about their relative importance. Source data are provided as a Source data file.

Genetic support for repurposing opportunities in IBD

We propose entrectinib (ENB) as a promising repurposing candidate for CD. Entrectinib is approved to treat solid tumors driven by specific oncogenic tyrosine kinases⁵¹. It is safe and well supported by patients⁵⁴. It was recently shown in vitro to inhibit the NLRP3 inflammasome by binding to NEK7⁵², one of our DAP-matching e-genes that is affected by the disease process in a consistent manner, at least in some cell types. Moreover, it was shown in in vivo models that shRNA-mediated NEK7 inhibition or ENB administration affects the outcome of inflammatory conditions including colitis^{52,53}. Thus, it appears that clinical trials to test the efficacy of ENB for IBD could be undertaken. On a cautionary note, however, it appears that the CD/IBD risk variants that are associated with an increase in *NEK7* expression in circulating naïve B lymphocytes may have the opposite effect (i.e., decrease *NEK7* expression) in gut-resident T and B lymphocytes (including plasmacytes) (albeit not statistically significant). If these effects prove to be true, and if they contribute to the risk to develop IBD, pharmacological down-regulation of *NEK7* may in fact have adverse effect. Also, it was demonstrated recently that the effect of NEK7 degradation on IL1 β production (in this case using a cereblon glue degrader), may be less potent in primates than in the mouse⁷⁶.

We identify at least two instances where IBD risk variants, rather counterintuitively, decrease the expression levels of what are assumed to be positive mediators of inflammation, in particular *IL18R1* and *IL18RAP* on the one hand, and *PTGIR* on the other (Supplemental material). Thus, “hypomorphic” IL18 and prostacyclin pathways may increase the risk to develop IBD, despite the fact that these pathways participate actively in the inflammatory process once initiated. These observations obviously call for caution when intending to use activators of the corresponding pathways to treat active IBD patients. It suggests, however, that treatment of IBD patients in active phase versus relapse should maybe be differentiated. It is conceivable that activation of the IL18 and prostacyclin pathways might help to prevent relapse, particularly after disease remission has been achieved following surgery.

It is now well established that the genetic architecture of CCD involves hundreds if not thousands of risk loci. Moreover, the architecture of individual risk loci may itself be fractionable, involving multiple genes perturbed by the same or distinct variants, and affecting disease outcome through multiple pathways. The multiple risk loci and genes therein are unlikely to affect disease outcome independently. It seems more likely that they trigger secondary effects along pathways that progressively merge into a decreasing number of “highways”, that ultimately determine disease. Such “highways” would involve the “core genes” as defined in the omnigenic model^{73,74}. If the molecular species that underpin these secondary effects can be quantified, they should be amenable to QTL analysis, and reveal DAP-matching EAP for one risk locus if operating close to the *cis*-effects, or multiple risk loci if operating closer to the disease. This should, in theory, allow reconstruction of the topology of the pathways linking the multiple risk variants with the disease. It seems that the components of the upper parts of the network (i.e., the “highways”) would make for better drug targets than those at the bottom of the tree. Of note, the abundance of the components of the upper part of the network is expected to be most strongly correlated with polygenic risk scores for the corresponding disease and—if conveniently measurable—may constitute biomarkers capturing both genetic and environmental effects.

It is a widely held and supported view that causative genes underpinning risk loci identified by GWAS constitute preferred drug targets^{7–10}. Yet, as already stated in results, it is not obvious that pharmacological targeting of a gene product that, upon perturbed expression by risk variants, increases the risk for a still healthy individual to develop a disease will necessarily reverse the disease process, i.e., be curative. Indeed, once the disease process is initiated, it's maintenance may rely less on factors that triggered it in the first place. A more straightforward expectation is that targeting such causative gene products in still healthy individuals may have a preventive effect. Yet, measuring the efficacy of such (personalized) preventive drug treatments will be difficult, and its implementation is hard to envisage

to date. As mentioned before, part of the missing DAP-matching eQTL may only manifest themselves after initiation of the disease process. It may be argued that the causative genes that underpin these context-specific eQTL may constitute even better drug targets. eQTL studies conducted in cohorts of affected individuals may therefore advantageously complement present eQTL studies which are primarily using healthy individuals.

Methods

Identifying cis-eQTL modules in bulk RNA-Seq data of 27 sorted circulating immune cell populations

Sample collection. We collected 40 ml of venous blood (EDTA) from 251 healthy European subjects at the academic hospital of the University of Liège (CHU) between October 2018 and November 2022. Written informed consent was obtained prior to donation in agreement with the recommendations of the declaration of Helsinki for experiments involving human subjects. The experimental protocol was approved by the Ethics Committee of the CHU Liège (reference number: 2017/214). Data collected from the electronic medical records (EMR) included birth date, age at sampling, ancestry, sex, weight, height, smoking and alcohol history, declared ethnicity, family history of disease, surgical and medication history, blood type when available, and known allergies. The following hematological parameters were measured: counts of white blood cells, neutrophils, lymphocytes, monocytes, eosinophils, basophils, platelets, red blood cells, hemoglobin concentration, hematocrit, mean cell volume, mean cell hemoglobin, mean cell hemoglobin concentration, red cell distribution width and mean platelet volume (Supplementary Data 1).

SNP genotyping. Genomic DNA was isolated from frozen EDTA-blood using the NucleoMag Blood 200 μ L kit (Macherey-Nagel) on a King-Fisher robot (Thermo Fisher Scientific). Individuals were genotyped for 713,606 SNPs using Illumina's Human OmniExpress BeadChips, an iScan system, and the Genome Studio software following the guidelines of the manufacturer. We confirmed the European ancestry of the participants by PCA (using the HapMap population as reference) as well as the absence of duplicated or related individuals (π hat > 0.185). All individuals had less than 3% of missing genotypes. We excluded variants with call rate ≤ 0.95 or deviating from Hardy-Weinberg equilibrium ($p \leq 3 \times 10^{-4}$) using plink (v1.9), leaving 689,223 quality-controlled variants. After lifting over to GRCh38 using Picard LiftoverVcf (v2.7.1), we phased and imputed to whole genome using the TOPMed Imputation Server (v1.6.6) and the TOPMed r^2 reference panel. We removed variants with imputation score (R^2) ≤ 0.7 or minor allele frequency (MAF) ≤ 0.05 using bcftools (v1.11), leaving genotypes at 6,299,998 QC-ed variants.

Cell sorting. Granulocytes were isolated from blood using the EasySep™ Direct Human Pan-Granulocyte Isolation Kit (StemCell Technologies #19659) within one hour after collection. Neutrophils and eosinophils were then isolated from the recovered granulocytes using the EasySep™ Human Neutrophil Isolation Kit (StellCell Technologies #17957) and EasySep™ Human Eosinophil Isolation Kit (StemCell Technologies #17956), respectively. Cells were recovered in cell homogenization buffer provided with the Maxwell 16 LEV simply RNA Tissue Kit (Promega #AS1280) or the AllPrep DNA/RNA Micro Kit (Qiagen #80284), and immediately frozen at -80°C until use.

Peripheral blood mononuclear cells (PBMC) were isolated from fresh blood using SepMate™-50 (IVD) collection tubes (StemCell Technologies #85460) and Lymphoprep™ density gradient medium (StemCell Technologies #07861). After two washing steps, PBMC were stained with two panels of antibodies for 30 min at 4°C . Cells suspended in PBS were then filtered using a $100\text{ }\mu\text{m}$ CellTrics filter (Sysmex #04004-2328). Cell sorting was performed on a FACS Aria III instrument (BD Biosciences) calibrated using CS&T beads (BD

Biosciences). Fluorescence compensations were performed using CompBeads (BD Biosciences #552843). After exclusion of debris and doublets, we targeted monocytes (classical, non-classical and intermediate), T lymphocytes (including $\gamma\delta$, mucosal-associated invariant T cells (MAIT) cells, naive and memory regulatory T cells, naive and memory CD8+ T cells, naive and memory CD4+ T cells, Th1, Th2, Th17 and Th1/17 helper T cells), B lymphocytes (naive, memory and plasmacytes), dendritic cells (plasmacytoid and myeloid), natural killer cells (NK and NKT) and innate lymphoid cells (ILC). Panels of antibodies used, definition of sorted cells, and gating strategies are described in SFig. 1 and Supplementary Data 29. Purity of the sorted cell populations ranged from 78% to 99%. Up to 20,000 cells of each cell population were sorted directly on the cell homogenization buffer and frozen immediately at -80°C until use. No residual samples remain.

Bulk RNA sequencing. Total RNA was purified from sorted cells using the Maxwell 16 LEV simplyRNA Tissue Kit (Promega #AS1280) on a Maxwell 16 instrument (Promega) or manually using the AllPrep DNA/RNA Micro Kit (Qiagen #80284) with the QIAshredder (Qiagen #79656), according to their respective manufacturer's instructions. RNA quantity was determined for all samples using the Quant-iT RiboGreen RNA Assay Kit (ThermoFisher Scientific #R11490), while the RNA quality was evaluated for a subset of samples using the RNA 6000 Pico Kit on a 2100 Bioanalyzer instrument (Agilent #5067-1513). Reagents for cDNA and library preparation were dispensed with a Mantis Liquid Handler (Formulatrix) using half the recommended volumes. Full-length cDNA was generated from 1 ng of total RNA using the SMART-Seq HT Kit (Takara #634436), which poly-A selects mRNAs. Obtained cDNA quantity and quality were determined for all samples using the Quant-iT PicoGreen dsDNA Assay Kit (ThermoFisher Scientific #P7589), and for some using the High Sensitivity DNA kit on a 2100 Bioanalyzer instrument (Agilent #5067-4626), respectively. Uniquely indexed libraries were constructed from 300 pg of cDNA using the Nextera XT DNA Library Preparation Kit (Illumina #FC-131-1096) and custom-made 24 forward and 16 reverse primers. The libraries' quantity was assessed for all samples by qPCR using a KAPA Library Quantification Kit (Roche#07960140001), while the libraries' quality was checked for all samples using a QIAxcel Advanced technology (Qiagen). The libraries were pooled and sequenced on a NovaSeq 6000 instrument (Illumina), to an average read depth of 11 ± 5 million paired-end reads per sample (University of Geneva core facility (Geneva): 2×50 bp (2112 samples); GIGA Genomics platform (Liège): 2×150 bp (3180 samples)). We performed RNA-seq for 5292 samples, corresponding to 27 cell types from 196 individuals. No residual samples remain.

