Integrated cell type-specific analysis of blood and gut identifies matching *cis*-eQTLs for 140 IBD risk loci

Hélène Perée

Supervisor: Dr. Souad Rahmouni
Co-supervisor: Pr. Michel Georges

University of Liège
Faculty of Medicine
GIGA-Molecular and Computational Biology
Unit of Animal Genomics

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor in Biomedical and Pharmaceutical Sciences

Academic year 2024-2025

ACKNOWLEDGMENTS

First, I would like to express my heartfelt gratitude to my supervisors, Souad Rahmouni and Michel Georges, for welcoming me into your laboratory and supporting me financially throughout these six years. I deeply appreciate your availability, patience, and dedication to the project. Souad, thank you for believing in me, for your mentorship, and for your constant presence from day one. Michel, I am grateful for your meticulous approach to analysis and for always pushing me forward with kindness. My sincere thanks to both of you for teaching me so much and for allowing me to continue in your lab after my thesis.

I would also like to extend my thanks to the members of my thesis jury, who have been involved from the beginning and have guided me throughout the six thesis committees: Michel Moutschen, Wouter Coppieters, and Édouard Louis. I am equally grateful to those who kindly joined the jury for the private and public defenses: Isabelle Cleynen, Philip Rosenstiel, and Alain Chariot. I consider myself fortunate to have had a multidisciplinary committee with expertise in genetics, immunology, and gastroenterology. Thank you all for the insightful conversation during my private defense.

I am extremely thankful to my colleagues for their help throughout this project. Haruko, thank you for training me in library preparation, for your invaluable contribution to the article, and for all the thoughtful gifts you brought me from Japan. Slava, I appreciate your work on developing the analysis pipeline and your assistance with troubleshooting. Yumie, your support has been invaluable, both professionally and personally, and I miss you greatly. Sophie V., thank you for being a model of kindness and perseverance. Myriam and Marie, I am grateful for your training in cell purification, flow cytometry, and RNA extraction, and I thank you both, along with Sarah and Nicolas, for executing so many of the experiments. Samira, thank you for taking on the challenging task of recruiting study participants. Sophie J., I appreciate you starting this project and inspiring me to complete my thesis. I also want to thank you all, along with Sam, Yasu, Alex Z., Alex T., Salomé, Loïc, and Lev, for creating such a positive atmosphere in the lab. Alex K., I am grateful for your work on developing the web application.

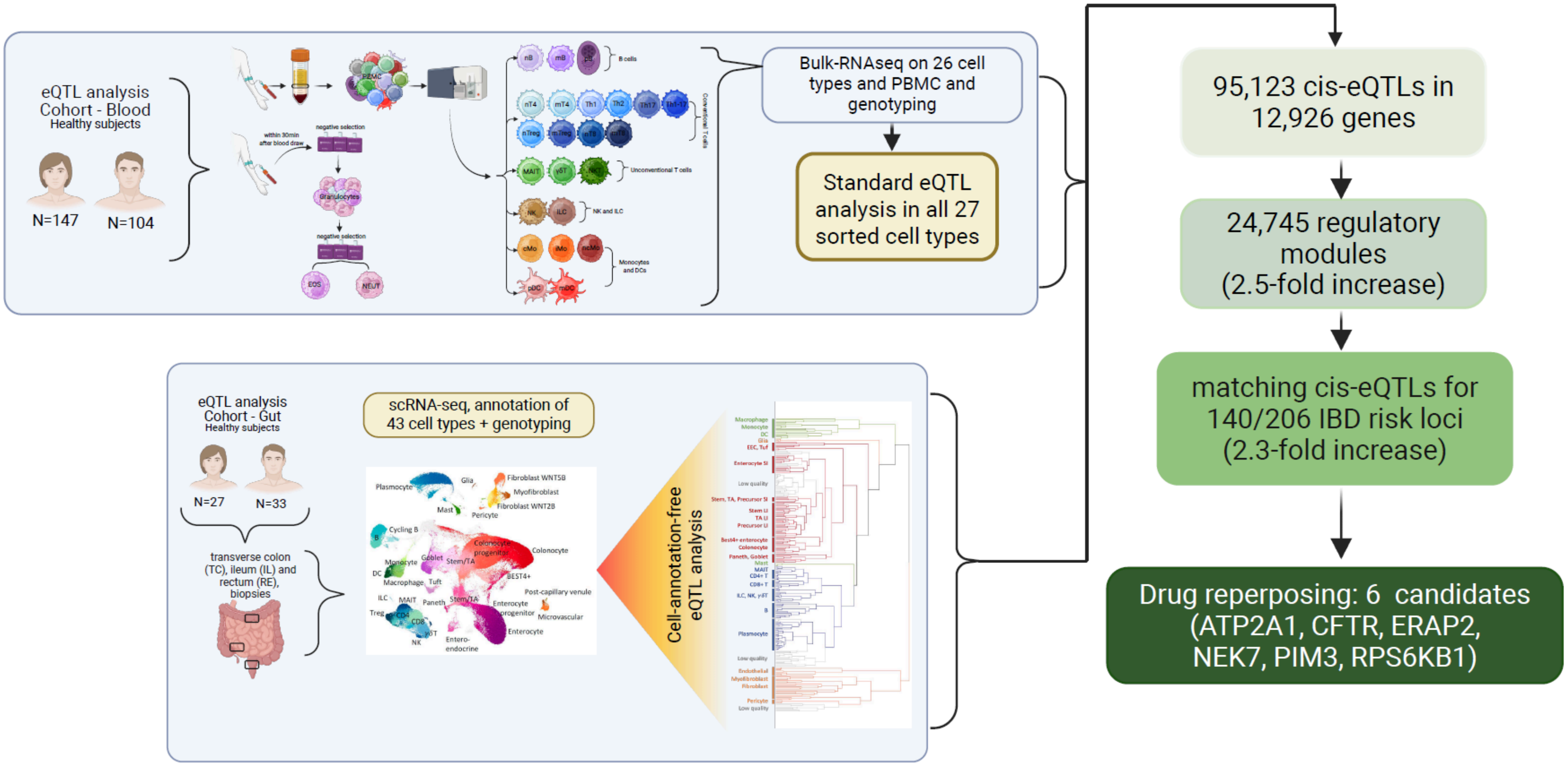
My sincere thanks go to the other members of the Unit of Animal Genomics for the enjoyable lunch times and lab activities: Miyako, Caroline, Lijing, Yefang, Can, Gabriel, Anne-So, Jérôme, Naveed, Damien, Snehal, Arthur, Vittorio, Vincent, Natalia, Maulana, Jose-Luis, and Arina. I would also like to express my appreciation to the principal investigators, Carole Charlier, Anne Van den Broeke, and Tom Druet, for your insightful questions during the seminars.

I am also deeply thankful to the members of the core facilities. Lati, Émilie, Manon, Justine, and Khadija, thank you for your training in library preparation and genotyping, as well as your considerable work. Alice, I truly appreciate your training in analysis. I thank you, along with David, for the enriching discussions we had regarding bioinformatics analyses, clusters, and mass storage. Raafat and Céline, thank you for your training in flow cytometry, for performing some experiments, and for your assistance with troubleshooting.

Finally, I wish to express my profound gratitude to my family, my in-laws, and my friends for your constant encouragement and support. Last but certainly not least, a heartfelt thank you to Denis, my wonderful husband, for illuminating my days and offering perpetual support during stressful times. I have no regrets about moving to Liège for this thesis, as it led me to meet you.

ABSTRACT

Inflammatory bowel disease (IBD), which includes Crohn's disease and ulcerative colitis, is a common complex disease (CCD) characterized by chronic inflammation of the gastrointestinal tract. The etiology of IBD is strongly influenced by genetics, with genome-wide association studies (GWAS) identifying 241 risk loci associated with the disease. The typical span of a risk locus is approximately 250 kb, encompassing between 0 and 100 genes. Coding variants are responsible for disease associations at only 14 of these risk loci. It has been shown that the majority of observed associations can be attributed to regulatory variants that perturb the expression of one or more neighboring genes in *cis*. Expression quantitative trait locus (eQTL) analyses in disease-relevant cell types obtained from healthy individuals are one method for identifying causal genes. Indeed, it seems reasonable to postulate that common regulatory variants causing eQTL effects can be detected in healthy individuals. Colocalization methods have been developed to quantify the similarity between disease association patterns (DAP) and expression association patterns (EAP) derived from GWAS and eQTL analyses, respectively. However, matching *cis*-QTLs to uncover putative causal genes has been successful for only 31.5% (63/200) of the IBD risk loci analyzed. The large proportion of "orphan" risk loci may be explained by the fact that the matching eQTLs have not yet been identified. Possible reasons include the absence or underrepresentation of disease-relevant cell types in existing datasets, that some eQTLs are only active in specific contexts (such as the disease process), or that the eQTLs driving CCDs are different from those discovered so far, which often have limited sample sizes compared to GWAS. The aim of this thesis is to investigate the hypothesis that disease-relevant cell types may have been absent or underrepresented in existing datasets. To discover novel eQTLs driving inherited predisposition to IBD, we established the CEDAR-2 cohort with healthy Europeans, dividing it into two parts. In the first part, we collected peripheral blood from 200 individuals and isolated peripheral blood mononuclear cells (PBMCs), as well as 26 circulating immune cell types using fluorescence-activated cell sorting (FACS) and magnetic-activated cell sorting (MACS). We then performed ultra-low input RNA sequencing (RNA-Seq) on each of the more than 5,000 samples. In the second part, we collected intestinal biopsies at three locations (ileum, colon, and rectum) from 60 individuals, performed single-cell RNA-Seq (scRNA-Seq), and defined 43 intestinal cell populations. All individuals were genotyped for approximately 700,000 SNPs and imputed to 6.3 million variants. We developed and applied a novel cell type annotation-free eQTL analysis approach for scRNA-Seq data. We identified 60,113 and 35,010 *cis*-eQTLs in blood and gut tissues, respectively, affecting 12,926 genes. We merged these eQTLs into 24,745 regulatory modules (RMs), most of which are cell type-specific, location-specific (small versus large intestine), or compartment-specific (blood versus gut). Next, we compared our catalogue of EAP with the DAP in 206 risk loci for IBD using published GWAS summary statistics, employing colocalization methods. We identified matching patterns for 140 risk loci, implicating 556 genes, of which 366 have not previously been associated with IBD. We identify 3.5 times more risk loci and 9.4 times more eQTL genes with DAP-matching EAP with our combined blood and gut dataset, than with whole blood eQTL information from 31,684 individuals. We examined the effect of disease status on the expression of our catalogue of 556 DAP-matching candidate genes in blood and gut tissues, showing that the expression of 119 genes is significantly affected in a manner consistent with the sign of the eQTL effect of risk variants. Notably, eight genes in our list of candidates are or have been clinically tested for treating IBD. We also identified ten genes targeted by drugs that are at least in clinical phase III for diseases other than IBD, which act on their targets in a manner consistent with the effects of IBD risk variants, with the expression of four of these genes being affected by disease status. Additionally, we developed a web browser to visualize our data.



eQTL analysis Cohort - Blood Healthy subjects

N=147 N=104

Bulk-RNAseq on 26 cell types and PBMC and genotyping

Standard eQTL analysis in all 27 sorted cell types

eQTL analysis Cohort - Gut Healthy subjects

N=27 N=33

transverse colon (TC), ileum (IL) and rectum (RE), biopsies

scRNA-seq, annotation of 43 cell types + genotyping

Cell-annotation-free eQTL analysis

95,123 cis-eQTLs in 12,926 genes

24,745 regulatory modules (2.5-fold increase)

matching cis-eQTLs for 140/206 IBD risk loci (2.3-fold increase)

Drug reperposing: 6 candidates (ATP2A1, CFTR, ERAP2, NEK7, PIM3, RPS6KB1)

RÉSUMÉ

Les maladies inflammatoires chroniques de l'intestin (IBD), qui incluent la maladie de Crohn et la rectocolite hémorragique, sont des maladies complexes fréquentes (CCD) caractérisées par une inflammation chronique du tractus gastro-intestinal. L'étiologie des MICI est fortement influencée par la génétique, les études d'association pangénomique (GWAS) ayant identifié 241 locus de risque associés à la maladie. La taille typique d'un locus de risque est d'environ 250 kb, englobant entre 0 et 100 gènes. Des variants codants ne sont responsables d'associations avec la maladie que pour 14 de ces locus. Il a été démontré que la majorité des associations observées peuvent être attribuées à des variants régulateurs perturbant l'expression d'un ou plusieurs gènes voisins en *cis*. Les analyses des locus de traits quantitatifs d'expression (eQTLs) dans des types cellulaires pertinents pour la maladie, obtenus chez des individus sains, constituent une méthode d'identification des gènes responsables. En effet, il semble raisonnable de postuler que les variants régulateurs courants responsables des effets eQTLs peuvent être détectés chez des individus sains. Des méthodes de colocalisation ont été développées pour quantifier la similarité entre les profils d'association de la maladie (DAP) et les profils d'association d'expression (EAP) dérivés respectivement des analyses GWAS et eQTL. Cependant, la correspondance des *cis*-eQTLs pour identifier les gènes potentiellement responsables n'a été efficace que pour 31,5 % (63/200) des loci à risque de MICI analysés. La forte proportion de loci à risque « orphelins » pourrait s'expliquer par le fait que les eQTLs correspondants n'ont pas encore été identifiés. Parmi les raisons possibles, on peut citer l'absence ou la sous-représentation de types cellulaires pertinents pour la maladie dans les ensembles de données existants, le fait que certains eQTLs ne soient actifs que dans des contextes spécifiques (comme le processus pathologique), ou que les eQTLs responsables des CCD soient différents de ceux découverts jusqu'à présent, dont les échantillons sont souvent limités par rapport aux GWAS. L'objectif de cette thèse est d'examiner l'hypothèse selon laquelle les types cellulaires pertinents pour la maladie pourraient avoir été absents ou sous-représentés dans les bases de données existantes. Afin de découvrir de nouveaux eQTLs responsables de la prédisposition héréditaire aux MICI, nous avons établi la cohorte CEDAR-2 auprès d'Européens sains, en la divisant en deux parties. Dans la première partie, nous avons prélevé du sang périphérique auprès de 200 individus et isolé des cellules mononucléaires du sang périphérique (PBMC), ainsi que 26 types de cellules immunitaires circulantes par tri cellulaire activé par fluorescence (FACS) et tri cellulaire activé par magnétique (MACS). Nous avons ensuite réalisé un séquençage d'ARN à très faible apport (RNA-Seq) sur chacun des plus de 5 000 échantillons. Dans la seconde partie, nous avons prélevé des biopsies intestinales à trois endroits (iléon, côlon et rectum) chez 60 individus, réalisé un RNA-Seq sur cellule unique (scRNA-Seq) et défini 43 populations de cellules intestinales. Tous les individus ont été génotypés pour environ 700 000 SNP et imputés à 6,3 millions de variants. Nous avons développé et appliqué une nouvelle approche d'analyse eQTL sans annotation de type cellulaire

pour les données scRNA-Seq. Nous avons identifié 60 113 et 35 010 *cis*-eQTLs dans le sang et les tissus intestinaux, respectivement, affectant 12 926 gènes. Nous avons fusionné ces eQTLs en 24 745 modules de régulation (MR), la plupart spécifiques au type cellulaire, à la localisation (intestin grêle versus gros intestin) ou au compartiment (sang versus intestin). Nous avons ensuite comparé notre catalogue d'EAP avec le DAP dans 206 loci à risque de MICI à l'aide de statistiques récapitulatives GWAS publiées, en employant des méthodes de colocalisation. Nous avons identifié des profils de correspondance pour 140 loci à risque, impliquant 556 gènes, dont 366 n'avaient jamais été associés aux MICI. Nous avons identifié 3,5 fois plus de loci à risque et 9,4 fois plus de gènes eQTLs avec des EAP compatibles DAP grâce à nos données combinées sanguines et intestinales qu'avec les données eQTLs du sang total de 31 684 personnes. Nous avons examiné l'effet du statut pathologique sur l'expression de notre catalogue de 556 gènes candidats compatibles DAP dans le sang et les tissus intestinaux, montrant que l'expression de 119 gènes est significativement affectée, d'une manière cohérente avec le signe de l'effet eQTL des variants à risque. Notamment, huit gènes de notre liste de candidats sont ou ont été cliniquement testés pour le traitement des MICI. Nous avons également identifié dix gènes ciblés par des médicaments au moins en phase clinique III pour des maladies autres que les MICI, qui agissent sur leurs cibles d'une manière cohérente avec les effets des variants à risque des MICI, l'expression de quatre de ces gènes étant affectée par le statut pathologique. De plus, nous avons développé un navigateur web pour visualiser nos données.

ACRONYMS

ANOVA	analysis of variance
APC	antigen-presenting cell
AMP	antimicrobial peptide
bp	base pair
CCD	common complex disease
CD	Crohn’s disease
CEDAR	correlated expression & disease association research
CPM	counts per million
CRM	*cis*-regulatory module
DAP	disease association pattern
DC	dendritic cell
DSS	dextran sulfate sodium
e-gene	expression quantitative trait locus gene
EAP	expression association pattern
eQTL	expression quantitative trait locus
FACS	fluorescence-activated cell sorting
FDR	false discovery rate
FPR	false positive rate
GALT	gut-associated lymphoid tissue
GI	gastrointestinal
GO	gene ontology
GTEx	genotype-tissue expression
GWAS	genome-wide association study
H0	null hypothesis
H1	alternative hypothesis

HLA	human leukocyte antigen
HT	high-throughput
IBD	inflammatory bowel disease
IEC	intestinal epithelial cell
IEL	intraepithelial lymphocyte
IFN	interferon
IL	ileum or interleukin
ILC	innate lymphoid cell
indel	insertion/deletion
INT	inverse normal transformation
ITG	integrin
JAK	Janus kinase
kb	kilobase
LD	linkage disequilibrium
LI	large intestine
LP	lamina propria
LPS	lipopolysaccharide
MACS	magnetic-activated cell sorting
MAF	minor allele frequency
MAIT	mucosal-associated invariant T cell
MAP	*Mycobacterium avium paratuberculosis*
MBV	match BAM to VCF
mDC	myeloid dendritic cell
MHC	major histocompatibility complex
MLN	mesenteric lymph node
MRCA	most recent common ancestor

NGS	next-generation sequencing
NK	natural killer cell
NKT	natural killer T cell
NLRP3	NOD-like receptor pyrin domain-containing protein 3
OR	odds ratio
ORF	open reading frame
PBMC	peripheral blood mononuclear cell
PC	principal component
PCA	principal component analysis
pDC	plasmacytoid dendritic cell
PEER	probabilistic estimation of expression residuals
pQTL	protein quantitative trait locus
PRS	polygenic risk score
PWAS	proteome-wide association study
QC	quality control
QQ	quantile-quantile
QTL	quantitative trait locus
RAF	risk allele frequency
RCF	relative centrifugal force
RE	rectum
RM	regulatory module
RNA-Seq	RNA sequencing
rs	reference single nucleotide polymorphism
SCFA	short-chain fatty acid
scRNA-Seq	single-cell RNA sequencing
SI	small intestine

SMR	summary Mendelian randomization
SNP	single nucleotide polymorphism
t-SNE	t-distributed stochastic neighbor embedding
TC	transverse colon
TCR	T cell receptor
TF	transcription factor
Tfh	T follicular helper cell
TGF	tumor growth factor
Th	T helper cell
TLR	toll-like receptor
TNF	tumor necrosis factor
TPM	transcripts per million
Treg	T regulatory cell
TSS	transcription start site
TWAS	transcriptome-wide association study
TYK	tyrosine kinase
UC	ulcerative colitis
UMAP	uniform manifold approximation & projection
UMI	unique molecular identifier

TABLE OF CONTENTS

INTRODUCTION

1. Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) is a common complex disease (CCD) that is imposing a growing burden on health care systems worldwide (Alatab *et al.*, 2020). It is characterized by a relapsing and remitting chronic inflammation of the gastrointestinal (GI) tract (Wallace, 2014). It is thought to result from a dysregulated immune response to the commensal microbiota in genetically predisposed individuals exposed to environmental risk factors. There are two main forms of IBD: Crohn's disease (CD) and ulcerative colitis (UC) (Figure 1). Patients with CD exhibit patchy, transmural inflammation in the intestinal wall, which can occur anywhere along the GI tract, from the mouth to the anus. By contrast, patients with UC present a continuous mucosal and submucosal inflammation that is restricted to the colon (Khor *et al.*, 2011).

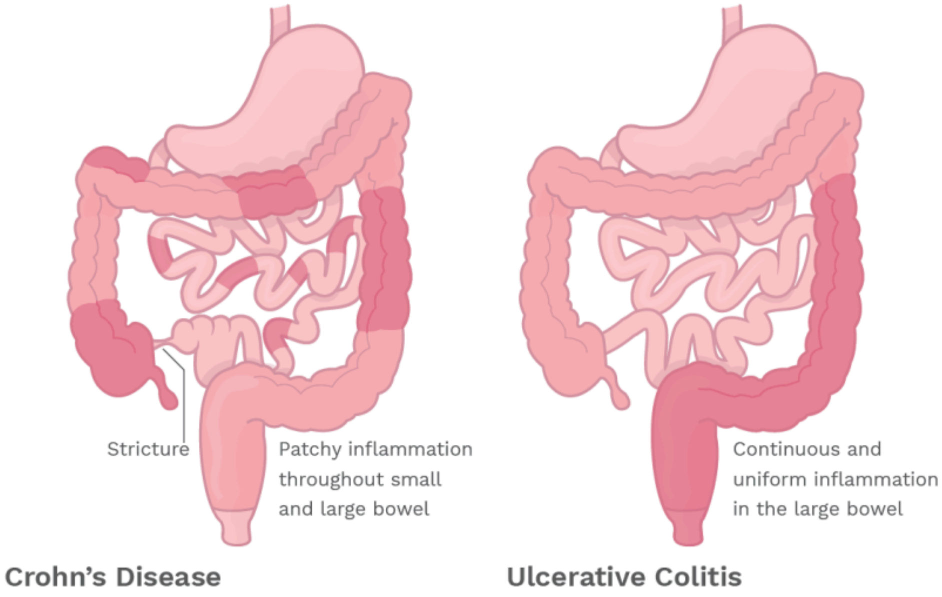
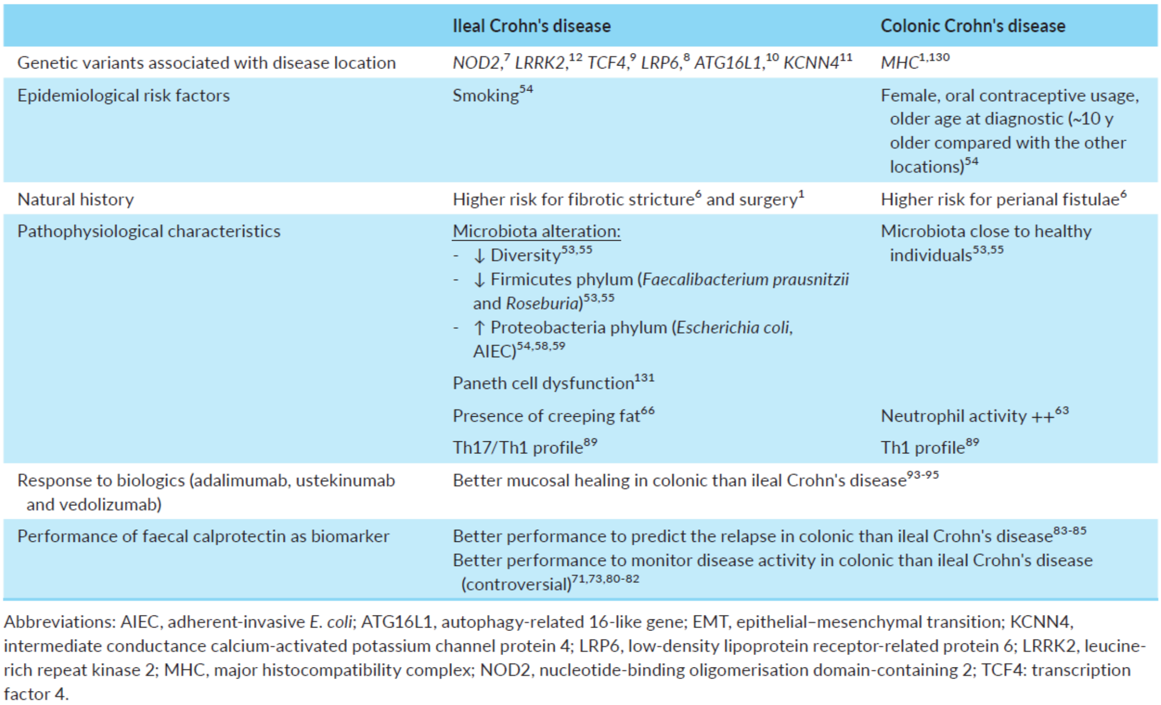

Stricture
Patchy inflammation throughout small and large bowel
Crohn's Disease
Continuous and uniform inflammation in the large bowel
Ulcerative Colitis

Figure 1: Comparison of CD and UC, the two main forms of IBD (from https://allmyfriendsareeggs.com/crohns-vs-colitis).

Recent studies distinguish ileal from colonic CD based on variations in the host's genetic profile and clinical manifestations (Table 1). For example, ileal CD has been linked to single nucleotide polymorphisms (SNPs) in genes such as *NOD2* and *ATG16L1*, while colonic CD is associated with SNPs in the major histocompatibility complex (*MHC*), specifically the human leukocyte antigen (*HLA*) system (Cleynen *et al.*, 2016, Pierre *et al.*, 2021).



	Ileal Crohn's disease	Colonic Crohn's disease
Genetic variants associated with disease location	*NOD2*,[7] *LRRK2*,[12] *TCF4*,[9] *LRP6*,[8] *ATG16L1*,[10] *KCNN4*[11]	*MHC*[1,130]
Epidemiological risk factors	Smoking[54]	Female, oral contraceptive usage, older age at diagnostic (~10 y older compared with the other locations)[54]
Natural history	Higher risk for fibrotic stricture[6] and surgery[1]	Higher risk for perianal fistulae[6]
Pathophysiological characteristics	Microbiota alteration: - ↓ Diversity[53,55] - ↓ Firmicutes phylum (*Faecalibacterium prausnitzii* and *Roseburia*)[53,55] - ↑ Proteobacteria phylum (*Escherichia coli*, AIEC)[54,58,59]	Microbiota close to healthy individuals[53,55]
	Paneth cell dysfunction[131]	
	Presence of creeping fat[66]	Neutrophil activity ++[63]
	Th17/Th1 profile[89]	Th1 profile[89]
Response to biologics (adalimumab, ustekinumab and vedolizumab)	Better mucosal healing in colonic than ileal Crohn's disease[93-95]	
Performance of faecal calprotectin as biomarker	Better performance to predict the relapse in colonic than ileal Crohn's disease[83-85] Better performance to monitor disease activity in colonic than ileal Crohn's disease (controversial)[71,73,80-82]	

Abbreviations: AIEC, adherent-invasive *E. coli*; ATG16L1, autophagy-related 16-like gene; EMT, epithelial–mesenchymal transition; KCNN4, intermediate conductance calcium-activated potassium channel protein 4; LRP6, low-density lipoprotein receptor-related protein 6; LRRK2, leucine-rich repeat kinase 2; MHC, major histocompatibility complex; NOD2, nucleotide-binding oligomerisation domain-containing 2; TCF4: transcription factor 4.

Table 1: Comparison of ileal and colonic CD (Pierre *et al.*, 2021).

Patients with colonic CD or UC have a higher risk of developing colorectal cancer (Gordon *et al.*, 2023).

a. Clinical manifestations

During flare-ups, IBD patients experience intestinal symptoms such as diarrhea, abdominal pain, rectal bleeding, weight loss, fever, and fatigue (Wallace, 2014). They can also suffer from extraintestinal manifestations involving in particular the joints (spondyloarthritis), eyes (episcleritis, scleritis, uveitis), and skin (erythema nodosum, pyoderma gangrenosum, Sweet syndrome) lesions (Hedin *et al.*, 2019). These debilitating and lifelong symptoms affect the quality of life and increase risk of anxiety and depression (Mitropoulou *et al.*, 2022), which in turn may worsen the disease (Bonaz & Bernstein, 2013).

Chronic inflammation disrupts the epithelial barrier which triggers chronic wound healing that can lead to intestinal fibrosis (Tavares de Sousa & Magro, 2023), to stricturing and/or penetrating (fistulas and abscesses) complications in CD (Khor *et al.*, 2011), and to abnormalities of motility and rectal urgency in UC (Tavares de Sousa & Magro, 2023). Strictures or stenoses are narrowed sections of the intestines that are caused by thickening of the bowel wall (Grajo *et al.*, 2021), that can progress to obstructions (Lichtenstein *et al.*, 2006). Fistulas are aberrant connections between two loops of the intestine or between the intestine and another organ such as the skin, bladder or vagina, which can develop into abscesses if infected, or into perforations (Hirten *et al.*, 2018).

To make a diagnosis, clinicians rely on clinical manifestations, imaging, endoscopy, histology and high levels of inflammatory markers such as erythrocyte sedimentation rate, serum C-reactive protein and fecal calprotectin (Hong & Baek, 2024).

b. Epidemiology

The prevalence of IBD is highest in westernized countries (Europe, North America and Oceania) with a prevalence exceeding 0.3% of the population. Indeed, it was shown that 322 (Germany and Italy) and 505 (Norway) per 100,000 people were affected by CD and UC, respectively. In total, it affects more than 1.5 million North Americans and 2 million Europeans. While the incidence has now stabilized in these countries, other parts of the world (Africa, Asia and South America) have seen a rapid increase (Molodecky *et al.*, 2012; Ng *et al.*, 2017). Four epidemiological stages have been established to classify the progression of IBD. The initial phase pertains to developing countries where the disease begins to surface. The second phase relates to newly industrialized nations where both incidence and prevalence are on the rise. The third phase involves westernized countries, where the increase in incidence halts, but prevalence continues to grow as mortality rates decline. In the fourth phase, both incidence and prevalence stabilize, with projections indicating that by 2030, IBD could impact 1% of the population (Kaplan & Windsor, 2020). IBD is mainly diagnosed between the second and fourth decade of life and impacts approximately the same number of men and women (Molodecky *et al.*, 2012). The total cost of IBD was estimated in the United States at 2.2 billion per year (Everhart & Ruhl, 2009).

i. Environmental risk factors

The increase in IBD incidence in parallel with urbanization suggests that environmental factors must play a role in disease onset. These include cesarean delivery, use of antibiotics in childhood, oral contraceptives or non-steroidal anti-inflammatory drugs, vitamin D deficiency, air pollution, stress and a diet high in fat and sugar but low in fiber. Conversely, being physically active or ever breastfed, having pets or farm animals or at least two siblings appear to be protective against IBD. Surprisingly, current smoking and appendectomy are both associated with an increased risk of CD but a reduced risk of UC (Loftus, 2004; Zhang, 2014; Piovani *et al.*, 2019). Considering all this, it has been suggested that lower exposure to antigenic stimuli in childhood is associated with a higher risk of IBD. This is called the hygiene hypothesis (Cholapranee & Ananthakrishnan, 2016).

Interestingly, the idea that IBD could be an infectious disease is still debated in the literature. For example, anti-*Mycobacterium avium paratuberculosis* (MAP) therapy is being considered to treat CD. This bacterium is responsible for Johne's disease in cattle, which is equivalent to CD in humans, and can be transmitted to humans through the consumption of milk and meat. Although difficult to detect, MAP appears to be more prevalent in CD patients than in healthy individuals (Honap *et al.*, 2021). Furthermore, *Mycobacterium bovis* infection has been shown to induce gastrointestinal tuberculosis in an immunocompromised patient with systemic lupus erythematosus and her symptoms were similar to those of CD: diarrhea, abdominal pain, weight loss, fever, fatigue, and enterocolitis (Winger *et al.*, 2016).

ii. Genetic risk factors

Environmental factors alone are not sufficient to trigger the disease. Ethnic differences, family studies and twin studies have underscored the hereditary nature of IBD (Ek *et al.*, 2014). Indeed, researchers have noted higher incidence of IBD among first- and second-degree relatives (Orholm *et al.*, 1991) and an increased concordance of the IBD phenotype in monozygotic twins compared to dizygotic twins of the same sex, reducing sex bias (Tysk *et al.*, 1988). Since then, numerous research teams have studied the genetic underpinnings of IBD (El Hadad *et al.*, 2024). Heritability, the proportion of phenotypic variance due to genetic variance, was estimated from twin studies to be 0.75 for CD and 0.67 for UC (Chen *et al.*, 2014). Except for some monogenic forms of IBD in children, IBD is mainly a polygenic disease (Jans & Cleynen, 2023). Several genes associated with IBD susceptibility have been identified, including *NOD2*, *ATG16L1*, *CARD9* and *IL23R* (Huang *et al.*, 2017), all of which play a role in the immune system.

The most well-known susceptibility gene for IBD is *NOD2* (nucleotide-binding oligomerization domain-containing protein 2), also known as *CARD15* (caspase recruitment domain-containing protein 15). It is expressed in intestinal epithelial cells and encodes an intracellular receptor that recognizes a degradation product of peptidoglycan, a component of the cell wall of almost all bacteria. Upon recognition, NOD2 activates the nuclear factor kappa B (NF-κB) pathway, promoting the production of defensins that help eliminate intracellular bacteria. Additionally, NOD2 stimulates ATG16L1 (autophagy-related 16-like 1), leading to autophagy. However, NOD2 is in cross-talk with Toll-like receptor 2 (TLR2), an extracellular receptor also stimulated by peptidoglycan, because ATG16L1 appears to inhibit TLR2-dependent NF-κB activation. In cases of NOD2 loss-of-function mutations, NOD2 can no longer activate NF-κB. Paradoxically, this deficiency leads to heightened inflammation because NOD2 can no longer activate ATG16L1 to prevent TLR2-dependent NF-κB activation (Graham *et al.*, 2020; Okai *et al.*, 2023) (Figure 2).

CARD9 is another pattern recognition receptor, but it binds to the fungal cell wall (Glassner *et al.*, 2020). Another example is *IL23R* (interleukin 23 receptor), which is activated by IL23, a cytokine secreted by activated macrophages and dendritic cells to promote the differentiation of naive CD4$^+$ T cells into T helper 17 (Th17) cells (Xu *et al.*, 2015). The second chapter of the introduction will be devoted to the genetic analysis of IBD predisposition.

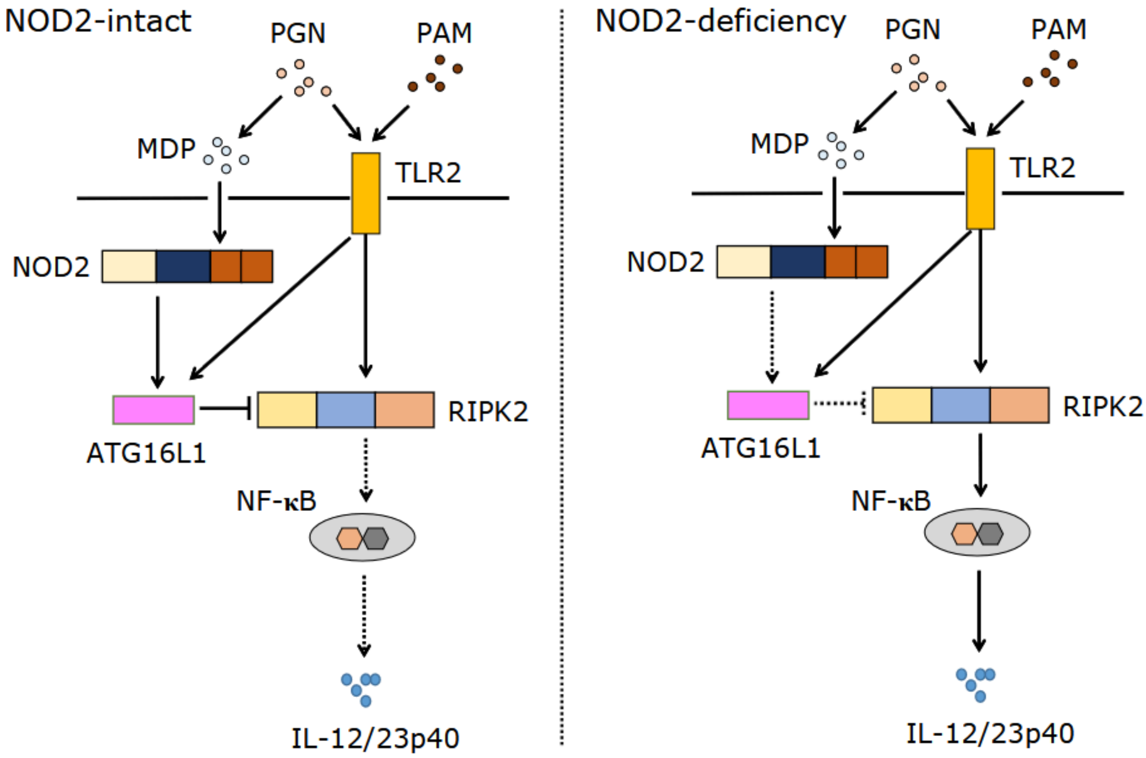


Figure 2: Cross-talk of NF-κB regulation following peptidoglycan recognition by NOD2 and TLR2 in the context of intact NOD2 (left) versus NOD2 deficiency (right) (from Okai *et al.*, 2023).

c. Pathogenesis

IBD impacts three components of the immune system: the intestinal epithelium, the innate immune system and the adaptive immune system (Silva *et al.*, 2016).

i. Microbiota

Commensal microorganisms that inhabit the gut include bacteria, viruses and fungi, which increase in number and diversity from the stomach to the colon (Mentella *et al.*, 2020). It was shown that IBD susceptibility genes such as *NOD2*, *ATG16L1*, and *CARD9* play a role in host-microbiome interactions, as discussed above. Environmental factors such as smoking, diet, breastfeeding and antibiotic use can influence the gut microbiota (Glassner *et al.*, 2020). Dysbiosis, altered composition of the gut microbiome, called dysbiosis, has been implicated in IBD. Compared to healthy individuals, IBD patients exhibit a significant reduction in the proportion of bacterial species from the Bacteroidetes phylum and a smaller decrease in species from the Firmicutes phylum, leading to an increased Firmicutes to Bacteroidetes ratio. These changes are counterbalanced by a large and small increase in the proportion of bacterial species from the Proteobacteria and Actinobacteria phyla, respectively (Hall *et al.*, 2017). Specifically, anti-inflammatory bacteria that produce short-chain fatty acids (SCFAs) such as *Faecalibacterium prausnitzii* from the Firmicutes phylum are depleted in IBD patients while pro-inflammatory bacteria such as pathogenic *Escherichia coli* from the Proteobacteria phylum and MAP from the Actinobacteria phylum are enriched in IBD patients. It is not yet clear whether this is a cause or a consequence of intestinal inflammation (Ramos & Papadakis, 2019; Piovani *et al.*, 2019; Shin *et al.*, 2023; Diez-Martin *et al.*, 2024). Viruses and fungi are rarely studied in IBD because they are difficult to sequence, but they are also important (Arora *et al.*, 2024). For example, it was shown that mice carrying an *ATG16L1* mutation only develop CD-like symptoms when infected with an enteric virus (Cadwell *et al.*, 2010; Kostic *et al.*, 2014; Brown *et al.*, 2019) and that an increase of *Candida* species (*C. albicans*, *C. tropicalis* and *C. glabrata*) was observed in the stools of CD patients (Underhill & Braun, 2022). Further evidence for the role of the microbiome in IBD pathogenesis is that fecal microbiota transplantation shows promise in inducing remission of active UC (Paramsothy *et al.*, 2017).

ii. Intestinal epithelium

Intestinal epithelium, along with its mucus layer, constitute the first barrier that separates the lamina propria (LP) (the host) from the gut lumen containing commensal bacteria, potential pathogens and food antigens (the environment). It consists of a monolayer of intestinal epithelial cells (IECs) with one type of absorptive cell, called enterocytes in the small intestine and colonocytes in the large intestine, and three to four types of secretory cells. These secretory cells include goblet cells (which produce mucus), enteroendocrine cells (which release hormones), tuft cells, and Paneth cells (which secrete antimicrobial peptides (AMP) such as α-defensins and are found only in the small intestine). All of these cells are replenished from stem cells and transit-amplifying cells located at the base of the crypts (Noah *et al.*, 2011; Coskun, 2014; Neurath, 2014; Sylvestre *et al.*, 2023). Intraepithelial lymphocytes (IELs) are also present among the IECs (Kong *et al.*, 2024). The LP is made up of stromal cells (such as fibroblasts, endothelial cells that make up blood vessels and glial cells that are part of the enteric nervous system) as well as myeloid and lymphoid immune cells (Xu, 2014; Ghilas *et al.*, 2022; Hickey *et al.*, 2023).

IBD is associated with down-regulation of epithelial cadherin (E-cadherin) in tight junctions and lower secretion of mucin 2 (Muc2) and resistin-like molecule β (RELMβ) proteins that compose mucus by goblet cells, which compromises the permeability of the intestinal barrier. Abnormal mechanisms associated with *NOD2* and *ATG16L1* genes and lower production of AMPs were also observed in Paneth cells (Khor *et al.*, 2011; Ramos & Papadakis, 2019) (Figure 3).

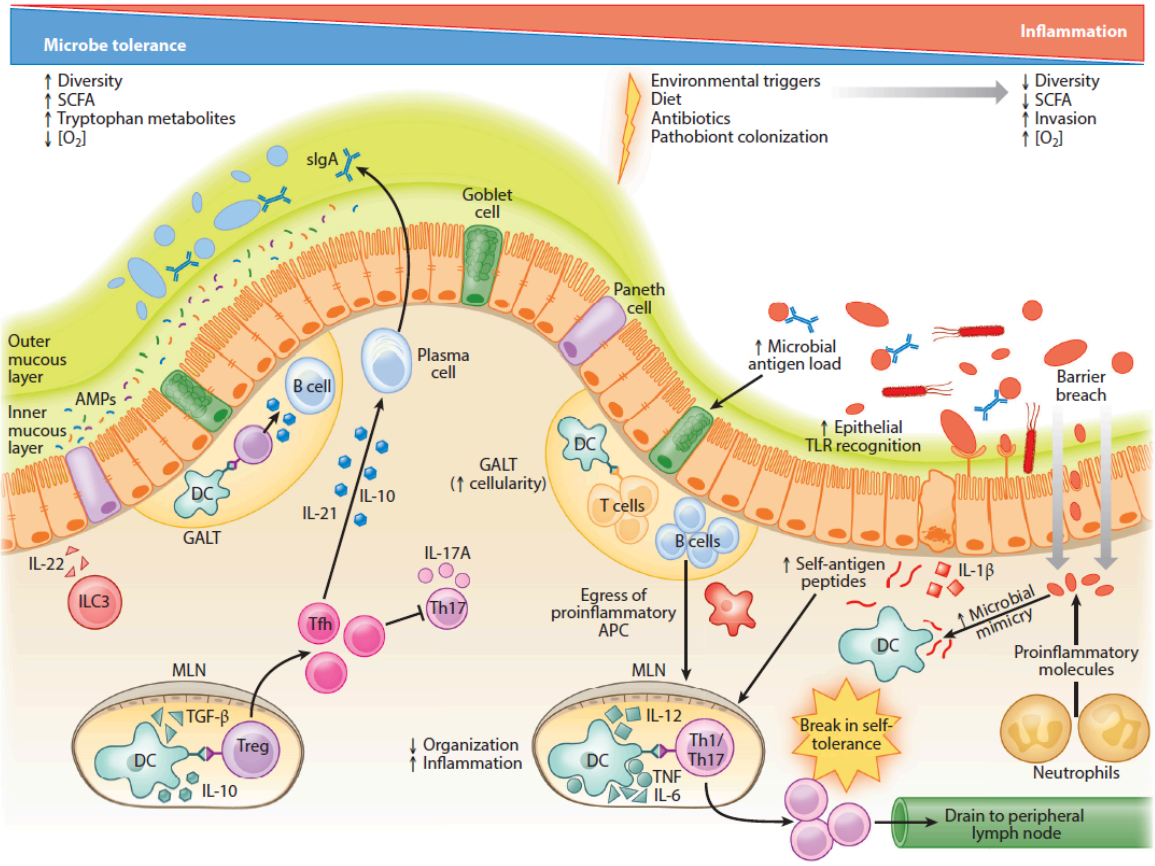


Figure 3: Comparison between healthy intestinal epithelium on the left and inflamed epithelium on the right (from Brown *et al.*, 2019). In a healthy intestinal epithelium, goblet cells (green) produce mucus and Paneth cells (purple) secrete AMPs. DCs (teal) located in gut-associated lymphoid tissues (GALT) and mesenteric lymph nodes (MLN) present microbial antigens to T cells and release tolerogenic signals to promote the differentiation of T cells into anti-inflammatory Tregs (purple). Tregs can be converted in Tfh (pink) to stimulate B cells (blue) to release IgA antibodies against the microbiota. In a damaged epithelium, neutrophils arrive to secrete pro-inflammatory molecules. DCs become pro-inflammatory and promote the differentiation of T cells into Th1 or Th17 cells (magenta). Injured epithelial cells release additional potential self-antigens that may be displayed to T cells. There is also an increase in the amount of microbial antigens that can imitate self-antigens or enhance TLR recognition by epithelial cells. If memory CD4$^+$ T cells migrate from the MLNs to the peripheral lymph nodes, inflammation can become systemic.

iii. Innate immune system

The innate immune system provides a non-specific, non-durable, but immediate response against pathogens. It is composed of monocytes/macrophages (called monocytes in blood and macrophages in tissues), granulocytes (neutrophils, eosinophils and basophils), dendritic cells (DC), natural killer cells (NK) and innate lymphoid cells (ILC). Innate immune cells recognize microbial antigens such as lipopolysaccharides via pattern recognition receptors (PRRs) such as Toll-like receptors (TLRs) and NOD-like receptors (NLRs), which leads to activation of NF-κB and secretion of chemokines and pro-inflammatory cytokines such as TNF-α and

IL6. Chemokines are signaling proteins that guide the integrin-dependent migration of neutrophils, macrophages, and T cells from the blood vessels to the site of inflammation where macrophages are responsible for the phagocytosis of infected cells and pathogens while DCs are antigen-presenting cells (APCs) able to display antigens with molecules of the MHC class II to activate naive T cells. In IBD, these innate immune cells identify the gut microbiota as harmful and fail to resolve inflammation. Consequently, they become dysregulated and contribute to chronic inflammation. Alterations in the genes that encode PRRs (such as *NOD2*) or in autophagy-related genes (*ATG16L1* and *IRGM*) contribute to the inability of innate immune cells to properly eliminate intracellular pathogens or damaged cellular components, leading to the development of IBD (Wallace, 2014; Geremia *et al.*, 2014).