Read mapping and quantification. Demultiplexing and FASTQ conversion were performed using bcl2fastq (v2.20). Read quality was assessed with FastQC (v0.12.1) and multiQC (v0.9). Reads were mapped to the GRCh38 (Ensembl release 105) human genome build using STAR (v2.7.1a). The STAR re-implementation of the WASP algorithm was used with a custom VCF file containing both reference and alternative alleles for all SNPs. The alignments that did not pass WASP filtering or that overlapped indels were removed from the resulting BAM files using samtools (v1.9). Alignment metrics were collected using Picard CollectRnaSeqMetrics (v2.7.1) (Supplementary Data 3). Matching of genomic and transcriptome genotypes was evaluated with QTLtools mbv (v1.3.1) (SFig. 2A). Unstranded gene counts were generated with HTSeq (v0.6.1p1). If samples were split across two sequencing lanes, we summed up the respective counts. To detect mislabeled samples, reads counts were normalized using the variance stabilizing transformation from the DESeq2 R package, and a t-SNE analysis was conducted using the 500 most variable genes (SFig. 2).

Supplementary Data 3 reports the characteristics of the 5030 RNA-seq libraries that passed the quality control procedure.

Transcriptome-based hierarchical clustering of circulating immune cell types. A dendrogram was constructed based on the RNA-seq data, using 1-|Spearman's correlation| between average (across all individuals) gene expression levels, using the “average”, “ward.d” and “ward.d2” methods. For each method, we built 32 dendrograms with from 250 to 8000 (by steps of 250), genes with best F statistic (cell type effect) from ANOVA. Within each method, we assessed the reliability of each dendrogram in a way inspired by the bootstrap procedure: for each node in the dendrogram, we computed the proportion among the 31 other trees that shared the same split. Within each method, we selected the dendrogram(s) with the highest sum of “bootstrap-like” values. Memory CD4 T cells, PBMCs, and granulocytes were ignored in this analysis as they encompass multiple cell types.

Cis-eQTL analyses. For each blood cell type, we filtered out the genes with less than 5 counts in more than 80% of samples, normalized the raw read counts using the DESeq2 R package and residualized them for age, sex, RNA extraction method, proportion of reads in each sequencing batch, top 3 genotype principal components (PCs) and top 13–36 expression PCs to maximize the number of cis-eQTLs. Supplementary Data 4 reports the number of samples, genes, and top expression PCs for each cell type. We removed variants with $MAF \leq 0.05$ in the retained samples using bcftools (v1.9) for each cell type. We performed eQTL mapping using QTLtools¹⁹ in a 2 Mb cis-window centered at the transcription start site and with the integrated rank normal transformation of the phenotypes. QTLtools applies two sequential corrections for multiple testing. The first corrects for the multiple markers that are tested in each cis-window. To account for the linkage disequilibrium (LD) that may exist between these markers, correcting for multiple testing is accomplished by applying a permutation scheme in which the expression phenotypes are randomized with respect to the unaltered vectors of marker genotypes (to maintain the LD structure). For each permutation pass, the best nominal p -value (across all markers in the window) is stored, yielding a list of N best p -values (where N is the number of permutations) under the null. The significance of the nominal p -values for the real (unpermuted) data, accounting for the testing of multiple markers per window, is then estimated as the proportion of equally low or lower p -values in the list of N best p -values under the null. To obtain accurate estimates of these corrected p -values, even when inferior to $1/N$, while limiting the number of permutations to run, QTLtools uses an approximation method based on the beta distribution⁷⁷, using data from 10,000 permutations. For each gene, QTLtools keeps the marker with the best corrected p -value in the window, yielding a list of window-corrected best p -values (one per gene) across the genome. These corrected p -values are then subjected to a second round of correction (for the testing of multiple genes) using the false discovery rate (FDR). eQTL with “within cell type FDR” ≤ 0.05 were considered significant.

Proportion of expression variance explained by cis-eQTLs. The proportion of expression variance was computed as $2pq\beta^2$, where p and q are the allelic frequencies of reference and alternate allele, respectively, and β is the slope of the regression line of standardized gene expression (mean: 0, variance: 1) on dosage of the alternate allele, i.e., the allele substitution effect. This gave near identical results as QTLtools (r-squared values).

Agglomerating cis-eQTL in gene-specific and across-genes cis-acting regulatory modules (RM). To assess if the expression levels of a given gene are affected by the same regulatory variants in a given pair of tissues, we compared the corresponding eQTL association patterns (EAP) using the θ metric devised in Momozawa et al.¹⁵ Two EAP were

compared if at least one of them was significant ($FDR \leq 0.05$). We only used variants that had a p -value below 0.05 in at least one of the two EAP to compute θ . Two EAP were assumed to be part of the same cis-acting regulatory module (RM) if $|\theta| \geq 0.6$ (cfr. ref.¹⁵), and if the window-adjusted p -value of the eQTL (for non-significant eQTL) was ≤ 0.12 , as these parameter values were shown to yield an agglomeration $FDR \leq 0.05$ in a permutation test (see DAP-EAP matching section, hereafter). To better describe the connectivity of the modules, we retrospectively also computed θ for pairs of non-significant EAP if they were assigned to the same module. We first constructed gene-specific modules, i.e., we only confronted EAP from the same gene yet from different cell types. In a second stage, we constructed across-gene modules by evaluating the similarity of the EAP of different genes in the same or in different tissues, using the union of the overlapping 2 Mb windows, and using the same threshold values as above. We defined a representative EAP for each module as the EAP with the lowest adjusted p -value for β . For all EAP in a module, the sign of β was compared to that of the representative EAP. If the modules encompassed more than one EAP, we performed a meta-analysis to combine the constituent EAP in a consensus EAP representing the module. For each variant of the complete window (2–4 Mb), we converted nominal p -values to z -score which we squared and summed across all EAP in the module. The corresponding sum was assumed to have a chi-squared distribution with degrees of freedom equal to the number of EAP in the module. When arithmetic underflow was reached for the p -values, the $-\log_{10}(p)$ values were predicted from the z -scores using a local polynomial regression. RMs were then curated to remove the connections between non-similar EAP. The similarity between the EAP and their consensus EAP was assessed using θ . If $|\theta| < 0.6$, the EAP were excluded from the module and the consensus EAP reconstructed. Similarity between excluded EAP was tested to allow them to form distinct “sub-RM”.

Quantifying the statistical significance of the overdispersion of module activity across cell types. We quantified the dispersion for the real data, as the variance of the sum of active cell types across modules. We then performed “permutations”, in the sense that—for a given cell type—the activity states (1's and 0's) were permuted between modules. This was repeated for all cell types, and the dispersion of that permutation computed as the variance of the sum of “active” cell types across modules. The statistical significance of the real dispersion was then defined as the proportion of permutations that yielded as large or larger variance than the real data.

Assigning modules to nodes and leaves in the ontogenic dendrogram. Gene-specific modules were assigned to nodes and leaves of the best supported ontogenic tree (SFig. 2C). Modules that encompassed only one cell type were assigned to the leaves of the tree, while modules that encompassed more than one cell type were assigned to the node corresponding to the most recent common ancestor (MRCA) of those cell types. Descendants of such MRCA nodes that were not encompassed by the module were assumed to have lost the eQTL effect. If these were part of another regulatory module for the same gene, the “loss” was converted to a “switch”. Memory CD4 T cells, PBMC and granulocytes were ignored in this analysis (cfr. above).

Testing for an excess sharing of RM between cell types. We expect that an eQTL detected in cell type i (f.i. B lymphocytes) will be shared more often with cell type j (f.i. T lymphocytes) that is ontogenetically closely related to i , than with cell type k (f.i. neutrophils) that is less closely related. In our regulatory module-based analysis framework, this means that monogenic regulatory modules (cRM) active in cell type i will more often be active in cell type j (and hence shared between i and j) than in cell type k . This constitutes our alternative hypothesis (H1). The corresponding null hypothesis (H0) assumes that a

monogenic cRM that is active in cell type i , has equal probability to be active in cell type j as in cell type k . For various reasons, the number of eQTL detected differs between cell types (SFig. 3A), and this will affect the sharing probabilities. The null hypothesis should therefore be reformulated as follows: the number of cRM active in cell type i that are shared with cell type j (as opposed to all $k \neq j$), should be proportionate to the number of eQTL detected in cell type j relative to the number of eQTL detected in cell type k . Thus, under H_0 , if the number of eQTL detected in cell type k is twice that number in cell type j , the expected number of modules shared between i and k is expected to be twice the number of modules shared between i and j . We tested this following Momozawa et al.¹⁵. Formally, assume that n_{ij}^S is the number of modules shared between cell type i and j , and that $n_{iT}^S = \sum_{j \neq i} n_{ij}^S$. Assume also that n_j^{eQTL} is the number of eQTL detected in cell type j , and that $n_{iT}^{eQTL} = \sum_{j \neq i} n_j^{eQTL}$. Under H_0 , the number of cRM active in cell type i that are also active in cell type j is expected to follow a binomial distribution with n_{iT}^S independent trials and probability of success $p_{succ} = n_j^{eQTL} / n_{iT}^{eQTL}$. The p -value of the observed number of successes (n_{ij}^S) was computed in R as $\text{pbinom}(c(n_{ij}^S, n_{iT}^S - n_{ij}^S), p = c(n_j^{eQTL} / n_{iT}^{eQTL}, 1 - n_j^{eQTL} / n_{iT}^{eQTL}), \text{lower.tail} = \text{FALSE}) + \text{dbinom}(c(n_{ij}^S, n_{iT}^S - n_{ij}^S), p = c(n_j^{eQTL} / n_{iT}^{eQTL}, 1 - n_j^{eQTL} / n_{iT}^{eQTL}))$. The first term yields $P[n_{succ} > n_{ij}^S]$ and the second $P[n_{succ} = n_{ij}^S]$. We thus perform a one-tailed test, generating the probability (under H_0) to observe the sharing of as many or more cRM between cell types i and j . The test is performed using one cell type as reference (i in the above). For each pair of cell types i and j , we performed two tests using first i and then j as the reference. The correlation between the two corresponding $-\log(p)$ -values was 0.96. The test was first performed using cRM that are active in only two cell types, then using cRM that are active in up to 3 cell types, then using cRM that are active in up to 4 cell types, etc., for a total of 26 tests.

We realize that the excess of cRM sharing between cell types, computed as described above, is largely confounded with the excess sharing of gene expression between cell types. For example, B and T lymphocytes may share more cRM than B lymphocytes and neutrophils, just because there are more genes that are expressed in both B and T lymphocytes, than genes that are expressed in both B lymphocytes and neutrophils. We therefore devised a test to evaluate whether closely related cell types were characterized by an excess sharing of polymorphic gene regulation on top of the excess sharing of gene expression profiles. To do this, we restricted the analyses to genes that were assigned to more than one cRM (say cRM “A”, “B” and “C”) (indicating that the genetic variants underpinning the observed eQTL differ between cell types), of which one was active in reference cell type i (say cRM “A”). We then identified the cell types in which cRM “B” or “C” were active. Thus, we identified the cell types with an eQTL for gene “x”, yet driven by distinct variants than those driving the eQTL for gene “x” in cell type i . Our prediction (and hence H_1) was that such events (module switching) would be more common for distantly related than for closely related cell types. As control, we also identified the cell types in which cRM “A” is active (as it is in cell type i). Such module sharing events (for genes that are subject to more than one cRM) are expected to be more common for closely related than for distantly related cell types. The corresponding null hypotheses were the same as before. Formally, assume that n_{ij}^{MS} is the number of genes for which we observed a module switch (MS) between cell type i and j , and that $n_{iT}^{MS} = \sum_{j \neq i} n_{ij}^{MS}$. Assume also that n_j^{eQTL} is the number of eQTL detected in cell type j , and that $n_{iT}^{eQTL} = \sum_{j \neq i} n_j^{eQTL}$. Under H_0 , the number of genes with module switch between cell type i and cell type j is expected to follow a binomial distribution with n_{iT}^{MS} independent trials and probability of success $p_{succ} = n_j^{eQTL} / n_{iT}^{eQTL}$. The p -value of the observed number of successes (n_{ij}^{MS}) was computed in R as $\text{pbinom}(c(n_{ij}^{MS}, n_{iT}^{MS} - n_{ij}^{MS}), p = c(n_j^{eQTL} / n_{iT}^{eQTL}, 1 - n_j^{eQTL} / n_{iT}^{eQTL}), \text{lower.tail} = \text{FALSE}) + \text{dbinom}(c(n_{ij}^{MS}, n_{iT}^{MS} - n_{ij}^{MS}), p = c(n_j^{eQTL} / n_{iT}^{eQTL}, 1 - n_j^{eQTL} / n_{iT}^{eQTL}))$, as before.

+dbinom($c(n_{ij}^{MS}, n_{iT}^{MS} - n_{ij}^{MS}), p = c(n_j^{eQTL} / n_{iT}^{eQTL}, 1 - n_j^{eQTL} / n_{iT}^{eQTL})$), as before.

Probing the causes of the cell type-specificity of gene-specific RM. Why is a *cis*-eQTL for gene “X” detected in cell type a but not b ? We distinguished three possible scenarios: (i) Module switch: gene “X” is subject to a *cis*-eQTL effect in both cell type a and b , but the variants involved are distinct (i.e. dissimilar EAP), and the two *cis*-eQTL are assigned to different RM, (ii) Lack of expression cell type b : gene “X” is expressed at too low levels in cell type b to allow for the detection of a *cis*-eQTL effect, and (iii) Conditional eQTL effects: gene “X” is expressed at sufficient levels in cell type b , but there is no evidence for a *cis*-eQTL effect (significant variants $\times$ cell type interaction effect). To test the statistical significance of the third scenario, we performed an interaction test using a linear model including genotype effect, cell type effect, genotype $\times$ cell type interaction effect, and a random intercept (to account for individual, across-cell type effects on a gene’s expression level), in R (lmer function of lme4 and lmerTest packages). To mitigate scale-dependent effects, we used within-cell-type, QQ normalized expression values (the same as the ones used for eQTL analysis). This removes cell type effects but doesn’t preclude testing the interaction effect. We estimated the proportion of true alternative hypotheses amongst the realized tests (i.e., π_1 as defined by Storey and Tibshirani²⁰) separately for genes whose expression level was higher in cell type a than in cell type b on the one hand and genes whose expression level was higher in cell type b than in cell type a on the other hand, to evaluate to what extent lower (and hence noisier) expression levels in cell type b may inflate the number of significant interactions.

Identifying *cis*-eQTL modules in single-cell RNA-Seq data from intestinal biopsies at three anatomical locations

Sample collection. Gut biopsies were obtained from 60 healthy adults (27 females and 33 males, average age was 54 years, ranging from 23 to 75) that were visiting the University Hospital of the University of Liège as part of a screening campaign for colon cancer between June 2019 and December 2021. Written informed consent was obtained prior to donation in agreement with the recommendations of the Declaration of Helsinki for experiments involving human subjects. The experimental protocol was approved by the Ethics Committee of CHU Liège (reference number: 2017/214). Two to four biopsy “bites” were collected from rectum (RE) and transverse colon (TC) for all participants, while biopsies from the terminal ileum (IL) were obtained for 52. Biopsies were collected in 40 ml of RPMI-1640 culture medium (Lonza Bioscience, 12–167 F) supplemented with 2 mM L-glutamine (Thermo Fisher Scientific, 25030024) and 10% FBS (Sigma, F7524) on ice, and processed freshly within one hour from the time of colonoscopy. Data collected from the electronic medical record (EMR) were the same as for the individuals providing blood samples (Supplementary Data 1). No residual samples remain.

SNP genotyping. Was conducted using the same procedure as for the circulating immune cell population samples (see above). We kept 686,493 SNPs interrogated by Illumina’s OmniExpress array after quality control, while genotypes at 6,352,658 variant positions were kept after imputation and quality control.

Single-cell RNA sequencing. Biopsies from the three locations (IL, TC, RE) were processed in parallel using so-called two-step protocols^{22,28} with some modifications. Fractionation of epithelial (EC) and lamina propria (LP) cell layers: The biopsies delivered in the transport media were collected by passing the media through a 100 μ m cell strainer (Pluriselect Life Science, 43-50100-50), and transferred to a 50 ml EZFlip tube (Thermo Fisher Scientific, 10571663) with 25 ml of pre-warmed Epithelial Strip Buffer consisting of HBSS (Thermo Fisher Scientific, 14170088), 5 mM EDTA (Thermo Fisher

Scientific, AM9260G), 15 mM HEPES (Lonza Bioscience, 17-737E) and 5% FBS and a magnetic stirring bar. The sample was agitated with gentle stirring (130 rpm) for 10 min at 37 °C by placing the tube upside down on a magnetic stirrer (Thermo Fisher Scientific, 50088009) in a 37 °C incubator. After adding DTT to a final concentration of 1 mM (VWR, 443852A), the sample was incubated for another 10 min with agitation at 37 °C. The sample was taken out of the incubator and shaken by hand vigorously for 10–15 s. It was passed through a 100 µm cell strainer to fractionate EC in the flow-through in a new 50 ml canonical tube, while the LP remained on the cell strainer. The LP sample on the strainer was rinsed with Washing Buffer (HBSS supplemented with 1 mM EDTA and 1% FBS) and kept on ice in a 6-well containing Wash Buffer. **Dissociation of EC:** The tube with the EC fraction was filled with ice-cold Washing Buffer up to 50 ml and centrifuged at 500 rcf for 5 min at 4 °C. After carefully removing the supernatant, the sample was transferred to a 1.5 ml siliconized microtube (Sigma, T4816) using 100 µl of TrypLE Express enzyme solution (Thermo Fisher Scientific, 12604-013). The sample was then mixed ten times using a 200 µl tip and incubated in a water bath at 37 °C for 5 min. The reaction was stopped by adding 1 ml of ice-cold Wash Buffer. ECs were collected by centrifugation at 500 rcf for 5 min at 4 °C, and resuspended in 100 µl of PBS (Lonza Bioscience, 17-516 F) with 10% FBS. Cell concentration and viability were estimated by staining 10 µl of cell suspension with an equal volume of 0.4% Trypan blue solution (Lonza Bioscience, 17-942E) using either TC20 (Bio-Rad) or Countess 3 (Thermo Fisher Scientific) automated cell counters. We obtained an average of $4.1\text{E} + 05$, $2.6\text{E} + 05$, and $1.7\text{E} + 05$ EC cells with viability of 58%, 47%, and 47% for IL, TC, and RE, respectively. **Dissociation of LP cells:** The LP remaining on the cell strainer was transferred to a gentleMACS C tube (Miltenyi Biotec, 130-093-237) using 15 ml of pre-warmed Enzyme Solution consisting of HBSS supplemented with 2.5 mg of Liberase TL (Roche, 05401020001), 7.5 U of DNase I (Thermo Fisher Scientific, EN052), and 2% FBS. The LP tissue was dissociated using a gentleMACS Octo Dissociator (Miltenyi Biotec) with “37C_m_LPK” program (~ 25 min). Large debris were filtered out by passing the cell suspension through a 100 µm cell strainer into a new 50 ml tube. The tube was filled with ice-cold Washing Buffer up to 50 ml and spun at 500 rcf for 5 min at 4 °C. After carefully removing the supernatant, the sample was treated with TrypLE Express enzyme and resuspended in 100 µl of PBS with 10% FBS as described above. We obtained an average $9.1\text{E} + 05$, $8.6\text{E} + 05$, and $5.5\text{E} + 05$ LP cells with viability of 74%, 78%, and 77% for IL, TC, and RE, respectively. **Cell hashing:** To reduce technical batch effects and costs of droplet-based scRNA-Seq⁷⁸, we labeled each fraction of cells with distinct oligo-tagged antibodies and performed droplet formation using all fractions from a donor together in a single well of a 10X Genomics Chromium system. As we generally obtained larger numbers of cells from LP than EC, the LP cell suspension was divided into two or three tubes, while EC was kept in one tube (total 10 tubes). The total volume of the cell suspension per tube was adjusted to 90 µl using PBS with 10% FBS. The cell suspensions were first incubated with 5 µl of Human TruStain FcX Fc receptor blocking solution (BioLegend, 422301) for 10 min at 4 °C, then mixed with 2 µl of 10 times diluted unique TotalSeq-B anti-human Hashtag antibodies (BioLegend; see also Supplementary Data 10) and incubated at 4 °C for 30 min. The cells were washed twice by adding 1 ml of PBS with 10% FBS and centrifuged at 400 rcf for 5 min at 4 °C. The cells were resuspended in 100 µl of PBS with 10% FBS, and cell density and viability were estimated as described above. Equalized numbers of cells (average of 70,000 cells per tube) were pooled into one tube. In addition, 20,000 non-labeled cells were added in the pool, to provide base lines of hashtag reads when demultiplexing sequencing data. The cell pool was washed once more and re-suspended in 100–400 µl of PBS with 10% FBS. After filtering through a 70 µm filter followed by a 40 µm cell strainer (Thermo Fisher Scientific, 22363548 and 22363547), cell density and viability were measured as described