Another mechanism implicated in the pathogenesis of IBD is pyroptosis, a form of inflammatory cell death. This process is initiated by the formation of inflammasomes, particularly the NOD-like receptor pyrin domain-containing protein 3 (NLRP3) inflammasome. When lipopolysaccharides (LPS) are recognized by TLR4, it activates the NF-κB pathway, which in turn increases the expression of NEK7. NEK7 then triggers the activation of the NLRP3 inflammasome and the subsequent activation of caspase-1. Caspase-1 processes pro-IL1β into its active form, IL1β, and cleaves gasdermin D (GSDMD) into its N- and C-terminal fragments. The N-terminal fragment of GSDMD forms pores in the cell membrane, leading to cell swelling and eventual cell death. Studies have shown elevated levels of caspase-1, NLRP3, and GSDMD in the tissues of IBD patients (Chen *et al.*, 2019) (Figure 4).

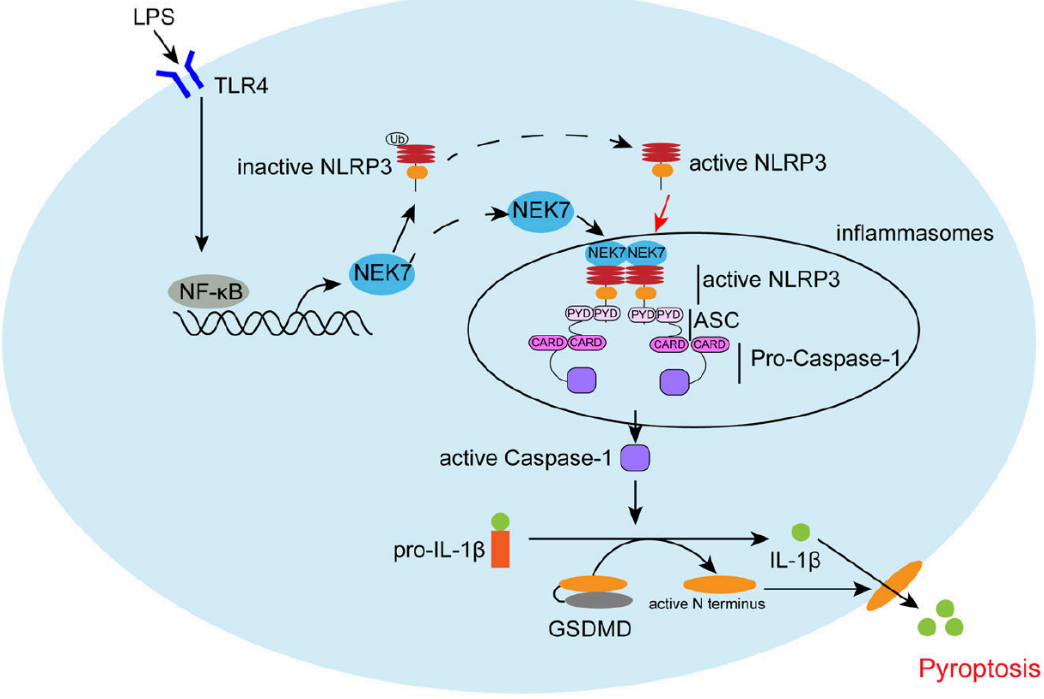


Figure 4: Formation of the NLRP3 inflammasome (from Chen *et al.*, 2019).

iv. Adaptive immune system

The adaptive/acquired immune system provides a specific, durable, but slow protection. It is activated by microbial or self-antigens displayed by APCs and is composed of B lymphocytes that secrete antibodies (humoral immunity) and T lymphocytes (cellular immunity). T lymphocytes can be subdivided in conventional αβ T cells, including CD8$^+$ cytotoxic T cells, CD4$^+$ T helper cells (Th), and T regulatory cells (Treg), and non-conventional T cells, including γδ T cells, natural killer T cells (NKT), and mucosal-associated invariant T cells (MAIT) (Graham *et al.*, 2020). IBD is associated with an imbalance between pro-inflammatory Th1 and Th17 and anti-inflammatory Treg. Th1 (activated by IL12 and secreting IFN-γ) and Th17 cells (activated by IL23 and secreting IL17) are central in driving inflammation in IBD, with Th1 cells involved in the response against intracellular pathogens and Th17 cells playing a role in mucosal immunity and tissue damage. CD is thought to be primarily driven by a Th1 response while UC is rather associated with a Th2 response. Variants in the *IL23R* gene have been identified in IBD patients, highlighting the role of the adaptive immune system in the development of IBD (Wallace, 2014; Geremia *et al.*, 2014; Diez-Martin *et al.*, 2024).

d. Treatment

Several classes of treatments have been approved for IBD, including anti–tumor necrosis factor (TNF) therapies, anti-integrins, anti-interleukins, Janus kinase (JAK) inhibitors, sphingosine-1-phosphate receptor (S1PR) modulators and other uncategorized therapies (Table 2). The primary treatment for IBD involves antibodies targeting TNFα. Two anti-TNFα antibodies (infliximab and adalimumab) are used to treat both CD and UC, while certolizumab pegol is used to treat CD only and golimumab is used to treat UC only (Vieujean *et al.*, 2025). TNFα inhibitors reduce inflammation but do not appear to prevent tumor necrosis, as they do not significantly elevate the risk of developing cancer (Muller *et al.*, 2021). While TNFα antagonists are effective in treating IBD, non-response and loss of response have been observed in 10-30% and 23-46% of patients, respectively (Roda *et al.*, 2016; Pierre *et al.*, 2021). However, a prospective study in IBD patients starting their first anti-TNFα treatment did not find any biomarker of remission (Mishra *et al.*, 2022). Antibodies targeting integrins include vedolizumab (anti-α4β7), used for both CD and UC, and natalizumab (anti-α4), used for CD only. Interleukins-targeting antibodies include risankizumab (anti-IL23) and ustekinumab (anti-IL12/23), both of which are used for CD and UC, as well as mirikizumab and guselkumab (both anti-IL23), which are currently used for UC and in clinical phase 3 for CD. Ustekinumab was initially developed for psoriasis but has been successfully repurposed for CD and UC because of the pivotal role of IL-23 in IBD, supported by genetic evidence, and its effectiveness in anti-TNFα non-responders (Liefferinckx *et al.*, 2019; Trajanoska *et al.*, 2023; Uchida *et al.*, 2023). JAK inhibitors include upadacitinib (JAK1 inhibitor), available for both CD and UC, while filgotinib (JAK1 inhibitor) and tofacitinib (pan-JAK inhibitor) are available for UC only. S1PR modulator therapies consist of etrasimod (S1PR1/4/5 modulator) and ozanimod (S1PR1/5 modulator) are in clinical use for UC and in phase 3 clinical trials for CD. Additional approved therapies include corticosteroids, thiopurines and cellular therapy for both CD and UC, methotrexate for CD only as well as 5-aminosalicylic acid and cyclosporine for UC only (Vieujean *et al.*, 2025).

Class of treatment	CD	UC
anti-TNFs	infliximab (anti-TNFα) adalimumab (anti-TNFα) certolizumab pegol (anti-TNFα)	infliximab (anti-TNFα) adalimumab (anti-TNFα) golimumab (anti-TNFα)
anti-integrins	vedolizumab (anti-α4β7) natalizumab (anti-α4)	vedolizumab (anti-α4β7)
anti-interleukins	risankizumab (anti-IL23) ustekinumab (anti-IL12/23)	risankizumab (anti-IL23) ustekinumab (anti-IL12/23) mirikizumab (anti-IL23) guselkumab (anti-IL23)
JAK inhibitors	upadacitinib (JAK1 inhibitor)	upadacitinib (JAK1 inhibitor) filgotinib (JAK1 inhibitor) tofacitinib (pan-JAK inhibitor)
S1PR modulators		etrasimod (S1PR1/4/5 modulator) ozanimod (S1PR1/5 modulator)
other therapies	corticosteroids thiopurines cellular therapy methotrexate	corticosteroids thiopurines cellular therapy 5-aminosalicylic acid cyclosporine

Table 2: Approved treatments for IBD (adapted from Vieujean *et al.*, 2025)

Patients treated with immunosuppressive agents, i.e. all drugs mentioned above except 5-aminosalicylic acid, are at increased risk of developing opportunistic infections. For example, corticosteroids have been linked to increased risk of fungal infections, thiopurines to a higher incidence of viral infections and anti-TNFs to both fungal and mycobacterial infections. Patients should be encouraged to get vaccinated against certain viral and bacterial infections such as hepatitis B virus and *Streptococcus pneumoniae*, respectively (Kucharzik *et al.*, 2021). Patients may also experience other adverse events: approximately 5% of patients starting infliximab therapy develop anti-TNF-induced lupus (Picardo *et al.*, 2020).

If medical treatment is not enough to reduce the inflammation, surgical procedures are used to dilate strictures, close fistulas, drain pus from abscesses or even resect parts of the intestines. These procedures are associated with a high risk of recurrence and complications (Grajo *et al.*, 2021). Massive or repeated resections should be limited as they can lead to short bowel syndrome (Thompson, 2000). The risk of needing surgery within 10 years after diagnosis decreased over time but it is still at 46.6% for CD and 15.6% for UC (Frolkis *et al.*, 2013). Therefore, there is a need for more drugs.

2. Forward genetic dissection of IBD predisposition

Since epidemiological studies concluded that IBD is hereditary, genetic dissection has begun to identify genetic variants associated with disease risk.

a. Linkage analyses

The identification of causal genetic variants in IBD started more than two decades ago with linkage analyses in families. The goal was to find genomic regions that are more frequently shared among affected individuals than among unaffected ones in the family (Jans and Cleynen, 2023). Linkage analyses started with the study of 270 markers that were not in linkage disequilibrium (LD) with each other, revealing that genetic variants on chromosome 16 were associated with CD (Hugot *et al.*, 1996). In particular, a frameshift mutation in the *NOD2* gene, located on this chromosome, was more often found in CD patients than in controls. As discussed in the "Pathogenesis" section, *NOD2* is involved in the innate immune response by recognizing bacterial components and activating appropriate immune responses. No control individuals were homozygous for this mutation. Since its frequency was rare in cases, this suggested that other *NOD2* variants may contribute to disease risk (Ogura *et al.*, 2001).

b. Genome-wide association studies

The discovery of only one gene was attributed to the small sample size and low number of variants. It was predicted that more risk loci could be identified by replacing linkage analyses in 20 families with association studies in populations (thousands of individuals) and by increasing the number of variants from 1000 to 1 million (Risch & Merikangas, 1996). It took 10 years for Haplotype Map (HapMap) to genotype 270 individuals for 600,000 common SNPs (The International HapMap Consortium, 2003). Then, it took a few more years before the first genome-wide association study (GWAS) in 2006 that identified *IL23R* for IBD (Duerr *et al.*, 2006). For more than 15 years, GWAS have identified tens to hundreds of loci associated with CCD such as IBD, diabetes and schizophrenia but also with quantitative traits such as height, body mass index and blood pressure in up to a million individuals (Abdellaoui *et al.*, 2023).

To perform a GWAS, phenotype information (such as disease status) and genotype data are required for a large number of unrelated participants, preferably of the same ancestry. Genotypes are obtained from DNA samples using microarrays for common SNPs, quality-controlled, phased and imputed (Anderson *et al.*, 2010; Uffelmann *et al.*, 2021). Phasing has to be performed before imputation and consists of using SNP array data to reconstruct the maternal and paternal copies of each haplotype. Then, imputation enables us to predict millions of untyped variants by comparing the reconstructed haplotypes with the haplotypes of a large reference panel with a certain accuracy (Hoffmann & Witte, 2015) (Figure 5). Typical reference panels include HapMap (The International HapMap Consortium, 2005; The International HapMap Consortium, 2007; The International HapMap 3 Consortium, 2010), 1000 Genomes Project (The 1000 Genomes Project Consortium, 2010; The 1000 Genomes Project Consortium, 2012; The 1000 Genomes Project Consortium, 2015) and Trans-Omics for Precision Medicine (TOPMed, Taliun *et al.*, 2021).

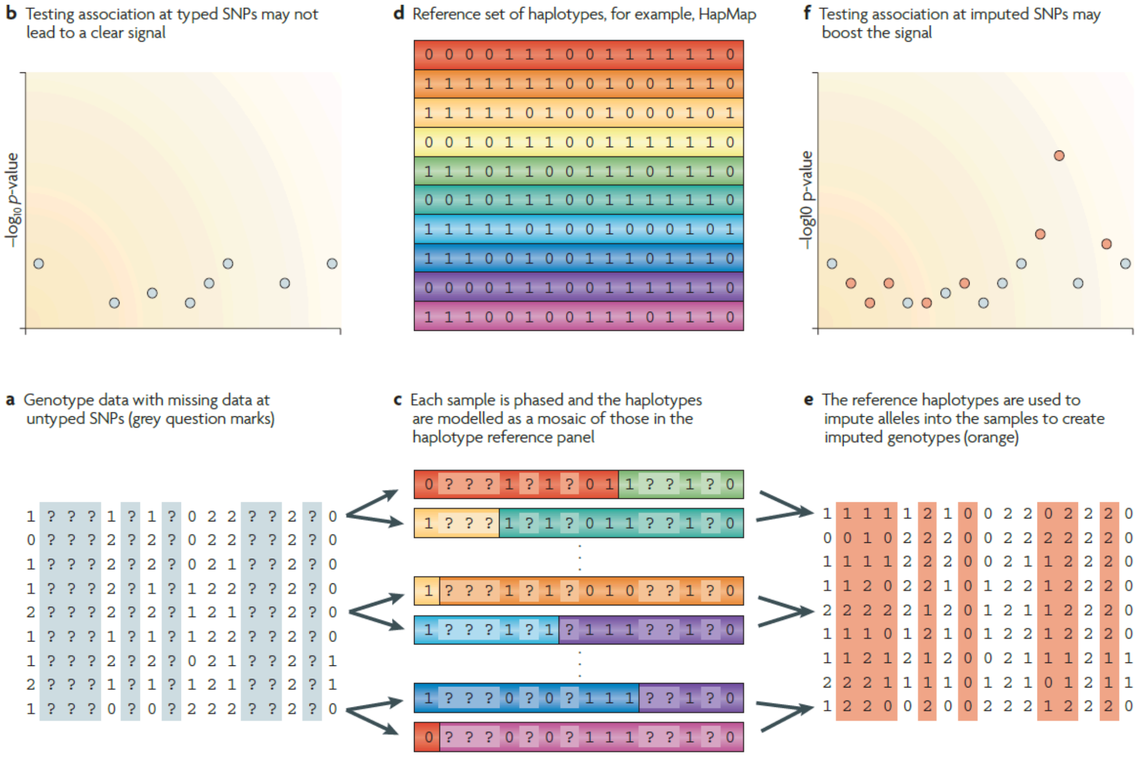

b Testing association at typed SNPs may not lead to a clear signal
−log10 p-value
d Reference set of haplotypes, for example, HapMap
f Testing association at imputed SNPs may boost the signal
−log10 p-value
a Genotype data with missing data at untyped SNPs (grey question marks)
c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel
e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

Figure 5: How phasing and imputation work (from Marchini & Howie, 2010). Panel a: data containing both genotyped and untyped (missing) SNPs. Panel b: association pattern between genotyped variants and a trait. No variant is associated with the trait. Panel c: haplotype reconstruction of three individuals by combining haplotypes from the reference panel. Panel d: haplotypes contained in the reference panel. Panel e: data containing both genotyped and imputed SNPs. Panel f: association pattern between genotyped variants (in blue) and imputed variants (in red) with a trait. One imputed variant is associated with the trait.

The correlation between phenotype and genotype is usually assessed at the SNP level but can be assessed at the haplotype level (Yang *et al.*, 2010). In the case of continuous traits, individuals are sorted by genotype and the three groups are compared by linear regression or analysis of variance (ANOVA) (Uffelmann *et al.*, 2021). Groups can be compared based on several inheritance patterns, such as dominance, semidominance (or additive), codominance, overdominance, and recessivity (Palmer *et al.*, 2023) (Figure 6). An additive model considers that the mean phenotype of heterozygous is positioned halfway between the two homozygotes. Linear regression assumes an additive model and estimates two parameters (slope and intercept) of a linear relationship between genotype and phenotype. ANOVA allows all models and estimates three parameters (the mean phenotypes of the three groups). The more parameters are estimated, the lower the power. In the case of binary traits, individuals are sorted by phenotype and allele frequencies are compared between groups by chi-square or likelihood ratio test (Uffelmann *et al.*, 2021). The chi-square test analyzes the contingency table (allele frequencies versus disease status). The likelihood ratio test evaluates the ratio of a model where allele frequencies are the same in cases and controls (H0) to a model where there is a difference (H1). The likelihood ratio roughly follows a chi-square distribution (Qian & Shao, 2013).

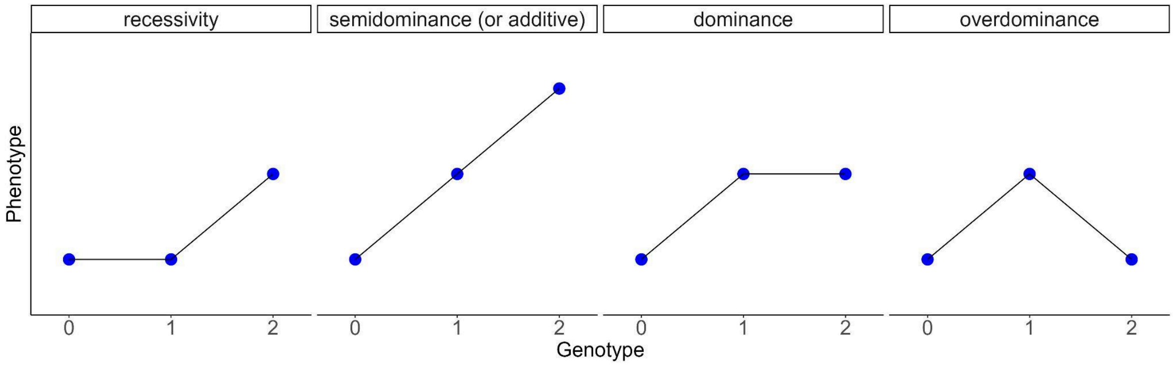


Figure 6: Some inheritance patterns (adapted from Palmer *et al.*, 2023). Blue dots correspond to the mean phenotype of homozygotes for the reference allele (0), heterozygotes (1) and homozygotes for the alternate allele (2).

Since millions of variants are tested, we account for multiple testing by defining significantly associated variants as those with a p-value less than 5×10^{-8} or a $-\log_{10}$(p-value) greater than 8.3. This Bonferroni correction was computed as 0.05 divided by one million independent tests (The International HapMap Consortium, 2005). However, it should be noted that this threshold is valid for the study of common variants in a European cohort, but it needs to be more stringent when analyzing variants with a minor allele frequency (MAF) below 5% (Fadista *et al.*, 2016). GWAS summary statistics include p-values, effect sizes and effect directions for each variant tested. The strength of association with phenotype ($-\log_{10}$(p-value)) depends on two parameters: the effect size and the allele frequency in the population. It was observed in GWAS studies that there are many variants with small effects and few variants with big effects. Indeed, variants with strong effects are quite rare due to natural selection (Uffelmann *et al.*, 2021). Results are usually visualized in two ways: Manhattan plots and quantile-quantile (QQ) plots (Figure 7).

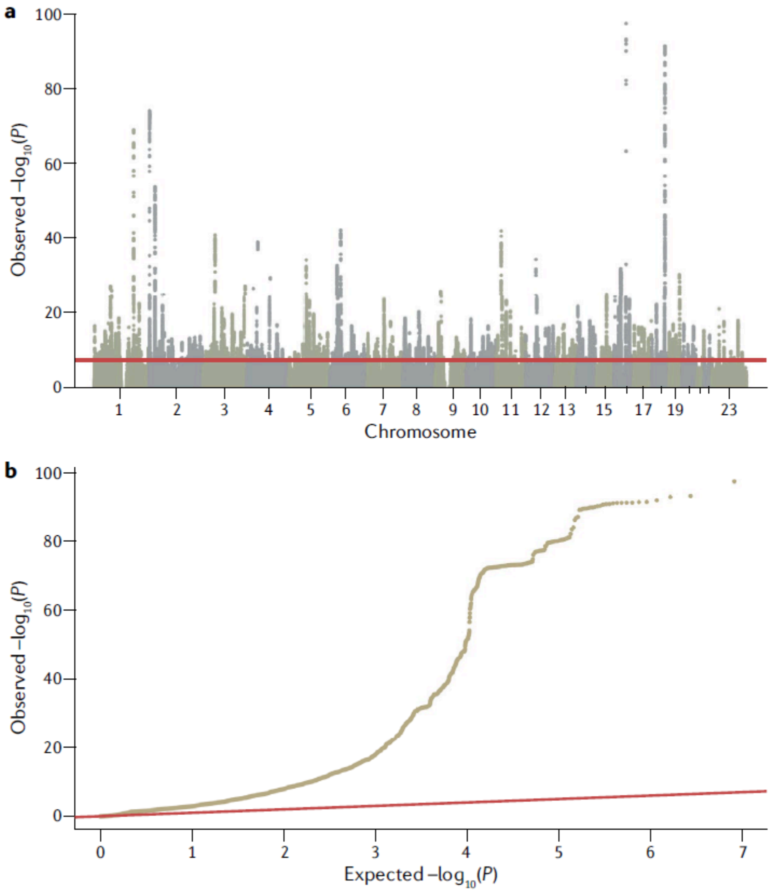


Figure 7: Common ways of visualizing GWAS results (from Uffelmann *et al.*, 2021). Panel a: Manhattan plot where the strength of association of each variant is plotted against its position on the chromosome. Two colors alternate to show even and odd chromosomes. The genome-wide significance threshold is represented by the red line. Panel b: QQ plot where the observed -log$_{10}$(p-values) are sorted and plotted against the expected -log$_{10}$(p-values).

In Manhattan plots (Figure 7a), each peak represents a genomic region or locus where multiple variants exceed the genome-wide significance threshold. One or several variants can drive the GWAS signal. If there are several causal variants, it is called allelic heterogeneity. However, many neutral variants are also associated with the phenotype because they are in LD with the causal variant. There are several reasons why neutral variants can be even more significant than causal variants, such as the fact that they can be better imputed or be in LD with two causal variants (that can cancel each other) or occur by chance, especially in the case of a small cohort (Uffelmann *et al.*, 2021). For instance, simulations demonstrated that the lead SNP was causal in 79% of cases for an odds ratio (OR) of 1.5 and a risk allele frequency (RAF) of 0.5 while it dropped to 5% of cases for an OR of 1.1 and a RAF of 0.05,, highlighting the importance of cohort size (Van De Bunt *et al.*, 2015). In QQ plots (Figure 7b), we check whether the majority of observed -log$_{10}$(p-values) are systematically higher than expected -log$_{10}$(p-values) from the chi-square distribution. Indeed, most of the observed -log$_{10}$(p-values) should be true null hypotheses (H0) and be positioned along a diagonal with a slope of 1 while only some should be true alternative hypotheses (H1) and be far from the diagonal. This deviation from expectation can be estimated with the lambda inflation factor which is obtained by dividing the median of the observed chi-squared values by the median of the expected chi-squared values. Inflation from expectation may indicate the presence of population stratification or cryptic relatedness. It means that allele frequencies differ between groups for a reason other than the phenotype of interest, such as demographic origin or relatedness (Uffelmann *et al.*, 2021). This can be addressed in three ways. The first way is to force the majority of observed -$\log_{10}$(p-values) to follow the diagonal with a slope of 1, a process called genomic control. The second way is to add genetic principal components (PCs) to the covariates to account for population stratification (Wu *et al.*, 2011). The third way is to add polygenic effects to the covariates to account for relatedness (Sillanpää, 2011). However, inflation from expectation may also suggest that there are a large, if not infinite, number of variants with small effects (polygenicity) (Uffelmann *et al.*, 2021), as suggested by Fisher in 1919 in what is now called the "infinitesimal model". These variants would not reach genome-wide significance due to lack of power resulting from the insufficient cohort sizes. This is one of the possible causes of missing heritability. LD score regression was developed to quantify the relative importance of population stratification versus polygenicity in a given GWAS. It assesses for all variants whether their strength of association is related to the size of the haplotype in which they are found: this is the case for polygenicity but not for population stratification. In practice, a regression analysis is performed between effect sizes and LD scores (the sum of r^2 with all other variants in a given window) of the variants. The slope of the regression line estimates polygenicity and heritability while the intercept estimates population stratification and cryptic relatedness. It was shown that inflation from expectation is mostly due to polygenicity rather than population stratification (Bulik-Sullivan *et al.*, 2015).

GWAS studies for IBD began in Europeans (Duerr *et al.*, 2006; Libioulle *et al.*, 2007; Burton *et al.*, 2007; Parkes *et al.*, 2007; Barrett *et al.*, 2008; Imielinski *et al.*, 2009; McGovern *et al.*, 2010; Ellinghaus *et al.*, 2016) before being applied to other populations such as Japanese (Yamazaki *et al.*, 2013), Koreans (Yang *et al.*, 2014) and North Indians (Juyal *et al.*, 2015). However, most associations could not be replicated across studies. It has been shown that in the case of small cohort size and strict threshold of multiple testing, detected associations are actually reinforced by noise and overestimated and are therefore weaker in replication studies. This is called the winner’s curse (Nakaoka & Inoue, 2009; Palmer & Pe’er, 2017). As power depends on cohort size, meta-analyses that aggregate the results of several individual GWAS were conducted (Franke *et al.*, 2010; Anderson *et al.*, 2011; Jostins *et al.*, 2012; Julià *et al.*, 2014). Then, trans-ancestry GWAS were also performed to increase the power and resolution of shared loci (Liu *et al.*, 2015, Peterson *et al.*, 2019). Moreover, a cross-disease GWAS for IBD and systemic sclerosis was performed to increase power, as multiple variants can be associated with multiple traits, which is called pleiotropy (González-Serna *et al.*, 2020). In total, 241 IBD risk loci have been identified for IBD (De Lange *et al.*, 2017).

Since LD limits the resolution of association studies, risk loci span genomic regions of approximately 250 kilobases that encompass thousands of variants and from 0 (gene deserts) to more than 50 genes. Therefore, post-GWAS analyses should be conducted to identify causal variants and genes in the GWAS peaks identified for IBD (Momozawa *et al.*, 2018).

c. Fine-mapping

Once GWAS have identified genomic loci of interest for a disease, fine-mapping is used to identify causal variants by exploring the LD structure and fitting multiple variants simultaneously. As mentioned above, few causal variants but also many neutral variants exceed the genome-wide threshold. Therefore, a large number of densely genotyped individuals is necessary to distinguish them. For IBD, fine-mapping analyses focused on 94 risk loci and were able to identify 18 causal variants with one base pair (bp) resolution as well as 51 causal variants with ≤10 bp resolution. It has been shown that few causal variants are protein-coding, meaning that they alter the primary amino acid structure and the resulting protein function, and that many causal variants are regulatory (non-coding). Therefore, monogenic diseases (such as cystic fibrosis) are dominated by coding variants while polygenic diseases (such as IBD) are dominated by regulatory variants. It is straightforward to identify causal genes when they are impacted by coding variants, especially if there are several (this is called allelic heterogeneity) (Huang *et al.*, 2017). Causal genes have been identified for 14 IBD risk loci such as *NOD2*, *ATG16L1*, *IL23R*, *CARD9*, *FUT2*, and *TYK2* (Sazonovs *et al.*, 2022). However, the identification of causal genes is less obvious in the case of regulatory variants.

Regulatory variants can affect transcriptional, post-transcriptional, translational or post-translational mechanisms of gene regulation. Since transcriptional mechanisms correspond to the layer with the largest proportion of the genome devoted to it, most regulatory variants are expected to affect gene switches, which perturbs the transcription rate of target genes (Huang *et al.*, 2017). These regulatory elements include promoters, enhancers, silencers and insulators. With the help of transcription factors (TFs), transcription begins with the recruitment of the RNA polymerase to the promoter, near the transcription start site (TSS) of the gene (Rojano *et al.*, 2019). There are three types of RNA polymerases in eukaryotes: the type I for rRNA genes, the type II for mRNA, miRNA, snRNA, and snoRNA genes, and the type III for tRNA and 5S rRNA genes (Carter & Drouin, 2009). The rate of transcription can be increased when TFs called activators bind distant regions called enhancers or decreased when TFs called repressors bind distant regions called silencers. DNA loops enable them to approach the promoter region (Rojano *et al.*, 2019). If TFs bind to insulators, this prevents interaction between the enhancer or silencer and the promoter (Raab & Kamakaka, 2010; Kolovos *et al.*, 2012). Multiple genes often share the same regulatory elements (Thurman *et al.*, 2012). It should be noted that the number of steady-state transcripts is determined not only by the transcription rate, but also by the decay rate (Pai *et al.*, 2012). Fine-mapping studies have shown that trait-associated variants are indeed overrepresented in TF binding sites and epigenetic signatures of gene switches. Follow-up analyses are needed to identify causal genes affected by regulatory variants and will be discussed in another section (Huang *et al.*, 2017).

d. Gene causality tests

Several methods have been developed to identify trait-causing genes. As a reminder, causal genes can be perturbed by one or several causal variants. Three genetic tests can be used in model organisms to identify genes responsible for trait variation between different strains or species (Figure 8). However, they are not feasible in humans. The first test is homologous recombination where the phenotype of two organisms is compared: one that is unedited and the other whose gene has been replaced by a homologous copy from a different strain or species. This remains a challenge in most species, even with CRISPR–Cas9-based genome editing technologies. The second test is transgenesis where two alleles carried by plasmids are each inserted into the genome of an organism. The third test is the reciprocal hemizygosity test and is the only one that does not require any understanding of the function of the candidate gene. It consists of comparing the phenotype of two F1 hybrids whose genomes differ only at a locus of interest. This locus is made hemizygous by genetic crosses, meaning that one of the two parental copies is deleted. This test can be performed for each putative causal gene in a quantitative trait locus (QTL) identified by linkage or association mapping (Stern, 2014). This assay was created in Drosophila (Stern, 1998) but has mainly been performed in the yeast *Saccharomyces cerevisiae* and got its name from a study where they pinpointed three closely linked genes influencing growth at high temperature (Steinmetz *et al.*, 2002). However, it has also been used in mice (Yalcin *et al.*, 2004) and cows (Karim *et al.*, 2011).

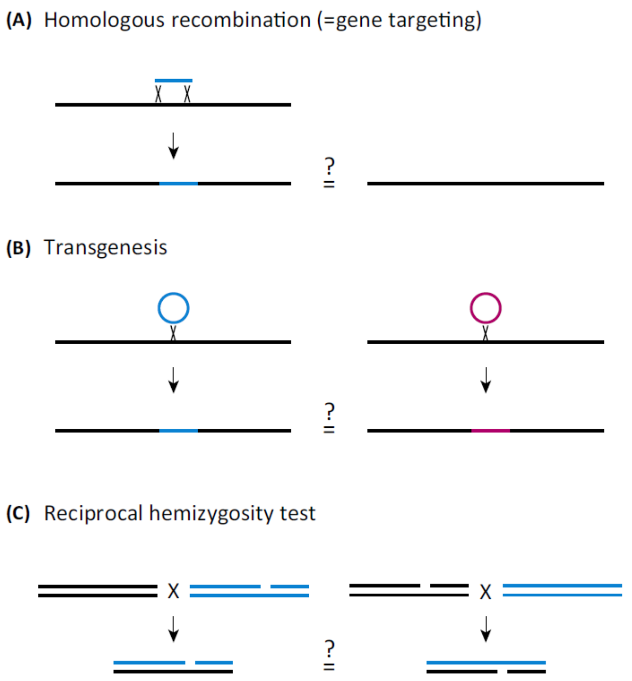

(A) Homologous recombination (=gene targeting)
?
=
(B) Transgenesis
?
=
(C) Reciprocal hemizygosity test
X
X
?
=

Figure 8: Comparison of three genetic tests to demonstrate gene causality (from Stern, 2014).

The burden test serves as an alternative test of genetic causality in humans. It consists of sequencing putative causal genes in a large number of cases and controls and investigating whether there are significantly more disruptive risk variants in cases (or more disruptive protective variants in controls), only for causal genes. The disruptive variants tested for are coding variants grouped into loss-of-function variants (frameshift, splice site and nonsense) and missense variants (Nicolae, 2016; Momozawa *et al.*, 2018). The hypothesis behind this test is that allelic heterogeneity, the fact that multiple, mostly rare, mutations in the same gene can cause the same disease, is not only applicable to Mendelian diseases but also to complex diseases (Pritchard, 2002). Several burden tests have been performed for putative IBD-causing genes. In the case of the *NOD2* gene, it was shown that CD patients carry 17% of the missense variants studied while controls and UC patients carry only 5% and 4%, respectively (Hugot *et al.*, 2001; Lesage *et al.*, 2002). In the case of the *IL23R* gene, it seems that three rare coding variants, enriched in controls, are protective for IBD. However, causality was suggested but not demonstrated for the other 62 positional candidate genes tested (Momozawa *et al.*, 2011). After that, they have sequenced 45 putative causal genes in 6,600 CD patients and 5,500 controls but, despite encouraging results, they concluded that the cohort size was still not sufficient to demonstrate causality, except for *NOD2* and *IL23R*. Another possible explanation for the low success could be that the causal variants are regulatory variants located in gene switches rather than coding variants located in open reading frames (Momozawa *et al.*, 2018). However, another study performed exome sequencing on 30,000 CD patients and 80,000 controls and confirmed the causality of a novel gene, ATG4C, again highlighting the role of autophagy (Sazonovs *et al.*, 2022). Therefore, burden tests do not seem to work in CCD.

e. Colocalization with expression quantitative trait loci

As mentioned above, non-coding variants are expected to mainly regulate transcription rates. Accordingly, they are enriched in proximal (promoter) or distant (enhancer or silencer) regulatory elements (Huang *et al.*, 2017). Variants that perturb the transcription rate of one (or more) genes are called expression quantitative trait loci (eQTLs). To determine if a variant influences the transcription rate of a target gene in a cohort of individuals, individuals need to be sorted by genotype: they carry 0, 1 or 2 alleles of interest (usually the alternate allele). Then, a regression of the number of transcripts of the gene of interest can be performed on the allelic dosage. If the regulatory variants are located on the same DNA molecule as the affected alleles, we speak of *cis*-eQTLs. Typically, they are located within one megabase of the TSS of the genes. If they are located on another DNA molecule, they are referred to as *trans*-eQTLs (Nica & Dermitzakis, 2013). Another characteristic that distinguishes *cis*-eQTLs from *trans*-eQTLs is allelic imbalance (Figure 9). It consists of studying the sequenced reads of samples from individuals heterozygous for the variant of interest. If we observe that more reads correspond to one of the two alleles, it is a *cis*-eQTL. Otherwise, there is no allelic imbalance and it is a *trans*-eQTL, even if the regulatory variants are located on the same DNA molecule as the target gene (Castel *et al.*, 2020). Genetic differences modulate gene expression levels of most genes (GTEx Consortium, 2017). Some genes can be affected by multiple eQTLs with independent effects and these can be detected using conditional analyses. One approach consists of using the top variant as covariate and repeating the eQTL mapping until no new significant variant is detected (Delaneau *et al.*, 2017).

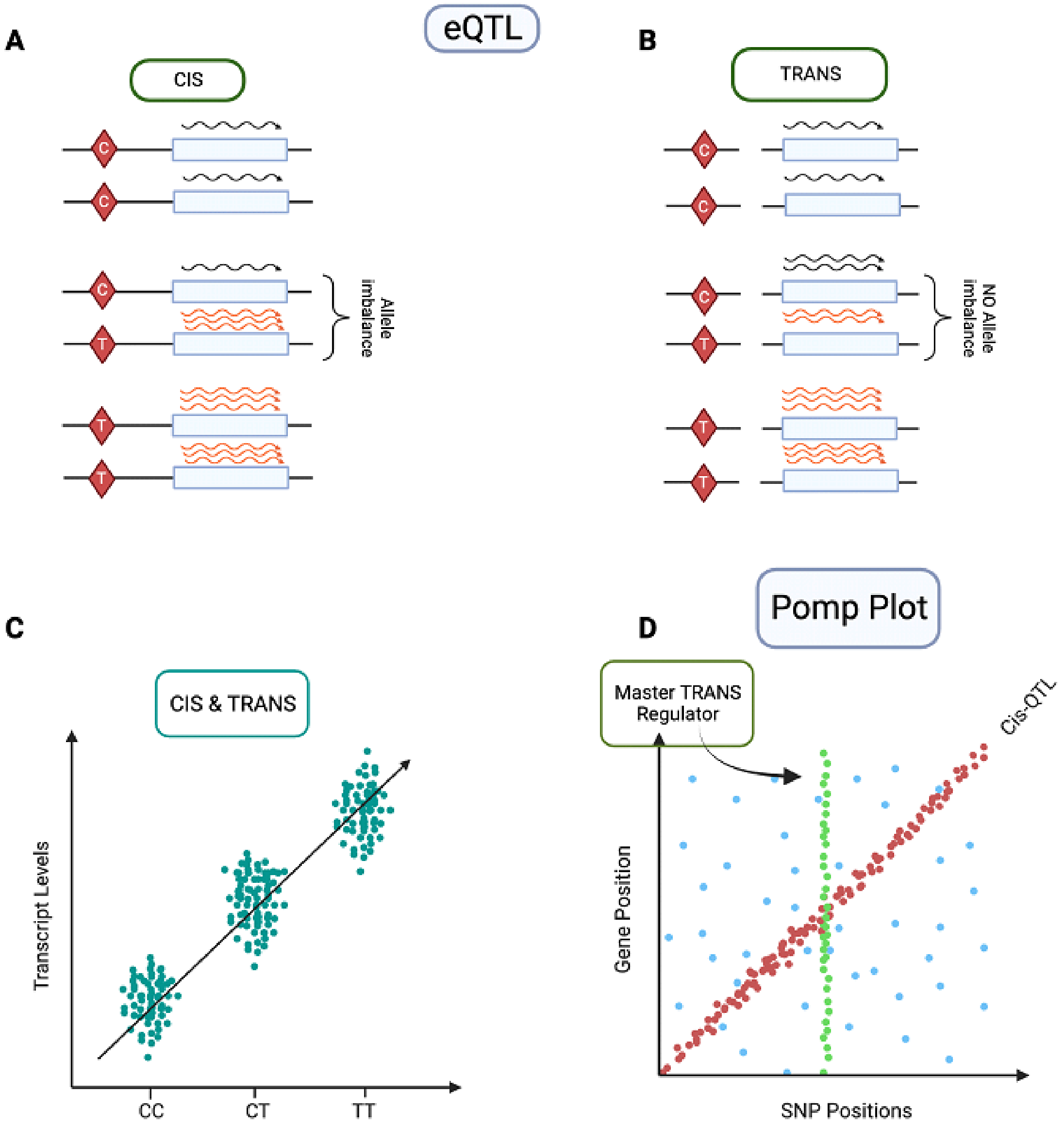


Figure 9: Comparison between *cis*- and *trans*-eQTL analyses. An enhancer region (diamond) controls the transcription rate of a target gene (rectangle) and contains a SNP where C is the reference allele and T the alternate allele. Panel A: the SNP influences the transcription rate of a gene located on the same DNA molecule and is called *cis*-eQTL. It is characterized by allelic imbalance in heterozygotes. Panel B: the SNP influences the transcription rate of a gene located on another DNA molecule and is called *trans*-eQTL. It is not characterized by allelic imbalance in heterozygotes. Panel C: in the case of eQTLs, the transcription rate is influenced by the genotype of the individual. Panel D: Pomp plot that summarizes *cis*- and *trans*-eQTL analyses. Each dot corresponds to a significant association between a SNP and the expression level of a gene. In the case of *cis*-eQTLs, we see a diagonal line because the SNPs are located on the same DNA molecule as the genes. In the case of *trans*-eQTLs, we observe dots distributed everywhere and vertical lines that correspond to master regulators.

eQTL analyses require establishing a cohort of unrelated individuals of the same ancestry to study the effect of common variants on target genes. Expression data are typically produced by RNA sequencing of cell type-specific samples (that are relevant for the phenotype of interest) and genotype data are generated by SNP genotyping (or DNA sequencing but it is still more expensive). These analyses are generally performed in healthy individuals. Indeed, the risk variants underlying the GWAS signals are common (MAF ≤ 0.05) and segregate in the population, including in unaffected individuals. However, these individuals remain healthy because they do not have a sufficient number of risk factors (both environmental and genetic) to exceed a threshold value of liability (Momozawa *et al.*, 2018). Concerning genotyping data, variants should not deviate from Hardy–Weinberg equilibrium because it could suggest a genotyping error (Anderson *et al.*, 2010). Regarding sequencing data, genes should be normalized after removing low expressed genes that are too noisy (Table 3) (Wang *et al.*, 2021). Besides these concerns, other aspects should be considered in eQTL analyses (Ko *et al.*, 2024).

Method	**Correction**	**Steps**
Counts per million (CPM)	sequencing depth (within-sample)	divide the number of reads of each gene by the total number of reads in this sample and by a million
Transcripts per million (TPM) (Zhao *et al.*, 2021)	sequencing depth and gene length (within-sample)	1. divide the number of reads by the gene length 2. divide the number of reads of each gene by the total number of reads in this sample and by a million
Median of ratios normalization implemented in DESeq2 (Anders and Huber, 2010; Love *et al.*, 2014)	library size and RNA composition (between-sample)	1. compute the geometric mean per gene (genes without 0s) 2. divide the raw counts by the geometric mean 3. compute the median of ratios per sample 4. divide the raw counts by the median of ratios
Rank-based inverse normal transformation (INT) (Beasley *et al.*, 2009)	make the phenotype normally distributed (within-sample)	1. take the ranks of the counts 2. subtract 0.5 from the ranks 3. divide the ranks by the number of observations 4. take their normal quantile

Table 3: Steps to perform several normalization methods.

First, we need to ensure that no bias is introduced when the sequenced reads are mapped to a reference genome. It is particularly important because we want to estimate the effect of reference and alternate alleles on transcription rate. Otherwise, this could lead to false signals in eQTL analyses. Indeed, reads that contain alternate alleles are often not mapped uniquely or mapped to an incorrect region. The method WASP was developed to correct this potential allelic bias (Figure 10). The software evaluates whether reads map to the reference genome, regardless of whether they contain the reference allele or the alternate allele. Reads are kept if they map in both cases and discarded if they map in one case (Van De Geijn *et al.*, 2015). The WASP algorithm has been re-implemented in the well-known STAR mapping algorithm (Dobin & Gingeras, 2015; Asiimwe & Dobin, 2024). It should be noted that ASElux is a software capable of counting allele-specific reads in a sample, but not of generating mapped reads (Miao *et al.*, 2018).

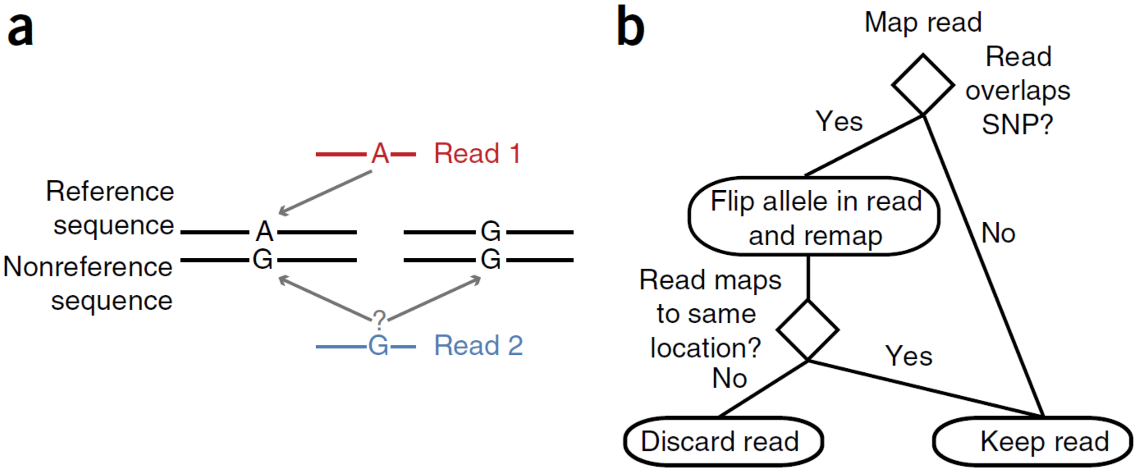


Figure 10: How the WASP method corrects mapping bias of allele-specific reads (from Van De Geijn *et al.*, 2015). Panel a shows that the red read containing the reference allele maps to the correct location while the blue read containing the alternate allele maps to an incorrect location. Panel b shows the WASP pipeline: reads are discarded if they map to the genome when they contain the reference allele but not when they contain the alternate allele.

Second, an individual's genotype needs to match the sequenced reads of their samples. For example, a method in QTLtools is dedicated to match BAM files to VCF files (MBV). The BAM file contains the sequenced reads for a particular cell type from a given individual while the VCF file contains the genotypes of all individuals. The method uses these two types of files to detect mislabelled individuals, contamination between samples from different individuals and PCR amplification bias (Figure 11). It computes for each individual present in the VCF file the fraction of homozygous and heterozygous genotypes which are concordant with the BAM file. The proportion of concordant homozygous sites is then plotted against the proportion of concordant heterozygous sites. The RNA sample can be of good quality but match an unexpected individual, indicating a mislabelling. If the concordance at homozygous sites decreases for the contaminated individual while the concordance at heterozygous sites increases for the contaminating individual, it indicates a problem of contamination. On the contrary, if the concordance at heterozygous sites decreases for the real individual while the concordance at homozygous sites increases for all other individuals, it indicates a problem of amplification bias (Fort *et al.*, 2017).