above. **Single cell RNA-Seq:** 37,000 cells (range: 15,000–40,000) were loaded into one well of 10X Genomics droplet-based scRNA-Seq system, and libraries were constructed by following the manufacturer’s protocol “Chromium Next GEM Single Cell 3’ Reagent Kits with Feature Barcoding technology for Cell Surface Protein, v3.0 or v3.1”. The libraries were sequenced for 546 million paired-end fragments on average for cDNA and 96 million fragments for Hashtags using either Illumina NextSeq 500 or NovaSeq 6000. The number of recovered cells was 12,436 on average (ranging 5208–44,126)(Supplementary Data 10). **Variations on the main protocol:** 89 of 172 samples were treated using the procedure described above (“protocol 4”). 28 Samples were treated using “protocol 1” with the following specificities. After dissociating biopsies into cell suspensions, EC cell fractions were sorted by FACS (BD Biosciences, FACSaria III Cell Sorter) to enrich living EC and intraepithelial lymphocytes (IEL) using anti-human CD326 (Biolegend, 324226), CD45 (Biolegend, 304037), CD19 (Biolegend, 363034) and CD11b antibodies (Biolegend, 301348) along with Zombie Green Fixable Viability Dye (Biolegend, 423112). In parallel, LP cell fractions were sorted for living lymphocytes (LP-LC) and myeloid cells (LP-MC) by staining with anti-human CD326, CD45 (Biolegend, 304032), and CD11b antibodies and Zombie Green Fixable Viability Dye. The 12 fractions of sorted cells (EC, IEC, LP-LC and LP-MC for 3 locations) were labeled using distinct TotalSeq-A anti-human Hashtag antibodies (Biolegend), and all cell fractions loaded together on a single well of 10X Genomics Chromium system. 36 samples were treated using “protocol 2” with the following specificities: living cells from EC and LP fractions of cells were enriched using EasySep Dead Cell Removal Annexin V kit (Stemcell Technologies, 17899). 19 samples were treated using “protocol 3”, with following specificities: living cells from EC and LP fractions of cells were enriched using FACS by staining cells with Zombie Green Fixable Viability Dye. **General precautions:** Samples were manipulated on ice unless described, and using low retention filter tips (e.g., RAININ, 30389213). Cell strainers were dipped in FBS before use. Samples retained on a cell strainer were transferred by flipping the cell strainer on a collection tube and flushing it with a buffer.

Preprocessing of scRNA-Seq data. Raw sequencing data were pre-processed with the Cellranger software version 7.1.0 with standard parameters and GRCh38 (Ensembl release 103) human genome build as reference. Processed counts were further analyzed in R (version 4.3.1) within the Seurat tools ecosystem (Seurat version 4.1.3²⁴). For each sample, mRNA read counts and hashtag barcode read counts were loaded to R and quality-checked. We excluded (i) hashtag barcodes with <300 reads per individual, (ii) cells with $\geq 50\%$ mitochondrial reads, (iii) cells with <200 different genes expressed, (iv) cells with <5 hashtag barcode reads. Demultiplexing of cells was done using two different algorithms implemented in Seurat: HTODemux⁷⁹ (positive.quantile 0.999, clusterization function - kmeans) and MULTISEQDemux⁷⁷ with automated threshold finding. Cells identified as singlets of the same tissue type by the two methods were kept for further analysis. To meet the RAM usage requirements, variable features selection was done in five sample batches with the “mean-var.plot” algorithm (3000 features per batch). We retained the intersection between the five batches. We then integrated the sample batches in the space defined by the first 50 principal components using Harmony²¹. The UMAP plots shown throughout the manuscript (f.i. Fig. 2D, E) are in this coordinate system. Yet, to better differentiate cell types and improve the clustering, we further split the data in two stages. We first used Hashtag information to separate cells by anatomical location (IL, TC, RE). Within each anatomical location we repeated the variable feature selection and integration process. In each of these three sub-datasets, we identified cell clusters with the Louvain algorithm. Based on marker gene expression, we assigned clusters to three groups: immune (expressing *PTPRC*), epithelial (expressing

EPCAM), and endothelial plus other cells (expressing *VWF*, *PECAM1*, *CDHS*, or not included in previous groups). Within each one of these nine sub-groups, we again repeated variable features selection, integration, and clustering (Louvain clustering algorithm with resolution parameter 1.5). This yielded a total of 276 clusters across the nine data sets (SFig. 5).

Constructing a hierarchical tree of cell clusters. We computed, for each of the 276 location- and cell-type specific cell clusters (obtained as described above), the mean coordinate vector in the space of the first 50 Harmony coordinates. Then we computed the Euclidean distance between these vectors followed by hierarchical clustering (function *hclust*, stats R package, “complete” agglomeration method). The final dendrogram was constructed with the *dendextend*⁸⁰ R package.

Cis-eQTL analysis. We performed eQTL analysis for each leaf and node in the hierarchical tree, provided that the median cells per patient was ≥ 5 , and that number of patients with cells in the leaf/node was ≥ 30 . This left 401 leaves/nodes for eQTL analysis. Within each analyzed leaf or node, all cells were treated as pseudo-bulk, i.e., as if all reads were derived from one mega-cell. Resulting gene expression data were normalized using DESeq2, residualized for age, sex, and five genotype PCs, and corrected for hidden confounders utilizing the probabilistic estimation of expression residuals (PEER) algorithm³⁰. We restricted eQTL analysis to genes expressed in more than 10% of samples and with mean proportion of reads $> 5 \times 10^{-7}$. eQTL analyses, window-specific multiple testing correction of nominal *p*-values, and computation of leaf/node-specific *q*-values followed the same procedures used for blood.

Agglomerating cis-eQTL in gene-specific and across-genes cis-acting regulatory modules (RM). The construction of cis-acting regulatory modules (RM) was done in a similar way as for the circulating immune cell populations, across all 276 leaves and 275 nodes of the hierarchical tree. For non-genome-wide-significant eQTL to be added to modules, their window-corrected *p*-value had to be ≤ 0.012 , as this yielded mergers with FDR ≤ 0.05 (see DAP-EAP part, hereafter).

Cell type annotation and cell type to hierarchical tree map. Cell type annotation was largely done by visually inspection of the expression profiles of 49 cell type-specific gene signatures obtained from the literature^{22–29} using the Seurat *AddModuleScore* function⁸¹, and the Azimuth human PBMC reference mapping program for immune cells²⁴, on our global UMAP (i.e., all 293,801 QC-ed cells) (Fig. 2D). In essence, the distribution of a cell type-specific gene-signature on the UMAP was confronted with the distribution of the cells from each node/leaf of our hierarchical tree, looking for best matches. A given cell-type was assigned to the best matching node (and hence all descendent nodes/leaves). We further distinguished location-specific (i.e., ileum, colon and rectum) nodes and leaves within cell type-specific sections of the tree. The workflow and outcome of this analysis are shown in SFig. 8 and Supplementary Data 11.

Assigning regulatory modules to intestinal cell types. Regulatory modules encompass one or more EAP that can be active in one or more nodes/leaves of the hierarchical tree. If all nodes/leaves in which a module is active belong to the same cell type (see previous section), the module was assigned to that cell type (and possibly anatomical location within cell type). If nodes/leaves belonging to a module correspond to multiple cell types, the module was assigned to one of 29 supergroups, listed in Supplementary Data 11 and 17. As an example, if a module was active in CD4 and CD8, it was assigned to the T lymphocyte supergroup.

Exploring the distribution of the number of nodes/leaves in which regulatory modules are active. As for blood, each module is characterized by a vector of 0's and 1's informing us about the nodes/leaves in which the module is active. In the case of the intestinal scRNA-Seq data, the length of the vector is 155 leaves + 246 nodes = 401 elements. The length of the vector is less than the sum of the total number of leaves and nodes in the module, because some nodes/leaves didn't have any active module in them. The sum of 1's in a module vector, i.e., the total number of leaves/nodes in which the module is active, is what we are looking at in Fig. 3C. More specifically, we are looking at the distribution of this sum across all modules. Conversely, the activity of a leaf/node can be summarized by a vector of 3345 (gene-specific modules) or 3081 (across-gene modules) 0's and 1's, indicating which module is active in the corresponding leaf/node. To verify whether the observed distribution differs from the expected one, assuming that the activity of a module in a given leaf/node is independent of its activity in other leaves/nodes, we assigned the 0's and 1's of a given leaf/node randomly to the modules, i.e., we randomly permuted module id within leaf/node. We then summed the number of 1's across the modules and examined the distribution of this sum across modules (gray bars in Fig. 3C). We know that the activities of a module in adjacent nodes/leaves in the tree are not independent, as these have cells in common. To properly evaluate the statistical significance of the apparent “overdispersion” of module activity (too many modules either active in few nodes/leaves, or in many nodes/leaves), we restricted to the analysis to “parent” nodes of the 49 distinct cell types, selected such that no cell could be part of more than one such cell type (i.e., none of the selected nodes is ancestor of any other one). We quantified the dispersion for the real data, as the variance of the sum of active nodes across modules. We then performed “permutations”, in the sense that—for a given node—the activity states (1's and 0's) were permuted between modules. This was repeated for all nodes, and the dispersion of that permutation computed as the variance of the sum of “active” nodes across modules. The statistical significance of the real dispersion was then defined as the proportion of permutations that yielded as large or larger variance than the real data.

3D plots of cis-eQTL activity. We developed an application to visualize the activity of an eQTL of interest on a 3D UMAP plot. Briefly, for each cell in the dataset, we identified the 100 nearest neighbors in the space defined by the 50 first expression principal components using the Annoy algorithm [<https://github.com/spotify/annoy>] implemented in the Seurat *FindNeighbors* function. We eliminated cell-centered neighborhoods encompassing cells from only one individual, less than five cells with non-null expression for the e-gene of interest, and MAF < 0.05 for the eQTL's top SNP amongst the individuals with cells in the neighborhood. We then performed eQTL analysis in the remaining neighborhoods, one at a time, using a mixed model^{82,83} including allele dosage (for the eQTL's top SNP) as fixed regression, and individual as random effect. For each neighborhood, we then multiplied $-\log(p\text{-value})$ of the eQTL effect by the sign of the regression coefficient (β), assigned it to the cell defining the neighborhood, and plotted it as the third, *z* dimension of a 3D plot, at the *x-y* coordinate position corresponding to the position of the reference cell in 2D UMAP space.

Merging blood and intestinal cis-eQTL modules reveals eQTLs that are specific for gut-resident immune cells

Module construction. Constructing modules integrating blood and intestinal data followed the same procedure as for the blood- and gut-specific modules. Requirements for an EAP to join a module were the same as before, i.e., $|\theta| \geq 0.6$ for both blood and gut EAP, $p_\theta \leq 0.05$ for both blood and gut, $p_{\text{eQTL, window adj}} \leq 0.12$ for blood and ≤ 0.012 for gut. These thresholds ensured an FDR ≤ 0.05 in the corresponding DAP-EAP confrontations (see hereafter).

Test of independence of cell type annotation in blood and gut. Modules were assigned to cell types separately for blood cell type populations (obvious) and intestinal cell type populations (using the same approach as for the intestinal modules) (Supplementary Data 17). One expects a certain degree of coherence with regards to cell type assignment in both datasets. As an example, modules that are assigned to lymphocytes in blood are expected to be assigned to the lymphoid compartment in the intestine as well. We verified this concordance by performing statistical tests of independence. Cell types were grouped in a limited number of mutually exclusive “supergroups” (Blood: lymphoid, myeloid other than granulocytes, granulocytes, multiple cell types, undetected; Gut: lymphoid, myeloid, enterocyte precursor, mature enterocyte, stromal, multiple cell types, undetected; see also Fig. 4B). We first analyzed the 2170 modules that were assigned to a supergroup (hence active) in both blood and gut. Our aim was to test whether specific supergroup combinations (f.i., active in lymphoid cells in blood and gut) occurred more often than expected by chance, where “chance” corresponds to the product of the frequencies of the corresponding supergroups in blood and gut. Thus, we determined the proportion of modules assigned to each supergroup separately for blood (p_j^{Blood}) and gut (p_j^{Gut}). We then performed a goodness-of-fit test, for each ij combination, using the `chisq.test(c( $n_1$ ,  $n_2$ ), c( $p_1$ ,  $p_2$ ))` function in R. In this, n_1 corresponds to the number of modules in the corresponding supergroup combination ij (f.i. lymphoid in blood and lymphoid in gut), n_2 corresponds to $2170 - n_1$, p_1 corresponds to $p_i^{\text{Blood}} \times p_j^{\text{Gut}}$, and p_2 corresponds to $1 - p_1$. Thus, as an example, we test whether the number of modules that are active in lymphoid cells in blood as well as in lymphoid cells in gut (n_1), differs significantly from the number predicted by the product (hence independence) of the proportion of blood modules active in lymphoid cells times the proportion of gut modules active in lymphoid cells ($(n_1 + n_2) \times p_{\text{lymphoid}}^{\text{Blood}} \times p_{\text{lymphoid}}^{\text{Gut}}$).

We then asked the question whether modules that were active in a specific supergroup in gut, had more (or less) chance to be inactive in blood. For that we considered the 4579 modules active in the gut, and determined, for those, the proportions that were active in the different gut supergroups (p_j^{Gut}), as well as the proportion that were not active in blood ($p_{\text{ND}}^{\text{Blood}}$). We then tested, for each supergroup j , whether the numbers that were active in the gut, yet not in the blood, accounted for a higher (or lower) proportion than expected by chance, where chance corresponds to product of the frequency of the corresponding supergroup in gut (p_j^{Gut}), and the probability for a gut module to be inactive in blood ($p_{\text{ND}}^{\text{Blood}}$). We did this using the `chisq.test(c( $n_1$ ,  $n_2$ ), c( $p_1$ ,  $p_2$ ))` function in R. In this, n_1 corresponds to the number of modules that are active in gut supergroup j (f.i. lymphoid cells) yet not active in blood, n_2 corresponds to $4579 - n_1$, p_1 corresponds to $p_j^{\text{Gut}} \times p_{\text{ND}}^{\text{Blood}}$, and p_2 corresponds to $1 - p_1$. Finally, we repeated the equivalent analysis for the 22,336 modules that were found to be active in blood yet not in the gut.

FLEX single-cell RNA sequencing. Gut biopsies from terminal ileum, transverse colon and rectum (~30 mg each) were collected in 40 ml of ice-cold Advanced DMEM/F12 medium (Gibco 12634-010). They were processed within 30 min using Chromium Next GEM Single Cell Fixed RNA Sample Preparation Kit (10x Genomics, 1000414) by following the manufacturer’s protocol: Tissue Fixation & Dissociation for Chromium Fixed RNA Profiling (CG000553/RevB). Briefly, the biopsy was fixed at 4 °C for 23 h in 2 ml of fixation buffer supplemented with 4% formaldehyde (Thermo Fisher, BP531-25). The fixed biopsy was transferred in a Miltenyi dissociation C tube (Miltenyi Biotec, 130-093-237) containing 2 ml of prewarmed RPMI 1640 medium (Corning, 10-040-CV) supplemented with 0.2 mg/ml of Liberase TL (Roche, 5401020001). The sample was dissociated using a gentleMACS Octo dissociator with heaters (Miltenyi Biotec) with a program (at 37 °C, spin at 50 rpm for 30 min, spin at 1000 rpm for 45 s, and spin at –1000 rpm

for 45 s). The cell suspension was passed through a cell strainer with 30 µm pores (Sysmex, 04-004-2326). The fixed nuclei were collected by a centrifugation and resuspended in Quenching buffer (10x Genomics). Nuclear density was estimated using Countess 3 fluorescence cell counter (Fisher) with AO/PI Cell Viability Kit (Westburg, LB F23001). Approximately 200,000 nuclei were obtained from each of gut biopsies. The samples were stored at –80 °C until a library construction after adding Enhancer buffer (10x Genomics) and 10% glycerol (Sigma, G5516) by following the manufacturer’s protocol. Peripheral blood from the same donor was collected using a K2-EDTA blood collection tube (BD Vacutainer, 367525) and processed by following the manufacturer’s protocol: Blood Fixation and Cell Isolation for Chromium Fixed RNA Profiling (CG000721/RevB). Briefly one milliliter of blood was mixed with 4 ml of 1x PBS (Corning, 21-040-CV) supplemented with 4% formaldehyde and fixed at room temperature for 23 h. Leukocytes were enriched using EasySep RBC Depletion Reagent (STEMCELL Technologies, 18170). We obtained approximately 7 million of fixed leukocytes from 1 ml of blood sample. The fixed leukocytes were stored at –80 °C until a library construction. A sequencing library was constructed by following the manufacturer’s protocol: Chromium Fixed RNA Profiling Reagent Kits for Multiplexed Samples (CG000527/RevE) using Chromium Fixed RNA Human Transcriptome Kit (10x Genomics, 1000475), Chromium Next GEM Chip Q Single Cell Kit (10x Genomics, 1000251) and Dual Index Kit TS Set A (10x Genomics, 1000251). Briefly, the stored fixed samples (~200,000 fixed nuclei for IL, TC, and RE and two million fixed leukocytes) were hybridized with probe pairs detecting 18,099 human mRNA as well as samples. After washing, equal numbers of nuclei/cells from each sample were mixed (total 26,000 cells) and run for a single GEM (Gel beads-in-emulsion) reaction. The library was sequenced for 250.9 million reads using Illumina NovaSeq 6000 (S4 Reagent Kit v1.5). The data was analyzed using Cellranger multi (10x Genomics, version 8.0.1) with a reference (refdata-gex-GRCh38-2024-A, Ensembl v110 annotations) and Seurat (version 5.3.0). Briefly, high-quality cells were kept based on QC metrics (mitochondria read ≤ 10% and unique feature count ≤ 10000). The data were normalized using the `sctransform` function, subjected to dimensionality reduction by PCA (`RunPCA(object, npcs = 100, assay = “SCT”)`). The four datasets (IL, TC, RE, BL) were integrated using `Harmony` (`IntegrateLayers(object, method = “HarmonyIntegration”, orig.reduction = “pca”, new.reduction = “harmony”, assay = “SCT”)` and `JoinLayers(object[["RNA"]])`). Cells were clustered by a graph-based approach and a modularity optimization (`FindNeighbors(object, dims = 1:30, reduction = “harmony”, assay = “SCT”)` and `FindClusters(object)`) and visualized using UMAP non-linear dimensional reduction (`RunUMAP(object, dims = 1:30, reduction = “harmony”, assay = “SCT”)`). A total of 36 clusters were annotated into 17 cell types based on expression of marker genes (Stromal: *BMP4*, *ADAMDECI*; Endothelial: *VWF*, *PLVAP*; Epithelial: *EPCAM*, *KRT19*; T: *CD3D*, *CD3E*; Naïve CD4 CD8: *CCR7*, *SELL*, *LEFT1*; Memory CD4: *CD4*; Memory CD8: *CD8A*, *GZMA*, *GZMK*; Treg: *FOXP3*, *IKZF2*, *RTKN2*; NK gdT: *NCAM1*, *FCER1G*, *TRDC*; dnT: *TRAC*; B: *CD19*, *MS4A1*, *PAX5*; Naïve B: *IL4R*, *FCER2*, *IGHM*, *IGHD*; Memory B: *AIM2*, *IGHA1*; Plasmacyte: *MZB1*, *XBPI*, *IGKC*; Myeloid: *LYZ*, *TYROBP*, *AIFI*; CD14 Monocyte: *CD14*, *VCAN*, *S100A8*; Macrophage: *GPNMB*, *CD36*, *CD68*, *AIFI*; Dendritic: *CST3*, *FLT3*; Neutrophil: *NAMPT*, *FPRI*, *FCGR3B*; MAST: *CPA3*, *FCERIA*). The experimental protocol was approved by the Ethics committee of CHU Liège (reference number: 2017/214).