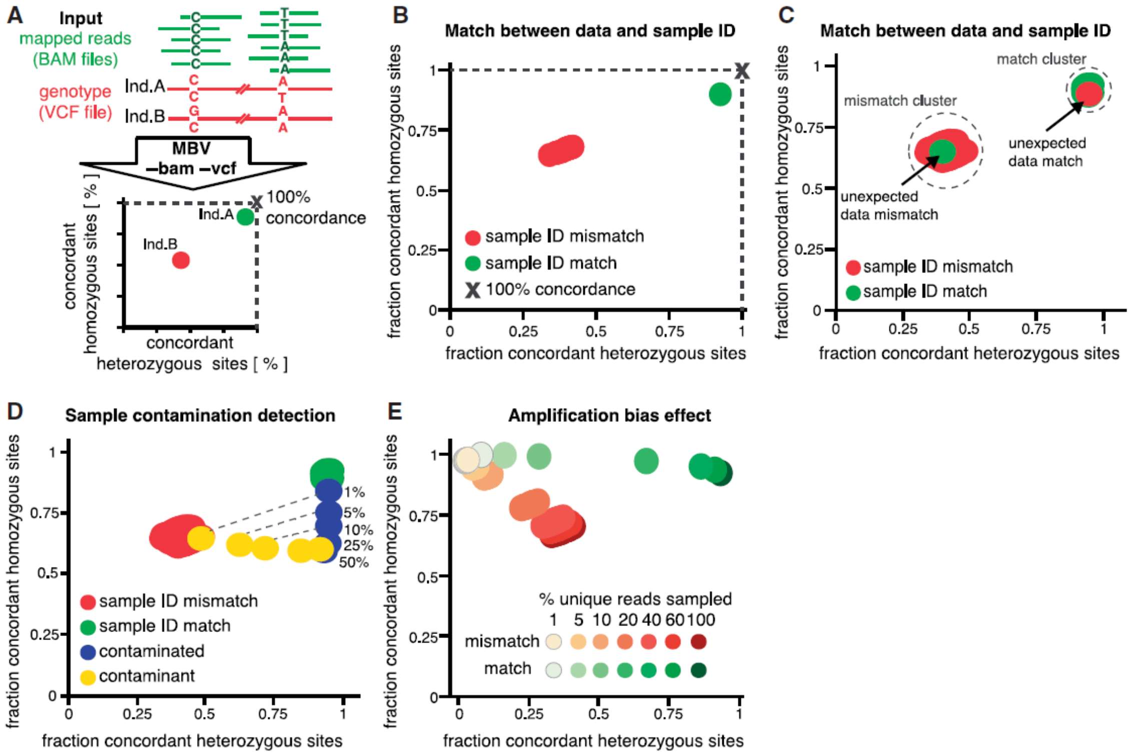


Figure 11: How the MBV method detects sample mislabelling, contamination and amplification bias (from Fort *et al.*, 2017). Each dot corresponds to the BAM file of an RNA sample. A match corresponds to a concordance at homozygous and heterozygous sites of 1 while a mismatch corresponds to a concordance at homozygous and heterozygous sites of 0.5. Matches and mismatches can be expected or not. The presumed donor is colored in green while the other individuals are colored in red. Panel A shows the content of a BAM file and a VCF file for two SNPs and two individuals as well as the resulting plot. Panel B shows an example where no problem is detected: as expected, the RNA sample matches the DNA of its presumed donor. Panel C shows an example of mislabelling: unexpectedly, the RNA sample matches the DNA of an individual that is not the presumed donor. Panel D shows an example of contamination. Panel E shows an example of amplification bias.

Third, another important aspect in eQTL analyses is to account for measured confounders (such as batch, sex and age) and unmeasured technical, environmental or demographic confounders (Leek & Storey, 2007). The unknown confounders are inferred from the genotype and phenotype data and included as covariates in the linear regression. Concerning the genotype data, it is widely accepted to use the 3-5 top genetic PCs to account for population stratification (Wu *et al.*, 2011). Regarding the phenotype data (i.e. gene expression levels), two methods are often used to infer the unknown covariates: the principal component analysis (PCA) and the probabilistic estimation of expression residuals (PEER) (Stegle *et al.*, 2010; Stegle *et al.*, 2012). A comparison between PCA and PEER methods showed that PCA performed better in addition to being significantly faster and easier to use and to interpret. The number of top expression PCs included in the model is the one that maximizes the detection of eQTLs. PEER has the advantage of being more flexible: it allows to include known covariates and returns either the inferred covariates (factor approach) or the expression data after removing the covariates (residual approach). However, there is no consensus on whether to use known covariates and the choice of factor or residual approach. The fact that different choices are made in different published studies decreases reproducibility (Zhou *et al.*, 2022). It is important to mention that, for *trans*-eQTLs, expression PCs or PEER factors should not be associated with genotype for inclusion in the model, as this may indicate a *trans*-eQTL hotspot (Stegle *et al.*, 2012).

Fourth, *cis*-eQTL mapping is performed in a 1 or 2 Mb window centered on each gene TSS. This means that the linear regression is repeated for each variant in the window and for each gene in each cell type. In the case of *trans*-eQTLs, the number of tests is incredibly higher because the effect of each variant on any chromosome is tested for each gene in each cell type. Each linear regression outputs the effect size (corresponding to the slope or beta) and the (nominal) p-value. Multiple testing is accounted for by correcting p-values first in each window and then in each cell type (Momozawa *et al.*, 2018). To correct the nominal p-values in each window, we perform 10,000 permutations of the phenotype and keep the best (adjusted) p-value for each gene in each cell type (Delaneau *et al.*, 2017). To correct the p-values in each cell type, the false discovery rate (FDR) test is applied on a list of adjusted p-values, sorted from smallest to largest, that correspond to the top variants of each gene tested. The expectation is that the p-values follow a uniform distribution when all tests are true H0. However, the p-values are shifted to low p-values when some tests are true H1 (Figure 12). The FDR or q-value is defined as the proportion of false positives among the equally or more significant tests. Therefore, p-values are related to the false positive rate (FPR) while q-values are related to the FDR (Benjamini & Hochberg, 1995; Storey & Tibshirani, 2003). In our example, an FPR of 5% indicates that 5% of neutral variants are significant while an FDR of 5% indicates that 5% of significant variants are neutral. The definition of these rates is provided in Table 4.

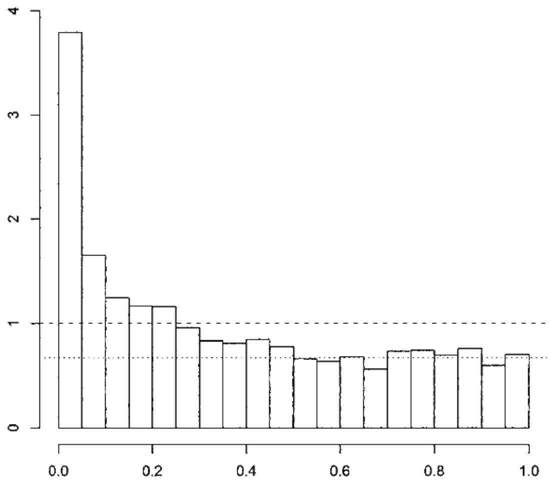


Figure 12: Density histogram of p-values shifted to low p-values (from Storey & Tibshirani, 2003). The dashed and dotted lines correspond to the expected and observed proportions of null p-values, respectively.

	H0 true	**H0 false**	
H0 rejected	false positive (FP)	true positive (TP)	FDR=FP/(FP+TP)
H0 not rejected	true negative (TN)	false negative (FN)	
	FPR=FP/(FP+TN)		

Table 4: Formulas for calculating false positive and discovery rates (adapted from Storey & Tibshirani, 2003).

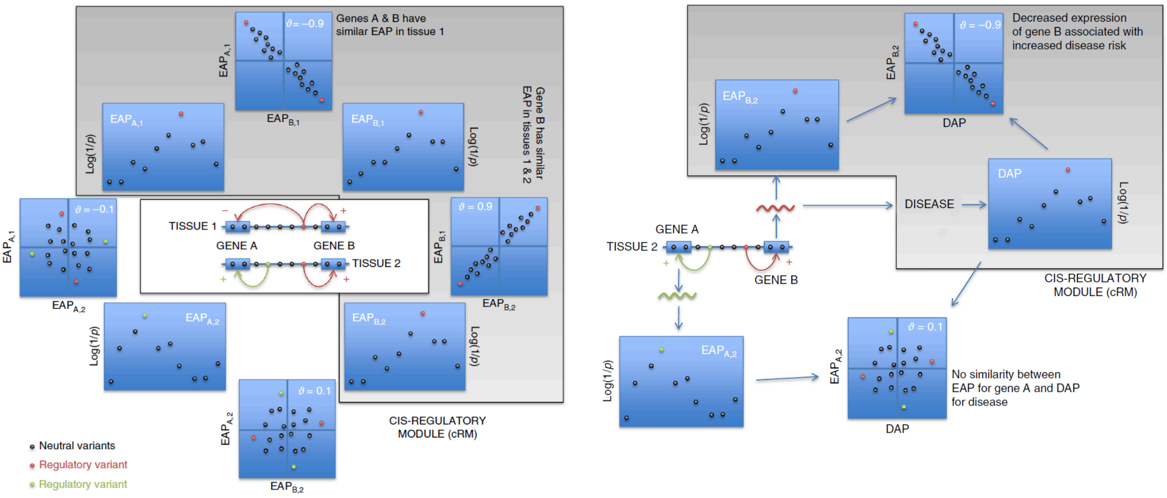
Relevant cell types for IBD include intestinal cell types as well as circulating immune cell types. Three methods can be used to detect cell type-specific eQTLs: bulk RNA sequencing (RNA-Seq) of purified tissues, single-cell RNA sequencing (scRNA-Seq) of bulk tissues and deconvolution to predict cell type-specific eQTLs in whole blood samples (Zhang & Zhao, 2023). A non-exhaustive list of published eQTL datasets is provided in Table 5.

Dataset	**Tissue**	**Method**	**Reference**
GEUVADIS	lymphoblastoid cell lines	RNA-Seq	The Geuvadis Consortium *et al.*, 2013
DGN	whole blood	RNA-Seq	Battle *et al.*, 2014
GTEx	49 tissues from post-mortem	RNA-Seq	Lonsdale *et al.*, 2013

	healthy individuals		The GTEx Consortium *et al.*, 2020
CEDAR	6 blood immune cell types and biopsies of 3 intestinal locations from healthy individuals	arrays	Momozawa *et al.*, 2018
DICE	15 blood immune cell types from healthy individuals	RNA-Seq	Schmiedel *et al.*, 2018
eQTLGen	meta-analysis in whole blood	RNA-Seq	Võsa *et al.*, 2021
ImmuNexUT	28 blood immune cell types from healthy individuals and 10 classes of immune-mediated diseases	RNA-Seq	Ota *et al.*, 2021
OneK1K	1 million of PBMCs	scRNA-Seq	Yazar *et al.*, 2022
/	whole blood	deconvolution	Aguirre-Gamboa *et al.*, 2020

Table 5: Non-exhaustive list of published eQTL datasets.

The goal of colocalization is to determine whether the association signals of two traits at a given locus are driven by the same or distinct variants (Zuber *et al.*, 2022). The two cohorts do not need to come from the same cohort, provided that the two cohorts are of the same ethnicity so that they share the same LD structure. For example, the two different datasets can be two eQTL analyses or one GWAS study and one eQTL analysis. In the first example, we look for variants that perturb the expression of one or several genes in one or several tissues. In the second example, we look for the variants that increase disease risk by perturbing transcription rates. We will focus on the second example in the explanation below. The colocalization method used in this study is the theta (Φ) metric (Momozawa *et al.*, 2018). In brief, the -log$_{10}$(p) values coming from GWAS studies (referred to as the disease association pattern ; DAP) are compared to the -log$_{10}$(p) values coming from eQTL analyses (the expression association pattern ; EAP) for the variants which are significant (-$\log_{10}$(p) values ≥ 1.3) in at least one of the two association patterns. The correlation is weighted and signed, meaning that it takes into account the significance of each variant (the weight) as well as its direction of effect (the sign) in each association pattern. We can then evaluate if the two association patterns are similar and in that case, if the regulatory variants increase the disease risk by increasing gene expression (theta close to 1) or by decreasing gene expression (theta close to -1). An absolute value of theta greater or equal to 0.6 was the threshold used in the Momozawa' study. A p-value for theta can be obtained by comparing the obtained theta with real and permuted data. In practice, we calculate theta when permuting only the EAP and then only the DAP, after which we average the two p-values. Momozawa *et al.* (2018) used the same metric to assess whether two eQTL association signals for the same gene in two different cell types or for two genes in the same or different cell type(s) are driven by the same variants in order to build regulatory modules (Momozawa *et al.*, 2018) (Figure 13). It should be noted that it is important to subset for the variants that are present in both datasets and to ensure that the tested alleles are the same in both datasets. If not, the tested allele may need to be adjusted in one study by reversing the sign of the z-score and the effect size (Zhao *et al.*, 2024).



$\vartheta = -0.9$
Genes A & B have similar EAP in tissue 1
$EAP_{A,1}$
$EAP_{B,1}$
Gene B has similar EAP in tissues 1 & 2
Log(1/p)
$\vartheta = -0.1$
$EAP_{A,2}$
– TISSUE 1 +
GENE A
GENE B
TISSUE 2
+
+
$\vartheta = 0.9$
$EAP_{B,2}$
$\vartheta = 0.1$
CIS-REGULATORY MODULE (cRM)
Neutral variants
Regulatory variant
Regulatory variant
Decreased expression of gene B associated with increased disease risk
DAP
DISEASE
TISSUE 2
No similarity between EAP for gene A and DAP for disease

Figure 13: Two examples of colocalization tests (from Momozawa *et al.*, 2018). Left panel: four EAPs (eQTL in gene A or B from tissue 1 or 2) are compared two by two. The eQTL in gene A from tissue 1 is negatively correlated with the eQTL in gene B from tissue 1. It means that the variants increase the expression of one gene while they decrease the expression of the other gene. The eQTL in gene B from tissue 1 is positively correlated with the eQTL in gene B from tissue 2. It means that the variants increase or decrease the expression in both tissues. As these three eQTLs are driven by the same variants (as assessed by theta), they are therefore clustered into the same *cis*-regulatory module (cRM). Right panel: a DAP is compared to two EAPs (eQTL in gene A or B from tissue 2). The eQTL in gene B is negatively correlated with the DAP. It means that variants increase disease risk by decreasing the expression of this gene. They are therefore clustered into the same cRM.

This method gave comparable results to a very similar approach called summary Mendelian randomization (SMR) (Momozawa *et al.*, 2018). The difference is that the SMR method compares the effect sizes instead of the -$\log_{10}$(p) values (Zhu *et al.*, 2016; Wu *et al.*, 2018). Both methods determine whether two association patterns are driven by the same variants, but they are unable to distinguish causal variants from neutral variants (Momozawa *et al.*, 2018). Both approaches are different from other colocalization methods such as RTC (Nica *et al.*, 2010), Sherlock (He *et al.*, 2013), coloc (Giambartolomei *et al.*, 2014; Wallace, 2020; Wallace, 2021) and eCAVIAR (Hormozdiari *et al.*, 2016). The RTC method performs the eQTL analysis with the GWAS top SNP as covariable to determine if the association with gene expression is still present (they are driven by the same variants) or not (they are driven by distinct variants). This is repeated with random SNPs from the same region to confirm that this is not due to chance (Nica *et al.*, 2010). The coloc method evaluates whether two traits are driven by the same causal variants (identified by fine-mapping) and requires p-values and MAFs. It can distinguish four alternative hypotheses as shown in Figure 14 (Giambartolomei *et al.*, 2014).

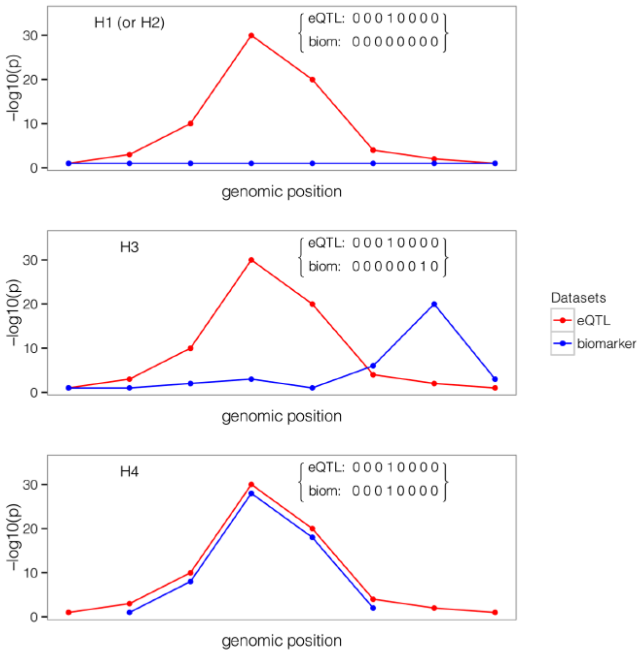


Figure 14: The four alternative hypotheses for a pair of vectors representing eight variants in a genomic region that are causally involved (value of 1) or not (value of 0) in two traits (from Giambartolomei *et al.*, 2014). In the first graph, one causal variant drives the association of one trait (H1 or H2). In the second graph, two causal variants each drive the association of one trait (H3). In the third graph, one causal variant drives the association of both traits (H4).

f. The omnigenic model

Very large GWAS studies have shown that most of the genome contributes to the inheritance of most traits. Pritchard has therefore proposed the omnigenic model. He stated that most causal variants are indeed regulatory variants disrupting genetic switch components. However, they would not cause the trait through *cis*-eQTL effects but by affecting gene regulatory networks (peripheral genes) in a way that converges on some disease-specific core genes (Boyle *et al.*, 2017). Most heritability would be driven by weak *trans*-eQTLs variants in a way that could be amplified if the core genes are co-regulated. There would be approximately 100 peripheral genes for one core gene. The goal would be to identify these core genes because they would be those involved in the pathogenesis of the disease (Liu *et al.*, 2019). Nevertheless, other researchers suggest that we can not exclude that some common diseases would be affected by many core genes, not just a few and that the priority remains to increase cohort sizes (Wray *et al.*, 2018).

g. Mendelian randomization

Mendelian randomization (MR) is an additional post-GWAS analysis method that was developed to distinguish correlation from causation. Indeed, observational epidemiological studies can not make this distinction because they face biases such as confounding (in which an exposure and an outcome are associated because they are influenced by another variable) or reverse causation (what we thought was the outcome is actually the exposure). Their assumptions are therefore assessed in expensive randomized controlled trials in which individuals are randomly assigned to an exposure group (control or treatment) by an investigator to assess a causal effect on an outcome. In comparison (Figure 15), MR is considered a natural experiment because individuals are assigned to an exposure group based on genetic variants randomly inherited from their parents (Richmond & Davey Smith, 2022; Zuber *et al.*, 2022; Burgess *et al.*, 2023).

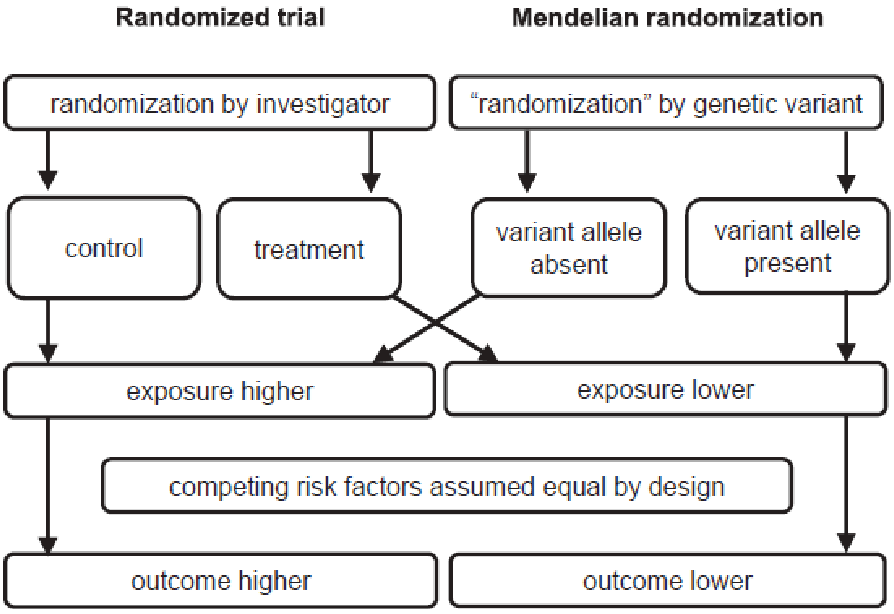


Figure 15: Comparison of randomized trial and Mendelian randomization (from Burgess *et al.*, 2023).

Typically, the exposure in MR is a modifiable environmental factor and the outcome is a disease. Genetic variants are considered instrumental variables and should only be associated with the exposure (relevance assumption), not the outcome (independence assumption). They should affect the outcome through exposure rather than through confounding factors (exclusion restriction assumption). This idea that variants should be correlated with the phenotype of interest but not with confounding factors comes from Mendel's second law, which states that alleles segregate independently during meiosis. MR studies have two advantages: reverse causation is impossible with variants, and they can be performed quickly with currently available data resources. However, a possible concern regarding the validity of MR analyses is that pleiotropy, the fact that multiple variants can be associated with multiple traits, threatens the exclusion restriction assumption. It remains valid in the case of vertical pleiotropy where the variants influence the outcome solely through the exposure but it is not applicable in the case of horizontal pleiotropy where the variants impact both the exposure and the outcome through distinct pathways (Richmond & Davey Smith, 2022; Zuber *et al.*, 2022; Burgess *et al.*, 2023). In practice, the effect sizes of the significant variants associated with trait 1 are plotted against the effect sizes of the same variants with trait 2. This is repeated when reversing traits 1 and 2. A good correlation is observed when we include the variants associated with exposure while a poor correlation is observed when we include the variants associated with outcome (Pingault *et al.*, 2018).

The causal relationship between multiple exposures and IBD has been assessed. The impact of gut microbiome composition was evaluated to determine whether particular microbial species play a causal role in IBD. It has been shown that six gut bacterial genera are causally associated with IBD (half of them with increased risk and the other half with decreased risk) but no reverse causation was observed (Liu *et al.*, 2022). An even more recent study revealed that 23 microbial taxa are linked to a higher risk of IBD, while 17 microbial taxa are linked to a lower risk of IBD (Li *et al.*, 2024). Another application was to investigate whether specific inflammatory cytokines are causally involved in IBD pathogenesis or are simply markers of disease activity. It was suggested that IL6 (Wang *et al.*, 2024), IL13 (Song *et al.*, 2024) and IL17 (Cai *et al.*, 2023; Liu *et al.*, 2023) have a causal effect on IBD. However, the study of the causal effect of smoking and dietary factors has shown that IBD does not appear to be causally affected by smoking (Jones *et al.*, 2020; Georgiou *et al.*, 2021) or vitamin D deficiency (Lund-Nielsen *et al.*, 2018). Other MR studies have examined the causal relationship between IBD and certain other diseases. For instance, it has been observed that CD is causally associated with an increased risk of psoriasis and psoriatic arthritis (Freuer *et al.*, 2022) as well as osteoporosis (Dai *et al.*, 2023).

h. Transcriptome-wide association studies

Another approach for linking gene expression to disease risk is through transcriptome-wide association studies (TWAS). It involves predicting gene expression from a large GWAS dataset using a smaller eQTL dataset and then assessing the association of the imputed gene expression with any phenotype of interest. The main advantage of TWAS is to base the prediction on all the independent *cis*- and *trans*-eQTLs that affect the transcription rate of the gene. Another advantage of TWAS is that it relies less on large sample sizes. Indeed, it is less influenced by multiple testing because it performs gene-level association tests rather than variant-based association tests. Two different input data can be used in TWAS analyses: the genotypes of each GWAS participant or the GWAS summary statistics. Individual-level genotype data are rarely made available and require more computation time (Li & Ritchie, 2021; Mai *et al.*, 2023).

TWAS analyses have been quite successful for IBD. In the first TWAS analysis, the PrediXcan method was developed on individual-level genotype data and GTEx (Lonsdale *et al.*, 2013), GEUVADIS (The Geuvadis Consortium *et al.*, 2013) and DGN (Battle *et al.*, 2014) eQTL datasets and enabled to identify that a higher expression of *ATG16L1*, *IL23R* and *APEH* and a lower expression of *ZNF300*, *NKD1*, *BSN*, *GPX1* and *SLC22A5* were associated with CD risk (Gamazon *et al.*, 2015). Several years later, the Summary-PrediXcan and Summary-MultiXcan methods were used on University of Barcelona and University of Virginia RNA sequencing project (BarcUVa-Seq) and correlated expression and disease association research (CEDAR) eQTL datasets and enabled to identify 186 new susceptibility genes for IBD, including 39 that are specific to the colon (Díez-Obrero *et al.*, 2022).

i. Proteome-wide association studies

Recently, proteome-wide association studies (PWAS) have been developed to investigate the association between protein levels and traits. This approach attempts to colocalize protein quantitative trait loci (pQTLs) with GWAS results to discover genomic loci that increase disease risk by perturbing protein abundance. The benefit of analyzing the proteome instead of the transcriptome is that proteins are the end products of protein-coding genes and can serve as potential drug targets and biomarkers. A PWAS analysis of plasma samples using MR identified five proteins associated with an increased risk of IBD (ERAP2, RIPK2, TALDO1, CADM2 and RHOC), two proteins with an increased risk of CD (MST1 and FLRT3) and three proteins with an increased risk of UC (VSIR, HGFAC and CADM2). CADM2, a cell adhesion molecule, is a susceptibility protein for both IBD and UC (Bai *et al.*, 2024).

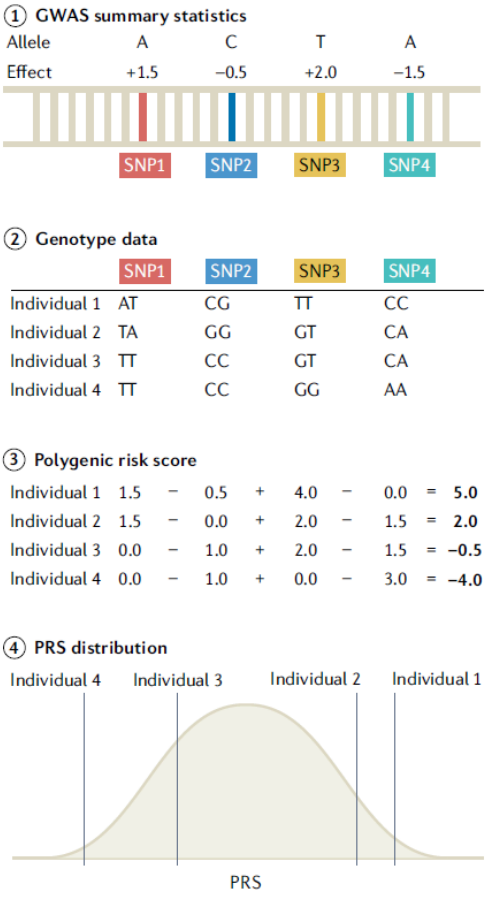
Another PWAS analysis using MR, this time performed on brain samples, showed that five proteins were associated with a decreased risk of IBD (GPSM1, AUH, TYK2, SULT1A1 and FDPS), three proteins with a decreased risk of CD (FDPS, SULT1A1, and PDLIM4) and one protein with a decreased risk of UC (AUH) (Xu *et al.*, 2024).

3. Towards personalized IBD treatment

We have already seen in the previous chapter that the identification of genetic variants associated with a disease allows us to better understand its genetic basis. However, GWAS results can be used for two other applications: the development of polygenic risk scores to identify individuals at increased risk of developing complex diseases and the discovery of targets for new (drug development) or existing (drug repurposing) therapies (Uffelmann *et al.*, 2021).

a. Polygenic risk scores

A first application of GWAS results is the development of polygenic risk scores (PRS) to provide personalized genetic prediction of CCD risk. As previously discussed, CCD are polygenic, meaning that their etiology depends on many variants with small effects. Each one of us carries some risk alleles for all CCD in a unique combination and the more risk alleles we carry, the more likely we are to develop the disease. A simple way to calculate a PRS for an individual is to take the cumulative sum of risk alleles multiplied by their effect sizes as shown in Figure 16. Many different tools exist for selecting SNPs to include in the prediction and for applying weights to them (Uffelmann *et al.*, 2021). Interestingly, non-additive models have started to be developed to consider interactions among SNPs (Elgart *et al.*, 2022). This concept is called genomic breeding value in livestock.



	SNP1	SNP2	SNP3	SNP4
Individual 1	AT	CG	TT	CC
Individual 2	TA	GG	GT	CA
Individual 3	TT	CC	GT	CA
Individual 4	TT	CC	GG	AA

Individual 1	1.5	−	0.5	+	4.0	−	0.0	=	**5.0**
Individual 2	1.5	−	0.0	+	2.0	−	1.5	=	**2.0**
Individual 3	0.0	−	1.0	+	2.0	−	1.5	=	**−0.5**
Individual 4	0.0	−	1.0	+	0.0	−	3.0	=	**−4.0**

Figure 16: The four steps to calculate PRS (from Uffelmann *et al.*, 2021). Step 1: Get GWAS summary statistics to know the effect size of each variant in a given trait. Step 2: Genotype a cohort of individuals to know the distribution of PRS in that population (in this example, four individuals were genotyped for four SNPs). Step 3: Calculate PRS for each individual as the cumulative sum of risk alleles (0, 1, 2) multiplied by their effect size. Step 4: Look at where each individual falls in terms of risk.

PRS have two advantages: the fact that the identification of causal variants is not mandatory and the ability to use the same genotyping data (collected once in a lifetime) for different diseases. Once the risk prediction is confirmed in individuals whose disease status is known, PRS can be calculated for the individuals for whom we wish to predict disease risk and compared to the PRS distribution in a cohort of the same ethnicity to obtain a relative risk. More cohorts of ethnic groups other than European are thus required. It is worth noting that PRS do not take into account environmental factors and these may change with age. Moreover, PRS explain less than the proportion of phenotypic variance due to genetic variance (heritability) because only risk variants present in at least 1% of the population are included in the calculation. This will be improved with larger GWAS cohorts (Wray *et al.*, 2021).

PRS can be used for patient stratification. For people at high risk of developing the disease, we could encourage them to get screened earlier and more frequently, to make lifestyle changes or to take preventive treatments. For example, an early detection of increased intraocular pressure and intervention can lead to prevention of glaucoma. In coronary artery disease and breast cancer, PRS has been shown to be as predictive as the screening of rare variants with large effects. For people at low risk of developing the disease, we could reduce invasive testing and preventive treatments and therefore costs. Unless they have a family history for the disease, in which case we could test for other genetic risk factors such as structural variants and chromosomal rearrangements (Wray *et al.*, 2021).

PRS could also help with clinical decisions. For instance, it could guide a diagnosis in cases of unclear symptoms, such as distinguishing between type 1 and type 2 diabetes, or advise on when to perform a mastectomy to prevent breast cancer in carriers of BRCA1 mutations. In the future, PRS could be useful in treatment choices. Once cohorts of patients treated with different drugs are established, prediction of treatment response and development of adverse events will be assessed (Wray *et al.*, 2021).

b. Drug development and repurposing

A second application of GWAS and eQTL analyses is the genetically informed target prioritization, namely the discovery of disease-causing genes that can become targets for new or existing drugs. On the one hand, the process of discovering and developing new drugs is long, risky and expensive. It has been shown that about

90% of drugs that undergo clinical trials fail to obtain approval, a phenomenon known as attrition, mainly due to insufficient efficacy and safety. Reasons could include a lack of understanding of the molecular basis of human diseases and overuse of non-human models. Better understanding through large-scale human genomic studies should help overcome these limitations (Trajanoska *et al.*, 2023). Indeed, it has been shown that drugs supported by genetics are 2.6 times more likely to succeed in the development process (King *et al.*, 2019; Minikel *et al.*, 2024). It has also been suggested that individuals most likely to respond to therapies should be selected in drug development, which would require additional consent for the use of their genetic information. On the other hand, the process of repurposing existing drugs allows for skipping the drug discovery phase, which lasts 5 to 15 years, and entering directly into clinical trials with drugs that are based on genetic evidence and that have been proven safe. Drug repurposing takes advantage of pleiotropy, the fact that multiple risk loci and genes are associated with multiple traits (Trajanoska *et al.*, 2023).

One drug previously approved for another indication has already been successfully repurposed for IBD. An antibody that targets IL12 and IL23, ustekinumab, was initially indicated for the treatment of psoriasis. Following the identification of an association between *IL23R* variants and IBD through GWAS, indication was subsequently extended for the treatment of CD (Trajanoska *et al.*, 2023) and UC (Uchida *et al.*, 2023). Other successes include 5α reductase inhibitors that have been repurposed to prevent hair loss in addition to treating benign prostatic hyperplasia, sulfonylureas that have been repurposed for a rare type of neonatal diabetes in addition to type 2 diabetes, FGFR tyrosine kinase inhibitors that have been repurposed for achondroplasia in addition to cholangiocarcinoma and bladder cancer, and IL17A signaling inhibitors that have been repurposed for ankylosing spondylitis in addition to psoriasis, rheumatoid arthritis and uveitis (Trajanoska *et al.*, 2023).

Two approved drugs are under development for CD. An antibody that targets TNFRSF11A, denosumab, is indicated to treat osteoporosis in postmenopausal women at high risk of fractures and is under development for CD after the identification of an association between TNFRSF11A variants and CD through GWAS. Moreover, an antibody that binds to IL18, which proved ineffective in treating diabetes, is now undergoing clinical trials for the treatment of CD and atopic dermatitis. Interestingly, an inhibitor of TYK2, baricitinib, was approved for rheumatoid arthritis and entered a clinical trial to treat COVID-19 one year after the identification of an association between TYK2 variants and COVID-19 through GWAS. This gene, which is part of the JAK family, is also associated with systemic lupus erythematosus and IBD (Trajanoska *et al.*, 2023). Three JAK inhibitors (upadacitinib, filgotinib and tofacitinib) are already used to treat IBD and five TYK2 inhibitors (brepocitinib, deucravacitinib, zasocitinib, VTX958 and OST-122) are in phase 2 clinical trials (Vieujean *et al.*, 2025).

OBJECTIVES

IBD is a CCD characterized by chronic inflammation of the GI tract resulting from a dysregulated immune response to the commensal microbiota in genetically predisposed individuals exposed to environmental risk factors. The etiology of IBD is strongly influenced by genetics, with GWAS identifying 241 risk loci associated with the disease. The typical span of a risk locus is approximately 250 kb, encompassing between 0 and 100 genes. Coding variants are responsible for disease associations at only 14 of these risk loci. It has been shown that the majority of observed associations can be attributed to regulatory variants that perturb the expression of one or more neighboring genes in *cis*.

eQTL analyses in disease-relevant cell types obtained from healthy individuals are one method for identifying causal genes. Indeed, it seems reasonable to postulate that common regulatory variants causing eQTL effects can be detected in healthy individuals. Colocalization methods have been developed to quantify the similarity between DAP and EAP derived from GWAS and eQTL analyses, respectively. Both studies should be conducted in cohorts of the same ethnicity so that they share the same LD structure. However, matching *cis*-QTLs to uncover putative causal genes has been successful for only 31.5% (63/200) of the IBD risk loci analyzed. The large proportion of "orphan" risk loci may be explained by the fact that the matching eQTLs have not yet been identified. Possible reasons include the absence or underrepresentation of disease-relevant cell types in existing datasets, that some eQTLs are only active in specific contexts (such as the disease process), or, according to Mostafavi *et al*. (2023), that the eQTLs driving CCDs are different from those discovered so far, which often have limited sample sizes compared to GWAS.

The aim of this thesis is to investigate the first hypothesis assuming why some eQTLs underlying GWAS-identified IBD peaks were missed: the disease-relevant cell types may have been absent or underrepresented in existing datasets. To discover novel eQTLs driving inherited predisposition to IBD, we established the CEDAR-2 cohort with healthy Europeans, dividing it into two parts. In the first part, we collected peripheral blood from 200 individuals and isolated PBMCs, as well as 26 circulating immune cell types using FACS (positive selection using fluorescent antibodies) and MACS (negative selection using magnetic antibodies). We then performed ultra-low input RNA-Seq on each of the more than 5,000 samples. In the second part, we collected intestinal biopsies at three locations (ileum, colon, and rectum) from 60 individuals, performed scRNA-Seq, and defined 43 intestinal cell populations. All individuals were genotyped for approximately 700,000 SNPs and imputed to 6.3 million variants. We performed *cis*-eQTL analyses to explore the genome-transcriptome association. Then, we integrated these results with 206 risk loci for IBD using published GWAS summary statistics, employing colocalization methods. Additionally, we developed a web browser to visualize our data. The novel causal genes represent preferred drug targets that could be considered for drug repurposing.

METHODS

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

1.1 Sample collection

We collected 40 mL of venous blood (EDTA) from 251 healthy European subjects (147 females and 104 males, average age was 48 years, ranging from 19 to 79) 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 (STable 1).

1.2 SNP genotyping

Genomic DNA was isolated from frozen EDTA-blood using the NucleoMag Blood 200 µL kit (Macherey-Nagel) on a KingFisher 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 ($\hat{\pi} > 0.185$). 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 r2 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 (5,896,787 SNPs and 403,211 indels).

1.3 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 (STable 28). Cells suspended in PBS were then filtered using a 100 µm 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 plasmocytes), 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 STable 28. 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.

1.4 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 has been 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): 2x50 bp (2,112 samples); GIGA Genomics platform (Liège): 2x150 bp (3,180 samples)). We performed RNA-seq for 5,292 samples, corresponding to 27 cell types from 196 individuals.

1.5 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 to detect reads that fail to align to the reference genome due to overlapping SNPs, as these variants are misinterpreted as mismatches (Dobin & Gingeras, 2015; Asiimwe & Dobin, 2024). The same VCF file—containing 5,896,787 imputed SNPs, all set to heterozygous—was provided for all samples to ensure unbiasedness across the entire cohort. 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) (STable 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 mislabelled samples, reads counts were normalized using the variance-stabilizing transformation provided by the DESeq2 R package and a t-SNE analysis was conducted using the 500 most variable genes (SFig. 2). STable 3 reports the characteristics of the 5,030 RNA-seq libraries that passed the quality control procedure.

1.6 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, PBMC and granulocytes were ignored in this analysis as they encompass multiple cell types.

1.7 *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. PCA was chosen over the PEER method because it yielded a higher number of eQTLs. STable 4 reports the number of samples, genes and top expression PCs for each cell type. We removed variants with MAF ≤ 0.05 in the retained samples using bcftools (v1.9) for each cell type. We performed eQTL mapping using QTLtools in a 2 Mb window centered at the transcription start site and with the integrated rank normal transformation of the phenotypes. The p-values were corrected for multiple testing within each window by permutation (10,000 permutations) and within each cell type using the false discovery rate (FDR). eQTLs with a "within cell type FDR" ≤ 0.05 were considered significant.

1.8 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).

1.9 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.*, 2018. Two EAP were compared if at least one of them was significant (FDR ≤ 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 ≤ 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 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".

1.10 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.

1.11 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).

1.12 Testing for an excess sharing of RM between cell types

We followed the methodology of Momozawa *et al*. (2018) to test whether specific cell types shared *cis*-eQTLs more frequently than expected by chance (expected for ontogenically related cell types). In our analysis framework, this would manifest itself by the fact that the corresponding cell types would be co-included in the same RM more often than expected by chance. The number of *cis*-eQTL detected by cell type differs, and this has to be taken into account when measuring enrichment. We assumed that if the sharing of eQTLs was equally likely between any pair of cell types (accounting for differing numbers of eQTL per cell type), the proportion of sharing events between cell type i and j should correspond to the proportion of eQTL detected in cell type j (n_{jT}) out of all eQTL detected in all cell types other than i ($\sum_{k \neq i}^{27} n_{kT}$). Imagine that cell type i is characterized by a total of n_{iS} sharing events, where $n_{iS} = \sum_{k \neq i}^{27} n_{ik}$, and n_{ik} is the observed number of sharing events between cell type i and k. We determined the probability to obtain n_{ij} or more "successes" under the null, by sampling n_{iS} events with a probability of success of $\frac{n_{jT}}{\sum_{k \neq i}^{27} n_{kT}}$ by simulation (n = 5,000). Of note, this process yields two p-values for every pair i, j: one obtained when considering i as reference, the other when considering j as reference. As in Momozawa *et al.*, we performed the analysis 26 times: first considering RM with no more than two cell types (hence eQTL that are shared by two cell types only), then considering RM with no more than three cell types, etc., until considering RM encompassing all cell types.

1.13 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 x cell type interaction effect). To test the statistical significance of the third scenario, we first measured the *cis*-eQTL effect at the top variant for cell type *a*, in cell types *a* and *b*, yielding β_a and β_b. We then generated bootstrap samples from cell type *a*, and computed 1,000 β_s's. The p-value of the variant x cell type interaction was determined as the number of β_s-values that would be equal or lower than β_b if β_a was positive, or equal or higher than β_b if β_a was negative.

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

2.1 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 the rectum (RE) and the 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, 12-167F) 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 (STable 1). 22 participants also provided blood samples for eQTL detection in circulating immune cell populations (see above).

2.2 SNP genotyping

This 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 (5,945,740 SNPs and 406,918 indels).

2.3 Single-cell RNA sequencing

Biopsies from the three locations (IL, TC, RE) were processed in parallel using so-called two-step protocols (Smillie *et al.*, 2019; Kong *et al.*, 2023) 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 µm 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 seconds. 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 relative centrifugal force (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. EC were collected by centrifugation at 500 RCF for 5 min at 4°C, and resuspended in 100 µl of PBS (Lonza Bioscience, 17-516F) with 10% FBS. Cell concentration and viability was 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 x 10^5, 2.6 x 10^5 and 1.7 x 10^5 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_LPDK" 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 of 9.1 x 10^5, 8.6 x 10^5 and 5.5 x 10^5 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-Seq[69], 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 STable 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 re-suspended in 100 µl of PBS with 10% FBS and cell density and viability was 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 5,208 ~ 44,126)(STable 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.

2.4 Preprocessing of scRNA-Seq data

Raw sequencing data were preprocessed 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 4.1.3 (Hao *et al.*, 2021)). 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 ≥ 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: HTODemux (Stoeckius *et al.*, 2018) (positive.quantile 0.999, clusterization function - kmeans) and MULTIseqDemux (McGinnis *et al.*, 2019) 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 (3,000 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 Harmony (Korsunsky *et al.*, 2019). The UMAP plots shown throughout the manuscript (f.i. Fig. 2D and 2E) 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*, *CDH5* 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).

2.5 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 R package (Galili, 2015).

2.6 *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 the 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) algorithm (Stegle *et al.*, 2010). The PEER method was chosen over PCA because it was more practical with the large number of cell types. We restricted eQTL analysis to genes expressed in more than 10% of samples and with mean proportion of reads > 5 x 10^{-7}. Association between gene expression and alternate allele dosage was conducted for each SNP in a 2Mb-window centered on the gene's transcription start site using QTLtools (Delaneau *et al.*, 2017). Nominal *p*-values were corrected for the realization of multiple tests in this *cis*-window by permutation, yielding a window-adjusted *p*-value for one lead SNP for every gene x leaf/node combination. Window-adjusted lead SNP p-values for all genes in a leaf/node were jointly used to compute a *q*-value (Storey and Tibshirani, 2003).

2.7 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. We first build gene-specific and then across-gene modules. For gene-specific modules, we computed θ (Momozawa *et al.*, 2018) between EAP obtained, for that gene, in the different nodes/leaves. θ was computed in the 2Mb *cis*-window centered on the gene's transcription start site (TSS) (the window used for *cis*-eQTL analysis). For across-gene modules, we additionally computed θ between EAP of different genes, provided that their 2Mb *cis*-windows overlapped. θ was then computed for the union between the two 2Mb *cis*-windows. For both gene-specific and across-gene windows, we only considered EAP pairs for which at least one corresponded to a significant eQTL (within leaf/node FDR ≤ 0.05). We only considered SNPs with eQTL nominal *p*-value ≤ 0.05 for at least one of the two EAP. EAP (from the same or different genes) were merged in the same module if $|\theta| \geq 0.6$, and (in the case one of the EAP pairs did not correspond to a significant eQTL) a *p*-value of the eQTL corrected for the multiple SNPs tested in the window (by permutation, see above) ≤ 0.012, as this yielded mergers with FDR ≤ 0.05 (see DAP-EAP part, hereafter). To be part of a module, an EAP had to satisfy these criteria with at least one significant member of the module. Once a module was assembled (all member EAP determined), θ was computed between non-significant members of the module to evaluate the tightness of the module (ideally one hopes for $|\theta| \geq 0.6$ between all members; observed: "Proportion of links" in STable 13 and STable 14). Links between pairs of EAP within a module were given a sign (positive or negative) depending of the value of θ. Constituent EAP were given a positive sign if their θ with the module's representative EAP (the one with the most significant eQTL) was positive, a negative sign otherwise. For modules that comprised more than one EAP, we computed a consensus EAP (2-4 Mb window) by "meta-analysis". *P*-values for all SNPs in the window were converted to z-scores, z-scores (for a given SNP) squared and summed over all members of the module. The resulting sum was converted back to a *p*-value assuming that it had a chi-squared distribution with numbers of degrees of freedom equal to the number of EAP in the module. The EAP of all constituent eQTL were confronted to the consensus EAP in the full window. An EAP was only maintained in the module if its $|\theta|$-value with the consensus was ≥ 0.6. Otherwise, it was ejected from the module. Ejected EAP were confronted to each other and given the possibility to assemble in new "sub-modules".

2.8 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 (Smillie *et al.*, 2019; Franzén *et al.*, 2019; Hao *et al.*, 2021; Burclaff *et al.*, 2022; Ishikawa *et al.*, 2022; Hickey *et al.*, 2023; Kong *et al.*, 2023; Krzak *et al.*, 2023) using the Seurat AddModuleScore function (Tirosh *et al.*, 2016), and the Azimuth human PBMC reference mapping program for immune cells (Hao *et al.*, 2021), 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 STable 11.

2.9 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 STable 11&17. As an example, if a module was active in CD4 and CD8, it was assigned to the T lymphocyte supergroup.

2.10 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 3,345 (gene-specific modules) or 3,081 (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.

2.11 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 (Bates *et al.*, 2015; Kuznetsova *et al.*, 2017) 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*-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.

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

3.1 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, $p_{eQTL,window\ adj} \leq 0.12$ for blood and ≤ 0.012 for gut. These thresholds ensured an FDR $\leq$ 0.05 in the corresponding DAP-EAP confrontations (see hereafter).