Identifying cis-eQTL driving inherited predisposition to IBD

Comparing EAP and DAP using theta. We performed a colocalization analysis of our EAP with IBD, CD and UC risk loci coming from de Lange et al.³². We lifted their data over from GRCh37 to GRCh38 and defined disease association patterns (DAP) in genomic locations where a risk locus was described in at least one of the diseases and where at least one variant had a p -value less or equal to 10^{-5} . We established the limits

of the DAPs manually to surround the peaks. In total, we tested 455 DAP corresponding to 206 risk loci (157 for CD, 173 for IBD, and 125 for UC). Some DAP were subdivided into two or three parts. We evaluated the similarity between DAP and EAP using θ following ref. 15, for all EAP for which the top eQTL SNP was within the boundaries of the analyzed disease interval. θ was computed for all variants located within the limits of the disease interval, provided that their nominal association p -value ≤ 0.05 either in the EAP, DAP or both. To determine the statistical significance of θ (i.e., p_θ), we performed up to (adaptive) 10,000 permutations (without replacement) of gene expression levels before recomputing θ . We defined p_θ as the proportion of permutations where the obtained $|\theta|$ is greater or equal to the observed.

To further define appropriate thresholds to declare a DAP-EAP match of interest, we repeated the full, genome-wide *cis*-eQTL analysis, both in blood and gut (i.e., in the 27 cell types and 551 leaves/nodes), after randomly disconnecting (i.e., permuting) the genotype (all variants) and expression (all genes) vectors. Genotype and expression vectors were maintained unaltered (hence LD structure on the one hand, and correlation structure between gene expression on the other hand, were conserved). We then repeated the colocalization exercise between all 455 DAPs, and the overlapping EAP obtained with the permuted data exactly as we did with the real data. Assume that a DAP matches a real EAP with $|\theta| = x \geq 0.6$, $p_\theta = y$, and $p_{\text{eQTL}, \text{window adj}} = z$, we would determine how many matches satisfying $|\theta| \geq x$, $p_\theta \leq y$, and $p_{\text{eQTL}, \text{window adj}} \leq z$, were obtained with the real data ($=N_R$) and how many with the permuted data ($=N_P$). The FDR of the corresponding DAP-EAP match was then computed as N_P/N_R . We defined two FDR thresholds of significance: Tier 1 ($\text{FDR} \leq 0.05$) and Tier 2 ($0.05 < \text{FDR} \leq 0.10$).

Comparing the yield of DAP-EAP matches between IBD and dilated cardiomyopathy (DCM). To evaluate the relevance of the DAP-EAP matches observed for IBD, we repeated the same analyses for dilated cardiomyopathy (DCM). DCM has a genetic architecture that is comparable to that of IBD, yet it is expected to be driven primarily by eQTL operating in the cardiovascular system rather than in immune cells or in the gut. Thus, we would expect that the yield of matching DAP-EAP in the CEDAR-2 set will be substantially lower for DCM than for IBD. Using the data from a large DCM GWAS conducted in Europeans³³, we defined the boundaries of 58 DCM risk loci using the exact same procedure as the one used for IBD. We did not split risk loci in sub-loci and excluded, for the comparison, these splitting events from the IBD results. We then conducted DAP-EAP colocalization using our θ -based approach following the exact same procedure and significance thresholds as for IBD. We compiled the number of DAP-EAP confrontations realized for DCM as well as for IBD, CD and UC. The results from IBD, CD and UC were kept separate (rather than aggregated as in Fig. 5A) for comparison. The outcome of interest was the proportion of realized DAP-EAP confrontations that yielded a match. The success rates of DCM and IBD, DCM and CD, DCM and UC were compared using a 2×2 contingency table test implemented with the `chisq.test(M)` function in R.

Defining the activity profile of DAP-EAP matching, gene-specific regulatory modules. We selected, by risk locus, disease, and gene, all DAP-matching EAP with $\text{FDR} \leq 0.10$. All selected EAP were considered to be part of the same gene-specific module as they all matched a common DAP. This module was then assigned to one of five categories in blood (lymphoid, myeloid, granulocytes, multiple cell types or undetected) depending on the cell type(s) in which the corresponding eQTL were detected, and to one of seven categories in the gut (lymphoid, myeloid, enterocyte precursors, mature enterocytes, stromal, multiple cell types, undetected) depending on the nodes/leaves of the hierarchical tree in which the corresponding eQTL were detected. For the gut, and as before, the eQTL was assigned to the most recent

common ancestor (MRCA) of all active nodes/leaves, and the corresponding MRCA node/leave was then assigned to one of the seven categories. We then compared the distribution of the DAP-EAP matches in this 5×7 grid with the equivalent distribution of all monogenic modules. We performed a goodness-of-fit test on a cell-by-cell basis as before.

Comparing the proportion of genes that upon genetic perturbation affects susceptibility to colitis in rodent models between DAP-matching and not-DAP-matching protein-coding e-genes. If the genetic perturbation of genes selected by virtue of a DAP-EAP match have a higher probability to affect the susceptibility of mice to induce colitis (whether increased or decreased susceptibility) than randomly selected genes, this would support the fact that the DAP-matching e-genes are enriched in causative genes for IBD. To test this as objectively as possible, we submitted our list of 481 protein-coding DAP-matching e-genes, mixed with 481 protein-coding non-DAP-matching genes, “blindly” to two scientists trained in the life sciences. We asked them to review the literature for all these genes with the objective to find reports of an effect of gene perturbation of the corresponding genes on the susceptibility to induced colitis in a rodent model. We defined genetic perturbation as: constitutive or conditional germ-line knock-out, constitutive or conditional overexpressing germ-line transgene, or knock-out in immune cells transferred to RAG^{-/-} mice. We did not consider knock-down with siRNA or pharmacological perturbation, as the specificity of those interventions were considered to be more difficult to prove. The 481 “control” genes were matched for expression profile across the 27 blood cell types, and the 43 gut cell types. For each DAP-matching e-gene we identified as non-DAP-matching control, the e-gene (hence with a *cis*-eQTL) that was the closest (Euclidian distance) in multi-tissular expression space. The statistical significance of the difference in “success” rate (i.e., the proportion of genes with such report) between DAP-matching and non-DAP-matching e-genes was determined using a standard goodness-of-fit test.

Comparing the number of DAP-EAP matches with the CEDAR2 cell-type specific information from ≤ 200 individuals versus the eQTLGen PBMC information from 35 K individuals. Summary statistics from eQTLGen, a meta-analysis of *cis*-eQTL results from 37 studies of blood and PBMC samples totaling 35 K individuals, were downloaded⁵⁸. We lifted their data over from GRCh37 to GRCh38. When arithmetic underflow was observed for the p -values, the $\log_{10} p$ -values were predicted from the z -scores using a local polynomial regression. We compared their 871 EAP (significant or not) with the 455 DAPs if the top variants were located within the disease interval. However, it was not possible to recompute the missing part of the EAP nor to compute a p -value for θ . Colocalized EAP correspond to EAP that have a $|\theta| \geq 0.6$ with a DAP (hence more permissive than the analysis of the CEDAR2 dataset).

Identifying drug candidates for repurposing

Effects of IBD on gene expression in circulating immune cell populations. We collected blood from 55 active CD patients and 100 healthy controls, fractionated 26 circulating immune cell populations by FACS or MACS following the same protocol as for eQTL analysis, and performed RNA-Seq on the 26 fractionated cell populations as well as on PBMC as for eQTL analysis (*vide supra*), all as one batch. Differential expression analysis between controls and patients was performed for all expressed genes by cell type using the DESeq2 R package^{84,85}, with the `apeglm` method for effect size shrinkage⁸³. Genes with a fold change ≥ 1.5 and $\text{FDR} \leq 0.05$ were considered differentially expressed (Supplementary Data 27). Supplemental Table 22 lists, for 556 genes with DAP-matching EAP, the blood cell types for which CD has a significant effect on gene expression ($\text{FDR} \leq 0.05$, $\log_2$ fold

change ≥ 1.5). For each such significant effect, Supplementary Data 22 provides cell type, effect (log2 fold change), and FDR. If, for a given gene, the majority sign of the effects (across cell type) corresponds to the majority sign of theta (across cell types) (f.i. CD most often increases the gene's expression, and the theta between the gene's EAPs and the DAP is most often positive), the effect is labelled consistent ("TRUE"), otherwise not ("FALSE").

Effects of IBD on gene expression level in intestinal cells. Were obtained from Supplemental Table 3 in Kong et al.²⁸. Kong et al. performed scRNA-Seq on biopsies from terminal ileum (TI) and colon (CO) of 25 healthy controls and 46 CD patients. For active CD patients, biopsies were collected from inflamed as well as non-inflamed regions. Cells were assigned to 65 cell types/states. Differential expression (DE) analysis was conducted by location (TI or CO), cell type, and contrast (inflamed regions in CD patients versus healthy controls, and non-inflamed regions in CD patients versus healthy controls), using MAST⁸⁶ in either discrete mode (expression rate) or continuous mode (expression level). Supplementary Data 22 (this work) lists, for 556 genes with DAP-matching EAP, the cell types/locations for which CD has a significant effect on gene expression (discrete and/or continuous FDR ≤ 0.001 ; if both FDR ≤ 0.001 , the best is shown). For each such significant effect, Supplementary Data 22 provides effect (log2 fold change), FDR, cell type, and location. If, for a given gene, the majority sign of the effects (across cell type/locations) corresponds to the majority sign of theta (across cell types) (f.i. CD most often increases the gene's expression, and the theta between the gene's EAPs and the DAP is most often positive), the effect is labelled consistent ("TRUE"), otherwise not ("FALSE").

Effects of IBD on circulating (plasma) protein levels. Were obtained from Supplementary Data 18 in Eldjarn et al.⁴⁸. Eldjarn et al. analyzed the association with IBD, of concentrations of 2931 circulating proteins (Olink Explore 3072 assay) in ~49 K individuals from the UK Biobank (including 900 IBD cases), and of 4,907 circulating proteins (SomaScan v4 assay) in ~36 K Icelanders (including 618 IBD cases), using logistic regression. Supplementary Data 22 (this work) lists, for 556 genes with DAP-matching EAP, significant (nominal p -value ≤ 0.01) associations between IBD status and plasmatic protein concentrations. For each such significant association, Supplementary Data 22 provides assay, effect (odds ratio), and nominal p -value. If, for a given gene, the direction of the effect(s) corresponds to the majority sign of theta (across cell types) (f.i., CD increases the protein abundance, and the theta between the gene's EAPs and the DAP is most often positive), the effect is labelled consistent ("TRUE"), otherwise not ("FALSE").

Identifying IBD drugs targeting e-genes with DAP (IBD) matching EAP. To identify whether any of the 556 e-genes with IBD-matching EAP were targeted by drugs that are or have been used to treat IBD we consulted the list of IBD drugs that was recently established by Vieujean et al.⁵⁰.

Identifying non-IBD drugs targeting e-genes with DAP (IBD) matching EAP. To identify whether any of the 556 e-genes with IBD-matching EAP were targeted by drugs that are or have been used to treat diseases other than IBD, we consulted the Open Targets⁴⁹ database. Open Targets was queried for targeting drugs by e-gene. For each drug, we extracted name, type of drug, mechanism of action, and the most advanced clinical trial by disease. Corresponding data for entrectinib were compiled from references^{51–54}.

Reporting summary

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

Data availability

All data are available within the constraints imposed by GDPR (General Data Protection Regulation). Processed data (nominal, permutational, and conditional cis-eQTL summary statistics, as well as gene expression data) have been deposited in open access on Zenodo under accession numbers: <https://zenodo.org/records/21457760>, <https://zenodo.org/records/21495473>, <https://zenodo.org/records/21535308>, <https://zenodo.org/records/21538172>, <https://zenodo.org/records/21538768>, <https://zenodo.org/records/21538968>, <https://zenodo.org/records/21539705>, <https://zenodo.org/records/21539811>, <https://zenodo.org/records/21539922>, <https://zenodo.org/records/21540023>, <https://zenodo.org/records/21540192>, <https://zenodo.org/records/21540504>, <https://zenodo.org/records/21540671>, <https://zenodo.org/records/21540784>. Seurat objects used for analysing 10x Genomics 3' GEM and FLEX single-cell RNA-seq data were deposited in open access on Zenodo under accession number: <https://zenodo.org/records/21373443>. Source data for figures are provided with this paper. Source data are provided with this paper.

References

- Busse, R. et al. Tackling chronic disease in Europe: strategies, interventions and challenges. in *European Observatory on Health Systems and Policies. Observatory Studies Series n°20* (World Health Organization, 2010).
- Schumacher, A., Gassmann, O. & Hinder, M. Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* **14**, 105 (2016).
- Jakab, M. et al. Health systems respond to noncommunicable diseases: time for ambition. *Copenhagen: WHO Regional Office for Europe* (2018).
- Polubriaginof, F. C. G. et al. Disease heritability inferred from familial relationships reported in medical records. *Cell* **173**, 1692–1704 (2018).
- Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
- Abdelloui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: realizing the promise. *Am. J. Hum. Genet.* **110**, 179–194 (2023).
- King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
- Burgess, S. et al. Using genetic association data to guide drug discovery and development: review of methods and applications. *Am. J. Hum. Genet.* **110**, 195–214 (2023).
- Trajanoska, K. et al. From target discovery to clinical drug development with human genetics. *Nature* **620**, 737–745 (2023).
- Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. *Nature* **629**, 624–629 (2024).
- Sazonovs, A. et al. Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. *Nat. Genet.* **54**, 1275–1283 (2022).
- Giambertolomei, C. et al. Bayesian test for colocalization between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
- Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
- Momozawa, Y. et al. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. *Nat. Commun.* **9**, 2427 (2018).

16. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
17. Umans, B. D., Battle, A. & Gilad, Y. Where are the disease-associated eQTLs? *Trends Genet.* **37**, 109–124 (2021).
18. Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat. Genet.* **55**, 1866–1875 (2023).
19. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
20. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
21. Korsunsky, I. et al. Raychaudhuri S. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
22. Smillie, C. S. et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**, 714–730.e22 (2019).
23. Franzén, O., Gan, L.-M., & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019: baz046 <https://doi.org/10.1093/database/baz046> (2019).
24. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
25. Burclaff, J. et al. A proximal-to-distal survey of healthy adult human small intestine and colon epithelium by single-cell transcriptomics. *Cell Mol. Gastroenterol. Hepatol.* **13**, 1554–1589 (2022).
26. Ishikawa, K. et al. Identification of quiescent LGR5+ stem cells in the human colon. *Gastroenterology* **163**, 1391–1406 (2022).
27. Hickey, J. W. et al. Organization of the human intestine at single-cell resolution. *Nature* **619**, 572–584 (2023).
28. Kong, L. et al. The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity* **56**, 444–458.e5 (2023).
29. Krzak, M. et al. Single-cell RNA sequencing reveals dysregulated cellular programmes in the inflamed epithelium of Crohn's disease patients. *medRxiv* <https://doi.org/10.1101/2023.09.06.23295056> (2024).
30. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
31. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J. K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat. Methods* **12**, 1061–1063 (2016).
32. de Lange, K. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
33. Jurgens, S. J. et al. Genome-wide association study reveals mechanisms underlying dilated cardiomyopathy and myocardial resilience. *Nat. Genet.* **56**, 2636–2645 (2024).
34. Yildiz, Ç et al. A possibly new autoinflammatory disease due to compound heterozygous phosphomevalonate kinase gene mutation. *Jt. Bone Spine* **90**, 105490 (2023).
35. Zou, D., Zhou, S., Wang, H., Gou, J. & Wang, S. Knee joint swelling at presentation: a case of pediatric Crohn disease with a TNFAIP3 mutation. *Pediatrics* **146**, e20193416 (2020).
36. Taniguchi, K. et al. Novel TNFAIP3 microdeletion in a girl with infantile-onset inflammatory bowel disease complicated by a severe perianal lesion. *Hum. Genome Var.* **8**, 1 (2021).
37. Roncagalli, R. et al. The scaffolding function of the RLTPR protein explains its essential role for CD28 co-stimulation in mouse and human T cells. *J. Exp. Med.* **213**, 2437–2457 (2016).
38. Magg, T. et al. CARMIL2 deficiency presenting as very early onset inflammatory bowel disease. *Inflamm. Bowel Dis.* **25**, 1788–1795 (2019).
39. Bosa, L. et al. Novel CARMIL2 loss-of-function variants are associated with pediatric inflammatory bowel disease. *Sci. Rep.* **11**, 5945 (2021).
40. Liu, F. et al. TLR-induced SMPD3 defects enhance inflammatory response of B cell and macrophage in the pathogenesis of SLE. *Scand. J. Immunol.* **86**, 377–388 (2017).
41. Al-Rashed, F. et al. Neutral sphingomyelinase 2 regulates inflammatory responses in monocytes/macrophages induced by TNF- α . *Sci. Rep.* **10**, 16802 (2020).
42. Li, S. et al. Sedanolid alleviates DSS-induced colitis by modulating the intestinal FXR-SMPD3 pathway in mice. *J. Adv. Res.* <https://doi.org/10.1016/j.jare.2024.03.026> (2024).
43. Yu, M. et al. International inflammatory bowel disease genetics consortium; Anderson CA, Daly MJ, Huang H. Cystic fibrosis risk variants confer protection against inflammatory bowel disease. *medRxiv* <https://doi.org/10.1101/2024.12.02.24318364> (2024).
44. Hallows, K. R., Kobinger, G. P., Wilson, J. M., Witters, L. A. & Foskett, J. K. Physiological modulation of CFTR activity by AMP-activated protein kinase in polarized T84 cells. *Am. J. Physiol. Cell Physiol.* **284**, C1297–C1308 (2003).
45. Prieto, S. et al. CDK8 and CDK19 act redundantly to control the CFTR pathway in the intestinal epithelium. *EMBO Rep.* **24**, e54261 (2022).
46. Belinky, F. et al. PathCards: multi-source consolidation of human biological pathways. *Database* **2015**, bav006 (2015).
47. Dimitrov, G. et al. Congenital Sodium Diarrhea by mutation of the SLC9A3 gene. *Eur. J. Med. Genet.* **262**, 103712 (2019).
48. Eldjarn, G. H. et al. Large-scale plasma proteomics comparisons through genetics and disease associations. *Nature* **622**, 348–358 (2023).
49. Mountjoy, E. et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet.* **53**, 1527–1533 (2021).
50. Vieujean, S. et al. Understanding the therapeutic toolkit for inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **22**, 371–394 (2024).
51. Liu, D., Offin, M., Harnicar, S., Li, B. T. & Drilon, A. Entrectinib: an orally available, selective tyrosine kinase inhibitor for the treatment of *NTRK*, *ROS1*, and *ALK* fusion-positive solid tumors. *Ther. Clin. Risk Manag* **14**, 1247–1252 (2018).
52. Jin, X. et al. Entrectinib inhibits NLRP3 inflammasome and inflammatory diseases by directly targeting NEK7. *Cell Rep. Med.* **4**, 101310 (2023).
53. Chen, X. et al. NEK7 interacts with NLRP3 to modulate the pyroptosis in inflammatory bowel disease via NF- κ B signaling. *Cell Death Dis.* **10**, 906 (2019).
54. Jiang, Q., Li, M., Li, H. & Chen, L. Entrectinib, a new multi-target inhibitor for cancer therapy. *Biomed. Pharmacother.* **150**, 112974 (2022).
55. Schmiedel, B. J. et al. Impact of genetic polymorphisms on human immune cell gene expression. *Cell* **175**, 1701–1715 (2018).
56. Ota, M. et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006–3021.e17 (2021).
57. Maecker, H. T., McCoy, J. P. & Nussenblatt, R. Standardizing immunophenotyping for the Human Immunology Project. *Nat. Rev. Immunol.* **12**, 191–200 (2012).
58. Vösa, 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).
59. Hu, S. et al. Inflammation status modulates the effect of host genetic variation on intestinal gene expression in inflammatory bowel disease. *Nat. Commun.* **12**, 1122 (2021).
60. Loefflerinckx, C. et al. The identification of blood-derived response eQTL reveals complex effects of regulatory variants on inflammatory and infectious disease risk. *PLoS Genet.* **21**, e1011599 (2025).