3.2 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). 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 an empirical test of independence. Cell types were groups in a limited number of "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 performed a test of independence within the group of 2,170 modules that were assigned to a supergroup in both data sets. We determined the proportion of modules assigned to each supergroup separately for blood (p_i^{Blood}) and gut (p_j^{Gut}). The observed number of modules assigned to supergroup i in blood and j in gut was then compared to the expected number computed as 2,170 $\times\, p_i^{Blood} \times p_j^{Gut}$. The probability to obtain an as large or larger deviation between expected and observed numbers by chance was determined from 1,000 replicates of 2,170 samplings with replacement with a probability of success of $p_i^{Blood} \times p_j^{Gut}$. Secondly, within the group of 4,579 modules that were active in gut alone, we determined the proportion that were assigned to each one of the gut supergroups (p_j^{Gut}) as well as the proportion that were not active in blood (p_{ND}^{Blood}). The observed number of modules assigned to supergroup j in gut yet undetected in blood was then compared to the expected number computed as

4,579 $\times p_{ND}^{Blood} \times p_{j}^{Gut}$. The probability to obtain an as large or larger deviation between expected and observed numbers by chance was determined from 1,000 replicates of 4,579 samplings with replacement with a probability of success of $p_{ND}^{Blood} \times p_{j}^{Gut}$. Finally, within the group of 22,336 modules that were active in blood alone, we determined the proportion that were assigned to each one of the gut supergroups (p_{i}^{Blood}) as well as the proportion that were not detected in gut (p_{ND}^{Gut}). The observed number of modules assigned to supergroup i in blood yet undetected in gut was then compared to the expected number computed as 22,336 $\times p_{i}^{Blood} \times p_{ND}^{Gut}$. The probability to obtain an as large or larger deviation between expected and observed numbers by chance was determined from 1,000 replicates of 22,336 samplings with replacement with a probability of success of $p_{i}^{Blood} \times p_{ND}^{Gut}$.

4. Identifying new *cis*-eQTL driving inherited predisposition to IBD

4.1 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.*, 2017. 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 |θ| 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 exer*cis*e 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 $p_{eQTL,window\ adj} = z$, we would determine how many matches satisfying $|\theta| \geq x$, $p_{\theta} \leq y$, and $p_{eQTL,window\ adj} \leq z$, 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 $\frac{N_P}{N_R}$. We defined two FDR thresholds of significance: Tier 1 (FDR ≤ 0.05) and Tier 2 (0.05 < FDR ≤ 0.10).

4.2 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 35K individuals, were downloaded (Võsa *et al.*, 2021). We lifted their data over from GRCh37 to GRCh38. When arithmetic underflow was observed for the *p*-values, the log10 *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 or to compute a *p*-value for theta. Colocalized EAPs correspond to EAPs that have a |θ| ≥ 0.6 with a DAP (hence more permissive than the analysis of the CEDAR2 dataset).

5. Identifying drug repurposing candidates

5.1 Effects of disease (CD) on the expression of DAP-matching e-genes in circulating immune cell populations

We collected blood from 55 active CD patients (25 females and 30 males), 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 (see above). Differential expression analysis between controls and patients was performed by cell type using the DESeq2 R package (Love *et al.*, 2014), using the apeglm method for effect size shrinkage (Zhu *et al.*, 2019). Genes with a fold change > 1.5 and an FDR < 0.05 were considered differentially expressed (STable 26). STable 22 lists, for 556 genes with DAP-matching EAP, the blood cell types for which CD has a significant effect on gene expression (log2 fold change > 0.58 and FDR < 0.05). For each significant effect, STable 22 provides cell type, effect (log2 fold change), and FDR. If, for a given gene, the majority sign of the effects (across cell types) 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 labeled as consistent ("TRUE"); otherwise, it is labeled as inconsistent ("FALSE").

5.2 Effects of disease (CD) on the expression of DAP-matching e-genes in intestinal cells

Data were obtained from Supplemental Table 3 in Kong *et al.* (2023). 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 MAST (Finak *et al.*, 2015) in either discrete mode (expression rate) or continuous mode (expression level). STable 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, STable 22 provides effect (log2 fold change), FDR, cell type and location. If, for a given gene, the majority sign of the effects (across cell types/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 labeled as consistent ("TRUE"); otherwise, it is labeled as inconsistent ("FALSE").

5.3 Effects of disease (IBD) on plasma protein levels for DAP-matching e-genes

Data were obtained from STable 18 in Eldjarn *et al.* (2023). Eldjarn *et al.* analyzed the association with IBD, of concentrations of 2,931 circulating proteins (Olink Explore 3072 assay) in 49K individuals from the UK Biobank (including 900 IBD cases), and of 4,907 circulating proteins (SomaScan v4 assay) in 36K Icelanders (including 618 IBD cases), using logistic regression. STable 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, STable 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 labeled as consistent ("TRUE"); otherwise, it is labeled as inconsistent ("FALSE").

5.4 Identifying DAP-matching e-genes that are the targets of known IBD drugs

A list of drugs that are or have been used to treat IBD was obtained from Vieujean *et al.* (2025), and is reported in STable 27. The corresponding sources of information were either ClinicalTrials.gov, the FDA, DailyMed, European Medicines Agency, or the literature, as listed in STable 27.

5.5 Identifying DAP-matching e-genes that are the targets of drugs (phase 1 to approved) that are used to treat diseases other than IBD

Drugs targeting DAP-matching e-genes were retrieved from the OpenTarget platform (https://platform.opentargets.org) and the literature. Their *modus operandi* (activator or inhibitor) was obtained from DrugBank (https://go.drugbank.com/) and/or ChEMBL (https://www.ebi.ac.uk/chembl/). Their status with regards to indication, clinical trials and/or approval was obtained from DailyMed (https://dailymed.nlm.nih.gov/dailymed/) and ClinicalTrials.gov (https://clinicaltrials.gov/).

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 (STable 1). We genotyped all individuals with Illumina's OmniExpress array interrogating ~700K 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; STable 2).

We produced next generation sequencing (NGS) libraries for each of the 27 cellular fractions of each individual (5,292 mRNA SMART-Seq HT libraries), and generated an average of ~12.6M, 2 x 150 bp (3,180 libraries) or ~13M, 2 x 50 bp (2,112 libraries) paired-end reads per library on Illumina instruments (STable 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) (STable 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 QTLtools (Delaneau *et al.*, 2017), regressing expression level on alternate allele dosage and including an average of 24.4 expression principal components (PC) in the model (STable 4; M&M). We identified 60,113 eQTL (within cell type FDR $\leq$ 0.05), influencing 11,874 eQTL genes (e-genes) (STable 5). The number of eQTL detected per cell type (average: 2,226; range: 608 – 3,448) 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 plasmocytes was lower than expected. Controlling for the variable abundance of immunoglobulin transcripts did not increase the number of detected plasmocyte 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).

eQTLs affecting the same gene in multiple cell types were merged if sharing a similar association pattern (EAP) as evaluated using theta ($|\theta| \geq 0.6$) (Momozawa *et al.*, 2018), yielding 23,831 gene-specific modules (STable 6). Modules were augmented with 5,840 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). Sharing of modules by cell types was largely determined by their ontogenic proximity ($p = 9 \times 10^{-68}$; SFig. 3E). 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 devised an interaction test to evaluate the importance of scenario 2 (see M&M) in the remaining 73.8% of cases. We estimated the proportion of alternative hypothesis (π_1) following Storey and Tibshirani, 2003 at 97.3%, hence indicating that the eQTL was indeed not active in cell type B in 71.8% of cases despite the fact that the gene was expressed in cell type B (Fig. 1B-C; SFig. 3F). For modules active in more than one cell type, the direction of the effects was the same across cell types in 96.1% of cases, yet there were 250 cases for which the sign of the effect switched between cell types (Fig. 1D-E; STable 6). The proportion of shared eQTL with opposite sign increased with ontogenic distance between cell types ($p_{logit} = 4 \times 10^{-33}$; STable 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=2,776) 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=9,201) did not show enrichment in any GO terms (STable 8).

We merged gene-specific modules with matching EAP "across genes", yielding a total of 22,067 regulatory modules. Modules were augmented with 7,402 tier-2 eQTL matching at least one significant *cis*-eQTL in the module, recruiting 759 extra genes (M&M; STable 9). We observed 2,470 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. 3G). 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. 3H). 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).

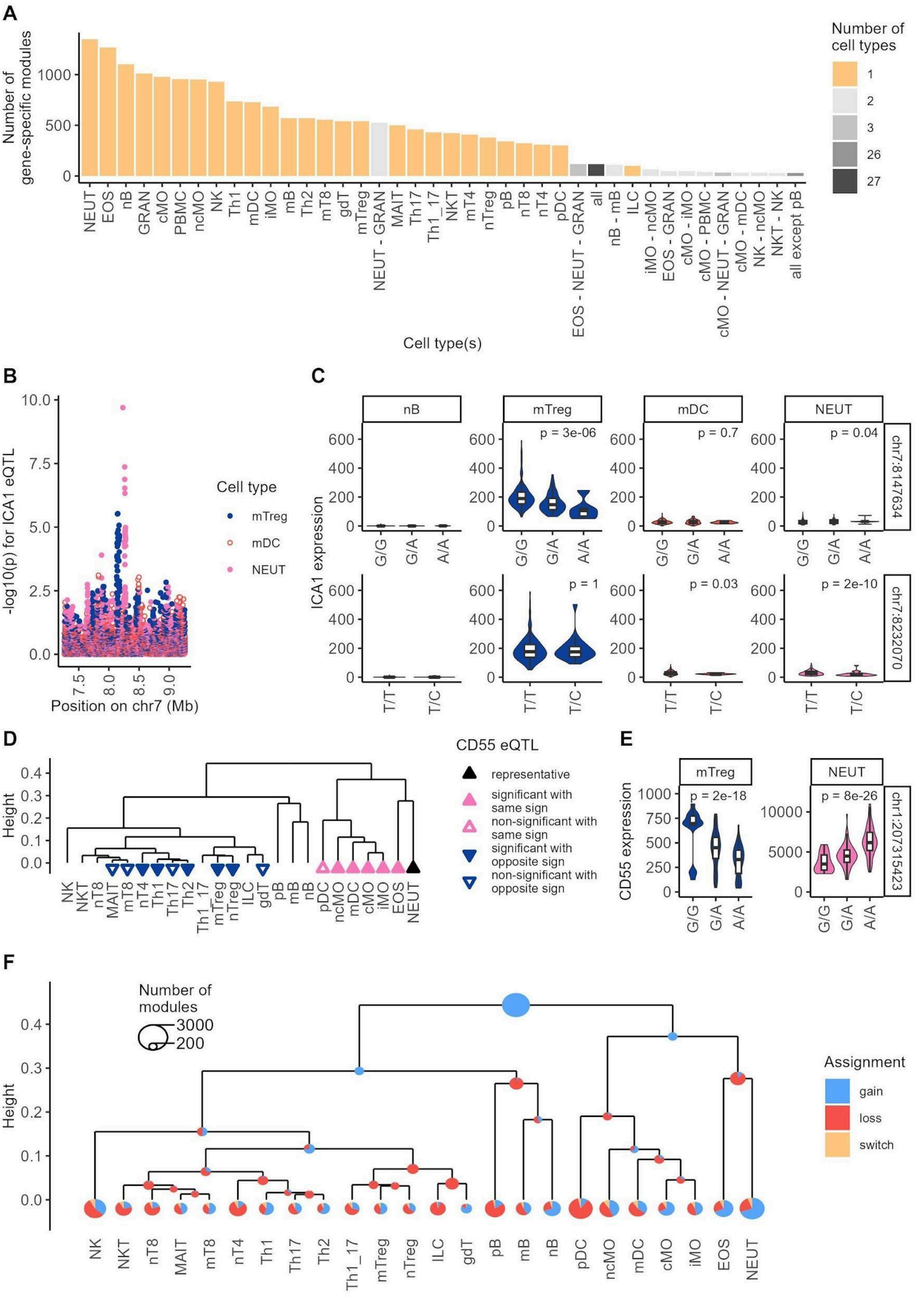
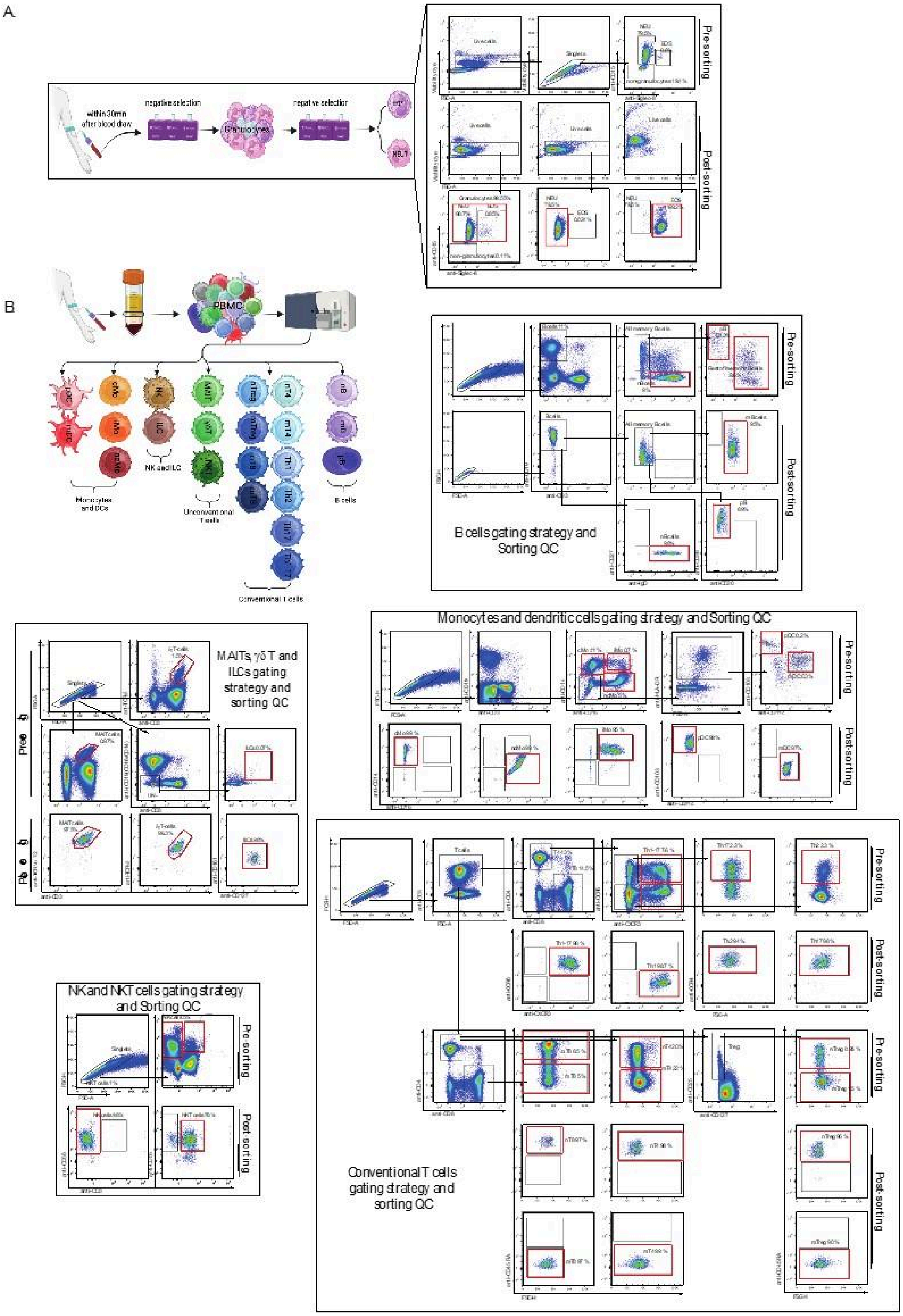
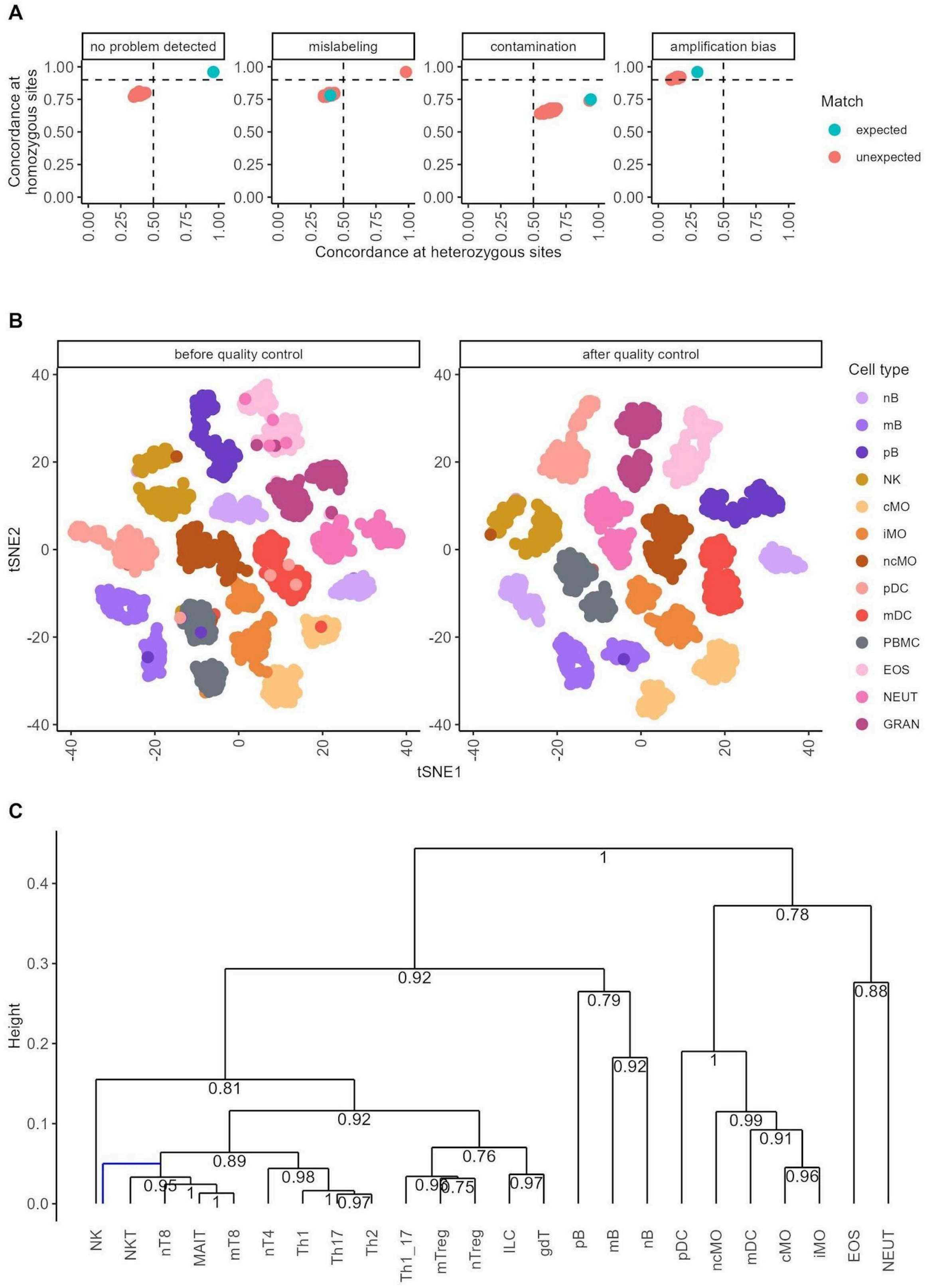


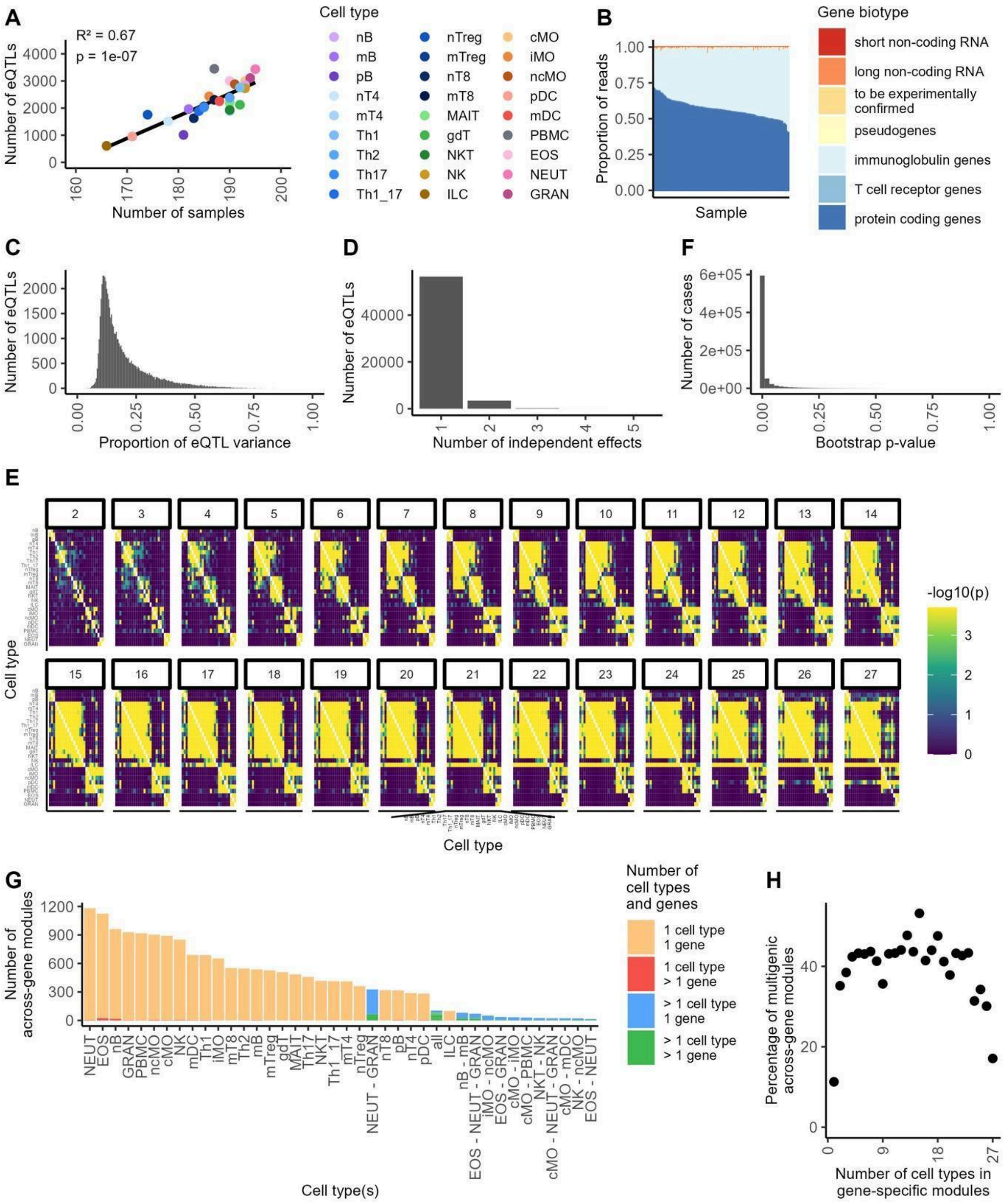
Figure 1: (A) Gene-specific modules grouped by cell type combinations in which they are active, and ordered according to the frequency of occurrence (40 most frequent combinations out of 3,572 observed). The most common modules are dominated cell type-specific ones, i.e., only active in one cell type (27 yellow bars). Also, amongst the top 40 combinations, are those corresponding to modules that are active in all (i.e., ubiquitous eQTL; black bar). The remaining combinations that are shown correspond to modules that are active in 2 or 3 closely related cell types (hence still very cell type-specific). Cell type abbreviations are as in STable 2. **(B-C)** Example of a gene (*ICA1*) that is controlled by two distinct 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 of these is active in myeloid dendritic cells (mDC, orange EAP and violin plot) despite the fact that the gene is expressed at relatively high levels in this cell type (scenario 2). The gene is expressed at very low levels in naïve B cells (nB) in which consequently neither module is active either (scenario 1). Individuals are sorted by genotype at the lead SNPs for modules #12,605 (upper row) and #16,344 (lower row). **(D-E)** Example of a gene-specific eQTL module (*CD55*, module #678) that is active in 16 of the 24 shown 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). However, the sign of the eQTL (effect of the alternate allele) 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 expressions) in memory regulatory T cells (left, blue) and neutrophils (right, pink) for individuals sorted by genotype at the lead variant for the consensus EAP. **(F)** Gains of gene-specific modules were assigned to the "most recent common ancestor" (MRCA) node of all nodes/leaves in which the module is active (blue segments). Losses of gene-specific modules were assigned to the MRCA node of all descendent (of a node to which a module was assigned) 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.



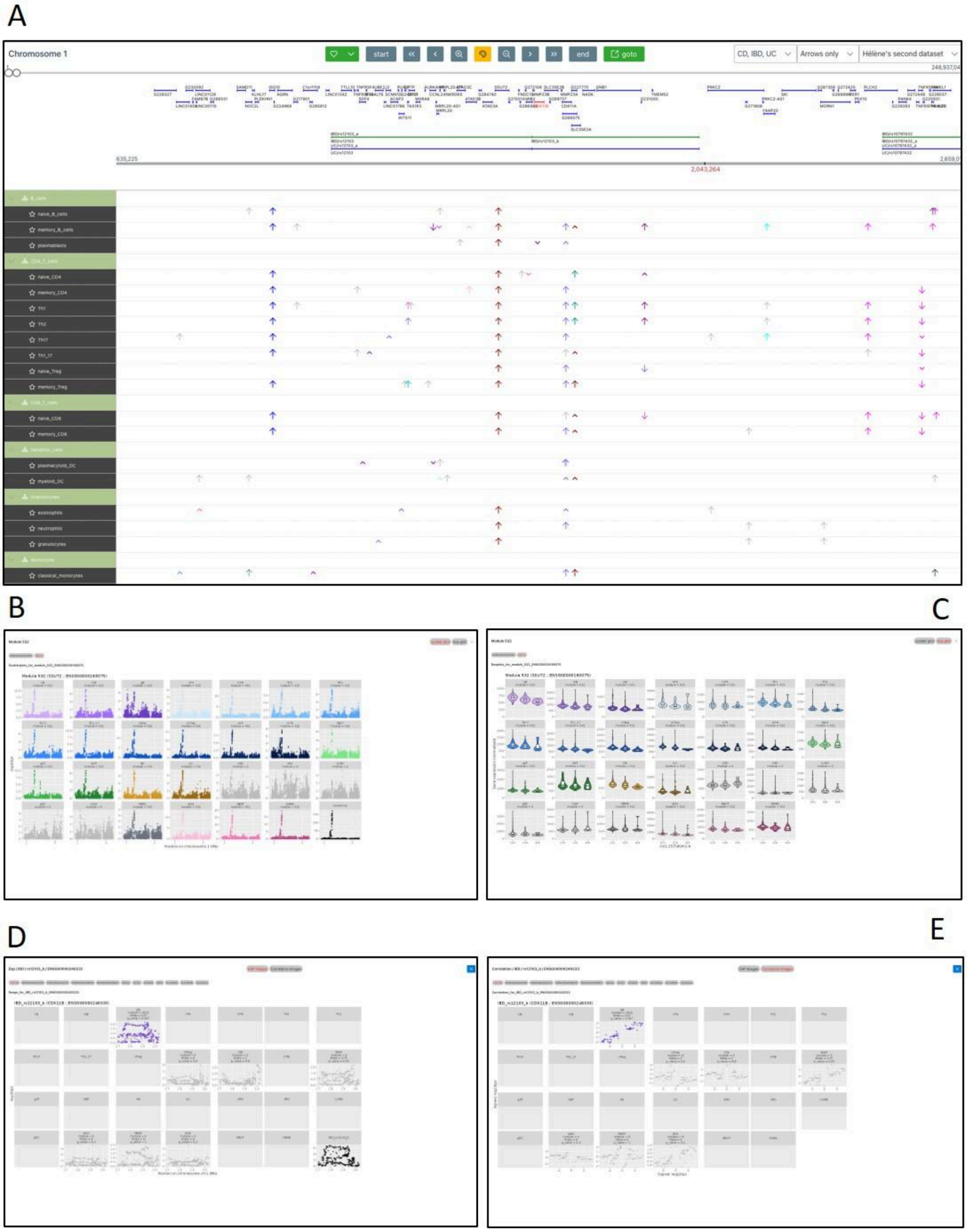
Supplemental Figure 1: Cell sorting and QC. (A) Granulocytes magnetic cell sorting from EDTA-anticoagulated peripheral whole blood. Blood was processed for granulocytes negative cell sorting within 30 minutes after collection. ¼ and ½ of the sorted granulocytes were further processed separately and respectively for Neutrophils and Eosinophils selection. The remaining ¼ was immediately frozen until use. For QC of the cell sorting, an aliquot before and after each sorted fraction were stained using anti-CD15 and anti-Siglec-8 antibodies, together with viability dye. Left panel shows the percentage of each sorted population before and after sorting. **(B)** Flow cytometry cell sorting and QC of PBMC derived cell populations. Representative FACS plots illustrating gating strategies and QC after cell sorting for B cells (upper right panel), MAITs, gd T and ILC cells (middle left panel), monocytes and dendritic cells (middle right panel), NK and NKT cells (lower left panel), and conventional CD8$^+$ and CD4$^+$ T cells (lower right panel). For a detailed description of subset names, see STable 2. For information on the flow-cytometry antibodies used in this study, see STable 29. Red quadrants indicate the target sorted cell types.



Supplemental Figure 2: Quality control of the bulk RNA-Seq data on 27 circulating immune cell populations. (A) Examples of issues detected by confronting genome and transcriptome genotypes. From left to right: (i) Example of an RNA sample that matches its presumed donor (green dot); the red dots correspond to the confrontation of that RNA sample with the DNA of all other individuals. (ii) Example of an RNA sample that matches the DNA of an individual that is not the presumed donor. (iii) Example of an RNA sample that is likely contaminated (mixture between two samples), as it is "too heterozygous" and hence shows a drop in concordance at homozygous genotypes for both the presumed donor (green dot) and the contaminating individual (co-positioned red dot). (iv) Example of a likely PCR issue; the RNA appears "too homozygous" and hence shows a drop in concordance at heterozygous genotypes. **(B)** Example of issues detected when confronting presumed cell type vs inferred transcriptome-based cell type: tSNA map of 2,444 samples from 13 cell types labeled by presumed cell type of origin before (left) and after (right) quality control. Based on the expression level of 500 most variable genes. **(C)** Hierarchical tree for 24 of the 27 cell types (PBMC, granulocytes and memory CD4$^+$ T cells were not included because being mixtures of cell types) obtained from 1-|Spearman's correlation| between the mean (across all samples) expression levels of the 4,000 most variable (ANOVA F value of cell type effect) coding genes using "average" method. The numbers on the nodes correspond to the proportion of trees recapitulating the same split when varying the number of variable genes considered from 250 to 8,000 (32 analyses) and using the same method. The tree shown is the best supported tree in the method. Nearly identical (most supported) trees were obtained when applying the "ward.d" and "ward.d2" methods. The blue line marks the only difference that was observed with the "ward.d2" method.



Supplemental Figure 3: eQTL mapping and construction of regulatory modules in 27 circulating blood cell populations. (A) Number of detected eQTL as a function of the number of usable samples for the corresponding (color labeled) cell type. **(B)** Demonstration of the inter-individual variation in the abundance of immunoglobulin reads (light blue) in plasmocytes of 181 individuals. **(C)** Distribution of the proportion of variation in gene expression level accounted for by the (primary) detected *cis*-eQTL. **(D)** Results of the conditional analysis for the detection of independent secondary *cis*-eQTL effects. **(E)** Log(1/p-value) of the excess number of shared modules over expectations (given number of detected eQTLs in the respective cell types) for all possible pairs of cell types. Computations are performed by including modules with two active cell types (upper left facet), three or less active cell types, four or less active cell types, …, 27 or less active cell types (lower right facet). **(F)** Frequency distribution of the p-values obtained using an empirical cell type x eQTL effect "interaction" test. We determined the probability to obtain a *cis*-eQTL effect that is as low or lower than the one observed in cell type B, by sampling expression levels (bootstrapping) by variant genotype from cell type A. The large excess of low p-values is striking (uniform distribution expected under the null), yielding an estimate of π_1 [Storey & Tibshirani, 2003] of 99.9%. **(G)** Across-gene regulatory modules grouped by the combination of cell types in which they are active and ordered by the frequency of occurrence of the corresponding pattern. Modules encompassing one gene and active in one cell type only are labeled in yellow (n = 15,926). Modules encompassing more than one gene yet active in only one cell type are labeled in red (n = 84). Modules encompassing one gene yet active in more than one cell type are labeled in blue (n = 3,433). Modules encompassing more than one gene and active in more than one cell type are labeled in green (n = 1,451). Only the 40 more common activity patterns (out of 2,959 in total) are shown. These include the 27 patterns corresponding to modules active in only one cell type. **(H)** Estimation of the proportion of gene-specific modules that merge with another gene-specific module to create a multigenic module when considering gene-specific modules active in 1, 2, 3, … 27 cell types (x-axis). Interestingly, gene-specific modules active in one cell type rarely (11.3%) fuse with others to form multigenic modules. Likewise, gene-specific modules that are active in all 27 cell types rarely (17.1%) become multigenic modules. A contrario, gene-specific modules that are active in 2 to 26 cell type more often (average 41.4%; range: 30.1% - 53.2%) fuse with others to form multigenic modules. To make sure that the probability of merging between gene-specific modules was not affected by the number of EAPs in the modules, we exclusively used one EAP per gene-specific module, namely the consensus EAP for modules with more than one EAP.



Supplemental Figure 4: Screenshots from the CEDAR website (https://tools.giga.uliege.be/cedar) for the 27 circulating immune cell populations. (A) Top panel: genome browser showing the position of genes and risk loci for inflammatory bowel disease (IBD). Bottom panel: activity of the eQTL and *cis*-acting regulatory modules (RM). Significant eQTL are marked by large arrows (↑). Arrows for eQTL that are part of the same module have the same color. eQTL in a module with effect sign opposite to the most significant eQTL in the module ("representative") are marked by a downward pointing arrow (↓). Non-significant eQTL that are part of a module by virtue of matching

EAP ($|\theta| \geq 0.6$) are marked by small arrows (∧ or ∨). eQTL that are alone in their own module are in grey. Modules can be highlighted by clicking on them. **(B)** By clicking on a module, one has access to the EAP graphs showing the activity of the corresponding eQTL in all 27 cell types by gene (one page per gene). EAPs that are part of the selected module are colored, others are shown in gray. If the gene was not expressed in a given cell type (and the eQTL could not be tested) the EAP graph is "empty". The last facet (lower right) corresponds to the "consensus" EAP of the module. **(C)** By clicking on a module, one also has access to violin plots at the lead SNP of the consensus EAP showing the distribution of DESeq2 normalized gene expression values for individuals sorted by SNP genotype. **(D)** By clicking on the bars corresponding to the disease risk loci, one has access to the EAP in all 27 cell types of all genes (one page per gene) that have a matching DAP-EAP ($|\theta| \geq 0.6$, $p \leq 0.05$) with the corresponding DAP (lower right facet). **(E)** By clicking on the bars corresponding to the disease risk loci, one also has access to the theta-plots [Momozawa *et al.* 2018] in all 27 cell types of all genes (one page per gene) that have a matching DAP-EAP ($|\theta| \geq 0.6$, $p \leq 0.05$).

Single-cell RNA Seq analysis of intestinal biopsies reveals 35,010 eQTL affecting 3,007 genes and clustering in 3,337 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 (5,154 cells per individual on average). The number of reads per cell averaged 48,628, the number of unique molecular identifiers (UMI) per cell 7,422, and the number of genes detected per cell 2,035 (STable 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 Harmony (Korsunsky *et al.*, 2019), 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 (Smillie *et al.*, 2019; Franzén *et al.*, 2019; Hao *et al.*, 2021; Burclaff *et al.*, 2022; Ishikawa *et al.*, 2022; Hickey *et al.*, 2023; Kong *et al.*, 2023; Krzak *et al.*, 2023) (Fig. 2; SFig. 5&6; STable 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 (average number of genes detected of 13,036 per leaf/node, ranging from 8,910 to 15,382), and leaf/node-specific PEER factors (Stegle *et al.*, 2010) 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 3,007 e-genes (STable 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 ~1,100 presumably limited by sample size (57 individuals) (Fig. 3A). A second independent effect was detected for 259 *cis*-eQTL, and a third for two (STable 12). As for circulating immune populations, we merged *cis*-eQTL in 3,345 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) (STable 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 distinct effect depending on cell type (Fig. 3F; STable 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 2,415 modules that remained monogenic. The number of genes in multigenic modules averaged 2.8, ranging from 2 to 16 (STable 14). The proportion of multigenic modules encompassing eQTL effects with opposite signs 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).

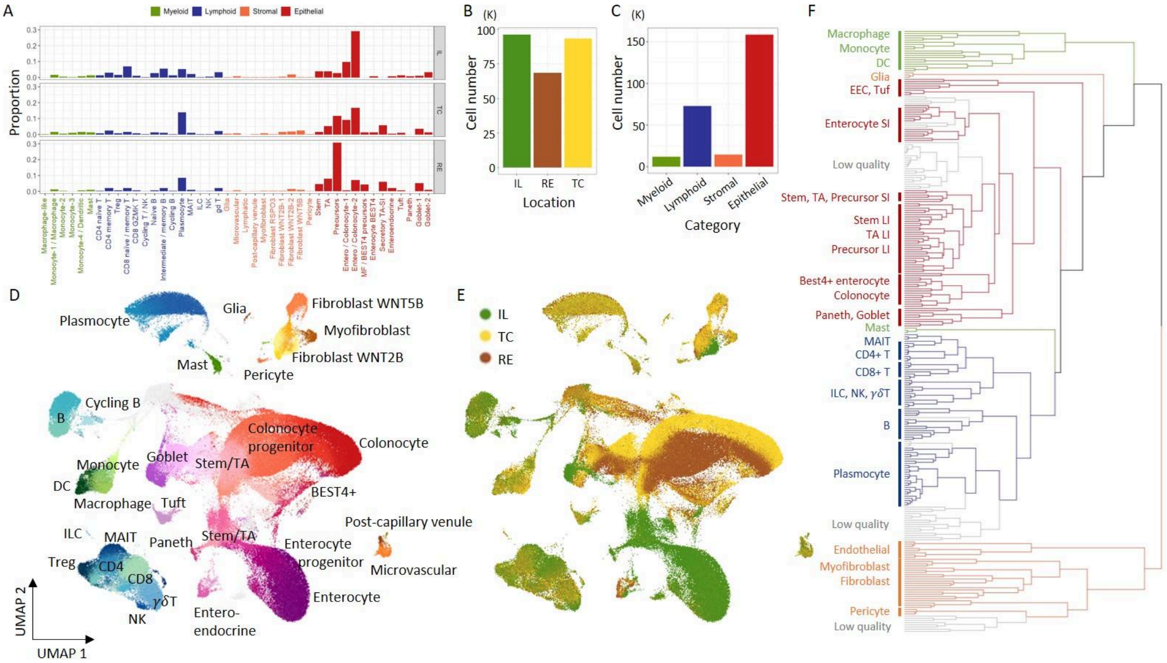


Figure 2: (A) Proportion of cells from ileal (IL), colonic (TC), and rectal (RE) biopsies assigned to 6 myeloid cell types (green), 14 lymphoid cell types (blue), 10 stromal cell types (orange) and 13 epithelial cell types (red). **(B)** Number of recovered quality-controlled (QC-ed) cells by anatomical 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 and C. The main cell types within categories are labeled.

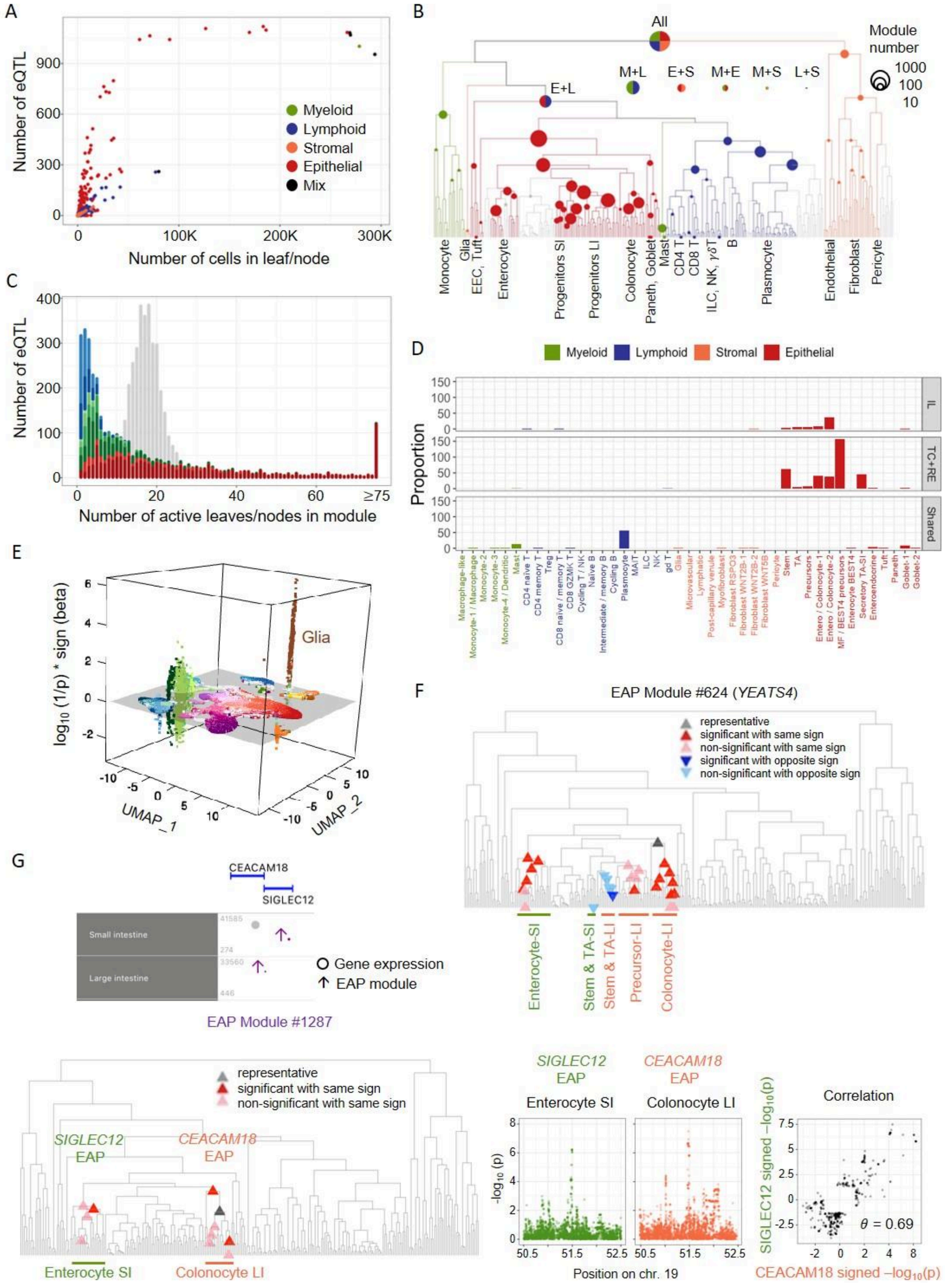
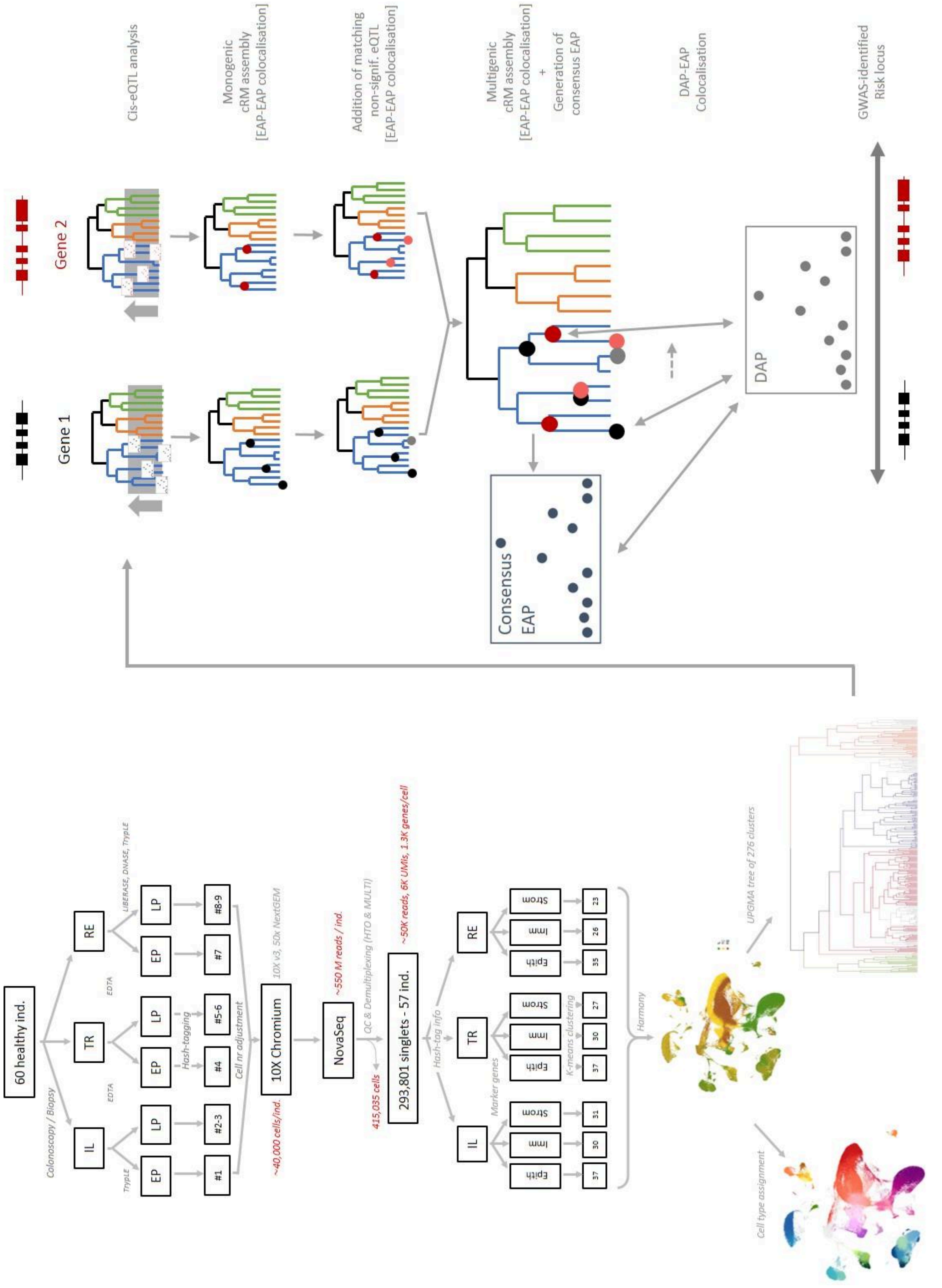


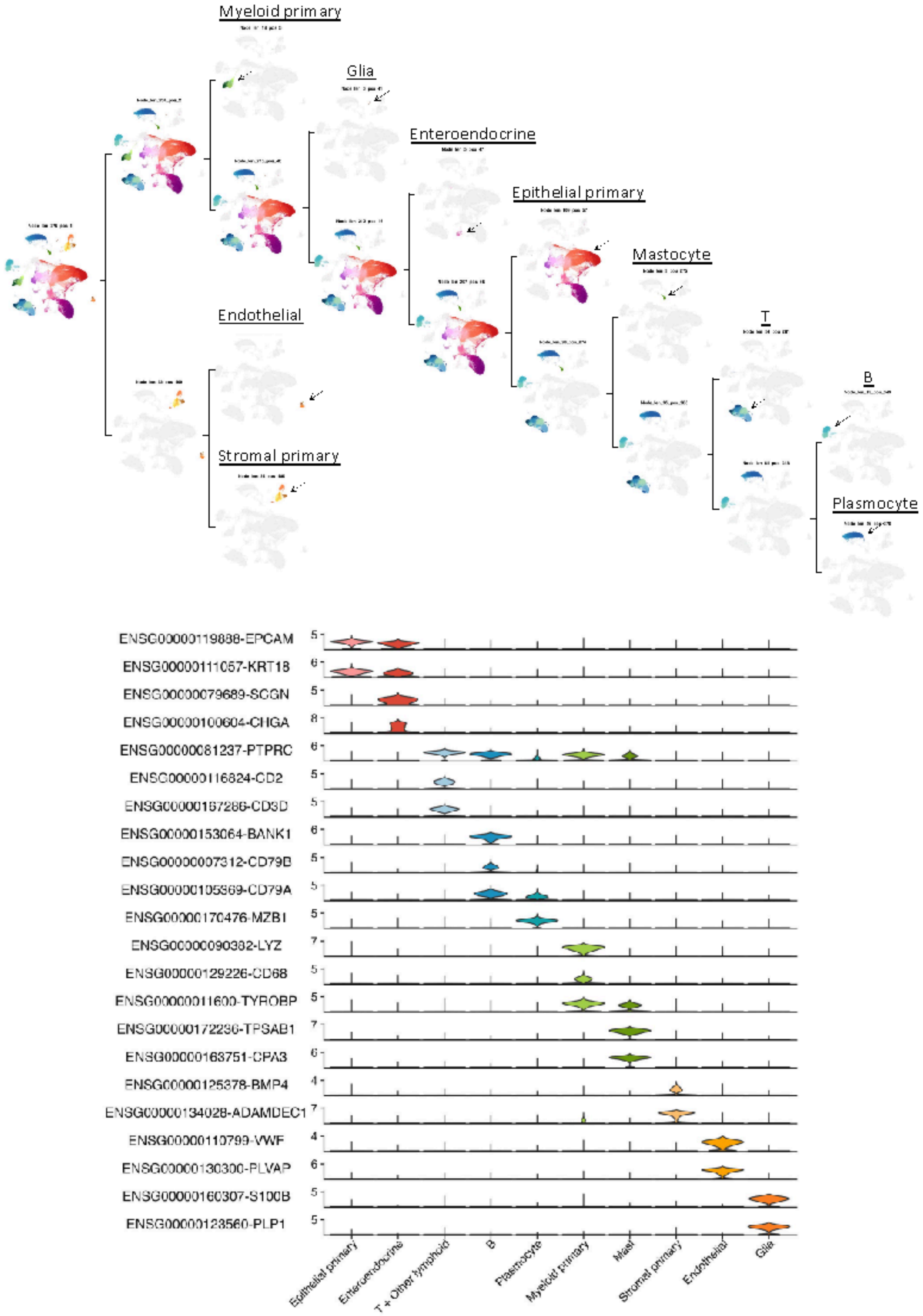
Figure 3: (A) Number of eQTL detected (within leaf/node FDR ≤ 0.05) as a function of the number of cells in the corresponding leaf/node. Leaves/nodes are colored by cell category (myeloid: green, lymphoid: blue, stromal: orange, epithelial: red, mix (=multiple categories): black). **(B)** Assignment of 3,345 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 corresponding leaf/node. Modules initially assigned near the tree's root were assigned to pairs of cell categories when possible, corresponding to the bisected circles ranked by size. **(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 as for the real data (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 effect (*GFRA2* in glia (brown)). The x- and y-axes correspond to the 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 the absorptive intestinal epithelium in the 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).

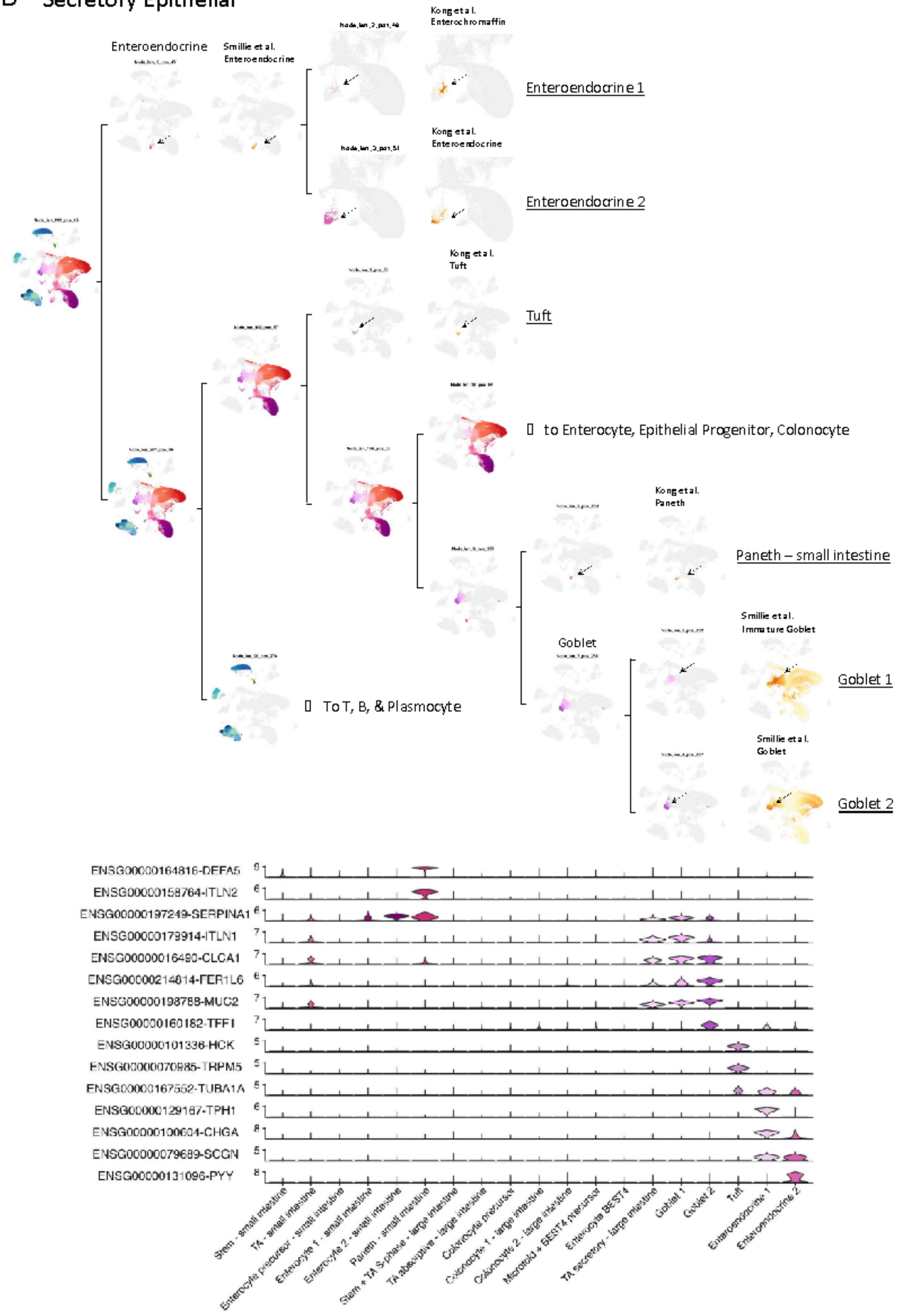


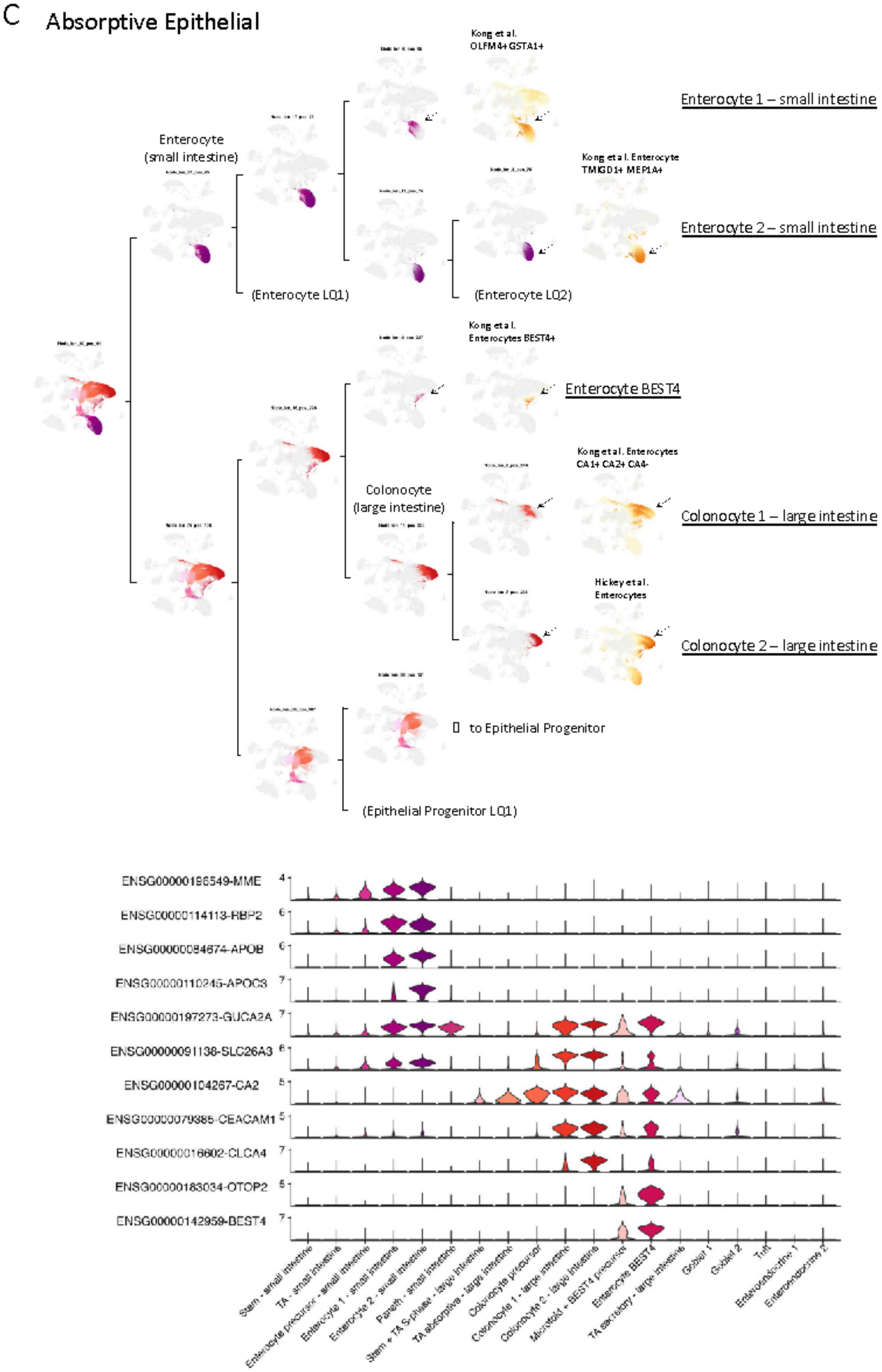
Supplemental Figure 5: Graphical abstract of workflow of *cis*-eQTL analyses in scRNA-Seq data from intestinal biopsies.

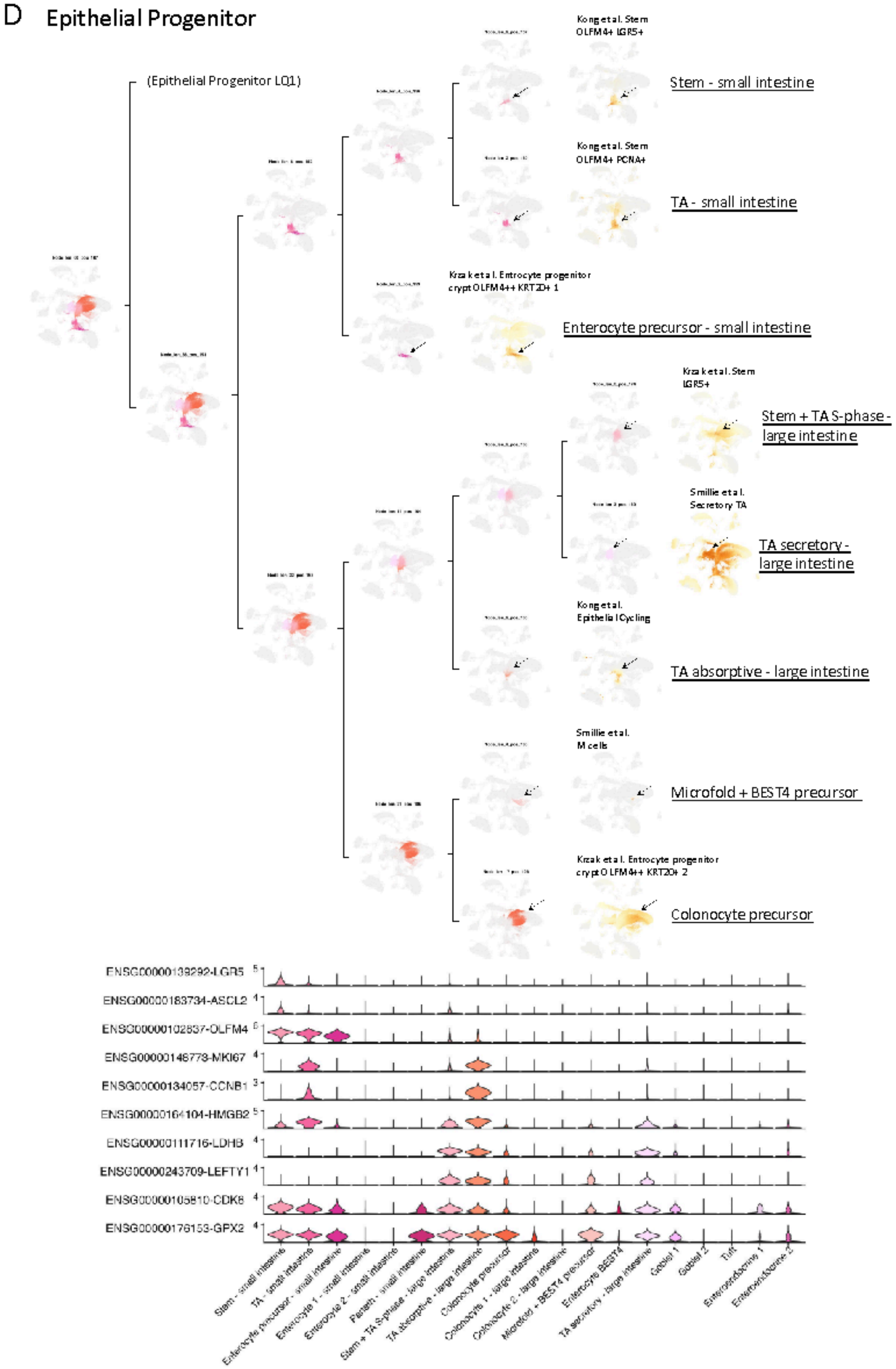
A 10 grand clusters



B Secretory Epithelial

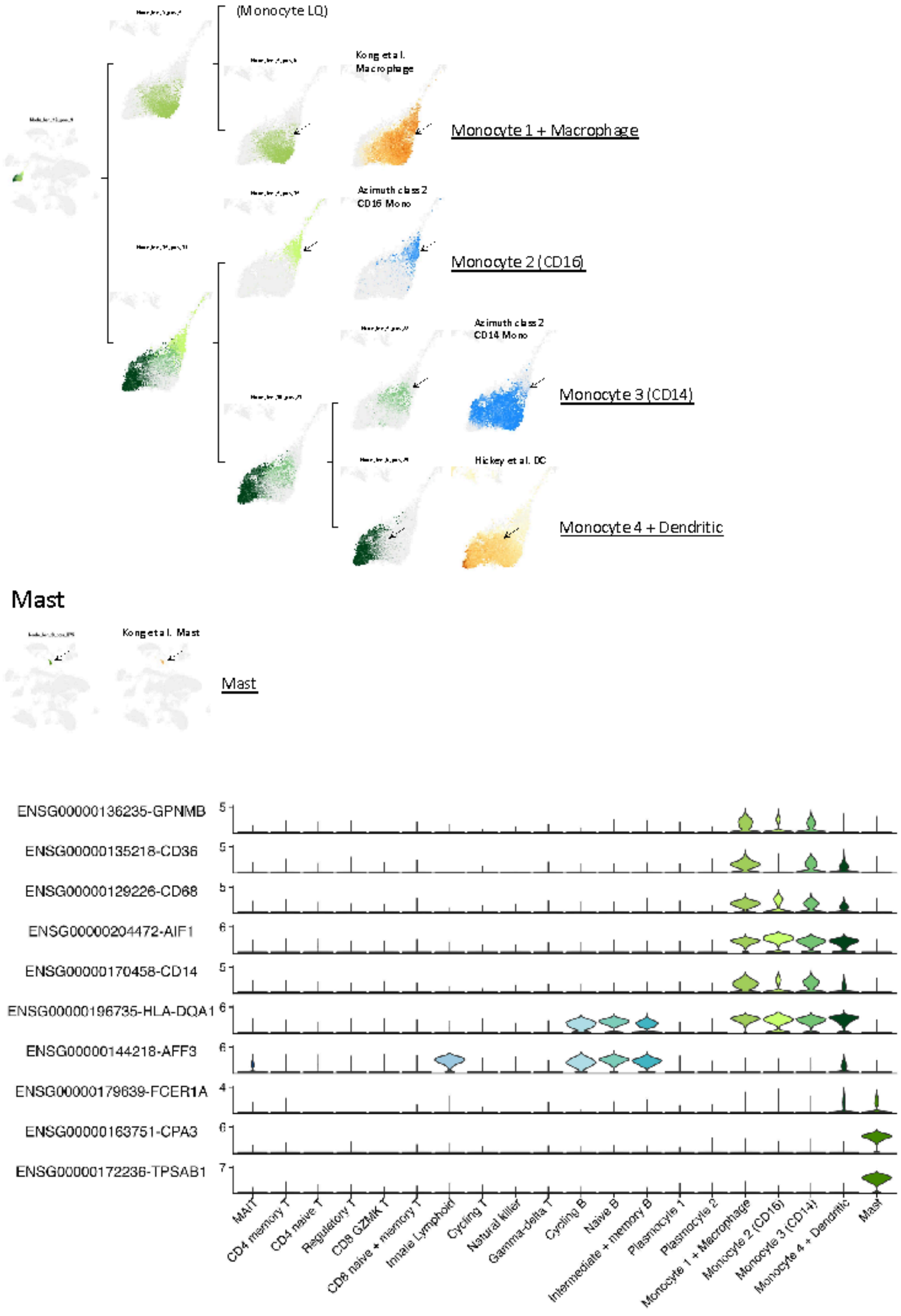


C
Absorptive Epithelial
Kong et al.
OLFM4+ GSTA1+
Enterocyte 1 – small intestine
Enterocyte
(small intestine)
Kong et al. Enterocyte
TMIGD1+ MEP1A+
Enterocyte 2 – small intestine
(Enterocyte LQ1)
(Enterocyte LQ2)
Kong et al.
Enterocytes BEST4+
Enterocyte BEST4
Kong et al. Enterocytes
CA1+ CA2+ CA4-
Colonocyte
(large intestine)
Colonocyte 1 – large intestine
Hickey et al.
Enterocytes
Colonocyte 2 – large intestine
□ to Epithelial Progenitor
(Epithelial Progenitor LQ1)
ENSG00000196549-MME
ENSG00000114113-RBP2
ENSG00000084674-APOB
ENSG00000110245-APOC3
ENSG00000197273-GUCA2A
ENSG00000091138-SLC26A3
ENSG00000104267-CA2
ENSG00000079385-CEACAM1
ENSG00000016602-CLCA4
ENSG00000183034-OTOP2
ENSG00000142959-BEST4
4
6
6
7
7
6
5
5
7
5
7
Stem - small intestine
TA - small intestine
Enterocyte precursor - small intestine
Enterocyte 1 - small intestine
Enterocyte 2 - small intestine
Paneth - small intestine
Stem + TA S-phase - large intestine
TA absorptive - large intestine
Colonocyte precursor
Colonocyte 1 - large intestine
Colonocyte 2 - large intestine
Microfold + BEST4 precursor
Enterocyte BEST4
TA secretory - large intestine
Goblet 1
Goblet 2
Tuft
Enteroendocrine 1
Enteroendocrine 2

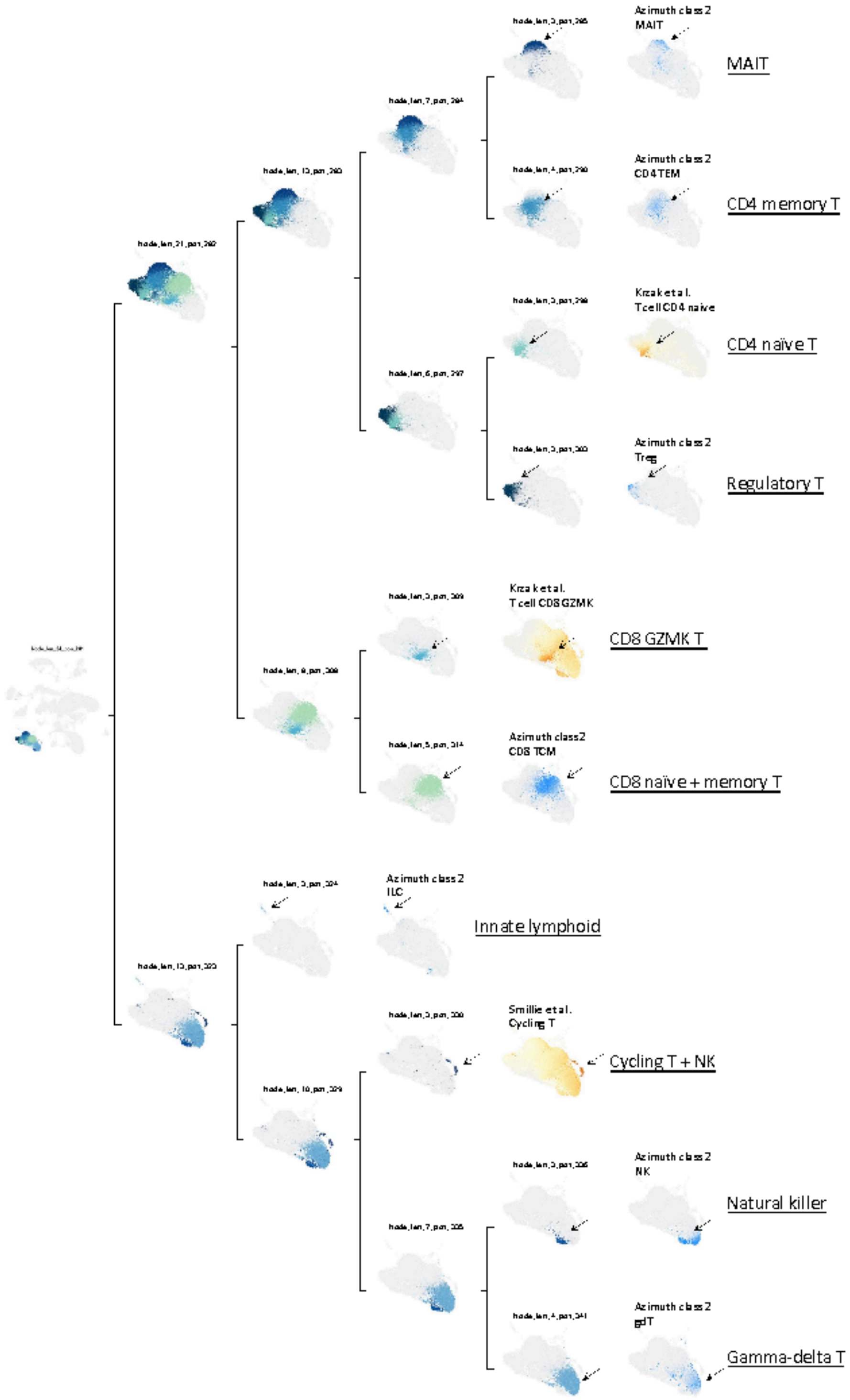


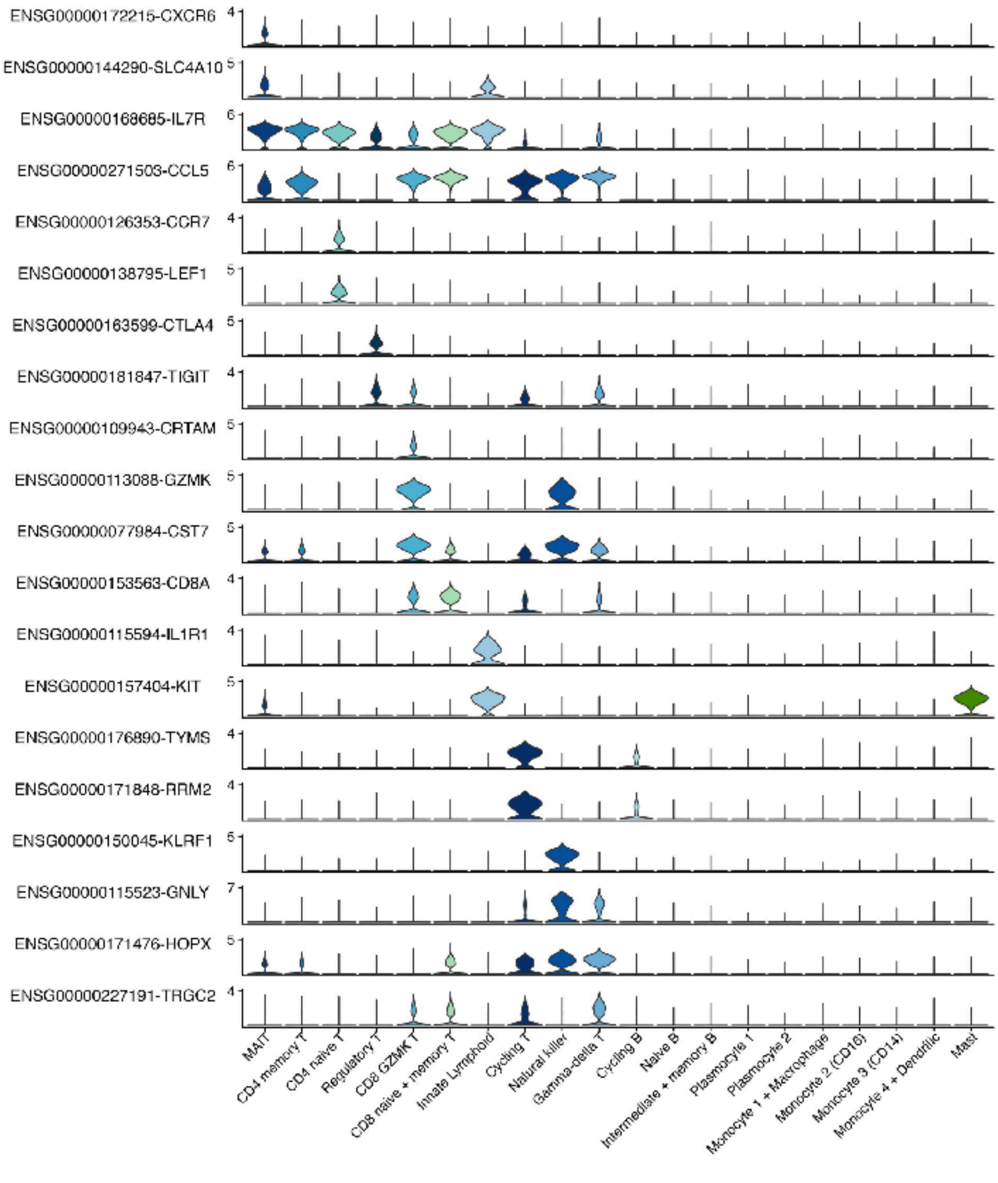
D
Epithelial Progenitor
(Epithelial Progenitor LQ1)
Kong et al. Stem
OLFM4+ LGR5+
Stem - small intestine
Kong et al. Stem
OLFM4+ PCNA+
TA - small intestine
Kraiczy et al. Entrocyte progenitor
crypt OLFM4++ KRT20+ 1
Enterocyte precursor - small intestine
Kraiczy et al. Stem
LGR5+
Stem + TA S-phase - large intestine
Smillie et al.
Secretory TA
TA secretory - large intestine
Kong et al.
Epithelial Cycling
TA absorptive - large intestine
Smillie et al.
M cells
Microfold + BEST4 precursor
Kraiczy et al. Entrocyte progenitor
crypt OLFM4++ KRT20+ 2
Colonocyte precursor
ENSG00000139292-LGR5
5
ENSG00000183734-ASCL2
4
ENSG00000102837-OLFM4
6
ENSG00000148773-MKI67
4
ENSG00000134057-CCNB1
3
ENSG00000164104-HMGB2
5
ENSG00000111716-LDHB
4
ENSG00000243709-LEFTY1
4
ENSG00000105810-CDK6
4
ENSG00000176153-GPX2
4
Stem - small intestine
TA - small intestine
Enterocyte precursor - small intestine
Enterocyte 1 - small intestine
Enterocyte 2 - small intestine
Paneth - small intestine
Stem + TA S-phase - large intestine
TA absorptive - large intestine
Colonocyte precursor
Colonocyte 1 - large intestine
Colonocyte 2 - large intestine
Microfold + BEST4 precursor
Enterocyte BEST4
TA secretory - large intestine
Goblet 1
Goblet 2
Tuft
Enteroendocrine 1
Enteroendocrine 2

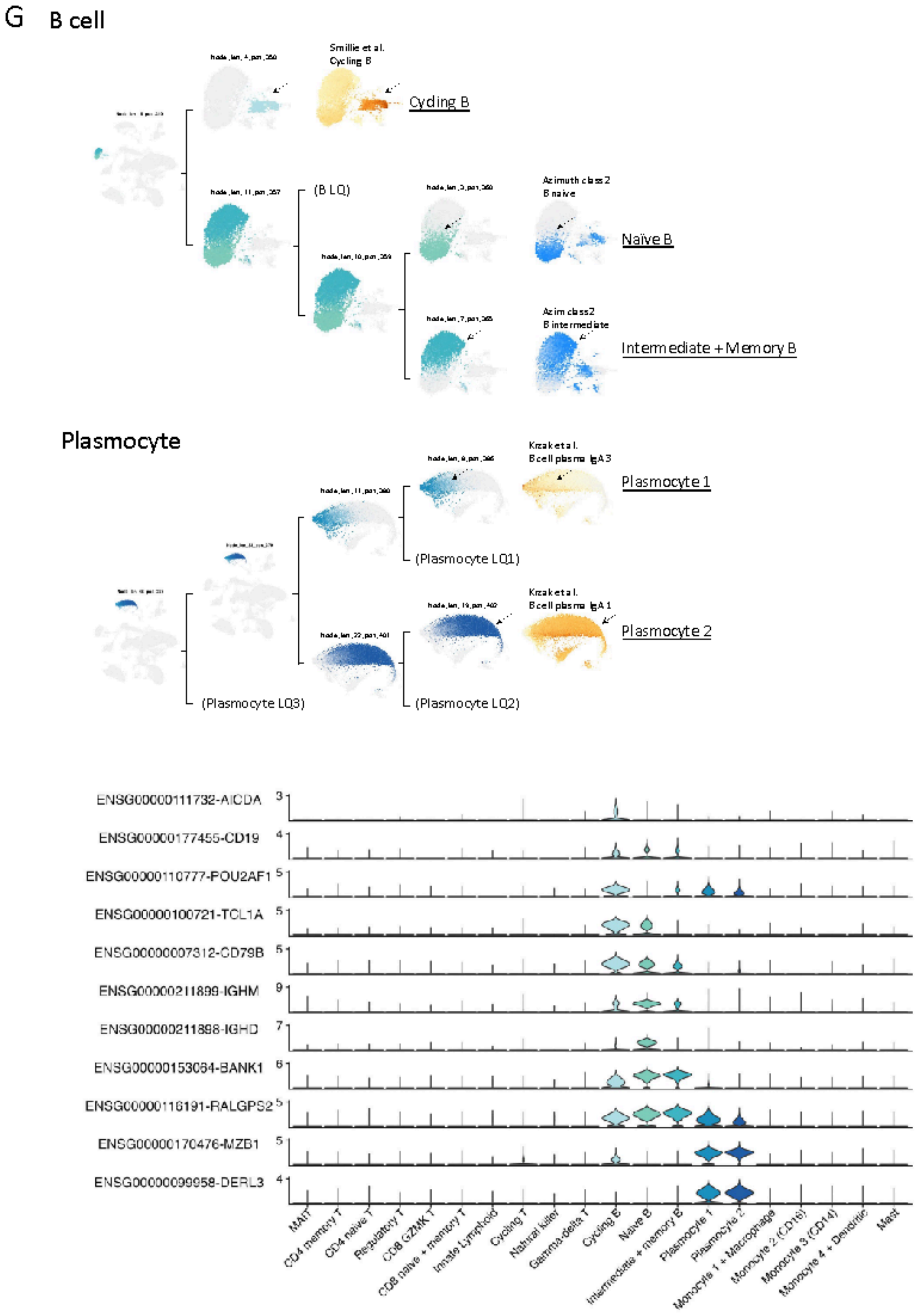
E Myeloid primary



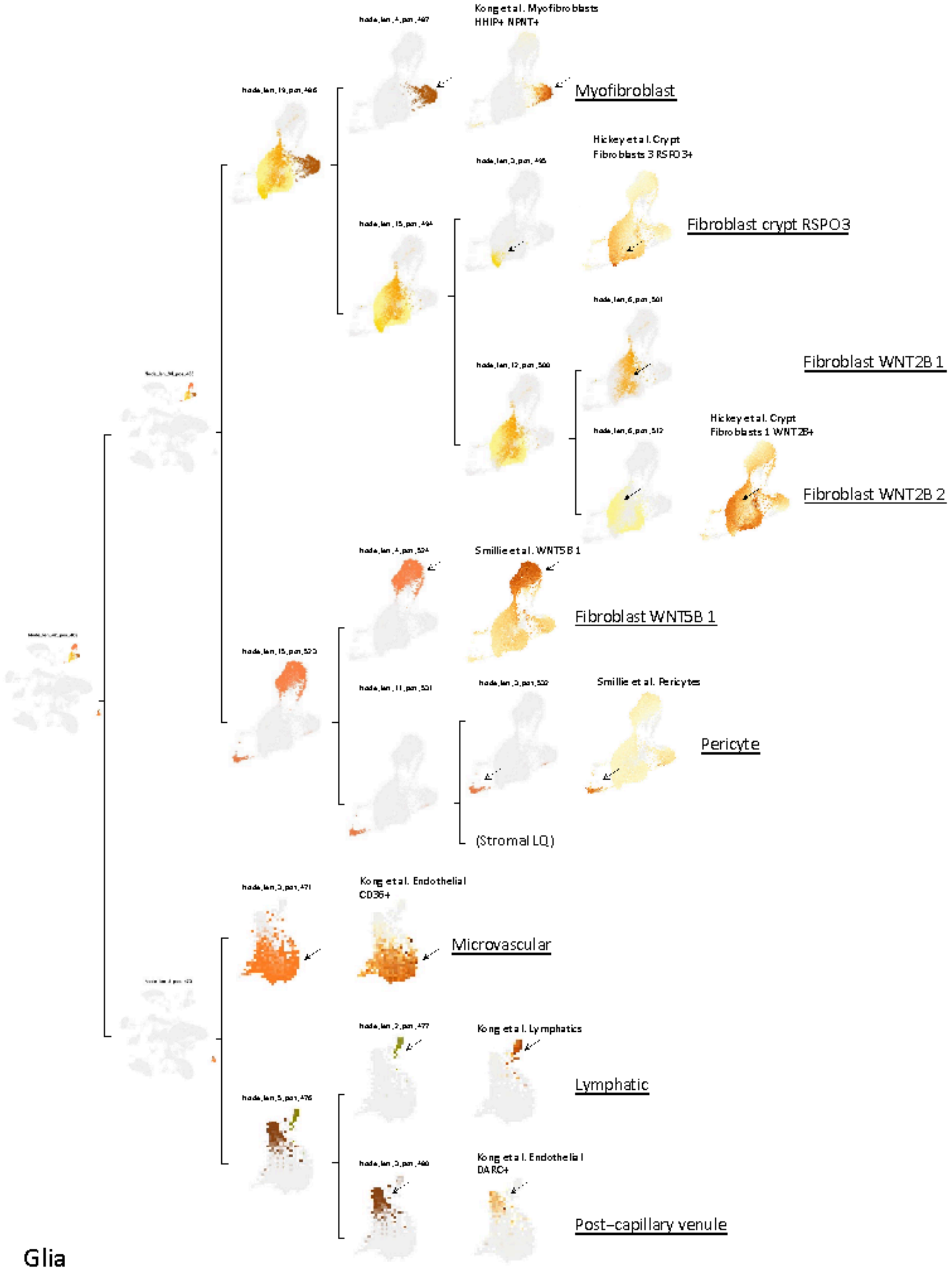
F T + Other lymphoid




ENSG00000172215-CXCR6
4
ENSG00000144290-SLC4A10
5
ENSG00000168685-IL7R
6
ENSG00000271503-CCL5
6
ENSG00000126353-CCR7
4
ENSG00000138795-LEF1
5
ENSG00000163599-CTLA4
5
ENSG00000181847-TIGIT
4
ENSG00000109943-CRTAM
5
ENSG00000113088-GZMK
5
ENSG00000077984-CST7
5
ENSG00000153563-CD8A
4
ENSG00000115594-IL1R1
4
ENSG00000157404-KIT
5
ENSG00000176890-TYMS
4
ENSG00000171848-RRM2
4
ENSG00000150045-KLRF1
5
ENSG00000115523-GNLY
7
ENSG00000171476-HOPX
5
ENSG00000227191-TRGC2
4
MAIT
CD4 memory T
CD4 naive T
Regulatory T
CD8 GZMK T
CD8 naive + memory T
Innate Lymphoid
Cycling T
Natural killer
Gamma-delta T
Cycling B
Naive B
Intermediate + memory B
Plasmocyte 1
Plasmocyte 2
Monocyte 1 + Macrophage
Monocyte 2 (CD16)
Monocyte 3 (CD14)
Monocyte 4 + Dendritic
Mast

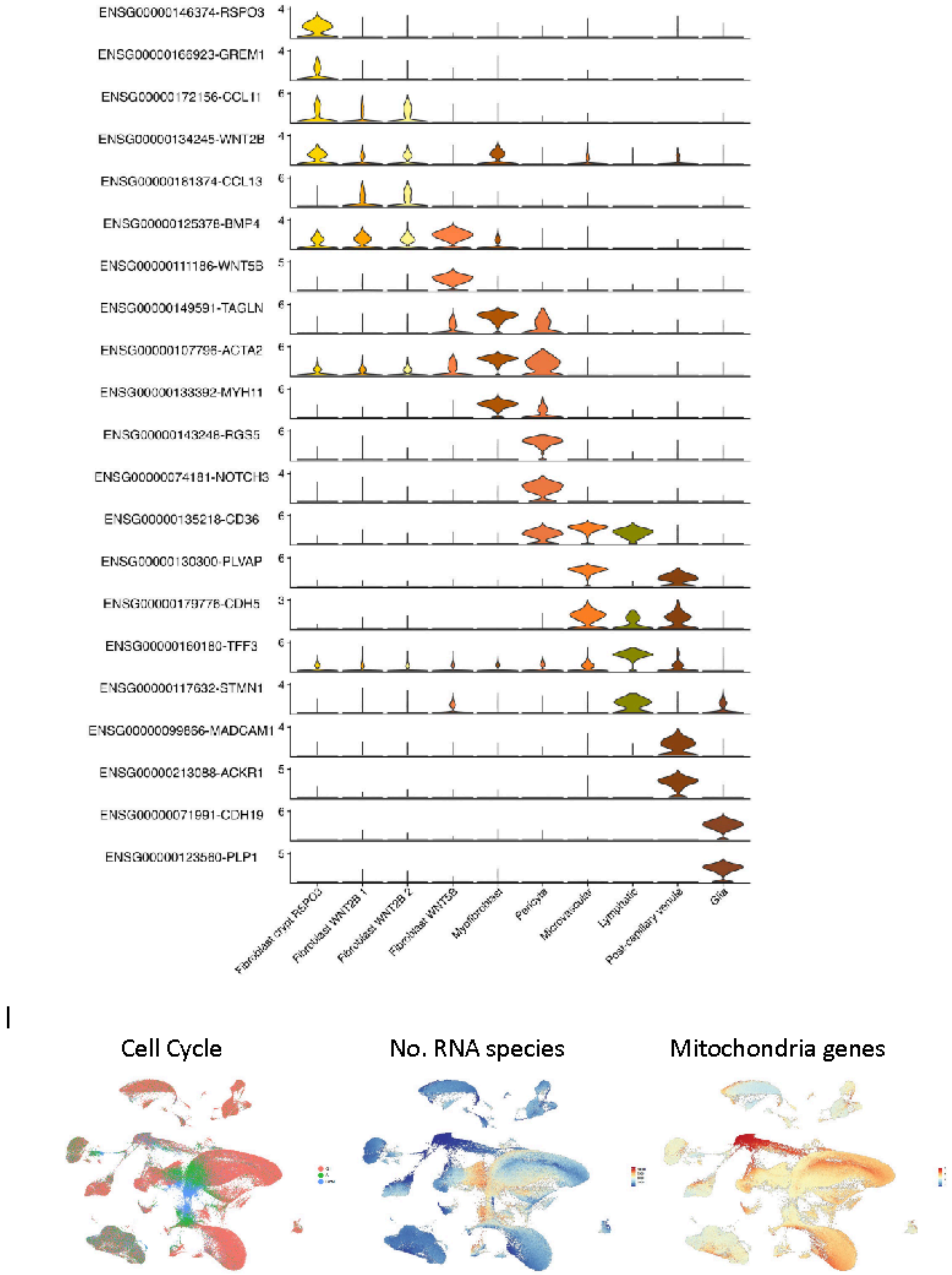

G
B cell
Smillie et al.
Cycling B
Cyding B
(B LQ)
Azimuth class2
B naive
Naïve B
Azim class2
B intermediate
Intermediate + Memory B
Plasmocyte
Kraket al.
B cell plasma IgA3
Plasmocyte 1
(Plasmocyte LQ1)
Kraket al.
B cell plasma IgA1
Plasmocyte 2
(Plasmocyte LQ3)
(Plasmocyte LQ2)
ENSG00000111732-AICDA 3
ENSG00000177455-CD19 4
ENSG00000110777-POU2AF1 5
ENSG00000100721-TCL1A 5
ENSG00000007312-CD79B 5
ENSG00000211899-IGHM 9
ENSG00000211898-IGHD 7
ENSG00000153064-BANK1 6
ENSG00000116191-RALGPS2 5
ENSG00000170476-MZB1 5
ENSG00000099958-DERL3 4
MAIT
CD4 memory T
CD4 naive T
Regulatory T
CD8 GZMK T
CD8 naive + memory T
Innate Lymphoid
Cycling T
Natural killer
Gamma-delta T
Cycling B
Naive B
Intermediate + memory B
Plasmocyte 1
Plasmocyte 2
Monocyte 1 + Macrophage
Monocyte 2 (CD16)
Monocyte 3 (CD14)
Monocyte 4 + Dendritic
Mast