61. Nishiyama, N. C. et al. eQTL in diseased colon tissue identifies potential target genes associated with IBD. *Nat. Commun.* **17**, 2736 (2026).
62. Alegbe, T. et al. Cell-type-resolved genetic variation shapes inflammatory bowel disease risk. *Nature* <https://doi.org/10.1038/s41586-026-10627-z> (2026).
63. Connally, N. J. et al. The missing link between genetic association and regulatory function. *eLife* **11**, e74970 (2022).
64. Steinmetz, L. et al. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326–330 (2002).
65. Stern, D. L. Identification of loci that cause phenotypic variation in diverse species with the reciprocal hemizyosity test. *Trends Genet.* **30**, 547–554 (2014).
66. Hugot, J. P. et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
67. Lesage, S. et al. CARD15/NOD2 mutational analysis and genotype-phenotype correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* **70**, 845–857 (2002).
68. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
69. Visscher, P. M. & Haley, C. S. Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. *Theor. Appl. Genet.* **93**, 691–702 (1996).
70. Yan, J., Hedl, M. & Abraham, C. An inflammatory bowel disease-risk variant in INAVA decreases pattern recognition receptor-induced outcomes. *J. Clin. Invest.* **127**, 2192–2205 (2017).
71. Mohanan, V. et al. *C1orf106* is a colitis risk gene that regulates stability of epithelial adherens junctions. *Science* **359**, 1161–1166 (2018).
72. Hamaoui, D. & Subtil, A. ATG16L1 functions in cell homeostasis beyond autophagy. *FEBS J.* **289**, 1779–1800 (2022).
73. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
74. Liu, X., Li, Y. I. & Pritchard, J. K. Trans effects on gene expression can drive omnigenic inheritance. *Cell* **177**, 1022–1034 (2019).
75. Graham, D. B. & Xavier, R. J. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* **578**, 527–539 (2020).
76. Sylvain, A. et al. A cereblon-based glue degrader of NEK7 regulates NLRP3 inflammasome in a context-dependent manner. *Cell Chem. Biol.* **32**, 955–968.e13 (2025).
77. Ongen, H., Buil, A., Anand Brown, A., Dermizakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
78. McGinnis, C. S. et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat. Methods* **16**, 619–626 (2019).
79. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
80. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
81. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
82. Bates, D., Mächler, M., Bolker, B. & Walker, S. “Fitting linear mixed-effects models using lme4.”. *J. Stat. Softw.* **67**, 1–48 (2015).
83. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. “lmerTest package: tests in linear mixed effects models.”. *J. Stat. Softw.* **82**, 1–26 (2017).
84. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
85. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).
86. Finak, G. et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).

Acknowledgements

We thank Sandra Ormenese, Celine Vanwinge and the GIGA-Flow cytometry core facility for their support, as well as Naima Ahariz, Azeddine Bentaib and the other members of the GIGA genomics platform. Souad Rahmouni is a senior research associate of the FRS-FNRS. We are grateful to Yurii Aulchenko for many stimulating discussions.

Author contributions

H.P.: performed experiments, analyzed data, wrote the manuscript. VAP, AM: analyzed data. Y.T.: performed experiments, analyzed data. A.K.: developed the CEDAR2 website. F.F.: supervised statistical analyses. S.V.: provided samples, analyzed data. SR: performed experiments, analyzed literature. M.M., M.W., S.A., S.J., N.F., L.K., M.D., E.D., R.S., K.H., and Y.M.: performed experiments. C.R., C.V.K., O.W., V.L., S.K., M.P., A.C.M., B.S., and J.P.L.: provided samples. W.C.: supervised experiments. E.D.: supervised experiments. E.L.: supervised sample collection, designed experiment. M.G.: designed experiments, analyzed data, supervised analyses, wrote the manuscript. H.T.: designed, supervised and performed experiments, analyzed data, wrote the manuscript. S.R.: designed and supervised experiments, analyzed data, wrote the manuscript. Members of the SYSCID and BRIDGE consortium provided feedback on experiment, results and manuscript.

Funding

This project was conducted with funding from the SYSCID and PerPrev H2020 grants (refs. 733100 and 101156542), the MyQuant (ref. 30770923) and BRIDGE (O.0006.22 – RG3124) projects from the Excellence of Science (EOS) program (FNRS, Fédération Wallonie-Bruxelles and FWO, Flemish Community), the CLIMAX (WELBIO-CR-2022 A) project from WELBIO (Walloon Region), the RHEAQT (T.0096.19) and IBD-GI-Seq (T.0190.19) projects from the FNRS (Fédération Wallonie-Bruxelles), the ARC RHEACT WITH HSPC project from the University of Liège. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-026-76672-4>.

Correspondence and requests for materials should be addressed to Michel Georges or Souad Rahmouni.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

© The Author(s) 2026

SYSCID Consortium

Konrad Aden⁹, Vibeke Andersen¹⁰, Diana Avalos¹¹, Aggelos Banos¹², George Bertsias¹², Marc Beyer¹³, Johanna I. Blase⁹, Dimitrios Boumpas¹², Emmanouil T. Dermitzakis¹¹, Axel Finckh¹⁴, Andre Franke⁹, Gilles Gasparoni¹⁵, Michel Georges^{1,8}✉, Wei Gu¹⁶, Robert Häslér⁹, Mohamad Jawhara¹⁰, Amy Kenyon¹⁷, Christina Kratsch¹⁸, Roland Krause¹⁶, Gordan Lauc¹⁹, Paul A. Lyons²⁰, Massimo Mangino²¹, Neha Mishra⁹, Gioacchino Natoli¹⁷, Marek Ostaszewski¹⁶, Silja H. Overgaard¹⁰, Marija Pezer¹⁹, Jeroen Raes²², Souad Rahmouni¹✉, Benedikt Reiz¹⁸, Elisa Rosati⁹, Philip Rosenstiel⁹, Despina Sanoudou¹², Venkata Satagopam¹⁶, Reinhard Schneider¹⁶, Jonas Schulte-Schrepping¹³, Joachim L. Schultze¹³, Prodromos Sidiropoulos¹², Kenneth GC Smith²³, Signe B. Sørensen¹⁰, Timothy Spector²¹, Doris Vandeputte²², Sara Vieira-Silva²², Aleksandar Vojta¹⁹, Jörn Walter¹⁵, Stefanie Wernat-Herresthal¹³, Vlatka Zoldoš¹⁹

BRIDGE Consortium Inna Afonina²⁴, Rudi Beyaert²⁴, Laure Dumoutier²⁵, Denis Franchimont²⁶, Michel Georges^{1,8}✉, Claude Libert²⁴, Claire Liefferinckx²⁶, Natalia Ferreras Moreno²⁴, Souad Rahmouni¹✉ & Ramnik Xavier²⁷

⁹Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel and University Medical Center Schleswig-Holstein, Kiel, Germany. ¹⁰Diagnostics and Clinical Research Unit, Institute of Regional Health Research, University Hospital of Southern Denmark, Aabenraa, Denmark. ¹¹Institute of Genetics and Genomics of Geneva (IGE3), University of Geneva, Geneva, Switzerland. ¹²Rheumatology and Clinical Immunology, Medical School, University of Crete, Heraklion, Greece/Laboratory of Rheumatology, Autoimmunity and Inflammation, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology (IMBB-FORTH), Heraklion, Greece. ¹³Systems Medicine, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany & Genomics & Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany. ¹⁴Rheumatology Division, Geneva University Hospital (HUG), Geneva, Switzerland & Geneva Center for Inflammation Research (GCIR), University of Geneva (UNIGE), Geneva, Switzerland. ¹⁵Department of Genetics/Epigenetics, Universität Des Saarlandes, Saarbrücken, Germany. ¹⁶Bioinformatics Core, Luxembourg Center for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg. ¹⁷Department of Experimental Oncology, IEO, European Institute of Oncology IRCCS, Milan, Italy. ¹⁸Comma Soft AG, Pützchens Ch 202-204a, Bonn, Germany. ¹⁹Genos Glycoscience Research Laboratory, 10000 Zagreb, Croatia/Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia. ²⁰Department of Medicine, University of Cambridge School of Clinical Medicine, Cambridge, United Kingdom & Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, Cambridge, UK. ²¹Department of Twin Research and Genetic Epidemiology, King's College, London, UK. ²²Laboratory of Molecular Bacteriology, Department of Microbiology and Immunology, KU Leuven, Leuven, Belgium & Center for Microbiology, VIB-KU Leuven, Leuven, Belgium. ²³Cambridge Institute of Therapeutic Immunology and Infectious Disease, Jeffrey Cheah Biomedical Centre, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK. ²⁴Unit of Molecular Signal Transduction in Inflammation, VIB Center for Inflammation Research, VIB, Ghent, Belgium & Ghent University, Department of Biomedical Molecular Biology, Ghent, Belgium. ²⁵Experimental Medicine, Unit de Duve Institute UCLouvain, Brussels, Belgium. ²⁶Department of Gastroenterology, Hepatopancreatology, and Digestive Oncology, HUB Hôpital Erasme, Université Libre de Bruxelles, Brussels, Belgium & Laboratory of Experimental Gastroenterology, Université Libre de Bruxelles, Brussels, Belgium. ²⁷Broad Institute of MIT and Harvard, Cambridge, USA & Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA.

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.