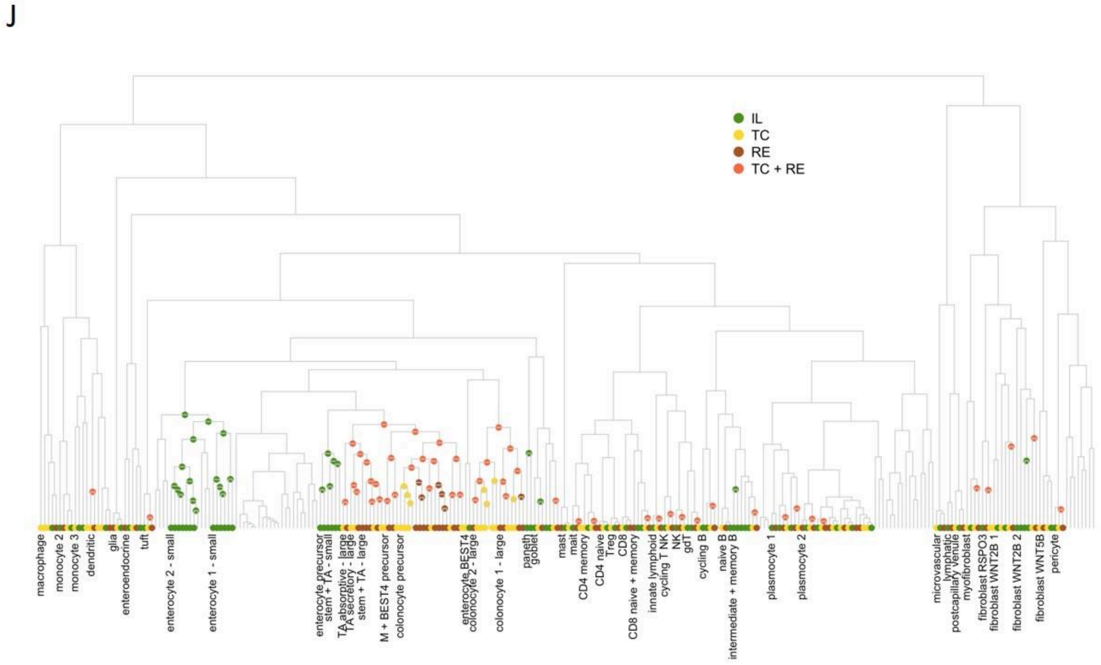
H Stromal primary



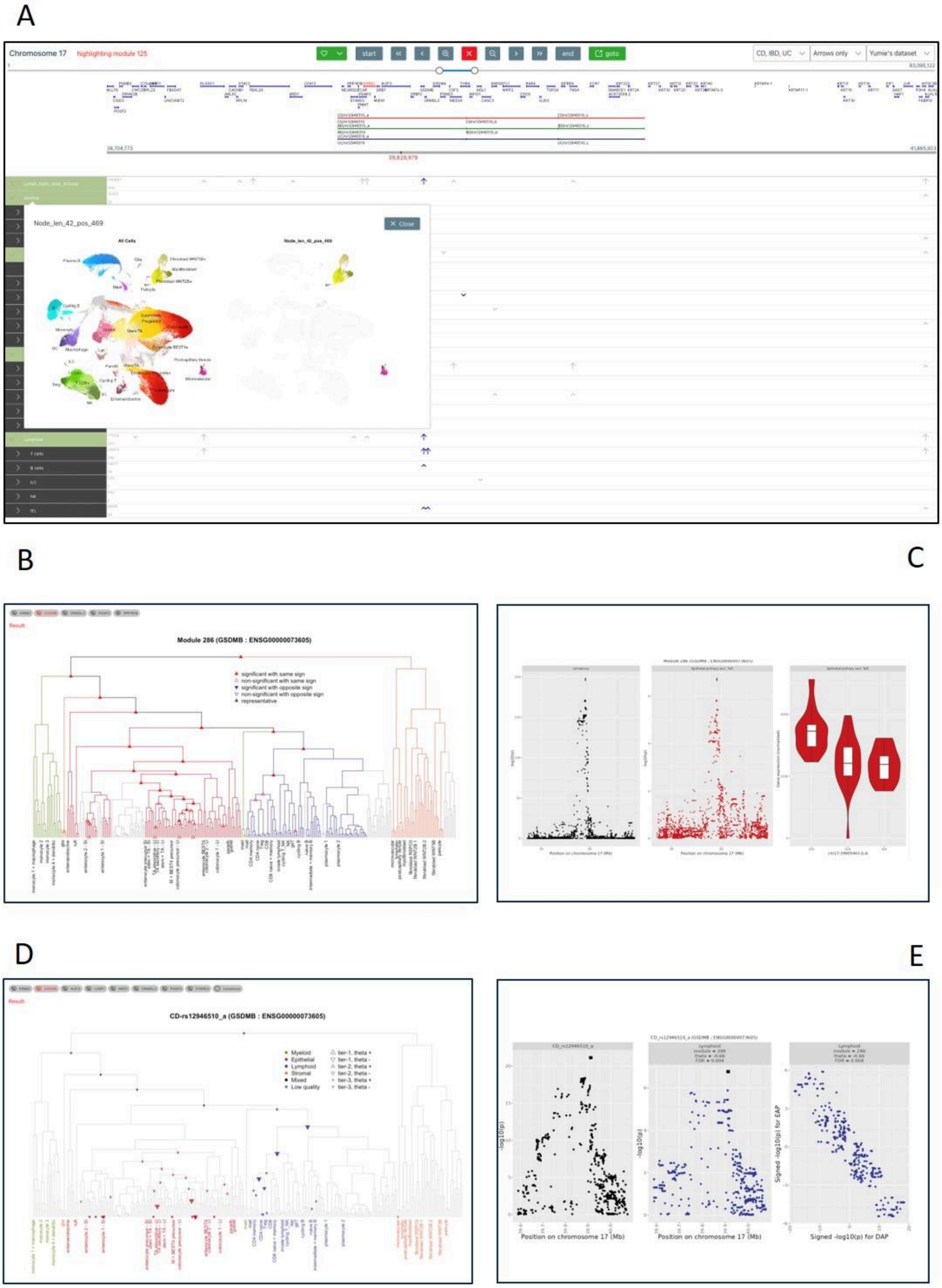
Glia

Kong et al. Glial
Glia


ENSG00000146374-RSPO3
ENSG00000166923-GREM1
ENSG00000172156-CCL11
ENSG00000134245-WNT2B
ENSG00000181374-CCL13
ENSG00000125378-BMP4
ENSG00000111186-WNT5B
ENSG00000149591-TAGLN
ENSG00000107796-ACTA2
ENSG00000133392-MYH11
ENSG00000143248-RGS5
ENSG00000074181-NOTCH3
ENSG00000135218-CD36
ENSG00000130300-PLVAP
ENSG00000179776-CDH5
ENSG00000160180-TFF3
ENSG00000117632-STMN1
ENSG00000099866-MADCAM1
ENSG00000213088-ACKR1
ENSG00000071991-CDH19
ENSG00000123560-PLP1
4
4
6
4
6
4
5
6
6
6
6
4
6
6
3
6
4
4
5
6
5
Fibroblast crypt RSPO3
Fibroblast WNT2B 1
Fibroblast WNT2B 2
Fibroblast WNT5B
Myofibroblast
Pericyte
Microvascular
Lymphatic
Post-capillary venule
Glia
I
Cell Cycle
No. RNA species
Mitochondria genes



Supplemental Figure 6: Cell type assignment of CEDAR2 intestinal scRNA-Seq data. (A-H) Illustration of the cell content of the nodes and leaves of the hierarchical tree of cell clusters. The different panels sequentially walk the reader through the tree from the root towards the leaf. In terms of cellular content, each node in the tree corresponds to the agglomeration of the cells of all dependent leaves. For each node, we show the global UMAP yet color only the cells that are part of the node. The colors used correspond to the colors of Fig. 2D, with epithelial cells in reddish tones, lymphoid cells in bluish tones, myeloid cells in greenish tones, and stromal cells in orange tones. For part of the nodes, corresponding to what were, in the end, nodes identified as the upper levels of one of 49 specific cell types, we add the same UMAP yet this time labeled in pseudo-color measuring the intensity of expression of a set of signature genes. The source of the corresponding gene signature is provided, as well as the name of the corresponding cell type (underlined). If the source is one of the following references [Smillie *et al.* 2019; Hickey *et al.* 2023; Kong *et al.* 2023; Krzak *et al.* 2023], we use yellowish tones. If the source is the Azimuth human PBMC reference mapping program [Hao *et al.*, 2021], we use bluish tones. The reason why we herein mention 49 cell types as opposed to 43 in f.i. Fig. 2A, is because we herein (SFig. 6) separately highlight ileal, colonic and rectal cell clusters for some epithelial cell type, while these are considered as one in Fig. 2A. Each lower panel shows expressions of a subset of marker genes (rows) characteristic for the indicated cell types (columns) in references [Franzén *et al.* 2019; Smillie *et al.* 2019; Burclaff *et al.* 2022; Ishikawa *et al.* 2022; Hickey *et al.* 2023; Kong *et al.* 2023; Hao *et al.*, 2024]. **(I)** Seurat cell cycle phase scores, the number of RNA detected and module scores for mitochondrial genes are shown on the UMAP. **(J)** Nodes, in the hierarchical tree, for which all dependent leaves derive from the same anatomical location (terminal ileum, transverse colon, rectum) are color-labeled accordingly. It appears that one has to go quite "high" (i.e., towards the leaves) in the tree, before colors appear. This means that for most cell types, the cells have very similar transcriptomes across anatomical locations. This is especially true for myeloid and lymphoid cell types. Separation by anatomical location occurs a bit earlier in the tree for epithelial and, to a lesser extent, for stromal cells.



Supplemental Figure 7: Screenshots from the CEDAR website (https://tools.giga.uliege.be/cedar) for the scRNA Seq analyses of the intestinal biopsies. **(A)** The X axis of the main page corresponds to the genome coordinates showing, in the top panel, the position of the genes as well as the boundaries of color-coded disease risk loci (IBD, CD and UC in the example), and in the lower panel the expression levels (marbles, not shown) and eQTL activity (arrows, as in SFig. 4 for blood) for the different genes. The Y axis corresponds to the cell type hierarchy guided by the tree resulting from hierarchical clustering of the cell clusters and their annotations to specific intestinal cell types. Clicking on a level in the cell type hierarchy shows the corresponding cell selection on the 2D UMAP. **(B)** Clicking on an arrow (corresponding to an eQTL effect) calls a set of pages for each gene in the corresponding module, each with the hierarchical tree with symbols showing the activity of the eQTL (full circles: significant eQTL, empty circles: non-significant eQTL yet matching the consensus EAP of the module). **(C)** Clicking on the symbols calls a next page that shows the consensus EAP of the module, the EAP for the selected gene for the selected position in the tree and the corresponding violin plot for the top variants of the (consensus) EAP. **(D)** Clicking on a disease locus calls a set of images corresponding to trees for all genes that show a matching EAP-DAP for the selected risk locus. **(E)** Clicking on the corresponding symbols in the image, calls gene-specific pages showing the DAP for the corresponding risk locus, the EAP for the corresponding eQTL and the theta plot.

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 (STable 15 and 16). This yielded 24,745 across-gene modules, of which 8,472 were active in more than one cell type. Of the latter, 2,172 were found to be active in both blood and biopsies (Fig. 4A). Amongst those, there was a significant excess of modules that were active in multiple cell types in both blood and biopsies ($p < 10^{-5}$), as well as modules that were lymphoid-specific in both blood and biopsies ($p = 7 \times 10^{-4}$). 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 ($p < 10^{-5}$), while modules that were enterocyte-specific in biopsies had less chance to be detected in blood ($p < 10^{-5}$), all as expected (Fig. 4B; STable 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. 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 they become gut-resident. One hundred thirty-one modules matched this pattern, of which 57 were dropped after visual EAP inspection. Of the remaining candidates: (i) 8 were mastocyte-specific eQTL, a myeloid cell type not present in blood, (ii) 39 were expressed at too low level in the circulating blood population to warrant eQTL analysis (and hence likely differentially expressed between cognate circulating and resident cells, reminiscent of scenario 1 above), and (iii) 26 were subject to gut-specific eQTL not active in the cognate circulating population despite the genes being detectable (hence reminiscent of scenario 2 above) (STable 18). 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 others (circulating). *CCL20* and *CCL24* constitute two interesting such examples (Fig. 4C-F).

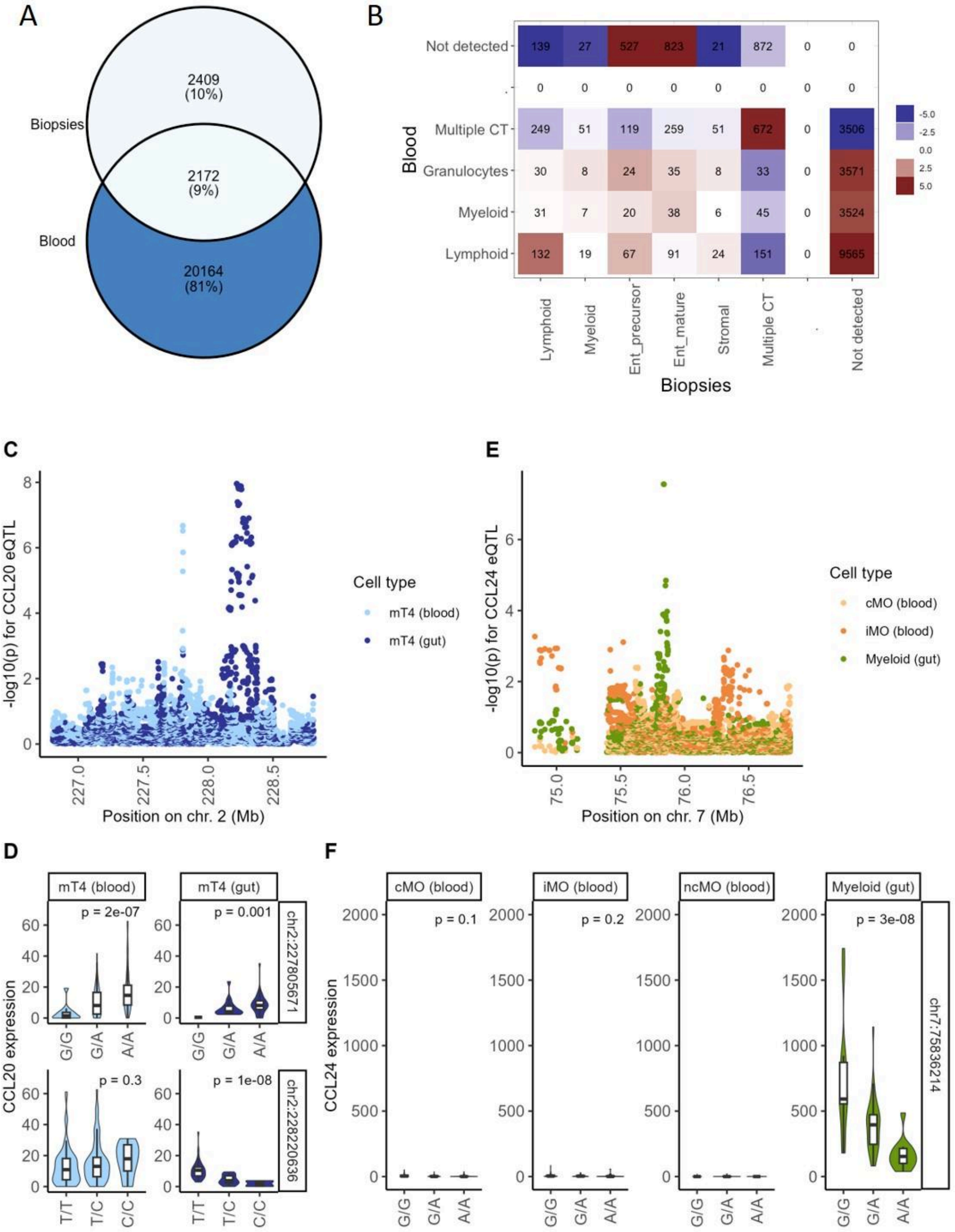


Figure 4: Merging regulatory modules across blood and gut samples. (A) When merging all blood (n=60,113) and gut (n=35,010) eQTL jointly in regulatory modules, we obtained a total of 24,745 across-gene modules, including 8,472 that are active in several cell types of which 2,172 (9%) encompassed eQTL 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 (i.e., active in several of the other categories). 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 an empirical 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). Red: excess. Blue: depletion. **(C)** 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). However, the gene is subject to two clearly distinct eQTL in these two compartments (light blue: blood mT4 EAP, dark blue: gut mT4 EAP). Of note 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 very well correspond to blood present in the biopsy. **(D)** 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) and the top SNP of the dark blue gut EAP (lower panels). **(E)** Example of a gene (*CCL24*) that is strongly expressed in gut-resident myeloid cells (including monocytes and dendritic cells) but nearly 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 (as the gene is virtually not expressed) (yellowish EAP). **(F)** Violin plots showing the expression of the *CCL24* gene in circulating monocytes (three 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).

Identifying new *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.*, 2017, 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) ~25K cases and ~35K controls (STable 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 described (Momozawa *et al.*, 2018). 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;

STable 20-22). No credible extra DAP-EAP matches were detected when using the modules' consensus EAP. To the best of our knowledge, no eQTL-based connection with IBD was previously reported for 366 of the 556 genes (STable 22). 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%). Thus, the scRNA-Seq data yielded a comparable number of DAP-EAP matches despite its limited sample size. For risk loci with matching DAP-EAP in both blood and gut, the e-genes involved were generally different. 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 ~25% of the differences in the number of matching e-genes per risk locus. 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. 8C and 8D). In circulating immune cells, regulatory modules active in all cell types contributed disproportionately to DAP-EAP matches, reminiscent of a previous report (Momozawa *et al.*, 2018). Concomitantly, in biopsies, modules assigned to the root of the tree accounted for the largest proportion of DAP-EAP matches (SFig. 8H and 8I). We note 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. 8F and 8G).

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*) (STable 23). 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 216 of the 556 genes (Table 1; STable 24 and 25). This included four genes associated with monogenic forms of human inflammatory bowel disease (*CARMIL2*, *PMVK*, *TMEM50B* and *TNIP1*), and 69 genes that upon perturbation affect susceptibility to colitis in a rodent model (STable 25). 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 heterozygote (Yıldız *et al.*, 2023). 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 (Zou *et al.*, 2020; Taniguchi *et al.*, 2021). Similarly, the chr16:rs28449958 locus encompasses *CARMIL2* causing very early onset IBD (Roncagalli *et al.*, 2016; Magg *et al.*, 2019; Bosa *et al.*, 2021), and also *SMPD3*, known to regulate TNF-α response in macrophages and B cells, and to influence the severity of dextran sulfate sodium (DSS)-induced colitis when modulated in mice (Liu *et al.*, 2017; Al-Rashed *et al.*, 2020; Li *et al.*, 2024). 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 CD (Yu *et al.*, 2024). In addition, we found that variants affecting the expression of *PRKAA1* (shown to phosphorylate and modulate *CFTR* activity (Hallows *et al.*, 2003)), *CDK19* (shown to control the *CFTR* pathway in the intestinal epithelium of mice (Prieto *et al.*, 2022)), *ADCY3* and *PRKAR2A* (both linked to β2 adrenergic-dependent *CFTR* expression (Belinky *et al.*, 2015)) 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 8 (Dimitrov *et al.*, 2019).

Recently, the whole exome of ~30,000 IBD patients was sequenced and rare-variant burden tests (MAF < 0.001) conducted for 11,978 genes (Sazonovs *et al.*, 2022). 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 Sazonovs' list of nine: *CCR7* (memory CD4 in gut) and *RELA* (plasmocytes 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, 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 (STable 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. 4D). This calls for caution when defining the desired effect of a drug (activator or inhibitor) based on the sign of θ.

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

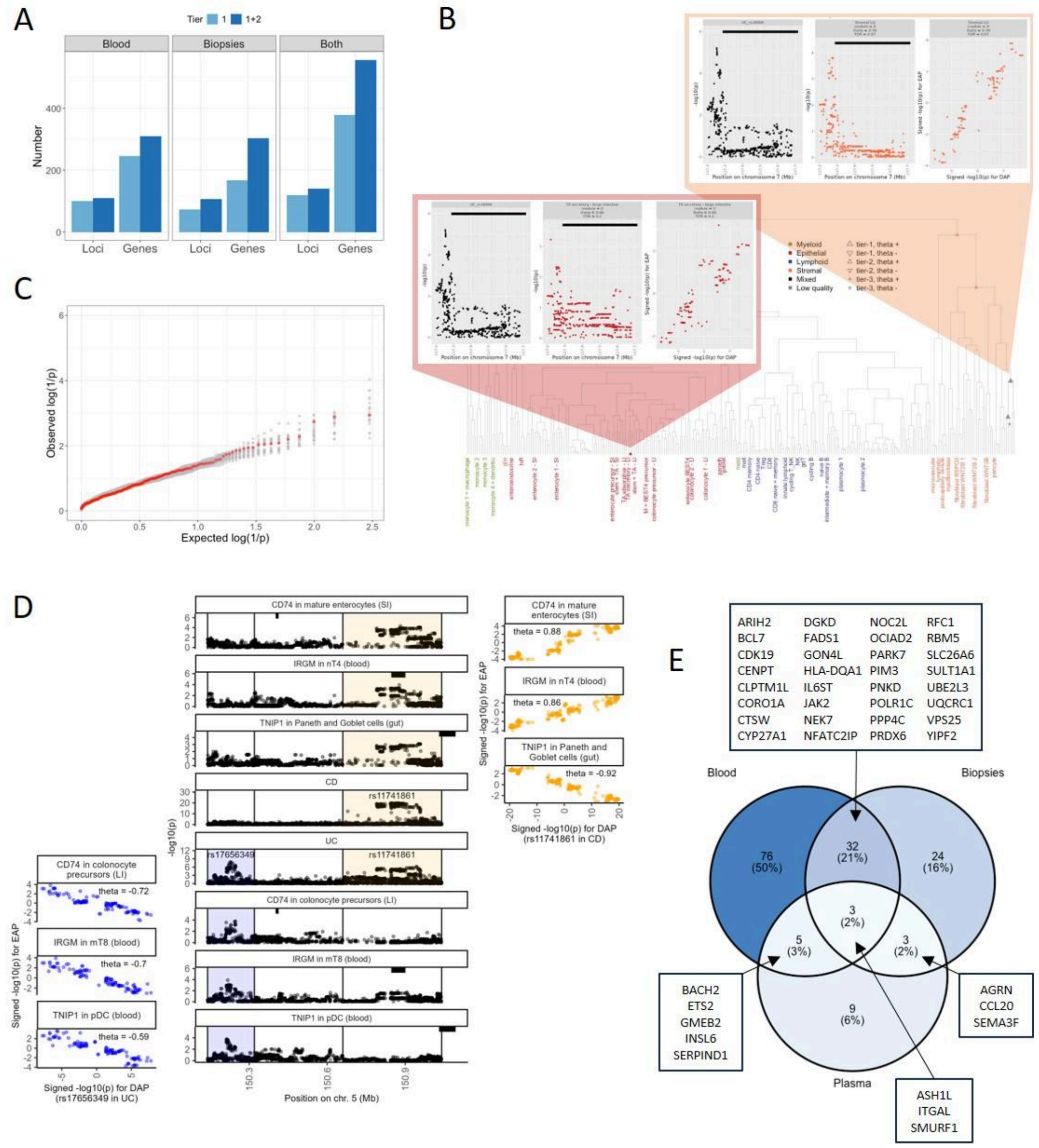
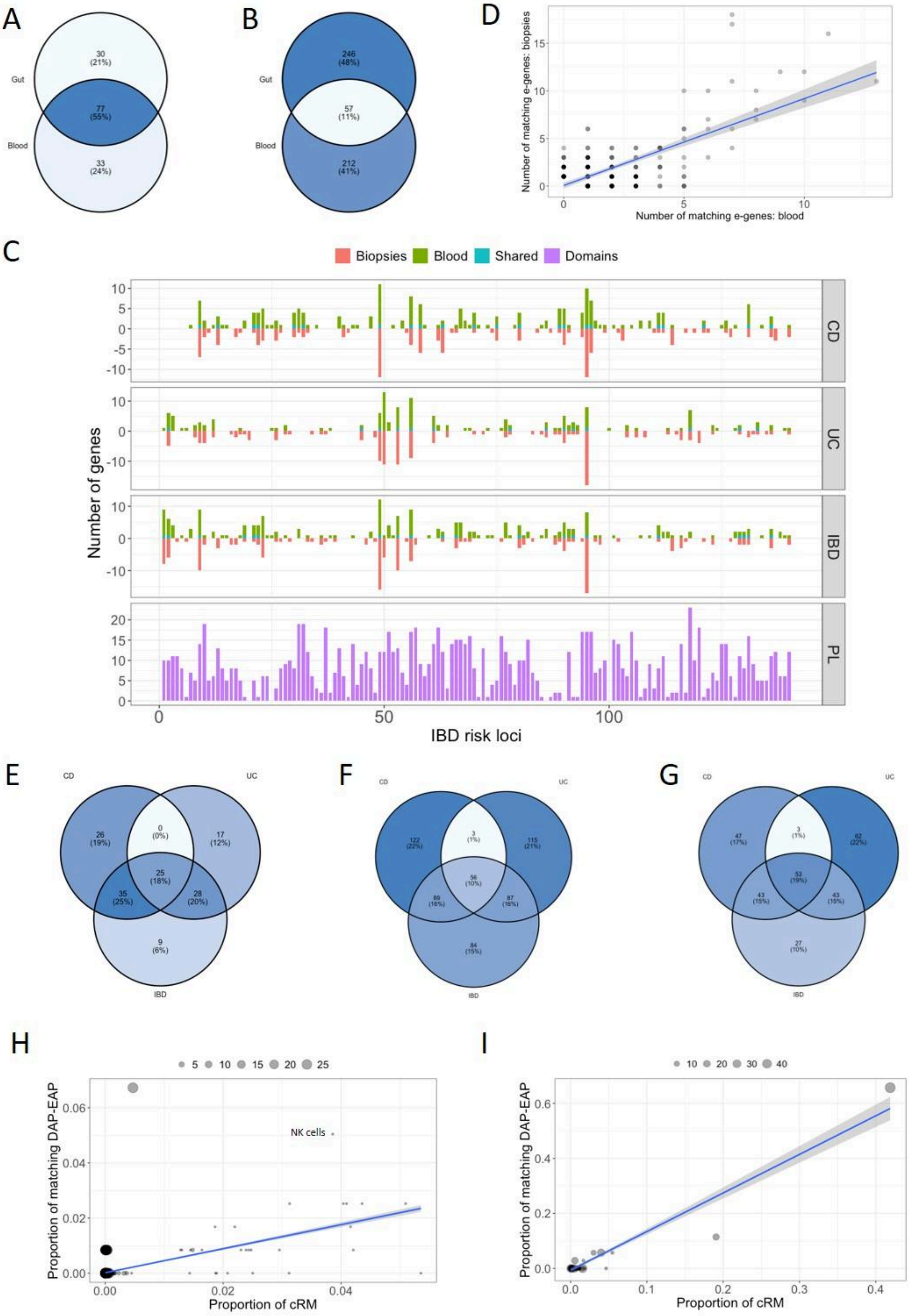


Figure 5: (A) Numbers of IBD risk loci and genes identified with matching DAP-EAP in the 27 circulating immune cell populations (blood), in the 43 intestinal cell populations (biopsies), and when combining both datasets (Both). Tier 1 corresponds to matching DAP-EAP with $|\theta| \geq 0.6$ and FDR ≤ 0.05. Tier 2 corresponds to matching DAP-EAP with $|\theta| \geq 0.6$ and $0.05 <$ FDR ≤ 0.10. **(B)** EAP of the cystic fibrosis *CFTR* gene matching the DAP of a UC risk locus on chromosome 7 (rs38904) in secretory TA precursor cells (red) and intestinal stromal cells (orange). The nodes/leaves with DAP-matching EAP are marked by triangles (large: FDR < 0.10, small: FDR > 0.10). Insets: (left) DAP for UC, (middle) EAP for *CFTR*, (right) θ plot, (red) secretory TA precursor cells, (orange) stromal cells. The boundaries of the CFTR gene are marked by the thick 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 CD[41]. **(C)** QQ plot generated with the burden test p-values obtained for 298 of our 556 DAP matching e-genes by (Sazonovs *et al.*, 2022) (red dots). Grey dots correspond to QQ plots obtained with randomly sampled sets of 298 genes from the list of genes with data in Sazonovs *et al.*, 2022. **(D)** 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. **(E)** DAP matching e-genes whose expression levels are affected by the IBD disease process 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 biopsies (Kong *et al.*, 2023), or in plasma (Eldjarn *et al.*, 2023).



Supplemental Figure 8: (A) Number of IBD risk loci with matching DAP-EAP in intestinal biopsies only (gut, n=30), in circulating immune populations only (blood, n=33), and both in gut and blood (n=77). **(B)** Number of e-genes with matching DAP-EAP in intestinal biopsies only (gut, n=246), in circulating immune populations only (blood, n=212), and both in gut and blood (n=57). **(C)** Number of e-genes with matching DAP-EAP in 206 IBD risk loci (ordered by chromosomal position along the X-axis) in blood (green bars) and biopsies (orange bars, numbers x -1). The blue bards correspond to the number of e-genes shared between blood and biopsies for a given risk locus. Numbers are given separately for CD, UC and IBD risk loci (see STable 21). Intestinal biopsies only (gut, n=246), in circulating immune populations only (blood, n=212), and both in gut and blood (n=57). The panel with the violet bars at the bottom, labeled "PL" (-eiotropy), reports the sum of the number of "trait domains" that are influenced by the pleiotropic genomic loci, defined by Watanabe *et al.*, 2019, that overlap with the corresponding IBD risk locus. **(D)** The number of matching e-genes (i.e., with matching DAP-EAP) in blood (X-axis) and biopsies (Y-axis) is highly correlated across the IBD risk loci, despite the fact that the corresponding e-genes mostly differ between blood and biopsies (see blue bars in panel C). **(E)** Numbers of risk loci with matching e-genes for CD, UC and IBD, with corresponding overlaps (i.e., risk loci with matching e-genes for CD and UC but not IBD (n=0), CD and IBD but not UC (n=35), UC and IBD but not CD (n=28), CD, UC and IBD (n=25)). **(F)** Numbers of matching e-genes for CD, UC and IBD, with corresponding overlaps (i.e., e-genes matching the DAP for CD and UC but not IBD (n=3), CD and IBD but not UC (n=89), UC and IBD but not CD (n=87), CD, UC and IBD (n=56)). **(G)** As for (F), yet restricting the analysis to 81 loci that are risk loci for both CD and UC. **(H)** The proportion of matching DAP-EAP involving "blood" *cis*-acting regulatory modules (= cRM) with a given "activity vector" (i.e., the combination of the 27 cell types in which the module is active) is correlated with the proportion of regulatory modules (= RM) with that "activity vector". Yet, some types of modules stand out. Regulatory modules that are active in all cell types account for only 0.0047 of all across-gene modules, yet account for 0.0672 of DAP-EAP matches (> 14-fold excess), reminiscent of Momozawa *et al.* [2018]. Each dot corresponds to one of 3,572 observed activity vectors. The size of the dot corresponds to the number of "1"'s in the vector (i.e., the number of cell types out of 27 in which the module is active). The graph was obtained considering across-gene modules and the three diseases (CD, UC and IBD) combined. **(I)** The proportion of matching DAP-EAP involving "biopsy" regulatory modules encompassing a number of cell types (i.e., out of the 43 distinguished cell types; see also Fig. 2A) is correlated with the overall proportion of cognate regulatory modules (RM = *cis*-acting regulatory module). Yet regulatory modules assigned to the root of the hierarchical tree (Fig. 2F), and hence encompassing mostly modules that are active in many cell types, account for the highest proportion of DAP-EAP matches. Each dot corresponds to one of 64 observed cell type groups. The size of the dot corresponds to the number of cell types (out of 43) included in the group. The graph was obtained considering gene-specific modules and the three diseases (CD, UC and IBD) combined.

Repurposing candidates 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 (STable 26). We additionally consulted lists of genes that were shown, from scRNA-Seq data, to be differentially expressed between intestinal biopsies of IBD cases and controls (Kong *et al.*, 2023), as well as lists of proteins that were differentially abundant in plasma of IBD cases and controls (Eldjarn *et al.*, 2023). 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. 5E).

Hundred eighty-one drugs targeting 180 e-genes are or have been used/tested to treat IBD (Mountjoy *et al.*, 2021; Vieujean *et al.*, 2025) (STable 27). 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* (ritlecitinib: phase 2 completed with positive results for CD and UC) and *TNFSF15* (several: phase 2 ongoing) (STable 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 (STable 22).
We further identified 40 genes in our list of candidates, targeted by known drugs (in phase 1 or higher) that – to the best of our knowledge – have not been tested in the context of IBD (STable 22). For 16 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 two (*PIM3, RPS6KB1*), phase 2 for six (*ATP2A1, CDK11B, ERAP2, HLA-DRB1, KIR2DL1, PPP5C*), phase 3 for one (*INPP5D*), and approved for at least one disease other than IBD for seven (*CFTR, CYP3A5, IL18R1, IL18RAP, LAMA2, NEK7* and *PTGIR*). For 6 of the 16 genes (underlined), expression was affected by the disease process in a manner consistent with θ, at least in one of the three datasets. Detailed examination of the drugs that were in phase 3 or higher (Supplemental Material 1), identifies entrectinib (ENB) as a possible 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 tumors (Liu *et al.*, 2018). It was shown to bind to arginine 121 of NEK7, thereby inhibiting its interaction with NLRP3 (Jin *et al.*, 2023). Downregulation of *NEK7* by intraperitoneal injection of lentiviruses expressing anti-*NEK7* shRNAs was shown to attenuate DSS-induced colitis (Chen *et al.*, 2019). In mouse models, ENB effectively reduced symptoms of NLRP3 inflammasome-related diseases (other than colitis) (Jin *et al.*, 2023). ENB has a high safety profile and is well tolerated by almost all patients without cumulative toxicity

(Jiang *et al.*, 2022). Genetic variants that increase *NEK7* expression in naïve B cells increase the risk for IBD (θ=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 controls (Kong *et al.*, 2023). 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 up-regulated in some circulating immune cells of active CD patients (f.i. eosinophils, STable 22), and these findings require further scrutiny.

Function		Genes
Epithelial homeostasis & barrier function	Epithelial transporters	ABCC2, ADCY3, CDK19, CFTR, MFSD4B, PRKAA1, *PRKAR2A,* SLC5A6, SLC9A3
	Epithelial integrity & repair	*CD74*, CDH3, CERS2, *CTSK, ERBB2,* HSPD1, *INAVA*, IRF6, KSR1, LASP1, MASTL, PARD6A, PEX6, PLAU, RAB13, *TKT*, ZEB2-AS1
	Epithelial stem cells & development	CDX1, HSPG2, *LAMA2*, NRBP1, SETDB1, SRCAP
Signalling in innate immunity	Pathogen-recognition receptors	ANKRD17-DT, ARIH2, *ATG16L1*, CARD9, CYLD, IL1R1, IL1R2, IL18R1, IL18RAP, *INAVA*, *IPMK,* IRF5, *LACC1*, *LAMA2*, NEK7, ORMDL3, PRKDC, *RFTN2*, RNASET2, *RNF123*, *SEC24C*, SSC5D, SENP7, *TRAF3IP3*, *TRAIP*, TRIM37, *IRGM*, UBR2, *USP19*, ZC3H15, ZDHHC11B, ZFP91
	TNF signalling	*ADCY7*, BABAM2, LTBR, MST1, MST1R, RASSF1, TNFRSF14, TNFRSF9, TNFSF15, TRAIP, ZFP36L2
	JAK-STAT signalling	FAM220A, JAK2, OCIAD2, *PRKAR2A*, PTPN2, TYK2
	IL6, SLAMF, Others	ASH1L, *DGKD*, EDC4, IL26, IL6R, IL6ST, *Ly9*, OSM, *RASGRP1*, *SLMAF1, SLMAF4,*
Cells in innate immunity	Leucocyte extravasation & trafficking	CDH26, *CORO1A*, CXCR1, CXCR2, CXCL2, CXCL5, CXCL1, CXCL8, GCA, ITGA4,ITGAL, LSP1, NEU4, SEMA3F, TLN1, TMEM87A, TSPA17
	Macrophages	AUH, *CCR2,* ETS2, *LACC1*, PPM1F, PROK2, QPCTL, SMPD3, *SP140*
	Other or multiple myeloid cells	ARNT, FOXP1-AS1, HINT1, *INPP5D*, KIR3DL1, KIR3DL3, PLPP6, *PPP3CB*
Autophagy & ubiquitin-dependent processes	Autophagy	*ATG16L1*, CDK11B, CLTC, *CORO1A,* CUL1, DAP, IPMK, *IRGM*, NR1D1, PCNP, PIM3, *PPP3CB*, PTPN23, SHSA5, TFAM, TUFM, *USP19,* VPS37C
	Ubiquitin-dependent processes	BBC3, CSID1, CUL2, FBXL19, KBTBD6, *LTBR*, MIB2, OTUD7B, PLA2G15, QRICH1, *RNF123*, *SEC24C*, SMURF1, TOM1, UBE2L3,
Other inflammation mediators	Prostaglandins & glycocorticoids	FADS1, FADS2, HSD11B1, PTGIR, PRXL2B, *PRKAR2A*
	Complement system	CD46, CR2, GNA12
Antigen presentation & MHC		*CD74*, CIITA, CTSS, ERAP2, LNPEP, RFX5, (+MHC)
Signalling in adapative immunity	BCR &/or TCR	CD19, LAT, LIME1, PTPRC, *PTPRH*, *RFTN2*, SLAMF8, THEMIS, *TRAF3IP3*
	CCR6, Others	*ADCY7*, CCL20, CCR6, *DGKD*, DUSP8, ELF3, *Ly9*, *RASGRP1*, SH2B3, *SLAMF4*
Cells in adaptive immunity	T cells	ATXN2L, BCL2L12, *CCR2*, *CORO1A*, CREM, *CTSK*, CTSW, FOSL2, GMEB2, GSDMB, MED1, NFATC2IP, PDCD1, PLXNB1, *PPP3CB*, PPP4C, SEMA3G, TAGAP, TEC, *TKT*, USP4
	B cells and multiple	BACH2, BRWD1, HHEX, IKZF3, *INPP5D*
Other	Other mechanisms	COQ10B, CYP3A5, FUT2, GPX1, HYAL1, LIG1, NADK, OAT, *QPRT*, RXFP4, SLC22A5, SLC25A15,*SP140*, SPHK2, TAS1R3, THRA
	Monogenic forms of IBD	CARMIL2, PMVK, TMEM50B, TNIP1 (*per* TNFAIP3)

Table 1: List of 211 e-genes with DAP-matching EAP and published evidence (documented in STable 25 and Suppl. 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.

DISCUSSION & PERSPECTIVES

Regulatory modules: sensible (e)QTL analysis framework?

It is generally admitted that most CCD risk variants act by perturbing the expression of causal genes in one or more disease-relevant cell types, and that it should be possible to pick up many of these regulatory effects as *cis*-eQTL in these cell types. *Cis*-eQTL effects are pervasive and it is therefore not sufficient to identify a *cis*-eQTL overlapping a risk locus to assume that it affects risk. By definition, causal *cis*-eQTL are determined by the same variant(s) that affect disease risk. As a consequence, the pattern of association between regional variants and gene expression (EAP) should be the same as that for the disease (DAP). This is certainly the case if disease and gene expression are measured in the same individuals. It is also the case if disease and gene expression are measured in distinct cohorts, provided that they share local LD structure. The DAP-EAP similarity applies to the causal variant(s) *per se*, but also to passenger variants whose association with disease and gene expression is indirect, reflecting their LD with the causal variant(s). The expression of a given gene may be controlled by distinct sets of regulatory variants in distinct cell types. This will yield distinct EAP for the same gene, even if the respective sets of regulatory variants partially overlap. The DAP will only match the gene's EAP in the disease-relevant cell type. It is conceivable that the expression of a causal gene in more than one cell type (with distinct EAP) influences disease risk. If the significant variants for the corresponding EAP are sufficiently distant, multiple DAP-matching EAP may be found for the same gene in different cell types. We have observed several such instances in this work, for adjacent sub-risk loci or even adjacent risk loci considered separate thus far. We have also observed instances where distinct EAP (corresponding to different cell types) for the same gene match the DAP for different diseases (in this case CD and UC). If the significant variants of the distinct EAP overlap, the DAP may not match either EAP. More advanced approaches would be needed to dissect such cases, converging towards fine-mapping of multiple independent variant effects. This, however, requires larger sample sizes than what is presently available for multi-tissue eQTL studies.

Causal *cis*-eQTL are the first links in the chain connecting risk variants with disease, the final outcome. The same regulatory module-based approach can in theory be used to uncover the intermediate molecular links between risk variants and disease, provided that the abundance or state of the corresponding molecular species can be quantified. The corresponding *trans*-QTL (whether expression QTL or any other quantifiable molecular phenotype) should be characterized by DAP-matching EAP. CCD are highly polygenic, influenced by hundreds of risk loci or more. It is likely that at least some pathways linking variants with disease converge prior to disease outcome. Thus, some intermediate molecular phenotypes will have matching EAP with multiple DAP (distinct risk loci). This should allow reconstruction of the topology of the pathways linking the multiple risk variants with the disease.

Cell type specificity of regulatory modules operating in immune cell populations

One striking observation of this work is that 73.2% of blood regulatory modules were found to be active in only one of the 27 studied immune cell populations. Most of the time (100-8.5=91.5%) the corresponding genes did not appear to be under marked genetic *cis*-control in the 26 other cell types (i.e., no module switch). Hence, a large proportion of eQTL appear to be very cell type-specific at the chosen level of granularity, at least in blood. One could argue that this is a power issue: the eQTLs may have existed in some other cell type, but remained under the radar of statistical significance because of insufficient sample size. We therefore expanded our search for matching EAP to non-significant eQTL, i.e., we allowed for tier-2 eQTL to enter into the modules if their EAP matched significant ones with $|\theta| \geq 0.6$ and a combination of *p*-values of match and eQTL ensuring an FDR ≤ 0.05. This had virtually no effect on the proportion of modules that were active in only one cell type (from 76% to 73.2%). We therefore think that the observed eQTL cell type specificity is genuine. These findings are reminiscent of those reported in circulating immune cells (Schmiedel *et al.*, 2018; Ota *et al.*, 2021), and across multiple tissues (The GTEx Consortium *et al.*, 2020).

The eQTL specificity may even be more pronounced for at least some immune cell populations, being in addition context-dependent. Indeed, the availability of scRNA-Seq data for intestinal biopsies allowed us to compare eQTL activity for the same cell type yet circulating in blood on the one hand, and resident in the intestine on the other hand. To detect such compartment-specific eQTL, we focused our attention on modules that were significantly active in (i) lymphocytes, or (ii) macrophages/monocytes/dendritic cells isolated from the intestinal biopsies, but not in the equivalent cell types isolated from blood. We didn’t add the mirror comparison, i.e., modules active in blood but not in gut, because we assumed that we had more (statistical) power to detect eQTL in blood than in gut. The absence of a detectable eQTL effect in the gut could more often be a trivial power issue. We detected several instances supporting the existence of such compartment-dependent eQTL. Part of these seem to involve induction of gene expression upon entering the intestinal compartment (scenario 1), others seem to be independent of gene expression level but a genuine conditional effect of the regulatory variants (scenario 2). We illustrate both scenarios using *CCL24* in monocytes/macrophages and *CCL20* in T lymphocytes, two CC-motif chemokines with chemotactic and antimicrobial activity whose genes show to have higher expression and be under specific genetic *cis*-control in the gut.

Using inferred cell type ontogeny to effectively map eQTL using scRNA-Seq data

Obviously, the observed degree of eQTL cell type specificity will depend on the chosen cell type granularity. This becomes particularly pertinent when working with single cell (ultimate granularity) RNA-Seq data: if one splits a cell cluster in sub-clusters, what was an eQTL specific for the cluster may now become shared by the sub-clusters (and hence apparently less cell type specific). Deciding at what cluster resolution to perform eQTL analysis will also have a considerable impact on eQTL detection: considering two clusters that share an eQTL separately rather than together, may decrease the detection power if the number of cells in each cluster is power limiting. The optimal cell partitioning strategy to detect a given eQTL/module will depend on where (i.e., in which cell types) the eQTL/module is active. If an eQTL is active in all cell types, the best strategy is to consider all cells jointly. If an eQTL is specific to a very small subset of cells, merging these cells with others that do not express the eQTL will reduce detection power. To address these issues in a generic way, we decided to construct a hierarchical tree of cell clusters based on the similarity of their transcriptomes. We assumed and demonstrated that this tree largely reflects cellular ontogeny. We also assumed that regulatory modules are turned on at specific developmental stages (i.e., nodes or leaves of the tree) and affect at least part of downstream branches along variable length. Accordingly, we performed eQTL analysis separately for all leaves and nodes of the tree. This effectively guides and limits the cell pooling options in an ontogenic framework, yet allows informed exploration of many possible scenarios. Of note, the proposed method disconnects eQTL mapping from cell type annotation.

The eQTL detection step was followed by the merging of similar EAP into modules, and the visualization of where a given module is active along the hierarchical tree. Assuming that the module has been turned on once along the ontogenic tree, the signal should be the strongest for the node corresponding to the ontogenic stage where it was turned on. The signal should become weaker as one moves towards the root as the cells in which the module is active are progressively diluted by cells in which it is not. It should also become weaker as one moves towards the leaves as the number of cells for analysis decreases. Thus, one could assign the module to the segment in the tree where the detection signal is maximal. In reality we didn’t see the predicted smooth decrease of signal strength up and down-wards (in the tree) from a point of maximum significance, presumably because signal strength is affected by a multitude of other factors. We therefore chose to assign the module to the most recent common ancestor (MRCA) of all active leaves/nodes, which may result in positioning the modules too much towards the root.

The number of nodes/leaves in which modules were active showed a clear sign of overdispersion with many modules being either active in fewer nodes/leaves than expected by chance, or active in more nodes/leaves than expected by chance. As for blood, the excess of modules active in fewer than expected nodes was unlikely due to statistical power issues as augmenting modules with matching yet less significant eQTL didn’t affect this pattern. We believe that this indicates that many regulatory modules are cell type-specific in the gut as previously observed for the FACS/MACS sorted circulating immune cell populations. In particular, enterocytes and stromal cells from the small intestine differ considerably from those of the large intestine, both with regards to transcriptome and eQTL activity.

Increasing cell type granularity uncovers new IBD-driving regulatory modules

The initial premise of this work was that a large fraction of risk loci remained "orphan" thus far, because the disease driving-eQTL were active in cell types that were absent or underrepresented in previous eQTL datasets. The observation that a large fraction of eQTL/modules indeed appear to be highly cell type and even context specific supports this hypothesis. Accordingly, this work in essence doubles the number of IBD risk loci with DAP-matching eQTL to 140, i.e., ~70% of the 206 studied risk loci. For approximately half of risk loci (55%), DAP-matching EAP were detected in both blood and gut. For the remaining half, there were slightly more matches in blood (24%) than in the gut (21%). Thus, scRNA-Seq-based eQTL detection in the gut made a considerable contribution to DAP-EAP matching, despite analyzing samples from only 57 individuals (yet in three locations) and detecting a lower number of genes. It seems reasonable to assume that increasing intestinal scRNA-Seq sample size will uncover DAP-matching EAP for additional risk loci, and meta-analyses towards that goal are in progress.

An additional factor that has contributed to the marked increase in the number of IBD risk loci with DAP-matching EAP is the splitting of risk loci into sub-risk loci. This was done for 32 of the 206 studied risk loci, because we assumed that "multimodal" DAP might be the sum of multiple independent EAP. This strategy allowed us to detect an extra 28 DAP-matching e-genes, covering an additional nine IBD risk loci.

It has been argued that it may be more effective to perform eQTL analyses in fewer, easily collectable sample types (f.i. whole blood) but from many more individuals, than to increase sample types yet remain limited in the number of individuals, to uncover more DAP-matching EAP. To verify this, we used PBMC eQTL summary statistics from ~30,000 healthy individuals of European descent (Võsa *et al.*, 2021), and searched for DAP-matching EAP using a slightly more permissive procedure as the one used with our own expression data. Using Võsa’s data, we identified DAP-matching EAP for 39 loci involving 59 genes. Using our own PBMC information (i.e., 187 individuals), we identified DAP-matching EAP for 32 IBD risk loci involving 42 genes, with 12 overlapping loci and three 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 36 loci and 27 genes overlapping with Võsa. Using our complete data set (blood and gut), DAP-matching EAP for 140 loci involving 556 genes, of which 38 loci and 30 genes overlapping with Võsa. Thus, the Võsa data enabled the identification of matching EAP for one IBD risk locus (rs1479918), and 29 e-genes that were missed with our data. However, we identified matching EAP for 102 IBD risk loci and 526 e-genes that were missed with Võsa's data. These findings suggest that a large number of additional matches are uncovered when performing eQTL analyses in isolated cell types. This also indicates that increasing the cell type granularity in our blood RNA-Seq data—by applying deconvolution methods based on our intestinal scRNA-Seq data—could help uncover DAP-matching EAP for additional risk loci.

Are the e-genes controlled by IBD-driving regulatory modules causal?

The initial assumption, when eQTL studies to identify causal genes in risk loci were initiated, was that DAP-matching EAP would occur rather exceptionally, yet - when detected - would provide strong evidence for gene causality. It now appears that DAP-matching EAP are rather common, and that for many risk loci, DAP-matching EAP are found for multiple genes. Moreover, we find in this work that when, for a given risk locus, DAP-matching EAP are found in both blood and gut, the genes involved are largely different. Are all of these genes, one way or the other, causally involved in disease risk, only some, or none? In other words, what is the proportion of "red herring" eQTL (Connally *et al.*, 2022) amongst DAP-matching EAP? All scenarios are plausible, and each one likely applies to at least some risk loci.

It seems possible that - because *cis*-acting regulatory elements are often shared by multiple genes [f.i. Thurman *et al.*, 2012] – many regulatory variants will affect the expression of neighboring genes including some that have no bearing on disease risk. The relatively modest signals obtained by pathway enrichment analysis supports the assertion that a sizable proportion of DAP-matching e-genes are not directly influencing disease risk. It is tempting to assume that DAP-matching information will be more specific for risk loci with fewer matching e-genes.

On the other hand, when scanning the literature for functional evidence supporting the causal involvement of genes in our candidate list, we were struck by the large number of DAP-matching e-genes with such evidence (212 out of the 483 examined coding genes). Several risk loci harbor more than one gene with quite enticing support (see examples, in results, of loci with strong functional candidates that additionally harbor less well-known genes underpinning early onset, Mendelian forms of IBD). Thus, it seems likely that – at least for some risk loci – the risk variants are affecting the expression of multiple genes that jointly affect the risk to develop IBD. In other words, the notion of polygenicity may not be limited to the fact that disease risk is affected by multiple loci in the genome, but additionally that (at least some) risk loci harbor multiple causal genes controlled by the same or distinct regulatory modules. 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 cells (Yan *et al.*, 2017), yet at the same time to regulate the stability of adherens junctions of intestinal epithelial cells (where we see the best DAP-match) (Mohanan *et al.*, 2018). Also, *ATG16L1* is increasingly understood to affect disease outcome through autophagy-dependent but also -independent mechanisms (Hamaoui and Subtil, 2022), in agreement with our observation of a DAP-match in eosinophils but also colonocyte precursors. The suggestion that specific risk loci may affect disease outcome through multiple genes and pathways is also well illustrated by the *CCR6*/*RNASET2* and *IRGM/CD74/TNIP1* gene sets.

Is it possible that for some risk loci, none of the DAP-matching e-genes are causal? Positional cloning has been pursued very successfully during the last 30 years, under the assumption that causal mutations always affect the function of a causal gene in *cis*. A key assumption of the omnigenic model (Boyle *et al.*, 2017; Liu *et al.*, 2019) is that variants can affect phenotype without having any causal *cis*-effect on the expression of neighboring genes. That doesn’t mean that one will not see *cis*-eQTL effects on neighboring genes, but rather that these do not influence the phenotype: a sobering thought for positional cloners.

Of note, we didn’t see an effect of the number of DAP-matching genes, with or without reported function, per risk locus on the magnitude of its effect on disease (measured by the odds ratio (OR)). This is not surprising given that many factors will affect OR, but a positive result would have been in support of the "multigenic" nature of individual risk loci whether being causal *per se* or not (i.e., omnigenic model).

Are multi-genic and multi-tissular regulatory modules underpinning pleiotropy?

The number of DAP-matching e-genes is remarkably high for some risk loci, often despite any obvious functional theme or coherence. This suggests that risk variants may hit master *cis*-regulators that control the expression of large chromosome domains and many genes therein, irrespective of function. We reasoned that such variants might, because of the number of affected genes, influence multiple traits, i.e., be pleiotropic. Of interest, under this scenario, the mechanism accounting for pleiotropy would only involve the variants, not the e-genes affected in *cis*, as causal genes would differ between traits. It is well established that as many as 90% of GWAS-identified risk loci affect multiple traits that can even belong to distinct "trait domains". We searched for a correlation between the number of DAP-matching (IBD) e-genes and the number of trait domains pleiotropically affected by the corresponding risk loci as reported in Watanabe *et al.* [2019]. There was no obvious relation between these two statistics. Thus, at first glance, it does not seem that risk loci with high numbers of DAP-matching e-genes make a disproportionate contribution to pleiotropy.

Regulatory modules that are active in all cell types, make a disproportionate contribution to the risk loci with DAP-matching EAP, particularly in blood. We made a similar observation for IBD risk loci with the less-granular CEDAR1 data set (Momozawa *et al.*, 2018). This is possibly due to the fact that detecting the DAP-match is less dependent on analyzing the correct (i.e., disease-relevant) cell type for these risk loci.

Why are DAP-matching e-genes not showing an excess burden of coding variants in patients?

As discussed before, being a DAP-matching e-gene does not prove that it is causally involved in disease risk. There are two formal tests of gene causality. The first is the reciprocal hemizygosity test, which is very difficult to apply in mammals, let alone humans (Steinmetz *et al.*, 2002; Stern, 2014). The second is the enrichment of disruptive coding variants in cases. For example, this test was used to demonstrate the causality of *NOD2* in CD (Hugot *et al.*, 2001; Lesage *et al.*, 2002). Therefore, a logical next step after the identification of DAP-matching e-genes, is to sequence the corresponding genes in large case-control cohorts and to perform rare variant-based burden tests [f.i. Momozawa *et al.*, 2018]. As a matter of fact, this approach has now been applied genome-wide (i.e., without preselection of target genes using eQTL information) for several common complex diseases on very large cohorts including IBD (Sazonovs *et al.*, 2022). Although some new causal genes have been identified using this approach (including some that overlap with DAP-matching candidates, including from this study), the yield has been, arguably, somewhat disappointing given the magnitude of the effort. The same applies to most diseases for which this approach has been used [f.i. Flannick *et al.*, 2019]. We herein have used Sazonovs' resequencing data to look for the distribution of burden test *p*-values for 298 candidate e-genes out of our list of 556. There was no evidence from a departure from expectation under the null. Does that mean that none of our DAP-matching e-genes are affecting IBD risk? Although we cannot completely exclude that possibility, alternative explanations exist. One is a power issue: sequencing 35,000 cases is a lot but may not be sufficient, especially for small genes, and given the fact that predicting the effect of coding variants other than stop-gains remains difficult. The other is that coding variants in the corresponding genes do not cause IBD but possibly other diseases. A fundamental difference between coding and regulatory variants is that coding variants affect the function of the gene equally in all tissues where the gene is used, while regulatory variants likely affect the function of the gene in a restricted set of tissues. It is increasingly apparent that most genes are utilized in many tissues and cell types, as testified by the many synonyms existing for most gene names. As indicated by their denomination, diseases such as IBD are organ-restricted in their manifestations. It is very possible therefore that to be risk variants for IBD the effects have to be organ restricted (gut and immune cells), and that – for many genes - this does not apply to coding variants. The target space to resequence may therefore have to be redirected to (cell type-specific) *cis*-acting regulatory elements, yet these remain difficult to identify, the effects of variants on their functionality difficult to predict, and the corresponding genome space limited (hence affecting power).

Genetic support for new repurposing opportunities in IBD

We identify entrectinib as a promising repurposing candidate for CD. Entrectnib (ENB) is a small molecule that blocks several tyrosine kinases including oncogenic ones. It has been approved by the FDA in 2019 for the treatment of several solid tumors. It is administered orally and has proven safe and well tolerated even at high doses and prolonged administration. More recently, a screening of FDA-approved kinase inhibitors, showed that ENB specifically blocks the NLRP3 inflammasome. This was shown to result from the reversible binding of ENB to R121 of the NEK7, thereby inhibiting NLRP3 activation. Paradoxically, ENB does not affect NEK7’s kinase activity, which increases the effect’s specificity. *In vivo* tests show that shRNA-mediated downregulation of NEK7 protects mice against DSS-induced colitis (Chen *et al.*, 2019), while ENB was shown to protect mice against various NLRP3-dependent inflammatory conditions (Jin *et al.*, 2023). Thus, there is considerable prior functional and preclinical in vivo evidence supporting the use of ENB to treat IBD. In here, we show that CD risk variants increase NEK7 transcript levels in circulating naïve B cells, with very strong support for "colocalization" (θ_{IBD} =0.94). We also show that expression levels of *NEK7* are affected in multiple circulating immune populations of active CD patients, and – using the data from Kong *et al.,* [2023] – are increased in the colonic epithelium of IBD patients. In addition to supporting a contribution of *NEK7* expression levels in influencing predisposition to IBD, this supports NEK7’s role in the disease process *per se*, and hence a more likely curative effect of ENB. We note, however, that the IBD risk variants appear to decrease *NEK7* expression in some gut-resident cells of healthy individuals, while *NEK7* expression is decreased in some circulating immune populations, and these findings deserve further scrutiny.

We further identify at least two instances, amongst the targets with approved drugs, where IBD risk variants, rather counterintuitively, increase 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. 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 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.

Possibly the most important outcome from this work is the suggestion that the notion of polygenicity extends *within* risk loci, and that causal genes in risk loci may influence disease through their effects on multiple cell types. It supports the notion that the genetic determinism of CCD is "quasi-infinitesimal". This raises the question as to whether targeting individual components of this genetic architecture to treat the disease is the most effective strategy. As mentioned before, it seems unlikely that the effect of the many underpinning risk variants on disease are independent. They must perturb a series of pathways that progressively converge into a limited number of "highways", that ultimately determine disease outcome. Such "highways" would involve the "core genes" as defined in the omnigenic model (Boyle *et al.*, 2017). It should, in theory, be possible to genetically reconstruct the topology of the corresponding network. Indeed, just as the disease is associated with multiple risk loci (multiple DAP) "in *trans*", components of the upper part of the network (the "highways") are also expected to be associated with multiple risk loci as *trans*-QTL. On the other hand, components of the lower parts of the tree will be associated with fewer risk loci, and ultimately, i.e., at the bottom of the network, only with one (for at least some, as a *cis*-eQTL). 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 are 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.

SUPPLEMENTAL MATERIAL

Comments on candidate drugs for repurposing to treat IBD with genetic support from this study.

IL18R1 and IL18RAP. IL18R1 and IL18RAP code for the IL18 receptor and its accessory subunit (enhancing IL18 binding), respectively. With IL1, IL18 is activated upon PAMP-recognition by the inflammasomes. Our data indicates that downregulation (predominantly) of both genes (IL18R1 and IL18RAP), primarily (but not exclusively) in circulating immune cells (including ILC, MAIT, and Th17), increases the risk to develop CD (predominantly), hence that "activating" both might have a beneficial effect. In agreement with these findings, IL18R KO mice are more susceptible to DSS-induced colitis [Takagi *et al.*, 2003], and pre-treatment of mice with IL18 protected against DSS-induced colitis [Pu *et al.*, 2019]. However, IL18 administration at later stages of DSS-induced colitis exacerbates symptoms [Pu *et al.*, 2019], while administration of IL18 neutralizing antibodies reduces the severity of DSS-induced colitis [f.i. Ikegami *et al.*, 2024]. In humans, abundance of IL18 is generally increased in IBD patients, both in blood and mucosa [Leach *et al.*, 2008], while mendelian randomization indicates that increase in IL18 concentrations increases IBD risk [Mokry *et al.*, 2019]. We observe an increase of IL18R1 mRNA abundance in naïve B cells of active CD patients, while Eldjarn *et al.* [2023] report an increased abundance of IL18R1 in IBD patients. Thus, the IL18 axis appears to have an anti-inflammatory effect prior to disease declaration, yet a pro-inflammatory effect after. Iboctadekin is reported as an agonist of IL18R1 and IL18RAP, and is in fact recombinant human IL18. Iboctadekin has been tested in the context of infectious diseases and oncology, but with mediocre results, and its development was discontinued. Given the apparent dual function of the IL18 axis in inflammation, the use of iboctadekin in the treatment of IBD appears difficult to support. Nevertheless, an iboctadekin-based treatment that might be considered in the context of IBD, is to prevent relapse, particularly after disease remission has been achieved following surgery, rather than its use during the active phase of the condition.

CFTR. Well-studied chloride channel essential for ion transport and fluid secretion in tissues, including the gastrointestinal tract, and playing an important role in mucin maturation in the intestine [Gustafsson *et al.*, 2012]. Nullizygosity for *CFTR* causes cystic fibrosis (CF), while hemizygosity may increase resistance to infectious diseases [Bradbury, 1998]. The prevalence of IBD was shown to be 7 times higher in CF patients compared to controls [Lloyd-Still, 1994]. *A contrario*, CFTR hemizygosity was recently shown to protect against CD [Yu *et al.*, 2024]. Also, glibenclamide, a (CFTR) inhibitor, was reported to protect rats against 2,4-dinitrobenzene sulfonic acid (DNBS)-induced gastrointestinal inflammation [Chidrawar & Alsuwayt, 2021]. In agreement with this, we herein shown that increased CFTR expression in intestinal epithelial and stromal cells is associated with increased risk for UC. Note, however, that there is some evidence (albeit non-significant) from our data that decreased expression of CFTR in specific enterocyte populations may increase CD risk, deserving further analysis in a larger dataset. Also, plasma levels of CFTR were found to be increased in IBD patients. This suggests that downregulating CFTR may improve IBD. Of note, CFTR has been shown to act as a pathogen recognition molecule that can induce NF-kB activation [Schroeder *et al.*, 2002], and serves as epithelial receptor for *S. Typhi* transluminal migration [Pier *et al.*, 1998]. Also, CFTR KO mice secreted more intestinal mucus [Norkina *et al.*, 2004; Hodges *et al.*, 2008], which plays a protective role in IBD. While most drug development targeting CFTR aims at restoring function in the context of CF, there exists at least one presumed CFTR inhibitor, namely crofelemer, approved for HIV-associated diarrhea [Tradrantip *et al.*, 2010]. Crofelemer is a natural, (not so) small molecule derived from the *Croton lechleri* tree. It is taken orally to treat secretory diarrhea (as opposed to diarrhea resulting from an underlying inflammatory process) in various contexts. It has proven effective in HIV-seropositive patients on stable antiretroviral therapy [Macarthur *et al.*, 2013]. Being primarily used for non-inflammatory diarrheal conditions, crofelemer may alleviate diarrheal symptoms IBD patients but it is unclear whether it also affects the underlying inflammatory process. Of note, inhibition of CFTR-mediated intestinal chloride secretion could be a potential therapy for bile acid diarrhea, a condition frequently present in IBD [Duan *et al.*, 2019]. It seems important to determine whether the apparent effect of *CFTR* expression on IBD risk depends on CFTR's chloride channel activity (affected by crofelemer) or of distinct functions of this gene (such as pathogen recognition).

LAMA2. Codes for a subunit of laminins, an essential component of the basement membrane with important roles in tissue organization. LAMA2 deficiency causes a muscular dystrophy accompanied by an exacerbated innate immunity response, autophagy and cell death [Jeudi *et al.*, 2011]. Intriguingly, somatic LAMA2 mutations were recurrently observed in colonic mucosa of DSS-treated mice [He *et al.*, 2021]. LAMA2 may play a role in intestinal barrier homeostasis. Along those lines, mutations in laminins have been associated with skin blistering [Schéele *et al.*, 2007]. We observe that variants that increase *LAMA2* expression, including in precursor and mature enterocytes, consistently increase CD risk. However, *LAMA2* expression was found to be downregulated in fibroblasts of IBD inflamed terminal ileum. Ocriplasmin is an injectable (intravitreally) protease inhibitor acting on laminins (and fibronectin) that is used to treat vitreomacular adhesion [Stalman *et al.*, 2012; Mastrapasqua *et al.*, 2023]. Whether it is the down- or up-regulation, or any, that is desirable in this case is uncertain. There is a need for further exploration of the link between LAMA2 and IBD, which may also involve its role in autophagy [Mastrapasqua *et al.*, 2023].

CYP3A5. Codes for a member of the cytochrome P450 monooxygenases metabolizing various endo- and xenobiotics. In particular CYP3A enzymes participate in bile acid biotransformation, which – if perturbed – could contribute to the disruption of gut homeostasis in colitis patients including by affecting microbiota composition [Hayes *et al.*, 2016; Zhang *et al.*, 2019; Lin *et al.*, 2020]. We observe that an increase of CYP3A5 expression in naïve CD8 increases CD risk (tier 1).

However, in enterocytes of the small intestine, the opposite seems to occur (negative theta), albeit with FDR above 0.10 but potentially more relevant. There was no detectable effect of disease status on CYP3A5 expression/abundance in blood, biopsies and plasma. There exist at least two small molecule inhibitors of CYP3A5: cobicistat and ritonavir, used to boost the effectiveness of other drugs (particularly antiretroviral). Ritonavir may have independent anti-inflammatory action, as well as affect the microbiome but this appears unrelated to the observed DAP-EAP match. Genotype at CYP3A5 variants has been proposed as a guide to tacrolimus dosage in the treatment of ulcerative colitis [Okabayashi *et al.*, 2019; Yamamoto *et al.*, 2020], yet this bears no relationship with DAP-EAP matching.

PTGIR. Encodes the G protein coupled receptor of prostacyclin (PGI2). Prostacyclin is mainly known as a potent vasodilator (and inhibitor of platelet activation). Concomitantly, small molecule agonists of PTGIR (often prostacyclin mimics) are primarily used in the treatment of pulmonary arterial hypertension. However, prostaglandins, including prostacyclin, are known to affect inflammation, having either pro-inflammatory (f.i. rheumatoid arthritis) or anti-inflammatory (f.i. allergic inflammation) effects depending on context [Stitham *et al.*, 2011; Dorris & Peebles, 2012]. PTGIR has well-documented anti-inflammatory properties by inhibiting the production of pro-inflammatory cytokines and the immune response [Ricciotti *et al.*, 2011]. Paradoxically, stimulating PTGIR signaling in cultured human monocytes augments calprotectin expression, a marker of inflammation [Karaky *et al.*, 2022]. PTGIR may play a role in maintaining the intestinal epithelial barrier [Turner, 2009], and in regulating intestinal microcirculation which is impaired in IBD [f.i. Hatoum *et al.*, 2003], and tissue repair. Of note, increased expression of the PTGER4 receptor for PGE2 (known to affect the immune system) has been associated with an increased risk for CD [Libioulle *et al.*, 2007], yet PTGER4 knock-out mice are more susceptible to DDS-induced colitis, while PTGER4 agonists are protective [Kabashima *et al.*, 2002]. We observe that reduced expression of PTGIR in multiple circulating T lymphocyte populations as well as resident monocytes is consistently associated with increased risk to develop UC. We don't see a clear effect of disease status on PTGIR expression/levels whether in circulating leukocytes, biopsies or plasma. We note that the use of cyclooxygenase inhibitors reportedly exacerbates symptoms in some IBD patients. As for IL18R1/IL18RAP, PTGIR activator-based treatments that could be considered in the context of IBD, are to prevent relapse, particularly after disease remission has been achieved following surgery, rather than their use during the active phase of the condition.

INPP5D. Encoding inositol Polyphosphate-5-Phosphatase D (=SHIP1). SHIP acts as a multifunctional protein controlled by multiple regulatory inputs, and influences downstream signaling via both phosphatase-dependent and -independent means. SHIP1 has functions in B cells, T cells, NK cells, dendritic cells, mast cells, and macrophages [Pauls & Marshall, 2017]. Downregulation is generally considered to be pro-inflammatory. SHIP deficiency causes Crohn's disease-like ileitis in the mouse [Kerr *et al.*, 2011], while deficiency in Inflammatory Bowel Disease is associated with severe Crohn's disease and peripheral T cell reduction [Fernandes *et al.*, 2018]. Accordingly, we observe that *INPP5D* downregulation in circulating T lymphocytes, MAIT and NK cells, as well as resident monocytes increases CD risk. Also, INPP5D expression is decreased in circulating monocytes from IBD patients. The small molecule rosiptor, the most developed and characterized allosteric activator of INPP5D, has been unsuccessfully tested as anti-inflammatory drug to treat interstitial cystitis and bladder pain syndrome and its development has been discontinued [https://www.pharmaceutical-technology.com/news/aquinox-halts-development-drug-bladder-pain-syndrome/]. Binding of rosiptor to SHIP1 was shown to be weak when compared to other agonists which may in part explain poor clinical results [Chamberlain *et al.*, 2020; Pedicone *et al.*, 2021]. We conclude that rosiptor is probably not an optimal candidate for repurposing.

NEK7. NIMA Related Kinase 7, plays a central role in assembly of the NLRP3 inflammasome [Shi *et al.*, 2016]. Missense *NLRP3* mutation cause dominant congenital multisystem inflammatory diseases (Cinca (MIM 607115) and Muckle-Wells syndromes (MIM 191900)). Downregulating NEK7 by intraperitoneal injection of lentiviruses expression anti-NEK7 shRNAs attenuates DSS-induced colitis [Chen *et al.*, 2019]. We observe that variants that increase NEK7 expression in circulating naïve B cells increase CD risk. Of note, however, these same variants appear to decrease NEK7 expression in gut-resident lymphoid and myeloid immune cells, albeit with FDR > 0.10, hence inversing the association with CD (decreased expression increases risk). Disease affects NEK7 expression but the sign of the affect depends on cell type: expression is increased in enterocytes from non-inflamed colon of patients when compared to controls, while expression is increased in some and decreased in other circulating immune cells of patients. 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 tumors [Liu *et al.*, 2018]. It was recently shown to bind to NEK7 R121, and interferes with its interaction with NLRP3. In mouse models, ENB effectively reduced symptoms of NLRP3 inflammasome-related diseases (other than colitis), indicating its potential as a therapeutic candidate [Jin *et al.*, 2023]. ENB has a high safety profile and is well tolerated by almost all patients without cumulative toxicity [Jiang *et al.*, 2022]. Inhibition of the inflammasome has shown beneficial effects in various inflammatory pathologies. ENB appears to be a repurposing candidate for IBD.

LIST OF PUBLICATIONS

Related to the thesis:

- Perée, H., Petrov, V.A., Tokunaga, Y., Kvasz, A., Vieujean, S., Regimont, S., Mni, M., Wéry, M., Azarzar, S., Jacques, S., Fouillien, N., Karim, L., Deckers, M., Detry, E., Mayer, A., Stephan, R., Harshman, K., Mizoro, Y., Reenaers, C., Van Kemseke, C., Warling, O., Labille, V., Kropp, S., Poncin, M., Moreau, A.C., Servais, B., Joly, J.-P., SYSCID Consortium, BRIDGE Consortium, Coppieters, W., Dermitzakis, E., Louis, E., Georges, M., Takeda, H., Rahmouni, S., 2024. Integrated cell type-specific analysis of blood and gut identifies matching eQTL for 140 IBD risk loci and entrectinib as possible repurposing candidate. https://doi.org/10.1101/2024.10.14.24315443 (under revision in Nature Communications)

Not related to the thesis:

- COVID-19 Host Genetics Initiative, *et al.*, 2021. Mapping the human genetic architecture of COVID-19. Nature 600, 472–477. https://doi.org/10.1038/s41586-021-03767-x
- Desmecht, S., Tashkeev, A., El Moussaoui, M., Marechal, N., Perée, H., Tokunaga, Y., *et al.*, 2022. Kinetics and Persistence of the Cellular and Humoral Immune Responses to BNT162b2 mRNA Vaccine in SARS-CoV-2-Naive and -Experienced Subjects: Impact of Booster Dose and Breakthrough Infections. Front. Immunol. 13, 863554. https://doi.org/10.3389/fimmu.2022.863554
- Jacques, S., Arjomand, A., Perée, H., Collins, P., Mayer, A., Lavergne, A., *et al.*, 2021. Dual-specificity phosphatase 3 deletion promotes obesity, non-alcoholic steatohepatitis and hepatocellular carcinoma. Sci Rep 11, 5817. https://doi.org/10.1038/s41598-021-85089-6
- Liefferinckx, C., De Grève, Z., Toubeau, J.-F., Perée, H., Quertinmont, E., Tafciu, V., *et al.*, 2021. New approach to determine the healthy immune variations by combining clustering methods. Sci Rep 11, 8917. https://doi.org/10.1038/s41598-021-88272-x
- Liefferinckx, C., Stern, D., Perée, H., Bottieau, J., Mayer, A., Dubussy, C., *et al.*, 2025. The identification of blood-derived response eQTLs reveals complex effects of regulatory variants on inflammatory and infectious disease risk. PLoS Genet 21, e1011599. https://doi.org/10.1371/journal.pgen.1011599

REFERENCES

1. Abdellaoui, A., Yengo, L., Verweij, K.J.H., Visscher, P.M., 2023. 15 years of GWAS discovery: Realizing the promise. The American Journal of Human Genetics 110, 179–194. https://doi.org/10.1016/j.ajhg.2022.12.011

2. Aguirre-Gamboa, R., De Klein, N., Di Tommaso, J., Claringbould, A., Van Der Wijst, M.G., De Vries, D., *et al.*, 2020. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. BMC Bioinformatics 21, 243. https://doi.org/10.1186/s12859-020-03576-5

3. Alatab, S., Sepanlou, S.G., Ikuta, K., Vahedi, H., Bisignano, C., Safiri, S., *et al.*, 2020. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Gastroenterology & Hepatology 5, 17–30. https://doi.org/10.1016/S2468-1253(19)30333-4

4. Al-Rashed, F., Ahmad, Z., Thomas, R., Melhem, M., Snider, A.J., Obeid, L.M., *et al.*, 2020. Neutral sphingomyelinase 2 regulates inflammatory responses in monocytes/macrophages induced by TNF-α. Sci Rep 10, 16802. https://doi.org/10.1038/s41598-020-73912-5

5. Anders, S., Huber, W., 2010. Differential expression analysis for sequence count data. Genome Biol 11, R106. https://doi.org/10.1186/gb-2010-11-10-r106

6. Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P., Zondervan, K.T., 2010. Data quality control in genetic case-control association studies. Nat Protoc 5, 1564–1573. https://doi.org/10.1038/nprot.2010.116

7. Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., *et al.*, 2011. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43, 246–252. https://doi.org/10.1038/ng.764

8. Arora, U., Kedia, S., Ahuja, V., 2024. The practice of fecal microbiota transplantation in inflammatory bowel disease. Intest Res 22, 44–64. https://doi.org/10.5217/ir.2023.00085

9. Asiimwe, R., Dobin, A., 2024. STAR+WASP reduces reference bias in the allele-specific mapping of RNA-seq reads. https://doi.org/10.1101/2024.01.21.576391

10. Bai, Z., Hao, J., Chen, M., Yao, K., Zheng, L., Liu, L., *et al.*, 2024. Integrating plasma proteomics with genome-wide association data to identify novel drug targets for inflammatory bowel disease. Sci Rep 14, 16251. https://doi.org/10.1038/s41598-024-66780-w

11. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., *et al.*, 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40, 955–962. https://doi.org/10.1038/ng.175

12. Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using **lme4**. J. Stat. Soft. 67. https://doi.org/10.18637/jss.v067.i01

13. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., *et al.*, 2014. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24, 14–24. https://doi.org/10.1101/gr.155192.113

14. Beasley, T.M., Erickson, S., Allison, D.B., 2009. Rank-Based Inverse Normal

Transformations are Increasingly Used, But are They Merited? Behav Genet 39, 580–595. https://doi.org/10.1007/s10519-009-9281-0

15. Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., Lancet, D., 2015. PathCards: multi-source consolidation of human biological pathways. Database 2015. https://doi.org/10.1093/database/bav006

16. Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B: Statistical Methodology 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

17. Bonaz, B.L., Bernstein, C.N., 2013. Brain-Gut Interactions in Inflammatory Bowel Disease. Gastroenterology 144, 36–49. https://doi.org/10.1053/j.gastro.2012.10.003

18. Bosa, L., Batura, V., Colavito, D., Fiedler, K., Gaio, P., Guo, C., *et al.*, 2021. Novel CARMIL2 loss-of-function variants are associated with pediatric inflammatory bowel disease. Sci Rep 11, 5945. https://doi.org/10.1038/s41598-021-85399-9

19. Boyle, E.A., Li, Y.I., Pritchard, J.K., 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169, 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

20. Bradbury, J., 1998. Why are cystic-fibrosis mutations so common? The Lancet 351, 1409. https://doi.org/10.1016/S0140-6736(98)23019-4

21. Brown, E.M., Kenny, D.J., Xavier, R.J., 2019. Gut Microbiota Regulation of T Cells During Inflammation and Autoimmunity. Annu. Rev. Immunol. 37, 599–624. https://doi.org/10.1146/annurev-immunol-042718-041841

22. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., *et al.*, 2015. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47, 291–295. https://doi.org/10.1038/ng.3211

23. Burclaff, J., Bliton, R.J., Breau, K.A., Ok, M.T., Gomez-Martinez, I., Ranek, J.S., *et al.*, 2022. A Proximal-to-Distal Survey of Healthy Adult Human Small Intestine and Colon Epithelium by Single-Cell Transcriptomics. Cellular and Molecular Gastroenterology and Hepatology 13, 1554–1589. https://doi.org/10.1016/j.jcmgh.2022.02.007

24. Burgess, S., Mason, A.M., Grant, A.J., Slob, E.A.W., Gkatzionis, A., Zuber, V., *et al.*, 2023. Using genetic association data to guide drug discovery and development: Review of methods and applications. The American Journal of Human Genetics 110, 195–214. https://doi.org/10.1016/j.ajhg.2022.12.017

25. Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., *et al.*, 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678. https://doi.org/10.1038/nature05911

26. Cadwell, K., Patel, K.K., Maloney, N.S., Liu, T.-C., Ng, A.C.Y., Storer, C.E., *et al.*, 2010. Virus-Plus-Susceptibility Gene Interaction Determines Crohn's Disease Gene Atg16L1 Phenotypes in Intestine. Cell 141, 1135–1145. https://doi.org/10.1016/j.cell.2010.05.009

27. Cai, Y., Jia, X., Xu, L., Chen, H., Xie, S., Cai, J., 2023. Interleukin-17 and

inflammatory bowel disease: a 2-sample Mendelian randomization study. Front. Immunol. 14, 1238457. https://doi.org/10.3389/fimmu.2023.1238457

28. Carter, R., Drouin, G., 2009. Structural differentiation of the three eukaryotic RNA polymerases. Genomics 94, 388–396. https://doi.org/10.1016/j.ygeno.2009.08.011

29. Castel, S.E., Aguet, F., Mohammadi, P., GTEx Consortium, Aguet, F., Anand, S., *et al.*, 2020. A vast resource of allelic expression data spanning human tissues. Genome Biol 21, 234. https://doi.org/10.1186/s13059-020-02122-z

30. Chamberlain, T.C., Cheung, S.T., Yoon, J.S.J., Ming-Lum, A., Gardill, B.R., Shakibakho, S., *et al.*, 2020. Interleukin-10 and Small Molecule SHIP1 Allosteric Regulators Trigger Anti-inflammatory Effects through SHIP1/STAT3 Complexes. iScience 23, 101433. https://doi.org/10.1016/j.isci.2020.101433

31. Chen, G.-B., Lee, S.H., Brion, M.-J.A., Montgomery, G.W., Wray, N.R., Radford-Smith, G.L., *et al.*, 2014. Estimation and partitioning of (co)heritability of inflammatory bowel disease from GWAS and immunochip data. Human Molecular Genetics 23, 4710–4720. https://doi.org/10.1093/hmg/ddu174

32. Chen, X., Liu, G., Yuan, Y., Wu, G., Wang, S., Yuan, L., 2019. NEK7 interacts with NLRP3 to modulate the pyroptosis in inflammatory bowel disease via NF-κB signaling. Cell Death Dis 10, 906. https://doi.org/10.1038/s41419-019-2157-1

33. Chidrawar, V., Alsuwayt, B., 2021. Defining the role of CFTR channel blocker and ClC-2 activator in DNBS induced gastrointestinal inflammation. Saudi Pharmaceutical Journal 29, 291–304. https://doi.org/10.1016/j.jsps.2021.02.005

34. Cholapranee, A., Ananthakrishnan, A.N., 2016. Environmental Hygiene and Risk of Inflammatory Bowel Diseases: A Systematic Review and Meta-analysis. Inflammatory Bowel Diseases 22, 2191–2199. https://doi.org/10.1097/MIB.0000000000000852

35. Cleynen, I., Boucher, G., Jostins, L., Schumm, L.P., Zeissig, S., Ahmad, T., *et al.*, 2016. Inherited determinants of Crohn’s disease and ulcerative colitis phenotypes: a genetic association study. The Lancet 387, 156–167. https://doi.org/10.1016/S0140-6736(15)00465-1

36. Connally, N.J., Nazeen, S., Lee, D., Shi, H., Stamatoyannopoulos, J., Chun, S., *et al.*, 2022. The missing link between genetic association and regulatory function. eLife 11, e74970. https://doi.org/10.7554/eLife.74970

37. Coskun, M., 2014. Intestinal Epithelium in Inflammatory Bowel Disease. Front. Med. 1. https://doi.org/10.3389/fmed.2014.00024

38. Dai, Z., Xu, W., Ding, R., Peng, X., Shen, X., Song, J., *et al.*, 2023. Two-sample Mendelian randomization analysis evaluates causal associations between inflammatory bowel disease and osteoporosis. Front. Public Health 11, 1151837. https://doi.org/10.3389/fpubh.2023.1151837

39. De Lange, K.M., Moutsianas, L., Lee, J.C., Lamb, C.A., Luo, Y., Kennedy, N.A., *et al.*, 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat Genet 49, 256–261. https://doi.org/10.1038/ng.3760

40. Delaneau, O., Ongen, H., Brown, A.A., Fort, A., Panousis, N.I., Dermitzakis, E.T., 2017. A complete tool set for molecular QTL discovery and analysis. Nat Commun 8,

15452. https://doi.org/10.1038/ncomms15452

41. Diez-Martin, E., Hernandez-Suarez, L., Muñoz-Villafranca, C., Martin-Souto, L., Astigarraga, E., Ramirez-Garcia, A., Barreda-Gómez, G., 2024. Inflammatory Bowel Disease: A Comprehensive Analysis of Molecular Bases, Predictive Biomarkers, Diagnostic Methods, and Therapeutic Options. IJMS 25, 7062. https://doi.org/10.3390/ijms25137062

42. Díez-Obrero, V., Moratalla-Navarro, F., Ibáñez-Sanz, G., Guardiola, J., Rodríguez-Moranta, F., Obón-Santacana, M., *et al.*, 2022. Transcriptome-Wide Association Study for Inflammatory Bowel Disease Reveals Novel Candidate Susceptibility Genes in Specific Colon Subsites and Tissue Categories. Journal of Crohn's and Colitis 16, 275–285. https://doi.org/10.1093/ecco-jcc/jjab131

43. Dimitrov, G., Bamberger, S., Navard, C., Dreux, S., Badens, C., Bourgeois, P., *et al.*, 2019. Congenital Sodium Diarrhea by mutation of the SLC9A3 gene. European Journal of Medical Genetics 62, 103712. https://doi.org/10.1016/j.ejmg.2019.103712

44. Dobin, A., Gingeras, T.R., 2015. Mapping RNA‐seq Reads with STAR. CP in Bioinformatics 51. https://doi.org/10.1002/0471250953.bi1114s51

45. Dorris, S.L., Peebles, R.S., 2012. PGI_2 as a Regulator of Inflammatory Diseases. Mediators of Inflammation 2012, 1–9. https://doi.org/10.1155/2012/926968

46. Duan, T., Cil, O., Tse, C.M., Sarker, R., Lin, R., Donowitz, M., Verkman, A.S., 2019. Inhibition of CFTR‐mediated intestinal chloride secretion as potential therapy for bile acid diarrhea. FASEB j. 33, 10924–10934. https://doi.org/10.1096/fj.201901166R

47. Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., *et al.*, 2006. A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene. Science 314, 1461–1463. https://doi.org/10.1126/science.1135245

48. Ek, W.E., D'Amato, M., Halfvarson, J., 2014. The history of genetics in inflammatory bowel disease. Ann Gastroenterol 27, 294–303.

49. El Hadad, J., Schreiner, P., Vavricka, S.R., Greuter, T., 2024. The Genetics of Inflammatory Bowel Disease. Mol Diagn Ther 28, 27–35. https://doi.org/10.1007/s40291-023-00678-7

50. Eldjarn, G.H., Ferkingstad, E., Lund, S.H., Helgason, H., Magnusson, O.Th., Gunnarsdottir, K., *et al.*, 2023. Large-scale plasma proteomics comparisons through genetics and disease associations. Nature 622, 348–358. https://doi.org/10.1038/s41586-023-06563-x

51. Elgart, M., Lyons, G., Romero-Brufau, S., Kurniansyah, N., Brody, J.A., Guo, X., *et al.*, 2022. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. Commun Biol 5, 856. https://doi.org/10.1038/s42003-022-03812-z

52. Ellinghaus, D., Jostins, L., Spain, S.L., Cortes, A., Bethune, J., Han, B., *et al.*, 2016. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. Nat Genet 48, 510–518. https://doi.org/10.1038/ng.3528

53. Everhart, J.E., Ruhl, C.E., 2009. Burden of Digestive Diseases in the United States Part I: Overall and Upper Gastrointestinal Diseases. Gastroenterology 136, 376–386. https://doi.org/10.1053/j.gastro.2008.12.015

54. Fadista, J., Manning, A.K., Florez, J.C., Groop, L., 2016. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Eur J Hum Genet 24, 1202–1205. https://doi.org/10.1038/ejhg.2015.269

55. Fernandes, S., Srivastava, N., Sudan, R., Middleton, F.A., Shergill, A.K., Ryan, J.C., Kerr, W.G., 2018. SHIP1 Deficiency in Inflammatory Bowel Disease Is Associated With Severe Crohn's Disease and Peripheral T Cell Reduction. Front. Immunol. 9, 1100. https://doi.org/10.3389/fimmu.2018.01100

56. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., *et al.*, 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 16, 278. https://doi.org/10.1186/s13059-015-0844-5

57. Fisher, R.A., 1919. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. R. Soc. Edinb. 52, 399–433. https://doi.org/10.1017/S0080456800012163

58. Flannick, J., Mercader, J.M., Fuchsberger, C., Udler, M.S., Mahajan, A., Wessel, J., *et al.*, 2019. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature 570, 71–76. https://doi.org/10.1038/s41586-019-1231-2

59. Fort, A., Panousis, N.I., Garieri, M., Antonarakis, S.E., Lappalainen, T., Dermitzakis, E.T., Delaneau, O., 2017. *MBV* : a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. Bioinformatics 33, 1895–1897. https://doi.org/10.1093/bioinformatics/btx074

60. Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., *et al.*, 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet 42, 1118–1125. https://doi.org/10.1038/ng.717

61. Franzén, O., Gan, L.-M., Björkegren, J.L.M., 2019. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database 2019. https://doi.org/10.1093/database/baz046

62. Freuer, D., Linseisen, J., Meisinger, C., 2022. Association Between Inflammatory Bowel Disease and Both Psoriasis and Psoriatic Arthritis: A Bidirectional 2-Sample Mendelian Randomization Study. JAMA Dermatol 158, 1262. https://doi.org/10.1001/jamadermatol.2022.3682

63. Frolkis, A.D., Dykeman, J., Negrón, M.E., deBruyn, J., Jette, N., Fiest, K.M., *et al.*, 2013. Risk of Surgery for Inflammatory Bowel Diseases Has Decreased Over Time: A Systematic Review and Meta-analysis of Population-Based Studies. Gastroenterology 145, 996–1006. https://doi.org/10.1053/j.gastro.2013.07.041

64. Galili, T., 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics 31, 3718–3720. https://doi.org/10.1093/bioinformatics/btv428

65. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J., *et al.*, 2015. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47, 1091–1098. https://doi.org/10.1038/ng.3367

66. Georgiou, A.N., Ntritsos, G., Papadimitriou, N., Dimou, N., Evangelou, E., 2021. Cigarette Smoking, Coffee Consumption, Alcohol Intake, and Risk of Crohn's

Disease and Ulcerative Colitis: A Mendelian Randomization Study. Inflammatory Bowel Diseases 27, 162–168. https://doi.org/10.1093/ibd/izaa152

67. Geremia, A., Biancheri, P., Allan, P., Corazza, G.R., Di Sabatino, A., 2014. Innate and adaptive immunity in inflammatory bowel disease. Autoimmunity Reviews 13, 3–10. https://doi.org/10.1016/j.autrev.2013.06.004

68. Ghilas, S., O'Keefe, R., Mielke, L.A., Raghu, D., Buchert, M., Ernst, M., 2022. Crosstalk between epithelium, myeloid and innate lymphoid cells during gut homeostasis and disease. Front. Immunol. 13, 944982. https://doi.org/10.3389/fimmu.2022.944982

69. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., Plagnol, V., 2014. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet 10, e1004383. https://doi.org/10.1371/journal.pgen.1004383

70. Glassner, K.L., Abraham, B.P., Quigley, E.M.M., 2020. The microbiome and inflammatory bowel disease. Journal of Allergy and Clinical Immunology 145, 16–27. https://doi.org/10.1016/j.jaci.2019.11.003

71. González-Serna, D., Ochoa, E., López-Isac, E., Julià, A., Degenhardt, F., Ortego-Centeno, N., *et al.*, 2020. A cross-disease meta-GWAS identifies four new susceptibility loci shared between systemic sclerosis and Crohn's disease. Sci Rep 10, 1862. https://doi.org/10.1038/s41598-020-58741-w

72. Gordon, H., Biancone, L., Fiorino, G., Katsanos, K.H., Kopylov, U., Al Sulais, E., *et al.*, 2023. ECCO Guidelines on Inflammatory Bowel Disease and Malignancies. Journal of Crohn's and Colitis 17, 827–854. https://doi.org/10.1093/ecco-jcc/jjac187

73. Graham, D.B., Xavier, R.J., 2020. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature 578, 527–539. https://doi.org/10.1038/s41586-020-2025-2

74. Grajo, J.R., Huang, C., Dillman, J.R., Gee, M.S., Jaffe, T.A., Soto, J.A., Baker, M.E., 2021. MR Enterography of Complicated Crohn Disease: Stricturing and Penetrating Disease. Topics in Magnetic Resonance Imaging 30, 23–30. https://doi.org/10.1097/RMR.0000000000000266

75. GTEx Consortium, 2017. Genetic effects on gene expression across human tissues. Nature 550, 204–213. https://doi.org/10.1038/nature24277

76. Gustafsson, J.K., Ermund, A., Ambort, D., Johansson, M.E.V., Nilsson, H.E., Thorell, K., *et al.*, 2012. Bicarbonate and functional CFTR channel are required for proper mucin secretion and link cystic fibrosis with its mucus phenotype. Journal of Experimental Medicine 209, 1263–1272. https://doi.org/10.1084/jem.20120562

77. Hall, A.B., Tolonen, A.C., Xavier, R.J., 2017. Human genetic variation and the gut microbiome in disease. Nat Rev Genet 18, 690–699. https://doi.org/10.1038/nrg.2017.63

78. Hallows, K.R., Kobinger, G.P., Wilson, J.M., Witters, L.A., Foskett, J.K., 2003. Physiological modulation of CFTR activity by AMP-activated protein kinase in polarized T84 cells. American Journal of Physiology-Cell Physiology 284, C1297–C1308. https://doi.org/10.1152/ajpcell.00227.2002

79. Hamaoui, D., Subtil, A., 2022. ATG16L1 functions in cell homeostasis beyond

autophagy. The FEBS Journal 289, 1779–1800. https://doi.org/10.1111/febs.15833

80. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., *et al.*, 2021. Integrated analysis of multimodal single-cell data. Cell 184, 3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048

81. Hatoum, O.A., Miura, H., Binion, D.G., 2003. The vascular contribution in the pathogenesis of inflammatory bowel disease. American Journal of Physiology-Heart and Circulatory Physiology 285, H1791–H1796. https://doi.org/10.1152/ajpheart.00552.2003

82. Hayes, M.A., Li, X.-Q., Grönberg, G., Diczfalusy, U., Andersson, T.B., 2016. CYP3A Specifically Catalyzes 1β-Hydroxylation of Deoxycholic Acid: Characterization and Enzymatic Synthesis of a Potential Novel Urinary Biomarker for CYP3A Activity. Drug Metabolism and Disposition 44, 1480–1489. https://doi.org/10.1124/dmd.116.070805

83. He, X., Fuller, C.K., Song, Y., Meng, Q., Zhang, B., Yang, X., Li, H., 2013. Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. The American Journal of Human Genetics 92, 667–680. https://doi.org/10.1016/j.ajhg.2013.03.022

84. He, J., Han, J., Liu, J., Yang, R., Wang, J., Wang, X., Chen, X., 2021. Genetic and Epigenetic Impact of Chronic Inflammation on Colon Mucosa Cells. Front. Genet. 12, 722835. https://doi.org/10.3389/fgene.2021.722835

85. Hedin, C.R.H., Vavricka, S.R., Stagg, A.J., Schoepfer, A., Raine, T., Puig, L., *et al.*, 2019. The Pathogenesis of Extraintestinal Manifestations: Implications for IBD Research, Diagnosis, and Therapy. Journal of Crohn’s and Colitis 13, 541–554. https://doi.org/10.1093/ecco-jcc/jjy191

86. Hickey, J.W., Becker, W.R., Nevins, S.A., Horning, A., Perez, A.E., Zhu, C., *et al.*, 2023. Organization of the human intestine at single-cell resolution. Nature 619, 572–584. https://doi.org/10.1038/s41586-023-05915-x

87. Hirten, R.P., Shah, S., Sachar, D.B., Colombel, J.-F., 2018. The Management of Intestinal Penetrating Crohn’s Disease. Inflammatory Bowel Diseases 24, 752–765. https://doi.org/10.1093/ibd/izx108

88. Hodges, C.A., Cotton, C.U., Palmert, M.R., Drumm, M.L., 2008. Generation of a conditional null allele for *Cftr* in mice. Genesis 46, 546–552. https://doi.org/10.1002/dvg.20433

89. Hoffmann, T.J., Witte, J.S., 2015. Strategies for Imputing and Analyzing Rare Variants in Association Studies. Trends in Genetics 31, 556–563. https://doi.org/10.1016/j.tig.2015.07.006

90. Honap, S., Johnston, E., Agrawal, G., Al-Hakim, B., Hermon-Taylor, J., Sanderson, J., 2021. Anti- *Mycobacterium paratuberculosis* (MAP) therapy for Crohn’s disease: an overview and update. Frontline Gastroenterol 12, 397–403. https://doi.org/10.1136/flgastro-2020-101471

91. Hong, S.M., Baek, D.H., 2024. Diagnostic Procedures for Inflammatory Bowel Disease: Laboratory, Endoscopy, Pathology, Imaging, and Beyond. Diagnostics 14, 1384. https://doi.org/10.3390/diagnostics14131384

92. Hormozdiari, F., van de Bunt, M., Segrè, A.V., Li, X., Joo, J.W.J., Bilow, M., *et al.*, 2016. Colocalization of GWAS and eQTL Signals Detects Target Genes. The

American Journal of Human Genetics 99, 1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003

93. Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C.A., *et al.*, 2017. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature 547, 173–178. https://doi.org/10.1038/nature22969

94. Hugot, J.-P., Laurent-Puig, P., Gower-Rousseau, C., Olson, J.M., Lee, J.C., Beaugerie, L., *et al.*, 1996. Mapping of a susceptibility locus for Crohn's disease on chromosome 16. Nature 379, 821–823. https://doi.org/10.1038/379821a0

95. Hugot, J.-P., Chamaillard, M., Zouali, H., Lesage, S., Cézard, J.-P., Belaiche, J., *et al.*, 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411, 599–603. https://doi.org/10.1038/35079107

96. Ikegami, S., Maeda, K., Urano, T., Mu, J., Nakamura, M., Yamamura, T., *et al.*, 2024. Monoclonal Antibody Against Mature Interleukin-18 Ameliorates Colitis in Mice and Improves Epithelial Barrier Function. Inflammatory Bowel Diseases 30, 1353–1366. https://doi.org/10.1093/ibd/izad292

97. Imielinski, M., Baldassano, R.N., Griffiths, A., Russell, R.K., Annese, V., Dubinsky, M., *et al.*, 2009. Common variants at five new loci associated with early-onset inflammatory bowel disease. Nat Genet 41, 1335–1340. https://doi.org/10.1038/ng.489

98. Ishikawa, K., Sugimoto, S., Oda, M., Fujii, M., Takahashi, S., Ohta, Y., *et al.*, 2022. Identification of Quiescent LGR5+ Stem Cells in the Human Colon. Gastroenterology 163, 1391-1406.e24. https://doi.org/10.1053/j.gastro.2022.07.081

99. Jans, D., Cleynen, I., 2023. The genetics of non-monogenic IBD. Hum Genet 142, 669–682. https://doi.org/10.1007/s00439-023-02521-9

100. Jiang, Q., Li, M., Li, H., Chen, L., 2022. Entrectinib, a new multi-target inhibitor for cancer therapy. Biomedicine & Pharmacotherapy 150, 112974. https://doi.org/10.1016/j.biopha.2022.112974

101. Jin, X., Liu, D., Zhou, X., Luo, X., Huang, Q., Huang, Y., 2023. Entrectinib inhibits NLRP3 inflammasome and inflammatory diseases by directly targeting NEK7. Cell Reports Medicine 4, 101310. https://doi.org/10.1016/j.xcrm.2023.101310

102. Jones, D.P., Richardson, T.G., Davey Smith, G., Gunnell, D., Munafò, M.R., Wootton, R.E., 2020. Exploring the Effects of Cigarette Smoking on Inflammatory Bowel Disease Using Mendelian Randomization. Crohn's & Colitis 360 2, otaa018. https://doi.org/10.1093/crocol/otaa018

103. Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., *et al.*, 2012. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491, 119–124. https://doi.org/10.1038/nature11582

104. Julià, A., Domènech, E., Chaparro, M., García-Sánchez, V., Gomollón, F., Panés, J., *et al.*, 2014. A genome-wide association study identifies a novel locus at 6q22.1 associated with ulcerative colitis. Human Molecular Genetics 23, 6927–6934. https://doi.org/10.1093/hmg/ddu398

105. Juyal, G., Negi, S., Sood, A., Gupta, A., Prasad, P., Senapati, S., *et al.*, 2015. Genome-wide association scan in north Indians reveals three novel

HLA-independent risk loci for ulcerative colitis. Gut 64, 571–579. https://doi.org/10.1136/gutjnl-2013-306625

106. Kabashima, K., Saji, T., Murata, T., Nagamachi, M., Matsuoka, T., Segi, E., *et al.*, 2002. The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. J. Clin. Invest. 109, 883–893. https://doi.org/10.1172/JCI0214459

107. Kaplan, G.G., Windsor, J.W., 2021. The four epidemiological stages in the global evolution of inflammatory bowel disease. Nat Rev Gastroenterol Hepatol 18, 56–66. https://doi.org/10.1038/s41575-020-00360-x

108. Karaky, M., Boucher, G., Mola, S., Foisy, S., Beauchamp, C., Rivard, M.-E., *et al.*, 2022. Prostaglandins and calprotectin are genetically and functionally linked to the Inflammatory Bowel Diseases. PLoS Genet 18, e1010189. https://doi.org/10.1371/journal.pgen.1010189

109. Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J.A.C., Baurain, D., *et al.*, 2011. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nat Genet 43, 405–413. https://doi.org/10.1038/ng.814

110. Kerr, W.G., Park, M.-Y., Maubert, M., Engelman, R.W., 2011. SHIP deficiency causes Crohn’s disease-like ileitis. Gut 60, 177–188. https://doi.org/10.1136/gut.2009.202283

111. Khor, B., Gardet, A., Xavier, R.J., 2011. Genetics and pathogenesis of inflammatory bowel disease. Nature 474, 307–317. https://doi.org/10.1038/nature10209

112. King, E.A., Davis, J.W., Degner, J.F., 2019. 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. https://doi.org/10.1371/journal.pgen.1008489

113. Ko, B.S., Lee, S.B., Kim, T.-K., 2024. A brief guide to analyzing expression quantitative trait loci. Molecules and Cells 47, 100139. https://doi.org/10.1016/j.mocell.2024.100139

114. Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R., Papantonis, A., 2012. Enhancers and silencers: an integrated and simple model for their function. Epigenetics & Chromatin 5, 1. https://doi.org/10.1186/1756-8935-5-1

115. Kong, L., Pokatayev, V., Lefkovith, A., Carter, G.T., Creasey, E.A., Krishna, C., *et al.*, 2023. 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. https://doi.org/10.1016/j.immuni.2023.01.002

116. Kong, L., Chen, S., Huang, S., Zheng, A., Gao, S., Ye, J., Hua, C., 2024. Challenges and opportunities in inflammatory bowel disease: from current therapeutic strategies to organoid-based models. Inflamm. Res. 73, 541–562. https://doi.org/10.1007/s00011-024-01854-z

117. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., *et al.*, 2019. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

118. Kostic, A.D., Xavier, R.J., Gevers, D., 2014. The Microbiome in Inflammatory

Bowel Disease: Current Status and the Future Ahead. Gastroenterology 146, 1489–1499. https://doi.org/10.1053/j.gastro.2014.02.009

119. Krzak, M., Alegbe, T., Taylor, D.L., Jones, G.-R., Ghouraba, M., Strickland, M., *et al.*, 2023. Single-Cell RNA Sequencing of Terminal Ileal Biopsies Identifies Signatures of Crohn's Disease Pathogenesis. https://doi.org/10.1101/2023.09.06.23295056

120. Kucharzik, T., Ellul, P., Greuter, T., Rahier, J.F., Verstockt, B., Abreu, C., *et al.*, 2021. ECCO Guidelines on the Prevention, Diagnosis, and Management of Infections in Inflammatory Bowel Disease. Journal of Crohn's and Colitis 15, 879–913. https://doi.org/10.1093/ecco-jcc/jjab052

121. Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. **lmerTest** Package: Tests in Linear Mixed Effects Models. J. Stat. Soft. 82. https://doi.org/10.18637/jss.v082.i13

122. Leach, S.T., Messina, I., Lemberg, D.A., Novick, D., Rubenstein, M., Day, A.S., 2008. Local and systemic interleukin-18 and interleukin-18-binding protein in children with inflammatory bowel disease: Inflammatory Bowel Diseases 14, 68–74. https://doi.org/10.1002/ibd.20272

123. Leek, J.T., Storey, J.D., 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet 3, e161. https://doi.org/10.1371/journal.pgen.0030161

124. Lesage, S., Zouali, H., Cézard, J.-P., Colombel, J.-F., Belaiche, J., Almer, S., *et al.*, 2002. CARD15/NOD2 Mutational Analysis and Genotype-Phenotype Correlation in 612 Patients with Inflammatory Bowel Disease. The American Journal of Human Genetics 70, 845–857. https://doi.org/10.1086/339432

125. Li, S., Wang, H., Peng, B., Zhang, M., Zhang, D., Hou, S., *et al.*, 2009. Efalizumab binding to the LFA-1 α_L I domain blocks ICAM-1 binding via steric hindrance. Proc. Natl. Acad. Sci. U.S.A. 106, 4349–4354. https://doi.org/10.1073/pnas.0810844106

126. Li, B., Ritchie, M.D., 2021. From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries. Front. Genet. 12, 713230. https://doi.org/10.3389/fgene.2021.713230

127. Li, F., Yu, C., Zhao, Q., Wang, Zhaodi, Wang, Zhi, Chang, Y., *et al.*, 2024. Exploring the intestinal ecosystem: from gut microbiota to associations with subtypes of inflammatory bowel disease. Front. Cell. Infect. Microbiol. 13, 1304858. https://doi.org/10.3389/fcimb.2023.1304858

128. Li, S., Zhuge, A., Chen, H., Han, S., Shen, J., Wang, K., *et al.*, 2024. Sedanolide alleviates DSS-induced colitis by modulating the intestinal FXR-SMPD3 pathway in mice. Journal of Advanced Research S2090123224001280. https://doi.org/10.1016/j.jare.2024.03.026

129. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., *et al.*, 2007. Novel Crohn Disease Locus Identified by Genome-Wide Association Maps to a Gene Desert on 5p13.1 and Modulates Expression of PTGER4. PLoS Genet 3, e58. https://doi.org/10.1371/journal.pgen.0030058

130. Lichtenstein, G.R., Olson, A., Travers, S., Diamond, R.H., Chen, D.M., Pritchard, M.L., *et al.*, 2006. Factors Associated with the Development of Intestinal Strictures or

Obstructions in Patients with Crohn’s Disease. Am J Gastroenterology 101, 1030–1038. https://doi.org/10.1111/j.1572-0241.2006.00463.x

131. Liefferinckx, C., Verstockt, B., Gils, A., Noman, M., Van Kemseke, C., Macken, E., De Vos, M., Van Moerkercke, W., Rahier, J.-F., Bossuyt, P., Dutré, J., Humblet, E., Staessen, D., Peeters, H., Van Hootegem, P., Louis, E., Franchimont, D., Baert, F., Vermeire, S., Belgian Inflammatory Bowel Disease Research and Development Group [BIRD group], 2019. Long-term Clinical Effectiveness of Ustekinumab in Patients with Crohn’s Disease Who Failed Biologic Therapies: A National Cohort Study. Journal of Crohn’s and Colitis 13, 1401–1409. https://doi.org/10.1093/ecco-jcc/jjz080

132. Lin, Q., Tan, X., Wang, W., Zeng, W., Gui, L., Su, M., *et al.*, 2020. Species Differences of Bile Acid Redox Metabolism: Tertiary Oxidation of Deoxycholate is Conserved in Preclinical Animals. Drug Metabolism and Disposition 48, 499–507. https://doi.org/10.1124/dmd.120.090464

133. Liu, J.Z., Van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., *et al.*, 2015. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat Genet 47, 979–986. https://doi.org/10.1038/ng.3359

134. Liu, F., Li, X., Yue, H., Ji, J., You, M., Ding, L., *et al.*, 2017. TLR -Induced SMPD 3 Defects Enhance Inflammatory Response of B Cell and Macrophage in the Pathogenesis of SLE. Scand J Immunol 86, 377–388. https://doi.org/10.1111/sji.12611

135. Liu, D., Offin, M., Harnicar, S., Li, B.T., Drilon, A., 2018. Entrectinib: an orally available, selective tyrosine kinase inhibitor for the treatment of NTRK, ROS1, and ALK fusion-positive solid tumors. TCRM Volume 14, 1247–1252. https://doi.org/10.2147/TCRM.S147381

136. Liu, X., Li, Y.I., Pritchard, J.K., 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177, 1022-1034.e6. https://doi.org/10.1016/j.cell.2019.04.014

137. Liu, B., Ye, D., Yang, H., Song, J., Sun, X., Mao, Y., He, Z., 2022. Two-Sample Mendelian Randomization Analysis Investigates Causal Associations Between Gut Microbial Genera and Inflammatory Bowel Disease, and Specificity Causal Associations in Ulcerative Colitis or Crohn’s Disease. Front. Immunol. 13, 921546. https://doi.org/10.3389/fimmu.2022.921546

138. Liu, B., Qian, Y., Li, Y., Shen, X., Ye, D., Mao, Y., Sun, X., 2023. Circulating levels of cytokines and risk of inflammatory bowel disease: evidence from genetic data. Front. Immunol. 14, 1310086. https://doi.org/10.3389/fimmu.2023.1310086

139. Lloyd-Still, J.D., 1994. Crohn’s disease and cystic fibrosis. Digest Dis Sci 39, 880–885. https://doi.org/10.1007/BF02087437

140. Loftus, E.V., 2004. Clinical epidemiology of inflammatory bowel disease: incidence, prevalence, and environmental influences. Gastroenterology 126, 1504–1517. https://doi.org/10.1053/j.gastro.2004.01.063

141. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., *et al.*, 2013. The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580–585. https://doi.org/10.1038/ng.2653

142. Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550. https://doi.org/10.1186/s13059-014-0550-8

143. Lund-Nielsen, J., Vedel-Krogh, S., Kobylecki, C.J., Brynskov, J., Afzal, S., Nordestgaard, B.G., 2018. Vitamin D and Inflammatory Bowel Disease: Mendelian Randomization Analyses in the Copenhagen Studies and UK Biobank. The Journal of Clinical Endocrinology & Metabolism 103, 3267–3277. https://doi.org/10.1210/jc.2018-00250

144. MacArthur, R.D., Hawkins, T.N., Brown, S.J., LaMarca, A., Clay, P.G., Barrett, A.C., *et al.*, 2013. Efficacy and Safety of Crofelemer for Noninfectious Diarrhea in HIV-Seropositive Individuals (ADVENT Trial): A Randomized, Double-Blind, Placebo-Controlled, Two-Stage Study. HIV Clinical Trials 14, 261–273. https://doi.org/10.1310/hct1406-261

145. Magg, T., Shcherbina, A., Arslan, D., Desai, M.M., Wall, S., Mitsialis, V., *et al.*, 2019. CARMIL2 Deficiency Presenting as Very Early Onset Inflammatory Bowel Disease. Inflammatory Bowel Diseases 25, 1788–1795. https://doi.org/10.1093/ibd/izz103

146. Mai, J., Lu, M., Gao, Q., Zeng, J., Xiao, J., 2023. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol 6, 899. https://doi.org/10.1038/s42003-023-05279-y

147. Major, E.O., 2010. Progressive Multifocal Leukoencephalopathy in Patients on Immunomodulatory Therapies. Annu. Rev. Med. 61, 35–47. https://doi.org/10.1146/annurev.med.080708.082655

148. Marchini, J., Howie, B., 2010. Genotype imputation for genome-wide association studies. Nat Rev Genet 11, 499–511. https://doi.org/10.1038/nrg2796

149. Mastrapasqua, M., Rossi, R., De Cosmo, L., Resta, A., Errede, M., Bizzoca, A., *et al.*, 2023. Autophagy increase in Merosin-Deficient Congenital Muscular Dystrophy type 1A. Eur J Transl Myol 33. https://doi.org/10.4081/ejtm.2023.11501

150. McGinnis, C.S., Patterson, D.M., Winkler, J., Conrad, D.N., Hein, M.Y., Srivastava, V., *et al.*, 2019. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. Nat Methods 16, 619–626. https://doi.org/10.1038/s41592-019-0433-8

151. McGovern, D.P.B., Jones, M.R., Taylor, K.D., Marciante, K., Yan, X., Dubinsky, M., *et al.*, 2010. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn’s disease. Human Molecular Genetics 19, 3468–3476. https://doi.org/10.1093/hmg/ddq248

152. Mentella, M.C., Scaldaferri, F., Pizzoferrato, M., Gasbarrini, A., Miggiano, G.A.D., 2020. Nutrition, IBD and Gut Microbiota: A Review. Nutrients 12, 944. https://doi.org/10.3390/nu12040944

153. Miao, Z., Alvarez, M., Pajukanta, P., Ko, A., 2018. ASElux: an ultra-fast and accurate allelic reads counter. Bioinformatics 34, 1313–1320. https://doi.org/10.1093/bioinformatics/btx762

154. Minikel, E.V., Painter, J.L., Dong, C.C., Nelson, M.R., 2024. Refining the impact of genetic evidence on clinical success. Nature 629, 624–629. https://doi.org/10.1038/s41586-024-07316-0

155. Mishra, N., Aden, K., Blase, J.I., Baran, N., Bordoni, D., Tran, F., *et al.*, 2022. Longitudinal multi-omics analysis identifies early blood-based predictors of anti-TNF therapy response in inflammatory bowel disease. Genome Med 14, 110. https://doi.org/10.1186/s13073-022-01112-z

156. Mitropoulou, M.-A., Fradelos, E.C., Lee, K.Y., Malli, F., Tsaras, K., Christodoulou, N.G., Papathanasiou, I.V., 2022. Quality of Life in Patients With Inflammatory Bowel Disease: Importance of Psychological Symptoms. Cureus. https://doi.org/10.7759/cureus.28502

157. Mohanan, V., Nakata, T., Desch, A.N., Lévesque, C., Boroughs, A., Guzman, G., *et al.*, 2018. *C1orf106* is a colitis risk gene that regulates stability of epithelial adherens junctions. Science 359, 1161–1166. https://doi.org/10.1126/science.aan0814

158. Mokry, L.E., Zhou, S., Guo, C., Scott, R.A., Devey, L., Langenberg, C., *et al.*, 2019. Interleukin-18 as a drug repositioning opportunity for inflammatory bowel disease: A Mendelian randomization study. Sci Rep 9, 9386. https://doi.org/10.1038/s41598-019-45747-2

159. Molodecky, N.A., Soon, I.S., Rabi, D.M., Ghali, W.A., Ferris, M., Chernoff, G., *et al.*, 2012. Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review. Gastroenterology 142, 46-54.e42. https://doi.org/10.1053/j.gastro.2011.10.001

160. Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., *et al.*, 2011. Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat Genet 43, 43–47. https://doi.org/10.1038/ng.733

161. Momozawa, Y., Dmitrieva, J., Théâtre, E., Deffontaine, V., Rahmouni, S., Charloteaux, B., *et al.*, 2018. IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes. Nat Commun 9, 2427. https://doi.org/10.1038/s41467-018-04365-8

162. Mostafavi, H., Spence, J.P., Naqvi, S., Pritchard, J.K., 2023. Systematic differences in discovery of genetic effects on gene expression and complex traits. Nat Genet 55, 1866–1875. https://doi.org/10.1038/s41588-023-01529-1

163. Mountjoy, E., Schmidt, E.M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., *et al.*, 2021. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat Genet 53, 1527–1533. https://doi.org/10.1038/s41588-021-00945-5

164. Muller, M., D’Amico, F., Bonovas, S., Danese, S., Peyrin-Biroulet, L., 2021. TNF Inhibitors and Risk of Malignancy in Patients with Inflammatory Bowel Diseases: A Systematic Review. Journal of Crohn’s and Colitis 15, 840–859. https://doi.org/10.1093/ecco-jcc/jjaa186

165. Nakaoka, H., Inoue, I., 2009. Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner’s curse. J Hum Genet 54, 615–623. https://doi.org/10.1038/jhg.2009.95

166. Neurath, M.F., 2014. Cytokines in inflammatory bowel disease. Nat Rev Immunol 14, 329–342. https://doi.org/10.1038/nri3661

167. Ng, S.C., Shi, H.Y., Hamidi, N., Underwood, F.E., Tang, W., Benchimol, E.I., *et al.*,

2017. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. The Lancet 390, 2769–2778. https://doi.org/10.1016/S0140-6736(17)32448-0

168. Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., Dermitzakis, E.T., 2010. Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet 6, e1000895. https://doi.org/10.1371/journal.pgen.1000895

169. Nica, A.C., Dermitzakis, E.T., 2013. Expression quantitative trait loci: present and future. Phil. Trans. R. Soc. B 368, 20120362. https://doi.org/10.1098/rstb.2012.0362

170. Nicolae, D.L., 2016. Association Tests for Rare Variants. Annu. Rev. Genom. Hum. Genet. 17, 117–130. https://doi.org/10.1146/annurev-genom-083115-022609

171. Noah, T.K., Donahue, B., Shroyer, N.F., 2011. Intestinal development and differentiation. Experimental Cell Research 317, 2702–2710. https://doi.org/10.1016/j.yexcr.2011.09.006

172. Norkina, O., Burnett, T.G., De Lisle, R.C., 2004. Bacterial Overgrowth in the Cystic Fibrosis Transmembrane Conductance Regulator Null Mouse Small Intestine. Infect Immun 72, 6040–6049. https://doi.org/10.1128/IAI.72.10.6040-6049.2004

173. Ogura, Y., Bonen, D.K., Inohara, N., Nicolae, D.L., Chen, F.F., Ramos, R., *et al.*, 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature 411, 603–606. https://doi.org/10.1038/35079114

174. Okabayashi, S., Kobayashi, T., Saito, E., Toyonaga, T., Ozaki, R., Sagami, S., *et al.*, 2019. Individualized treatment based on CYP3A5 single-nucleotide polymorphisms with tacrolimus in ulcerative colitis. Intest Res 17, 218–226. https://doi.org/10.5217/ir.2018.00117

175. Okai, N., Masuta, Y., Otsuka, Y., Hara, A., Masaki, S., Kamata, K., Minaga, K., Honjo, H., Kudo, M., Watanabe, T., 2024. Crosstalk between NOD2 and TLR2 suppresses the development of TLR2-mediated experimental colitis. J. Clin. Biochem. Nutr. 74, 146–153. https://doi.org/10.3164/jcbn.23-87

176. Orholm, M., Munkholm, P., Langholz, E., Nielsen, O.H., Sørensen, T.I.A., Binder, V., 1991. Familial Occurrence of Inflammatory Bowel Disease. N Engl J Med 324, 84–88. https://doi.org/10.1056/NEJM199101103240203

177. Ota, M., Nagafuchi, Y., Hatano, H., Ishigaki, K., Terao, C., Takeshima, Y., *et al.*, 2021. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell 184, 3006-3021.e17. https://doi.org/10.1016/j.cell.2021.03.056

178. Pai, A.A., Cain, C.E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.-B., *et al.*, 2012. The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels. PLoS Genet 8, e1003000. https://doi.org/10.1371/journal.pgen.1003000

179. Palmer, C., Pe'er, I., 2017. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLoS Genet 13, e1006916. https://doi.org/10.1371/journal.pgen.1006916

180. Palmer, D.S., Zhou, W., Abbott, L., Wigdor, E.M., Baya, N., Churchhouse, C., *et al.*, 2023. Analysis of genetic dominance in the UK Biobank. Science 379,

1341–1348. https://doi.org/10.1126/science.abn8455

181. Paramsothy, S., Kamm, M.A., Kaakoush, N.O., Walsh, A.J., Van Den Bogaerde, J., Samuel, D., *et al.*, 2017. Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. The Lancet 389, 1218–1228. https://doi.org/10.1016/S0140-6736(17)30182-4

182. Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M., Anderson, C.A., Fisher, S.A., *et al.*, 2007. Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39, 830–832. https://doi.org/10.1038/ng2061

183. Pauls, S.D., Marshall, A.J., 2017. Regulation of immune cell signaling by SHIP1: A phosphatase, scaffold protein, and potential therapeutic target. Eur J Immunol 47, 932–945. https://doi.org/10.1002/eji.201646795

184. Pedicone, C., Meyer, S.T., Chisholm, J.D., Kerr, W.G., 2021. Targeting SHIP1 and SHIP2 in Cancer. Cancers 13, 890. https://doi.org/10.3390/cancers13040890

185. Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.-Y., Popejoy, A.B., *et al.*, 2019. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell 179, 589–603. https://doi.org/10.1016/j.cell.2019.08.051

186. Picardo, S., So, K., Venugopal, K., 2020. Anti-TNF-induced lupus in patients with inflammatory bowel disease. JGH Open 4, 507–510. https://doi.org/10.1002/jgh3.12291

187. Pier, G.B., Grout, M., Zaidi, T., Meluleni, G., Mueschenborn, S.S., Banting, G., *et al.*, 1998. Salmonella typhi uses CFTR to enter intestinal epithelial cells. Nature 393, 79–82. https://doi.org/10.1038/30006

188. Pierre, N., Salée, C., Vieujean, S., Bequet, E., Merli, A., Siegmund, B., *et al.*, 2021. Review article: distinctions between ileal and colonic Crohn's disease: from physiology to pathology. Aliment Pharmacol Ther 54, 779–791. https://doi.org/10.1111/apt.16536

189. Pingault, J.-B., O'Reilly, P.F., Schoeler, T., Ploubidis, G.B., Rijsdijk, F., Dudbridge, F., 2018. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet 19, 566–580. https://doi.org/10.1038/s41576-018-0020-3

190. Piovani, D., Danese, S., Peyrin-Biroulet, L., Nikolopoulos, G.K., Lytras, T., Bonovas, S., 2019. Environmental Risk Factors for Inflammatory Bowel Diseases: An Umbrella Review of Meta-analyses. Gastroenterology 157, 647-659.e4. https://doi.org/10.1053/j.gastro.2019.04.016

191. Prieto, S., Dubra, G., Camasses, A., Aznar, A.B., Begon-Pescia, C., Simboeck, E., *et al.*, 2023. CDK8 and CDK19 act redundantly to control the CFTR pathway in the intestinal epithelium. EMBO Reports 24, e54261. https://doi.org/10.15252/embr.202154261

192. Pritchard, J.K., 2002. The allelic architecture of human disease genes: common disease-common variant... or not? Human Molecular Genetics 11, 2417–2423. https://doi.org/10.1093/hmg/11.20.2417

193. Pu, Z., Che, Y., Zhang, W., Sun, H., Meng, T., Xie, H., *et al.*, 2019. Dual roles of IL-18 in colitis through regulation of the function and quantity of goblet cells. Int J Mol

Med. https://doi.org/10.3892/ijmm.2019.4156

194. Qian, M., Shao, Y., 2013. A Likelihood Ratio Test for Genome‐Wide Association under Genetic Heterogeneity. Annals of Human Genetics 77, 174–182. https://doi.org/10.1111/ahg.12005

195. Raab, J.R., Kamakaka, R.T., 2010. Insulators and promoters: closer than we think. Nat Rev Genet 11, 439–446. https://doi.org/10.1038/nrg2765

196. Ramos, G.P., Papadakis, K.A., 2019. Mechanisms of Disease: Inflammatory Bowel Diseases. Mayo Clinic Proceedings 94, 155–165. https://doi.org/10.1016/j.mayocp.2018.09.013

197. Ricciotti, E., FitzGerald, G.A., 2011. Prostaglandins and Inflammation. ATVB 31, 986–1000. https://doi.org/10.1161/ATVBAHA.110.207449

198. Richmond, R.C., Davey Smith, G., 2022. Mendelian Randomization: Concepts and Scope. Cold Spring Harb Perspect Med 12, a040501. https://doi.org/10.1101/cshperspect.a040501

199. Risch, N., Merikangas, K., 1996. The Future of Genetic Studies of Complex Human Diseases. Science 273, 1516–1517. https://doi.org/10.1126/science.273.5281.1516

200. Roda, G., Jharap, B., Neeraj, N., Colombel, J.-F., 2016. Loss of Response to Anti-TNFs: Definition, Epidemiology, and Management. Clinical and Translational Gastroenterology 7, e135. https://doi.org/10.1038/ctg.2015.63

201. Rojano, E., Seoane, P., Ranea, J.A.G., Perkins, J.R., 2019. Regulatory variants: from detection to predicting impact. Briefings in Bioinformatics 20, 1639–1654. https://doi.org/10.1093/bib/bby039

202. Roncagalli, R., Cucchetti, M., Jarmuzynski, N., Grégoire, C., Bergot, E., Audebert, S., *et al.*, 2016. The scaffolding function of the RLTPR protein explains its essential role for CD28 co-stimulation in mouse and human T cells. Journal of Experimental Medicine 213, 2437–2457. https://doi.org/10.1084/jem.20160579

203. Sazonovs, A., Stevens, C.R., Venkataraman, G.R., Yuan, K., Avila, B., Abreu, M.T., *et al.*, 2022. Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. Nat Genet 54, 1275–1283. https://doi.org/10.1038/s41588-022-01156-2

204. Schéele, S., Nyström, A., Durbeej, M., Talts, J.F., Ekblom, M., Ekblom, P., 2007. Laminin isoforms in development and disease. J Mol Med 85, 825–836. https://doi.org/10.1007/s00109-007-0182-5

205. Schmiedel, B.J., Singh, D., Madrigal, A., Valdovino-Gonzalez, A.G., White, B.M., Zapardiel-Gonzalo, J., *et al.*, 2018. Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. Cell 175, 1701-1715.e16. https://doi.org/10.1016/j.cell.2018.10.022

206. Schroeder, T.H., Lee, M.M., Yacono, P.W., Cannon, C.L., Gerçeker, A.A., Golan, D.E., Pier, G.B., 2002. CFTR is a pattern recognition molecule that extracts *Pseudomonas aeruginosa* LPS from the outer membrane into epithelial cells and activates NF-κB translocation. Proc. Natl. Acad. Sci. U.S.A. 99, 6907–6912. https://doi.org/10.1073/pnas.092160899

207. Shi, H., Wang, Y., Li, X., Zhan, X., Tang, M., Fina, M., *et al.*, 2016. NLRP3

activation and mitosis are mutually exclusive events coordinated by NEK7, a new inflammasome component. Nat Immunol 17, 250–258. https://doi.org/10.1038/ni.3333

208. Shin, Y., Han, S., Kwon, J., Ju, S., Choi, T., Kang, I., Kim, S., 2023. Roles of Short-Chain Fatty Acids in Inflammatory Bowel Disease. Nutrients 15, 4466. https://doi.org/10.3390/nu15204466

209. Sillanpää, M.J., 2011. Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. Heredity 106, 511–519. https://doi.org/10.1038/hdy.2010.91

210. Silva, F.A.R., Rodrigues, B.L., Ayrizono, M.D.L.S., Leal, R.F., 2016. The Immunological Basis of Inflammatory Bowel Disease. Gastroenterology Research and Practice 2016, 1–11. https://doi.org/10.1155/2016/2097274

211. Smillie, C.S., Biton, M., Ordovas-Montanes, J., Sullivan, K.M., Burgin, G., Graham, D.B., *et al.*, 2019. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell 178, 714-730.e22. https://doi.org/10.1016/j.cell.2019.06.029

212. Song, Z., Li, X., Xie, J., Han, F., Wang, N., Hou, Y., Yao, J., 2024. Associations of inflammatory cytokines with inflammatory bowel disease: a Mendelian randomization study. Front. Immunol. 14, 1327879. https://doi.org/10.3389/fimmu.2023.1327879

213. Stegle, O., Parts, L., Durbin, R., Winn, J., 2010. 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. https://doi.org/10.1371/journal.pcbi.1000770

214. Stegle, O., Parts, L., Piipari, M., Winn, J., Durbin, R., 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc 7, 500–507. https://doi.org/10.1038/nprot.2011.457

215. Steinmetz, L.M., Sinha, H., Richards, D.R., Spiegelman, J.I., Oefner, P.J., McCusker, J.H., Davis, R.W., 2002. Dissecting the architecture of a quantitative trait locus in yeast. Nature 416, 326–330. https://doi.org/10.1038/416326a

216. Stern, D.L., 1998. A role of Ultrabithorax in morphological differences between Drosophila species. Nature 396, 463–466. https://doi.org/10.1038/24863

217. Stern, D.L., 2014. Identification of loci that cause phenotypic variation in diverse species with the reciprocal hemizygosity test. Trends in Genetics 30, 547–554. https://doi.org/10.1016/j.tig.2014.09.006

218. Stitham, J., Midgett, C., Martin, K.A., Hwa, J., 2011. Prostacyclin: An Inflammatory Paradox. Front. Pharmacol. 2. https://doi.org/10.3389/fphar.2011.00024

219. Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., *et al.*, 2018. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol 19, 224. https://doi.org/10.1186/s13059-018-1603-1

220. Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. U.S.A. 100, 9440–9445. https://doi.org/10.1073/pnas.1530509100

221. Sylvestre, M., Di Carlo, S.E., Peduto, L., 2023. Stromal regulation of the intestinal barrier. Mucosal Immunology 16, 221–231. https://doi.org/10.1016/j.mucimm.2023.01.006

222. Taliun, D., Harris, D.N., Kessler, M.D., Carlson, J., Szpiech, Z.A., Torres, R., *et al.*, 2021. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299. https://doi.org/10.1038/s41586-021-03205-y

223. Taniguchi, K., Inoue, M., Arai, K., Uchida, K., Migita, O., Akemoto, Y., *et al.*, 2021. Novel TNFAIP3 microdeletion in a girl with infantile-onset inflammatory bowel disease complicated by a severe perianal lesion. Hum Genome Var 8, 1. https://doi.org/10.1038/s41439-020-00128-4

224. Tavares De Sousa, H., Magro, F., 2023. How to Evaluate Fibrosis in IBD? Diagnostics 13, 2188. https://doi.org/10.3390/diagnostics13132188

225. The 1000 Genomes Project Consortium, 2010. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073. https://doi.org/10.1038/nature09534

226. The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65. https://doi.org/10.1038/nature11632

227. The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature 526, 68–74. https://doi.org/10.1038/nature15393

228. The Geuvadis Consortium, Lappalainen, T., Sammeth, M., Friedländer, M.R., ‘T Hoen, P.A.C., Monlong, J., *et al.*, 2013. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511. https://doi.org/10.1038/nature12531

229. The GTEx Consortium, Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., *et al.*, 2020. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318–1330. https://doi.org/10.1126/science.aaz1776

230. The International HapMap Consortium, 2003. The International HapMap Project. Nature 426, 789–796. https://doi.org/10.1038/nature02168

231. The International HapMap Consortium, 2005. A haplotype map of the human genome. Nature 437, 1299–1320. https://doi.org/10.1038/nature04226

232. The International HapMap Consortium, 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449, 851–861. https://doi.org/10.1038/nature06258

233. The International HapMap 3 Consortium, 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467, 52–58. https://doi.org/10.1038/nature09298

234. Thompson, J.S., 2000. Comparison of massive vs. repeated resection leading to short bowel syndrome. Journal of Gastrointestinal Surgery 4, 101–104. https://doi.org/10.1016/S1091-255X(00)80039-6

235. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., *et al.*, 2012. The accessible chromatin landscape of the human genome. Nature 489, 75–82. https://doi.org/10.1038/nature11232

236. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., *et al.*, 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 352, 189–196. https://doi.org/10.1126/science.aad0501

237. Trajanoska, K., Bhérer, C., Taliun, D., Zhou, S., Richards, J.B., Mooser, V., 2023. From target discovery to clinical drug development with human genetics. Nature 620, 737–745. https://doi.org/10.1038/s41586-023-06388-8

238. Turner, J.R., 2009. Intestinal mucosal barrier function in health and disease. Nat Rev Immunol 9, 799–809. https://doi.org/10.1038/nri2653

239. Tysk, C., Lindberg, E., Jarnerot, G., Floderus-Myrhed, B., 1988. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. Gut 29, 990–996. https://doi.org/10.1136/gut.29.7.990

240. Uchida, G., Nakamura, M., Yamamura, T., Tsuzuki, T., Kawashima, H., 2023. Real-world effectiveness of ustekinumab for patients with ulcerative colitis: a systematic review and meta-analysis. https://doi.org/10.18999/nagjms.85.3.402

241. Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., *et al.*, 2021. Genome-wide association studies. Nat Rev Methods Primers 1, 59. https://doi.org/10.1038/s43586-021-00056-9

242. Underhill, D.M., Braun, J., 2022. Fungal microbiome in inflammatory bowel disease: a critical assessment. Journal of Clinical Investigation 132, e155786. https://doi.org/10.1172/JCI155786

243. Van De Bunt, M., Cortes, A., IGAS Consortium, Brown, M.A., Morris, A.P., McCarthy, M.I., 2015. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. PLoS Genet 11, e1005535. https://doi.org/10.1371/journal.pgen.1005535

244. Van De Geijn, B., McVicker, G., Gilad, Y., Pritchard, J.K., 2015. WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nat Methods 12, 1061–1063. https://doi.org/10.1038/nmeth.3582

245. Vieujean, S., Jairath, V., Peyrin-Biroulet, L., Dubinsky, M., Iacucci, M., Magro, F., Danese, S., 2025. Understanding the therapeutic toolkit for inflammatory bowel disease. Nat Rev Gastroenterol Hepatol. https://doi.org/10.1038/s41575-024-01035-7

246. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., *et al.*, 2021. 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. https://doi.org/10.1038/s41588-021-00913-z

247. Wallace, K.L., 2014. Immunopathology of inflammatory bowel disease. WJG 20, 6. https://doi.org/10.3748/wjg.v20.i1.6

248. Wallace, C., 2020. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet 16, e1008720. https://doi.org/10.1371/journal.pgen.1008720

249. Wallace, C., 2021. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet 17, e1009440. https://doi.org/10.1371/journal.pgen.1009440

250. Wang, T., Liu, Y., Ruan, J., Dong, X., Wang, Y., Peng, J., 2021. A pipeline for

RNA-seq based eQTL analysis with automated quality control procedures. BMC Bioinformatics 22, 403. https://doi.org/10.1186/s12859-021-04307-0

251. Wang, Z., Shu, Q., Wu, J., Cheng, Y., Liang, X., Huang, X., *et al.*, 2024. Evaluating the association between immunological proteins and common intestinal diseases using a bidirectional two-sample Mendelian randomization study. Cytokine 184, 156788. https://doi.org/10.1016/j.cyto.2024.156788

252. Watanabe, K., Stringer, S., Frei, O., Umićević Mirkov, M., De Leeuw, C., Polderman, T.J.C., *et al.*, 2019. A global overview of pleiotropy and genetic architecture in complex traits. Nat Genet 51, 1339–1348. https://doi.org/10.1038/s41588-019-0481-0

253. Winger, B.A., Foy, E., Sud, S.R., MacKenzie, J.D., Pua, H.H., Lau, A.H., *et al.*, 2016. *Mycobacterium bovis* Enterocolitis in an Immunocompromised Host. J. pediatr. gastroenterol. nutr. 63. https://doi.org/10.1097/MPG.0000000000000528

254. Wray, N.R., Wijmenga, C., Sullivan, P.F., Yang, J., Visscher, P.M., 2018. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell 173, 1573–1580. https://doi.org/10.1016/j.cell.2018.05.051

255. Wray, N.R., Lin, T., Austin, J., McGrath, J.J., Hickie, I.B., Murray, G.K., Visscher, P.M., 2021. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. JAMA Psychiatry 78, 101. https://doi.org/10.1001/jamapsychiatry.2020.3049

256. Wu, C., DeWan, A., Hoh, J., Wang, Z., 2011. A Comparison of Association Methods Correcting for Population Stratification in Case–Control Studies. Annals of Human Genetics 75, 418–427. https://doi.org/10.1111/j.1469-1809.2010.00639.x

257. Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., *et al.*, 2018. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun 9, 918. https://doi.org/10.1038/s41467-018-03371-0

258. Xu, X.-R., 2014. Dysregulation of mucosal immune response in pathogenesis of inflammatory bowel disease. WJG 20, 3255. https://doi.org/10.3748/wjg.v20.i12.3255

259. Xu, W.-D., Xie, Q.-B., Zhao, Y., Liu, Y., 2015. Association of Interleukin-23 receptor gene polymorphisms with susceptibility to Crohn's disease: A meta-analysis. Sci Rep 5, 18584. https://doi.org/10.1038/srep18584

260. Xu, Y., Yan, Z., Liu, L., 2024. Identification of novel proteins in inflammatory bowel disease based on the gut-brain axis: a multi-omics integrated analysis. Clin Proteom 21, 59. https://doi.org/10.1186/s12014-024-09511-7

261. Xu, X., Han, Y., Deng, J., Wang, S., Zhuo, S., Zhao, K., Zhou, W., 2024. Repurposing disulfiram with CuET nanocrystals: Enhancing anti-pyroptotic effect through NLRP3 inflammasome inhibition for treating inflammatory bowel diseases. Acta Pharmaceutica Sinica B 14, 2698–2715. https://doi.org/10.1016/j.apsb.2024.03.003

262. Yalcin, B., Willis-Owen, S.A.G., Fullerton, J., Meesaq, A., Deacon, R.M., Rawlins, J.N.P., *et al.*, 2004. Genetic dissection of a behavioral quantitative trait locus shows that Rgs2 modulates anxiety in mice. Nat Genet 36, 1197–1202. https://doi.org/10.1038/ng1450

263. Yamamoto, Y., Nakase, H., Matsuura, M., Maruyama, S., Masuda, S., 2020. CYP3A5 Genotype as a Potential Pharmacodynamic Biomarker for Tacrolimus

Therapy in Ulcerative Colitis in Japanese Patients. IJMS 21, 4347. https://doi.org/10.3390/ijms21124347

264. Yamazaki, K., Umeno, J., Takahashi, A., Hirano, A., Johnson, T.A., Kumasaka, N., *et al.*, 2013. A Genome-Wide Association Study Identifies 2 Susceptibility Loci for Crohn’s Disease in a Japanese Population. Gastroenterology 144, 781–788. https://doi.org/10.1053/j.gastro.2012.12.021

265. Yan, J., Hedl, M., Abraham, C., 2017. An inflammatory bowel disease–risk variant in INAVA decreases pattern recognition receptor–induced outcomes. Journal of Clinical Investigation 127, 2192–2205. https://doi.org/10.1172/JCI86282

266. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., *et al.*, 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42, 565–569. https://doi.org/10.1038/ng.608

267. Yang, S.-K., Hong, M., Zhao, W., Jung, Y., Baek, J., Tayebi, N., *et al.*, 2014. Genome-wide association study of Crohn’s disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut 63, 80–87. https://doi.org/10.1136/gutjnl-2013-305193

268. Yazar, S., Alquicira-Hernandez, J., Wing, K., Senabouth, A., Gordon, M.G., Andersen, S., *et al.*, 2022. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. Science 376, eabf3041. https://doi.org/10.1126/science.abf3041

269. Yıldız, Ç., Gezgin Yıldırım, D., Inci, A., Tümer, L., Cengiz Ergin, F.B., Sunar Yayla, E.N.S., *et al.*, 2023. A possibly new autoinflammatory disease due to compound heterozygous phosphomevalonate kinase gene mutation. Joint Bone Spine 90, 105490. https://doi.org/10.1016/j.jbspin.2022.105490

270. Yu, M., Zhang, Q., Yuan, K., Sazonovs, A., Stevens, C., Fachal, L., *et al.*, 2024. Cystic fibrosis risk variants confer protection against inflammatory bowel disease. https://doi.org/10.1101/2024.12.02.24318364

271. Zhang, Y.-Z., 2014. Inflammatory bowel disease: Pathogenesis. WJG 20, 91. https://doi.org/10.3748/wjg.v20.i1.91

272. Zhang, J., Gao, L.-Z., Chen, Y.-J., Zhu, P.-P., Yin, S.-S., Su, M.-M., *et al.*, 2019. Continuum of Host-Gut Microbial Co-metabolism: Host CYP3A4/3A7 are Responsible for Tertiary Oxidations of Deoxycholate Species. Drug Metabolism and Disposition 47, 283–294. https://doi.org/10.1124/dmd.118.085670

273. Zhang, J., Zhao, H., 2023. eQTL studies: from bulk tissues to single cells. Journal of Genetics and Genomics 50, 925–933. https://doi.org/10.1016/j.jgg.2023.05.003

274. Zhao, Y., Li, M.-C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., *et al.*, 2021. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. J Transl Med 19, 269. https://doi.org/10.1186/s12967-021-02936-w

275. Zhao, S., Crouse, W., Qian, S., Luo, K., Stephens, M., He, X., 2024. Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. Nat Genet 56, 336–347. https://doi.org/10.1038/s41588-023-01648-9

276. Zhou, H.J., Li, L., Li, Y., Li, W., Li, J.J., 2022. PCA outperforms popular hidden

variable inference methods for molecular QTL mapping. Genome Biol 23, 210. https://doi.org/10.1186/s13059-022-02761-4

277. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., *et al.*, 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet 48, 481–487. https://doi.org/10.1038/ng.3538

278. Zhu, A., Ibrahim, J.G., Love, M.I., 2019. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics 35, 2084–2092. https://doi.org/10.1093/bioinformatics/bty895

279. Zou, D., Zhou, S., Wang, H., Gou, J., Wang, S., 2020. Knee Joint Swelling at Presentation: A Case of Pediatric Crohn Disease With a TNFAIP3 Mutation. Pediatrics 146, e20193416. https://doi.org/10.1542/peds.2019-3416

280. Zuber, V., Grinberg, N.F., Gill, D., Manipur, I., Slob, E.A.W., Patel, A., *et al.*, 2022. Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. The American Journal of Human Genetics 109, 767–782. https://doi.org/10.1016/j.ajhg.2022.04.001